BIOSTATISTICS

For Health Students
With Manual on Software Applications

(Second Edition)

Muhammad Hanif
Munir Ahmad
Ezz H. Abdelfattah

BIOSTATISTICS
For Health Students
With Manual on Software Applications

Muhammad Hanif
Ph.D.

Munir Ahmad
Ph.D.

National College of Business Administration & Economics
Lahore, Pakistan

and

Ezz H. Abdelfattah
Ph.D.
Statistics Department, Faculty of Science
King Abdul Aziz University
21589, Jeddah 80203, Saudi Arabia

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ ۞

وَلَا تَقُوْلَنَّ لِشَأْىْءٍ اِنِّىْ فَاعِلٌ ذٰلِكَ غَدًا ۞

اِلَّاۤ اَنْ يَّشَآءَ اللّٰهُ ۚ وَاذْكُرْ رَّبَّكَ اِذَا نَسِيْتَ وَقُلْ

عَسٰۤى اَنْ يَّهْدِيَنِ رَبِّيْ لِاَقْرَبَ مِنْ هٰذَا رَشَدًا ۞

سُوْرَةُ الْكَهْفِ (٢٣ - ٢٤)

اور نہ کہنا کسی کام کو کہ میں یہ کل کرونگا۔ مگر یہ کہ اللہ چاہے تو اور یاد کر لے اپنے رب کو جب بھول جائے اور کہہ دو کہ امید ہے میرا رب مجھ کو اس سے زیادہ نزدیک نیکی کی راہ دکھلائے۔

سورة الکھف (23-24)

No say anything "I shall be sure to do so and so tomorrow", except "if ALLAH so wills"⊗ And remember your Lord when you forget [it] and say, "Perhaps my Lord will guide me to what is nearer than this to right conduct.⊗

Surat Al-Kahf (23-24)

FOREWORD

When I was a doctorate student at Johns Hopkins School of Public Health. I used to take Biostatistics as a course, which I have to accept and live with it. I did not have much of a problem with it, but I could have enjoyed it more if it were presented to me in more attractive way. I mean in relation to real life rather than abstracts of figures. With this innovative writing of Prof. Hanif and Prof. Ahmad, I can see that the science of numbers and ratios is being wisely integrated with epidemiology.

Through feedback from the learners, I am sure that more will be added to this healthy relation between Biostatistics and other medical and public health sciences.

Prof. Zohair Sebai
Saudi Arabia

مقــدمه

عندما كن اطلب العلم فى كلية الصحة العامة بجامعة جونز هو بكنزيامريكا، كنت مضطر الى ان اتقبل علم الاحصاء كمادة تفيله تتحدث عن الارقام بلهجة جافة، ولم يكن لدى خيار فى ذلك. اى نعم لم اكن اجد صعوبة فى دراستها ولكن كم تمنيت ان تعطى لى بشكل افضل، اى ان يكون فيها حياة اكثر مع هذه الكتابة المبدعة من البر وفسور حنيف والبر و فسور احمد اجدان مادة الاحصاء اكتسبت حياة بفضل و صلها بعلم الابيد ميولوجى. وانلوائق من ان ردود الافعال من الدار سين سوف تضفى عليها حياة اكثر و تجعلها اكثر صلة بعلوم الطلب والصحة العامة.

الاستاذ زهير اسباعى

PREFACE TO SECONED EDITION

In this Edition the analysis of statistical data have been done on the basis of IBM 22 SPSS Package. In logistic regression (Chapter 9) basic concept with analysis of ordinal logistic regression and multinomial logistic regression have been added. A new Chapter of survival analysis is included as Chapter 10. The previous Chapter 10 (Reliability Coefficient) from the old addition is now Chapter 11. We are thankful to Dr. Nadeem Shafique Butt of COMSATS Institute of Information Technology, Lahore for the addition of new material in this Edition. We are also thankful to Mr. M. Imtiaz and M. Iftikhar of Islamic Countries Society of Statistical Sciences (ISOSS) for excellent typesetting of this book.

<div align="right">

Muhammad Hanif
Munir Ahmad
Ezz H. Abdelfattah

</div>

PREFACE

The use of statistical techniques of data analysis has been observed to have dramatically increased recently, particularly for application in the biomedical and social sciences. This may be partially attributed to the developments during the last few decades of sophisticated methods for analyzing quantitative and categorical data. It also reflects the increasing methodological sophistication of scientists and applied statisticians. The Islamic Educational Scientific and Cultural Organization (ISESCO) realized that the knowledge of these statistical methods in health and medical research as well as in clinical practice was very important for dealing with uncertainty in diagnosis, treatment and prognosis. Moreover these methods are useful for health professionals, since they have to evaluate their day-to-day clinical data and research material. Such statistical analyses could improve their understanding and skills for treatment of patients, as well as planning, implementation and evaluation of health programs. Considering all these reasons, ISESCO formed a committee headed by Dr. Munir Ahmad in 1993 to develop a curriculum regarding Bio-statistics for medical colleges in the Islamic Countries. The senior author was also member of this committee. The curriculum was developed and circulated among the medical colleges of the Islamic Countries. Most of the Islamic Countries sent their comments and suggestions, which were incorporated in the curriculum before approval. Then we decided to write this manual for the medical, health and social sciences students. This is a self-reading manual written in a simple language, which can easily be comprehended and could be of use for health related and social studies, both at the undergraduate and postgraduate levels.

This manual consists of 10 chapters and presents the most important methods for analyzing quantitative and categorical data. It summarizes methods that have long played a prominent role, such as parametric and non-parametric tests; linear regression, chi square tests and measures of association including the tests of significance of relative risk, odds ratio and Mental-Haenszel odds ratio. A chapter on various types of sampling techniques and estimation of sample size has been added which is normally not included in common books on Bio-statistics. Various methods of reliability co-efficient with applications have been put together to facilitate the research workers. This manual puts special emphasis on logistic regression, a newly developed technique for qualitative data analysis. Another feature of this manual is that one can easily understand and use SPSS (Statistical Package for Social Sciences) software. Much emphasis has been given to the ability to select an appropriate test for the analysis of data with medical interpretation in the context of the problem.

The technical components of the manual have been explained in a way that does not require familiarity with mathematics such as calculus and matrix algebra. Examples relating to health problems have been solved using SPSS software. Permission has been taken for the examples and tables included in this manual.

In general most statistical methods require extensive computations. We have tried to avoid details of complex calculations, since software for data analyses are available. It is recommended for the users of this manual to use software, where possible, in solving the

problems. The data entry system has been explained either in the text or at the end of each chapter. However, for those who wish to solve problems manually, all the steps have been clearly demonstrated. At the end of each chapter the applications of SPSS software have been demonstrated in details.

Muhammad Hanif
Munir Ahmad

Contents

Basic Concepts and Data Presentation

1.1 Introduction

The word *statistics* seems to have been derived from the Latin word *status* or the Italian word *statist*. Both these words mean a *political state*. The word *statist* was also used by Shakespeare and Milton in the sense of a *statesman*, i.e. a person well versed in the affairs of the state. Modern concept of statistics was illustrated by Sir R.A. Fisher (1890-1962), J. Neyman (1894-1983), E.S. Pearson (1895-1981) and many others.

The word statistics is used in the plural sense to refer to *numerical facts in any field of study*. It concerns with collection, organization, summarization, analysis and drawing inferences from a data set. This word is also used in singular sense to refer to the *science* comprising methods, which are used in collection, presentation, analysis, and interpretation of numerical data.

Bio-statistics is the branch of statistics that concerns with the applications of statistical methods to medical and biological data.

In medical field, statistical methods enable us to study the effectiveness of different treatments in medicines. Recently, it has been found that applications of statistical methods in medical data are very effective. Testing of hypothesis, analysis of variance, chi-square, non-parametric methods, regression and correlation, logistic regression etc. are frequently used in the analysis of data in the health and medical sciences.

Knowledge of statistical methods is very important in health and medical research and in clinical practice for dealing with uncertainty in diagnosis, treatments and prognosis. These methods are useful and important both for clinicians as well as medical researchers, since they have to evaluate both clinical and research materials to improve their understanding and skills while treating patients.It is necessary to explain some basic terms and their definitions to understand statistical concepts in depth.

Here is a quick chart for the steps for scientific research:

Steps for Scientific Research
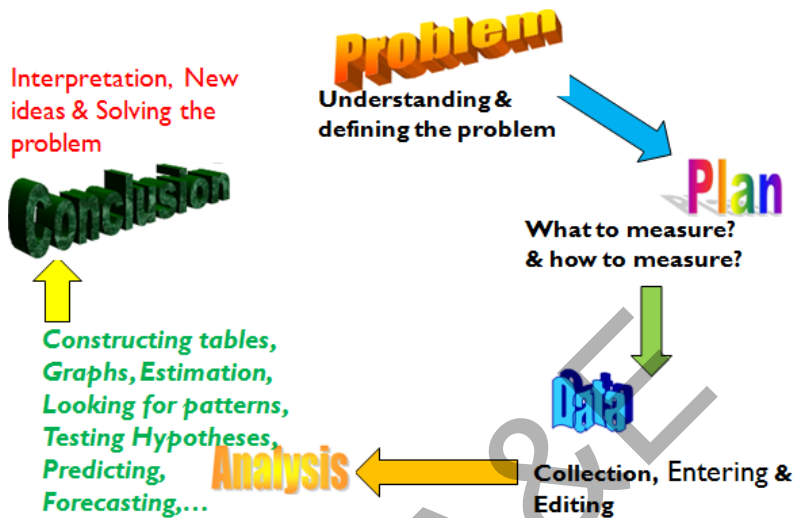
1.1.1 Population versus Sample

Population means an aggregate of individuals having a particular characteristic. In medical science it is generally human population but it may be a population of patients. The group of all patients in any hospital is known as a population of patients of that hospital. Population of smokers, population of cancer patients, etc. are some examples of population. In medical science we sometimes consider a *target population* about which inferences are to be drawn. Generally, population is of two types viz. *Finite* and *Infinite population.* A population is said to be finite if one can count individuals, otherwise, it is known as an infinite population. An infinite population comprises infinitely large number of elements. In statistics, if the number of individuals in a population is countable, it is known as a finite population and if it is not, then it is treated as infinite population.

A *sample* is defined as a representative part of any population. This representative part is not haphazard but some scientific method is used to select this part. At this stage, one should only remember that *random technique*, giving all members of the population an equal chance of selection, is applied to select the sample. Sample is considered to be *large* if the number of individuals in the sample is 30 or more, otherwise it is considered as a *small* sample. (Details of this will be discussed in Chapter 3).

1.1.2 Parameter versus Statistic

Parameter is a value (known or unknown) concerning some characteristic of a population. For example, average age of patients in a certain hospital admitted at a certain time is a parameter. It is a fixed quantity and always to be estimated.

Statistic is a value concerning some characteristic of a sample. For example, sample average can be defined as a statistic. Sample average may vary from sample to sample even drawn from the same population.

1.1.3 Descriptive versus Inferential Statistics

Descriptive statistics is a branch of statistics devoted to the organization, summarization and description of data. *Inferential statistics* is the branch of statistics concerned with using sample data to make inferences about a population. Proper sampling technique provides a *measure of reliability* for the inference. In inferential statistics, predictions are made and conclusions are drawn for the target population based on the sample.
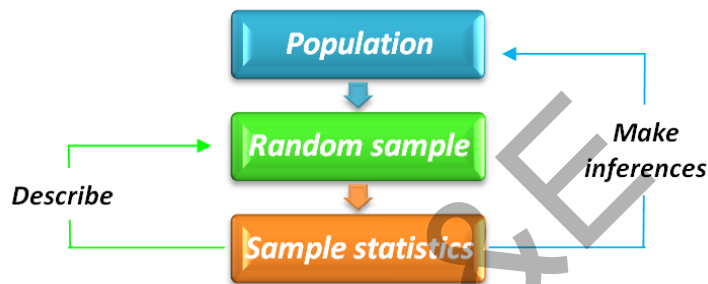
1.1.4 Descriptive versus Analytic Studies

A study has one of two objectives; either descriptive or analytic. In a *descriptive study*, statistical data is collected, organized and summarized according to one or more characteristics. The study of means, proportions, rates, standard deviations, graphic representations of data fall under the category of descriptive studies. Association or correlation is sought but no cause-effects are inferred. In fact no causal inference is involved in descriptive studies. Measuring of incidence, and most of the vital statistics, i.e. death rate, birth rate, fertility rate, etc. also come under descriptive study. Study of child growth and development comes under descriptive study. How many people are suffering from AIDS is an example of cross-sectional study. This study measures the prevalence of disease at a point in time and also determines the association between a factor and disease. Some other types of descriptive studies are *case-report*, *case-series* (analysis of cases) etc. In *analytic studies*, a sample data is studied to draw inference about the nature of the data set from which the sample is selected. The main objective of analytic studies is to draw inference.

1.1.5 Cohort Study

Cohort refers to the fact that the study group is followed forward in time to the future. A Cohort study is a follow up study in which people that are exposed (or not exposed) to the suspected causal factor or compared to the subsequent development of the disease. It determines the association between exposure and disease. Incidence of disease can be estimated in exposed and non-exposed groups. In a Cohort study, a long time period is required. It is very costly, and is conducted relatively on common diseases.

For example, consider a cohort of 1000 persons of which 400 are smokers and 600 are non-smokers. The entire cohort is followed for 15 years and it is found that 50 out of 1000 develop lung cancer. Of these 45 were smokers and 5 were not. The information is summarized in a 2x2 table.

	Disease		
	Lung cancer	**Without lung cancer**	**Total**
Smokers	45	355	400
Non-smokers	5	595	600
Total	50	950	1000

1.1.6 Case versus Control study

A case-control study is backward looking study. This starts with the outcome of a disease and goes back to suspected cause. People with the disease are compared with people who are free from disease (control). The term *case-control* study is often called a retrospective study. This is a short time study, relatively less expensive and suitable for rare disease however incidence rate cannot be determined.

Suppose we like to determine the association between smoking and lung cancer. Suppose 100 cases having lung cancer (case) and 100 cases free from lung cancer (control) are selected. Both cases and controls are asked if they are smokers or non- smokers.

The information is summarized in a 2 x 2 table as:

	Cases	**Control**
Smokers	90	40
Non-smokers	10	60
Total	100	100

Of 100 lungs cancer cases 90 were smokers and 10 were non- smoker. Of 100 persons who are free from cancer 40 were smokers and 60 were non- smokers. Study of such cases fall under the category of case-control study.
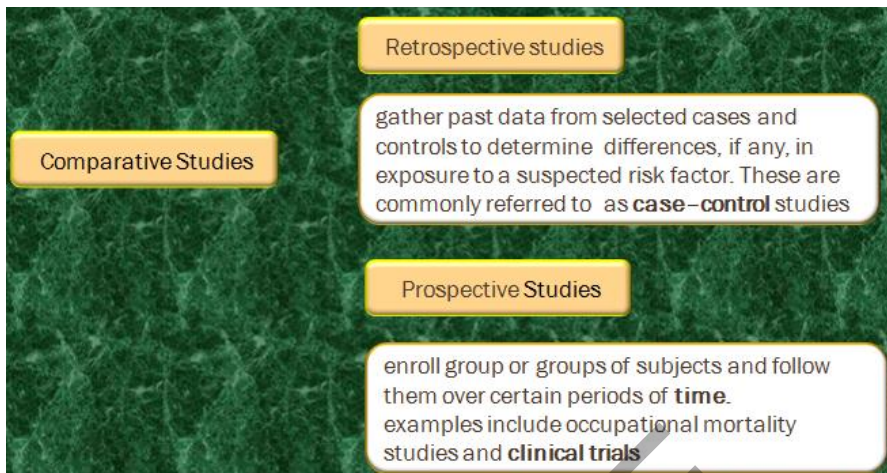
1.1.7 Experimental study

Experimental studies are considered special types of cohort studies where all conditions of the study are specified by an investigator, namely selection of treatment group, nature of interventions, management during follow up, etc. *The bearings of children, exposure to hazards, or personality type, are not normally subject to experiment.*

1.1.8 Intervention Studies

Epidemiological experiments that are designed to test cause-effect hypotheses may be termed *intervention studies*. Intervention studies may be group-based or individual-based. If the effect of fluoride on dental caries is investigated by fluoridating, the water supplies of some towns and comparing the subsequent occurrence of dental caries in these towns, it is a *group-based experiment*. On the other hand, when the administration of oxygen,[to premature infants causing retrolental fibroplasia (a blinding disease)], is tested by administering oxygen continuously to some babies then it is an *individual-based experiment.*

1.2 Variable

A variable is a characteristic of an individual which takes different values at different situations i.e. age, height and weight of patients, level of education, marital status, pulmonary blood flow (PBF), pulmonary blood volume (PBV), stage of a disease type of accidents, number of visits to a hospital, gestation age (weeks), smoking status etc. are a few examples of a variable. The values assumed by these variables are either *categorical or numerical.* A numerical variable may further be divided into two types: *discrete variable* or *continuous variable*.

1.2.1 Categorical Variable

A categorical variable is one for which the observations recorded result in a set of categories. For example, gender is a categorical variable as it falls into two categories only such as male and female. Recovery from disease is a categorical variable as it may

be recorded into three categories as, not recovered, partially recovered or completely recovered. Similarly level of education is a categorical variable. Categorical variable is often referred to as a *qualitative variable*.

1.2.2 Numerical Variable

A numerical variable is one for which the observations are recorded in numerical values such as, age, height, etc. It has further two types viz. discrete and continuous. A numerical variable is often referred to as a *quantitative variable*.

(a) Discrete Variable
A variable that is capable of taking a set of discrete numerical values such as 10, 15, 1, 199, etc. but not every possible value between two given numbers, is termed as *discrete variable*. The number of heart beats in a fixed time period, number of successful operations in a hospital; number of cases reported at a casualty ward of a certain hospital etc. are a few examples of *discrete variables*.

(b) Continuous Variable
A variable, which is capable of taking every possible value between two given number is termed as a *continuous variable*. Age, weight, length, etc. are a few examples of continuous variables.

1.2.3 Dependent and Independent Variables

Variables can further be divided into *dependent* (response) and an *independent* (predictor or explanatory) variable. Some examples of dependent and independent variables are as follows:

a. In a study of a prevalence of a disease in different age groups, the presence of the disease may be referred to as a *dependent variable*, whereas age is an *independent* variable.

b. In the study of the effect of smoking on lungs, smoking is an *independent variable;* whereas effect of smoking on the lungs is a *dependent variable*.

c. In a study of an association between birth weight of a child gestation period (weeks) and smoking status are possible factors that may influence the birth weight of a child. *Birth weight* is *dependent variable* whereas smoking status and gestation period are *independent variables*.

d. In the study of early sitting, smiling and walking of a child, the factors such as age, gender, birth weight, type of feeding, education of mother and father, birth order, number of siblings, etc. are *independent variables*.

e. In a study of mental disorders among elderly population; gender, age, family type, education level, income, family history, etc. may be taken as *independent variables*.

(a) An Interval Scale

This scale considers as pertinent information not only the relative order of the measurements as in the ordinal scale but also the size of the interval between measurements, that is the size of the difference (in a subtraction sense) between two measurements. We know, for example, that the difference between measurement of 10 and a measurement of 20 is equal to the difference between measurements of 20 and 30. The ability to do this implies the use of a unit distance and a zero point, both of which are arbitrary but it is not important which measurement is declared to be zero or which distance is defined to be the unit distance. Temperature has been measured quite adequately for some time by both the Fahrenheit and Centigrade scales, which have different origin and scale. The principle of interval measurement is not violated by a change in scale or location or both. In simple words, we can say that an interval scale may have an arbitrary zero unit, for example, temperature measured on a Celsius scale is an *interval scale* as 25°C = 72°F and 50°C = 112°F but the intervals of Celsius scale and Fahrenheit scale are not equal, e.g. [25 , 50] ≠ [72 , 112].

(b) The Ratio Scale

Unlike, the interval scale, the ratio scale has an absolute zero point, for example, weight measured on metric scale is a *ratio scale* because

   1 ton = 1016 Kg; and 2 tons = 2032 Kg therefore [1 : 2] = [1016 : 2032]

The ratio scale of measurement is used when the order and interval size are important, and the ratio between two measurements is meaningful. The ratio scale is appropriate for measuring crop yields, distances, weights, heights, income, length, time, mass, volume, etc.

1.4 Types of Statistical Data

An observation recorded or measurement taken in a planned study with some objectives in mind may result in a letter like "A" type blood or number like "120 mmHg" blood pressure. A collection of such observations may be termed as *data* or *statistical data*. Data may be classified into two types, viz. *Qualitative Data* and *Quantitative Data.*

1.4.1 Qualitative Data

When a population is classified into several categories, it is possible to count the number of individuals in each category. These counts are the *qualitative data*. A diagnostic test for pregnancy gives either positive (+) or negative(-) result. Colour of hair, colour of eyes, gender, non-resident, vaccinated or not, blood types, etc. are few examples of qualitative data. Observations recorded qualitatively (non-numerical measurements) give rise to qualitative data.

1.4.2 Quantitative Data

Observations, which are measured quantitatively (numerical measurements) give rise to *quantitative data*, such as measurement of serum cholesterol level, systolic blood pressure, blood urea nitrogen (BUN), etc. are some examples of quantitative data.

S1 Introduction to IBM-SPSS

S1.1 The origins of SPSS

In 1968, Norman H. Nie, C. Hadlai (Tex) Hull and Dale H. Bent, three young men from disparate professional backgrounds, developed a software system based on the idea of using statistics to turn raw data into information essential to decision-making.

Nie, a social scientist and Stanford doctoral candidate, represented the target audience and set the requirements; Bent, a Stanford University doctoral candidate in operations research, had the analysis expertise and designed the SPSS system file structure; and Hull, who had recently graduated from Stanford with a master of business administration degree, programmed.

This revolutionary statistical software system was called SPSS, which stood for the Statistical Package for the Social Sciences. SPSS is renamed as PASW (Predictive Analytic Soft Ware) in version 18 after owned by IBM in 2009. Starting from version 19 IBM gave the name IBM-SPSS for the statistical package.

Today: IBM-SPSS is recognized as a leader in the predictive analytics market space. Predictive analytics, combines advanced analytics and decision optimization.

The symbols used for data according to the measurement level:

Measurment level	Data Type			
	Time	Date	String	Numeric
Scale			N/A	
Ordinal				
Nominal				

S1.2 The Views of IBM-SPSS

SPSS has two views, the **Data view** and the **Variable view**. The Data view displays the actual data values or defined value labels, while in the Variable view, the variables are defined with label, measurement levels and other important features.

Many of the features of Data view are similar to the features that are found in spreadsheet applications. There are, however, several important distinctions:

- **Rows are cases**. Each row represents a case or an observation. For example, each individual respondent to a questionnaire is a case.
- **Columns are variables**. Each column represents a variable or characteristic that is being measured. For example, each item on a questionnaire is a variable.

S1.3 The Toolbar

Icon	Use	Function
	Open file	In addition to files saved in SPSS format, we can open Excel, SAS, and Stata, tab-delimited and other files without converting the files to an intermediate format or entering data definition information.
	Save file	In addition to saving data files in SPSS format, we can save data from SPSS in a wide variety of external formats
	Print	PRINT displays the values of variables for each case in the data.
	Recall	Recall recently used dialogs
	Undo	Undo a user action

Icon	Use	Function
	Redo	Redo a user action
	Go to case	To go to a specific case
	Go to variable	To go to a specific variable
	Variables	To see the definition for a specific variable
	Find	To search for a specific word or number
	Insert case	To insert case between two cases
	Insert variable	To insert variable between two variables
	Split file	Split File splits the data file into separate groups for analysis based on the values of one or more grouping variables. If we select multiple grouping variables, cases are grouped by each variable within categories of the preceding variable on the Groups Based On list.
	Weight cases	Weight Cases gives cases different weights (by simulated replication) for statistical analysis.
	Select cases	Select Cases provides several methods for selecting a subgroup of cases based on criteria that include variables and complex expressions. We can also select a random sample of cases.
	Value labels	When labels have been assigned to the category codes of a categorical variable, these can be displayed by checking Value Labels
	Use variable sets	This is to define sets for group of variables e.g. to construct a set called "demography" to contain all demographic variables only.
	Show all variables	This is to show all variables, not only the pre-defined variable sets
	Spell check	This for checking the spelling.

S1.4 The Menu

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help	
File	We use the file menu to read the data, an existing SPSS data file, spreadsheet, text, or database files created by other software.
Edit	Used to perform the standard Windows functions to cut copy & paste selections & to find data values.
View	Used to display gridlines, labels, the status bar & toolbars, & to change the display font.
Data	Used to access the SPSS facilities that make global changes to SPSS data files.
Transform	Used to access SPSS facilities that modify or create new variables in the data file. We can compute new variables, bin values of scale variables, manipulate date/time variables, & record variables from this menu.
Analyze	Used to analyse the SPSS statistical & reporting procedures we have installed with SPSS. This menu contains all of the SPSS procedures included in the SPSS base system. EX .frequencies, cross tabs as well as other descriptive procedures, regression, analysis of variance & many more.
Direct Marketing	Has some recent applications in Marketing researches
Graphs	Used to create charts using the Chart Builder or the Interactive Graphics system. Some statistical procedures also optionally generate charts.
Utilities	Used to display variable information, to define & use variable sets to control the variables that appear in the Data Editor & in the variable lists of dialogue boxes.
Add-ons	Used to add new products of SPSS, not included in SPSS base system.
Window	We use the Window menu to switch between SPSS windows & manipulate how they appear on the screen.
Help	Used to provide access to the many Help features of SPSS.

S1.5 The Variable View

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role

Name	Each variable must be assigned a unique name no longer than 64 characters.
Type	the type or format of the variable(numeric , string, dollar, etc.)
Width	the total number of columns(width) of the variable values
Decimals	the number of decimal positions of the variable value (should be set to 0 with nominal or ordinal variables)
Label	Variable label for the variable
Values	Value label for any nominal or ordinal variable
Missing	the values which should be flagged as user-missing and excluded by default from most analysis
Columns	Changes the display width of the column in the data view.
Align	Placement of the report relative to its margins. LEFT, CENTER, or RIGHT can be specified in the parentheses following the keyword.
Measure	The level of measurement for the variable
Role	Used to define the dependent variable (target) and independent variables (input) to be used automatically :

- Input
- Target
- Both
- None
- Partition
- Split
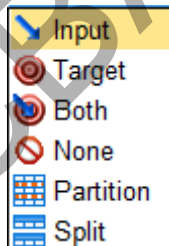
Example S1-1

Suppose for example, we have the following simple questionnaire,

It is clear that:

Variable	Measure	Symbol	Value
Serial No	Scale	(any)	
Age	Scale		
Gender	Nominal		1=male 2=female
Pain level	Ordinal		1=Mild 2=Moderate 3=Severe
Preferred medicine	Nominal		0=No 1=Yes

- Each variable has a column
- For the Preferred medicine, each choice is considered a variable, so that it has three variables (columns), this is known as "Multiple response".

Now suppose that the 1st patient's response was as follows:
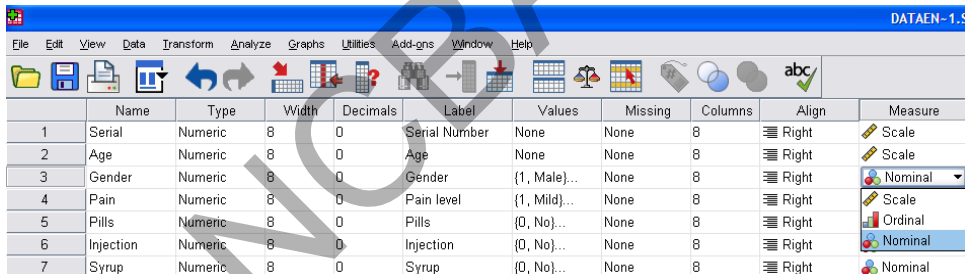
Then the corresponding data should be entered as a raw as follows:

Serial	Age	Gender	Pain level	Pills	Injection	Syrup
1	26	female	Mild	Yes	No	Yes

Now suppose that we have 10 patients with the following responses:

Serial	Age	Gender	Pain level	Pills	Injection	Syrup
1	26	female	Mild	Yes	No	Yes
2	21	female	Moderate	Yes	No	No
3	18	male	Moderate	No	No	Yes
4	35	male	Mild	Yes	Yes	No
5	41	female	Severe	Yes	Yes	Yes
6	22	male	Severe	Yes	No	No
7	22	male	Moderate	Yes	No	No
8	31	female	Mild	Yes	Yes	No
9	19	male	Severe	No	Yes	Yes
10	26	male	Severe	Yes	No	No

Variables in Variable View:
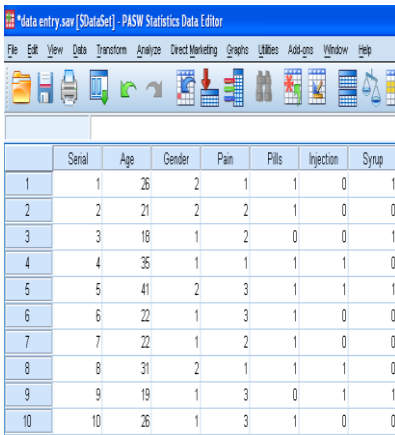
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Serial	Numeric	8	0	Serial Number	None	None	8	Right	Scale
2	Age	Numeric	8	0	Age	None	None	8	Right	Scale
3	Gender	Numeric	8	0	Gender	{1, Male}...	None	8	Right	Nominal
4	Pain	Numeric	8	0	Pain level	{1, Mild}...	None	8	Right	Scale
5	Pills	Numeric	8	0	Pills	{0, No}...	None	8	Right	Ordinal
6	Injection	Numeric	8	0	Injection	{0, No}...	None	8	Right	Nominal
7	Syrup	Numeric	8	0	Syrup	{0, No}...	None	8	Right	Nominal

Data in Data View:

Showing the numbers

Showing the values of the numbers (for Nominal or Ordinal)
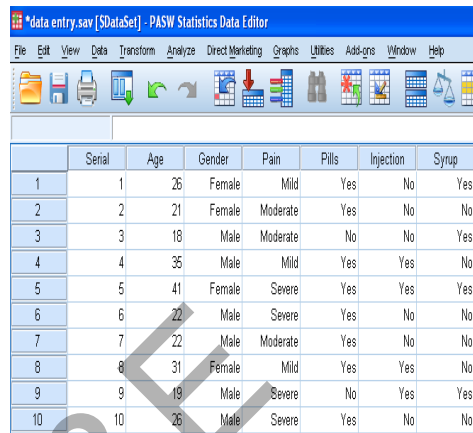
Push

	Serial	Age	Gender	Pain	Pills	Injection	Syrup
1	1	26	2	1	1	0	1
2	2	21	2	2	1	0	0
3	3	18	1	2	0	0	1
4	4	35	1	1	1	1	0
5	5	41	2	3	1	1	1
6	6	22	1	3	1	0	0
7	7	22	1	2	1	0	0
8	8	31	2	1	1	1	0
9	9	19	1	3	0	1	1
10	10	26	1	3	1	0	0

	Serial	Age	Gender	Pain	Pills	Injection	Syrup
1	1	26	Female	Mild	Yes	No	Yes
2	2	21	Female	Moderate	Yes	No	No
3	3	18	Male	Moderate	No	No	Yes
4	4	35	Male	Mild	Yes	Yes	No
5	5	41	Female	Severe	Yes	Yes	Yes
6	6	22	Male	Severe	Yes	No	No
7	7	22	Male	Moderate	Yes	No	No
8	8	31	Female	Mild	Yes	Yes	No
9	9	19	Male	Severe	No	Yes	Yes
10	10	26	Male	Severe	Yes	No	No

1.5 Graphical Presentation of Qualitative Data

The medical scientists while writing their papers or reports always present their information in the forms of diagrams and graphs as they are made to *summarize the data and a guide to further analysis*. Graphs are used to compare two or more than two sets of data. Every graph or chart should have a title that should give a clear description of the diagram or chart. A suitable scale should be used. The horizontal and vertical axes should be marked so that the graph or chart should be self-explanatory. There are many ways to present the data by charts and diagrams. We will discuss only *commonly* used charts or diagrams. Data involving a categorical variable measured on a nominal or ordinal scale can be displayed by (i) Simple Bar Charts (ii) Subdivided and Multiple Bar Charts and (iii) Pie Charts.

When representing the data graphically, we can use the "**Graphs → Chart Builder**"
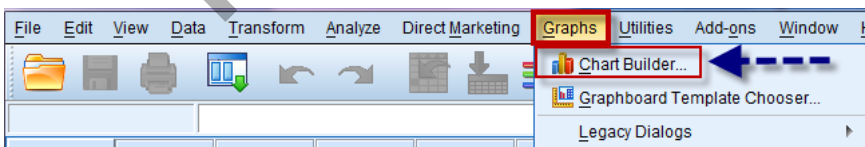
1.5.1 Bar Charts
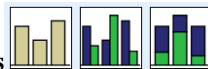
Bar chart is mainly used for graphical presentation of categorical data. Bar chart is obtained by plotting categories (of some constant widths) along X-axis and erecting bars of the heights equal to the corresponding numbers along Y-axis. Usually some fixed gap is left between two bars. Some non statisticians make the bar diagram for the data which relate to time, which in fact is not an appropriate chart.

Example 1.1:

Table 1.1 shows the blood groups of 230 patients visiting in January 1994 in the Blood Bank of King Fahd Teaching Hospital of the King Faisal University at Al-Khobar.

Table 1.1:
Blood groups of patients

Blood Group	A$^+$	A$^-$	B$^+$	B$^-$	AB$^+$	O$^+$	O$^-$
No. of Patients	35	10	45	5	20	105	10

Draw a suitable diagram for these data.

Solution:

Since the data given in the table are categorical, the most appropriate diagram is Bar Chart. There are 230 patients falling in 7 categories of various blood groups and each category is presented by a bar of height equal to the number of patients in that category as shown in Figure 1.1 presents each.
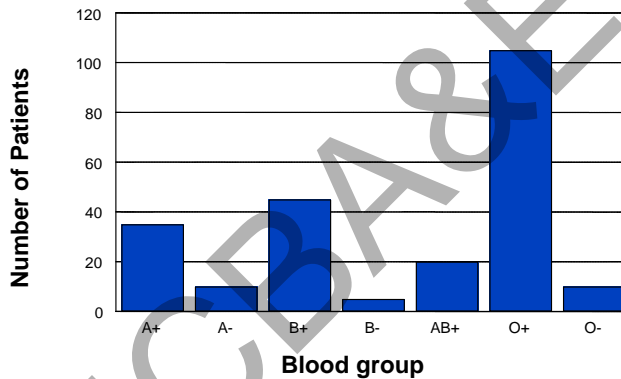
Fig. 1.1: Bar chart of blood groups

Example S1-2

To obtain the simple bar chart using IBM-SPSS, we enter the data and follow the following steps: **Graphs➔ Chart Builder:**
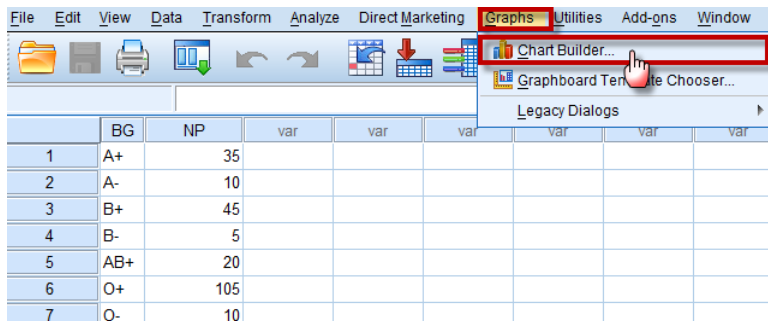
We drag or double click the icon , then we move the variables as follows:
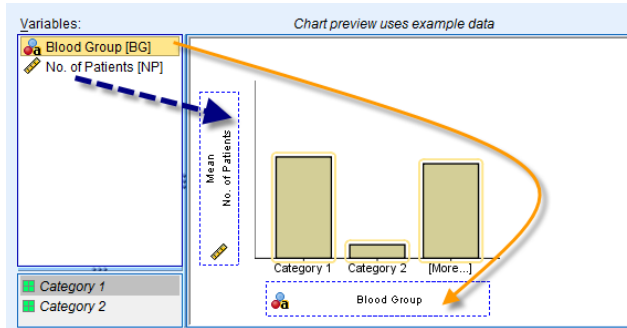
Once we click on , we get the following chart:
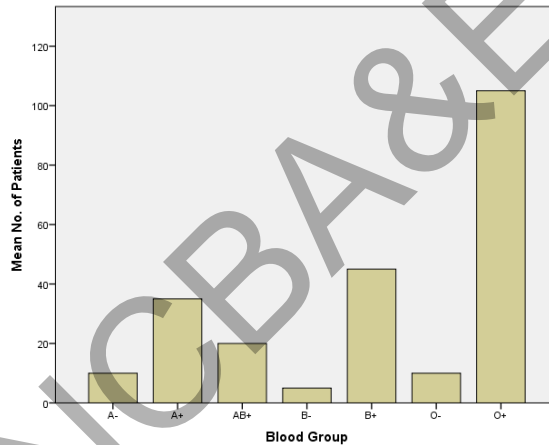
Once we click on the graph twice, we will change to the "Chart editor" then we can manipulate the figure (e.g. change the color, change it to 3D, etc…) and then using the icon , we can add the numbers.
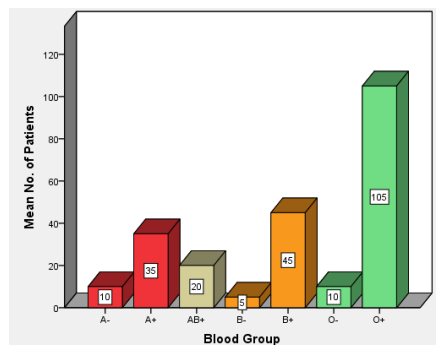
1.5.2 Subdivided and Multiple Bar Charts

If the data is grouped on the basis of two categorical variables then categories of one variable are displayed by erecting bars of height which corresponds to the values of these categories and the categories of second variable are displayed by dividing each bar into parts of size equal to the values of the sub-categories, whereas in multiple bar charts two bars for each category are constructed side by side.

Example 1.2:
Table 1.2 shows the type of investigation conducted on patients with breast disease for study 1 and study 2, in a New Bury Hospital of Berkshire from October 1 to December 31, 1989 (study 1) and from April 16 to July 19, 1990 (study-2)

Table 1.2:
Type of investigations by study type

No.	Type of Investigation	Study 1	Study 2	Total	%
1	Mammogram	11	15	26	23.9
2	FNAC*	5	8	13	11.9
3	FNAC + Mammogram	17	25	42	38.5
4	Cyst Aspiration	2	2	4	3.7
5	Cyst Aspiration + Mammogram	3	6	9	8.3
6	NIL	8	7	15	13.7
	Total	46	63	109	100

*Fine needle aspiration for catalogue

Prepare suitable charts.

Solution (a):
Subdivided Bar Chart - The numbers in each category are added and bar chart is prepared for each category. Further, each bar is divided into two types of study as shown in Fig. 1.2.

To obtain the subdivided bar chart using IBM-SPSS, we enter the data and follow the following steps: **Graphs➔ Chart Builder:**
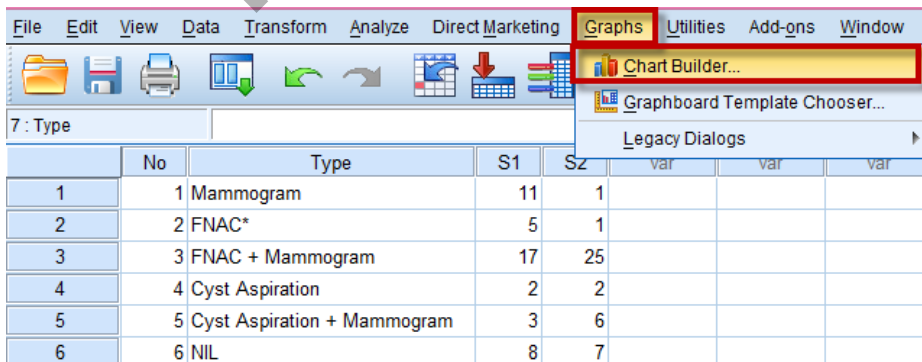
We drag or double click the icon �_▮ı▮▮▮, then we move the variables as follows:
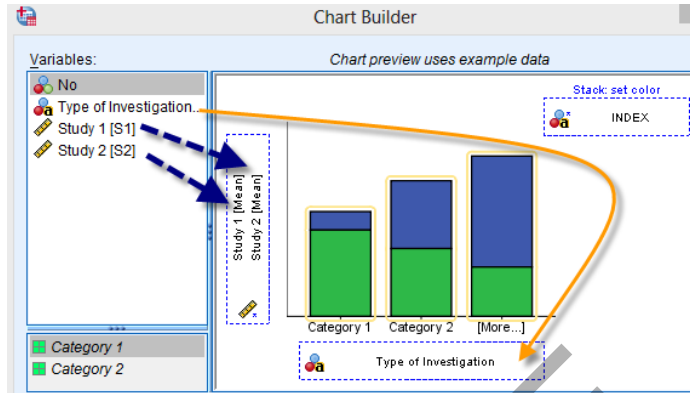
Once we click on **OK**, we get the following chart:
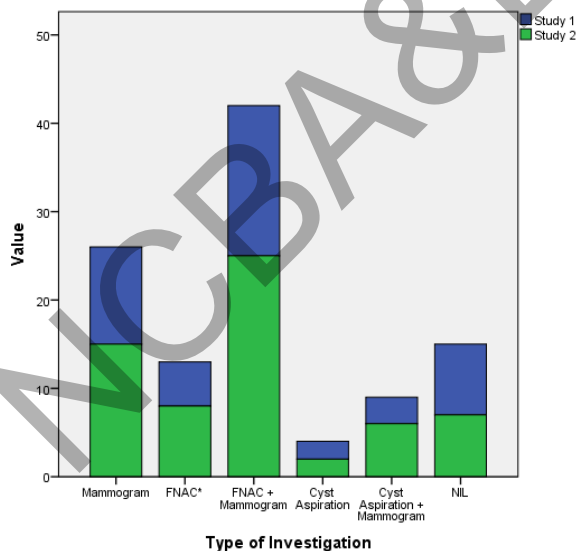
Fig. 1.2: Subdivided bar chart for Types of investigations performed

Solution (b):
Multiple Bar Chart - In this diagram, same data is used and two bars for each type of investigations of both studies are placed side by side as shown in Figure 1.3.

The advantage of the multiple bar chart is that comparison can be made easily. If there could be more than two studies, more than two bars are created side by side.

To obtain the subdivided bar chart using IBM-SPSS, we enter the data and follow the following steps: **Graphs➜ Chart Builder:**

We drag or double click the icon, then we move the variables as follows:

Once we click on OK, we get the following chart:

Fig. 1.3: Multiple bar chart for Types of investigations performed

1.5.3 Pie Chart

Pie chart is a pictorial presentation of the data. If a set of observation has K categories, it is represented by pies i.e. K sectors in a circle. The angle of the ith sector at the center of the circle, denoted by A_i, is proportional to the number in that category. It is given by:

$$A_i = \frac{\text{Value of the } i^{th} \text{ category}}{\text{Total value of all categories}} \times 360°; i = 1, 2, 3, \ldots, K$$

This is explained by the following example.

Example 1.3:

Table 1.3 shows the reported cases of AIDs in the 5 continents as of 17 Jan. 1992 (WHO).

Table 1.3:
Number of cases of AIDs by continents

Continents	No. of Cases
America	252,977
Africa	129,066
Europe	60,195
Oceanic	3,189
Asia	1,254
Total	446,681

Prepare a suitable chart for the given data.

Solution:

One can say that this data may be represented by bar charts, the answer is no, as the difference between the minimum value and maximum value is so much (more than 1:10) that bar charts for these data cannot be presented on normal paper. Besides we may be interested in the proportional share of each continent ratio than actual numbers. Therefore we look for another solution. The appropriate chart for this type of data is, Pie Chart that is shown in Fig. 1.4.

The angles and percentages are calculated as follows:

Table 1.4:
Computation of AIDS case by continent for Pie Diagram

Continents	No. of Cases of AIDS	A_i	Cumulative A_i
America	252977	204°	204°
Africa	129066	104°	308°
Europe	60195	48°	356°
Oceanic	3189	3°	359°
Asia	1254	1°	360°
Total	446681	360°	

American Continent $= A_1 = \dfrac{252,977}{446,681}$ x $360° = 204°$

African Continent $= A_2 = \dfrac{129,066}{446,681}$ x $360° = 104°$

Europe Continent $= A_3 = \dfrac{60,195}{446,681}$ x $360° = 48°$

Australia Continent $= A_4 = \dfrac{3,189}{446,681}$ x $360° = 3°$

Asia Continent $= A_5 = \dfrac{1,254}{446,681}$ x $360° = 1°$
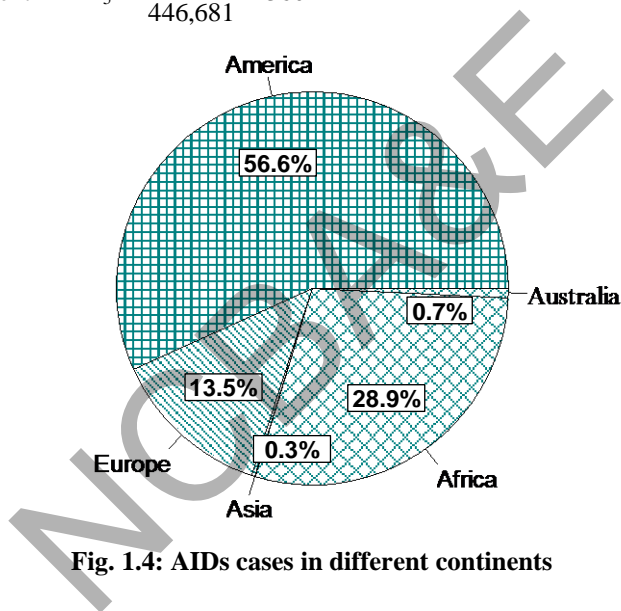
Fig. 1.4: AIDs cases in different continents

Note that it will be convenient to draw the chart if you calculate cumulative A_i. If one is using computer then there is no need of this column.

Example S1-3

To obtain the pie chart using IBM-SPSS, we enter the data and follow the following steps: **Graphs→ Chart Builder:**
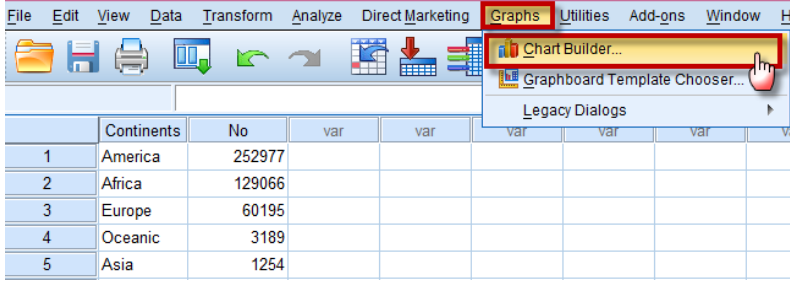
We drag or double click the icon , then we move the variables as follows:
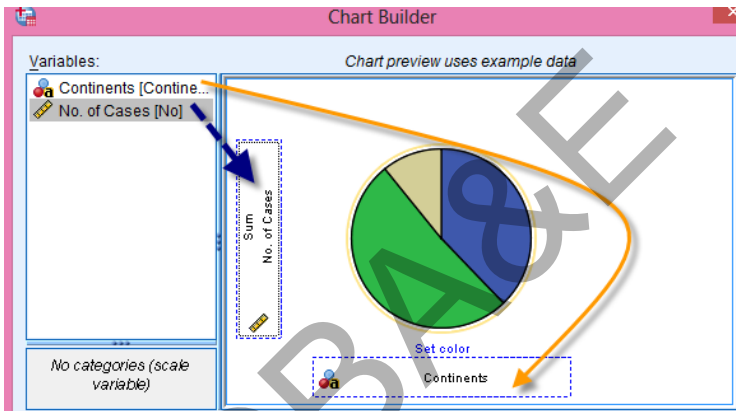
Once we click on , we get the following chart:
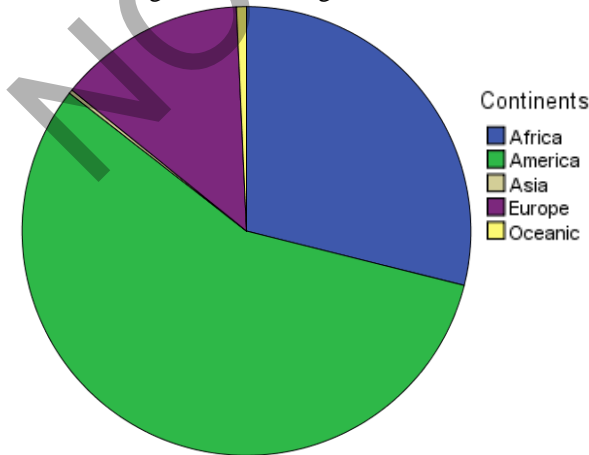
Once we click on the graph twice, we will change to the "Chart editor" then using the icon , we can add the percentages.

1.6 Summarization of Quantitative Data

In this section construction of grouping frequencies into tables, is explained. Relative frequency, and relative cumulative frequency have also been defined and are calculated. Their uses have also been discussed.

1.6.1 Frequency Table and Frequency Distribution

Frequency table is a two-column tabular presentation of the data. First column shows the different values of variable and second column the corresponding frequencies. To explain this, suppose we take 120 students from King Faisal University and record their weights to the nearest Kg.

Table 1.5:
Weights of 120 students in Kg

67	63	57	85	67	60	75	55	67	68	51	54
45	57	64	68	67	86	63	60	98	83	76	70
56	50	74	74	67	77	61	85	66	66	60	61
58	56	56	57	60	60	63	64	85	80	75	75
57	58	59	58	58	61	62	91	74	72	57	73
61	86	64	91	64	64	61	62	69	57	81	66
65	81	82	76	77	81	76	66	62	63	62	63
60	60	72	72	79	70	70	58	78	58	71	76
60	60	65	65	66	65	73	73	71	73	66	73
67	68	69	68	73	68	74	68	67	76	52	79

This is known as *raw or ungrouped data*. As the data is presented, it is difficult to understand how the weights of students are distributed. Only after some search, we can find that the minimum value is 45 and maximum value is 98. One can say that the weight of the 120 students of this University varies from 45 Kg to 98 Kg. Therefore, for better understanding we need some more manipulation of raw data.

In order to get a clear picture of the data, the data are presented in a condensed form, which is only possible if the data are grouped into a number of classes. If someone is working on the statistical packages, like *SPSS or SAS* he can directly condense the data into sufficient number of groups or classes.

How many groups should be there and how to make groupings? These two questions are very common for medical scientists. Let us deal with these, one by one.

Before grouping the data, it is important to decide upon the number of groups to be made. *As a general rule, the number of groups should neither be too small so that all the information is lost nor should be so large that no useful summarization is obtained. Usually the number of groups is taken from 5 to 15 and preferably from 5 to 10.*

Regarding second question, let K be the number of groups to be made, d the width of each of the group. The number K may be obtained by using Sturge's Rule as:

$$K = 1 + 3.322 \, (\log_{10} n),$$

where d = R/K, and R = maximum - minimum value of the data, n is the total number of observations. *Smallest value in the data set may be taken as the lower limit of the first group.* If, however, it is not an integer the next higher integer value is selected. *Note that this formula provides a guideline only and the value of K thus obtained, can be increased or decreased, for better presentation.* In the above data, maximum value is 98 and minimum value is 45, thus R = 98 - 45 = 53, n = 120.

Using the Sturge's Rule

$$K = 1 + 3.322 \ (\log_{10}120) = 1 + 3.322 \ (2.079) = 7.906 \sim 8$$

$$R = 53, \text{ then d (width)} = \frac{53}{8} = 6.6 \sim 7$$

Most statisticians prefer to group the data starting with a number with a multiple of 2 or 5 or 10 as the class may be.

Select 45 as the lower of the class limit and make the following groupings called *class intervals*:

45 to 51, 52 to 58, 59 to 65, 66 to 72, 73 to 79, 80 to 86, 87 to 93 and 94 to 100.

Table 1.6:
Distribution of students by weights in Kg.

Weights (class-limits/intervals)	Number of students
45 - 51	3
52 - 58	18
59 - 65	33
66 - 72	29
73 - 79	23
80 - 86	11
87 - 93	2
94 - 100	1
Total	120

This is known as *grouped data*. This table is known as *frequency table or frequency distribution.* To make frequency distribution by using SPSS package proceed as follows:

(i)   Enter raw data

(ii)   Click tool and then click recode, and click recode into different variable

(iii)   Bring the original variable to the right hand side and create a new variable (say, x) and change variable, finally

(iv)   Click old and new variable, recode data according to the groups you want to make.

Note that, the *class intervals* given in table 1.6 are called discrete class intervals. If someone is interested to present this data in form of appropriate diagram then one cannot, as the groups are discrete. Therefore continuous groups are must. To make it continous see the upper limit of the first group and lower limit of the second group, find their

difference and divide by 2. Add this number in the upper limit of the group and subtract from the lower limit of the group i.e. 45 – 0.5 = 44.5 and 51 + 0.5 = 51.5. Now these class limits will be called *class boundries* The class limits of table 1.6 is rewritten as class boundaries in table 1.7 (Column 1).

Table 1.7:
Distribution of Students by Weights in Kg. Percentage

Class Boundaries (1)	Number of students (2)	Relative frequencies or Proportion (3)	Percentage (4)	Cumulative frequencies (5)	Relative Cumulative frequencies (6)
44.5 - 51.5	3	0.025	2.5	3	0.025
51.5 - 58.5	18	0.150	15.0	3 + 18 = 21	0.175
58.5 - 65.5	33	0.275	27.5	21 + 33 = 54	0.450
65.5 - 72.5	29	0.242	24.2	54 + 29 = 83	0.692
72.5 - 79.5	23	0.191	19.1	83 + 23 = 106	0.883
79.5 - 86.5	11	0.092	9.2	106 + 11 = 117	0.975
86.5 - 93.5	2	0.017	1.7	117 + 2 = 119	0.995
93.5 - 100.5	1	0.008	0.8	119 + 1 = 120	1.000
Total	120	1			

If we do not know as to how many grouping there should be by using the given formula, we can use the following rule to calculate class interval.

Find the maximum and minimum values from the data. Calculate the range i.e. difference between maximum and minimum value. Divide the difference by the number of groups one likes to make. For example, in the above data maximum value is 98 and minimum value is 45, the range is 98 – 45 = 53. Suppose we like to make 10 groups then 53/10 = 5.3, roughly the groups will be made with an interval of 5 or 6. We shall prefer the interval to be 5

1.6.2 Relative Frequency

Relative frequency of a class interval is proportion of the class frequency relative to the total frequencies. Relative frequencies are in column (3), Table 1.7. The purpose ofcalculating the relative frequencies is to obtain the idea of proportion, and percentage which are, in fact, useful to understand the basic concept of different types of rates, ratios and consequently the idea of probability. From the Table 1.7, we can immediately say that there are about 27.5% students whose weight lies in the weight group 58.5 - 65.5 Kg.

1.6.3 Cumulative Frequency

The cumulative class frequency of class interval is the total number of observations having values less than the upper limit of that class interval. One of the advantages of the construction of cumulative frequency table is that, one gets immediately the picture, how many students have weight less than or equal to a certain point. For example there are 117 students whose weights are less than or equal to 86.5 Kg. The cumulative frequencies are given in column 5 of Table 1.7.

1.6.4 Relative Cumulative Frequency

The cumulative frequency of a class interval divided by the total frequencies is called *relative cumulative frequency*. It is generally expressed in the form of percentages and is known as *percentage cumulative frequency*. One of its advantages is that one can immediately get an idea, of the *percentage of the students whose weight is less than or equal to a certain point*. For example 69.2% students have weight less than or equal to 72.5 Kg. In other words one can say that about 31% students have weight above 72.5 Kg. The relative cumulative frequencies are given in column (6), Table 1.7.

1.7 Graphical Presentation of Quantitative Data

A grouped data involving a quantitative variable may be presented by various graphs. Some commonly used graphs are histogram, frequency polygon, frequency curve and cumulative frequency curve.

Example S1-4

We can use the IBM-SPSS, to change the raw date into a frequency table, then to obtain the frequency and cumulative table, through the following steps:
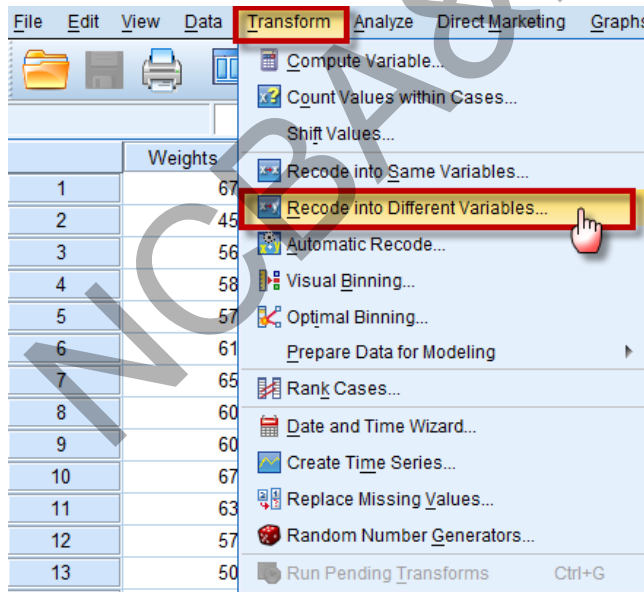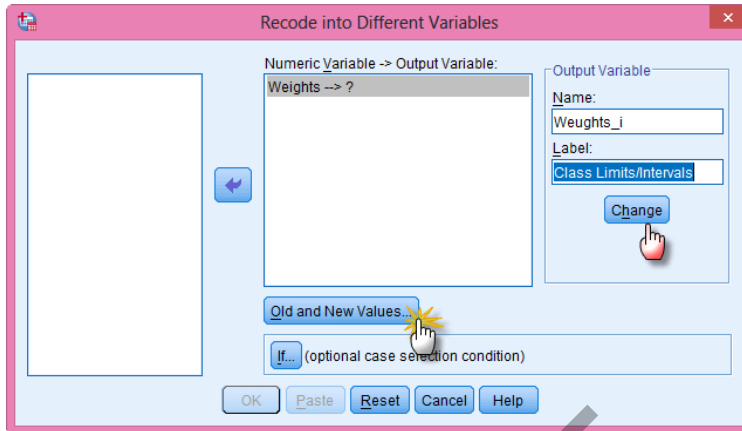
Now, we define the classes as follows:
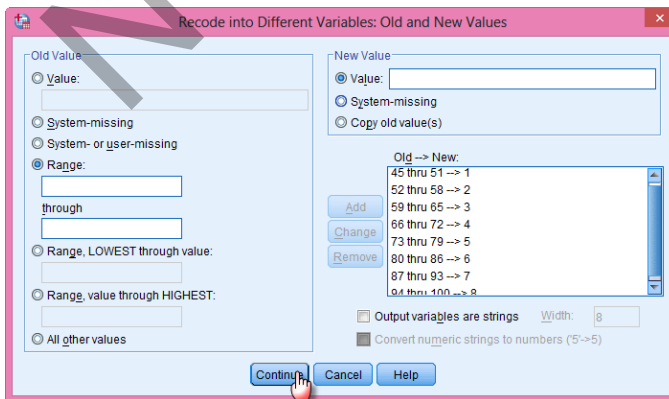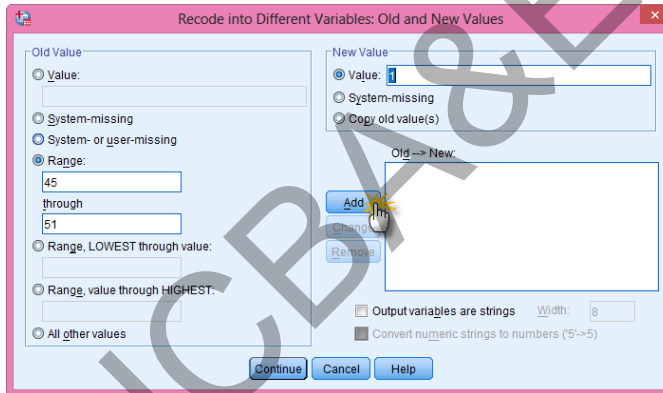
Now, in the "Variable View" we use the Values" to define the "Value Lables"
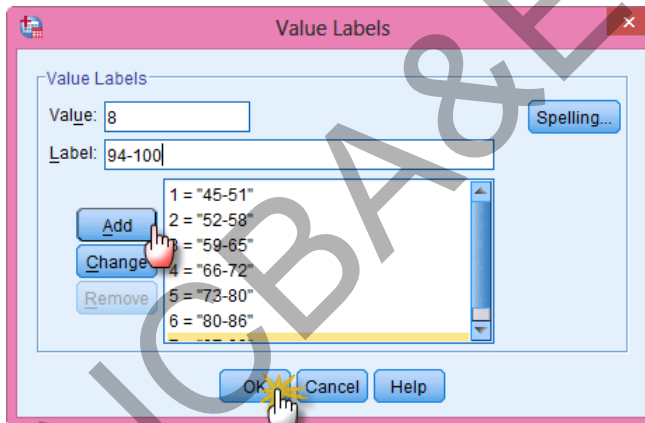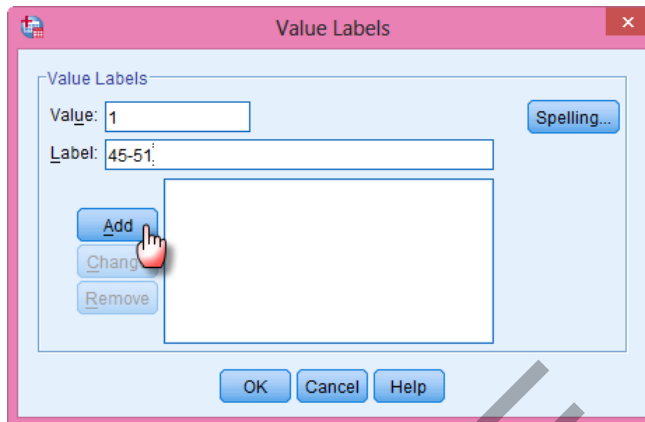
Then we can obtain the frequency table through:

Analyze→ Descriptive Statistics → Frequencies,
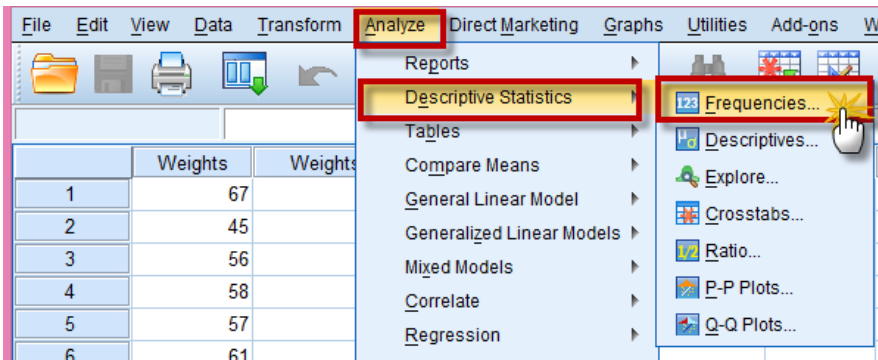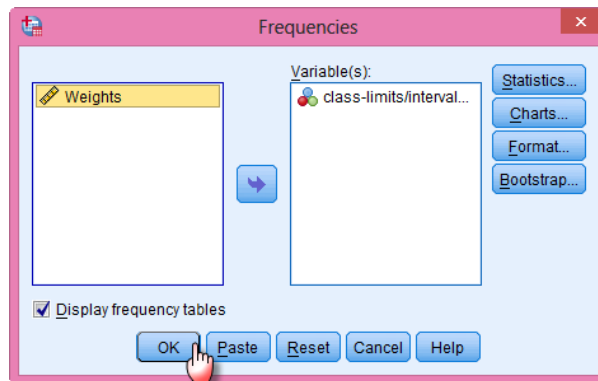
Once we click on OK, we get the following table:

class-limits/intervals

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	45-51	3	2.5	2.5	2.5
	52-58	18	15.0	15.0	17.5
	59-65	33	27.5	27.5	45.0
	66-72	29	24.2	24.2	69.2
	73-80	23	19.2	19.2	88.3
	80-86	11	9.2	9.2	97.5
	87-93	2	1.7	1.7	99.2
	94-100	1	.8	.8	100.0
	Total	120	100.0	100.0	

1.7.1 Histogram

Histogram is a graphical display of a frequency distribution and is obtained by plotting the class intervals along the X-axis and frequencies along the Y-axis. On each class interval (taken as width), we draw adjacent vertical bars of the heights equal to the corresponding frequencies. The graph thus obtained is called histogram. Histogram is constructed by using the data given in Table 1.7 and is shown in Figure1.5.

1.7.2 Frequency Polygon and Frequency Curve

Frequency Polygon is a graph obtained by joining by straight lines the mid points of the tops of the bars of the histogram. Frequency curve is a smoothed curve, which does not necessarily pass through the mid points like frequency polygon. The ends of the graph drawn in this way do not meet the X-axis, but remain open ended. This curve is very important as analysis of the data depends on the shape of the curve drawn. Frequency curve is plotted by using the data given in Table 1.7 and is shown in Fig.1.5.

To draw histogram we proceed as follows:

    i.   Enter the mid-points of groups in the first column
    ii.   Enter the frequencies in the second column
    iii. On data menu, click weight cases
    iv. Bring the frequency to the right hand side in frequency variable and click Ok
    v.   On graphs menu, click Histogram

Note: the histogram is ready but may not be according to your requirements

    vi. Click at *any point on X-axis* of the diagram a new histogram will appear, click any point on the X-axis
    vii. Click *custom* and then click *define*
    viii. Adjust interval and interval width as per your data
    ix. Histogram can be made directly from the raw data. For this purpose
    x.   Enter the required data
    xi. Click *graph* and click *histogram*, then follow steps vi-viii
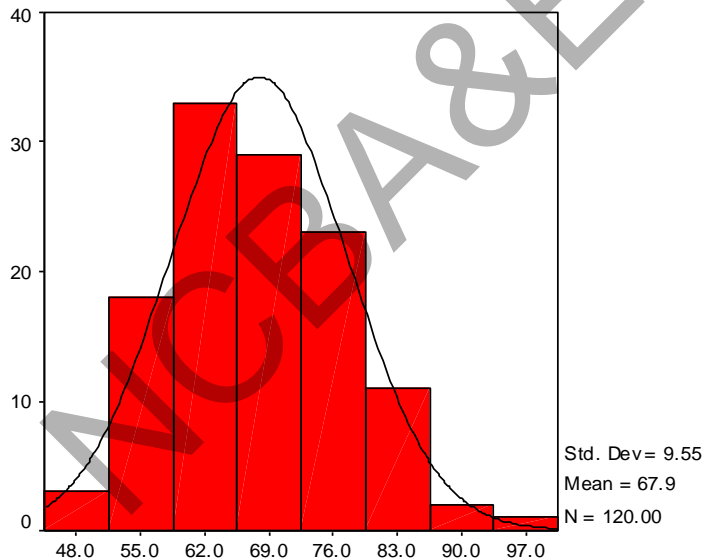
Fig. 1.5: Histogram frequency polygon and frequency curve

Example S1-5

To obtain the Histogram (automatically) using IBM-SPSS, we enter the data and follow the following steps: **Graphs➜ Chart Builder:**
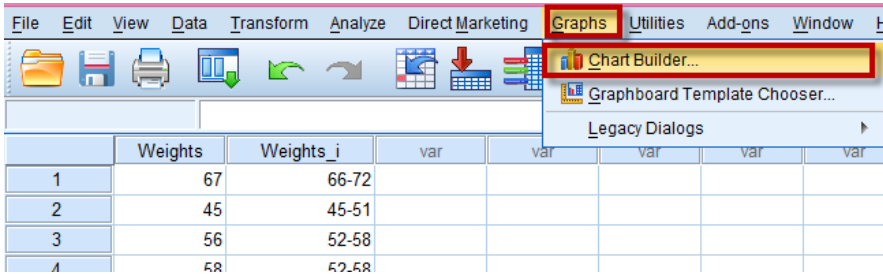
We drag or double click the icon , then we move the variable as follows:
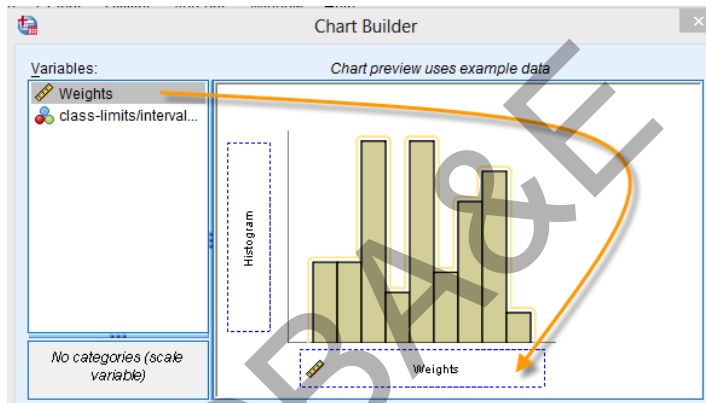
Once we click on , we get the following chart:
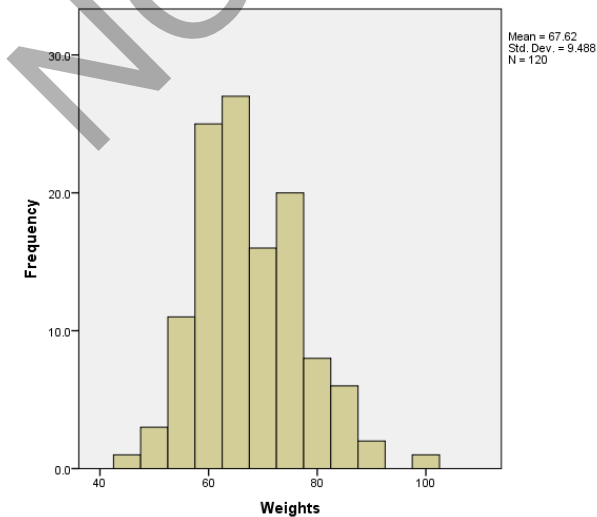
Also, we can obtain the Histogram (automatically) along with the Normal curve using IBM-SPSS, through the following alternative steps:

Analyze→ Descriptive Statistics → Frequencies,
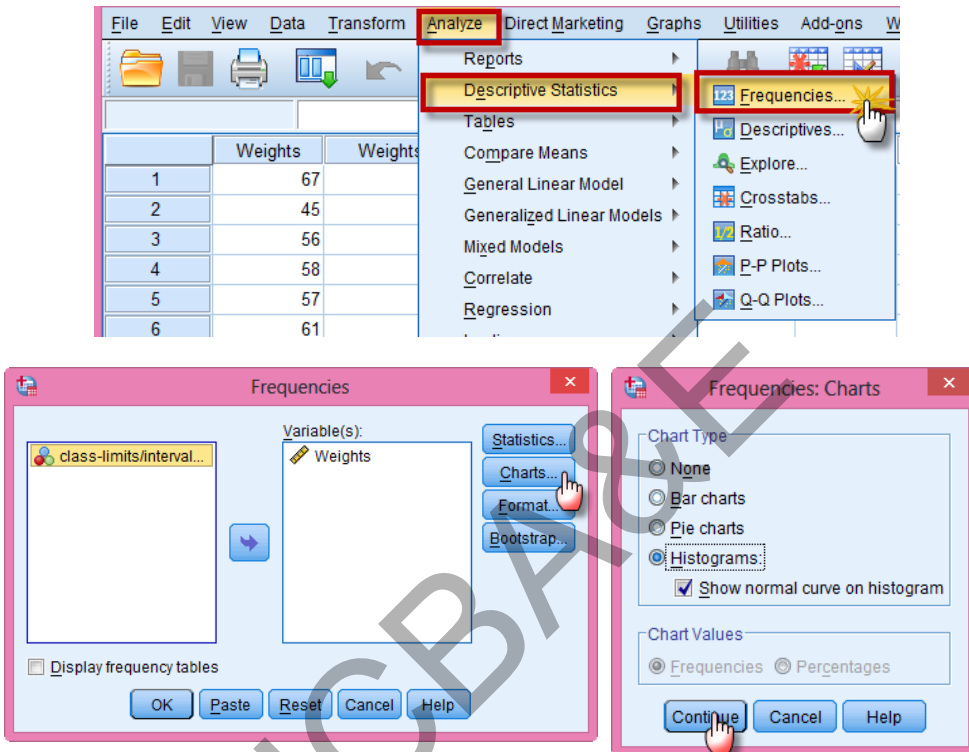
Once we click on [OK], we get the following chart:
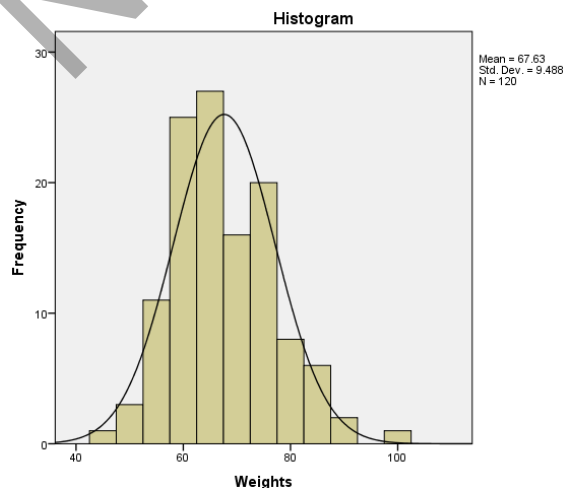
1.7.3 Types of Frequency Curve

Frequency curves are generally of two types; (i) *symmetrical* and (ii) *asymmetrical* or *skewed.* Asymmetrical or skewed curve is either positively skewed or negatively skewed. In symmetrical curves, observations are equidistant from the central maximum.

Normal curve (to be discussed later) is an important example of this type. In asymmetrical curves, the tails of the curves is longer on one side than the other side. If the longer tail is to the right, the curve is said to be *positively skewed.* If the longer tail is to the left, the curve is said to be *negatively skewed*.

(i) Symmetrical   (ii) Positively Skewed   (iii) Negatively Skewed
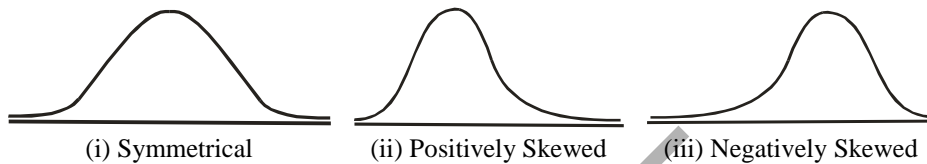
Fig. 1.6: Symmetrical and Asymmetrical curves

1.7.4 Cumulative Frequency Curve

Cumulative frequency curve is a graph obtained by plotting the upper limits on X-axis and the corresponding cumulative frequencies along Y-axis and joining the points by freehand. The graph of cumulative frequency using the data given in Table 1.7 is shown in Figure 1.7.
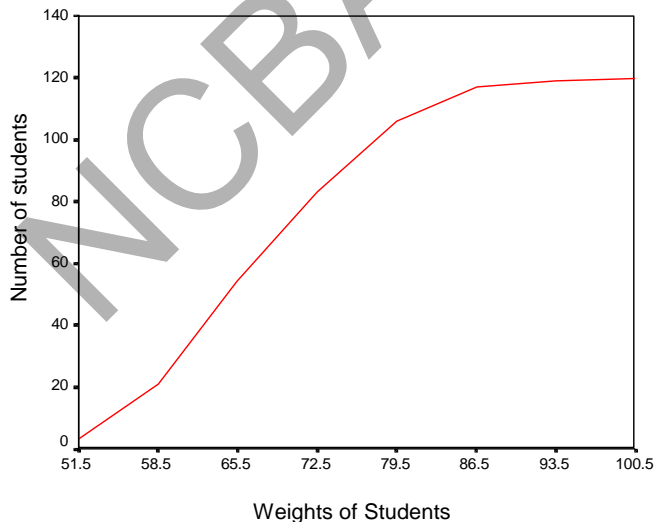
Fig. 1.7: Cumulative frequency curve

If we use SPSS package we proceed as follow:

1. Enter the upper limits of group in one column
2. Enter cumulative frequencies in the second column
3. Follow the guidelines given on

1.8 Historigram: Graphical Presentation of Data Relating to Time

Sometimes data is relating to time. People without going into details of the nature of data draw either bar diagram or pie charts for this type of data. In fact bar diagram or pie charts are not appropriate. The line diagram is drawn for the data relating to time. This graph is known as *Historigram*. One can see the trend of the data and may guess which type of analysis for this type of data.

Below are the data relating to number of students (males and females) admitted in King Faisal University from 1975-1976 to 1993-1994 in medical college. We are interested to present this data in an appropriate diagram.

Example 1.4:
Table 1.8 shows the data relating to admission of students in King Faisal University. Draw a suitable graph for this data.

Table 1.8:
Distribution of students by gender admitted in King Faisal University from 1975 to 1994

Year	Male	Female	Total
1975-76	170	0	170
1976-77	316	35	351
1977-78	537	77	614
1978-79	702	170	872
1979-80	910	248	1158
1980-81	1096	334	1430
1981-82	1269	544	1813
1982-83	1439	770	2209
1983-84	1577	1018	2595
1984-85	1876	1371	3247
1985-86	1898	1608	3506
1986-87	2088	1760	3848
1987-88	2146	1880	4026
1988-89	2234	2126	4360
1989-90	2226	2371	4597
1990-91	2259	2725	4984
1991-92	2430	2704	5134
1992-93	2681	3000	5681
1993-94	3145	3120	6265

Solution:
Fig (1.8) shows time series graphs of years and students by gender This Fig is given on next page.
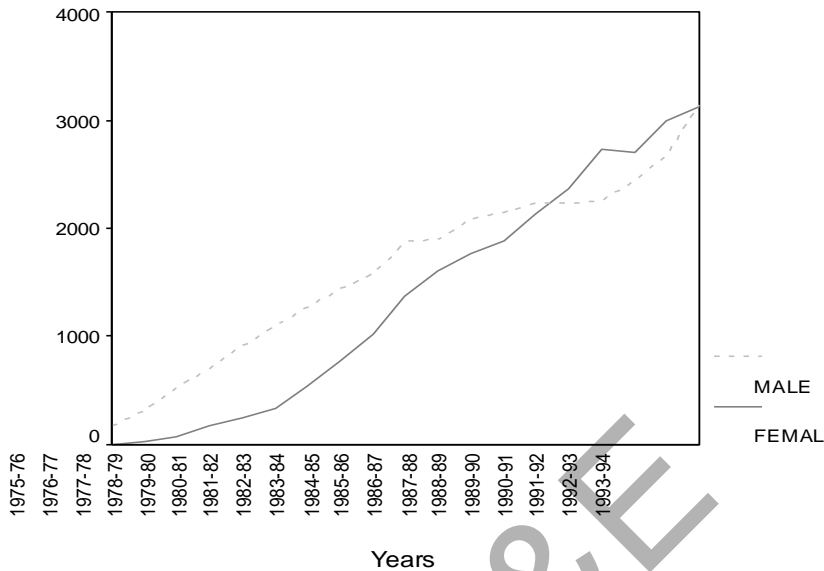
Fig. 1.8: Number of students admitted in King Faisal University

Example S1-6

For the data given in example S1-1, represent each of the age, gender and pain level using IBM-SPSS:

Variable	Measure	Symbol	Value	Graph
Age	Scale			Histogram
Gender	Nominal		1=male 2=female	Pie
Pain level	Ordinal		1=Mild 2=Moderate 3=Severe	Bar

We have two ways for representing data, either through the "**Graphs → Chart Builder**"
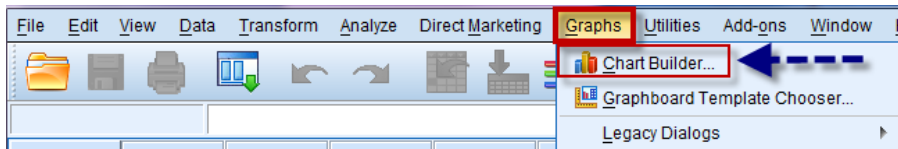
Or, we can Graph using Descriptive as follows:

(Analyze→ Descriptive Statistics→ Frequencies)
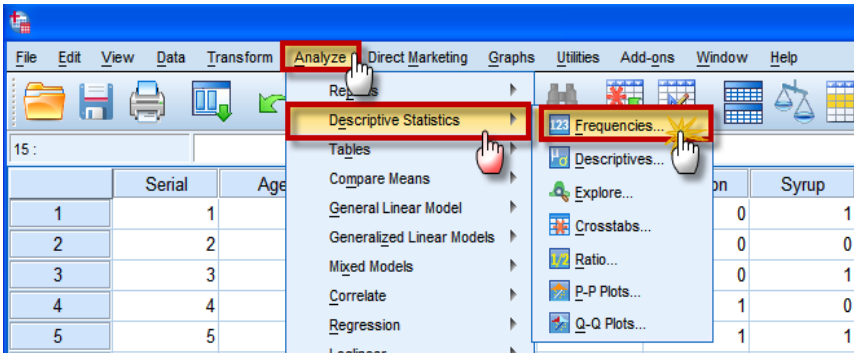
For the scale variable "Age":
Move the "Age" into variable,
Push on "Charts"
Select "Histogram"
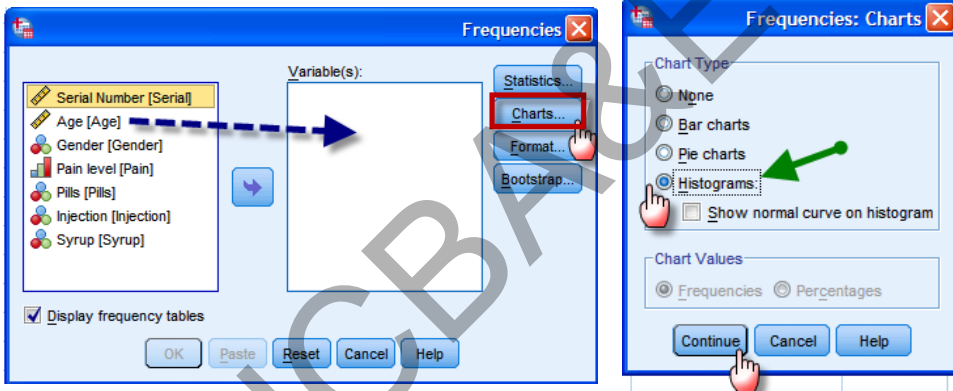
Push on " Continue " then " OK ", to get:
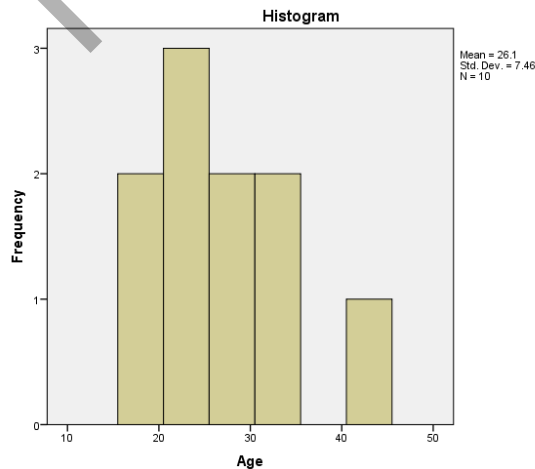
For the Nominal variable "Gender":
Move the "Gender" into variable,
Push on "Charts"
Select "Pie charts"
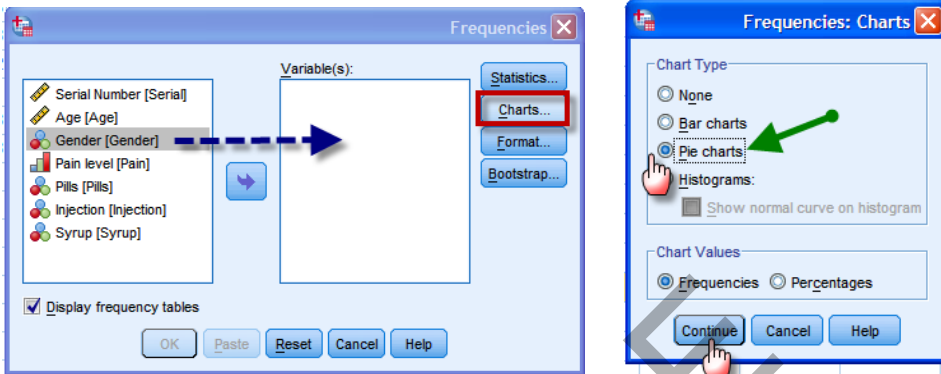
Push on " Continue " then " OK ", to get:

	Frequency	Percent
Male	6	60.0
Female	4	40.0
Total	10	100.0
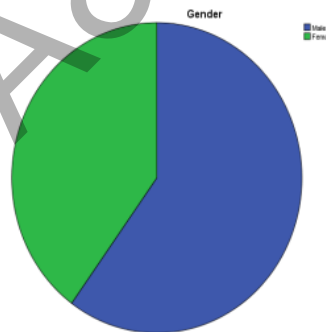
For the ordinal variable "Pain level":
Move the "pain" into variable,
Push on "Charts"
Select "Bar charts"
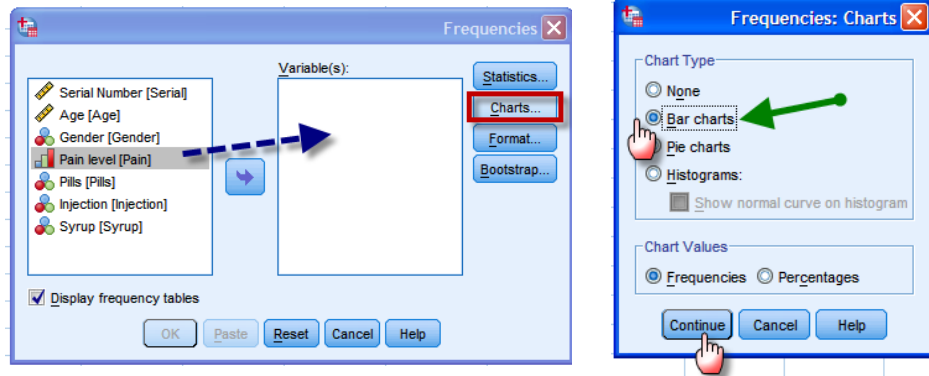
Push on " [Continue] " then " [OK] ", to get:

	Frequency	Percent
Mild	3	30.0
Moderate	3	30.0
Severe	4	40.0
Total	10	100.0
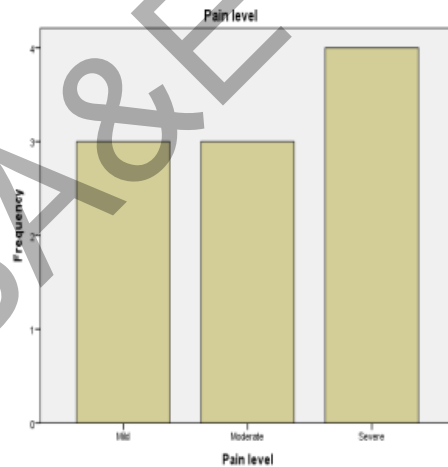
1.9 Descriptive Statistics

After the graphical presentation and summarization of statistical data, the next step is to proceed to different measures for statistical analysis. The methods of statistical analysis for qualitative and quantitative data are different. Proportion, percentage, ratio, indices, ranks, association, test of independence, etc. are possible methods of statistical analyses for qualitative data whereas percentage, indices, averages, variations, correlation, regression, analysis of variance, etc. are possible methods of analysis for quantitative data. For qualitative data, we shall describe the methods wherever it is necessary but we begin with quantitative data analysis.

1.9.1 Rates

Suppose, in a specified population, n events occur during a fixed period of time. If n(A) of these events possess some characteristic, say A, then rate of the event having the characteristic A is given by

$$R(A) = \frac{n(A)}{n} \cdot \text{base (K)}$$

per base (K) unit, where base is usually taken as 1,100,1000, or 100000, etc.

* If base is 1 then R(A) becomes proportion of A as given in column 3 of Table 1.6.

* If base is 100 then R(A) becomes percentage of A as given in column 4 of Table 1.6.

* In some of the cases base is either 100 or 1000 or 100000, like the death rate, birth rate. For very small proportions such as cancer patients base may be 10,000 or even 100,000.

(i) Prevalence Rate (P.R.)

Prevalence rate of an attribute or a disease in any group, is the proportion of individuals in the groups having that attribute at one point in time. This is also known a prevalence ratio.

$$\text{P.R} = \frac{\text{Number of individuals with disease at a given time}}{\text{Total number of individuals exp osed to the disease}} \times K$$

(ii) Incidence Rate (I.R.)

The risk of developing the disease over a period of time is called incidence rate and is calculated as:

$$\text{I.R} = \frac{\text{Number of new cases of disease over a period of time}}{\text{Population at risk of developing the disease}} \times K$$

(iii) Crude Death Rate (CDR)

$$\text{CDR} = \frac{\text{Total deaths during a calander year}}{\text{Total population on mid year July 01}} \times K$$

K is either 1000 or 100000.

(iv) Specific Death Rate (SDR)

$$\text{SDR} = \frac{\text{Total deaths in specific sub - group during a calander year}}{\text{Total population in the specific group on July 01}} \times K$$

(v) Crude Birth Rate (CBR)

$$\text{CBR} = \frac{\text{Total live births during the year}}{\text{Total populaiton on july 01}} \times K$$

(vi) Maternal Mortality Rate (MMR)

$$\text{MMR} = \frac{\text{Deaths from all puerperal causes during a year}}{\text{Total live births during the year}} \times K$$

The preferred denominator for this rate is the number of pregnant women during the year but it is difficult to determine. *A death from a puerperal is a death that can be ascribed to some phase of child bearing i.e. pregnancy or puerperal.*

(vii) Infant Mortality Rate (IMR)

$$IMR = \frac{\text{Deaths under one year of age during a year}}{\text{Total of live births during the year}} \times K$$

(viii) Neo-natal Mortality Rate(NNMR)

$$NNMR \frac{\text{Deaths from 0 to 28 days during a year}}{\text{Total of live births during the year}} \times K$$

(ix) Fetal Death Rate (FDR)

$$FDR = \frac{\text{Total fetal deaths during a year}}{\text{Total deliveries during the year}} \times K$$

A fetal death is defined as a product of conception that shows no sign of life after complete birth.

(x) Pre-Natal Mortality Rate (PMR)

$$PMR = \frac{\text{Total fetal deaths of 20(24) weeks or more} + \text{Infant deaths under 7 days}}{\text{Total births (alive and dead)}} \times K$$

(xi) General Fertility Rate (GFR)

$$GFR = \frac{\text{Total live birth to women aged } 15-44 \text{ years}}{\text{Total population of women aged } 15-44 \text{ years}} \times K$$

(xii) Body Mass Index (Quetelet's Index)

$$BMI = \frac{\text{Weight of } a \text{ person}}{\left(\text{Height of the person}\right)^2}$$

(xiii) Ponderal Index

$$= \frac{\text{Height}}{\left(\text{Weight}\right)^{1/3}}$$

Note: Units for weight and height are arbitrarily assigned.

1.9.2 Ratios

Suppose in a specific population, n events occur during a fixed period of time and n(A) of these events possess some characteristic "A" and n - n(A) of these events do not possess this characteristic, then the ratio of these events possessing the characteristic "A" is given as

$$\text{Ratio (A)} = \frac{n\ (A)}{n - n(A)}$$

For example, gender *ratio*, which is commonly used, is defined as

$$\text{Gender Ratio} = \frac{\text{Number of females}}{\text{Number of males}}$$

Some more examples are:

(i) Fetal Death Ratio

$$\text{FDR} = \frac{\text{Total number of fetal deaths during a year}}{\text{Total number of live births during a year}}$$

(ii) Immaturity Ratio

$$\text{IR} = \frac{\text{Number of live births under 2500 grams during a year}}{\text{Total number of live births during a year}}$$

(iii) Case-Fatality Ratio

$$\text{CFR} = \frac{\text{Total number of deaths due to disease}}{\text{Total number of cases due to disease}}$$

1.9.3 Odds Ratio

Suppose the number of observations possessing a characteristic "A" say case and control and is further classified according to another factor "B" called diseaseed and not diseased and we make a cross tabulation then these information may be presented 2 x 2 table also called contingency table as:

Table 1.9:
2 x 2 table for case-control versus disease-non-disease

Characteristics			
	Case	Control	
	A	$\overline{A}$	Total
Disease B	a	b	a + b
Non-disease $\overline{B}$	c	d	c + d
Total	a + c	b + d	a + b + c + d

A = exposed $\overline{A}$ = non-exposed

B = disease (case) $\overline{B}$ = no disease (control)

then the rate of diseased persons among exposed = $\dfrac{a}{a+c}$. The rate of non-diseased persons among exposed persons is $\dfrac{a}{a+c}$. Then the rate of exposure

among exposed case is $\dfrac{a}{a+c} \div \dfrac{c}{a+c} = \dfrac{c}{a}$. Similarly the rate of exposure among controls

$= \dfrac{b}{b+d} \div \dfrac{d}{b+d} = \dfrac{b}{d}$,

The odds ratio is the ratio of these odds and is given by

$$OR = \dfrac{a}{c} \div \dfrac{b}{d} = \dfrac{ad}{bc}$$

If any cell is zero, the odd ratios can be calculated by adding $\dfrac{1}{2}$ *to each cell.*The details

of odd ratio with its statistical meaning attached to it along with its statistical significance will be discussed in Chapter 7.

Example 1.5:
In a case-control study, let us take artificial example of alcohol and liver cirrhosis. The data are given table 1.10.

Table 1.10:
Liver cirrhosis

Alcohol	Case	Control	Total
	LC	$\overline{LC}$	
A	400 a	333 b	733
$\overline{A}$	100 c	167 d	267
Total	500	500	1000

LC = liver cirrhosis $\overline{LC}$ = no liver cirrhosis

A = alcohol drinking $\overline{A}$ = no alcohol drinking

The different indices are calculated as:

i) Odds of alcohol among cases $= \dfrac{400}{500} / \dfrac{100}{500} = \dfrac{400}{100} = \dfrac{a}{c}$

ii) Odds of alcohol among controls $= \dfrac{333}{500} / \dfrac{167}{500} = \dfrac{333}{167} = \dfrac{b}{d}$

iii) Odd ratio (OR) $= \dfrac{a/c}{b/d} = \dfrac{ad}{bc} = \dfrac{400 \times 167}{333 \times 100} = 2.006$

Example 1.6:
Consider the following example of the relationship between smoking and lung cancer in a case-control study:

	Cases	Controls
Smokers	145	107
Non- or ex-smokers	55	93

Calculate the odds ratio and give its meaning.

Solution:

$$\textbf{Odds ratio} = \frac{A/B}{C/D} = \frac{145/107}{55/93} = \textbf{2.29}$$

The value 2.29 can be interpreted as an estimate of the ratio of the odds, in the population, of smoker developing lung cancer to the odds of a non-smoker developing this disease.

In other words we can say that a smoker has 2.29 times more risk of developing lung cancer, than a non-smoker.

Example S1-7

Consider the following example of the relationship between smoking and lung cancer in a case-control study:

	Cases	Controls
Smokers	145	107
Non- or ex-smokers	55	93

Calculate the odds ratio and interpret its meaning.

Solution: The data in the IBM-SPSS file is as follows:

The variable view is:

We define the values of the variable as follows
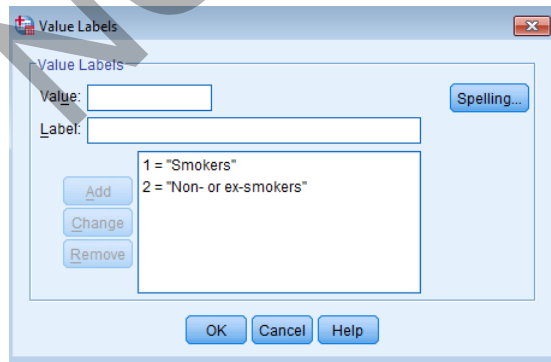
The Data are entered in two columns. The 1st column is for smoking, in which the value "1" which is corresponding to "Smokers" is entered 252 times, the value "2", which is corresponding to "Non – or-ex-smokers" is entered 148 times.

The 2nd column is for disease, in which the value "1" which is corresponding to "case" and the value "2" which is corresponding to "control", is entered as follows: "1" 145 times than "2" 105 times, then "1" 55 times than "2" 93 times. A part of the data view is:

To calculate the odds ratio, we follow the following steps:

Analyze→ Descriptive Statistics→ Crosstabs
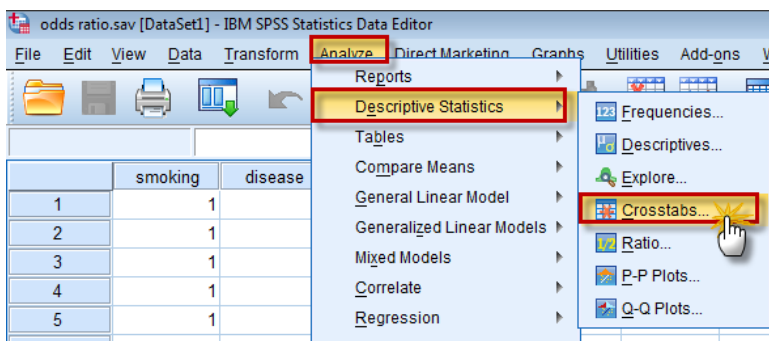
Then we will get the following windows, in which we will move the variable "smoke" to "Row" and the variable "disease" to "Columns". Then, we push on Statistics, as follows:
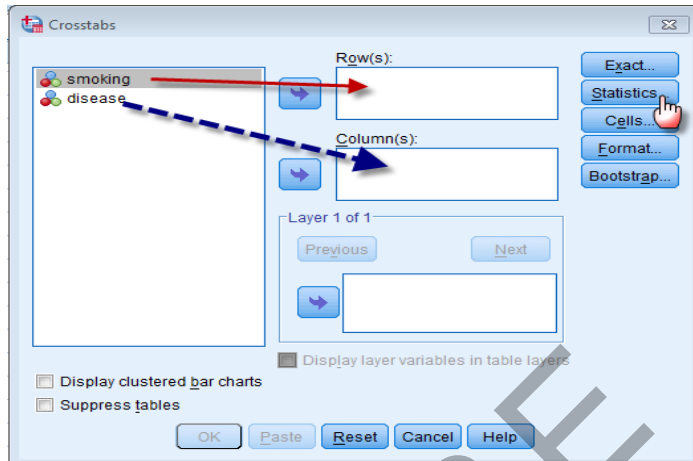
We mark on "**Cochran's and Mantel-Haenszel statistics**", then we push on continue, as shown in the following figure:
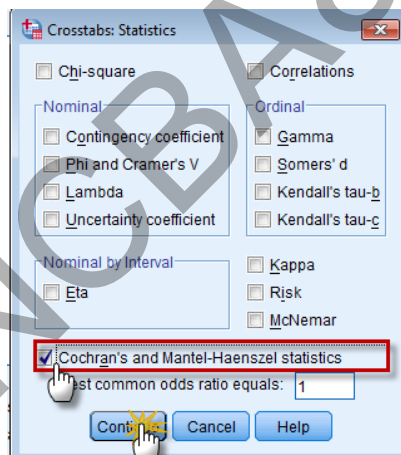
Now Click on [OK] to get the following results:

		disease		Total
		Case	Control	
smoking	Smokers	145	107	252
	Non- or ex-smokers	55	93	148
Total		200	200	400

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			2.291
ln(Estimate)			.829
Std. Error of ln(Estimate)			.213
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1.511
		Upper Bound	3.476
	ln(Common Odds Ratio)	Lower Bound	.413
		Upper Bound	1.246

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

The first table gives the observed values. The second table gives the odds ratio.

Note: The table also gives the 95% confidence value, with lower value equals 1.511 and upper value equals 3.476. It means that with 95% confidence, a smoker has (at least) about 1.5 times the risk of developing lung cancer than a non-smoker.

1.9.4 Measures of Central Tendency

Central tendency is a characteristic of a data set that relates to its average value. It is the central value in the sense that it is located in the middle and the data points cluster around it. *Since it is the most representative point of the data and a comparison between two or more data sets may, therefore, be made by their respective central points. In simple way, it can be said that methods of measures of central tendency are useful for the purpose of comparison of two or more similar types of data sets*. Most commonly used measures are, arithmetic mean, median and mode. Quartiles, deciles and percentiles are also position indicators and useful for comprehensive comparison of two or more sets of data.

(i) Arithmetic mean
Arithmetic mean or simply *mean* is most commonly used measure of central tendency. It has a very important property, viz., when it is subtracted from all the values of data, the sum of the differences of mean from observations is zero. It uses all observations fully in its calculation.

(a) Mean for ungrouped data
Add all the observations in a set of data and divide by the total number of observations, i.e.

$$\text{Mean} = \frac{\text{sum of all the observations of data set}}{\text{total number of observations}}$$

If "x_i" denotes the value of the ith observation and "n" the number of observations, then the mean ($\bar{x}$) is

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_1}{n} = \frac{\sum x_i}{n} \tag{1.1}$$

Example:
Suppose the weights of 14 patients are 62, 64, 65, 66, 68, 70, 70, 70, 70, 74, 74, 77, 77, 79 in kg, the mean for this data is

$$\text{Mean} = \frac{62 + 64 + 65 + \ldots + 79}{14} = \frac{1,036}{14} = 74\text{kg}$$

(b) Mean for grouped data
Given a grouped data, we first find the midpoints of the groups, which are multiplied by the corresponding frequencies of those groups. All these products are added. This sum is divided by the sum of all the frequencies. Suppose the weights of 14 patients is given, the mean can be calculated as:

Table 1.11:
Distribution of patient by Weights

Weight of patients (kg) (1)	Number of patients (f_i) (2)	Mid-points of groups (x_i) (3)	$f_i\,x_i$
60-64	2	62	124
65-69	4	67	268
70-74	6	72	432
75-79	2	77	154
Total	14		978

$$\text{Mean} = \frac{978}{14} = 69.857 \text{ kg}$$

If x_i are the mid values of the groups and f_i are the frequencies, then the mean ($\bar{x}$) is

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum f_i x_i}{\sum f_i} \tag{1.2}$$

The mean obtained from grouped data may be different from the mean obtained from ungrouped data. This is because in grouped data we assume that all the values in that group is placed at the mid-value of the class interval.

(ii) *Median, quartile, decile and percentile* [quantile]
The median of data set arranged in order of magnitude is the middle most value. If the numbers of observation are odd, the middle value is the median. If the numbers of observations are even, the arithmetic mean of the two middle most values is the median value. Median tells us that 50% of the observations are on both sides of the median point. *Median is a suitable measure for a data set which is measured on an ordinal or a ratio scale.* Like median, *quartiles* are points dividing an ordered data set into 4 equal parts, *deciles* divide ordered data set into 10 equal parts and *percentiles* divide an ordered data set into 100 equals parts. Note that for comprehensive comparison for two or more than two data sets of the same type, percentile *is relatively a better measure*. Since *SPSS* package will be used for the calculation of all these measures, therefore, detailed discussion on this topic will not be useful. By using *SPSS* package (as explained at the end of the chapter) median and other measures can be calculated easily. If we use SPSS

package on the raw data given in Table 1.5 the median comes out to be 66.5 kg and lower (first) quartile is 60.0 kg and upper (third) quartile is 74.9 kg whereas for different deciles or percentiles the values are as:

Percentiles	10	20	30	40	50	60	70	80	90
Value (kg)	57	60	61.3	64	66.5	68	74	75.8	81

(iii) Mode

Mode is the most frequently occuring number in the data set. The mode of the given data is 60, as 60 has occurred more times in the data set than anyother number. It is not an effective measure. Sometimes, there is no mode and sometimes there are more than one modal values. Sometimes, the distribution is bi-modal or is multimodal. In such cases, mode does not provide true picture of the central tendency. It is not generally done, but one way of finding a mode in multimodal data, is to find the average of all modes.The average mode may be considered as mode of the data set. For example the scores of medical students in a test are 2, 2, 2, 3, 5, 5, 5, 6, 6 in this case 2 and 5 are two modes. The average mode is $(2 + 5)/2 = 3.5$

1.9.5 Measures of Dispersion

The average value of a set of observations fails to describe the distribution without some degree of variation of the observations about the averages. Statistical measures of dispersion are used to measure the extent to which individual observations disperse or cluster around the average. They, like mean are also used to compare two or more data sets of same nature. Here only two measures, which are commonly used in medical science, will be described. These are *range* and *standard deviation*

(i) Range

Range is the difference between maximum and minimum values of data set, such as blood pressure, blood cholesterol level, hemoglobin (Hg/dl) etc. This is a useful but a crude measure in medical sciences as it provides a quick value of variation. The range of the data set, given in Table is 1.5, 98 - 45 = 53 kg. [Maximum Value – Minimum Value].

(ii) Standard Deviation

The most widely used and stable measure of dispersion is the *standard deviation*. This is a square root of *variance*. The variance is defined as mean squared deviation about the mean. The standard deviation (s.d) for the ungrouped data is calculated as:

$$\text{Standard Deviation } (\sigma) \;=\; \sqrt{\frac{1}{n}\left[\sum x_i^2 \;-\; \frac{(\sum x_i)^2}{n}\right]} \tag{1.3}$$

For dealing with frequency table we have

$$\text{Standard Deviation } (\sigma) \;=\; \sqrt{\frac{1}{\sum f_i}\left[\sum f_i x_i^2 - \frac{\left(\sum f_i x_i\right)}{\sum f_i}\right]} \tag{1.4}$$

The computation of standard deviations for grouped data for population and sample is shown as:

Table 1.12:
Computation of mean, variance and standard deviation

Weight (kg)	Number of students (f)	Mid-Points of weight (x_i)	frequency & Mid-Points ($x_i f_i$)	Frequency * (Mid-Points)^2
44.5 - 51.5	3	48	144	6912
51.5 - 58.5	18	55	990	54450
58.5 - 65.5	33	62	2046	126852
65.5 - 72.5	29	69	2001	138069
72.5 - 89.5	23	76	1748	132848
89.5 - 86.5	11	83	913	75779
86.5 - 93.5	2	90	180	16200
93.5 - 100.5	1	97	97	9409
Total	120	580	8119	560519

$$\text{Mean} = \frac{8119}{120} = 67.658 \text{ kg}$$

Using (1.4) the standard deviation comes out as:

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{1}{120}\left[560519 - \frac{(8119)^2}{120}\right]} = 9.661 \text{ kg (Population)}.$$

Assuming that this data is a sample from a certain population.

$$\text{Standard Deviation } (s) = \sqrt{\frac{1}{120-1}\left[560519 - \frac{(8119)^2}{120}\right]} = 9.702 \text{ kg (Sample)}$$

The variance of population $\sigma^2 = (9.661)^2 = 93.334$ kg, whereas the variance of sample $s^2 = (9.70)^2 = 94.129$. Difference will be only marginal if $\sum f_i$ is large. The only difference between population standard deviation and sample standard deviation is that in sample standard deviation the divisor is total number of observations minus 1, i.e. (n - 1) or $\sum f_i - 1$.

Note that Mean, Median, Mode, variance and standard deviation may be calculated directly from the grouped data by using IBM-SPSS package. For this purpose one should follow the following steps.

Example S1-8

1. enter the required mid points in one column and enter frequencies in another column

2. Click *DATA* and click *weight cases*, bring the frequencies to right hand side and click *ok*

3. Click *analysis*, then *frequencies*, mark mean median mode etc.

Now Click on to get the following results:

Mid-Points of weight

N	Valid	120
	Missing	0
Mean		67.66
Median		69.00
Mode		62
Std. Deviation		9.702
Variance		94.126

Note also that the mean of ungrouped data (raw data) is 67.625 kg and standard deviation is 9.488 kg whereas in grouped data the mean is 67.658 kg and the standard deviation is 9.661 kg. Note that grouped and ungrouped data results are close to each other. The difference (*error*) coming in the results is due to the grouping. When raw data is grouped, it loses some information. If a different grouping of the same is made then the mean and standard deviation are different. In grouped data it is assumed that all the values lying in that group correspond to the mid-value of the group. When a statistical package is used, these measures are calculated from raw data directly. *Note that when you transfer the observations from one media to another one, some information are lost[1]*.

From the example given in table 1.5, new groups are formed as in table 1.13:

Table 1.13:
Computation of mean

Weight (1)	Number of students (2)	Mid-points of groups (3)	2 x 3
45-52	3	48.5	145.5
52-59	18	55.5	999.0
59-66	32	62.5	2000.0
66-73	28	69.5	1946.0
73-80	24	76.5	1836.0
80-87	12	83.5	1002.0
87-94	2	90.5	181.0
95-101	1	97.5	97.5
Total	120	584.0	8207.0

$$\text{Mean} = \frac{8207}{120} = 68.392 \text{ kg.}$$

The mean with the first grouping is 67.658 kg whereas the mean with new grouping is = 68.392 kg. Therefore, we see if the groups are changed the mean is also different. *The mean and standard deviation calculated from a raw data are always exact.*

1.9.6 Relative Measure

All the measures we have so far discussed are called absolute measures, that is, these are measured *in terms of their basic units*. Suppose there are two sets of data of the same type but these are measured in different units (weights in kilograms and in pounds) and we want to compare two sets of data. Even if the standard deviation of one set of data is less than the standard deviation of another set of data, we cannot say that the first set of data is less scattered than the second set of data. We cannot make such comparison, as the basic units are different. Measures, which enable us to make such comparisons, are free of units and are called *Relative Measures*. Some of the useful and commonly used relative measures are: (i) Coefficient of variation (ii) Z-score.

---

[1]*This point will be explained while applying the logistic regression (Chapter 9).*

(i) Coefficient of Variation

We know that in central tendency mean is the best measure among the group and in measure of dispersion standard deviation is the best measure then these two measures are used to establish an index called coefficient of variation. If the units of the two or more data sets are different then coefficient of variation is the best method for comparison.

Coefficient of variation (C.V) is a relative measure of variation in any variable and is defined by

$$C.V = \frac{\text{sample standard deviation}}{\text{sample mean}} \times 100 = \frac{s}{\bar{x}} \times 100$$

(1.5)

$$C.V = \frac{\text{population standard deviation}}{\text{population mean}} \times 100 \text{ (for population)} = \frac{\sigma}{\mu} \times 100 \qquad (1.6)$$

Note that, if one is comparing two or more data sets, then, a data set, which has less Coefficient of Variation is more *consistent*, more *homogeneous* and more *stable* than a data set that has larger C.V.

The coefficient of variation is a useful measure of relative spread in data and is used frequently in the biological sciences. For example, suppose the authors of the study on diet and lipoproteins want to compare the variability in the ratio of total/HDL cholesterol with the variability in vessel diameter change for the 18 patients who had no lesion growth. The mean and the standard deviation of total/HDL cholesterol (in mill moles per liter) are 5.81 and 1.20, respectively; for the vessel diameter change (in millimeters), they are 0.12 and 0.29, respectively. A comparison of 1.20 and 0.29 makes no sense because cholesterol and vessel diameter are measured on different scales. The coefficient of variation adjusts the scales so that a sensible comparison can be made.

Variation, as measured by the standard deviation, is small relative to the mean. Therefore, readers of their article can be confident that the assay results were consistent. From this formula, the CV for total/HDL cholesterol is (1.20/5.81) (100) = 20.7%, and the CV for vessel diameter change is (0.29/0.12) (100) = 241.7%. Therefore, we can conclude that the relative variation in vessel diameter change is much greater than (more than 10 times as great as) that in cholesterol ratio.

A frequent application of the coefficient of variation is in laboratory testing and quality control procedures. For example, screening for neural tube defects is accomplished by measuring maternal serum alpha fetoprotein. DiMaio et al. (1987) evaluated the use of this test in a prospective study of 34,000 women. The reproducibility of the test procedure was determined by repeating the assay ten times in each of four pools of serum. They calculated the mean and the standard deviation of the ten assays in each pool of serum and then used them to find the coefficient of variation for each pool. The coefficients of variation for the four pools were 7.4%, 5.8%, 2.7%, and 2.4%. These values indicate relatively good reproducibility of the assays because the variation, as measured by standard deviation, is small relative to the mean. Therefore readers of their articles can be confident that the assay results were consistent

Example 1.7:

In the following table, data are given relating to collection of blood and to compare two methods of coagulation. The data are related to the arterial activated partial ehromboplastin time (APTT). Values are recorded for 30 patients in each of two groups. Do these data indicate the difference in the distribution of APTT times?

Table 1.14:

METHOD 1:					
20.7	29.6	34.4	56.6	22.5	29.7
31.2	38.3	28.5	22.8	44.8	41.6
24.9	29.0	30.1	33.9	39.7	45.3
22.9	20.3	28.4	35.5	22.8	54.7
52.4	20.9	46.1	35.0	46.1	22.1
METHOD 2:					
23.9	23.2	56.2	30.2	27.2	21.8
53.7	31.6	24.6	49.8	22.6	48.9
23.1	34.6	41.3	34.1	26.7	20.1
38.9	24.2	21.1	40.7	39.8	21.4
41.3	23.7	35.7	29.2	27.4	23.3

Solution:

These information relate to two data sets (groups), and these two groups are not selected from any population(s). We like to see which method is better than the other by comparing two data sets. All the basic measures are calculated using IBM-*SPSS package* regarding two methods through:

Analyze → Descriptive Statistics → Frequency →
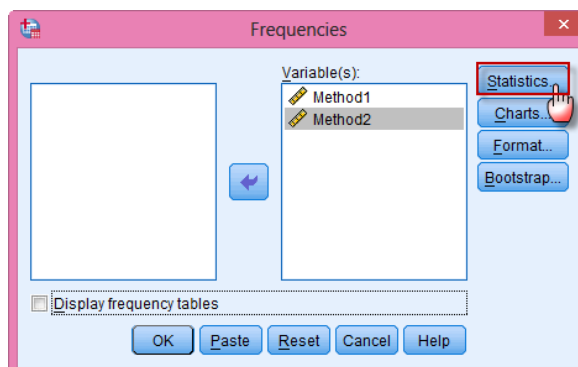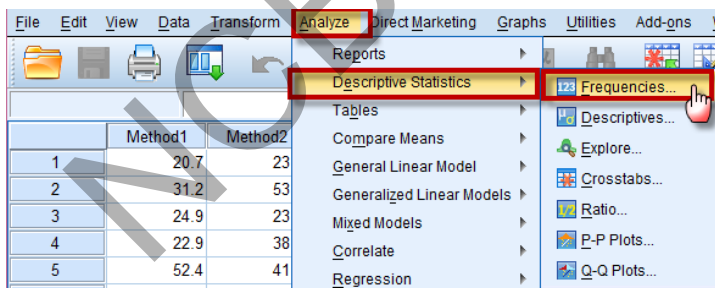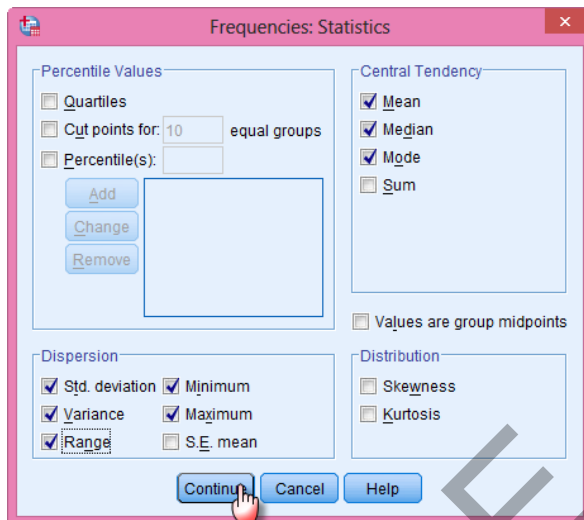
and the output is given on next table.

SPSS output for Descriptive Measures

Table 1.15:
Different values of the descriptive measures
Statistics

		Method1	Method2
N	Valid	30	30
	Missing	0	0
Mean		33.693	32.010
Median		30.650	28.300
Mode		22.8[a]	41.3
Std. Deviation		10.7298	10.4586
Variance		115.130	109.383
Range		36.3	36.1
Minimum		20.3	20.1
Maximum		56.6	56.2

a. Multiple modes exist. The
smallest value is shown

If we cannot reach any decision by using mean and standard deviation, we go ahead for coefficient of variation.

We know that by looking at the mean we cannot reach any conclusion unless we go ahead for standard deviation. Method 2 has less standard deviation than the data collected by Method 1, therefore, we say that Method 2 of taking the blood is better than Method 1. To be sure we go ahead to relative measure (coefficient of variation COV). The coefficient of variation of data collected by Method 2 is less than the coefficient of

variation of the data collected by Method 1. Therefore, we confirm our decision that data collected by Method 2 is more consistent, more homogeneous and more stable than Method 1. One can calculate the C.V by hand to get:

Table 1.16:
Coefficient variation

	Method 1	Method 2
Mean	33.693	32.010
Standard deviation	10.730	10.459
Coefficient variation	31.85%	30.67%

Note that Coefficient of Variation is not available directly in IBM-SPSS Package, unless we add a dummy variable of 1's and use the Ratio. For example, to calculate the C.Vfor Method 1, we do as follows:

Now Click on **OK** to get the following results:

Ratio Statistics for Method1 / Dummy

	95% Confidence Interval for Mean			95% Confidence Interval for Median							Coefficient of Variation
Mean	Lower Bound	Upper Bound	Median	Lower Bound	Upper Bound	Actual Coverage	Minimum	Maximum	Std. Deviation	Range	Mean Centered
33.693	29.687	37.700	30.650	28.400	38.300	95.7%	20.300	56.600	10.730	36.300	31.8%

The confidence interval for the median is constructed without any distribution assumptions. The actual coverage level may be greater than the specified level. Other confidence intervals are constructed by assuming a Normal distribution for the ratios.

(ii) Z-Score
Z-score is also a relative measure of a variable and is defined as

$$Z = \frac{\text{Value of variable (x)} - \text{Population mean}}{\text{Population standard deviation}} \qquad (1.7)$$

Example 1.8:
A student's average grade in Pharmacology is 67 and in Bio-statistics is 87. If the class means and standard deviation in Bio-statistics is 80 and 5 respectively, whereas in Pharmacology the mean and standard deviation is 79 and 8 respectively, then find the Z-scores in these subjects and interpret the results.

Solution:
The Z-scores in these two subjects are:

Subject	Z-score
Bio-statistics	$\frac{87 - 80}{5} = 1.4$
Pharmacology	$\frac{67 - 79}{8} = -1.5$

Z-score in Bio-statistics is 1.4, i.e. 1.4 times standard deviation above the mean of the class whereas in Pharmacology the Z-score is -1.5, which means 1.5 times standard deviation below the mean of the class. Thus Z-score measures his ability in relation to his class and is free of unit measure. *Note that the variable* Z *has mean = 0 and standard deviation =1 (details will be given later)*

Note that we can obtain the Z-score for all the values as in the following steps:
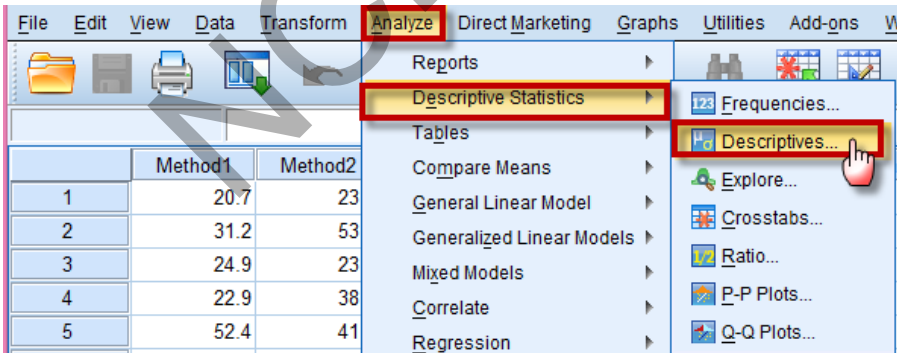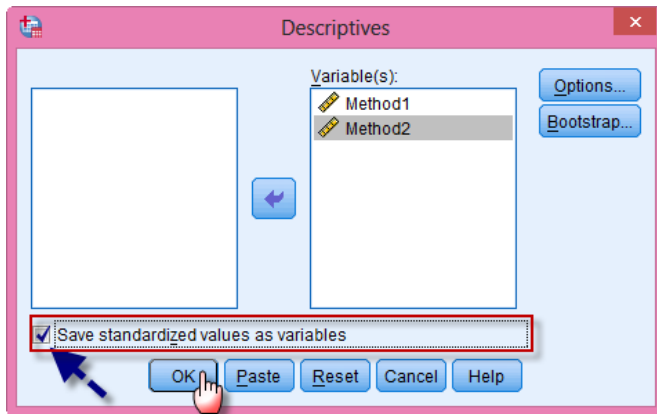
The default outcome is as in the following table:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Method1	30	20.3	56.6	33.693	10.7298
Method2	30	20.1	56.2	32.010	10.4586
Valid N (listwise)	30				

And the Z-scores are added directly to the data. Here is first ten for both variables:

	Method1	Method2	ZMethod1	ZMethod2
1	20.7	23.9	-1.21095	-.77544
2	31.2	53.7	-.23237	2.07388
3	24.9	23.1	-.81952	-.85193
4	22.9	38.9	-1.00592	.65879
5	52.4	41.3	1.74342	.88826
6	29.6	23.2	-.38149	-.84237
7	38.3	31.6	.42933	-.03920
8	29.0	34.6	-.43741	.24764
9	20.3	24.2	-1.24823	-.74675
10	20.9	23.7	-1.19231	-.79456

1.10 Mean ± k × Standard Deviation

What percentage of observations falls within mean ± k × s.d. (k = 1, 2, 3). Empirically it is known that, for a reasonably large set of data having a bell shaped frequency curve (symmetrical curve), about 68% of the observations fall within mean ± 1 s.d, about 95% of the observations fall within mean ± 2 s.d and 99% of the observation fall within mean ± 3 × s.d. (details will be discussed at a later stage).

The advantage of this empirical rule is, if we do not have the data and only mean and standard deviation are known, then one can calculate the ranges where 64% to 68%, 95% and 99% of the data are lying. For example, the mean and standard deviation of raw data are, mean is 67.625 and s.d = 9.488 respectively, then the weight of about 68% of the students is lying between 58 to 77Kg., and the weight of 95% of the students will be lying between 49 to 87 Kg etc.

Chapter 2

Basic Concepts of Probability and Probability Distributions

2.1 Introduction

Different methods of summarizing the statistical data along with their graphical presentations have been discussed in Chapter 1. This chapter deals with the basic concepts of probability and probability distributions. The purpose of this chapter is not to teach probability to medical students and research workers but to clarify some of the basic concepts involved in understanding the interpretation of the results. For example, a major reason for performing clinical research, however, is to generalize the findings from a set of observations on one group of subjects to other similar groups of subjects. If we are interested to study whether smoking causes lung cancer, or it leads to cardiac problems, it is not possible to study all the persons who smoke. We investigate a small group of smokers selected from a larger group. The conclusion may indicate that smokers run a greater risk of lung cancer or a myocardial infarction. We say that smokers may have more chance of lung cancer than non-smokers. The term chance in the statistical language is designated as *probability*. A sample rarely tells us precise story about the population from which it is selected. There is always uncertainty about how far the sample estimate will depart from the true population value. Measures of the amount of uncertainty associated with estimate play a major role in statistical inference. How do we measure the uncertainty associated with events? The answer is probability. The concept of probability is very useful in understanding and interpreting statistical data. It helps us to understand the *confidence limits, p-value* (will be discussed in chapter 4) and the terms like *significance* and *non-significance*.

Whenever one deals with the probability, one faces the word *experimen*t. This word has very broad meaning. An experiment is a process of making observations or taking measurements on one or more experimental units. An experiment can be repeated many times. Each replication is called a *trial.* One or more outcomes can result from each trial. Consider a large number of trials. The probability of a specific outcome is the number of times that the specific outcome occurs divided by the total number of trials. If E is an event then the probability of an event will be defined as:

$$P(E) = \frac{\text{Number of times E occurs in an experiment}}{\text{Total number of trials in an experiment}}$$

If the number of trials is very large this ratio is generally seen to be fairly stable from one instance to another.

An estimate of the probability may be determined empirically or it may be based on theoretical model. If we flip a coin, the chance of getting a head or a tail is 50%. If this coin is flipped, say 20 times, there is no guarantee that exactly 10 heads will be observed.

Then again if the coin is tossed 2,000 times the ratio of number of heads to total number of trial will be very near to $1/2$. The frequency of the heads may vary from 0 to 2000, though in each case the chance of getting the head is 50%.

2.2 Definition and rules of probability

We will describe some examples to illustrate the concept of probabilities.

Example 2.1

Following data relate to total circulating albumin (gm) for 30 normal males aged 20-29.

Table 2.1
Distribution of males by total circulating albumin(gm)

Total circulating albumin (gm)	Number of males	Relative frequency
99.5-109.4	2	2/30
109.5-119.4	6	6/30
119.5-129.4	6	6/30
129.5-139.4	7	7/30
139.5-149.4	8	8/30
149.5-159.4	1	1/30
Total	**30**	**1.00**

Suppose a person is picked up at random, the probability that the person belongs to the group 119.5-129.4 is 6/30, which in fact is a relative frequency of this group. It means that of 30 persons, 6 belong to the group 119.5 - 129.4.

Example 2.2

In a study of the relation between blood type and disease, a sample of patients with peptic ulcer, patients with gastric cancer and control persons that are free from these diseases are classified into the blood type (O,A,B). The data are given in table 2.2:

Table 2.2
Distribution of patients by disease and blood

Blood type	Peptic ulcer	Gastric cancer	Controls	Total	Probability
O	983	383	2892	4258	0.486
A	679	416	2625	3720	0.424
B	134	84	570	788	0.090
Total	1796	883	6087	8766	1.0
Probability	0.205	0.101	0.694	1.0	

Source: Snedecor and Cochran (1980)

In presenting this problem one can easily determine the probability that a patient selected at random will fall in blood group O or A or B or he/she is suffering from peptic ulcer or gastric cancer.

The probability that a person selected at random from 8766 cases falls in blood type O group will be

$\textbf{P}$(blood type O) = 4258/8766 = 0.486.

Again the probability that a person selected at random from 8766 cases belongs to peptic ulcer group will be

$\textbf{P}$(peptic ulcer) = 1796/ 8766 = 0.205.

If we add the probabilities of blood type O, A, and B it comes out to be 1 0

(see Table-2.2). The value can be zero if no patient is in a group and can be 1 if all the patients fall in that group. Therefore two important results can be drawn from this:

i)   The sum of all the probabilities of all possible outcomes of an experiment is equal to 1.
ii) The probability of each outcome (blood type or type of disease) is greater than or equal to zero but cannot be greater than 1 or less than zero.

Therefore a general rule can be stated that the probability of any outcome lies between 0 and 1, both ends inclusive.

$$0 \le \textbf{P}(A) \le 1. \tag{2.1}$$

The probability that a selected person does not belong to blood type O, will be $1 - \textbf{P}$ (with blood type O) = 1- 0.486 = 0.514, as the total probability is 1 this is such because a person either falls in blood type O group or does not falls in blood type O group.

2.2.1 Additive Rule of Probability for Mutually Exclusive Events

Before we explain the additive law of probability it is essential to understand an *event and mutually exclusive events*. *An event may be defined as either a single outcome or a set of outcomes of an experiment*. Two or more events are *mutually exclusive* if the occurrence of one event precludes the occurrence of another event. In the above example, a person cannot have a blood type O or A at the same time, therefore blood type O and A are mutually exclusive events.

Suppose the probability of blood type O = 0.486 whereas the probability of blood type A = 0.424. The probability of blood type O or A will be

$$\textbf{P}(\text{O or A}) = \textbf{P}(\text{O}) + \textbf{P}(\text{A}) = 0.486 + 0.424 = 0.91. \tag{2.2}$$

This is known as an additive law of probability for mutually exclusive events.

2.2.2 Independent Events and Multiplicative Rule of Probability

If the outcome of one event does not affect the outcome of another event then these events are called independent events. If two events A and B are independent then the probability that both A and B occur is equal to the product of their respective probabilities i.e.

$$\textbf{P}(\text{A and B}) = \textbf{P}(A)\,\textbf{P}(B). \tag{2.3}$$

Suppose two coins are tossed. The probability that heads occur on both coins i.e. $\textbf{P}$ (two heads) = $\textbf{P}$ (H_1 and H_2) = $\textbf{P}(H_1)\,\textbf{P}(H_2)$, where H_1 denotes the head on first coin and H_2 head on the second coin. Since $\textbf{P}(H_1) = \textbf{P}(H_2) = \dfrac{1}{2}$ therefore $\textbf{P}(H_1 \text{ and } H_2) = \dfrac{1}{4}$.

2.2.3 Additive Rule for non Mutually Exclusive Events

Let us now examine the situation for finding out the probability that either of the two events occur, when they are not mutually exclusive. For example, type of peptic ulcer and patients with blood group O is not mutually exclusive. The additive rule of the probability can be modified otherwise the probability that both events occur will be added twice into the calculated probability. The probability that a randomly selected person has a peptic ulcer = 1796/8766 = 0.205 and the person has blood type O = 4258/8766 = 0.486. Here the joint probability of being ulcer and has a blood type O has been added twice. This joint probability of being peptic ulcer and have a blood type O = 983/8766 = 0.112 must be subtracted from the calculated probability, i.e.

$$P(\text{ peptic ulcer or blood type O}) = P(\text{peptic ulcer}) + P(\text{blood type O})$$
$$- P(\text{peptic ulcer and blood type O})$$

$$= 0.205 + 0.486 - 0.112 = 0.579.$$

Therefore the additive law of probability for non-mutually exclusive events may be stated as:

The probability that either event A or an event B or both occur is

$$\mathbf{P}(A \text{ or } B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \text{ and } B). \tag{2.4}$$

If A and B are mutually exclusive then the $\mathbf{P}(A \text{ and } B) = 0$.

2.2.4 Conditional Probability

The probability of an event A, given that an event B i.e. $\mathbf{P}(A|B)]$ has occurred, is called the conditional probability of A given B, is defined as:

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \text{ and } B)}{\mathbf{P}(B)}, \tag{2.5}$$

and

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(A \text{ and } B)}{\mathbf{P}(A)}. \tag{2.6}$$

The probability of a person selected at random has a peptic ulcer given that he has blood type O.

$$\mathbf{P}(\text{peptic ulcer} \mid \text{blood type O}) = 983/ 4258 = 0.231.$$

This may also be calculated using (2.5).

$$\mathbf{P}(\text{peptic ulcer and blood type O})$$

$$= \mathbf{P}[\text{Peptic ulcer} \mid \text{blood type O}] \times P[\text{blood type O}]$$

$$= \frac{983}{4258} \times \frac{4258}{8766} = \frac{983}{8766} = 0.11.$$

Like wise **P** (peptic ulcer | blood type O) = $\dfrac{983}{8766} \times \dfrac{8766}{4258} = \dfrac{983}{4258} = 0.231$.

2.2.5 Rule of multiplication for non-independent events

The probability of A and B

$$\mathbf{P}(A \text{ and } B) = \mathbf{P}(A)\,\mathbf{P}(B|A) = \mathbf{P}(B)\,\mathbf{P}(A|B) \tag{2.7}$$

P(gastric cancer and blood type B) = **P**(gastric cancer) **P**(blood type B | gastric cancer)

$$\left(\frac{883}{8766}\right)\left(\frac{84}{883}\right) = 0.0096 \ .$$

2.2.6 Properties of Probability

1. The probability of any event always lies from 0 and 1.

2. If we list all possible events mutually exclusive and exhaustive, the sum of their probabilities is always 1.

3. If two events A and B are mutually exclusive, then the probability that either A or B occurs is equal to **P**(A) + **P**(B).

4. If two events A and B are independent then the probability of both A and B occurring together is equal to the product of their probabilities i.e. **P**(A and B) = **P**(A) **P**(B).

5. If two events A and B are not mutually exclusive, the probability that either A or B or both occur is equal to **P**(A) +**P**(B) -**P**(AB). If A and B are mutually exclusive then **P**(A and B) = 0.

6. The probability of an event A, given that B has already occurred, is called the conditional probability of A given B i.e. **P**(A|B).

7. The probability that both events A and B occur is

$$\mathbf{P}(A \text{ and } B) = \mathbf{P}(A)\,\mathbf{P}(A|B) = \mathbf{P}(B)\,\mathbf{P}(B|A) \tag{2.7}$$

Example 2.3:
The following data relates to Chinese smoking and lung cancer study in Beijing during 1990. Various types of probabilities can be calculated based on the data.

Table 2.3 Status of lung cancer by smoking

		Lung Cancer		
		Yes	No	Total
Smoking	Yes	126	100	226
	No	35	61	96
		161	161	322

(i)   The probability that a selected person has a lung cancer $= \dfrac{161}{322} = 0.50$.

(ii)   The probability that a selected person is smoker $= \dfrac{226}{322} = 0.702$.

(iii)   The probability that a man has a lung cancer given that he is smoker

$$= \frac{126}{226} = 0.56 = \frac{\left[\frac{126}{322} \right]}{\left[\frac{226}{322} \right]}.$$

(iv)   The probability that a man is smoker given that he has lung cancer

$$= \frac{126}{161} = 0.78 = \frac{\left[\frac{126}{322} \right]}{\left[\frac{161}{322} \right]}.$$

(v)   The probability that a man does not have lung cancer given that he is not

smoker $= \dfrac{61}{96} = 0.64 = \dfrac{\left[\frac{61}{322} \right]}{\left[\frac{96}{322} \right]}.$

(vi)   The probability that a man is not smoker given that he has lung cancer

$$= \frac{35}{161} = 0.22 = \frac{\left[\frac{35}{322} \right]}{\left[\frac{161}{322} \right]}.$$

(vii)   The probability that a man is smoker and does not have lung cancer

$$= \frac{100}{322} = 0.31.$$

(viii) The probability that a man is either smoker or lung cancer or both

$$= \frac{226}{322} + \frac{161}{322} - \frac{126}{322} = 0.81.$$

(ix)   The probability that a man is smoker and has lung cancer $= \dfrac{126}{322} = 0.39$.

        Or P (smoker and cancer) = P(smoker) P(cancer|smoker)

$$= \left(\frac{226}{322} \right) \left(\frac{126}{226} \right) = 0.39$$

(x)   The probability that a man is not smoker and does not have lung cancer

$$= \frac{61}{322} = 0.19.$$

(xi)   P (no smoker and no cancer) =P(no smoker) P(no cancer | no smoker)

$$= \left(\frac{96}{322} \right) \left(\frac{61}{96} \right) = 0.19 \cdot$$

2.3 Probability distribution

In order to understand the concept of probability distribution, the explanation of some terms is necessary.

(a) A *random variable* is a quantity whose value depends upon the outcome of an experiment. Random variable has two types (i) A *discrete random variable* is one that assumes a countable number of values and (ii) A *continuous random variable* assumes any value on an interval on a line.

(b) *Probability distribution* is a table or formula listing all possible values that a random variable can take alongwith associated probabilities. If the random variable is discrete then this distribution is called *discrete probability distribution* otherwise it is called *continuous probability distribution.* While discussing continuous random variable the number of possible values become infinite and cannot be listed. This is taken care of by considering probability density function, which we will discuss later. Binomial, Poisson and Normal distributions are some examples of probability distributions.

Regardless of whether a random variable is continuous or discrete its probability distribution must conform to the basic rules of probability (i) $0 \leq P (A) \leq 1$ and (ii) the sum of the probabilities of all the values of random variable must be 1.

2.3.1 The Binomial Probability Distribution

Frequently in health sciences, investigations are made in which the investigator is interested in one of the two possible outcomes; test is positive or negative, a patient is suffering with diabetes or not, or in general a person is suffering with some disease or not. The outcome may be called *success* and *failure*. When a single trial of some experiment can result in only one of the two mutually exclusive outcomes then the trial is called a *Bernoulli trial.* The probability of positive test is denoted by p whereas the probability of negative with q. Note that (q+p = 1). When such experiment is repeated n times under same conditions and X of them has some specific proposition then this distribution is known as *binomial probability distribution.* This distribution is named after a Swiss mathematician James Bernoulli (1654-1705).

A binomial experiment is one that possesses the following properties.

(i)   Each experimental unit results in an outcome that may be classified as a success or a failure.

(ii)   The random variable X counts the number of successes or failures in n trials.

(iii)   The probability of single experimental unit of success denoted by p, remains same (constant) from trial to trial.

(iv)   The outcome for any one experimental unit is independent of the outcome of another experimental unit (draws are independent).

The binomial distribution gives the probability that a specified outcome occurs in a given number of independent trials. The binomial distribution can be used to model the inheritability of a particular trait in genetics, to estimate the occurrence of a specific reaction, such as the single packet (quantal release) of acetylcholine at the neuromuscular junction, or to estimate the death of a cancer cell in an *in vitro* test of a new chemotherapeutic. Binomial distribution is useful in understanding the relative risk, odds ratio, sensitivity (true positive), specificity (true negative), false negative and false positive etc. (all these terms will be discussed in Chapter 7).

To develop the concept of binomial distribution let 5 coins be flipped. Suppose there are three heads and two tails. The outcome of a head is considered as a success whereas an outcome of a tail is a failure. The probability of success (S) is denoted by p whereas the probability of failure (F) by q (q=1–p). Since the trials are independent, according to multiplicative law of probability, the probability of a sequence S, S, F, F, S is:

$$\mathbf{P}(S, S, F, F, S) = p\,p\,q\,q\,p = p^3 q^2$$

The probability of a head or a tail of a coin is equal and is 0.50, therefore

$$\mathbf{P}(S, S, F, F, S) = 0.5^3\,0.5^2 = 0.03125.$$

If we make all possible arrangements of 3 heads and 2 tails it will appear in 10 possible ways. Therefore the probability of 3 heads when 5 coins are flipped will be

$$\mathbf{P}\ (3\ \text{heads and 2 tails}) = 10\ (0.5)^3 (0.5)^2 = 0.3125$$

If we take x = 3 and n = 5 (5 coins are tossed and 3 heads appeared) then we may easily write the formula to calculate the probability of x successes in an n trials as

$$\mathbf{P}\ (X\ \text{successes}) = \binom{n}{x} p^x q^{n-x}, \text{ for } x = 0,1,2,3,\ldots, n. \qquad (2.8)$$

$$= 0, \text{ otherwise}$$

where $\binom{n}{x}$ means that x things are taken from n

and $\binom{n}{x} = \dfrac{n!}{x!\ (n - x)!}$, and $n! = n(n\text{-}1)\ (n\text{-}2)\ \ldots\ldots\ (2)(1)$,

and 0!=1.

If, in this formula, we put x=3 and n= 5 we get the required probability.

$$P(X = 3) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{5}{3}(0.5)^3 (1-0.5)^2 = 0.3125,$$

When sample size is large, it is difficult to calculate the probability. In this case we can use Tables given at the end of the Chapter. Without going into details of derivation, the mean and standard deviation of the binomial distribution are:

$$\mu = np \text{ and } \sigma = \sqrt{np(1-p)} \tag{2.9}$$

Example 2.4:

The probability of death with certain disease is 40%. Five such patients are admitted in the hospital, what is the probability that exactly 3 of them die?

Solution:

Here p = 0.4, q = 1-0.4 = 0.6, and n = 5 the probability that exactly 3 of them will die is

$$\mathbf{P}(X=3) = \binom{5}{3} (0.4)^3 (0.6)^2 = 10 (0.4)^3 (0.6)^2 = 0.2304$$

Instead of calculating the probability, table of cumulative binomial probability can be consulted to find the probability. These tables are available in any book on statistics. For ready reference a portion of the table has been reproduced at the end of the chapter.

Probability for n=5, x = 3 and p = 0.4 is 0.2304. Since in the table, cumulative probability is given therefore,

$$\mathbf{P}(X = 3) = \mathbf{P}(X \le 3) - \mathbf{P}(X \le 2)$$

From the table for n=5, x=3 and p=0.4, we get

$$\mathbf{P}(X=3) = 0.9130 - 0.6826 = 0.2304.$$

Example 2.5:

The dairy industry is capitalizing on new medical research in the field of osteoporosis (an age related condition characterized by decreased bone mass and increased susceptibility to fractures) to promote its product. According to the National Institute of Health, by the age of 90, 32% of women and 17% of men will suffer a hip fracture because of osteoporosis (American Demographics, Oct. 1985). Find the probability that (a) in a random sample of 5 women aged 90, exactly three have suffered a broken hip due to osteoporosis, (b) at least two of the 5 women have suffered a broken hip due to osteoporosis, and (c) at most three have suffered a fractured hip due to osteoporosis.

Solution:

$\mathbf{P}$(women with hip fracture) = 0.32 $\mathbf{P}$(men with hip fracture)=0.17

(i) n = 5, x = 3;   (ii) n = 5, x ≥ 2

(a) For women with hip fracture

   (i) n = 5, x = 3 , p =.32

$$\mathbf{P}(X = 3) = \mathbf{P}(X \le 3) - \mathbf{P}(X \le 2) = 0.9610 - 0.809 = 0.1515. \text{ (from the table)}$$

   (ii) n = 5, x ≥ 2 , p = 0.32

The **P** $(X \geq 2) = $ **P**$(X = 2) + $ **P**$(X = 3) + $ **P**$(X = 4) + $ **P**$(X = 5)$

since the total probability $= 1$, therefore the

$\quad$ **P**$(X \geq 2) = 1 - $ **P** $(X \leq 1) = 1 - 0.4875 = 0.5125$ (from the Table 2.4).

(b) for men with hip fracture

$\quad$ n = 5 and x $\leq$ 3, p = 0.17 we are interested to find **P**(0) +**P**(1))+**P**(2)+**P**(3) for this one can consult the Table 2.4 directly against x = 3 and p = 0.17 which is 0.9964.

Example 2.6:

One of the most comprehensive studies of drug used in junior high school was conducted in U.S.A. The survey of 1,532 eighth-grade students found that 25% believed they would use marijuana and 11% believed they would use cocaine by the time they enter high school (Alligator, Sept. 27, 1984). A representative of the community group that conducted the study claims that these results are applicable nationwide. Consider a random sample of 10 eighth- graders selected from a school. Assume that the result is applicable nationwide, find the probability that (a) exactly 5 of the eighth-graders believe they will use marijuana before entering high school (b) at least 2 of the eighth- graders believe they will use marijuana before entering high school and (c) at most three of the eighth-graders believe they will use cocaine before entering school.

Solution:

$\quad$ (a) The Probability of students using marijuana = **P**(marijuana) = 0.25,

$\quad$ (b) The probability of students using cocaine = **P**(cocaine) = 0.11

(a) p = 0.25, n = 10, x = 5, then

$\quad$ **P** $(X = 5) = $ **P** $(X \leq 5) - $ **P** $(X \leq 4) = 0.9803 - 0.9219$

$\quad\quad = 0.0584$ (from the table 2.4)

(b) p =.25, n=10, **P**$(X \geq 2) = 1 - $ **P** $(X \leq 1) = 1 - 0.2440 = 0.776$ (Table 2.4)

(c) p = 0.11, n= 10, x=3, then **P**$(X = 3) = $ **P**$(X \leq 3) + $ **P**$(X \leq 2)$

$\quad\quad = 0.9822 - 0.9116 = 0.706$ (Table 2.4)

Example 2.7:

A physician claims that only 10% of all American adults suffer from high blood pressure. The American Medical Association conducted a study involving 1,200 randomly selected American adults. Find the mean number of adults in the sample who suffer from high blood pressure, and standard deviation of adults with high blood pressure if the physician's claim is true.

Solution:

The probability of adults having blood pressure p = 0.1 and n = 1200;

(i) The mean number of adults who suffer from high blood pressure

$\quad\quad = np = 1200 \times 0.10 = 120$

(ii) The standard deviation is

$$s = \sqrt{np(1-p)} = \sqrt{1200(0.10)(1-0.10)} = 10.39$$

Using empirical rule, the limits will be $120 \pm 2 \times 10.39 \sim (99 \sim 141)$ there are about 95% chances that people suffering with blood pressure will lie between 8.25% ~ 11.75% in a population.

2.3.2 The Poisson Probability Distribution

Like the binomial distribution, Poisson distribution is also a discrete probability distribution. This distribution is named after the French mathematician S.D. Poisson. The use of this distribution is extensive in biology and medicine. Poisson distribution is used to determine the probability of rare events: i.e. it gives the probability that an outcome occurs a specified number of times when the number of trials is large and the probability of occurrence is very small.

Poisson distribution is used to plan the number of beds a hospital needs in the intensive care unit; the number of ambulances needed on call in a certain hospital. This is a useful distribution for estimation of bacteria in colonies. It can also be used to model the number of cells in a given volume of fluid; the number of bacterial colonies growing in a certain amount of medium.

A Poisson experiment is one that possesses the following three properties:

(i)   The number of outcomes occurring in one time interval is independent of the number in any disjoint time interval,

(ii) The probability that a single outcome will occur during a very short time interval is proportional to the length of the time interval and does not depend on the number of outcomes occurring before this time or on the past history of the process,

(iii) The probability that more than one outcome will occur in such a short time interval is negligible

A random variable X taking on one of the values 0,1,2 …. is said to be a Poisson random variable with parameter μ if for some $\mu > 0$, its probability distribution is

$$P(X) = \frac{e^{-\mu}\mu^{x}}{x!}, \quad x = 0,1,2,....\infty \tag{2.10}$$

where e stands for constant and is approximately 2.7183, and μ is the parameter of the distribution and is the average number of outcomes occurring in a given time interval. Some examples of random variables are given which usually follow Poisson distribution:

(i)   The number of people in a community living up to 100 years of age.

(ii) The number of α - particles discharged in a fixed period of time from some radioactive material.

    (iii) The number of wrong telephone numbers that are dialed in a small interval of time.

    (iv) The number of sudden deaths of healthy men in a small interval of time period.

Note that the both mean and standard deviation of Poisson distribution is μ.

Example 2.8

The probability that a person dies from certain respiratory infection is 0.002. Find the probability that (i) less than 5 of the next 2000 persons so infected will die (ii) exactly 5 will die.

Solution:

    p= 0.002, n= 2000, x = 5, mean = $\mu = np = 2000 \times 0.002 = 4$

(a) $\mathbf{P}(X < 5) = \mathbf{P}(X \leq 4) = 0.629$ [table 2.5]

(b) $\mathbf{P}(X = 5) = \dfrac{e^{-4}4^5}{5!} = \dfrac{0.0183 \times 124}{120} = 0.156.$

Like binomial distribution, probability for the Poisson distribution may also be calculated using the cumulative probability table. For this purpose a portion of the table has been reproduced for ready reference at the end of this chapter (Table 2.5).

We consult the table to see the probability for μ

    $\mathbf{P}(X = 5) = \mathbf{P}(X \leq 5) - \mathbf{P}(X \leq 4) = 0.785 - 0.629 = 0.156$

Example 2.9:

The probability that a student fails the screening test for scoliosis (curvature of the spine) at a local high school is known to be 0.004. 1500 students are selected for such a test. Find the probability that (i) less than 5 will fail the test (ii) not more than 4 will fail the test.

Solution:

p= 0.004, n = 1500, (i) we find $\mathbf{P}(X < 5)$,

              (ii) $\mathbf{P}(X \geq 4)$ mean $= \mu = np = (0.004)(1500) = 6.$

(i) $\mathbf{P}(X < 5) = \mathbf{P}(X = 0) + \mathbf{P}(X=1) + \mathbf{P}(X=2) + \mathbf{P}(X=3) + \mathbf{P}(X=4) = \mathbf{P}(X \leq 4)$

      $= 0.285$ [table 2.5]

(ii) $\mathbf{P}(X \geq 4) = 1 - \mathbf{P}(X \leq 3) = 1 - 0.151 = 0.849$[table 2.5]

2.3.3 The Normal Probability Distribution

One of the most useful models frequently used is the Normal probability model. This model has not only wide application in mathematics and statistics but also in medical and social sciences. This distribution is continuous unlike binomial and Poisson distributions. The graph of the normal distribution is known as *normal curve*. The shape of the normal distribution is shown as:

μ - 3σ   μ - 2σ   μ - σ   μ   μ + σ   μ + 2σ   μ + 3σ
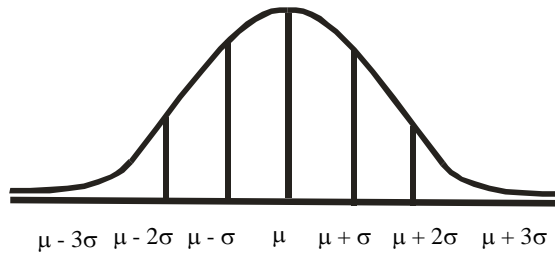
Fig. 2.1: The shape of the normal distribution

The area under the normal curve is always used as a reference value in order to draw any conclusion about any experiment. The laboratory investigations of any suspected patient are always compared with the standard (healthy person) value in order to draw any conclusion. If the readings of the investigation fall within the limits of the standard value, it is always considered that a suspected patient is out of the dangerous zone. Exactly in the same way the findings of an experiment are compared with the values of normal distribution and conclusions are drawn accordingly. This concept will be explained in the remaining chapters.

This distribution was discovered by DeMoivre in 1733 and was developed by Gauss (1777-1855). Sometimes this probability distribution is known as Gaussian distribution. We will use the word normal for this distribution, as it is very familiar to social and medical scientists.

Much can be discussed regarding normal distribution but we will limit ourselves with the application for medical scientists. If X is a continuous random variable with mean μ and standard deviation σ then the probability density function of the normal distribution will be

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \ e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{2.11}$$

where $\pi = 3.14159$, $e = 2.71828$, $-\infty < X < +\infty$ and $\sigma > 0$.

The mean measures the location of the distribution and standard deviation measures the spread. The mathematical equation of the normal distribution depends on two parameters μ and σ. It is usually written as $X \sim N\ (\mu, \sigma^2)$ and read as, X is normally distributed with mean $= \mu$ and variance $= \sigma^2$.

Since the values of μ and σ vary from one normal distribution to another, the easiest way to express a distance from mean is in terms of a Z- score,

$$Z = \frac{X - \mu}{\sigma} \tag{2.12}$$

This is distance between X and μ, expressed in units of σ, Z is commonly known as *standard normal variable (variate)* with mean $= 0$ and variance $= 1$ and is written as $Z \sim N\ (0, 1)$. The equation of the standard normal distribution is

$$f(z) \ = \ \frac{1}{\sqrt{2\pi}} \ e^{-\frac{1}{2}z^2} \quad -\infty < z < +\infty. \tag{2.13}$$

Probability of any part of the curve can be calculated by the method of integration. Since this is difficult to calculate for medical scientists, therefore table for the standard normal distribution has been provided at the end of this chapter (Table 2.6). In order to calculate the area under the curve of the normal distribution the general equation is converted into standard equation by using the standard normal variable and the table is consulted to calculate the probability.

This curve is symmetric about the mean value. About 68% of the area lies between $\mu \pm 1 \sigma$, about 95% between $\mu \pm 2 \sigma$ and about 99% lies between $\mu \pm 3 \sigma$. This approximately agrees with the empirical rule stated in Chapter 1. Note that areas under the normal curve have a probabilistic interpretation. If a population of measurements has approximately normal distribution, then the probability that a randomly selected observation falls in the interval $\mu \pm 2\sigma$ is approximately 95%, but area between $\mu \pm 1.96$ σ is exactly 95% (Fig. 2.4). Medical scientists usually use the value 2 rather than 1.96 because of convenience. The area under normal curve beyond a value of Z is known as p-value. For a given Z=1.3, the p-value is P $(Z \geq 3) = 0.0968$. Some of the properties of the normal distribution are as follows:

(i)   It is symmetrical about the mean value therefore half of the probability of this distribution is on the right of the mean and half on the left of the mean.

(ii) The total area under the curve is equal to 1.

(iii) Mean, median and mode are equal.

(iv) It is completely determined by mean and standard deviation.

Example 2.10:
Given the standard normal distribution $\sim N (0, 1)$, calculate the probability that

(a) (i) **P**$(Z \leq -1.96)$ (ii) **P**$(Z \geq 1.96)$ (iii) **P**$(-1.96 \leq Z \leq +1.96)$

(b) (i) **P**$(Z \leq -2.58)$ (ii) **P**$(Z \geq 2.58)$ (iii) **P**$(-2.58 \leq Z \leq +2.58)$

(c) (i) **P**$(Z \leq -2.33)$ (ii) **P**$(Z \leq -1.65)$.

Solution:
It is always advised to sketch a diagram of normal distribution before solving the problem as it makes things easier and also errors in calculation are avoided.

(a) (i) **P**$(Z \leq -1.96)$ = probability from $-\infty$ to -1.96. In the Table 2.6 cumulative probability is given, therefore we can see the table directly and set the value 0.0250.

Fig. 2.2
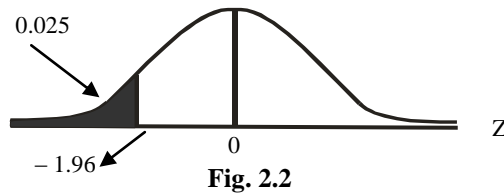
(ii) The curve is symmetrical, therefore **P** (Z ≥ 1.96) = **P** (Z ≤ -1.96) = 0.0250 or it may be calculated as

P (Z ≥1.96) = 1 − **P** (Z ≤ 1.96) = 1 - 0.9750 = 0.0250.

Fig. 2.3
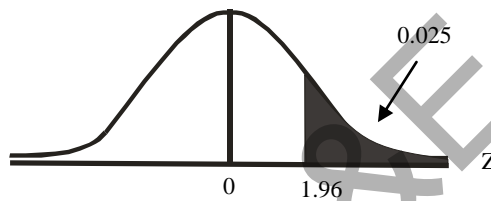
(iii) **P** (-1.96 ≤ Z ≤ +1.96) =1 - [**P**(Z ≤ -1.96) + **P**(Z ≥ 1.96)] = 1- [0.0250 + 0.0250] = 0.95

or this may be calculated as

= **P**(Z ≤ 1.96) - **P** (Z ≤ -1.96) = 0.9750 - 0.0250 = 0.95

Fig. 2.4
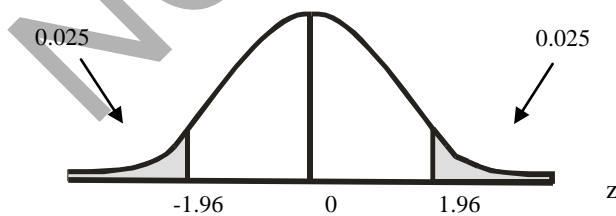
Therefore 95 % of the probability of the normal distribution is between -1.96 to 1.96, 5 % is lying beyond these limits. In other words if p = 0.0250, then either z is greater than or equal to 1.96 or less than or equal to -1.96. This probability is usually referred to as two-tailed probability.

(b) (i) $\mathbf{P}(Z \le -2.58) = .0049 \approx 0.005,$
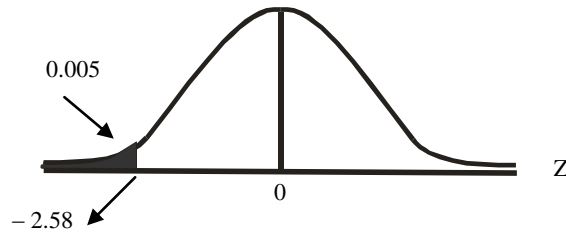
Fig. 2.5

(ii) $\mathbf{P}(Z \ge 2.58) = \mathbf{P} (Z \le -2.58)$ or $1 - \mathbf{P} (Z \le 2.58) = 1 - 0.9951 = 0.049 = .5049 \approx$ 0.005.
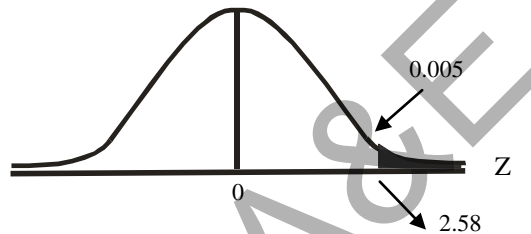
Fig. 2.6

(iii) $\mathbf{P} (-2.58 \le Z \le 2.58) = \mathbf{P} (Z \le 2.58) - \mathbf{P}(Z \le -2.58) = 0.9951 - 0.0049$ $= 0.9902 \approx 99\%.$

Therefore 99% of the probability of the normal distribution is between - 2.58 and 2.58 and only 1% probability is beyond these two points.
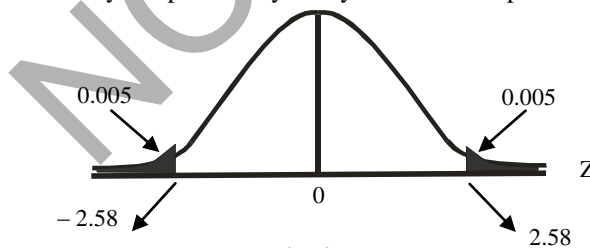
Fig. 2.7

(c) (i) $\mathbf{P} (Z \le -2.33) = 0.0099 \approx 1\%$
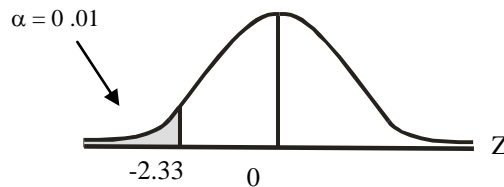
Fig. 2.8

(ii) $\mathbf{P}(Z \le -1.65) = 0.0495$

0.0495

-1.65   0

Z
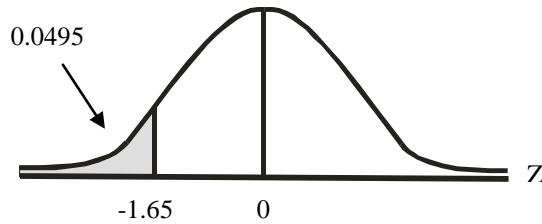
Fig. 2.9

From the above calculations, it is clear as the probability {p-value} decreases z-value increases and vice versa.

Example 2.11:

Medical research has linked excessive consumption of salt to hypertension. The average amount of salt consumed per day by an American is 15 gram, although the actual physiological minimum daily requirement for salt is only 220 milligrams. Suppose that the amount of salt per day is approximately normally distributed with a standard deviation of 5 grams. What proportion of all Americans consume between 14 and 22 grams of salt per day?

Solution:

The proportion of Americans who consume between x = 14 and x = 22 grams salt is shown in the shaded area of the graph 2.10.

p

14 15   22

x
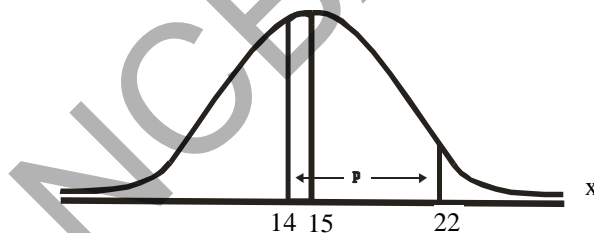
Fig. 2.10

Since mean = 15 grams with standard deviation = 5. This does not follow standard normal distribution. In order to find the proportion, it is to be converted into the standard normal distribution by using standardized normal variable (z - variate)

$\mu = 15, \sigma = 5,$

$$z_1 = \frac{14 - 15}{5} = -0.20 \qquad z_2 = \frac{22 - 15}{5} = 1.40$$

This can be shown by the diagram

Fig. 2.11

$P(-0.20 \leq Z \leq 1.40) = P(Z \leq 1.40) - P(Z \leq -0.20)$

$= 0.9192 - 0.4207 = 0.4985$

49.85% ≈ 50% of the Americans consume between 14 to 22 grams of salt per day.

Table 2.4
Cumulative Probabilities of Binomial Distribution for n=5 and n=10

(i) n=5

x↓ p→	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20
0	.5584	.5277	.4984	.4704	.4437	.4182	.3939	.3707	.3487	.3277
1	.9035	.8875	.8704	.8533	.8352	.8165	.7973	.7776	.7576	.7373
2	.9888	.9875	.9821	.9780	.9734	.9682	.9625	.9563	.9495	.9421
3	.9993	.9991	.9987	.9983	.9978	.9971	.9964	.9955	.9945	.9933
4	1.000	1.000	1.000	.9999	.9999	.9999	.9999	.9998	.9998	.9997
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

x↓ p→	.31	.32	.33	.34	.35	.36	.37	.38	.39	.40
0	.1564	.1454	.1350	.1252	.1160	.1074	.0092	.0916	.0845	.0778
1	.5077	.4875	.4675	.4478	.4284	.4094	.3907	.3724	.3545	.3370
2	8234	.8095	.7950	.7801	.7648	.7491	.7330	.7165	.6997	.6826
3	.9653	.9610	.9564	.9514	.9460	.9402	.9340	.9274	.9204	.9130
4	.9971	.9966	.9961	.9955	.9947	.9940	.9931	.9921	.9910	.9898
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

(ii) n=10

x↓ p→	.22	.23	.24	.25	.26	.27	.28
0	.0834	.0733	.0643	.0563	.0492	.0430	.0374
1	.3185	.2921	.2673	.2440	.2222	.2019	.1830
2	.6169	.5863	.5558	.5256	.4958	.4665	.4378
3	.8413	.8206	.7988	.7759	.7521	.7274	.7021
4	.9521	.9431	.9330	.9219	.9096	.8963	.8819
5	.9896	.9870	.9839	.9803	.9761	.9713	.9658
6	.9984	.9979	.9973	.9965	.9955	.9944	.9930
7	.9998	.9998	.9997	.9996	.9994	.9993	.9990
8	1.0000	1.0000	1.0000	1.0000	1.0000	.9999	.9999
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2.5
Cumulative probability for Poisson distribution

x↓ μ→	3.6	3.8	4.0	4.2	4.4	4.6	4.8	5.0
0	.027	.022	.018	.015	.012	.010	.008	.007
1	.126	.107	.092	.078	.066	.056	.048	.040
2	.303	.269	.238	.210	.185	.163	.143	.125
3	.515	.473	.433	.395	.359	.326	.294	.265
4	.706	.668	.629	.590	.551	.513	.476	.440
5	.844	.816	.785	.753	.720	.686	.651	.616
6	.927	.909	.889	.867	.844	.818	.791	.762
7	.969	.960	.949	.963	.921	.905	.887	.867
8	.988	.984	.979	.972	.964	.955	.944	.932
9	.996	.994	.992	.989	.985	.980	.975	.968
10	.999	.998	.997	.996	.994	.992	.990	.986
11	1.000	.999	.999	.999	.998	.997	.996	.995
12	-	1.000	1.000	1.000	.999	.999	.999	.998
13	-	-	-	-	1.000	1.000	1.000	.999
14	-	-	-	-	-	-	-	1.000

x↓ μ→	5.2	5.4	5.6	5.8	6.0	6.2	6.4	6.6
0	.006	.005	.024	.003	.002	.002	.002	.001
1	.034	.029	.024	.021	.017	.015	.012	.010
2	.109	.095	.082	.072	.062	.054	.046	.040
3	.238	.213	.191	.170	.151	.134	.119	.105
4	.406	.373	.342	.331	.285	.259	.235	.213
5	.581	.546	.512	.478	.446	.414	.384	.355
6	.732	.702	.670	.638	.606	.574	.542	.511
7	.845	.822	.797	.771	.744	.716	.687	.658
8	.918	.903	.886	.867	.847	.826	.803	.780
9	.960	.951	.941	.929	.916	.902	.886	.869
10	.982	.977	.972	.965	.975	.949	.939	.927
11	.993	.990	.988	.984	.980	.975	.969	.963
12	.997	.996	.995	.993	.991	.989	.986	.982
13	.999	.999	.998	.997	.996	.995	.994	.992
14	1.000	.999	.999	.999	.999	.998	.997	.997
15	-	1.000	1.000	1.000	.999	.999	.999	.999
16	-	-	-	-	1.000	1.000	1.000	.999
17	-	-	-	-	-	-	-	1.000

Table 2.6
Probabilities of the Normal Distribution
(Areas between - ∞ and z)

z	-.09	-.08	-.07	-.06	-.05	-.04	-.03	-.02	-.01	-.00
-3.80	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.70	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.60	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
-3.50	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.40	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.30	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005
-3.20	.0005	.0005	.0005	0006	.0006	.0006	.0006	.0006	.0007	.0007
-3.10	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010
-3.00	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013
-2.90	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019
-2.80	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026
-2.70	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035
-2.60	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047
-2.50	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062
-2.40	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082
-2.30	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107
-2.20	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139
-2.10	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179
-2.00	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228
-1.90	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287
-1.80	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359
-1.70	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446
-1.60	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548
-1.50	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668
-1.40	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808
-1.30	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968
-1.20	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151
-1.10	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357
-1.00	.1379	.1401	.1423	.1446	.1469	.1492	.1515	.1539	.1562	.1587
-0.90	.1611	.1635	.1660	.1685	.1711	.1736	.1762	.1788	.1814	.1841
-0.80	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119
-0.70	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420
-0.60	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743
-0.50	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085
-0.40	.3121	.3156	.3192	.3228	.3264	.3300	.3336	.3372	.3409	.3446
-0.30	.3483	.3520	.3557	.3594	.3632	.3669	.3707	.3745	.3783	.3821
-0.20	.3859	.3897	.3936	.3974	.4013	.4952	.4090	.4129	.4168	.4207
-0.10	.4247	.4286	.4325	.4364	.4404	.4443	.4483	.4522	.4562	.4602
0.00	.4641	.4681	.4721	.4761	.4801	.4840	.4880	.4920	.4960	.5000

Table 2.6 (contd.)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.10	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.20	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.30	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.5617
0.40	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.50	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.60	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.70	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.80	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.90	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.00	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.10	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.20	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.30	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.40	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.50	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.60	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.70	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.80	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.90	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.00	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.10	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.20	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.30	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.40	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.50	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.60	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.70	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.80	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.90	.9981	.9982	.9982	.9983	.9984	.9985	.9985	.9985	.9986	.9986
3.00	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.10	.9990	.9991	.9991	.9992	.9992	.9992	.9992	.9992	.9993	.9993
3.20	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.30	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.40	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.50	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.60	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.70	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.80	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Chapter 3

Sampling Procedures and Sample Size Estimation

3.1 Introduction

Most survey work involves sampling from finite populations. There are two parts to any sampling strategy (design). First, there is a *selection procedure*, the manner in which sampling units are selected from a population. Second, there is an *estimation procedure* that prescribes how inferences are to be drawn from sample to the population

Sampling is procedure or process of selecting some units from the population with some common characteristics and is primarily concerned with the collection of data of some selected units of the population. *Census* is another method of data collection and is defined as a complete enumeration of the population. A list of population units from which the sample is selected is called a *sampling frame*.

Since sample is a part of population, the result based on the sampled observations will not be equal to that of population values. There must be some difference, which is inevitable. This difference is known as *error*. This error is arising due to drawing inferences about the population on the basis of sampled observations, therefore, it is termed as *sampling error*, e.g. the prevalence of tuberculosis based on a sample cannot be identical to its prevalence in the population. The sampling error usually decreases as the sample size increases. In many situations, the decrease is inversely proportional to the sample size, in fact, to the square root of the sample size. The sampling error is reduced to minimum if the choice of the sampling unit, sampling design, selection procedure, sample size and method of data analysis are appropriate. Note that in the reduction of sampling error, sample size plays an important role.

Error arising from the causes not associated with the sampling process is known as *non-sampling error*, which is common, both to complete enumeration and sample surveys. Non-sampling error includes (i) response error (ii) non-response error (iii) measurement and coding error, (iv) improper method for statistical analysis (v) non- coverage of population, (vi) interviewers error, (vii) data entry error etc. As the sample size increases, non-sampling error increases. Generally if the sample is proper representative of a population, sampling error is minimum. A representative sample must possess all the important characteristics of the population under study. If one is to investigate malnutrition in children under five, then our population will be all children from 0 to 4 years of age.

A question naturally arises why sampling? The answer is as follows:

There are some advantages to select a sample from a population. These are:

    (i)   A sample is a part of population; the information can be collected more *cheaply* and more *rapidly* as compared to complete enumeration.

(ii)   A sample makes it possible to concentrate on individual units and to obtain relevant information *comprehensively and accurately*.

(iii) Selection of appropriate sampling design reduces non-sampling error.

(iv)   More precise results can be obtained by survey and sampling experts.

3.2 Types of Sampling

There are, generally, two types of sampling, i.e. (i) *probability sampling* and (ii) *non-probability sampling*.

3.2.1 Probability Sampling

A probability sample or *a random sample* is one in which the probability of selection of each unit in the population is known. The probability of selection of each unit may or may not be independent. If a sample is selected at random then it is known as a probability sample. In fact probability sampling is a general name given to the sampling plan in which

(a) every individual in the sampled population has a known probability of entering in the sample, (b) the sample is chosen by a process involving one or more steps of automatic randomization, (c) in the analysis of the samples, weights (probabilities) appropriate to the probabilities given in (a) above are used.

3.2.2 Non-Probability Sampling

A sample selected by a non-random process is termed as a non-probability sample. Judgment samples, purposive samples and quota samples are examples of non-probability samples. These types of selection procedures are useful when the population units are highly variable and the sample is small. In these selection procedures, there is no way to check the *precision* and to obtain the precise *estimates*. There is no way to determine the *sampling, non-sampling errors.*

3.3 Some Commonly Used Selection Procedures

In this section some commonly used selection procedures of probability sampling and estimation of mean, variance, confidence intervals are described.

3.3.1 Simple Random Sampling

Random sampling or more precisely *simple random sampling* is a term covering two of the most straightforward selection procedures used in the probability sampling. In both these procedures population units are drawn (selected) one by one with equal probability until the sample is achieved of the required size. If unit once selected is not allowed to be selected again, the procedure is known as *simple random sampling without replacement (srswor)*. If the selection at each draw is from the whole population, the procedure is known as *simple random sampling with replacement (srswr)*. Selection of units using *srswr* is independent from draw to draw, but if *srswor* is used the selections are not independent. This is because in *srswor* the probability of selection of a population unit at any given draw depends on whether or not it has been selected at some previous draws. It

is generally assumed that the characteristic for which the sample is selected does not change during sampling operation and selection must be independent of the characteristic under investigation. This selection procedure is explained in the following example:

Example 3.1:

Suppose there are 500 households in a certain area and we are interested in holding a tuberculosis (TB) survey to check the prevalence of TB in that area. First, we get a map of that area. We will allot our own numbers starting from 001 to 500. Suppose we want to select 5 percent sample from this population, which comes out to be 25 households. Then select any three columns from the random number tables (table 3.10) as population is of three digits.

Include all those numbers, which are between 001 and 500 both ends inclusive and reject all others. If any number previously selected is repeated ignore it. As an example a sample of 25 houses has been drawn using the random digits. These random digits are given in Table 3.17. Note, if any number is repeated ignore it.

In cases where respondents do not cooperate or household is closed, we need to have some randomly selected reserve sample so that it can be utilized if any non-response occurs. It has been observed that 5 to 10 percent is the non-response rate, so it is advisable while selecting a sample, to select a reserve sample at that time. If, for example, the 9th house (house number 466) in our actual sample is not co-operating or is closed then it can be substituted by the 26th house (house number 270), which is the first house in our reserved sample and so on. In any case the interviewer has no personal choice to select the house.

Table 3.1
Selected actual and reserved samples

Sr. No.	Random number/House number in our list	Sr. No.	Random number/House number in our list
Actual Sample			
1	427	16	218
2	275	17	014
3	356	18	146
4	463	19	292
5	112	20	174
6	497	21	405
7	054	22	094
8	163	23	158
9	308	24	103
10	062	25	122
11	466	26	270
12	143	27	104
13	465	28	120
14	078	29	030
15	467	30	476

If we like to investigate the quality of the X-ray films in a certain laboratory, then all X-ray films will be our study population. Each x-ray film must have ID number and required sample will be selected accordingly.

Simple random sampling selection procedure is very simple and easily understandable as each unit of population has an equal chance to be in the sample and also each selected sample has an equal probability. This design is ineffective if the population units are highly variable.

Many samples can be selected but in practical life, only one sample is selected and it is assumed that this sample will be the representative sample of the population under study. The sample mean or sample proportion is assumed to be the estimated value of population mean or population proportion.

3.3.2 Estimation of mean and variance for sample mean and sample proportion

An unbiased estimator of population mean is

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{3.1}$$

A sample will yield unbiased estimate under the following conditions.

(i)   All the units of the population to be sampled are listed . Failure to do so causes bias, known as *coverage bias*.

(ii)   Each unit of the population to be sampled must have a known probability, other than zero. Failure to do so causes bias known as *sample selection bias*.

(iii) The measurements, observations or responses must be obtained from each sample unit. Failure to do so causes bias known as *non-response bias.*

(iv)   Actual values of measurements, or observations or responses are obtained. Failure to do so causes bias known as *response bias*.

(v)   Appropriate sample design must be used. Failure to do so causes bias known as *sample design bias*.

(vi)   Appropriate method of estimation is to be used. Failure to do so causes bias known as *sample estimation bias*.

(vii) One should not collect information from the next door if sampled unit is not available. Failure to do so causes bias known as *substitution bias*.

(viii) Finally, all the arithmetic, clerical and other operations entailed in sample selection and estimation must be performed properly. Failure to do so causes bias known as *operational bias*.

The variance expressions of sample mean for without and with replacement sampling are respectively given as:

$$\text{Var}(\bar{y}_{wor}) = \frac{N-n}{N} \frac{S^2}{n} = (1-f) \frac{S^2}{n}, \tag{3.2}$$

f = n/N and

$$Var\left(\overline{y}_{wr}\right) = \frac{N-1}{N}\frac{S^2}{n} = \left(1 - \frac{1}{N}\right)\frac{S^2}{n} \qquad (3.3)$$

where $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$ \qquad (3.4)

For large N correction factor is ignored and we get the same expression for sampling with and without replacement i.e.

$$Var(\overline{y}) = S^2/n \qquad (3.5)$$

An unbiased variance estimator for without replacement and with replacement sampling are given respectively as:

$$var(\overline{y}_{wor}) = (1-f)\frac{s^2}{n}. \qquad (3.6)$$

and

$$var(\overline{y}_{wr}) = \frac{s^2}{n}. \qquad (3.7)$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$ \qquad (3.8)

For large N the unbiased variance expression for with and without replacement is:

$$var(\overline{y}) = s^2/n \qquad (3.9)$$

In case of qualitative data such as smoker and non-smoker, educated and non-educated etc. the proportion (p) of smokers, educated etc. is calculated. If p is an unbiased proportion of population proportion P, then the variance expressions of sample proportion for with and without replacement sampling are respectively given as:

$$Var(p_{wor}) = \frac{N-n}{N-1}\frac{PQ}{n}, \ \ (srswor) \qquad (3.10)$$

and

$$Var(p_{wr}) = PQ/n \ \ (srswr) \qquad (3.11)$$

For large N, N-1 approaches to N, and if fpc is ignored than we get the same expression for with and without replacement given as:

$$Var(p_{wr}) = PQ/n \qquad (3.12)$$

An unbiased variance expression for with and without replacement is:

$$var(p_{wor}) = \frac{N-n}{N}\frac{pq}{n-1}, \qquad (3.13)$$

and

$$var(p_{wr}) = \frac{N-1}{N}\frac{pq}{n-1} \qquad (3.14)$$

Ignoring fpc we get:

$$\text{var}(p) = pq/(n-1) \tag{3.15}$$

For large n, if n/N is small, we get:

$$\text{var}(p) = pq/n \tag{3.16}$$

3.3.3 Estimation of Sample Size

The most important question for statisticians and non-statisticians is how large a sample should be? In a survey sampling, it is always a problem for an experimenter to know or to determine the size of the sample when the result is required with *least sampling error*. Should a sample be 2% or 5% or 10% or any other fraction? Although the sample size is a matter of choice with the planner, yet great care and weight is needed in its determination. Since sample is a proportion of the population, it should neither be too large to involve a lot of expenditure and non-sampling error nor too small to make the results less reliable. In fact the sample size depends on the cost involved and time and precision required. Optimal sample size minimizes sampling error. Although sampling error is decreased by the increase of sample size but without optimal sample size there is a danger of large non-sampling error.

The following formula may be used for different situations.

(a) Sampling for Proportions

(i) Sample size for absolute precision

$$n_0 = \frac{Z_{(1-\alpha/2)}^2\ p(1-p)}{d^2}, \tag{3.17}$$

where d is the difference between estimated and actual value. i.e. absolute precision required on either side of the proportion p. It is usually taken as 5%. If sample size is large, then for 95% probability level or confidence level $Z_{1-\alpha/2}$ is taken as 1.96, for 99% level, 2.58, and for 90% the confidence level is 1.645. For convenience, sample size has been calculated for different values of p and d, [Tables 3.12 and 3.13 are given at the end of this chapter].

Example 3.2:
The Ministry of Health wishes to estimate the prevalence of tuberculosis among children under 5 years of age. How many children should be there in the sample so that the prevalence may be estimated within 5% points of the true value with 95% or 99% confidence level, if it is known that the true rate will not exceed 15%.

Solution:
In This exemple we have

$$p = 0.15,\ 1-p = 0.85$$

Probability level or confidence level $(1 - \alpha) = 95\%$ or 99%.

$$d = 5 \text{ percentage points}$$

$$Z_{1-\alpha/2} = 1.96 \text{ for } \alpha = 0.05 \text{ and } Z_{1-\alpha/2} = 2.58 \text{ for } \alpha = 0.01$$

Using the formula, we have

$$n_0 = \frac{(1.96)^2 \ (0.15) \ (0.85)}{(0.05)^2} = 196 \text{ for } 1-\alpha = 95\%$$

and

$$n_0 = \frac{(2.58)^2 \ (0.15) \ (0.85)}{(0.05)^2} = 339 \text{ for } 1-\alpha = 99\%$$

If population is finite then an approximation of sample size can be obtained as $n_1 = \dfrac{n_0}{1 + (n_0 - 1)/N}$. If the population of children less than 5 years of age is 20,000, then the sample size may be estimated as, by an approximation,

$$n_1 = \frac{196}{1 + (196 - 1)/20000} = \frac{196}{1.00975} = 194$$

This is not different from 196, so 196 or 194 may be taken as a sample size.

Example 3.3:
Ministry of Health would like to estimate the proportion of children who are receiving medical care regularly. How large should be the sample if the estimate falls within 5% of true proportion with 95% confidence level.

Solution:
In this question, the assumption regarding proportion of children who are receiving regularly medical care is that 50% of the population of children is receiving medical care. Using p = 0.50, maximum sample size will be obtained.

If we take

$$p = 0.5; \ 1 - \alpha = 0.95, 0.99; \ d = 0.05$$

then

$$n_0 = \frac{(1.96)^2 \ (0.5) \ (0.5)}{(0.05)^2} = 384 \text{ for } 95\%$$

$$n_0 = \frac{(2.58)^2 \ (0.5) \ (0.5)}{(0.05)^2} = 666 \text{ for } 99\%$$

Suppose N = 600 then, then the sample size for 95% level comes to be:

$$n_1 = \frac{384}{1 + (384 - 1)/600} = \frac{384}{1.638} = 234 \text{ (2nd approx.)}$$

$$n_2 = \frac{234}{1 + (234 - 1)/600} = \frac{234}{1.388} = 169 \text{ (3rd approx.)}$$

$$n_3 = \frac{169}{1 + (169 - 1)/600} = \frac{169}{1.280} = 132 \text{ (4th approx.)}$$

This process will continue till difference between the last two approximations becomes minimal.

(ii) Sample size for relative precision

If the coefficient of variation (or for relative precision) is given, the formula for the determination of sample size is

$$n = \frac{z_{1-\alpha/2}^2 \ (1-p)}{D^2 \, p} \, , \tag{3.18}$$

where D denotes coefficient of variation or relative precision.

For convenience, sample sizes have been calculated for different values of p and D. [see Tables 3.14 and 3.15]

Example 3.4:
Ministry of Health of Eastern Province would like to conduct a survey regarding hypertension of elderly persons (above the age of 60). It is known from the past experience that the prevalence of hypertension is 25%. How large a sample should be so that the resulting estimates falls within 10% (not 10% points) of the true proportion with 95% confidence level?

Solution:
In this question p = 0.25, Confidence level = 95% and relative precision is 10% of 25%. There are two ways to solve this problem.

(i) Using relative precision formula

$$n = \frac{(1.96)^2 \ (0.75)}{(0.05)^2 \ (0.25)} = 4610$$

(ii) Using absolute precision formula

Since d = 0.05 x 0.25 = 0.0125

$$n = \frac{(1.96)^2 (0.25) \ (0.75)}{(0.0125)^2} = 4610$$

If population size is known to be 2000, then

$$n_1 = \frac{4610}{1 + (4610 - 1)/2000} = \frac{4610}{3.3045} = 1395$$

$$n_2 = \frac{1395}{1+(1395-1)/2000} = \frac{1395}{1.697} = 822$$

$$n_3 = \frac{822}{1+(822-1)/2000} = \frac{822}{1.4105} = 583$$

This process will continue till there is not much difference between the last two approximations. We see that after 10th approximation, we get the sample size of 212.

If p = 25% to 40% and relative precision D = 0.05 then for different values of p and with 95% confidence level, the sample size will be:

Table 3.2

p	0.25	0.30	0.35	0.40
n	4610	3585	2854	2305

The relative precision (D) may be converted into absolute precision (d) as

$$d = p \times D = \begin{bmatrix} .25 \times .05 = 0.0125 \\ .30 \times .05 = 0.0150 \\ .35 \times .05 = 0.0175 \\ .40 \times 0.5 = 0.0200 \end{bmatrix}$$

The sample sizes for different values of d and p and for 95% confidence level are given as:

Table 3.3
Sample sizes for different values of p and d

p↓ d→	0.0125	0.0150	0.0175	0.0200
0.25	**4610**	3201	2352	1801
0.30	5163	**3585**	2634	2017
0.35	5593	3884	**2854**	2184
0.40	5901	4098	3010	**2305**

If the range is given, i.e. the prevalence is 10 to 25%, then it is always advisable to use prevalence 25% for precision. If the range is 45% to 55% then for precision use p = 50% but for relative precision use 55%.

(b) *Sampling with Continuous Data (absolute precision)*

If mean and sample variance is known then the formula for determination of sample size

$$n_0 = \frac{z_{1-\alpha/2}^2 \, s^2}{d^2} \quad \text{(Ist approx.)} \tag{3.19}$$

If the population size is known then

$$n_1 = \frac{n_0}{1+\dfrac{n_0}{N}} \qquad \text{(2nd approx.)}$$

or

$$n_2 = \frac{n_1}{1+\dfrac{n_1}{N}} \qquad \text{(3rd approx.)}$$

and so on.

Example 3.5:

A physician would like to know the mean fasting blood glucose of patients seen in the diabetes clinic over the past 10 years. Determine the number of records the physician should examine in order to obtain 90% and 95% confidence level for population if the desired width of the interval is 8 units and pilot sample yields a standard deviation of 60 units.

Solution:

Here s = 60, D = 4, as the total width is 6 which is on the both sides of the mean. Therefore, the sample size for 90% confidence will be

$$n = \frac{(1.645)^2 (60)^2}{(4)^2} = 609 \text{ for } 90\%$$

and for 95 %

$$n = \frac{(1.96)^2 (60)^2}{(4)^2} = 864 \text{ for } 95\%$$

3.3.4 Standard Deviation and Standard Error

When numerical findings are reported in research articles or medical dissertation, regardless of whether or not their statistical significance is quoted, they are often presented with additional statistical information. The distinction between standard deviation and the standard error is often misunderstood. By contrast, the standard error is a *measure of the uncertainty in a sample statistic*.

The standard deviation is relevant when variability between individuals is of interest whereas the standard error is relevant to summary statistics such as mean, proportions, differences between means and proportions, etc.

The standard error of the sample statistic, which depends on both the standard deviation and the sample size, is recognition that a sample is most *unlikely* to determine the population value exactly. In fact, if a further sample is taken in identical circumstances, it will almost certainly produce different estimates of the same population. The sample statistic is therefore imprecise and the standard error is a measure of this imprecision.

The standard error of sampling mean is given as:

$$\text{SE}(\overline{y}) = \sqrt{\text{var}(\overline{y})} \qquad\qquad\qquad (3.20)$$

3.3.5 Confidence Limits

It is not possible for a sample to evaluate characteristics of a population exactly, but it estimates the characteristics as accurately as possible. One way out may be to find intervals which are functions of observations and which cover the parameter with pre-assigned probabilities. In case the variable is normally distributed with known variance, the sampling distribution of means is also normally distributed. The interval $\bar{x} \pm 1.96$ SE ($\bar{x}$) will cover sample means in 95% of the cases.

The confidence intervals are calculated whenever an inference is to be made from the sample to the population from which the sample has been drawn. The calculated interval provides a range of values within which lies the population value. Confidence limits are calculated with $(1 - \alpha)$% confidence coefficient. The width of the confidence coefficient intervals depends on three factors. Firstly the size of sample (large sample sizes give narrower confidence intervals), secondly the standard deviation of the characteristic being studied (smaller the standard deviation, narrower the confidence interval) and finally the degree of confidence is required.

The confidence limits for sample mean are:

$$\text{mean} \pm Z_{1-\alpha/2} \text{ S.E (mean)} \tag{3.21}$$

For 95% reliability the confidence limits will be:

$$\bar{y} - 1.96 \text{ S.E}(\bar{y}) \text{ and } \bar{y} + 1.96 \text{ S.E}(\bar{y}) \tag{3.22}$$

For sample proportion the confidence limits will be

$$p \pm Z_{1-\alpha/2} \sqrt{\frac{P(1-P)}{n}} \tag{3.23}$$

(i) Confidence limits for large sample
This is explained with the following example.

Example 3.6:
The serum cholesterol level of healthy persons is given. Select a sample of 30 persons from the population of 90 persons and estimate the average cholesterol level of persons in the population. Construct 95% confidence limits for the mean of the population (the data is given on in Table 3.4)

Solution:
We have 90 persons in the population, or we say N = 90. The purpose is to select a random sample of 30 persons from the given population of 90 persons. One should remember that random number table or a computer psuedo random numbers are used to select 30 persons out of 90.The mean cholesterol level of persons in the population is 19316/90 = 21462. One should remember that population mean or proportion is never known before it is always to be estimated on the basis of sample. In this example actual population values are given and sample has been selected so that comparison could be made. We can calculate the mean of the selected sample and can compare it with population mean.

Table 3.4
Serum cholesterol level of 90 healthy persons

Persons	Cholesterol level	Persons	Cholesterol level	Persons	Cholesterol level
1	154	31	172	61	235
2	212	32	219	62	253
3	222	33	247	63	263
4	259	34	186	64	266
5	239	35	257	65	200
6	201	36	222	66	200
7	204	37	208	67	223
8	208	38	170	68	155
9	197	39	202	69	201
10	205	40	222	70	234
11	196	41	236	71	263
12	212	42	248	72	233
13	218	43	186	73	223
14	196	44	259	74	198
15	169	45	218	75	177
16	179	46	208	76	197
17	210	47	226	77	221
18	204	48	160	78	220
19	212	49	171	79	231
20	191	50	238	80	222
21	239	51	175	81	200
22	251	52	208	82	225
23	160	53	239	83	279
24	211	54	255	84	283
25	188	55	221	85	258
26	236	56	160	86	253
27	248	57	224	87	234
28	189	58	156	88	276
29	174	59	230	89	265
30	138	60	262	90	221
total					19316

A random sample of 30 using the random digits given at the end of the chapter has been selected and the values of the sample are given in Table 3.5.

Table 3.5
Selected sample of 30 persons

Sr. No.	Random number	Cholesterol level x	x^2
1	88	276	76176
2	25	188	35344
3	56	160	25600
4	07	204	41616
5	31	172	29584
6	47	226	51076
7	73	223	49729
8	16	179	32041
9	89	265	70225
10	03	222	49284
11	72	233	54289
12	74	198	39204
13	43	186	34596
14	17	210	44100
15	83	279	77841
16	62	253	64009
17	37	208	43264
18	65	204	41616
19	79	231	53361
20	06	201	40401
21	33	247	61009
22	32	219	47961
23	12	212	44944
24	02	212	44944
25	45	218	47524
26	13	218	47524
27	66	200	40000
28	23	160	25600
29	20	191	36481
30	35	257	66049
Total		6452	1415392

The sample and population means are $6452/30 = 215.07$ and $19316/90 = 214.62$, respectively. We see that one random sample has been selected and mean cholesterol level on the basis of the sample is 215.07 whereas mean cholesterol level of the population is 214.62. The difference between sample and population mean is not much. As mentioned before, in practical life, we never know population mean and proportion this is assumed to be an estimate of the population mean. The sample mean, 215.07 is an estimated value of population mean, 214.62. To locate the position of population mean, we construct 95% or 99% confidence limits, then we say with confidence that is, we are 95% or 99% confident that these two limits contain population mean. For this purpose we calculate first sample standard deviation and then standard error. The sample standard deviation is: (using Equation 3.8)

$$s = \sqrt{\frac{1}{29}\left[1415392 - \frac{(6452)^2}{30}\right]} = 30.952$$

If a sample is large, we divide by 30 = n or 29 = (n - 1) which does not make much difference but remember if the sample size is less than 30, it is essential that the divisor for standard deviation is n-1.

The standard error of sample mean is:

$$\text{S.E} \left(\bar{x} = \text{mean}\right) = \frac{s}{\sqrt{n}} = \frac{30.95}{\sqrt{30}} = \frac{30.95}{5.477} = 5.650$$

The confidence limits of μ is

mean $\pm Z_{1-\alpha/2}$ S.E (mean)   (3.21)

Were $Z_{1-\alpha/2}$ is taken 1.645 for 90%, 1.96 for 95% and 2.58 for 99% confidence level [Table 3.18]. The 90%, 95% and 99% confidence limits respectively are:

215.07 ± 1.645 x 5.650 = (205.80 - 224.36) is a 90% confidence limits
215.07 ± 1.96 x 5.650 = (203.996 - 226.074) is a 95% confidence limits
215.07 ± 2.58 x 5.650 = (200.493 ~ 229.647) is a 99% confidence limits

In this example, we state that population mean is 214.62. Therefore, we say with 90% or 95% or 99% confidence that these limits contain population mean. If the population mean is not known, even then we say with confidence that above statement is true.

Example S3-1 (Selecting a Simple Random Sample using IBM-SPSS)

To select a random sample of size 30 from the data in table 3.4, using IBM-SPSS, we follow the following steps: **Data➔ Select Cases:**

(we can either chose **Filter out unselected cases**, or **Copy selected cases to a new dataset**, or **Delete unselected cases**):
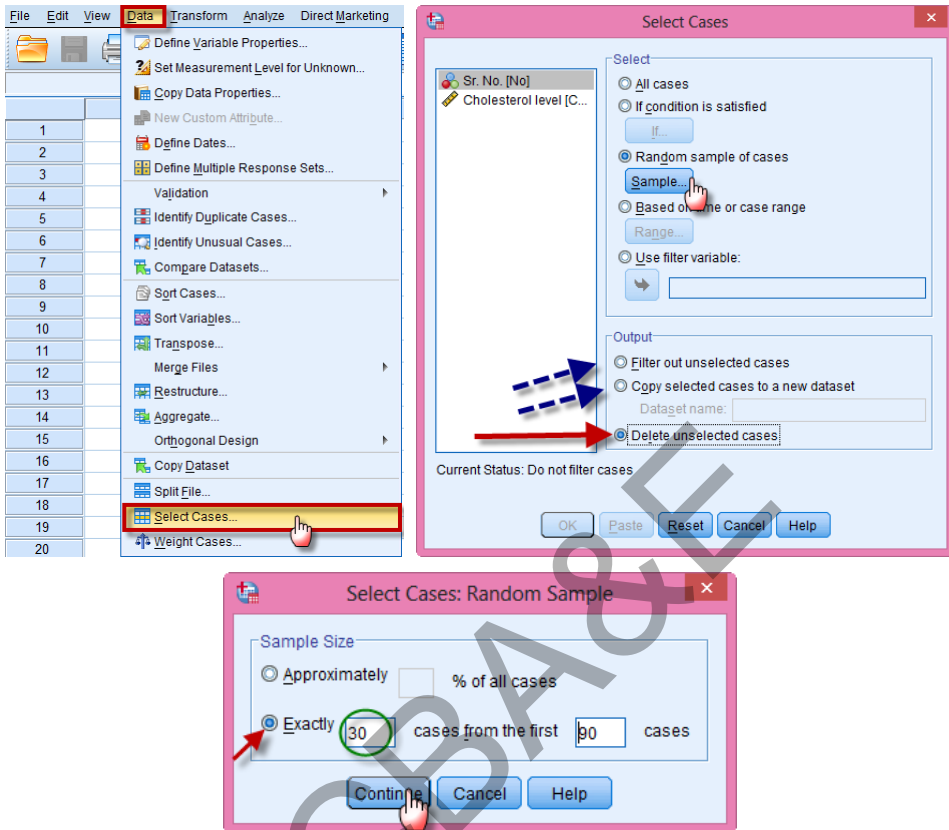
Then, click on OK to get directly the desired random sample

Example S3-2

To obtain the Confidence limits for the mean using IBM-SPSS, for the data in table 3.5, we enter the data and follow the following steps:

Analyze→ Descriptive Statistics → Explore:
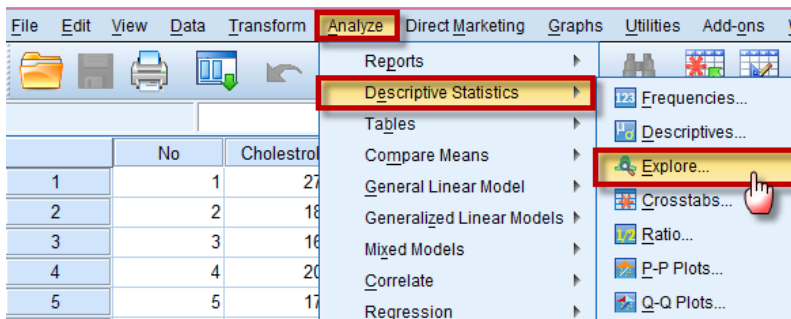
We move the variable into Dependent List and do as follows:
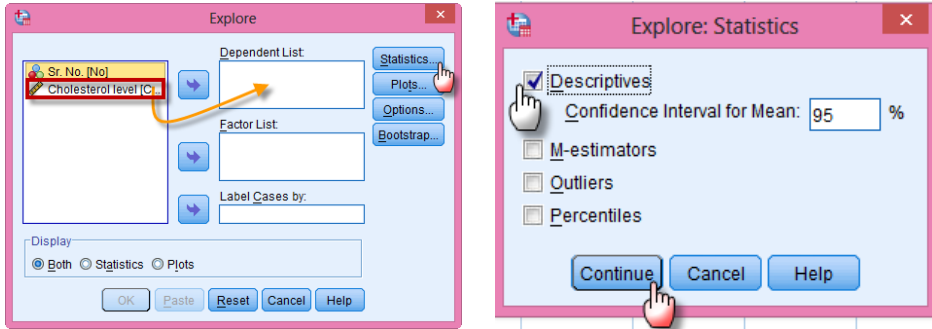
(Note that we can change the 95% to any other value, e.g. 90% or 99%).

Once we click on OK , we get the following output for the 95%:

			Statistic	Std. Error
Cholesterol level	Mean		215.07	5.651
	95% Confidence Interval for Mean	Lower Bound	203.51	
		Upper Bound	226.62	
	5% Trimmed Mean		214.63	
	Median		212.00	
	Variance		957.995	
	Std. Deviation		30.952	
	Minimum		160	
	Maximum		279	
	Range		119	
	Interquartile Range		35	
	Skewness		.313	.427
	Kurtosis		-.175	.833

(ii) Confidence limits for small sample
The values of 90% or 95% or 99% are only used if the sample size is large.

Example 3.7:
A sample of size 10 is drawn from the population given in example 3.4 is given on next page table 3.6:

The sample mean = 2221/10 = 222.1, and the sample standard deviation using (3.8) is

$$s = \sqrt{\frac{1}{10-1}\left[506321 - \frac{(2221)^2}{10}\right]} = 38.060.$$

(The divisor is (10 - 1) and not 10.)

The confidence limits are $222.1 \pm t_{1-\alpha/2} \dfrac{38.060}{\sqrt{10}} = 222.1 \pm 2.262 \times \dfrac{38.060}{3.162}$ or [194.875,

249.325]. 2.262 is value from the t-table [Table 3.17].

Table 3.6
Selected sample of 10 persons

Random members	Value of cholesterol level X	X^2
82	225	50625
44	259	67081
53	239	57121
60	262	68644
06	201	40401
07	204	41616
30	138	19044
65	200	40000
61	235	55225
85	258	66564
Total	2221	506321

The question is how to see the table. If it is 95% confidence limit then subtract 0.95 from 1, i.e. 1 - 0.95 = 0.05, divide 0.05 by 2, i.e. = 0.025, subtract 0.025 from 1 we will get 0.975, consult the t-table under 0.975 and against 9 = (n - 1). This value is used at the place of $t_{1-\alpha/2}$. (n - 1) is called the degree of freedom and 0.05 (5%) is called level of significance. This will be explained in the next Chapter.

Example 3.8*:*
A sample of 25 physically active adult males was selected and arterial blood gas analysis was performed. The results are given in terms of P_aQ_2 values i.e. 75, 88, 75, 88, 72, 83, 83, 72, 87, 78, 78, 77, 79, 80, 80, 83, 79, 79, 72, 83, 76, 85, 86, 84, 75. Compute 95% confidence limits for the mean.

***Solution*:**
Mean = 79.88 and sample standard deviation = 4.969. The 95% confidence limits will be

$$\text{mean} \pm t_{1-\alpha/2} \frac{\text{sample s.d (s)}}{\sqrt{25}}$$

$$79.88 \pm 2.0639 \times \frac{4.969}{5} = 79.88 \pm 2.049 \text{ or } [77.830 \sim 81.929].$$

(The table value at 5% level of significance with 24 degrees of freedom 2.0639)

The confidence interval is narrow and therefore, we say our sample estimate is close to population parameter.

Example S3-3

To obtain the Confidence limits for the mean using IBM-SPSS, for the data in example 3.8, we enter the data and follow the following steps:

Analyze➔ Descriptive Statistics ➔ Explore:
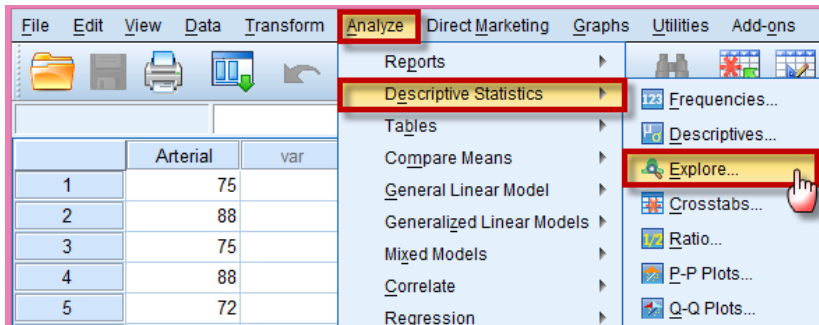
We move the variable into Dependent List and do as follows:
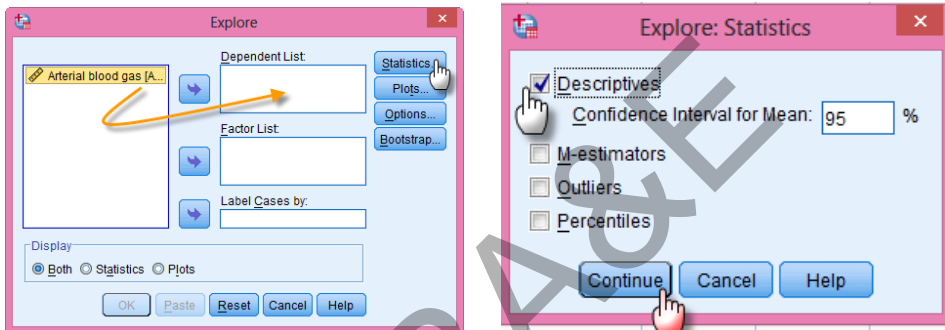
Once we click on OK, we get the following output for the 95%:

Descriptives

			Statistic	Std. Error
Arterial blood gas	Mean		79.88	.994
	95% Confidence Interval for Mean	Lower Bound	77.83	
		Upper Bound	81.93	
	5% Trimmed Mean		79.87	
	Median		79.00	
	Variance		24.693	
	Std. Deviation		4.969	
	Minimum		72	
	Maximum		88	
	Range		16	
	Interquartile Range		8	
	Skewness		.027	.464
	Kurtosis		-.991	.902

Example 3.9:
Among Saudi male children 7% asthma was found during a survey held at Yumboo. The sample size was 200. Estimate 95% confidence limits for population proportion of Yumboo city.

Solution: p   =   0.07

   1 - p   =   0.093

$$p \pm Z_{1-\alpha/2} \; \sqrt{\frac{P\,(1-P)}{n}} \qquad\qquad (3.23)$$

Since sample (n) is large we use 1.96 for 95% confidence level.

$$0.07 \pm 1.96 \; \sqrt{\frac{0.07\,(1-0.07)}{200}}$$

or

$$0.07 \pm 0.035 \quad [0.035,\ 0.105]$$

These limits contain the proportion of children suffering from asthma in the city of Yumboo.

Example S3-4

To obtain the Confidence limits for the proportion using IBM-SPSS, for the data in example 3.9, we enter the data (14 of 1's and 186 of 0's) and follow the following steps:

Analyze➔ Descriptive Statistics ➔ Frequency:
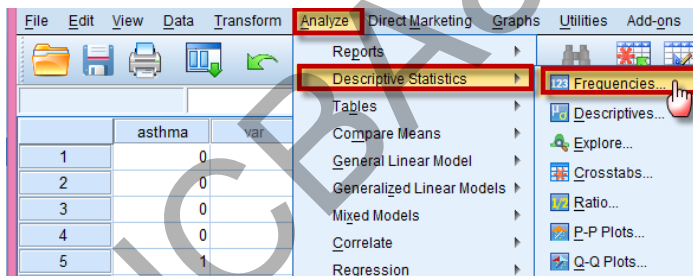
We move the variable into Dependent List and use Bootstrap as follows:

Once we click on OK, we get the following output:

asthma
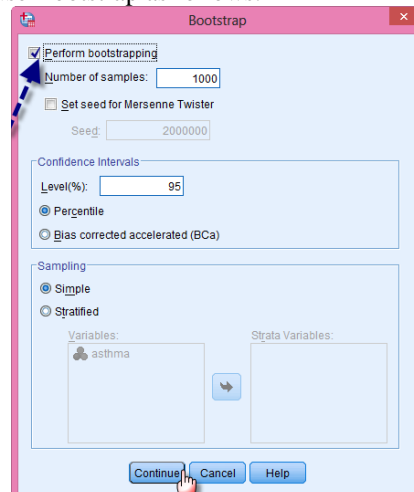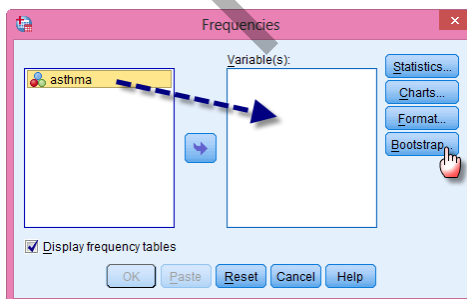
		Frequency	Percent	Valid Percent	Cumulative Percent	Bootstrap for Percent[a]			
						Bias	Std. Error	95% Confidence Interval	
								Lower	Upper
Valid	0	186	93.0	93.0	93.0	.0	1.9	89.5	96.5
	1	14	7.0	7.0	100.0	.0	1.9	3.5	10.5
	Total	200	100.0	100.0		.0	.0	100.0	100.0

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

3.3.6 Stratified Random Sampling

As has been mentioned before, if there is a large variation among the population units, then simple random sampling selection procedure will be less precise, i.e. estimates obtained from using this selection procedure will not be a good estimate of population parameter. If relatively more precise results are to be obtained, then the population is to be divided into different *homogeneous groups*, called *strata*. The strata are formed so that inside each stratum, units are as homogeneous as far as possible. *Stratification* is a process of dividing the population into different strata and selecting a sample of the required number of units within strata, using the simple random sampling selection procedure. Estimates (i.e. mean, proportion, etc.) of each stratum are aggregated to produce an estimate for the whole population using a method of *weighted mean*. There are number of reasons for using this type of selection procedure, i.e. (i) it may increase precision by reducing the variation, (ii) information may be needed for individual strata, (iii) it is easy to control the execution of survey, and (iv) simultaneous work can be started by independent teams. Stratification can be done by area, age, gender, race, area, nationality, type of patients admitted in the hospital, etc. Sample may be selected using a method of proportional allocation. This method of allocation is more scientific and easily under stable by all. This allocation is highly useful if there is a considerable difference between strata averages or proportions and not many differences between the variances within the strata. In the study of population of smokers, the physician may wish to stratify according to type of smokers (light, medium or heavy smokers). The population of smokers may be divided into light smokers, medium smokers or heavy smokers.

An unbiased estimator for population mean for stratified random sampling is

$$\overline{Y}_{st} = \sum_{h=1}^{k} N_h \overline{Y}_h / N \tag{3.24}$$

The variance of sample mean of stratified random sampling is as:

$$\text{Var}(\overline{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{k} [N_h (N_h - n_h) \frac{S_h^2}{n_h}] \tag{3.25}$$

If the allocation of the sample size is proportional then the variance of sample mean will be

$$Var_{prop}(\bar{y}_{st}) = \frac{N-n}{Nn} \sum_{h=1}^{k} W_h S_h^2 \qquad (3.26)$$

If correction factor is ignored then (3.26) takes the following form

$$Var_{prop}(\bar{y}_{st}) = \sum_{h=1}^{k} W_h S_h^2 / n \qquad (3.27)$$

The optimum allocation of sample size when the cost in involved is

$$n_h = \frac{n\ W_h\ S_h / \sqrt{C_h}}{\sum_{h=1}^{k} W_h S_h / \sqrt{C_h}} \qquad (3.28)$$

If the cost is ignored then the above formula takes the following form

$$n_h = \frac{n\ W_h\ S_h}{\sum_{h=1}^{k} W_h\ S_h} \qquad (3.29)$$

The variance of the sample mean for optimum allocation when cost is involved

$$Var_{min}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{k} W_h\ S_h / \sqrt{C_h}\right)\left(\sum_{h=1}^{k} W_h\ S_h \sqrt{C_h}\right) - \frac{1}{N}\sum_{h=1}^{k} W_h S_h^2 \qquad (3.30)$$

If the cost factor is ignored then (3.30) will be

$$Var_{min}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{k} W_h\ S_h / \sqrt{C_h}\right)\left(\sum_{h=1}^{k} W_h\ S_h \sqrt{C_h}\right) \qquad (3.31)$$

Example 3.10:
The smoking information given in the following table and is obtained from census of an Australian City during 1966.

Table 3.7
Stratification with respect to number of cigarette smoking

Type of Smoking	Population Size of adult males
Light smoker < 10	28,900
Medium smoker 10 – 20	38,300
Heavy smoker > 20	52,800
Total	120,000

In order to examine the current smoking habits of adult males in the city, using the information, a sample survey was planned for 1968. It was further decided to use a sample size of 800 adult males.

Solution:

The sample size is allocated to each stratum by using proportional allocation method as:

$$\text{Light smoker} = \frac{28900}{120000} \times 800 = 192.6 \sim 193$$

$$\text{Medium smoker} = \frac{38300}{120000} \times 800 = 255.3 \sim 255$$

$$\text{Heavy smoker} = \frac{52800}{120000} \times 800 = 352 = 352$$

3.3.7 Sytematic Sampling

This selection procedure is different from simple random sampling selection procedure. In simple random sampling procedure every unit is selected by using random numbers table whereas in systematic selection procedure, only the first unit is selected at random and the rest of the units are automatically determined. Suppose there are 500 households in a population and 5 percent sample is to be selected from this population using systematic selection procedure. The sample size comes out to be 25 units. What we do is to calculate N/n = 500/25 = 20 (K), this is called *skip interval*. Note that 25 is the size of the sample. Select one unit randomly from first 20 units, using simple random sampling selection procedure. For this purpose, we will adopt the same procedure as it was done in case of simple random sampling selection procedure. Choose two columns of random number tables, and take the first number that is less or equal to 20 (00 is not considered). By using the random numbers table, 12th household is chosen from first twenty households, then remaining households will be chosen automatically with the skip interval as 12 + 20, 12 + 2(20), 12 + 3(20) and so on. The sample will consist of the following households.

12, 32, 52, 72, 92, 112, 132, 152, 172, 192, 212, 232, 252, 272, 292, 312, 332, 352, 372, 392, 412, 432, 452, 472 and 492.

This procedure of selecting the sample is called systematic selection procedure. The probability of the selection of the sample is 1/K= 1/20, which is in fact the probability with which any member of the group is selected in the sample. This type of selection procedure is very useful when the population size is unknown or sampling frame is not possible. If the population size is known, it is advisable to use simple random sampling selection procedure. In summary, the following remarks are useful for systematic sampling procedure.

  i)   Selection is simple, easier and quicker.
  ii) It involves less cost as compared to simple random sampling.
  iii) A complete and up to date frame is not strictly needed, but the idea of the population is necessary, whereas in simple random sampling selection, procedure a complete and up to date frame is necessary.

In practical situation N/n is not an integer. If population units are 1012 and sample of size 40 is to be selected, the skip interval comes out to be as 1012/40 = 25.3, take 25 as skip interval. If population units are 1025 and a sample of size 40 is to be selected, the skip interval comes out to be as 1025/40 = 25.6, take 26 as skip interval, etc. In most of situations population size is not known, then skip interval is the choice of an experienced sampling statistician. Note that, if sampling frame is available then simple or stratified random sampling is a better choice.

An unbiased estimator for population mean is

$$\bar{y}_{sy} = \frac{1}{nk} \sum_{r=1}^{k} \sum_{i=1}^{n} y_{ri}$$ (3.34)

The variance of sample mean is

$$Var(\bar{y}_{sy}) = \frac{1}{k} \sum_{r=1}^{k} (\bar{y}_r - \bar{Y})^2$$ (3.35)

Other form of variance is

$$Var(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_w^2,$$ (3.36)

where S^2 is total sum of square and S_w^2 is within sum of squares i.e.

$$(nk-1)S^2 = \sum_{r=1}^{k} \sum_{i=1}^{n} [y_{ri} - \bar{Y}]^2$$ (3.37)

and

$$S_w^2 = \frac{1}{k(n-1)} \sum_{r=1}^{k} \sum_{i=1}^{n} (y_{ri} - \bar{y}_r)^2.$$ (3.38)

3.3.8 Single Stage Cluster Sampling

The word cluster was used by Hansen and Hurwitz (1942) to describe a group of elements that constitute a sampling unit. When the entire area containing the population under study is sub-divided into smaller areas and each element of the population is associated with one and only one such small area, the procedure is alternatively called *area sampling*. Cluster sampling is a selection procedure in which population units (elements) are divided into convenient number of groups, called clusters. Each cluster contains some elements. A random sample of some clusters is selected using a simple random sampling procedure or probability proportional to size selection procedure (see next section). Each selected cluster is studied in full. Since all the elements in the sampled cluster are examined in full, therefore it is known as a single stage cluster sampling. Sometimes clusters are known as *primary units* in the context of multistage sampling and elements within each cluster are called *secondary units*.

The concept of cluster was developed for the cases, where the list of elements is not available. For example, in a population survey, list of households is available whereas a list of persons is not. Since cluster sampling consists of groups of elements, approach to

the elements is faster, easier and more convenient than other sampling procedures. Cost will be less if the elements are grouped in a cluster rather than randomly dispersed throughout the area. Since cluster sampling is not a true representative sampling method as compared to simple random method, therefore, the efficiency will be less. The efficiency of clustering sampling depends on the size of the cluster. If the size of clusters is large and the number of clusters is less the efficiency will also be decreased, but if the size is small and number of clusters is more, the efficiency will be increased. Cluster sampling procedure is different from stratified sampling in the sense that in the former case all elements within groups (clusters) are studied.

The cluster sampling procedure is explained below:

Suppose we would like to hold a TB survey in Dammam City and the list of households and list of persons are not known to us. We can divide the whole city into different sectors (clusters) say (40). We try to divide the population into equal size clusters as far as possible. Suppose 10 sectors (clusters) are likely to be selected. We will use simple random sampling procedure to select 10 clusters. Then all the 10 selected clusters will be examined fully to check the prevalence of TB.

If the clusters vary in size then, simple random selection procedure will not be appropriate method of selection. We will select the sample keeping in view, the size of the clusters. The selection used in these situations will be known as *probability proportional to size sampling selection procedure*.

An unbiased estimator of population mean is

$$\bar{y}_e = \frac{1}{nM} \sum_{i=1}^{n} \sum_{j=1}^{M} y_{ij} \tag{3.39}$$

The variance of sample mean is

$$\text{Var}(\bar{y}_e) = \frac{N-n}{Nn(N-1)} \sum_{i=1}^{N} (\bar{Y}_i - \bar{Y})^2 \tag{3.40}$$

An unbiased variance estimator of (3.40) is

$$\text{var}(\bar{y}_e) = \frac{N-n}{Nn} \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_e)^2 \tag{3.41}$$

where $\bar{y}_i$ is mean of the cluster of population and $\bar{Y}$ the mean of population.

3.3.9 Probability Proportional to Size Sampling Procedure

In all the above selection procedures, equal probability of selection was involved i.e. each unit or each cluster has equal chance to be in the sample, but in probability proportional to size sampling procedure, units are selected keeping in mind the size of units. This method is also known as sampling with unequal probabilities of selection procedure. Hansen and Hurwitz (1943) suggested this selection procedure.

An unbiased estimator for population total is given as

$$y'_{HH} \text{ or } y'_{PPS} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}, \tag{3.42}$$

where p_i is the probability of selection of the ith population unit to be in sample.

The variance and unbiased variance estimator are given respectively

$$Var(y'_{HH}) = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{Y_i^2}{p_i} - Y^2\right) \tag{3.43}$$

and

$$var(y'_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - y'_{HH}\right)^2 \tag{3.44}$$

Here only brief introduction is given, if anyone is interested, he may refer to a monograph on *sampling with unequal probabilities by Brewer and Hanif (1983)*. This selection procedure is explained as:

Example 3.11:
Areas of 20 sectors and numbers of households in each area are given. Select a sample of 5 sectors.

Solution
To select a sample, some basic calculations are required. In column 5, proportions (probability) of the area of each sector, and in column 6 cumulative size of the area have been calculated. In column 7 range of each sector is given. The ranges are given only for convenience otherwise it is not essential. Suppose we like to select a sample of 5 sectors under this selection procedure. Five random numbers are selected between 001 and 448. These random numbers are 153, 52, 414, 283 and 177. They fall in the ranges 151 - 156, 43 - 58, 316 - 438, 257 - 310 and 162 - 256. Therefore, sector numbers, 8, 4, 16, 11, and 10 are in the sample as given in Table 3.9.

Table 3.8:
Population of house hold along with area

Sector No. (1)	Area Z_i (2)	No. of Households (3)	Z_i/Z (4)	Proportion Z_i/Z (5)	Cumulative Z_i (5)	Range (6)
1	33	2328	33/448	0.074	33	1 - 33
2	8	754	8/448	0.018	41	34 - 41
3	1	105	1/448	0.002	42	42
4	16	949	16/448	0.035	58	43 - 58
5	43	3091	43/448	0.096	101	59 - 101
6	40	1736	40/448	0.089	141	102 - 141
7	9	840	9/448	0.020	150	142 - 150
8	6	311	6/448	0.014	156	151 - 156
9	5	0	5/448	0.011	161	157 - 161
10	95	3044	95/448	0.212	256	162 - 256
11	54	2483	54/448	0.121	310	257 - 310
12	1	128	1/448	0.002	311	311
13	1	102	1/448	0.002	312	312
14	2	60	2/448	0.005	314	313 - 314
15	1	0	1/448	0.002	315	325
16	123	11799	123/448	0.275	438	316 - 438
17	1	26	1/448	0.002	439	439
18	3	317	3/448	0.007	442	440 - 442
19	4	190	4/448	0.009	446	433 - 446
20	2	180	2/448	0.005	448	447 - 448
Total	448 = Z	28443				

Table 3.9:
Sample selected from Table 3.3

Random Numbers	Sr. No. of Sector	Number of Houses	Probability of Selection
52	4	949	0.035
153	8	311	0.014
177	10	3044	0.212
283	11	2483	0.121
414	16	11799	0.275

This selection procedure is with replacement and a cluster can be selected twice.

There are over 100 selection procedures relating to probability proportional to size without replacement. Here only one selection procedure that is most frequently used by non-statisticians is described.

3.3.10 Random Systematic Selection Procedure

This selection procedure is simple and easy for the selection of a sample. It is commonly used in a large-scale survey. In this selection procedure the population units (sectors) are randomly arranged. The size of each population is mentioned against every unit. The size may be area or may be total number of households in that sector.

Example 3.12:

Suppose we have a population of 8 sectors. Select a sample of 3 sectors. These sectors are arranged randomly before the selection of sample. Against each sector, the size of sector is given.

Table 3.10:
Population of 8 Sectors

Sectors	Size of Sector	Cumulative Size	Cluster to be Selected
1	15	15	
2	81	96	36
3	26	112	
4	42	164	136
5	20	184	
6	16	200	
7	45	245	236
8	55	300	
Total	300		

A sample of 3 sectors is to be selected. Divide the total by the sample size to obtain skip interval, i.e. $300/3 = 100$. Select a random start from 001 to 300. Let the random start be 36, so the first sector selected will be the 2nd one. For the selection of second and third sectors, we proceed as: add $36 + 100 = 136$ and $36 + 2(100) = 236$. 136 falls against 164 and 236 falls against 245. So, 4th and 7th sectors are selected. As a result 2nd, 4th and 7th sectors are in the sample.

3.3.11 Multistage Sampling

Simple random sampling and stratified random sampling selection procedures described above may be considered as a single stage sampling procedure. In a single stage selection procedure, a sample is drawn from a population and informations are obtained from the sampling units. In multistage sampling, a population is divided into a number of large units and a sample of large units is selected either using equal probability selection procedure or using probability proportional to size selection procedure. Each of selected large unit is further subdivided into smaller units, and a sample of these units is selected from each of the selected large units. Kendall and Bukland (1980) in the Dictionary of Statistical Terms define a multistage sample as one *which is selected by stages, the sample units at each stage being sub-sampled from the (larger) units chosen at the previous stage* or in multistage sampling selection is carried out in a succession of stages. Typical example of multistage sampling may be a health survey in Eastern Province, Saudi Arabia where the Eastern Province is divided into primary care centers as the first stage units. A sample may be selected from primary care centers as primary sampling units (P.S.U.) From each primary care centers; sample of patients may be selected as second stage units (SSU) and so on.

Multistage sampling is most frequently used in field surveys where the list of last stage units is difficult to get. Though by using multistage sampling precision is lost but it is much cheaper and quicker than any other design.

Table 3.11 (Random Digits)

57780	97609	52482	12783	88768	12323	64967	22970	11204	37576
68327	00067	17487	49149	25894	23639	86557	04139	10756	76285
55888	82253	67464	91628	88764	43598	45481	00331	15900	97699
84910	44827	31173	44247	56573	91759	79931	26644	27048	53704
35654	53638	00563	57230	07395	10813	99194	81592	96834	21374
46381	60071	20835	43110	31842	02855	73446	24456	24268	85291
11212	06034	77313	66896	47902	63483	09924	83635	30013	61791
49703	07226	73337	49223	73312	09534	64005	79267	76590	26066
05482	30340	24606	99042	16536	14267	84084	16198	94852	44305
92947	65090	47455	90675	89921	13036	92867	04786	76776	18675
51806	61445	32437	01129	03644	70024	07629	55805	85616	59569
16383	30577	91319	67998	72423	81307	75192	80443	09651	30068
30893	85406	42369	71836	74479	68273	78133	34506	68711	58725
59790	11682	63156	10443	99033	76460	36814	36917	37232	66218
06271	74980	46094	21881	43525	16516	26393	89082	24343	57546
93325	61834	40763	81178	17507	90432	50973	35591	36930	03184
46690	08927	32962	24882	83156	58597	88267	32479	80440	41668
82041	88942	57572	34539	43812	58483	43779	42718	46798	49079
14306	04003	91186	70093	62700	99408	72236	52722	37531	24590
63471	77583	80056	59027	37031	05819	90836	19530	07138	36431
68467	17634	84211	31776	92996	75644	82043	84157	10877	12536
94308	57895	08121	07088	65080	51928	74237	00449	86625	06626
52218	32502	82195	43867	79935	34620	37386	00243	46353	44499
46586	08309	52702	85464	06670	18796	74713	81632	34056	56461
07869	80471	69139	82408	33989	44250	79597	15182	14956	70423
46719	60281	88638	26909	32415	31864	53708	60219	44482	40004
74687	71227	59716	80619	56816	73807	94150	21991	22901	74351
42731	50249	11685	54034	12710	35159	00214	19440	61539	25717
71740	29429	86822	01187	96497	25823	18415	06087	05886	11205
96746	05938	11828	47727	02522	33147	92846	15010	96725	67903
27564	81744	51909	36192	45263	33212	71808	24753	72644	74441
21895	29683	26533	14740	94286	90342	24671	52762	22051	31743
01492	40778	05988	65760	13468	31132	37106	02723	40202	15824
55846	19271	22846	80425	00235	34292	72181	24910	25245	81239
14615	75196	40313	50783	66585	39010	76796	31385	26785	66830
77848	15755	91938	81915	65312	86956	26195	61525	97406	67988
87167	03106	52876	31670	23850	13257	77510	42393	53782	32412
73018	56511	89388	73133	12074	62538	57215	23476	92150	14737
29247	67792	10593	22772	03407	24319	19525	24672	21182	10765
17412	09161	34905	44524	20124	85151	25952	81930	43536	39705
68805	19830	87973	99691	25096	41497	57562	35553	77057	06161
40551	36740	61851	76158	35441	66188	87728	66375	98049	84604
90379	06314	21897	42800	63963	44258	14381	90884	66620	14538
09466	65311	95514	51559	29960	07521	42180	86677	94240	59783
15821	25078	19388	93798	50820	88254	20504	74158	35756	42100
10328	60890	05204	30069	79630	31572	63273	13703	52954	72793
49727	08160	81650	71690	56327	06729	22495	49756	43333	34533
71118	41798	34541	76132	40522	51521	74382	06305	11956	30611
53253	23100	03743	48999	37736	92186	19108	69017	21661	17175
12206	24205	32372	46438	67981	53226	24943	68659	91924	69555

Tables 3.12
Estimation of sample size with absolute precision (95%).

P↓d→	.01	.02	.03	.04	.05	.06	.07	.08	.09	.1
.01	380	95	42	24	15	11	8	6	5	4
.02	753	188	84	47	30	21	15	12	9	8
.03	1118	279	124	70	45	31	23	17	14	11
.04	1475	369	164	92	59	41	30	23	18	15
.05	1825	456	203	114	73	51	37	29	23	18
.06	2167	542	241	135	87	60	44	34	27	22
.07	2501	625	278	156	100	69	51	39	31	25
.08	2827	707	314	177	113	79	58	44	35	28
.09	3146	787	350	197	126	87	64	49	39	31
.1	3457	864	384	216	138	96	71	54	43	35
.15	4898	1225	544	306	196	136	100	77	60	49
.2	6147	1537	683	384	246	171	125	96	76	61
.25	7203	1801	800	450	288	200	147	113	89	72
.3	8067	2017	896	504	323	224	165	126	100	81
.35	8740	2185	971	546	350	243	178	137	108	87
.4	9220	2305	1024	576	369	256	188	144	114	92
.45	9508	2377	1056	594	380	264	194	149	117	95
.5	9604	2401	1067	600	384	267	196	150	119	96

Table 3.13
Estimation of sample size for absolute precision (99%)

P↓d→	.01	.02	.03	.04	.05	.06	.07	.08	.09	.1
.01	658	165	73	41	26	18	13	10	8	7
.02	1305	326	145	82	52	36	27	20	16	13
.03	1937	484	215	121	77	54	40	30	24	19
.04	2556	639	284	160	102	71	52	40	32	26
.05	3162	790	351	198	126	88	65	49	39	32
.06	3754	939	417	235	150	104	77	59	46	38
.07	4333	1083	481	271	173	120	88	68	53	43
.08	4899	1225	544	306	196	136	100	77	60	49
.09	5452	1363	606	341	218	151	111	85	67	55
.1	5991	1498	666	374	240	166	122	94	74	60
.15	8487	2122	943	530	339	236	173	133	105	85
.2	10650	2663	1183	666	426	296	217	166	131	107
.25	12481	3120	1387	780	499	347	255	195	154	125
.3	13978	3495	1553	874	559	388	285	218	173	140
.35	15143	3786	1683	946	606	421	309	237	187	151
.4	15975	3994	1775	998	639	444	326	250	197	160
.45	16475	4119	1831	1030	659	458	336	257	203	165
.5	6641	4160	1849	1040	666	462	340	260	205	166

Table 3.14
Estimation of sample size with relative precision (95%)

P↓D→	.01	.02	.03	.04	.05	.06	.07	.08	.09
.01	3803184	950796	422576	237699	152127	105644	77616	59425	46953
.02	1882384	470596	209154	117649	75295	52288	38416	29412	23239
.03	1242117	310529	138013	77632	49685	34503	25349	19408	15335
.04	921984	230496	102443	57624	36879	25611	18816	14406	11383
.05	729904	182476	81100	45619	29196	20275	14896	11405	9011
.06	601851	150463	66872	37616	24074	16718	12283	9404	7430
.07	510384	127596	56709	31899	20415	14177	10416	7975	6301
.08	441784	110446	49087	27611	17671	12272	9016	6903	5454
.09	388428	97107	43159	24277	15637	10790	7927	6069	4795
.1	345744	86436	38416	21609	13830	9604	7056	5402	4268
.15	217691	54423	24188	13606	8708	6047	4443	3401	2688
.2	153664	38416	17074	9604	6147	4268	3136	2401	1897
.25	115248	28812	12805	7203	4610	3210	2352	1801	1423
.3	89637	22409	9960	5602	3585	2490	1829	1401	1107
.35	71344	17836	7927	4459	2854	1982	1456	1115	881
.4	57624	14406	6403	3601	2305	1601	1176	900	711
.45	46953	11738	5217	2935	1878	1304	968	734	580
.5	38416	9604	4268	2401	1537	1067	784	600	474
.55	31431	7858	3492	1964	1257	873	641	491	388
.6	25611	6043	2846	1601	1024	711	523	400	316
.65	20686	5171	2298	1293	827	575	422	323	256
.7	16464	4116	1829	1029	669	457	336	257	203
.75	12805	3201	1423	800	512	366	261	200	158
.8	9604	2401	1067	600	364	267	196	150	119
.85	6779	1695	753	424	271	188	138	106	84
.9	4268	1067	474	267	171	119	87	67	53
.95	2022	505	225	126	81	56	41	32	25

Table 3.15
Estimation of sample size for relative precision (99%)

P↓D→	.01	.02	.03	.04	.05	.06	.07	.08	.09
.01	6589836	1647459	732204	411865	263583	183051	134486	102966	81356
.02	3261636	815409	362404	203852	130465	90601	66564	50963	40267
.03	2152236	538059	239137	134515	86089	59784	43923	33629	26571
.04	1597536	399384	177504	99846	63901	44376	32603	24962	19723
.05	1264716	316179	140524	79045	50589	36131	25811	19761	15614
.06	1042836	260709	115871	66177	41713	28968	21282	16294	12875
.07	884350	221088	96261	55272	35374	24565	18048	13818	10918
.08	765486	191372	85054	47843	30619	21264	15622	11961	9450
.09	673036	168259	74782	42065	26921	18695	13735	10516	8309
.1	599076	149769	66564	37442	23963	16641	12226	9361	7396
.15	377196	94299	41911	23575	15088	10478	7698	5894	4657
.2	266256	66564	29684	16641	10650	7396	5434	4160	3287
.25	199692	49923	22188	12481	7988	5547	4075	3120	2465
.3	156316	36829	17257	9707	6213	4314	3170	2427	1917
.35	123619	30905	13735	7726	4945	3434	2523	1932	1526
.4	99846	24961	11094	6240	3994	2774	2038	1560	1233
.45	81366	20339	9040	5085	3254	2260	1660	1271	1004
.5	66564	16641	7396	4160	2663	1849	1358	1040	822
.55	54461	13615	6051	3404	2178	1513	1111	851	672
.6	44376	11094	4931	2774	1775	1233	906	693	548
.65	35842	8961	3982	2240	1434	996	731	560	442
.7	28527	7132	3170	1783	1141	792	582	446	352
.75	22198	5547	2465	1387	888	616	453	347	274
.8	18641	4160	1849	1040	666	452	340	260	205
.85	11747	2937	1305	734	470	326	240	184	145
.9	7396	1849	822	452	296	205	151	116	91
.95	3503	876	389	219	140	97	71	55	43

Table-3.16: (Percentile of t-distribution)
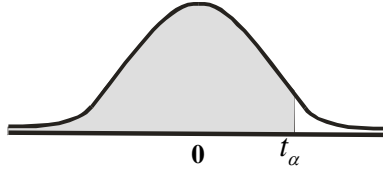
d.f/α	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	3.078	6.3138	12.706	31.821	63.6570
2	1.886	2.9200	4.3027	6.965	9.9248
3	1.638	2.3534	3.1825	4.541	5.8409
4	1.533	2.1318	2.7764	3.747	4.6041
5	1.476	2.0150	2.5706	3.365	4.0321
6	1.440	1.9432	2.4469	3.143	3.7074
7	1.415	1.8946	2.3646	2.998	3.4995
8	1.397	1.8595	2.3060	2.896	3.3554
9	1.383	1.8331	2.2622	2.821	3.2498
10	1.372	1.8125	2.2281	2.764	3.1693
11	1.363	1.7959	2.2010	2.718	3.1058
12	1.356	1.7823	2.1788	2.681	3.0545
13	1.350	1.7709	2.1604	2.650	3.0123
14	1.345	1.7613	2.1448	2.624	2.9768
15	1.341	1.7530	2.1315	2.602	2.9467
16	1.337	1.7459	2.1199	2.583	2.9208
17	1.333	1.7396	2.1098	2.567	2.8982
18	1.330	1.7341	2.1009	2.552	2.8784
19	1.328	1.7291	2.0930	2.539	2.8609
20	1.325	1.7247	2.0860	2.528	2.8453
21	1.323	1.7207	2.0796	2.518	2.8314
22	1.321	1.7171	2.0739	2.508	2.8188
23	1.319	1.7139	2.0687	2.500	2.8073
24	1.318	1.7109	2.0639	2.492	2.7969
25	1.316	1.7081	2.0595	2.485	2.7874
26	1.315	1.7056	2.0555	2.479	2.7787
27	1.314	1.0733	2.0518	2.473	2.7707
28	1.313	1.1701	2.0484	2.467	2.7633
29	1.311	1.6991	2.0452	2.462	2.7564
30	1.310	1.6973	2.0423	2.457	2.7500
35	1.3602	1.6896	2.0301	2.438	2.7239
40	1.3031	1.6839	2.0211	2.423	2.7045
45	1.3007	1.6794	2.0141	2.412	2.6896
50	1.2987	1.6759	2.0086	2.403	2.6778
60	1.2959	1.6707	2.0003	2.390	2.6603
70	1.2938	1.6669	1.9945	2.381	2.6480
80	1.2922	1.6641	1.9901	2.374	2.6388

Chapter 4

Hypothesis Testing Procedures

4.1 Introduction

Generally there are two methods available and widely used for making inferences about the population parameters i.e.

(a) Inference may be drawn through *confidence limits*.

(b) Inference may be drawn about specific value of the population through *testing of hypotheses*.
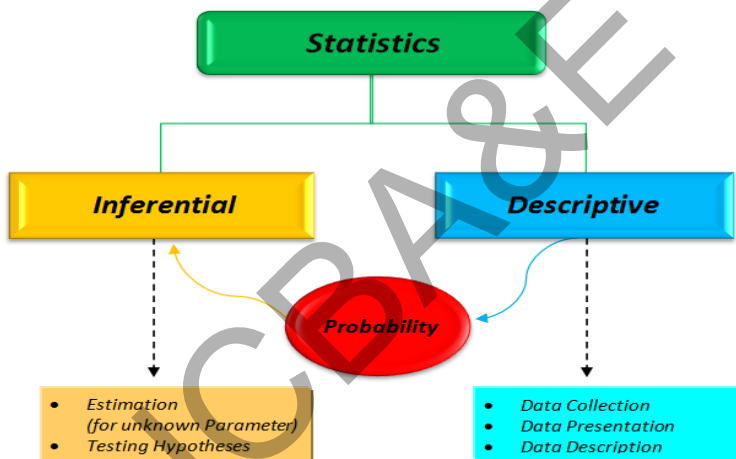
Though the confidence intervals and testing of hypotheses are related and either can be used in making decision about the population parameters yet the decision can be made in a more effective way by the use of testing of hypothesis procedure. Two examples are given to explain how the method of confidence intervals is used to make decision about a parameter.

Example 4.1:
Suppose a research worker, working for the Environmental Protection Agency [EPA] wants to determine whether the *mean level* of a certain type of pollutant released into the atmosphere by a certain chemical company meets the guidelines set by the EPA. If 4 parts per million is the upper limit allowed by the EPA then the research worker will use a sample data (i.e. daily pollution measurements) to decide whether the *mean is greater than 4*. If, for example, 95% confidence interval for mean contains numbers greater than 4, then the research worker would suspect that the mean exceeds the established limits.

Example 4.2:
Suppose that a certain hospital purchases some syringes from a manufacturer. The manufacturer claims that not more than 1% of the equipments are defective. It is not

possible for hospital authority to test each and every syringe; they will take a random sample to test the defective items. The hospital authority wants to see whether the proportion of defective items exceeds 1% or not, based on the information contained in the sample. If the sample proportion falls inside the confidence limits of *1%* then the hospital authority will accept the lot, otherwise, the lot will not be accepted.

This is how inferences are drawn through confidence intervals.

Whenever any research worker in any field wants to test a new theory, he always first formulates a hypothesis that provides an explanation of his experience. He makes some assumptions about some characteristic of a population, tries to support it by information obtained from sample data. These assumptions are called *hypotheses*. This is the beginning of the concept of testing of hypotheses. The purpose of hypothesis testing is to help the research worker in making decision for the population on the basis of the information collected through sample. For example, we may examine a manufacturer's claim that his drug on the average is more effective than an alternative drug already available in the market. We will reach the decision through a sample of patients on whom the drugs are tried.

Before we pass on to the application of testing of hypotheses it is useful and important to explain some basic terms to understand the concept of testing of hypothesis. More precisely one must understand what statistical hypothesis is? How should the tests be performed? What types of errors one can face? How to draw conclusion(s) regarding parameter(s) on the basis of sampled observations? What p-value is?

4.1.1 Hypothesis or a Statistical Hypothesis

As mentioned earlier, a research worker always makes certain assumptions, when he wants to test a new theory. In statistics, it is known as a *hypothesis*. A *hypothesis* or a *statistical hypothesis* is a statement about the *specified value(s)* of the parameter(s). In its most general form a statistical hypothesis tells us something about this distribution of an *observed random variable*. This statement may be true or may not be true. In fact this is a baseline to start the experiment. We set up two types of statistical hypotheses, viz.

(i) Null hypothesis H_0 and (ii) Alternative hypothesis H_1

The *Null Hypothesis* states that there is no difference between the *specified* or *stated value* (μ_0 = mean or P_0 = proportion) and actual unknown values of **μ, or P** of the parameters. An initial hypothesis of equivalence of two statements is called *Null Hypothesis*. For example, a manufacturer of some brand of cigarette claims that 30% of the smokers prefer his brand of cigarettes. The null hypothesis will be, that the claim of the manufacturer is correct. A manufacturer of a drug claims that the drug manufactured by him is more effective than the drug already available in the market. The null hypothesis states that there is no difference between the efficacies of the two drugs.

An *alternative hypothesis* states that the specified or stated value and an *actual unknown value* of the parameter are not equivalent or the null hypothesis is not true. In the first case

H_0: P = 0.30   (null hypothesis)
H_1 : P > 0.30   (alternative hypothesis)

and in the second case

H_0 : $P_1 = P_2$ (There is no difference between two types of drugs: null hypothesis)
H_1: $P_1 > P_2$ (Drug one is superior to that of the second drug: alternative hypothesis)

There is an unstated willingness on this part of the investigator to accept H_1 in case he/she rejects H_0.

An accepted convention in the simple testing of hypotheses is to write null hypothesis (H_0) with an equality (=) sign and the alternative could be greater (>) or less (<) or not equal (≠) depending on the problem. If not equal (≠) then it is called *two-tail* test otherwise it is known as *one-tail* test. The one-tail and two-tail tests are explained in the following subsections:

4.1.2 One-tail and Two-tail Test

One-tail test is that in which alternative hypothesis is directional. This includes either less (<) or greater (>), i.e. unknown mean or proportion is either greater or less than specified or stated mean or proportion. *Two- tail test is one in which the alternative hypothesis does not specify departure from null hypothesis in particular direction.* One-tail and two-tail tests are explained in Table 4.1 and Table 4.2

Table 4.1
One-tail test of mean and proportion for one sample and two samples

	Mean	Proportion
One sample	H_0 : $\mu = \mu_0$ H_1 : $\mu > \mu_0$	H_0 : $P = P_0$ H_1 : $P > P_0$
Two samples	H_0 : $\mu_1 = \mu_2$ H_1 : $\mu_1 > \mu_2$	H_0 : $P_1 = P_2$ H_1 : $P_1 > P_2$

Table 4.2
Two-tailed test of mean and proportion for one and two samples

	Mean	Proportion
One sample	H_0 : $\mu = \mu_0$ H_1 : $\mu \neq \mu_0$	H_0 : $P = P_0$ H_1 : $P \neq P_0$
Two Samples	H_0 : $\mu_1 = \mu_2$ H_1 : $\mu_1 \neq \mu_2$	H_0 : $P_1 = P_2$ H_1 : $P_1 \neq P_2$

An incidence of tuberculosis among people living in Eastern Province of Saudi Arabia is known *to be not more than 0.03*. After conducting a medical survey, the researcher believes that the incidence is much higher. The researcher is interested in detecting whether true incidence of tuberculosis is larger than 0.03. He forms the null and alternative hypotheses (one-tail) as:-

H_0 : P = 0.03
H_1 : P > 0.03

If the researcher is interested in detecting that there is no difference between incidences of two provinces of Saudi Arabia, then his null and alternative hypotheses (two-tail) are

$H_0 : P_1 = P_2$
$H_1 : P_1 \neq P_2$

4.1.3 Level of Significance (α)

The probability of rejecting the null hypothesis, when the null hypothesis is true is called the *level of significance or probability of type I error*. This probability is generally specified before the sample is drawn. Level of significance is generally chosen either 1% or 5%. In medical trials, because human lives are involved therefore, sometimes level of significance may go as low as 0.1% or even 0.05%. When we say that the level of significance is 5%, we mean that there are 5 in 100 chances that the null hypothesis is rejected when it is in fact is true and we are 95% confident regarding our decision. Commonly, the level of significance is denoted by the Greek letter α (Alpha).

4.1.4 Confidence Level (1 - α)

The complement of probability α is (1 - α) that is called *confidence level* or *confidence coefficient*. It gives the probability of accepting H_0 whenever it is true.

4.1.5 A Critical Value

A critical value is a boundary or separation point between rejection and acceptance regions. For example if we choose 5% level of significance, then the boundary points for a two-tailed test (critical values) at 5% level of significance are -1.96 and 1.96, see Fig. 4.1.
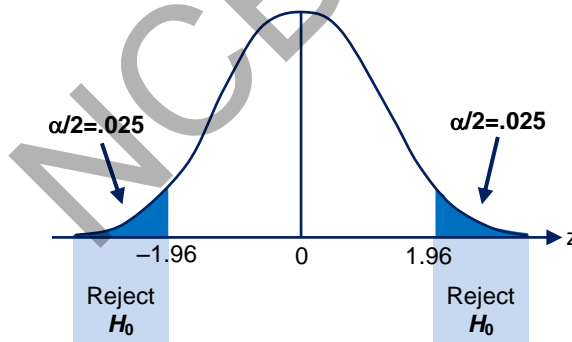
Fig. 4.1: Critical values

The points beyond 1.96 and -1.96 are called *rejection regions* and points between -1.96 to 1.96 is known as *acceptance region* for two-tail- test. The points -1.96 and 1.96 are called *critical values*. If it is a one-tail-test then for the same level of significance, the rejection and acceptance regions are shown in Fig. 4.2. A critical value depends on the level of significance of the test. For large sample, the critical values or a critical z-value for one-tail and two-tail tests, commonly used are as given in Table 4.3.

Table 4.3
Level of Significance for Acceptance Region for
One-tailed and two tailed tests

Level of Significance	Two-tail test	One-tail test
1%	-2.58 to +2.58	-2.33 to +∞ or -∞ to 2.33
5%	-1.96 to +1.96	-1.645 to +∞ or -∞ to 1.645

Rejection Region   Acceptance Region

Acceptance Region   Rejection Region
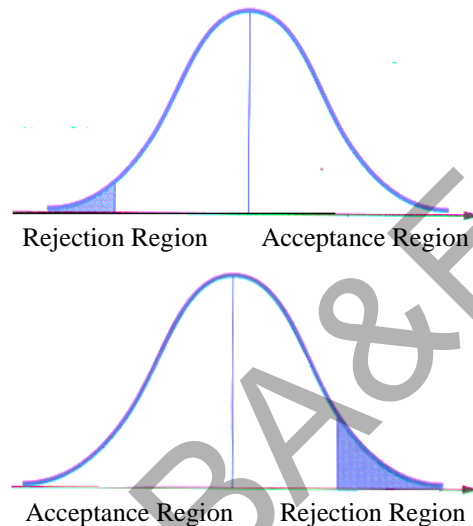
Fig. 4.2: Rejection and acceptance regions for given level of significance

4.1.6 Test Statistic

A decision based on a sample, is made to reject or accept null hypothesis. These decisions depend on the value of some statistic with a probability distribution Such a statistic is called a *test- statistic*.

4.1.7 Type I and Type II Errors

The main aim of the testing of the hypotheses is to make decision whether to accept or not to accept the null hypothesis in favour of an alternative hypothesis. We always like to make correct decision, but this decision depends on the sampled observations. In spite of every precaution taken, there is a chance of committing an error. *We may reject null hypothesis when it is true or we may accept the null hypothesis when it is false. Therefore, two types of errors may be committed during the process of testing of hypothesis, which are known as Type I and Type II errors.*

Type I error occurs when the null hypothesis is true and it is not accepted whereas Type II error occurs when the null hypothesis is false and it is accepted. The probability of committing Type I error is denoted by α (Alpha) whereas the probability of committing Type II error is denoted by β (Beta). There is an interesting relationship between the

probabilities of two types of errors *for a fixed sample size*. If one increases the other decreases and if one decreases, the other increases.

There are four possibilities regarding the correctness of the decision in any hypothesis test. These possibilities are explained in Table 4.4 on next page.

We see in Table 4.4 that false positive corresponds to Type I error and false negative corresponds to Type II error

<div align="center">

Table 4.4
Types of Errors

</div>

Decision $\rightarrow$ Hypothesis $\downarrow$	Accept H_0	Reject H_0
H_0 is true	True + Correct decision	False + Type I error
H_0 is false	False - Type II error	True - Correct decision

Note that Type I error is more serious than Type II error. If H_0 is rejected then usually one is not clear about what to substitute in its place. So we want to avoid unnecessary rejection of a true H_0. The conventional practice is to ensure that probability of Type I error is controlled below a predetermined level of tolerance and then to choose among these tests, the one with the smallest possible probability of Type II error i.e. to fix probability of Type I error and then select an appropriate test which minimizes probability of Type II error.

In practice, we are very careful in stating the decision. If sampled observations do not provide sufficient evidence to support the null hypothesis, we prefer the decision, and say, *we fail to reject* the null hypothesis. If we were to accept the null hypothesis, the reliability of the conclusion is measured by the probability of Type II error. The power of test for testing the hypothesis under consideration where $\bar{x} \geq A$ is unknown. For given α and β, we have the following two equalities for determining these values.

$$\frac{\sqrt{n}}{\sqrt{2\pi}} \int_{A}^{\infty} \exp\left[-\frac{n(\bar{x} - \mu_o)^2}{2}\right] d\bar{x} = \alpha \tag{4.1}$$

and

$$\frac{\sqrt{n}}{\sqrt{2\pi}} \int_{A}^{\infty} \exp\left[-\frac{n(\bar{x} - \mu_1)^2}{2}\right] d\bar{x} = 1-\beta \tag{4.2}$$

hold.

Let us write $A = \mu_0 + z_\alpha / \sqrt{n}$ where z_α is chosen in such a way that for a random variable y with normal distribution $N(0,1)$, $\mathbf{P}(Z \geq z_\alpha) = \alpha$. From (4.2), we have $A = \mu_1 + z_\beta / \sqrt{n}$, where z_β is chosen in such a way that $\mathbf{P}[Z \leq z_\beta] = \beta$.

From the equality:

$$\mu_0 + \frac{z_\alpha}{\sqrt{n}} = \mu_1 + \frac{z_\beta}{\sqrt{n}} \ ,$$

we obtain

$$n = \frac{\left(z_\alpha - z_\beta\right)^2}{\left(\mu_1 - \mu_0\right)^2} \qquad\qquad (4.3)$$

and

$$A = \frac{z_\alpha \mu_1 - z_\beta \mu_0}{z_\alpha - z_\beta} \qquad\qquad (4.4)$$
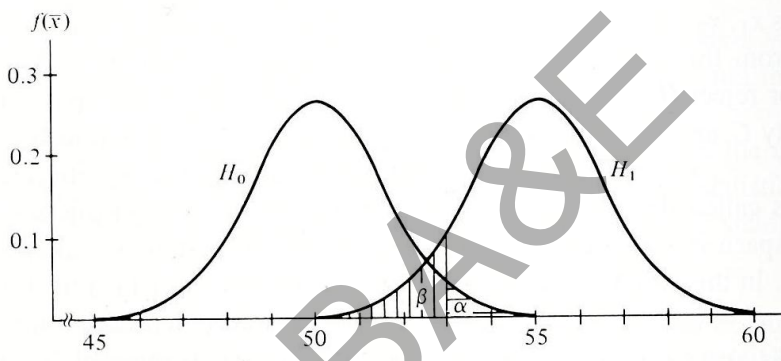
Fig. 4.3: The power of a test

The power 1-β is defined as the probability of acceptance of H_1 when it is true or rejecting of H_0 when it is false. Unfortunately the probability of Type II error is not constant, but depends on the specific alternative value of the parameter. In order to calculate Type I error and Type II error some examples are given

The purpose of presenting examples is not that health scientists should calculate and find amounts of probabilities of Type I and Type II errors but the main objective is to show how the probabilities of Type I and Type II errors behave with the increase of sample size, so that one should be careful in testing of hypotheses.

Example 4.3.
In a large school of USA, the proportion of high school students that regularly use some form of illegal drug is reported to be 0.50. The school authority took a random sample of 200 students and it was found that 45% of the students were using illegal drug. If the rule of rejection is to calculate Z and reject H_0 whenever $Z \le -1.60$

   i.   Would you reject H_0?.

   ii. Calculate the probability of Type I error.

Solution:

$H_0 : P = 0.50$

$H_1 : P < 0.50$

Since $p = 0.45$ therefore, $\mathbf{P}$(Type I error) $= \mathbf{P}(p < 0.45)$

We know from Chapter 2

$$Z = \frac{0.45 - 0.50}{\sqrt{\dfrac{0.50 \times 0.50}{200}}} = \frac{-0.05}{0.035} = -1.43$$

(i) We will not reject as the calculated value of Z, is less than -1.60

(ii) The probability of type I error for Z=+1.43 will be calculated as:

$\alpha = \mathbf{P}[p < 0.45] = \mathbf{P}[Z > 1.43] = 0.076$ (Table 2.6).

Therefore, one is planning to use about 7.6% level of significance.

Probability of Type I or Type II error relate to a well-defined rule of rejection. For example if you decide that you will reject H_0 whenever Z calculated exceeds a given value (say 1.64). Then correspondingly to this you say probability of Type I error is such and such.

Example 4.4:

For a certain hypothesis $H_0 : \mu = 50$ versus $H_1 : \mu > 50$. Suppose $\sigma = 9.0$. Calculate probability of Type I error for the following cases:

i)   A random sample of 40 observations was taken and found that sample mean is 52.0.

ii) A random sample of 60 observations was taken and found that sample mean is 52.0.

iii) A random sample of 120 observations was taken and found that sample mean is 52.0.

Comment what happens if sample mean is fixed and sample size is increasing.

Solution:

(i) $H_0 : \mu = 50$

$H_1 : \mu > 50$

Sample mean ($\bar{x}$) = 52, n=40, and σ =9.0)

$$Z = \frac{52 - 50}{\dfrac{9}{\sqrt{40}}} = \frac{2}{9} \times \sqrt{40} = 1.40$$

Therefore, the probability of type I error is

$\alpha = \mathbf{P}[Z > 1.40] = 0.0808$ (From Table 2.6) $= 8\%$

(ii) $H_0 : \mu = 50$

$H_1 : \mu > 50$

$$Z = \frac{52 - 50}{\frac{9}{\sqrt{60}}} = \frac{2}{9} \times \sqrt{60} = 1.72$$

Therefore, the probability of type I error is

$\alpha = \mathbf{P}[Z > 1.72] = 0.0427$ (From Table 2.6) $= 4.3\%$

iii) $H_0 : \mu = 50$

$H_1 : \mu > 50$

$$Z = \frac{52 - 50}{\frac{9}{\sqrt{120}}} = \frac{2}{9} \times \sqrt{120} = 2.43$$

Therefore, the probability of type I error is

$\alpha = \mathbf{P}[Z > 2.43] = 0.0075$ (From Table 2.6) $= 0.75\%$

We find that if sample size increases probability of Type I error decreases provided variance is the same.

Example 4.5:

A quality control worker is going to check a large production of drug. If the lot has 5% or fewer defectives than the lot is of acceptable quality. He took a random sample of 100 tablets of certain drug and found that the defective rate is 12%. For 1% level of significance, calculate probability of Type II error (β).

Solution:

H_0 $P = 0.05$

H_1 $P > 0.05$

Z value is 2.33 at 1% $n = 100$

Actual sample proportion (p) $= 0.12$

The calculated proportion comes out to be

$$2.33 = \frac{\hat{p} - 0.05}{\sqrt{\frac{0.05 \times 0.95}{100}}}$$

Solving this, we get:

$$\hat{p} = 0.101$$

or

$$\hat{p} = 0.05 + 2.33 \sqrt{\frac{0.05 \times 0.95}{100}} = 0.101.$$

Now

$$Z = \frac{0.101 - 0.120}{\sqrt{\frac{0.12 \times 0.88}{100}}} = -0.59$$

Therefore, the probability of type II error is

$$\beta = \mathbf{P}\,[\,\hat{p} < 0.101\,] = P\,[Z < -0.59] = 0.2776$$

The probability is about 28% that the quality worker will fail to detect that the proportion of defectives for this production is actually larger than 0.05 (5%).

Note that $1 - \beta$ is **the power of the test**, this represents the probability that null hypothesis is rejected when it is false. In the above example, power of the test will be $1 - 0.2776 = 0.7224$. There is about 72% probability that null hypothesis is rejected when null hypothesis is *false*. Note that for fixed sample size power increases as α increases and for fixed level of significance, power increases as n increases. The power of the test may be stated as:

The power of a test is the probability that the test will lead to rejection of the H_0 when, in fact, H_1 is true.

Example 4.6:
For hypothesis test $H_0 : \mu = 50.0$ against $H_1 : \mu < 50.0$ and $\alpha = 0.05$, $\sigma = 9.0$.

   (a) Calculate β if $\mu = 48.0$ and $n = 36$
   (b) Calculate β if $\mu = 48.0$ and $n = 81$

How does β behave with the sizes of samples?

Solution:

   (a) $H_0 : \mu = 50.0$
        $H_1 : \mu < 50.0$ $\sigma = 9.0$
        $\alpha = 0.05$
        $Z = -1.645$ for 95% one-tailed test.

   Sample mean $(\overline{x}) = 50.0 - 1.645 \,\frac{9}{\sqrt{36}} = 47.53$

$$Z = \frac{47.53 - 48}{\frac{9}{\sqrt{36}}} = -0.31$$

$$\beta = \mathbf{P}\,[\,\bar{x} > 47.53] = \mathbf{P}[Z > -0.31] = \mathbf{P}[Z < 0.31] = 0.6217$$

(b) $\bar{x} = 50.0 - 1.645 \dfrac{9}{\sqrt{81}} = 48.355$

$$Z = \frac{48.355 - 48}{\frac{9}{\sqrt{81}}} = 0.355$$

$$\beta = \mathbf{P}[\,\bar{x} > 48.355] = \mathbf{P}[Z > 0.355] = P(Z < -0.355) = 0.3632$$

As the size of the sample increases, β decreases.

4.2 Estimation of Sample size when Probability of Type I Error and Power of the test are known

We know that type I and Type II errors cannot be controlled simultaneously. If we try to control Type I error then type II will go up and vice versa. In Chapter 3 we described the methods of estimation of sample size by fixing the type I error and Type II error was controlled by large sample size. In medical science, sometimes we are forced to a small sample size. What we do, we fix the probability of type I error and also fix the probability of Type II error in term of Power of the test then the calculation of sample size is made. Since calculations are bit cumbersome, therefore for the convenience of the users they are given in different tables at the end of the Chapter.

4.2.1 Sample size for comparing proportions

$$n = \frac{\left\{ A\sqrt{P(a)\left[1 - P(a)\right]} + B\sqrt{P(0)\left[1 - P(0)\right]} \right\}^2}{\left[P(0) - P(a) \right]^2}, \tag{4.5}$$

where A and B are given for various level of significance. **P**(0)= present proportion, **P**(a)= anticipated proportion. Find sample size n from (4.5).

  (i)   For 5% level of significance and 90% power (two sided), A=1.96,B=1.28 (Table 4.10)

  (ii) For 1% level of significance and 90% power (two sided), A=2.58,B=1.28 (Table 4.11).

  (iii) For 5% level of significance and 90% power(one sided), A=1.645, B=1.28 (Table 4.12)

  (iv) For 1% level of significance and 90% power(one sided), A=2.58, B= 1.28 (Table 4.13)

(v)   For 5% level of significance and 80% power(two sided), A=1.96, B=0.84
      (Table 4.14)

(vi) For 1% level of significance and 80% power(two sided) A=2.33, B=0.84
      (Table 4.15)

(vii) For 5% level of significance and 80% power (one sided), A=1.645, B=0.84
      (Table 4.16)

(viii) For 1% level of significance and 80% power (one sided), A=2.33, B=0.84
      (Table 4.15)

Some more examples are given below:

Example 4.7:

An investigator wants to know the size of the sample in his study if he uses intermittent
pneumatic (IPC) to prevent Deep Venous Thrombosis (DVT) following total hip
replacement. He states that 70 patients in each group gives a probability of 80% of
detecting a 20% difference (from the estimated frequency of 10%) between the three
therapies groups when p is less than 5%. How large sample size is needed in the study in
order to detect an overall reduction from previous studies that indicate 20-50% of patients
develop DVT? It is assumed that investigator wants 80% power of detecting a decrease in
rate of DVT from 20% to 10%.

Solution:

From this example we can easily extract following information.

Test rate = 20% = P(0); anticipated rate = 10% = P(a); level of significance = 5% or 1%;
power of the test = 80% (probability of type II error is 20%). The size of sample may be
seen from the corresponding table(4.14), given at the end of the chapter. The sample sizes
are reproduced below.

Level of significance and sample sizes

	5%	1%
Two-tailed	108	165
One-tailed	83	141

Example 4.8:

The five years cure rate for a particular cancer (the proportion of patients free from
cancer five years after treatment) is reported in the literature to be 50%. An investigator
wishes to test the hypothesis that his cure rate applies in a certain local health district.
What minimum sample size would be needed if the investigator was interested in
rejecting the null hypothesis only if the true rate was less than 50% and wanted to be 90%
sure of detecting a true rate of 40% at 5% level of significance?

Solution:

True cure rate = 50%=P(0), anticipated cure rate = 40% = P(a); level of significance 5%
or 1% and power of the test =90% (probability of type II error is 10%)

The sample size for various levels of significance may be seen from the tables. The
sampling sizes for all these cases are reproduced below.

Level of significance and sample sizes

	5%	1%
Two-tailed	259	368
One-tailed	211	322

Example 4.9:

Previous surveys have demonstrated that the usual prevalence of dental caries among school children in a particular community is about 25%. How many children should be included in a new survey design to test for decrease in the prevalence of dental carries, if it is designed to be 90% sure of detecting a rate of 20% at 5% level of significance?

Solution:

Test caries rate = 25%=$P(0)$; anticipated rate = 20%=$P(a)$; power of the test = 90%; level of significance 5% or 1%. The tables are used to find the sizes of the samples:

Level of significance and sample sizes

	5%	1%
two-tailed	741	1062
one-tailed	600	926

4.2.2 Sample size for a single mean

We know that:

$$Z_\alpha = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \text{ and } Z_\beta = \frac{\bar{x} - \mu_1}{\sigma / \sqrt{n}}$$

Solving these two critical ratio for sample size n, we get:

$$n = \left[\frac{\{Z(\alpha) - Z(\beta)\}\sigma}{\mu_1 - \mu_0} \right]^2 \tag{4.6}$$

where $Z(a)$ =value against given level of significance; $Z(\beta)$ =table value against given power; μ_1 = given mean; μ_0= expected mean and σ = standard deviation. If $\sigma = 1$, then (4.6) is identical to (4.3).

Example 4.10:

Suppose the investigator wants to know whether PIMAX (maximal inspiratory mouth pressure) is the same in patients with kyphoscoliosis and in normal patients without kyphoscoliosis. Suppose the investigator wants the type I error to be 0.05 and he wants a 0.90 probability of detecting a true difference. His past experience is that the mean PIMAX is 110 cmH$_2$O in normal patients with a standard deviation of 20cm H$_2$O. Suppose the investigator wants to be able to say that mean PIMAX of 80cm H$_2$O or less in kyphoscoliosis patients is significantly different from normal. What would be the sample size to achieve this target?

Solution:
Level of significance for 5%, then $Z(\alpha = 0.05) = 1.96$. Lower tail z-value (Power) for 90%, $Z(0.10) = -1.28$. Given mean = 110; expected mean 80 and standard deviation = 20.

Then using (4.2) we get that sample size is 5.

4.2.3 Sample size for Comparing of two proportions

$$n = \frac{\left[A\sqrt{P(c)\,(1-P(c))} + B\sqrt{P(t)(1-P(t)) + P(c)\,(1-P(c))}\right]^2}{\left[P(t) - P(c)\right]^2}, \qquad (4.7)$$

where A and B are defined in (4.1), p(c) = proportion of control group and P(t) = proportion of treatment group.

Example 4.11:
A randomized trial was used to evaluate the efficacy of J5 antiserum in presenting the serious consequences of gram-negative infection. This study involved a trial J5 antiserum in surgical patients to determine whether it is effective in preventing gram-negative infections. The actual study utilized 126 patients in the treatment group and 136 in the control group. Let us suppose that an investigator prior to doing the study wants to estimate the sample size needed to detect a reduction in proportion of patients who experience shock from 10% level according to the investigator's previous experience to 5% or less if patients are given transfusions from donors treated with J5. He is willing to accept a type I error of 0.05 and wants a 90% probability of detecting a true difference. Determine the sample size under this situation for each group.

Solution:
Proportion in the control group = P(c) = 10%; proportion in the treatment group = P(t) = 5%; level of significance = 5% [1.96 = Z(a)]; power of the test = 90% [table value = 1.28 = Z(b)]. Using (4.3) we get n = 682, the sample size for each group. Suppose the sample size is large and the chances are that the investigator will compromise and recalculate the sample size with less power or a larger difference. If we take the same difference and reduce the power from 90% to 70% (table value for 70% is approximately 0.52) the sample size comes out to be 420 for each group. Again if he needs to detect a drop in the infection rate from 10% to 3% with power 70% then the sample size will be 208 for each group.

4.3 Diagnosing a Test-Statistic for Testing of Hypotheses and p-Value

4.3.1 Diagnosing a Test-Statistic

The manner in which the test-statistic is actually used depends on the parameter of interest. For example, if for large sample, we are interested to test population mean or proportion, and then the test-statistic *for both will not be the same*. If a variance is to be tested, then different test-statistic will be used. How to proceed to diagnose a test-statistic, is first to determine the parameter of interest. What the researcher needs is very important. Three steps will be useful to diagnose a test-statistic.

i) First, try to understand the objectives for which the data are collected or measurements are taken.

ii) Second, try to identify the type of variable(s), whether measurements are qualitative or quantitative in nature.

iii) Third, try to identify the parameter(s) to be tested.

Note that, if the variable is quantitative, then parameter may be either population mean or population variance and if it is qualitative, the parameter may be population proportion.

If one looks into your objectives minutely, the problems can be solved easily. If it is a written statement then there is certainly an indication, and the hypothesis can be formulated easily. Let us try to guide how to formulate the hypothesis through these examples.

There is one glass of Pepsi and another glass of Mecca-Cola and it is required to select one, which tastes best. Here experimental units are the consumers and the variable under study is qualitative. Therefore, the parameter of interest is the *proportion* of population who favor Mecca-Cola over PEPSI or vice-versa.

A dietician would like to see whether a new diet is effective in reducing weight of an obese woman. Here the experimental women will be obese women and the variable to be measured is quantitative. The dietician will be comparing mean weight before and after the completion of course.

A manufacturer of a new drug claims that his drug is more effective than the one already available in the market. Naturally the experimenter will select two groups to see the effectiveness of these two types of drugs in terms of proportions and these proportions will be compared.

4.3.2 p -Value

Since it is difficult to understand the concept of p-value for non-statisticians, therefore, some remarks on p-value is devoted in this section. We know that in testing of hypotheses we choose the level of significance beforehand. The null hypothesis is accepted if the calculated value of test-statistic is less than the corresponding value at the level of significance. If both values are equal, we say that one is in a critical situation. There is one drawback that the test be conducted in this manner. A measure of the *level of significance* of the test results is not readily available. If the value of the test-statistic falls in the rejection region, we have no measure of the extent to which the data disagree with the null hypothesis.

Consider the null hypothesis that the average weight of the university students is 68.5 kg to be tested against alternative hypothesis that the average weight is greater than 68.5 at fixed 5% level of significance. Consider the following possible values of the computed test-statistic (z-statistic)

$$Z_c = 2.01 \text{ and } Z_c = 3.87,$$

which of these values of test-statistic provides stronger evidence for the rejection of null hypothesis? How can we measure the extent of disagreement between the sample data and null hypothesis for each of the computed value?

We know that at 5% level of significance the Z-value for one-tailed test is 1.645. Both computed values are greater than 1.645 and falls in the rejection region, therefore the result in each case is statistically significant.

Note that Z-test-statistic of population mean is simply Z-score (Chapter 1, Section 1.9.6). Therefore, Z-score of 3.87 would present strong evidence *that the true mean is larger than 68.5 kg.*

One way of measuring the amount of disagreement between sample mean and the value of population mean or proportion in the null hypothesis is to calculate the probability that the *observed value* of the test-statistic equals to or greater or less than the actual computed value under null hypothesis. The disagreement between sample statistic and population parameter H_0 can be measured as:

p-value = $P[Z > Z_c]$ upper one-tailed

p-value = $P[Z < Z_c]$ lower one-tailed

p-value = $P[Z \neq |Z_c|]$ two-tailed

where Z_c is the computed value of the test-statistic. From Table 2.1 we can calculate the probability.

$P(Z \geq 2.01) = 1 - 0.9778 = 0.0222$ and

$P(Z \geq 3.87) = 1 - 0.9999 = 0.0001$

We can draw a conclusion that smaller the probability (p-value), greater is the extent of disagreement between sample statistic and population parameter (mean or proportion). Note that the p-value for the two-tailed test is twice the p-value of one-tailed test.

Thus we can say that p-value is the maximum probability of rejecting the null hypothesis, when null hypothesis is true. Some statisticians referred to p-value as the observed level of significance of the test under consideration. In fact, for computer it is easy to calculate p-value but it takes much longer time to calculate the test-statistic value for a given α-value.

In most of the medical journals, dissertations and technical reports test-statistics and p-values associated with the tests are mentioned and it is left to the research workers to draw conclusions whether to accept or not to accept the null hypothesis.

There are two advantages of reporting the results in the form of *test-statistic* and p-*value*:

a) Most software packages (like SPSS or SAS) present a p-value. This makes it easy for the researcher to decide whether to accept or not to accept the null hypotheses.

b) Researchers are allowed to select the maximum value of the level of significance that they would be willing to tolerate in carrying out standard tests of hypothesis.

One should follow two points to decide whether to accept the null hypothesis or not, when the results are presented in the form of p-values.

i)   Choose the maximum value of the level of significance (1%, 5%, 10%, ...) that one is willing to tolerate.

ii) If p-value of the test is less than the stated α-value (given level of significance) then do not accept the null hypothesis otherwise accept the null hypothesis.
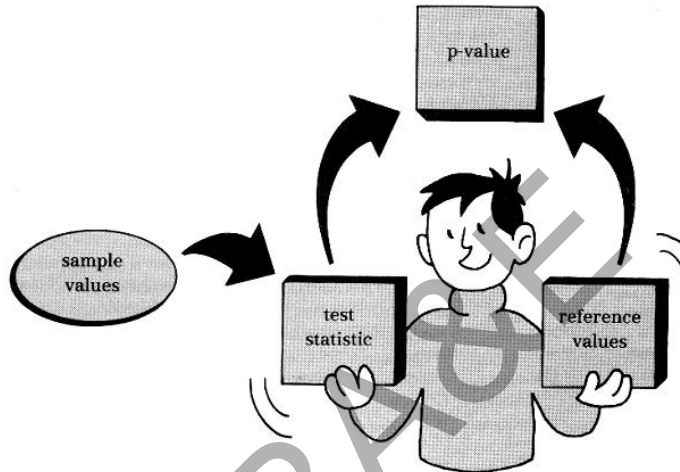
Fig. 4.4: Performing a hypothesis test

4.4 General Procedure of Testing of Hypothesis

There are several steps in testing of hypothesis, which lead to a conclusion to accept or not to accept the hypothesis. These steps are common for all types of tests of significance. These general steps lead us to the final decision about the null hypothesis.

Step 1:   Write two statements, which are appropriate concerning value of the parameter i.e. to state null and alternative hypotheses.

Step 2:   State whether the test is a one-tailed or a two-tailed test.

Step 3:   Choose the level of significance. Usually 1% or 5% level of significance is chosen.

Step 4:   State an appropriate test-statistic to be used.

Step 5:   Calculate the value using the test-statistic mentioned in Step 4.

Step 6:   State the decision rule for the acceptance of null hypothesis. The decision rule is to accept the null-hypothesis if calculated value is less (larger) than table value at a given level of significance otherwise do not accept the null hypothesis.

Health scientists usually interpret the result in terms of p-value (observed level of significance). If the observed p-value is less than the stated p-value (given level of significance), then the null hypothesis is not accepted.

Step 7:   Draw the inference about the parameter on the basis of the above steps.

All these steps are given in the flow chart Fig. 4.5 (next page)

4.5 Tests of Significance

The following tests of significance will be discussed in this section.

Tests for mean and proportion	Tests for variance
Z-test	χ^2-test
t-test	F-test

If the condition of normality is satisfied we use parametric tests. If the responses are distribution free then we use non-parametric tests (non-parametric tests will be discussed in Chapter 8). The lay out for the tests of means and proportions (Z and t) on next page:

Layout of the Test of Significance - I

Test for Means and Proportions
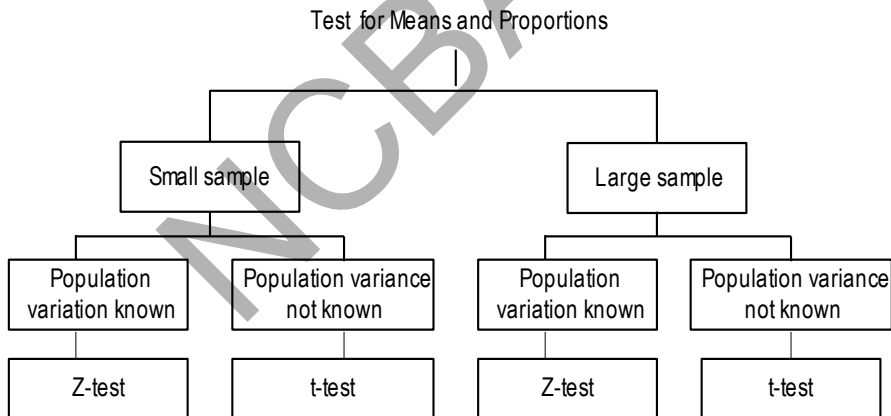
Steps of General Procedure of the testing of Hypothesis

1
| Assumptions(s) |

2
| Write null and alternative hypotheses |

3
| One tailed test or two tailed test |

4
| Level of significance |

5
| Test-statistic to be used |

6
| Calculate p-value |

7
| Comparison of observed p-value (table value) and calculated p-value |

8
| If calculated p-value is less than observed p-value do not accept the hypothesis otherwise accept |

9
| Conclusion(s) and Interpretation |
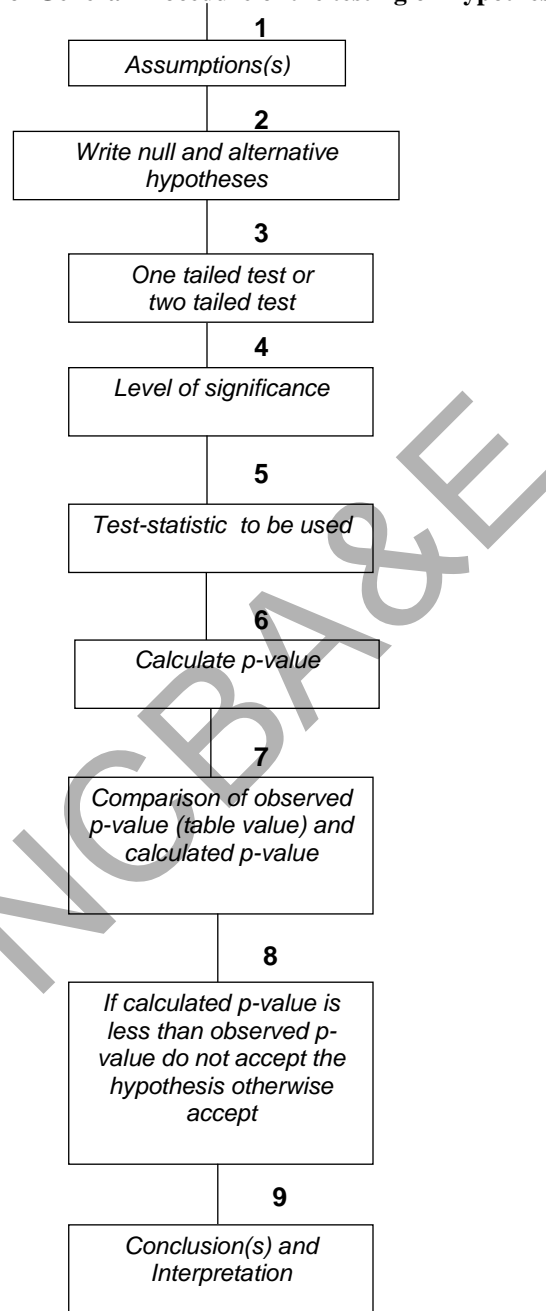
Fig. 4.5: Flow chart of the testing of hypotheses

Layout of the test of significance II

Test for Variance
(for large and small samples)

Single sample	Two samples
Chi-square	F-test
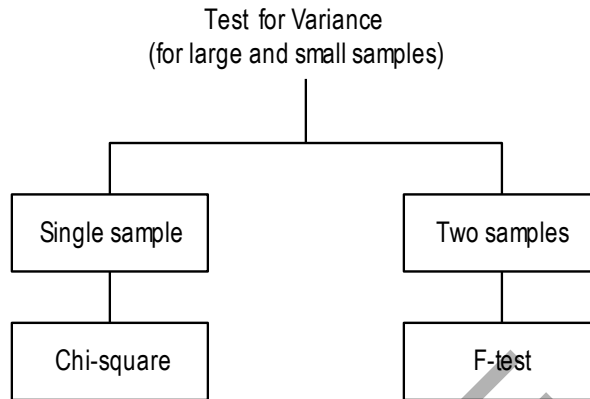
Fig. 4.7: Layout plan of the test of significance II

Before the application of t-test, test of homogeneity (equality of variance) is applied, if this condition is satisfied, t-test is used otherwise non-parametric tests or some other alternatives are used. Five tests are available to test the homogeneity of samples. These are:

(a) Bartlett's test (1936)

(b) F-test

(c) Levene's test (1962)

(d) Cochran's test (1962)

(e) Samiuddin-Hanif-Asad cube root test (1978)

Cochran's test is a special test as it is applicable for equal number of observations in the samples. Only Levene's test of homogeneity is available in SPSS package, therefore, we stick to it. Note that in EPI-INFO package, Bartlett's test is available. Samiuddin-Hanif-Asad test is very simple to calculate and more or less identical to Bartlett's test. When t-test is used the SPSS package automatically test the homogeneity (equality) of variance. The flow chart (layout) of tests of significance for parametric and non-parametric situations follows (Fig. 4.8).

Fig. 4.8: Flow chart from tests of significance.

Layout of Test of Significance-III
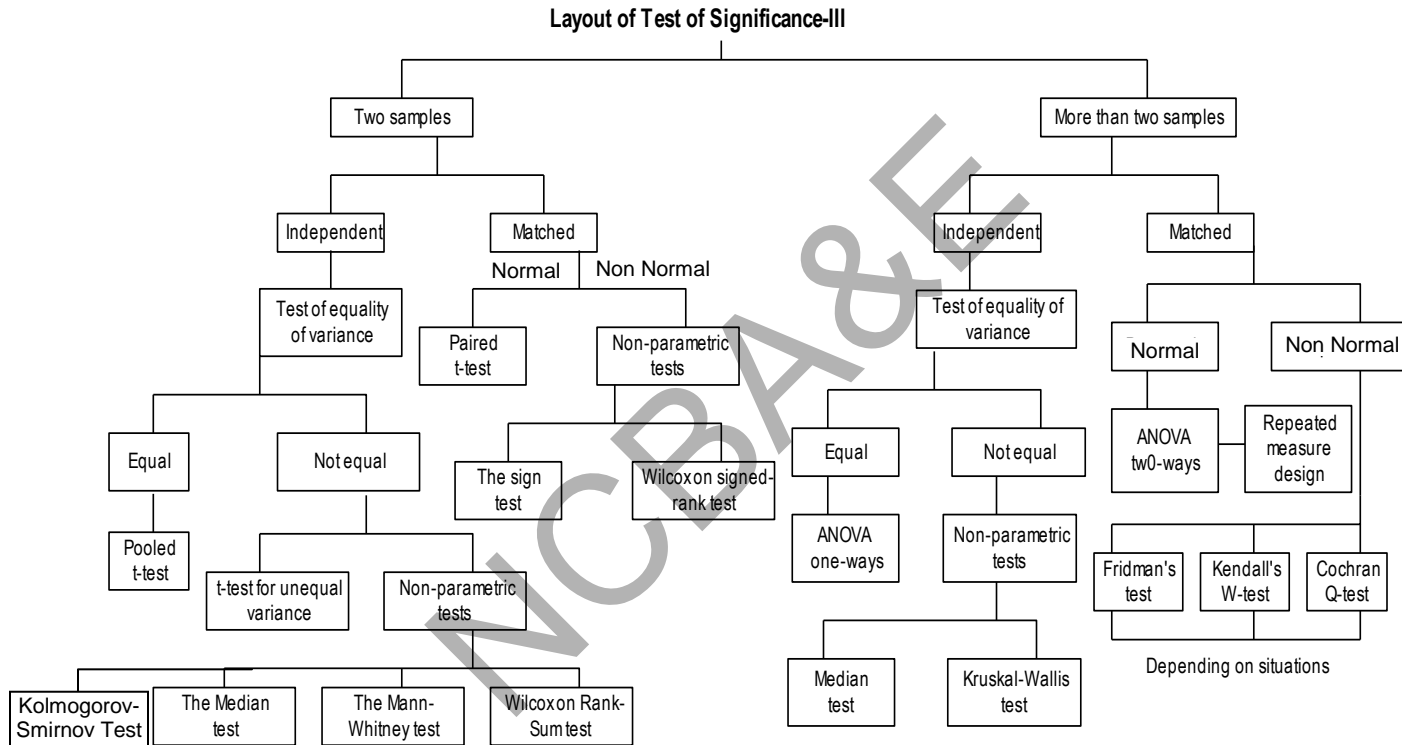
If t-test for unequal variance is there in computer printput, it is always advised to choose this instead of non-parametric tests.

4.5.1 Z-Test for one and two samples for means and proportions

This test is used to test mean and proportion for one sample and to test the difference between two sample means and proportions. Followings are the assumptions and conditions to apply Z-test.

(i)   Sampled population should be normal.

(ii)   Sample must be random.

(iii)   Sample size is large and population variance is known. If population variance is not known, sample variance may be used when sample size is large.

(iv)   If sample size is small and/or population variance is known, this test is also applicable.

(v)   Samples must be independent.

Since in practice population variance is never known, we always use either t-test or its equivalent non-parametric test as the circumstances occur.

(i) *Z-test for one sample mean*

This is used to test whether a given sample has been selected from the population whose mean and variance are known. Since sample mean ($\bar{x}$) is representative of the population mean (μ), we find the difference between sample mean and population mean. If there is no difference, we say that the sample has been selected from the population whose mean and variance are given.

Some examples for Z-test are given. The purpose of these examples is to demonstrate how Z-test is used to test the mean and proportion. Later on it will be demonstrated how SPSS package is used to solve the problems.

Example 4.12:
Family and Community Medicine Department feels through a study that patients in an area spend on the average 12 minutes with the doctors in the Family Care Centers. Ministry of Health feels that doctors should spend more time with the patients. For this, the Ministry took a random sample of 50 patients from the Family Care Centers of the area and found that doctors are spending on the average 13.6 minutes. The population standard deviation is 8.2 minutes. Use 5% level of significance to test that doctors are spending on the average more than 12 minutes with the patient.

Solution:

(1) $H_0 : \mu = 12$ minutes $\bar{x} = 13.6$, $\sigma = 8.2$

   $H_1 : \mu > 12$ minutes

(2) $\alpha = 0.05$ and $n = 50$

   Since sample is large, Z-test is used.

(3) Test-statistic: $Z_c = \left\{ (\bar{x} - \mu)\sqrt{n} \right\} / \sigma$,   (4.8)

where:

$\bar{x}$ = Sample mean μ = population mean

σ = Population standard deviation   n = sample size

$$z_c = \frac{(13.6 - 12)\sqrt{50}}{8.2} = 1.38$$

(4) Since it is a one-tail test, the Z-value for 5% level of significance is 1.645.

(5) The calculated value is 1.38 which is less than table value, therefore, the result is non-significant and the null hypothesis is not rejected, we say with 95% confidence that the study conducted by the Family and Community Department shows that doctors are spending on the average 12 minutes with the patients. This conclusion may also be shown through p-value.

Stated p-value	Observed p-value
	$\mathbf{P}[Z \geq 1.38]$
0.05	= 1 - 0.9162 = 0.0838

Since observed p-value is more than stated p-value, it falls in the acceptance region; therefore, we are 95% confident that the conclusion is correct.

The virtue of the p-value in computation is, that one can simply report the p-value and different workers can make their decisions.

95% confidence limits may be calculated as:

$$13.6 \pm 1.645 \ \frac{8.2}{\sqrt{50}} \ \text{or} \ (11.692 \ , \ 15.508)$$

We say with 95% confidence that these two limits contain population mean (which in this case is 12). Since these limits do not contain zero, therefore, we can also say that there is significance difference between sample and population means. Note that in practice population mean or proportion is never known to us. That is why, we construct confidence limits to see the location of the population mean or proportion (see Chapter 3).

Example 4.13:
An article published in Medical Journal where it was claimed that by better nutrition the mean weight of adult women in USA had increased to 79.5 kg. The authority of weight control felt that the figure was too high for the females. A sample of 45 women was taken and found that average weight was 76.6 kg with standard deviation 11.7. Perform a test that $H_0 : \mu = 79.5$ against $\mu < 79.5$ at 5% level of significance and give interpretation about conclusion.

Solution:

(1) $H_0 : \mu = 79.5$ $\bar{X} = 76.6,$

   $H_1 : \mu < 79.5$ $s = 11.7,$

(2) $\alpha = 0.05$ $n = 45$

Since sample is large, Z-test is applied.

(3) test-statistic: $Z_c = \dfrac{76.6 - 79.5}{\dfrac{11.7}{\sqrt{45}}} = -1.663$

(4) Since it is a one-tailed test, the Z-value is -1.645 at 5% level of significance.

(5) The calculated absolute value of Z_C is more than the table value of Z, therefore, the result is significant and the null hypothesis is not accepted. We say with 95% confidence level that the average weight of the women is less than 79.5 kg. Conclusion may also be drawn by the use of p-value as:

Stated p-value	Observed p-value
0.05	$P[Z \le -1.66] = 0.0485$

Since the observed p-value is less than the stated p-value, the statistic value falls in the rejection region.

95% confidence limits may be calculated as:

$$76.6 \pm 1.645 \ \dfrac{11.7}{\sqrt{45}} \ , \text{ or } \ [73.73, \ 79.47]$$

We are 95% confident that these two limits do not contain the average weight of the women. The average weight of 95% women would not lie in (73.73, 79.47).

(ii) Z-test for one sample proportion

Example 4.14:
It was reported in the Journal of the American Geriatric Society (1990) that hospital patients over the age of 65 apparently face high risk of serious treatment errors. The records of 122 elderly patients were randomly selected and 30 out of them found to have at least one erroneously prescribed medication. (They received unneeded drug or they failed to receive necessary drug). The researcher did not expect such a high rate. Test at 5% level of significance that the true proportion of elderly patients who have at least one erroneously prescribed drug exceeds 20%.

Solution:

(1) $H_0 : P = 0.20$ $\hat{p} = \dfrac{30}{122} = 0.246,$

$\quad H_1 : P > 0.20$ $n = 122$

(2) $\alpha = 0.05$

Since proportion is to be tested and sample is large, therefore, Z-test for the testing of proportion will be used.

(3) test-statistic: $Z_c = \dfrac{\left|\hat{p} - P\right|}{\sqrt{\dfrac{P(1 - P)}{n}}}$ \qquad\qquad (4.9)

where: $\hat{p}$ = sample proportion   P = population proportion

Note that the denominator of the Z-statistic contains the population proportion.

$Z_c = \dfrac{\left|0.246 - 0.20\right|}{\sqrt{\dfrac{0.20 \times 0.80}{122}}} = 1.27$

(4)   Since it is a one-tailed test, Z-value is 1.645 at 5% level of significance.

(5) The calculated value of Z_c is less than the table value of Z at 5% level of significance. Therefore, the result is non-significant and the null hypothesis is accepted. We can say with 95% confidence that the true proportion of elderly patients who received at least one erroneously prescribed drug does not exceed 20%.

The p-value will be calculated as:

$P[Z > 1.27] = 1 - 0.8980 = 0.1020$ (observed level of significance)

Since observed p-value is more than stated p-value (0.05), therefore, the null hypothesis is accepted. The 95% confidence limits are

$0.246 \pm 1.645 \sqrt{\dfrac{0.2 \times 0.8}{122}}$ or [0.186 , 0.306].

The proportions of elderly patients who have at least one erroneously prescribed drug vary from 0.186 to 0.306. Since the value of H_0 lies inside the interval, 0.186 to 0.306, the null hypothesis is accepted.

Example 4.15:

Prior to the Polio immunization program in the Eastern Province of Saudi Arabia, a survey revealed that 180 out of a random sample of 400 elementary school children have been immunized against Polio. Can we say at 5% level of significance that 50% of the elementary school children in this area had been immunized?

Solution:

(1) $H_0 : P = 0.50$ \qquad $\hat{p} = \dfrac{180}{400} = 0.45$

$H_1 : P \neq 0.50$ \qquad $n = 400$

(2) $\alpha = 0.05$

Since sample is large, Z-test will be used.

(3) test-statistic: $Z_c = \dfrac{|\, 0.45 - 0.50\,|}{\sqrt{\dfrac{0.50 \times 0.50}{400}}} = 2.0$

(4) Since it is a two-tailed test, Z-value at 5% level of significance is 1.96.

(5) The calculated value of Z_c is more than the Z-value of the table, it falls in the rejection region. Therefore, it is significant. We may say with 95% confidence that the null hypothesis is not accepted and say that 50% of the children were not immunized.

Conclusion may also be drawn by the use of p-value as:

Stated p-value	Observed p-value
0.05	**2 P** $[Z \leq |2.0|]$ $0.0228 + 0.0228 = 0.0456$

Since observed p-value is less than stated p-value, therefore, it falls in the rejection region and null hypothesis is rejected.

The 95% confidence limits will be

$0.45 \pm 1.96 \sqrt{\dfrac{0.5 \times 0.5}{400}}$ or $[0.401\ ,\ 0.499]$

Since the interval does not contain 0.5, there is significance difference.

(iii) Z-test for two samples (means)

In case of Z - test for two samples, two random samples are selected independently, one from each population (case- control study) and the main purpose is to see whether two populations are different or not. Since samples are representative of two populations, we compare two sample means to compare two populations.

Example 4.16:

A study was conducted to compare percentage of body fat for rural and urban college male students. For this purpose, two random samples one from each area were selected. The percentage of body fat for each sample was measured. Can we say at 5% level of significance that there is no difference in body fat in two groups? The data are given as:

	Urban	Rural
Sample	193	188
Mean	12.07	11.04
s.d	3.04	2.63

(American Journal of Physical Anthropology, 1993, Vol. 54, pp. 119-112)

Solution:

(1) $H_0 : \mu_1 = \mu_2$ (There is no difference between population means)

   $H_1 : \mu_1 \neq \mu_2$ (There is difference between population means)

(2) $\alpha = 0.05$

Since two samples are given and sample size is large, then Z-test can be used to test the difference between means of two samples.

(3) test-statistic: $Z_c = \dfrac{\left|\overline{x}_1 - \overline{x}_2\right|}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ (4.10)

where: $\overline{x}_1$ = mean of first sample, $\overline{x}_2$ = mean of second sample

   s_1^2 = variance of first sample, s_2^2 = variance of second sample

   n_1 = size of first sample, n_2= size of second sample

   $Z_c = \dfrac{\left|12.07 - 11.04\right|}{\sqrt{\dfrac{(3.04)^2}{193} + \dfrac{(2.63)^2}{188}}}$ $= 3.54$

(4) Since it is a two-tailed test, Z-value at 5% level of significance is ± 1.96.

(5) The calculated value of Z_c is greater than the Z-value of the Table; it falls in the rejection region. Therefore, it is significant. We say with 95% confidence that mean fat of two groups is different, i.e. two samples are different and consequently two populations are different.

Since mean of the urban group is higher than rural group, therefore, we say that the average fat in urban males is more as compared to that of rural males.

p-value may be calculated as:

Stated p-value	Observed p-value
	$2P\ [Z \geq |3.56|]$
0.05	$0.0002 + 0.0002 = 0.0004$

The 95% confidence limits are

$$(12.07 - 11.04) \pm 1.96 \sqrt{\dfrac{(3.04)^2}{193} + \dfrac{(2.63)^2}{188}} , \quad (0.46 , 1.6)$$

So we can say with 95% confidence that these limits contain the difference of two population means from which these samples are selected. Since these limits do not

contain zero, therefore, there is significance difference between two samples as we should expect.

(iv) Z-test for two sample proportions

Example 4.17:

An epidemiologist compared a sample of 100 adult cases that were suffering from certain diseases with a sample of 120 controls (free from diseases). It was found that 69 of the diseased and 80 of the controls were employed in subsistence occupations. Can the epidemiologist say on the basis of this information at 5% level of significance that two population proportions differ with respect to the proportion employed in subsistence occupations?

Solution:

(1) $H_0 : P_1 = P_2$ $\hat{p}_1$ (diseased) $= \dfrac{69}{100} = 0.69$

$\quad\;\; H_1 : P_1 \neq P_2$ $\hat{p}_2$ (controls) $= \dfrac{80}{120} = 0.67$

(2) $\alpha = 0.05$

Since sample size is large, Z-test for proportion is used.

(3) test-statistic

$\quad$ c: $Z_C = \dfrac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\dfrac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}}$ (4.11)

where: $\hat{p}_1 =$ proportion of first sample and $\hat{p}_2 =$ proportion of second sample

$\quad Z_c = \dfrac{\left|.69 - .67\right|}{\sqrt{\dfrac{(.69)(.31)}{100} + \dfrac{(.67)\,(.33)}{120}}} = \dfrac{.02}{.102} = 0.196$

(4) Since it is a two-tailed test, Z-value at 5 percent level is 1.96.

(5) The calculated value is far less than the table value, the result is non-significant, we say with 95% confidence that there is no difference between two groups.

4.5.2 t-test for single and two samples

This is known as Student's t-distribution or t-test and was discovered by British Chemist, W.S. Gosset. He published his work under the pseudo-name *Student* in 1908. When sample size is small, t-test is applied. It is also used for one sample and two samples to test the mean and proportion like Z-test. Followings are the assumptions and conditions for the application of t-test.

(i)   Sampled population should be normal.

(ii) The sample must be random, so that the observations are independently distributed.

(iii) Sample is small and population variance is not known, it can also be applied when sample is large and population variance is not known. For large sample Z-test and t-test are almost identical. (We have seen that in all the statistical packages only t-test is given).

(iv) In case of two samples, it is generally assumed that population variances are equal and samples are independent.

(i) t-test for one sample mean

Example 4.18:

A new brand of oatmeal cereal claims that a 1.5-ounce serving of the cereal has 140 calories. The staff of the laboratory analyzed the 12 different servings of 1.5-ounces each. The result yielded the mean equal to 153 calories with standard deviation of 21 calories. Can the company's claim of 140 calories be rejected based on the data collected? Use 1% level of significance.

Solution:

(1) $H_0 : \mu = 140$ calories $\bar{X} = 153$ calories

   $H_1 : \mu \neq 140$ calories $n = 12$ $s = 21$ calories

(2) $\alpha = 0.01$

since the sample is small and also population standard deviation is not known, therefore, t-test is to be used.

(3) test-statistic: $t_c = \dfrac{|\bar{x} - \mu|}{s} \sqrt{n}$ (4.12)

$$= \frac{|153 - 140|}{21/\sqrt{12}} = 2.144$$

(4) Since the sample size is small, we will see the t-table (Table 3.19). How to see the table? Since it is a two-tailed test, divide 0.01 by 2. We will get 0.005. Subtract 0.005 from 1 which gives 0.995. Now see the table under 0.995 against $(12 - 1) = 11$ degrees of freedom. This gives 3.1058. (This was explained in Chapter 3 as well).

(5) Since our calculated value is less than table value, therefore, the result is non-significant and we do not reject the null hypothesis. We conclude that the sample mean calories is not different from the population mean. The 95% confidence limits are:

   $153 \pm 3.1058 \, [21/\sqrt{12}\,]$, or [134.17, 171.83]

Note that the H_0 value lies in the interval.

Example 4.19:

A series of 10 blood tests were run on a particular patient over several days. The variable monitored in the total protein level. Since the blood protein level should be neither too large nor too small, it is desirable to detect either situation $\mu = 7.25$ or $\mu \neq 7.25$ based on a sample of size 10. The sample values are,

7.23, 7.24, 7.25, 7.28, 7.31, 7.29, 7.32, 7.26, 7.27, 7.24

Test at 5% level of significance whether population mean is 7.25,

1- H_0: $\mu = 7.25$

2- H_1: $\mu \neq 7.25$

3- $\alpha = 5\%$

4- Test Statistic: t-test for single sample.

Solution:

(1) $H_0 : \mu = 7.25$

$H_1 : \mu \neq 7.25$

(2) By simple calculation, we get:

$\bar{X} = 7.269$,   s = 0.0307

(3) $\alpha = 0.05$, n=10

since the sample is small and also population standard deviation is not known, therefore, t-test is be used.

(4) test-statistic: $t_c = \dfrac{|\bar{x} - \mu|}{s} \sqrt{n} = \dfrac{|7.269 - 7.25|}{0.0307 / \sqrt{10}} = 1.956$

(5) Since the sample size is small, we will see the t-table (Table 3.19). Since it is a two-tailed test, divide 0.05 by 2. We will get 0.025. Subtract 0.025 from 1 which gives 0.975. Now see the table under 0.975 against (10 - 1) = 9 degree of freedom. This gives 2.2622.

(6) Since our calculated value is less than table value, therefore, the result is non-significant and we do not reject the null hypothesis. We conclude that the sample mean value is not different from the population mean, therefore the hypothesis is not rejected and one can say with 95% confidence level that on the average blood protein level is not different than 7.25

This example can be solved by using IBM-SPSS package as follows:

Example S4-1

To test for the mean using IBM-SPSS, for the data in example 4.19, we enter the data and follow the following steps:

Analyze➜Compare Means➜One Sample T-Test:
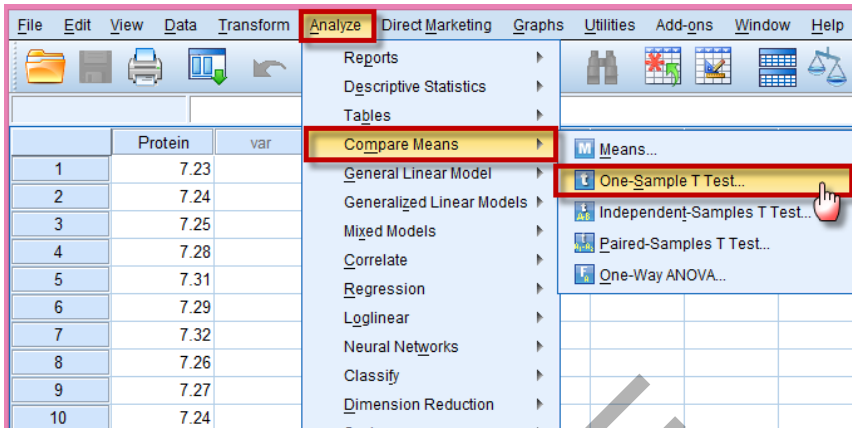
We move the variable into Test Variable(s) and change the Test Value from 0 to 7.25, as follows:
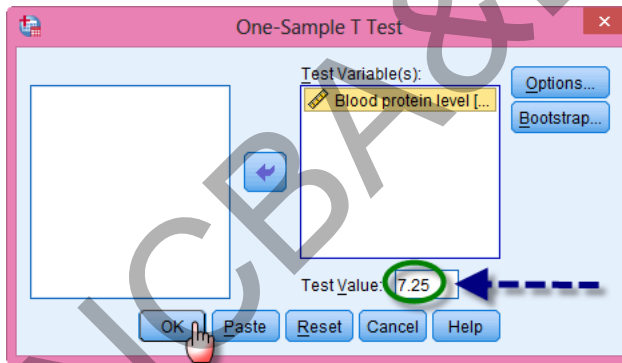
Once we click on [OK], we get the following output:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Blood protein level	10	7.2690	.03071	.00971

One-Sample Test

	Test Value = 7.25					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Blood protein level	1.956	9	.082	.01900	-.0030	.0410

P-Value

Since Sig. (2-tailed)- p-value = .082 which is greater than 0.05, therefore hypothesis is not rejected and one can say with 95% confidence level that on the average blood protein level is not different than 7.25.

(ii) t-test for one sample proportion

t-test is also applied for test of sample proportions. This has been explained in the following example.

Example 4.20:

A report claims that at least one-half of the patients with back pain who receive acupuncture treatments obtain relief. The doctors at a major hospital in New York City feel that the estimate of 0.50 is too high. They check the records of 25 patients at their hospital that received similar treatment for back pain. If 12 of these patients got relief, can figure of 0.50 be rejected as too high for patients at this hospital? Use 5% level of significance.

Solution:

(1) $H_0 : p = 0.50 \qquad \hat{p} = .47$

   $H_1 : p < 0.50 \qquad n = 225$

(2) $\alpha = 0.05$ (one tailed test)

(3) Test-statistic: $Z = \dfrac{|\hat{p} - P|}{\sqrt{\dfrac{P(1-P)}{n}}}$ \hfill (4.13)

(4) $Z_{cal} = \dfrac{|0.47 - 0.50|}{\sqrt{\dfrac{0.50\,(1-0.50)}{25}}} = \dfrac{-0.03}{0.01} = -3$

(5) The table value for 95% confidence level is 1.645. The calculated value is more than the table value; therefore, we do not accept H_0 and say that 50% of the patients receiving the treatment are not getting relief.

This example can be solved be using IBM-SPSS package as follows:

Example S4-2

To test for the proportion using IBM-SPSS, for the information in example 4.20, for the alternative $H_1 : \mu \neq 0.50$ we enter the data (twelve 1's and thirteen 0's) and follow the following steps:

Analyze➜Nonparametric➜Legacy Dialog ➜ Binomial:
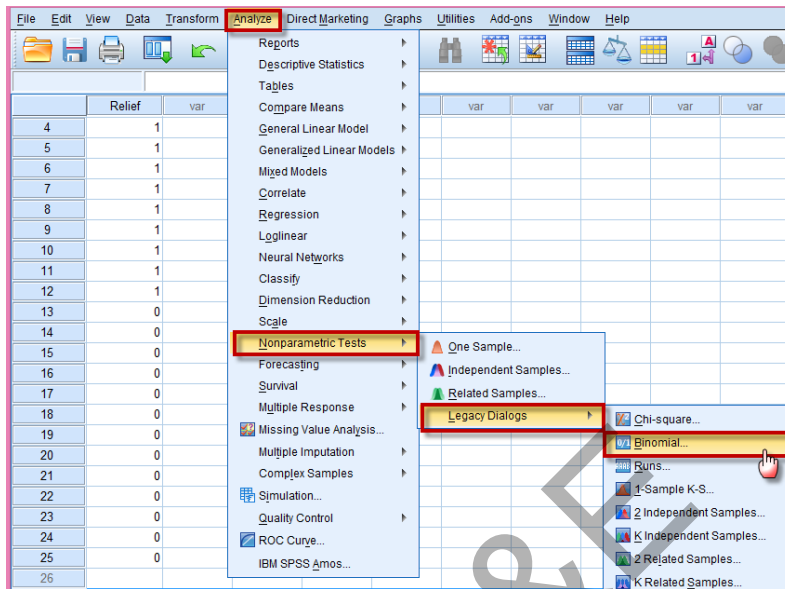
We move the variable into Test Variable List and be sure that the Test Proportion is 0.50, as follows:
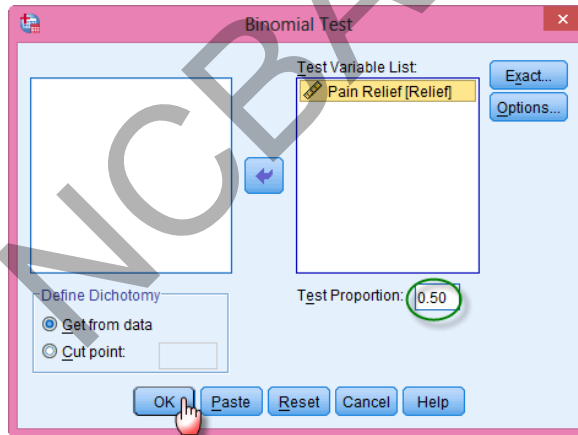
Once we click on OK, we get the following output:

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Pain Relief	Group 1	yes	12	.48	.50	1.000
	Group 2	No	13	.52		
	Total		25	1.00		

P-value

Since the Exact Sig. (2-tailed)- p-value = 1.000 which is greater than 0.05, therefore the hypothesis $H_0 : \mu = 0.50$ is not rejected.

(iii) t-test for two sample means

t-test may be used to test the difference of two population means as:

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{4.14}$$

$$\text{where: } s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{4.15}$$

is known as pooled standard deviation. We assume that variances are the same.

Example 4.21:

A study is conducted to compare the performances of two groups of non-handicapped children. One group is selected from those non-handicapped children who are studying with handicapped children and one group is selected from non-handicapped children studying in normal school. Each group contains 16 children. A test of skill development is administered to them the result is given as:

	Children in handicapped school	Children in non-handicapped school
Sample size	16	16
Mean score	122.69	124.85
s.d	10.50	10.50

Can we conclude at 5% level of significance that there is no difference between the mean scores of two groups? (Journal of Exceptional Children, Vol. 51(1), pp. 41-48).

Solution:

(1) $H_0 : \mu_1 = \mu_2$

    $H_1 : \mu_1 \neq \mu_2$

(2) $\alpha = 0.05$

Since the sample in each group is small, t-test is applied. In the application of t-test, it is assumed that the variances of the two populations are same. The pooled variance is:

$$s_p^2 = \frac{(16-1)(10.5)^2 + (16-1)(10.5)^2}{16+16-2} = 10.5$$

$$s_p = 3.240$$

(3) Test-statistic: $t = \dfrac{|122.69 - 124.85|}{3.240\sqrt{\dfrac{1}{16} + \dfrac{1}{16}}} = \dfrac{2.16}{1.146} = 1.88$

(4) $n_1 = 16$ and $n_2 = 16$, d.f. $= n_1 + n_2 - 2 = 16 + 16 - 2 = 30$. The table value is 2.0423.

(5) Since our calculated value is less than the table value, the result is insignificant and we do not reject the null hypothesis and say with 95% confidence that the performance of two groups is the same.

The 95% confidence limits are

(122.69 - 124.85) ± 2.0423 x 1.146, [-4.4818, 0.1618]

Example 4.22:

The objective of the study was to see whether the risk of coronary heart disease (CHD) could be reduced by an increased consumption of fish. For this purpose, two groups of men were selected, one consisting of 159 men who did not use the fish and other consisting of 79 men who were using more than 45 gram fish per day. After 25 years, their level of dietary cholesterol (one of the risk factors for coronary disease) present in each was recorded. The mean levels of dietary cholesterol along with the standard deviation for each group are given below. Test at 5% level of significance whether consumption of fish has real effect on the level of dietary cholesterol?

	Consumption of fish	Consumption of fish
Sample size	159	79
mean	146	158
s.d	66	75

(Source: New England Journal of Medicine, Vol. 312, pp. 1205-1209, 1985)

Solution:

(1) $H_0 : \mu_1 = \mu_2$
     $H_1 : \mu_1 \neq \mu_2$

(2) $\alpha = 0.05$

Since the sample is large, therefore Z-test should be applied. (Here we will use Z-test and t-test and see how much these two differ when the size of the sample is large. It has been stated before that for large sample Z-test and t-test are almost identical)

(3) (a) test-statistic: $Z_c = \dfrac{|146 - 158|}{\sqrt{\dfrac{(66)^2}{159} + \dfrac{(75)^2}{79}}} = \dfrac{12}{9.930} = 1.208$

(b) test-statistic: $t_c = \dfrac{|146 - 158|}{69.104\sqrt{\dfrac{1}{159} + \dfrac{1}{79}}} = \dfrac{12}{9.536} = 1.258$

(4) Since the sample is large and it is a two-tailed test, the table value will be 1.96 for 5% level of significance. (Even t-table gives the same value).

(5) Our calculated values under both test-statistic are less than the table value. Therefore, under both tests the result is non-significant and we accept the hypothesis and say with 95% confidence that on the average, there is no difference in the level of dietary cholesterol in both the groups. We say eating fish has no effect in reducing the risk of coronary heart disease.

(6) If Z-statistic and t-statistic give different values where as in one case it is rejected and in other case it is accepted, then if sample size is small, we make decision on the basis of t-test.

The 95% confidence limits are

(146-158) ± 1.96 × 9.536, [-30.69 6.690]. These two limits, contain the difference of two population means, therefore we accept the hypothesis.

(iv) t-test for testing two sample proportions

t-test for testing the difference of two proportions can be used as:

$$t_c = \frac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{p_c(1 - p_c)}\ \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{4.16}$$

where: p_c(pooled proportions) $= \dfrac{x_1 + x_2}{n_1 + n_2}$ $\qquad$ (4.17)

where x_1 and x_2 are the number of cases from the total in favor of certain characteristics.

Example 4.23:

Two preparations of drug, presented in the same table form are tested for their efficacy in alleviating headache. Preparation A is given to 25 patients, 17 claiming it effective, while B has been given to 20 patients, 16 claiming it effective. Does this provide evidence of a difference between A and B? Use 5% level of significance.

Solution:

(1) $H_0 : P_1 = P_2$ $\qquad$ $\hat{p}_1 = \dfrac{17}{25} = 0.68$

$\quad$ $H_1 : P_1 \neq P_2$ $\qquad$ $\hat{p}_2 = \dfrac{16}{20} = 0.80$

(2) $\alpha = 0.05$

Samples are small, and difference between two proportions is to be tested, therefore, t-test for proportions is used. The pooled proportion is,

$$p_c(\text{pooled proportion}) \;=\; \frac{17+16}{25+20} \;=\; 0.733$$

(3) test-statistic $\;t_c \;=\; \dfrac{|0.68-0.80|}{\sqrt{0.733(1-0.733)}\sqrt{\dfrac{1}{25}+\dfrac{1}{20}}} \;=\; 0.905$

(4) Since it is a two-tailed test, therefore, for 5% level of significance, we see the table value under $t_{0.975}$ and against $(25 - 1 + 20 - 1) = 43$ degrees of freedom. The table value is 2.023.

(5) The calculated value is less than the table value. It falls in the acceptance region, the result is non-significant, and we therefore, do not reject the null hypothesis and say with 95% confidence that there is no difference in the preparation of A and B.

p-value for the small-sample tests are computed in the same way as those for large sample test. Since SPSS package automatically gives p-value for two-tailed test.

95% confidence limits for the difference of proportions will be:

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm t_{1-\alpha/2} \times \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \qquad (4.18)$$

$$(0.68 - 0.80) \pm 2.023 \sqrt{\frac{0.733(1-0.733)}{25} + \frac{0.733(1-0.733)}{20}} = [-0.37, 0.13]$$

4.5.3 Application of SPSS package

If we have two groups and the two groups are Independent's we have to use

Analyze→Compare Means→Independent Sample T-Test…

How to do the test:

1- Move the variable to be tested to $\boxed{\text{Test Variable(s):}}$ and its scale variable $\diagup$ Scale

2- Move the variable Which determines the two groups to $\boxed{\text{Grouping Variable:}}$ and it have to be 🔹 Nominal or ▫ Ordinal

3- Definition of the two groups using $\boxed{\text{Define Groups...}}$ and we use number 1 as definition for group 1 and number 2 as definition for group 2 then click on $\boxed{\text{Continue}}$ and $\boxed{\text{OK}}$

Example S4-3

Two random samples each of 50 children were selected from two different populations. Population A had iron deficiency anemia while population B have healthy children in the same age group as population A. The hemoglobin (Hb) measurements was collected for

each child. Can we say at 5% level of significance that mean Hb is different in the two populations? The data is given in Table 4.5 taken from (Daniel, Biostat - 1991):

Table 4.5:
Healthy and anemic children
Sample 1
Children with iron deficiency anemia

9.6	2.2	3.6	5.5	3.9	5.5
4.9	5.5	8.7	7.9	6.0	3.5
7.6	9.4	9.1	11.9	6.4	9.9
6.9	7.8	6.6	7.8	4.8	10.2
2.6	5.2	6.9	4.7	6.7	8.4
10.5	3.7	5.3	7.5	4.7	6.2
3.3	6.9	6.9	5.9	10.6	4.3
7.4	4.2	7.1	6.9	6.7	8.0
				7.4	5.8

Sample 2
Healthy Children

14.6	12.7	10.1	11.8	13.2	12.5
14.1	12.9	14.0	15.2	13.4	14.6
11.6	12.6	13.4	13.3	14.6	13.0
16.0	10.6	10.5	13.4	11.3	11.8
10.3	14.1	10.2	14.9	9.6	11.9
14.5	14.4	12.3	9.9	14.0	15.6
14.6	13.1	14.1	10.6	15.2	14.3
12.7	13.9	12.3	11.4	13.9	13.5
				10.5	13.7

Is there a difference between the Children with iron deficiency anemia and Healthy Children in the proportion of hemoglobin in the blood?:

Solution:
To test this Hypothesis we follow the following steps :

Analyze→Compare Means→Independent Sample T-Test…
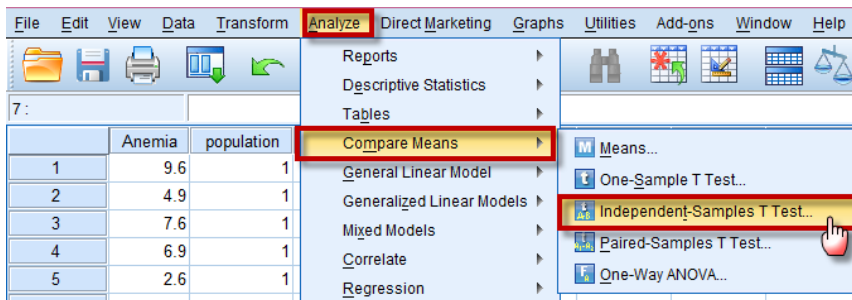
Move the variable Anemia to Test Variable(s):

Move the variable population to Grouping Variable:

Then click on Define Groups...
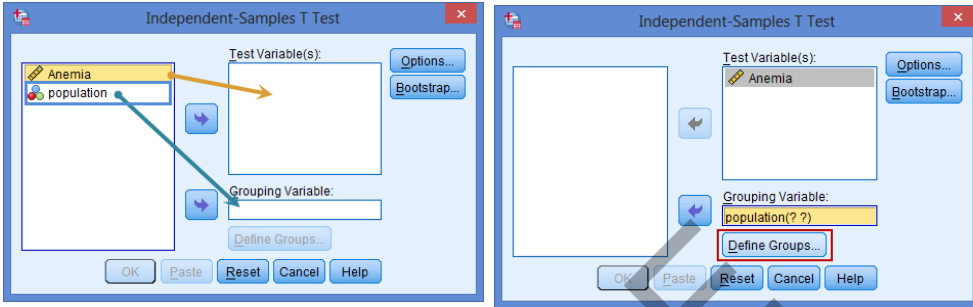
use number 1 as definition for group 1 and number 2 as definition for group 2 then click on Continue
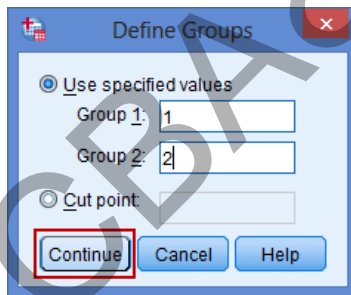
Once we click on OK, we get the following output:

Group Statistics

	population	N	Mean	Std. Deviation	Std. Error Mean
Anemia	Children with iron deficiency anemia	50	6.5800	2.20454	.31177
	Healthy Children	50	12.9340	1.66226	.23508

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
							Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)			Lower	Upper
Anemia	Equal variances assumed	2.624	.108	-16.273	98	.000	-6.35400	.39046	-7.12886	-5.57914
	Equal variances not assumed			-16.273	91.107	.000	-6.35400	.39046	-7.12960	-5.57840

- In this example, the p-value for Levene's test is 0.108, therefore the result is not significant, which means that both samples have equal variances. Therefore, we

choose t-test for equal variances for equal variances assumed. The p-value for t-test is 0.000, which is less than stated p-value, i.e. 0.05. It falls in the rejection region and the test is significant.

- We say with 95% confidence that the means Hb of two samples are different. Consequently the means Hb of two populations are different.

Example S4-4

Do we conclude that, on the average, lymphocytes and tumor cells differ in size? The followings are the cell diameters (µm) of 40 lymphocytes and 50 tumor cells obtained from biopsies of tissue from patients with melanoma.

Table 4.5
Data relating to Lymphocytes and tumor cells

Lymphocytes			
9.0	4.7	8.9	8.4
6.3	5.0	7.8	8.0
8.6	6.8	5.7	6.2
7.4	4.9	6.4	6.3
8.8	7.1	4.7	6.4
9.4	4.8	4.9	5.9
5.7	3.5	10.4	8.0
7.0	7.1	7.6	7.1
8.7	7.4	7.1	8.8
5.2	5.3	8.4	8.3

Tumor cells				
12.6	16.2	23.3	20.0	19.1
16.7	15.8	17.9	19.1	18.9
20.0	13.9	13.9	22.8	17.9
17.7	16.9	22.8	19.6	18.2
16.3	18.1	11.2	18.6	16.1
14.6	23.9	17.1	21.0	19.4
15.9	16.0	13.4	16.6	18.7
17.8	22.1	18.3	13.0	15.2
15.1	16.4	19.4	18.4	20.7
17.7	24.3	19.5	16.4	21.5

Can we say at 5% level of significance that on the average tumor cells differ in size? (source Daniel, 1991)

To test this Hypothesis we follow the following steps :

Analyze→Compare Means→Independent Sample T-Test…
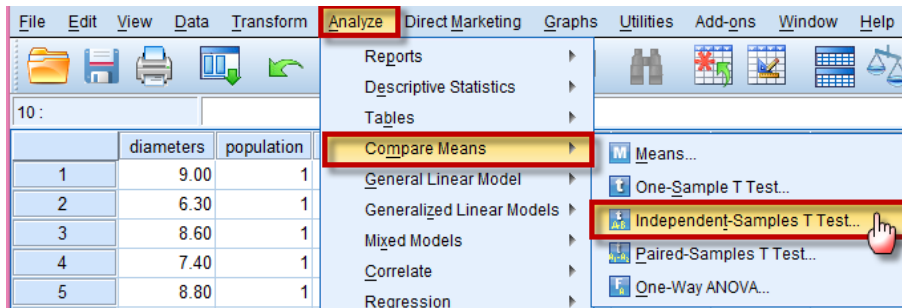
Move the variable diameters to **Test Variable(s):**

Move the variable population to **Grouping Variable:**

Then click on **Define Groups...**
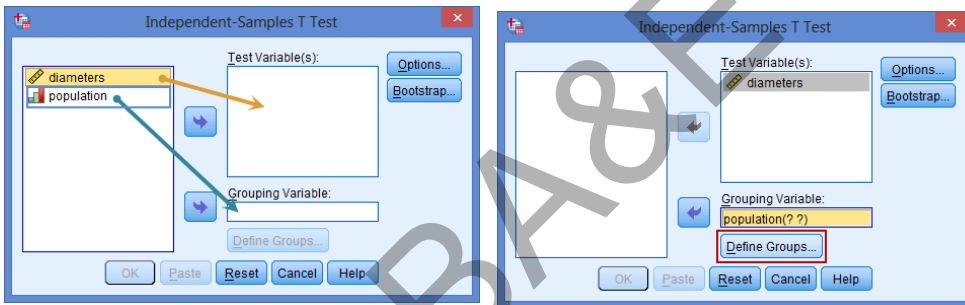
use number 1 as definition for group 1 and number 2 as definition for group 2 then click on **Continue**
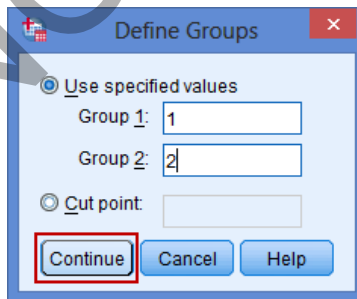
Once we click on **OK**, we get the following output:

Group Statistics

	population	N	Mean	Std. Deviation	Std. Error Mean
diameters	Lymphocytes	40	6.9500	1.59583	.25232
	Tumor cells	50	17.9200	2.96861	.41983

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
							Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)			Lower	Upper
diameters	Equal variances assumed	9.684	.003	-21.049	88	.000	-10.97000	.52116	-12.00569	-9.93431
	Equal variances not assumed			-22.396	78.005	.000	-10.97000	.48982	-11.94515	-9.99485

- In this example, the p-value for Levene's test is 0.003, therefore the result is significant, which means that we may consider both samples have different variances. Therefore, we choose t-test for equal variances for equal variances not assumed. The p-value for t-test is 0.000, which is less than stated p-value, i.e. 0.05. It falls in the rejection region and the test is significant.

- We say with 95% confidence that on the average tumor cells of both the samples differ in size.

NOTE: When the condition of normality is not satisfied, we go for non-parametric-tests (to be discussed in Chapter 8). When the sampled populations are decidedly non-normal, any inference derived from the small samples (t-test) for $\mu_1 = \mu_2$ is not reliable. In this case, one alternative is to use Wilcoxon Rank-sum test.

4.5.4 t-test for Paired Observations

Till now, tests were used to find the difference between two independent samples. In this section, t-test will be used for paired observations. Let us first examine the potential drawback in using the t-test for two independent samples.

Suppose an elementary school teacher wants to compare two methods of teaching of reading skills of first graders. One way is to choose randomly 40 students from the available first graders. Two equal groups are formed randomly and reading achievement test scores are obtained after completion of the experiment. t-test is used to test the difference between two methods. A potential drawback to this method is that IQ, reading ability, socio- economic of the elementary graders are not taken into consideration before dividing into two groups.

A better method of forming the group is to remove the variation of extraneous factors such as IQ, reading ability, etc. One way to do this is to match the first graders in pairs according to IQ, socio-economic status, etc. and from each pair one member is selected randomly to be taught by Method-I and other member to be taught by Method II, then the difference between the *matched pairs* of achievement test scores would provide a clear picture of the true difference in achievement for the two rating methods as the matching would cancel the effects of the extraneous factors that formed the basis of matching. Groups formed in such a way are called *matched groups*. In medical trial it is all the matched frequency. Matching is done on age; on blood pressure (B.P) levels sometimes experiments are conducted on identical timings etc.

The objective of the paired comparison test is to eliminate the effect of extraneous factors by making the pairs similar with respect to as many variables as possible. It gives an excellent result if one can do this but in the presence of many factors, it is not an easy task. Therefore, the research worker prefers to form independent groups. Here we do not

perform the analysis on individual observations but we use the differences between individual pairs of observations. Because of this reason, the condition of *the equality of variances* is not strictly required.

In this type of problems, our hypothesis is, that there is no difference between two informations taken before and after the application of a treatment. This type of test is commonly used in medical science. If one wants to see the effect of medicine or diet on serum cholesterol levels, one will select a group of patients, measure their serum cholesterol levels, apply some medicine or diet and after completion of the course again measure the serum cholesterol levels and see the difference.

To test the significance, we proceed as:

$$t = \frac{\overline{d} - 0}{s_d / \sqrt{n}} \tag{4.19}$$

where $\overline{d}$ is the average of the differences between two paired observations. s_d is the standard deviation of the differences.

Example 4.24:

Thirty-six children were selected at random from a school and an intelligence test was given on the day they had breakfast. The same children were given a similar test on the day they did not have the breakfast. Test, whether fasting affects the test performance. The result of the two tests are given in the following Table:

With breakfast	17	16	21	20	21	19	20	14	13
Without breakfast	14	15	18	15	16	15	16	17	15
With breakfast	10	23	21	12	19	14	15	18	13
Without breakfast	9	21	18	13	18	10	15	15	13
With breakfast	24	20	18	18	11	19	10	15	17
Without breakfast	23	18	13	16	7	15	8	11	13
With breakfast	24	13	15	14	17	19	16	18	24
Without breakfast	24	12	14	12	12	18	19	16	22

Is there a difference between the scores with and without breakfast?

Solution:

From Table 4.7, we can see that:

Sum of the difference $= \Sigma d = 72$

Average of the difference $= \overline{d} = \dfrac{72}{36} = 2$

$\Sigma d^2 = 306$

$$s = \sqrt{\frac{1}{36-1}\left[306 - \frac{(72)^2}{36}\right]} = 2.151$$

(1) H_0 : There is no difference between two groups

   H_1 :There is difference between two groups.

(2) Since the observations are paired, therefore, t-test for paired observations will be applied

(3) Test –Statistics: $t = \dfrac{\overline{d}-0}{s/\sqrt{n}} = \dfrac{2}{2.151/\sqrt{36}} = 5.57$

(4) The table value at 5% level of significance for 35 d.f is 1.6896 (note that is a one-tail test)

(5) Calculated value is much greater than table value. Therefore, the null hypothesis is not accepted, we say with 95% confidence that breakfast has positive effect on the students' performance as we see that experiment has mean increase of 2.0 points on the test scores with standard deviation of 2.15.

Table 4.7:
Data relating to scores with and without breakfast

With breakfast	Without breakfast	Difference d	d^2	With Breakfast	Without Breakfast	Difference d	d^2
17	14	3	9	24	23	1	1
16	15	1	1	20	18	2	4
21	18	3	9	18	13	5	25
20	15	5	25	18	16	2	4
21	16	5	25	11	7	4	16
19	15	4	16	19	15	4	16
20	16	4	16	10	8	2	4
14	17	-3	9	15	11	4	16
13	15	-2	4	17	13	4	16
10	9	1	1	24	24	0	0
23	21	2	4	13	12	1	1
21	18	3	9	15	14	1	1
12	13	-1	1	14	12	2	4
19	18	1	1	17	12	5	25
14	10	4	16	19	18	1	1
15	15	0	0	16	19	-3	9
18	15	3	9	18	16	2	4
13	13	0	0	24	22	2	4
						72	306

This Problem has been also solved by using IBM-SPSS Package. Before we proceed further the normality of the observations with breakfast and without breakfast has been checked using Kolmogorov-Smirnov method. If the observations will be normal then the different of these observation will be normal

Example S4-5

To test this Hypothesis that there is a difference between the scores with and without breakfast for the data given in example 4.26, we follow the following steps :
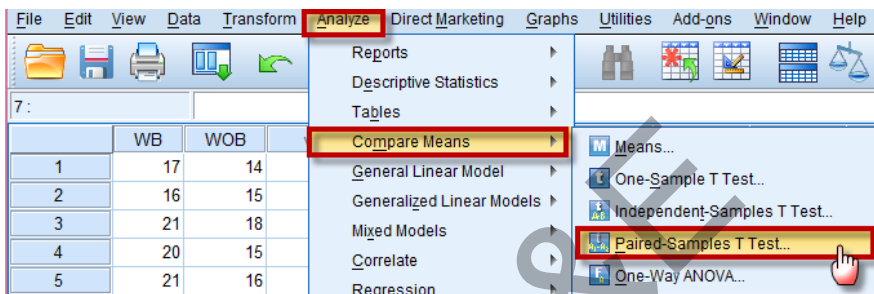
Analyze→Compare Means→Paired Sample T-Test…

Then we move the variables as follows:
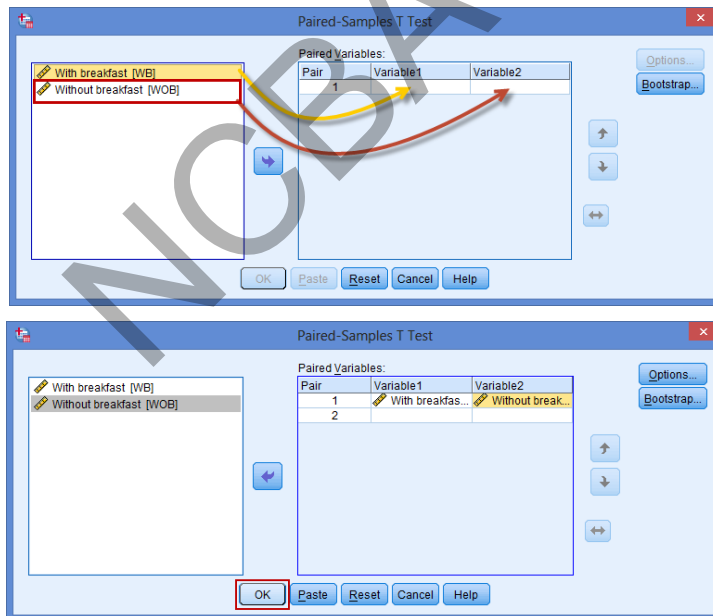
Once we click on , we get the following outputs:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	VAR00001	17.1667	36	3.87298	.64550
	VAR00002	15.1667	36	3.90238	.65040

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	VAR00001 & VAR00002	36	.847	.000

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	VAR00001 - VAR00002	2.00000	2.15141	.35857	1.27207	2.72793	5.578	35	.000

Note that the correlation between matched pair is high and significant (r=0.847 and p-value = 0.000). Since paired test is one-tailed test, we divide the p-value by 2 and we get p=.000, therefore, the null hypothesis is not accepted. We see that mean ($\bar{d}$) is positive (d = with breakfast - without breakfast), therefore, we can say with 95% confidence that fasting has bad effect on the test score.

Before we proceed to apply t-test for paired observations it is advised that one should test the normality of the observations using Kolmogorov-Smirnov Z test. If the condition of normality is satisfied then one should apply t-test for paired observations otherwise one must use non-parametric tests equivalent to t-test for paired observations. If one is not aware of Kolmogorov- Smirnov test, one can see the significance of correlation coefficient and can apply paired t-test if the correlation coefficient is significant.

Example S4-6

Sixteen students were selected at random, their rates of heartbeat were taken while taking a final examination and while they were in relaxing situation. The results are noted and given in table 4.8. Test at 5% level of significance, whether examination has an effect on the heartbeat?

Table 4.8
Data relating to heartbeats during examination and relaxing situations

During examination (x)	Relaxing situation (y)	During examination (x)	Relaxing situation (y)
98	78	102	80
112	76	105	74
85	80	120	86
89	76	83	78
106	82	97	74
110	85	90	80
92	75	101	87
86	76	88	72

Solution:

We note the number of variables is 16 (small sample)

Therefor we have to test the normality before testing t-test for paired samples

And to test the normality follow the following steps:

Analyze → Nonparametric tests → Legacy Dialogs → 1-Sample K-S …
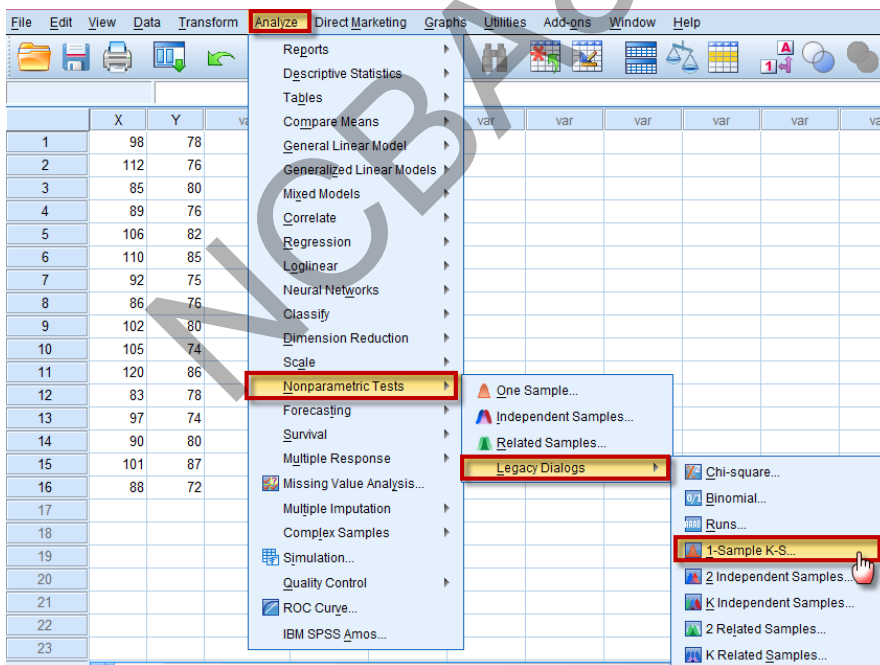
We move the two variables to Test Variable List: and select ✓ Normal then we click on OK

In the output we will be:

We note the p value for both variables is greater than 0.05 which means that both variables follow normal distribution so we can use t-paired test (no need for a non-parametric test)

Now to test the hypothesis that there is a difference between the during examination and relaxing situation? We follow the following steps: 4

Analyze ➔ Compare Means ➔   Paired Sample T-Test…
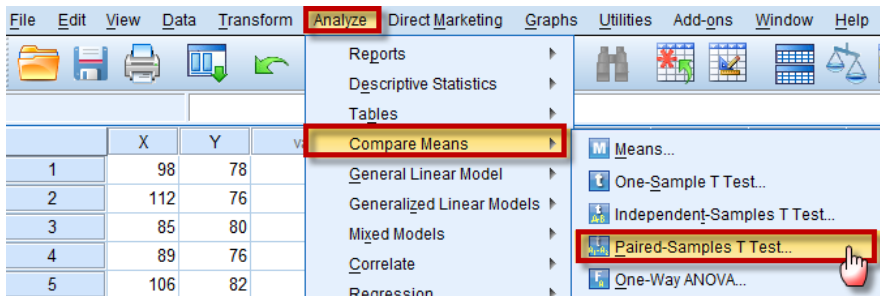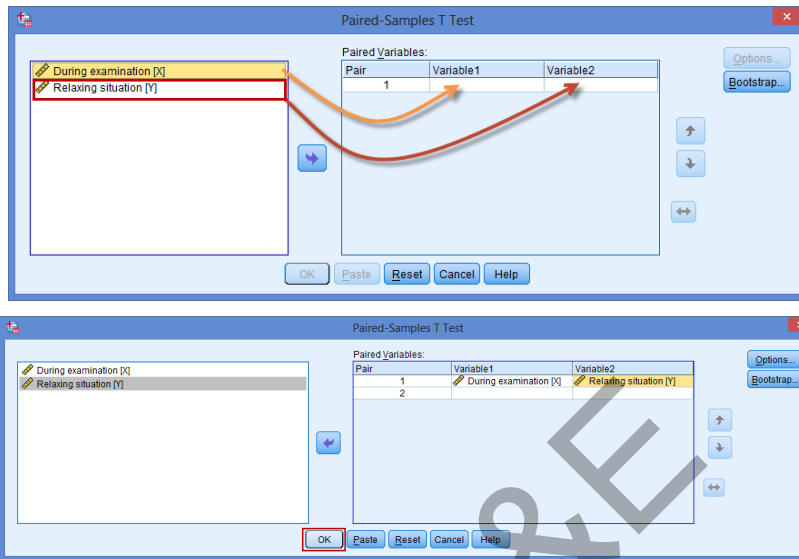
Then we move the variables as follows:

Once we click on OK, we get the following outputs:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	During examination	97.7500	16	10.89648	2.72412
	Relaxing situation	78.6875	16	4.49768	1.12442

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	During examination & Relaxing situation	16	.484	.058

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	During examination - Relaxing situation	19.06250	9.56709	2.39177	13.96456	24.16044	7.970	15	.000

We can see that on the average 19 points heartbeat is more while students are in the examination hall with s.d. 9.6. p-value < 0.000 (one-tailed is half of two-tailed). We do not accept the null hypothesis and say that students have greater heart beat during examination.

4.6 Testing a Population Variance for Single Samples

Hypothesis testing about a population variance may be carried out using chi-square $\left(\chi^2\right)$ distribution. Note that in the application of chi-square, the assumption of normality is required whether the sample is small or large and samples selected from the population must be random. Like t or z-tests, this can also be conducted as one-tailed and two-tailed tests.

Table 4.9

	One-tail	Two-tail
H_0	$\sigma^2 = \sigma_0^2$	$\sigma^2 = \sigma_0^2$
H_1	$\sigma^2 \overset{>}{<} \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$

where σ_0^2 is the specified value of σ^2 (population variance)

The test-statistic is

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \tag{4.20}$$

where s^2 is the sample variance. The degree of freedom for χ^2 is n – 1.

χ^2 distribution tends to normality as the sample size increases (see Figure 4.9).
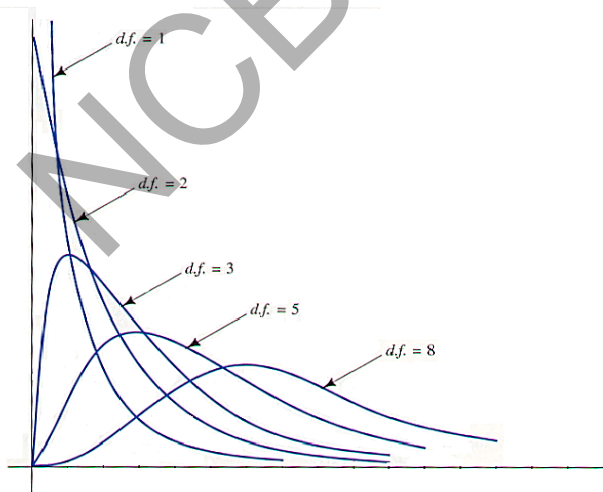
Fig. 4.9: Behavior of χ^2-distribution as sample size increases.

The general principle of testing the hypothesis is the same as mentioned before. 95% confidence limits for population variance may be calculated as

$$P\left[\chi^2_{0.025} < \frac{(n-1)S^2}{\sigma^2_{0.025}} < \chi^2_{0.975}\right] = 95\%$$

(4.21)

$$P\left[\frac{(n-1)s^2}{\sigma^2_{0.025}} < \sigma^2 < \frac{(n-1)s^2}{\sigma^2_{0.975}}\right] = 95\%$$

(4.22)

This provides confidence interval for σ^2 as

$$\frac{(n-1)s^2}{\sigma^2_{0.975}} \text{ and } \frac{(n-1)s^2}{\sigma^2_{0.025}}$$

(4.23)

If we are interested in constructing confidence limits for σ then these may be approximately calculated as:

$$\sqrt{\frac{(n-1)s^2}{\sigma^2_{0.975}}} \text{ and } \sqrt{\frac{(n-1)s^2}{\sigma^2_{0.025}}}$$

(4.24)

Example 4.25:
A hospital conducted a study of acute leukemia. For this purpose a random sample of 25 patients was selected from an approximate normal population. The Hemoglobin (gm%) values were recovered. The variance of these observations was 4.6. Can we say at 5% level of significance that the variance of population from which the sample has been selected is 5?

Solution:

(1) $H_0 : \sigma^2 = 5 \qquad s^2 = 4.6$

$\quad H_1 : \sigma^2 \neq 5 \qquad \sigma^2 = 5 \qquad n = 25$

(2) $\alpha = 0.05$

There is a single sample. It is required to *test variance,* therefore chi-square test for single sample will be used (using 4.15).

(3) test-statistic: $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

$$= \frac{24 \times 4.6}{5} = 22.08$$

(4) Since it is a two- tail test, to see table value we will divide 0.05 by 2 (as in the case of t or z) which come out to be 0.025. Subtract 0.025 from 1 which is 0.975. (See Chi square table-Chapter 8) under $\chi^2_{0.975}$ and against $(25 - 1) = 24$ d.f. The table value is 39.364.

(5) The calculated value of χ^2 is 22.08 for one tail test which is less than the table value. So we do not *reject* the null hypotheses and say with 97.5% confidence that this sample has been selected from a population whose variance is 5.

Note that for one tailed test (less than or greater than) we see the table directly under $\chi^2_{0.975}$ and against the desired degrees of freedoms.
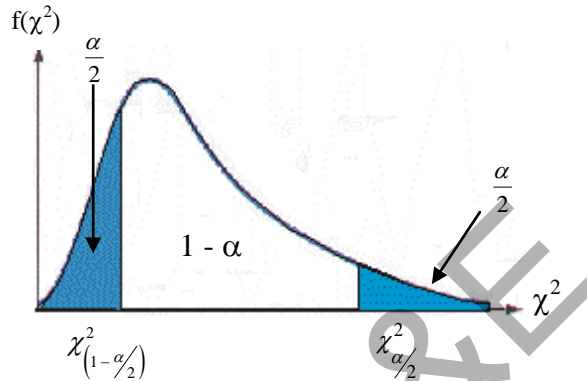
Fig. 4.10: The location of $\chi^2_{(1-\alpha/2)}$ and $\chi^2_{\alpha/2}$ for Chi-square distribution.

The confidence limits for σ^2 (population variance) may be calculated as:

(a) We see from the table

   (i) $\chi^2_{0.025}$ at $24 = (25 - 1)$ degrees of freedom $= 12.401$

   (ii) $\chi^2_{0.975}$ at $24 = (25 - 1)$ degrees of freedom $= 39.364$

(b) We calculate:

   (i) $\dfrac{(n-1)s^2}{\chi^2_{0.025}} = \dfrac{(24)\,(4.6)}{12.401} = 8.9$

   (ii) $\dfrac{(n-1)s^2}{\chi^2_{0.975}} = \dfrac{24 \times 4.6}{39.364} = 2.8$

Therefore, the confidence limits for σ^2 are [2.8, 8.9][see Fig 4.11}

Note that the sample value 4.6 is covered by the interval and confidence limits for σ are given by

$(\sqrt{2.8}\ ,\ \sqrt{8.9}) = [1.67 \sim 2.98]$

4.7 Testing the Ratio of Two Population Variances

Variance test should invariably be applied before conducting a small-sample t-test, for the difference of two means, as the condition of equality of variances is required under its assumptions. In other words, the application of t-test for two independent samples requires the assumption that the variances of the two populations are equal. Sometimes, the assumptions of equality of variances need to be tested. If the variances are significantly different than any inference based on the t-test becomes suspected. Therefore, it is essential that we detect the significance difference between two variances before applying the small-sample t-test for two independent samples. These variances may also be tested through *variance ratio test* commonly known as F-test, i.e.

$$F = \sigma_1^2 \,/\, \sigma_2^2 \tag{4.25}$$

If two variances are equal then $F = 1$.

We know that population variances are never known and we also know that for large samples, s_1^2 (variance of the first sample) and s_2^2 (variance of the second sample) are unbiased estimates of population variances respectively, therefore, F-is defined as:

$$\begin{aligned} F &= S_1^2 / S_2^2 \quad \text{if} \quad S_1^2 > S_2^2 \\ &= S_2^2 / S_1^2 \quad \text{if} \quad S_2^2 > S_1^2 \end{aligned} \tag{4.26}$$

Here the null hypothesis is $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1 : \sigma_1^2 > \sigma_2^2 \ or \ \sigma_1^2 < \sigma_2^2$. Samples are randomly and independently selected from two normal populations. Note that F takes only non-negative values, as it is the ratio of two variances. The range of the F is from zero to infinity.

Example 4.26:
An experiment was conducted to examine the diet metabolizable energy content of commercial cat foods. Fifty-seven domestic short hair cats were selected. Twenty eight were fed on a diet of commercial canned cat food whereas 29 cats were fed on a diet of dry cat food. This experiment was completed in three weeks. At the end of the experiment, metabolizable energy content was determined for each cat. Do you say at 5% level of significance that variation in metabolizable energy content in cats fed on two types of food were different. The data is given as:

	Canned food	Dry food
Sample size Sample	28	29
standard deviation	0.96	3.70

(Feline Practice Vol. 15(2), 1986)

Solution:

(1) $H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Table (4.10)

$$n = \frac{\left[A\sqrt{P_0(1-P_0)} + B\sqrt{P_a(1-P_a)}\right]^2}{(P_0 - P_a)^2}$$

For 5% level of significance and 90% power (Two sided), A = 1.96, B = 1.28

P_a	P_o																		
	0.05	0.07	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.05		1518	301	96	51	32	23	17	13	10	8	7	6	5	4	3	3	2	2
0.07	1421		930	165	73	43	29	21	16	12	10	8	6	5	4	3	3	2	2
0.10	264	869		470	137	68	42	28	21	16	12	10	8	6	5	4	3	3	2
0.15	79	144	437		617	171	82	49	33	23	17	13	10	8	7	5	4	3	2
0.20	40	61	121	588		741	199	94	55	36	25	19	14	11	8	7	5	4	3
0.25	25	35	58	158	717		844	222	102	59	38	26	19	14	11	8	6	5	4
0.30	17	23	35	74	188	825		926	240	109	62	40	27	19	14	11	8	6	4
0.35	12	16	23	43	87	213	911		987	252	113	63	40	27	19	14	10	7	5
0.40	10	12	17	29	50	97	233	976		1027	259	115	63	40	26	18	13	9	6
0.45	8	9	13	20	33	56	105	248	1021		1046	260	114	62	38	25	17	12	8
0.50	6	8	10	15	23	36	60	111	257	1044		1044	257	111	60	36	23	15	10
0.55	5	6	8	12	17	25	38	62	114	260	1046		1021	248	105	56	33	20	13
0.60	4	5	6	9	13	18	26	40	63	115	259	1027		976	233	97	50	29	17
0.65	3	4	5	7	10	14	19	27	40	63	113	252	987		911	213	87	43	23
0.70	3	3	4	6	8	11	14	19	27	40	62	109	240	926		825	188	74	35
0.75	2	3	4	5	6	8	11	14	19	26	38	59	102	222	844		717	158	58
0.80	2	2	3	4	5	7	8	11	14	19	25	36	55	94	199	741		588	121
0.85	2	2	2	3	4	5	7	8	10	13	17	23	33	49	82	171	617		437
0.90	-	2	2	3	3	4	5	6	8	10	12	16	21	28	42	68	137	470	

Table-(4.11)

$$n = \frac{\left[A\sqrt{P_0(1-P_0)} + B\sqrt{P_a(1-P_a)}\right]^2}{(P_0 - P_a)^2}$$

For 1% level of significance and 90% power (Two sided), A = 2.58, B = 1.28

P_a	P_o																		
	0.05	0.07	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.05		2197	444	145	77	49	35	26	20	16	13	10	8	7	6	4	4	3	2
0.07	1976		1346	244	110	65	44	31	24	18	15	12	10	8	6	5	4	3	2
0.10	359	1208		682	201	101	62	42	31	23	18	14	11	9	7	6	5	4	3
0.15	104	195	607		887	248	120	72	48	34	25	19	15	12	9	7	6	4	3
0.20	52	82	166	822		1062	288	135	79	52	37	27	20	15	12	9	7	5	4
0.25	32	46	79	218	1007		1207	319	147	85	55	38	27	20	15	12	9	7	5
0.30	22	30	47	102	262	1162		1321	343	156	89	56	39	27	20	15	11	8	6
0.35	16	21	31	59	120	299	1286		1406	359	161	90	57	38	27	19	14	10	7
0.40	12	16	22	39	69	136	328	1381		1461	368	163	90	56	37	25	18	12	8
0.45	9	12	17	27	45	77	148	349	1446		1486	369	161	88	53	35	23	16	10
0.50	8	10	13	20	32	50	84	156	363	1480		1480	363	156	84	50	32	20	13
0.55	6	8	10	16	23	35	53	88	161	369	1486		1446	349	148	77	45	27	17
0.60	5	6	8	12	18	25	37	56	90	163	368	1461		1381	328	136	69	39	22
0.65	4	5	7	10	14	19	27	38	57	90	161	359	1406		1286	299	120	59	31
0.70	4	4	6	8	11	15	20	27	39	56	89	156	343	1321		1162	262	102	47
0.75	3	4	5	7	9	12	15	20	27	38	55	85	147	319	1207		1007	218	79
0.80	3	3	4	5	7	9	12	15	20	27	37	52	79	135	288	1062		822	166
0.85	2	3	3	4	6	7	9	12	15	19	25	34	48	72	120	248	887		607
0.90	2	2	3	4	5	6	7	9	11	14	18	23	31	42	62	101	201	682	

Table-(4.12)

$$n = \frac{\left[A\sqrt{P_0(1 - P_0)} + B\sqrt{P_a(1 - P_a)}\right]^2}{(P_0 - P_a)^2}$$

For 5% level of significance and 90% power (Two sided), A = 1.645, B = 1.28

P_a	P_o																		
	0.05	0.07	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.05		1221	239	76	40	25	18	13	10	8	6	5	4	4	3	3	2	2	-
0.07	1174		748	131	58	34	23	16	12	10	8	6	5	4	3	3	2	2	-
0.10	221	718		378	109	54	33	22	16	12	10	8	6	5	4	3	3	2	2
0.15	67	121	362		498	137	66	39	26	19	14	11	8	7	5	4	3	3	2
0.20	34	52	102	484		600	161	75	44	29	20	15	11	9	7	5	4	3	3
0.25	21	30	49	131	588		685	180	83	48	31	21	16	12	9	7	5	4	3
0.30	15	20	30	62	155	675		753	194	88	50	32	22	16	12	9	7	5	4
0.35	11	14	20	36	72	175	745		803	205	92	52	33	22	16	11	8	6	5
0.40	8	11	14	24	42	80	191	798		836	211	93	52	32	22	15	11	8	6
0.45	7	8	11	17	27	46	86	203	833		852	212	93	51	31	21	14	10	7
0.50	5	7	9	13	19	30	49	91	210	851		851	210	91	49	30	19	13	9
0.55	4	5	7	10	14	21	31	51	93	212	852		833	203	86	46	27	17	11
0.60	4	4	6	8	11	15	22	32	52	93	211	836		798	191	80	42	24	14
0.65	3	4	5	6	8	11	16	22	33	52	92	205	803		745	175	72	36	20
0.70	3	3	4	5	7	9	12	16	22	32	50	88	194	753		675	155	62	30
0.75	2	3	3	4	5	7	9	12	16	21	31	48	83	180	685		588	131	49
0.80	2	2	3	3	4	5	7	9	11	15	20	29	44	75	161	600		484	102
0.85	2	2	2	3	3	4	5	7	8	11	14	19	26	39	66	137	498		362
0.90	-	-	2	2	3	3	4	5	6	8	10	12	16	22	33	54	109	378	

Table-(4.13)

$$n = \frac{\left[A\sqrt{P_0(1-P_0)} + B\sqrt{P_a(1-P_a)}\right]^2}{(P_0 - P_a)^2}$$

For 1% level of significance and 90% power (Two sided), A = 2.33, B = 1.28

P_a	P_o																		
	0.05	0.07	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.05		1908	383	124	66	42	30	22	17	13	11	9	7	6	5	4	3	2	2
0.07	1741		1169	210	94	56	37	27	20	16	13	10	8	7	5	4	3	3	2
0.10	319	1064		592	174	87	53	36	26	20	15	12	10	8	6	5	4	3	2
0.15	94	173	535		772	215	104	62	41	30	22	17	13	10	8	6	5	4	3
0.20	47	73	147	723		926	250	118	69	45	32	23	18	14	10	8	6	5	3
0.25	29	41	70	193	884		1053	278	128	74	48	33	24	18	13	10	8	6	4
0.30	20	27	42	90	231	1019		1154	299	136	77	49	34	24	18	13	10	7	5
0.35	14	19	28	53	106	263	1127		1228	314	141	79	50	33	23	17	12	9	6
0.40	11	14	20	35	61	119	288	1209		1277	322	142	79	49	32	22	16	11	8
0.45	9	11	15	24	40	68	130	306	1265		1299	323	141	77	47	31	21	14	9
0.50	7	9	12	18	28	44	73	137	318	1295		1295	318	137	73	44	28	18	12
0.55	6	7	9	14	21	31	47	77	141	323	1299		1265	306	130	68	40	24	15
0.60	5	6	8	11	16	22	32	49	79	142	322	1277		1209	288	119	61	35	20
0.65	4	5	6	9	12	17	23	33	50	79	141	314	1228		1127	263	106	53	28
0.70	3	4	5	7	10	13	18	24	34	49	77	136	299	1154		1019	231	90	42
0.75	3	3	4	6	8	10	13	18	24	33	48	74	128	278	1053		884	193	70
0.80	2	3	3	5	6	8	10	14	18	23	32	45	69	118	250	926		723	147
0.85	2	2	3	4	5	6	8	10	13	17	22	30	41	62	104	215	772		535
0.90	2	2	2	3	4	5	6	8	10	12	15	20	26	36	53	87	174	592	

Table-(4.14)

$$n = \frac{\left[A\sqrt{P_0(1 - P_0)} + B\sqrt{P_a(1 - P_a)}\right]^2}{(P_0 - P_a)^2}$$

For 5% level of significance and 80% power (Two sided), A = 1.96, B = 0.84

P_a	P_o																		
	0.05	0.07	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.05		1167	238	78	42	27	19	14	11	9	7	6	5	4	3	3	2	2	-
0.07	1029		716	131	59	35	24	17	13	10	8	7	5	4	4	3	2	2	-
0.10	185	629		363	108	54	34	23	17	13	10	8	6	5	4	3	3	2	2
0.15	53	101	316		470	132	64	39	26	19	14	11	8	7	5	4	3	3	2
0.20	26	42	86	430		562	153	72	43	28	20	15	11	8	7	5	4	3	2
0.25	16	24	41	114	527		637	169	78	45	29	20	15	11	8	6	5	4	3
0.30	11	15	24	53	137	609		697	181	83	47	30	21	15	11	8	6	4	3
0.35	8	11	16	31	63	157	675		741	190	85	48	30	20	14	10	7	5	4
0.40	6	8	12	20	36	71	172	726		770	194	86	48	30	20	13	9	7	4
0.45	5	6	9	14	24	41	77	183	760		782	195	85	46	28	18	12	8	5
0.50	4	5	7	11	17	26	44	82	191	779		779	191	82	44	26	17	11	7
0.55	3	4	5	8	12	18	28	46	85	195	782		760	183	77	41	24	14	9
0.60	3	3	4	7	9	13	20	30	48	86	194	770		726	172	71	36	20	12
0.65	2	3	4	5	7	10	14	20	30	48	85	190	741		675	157	63	31	16
0.70	2	2	3	4	6	8	11	15	21	30	47	83	181	697		609	137	53	24
0.75	2	2	3	4	5	6	8	11	15	20	29	45	78	169	637		527	114	41
0.80	2	2	2	3	4	5	7	8	11	15	20	28	43	72	153	562		430	86
0.85	-	2	2	3	3	4	5	7	8	11	14	19	26	39	64	132	470		316
0.90	-	-	2	2	3	3	4	5	6	8	10	13	17	23	34	54	108	363	

Table-(4.15)

$$n = \frac{\left[A\sqrt{P_0(1 - P_0)} + B\sqrt{P_a(1 - P_a)}\right]^2}{(P_0 - P_a)^2}$$

For 1% level of significance and 80% power (Two sided), A = 2.58, B = 0.84

P_a	P_o																	
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.05		367	122	66	43	30	23	18	14	11	9	7	6	5	4	3	2	2
0.10	266		551	165	84	52	36	26	20	15	12	10	8	6	5	4	3	2
0.15	75	462		710	201	98	59	40	28	21	16	13	10	8	6	5	4	3
0.20	36	124	633		845	231	110	64	42	30	22	16	13	10	7	6	4	3
0.25	22	58	166	780		957	255	118	68	44	31	22	16	12	9	7	5	4
0.30	15	34	76	201	903		1044	272	124	71	45	31	22	16	12	9	6	4
0.35	11	23	44	92	231	1003		1109	284	128	71	45	30	21	15	11	7	5
0.40	8	16	29	53	104	255	1079		1150	290	128	71	44	29	20	14	9	6
0.45	7	12	20	34	59	114	272	1132		1167	290	126	68	41	27	18	12	8
0.50	5	9	15	24	38	65	122	284	1161		1161	284	122	65	38	24	15	9
0.55	4	8	12	18	27	41	68	126	290	1167		1132	272	114	59	34	20	12
0.60	4	6	9	14	20	29	44	71	128	290	1150		1079	255	104	53	29	16
0.65	3	5	7	11	15	21	30	45	71	128	284	1109		1003	231	92	44	23
0.70	3	4	6	9	12	16	22	31	45	71	124	272	1044		903	201	76	34
0.75	2	4	5	7	9	12	16	22	31	44	68	118	255	957		780	166	58
0.80	2	3	4	6	7	10	13	16	22	30	42	64	110	231	845		633	124
0.85	2	3	4	5	6	8	10	13	16	21	28	40	59	98	201	710		462
0.90	-	2	3	4	5	6	8	10	12	15	20	26	36	52	84	165	551	

Table-(4.16)

$$n = \frac{\left[A\sqrt{P_0(1-P_0)} + B\sqrt{P_a(1-P_a)}\right]^2}{(P_0 - P_a)^2}$$

For 5% level of significance and 80% power (One sided), A = 1.645, B = 0.84

P_a	P_o																	
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.05		184	60	32	21	15	11	8	7	5	5	4	3	3	2	2	-	-
0.10	150		282	83	42	26	18	13	10	8	6	5	4	3	3	2	2	1
0.15	44	252		368	103	50	30	20	14	11	8	7	5	4	3	3	2	2
0.20	22	69	342		440	119	56	33	22	15	11	9	7	5	4	3	3	2
0.25	14	33	91	418		500	132	61	35	23	16	12	9	7	5	4	3	2
0.30	9	20	43	109	482		548	142	65	37	24	16	12	9	6	5	4	3
0.35	7	13	25	50	124	534		583	149	67	38	24	16	11	8	6	4	3
0.40	5	10	16	29	57	136	573		606	153	68	38	23	16	11	8	5	4
0.45	4	7	12	19	32	62	145	600		616	153	67	37	22	15	10	7	5
0.50	3	6	9	13	21	35	65	151	614		614	151	65	35	21	13	9	6
0.55	3	5	7	10	15	22	37	67	153	616		600	145	62	32	19	12	7
0.60	2	4	5	8	11	16	23	38	68	153	606		573	136	57	29	16	10
0.65	2	3	4	6	8	11	16	24	38	67	149	583		534	124	50	25	13
0.70	2	3	4	5	6	9	12	16	24	37	65	142	548		482	109	43	20
0.75	2	2	3	4	5	7	9	12	16	23	35	61	132	500		418	91	33
0.80	-	2	3	3	4	5	7	9	11	15	22	33	56	119	440		342	69
0.85	-	2	2	3	3	4	5	7	8	11	14	20	30	50	103	368		252
0.90	-	-	2	2	3	3	4	5	6	8	10	13	18	26	42	83	282	

Table-(4.16)

$$n = \frac{\left[A\sqrt{P_0(1 - P_0)} + B\sqrt{P_a(1 - P_a)}\right]^2}{(P_0 - P_a)^2}$$

For 1% level of significance and 80% power (One sided), A = 2.33, B = 0.84

P_a	P_o																	
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.05		312	104	56	36	26	19	15	12	9	8	6	5	4	3	3	2	2
0.07	1304	927	172	78	47	32	23	17	14	11	9	7	6	5	4	3	2	2
0.10	231		471	141	71	44	30	22	17	13	10	8	7	5	4	3	3	2
0.15	66	400		608	172	84	50	34	24	18	14	11	8	7	5	4	3	2
0.20	32	108	546		724	198	94	55	36	26	19	14	11	8	6	5	4	3
0.25	19	51	143	672		820	218	101	58	38	26	19	14	11	8	6	4	3
0.30	13	30	66	174	778		896	233	106	61	39	26	19	14	10	7	5	4
0.35	10	20	38	79	199	863		952	244	109	61	39	26	18	13	9	7	4
0.40	7	14	25	46	90	219	928		987	249	110	61	38	25	17	12	8	5
0.45	6	11	18	30	51	99	234	973		1003	249	109	59	36	23	15	10	7
0.50	5	8	13	21	33	56	105	244	998		998	244	105	56	33	21	13	8
0.55	4	7	10	15	23	36	59	109	249	1003		973	234	99	51	30	18	11
0.60	3	5	8	12	17	25	38	61	110	249	987		928	219	90	46	25	14
0.65	3	4	7	9	13	18	26	39	61	109	244	952		863	199	79	38	20
0.70	2	4	5	7	10	14	19	26	39	61	106	233	896		778	174	66	30
0.75	2	3	4	6	8	11	14	19	26	38	58	101	218	820		672	143	51
0.80	2	3	4	5	6	8	11	14	19	26	36	55	94	198	724		546	108
0.85	2	2	3	4	5	7	8	11	14	18	24	34	50	84	172	608		400
0.90	-	2	3	3	4	5	7	8	10	13	17	22	30	44	71	141	471	

Chapter 5

Analysis of Variance

5.1 Introduction

In Chapter 4, we have studied the testing of hypothesis procedure with two independent samples and for paired observations. In most practical situations, we study, more than two populations. In such cases the application of t-test is not appropriate. Sir R. A. Fisher and his colleagues developed designs of experiments and a statistical technique known as *analysis of variance* (*ANOVA) technique*. In medical research usually, observational and experimental studies are made. Observational studies are based on surveys whereas clinical case studies are based on experiments. Experimental studies are laboratory-controlled experiments where each experiment is designed to compare factors. The experiments that concern clinicians are clinical trials. We allocate drugs or treatments to patients and observe the outcome. Suppose we have two new drugs to be tested along with a control drug, a placebo. There are various ways of performing the experiments depending on an objective. If we are interested in drugs efficacy only, then drugs are randomly assigned to patients and their response noted. A more controlled experiment may form blocks of patients given same age group and select randomly as many patients from a group as the number of drugs or multiple patients per drug. One drug to each patient in the age group called blocks is the randomly given to patients. This way each drug will get as many patients (an equal number for all drugs) as there are blocks. The idea is to make the units in a block as similar as possible. The first experiment is called Completely Randomized Design and the analysis of this design is made by using Analysis of Variance with One-Way Classification The second one is called Randomized Block Design and analysis of this design may be made by using Analysis of variance Two-Way Classification. Similarly other types of designs can be adopted depending on the objectives and resources available. The main purpose of analysis of variance technique is to see, whether there is any difference among k population means in (Note that ANOVA can also be applied on two samples). In this chapter only analysis of variance for one-way classification, two-way classification, repeated measure design, Multivariate Analysis of Variance (MANOVA) and simple factorial design will be discussed. The classification of observations on the basis of single criterion is called one-way classification whereas the classification of observations according to two criteria is called two-way classification. If the classifications are based on multi-way classification with more than two factors then analysis is made using MANOVA (multivariate analysis of variance) and repeated measure design.

5.2 Analysis of Variance with One- Way classification

Suppose there are k treatments (drugs) that are randomly assigned to experimental units. Random allocation of treatments to experimental units is known as *completely randomized design*. For the analysis of such type of data, analysis of variance with one-

way classification is used. What we do, we select independent random samples from different populations to make inferences about the population means associated with various treatments.

The null hypothesis to be tested is

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_k$$

at a particular level of significance. Where μ_1, μ_2, μ_3, ... μ_k are the means of k populations. The alternative hypothesis will be that at least two means differ. The following assumptions must be made:

In an additive model

$$y_{ij} = \mu + \xi_i + \varepsilon_{ij} \quad i = 1, 2, 3,\ldots.n, j = 1, 2, 3,\ldots\ldots, k \tag{5.1}$$

where μ is the general mean response and ξ_i is the effect of the ith drug.

$\Sigma \xi_i = 0$ and $\varepsilon_{ij} \sim NID (0, \sigma^2)$

The assumptions are as:

(i)   The observed values are all independent random variables selected from each sampled population

(ii)   Each sampled population is normally distributed

(iii) The variances of all the populations are same and constant.

When these assumptions are violated, the inferences become doubtful.

One way analysis of variance technique partitions the total sum of square (TSS) into two components called, between sum of squares [SS(B)] and within sum of squares[SS(W)] as shown is diagram 5.1.
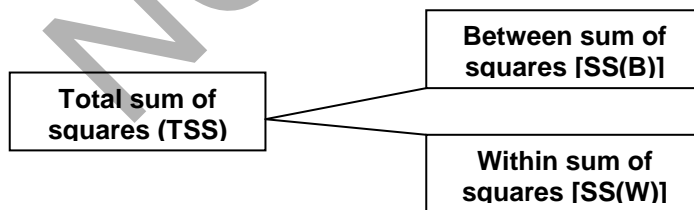
Fig. 5.1: Partitioning of total sum of squares into different components.

If H_0 is true then the two components are used to provide independent estimates of σ^2.

We compare the source of variability by forming F-test i.e.

$$F = \frac{\text{Between mean squares [MS(B)]}}{\text{Within mean squares [MS(W)]}}$$

In the definition of F both the numerator and denominator estimate the σ^2 and consequently if H_0 is true F should be close to 1.

F is based on $\nu_1 = k-1$ and $\nu_2 = n-k$ degrees of freedom where k are treatments and n number of observations. If computed value exceeds the table value, we reject the null hypothesis and conclude that at least two treatment-means differ with each other. The results of the analysis of variance are usually summarized and presented in an analysis of variance table (ANOVA table). The table shows the sources of variation, their respective degrees of freedom, sum of squares, mean sum of squares and computed F-statistic, (in SPSS output, p-value is also given). If there are k treatments with n observations then the output may be displayed in the table 5.1.

Table 5.1:
ANOVA- ONE WAY

Source of variation	d.f	SS	MSS	F-statistic
Between treatments	k-1	SS(B)	SSB/(k-1) = MSB	
Within treatments	n-k	SS(W)	SSW/(n-k) = MSW	F = MS(B)/MS(W)
Total	n-1	TSS		

We have further tests to determine which pairs are significantly different. For this purpose Multiple Range Tests are used and are given as:

(i)	LSD test	(v)	Tukey -HSD test
(ii)	Modified LSD (Bonferroni) test	(vi)	Tukey-B test
(iii)	Duncan's test	(vii)	Scheffe's test
(iv)	Student - Newman-Keuls test	(viii)	Dunnett's test

Any one of the above tests can be applied to test the difference between two samples. LSD and Duncan's tests are commonly applied.

Example 5.1:
Suppose we wish to determine the usefulness of the measurement of serum Lipid-bound Silica Acid (LSA) in the detection of breast cancer. For this purpose, four populations are selected as:

Population A: Normal/control. (Healthy subjects)

Population B: Patients with benign breast cancer

Population C: Patients with benign primary cancer

Population D: Patients with recurrent meta-static breast cancer

One sample from each population is selected randomly and LSA measurements (mg/dl) are recorded. We compare these samples to find out the difference between the means.

The data regarding LSA measurements (mg/dl) are given in Table 5.2.

Table 5.2:
Measurements of Lipid bound silica acid (LSA)

Normal/ Control	Patients with benign breast cancer	Patient with primary breast cancer	Patient with meta-static breast cancer
18.80	24.30	18.00	22.30
18.80	18.60	16.40	22.90
20.10	24.70	22.50	22.70
14.50	22.50	18.20	22.40
15.80	23.00	17.50	25.20
18.20	14.90	21.00	18.70
15.70	22.70	23.20	22.20
20.90	18.60	19.90	23.00
20.40	20.60	19.80	25.50
16.90	24.60	16.20	19.70
180.1	214.5	192.7	224.6

Grand total = 180.1+214.5+192.7+224.6= 811.9

Test at 5% level of significance that there is no difference between 4 groups.

Solution:

(1) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

   H_1 : At least two sample means are not equal

(2) $\alpha = 0.05$

(3) Test-statistic: F-test in one-way ANOVA.

   For the calculation proceed as follows:

(i)   Correction factor = $(811.9)^2/40 = 16479.5402$

(ii)   Total sum of squares = $18.8^2 + 18.8^2 + \ldots + 19.7^2 - 16479.5402 = 352.2097$

(iii) Between sum of squares

$$= \frac{(180.1)^2 + (214.5)^2 + (192.7)^2 + (224.6)^2}{10} - 16479.5402 = 122.9308$$

   Note that the divisor (10) is the number of observations on which the column or group totals are based.

(iv)   Within sum of squares = TSS-SS(B) = 352.2097 - 122.9308 = 229.2789

(v)   Mean sum of squares (B) = 122.9308/(4-1) = 40.9769

(vi)   Mean sum of squares (W) = 229.2789/36 = 6.3689

(vii) F-statistic = 40.9769/ 6.3689 = 6.4339

These may be presented in the ANOVA table

Table 5.3:
One- Way ANOVA

Source of variation	d.f	SS	MSS	F_cal	F_tab
Between groups	3	122.9308	40.9769	40.9769/6.3689 = 6.4339	3.46
Within groups	36	229.2790	6.3689		
Total	39	352.2097			

(4) Table value against 3 and 36 degree of freedom at 5% level of significance is 3.46.

(5) Calculated value is more than table value, result is significant, therefore we do not accept the null hypothesis and say that at least two sample means differs with each other.

IBM-SPSS package may be used for the calculations, as explained by the following example:

Example S5-1

To test that there is no difference between the 4 groups for the data given in table 5.2, the data are entered in one column and we add another grouping variable with the numbers 1,2,3 and 4 corresponding to the 4 (independent) groups, then we follow the following steps:

Analyze→Compare Means→One-Way ANOVA…

Move the variable measurement to Dependent List:

Move the variable population to Factor:
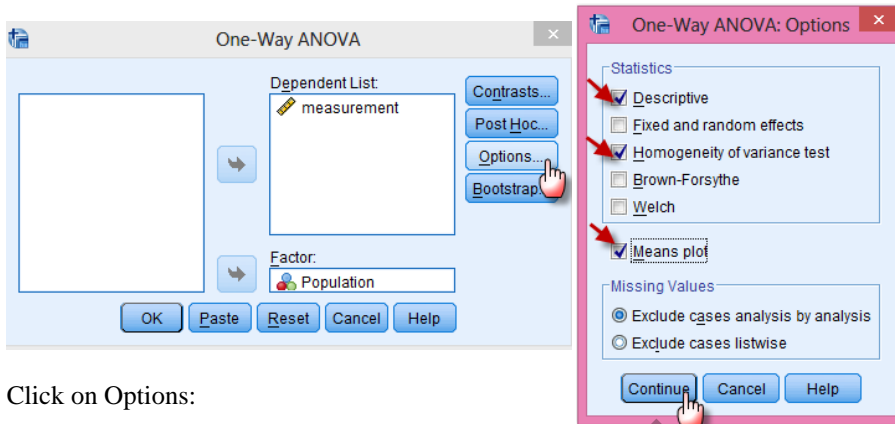
Click on Options:

We click on Continue and on OK, to get the following outputs:

SPSS output for ANOVA One-way

Analysis of Variance

Descriptives

measurement

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Normal/ Control	10	18.0100	2.19821	.69513	16.4375	19.5825	14.50	20.90
Patients with benign breast cancer	10	21.4500	3.22396	1.01950	19.1437	23.7563	14.90	24.70
Patient with primary breast cancer	10	19.2700	2.42901	.76812	17.5324	21.0076	16.20	23.20
Patient with meta-static breast cancer	10	22.4600	2.08551	.65949	20.9681	23.9519	18.70	25.50
Total	40	20.2975	3.00517	.47516	19.3364	21.2586	14.50	25.50

Test of Homogeneity of Variances

measurement

Levene Statistic	df1	df2	Sig.
1.373	3	36	.267

If the p-value of Levene's test of homogeneity of variance is greater than 0.05, then the condition of homogeneity is satisfied and ANOVA technique can be applied to test the difference between different groups. In this example, condition of homogeneity is satisfied (see Levene's test p-value = 0.267), so ANOVA technique is appropriate.

ANOVA

measurement

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	122.931	3	40.977	6.434	.001
Within Groups	229.279	36	6.369		
Total	352.210	39			

See the results of F-statistic from ANOVA table, p = 0.001, which is less than the p-value of 0.05, therefore, the null hypothesis is not accepted. We can say with 95% confidence that at least two sample means are different.

POST HOC Test: Since samples are different, we apply any one of the multiple range tests to see which samples (groups) are homogeneous. We have applied LSD test and modified LSD test (Bonforroni) to see the differences between two means:

Analyze→Compare Means→One-Way ANOVA…

We chose Post Hoc:
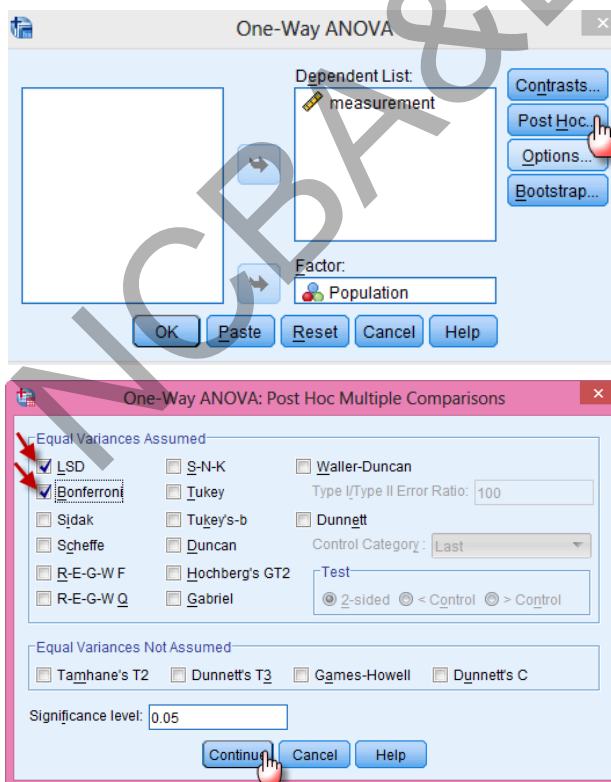
We click on Continue and on OK , to get the following outputs:

SPSS output for multiple range tests

LSD and Modified LSD Tests with 5% level of significance

	(I) CODES	(J) CODES	Mean Difference (I-J)	Std. Error	Sig.
LSD	Control	Benign	-3.4400*	1.1286	.004
		Primary	-1.2600	1.1286	.272
		Meta-Static	-4.4500*	1.1286	.000
	Benign	Control	3.4400*	1.1286	.004
		Primary	2.1800	1.1286	.061
		Meta-Static	-1.0100	1.1286	.377
	Primary	Control	1.2600	1.1286	.272
		Benign	-2.1800	1.1286	.061
		Meta-Static	-3.1900*	1.1286	.008
	Meta-Static	Control	4.4500*	1.1286	.000
		Benign	1.0100	1.1286	.377
		Primary	3.1900*	1.1286	.008
Bonferroni	Control	Benign	-3.4400*	1.1286	.026
		Primary	-1.2600	1.1286	1.000
		Meta-Static	-4.4500*	1.1286	.002
	Benign	Control	3.4400*	1.1286	.026
		Primary	2.1800	1.1286	.368
		Meta-Static	-1.0100	1.1286	1.000
	Primary	Control	1.2600	1.1286	1.000
		Benign	-2.1800	1.1286	.368
		Meta-Static	-3.1900*	1.1286	.046
	Meta-Static	Control	4.4500*	1.1286	.002
		Benign	1.0100	1.1286	1.000
		Primary	3.1900*	1.1286	.046

*. The mean difference is significant at the .05 level.

(a) The result of LSD test

(i) Control group and Primary group are homogeneous

(ii) Benign group and Primary group are homogeneous

(iii) Benign group and Meta-Static group are homogeneous

(b) The result of Bonferroni's test (Modification of LSD)

Results of Bonferroni test are the same as for LSD.

The following figure (obtained through the Means Plot) may reflect the results:
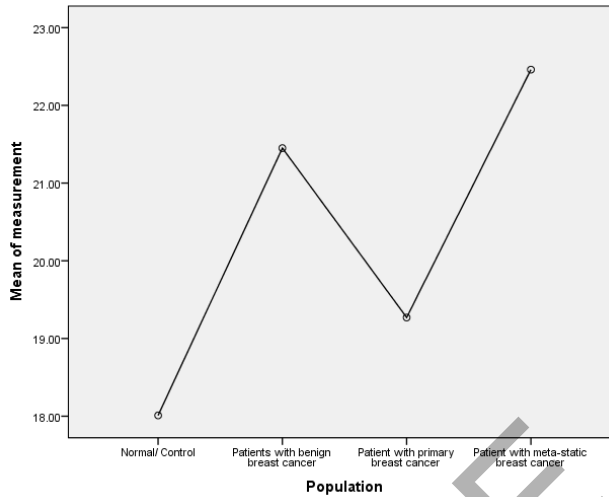
In the previous example, the sample sizes were equal in the different groups (this is known as a balanced model). We can use the same procedure – under the same conditions in case of the unbalanced model, as can be seen in the following example:

Example 5.2:

Anionwu et al. (1981) reported data on steady-state hemoglobin levels for patients with different types of sickle cell disease. The question of interest is whether the steady-state hemoglobin levels differ significantly between patients with different types. The data are given as follows.

Table 5.4

Type of Sickle Cell Disease		
HB SS	**HB S/-Thalassaemia**	**HB SC**
7.2	8.1	10.7
7.7	9.2	11.3
8.0	10.0	11.5
8.1	10.4	11.6
8.3	10.6	11.7
8.4	10.9	11.8
8.4	11.1	12.0
8.5	11.9	12.1
8.6	12.0	12.3
8.7	12.1	12.6
9.1		12.6
9.1		13.3
9.1		13.3
9.8		13.8
10.1		13.9
10.3		

Source: Anionwu et al. (1981)

By using analysis of variance technique, test whether there is any significant difference between three types of Sickle all disease at 5% level of significance.

Example S5-2

To test that there is no difference between the 3 groups for the data given in table 5.4, that is to test:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \text{At least two sample means are not equal,}$$

the data are entered in one column and we add another grouping variable with the numbers 1,2 and 3 corresponding to the 3 (independent) groups, then we follow the following steps :

Analyze→Compare Means→One-Way ANOVA…

Move the variable measurement to Dependent List:

Move the variable population to Factor:
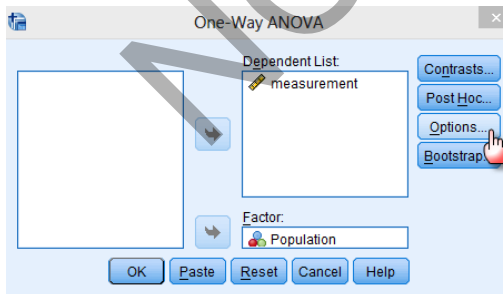
Click on Options:
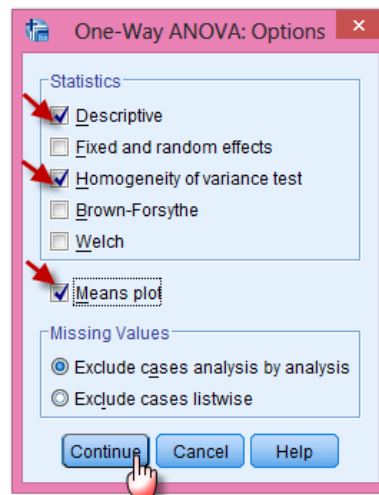
We click on [Continue] and on [OK], to get the following outputs:

SPSS output for ANOVA One-way

Analysis of Variance

Descriptives

measurement

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
HB SS	16	8.713	.8445	.2111	8.263	9.162	7.2	10.3
HB S/-Thalassaemia	10	10.630	1.2841	.4061	9.711	11.549	8.1	12.1
HB SC	15	12.300	.9419	.2432	11.778	12.822	10.7	13.9
Total	41	10.493	1.8564	.2899	9.907	11.079	7.2	13.9

Test of Homogeneity of Variances

measurement

Levene Statistic	df1	df2	Sig.
.902	2	38	.414

The p-value of Levene's test of homogeneity of variance is greater than 0.05, (p-value = 0.414), then the condition of homogeneity is satisfied and ANOVA technique can be applied to test the difference between different groups.

ANOVA

measurement

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	99.889	2	49.945	49.999	.000
Within Groups	37.959	38	.999		
Total	137.848	40			

See the results of F-statistic from ANOVA table, $p < 0.001$, therefore, the null hypothesis is not accepted. We can say with 99% confidence that at least two sample means are different (we may say that the test is highly significant).

POST HOC Test: Since samples are different, we apply any one of the multiple range tests to see which samples (groups) are homogeneous. We have applied LSD test and modified LSD test (Bonforroni) to see the differences between two means:

Analyze→Compare Means→One-Way ANOVA…

We chose Post Hoc:
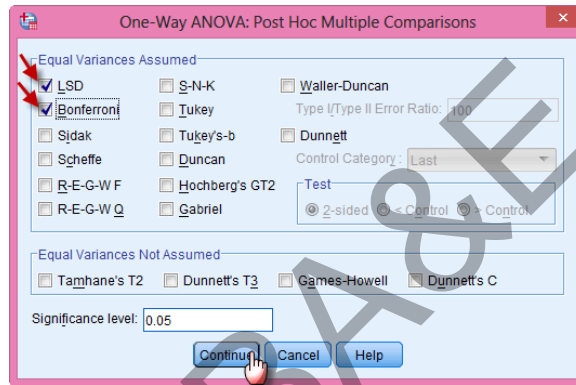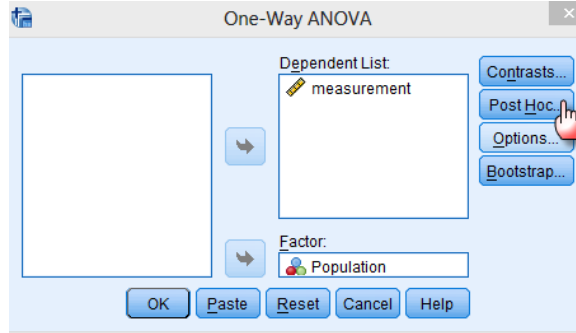
We click on Continue and on OK, to get the following outputs:

SPSS output for multiple range tests

LSD and Modified LSD Tests with 5% level of significance

Multiple Comparisons

Dependent Variable: measurement

	(I) Population	(J) Population	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
LSD	HB SS	HB S/-Thalassaemia	-1.9175*	.4029	.000	-2.733	-1.102
		HB SC	-3.5875*	.3592	.000	-4.315	-2.860
	HB S/-Thalassaemia	HB SS	1.9175*	.4029	.000	1.102	2.733
		HB SC	-1.6700*	.4080	.000	-2.496	-.844
	HB SC	HB SS	3.5875*	.3592	.000	2.860	4.315
		HB S/-Thalassaemia	1.6700*	.4080	.000	.844	2.496
Bonferroni	HB SS	HB S/-Thalassaemia	-1.9175*	.4029	.000	-2.927	-.908
		HB SC	-3.5875*	.3592	.000	-4.487	-2.688
	HB S/-Thalassaemia	HB SS	1.9175*	.4029	.000	.908	2.927
		HB SC	-1.6700*	.4080	.001	-2.692	-.648
	HB SC	HB SS	3.5875*	.3592	.000	2.688	4.487
		HB S/-Thalassaemia	1.6700*	.4080	.001	.648	2.692

*. The mean difference is significant at the 0.05 level.

(a) The result of LSD test all groups are different than each other

(b) The result of Bonferroni's test (Modification of LSD)

Results of Bonferroni test are the same as for LSD.

The following figure (obtained through the Means Plot) may reflect the results:

Means Plots
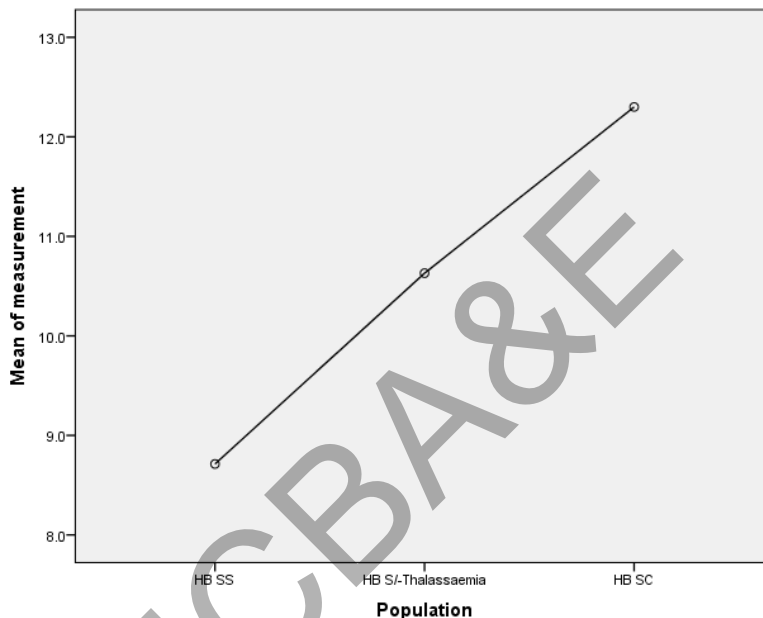
Example 5.3:

Vanadium is recently recognized essential trace element. An experiment was conducted to compare the concentration of vanadium in biological materials using isotope dilution mass spectrometry. The following table gives the quantities of vanadium (measured in nano-grams per gram) in dried samples of oyster tissue, citrus leaves, bovine liver and human serum. Use an appropriate method of analysis to determine whether the distribution of vanadium concentrations for the four biological materials differ in locations. The data is given in Table 5.5. Use 5% level of significance.

Table 5.5

Oyster tissue	Citrus tissue	Bovine lever	Human serum
2.35	2.32	0.39	0.10
1.30	3.07	0.54	0.17
0.34	4.09	0.30	0.14
			0.16
			0.16

(Source: Analytical Chemistry, Vol. 57(13), 1985, pp. 2475).

Solution:

(1) H_0 : There is no difference between the Vanadium concentrations for the four biological materials.

H_1 : At least two differ.

(2) $\alpha = 0.05$

(3) Test-statistic. Analysis of Variance

Before applying the Analysis of Variance, test of Homogeneity is applied whether we can apply this test or not.

Test of Homogeneity of Variances

concentration

Levene Statistic	df 1	df 2	Sig.
3.955	3	10	.043

(4) Since the p-value of the homogeneity of test is less than 0.05, therefore, the condition for the equality of variances is not met. We may not apply Analysis of Variance technique to find out whether there is any difference between concentrations of four groups.

To solve this problem and find out significant difference, we will apply non-parametric method called Kruskal-Wallis. This will be discussed in Chapter 8.

5.3 Analysis of variance for two-Way classification

Suppose there are k treatments (drugs) and b blocks (age groups). If k treatments are compared within each of b blocks k treatments are randomly assigned within each block. This is known as randomized block design. For the analysis of such data, two-way analysis of variance is appropriate. In simple language, if the data are given according to two criteria then analysis of variance for two-way classification is the proper method for analysis. Suppose we want to compare three types of drugs (A, B, C) on patients of different age groups and would like to see how these different types of drugs have an effect on patients of different age groups. To compare k drugs (treatments) on b blocks (age groups), our hypotheses will be:

(a) H_0 : $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$
(i.e. there is no difference in the treatment means)

H_1 : at least two treatments means differ significantly.

(b) H_0 : $\beta_1 = \beta_2 = \ldots = \beta_b$
(i.e. there is no difference among means of the blocks)

H_1 : at least two block means differ significantly.

In the additive model

$$\mathbf{y_{ij}} = \mathbf{\mu_{ij}} + \mathbf{\varepsilon_{ij}} = \mathbf{\mu} + \mathbf{\zeta_i} + \mathbf{\beta_j} + \mathbf{\varepsilon_{ij}} \tag{5.2}$$

$$i = 1,2,3,\ldots,k; \ j = 1,2,3,\ldots,b; \ n = bk,$$

where $\Sigma \, \xi_i = 0$ and $\Sigma \, \beta_j = 0$, where ξ_i is the net effect of the ith drug and β_j is the net effect of the jth age group and $\varepsilon_{ij} \sim \mathbf{NID} \ (\mathbf{0, \sigma^2})$

The assumptions are as:

(i)   The population distribution of the difference between pairs of treatment observations within a block is approximately normal.

(ii)   (ii)   The variance of the probability distributions is constant and same for all pairs of observations.

(iii)   (iii) The treatments (drugs) are randomly assigned to the experimental units (age) within each block.

When the assumptions are violated, an alternative technique known as Friedman's test (Chapter-8) may be used instead of ANOVA.

Like one way analysis of variance, two-way analysis of variance partitions the total sum of squares (TSS) into three components i.e. treatment sum of squares [SS(T)]; Block sum of squares [SS(B)]; and error sum of squares (SSE). This is shown in Fig. 5.2.

We compare the three sources of variation by the F-statistic

F_1 = (mean squares treatments)/(Mean squares Error)

F_2 = (mean squares blocks)/(Mean squares error).

If there are k treatments and b blocks then F_1 is based on $v = k-1$ and $v = (b-1)(k-1)$ degrees of freedom whereas F_2 is based on $v = (b-1)$ and $v = (b-1)(k-1)$ degrees of freedom.
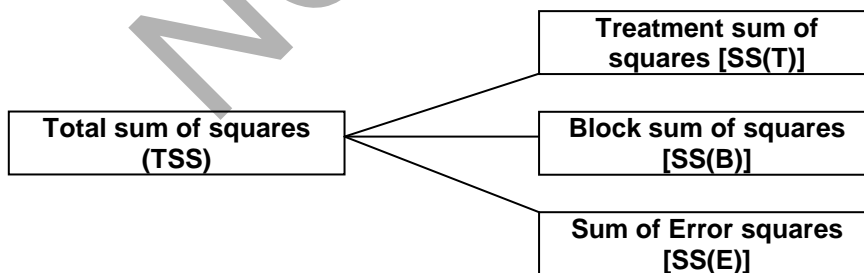
Fig. 5.2: Partitioning of total sum of squares into different components

The results of the analysis of variance two- way classification are usually summarized and presented in an analysis of variance (ANOVA) table and this table shows the sources of variation, their respective degrees of freedom, sum of squares, mean sum of squares and F-statistics. If SPSS package is used the p-value also appears in the table. If there are k treatments and b blocks then the output is displayed as:

Table 5.6
Two-way ANOVA

Sources of variations	d.f	SS	MSS	F(cal)
Between treatments	k-1	SS(T)	SS(T)/(k-1)=MST	MST/ MSE
Between blocks	b-1	SS(B)	SS(B)/(b-1)=MSB	MSB/ MSE
Errors	(k-1)(b-1)	SS(E)	SSE/ (k-1)(b-1)=MSE	
Total	Nk–1 = n-1	TSS		

Example 5.4:

The pharmaceutical project manager decides to replicate the study, comparing the ACC inhibitors, grouping the subject into blocks on the basis of age. It is known that age affects systolic blood pressure systematically. The data regarding age and the use of the drug are as in Table 5.7

Table 5.7:
Measurements of systolic blood pressure

Age	Drug A	Drug B	Drug C	Total
20 - 30	100	90	110	300
30 - 40	105	80	90	275
40 - 50	95	80	80	255
50 - 60	110	75	100	285
> 60	90	90	95	275
Total	500	415	475	1390

Use the analysis of variance technique to find difference between the effect of drugs and between age groups.

Solution:

Here the data is given according to two criteria, i.e. use of drugs and age groups. Two-way ANOVA technique is applied to see the difference between drugs and between age groups. Our null and alternative hypotheses are:

(a) H_o: All the drugs are equally effective.

   H_1: At least there is difference between two drugs.

(b) H_o : There is no difference in age groups.

   H_1 : At least there is difference between two age groups.

(2) $\alpha = 0.05$

(3) Test-statistic: F-test in Two-way ANOVA

For calculation we proceed as:

(i) Correction factor = $(1390)^2 /15 = 128806.667$

(ii) Total sum of squares = $100^2 + 105^2 + \ldots + 95^2 - 128806.667 = 1693.333$

(iii) Sum of squares of treatments.(drugs)

$$= \frac{500^2 + 415^2 + 475^2}{5} - 128806.667 = 763.333$$

(iv) Sum of squares of blocks.(age groups)

$$= \frac{300^2 + 275^2 + \ldots + 275^2}{3} - 128806.667 = 360.0$$

Note that the divisor is the number of observations in which the totals are based.

(v) Sum of squares of residuals.(error) = TSS – SS(T)- SSB

$$= 1693.33 -763.33 - 360.0 = 570.0$$

This can be presented in the standard ANOVA table

Table 5.8
ANOVA two-way

Source of Variation	Df	SS	MSS	F(cal)	F(tab) 5%
Between Drugs	2	763.333	(763.333)/2= 381.667	381.667/71.25 = 5.357	4.46
Between Age	4	360.00	(360)/4= 90.00	90.00/71.25 = 1.263	3.81
Error	2×4=8	570.00	570.0/8=71.25		
Total	14	1693.33			

Interpretation

i) **Between drugs**: $F_{cal} = 5.357$ for drugs whereas F_{tab} (2,8) = 4.46. The calculated value is more than the table value therefore at 5% level of significance we do not accept the hypothesis and say with 95% confidence that effect of at least two drugs is not the same.

ii) **Between Age groups:** $F_{cal} = 1.263$ for blocks; F_{tab} = (4,8) = 3.81. The calculated value is less than table value therefore at 5% level of significance we accept the hypothesis and say that the effect of drugs on all the age groups is the same.

Example S5-3

To test that the null hypotheses:

(a) H_o: All the drugs are equally effective.

H_1: At least there is difference between two drugs.

(b) H_o : There is no difference in age groups.

H_1 : At least there is difference between two age groups.

for the data given in table 5.7,

the data are entered in one column and we add another two grouping variables with the numbers 1,2 and 3 corresponding to the 3 (independent) drug groups, and the numbers 1,2,3,4 and 5 corresponding to the 5 (independent) age groups, the data and the value labels will look as :
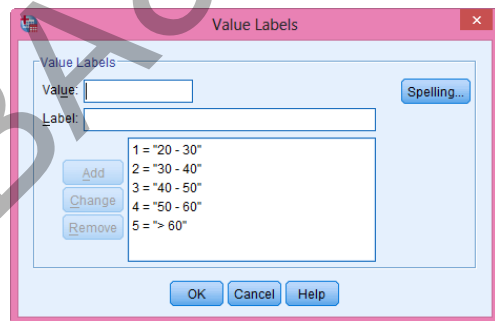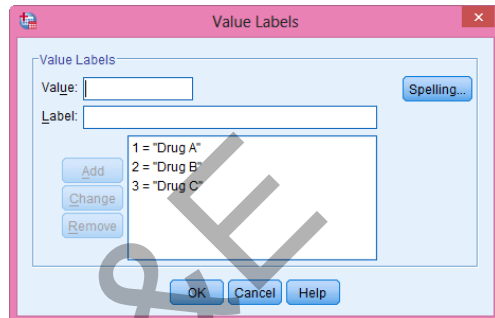
Note that, since the data for each drug doesn't appear more than once with respect to each age group, we say that there is "no interaction".

Now, we follow the following steps :

Analyze→General Linear Model→Univariate…
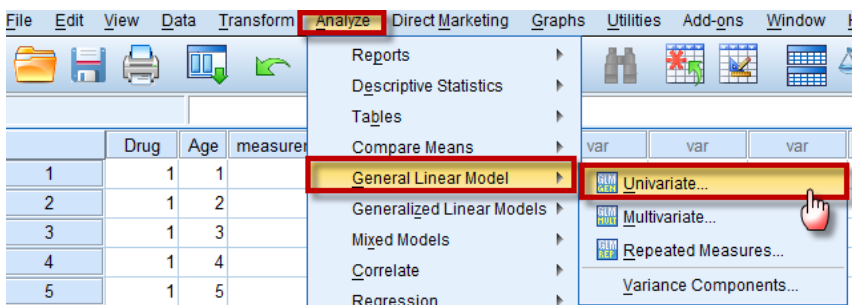
Move the variable measurement to Dependent Variable:

Move the grouping variables to Fixed Factor(s):
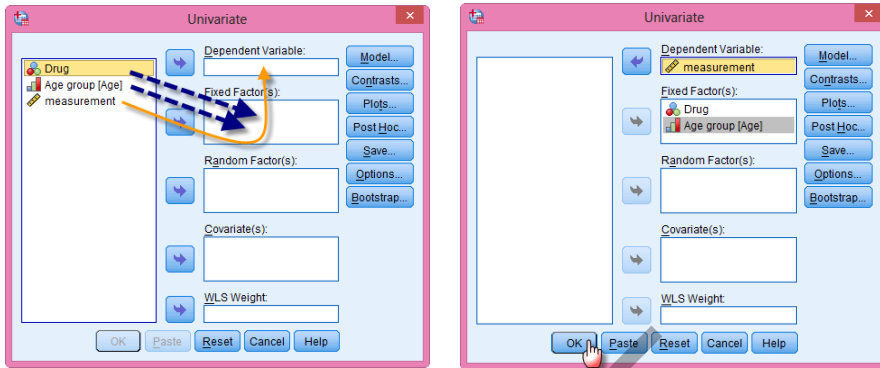
We click on **Model...**, chose Custom, click on Type and chose Main effects to remove the interaction option (also we remove the "include intercept model"):
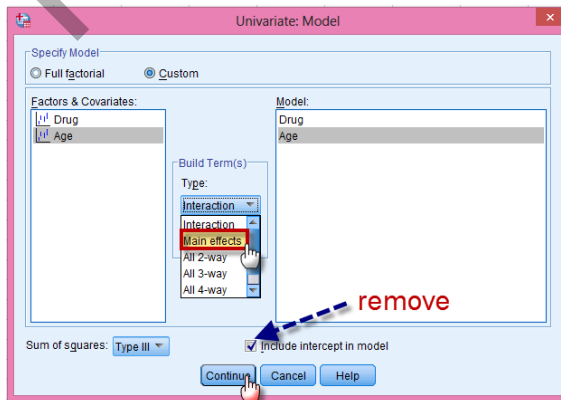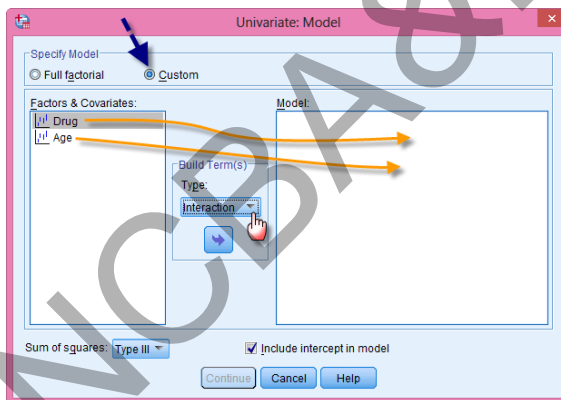
We click on [Continue] and on [OK], to get the following output:

SPSS output for ANOVA Two-way Classifications

Tests of Between-Subjects Effects

Dependent Variable: measurement

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	129930.000[a]	7	18561.429	260.511	.000
Drug	763.333	2	381.667	5.357	.033
Age	360.000	4	90.000	1.263	.360
Error	570.000	8	71.250		
Total	130500.000	15			

a. R Squared = .996 (Adjusted R Squared = .992)

Interpretation

(i) Calculated p-value for the age group is 0.360 which is more than 0.05, therefore, result is non-significant and we say with 95% confidence that there is no difference in the systolic blood pressure on age groups regarding the effect of drugs.

(ii) The p-value of the drug is 0.033, which is less than 0.05; therefore the result is significant, we say with confidence that there is a significant difference in the effect of drugs. At least two of the drugs do not have the same effect.

POST HOC Test: Since there is a significant difference w.r.t. the Drug, we have to apply a Post Hoc test, say LSD test to see the differences between each two drug means:

Analyze→General Linear Model→Univariate…

We chose Post Hoc:

We move the variable Drug only
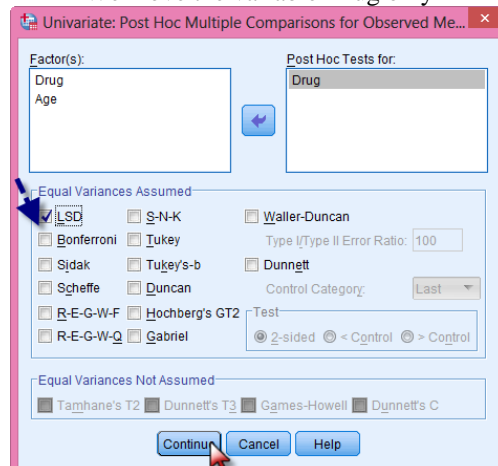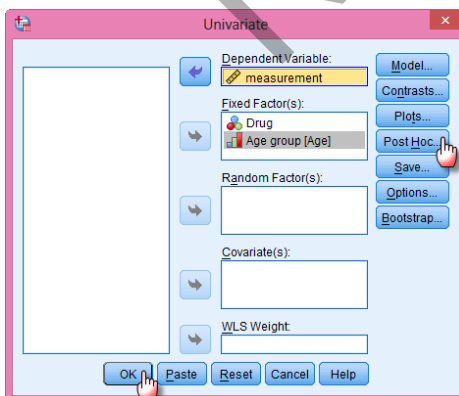
We click on `Continue` and on `OK`, to get the following outputs:

Multiple Comparisons

Dependent Variable: measurement

LSD

(I) Drug	(J) Drug	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Drug A	Drug B	17.00*	5.339	.013	4.69	29.31
	Drug C	5.00	5.339	.376	-7.31	17.31
Drug B	Drug A	-17.00*	5.339	.013	-29.31	-4.69
	Drug C	-12.00	5.339	.055	-24.31	.31
Drug C	Drug A	-5.00	5.339	.376	-17.31	7.31
	Drug B	12.00	5.339	.055	-.31	24.31

The result of LSD test show that the difference is between Drug A and Drug B only

Example 5.5:

Sixteen overweight females participated in a study to compare four types of diets for weight reduction. Females were grouped according to initial weight and randomly distributed to one of the four types of diets. At the end of the experiment the following weight losses in pounds were recorded.

Table 5.9:
Type of diet and weight loss in pounds

Initial weight(pounds)	Diet 1	Diet 2	Diet 3	Diet 4
150-174	12	26	24	23
175-199	15	29	23	25
200-225	15	27	25	24
> 225	16	38	33	31

After eliminating differences due to initial weight, do these data provide evidence to indicate that there is no difference in different types of diets?

Solution:

This is a question of randomized block design where types of diet are treatments and initial weight groups are blocks, therefore two-way analysis of variance technique is applied to see the difference in different types of diet. Our null and alternative hypotheses are:

H_0 : there is no difference in the types of diet.

H_1 : at least there is a difference in two types of diet.

The SPSS package is used to solve this problem and the output is as:

SPSS output for ANOVA two-way classification

Tests of Between-Subjects Effects

Dependent Variable: Type of Diet and Weight Loss in Pounds

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	698.500	6	116.417	26.694	.000
Intercept	9312.250	1	9312.250	2135.293	.000
Weight	161.250	3	53.750	12.325	.002
Diet	537.250	3	179.083	41.064	.000
Error	39.250	9	4.361		
Total	10050.000	16			
Corrected Total	737.750	15			

Interpretation

Diet: The p-value $= 0.000$ which is less than 0.05, therefore the result is significant. We say with 95% confidence that effect of the types of diet in reducing the weight is not the same.

Weight: The p-value $= 0.002$ which is less than 0.05, the result is significant. We say with 95% confidence that the effect of diets has significant effect in weight losses.

5.4 Repeated Measure Design or Repeated Measure Analysis of Variance

In Chapter 4, analysis was made when each subject was measured twice by using t-test for paired observations. Repeated measure design is an extension of this problem. *Any design involving successive measurements on the same subject is called a repeated measures design.* In a repeated measures designs, units are subject to repeated measures; for example blood pressures may be measured at successive intervals, say, once a week, for a group of patients attending a clinic. In this design, measurements on the same variable are made on two or more different occasions. Such data can be collected either prospectively, following subjects forward in time, or retrospectively, by extracting measurements on each person from historical records.

In repeated measure design, each subject acts as its own control. This helps to control the variability between subjects since the same subject is measured repeatedly. This design has the ability to control for extraneous variation among subjects. *Of course, when repeated measures are taken in different time sequences, it is not possible to include randomization.*

There are four important classes of repeated measures studies i.e. split-plot experiment; longitudinal studies, changeover studies and sources of variability studies. Because of the limited scope of this book it is not possible to discuss all of these. Some examples of longitudinal studies and changeover studies are given. In health sciences one can face such types of examples.

1. Two treatments for chronic pain are randomly assigned to subjects, and the extent of pain relief is evaluated at weekly visits for six weeks.

2. Boys and girls from a cohort of one year old are observed every six months for five years to assess their ability to perform a manual dexterity task (or measurements of height, weight, or physical fitness might be made)

3. Two treatments for a dental problem are randomly assigned to children. The status of teeth on the upper and lower jaws is evaluated every three months for one year.

4. Information on smoking is obtained for each subject by two different methods: one was the subject's self-report to a direct question and the other is biochemical determination based on carbon monoxide levels in the blood. Subjects are randomly assigned to one of the two sequence groups: for one group, the self-report preceded the biochemical determination: and for the second group, the self-report followed the biochemical determination.

5. The relative potency of two drugs that influence cardiovascular function is assessed through a changeover design. Volunteers are randomly assigned to one of two sequence groups. One group receives drug A during the first six-week study period and drug B during the second, and the other group receives the opposite regimen. A two -week washout period separates the two-treatment period. During each treatment period, three doses of the drug are tested with the drug dose being successively increased every two weeks. At the beginning of the treatment and at the end of each two-week dose interval, heart rate is measured before and after a treadmill exercise test.

Most of the times health scientists use single-factor repeated measure design. This can be easily extended to two or more factors.

Before we proceed for the discussion of repeated measure design it is necessary to explain the concept of *Sphericity*. Sphericity refers to the equality of variance of the difference between treatment levels. So, if you were to take each pair of treatment levels, and to calculate their differences, then it is necessary that these differences have equal variance. For any data, sphericity will hold when:

$$\text{Variance}_{A-B} = \text{Variance}_{A-C} = \text{Variance}_{B-C}$$

Assumptions of sphericity must hold; in other words we assume that relationship between pairs of experimental conditions is similar i.e. level of dependence between expert, mental conditions is roughly similar. This assumption is called the assumption of sphericity. Sphericity is denoted by ε. This can be tested by Mauchly's test. If the p-value of Mauchly's test is less than 0.05 we say that there is a significant difference between the variances of difference of each pair and say that condition of sphericity is not met. If p-value of Mauchly's test is greater than 0.05 we say that variance of difference are equal. Violation of the sphericity assumption makes the usual F-test inaccurate. We can use the corrected value of F by using either of the methods given by Greenhouse – Geisser (1959), Huynh- Feldt (1976) and lower bound (Milliken and Johnson-1984) for decision or multivariate analysis technique can be used. All these methods are given in SPSS. If the condition of sphericity does not hold then we look into the p-values of

Greenhouse–Geisser and Huynh–Feldt and take the average of these two. If the two corrections give rise to the same conclusion it makes little difference, which method you chose to draw inference.

The additive model for fixed- effect single factor repeated measure design is

$$y_{ij} = \mu + \zeta_i + \beta_j + \varepsilon_{ij} \quad i = 1, 2, 3,\ldots, k, j = 1, 2, 3,\ldots, b. \tag{5.3}$$

where μ denotes over all mean. Also ξ_i is the net effect of ith treatment and β_j is the net effect of jth block. $\varepsilon_{ij} \sim \textbf{NID } (\textbf{0}, \sigma^2).$

The simplest repeated measure design is one in which, in addition to the treatment variable, one additional variable is considered. This is known as single- factor repeated measure design.

Example 5.6:

The purpose of the study is to determine the pharmacokinetics of phenytoin in the presence and absence of concomitant fluconazole therapy. Blum et al. (1991) collected the data (reproduced below in Table 5.10) during the course of the study on trough serum concentration fluconazole for 10 healthy males at different points in time. By using a method of repeated measure design analyze the data and see if at different times there is any significant difference in the mean serum concentration of fluconazole.

Table 5.10:
Data relating to mean serum concentration of fluconazole

	Day 14C_{min}(μg/ml)	Day 18C_{min}(μg/ml)	Day 21C_{min}(μg/ml)
1	8.28	9.55	11.21
2	4.71	5.05	5.20
3	9.48	11.33	8.45
4	6.04	8.08	8.42
5	6.02	6.32	6.93
6	7.34	7.44	8.12
7	5.86	6.19	5.98
8	6.08	6.03	6.45
9	7.50	8.04	6.26
10	4.92	5.28	6.17

Solution:

(1) $H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : At least two differ

(2) $\alpha = 0.05$

(3) Test Statistic; Repeated Measure Design

Now, to perform the analysis we enter above data in IBM-SPSS just like the paired samples t-test (but more than two variables) and proceeds are as under:

Example S5-4

The data will look as:

	serial	Day14	Day18	Day21
1	1	8.28	9.55	11.21
2	2	4.71	5.05	5.20
3	3	9.48	11.33	8.45
4	4	6.04	8.08	8.42
5	5	6.02	6.32	6.93
6	6	7.34	7.44	8.12
7	7	5.86	6.19	5.98
8	8	6.08	6.03	6.45
9	9	7.50	8.04	6.26
10	10	4.92	5.28	6.17

Now, we follow the following steps:

Analyze→General Linear Model→Repeated Measures…
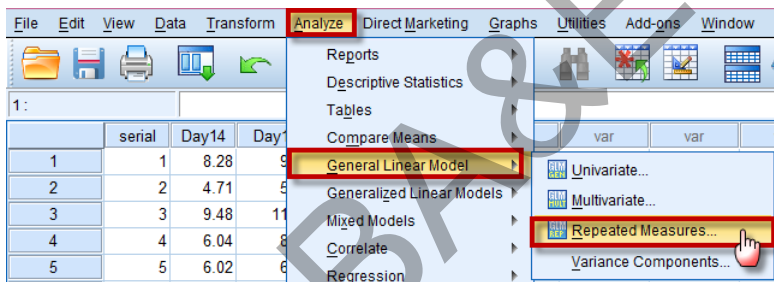
We Enter Within-Subject Factor Name "day" (By default it is "factor 1")
Number of Levels in this case are "3", then click on Add.
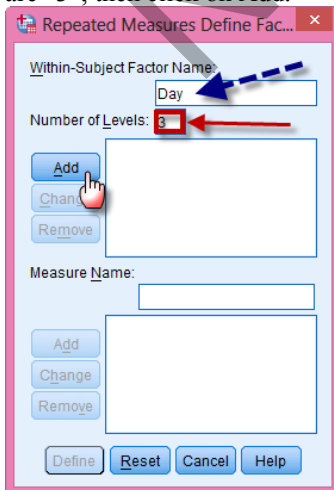
We click [Define]

A new dialogue box will be opened named as (Repeated Measures), Select all the 3 variables and bring them to the right side in the space available for Within-Subjects Variables (day).
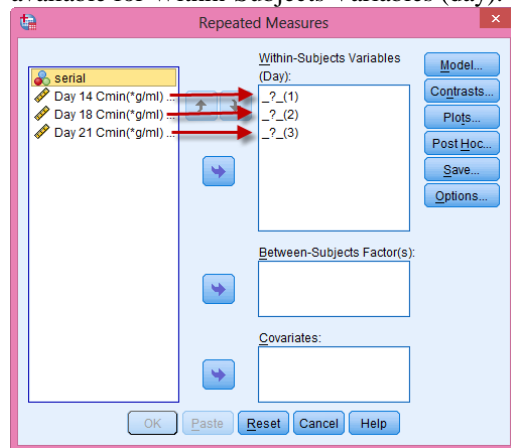
We click on [OK], to get the following outputs:

SPSS Output for Repeated Measures Design

Mauchly's Test of Sphericity

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
DAY	.645	3.513	2	.173	.738	.848	.500

(4) Since the calculated p-value of Mauchly's test of sphericity is 0.173 for 5% significance level, which is more than 0.05 therefore assumption of Sphericity is met.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
DAY	Sphericity Assumed	3.286	2	1.643	2.429	.116
	Greenhouse-Geisser	3.286	1.476	2.227	2.429	.135
	Huynh-Feldt	3.286	1.695	1.938	2.429	.127
	Lower-bound	3.286	1.000	3.286	2.429	.154
Error(DAY)	Sphericity Assumed	12.176	18	.676		
	Greenhouse-Geisser	12.176	13.280	.917		
	Huynh-Feldt	12.176	15.258	.798		
	Lower-bound	12.176	9.000	1.353		

Looking into the above table we can interpret the result

(5) The calculated p-value is 0.116 which is greater than 0.05. We conclude that there is insignificant difference in the mean serum concentration of fluconazole, taken at different time.

***Example* 5.7:**

A group of students investigated the consistency of marking by submitting the same assignments to four different tutors. The marks given by each tutor was recorded for each of the eight assignments. Data for the assignments marks is given in table 5.11.

Table 5.11

Assignments	Tutor 1	Tutor 2	Tutor 3	Tutor 4
1	62	58	63	64
2	63	60	68	65
3	65	61	72	65
4	68	64	58	61
5	69	65	54	59
6	71	67	65	50
7	78	66	67	50
8	75	73	75	45

Solution:

(1) H_0 : On the average all the four tutors are equal in marking the assignments.

   H_1 : At least two differ

(2) α : 0.05

(3) Test Statistic : Repeated Measure Design

The output for the repeated measures design using IBM-SPSS package is given as follows.

Mauchly's Test of Sphericity[b]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
TUTOR	.131	11.628	5	.043	.558	.712	.333

(4) The calculated p-value of Mauchly's test of sphericity is 0.043, which is less than 0.05 therefore assumption of sphericity is violated.

Looking into the following table we can interpret the results.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TUTOR	Sphericity Assumed	554.125	3	184.708	3.700	.028
	Greenhouse-Geisser	554.125	1.673	331.245	3.700	.063
	Huynh-Feldt	554.125	2.137	259.329	3.700	.047
	Lower-bound	554.125	1.000	554.125	3.700	.096
Error(TUTOR)	Sphericity Assumed	1048.375	21	49.923		
	Greenhouse-Geisser	1048.375	11.710	89.528		
	Huynh-Feldt	1048.375	14.957	70.091		
	Lower-bound	1048.375	7.000	149.768		

(5) Since the assumption of Sphericity is violated at 5% level of significance then according to the suggestion given by Stevens (1992) we have to check the p-values for the Greenhouse-Geisser test and Huynh-Feldt test simultaneously. The above table gives the calculated p-values for these two tests, which are 0.063 and 0.047 respectively. In this example, one interesting thing is that both these p-values do not lead to the same conclusion because calculated p-value for the Greenhouse-Giesser is 0.063 which is greater than 0.05 but the calculated p-value for the Huynh-Feldt is 0.047 that is less than 0.05, so as suggested by Stevens (1992) the average value of these two p-values should be taken which comes out to be $\dfrac{0.063 + 0.047}{2} = 0.055$, which is more than 0.05, so we choose the results of Greenhouse-Geisser and say that at 95% confidence level there is no significant difference among four tutors in marking the assignment of the students.

If the condition of the sphericity is violated then other way is to go to Multivariate Analysis of Variance (MANOVA):

Example S5-5

The data will look as :

	Assignments	Tutor1	Tutor2	Tutor3	Tutor4
1	1	62	58	63	64
2	2	63	60	68	65
3	3	65	61	72	65
4	4	68	64	58	61
5	5	69	65	54	59
6	6	71	67	65	50
7	7	78	66	67	50
8	8	75	73	75	45

Now, we follow the following steps :

Analyze→General Linear Model→Repeated Measures…
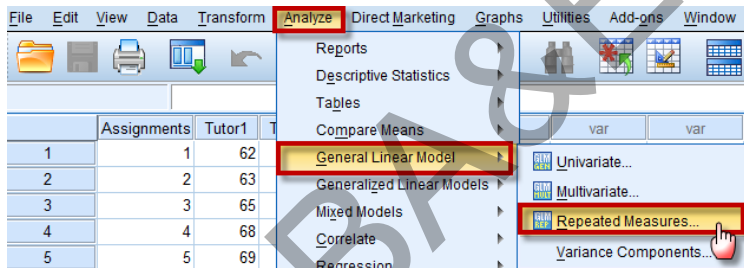
We Enter Within-Subject Factor Name "Tutor" Number of Levels in this case are "4", then click on Add.
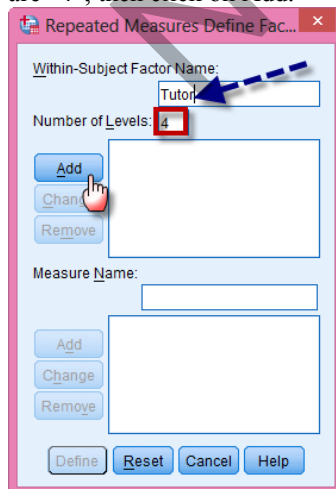
We click Define

A new dialogue box will be opened named as (Repeated Measures), Select all the 4 variables and bring them to the right side in the space available for Within-Subjects Variables (Tutor). Click on Options:
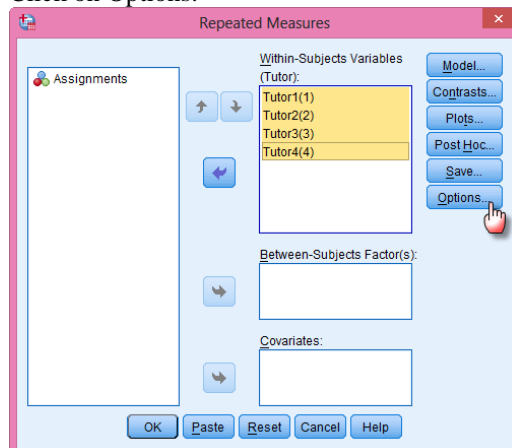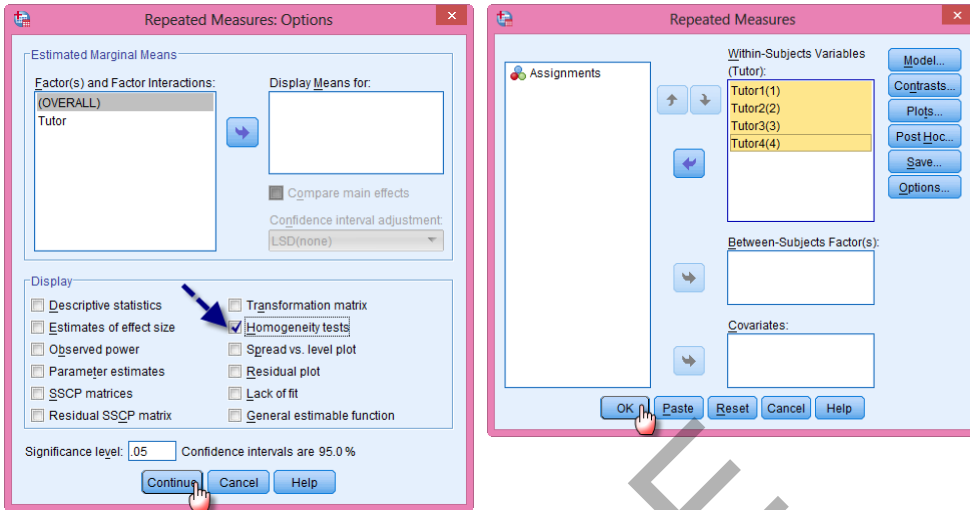
We click on OK, to get the following outputs:

SPSS Output for MANOVA

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Tutor	Pillai's Trace	.741	4.760[b]	3.000	5.000	.063
	Wilks' Lambda	.259	4.760[b]	3.000	5.000	.063
	Hotelling's Trace	2.856	4.760[b]	3.000	5.000	.063
	Roy's Largest Root	2.856	4.760[b]	3.000	5.000	.063

a. Design: Intercept
   Within Subjects Design: Tutor

b. Exact statistic

(6) In the output for Multivariate Tests four kinds of tests are used to test the significance of the model, but Wilks' Lambda is more powerful and is frequently used. The p-value of Wilks' Lambda is 0.063 which is greater than 0.05. We can say with 95% confidence that there is no significant difference in evaluating the assignments of the students.

(the other outputs are the same given in Example 5.7)

Multiple Comparison test:

If the null hypothesis is not accepted we may use the multiple comparison test to see which two groups differ. The procedure for the multiple comparison tests is as:

- Click *defines* and then click *options*.
- Bring the factor name on the right side *(Display Means for)*.
- Click *compare* main effects.
- Click *confidence* interval estimation (Bonferroni Test).
- Click *Continue*.
- Click *ok*.

5.5 Multivariate Analysis of Variance (MANOVA)

In previous sections we have studied the methods to compare several groups; each measured on single variable of interest; by using simple and repeated measures ANOVA. There are several situations where we have to compare several groups; each measured on more than one variable. For example we may be interested in comparing the effectiveness of four medicines when reduction in blood pressure level and increase in sugar level is obtained after applying each medicine. In these situations simple or repeated measures ANOVA does not solve the problem and we have to use the technique known as Multivariate Analysis of Variance (MANOVA). The MANOVA technique is used to compare several groups and each group constitute several variables. In MANOVA the hypothesis of preliminary interest is that mean vectors of several groups are equal. Before we proceed for procedure to carry out MANOVA in SPSS it is worthwhile to discuss its assumptions. These assumptions are given as under.

5.5.1 Assumptions

The following assumptions must hold for applying MANOVA.

1. Samples must be random.

2. Condition of normality must hold.

3. Errors covariance should be equal across various groups; [test of Homogeneity (Box's Test)].

4. Condition of additivity must hold.

5. Condition of sphericity (Bartlett's test) may not hold.

6. There should be several dependent variables.

Example 5.8:
Forty-five patients suffering from cancer were given the radiation therapy and the effects were recorded. The patients were grouped in four groups. The average score for the first three days following radiation therapy are given below Test the null hypothesis that four radiations therapy have equal average score for three days.

Table 5.12

Control			25-50R			75-100R			125-250R		
1	2	3	1	2	3	1	2	3	1	2	3
223	214	224	60	95	103	216	187	239	198	245	237
72	80	65	45	45	76	210	176	139	167	259	185
172	175	170	95	95	98	206	218	276	158	168	196
180	175	165	175	175	167	198	225	216	176	168	244
195	200	185	203	203	218	198	203	203	187	217	224
35	25	25	191	191	116	118	181	198	260	234	238
			114	114	123	248	245	187	214	267	265
			35	35	76	260	206	167	216	248	265
			55	55	45	95	116	214	234	248	259
			106	106	121	238	214	255	158	269	268
			264	264	216	234	243	167			
			210	210	216	95	103	34			
			34	34	56	134	147	168			
			255	255	270	136	138	234			
						98	89	201			

Solution:

(1) H_0 : On the average radiation therapy has equal effect on the four groups

H_1 : At least two groups differ

(2) $\alpha = 0.05$

(3) Test Statistic : Multivariate Analysis of Variance.

Example S5-6

The data will look as:

	Group	Day1	Day2	Day3
1	1	223	214	224
2	1	72	80	65
3	1	172	175	170
4	1	180	175	165
5	1	195	200	185
6	1	35	25	25
7	2	60	95	103
8	2	45	45	76
9	2	95	95	98
10	2	175	175	167
11	2	203	203	218
12	2	191	191	116
13	2	114	114	123
14	2	35	35	76
15	2	55	55	45
16	2	106	106	121
17	2	264	264	216
18	2	210	210	216
19	2	34	34	56
20	2	255	255	270
21	3	216	187	239
22	3	210	176	139
23	3	206	218	276
24	3	198	225	216
25	3	198	203	203
26	3	118	181	198
27	3	248	245	187
28	3	260	206	167
29	3	95	116	214
30	3	238	214	255
31	3	234	243	167
32	3	95	103	34
33	3	134	147	168
34	3	136	138	234
35	3	98	89	201
36	4	198	245	237
37	4	167	259	185
38	4	158	168	196
39	4	176	168	244
40	4	187	217	224
41	4	260	234	238
42	4	214	267	265
43	4	216	248	265
44	4	234	248	259
45	4	158	269	268

the steps for applying MANOVA are as:-

Analyze→General Linear Model→ Multivariate …
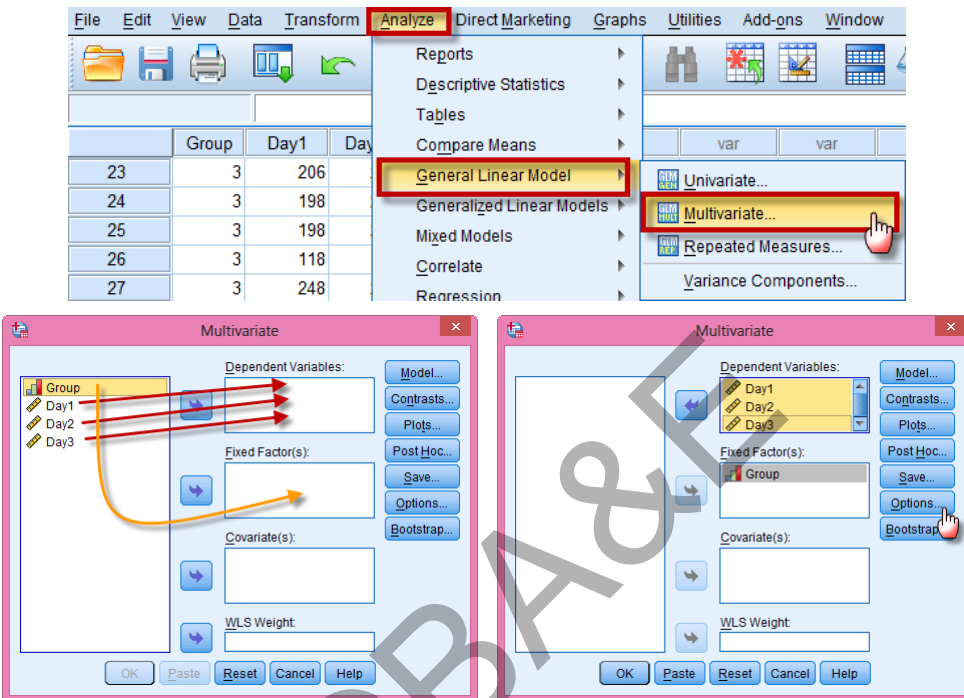
(i) Click *Option* then click Homogeneity of variance

(ii) For multiple Comparison, bring the code in right window, then click compare main effect finally choose the method for comparison

We click on OK , to get the following outputs:

SPSS OUTPUT OF MANOVA

Box's Test of Equality of Covariance Matrice

Box's M	71.735
F	3.323
df 1	18
df 2	1962.936
Sig.	.000

Bartlett's Test of Sphericity [a]

Likelihood Ratio	.000
Approx. Chi-Square	105.558
df	5
Sig.	.000

(4) Since the p-value of Bartlett's Test of sphericity is less than 0.05 therefore MANOVA can be applied.

Multivariate Tests[d]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.901	118.612[b]	3.000	39.000	.000
	Wilks' Lambda	.099	118.612[b]	3.000	39.000	.000
	Hotelling's Trace	9.124	118.612[b]	3.000	39.000	.000
	Roy's Largest Root	9.124	118.612[b]	3.000	39.000	.000
CODES	Pillai's Trace	.517	2.843	9.000	123.000	.004
	Wilks' Lambda	.522	3.234	9.000	95.066	.002
	Hotelling's Trace	.842	3.524	9.000	113.000	.001
	Roy's Largest Root	.745	10.175[c]	3.000	41.000	.000

(5) Since the p-value of Wilks' Lambda is less than 0.05, therefore there is a significant difference between groups, regarding the effect of radiation therapy.

(6) Multiple comparison test can be performed to see which groups differ. One can see the p-value (sig.), if it is less than 0.05 for any pair then these two groups differ.

Example 5.9:

Thirty individuals were randomly assigned to three different exercise types viz. at rest, walking leisurely and running. Each group was given two different types of diets; low-fat and high-fat. The pulse rate of these individuals was recorded at three different times during their exercise. The data obtained is given in table 5.14:

Table 5.14

		Low-Fat			High-Fat		
		1 minute	15 minute	30 minute	1 minute	15 minute	30 minute
Rest		85	85	88	83	83	84
		90	92	93	87	88	90
		97	97	94	92	94	95
		80	82	83	97	99	96
		91	92	91	100	97	100
Walking Leisurely		86	86	84	84	86	89
		93	103	104	103	109	90
		90	92	93	92	96	101
		95	96	100	97	98	100
		89	96	95	102	104	103
Running		93	98	110	95	126	143
		98	104	112	100	126	140
		98	105	99	103	124	140
		87	132	120	94	135	130
		94	110	116	99	111	150

Solution:

The data entry is explained on next page.

    (1) H_{01} : On the average the pulse rate is equal at various time.

       H_{02} : On the average the pulse rate is equal for various exercises.

       H_{03} : On the average the pulse rate is equal for various diets.

       H_1 : At least two groups differ

    (2) $\alpha = 0.05$

    (3) Test Statistic: MANOVA Repeated Measure Design (Between and Within Effects)

Example S5-7

The data will look as:

Table 5.15

	Exertype	Diet	Time1	Time2	Time3						
1	1	1	85	85	88	16	2	2	84	86	89
2	1	1	90	92	93	17	2	2	103	109	90
3	1	1	97	97	94	18	2	2	92	96	101
4	1	1	80	82	83	19	2	2	97	98	100
5	1	1	91	92	91	20	2	2	102	104	103
6	1	2	83	83	84	21	3	1	93	98	110
7	1	2	87	88	90	22	3	1	98	104	112
8	1	2	92	94	95	23	3	1	98	105	99
9	1	2	97	99	96	24	3	1	87	132	120
10	1	2	100	97	100	25	3	1	94	110	116
11	2	1	86	86	84	26	3	2	95	126	143
12	2	1	93	103	104	27	3	2	100	126	140
13	2	1	90	92	93	28	3	2	103	124	140
14	2	1	95	96	100	29	3	2	94	135	130
15	2	1	89	96	95	30	3	2	99	111	150

the steps for applying MANOVA are as:-

Analyze→General Linear Model→ Multivariate …

The steps for applying this design are as:

click:-

    a. Against Within Subject Factor Name enter time

    b. Against Number of Levels enter 3 and click [Add] then [Define]

    c. Take variables time1, time2 and time3 to Within Subject Variable box.

    d. Take variables Exertype (exercise) and diet to Between Subject Factor(s) box.
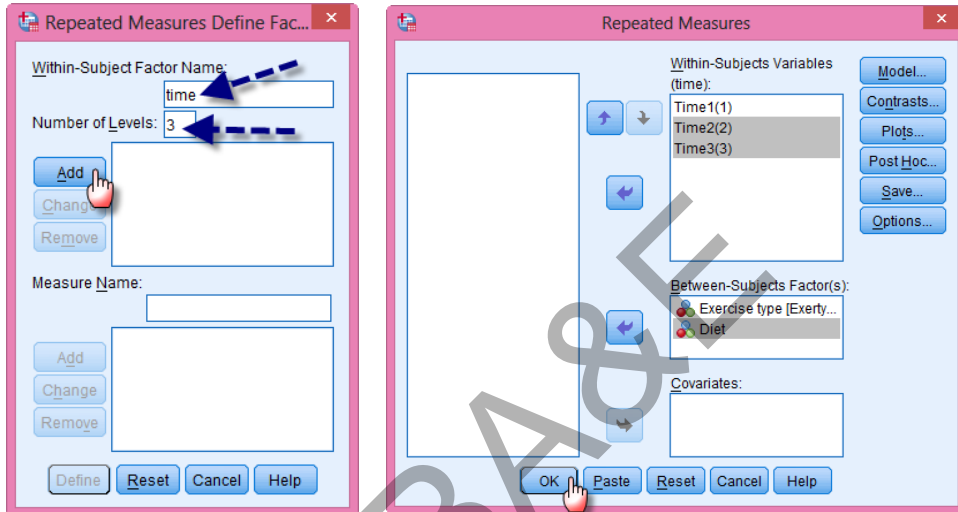
We click on [OK], to get the following outputs:

Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.
TIME	Pillai's Trace	.782	41.209	2.000	23.000	.000
	Wilks' Lambda	.218	41.209	2.000	23.000	.000
	Hotelling's Trace	3.583	41.209	2.000	23.000	.000
	Roy's Largest Root	3.583	41.209	2.000	23.000	.000
TIME * EXERTYPE	Pillai's Trace	.836	8.611	4.000	48.000	.000
	Wilks' Lambda	.172	16.214	4.000	46.000	.000
	Hotelling's Trace	4.762	26.193	4.000	44.000	.000
	Roy's Largest Root	4.753	57.035	2.000	24.000	.000
TIME * DIET	Pillai's Trace	.252	3.865	2.000	23.000	.036
	Wilks' Lambda	.748	3.865	2.000	23.000	.036
	Hotelling's Trace	.336	3.865	2.000	23.000	.036
	Roy's Largest Root	.336	3.865	2.000	23.000	.036
TIME * EXERTYPE * DIET	Pillai's Trace	.518	4.189	4.000	48.000	.005
	Wilks' Lambda	.483	5.047	4.000	46.000	.002
	Hotelling's Trace	1.069	5.881	4.000	44.000	.001
	Roy's Largest Root	1.068	12.819	2.000	24.000	.000

Mauchly's Test of Sphericity

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse -Geisser	Huynh-Feldt	Lower-bound
TIME	.924	1.814	2	.404	.930	1.000	.500

(4) Since the p-value of Wilks' Lambda is less than 0.05 for Time therefore there is significant difference between pulse rate at various exercise time.

(5) Since the p-value of Mauchly's Test is greater than 0.05 therefore the errors are spherical.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
TIME	Sphericity Assumed	2066.600	2	1033.300	31.721	.000
	Greenhouse-Geisser	2066.600	1.859	1111.668	31.721	.000
	Huynh-Feldt	2066.600	2.000	1033.300	31.721	.000
	Lower-bound	2066.600	1.000	2066.600	31.721	.000
TIME * EXERTYPE	Sphericity Assumed	2723.333	4	680.833	20.900	.000
	Greenhouse-Geisser	2723.333	3.718	732.469	20.900	.000
	Huynh-Feldt	2723.333	4.000	680.833	20.900	.000
	Lower-bound	2723.333	2.000	1361.667	20.900	.000
TIME * DIET	Sphericity Assumed	192.822	2	96.411	2.960	.061
	Greenhouse-Geisser	192.822	1.859	103.723	2.960	.066
	Huynh-Feldt	192.822	2.000	96.411	2.960	.061
	Lower-bound	192.822	1.000	192.822	2.960	.098
TIME * EXERTYPE * DIET	Sphericity Assumed	613.644	4	153.411	4.709	.003
	Greenhouse-Geisser	613.644	3.718	165.046	4.709	.004
	Huynh-Feldt	613.644	4.000	153.411	4.709	.003
	Lower-bound	613.644	2.000	306.822	4.709	.019
Error(TIME)	Sphericity Assumed	1563.600	48	32.575		
	Greenhouse-Geisser	1563.600	44.616	35.046		
	Huynh-Feldt	1563.600	48.000	32.575		
	Lower-bound	1563.600	24.000	65.150		

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	894608.100	1	894608.100	10296.660	.000
EXERTYPE	8326.067	2	4163.033	47.915	.000
DIET	1261.878	1	1261.878	14.524	.001
EXERTYPE * DIET	815.756	2	407.878	4.695	.019
Error	2085.200	24	86.883		

(6) Since the errors are spherical therefore the Willk's Lambda statistics is appropriate for testing the significance of various factors.

(7) Since the p-value of exertype and diet are less than 0.05 therefore there is significant difference among pulse rate at various exercise and diet levels.

5.6 Simple Factorial Experiment

An experiment in which two or more factors and each factor at different levels (variables) are investigated is called a factorial experiment. The model for the two-way factorial experiment with interaction is given below.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \tag{5.4}$$

$$i = 1, 2, …, n_{jk}, j = 1, 2, …, m_k, k = 1, 2, …, p$$

The data for a two-factor factorial experiment are presented in a two-way table with rows corresponding to levels of one factor and columns corresponding to levels of another factor.

Example 5.10:

A study was made as to how the concentration of a certain drug in the blood, 24 hours after being injected, is influenced by age(B) and gender(A). An analysis of the blood samples of 40 patients yielded the following concentrations (in milligrams per cubic centimeter).

Table 5.16:
Age groups(B)

			11-25	26-40	41-65	Over 65
			B_1	B_2	B_3	B_4
Gender (A)	Male	A_1	52.0	52.5	53.2	82.4
			56.6	49.6	53.6	86.2
			68.2	48.7	49.8	101.3
			82.5	44.6	50.0	92.4
			85.6	43.4	51.2	78.6
	Female	A_2	68.6	60.2	58.7	82.2
			80.4	58.4	55.9	79.6
			86.2	56.2	56.0	81.4
			81.3	54.2	57.2	80.6
			77.2	61.1	60.0	82.2

(1) test the hypothesis that gender does not affect the blood concentration

(2) test the hypothesis that age does not affect blood concentration

(3) test the hypothesis that there is no interaction between age and gender

Here, there are 4 types of age groups and two types of gender. This experiment involves two factors. Factor "A" has two levels (A_1, A_2) whereas factor "B" has 4 levels (B_1, B_2, B_3, B_4), Each of the 2×4 combinations of this table represent the treatments of the experiment. For this reason the experiment is referred as 2×4 factorial experiment.

In factorial experiment, when the difference between the mean levels of factor "A" depends on the different levels of factor "B", we say that factors A and B interact. If the difference is independent of the levels of "B", then there is no interaction between factors A and B.

Following assumption should be kept in mind while applying factorial experiment

1. The population of the observations for any factor level combination is approximately normal.

2. The variance of the probability distribution is constant and same for the factor level combinations.

3. The treatments, factor level combinations, are randomly assigned to the experimental units.

4. The observations for each factor level combination represent independent random samples.

When the assumptions for the factorial experiment are violated, then we use non-parametric test equivalent to simple factorial experiment.

The hypotheses for the simple factorial experiment are:

(1) Factor A (main effect)

H_0 : there is no difference among the means for main effect " A"

H_1 : At least two of the main effect differ

(2) Factor B (main effect)

H_0 : there is no difference among the means for main effect "B"

H_1 : at least two of the main effect B means differ

(3) Interaction factor (AB)

H_0 : there is no interaction between factors A and B

H_1 : factors A and B interact.

An easy graphical representation is sometimes illuminating and can also throw light on the presence or absence of interaction. Plot levels of one factor on the x-axis and y-axis represent observations. Each line indicates the changes in responses in arrange Y for the different levels of factor A. See Fig. 5.3.
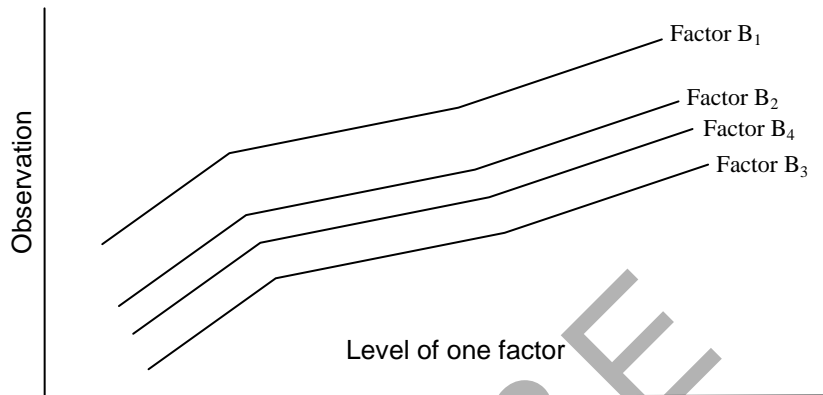
Fig. 5.3: Main effect and interaction plots

A diagram (5.3) showing parallel lines indicates absence of interaction. Intersecting lines estimates presence of interaction etc.

The analysis of variance for the two-factor factorial experiment is very similar to the analysis of variance of two-way classification. The sum of squares of rows and blocks are now replaced by sum of squares of two factors, SS(A), and SS(B), called main effect sum of squares and the interaction sum of squares, SS(AB).

Finally, because we have more than one observation per cell for the two-way table, we calculate a sum of squares of error, called SS(E).

The partitioning of the sum of squares of the total into different components is shown in Figure 5.4.
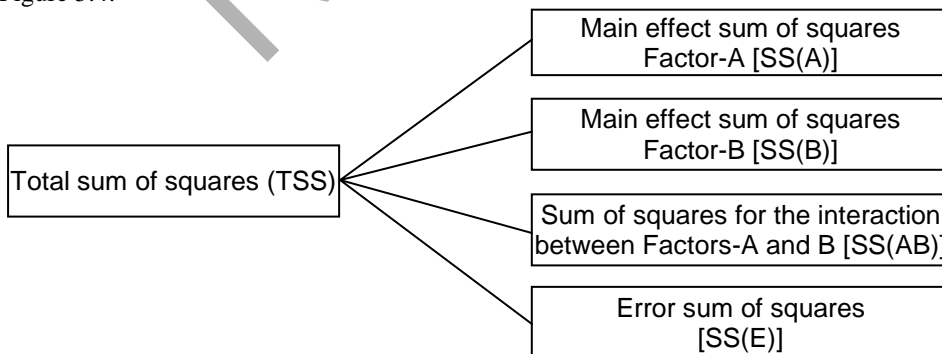
Fig. 5.4: Partitioning of total sum of squares into different components

These results are usually presented in the two-way factorial experiment as:

Table 5.17 ANOVA table for two- factor factorial experiment

Source of Variation	Df	SS	MS	F_cal	F_tab
A	a-1	SS(A)	MS(A)=SS(A)/(a-1)	MS(A)/MSE	
B	b-1	SS(B)	MS(B)=SS(A)/(b-1)	MS(B)/SSE	
2-way interaction A× B	(a-1) (b-1)	SS(AB)	MS(AB) = SS(AB)/(a-1)(b-1)	MS(AB)/MSE	
Residual (error)	ab(r-1)	SSE	MSE = SSE/[ab(r-1)]		
Total	abr-1=n-1	TSS			

r = replication, n= a ×b× r.

Note that in running the SPSS package, one should follow exactly the same procedure as has been suggested by ANOVA two-way classification except that two ways interaction should be clicked instead of no interaction.

Example S5-8

The data will look as:

Table 5.18

	Gender	Age	ID	Concentration
1	1	1	1	52.0
2	1	1	2	56.6
3	1	1	3	68.2
4	1	1	4	82.5
5	1	1	5	85.6
6	1	2	1	52.5
7	1	2	2	49.6
8	1	2	3	48.7
9	1	2	4	44.6
10	1	2	5	43.4
11	1	3	1	53.2
12	1	3	2	53.6
13	1	3	3	49.6
14	1	3	4	50.0
15	1	3	5	51.2
16	1	4	1	82.4
17	1	4	2	86.2
18	1	4	3	101.3
19	1	4	4	92.4
20	1	4	5	78.6
21	2	1	1	68.6
22	2	1	2	80.4
23	2	1	3	86.2
24	2	1	4	81.3
25	2	1	5	77.2
26	2	2	1	60.2
27	2	2	2	58.4
28	2	2	3	56.2
29	2	2	4	54.2
30	2	2	5	61.1
31	2	3	1	58.7
32	2	3	2	55.9
33	2	3	3	56.0
34	2	3	4	57.2
35	2	3	5	60.0
36	2	4	1	82.2
37	2	4	2	79.6
38	2	4	3	81.4
39	2	4	4	80.6
40	2	4	5	82.2

the steps for applying the two-way factorial experiment are as:-

Analyze→General Linear Model→ Univariate …
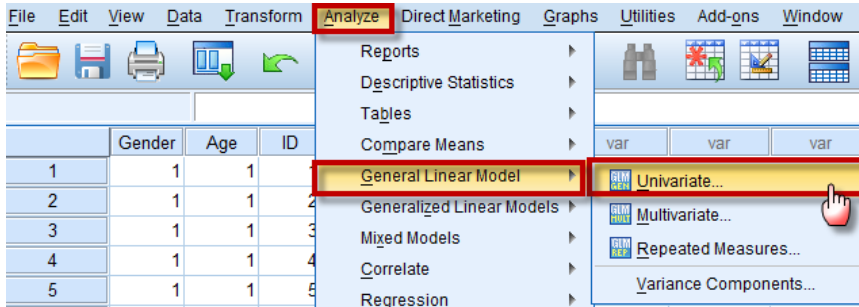
The steps for applying this design are as:
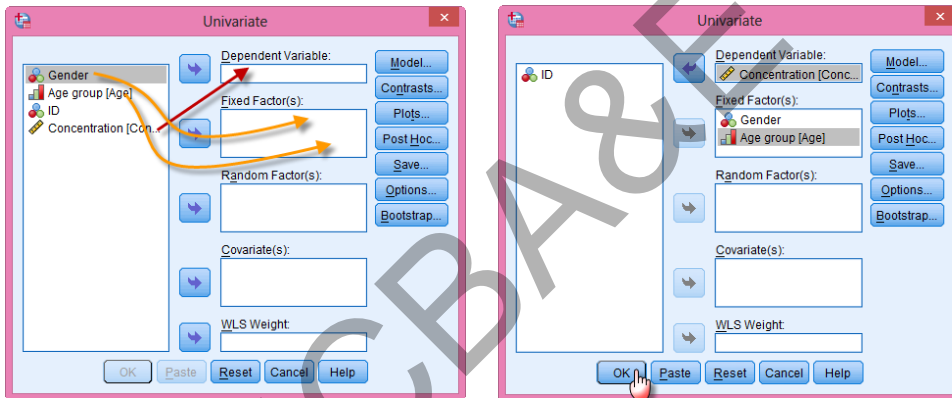
We click on OK , to get the following outputs:

SPSS Output for Simple Factorial Experiment (Univariate Analysis)

ANALYSIS OF VARIANCE

Tests of Between-Subjects Effects

Dependent Variable: Concentration

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7847.511[a]	7	1121.073	23.770	.000
Intercept	176863.401	1	176863.401	3750.065	.000
Gender	227.529	1	227.529	4.824	.035
Age	7133.193	3	2377.731	50.415	.000
Gender * Age	486.789	3	162.263	3.440	.028
Error	1509.208	32	47.163		
Total	186220.120	40			
Corrected Total	9356.719	39			

a. R Squared = .839 (Adjusted R Squared = .803)

The interpretation of the results is given as:

Gender (Factor A)

p-value = 0.035, which is less than 0.05 there is a significant difference between gender regarding concentration of drug in the blood, .. Gender does affect the blood concentration.

Age groups (Factor B)

p- value = 0.000, which is less than 0.05 the result is significant, therefore, there is a significant difference between age groups with respect to concentration of blood.

Interaction between age and gender (AB)

p-value = 0.028, which is less than 0.05 which shows that age and gender interact. This simply means that the response of treatment at different ages does not show the same pattern for both males and females.

Example 5.11:
An experiment is devised to test the hypothesis that an elderly person's memory retention can be improved by a set of oxygen treatments. A group of scientists administer these treatments to men and women. The men and women are each randomly divided into 4 groups of 1, 2, 3, 4 (the two groups not given any treatments are served as control). The treatments are set up in such a manner so that all individuals thought they are receiving the oxygen treatments for the total three weeks. After the treatment ended, a memory retention test was administered. The result (higher scores indicating higher memory retention) are as follows:

Table 5.19
Number of week's oxygen treatments (scores)

		0	1	2	3
		42	39	38	42
		54	52	50	55
	Male	46	51	47	39
		38	50	45	38
		51	47	43	51
Gender		49	48	27	61
		44	51	42	55
	Female	50	52	47	45
		45	54	53	40
		43	40	58	42

i) Test the hypothesis that length of treatment does not affect the memory retention.
ii) Test the hypothesis that there is no difference in gender.
iii) Test whether or not there is interaction effect.

Solution:
SPSS package is used and output is on next page:

SPSS output for simple factorial experiments

Tests of Between-Subjects Effects

Dependent Variable: DATA

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	86862.400	1	86862.400	1589.431	.000
	Error	218.600	4	54.650[a]		
GENDER	Hypothesis	19.600	1	19.600	.441	.543
	Error	177.900	4	44.475[b]		
WEEKS	Hypothesis	60.000	3	20.000	.311	.817
	Error	772.000	12	64.333[c]		
SCORES	Hypothesis	218.600	4	54.650	.790	.579
	Error	341.288	4.933	69.183[d]		
GENDER * WEEKS	Hypothesis	18.000	3	6.000	.151	.927
	Error	475.500	12	39.625		
GENDER * SCORES	Hypothesis	177.900	4	44.475	1.122	.391
	Error	475.500	12	39.625		
					.	.

Interpretation

Length of treatment: p = 0.817, which is greater than 0.05 the result is not significant, therefore, we say with 5% level of significance that the length of treatment does not affect the memory retention.

Gender: p = 0.543, which is greater than 0.05 therefore the result is not significant. We say with 95% confidence that there is no difference in males and females regarding memory retention.

Interaction age and gender: p=0.927, which is greater than 0.05 therefore gender and time period have no interaction.

Example 5.12:

Twenty overweight individuals, each more than 40 pounds over-weight, were randomly assigned to one of 2 diets. After 10 weeks, the total weight loss (in pounds) of the individuals on each of the diets was as in Table 5.20:

Table 5.20

diet1	22.2	23.4	24.2	16.1	9.4	12.5	18.6	32.2	8.8	7.6
diet2	24.2	16.8	14.6	13.7	19.5	17.6	11.2	9.5	30.1	21.5

Test at 5% level of significance that two diets have equal effect.

Suppose 10 people placed on each diet consisted of 5 men and 5 women. The data are given in Table 5.21.

Table 5.21

	Diet 1	Diet 2
Women	7.6	19.5
	8.8	17.6
	12.5	16.8
	16.1	13.7
	18.6	21.5
Men	22.2	30.1
	23.4	24.2
	24.2	9.5
	32.2	14.6
	9.4	11.2

(i) Test the hypothesis that the diet has the same effect on men and women

(ii) Test the hypothesis that there is no interaction between gender and diet.

Solution:
This question is left to the students to solve by using SPSS Package.

5.7 "n of 1 Trials": Controlled Trials in Single Subjects

Controlled trials in individual patients have long been used in behavioral science and have recently been discussed and used by many authors. March et al. (1994) show that controlled trials offer a methodology for informed decision making. Johansson (1991) argues that "n of 1 trials" may be more economical and speedy in new drug development than the conventional clinical trials. Mahon et al. (1996) show that 'n of 1 trials' lead to better outcome over standard practice in terms of use of less medication.

In Statistics, we need a sample of reasonable size to draw valid inference for the population from which a sample is drawn. Statistics do not deal with individual units. However, in Fisher tea testing problem, a woman was asked to detect whether milk had been added before or after a tea infusion. She was given a number of cups of tea purely in random order. It was not envisaged for this tea testing experiment whether women in general could detect the difference between milk added before or after tea infusion. If a group of such individuals is involved in the tea testing experiment, the results can be generalized.

In medical sciences and other areas like Psychology, behavioral medicine, etc. doctors are interested in the individual patients, and as such single case studies are more relevant to subjects of researchers.

In order to deal with individual units, a method of 'n of 1' trial or "controlled trial in single subjects" has been developed. The basic concept of "n of 1" trial is that two treatments can be compared on the same patient and that "n of 1" trials have been developed to find appropriate treatment for individual patients.

It is true that observations on one individual are not independent and so many conventional statistical techniques are inapplicable but Campbell (1994) professes that data measured serially are not necessarily dependent. He gave an example of

independence in "randomly generated numbers purporting to be blood pressure recordings at 5 minutes intervals 20 minutes before and 20 minutes into a psychological stress test". The example seems to contradict itself as 'randomly generated numbers' cannot represent blood pressure recordings in individual patients. In statistics, particularly in Business and Economic Statistics, methods that can be applied to serially dependent observations are available and so these methods can be applied to data measured serially in medicine.

5.7.1 Statistics in "n of 1 trials"

In "n of 1 trial" experiments, treatments and/or treatment periods are randomly allocated to a single subject. The outcomes of such an experiment are observations that are not generally independent.

A study was carried out by March et al. (1994) on individual patients where each patient was treated with a particular dose. Patients, doctors and research assistants were all blinded so far as treatment was concerned. Besides basic statistics, graphs of daily scores were plotted. Values from the second weeks are compared over the cycle by a paired t-test with 2 degree of freedom. The sign test was also used to assess the effect of the dose.

Because of danger of one dose over the other in a particular types of patients, the dose is not prescribed without an "n of 1 trial" to each patient. It was seen that 'n of 1 trials" provided useful decision about the patients. It further avoided unnecessary treatment with a particular medicine.

The main idea of "n of 1 trials" is that each patient is his own control as well as treated subject. Each treatment and treatment periods are randomly assigned to individual patient. Responses to each treatment and treatment periods are recorded. Many clinicians are confident that controlled single-subject-trial can be used to solve difficult issues. See Guyatt et al. (1990), Johann Essen (1991), Levis (1991), etc. However not all clinical drugs are appropriate for n of 1 trials. (Guyatt et al. 1988, Johannessen et al. 1991).

Group trials or "n of 1 trials" are similar in nature as in statistics. Treatments and treatment periods are randomly allocated to subjects. With single subject, the number of treatment periods (sample size) is minimal or very low giving rise to large type II error but the "n of 1 trial" violates some of the assumptions needed in statistical tests.

There is no reason to conduct "n of 1 trial" or for this purpose any experiment, if drug effect is well known and works for all patients. The "n of 1 trial" should be adopted in those cases where the efficiency of a drug is intended to be used in long-term management.

In research where drugs efficiency needs to be tried, "n of 1 trials" may give new insight into the problems. In development of new drug, "n of 1 trials" could prove very useful instead of experiments run over many subjects. March et al. (1994) says, "In conclusion, the single subject trial bridges gap between research and clinical practice. It may provide new insight into vaguely defined conditions, improve therapeutic decisions, strengthen the doctor-patient relationship and create a more critical attitude towards drug treatment both among patients and doctors".

The "n of 1 trial" avoids one of the biggest problems of finding enough suitable patients for clinical research. 'n of 1' trials are advocated for such clinical conditions that are chronic and curable with repeated doses and that an individual patient responds to a particular treatment.

In large number of cases in "n of 1 trials", determination of variations within and between patients is possible and could provide information about the average effect.

In an experiment, a patient is treated with a placebo and a drug over 12 treatment periods. The drug along with placebo is administered to particular patient in a double blind, randomized multiple cross over sequence. Each treatment period is randomly assigned. Patient is asked to give score in a scale of 6 for pain for each treatment period. Measures of responses are obtained for each treatment period.

Table 5.22
Scores given by a patient by treatment and treatment period

Patient Treatment Period	Drug	Score out of 6
1	Drug (D)	4
1	Placebo (P)	2
1	P	2
1	D	4
1	D	5
1	P	3
1	P	1
1	D	4
1	P	2
1	D	5
1	P	3
1	D	4

The data is summarized in Table 5.17 (next page) and is represented by single bar charts as given in Fig. 5.5.
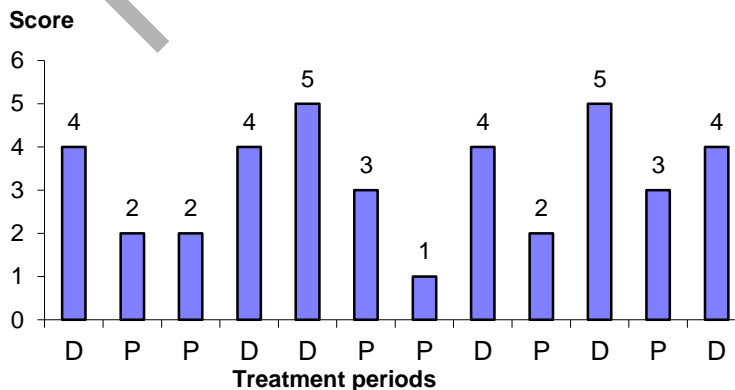
Fig. 5.5: Bar diagram of drug/placebo

Table 5.23
Summary Statistics

Drug Score	Placebo Score
4	2
4	2
5	3
4	1
5	2
4	3
4.33	2.17

Average (Standard Deviation) is 2.17 ($\pm$ 0.75) for Placebo and 4.33 ($\pm$ 0.52) for Drug per treatment period. The ratio of drug to placebo seems to be 2 to 1.

5.7.2 Use of Analysis of Variance for "n of 1 trials"

The 'n of 1 trials' is a special case of cross over design or repeated measure designs. The research unit is a human or an animal subject. Each subject is measured under several conditions, or at different points of time.

Suppose we have n patients and each patient is subject to p treatments or each patient is administered a drug p times (viz. days) and each time a measurement of some character is made. The data format is as follows:

Subjects	Repeated Measures
S_1	$y_{11}\ y_{12}$ ...... y_{1p}
S_2	$y_{21}\ y_{22}$ ...... y_{2p}
.	. . . .
.	. . . .
S_n	$y_{n1}\ y_{n2}$ ...... y_{np}

The correct analysis of such data is more complex than if each patient is measured once.

A simple additive model is applicable with usual conditions.

$$\mathbf{y_{ij} = m + b_i + \varepsilon_{ij}} \qquad i = 1, 2, ..., n$$

$$j = 1, 2, ..., p$$

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & ... & \rho \\ \rho & 1 & ... & \rho \\ . & . & . & . \\ \rho & \rho & ... & 1 \end{pmatrix} = s^2 (1-\rho)I + \rho J$$

where $\Sigma_i\ t_i = 0$, $e_{ji} \sim$ NID $(0, \Sigma)$, J is a square matrix of 1's and I is an identity matrix. Bock (1963) and Huyuh and Feldt (1970) showed that the most general condition under which univeriate F-type remains valid is that $C\ \Sigma\ C^{/} = \sigma_I^2$ where C is (p-1) xp matrix whose rows are orthogonal contrasts. In clinical trials where n and p are sufficiently

large, usual model conditions are met. When $\Sigma \neq \sigma^2 I$, an approximate F-test for repeated measures is applicable with reduced degree of freedom

$$\nu = \frac{\left[t_r(\Sigma - J\Sigma/p) \right]^2}{(p-1)t_r(\Sigma - J\Sigma/p)} . \tag{5.5}$$

Cases dealing with missing data can also be dealt with (Crepean et al, 1985). Bland and Altmar (1994) generated simulated data on 5 subjects with un-correlated pairs of measurements:

Table 5.24

	Subject	1	Subject	2	Subject	3	Subject	4	Subject	5
	A	B	A	B	A	B	A	B	A	B
	48	58	63	28	38	40	51	46	55	62
	56	53	74	24	56	41	46	36	51	50
	49	44	69	26	46	40	36	41	54	66
	38	53	55	19	43	41	49	43	46	51
	50	56	73	22	52	34	46	45	55	52
Subject mean	48.2	52.8	66.8	23.8	47.0	39.2	45.6	42.2	52.2	56.2
Correlation coefficient	r = -0.02 p = 0.97		r = 0.32 p = 0.59		r = -0.30 p = 0.63		r = 0.37 p = 0.55		r = 0.55 p = 0.33	

A and B may be two drugs. Each drug is administered 5 times to each subject.

Bland and Altma (1994) made a correlation analysis on the repeated data. The same can be used to study variation between subjects, and A and B within subjects.

There are 5 subjects and two types of drugs. It is a crossover design. Subjects and drugs cannot be randomized. However, drugs A and B can be randomized within subjects. Each subject is given the two drugs 5 times at random with all the medical conditions like drug A can be given say after drug B is given. We have 5 observations from each of the drugs for each of the subjects.

Drugs	x1	x2	x3	x4	x5
1	48	63	38	51	55
1	56	74	56	46	51
1	49	69	46	36	54
1	38	55	43	49	46
1	50	73	52	46	55
2	58	28	40	46	62
2	53	24	41	36	50
2	44	26	40	41	66
2	53	19	41	43	51
2	56	22	34	45	52

Analysis is done and the result is given on the next page.

ANOVA TABLE

Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.
SUBJECTS	Pillai's Trace	.965	34.475	4.000	5.000	.001
	Wilks' Lambda	.035	34.475	4.000	5.000	.001
	Hotelling's Trace	27.580	34.475	4.000	5.000	.001
	Roy's Largest Root	27.580	34.475	4.000	5.000	.001
SUBJECTS * DRUGS	Pillai's Trace	.996	303.053	4.000	5.000	.000
	Wilks' Lambda	.004	303.053	4.000	5.000	.000
	Hotelling's Trace	242.443	303.053	4.000	5.000	.000
	Roy's Largest Root	242.443	303.053	4.000	5.000	.000

Mauchly's Test of Sphericity

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon	
					Greenhouse-Geisser	Huynh-Feldt
SUBJECTS	.104	14.499	9	.116	.679	1.000

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
SUBJECTS	Sphericity Assumed	910.000	4	227.500	8.732	.000
	Greenhouse-Geisser	910.000	2.714	335.261	8.732	.001
	Huynh-Feldt	910.000	4.000	227.500	8.732	.000
	Lower-bound	910.000	1.000	910.000	8.732	.018
SUBJECTS * DRUGS	Sphericity Assumed	3856.720	4	964.180	37.01	.000
	Greenhouse-Geisser	3856.720	2.714	1420.890	37.01	.000
	Huynh-Feldt	3856.720	4.000	964.180	37.01	.000
	Lower-bound	3856.720	1.000	3856.720	37.01	.000
Error(SUBJECTS)	Sphericity Assumed	833.680	32	26.052		
	Greenhouse-Geisser	833.680	21.714	38.393		
	Huynh-Feldt	833.680	32.000	26.052		
	Lower-bound	833.680	8.000	104.210		

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	112338.000	1	112338.000	1979.873	.000
DRUGS	1039.680	1	1039.680	18.324	.003
Error	453.920	8	56.740		

Repeated measurements are assumed independent as drugs were randomly administered to patients having no knowledge what drug is being administered to them. Subjects are significantly different whereas drugs are effective.

'n of 1' with at least singly-blinded, can be easily analyzed and valid conclusion can be drawn. The results of 5 subjects can be pooled provided between subjects variation is not significant. Power of test and type I error can be usually calculated.

Chapter 6

Regression and Correlation

6.1 Introduction

In this Chapter, we will discuss and analyze the relationship between two and more than two variables. For example, a medical researcher may be interested in the relationship between a patient's blood pressure, X, and heart rate, Y; he may be interested to see the relationship of a certain drug and its effect in lowering the heart rate in adults; he may be interested in the relationship between the increase in age or weight and its effect on systolic blood pressure and so on. In each case, the objective of his interest is not merely academic but the medical researchers wish to determine whether blood pressure is a good indicator of a patient's heart rate or increase in weight.

One of the methods to investigate the increase (decrease) in one variable with the increase (decrease) in another variable is a regression method. Regression method refers to a set of techniques for studying the straight-line relationship among two or more than two variables, one of them is dependent (response) variable and others are all independent (explanatory) variables(s). The terms *dependent* and *independent* do not imply any cause and effect relationship between the two variables. It simply means that one variable is independent and the other variable depends on the first one. In the example of blood pressure and weight of patients, blood pressure is the response variable that depends on the weight, which is the explanatory variable. In case, regression is used for prediction, blood pressure is the outcome and weight is the predictor. Possibly the simple line could be Y = a + b X, where a and b are constant numbers, a is called intercept, b is slope of the straight line. It is not possible to determine a unique line that fits all points. We find the best possible line that passes through the nearest places of all these points.

If we are interested in finding whether some sort of relationship exists between two or more than two variables, then it is a study of *correlation*. In fact correlation indicates relationship between two variables. The correlation refers to measurements of the strength of relationship between two or more than two variables. A numerical value of correlation is *called a correlation coefficient*.

Note that in linear regression the dependent variable is always quantitative.

6.2 Simple Linear Regression Analysis

We explain the concept of simple regression analysis, with an example:

Example 6.1:
The following data and Table 6.1 show the age (X) and blood pressure B.P (Y) of 20 healthy persons taken from a large population.

Table 6.1

Age (x)	B.P. (y)	Age (x)	B.P. (y)
20	120	46	128
43	128	53	136
63	141	70	146
26	126	20	124
53	134	63	143
31	128	43	130
58	136	26	124
46	132	19	121
58	140	31	126
70	144	23	123

We can visualize the bivariate relationship by constructing a *scatter diagram* for this sample data.

The scatter diagram is a useful aid in studying the relationship between two variables. The basic purpose of scatter diagram is to see whether there is any relationship between the two variables. The scatter diagram [6.1] allows visual examination whether there is a linear, non-linear or no relationship between variables. Plotting pairs of sample observations on two-dimensional graph paper construct a scatter diagram, i.e. age (independent variable) on the x-axis and blood pressure (dependent variable) on y-axis. If we draw a straight line through these points as shown in Figure 6.2, the line will not pass through all these points. It can be seen that blood pressure increases linearly as the age increases. Thus we could select a model that proposes a straight line relationship between age and blood pressure. We do not expect that the relationship, $Y = \alpha + \beta X$ will hold exactly for every healthy person. This model will be adequate if all the points fall exactly on the straight line. This model is known as a *deterministic model*. This ideal situation generally never occurs in practice.

A more reasonable model is one that allows *unexplained variation* in blood pressure caused simply by random phenomena.
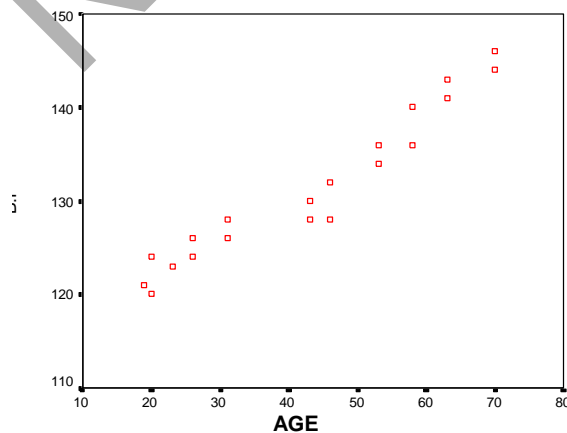
Fig. 6.1: Scatter diagram of age and blood pressure

A model that accounts for this random error is called a *probabilistic model*, i.e.

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \, , i = 1, 2, ...n,$$ (6.1)

where α and β are constants and ε_i are the deviations of points from the line.
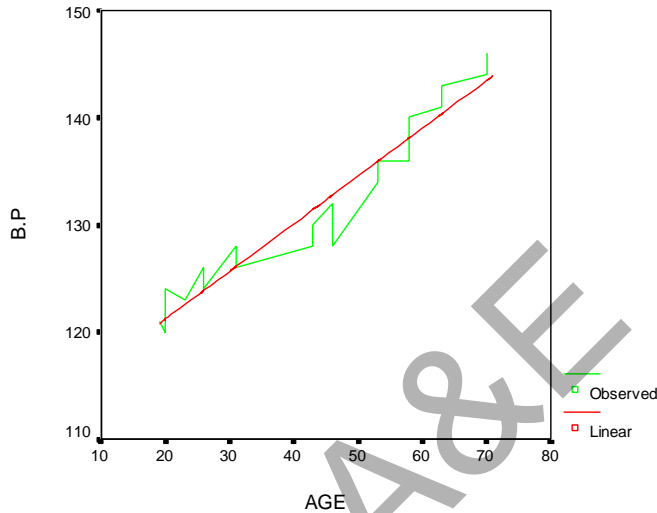
Fig. 6.2:   Straight line by the method of least squares

This is known as a full linear regression model. We assume that ε_i follows a normal distribution with mean = 0 and variance = σ^2 i.e. N(0, σ^2), α is the intercept and β is called the slope of the line. The slope shows the amount of increase (or decrease) in the deterministic component of Y for every 1-unit increase (or decrease) in X.

One interpretation of the regression line is that for a healthy person with age (X), the corresponding blood pressure (Y) will be normally distributed with mean $= \alpha + \beta X$ and variance σ^2. If σ^2 were 0, then every point would fall exactly on the regression line. However, the larger the σ^2, the greater the deviations of points from the regression line.

How can we interpret β ?

(a) $\beta > 0$ (a) $\beta < 0$ (a) $\beta = 0$
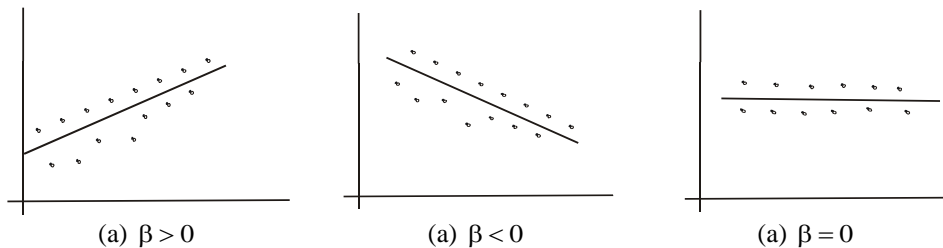
Fig. 6.3: Regression lines for different values of β

If β is greater than 0 then as x increases, the expected value of $y = \alpha + \beta x$ increases [see Fig. 6.3(a)]. If β is less than zero then as x increases, the expected value of y decreases [see Fig. 6.3(b)]. If $\beta = 0$ then there is no relationship between x and y [see Fig. 6.3(c)] and y-points lie around a line parallel to x-axis.

Moreover the effect of σ^2 on a regression line may be seen from Figure 6.4.

$\sigma^2 > 0$ (imperfect fit) $\sigma^2 = 0$ (perfect fit)
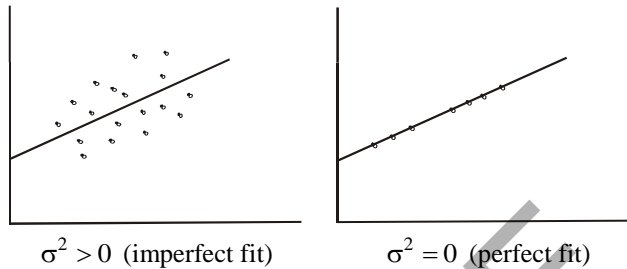
Fig. 6.4: The effect of σ^2 on a regression line

As with most statistical procedures, the validity of the inferences depends on certain assumptions being satisfied. The assumptions about the random error, ε, required for a linear regression analysis are as:

(i) The probability distribution of ε is normally distributed with "zero" mean and "σ^2" variance.

(ii) The errors associated with any two observations are independent, i.e. the error associated with one value of y has no effect on the errors associated with other values of y.

Note that there are some more assumptions i.e. non-zero variance of independent variable, Additivity, multi-colinearity, homo-scedasticity and normality; these are not mentioned here. The outcome variable must be quantitative.

6.2.1 Method of Least Squares

One way to use regression is to fit a straight line through a set of points. Many straight lines can be drawn, but a straight line fitted by the *method of least squares is the best fitted straight line*.

The best line is that which passes as nearly as possible through the points i.e. deviations of points from the straight line is smallest. If sum of squares of all deviations of all the points from y of the straight line is minimized, then the line obtained through this process shall be the best-fitted line for the data. This method is called the Method of Least Squares. If the regression line of Y on X is linear, we have an equation (6.1), where ε_i represent measurement errors in Y but not in X.

By the method of Least Squares, we minimize $\sum\limits_{i} \varepsilon_i^2$ (sum of the squares of errors) with respect to α and β. We get two least squares equations. If we solve them, we get

$$\mathbf{a} = \overline{y} - \mathbf{b_{yx}}\ \overline{x}$$

and

$$\mathbf{b_{yx}} = \dfrac{\dfrac{\Sigma xy}{n} - \dfrac{\Sigma x}{n}\dfrac{\Sigma y}{n}}{\dfrac{\Sigma x^2}{n} - \left(\dfrac{\Sigma x}{n}\right)^2}\ , \tag{6.2}$$

where "a" is an estimate of α and b_{yx} is an estimate of β_{yx}.

The derivation of the formula is not given here.

6.2.2 Some Applications of Simple Regression

(i)   In studying the effect of a certain drug in reducing heart rate in adults.

(ii)   In studying the relationship between an objective measurement of anxiety and heart rate in adults.

(iii)   In studying the relationship between age and systolic blood pressure.

(iv)   In studying the relationship between birth weight and cholesterol level in pregnant women near term.

(v)   In studying the relationship between HDL cholesterol and alcohol consumption.

The solution of example 6.1 is first explained by manual process, then by using SPSS Package.

No.	Age (x)	b.p. (y)	x^2	xy	No.	Age (x)	b.p. (y)	x^2	xy
1	20	120	400	2400	11	46	128	2116	5888
2	43	128	1849	5504	12	53	136	2809	7208
3	63	141	3969	8883	13	70	146	4900	10220
4	26	126	676	3276	14	20	124	400	2480
5	53	134	2809	7102	15	63	143	3969	9009
6	31	128	961	3968	16	43	130	1849	5590
7	58	136	3364	7888	17	26	124	676	3224
8	46	132	2116	6072	18	19	121	361	2299
9	58	140	3364	8120	19	31	126	961	3906
10	70	144	4900	10080	20	23	123	529	2829
					Total	862	2630	42978	115946

$\sum y = 2630 \quad \sum x = 862, \quad \bar{y} = 131.50 \quad \bar{x} = 43.10$

$\sum x^2 = 42978 \quad \sum xy = 115946$

$$b_{yx} = \dfrac{\dfrac{115946}{20} - \dfrac{862}{20}\dfrac{2630}{20}}{\dfrac{42978}{20} - \left(\dfrac{862}{20}\right)^2} = 0.445089 \approx 0.445$$

The linear regression equation is

$a = \bar{y} - b_{yx}\bar{x} = 112.317$

The fitted Regression line will be

$Y - \bar{y} = b_{yx}\left(X - \bar{x}\right)$ (6.3)

$\hat{y} - 131.50 = 0.445\,(x - 43.10)$

or

$\hat{y} = 112.317 + 0.445\,x$

Regression line may be fitted, alternatively, by using the SPSS package.

How to use the IBM-SPSS package? And how to enter the data to fit linear regression line? It has been explained at the end of the Chapter. The IBM-SPSS package has been used:

Example S6-1

To see how we plot the scatter diagram and construct the regression equation, draw the regression line, we follow the following steps:

The data will be in columns as follows:

	No	Age	B.P
1	1	20	120
2	2	43	128
3	3	63	141
4	4	26	126
5	5	53	134
6	6	31	128
7	7	58	136
8	8	46	132
9	9	58	140
10	10	70	144
11	11	46	128
12	12	53	136
13	13	70	146
14	14	20	124
15	15	63	143
16	16	43	130
17	17	26	124
18	18	19	121
19	19	31	126
20	20	23	123

We plot the scatter diagram as follows:

Graphs→Chart Builder…

From the Gallery select **"Scatter/Dot"**

Double click or move the icon

Move the variable Independent variable "Age" to the X-axis:

Move the variable Dependent variable "Blood Pressure" to the Y-axis:
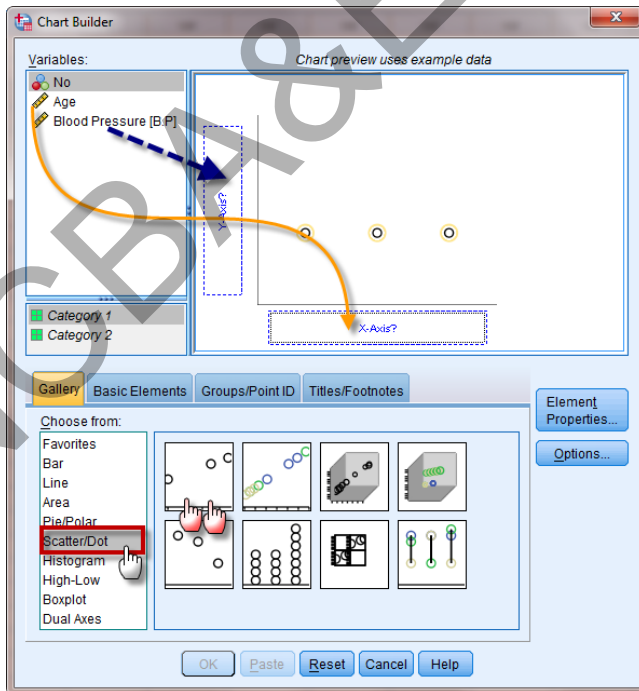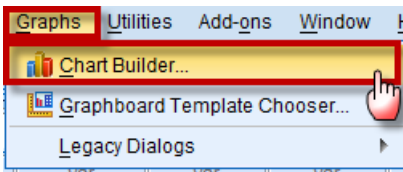
We click on [OK], to get the following Figure:
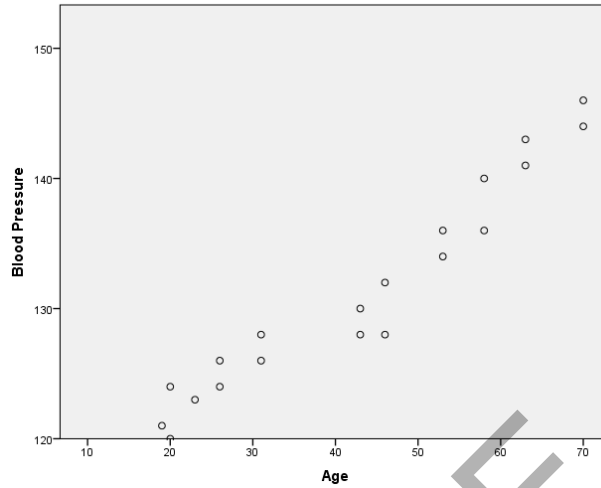
Fig. 6.5: The Scatter Diagram

We obtain the regression equation as follows:

Analyze→ Regression→ Linear …

Move the variable Independent variable "Age" to the Independent(s):

Move the variable Dependent variable "Blood Pressure" to the Dependent:
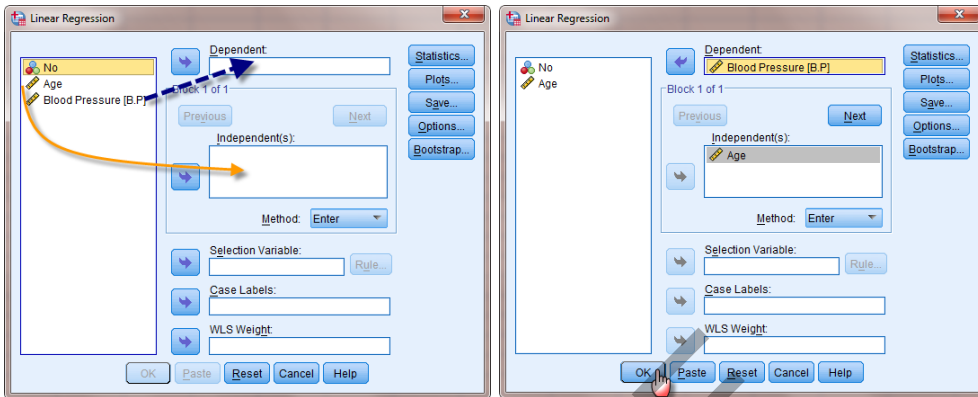
We click on [OK], to get the following outputs:

SPSS output for simple regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.967[a]	.935	.931	2.12

a. Predictors: (Constant), Age of the Patients

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1154.116	1	1154.116	256.838	.000[a]
	Residual	80.884	18	4.494		
	Total	1235.000	19			

a. Predictors: (Constant), Age of the Patients

b. Dependent Variable: Blood Pressure

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	112.317	1.287		87.241	.000
	Age of the Patients	.445	.028	.967	16.026	.000

a. Dependent Variable: Blood pressure

The output is divided into *three general parts*:

   (a) R, R^2 and Adjusted R^2
   (b) ANOVA Table
   (c) Parameters in the equation

These need some explanations.

a) R, R^2 and Adjusted R^2

(i) Simple or Multiple Correlation R:

The basic objective of correlation is to obtain a measure of the degree of relationship that exists between two or more than two variables. This is an index of correlation coefficient. For simple linear regression, it is the simple correlation coefficient, but if independent variables are more than one, it is a study of multiple correlations. *Multiple correlations are the combined effect of all independent variables on dependent variable.* The range of simple correlation coefficient is from -1 to 1 and for multiple correlation coefficients, R varies from 0 to +1.

(ii) R^2 (coefficient of determination):

R^2, which is commonly known as coefficient of determination, is the proportion of the variance of dependent Y that can be explained by the independent variable X. R^2 ranges from 0 to 1. The closer the value of R^2 to 1 the better the model is that accounts for the variation in the data. If $R^2 = 1$, then all the variation in the dependent variable Y can be explained by the variation in independent variable X and all the points fall on the regression line. In this situation, once we know X, we can predict Y, exactly with no error in prediction. If $R^2 = 0$ then independent variable does not give any information about dependent variable.

R^2 can also be calculated from the ANOVA Table as:

$$R^2 = \frac{\text{Regression sum of squares}}{\text{Total (Regression + Residuals) sum of squares}} \qquad (6.4)$$

For this example the value of R^2 is

$$R^2 = \frac{1154.116}{1154.116 + 80.884} = 0.93451$$

R^2 depends on the value of the sum of squares of the residuals. If sum of the squares of the residuals are zero then $R^2 = 1$. This means all the points will fall on the regression line. As the sum of the squares of the residuals increases, the R^2 decreases. In this table R^2 is about 0.935 which means that 93.5% of the variation in Y (blood pressure) is explained by the X (age), or in other words we can say that 93.5% of the sum of squares of deviations of the y-values about their mean is attributable to the linear relationship between Y and X. The practical interpretation of the coefficient of determination, R^2 is briefly described as:

About 100% (R^2) of the information in X explains Y.

(iii) Adjusted R^2:

This value indicates the loss of predictive power or shrinkage. This tells us how variation in Y would be accounted for if the model has been derived from the population from which the sample has been taken. In this example $R^2 = 0.935$ and adjusted $R^2 = 0.931$, therefore the shrinkage is about 0.4 % ($0.935 - 0.931$). This means if the model were derived from the population rather than sample, it would be approximately 0.4% less variance in the outcome variable. This can be calculated by using Stein's formula reported by Stevens (1992).

$$\text{Adjusted}\,R^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\left(1-R^2\right)\right], \quad (6.4)$$

where k is the number of predictors.

(b) ANOVA Table:

The terms in ANOVA table are defined below.

(i) Degrees of Freedom (df)

Degrees of freedom (df) is always 1 for a straight line model and the degrees of freedom of the total is one less than the total number of observations minus the number of parameters (in regression) estimated, (in this example, 20 - 1 = 19), whereas the degrees of freedom for the residual is, 19 - 1 = 18 (degrees of freedom of total - degrees of freedom of regression model).

(ii) Sum of squares

Sum of Squares column separates the variation in the data into portions that are attributable to the regression model and to the residual (error).

(iii) Total sum of squares

= Regression sum of squares + Residual sum of squares

(iv) Mean sum of squares

This is equal to the sums of squares of regression, divided by the degree of freedom. The Mean Square Error equals the sum of squares of errors divided by the error degrees of freedom.

$$\text{MS}\,(\text{Regression Model}) = \frac{\text{Sum of squares of regression of errors}}{\text{Degrees of freedom}}$$

$$= \frac{1154.116}{1} = 1154.116$$

and

$$\text{MS}\,(\text{Error}) = \frac{\text{Sum of squares of errors}}{\text{Error degrees of freedom}} = \frac{80.884}{18} = 4.9$$

(v) t-statistic
t is test-statistic and p-value is associated with the test of the hypothesis. For example, the value of t-statistic from the t-table at 5% significance level is 2.10 for 9 d.f. whereas calculated t-value is 16.06.

(c) Parameters in the Equation

(i) Intercept
One constant term is the intercept of the line. Positive value of the intercept indicates that the line is passing through a point above the origin whereas negative constant value indicates that the line is passing through a point below the origin on the x-axis.

(ii) p-Value
p-value is the level of significance at the observed value of the test- statistic. It is the probability of observing a value beyond the value of test- statistic. It is sometimes matched with the given level of significance. The calculated p-value is 0.0000, which is less than 0.05 (table value). (This has been explained in details in Chapter 4).

(iii)   Slope
The second parameter is the slope of the line. If $\beta = 0$, y is constant. If $\beta > 0$, then y increases (decreases) when x increases (decreases) and if $\beta < 0$, y decreases (increases) when x increases (decreases). These values are the coefficients of independent variable. The interpretation of the regression lines depends on the positive or negative values of B(β). If $\beta = 0$ then there is no relationship between two variables. If $p < 0.05$, the variables are significant and if $p \geq 0.05$, then the variables are non-significant.

Suppose that variables are significant, the results are interpreted as:

(i)   If the coefficient (β) of independent variable is positive then we say that independent variable has a positive effect on the dependent variable.

(ii)   If the coefficient is negative then we say that independent variable has negative effect on the dependent variable.

(iii) The coefficient of independent variable tells us about the rate of change per unit in the dependent variable.

We can draw inference from this example as:

The coefficient of X is about 0.45, and is positive. The increase of one year in age there is 0.45 points increase in blood pressure. To see the increase in blood pressure in 10 years in age, multiply the coefficient of X by 10, which gives 4.5. We say that with the increase of 10 years in age, there is 4.5 points increase in the blood pressure.

Note: We obtain the regression line over the scatter diagram as follows:

Double click on the scatter diagram in Fig 6.5 to open the **Chart editor**
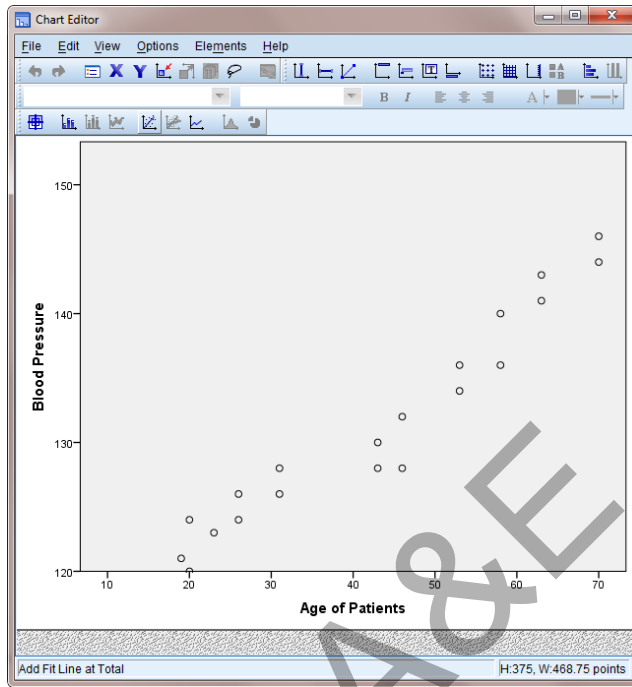
Click on and chose Linear
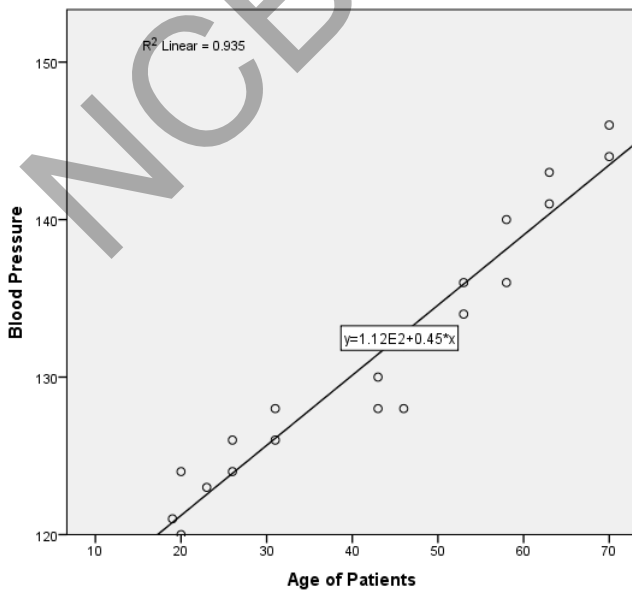
Fig. 6.6: The Scatter Diagram with Regression line

Example 6.2:

An experiment was conducted to study the relationship between an objective measurement of anxiety and heart rate in adults. The data relate to 12 normal adults and is given in Table 6.2. Fit a linear relationship between heart rate per minute and objective measurement of anxiety by using the method regression and interpret the result.

Solution:

Here X is independent and Y is considered as a dependent variable so a regression line $E(Y) = \alpha + \beta X$ is fitted.

Table 6.2

Heart rate per minute (X)	Objective measurements of anxiety (Y)
50	48
55	41
60	45
65	41
70	42
75	36
80	38
85	36
90	30
95	32
100	34
105	25

The IBM-SPSS package is used to solve the problem as explained in the following Example:

Example S6-2

The data will be in columns.

We obtain the regression equation as follows:

Analyze→ Regression→ Linear …
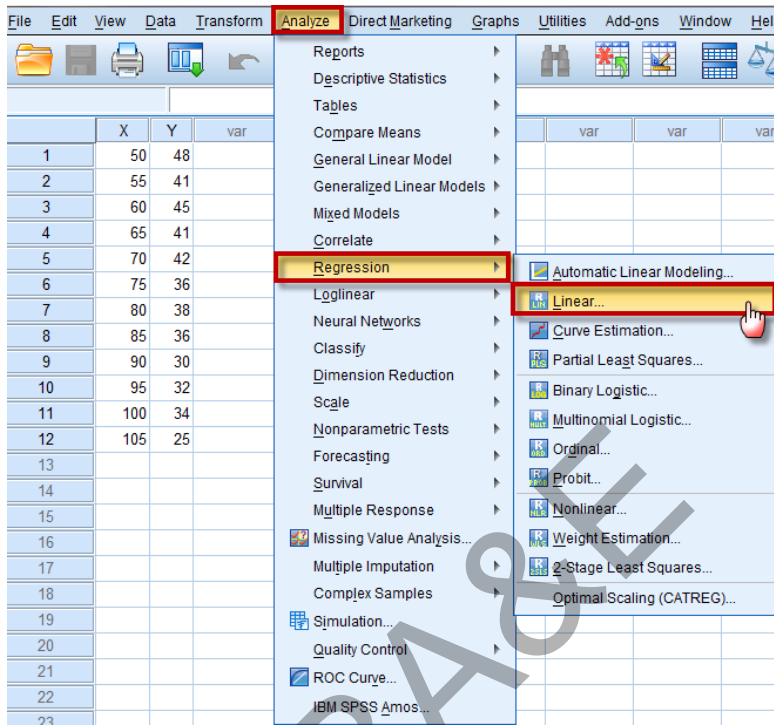
Move the variable Independent variable "X" to the Independent(s):
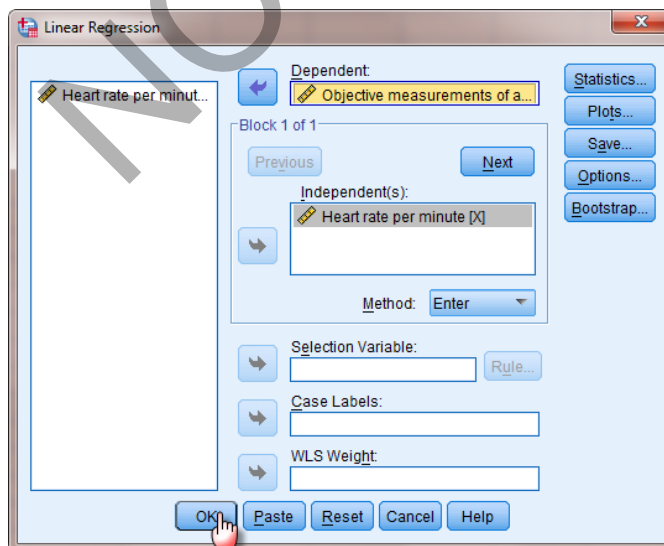
Move the variable Dependent variable "Y" to the Dependent:

We click on OK , to get the following outputs:

SPSS output for simple regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.921a	.849	.834	2.67

a. Predictors: (Constant). Heart Rate Per Minute

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	399.448	1	399.448	56.087	.000a
	Residual	71.219	10	7.122		
	Total	470.667	11			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	63.239	3.544		17.845	.000
	Heart Rate Per Minute	-.334	.045	-.921	-7.489	.000

(i) $R^2 = 0.849$, so about 85% of variation in objective measurement of anxiety has been explained by heart rate per minute.

(ii) Adjusted $R^2 = 0.834$, one can say that a loss of predicted power by using this model is 1.5% (0.849 – 0.834).

(iii) Constant = 63.24

(iv) Slope (β) = -0.334 (negative)

Therefore, the regression line takes the following form.

$$\hat{y} = 63.24 - 0.334\ X$$

The p-value of heart rate per minute is 0.000, which is significant; therefore one can say that heart rate has an effect on anxiety.

Moreover $B = -0.33427 \approx -0.33$, we say that with one unit increase in heart rate, the anxiety decreases by 0.33 units, i.e. for common understanding we multiply -0.334 by 10, which comes out to be -3.34. This means that with 10 points increase in heart rate, anxiety decreases by 3.34 points.

6.3 The Coefficient of Correlation

In Section 6.2 we have discussed that least squares slope b_{yx} (β_{yx}) = b provides useful information between two variables Y and X. Another way to measure relationship is to compute the Pearson product moment correlation coefficient. This is commonly known as r. The correlation coefficient provides a quantitative measure of the strength of the linear relationship between two variables. Note that unlike the slope, the correlation coefficient r is *scale less*. The value of r is always between -1 and +1, no matter what the units of two variables are. Since r and β provide information about the utility of the model, it is not surprising that there is a similarity in computation. The correlation coefficient r is calculated as:

$$ r = \frac{\dfrac{\Sigma xy}{n} - \dfrac{\Sigma x}{n}\dfrac{\Sigma y}{n}}{\sqrt{\dfrac{\Sigma x^2}{n} - \left(\dfrac{\Sigma x}{n}\right)^2}\sqrt{\dfrac{\Sigma y^2}{n} - \left(\dfrac{\Sigma y}{n}\right)^2}} \tag{6.5} $$

The correlation coefficient is symmetrical in x and y. The derivation of the formula (6.5) may be seen in any textbook on statistics. If r = 1 or -1 then we say that there is a perfect positive or a perfect negative correlation. Positive value of r implies that y-value increases as x-value increases. Negative value of r implies that y-value decreases as x-value increases. r = 0 means that there is no correlation. It can be seen from the Fig. 6.5.

A correlation coefficient measures the linear relationship between two variables. A coefficient of +1 means that a higher value of one variable is always associated with a higher value of another, and a coefficient of -1 means that a higher value of one is always associated with a lower value of the other and this relationship is perfect linear. *The correlation coefficient does not indicate how much each variable changes but it indicates the degree of relationship between two variables.*
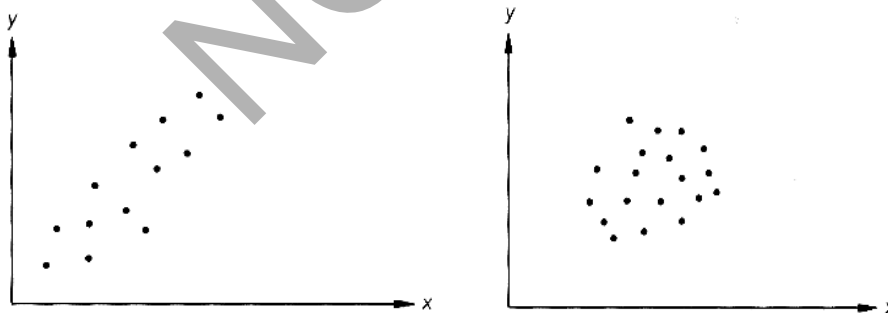
a) Positive r: Y increases as X increases   b) r near 0: little or no linear relationship between Y and X
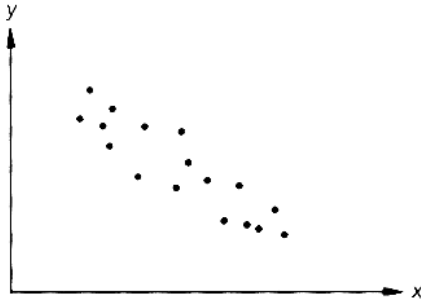
c) Negative r: Y decreases as X increases
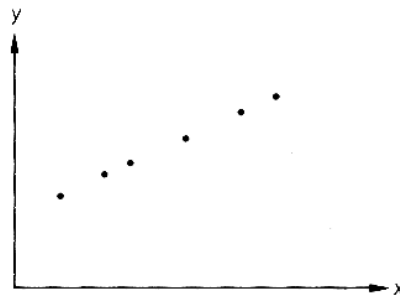
d) r = 1: a perfect positive relationship between Y and X
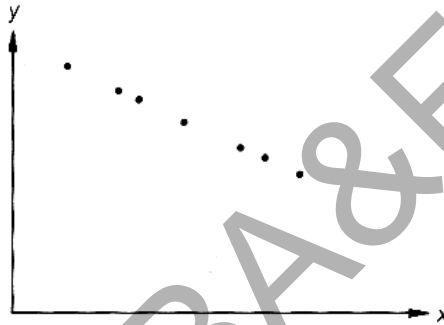
e) r = -1: a perfect negative relationship between Y and X

Fig. 6.7: Value of the correlation coefficient for different pattern of variables

Since the two numerical descriptive measures r and R^2 are very closely related, there may be some confusion as to when each should be used. The recommendations are as:

Coefficient of correlation measures relationship between two variables X and Y, whereas the coefficient of determination (R^2) determines how well the least squares straight-line model fits the data.

Example 6.3:
The followings are the systolic blood pressure of each of 25 pairs of identical twins.

Table 6.3

First twin (x)	118	116	118	120	122	122	122	120	124	125	138	140
Second twin (y)	115	119	116	119	118	138	124	128	126	130	130	125

First twin (x)	142	144	145	162	180	180	182	185	170	172	150	152	155
Second twin (y)	164	160	158	145	184	190	188	180	174	170	160	155	160

Calculate the correlation coefficient and interpret the result.

Solution:

We can proceed with the calculations as:

$\sum(x)$ = sum of the x-values   = 3604

$\sum(y)$ = sum of the y-values   = 3676

$\sum x^2$ = sum of the squares of x-values   = 532832

$\sum y^2$ = sum of the squares of y-values   = 555618

$\sum xy$ = sum of the product of xy   = 543120

n = 25

$$r = \frac{\dfrac{543120}{25} - \dfrac{3604}{25}\dfrac{3676}{25}}{\sqrt{\dfrac{532832}{25} - \left(\dfrac{3604}{25}\right)^2}\sqrt{\dfrac{555618}{25} - \left(\dfrac{3676}{25}\right)^2}} = 0.93$$

Alternatively IBM-SPSS package may be used to solve this problem as explained in the following Example:

Example S6-3

The data will be in columns.

We obtain the Correlation coefficient:

Analyze→ Correlate→ Bivariate …
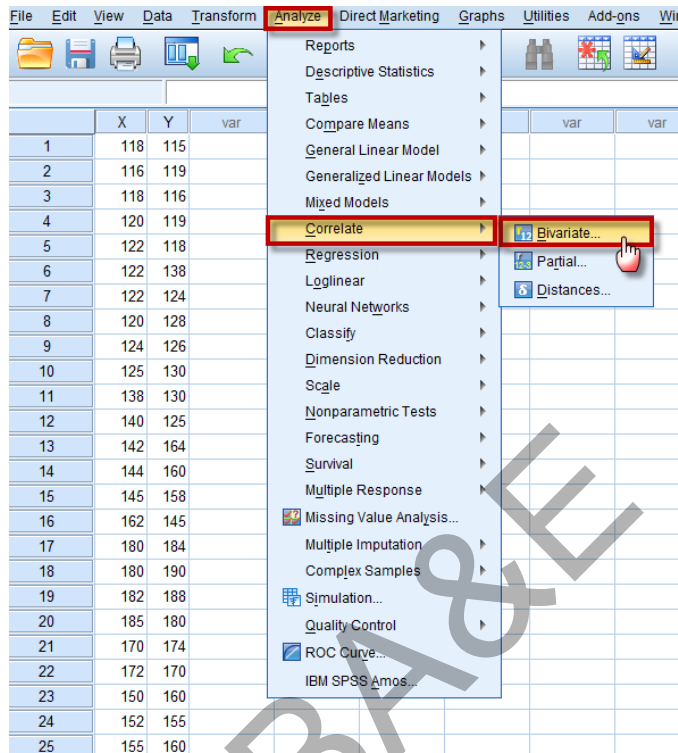
Move the two variables to the Variables:
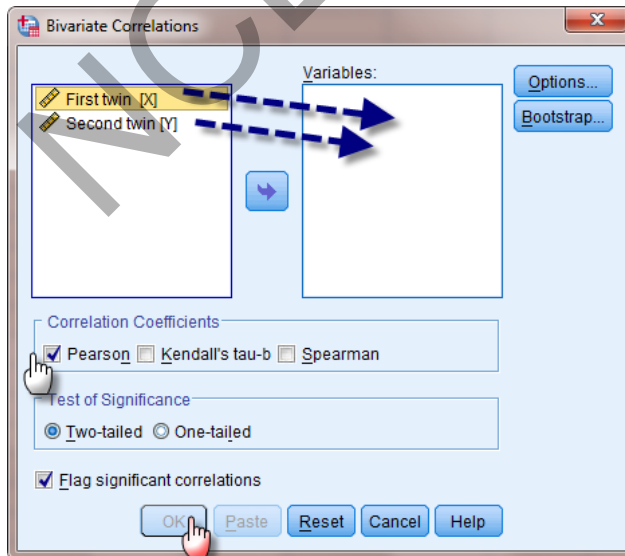
We click on [OK], to get the following outputs:

SPSS output for correlation coefficient

Correlations

		First Twin	Second Twin
First Twin	Pearson Correlation	1.000	.931**
	Sig. (2-tailed)	.	.000
	N	25	25
Second Twin	Pearson Correlation	.931**	1.000
	Sig. (2-tailed)	.000	.
	N	25	25

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

r = 0.931, we can say there is about 93% correlation between two like twin. Since p-value = 0.00, therefore, it is highly significant. This means that the population from which this sample has been taken is highly correlated with respect of identical twins.

6.4 Regression Model for Prediction

After we have statistically checked the usefulness of the straight-line model and are satisfied that X contributes information for the prediction of Y, we are ready to accomplish our original objective using the model for estimation and prediction. The probabilistic model for making inferences can be divided into two categories, viz.

(i) Estimating the mean value of Y, i.e. E (Y), for a specific value of X.

(ii) Predicting Y value for a given value of X.

In the first case, we want to estimate the mean value of Y for a very large number of experiments at a given X value. For example, the psychologist may want to estimate the *mean creativity score* for all mentally retarded children with flexibility score of 3. In the second case, we wish to predict the outcome of a single experiment at a given X value. For example, he may want to predict the creativity score of a particular mentally retarded child who exceeds 3 on the flexibility test. We use the least squares model

$$\hat{y} = a + bX ,$$ (6.6)

both to estimate the mean value of Y, i.e. E(Y), and to predict a value of Y for given X. For this, consider an hypothetical data given in Table 6.4:

Table 6.4

Child	Flexibility score (X)	Creativity score (Y)
1	2	2
2	3	5
3	4	7
4	5	10
5	6	11

Suppose we fit least squares model relating creativity score, y, to flexibility score, x to be

$\hat{y} = -2.2 + 2.3\ x$

We estimate for the mean creativity score of all mentally retarded children that have a flexibility score of 3.

We need to find estimate of E(Y). On the basis of least squares model, our estimate is simply $\hat{y}$. Then, when x = 3, we have

$\hat{y} = -2.2 + (2.3)(3) = 4.7$

Thus, the estimated mean creativity score for all mentally retarded children with flexibility score 3 is 4.7.

We also use the least squares model to predict the creativity score of a particular retarded child whose flexibility score is 3. Just as we use $\hat{y}$ from the least squares model to estimate E(y), we also use $\hat{y}$ to predict a particular value of y for a given value of x. Again when x = 3, we obtain $\hat{y}_s = 4.7$. Thus we predict that a retarded child with a flexibility score of 3 would have a creativity score of 4.7.

Since the least squares model is used to obtain both the estimator of E (Y) and the predictor of y, then how do these two methods differ. The difference lies in the accuracy with which the estimate and prediction are made. This accuracy is best measured by the repeated sampling errors of the least squares line when it is used as an estimator and predictor, respectively. The 95% confidence interval for the mean creativity score for all mentally retarded children with a flexibility score of 3, will be 3.645 to 5.755 whereas the 95% prediction interval, predict the creativity score of a particular retarded child if his flexibility score is 3 will be 2.503 to 6.897. (These limits can be calculated by using SPSS packages easily see Chapters 4 and 5.) It is important to note that the prediction interval for an individual mentally retarded child is wider than the corresponding confidence interval for the mean creativity score. (Note that this will always be true). Over the range of the sample data, the widths of both intervals increase as the value of x gets farther from $\bar{x}$. Thus, *the more* x *deviates from* $\bar{x}$, *the less useful the interval will be in practice*. In fact, when x is selected far away from $\bar{x}$ so that it falls outside the range of the sample data, it is dangerous to make any inference about E(y) or y.

Example S6-4

The data will be in columns.

We obtain the Correlation coefficient:

Analyze→ Regression→ Linear …
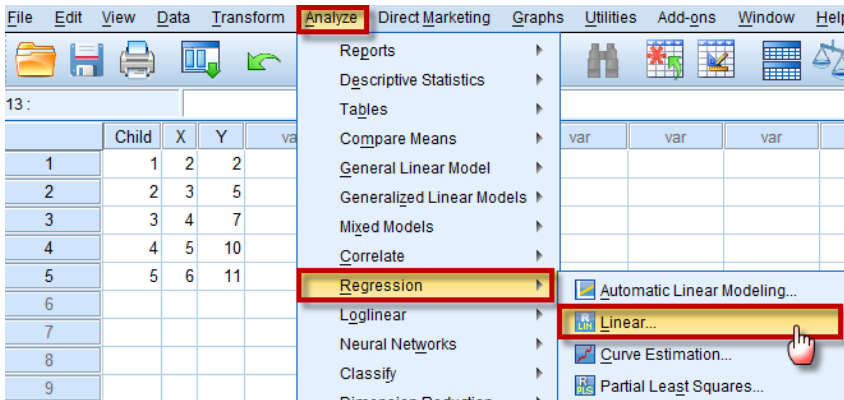
Move the X variable to the Independent(s):

Move the Y variable to the dependent(s):

Click on Save
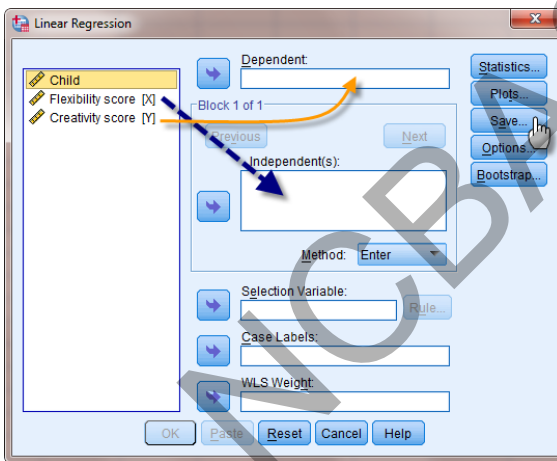
For predicted values Mark on Unstandardized
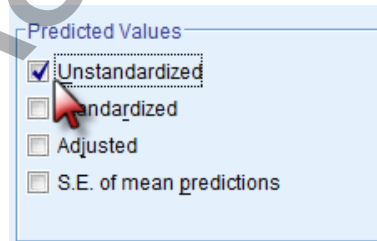
We click on Continue then OK, to get the following output:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.200	.812		-2.708	.073
	Flexibility score	2.300	.191	.990	12.011	.001

a. Dependent Variable: Creativity score

And the predicted values for Y will be added to the data file:

(**Note that** if the value wanted to be predicted is not one of the X values, we add it to the Data file and repeat the same steps and the predicted value of Y will be add automatically).

6.5 Multiple Regression Analysis

This is more complex than simple regression model. In example 6.1, two variables such as weight and blood pressure were used, additional variables such as age, family history, diet, etc. might also be related to blood pressure. Thus we would want to incorporate these and other potential variables into the model if we need to make accurate predictions of blood pressure. A more complex model relating blood pressure to various independent variables such as age, weight, family history is called a general linear statistical model.

The general linear model is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \qquad (6.7)$$

where X_1, X_2, X_3, ... could be weight, height and family history etc. Here Y is dependent and X_1, X_2, X_3, ... are independent variables. $\beta's$ determine the contribution of the independent variable X's and ε as usual is random error component of the model.

6.5.1 Applications of multiple-regression

Some applications of regression

    (i)   Relationship between age, HDL cholesterol and alcohol consumption

    (ii)   Relationship between hypertension (mean arterial blood pressure) and age, weight, body surface area, duration of hypertension, basal pulse and measure of stress.

    (iii) Relationship between birth weight of a child and gestation period and smoking (note that smoking is a qualitative variable).

    (iv) Relationship of systolic blood pressure, birth weight and age of infants.

Method of least squares will also be used to fit linear model to a set of data. This process, along with the estimation and test procedure associated with it, is called a *multiple regression analysis*. Since computations involved in the multiple regression are complex, therefore, all calculations will be made on the computer by using SPSS package. We will follow the same steps as in case of simple model, i.e. the assumptions about the random error term ε in the general linear model are same as in case of simple model.

6.5.2 Fitting the model and interpretation of coefficients

Several cases will be discussed as:

(i)   All independent variables are quantitative.

(ii)   Some independent variables are quantitative and some are qualitative of *two* levels.

(iii)   Some independent variables as quantitative and some are qualitative of *three* levels.

Case 1: All the independent variables are quantitative

Example 6.4:
The data given in Table 6.5 were collected using a simple random sample of 20 hypertensive patients.

Y = mean arterial blood pressure (mmHg)

X_1 = age (years), X_2 = weight (kg), X_3 = body surface area (sqms)

X_4 = duration of hypertension (years), X_5 = basal pulse (beats/min)

X_6 = measures of stress

Table 6.5

Patient	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	105	47	85.4	1.75	5.1	63	33
2	115	49	94.2	2.10	3.8	70	14
3	116	49	95.3	1.98	8.2	72	10
4	117	50	94.7	2.01	5.8	73	99
5	112	51	89.4	1.89	7.0	72	95
6	121	48	99.5	2.25	9.3	71	10
7	121	49	99.8	2.25	2.5	69	42
8	110	47	90.9	1.90	6.2	66	8
9	110	49	89.2	1.83	7.1	69	62
10	114	48	92.7	2.07	5.6	64	35
11	114	47	94.4	2.07	5.3	74	90
12	115	49	94.1	1.98	5.6	71	21
13	114	50	91.6	2.05	10.2	68	47
14	106	45	87.1	1.92	5.6	67	80
15	125	52	101.3	2.19	10.0	76	98
16	114	46	94.5	1.98	7.4	69	95
17	106	46	87.0	1.87	3.6	62	18
18	113	46	94.5	1.90	4.3	70	12
19	110	48	90.5	1.88	9.0	71	99
20	122	56	95.7	2.09	7.0	75	19

Discuss the effect of all the independent variables on mean arterial blood pressure, by using the method of multiple regression. Comment on the individual variable. (Source Daniel, 1981).

Example S6-5

The data will be in columns.

We obtain the multiple regression coefficients as follows:
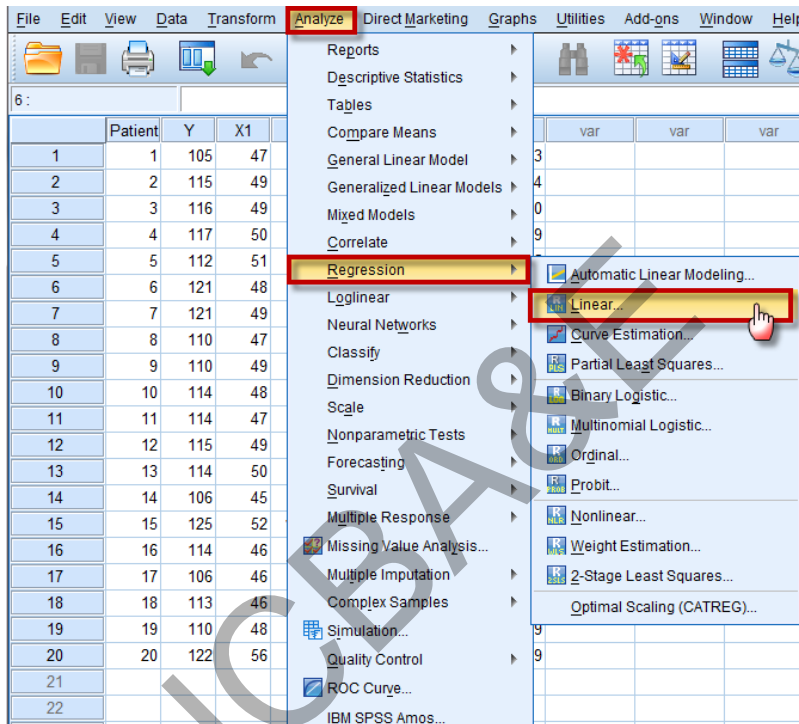
Analyze→ regression → Linear …

Move the variable Independent variables "X1,…,X6" to the Independent(s):

Move the variable Dependent variable "Y" to the Dependent:
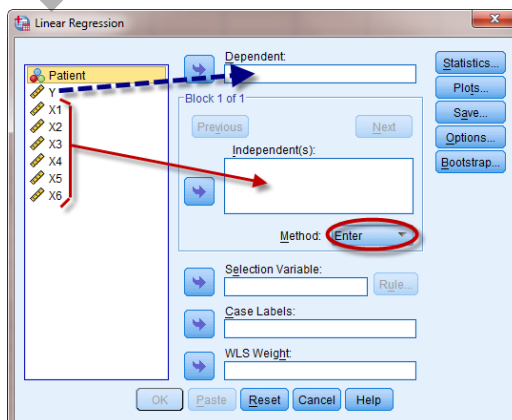
We click on [OK], to get the following outputs (for the "**Enter**" Method):

SPSS output for multiple regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-13.845	2.687		-5.153	.000
	age (years)	.729	.054	.336	13.403	.000
	weight (kg)	.957	.063	.757	15.095	.000
	body surface area (sqms)	3.923	1.621	.099	2.419	.031
	duration of hypertension (years)	.063	.051	.025	1.244	.235
	basal pulse (beats/min)	-.074	.052	-.052	-1.418	.180
	measures of stress	.004	.003	.029	1.318	.210

a. Dependent Variable: Mean arterial blood pressure

We may note that X1, X2 and X3 are Not Significant (as the P-values > 0.05). Here we advise to use an alternative method than the **Enter** method. We will use the **Stepwise** method, *which not only select the significant variables, but also it select them in order of importance* as follows:

Move the variable Independent variables "X1,…,X6" to the Independent(s):

Move the variable Dependent variable "Y" to the Dependent:

Chose the Stepwise Method:
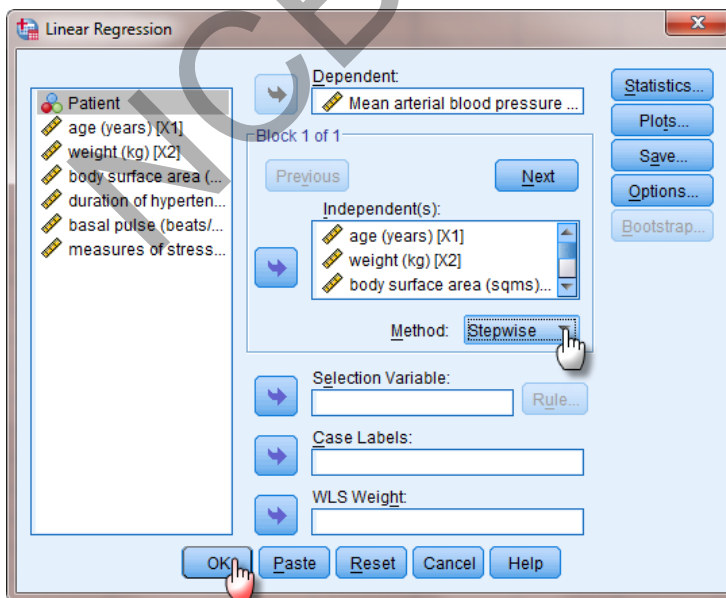
We click on [OK], to get the following outputs (for the "**Stepwise**" Method):

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	Sig. F Change	Durbin-Watson
							Change Statistics	
3	.997	.995	.994	.43705	.003	9.255	.008	1.896

In multiple regression analysis, the value of R^2 is used as how much variation in the dependent variables has been explained by independent variable. As an alternative to using R^2 as a measure of model accuracy, the adjusted R^2 is computed. Unlike R^2, adjusted R^2 takes into account the loss of predictive power by this model, if the model were derived from the population rather than sample. Adjusted R^2 will always be smaller than R^2 and cannot be forced to 1 by simply adding more and more independent variables to the model as the case with R^2. Consequently, analysts prefer more conservative adjusted R^2, when choosing the measure of model accuracy. The value of adjusted R^2 = 0.99 which is slightly smaller than R^2. Our interpretation is that after adjusting for sample size and number of parameters in the model, approximately 99% of sample variation in means arterial blood pressure has been explained by the linear model and loss of predictive power or shrinkage is about 0.3% (0.997 – 0.994).

(1) We see that R^2 (coefficient of determination) = 0.995, this implies that by using these independent variables (age, weight and body surface area) in a first order model to predict y, 99.5% variation has been explained of mean arterial blood pressure by age, weight, body surface area, whereas duration of hypertension, based pulse and measure of stress are not playing part in explaining the variation of mean arterial blood pressure as they are non-significant. Adjusted R^2 is 0.994, the loss of predictive power is 0.6% if this model will be used for the purpose of forecasting.

ANOVA

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	556.944	3	185.648	971.934	.000
Residual	3.056	16	.191		
Total	560.000	19			

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-13.667	2.647		-5.164	.000
Weight (kg)	.906	.049	.717	18.490	.000
Age (years)	.702	.044	.323	15.961	.000
Body surface area (sqm)	4.627	1.521	.116	3.042	.008

Excluded Variables

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
	duration of hypertension (years)	.026	1.359	.194	.331	.866
	basal pulse (beats/min)	-.014	-.452	.658	-.116	.355
	measures of stress	.018	.988	.339	.247	.992

(2) Variables age, weight and body surface area are in the equation. They are highly significant ($p < 0.0001$), we say that these variables have very strong effect on the mean arterial blood pressure.

(3) Since the variables, basal pulse, duration of hypertension and measure of stress are not in the equation, these are non-significant ($p > 0.05$) [can be seen in SPSS output]. Therefore, we say that these variables have no effect on mean arterial blood pressure. This does not mean that these variables are less important.

The general model takes the following form:

Mean arterial blood pressure
$$= -13.667 + 0.702 \text{ age} + 0.906 \text{ weight} + 4.627 \text{ body surface area}$$

or

$$\hat{y} = -13.667 + 0.702X_1 + 0.906X_2 + 4.627X_3 \tag{6.8}$$

Coefficients of age, weight and body surface area are positive, therefore, these factors have positive effect on mean arterial blood pressure. These can be interpreted as:

Age: with 10 years increase in age the mean arterial blood pressure is increased by 7 points provided all other variables are held constant.

Weight: with 10 kg increase in weight the mean arterial blood pressure is increased by 9 points provided all other variables are kept constant.

Body surface area: with one square meter increase in the body the mean arterial blood pressure is increased by 4.6 points when all other variables are kept constant.

Starting from Version 19, The IBM-SPSS add the "Automatic Linear Modeling" for the regression. Here, we will show the steps for using it:

Example S6-5b (Automatic Linear Modeling)

Before we use the Automatic Linear Model, we have to be sure that we define the Dependent Variable "Target" and the independent variable(s) "Input". We Change the **Role** as follow:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Patient	Numeric	8	0		None	None	4	≡ Right	🕸 Nominal	🚫 None
2	Y	Numeric	8	0	Mean arterial bl...	None	None	4	≡ Right	🖉 Scale	◎ Target
3	X1	Numeric	8	0	age (years)	None	None	4	≡ Right	🖉 Scale	⬊ Input
4	X2	Numeric	8	1	weight (kg)	None	None	4	≡ Right	🖉 Scale	⬊ Input
5	X3	Numeric	8	2	body surface ar...	None	None	4	≡ Right	🖉 Scale	⬊ Input
6	X4	Numeric	8	1	duration of hyp...	None	None	4	≡ Right	🖉 Scale	⬊ Input
7	X5	Numeric	8	0	basal pulse (be...	None	None	4	≡ Right	🖉 Scale	⬊ Input
8	X6	Numeric	8	0	measures of str...	None	None	4	≡ Right	🖉 Scale	⬊ Input

We obtain the Automatic Linear Modeling as follows:

Analyze→ regression → Automatic Linear Modeling …

The dependent variable (Target) and the independent variables (Predictors or inputs) will be chosen in an automatic manner:
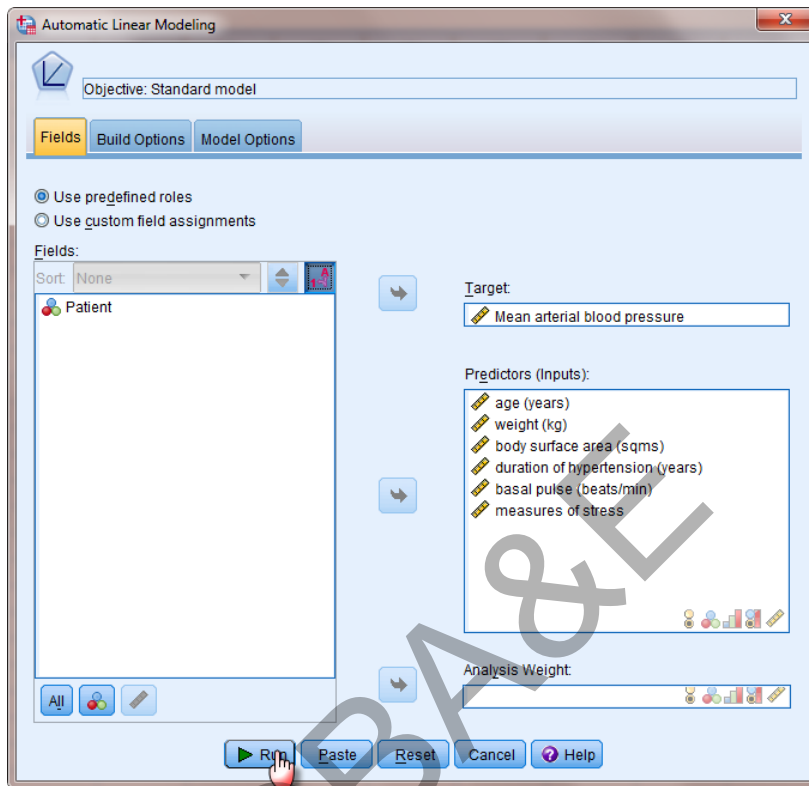
We just click on Run to get the following results:

The Forward Stepwise was chosen automatically and the R^2 is given as the "Accuracy" with the value of 99.4%.

Many features can be study from the Automatic Linear Modeling, we will mention the most important two of them:

When we click on , the significant independent variables will be shown with corresponding Importance as predictors for the Target (dependent variable):

When we click on , and chose "Table", we get the Coefficients of the model:

Coefficients
Target: Mean arterial blood pressure

Model Term	Coefficient ▶	Sig.	Importance
Intercept	-13.667	.000	
X2_transformed	0.906	.000	0.564
X1_transformed	0.702	.000	0.420
X3_transformed	4.627	.008	0.015

Case 2: Multiple regression analysis when qualitative variables are involved as independent variable(s)

Multiple regression analysis can also be performed if in the data, qualitative (non-metric) independent variables are also involved. Qualitative variables, unlike quantitative (metric) variables, cannot be measured on a numerical scale. Therefore, we need to code the values of the qualitative variable (called levels) before we perform regression analysis. These coded variables are called *dummy variables,* since the numbers assigned to various levels are selected arbitrarily.

A convenient method of coding the values of a qualitative variable at two levels involves assigning a *value one* to one of the levels and a *value zero* to another. For example, the dummy variable used to describe smoking status could be coded as follows:

$$\text{Smoking status } X = \begin{bmatrix} 1 = \text{smoker} \\ 0 = \text{non} - \text{smoker} \end{bmatrix}$$

The choice of which level is assigned to 1 and which is assigned to 0 is arbitrary (nominal scale). The advantage of using a, (0, 1) coding scheme is that the β -coefficients are easily interpreted. This is explained as:

It is a common observation that smoker mothers give birth to babies with low weight as compared to non-smoker mothers. We can write a model for average weight of babies as

$$E(Y) = \beta_0 + \beta_1 X$$

The dummy variable used to describe smoking status could be coded as:

$$X = \begin{bmatrix} 1 = & \text{smoker} \\ 0 = & \text{non} - \text{smoker} \end{bmatrix}$$

The model allows us to compare the *average weight* of smoker and non-smoker mothers.

$$\text{Smoker mother } (X = 1): E(Y) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$$
$$\text{Non-smoker mother } (X = 0): E(Y) = \beta_0 + \beta_1(0) = \beta_0$$

First note that β_0 represents the average weight of babies with non-smoker mothers. When a 0-1 coding convention is used, β_0 will always represent the mean response associated with the level of the qualitative variable assigned to value 0 (called the base level). The difference between the mean weight of the babies between smoker and non-smoker mothers is β_1, i.e.

$$\mu_{(NS)} - \mu_{(S)} = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

Therefore, with the 0-1 coding convention, β_1 will always represent the difference between mean responses for level assigned the value 1 and the mean for the base level.

For models that involve the qualitative independent variable at more than two levels,

additional dummy variables must be created. *In general, the number of dummy variables used to describe a qualitative variable will be one less than the number of levels of the qualitative variable, i.e.*

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$X_1 = \begin{cases} 1 & \text{if level A} \\ 0 & \text{if not} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if level B} \\ 0 & \text{if not} \end{cases}$$

Base level = level C.

Interpretation of β's will be as:

β_0 = mean level of base level

β_1 = mean level of base A - mean level of base C

β_2 = mean level of base B - mean level of base C

To interpret β's we write:

Level 1: $X_1 = 1$ if A otherwise 0

$$E(Y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

Level 2: $X_2 = 1$ if B otherwise 0

$$E(Y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Level 3: $X_1 = 0$ $X_2 = 0$

$$E(Y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

β_0 = mean of the base (level 3)

β_1 = mean of the base level (1) - mean of the base level 3

β_2 = mean of the base level (2) - mean of the base level 3

Example 6.5:

Following data based on a random sample of 32 births regarding smoking and non-smoking mothers. The birth weight of each baby at the time of birth and gestation period for each mother was recorded. Using multiple-regression, analyze the data and interpret the results. Data is given on next page.

Solution:

In this problem there are three variables, one dependent (birth weight = Y) and two independent variables (gestation period = X_1 and smoking status = X_2). Smoking status is a qualitative variable.

Table 6.6

Case	Birth weight (grams) Y	Gestation (weeks) X_{1z}	Smoking status of mothers X_2	Dummy code S = 1 NS = 0	Dummy code S = 0 NS = 1
1	2940	38	S	1	0
2	3130	38	N	0	1
3	2420	36	S	1	0
4	2450	34	N	0	1
5	2760	39	S	1	0
6	2440	35	S	1	0
7	3226	40	N	0	1
8	3301	42	S	1	0
9	2729	37	N	0	1
10	3410	40	N	0	1
11	2715	36	S	1	0
12	3095	39	N	0	1
13	3130	39	S	1	0
14	3244	39	N	0	1
15	2520	35	N	0	1
16	2928	39	S	1	0
17	3523	41	N	0	1
18	3446	42	S	1	0
19	2920	38	N	0	1
20	2957	39	S	1	0
21	3530	42	N	0	1
22	2580	38	S	1	0
23	3040	37	N	0	1
24	3500	42	S	1	0
25	3200	41	S	1	0
26	3322	39	N	0	1
27	3459	40	N	0	1
28	3346	42	S	1	0
29	2619	35	N	0	1
30	3175	41	S	1	0
31	2740	38	S	1	0
32	2841	36	N	0	1

(Source: Daniel, 1991)

For smoking status, the answer is either smoker or not smoker. These are coded as:

$$X_2 = \begin{cases} 1 & \text{smoker} \\ 0 & \text{otherwise} \end{cases}$$
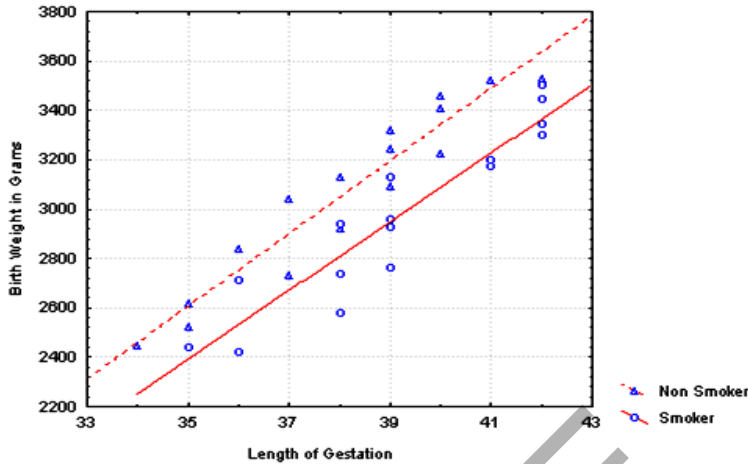
Fig. 6.6: Birth weight length of gestation (weeks)

Fitted regression lines for smoking (Δ) and non-smoking mothers (●).

SPSS package was used to fit multiple-regression and the output is as:

SPSS output for multiple regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.947[a]	.896	.889	115.5302

$R^2 = 0.896$, therefore one can say that 89.6% variation of birth weight of babies has been explained by gestation period and smoking status.

Adjusted $R^2 = 0.889$, the loss of predictive power by using this model is 0.3% [0.889 – 0.896]. Since $R^2 = 0.896$ and is closer to 1 therefore fitted model is reasonably reliable for prediction.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3348720	2	1674359.837	125.446	.000[a]
	Residual	387069.8	29	13347.235		
	Total	3735790	31			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2389.573	349.206		-6.843	.000
	Gestation (weeks)	143.100	9.128	.963	15.677	.000
	Smoking Status 1	-244.544	41.982	-.358	-5.825	.000

The fitted linear model will be:

Expected birth weight = -2389.573 +143.1(gestation)-244.544 (smoking status).

$$\hat{y} = -2389.573 + 143.1 X_1 - 244.544 X_2 \qquad (6.9)$$

If we wish to consider only the birth to smoking mothers, then put $X_2 = 1$ in the equation (6.9) then.

$$\hat{y} = -2389.573 + 143.1 X_1 - 244.544 (1)$$

or

$$\hat{y} = -2634.117 + 143.1 X_1 \qquad (A)$$

If we wish to consider only the births to non-smoking mothers, then put $X_2 = 0$ in the model as:

$$\hat{y} = -2389.573 + 143.1 X_1 - 244.544 (0)$$

or

$$\hat{y} = -2389.573 + 143.1 X_1 \qquad (B)$$

The slope of the equations (A) and (B) is the same, but there is difference in intercepts for smoking and non-smoking mothers. The intercept for the equations associated with non-smoking mothers is larger than smoking mothers. Therefore, we conclude from this sample that babies born to mothers who do not smoke, weighed, on the average, more than babies born to mothers who smoke, provided there is no change in gestation period. On the average, the amount of difference in weight is about 245 grams (2634.1 - 2389.5).

A general rule is stated below to interpret the result for qualitative variables.

General Rule

    (i) If the coefficient is negative, the higher code has negative effect.

    (ii) If the coefficient is positive, the higher code has positive effect.

Let us reconsider the equation (6.9)

$$\hat{y} = -2389.573 + 143.100 X_1 - 244.544 X_2$$

X_2 is a qualitative variable and coded as smoker = 1, non-smoker = 0. The coefficient of X_2 is negative and the code of smoker is 1, therefore, the mothers who smoke will give

birth to babies, who on the average will be less in weight than those babies born to non-smoking mothers.

If the smoker is coded as 0 and non-smoker as 1, then the output for multiple-regression, using SPSS is as:

SPSS output for multiple regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.947[a]	.896	.889	115.5302

R^2 and adjusted R^2 are the same as in the previous analysis.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3348720	2	1674359.837	125.446	.000[a]
	Residual	387069.8	29	13347.235		
	Total	3735790	31			

The result for the ANOVA is the same as the previous analysis.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-2634.117	358.872		-7.340	.000
	Gestation (weeks)	143.100	9.128	.963	15.677	.000
	Smoking Status 2	244.544	41.982	.358	5.825	.000

The regression equation is as:

$$\hat{y} = -2634.117 + 143.100\, X_1 + 244.544\, X_2 \qquad (6.10)$$

For smoking mothers put $X_2 = 0$ as:

$$\hat{y} = -2634.117 + 143.100\, X_1 + 244.544\,(0)$$
$$\hat{y} = -2634.117 + 143.100\, X_1 \qquad (C)$$

For non-smoking mothers, put $X_2 = 1$ as:

$$\hat{y} = -2634.117 + 143.100\, X_1 + 244.544\,(1)$$
$$\hat{y} = -2389.57 + 143.100\, X_1 \qquad (D)$$

The slopes of equations (C) and (D) are the same but there is difference in intercepts. The intercept for non-smoking mothers is greater than smoking mothers, therefore, non-smoking mothers, will give birth to a child on the average more than smoking mothers and again the difference in weight is 245 grams.

We can apply the general rule, mentioned before, to fitted regression equation (6.9). The code for non-smoker is 1, the coefficient of X_2 is positive, therefore, higher code has positive effect. Therefore, non-smoker mothers give birth to babies, who on the average are more in weight than smoking mothers. This rule can be applied to any qualitative variable when they are coded.

Example S6-6

The data will be in columns were X2 has a Nominal measurement level.

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Case	Numeric	8	0		None	None	4	Right	Scale	None
Y	Numeric	8	0	Birth weight (grams)	None	None	3	Right	Scale	Target
X1	Numeric	8	0	Gestation (weeks)	None	None	2	Right	Scale	Input
X2	Numeric	8	0	Smoking status of mothers	{0, Non Sm…	None	2	Right	Nominal	Input
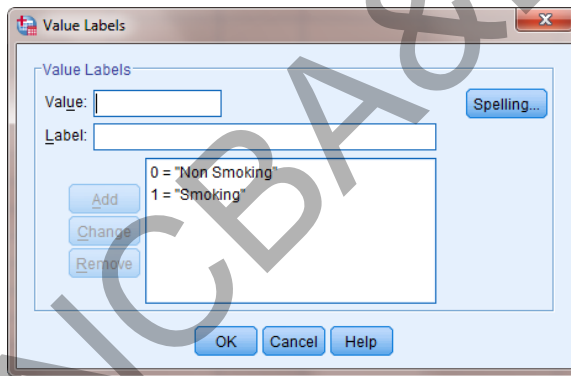
The values of X2 are as follow:

We obtain the multiple regression coefficients as follows:

Analyze→ regression → Linear …
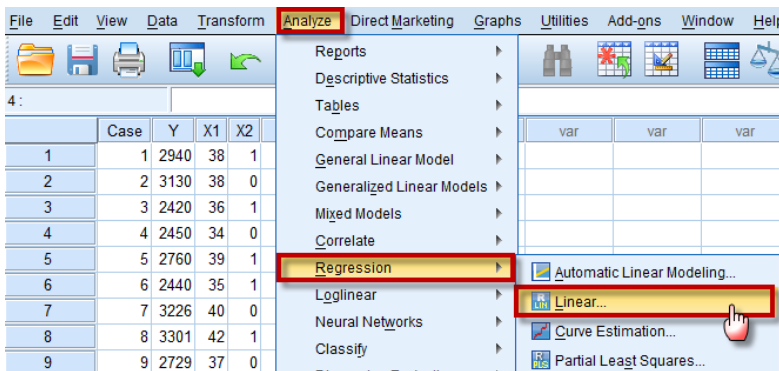
Move the variable Independent variables "X1, X2" to the Independent(s):

Move the variable Dependent variable "Y" to the Dependent:
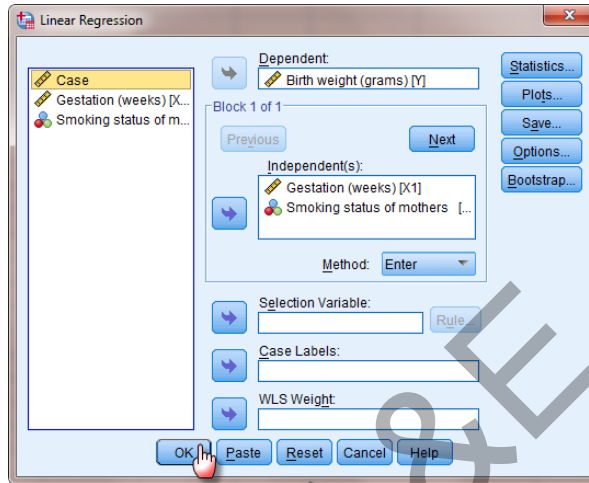
We click on OK , to get the following output (as before):

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2389.573	349.206		-6.843	.000
	Gestation (weeks)	143.100	9.128	.963	15.677	.000
	Smoking status of mothers	-244.544	41.982	-.358	-5.825	.000

a. Dependent Variable: Birth weight (grams)

We obtain the figure through the following steps:
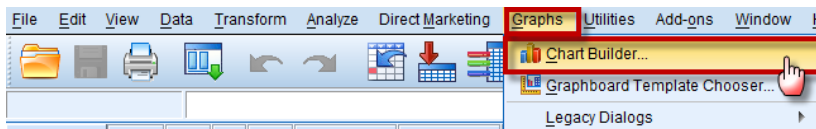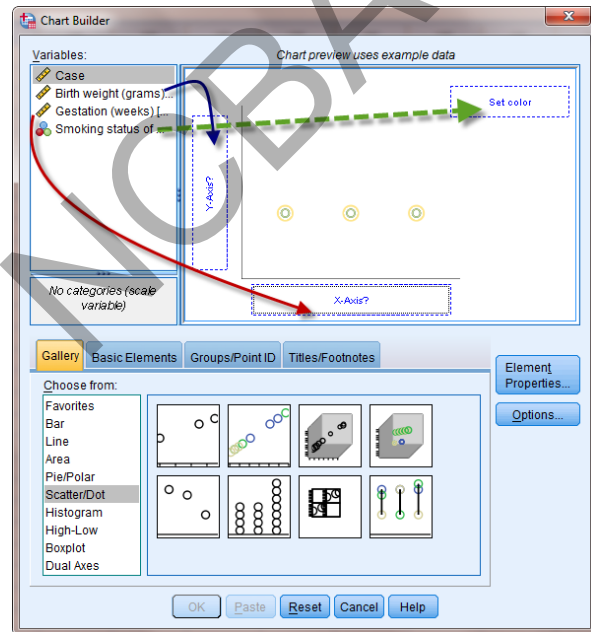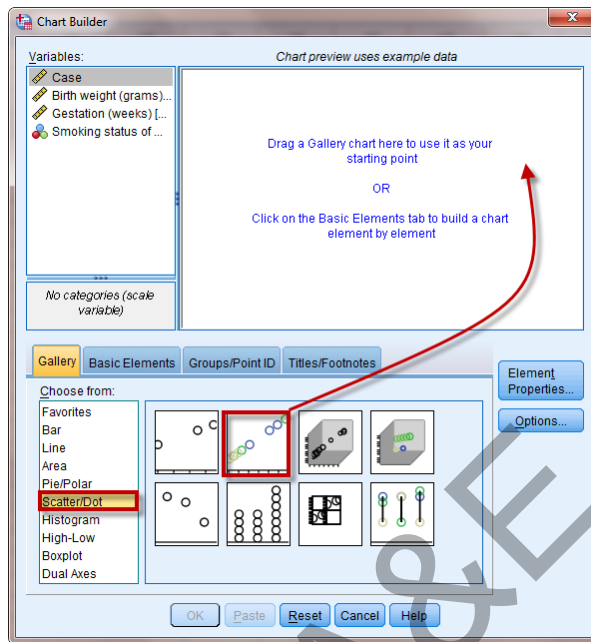
Graphs → Chart Builder …

We click on [OK], to get the following figure:
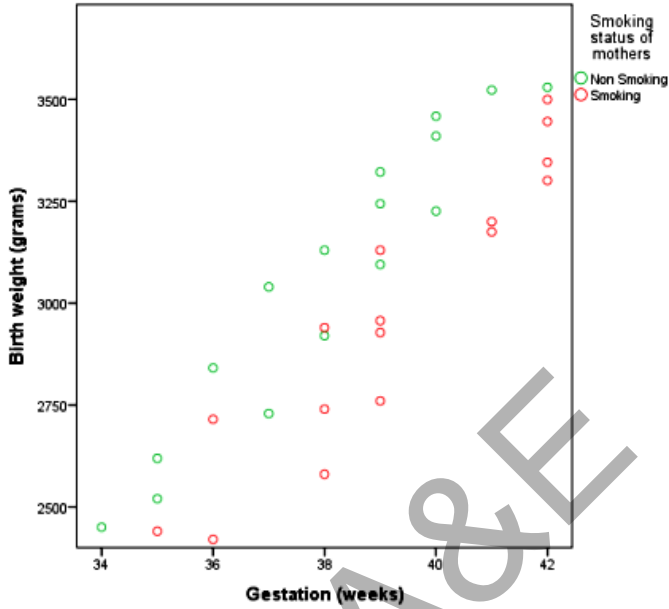
Double click the Fig to add change colors

At Chart Editor add lines through [icon]
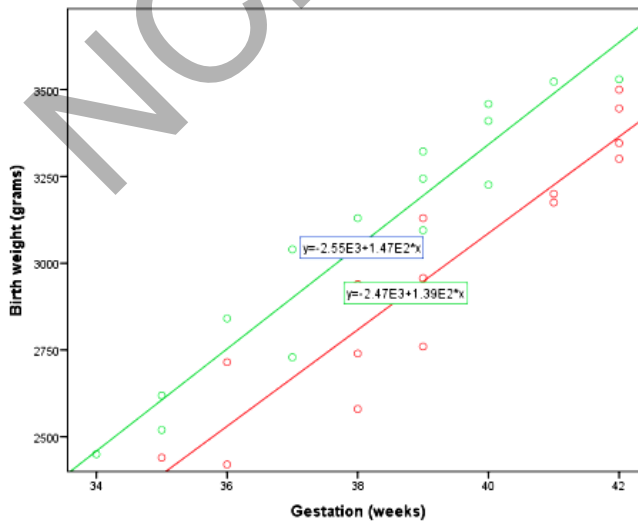
Case 3:   Multiple regressions when qualitative variable is of three levels

We, now consider the situations where the independent qualitative variable is of three levels.

Example 6.6:

A team of mental health researcher wishes to compare three methods A, B, C of treating severe depression. They took a sample of 36 patients and randomly assign the method of treatment.

> Y = measure of effectiveness
> X_1 = age of the patients
> X_2 = method of treatment

Use the method of regression to study (1) the relationship between age and treatment effectiveness (ii). The relationship between age and treatment effectiveness as well as interaction (if any) between age and treatment the data are given in Table (6.7)

Solution:

There are two variables.

  (i)   Age = X_1 (quantitative)
  (ii)   Method of treatment = X_2 (qualitative)

There are three levels A, B, C, therefore create two dummy variables say X_3 and X_4 as:

  (iii) If X_2 = A then $X_3 = 1$ and $X_4 = 0$
  (iv) If X_2 = B then $X_4 = 1$ and $X_3 = 0$
  (v)   If X_3 = C then $X_3 = 0$ and $X_4 = 0$

We want to consider the relationship between age and treatment effectiveness as well as an interaction (if any) between age and treatment.

Table 6.7

Y	X_1	X_2	y	X_1	X_2
56	21	A	65	43	A
41	23	B	55	45	B
40	30	B	57	48	B
28	19	C	59	47	C
55	28	A	64	48	A
25	23	C	61	53	A
46	33	B	62	58	B
71	67	C	36	29	C
48	42	B	69	53	A
63	33	A	47	29	B
52	33	A	73	58	A
62	56	C	64	66	B
50	45	C	60	67	B
45	43	B	62	63	A
58	38	A	71	59	C
46	37	C	62	51	C
58	43	B	70	67	A
34	27	C	71	63	C

Source (Daniel 1985)

If we use dummy variable the data will take the following form.

Table 6.8

Measure of effectiveness	Age X₁	Method of treatment X₂	Dummy variables X₃	X₄	Measure of effectiveness	Age X₁	Method of treatment X₂	Dummy variables X₃	X₄
56	21	A	1	0	65	43	A	1	0
41	23	B	0	1	55	45	B	0	1
40	30	B	0	1	57	48	B	0	1
28	19	C	0	0	59	47	C	0	0
55	28	A	1	0	64	48	A	1	0
25	23	C	0	0	61	53	A	1	0
46	33	B	0	1	62	58	B	0	1
71	67	C	0	0	36	29	C	0	0
48	42	B	0	1	69	53	A	1	0
63	33	A	1	0	47	29	B	0	1
52	33	A	1	0	73	58	A	1	0
62	56	C	0	0	64	66	B	0	1
50	45	C	0	0	60	67	B	0	1
45	43	B	0	1	62	63	A	1	0
58	38	A	1	0	71	59	C	0	0
46	37	C	0	0	62	51	C	0	0
58	43	B	0	1	70	67	A	1	0
34	27	C	0	0	71	63	C	0	0

The SPSS package is used and the output is as:

SPSS output for multiple regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.956[a]	.914	.900	3.92

a. Predictors: (Constant), X1X4, Age, X3, X4, X1X3

$R^2 = 0.914$, therefore about 91% of the variation of dependent variable, measure of effectiveness, has been explained by the independent variables.

Adjusted $R^2 = 0.900$ therefore one can say that the loss of prediction power by using this model is 0.14%.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4932.852	5	986.570	64.043	.000[a]
	Residual	462.148	30	15.405		
	Total	5395.000	35			

a. Predictors: (Constant), X1X4, Age, X3, X4, X1X3

b. Dependent Variable: Measure of Effectiveness

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	6.211	3.350		1.854	.074
	Age	1.033	.072	1.218	14.288	.000
	X3	41.304	5.085	1.591	8.124	.000
	X4	22.707	5.091	.874	4.460	.000
	X1X3	-.703	.109	-1.298	-6.451	.000
	X1X4	-.510	.110	-.922	-4.617	.000

a. Dependent Variable: Measure of Effectiveness

The fitted regression line will be

$$\hat{y} = 6.211 + 1.033 \text{ age} + 41.304X_3 + 22.707X_4 - 0.703X_1X_3 - 0.510X_1X_4$$

X_1X_3 and X_1X_4 are the interactions terms between quantitative and qualitative variables. X_3X_4 will be zero as when $X_3 = 1$, $X_4 = 0$ and $X_4 = 1$, $X_3 = 0$.

Put $X_3 = 1$, $X_4 = 0$, we get

$$\hat{y} = 47.515 + 0.33 \ X_1$$

If we put $X_3 = 0$ and $X_4 = 1$, then

$$\hat{y} = 28.918 + 0.5233 \ X_1$$

If we put $X_3 = 0$ and $X_4 = 0$, then

$$\hat{y} = 6.211 + 1.033 \ X_1$$

In order to draw the conclusion, one can look into slopes and the constants, i.e.

Table 6.9

Intercept (constant)	Slope	$\text{Tan}^{-1} \theta$
47.52	0.327	18.11°
28.91	0.523	27.61°
6.21	1.020	45.57°

We draw graph with the given angles and intercepts as given in Fig. 6.7:

Slope of A and B are not much different but there is much difference in the intercept. Looking at the graph, we can say:

(i) Treatment A is better than treatment B up till the age of 65 but this difference is very small after age 65.

(ii) Treatment C is less effective at younger age but it is as effective as treatment A and treatment B at higher age.

Now look at intercepts: Treatment A has higher intercept value than B and C. C has the minimum intercept. We can say that on the average treatment A is more effective than B, C is less effective at younger ages.

Now we look at the slopes: Treatment C has slope 1.033 which is higher than the other two, so one can say that at later stage this treatment is more effective than B and C. The difference in slopes of B and C are not much, therefore, at later stage both have almost equal effect.
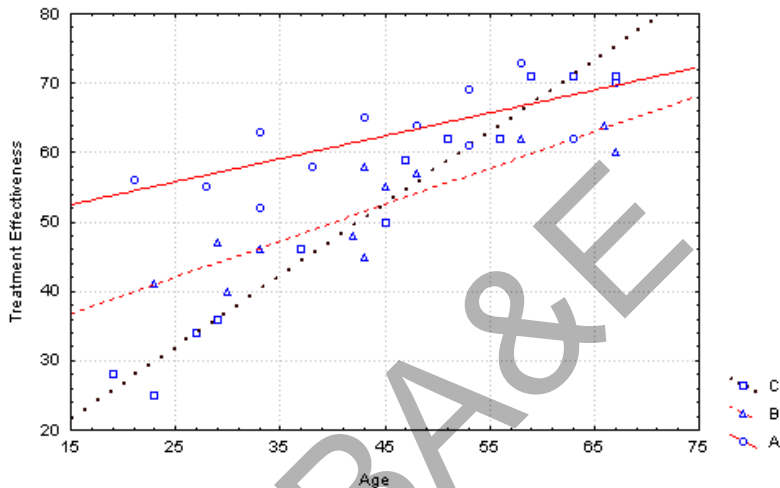
Fig. 6.7: Treatment effect

Example S6-7

The data will be in columns were X2 has a Nominal measurement level.

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Case	Numeric	8	0		None	None	4	Right	Nominal	None
Y	Numeric	8	0	Measure of effectiveness	None	None	3	Right	Scale	Target
X1	Numeric	8	0	Age	None	None	2	Right	Scale	Input
X2	Numeric	1	0	Method of treatment	{1, A}…	None	3	Right	Nominal	Input
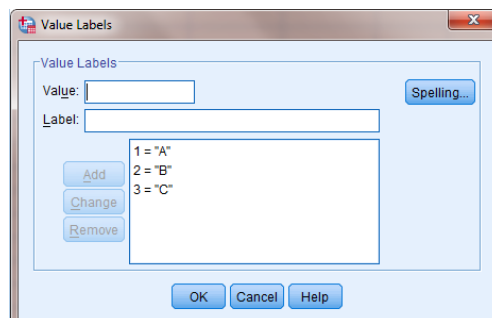
The values of X2 are as follow:

We obtain the figure through the following steps:
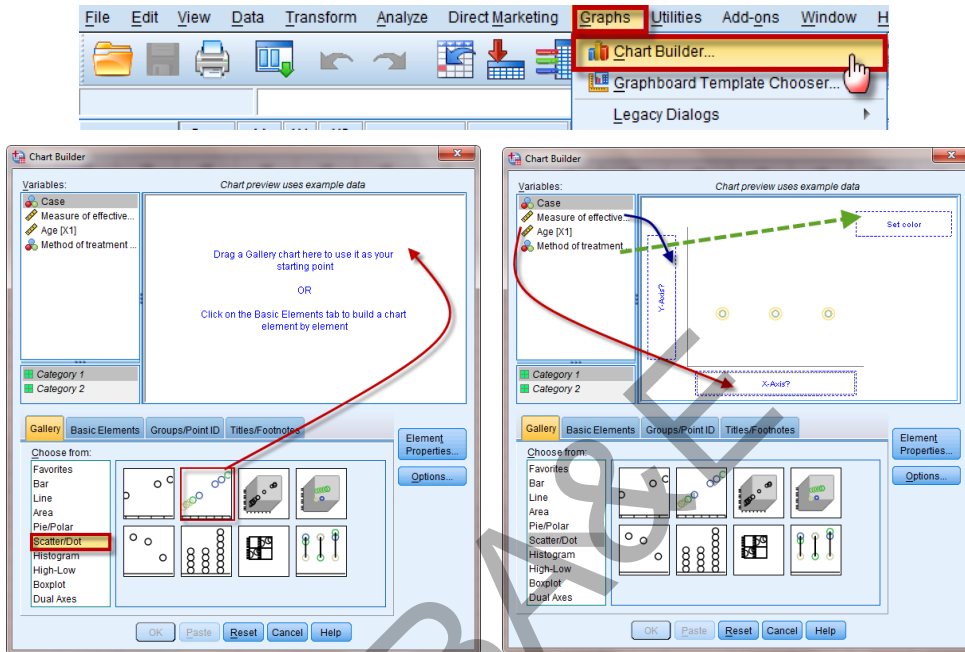
Graphs → Chart Builder …

We click on [OK], to get the following figure:
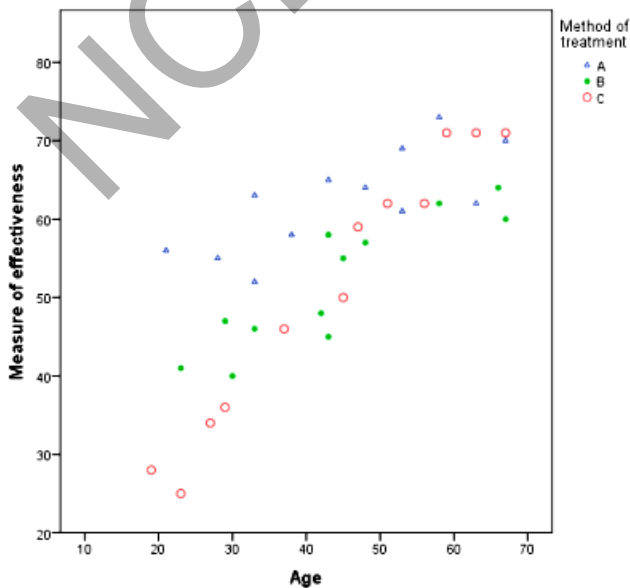
Double click the Fig to add change colors

At Chart Editor add lines through 📈
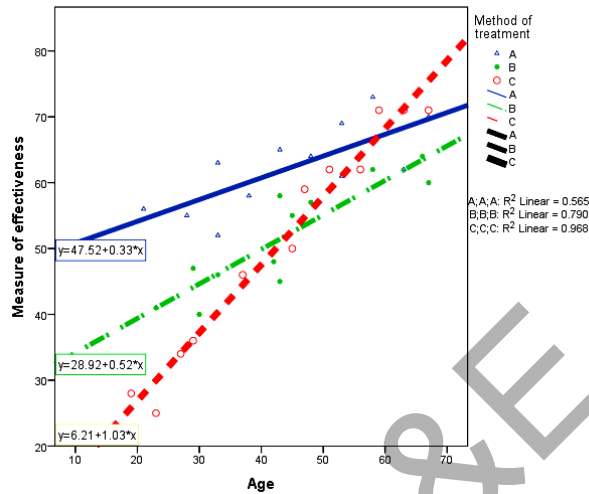
Example 6.7:
Data for the risk factors given in the <u>Appendix</u> associated with low infant birth weight are given at the end of Chapter. Data were collected at Baystate Medical Center, Spring Field, Massachusetts, during 1986 for 189 females. The code sheet for these data is provided as:

Variables and Code	Abbreviation
Age of mother in years	AGE
Weight of mothers at the last menstrual period (pounds)	LWT
Smoking status (1 = yes; 0 = no)	SMOKE
Race (1 = white, 2 = black, 3 = others)	RACE
History of premature labor (0 = none, 1 = one)	PTL
History of hypertension (1 = yes, 0 = no)	HT
Pressure of uterine irritability (1 = yes, 0 = no)	UI
Number of physician visits (0 = none, 1 = one, 2 = two)	FTV
Birth weight in grams	BWT

Use the multiple-regression to analysis the data and interpret the result. Data are given in the Appendix at the end of this chapter.

Solution:
In this example, there are 9 variables. Birth weight in grams (BWT) is dependent variable whereas all others are independent variables. Age and weight of mother are quantitative variables whereas all others are categorical variables. Race and number of visits of physicians have more than two categories, therefore, dummy variables will be created for these two variables, such as, two dummy variables for race and three dummy variables for number of visits of physicians. For the race two dummy variables may be created as:

$$Race\ 1 = \begin{cases} 1 & \text{if race = white} \\ 0 & \text{otherwise} \end{cases}$$

$$Race\ 2 = \begin{cases} 1 & \text{if race = black} \\ 0 & \text{otherwise} \end{cases}$$

Race = others If race 1 = 0 and race 2 = 0

Similarly dummy variables may be created for FTV. Because of complex calculation, SPSS package has been used and output is given as:

SPSS output for multiple regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics R Square Change	Sig. F Change	Durbin-Watson
5	.492	.242	.222	643.26688	.026	.013	.556

ANOVA

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	24203675.763	5	4840735.153	11.698	.000
Residual	75723987.549	183	413792.282		
Total	99927663.312	188			

Coefficients

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
	(Constant)	3041.608	75.924		40.061	.000
	Pressure of uterine irritability	-539.739	133.708	-.264	-4.037	.000
	History of hypertension	-656.615	201.852	-.211	-3.253	.001
	History of premature labor	-250.013	114.848	-.160	-2.177	.031
	RACE1	383.822	98.436	.264	3.899	.000
	Smoking Status	-283.067	112.951	-.190	-2.506	.013

Excluded Variables

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
Age of mother in years	.005	.068	.946	.005	.939
Weight of mothers at the last menstrual period (pounds)	.121	1.861	.064	.137	.967
Number of Physician visits	-.013	-.198	.843	-.015	.973
RACE2	-.009	-.130	.896	-.010	.812

$R^2 = 0.242$, this means that about 24% variation of dependent variable (birth weight) has been explained by the independent variables.

If we look at the output, hypertension (HT), premature birth (PTL), race1, smoking and uterine irritability (UI) appeared as significant variables, whereas age, number of physicians' visits (FTV), weight at the last menstrual period (LWT) and race2, appeared as non-significant variables.

The fitted regression model is

$$\hat{y} = 3596.619 - 530.610 \text{ HT} - 276.611 \text{ PTL} + 383.822 \text{ race 1}$$
$$- 283.067 \text{ smoke} - 539.739 \text{ UI}$$

The interpretation of these coefficients is as:

(i) History of hypertension (HT)
Since higher code is assigned for hypertensive cases and the coefficient for this variable is negative, therefore, all hypertensive cases will have a low weight at the time of birth on the average, provided all other variables are held constant.

(ii) History of premature (PTL)
Since higher code is assigned to premature cases and the coefficient is negative, therefore, all cases who have premature labour will have babies which will have less weight on the average, provided all other variables are held constant.

(iii) Race (1)
The coefficient of race (1) is positive. This indicates that white race will have the babies which on the average are more in weight than black and others provided all other variables are held constant. Note that other race is our reference point.

(iv) Smoking (smoke)
Since higher code is for non-smoker and the coefficient is negative, therefore, smoking mothers will give birth with low weight on the average, provided all other variables are held constant.

(v) Presence of urine irritability (UI)
Since higher code is for the presence of irritability and the coefficient is negative, therefore, all those cases who have urine irritability will have the babies with low weight on the average, provided all other variables are held constant.

6.6 Partial Correlation

It is a linear relationship between two variables when the effect of other variables has been removed (or kept constant). Here we stick to three variables only and it is explained by the following examples.

Example 6.8:
The following data obtained on 12 males between the ages of 12 and 18 years. Calculate all partial correlation coefficients.

Table 6.10

Height (1)	Radius length (2)	Femur length (3)
149.00	21.00	42.50
152.00	21.79	43.70
155.70	22.40	44.75
159.00	23.00	46.00
163.30	23.70	47.00
166.00	24.30	47.90
169.00	24.92	48.95
172.00	25.50	49.90
174.50	25.80	50.30
176.10	26.01	50.90
176.50	26.15	50.85
179.00	26.30	51.10

Solution:
Using SPSS package, partial correlation coefficients calculated and the SPSS output is as:

Variable	Mean	s. d.	cases
Height	166.0083	10.2065	12
Radius length	24.2392	1.8396	12
Femur length	47.8208	3.0132	12

$r_{23.1} = 0.9011$, $r_{13.2} = -0.0856$ $r_{12.3} = 0.5080$

$r_{12.3}$ = partial correlation between 1 and 2 while 3 is kept as constant.

Example S6-8

We obtain r23.1 = partial correlation between X2 (Radius length) and X3 (Femur length) while X1 (Height) is kept as constant, through the following steps:

Analyze → Correlate→ Partial …
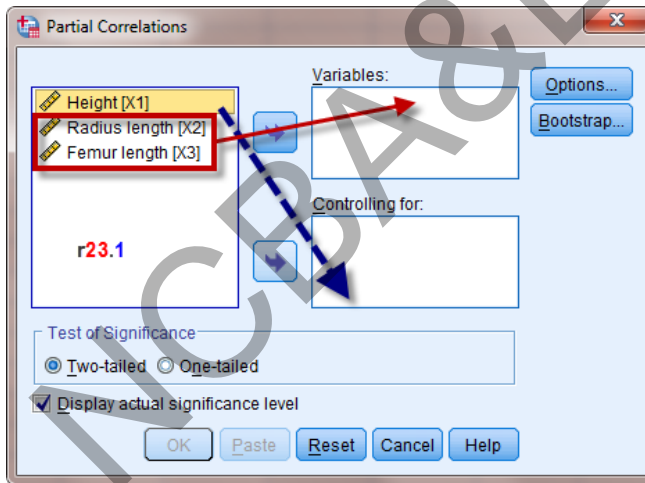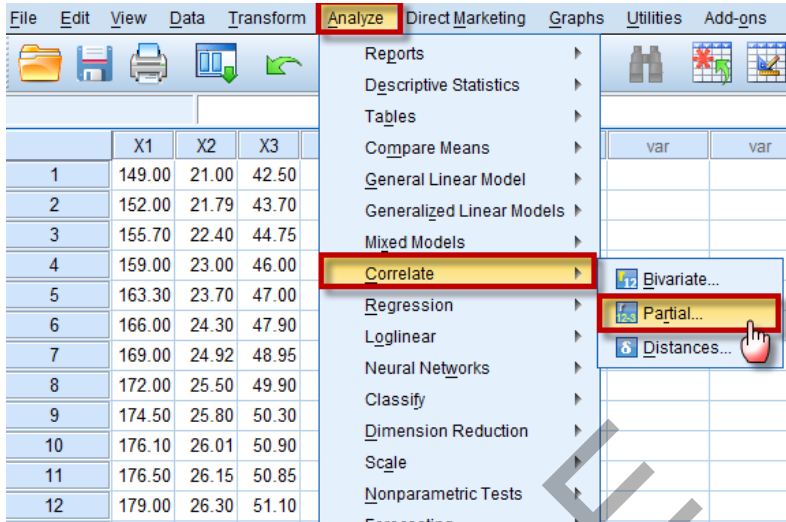
We click on OK, to get the following output:

Correlations

r23.1

Control Variables			Radius length	Femur length
Height	Radius length	Correlation	1.000	.901
		Significance (2-tailed)	.	.000
		df	0	9
	Femur length	Correlation	.901	1.000
		Significance (2-tailed)	.000	.
		df	9	0

6.7 Intra-Class Correlation Coefficient

In the previous section, we have discussed simple correlation coefficient (Pearson's correlation coefficient). Pearson correlation is based on regression analysis and is a measure of the extent to which the relationship between two variables can be described by a regression line. One of the properties of the correlation is that it provides a relative, rather than absolute, measure of agreement between pairs of scores for the same person. If the differences between the scores for the same persons are small relative to the differences between scores of different persons, then the test will tend to show a high reliability (Chapter 10). Conversely, if the differences between scores for the same persons are large relative to the scores of different persons, then the scores will show low reliability. Moreover, the perfect fit is obtained resulting in a Pearson correlation coefficient of 1.0 despite the fact that the intercept is non-zero and the slope is not equal to 1.

Let us consider the use of correlation coefficient to quantify measurement error. *Measurement error is the variation between measurements of the same quantity on the same individual.* A common design for the investigation of measurement error is to take pairs of measurements on a group of subjects. Following data relate to pairs of measurements of FEV (liters) a few weeks apart from 20 Scottish children taken from a large study

Table 6.11 (Measurements)

Subject No.	1st	2nd	Subject No.	1st	2nd
1	1.19	1.37	11	1.54	1.57
2	1.33	1.32	12	1.59	1.60
3	1.35	1.40	13	1.61	1.53
4	1.36	1.25	14	1.61	1.61
5	1.38	1.29	15	1.62	1.68
6	1.38	1.37	16	1.78	1.76
7	1.38	1.40	17	1.80	1.82
8	1.40	1.38	18	1.85	1.89
9	1.43	1.38	19	1.94	2.10
10	1.43	1.51	20	2.10	2.20

One way for the investigation of measurement error is to calculate the correlation coefficient between pairs of measurement. We know that in general, the correlation coefficient between repeated measurements depends on the variability between subjects. Samples containing subjects who differ greatly will produce *larger* correlation coefficients than will samples containing *similar* subjects. The correlation coefficient between the pairs of the above data is 0.96. Suppose we split this group in which we have measured forced expiatory volume in one second (FEV_1) into two sub samples, the first 10 subjects and the second 10 subjects. We see that the correlation coefficient for the first sub sample is r = 0.26 and for the second is r = 0.97. These values are not equal to full sample. Moreover if we change the order of even number of the sample then r = 0.94 which is not equal to 0.96. The Pearson correlation coefficient depends on the way the sample is chosen. If we select subjects to give a wide range of the measurements, the natural approach when investigating measurement error, this will inflate the correlation coefficient. The correlation coefficient between repeated measurements is often called the *reliability* of the measurement method. It is widely used in the validation of psychological measures such as scales of anxiety and depression, where it is known as the test-retest method of reliability (see Chapter 10).

Another problem with the use of correlation coefficient between the first and second measurements is that there is no reason to suppose that their order is important. If the order were important the measurement would not be repeated observations of the same thing. We have seen that reversing the order of some subjects the correlation coefficient is changed.

To avoid this problem we study *intra-class* correlation. *Intra-class correlation is the proportion of the total variance of an observation that is associated with the class to which it belongs.*

As already stated that perfect fit is obtained resulting in a Pearson correlation coefficient of 1.0 despite the fact that the intercept is non-zero and the slope is not equal to one, by contrast the intra-class correlation coefficient will yield a value 1.0 only if the observations on each subject are identical which indicate slope of 1 and intercept is zero. This suggests that Pearson correlation coefficient is an inappropriate and a liberal measure of reliability. The intra-class correlation coefficient estimates the average correlation among all possible orderings of pairs. It also extends easily to the case of more than two observations per subject, whereas it estimates the average correlation between all possible pairs of observations. The best way to calculate the intra-class correlation coefficient is via analysis of variance one way classification. In the above data there are 20 subjects and each subject has 2 observations. We have used SPSS package to perform ANOVA-one way. The results are given as:

Analysis of variance (one way)

Sources of Variables	df	Sum of squares	Mean sum of squares	F ratio	p-value
Between subjects σ_b^2	19	2.3638	0.1244	43.65	0.0000
Within subjects σ_W^2	20	0.0570	0.0029	-	-
Total σ_T^2	39	2.4208	-	-	-

The intra-class correlation may be calculated as:

$$R_I = \frac{m\, s_b^2 - s_T^2}{(m-1)\, s_T^2} \tag{6.11}$$

where m is number of observations per subject. Using (6.11)

$$R_I = \frac{2(2.3638) - 2.4208}{(2-1)\, 2.4208} = 0.953$$

The intra-class correlation coefficient is 0.953 with p = 0.000.

In practice, there will be not much difference between Pearson correlation coefficient and intra-class correlation coefficient for true measurements. If, however, there is a systematic change from the first measurement to the second, as might be caused by a learning effect, intra-class correlation coefficient will be less than Pearson correlation coefficient. If there were such an effect the measurements would not be made under the same conditions and so we would not measure reliability.

The correlation coefficient can be used to compare measurements of different quantities, such as different scales for measuring anxiety. We could make repeated measurements of all the quantities on the same subjects and calculate intra-class correlation coefficients.

Example 6.9:
The data in Table 6.14 relate to the repeated peak expiatory flow rate (PEFR) measurements for 20 school children. Use the method of intra-class correlation coefficient to quantify the measurement error.

Table 6.12

Child No.	PEFR (1/MIN)				Child No.	PEFR (1/MIN)			
	1st	2nd	3rd	4th		1st	2nd	3rd	4th
1	190	220	200	200	11	300	300	310	300
2	220	200	240	230	12	270	250	330	370
3	260	260	240	280	13	320	330	330	330
4	210	300	280	265	14	335	320	335	375
5	270	265	280	270	15	350	320	340	365
6	280	280	270	275	16	360	320	350	345
7	260	280	280	300	17	330	340	380	390
8	275	275	275	305	18	334	385	360	370
9	280	290	300	290	19	400	420	425	420
10	320	290	300	290	20	430	460	480	470

Solution:
There are 20 subjects and each subject has 4 items. We have performed analysis of variance one way classification to calculate intra-class correlation coefficient. SPSS package was used and the result for ANOVA was as follows:

SPSS out put
Analysis of variance (one way)

Source	df	Sum of squares	Mean sum of squares	F ratio	p-value
Between children σ_b^2	19	285318.4375	15016.7599	32.608	0.0000
Within children σ_W^2	60	27631.2500	460.5208	-	-
Total σ_T^2	79	312949.6875	-	-	-

Using (6.12), the intra-class correlation coefficient is

$$R_I = \frac{4(285318.4375) - 312949.6875}{(4 - 1)\,312949.6875} = 0.882$$

Therefore, the measurement error is $(1 - R_I)\,100 = (1 - 0.882)\,100 = 11.8$.

APPENDIX

age	Lwt	race	pti	smoke	ht	ut	ftv	bwt
19	182	2	0	0	0	1	0	2523
33	155	3	0	0	0	0	3	2551
20	105	1	0	1	0	0	1	2557
21	108	1	0	1	0	1	2	2594
18	107	1	0	1	0	1	0	2600
21	124	3	0	0	0	0	0	2622
22	118	1	0	0	0	0	1	2637
17	103	3	0	0	0	0	1	2637
29	123	1	1	1	0	0	1	2663
26	113	1	1	1	0	0	0	2665
19	95	3	0	0	0	0	0	2722
19	150	3	0	0	0	0	1	2733
22	95	3	0	0	1	0	0	2750
30	107	3	0	0	0	1	2	2750
18	100	1	1	1	0	0	0	2769
18	100	1	1	1	0	0	0	2769
15	98	2	0	0	0	0	0	2776
25	118	1	1	1	0	0	3	2782
20	120	3	0	0	0	1	0	2807
28	120	1	1	1	0	0	1	2821
32	121	3	0	0	0	0	2	2835
31	100	1	0	0	0	1	3	2835
36	202	1	0	0	0	0	1	2836
28	120	3	0	0	0	0	0	2863
25	120	3	0	0	0	1	2	2877
28	167	1	0	0	0	0	0	2877
17	122	1	1	1	0	0	0	2906
29	150	1	0	0	0	0	2	2920
26	168	2	1	1	0	0	0	2920
17	113	2	0	0	0	0	1	2920
17	113	2	0	0	0	0	1	2920
24	90	1	1	1	0	0	1	2948
35	121	2	1	1	0	0	1	2948
25	155	1	0	0	0	0	1	2977
25	125	2	0	0	0	0	0	2977
29	140	1	0	1	0	0	2	2977
19	138	1	0	1	0	0	2	2977
27	124	1	0	1	0	0	0	2992
31	215	1	0	1	0	0	2	3005
33	109	1	0	1	0	0	1	3033
21	185	2	0	1	0	0	2	3042
19	189	1	0	0	0	0	2	3062

age	Lwt	race	pti	smoke	ht	ut	ftv	bwt
23	130	2	0	0	0	0	1	3062
21	160	1	0	0	0	0	0	3062
18	90	1	0	1	0	1	0	3076
18	90	1	0	1	0	1	0	3076
32	132	1	0	0	0	0	3	3080
19	132	3	0	0	0	0	0	3090
24	115	1	0	0	0	0	2	3090
22	85	3	0	1	0	0	0	3090
22	120	1	0	0	1	0	1	3100
23	128	3	0	0	0	0	0	3104
22	130	1	0	1	0	0	0	3132
30	95	1	0	1	0	0	2	3147
19	115	3	0	0	0	0	0	3175
16	110	3	0	0	0	0	0	3175
21	110	3	0	1	0	1	0	3203
30	153	3	0	0	0	0	0	3203
20	103	3	0	0	0	0	0	3203
17	119	3	0	0	0	0	0	3225
17	119	3	0	0	0	0	0	3225
23	119	3	0	0	0	0	2	3232
24	110	3	0	0	0	0	0	3232
28	140	1	0	0	0	0	0	3234
26	133	3	2	1	0	0	0	3260
20	169	3	1	0	0	1	1	3274
24	115	3	0	0	0	0	2	3274
28	250	3	0	1	0	0	3	3303
20	141	1	2	0	0	1	1	3317
22	158	2	1	0	0	0	2	3317
22	112	1	2	1	0	0	0	3317
31	150	3	0	1	0	0	2	3321
23	115	3	0	1	0	0	1	3331
16	112	2	0	0	0	0	0	3374
16	135	1	0	1	0	0	0	3374
18	229	2	0	0	0	0	0	3402
25	140	1	0	0	0	0	1	3416
32	134	1	1	1	0	0	3	3430
20	121	2	0	1	0	0	f	3444
23	190	1	0	0	0	0	0	3459
22	131	1	0	0	0	0	1	3460
32	170	1	0	0	0	0	0	3473
30	110	3	0	0	0	0	0	3475
20	127	3	0	0	0	0	0	3487
23	123	3	0	0	0	0	0	3544
17	120	3	0	1	0	0	0	3572
19	105	3	0	0	0	0	0	3572

age	Lwt	race	pti	smoke	ht	ut	ftv	bwt
23	130	1	0	0	0	0	0	3586
36	175	1	0	0	0	0	0	3600
22	125	1	0	0	0	0	1	3614
24	133	1	0	0	0	0	0	3614
21	134	3	0	0	1	0	2	3629
19	235	3	0	1	0	0	0	3629
25	95	1	1	1	0	1	0	3637
16	135	1	0	1	0	0	0	3643
29	135	1	0	0	0	0	1	3651
29	154	1	0	0	0	0	1	3651
19	147	1	0	1	0	0	0	3651
19	147	1	0	1	0	0	0	3651
30	137	1	0	0	0	0	1	3699
24	110	1	0	0	0	0	1	3728
19	184	1	0	1	0	0	0	3756
24	110	3	1	0	0	0	0	3770
23	110	1	0	0	1	0	1	3776
20	120	3	0	0	0	0	0	3770
25	241	2	0	0	0	0	0	3790
30	112	1	0	0	0	0	1	3799
22	169	1	0	0	0	0	0	3827
18	120	1	0	1	0	0	2	3856
16	170	2	0	0	0	0	3	3860
32	186	1	0	0	0	0	2	3860
18	120	3	0	0	0	0	1	3884
29	130	1	0	1	0	0	2	3884
33	117	1	0	0	0	1	1	3912
20	170	1	0	1	0	0	0	3940
28	134	3	0	0	0	0	1	3941
14	135	1	0	0	0	0	0	3941
28	130	3	0	0	0	0	0	3969
25	120	1	0	0	0	0	2	3983
16	95	3	0	0	0	0	1	3997
20	158	1	0	0	0	0	1	3997
26	160	3	0	0	0	0	0	4054
21	115	1	0	0	0	0	1	4054
22	129	1	0	0	0	0	0	4111
25	130	1	0	0	0	0	2	4153
31	120	1	0	0	0	0	2	4167
35	170	1	1	0	0	0	1	4174
19	120	1	0	1	0	0	0	4238
24	116	1	0	0	0	0	1	4593
45	123	1	0	0	0	0	1	4990
28	120	3	1	1	0	1	0	709

age	Lwt	race	pti	smoke	ht	ut	ftv	bwt
29	130	1	0	0	0	1	2	1021
34	187	2	0	1	1	0	0	1135
25	105	3	1	0	1	0	0	1330
25	85	3	0	0	0	1	0	1474
27	150	3	0	0	0	0	0	1588
23	97	3	0	0	0	1	1	1588
24	128	2	1	0	0	0	1	1701
24	132	3	0	0	1	0	0	1729
21	165	1	0	1	1	0	1	1790
32	105	1	0	1	0	0	0	1818
19	91	1	1	1	0	1	0	1885
25	115	3	0	0	0	0	0	1893
16	130	3	0	0	0	0	1	1899
25	92	1	0	1	0	0	0	1928
20	150	1	1	1	0	0	2	1928
21	200	2	0	0	0	1	2	1928
24	155	1	1	1	0	0	0	1936
21	103	3	0	0	0	0	0	1970
20	125	3	0	0	0	1	0	2055
25	89	3	0	0	0	0	1	2055
19	102	1	0	0	0	0	2	2082
19	112	1	1	1	0	1	0	2084
26	117	1	1	1	0	0	0	2084
24	138	1	0	0	0	0	0	2100
17	130	3	1	1	0	1	0	2125
20	120	2	1	1	0	0	3	2126
22	130	1	1	1	0	1	1	2187
27	130	2	0	0	0	1	0	2187
20	80	3	1	1	0	1	0	2211
17	110	1	1	1	0	0	0	2225
25	105	3	0	0	0	0	1	2240
20	109	3	0	0	0	0	0	2240
18	148	3	0	0	0	0	0	2282
18	110	2	1	1	0	0	0	2296
20	121	1	1	1	0	1	0	2296
21	100	3	0	0	0	0	3	2301
26	96	3	0	0	0	0	0	2325
31	102	1	1	1	0	0	1	2353
15	110	1	0	0	0	0	0	2353
23	187	2	1	1	0	0	1	2367
20	122	2	1	1	0	0	0	2381
24	105	2	1	1	0	0	0	2381
15	115	3	0	0	0	1	0	2381
23	120	3	0	0	0	0	0	2395

age	Lwt	race	pti	smoke	ht	ut	ftv	bwt
30	142	1	1	1	0	0	0	2410
22	130	1	1	1	0	0	1	2410
17	120	1	1	1	0	0	3	2414
23	110	1	1	1	0	0	0	2424
17	120	2	0	0	0	0	2	2438
26	154	3	0	0	1	0	1	2442
20	105	3	0	0	0	0	3	2450
26	190	1	0	1	0	0	0	2466
14	101	3	1	1	0	0	0	2466
28	95	1	0	1	0	0	2	2466
14	100	3	0	0	0	0	2	2495
23	94	3	0	1	0	0	0	2495
17	142	2	0	0	1	0	0	2495
21	130	1	0	1	1	0	3	2495

Chapter 7

Analysis of Categorical Data

7.1 Introduction

The chi-square test is often used in experimental work where the data consist of frequencies or counts. For example, the number of boys and number of girls in a class who have had their tonsils out is distinct from quantitative data obtained from the measurement of continuous variable such as height, weight, temperature and so on.

The most common use of the test is probably with categorical data such as level of education, marital status, etc. The test can also be used in experiments designed to assess the effect of inoculation in immunizing people against disease and in clinical trials involving drugs.

The test is frequently employed to determine if there is an *association between variables*. When the word *association* is used in the statistical sense, a comparison is implied. For example, if we say that there is an association between inoculation and immunization against some disease, we mean that *proportion of inoculated people* who contracted disease is different from the proportion of not inoculated people who do so. Of course the two proportions might be expected to differ in some measure due to chance factor of sampling, and for other reasons which might be attributed to *random causes*, but the test enables us to calculate the probability that a difference as great as or greater than that obtained could have arisen in this way.

Before we introduce the test, it would be better to illustrate the word *classification.* It is possible to classify a population in many different ways. For instance, population may be classified as males and females, married and unmarried, smokers and nonsmokers, etc. These classifications are known as *dichotomous* classifications. If the population is divided into more than two groups, like poor, good, very good, and high, medium, low education, etc., then these classifications are known as *multiple (polychotomous) classifications.* If the classification is dichotomous or multiple, it must be exhaustive and mutually exclusive. An example of dichotomous classification is given in Table 7.1.

Table 7.1
Number of patients with hypertension and no hypertension

		Stroke history		
		Present (+)	**not-present (-)**	
Hyper tension History	**Present (+)**	a 15	b 185	a+b = 200
	Not-present (-)	c 5	d 795	c+d = 800
		a+c = 20	b+d = 980	a+b+c+d = 1000

This table is known as 2x2 contingency table or two-dimensional table or fourfold *contingency table*. The entries in the cells of the Table (7.1) may be *frequencies* and may be transformed into *proportions* or *percentages*. The frequencies of four cells may be represented by a, b, c, d. An example of multiple classification, which is called 2×4 contingency table is given in Table 7.2:

Table 7.2
Distribution of patients by Diet and cancer tumor

		Diet				
		high fat no fiber	**High fat Fiber**	**low fat no fiber**	**low fat fiber**	**Total**
Cancer	**Yes**	27	20	19	14	80
Tumor	**No**	3	10	11	16	40
	Total	30	30	30	30	120

Note that contingency table is always read as row (r) by column (c) i.e. r×c. It is important to note that, in whatever form the entries are presented; the data are originally frequencies or counts. Of course, for the application of the chi-square test, continuous data can often be put into discrete form. For example, weight is a continuous variable, but if population is classified into different weight groups then different weight groups can be treated as if they were discrete groups. Below are given some examples where the chi-square test is applicable to test the association.

   (i)   Cigarette smoking and premature death from cardiovascular disease.

   (ii)   Smoking and lung cancer

   (iii) Smoking and myocardial infarction.

   (iv)   Post laparotomy wound infection in patients receiving antibiotic versus placebo.

   (v)   Two chemotherapy regiments for advanced acute lymphoblastic leukemia in children.

   (vi)   Nutritional status and academic performance.

   (vii) Incidence of miscarriage among woman exposed to agricultural pesticides.

   (viii) Fat diet and cancer tumor.

7.2 Assumptions

   (i)   The sample must be random so that the observations are independently distributed.

   (ii)   Each individual or unit in the sample has the same probability being from a particular cell and the sample is large.

   (iii) Each observation may be categorized either into class 1 or class 2, etc.

7.3 Uses of Chi-Square Test

The chi-square test can be used in different forms to test:

(i)   The variance for a single sample. This has been discussed in Chapter 4.

(ii) Goodness of fit. (This is not described here as health scientists use it very rarely).

(iii) Independence of attribute and homogeneity of groups.

(iv) Association when the data have linear trend (Mantel-Haenszel).

(v)   Association in matched samples.

(vi) The significance of relative risk and odds ratio.

In the application of chi-square test, there are two sets of frequencies, one set is called *observed (actual)* frequencies and other set is called *expected* frequencies. Observed frequencies are those which we get from a sample and are categorized into two or more than two classifications. Expected frequencies are the number of observations in our sample that we would expect to observe if some null hypothesis about the variable is true. For example, if we have a sample of 39 patients, who visit the hospital in a particular time, 13 out of them are old, 15 are young and 11 are children. These will be known as observed frequencies while in this case we would expect that sample must contain 13 old persons, 13 young persons and 13 children. This distribution gives us expected frequencies. Since expected frequencies are not known, we can estimate them from observed frequencies under the same hypothesis. An example, showing calculations of expected frequencies, is given as:

Example 7.1:
In a study of the relation between blood type and disease, large samples of patients with peptic ulcer, patients with gastric cancer and control persons free from these diseases were classified as to blood type (O, A, B). The observed frequencies are as follows:

Table 7.3
Distribution of patients by blood type and Disease
Disease

Blood type	Peptic ulcer	Gastric cancer	Controls	Total
O	$983 = O_{11}$	$383 = O_{12}$	$2892 = O_{13}$	$R_1 = 4258$
A	$679 = O_{21}$	$416 = O_{22}$	$2625 = O_{23}$	$R_2 = 3720$
B	$134 = O_{31}$	$84 = O_{32}$	$570 = O_{33}$	$R_3 = 788$
Total	$C_1 = 1796$	$C_2 = 883$	$C_3 = 6087$	$n = 8766$

(Source: Snedecor and Cochran, 1980)

O_{ij} is the observed frequencies in the (i,j)th cell.

If we assume that disease and blood type are independent then the expected frequencies are calculated as:

$$E_{ij} = \frac{R_i \times C_j}{n}$$

where R_i are the ith row total and C_j the jth column total.

Thus we have:

$$E_{11} = \frac{1796 \times 4258}{8766}, \quad E_{22} = \frac{883 \times 3720}{8766}, \text{ etc.}$$

Table 7.4
Distribution of expected patient by blood type and Disease

Blood type	Peptic ulcer	Gastric cancer	Controls	Total
O	$872.39 = E_{11}$	$428.91 = E_{12}$	$2956.70 = E_{13}$	4258
A	$762.16 = E_{21}$	$374.72 = E_{22}$	$2583.12 = E_{23}$	3720
B	$161.45 = E_{31}$	$79.37 = E_{32}$	$547.18 = E_{33}$	788
Total	1796	883	6087	8766

7.4 Independence and Homogeneity

7.4.1 2x2 Contingency Table

This test can also be thought of as a test of difference between two proportions.

Example 7.2:
Following data relate to deaths of males and females due to T.B.

Table 7.5
Observed frequencies of deaths by gender and form of T.B. Gender

Form of T. B	Males	Females	Total
T.B. of respiratory system	3534	1319	4853
Other form of T.B.	270	250	520
Total	3804	1569	5373

Are the two classifications of the people in the sample independent? (Maxwell, 1961)

Solution:
(1) H_0: Classification of people and the form of T.B. from which people die are independent.

   H_1: They are not independent. (There is association)

(2) $\alpha = 0.05$

(3) test-statistic: Chi-square   (data is qualitative)

   (i) Chi-square $(\chi^2) = \sum_i \sum_j \frac{\left(O_{ij} - E_{ij}\right)}{E_{ij}}$,   (7.1)

   (where O_{ij} = observed and E_{ij} = expected frequencies).

(ii) If it is a 2 x 2 contingency table, then calculation may be simplified by using the following formula:

$$\chi^2 = \frac{(ad - bc)^2 n}{(a + b)(c + d)(a + c)(b + d)} \qquad (7.2)$$

where $n = a + b + c + d$: The placement of a, b, c and d is shown in Table 7.1.

(4) To calculate chi-square we need expected frequencies, the calculations of expected frequencies have been explained in table 7.4 and for this example are given in Table 7.5.

Table 7.6
Expected Frequencies of deaths by gender and form of T.B. Gender

Form of T.B	Males	Females	Total
T.B. of Respiratory system	$E_{11} = \dfrac{3804 \times 4853}{5373}$ $= 3435.8$	$E_{12} = \dfrac{1569 \times 4853}{5373}$ $= 1417.2$	4853
Other form of T.B.	$E_{21} = \dfrac{3804 \times 520}{5373}$ $= 368.2$	$E_{22} = \dfrac{1569 \times 522}{5373}$ $= 151.8$	520
Total	3804	1569	5373

The chi-square value is calculated as:

(O - E)	$(O - E)^2$	$(O - E)^2 / E$
98.2	9643.24	2.807
-98.2	9643.24	6.804
-98.2	9643.24	26.190
98.2	9643.24	63.526
Total		99.327

$$\chi^2 = 99.326$$

If we use (7.2), then there is no need to calculate expected frequencies. We can use the observed frequencies directly to calculate chi-square.

$$\chi^2 = \frac{(3534 \times 250 - 270 \times 1319)^2 5373}{(4853)(520)(3804)(1569)} = 99.213$$

(There is a difference in result between two methods. This is because in first method approximation is involved. So it may be better to use the second form).

(5) Since it is a one-sided test we can see the table value for the desired degree of freedom under chi-square 0.95 for 5% level of significance. The degree of freedom is determined as (r - 1) (c - 1) = (2 - 1) (2-1) = 1. (Note that in 2 x 2 table, degree of freedom is always 1). (See table of χ^2 given at the end of this Chapter).

(6) The calculated value is 99.213, which is greater than table value (3.841) for one degree of freedom Therefore, the data do not show that the two variables are independent and we say with 95% confidence that two classifications of the people in our sample are not independent (see Fig. 7.1).

$$\chi^2_{0.95}$$
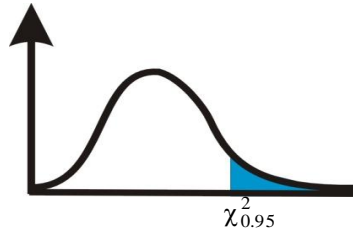
Fig. 7.1

To put it differently we may say that distribution of type of TB does depend on sex. In the application of chi-square, one point to be noted about the magnitude of the expected frequencies. *If the expected frequencies are too small then chi-square will not reflect the departure of observed from expected frequencies.*

There is no general rule regarding the minimum value of the expected/observed frequencies, but values of 3, 4 or 5 are widely used as minimum. If one should get expected/observed frequencies too small, it can be combined with expected/observed frequencies in an adjacent class interval. Generally, if it is less than 5 then Pearson's chi-square is not strictly valid.

7.4.2 Phi Coefficient

The Phi coefficient is a degree of association between two attributes and is calculated as:

$$\text{Phi} = \phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \sqrt{\frac{\chi^2}{n}} \tag{7.3}$$

$$= \sqrt{\frac{99.21346}{5373}} = 0.13589$$

The degree of association between death of people and form of T.B., with which people die, is about 13.6%. The range of ϕ is from -1 to 1. If ϕ is 0, the attributes are independent. If $\phi = 1$, there is complete positive association and for -1 there is complete negative association. This happens only when entries are only in the leading diagonal when $b = c = 0$ and consequently $\phi = 1$ (or $a = d = 0$). This measure is not very satisfactory since it does not necessarily have an upper limit of 1. This is used when scale is nominal.

7.4.3 Contingency coefficient (C)

It also measures the degree of association. This coefficient lies between 0 and 1 and attains its lower limit in case of complete independence, that is when $\chi^2 = 0$. It is also calculated when scale is nominal. It is calculated as:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{7.4}$$

$$= \sqrt{\frac{99.21346}{99.21346 + 5373}} = 0.1346 \text{ or } 13.5\%$$

C cannot attain its upper limit even in case of complete association.

7.4.4 Cramer's-V (V)

This coefficient also measures the degree of association. For 2x2 table Cramer's-V is identical to Phi. It is designed in such a way that it can attain upper bound 1. This is often used for general contingency table of size $r \times c$. It is calculated as:

$$V = \sqrt{\frac{\chi^2 / n}{\min(r-1, c-1)}} \tag{7.5}$$

$$= \sqrt{\frac{\dfrac{99.21346}{5373}}{(2-1)}} = 0.13589$$

7.4.5 Adjusted Chi-square (Yates' Correction)

Some times in 2 x 2 contingency table, expected frequency is less than 5 where pooling of data is impossible. Yates (1934) recommended an adjustment as *correction for continuity* known as *Yates' correction*. This is done by subtracting 1/2 from the positive discrepancies (O - E) and adding 1/2 to the negative discrepancies (O - E) before these values are squared. For this (7.1) takes the following form.

$$\chi^2 = \sum_i \sum_j \frac{\left[\left|O_{ij} - E_{ij}\right| - 0.5\right]^2}{E_{ij}} \tag{7.6}$$

Alternatively, this correction can be adjusted in (7.2).

$$\chi^2 = \frac{\left[|ad - bc| - 0.5n\right]^2 \ n}{(a+b)(c+d)(a+c)(b+c)}, \tag{7.7}$$

where |A| means absolute value of A. *(It is desirable to apply the Yates' correction at all times, whether or not expected frequencies are greater than 5, but it is essential to do so in cases when expected frequencies are less than 5 and sample size is small).* If the sample size is reasonably large, the correction will have little effect on the value of χ^2.

The same result may be obtained using IBM-SPSS package. The entry of the data for the calculation of chi-square has been explained in the next example.

Example S7-1

A part of the data will be in columns as follows:

	id	FormofT.B	gender
1	1	1	1
2	2	1	1
3	3	1	1
4	4	1	1
5	5	1	1
6	6	1	1
7	7	1	1
8	8	1	1
9	9	1	1
10	10	1	1

The Variable View is as follows:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	12	0		None	None	12	Right	Scale	Input
2	FormofT.B	Numeric	12	0	Form of T.B.	{1, T.B. of r...	None	12	Right	Nominal	Input
3	gender	Numeric	12	0		{1, Male}...	None	12	Right	Nominal	Input

The labels are defined as:
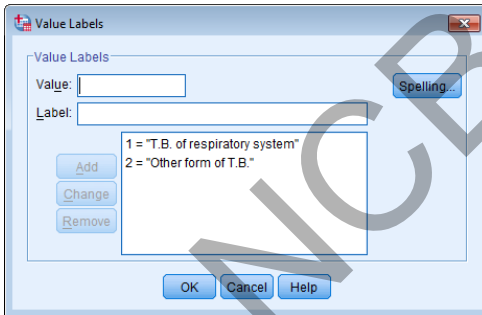
Form of T.B.:   Gender
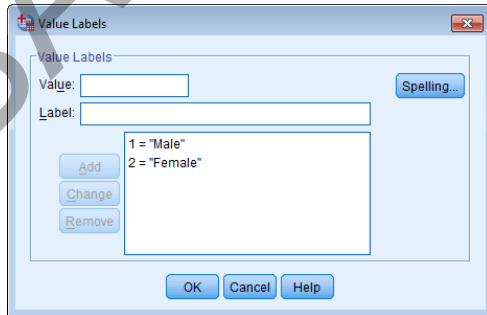
We apply the Chi-square test for Independence as follows:

Analyze→ Descriptive Statistics→ Crosstabs …
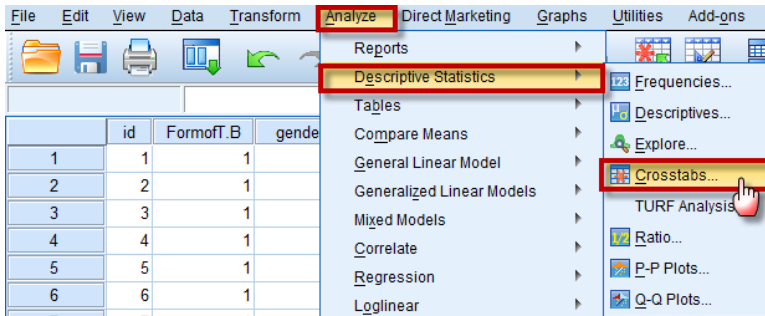
Move the variable "Form of T.B" to the Row(s):

Move the variable "gender" to the Column(s):

We click on [Statistics...] and mark on "Chi-square",

We also mark on "Contingency coefficient" and "Phi and Cramer's V" then click on [Continue]
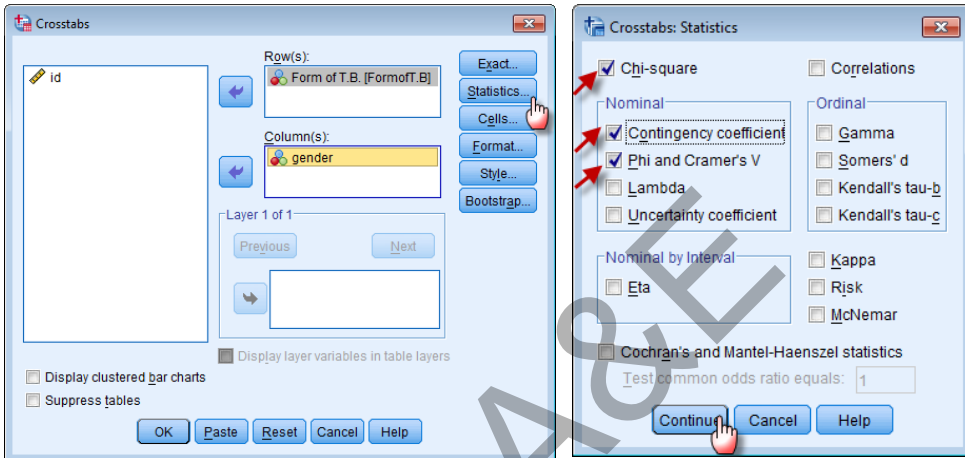
Now to show the expected values beside the observed values, and the percentages,

We click on [Cells...] and mark on Expected, Rows and Column, then click on [Continue]
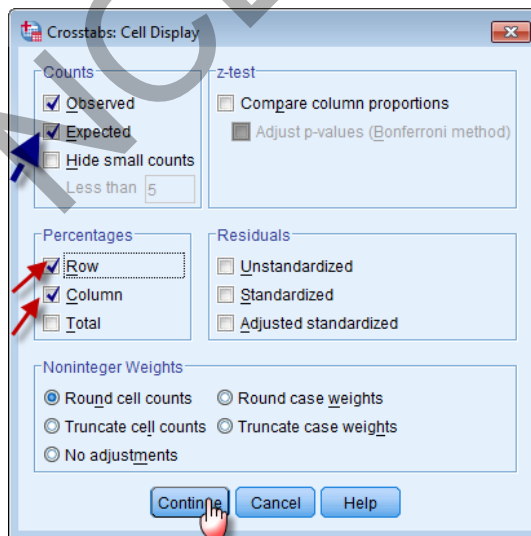
Now click on [OK], to get the following outputs:

SPSS output for chi-square

Form of T.B. * gender Crosstabulation

				gender		Total
				Male	Female	
Form of T.B.	T.B. of respiratory system	Count		3534	1319	4853
		Expected Count		3435.8	1417.2	4853.0
		% within Form of T.B.		72.8%	27.2%	100.0%
		% within gender		92.9%	84.1%	90.3%
	Other form of T.B.	Count		270	250	520
		Expected Count		368.2	151.8	520.0
		% within Form of T.B.		51.9%	48.1%	100.0%
		% within gender		7.1%	15.9%	9.7%
Total		Count		3804	1569	5373
		Expected Count		3804.0	1569.0	5373.0
		% within Form of T.B.		70.8%	29.2%	100.0%
		% within gender		100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	99.213[b]	1	.000		
Continuity Correction[a]	98.205	1	.000		
Likelihood Ratio	91.588	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	99.195	1	.000		
N of Valid Cases	5373				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 151.85.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.136	.000
	Cramer's V	.136	.000
	Contingency Coefficient	.135	.000
N of Valid Cases		5373	

Four results of chi-square are given at the end of the IBM-SPSS output, i.e. (i) chi-square Pearson, (ii) chi-square continuity correction, (iii) chi-square likelihood ratio and (iv) linear trend (Mantel-Haenszel). The important point is to choose the appropriate result, here we choose Pearson chi-square as the scale is nominal and no frequency in the cell is less than 5 (minimum expected frequency = 151.85), p-value = 0.000, which is less than 0.05 (observed p-value). We confirm our previous result. Note that there is small difference between the results of chi-square in our manual and computer calculations. Other forms of chi-square will be explained later. Phi, Cramer's V and contingency coefficient measure degree of association between two attributes and are calculated when scale is nominal.

Example 7.3:

The following data relate to suicidal feelings in samples of psychotic and neurotic patients:

Table 7.7
Distribution of Patients by type of patents and suicidal feelings

	Psychotics	Neurotics	Total
suicidal feelings	2	6	8
no suicidal feelings	18	14	32
Total	20	20	40

Test at 5% level of significance whether there is an association between two psychotics groups and the presence or absence of suicidal feelings.

Solution:

(1) H_0 : Two groups are independent with presence and absence of suicidal feelings.

   H_1 : Two groups are not independent.

(2) $\alpha = 0.05$

(3) test-statistic: Chi-square is applied, but we compute expected frequencies to see if Yates' correction can be applied?

Expected frequencies

	Psychotics	Neurotics	Total
Suicidal feelings	4	4	8
No suicidal feelings	16	16	32
Total	20	20	40

In two cells, expected frequencies are less than 5, so Yates' correction is applicable. Using (7.6), we have:

O - E	Corrected discrepancy	$(O - E)^2$	$(O - E)^2 / E$
-2	|-2| - 0.5 = 1.5	2.25	0.5625
2	| 2| - 0.5 = 1.5	2.25	0.5625
2	| 2| - 0.5 = 1.5	2.25	0.140625
-2	|-2| - 0.5 = 1.5	2.25	0.140625
Total			1.40625

The calculated value of $\chi^2 = 1.40625$

This can be solved by using the formula given in expression (7.2) as:

$$\chi^2 = \frac{\left[|\,2 \times 14 - 18 \times 6\,| - 0.5 \times 40\right]^2 (40)}{8 \times 32 \times 20 \times 20} = 1.40625$$

which is the same as above.

(4) The table value for 5% level of significance at 1 degree of freedom is $\chi^2_{0.95} = 3.841$.

(5) Calculated value is less than the table value, therefore, we say with 95% confidence that there is no evidence that psychotics and neurotics groups differ with respect to symptoms.

Note that minimum value of chi-square is zero. It is only possible when the expected minus observed value in each cell is zero.

7.4.6 Fisher's exact test

The method of Yates' correction was useful when manual calculations were done. Now different types of statistical packages are available. Therefore, it is better to use Fisher's exact test rather than Yates' correction as it gives exact result. It is used when expected frequency in the cell is less than 5 and sample size is small. The formula of Exact Test is

$$\text{Fisher's Exact test} = \frac{R_1! \ \ R_2! \ \ C_1! \ \ C_2!}{n! \ \ a! \ \ b! \ \ c! \ \ d!}, \tag{7.8}$$

where R_1, R_2 are rows totals and C_1, C_2 are columns totals. Note that Fisher's exact test for 2x2 contingency table does not use the chi-square approximation.

IBM-SPSS package has been used for the above data and computer output is given below. Since expected frequencies are less than 5 in two cells we do not choose Pearson chi-square, we either choose chi-square with Yates' correction (continuity correction) or Fisher's exact-test.

SPSS output for chi-square
(Yates' correction and Fisher's exact test)
suicidal feeling * type of disease Crosstabulation

			type of disease		Total
			1.00	2.00	
suicidal feeling	1.00	Count	2	6	8
		Expected Count	4.0	4.0	8.0
		% within suicidal feeling	25.0%	75.0%	100.0%
		% within type of disease	10.0%	30.0%	20.0%
		% of Total	5.0%	15.0%	20.0%
	2.00	Count	18	14	32
		Expected Count	16.0	16.0	32.0
		% within suicidal feeling	56.3%	43.8%	100.0%
		% within type of disease	90.0%	70.0%	80.0%
		% of Total	45.0%	35.0%	80.0%
Total		Count	20	20	40
		Expected Count	20.0	20.0	40.0
		% within suicidal feeling	50.0%	50.0%	100.0%
		% within type of disease	100.0%	100.0%	100.0%
		% of Total	50.0%	50.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.500[b]	1	.114		
Continuity Correction[a]	1.406	1	.236		
Likelihood Ratio	2.594	1	.107		
Fisher's Exact Test				.235	.118
Linear-by-Linear Association	2.437	1	.118		
N of Valid Cases	40				

a. Computed only for a 2x2 table

b. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 4.00.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-.250	.114
	Cramer's V	.250	.114
	Contingency Coefficient	.243	.114
N of Valid Cases		40	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Since the p-value for Yates' correction (continuity correction) is 0.236(two tailed) and for Fisher's exact test is 0.235 (two tailed) which is greater than 0.05 therefore the data give no evidence that psychotics and neurotics differ with respect to symptoms (we confirm our above findings). Note that the p-value found in SPSS output by the use of Pearson Chi-square test, the Yates' correction and Fisher's test reflect a number of general points about the three tests when applied to small and moderate sized samples.

(i)   Yates' correction and Fisher's test give similar results.

(ii) p-value obtained by Yates' correction and Fisher's test are higher than those given by Pearson's chi-square method.

(iii) In large samples, it is well known that all three methods give almost identical results.

Fisher's exact test is also available in statistical packages for 3×3, 4×4 etc. contingency tables.

Example 7.4:
An interaction study of two social groups of children was conducted. Two independent random samples of 15 children each were selected with and without development delays (mild mental retardation). After observing in a control playground environment, the children during *free play* the researcher recorded the number of children for each group who exhibited disruptive behavior (i.e. ignoring, rejecting other children, taking toys from another child). The data are summarized in the two-way table. Analyze the data given in Table 7.7 and interpret the results.

Table 7.8
Behavior

	Disruptive Behavior	Non-disruptive Behavior	Total
With development delay	12	3	15
Without development delay	5	10	15
Total	17	13	30

(Doop, Baker and Brown, American Journal on Mental Retardation, Vol. 96(4), 1992.

Solution:

(1) H_0 :   There is no difference between with development delay and without development disruptive behavior.

    H_1 :   There is difference.

(2) $\alpha = 0.05$

(3) Test-statistic: Chi-square

   After the calculations of expected frequencies, we will decide whether we apply Pearson chi-square or adjusted chi-square (Yates' correction). The expected frequencies as:

	Disruptive Behavior	non-disruptive behavior	total
with development delay	8.5	6.5	15
without development delay	8.5	6.5	15
Total	17	13	30

   Since no expected cell is less than 5, we use the method of Pearson chi-square (7.2)

$$\chi^2 = \frac{(12 \times 10 - 3 \times 5)^2}{15 \times 15 \times 17 \times 13} = 6.65$$

(4) Table value for 5% level of significance against one degree of freedom is 3.841, which is less than calculated value. The result is significant and we say that there is difference between with development delay and without development delay in disruptive behavior.

7.4.7 R x C contingency table

It is a generalization of the 2x2 contingency table. The case, where there are r rows and c columns, called the r x c contingency table. Suppose we have r populations and one random sample from each population is drawn. Each observation in each sample is classified into one of $r \times c$ different categories. The assumptions are:

(i)   Each sample is random.
(ii)   The outcomes of various samples are all mutually independent.
(iii)   Each observation may be categorized into exactly one of the categories or classes.

Example 7.5:

The researchers randomly divided 120 laboratory rats into four groups of 30 each. All rats were injected with a drug that causes breast cancer, then each rat was fed a diet of fat and fiber for 15 weeks. However, the levels of fat and fiber varied from group to group. At the end of the feeding period, the number of rats with cancer tumor was determined for each group. The data are given in Table 7.9.

Table 7.9
Diet

		high fat with no fiber	High fat with fiber	low fat with no fiber	low fat with fiber	total
Cancer	Yes	27 (22.5%)	20 (16.7%)	19 (15.8%)	14 (11.7%)	80
Tumor	No	3	10	11	16	40
	Total	30	30	30	30	120

Is there any evidence to indicate that diet and presence/absence of cancer are independent? Use 5% level of significance.(source: *Journal of the National Cancer Institute, 1991*)

Solution:

(1) H_0 : Diet and presence/absence of cancer are independent.

   H_1 : They are not independent.

(2) $\alpha = 0.05$

(3) test-statistic: χ^2

SPSS output for Chi-square

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	12.900[a]	3	.005
Likelihood Ratio	14.183	3	.003
Linear-by-Linear Association	11.900	1	.001
N of Valid Cases	120		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.00.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.328			.005
	Cramer's V	.328			.005
	Contingency Coefficient	.312			.005
Ordinal by Ordinal	Kendall's tau-b	.289	.074	3.810	.000
	Kendall's tau-c	.333	.087	3.810	.000
N of Valid Cases		120			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

(4) Calculated p-value is less than the observed p-value (0.05). The result is significant therefore, the null hypothesis is not accepted. We can say with 95% confidence that diet and presence/absence of cancer are not independent and there is about 33% association between these two factors (Cramer's V). We can see from the table that direction of departure from low fat fiber to high fat fiber leading to less cases of cancer.

Example 7.6:

In a study of the relation between blood type and disease, large samples of patients with Peptic ulcer, patients with gastric cancer and control group were classified as to blood type O, A and B. The data are given in Example 7.1 and represent in Table 7.10.

Table 7.10
Disease

Blood type	Peptic ulcer	Gastric cancer	Controls	Total
O	983	383	2892	4258
A	679	416	2625	3720
B	134	84	570	788
Total	1796	883	6087	8766

(Source: Snedecor and Cochran, 1980)

Test the hypothesis that the blood type is the same for the three samples.

Solution:

(1) H_0 = All the blood types are same

H_1 = blood type are not same

(2) $\alpha = 0.05$

(3) Test-statistic χ^2

By using SPSS package, we get $\chi^2 = 40.54339$, p-value = 0.0000

(4) Since p-value is less than 0.05, therefore, the result is significant. The hypothesis is not accepted and we can say that the blood types are not dependent.

If we look into the data carefully and convert into percentage as

Blood type	Peptic ulcer (%)	Gastric cancer (%)	Controls (%)
O	983 (54.7)	383 (43.4)	2892 (47.5)
A	679 (37.8)	416 (47.1)	2625 (43.1)
B	134 (7.5)	84 (9.5)	570 (9.4)

We see that there is not much difference between blood type distributions for gastric cancer patients and controls but peptic ulcer patients differ from both in blood type O. We go back to the data and see if there is any difference between the blood type in gastric cancer patients and control.

Blood type	Gastric cancer	Controls	Total
O	383	2892	3275
A	416	2625	3041
B	84	570	654
Total	883	6087	6970

The calculated value of χ^2 = 5.6361 with p-value = 0.05972. Therefore, there is no difference in blood types between gastric cancer patients and controls.

Further we combine the gastric cancer and controls and omit blood type O and try to test whether the distribution of blood type A and B is the same or different. By doing so, we get the table as:

Blood type	Peptic ulcer	Gastric + Controls
A	679	3041
B	134	654

The calculated value of χ^2 = 0.68471, p-value = 0.408. The result is insignificant and there is no difference between blood type A and B.

We further test whether the proportion of O type versus A + B type in the sample is the same. We get

Blood type	Peptic ulcer	Gastric + Controls	Total
O	983	3275	4258
A + B	813	3695	4508
Total	1796	6970	8766

The calculated value of χ^2 = 34.298 with p-value = 0.000. The result is significant, therefore, we conclude that low p-value or high value of χ^2 is due primarily to an excess of O type blood among the peptic ulcer.

7.4.8 Application of Kendall's Tau b (τ_b)

It takes into considerations the ties and is based on the number of concordant and discordant pairs. An example is presented where the application of Kendall's Tau b coefficient is fruitful. *The solution of the following example will be given using IBM-SPSS:*

Example S7-2

An animal epidemiologist tested dairy cows for the presence of a bacterial disease. The disease is detected by the analysis of blood samples, and the disease severity for each animal was classified as None (0), Low (1) and High (2). Moreover, the size of the herd that each cow belongs to a category is classified as Large (1), Medium (2) and Small (3). The number of animals in each of the 9 cells are recorded as:

Table 7.11
Disease severity

Size of the herd	None (0)	Low (1)	High (2)	Total
Large (1)	11	88	136	235
Medium (2)	18	4	19	41
Small (3)	9	5	9	23
Total	38	97	164	299

The disease is transmitted from cow to cow by bacteria, so the epidemiologist wants to know if disease severity depends on herd size.

Does disease severity increase as herd size increases?

Solution:
Since the categories for *herd size* and for disease severity are *ordered*, therefore, both characteristics are *ordinal*.

The χ^2–statistic tests the independence of herd size and disease severity, but *the test does not show whether there is a trend in disease severity related to increasing herd size and as such Kendall's Tau-b can be used.*

The Variable View is as follows:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	8	0		None	None	2	Right	Unknown	Input
2	Size	Numeric	8	0	Size of the herd	{1, Large}...	None	5	Right	Ordinal	Input
3	Disease	Numeric	8	0	Disease severity	{0, None}...	None	6	Right	Nominal	Input

The labels are defined as:

Size of the herd   Disease severity

We apply the Chi-square test for Independence and calculate *Kendall's Tau-b* as follows:

Analyze→ Descriptive Statistics→ Crosstabs …

Now click on OK , to get the following outputs:

SPSS output for Chi-square and related indices

Size of the herd * Disease severity Crosstabulation

			Disease severity			Total
			None	Low	High	
Size of the herd	Large	Count	11	88	136	235
		Expected Count	29.9	76.2	128.9	235.0
		% within Size of the herd	4.7%	37.4%	57.9%	100.0%
		% within Disease severity	28.9%	90.7%	82.9%	78.6%
	Medium	Count	18	4	19	41
		Expected Count	5.2	13.3	22.5	41.0
		% within Size of the herd	43.9%	9.8%	46.3%	100.0%
		% within Disease severity	47.4%	4.1%	11.6%	13.7%
	Small	Count	9	5	9	23
		Expected Count	2.9	7.5	12.6	23.0
		% within Size of the herd	39.1%	21.7%	39.1%	100.0%
		% within Disease severity	23.7%	5.2%	5.5%	7.7%
Total		Count	38	97	164	299
		Expected Count	38.0	97.0	164.0	299.0
		% within Size of the herd	12.7%	32.4%	54.8%	100.0%
		% within Disease severity	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asy mp. Sig. (2-sided)
Pearson Chi-Square	67.041[a]	4	.000
Likelihood Ratio	56.642	4	.000
Linear-by-Linear Association	23.636	1	.000
N of Valid Cases	299		

a. 1 cells (11.1%) have expected count less than 5. The minimum expected count is 2.92.

Symmetric Measures

		Value	Asy mp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	-.217	.061	-3.402	.001
	Kendall's tau-c	-.148	.044	-3.402	.001
	Spearman Correlation	-.233	.066	-4.131	.000[c]
Interv al by Interv al	Pearson's R	-.282	.066	-5.058	.000[c]
N of Valid Cases		299			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.
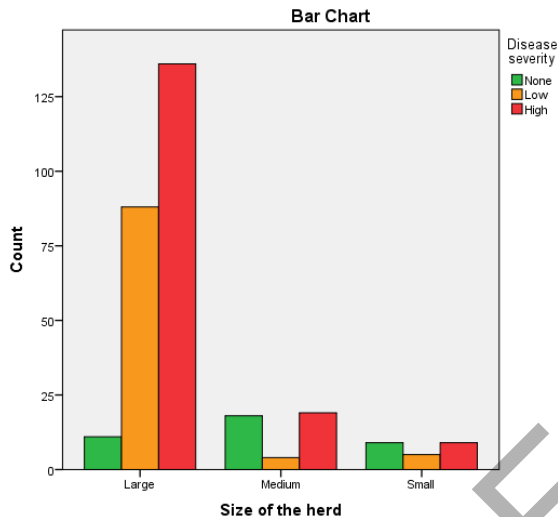
The value of the Kendall's Tau-b is -0.217, which is a measure of association between disease severity and herd size. A negative value means that as one variable decreases, the other increases. In this example, -0.217 means that the disease severity increases as the herd size decreases. This is exactly what one may conclude looking at the observed cell and expected cell frequencies.

One can get incorrect result if the order of the values does not match an increasing or decreasing trend. One can associate large with 1, medium with 2 and small with 3. However, one could receive incorrect result if the order of the values do not match an increasing or decreasing trend. For example, if we associate large as 2, medium as 1 and small with 3, Kendall's Tau-b is meaningless. In general, one needs to look at the values of the variables (both character and numeric) when using Kendall's Tau-b, and make sure that the "order" of values is one that makes sense. The approximate 95% confidence limits are:

$$- 0.21731 \pm 1.96 (0.06065) \text{ or } - 0.21731 \pm 0.11887 \text{ [-0.984 ~ -0.336]}$$

Since the confidence limits do not include zero (0), one can be fairly sure that the association between disease severity and herd size is an increasing one.

Example 7.7:
A simple random sampling procedure was used to select 5 primary health care (PHC) centers out of 9 from Al-Khobar area. Within each selected PHC center, a systematic sampling scheme was applied and 659 patients were selected to determine the pattern of laboratory (Lab) utilization. The data of lab utilization (proper and improper) are as:

Table 7.12
Primary Health Care Centers

Utilization	1	2	3	4	5	Total
Proper	48	51	44	103	77	323
Improper	67	51	37	96	85	336
Total	115	102	81	199	162	659

Moreover, these data are further divided as over utilization, proper utilization and under-utilization. These data are given as:

Table 7.13
Primary Health Care Centers

Utilization	1	2	3	4	5	Total
Over	18	4	15	21	29	87
Proper	48	51	44	103	77	323
Under	49	47	22	75	56	249
Total	115	102	81	199	162	659

Use a statistical technique to analyze the data and to see the difference, if any, between primary health care centers regarding lab utilization.

Solution:
In the first table, we will apply chi-square test as the rows are divided into two categories "yes" and "no". In the second table, the rows are ordinal and columns are nominal. There should be no longer any hesitation in applying the rank test to situations that have many ties. The alternative and frequently used method is Kruskal-Wallis. (This will be described in Chapter 8). In fact the Kruskal-Wallis-H test is excellent test to use in contingency tables, where rows represent ordered scale and columns represent nominal scale.

7.4.9 2 x 2 x K Tables (Meta Analysis)

Sometimes it is possible that a number of 2x2 tables, all bearing on the same question may be available. It becomes of interest how to combine all the tables so that meaningful results may be derived. For example, in an investigation into occurrence of lung cancer among smokers and non-smokers, data may be obtained from several different areas and for each area the data might be arranged in 2x2 table. Again, in an investigation of the occurrence of lung cancer in smokers and non-smokers in different parts of China, data may be obtained from each of several different areas. The question is how this separate information may be pooled? This is explained in the following example. Firstly it is solved manual process and then by using SPSS Package.

Example 7.8:
The following data relating to Chinese smoking and lung cancer study in different parts of China (S = smoker; $\overline{S}$ = non-smoker). Analyze the data to find out whether there is any association between smoking and lung cancer.

Table 7.14

City	Smoking Status	Lung Cancer		Total	Proportion of lung cancer	χ^2	p-value	Phi	χ
		Yes	No						
Beijing	S	126	100	226	0.558	10.033	0.002	0.177	3.17
	$\bar{s}$	35	61	96	0.365				
		161	161	322					
Shanghi	S	908	688	1596	0.569	101.326	0.000	0.187	10.07
	$\bar{s}$	497	807	1304	0.381				
		1405	1495	2900					
Shenyang	S	913	747	1660	0.550	86.660	0.000	0.183	9.31
	$\bar{s}$	336	598	934	0.360				
		1249	1345	2594					
Nanjing	S	235	172	407	0.577	31.925	0.000	0.233	5.63
	$\bar{s}$	58	121	179	0.324				
		293	293	586					
Harbin	S	402	308	710	0.566	38.743	0.000	0.192	6.22
	$\bar{s}$	121	215	336	0.360				
		523	523	1046					
Zhebzou	S	182	156	338	0.538	5.976	0.014	0.108	2.44
	$\bar{s}$	72	98	170	0.423				
		254	254	508					
Taiyuan	S	60	99	159	0.377	5.470	0.018	0.160	2.34
	$\bar{s}$	11	43	54	0.204				
		71	142	213					
Nanchang	S	104	89	193	0.539	5.113	0.023	0.143	2.26
	$\bar{s}$	21	36	57	0.368				
		125	125	250		285.246			41.46

Source: Liu, Z (1992) smoking and lung cancer in China. Inter. J. Epidemiology, Vol. 21, 197-201

Solution:
There are several methods to pool the data.

(i) Pooling the data into 2x2 table

One way is to pool the data in a single table and usual chi-square is calculated. This procedure is applicable or legitimate if *the corresponding proportions in the various tables are alike.* If the proportions vary from table to table, or we suspect that they vary, then this procedure should not be used, *as the combined data will not accurately reflect the information contained in the original tables.* In fact in some cases it so happens that combining several tables each having the two attributes are highly associated and results in a table shows no association. For example, in the lung cancer and smoking study conducted in China at eight places, it may well be the case that the occurrence of lung cancer is more frequent in some areas than the other. If we combine the data into 2x2 table, we get

	C	$\bar{c}$
s	2930	2359
$\bar{s}$	1151	1979

$$\chi^2 = 273.091 \qquad Phi = 0.180 \qquad p < 0.0001$$

The occurrence of lung cancer is associated with localities of China. Since there is variation in proportion of lung cancer, we may not combine all the groups in a single 2x2 table.

(ii) Adding the value of χ^2

The second technique that is often used is to compute the usual chi-square value separately for each table and then add them together. The resulting value may then be compared with the value of chi-square from tables with k degrees of freedom, *where k is the number of separate tables. This is not a good method since it does not take into account the direction of the difference between the proportions in various tables and consequently lacks power in detecting a difference that show up consistently in the same direction in all or most of the individuals.* If we use this method, we get $\chi^2_{pooled} = 285.246$. The table value for 8 degrees of freedom (since there are 8 tables) is 15.507. Since table value is much less than calculated value, therefore the result is significant and we can say that there is association between smoking and lung cancer.

(iii) The method of summing χ rather than χ^2

If the sample sizes of the individual tables do not differ greatly (say by more than a ratio of 2 to 1) and the values taken by the proportions are between approximately 0.2 and 0.8, then a method based on the sum of the square root of the χ^2 statistic, taking account of the signs of the differences in proportions, may be used. This will be normally distributed with mean zero and standard deviation $\sqrt{K}$ if the sample is large. Then

$$Z = \sum_{i=1}^{K} \frac{\chi_i}{\sqrt{k}} = \frac{41.46}{\sqrt{8}} = 14.66$$

Using Table 7.12 we have Z = 1.96. Since calculated Z is much more than 1.96 (table value at 5% level of significance), therefore result is significant and there is a strong association between smoking and lung cancer.

If the sample sizes and the proportions do not satisfy the conditions mentioned above, the addition of the χ value tends to lose power. Tables that arise from very small sample size cannot be expected to be as much of use as those where the sample size is moderate to large in detecting the difference in the proportions, yet in the $\sqrt{\chi^2}$ method all tables receive the same weight. *When differences in the sample sizes are extreme, some method of weighting the results from individual tables is needed.* Cochran (1954) suggested a test to solve this problem. Another test procedure for examining series of 2x2 tables is that suggested by Mantel and Haenszel (1959). By combining this test, it is known as Cochran -Mantel -Haenszel test.

(iv) The Cochran- Mantel-Haenszel test

To apply Cochran-Mantel-Haenszel test, some further calculations are required. These calculations are made in the Table 7.15.

Table 7.15

City	Smo-king Status	Lung Cancer Yes	Lung Cancer No	Total	$\dfrac{(a+b)(a+c)}{n} = E(a)$	$\dfrac{(a+b)(c+d)(a+c)(b+d)}{n^2(n-1)}$	a − E(a)
1	S	126	100	226			
	$\bar{S}$	35	61	96	113.0	16.9	13.0
		161	161	322			
2	S	908	688	1596			
	$\bar{S}$	497	807	1304	773.2	179.3	134.8
		1405	1495	2900			
3	S	913	747	1660			
	$\bar{S}$	336	598	934	799.3	149.3	113.7
		1249	1345	2594			
4	S	235	172	407			
	$\bar{S}$	58	121	179	203.5	31.1	31.5
		293	293	586			
5	S	402	308	710			
	$\bar{S}$	121	215	336	355.0	57.1	47.0
		523	523	1046			
6	S	182	156	338			
	$\bar{S}$	72	98	170	169.0	28.3	13.0
		254	254	508			
7	S	60	99	159			
	$\bar{S}$	11	43	54	53.0	9.0	7.0
		71	142	213			
8	S	104	89	193			
	$\bar{S}$	21	36	57	96.5	11.0	7.5
		125	125	250		482.0	367.5

Cochran-Mantel-Haenszel (CMH) test = $\dfrac{(367.5)^2}{482.0}$ = 280.2. This test has a large sample chi-squared distribution with 1 d.f. We can see that the result is highly significant. A statistical analysis that combines information from several studies is called *meta analysis*. This meta analysis may provide stronger evidence of an association than any single partial table.

Calculation of Cochran-Mantel-Haenszel (CMH) test to perform Meta Analysis using IBM-SPSS package.

Example S7-3

In example 7.8 there are eight study areas regarding smoking and lung cancer. In order to apply Cochran-Mantel-Haenszel (CMH) technique these informations will be entered in the following way.

- Enter the data in the following manner.

	City	Row	Column	Data		City	Row	Column	Data
1	1	1	1	126	1	Beijing	Smokers	Yes	126
2	1	1	2	100	2	Beijing	Smokers	No	100
3	1	2	1	35	3	Beijing	Non Smokers	Yes	35
4	1	2	2	61	4	Beijing	Non Smokers	No	61
5	2	1	1	908	5	Songhai	Smokers	Yes	908
6	2	1	2	688	6	Songhai	Smokers	No	688
7	2	2	1	497	7	Songhai	Non Smokers	Yes	497
8	2	2	2	807	8	Songhai	Non Smokers	No	807
9	3	1	1	913	9	Shenyang	Smokers	Yes	913
10	3	1	2	747	10	Shenyang	Smokers	No	747

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
City	Numeric	8	0		{1, Beijing}...	None	8	Right	Nominal	Input
Row	Numeric	8	0	Smoking Status	{1, Smokers...	None	10	Right	Nominal	Input
Column	Numeric	8	0	Lungs Cancer	{1, Yes}...	None	8	Right	Nominal	Input
Data	Numeric	8	0		None	None	8	Right	Scale	Input

The labels are defined as:
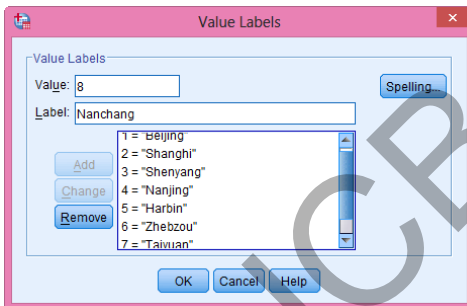
City   Smoking status
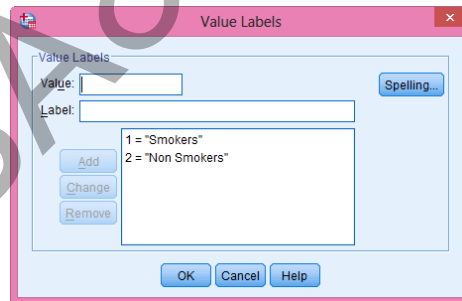
Lungs Cancer
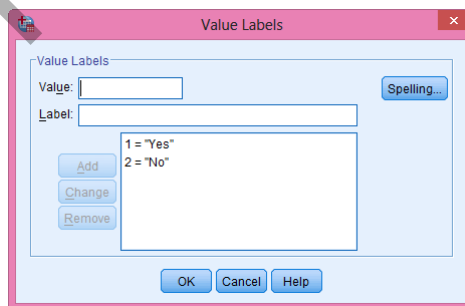
To proceed for analysis

1. Click *Data* and then click *Weight Cases* (Weight the cases by the variable data)

Move the variable "row" to the Row(s):

Move the variable "column" to the Column(s):
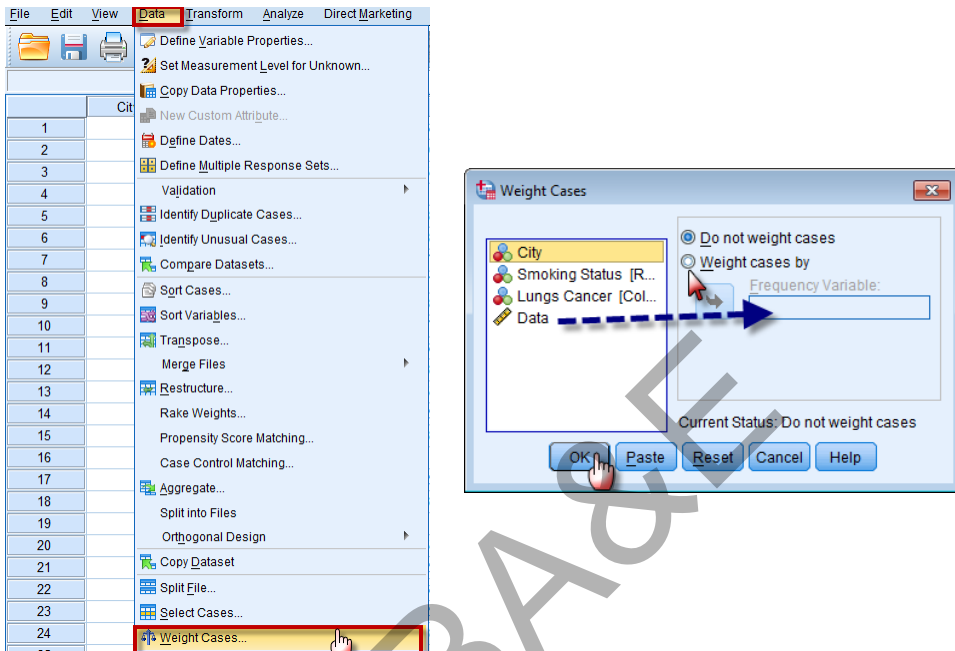
2. Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab*.

Analyze→ Descriptive Statistics→ Crosstabs …
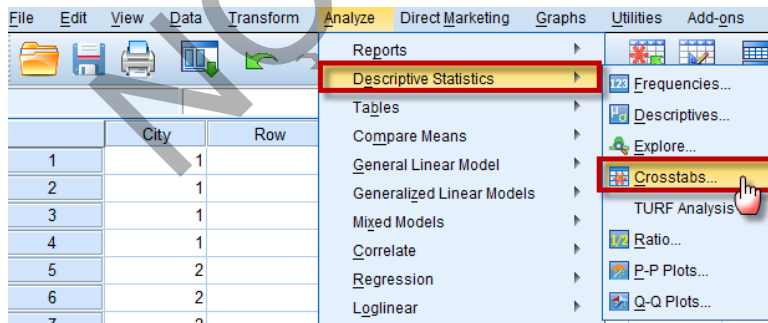
Move the variable "row" to the Row(s):

Move the variable "column" to the Column(s):

We click on Statistics... and mark on "Chi-square",

We also mark on "*Cochran-Mantel-Haenszel*" then click on Continue
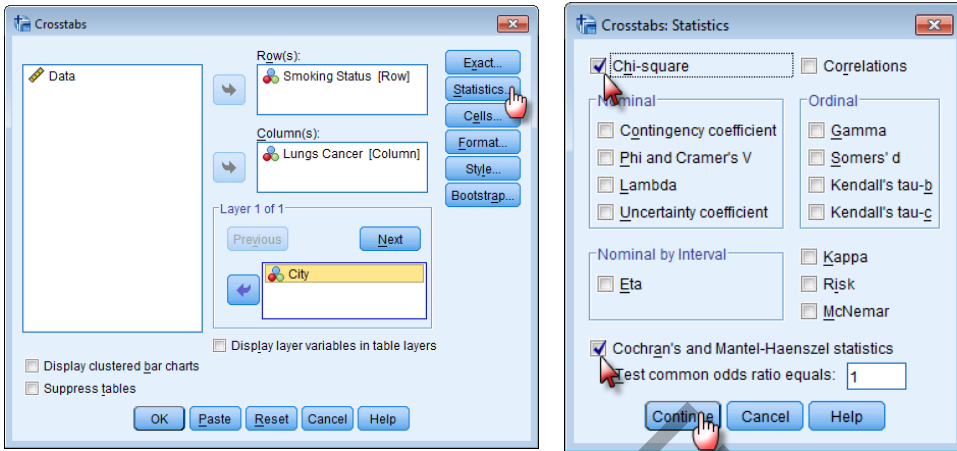
Now click on Continue and on OK, to get the following outputs:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Smoking Status * Lungs Cancer * CITY	8419	100.0%	0	.0%	8419	100.0%

Smoking Status * Lungs Cancer * CITY Crosstabulation

Count

CITY			Lungs Cancer		Total
			Yes	No	
Beijing	Smoking Status	Smokers	126	100	226
		Non-Smokers	35	61	96
	Total		161	161	322
Shanghi	Smoking Status	Smokers	908	688	1596
		Non-Smokers	497	807	1304
	Total		1405	1495	2900
Shenyang	Smoking Status	Smokers	913	747	1660
		Non-Smokers	336	598	934
	Total		1249	1345	2594
Nanjing	Smoking Status	Smokers	235	172	407
		Non-Smokers	58	121	179
	Total		293	293	586
Harbin	Smoking Status	Smokers	402	308	710
		Non-Smokers	121	215	336
	Total		523	523	1046
Zhebzou	Smoking Status	Smokers	182	156	338
		Non-Smokers	72	98	170
	Total		254	254	508
Taiyuan	Smoking Status	Smokers	60	99	159
		Non-Smokers	11	43	54
	Total		71	142	213
Nanchang	Smoking Status	Smokers	104	89	193
		Non-Smokers	21	36	57
	Total		125	125	250

The Chi-Square along with significance level (p-value) for each table is given as

CITY	Chi-Square Tests	df	Asymp. Sig. (2-sided)
Beijing	10.033	1	0.002
Shanghi	101.327	1	0.000
Shenyang	86.661	1	0.000
Nanjing	31.925	1	0.000
Harbin	38.743	1	0.000
Zhebzou	5.976	1	0.014
Taiyuan	5.470	1	0.019
Nanchang	5.113	1	0.024

Tests for Homogeneity of the Odds

Statistic		Chi-Squared	df	Asymp. Sig. (2-sided)
Conditional-	Cochran's-	280.38	1	.000
Independence	Mantel-Haenszel	279.37	1	.000
Homogenei	Breslow-Day-	5.200	7	.636
	Tarone'	5.200	7	.636

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

There is a small difference in Chi- Square value in manual calculation and in computer application as approximation is involved in manual process.

Example 7.9:
Data regarding incidence of tumors in the two hemispheres for three sites in the cortex is available as:

Table 7.16

Sr. No.	Site of tumor	Benign tumors	Malignant tumors	Total	Proportion of malignant tumors
1	Left hemisphere	17	5	22	0.2273
	Right hemisphere	6	5	11	0.4545
		23	10	33	
2	Left hemisphere	12	3	15	0.2000
	Right hemisphere	7	5	12	0.4167
		19	8	27	
3	Left hemisphere	11	3	14	0.2143
	Right hemisphere	11	9	20	0.4500
		22	12	34	

Can we say that there is association between type of tumor and among hemisphere?

Solution:
We left this problem to the students to solve by using IBM-SPSS package on the lines suggested in Example 7.9.

7.5 Matched Samples (McNemar test)

One to one matching is frequently used by research workers to increase the precision of the comparison. This point has also been discussed in Chapter 4 in details as well. The matching is usually done on variable such as age, sex, weight, etc. and like information about which data can be obtained easily. *Two samples matched in a one-to-one way must*

be thought of correlated samples and consequently are not independent. As a result, the ordinary chi-square test is not strictly applicable for assessing the difference between frequencies obtained with reference to these samples. The appropriate test for comparing frequencies in matched samples is one due to McNemar (1955). This is a special case of Cochran-Mantel-Haenszel test.

Suppose the data are nominal with two categories that we call 1 and 0, i.e. $X_i = 1$ or 0 and $Y_i = 1, 0$, i.e.

Table 7.17
Sample 1

	X_i	Y_i		Total
		Yes = 1	No = 0	
	Yes = 1	(1, 1) a	(1, 0) b	a + b
Sample 2	No = 0	c	d	c + d
		(0, 1)	(0, 0)	
	Total	a + c	b + d	a + b + c + d

Since we are concerned with the difference between sample 1 and sample 2. There is no difference in the cells of the table corresponding to cell a and cell d therefore, the comparison is confined to cells b and *c* only. In these situations, our null hypothesis will be that the two samples do not differ as regards to the attribute. We would expect cell b and cell c to be equal. We expect that the values in these two cells would each be $(b + c)/2$. Then the null and alternative hypotheses are

(1) H_0 : Two samples do not differ with regards to the attributes

   H_1 : These are not equal.

(2) $\alpha = 0.05$

(3) test-statistic: $\chi^2_{McNemar}$

$$\chi^2_{McNemar} = \frac{(b-c)^2}{b+c} \tag{7.9}$$

If the frequency in the cell b or c or in both is less than 5 then corrected value of McNemar test will be calculated as:

$$\chi^2_{McNemar(c)} = \frac{\left[|b-c|-1\right]^2}{b+c} \tag{7.10}$$

7.5.1 Layout of Tests of Significance

The following layout will be useful to understand the applications of chi-squares and McNemar's tests.

LAYOUT OF TEST OF SIGNIFICANCE
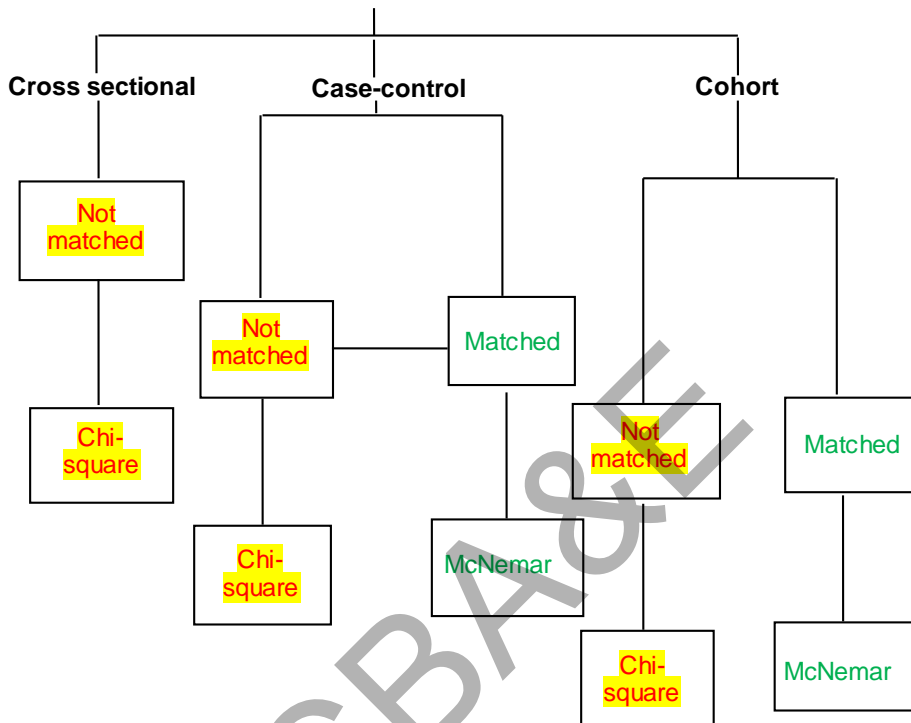
McNemar test is applicable in case-control and cohort matched samples.

Example 7.10:

Following data relate to 400 study subjects, consisting of 200 matched-pairs. For 7 pairs both the smokers and non-smokers developed myocardial infarction (MI) and for 150 pairs, neither did. In 14 pairs only the non-smoker have myocardial infarction whereas in 29 pairs only the smoker did. This data relate to the results of a cohort study of myocardial infarction in 200 smoking and 200 non-smoking men matched by age, blood pressure and serum cholesterol concentration. Cells a and d represent those matched pairs in which both exposed and non-exposed members develop the same outcome whereas b and c represent those matched pairs in which members experience opposite results. The data are given as:

Table 7.18
Smokers

Non-Smokers	MI	not MI	Total
MI	7 (1, 1) a	14 (1,0) b	21
Not MI	c (0,1) 29	d (0, 0) 150	179
Total	36	164	200

Test the significance between smoking and myocardial infarction at 5% level of significance.

Solution:

(1) H_0 : Smoking has no effect on myocardial infarction.

   H_1 : Smoking and myocardial infraction are associated.

(2) $\alpha = 0.05$

(3) test-statistic: Since the pairs are matched, the value of chi-square depends on the observed frequencies in the two discordant cells b and c. It is interpreted in the same way as the usual χ^2 with 1 d.f. McNemar chi-square procedure is used below (using equation 7.1):

$$\chi^2_{\text{McNemar}} = \frac{[29-14]^2}{29+14} = 5.233$$

(4) Table value of χ^2 for 5% level of significance and for 1 degree of freedom is 3.841

(5) Since the calculated value is more than the table value, we do not accept the null hypothesis and say that smokers are indeed at risk for subsequent myocardial infarction.

McNemar test is also applicable to situations in which the same subjects are observed on two occasions.

Example S7-4

In Re-solving example 7.10 using IBM-SPSS,

• Enter the data in the following manner.

	Non_smokers	Smokers
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	2
9	1	2
10	1	2

	Non_smokers	Smokers
1	MI	MI
2	MI	MI
3	MI	MI
4	MI	MI
5	MI	MI
6	MI	MI
7	MI	MI
8	MI	Not MI
9	MI	Not MI
10	MI	Not MI

(up to row 200)

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Non_smokers	Numeric	8	0	Non smokers	{1, MI}...	None	9	Right	Nominal	Input
Smokers	Numeric	8	0	Smokers	{1, MI}...	None	8	Right	Nominal	Input

The labels are defined as:

Non smokers   Smokers
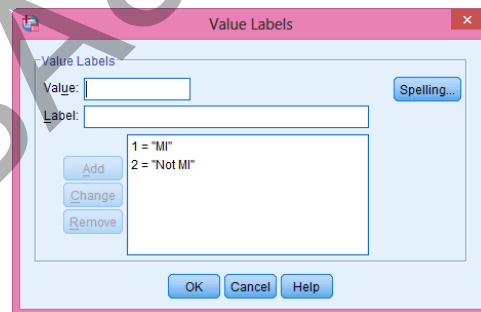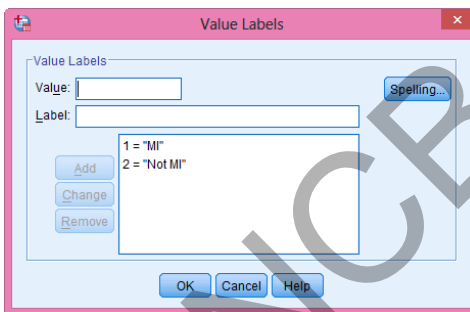
To proceed for analysis

Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab.*

Analyze→ Descriptive Statistics→ Crosstabs …
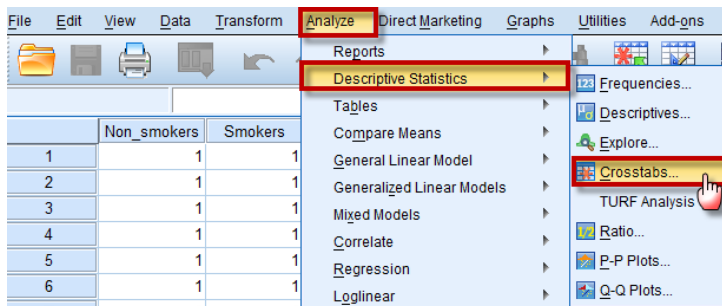
Move the variable "Non-smokers" to the Row(s):

Move the variable "Smokers" to the Column(s):

We click on [Statistics...] and mark on "McNemar",
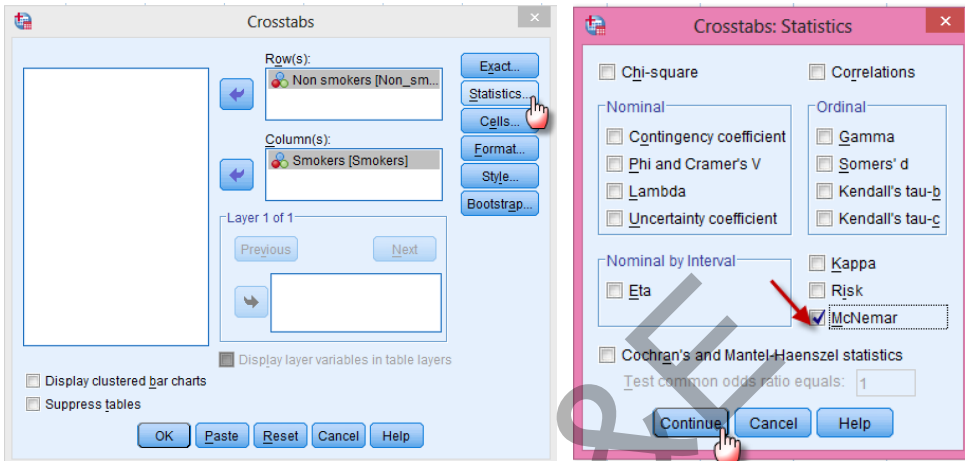
Now click on [Continue] and on [OK], to get the following output:

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		.032[a]
N of Valid Cases	200	

a. Binomial distribution used.

Since the P-value= 0.032 and is less than 0.05, we do not accept the null hypothesis and say that smokers are indeed at risk for subsequent myocardial infarction.

Example 7.11:
Two drugs A and B are used to same patients on two different occasions in the treatment of depression and are compared in terms of possible side-effects, nausea. The drugs are given to the patients on two different occasions and the incidence of nausea recorded in the following table.

Table 7.19

		Drug A		
		Nausea	No-Nausea	Total
Drug B	Nausea	9	3	12
	No-Nausea	13	75	88
	Total	22	78	100

Compare the effect of two drugs.

Solution:
Here we are dealing with correlated rather than independent observations since the same group receives both drugs A and B, the comparison of the drugs will be made by using McNemar's test.

(1) H_0 : Incidence of nausea is same for the two drugs

   H_1 : Incidence of nausea is different for the two drugs

(2) $\alpha = 0.05$

(3) test-statistic: McNemar

Since one of the values in the one cell is less than 5, therefore we will apply (7.10) to calculate test-statistic, i.e.

$$\chi^2_{McNemar} = \frac{(|3-13|-1)^2}{3+13} = 5.06$$

(4) Table value of χ^2 for 5% level of significance and for 1 degree of freedom is 3.841.

(5) Since calculated value is greater than the table value, we do not accept the hypothesis and say that incidence of nausea is different for the two groups of drugs.

(Note that the P-value when using IBM-SPSS will be equal 0.021 which gives the same result for significance)

7.6 Mantel-Haenszel Test for Linear Association

If the exposure variable is ordinal, then the ordinary chi-square test does not take into account the inherent order among the categories. It merely tests the overall departure of observed from expected across the $r \times 2$ cells of the table. A test of linear association between columns and rows will be statistically inefficient, because it fails to distinguish between one-and two-category differences. Following example is given to explain this concept.

Example 7.12:
The following table gives a summary of the results of a cohort study in which children with otitis media (Middle-ear infection) were treated with oral amoxicillian in either the dosage range recommended (RD) by the manufacturer, a dosage above that recommended dose (HD), or a dosage below the recommended dose (LD). The children were followed for the duration of their 10-days course of treatment for the occurrence of diarrhea, a well-known side effect of oral amoxicillian.

Table 7.20
Response

		Diarrhea	**no diarrhea**	**total**
	high dose	12	38	50
Dose	**recommended dose**	13	87	100
	Low dose	4	46	50
	Total	29	171	200

Is the dose response relation significant?

Solution:

Health Scientists will immediately apply Pearson chi-square (ordinary chi-square) to see the association between dose and response. The test is not applicable because of existence of linearity in one of the categories.

Suppose we apply ordinary chi-square.

Response

Dose	**D**	$\bar{D}$	**Total**
HD	12 (7.25)	38 (42.75)	50
RD	13 (14.50)	87 (85.50)	100
LD	4 (7.25)	46 (42.75)	50
Total	29	171	200

where: D   = Diarrhea; $\bar{D}$ = Not diarrhea
           HD   = High dose;   RD   = Recommended dose
           LD   = Lower dose

The $\chi^2 = 5.5253$ with p-value $= 0.06312$, but at 5% level of significance, there is no association between dose - response.

For this type of problem a preferable test is chi-square with linear trend (Mental-Haenszel). The formula for chi-square for linear trend is

$$\chi^2_{MH} = \frac{n[n\Sigma t_i w_i - t\Sigma n_i w_i]^2}{t(n-t)[n\,\Sigma t_i w_i^2 - (\Sigma n_i w_i)^2]} \; , \qquad\qquad (7.11)$$

where: n = sum of all the frequencies

     w_i = weight (score) assigned to ith category

     t_i = number of subjects within the ith category

who experience the target outcome

     n_i = number of subjects in the ith exposure category

     t = total number who experience the outcome

χ^2_{MH} has one degree of freedom. To solve this, the table can be rearranged as

Weight

	D	$\overline{D}$	**Total**
w$_1$ = +1 HD	t$_1$ = 12	38	50 = n$_1$
W$_2$ = 0 RD	t$_2$ = 13	87	100 = n$_2$
W$_3$ = -1 LD	t$_3$ = 4	46	50 = n$_3$
Total	t = 29	171 (n – t)	n = 200

(Note that if there are four categories the weight will be assigned as +3, +1, -1, -3 and so on).

Using the above formula of chi-square for linear trend, we get $\chi^2 = 5.137$ at 1 degree with p-value = 0.023 (two tailed). The result is significant and we can say with 95% confidence that there is dose-response relationship.

This example is solved by using IBM-SPSS package and the steps are as follows:

Example S7-5

- Enter the data in the following manner.

	Dose	Response			Dose	Response
1	1	1		1	High	Diarrhea
2	1	1		2	High	Diarrhea
3	1	1		3	High	Diarrhea
4	1	1		4	High	Diarrhea
5	1	1		5	High	Diarrhea
6	1	1		6	High	Diarrhea
7	1	1		7	High	Diarrhea
8	1	1		8	High	Diarrhea
9	1	1		9	High	Diarrhea
10	1	1		10	High	Diarrhea
11	1	1		11	High	Diarrhea
12	1	1		12	High	Diarrhea
13	1	2		13	High	No diarrhea
14	1	2		14	High	No diarrhea

(up to row 200)

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Dose	Numeric	8	0	Dose	{1, High}...	None	9	Right	Nominal	Input
Response	Numeric	8	0	Response	{1, Diarrhea}...	None	8	Right	Nominal	Input

The labels are defined as:

Dose
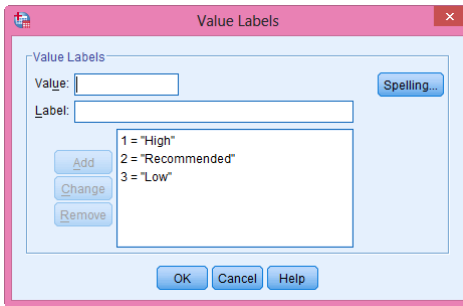
Response
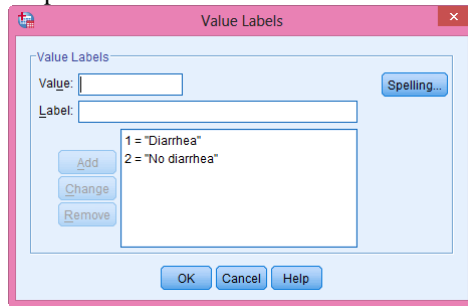
To proceed for analysis

Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab.*

Analyze→ Descriptive Statistics→ Crosstabs …
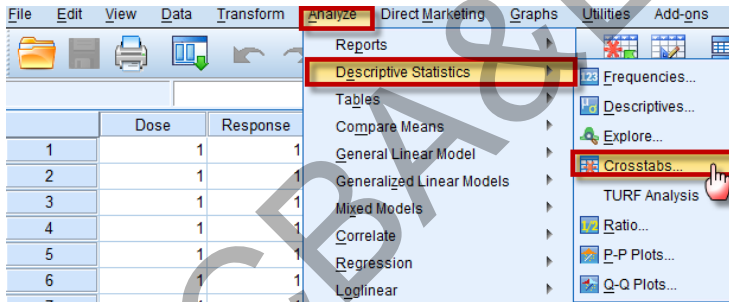
Move the variable "Dose" to the Row(s):

Move the variable "Response" to the Column(s):

We click on Statistics… and mark on "McNemar",
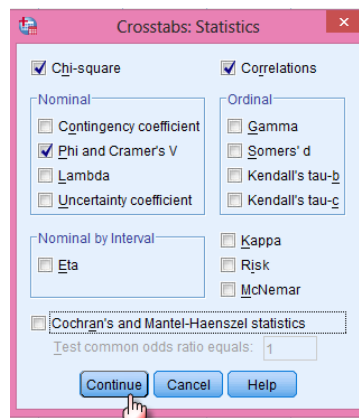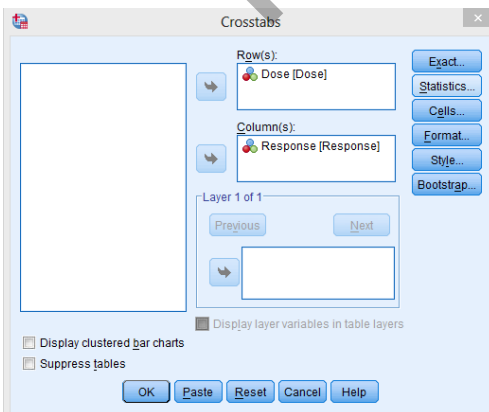
Now click on [Continue] and on [OK], to get the following output:

SPSS output
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5.525[a]	2	.063
Likelihood Ratio	5.313	2	.070
Linear-by-Linear Association	5.137	1	.023
N of Valid Cases	200		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.25.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.166			.063
	Cramer's V	.166			.063
Interval by Interval	Pearson's R	.161	.069	2.290	.023[c]
Ordinal by Ordinal	Spearman Correlation	.161	.069	2.290	.023[c]
N of Valid Cases		200			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

There is a simple way to calculate χ^2_{MH} is as;

$$\chi^2_{MH} = (n - 1) \, r^2, \tag{7.12}$$

where r is the correlation coefficient between two attributes, n is the total number of frequencies. Here n = 200, r = 0.161, thus

$$\chi^2_{MH} = (200 - 1) \, (0.161)^2 = 5.158$$

which gives little different result as in manual calculation some approximations are involved. We can also apply if both variables are linear or on ordinal scale.

Note, that failure to consider the ordinal nature of the exposure variable in the analysis would thus have led to a loss of statistical efficiency. In these types of situations, Mann-Whitney-U-test can be used. This will be discussed in Chapter 8.

7.7 Testing the Statistical Significance of Relative Risk and Odds Ratio

In this section, a great deal of discussion is devoted to definition, estimation and statistical significance of relative risk and odds ratio. The theoretical background of the relative risk and odds ratio are not discussed as this has been given in detail in books on epideomology.

7.7.1 Relative Risk (RR) Estimate

Relative risk is a measure of the association between exposure to a particular factor and risk of a certain outcome. For two dichotomous variables viz. exposure (E) and disease (D), the relative risk (RR) estimate in 2x2 table is defined as

$$RR = \frac{P[D = yes \, / \, E = yes]}{P[D = yes \, / \, E = No]}$$

$$= \frac{\text{risk of disease if exposed to the risk}}{\text{risk of disease if not exposed to the risk}}$$

$$= \frac{\text{incidence of disease in exposed group}}{\text{incidence of disease in non - exposed group}}$$

Consider a 2x2 table

Table 7.21
Disease

		D	$\overline{D}$	Total
	E	a	b	a + b
Exposed	$\overline{E}$	c	d	c + d
	Total	a + c	b + d	a + b + c + d

where: E = exposed; $\overline{E}$ = not exposed
       D = disease; $\overline{D}$ = no disease

and a, b, c, d are frequencies in the relevant cells.

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{p_1}{p_2} \text{ where } p_1 = \frac{a}{a+b} \text{ and } p_2 = \frac{c}{c+d} \qquad (7.13)$$

Note that relative risk is calculated for cohort, longitudinal or experimental studies. Relative risk *does not measure the probability that someone with this factor will develop the disease but it measures the strength or magnitude of exposed-outcome association.* The greater the value of RR the stronger the association between exposure and disease to risk factor. If the value of RR is 1, this indicates that exposure and disease are unrelated. If the value of RR is less than 1, this indicates that there is a negative association between exposure and the disease. If the value of RR is greater than 1, this indicates that there is a positive association between exposure and disease. *In case-control study, the relative risk cannot be calculated directly. Therefore, in case-control study risk can be estimated by the odds ratio. It acts as an approximation to the relative risk*.

7.7.2 Odds ratio

If the two possible states of the variable are labeled *success* and *failure,* then the odds ratio is a measure of the odds of a success in one group relative to that in the other.

The steps in the calculation of odds ratio are given below:

Consider the data in Table 7.22.

Table 7.22

	Case	Control	Total
E	a	b	a + b
$\overline{E}$	c	d	c + d
Total	a + c	b + d	a + b + c + d

(i)   Rate of exposure in cases $= \dfrac{a}{a+c}$ (7.14)

(ii)   Rate of exposure in controls $= \dfrac{b}{b+d}$ (7.15)

(iii)   The odds that an individual exposed to the risk has the disease is

$$\frac{a/(a+b)}{b/(a+b)} = a/b \qquad (7.16)$$

(iv)   The odds that an individual who has not been exposed to the risk factor has the disease.

$$\frac{c/(c+d)}{d/(c+d)} = c/d \qquad (7.17)$$

(v)   Odds of exposure in cases = a/c.

(vi)   Odds of exposure in controls = b/d.

(vii) $OR = \dfrac{a/(a+b)}{b/(a+b)} \,/\, \dfrac{c/(c+d)}{d/(c+d)} = \dfrac{ad}{bc}$ (7.18)

The odds ratio can directly be calculated from the table by using

$$OR = \frac{ad}{bc} \qquad (7.19)$$

Note that relative *risk is a ratio of two probabilities* and the *odds ratio is a ratio of two odds.*

7.7.3 Attributable risk (Risk difference, Rate difference)

It is a measure of association between exposure to a particular factor and the risk of a particular outcome and is calculated as:

    Incidence rate among exposed – Incidence rate among non-exposed

In terms of a 2 x 2 table, it is calculated as:

$$A. R = \frac{a}{a+c} - \frac{c}{c+d} \qquad (7.20)$$

It measures the amount of the incidence that can be attributed to one particular factor.

Before we pass on to the statistical significance of relative risk and odds ratio, the following steps should be kept in mind.

(a) General results

    (i)   If RR or OR is greater than 1, exposure is associated with increased risk of outcome (positive association).

    (ii)   If RR or OR is less than 1, it indicates that exposure protects against the development of the outcome (negative association).

    (iii) If RR or OR is equal to 1, exposure and outcome are independent (no association).

(b) Warning

If any cell has zero frequency, then 0.5 is added to each cell and odds ratio can be calculated.

(c) Test of significance for relative risk and odds ratio

RR or OR may occur greater or less than 1 by chance, if H_0 is true. For this purpose, it is advisable to test the significance as:

(i) Chi-square

    (i)   If RR or OR is greater than 1 and chi-square gives significant result, then exposure is associated significantly with increased risk of the outcome.

    (ii)   If RR or OR is less than 1 and chi-square is significant, there is a protection of exposure against outcome.

    (iii) If RR or OR is less than or greater than 1 and chi-square is non-significant then RR or OR is by chance.

(ii) Confidence limits
The confidence limits of RR and OR are derived by Miettinen (1969). We may construct 95% or 99% confidence limits for RR or OR. If the interval does not include 1, then RR or OR is statistically significant. The result can be interpreted on the basis of the values of the RR and OR.

Example 7.13:
The data regarding cohort study of 200 smokers (cases) and 200 non-smokers (controls) for occurrence of myocardial infarction (MI) are given in Table 7.23.

Table 7.23

	MI	$\overline{\text{MI}}$	Total
Smoker	32	168	200
	a	b	
Nonsmoker	c	d	
	15	185	200
Total	47	353	400

where MI = myocardial infraction and $\overline{\text{MI}}$ = no myocardial infraction

Calculate relative risk of myocardial infarction in smokers.

Solution:

MI in smokers = 32/200 = 0.16 (16%)

MI in non-smokers = 15/200 = 0.075 (7.5%)

$$RR = \frac{32/200}{15/200} = \frac{a/(a+b)}{c/(c+d)} = 2.13$$

This indicates that those who smoke have 2.13 times more chance of myocardial infarction than those who do not smoke.

$$AR = \frac{32}{47} - \frac{168}{353} = 0.681 - 0.476 = 0.205$$

(i) Testing of significance of relative risk

The significance of relative risk may be tested by the method of chi-square. Confidence limits can also be constructed for RR and AR.

Commonly, health scientists use the confidence limits to draw inference. However, it is advisable that method of chi-square be used as this method has a general application and is commonly understandable. Using formula (7.2)

$$\textbf{(a)} \ \chi^2 \ = \ \frac{[32 \times 185 - 168 \times 15]^2 \times 400}{200 \times 200 \times 47 \times 353} \ = \ 6.97$$

The table value for 1 degree of freedom at 5% level of significance is 3.841. Since our calculated value is more than the table value, therefore, the result is significant. Since relative risk is greater than 1 and χ^2 gives significant result, therefore, smoking has positive effect on myocardial infarction.

(ii) Confidence limits

95% confidence limits of RR are

(i) $\left(RR\right)^{1\pm1.96\sqrt{\chi^2}} = \left(2.13\right)^{1\pm1.96\sqrt{6.97}} = \left[1.22, \ 3.73\right]$ (7.21)

$$\text{(ii)} \quad RR\, e^{+1.96\sqrt{\frac{1-\frac{a}{a+b}}{a}+\frac{1-\frac{c}{c+d}}{c}}} = \big[1.19,\ 3.80\big] \qquad\qquad (7.22)$$

These limits do not include 1, so the value of the relative risk is not by chance. This can also be calculated by using IBM-SPSS package. The entry of data is just like, the entry of data for the calculations of χ^2.

Example S7-6

- Enter the data in the following manner.

	Risk_factor	Response
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1

	Risk_factor	Response
1	Smoker	MI
2	Smoker	MI
3	Smoker	MI
4	Smoker	MI
5	Smoker	MI
6	Smoker	MI
7	Smoker	MI
8	Smoker	MI
9	Smoker	MI
10	Smoker	MI

(up to row 400)

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Risk_factor	Numeric	8	0	Smoking	{1, Smoker}...	None	9	Right	Nominal	Input
Response	Numeric	8	0	Myocardial infarction	{1, MI}...	None	8	Right	Nominal	Input

The labels are defined as:

Risk Factor   Response
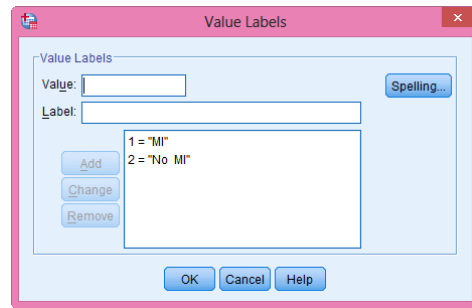
Value Labels	Value Labels
Value: Label: 1 = "Smoker" 2 = "Non Smoker" Add Change Remove OK Cancel Help	Value: Label: 1 = "MI" 2 = "No MI" Add Change Remove OK Cancel Help

To proceed for analysis

Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab.*

Analyze→ Descriptive Statistics→ Crosstabs …

Move the variable "Risk factor (smoking)" to the Row(s):

Move the variable "Response (MI)" to the Column(s):

We click on Statistics... and mark on "Phi and Cramer's V" and on Risk",
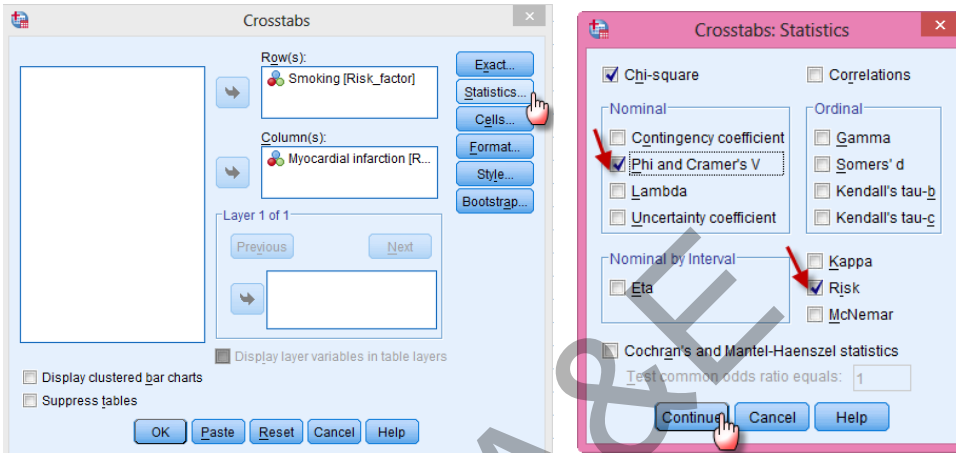
Now click on Continue and on OK, to get the following output:

SPSS output for Relative Risk

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.968[b]	1	.008		
Continuity Correction[a]	6.172	1	.013		
Likelihood Ratio	7.110	1	.008		
Fisher's Exact Test				.012	.006
Linear-by-Linear Association	6.950	1	.008		
N of Valid Cases	400				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 23.50.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.132	.008
	Cramer's V	.132	.008
N of Valid Cases		400	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for SMOKING (1 / 2)	2.349	1.229	4.491
For cohort myocardial = 1	2.133	1.193	3.815
For cohort myocardial = 2	.908	.845	.976
N of Valid Cases	400		

The confidence limits calculated on the basis of equation 7.22 matches with computer output.

Example 7.14:

The director of community health for a certain state observes that women living in rural parts of the state have a high rate of miscarriage than women living in urban areas as they are exposed to pesticides. The director takes 100 cases from the rural parts and 100 from urban areas and both groups are followed up. The results are in Table 7.24.

Table 7.24

	Miscarriage	Not Miscarriage	total
Exposed	30	70	100
Not-exposed	10	90	100
Total	40	160	200

Calculate relative risk for the women who are exposed to pesticide.

Solution:

Miscarriage in exposed group $= \dfrac{30}{100} = 0.3$ (30%)

Miscarriage in not exposed group $= 10/100 = 0.1$ (10%)

Relative risk in exposed group $= \dfrac{30/100}{10/100} = 3$

Those women who are exposed to pesticide, have 3 times more chance of miscarriage than those women who are not exposed to the pesticide.

The significance of relative risk may be tested by using formula Chi-Square, (7.2)

(a) $\chi^2 = \dfrac{[30\times90 - 70\times10]^2 \times 200}{100\times100\times40\times160} = 12.5$

The 5% table value of chi-square with 1 degree of freedom is 3.841. The calculated value is much greater than table value, therefore, the incidence of miscarriage in women exposed to the pesticide differs significantly. Since relative risk is 3 and the value of chi-square gives significant result, therefore, exposure to pesticides has three times more chances of miscarriage.

The confidence limits for relative risks may be used to test the significance.

(i) $3^{1\pm1.96/\sqrt{12.5}} = [1.63 , 5.52]$

(ii) $3e^{\pm1.96\sqrt{\dfrac{1-\frac{30}{100}}{30}+\dfrac{1-\frac{10}{100}}{10}}} = [1.55 , 5.80]$

Both sets of confidence limits do not include 1, so the value of relative risk is not by chance. IBM-SPSS Package may be used for calculations

7.7.4 Relative risk of matched-pairs

Paired matching is often used in observational studies to reduce confounding. If pair matching is used in the design, the statistical analysis will be more efficient (have greater power of the test).

When both exposure and outcome are dichotomous and the matching is by pairs, the result can be expressed as in Table 7.25.

Table 7.25

Non Smokers	Smokers		
	MI	$\overline{\text{MI}}$	
MI	A	b	a + b
$\overline{\text{MI}}$	c	d	c + d
	a + c	b + d	a+b+c+d

where MI = myocardial infraction and $\overline{\text{MI}}$ = no myocardial infraction

Cells a and d represent those matched pairs in which both the exposed and non-exposed members develop the same outcome, whereas cells b and c represent those matched pairs in which the members experience opposite results.

The relative risk of matched pair is = $\dfrac{a + c}{a + b}$ is a ratio of exposed to non-exposed matched pairs. The chi-square relative risk of matched pairs may be calculated by using (7.9):

The matched-pair χ^2 -test, [also called McNemar χ^2 test], is the test generally used for comparing proportions in two pair matched groups. It is analogous to categorical data of the paired t-test (discussed in Chapter 4) for continuous variables.

Example 7.15:
Calculate the Relative Risk from Example 7.9 and test its significance.

Solution:
The relative risk of matched pairs is

$$RR_{Matched} = \frac{7+29}{7+14} = \frac{36}{21} = 1.71$$

Therefore, smokers have 1.71 time more chance of myocardial infarction than non-smokers.

(i) $\chi^2_{McNemar}$ $= \dfrac{(b-c)^2}{b+c}$ (7.9)

$$= \frac{(14-29)^2}{14+29} = 5.233$$

The 5% table value of χ^2 at 1 degree of freedom is 3.841. The calculated value of chi-square test, is greater than the table value, therefore, result is significant. Since RR is greater than 1 and the value of chi-square gives significant result, therefore, smokers have 1.71 times more chance of myocardial infarction than non-smokers.

(ii) Confidence limits (using formula 7.21)

$$(1.71)^{1 \pm 1.96\sqrt{5.233}} \text{ or } [1.08, 2.69]$$

This does not include 1, therefore. The result is significant and smokers have 1.71 times more chance of Myocardial Infarction than non-smokers.

When the expected frequency in any cell is less than five, then correction factor (Yates' Correction) may be used in the calculation of chi-square as explained in sub-section (7.4.5).

7.7.5 Odds ratio and tests of significance

As we know that odds ratio is calculated for case-control study. It is assumed that the [exposure = yes, disease = yes] cell is on the main diagonal of a matrix.

Example 7.16:
We have taken an hypothetical example to show how odds ratio is calculated. The data is given in Table 7.26.

Table 7.26

Non Smokers	Smokers		Total
	Case MI	Control $\overline{\text{MI}}$	
MI	a = 90	b = 40	a + b = 130
$\overline{\text{MI}}$	c = 10	d = 60	c + d = 70
Total	a + c = 100	b + d = 100	a+b+c+d=200

Solution:

Rate of exposure in cases: $a/(a + c) = 90/100 = 90\%$
Rate of exposure in controls: $b/(b + d) = 40/100 = 40\%$

Using (7.19)

$$\text{Odds ratio} = \frac{90 \times 60}{40 \times 10} = 13.5$$

This shows that smokers have 13.5 times more chance of developing myocardial infarction than non-smokers.

(i) Test of significance

(a) Using the method of chi-square

$$\chi^2 = \frac{(90 \times 60 - 40 \times 10)^2\ 200}{130 \times 70 \times 100 \times 100} = 54.9$$

Since calculated value of chi-square gives significant result, therefore, we say with 95% confidence that smokers have 13.5 times more chance of myocardial infarction than non-smokers.

(ii) Confidence limits (using 7.21), we get

(i) $(13.5)^{1 \pm 1.96/\sqrt{54.9}}$ or $[6.87, 26.55]$

(ii) (OR) $e^{\pm 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$ (7.23)

$(13.5)e^{\pm 1.96 \sqrt{\frac{1}{90} + \frac{1}{40} + \frac{1}{10} + \frac{1}{60}}}$ or $[6.275 \sim 29.04]$

This does not include 1, therefore, we confirm our previous result.

Example S7-7

- Enter the data in the following manner.

	Non_smokers	Smokers
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1

	Non_smokers	Smokers
1	Case	MI
2	Case	MI
3	Case	MI
4	Case	MI
5	Case	MI
6	Case	MI
7	Case	MI
8	Case	MI
9	Case	MI
10	Case	MI

(up to row 200)

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Non_smokers	Numeric	8	0	Myocardial infarction	{1, Case}...	None	9	Right	Nominal	Input
Smokers	Numeric	8	0	Myocardial infarction	{1, MI}...	None	8	Right	Nominal	Input

To proceed for analysis

Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab*.

Analyze→ Descriptive Statistics→ Crosstabs …

Move the variable "Non-smoking" to the Row(s):

Move the variable "Smoking" to the Column(s):
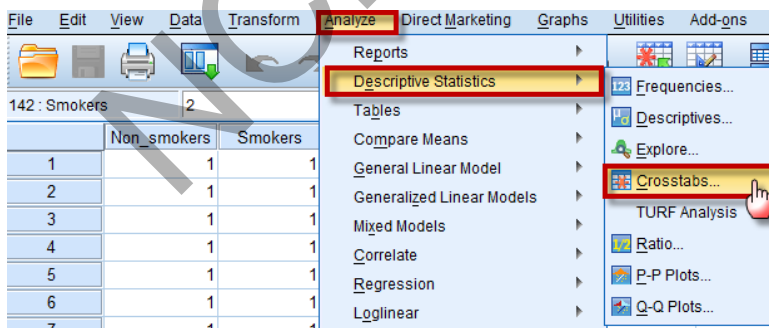
We click on Statistics... and mark on "Cochran's and Mentel-Haenszel statistics",
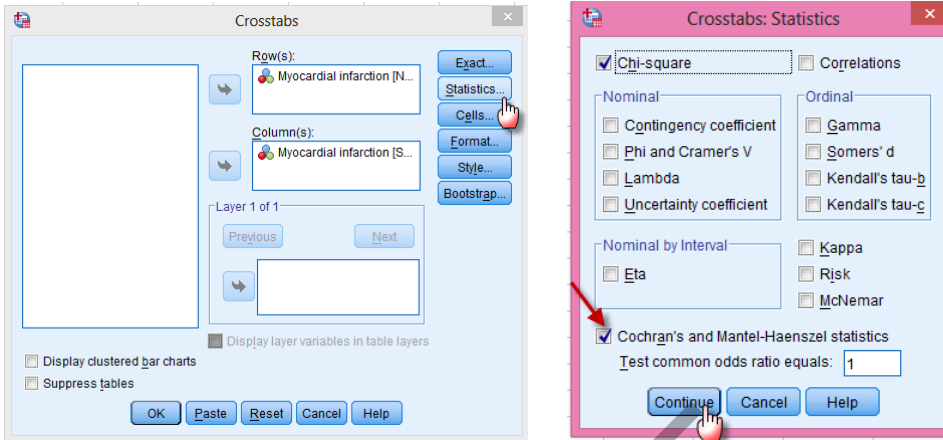
Now click on **Continue** and on **OK**, to get the following output:

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			13.500
In(Estimate)			2.603
Std. Error of In(Estimate)			.391
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	6.275
		Upper Bound	29.043
	In(Common Odds Ratio)	Lower Bound	1.837
		Upper Bound	3.369

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

7.7.6 Matched analysis in case-control study

A matched analysis in case-control is similar to the analysis of matched Cohort studies with dichotomous exposure and outcome.

Example 7.17:

Data regarding case-control study of breast feeding (BF) as a possible protective factor against subsequent gastroenteritis (intestinal infection) in first year of life in 100 pairs (200 total subjects) of infant matched for age, sex and socio-economic status is given in Table 7.27.

Table 7.27
Cases

		BF	$\overline{\text{BF}}$	Total
Controls	**BF**	6	26	32
		a	b	
		c	d	
	$\overline{\text{BF}}$	9	59	68
	Total	15	85	100

Calculate the odds ratio for case-control study.

Solution:
The matched odds ratio is defined as the ratio of the number of pairs discordant for exposure history i.e.

$$\text{OR}_{\text{Matched}} = \frac{c}{b} = \frac{9}{26} = 0.35 \tag{7.24}$$

Since OR is less than 1, so we say that breast-feeding has a protective effect against gastroenteritis.

(i) Test of significance

(a) **Using chi-square**
Matched pairs: McNemar test is used to calculate chi-square.

$$\chi^2_{\text{McNemar}} = \frac{(b-c)^2}{b+c} = \frac{(9-26)^2}{9+26} = \frac{289}{35} = 8.25$$

(ii) Confidence limits

$$(0.35)^{1 \pm 1.96\sqrt{8.25}} \quad \text{or} \quad [0.17, 0.71]$$

This does not include 1, therefore, result is significant and we confirm our above findings.

7.8 Relation between odds ratio and relative risk

The physicians' health study research group at Harvard Medical School takes the following data from a report on the relationship between aspirin use and myocardial infarction. The physicians' study was a five-year randomized study testing whether intake reduces mortality from cardiovascular disease. Physicians were blind in the study and did not know which type of pill they were taking. The results are given in Table 7.28.

Table 7.28

Group	Myocardial Infarction		
	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

Solution:

$$p_1 = 189/11034 = 0.0171 ; \quad p_2 = 104/11037 = 0.0094$$

The estimated standard error is (see Chapter 3)

$$\sqrt{\frac{0.0171 \times 0.9829}{11034} + \frac{0.0094 \times 0.9906}{11037}} = 0.0015$$

The 95% confidence limits are

$$0.0171 - 0.0094 \pm 1.96 \times 0.0015 \text{ or } [0.005, 0.011]$$

Since this interval contains only positive values, we conclude that taking aspirin reduces the risk of myocardial infarction.

The odds ratio for aspirin study is

$$OR = \frac{(189)(10933)}{(104)(10845)} = 1.832$$

The estimated odds of myocardial infarction for physicians taking placebo equal 1.832 times the estimated odds for physicians taking aspirin. The estimated odds were 83.2% higher for the placebo group.

A sample odds ratio of 1.832 does not mean that p_1 is 1.832 times p_2; that would be the interpretation of a relative risk. The relative risk will:

$$RR = \frac{189/11034}{104/11037} = \frac{0.0171}{0.0094} = 1.819$$

The relationship between odds ratio and relative risk is given as:

$$\text{Consider } OR = \frac{a}{c} \cdot \frac{d}{b} \text{ and } RR = \frac{a/(a+b)}{c/(c+d)}$$

$$\text{Then odds ratio } = \frac{ad}{bc} = \text{Relative risk } \frac{d/(c+d)}{b/(a+b)} \qquad (7.25)$$

$$= 1.832 = 1.819 \times \frac{0.9906}{0.9829} = 1.833$$

When the proportion of success is close to zero for both the groups, the fraction in the last term of this expression approximately equals to 1.0, then odds ratio and relative risk take similar values. In the above table for each group, the sample preparation of myocardial infarction cases is close to zero. Thus, the sample odds ratio of 1.83 is similar to the sample relative risk of 1.82. In such a case, an odds ratio of 1.83 does mean that $p_1 \left[= a/(a+b) \right]$ is about 1.83 times $p_2 \left[= c/(c+d) \right]$. The relationship between the odds ratio and the relative risk is useful as for some data sets, calculation of relative risk is not possible, yet one can calculate the odds ratio and use it to approximate the relative risk.

7.9 Mantel-Haenszel Procedure for Relative Risk and Odds Ratio

When exposure and outcome variables are all categorical and the number of variables is small, stratification is usually the procedure of choice. We have seen in Chapter-3 that stratification controls sampling error. Here a more commonly used approach is the Mantel-Haenszel procedure in which the result from each stratum are weighted approximately according to the sample size of stratum to yield an overall relative risk or odds ratio.

The Mantel-Haenszel procedure is the most appropriate and widely used technique for controlling a small number of categorical confounding factors. As the number of confounding factors increases, the computations become difficult, moreover, there may be some loss of control when continuous confounding variables are arbitrarily categorized. For these situations multiple logistic regressions (to be discussed in Chapter-8) is commonly used for multiple confounding factors. Note that Mantel-Haenszel tests are generally not affected by tables with zero cell.

Example 7.18:
For a Cohort study, data of success (S) and failure (F) for two medical treatments (T_1 and T_2) which may control confounding variable. (gender) are given below in Table 7.29.

Table 7.29
Outcome

Treatment	S	F	Total
T_1	40	60	100
T_2	60	40	100
Total	100	100	200

Compare T_1 and T_2 and test its significance.

Solution:

Using (7.13), the relative risk of success $[T_1, T_2] = \dfrac{40/100}{60/100} = 0.667$.

This shows that T_1 is less efficient than treatment T_2. The crude relative success of T_1 versus T_2 is 0.667, which may be biased by the confounding effect of sex. To test its significance, the chi-square is calculated using (7.2).

$$\chi^2_{\text{Pearson}} = \frac{(40 \times 40 - 60 \times 60)^2\, 200}{100 \times 100 \times 100 \times 100} = 8.000$$

Since 5% table value for 1 degree of freedom is 3.841, therefore, the result is significant. We say that success of T_1 is as less efficient than T_2.

If the data is stratified by sex, then relative risks for each gender are as follows:

Table 7.30
Stratification of data by sex

Treatment	Males			Females		
	S	F	Total	S	F	Total
T_1	24	3	27	16	57	73
T_2	58	30	88	2	10	12
Total	82	33	115	18	67	85

$$\text{RR (males) of success } (T_1, T_2) = \frac{24/27}{58/88} = 1.349$$

$$\text{RR (females) of success } (T_1, T_2) = \frac{16/73}{2/12} = 1.315$$

In males, treatment T_1 is 1.349 times more effective than T_2. In females, treatment T_1 is 1.315 times more effective than T_2.

This means that relative success rate T_1 versus T_2 has almost equal effect on both sexes.

Example S7-8

- For the data given in table 7.30, Enter the data in the following manner.

	Gender	Treatments	Result		Gender	Treatments	Result
1	1	1	1	1	Male	T1	Success
2	1	1	1	2	Male	T1	Success
3	1	1	1	3	Male	T1	Success
4	1	1	1	4	Male	T1	Success
5	1	1	1	5	Male	T1	Success
6	1	1	1	6	Male	T1	Success
7	1	1	1	7	Male	T1	Success
8	1	1	1	8	Male	T1	Success
9	1	1	1	9	Male	T1	Success
10	1	1	1	10	Male	T1	Success

(up to row 200)

The Variable View is as follows:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Gender	Numeric	8	0		{1, Male}...	None	8	Right	Nominal	Input
2	Treatments	Numeric	8	0	Medical treatments	{1, T1}...	None	9	Right	Nominal	Input
3	Result	Numeric	8	0		{1, Success...	None	8	Right	Nominal	Input

To proceed for analysis for table 7.29 (regardless of gender),

Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab.*

Analyze→ Descriptive Statistics→ Crosstabs …
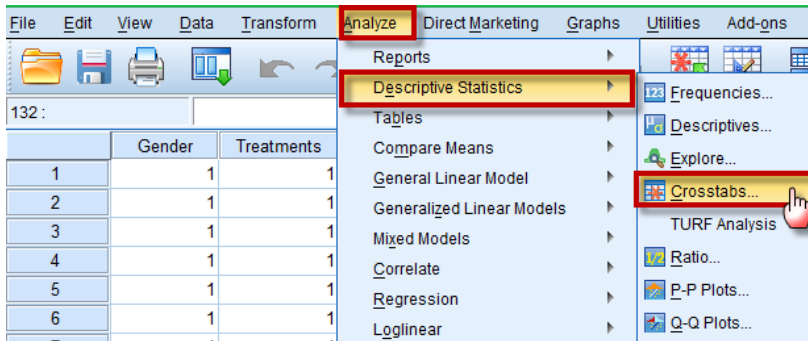
Move the variable "Treatment" to the Row(s):

Move the variable "Result" to the Column(s):

We click on [Statistics...] and mark on "Chi-square" and "Risk",
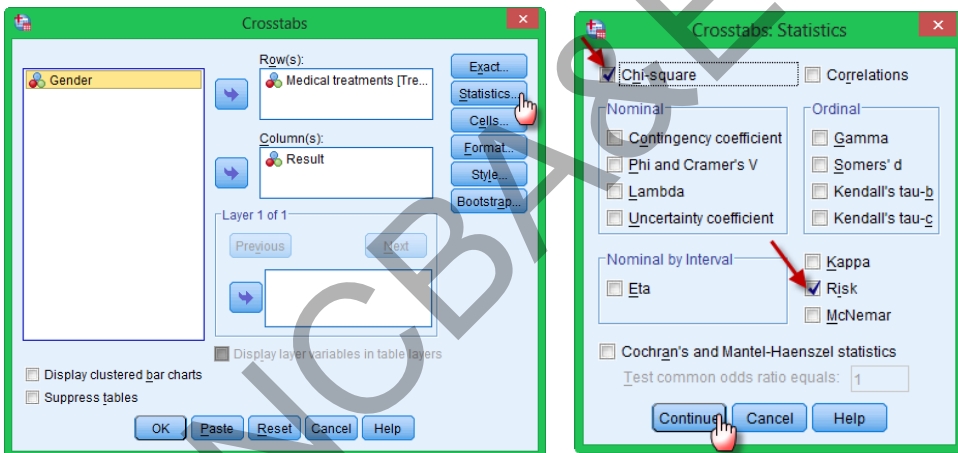
Now click on [Continue] and on [OK], to get the following outputs:

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Medical treatments (T1 / T2)	.444	.252	.783
For cohort Result = Success	.667	.500	.890
For cohort Result = Failure	1.500	1.124	2.002
N of Valid Cases	200		

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8.000[a]	1	.005		
Continuity Correction[b]	7.220	1	.007		
Likelihood Ratio	8.054	1	.005		
Fisher's Exact Test				.007	.004
Linear-by-Linear Association	7.960	1	.005		
N of Valid Cases	200				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 50.00.

b. Computed only for a 2x2 table.

The results are exactly as given by hand calculation.

Now to proceed for analysis for table 7.30, we first split the file using ▦ according to the gender, as follows:
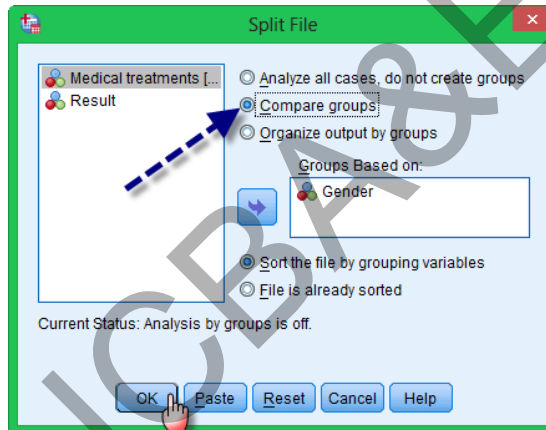
Now, click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab*.

Analyze➔ Descriptive Statistics➔ Crosstabs …
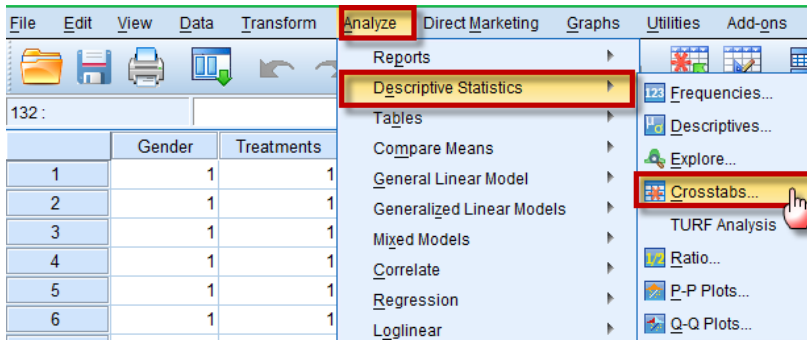
Move the variable "Treatment" to the Row(s):

Move the variable "Result" to the Column(s):

We click on [Statistics...] and mark on "Risk",
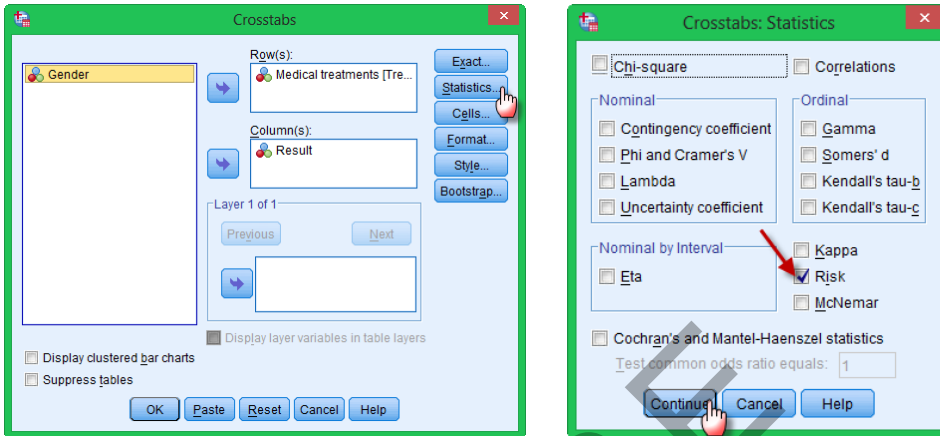
Now click on [Continue] and on [OK], to get the following outputs:

Medical treatments * Result Crosstabulation

Count

Gender			Result		Total
			Success	Failure	
Male	Medical treatments	T1	24	3	27
		T2	58	30	88
	Total		82	33	115
Female	Medical treatments	T1	16	57	73
		T2	2	10	12
	Total		18	67	85

Risk Estimate

Gender		Value	95% Confidence Interval	
			Lower	Upper
Male	Odds Ratio for Medical treatments (T1 / T2)	4.138	1.152	14.862
	For cohort Result = Success	1.349	1.103	1.649
	For cohort Result = Failure	.326	.108	.985
	N of Valid Cases	115		
Female	Odds Ratio for Medical treatments (T1 / T2)	1.404	.279	7.066
	For cohort Result = Success	1.315	.345	5.008
	For cohort Result = Failure	.937	.708	1.241
	N of Valid Cases	85		

The results are exactly as given by hand calculation.

7.9.1 Mantel-Haenszel relative risk

The relative risk is calculated as

$$RR_{MH} = \frac{\Sigma a_i(c_i + d_i)/n_i}{\Sigma c_i(a_i + b_i)/n_i} \tag{7.26}$$

The Mantel-Haenszel relative risk analysis combines the stratum-specific result to yield an un-confounded overall result. Using (7.28) we get

$$RR_{MH} = \frac{\dfrac{24(58+30)}{115} + \dfrac{16(2+10)}{85}}{\dfrac{58(24+3)}{115} + \dfrac{2(16+57)}{85}} = 1.34$$

This is not very much different from the relative risk of males and females.

7.9.2 Mantel-Haenszel chi-square

As we know that this is a method of controlling confounding in stratification. This requires that the confounder be categorical variable. If it is continuous, categorized, the formula of chi-square given by Mantel-Haenszel for the significance of Mantel-Haenszel relative risk is

$$\chi^2_{MH} = \frac{\left[\Sigma \dfrac{a_i d_i - b_i c_i}{n_i}\right]^2}{\Sigma \dfrac{r_{1i} r_{2i} c_{1i} c_{2i}}{(n_i - 1)n_i^2}}, \tag{7.27}$$

with 1 df, where r_{1i} and r_{2i} are row totals for different strata and c_{1i} and c_{2i} are column totals for different strata. Using (7.27), we get.

$$\chi^2_{MH} = \frac{\left[\dfrac{24 \times 30 - 58 \times 3}{115} + \dfrac{16 \times 10 - 57 \times 2}{85}\right]^2}{\dfrac{27 \times 88 \times 82 \times 33}{114 \times (115)^2} + \dfrac{73 \times 12 \times 18 \times 67}{84 \times (85)^2}} = 4.658$$

which is more than 3.841 (table value). The result is significant, we say that gender does not play role as confounder. We conclude that the higher success rate of T_1 observed in the sample arose was by chance.

The output of SPSS Package is given as below there are some minor difference in the result, which is due to approximation in manual calculations.

Tests for Homogeneity of the Odds

Statistics		Chi-Squared	df	Asymp. Sig. (2-sided)
Conditional Independence	Cochran's-Mantel-Haenszel	4.703	1	.030
		3.819	1	.051
Homogeneity	Breslow-Day-Tarone's	1.096	1	.295
		1.092	1	.296

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

The p-value for 5% degree of freedom for two tailed is 0.030 for one tailed will be $2 \times 0.030 = 0.06$

Risk Estimate

Sex		Value	95% Confidence Interval	
			Lower	Upper
Males	Odds Ratio for Treatment (T1 / T2)	4.138	1.152	14.862
	For cohort Males = S	1.349	1.103	1.649
	For cohort Males = F	.326	.108	.985
	N of Valid Cases	115		
Females	Odds Ratio for Treatment (T1 / T2)	1.404	.279	7.066
	For cohort Females = S	1.315	.345	5.008
	For cohort Females = F	.937	.708	1.241
	N of Valid Cases	85		

Mantel-Haenszel Common Odds Ratio

Estimate			2.853
ln(Estimate)			1.048
Std. Error of			.507
Asymp. Sig. (2-sided)			.038
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1.057
		Upper Bound	7.699
	ln(Common Odds Ratio)	Lower Bound	.056
		Upper Bound	2.041

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption.

7.9.3 Mantel-Haenszel odds ratio

For case-control study the odds ratio is calculated as:

$$OR_{MH} = \frac{\Sigma a_i d_i / n_i}{\Sigma b_i c_i / n_i} \tag{7.28}$$

Example 7.19:
A hypothetical data regarding coffee drinkers and renal cancer are as:

Table 7.31
Renal Cancer

Coffee drinker	*RC*	$\overline{RC}$	Total
CD	400	333	733
$\overline{CD}$	100	167	267
Total	500	500	1000

Solution:
Using (7.19), the odds ratio is

$$OR = \frac{400 \times 167}{100 \times 333} = 2.006$$

Using (7.2), the chi-square is

$$\chi^2 = \frac{(66800 - 33300)^2 1000}{733 \times 500 \times 500 \times 267} = 22.9$$

Since χ^2 is significant, OR is 2; therefore, coffee drinkers have double the risk of renal cancer than non-coffee drinkers.

We take smoking as confounding factor. The data for smokers and non-smokers are given as:

Table 7.32
Stratification of data according to smokers and non-S smokers

	Smokers			non-smokers		
	RC	$\overline{RC}$	Total	*RC*	$\overline{RC}$	Total
CD	350	80	430	50	253	303
$\overline{CD}$	75	20	95	25	147	172
Total	425	100	525	75	400	475

The odds ratios for smokers and non-smokers are 1.17 and 1.16. The Mantel- Haenszel the odds ratio using (7.28) is:

$$OR_{MH} = \frac{\dfrac{350 \times 20}{525} + \dfrac{50 \times 147}{475}}{\dfrac{80 \times 75}{525} + \dfrac{253 \times 25}{475}} = 1.16$$

Using (7.27), the chi-square is

$$\chi^2_{MH} = 0.619$$

Since at 5% level of significance, the calculated value of χ^2 value is less than the table value, therefore, OR > 1 is by chance, therefore smoking does not play any role as confounder.

Here is the results using IBM-SPSS:

Example S7-9

- For the data given in table 7.32, Enter the data in the following manner.

	Smoking	Coffee	Renal
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1

	Smoking	Coffee	Renal
1	Smoker	Yes	Yes
2	Smoker	Yes	Yes
3	Smoker	Yes	Yes
4	Smoker	Yes	Yes
5	Smoker	Yes	Yes
6	Smoker	Yes	Yes
7	Smoker	Yes	Yes
8	Smoker	Yes	Yes
9	Smoker	Yes	Yes
10	Smoker	Yes	Yes

(up to row 1000)

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Smoking	Numeric	8	0		{1, Smoker}...	None	8	Right	Nominal	Input
Coffee	Numeric	8	0	Coffee Drinker	{1, Yes}...	None	8	Right	Nominal	Input
Renal	Numeric	8	0	Renal Cancer	{1, Yes}...	None	8	Right	Nominal	Input

To proceed for analysis for table 7.31 (regardless of Smoking),

Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab.*

Analyze→ Descriptive Statistics→ Crosstabs …
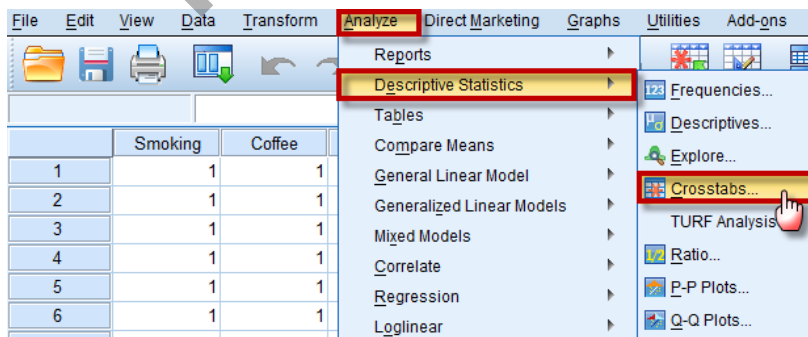
Move the variable "Coffee" to the Row(s):

Move the variable "Renal" to the Column(s):

We click on [Statistics...] and mark on "Chi-square" and "Risk",
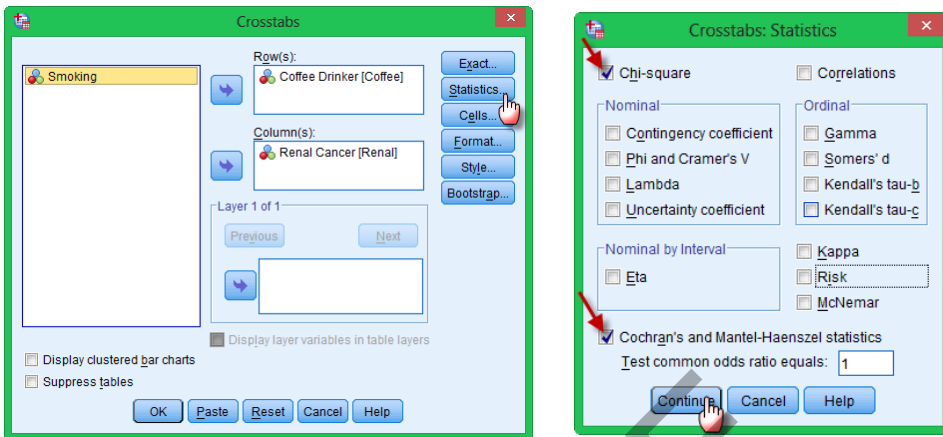
Now click on [Continue] and on [OK], to get the following outputs:

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			2.006
ln(Estimate)			.696
Std. Error of ln(Estimate)			.147
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1.505
		Upper Bound	2.674
	ln(Common Odds Ratio)	Lower Bound	.409
		Upper Bound	.983

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	22.937[a]	1	.000		
Continuity Correction[b]	22.257	1	.000		
Likelihood Ratio	23.126	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	22.914	1	.000		
N of Valid Cases	1000				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 133.50.

b. Computed only for a 2x2 table

The results are exactly as given by hand calculation.

Now to proceed for analysis for table 7.32, we first split the file using ⊞ according to the Smoking, as follows:
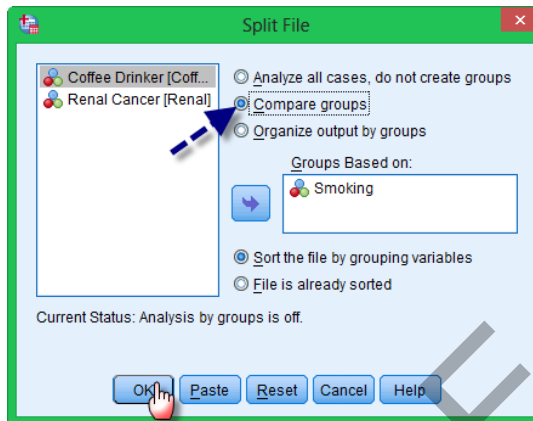
Now, click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab*.

Analyze→ Descriptive Statistics→ Crosstabs …

We click on **Statistics...** and mark on "Risk",
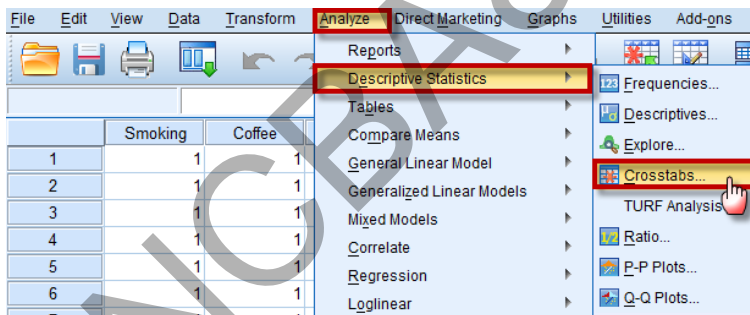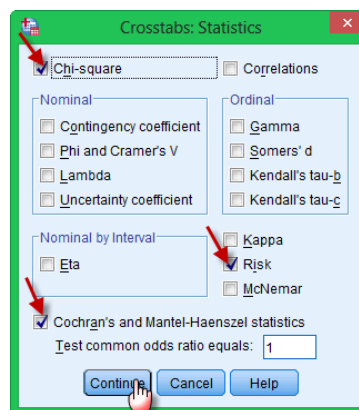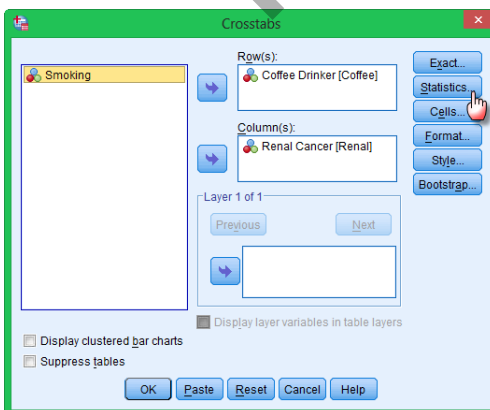
Now click on Continue and on OK, to get the following outputs:

Risk Estimate

Sex		Value	95% Confidence Interval	
			Lower	Upper
Smokers	Odds Ratio for Treatment (Coffee Drinker / Non-Coffee Drinker)	1.167	.673	2.022
	For cohort Males = Ranal Cancer	1.031	.921	1.155
	For cohort Males = 2	.884	.571	1.368
	N of Valid Cases	525		
Non-Smokers	Odds Ratio for Treatment (Coffee Drinker / Non-Coffee Drinker)	1.162	.690	1.957
	For cohort Males = Ranal Cancer	1.135	.730	1.767
	For cohort Males = 2	.977	.902	1.058
	N of Valid Cases	475		

Tests for Homogeneity of the Odds Ratio

Statistics		Chi-Squared	df	Asymp. Sig. (2-sided)
Conditional Independence	Cochran's	.621	1	.431
	Mantel-Haenszel	.476	1	.490
Homogeneity	Breslow-Day	.000	1	.992
	Tarone's	.000	1	.992

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			1.164
ln(Estimate)			.152
Std. Error of ln(Estimate)			.193
Asymp. Sig. (2-sided)			.431
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	.797
		Upper Bound	1.700
	ln(Common Odds Ratio)	Lower Bound	-.226
		Upper Bound	.530

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

Example 7.20:
Calculate the odds ratio from the data given in Example 7.10. Also calculate odds ratio using Mantel-Haenszel method.

Solution:

Table 7.33

City	Smoking Status	Lung Cancer Yes	Lung Cancer No	Total	Odds ratio
1	S	126	100	226	
	S̄	35	61	96	2.20
		161	161	322	
2	S	908	688	1596	
	S̄	497	807	1304	2.14
		1405	1495	2900	
3	S	913	747	1660	
	S̄	336	598	934	2.18
		1249	1345	2594	
4	S	235	172	407	
	S̄	58	121	179	2.85
		293	293	586	
5	S	402	308	710	
	S̄	121	215	336	2.32
		523	523	1046	
6	S	182	156	338	
	S̄	72	98	170	1.59
		254	254	508	
7	S	60	99	159	
	S̄	11	43	54	2.37
		71	142	213	
8	S	104	89	193	
	S̄	21	36	57	2.00
		125	125	250	

$$OR_{MH} = \frac{126 \times 61/322 + ... + 104 \times 36/250}{35 \times 100/322 + ... + 21 \times 89/250} = 2.17$$

Testing the Significance

(i) **Using χ^2 method**

The Mantel-Haenszel chi-square has been calculated in Example 7.10 and is 280.2 which is much more than the table value of χ^2 for 1 df. Therefore, there is a strong evidence that smoking causes cancer.

(ii) **Using confidence limits**

The formula for the calculation of standard error is very complex [Robinson et al. (1996)] but SPSS Package is used to compute the standard error. 95% confidence interval is (1.98, 2.38) which does not include 1. Therefore one can conclude that smoking causes cancer.

The IBM-SPSS output for odds ratios and confidence limits is as:

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			2.174
ln(Estimate)			.777
Std. Error of ln(Estimate)			.047
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	1.984
		Upper Bound	2.383
	ln(Common Odds Ratio)	Lower Bound	.685
		Upper Bound	.868

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

(The student has to check the results using IBM-SPSS)

7.10 Sensitivity, Specificity and Kappa-Statistic

7.10.1 Screening test

A test is reliable if it provides consistent result when performed more than once. The test is valid if it correctly identifies those who probably have the disease (true positive) and those who are probably free from disease (true negative). Validity is measured by both sensitivity and specificity.

7.10.2 Validity of a screening test

Consider the screening test results of patients in the 2x2 contingency table, where P = patients with disease, $\overline{P}$ = patients with no diseases.

Table 7.34
Disease status

Screening test	P	$\overline{P}$	Total
Positive	a = **TP**	b = FP	a + b
Negative	c = FN	d = **TN**	c + d
Total	a + c	b + d	a + b + c + d

TP = true positive;   The result is positive and patient possesses the disease.
TN = true negative;   The result is negative and patient possesses no disease.
FP = false positive;   The result is positive and patient doesn't possess the disease.
FN = false negative; The result is negative and patient has disease.

(i) *Sensitivity* is the *proportion* of truly ill people in the screened population who are identified as ill by the screening test. It is the ability of the test to identify accurately those who have the disease. It is calculated as a/(a + c).

(ii) **Specificity** is the proportion of truly healthy people who are so identified by the screening test. It is the ability of the test to identify accurately those who do not have the disease. It is calculated as d/(b + d).

(iii) **Positive predictive value (rate)** is the probability of a person having the disease when the test is positive. (This is also called predictive value of a positive test) $\boxed{a/(a + b)}$.

(iv) **Negative predictive value (rate)** is the probability of a person not having the disease when the test is negative. (This is also called predictive value of a negative test) $\boxed{d/(c + d)}$.

(v) **False positive rate** is the proportion that a *disease-free* person has a positive test result $\boxed{b/(b + d)}$.

(vi) **False negative rate** is the proportion that a *diseased individual* will have a negative test result $\boxed{c/(c + d)}$.

(vii) Prevalence of disease = (a + c) / (a + b + c + d)

Example 7.21:
In a BCP screening test of 1600 cancer for breast patients, the results are given below:

Table 7.35

Test	Disease		Total
	D⁺	$\overline{D}$	Total
Positive	a 570 TP	b 150 FS	720
Negative	c 30 FN	d 850 TN	880
Total	600	1000	1600

Compute validity of screening test and discuss the result.

Test = BCP, Disease = Breast cancer

Solution:

(i) **Sensitivity** = $\dfrac{a}{a+c}$ = $\dfrac{570}{600}$ = 0.95×100 = 95%

It shows that 95% of patients are correctly identified as cases of disease and 5% are incorrectly identified as cancer patients.

(ii) **Specificity** $= \dfrac{d}{b+d} = \dfrac{850}{1000} = 0.85 \times 100 = 85\%$

It shows that 85% of patients correctly identified as cases of free from disease.

(iii) **Positive predictive value** $= \dfrac{a}{a+b} = \dfrac{570}{720} = 0.792 \times 100 = 79.2\%$

0.792 is the probability of patients having the disease as the test result is positive.

(iv) **Negative predictive value** $= \dfrac{d}{c+d} = \dfrac{850}{880} = 0.966 \times 100 = 96.6\%$

Since the test is negative 0.966 is the probability of not having the disease.

(V) **False positive rate** $= \dfrac{b}{b+d} = \dfrac{150}{1000} = 0.15 \times 100 = 15\%$

15% of the patients that are diseased free have a positive test result.

(vi) **False negative rate** $= \dfrac{c}{c+d} = \dfrac{30}{880} = 0.034 \times 100 = 3.4\%$

3.4% of the patients that are diseased individual and have negative result.

(vii) **Prevalence of disease** $= \dfrac{a+c}{a+b+c+d} = \dfrac{570+30}{570+150+30+850} = \dfrac{600}{1600} = 37.5\%$

Since the sensitivity and specificity are both large whereas false positive and false negative are small, therefore, the test is useful and valid.

The IBM-SPSS package results are as follows:-

Example S7-10

- For the data given in table 7.35, Enter the data in the following manner.

	Test	Disease			Test	Disease
1	1	1		1	Positive	D+
2	1	1		2	Positive	D+
3	1	1		3	Positive	D+
4	1	1		4	Positive	D+
5	1	1		5	Positive	D+
6	1	1		6	Positive	D+
7	1	1		7	Positive	D+
8	1	1		8	Positive	D+
9	1	1		9	Positive	D+
10	1	1		10	Positive	D+

(up to row 1600)

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Test	Numeric	8	0		{1, Positive}...	None	8	Right	Nominal	Input
Disease	Numeric	8	0		{1, D+}...	None	8	Right	Nominal	Input

Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab.*

Analyze→ Descriptive Statistics→ Crosstabs …
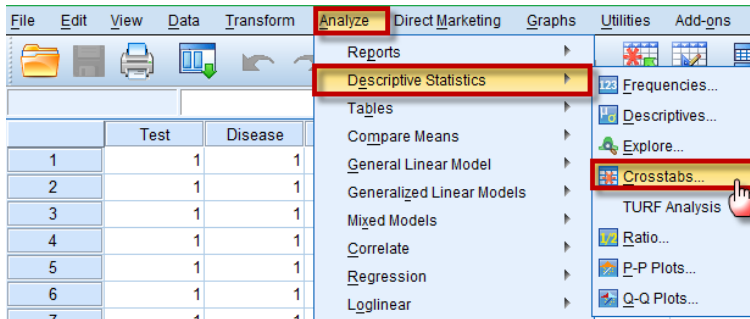
We click on ▭Statistics... and mark on "Row" and "Column" Percentages;

Now click on ▭Continue and on ▭OK , to get the following outputs:

7.10.3 Diagnostic Tests (Sensitivity and Specificity)

The simplest diagnostic test is one where the results of an investigation, such as an x-ray examination or biopsy, are used to classify patients into two groups according to the presence and absence of symptom. For example, the following table (7.32) shows the results of a test on a liver scan and the correct diagnosis based on necropsy, biopsy, or surgical inspection. The data is given in Table 7.36.

Table 7.36
Results of liver scan and correct diagnosis

Liver scan	Pathology		Total
	Abnormal (+)	Normal (-)	
Abnormal (+)	231	32	263
Normal (-)	27	54	81
Total	258	86	344

How good is the liver scan as diagnosis of abnormal pathology?

One approach is to calculate the proportions of patients with normal and abnormal liver scans who are correctly diagnosed by the scan. The terms positive and negative are used to refer to the presence or absence of the condition of interest, here abnormal pathology. Thus there are 258 true positives and 86 true negative. The proportion of these two groups that were correctly diagnosed by the scan were $231/258 = 0.895$ and $54/86 = 0.628$. These two proportions are known as *sensitivity* and *specificity* respectively. We can thus say that, based on the sample studies, we would expect about 90% of patients with abnormal pathology to have abnormal liver scans, while about 63% of those with normal pathology would have normal lever scans.

Sensitivity and specificity are one approach to quantify the diagnostic ability of the test. In clinical practice, however, the test result is all that is known, so we want to know *how good the test is at predicting abnormality .In other words, what proportion of patients with abnormal test results are truly abnormal?* The whole point of a diagnostic test is to use it to make a diagnosis, so we need to know the probability that the test will give the correct diagnosis. The sensitivity and specificity do not give us this information. Instead we must approach the data from the direction of the test results, using predictive values, i.e. positive predictive value and negative predictive value.

If we go back to the above table we see that 231 from 263 patients with abnormal liver Scans had abnormal pathology, giving the proportion of correct diagnoses as $231/263 = 0.878 \approx 88\%$. Similarly, among the 81 patients with normal liver scans, the proportion of correct diagnoses was $54/81 = 0.667 \approx 67\%$. These proportions are of limited validity, however, the predictive values of a test in clinical practice depend critically on the prevalence of the abnormality in the patients being tested. This may well differ from the prevalence in a published study assessing the usefulness of the test.

In the liver scan study, the prevalence of abnormality is $258/344 = 0.75 \approx 75\%$. If the same test was used in a different clinical setting where the prevalence of abnormality was 0.25 (25%), we would have positive predictive value of 0.45 and a negative predictive value of 0.95. The rare the abnormality the more sure we can be that a negative test

indicates no abnormality and the less sure that a positive result really indicates an abnormality. Predictive values observed in one study do not apply universally. The other ways of calculating the positive and negative predictive values (PPV and NPV) are:

$$\text{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity})(1 - \text{prevalence})} \qquad (7.29)$$

$$\text{NPV} = \frac{\text{specificity} \times (1 - \text{prevalence})}{(1 - \text{sensitivity}) \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence})} \qquad (7.30)$$

If the prevalence of the disease is very low, the positive predictive value will not be close to 1 even if both the sensitivity and specificity are high. Thus in screening the general population it is inevitable that many people with positive test results will be false positive.

The prevalence can be interpreted as the probability that the subject has the disease, before the test is carried out, known as the prior probability of disease. The positive and negative predictive values are the revised estimates of the same probability for those subjects who are positive and negative on the test and are known as posterior probabilities. The difference between the prior and posterior probabilities is one way of assessing the usefulness of the test.

For any test result, we compare the probability of having a positive result the patient is truly diseased with the corresponding probability if he or she were healthy. The ratio of these probabilities is called *likelihood ratio* and is calculated as

$$\text{LR} = \frac{\text{sensitivity}}{1 - \text{specificity}} \qquad (7.31)$$

The likelihood ratio indicates the value of the test for increasing certainty about a positive diagnosis. For the lever scan data the prevalence of abnormal pathology was 0.75, so the pretest odds of disease were 0.75/(1 - 0.75) = 3.0. The sensitivity was 0.895 and the specificity was 0.628. The post-test odds of disease given a positive test is 0.878/(1 - 0.878) = 7.22 and the likelihood ratio is 0.895/(1 - 0.628) = 2.41. The posttest odds of having the disease can be calculated as:

Pretest odds x likelihood ratio = 3.0 x 2.41 = 7.23.

A high likelihood ratio may show that the test is useful, but it does not necessarily follow that the positive test is a good indicator of the presence of disease.

7.10.4 Kappa (Cohen's Kappa)-Statistic

In Chapter 6, we have discussed the method of correlation that is used to measure the degree of agreement between two variables. Pearson's correlation coefficient is calculated when the variables are continuous whereas Spearman's rank correlation (Chapter-8) coefficient is used when the variables are ordinal.

For qualitative variables, a frequently used index of agreement between observers is known as Cohen's Kappa coefficient (Cohen-1960). This measure has the desirable

feature of showing how much more agreement there is than would be expected by chance. Kappa has been extended to situations where more than one rater is to be compared and where the variable is polychotomous rather than dichotomous (Fliess-1981).

Kappa (K) statistic is calculated as:

$$K = \frac{P_0 - P_c}{1 - P_c} \tag{7.32}$$

where P_0 = observed proportion of agreement, and P_c = expected proportion of agreement under the assumption of independence.

Landis and Koch (1977) provided the following guidelines for the evaluation of Kappa. These guidelines are arbitrary but potentially useful *benchmarks* for evaluating observed values of the Kappa coefficient. They are as follows:

Table 7.37

K	Strength of agreement
0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost perfect

In general practice K > 0.75 means excellent agreement; $0.4 \leq K \leq 0.75$ means good agreement; less than 0.4 means poor agreement (Fleiss 1981). This method is often used to investigate the reliability of the categorical scale usually by evaluating agreement between the two observers. It is not, however, an adequate measure of agreement since it ignores agreement between the observers that might be due to chance. To illustrate the problem we take some examples.

Example 7.22:
A diet questionnaire was administered by male to 537 females on two different occasions several months apart regarding beef consumption. The data regarding beef consumption reported by 537 females at two different surveys are as:

Table 7.38
Consumption of beef

		Survey 2		Total
		= 1 serving/week	> 1 serving/week	
Survey 1	= 1 serving/week	136	92	228
	> 1 serving/week	69	240	309
	Total	205	332	537

Solution:
Since in the calculations expected frequencies are involved, therefore, these are calculated as:

Expected Frequencies			
	Survey 2		Total
	= 1 serving/week	> 1 serving/week	
Survey 1 = serving/week	87	141	228
> serving/week	118	191	309
Total	205	332	537

$$P_0 = \frac{136 + 240}{537} = \text{(observed proportion of agreement)} = 0.70$$

$$P_c = \frac{87 + 191}{537} = \text{(expected proportion of agreement)} = 0.52$$

$$K = \frac{0.70 - 0.52}{1 - 0.52} = 0.375 = 37.5\%$$

IBM-SPSS package is used to calculate Kappa, the data are entered as for χ^2 statistic, as in the following example;

Example S7-11

- For the data given in table 7.38, Enter the data in the following manner.

	S1	S2			S1	S2
1	1	1	1		= serving/week	= 1 serving/week
2	1	1	2		= serving/week	= 1 serving/week
3	1	1	3		= serving/week	= 1 serving/week
4	1	1	4		= serving/week	= 1 serving/week
5	1	1	5		= serving/week	= 1 serving/week
6	1	1	6		= serving/week	= 1 serving/week
7	1	1	7		= serving/week	= 1 serving/week
8	1	1	8		= serving/week	= 1 serving/week
9	1	1	9		= serving/week	= 1 serving/week
10	1	1	10		= serving/week	= 1 serving/week

(up to row 537)

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
S1	Numeric	8	0	Survey 1	{1, = serving...	None	8	Right	Nominal	Input
S2	Numeric	8	0	Survey 1	{1, = 1 servi...	None	8	Right	Nominal	Input

Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab*.

Analyze→ Descriptive Statistics→ Crosstabs …
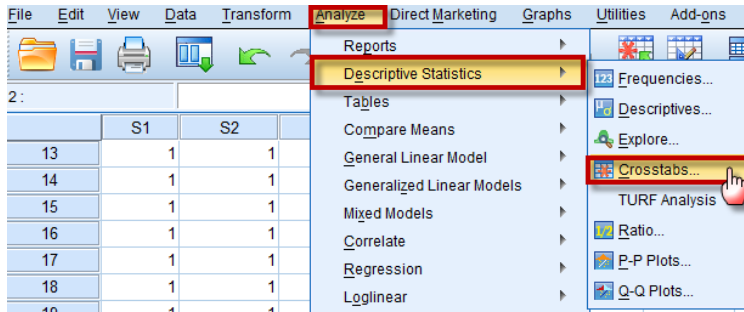
We click on [Statistics...] and mark on "Chi-square", "Phi and Cramer's V" and "Kappa";
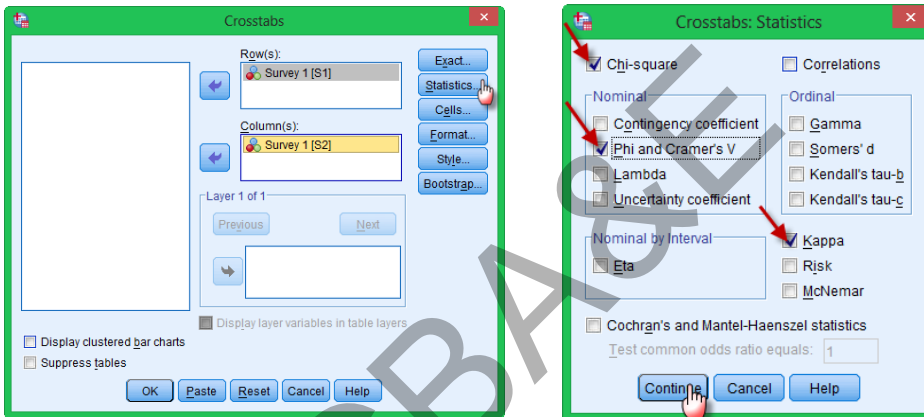
Now click on [Continue] and on [OK], to get the following outputs:

SPSS output for KAPPA

Survey 1 * Survey 1 Crosstabulation

Count

		Survey 1		Total
		= 1 serving/week	> 1 serving/week	
Survey 1	= serving/week	136	92	228
	> serving/week	69	240	309
Total		205	332	537

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	77.417[b]	1	.000		
Continuity Correction[a]	75.844	1	.000		
Likelihood Ratio	78.396	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	77.273	1	.000		
N of Valid Cases	537				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 87.04.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.380			.000
	Cramer's V	.380			.000
Measure of Agreement	Kappa	.378	.040	8.799	.000
N of Valid Cases		537			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Since K = 37.8 %, therefore, according to scale suggested by Fliess (1981), there is a poor agreement between the two related information. The details of Kappa-statistic will be discussed in Chapter 10. Phi (ϕ)= 0.37969 is almost identical to Kappa.

For some research workers, informal evaluation of observed Kappa values will not be sufficient, instead they will be interested in testing hypotheses about the population Kappa. For this purpose standard error of Kappa needed to be calculated. Fleiss, Cohen and Everitt (1969) derived an asymptotic large sample variance of K. This is beyond the scope of this book. The standard error and confidence limits for population Kappa may be derived by using SPSS Package. This may be calculated as:

In the above example, the value of SE (K) = 0.040. Approximate 95% confidence limits for population Kappa are

$$0.375 \pm 1.96 \text{ x } 0.040 \approx [\ 0.297\ ,\ 0.453].$$

Example 7.23:
The data regarding the agreement about the severity of byssinosis for first and second examinations for 183 patients are given below. Calculate the agreement index between two examinations.

Table 7.39
Agreement about the severity of byssinosis

		2nd examination			
		Normal	**Grade 1**	**Grade 2**	**Total**
1st Examination	**Normal**	72	6	0	78
	Grade 1	6	47	17	70
	Grade 2	1	14	20	35
	Total	79	67	37	183

Solution:

The expected frequencies (as required) are

		2nd Examination			
		Normal	**Grade 1**	**Grade 2**	**Total**
1st	**Normal**	33.7			78
Examination	**Grade 1**		25.6		70
	Grade 2			7.1	35
	Total	79	67	37	183

Observed proportion of agreement = $\dfrac{72 + 47 + 20}{183} = 0.76$

Expected proportion of agreement = $\dfrac{33.7 + 25.6 + 7.1}{183} = 0.36$

$K = \dfrac{0.76 - 0.36}{1 - 0.36} = 62.5\%$

According to the scale suggested by Fleiss (1981), there is 62.5% agreement that is considered as good agreement. This table is 3x3, we can calculate Cramer's V which is 0.6227 (almost identical). This difference is because of zero frequency in one cell.

IBM-SPSS package is used to calculate Kappa-statistic as follows:

Example S7-12

- For the data given in table 7.33, Enter the data in the following manner.

	first	second
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1

	first	second
1	Normal	Normal
2	Normal	Normal
3	Normal	Normal
4	Normal	Normal
5	Normal	Normal
6	Normal	Normal
7	Normal	Normal
8	Normal	Normal
9	Normal	Normal
10	Normal	Normal

(up to row 183)

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
first	Numeric	8	0	1st Examination	{1, Normal}...	None	6	Right	Nominal	Input
second	Numeric	8	0	2nd Examination	{1, Normal}...	None	8	Right	Nominal	Input

Click *Analyze* then click *Descriptive Statistics* and then click *Cross-tab.*

Analyze→ Descriptive Statistics→ Crosstabs …
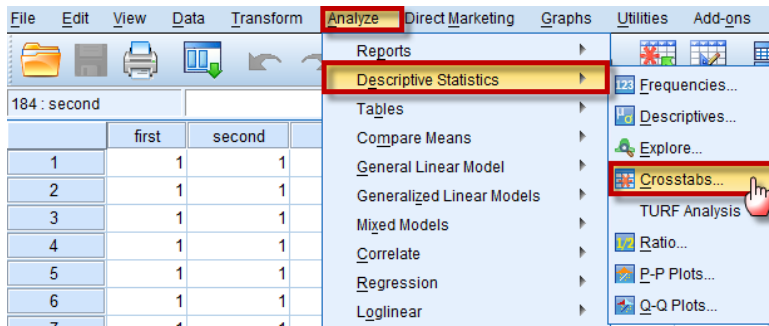
We click on ⬚Cells... and mark on "Observed" and "Expected";
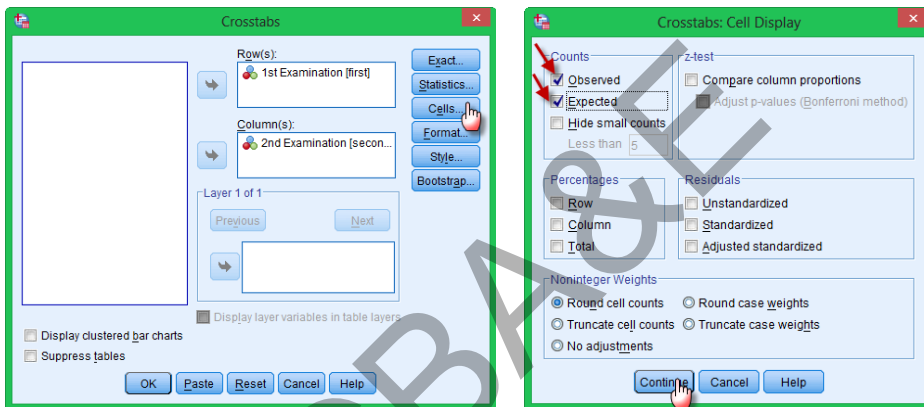
Also, click on ⬚Statistics... and mark on "Chi-square", "Phi and Cramer's V" and "Kappa";
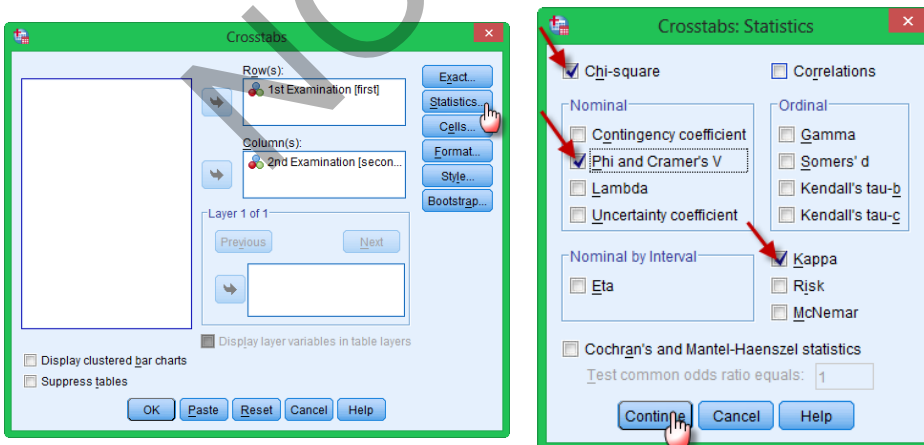
Now click on ⬚Continue and on ⬚OK, to get the following outputs:

SPSS output for Kappa-statistic

EXAM1 * EXAM2 Crosstabulation

			EXAM2			
			1	2	3	Total
1st Exam	1	Count	72	6	0	78
		Expected Count	33.7	28.6	15.8	78.0
	2	Count	6	47	17	70
		Expected Count	30.2	25.6	14.2	70.0
	3	Count	1	14	20	35
		Expected Count	15.1	12.8	7.1	35.0
Total		Count	79	67	37	183
		Expected Count	79.0	67.0	37.0	183.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	151.907[a]	4	.000
Likelihood Ratio	173.160	4	.000
Linear-by-Linear Association	109.463	1	.000
N of Valid Cases	183		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.08.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.911			.000
	Cramer's V	.644			.000
Measure of Agreement	Kappa	.623	.048	11.541	.000
N of Valid Cases		183			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The 95% confidence limits for population Kappa is

$$0.623 \pm 1.96 \times 0.048 \approx (0.53, 0.72).$$

Example 7.24:

The joint ratings of the two clinicians (psychiatrists) regarding 118 patients have been displayed in Table 7.40.

Table 7.40
Rating of two cliniciens

		Psychiatrist 1					
		D1	D2	D3	D4	D5	Total
	D1	22	2	2	0	0	26
	D2	5	7	14	0	0	26
Psychiatrist 2	D3	0	2	36	0	0	38
	D4	0	1	14	7	0	22
	D5	0	0	3	0	3	6
	Total	27	12	69	7	3	118

Calculate the degree of agreement between the two clinicians.

Solution:

$$\text{Observed proportion of agreement} = \frac{22 + 7 + 36 + 7 + 3}{118} = 0.636$$

$$\text{Expected proportion of agreement} = \frac{5.91 + 2.6 + 22.2 + 1.3 + 0.2}{118} = 0.273$$

$$K = \frac{0.636 - 0.273}{1 - 0.273} = 0.499 = 49.9\%$$

The IBM- SPSS package is used and the output is as:

SPSS output for Kappa-statistic

PSYCH1 * PSYCH2 Crosstabulation

			PSYCH2					
			1	2	3	4	5	Total
PSYCH1	1	Count	22	2	2	0	0	26
		Expected Count	5.9	2.6	15.2	1.5	.7	26.0
	2	Count	5	7	14	0	0	26
		Expected Count	5.9	2.6	15.2	1.5	.7	26.0
	3	Count	0	2	36	0	0	38
		Expected Count	8.7	3.9	22.2	2.3	1.0	38.0
	4	Count	0	1	14	7	0	22
		Expected Count	5.0	2.2	12.9	1.3	.6	22.0
	5	Count	0	0	3	0	3	6
		Expected Count	1.4	.6	3.5	.4	.2	6.0
Total		Count	27	12	69	7	3	118
		Expected Count	27.0	12.0	69.0	7.0	3.0	118.0

Symmetric Measures

	Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement Kappa	.498	.057	10.335	.000
N of Valid Cases	118			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The 95% confidence limits for population K may be calculated as:

$$0.49842 \pm 1.96 \,(0.05660)$$

or $(0.387 \sim 0.609)$

the agreement between psychiatrist 1 and psychiatrist 2 is about 50% which according to Landis and Koch (1977) is moderate.

Table 7.42:
Percentage points of the χ^2-distribution
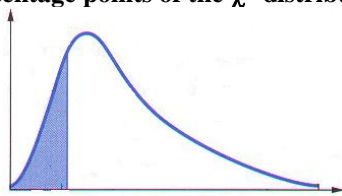
$P(\chi^2_{20} \leq 31.410) = .95$

df	$\chi^2_{.005}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.0000393	.000982	.00393	2.706	3.841	5.024	6.635	7.879
2	.0100	.0506	.103	4.605	5.991	7.378	9.210	10.597
3	.0717	.216	.352	6.251	7.815	9.348	11.345	12.838
4	.207	.484	.711	7.779	9.488	11.143	13.277	14.860
5	.412	.831	1.145	9.236	11.070	12.832	15.086	16.750
6	.676	1.237	1.635	10.645	12.592	14.449	16.812	18.548
7	.989	1.690	2.167	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	2.733	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	3.325	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	3.940	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	4.575	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	5.226	18.549	21.026	23.336	26.217	28.300
13	3.565	5.009	5.892	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	6.571	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	7.261	22.307	24.996	27.488	30.578	32.801
16	5.142	6.908	7.962	23.542	26.296	28.845	32.000	34.267
17	5.697	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	29.615	32.671	35.479	38.932	41.401
22	8.643	10.982	12.338	30.813	33.924	36.781	40.289	42.796
23	9.260	11.688	13.091	32.007	35.172	38.076	41.638	44.181
24	9.886	12.401	13.848	33.196	36.415	39.364	42.980	45.558
25	10.520	13.120	14.611	34.382	37.652	40.646	44.314	46.928
26	11.160	13.844	15.379	35.563	38.885	41.923	45.642	48.290
27	11.808	14.573	16.151	36.741	40.113	43.194	46.963	49.645
28	12.461	15.308	16.928	37.916	41.337	44.461	48.278	50.993
29	13.121	16.047	17.708	39.087	42.557	45.722	49.588	52.336
30	13.787	16.781	18.493	40.256	43.773	46.979	50.892	53.672
35	17.192	20.569	22.465	46.059	49.802	53.203	57.342	60.275
40	20.707	24.433	26.509	51.805	55.758	59.342	63.691	66.766
45	24.311	28.366	30.612	57.505	61.656	65.410	69.957	73.166
50	27.991	32.357	34.764	63.167	67.505	71.420	76.154	79.490
60	35.535	40.482	43.188	74.397	79.082	83.298	88.379	91.952
70	43.275	48.758	51.739	85.527	90.531	95.023	100.425	104.215
80	51.172	57.153	60.391	96.578	101.879	106.629	112.329	116.321
90	59.196	65.647	69.126	107.565	113.136	118.136	124.116	128.299
100	67.328	74.222	77.929	118.498	124.342	129.561	135.807	140.169

Chapter 8

Non-Parametric Tests

8.1 Introduction

The application of some parametric tests has been discussed in Chapter 4. It dealt with the comparison of means or proportions of two or more than two samples, paired or independent. This Chapter presents a number of alternative methods relating to the same problems when the conditions for parametric tests are not met. Suppose that a researcher wants to study the population and needs to draw inference about a measure of central tendency, i.e. mean, proportion, median *based on a small sample then h*e has to have the assumption of an approximately normal population needed to justify using a t-test for a hypothesis or construct confidence limits. In absence of this assumption, the t-test would be inappropriate and as such one would not apply the parametric tests. In this Chapter, we will study some statistical tests that may be used to draw inferences about the population when assumption of normality is not met. These include some of the statistical methods that are collectively referred to as *non-parametric methods or distribution free methods*.

These methods use, for example ranks of observations to perform tests rather than observations. Since these methods are using ranks rather than actual observations, the result obtained through these methods will not be as robust as by the methods used in Chapter 4. In brief, these methods are applied when; (i) data are in the form of ranks or the data are converted into ranks, and (ii) data do not satisfy the condition of normality.

Non-parametric tests are *distribution-free,* that is, they rely on very few assumptions about the probability distributions of sampled population. These methods are commonly used in medical and health sciences, as their samples are always small. Sometimes they are forced by the situations to take small samples because of non-availability of patients and expenditure involved. These methods are used, as they are relatively easy to apply as compared to the parametric tests.

One of the advantages of non-parametric statistical procedures is that they can be used with data that are based on a week measurement scale. These scales have been discussed in detail in Chapter 1.

Note: *We use the non-parametric tests if the measurement level of the dependent variable has either nominal or ordinal scale level, or if its measurement level is scale, but not drawn from a normal population specially for the case of small samples.*
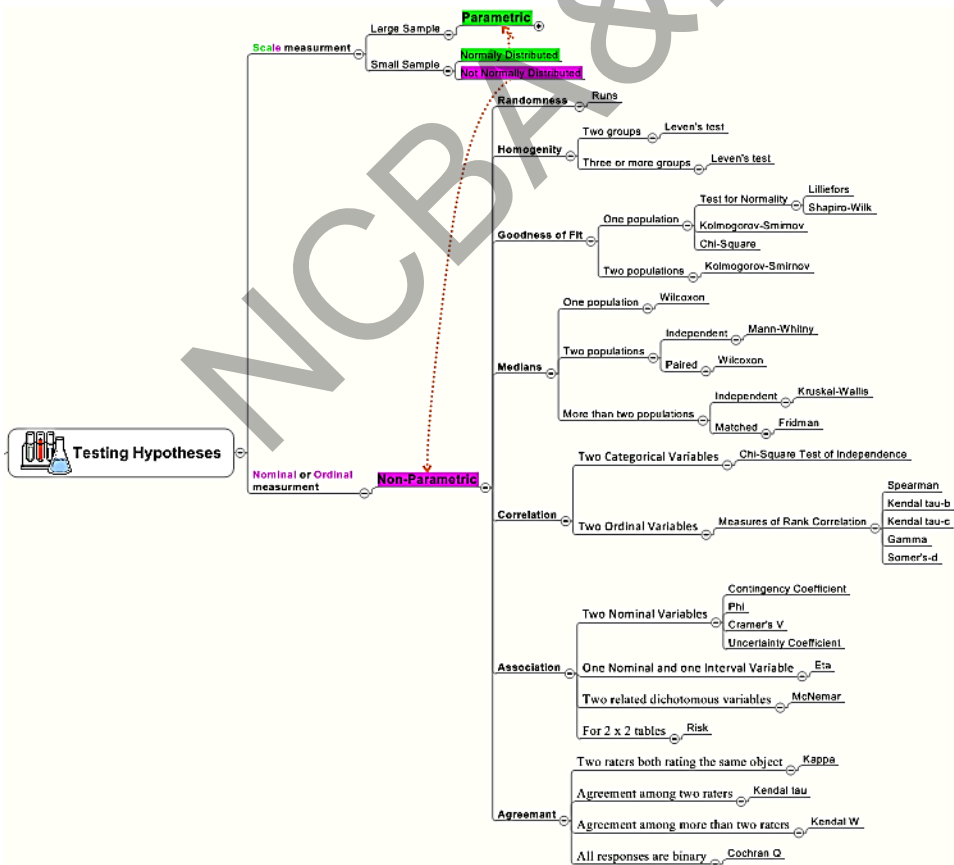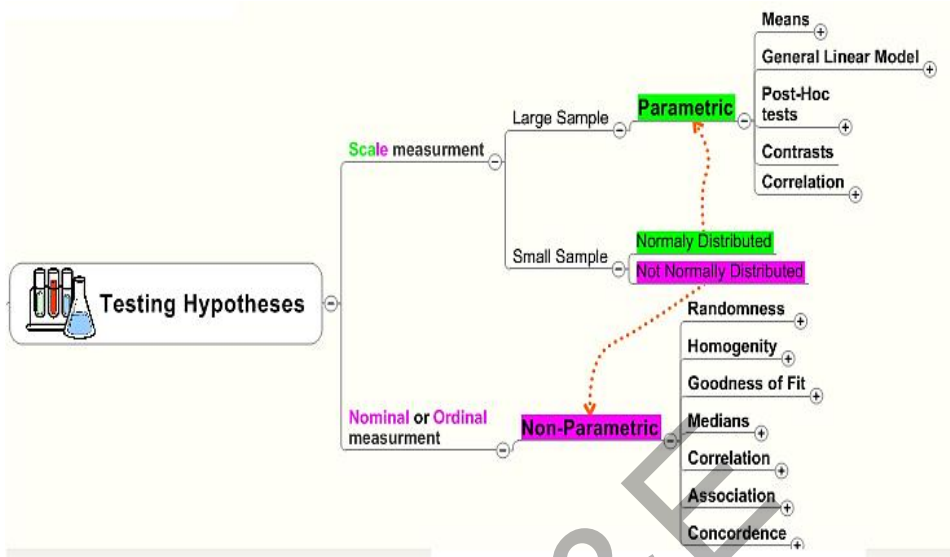
The following non-parametric tests are discussed in this Chapter.

(i)	The sign test (one sample)	(ii)	The sign test (two samples)
(iii)	The Wilcoxon signed-rank test	(iv)	McNemar test
(v)	The Wilcoxon rank-sum W-test	(vi)	Mann-Whitney U-test
(vii)	The Median test	(viii)	The Kruskal-Wallis H-test
(ix)	Fridman's test	(x)	Kendall's W-test and
(xi)	Cochran Q-test.	(xii)	Kolmogorov-Smirnov test

8.2 The Sign Test

When the population is non-normal and the size of the sample is less than 30, the t-test is not valid. We look for a non-parametric test. The simplest non-parametric test to apply in this situation is the sign test. This test is specifically designed for testing hypotheses about the median of any continuous population. Like mean, median is also a measure of central tendency, because of this the sign test is sometimes referred to as a *test for location*. The only assumption underlying the test is that the distribution of a variable of interest is continuous. The sign test gets its name from the fact that plus and minus signs, rather than numerical values, provide the raw data used in the calculations. Since the signs are either yes (+) or no (-), and trials are independent, the properties of a binomial experiment listed in Chapter 2 are satisfied. We use binomial probability table to calculate the p-value. The sign test is explained first for one sample then for paired observations (paired samples). The following points should be kept in mind while using the sign test?

(a) The sample is randomly selected from the population.

(b) If any sign is zero, it is ignored and the number of trials are counted on the basis of (+) and (-) signs only.

8.2.1 The Sign test for a single sample

Example 8.1:
The Environmental Protection Agency (EPA) sets certain pollution guidelines for major industries. For a particular company that discharges waste water into a nearby river, the EPA criterion is that the median amount of pollution in water from the river may not exceed 5 parts per million (ppm). Responding to numerous complaints, the EPA takes 10 water samples from the river at the discharge point and measures the pollution level in each sample. The results (in ppm) are as:

$$5.1, 4.3, 5.3, 6.2, 5.6, 4.7, 8.4, 5.9, 6.8, 3.0$$

Do the data provide sufficient evidence to indicate that median pollution level in water discharged at the plant exceeds 5 ppm? Use 5 percent level of significance.

Solution:

(i) H_0: M_0 (median) = 5

       H_1: $M_0 > 5$

(ii) $\alpha = 5\%$ (This is the one tailed test. Note that one-tailed and two-tailed tests have been explained in detail in Chapter 4).

(iii) Test-statistic: The sign test for a single sample:

       To apply the sign test, we calculate the scores above (+) and below (-) the specified value of the <u>median</u> (in our case it is 5).

Table 8.1

Epm	5.1	4.3	5.3	6.2	5.6	4.7	8.4	5.9	6.8	3
Score	+	-	+	+	+	-	+	+	+	-

It is expected that p(+) = p(-) = 0.5. In this example total number of (+) scores are 7 and (-) scores are 3. There is no zero, therefore, n = 10. Suppose one of the scores is zero then n will be 9 instead of 10. The p-value will be calculated by using the binomial probability table for (p = 0.5, n = 10, X ≥ 7).

Note: We calculate the probability (p-value) of the number of pluses or minuses that is larger than the observed pluses or minuses.

   p-value = $P(\geq 7)$ = $P(7)+P(8)+P(9)+P(10)$ = $1–P(\leq 6)$ = $1–0.8281 = 0.1719$.

We can also calculate it for number of minuses using binomial distribution as follows:

$$\binom{10}{x} p^X (1-p)^{10-x}, \text{ where } x = 0, 1, 2, 3 \text{ and } p = 0.5, \text{ then} \qquad (8.1)$$

$$\text{p-value} = \binom{10}{0}(0.5)^0(0.5)^{10-0} + \binom{10}{1}(0.5)^1(0.5)^9$$

$$+ \binom{10}{2}(0.5)^2(0.5)^8 + \binom{10}{3}(0.5)^3(0.5)^7 = 0.1719$$

or directly we see binomial table for p = 0.5, n = 10, X = 3, we get 0.1719.

(iv)   Stated p-value (α-value) = 0.05, observed p-value = 0.1719

Since observed (calculated p-value) is more than stated p-value, therefore, result is non-significant, we cannot reject the null hypothesis. (See the rule for rejection and acceptance of null hypothesis using p-value in Chapter 4). That is, there is insufficient evidence to indicate that median pollution level of water discharge from the plant exceeds 5 or the permissible level.

Like parametric test, it can be one-tailed or two-tailed test as:

One-tailed	Two-tailed
H_0: $M = M_0$	H_0: $M = M_0$
H_1: $M > M_0$ or $M < M_0$	H_1: $M \neq M_0$
Observed p-value	Observed p-value
= $P[X \geq$ number of "+" signs]	$2P[X \geq$ number of "+" signs]

The method of acceptance and rejection is as follows

(i)   Reject the null hypothesis if, p-value (observed p-value) < (stated p-value) = α.

(ii)   If n exceeds 10 then we may use test statistic:

$$Z = \frac{X - np}{\sqrt{npq}} ,$$ (8.2)

where X is the maximum number of "+" signs. Then, the null hypothesis is rejected on the basis of Z-value from the table. For example, in this case n = 10 then p = 0.5, np = 5 and $\sqrt{npq} = 1.58$, X = 7, then using (7.2) we get

$$Z = \frac{7-5}{0.5\sqrt{10}} = \frac{2}{1.58} = 1.27$$

where X is the number of sample observations that exceeds the median. In this case X=7. The p-value can be seen from the normal Table 2.6 given in Chapter-2, which is 0.102. This more than stated p-value, therefore, we cannot reject the hypothesis. As n increases, binomial distribution tends to normality. When p = 0.5 .The normal approximation performs reasonably well even for n as small as 10 if p is near $\frac{1}{2}$. Thus for n ≥ 10, we can conduct the sign test using the formula (8.2).

(iii)   For two-tailed test one may calculate the test statistic as either $x_1 =$ number of observations greater than M_0 for number of successes in n-trials. $x_2 =$ number of observations less than M_0, the number of failures in n-trials. Note that $x_1 + x_2 = n$.

Note: We can obtain the p-value for the Sign test through IBM-SPSS by one of two methods; 1st by the choice is manually through Legacy Dialogs while the 2nd method will be automatically which gives also the decision rule of rejecting or not rejecting the null hypothesis as follows:

Example S8-1

The data will be in columns as follows (we add a column for the median):

	EPA	Median
1	5.10	5.00
2	4.30	5.00
3	5.30	5.00
4	6.20	5.00
5	5.60	5.00
6	4.70	5.00
7	8.40	5.00
8	5.90	5.00
9	6.80	5.00
10	3.00	5.00

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
EPA	Numeric	8	2		None	None	8	Right	Scale	Input
Median	Numeric	8	2		None	None	8	Right	Scale	Input

A (the Sign test manually)

We apply the Sign test manually as follows:

Analyze→ Nonparametric Tests→ Legacy Dialogs→ 2 Related Samples …
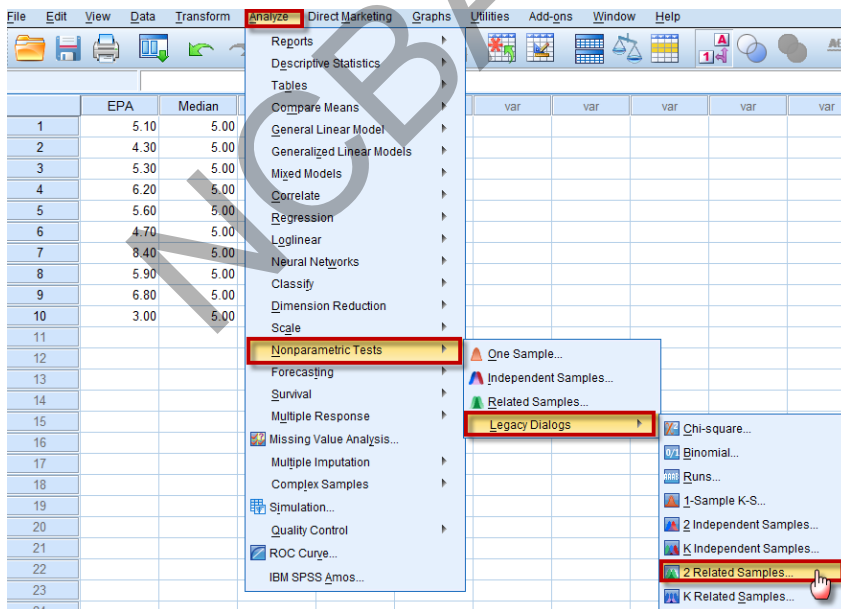
Move the variable "EPA" to Variable1:

Move the variable "Median" to Variable2:
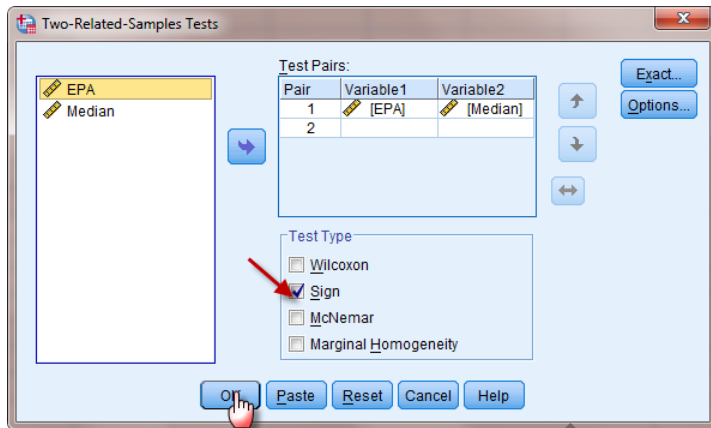
Now click on OK, to get the following output:

SPSS output for Sign test

Frequencies

		N
Median - EPA	Negative Differences[a]	7
	Positive Differences[b]	3
	Ties[c]	0
	Total	10

a. Median < EPA

b. Median > EPA

c. Median = EPA

Test Statistics[a]

	Median - EPA
Exact Sig. (2-tailed)	.344[b]

a. Sign Test

b. Binomial distribution used.

Note: The p-value for one tailed test will be 0.344/2 = 0.172, as given before.

B (the Sign test automatically)

We apply the Sign test automatically as follows:

Analyze→ Nonparametric Tests→ 2 Related Samples …
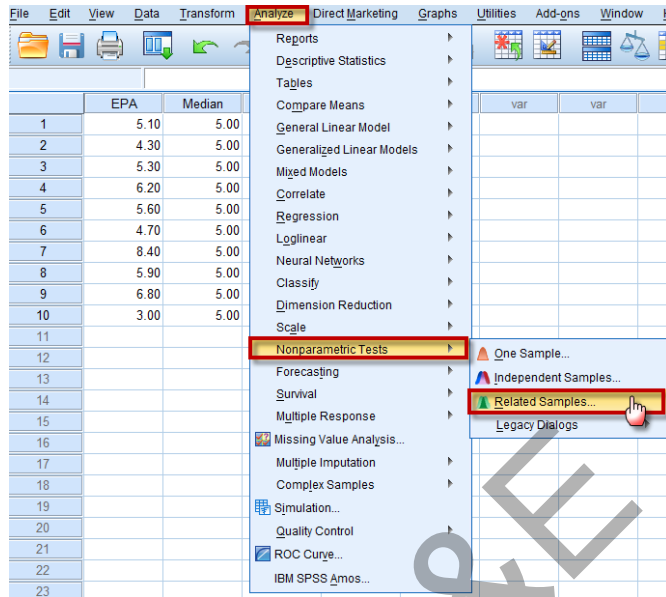
We may choose either **Automatically compare observed data to hypothesized** for a complete automation , Or **Customize analysis** , as follows:
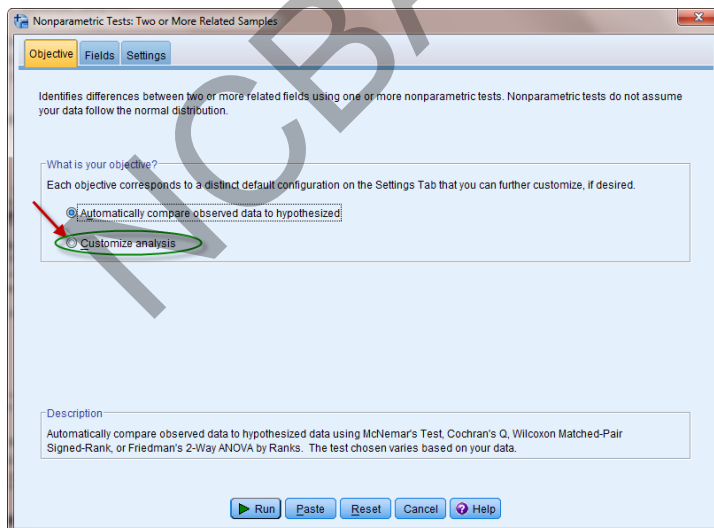
We choose the Sign test and click on **Fields** to move the variables:
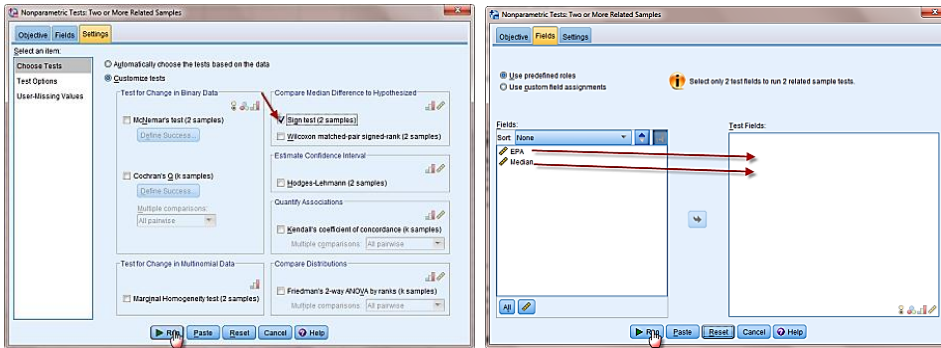
We click on to get the following final result:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between EPA and Median equals 0.	Related-Samples Sign Test	.344[1]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

[1]Exact significance is displayed for this test.

Note: We obtain the p-value and the decision rule of not rejecting (Retain) the null hypothesis.

8.2.2 The Sign test for samples of paired observation

The sign test may also be used with samples of paired observations in which each member of one sample is matched with a member of the other sample to form a sample of matched pairs. This is equivalent to t-test for paired observations.

Example 8.2:

A sample of 15 patients suffering from asthma participated in an experiment to study the effect of a new treatment on pulmonary function. Among various measurements recorded were those of forced expiratory volume (liters) in one second (FEV_1) before and after application of the treatment. The results are given in Table 8.2. Can we conclude that treatment is effective in increasing the FEV_1 level? Use 5% level of significance.

Table 8.2

Subject	Before	After	Subject	Before	After
1	1.69	1.69	9	2.58	2.44
2	2.77	2.22	10	1.84	4.17
3	1.00	3.07	11	1.89	2.42
4	1.66	3.35	12	1.91	2.94
5	3.00	3.00	13	1.75	3.04
6	0.85	2.74	14	2.46	4.62
7	1.42	3.61	15	2.35	4.42
8	2.82	5.14			

Solution:

(1) H_0 : Median (before) = Median (after)

   H_1 : Median (after) > Median (before)

(2) $\alpha = 0.05$

(3) Test-statistic: The sign test for paired observations.

   For the purpose of calculations, we proceed as follows:

Subject	Before	After	Before-After	Subject	Before	After	Before-After
1	1.69	1.69	0	9	2.58	2.44	+
2	2.77	2.22	+	10	1.84	4.17	-
3	1.00	3.07	-	11	1.89	2.42	-
4	1.66	3.35	-	12	1.91	2.94	-
5	3.00	3.00	0	13	1.75	3.04	-
6	0.85	2.74	-	14	2.46	4.62	-
7	1.42	3.61	-	15	2.35	4.42	-
8	2.82	5.14	-				

X_1 = total plus signs = 2; X_2 = total minus signs = 1 and there are two are zeros and zeros are ignored, therefore, n = 15 - 2 = 13. The p-value may be calculated using (8.1) when n = 13, and X = 2. The p = value

$$= P\,[X \le 2] = \binom{13}{0}(0.5)^0(0.5)^{13} + \binom{13}{1}(0.5)^1(0.5)^{13} + \binom{13}{2}(0.5)^2(0.5)^{13} = 0.0112$$

This p-value may directly be seen from binomial probability table when n = 13, p = 0.5 and X ≤ 2.

(4) The stated p-value is 0.05 (one-tailed test). The observed p-value is 0.0112 (calculated p-value). Since observed p-value is less than the stated p-value, we do not accept the hypothesis, therefore, new treatment is effective.

Since n = X_1 + X_2 is ≥ 10, therefore, the sign test can also be carried out using normal approximation to the binomial distribution, i.e. $\mu = np = 13 \times 0.5 = 6.5$ and $\sigma = \sqrt{13 \times 0.5 \times 0.5} = 1.80$

$$Z = \left| \frac{\left(x + \frac{1}{2}\right) - \frac{n}{2}}{0.5\sqrt{13}} \right| \qquad\qquad (8.3)$$

$$Z = \left| \frac{\left(2 + \frac{1}{2}\right) - \frac{13}{2}}{1.80} \right| \quad Z = \left| \frac{2.5 - 6.5}{1.80} \right| = 5.0$$

which is more than 1.64, therefore, we reject the null hypothesis, we say with 95% confidence that new treatment is effective.

The p-value may also be found using Z-table, p-value = 0.0091, which is less than 0.05, we confirm our previous result.

Another possible test to test the hypothesis P[+] = P[-] = 1/2 is the chi-square test. Given observed values, X_1 and X_2, the expected values are calculated as:

Observed	X_1	X_2
Expected	$\dfrac{X_1 + X_2}{2}$	$\dfrac{X_1 + X_2}{2}$

Now

$$\chi^2 = \frac{(X_1 - X_2)^2}{X_1 + X_2} , \tag{8.4}$$

where X_1 and X_2 represent the number of " +" and " –" signs. In this example $X_1 = 2$ and $X_2 = 11$, then chi-square will be

$$\chi^2 = \frac{(2-11)^2}{2+11} = \frac{81}{13} = 6.23$$

Since it is one-tailed test, the table value of chi-square for one degree of freedom is 5.024 (see Chapter-4). Therefore, we reject the null hypothesis and confirm our above findings. IBM-SPSS package has been used to solve this problem and the output has been given in the following example (using the automated way).

Example S8-2

The data will be in columns as follows:

serial	before	after
1	1.69	1.69
2	2.77	2.22
3	1.00	3.07
4	1.66	3.35
5	3.00	3.00
6	.85	2.74
7	1.42	3.61
8	2.82	5.14
9	2.58	2.44
10	1.84	4.17
11	1.89	2.42
12	1.91	2.94
13	1.75	3.04
14	2.46	4.62
15	2.35	4.42

The Variable View is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
serial	Numeric	8	0		None	None	8	Right	Nominal	Input
before	Numeric	8	2		None	None	8	Right	Scale	Input
after	Numeric	8	2		None	None	8	Right	Scale	Input

(The Sign test for samples of paired observation automatically)

We apply The Sign test for samples of paired observation automatically as follows:

Analyze→ Nonparametric Tests→ 2 Related Samples …
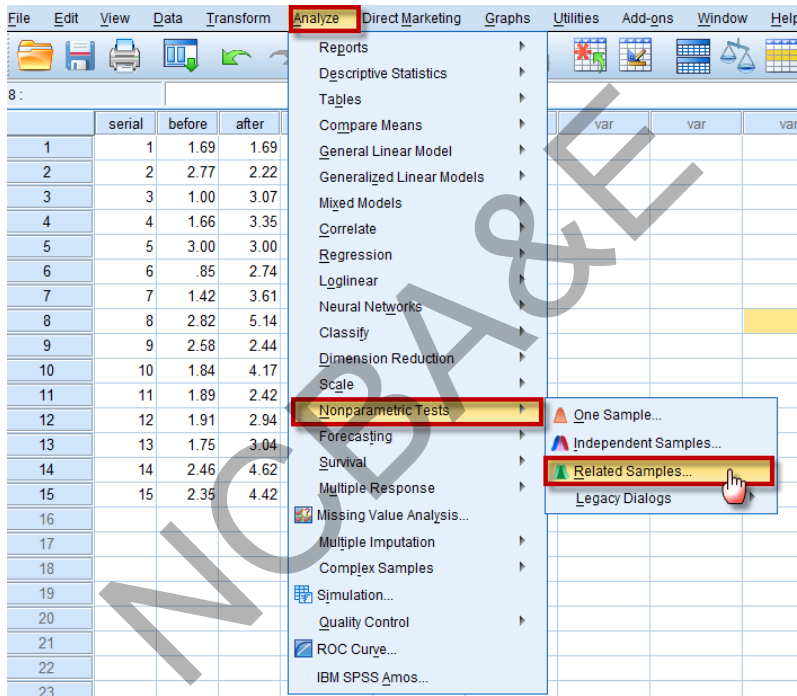
We will choose Customize analysis , as follows:
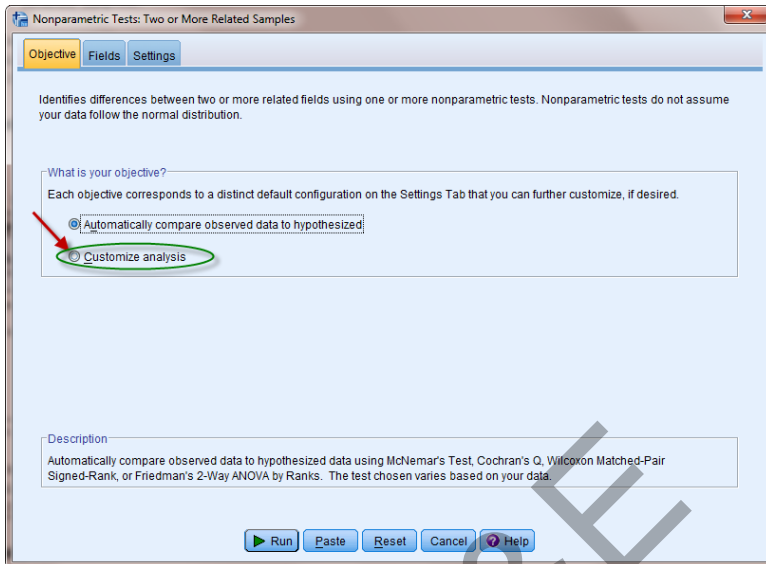
We choose the Sign test and click on Fields to move the variables:
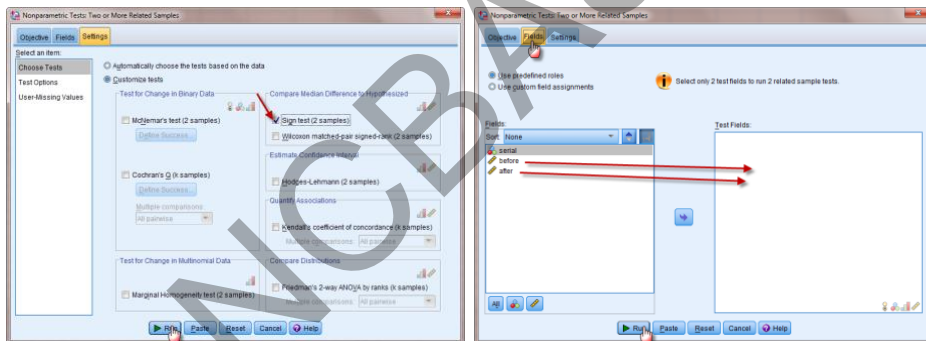
We click on ▶ Run to get the following final result:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between before and after equals 0.	Related-Samples Sign Test	.022¹	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

¹Exact significance is displayed for this test.

2 P-value

Note: The p-value for one tailed test will be $0.022/2 = 0.011$, and the decision rule is to reject the null hypothesis (as before).

8.3 The Wilcoxon Signed-rank test

The test is applied to paired observations when the condition of normality is not met. For the application of this test, we have random sample like all other non-parametric tests. The variable must be continuous. The measurement scale is *interval*. *This test is better than the sign test as the sign test completely ignores the magnitude of the differences between paired observations whereas this test takes into consideration this point.* The Wilcoxon signed-rank test for matched pairs for one-tailed and two-tailed tests is explained below:

Let X and Y represent the population variables then

	One-tailed test	**Two-tailed test**
1.	H_0 : X and Y are identical	H_0 : X and Y are identical
	H_1 : X is shifted to the left of X or Y is shifted to the left of X	H_1 : X shifted either to the right or to the left

2.   Calculate the difference between the n matched pairs of observations. Take absolute value of differences. Then rank the absolute values from the smallest to the highest. Attach sign to ranks based on the signs of differences.

3.   T(-) or T(+)   T, the smaller of T(-) or T(+)

4.   Rejection region
      $T(-) \leq T_0$ (table value)
      or $T(+) \leq T_0$ (table value) $T < T_0$ (table value)

5.   Note that zero is eliminated and matched pairs are counted without zero.

Example 8.3:
Use the data given in Example 8.2 (Table 8.2) and apply Wilcoxon -Signed-rank test to see whether the treatment is effective in existing the FEV_1 level?

Solution:
To solve this question follow these steps (table given below):

(i)   Take the differences between the paired observations i.e. y − x = d. These differences are calculated in column 4 of the above table.

(ii)   Take the absolute values of the differences (discard the algebraic sign). This is done in column 5 of the above table.

(iii) Assign the ranks to differences (as in column 6) assigning rank 1 to the smallest observed differences. If there is a tie then use the method of tied rank and ignore zero. This step is completed in column 7.

(iv)   Sum of positive ranks is 87 and sum of the negative ranks is 4.

(v)   The table against number of matched pairs 13 (excluding zeros), at 5% level of significance is 17.

1	2	3	4	5	6	Rank 7	
Subject	Before (x)	After (y)	d	|d|	Ranks	Positive ranks	Negative ranks
1	1.69	1.69	0.0	0	-	-	-
2	2.77	2.22	-0.55	0.55	3	-	-3
3	1.00	3.07	2.07	2.07	8.5	8.5	-
4	1.66	3.35	1.69	1.69	6	6	-
5	3.00	3.00	0.0	0.0	-	-	-
6	0.85	2.74	1.89	1.89	7	7	-
7	1.42	3.61	2.19	2.19	11	11	-
8	2.82	5.14	2.32	2.32	12	12	-
9	2.58	2.44	-0.14	0.14	1	-	-1
10	1.84	4.17	2.33	2.33	13	13	-
11	1.89	2.42	0.35	0.35	2	2	-
12	1.91	2.94	1.03	1.03	4	4	-
13	1.75	3.04	1.29	1.29	5	5	-
14	2.46	4.62	2.16	2.16	10	10	-
15	2.35	4.42	2.07	2.07	8.5	8.5	-
				20.08		87	-4

(vi) Reject H_0 if calculated value is less than table value. In this example calculated value is 4 which is smaller than 87 and 4, and table value is 17, so the null hypothesis is rejected and we say that the new treatment is better than the old one.

In IBM-SPSS package, the data are entered like t-test for paired observations. The difference between the calculation of these tests and t-test for paired observations is that in the former case we click non-parametric rather than click *compare means*. The IBM-SPSS package is used and the results (using **Analyze→ Nonparametric Tests→ Legacy Dialogs→ 2 Related Samples …**) are given as:

SPSS output for Wilcoxon Singed-Rank Test and the Sign Test

Ranks

		N	Mean Rank	Sum of Ranks
AFTER - BEFORE	Negative Ranks	2[a]	2.00	4.00
	Positive Ranks	11[b]	7.91	87.00
	Ties	2[c]		
	Total	15		

a. AFTER < BEFORE

b. AFTER > BEFORE

c. AFTER = BEFORE

Test Statistics[b]

	AFTER - BEFORE
Z	-2.901[a]
Asymp. Sig. (2-tailed)	.004

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

The calculated p-value is 0.008 for one tailed test, which is less than 0.05, therefore we do not accept the null hypothesis and say with 95% confidence treatment is effective.

Example S8-3

The data will be in columns as in example S8-2.

(The Wilcoxon signed-rank automatically)

We apply the Wilcoxon signed-rank for samples of paired observation automatically as follows:

Analyze→ Nonparametric Tests→ Related Samples …
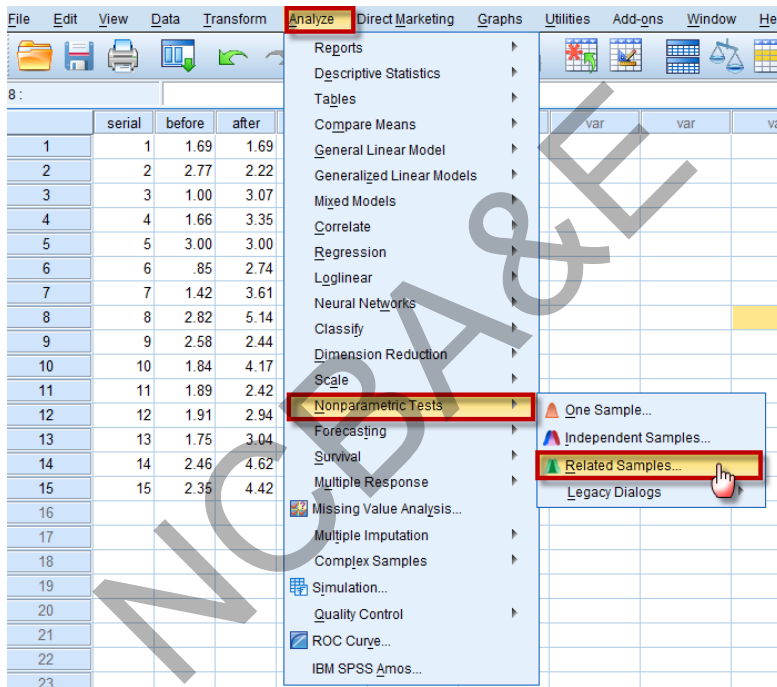
We will choose Customize analysis , as follows:
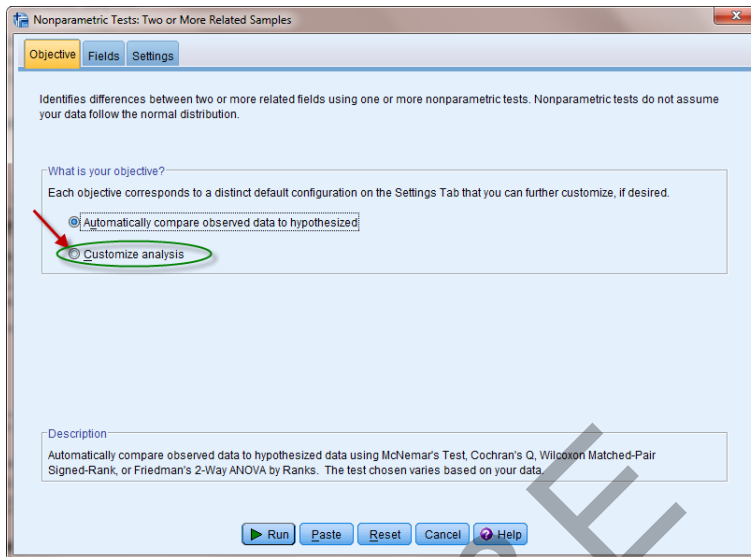
We choose the Sign test and click on **Fields** to move the variables:
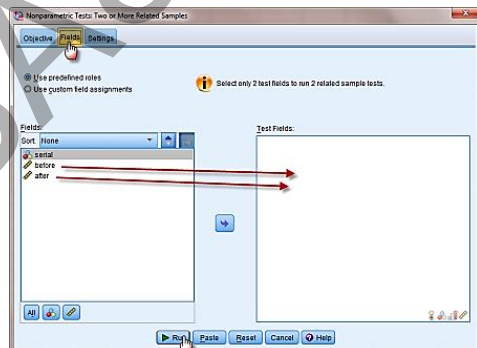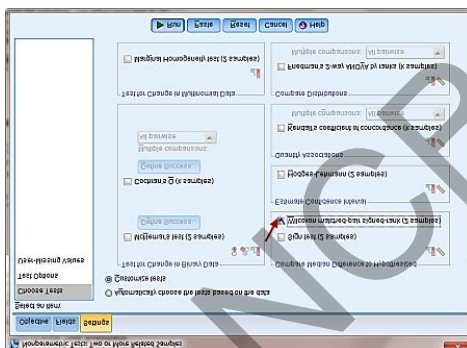
We click on ▶ Run to get the following final result:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between before and after equals 0.	Related-Samples Wilcoxon Signed Rank Test	.004	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Note: The p-value for one tailed test will be 0.004/2 = 0.002, and the decision rule is to reject the null hypothesis.

8.4 Test for Two Independent Samples

In Chapter 4, we have discussed t-test for two independent samples. When the conditions for t-test are not met then any one of the following alternative tests may be used.

 (i) The median test,   (ii) The Mann-Whitney test
 (iii) Wilcoxon test   (iv) Kolmogorov-Smirnov Test

8.4.1 The median test

The median test can be used for two or more than two independent samples to test whether two or more than two populations have the same median. This is a replacement for t-test for two independent samples and one way-ANOVA technique. For this, 2x2 or r x c contingency table is constructed. The number in each cell is the number that is below or above the median (the median of all observations in two or more than two samples). Commonly the median test is used for t-test for two independent samples. If samples are more than two then Kruskal-Wallis test is used. Kruskal-Wallis test will be discussed in next section.

Assumptions:

 (i)   Sample is a random sample

 (ii)   Samples are independent.

 (iii) The measurement scale is at least ordinal.

 (iv) If any cell has zero frequency, then this test cannot be used. The null and alternative hypotheses are

H_0 : two (or more) populations have the same median.

H_1 : at least two of the populations have different medians.

Example 8.4:

A study was conducted to compare the amount of time (in minutes) spent watching television each day by rural and urban elementary school children in Eastern Province of Saudi Arabia. Eight urban and nine rural children were randomly selected from elementary schools. The results are given in Table 8.3.

Table 8.3

Urban children	Rural children
60	140
240	80
190	45
75	210
30	120
150	135
220	30
190	120
	200

Is there any difference between two types of elementary school children in television viewing habits? Use 5% level of significance.

Solution:

(1) H_0 : The median time for two types of children is the same.

   H_1 : The median times are not equal.

(2) $\alpha = 0.05$

(3) Test-statistic: Since two samples are independent, one possible test is the median test. To apply median test, we proceed as:

   (i) Arrange the observations in order in the combined samples, i.e. 30, 30, 45, 60, 75, 80, 120, 120, 135, 140, 150, 190, 190, 200, 210, 220, 240. The median = 135.

   (ii) Prepare 2x2 contingency table as:

   If H_0 is true then the common median may be estimated from the combined sample this is precisely what the test does. Testing the equality of proportions can therefore test any difference in the Urban and rural pattern.

	Urban	Rural	Total
Above median	5 a	3 b	8
	c	d	
Below median	3	6	8
Total	8	9	17 = n

   Note that if any observation is equal to median, it may be ignored in analysis.

   (iii) Apply chi-square

$$\chi^2 = \frac{\left[|ad - bc| - \frac{n}{2} \right]^2 n}{(a + b)(c + d)(a + c)(b + d)} \qquad (8.5)$$

(number in the cells is less than 5)

$$\chi^2 = \frac{\left[|30 - 9| - \frac{17}{2} \right]^2 17}{9 \times 8 \times 8 \times 8} = 0.576$$

(4) The table value of chi-square for 5% level of significance is 5.024 which is more than calculated value, therefore, we say that there is no difference between two types of children regarding watching the television

SPSS package can be used and one can follow these steps:

   (i) Enter the data on SPSS package like t-test for two independent samples.

(ii)   Choose a non-parametric test.

(iii)   Choose "more than two independent samples".

(iv)   There are two tests:

        (a) | Kruskal Wallis | (b) | Median |

Choose either of them; you will get the same result.

It is advised that the median test should be used for two samples and the Kruskal-Wallis is to be used for more than two independent samples. The IBM-SPSS package is used and the result (using **Analyze→ Nonparametric Tests→ Legacy Dialogs→ 2 Independent Samples …**) is as follows:

SPSS output for Median test

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
TIME	17	131.47	69.39	30	240
CATEGORY	17	1.53	.51	1	2

Median Test

Frequencies

		CATEGORY	
		1 (Urban)	2 (Rural)
TIME	> Median	5	3
	<= Median	3	6

Test Statistics[a]

	TIME
N	17
Median	135.00
Exact Sig.	.347

a. Grouping Variable: CATEGORY

Calculated p = 0.347, which is more than 0.05, the result is non-significant. Therefore, there is no difference between two types of children belonging to urban and rural facilities of watching the television.

Example S8-4

(The Median test automatically)

The data will be in columns as follows:

	Group	time
1	1	60
2	1	240
3	1	190
4	1	75
5	1	30
6	1	150
7	1	220
8	1	190
9	2	140
10	2	80
11	2	45
12	2	210
13	2	120
14	2	135
15	2	30
16	2	120
17	2	200

	Group	time
1	Urban children	60
2	Urban children	240
3	Urban children	190
4	Urban children	75
5	Urban children	30
6	Urban children	150
7	Urban children	220
8	Urban children	190
9	Rural children	140
10	Rural children	80
11	Rural children	45
12	Rural children	210
13	Rural children	120
14	Rural children	135
15	Rural children	30
16	Rural children	120
17	Rural children	200
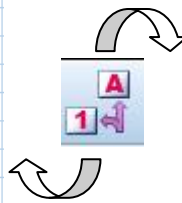
We apply the Wilcoxon signed-rank test for samples of paired observation automatically as follows:

Analyze→ Nonparametric Tests→ Independent Samples …

We may choose either ◎ Compare median$\underline{s}$ across groups , or Customize analysis . Both will give the same result. Choosing ◎ Compare median$\underline{s}$ across groups will give the result of the median test directly. We will choose Customize analysis , as follows:
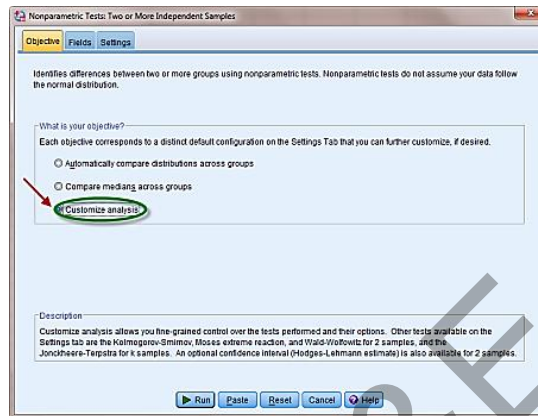
We choose the Median test and click on Fields to move the variables:

We click on ▶ Run to get the following final result:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The medians of amount of time (in minutes) are the same across categories of Group.	Independent-Samples Median Test	.347[1,2]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

[1]Exact significance is displayed for this test.

[2]Fisher Exact Sig.

2 P-value

Note: The p-value for two tailed test is 0.347, and the decision rule is to reject the null hypothesis.

8.4.2 The Mann-Whitney and Wilcoxon Rank sum-W tests

Two tests are given in this section.

(a) The Mann-Whitney test

This test is based on two independent random samples.

Assumptions

   (i) These samples are random and independent.

   (ii) The measurement scale is at least ordinal.

Example 8.5:

In a controlled environment laboratory, 10 men and 10 women were tested to determine the room temperature (in Fahrenheit) they found to be the most comfortable. The results are given in Table 8.4:

Table 8.4

Men	74	72	77	76	76	73	75	73	74	75
Women	75	77	78	79	77	73	78	79	78	80

Assuming that these temperatures resemble a random sample from their respective populations. Is the average comfortable temperature the same for men and women? Use 5% level of significance.

Solution:

(1) H_0 : The average (median) comfortable temperature for men and women is the same.

   H_1 : The average comfortable temperature is not the same.

(2) $\alpha = 0.05$

(3) test-statistic: Mann-Whitney test

   To apply the Mann-Whitney test, we will proceed as:

   (i)   Arrange the observations of two samples together in ascending order, like the Median test, i.e. 72, 73, 73, 73, 74, 74, 75, 75, 75, 76, 76, 77, 77, 77, 78, 78, 78, 79, 79, 80

   (ii)   Rank these observations as:

      1, 3, 3, 3, 5.5, 5.5, 8, 8, 8, 10.5, 10.5, 13, 13, 13, 16, 16, 16, 18.5, 18.5, 20.

   (iii) R_1 (sum of the ranks of first sample)
         $= 5.5 + 1 + 13 + 10.5 + 10.5 + 3 + 8 + 3 + 5.5 + 8.5 = 68.5$.

      R_2 (sum of the ranks of second sample)

      $= 8.5 + 13 + 16 + 19 + 13 + 16 + 3 + 16 + 19 + 20 = 143.5$.

   (iv)   Calculate:

$$\mu_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1, \quad \mu_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \qquad (8.6)$$

$$\mu_1 = 10 \times 10 + 55 - 68.5 = 86.5, \quad \mu_2 = 10 \times 10 + 55 - 143.5 = 9.5$$

    (v)   Take the smaller value, which is 9.5.

(4) The table value for 10 by 10 at 5% level of significance is 28.

(5) Our calculated value 9.5 does not fall in the acceptance region, therefore, the average comfortable temperatures for the men and women are not equal.

IBM-SPSS package is used for Mann-Whitney U-Test and Wilcoxon Rank Sum W-Test and the result (using **Analyze→ Nonparametric Tests→ Legacy Dialogs→ 2 Independent Samples …**) is as follows:

SPSS output for Mann-Whitney U-Test and Wilcoxon Rank Sum W-Test

Test Statistics[b]

	temprature
Mann-Whitney U	13.000
Wilcoxon W	68.000
Z	-2.817
Asymp. Sig. (2-tailed)	.005
Exact Sig. [2*(1-tailed Sig.)]	.004[a]

a. Not corrected for ties.

b. Grouping Variable: CATEGORY

Median test is not as robust as Man-Whitney test because median test loses information of equal ranks whereas Man-Whitney use these information.

IBM-SPSS package is used for Two samples Kolmogorov-Smirnov test and the result (using **Analyze→ Nonparametric Tests→ Legacy Dialogs→ 2 Independent Samples …**) is as follows:

Two-Sample Kolmogorov-Smirnov Test

Test Statistics[a]

		temprature
Most Extreme Differences	Absolute	.700
	Positive	.700
	Negative	.000
Kolmogorov-Smirnov Z		1.565
Asymp. Sig. (2-tailed)		.015

a. Grouping Variable: CATEGORY

The p-value for this test is 0.015, which is less than 0.05; hence we confirm our previous findings.

Example S8-5

(The Mann-Whitney U-test automatically)

The data for example 8.5 will be in columns as follows:

	Group	Temperature
1	1	74
2	1	72
3	1	77
4	1	76
5	1	76
6	1	73
7	1	75
8	1	73
9	1	74
10	1	75
11	2	75
12	2	77
13	2	78
14	2	79
15	2	77
16	2	73
17	2	78
18	2	79
19	2	78
20	2	80

	Group	Temperature
1	Men	74
2	Men	72
3	Men	77
4	Men	76
5	Men	76
6	Men	73
7	Men	75
8	Men	73
9	Men	74
10	Men	75
11	Women	75
12	Women	77
13	Women	78
14	Women	79
15	Women	77
16	Women	73
17	Women	78
18	Women	79
19	Women	78
20	Women	80

We apply the Mann-Whitney U-test for samples of paired observation automatically as follows:

Analyze→ Nonparametric Tests→ Independent Samples …

We may choose either ◎ Compare medians across groups , or Customize analysis . Both

will give the same result. Choosing ◎ Compare medians across groups will give the result

of the median test directly. We will choose Customize analysis , as follows:
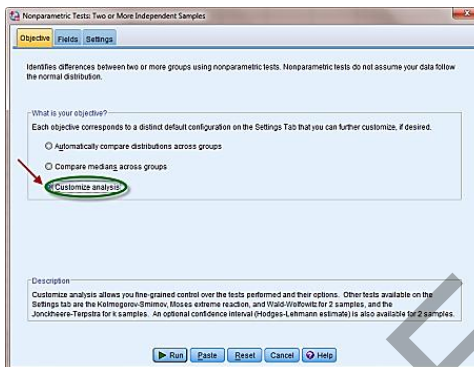
We choose the Median test and click on Fields to move the variables:

We click on ▶ Run to get the following final result:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Temperature is the same across categories of Group.	Independent-Samples Mann-Whitney U Test	.004[1]	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

[1]Exact significance is displayed for this test.

2 P-value

Note: The decision rule is to reject the null hypothesis.

(b) The Wilcoxon Rank-sum-W test

This test is based on two independent random samples. The Mann-Whitney and Wilcoxon-Rank sum tests are identical. Any one of the tests can be applied.

Assumptions

The assumptions of this test are the same as in case of the Mann-Whitney test.

Example 8.6:

A preliminary study was conducted to obtain information on the background levels of the toxic substance polychlorinated biphenyl (PCB) in soil sample in the United Kingdom. Such information could then be used as a benchmark against which PCB levels at waste disposal facilities in the United Kingdom can be compared. Table 8.5 contains the measured PCB levels of soil samples taken at 14 rural and 15 urban locations in the United Kingdom. (PCB concentration is measured in 0.0001 gram per kilogram of soil). From these preliminary results, the researchers reported "a significant difference between (PCB levels) for rural areas and for urban areas". Do the data support the researcher's conclusion regarding significance difference? Test using 5% level of significance (source: *Chemosphere*, Feb. 1986).

Solution:

(1) H_0 : There is no difference in PCB levels in two areas.

   H_1 : There is a difference in PCB levels in two areas.

(2) $\alpha = 0.05$

(3) test-statistic: Three possible tests can be used.

   (i) The Median test, (ii) The Mann-Whitney test, and (iii) Wilcoxon rank sum-W test

Table 8.5

Rural	R_1	Urban	R_2
5.3	5.5	24	24.0
8.1	7.0	29	25.0
1.8	4.0	16	18.0
9.0	9.0	21	21.0
1.6	3.0	107	28.0
23.0	23.0	94	27.0
1.5	2.0	141	29.0
5.3	5.5	11	12.5
9.8	11.0	11	12.5
15.0	17.0	49	26.0
12.0	14.5	22	22.0
8.2	8.0	13	16.0
9.7	10.0	18	19.5
1.0	1.0	12	14.5
		18	19.5
	$120.5 = T_1$		$314.5 = T_2$

The Median test and Mann-Whitney test have been explained before. Now, here we demonstrate the application of Wilcoxon rank sum test. The test-statistics is

$$T = S - \frac{n(n+1)}{2}, \qquad (8.7)$$

where S is the smaller sum of the ranks of the rural and urban areas.

(i) Rank the rural (sample-1) and urban (sample-2), considering it as one sample. This has been done in the above table.

(ii) Add the ranks for each sample.

$T_1 = 120.5$ and $T_2 = 314.5$

(iii) $S = $ Smaller $\{T_1, T_2\}$. Since T_1 is less than T_2 then, we calculate the test-statistic using n=14, $S = T_1$

$$T = 120.5 - \frac{14(14+1)}{2} = 15.5$$

(4) Table value for Wilcoxon Rank against $n_1 = 14$ and $n_2 = 15$ for 5% level of significance is 67.

Since calculated value of 15.5 is less than the table value, therefore, we do not accept the null hypothesis and say with 95% confidence that there is a significance difference between PCB levels for rural and urban areas. The IBM-SPSS package is used and the results for Mann-Whitney U and Wilcoxon Rank sum tests and the result (using **Analyze→ Nonparametric Tests→ Legacy Dialogs→ 2 Independent Samples ...**) is as follows:

SPSS output for Mann-Whitney U test and Wilcoxon Rank-Sum W test

Test Statistics[b]

	PCB
Mann-Whitney U	15.500
Wilcoxon W	120.500
Z	-3.908
Asy mp. Sig. (2-tailed)	.000
Exact Sig. [2*(1-tailed Sig.)]	.000[a]

a. Not corrected f or ties.

b. Grouping Variable: CATEGORY

Test Statistics^a

		PCB
Most Extreme Differences	Absolute	.786
	Positive	.000
	Negative	-.786
Kolmogorov -Smirnov Z		2.114
Asymp. Sig. (2-tailed)		.000

a. Grouping Variable: CATEGORY

Observed p-value < 0.000, which is less than stated p-value (0.05), therefore, we confirm our above findings.

For the Median test SPSS output is as:

Median Test

Frequencies

		CATEGORY	
		1(rural)	2(urban)
PCB	> Median	2	12
	<= Median	12	3

Test Statistics^a

	PCB
N	29
Median	12.0000
Exact Sig.	.001

a. Grouping Variable: CATEGORY

p-value is less than 0.05 (observed p-value), again we confirm the previous findings.

8.5 Test for K-Independent Samples

There are two possible tests that can be used for K-independent samples.

(a) The Median test
(b) The Kruskal-Wallis-H test

The Median test has already been explained for two samples in section 8.4, here SPSS package will also be applied to more than two samples. We describe the Kruskal-Wallis-H test in details first and application of the Median test later on in this section.

8.5.1 The Kruskal-Wallis test (or H-test)

The Kruskal-Wallis test provides a non-parametric alternative to the one-way ANOVA for comparing more than two independent samples. Like Median test, the Mann-Whitney

test and Wilcoxon test, no assumption regarding the normality or equality of variances of sampled populations is required.

Assumptions

   (i)   The K-samples are randomly and independently selected from their respective populations.

   (ii)   In addition to randomness within each sample, there is mutual independence among various samples.

   (iii) The measurement scale is *ordinal*.

   (iv)   For the chi-square approximation to be adequate, there should be five or more observations in each sample.

Following rules must be taken into consideration to see the significance of the Kruskal-Wallis test.

   (i)   If there are two or three groups, all groups are 5 or less in size and there are no ties, ties determine the significance of computed table.

   (ii)   If there are three groups and number of observations in each group are five or more consult chi-square table.

   (iii) If there are four or more groups, consult chi-square table for the significance of the result.

Example 8.7:

Vanadium is recently recognized essential trace element. An experiment was conducted to compare the concentration of vanadium in biological materials using isotope dilution mass spectrometry. The following table gives the quantities of vanadium (measured in nanograms per gram) in dried samples of oyster tissue, citrus leaves, and bovine liver and human serum. Use an appropriate method of analysis to determine whether the distribution of vanadium concentrations for the four biological materials differ in locations. The data is given in Table 8.6. Use 5% level of significance.

Table 8.6

Oyster tissue	Ranks	Citrus tissue	Ranks	Bovine lever	Ranks	Human serum	Ranks
						0.10	1
2.35	12	2.32	11	0.39	8	0.17	5
1.30	10	3.07	13	0.54	9	0.14	2
0.34	7	4.09	14	0.30	6	0.16	3.5
						0.16	3.5
Total	$T_1 = 29$		$T_2 = 38$		$T_3 = 23$		$T_4 = 15$

(Source: Analytical chemistry, Vol. 57(13), 1985, pp. 2475).

Solution:

(1) H_0 : There is no difference between the Vanadium concentrations for the four biological materials. (Population distributions are all identical).

   H_1 : They are different.

(2) $\alpha = 0.05$

(3) Test-statistic: Since there are more than two independent samples, therefore, the Kruskal-Wallis (H) test is used. We proceed as follows:

   (i) *Rank all the observations as if it were a one sample.* This is done in the above table.

   (ii) The sum of the ranks in each sample is also given.

   $$T_1 = 29, \ T_2 = 38, \ T_3 = 23 \text{ and } T_4 = 15$$

   (iii) test-statistic

   $$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3\,(n+1) , \qquad (8.8)$$

   where:

   $n = n_1 + n_2 + n_3 + n_4 = 3 + 3 + 3 + 5 = 14$

   k = number of groups = 4

   T_i = sum of the ranks in the ith group ($T_1 = 29$, $T_2 = 38$, $T_3 = 23$, $T_4 = 15$.

   $$H = \frac{12}{14 \times 15} \left(\frac{841}{3} + \frac{1444}{3} + \frac{529}{3} + \frac{225}{5} \right) - 3\,(14 + 1) = 11.17$$

(4) Rejection region is calculated as:

   There are $k = 4$ samples. The degrees of freedom are $k - 1 = 3$.
   The table value of chi-square for 5% level of significance is 9.348.
   (See the χ^2-Table 7.1, Chapter 7).

(5) The calculated value is more than the table value so we do not accept the null hypothesis and say that there is difference between the vanadium concentrations for four biological materials, or we say that populations are not identical.

Note that the entry of data in SPSS package is like one-way ANOVA and we click non-parametric methods for K-independent samples. The following methods appear on monitor:

(i) The Kruskal-Wallis   (ii) The Median

We choose one of them. If any cell is zero, the Median test fails. The Kruskal-Wallis test is usually more powerful than the Median test The IBM-SPSS package is used and (using **Analyze→ Nonparametric Tests→ Legacy Dialogs→ K Independent Samples …**) is as follows:

SPSS output for the Kruskal-Wallis and the Median tests

	CATEGORY	N	Mean Rank
Concentration	1 (Oyster)	3	9.50
	2(citrus)	3	12.67
	3(bovine)	3	7.83
	4(Human)	5	3.00
	Total	14	

Test Statistics[a,b]

	Concentration
Chi-Square	11.116
df	3
Asymp. Sig.	.011

a. Kruskal Wallis Test

b. Grouping Variable: CATEGORY

(a) p-value = 0.011 which is less than 0.05, we reject the null hypothesis and confirm our above findings.

(b) We apply the median test, as the frequencies in the cell are less than 5 and two cells have zero frequency.

So far, we have seen only one picture of the application of the Kruskal-Wallis H-test, which is a substitute of one way-ANOVA. There are recent advances in the theory of rank tests. There should no longer be any hesitation in applying the rank test to situations that have many ties. In fact Kruskal-Wallis H-test also gives an excellent performance in contingency table, where rows represent ordered category (rows are ordinal) and columns represent different populations (columns are nominal).

The IBM-SPSS Package is used to apply Median test for the Example 8.8 and the output (using **Analyze→ Nonparametric Tests→ Legacy Dialogs→ K Independent Samples …**) is as follows:

SPSS Output for Median Test

		identification			
		1.00	2.00	3.00	4.00
concentration	> Median	2	3	2	0
	<= Median	1	0	1	5

Test Statistics^b

	concentration
N	14
Median	.3650
Chi-Square	8.667^a
df	3
Asymp. Sig.	.034

a. 8 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1.5

b. Grouping Variable: identification

p-value for median test is 0.034 which is less than 0.05, we can conclude at 5% level of significance that there is significant difference in concentration of different categories.

Example S8-6

(The Kruskal-Wallis H-test automatically)

The data for example 8.7 will be in columns as follows:

	Group	vanadium
1	1	2.35
2	1	1.30
3	1	.34
4	2	2.32
5	2	3.07
6	2	4.09
7	3	.39
8	3	.54
9	3	.30
10	4	.10
11	4	.17
12	4	.14
13	4	.16
14	4	.16

	Group	vanadium
1	Oyster tissue	2.35
2	Oyster tissue	1.30
3	Oyster tissue	.34
4	Citrus tissue	2.32
5	Citrus tissue	3.07
6	Citrus tissue	4.09
7	Bovine lever	.39
8	Bovine lever	.54
9	Bovine lever	.30
10	Human serum	.10
11	Human serum	.17
12	Human serum	.14
13	Human serum	.16
14	Human serum	.16

We apply the Kruskal-Wallis H-test for independent samples automatically as follows:
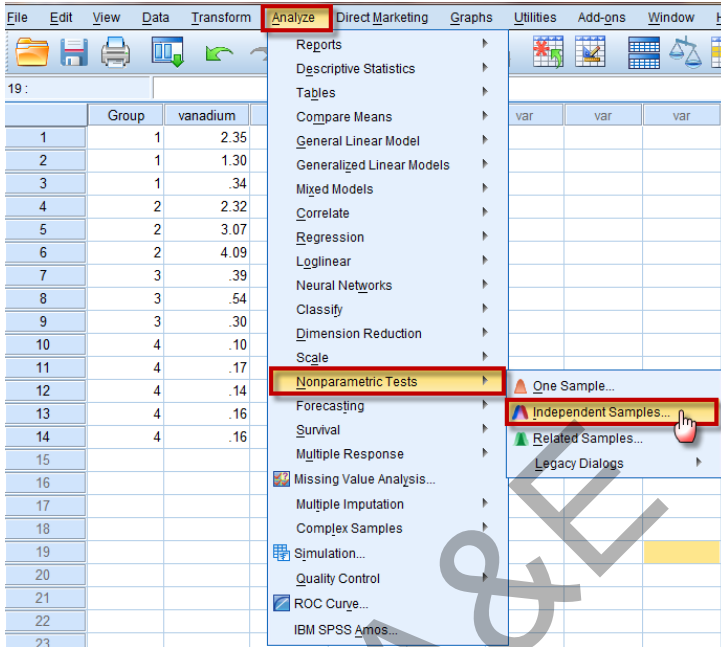
Analyze→ Nonparametric Tests→ Independent Samples …

We may choose either ⊙ Compare medians across groups , or Customize analysis . Both will give the same result. Choosing ⊙ Compare medians across groups will give the result of the Kruskal-Wallis H-test directly. We will choose Customize analysis , as follows:
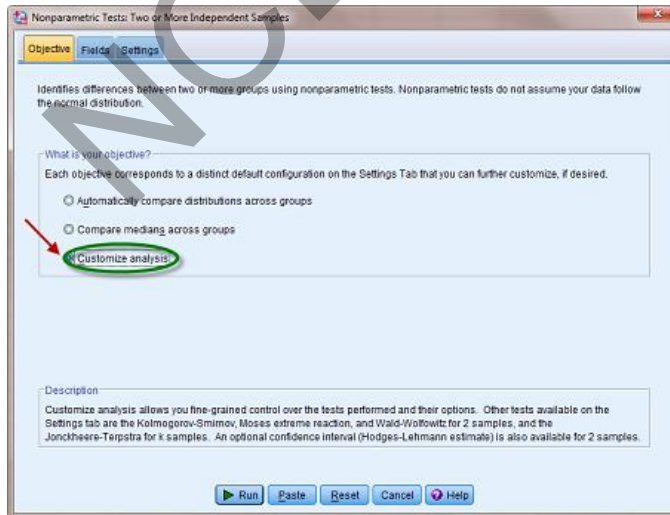
We choose the Kruskal-Wallis H-test and click on Fields to move the variables:
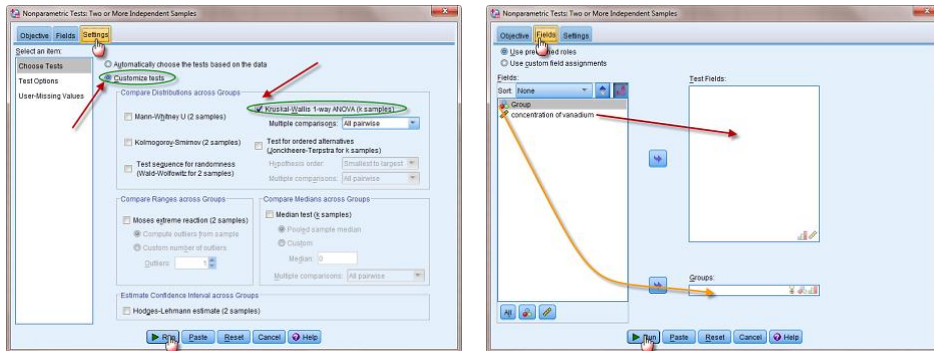
We click on ▶ Run to get the following final result:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of concentration of vanadium is the same across categories of Group.	Independent-Samples Kruskal-Wallis Test	.011	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

P-value

Note: The decision rule is to reject the null hypothesis. Double click on the output will yield the following comparisons:

Independent-Samples Kruskal-Wallis Test
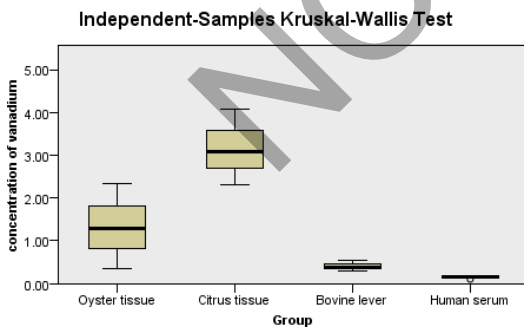
Total N	14
Test Statistic	11.196
Degrees of Freedom	3
Asymptotic Sig. (2-sided test)	.011

1. The test statistic is adjusted for ties.

Example 8.8:
Three instructors gave the grades to students. They assigned scores over the past semester and to see if some of them tend to give lower grades than others. The data is given below:

Table 8.7

Grades	Instructors		
	I_1	I_2	I_3
A	4	10	6
B	14	6	7
C	17	9	8
D	6	7	6
E	2	6	1

Can we say at 5% level of significance that three instructors graded evenly with each other?

Solution:

(i) H_0 : There is no difference in 3 instructors in grading the students.

         H_1 : At least two differ.

(ii) $\alpha = 0.05$

(iii)   test statistics : Kruskal-Wallis, using (8.8), we have $H = 0.845$

(iv)   p-value = 0.6447, which is greater than 0.05. Therefore, there is no difference in these instructors in assigning the grades.

The IBM-SPSS package is used and the result is:

SPSS output for the Kruskal-Wallis H Method

Test Statistics[a,b]

	score
Chi-Square	.878
df	2
Asymp. Sig.	.645

a. Kruskal Wallis Test

b. Grouping Variable: CATEGORY

Example 8.9:

A simple random sampling procedure was used to select 5 primary health care centers out of 9 from Al-Khobar area. The data regarding lab utilization are given as:

Table 8.8
Primary health care centers

Utilization	1	2	3	4	5	Total
Over	18	4	15	21	29	87
Proper	48	51	44	103	77	323
Under	49	47	22	75	56	249
Total	115	102	81	199	162	659

Use proper method of analysis to the data and to see the difference, if any, between primary health care centers regarding laboratory utilization.

Solution:

   (i) H_0 : There is no difference in lab utilization

          H_1 : At least two differ.

   (ii) $\alpha = 0.05$

   (iii)   test statistics : Kruskal-Wallis
          (as rows are ordinal and columns are nominal)

   (iv)   The IBM-SPSS package is used and output is given as:

SPSS output for Kruskal-Wallis H-test

Test Statistics[a,b]

	Lab Utilization
Chi-Square	4.133
df	4
Asymp. Sig.	.388

a. Kruskal Wallis Test

b. Grouping Variable: CATEGORY

Calculated p-value $= 0.388$, we say that there is no difference in all PHC centers in utilization of laboratory facilities.

8.6 K-Related Samples

In Section 8.2, we have discussed two tests for related samples but in this section, we consider some tests for more than two related samples. These are:

   (i)   The Friedman test

   (ii)   Kendall's coefficient of concordance (Kendall's W-test)

   (iii)   Cochran's test

8.6.1 The Friedman test

It is an extension of sign test for two related samples. This is a better-known test for the experimental situation, but it has less power in some situations. The test is appropriate whenever the data are measured on ordinal scale and can be meaningfully arranged in a two-way ANOVA classification. The problem of several related samples arises in an experiment that is designed to detect differences in k possibly different treatment $(k \geq 2)$. The observations are arranged in blocks, which are groups of k experimental units.

Assumptions

  (i) The variables are mutually independent.

 (ii) Within each block the observations may be ranked according to some criterion of interest.

    (1) H_0 : The k-populations are identical.

      H_1 : At least two of the k-populations are different.

    (2) $\alpha = 0.05$

    (3) test-statistic:

$$\chi_F^2 \ = \ \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 \ - \ 3n(k+1) , \tag{8.9}$$

    where k = number of samples or treatments

      n = number of blocks

      R_i = sum of the ranks for the ith treatment

Example 8.10:

There are three observers who assess a total of 10 patients for some attributes, say, sadness on a 10-point scale. Their scores are shown on Table 8.9:

Table 8.9

Patients	Observer 1	Observer 2	Observer 3
1	6	7	8
2	4	5	6
3	2	2	2
4	3	4	5
5	5	4	6
6	8	9	10
7	5	7	9
8	6	7	8
9	4	6	8
10	7	9	8

Can we say that there is a difference in three observers in assessing the sadness on 10-point scale? Use 5% level of significance.

Solution:

We proceed as;

(i)   Rank the observations according to rows as in the following table:

Patients	Observer 1	Ranks	Observer 2	Ranks	Observer 3	Ranks
1	6	1	7	2	8	3
2	4	1	5	2	6	3
3	2	2	2	2	2	2
4	3	1	4	2	5	3
5	5	2	4	1	6	3
6	8	1	9	2	10	3
7	5	1	7	2	9	3
8	6	1	7	2	8	3
9	4	1	6	2	8	3
10	7	1	9	3	8	2
Sum		$R_1 = 12$		$R_2 = 20$		$R_3 = 28$

(ii)   Sum the ranks in each column and calculate

$$\chi_F^2 = \frac{12}{10 \times 3(3+1)} [12^2 + 20^2 + 28^2] - 3(10)(3+1) = 12.8$$

(iii)   The table value of chi-square for 2 degree of freedom at 5% level of significance is 3.841.

(iv)   The calculated value of χ^2 is much greater than the table value, therefore, we reject the null hypothesis and say that the observers are different in assessing the sadness rank on 10-point scale.

Example S8-7

(The Friedman test automatically)

The data for example 8.9 will be in columns as follows:

	Patients	Observer1	Observer2	Observer3
1	1	6	7	8
2	2	4	5	6
3	3	2	2	2
4	4	3	4	5
5	5	5	4	6
6	6	8	9	10
7	7	5	7	9
8	8	6	7	8
9	9	4	6	8
10	10	7	9	8

We apply the Friedman test for related samples automatically as follows:
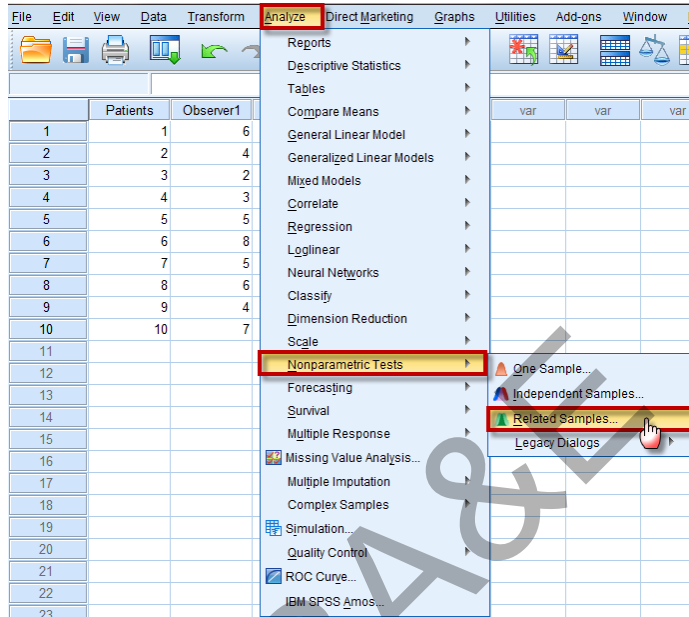
Analyze→ Nonparametric Tests→ Related Samples …

We may choose either ⊙ Automatically compare observed data to hypothesized , or Customize analysis
to choose Friedman test manually. Both will give the same result. Choosing
⊙ Automatically compare observed data to hypothesized will give the result of the Friedman test
directly, as follows:

We click Fields to move the variables, and then we click on Objective and choose
⊙ Automatically compare observed data to hypothesized :
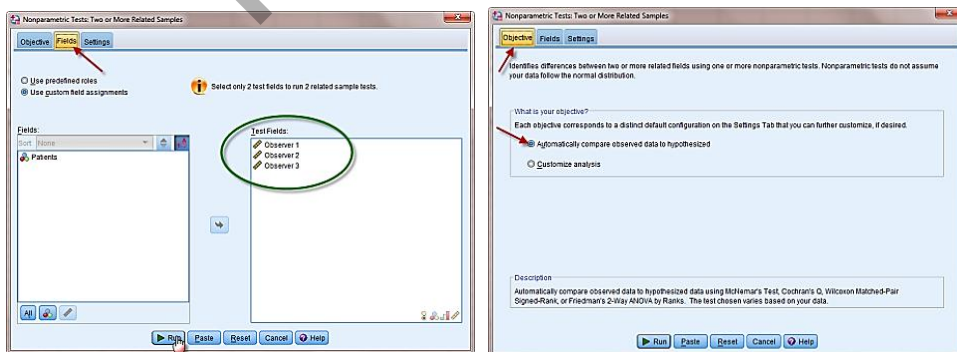
We click on ▶ Run to get the following final result:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Observer 1, Observer 2 and Observer 3 are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.001	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05. **P-value**

Note: The decision rule is to reject the null hypothesis. Double click on the output will yield the following comparisons:

Related-Samples Friedman's Two-Way Analysis of Variance by Ranks

Observer 1	Observer 2	Observer 3
Mean Rank = 1.20	Mean Rank = 2.00	Mean Rank = 2.80

Total N	10
Test Statistic	14.222
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	.001

8.6.2 Kendall's coefficient of concordance or W-statistic

A statistic, called Kendall's W coefficient was introduced by Kendall (1939). It may be used in the same situation where Friedman's test statistic is applicable. It has a special advantage that it gives the index of agreement. It is calculated as:

$$W = \frac{12}{n^2 k\,(k+1)\,(k-1)} \sum_{i=1}^{k} \left[R_i - \frac{n\,(k+1)}{2} \right]^2 , \qquad (8.10)$$

where n, k an R_i has been defined in (8.9).

If there is *perfect agreement* in the observers in all the blocks, the result of W is 1.0. If there is a perfect disagreement among observers then W is 0 or very close to zero. W can be easily calculated using Example 8.10. If IBM SPSS package is to be used, the entry of data is like the previous example.

SPSS output for Friedman and Kendall's Coefficient W

Friedman Two-Way ANOVA

Cases	Chi-Square	D.F.	Significance
10	12.80	2	.0017

Kendall Coefficient of Concordance

Cases	W	Chi-Square	D.F.	Significance
10	.711	14.22	2	.0008

We conclude at 5% level of significance that there is disagreement between the observers. Since W = 0.71, we say that there is 71% agreement.

Example S8-8

(The Kendall's coefficient of concordance automatically)

The data for example 8.9 will be in columns as follows:

	Patients	Observer1	Observer2	Observer3
1	1	6	7	8
2	2	4	5	6
3	3	2	4	2
4	4	3	4	5
5	5	5	4	6
6	6	8	9	10
7	7	5	7	9
8	8	6	7	8
9	9	4	6	8
10	10	7	9	8

We apply the Kendall's coefficient of concordance automatically as follows:

Analyze→ Nonparametric Tests→ Related Samples …

We may choose either ○ Automatically compare observed data to hypothesized , or Customize analysis to choose Kendall's coefficient of concordance test manually. Both will give the same result. We will Choose Customize analysis and click on Kendall's coefficient of concordance, as follows:

We click on ▶ Run to get the following final result:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Observer 1, Observer 2 and Observer 3 are the same.	Related-Samples Kendall's Coefficient of Concordance	.001	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Note: The decision rule is to reject the null hypothesis. Double click on the output will yield the following comparisons:

Related-Samples Kendall's Coefficient of Concordance
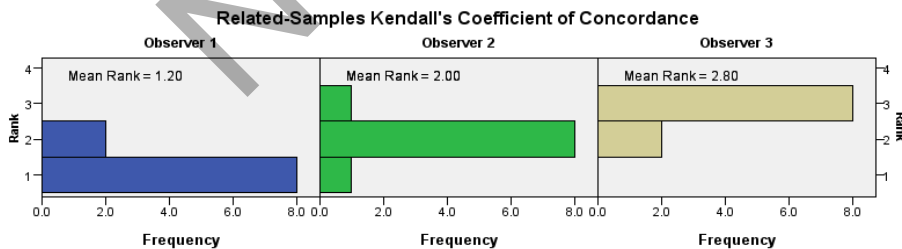
Total N	10
Kendall's W	.711
Test Statistic	14.222
Degrees of Freedom	2
Asymptotic Sig. (2-sided test)	.001

8.6.3 Cochran's Q test

Sometimes the use of a treatment results in one of two possible outcomes, i.e. the medicine is *effective* or *not effective*, a certain treatment may result in a *success* or a *failure*. If there are several treatments and each is applied in several different trials, the result is given in the form of a 2x2 contingency table and treatment differences may be tested using chi-square method. If the treatment result is classified into one of two categories then Cochran (1950) proposed a test known as Cochran's Q-test. This is an extension of McNemar test, which has been discussed in Chapter 7.

Each of k treatments is applied independently to each of n blocks and the result of each treatment is recorded as either 1 or 0, i.e. success or failure. Then the table takes the following form:

Table 8.10

Blocks (patients)	Treatment					Total
	1	**2**	**3**	**...**	**K**	
1	1	0	1		1	R_1
2	0	1	1		1	R_2
3	0	0	0		0	⋮
4	1	0	0		1	R_n
⋮	⋮	⋮	⋮		⋮	
n						
Total	C_1	C_2			C_k	

Assumptions:

Responses within blocks are correlated and the blocks (patients) are independent and as such are randomly selected.

The outcome of the treatment within each block may be dichotomized so the outcomes are tested as either 0 or 1.

(i) H_0 : All the treatments are equally effective.
        H_1 : There is difference in effectiveness.

(ii) $\alpha = 0.05$

(iii)   test-statistic: Cochran's test.

$$\chi_C^2 = \frac{k(k-1)\sum\limits_{j=1}^{k} C_j^2 - (k-1)N^2}{kN - \sum\limits_{i=1}^{n} R_i^2} , \qquad (8.11)$$

where k is the number of treatments or samples

   C_i is the sum of the columns
   R_i is the sum of rows

(iv) χ_C^2 is calculated and is compared with table value of χ^2 with (k - 1) degree of freedom and significance is determined accordingly.

Example 8.11:

One hundred people were asked to taste four new brands of cough syrup and state which new brands taste better than the present formula and which brands do not. As indicated in the following table, 15 subjects preferred the new taste to the old for all four brands, 3 subjects preferred brands A, B and C over the old brand but did not prefer brand D over the present formula, and so on. Test the null hypothesis that there is no significant difference in preferences among the four new brands of cough syrup.

Table 8.11

Brand				Number of subjects with this response
A	B	C	D	
1	1	1	1	15
1	1	1	0	3
1	1	0	1	3
1	0	1	1	6
0	1	1	1	21
1	1	0	0	1
1	0	1	0	1
0	1	1	0	1
1	0	0	1	2
0	1	0	1	2
0	0	1	1	19
1	0	0	0	3
0	1	0	0	3
0	0	1	0	2
0	0	0	1	13
0	0	0	0	5
Total 8	8	8	8	100

Solution:

(1) H_0 : There is no difference among A, B, C and D

   H_1 : There is difference.

(2) $\alpha = 0.05$

(3) test-statistic: Cochran's Q-test.

The SPSS package is used and the result is given as

SPSS output for Cochran's Q test

Cases	Cochran's Q-test	d.f	Significance
100	58.015	3	0.0000

Note that the data entry on SPSS package is like data for t-test (paired).

The result is significant at 5% level of significance and we conclude that there is difference in the taste of all the four brands of cough syrup. Note that at the time of entering the data, each set is entered a number of times mentioned against each set (see

application of SPSS package). If there are two treatments then the experimenter has a choice to use either Cochran's Q-test or McNemar test. Algebraically for two treatments Cochra's Q-test is identical to McNemar test and these are approximated by χ^2 with one degree of freedom. The McNemar test is used for brand A and B. The result is significant at 5% level of significant; therefore, we conclude that there is a difference in taste in two brands of syrup A and B.

SPSS output for McNemar test

Var 1 by Var 2

		Var 2		
		0	1	Total
Var 1	0	27	39	66
	1	22	12	34
			Cases	100

Chi- square = 5.0256 Significance = 0.0250

Example S8-9

(The Cochran's Q test automatically)

The data for example 8.11 will be in 4 columns and 100 rows. The 1st 18 case, as a part of the data is as follows:

File	Edit	View	Data	Transform	Analyze	Direct Marketing

75 : A				
	A	B	C	D
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
13	1	1	1	1
14	1	1	1	1
15	1	1	1	1
16	1	1	1	0
17	1	1	1	0
18	1	1	1	0

We apply the Cochran's Q test automatically as follows:

Analyze→ Nonparametric Tests→ Related Samples …

We may choose either ○ Automatically compare observed data to hypothesized , or Customize analysis to choose Kendall's coefficient of concordance test manually. Both will give the same result. We will Choose Customize analysis and click on Kendall's coefficient of concordance, as follows:

We click on ▶ Run to get the following final result:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Brand A, Brand B, Brand C and Brand D are the same.	Related-Samples Cochran's Q Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

P-value

Note: The decision rule is to reject the null hypothesis. Double click on the output will yield the following comparisons:

Related-Samples Cochran's Q Test

Total N	100
Test Statistic	58.015
Degrees of Freedom	3
Asymptotic Sig. (2-sided test)	.000
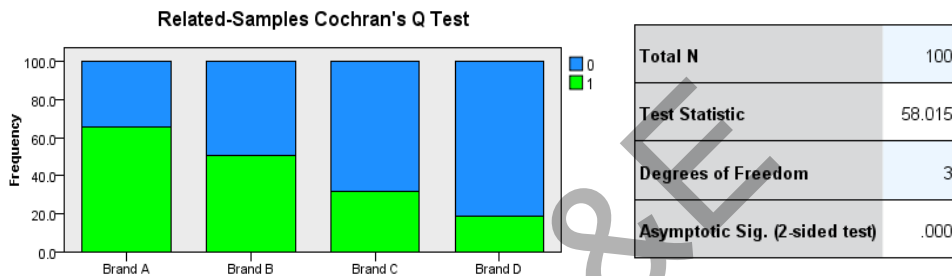
8.7 Measures of Rank Correlation

It is commonly known as Spearman Rank Correlation. This is frequently used because of its simplicity. This measure of correlation may be used with ordered data or data transformed to ranks without any requirements concerning the scale of measurement although it is difficult to interpret unless the scale of measurement is interval. The measure of correlation as given by Spearman (1904) is usually designated by ρ (rho) and if there is no tie, then

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)} \tag{8.12}$$

where Σd_i^2 is the sum of square of the differences. If there are not many ties, the procedure for calculation is as:

(i)   Rank the values of one set (say x) from 1 to n and also rank the value of second set (say y) from 1 to n.

(ii)   Find the differences (d_i) between the ranks of first set and the second set.

(iii)   Find Σd_i^2.

Example 8.12:
Twelve sets of identical twins were given psychological tests to measure their aggressiveness. The emphasis is on examination of the degree of similarity between twins within the set. The data were measures of aggressiveness and are given in Table 8.12:

Table 8.12

Twin Set	1	2	3	4	5	6	7	8	9	10	11	12
First Born	86	72	77	68	91	73	75	92	70	71	88	87
Second Born	88	77	76	64	96	72	65	90	66	80	81	73

Calculate the rank correlation coefficients between the two measures aggressiveness and test the significance of this correlation coefficient.

Solution:

First Born (x)	R1	Second Born (y)	R2	(R1 - R2) = d_i	d_i^2
(1)	(2)	(3)	(4)	(5)	(6)
86	8	88	10	-2	4
72	4	77	7	-3	9
77	7	76	6	+1	1
68	1	64	1	0	0
91	11	96	12	-1	1
73	5	72	4	1	1
75	6	65	2	4	16
92	12	90	11	1	1
70	2	66	3	-1	1
71	3	80	8	-5	25
88	10	81	9	1	1
87	9	73	5	4	16
Total					76

The rank of x and y are given in column 2 and 4 respectively of the above table.

$$N = 12 \qquad \sum d_i^2 = 76$$

Using formula (8.12), we can calculate ρ as

$$\rho = 1 - \frac{6(76)}{12(12^2 - 1)} = 0.7343$$

with $p < 0.007$

If we calculate Pearson correlation coefficient then $r = 0.7215$ with $p < 0.003$

The significance of this can be tested as:

(i) H_0 : The measure of aggressiveness of two identical twins are mutually independent.

   H_1 : There is either a positive correlation or a negative correlation between the two measures of aggressiveness.

(ii) $\rho = 0.05$

(iii)   It is a two tailed-test, the table value at 5% level of significant for 11 df is 0.623.

(iv)   The calculated value of ρ is 0.7343 which is greater than the table value. Therefore, it is significant and we say that there is a relation between the measures of aggressiveness. The SPSS package is used to calculate the rank correlation coefficient. The entry of data is like Pearson's correlation.

Example S8-10

The data for example 8.12 will be as follows:

	set	first	second
1	1	86	88
2	2	72	77
3	3	77	76
4	4	68	64
5	5	91	96
6	6	73	72
7	7	75	65
8	8	92	90
9	9	70	66
10	10	71	80
11	11	88	81
12	12	87	73

We apply the Spearman Rank Correlation as follows:

Analyze→ Correlate→ Bivariate …
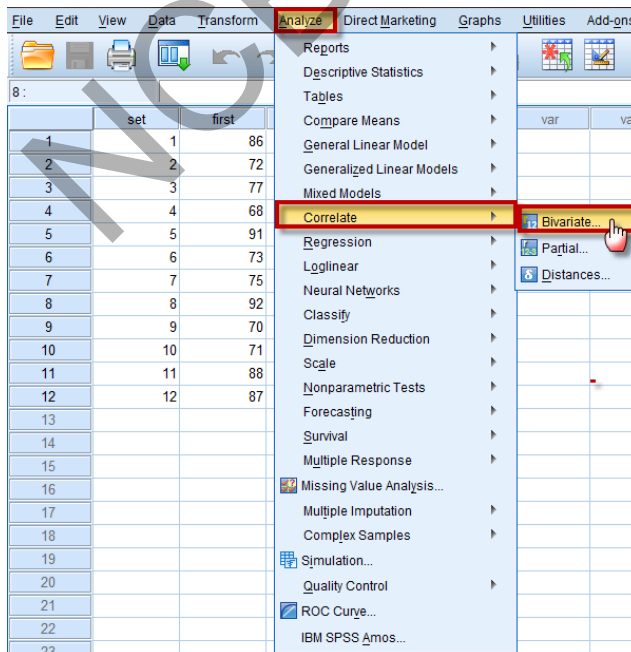
Move the two variables to "Variables". Mark on Spearman.
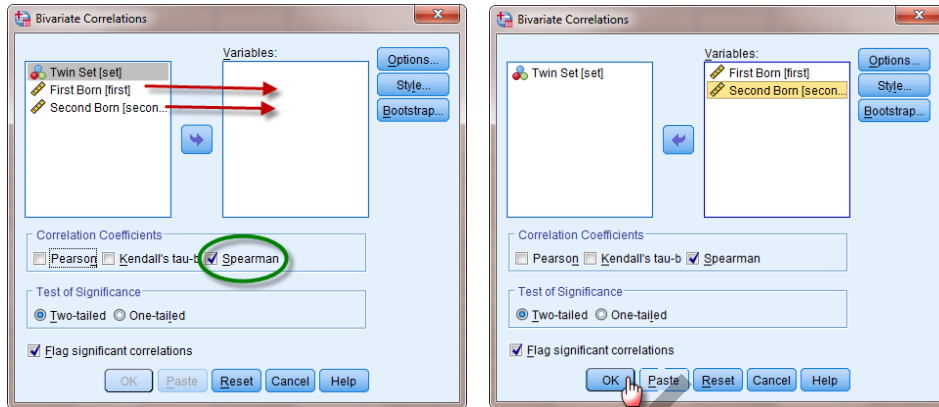
We click on OK to get the following final result:

Correlations

			First Born	Second Born
Spearman's rho	First Born	Correlation Coefficient	1.000	.734**
		Sig. (2-tailed)	.	.007
		N	12	12
	Second Born	Correlation Coefficient	.734**	1.000
		Sig. (2-tailed)	.007	.
		N	12	12

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

The p-value = 0.007, the result is significant therefore we can say with 95% level of confidence that there is 73.4% correlation between the measure of aggressiveness in the population where from this sample has been selected.

Chapter 9

Logistic Regression

9.1 Introduction

In Chapter 6, we studied linear regression but this method of analysis is generally not applicable when the dependent variable is **binary** or has only two values (yes, no), or has a nominal measurement level with more than two values. An other method known as logistic regression is commonly used for such situations. Before this, a method of discreminatnt analysis was also in practice but this allows direct prediction of group membership but the assumptions of multivariate normality if independent variables is required for prediction rule to be optimal. Logistic regression model requires fewer assumptions than discriminant analysis and even when the assumptions required for discriminant analysis are not met, logistic regression, still performs well. [see Hosmer and Lemesho (1989) and Kleinbaum (1992).] In logistic regression one can directly estimate the probability of an event whereas in linear regression it is not possible as they do not fall in the inteval 0 to 1.

The method of logistic regression has become the standard method of analysis for the last three decades, when the dependent variable is binary or dichotomous (yes, no). The difference between logistic and linear regression lies both in the choice of a model and assumptions. Once the difference is accounted for, then logistic method of analysis follows the same general principles as used in linear regression. To illustrate logistic regression, let us consider a dichotomous disease outcome with zero representing *not diseased* and 1 representing *diseased*, i.e. coronary heart disease (CHD) may be classified as either zero (without CHD) or 1 (with CHD). The CHD is an outcome of some cause, so we call CHD as dependent variable. Suppose we are interested in a single dichotomous exposure variable, i.e. smoking which is classified as "yes" for smoker and "no" for non-smoker. To evaluate the extent to which smoking is associated with CHD, we perform analysis by the method of logistic regression. We can take into consideration some control variables, if we like, such as age, race, sex, etc. The difference between logistic regression and odds ratio is:

i)   The method of logistic regression, is applicable in even elementary analysis.

ii) The probability of an event is calculated by the use of logistic method, whereas we cannot calculate the probability of an event by the method of odds ratio.

iii) Odds ratio tells us only how much risk of CHD is involved after a certain period but does not explain how much the risk of CHD is involved with the increase in age whereas the method of logistic regression explains this point also in an elegant way.

In fact, logistic regression is needed by health scientists and others despite the fact that some approximation is involved because of the tranformation of the data from one mode to another mode. This subject is very vast and it is not possible to cover all the aspects of

logistic regression in this book. We have tried to summarize the necessary points which are useful for health scientists.

The logistic regression model is given as:

$$f(z) = \text{Prob (event)} = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}} \tag{9.1}$$

where $z = \beta_0 + \beta_1 X_1$ (simple model) and $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$ for multiple model and $\beta_0, \beta_1, \beta_2, \dots$ are coefficients,

$\text{Prob (Event)} = \log_e \left(\dfrac{p}{1 - p} \right)$, p is the proportion of the event of "yes" or "no" and e is the

base of natural logarithms.

The probability of the event not occurring is estimated as:

Prob (no event) = 1 - Prob (event) (recall binomial distribution)

Many distribution functions have been proposed for use in the analysis of a dichotomous outcome (See Cox-1970) but logistic regression method is very popular for the following reasons.

(1) Logistic function [f(z)] ranges between 0 and 1 and is the primary reason for its popularity. The model is designed to describe probabilities, which is always some number between 0 and 1. In epidemiological terms, such a probability gives the *risk* of an individual getting a disease, i.e. individual risk is measured by $0 \leq \text{Prob} \leq 1$. By using the logistic model, we can never get a risk estimate either above 1 or below 0. This is the primary reason why logistic method is the first choice.

(2) The shape of the logistic model f(z) is s-shaped. This is considered to be widely applicable for the multivariable nature of an epidemiological research. The s-shape of f(z) indicates that the effect f(z) on an individual's risk is minimal for low z's until some threshold is reached. This risk then rises rapidly over a certain range of intermediate z values, and then remains extremely high around 1. The shape is indicated in Fig. 9.1.
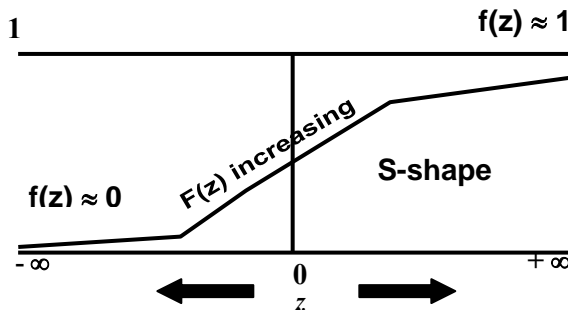
Fig.9.1: Shape of the logistic regression

By the use of logistic model, we can estimate the probability that the disease will develop during a defined period say t_0 to t_1.

9.2 Fitting of Simple Logistic Model

For fitting of logistic regression following example is given.

***Example 9.1*:**
In a study of 100 subjects that participated in the study, the age in years alongwith the presence (yes=1) and absence (no = 0) of evidence of coronary heart disease is recorded.

Table 9.1
Age and Coronary Heart Disease Status (CHD) of 100 Subjects

ID	AGE	CHD	ID	AGE	CHD	ID	AGE	CHD
1	20	0	35	38	0	68	51	0
2	23	0	36	39	0	69	52	0
3	24	0	37	39	1	70	52	1
4	25	0	38	40	0	71	53	1
5	25	1	39	40	1	72	53	1
6	26	0	40	41	0	73	54	1
7	26	0	41	41	0	74	55	0
8	28	0	42	42	0	75	55	1
9	28	0	43	42	0	76	55	1
10	29	0	44	42	0	77	56	1
11	30	0	45	42	1	78	56	1
12	30	0	46	43	0	79	56	1
13	30	0	47	43	0	80	57	0
14	30	0	48	43	1	81	57	0
15	30	0	49	44	0	82	57	1
16	30	1	50	44	0	83	57	1
17	32	0	51	44	1	84	57	1
18	32	0	52	44	1	85	57	1
19	33	0	53	45	0	86	58	0
20	33	0	54	45	1	87	58	1
21	34	0	55	46	0	88	58	1
22	34	0	56	46	1	89	59	1
23	34	1	57	47	0	90	59	1
24	34	0	58	47	0	91	60	0
25	34	0	59	47	1	92	60	1
26	35	0	60	48	0	93	61	1
27	35	0	61	48	1	94	62	1
28	36	0	62	48	1	95	62	1
29	36	1	63	49	0	96	63	1
30	36	0	64	49	0	97	64	0
31	37	0	65	49	1	98	64	1
32	37	1	66	50	0	99	65	1
33	37	0	67	50	1	100	69	1
34	38	0						

*AG = Age groups

It is of interest to explore the relationship between age and presence or absence of CHD in this study.

Solution:

The outcome (dependent) variable is CHD, which is dichotomous, therefore, multiple linear regression cannot be fitted, instead logistic model will be fitted. Because of the complexity in calculations the IBM SPSS package is used to fit the logistic regression, as can be seen in the following steps:

Example S9-1

The data will be in 3 columns and a part of the data is as follows:

	ID	Age	CHD
1	1	20	0
2	2	23	0
3	3	24	0
4	4	25	0
5	5	25	1
6	6	26	0
7	7	26	0
8	8	28	0
9	9	28	0
10	10	29	0

The Variable View is as follows:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ID	Numeric	8	0		None	None	4	Right	Nominal	None
2	Age	Numeric	8	0		None	None	4	Right	Scale	Input
3	CHD	Numeric	8	0	Coronary Heart...	None	None	6	Right	Nominal	Target

We apply the Binary logistic as follows:
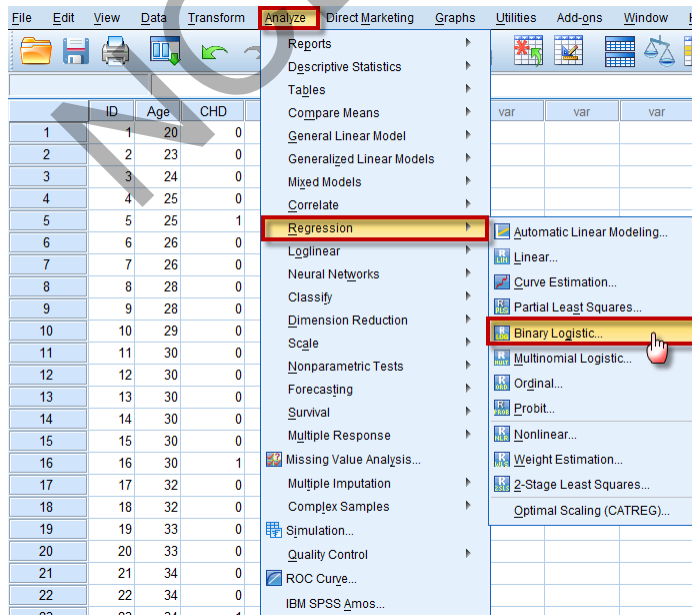
Analyze➔ Regression➔ Binary Logistic…

Move the variable "CHD" to Dependent:

Move the variable "Age" to Covariates:
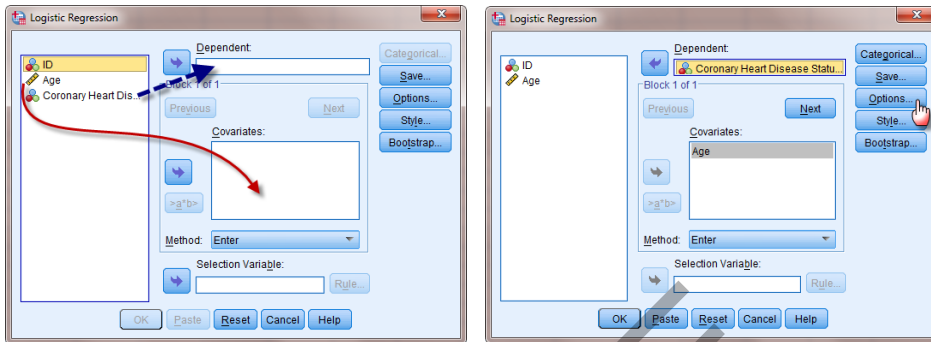
Click on and select the following:
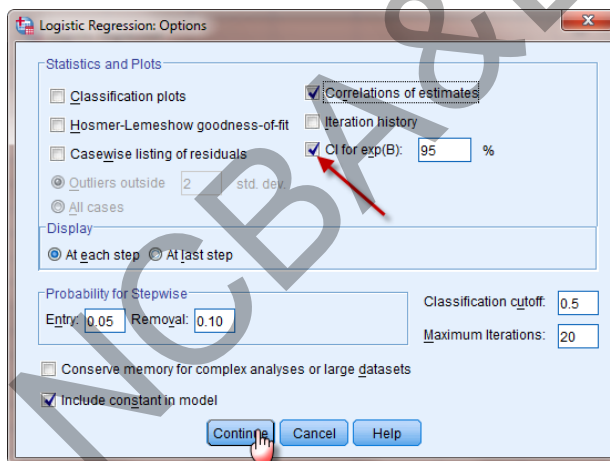
Now click on then , to get the following outputs:

SPSS output for logistic regression

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	107.353	.254	.341

The value -2 log likelihood for model containing independent variables = 107.353.

Classification ^a

		Predict		
		CH		Percenta
Observe		0	1	Corre
CH	0	45	12	78.
	1	14	29	67.
Overall				74.

a. The cut value is

From the above Classification Table for CHD we see that 45 patients without CHD were correctly predicted by the model not to have CHD. Similarly 29 men with CHD were correctly predicted to have CHD. A total of 26 (12 + 14) men were miss classified in the analysis- 12 men with negative CHD and 14 men with positive CHD, whereas 78.95% of the men were correctly classified without disease and 67.44% were correctly classified as with CHD. Overall 74% of the 100 men were correctly classified.

Omnibus Tests of Model

	Chi-square	df	Sig.
Step	29.310	1	.000
Block	29.310	1	.000
Model	29.310	1	.000

The value of model chi-square is 29.31 with p = 0.000. This is highly significant. Therefore we are 95% confident that the fitted model is appropriate.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.f or EXP(B)	
								Lower	Upper
Step 1^a	AGE	.111	.024	21.25	1	.000	1.117	1.066	1.171
	Constant	-5.309	1.134	21.94	1	.000	.005		

a. Variable(s) entered on step 1: AGE.

Interpretation of results

(1) $\exp(e^{\beta}) = OR = e^{0.111} = 1.1173$

A value of 1.12 of odds ratio means that with the increase of one year in age the risk of CHD is increased 1.12 times provided all other factors are kept constant. Since one year increase does not give any significant change, therefore, we can see the significant change after 10 years. This is calculated as:

$$e^{years \times \beta} = e^{10 \times 0.1109} = 3.03$$

This indicates that with an increase of 10 years in age the risk of CHD increases 3.03 times.

(2) Wald's statistic, W

$$W = \left[\hat{\beta}/S.E(\hat{\beta})\right]^2 = \left(\frac{0.1109}{0.0241}\right)^2 = 21.18, \tag{9.2}$$

This estimate, under the hypothesis that $\beta_1 = 0$, follows a standard normal distribution, $N(0, 1)$. In this example, Wald statistic shows that age has significant affect on CHD, i.e. as age increases, chances of CHD increases. Hauck and Donner (1977), examined the performance of Wald statistic and found that it behaved in an aberrant manner, after failing to reject when the coefficient is significant. Moreover, it has an undesirable property, i.e. this method fails when the coefficient $(\hat{\beta})$ is large. If the coefficient is large, the $SE(\hat{\beta})$ is too large, then the Wald-statistic is too small, to reject the null hypothesis, when in fact the null hypothesis should be accepted. Therefore, when coefficient is large, one should not rely on Wald-statistic, instead one should build a model with and without that variable and base the hypothesis test on chi-square test.

(3) Partial Correlation Coefficient(R)
R ranges from -1 to +1. A positive value of R indicates that as the variable increases in value so does the likelihood of the event occurring. If R is negative, the opposite is true. Small value of R indicates that the *variable* has little contribution to the model.

9.2.1 Application of simple logistic model for prediction

We can apply the simple logistic model to find the chances of a disease of a person at a given age. *If the probability is less than 0.5, we say that the event is not likely to occur but if the probability is 0.5 or more we say that there is a chance of the occurrence of an event. The higher the probability the greater the chance of occurrence of the disease.* Using the results of Example 9.1. in (9.2)

$$Z = -5.31 + 0.111 \text{ (age)}$$

The probability of the occurrence of an event (CHD) may be calculated as:

$$P(CHD) = \frac{1}{1 + e^{-Z}}$$

Suppose the age is 40 years then

$$Z = -5.31 + 0.111(40) = -0.87 \text{ and } e^{-(-0.87)} = 2.39$$

Using (9.1) the probability of CHD will be

$$P(CHD) = \frac{1}{1 + 2.39} = 0.29$$

On the basis of data given if the age of a person is 40, there is a small chance of CHD as the probability is less than 0.50 or we say that there is only 29% chance of CHD.

Again suppose the age = 60, then from the model Z = 1.35 and e$^{-1.35}$ = 0.26. The probability of the occuring of CHD will be

$$\mathbf{P}(\text{CHD}) = \frac{1}{1 + 0.26} = 0.79$$

Since the probability is high, so a person who is approaching the age of 60 has about 80% chances of CHD.

We will show how to calculate the probabilites directely through the IBM SPSS in the following example:

Example S9-2

We will add the age of 40 and age of 60 to the data and apply the Binary logistic and get the predicted values directly, as follows:

Analyze→ Regression→ Binary Logistic…
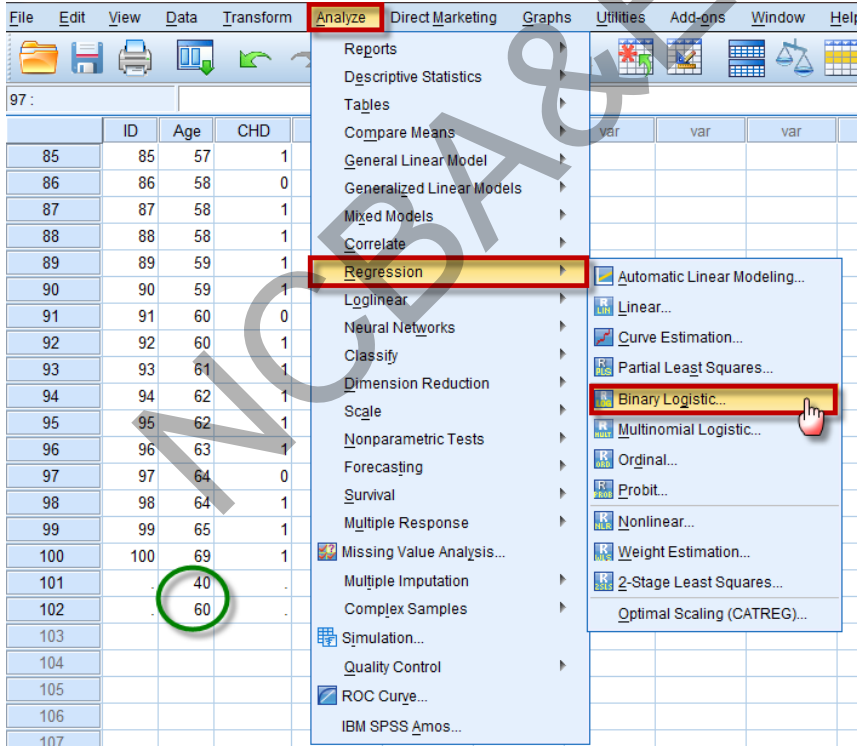
Move the variable "CHD" to Dependent:

Move the variable "Age" to Covariates: choose "save"

Click on Save... and choose "Probabilites" for the Predicted Values, as follows:
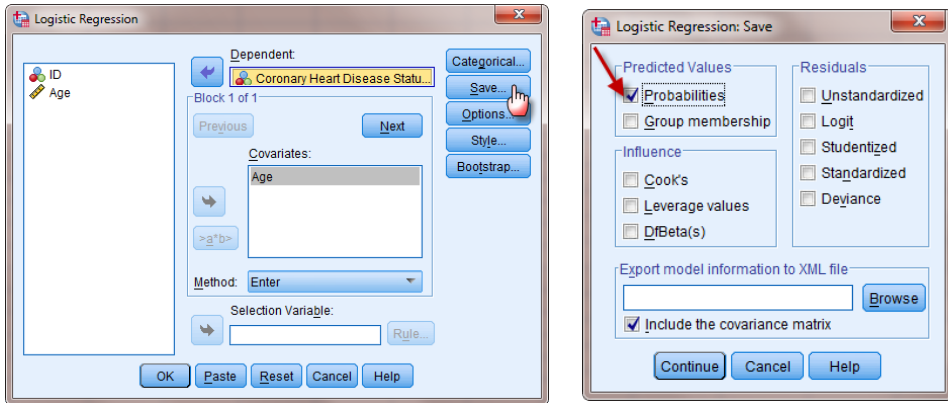
Now click on Continue then OK , to find out the predicted values added to the data directly, as:

90	90	59	1	.77467
91	91	60	0	.79344
92	92	60	1	.79344
93	93	61	1	.81103
94	94	62	1	.82745
95	95	62	1	.82745
96	96	63	1	.84272
97	97	64	0	.85687
98	98	64	1	.85687
99	99	65	1	.86994
100	100	69	1	.91246
101		40		.29471
102		60		.79344

9.2.2 Confidence limits for odds ratio

95% confidence limit may be calculated as:

$$e^{\hat{\beta} \pm 1.96 \, S.E\left(\hat{\beta}\right)} \tag{9.3}$$

$$= e^{0.111 \pm 1.96 \, (0.0241)} \text{ or } [1.07 \, , \, 1.17]$$

The odds ratio is greater than 1 and the confidence limits does not include 1 so age is playing a significant role in the CHD. We can say as the age increases there are more chances of CHD. The other formula for the calculation of confidence limits is as:

$$(OR \div e^{1.96 \, S.E\left(\hat{\beta}\right)}, \, OR \times e^{1.96 \, S.E\left(\hat{\beta}\right)}) = [1.07 \, , \, 1.17] \tag{9.4}$$

Anyone of the above formula can be used for the calculation of confidence limits.

We will show how to calculate the confidence limits *for the odds ratio, through the IBM SPSS in the following example:*

Example S9-3

We apply the Binary logistic for the data in example S9-1, as follows:

Analyze→ Regression→ Binary Logistic…
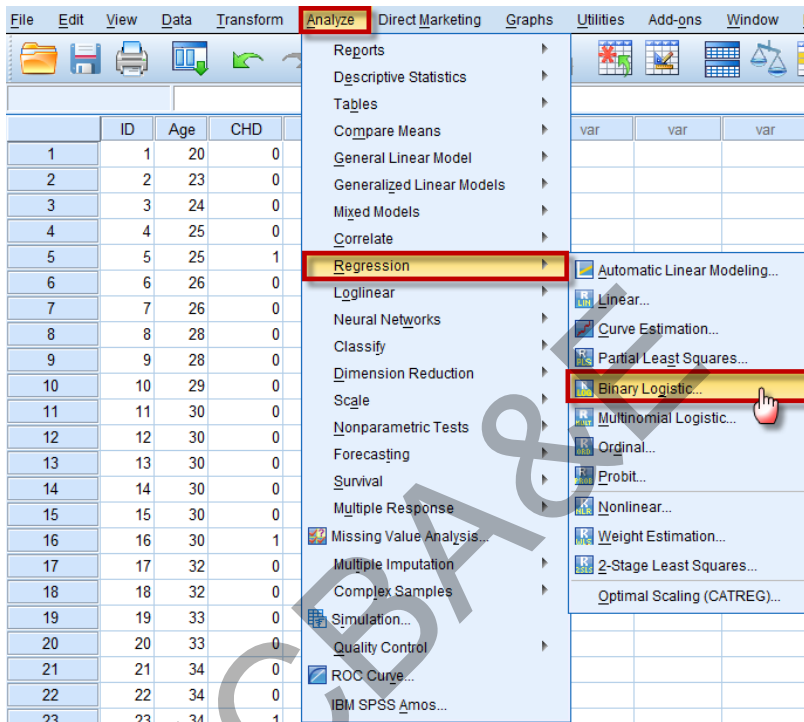
Move the variable "CHD" to Dependent:

Move the variable "Age" to Covariates: choose "Options"

Mark on "CI for exp(B), at 95%;
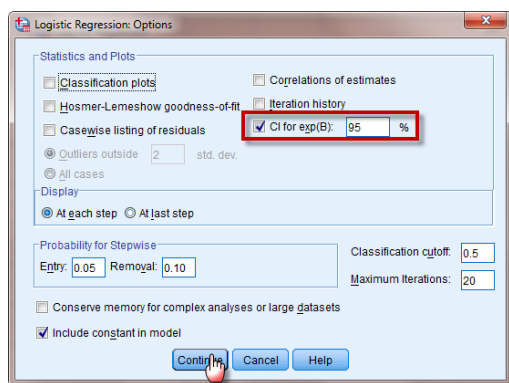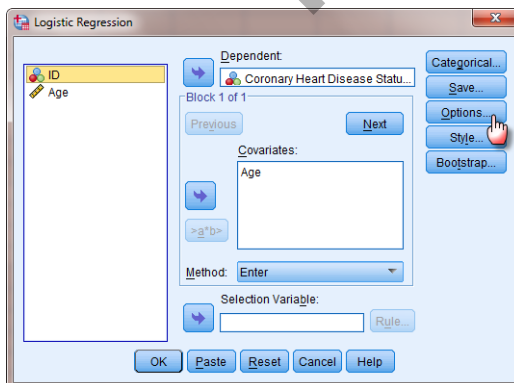
Now click on [Continue] then [OK], to find out the 95% Confidence limits for odds ratio, as:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Age	.111	.024	21.254	1	.000	1.117	1.066	1.171
	Constant	-5.309	1.134	21.935	1	.000	.005		

a. Variable(s) entered on step 1: Age.

9.3 The Multiple Logistic Model

Like linear regression model we will generalize the simple logistic regression model to the case of multiple logistic regression model. This has been defined before and is as:

$$\text{Prob(event)} = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}}$$

where:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K$$

Example 9.2:
Suppose the disease of interest is CHD. Here CHD is coded as 1 if a person has the disease and 0 otherwise. There are three independent variable such as X_1 = Age (quantitative); X_2 = ECG (electro-cardiogram status) is 1 if abnormal and 0 if normal; X_3 = CAT (catecholamine level) is 1 if high and 0 if low. The data are of 609 white males [Kleinbaum (1992)].

Solution:
Here CHD is a dependent variable and dichotomous. In order to see the effect of Age, ECG and CAT, we fit multiple logistic regression model taking Age, ECG and CAT as independent variables. These 609 people are followed for 9 years to determine CHD status.

Multiple logistic regression model was fitted using IBM SPSS package and the coefficients are obtained

$$\hat{\beta}_0 = \text{-3.911, } \hat{\beta}_1 = 0.029, \hat{\beta}_2 = 0.342 \text{ and } \hat{\beta}_3 = 0.652 \text{, therefore}$$

$$Z = \text{-3.911} + 0.029(\text{Age}) + 0.342(\text{ECG}) + 0.652(\text{CAT}) \tag{9.5}$$

The odds ratio for the variables alongwith coefficients are as:-

Table 9.3

Variable	Coefficient	OR
Age	0.029	1.03
ECG	0.342	1.41
CAT	0.652	1.92
Constant	-3.911	

Since the odds ratio are greater than 1 in all cases, therefore, Age, abnormal ECG, (code is 1 if abnormal) and high CAT (catecholamine level is 1 of high) will have significant role in CHD. The odds ratio for the age is 1.03, therefore the increase of one year in age increases the risk of CHD by 1.03 times more. The odds ratio for ECG is 1.41 and the code is 1 if ECG is positive, therefore the risk of CHD is 1.41 more if the ECG is negative. Similarly the odds ratio for the CAT is 1.92 and code for abnormal CAT is 1, the risk of CHD is 1.92 time more if CAT is abnormal.

Suppose we want to use our fitted model, to obtain the predicted risk for a certain individual. For this purpose, we would like to specify the values of Age, ECG and CAT. suppose the Age is 45, ECG = 1 and CAT = 0, then from (9.5)

$$Z = -3.911 + 0.029(45) + 0.342(1) + 0.652(0) = -2.264$$

$$\mathbf{P}(\text{predicted risk}) = \frac{1}{1 + e^{-Z}} = \frac{1}{1 + e^{2.264}} = \frac{1}{10.62} = 9.4\%$$

Then the individual has 9.4% risk of CHD over the period of follow up study. If we say that Age = 45, ECG = 1 and CAT = 1, then we have

$$Z = -3.911 + 0.029(45) + 0.342(1) + 0.652(1) = -1.16$$

$$\mathbf{P}(\text{predicted risk}) = \frac{1}{1 + e^{-(-1.162)}} = \frac{1}{6.013} = 16.6\%$$

The person has 16.6% risk of CHD over the period of follow up study.

From the above example we conclude that a person whose age is 45, ECG is abnormal (1) but CAT is low (0), the risk of CHD is 9.4% whereas, the same person whose CAT is also high the risk of CHD is 16.6%.

The risk ratio can be calculated as:

$$PR = \frac{P(CAT = 1)}{P(CAT = 0)} = \frac{0.166}{0.094} = 1.77 \tag{9.6}$$

Thus using a fitted model, we find that the person with high CAT has 1.77 times more risk than a person with low CAT.

Note that two conditions must be satisfied to estimate *risk ratio* (RR) directly. First that we must have *follow up study* so that we can legitimately estimate individual risk. Second, for the two individuals being compared, we must specify values for all the independent variables in our fitted model to compute risk for each individual. If either of the above condition is not satisfied we cannot estimate risk ratio directly but it may be possible to estimate risk ratio indirectly. For this purpose odds ratio is computed. In fact the odds ratio is the only measure of association directly estimated from a logistic model, regardless of whether the study design is follow up, case-control or cross-sectional. Though logistic model is applicable to case-control and cross-sectional studies, there is one important limitation in the analysis of such studies. This model cannot be used to

predict individual risk for case-control or cross-sectional studies whereas in follow-up studies a fitted logistic model can be used with specified independent variables. In fact estimates of odds ratio can be obtained for case-control and cross-sectional studies.

For a 2x2 table, risk estimates can be used only if the data are derived from a follow-up study, whereas odds ratio is appropriate if the data are derived from case-control or cross-sectional study.

Example 9.3:
The treatment and prognosis depends how much the disease has spread. One of the regions to which a cancer may spread is the lymph nodes. If the lymph nodes are involved the prognosis is generally poorer than if they are not, that is why it is desirable to establish as early as possible whether the lymph nodes are cancerous. For certain cancers exploratory surgey is done to determine whether the nodes are cancerous, since this will determine what treatment is needed. If one could predict whether the nodes are affected or not on the basis of data, then surgery is not required. By doing so considerable discomfort and expense could be avoided. For this purpose Brown (1982) took a sample of 53 men with possible prostate cancer.

For each patient age, serum acid phophate(ACID). the stage of the disease (STAGE); an indication how advanced the disease is, the grade of the tumor; an indication of malignancy, X-Ray, as well as the cancer has spread to the regional lymph nodes at the time of surgery was recorded. This data is given in Table 9.4 and has been analysed using logistic model and prediction whether nodes have been affected are made.

Solution:
X-Ray, STAGE, GRADE are qualitative(0, 1) variable and are coded as 1 if X-Ray indicates positive result, the value is 1 if the Stage is advanced, the value is 1 if it is malignant tumor. Node involvement is dependent variable coded as yes or no or 1 or 0. The result of the Logistic regression model using IBM **SPSS** package is given as :

(i) 2×2 Table

Table 9.4: predicted

		0	1	
Observed	0	28	5	84.85%
	1	7	13	65.00%
		35	18	77.36%

-2log likelihood = 70. 252

It can be seen from the table that 28 men with negative nodes are predicted correctly by the logistic model; 13 men with positive nodes were correctly predicted to have positive nodes. The off diagonal entries (12) of the table were missclassified, 5 men with negative nodes and 7 men with positive nodes; 84.85% were correctly classified without diseases. 65% were correctly classified with diseased nodes. Overall 77.36% of 53 men were correctly classified.

(ii) Coefficients

Table 9.5

variable	coeff.	S.E	Wald statistic		p-value	R	OR
			T	F			
age	-.069	0.058	1.20	1.44	0.23	0.00	0.93
ACID	0.024	0.013	1.84	3.39	0.06	0.14	1.02
X-Ray	2.045	0.807	2.53	6.40	0.01	0.25	7.73
GRADE	0.761	0.771	0.99	0.98	0.32	0.00	2.14
STAGE	1.564	0.774	2.02	4.08	0.04	0.17	4.78
Constant	0.618	3.460	-	-	-	-	-

Let us first interpret the result through Wald's statistic. In the table given above only *X-ray* and *stage* appear as significant as the t- values are more than 1.96 at 5% level of significant variables. We conclude that positive result of X-ray and Stage will indicate that nodes are affected. As mentioned earlier one cannot rely on the results of Wald's statistic as this method fails when coefficients are large. The p-values of *X-Ray* and *Stage* also indicate that variables have significant contribution. All other variables appear as non significant. This can be interpreted through odds ratio as:

Since the coefficient of X-ray is positive, and high code is 1 for X-ray which indicates positive result, odds ratio is 7.33, therefore a man whose X-ray report is positive has 7.33 times more chances that nodes are affected than the person whose X-ray result is negative. Again the coefficient of the Stage is positive and high code is 1 if the stage of the disease is advanced,odds ratio is 4.7, therefore a person whose stage is advanced has about 5 times more chances that nodes are affected.

The probability(predicted) of the involvement of nodes will be calculated

$$\textbf{P}(\text{nodal involvement}) = \frac{1}{1 + e^{-z}}$$

where:

$$z = 0.618-0.069(\text{age}) + 0.024(\text{ACID}) + 2.045(\text{X-Ray})$$
$$+ 0.761(\text{GRADE}) + 1.564(\text{STAGE}).$$

Case 1

Suppose the age of a person is 66 years; his serum phosphatase level is 48 and all other have zero values then

$$z = 0.0618 - 0.693(66) + 0.0243(48) = -3.346,$$

The probability of nodal involvement may be calculated using (9.3)

$$\textbf{P}(\text{nodal involvement}) = \frac{1}{1 + e^{-(-3.346)}} = 0.034 = 3.4\%$$

Since the probability is very low, it can be predicted that nodes are unlikely to be malignant.

Case 2

Age = 60 years; serum Acid Phophatase = 62; X-ray =1(positive)

the z value will be

z = 0.0618 - 0.0693(60) + 0.0243(62) + 2.0453(1) = -0.54

The estimated probability will be = **P** (malignant node) = 0.37

Again the probability is less than 0.50, therefore we conclude that nodes are unlikely to be malignant.

Case 3

Age = 60, ACID = 62, X-Ray = 1, Grade = 1 Stage = 0

The z is 0.22. Therefore the estimated probability will be = **P**(malignant node) = 0.554

Since it is more than 50% we say under the rule that nodes are likely to be malignant.

Case 4

Age = 60, ACID = 62, X-Ray = 1, Grade = 1, Stage = 1

The estimated probability will be

P(malignant node) = 0.73

There is a high chance that nodes are likely to be malignant..

Example 9.4:
Data for the risk factors associated with low infant birth weight were given in example Chapter-6 alongwith code sheet. The dependent variable is low birth weight. It is 1 if weight is less than 2500 pounds, otherwise = 0, the independent variables are Age of the mother (Age); weight in pounds at the last menstural period (LWT); smoking status (yes = 1, no = 0); race (white = 1, black = 2, other = 3); History of premature labor (none = 0, yes = 1), history of hypertension (yes = 1, no = 0), presence of uterine irritability (yes = 1, no = 0), number of physician visit (none = 0, one = 1). Fit the multiple logistic regression and interpret the result.

Solution:
The data are entered like multiple linear regression, instead of clicking linear regression we now go to logistic regression. Here Low birth weight with coding system is dependent variable where age, number of visits of physicians, history of hypertension, weight at the last menstrual period, history of premuature labor, race, smoking, and uterine irritability are independent variables.

Because of the complexity of the data the caculations are done using SPSS package and the output is given below

SPSS output for Logistic Regression
Model Summary

-2 Log Likelihood	Cox & Snell R Square	Nagelkerke R Square
201.614	0.160	0.226

The value of –2log likelihood for model containing independent variable = 201.614.

Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step	33.058	8	.000
Block	33.058	8	.000
Model	33.058	8	.000

Model chi-square gives significant result with p = 0.000, therefore the model is an appropriate one.

Classification Table

	Predicted		
Observed	0	1	**Percentage Correct**
0	120	10	92.3
1	37	22	37.3
Overall Percentage			75.1

From the above Classification Table we can see that 120 children with high birth weight were correctly predicted 22 children with low birth weight were predicted correctly classified a total of 47 children were miss-classified 10 with high birth weight, and 37 with low birth weight. 92.03% of high birth weights are correctly classified, whereas 37.3% with low birth weight were correctly classified overall 75.1% children were correctly classified.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp (B)	95.0% C.I.f or EXP(B) Lower	Upper
AGE	-.033	.036	.798	1	.372	.968	.901	1.040
LWT	-.010	.007	2.324	1	.127	.990	.977	1.003
RACE	.482	.217	4.934	1	.026	1.620	1.058	2.480
PTL	.926	.399	5.388	1	.020	2.523	1.155	5.513
SMOKE	.694	.431	2.599	1	.107	2.002	.861	4.656
HT	1.933	.685	7.972	1	.005	6.911	1.806	26.442
UI	.799	.457	3.065	1	.080	2.224	.909	5.443
FTY	.055	.189	.086	1	.770	1.057	.729	1.532
Constant	-.563	1.27	.198	1	.656	.569		

Variable(s) entered are AGE, LWT, RACE, PTL, SMOKE, HT, UI, FTV.

If we look into the result of Wald's statistics, *hypertension (HT), weight at the last menstrual period (LWT), history of premature labor (PTL), race, smoking (Smoke) and presence of uterine irritability (UI) are appearing as significant variables.*. Since Wald's statistic does not provide reliable result, therefore we try to inference through odds ratio. The elimination process will be used by the SPSS package and the results are as follows:

Note that step 4 is the final answer.

SPSS output after Elinimation Process

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	224.638	.052	.073
2	217.790	.085	.120
3	212.363	.111	.157
4	208.303	.130	.183

Classification Table[a]

		Predicted		
		BWT1		Percentage Correct
Observed		0	1	
Step 4		120	10	92.3
		44	15	25.4
				71.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp (B)	95.0% C.I. for EXP(B) Lower	Upper
Step 1	PTL	1.035	.332	9.750	1	.002	2.816	1.470	5.393
	Constant	-1.074	.188	32.49	1	.000	.342		
Step 2	PTL	1.137	.339	11.25	1	.001	3.117	1.604	6.056
	HT	1.693	.661	6.562	1	.010	5.434	1.488	19.843
	Constant	-1.217	.202	36.37	1	.000	.296		
Step 3	RACE	.434	.189	5.270	1	.022	1.544	1.066	2.237
	PTL	1.321	.361	13.36	1	.000	3.747	1.845	7.608
	HT	1.690	.678	6.220	1	.013	5.421	1.436	20.459
	Constant	-2.096	.451	21.55	1	.000	.123		
Step 4	RACE	.412	.192	4.604	1	.032	1.510	1.036	2.199
	PTL	1.249	.367	11.58	1	.001	3.488	1.698	7.164
	HT	1.835	.681	7.255	1	.007	6.264	1.648	23.805
	UI	.909	.447	4.142	1	.042	2.481	1.034	5.953
	Constant	-2.194	.460	22.70	1	.000	.112		

Here Hypertension, History of premature labor, race and uterine irritability are appearing as significant variables with odds ratio 6.3, 3.5, 1.5, and 2.5 respectively.

Now dummy variables can be created by the automatic process of the logistic regression model and the result is given as:

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
RACE	1	95	1.000	.000
	2	26	.000	1.000
	3	68	.000	.000

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	10.034	1	.002
	Block	10.034	1	.002
	Model	10.034	1	.002
Step 2	Step	6.847	1	.009
	Block	16.882	2	.000
	Model	16.882	2	.000
Step 3	Step	4.763	1	.029
	Block	21.644	3	.000
	Model	21.644	3	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	224.638	.052	.073
2	217.790	.085	.120
3	213.028	.108	.152

Classification Table[a]

			Predicted BWT1 0	Predicted BWT1 1	Percentage Correct
	Observed		0	1	
Step 3	BWT1	0	121	9	93.1
		1	45	14	23.7
	Overall Percentage				71.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1	PTL	1.035	.332	9.750	1	.002	2.816	1.470	5.393
	Constant	-1.074	.188	32.49	1	.000	.342		
Step 2	PTL	1.137	.339	11.25	1	.001	3.117	1.604	6.056
	HT	1.693	.661	6.562	1	.010	5.434	1.488	19.843
	Constant	-1.217	.202	36.37	1	.000	.296		
Step 3	PTL	1.075	.345	9.725	1	.002	2.930	1.491	5.758
	HT	1.852	.666	7.730	1	.005	6.372	1.727	23.509
	UI	.967	.439	4.863	1	.027	2.630	1.114	6.212
	Constant	-1.373	.220	38.92	1	.000	.253		

Note that race is not appearing as significant whereas Race1 appear as significant in set variables not in the equation. It happens so as the race race2 is not significant.

Variables not in the Equation

			Score	df	Sig.
Step 3	Variables	AGE	2.354	1	.125
		LWT	3.771	1	.052
		RACE	5.495	2	.064
		RACE(1)	5.492	1	.019
		RACE(2)	1.190	1	.275
		SMOKE	.857	1	.355
		FTY	.242	1	.623

Note that in automatic process Race1 is not appearing as significant variable

Interpretation of the Coefficients

1. Hypertension (HT)

The odds ratio for hypertension is 6.4 and code for hypertension is high, therefore hypertensive mothers have 6.4 times more chance of having low weight babies on the average. The confidence limits for this variable 1.727 to 23.509. This does not incude 1, so hypertension plays a significant roll.

2. History of premature labor (PTL)

Since the odds ratio is 2.93, therefore all those cases which have premature labor will have 2.93 times chance of having low birth weight than those who do not have premature labor. The confidence limts for the PTL are 1.491 ~ 3.758 which does not include 1 so this factor plays a significance roll.

3. Presence of Uterine Irritability

The odds ratio for uterine irritability is about 2.6, therefore all those mothers who have problem of uterine irritability will have 2.6 times more chance of having low weight babies at birth.

4. Race

Before the interpretation of the result one should look into the coding system of the race. After the creation of dummy variables the odds ratio for white race is 0.44. If we recall Chapter 6, the code is 1 for white race and the odds ratio is less than 1 therefore *white race* has protection against low birth weight. In simple language *other race* will have babies less than average weight. If we look into the analysis without creating the dummy variables we see that the coefficient of race is positive and the odds ratio for the race is 1.5; code for other race is 3 therefore other race will have the babies low in weight on the average,than the black and white repectively.

The method of multiple regression analysis was also used to analyse this data in Chapter-6 and was found that variables like hypertension,history of premature labor, race, uterine irribability and smoking turned out to be significant. In logistic regression hypertension, premaiure labour, race and uterine irritability are significant factors, where smoking is insignificant. The reason is very simple as multiple regression uses actual birth weight whereas in logistic regression we used binary system for birth weight, therefore some information is lost. It is recommended that logistic regression be used binary data is to be analysed.

Example 9.5:

The variables given in Table 9.7 relate to the study of risk factors associated with ICU mortality. Data were collected at Baystate Medical center, Sprinng Field, Massacuhusetts U.S.A. The primary out come (dependent variable) is vital status (live or dead) at hospital discharge (STA). The major goal of this study was to develop a logistic model to predict the probability of survival to hospital discharges of patients. The variables associated with this study and code sheet are given below. Analyze the data by logistic regression and interpret the results. The data is given at the end of this Chapter. Analyze the data and interpret the result.

Table 9.7

S#	Variable	Code Number	ID
1	vital status	0=live,1=dead	ST
2	Age	Years	AGE
3	Gender	0=male,1=female	GE
4	Race	1=white,2= black,3=other	RA
5	service at ICU	0=medical, 1= surgical	SE
6	Cancer	0=no, 1=yes	CA
7	history of chronic renal failure	0=no, 1= yes	CR
8	infection probable at ICU admission	0=no, 1=yes	IN
9	CPR prior to ICU	0=no,1=yes	CP
10	systolic blood pressure	mmHg	BP
11	heart rate at ICU admission	beat/min	HR
12	previous admission to an ICU within 6 months	No=0, yes=1	PA
13	type of admission	0= elective, 1= emergency	TY
14	long bone, multiple, neck, single area, or hip fracture	0=no, 1=yes	FR

S#	Variable	Code Number	ID
15	PO2 from initial blood	0 if > 6o, 1 if ≤ 60	PO
16	Ph from inital blood gases	0 if ≥ 7.25, 1 if < 7.25	PH
17	POC2 from initial blood gases	0 if ≤ 45, 1 if > 45	PC
18	bicarbonate from initial blood gases	0 if ≥ 18, 1 if < 18	BI
19	creatinine from initial blood gases	0 if ≤2, 1 if >2	CE
20	Level of consciousness at ICU admission	0 = no coma or stupor, 1= deep stupor, 2= coma	LO

Example S9-4

The data will be in columns and a part of the data is as follows:

	age	gen	ra	se	ca	cr	in	cp	bt	hra	pa	ty	fr	po	ph	pc	bi	ce	lo	st
1	27	1	1	0	0	0	1	0	142	88	0	1	0	0	0	0	0	0	0	0
2	59	0	1	0	0	0	1	0	112	80	1	1	0	0	0	0	0	0	0	0
3	77	0	1	1	0	0	0	0	100	70	0	0	0	0	0	0	0	0	0	0
4	54	0	1	0	0	0	0	0	142	103	0	1	1	0	0	0	0	0	0	0
5	87	1	1	1	0	0	1	0	110	154	1	1	0	0	0	0	0	0	0	0
6	69	0	1	0	0	0	1	0	110	132	0	1	0	1	0	0	1	0	0	0
7	63	0	1	1	0	0	1	0	104	66	0	0	0	0	0	0	0	0	0	0
8	30	1	1	0	0	0	0	0	144	110	0	1	0	0	0	0	0	0	0	0
9	35	0	2	0	0	0	0	0	108	60	0	1	0	0	0	0	0	0	0	0
10	70	1	1	1	1	0	0	0	138	103	0	0	0	0	0	0	0	0	0	0

We apply the Binary logistic as follows:

Analyze→ Regression→ Binary Logistic…
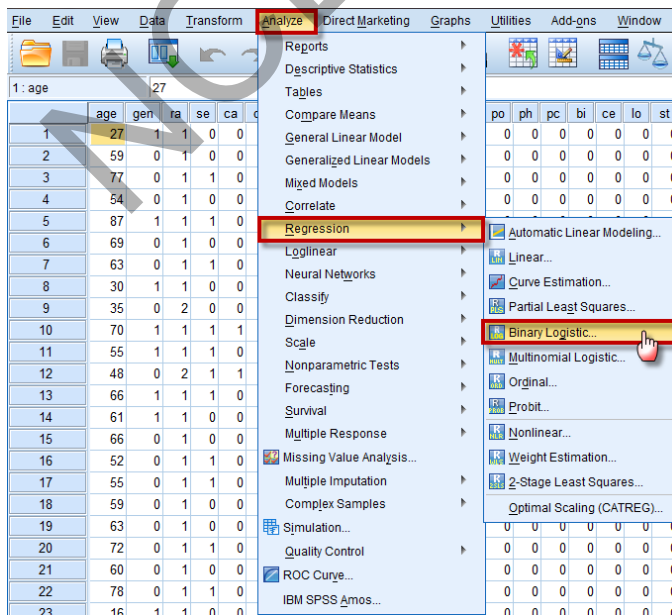
Move the variable "st" to Dependent:

Move all other variables to Covariates:
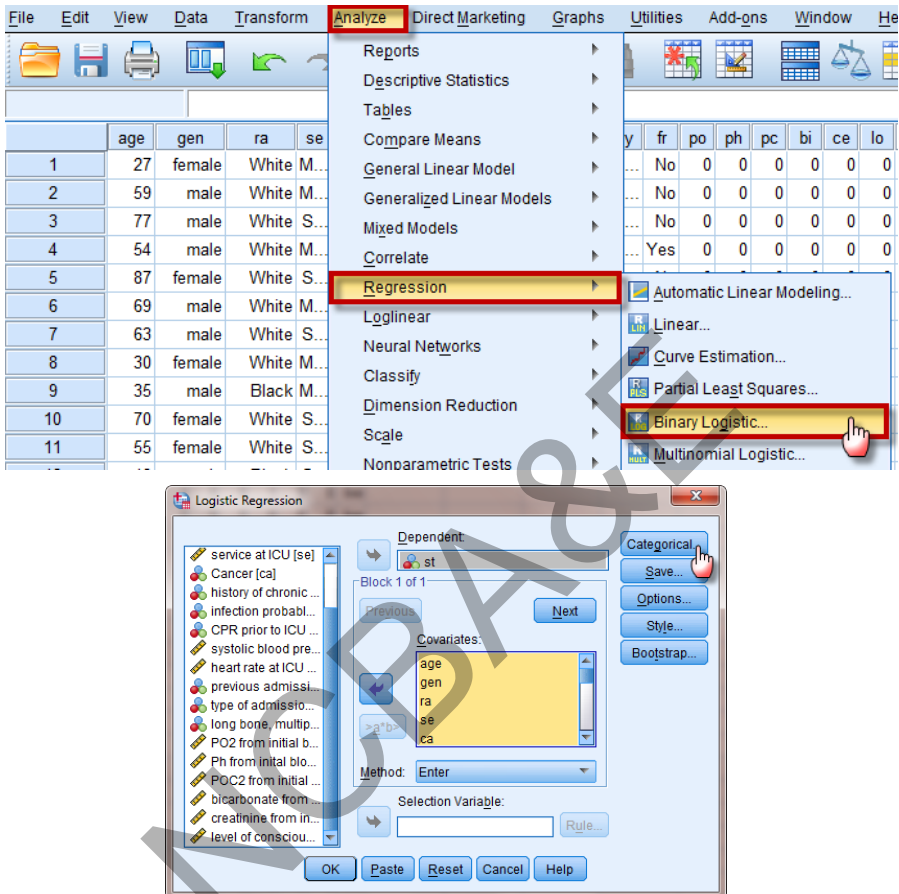
Click on Categorical to specify the categorical variables (i.e with Nominal or Ordinal measurements)

Now click on Method then choose Forward LR (to select the best Model):
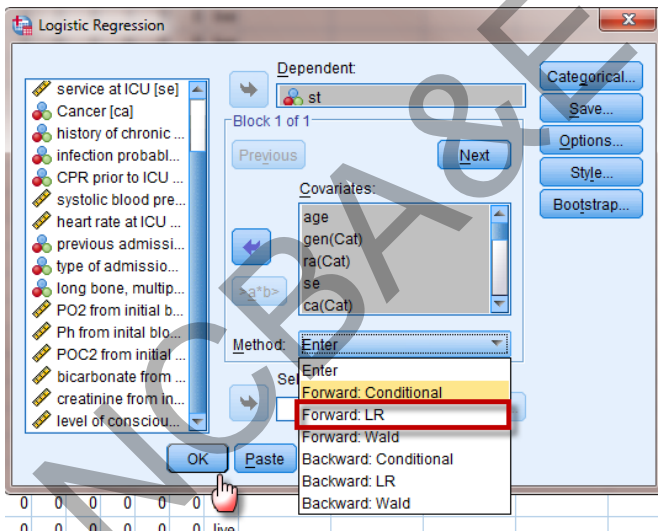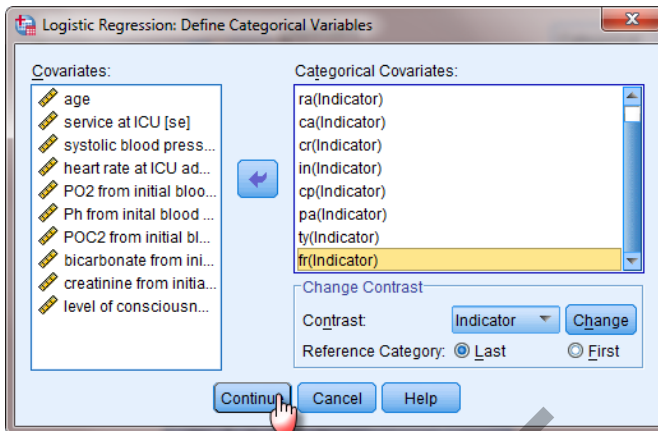
Now click on [OK], to get the following outputs:

SPSS output after the creation of dummy variables by automatic process

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	172.697	.128	.203
2	163.558	.167	.264
3	151.540	.216	.341
4	144.907	.241	.382

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
2	.003	1	.955
3	.167	2	.920
4	5.496	8	.703

Classification Table [a]

			Predicted		
			Vital		Percentage
Observed			0	1	Correct
Vital	0		153	7	95.6
	1		29	11	27.5
Overall					82.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 4	AGE	.028	.012	5.878	1	.015	1.028	1.005	1.052
	CR	1.191	.546	4.756	1	.029	3.291	1.128	9.599
	TY	2.742	1.041	6.945	1	.008	15.526	2.020	119.356
	LO			4.657	2	.097			
	LO(1)	-1.949	.924	4.446	1	.035	.142	.023	.872
	LO(2)	8.517	22.75	.140	1	.708	4998.533	.000	1.2E+23
	Constant	-3.927	1.576	6.212	1	.013	.020		

Interpretation of the variables

Age, history of chronic renal failure (CR), level of consciousness (LO) and type of admission (TY) are appearing as significant variables with odds ratio 1.03,3.3, .016, 0.14 and 15.5. The interpretation of individual variable is given below:

Age

The coefficient is positive and odds ratio is 1.03, therefore as the age increases by one year the chances of death of the patient is increased 1.03 time.

History of chronic renal failure

The coefficient is positive, therefore a patient who is suffering with this problem has more chance of death. The odds ratio is 3.3. The chances of death of the patient suffering from renal failure appears 3.3 times more than those who are not suffering with renal failure.

Type of admission

The coefficients is positive and odds ratio is 15.5, therefore a patient admitted under emergency has 15.5 times more chances of death.

Level of Consciousness at ICU admission

If we look into the coding sheet,low code is for that patient who has no *coma stupor* and high code for the coma patient at the time of admission. We created dummy variables with base zero i.e a patient admitted in the hospital without coma. The odds ratio is 0.142 which is less than 1 and the coefficient is negative. We say that a patient without coma has about 86% chances that he would be discharged alive.

9.4 The Ordinal Regression

Ordinal Regression allows us to model the dependence of a polytomous ordinal response on a set of predictors, which can be factors or covariates. The design of Ordinal Regression is based on the methodology of McCullagh (1980, 1998.

Standard linear regression analysis involves minimizing the sum-of-squared differences between a response (dependent) variable and a weighted combination of predictor (independent) variables. The estimated coefficients reflect how changes in the predictors affect the response. The response is assumed to be numerical, in the sense that changes in the level of the response are equivalent throughout the range of the response. For example, the difference in weight between a person who is 70 kg weight and a person who is 60 kg weight is 10 kg, which has the same meaning as the difference in weight between a person who is 90 kg weight and a person who is 80 kg weight. These relationships do not necessarily hold for ordinal variables, in which the choice and number of response categories can be quite arbitrary.

As an example, Ordinal Regression could be used to study patient reaction to drug dosage. The possible reactions may be classified as *none*, *mild*, *moderate*, or *severe*. The difference between a mild and moderate reaction is difficult or impossible to quantify and is based on perception. Moreover, the difference between a mild and moderate response may be greater or less than the difference between a moderate and severe response.

Generalized linear models. An alternative approach uses a generalization of linear regression called a **generalized linear model** to predict *cumulative probabilities* for the categories. With this method, we fit a separate equation for each category of the ordinal dependent variable. Each equation gives a predicted probability of being in the corresponding category or any lower category.

Generalized linear models are a very powerful class of models, which can be used to answer a wide range of statistical questions. The basic form of a generalized linear model is shown in the following equation:

$$link(\gamma_{ij}) = \theta_j - \left[b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} \right]$$

where

link( )	is the link function
γ_{ij}	is the cumulative probability of the jth category for the ith case
θ_j	is the threshold for the jth category
P	is the number of regression coefficients
$x_{i1}...x_{ip}$	are the values of the predictors for the ith case
$\beta_1...\beta_p$	are regression coefficients

Link function. The link function is a transformation of the cumulative probabilities that allows estimation of the model. Five link functions are available, summarized in the following table.

Function	Form	Typical application
Logit	log(x / (1−x))	Evenly distributed categories
Complementary log-log	log(−log(1−x))	Higher categories more probable
Negative log-log	−log(−log(x))	Lower categories more probable
Probit	F−1(x)	Latent variable is normally distributed
Cauchit (inverse Cauchy)	tan(π(x−0.5))	Latent variable has many extreme values

Note: If we didn't chose the link function then the default is the (logit)

Example 9.6:
Data for a study done to predict a baby's weight category, given various medical and personal characteristics for 189 women. From their database, the Birth Weight Category is the (dependent) variable, with four ordinal levels: >3500 grams, 3000-3500 grams, 2500-3000 grams , and <2500 grams. Potential predictors consist of various medical and personal characteristics of women, including age, race (white = 1, black = 2, other = 3),); smoking status (yes = 1, no = 0), premature labor (none = 0, yes = 1), hypertension (yes = 1, no = 0), and Uterine Irritability (yes = 1, no = 0)

Example S9-5

The data will be in 7 columns and a part of the data is as follows:

	AGE	RACE	SMOKE	PTL	HT	UI	BWC
1	28	3	1	1	0	1	4
2	29	1	0	0	0	1	4
3	34	2	1	0	1	0	4
4	25	3	0	1	1	0	4
5	25	3	0	0	0	1	4
6	27	3	0	0	0	0	4
7	23	3	0	0	0	1	4
8	24	2	0	1	0	0	4
9	24	3	0	0	1	0	4
10	21	1	1	0	1	0	4

The variable view is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
AGE	Numeric	11	0		None	None	11	Right	Scale	Input
RACE	Numeric	11	0		{1, white}...	None	11	Right	Nominal	Input
SMOKE	Numeric	11	0	Smoke status	{0, No}...	None	11	Right	Nominal	Input
PTL	Numeric	11	0	premature labor	None	None	11	Right	Scale	Input
HT	Numeric	11	0	Hypertension	{0, No}...	None	11	Right	Nominal	Input
UI	Numeric	11	0	Uterine Irritability	{0, No}...	None	11	Right	Nominal	Input
BWC	Numeric	8	0	Birth Weight C...	{1, >3500 gr...	None	12	Right	Ordinal	Target

The target variable is the baby's weight category and we apply the Ordinal regression as follows:

Analyze→ Regression→ Ordinal …
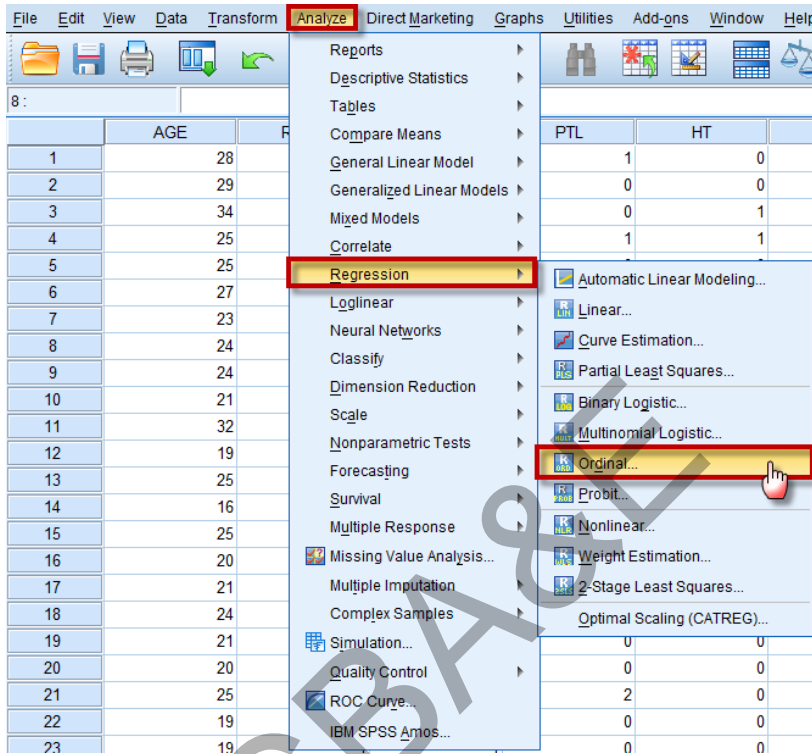
Move the Target ordinal variable "BWC" to Dependent:

Move the Categorical variables to Factors:

Move the Containous variables to Covariate(s):
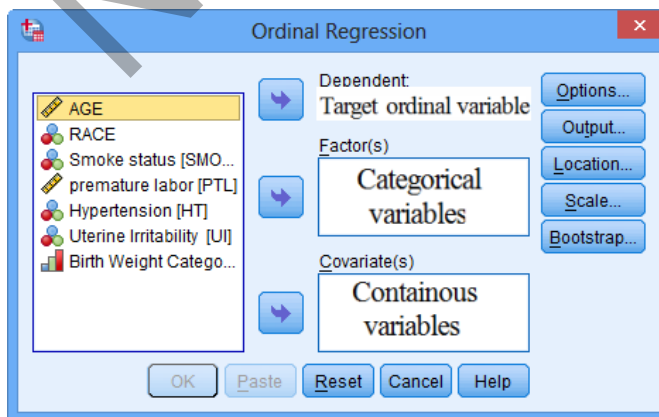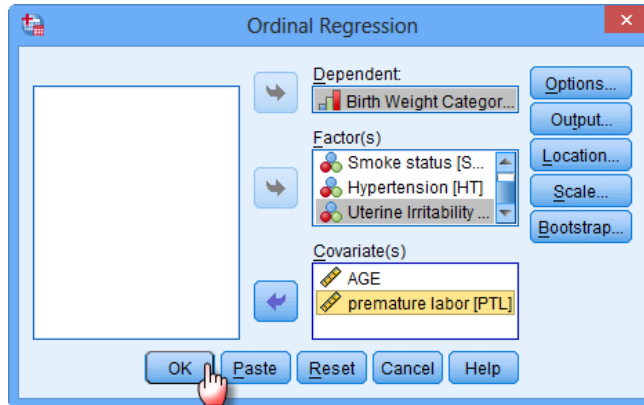
Now click on [OK], to get the following outputs:

SPSS outputs
Case Processing Summary

		N	Marginal Percentage
Birth Weight Category	>3500 grams	46	24.3%
	3000-3500 grams	46	24.3%
	2500-3000 grams	38	20.1%
	<2500 grams	59	31.2%
RACE	white	96	50.8%
	black	26	13.8%
	other	67	35.4%
Smoke status	No	115	60.8%
	Yes	74	39.2%
Hypertension	No	177	93.7%
	Yes	12	6.3%
Uterine Irritability	No	161	85.2%
	Yes	28	14.8%
Valid		189	100.0%
Missing		0	
Total		189	

N -N provides the number of observations fitting the description in the first column. For example, the first four values give the number of observations for which the "Birth Weight Status" is >3500 grams, 3000-3500 grams, 2500-3000 grams and <2500 grams, respectively.

Marginal Percentage - The marginal percentage lists the proportion of valid observations found in each of the outcome variable's groups. This can be calculated by dividing the N for each group by the N for "Valid". Of the 189 subjects with valid data, 46 were categorized as birth weight >3500 grams. Thus, the marginal percentage for this group is (59/189) * 100 = 31.2 %.

Valid - This indicates the number of observations in the dataset where the outcome variable and all predictor variables are non-missing.

Missing - This indicates the number of observations in the dataset where data are missing from the outcome variable or any of the predictor variables.

Total - This indicates the total number of observations in the dataset--the sum of the number of observations in which data are missing and the number of observations with valid data.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	423.904			
Final	386.608	37.295	7	.000

Link function: Logit.

Model - This indicates the parameters of the model for which the model fit is calculated. "Intercept Only" describes a model that does not control for any predictor variables and simply fits an intercept to predict the outcome variable. "Final" describes a model that includes the specified predictor variables and has been arrived at through an iterative process that maximizes the log likelihood of the outcomes seen in the outcome variable. By including the predictor variables and maximizing the log likelihood of the outcomes seen in the data, the "Final" model should improve upon the "Intercept Only" model. This can be seen in the differences in the -2(Log Likelihood) values associated with the models.

-2(Log Likelihood) - This is the product of -2 and the log likelihoods of the null model and fitted "final" model. The likelihood of the model is used to test of whether all predictors' regression coefficients in the model are simultaneously zero and in tests of nested models.

Chi-Square - This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero in the model. The LR Chi-Square statistic can be calculated by -2*L(null model) - (-2*L(fitted model)) = 423.904-386.608= 37.295

df - This indicates the degrees of freedom of the Chi-Square distribution used to test the LR Chi-Square statistic and is defined by the number of predictors in the model.

Sig. - This is the probability of getting a LR test statistic as extreme as, or more so, than the observed under the null hypothesis; the null hypothesis is that all of the regression coefficients in the model are equal to zero. In other words, this is the probability of obtaining this chi-square statistic (37.295) if there is in fact no effect of the predictor variables. This p-value is compared to a specified alpha level, our willingness to accept a type I error, which is typically set at 0.05 or 0.01. The small p-value from the LR test, <0.00001, would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero. The parameter of the Chi-Square distribution used to test the null hypothesis is defined by the degrees of freedom in the prior column.

Pseudo R-Square

Cox and Snell	.179
Nagelkerke	.191
McFadden	.072

Link function: Logit.

Pseudo R-Square - These are three pseudo R-squared values. Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many people have tried to come up with one. There are a wide variety of pseudo R-squared statistics which can give contradictory conclusions. Because these statistics do not mean what R-squared means in OLS regression (the proportion of variance for the response variable explained by the predictors), we suggest interpreting them with great caution.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[BWC = 1]	-4.437	.957	21.487	1	.000	-6.313	-2.561
	[BWC = 2]	-3.199	.931	11.796	1	.001	-5.025	-1.374
	[BWC = 3]	-2.236	.917	5.949	1	.015	-4.033	-.439
Location	AGE	-.010	.027	.141	1	.707	-.062	.042
	PTL	.367	.300	1.500	1	.221	-.220	.955
	[RACE=1]	-1.014	.329	9.500	1	.002	-1.659	-.369
	[RACE=2]	.275	.438	.392	1	.531	-.585	1.134
	[RACE=3]	0[a]	.	.	0	.	.	.
	[SMOKE=0]	-1.093	.312	12.296	1	.000	-1.704	-.482
	[SMOKE=1]	0[a]	.	.	0	.	.	.
	[HT=0]	-1.023	.573	3.181	1	.075	-2.147	.101
	[HT=1]	0[a]	.	.	0	.	.	.
	[UI=0]	-1.023	.408	6.281	1	.012	-1.823	-.223
	[UI=1]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Threshold - This represents the response variable in the ordered logistic regression. The threshold estimate for [birth weight category= 1.00] is the cutoff value between birthweight<2500 and birthweight 2500-3000 grams and the threshold estimate for [birthweight_stauts = 2.00] is the cutoff value between 2500-3000 grams and 3000-3500 grams and so on**.** Underneath **Threshold** are the predictors in the model.

Estimate - These are the ordered log-odds (logit) regression coefficients. Standard interpretation of the ordered logit coefficient is that for a one unit increase in the predictor, the response variable level is expected to change by its respective regression coefficient in the ordered log-odds scale while the other variables in the model are held constant. Interpretation of the ordered logit estimates is not dependent on the ancillary parameters; the ancillary parameters are used to differentiate the adjacent levels of the response variable. However, since the ordered logit model estimates one equation over all levels of the outcome variable, a concern is whether our one-equation model is valid or a more flexible model is required. The odds ratios of the predictors can be calculated by exponentiating the estimate.

Age - This is the ordered log-odds estimate for a one unit increase in **Age** on the expected **birthweight category** given the other variables are held constant in the model. If age of mother were to increase by one point, then ordered log-odds of being in a higher **birthweight** category would increase by 0.001 while the other variables in the model are held constant.

[SMOKE=0] - This is the ordered log-odds estimate of comparing smoking status on expected **birthweight** given the other variables are held constant in the model. The ordered logit for **[SMOKE=0]** being in a higher **birthweight** category is 1.093 more than smoker mothers when the other variables in the model are held constant.

Wald - This is the Wald chi-square test that tests the null hypothesis that the estimate equals 0.

df - These are the degrees of freedom for each of the tests of the coefficients. For each **Estimate** (parameter) estimated in the model, one **df** is required, and the **df** defines the Chi-Square distribution to test whether the individual regression coefficient is zero given the other variables are in the model.

Sig.- These are the p-values of the coefficients or the probability that, within a given model, the null hypothesis that a particular predictor's regression coefficient is zero given that the rest of the predictors are in the model. They are based on the **Wald** test statistics of the predictors, which can be calculated by dividing the square of the predictor's estimate by the square of its standard error. The probability that a particular **Wald** test statistic is as extreme as, or more so, than what has been observed under the null hypothesis is defined by the p-value and presented here. The Wald test statistic for the predictor **age** is 0.002 with an associated p-value of 0.963. If we set our alpha level to 0.05, we would fail to reject the null hypothesis and conclude that the regression coefficient for **age** has not been found to be statistically different from zero in estimating **birthweight_status** given **other predictor(s)** are in the model. The Wald test statistic for the predictor **smoke** is 11.150 with an associated p-value of 0.001. If we set our alpha level to 0.05, we would fail to reject the null hypothesis and conclude that the regression coefficient for **smoke** has been found to be statistically different from zero in estimating **birthweight_status** given **other predictor(s)** are in the model.

95% Confidence Interval - This is the Confidence Interval (CI) for an individual regression coefficient given the other predictors are in the model. For a given predictor with a level of 95% confidence, we'd say that we are 95% confident that the "true" population regression coefficient lies in between the lower and upper limit of the interval.

9.5 The Multinomial Logistic Regression

Linear regression is not appropriate for situations in which there is no natural ordering to the values of the dependent variable. Multinomial Logistic Regression is useful for situations in which we want to be able to classify subjects based on values of a set of predictor variables.. This type of regression is similar to binary logistic regression, but it is more general because the dependent variable is not restricted to two categories.

For a dependent variable with k categories, consider the existence of k unobserved continuous variables, $Z_1, ... Z_K$, each of which can be thought of as the "propensity toward" a category. In the case of a many categories to chose from, Z_k represents a customer's propensity toward selecting the kth category, with larger values of Z_k corresponding to greater probabilities of choosing that category (assuming all other Z's remain the same).

We will generalize the binary logistic regression model to the case of multinomial

logistic regression model. This has been defined before and is as:

$$\text{Prob(event } Z_{ik}) = \frac{e^{Z_{ik}}}{\sum_{j=1}^{k} e^{Z_{ij}}} = \frac{e^{Z_{ik}}}{e^{Z_{i1}} + \cdots + e^{Z_{ij}} + \cdots + e^{Z_{ik}}} \quad ; i=1,2,...,k$$

where:

Prob(event Z_{ik}) is the probability the ith case falls in category k

Z_{ik} is the value of the kth unobserved continuous variable for the ith case

X_{ij} is the jth predictor for the ith case

b_{kj} is the jth coefficient for the kth unobserved variable

j is the number of predictors

If Z_k were observable, we would simply fit a linear regression to each Z_k and be done. However, since Z_k is unobserved, we must relate the predictors to the probability of interest by substituting for Z_k.

Z_k is also assumed to be linearly related to the predictors.

$Z_{ik} = b_{k0} + b_{k1}X_{i1} + b_{k2}X_{i2} + ... + b_{kj}X_{ij}$

Example 9.7:

In order to market new drugs, pharmacy copmany want to predict what to test two new drugs in compare with an old one. By performing a Multinomial Logistic Regression, the company can determine the strength of influence a person's age, gender, and marital status has upon the type of drug they used. The company can then slant the advertising campaign of a particular drug toward a group of people likely to use it. The variables given in Table 9.8 relate to the study: The variables associated with this study code and data sheet are given below. Analyze the data by logistic regression and interpret the results.

Table 9.8

S#	Variable	Code Number	ID
1	Age in years	Years	Age
2	Sex	1=male,2=female	Sex
3	Marital status	1=Unmarried,2= Married	Marital
4	Drug	1= Regular drug,2=Drug A, 3=Dug B	Drug

Age	Sex	Marital	Drug	Age	Sex	Marital	Drug	Age	Sex	Marital	Drug
50	1	2	3	59	1	2	3	26	2	2	3
23	2	2	3	70	2	2	2	61	1	2	2
30	2	1	3	62	1	2	2	41	2	2	2
44	1	2	3	30	1	1	1	67	1	2	3
32	2	1	1	25	1	1	1	44	1	2	3
65	1	2	2	61	2	1	2	28	1	2	3
36	2	2	1	28	2	2	3	29	1	1	1
39	2	2	1	48	2	2	2	52	2	2	2
46	2	2	3	66	2	1	2	22	2	1	3

Example S9-6

The data will be in 4 columns as follows:

Age	Sex	Marital	Drug
50	1	2	3
23	2	2	3
30	2	1	3
44	1	2	3
32	2	1	1
65	1	2	2
36	2	2	1
39	2	2	1
46	2	2	3
59	1	2	3
70	2	2	2
62	1	2	2
30	1	1	1
25	1	1	1
61	2	1	2
28	2	2	3
48	2	2	2
66	2	1	2
26	2	2	3
61	1	2	2
41	2	2	2
67	1	2	3
44	1	2	3
28	1	2	3
29	1	1	1
52	2	2	2
22	2	1	3

Age	Sex	Marital	Drug
50	Male	Married	Drug B
23	Female	Married	Drug B
30	Female	Unmarried	Drug B
44	Male	Married	Drug B
32	Female	Unmarried	Regular drug
65	Male	Married	Drug A
36	Female	Married	Regular drug
39	Female	Married	Regular drug
46	Female	Married	Drug B
59	Male	Married	Drug B
70	Female	Married	Drug A
62	Male	Married	Drug A
30	Male	Unmarried	Regular drug
25	1	Male	Unmarried
61	2	Female	Unmarried
28	2	Female	Married
48	2	Female	Married
66	2	Female	Unmarried
26	2	Female	Married
61	1	Male	Married
41	Female	Married	Drug A
67	Male	Married	Drug B
44	Male	Married	Drug B
28	Male	Married	Drug B
29	Male	Unmarried	Regular drug
52	Female	Married	Drug A
22	Female	Unmarried	Drug B

The target variable is the Drug and we apply the Multinomial Binary logistic as follows:
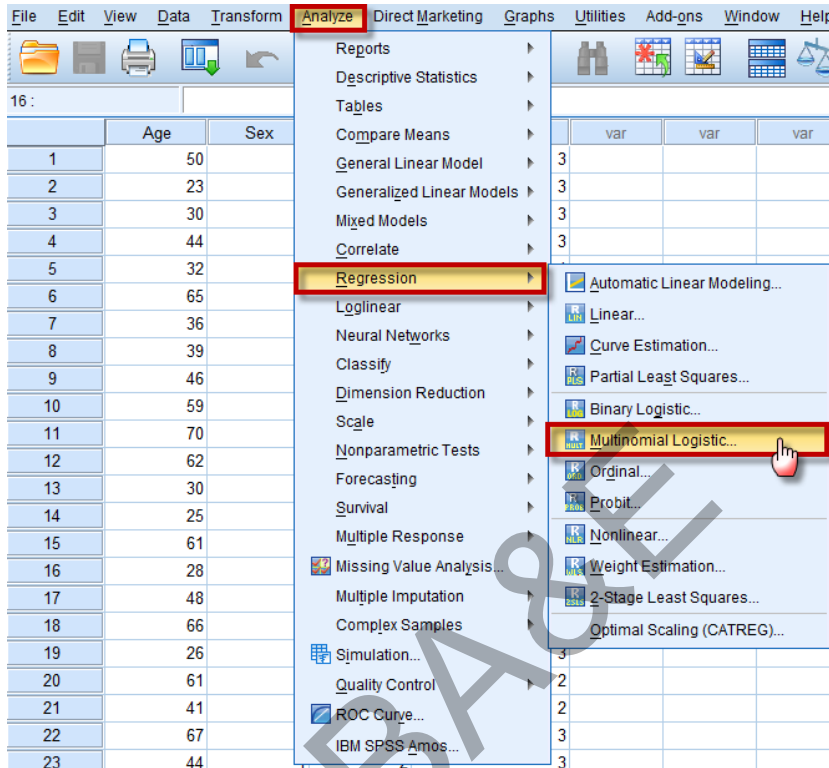
Analyze→ Regression→ Multinomial Logistic…

Move the variable "Drug" to Dependent:

Move the categorical variables to Factors:

Move the containous variable (Age) to Covariate(s):
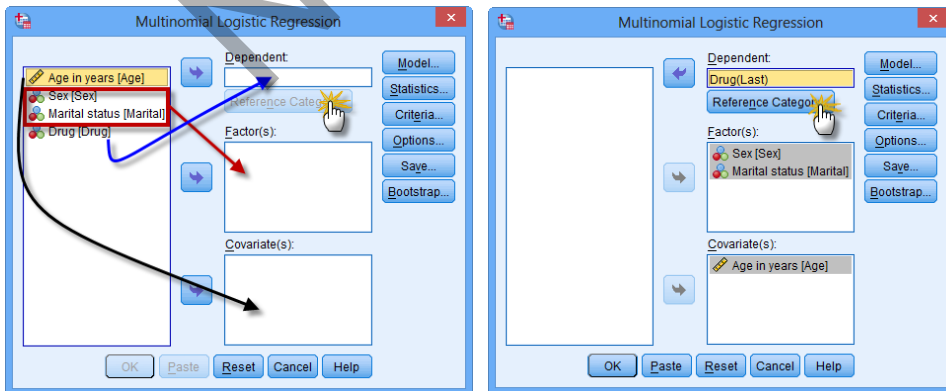
Click on Reference Category to specify the reference category (First Category, which is the Regular drug)
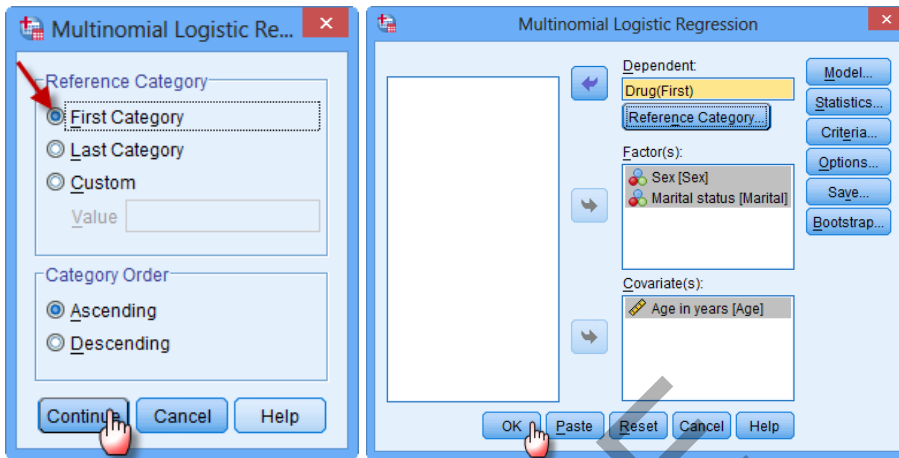
Now click on [OK], to get the following outputs:

SPSS output after the creation of dummy variables by automatic process
Case Processing Summary

		N	Marginal Percentage
Drug	Regular drug	6	22.2%
	Drug A	9	33.3%
	Drug B	12	44.4%
Sex	Male	12	44.4%
	Female	15	55.6%
Marital status	Unmarried	8	29.6%
	Married	19	70.4%
Valid		27	100.0%
Missing		0	
Total		27	
Subpopulation		26[a]	

According to the case processing summary, the modal category is the new Drug B, with 44.4% of the cases.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	57.286			
Final	32.164	25.122	6	.000

This is a likelihood ratio test of our model (Final) against one in which all the parameter coefficients are 0 (Null). The chi-square statistic is the difference between the -2 log-likelihoods of the Null and Final models.

Since the significance level of the test is less than 0.05, we can conclude the Final model is outperforming the Null.

Pseudo R-Square

Cox and Snell	.606
Nagelkerke	.688
McFadden	.439

Pseudo R-Squared Statistics. The r-squared statistic, which measures the variability in the dependent variable that is explained by a linear regression model, cannot be computed for multinomial logistic regression models. The pseudo r-squared statistics are designed to have similar properties to the true r-squared statistic.

In the linear regression model, the coefficient of determination, R^2, summarizes the proportion of variance in the dependent variable associated with the predictor (independent) variables, with larger R^2 values indicating that more of the variation is explained by the model, to a maximum of 1. For regression models with a categorical dependent variable, it is not possible to compute a single R^2 statistic that has all of the characteristics of R^2 in the linear regression model, so these approximations are computed instead. The following methods are used to estimate the coefficient of determination:

Cox and Snell's R^2 (Cox and Snell, 1989) is based on the log likelihood for the model compared to the log likelihood for a baseline model. However, with categorical outcomes, it has a theoretical maximum value of less than 1, even for a "perfect" model. Nagelkerke's R^2 (Nagelkerke, 1991) is an adjusted version of the Cox & Snell R-square that adjusts the scale of the statistic to cover the full range from 0 to 1. McFadden's R2 (McFadden, 1974) is another version, based on the log-likelihood kernels for the intercept-only model and the full estimated model. What constitutes a "good" R^2 value varies between different areas of application. While these statistics can be suggestive on their own, they are most useful when comparing competing models for the same data. The model with the largest R^2 statistic is "best" according to this measure, which is given here by Nagelkerke.

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	32.164[a]	.000	0	.
Age	51.582	19.418	2	.000
Sex	38.322	6.158	2	.046
Marital	34.940	2.776	2	.250

The likelihood ratio tests check the contribution of each effect to the model. For each effect, the -2 log-likelihood is computed for the reduced model; that is, a model without the effect. The chi-square statistic is the difference between the -2 log-likelihoods of the reduced model from this table and the Final model reported in the model fitting information table. If the significance of the test is small (less than 0.05) then the effect contributes to the model. And since the significance of the test is less than 0.001, we can say that the effect contributes to the model.

Some effects can be difficult to test. For example, the intercept cannot be tested in this model because removing the intercept simply causes one of the previously redundant factor levels to become non-redundant.

Parameter Estimates

Drug[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
Drug A	Intercept	-10.253	5.796	3.130	1	.077			
	Age	.278	.134	4.259	1	.039	1.320	1.014	1.718
	[Sex=1]	-4.774	2.867	2.773	1	.096	.008	3.063E-005	2.328
	[Sex=2]	0[b]	.	.	0	.	.	.	.
	[Marital=1]	-2.276	4.507	.255	1	.614	.103	1.499E-005	704.429
	[Marital=2]	0[b]	.	.	0	.	.	.	.
Drug B	Intercept	1.137	2.518	.204	1	.652			
	Age	.013	.065	.040	1	.841	1.013	.892	1.151
	[Sex=1]	-.106	1.266	.007	1	.933	.899	.075	10.746
	[Sex=2]	0[b]	.	.	0	.	.	.	.
	[Marital=1]	-2.145	1.383	2.407	1	.121	.117	.008	1.759
	[Marital=2]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Regular drug.

b. This parameter is set to zero because it is redundant.

The parameter estimates table summarizes the effect of each predictor. The ratio of the coefficient to its standard error, squared, equals the Wald statistic. If the significance level of the Wald statistic is small (less than 0.05) then the parameter is different from 0. Age is the only signficant. The odds ratio with its confidence intervals was also given.

• Notes: Parameters with significant **negative** coefficients **decrease** the likelihood of that response category with respect to the reference category. Parameters with **positive** coefficients **increase** the likelihood of that response category. The parameters associated with the last category of each factor is redundant given the intercept term.

Example S9-7

We will add a Married male case of age of 55 to the data and apply the Multinomial logistic and get the predicted values directly, also we will see how to calculate the correct percentage for the prediction as follows:

Analyze→ Regression→ Multinomial Logistic…

Move the variable "Drug" to Dependent:

Move the categorical variables to Factors:

Move the containous variable (Age) to Covariate(s):

Click on [Save...] and choose "Estimate response probabilites" and "Pridicted category", as follows:
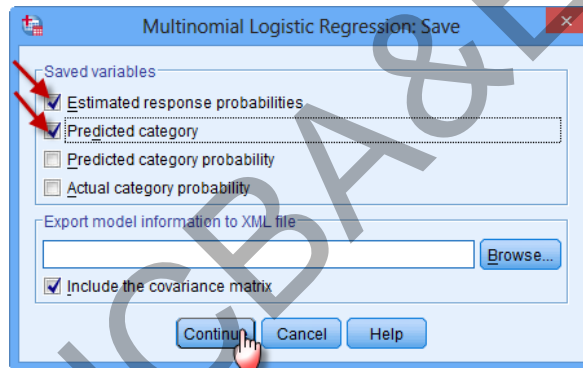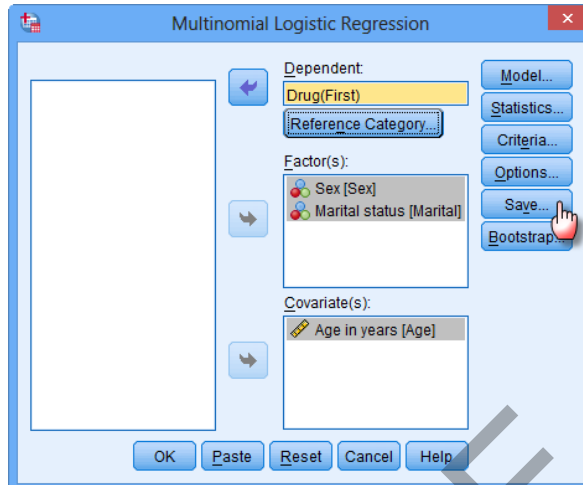
Now click on Continue then OK, to find out the predicted values added to the data directly, as:

	Age	Sex	Marital	Drug	EST1_1	EST2_1	EST3_1	PRE_1	var
22	67	1	2	3	.02	.82	.16	2	
23	44	1	2	3	.17	.01	.82	3	
24	28	1	2	3	.20	.00	.80	3	
25	29	1	1	1	.68	.00	.32	1	
26	52	2	2	2	.01	.90	.08	2	
27	22	2	1	3	.67	.00	.33	1	
28	55	1	2		.12	.16	.72	3	
29									

It can be seen that a Married male case of age of 55 is predicted to prefer Drug B.

Note also that using the Cross tabulation between the Predicted Response Category and the actual Drug will lead to the following crosstabulation table:

Predicted Response Category * Drug Crosstabulation

			Drug			Total
			Regular drug	Drug A	Drug B	
Predicted Response Category	Regular drug	Count	4	0	2	6
		% of Total	14.8%	0.0%	7.4%	22.2%
	Drug A	Count	0	8	2	10
		% of Total	0.0%	29.6%	7.4%	37.0%
	Drug B	Count	2	1	8	11
		% of Total	7.4%	3.7%	29.6%	40.7%
Total		Count	6	9	12	27
		% of Total	22.2%	33.3%	44.4%	100.0%

According to this table, we can see that the correct prediction percentage for this model equals to 14.8+29.6+29.6 = 74.0%

APPENDIX

age	gen	ra	se	ca	cr	in	cp	bp	hra	pa	ty	fr	po	ph	pc	bi	ce	lo	st
27	1	1	0	0	0	1	0	142	88	0	1	0	0	0	0	0	0	0	0
59	0	1	0	0	0	1	0	112	80	1	1	0	0	0	0	0	0	0	0
77	0	1	1	0	0	0	0	100	70	0	0	0	0	0	0	0	0	0	0
54	0	1	0	0	0	0	0	142	103	0	1	1	0	0	0	0	0	0	0
87	1	1	1	0	0	1	0	110	154	1	1	0	0	0	0	0	0	0	0
69	0	1	0	0	0	1	0	110	132	0	1	0	1	0	0	1	0	0	0
63	0	1	1	0	0	1	0	104	66	0	0	0	0	0	0	0	0	0	0
30	1	1	0	0	0	0	0	144	110	0	1	0	0	0	0	0	0	0	0
35	0	2	0	0	0	0	0	108	60	0	1	0	0	0	0	0	0	0	0
70	1	1	1	1	0	0	0	138	103	0	0	0	0	0	0	0	0	0	0
55	1	1	1	0	0	0	0	188	86	1	0	0	0	0	0	0	0	0	0
48	0	2	1	1	0	0	0	162	100	0	0	0	0	0	0	0	0	0	0
66	1	1	1	0	0	1	0	160	80	1	0	0	0	0	0	0	0	0	0
61	1	1	0	0	0	0	0	174	99	0	1	0	0	1	0	1	1	0	0
66	0	1	0	0	0	0	0	206	90	0	1	0	0	0	0	0	1	0	0
52	0	1	1	0	0	1	0	150	71	1	0	0	0	0	0	0	0	0	0
55	0	1	1	0	0	1	0	140	116	0	0	0	0	0	0	0	0	0	0
59	0	1	0	0	0	1	0	48	39	0	1	0	1	0	1	1	0	2	0
63	0	1	0	0	0	0	0	132	128	1	1	0	0	0	0	0	0	0	0
72	0	1	1	0	0	0	0	120	80	1	0	0	0	0	0	0	0	0	0
60	0	1	0	0	0	1	1	114	110	0	1	0	0	0	0	0	0	0	0
78	0	1	1	0	0	0	0	180	75	0	0	0	0	0	0	0	0	0	0
16	1	1	0	0	0	0	0	104	111	0	1	0	0	0	0	0	0	0	0
62	0	1	1	0	1	0	0	200	120	0	0	0	0	0	0	0	0	0	0
61	0	1	0	0	0	1	0	110	120	0	1	0	0	0	0	0	0	0	0
35	0	1	0	0	0	0	0	150	98	0	1	0	0	0	0	0	0	0	0
74	1	1	1	0	0	0	0	170	92	0	0	0	0	0	1	0	0	0	0
68	0	1	1	0	0	0	0	158	96	0	0	0	0	0	0	0	0	0	0
69	1	1	1	0	0	0	0	132	60	0	1	0	0	0	0	0	0	0	0
51	0	1	0	0	0	0	0	110	99	0	1	0	0	0	0	0	0	0	0
55	0	3	1	0	0	0	0	128	92	0	0	0	0	0	0	0	0	0	0
64	1	1	1	0	0	1	0	158	90	1	1	0	0	0	0	0	0	0	0
88	1	1	1	0	0	1	0	140	88	1	1	1	0	0	0	0	0	0	0
23	1	1	1	0	0	0	0	112	64	0	1	0	0	0	0	0	0	0	0
73	1	1	1	1	0	0	0	134	60	0	0	0	0	0	1	0	0	0	0
53	0	3	1	0	0	0	0	110	70	1	0	0	0	0	0	0	0	0	0
74	0	1	1	0	0	0	0	174	86	0	0	0	0	0	0	0	0	0	0
68	0	1	1	0	0	0	0	142	89	0	0	0	0	0	0	0	0	0	0
66	1	1	0	0	0	1	0	170	95	1	1	0	0	0	0	0	0	0	0
60	0	1	1	1	0	1	0	110	92	0	0	0	0	0	0	0	0	0	0
64	0	1	1	0	0	1	0	160	120	0	0	0	0	0	0	0	0	0	0
66	0	2	1	1	0	1	0	150	120	0	0	0	0	0	1	0	0	0	0

age	gen	ra	se	ca	cr	in	cp	bp	hra	pa	ty	fr	po	ph	pc	bi	ce	lo	st
19	1	1	1	0	0	1	0	142	106	0	1	1	0	0	0	0	0	0	0
18	1	1	0	0	0	0	0	146	112	0	1	0	0	0	0	0	0	0	0
63	0	1	1	0	0	1	0	162	84	1	1	0	0	0	0	0	0	0	0
45	0	1	0	0	0	0	0	126	110	0	1	0	0	0	0	0	0	0	0
64	0	1	0	0	0	0	0	162	114	0	1	0	0	0	0	0	0	0	0
68	1	1	0	0	0	0	0	200	170	1	1	0	0	0	0	0	0	0	0
64	0	1	0	0	0	0	0	126	122	0	1	0	1	0	1	0	0	0	0
82	0	1	1	0	0	0	0	135	70	0	0	0	0	0	0	0	0	0	0
73	0	1	1	0	0	0	0	170	88	0	0	0	0	0	0	0	0	0	0
70	0	1	0	0	0	0	0	86	153	1	1	0	0	0	1	0	0	0	0
61	0	1	1	0	0	1	0	68	124	0	1	0	0	0	0	0	0	0	0
64	0	1	1	1	0	1	0	116	88	0	0	0	0	0	0	0	0	0	0
47	0	1	1	1	0	1	0	120	83	0	0	0	0	0	0	0	0	0	0
69	0	1	1	0	0	0	0	170	100	0	0	0	0	0	0	0	0	0	0
67	1	1	0	0	0	1	0	190	125	0	1	0	0	0	0	0	0	0	0
18	0	1	1	1	0	0	0	156	99	0	0	0	0	0	0	0	0	0	0
77	0	1	1	0	0	1	0	158	107	0	0	0	0	0	0	0	0	0	0
32	0	2	1	0	0	0	0	120	84	0	1	0	0	0	0	0	0	0	0
19	1	1	1	0	0	1	0	104	121	1	0	0	0	0	0	0	0	0	0
72	1	1	1	0	0	0	0	130	86	0	1	0	0	0	0	0	0	0	0
49	0	1	0	0	0	1	0	112	112	0	1	0	0	0	0	0	0	0	0
68	1	1	1	0	0	0	0	154	74	0	0	0	0	0	0	0	0	0	0
82	0	1	1	0	1	1	0	130	131	0	1	0	0	0	0	0	0	0	0
32	1	3	0	0	0	1	1	110	118	0	1	0	0	0	0	0	0	0	0
78	1	1	1	0	0	1	0	126	96	0	1	0	0	0	0	0	0	0	0
57	0	1	0	0	0	1	0	128	104	0	1	0	0	0	1	0	0	0	0
46	1	1	1	1	0	0	0	132	90	0	1	0	0	0	0	0	0	0	0
23	0	1	0	0	0	1	0	144	88	0	1	0	0	0	0	0	0	0	0
55	0	1	0	0	0	0	0	132	112	0	1	0	0	0	0	0	0	0	0
18	0	1	1	0	0	0	0	112	76	0	1	1	0	0	0	0	0	0	0
20	0	1	1	0	0	0	0	164	108	0	1	0	0	0	0	0	0	0	0
75	1	1	1	0	0	0	0	100	48	0	0	0	0	0	0	0	0	0	0
79	0	1	0	0	0	1	0	112	67	0	0	0	0	0	0	0	0	0	0
40	0	1	1	0	0	0	0	140	65	0	1	1	0	0	0	0	0	0	0
76	0	1	1	0	0	1	0	110	70	0	1	0	0	0	0	0	0	0	0
66	1	1	1	0	0	1	0	139	92	0	0	0	0	0	0	0	0	0	0
76	0	1	0	0	0	1	0	190	100	0	1	0	0	0	0	0	0	0	0
80	1	1	1	0	0	0	0	162	44	0	1	0	0	0	0	0	0	0	0
23	1	1	0	0	0	1	0	120	88	0	1	0	0	0	0	0	0	0	0
48	0	2	1	0	0	1	0	92	162	1	1	0	0	0	0	0	0	0	0
67	0	2	1	0	0	0	0	90	92	1	0	0	0	0	0	0	0	0	0
69	1	1	1	0	0	0	0	150	85	0	1	0	0	0	0	0	0	0	0
65	0	3	1	0	0	0	0	208	124	0	0	0	0	0	0	0	0	0	0
72	0	1	1	0	0	0	0	126	88	0	0	0	0	0	0	0	0	0	0

age	gen	ra	se	ca	cr	in	cp	bp	hra	pa	ty	fr	po	ph	pc	bi	ce	lo	st
55	0	1	0	0	0	0	0	190	136	0	1	0	1	1	1	0	0	0	0
40	0	1	0	0	0	0	0	130	65	0	1	0	0	0	0	0	0	0	0
55	1	1	0	0	0	1	0	110	86	0	1	0	0	0	0	0	0	0	0
34	0	1	1	0	0	0	0	110	80	0	1	1	0	0	0	0	0	0	0
47	1	1	1	0	0	0	0	132	68	0	1	0	0	0	0	0	0	0	0
41	1	1	0	0	0	1	0	118	145	0	1	0	0	1	0	1	0	0	0
84	1	1	0	0	1	1	0	100	103	0	1	0	0	0	0	1	1	0	0
88	1	1	1	0	0	0	0	110	46	1	0	0	0	0	0	0	0	0	0
77	1	1	1	1	0	0	0	212	87	0	0	0	0	0	1	0	0	0	0
80	0	1	0	0	0	0	0	122	126	0	1	0	1	0	0	1	0	0	0
16	0	1	1	0	0	0	0	100	140	0	1	1	0	0	0	0	0	0	0
70	0	1	1	0	0	0	0	160	60	0	0	0	0	0	0	0	0	0	0
83	1	1	1	0	0	1	0	138	91	0	1	0	0	0	0	0	0	0	0
23	0	2	0	0	0	0	0	130	52	0	1	0	0	0	0	0	0	0	0
67	1	1	0	0	0	0	1	120	120	0	1	0	0	1	1	0	0	0	0
18	0	1	1	1	0	0	0	130	140	0	0	0	0	0	0	0	0	0	0
77	1	1	0	0	0	1	0	136	138	0	0	0	1	1	1	0	0	0	0
48	1	1	0	0	0	0	1	128	96	0	1	0	0	0	0	0	0	0	0
24	1	2	0	0	0	0	0	140	86	0	1	0	0	0	0	0	0	0	0
71	1	1	0	0	0	1	0	124	106	0	1	0	0	0	0	0	0	0	0
72	0	1	1	0	0	0	0	134	60	0	1	0	0	0	0	0	0	0	0
77	1	1	1	0	1	0	0	170	115	1	0	0	0	0	0	0	0	0	0
60	0	1	1	0	0	1	0	124	135	0	1	0	0	0	0	0	0	0	0
46	0	1	1	1	0	0	0	110	128	0	0	0	0	0	0	0	0	0	0
65	1	1	0	0	0	0	0	100	105	0	1	0	0	0	0	0	0	0	0
36	0	1	0	0	0	0	0	224	125	0	1	0	0	0	0	0	0	0	0
68	0	1	1	0	0	0	0	112	64	0	0	0	0	0	0	0	0	0	0
58	0	1	0	0	0	0	0	154	98	0	1	0	0	0	0	0	0	0	0
76	1	1	0	0	0	1	0	92	112	0	1	0	0	0	0	0	0	0	0
41	1	2	0	0	0	0	0	110	144	0	1	0	0	0	0	1	1	0	0
20	0	3	0	0	0	0	0	120	68	0	1	0	0	0	0	0	0	0	0
91	0	1	0	1	1	1	0	152	125	0	1	0	0	0	0	0	0	0	0
75	0	1	1	0	0	0	0	140	90	0	1	0	0	0	0	0	0	0	0
25	1	1	0	0	0	0	0	131	135	0	1	0	0	0	0	1	0	0	0
70	0	1	0	0	0	1	0	78	143	0	1	0	1	0	0	0	0	0	0
47	0	1	1	0	0	0	0	156	112	0	1	0	0	0	0	0	0	0	0
75	0	3	1	0	0	0	0	144	120	0	1	0	0	0	0	0	1	0	0
40	0	2	0	0	0	1	0	160	150	1	1	1	0	0	0	0	0	0	0
71	0	1	0	0	0	1	0	148	192	0	1	0	1	1	1	0	0	0	0
70	1	1	0	0	0	1	0	90	140	0	1	0	1	0	0	1	0	0	0
58	0	1	1	0	0	0	0	148	95	1	1	0	0	0	0	0	0	0	0
54	0	1	1	0	0	0	0	136	80	0	0	0	0	0	0	0	0	0	0
77	0	1	1	0	0	0	0	128	59	0	0	0	0	0	0	0	0	0	0
55	0	1	1	1	0	1	0	138	140	0	0	0	0	0	0	0	0	0	0

age	gen	ra	se	ca	cr	in	cp	bp	hra	pa	ty	fr	po	ph	pc	bi	ce	lo	st
21	0	1	1	0	0	0	0	120	62	0	1	0	0	0	0	0	0	0	0
53	0	2	0	0	1	0	1	170	115	0	1	0	0	0	0	0	0	0	0
31	1	1	0	1	1	1	1	146	100	0	1	0	0	1	1	0	0	0	0
71	0	1	1	1	0	0	0	204	52	0	0	0	1	0	0	0	0	0	0
49	0	2	0	0	0	0	0	150	100	0	1	1	0	0	0	0	0	0	0
60	1	2	0	0	0	1	0	116	92	1	1	0	0	0	0	0	0	0	0
50	0	1	0	0	0	1	0	156	99	0	1	0	0	0	1	0	0	0	0
45	1	1	1	0	0	0	0	132	109	0	1	1	0	0	0	0	0	0	0
21	0	1	1	0	0	0	0	110	90	0	1	1	0	0	0	0	0	0	0
73	1	1	1	0	0	0	0	130	83	0	1	0	0	0	0	0	0	0	0
28	0	1	1	0	0	1	0	122	80	1	0	0	0	0	0	0	0	0	0
17	0	1	1	0	0	0	0	140	78	0	1	0	0	0	0	0	0	0	0
17	1	3	0	0	0	0	0	130	140	0	1	0	0	0	0	0	0	0	0
21	1	1	1	0	0	0	0	142	79	0	1	0	0	0	0	0	0	0	0
68	1	1	1	0	0	0	0	91	79	0	0	0	0	0	0	0	0	0	0
17	0	3	1	1	0	0	0	136	78	0	1	0	0	0	0	0	0	0	0
60	0	1	0	0	0	1	0	108	120	0	0	0	0	0	0	0	0	0	0
69	0	1	1	0	0	0	0	169	73	0	1	0	0	0	0	0	0	0	0
88	1	1	0	0	1	0	0	190	88	0	1	0	0	0	0	0	0	0	0
20	0	1	1	0	0	0	0	120	80	0	1	0	0	0	0	0	0	0	0
89	1	1	1	0	0	0	0	190	114	0	1	0	0	0	1	0	0	2	0
62	1	1	0	0	0	0	0	110	78	0	1	0	0	0	0	0	0	0	0
46	0	1	0	0	1	1	0	142	89	0	1	0	0	1	0	1	0	0	0
19	0	1	1	0	1	1	0	100	137	0	1	0	0	0	0	0	0	0	0
71	0	1	0	0	1	1	0	124	124	0	1	0	1	1	1	0	0	0	0
67	0	1	1	0	0	0	0	152	78	0	0	0	0	0	0	0	0	0	0
20	0	1	1	0	0	0	0	104	83	0	1	0	0	0	0	0	0	0	0
73	1	2	0	0	0	0	0	162	100	0	1	0	0	0	0	0	0	0	0
59	0	1	0	0	0	0	0	100	88	0	1	0	0	0	0	0	0	0	0
42	0	1	1	0	0	0	0	122	84	0	1	1	0	0	0	0	0	0	0
87	1	1	1	0	1	1	0	80	96	0	1	1	1	1	1	0	0	0	1
76	1	1	1	0	0	1	0	128	90	1	1	0	0	0	0	0	0	0	1
78	0	1	0	0	0	1	0	130	132	0	1	0	0	0	0	1	0	0	1
63	0	1	0	1	1	1	0	112	106	1	1	0	1	0	0	0	0	0	1
19	0	1	1	0	0	0	0	140	76	0	1	0	0	0	0	0	0	0	1
67	1	1	0	0	0	1	0	62	145	0	1	0	0	0	0	0	1	0	1
53	1	1	0	0	0	1	0	148	128	0	1	0	0	1	1	0	0	0	1
92	0	1	0	0	0	1	0	124	80	0	1	0	0	0	0	1	0	0	1
57	0	1	0	0	0	1	1	110	124	0	1	0	0	0	0	0	0	2	1
75	1	1	1	0	0	0	0	130	136	0	0	0	0	0	0	0	0	0	1
91	0	1	0	0	0	1	0	64	125	0	1	0	0	0	1	0	0	0	1
70	0	1	1	0	0	0	0	168	122	0	0	0	1	0	0	0	0	1	1
88	0	1	0	0	0	1	1	141	140	0	1	0	0	0	0	0	0	0	1
41	0	1	1	0	0	1	0	140	58	0	1	0	0	0	0	0	0	2	1

age	gen	ra	se	ca	cr	in	cp	bp	hra	pa	ty	fr	po	ph	pc	bi	ce	lo	st
61	0	1	0	0	0	0	0	140	81	1	1	0	0	0	0	0	0	0	1
80	0	1	1	0	0	0	0	100	85	0	1	0	0	0	0	0	0	0	1
40	0	1	0	0	0	1	0	86	80	0	1	0	0	0	0	0	0	0	1
75	0	1	0	0	0	1	0	90	100	0	1	0	0	0	0	0	0	1	1
63	1	1	1	0	1	1	1	36	86	0	1	1	0	0	0	0	1	2	1
75	1	1	0	1	0	0	0	190	94	0	1	0	0	0	0	0	0	0	1
20	0	1	1	0	0	0	0	148	72	0	1	1	0	0	0	0	0	0	1
71	0	1	0	0	0	0	0	142	95	0	1	0	0	0	0	0	0	0	1
51	1	1	1	0	0	1	0	134	100	1	1	0	0	0	0	0	0	1	1
65	0	1	0	0	0	0	0	66	94	0	1	0	0	0	0	0	0	2	1
69	1	3	0	0	1	0	0	170	60	1	1	0	1	0	0	0	0	0	1
55	0	1	1	0	1	1	0	122	100	1	1	0	0	0	0	0	0	0	1
50	1	1	1	1	0	0	0	120	96	0	1	0	0	0	0	0	0	0	1
78	0	1	0	0	0	1	0	110	81	0	1	0	0	0	0	0	0	0	1
71	1	1	0	0	0	0	1	70	112	0	1	0	0	0	0	0	0	0	1
85	1	1	1	0	1	0	0	136	96	0	1	0	0	0	0	0	0	0	1
75	0	1	0	0	0	1	0	130	119	0	1	0	0	1	0	1	1	1	1
65	1	1	0	0	0	1	1	104	150	0	1	0	0	0	1	0	0	0	1
49	0	1	0	0	1	1	1	140	108	0	1	0	0	0	0	1	0	0	1
75	1	1	0	0	0	1	1	150	66	0	1	0	0	0	0	0	1	1	1
72	1	1	0	0	0	0	0	90	160	0	1	0	0	0	0	0	0	0	1
69	0	1	0	0	1	0	0	80	81	0	1	0	0	0	0	0	0	0	1
64	0	1	0	1	0	1	0	80	118	0	1	0	1	0	0	0	1	1	1
60	0	1	0	0	0	1	0	56	114	1	1	0	0	1	0	1	0	0	1
60	0	3	1	0	1	1	0	130	55	0	1	0	0	0	0	0	0	0	1
50	1	2	0	0	0	0	0	256	64	0	1	0	0	0	0	0	0	0	1

id	AGE	RACE	SMOKE	PTL	HT	UI	BWC
1	28	3	1	1	0	1	4
2	29	1	0	0	0	1	4
3	34	2	1	0	1	0	4
4	25	3	0	1	1	0	4
5	25	3	0	0	0	1	4
6	27	3	0	0	0	0	4
7	23	3	0	0	0	1	4
8	24	2	0	1	0	0	4
9	24	3	0	0	1	0	4
10	21	1	1	0	1	0	4
11	32	1	1	0	0	0	4
12	19	1	1	2	0	1	4
13	25	3	0	0	0	0	4
14	16	3	0	0	0	0	4
15	25	1	1	0	0	0	4
16	20	1	1	0	0	0	4
17	21	2	0	0	0	1	4
18	24	1	1	1	0	0	4
19	21	3	0	0	0	0	4
20	20	3	0	0	0	1	4
21	25	3	0	2	0	0	4
22	19	1	0	0	0	0	4
23	19	1	1	0	0	1	4
24	26	1	1	1	0	0	4
25	24	1	0	0	0	0	4
26	17	3	1	1	0	1	4
27	20	2	1	0	0	0	4
28	22	1	1	1	0	1	4
29	27	2	0	0	0	1	4
30	20	3	1	0	0	1	4
31	17	1	1	0	0	0	4
32	25	3	0	1	0	0	4
33	20	3	0	0	0	0	4
34	18	3	0	0	0	0	4
35	18	2	1	1	0	0	4
36	20	1	1	1	0	1	4
37	21	3	0	1	0	0	4
38	26	3	0	0	0	0	4
39	31	1	1	1	0	0	4
40	15	1	0	0	0	0	4
41	23	2	1	0	0	0	4
42	20	2	1	0	0	0	4
43	24	2	1	0	0	0	4
44	15	3	0	0	0	1	4
45	23	3	0	0	0	0	4
46	30	1	1	1	0	0	4
47	22	1	1	0	0	0	4

id	AGE	RACE	SMOKE	PTL	HT	UI	BWC
48	17	1	1	0	0	0	4
49	23	1	1	1	0	0	4
50	17	2	0	0	0	0	4
51	26	3	0	1	1	0	4
52	20	3	0	0	0	0	4
53	26	1	1	0	0	0	4
54	14	3	1	1	0	0	4
55	28	1	1	0	0	0	4
56	14	3	0	0	0	0	4
57	23	3	1	0	0	0	4
58	17	2	0	0	1	0	4
59	21	1	1	0	1	0	4
60	19	2	0	0	0	1	3
61	33	3	0	0	0	0	3
62	20	1	1	0	0	0	3
63	21	1	1	0	0	1	3
64	18	1	1	0	0	1	3
65	21	3	0	0	0	0	3
66	22	1	0	0	0	0	3
67	17	3	0	0	0	0	3
68	29	1	1	0	0	0	3
69	26	1	1	0	0	0	3
70	19	3	0	0	0	0	3
71	19	3	0	0	0	0	3
72	22	3	0	0	1	0	3
73	30	3	0	1	0	1	3
74	18	1	1	0	0	0	3
75	18	1	1	0	0	0	3
76	15	2	0	0	0	0	3
77	25	1	1	0	0	0	3
78	20	3	0	0	0	1	3
79	28	1	1	0	0	0	3
80	32	3	0	0	0	0	3
81	31	1	0	0	0	1	3
82	36	1	0	0	0	0	3
83	28	3	0	0	0	0	3
84	25	3	0	0	0	1	3
85	28	1	0	0	0	0	3
86	17	1	1	0	0	0	3
87	29	1	0	0	0	0	3
88	26	2	1	0	0	0	3
89	17	2	0	0	0	0	3
90	17	2	0	0	0	0	3
91	24	1	1	1	0	0	3
92	35	2	1	1	0	0	3
93	25	1	0	0	0	0	3
94	25	2	0	0	0	0	3
95	29	1	1	0	0	0	3

id	AGE	RACE	SMOKE	PTL	HT	UI	BWC
96	19	1	1	0	0	0	3
97	27	1	1	0	0	0	3
98	31	1	1	0	0	0	2
99	33	1	1	0	0	0	2
100	21	2	1	0	0	0	2
101	19	1	0	0	0	0	2
102	23	2	0	0	0	0	2
103	21	1	0	0	0	0	2
104	18	1	1	0	0	1	2
105	18	1	1	0	0	1	2
106	32	1	0	0	0	0	2
107	19	3	0	0	0	0	2
108	24	1	0	0	0	0	2
109	22	3	1	0	0	0	2
110	22	1	0	0	1	0	2
111	23	3	0	0	0	0	2
112	22	1	1	0	0	0	2
113	30	1	1	0	0	0	2
114	19	3	0	0	0	0	2
115	16	3	0	0	0	0	2
116	21	3	1	0	0	1	2
117	30	3	0	0	0	0	2
118	20	3	0	0	0	0	2
119	17	3	0	0	0	0	2
120	17	3	0	0	0	0	2
121	23	3	0	0	0	0	2
122	24	3	0	0	0	0	2
123	28	1	0	0	0	0	2
124	26	3	1	2	0	0	2
125	20	3	0	1	0	1	2
126	24	3	0	0	0	0	2
127	28	3	1	0	0	0	2
128	20	1	0	2	0	1	2
129	22	2	0	1	0	0	2
130	22	1	1	2	0	0	2
131	31	3	1	0	0	0	2
132	23	3	1	0	0	0	2
133	16	2	0	0	0	0	2
134	16	1	1	0	0	0	2
135	18	2	0	0	0	0	2
136	25	1	0	0	0	0	2
137	32	1	1	1	0	0	2
138	20	2	1	0	0	0	2
139	23	1	0	0	0	0	2
140	22	1	0	0	0	0	2
141	32	1	0	0	0	0	2
142	30	3	0	0	0	0	2
143	20	3	0	0	0	0	2

id	AGE	RACE	SMOKE	PTL	HT	UI	BWC
144	23	3	0	0	0	0	1
145	17	3	1	0	0	0	1
146	19	3	0	0	0	0	1
147	23	1	0	0	0	0	1
148	36	1	0	0	0	0	1
149	22	1	0	0	0	0	1
150	24	1	0	0	0	0	1
151	21	3	0	0	0	0	1
152	19	1	1	0	1	0	1
153	25	1	1	3	0	1	1
154	16	1	1	0	0	0	1
155	29	1	0	0	0	0	1
156	29	1	0	0	0	0	1
157	19	1	1	0	0	0	1
158	19	1	1	0	0	0	1
159	30	1	0	0	0	0	1
160	24	1	0	0	0	0	1
161	19	1	1	0	1	0	1
162	24	3	0	1	0	0	1
163	23	1	0	0	0	0	1
164	20	3	0	0	0	0	1
165	25	2	0	0	1	0	1
166	30	1	0	0	0	0	1
167	22	1	0	0	0	0	1
168	18	1	1	0	0	0	1
169	16	2	0	0	0	0	1
170	32	1	0	0	0	0	1
171	18	3	0	0	0	0	1
172	29	1	1	0	0	0	1
173	33	1	0	0	0	1	1
174	20	1	1	0	0	0	1
175	28	3	0	0	0	0	1
176	14	1	0	0	0	0	1
177	28	3	0	0	0	0	1
178	25	1	0	0	0	0	1
179	16	3	0	0	0	0	1
180	20	1	0	0	0	0	1
181	26	3	0	0	0	0	1
182	21	1	0	0	0	0	1
183	22	1	0	0	0	0	1
184	25	1	0	0	0	0	1
185	31	1	0	0	0	0	1
186	35	1	0	1	0	0	1
187	19	1	1	0	0	0	1
188	24	1	0	0	0	0	1
189	45	1	0	0	0	0	1

Answering a Statistical Question

Question 1	(Measurment level)		
	(Scale)	(Ordinal)	(Nominal)
How can we represent Data?	(Histogram), (Line) (Curve) (Boxplot) (Error bar) (Scatter plot)	(Bars), (Pie)	

Question 2	(Measurment level)		
	(Scale)	(Ordinal)	(Nominal)
How can we describe the variable?	(Mean), (Standard deviation)	(Median), (Interquartile range)	(Mode), (Proportions)

Question 3		(Measurement level)		
		(Scale)	(Ordinal)	(Nominal)
Is there is a relation between variables?	(Scale)	(Pearson)	(Ordinal Bi-serial)	(Point Bi-serial) (Eta)
	(Ordinal)	(Ordinal Bi-serial)	(Kendall) (Spearman) (Gamma)	(Bi-serial)
	(Nominal)	(Point Bi-serial) (Eta)	(Bi-serial)	(Phi) (Contingency Coefficient) (Lambda)

Note that we can use SPSS to calculate each of Ordinal Bi-serial, Point Bi-serial, Bi-serial, by the same way we calculate Pearson correlation coefficient.

Question 4	Measurment level of Dependent variable		
	(Scale)	(Ordinal)	(Nominal)
How Can we predict?	(Liner Regression) (Nonliear Regression)	(Ordinal regression)	(Logistic regression)

Question 5	(Measurment level)		
	(Scale) 📏	**(Ordinal)**	**(Nominal)**
How Can we Estimate?	**(CI for Mean)**	**(CI for Median)**	**(CI for Proportion)**

Question 6		Measurment level of Dependent variable		
		Scale 📏	**Ordinal**	**Nominal**
		Scale (from Normal Population)	**Rank, or Scale (from Non-Normal Population)**	**Binomial (Two Possible Outcomes)**
Is there a difference between groups?	**1 group**	**(One sample t test)**	**(Wilcoxon test)**	**(Binomial test)**
	2 independent groups	**(Independent sample t test)**	**(Mann-Whitney)**	**(Chi-square test)**
	2 matched groups	**(Paired sample t test)**	**(Wilcoxon test)**	**(McNemar)**
	3+ independent groups	**(One-way ANOVA)**	**(Kruskal-Wallis)**	**(Chi-square test)**
	3+ matched groups	**(Repeated Measurements)**	**(Friedman test)**	**(Chi-square test)**

Selecting a Statistical test using SPSS

Goal	Measurement level of Dependent variable		
	Scale (from Normal Population)	**Rank or Scale (from Non-Normal Population)**	**Binomial (Two Possible Outcomes)**
Describe one Group	Analyze→ Descriptive Statistics→ Descriptives…	Analyze→ Descriptive Statistics→ Frequencies → Statistics…	Analyze→ Descriptive Statistics→ Frequencies → Statistics…
Compare one group to a hypothetical value	Analyze→ Compare means→ One-sample T Test…	Analyze→ Nonparametric Tests→ 2 Related samples → Wilcoxon (after adding median value as the 2nd variable)	Analyze→ Nonparametric Tests→ Binomial…
Compare two unpaired groups	Analyze→ Compare means→ Independent-sample T Test…	Analyze→ Nonparametric Tests→ 2 Independent samples → Mann-Whitney U	Analyze→ Descriptive Statistics→ Crosstabs…→ Statistics…→ Chi-square
Compare two paired groups	Analyze→ Compare means→ Paired-sample T Test…	Analyze→ Nonparametric Tests→ 2 Related samples → Wilcoxon	Analyze→ Nonparametric Tests→ 2 Related samples → McNemar
Compare three or more unmatched groups	Analyze→ Compare means→ One-Way ANOVA…	Analyze→ Nonparametric Tests→ k Independent samples → Kruskal-Wallis H	Analyze→ Descriptive Statistics→ Crosstabs…→ Statistics…→ Chi-square
Compare three or more matched groups	Analyze→ General Linear Model→ Repeated Measures…	Analyze→ Nonparametric Tests→ k Related samples → Friedman	Analyze→ Nonparametric Tests→ k Related samples → Cochran's Q
Quantify association between two groups	Analyze→ Correlate→ Bivariate → Pearson	Analyze→ Correlate→ Bivariate → Spearman	Analyze→ Descriptive Statistics→ Crosstabs…→ Statistics…→ Contingency coefficient
Predict value from another measured variable	Analyze→ Regression→ Linear…	Analyze→ Regression→ Ordinal …	Analyze→ Regression→ Binary Logistic…
Predict value from several measured or binomial variables	Analyze→ Regression→ Linear… (chose more variables)	Analyze→ Regression→ Ordinal …	Analyze→ Regression→ Multinomial Logistic…

Chapter 10

Survival Analysis

10.1 Introduction

Study data may be collected in many different ways. In addition to surveys, which are cross-sectional, biomedical research data may come from different sources.

*The two fundamental designs being **retrospective** and **prospective**.*

Retrospective studies gather past data from selected cases and controls to determine differences, if any, in exposure to a suspected risk factor.

They are commonly referred to as case–control studies; each case–control study is focused on a particular disease.

In a typical case–control study, cases of a specific disease are ascertained as they arise from population-based registers or lists of hospital admissions, and controls are sampled either as disease-free persons from the population at risk or as hospitalized patients having a diagnosis other than the one under study.

The advantages of a retrospective study are that it is economical and provides answers to research questions relatively quickly because the cases are already available. Major limitations are due to the inaccuracy of the exposure histories and uncertainty about the appropriateness of the control sample; these problems sometimes hinder retrospective studies and make them less preferred than prospective studies.

Prospective studies, also called **cohort studies**, are epidemiological designs in which one enrolls a group of persons and follows them over certain periods of time; examples include occupational mortality studies and clinical trials.

The cohort study design focuses on a particular exposure rather than a particular disease as in case–control studies. Advantages of a longitudinal approach include the opportunity for more accurate measurement of exposure history and a careful examination of the time relationships between exposure and any disease under investigation.

An important subset of cohort studies consists of randomized clinical trials where follow-up starts from the date of enrollment and randomization of each subject.

Basic survival analysis and Cox's proportional hazards regression—were developed to deal with survival data resulting from <u>prospective</u> or cohort studies.
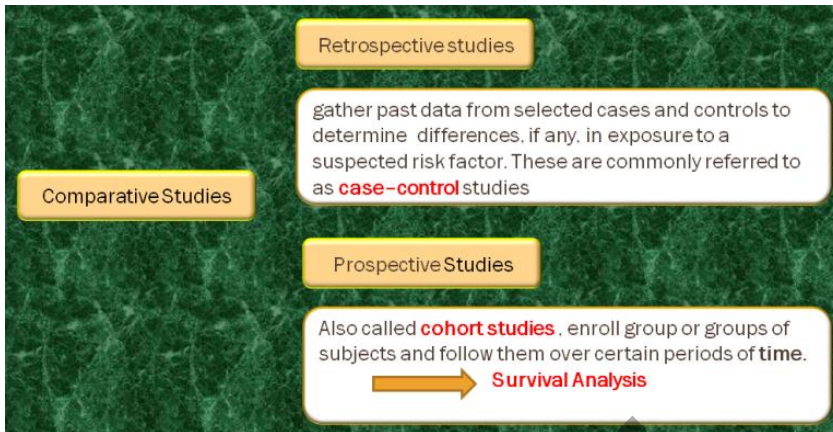
Survival analysis, which was developed to deal with data resulting from prospective studies, is also focused on the occurrence of an event, such as death or relapse of a disease, after some initial treatment—a binary outcome.

The basic difference with the logistic regression analysis is that:

a- For survival data, studies have staggered entry, and subjects are followed for varying lengths of time; they do not have the same probability for the event to occur even if they have identical characteristics, a basic assumption of the logistic regression model.

b- Second, each member of the cohort belongs to one of three types of termination:

1. Subjects still <u>alive</u> on the analysis date

2. Subjects who <u>died</u> on a known date within the study period

3. Subjects who are <u>lost</u> to follow-up after a certain date (This is known as Censoring).

That is, for many study subjects, the observation may be terminated before the occurrence of the main event under investigation: for example, subjects in types 1 and 3.

10.2 Survival analyses

Survival analyses or time to event analyses are frequently used in medical sciences where the interest is in observing time to death either of patients or of laboratory animals. There are certain aspects of survival analysis data, such as censoring and non-normality, that cause great difficulty when trying to analyze the data using traditional statistical methods such as t-test, ANOVA and linear regression etc.

A censored observation is defined as an observation with incomplete information. There are four different types of censoring possible: right truncation, left truncation, right censoring and left censoring.

Right truncation occurs when the entire study population has already experienced the event of interest (for example: a historical survey of patients on a cancer registry).

Left truncation occurs when the subjects have been at risk before entering the study (for example: a life insurance policy holder where the study starts on a fixed date, event of interest is age at death).

Right censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred. For example, we consider patients in a clinical trial to study the effect of treatments on stroke occurrence. The study ends after 5 years. Those patients who have had no strokes by the end of the year are censored. If the patient leaves the study at time te ; then the event occurs in (te,∞).

Left censoring is when the event of interest has already occurred before registration in study. This is very rarely encountered.

In this chapter we will focus exclusively on right censoring for a number of reasons. Most data used in analyses have only right censoring. Furthermore, right censoring is the most easily understood of all the four types of censoring and if a researcher can understand the concept of right censoring thoroughly it becomes much easier to understand the other three types. When an observation is right censored it means that the information is incomplete because the subject did not have an event during the time that the subject was part of the study. The point of survival analysis is to follow subjects over time and observe at which point in time they experience the event of interest. It often happens that the study does not span enough time in order to observe the event for all the subjects in the study. This could be due to a number of reasons. Perhaps subjects drop out of the study for reasons unrelated to the study (i.e. patients moving to another area and leaving no forwarding address). The common feature of all of these examples is that if the subject had been able to stay in the study then it would have been possible to observe the time of the event eventually.

Outcome Variable: Time until an Event occurs
Start Follow-up → Time → Event
      Event of Interest: Death
           Disease
            Relapse
              Recovery

years in study
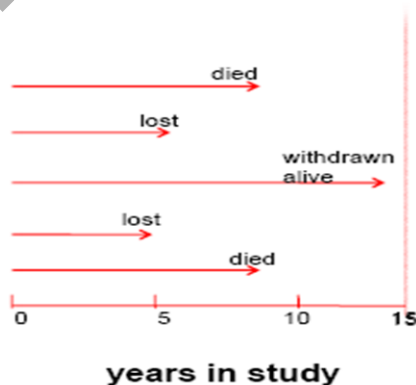
Censoring: Don't know survival time exactly

Reasons of Censoring?

  Study ends – no event
    Lost of follow-up
      Withdraws

Two Key Quantities of interest in survival analysis

1. S(t) = survivor function
2. h(t) = hazard function

Survivorship or Survival Function S(t)

Survivorship or Survival Function, S(t), is the probability that an individual's time, T, is greater than a specified time, t. In mathematical terms:

  S(t) = Prob(survives longer than t)
    = Prob(T>t)
    = 1 - F(t)

where F(t) is the cumulative distribution function of T.

Hazard Function h(t):

Hazard Function, h(t), is the conditional failure rate. It is the probability of failure during a small time interval given that the individual has survived until the beginning of the interval.

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left(t \le T < t + \Delta t \mid T \ge t\right)}{\Delta t}$$

Relationship between S(t) & h(t)
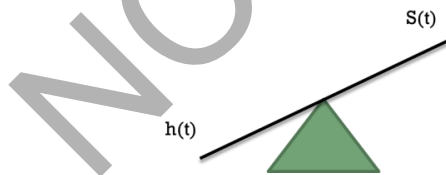
$$S(t) = e^{-\int_0^t h(u)du} \quad , \quad h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]$$

Life Table Analysis

A life table presents the proportion surviving, the cumulative hazard function, and the hazard rates of a large group of subjects followed over time. The analysis accounts for subjects who die (fail) as well as subjects who are censored (withdrawn). The life-table method competes with the Kaplan- Meier product-limit method as a technique for survival analysis. The life-table method was developed first, but the Kaplan-Meier method has been shown to be superior and with the advent of computers is now the method of choice. However, for large samples, the life-table method is still popular in that it provides a simple summary of a large set of data.

Example 10.1:

We will give a brief introduction to the subject in this section. For a complete account of life- table analysis, we suggest the books by Lee (1992) and Elandt-Johnson and Johnson (1980).

Lee (1992) constructs a life table. The survival experience of 2418 males with angina is recorded in years. The life table will use 16 intervals of one year each. (1=Events and 0=Censored).

Time	Event	Count	Time	Event	Count
0.5	1	456	1.5	0	39
1.5	1	226	2.5	0	22
2.5	1	152	3.5	0	23
3.5	1	171	4.5	0	24
4.5	1	135	5.5	0	107
5.5	1	125	6.5	0	133
6.5	1	83	7.5	0	102
7.5	1	74	8.5	0	68
8.5	1	51	9.5	0	64
9.5	1	42	10.5	0	45
10.5	1	43	11.5	0	53
11.5	1	34	12.5	0	33
12.5	1	18	13.5	0	27
13.5	1	9	14.5	0	23
14.5	1	6	15.5	0	30

The IBM-SPSS package is used as shown in the following Example:

Example S10-1

The data will be in 3 columns and a part of the data is as follows:

	Time	Event	Count			Time	Event	Count
1	.5	1	456		1	.5	Events	456
2	1.5	1	226		2	1.5	Events	226
3	2.5	1	152		3	2.5	Events	152
4	3.5	1	171		4	3.5	Events	171
5	4.5	1	135		5	4.5	Events	135
6	5.5	1	125		6	5.5	Events	125
7	6.5	1	83		7	6.5	Events	83
8	7.5	1	74		8	7.5	Events	74
9	8.5	1	51		9	8.5	Events	51
10	9.5	1	42		10	9.5	Events	42
11	10.5	1	43		11	10.5	Events	43
12	11.5	1	34		12	11.5	Events	34
13	12.5	1	18		13	12.5	Events	18
14	13.5	1	9		14	13.5	Events	9
15	14.5	1	6		15	14.5	Events	6
16	1.5	0	39		16	1.5	Censored	39
17	2.5	0	22		17	2.5	Censored	22
18	3.5	0	23		18	3.5	Censored	23
19	4.5	0	24		19	4.5	Censored	24
20	5.5	0	107		20	5.5	Censored	107

We first weight data by count as follows:

Data → Weight Case
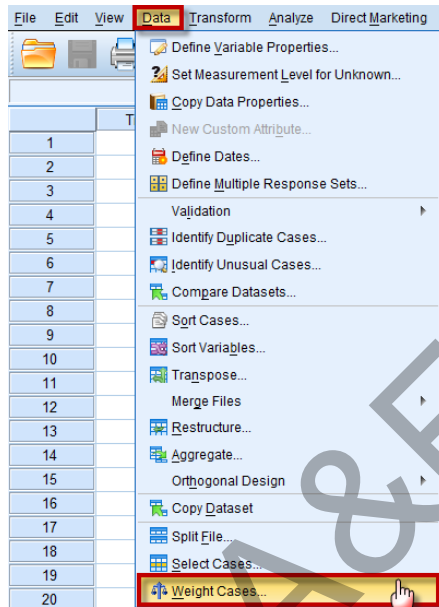
Weight Cases by: Select "count" as frequency variable
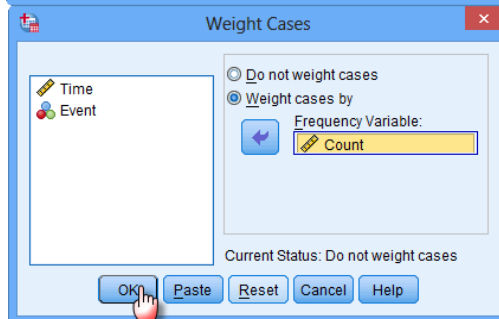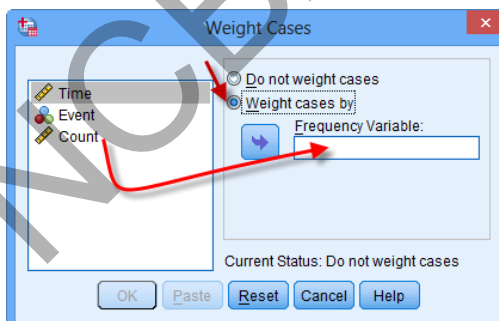
Now click on **OK**, to start Survival analysis as follows:

Analyze→ Survival→ Life Tables…
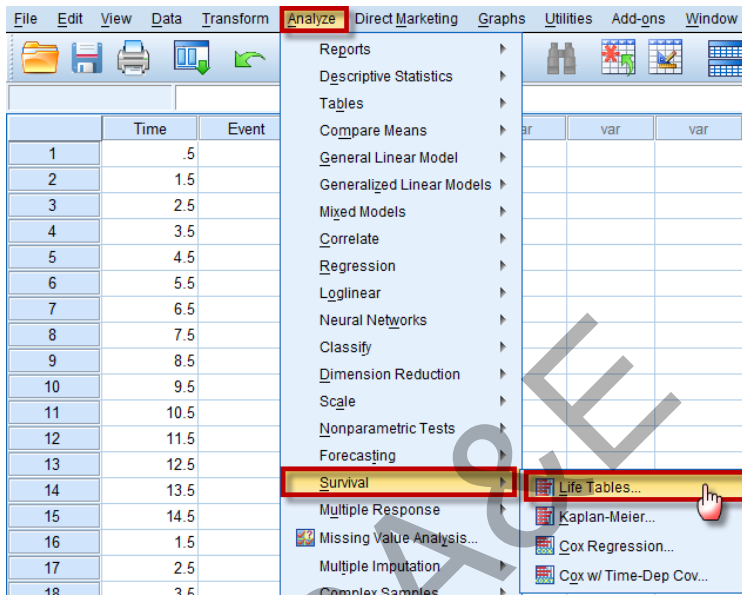
Move the Survival Time Variable (time) to Time

For: Display Time Intervals we define it from 0 through 16 by 1

Click on "Define Event" , mark on "Single value:" put 1

Click on "Continue"
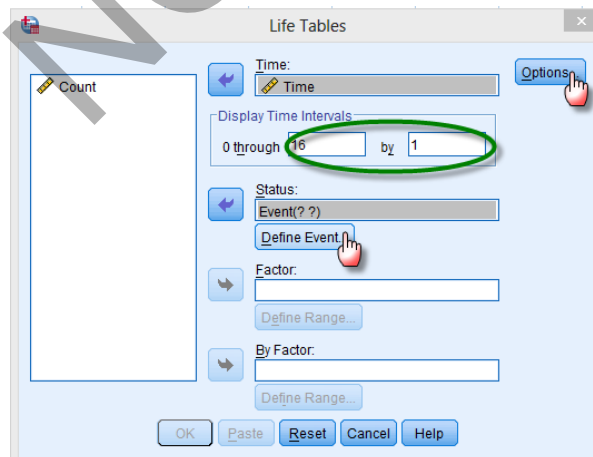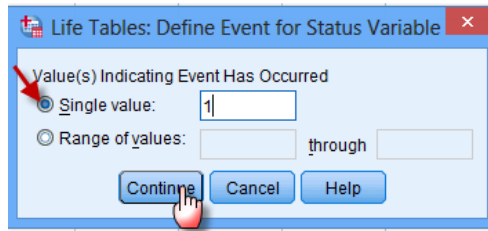
Also click on "Options" and mark on "Life Table"

For "Plot", mark on Survival, Hazard and Density

Now click on Continue then OK, to get the following outputs:

Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Number of Terminal Events	Proportion Terminating	Proportion Surviving	Cumulative Proportion Surviving at End of Interval	Std. Error of Cumulative Proportion Surviving at End of Interval	Probability Density	Std. Error of Probability Density	Hazard Rate	Std. Error of Hazard Rate
0	2418	0	2418.000	456	.19	.81	.81	.01	.189	.008	.21	.01
1	1962	39	1942.500	226	.12	.88	.72	.01	.094	.006	.12	.01
2	1697	22	1686.000	152	.09	.91	.65	.01	.065	.005	.09	.01
3	1523	23	1511.500	171	.11	.89	.58	.01	.074	.005	.12	.01
4	1329	24	1317.000	135	.10	.90	.52	.01	.059	.005	.11	.01
5	1170	107	1116.500	125	.11	.89	.46	.01	.058	.005	.12	.01
6	938	133	871.500	83	.10	.90	.42	.01	.044	.005	.10	.01
7	722	102	671.000	74	.11	.89	.37	.01	.046	.005	.12	.01
8	546	68	512.000	51	.10	.90	.33	.01	.037	.005	.10	.01
9	427	64	395.000	42	.11	.89	.30	.01	.036	.005	.11	.02
10	321	45	298.500	43	.14	.86	.26	.01	.043	.006	.16	.02
11	233	53	206.500	34	.16	.84	.21	.01	.042	.007	.18	.03
12	146	33	129.500	18	.14	.86	.18	.01	.030	.007	.15	.04
13	95	27	81.500	9	.11	.89	.16	.01	.020	.007	.12	.04
14	59	23	47.500	6	.13	.87	.14	.01	.021	.008	.13	.05
15	30	30	15.000	0	0.00	1.00	.14	.01	0.000	0.000	0.00	0.00

Life Table[a]

a. The median survival time is 5.3313

Column A

Interval start time

Indicates the intervals of supervision in years; 0 = up to 1 year, 1 = 1 year up to 2 years, etc.

Column B

Number entering this interval

The number of cases still alive up to the beginning of the interval (t).

$$B_t = (B_{t-1}) - (C_{t-1} + E_{t-1})$$
$$B_4 = (1697) - (22 + 152) = 1523$$

Column C

Number withdrawn during interval

Censored cases.

These censored cases are called "withdrawn" since they do not appear in later intervals (t).

Column D

Number exposed to risk

This is the average number exposed to risk in the interval and calculated as:

The number of cases entering the interval minus 1/2 the cases withdrawing (censored cases) during the interval.

$$D_t = (B_t) - (C_t) (0.5)$$
$$D_4 = (1523) - (23) (0.5) = 1511.5$$

Column E

Number of terminal events

The number of cases that were died in the interval, i.e. coded 1 in the database

Column F

Proportion Terminating

This is the proportion of cases that were died in the interval, which is the probability of death in the interval.

$$F_t = (E_t) / (D_t)$$
$$F_4 = (171) / (1511.5) =\sim 0.1131$$

Column G

Proportion surviving

The proportion of cases still alive through the end of the interval, the probability of being successful through the end of the interval

$G_t = (1.0 - F_t)$
$G_4 = (1.0 - 0.1131) = \sim 0.8869$

Column H

Cumulative proportion surviving at end

The probability of a case remaining alive up to and through the end of the Interval.

$H_t = (H_{t-1}) (G_t)$
$H_4 = (H_{4-1}) (G_4)$
$H_4 = (0.6524) (0.8869)$
$H_4 = \sim 0.5786$

Column I

Standard error of the cumulative proportion surviving.

The error associated with the estimated probability of a case surviving up to and through the end of the interval.

Column J

Probability density.

The estimated Probability of revocation during interval (t).

$J_t = (H_{t-1}) - (H_t)$

Column K

Standard error of the probability density.

Estimated error of the probability density estimate.

Column L

Hazard rate.

The proportion of case that have survived, i.e. been on probation, up to the interval (t) who are expected to fail in the interval.

$L_t = (E_t) / [D_t - E_t (0.5)]$

Column M

Standard error of the hazard rate.

Estimated error of the hazard rate.

The Median Survival Time

How many years elapse before half the survivors died? Median survival time = 5.3313 years means By 5.3313 years, half the patients in the sample were died.

Graphs Section:

Cases weighted by freq

Cases weighted by freq

Cases weighted by freq
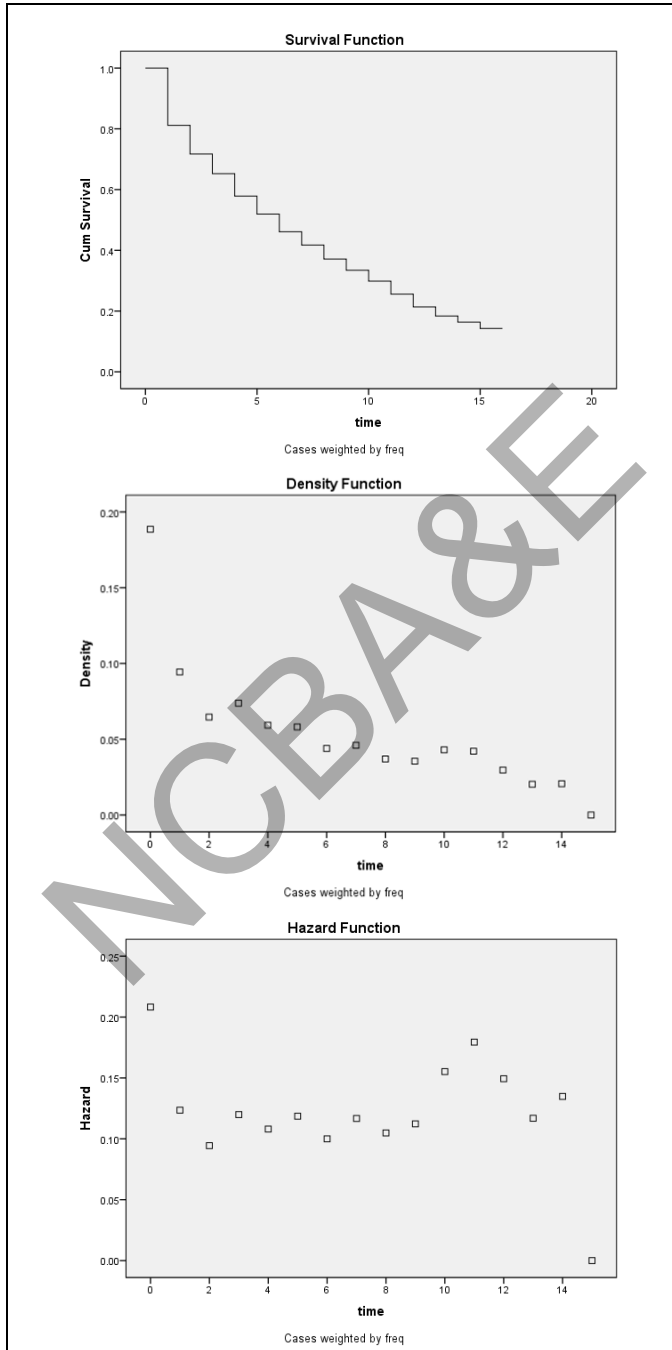
The Survival Function: A plot of the cumulative proportion of cases surviving up to the end of each interval of time.

The Density Function: A plot of the probability density associated with each interval of time. This illustrates the difference between the proportion of cases that began each interval and the proportion that survived to the end of the interval.

The Hazard Function: A plot of the hazard rate. This illustrates the proportion of cases that have survived up to the beginning of the interval that are expected to fail in the interval. As a rate, it can take values greater than 1.

Over all write up for the Life Table Analysis

This table shows the estimated survival probabilities within 15 intervals for a total of 2418 items. It shows how many items were at risk at the start of each interval, how many failed before the end of the interval, and how many were withdrawn (censored) during the interval. The column labeled Cumulative Survival shows the estimated probability of an item surviving as least as long as the beginning of the interval. The column labeled Hazard is the estimated hazard function (instantaneous failure rate) over each interval. Density shows an estimate of the density function of the corresponding lifetime distribution. Standard errors are also shown in parentheses for each of the three functions.

10.3 Kaplan Meier

There are many situations in which we would want to examine the distribution of times between two events, such as length of employment (time between being patient and leaving the hospital).

However, this kind of data usually includes some censored cases. Censored cases are cases for which the second event isn't recorded.

The Kaplan-Meier procedure is a method of estimating time-to-event models in the presence of censored cases.

The Kaplan-Meier model is based on estimating conditional probabilities at each time point when an event occurs and taking the product limit of those probabilities to estimate the survival rate at each point in time.

Kaplan-Meier Product Limit Estimation:

The life table method is the oldest and most commonly used technique for estimating the survival function (and the hazard and probability density functions). However, the exact estimates from the life table will depend on the choice of the number and widths of survival time intervals. The Kaplan-Meier product-limit method estimates the survival function directly from the survival times, without tabulation.

Kaplan-Meier product-limit estimator is defined as follows

$$\hat{S}(T) = \begin{cases} 1 & \text{if } T_{\min} > T \\ \displaystyle\prod_{A \leq T_i \leq T}\left[1 - \dfrac{d_i}{r_i}\right] & \text{if } T_{\min} \leq T \end{cases}$$

The variance of S(T) is estimated by Greenwood's formula

$$\hat{V}\left[\hat{S}(T)\right] = \hat{S}(T)^2 \sum_{A \leq T_i \leq T} \frac{d_i}{r_i\left(r_i - d_i\right)}$$

Nelson-Aalen Hazard Estimator

The Nelson-Aalen estimator is recommended as the best estimator of the cumulative hazard function, H(T). This estimator is give as

$$\hat{H}(T) = \begin{cases} 0 & \text{if } T_{min} > T \\ \sum_{A \leq T_i \leq T} \frac{d_i}{r_i} & \text{if } T_{min} \leq T \end{cases}$$

Example 10.2:
Dataset given below was reported by Crowley and Hu (1977) pertaining to the survival of heart transplant patients.

The data is as follows:

id	time	Censoring	hospital	age	antigen	mismatch	status
1	1	Censored	BINER	54	0	0.47	0
2	1	Censored	ST_AND	35	0	0.67	0
3	3	Censored	HILLVIEW	40	0	1.66	0
4	10	Complete	HILLVIEW	55	1	2.76	1
5	10	Complete					1
6	12	Censored	HILLVIEW	29	0	0.61	0
7	13	Censored	HILLVIEW	28	1	0.77	0
8	15	Censored	HILLVIEW	54	0	1.11	0
9	23	Censored	HILLVIEW	56	0	2.05	0
10	25	Complete	ST_AND	53	1	1.68	1
11	26	Censored	ST_AND	52	1	0.82	0
12	29	Complete	ST_AND	54	0	1.08	1
13	30	Censored	ST_AND	45	0	0.16	0
14	39	Complete					1
15	39	Complete	HILLVIEW	42	0	1.38	1
16	44	Censored	ST_AND	36	0	0	0
17	46	Complete	ST_AND	42	0	0.61	1
18	47	Complete	ST_AND	61	1	0.87	1
19	48	Censored	BINER	53	0	3.05	0
20	50	Complete	BINER	49	0	0.66	1
21	50	Complete	HILLVIEW	46	0	2.25	1
22	51	Complete	HILLVIEW	47	0	1.38	1
23	51	Complete	ST_AND	52	0	1.51	1
24	54	Complete	HILLVIEW	49	0	2.09	1
25	60	Complete	HILLVIEW	64	0	0.69	1
26	63	Complete	BINER	56	1	2.16	1
27	64	Complete	ST_AND	54	0	1.89	1

id	time	Censoring	hospital	age	antigen	mismatch	status
28	65	Complete	ST_AND	45	1	1.68	1
29	66	Complete	HILLVIEW	51	0	1.12	1
30	68	Complete	HILLVIEW	51	1	1.33	1
31	110	Censored	BINER	23	1	1.78	0
32	127	Censored	ST_AND	48	0	0.36	0
33	136	Complete	ST_AND	52	1	1.62	1
34	161	Complete	BINER	43	0	1.2	1
35	167	Censored	BINER	26	0	0.46	0
36	228	Censored	HILLVIEW	19	0	1.02	0
37	237	Censored	ST_AND	47	0	0.33	0
38	253	Complete	HILLVIEW	48	1	1.08	1
39	280	Complete	BINER	49	0	1.12	1
40	297	Complete	BINER	42	0	0.6	1
41	305	Censored	HILLVIEW	49	0	0.81	0
42	322	Complete	ST_AND	48	1	1.82	1
43	339	Censored	HILLVIEW	54	0	0.68	0
44	389	Censored	BINER	48	1	1.44	0
45	439	Censored	ST_AND	52	1	1.94	0
46	456	Censored	ST_AND	46	0	1.41	0
47	499	Censored	HILLVIEW	52	1	1.7	0
48	551	Censored	HILLVIEW	48	0	0.12	0
49	589	Censored	BINER	47	0	0.97	0
50	592	Censored	BINER	26	1	1.46	0
51	624	Complete	HILLVIEW	51	0	1.32	1
52	660	Censored	ST_AND	48	0	1.2	0
53	730	Complete	ST_AND	58	0	0.96	1
54	815	Censored	BINER	32	1	1.93	0
55	836	Complete	BINER	44	0	1.58	1
56	838	Censored	BINER	41	0	0.19	0
57	875	Censored	ST_AND	38	0	0.98	0
58	994	Complete	BINER	48	0	0.81	1
59	1024	Complete	BINER	43	0	1.13	1
60	1106	Censored	HILLVIEW	36	0	1.35	0
61	1264	Censored	BINER	45	0	0.98	0
62	1350	Complete	BINER	54	0	0.87	1
63	1367	Censored	BINER	48	0	0.75	0
64	1536	Censored	BINER	49	0	0.91	0
65	1549	Censored	HILLVIEW	40	0	0.38	0
66	1775	Censored	ST_AND	33	0	1.06	0

The first variable in this data set is survival time, that is, the date of the heart transplant and the date when the respective patient either died or dropped out of the study (could no longer be contacted). Variable Censored is the censoring indicator variable with the codes that identify whether a respective time represents an observation that is completely specified or a censored observation (0-Complete; 1-Censored). The variable Hospital is a (fictitious) grouping variable which identifies to which one of three different hospitals a respective case belongs.

Example S10-2

Given below is a part of the Data in SPSS Data sheet.

	time	Censoring	hospital	age	antigen	mismatch	status
1	1	1	3	54	0	.47	0
2	1	1	2	35	0	.67	0
3	3	1	1	40	0	1.66	0
4	10	0	1	55	1	2.76	1
5	10	0	.	.	.	.	1
6	12	1	1	29	0	.61	0
7	13	1	1	28	1	.77	0
8	15	1	1	54	0	1.11	0
9	23	1	1	56	0	2.05	0
10	25	0	2	53	1	1.68	1

We start Kaplan-Meier as follows:

Analyze→ Survival→ Kaplan-Meier…
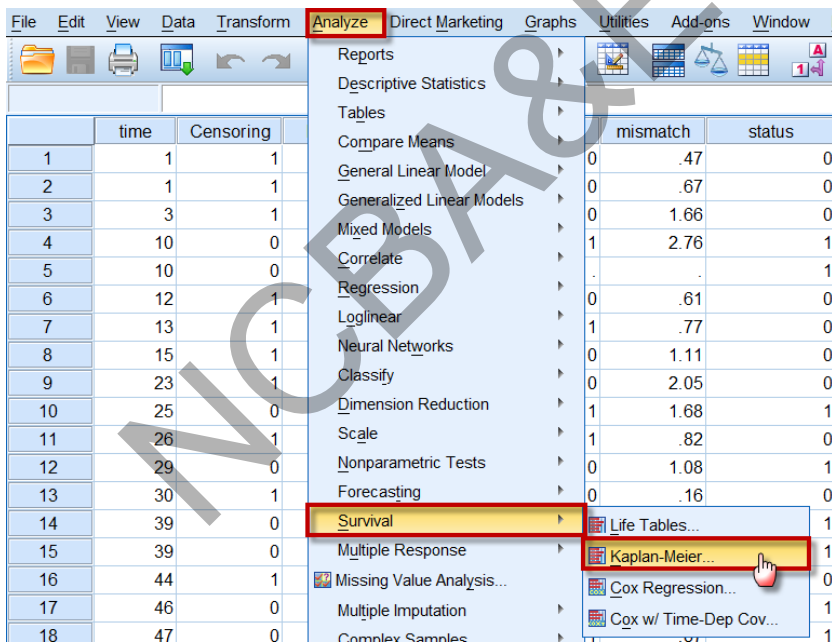
Move the Survival Time Variable (time) to Time

Click on "Define Event" , mark on "Single value:" put 0

Click on "Options" and mark on Survival table(s), Mean and median survival and Survival

Now click on **Continue** then **OK**, to get the following outputs:

Survival Table

	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	1.000	Censord	.	.	0	65
2	1.000	Censord	.	.	0	64
3	3.000	Censord	.	.	0	63
4	10.000	Complete	.	.	1	62
5	10.000	Complete	.968	.022	2	61
6	12.000	Censord	.	.	2	60
7	13.000	Censord	.	.	2	59
8	15.000	Censord	.	.	2	58
9	23.000	Censord	.	.	2	57
10	25.000	Complete	.951	.027	3	56
11	26.000	Censord	.	.	3	55
12	29.000	Complete	.934	.032	4	54
13	30.000	Censord	.	.	4	53
14	39.000	Complete	.	.	5	52
15	39.000	Complete	.899	.039	6	51
16	44.000	Censord	.	.	6	50
17	46.000	Complete	.881	.042	7	49
18	47.000	Complete	.863	.045	8	48
19	48.000	Censord	.	.	8	47
20	50.000	Complete	.	.	9	46
21	50.000	Complete	.826	.050	10	45
22	51.000	Complete	.	.	11	44
23	51.000	Complete	.789	.054	12	43
24	54.000	Complete	.771	.056	13	42
25	60.000	Complete	.753	.058	14	41
26	63.000	Complete	.734	.059	15	40
27	64.000	Complete	.716	.060	16	39
28	65.000	Complete	.698	.062	17	38
29	66.000	Complete	.679	.063	18	37
30	68.000	Complete	.661	.064	19	36
31	110.000	Censord	.	.	19	35
32	127.000	Censord	.	.	19	34
33	136.000	Complete	.641	.065	20	33
34	161.000	Complete	.622	.065	21	32
35	167.000	Censord	.	.	21	31
36	228.000	Censord	.	.	21	30
37	237.000	Censord	.	.	21	29
38	253.000	Complete	.601	.067	22	28
39	280.000	Complete	.579	.068	23	27

	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
40	297.000	Complete	.558	.068	24	26
41	305.000	Censord	.	.	24	25
42	322.000	Complete	.535	.069	25	24
43	339.000	Censord	.	.	25	23
44	389.000	Censord	.	.	25	22
45	439.000	Censord	.	.	25	21
46	456.000	Censord	.	.	25	20
47	499.000	Censord	.	.	25	19
48	551.000	Censord	.	.	25	18
49	589.000	Censord	.	.	25	17
50	592.000	Censord	.	.	25	16
51	624.000	Complete	.502	.073	26	15
52	660.000	Censord	.	.	26	14
53	730.000	Complete	.466	.076	27	13
54	815.000	Censord	.	.	27	12
55	836.000	Complete	.427	.079	28	11
56	838.000	Censord	.	.	28	10
57	875.000	Censord	.	.	28	9
58	994.000	Complete	.380	.083	29	8
59	1024.000	Complete	.332	.085	30	7
60	1106.000	Censord	.	.	30	6
61	1264.000	Censord	.	.	30	5
62	1350.000	Complete	.266	.090	31	4
63	1367.000	Censord	.	.	31	3
64	1536.000	Censord	.	.	31	2
65	1549.000	Censord	.	.	31	1
66	1775.000	Censord	.	.	31	0

This table shows estimated survival probabilities based on the data in Time. Each row of the table represents a single data value, displayed in increasing order. If the data value represents a failure or death, the status column indicates Event. If the data value represents a censored observation, the status column indicates Censored. The number at risk is the number of items which have survived up until each data value. For each unique failure time, the data displays the estimated survival probability, the standard error of that estimate, and the estimated hazard function.

Means and Medians for Survival Time

Mean[a]				Median			
Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound			Lower Bound	Upper Bound
783.645	109.726	568.582	998.708	730.000	312.810	116.893	1343.107

a. Estimation is limited to the largest survival time if it is censored.

Median Survival Time: This is not the conventional median, this is the time associated with the first case to have a cumulative survival probability ≤ 0.5

Survival Function
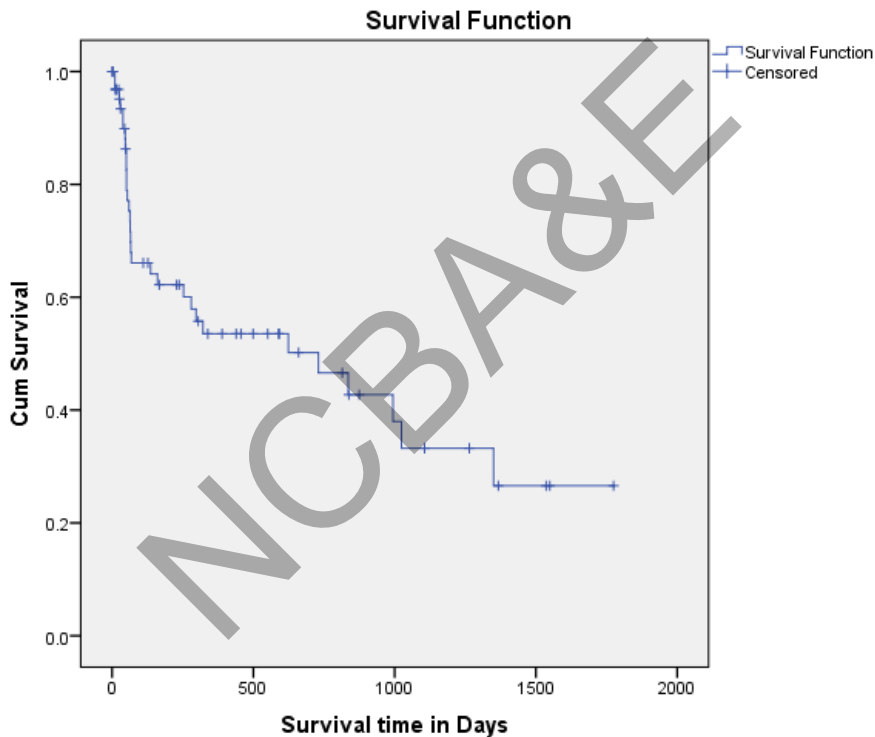
Test of Significance for comparison of Kaplan Meier Survival Curves

Are Kaplan Meier survival curves statistically equivalent? There are three Tests available in SPSS for the comparison of KM survival curves

1. Log-Rank (Mantel-Haenszel Test)

2. Breslow Generalized Wilcoxon Test

3. Tarone-Ware Test

Log-Rank Test

In survival analysis, the log-rank test is a hypothesis test to compare the survival distributions of two or more samples. It is a nonparametric test and appropriate to use when the data are right skewed and censored. Log-rank test is widely used in clinical trials to establish the efficacy of a new treatment compared to a control treatment when the measurement is the time to event (such as the time from initial treatment to a heart attack). The test is also called the Mantel–Cox test, named after Nathan Mantel and David Cox.

The log-rank (Mantel-Cox) test is the more powerful of the two tests if the assumption of proportional hazards is true. Proportional hazards means that the ratio of hazard functions (deaths per time) is the same at all time points. One example of proportional hazards would be if the control group died at twice the rate as treated group at all time points. Prism actually computes the Mantel-Haenszel method, which is nearly identical to the log-rank method (they differ only in how they deal with two subjects with the same time of death).

In Log-Rank test all cases weighted equally, log-rank is least conservative of the three tests available in SPSS

Breslow Test

The Gehan-Breslow-Wilcoxon method gives more weight to deaths at early time points. This often makes lots of sense, but the results can be misleading when a large fraction of patients are censored at early time points. In contrast, the log-rank test gives equal weight to all time points. The Gehan-Wilcoxon test does not require a consistent hazard ratio, but does require that one group consistently have a higher risk than the other.

You need to choose which P value to report. Ideally, this choice should be made before you collect and analyze your data.

If in doubt, report the log-rank test (which is more standard) and report the Gehan-Wilcoxon results only if you have a strong reason.

Tarone-Ware Test

Breslow test and Tarone ware test are identical the only difference is Tarone-Ware test uses Square root of the number of cases at risk at event time (t) as weights (i.e Weights earlier cases less heavily than the Breslow Test does). Tarone-Ware Test is mid-conservative of the three tests.

SPSS Procedure

Analyze → Survival → Kaplan Meier

> Time: Survival Time Variable (time)
>
> Status: Censored Variable
>
> Define: 0-Complete; 1-Censored
>
> Factor: Choose Hospital as factor variable

Options: Plots: Choose Survival
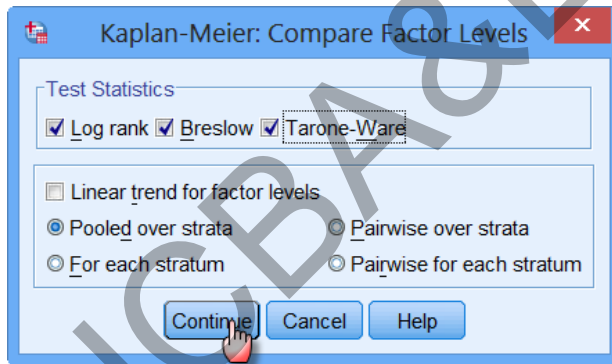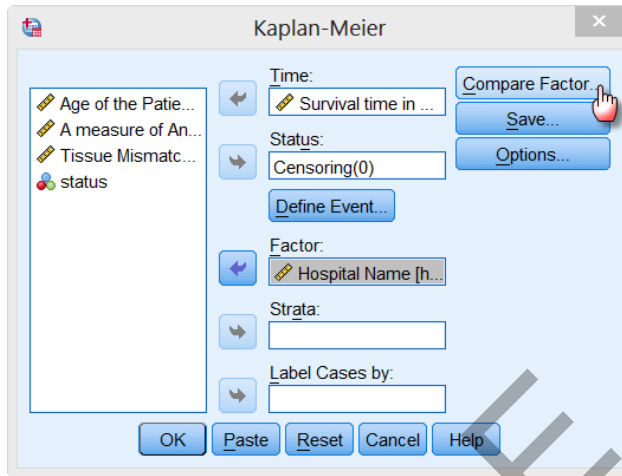
Now click on Continue then OK, to get the following outputs:

Case Processing Summary

Hospital Name	Total N	N of Events	Censored	
			N	Percent
HILLVIEW	22	10	12	54.5%
ST_AND	21	10	11	52.4%
BINER	21	9	12	57.1%
Overall	64	29	35	54.7%

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	3.013	2	.222
Breslow (Generalized Wilcoxon)	5.195	2	.074
Tarone-Ware	4.523	2	.104

Test of equality of survival distributions for the different levels of Hospital Name.

Three tests have also been performed to determine whether there is a statistically significant difference between the survival probabilities of the 3 groups (hospitals). Since the smallest P-value is greater than or equal to 0.05, there is not a statistically significant difference between the groups at the 95% confidence level.

10.4 Cox – Regression

(Proportional Hazards Model (PHM))

Cox Regression builds a predictive model for time-to-event data. The model produces a survival function that predicts the probability that the event of interest has occurred at a given time t for given values of the predictor variables. The shape of the survival function and the regression coefficients for the predictors are estimated from observed subjects; the model can then be applied to new cases that have measurements for the predictor variables.

Note that information from censored subjects, that is, those that do not experience the event of interest during the time of observation, contributes usefully to the estimation of the model.

Cox (proportional hazards) regression analysis models the relationship between a set of one or more covariates and the hazard rate. Covariates may be discrete or continuous. Cox Regression can be used to study the impact of various factors on survival. You may be interested in the impact of diet, age, amount of exercise, and amount of sleep on the survival time after an individual has been diagnosed with a certain disease such as cancer. Under normal conditions, the obvious statistical tool to study the relationship between a response variable (survival time) and several explanatory variables would be multiple regression. Unfortunately, because of the special nature of survival data, multiple regression is not appropriate. Survival data usually contain censored data and the distribution of survival times is often highly skewed. These two problems invalidate the use of multiple regression. Many alternative regression methods have been suggested. The most popular method is the proportional hazard regression method developed by Cox (1972).

The Cox (1972) expressed the relationship between the hazard rate and a set of covariates using the model

$$h(T) = h_0(T) e^{\sum_{i=1}^{p} x_i \beta_i}$$

$$\ln[h(T)] = \ln[h_0(T)] + \sum_{i=1}^{p} x_i \beta_i$$

$$\ln\left[\frac{h(T)}{h_0(T)}\right] = \sum_{i=1}^{p} x_i\beta_i$$

$$\frac{h(T)}{h_0(T)} = e^{\sum_{i=1}^{p} x_i\beta_i}$$

$$e^{x_1\beta_1} e^{x_2\beta_2} e^{x_3\beta_3} \ldots\ldots e^{x_p\beta_p}$$

The Regression Coefficients can thus be interpreted as the relative risk when the value of the covariate is increased by one unit. Unlike most regression models, this does not include and intercept term. This is because if an intercept term were included, it would became part of h0(t).

Example 10.3:
You have data on 48 participants in a cancer drug trial. Of these 48, 28 received treatment (drug=1) and 20 receive a placebo (drug=0). The participant range in age from 47 to 67 years. You wish to analyze time until death, measured in months. You have data given below.

study time	died	drug	age	Study time	died	drug	age
1	1	0	61	10	0	1	49
1	1	0	65	11	0	1	61
2	1	0	59	13	1	1	62
3	1	0	52	15	0	1	50
4	1	0	56	16	1	1	67
4	1	0	67	19	0	1	50
5	1	0	63	20	0	1	55
5	1	0	58	22	1	1	58
8	1	0	56	23	1	1	47
8	0	0	58	32	0	1	52
8	1	0	52	6	1	1	55
8	1	0	49	10	1	1	54
11	1	0	50	17	0	1	60
11	1	0	55	19	0	1	49
12	1	0	49	24	1	1	58
12	1	0	62	25	0	1	50
15	1	0	51	25	1	1	55
17	1	0	49	28	1	1	57
22	1	0	57	28	0	1	48
23	1	0	52	32	0	1	56
6	1	1	67	33	1	1	60
6	0	1	65	34	0	1	62
7	1	1	58	35	0	1	48
9	0	1	56	39	0	1	52

Example S10-3

Given below is the preview of cases in SPSS Data sheet.

Analyze → Survival → Cox Regression
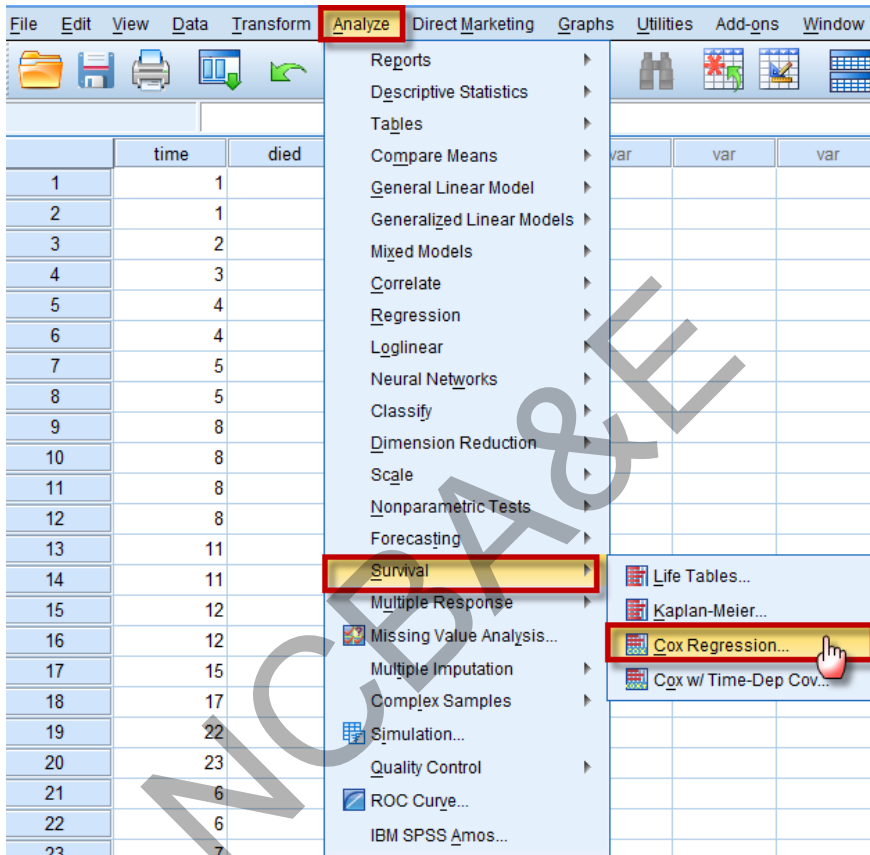
Time: Survival Time Variable (time)

Status: Censored Variable

Define: 1-Complete; 0-Censored

Covariates: Choose independent variable(s) (Drug, Age)

Categorical: Choose factor variable(s) (Drug)

Now click on [Continue] then [OK], to get the following outputs:
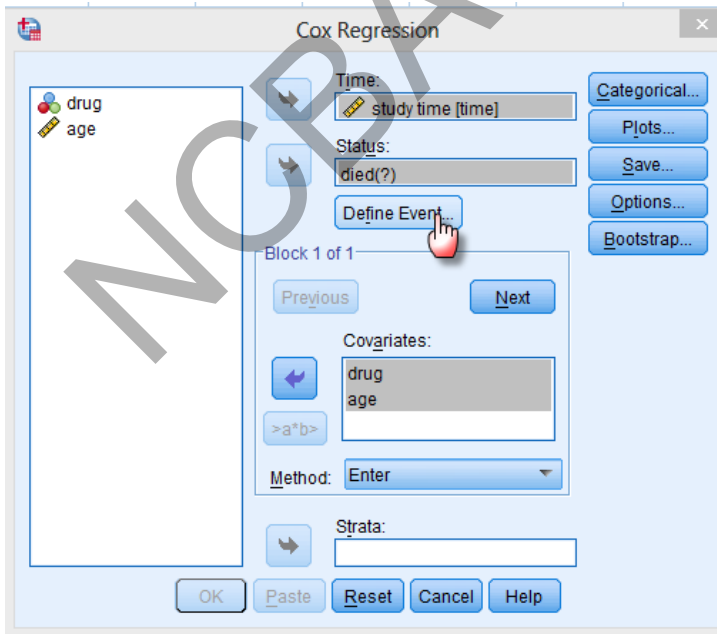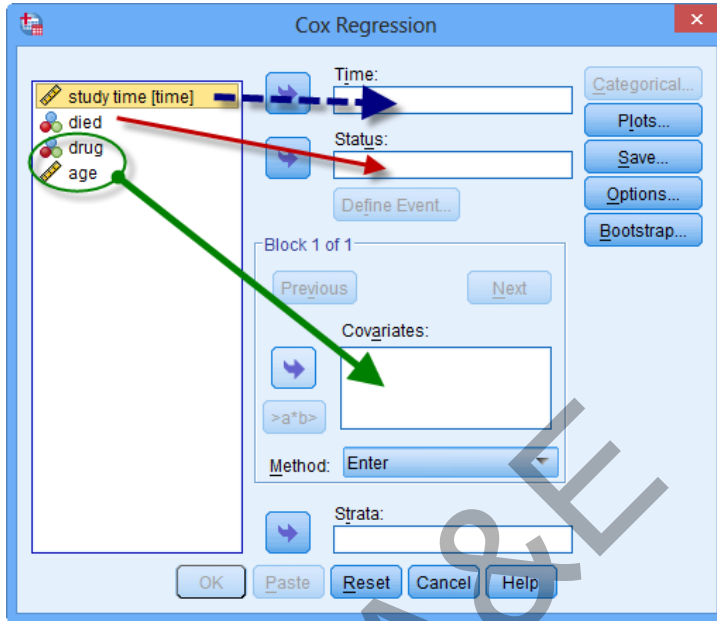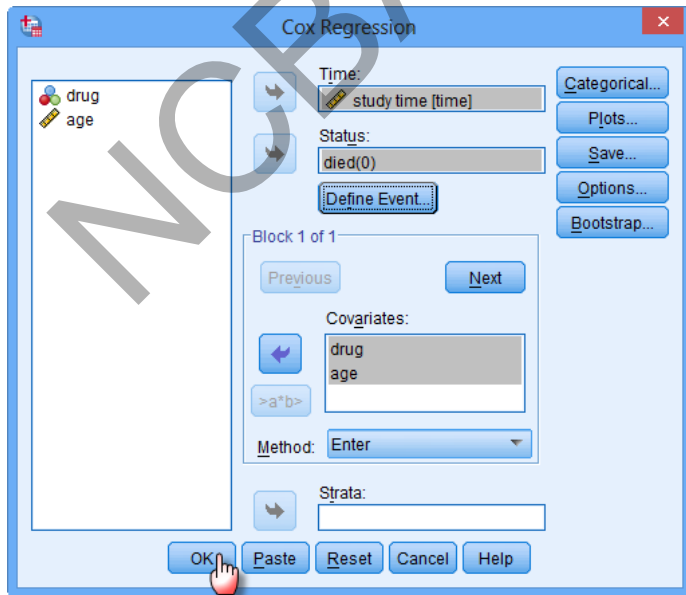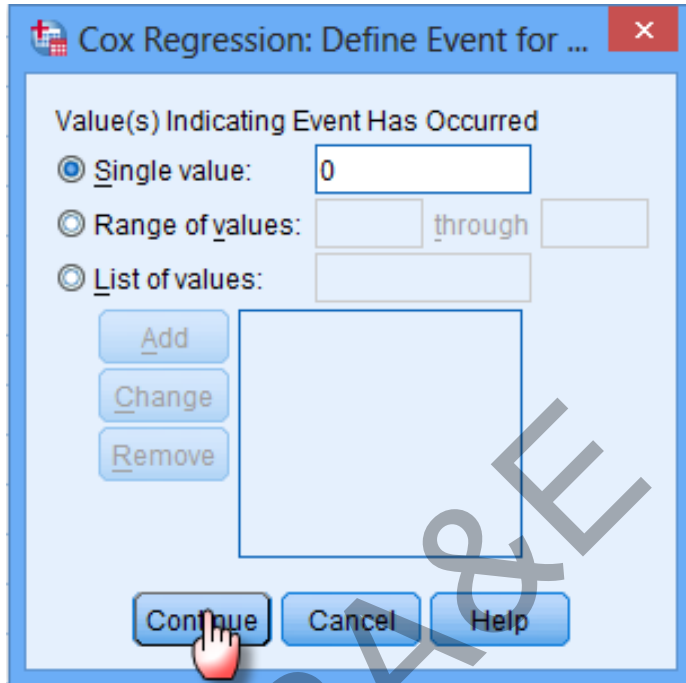
-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
83.061	.773	2	.679	.901	2	.637	.901	2	.637

Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
drug	.879	1.068	.678	1	.410	2.408
age	.008	.051	.026	1	.872	1.008

The output shows the results of fitting a cox regression model to describe the relationship between Time and 2 independent variable(s) drug and age. The hazard function at a selected combination of the input factors x is a multiple of the baseline hazard function $h(t|0)$, as shown below:

$$h(t|x)=h(t|0)*exp(0.00820993*Age + 0.878992*Drug=1)$$

In determining whether the model can be simplified, notice that the highest P-value for the likelihood ratio tests is 0.8723, belonging to Age. Because the P-value is greater or equal to 0.05, that term is not statistically significant at the 95.0% or higher confidence level. Consequently, you should consider removing Age from the model.

Chapter 11

Reliability Coefficient

11.1 Introduction

The degree of stability is exhibited when measurement is repeated under identical situation. Reliability refers to the closeness of measurements of observations obtained under identical situations. If the cholesterol concentration of two *portions* of the same serum specimen is measured in an automated chemical analyzer, ideally two results should be exactly the same.

Note that all the fluctuations in measurements or observations are attributable to lack of reliability. The attributes themselves usually vary in a variety of ways. Consider the distribution of blood pressure found in a community survey in which each subject has two measurements. The major components of variation in the distribution are as follows:

1. **Difference among subgroups**
   For example, older persons have higher blood pressure than younger ones.
2. **Difference among individuals within subgroups**
   For example, among old men aged 60, some individuals have higher blood pressure than the others.
3. **Difference within each individual**
   Due to variety of influences each individual's blood pressure varies from one moment to another.
4. **Measurement errors**
   Even if the blood pressure measured were exactly the same, it would appear to vary because of the observers' failure in accurate measurements.
5. **Sampling variations**
   We know if sample is small, sampling error is more whereas if sample is large, sampling error is less, moreover if repeated samples are selected from a population, the findings in each sample will differ from one to the other.

Daily experiences constantly remind us of measurement errors for instance, bath room scales are typically accurate to no better than ± 1 kg, home thermometer is accurate to about ± 0.2c etc. Therefore we can say that error of measurement is a relatively small fraction of the observations.

The definition of reliability is

$$\text{Reliability} = R_e = \frac{\text{Subject variation within groups}}{\text{Subject variation + Measurement error}} \tag{11.1}$$

or

$$\text{Reliability} = \frac{\text{Variance components among subjects } (\sigma_s^2)}{\text{Variance components among subjects } (\sigma_s^2) + \text{Variance of error } (\sigma_E^2)} \tag{11.2}$$

11.2 Reliability of a Test

The reliability coefficient for a test of scores from a group of examinees is the coefficient of correlation between that set of scores and another set of scores on an equivalent test obtained independently from the members of the same groups.

The analytical approach is based on the statistical technique called Analysis of Variance. This will be explained by an example.

Example 11.1:
The data given in Table 11.1 relate to degree of sadness of 10 patients rated by 3 observers.

Table 11.1

Patients	Observer 1	Observer 2	Observer 3
1	6	7	8
2	4	5	6
3	2	2	2
4	3	4	5
5	5	4	6
6	8	9	10
7	5	7	9
8	6	7	8
9	4	6	8
10	7	9	8
Mean	5.0	6.0	7.0

Calculate the reliability coefficient among patients with regards to three observers.

Solution:
This problem relates to TWO WAY ANOVA. This method of analysis has been explained in Chapter 5. There are three observers and 10 patients. There could be three sources of variations-Patients-Observers and Error.

The IBM-SPSS package is used as shown in the following Example:

Example S11-1

The data will be in 3 columns and a part of the data is as follows:

|	Observers	Patients	Score
1	1	1	6
2	1	2	4
3	1	3	2
4	1	4	3
5	1	5	5
6	1	6	8
7	1	7	5
8	1	8	6
9	1	9	4
10	1	10	7
11	2	1	7

We apply the TWO WAY ANOVA as follows:

Analyze→ General Linear Model→ Univariate…
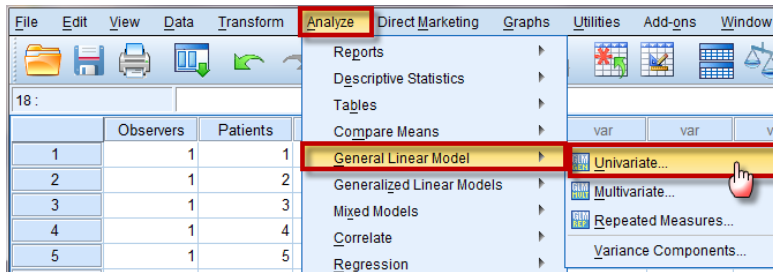
Move the variable "Score" to Dependent Variable:

Move both "Observers" and "Patients" to Fixed Factor(s):
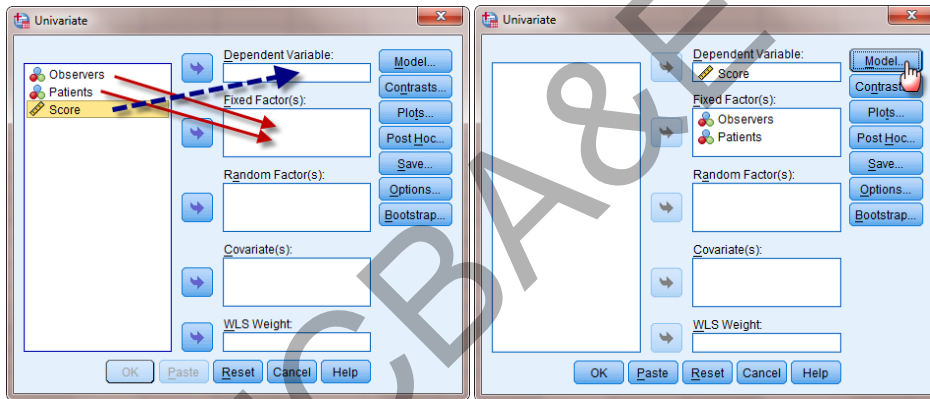
Choose "Custom";

Move both "Observers" and "Patients" to Model:

Select "Main effects" from the Type:

Now click on **Continue** then **OK**, to get the following outputs:

<div align="center">
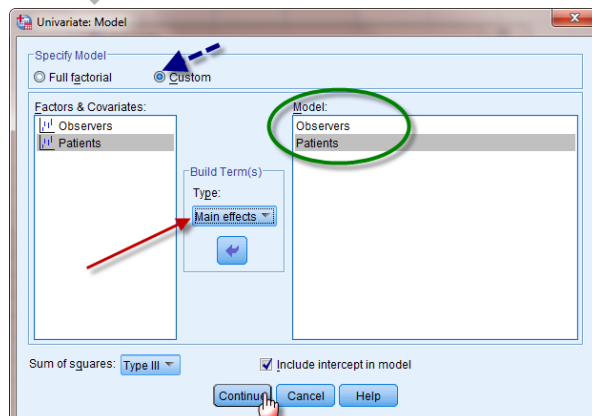
SPSS Output for ANOVA 2 Ways

ANALYSIS OF VARIANCE

</div>

scores

by observers

patients

unique sum of squares

all effects entered simultaneously

Source of Variation	E(MS)	Sum of Squares	DF	Mean Square	F	Sig
OBSERVER	$\sigma_E^2 + 10\sigma_O^2$	20.000	2	10.000	18.000	.000
PATIENTS	$\sigma_E^2 + 3\sigma_P^2$	114.000	9	12.667	22.800	.000
Residual	σ_E^2	10.000	18	0.556		
Total		144.000	29	4.966		

By simple calculations we get variance components:

$$\sigma_E^2 \ = \ \sigma^2 = MS(E) = 0.556$$

$$\sigma_P^2 \ = \ \sigma^2 \text{ (patients)} = \{MS(P) - MS(E) \}/3 \ = \ \{12.667\text{-}.556\}/3 \ = \ 4.037$$

$$\sigma_O^2 \ = \ \sigma^2 \text{ (observers)} = \{MS(O) \text{-}MS(E) \}/10 \ = \ \{10.00\text{-}0.556\}/10 \ = \ 0.94$$

σ^2 (patients) and σ^2 (observers) are called variance components of sources of variation. The reliability may be calculated using (11.2) as

$$R_e = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_E^2} = (4.037)/\{4.037 + 0.556\} \ = 0.88,$$

where R_e is the coefficient of reliability.

This shows that 88% of the variance in the scores results from true variance among patients. This coefficient is known as *reliability coefficient.*

11.3 Different Forms of Measuring Reliability Coefficients

Reliability is measured by performing two or more independent measurements and comparing the findings, using an appropriate statistical index. There has been a number of methods suggested in literature but no method is perfect. There are some drawbacks and good points in each method.

There has been a considerable debate in the literature regarding the most appropriate

choice of the reliability coefficients. Four tests of measuring the reliability are given and for each case an example is presented so that it should be very clear which method is applicable under what situation. However, more than one methods can be used for one problem. The methods are given as:

(i) Test -retest method

(ii) Split-half

(iii) Kuder and Richardson-20 method

(iv) Cronbach's Alpha (α)

11.3.1 Test-Retest Method

In test-retest method comparison may be based on observations by different observers or interviews by different observers or repeated measurements or interviews using the same questionnaire. Replicated tests may be made on the same blood specimens. A question may be repeated in the same questionnaire, or differently worded questions asking for the same information may be included. The results of test-retest comparison depend on the interval between the tests. *A questionnaire based measure of overall health, for example, was found to have test-retest reliability of about* 0.85 *over a 1-month period, but only about* 0.56 *over a 3-year interval* (Ware, J.E -1984).

The methods of correlation and KAPPA-Statistic may be used to test-retest method of testing reliability coefficient. They are as:

Continuous Data	Ordinal Data	Categorical Data
(i) Person correlation coefficient	(1) Spearman-Brown correlation coefficient	(1) KAPPA-Statistic
(ii) Intra-class correlation coefficient	(2) Kendall's Tau(τ)	(2) i- Phi ii- Cramer's V

(i) Person and intra-class correlation coefficients
This has been explained in Section 6.4. Spearson Brown formula (rank correlation) is used when data is ordinal. Kendall's tau can also be categorized in this rank.

(ii) Kappa-Statistic
There are many situations in medicine which has only two levels i.e. presence or absence, positive or negative, normal or abnormal. A straightforward approach is to calculate simple agreement: the proportion of responses in which the two observations agreed. For such types of qualitative variable a frequently used index of reliability or agreement between observers is known as Cohen's Kappa coefficients (Cohen-1960). This index or measure has the desirable feature of showing how much more agreement there is than would be expected by chance. This measure is very strongly influenced by the distribution of positive and negative values. If there is a preponderance of either normal or abnormal causes, there will be high agreement. The Kappa-Statistic explicitly deals with the situations by examining the proportion of

responses in the two agreement cells in relation to the proportion of responses in these cells, which would be expected by chance.

$$K = \frac{P(O) - P(C)}{1 - P(C)}$$

by using (7.34), where:

P(O) = observed proportion of agreement and

P(C) = expected proportion of agreement

This has already been explained in the context of Chi-square (Chapter 7)

Example 11.2:

Suppose we were to consider a judgment by two observers of the presence and absence of a Babinski sign, an up going toe following scratching of the bottom of the foot, on a series of neurological patients. The data are given in Table 11.2 by 2×2 table.

Table 11.2:

Observer 2	Observer 1 Present	Absent	Total
Present	20	15	35
Absent	10	55	65
Total	30	70	100

Calculate the agreement index between two observers using Kappa-statistic.

Solution:

Since in the calculation of Kappa index expected frequencies will be used therefore these are given as:

Observer 2	Observer 1 Present	Absent	Total
Present	10.5	24.5	35
Absent	19.5	45.5	65
Total	30	70	100

$$P(O) = \frac{20 + 55}{100} = 0.75$$

$$P(C) = \frac{10.5 + 45.5}{100} = 0.56$$

The Kappa index is

$$K = \frac{0.75 - 0.56}{1 - 0.56} = 0.43.$$

We say that there is good agreement between two observers.

The IBM-SPSS package is used as shown in the following Example:

Example S11-2

The data will be in 2 columns and a part of the data is as follows:

	Observer1	Observer2			Observer1	Observer2
1	1	1		1	Present	Present
2	1	1		2	Present	Present
3	1	1		3	Present	Present
4	1	1		4	Present	Present
5	1	1		5	Present	Present
6	1	1		6	Present	Present
7	1	1		7	Present	Present
8	1	1		8	Present	Present
9	1	1		9	Present	Present
10	1	1		10	Present	Present

The variable view is as follows:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Observer1	Numeric	8	0	Observer 1	{1, Present}...	None	8	Right	Nominal	Input
Observer2	Numeric	8	0	Observer 2	{1, Present}...	None	8	Right	Nominal	Input

We calculate Kappa index as follows:

Analyze→ Descriptive Statistics→ Crosstabs…
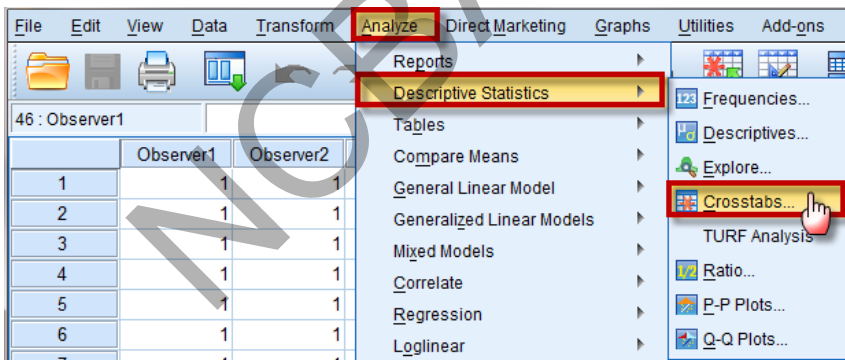
Move the variable "Observer1" to Row(s) and "Observer2" to Column(s):

Click on Statistics:

Mark on both "Kappa" and "Phi and Cramer's V":
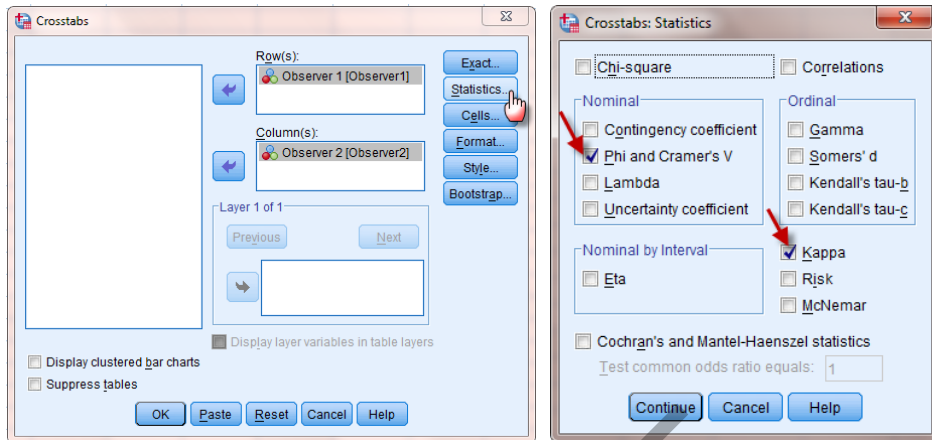
Now click on **Continue** then **OK**, to get the following outputs:

Observer 1 * Observer 2 Crosstabulation

Count

		Observer 2		Total
		Present	Absent	
Observer 1	Present	20	15	35
	Absent	10	55	65
Total		30	70	100

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.435			.000
	Cramer's V	.435			.000
Measure of Agreement	Kappa	.432	.095	4.346	.000
N of Valid Cases		100			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Note that phi, Cramer's V and coefficient of contingency are other methods of testing of association. The results obtained from these indices are almost identical with Kappa index. An alternative form of testing the reliability for such cases, without collecting information second time, has been suggested by Kuder-Richardson, given in Section 11.3.3.

Kappa index can also be used for multiple-classification (n x n table). This has been explained in Chapter 8. Cramer's V, which is close to Kappa, can be used in n×m table.

11.3.2 Split-half Method

Another approach to test the reliability or homogeneity of a scale is called split-half method. Here the items are randomly divided into two halves which are then correlated. The easiest way is to put all odd number items in one half and even number items in the second half randomly and calculate Pearson correlation coefficient and Guttman split-half coefficient. It also depends on the order in which observation are written down. One problem with method is that the resulting correlation coefficient under estimates of the true reliability of the scale, as the reliability of a scale is directly proportional to the number of items in it. Since the sub-scales being correlate are only half the length of the version that will be used in practice, the resulting correlation coefficient will be low or too low. The Pearson-Brown formula is used to correct this occurrence. The equation for correlation coefficient is

$$\rho(\text{Rho}) = \frac{kr}{1 + (k-1)\,r}\,, \tag{11.3}$$

where, k is the factor by which the scale is increased and r is the original correlation coefficient.

If we need only the reliability of a test twice as in the case of reliability estimation by split-half method, the formula is simple as

$$\rho(\text{Rho}) \quad = \frac{2r}{1 + r}\,. \tag{11.4}$$

If for example 40-items scale has been divided into two-half and found that correlation coefficient between two half is 0.82, we can use (11.4) to increase the reliability. This is known as Guttman Reliability Index. The revised index by using (11.4) will be 0.90.

It is not self- evident why this method should help, but the answer lies in the statistical theory. As long as the test items are not perfectly correlated, the true variance will increase as the square of the number of the items, whereas the error variance will increase only as the number of items decreases. So if the test length is doubled, the true variance will be 4 times as large and error variance 2 times as large as the original test.

This method of testing the reliability is commonly used when study of knowledge, attitude and practice is conducted and questions are in the form of Likert's scale (Likert-1952). In a Likert scale a person expresses an opinion by rating his agreement with a series of statements such as:

(i) Recent research doubled the association between smoking and lung cancer

Strongly agree	Somewhat agree	not sure	Somewhat disagree	Strongly disagree

(ii) Passive smoking is always harmful

Strongly agree	Somewhat agree	not sure	Somewhat disagree	Strongly disagree

The application of this method is shown below:

Example 11.3:

There are four questions and five students for an essay contest. Their scores are given below. (These scores may be regarded as the rating by four judges of the performances of five students).

Table 11.3

Students	Question 1	Question 2	Question 3	Question 4
1	2	1	1	3
2	6	4	5	6
3	3	2	1	1
4	6	3	3	3
5	6	4	4	3

Use the Split-half method and calculate the reliability index.

Solution:

In this question there are 4 items. We can combine odd-items together and even items together as:

Q 1

Q 3

Q 2

Q 4

and apply the method of split-half to calculate the reliability.

The SPSS package was used and the result is given as:

SPSS output for Split-Half Method
RELIABILITY ANALYSIS - SCALE (SPLIT)
Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between students	41.2000	4	10.3000		
Within questions	13.2000	3	4.4000	18.8571	.0001
Residual	2.8000	12	.2333		
Total	57.2000	19	3.0105		
Grand Mean	**3.2000**				

Reliability Coefficients

N of Cases   = 5.0   N of Items= 4

Correlation between forms   = .9529   Equal length Spearman-Brown = .9759

Guttman Split-half   = .9757   Unequal-length Spearman-Brown = .9759

2 Items in part 1   2 Items in part 2

Alpha for part 1   = .9320   Alpha for part 2   = .9815

Since there are many ways to divide a test into two halves, so there are in fact many possible coefficients of reliability. A 10-item test can be divided into 126 ways, a 12 item test 462 ways and so on. (These numbers represent the combination of n items taken n/2 at a time). The reliability coefficients may differ quite considerably from one split to another split. This can be seen as:

Table 11.4

Different halves	Person correlation coefficient	Split-half reliability
1,3 and 2,4	0.9011	0.9278
1,2 and 3,4	0.7863	0.8763
1,4 and 2,3	0.9423	0.9691

This is one of the major objection of the application of this test. A refined form of this test has been suggested by Cronbach, known as Cronbach's Alpha(α) (see section 11.3.4 below).

The IBM-SPSS package is used as shown in the following Example:

Example S11-3

The data will be in 5 columns and the data is as follows:

	Student	Q1	Q2	Q3	Q4
1	1	2	1	1	3
2	2	6	4	5	6
3	3	3	2	1	1
4	4	6	3	3	3
5	5	6	4	4	3

We calculate Split-half Method as follows:

Analyze→ Scale→ Reliability Analysis…
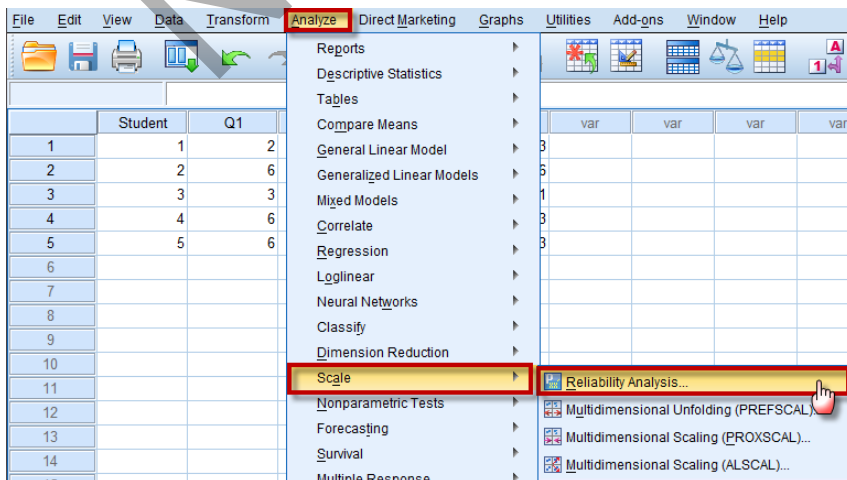
Move the variables "Question 1,…, Question 4" to Items:

Click on Model:

Chose Split-half:
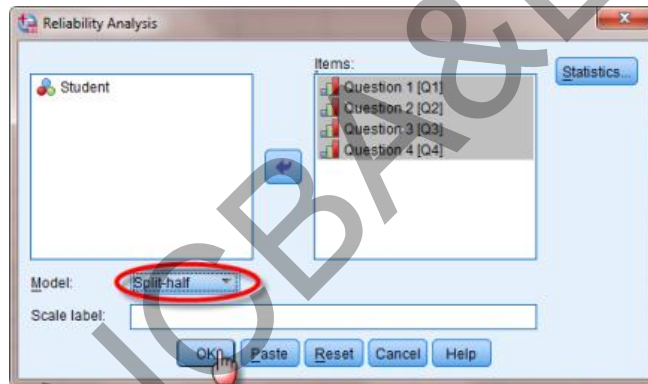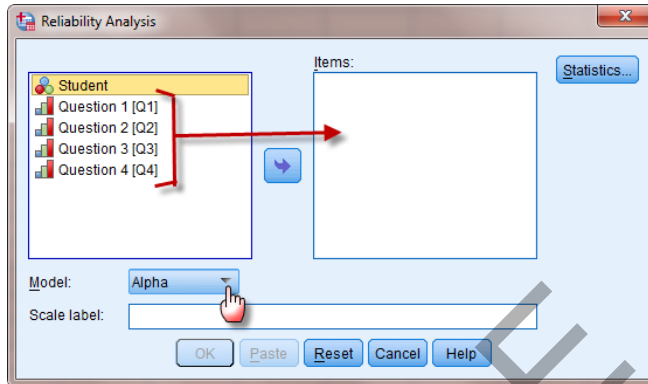
Now click on OK , to get the following output:

Reliability Statistics

Cronbach's Alpha	Part 1	Value	.932
		N of Items	2[a]
	Part 2	Value	.887
		N of Items	2[b]
	Total N of Items		4
Correlation Between Forms			.781
Spearman-Brown Coefficient	Equal Length		.877
	Unequal Length		.877
Guttman Split-Half Coefficient			.876

a. The items are: Question 1, Question 2.

b. The items are: Question 3, Question 4.

11.3.3 Kuder-Richardson Formula-20

Kuder-Richardson formula-20 is appropriate for scale items which are answered **dichotomously** such as 'true - false', 'yes - no', 'present - absent' etc. Their formula 20 is

$$r = \frac{k}{k-1}\left[1 - \frac{\Sigma pq}{\sigma^2}\right] \tag{11.5}$$

where k = the number of items in the test

    p = proportion of correct response to a particular item

    q = proportion of incorrect response to that item

    σ^2 = variance of the total scores of the test.

To compute the reliability we measure the proportion of the people answering positively to each of the questions and the variance of the scores must be known. This is explained with the following example.

Example 11.4:
Ten students took a six -item test. The results were as follows:

Student	Q1	Q2	Q3	Q4	Q5	Q6
1	1	1	1	1	1	1
2	1	1	1	1	1	0
3	1	1	1	1	0	0
4	1	1	1	1	0	0
5	1	1	1	0	0	0
6	1	1	0	0	0	1
7	1	1	0	1	0	0
8	1	0	0	0	1	0
9	0	0	0	0	1	1
10	0	0	0	0	0	1

where 1 means true answer and 0 means false answer.

The distribution of the scores of the students and item scores are given in Table 11.5. Calculate the reliability coefficient using Kuder-Richardson formula-20.

Table 11.5

Student scores		Item scores	
Score	**Frequency**	**Score**	**Frequency**
6	1	8	1
5	1	7	1
4	2	6	0
3	3	5	2
2	2	4	2
1	1	-	-
Total	10	Total	6

Solution:

Score of the students

Score (x)	Frequency (f)	fx	fx^2
6	1	6	36
5	1	5	25
4	2	8	32
3	3	9	27
2	2	4	8
1	1	1	1
Total	10	33	129

$\Sigma fx = 33$, $\Sigma fx^2 = 129$

$$\sigma^2 = \frac{129}{10} - \left(\frac{33}{100}\right)^2 = \boxed{2.01}$$

Item scores

Score	Frequency	p	q	p×q
8	1	0.8	0.2	0.16
7	1	0.7	0.3	0.21
6	0	0.6	0.4	0.24
5	2	0.5	0.5	0.25
4	2	0.4	0.6	0.24
Total				1.35

Similarly $\sigma^2 = 2.01$ and $\Sigma pq = 1.35$, $k = 6$. Using (11.5), we get the reliability coefficient as:

$$r = \frac{6}{5}\left[1 - \frac{1.35}{2.01}\right] = 0.394 \quad \text{which is low.}$$

Example S11-4

The data will be in 7 columns and the data is as follows:

Student	Q1	Q2	Q3	Q4	Q5	Q6
1	1	1	1	1	1	1
2	1	1	1	1	1	0
3	1	1	1	1	0	0
4	1	1	1	1	0	0
5	1	1	1	0	0	0
6	1	1	0	0	0	1
7	1	1	0	1	0	0
8	1	0	0	0	1	0
9	0	0	0	0	1	1
10	0	0	0	0	0	1

Where 1 means true answer and 0 means false answer.

We calculate Kuder-Richardson formula-20 as follows:

Analyze→ Scale→ Reliability Analysis…
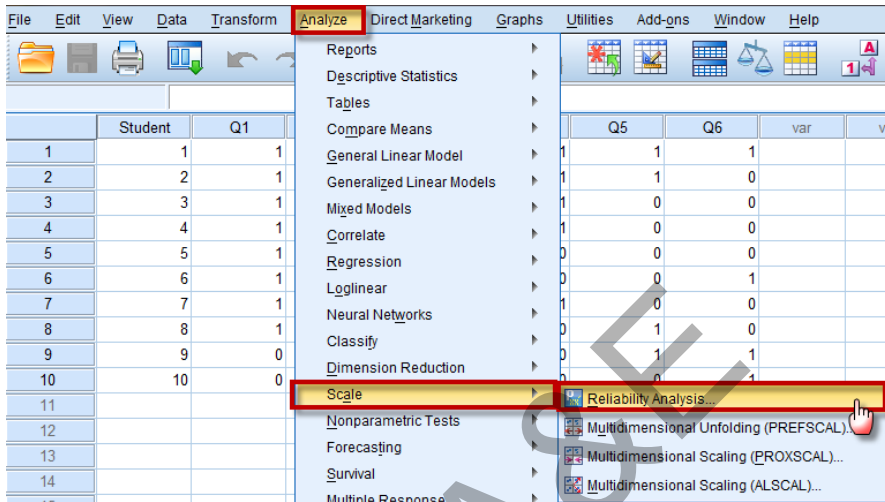
Move the variables "Question 1,…, Question 6" to Items:
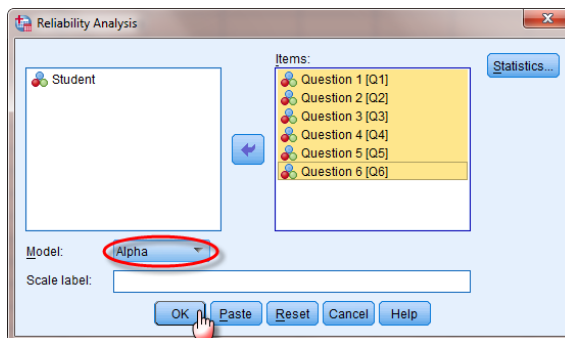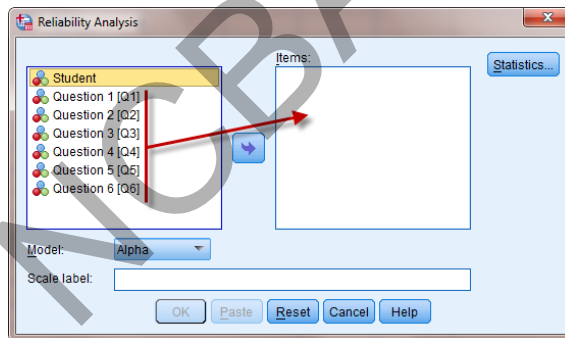
Now click on OK , to get the following output:

Reliability Statistics

Cronbach's Alpha	N of Items
.394	6

Kuder-Richardson

Note that in case of 0/1 response, Kuder-Richardson is the same as Cronbach's Alpha.

11.3.4 Cronbach's Alpha (α)

Cronbach's alpha is an extension of KR-20, allowing it to be used when there are more than two response alternatives. If alpha were used with dichotomous items, the result would be identical to KR-20. The formula for alpha is very similar to KR-20, except that the standard deviation for each item is substituted for p q

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\Sigma\,\sigma_i^2}{\sigma_T^2}\right] \tag{11.6}$$

where k = number of items

σ_i^2 = variance of the scores on a particular question or from a particular person

$\Sigma\sigma_i^2$ = sum of the rating variances for all persons

σ_T^2 = variance of the sum of the ratings from all the persons

Conceptually, both the formulas give the average of the possible split-half reliabilities of scale. This method is explained as:

Example 11.5:
This example was used as in split-half method. The data is given as:

Student	Question 1	Question 2	Question 3	Question 4	x	x^2
1	2	1	1	3	7	49
2	6	4	5	6	21	441
3	3	2	1	1	7	49
4	6	3	3	3	15	225
5	6	4	4	3	17	289
Total (y)	23	14	14	16	67	1053
y^2	529	196	196	256	1177	

Solution:

20 question scores squared $= 2^2 + 6^2 + ... + 3^2 = 283$

5 student totals squared $= 7^2 + 21^2 + ... + 17^2 = 1053$

4 question totals squared $= 23^2 + 14^2 + 14^2 + 16^2 = 1177$

For the solution we will calculate the variance of the total score as:

$$\sigma_t^2 = \frac{1053}{5} - \left(\frac{67}{5}\right)^2 = 31.0$$

$$\Sigma\sigma_i^2 = \frac{283}{5} - \frac{1177}{5^2} = 9.5$$

Using formula (11.6), we get

$$\frac{4}{3}\left[1 - \frac{9.5}{31.0}\right] = 0.924$$

Thus the reliability question 0.924 whereas in split-half method 0.9278.

The IBM-SPSS package were used as follows:

Example S11-5

The data will be in 5 columns and the data is as follows:

Student	Q1	Q2	Q3	Q4
1	2	1	1	3
2	6	4	5	6
3	3	2	1	1
4	6	3	3	3
5	6	4	4	3

We calculate Cronbach's Alpha as follows:

Analyze→ Scale→ Reliability Analysis…
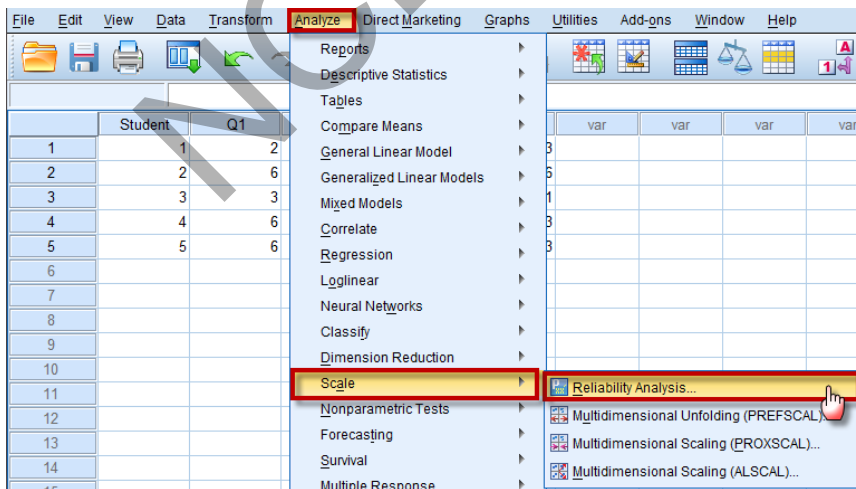
Move the variables "Question 1,…,Question 4" to Items:
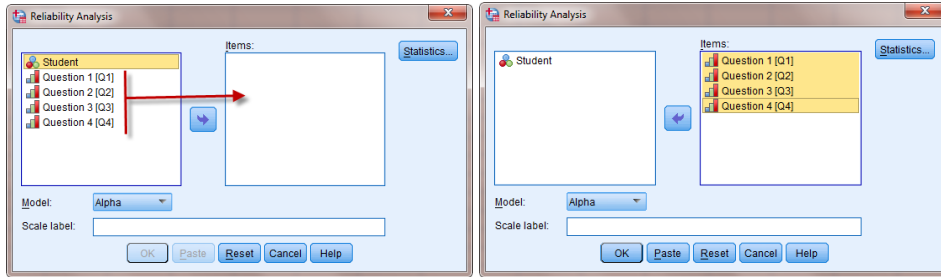
Now click on [OK], to get the following output:

Reliability Statistics

Cronbach's Alpha	N of Items
.924	4

Bibliography

1. Abramson, J.H. (1990). *Survey methods in community medicine.* Churchill Livingstone. London.

2. Abramson, J.H. (1994). *Making sense of data.* Oxford University Press. London.

3. Agresti, A. (1996). *An introduction to categorical data analysis.* John Wiley and Sons. New York.

4. Ahmad, A.M. (1998). Hyperthyroidism treatment with radioactive iodine: A 16-year retrospective analysis. *Unpublished report. Wansbeck General Hospital, Asbington. England* pp120.

5. Barlow D.H. and Hersen, M. (1984). *Single case experimental design: strategies for studying behaviour change.* Pergamon Press New York.

6. Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society.* A. 160, 268-282.

7. Bland, M. (1997). *An introduction to medical statistics 2nd edition.* Oxford University Press, England.

8. Blum, R.A., Wilton, J.H.Y., Hilligoss, D.M., Gardner, M.J., Henry, G.E., Harrison, N.J. and Schentag, J.J. (1991). Effect of fluconazole on the disposition of phenytoin. *Clin. Pharma. Therapeutics.* 49, 420-425.

9. Breslew, N.E., and Day, N.E. (1980). *Statistical methods in cancer research Vol. I: The analysis of case- control studies.* Lyon: International Agency for Research on Cancer.

10. Brewer, K.R.W. and Hanif, M. (1983). *Sampling with unequal probabilities.* Springer Verlag. New York.

11. Brown, C.C. (1982). On the goodness of fit for logistic model based on score statistics. *Commun. Statist.*, 11, 1087-1105.

12. Chatelier, G., Day M., Bobrie, G and Menard, J. (1995). Feasibility study of N of 1 trials with blood pressure self-monitoring in hypertension. *Hypertension,* 25, 294-301.

13. Cochran, W.G. (1950). The comparison of percentage in matched samples. *Biometrika,* 37, 256-266.

14. Cochran, W.G. (1977). *Sampling Techniques.* John Wiley and Sons. New York.

15. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measures.* 20, 37-46.

16. Conover, W.J. (1980). *Practical non-parametric methods.* John Wiley and Sons New York.

17. Cronbach, L.J. (1957). The two disciplines of the scientific psychology. *The American psychologist*, 12, 671-684.

18. Daniel, W.W. (1991). *Bio-statistics: A foundation for analysis in the health sciences.* John Wiley and Sons New York.

19. Drrbin, J. and Watson, G.S. (1951). Testing for serial correlation in least square regression, II. *Biometrika*, 30, 159-178.

20. Edgington, E.S. (1984). Statistics and single- case analysis. *Progress in Behaviour Modification,* 16, 83-119.

21. Everitt, B.S. (1992). *The analysis of contingency tables.* Chapman and Hall. London.

22. Fleiss, J.L. (1980). *Statistical methods for rates and proportions.* John Wiley and Sons. New York.

23. Friedman, H. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.*, 32, 675-701.

24. Geisser, S. and S. Greenhouse. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Sciences*, 29, 885-891.

25. Greenhouse, S.W. and Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24, 95-112.

26. Guyat, G.H., Keller, J.L., Jaeschke, R., Rosenbloom, D., Adachi, J.P. and Newhouse, M.T. (1990). The n– of – 1 randomized controlled trial: clinical usefulness, our three- year experience. *Ann. Intern Med.*, 112, 293-299.

27. Guyatt, G.H., Heyting, A., Jeaschke, R., Keller, J., Adachi, J. D and Roberts, R.S. (1990). N of 1 randomized trials for investigating new drugs. *Controlled Clin. Trial.,* 11, 88-100.

28. Guyat, G., Scakett, D. and Adachi, J. (1988). A clinician's guide for conducting randomized trials in individual patients. *Cana. Med. Assoc.*, 139, 497-503.

29. Guyatt, G., Sackett, D., Taylor, D.W., Chong, J., Roberts, R. and Pugsley, S. (1986). Determining optimal therapy-randomized trials in individual patients. *N. Engl. J. Med.*, 314(14), 889-92.

30. Hauck, W.W. and Donner, A. (1977). Wald's test as applied to hypothesis in logit analysis. *J. Amer. Statist. Assoc.*, 72, 851-853.

31. Hosmer, D.W. and Lemeshow, S. (1989) *Applied logistic regression.* John Wiley and Sons. New York.

32. Huynh, H. (1978). Some approximate tests for repeated measurement design. *Psychometrika*, 43, 161-175.

33. Huynh, H. and Feldt, L.S. (1976). Estimation of the Box correction for degrees of freedom form sample data in randomized and split plot designs. *J. Edu. Statist.*, 1(1), 69-82.

34. Jaeschke, R., Adachi, J., Guyatt, G., Keller, J. and Wong, W. (1991). Clinical usefulness of amitriptyline in fibromyalgia: the results of 23 n- of-1 randomized controlled trials. *J. Rheumatol.*, 18, 447-451.

35. Johannessen, T. (1991). Controlled trials in single subjects- value in clinical medicine. *B. Med. Jour.*, 303, 173-174.

36. Johannessen, T., Petersen, H., Kristensen, P. and Fosstvedt, D. (1991). The controlled single subject trial. *Scand. J. Prim. Health Care*, 9, 71-91.

37. Johannesses, T., Fosstvedt, D. and Petersen, H. (1991). Combined single subject trials. *Scand. J. Prim. Health Care*, 9, 23-27.

38. Kleinbaumn, D.G. (1994). *Logistic regression.* Springer Verlag. New York.

39. Knapp, R.G. and Miller, M.C. (1992). *Clinical epidemiology and bio- statistics.* Harwal Publishing Company. Malvern, Pennsylvania.

40. Kruskal, W.H. and Wallis, W.A. (1952). Use of rank in one- criterion analysis of variance. *J. Amer. Statist. Assoc.*, 47, 583-621.

41. Landis, J.R. and Koch, G.G. (1977). The measurements of observer agreement for categorical data. *Biometrices*, 33, 159-174.

42. Levene, H. (1960)**.** *In contribution to probability and statistics*. Stan University Press. p-278. USA.

43. Lewis, J.A. (1991). Controlled trials in single subjects- limitation of use. *B. Med. Journal.*, 147(suppl.) 40-45.

44. Link, E.D. and Boigo, I.N. (1992). N of 1 randomized control trial for BPD patients. *Can. J. Psychiatry*, 37, 148.

45. Mann, H.D. and Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Ann. Math.*, 18, 50-60.

46. Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Institute*, 22, 719-748.

47. Maxwell, A.E. (1961). *Analyzing qualitative data.* Methuen and Co. Ltd. London.

48. March, L., Irwig, L., Schwarz, J. and Simpson, C. (1994). n of 1 trials comparing a non- steroidal anti-inflammation with paracetamol in osteoarthritis. *B. Med. Journal*, 309, 1041-1045.

49. McLeod, R.S., Taylor, D.W. and Cohen, Z. (1986). Single-patient randomized clinical trial: use in determining optimum treatment for patient with inflammation of Kock continent ileostomy reservoir. *Lancet.* 1, 726-728.

50. McNemar, Q. (1955). *Psychological Statistics.* John Wiley and Sons. New York.

51. Miettinen, O.S. (1969). Individual matching with multiple controls in case of all- or- non responses. *Biometrices,* 26, 239-255.

52. Miliken, G.A. and Hohnson, D.E. (1984). *Analysis of Messy Data*, Volume 1: Designed Experiments. New York: Van Nostrand Reinhold.

53. Nuovo, J., Ellsworth, A.J. and Larson, E.B. (1986). Treatment of atopic dermatitis with antihistamines: lessons from a single- patient randomized clinical trial. *J. Amer. Board. Family Practice.* 5, 137-141.

54. Patel, A., Jaeschke, R., Guyatt, G.H., Keller, J.L. and Newhouse, M.T. (1991). Clinical usefulness of N- of- 1 randomized controlled trials in patients with nonreversible chronic airflow limitation. *Amer. Rew. Respir. Dis.*, 44, 962-964.

55. Robin, E.D. and Burke, C.M. (1986). Single- patient randomized clinical trial: opiates for intractable dyspnea. *Chest.* 90, 888-892.

56. Samiuddin, M., Hanif, M. and Asad, H. (1978). Some comparison of the Bartlett and cube root tests of homogeneity of variance. *Biometrika*, 65, 218-221.

57. Spiegelhalter, D.J. (1988). Statistical issues in studies in individual response. *Scand. J. Gastroenterol,* 147(suppl), 40-45.

58. Snedecor, G.W. and Coxhran, W.G. (1980). *Statistical methods.* John Wiley and Sons. New York.

59. Streinner, D.L. and Norman, G.R. (1995). *Health measurement scales.* Oxford University Press. London.

60. Yusuf, C.R. and Peto, R. (1984). Why do we need some large, simple randomized trials? *Stat. Med.,* 3, 409-420.

61. Wallenstein, S. Patel, H. and Fava, G. (1990). Two treatment crossover designs. In Peace K. Ed. *Statistical issues in drug research and development.* New York. Marcel Dekker.

Subject Index