PROCEEDINGS

OF

"WORKSHOP ON BIOLOGICAL DATA ANALYSIS WITH HANDS ON TRAINING ON STATISTICAL SOFTWARES"

ORGANIZED BY

UNIVERSITY OF HEALTH SCIENCES LAHORE

AND

HIGHER EDUCATION COMMISSION PAKISTAN

MARCH 22 – 25, 2011

WORKSHOP COMMITTEES

Patron

Prof. Malik Hussain Mubbashar, Vice Chancellor / Chief Executive, UHS Lahore.

Advisor

Col (retd) Jawaid Iqbal, Director (Admin & Coord), UHS Lahore.

Principal Organizer

Mr. Waqas Sami

Event Organizing Committee

Mr. Ijaz Hussain,

Mr. Ali Fayyaz,

Registration Committee

Mr. Harris Aziz

Mr. M. Nadeem

Publication In-charge

Mr. M. Atif

Proceedings of the Workshop

Editors

Mr. Waqas Sami

Mr. M. Atif

Mr. M. Hassan Hashmi

ACKNOWLEDGEMENT

University of Health Sciences Lahore organized the 04 days workshop on Biological Data Analysis with Hands on Training on Statistical Softwares.

We at University of Health Sciences express our heartiest gratitude to Higher Education Commission Pakistan for providing financial support. My sincere thanks to Prof. M. H. Mubbashar, Vice Chancellor / Chief Executive, UHS Lahore for his invaluable input and heartfelt advices in organizing this workshop. I also extend my sincere thanks to Col (retd) Jawaid Iqbal, Director (Admin & Coord), UHS Lahore, who played a pivotal role in organizing this event. I also thank the Organizing Committee members, resource persons, participants, other well – wishers and friends for their encouragement and input in making this workshop a success.

Waqas Sami

TABLE OF CONTENTS

ABOUT THE WORKSHOP

Medicine is a science in which chance plays a very significant role. Biostatistics, as science helps to quantify the contribution of chance and as an art helps medical researchers to make valid conclusions from their study, helps clinicians in making diagnostic, prognostic or therapeutic decisions and help policy makers to plan, monitor and evaluate public health initiatives.

Medical students / researchers / Clinicians are facing a growing flood of data that they need to digest in their research. However, their ability soon surpasses their capacity to analyze and make sense of the data generated in a timely fashion. Pakistan is a country where research culture has recently developed; medical researchers have limited knowledge of statistical methods used in analyzing data, interpreting results and use of statistical software like SPSS, STATA, Epi info, SAS and R, etc.

This workshop was designed to help the novice and existing researchers to develop a clear understanding of analyzing biological data using correct statistical methods including interpretation of results and reporting in statistical language. The workshop encompassed topics that were in line with the best international practices with special emphasis on “Sample Size Calculation”. The participants were given hands – on training on contemporary statistical softwares followed by pre-post assessment test.

OBJECTIVES

- To develop understanding on right approaches towards biological data analysis.
- To increase awareness of statistical considerations, stress on application and interpretation of results.
- To orient participants towards hypothesis testing, logistic regression techniques and analyzing categorical data.
- To give an understanding on the concepts of sample size and its calculation in health research.

- To provide hands on training on statistical softwares for analyzing qualitative and quantitative nature data.

WORKSHOP MATERIAL

The printed material was distributed among all the participants before the inception of each session. For every lecture, the material contained the objectives of the presentation and an outline of the topics to be covered.

SESSION FORMAT

- Presentations and panel discussions
- Live demonstration on how to analyze data
- Small group discussion sessions
- One-on-one sessions
- Special interest group sessions
- Hands – on training sessions

TARGET PARTICIPANTS

Students / researchers / faculty members who were working in various fields of health sciences and wanted to enhance their data analysis, statistical methods and data presentation techniques. There was no previous mathematics / statistics background required for this workshop.

NUMBER OF PARTICIPANTS

The number of participants were limited to 30 and nominations were sought from Postgraduate Medical Institute, Institute of Public Health and University of Health Sciences Lahore.

REGISTRATION

The deadline of registration was March 10, 2011.

Mr. Waqas Sami
Statistical Officer, UHS Lahore

Workshop Facilitator / Coordinator

EXECUTIVE SUMMARY

The 04 days workshop was organized by University of Health Sciences in collaboration with Higher Education Commission Pakistan. The workshop drew array of participants (postgraduate medical students, medical researchers and faculty members) from public postgraduate institutions.

A total of eight (08) presentations were made on topics such as Introduction to Biostatistics, Descriptive Statistics, Hypothesis testing (parametric and non-parametric approach), Regression analysis (qualitative & quantitative dependent variables), categorical data analysis and sample size calculation in health research. In sessions – III of all 04 days the participants were given hands on training on computers on all the covered topics.

The first day of the workshop witnessed the opening ceremony with addresses by Prof. John Biggs (*former Dean of Postgraduate Medicine in Cambridge University, Chairman of the ACT Health Human Research Ethics Committee, Australia and Adjunct Professor, Medical Education at University of Health Sciences Lahore).* Prof. Malik Hussain Mubbashar (*Vice Chancellor / Chief Executive University of Health Sciences Lahore)* and Dr. M. Al Rukban *(Dean, College of Medicine, Majmaah University Saudi Arabia).*

The chief guest of the occasion, Prof. John Biggs in his keynote address highlighted the importance of Biostatistics in medical research and shared his views on the misuse of statistics reported in medical journals. Dr. Biggs said that reliance on statistics during diagnosis might lead to wastage of resources and incorrect conclusions. He said that statistics are crucial. “They can affect judgment whether an individual will live or die, health will be protected or jeopardized, and whether the medical science will advance or get sidetracked," he added.

Prof Malik H Mubbashar, UHS Vice Chancellor, in his special address said that medicine is a science in which chance plays a very significant role. Biostatistics, as science helps to quantify the contribution of chance and as an art helps medical researchers to make valid conclusions from their study, helps clinicians in making diagnostic, prognostic or therapeutic decisions and help policy makers to plan, monitor and evaluate public health initiatives. On the other side, inappropriate uses of statistical methods are found in every stage of medical research related to data analysis; design of the experiment, data collection and pre-processing, analysis method, implementation, and interpretation and urged associating biostatisticians early in research projects lest the later stages of the studies were affected.

Dr. Rukban appreciated the efforts of University of Health Sciences in organizing such educational event for medical professionals. He said biostatistics is now the better half of medical sciences in which the thinking of a researcher is proved by the analytics of biostatistics.

Mr. Waqas Sami (facilitator & resource person) said that the general problem faced by medical students / professionals in learning biostatistics subject is the medium of instruction; for them the medium of instruction should be problem oriented instead of technique oriented. The technique oriented medium involves manual calculation, derivation of formulas etc which make the subject learning boring. He further added that Pakistani researchers lagged behind in research publications because of not having enough statistical knowledge. In Pakistan, 69 various categories of medical journals are being published which so far have never been audited for statistical flaws. Moreover, 80% of these medical journals are working without expert biostatisticians.

Mr. Waqas Sami offered the vote of thanks at the end of inaugural session.

After the inaugural session, the participants were asked to appear in the pre assessment test. Two presentations were made on the first day followed by hands on training on computers. The first session was on Introduction to Biostatistics and Data Entry in SPSS 18.0, whereas, the second session was on descriptive statistics.

Two presentations were made on the second day of the workshop followed by hands on training on statistical softwares. The first presentation was made on Hypothesis testing and tests of significance using parametric approach. The second presentation was made on Non-Parametric techniques.

Two presentations were made on the third day of the workshop followed by hands on training on statistical softwares. The first presentation was made on Regression Analysis (qualitative and quantitative dependent variables). The second presentation was made on categorical data analysis.

Two presentations were made on the fourth day of the workshop. The first presentation was made on Sample Size Calculation in Health Research. The second presentation covered the Sample Size Calculation using PASS 2008 software. The fourth day of the workshop concluded with distributing certificates to the participants. Prior to certificate distribution, participants were requested to fill in feedback forms and to take post-assessment test.

The four days workshop provided the medical professionals with the opportunity to learn the subject of Biostatistics and to analyze the data using correct statistical techniques. Special emphasis was laid on hands on training and sample size calculation. The pre-post assessment test results showed a dramatic difference highlighting that the workshop was successful in meeting its aims and objectives.

WORKSHOP PROGRAMME

Day 1 (March 22.03.2011)

Time	Activity	Facilitator(s)	Location
08:15 – 08:45 a.m.	Registration	Registration Committee	Reception Counter
09:00 – 9:50 a.m.	Recitation from the Holy Quran Chief Guest – Prof. John Biggs Special Address – Vice Chancellor		Senate Hall
10:00 – 10:20 a.m.	Tea Break	Caretaker	Dining Hall # 2
10:30 – 11:20 a.m.	Introduction to Biostatistics and Data Entry in SPSS 18.0	Mr. Waqas Sami	Workshop Room
11:30 – 01:00 p.m.	Descriptive Statistics	Mr. Waqas Sami	Workshop Room
01:00 – 01:55 p.m.	Lunch / Prayer Break	Caretaker	Dining Hall # 2
02:00 – 04:00 p.m.	Hands on Data Entry and Descriptive Statistics Training on Computers	Mr. Waqas Sami & Mr. Waqas Latif	Computer Lab

Day 2 (March 23.03.2011)

Time	Activity	Resource Person	Location
09:00 – 11:00 a.m.	Inferential Statistics	Mr. Waqas Sami	Workshop Room
11:00 – 11:20 a.m.	Tea Break	Caretaker	Dining Hall # 2
11:30 – 01:00 p.m.	Non-Parametric Techniques	Mr. M. Bilal	Workshop Room
01:00 – 01:55 p.m.	Lunch / Prayer Break	Caretaker	Dining Hall # 2
02:00 – 04:00 p.m.	Hands on Parametric and Non-Parametric Tests Training on Computers	Mr. Waqas Sami, Mr. M. Bilal & Mr. Waqas Latif	Computer Lab

Day 3 (March 24.03.2011)

Time	Activity	Resource Person	Location
09:00 – 10:30 a.m.	Regression Analysis **(Quantitative and Qualitative Dependent Variables)**	Dr. Qaiser Shahbaz	Workshop Room
10:35 – 10:55 a.m.	Tea Break	Caretaker	Dining Hall # 2
11:00 – 01:00 p.m.	Categorical Data Analysis	Mr. Nadeem Shafique	Workshop Room
01:00 – 01:55 p.m.	Lunch / Prayer Break	Caretaker	Dining Hall # 2
02:00 – 04:00 p.m.	Hands on Regression and Categorical Data Analysis Training on Computers	Mr. Nadeem Shafique, Mr. Waqas Sami & Mr. Waqas Latif	Computer Lab

Day 4 (March 25.03.2011)

Time	Activity	Resource Person	Location
09:00 – 11:00 a.m.	Sample Size Calculation in Health Research	Mr. Waqas Sami	Workshop Room
11:00 – 11:20 a.m.	Tea Break	Caretaker	Dining Hall # 2
11:30 – 01:00 p.m.	Sample Size Calculation Using P.A.S.S 2008 Software	Mr. Waqas Sami	Workshop Room
01:00 – 02:20 p.m.	Lunch / Prayer Break	Caretaker	Dining Hall # 2
02:30 – 02:50 p.m.	Certificate Distribution		Workshop Room
03:00 – 03:45 p.m.	MCQ Based Assessment Test	Mr. Waqas Sami	Workshop Room
03:45 – 04:00 p.m.	Closing Ceremony		Workshop Room

OPENING SESSION

Workshop began with the recitation from the Holy Quran by Mr. M. Hassan Hashmi. Mr. Waqas Sami (facilitator) after introducing himself welcomed the chief guest, vice chancellor, participants and resource persons at the inaugural session of the workshop.

Mr. Sami further informed the audience about the format, aims, objectives and format of the workshop and shared a famous Chinese quote "“Give a man a fish; feed him for a day. Teach a man to fish; feed him for a lifetime". Mr. Sami further added that the general problem faced by medical students / professionals in learning biostatistics subject is the medium of instruction; for them the medium of instruction should be problem oriented instead of technique oriented. The technique oriented medium involves manual calculation, derivation of formulas etc which make the subject learning boring.

Prof. John Biggs (Chief – Guest) *former Dean of Postgraduate Medicine in Cambridge University, Chairman of the ACT Health Human Research Ethics Committee, Australia and Adjunct Professor, Medical Education at University of Health Sciences Lahore)* said; "Worthy Vice Chancellor, Distinguish guests, dear participants, ladies and gentlemen, I am indeed honoured to be invited as Chief Guest to deliver a keynote speech at the workshop on *Biological Data Analysis with Hands on Training on Statistical Softwares.* I have spent my life watching and conducting research but I still feel we lack behind basic understanding of the tool that helps us in decision making i.e. biostatistics". Prof. Biggs was of the view that careful and accurate use of statistics in medical research was of major importance and, therefore, must be enforced emphatically. "The use of statistics in medical diagnosis and biomedical research may affect whether individuals live or die, whether their health is protected or jeopardized, and whether medical science advances or gets sidetracked", Prof. Biggs further said that problem was a serious one, as the inappropriate use of statistical analysis might lead to incorrect conclusions, artificial research results and a waste of valuable resources. He shared that the statistical errors are so common that it is believed that almost 50% of medical literature has statistical flaws *(Altman DG (1991). Improving Doctors' Understanding of Statistic...* In another article written by *Welch GE (1996),* "Statistical errors were found in 19% (27/145) articles published in an obstetrics and gynecology journal. Dr. Biggs regretted that many researches were conducted by doctors who were not adequately trained in research.

The Vice Chancellor, University of Health Sciences Lahore said that medical researchers had to be encouraged to learn more about biostatistics. Biostatisticians should be involved early in study design, as mistake at this point could have major repercussions, thus, negatively affecting all subsequent stages of medical research. He further added that inappropriate use of statistical methods are found in every stage of medical research related to data analysis; design of the experiment, data collection and pre-processing, analysis method, implementation, and interpretation. A part from being highly unethical this may lead to distorted results, incorrect conclusions, and substantial waste of financial resources and may have serious clinical consequences. He pointed out that a great number of published medical researches contain statistical errors by sharing eye opening statistics.

Statistical errors in 55 Eligible manuscripts submitted to Biochemia Medica during 2006-2009

Error	Error Rate in Percentage
Sample Size not Scientifically Calculated	55/55 (100.0 %)
Incorrect Choice of Statistical tests	46/55 (83.63 %)
Incorrect interpretation of correlation analysis	22/55 (40.00 %)
Incorrect presentation of descriptive analysis	19/55 (35.00 %)
Incorrect interpretation of p-value	13/55 (23.63 %)

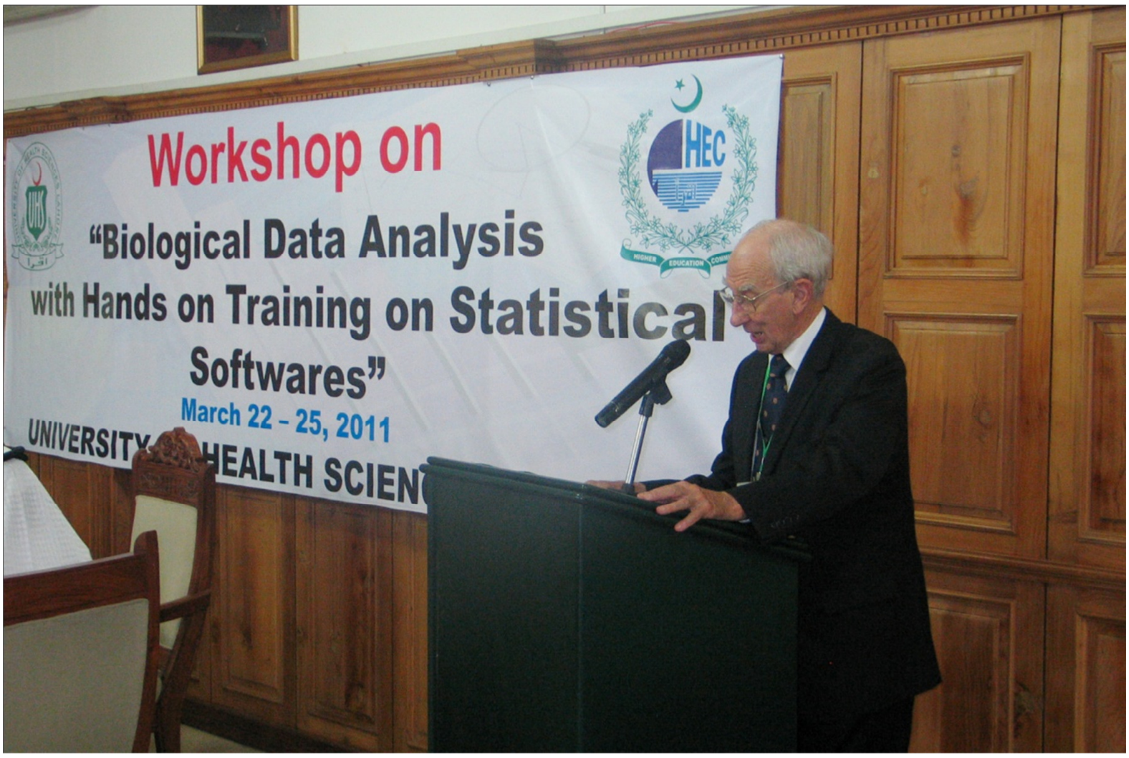
He further shared the misuse of standard error of the mean in critical evaluation of four anesthesia journals (2001).

Journal Name	Error Rate in Percentage
Anesthesia & Analgesia	112/405 (27.65%)
British Journal of Anesthesia	31/137 (22.62%)
Anesthesiology	48/257 (18.67%)
European Journal of Anesthesiology	07/61 (11.47%)

Dr. Rukban appreciated the efforts of University of Health Sciences in organizing such educational event for medical professionals. He said biostatistics is now the better half of medical sciences in which the thinking of a researcher is proved by the analytics of biostatistics. He also shared the situation of Biostatistical need in Saudi Arabia which is spending a lot of amount in gathering data on public health problems.

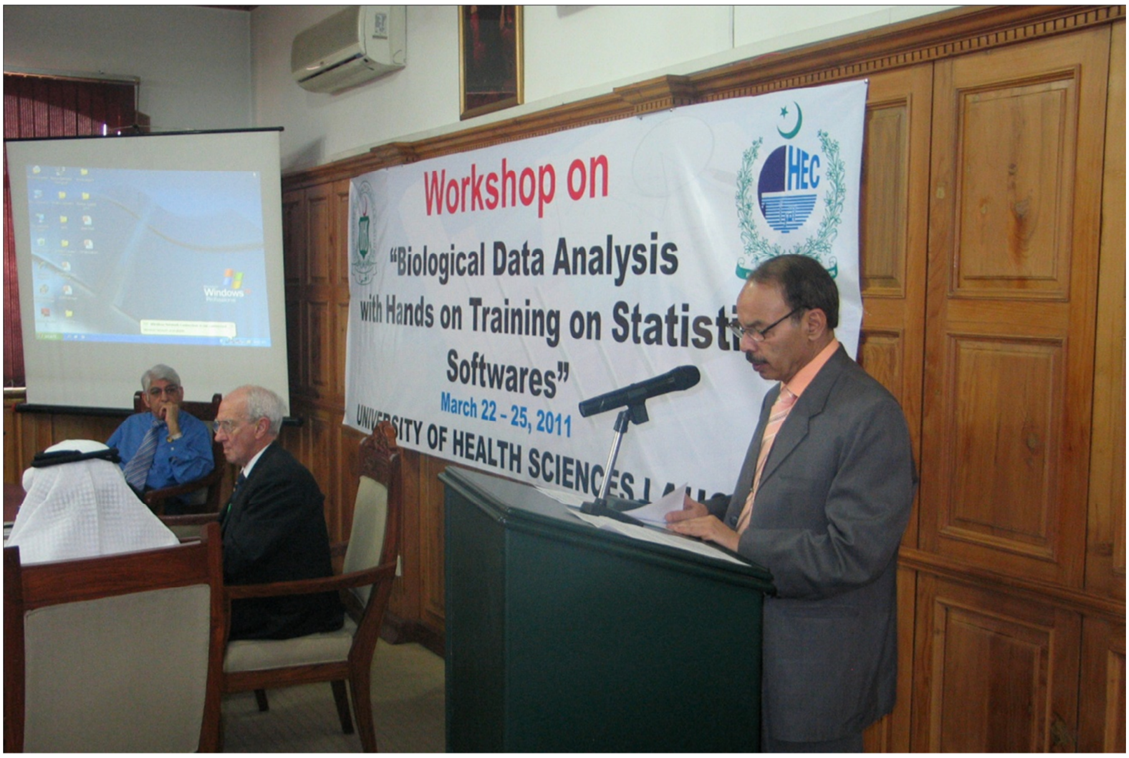
Mr. Waqas Sami (facilitator & resource person) said that Pakistani researchers lagged behind in research publications due to scarcity in statistical knowledge. In Pakistan, 69 various categories of medical journals are being published which so far have never been audited for statistical flaws. Moreover, 80% of these medical journals were without expert biostatisticians.

Mr. Waqas Sami delivered the vote of thanks at the end of inaugural session "Ladies and gentlemen! I, on behalf of University of Health Sciences, and indeed on personal note extend a very hearty vote of thanks to the Chief Guest – Prof. John Biggs for accepting our invitation as a chief guest and delivering a keynote speech; the Vice Chancellor, UHS for playing a mentoring role in organizing this event. I thank them both for gracing the inaugural session and for sharing their invaluable experiences! We all know that events cannot happen overnight. The wheels started rolling weeks ago that required meticulous planning and a birds eye for details. We had been fortunate enough to be backed by the office of Dir (Administration & Coordination) who in full swing helped in organizing this event.



Prof. John Biggs (Chief Guest) delivering the keynote speech

Participants from PGMI, IPH and UHS


Workshop on
"Biological Data Analysis
with Hands on Training on Statisti
Softwares"
March 22 - 25, 2011
UNIVERSITY OF HEALTH SCIENCES LAH
HEC

Prof. M. H. Mubbashar (Vice Chancellor / Chief Executive, University of Health Sciences Lahore) sharing importance of Biostatistics in medical sciences

Workshop on
"Biological Data Analysis
with Hands on Training on Statistic
Softwares"
March 22 - 25, 2011
HEC

Dr. Mohammed O. AlRukban, Dean College of Medicine, Majmaah University expressing views on importance of medical statistics

DAY 1

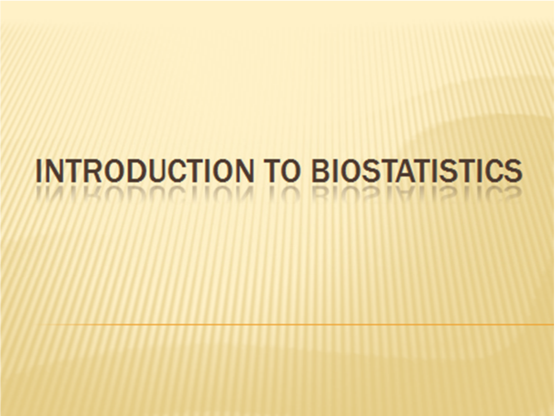
March 22, 2011

Session I : Introduction to Biostatistics & Data Entry in SPSS 18.0

Resource Person : Mr. Waqas Sami

Time : 10: 30 a.m. – 11: 20 a.m.

The resource person informed the participants that this lecture is the core of biostatistics subject, if you understand about the types of variables, can differentiate between types of scales, observations and variables then the biostatistics subject is extremely easy to understand.

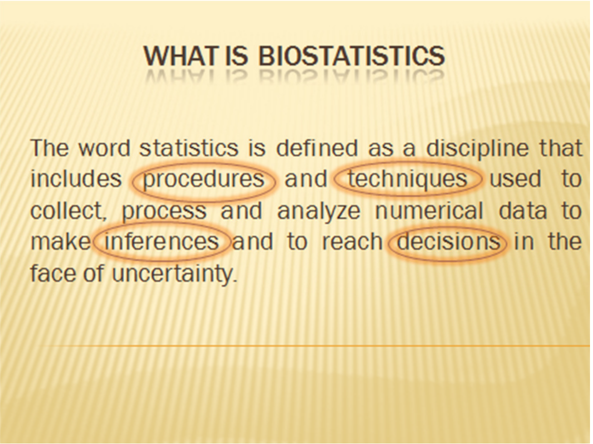


INTRODUCTION TO BIOSTATISTICS

The word biostatistics is defined as a discipline that includes procedures and techniques used to collect process and analyze numerical data to make inferences and to reach decisions in the face of uncertainty.

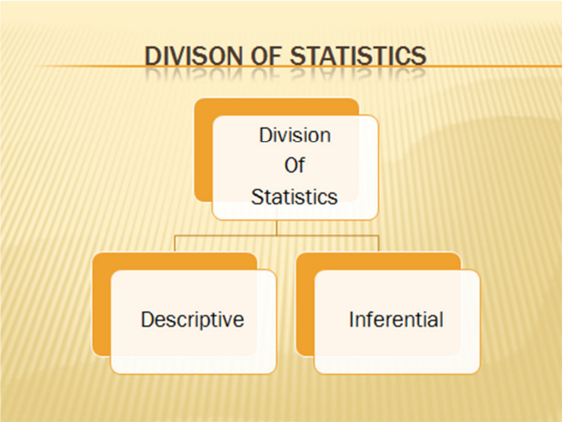
The importance of biostatistics is;

- It assists in summarizing the larger sets of data in a form that is easily understandable.
- It assists in a sound and effective planning in any field of inquiry.
- It assists in drawing general conclusions and in making predictions of how much of a thing will happen under given conditions.
- Biostatistical techniques being powerful tools for analyzing numerical data that are used in almost every branch of learning.
- A businessman, an industrialist and a research worker all employ statistical methods in their work. Banks, insurance companies and government all have their statistical departments.



WHAT IS BIOSTATISTICS

The word statistics is defined as a discipline that includes procedures and techniques used to collect, process and analyze numerical data to make inferences and to reach decisions in the face of uncertainty.



DIVISON OF STATISTICS

The further discussion highlighted the difference between descriptive statistics and inferential statistics.

Descriptive statistics is a branch of statistics which deals with concepts and methods concerned with summarization and description of the important aspects of numerical data. This area of study consists of condensation of data, their graphical displays and the computational of few numerical quantities that provide information about the centre of data.

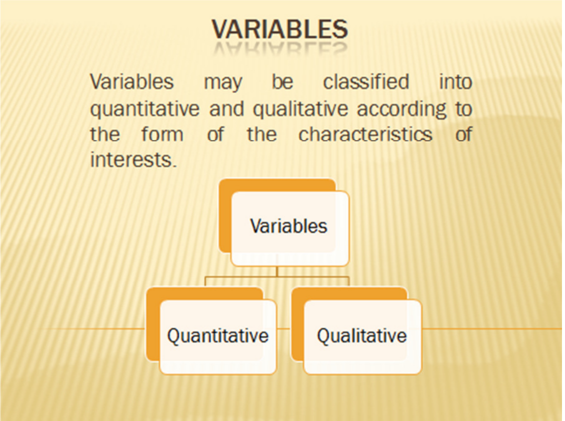
Inferential statistics deals with procedures for making inferences about the characteristics that describe the larger group of data or the whole called population, from the knowledge derived from only a part of the data known as a sample.

Observations and variables: In statistics observation means any sort of numerically recording of information, whether it is physical measurement such as height or weight, a classification such a heads or a tails, or an answer to a question such as yes or no.

Variables may be classified into quantitative and qualitative according to the form of the characteristics of interests.

A variable is said to be **qualitative** if it contains non-numerical characteristics, for example education, gender, marital status, colour of eye, satisfaction.

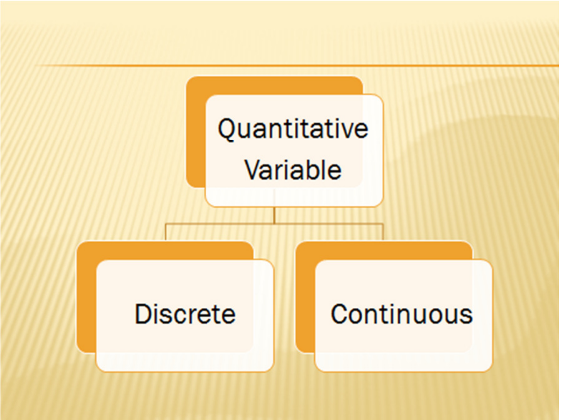
A variable is said to be **quantitative** when a characteristic can be expressed numerically such as age, weight, income etc.



VARIABLES

Variables may be classified into quantitative and qualitative according to the form of the characteristics of interests.

The resource person stressed upon that this concept is very important to understand as there are separate analysis techniques for qualitative and quantitative nature data.



Quantitative Variable

Discrete

Continuous

A **discrete variable** is one that can take only a discrete set of integers or whole numbers, a discrete variable represents count data such as the number of persons in a family, the number of rooms in a house, the number of deaths in an accident.

A variable is called a **continuous variable** if it can take on any value fractional, or within a given interval, a continuous variable represents measurement data such as the age of a person, the height of a plant, the weight of a commodity, the temperature at a place.

SCALES OF MEASUREMENTS:

Nominal scale

The type of scale which is only used to classify the variable, for example, students are classified as male and female, and number 1 and 2 can also be used to identify the categories, similarly rainfall may be classified as heavy, moderate, and light, numbers 1,2,and 3 can also be used to identify categories.

Ordinal

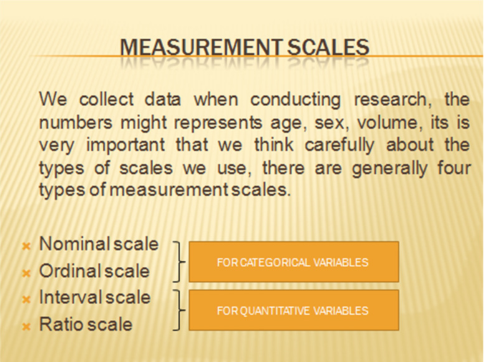
It includes the characteristics of a nominal scale and in addition has the property of ordering or ranking of measurements, for example the performance of students is rates as excellent, good, fair or poor etc, number 1,2,3 and 4 may also be used to order or rank the categories.

Interval

An interval scale is a scale of measurement where the distance between any two adjacent units of measurement (or 'intervals') is same. Scores on an interval scale can be added and subtracted but cannot be meaningfully multiplied or divided. For example, the time interval between the starts of years 1981 and 1982 is the same as that between 1983 and 1984, namely 365 days.

Ratio

It is a special kind of interval scale where the scale of measurement has a true zero point as its origin, the ratio scale is used to measure, age, distance, money etc, it can be added, subtracted, divided, multiplied etc.



MEASUREMENT SCALES

We collect data when conducting research, the numbers might represents age, sex, volume, its is very important that we think carefully about the types of scales we use, there are generally four types of measurement scales.

- Nominal scale
- Ordinal scale
- Interval scale
- Ratio scale

FOR CATEGORICAL VARIABLES

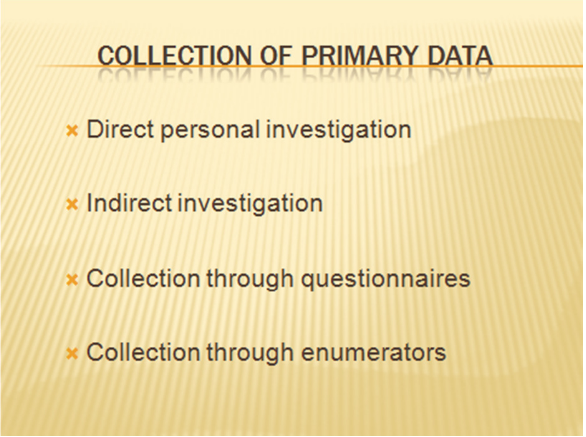
FOR QUANTITATIVE VARIABLES

TYPES OF DATA:

The most important part of statistical work is perhaps the collection of data, statistical data are collected either by complete enumeration called census which in many cases is too costly as it requires large number of enumerators and supervisory staff, or by partial enumeration associated with a sample which save much time and money.

PRIMARY DATA:

Data that have been originally collected (raw data) and have not gone any sort of statistical treatment are called primary data.



COLLECTION OF PRIMARY DATA

- Direct personal investigation
- Indirect investigation
- Collection through questionnaires
- Collection through enumerators

DIRECT PERSONAL INVESTIGATION:

In this method an investigator collects the information personally from the individuals concerned, since he interview the informants himself the information collected is generally considered quite accurate and complete, this method may prove to be very costly and time consuming when the area to be covered is vast.

INDIRECT INVESTIGATION:

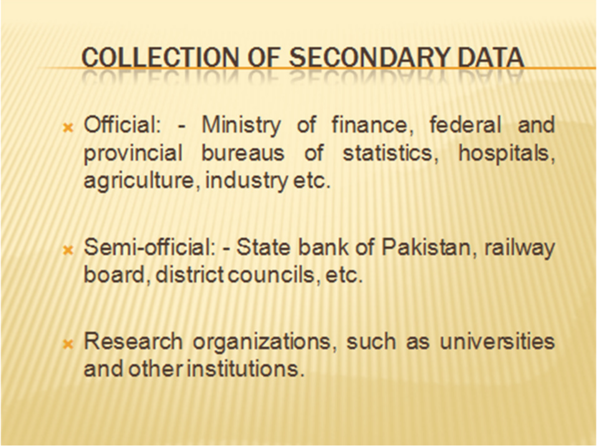
Sometimes the direct sources do not exist or the informants hesitate to respond, in such cases third parties having information are interviewed.

COLLECTION THROUGH QUESTIONNAIRE:

A questionnaire is an inquiry form comprising of a number of questions with space for entering the information asked. This method is cheap and good for extensive inquiries. This method is considered as the standard method for collection of data, it is important to note that the questions asked should be few, brief, very simple, easy for all respondents to answer and clearly worded.

SECONDARY DATA:

Data that have undergone any statistical treatment at least once i.e. the data have been classified, tabulated, or presented in some form for a certain purpose are called secondary data.

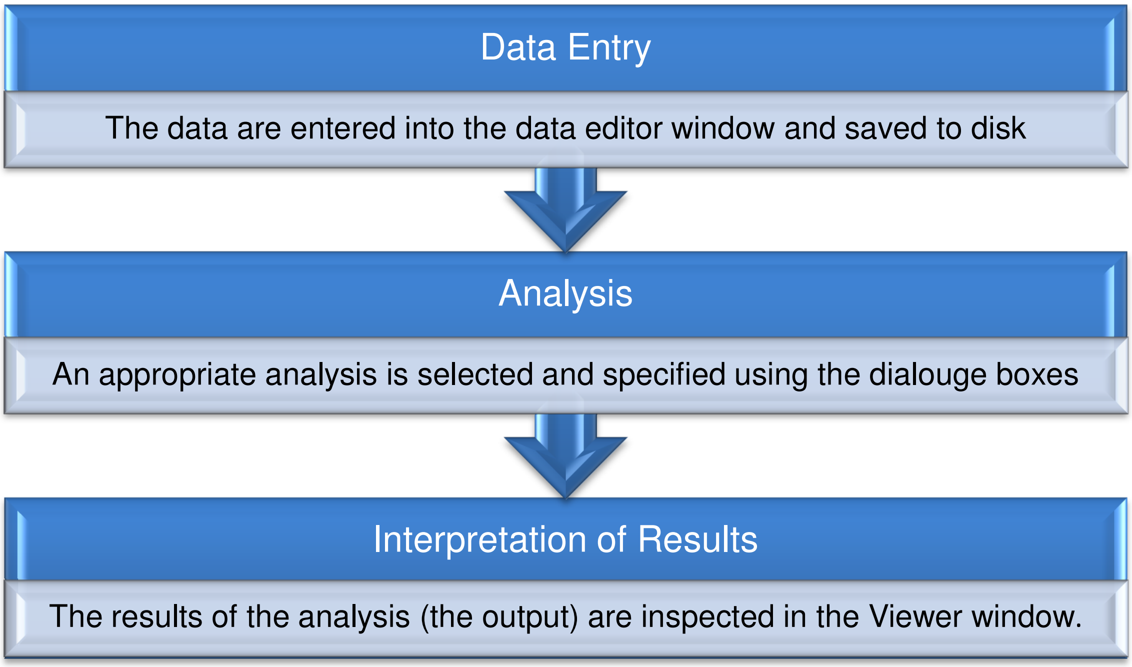


COLLECTION OF SECONDARY DATA

- Official: - Ministry of finance, federal and provincial bureaus of statistics, hospitals, agriculture, industry etc.
- Semi-official: - State bank of Pakistan, railway board, district councils, etc.
- Research organizations, such as universities and other institutions.

WORKING WITH SPSS

There are three basic concepts involved in data analysis using SPSS. Firstly you must enter a raw data and save to a file. Secondly, you must select and specify the analysis you require. Thirdly, you must examine the output produced by SPSS. These steps are illustrated below.



Data Entry

The data are entered into the data editor window and saved to disk

Analysis

An appropriate analysis is selected and specified using the dialouge boxes

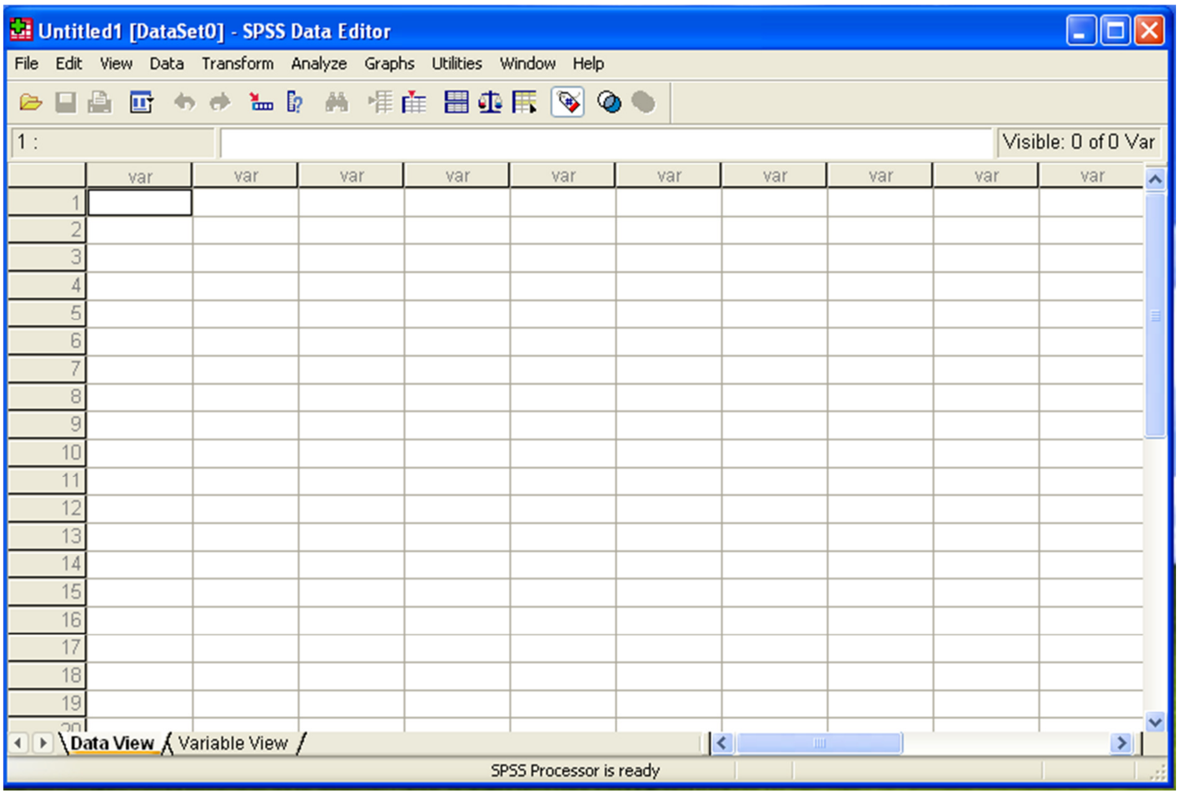
Interpretation of Results

The results of the analysis (the output) are inspected in the Viewer window.

STARTING SPSS

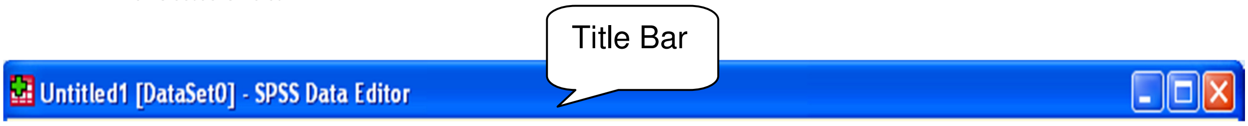
Once you have clicked on the SPSS icon, you will see the data editor window as shown below.

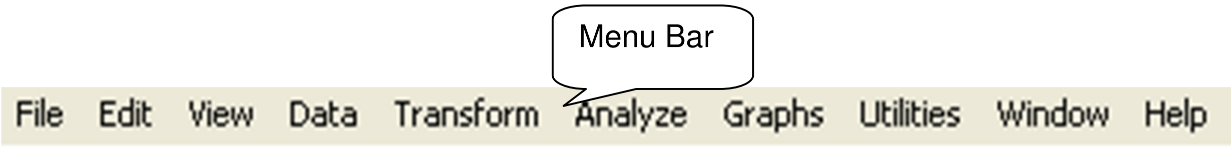
TIP: If you do not have an SPSS icon on your desktop then click on the start menu at the bottom left hand corner of the windows screen, then select programmes and then SPSS 18.0 or the version you have, drag the SPSS icon on your desktop.


Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 :
Visible: 0 of 0 Var
var var var var var var var var var var
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Data View
Variable View
SPSS Processor is ready

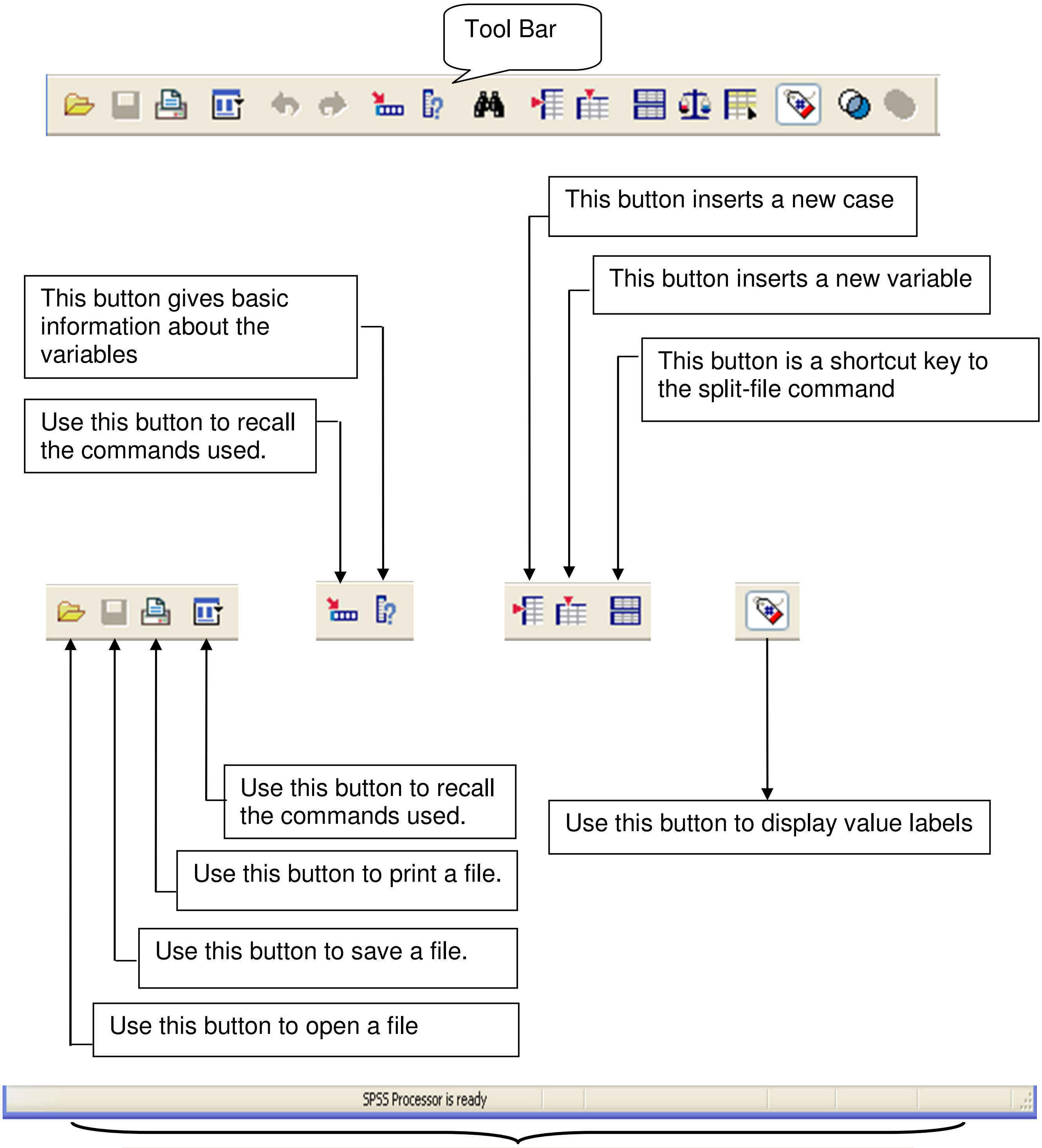
There are four main bars:

- The title bar
- The menu bar
- The tool bar
- The status bar




Menu Bar
File Edit View Data Transform Analyze Graphs Utilities Window Help

Action	Common Uses
File	❖ Open a new/existing file ❖ Open a new file into SPSS from an existing text file, Excel spreadsheet or Database ❖ Import data ❖ Save the data file ❖ Exit SPSS for Windows
Edit	❖ To make changes to the data, copy, paste, insert variables, insert cases etc
View	❖ Hide or show Status bar or Toolbar ❖ Change font or point size of the data ❖ Hide or show gridlines ❖ Switch between Data View and Variable View
Data	❖ To manipulate existing SPSS data files - Define variables, Sort cases, Merge files, Split files, Select cases, Weight cases etc.
Transform	❖ Perform computations on variables -Create new variables from existing ones. Recode old variables etc.
Analyze	❖ Contains extensive list of statistical analysis that can be conducted for example, Descriptive statistics, t-test, ANOVA, Logistic Regression, Survival Analysis etc.
Graphs	❖ To obtain high resolution plots and graphs, which can be edited in Chart Editor window.
Utilities	❖ Contains options to gives information about variables, you can write comments to a data file and set name for the variable sets etc.
Window	❖ To move to any open window or to see which window is active. The window with a check mark is the active one.
Help	❖ To get help on topics in SPSS via a Predefined List of Topics, Tutorial, Statistics Coach, Syntax Guide etc.



This button inserts a new case

This button inserts a new variable

This button is a shortcut key to the split-file command

This button gives basic information about the variables

Use this button to recall the commands used.

Use this button to display value labels

Use this button to recall the commands used.

Use this button to print a file.

Use this button to save a file.

Use this button to open a file

A status bar at the bottom of the SPSS application window indicates the current status of the SPSS processor. If the processor is running a command, it displays the command name and a case counter indicating the current case number being processed. When you first begin an SPSS session, the status bar displays the message Starting SPSS Processor. When SPSS is ready, the message changes to SPSS Processor is ready. The status bar also provides information such as command status, filter status, weight status, and split file status etc.

DATA ENTRY IN SPSS

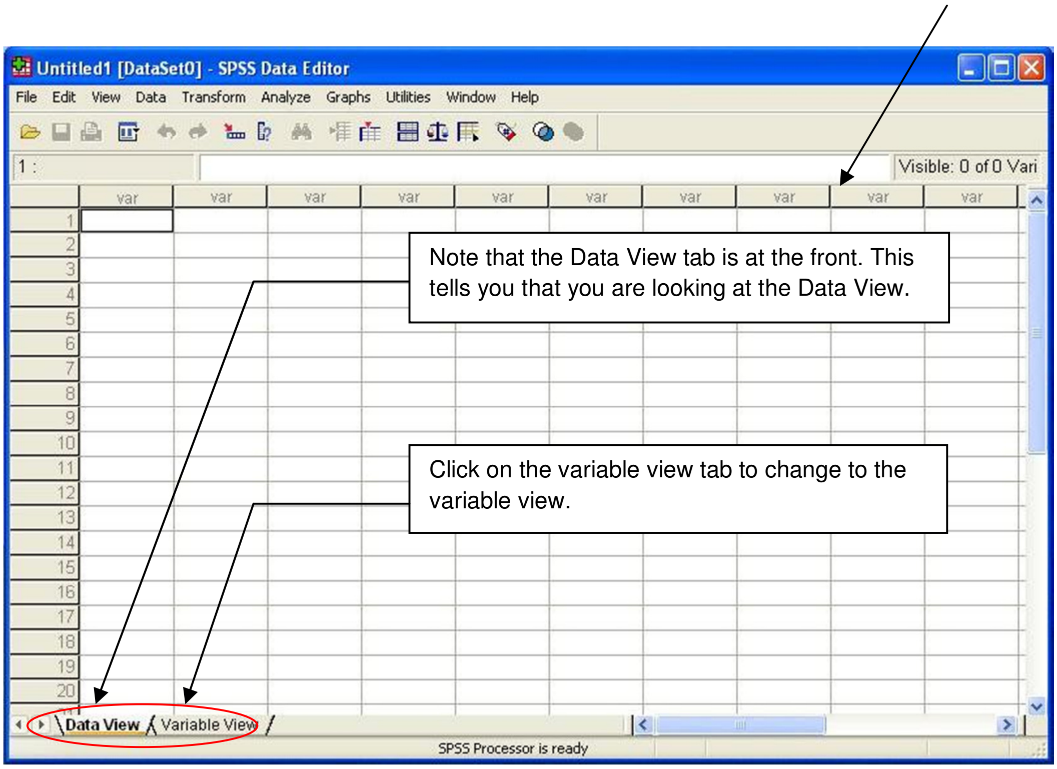
THE DATA EDITOR WINDOW

When we start SPSS, the Data Editor window is the active window. This window records all the data we want to analyze. The window is arranged as a table with a large number of cells in rows and columns, we can say this is a special sort of spreadsheet. The table can be very large, and only a small part of it can be seen through the Data Editor window at a time. You can move the scroll bars on the edges of the window to move round the table.

TIP: In the Data Editor window each row represents an individual participant and each column represents a variable.

DEFINING A VARIABLE IN SPSS

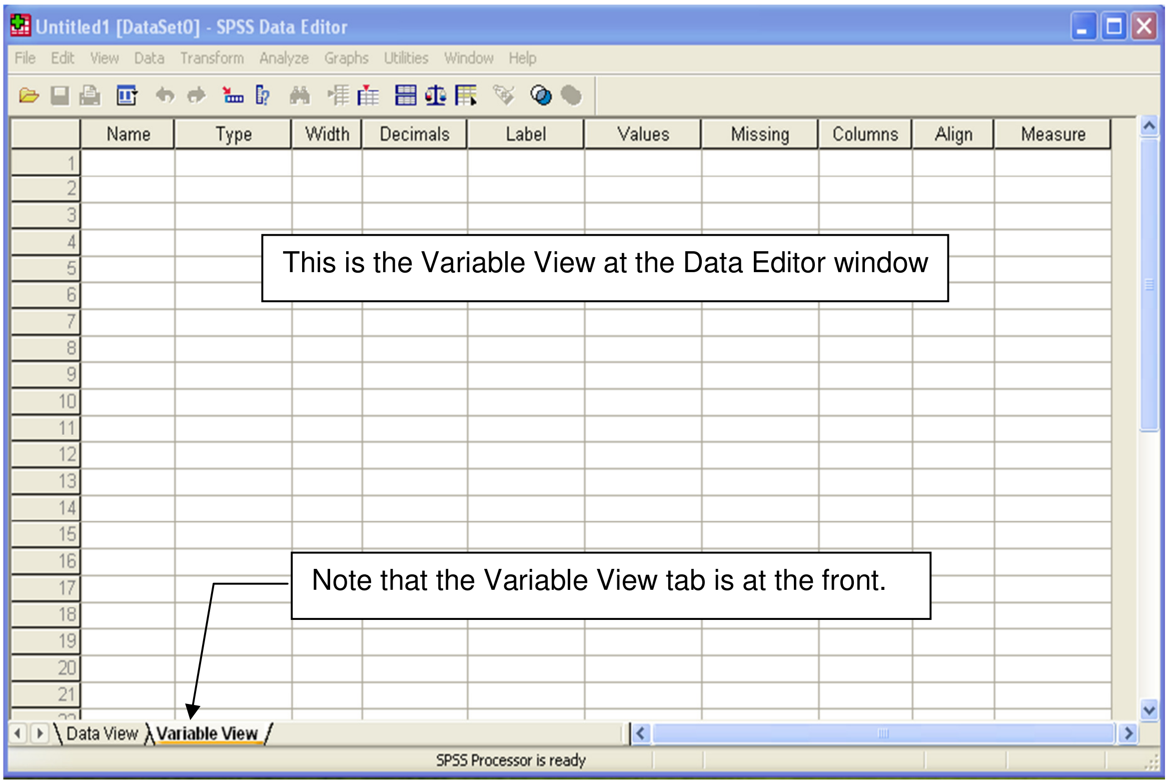
If you look at the bottom left hand corner of the Data Editor window you will notice two tabs. One tab is labeled “Data View” and the other is labeled “Variable View”. The Data View is the screen you will use to enter the data. At present this view shows an empty data table in which each of the variables is labeled **“var”.**


Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 :
Visible: 0 of 0 Vari
var var var var var var var var var var
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Data View
Variable View
SPSS Processor is ready

Note that the Data View tab is at the front. This tells you that you are looking at the Data View.

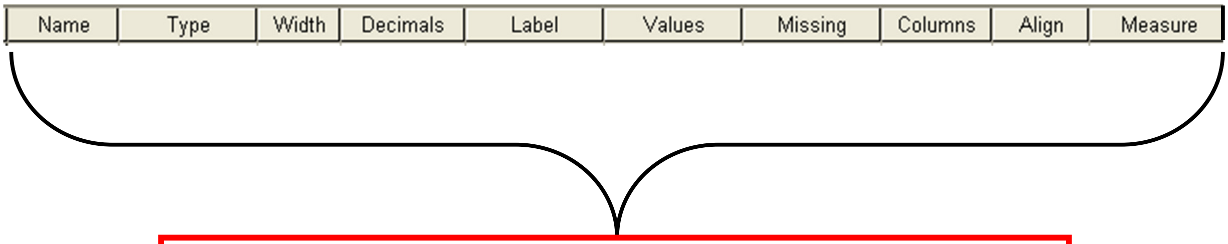
Click on the variable view tab to change to the variable view.

Before you can type your data into the data table, you need to give SPSS other important information about each of your variables. The process of defining the variables is undertaken in the Variable View. If you click on the variable view tab you will notice that in this view the columns are headed Name, Type, Width, Decimal etc.


Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Name Type Width Decimals Label Values Missing Columns Align Measure
Data View
Variable View
SPSS Processor is ready

This is the Variable View at the Data Editor window

Note that the Variable View tab is at the front.

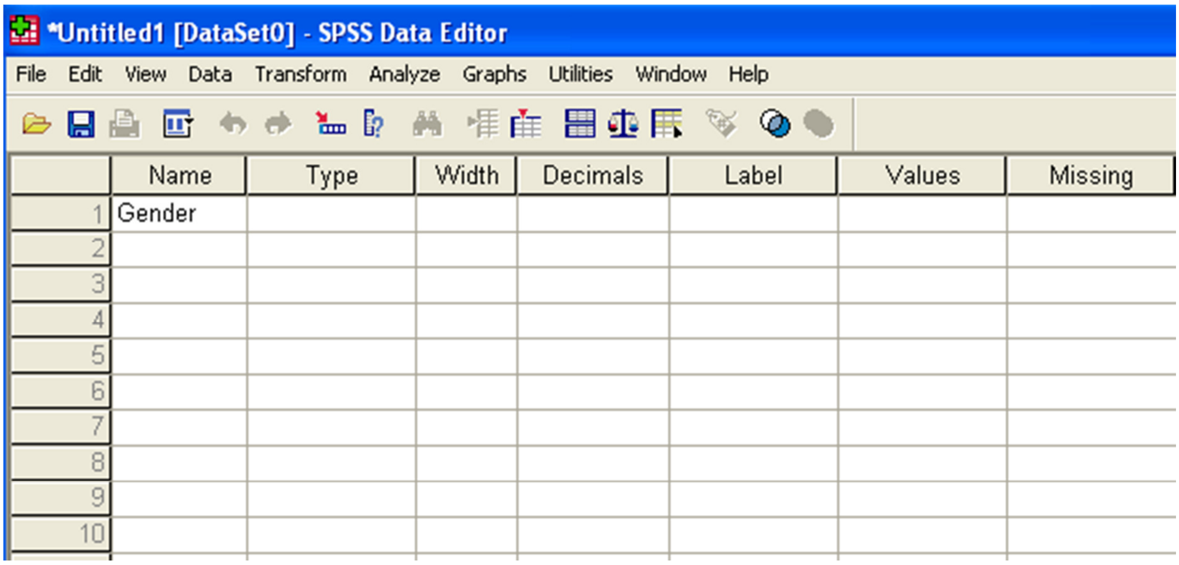


The columns represent specific characteristics of the variables. Each of them is discussed as below.

VARIABLE NAME

The first thing you need to do is to give the variable a name. Type the name of your first variable into the first row of the name column. Variable names can be 64 characters long and must start with a letter of the alphabet. Variable names cannot contain spaces or special characteristics such as colons, semicolons, hyphens or commas, full stops, @, #, and $ characteristics are not allowed. If you enter an invalid variable name SPSS will automatically warn you when you try to move from the name column.

TIP: The underline character (_) can be used in places of spaces in variable names. For example the name “Q1_1” might be used for a question type like Question 1 Part 1.

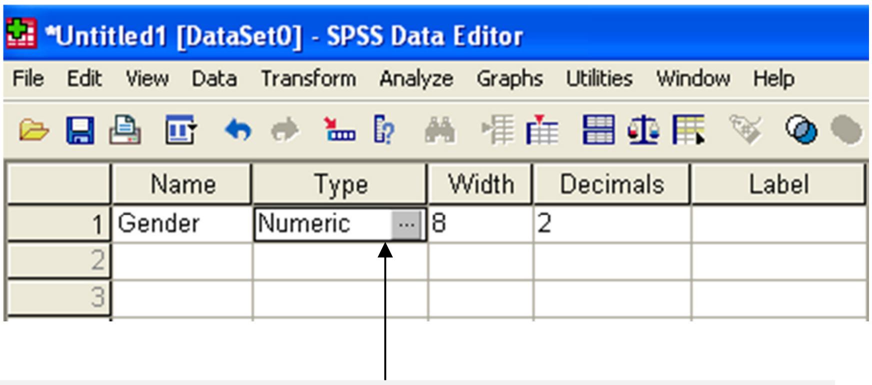


	Name	Type	Width	Decimals	Label	Values	Missing
1	Gender						
2							
3							
4							
5							
6							
7							
8							
9							
10							

Once you have entered the variable name, click the mouse or the tab key to move to the next column of the table. As you move the cursor, several other columns of the table will be filled with either words or numbers. These are the default SPSS settings. You can leave these settings as they are or you can change some or all of them before moving on to define the next variable.

VARIABLE TYPE

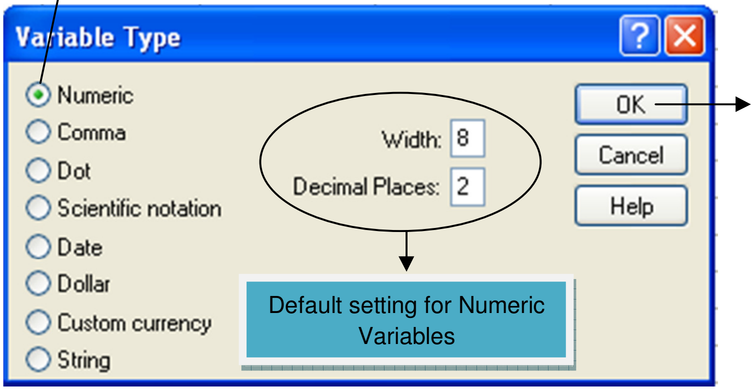
The second column in the Variable View is headed Type. SPSS can handle variables of different types. For example, variables can be numeric (containing numbers) or string (containing alphabets) or even dates in many formats. By default the Type column is set to numeric, if you want to change the variable type and click on the button that appears next to the default setting. This will call up the Define Variable Type dialogue box (see below).


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label
1	Gender	Numeric	8	2	
2					
3					

Move the cursor to the Type cell, click on this button to call up the Define Dialogue box

Select the variable type from this list. Here the type is numeric as shown by the filled circle.



● Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
○ String

Width: 8
Decimal Places: 2

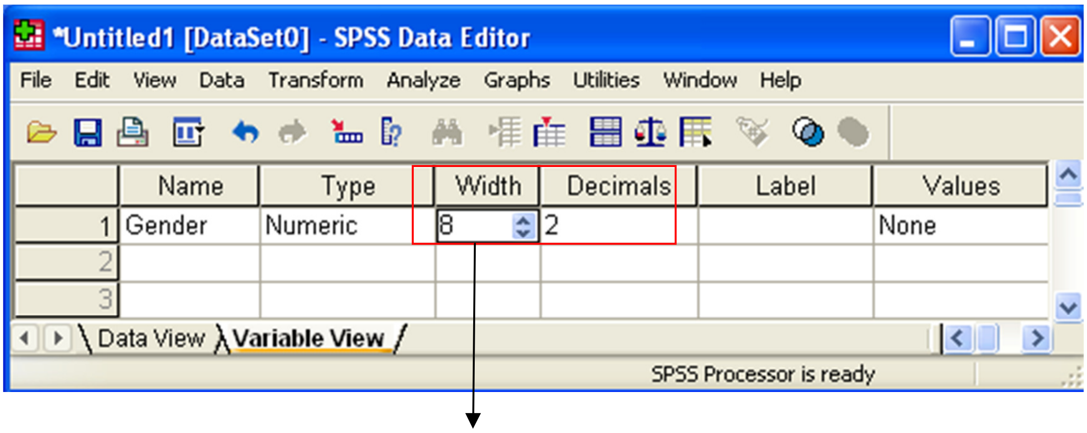
Default setting for Numeric Variables

Click the OK button to close this dialogue box.

WIDTH AND DECIMAL PLACES OF THE VARIABLES

The Define Variable Type dialogue box also allows to set the Width and Decimal Places of the variables. These settings adjust the number of characteristics before and after the decimal place used to display the variable in the Data Editor and Output windows. With numeric data the default settings are for a total Width of 8 with 2 Decimal Places for example (12345.78).

NOTE: If you attempt to input a data value that will not fit into the width, then SPSS will round it in order to display the value. However, the value you have entered is stored by SPSS and will be used in all the calculations.



*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

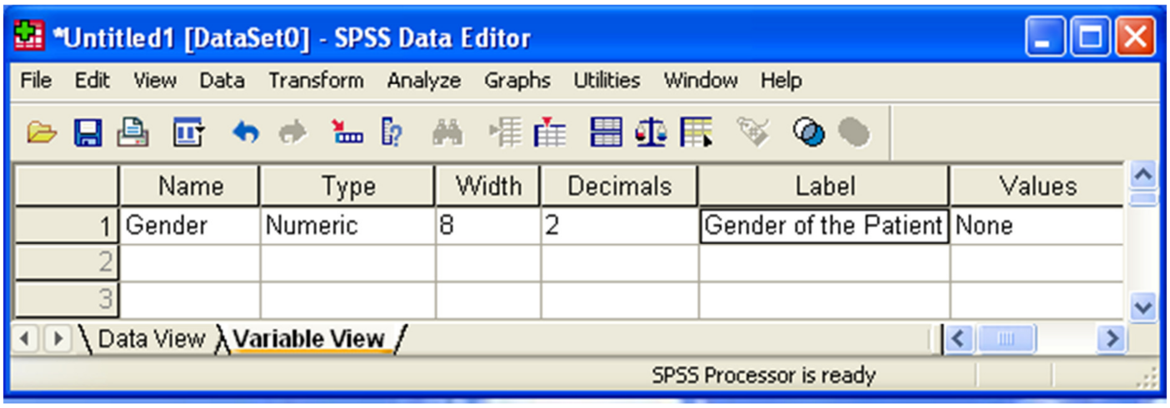
	Name	Type	Width	Decimals	Label	Values
1	Gender	Numeric	8	2		None
2						
3						

These settings can also be changed manually, by using the up and down buttons

VARIABLE LABEL

The fifth column in the Variable is headed Label. This column is used to enter a variable label. A variable label is simply a phrase that helps you to remember what data this variable contains. For example if you have entered a variable name as Gender, the Variable Label can be typed as Gender of the Patient. SPSS will not try to interpret this label, it will simply insert it into the output.

NOTE: If you have not assigned a Variable Label, SPSS will produce the output using the Name given to the variable.

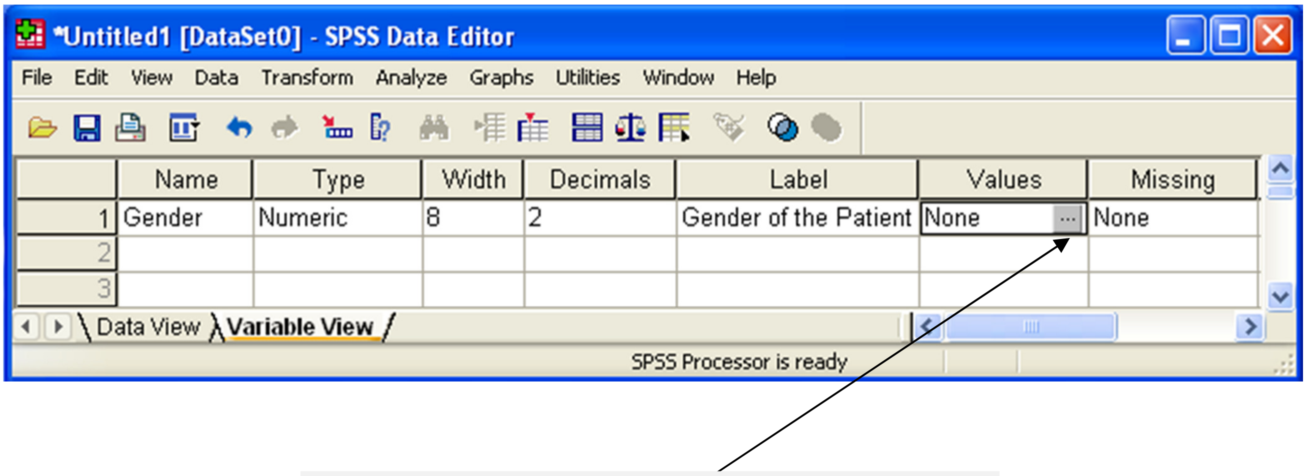


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

	Name	Type	Width	Decimals	Label	Values
1	Gender	Numeric	8	2	Gender of the Patient	None
2						
3						

VALUE LABEL

A value label is a label assigned to a particular value of a variable. Value Labels are used for coding the categorical variables measured on either nominal or ordinal scale. For example, Gender can be coded as 1 = Male, 2 = Female. Disease severity can be coded as 1 = Mild, 2 = Moderate, 3 = Severe. In SPSS Value Labels are entered using the Values column, at present this column contains the word none as no values have been assigned.

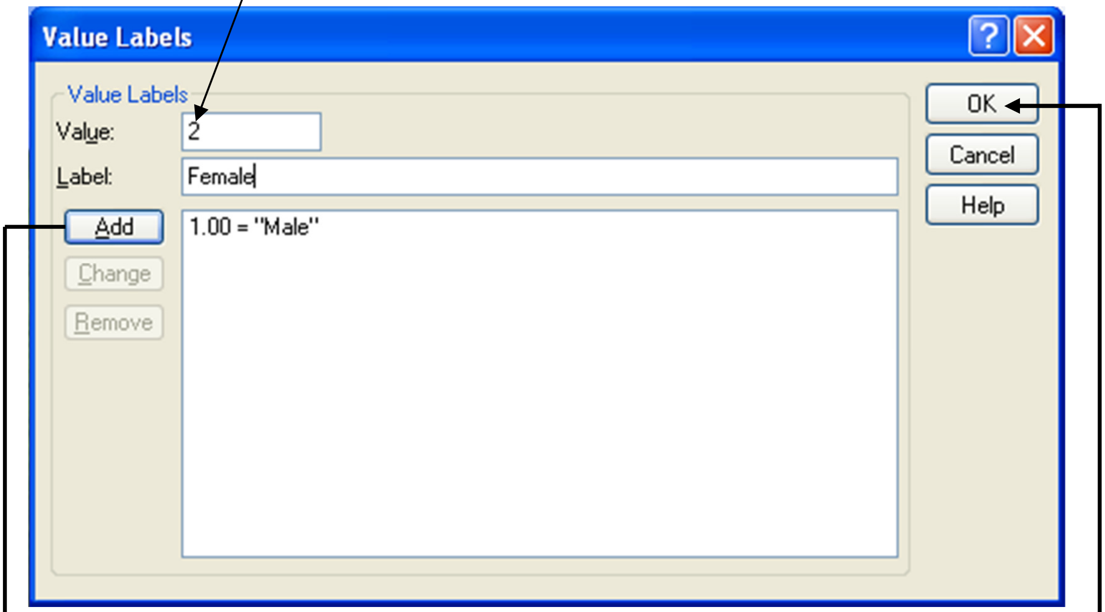


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing
1	Gender	Numeric	8	2	Gender of the Patient	None	None
2							
3							

Click on this button, this will call up the value labels dialogue box

Enter a coding value into the Value box, and then enter the label for this value into the Label box



Value: 2

Label: Female

1.00 = "Male"

Click on the Add button to add this value label for this variable. Repeat this step to add additional values and labels.

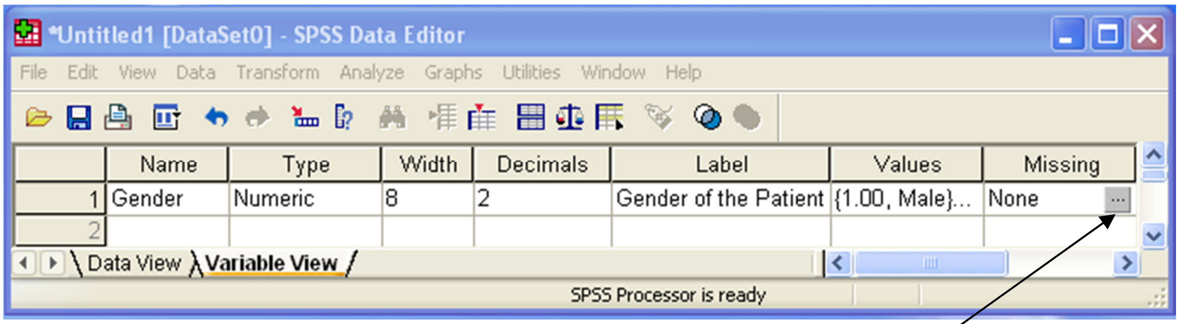
When you have added all the value labels for the variable, click on the OK button to close this dialogue box and return to the Variable View Table.

TIP: Value labels can a great help when interpreting the SPSS output. It would not be appropriate to add value labels for quantitative variables like Age, Cholesterol etc. In fact for presentation purposes, the value labels should be assigned to nominal and ordinal scale variables.

MISSING VALUES

Sometimes the datasheet may not have complete information i.e. it contains some missing observations, for example a patient might decline to tell his/her religion, missing laboratory values, lost to follow-up etc. When we have missing values we have to tell SPSS that we do not have valid data for this patient on this variable and we do this by choosing a value that cannot normally occur for the concerned variable. In the religion example above, we might choose to code religion as 9 if the patient does not state their religion.

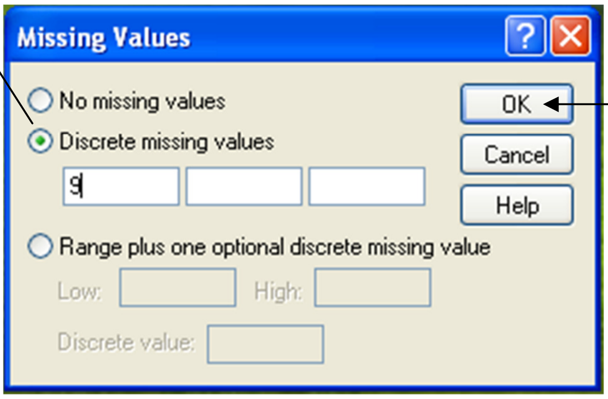
Before you specify any missing values, the cell in the Missing column of the Variable View table will contain the word **None** (see below). To specify a missing value click in the Missing column of the Variable View table, a button will appear at the right end of the cell. Click on the button to call up the Missing Value dialogue box.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing
1	Gender	Numeric	8	2	Gender of the Patient	{1.00, Male}...	None ...
2							

Click on the button in the Missing cell to call up the Missing Value dialogue box.

To include up to three different missing values click on the circle then enter the missing value(s) in the boxes.

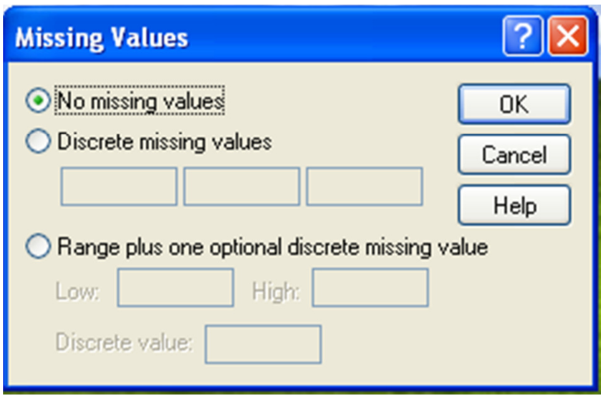


○ No missing values
● Discrete missing values
9
○ Range plus one optional discrete missing value
Low: High:
Discrete value:

Click on the OK button to close the dialogue box and return to the Variable View Table.

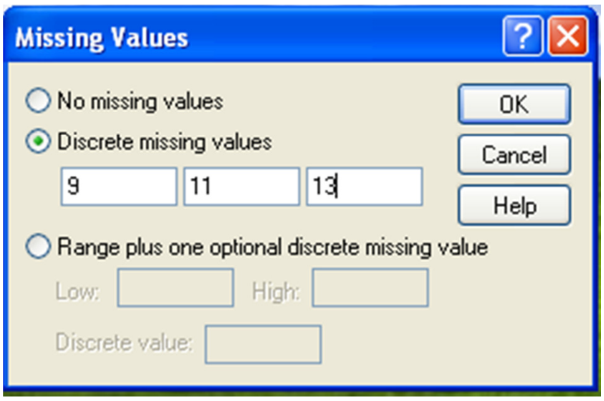
SPSS allows you to specify the missing values in several ways:

1. **NO MISSING VALUES:** This is the default setting for this dialogue box. If this option is selected then SPSS will treat all the observations for this variable as valid.



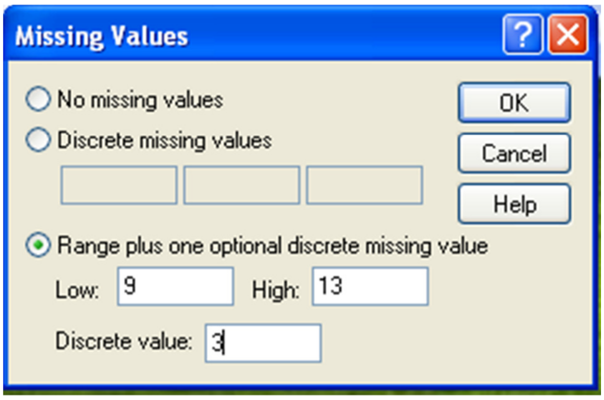
● No missing values
○ Discrete missing values
○ Range plus one optional discrete missing value
Low: High:
Discrete value:

2. **DISCRETE MISSING VALUES:** This option allows you to enter three discrete values. For example 9, 11, 13 could be set as missing values by selecting this option and entering the values in the three boxes. If you have only one missing value then enter it into the first of the three boxes.



○ No missing values
● Discrete missing values
9 | 11 | 13
○ Range plus one optional discrete missing value
Low: High:
Discrete value:

3. **RANGE PLUS ONE OPTIONAL DISCRETE MISSING VALUE:** This option allows you to indicate that a range of values if being used as missing values. For example selecting this option and entering the values 9 and 13 in the Low and High value boxes would instruct SPSS to treat the values 9, 10, 11, 12, 13 as missing values. If in addition to this range of values the value 3 were typed into the Discrete value box, then SPSS would treat the values 9, 10, 11, 12, 13 and 3 as missing. In practice we rarely need more than one missing value for a variable.



Missing Values

○ No missing values

○ Discrete missing values

● Range plus one optional discrete missing value

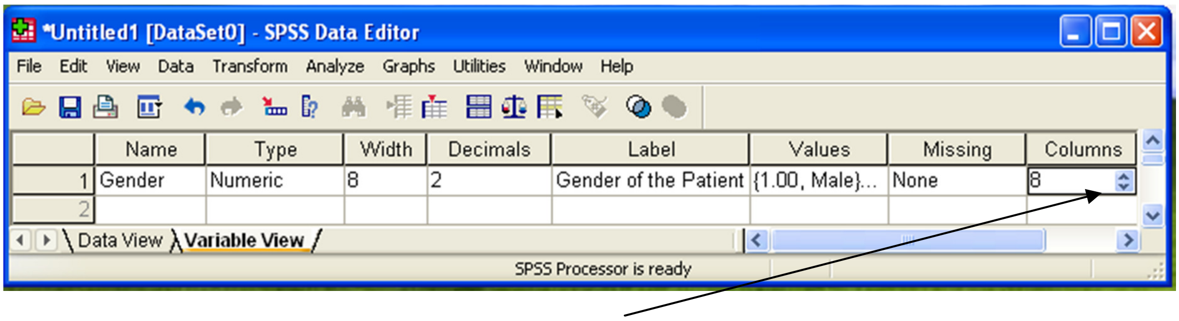
Low: 9 High: 13

Discrete value: 3

TIP: Missing value dialogue box does not allow you to label the missing values once you have entered them. However, you can label them in Value Labels dialogue box for example, 9 = unanswered, 10 = illegible etc

COLUMN FORMAT

The next column of the Variable View table is labeled Columns. This entry in the table is used to specify the width of the column that the variable occupies in the Data View table of the Data Editor window. You can leave this value at its default setting i.e. 8, unless you want to change the appearance of the Data View table. You may for example want to fit more columns onto the screen in order to see more variables without having to scroll. In this case you can reduce the width of each column.



*Untitled1 [DataSet0] - SPSS Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns
1	Gender	Numeric	8	2	Gender of the Patient	{1.00, Male}...	None	8
2								

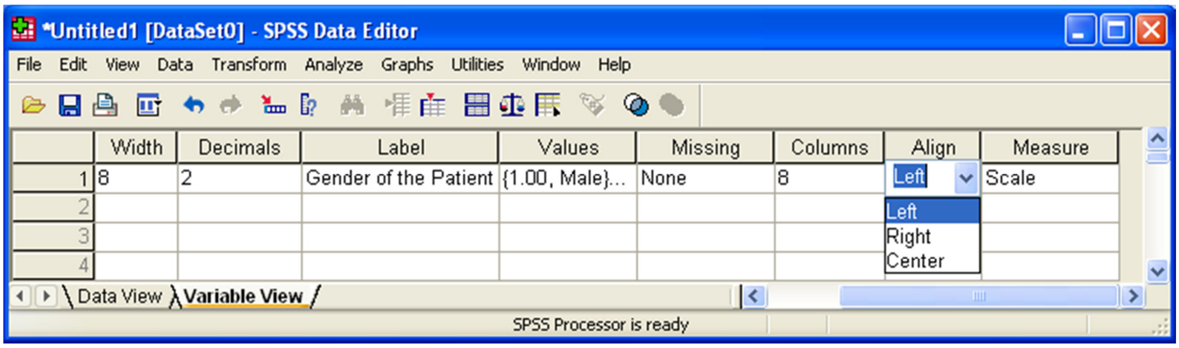
SPSS Processor is ready

Click on the cell and then use the up and down buttons to adjust the value.
You can look at the effect of the change you have made by switching to the Data View.

COLUMN ALIGNMENT

This column of the Variable View is labeled Align which allows you to specify the alignment of the text within the cells of the Data View of the Data Editor Window. This setting has no effect on the SPSS output, it is used just to change the appearance of the Data View table. The default setting is right alignment, in which the values are aligned on the right side. In left alignment the values are aligned to the left side. In centre alignment the values are centered in the cell.

To change the Column Alignment, click in the Align cell and then click on the menu button that will appear in the cell and select the required alignment from the drop down list (as shown below).



	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	8	2	Gender of the Patient	{1.00, Male}...	None	8	Left	Scale
2							Left Right Center	
3								
4								

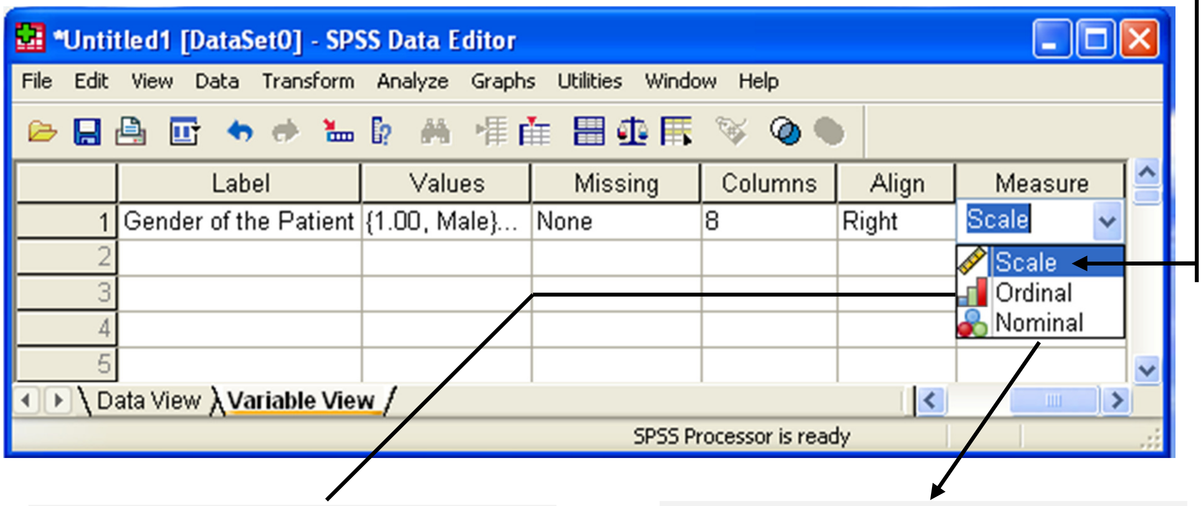
TIP: If you alter either the column width or alignment, remember you should switch to the Data View to see the effect of your changes.

MEASUREMENT

The final column of the Variable View is the labeled Measure. This column is used to specify the measurement scale for the variable(s). In SPSS we have 03 options, Nominal, Ordinal and Scale. SPSS does not distinguish between interval and ratio data and uses the term scale to cover both of them.

NOTE: You must specify appropriate measurement scale to the variables otherwise you can face difficulty in interpreting and editing the charts.

Select the Scale option for variables measured on either an Interval or Ratio scale

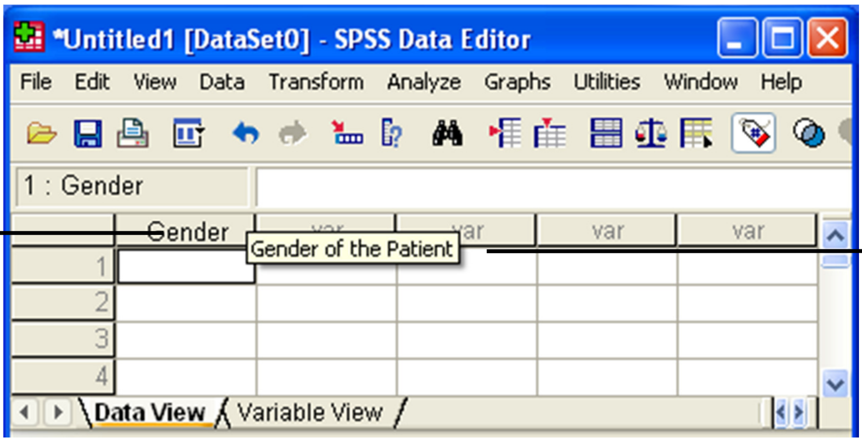

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View
SPSS Processor is ready

	Label	Values	Missing	Columns	Align	Measure
1	Gender of the Patient	{1.00, Male}...	None	8	Right	Scale
2						
3						
4						
5						

Select the Ordinal option for variables measured on Ordinal scale (e.g. Disease Stage)

Select the Nominal option for Nominal variables (e.g. Gender, Occupation, Religion etc)

Once you have completed the definition of your first variable, switch to the Data View window. You will see the name of your new variable at the top of the first column.

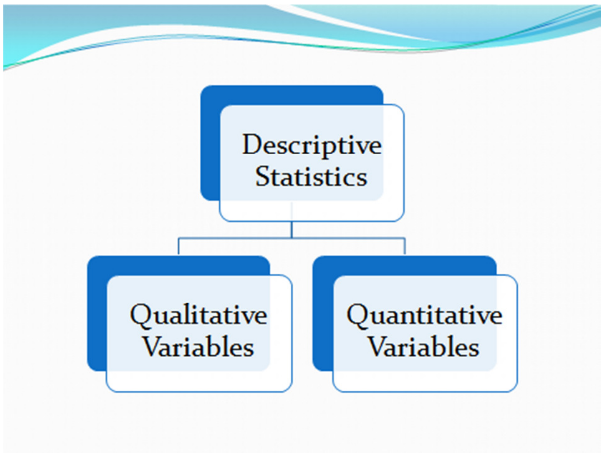

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Data View
Variable View

1 : Gender

	Gender	var	var	var	var
1					
2					
3					
4					

The new variable name appears at the top of the column. This column is now ready to accept data.

Note: If you move the mouse pointer over the variable name, a pop-up will appear, displaying the variable label (assigned in the Variable View under the heading of Label).



Descriptive Statistics

Qualitative Variables

Quantitative Variables

Descriptive statistics gives informtion about frequencies & percentages, central tendencies (mean, median and mode etc), dispersion (range, inter-quartile range, standard deviation, standard error of mean etc) and graphs. Its is very important to note that we cannot test any hypothesis in descriptive part. Moreover, the analysis for qualitative and quantitative nature data is also different.

ANALYSIS FOR QUALITATIVE NATURE DATA:

I. Frequencies and Percentages

II. Graphs (Bar-Chart, Pie-Chart etc)

Note: We cannot calculate any average such as mean, median or mode etc for qualitative variables because it does not give meaningful interpretation.

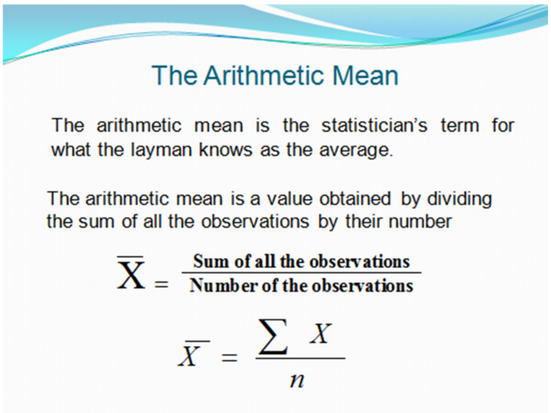
ANALYSIS FOR QUANTITATIVE NATURE DATA:

1. Center of data (Mean, Median, Mode etc)
2. Dispersion in data (S.E.M, S.D, IQR, Range etc)
3. Shape of the Distribution
   - ❖ Skewness
   - ❖ Kurtosis
4. Graphical Presentation
   - ❖ Histogram
   - ❖ Box – plot etc

DESCRIPTIVE STATISTICS

CENTRAL TENDENCIES:

ARITHMETIC MEAN:



The Arithmetic Mean

The arithmetic mean is the statistician's term for what the layman knows as the average.

The arithmetic mean is a value obtained by dividing the sum of all the observations by their number

$$\overline{X} = \frac{\text{Sum of all the observations}}{\text{Number of the observations}}$$

$$\overline{X} = \frac{\sum X}{n}$$

Example:

The ages of 30 patients admitted to a certain hospital during a particular week were as follows:

48, 31, 54, 37, 18, 64, 61, 43, 40, 71, 51, 12, 52, 65, 53, 42, 39, 62, 74, 48, 29, 67, 30, 49, 68, 35, 57, 26, 27, 58.

Calculate the mean age of the patients admitted?

Arithmetic Mean = 43.27 years

One big limitation to the use of Arithmetic Mean:

Every value in a data-set is included in the calculation of the mean, whether the value be high or low. Where there are a few very high or very low values in the series, their effect can be to *drag* the arithmetic mean towards them. This may make the mean unrepresentative.

PROPERTIES OF ARITHMETIC MEAN:

- ❖ Best understood average in statistics.
- ❖ Relatively easy to calculate.
- ❖ Takes into account every value in the series.

MEDIAN:

The median represents the middle value of the ordered sample data.

OR

A number such that 50% of the measurements are below it and 50% of the measurements are above it.

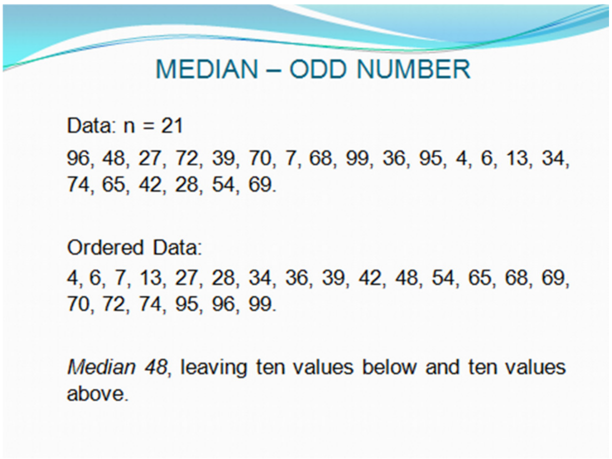
Median for Odd Number Data

When the sample size is odd, the median is the middle value.

Median for Even Number Data

When the sample size is even, the median is the average of the two middle values.

Example:



MEDIAN – ODD NUMBER

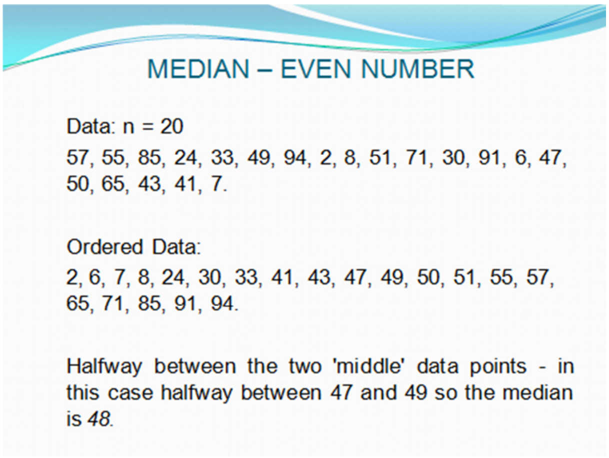
Data: n = 21

96, 48, 27, 72, 39, 70, 7, 68, 99, 36, 95, 4, 6, 13, 34, 74, 65, 42, 28, 54, 69.

Ordered Data:

4, 6, 7, 13, 27, 28, 34, 36, 39, 42, 48, 54, 65, 68, 69, 70, 72, 74, 95, 96, 99.

Median 48, leaving ten values below and ten values above.



MEDIAN – EVEN NUMBER

Data: n = 20

57, 55, 85, 24, 33, 49, 94, 2, 8, 51, 71, 30, 91, 6, 47, 50, 65, 43, 41, 7.

Ordered Data:

2, 6, 7, 8, 24, 30, 33, 41, 43, 47, 49, 50, 51, 55, 57, 65, 71, 85, 91, 94.

Halfway between the two 'middle' data points - in this case halfway between 47 and 49 so the median is *48*.

PROPERTIES OF MEDIAN:

- ❖ It is easily calculated and understood.
- ❖ It can be computed even for ordinal scale data.
- ❖ It is not affected by extreme values.
- ❖ In a highly skewed distribution median is an appropriate average to use.
- ❖ It necessitates the arrangement of data into an array which can be tedious and time consuming especially if the dataset is large.

MODE:

The mode is the value that occurs most frequently. It is least used of the three measures of central tendency.

EXAMPLE:

Date	High Temperature	
2-Jan	59	
3-Jan	60	
4-Jan	43	
5-Jan	42	
6-Jan	35	
7-Jan	32	**<===Mode**
8-Jan	32	**<===Mode**
9-Jan	46	
10-Jan	41	
11-Jan	52	

Mode = 32

NOTE:

- ➢ It is possible for a set of data values to have more than one mode.
- ➢ If there are two data values that occur most frequently, we say that the set of data values is **bimodal**.

PROPERTIES OF MODE:

- ❖ It is simply defined and easily calculated, in some cases it is extremely easy to calculate mode.
- ❖ It is not affected by outliers (small or large values).
- ❖ It can be determined for both qualitative and quantitative data.
- ❖ It is often in-determinable.
- ❖ It is not based on all values.

QUARTILES:

Divide the data into four equal parts.

Q1 = ¼ (n+1) th obs 25% obs

Q2 =2/4 (n+1) th obs. 50% obs

Q3 = ¾ (n+1) th obs. 75% obs

Example: If in a certain data the third quartile (Q3) is 40, it means that 25% of the measurements in the data are above 40 and otherwise 75% of the measurements are below 40.

DECILES:

Divide the data into 10 equal parts.

D1 = 1/10 (n+1) th obs. 10% observations

D2 = 2/10 (n+1) th obs. 20% observations

D3 = 3/10 (n+1) th obs. 30% observations

.

.

.

D9 = 9/10 (n+1) th obs. 90% observations

Example: if in a certain data the 4th Decile (D4) is 160, it means that 60% of the measurements in the data are above 160 and otherwise 40% of the measurements are below 160.

PERCENTILES:

Divide the data into 100 equal parts.

P1 = 1/100 (n+1) th obs.

P2 = 2/100 (n+1) th obs.

P3 = 3/100 (n+1) th obs.

.

.

.

P99 = 99/100 (n+1) th obs.

Example: If in a certain data the 85th percentile is 340 means that 15% of the measurements in the data are above 340. It also means that 85% of the measurements are below 340.

MEASURES OF DISPERSION:

Measures of dispersion are descriptive statistics that describe how similar a set of scores are to each other.

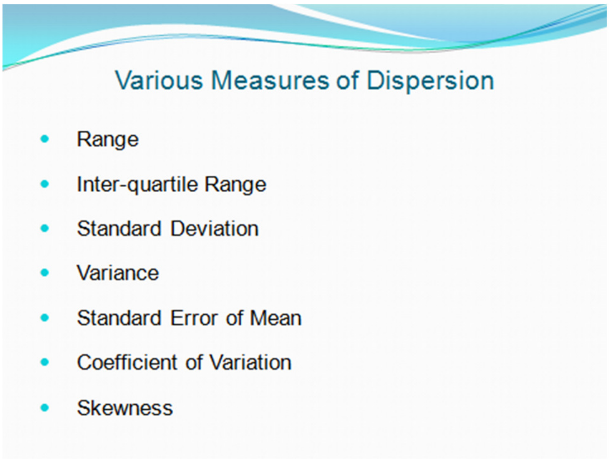
OR

Measures of dispersion indicate the extent to which the observed values are "spread out" around that center of data.

Note:

- The more similar the scores are to each other, the lower the measure of dispersion will be.
- The less similar the scores are to each other, the higher the measure of dispersion will be.
- In general, the more spread out a distribution is, the larger the measure of dispersion will be.

VARIOUS MEASURES OF DISPERSION:



Various Measures of Dispersion

- Range
- Inter-quartile Range
- Standard Deviation
- Variance
- Standard Error of Mean
- Coefficient of Variation
- Skewness

RANGE:

The range is a measure of the spread or the dispersion of the observations. It is the difference between the largest and the smallest observed value of some characteristic.

Range = XL – XS

Example:

What is the range of this Dataset:

65,73,89,56,73,52,47

Smallest Value (XS) = 47

Largest Value (XL) = 89

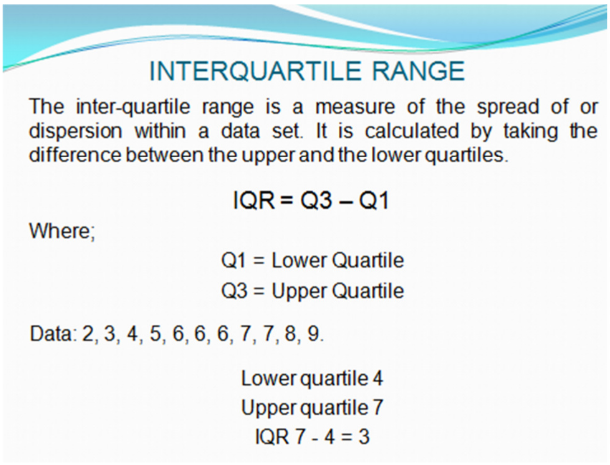
The range is 89-47 = 42.

USE OF RANGE AS MEASURE OF DISPERSION:

- The range is used when you have discrete, continuous or ordinal data.
- A great deal of information is ignored when computing the range since only the largest and the smallest data values are considered; the remaining data are ignored.
- The range value of a data set is greatly influenced by the presence of just one unusually large or small value in the sample (outlier).
- Two very different sets of data can have the same range:

  1 1 1 1 9 vs. 1 3 5 7 9

INTER-QUARTILE RANGE:



INTERQUARTILE RANGE

The inter-quartile range is a measure of the spread of or dispersion within a data set. It is calculated by taking the difference between the upper and the lower quartiles.

IQR = Q3 – Q1

Where;

Q1 = Lower Quartile

Q3 = Upper Quartile

Data: 2, 3, 4, 5, 6, 6, 6, 7, 7, 8, 9.

Lower quartile 4
Upper quartile 7
IQR 7 - 4 = 3

STANDARD DEVIATION:

The Standard Deviation is a number that measures how far away each number in a set of data is from their mean.

- If the Standard Deviation is *large*, it means the numbers are spread out from their mean.
- If the Standard Deviation is *small*, it means the numbers are close to their mean.

EXAMPLE:

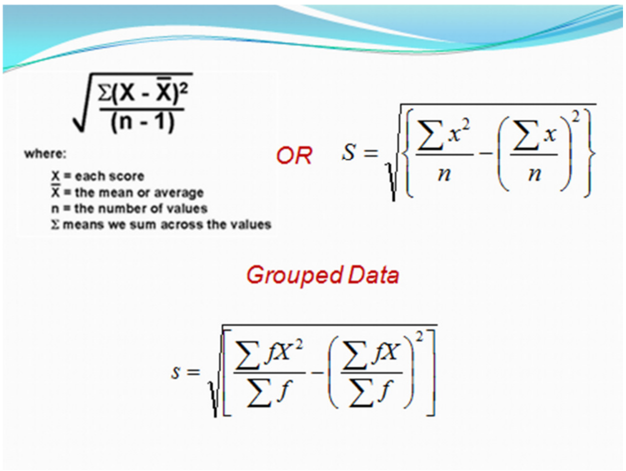
Two classes took a recent biostatistics quiz. There were 10 students in each class, and each class had an average score of 81.5.

Since the averages are the same, can we assume that the students in both classes all did pretty much the same on the exam?

The answer is… No.

The mean does not tell us anything about the variation in the grades.

FORMULA FOR CALCULATING SD:



$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

where:

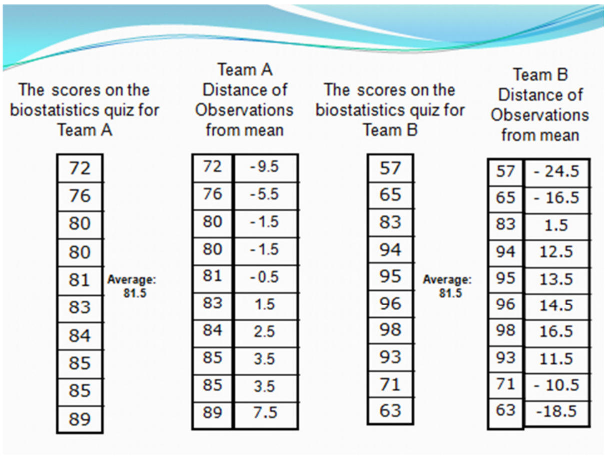
X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

OR $S = \sqrt{\left\{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2\right\}}$

Grouped Data

$$s = \sqrt{\left[\frac{\sum fX^2}{\sum f} - \left(\frac{\sum fX}{\sum f}\right)^2\right]}$$

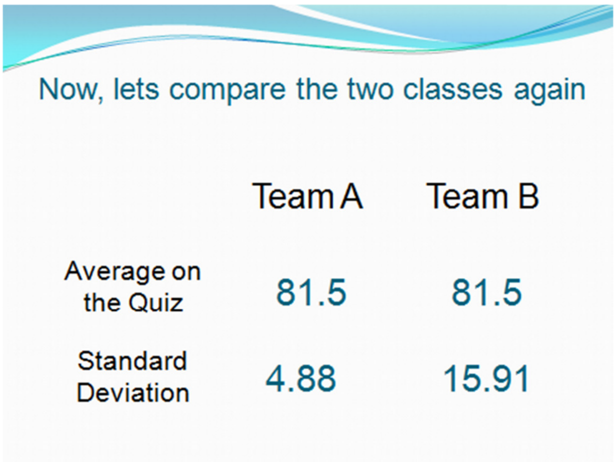
CALCULATION:



The scores on the biostatistics quiz for Team A	Team A Distance of Observations from mean		The scores on the biostatistics quiz for Team B	Team B Distance of Observations from mean	
72	72	-9.5	57	57	- 24.5
76	76	-5.5	65	65	- 16.5
80	80	-1.5	83	83	1.5
80	80	-1.5	94	94	12.5
81	81	-0.5	95	95	13.5
83	83	1.5	96	96	14.5
84	84	2.5	98	98	16.5
85	85	3.5	93	93	11.5
85	85	3.5	71	71	- 10.5
89	89	7.5	63	63	-18.5

Average: 81.5

Average: 81.5



Now, lets compare the two classes again

	Team A	Team B
Average on the Quiz	81.5	81.5
Standard Deviation	4.88	15.91

VARIANCE:

Variance is defined as the average of the square deviations:

$$s^2 = \frac{\sum\left(X - \overline{X}\right)^2}{N-1}$$

- ❖ The larger the variance is, the more the scores deviate, on average, away from the mean.
- ❖ The smaller the variance is, the less the scores deviate, on average, from the mean.

LIMITATIONS OF VARIANCE:

The calculation of variance involves squaring of the units in which the variables are measured.

Example: If people weights are measured in pounds, then the variance of the weights would be expressed in pounds^2 (or squared pounds)

Since squared units of measure are often awkward to deal with, the *square root of variance* is often used instead to interpret the results in the same units which is the standard deviation.

$$SD = \sqrt{Variance}$$

STANDARD ERROR OF MEAN:

The precision of the mean of a sample of data, as an estimate of some unknown or “true” value of the mean of the population, can be described using the standard error of the mean.

In other works, standard error of mean tells us how well the sample mean represents the population mean.

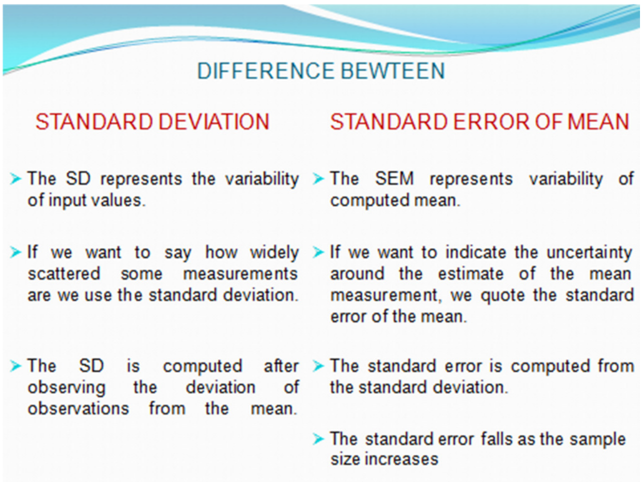
$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Where;

S is the standard deviation

N is the sample size

DIFFERENCE BEWTEEN SD & SEM:



DIFFERENCE BEWTEEN

STANDARD DEVIATION	STANDARD ERROR OF MEAN
➢ The SD represents the variability of input values.	➢ The SEM represents variability of computed mean.
➢ If we want to say how widely scattered some measurements are we use the standard deviation.	➢ If we want to indicate the uncertainty around the estimate of the mean measurement, we quote the standard error of the mean.
➢ The SD is computed after observing the deviation of observations from the mean.	➢ The standard error is computed from the standard deviation.
	➢ The standard error falls as the sample size increases

COEFFICIENT OF VARIATION:

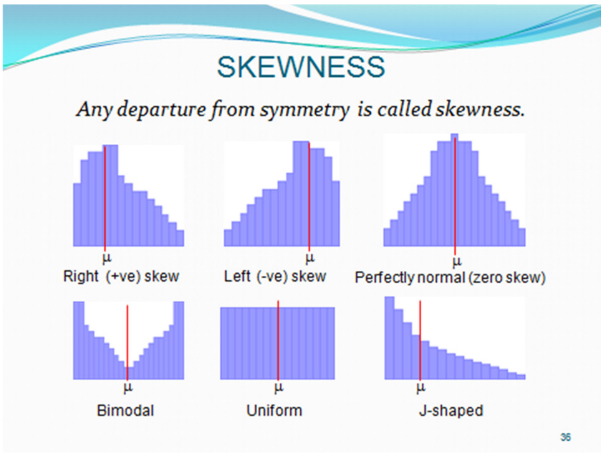
Coefficient of variation is used to compare the variation in two or more datasets that are measured in different units, for example, one variable may be measured in hours and other in kilogram etc.

$$CV = \frac{S}{\overline{X}} \times 100\%$$

Where:

S is the standard deviation
X bar is the mean

SKEWNESS:



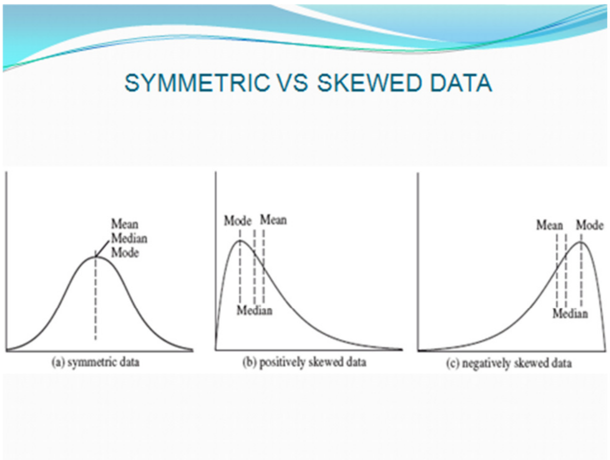
SKEWNESS

Any departure from symmetry is called skewness.

CRITERIA FOR DECLARING SKEWNESS:

- ❖ If s < 0, then the distribution has a negative skew.
- ❖ If s > 0 then the distribution has a positive skew.
- ❖ If s = 0 then the distribution is symmetrical.
- ❖ The more different s is from 0, the greater the skew in the distribution

SYMMETRICAL VS SKEWED DATA:



SYMMETRIC VS SKEWED DATA

GRAPHICAL PRESENTATION:

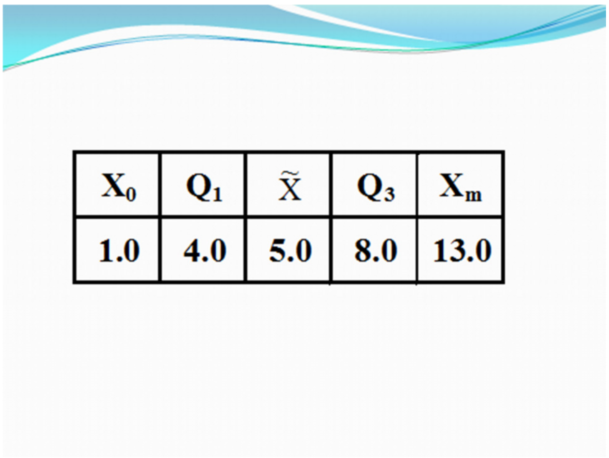
- ❖ Bar Chart – Nominal or ordinal scale data.
- ❖ Pie Chart – Nominal or ordinal scale data.
- ❖ Histogram – Ratio scale data.
- ❖ Box-plot or 5 Number Summary – Ratio scale data

CONSTRUCTION OF BOX-PLOT OR 5 NUMBER SUMMARY:

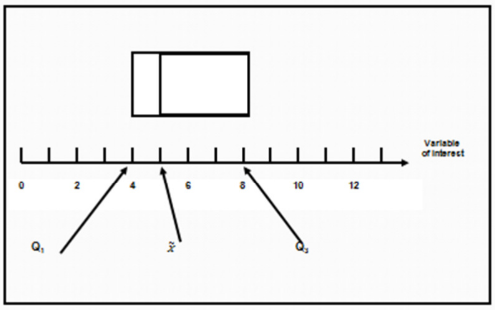
A five-number summary consists of

- ❖ X0 (minimum value)
- ❖ Q1 (Lower quartile)
- ❖ Q2 (Median)
- ❖ Q3 (Upper quartile)
- ❖ Xm (Maximum Value)

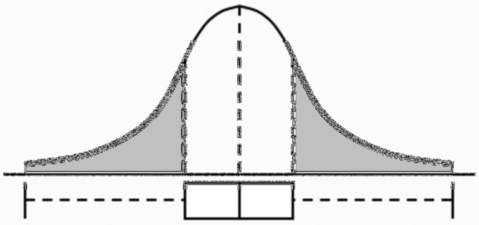
It provides us quite a good idea about the shape of the distribution.



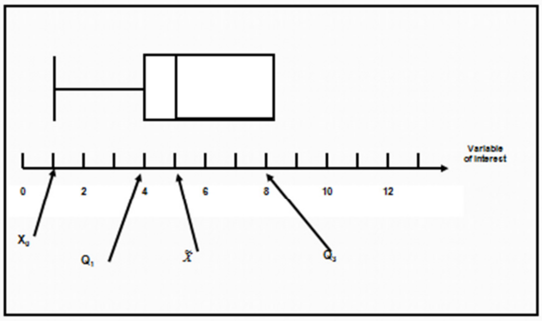
X_0	Q_1	$\tilde{X}$	Q_3	X_m
1.0	4.0	5.0	8.0	13.0



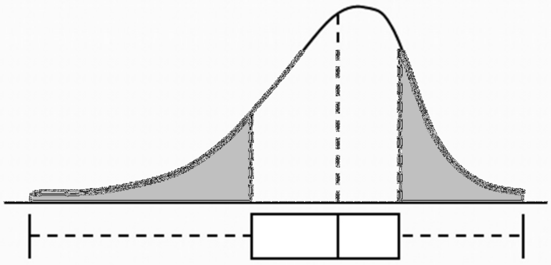
SYMMETRICAL DISTRIBUTION



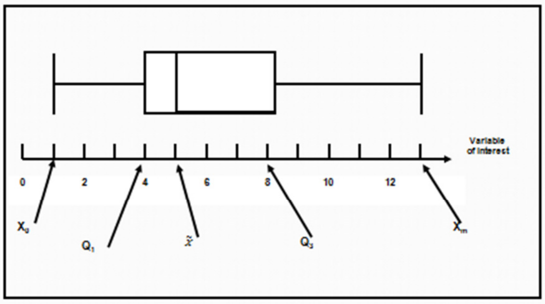
(a) Bell-shaped distribution



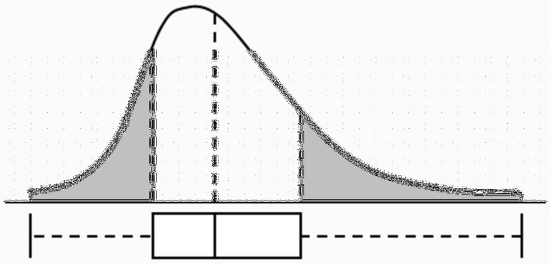
LEFT SKEWED DISTRIBUTION



Left-skewed distribution



RIGHT SKEWED DISTRIBUTION



Right-skewed distribution

Workshop on
“Biological Data Analysis
with Hands on Training on Statistical
Softwares”
March 22 – 25, 2011
LAHORE

Session I – Introduction to Biostatistics and Data Entry in Statistical Package

Session II – Descriptive Statistics

Session III : Hands on Training

Facilitator : Mr. Waqas Sami and Mr. Waqas Latif

Time: 02:00 p.m. – 03:30 p.m.

Session III was based on hands on training on computers in entering data in statistical package for social sciences 18.0 and calculating descriptive statistics. The participants were given a datasheet containing 10 variables with respective coding scheme:

Gender: 1 = Male, 2 = Female

Marital Status: 1 = Married, 2 = Unmarried

Education: 1 = Graduation, 2 = MBBS, 3 = M.Phil

College: 1 = King Edward Medical University, 2 = Allama Iqbal Medical College,

3 = Lahore Medical & Dental College, 4 = Lahore University

Job Nature: 1 = Permanent, 2 = Contract

Salary: 1 = Below 8000, 2 = 8000 – 15000, 3 = 15000 – 35000, 4 = Above 35000

Secondly, they were asked to:

1. Define Variables in the SPSS according to the nature of the variables.
2. Enter the data in the data view window.
3. Check for duplicate cases.
4. Make Pie Charts, Bar Charts and histogram.
5. Calculate measures of central tendencies and measures of dispersions.

DAY 2

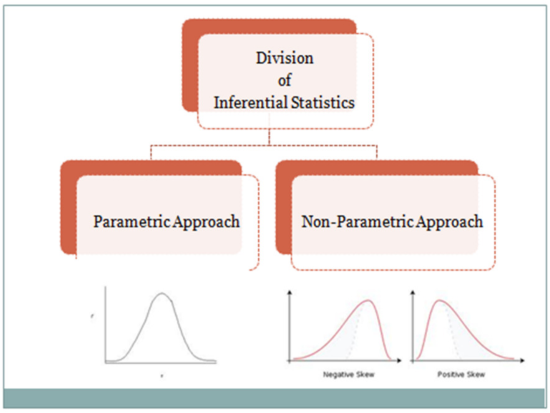
March 23, 2011

Session I : Parametric Approach Techniques

Resource Person: Mr. Waqas Sami

Time: 09:00 a.m. – 11:00 a.m.

Inferential statistics is divided into two approaches, parametric approach and non-parametric approach. When the assumption of normality of data is fulfilled we apply parametric approach tests and when the data is skewed (condition in which outliers are present) we apply non-parametric approach tests.



ASSUMPTIONS OF PARAMETRIC APPROACH TECHNIQUES:

- Randomization.
- Normality.
- Homogeneity of Variance (the population SD or Variance are equal or approximately equal).
- The data is measured on an interval or ratio scale.

HYPOTHESIS TESTING:

Hypothesis testing is a very important phase of statistical inference, it is a procedure which enables us to decide on the basis of information obtained from sample data whether to accept or reject a statement about population parameter. Such a statement or assumption which may or may not be true, is called statistical hypothesis, we accept the hypothesis when it is supported by sample data and reject it when the sample data fails to support it.

NULL HYPOTHESIS:

A null hypothesis is any hypothesis which is to be tested, it is generally denoted by the symbol Ho, the word *null* in the word null hypothesis implies that Ho is the hypothesis of no effect, a hypothesis should always be precise either we are testing means, observing associations or differences etc.

ALTERNATE HYPOTHESIS:

An alternative hypothesis is a hypothesis which we accept when the null hypothesis Ho is rejected, it is denoted by H1. A null hypothesis Ho is tested against an alternative hypothesis H1. This is also called the researchers hypothesis.

LEVEL OF SIGNIFICANCE:

Level of significance is the probability of rejecting null hypothesis Ho, it is denoted by α. The most frequently used values of α are 0.05, and 0.01 i.e. 5% and 1%, by 5% we mean that we are 95% confident in making correct decision and 5% is still chance of error.

TEST STATISTIC:

The formula which provides a basis for testing a null hypothesis is called *test statistic.*

Formula for testing One Sample t test

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

CALCULATIONS:

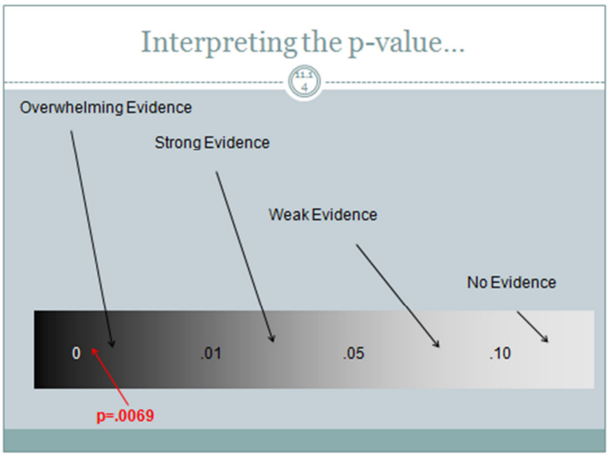
The required statistics are computed in this step and values are put in the formula to get the desired test statistic value. In classical hypothesis testing this part is very important but now due to availability of statistical packages we can easily have the calculations with other required quantities.

CRITICAL REGION:

In critical region we decide whether to accept or reject the null hypothesis, by comparing P-value and level of significance i.e. α.

A p-value is a Measure of strength of evidence the sample data provides against the null hypothesis.

If P-value is less than α, we reject our null hypothesis and if P-value is greater than α, we accept Ho or null hypothesis.



CONCLUSION:

Since P-value is less than α, we reject our Ho or null hypothesis and have sufficient evidence to conclude that, for example there is association between smoking and lung cancer.

Since P-value is greater than α, we accept our Ho or null hypothesis and have sufficient evidence to conclude that, for example there is no association between smoking and lung cancer.

PARAMETRIC APPROACH TESTS:

- One Sample t test.
- Two Independent Sample t test.
- Paired Sample t test.
- Analysis of Variance.
- Pearson Correlation.

ONE SAMPLE T TEST:

One Sample t-test is used to compare one group to a given standard on the basis of Arithmetic Average (Mean).

ASSUMPTIONS OF ONE SAMPLE T TEST:

- The data should be continuous.
- The data follows Normal distribution.
- Randomization.

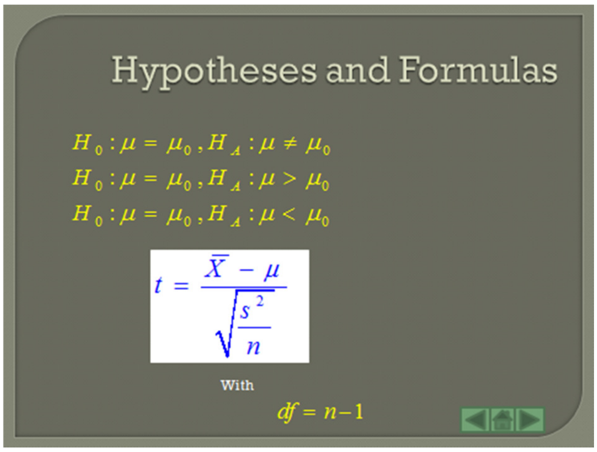
EXAMPLE:

For or the past five years, a zoologist has been involved in an extensive research-project regarding the animals of one particular species.

Based on his research-experience, the zoologist believes that the average height of the animals of this particular species is 66 centimeters. He selects a random sample of ten animals of this particular species, and, upon measuring their heights, the following data is obtained:

63, 63, 66, 67, 68, 69, 70, 70, 71, 71

In the light of these data, test the hypothesis that the mean height of the animals of this particular species is 66 centimeters.



Hypotheses and Formulas

$H_0 : \mu = \mu_0 , H_A : \mu \neq \mu_0$

$H_0 : \mu = \mu_0 , H_A : \mu > \mu_0$

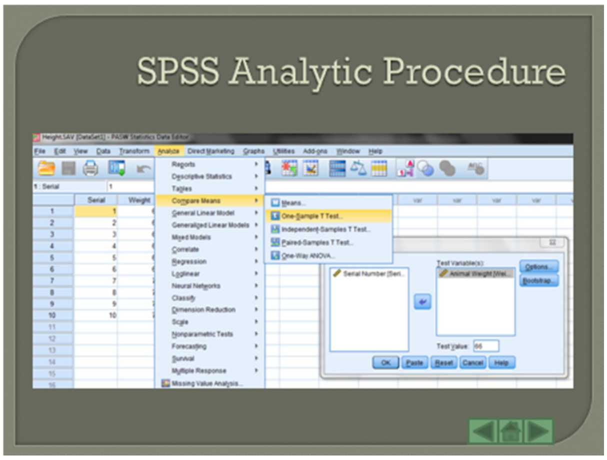
$H_0 : \mu = \mu_0 , H_A : \mu < \mu_0$

$$t = \frac{\overline{X} - \mu}{\sqrt{\frac{s^2}{n}}}$$

With

$df = n-1$

HOW TO PERFORM ONE SAMPLE T TEST IN SPSS:

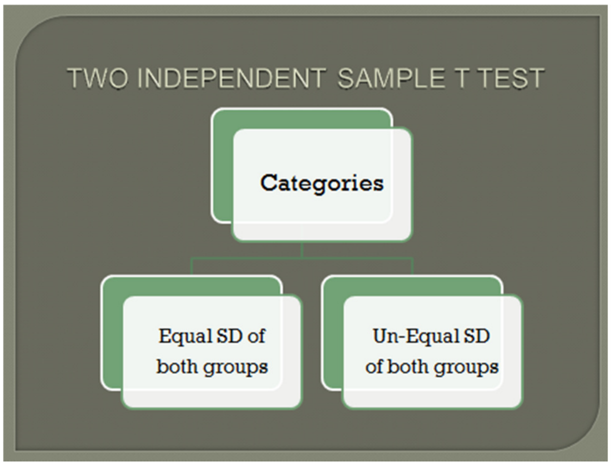


SPSS Analytic Procedure

Height.SAV [DataSet1] - PASW Statistics Data Editor

Test Variable(s): Animal Weight [Wei...

Test Value: 66



TWO INDEPENDENT SAMPLE T TEST

Categories

Equal SD of both groups

Un-Equal SD of both groups

TWO – INDEPENDENT SAMPLES T-TEST (EQUAL SDs):

Independent sample t – test is used to compare two groups on the basis of their averages.

ASSUMPTIONS OF TWO – INDEPENDENT SAMPLE T TEST:

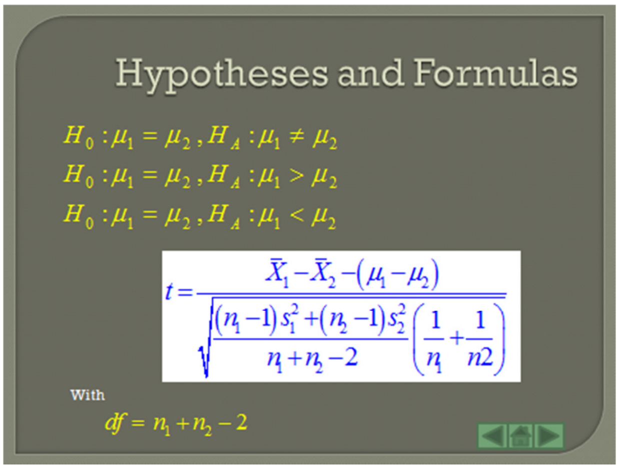
- The data are continuous.
- The data follows Normal distribution.
- The variances of the two populations are equal.
- The two samples are independent.
- Both samples are simple random samples from their respective populations.

EXAMPLE:

We are interested in determining which study supplement is more effective in reducing test anxiety. We have heard rumors that some people take Xanax when studying and some people drink a variety of caffeine laden substances while studying.

We decide that we will compare Xanax to Mountain Dew (as representative of the caffeine laden substances). We will use a biostatistics mid-term exam as the test criteria. 10 participants were randomly assigned to the Xanax group and 10 to the Mountain Dew group.

Test if there is a significant difference between the groups in their average BP during the exam.



Hypotheses and Formulas

$H_0 : \mu_1 = \mu_2 , H_A : \mu_1 \neq \mu_2$

$H_0 : \mu_1 = \mu_2 , H_A : \mu_1 > \mu_2$

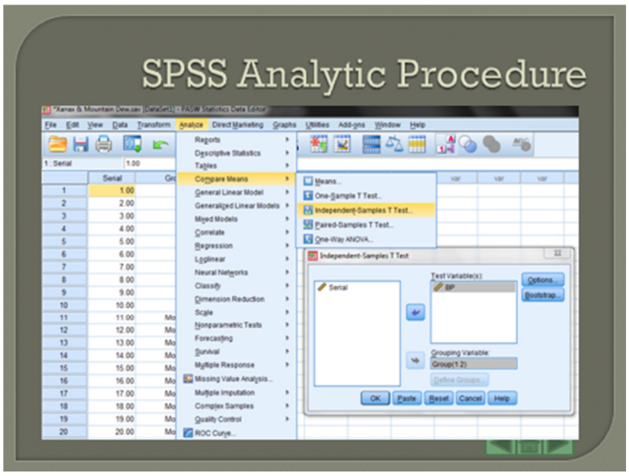
$H_0 : \mu_1 = \mu_2 , H_A : \mu_1 < \mu_2$

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n2}\right)}}$$

With

$df = n_1 + n_2 - 2$

TWO INDEPENDENT SAMPLE T TEST IN SPSS:



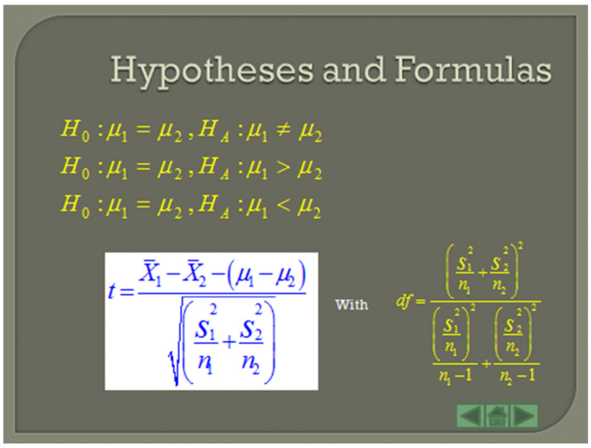
SPSS Analytic Procedure

TWO – INDEPENDENT SAMPLES T-TEST (UN-EQUAL SDs):

Independent Samples t-test is use to compare two independent groups on the basis of average. This test does not require homogeneity of the variances.

EXAMPLE:

The cell diameters (µm) of 40 Lymphocytes and 50 tumor cells obtained from biopsies of tissue from patients with melanoma. Can we conclude that, on the average, lymphocytes and tumor cells differ in size? At 5% level of significance.



Hypotheses and Formulas

$H_0: \mu_1 = \mu_2, H_A: \mu_1 \neq \mu_2$

$H_0: \mu_1 = \mu_2, H_A: \mu_1 > \mu_2$

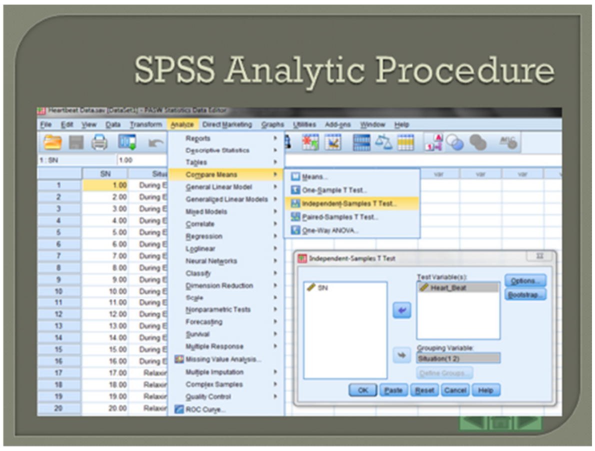
$H_0: \mu_1 = \mu_2, H_A: \mu_1 < \mu_2$

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

With

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

TWO INDEPENDENT SAMPLE T TEST IN SPSS:



SPSS Analytic Procedure

Test Variable(s): Heart_Beat

Grouping Variable: Situation(1 2)

PAIRED SAMPLE T TEST:

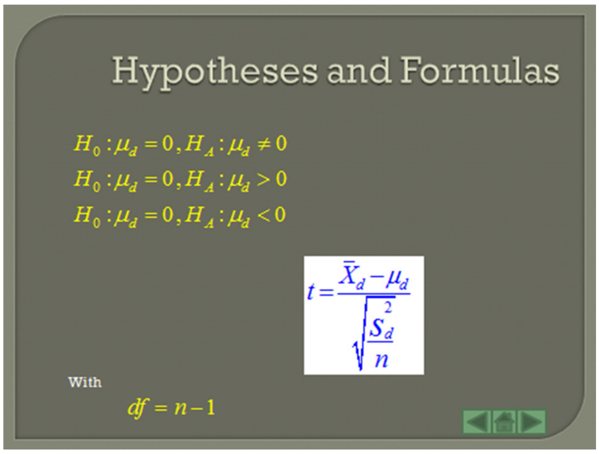
In the paired case, we take two measurements on same individual at different times, or we have one measurement on each individual of a pair.

ASSUMPTIONS OF PAIRED SAMPLE T TEST:

- ❖ The data are continuous.
- ❖ The data, i.e., the differences for the matched-pairs, follow a Normal distribution.
- ❖ The sample of pairs is a simple random sample from its population.

EXAMPLE:

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, weather fasting affects the test performance at 5% level of significance.



Hypotheses and Formulas

$H_0: \mu_d = 0, H_A: \mu_d \neq 0$

$H_0: \mu_d = 0, H_A: \mu_d > 0$

$H_0: \mu_d = 0, H_A: \mu_d < 0$

$$t = \frac{\bar{X}_d - \mu_d}{\sqrt{\frac{S_d^2}{n}}}$$

With

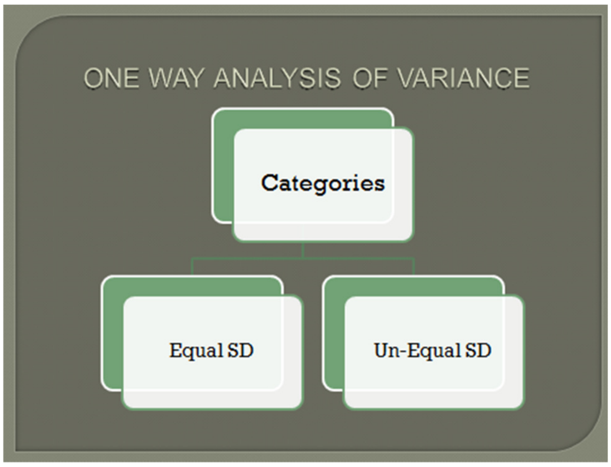
$df = n - 1$

ONE WAY ANALYSIS OF VARIANCE:

One Way Analysis of Variance is used to compare the means of more than two groups on the basis of their averages.

ASSUMPTIONS OF ONE WAY ANALYSIS OF VARIANCE:

- ❖ The data are continuous.
- ❖ The data follow the Normal distribution, each group is normally distributed.
- ❖ The variances of the populations are equal.
- ❖ The groups are independent.
- ❖ Each group is a simple random sample from its population.



ONE WAY ANALYSIS OF VARIANCE

Categories

Equal SD

Un-Equal SD

EXAMPLE: EQUAL SD'S

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, the following groups were selected:

Group A:
Healthy Subjects

Group B:
Patients with benign breast cancer

Group C:
Patients with benign primary cancer

Group D:
Patients with recurrent meta-static breast cancer

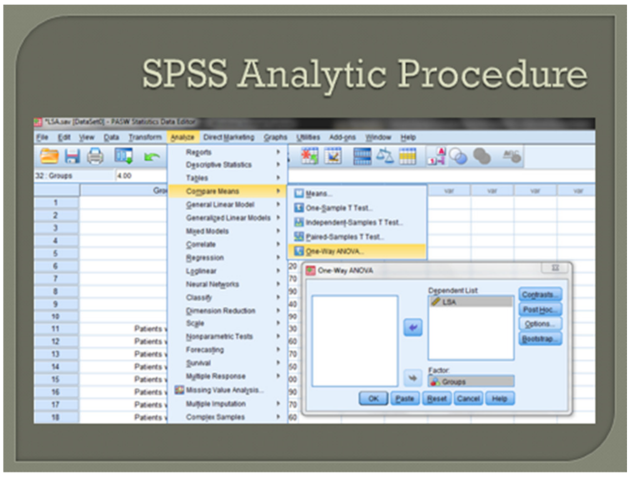
Test at 5% level of significance that is there any difference in the LSA values of 4 groups.

HYPOTHESIS AND FORMULA:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

$$H_A : Atleast\ one\ pair\ is\ significantly\ diffrent$$

HOW TO PERFORM ANOVA IN SPSS:



SPSS Analytic Procedure

One-Way ANOVA

Dependent List: LSA

Factor: Groups

POST – HOC TESTS:

The significant p-value of Analysis of Variance tells you that at least one of the mean differs, now to see which group men differs we apply post hoc tests and select the suitable under following conditions.

Different Situations

Equal Variances and equal sample size then use:

TUKEYS TEST

Equal Variances but slightly different sample sizes then use:

GABRIEL's TEST

Equal Variances but very different sample sizes then use:

HOCHBERG's GT2 TEST

Unequal Variances and sample sizes (Whether slightly different or very different)

GAMES – HOWELL TEST

ONE WAY ANALYSIS OF VARIANCE FOR UNEQUAL SD'S:

Welch ANOVA is used to compare means of more than two groups on the basis of averages. This test doest not require the data to follow homogeneity of variances.

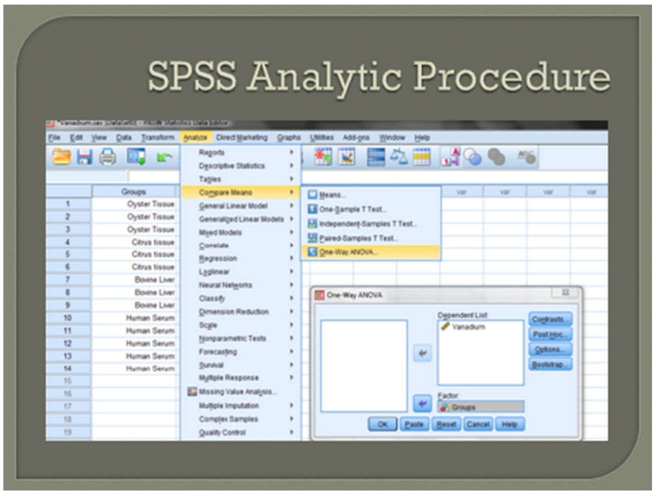
EXAMPLE:

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 quantities of vanadium in dried samples of oyster tissue, citrus leaves, bovine liver and human serum were observed. Test whether the distribution of vanadium concentrations for the four biological materials differ in locations at 5% level of significance.

HYPOTHESIS AND FORMULA:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

$$H_A : Atleast\ one\ pair\ is\ significantly\ diffrent$$



SPSS Analytic Procedure

	Groups
1	Oyster Tissue
2	Oyster Tissue
3	Oyster Tissue
4	Citrus tissue
5	Citrus tissue
6	Citrus tissue
7	Bovine Liver
8	Bovine Liver
9	Bovine Liver
10	Human Serum
11	Human Serum
12	Human Serum
13	Human Serum
14	Human Serum

One-Way ANOVA

Dependent List: Vanadium

Factor: Groups

CORRELATION:

A correlation is a relationship between two variables.

OR

A correlation indicates the extent to which two variables are related.

It ranges from -1.0 to +1.0

A positive correlation coefficient indicates a positive relationship, a negative coefficient indicates an inverse relationship.

TYPES OF CORRELATIONS:

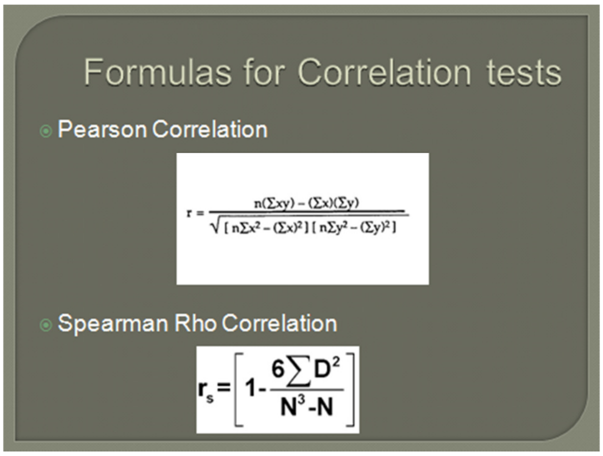
I. Pearson Correlation:

- ❖ Both the variables are quantitative and normally distributed.

II. Spearman Rho Correlation

- ❖ Both variables are quantitative but are non normally distributed.
- ❖ One normally distributed quantitative variable and the other one is non normally distributed.
- ❖ Both variables are measured at ordinal scale.

FORMULA FOR CORRELATION:



Formulas for Correlation tests

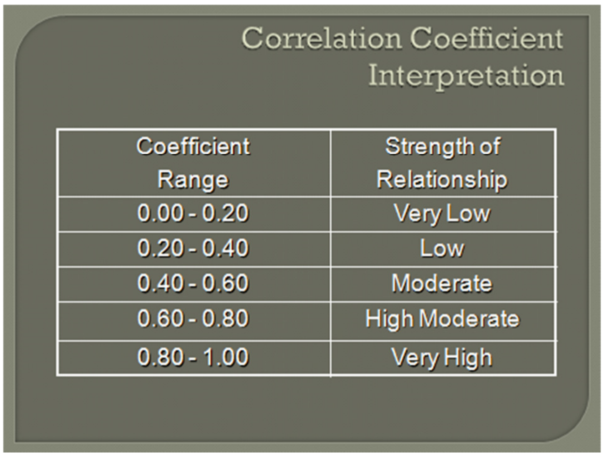
- Pearson Correlation

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

- Spearman Rho Correlation

$$r_s = \left[1 - \frac{6\sum D^2}{N^3 - N}\right]$$

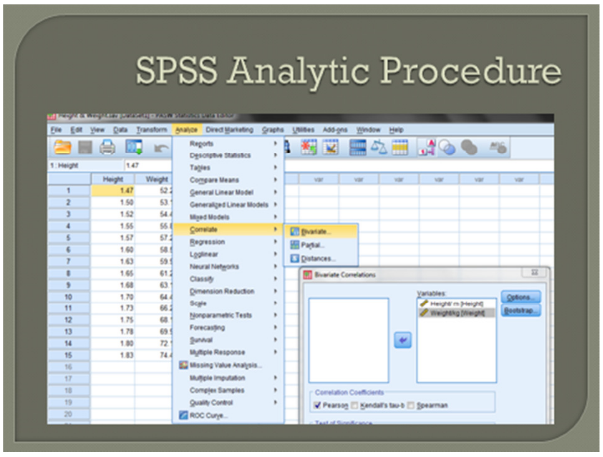
CORRELATION COEFFICIENT:



Correlation Coefficient Interpretation

Coefficient Range	Strength of Relationship
0.00 - 0.20	Very Low
0.20 - 0.40	Low
0.40 - 0.60	Moderate
0.60 - 0.80	High Moderate
0.80 - 1.00	Very High

CORRELATION IN SPSS:



SPSS Analytic Procedure

1 : Height 1.47

	Height	Weight
1	1.47	52.2
2	1.50	53.1
3	1.52	54.4
4	1.55	55.[illegible]
5	1.57	57.2
6	1.60	58.5
7	1.63	59.5
8	1.65	61.2
9	1.68	63.1
10	1.70	64.4
11	1.73	66.2
12	1.75	68.1
13	1.78	69.5
14	1.80	72.1
15	1.83	74.4

Variables: Height/ m [Height], Weight/kg [Weight]

Correlation Coefficients

☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Session II : Non – Parametric Techniques

Resource Person: Mr. M. Bilal

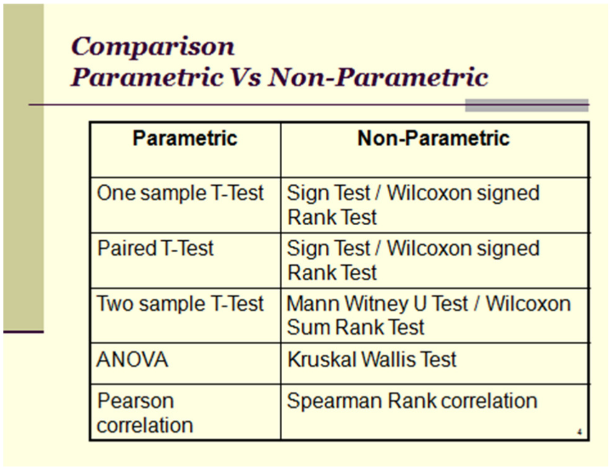
Time: 11:30 a.m. – 01:00 p.m.

REVIEW:

The parametric tests are applied when normality (and homogeneity of variance) assumptions are satisfied otherwise the equivalent non-parametric test are used.

Nonparametric statistics or distribution-free tests are those that do not rely on parameter estimates or precise assumptions about the distributions of variables.

PARAMETRIC VS NON-PARAMETRIC TESTS:



Comparison
Parametric Vs Non-Parametric

Parametric	Non-Parametric
One sample T-Test	Sign Test / Wilcoxon signed Rank Test
Paired T-Test	Sign Test / Wilcoxon signed Rank Test
Two sample T-Test	Mann Witney U Test / Wilcoxon Sum Rank Test
ANOVA	Kruskal Wallis Test
Pearson correlation	Spearman Rank correlation

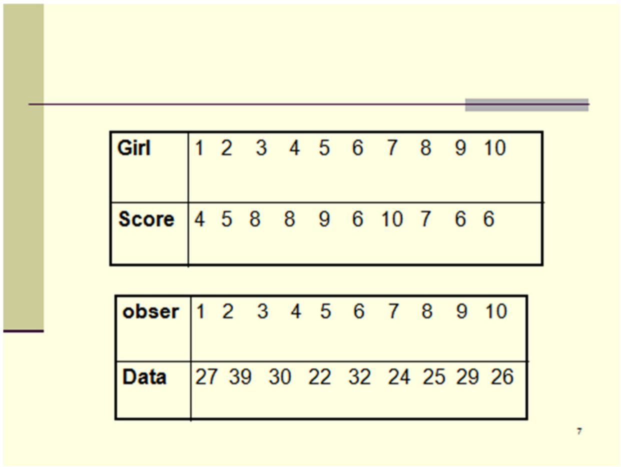
SIGN TEST:

The sign test may be used to compare two samples of observations when the populations from which the values are drawn are not independent. The sign test is used to evaluate the null hypothesis that in the underlying population of differences among pairs, median difference is equal to zero.

EXAMPLE:

Researchers wish to know if instruction in personal care and grooming would improve the appearance of mentally retarded girls. In a school for the mentally retarded 10 girls selected at random received special instruction in personal care and grooming. We wish to know if we conclude that the median score of the population from which we assume this sample to have been drawn is different from 5.

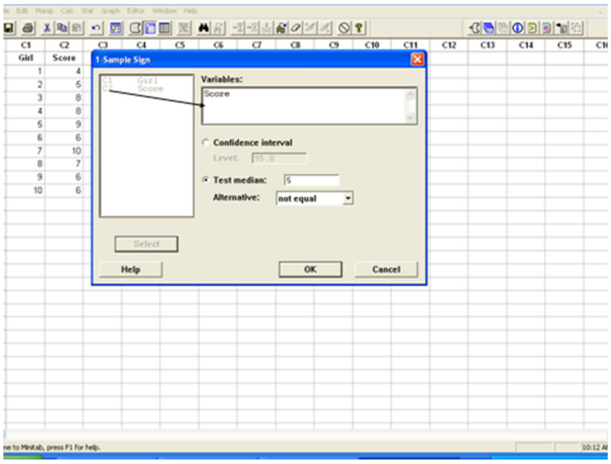
DATA:



Girl	1	2	3	4	5	6	7	8	9	10
Score	4	5	8	8	9	6	10	7	6	6

obser	1	2	3	4	5	6	7	8	9	10
Data	27	39	30	22	32	24	25	29	26	

RUNNING ONE SAMPLE SIGN TEST WITH MINITAB:



C1 Girl	C2 Score
1	4
2	5
3	8
4	8
5	9
6	6
7	10
8	7
9	6
10	6

1-Sample Sign

Variables: Score

○ Confidence interval

Level: 95.0

● Test median: 5

Alternative: not equal

THE WILCOXON SIGNED RANK TEST:

The data for analysis are measured on at least an interval scale, the Sign test may be undesirable since it would not make full use of the information contained in the data. A more appropriate Procedure might be the Wilcoxon Signed Ranked Test which makes use of the magnitudes of the differences between measurements and hypothesized location parameter rather than just the signs of the differences.

EXAMPLE:

A sample of 15 patients suffering from asthma participated in an experiment to study the effect of a new treatment on pulmonary function. Among the various measurements recorded were those of forced expiratory volume (liters) in 1 second before and after application of the treatment.

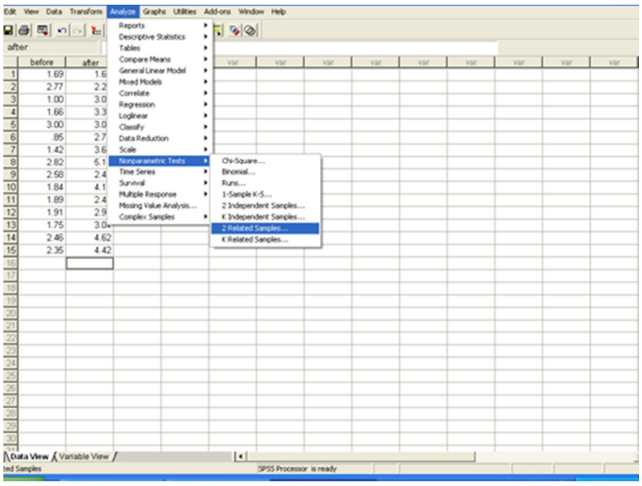
Before	1.69, 2.77, 1, 1.66, 3, 0.85, 1.42, 2.82, 2.58, 1.84, 1.89, 1.91, 1.75, 2.46, 2.35
After	1.69, 2.22, 3.07, 3.35, 3, 2.74, 3.61, 5.14, 2.44, 4.17, 2.42, 2.94, 3.04, 4.62, 4.42

FORMULA:

Test statistic: $$Z_T = \frac{T - \mu_T}{\sigma_T}$$

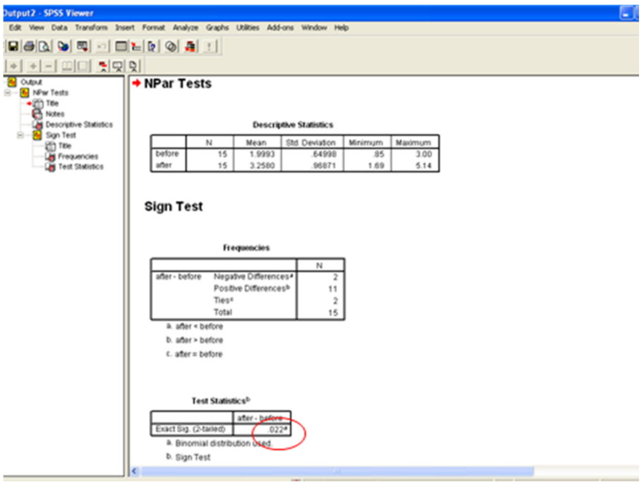
Where: $$\mu_T = \frac{n(n+1)}{4} \qquad \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

RUNNING THE TEST IN SPSS:



	before	after
1	1.69	1.6
2	2.77	2.2
3	1.00	3.0
4	1.66	3.3
5	3.00	3.0
6	.85	2.7
7	1.42	3.6
8	2.82	5.1
9	2.58	2.4
10	1.84	4.1
11	1.89	2.4
12	1.91	2.9
13	1.75	3.0
14	2.46	4.62
15	2.35	4.42

OUTPUT:



NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
before	15	1.9993	.64998	.85	3.00
after	15	3.2580	.96871	1.69	5.14

Sign Test

Frequencies

		N
after - before	Negative Differences[a]	2
	Positive Differences[b]	11
	Ties[c]	2
	Total	15

a. after < before

b. after > before

c. after = before

Test Statistics[b]

	after - before
Exact Sig. (2-tailed)	.022[a]

a. Binomial distribution used.

b. Sign Test

THE MANN WHITNEY U TEST :

Used to compare the ranks of two independent groups, comparable to the purpose of the t test.

It is more powerful than the median test since the latter only considers the number of cases above & below the median, not the rank order of the cases

ASSUMPTIONS:

- The dependent variable is ordinally scaled instead of interval or ratio.
- The assumption of normality has been violated in a t-test (especially if the sample size is small.)
- The assumption of homogeneity of variance has been violated in a t-test.

EXAMPLE:

A researcher designed an experiment to assess the effects of prolonged inhalation of cadmium oxide. Fifteen laboratory animals served as experimental subjects, while 10 similar animals served as controls. The variable of interest was hemoglobin level following the experiment. We wish to know if we can conclude that prolonged inhalation of cadmium oxide reduces hemoglobin level.

HYPOTHESIS AND FORMULA:

Hypothesis Formulation

$$H_0 : M_X = M_{Y,} \; H_1 : M_X \neq M_Y$$

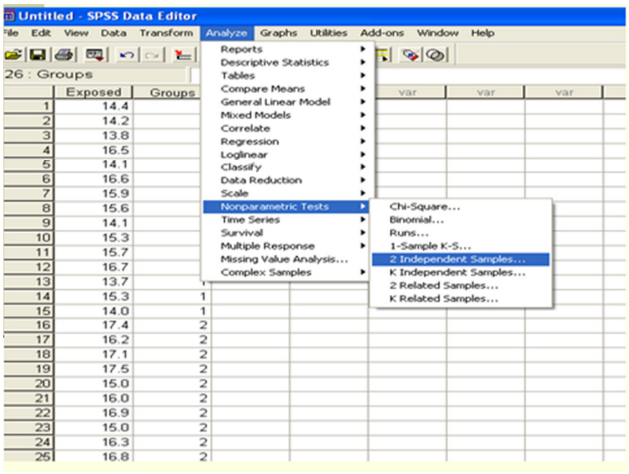
Test Statistic is:

$$T = S - \frac{n(n+1)}{2}$$

Large sample Approximation: when either n or m is greater than 20 we may use

$$Z = \frac{T - mn/2}{\sqrt{mn(n+m+1)/12}}$$

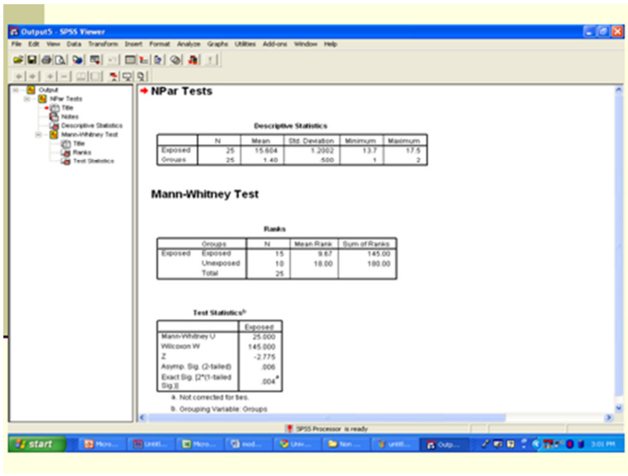
RUNNING THE TEST IN SPSS:



26 : Groups

	Exposed	Groups
1	14.4	
2	14.2	
3	13.8	
4	16.5	
5	14.1	
6	16.6	
7	15.9	
8	15.6	
9	14.1	
10	15.3	
11	15.7	
12	16.7	
13	13.7	1
14	15.3	1
15	14.0	1
16	17.4	2
17	16.2	2
18	17.1	2
19	17.5	2
20	15.0	2
21	16.0	2
22	16.9	2
23	15.0	2
24	16.3	2
25	16.8	2

OUTPUT:



NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Exposed	25	15.604	1.2002	13.7	17.5
Groups	25	1.40	.500	1	2

Mann-Whitney Test

Ranks

	Groups	N	Mean Rank	Sum of Ranks
Exposed	Exposed	15	9.67	145.00
	Unexposed	10	18.00	180.00
	Total	25		

Test Statistics[b]

	Exposed
Mann-Whitney U	25.000
Wilcoxon W	145.000
Z	-2.775
Asymp. Sig. (2-tailed)	.006
Exact Sig. [2*(1-tailed Sig.)]	.004[a]

a. Not corrected for ties.

b. Grouping Variable: Groups

KRUSKAL WALLIS H TEST:

A nonparametric alternative procedure to a one way analysis of variance of the F test for the testing the equality of several means, is the Kruskal Wallis test. This test is a generalization of the two sample Mann-Whitney U test.

EXAMPLE:

The following data represent the operating times in hours for three types of scientific calculators before a recharge is required:

Calculator	Operating times in hours
A	4.9, 6.1, 4.3, 4.6, 5.3
B	5.5, 5.4, 6.2, 5.8, 5.5, 5.2. 4.8
C	6.4, 6.8, 5.6, 6.5, 6.3, 6.6

TEST STATISTIC:

$$H = \frac{(n-1)(S_k^2 - C)}{S_r^2 - C}$$

Where:

$$S_k^2 = \sum \frac{R_i^2}{n_i}$$

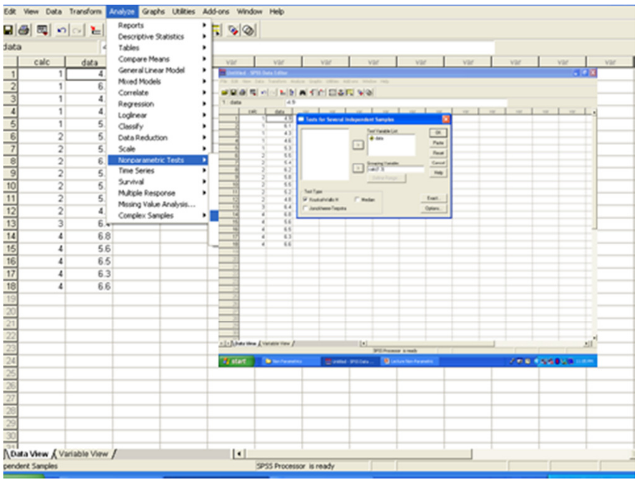
$$S_r^2 = \sum r_{ij}^2$$

$$C = \frac{n(n+1)^2}{4}$$

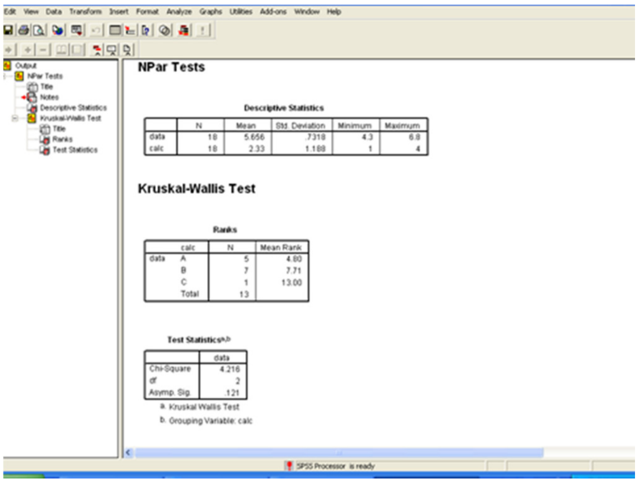
If No ties the statistic H is

$$H = \frac{12 S_k^2}{n(n+1)} - 3(n+1)$$

RUNNING THE TEST IN SPSS:



OUTPUT:



NPar Tests

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
data	18	5.656	.7318	4.3	6.8
calc	18	2.33	1.188	1	4

Kruskal-Wallis Test

Ranks

	calc	N	Mean Rank
data	A	5	4.80
	B	7	7.71
	C	1	13.00
	Total	13	

Test Statistics[a,b]

	data
Chi-Square	4.216
df	2
Asymp. Sig.	.121

[a] Kruskal Wallis Test

[b] Grouping Variable: calc

SPEARMAN RANK CORRELATION COEFFICIENT:

Several nonparametric measures of correlation are available to the researcher. Of these a frequently used procedure that is attractive because of the simplicity of the calculations involved is due to Spearman. The measure of correlation computed this method is called the Spearman Rank Correlation coefficient.

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Assumptions of Parametric Approach

- Randomization.
- Normality.
- Homogeneity of Variance (the population SD or Variance are equal or approximately equal).
- The data is taken from an interval or ratio scale.

Session I – Lecture on Parametric Approach

Session III: Hands on Training

Facilitator: Mr. M. Bilal, Mr. Waqas Sami and Mr. Waqas Latif

Time: 02:00 p.m. – 3: 30 p.m.

Session III was based on hands on training on computers. The participants were provided datasheets for practicing parametric and non-parametric tests application on SPSS followed by reporting of results in statistical language.

Exercise – Parametric Approach

Exercise 1:

The data contains fluoride concentrations from different towns in Lahore city, according to the guidelines by WHO concentration of fluoride in water should not exceed 1.5 mg, test the hypothesis at 5% significance level, and interpret the results.

1.9, 2.6, 1.8, 1.2, 1.4, 1.3, 0.9, 0.6, 0.5, 0.4, 0.3, 0.4, 0.2, 0.1, 0.2, 1.8, 1.3, 1.5, 0.8, 1.1

Exercise 2:

Twelve hogs were fed on diet A and Fifteen on diet B. The gains in weights for the individual hogs (in pounds) were as follows:

Diet A: 25, 30, 28, 34, 24, 25, 13, 32, 24, 30, 31, 35.

Diet B: 44, 34, 22, 08, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22.

Test at 5% level of significance that is there any difference in diet on which the hogs were fed.

Exercise 3:

Ten young recruits were put through a strenuous physical training programme by the Army. Their weights were recorded before and after the training with the following results:

Recruit	1	2	3	4	5	6	7	8	9	10
Weight before	125	195	160	171	140	201	170	176	195	139
Weight after	136	201	158	184	145	195	175	190	190	145

Using α = 0.05, would you say that the programme affects the average weight of recruits?

Exercise 4:

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. Test at 5% level of significance.

HB SS	7.2	7.7	8.0	8.1	8.3	8.4	8.4	8.5	8.6	8.7	9.1	9.1	9.8	10.1	10.3
HB S Thalassaemia	8.1	9.2	10.0	10.4	10.6	10.9	11.1	11.9	12.0	12.1					
HB SC	10.7	11.3	11.5	11.6	11.7	11.8	12.0	12.1	12.3	12.6	12.6	13.3	13.8	13.3	13.9

Exercise 5:

The following data is obtained on 12 males between the ages of 12 and 18 years. Calculate the Pearson correlation coefficient between Height & Radius Length, Height & Femur Length and Radius & Femur Length, also interpret the results.

Height	Radius Length	Femur Length
149.0	21.0	42.0
152.0	21.79	43.70
155.70	22.40	44.75
159.0	23.0	46.0
163.30	23.70	47.0
166.0	24.30	47.90
169.0	24.92	48.95
172.0	25.50	49.90
174.50	25.80	49.90
174.50	25.80	50.30
176.10	26.01	50.90
176.50	26.15	50.85
179.0	26.30	51.10

Exercise – Non – Parametric Approach

Exercise 1:

Researchers wish to know if instruction in personal care and grooming would improve the appearance of mentally retarded girls. In a school for the mentally retarded 10 girls selected at random received special instruction in personal care and grooming. We wish to know if we conclude that the median score of the population from which we assume this sample to have been drawn is different from 5.

Girl	1 2 3 4 5 6 7 8 9 10
Score	4 5 8 8 9 6 10 7 6 6

Exercise 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 the various measurements recorded were those of forced expiratory volume (liters) in 1 second before and after application of the treatment. The results are given below: Run the Wilcoxon Signed Rank Test and Sign Test.

Before	1.69, 2.77, 1, 1.66, 3, 0.85, 1.42, 2.82, 2.58, 1.84, 1.89, 1.91, 1.75, 2.46, 2.35
After	1.69, 2.22, 3.07, 3.35, 3, 2.74, 3.61, 5.14, 2.44, 4.17, 2.42, 2.94, 3.04, 4.62, 4.42

Exercise 3:

A researcher designed an experiment to assess the effects of prolonged inhalation of cadmium oxide. Fifteen laboratory animals served as experimental subjects, while 10 similar animals served as controls. The variable of interest was hemoglobin level following the experiment. We wish to know if we can conclude that prolonged inhalation of cadmium oxide reduces hemoglobin level.

Exposed	14.4	14.2	13.8	16.5	14.1	16.6	15.9	15.6	15.9	15.6	14.1	15.3
Unexposed	17.4	16.2	17.1	17.5	15.0	16.0	16.9	15.0	16.9	15.0	16.3	16.8

Exercise 4:

The effects of two drugs on reaction time to a certain stimulus were studied in three samples of experimental animals. Sample III served as a control while the animals in sample I were treated with drug A and those in sample II were treated with drug B prior to the application of the stimulus shows the reaction times in seconds of the 13 animals. Can we conclude that three population represented by three samples differ with respect to reaction time.

Reaction Time in Seconds of 13 Experimental Animals

Sample		
I	II	III
17	8	2
20	7	5
40	9	4
31	8	3
35		

DAY 3

March 24, 2011

Session I : Regression Analysis

Resource Person : Dr. Qaiser Shahbaz

Time: 09:00 a.m. – 10:30 p.m.

DEPENDENT VARIABLE: (QUANTITATIVE)

- Regression Analysis deals with prediction of one or more (*Dependent* variables) on the basis of one or more *Fixed/Random* variables (*Independent* Variables, *Regressors*).
- Purpose is to fit an optimum model that can be used for prediction with least possible error and most significant regressors.
- The models are collectively called the *Regression Models.*

Regression Analysis:

I. Model Identification

- Scatter Plots and Matrix Plots

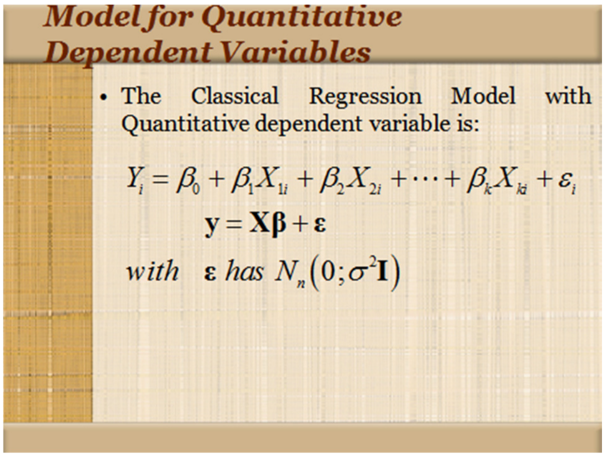
II. Models Estimation:

- Classical Least Squares Estimation
- Weighted Least Square Estimation
- Generalized Least Squares Estimation
- Iteratively Re–weighted Least Squares Estimation
- Maximum Likelihood Estimation

MODEL DIAGNOSIS:

- Outliers
- Residual Analysis
- Autocorrelations
- Heteroscedasticity
- Multicollinearity
- Leverage Values
- Influential Observations
- Model Validation

MODEL FOR QUANTITATIVE DEPENDENT VARIABLE:



Model for Quantitative Dependent Variables

- The Classical Regression Model with Quantitative dependent variable is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

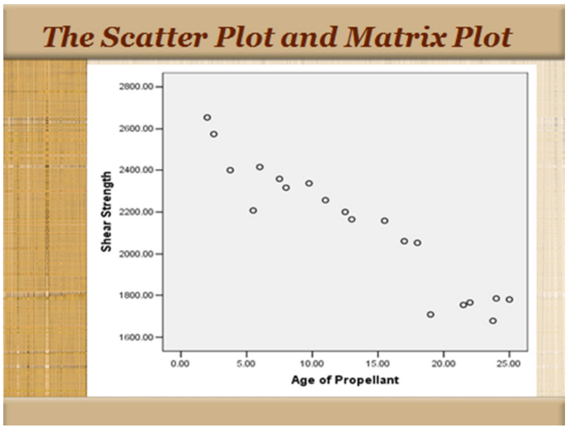
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\text{with} \quad \boldsymbol{\varepsilon} \text{ has } N_n\left(0;\sigma^2\mathbf{I}\right)$$

MODEL IDENTIFICATION:

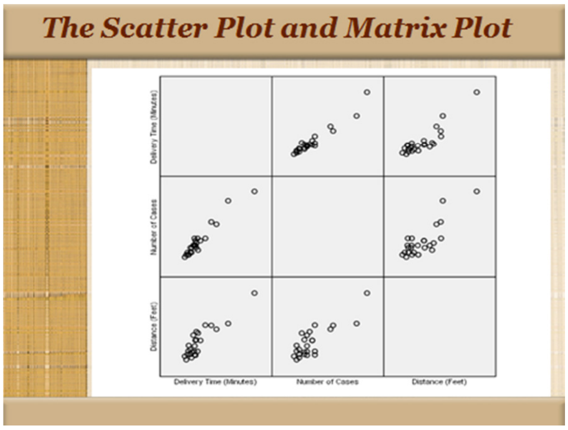
- The Scatter Plot and Matrix Plot can be used for model identification.
- If plots show linear trend then the Linear Regression Model is appropriate.
- If linear trend is not evident then a transformation of variables is done.
- Either Dependent or Independent Variables can be transformed.

SCATTER PLOT:



The Scatter Plot and Matrix Plot

MATRIX PLOT:



The Scatter Plot and Matrix Plot

INTERPRETATION OF MODEL PARAMETERS:

The estimated model is:

$$E\left(Y_i\right)=\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i}+\cdots+\hat{\beta}_k X_{ki}$$

The Coefficient is the Mean Value of *Y* when all *Independent* variables are zero. The Coefficient is the *Partial Effect* of j–th *Independent* variable.

IMPORTANT MEASURES IN REGRESSION ANALYSIS:

- R^2: Measures the Proportion of Variation explained by the regression model
- $S_{Y.X}$: Measures the amount of error in the predicted mean value of dependent variable
- Adjusted R^2: consider the number of explanatory variables in the model

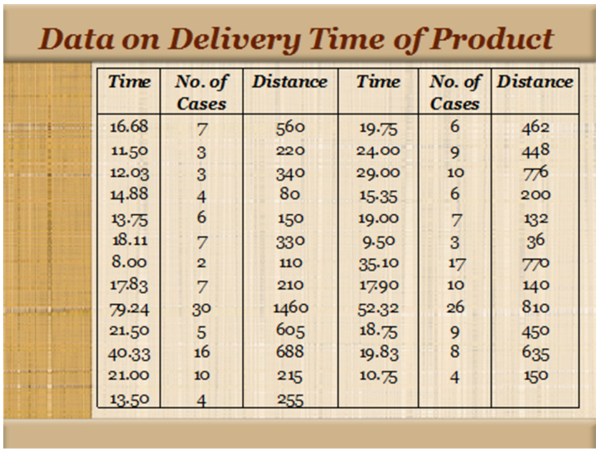
TEST OF SIGNIFICANCE FOR MODEL:

- Certain tests of significance for the model can be conducted.
- Significance of *Full Model* is tested by using the *F* – Statistic.
- Significance of *Individual* parameters is tested by using the *t* – Statistic.
- Confidence Intervals for parameters are constructed by using the *t* – Statistic.
- Confidence Intervals for the *Predicted Mean* value of dependent variable are constructed by using the *t* – Statistic.

EXAMPLE:

A soft drink bottler is analyzing the vending machine service rout in his distribution system. He is interested in predicting the amount of time require by the rout driver to service the vending machines in the outlet. Data on Delivery Time (Y), Product Stocked (X_1) and Distance Walked (X_2) is collected and is given:

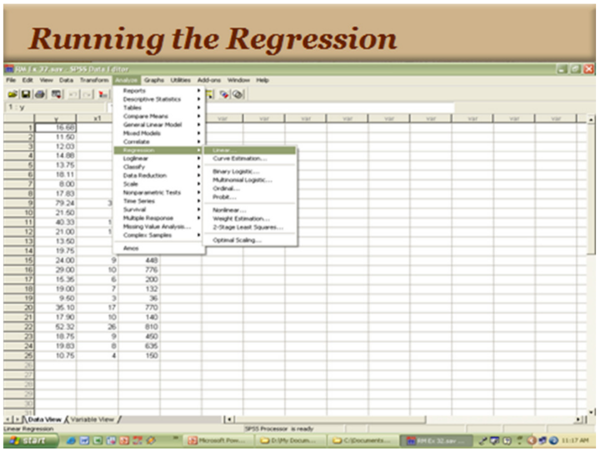
DATASHEET:



Data on Delivery Time of Product

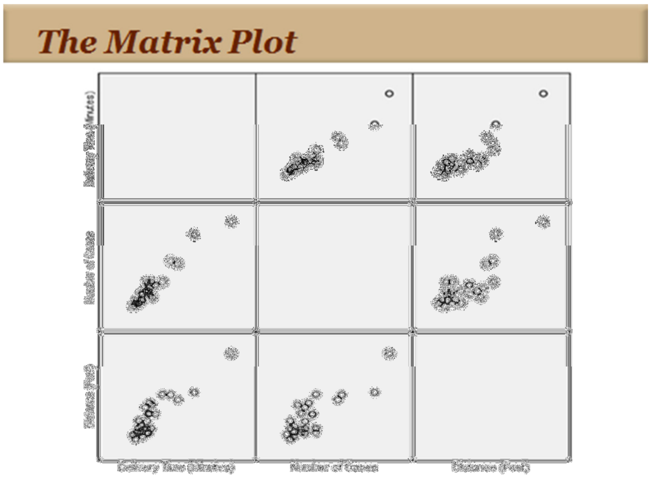
Time	No. of Cases	Distance	Time	No. of Cases	Distance
16.68	7	560	19.75	6	462
11.50	3	220	24.00	9	448
12.03	3	340	29.00	10	776
14.88	4	80	15.35	6	200
13.75	6	150	19.00	7	132
18.11	7	330	9.50	3	36
8.00	2	110	35.10	17	770
17.83	7	210	17.90	10	140
79.24	30	1460	52.32	26	810
21.50	5	605	18.75	9	450
40.33	16	688	19.83	8	635
21.00	10	215	10.75	4	150
13.50	4	255			

RUNNIG THE SIMPLE LINEAR REGRESSION IN SPSS:



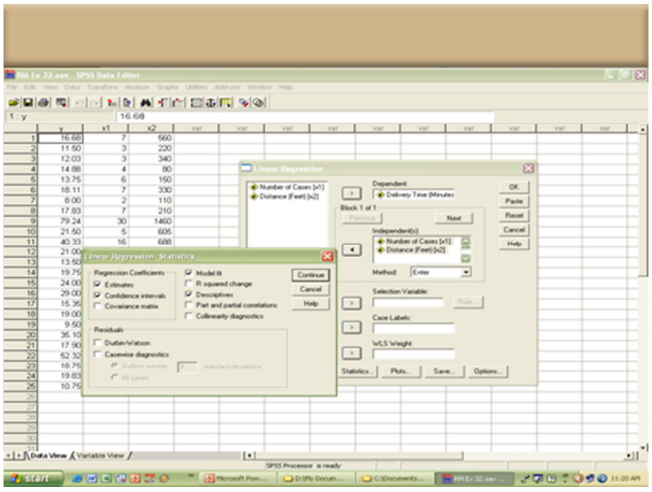
Running the Regression

MATRIX PLOT:

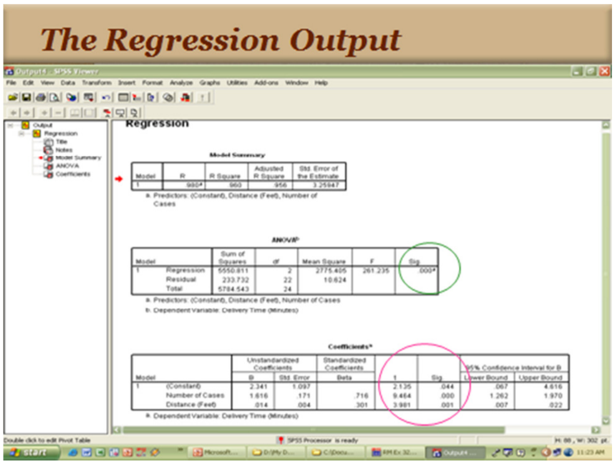


The Matrix Plot

MODEL DIAGNOSTICS:



REGRESSION OUTPUT:



The Regression Output

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.980[a]	.960	.956	3.25947

a. Predictors: (Constant), Distance (Feet), Number of Cases

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5550.811	2	2775.405	261.235	.000[a]
	Residual	233.732	22	10.624		
	Total	5784.543	24			

a. Predictors: (Constant), Distance (Feet), Number of Cases

b. Dependent Variable: Delivery Time (Minutes)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	2.341	1.097		2.135	.044	.067	4.616
	Number of Cases	1.616	.171	.716	9.464	.000	1.262	1.970
	Distance (Feet)	.014	.004	.301	3.981	.001	.007	.022

a. Dependent Variable: Delivery Time (Minutes)

REGRESSION ANALYSIS:
(QUALITATIVE DEPENDENT VARIABLE)

- The *Regression Analysis* deals with *prediction* of the *Mean* value of the *Dependent* variable by using information of *Independent* variables.

- *Nature* of the dependent variable plays very important role in the regression analysis.

- Major types of the dependent variable encountered in the regression analysis are Quantitative and Qualitative types.

- Estimation framework differ for both type of the dependent variables.

TYPES OF QUALITATIVE DEPENDENT VARIABLES:

Following major types of Qualitative Variables are met in practice:

- Binary Variable

- Categorical without Order

- Categorical with Order

REGRESSION WITH BINARY DEPENDENT VARIABLE:

- The dependent variable is *Binary*.

- Distribution of the Error is *Binomial*

- Several models are available depending upon the *Link Function;* two most popular are:

  - The Binary Logistic Regression

  - The Probit Regression

- The Models are used to predict the probability of falling in the success category given the information of explanatory variables.

BINARY LOGISTIC REGRESSION ANALYSIS:

- The Dependent variable is *Binary*, say for example, recovery from a disease (Yes, No) ; qualifying an entry test (Yes, No) etc.

- The "*Yes*" category is generally referred to as the success category.

- Used to model the *probability* of having in the *success* category given the information of independent variables.

- Can also be used to predict the *Logit* of the success category

- Commonly used in *Medical* sciences.

MODEL FOR BINARY LOGISTIC:

The *Logistic Regression* model; used to predict the *probability* of dependent variable to have in the success category given the information of explanatory variables; is given as:

$$P\left(Y=1|\mathbf{x}\right)=\frac{1}{1+\exp\left[-\left\{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k\right\}\right]}+\varepsilon$$

The *Logit* model; used to predict *Logit* of the dependent variable; is given as:

$$ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k+\varepsilon$$

TESTING ADEQUACY AND SIGNIFICANCE OF THE MODEL:

Adequacy of the *Logistic Regression* can be tested by using the *Deviance* statistic that measures difference between *saturated* model and the *fitted* model. An *insignificant* result indicates that the model is *adequate*.

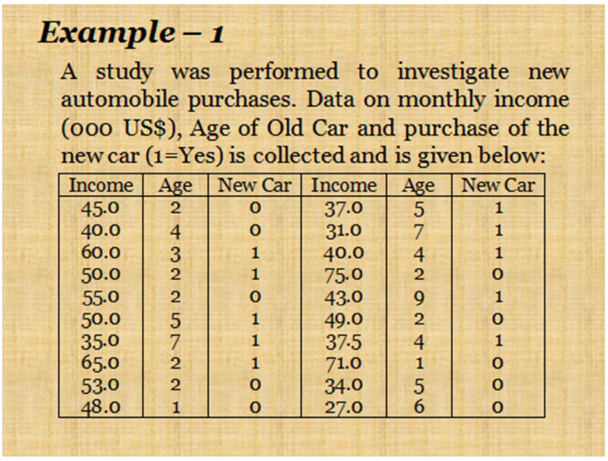
Significance of the model is tested by using the model Chi–Square. This test tests whether all of the regression coefficients are significantly different from zero. A *significant* result indicates that the coefficients are *different from zero*.

DATA FORMAT FOR LOGISTIC REGRESSION:

Two formats of data are available for *Logistic Regression*.

- ❖ The *Raw Format*: The data is entered as it is collected.
- ❖ *Covariate Class Format*: The data is entered in the form of groups.

Example:

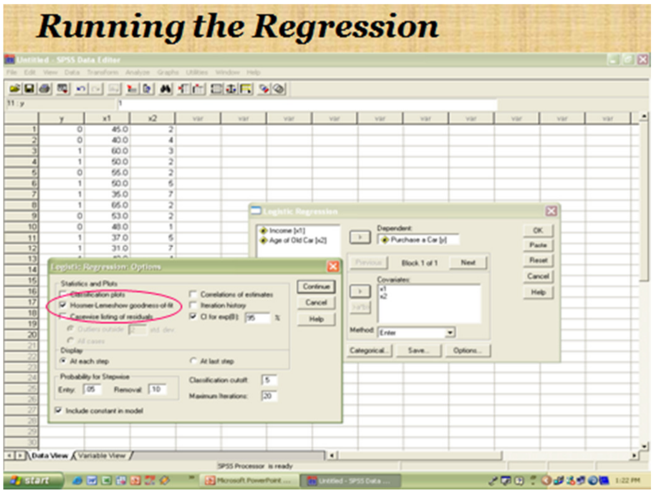


Example – 1

A study was performed to investigate new automobile purchases. Data on monthly income (000 US$), Age of Old Car and purchase of the new car (1=Yes) is collected and is given below:

Income	Age	New Car	Income	Age	New Car
45.0	2	0	37.0	5	1
40.0	4	0	31.0	7	1
60.0	3	1	40.0	4	1
50.0	2	1	75.0	2	0
55.0	2	0	43.0	9	1
50.0	5	1	49.0	2	0
35.0	7	1	37.5	4	1
65.0	2	1	71.0	1	0
53.0	2	0	34.0	5	0
48.0	1	0	27.0	6	0

DIAGNOISTIC MEASURES:



Running the Regression

Untitled - SPSS Data Editor

	y	x1	x2
1	0	45.0	2
2	0	40.0	4
3	1	60.0	3
4	1	50.0	2
5	0	55.0	2
6	1	50.0	5
7	1	35.0	7
8	1	65.0	2
9	0	53.0	2
10	0	48.0	1
11	1	37.0	5
12	1	31.0	7

Logistic Regression: Dependent: Purchase a Car [y]; Covariates: x1, x2; Method: Enter

Logistic Regression: Options

Statistics and Plots

- □ Classification plots
- ☑ Hosmer-Lemeshow goodness-of-fit
- □ Casewise listing of residuals
- □ Correlations of estimates
- □ Iteration history
- ☑ CI for exp(B): 95 %

Display: ● At each step ○ At last step

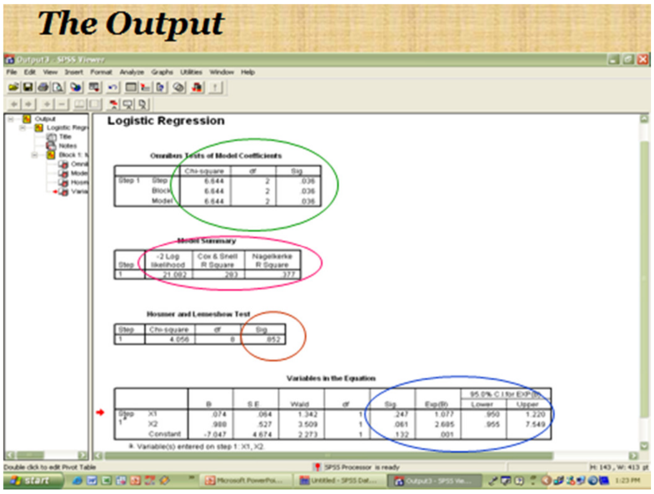
Probability for Stepwise: Entry: .05 Removal: .10

Classification cutoff: .5

Maximum Iterations: 20

☑ Include constant in model

BINARY LOGISTIC OUTPUT:



The Output

Logistic Regression

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	6.644	2	.036
	Block	6.644	2	.036
	Model	6.644	2	.036

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	21.082	.283	.377

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.056	8	.852

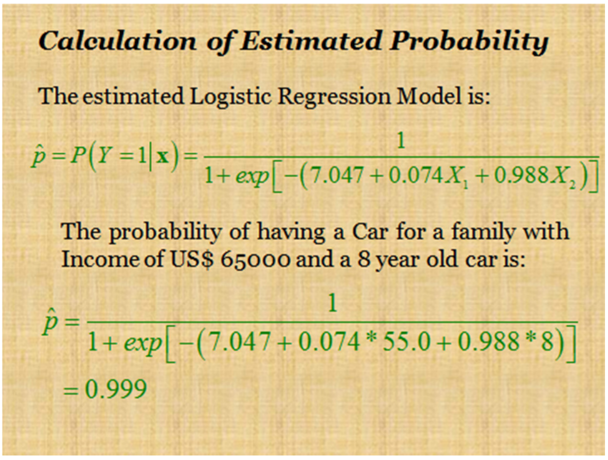
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1[a]	X1	.074	.064	1.342	1	.247	1.077	.950	1.220
	X2	.988	.527	3.509	1	.061	2.685	.955	7.549
	Constant	-7.047	4.674	2.273	1	.132	.001		

a. Variable(s) entered on step 1: X1, X2.

Output3 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
SPSS Processor is ready

ESTIMATED PROBABILITIES:



Calculation of Estimated Probability

The estimated Logistic Regression Model is:

$$\hat{p} = P(Y = 1|\mathbf{x}) = \frac{1}{1 + exp\left[-(7.047 + 0.074X_1 + 0.988X_2)\right]}$$

The probability of having a Car for a family with Income of US$ 65000 and a 8 year old car is:

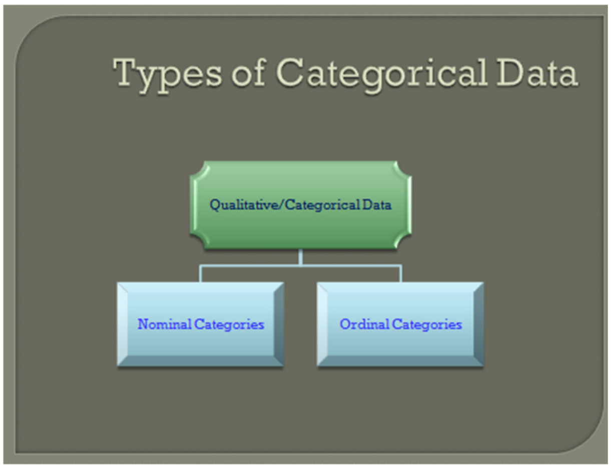
$$\hat{p} = \frac{1}{1 + exp\left[-(7.047 + 0.074 * 55.0 + 0.988 * 8)\right]}$$

$$= 0.999$$

Session II : Catagorical Data Analysis

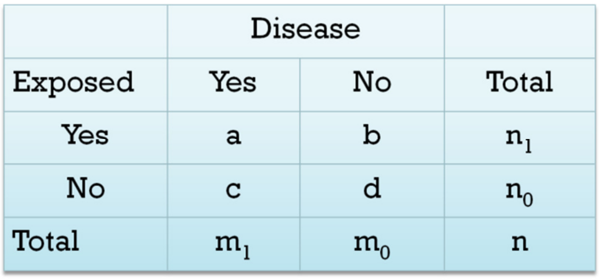
Resource Person: Mr. Nadeem Shafique Butt

Time: 11:30 a.m. – 01:00 p.m.



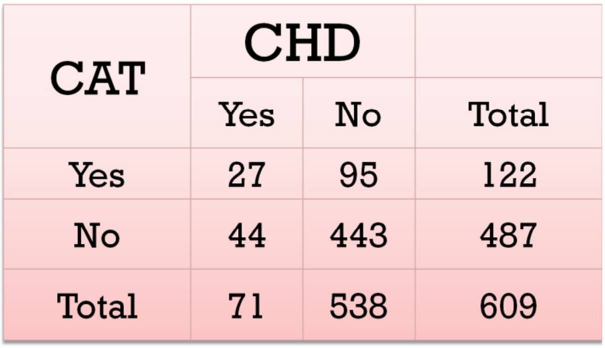
RATES & RATIOS:

2 by 2 Contingency Table Analysis



	Disease		
Exposed	Yes	No	Total
Yes	a	b	n_1
No	c	d	n_0
Total	m_1	m_0	n

EXAMPLE OF 2 X 2 TABLE:

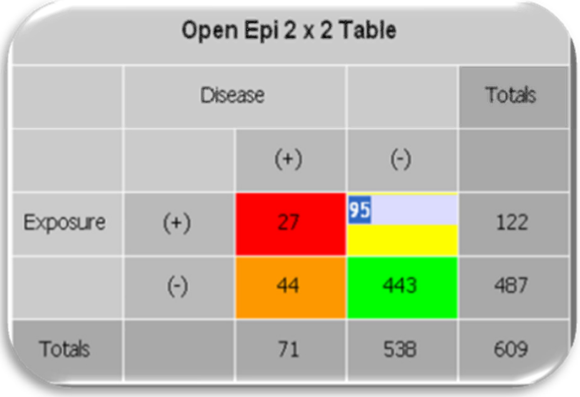


CAT	CHD		
	Yes	No	Total
Yes	27	95	122
No	44	443	487
Total	71	538	609

COMMON RATES & RATIOS:

- ❖ Risk Ratio/Relative Risk
- ❖ Risk Difference
- ❖ Odds Ratio

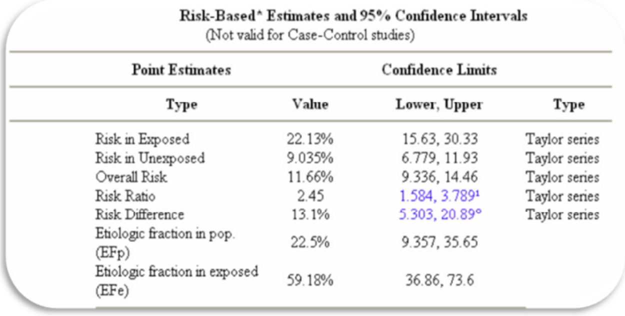
Example – Risk Ratio



Open Epi 2 x 2 Table

		Disease		Totals
		(+)	(-)	
Exposure	(+)	27	95	122
	(-)	44	443	487
Totals		71	538	609

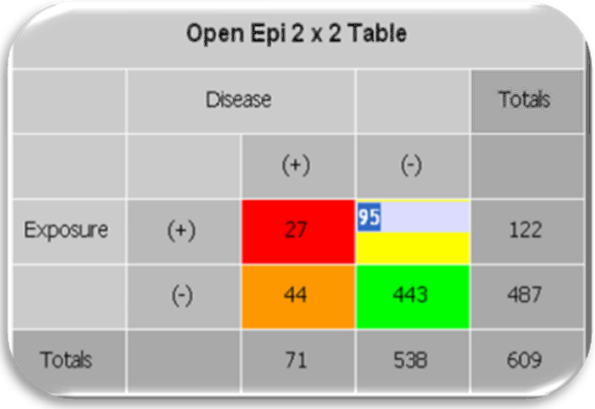
RESULTS:



Risk-Based* Estimates and 95% Confidence Intervals
(Not valid for Case-Control studies)

Point Estimates		Confidence Limits	
Type	Value	Lower, Upper	Type
Risk in Exposed	22.13%	15.63, 30.33	Taylor series
Risk in Unexposed	9.035%	6.779, 11.93	Taylor series
Overall Risk	11.66%	9.336, 14.46	Taylor series
Risk Ratio	2.45	1.584, 3.789^1	Taylor series
Risk Difference	13.1%	5.303, 20.89°	Taylor series
Etiologic fraction in pop. (EFp)	22.5%	9.357, 35.65	
Etiologic fraction in exposed (EFe)	59.18%	36.86, 73.6	

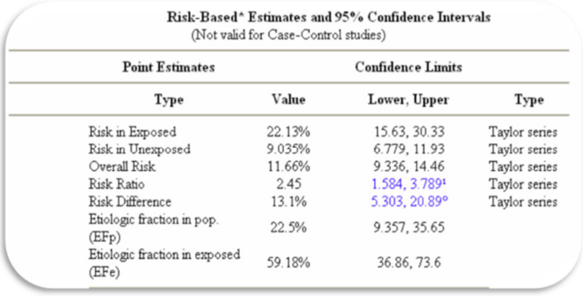
EXAMPLE – RISK DIFFERENCE:



Open Epi 2 x 2 Table

	Disease			Totals
		(+)	(-)	
Exposure	(+)	27	95	122
	(-)	44	443	487
Totals		71	538	609

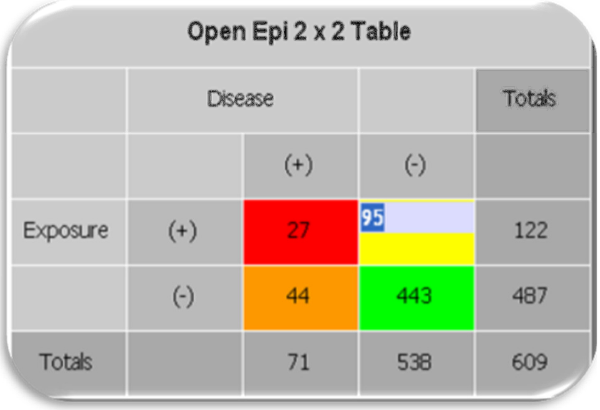
RESULTS:



Risk-Based* Estimates and 95% Confidence Intervals
(Not valid for Case-Control studies)

Point Estimates		Confidence Limits	
Type	Value	Lower, Upper	Type
Risk in Exposed	22.13%	15.63, 30.33	Taylor series
Risk in Unexposed	9.035%	6.779, 11.93	Taylor series
Overall Risk	11.66%	9.336, 14.46	Taylor series
Risk Ratio	2.45	1.584, 3.789¹	Taylor series
Risk Difference	13.1%	5.303, 20.89°	Taylor series
Etiologic fraction in pop. (EFp)	22.5%	9.357, 35.65	
Etiologic fraction in exposed (EFe)	59.18%	36.86, 73.6	

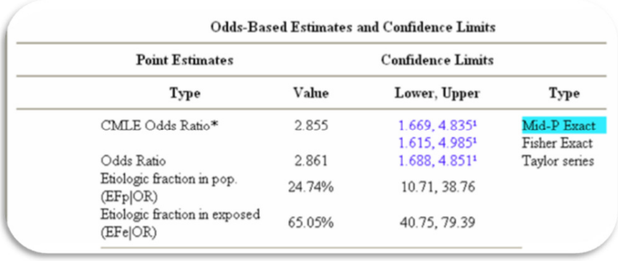
EXAMPLE – ODDS RATIO



Open Epi 2 x 2 Table

	Disease			Totals
		(+)	(-)	
Exposure	(+)	27	95	122
	(-)	44	443	487
Totals		71	538	609

RESULTS:



Odds-Based Estimates and Confidence Limits

Point Estimates		Confidence Limits	
Type	Value	Lower, Upper	Type
CMLE Odds Ratio*	2.855	1.669, 4.835¹	Mid-P Exact
		1.615, 4.985¹	Fisher Exact
Odds Ratio	2.861	1.688, 4.851¹	Taylor series
Etiologic fraction in pop. (EFp|OR)	24.74%	10.71, 38.76	
Etiologic fraction in exposed (EFe|OR)	65.05%	40.75, 79.39	

STRATIFIED 2X2 TABLE ANALYSIS:

General Layout

Exposed	Disease		
	Yes	No	Total
Yes	a_i	b_i	n_{1i}
No	c_i	d_i	n_{0i}
Total	m_{1i}	m_{0i}	n_i

EXAMPLE:

Age <= 35

Exposed CAT	Disease CHD		Total
	Yes	No	
Yes	23	74	97
No	20	134	154
Total	43	208	251

Age <= 35

Exposed CAT	Disease CHD		Total
	Yes	No	
Yes	4	21	97
No	24	309	154
Total	28	330	358

CONTINGENCY TABLE ANALYSIS:

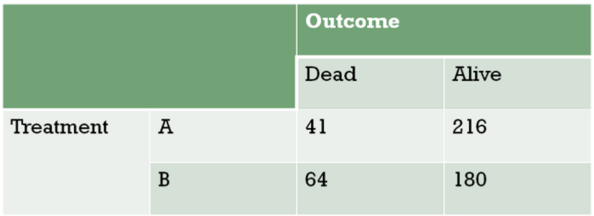
PEARSON'S CHI-SQUARE: (2 X 2)

Chi-Square test is 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 a 2 X 2 table chi-square statistic is calculated as:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

EXAMPLE:

The following represent mortality data for two groups of patients receiving different treatments, A and B.



		Outcome	
		Dead	Alive
Treatment	A	41	216
	B	64	180

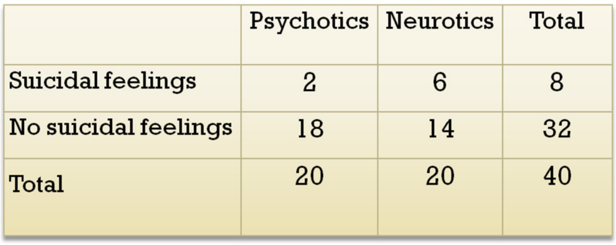
YATES CORRECTED CHI-SQUARE:

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's correction.*

$$\chi^2 = \frac{n\left[|ad-bc|-0.5n\right]^2}{(a+b)(c+d)(a+c)(b+c)}$$

EXAMPLE:

The following data relate to suicidal feelings in samples of psychotic and neurotic patients:



	Psychotics	Neurotics	Total
Suicidal feelings	2	6	8
No suicidal feelings	18	14	32
Total	20	20	40

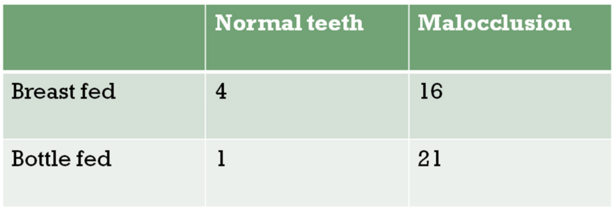
FISHER'S EXACT TEST:

The method of Yates's 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's correction as it gives exact result.

$$Fisher's\,Exact\,Test = \frac{R_1!R_2!C_1!C_2!}{n!a!b!c!d!}$$

EXAMPLE:

The following data compare malocclusion of teeth with method of feeding infants.



	Normal teeth	Malocclusion
Breast fed	4	16
Bottle fed	1	21

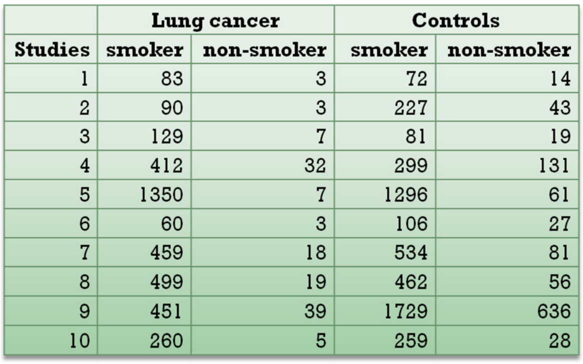
MENTAL-HAENSZEL CHI-SQUARE:

This is a method of controlling confounding in stratification. This requires that the confounder be categorical variable or if continuous, that be categorized. The formula of chi-square given by Mantel-Haenszel.

$$\chi^2_{MH} = \frac{\left[\sum \frac{a_i d_i - b_i c_i}{n_i}\right]}{\sum \frac{r_{1i} r_{2i} c_{1i} c_{2i}}{n_i^2 (n_i - 1)}}$$

EXAMPLE:

The following data compare the smoking status of lung cancer patients with controls. Ten different studies are combined in an attempt to improve the overall estimate of relative risk.



	Lung cancer		Controls	
Studies	smoker	non-smoker	smoker	non-smoker
1	83	3	72	14
2	90	3	227	43
3	129	7	81	19
4	412	32	299	131
5	1350	7	1296	61
6	60	3	106	27
7	459	18	534	81
8	499	19	462	56
9	451	39	1729	636
10	260	5	259	28

PEARSON'S CHI-SQUARE: (RXC)

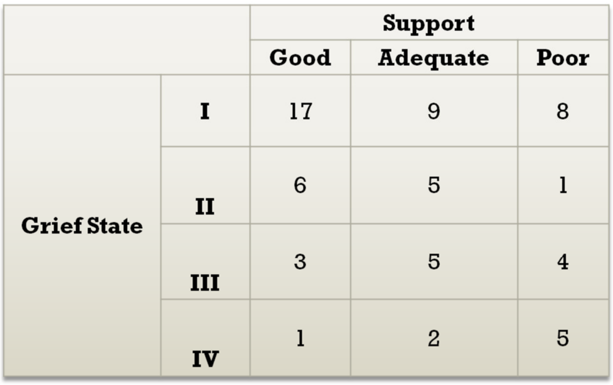
For an R x C table the Chi-Square Statistics can be calculated as:

$$\chi^2 = \sum_i \sum_j \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Yates Corrected Chi-Square for R X C contingency table can be calculated as:

$$\chi^2 = \sum \sum \frac{\left[\left|O_{ij} - E_{ij}\right| - 0.5\right]^2}{E_{ij}}$$

The following data (as above) describe the state of grief of 66 mothers who had suffered a neonatal death. The table relates this to the amount of support given to these women:



		Support		
		Good	Adequate	Poor
Grief State	I	17	9	8
	II	6	5	1
	III	3	5	4
	IV	1	2	5

FISHER'S EXACT TEST FOR RXC TABLES:

The generalized Fisher exact test is difficult to compute (Mehta and Patel, 1983, 1986a); it may take a long time and it may not be computed for the table that you enter. If the Fisher exact method cannot be computed practically then a hybrid method based upon Cochrane rules is used (Mehta and Patel, 1986b); this may also fail with large tables and/or numbers. The Fisher-Freeman-Halton result is quoted with just one P value as it is implicitly two-sided.

- Let t denote a table from the set of all tables with the same row and column margins.

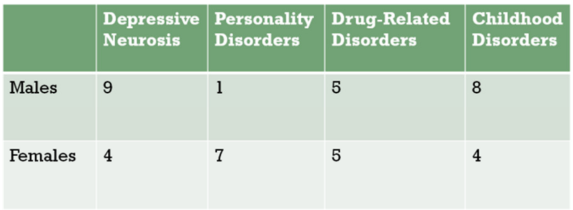
- Let D(t) be the measure of discrepancy.

- The exact two sided P value = P [D(t) >= D(t observed)] = sum of hypergeometric probabilities of those tables where D(t) is larger than or equal to the observed table.

- In large samples the distribution of D(t) conditional on fixed row and column margins converges to the chi-square distribution with (r-1)(c-1) degrees of freedom.

EXAMPLE:

Data is a classification of admissions to a mental hospital by diagnosis and gender:



	Depressive Neurosis	Personality Disorders	Drug-Related Disorders	Childhood Disorders
Males	9	1	5	8
Females	4	7	5	4

CO-EFFICIENT OF CONTINGENCY:

A measure of association based on chi-square. The value ranges between zero and 1, with zero indicating no association between the row and column variables and values close to 1 indicating a high degree of association between the variables.

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}$$

PHI & CRAMER'S V:

The Phi coefficient is a degree of association between two attributes and is calculated as:

$$Phi=\phi=\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}=\sqrt{\frac{\chi^2}{n}}$$

The range of phi is from 0 to 1. If phi is 0, the attributes are independent. If phi = 1, there is complete association.

$$V=\sqrt{\frac{\chi^2/n}{\min(r-1,c-1)}}$$

GOODMAN AND KRUSKAL'S LAMBDA:

Lambda: A measure of association which reflects the proportional reduction in error when values of the independent variable are used to predict values of the dependent variable.

A value of 1 means that the independent variable perfectly predicts the dependent variable.

A value of 0 means that the independent variable is no help in predicting the dependent variable.

$$\lambda=\frac{\sum_{j=1}^{c}\max_i\left(n_{ij}\right)-\max_i\left(n_{.j}\right)}{n-\max_i\left(n_{.j}\right)}$$

KENDALL'S TAU B:

A nonparametric measure of correlation for ordinal or ranked variables that take ties into account. The sign of the coefficient indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships. Possible values range from -1 to 1, but a value of -1 or +1 can only be obtained from square tables.

$$\tau_b = \frac{S}{\sqrt{(P+Q+X_0)(P+Q+Y_0)}}$$

Where S=P-Q
P=Concordant pairs of observation
Q=Discordant pairs of observation
m=min(r,c)

KENDALL'S TAU C:

A nonparametric measure of association for ordinal variables that ignores ties. The sign of the coefficient indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships. Possible values range from -1 to 1, but a value of -1 or +1 can only be obtained from square tables.

$$\tau_c = \frac{2mS}{n^2(m-1)}$$

Where S=P-Q
P=Concordant pairs of observation
Q=Discordant pairs of observation
m=min(r,c)

GOODMAN AND KRUSKAL'S GAMMA:

A symmetric measure of association between two ordinal variables that ranges between -1 and 1. Values close to an absolute value of 1 indicate a strong relationship between the two variables. Values close to zero indicate little or no relationship.

$$\gamma = \frac{P-Q}{P+Q}$$

SOMER'S D:

A measure of association between two ordinal variables that ranges from -1 to 1. Values close to an absolute value of 1 indicate a strong relationship between the two variables, and values close to 0 indicate little or no relationship between the variables. Somers' d is an asymmetric extension of gamma that differs only in the inclusion of the number of pairs not tied on the independent variable.

$$d_{yx} = \frac{S}{P+Q+Y_0}$$

MCNEMAR CHI-SQUARE:

Chi-Square requires that the numbers in cells are independently distributed if the problem is of matched pairs or repeated measurements the ordinary Chi-Square is not suitable. The modified form of Chi-Square for such cases is purposed by McNemar, and McNemar Chi-Square can be calculated as:

$$\chi^2_{MCNemar} = \frac{(b-c)^2}{(b+c)}$$

Dr. Qaiser Shahbaz delivering the lecture on Regression Analysis

Mr. Nadeem Shafique Butt delivering the lecture on Categorical Data Analysis

Session III: Hands on Training

Facilitator: Mr. Waqas Sami and Mr. Waqas Latif

Time: 02:00 p.m. – 3: 30 p.m.

Session III was based on hands on training on computers. The participants were provided datasheets for practicing regression analysis and categorical data analysis application on SPSS followed by reporting of results in statistical language.

Exercise – Regression Analysis

A University medical centre urology group was interested in the association between prostate – specific antigen (PSA) and number of prognostic clinical measurements in men with advanced prostate cancer. Data was collected on 97 men who were about to undergo radical prostectomies.

Required:

- Identify the dependent and independent variables.
- Check assumptions of normality, linearity, homocedasticity and dependence.
- Run a multiple regression model and interpret the results.

1. In a study of 100 subjects that participated in the study, the age in years along with the presence and absence of coronary heart disease is recorded. It is of interest to explore the relationship between age and presence or absence of coronary heart disease.
2. A case – control study was conducted to see the effect of coffee drinking (number of cups per day) on Myocardial infarction for a sample of 55 men under age of 55 years. Construct a logistic regression model and interpret the results.
3. A case – control study was conducted to see the effect of number of cigarettes smoked per day on Myocardial infarction for a sample of 55 men under age of 55 years. Construct a logistic regression model and interpret the results.

Exercise – Categorical Data Analysis

Exercise 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 group were classified as to blood type O, A and B. The data are given in following Table.

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

Test the hypothesis that the blood type is the same for the three samples.

Exercise 2:

One random sample of 50 students was selected from a private school and another sample of 100 students is selected from a public school. The aim is to see whether there is any difference in intelligence among the students studying in two different types of schools. They are given standardized achievement tests. The results are as:

Scores

	0 –100	100 - 200	200 - 300	300 - 400	total
Private school	6	18	20	6	50
Public school	30	45	20	5	100
Total	36	63	40	11	150

Test whether the distribution of the scores in private and public school is the same? Use 5% level of significance.

Exercise 3:

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 play ground environment, the children during free play the researcher recorded the number of children for each group who exhibited disruptive behavior. The data are summarized in the two-way table. Analyze the data given in Table and interpret the results.

BEHAVIOUR

	Disruptive Behavior	**non-disruptive behavior**	**total**
with development delay	12	3	15
without development delay	5	10	15
total	17	13	30

Exercise 4:

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).

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?

Exercise 5:

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 centers, 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:

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.

Exercise 6:

Comparison of Bell and Kato-Katz methods for detecting Schistosoma mansoni eggs in faces. The same 315 specimens were examined using each method.

		Kato-Katz		Total
		+	-	
Bell	+	184	54	238
	-	14	63	77
Total		198	117	315

Is there any difference in the abilities of the methods to detect Schistosoma mansoni eggs in faces?

Exercise 7:

Data regarding incidence of tumors in the two hemispheres for three sites in the cortex is available as:

Site in Cortex	**Site of tumor**	**Benign tumors**	**Malignant tumors**
1	Left hemisphere	17	5
	Right hemisphere	6	5
2	Left hemisphere	12	3
	Right hemisphere	7	5
3	Left hemisphere	11	3
	Right hemisphere	11	9

Can we say that there is association between type of tumor and among hemispheres?

DAY 4

March 25, 2011

Session I : Sample Size Calculation in Health Research + Application on PASS

Resource Person: Mr. Waqas Sami

Time: 9:00 a.m. – 01:00 p.m.

SAMPLE SIZE CALCULATION IN HEALTH RESEARCH:

"How many participants will I need?"

Issues related to sample size calculations in Health Research

SAMPLE SIZE DETERMINATION METHODS:

- ❖ Hypothesis Testing
- ❖ Confidence Interval or Level of Precision
- ❖ Study Designs

SAMPLE SIZE DETERMINATION INFORMATION:

- ❖ Objectives of study
- ❖ Variables of interest
  - Type of data e.g. qualitative, quantitative
- ❖ Desired significance level
- ❖ Desired power
- ❖ Effect/difference of clinical importance
- ❖ Standard deviations
- ❖ One or two-sided tests

DESIRED LEVEL OF SIGNIFICANCE:

Level of significance is the probability of rejecting null hypothesis Ho, it is denoted by α. The most frequently used values of α are 0.05, and 0.01 i.e. 5% and 1%, by 5% we mean that there are about 5 chances in 100 of incorrectly rejecting the null hypothesis or we can say that we are 95% confident in making the correct decision.

ALSO CALLED TYPE I ERROR

POWER OF STUDY:

The power (1 - β) is define as the probability of rejecting False Hypothesis i.e. either Null or Alternate.

ALSO CALLED TYPE II ERROR

EFFECT / DIFFERENCE OF CLINICAL IMPORTANCE:

This parameter is the measured difference between comparison groups that the investigator would like to detect.

Example:

Suppose a study is designed to compare a standard diagnostic procedure of 80% accuracy with a new procedure of unknown but potentially higher accuracy. Suppose that the investigator believes that it would be a clinically important improvement if the new procedure were 90% accurate. Therefore, the investigator would choose a minimum expected difference of 10% (0.10).

FROM WHERE TO GET THE EXPECTED DIFFERENCE VALUES:

- The setting of this parameter is subjective and is based on clinical judgment and experience with the problem being investigated.

- The results of literature review or a pilot study can also guide the selection of a minimum difference.

MEASUREMENT OF VARIABILITY:

- This parameter is represented by the expected SD in the measurements made within each comparison group.

- A review of the literature can provide estimates of this parameter.

- If preliminary data are not available, this parameter may have to be estimated on the basis of subjective experience, or a range of values may be assumed.

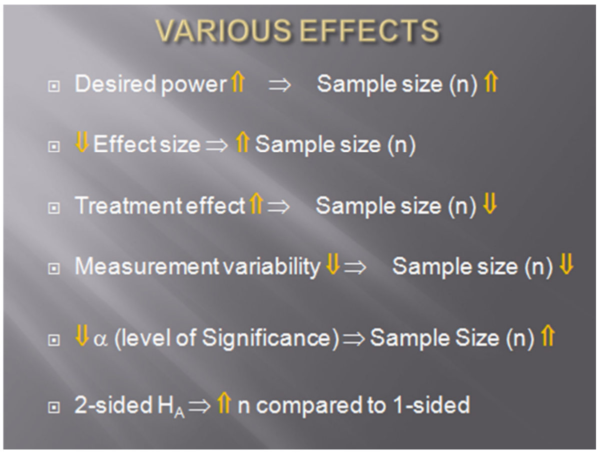
ONE TAILED AND TWO TAILED TESTS:

In a few cases, it may be known before the study that any difference between comparison groups is possible in only one direction.

In such cases, use of a one-tailed statistical analysis, which would require a smaller sample size for detection of the minimum difference than would a two-tailed analysis, may be considered.

Because of this simple relationship and truly the one-tailed analyses are rare, a two-tailed analysis is assumed in the remainder of this presentation.

VARIOUS EFFECTS:



VARIOUS EFFECTS

- Desired power ⇑ ⇒ Sample size (n) ⇑
- ⇓ Effect size ⇒ ⇑ Sample size (n)
- Treatment effect ⇑ ⇒ Sample size (n) ⇓
- Measurement variability ⇓ ⇒ Sample size (n) ⇓
- ⇓ α (level of Significance) ⇒ Sample Size (n) ⇑
- 2-sided H_A ⇒ ⇑ n compared to 1-sided

HYPOTHESIS TESTING:

- One Sample Mean
- Two Sample Mean
- Paired Sample Mean
- One Sample Proportion
- Two Sample Proportion
- Correlations
- Three (03) or more means Comparison

ONE SAMPLE MEAN:

We want to know whether birth weights of full-term infants who ultimately died of Sudden Infant Death Syndrome (SIDS) is significantly different from that of other full term births. A Sample of n=10 SIDS cases demonstrated that the Average birth weight of 10 cases was 2890.5 grams with a Standard Deviation = 720.0 grams. The average Birth Weight of other full term births was 3300 grams.

How large a sample is needed to test the SIDS data with 90% power at α = 0.05 (two sided). We want to detect the mean difference in birth weight of 300 grams. The standard deviation is 720.0.

SOLUTION:

- Desired Power = 90% (1.28)
- Desired Level of Significance =(1.96)
- Mean Difference = 300
- Standard Deviation = 720.0

$$n = \frac{\sigma 2(Z_{1-\alpha/2} + Z_{1-\beta})2}{(\mu_1 - \mu_2)2} = 60.46 \Rightarrow 61$$

TWO SAMPLE MEAN:

Situations:

- When common Variance or Standard Deviation is given for both groups.

$$n_1 = \frac{2\sigma^2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{\text{difference}^2}$$

- When separate Variance or Standard Deviation is given for both groups.

$$n_1 = \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2(\sigma_1 2 + \sigma_2 2)}{(\mu 1 - \mu 2)^2}$$

EXAMPLE – SITUATION 1:

A study tested the cholesterol-lowering potential of dietary linoleic acid in mildly hypercholesterolemia subjects and healthy subjects. Values are plasma cholesterol levels (mmol/L) in two independent groups.

Suppose that you are looking for a:

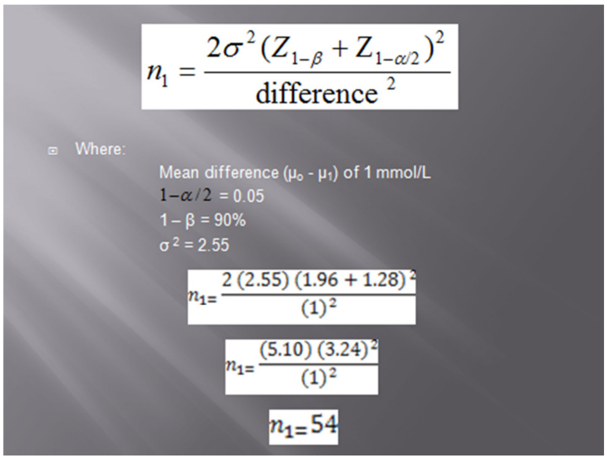
Mean difference (μ_o - μ_1) of 1 mmol/L

$\alpha = 0.05$

$1 – \beta = 90\%$

$\sigma^2 = 2.55$

How many individuals must be studied to achieve these conditions:



$$n_1 = \frac{2\sigma^2 (Z_{1-\beta} + Z_{1-\alpha/2})^2}{\text{difference}^2}$$

⊡ Where:

Mean difference (μ_o - μ_1) of 1 mmol/L

$1-\alpha/2 = 0.05$

$1 – \beta = 90\%$

$\sigma^2 = 2.55$

$$n_1 = \frac{2\,(2.55)\,(1.96 + 1.28)^2}{(1)^2}$$

$$n_1 = \frac{(5.10)\,(3.24)^2}{(1)^2}$$

$$n_1 = 54$$

EXAMPLE – SITUATION 2:

A clinical dietician wants to compare two different diets, A and B, for diabetic patients.

She hypothesizes that diet A (Group 1) will be better than diet B (Group 2), in terms of lowering blood glucose. She plans to get a random sample of diabetic patients and randomly assign them to one of the two diets. At the end of the experiment, which lasts 6 weeks, a fasting blood glucose test was conducted on each patient.

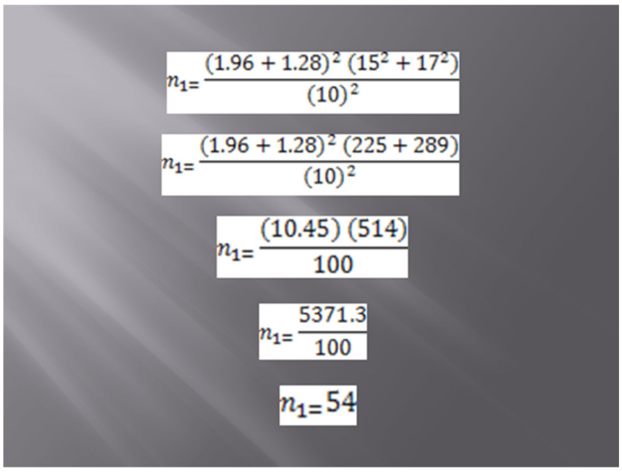
She also expects that the average difference in blood glucose measure between the two groups will be about 10 mg/dl. Furthermore, she also assumes the standard deviation of blood glucose distribution for diet A to be 15 and the standard deviation for diet B to be 17.

The dietician wants to know the number of subjects needed in each group.

SOLUTION:

Desired Power = 90%
Desired Level of Significance = 0.05
Mean Difference = 10 mg/dl
Standard Deviation of Group A = 15
Standard Deviation of Group B = 17

$$n_1 = \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2 (\sigma_1 2 + \sigma_2 2)}{(\mu 1 - \mu 2)^2}$$



$$n_1 = \frac{(1.96 + 1.28)^2\,(15^2 + 17^2)}{(10)^2}$$

$$n_1 = \frac{(1.96 + 1.28)^2\,(225 + 289)}{(10)^2}$$

$$n_1 = \frac{(10.45)\,(514)}{100}$$

$$n_1 = \frac{5371.3}{100}$$

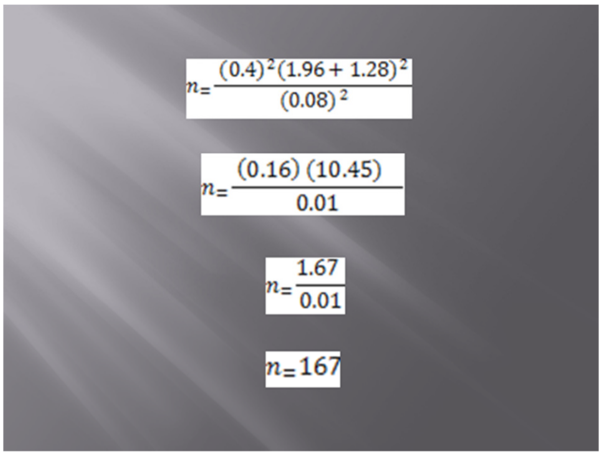
$$n_1 = 54$$

PAIRED SAMPLE MEAN:

A researcher sought to learn whether oat bran cereal lowered low – density lipoprotein (LDL) cholesterol in hypercholesterolemia men. How large the sample is required if he wants to detect a mean change of 0.8 mmol /L with a Standard Deviation of 0.4 mmo/L using level of significance = 0.05 and power = 90%.

SOLUTION:

$$n = \frac{\sigma_d{}^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$



$$n = \frac{(0.4)^2 (1.96 + 1.28)^2}{(0.08)^2}$$

$$n = \frac{(0.16)\,(10.45)}{0.01}$$

$$n = \frac{1.67}{0.01}$$

$$n = 167$$

ONE SAMPLE PROPORTION:

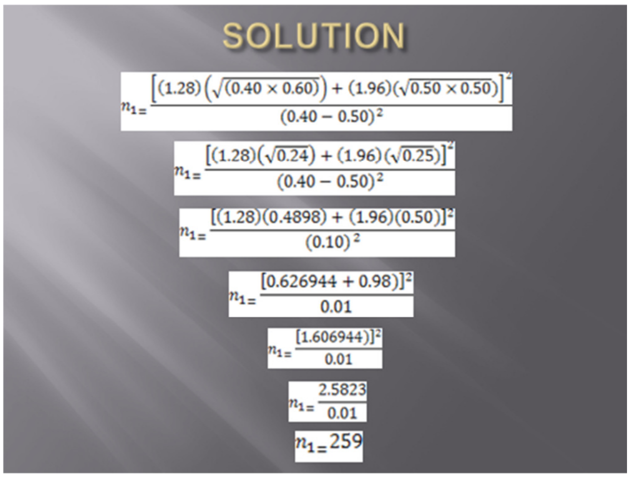
The five year cure rate for a particular cancer is reported in the literature to be 50%. An investigator wishes to test the hypothesis that this cure rate can be applied in a certain local health district. What is the required sample size if the investigator is interested in detecting a true rate of 40%. The level of significance is set at 5% with 90% power of study.

SOLUTION:

- ❖ Test cure rate (p_0) = 50%
- ❖ Anticipated C (p_1) = 40%
- ❖ Level of significance = 5%
- ❖ Power of study = 90%
- ❖ Alternate hypothesis is two sided

FORMULA:

$$n = \frac{\left(Z_{1-\beta}\left[\sqrt{p_1(1-p_1)}\right] + Z_{1-\alpha/2}\left[\sqrt{p_o(1-p_o)}\right]\right)2}{(p_1 - p_o)2}$$



SOLUTION

$$n_1 = \frac{\left[(1.28)\left(\sqrt{(0.40 \times 0.60)}\right) + (1.96)(\sqrt{0.50 \times 0.50})\right]^2}{(0.40 - 0.50)^2}$$

$$n_1 = \frac{[(1.28)(\sqrt{0.24}) + (1.96)(\sqrt{0.25})]^2}{(0.40 - 0.50)^2}$$

$$n_1 = \frac{[(1.28)(0.4898) + (1.96)(0.50)]^2}{(0.10)^2}$$

$$n_1 = \frac{[0.626944 + 0.98)]^2}{0.01}$$

$$n_1 = \frac{[1.606944)]^2}{0.01}$$

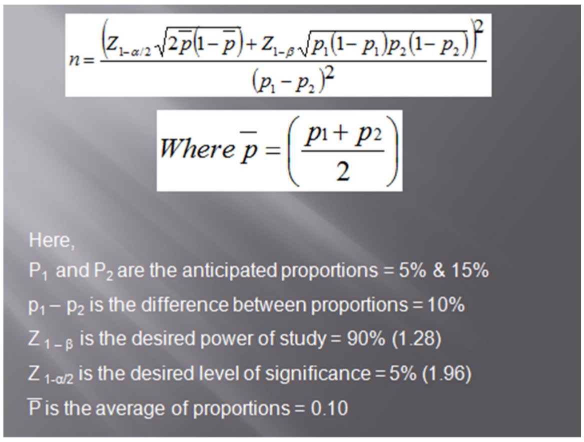
$$n_1 = \frac{2.5823}{0.01}$$

$$n_1 = 259$$

TWO SAMPLE PROPORTION:

It is believed that the proportion of patients who developed complications after undergoing one type of surgery is 5% while the proportion of patients who developed complications after second type of surgery is 15%.

How large should be the sample size in each of the two groups of patients if the investigator wishes to detect with a power of 90% at 5% level of significance.



$$n = \frac{\left(Z_{1-\alpha/2}\sqrt{2\overline{p}\left(1-\overline{p}\right)} + Z_{1-\beta}\sqrt{p_1(1-p_1)p_2(1-p_2)}\right)^2}{(p_1 - p_2)^2}$$

$$Where\ \overline{p} = \left(\frac{p_1 + p_2}{2}\right)$$

Here,

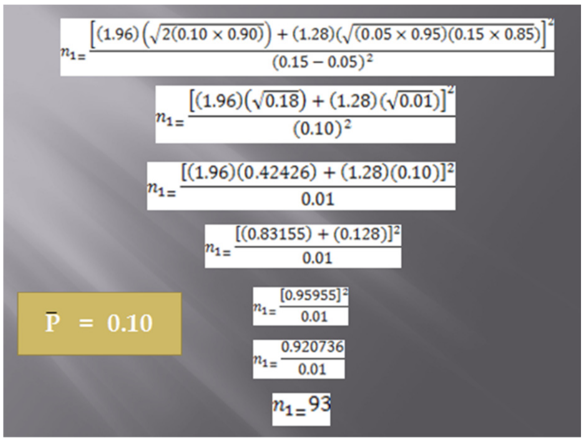
P_1 and P_2 are the anticipated proportions = 5% & 15%

$p_1 - p_2$ is the difference between proportions = 10%

$Z_{1-\beta}$ is the desired power of study = 90% (1.28)

$Z_{1-\alpha/2}$ is the desired level of significance = 5% (1.96)

$\overline{P}$ is the average of proportions = 0.10



$$n_1 = \frac{\left[(1.96)\left(\sqrt{2(0.10 \times 0.90)}\right) + (1.28)(\sqrt{(0.05 \times 0.95)(0.15 \times 0.85)}\right]^2}{(0.15 - 0.05)^2}$$

$$n_1 = \frac{\left[(1.96)(\sqrt{0.18}) + (1.28)(\sqrt{0.01})\right]^2}{(0.10)^2}$$

$$n_1 = \frac{[(1.96)(0.42426) + (1.28)(0.10)]^2}{0.01}$$

$$n_1 = \frac{[(0.83155) + (0.128)]^2}{0.01}$$

$\overline{P} = 0.10$

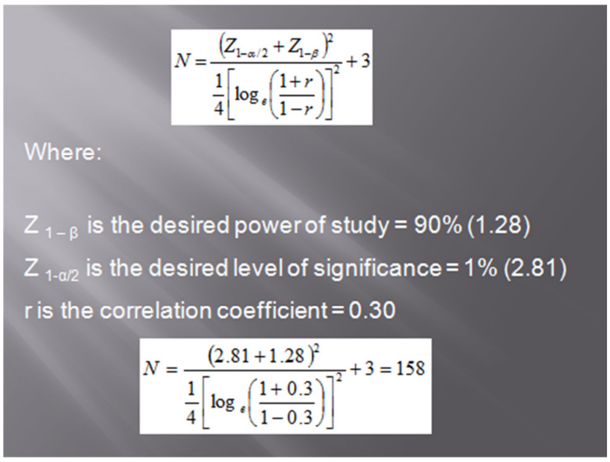
$$n_1 = \frac{[0.95955]^2}{0.01}$$

$$n_1 = \frac{0.920736}{0.01}$$

$$n_1 = 93$$

CORRELATION:

According to the literature, the correlation between salt intake and systolic blood pressure is around 0.3. A study is conducted to test the correlation in a population, with the significance level of 1% and power of 90%. How many participants should be there in the study.



$$N = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\frac{1}{4}\left[\log_e\left(\frac{1+r}{1-r}\right)\right]^2} + 3$$

Where:

$Z_{1-\beta}$ is the desired power of study = 90% (1.28)

$Z_{1-\alpha/2}$ is the desired level of significance = 1% (2.81)

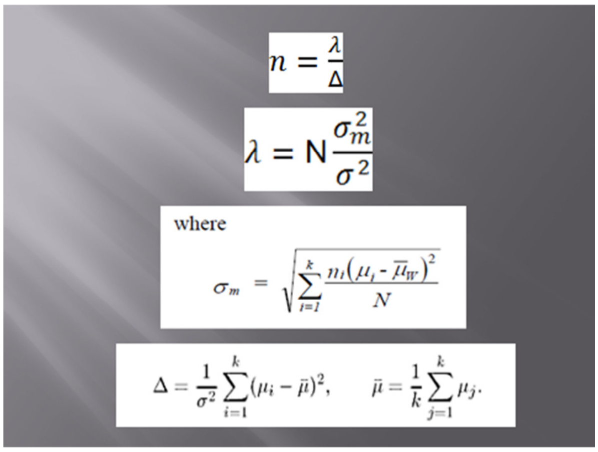
r is the correlation coefficient = 0.30

$$N = \frac{(2.81 + 1.28)^2}{\frac{1}{4}\left[\log_e\left(\frac{1+0.3}{1-0.3}\right)\right]^2} + 3 = 158$$

COMPARISON OF 03 OR MORE SAMPLE MEANS:

EXAMPLE – 04 GROUPS MEANS:

A four-arm (k=4) parallel group double blind randomized clinical trial is to be conducted to compare 04 treatments. The comparison is made at a significance level of α = 0.05. Assume that the standard deviation within each group is σ = 3.5 and the true mean responses for the four treatment groups are: $\mu_1 = 8.25$, $\mu_2 = 11.75$, $\mu_3 = 12.00$ and $\mu_4 = 13.00$. Calculate the sample size at 80% power of study.



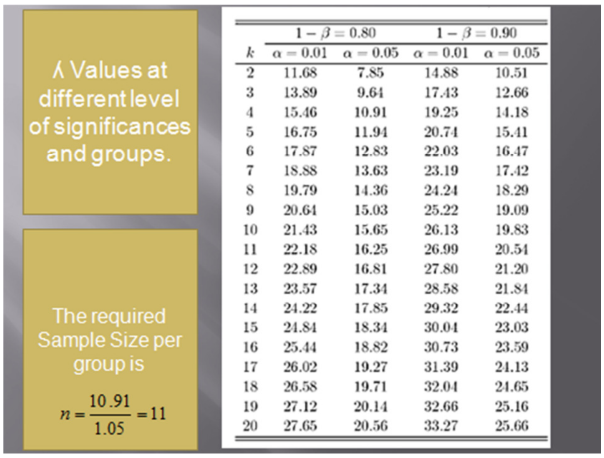
$$n = \frac{\lambda}{\Delta}$$

$$\lambda = \mathrm{N}\frac{\sigma_m^2}{\sigma^2}$$

where

$$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{n_i\left(\mu_i - \overline{\mu}_W\right)^2}{N}}$$

$$\Delta = \frac{1}{\sigma^2}\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2, \qquad \bar{\mu} = \frac{1}{k}\sum_{j=1}^{k}\mu_j.$$



ʎ Values at different level of significances and groups.

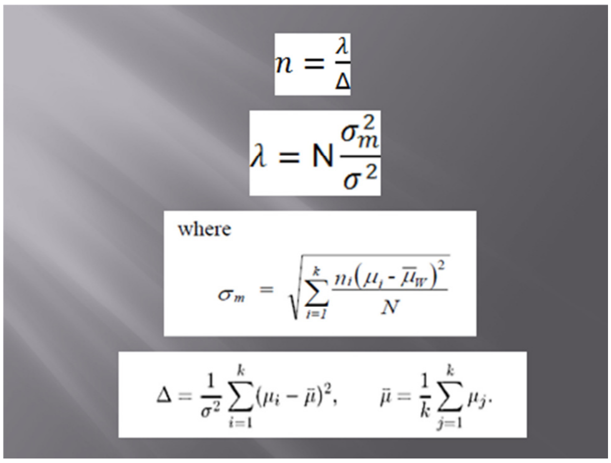
The required Sample Size per group is

$n = \frac{10.91}{1.05} = 11$

	$1-\beta=0.80$		$1-\beta=0.90$	
k	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$
2	11.68	7.85	14.88	10.51
3	13.89	9.64	17.43	12.66
4	15.46	10.91	19.25	14.18
5	16.75	11.94	20.74	15.41
6	17.87	12.83	22.03	16.47
7	18.88	13.63	23.19	17.42
8	19.79	14.36	24.24	18.29
9	20.64	15.03	25.22	19.09
10	21.43	15.65	26.13	19.83
11	22.18	16.25	26.99	20.54
12	22.89	16.81	27.80	21.20
13	23.57	17.34	28.58	21.84
14	24.22	17.85	29.32	22.44
15	24.84	18.34	30.04	23.03
16	25.44	18.82	30.73	23.59
17	26.02	19.27	31.39	24.13
18	26.58	19.71	32.04	24.65
19	27.12	20.14	32.66	25.16
20	27.65	20.56	33.27	25.66

EXAMPLE – 05 GROUP MEANS:

A completely randomized one way experiment is to be run with k= 5 different treatments. The process standard deviation is known to be σ = 13.5. If we want 90% probability of detecting a difference of δ = 10 between the treatments what is the required n?. Use a significance level of 0.05.



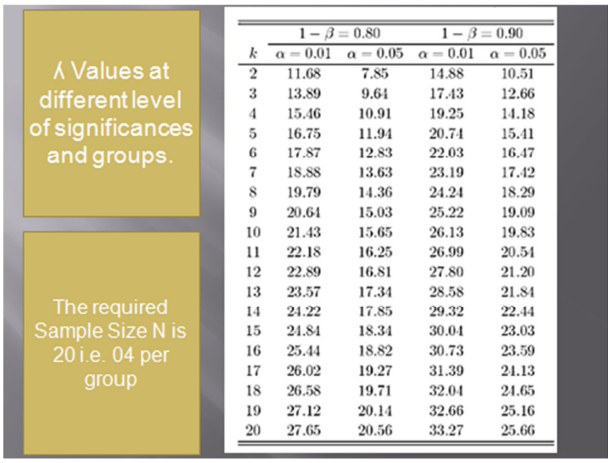
$$n = \frac{\lambda}{\Delta}$$

$$\lambda = \mathrm{N}\frac{\sigma_m^2}{\sigma^2}$$

where

$$\sigma_m = \sqrt{\sum_{i=1}^{k} \frac{n_i(\mu_i - \bar{\mu}_w)^2}{N}}$$

$$\Delta = \frac{1}{\sigma^2}\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2, \qquad \bar{\mu} = \frac{1}{k}\sum_{j=1}^{k}\mu_j.$$



ʎ Values at different level of significances and groups.

The required Sample Size N is 20 i.e. 04 per group

	$1-\beta=0.80$		$1-\beta=0.90$	
k	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$
2	11.68	7.85	14.88	10.51
3	13.89	9.64	17.43	12.66
4	15.46	10.91	19.25	14.18
5	16.75	11.94	20.74	15.41
6	17.87	12.83	22.03	16.47
7	18.88	13.63	23.19	17.42
8	19.79	14.36	24.24	18.29
9	20.64	15.03	25.22	19.09
10	21.43	15.65	26.13	19.83
11	22.18	16.25	26.99	20.54
12	22.89	16.81	27.80	21.20
13	23.57	17.34	28.58	21.84
14	24.22	17.85	29.32	22.44
15	24.84	18.34	30.04	23.03
16	25.44	18.82	30.73	23.59
17	26.02	19.27	31.39	24.13
18	26.58	19.71	32.04	24.65
19	27.12	20.14	32.66	25.16
20	27.65	20.56	33.27	25.66

CONFIDENCE INTERVAL:

- One Sample Proportion
- Two Sample Proportion

There is only one method of determining sample size that allows the researcher to PREDETERMINE the accuracy of the sample results “The Confidence Interval Method of Determining Sample Size”.

Requirement for Calculation of Sample Size is;

- Anticipated population proportion.
- Confidence Level.
- Absolute or Relative Precision

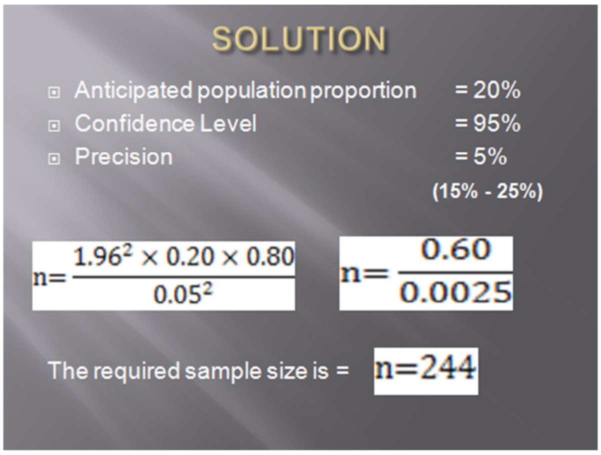
DESCRIPTION:

- Anticipated population proportion:
  - Is usually reported in the form of prevalence, incidence or rate etc.
- Confidence Level:
  - (less error = more accuracy) Conventional is + 5%
- Precision
  - Is usually reported in percentage points or percentage.

ONE SAMPLE PROPORTION:

A local health department wishes to estimate the prevalence rate of tuberculosis infection among children under five years of age in its locality. How many children should be included in the sample so that the rate may be estimated to within 5 percentage points of the true value with 95% confidence if its is known that the true rate is unlikely to exceed 20%.

$$n = \frac{z^2(pq)}{e^2}$$



SOLUTION

- Anticipated population proportion = 20%
- Confidence Level = 95%
- Precision = 5%

(15% - 25%)

$$n=\frac{1.96^2 \times 0.20 \times 0.80}{0.05^2} \qquad n=\frac{0.60}{0.0025}$$

The required sample size is = $n=244$

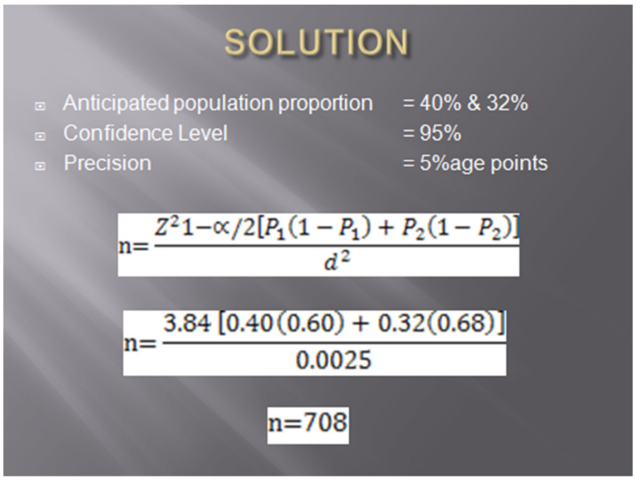
TIPS:

- ❖ For situations in which no anticipation is possible a figure of 0.5 should be used.
- ❖ If the anticipation is given in range the closest to 0.5 should be used.
- ❖ The sample size required will be largest when P is equal to 0.5.

TWO SAMPLE PROPORTION:

Suppose that in a pilot study of 50 agricultural workers in an irrigation project. It was observed that 40% had active schistosomiasis. A similar pilot study of 50 agricultural workers outside the project demonstrated that 32% had active schistosomiasis infection. If we would like to carry out a larger study to estimate the true schistosomiasis risk difference to within 5 percentage points of the true value of the confidence interval, how many people must be studied in each of the two groups.

$$n=\frac{Z^2 1{-}\propto/2[P_1(1-P_1)+P_2(1-P_2)]}{d^2}$$



SOLUTION

- Anticipated population proportion = 40% & 32%
- Confidence Level = 95%
- Precision = 5%age points

$$n=\frac{Z^2 1{-}\propto/2[P_1(1-P_1)+P_2(1-P_2)]}{d^2}$$

$$n=\frac{3.84\,[0.40(0.60)+0.32(0.68)]}{0.0025}$$

$$n=708$$

STUDY DESIGNS:

Case – Control Study

EXAMPLE:

Suppose a researcher wishes to study the effect of passive smoking in a group of BC cases compared to a group of controls. Available information indicates that 30% of the controls are exposed to passive smoking and the researcher wishes to have an 80% chance of detecting an odds ratio of 2. How large a sample should be included in each of the group.

$$\frac{\left[Z_{(1-\alpha/2)}\sqrt{2\bar{P}(1-\bar{P})}+Z_{(1-\beta)}\sqrt{P_1(1-P_1)+P_2(1-P_2)}\right]^2}{(P_1-P_2)^2}$$

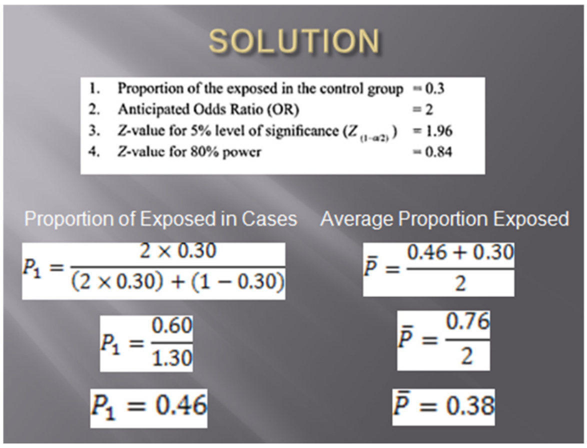
where,

$$P_1=\frac{(OR)P_2}{(OR)P_2+(1-P_2)}; \quad \bar{P}=\frac{P_1+P_2}{2}$$

P_2 = proportion of exposed among controls

P_1 = proportion of exposed among cases

OR = anticipated odds ratio



SOLUTION

1. Proportion of the exposed in the control group = 0.3
2. Anticipated Odds Ratio (OR) = 2
3. *Z*-value for 5% level of significance ($Z_{(1-\alpha/2)}$) = 1.96
4. *Z*-value for 80% power = 0.84

Proportion of Exposed in Cases

$$P_1 = \frac{2 \times 0.30}{(2 \times 0.30) + (1 - 0.30)}$$

$$P_1 = \frac{0.60}{1.30}$$

$$P_1 = 0.46$$

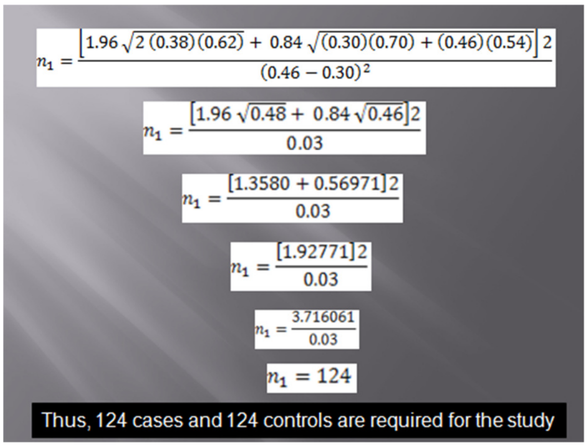
Average Proportion Exposed

$$\bar{P} = \frac{0.46 + 0.30}{2}$$

$$\bar{P} = \frac{0.76}{2}$$

$$\bar{P} = 0.38$$

CALCULATION:



$$n_1 = \frac{\left[1.96\sqrt{2\,(0.38)(0.62)} + 0.84\sqrt{(0.30)(0.70) + (0.46)(0.54)}\right]2}{(0.46-0.30)^2}$$

$$n_1 = \frac{\left[1.96\sqrt{0.48} + 0.84\sqrt{0.46}\right]2}{0.03}$$

$$n_1 = \frac{[1.3580 + 0.56971]2}{0.03}$$

$$n_1 = \frac{[1.92771]2}{0.03}$$

$$n_1 = \frac{3.716061}{0.03}$$

$$n_1 = 124$$

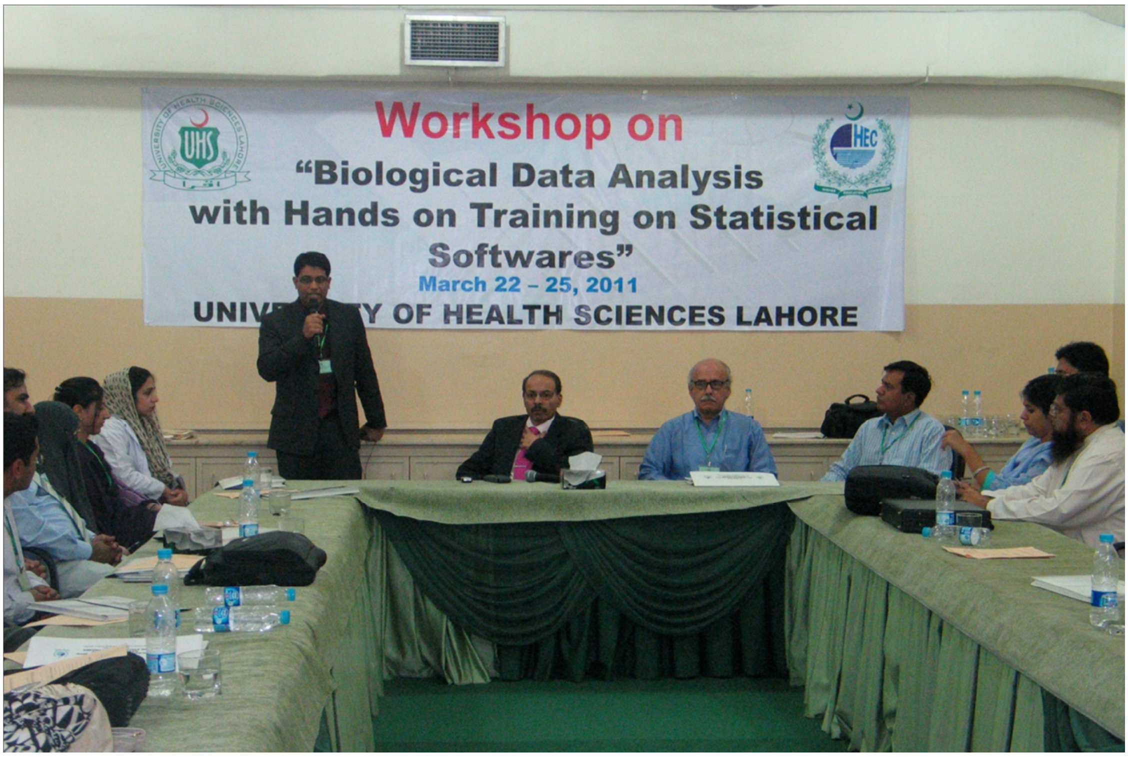
Thus, 124 cases and 124 controls are required for the study

Lecture on Sample Size Calculation in Health Research

Workshop on
"Biological Data Analysis
with Hands on Training on Statistical
Softwares"
March 22 – 25, 2011
UNIVERSITY OF HEALTH SCIENCES LAHORE

Question and Answers Session

CLOSING CEREMONY & CERTIFICATE DISTRIBUTION


UHS
Workshop on
“Biological Data Analysis
with Hands on Training on Statistical
Softwares”
March 22 – 25, 2011
UNIV Y OF HEALTH SCIENCES LAHORE
HEC

Closing Ceremony

Certificate Distribution

RESULT OF PRE – POST ASSESSMENT TEST

Name of Candidate	Institution	Pre-assessment marks	Post-assessment marks	Total Marks
Mr. Shahzad Khalil	PGMI	08	13	**20**
Dr. Abdullah Tariq	PGMI	10	15	**20**
Dr. Fozia Farzana	PGMI	09	15	**20**
Dr. Javeria Malik	PGMI	06	15	**20**
Dr. Balqis Akhtar	PGMI	10	14	**20**
Dr. Qamar Ishfaq Ahmad	PGMI	12	16	**20**
Dr. Ali Raza	PGMI	13	16	**20**
Dr. Abdul Hye	PGMI	05	17	**20**
Dr. Faiqa Arshad	PGMI	06	14	**20**
Dr. Hifza Noor Lodhi	PGMI	13	18	**20**
Dr. Zubaida Afshan	PGMI	06	11	**20**
Dr. Faiza Oje Bhatti	PGMI	09	17	**20**
Dr. Faiza Rafiq	PGMI	15	19	**20**
Dr. Rizwan Faisal	PGMI	07	13	**20**
Dr. Tasheen Ikram	PGMI	15	19	**20**
Dr. Shahid Mehmood	IPH	07	16	**20**
Dr. Taskeen Zahra Sajid	IPH	14	16	**20**
Dr. Anjum Razzaq	IPH	14	17	**20**
Dr. Syed Razi Haider	IPH	13	19	**20**
Dr. Syed Naqeeb Hussain	IPH	10	12	**20**
Dr. Romeeza Tahir	UHS	13	18	**20**
Dr. Sumbla Ghaznavi	UHS	10	16	**20**
Dr. Zafar Iqbal	UHS	10	14	**20**
Mr. Ikram Ullah	UHS	11	16	**20**
Dr. Asima karim	UHS	17	18	**20**
Dr. M. Nadir Iqbal	UHS	18	19	**20**
Dr. Ammad Hussain	UHS	09	17	**20**
Dr. Nadia Naseem	UHS	09	16	**20**
Dr. Kanwal Ashal Pal	UHS	12	17	**20**
Dr. Sumair Anwar	UHS	19	19	**20**

PGMI – Postgraduate Medical Institute
IPH – Institute of Public Health
UHS – University of Health Sciences Lahore

FEEDBACK OF WORKSHOP

	Strongly Agree n (%)	Somewhat Agree n (%)	Disagree n (%)	Total
The material was well organized	28 (93.3)	2 (6.7)	0 (0.0)	**30**
The ideas and skills demonstrated were useful	20 (66.7)	10 (33.3)	0 (0.0)	**30**
The information was practical	19 (63.3)	9 (30.0)	2 (6.7)	**30**
The presentations met the workshop objectives	20 (66.7)	10 (33.3)	0 (0.0)	**30**
The workshop addressed to my interests successfully	13 (43.3)	15 (50%)	2 (6.7)	**30**
The examples presented were relevant	25 (83.3)	5 (16.7)	0 (0.0)	**30**
The pace of the presentations were comfortable	19 (63.3)	8 (26.7)	3 (10.0)	**30**
I would recommend such workshops	28 (93.3)	2 (6.7)	2 (6.7)	**30**
The trainers grasped my interest	20 (66.7)	10 (33.3)	0 (0.0)	**30**
The trainers had a professional approach	28 (93.3)	2 (6.7)	0 (0.0)	**30**
The trainers stayed focused on the topic	25 (83.3)	5 (16.7)	0 (0.0)	**30**
The trainers effectively responded to the questions	24 (80.0)	6 (13.3)	0 (0.0)	**30**
The trainers used relevant examples	26 (86.7)	4 (13.3)	0 (0.0)	**30**
The trainers solicited the audience interaction	19 (63.3)	9 (30.0)	2 (6.7)	**30**
Overall, I found the trainers as experts	28 (93.3)	2 (6.7)	0 (0.0)	**30**
Would you like to attend similar workshops in future	30 (100.0)	0 (0.0)	0 (0.0)	**30**
Would you like to attend other workshops conducted by the same trainers	30 (100.0)	0 (0.0)	0 (0.0)	**30**

GROUP PHOTOGRAPH

Workshop on
"Biological Data Analysis
with Hands on Training on Statistical
Softwares"
March 22 - 25, 2011
UNIVERSITY OF HEALTH SCIENCES LAHORE