INTRODUCTION

PREFACE

What is Statistics?

Many if not most, of life's really important question involve partial information. The process of drawing good, reliable conclusions from partial information is central to the study of Statistics. Statistics is described as the study of how to collect, organize, analyze, and interpret information, especially numerical information. A definite advantage of statistical methods is that they can help us make decisions without prejudice. Moreover, Statistics can be used fort making decisions when we are faced with uncertainties. For instance, If we wish to estimate the proportion of people who will have a severe reaction to a knew flu shot without giving the shot to everyone who wants it, Statistics can provide appropriate methods. Before going any further let us define two important terms in statistical studies, namely Population and Sample.

POPULATION: this term refers to all measurements or observations of interest. For examples, if there are 600 students in the school that are classified according to blood type, we say we have a population of size 600. The numbers on the cards in a deck, the heights of residents in a certain city, and the lengths of fish in a particular lake, are all examples of populations. Whereas the term **SAMPLE:** is simply a part of the population. But not every sample is useful; the sample must represent the population.

Statistics has two major types that are Descriptive Statistics and Inferential Statistics,

Descriptive Statistics: consists of all methods for organizing and summarizing information, while **Inferential Statistics:** consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample from the population.

Observational and Experimental Studies

In many cases the purpose of a statistical study is to investigate whether a relationship exists between two variables or characteristics, such as smoking and lung cancer, height and weight, or educational attainment and annual income. For such studies it is important to distinguish between two types of procedures: observational studies and

experimental studies. In an *observational study*, researchers simply observe, in an experimental, researchers design and control. Observational studies can reveal only association, whereas *experimental* studies can establish causation (cause and effect).

Variables and Data

1. Variables

A characteristic that varies from one person or thing to another is called a variable. A non-numerical-valued variable is called qualitative variable and any numerically valued variable is called quantitative variable. Quantitative variables also can be classified as discrete or continuous. A discrete variable is one whose possible values form a finite (or countable infinite) set of numbers. A continuous variable is a variable whose possible values form some interval of numbers.

2. Data

Observing the values of a variable for one or more people or things yields data. Thus the information collected, organized, and analyzed by statisticians are data so Data can be defined as information obtained by observing values of a variable. Qualitative data, Quantitative data, Discrete data and Continuous data are the data obtained by observing the values of Qualitative, Quantitative, Discrete, and continuous variables, respectively.

Example 1. Humans are classified as having one of four blood types: A, B, AB ro O. What kind of data do you receive when you are told your blood type?
Solution: Your blood type is Qualitative data obtained by observing your value of the variable "Blood Type".

Example 2. The US bureau of the census collects data on household size and publishes the information in current population reports. What kind of data is the number of people in your household?
Solution: The number of people in your household is discrete, quantitative data; data obtained by observing the value of the variable "Household Size" for your household.

Example 3. In *Information Please Almanac* lists the world's highest waterfalls. The list shows that Angel falls in Venezuela is 3281 feet high, more than twice as high as Ribbon falls in Yosemite, California, which is 1612 feet high. What kind of data are these heights?

Solution: The waterfall heights are continuous, quantitative data; data obtained by observing the values of the variable "Height" for the two waterfalls.

Levels of Measurements

When we collect data, it is common to classify the information obtained according to one of the following four levels of measurement:

Nominal Level: Data at this level of measurement consist of names only or qualities, with no implied criteria by which the data can be identified as greater than or less than other data items. The Faculty of the student at Yarmouk University, Nationality of a person and Blood type are all examples of the Nominal level.

Ordinal Level: Data at this level may be arranged in some order, but actual difference between data values either cannot be determined or are meaningless. Faculty member rank (Prof., Assoc., Assis, … etc.), military rank and student rating at graduation (Excellent, V.Good, …etc) are all examples of the Ordinal level.

Interval Level: It is like the Ordinal level, but it has the additional property that meaningful differences between data values can be computed. Interval-level data may not have an intrinsic zero or starting point. Consequently, differences are meaningful but ratios of data are not. Body temperatures and Intelligence quotient are examples of such a level.

Ratio Level: Is similar to the interval level, but it includes an inherent zero as a starting point for all measurements. Consequently, at this level both differences and ratios are meaninggul. Age, Weight and Height of a person are all examples of the Ratio level.

Important Notes:

1. You should bring a 3.5-inch floppy disk, submit it to Mr. Baroud to download the projects necessary to your work through the semester, and bring it in each session. ALWAYS keep it clean and in excellent condition.
2. You should bring the booklet of STAT 105 in each session, the booklet is available ONLY at Rawa'ah bookshop opposite to the main gate of the university.
3. No Homework is valid to be submitted BEYOND the deadline.
4. No makeup exams under any circumstances.
5. The midterm exam will be in the 9^{th} week from the beginning of the semester.

6. The final exam will be in the last week of the semester just before the final exams.
7. The whole mark will be divided as follows:
   40% for HW's
   25% for midterm exam
   35% for final exam
8. All the exams will be theoretical and practical.
9. In the beginning of each session you will sit for a short practical QUIZ about the material to be explained, that is you should prepare your self well for the quiz in advance.
10. Write your NAME, SECTION NUMBER and SERIAL NUMBER on your disk and on each of the student work sheets in advance.

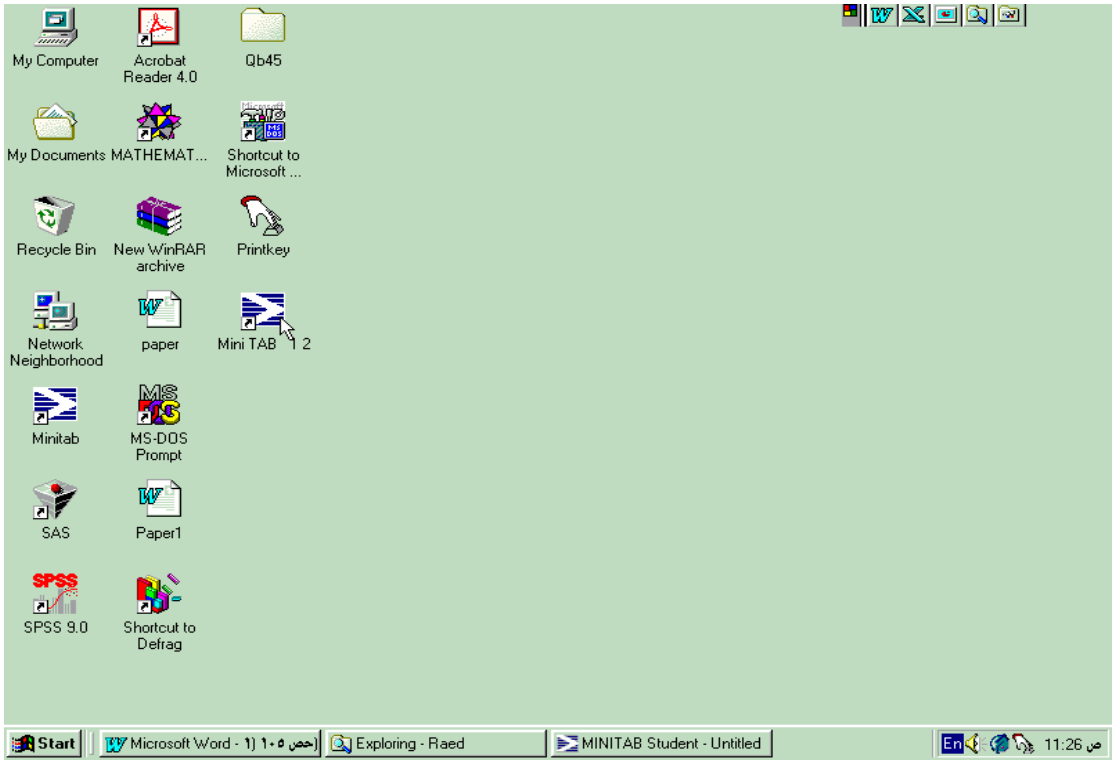
INTRODUCTION TO MINITAB

The most commonly used programs for statistical work are obtained from statistical software packages, collections of statistical computer programs written by some organization or individual. We have chosen Minitab to illustrate the basic ideas of statistical software. You can find the Minitab package is already downloaded on the computer system right in front of you.

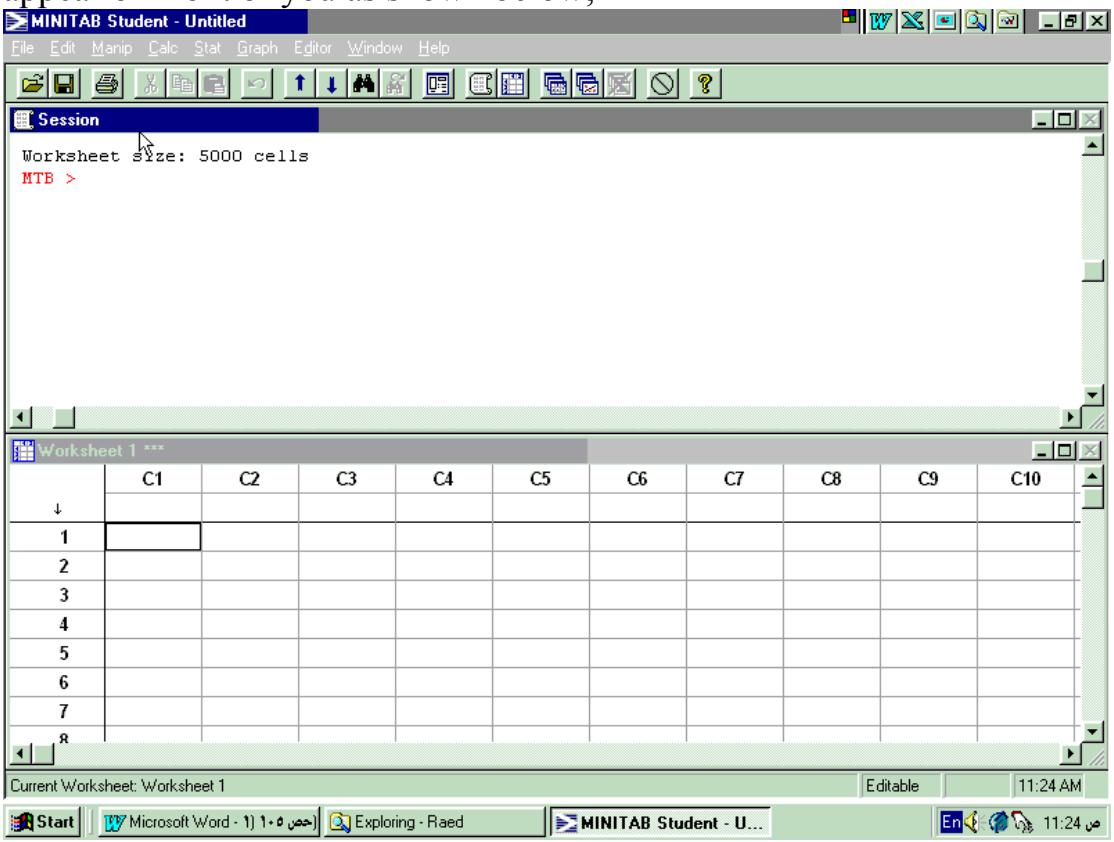
Traditionally, Minitab has been applied by typing commands on the `MTB>` prompt, so-called **session commands** which are available on all platforms. Minitab has recently introduced as menu interface as well, so that commands can also be executed by choosing them from menus and completing dialogue boxes, which are called **menu commands** that may not be available on all platforms. For each application of Minitab, the required session commands and the corresponding menu commands will be separately introduced.

Startup

To start running the Minitab package, you can find the Minitab icon on the start screen after a little while of switching on the computer in front of you as shown below;



By double clicking on the icon the Minitab program will immediately appear on front of you as shown below;



In the figure above you can see that there are two secondary windows within the main Minitab window, namely the upper one is **Session** and the lower one is **Work sheet 1 ***** or **Data**. You can see that the active window is the Session, so if you want to work on the Work sheet 1 *** instead you can simply click anywhere inside the needed window to make it active.

The Session window is that one, which you can type Minitab commands on and get results and graphs on also. Whereas, the Work sheet 1 *** window is that one, you can enter data only and get data entered from Minitab commands.

Storing Data Sets

To prepare you for Minitab applications, we will know explain how to store one or more data set. As a beginning you need to understand that the data set is stored in a column that is designated by a "C" followed by a number; thus C1 stands for column 1, C2 for column 2, … etc.

Two commands can be used to store data. One is the **Set** command; the other is the **Read** command.

Set command: is used to store one data set at a time.

Session commands: Follow the syntax;

```
MTB > SET   C1          (Enter)
DATA> # # #             (Enter)
DATA> # # # … etc.      (Enter)
DATA> END               (Enter)
MTB >
```

Menu commands: We simply enter data in column C1 in the data window directly after Minitab has started.

Read command: is used to store two or more data sets simultaneously, provided that the data set are all of the same size.

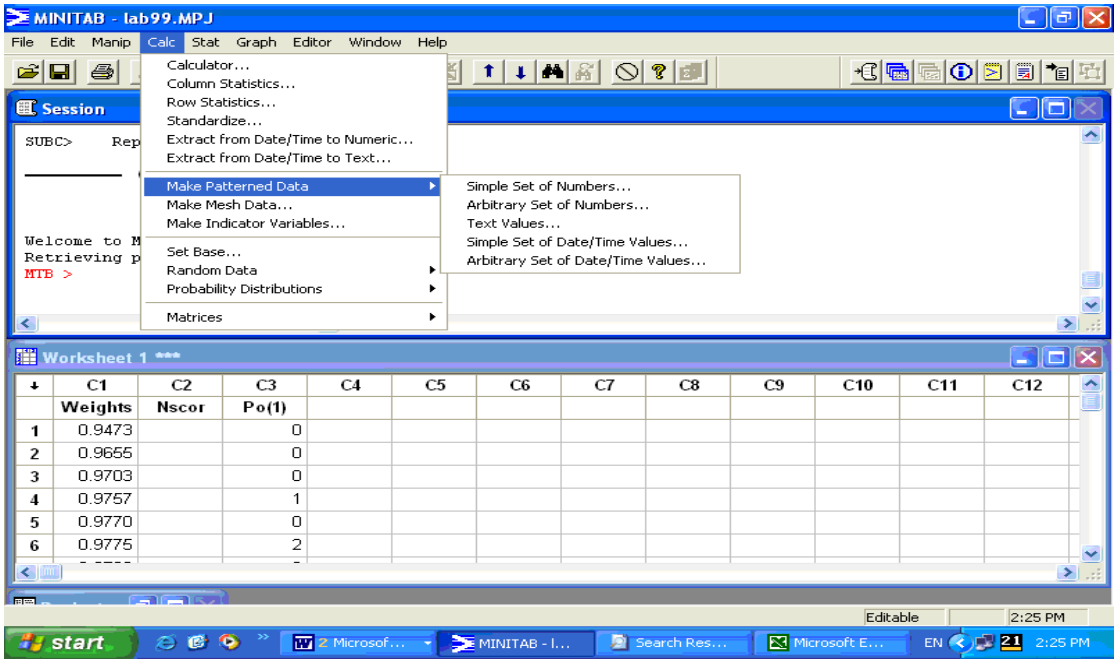
Session commands: Follow the syntax;

```
MTB> READ C1 C2 C3 . . . Cn    (Enter)
DATA> #1 #2 #3 . . . #n        (Enter)
DATA> #1 #2 #3 . . . #n        (Enter)
                .
                .
                .
DATA> #1 #2 #3 . . . #n        (Enter)
DATA> END                      (Enter)
MTB>
```

Menu commands: We simply enter data in columns C1, C2, C3 up to Cn in the data window directly after Minitab has started.

There is a more powerful tool for storing data sets especially patterned ones that enables you to enter data with some pattern like arithmetic sequences or constant values. It can be done as follows;

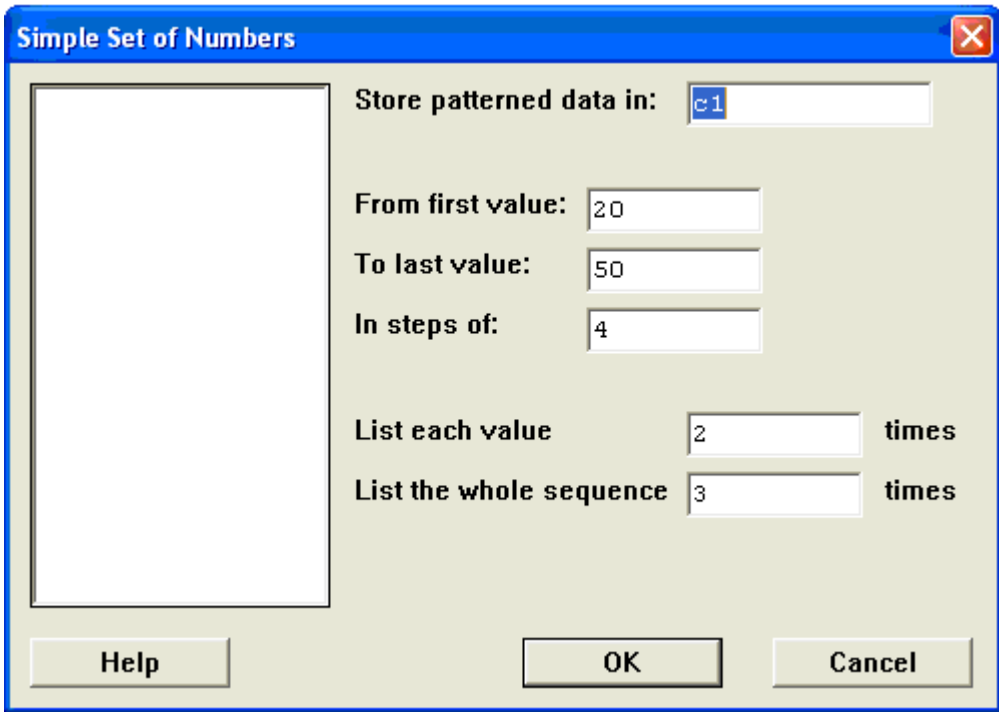
Menu commands: **Calc > Make Patterned Data…**



Example 1. If you want to store the multiples of 4 between 20 and 50 where each value appears twice and the whole sequence repeats 3 times in column C1, i.e. C1 will contain

20 20 24 24 28 28 32 32 36 36 40 40 44 44 48 48
20 20 24 24 28 28 32 32 36 36 40 40 44 44 48 48
20 20 24 24 28 28 32 32 36 36 40 40 44 44 48 48

this can be done as shown in the figure below;



Naming a Column

Naming a column is often useful because it allow us to refer to the column by its name instead of trying to remember its number. This is can be done easily either by using the command `Name C3 'Exam'` while you are in the session window, or simply write the column name directly above the first entry in the column while you are in the data window.

Note: Actually for the ease of dealing with the data, that is entering, deleting or modifying data, it's much better to use the data window instead of the session window.

Copying from Session to Data

You can copy any number of columns from the Session window to the Data window by using the following procedure,

1. Shade the numbers, in the columns, needed only using the mouse or the *shift* and *arrows* keys on the keyboard.
2. Press ctrl+c or click on the copy button to copy.

3. Click on the first cell in the first empty column.
4. Press ctrl+v or click on the paste button to paste.
5. Press OK.

Opening Files

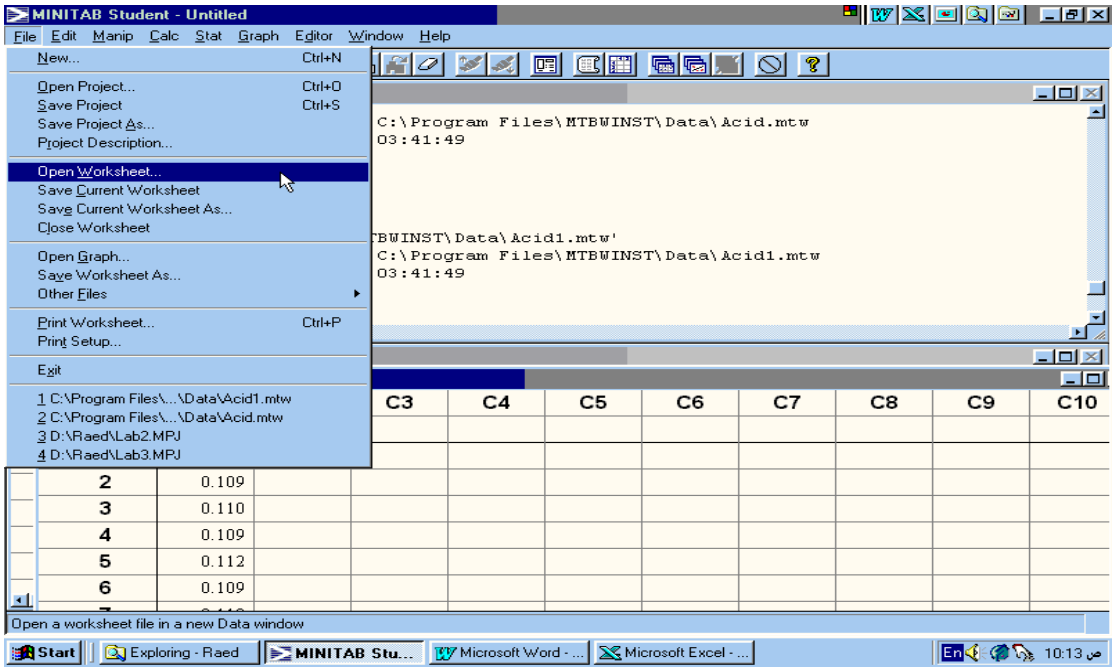
1. To Open an Old Worksheet

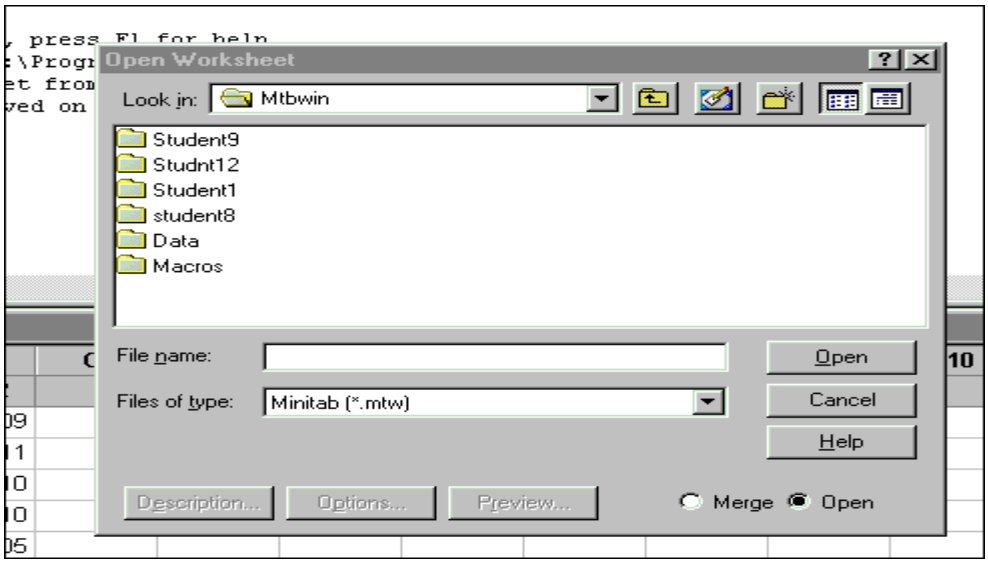
If you want to open an already existing worksheet from the hard disk you can do the following,

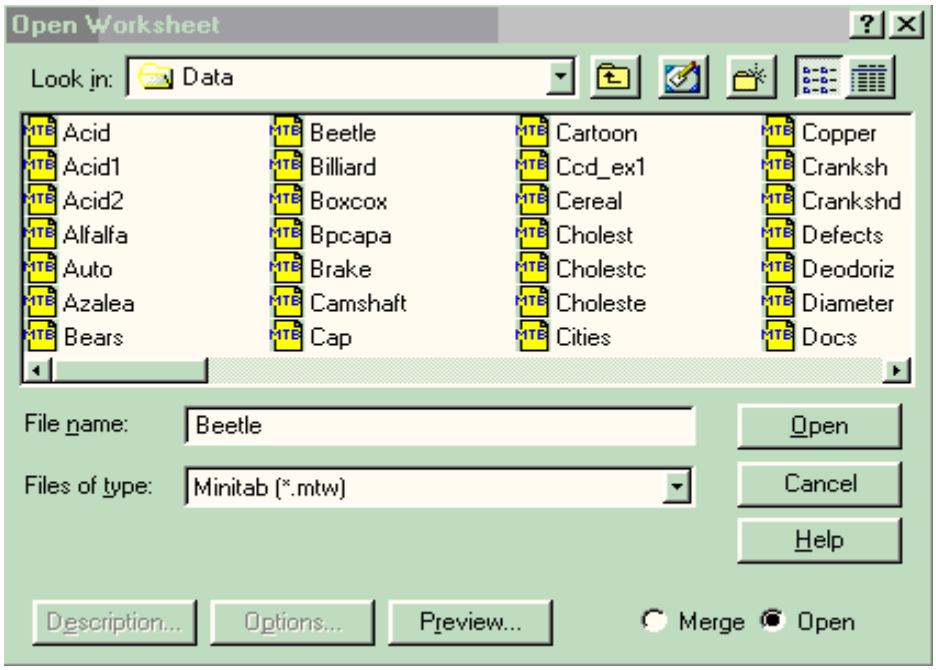
Session commands:

```
MTB > Retr 'C:\Program Files\MTBWINST\Data\Acid.mtw'
```

Menu commands: **File>Open Worksheet…**



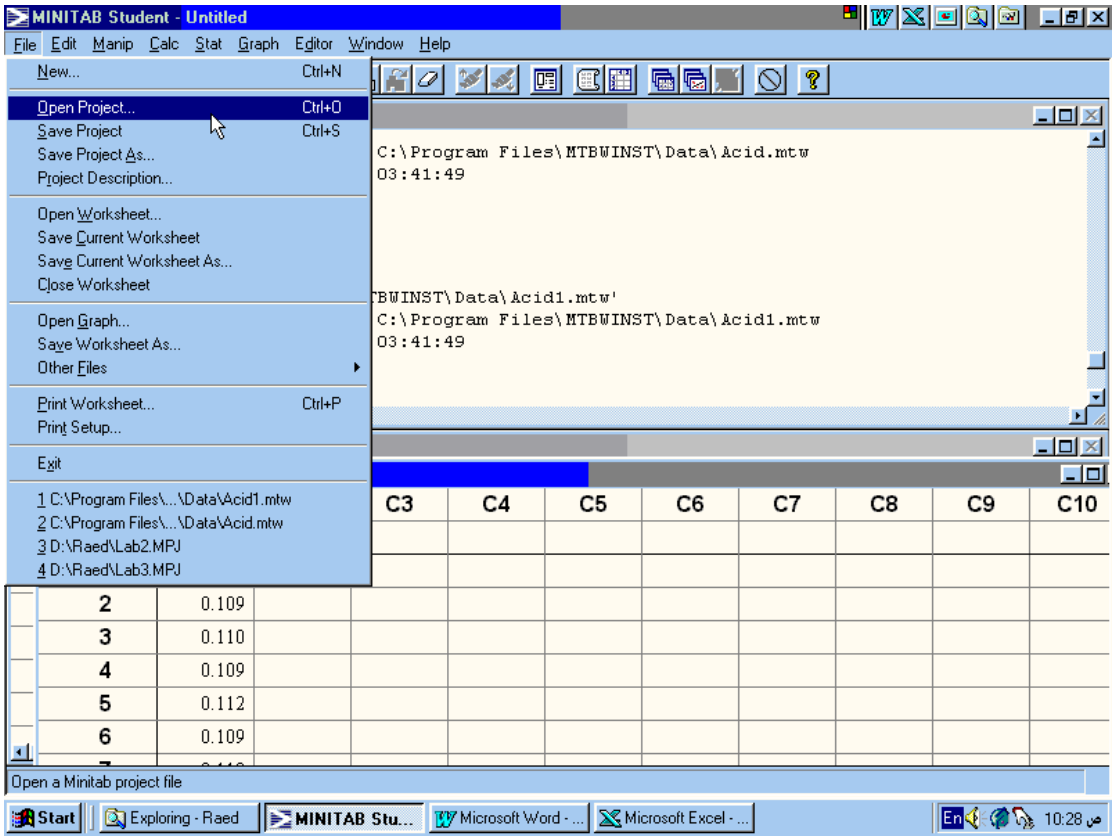




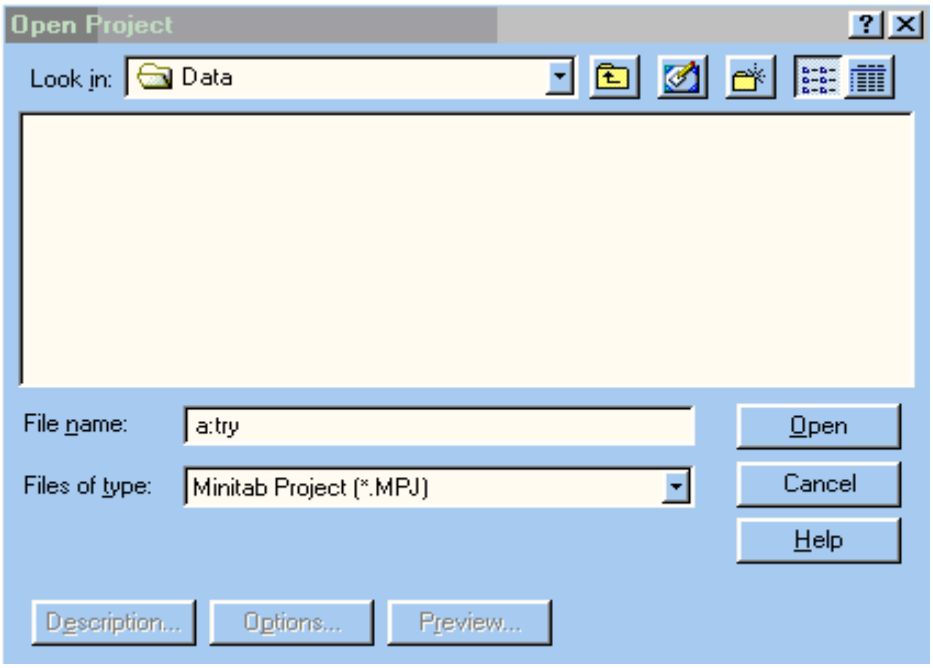
2. To Open an Old Project

If you want to open an already existing project from your own floppy drive you can do the following,

Menu commands: **File>Open Project…**



OR

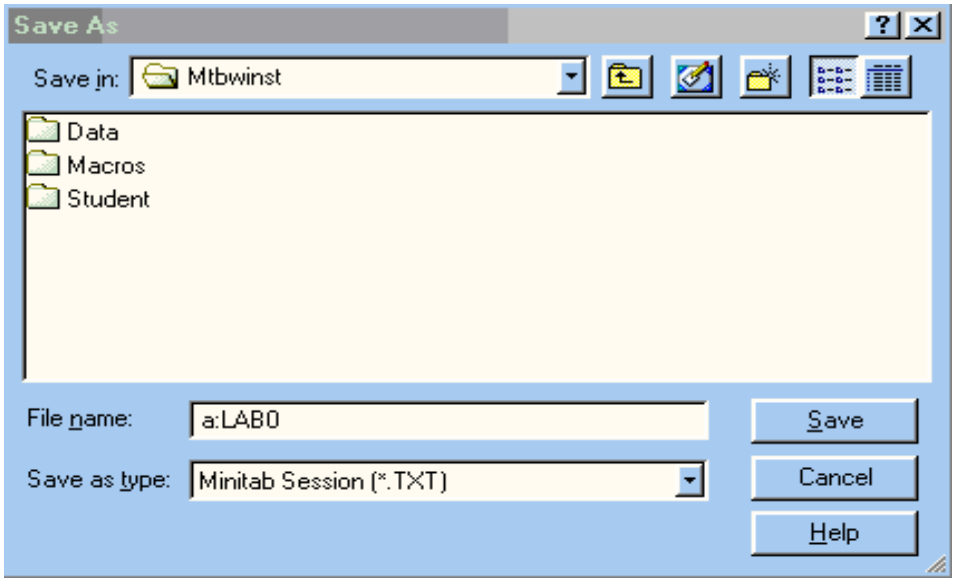


Saving Files

1. Saving Your Session

It's very important to save your commands and the results obtained from it in the end of the class in order to be able to print it on papers using the printer in the lab. This can be done as the following,

Menu commands: **File>Save Session Window As…**



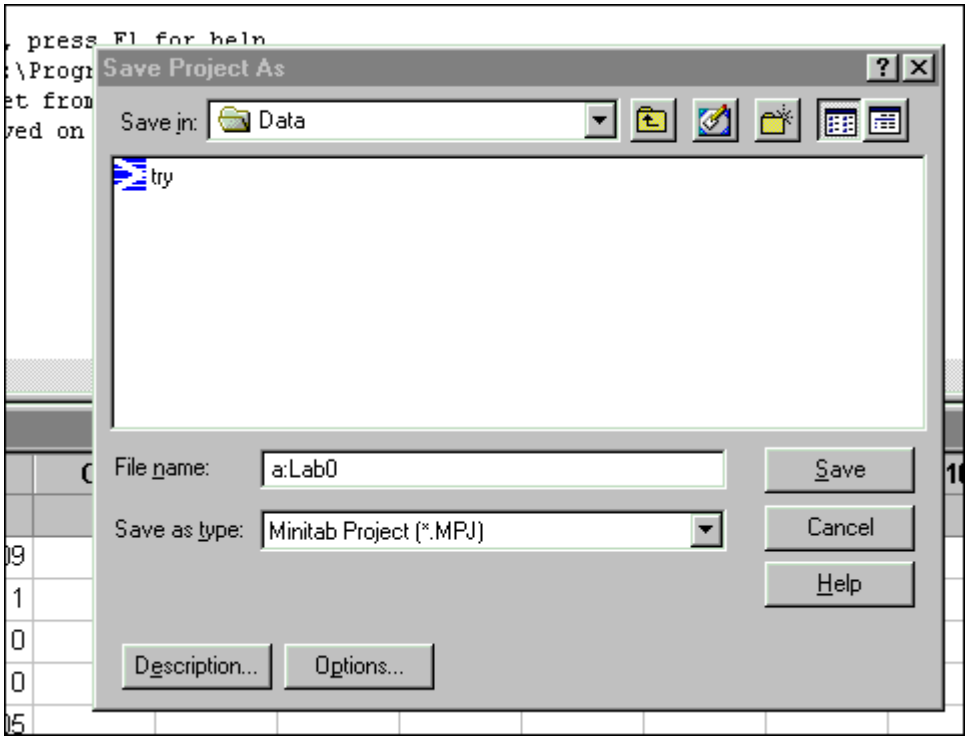
2. Saving Your Project

It's very important to save your whole work in the end of the class in case you need to modify on your project later. This can be done as the following,

Session commands:

```
MTB > Save  'a:lab0';
SUBC> Project.
```

Menu commands: **File>Save Project…**

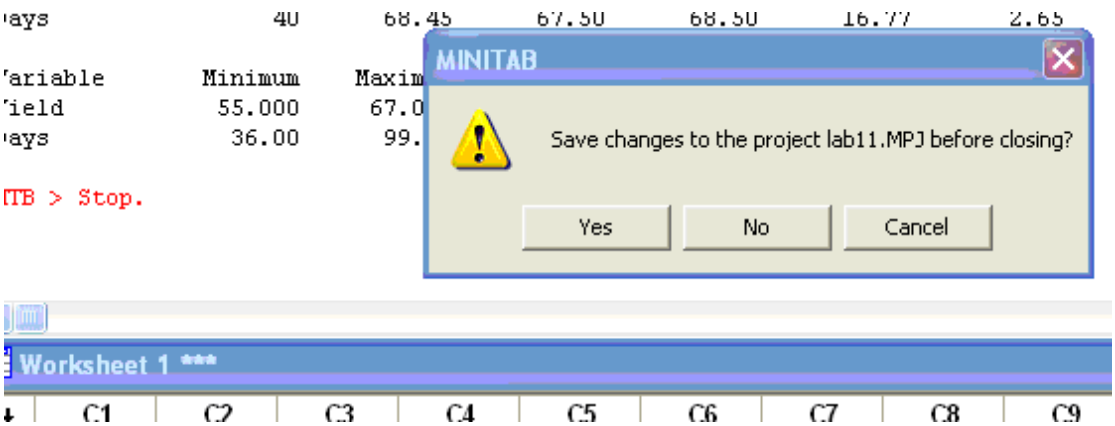


Exiting Minitab

When we are finished with Minitab, we must exit the software. This is accomplished in the following manner:

Session commands: Type *Stop* command then press (enter).

Menu commands: **File > Exit** then click on **NO**.



Printing Your Work

After exiting Minitab you need to print your work on the printer, this can be achieved by the following commands, after you insert your floppy in the floppy drive in the case of the printer's computer,

c:\> a: (Moves the reader head to drive a)
a:\> dir (Exhibits your files)
a:\> copy filename.extension prn

```
C:\WINDOWS>cd\

C:\>a:

A:\>dir

 Volume in drive A has no label
 Volume Serial Number is 3261-19FA
 Directory of A:\

WORKSHOP DOC        23.552  07/05/02 11:09a Workshop.doc
FALL2002 XLS       111.616  15/01/03  4:45p Fall2002.xls
LAB0     MPJ         8.192  23/01/03  1:09p Lab0.MPJ
PAPER2   TXT        16.332  29/07/02 12:15p Paper2.txt
         4 file(s)        159.692 bytes
         0 dir(s)       1.296.896 bytes free

A:\>copy paper2.txt prn
```

From know on you are able to start Minitab, deal with data and exit easily. The details of other commands will be given at the suitable location of this booklet when it is necessary.

UNIT I
DESCRIPTIVE STUDY OF DATA

Main aspects of describing a data set:

- Summarization and description of the overall pattern of the data by:

1. Presentation of tables and graphs
2. Examination of the overall shape of the graphed data for important features, including symmetry or departures from it.
3. Scanning the graphed data for any unusual observation that seems to stick far out from the major mass of the data.

- Computation of numerical measures for

1. A typical or representative value of the center of the data.
2. The amount of spread or variation present in the data.

LAB 1
REPRESENTATION OF DATA BY CHARACTER GRAPHS

Dot-plot

The dot-plot is a graphical display for quantitative data. Dot-plots are particularly useful for showing the relative positions of the data in a data set or for comparing two or more data sets, when the data consist of a small set of numbers (say, less than 20 or 25).

Dot-plot can be performed by Minitab using command **Dotplot**.

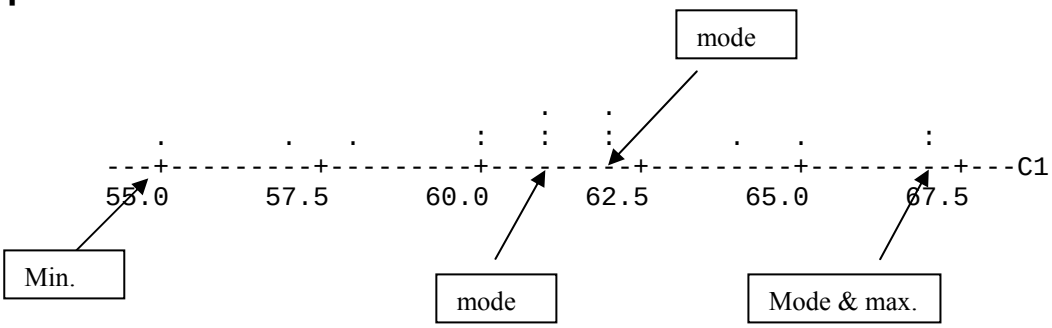
Open the Minitab project *Lab1.mpj* from your floppy disk.

Example 1. Construct a dot-plot for the data Yield stored in C1.

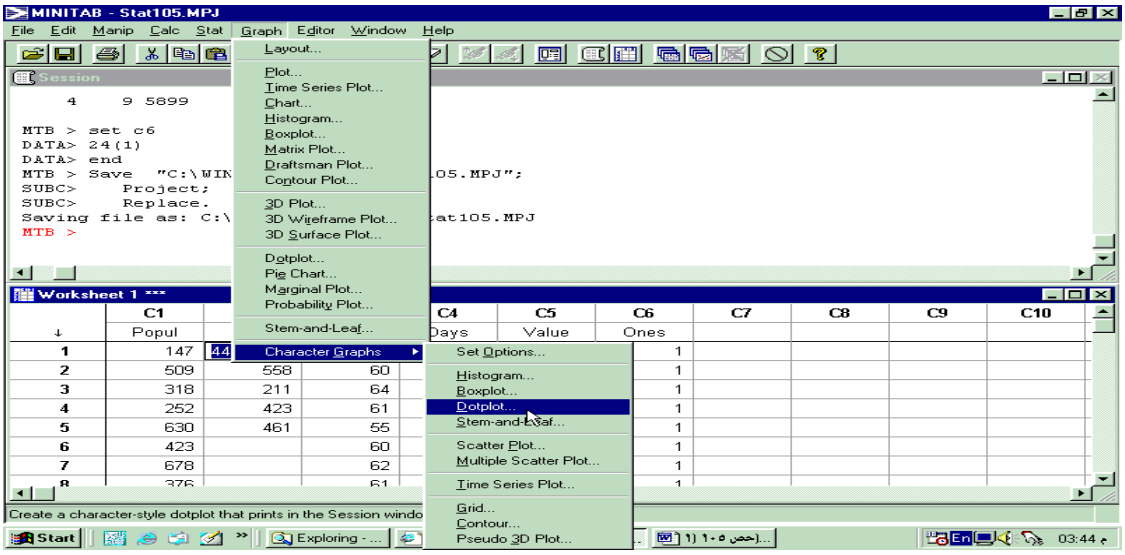
Session commands:

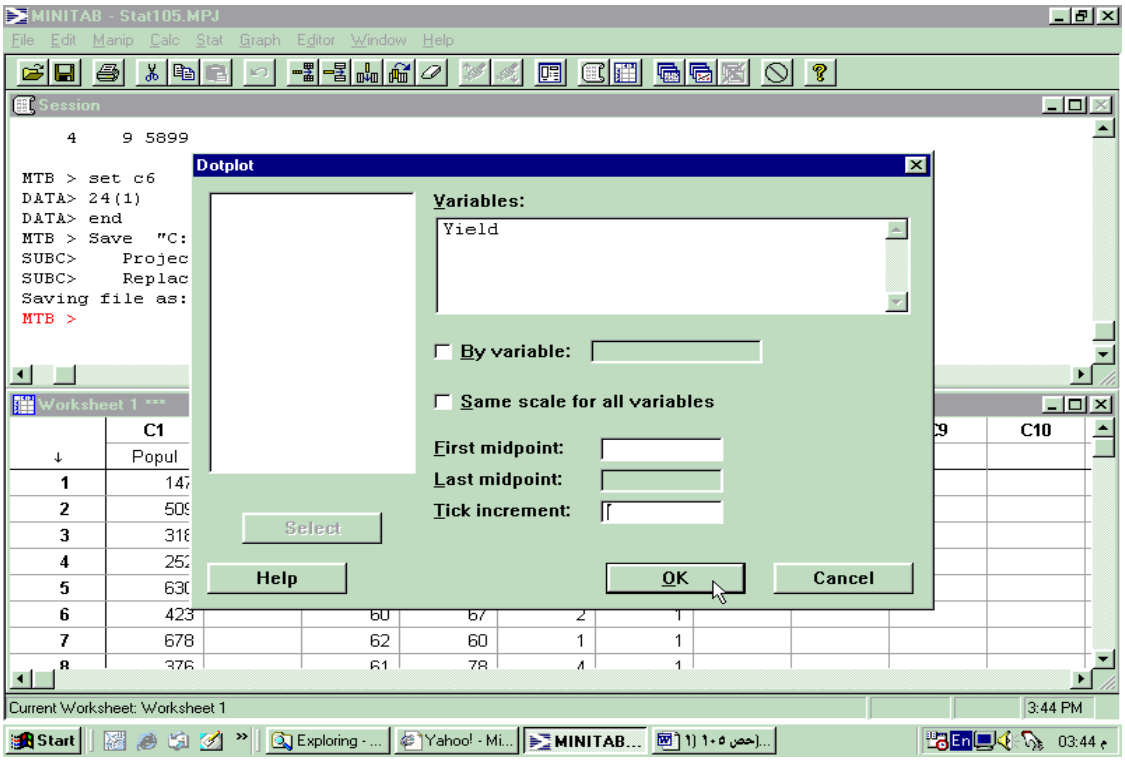
```
MTB > Dotp 'Yield'
```

Dotplot



Menu Commands: **Graph> Character Graphs> Dotplot…**





Note: The *Tick increment* option is used to control the width of the plot since the width of the plot is inversely proportional to the tick increment given that the width of the plot should not exceed the screen width.

Comment: from the previous dot plot it is clear that it has 3 modes (61, 62 and 67) and it is right skewed (or +ve skewed).

Box-Plot

Box- plot is a graphical technique to display the structure of the data set by using the quartiles and the extreme values of a sample.

To construct a Box plot we need to know the following definitions:

The Inter-Quartile Range (IQR) = O_3-Q_1

Where Q_1, Q_2 and Q_3 are the first, the second and the third quartiles respectively.

Inner fences: (Lower) LIF= Q_1-1.5IQR, (Upper) UIF= Q_3 + 1.5IQR
Outer fences: (Lower) LOF= Q_1-3IQR, (Upper) UOF= Q_3+3IQR

In Box plot, data values that lie between the inner and the outer fences are considered possible, or mild, outliers, whereas those lie outside outer fences are considered probable, or extreme, outliers. Box-plot can be obtained using Minitab by the command **Box-Plot**.

The advantages of the box-plot are the following:

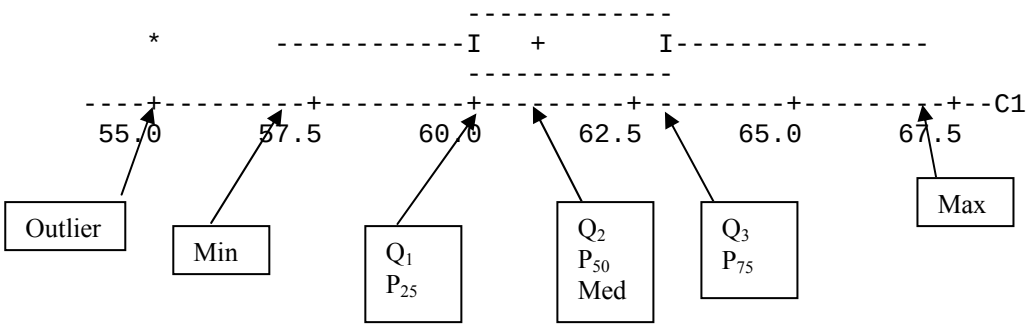
a. It gives the five-number summary.
b. It gives the outliers and distinguishes between the possible and the probable ones.
c. It determines, more accurately, the symmetry and skewness of the graph by the following:
   - If $Q_3 - Q_2 > Q_2 - Q_1$ then it is skewed to the right (+ve skewed).
   - If $Q_3 - Q_2 < Q_2 - Q_1$ then it is skewed to the left (-ve skewed).
   - If $Q_3 - Q_2 = Q_2 - Q_1$ then it is symmetric.

   Considering that the tails are approximately of equal length.

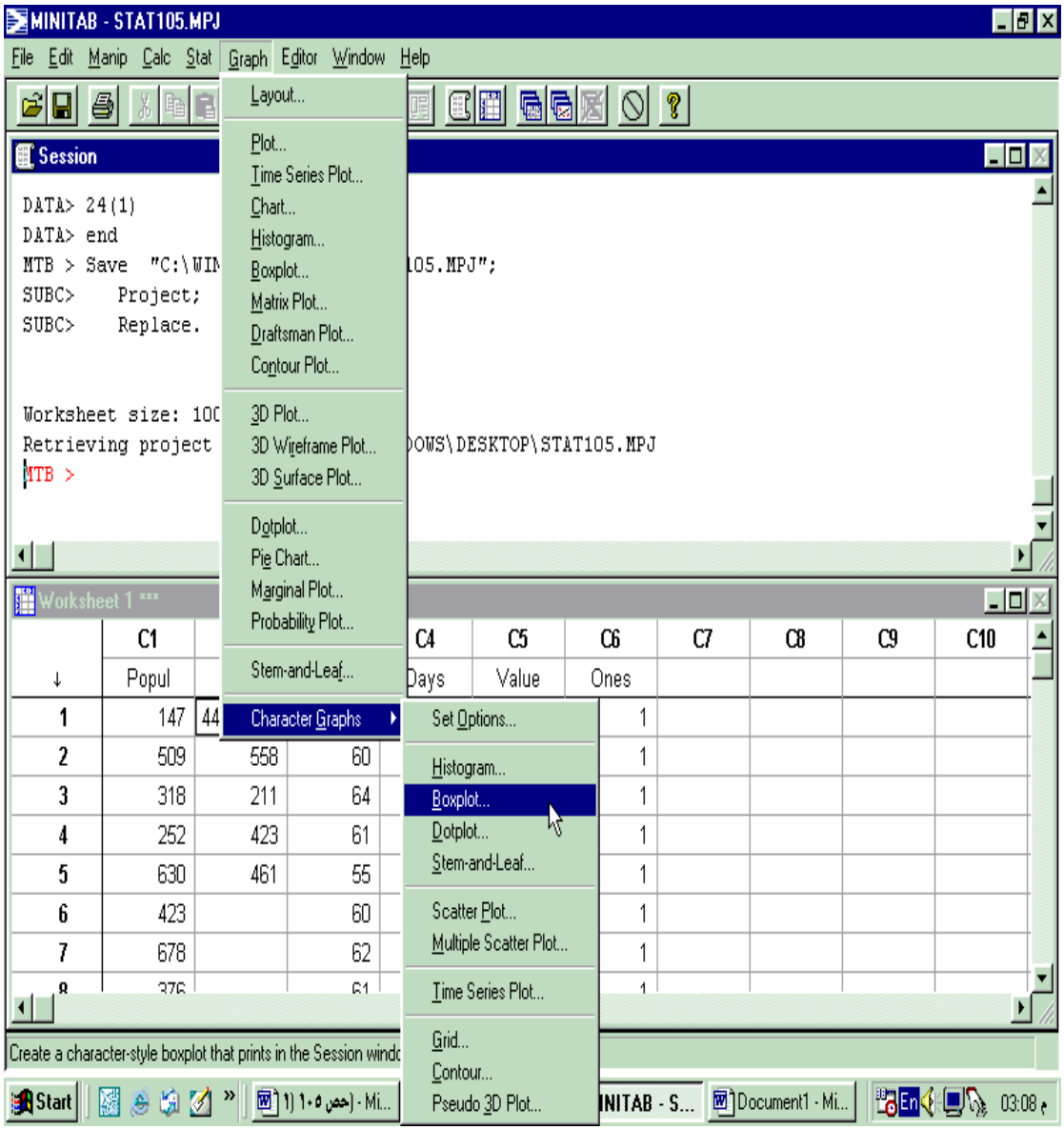
Example 2. For the previous sample, construct a box-plot.
Session commands:

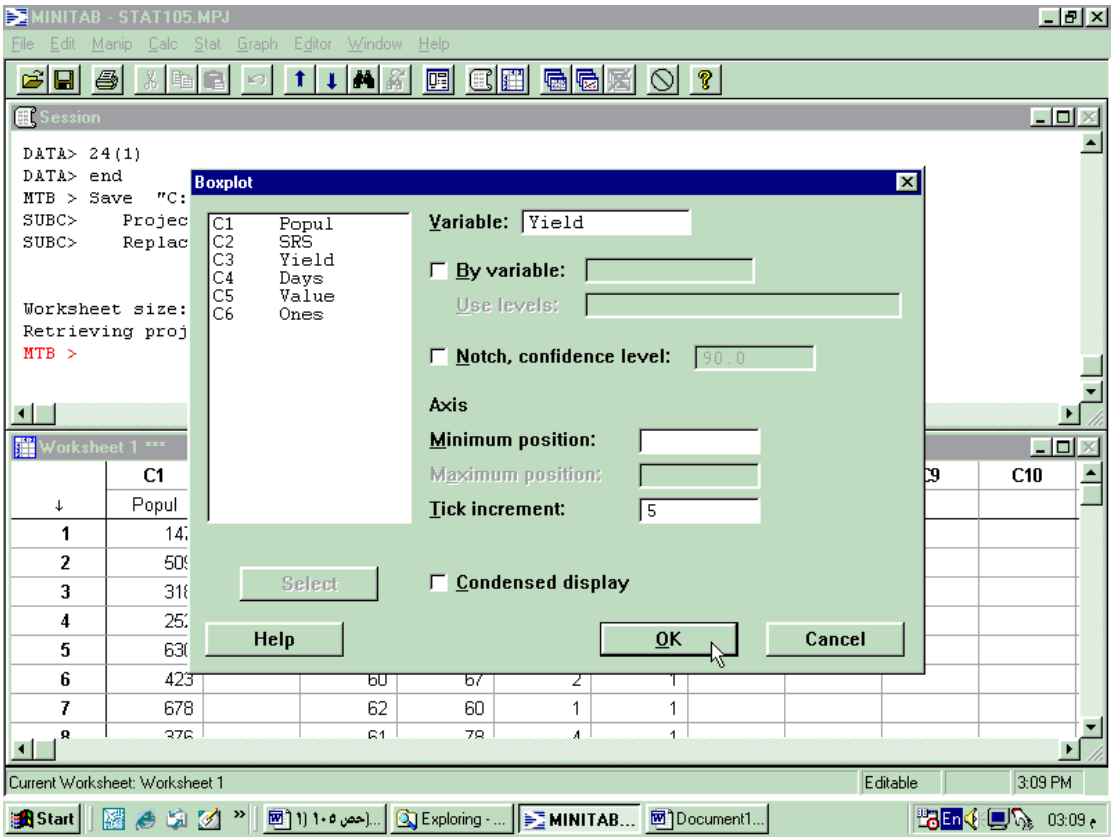
```
MTB > GStd.
MTB > Boxp C1.
```

Boxplot



Menu commands: **Graph> Character Graphs>Boxplot…**





In Minitab the asterisk (*) denotes a possible, or mild, outlier where the circle (o) denotes a probable, or extreme, outlier.

Stem-and-leaf diagram

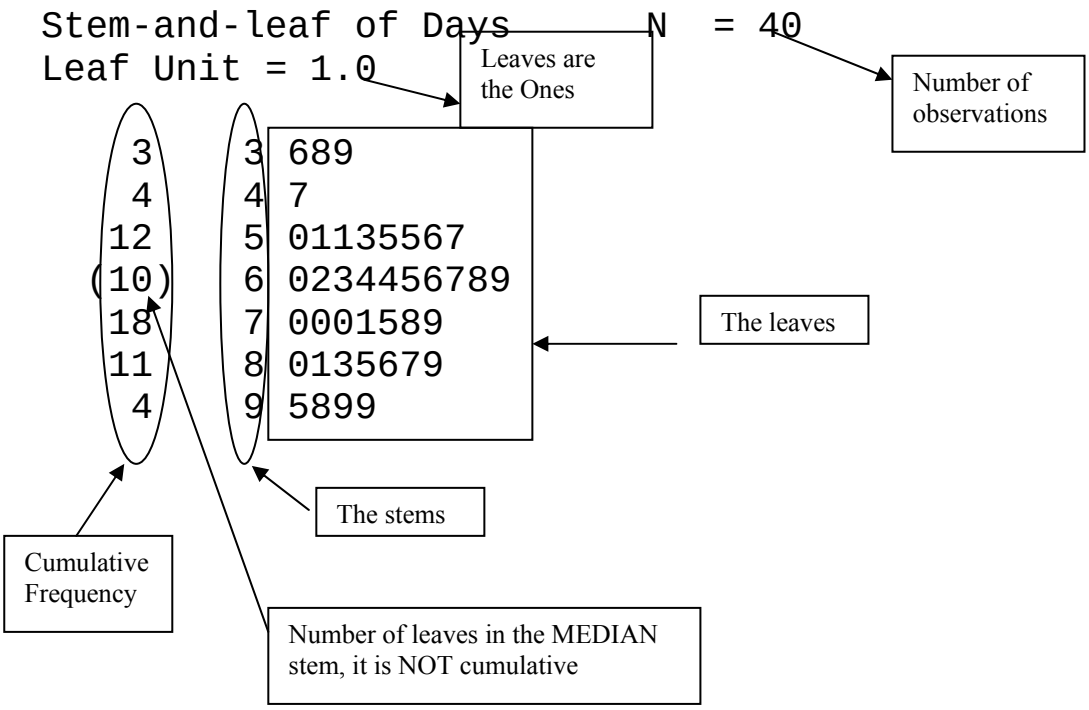
This method was developed in the late 1960's by Professor John Tukey. It is often easier to construct than either frequency distribution or histogram and generally displays more information. This can be performed in Minitab using the command **Stem.**

Example 3. Construct an ordered stem-and-leaf diagram for the data Days, one value for (ten leaves in) each stem, we let the increment to be 10;

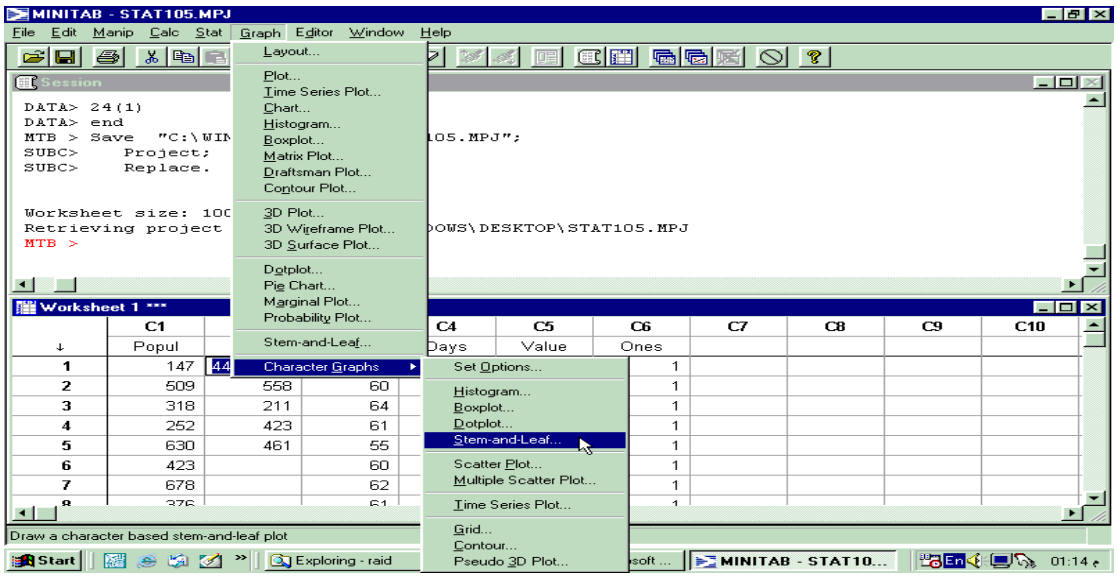
Session commands:

```
MTB > stem c2;
SUBC> incr 10.
```

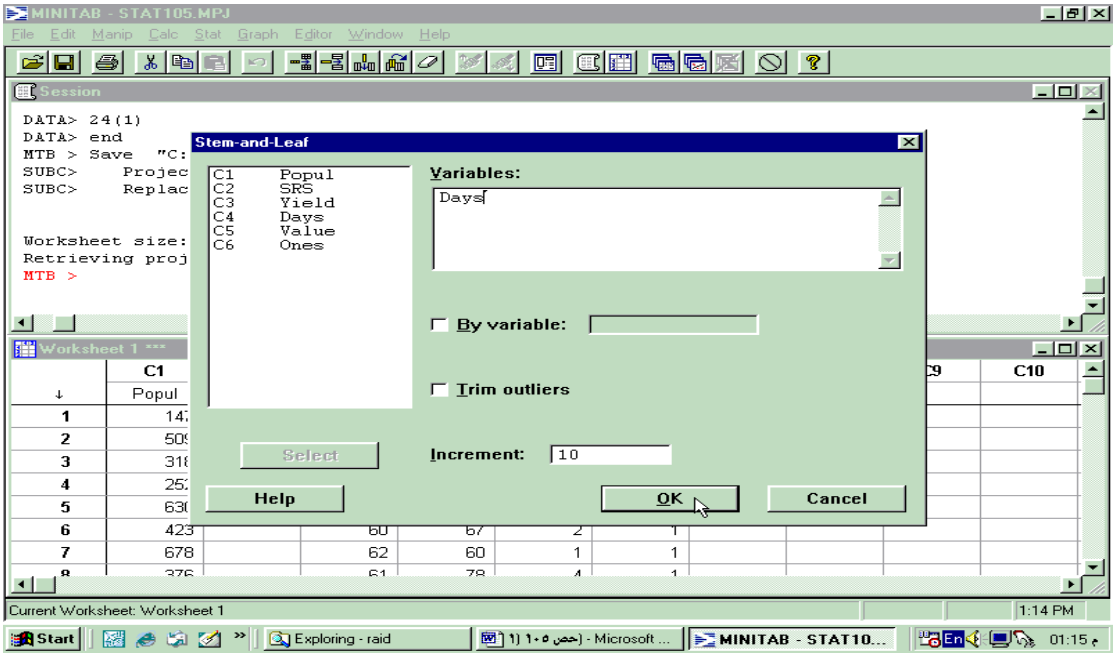
Character Stem-and-Leaf Display



Menu commands: **Graph>Character Graphs>Stem-and-Leaf…**



Comment: from the previous stem-and-leaf diagram it is clear that it is right skewed, since the number of leaves right to the median is greater than that left to the median. The number of modes will be the same regardless of the type of the stem-and-leaf diagram (ordered or modified).



Example 4. To obtain a modified stem-and leaf diagram, for the data Days, with two values for (five leaves in) each stem let the Increment be 5,

Session commands:

```
MTB > stem 'days';
SUBC> incr 5.
```

Character Stem-and-Leaf Display

```
Stem-and-leaf of Days       N   = 40
Leaf Unit = 1.0

  3     3 689
  3     4
  4     4 7
  8     5 0113
 12     5 5567
 17     6 02344
 (5)    6 56789
 18     7 0001
 14     7 589
 11     8 013
  8     8 5679
  4     9
  4     9 5899
```

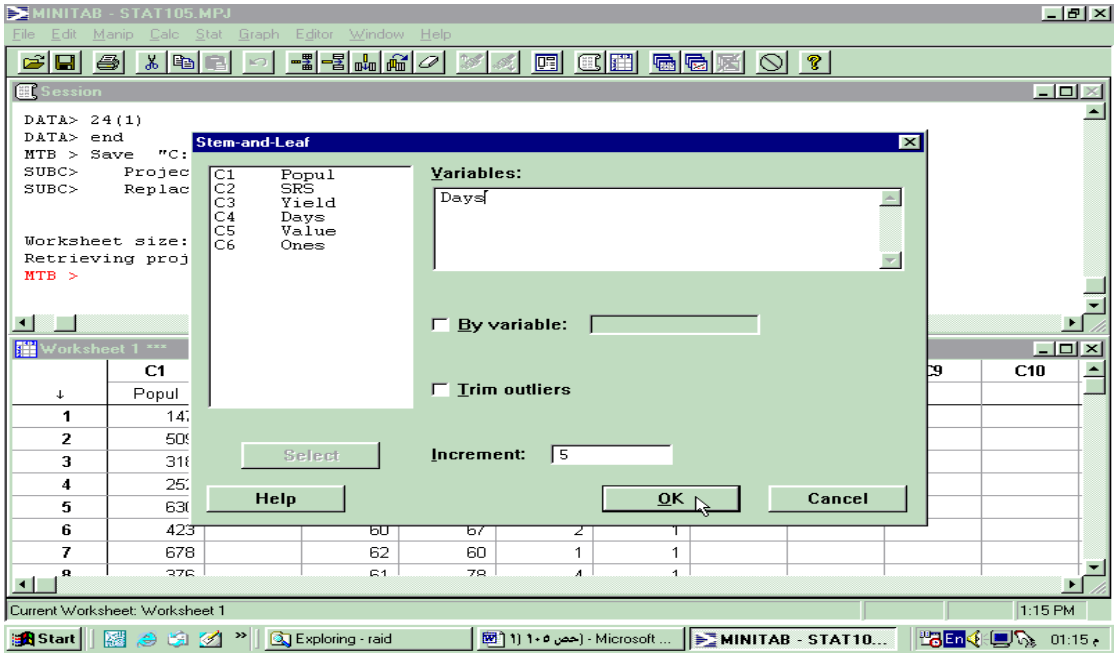
Each stem has 2 values

Each value of the stem has 5 leaves

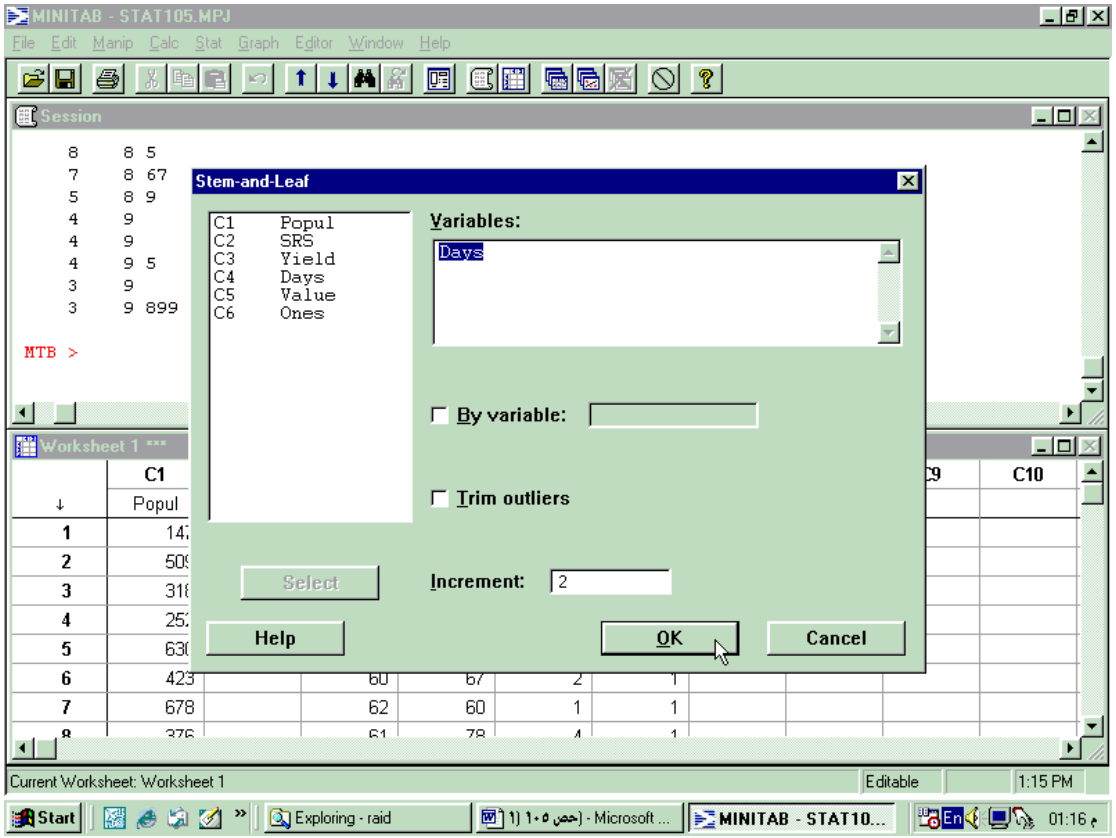
The median lies between 65 & 69

Notice that the leaves appear in ascending order in all the cases

Menu commands: Type 5 in **Increment**.



Example 5. To obtain a modified stem-and leaf diagram, for the data Yield, with five values for (two leaves in) each stem let the Increment be 2;
Menu commands: Type 2 in **Increment**.

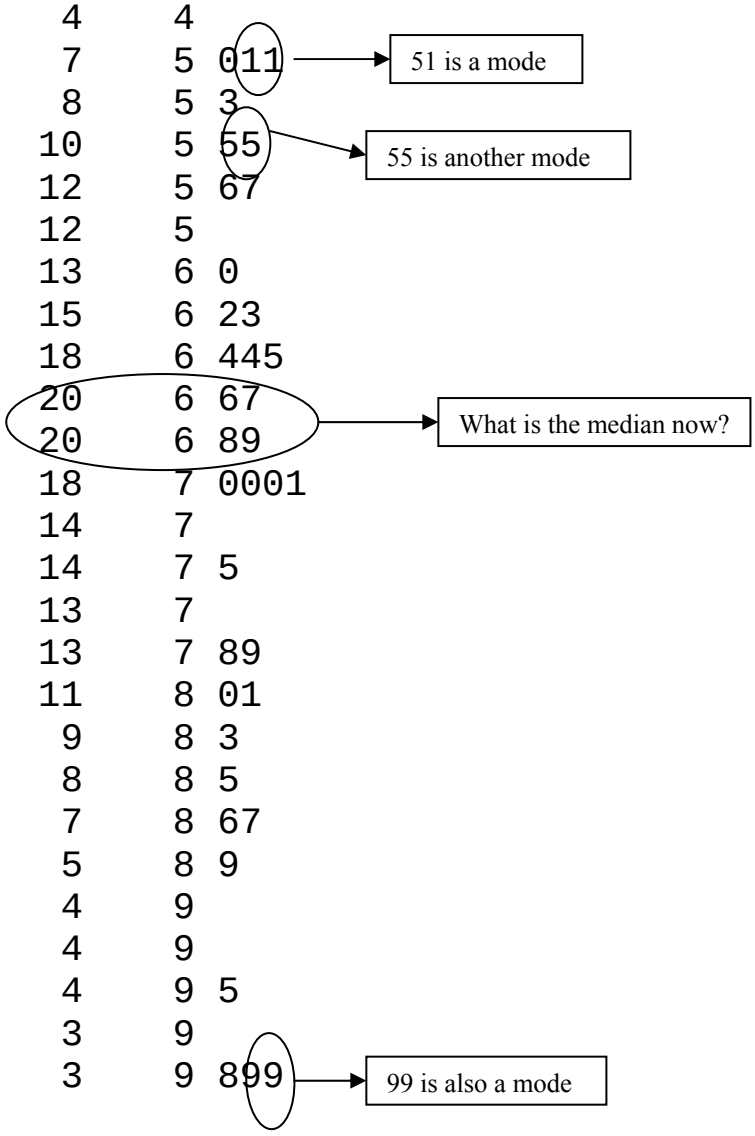


Session commands:

```
MTB > stem 'Days';
SUBC> incre 2.
```

Character Stem-and-Leaf Display

```
Stem-and-leaf of Days      N  = 40
Leaf Unit = 1.0

 1    3 6
 3    3 89
 3    4
 3    4
 3    4
 4    4 7
 4    4
 7    5 011
 8    5 3
10    5 55
12    5 67
12    5
13    6 0
15    6 23
18    6 445
20    6 67
20    6 89
18    7 0001
14    7
14    7 5
13    7
13    7 89
11    8 01
 9    8 3
 8    8 5
 7    8 67
 5    8 9
 4    9
 4    9
 4    9 5
 3    9
 3    9 899
```



LAB 2
RERESENTATION OF DATA BY HIGH RESOLUTION GRAPHS

Random Sample

If we are to use information obtained from a sample to draw conclusions about a population, the sample must be representative of the entire population. A type of sample that is representative of the entire population is a random sample so the random sample is defined as follows,

Definition: A simple random sample (SRS) of measurements from a population is a one selected in such a manner that every sample of size n from the population has equal chance, probability, of being selected, and every member of the population has equal chance of being included in the sample.

The traditional way to select a SRS is to use the random-number table, but here we will not use such a table because the Minitab software can be employed to obtain a SRS from a population.

Open the Minitab project *Lab2.mpj* from your floppy disk.

Example 1. Generate (draw or simulate) a SRS of size 5 from the population in C1 and store it in C2.

Session commands:

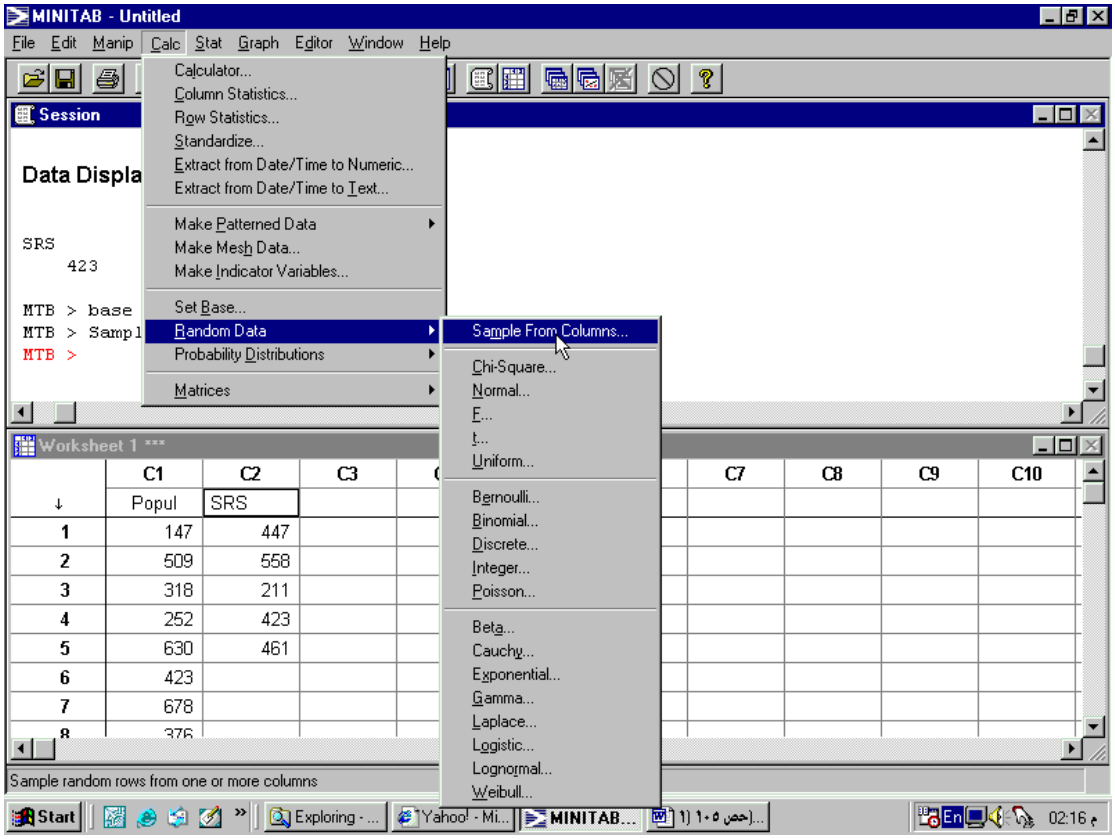
```
MTB > base 0
MTB > Sample 5 'Popul' 'SRS'
MTB > Print 'SRS'
```

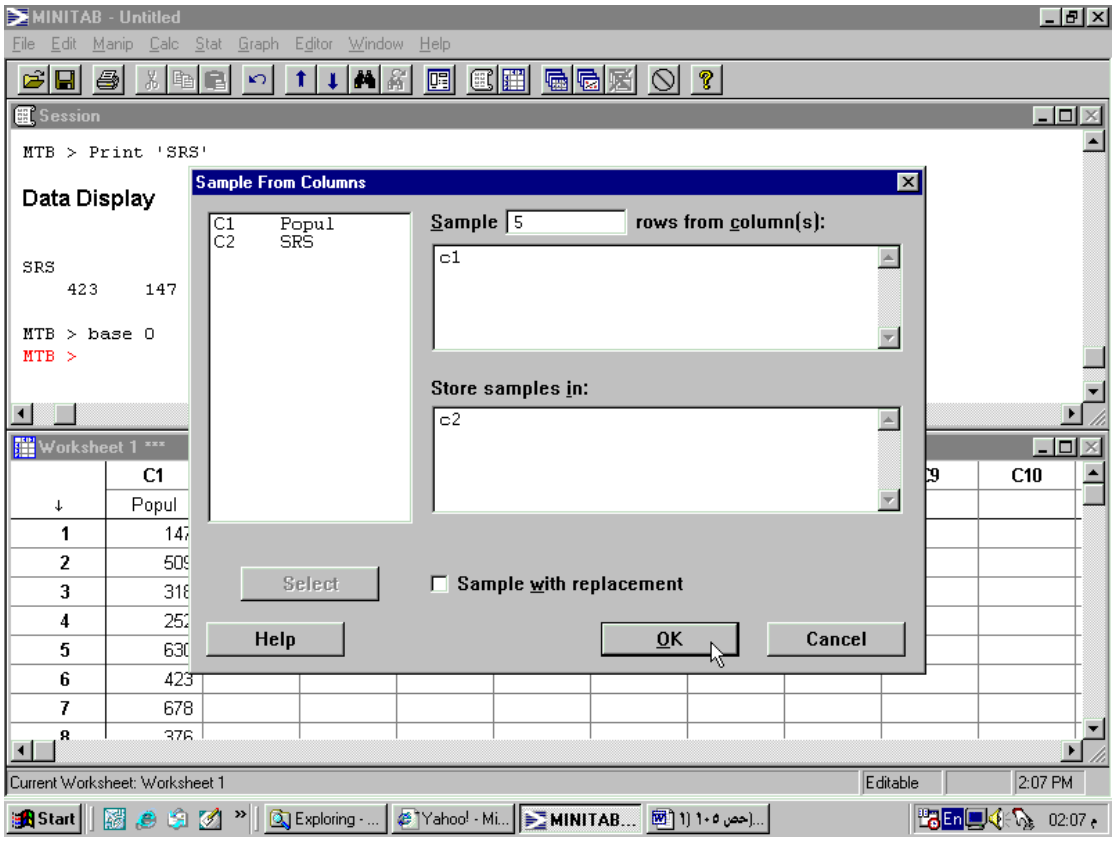
Compulsory in sessions and exams
Forbidden in HW's

Data Display

```
SRS
    423    147    509    211    577
```

Menu commands: **Calc> Random Data> Sample From Columns…**





Note: Each time you need to generate a random sample in the class you MUST execute the command (`MTB > base 0`) first.

Bar Chart and Pie chart

Both charts, bar and pie, are used to represent discrete and qualitative data in high resolution graphs.

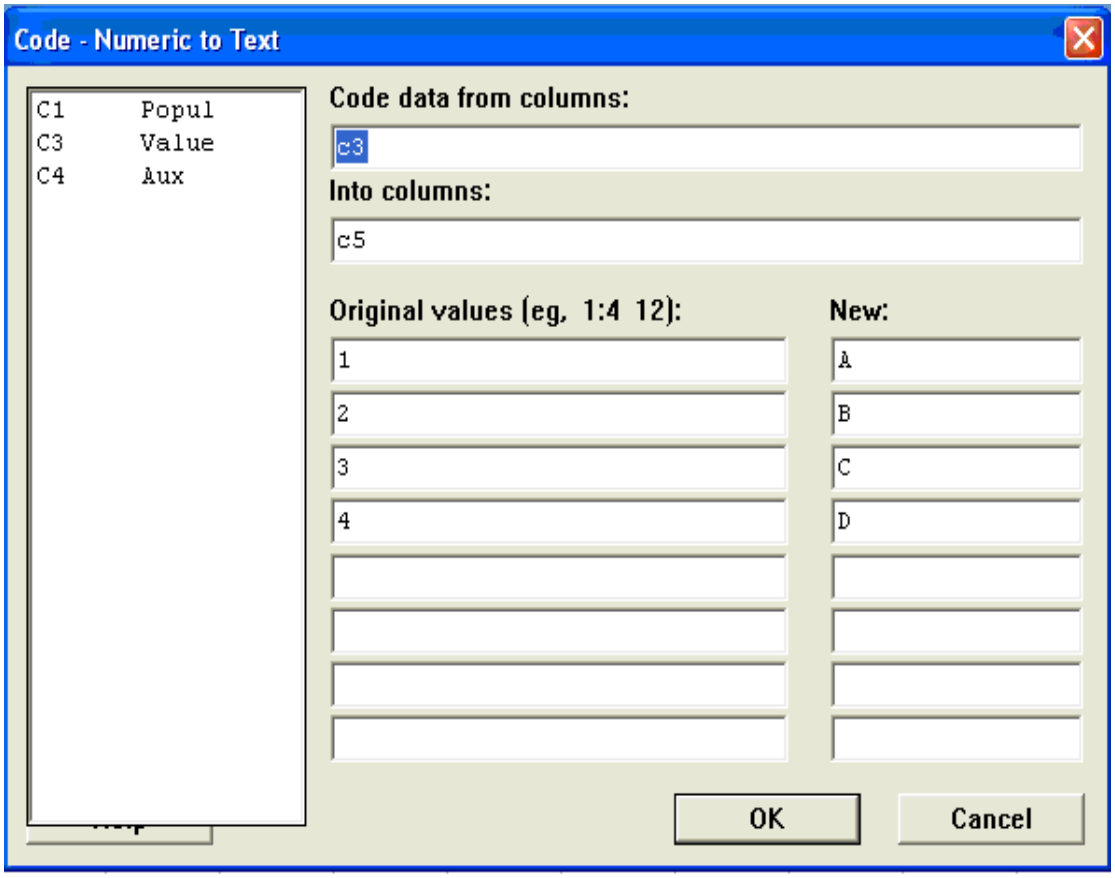
1. Bar Chart

To make a bar chart, the class values are marked a long the horizontal axis and a vertical line (or bar) of height equal to the class frequency is erected over the respective class values. This can be performed using Minitab, for the following data set, as follows:

Example 2. Construct a bar chart for the data Value and Coded.

Before constructing a bar chart for any set of data you must create a new column filled with a number of a constant digits equal to the number of observations in the column to be charted.

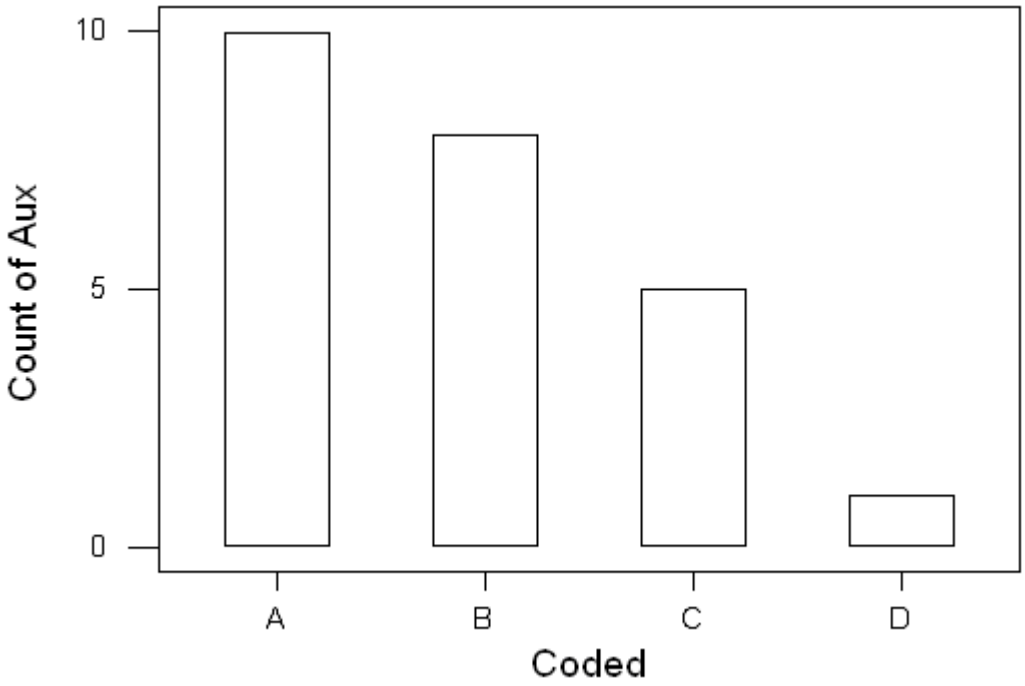
Note: Bar chart and Pie chart are the only graphs that can be used to represent qualitative, nonnumeric, data. So we will construct the bar chart and the pie chart for the variable Coded which is a coded column of the column Value as shown below;



Session commands:

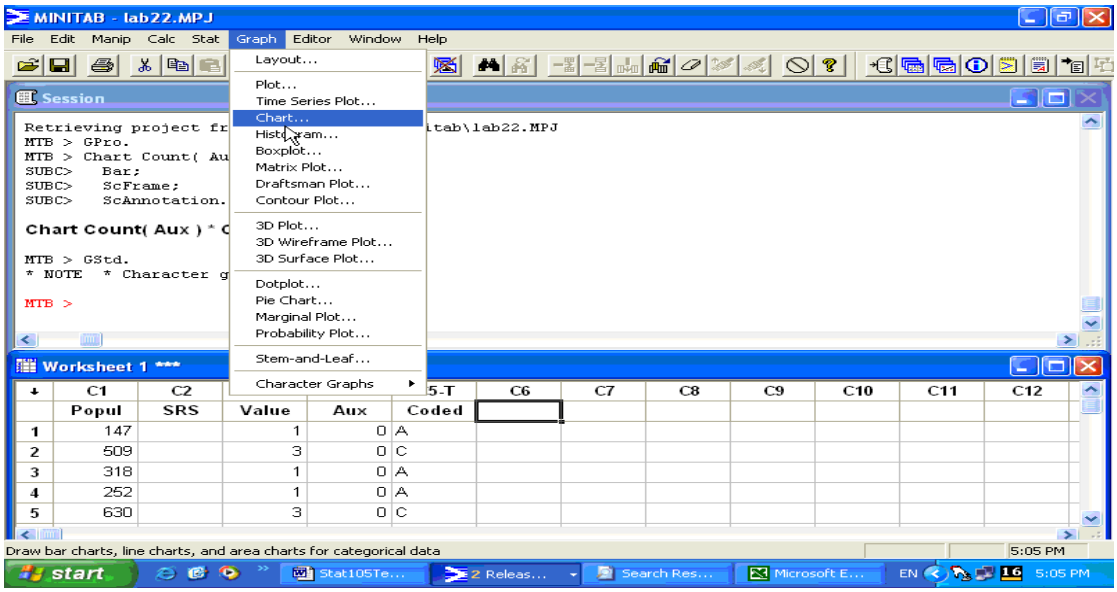
```
MTB > set c4
DATA> 24(0)
DATA> end
MTB > Chart Count (C4) * c5;
SUBC>   Bar.
```

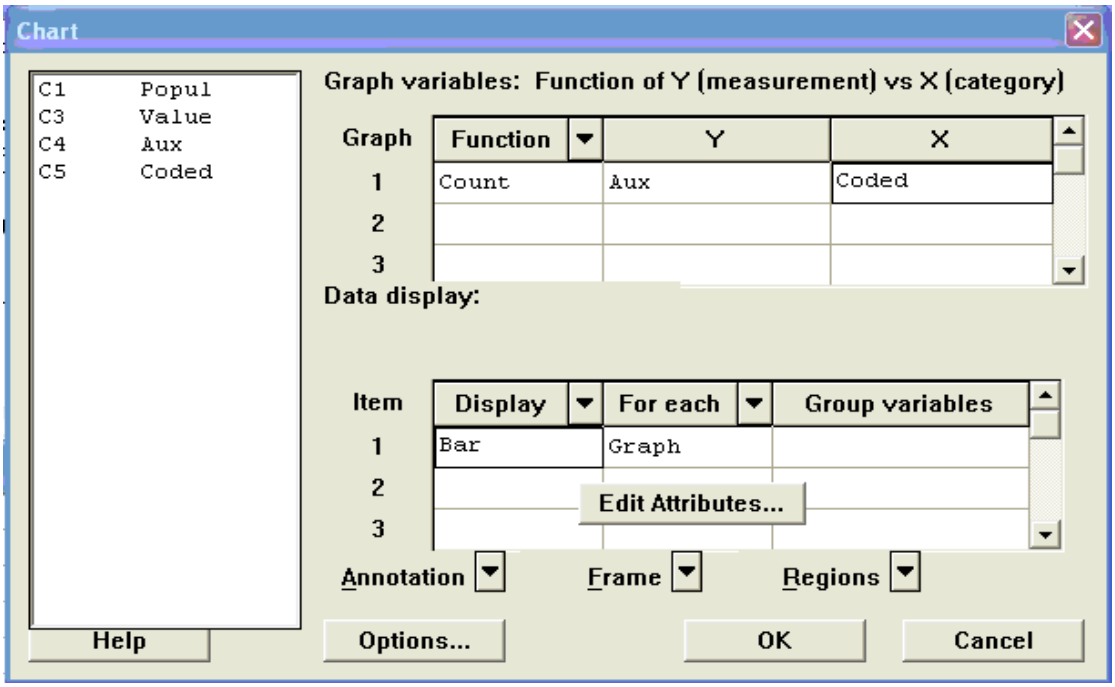
Here you can use any digit and not only 0



Comment: from the previous bar-chart it is clear that the chart is right skewed and it has one mode at the value A.

Menu commands: **Graph>Chart…**



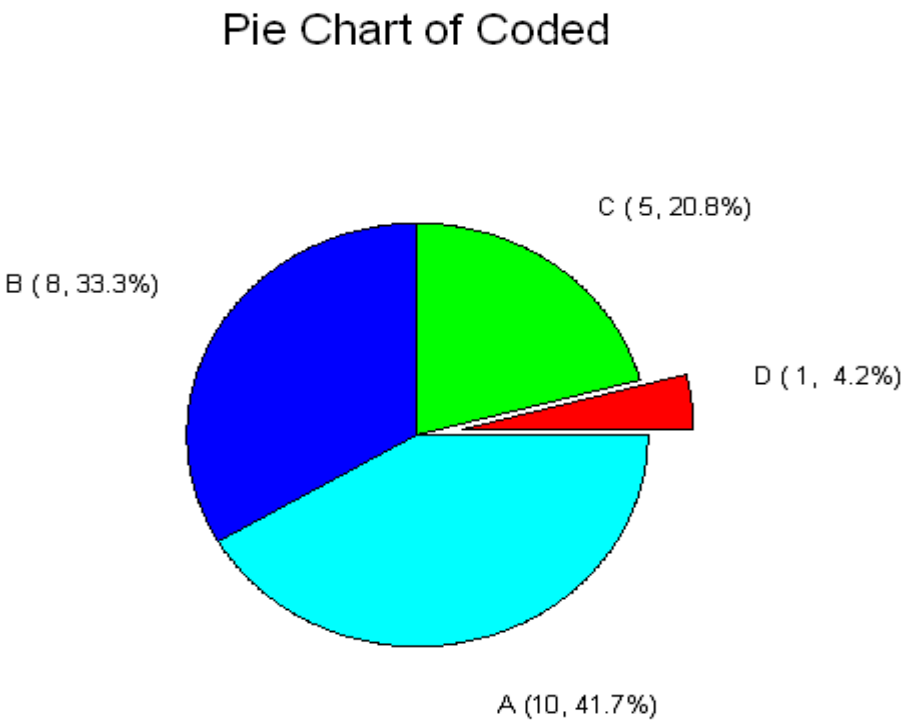


2. Pie chart

Example 3. For the previous example, the pie chart can be drawn using Minitab as follows:

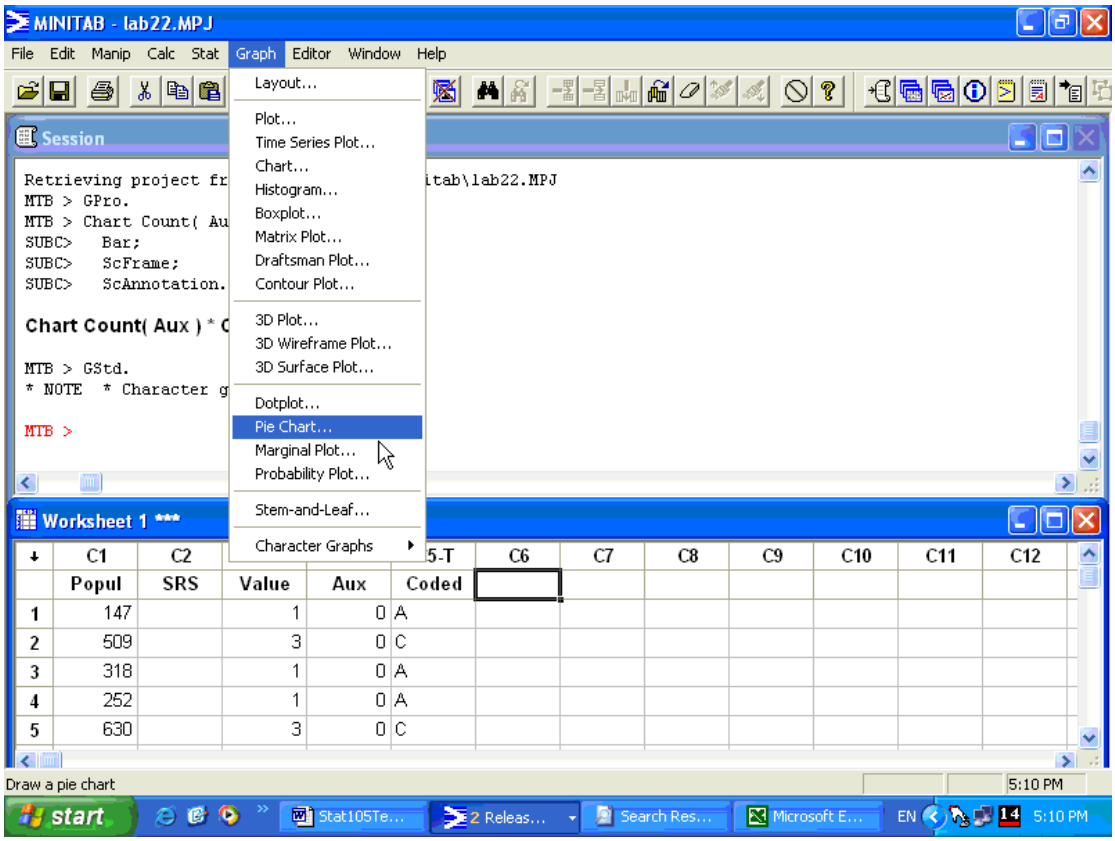
Session commands:

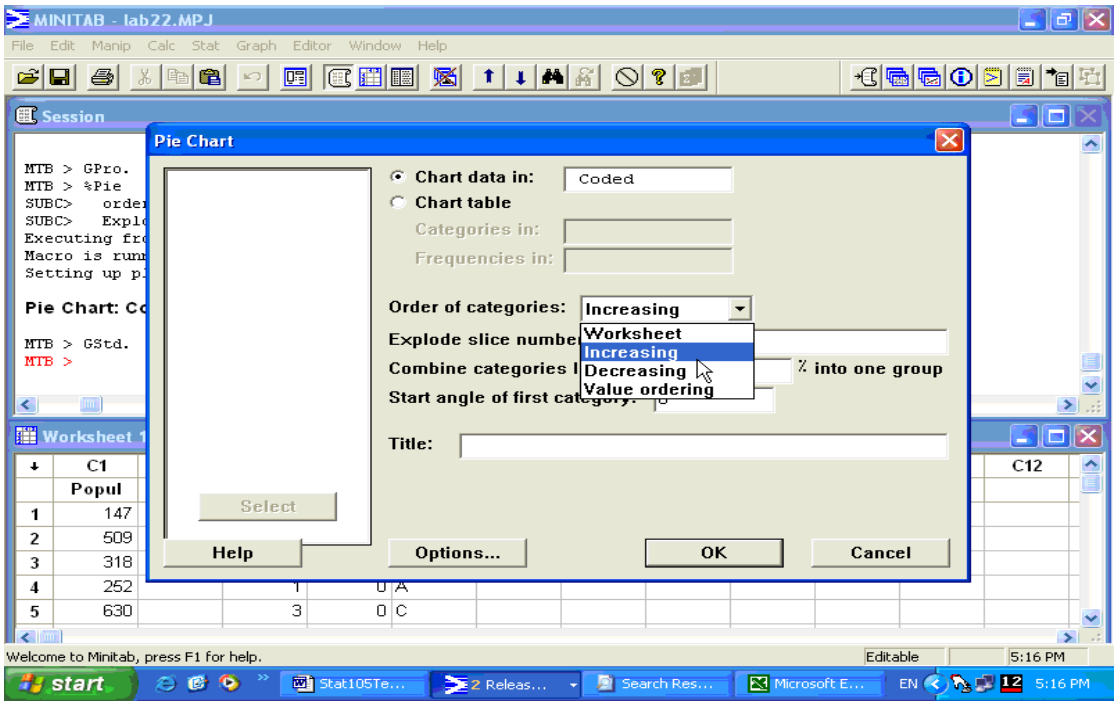
```
MTB > %Pie 'Value';
SUBC> orde 2;
SUBC> Expl 1.
```

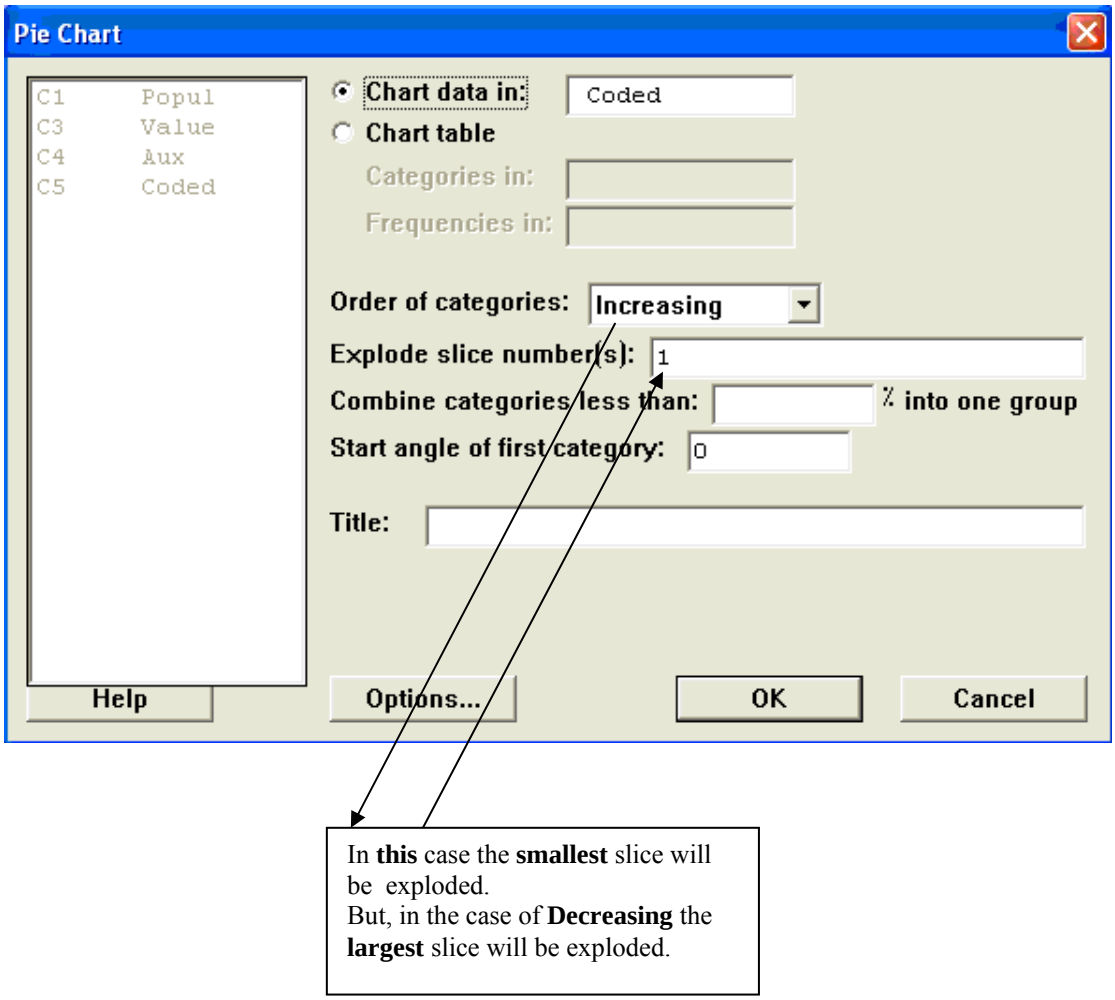


Comment: From the previous pie-chart it is clear that the LARGEST slice is for the value A and the SMALLEST slice is for the value D.

Menu commands: Choose **Graph>Pie Chart…**







Note: In the cases of increasing and decreasing the two extreme slices (the largest and the smallest) will always be adjacent since the pie chart is circular.

LAB 3
REPRESENTATION OF DATA BY FREQUENCY TABLES

Grouping Data

By suitably organizing data, we can often make a large and complicated set of data more compact and easier to understand. We will discuss grouping, which involves, as the term implies, putting data into groups rather than treating each piece of data individually. Grouping is one of the most common methods for organizing data.

Open the Minitab project *Lab3.mpj* from your floppy disk.

1. Single-value (Simple) Frequency Table (SFT)

In some cases it is more appropriate to use classes that each represent a single possible numerical value. This is often true for discrete data or qualitative data. Consider the following example;

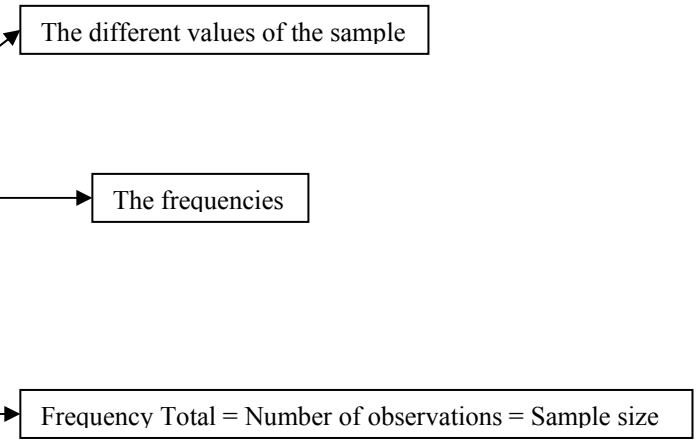
Example 1. To construct an **SFT** for the data Yield from Lab1 we use the Minitab command **Tally**,

Session commands:

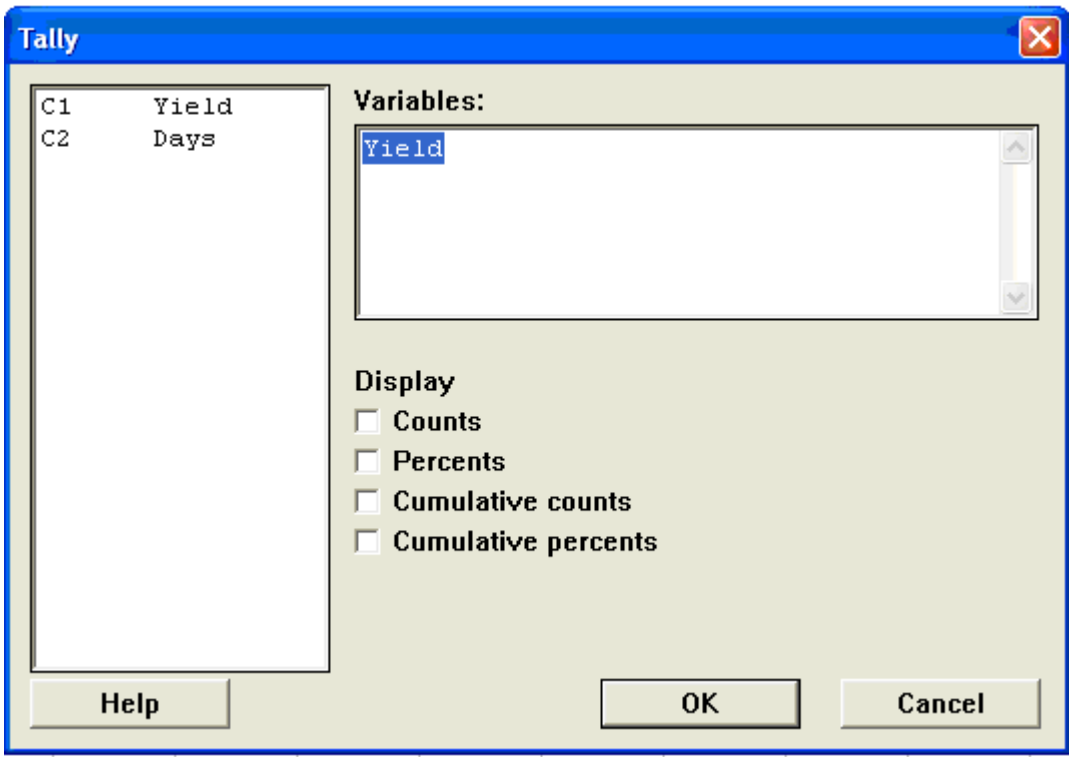
```
MTB > tall c1
```

Tally for Discrete Variables: Yield

Yield	Count
55	1
57	1
58	1
60	2
61	3
62	3
64	1
65	1
67	2
N=	15



Menu commands: **Stat>Tables>Tally…**



Exercise. (IMPORTANT) Construct an SFT for the data Days and copy the resulting columns onto the data window in the columns C2 and C3 respectively. (Hint: use the steps in the bottom of **p.8** in the introduction to the Minitab package to copy columns)

2. Grouped Frequency Table (GFT)

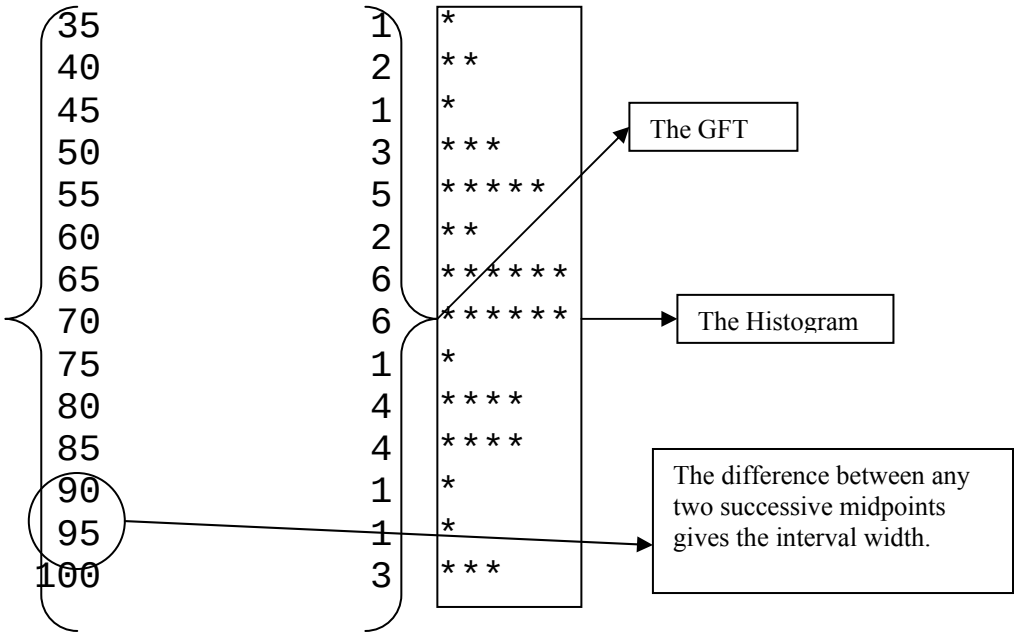
When summarizing a large set of data it is often useful to classify the data into classes or categories and to determine the number of individuals belonging to each class, called the class frequency. A tabular arrangement of data by classes together with the corresponding frequencies is called a frequency distribution or simply a frequency table.
Consider the following terminology,
Classes: Categories for grouping data.
Lower class limit: The smallest value that can go into a class.
Upper class limit: The largest value that can go into a class.
Class mark: The midpoint of a class.
Class width: The difference between the lower class limit of the given class and the lower class linit of the next higher class.
Frequency: The number of pieces of data in a class.
Frequency Distribution: A listing of classes and their frequencies.
Relative frequency: The ratio of the frequency of a class to the total number of pieces of data.
Relative Frequency Distribution: A listing of classes and their relative frequencies.

Cumulative Frequency: The total frequency of all values less than the upper class limit.
Cumulative Frequency distribution: A listing of the upper class limits and their cumulative frequencies.
Relative Cumulative Frequency: Is the cumulative frequency divided by the total frequency.
Relative Cumulative Frequency Distribution: A listing of the upper class limits and their realtive cumulative frequencies.
Frequency Histogram: Consists of a set of rectangles having bases on a horizontal axis with centers at the class marks and lengths equal to the class widths. Areas are proportional to class frequencies.
Example 3. To construct a **GFT** for the data Days we can use the command **Histogram**.
Session commands:

```
MTB > gstd
* NOTE  * Character graphs are obsolete.
* NOTE  * Standard Graphics are enabled.
          Professional Graphics are disabled.
MTB > hist c1
```

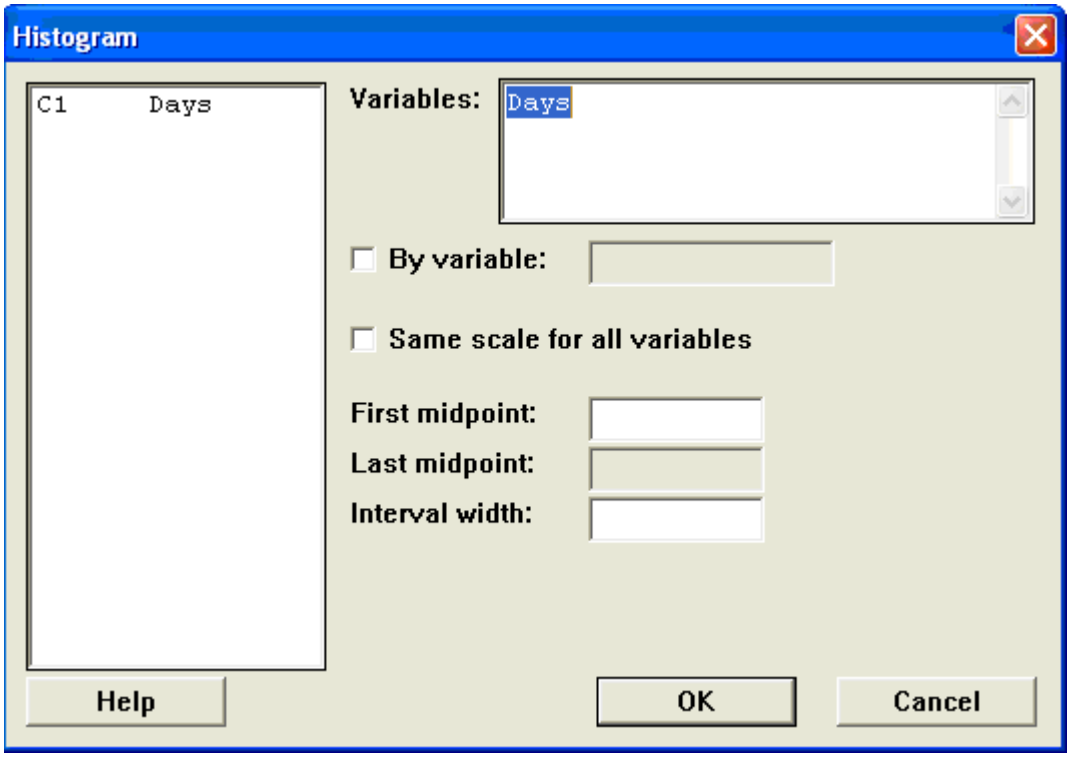
Histogram

```
Histogram of Days   N = 40

Midpoint        Count
      35            1  *
      40            2  **
      45            1  *
      50            3  ***
      55            5  *****
      60            2  **
      65            6  ******
      70            6  ******
      75            1  *
      80            4  ****
      85            4  ****
      90            1  *
      95            1  *
     100            3  ***
```



Note: we can comment in terms of modality and skewness as before.

Menu commands: **Graph>Character Graphs>Histogram…**



You can use the subcommand **Increment** to control, indirectly, the number of classes that you wish to use.
Session commands:

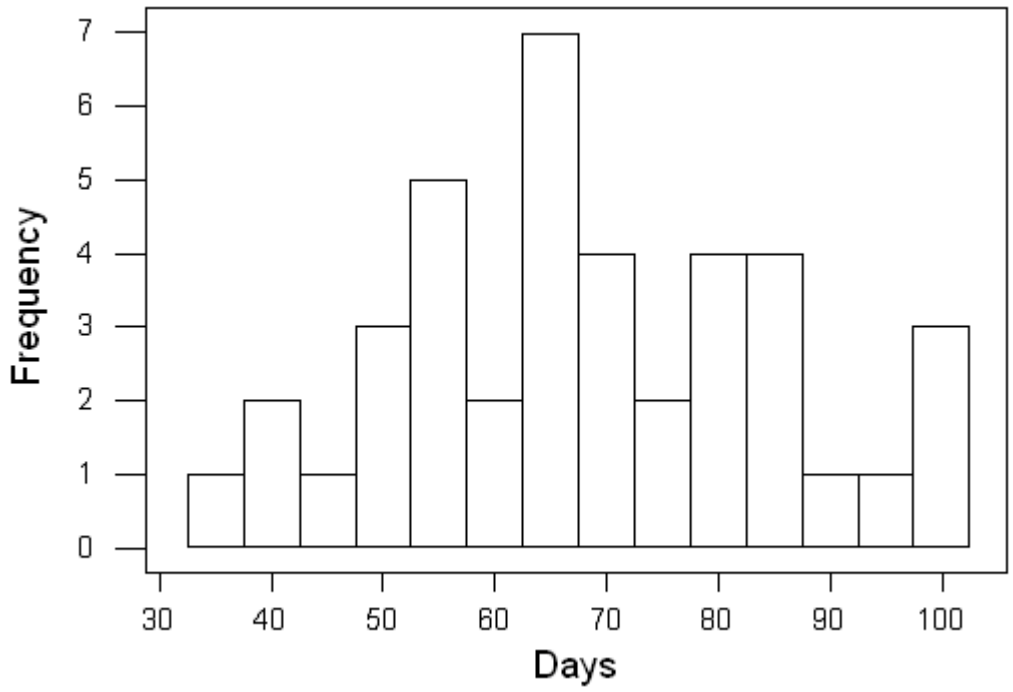
```
MTB > gstd
* NOTE  * Character graphs are obsolete.
MTB > hist c1;
Subc> incr 10.
```

Histogram

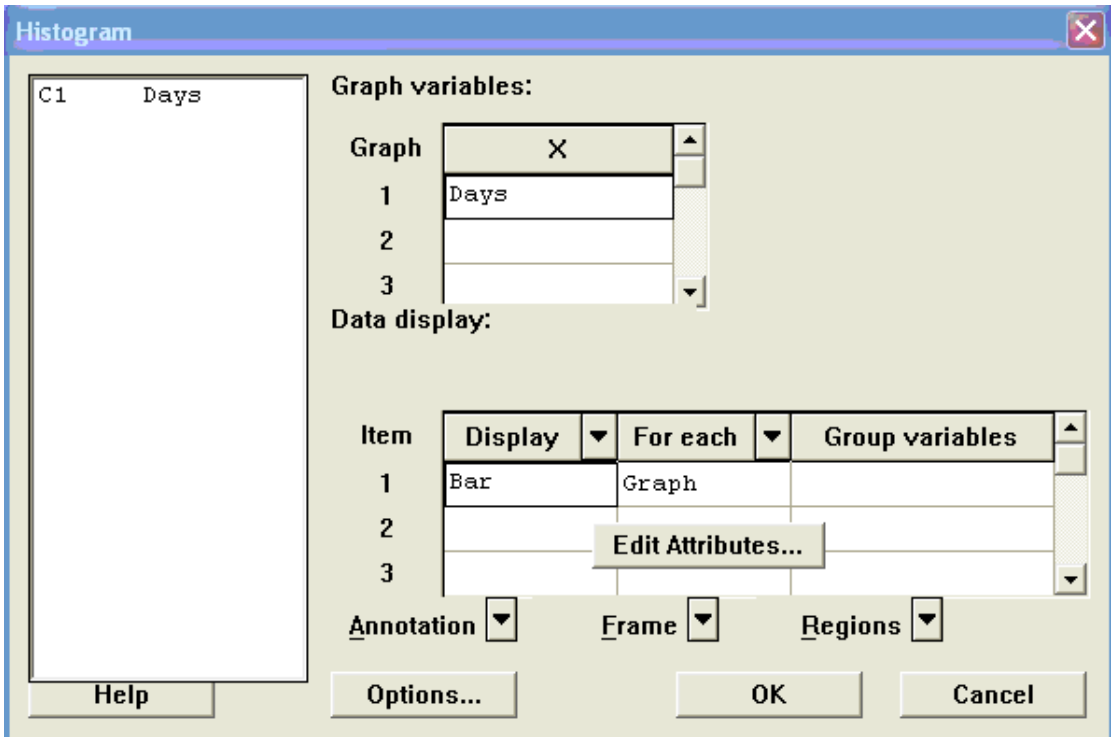
```
Histogram of Days   N = 40

Midpoint        Count
    40.0            3  ***
    50.0            5  *****
    60.0           10  **********
    70.0            7  *******
    80.0            7  *******
    90.0            4  ****
   100.0            4  ****
```

There is another type of histograms called High Resolution Histogram,
Session commands:

```
MTB > GPro.
MTB > Histogram C1;
SUBC> MidPoint;
SUBC> Bar;
```



Menu commands: **Graph> Histogram…**

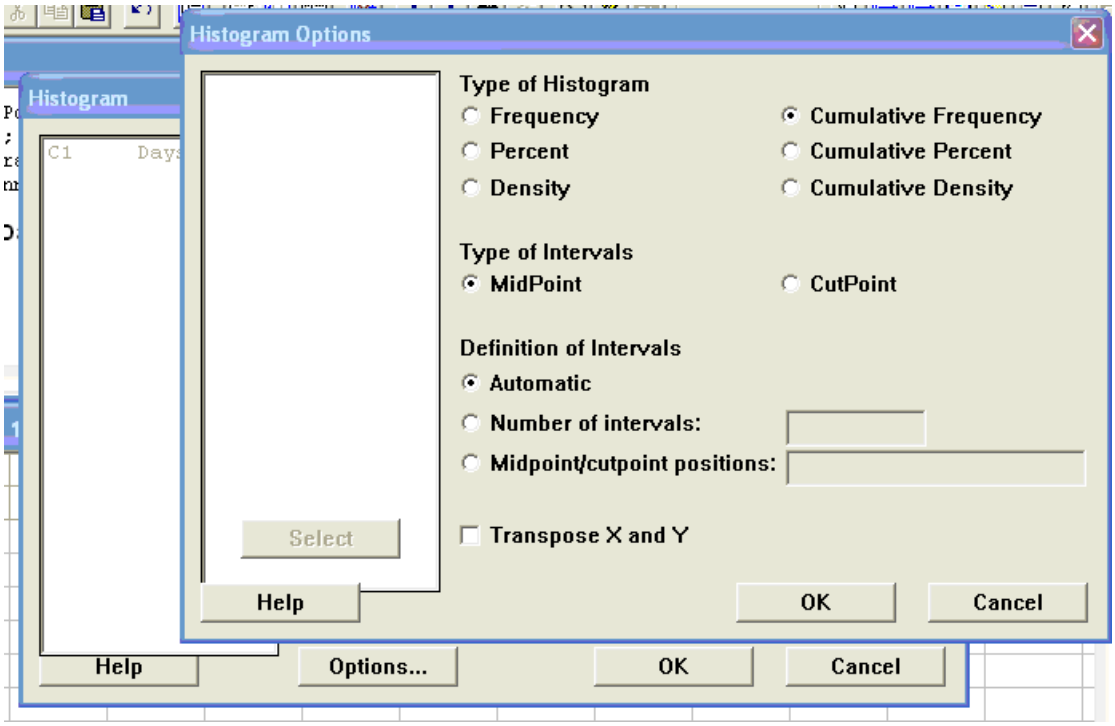


Different types of frequency distributions can be obtained from the high resolution Minitab graphs as follows,

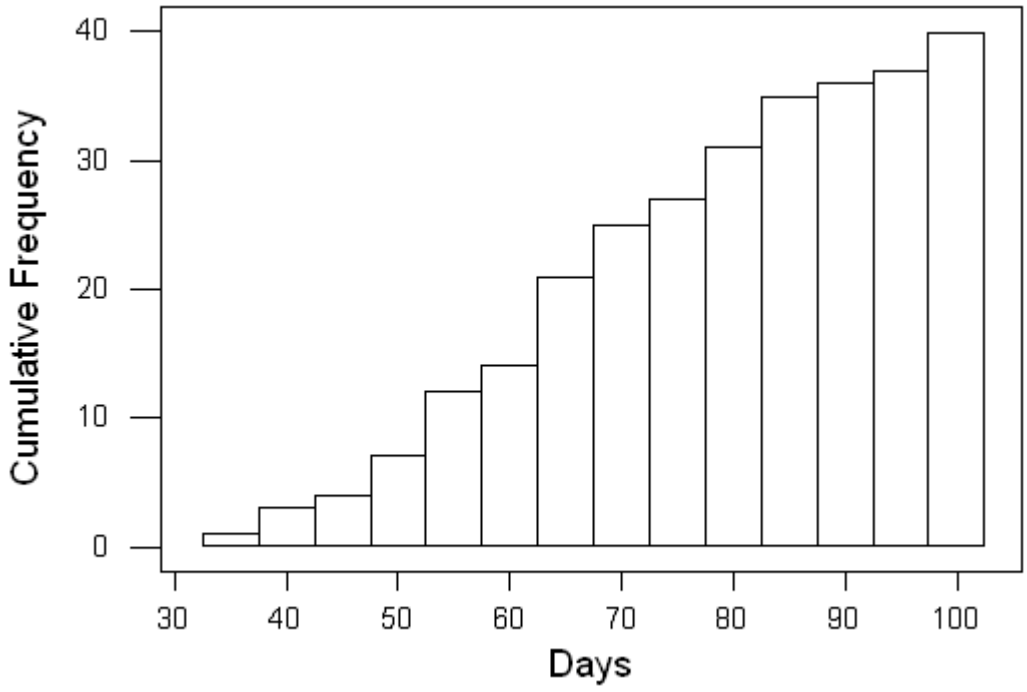
Menu commands: **Graph>Histogram…**

Select : from the dialogue box **Options**

Check: **Type of Histogram** required.



Executing the previous boxes gives;



UNIT II
DESCRIPTIVE MEASURES

LAB 4
SAMPLE MEAN AND STANDARD DEVIATION

Sometimes we are interested in one number which represents the typical number of the data set. In unit I we defined one such number viz., Q_2 which is called also the median of the data set. Another number frequently used as a location measure is the sample mean. Similarly, we defined the range of the sample which gives some ideas about the variation in the sample data. Another measure for dispersion is the sample standard deviation.

It is useful to know that the sample mean and the sample standard deviation are denoted by $\bar{x}$ and s respectively for the sample $x_1, x_2, \ldots, x_n$.

1. Calculation of $\bar{x}$ and s using the sample (raw) data:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad , s = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad , M = \text{Medain} = Q_2$$

where S^2 is called the sample variance.
The sample mean, standard deviation and median are calculated by Minitab using the commands Mean, Stan and Medi respectively.

Open the Minitab project *Lab4.mpj* from your floppy disk.

Example 1. Calculate the sample mean, standard deviation and median for the data stored in C1 named Days:
which are stored in C1,
Session commands:

```
MTB > name k1 'RawMean =' k2 'RawStDev ='
MTB > name k5 'GFTMean =' k6 'GFTStDev ='
MTB >
MTB > Mean 'Days' 'RawMean ='.
```

Mean of Days

```
Mean of Days = 68.300
```

```
MTB > StDev 'Days' 'RawStDev ='.
```

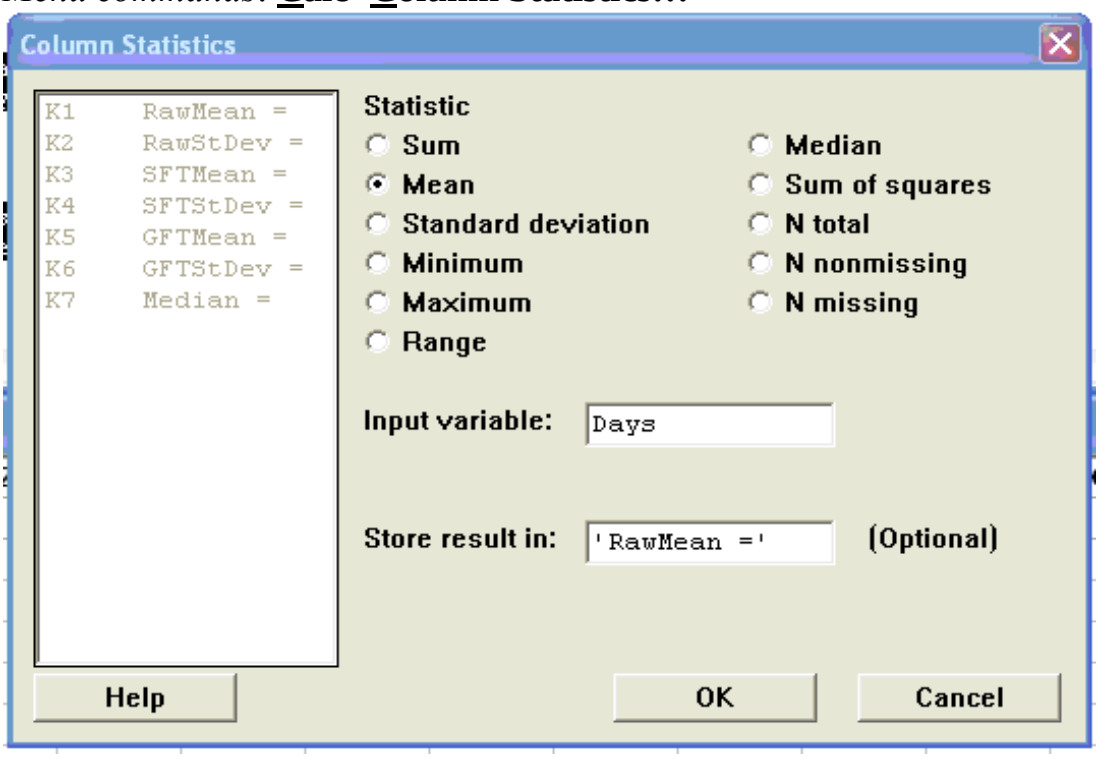
Standard Deviation of Days

```
   Standard deviation of Days = 16.786
MTB > Median 'Days' k7.
```

Median of Days

```
   Median of Days = 66.500
MTB > name k7 'Median ='
```

Menu commands: **Calc>Column Statistics…**



2. Calculation of $\bar{x}$ and s for SFT and GFT:

In case of grouped data,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i \quad ,s = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i} \quad ,n = \sum_{i=1}^{k} f_i \,, \quad k = \text{number of classes}$$

were xi's and fi's are the classmarks and their corresponding frequencies respectively.

Example 2. For the SFT of the data in C1, which is stored in the columns C2 and C3, calculate the mean and the standard deviation.

Note: It is much easier to achieve that using the session commands.

Session commands:

```
MTB > name k3 'SFTMean =' k4 'SFTStDev ='
MTB > let k3=sum(c2*c3)/sum(f)
MTB > let k4=sQrT(sum((x-k3)**2*f)/(sum(f)-1))
```

Example 3. For the GFT of the data in C1, which is stored in the columns C3 and C5, calculate the mean and the standard deviation.

Session commands:

```
MTB > name k5 'GFTMean =' k6 'GFTStDev ='
MTB > let k5=sum(c4*c5)/sum(fi)
MTB > let k6=sqrt(sum((xi-k5)**2*fi)/(sum(fi)-1))
MTB >
MTB > print k1-k6
```

Data Display

```
RawMean =     68.3000
RawStDev =    16.7855
SFTMean =     68.3000
SFTStDev =    16.7855
GFTMean =     69.5000
GFTStDev =    17.2389
```

Note: The raw measures is identical to the SFT ones, where the GFT measures differ slightly but obvious.

3. Descriptive Statistics:

To give more than one statistic for one or more columns at a time we can use the command **Desc**, and to store them in a column for each measurement we can use the command **Stat**;

Session commands:

```
MTB > Desc 'Days'-'fi'.
```

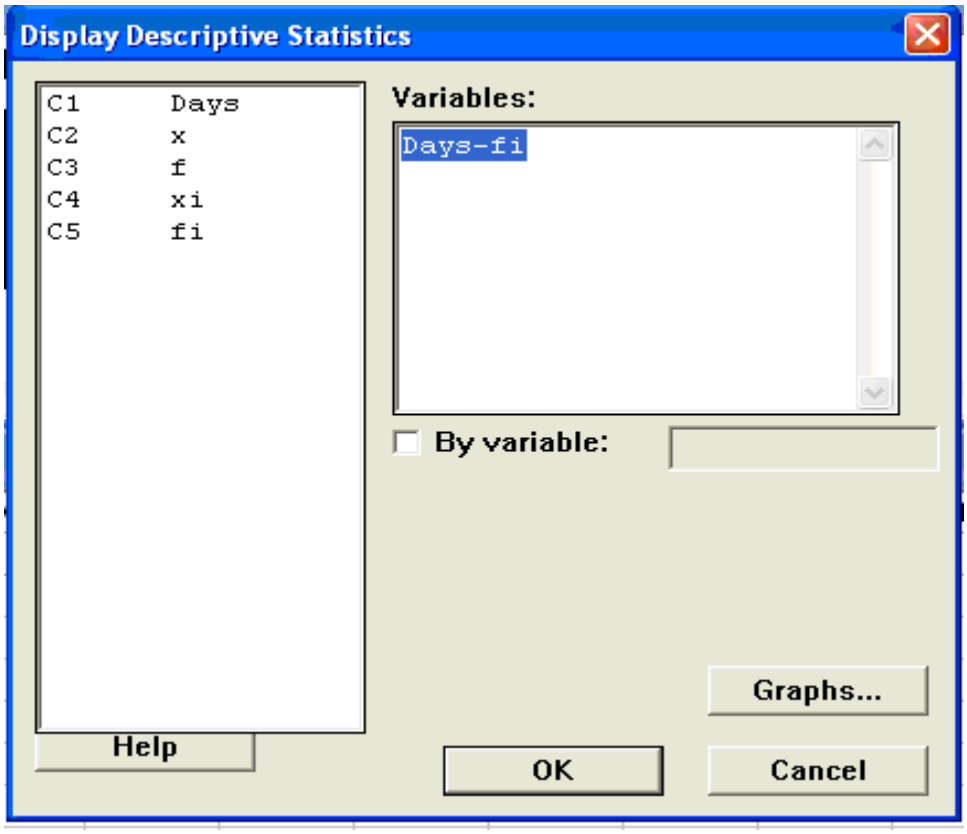
Descriptive Statistics: Days, x, f, xi, fi

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Days	40	68.30	66.50	68.33	16.79	2.65
x	35	68.54	68.00	68.65	16.74	2.83
f	35	1.1429	1.0000	1.0645	0.4300	0.0727
xi	7	70.00	70.00	70.00	21.60	8.16
fi	7	5.714	5.000	5.714	2.430	0.918

Variable	Minimum	Maximum	Q1	Q3
Days	36.00	99.00	55.25	80.75
x	36.00	99.00	56.00	81.00
f	1.0000	3.0000	1.0000	1.0000
xi	40.00	100.00	50.00	90.00
fi	3.000	10.000	4.000	7.000

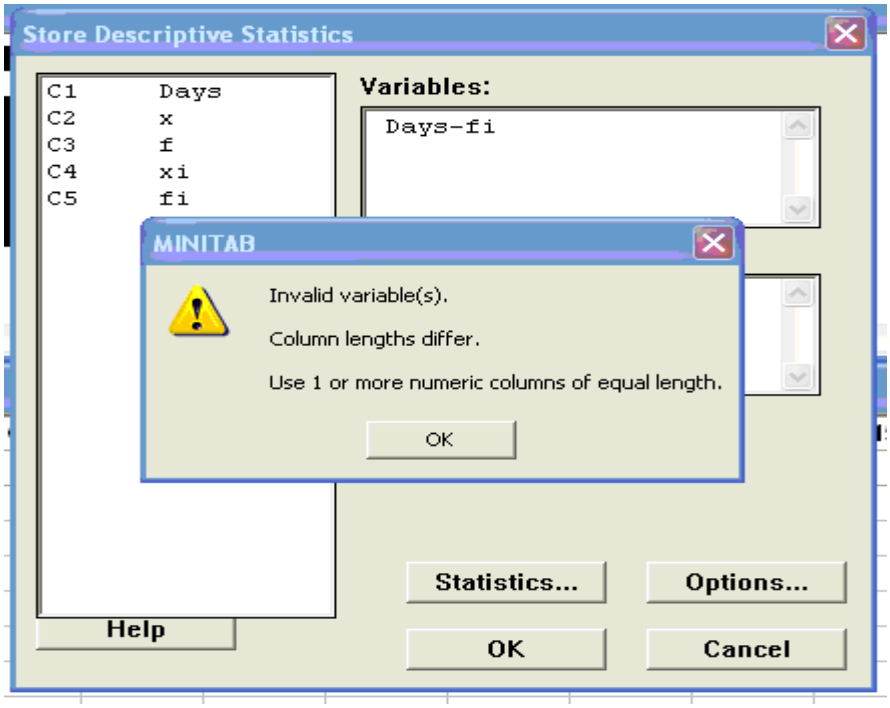
Note: TrMean means Trimmed Mean where Minitab removes the smallest 5% and the largest 5% of the values (rounded to the nearest integer), and then averages the remaining values. The SE Mean will be explained later, in Lab 6.

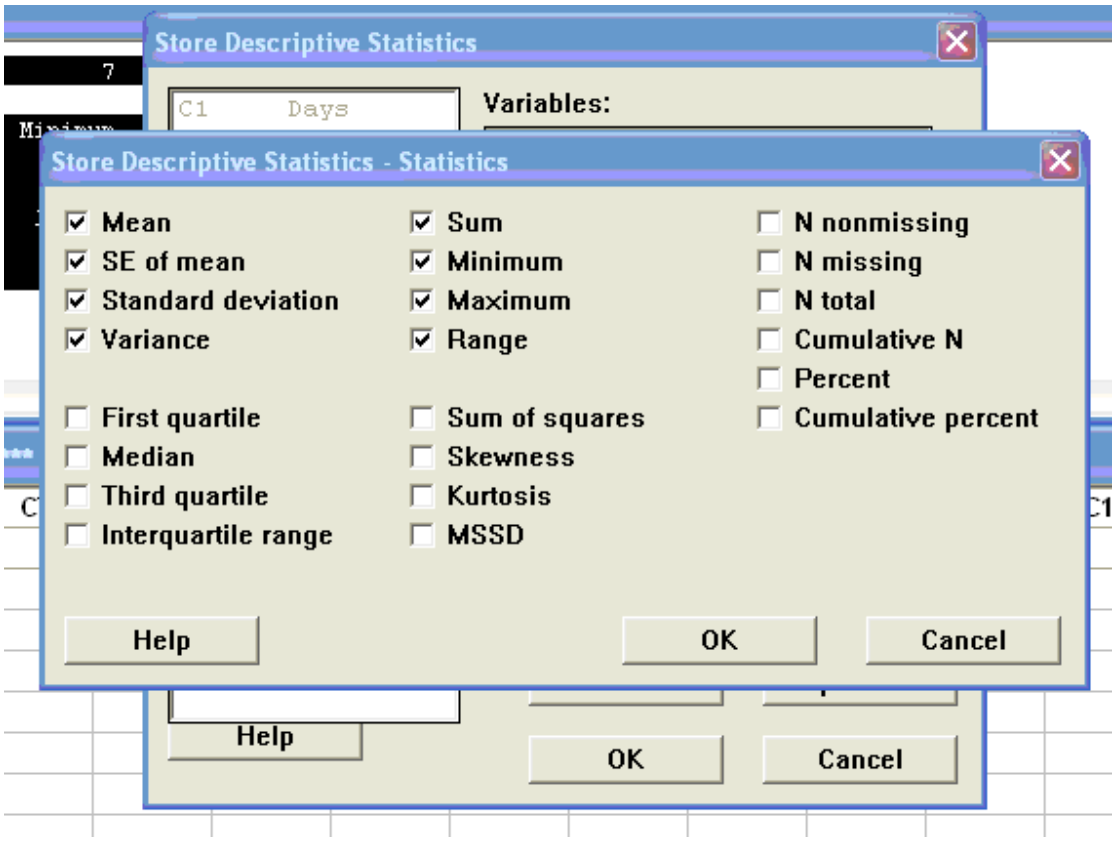
Menu commands: **Stat>Basic Statistics>Display Descriptive Statistics…**



Note: If you want to use the command **Stat** you should select columns of equal length ONLY and it is much easier to use the session commands to execute the **Stat** command than the menu ones;

Menu commands: **Stat>Basic Statistics>Store Descriptive Statistics…**


Store Descriptive Statistics
C1 Days
C2 x
C3 f
C4 xi
C5 fi
Variables:
Days-fi
MINITAB
Invalid variable(s).
Column lengths differ.
Use 1 or more numeric columns of equal length.
OK
Statistics...
Options...
Help
OK
Cancel


Store Descriptive Statistics
C1 Days
Variables:
Store Descriptive Statistics - Statistics
Mean
SE of mean
Standard deviation
Variance
Sum
Minimum
Maximum
Range
N nonmissing
N missing
N total
Cumulative N
Percent
Cumulative percent
First quartile
Median
Third quartile
Interquartile range
Sum of squares
Skewness
Kurtosis
MSSD
Help
OK
Cancel
Help
OK
Cancel

LAB 5
PERCENTILES

Definitions: *Minimum (min.):* Is the smallest sample observation.

Maximum (max.): Is the largest observation of the sample.
Range (R): Is the difference between max. and min.
P^{th} percentile: Is that value of the variable, below which p% of the sample lies.
Quartiles: 3 values (denoted by Q_1, Q_2 and Q_3) within the data divides the data into 4 quarters. That is Q_1 is the value that exceeds 25% of the data and Q_2 is the value exceeds 50% of the data … etc.
Deciles: 9 values (denoted by D_1, D_2 … D_9) within the data divides the data into 10 tenths. That is D_1 is the value that exceeds 10% of the data and D_2 is the value exceeds 20% of the data … etc.
Percentiles: 99 values (denoted by P_1, P_2 … P_{99}) within the data divides the data into 100 hundredths. That is P_1 is the value that exceeds 1% of the data and P_{34} is the value exceeds 34% of the data … etc.

Note that there is a one-to-one correspondence among the quartiles, the deciles and the percentiles;
$Q_1=P_{25}$, $Q_3=P_{75}$, $D_1=P_{10}$, $D_9=P_{90}$, **$Q_2=D_5=P_{50}$=The Median**.

Assume that the sample has observations $x_1,x_2,\ldots,x_n$. The increasingly (ascending) ordered sample is denoted by $x_{(1)},x_{(2)},\ldots x_{(n)}$. The P^{th} percentile is denoted by P_p where the subscript p denotes the *number* of the percentile.

Procedure:

1. Sort: the sample in ascending order.
2. Locate: the p^{th} percentile that is the *rank* of the percentile using the equation $q = \frac{p}{100}(n+1)$, where n>10, n is the sample size and m is the greatest integer less than or equal to q (i.e., $m \le q < m+1$).
3. Calculate: the *value* of the p^{th} percentile can be obtained by interpolating between the values $x_{(m)}$ and $x_{(m+1)}$ according to the formula: $P_p = x_{(m)} + (q-m)\,[x_{(m+1)} - x_{(m)}]$.

If p = 25 , 50 or 75 then the 25^{th} , 50^{th} or 75^{th} percentiles are called the first, the second and the third quartiles, denoted by Q_1, Q_2, and Q_3 respectively.

Open the Minitab project *Lab5.mpj* from your floppy disk.

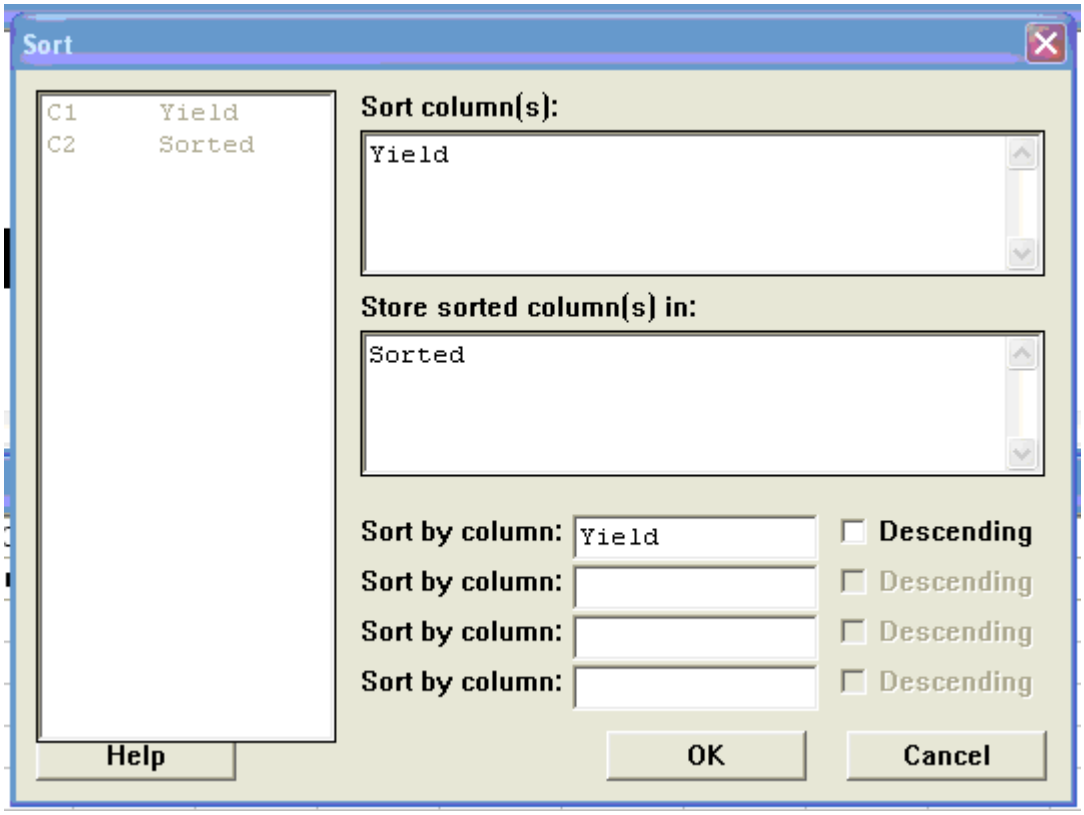
Sorting Data
We can sort the data in C1 increasingly in C2 as follows:
Session commands:

```
MTB > sort c1 c2
MTB > print c2
```

Data Display

```
Sorted
55      57      58      60      60      61      61    61
62      62      62      64      65      67      67
```

Menu commands: **Manip** > **Sort …**



Example 1. To compute the 35th percentile of the data Yield which are stored in C1 and sorted in C2, using Minitab, use only Session commands as follows:

```
MTB > sort c1 c2
MTB > # Calculation of the 35th percentile
MTB > name k1 'q35 =' k2 'P35 ='
```

```
MTB > let k1=35/100*(15+1)
MTB > prin k1
```

Data Display

```
q35 =    5.60000
```

then (the rank) q = 5.6 , m = 5 and m+1 = 6

```
MTB > let k2=c2(5)+(k1-5)*(c2(6)-c2(5))
MTB > print k2
```

Data Display

```
P35 =    60.6000
```

then (the value) P35 = 60.6

Example 2. To compute the 7^{th} decile = 70^{th} percentile of the previous data, do as follows:

```
MTB > # Calculating the 70th percentile
MTB > name k3 'q70 =' k4 'P70 ='
MTB > let k3=70/100*16
MTB > print k3
```

Data Display

```
q70 =    11.2000
MTB > let k4=c2(11)+(k3-11)*(c2(12)-c2(11))
MTB > prin k4
```

Data Display

```
P70 =    62.4000
```

Q_1, Q_2, Q_3, min and max are the main measures of the Five-number summary of the data that can be obtained from the command **Desc**:
Session commands:

```
MTB > desc c1
```

Descriptive Statistics: Yield

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Yield	15	61.467	61.000	61.538	3.378	0.872

Variable	Minimum	Maximum	Q1	Q3
Yield	55.000	67.000	60.000	64.000

UNIT III
PROBABILITY DISTRIBUIONS

LAB 6
DISCRETE AND BINOMIAL DISTRIBUTIONS

Let X be a discrete random variable assumes values $x_1,x_2\ldots,x_n$ with probabilities $p_1,p_2,\ldots p_n$ respectively. Any table, formula or graph gives the different values of the r.v. with their corresponding probabilities is called a probability distribution.

Probability Plot

To construct a probability plot we plot the probabilities of X versus the corresponding values using the command **Plot**.

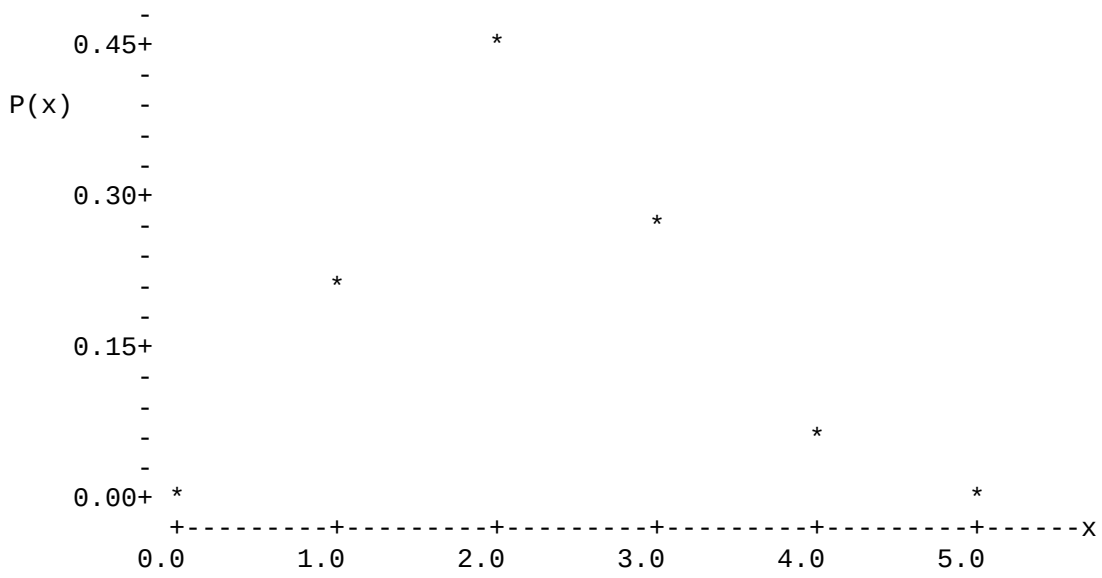
Open the Minitab project *Lab6.mpj* from your floppy disk.

Example 1. Construct a probability plot for the probability distribution of the r.v. X stored in C1 and C2.

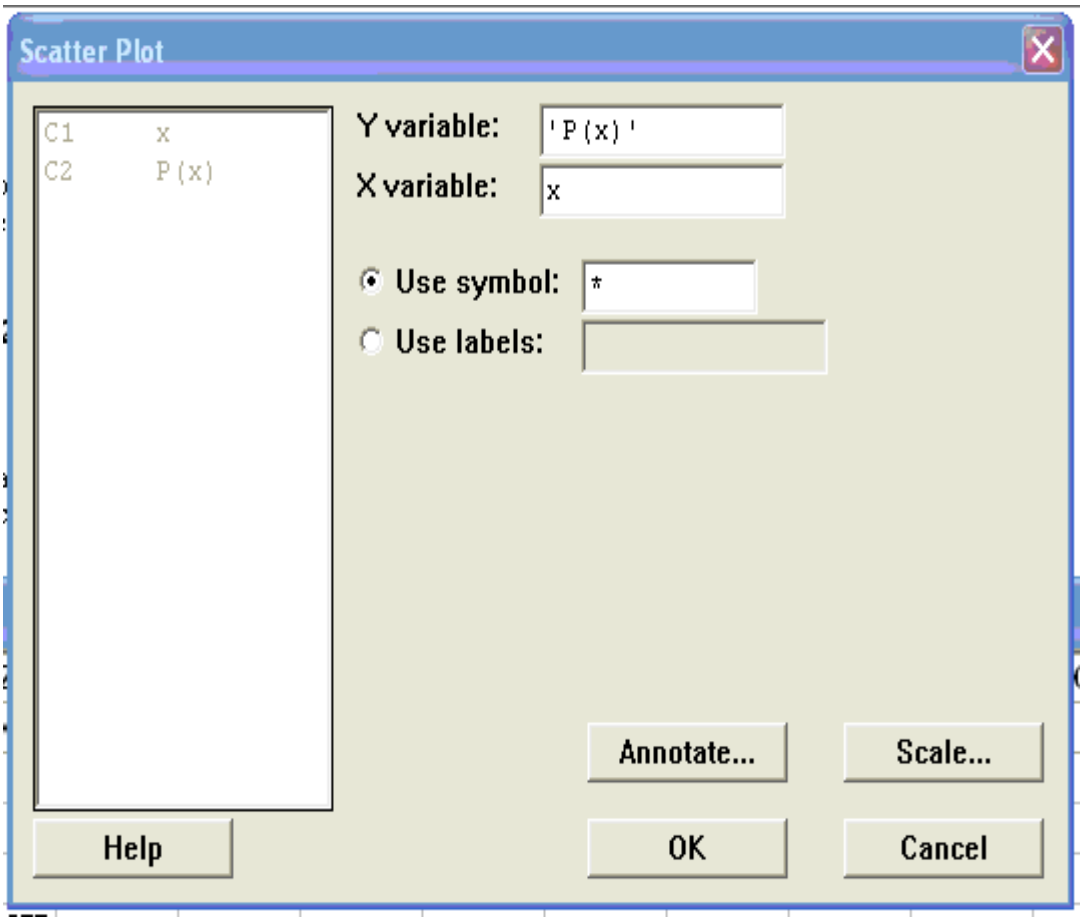
Session commands:

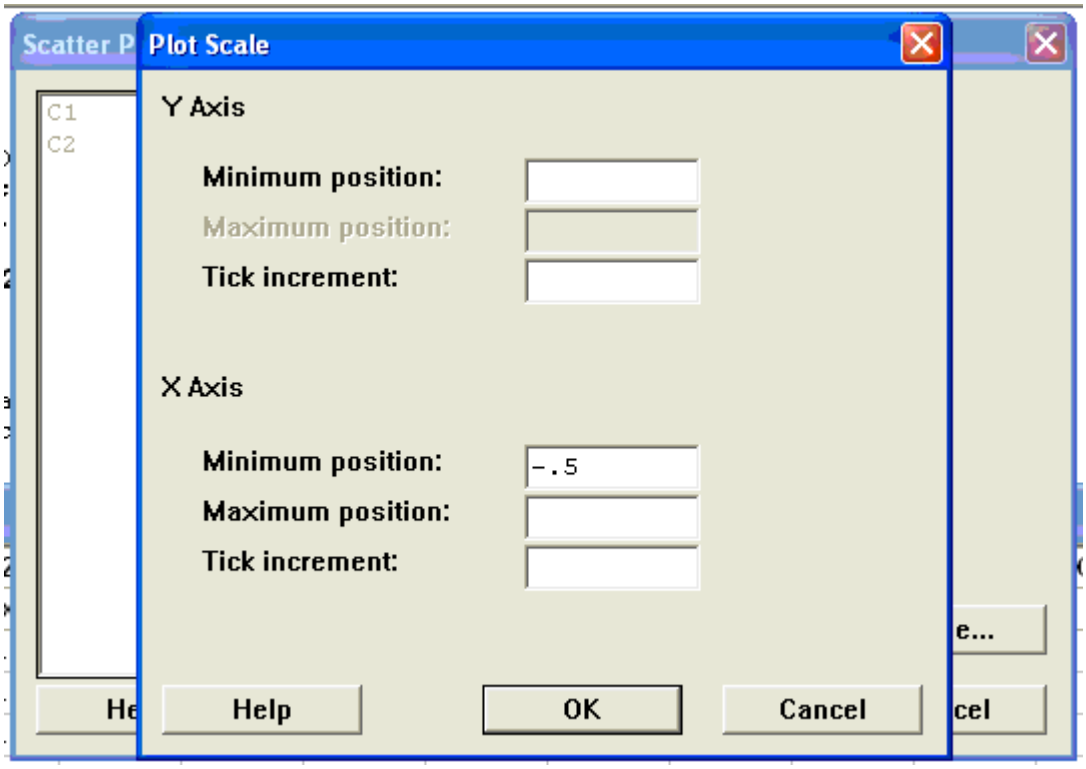
```
MTB > # Probability plot of discrete r.v.
MTB > gstd

* NOTE  * Character graphs are obsolete.

* NOTE  * Standard Graphics are enabled.
          Professional Graphics are disabled.
         Use the GPRO command to enable
         Professional Graphics.
MTB > plot c2 c1
```

Plot

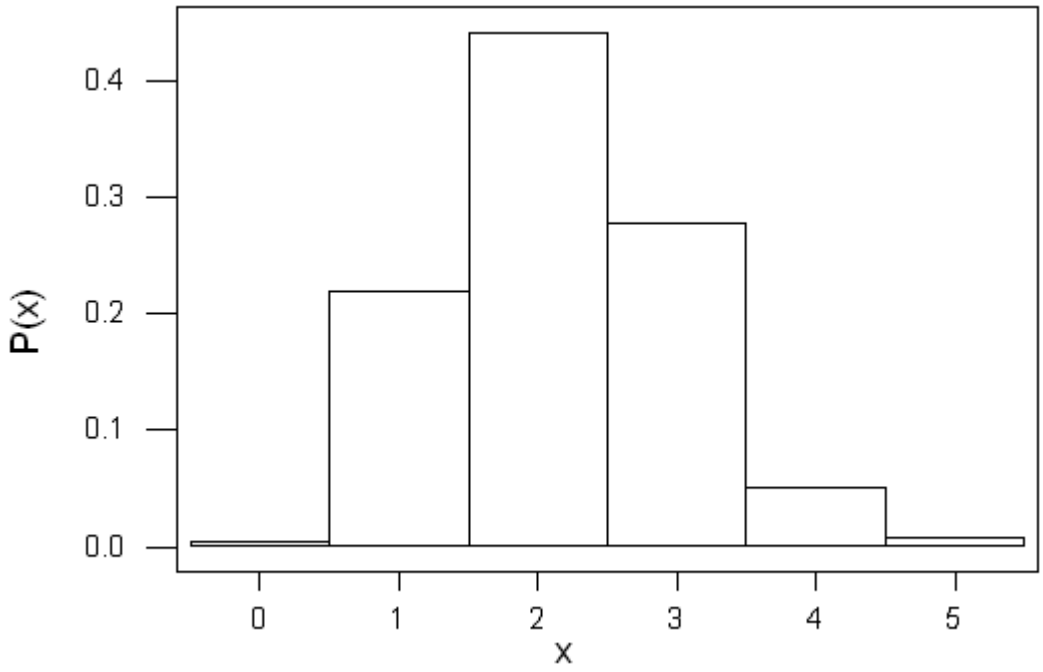


Menu commands: **Graph>Character Graph>Scatter Plot…** and select the option **Scale…** and fill in the *Minimum position* of *X Axis* by -0.5.

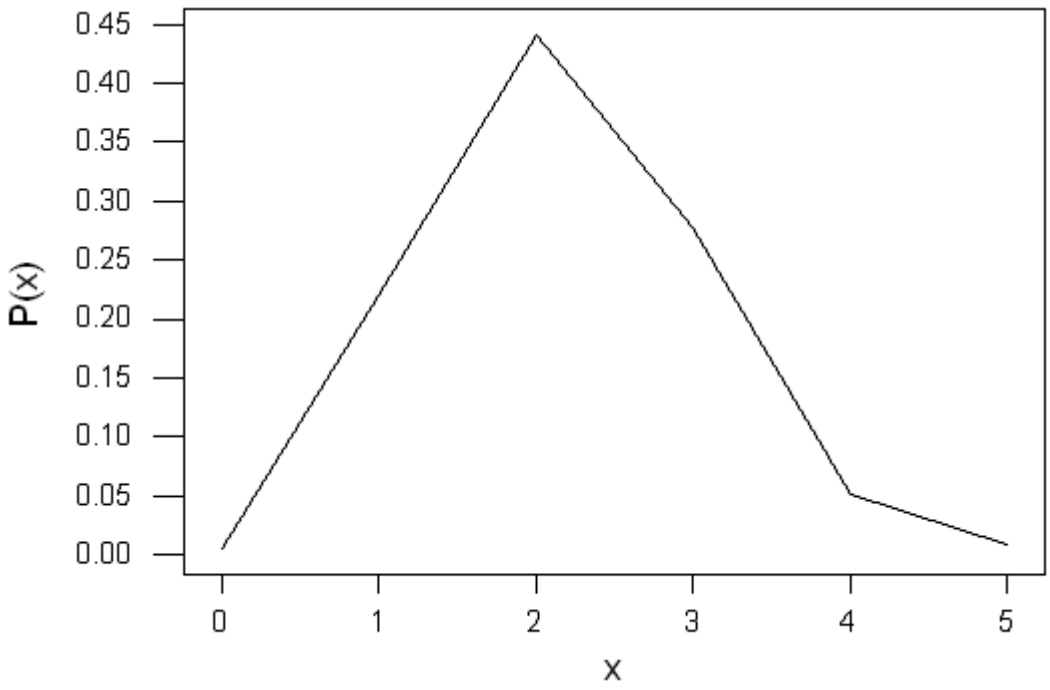




The previous probability plot can be changed into a probability histogram by erecting adjacent bars one over each value of the r.v. with a height equal to its probability as follows;



Also the probability plot can be changed into a probability polygon simply by connecting the asterisks with line segments as follows;



Discrete Random Samples

Minitab has the command **Random** that will simulate observations of a specified random variable. In particular, by employing the **Discrete** subcommand of the Random command, we can generate random observations from any (finite-valued) discrete random variable.

Example 2. For the previous example, generate a r.s. of size 1000 observations from the random variable X, store it in C3 and check weather the generated sample is representative for the r.v. X.

Session commands:

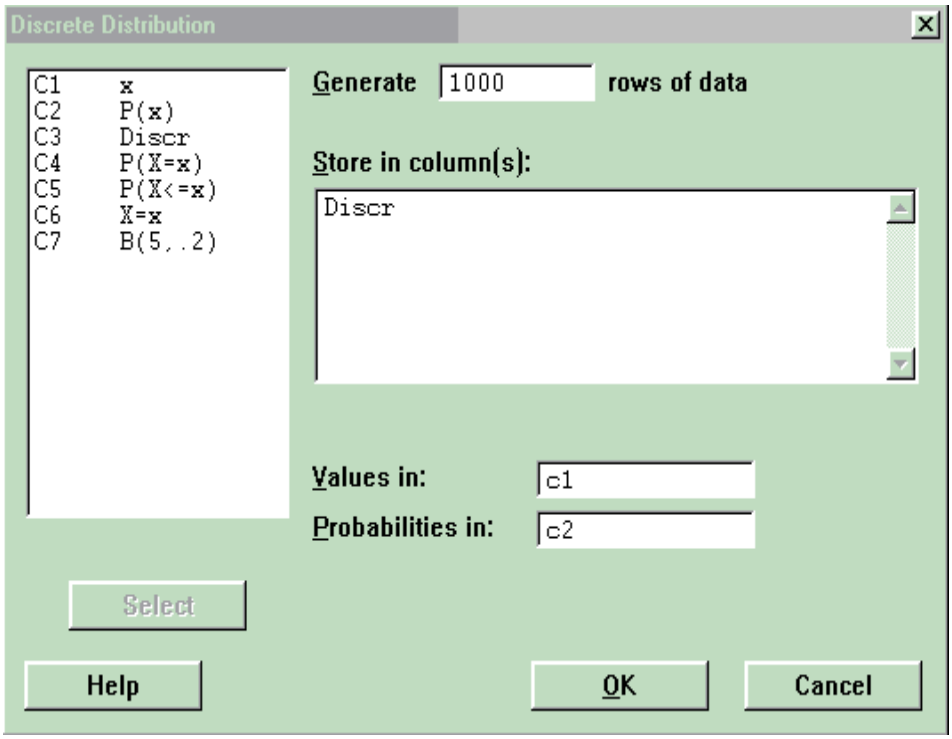
```
MTB > base 0
MTB > rand 1000 c3;
SUBC> disc c1 c2.
MTB > tall c3
```

Tally for Discrete Variables: Discr

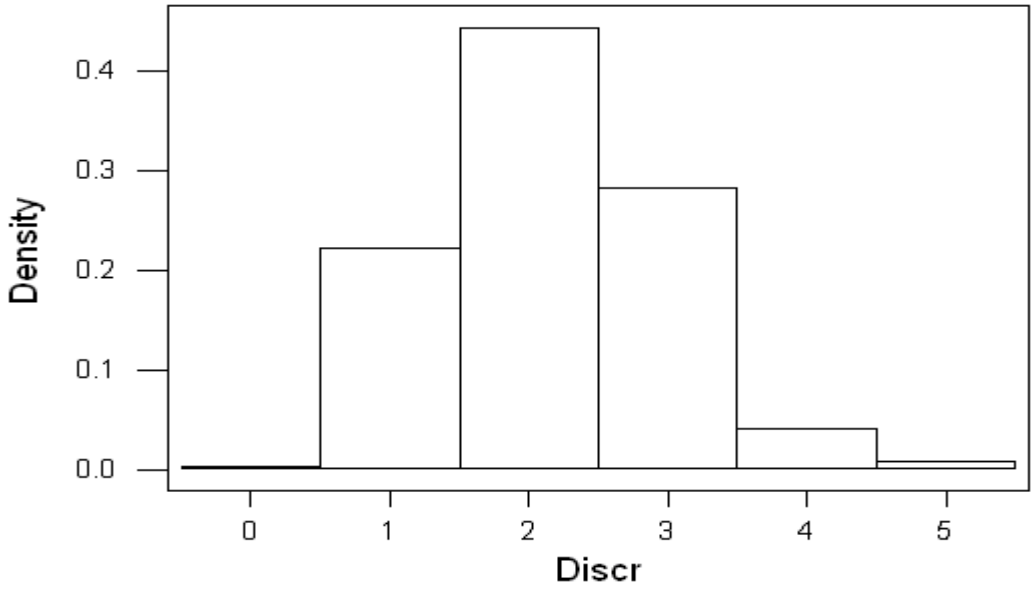
```
Discr  Count
    0      3
    1    222
    2    445
    3    283
    4     40
    5      7
  N=   1000
```

Note: constructing an SFT for the sample is a useful tool to check, roughly, the goodness of fit of the sample to the probability distribution. This can be achieved by simply comparing the relative frequencies in the sample by the probabilities in the distribution.

Menu commands: **Calc>Random Data>Discrete…**



Note: by comparing the relative frequencies in the SFT by the probabilities in the distribution it seems that the generated r.s. is a representative (good) one.



The Binomial Distribution

The random variable X is said to have a binomial distribution with parameters n, p, denoted by **X:B(n, p)**, if its probability law is given by

$$P(x) = f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad , \mathrm{x} = 0,1,2,...,\mathrm{n}$$

where the integer n denotes the number of trials and the fraction p denotes the probability of success, are both positive.

Calculating the Probabilities

To calculate the Bionmial probabilities using Minitab, we use the command **PDF** with the subcommand **Binomial**.

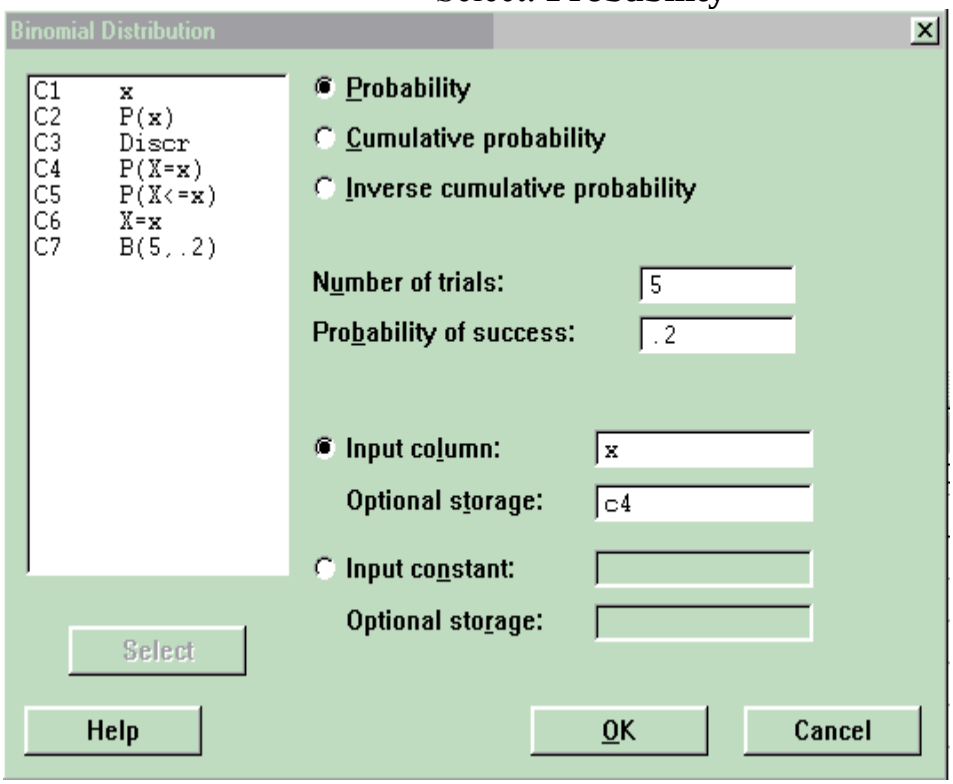
Example 1. For a binomial distribution with n =5 and p=0.2 calculate p(x), where x=0,1,2,….,5.

Session commands:

```
MTB > pdf c1 c4;
SUBC> bino 5 .2.
```

Menu commands: **Calc>Probability Distribution>Binomial…**

Select: **Probability**



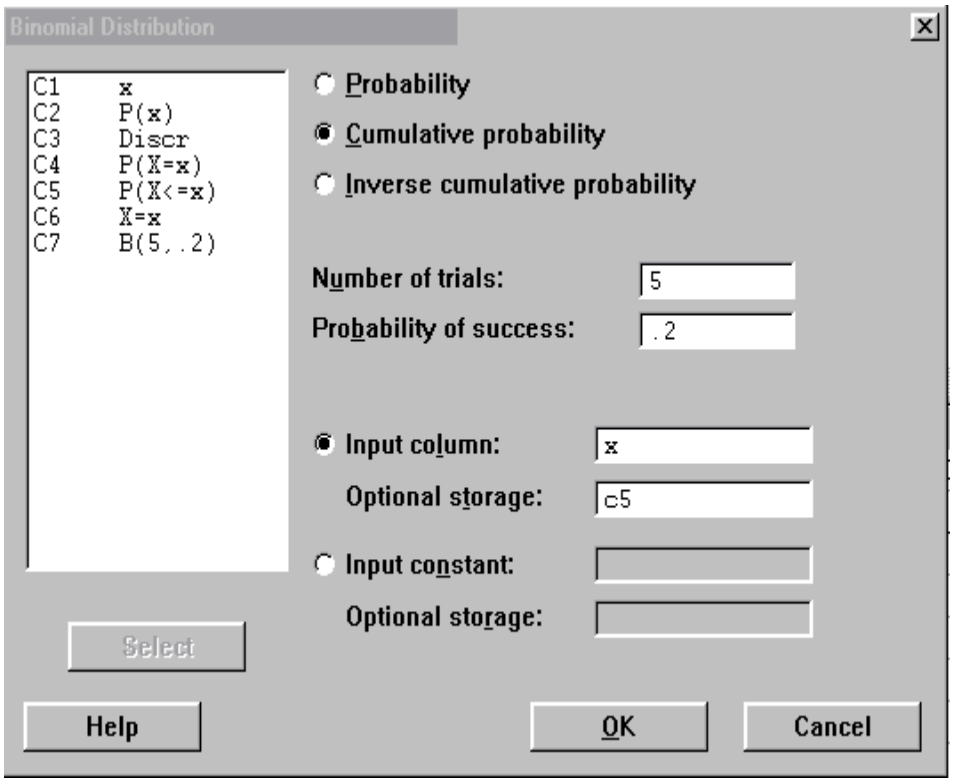
Calculating the Areas

To calculate the areas under the Binomial curve to the left of a point x using Minitab, we use the command **CDF** with the subcommand **Binomial**.

Example 2. From the last example calculate F(x)=P(X≤x), the area to the left (under) x, where x=0,1,2,….,5.
Session commands:

```
MTB > cdf c1 c5;
SUBC> bino 5 .2.
```

Menu commands: **Calc>Probability Distribution>Binomial…**
Select: **Cumulative probability**



Finding Points of Given Probabilities

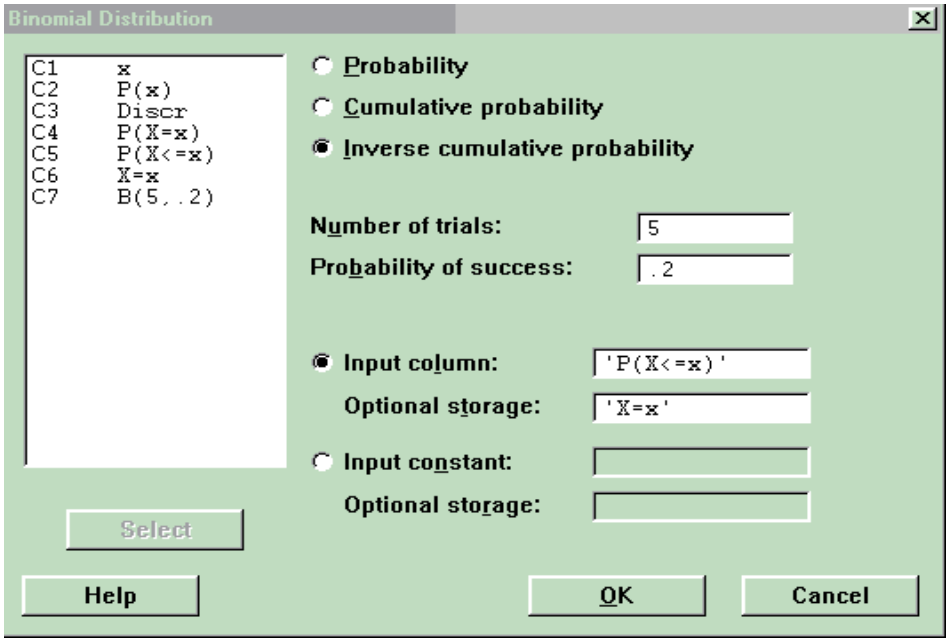
To find value of the Binomial r.v. with given probability to its left using Minitab, we use the command **InvCDF** with the subcommand **Binomial**.

Example 3. From the last example find the values of X that lie areas in column c5 to its left.

Session commands:

```
MTB > invcdf c5 c6;
SUBC> bino 5 .2.
```

Menu commands: **Calc>Probability Distribution>Binomial…**
Select: **Inverse cumulative probability**



Generating a Binomial r.s.

Example 4. Generate 1000 random observations from the random variable X, where X:B(5,.2), store it in c7 and check the goodness of fit of the sample to the distribution.

Session commands:

```
MTB > base 0
MTB > rand 1000 c7;
SUBC> bino 5 .2.
MTB > tall c7
```

Tally for Discrete Variables: B(5,.2)

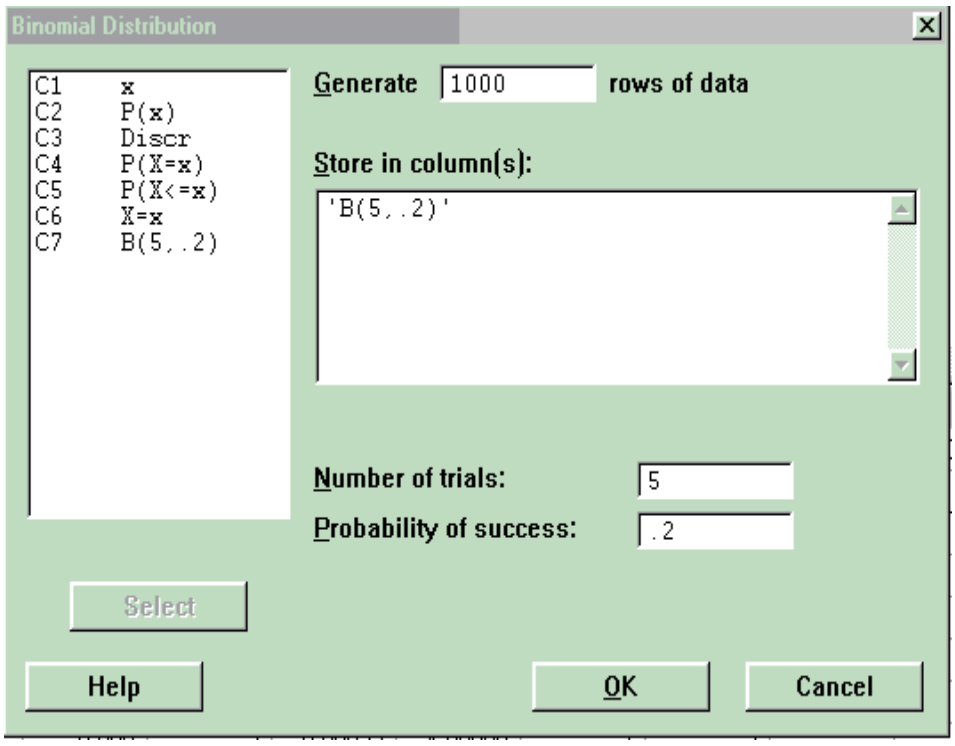
```
B(5,.2)  Count
      0    325
      1    408
      2    205
      3     55
      4      7
     N=   1000
```

```
MTB > print c1-c2 c4-c6
```

Data Display

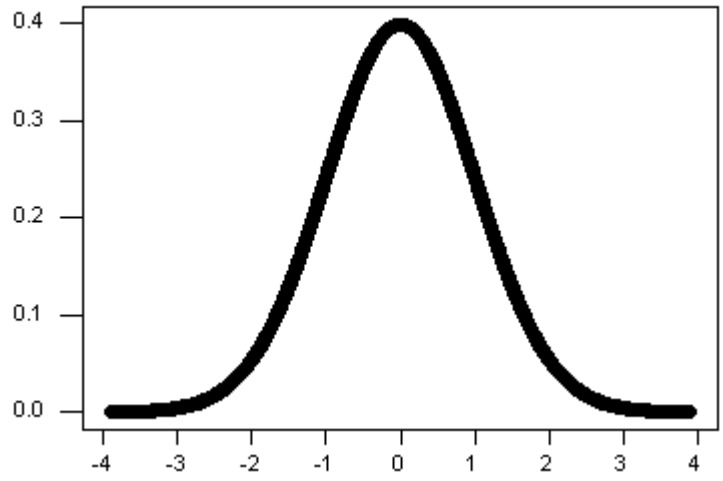
Row	x	P(x)	P(X=x)	P(X<=x)	X=x
1	0	0.004	0.32768	0.32768	0
2	1	0.219	0.40960	0.73728	1
3	2	0.442	0.20480	0.94208	2
4	3	0.277	0.05120	0.99328	3
5	4	0.050	0.00640	0.99968	4
6	5	0.008	0.00032	1.00000	5

Menu commands: **Calc>Random Data>Binomial…**



LAB 7
NORMAL DISTRIBUTION

In the world around us, we observe a variety of populations and random variables. Many are intrinsically different. But some such as aptitude test scores, heights of women and yield weight share an important characteristic: the probabilities associated with them are equal, at least approximately, to areas under a normal curve, that is, a bell-shaped curve like the one shown below.



The normal distribution has two parameters, μ and σ, for the mean and standard deviation of the normal distribution, denoted by X:N(μ, σ^2). The normal curve has the equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^3}} \quad , -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

Basic properties of the normal curve;

1. The curve is symmetric about the vertical line x=μ.
2. The area under the curve is unity.
3. Small values of σ make the curve peaked, where large values make it flat. The curve approaches the horizontal axis as x approaches ± ∞.
4. Most of the area under the curve lies between μ-3σ, μ+3σ.

Open the Minitab project *Lab7.mpj* from your floppy disk.

Calculating Densities Under the Normal Curve

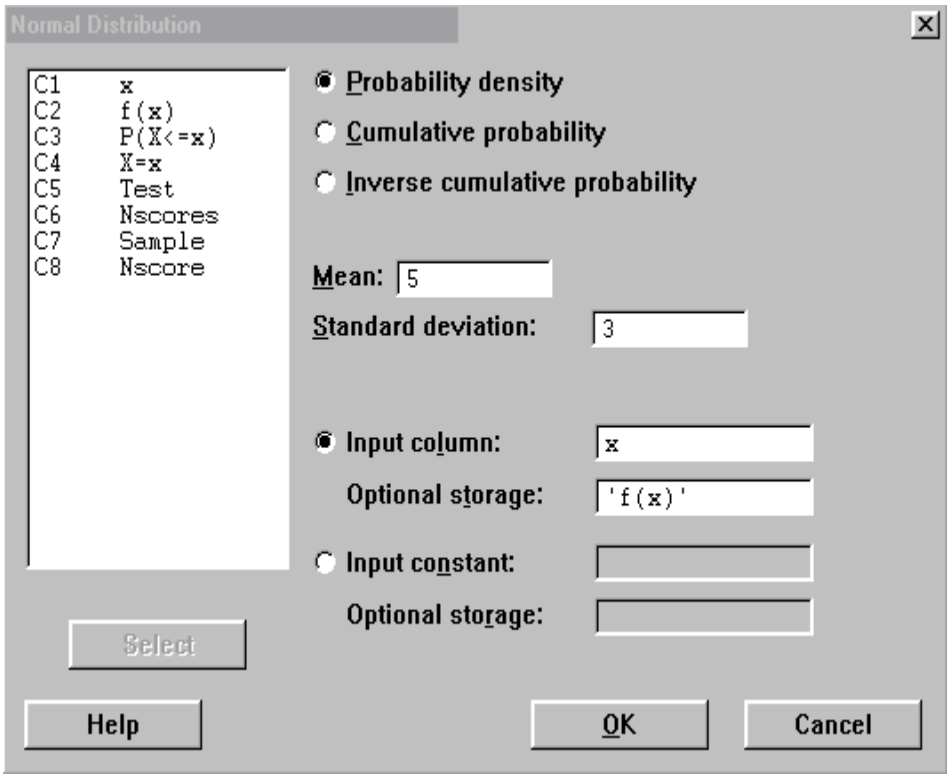
The densities under the normal curve can be computed using the following Minitab command **PDF**.

Example 1. To calculate the density under a normal curve with $\mu=5$, $\sigma=3$ at each of the points in C1.

Session commands:

```
MTB > # Calculation of f(x).
MTB > pdf c1 c2;
SUBC> norm 5 3.
```

Menu commands: **Calc>Probability Distribution>Normal…**
Select: **Probability density**



Calculating Areas Under the Normal Curve

The areas under the normal curve can be computed using the following Minitab command **CDF**.

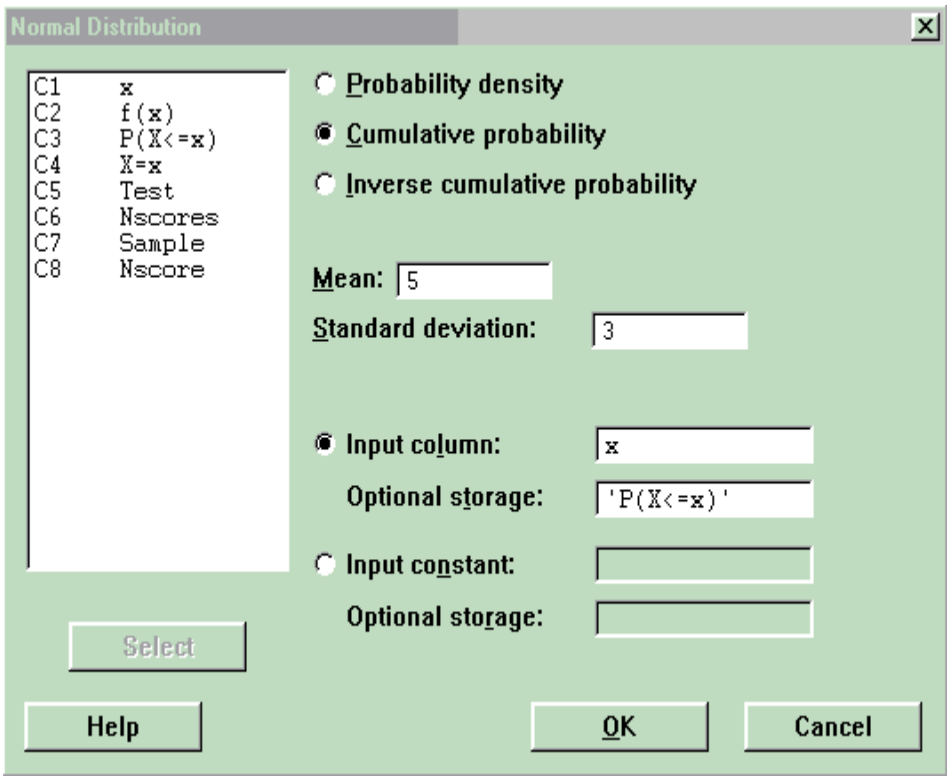
Example 2. To calculate the area under a normal curve with $\mu=5$, $\sigma=3$ that lies to the left of each of the points in C1.

Session command:

```
MTB > # Calculation of F(x).
MTB > cdf c1 c3;
```

```
SUBC> norm 5 3.
```

Menu commands: **Calc>Probability Distributions>Noraml…**
Select: **Cumulative probability**



Finding Points Under the Normal Curve

The values of the normal r.v. can be found, given the areas under the normal curve and to the left of each, using the following Minitab command **InvCDF**.

Example 3. Determine the x-values where X has a normal distribution with $\mu=5$ and $\sigma=3$, that has areas in C3 to its left.

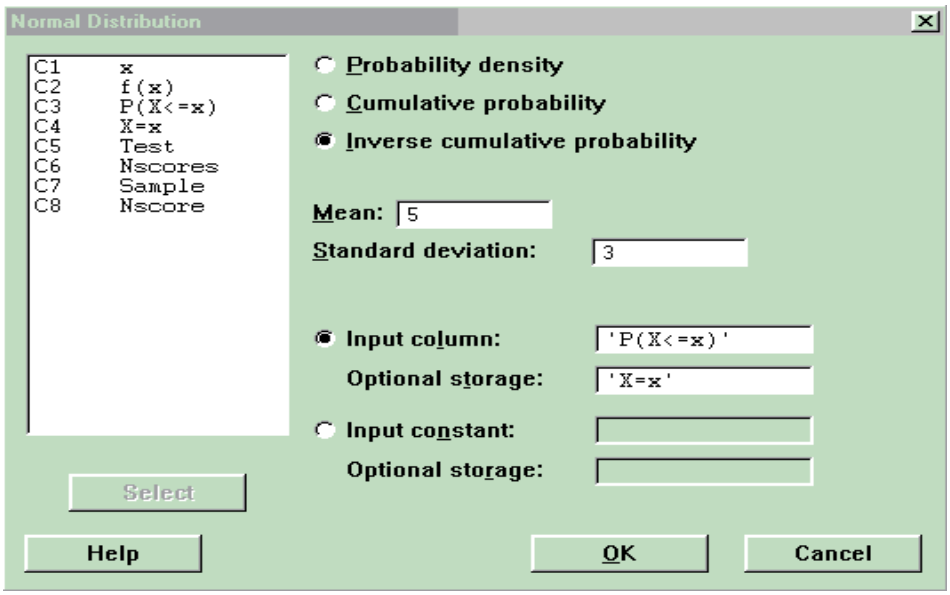
Session commands:

```
MTB > InvCDF  c3 c4;
SUBC>   Noraml 5 3.
MTB >
MTB > print c1-c4
```

Data Display

Row	x	f(x)	P(X<=x)	X=x
1	-4.0	0.001477	0.001350	-4.0000
2	-1.0	0.017997	0.022750	-1.0000
3	0.5	0.043173	0.066807	0.5000
4	2.0	0.080657	0.158655	2.0000
5	5.0	0.132981	0.500000	5.0000
6	8.0	0.080657	0.841345	8.0000
7	9.5	0.043173	0.933193	9.5000
8	11.0	0.017997	0.977250	11.0000
9	14.0	0.001477	0.998650	14.0000

Menu commands: **Calc>Probability Distribution>Normal…**
Select**: Inverse cumulative probability**

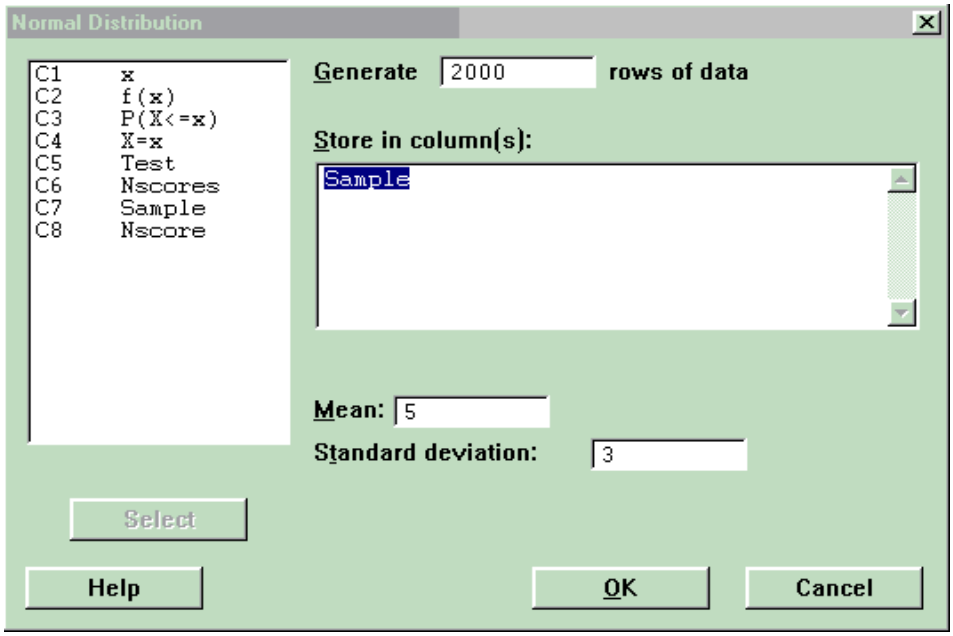


Generating a Normal r.s.

Example 4. Generate a sample of size 2000 from the r.v. X:N(5,9), store it in C7 and check the normality assumption of the sample.
Session commands:

```
MTB > base 0
MTB > rand 1000 c7;
SUBC> norm 5 3.
```

Menu commands: **Calc>Random Data>Normal…**



But to check the normality assumption is the question of the next section and will be revisited after introducing that section.

Normal Probability Plot (NPP)

Normal Probability Plot can be used, roughly, as an aid for deciding whether a sample is drawn from a population whose distribution is (approximately) normal. Minitab has a command called **Nscores** that can be used to determine the normal scores for a data set. We can employ the Minitab's **Plot** command to obtain a Normal Probability Plot.

Procedure for Character NPP:

1. Calculate the normal scores (ZScores) of the sample.
2. Plot the NScores versus the observations of the sample.

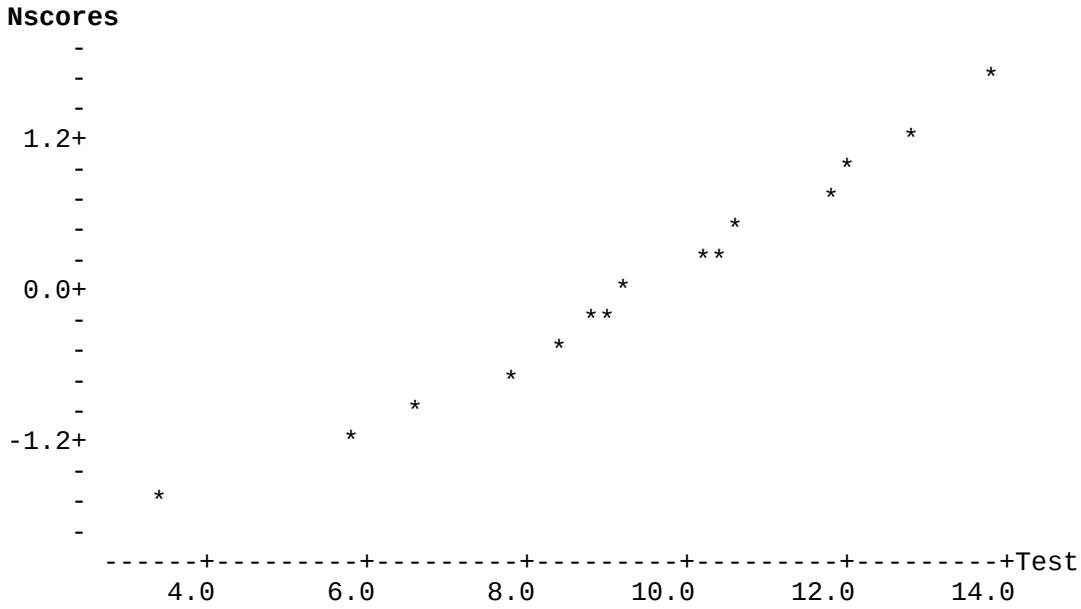
Example 5. A sample of 15 cars yields the following data on miles driven 10.2 9.2 11.8 10.3 13.7 6.6 8.9 7.7 8.7 12.7 3.3 5.7 8.3 10.6 12.0. Do the data imply that the miles driven by all the cars (population) are normally distributed?

This question can be answered by constructing an NPP for the sample stored in C5 using the following commands;

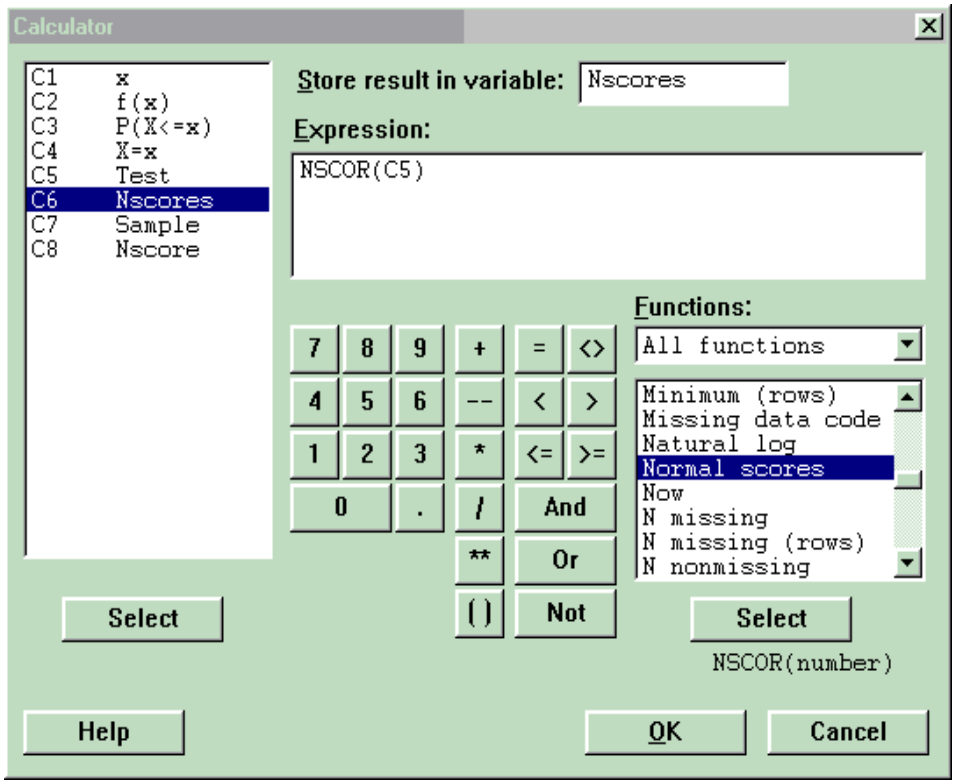
```
MTB > # Constructing an NPP for the Test
MTB > let c6=nsco(c5)
MTB > gstd
* NOTE  * Character graphs are obsolete.
* NOTE  * Standard Graphics are enabled.
           Professional Graphics are disabled.
```

```
            Use the GPRO command to enable
Professional Graphics.
MTB > plot c6 c5
```

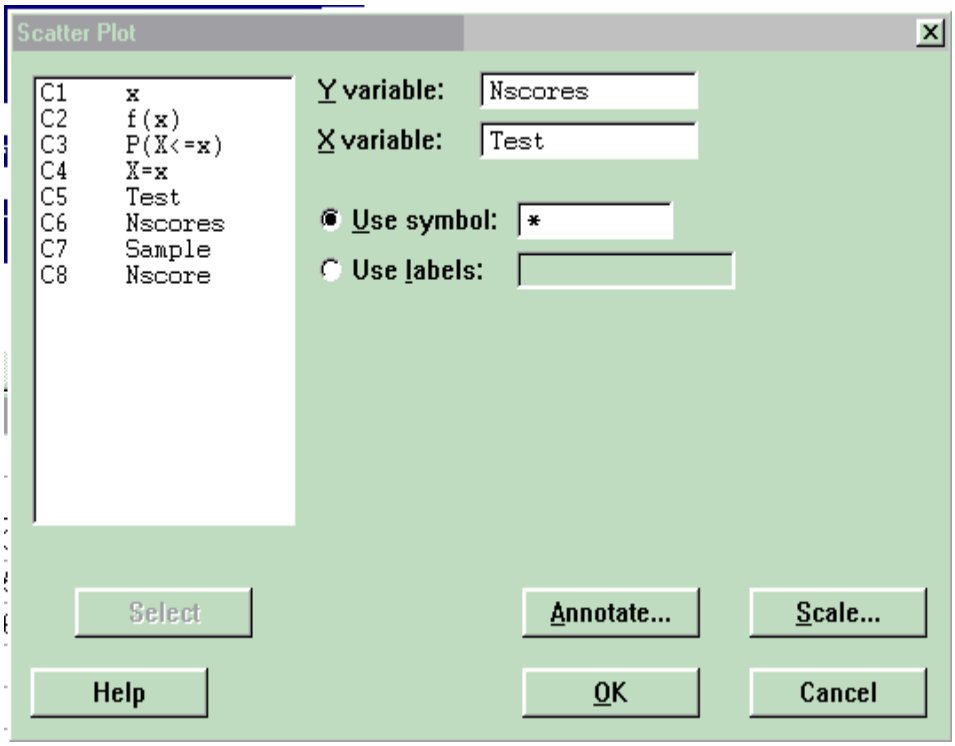
Plot



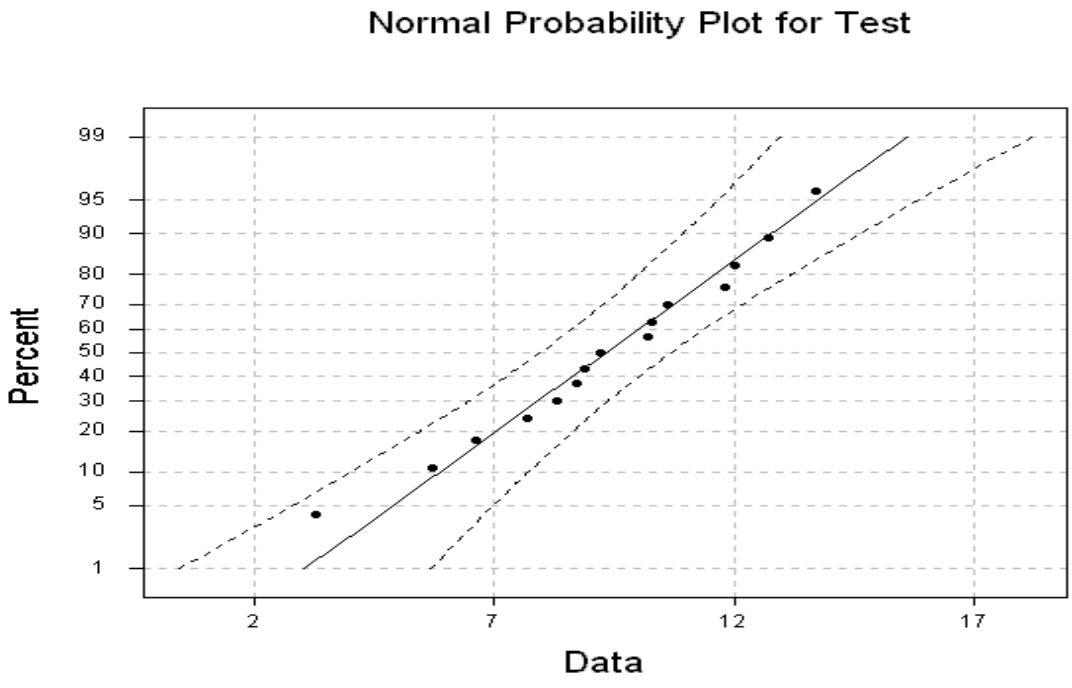
Comment: Since the asterisks formed a straight line approximately, we can conclude that the sample is drawn from a normal population.
Menu commands: **Calc>Calculator…**



Menu commands: **Graph>Character Graph>Scatter Plot…**



There is another effective and numerical method in Minitab which can be used to draw Normal Probability Plot for a given set of data;
Menu commands: **Graph>Probability Plot…**
Select**: Distribution** (Normal)

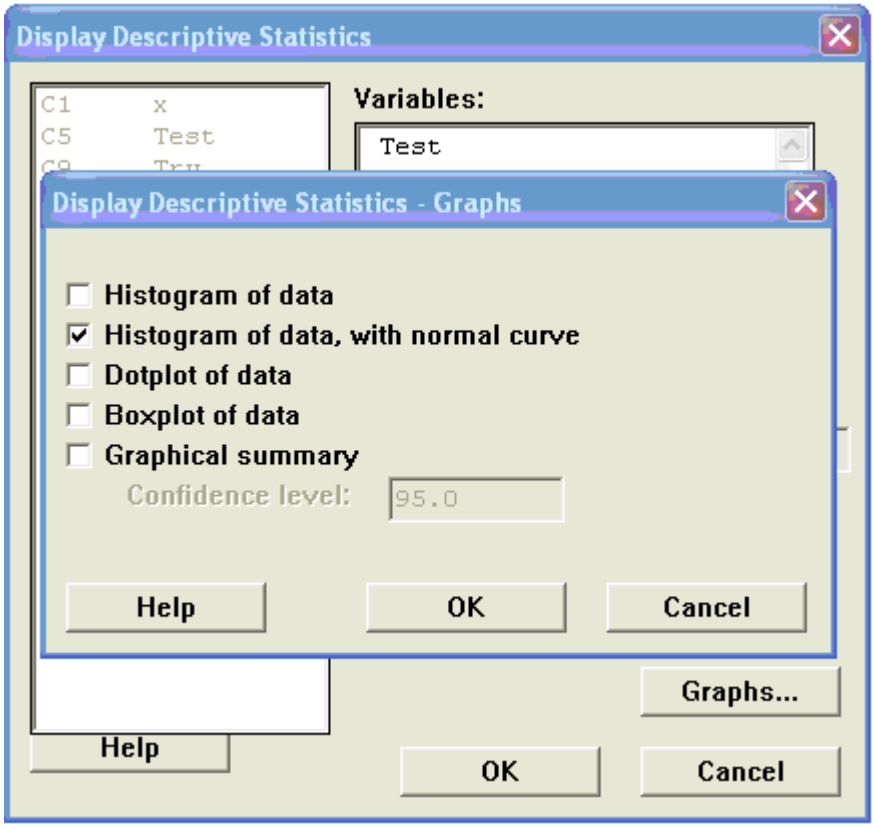


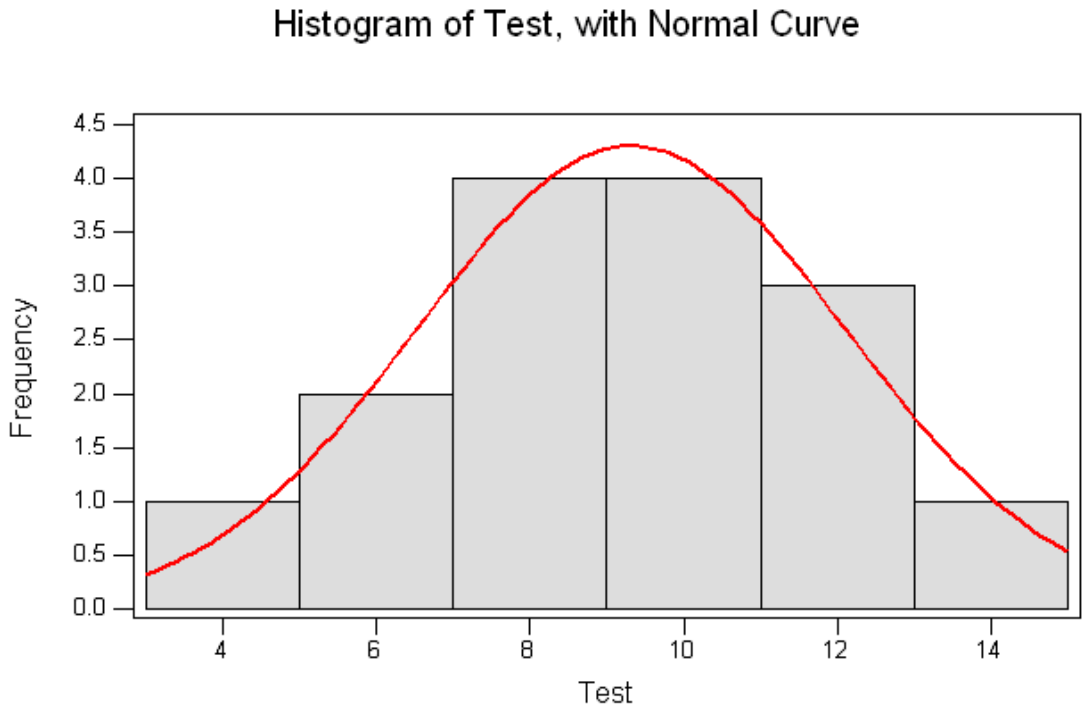
There is also another method in Minitab to check the normality assumption of a given data set.

Menu Commands: **Stat>Basic Statistics>Display Descriptive Statis…**

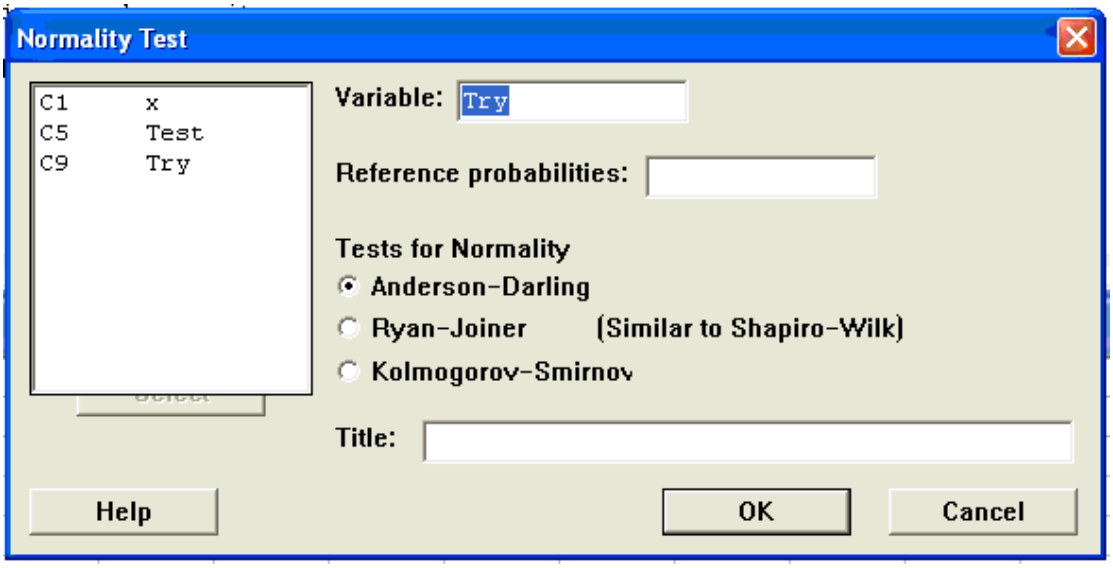
Select **Graph** option and tick on

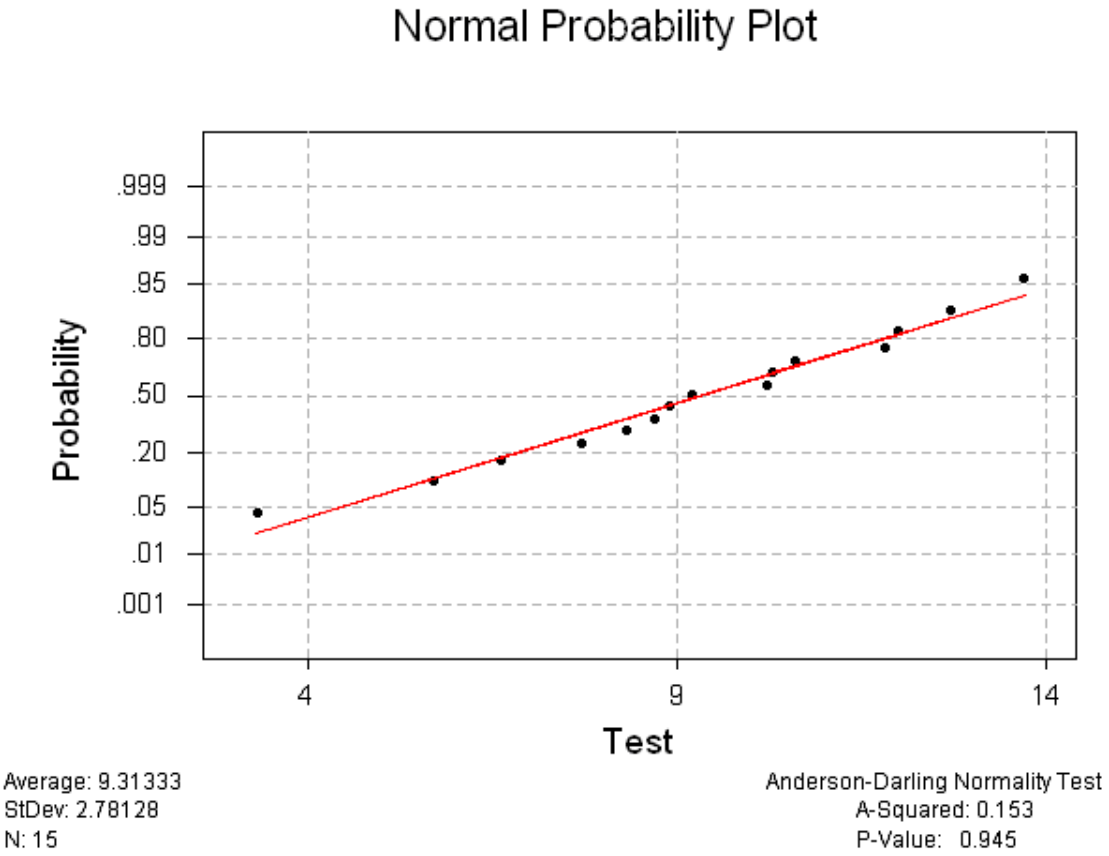
Histogram of data, with normal curve





The last numerical and the multipurpose, uses more than one test statistic, method for checking the normality assumption is the Normality Test given as follows;
Menu commands: **Stat>Basic Statistics>Normality Test…**

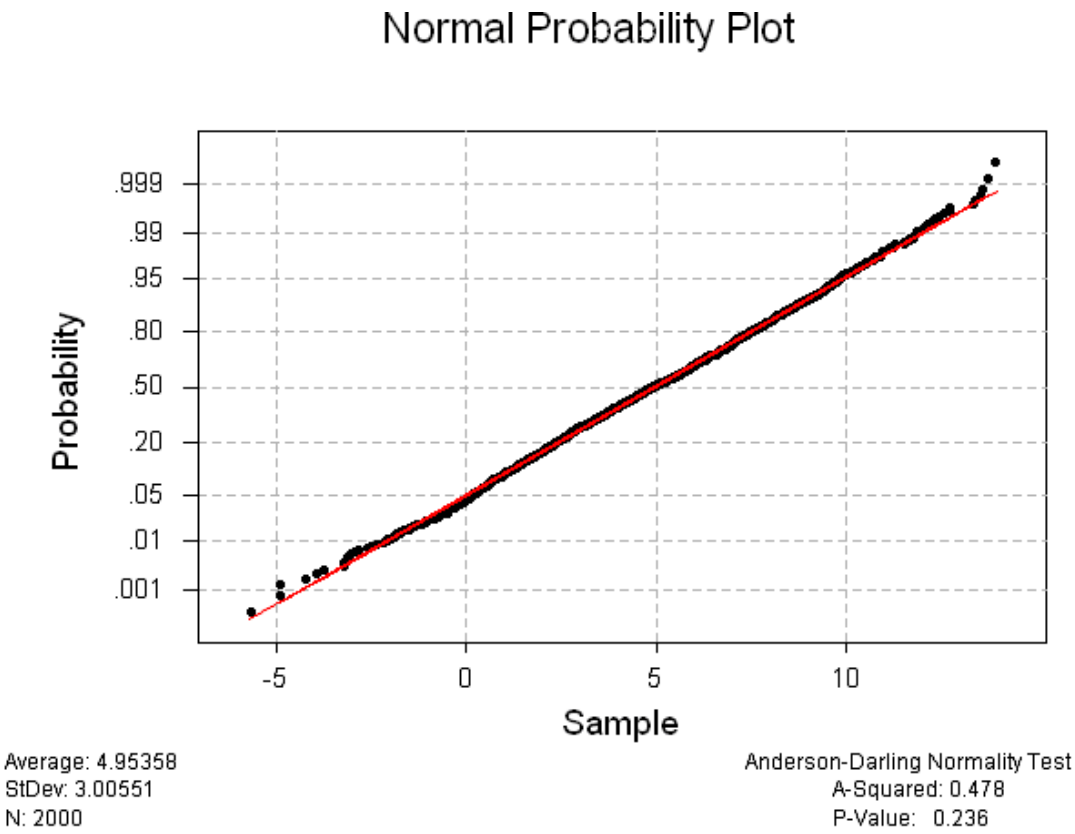




Rule: If P-Value < 0.05 we would Reject that the sample is drawn from a normal population with a confidence level of 95%.

Comment: Since the P-Value = 0.945 >> 0.05 then we conclude that the sample is drawn from a normal population.

Now return back to the generated sample in Example 4 which was named *Sample*, then the normality test gives;

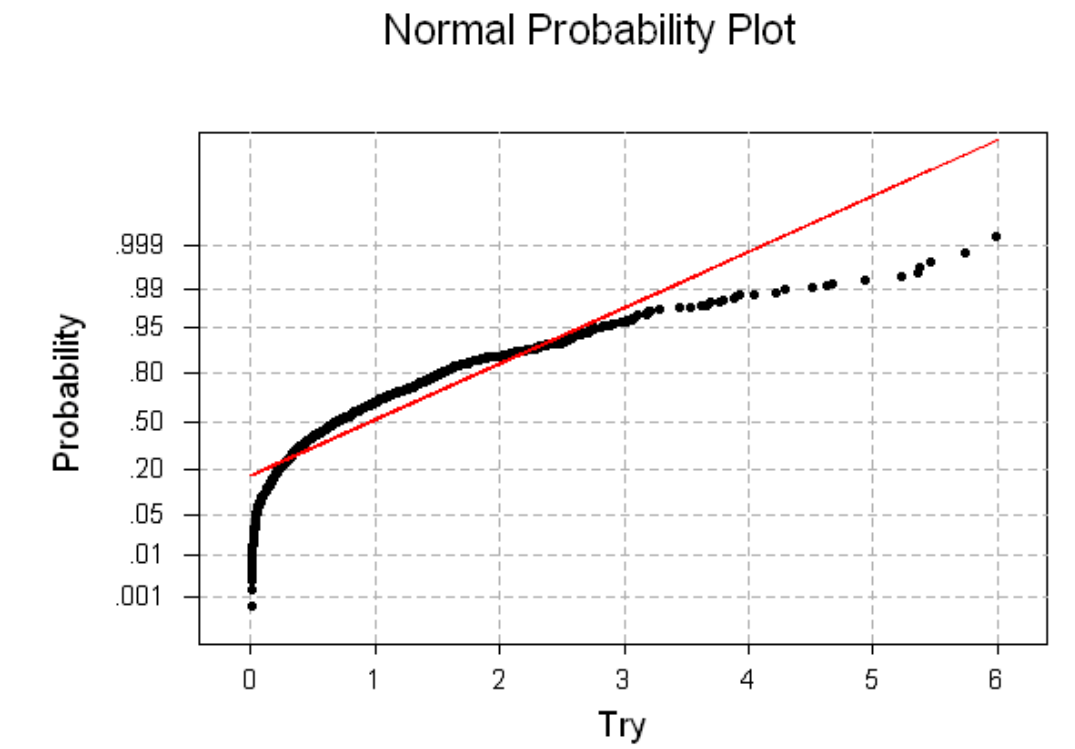


Comment: It is obvious that Sample was drawn from a normal population.

Now as an exercise we tested the normality assumption of the sample in Try and we concluded that;

Comment: Since P-Value = < 0.05 the sample Try was NOT drawn from a normal population.

Remember: in the home work the only method you can use to test the normality assumption is the Character NPP to be printed on the printer.


Normal Probability Plot
Probability
.999
.99
.95
.80
.50
.20
.05
.01
.001
0
1
2
3
4
5
6
Try
Average: 0.963901
StDev: 0.930082
N: 1000
Anderson-Darling Normality Test
A-Squared: 41.770
P-Value: 0.000

LAB 8
VERIFICATION OF THE CENTRAL LIMIT THEOREM

The Mean & Standard deviation of $\overline{X}$

Suppose a random sample of size n is to be drawn from a population with mean μ and standard deviation σ. Then $\mu_{\overline{X}}$ = μ, i.e., for each sample, the mean of the random variable $\overline{X}$ is equal to mean of the population. The standard deviation of $\overline{X}$ is

$$\sigma_{\overline{X}} = \begin{cases} \sqrt{\dfrac{N-n}{N-1}}\,\dfrac{\sigma}{\sqrt{n}} & \text{if sampling without replacement} \\ \dfrac{\sigma}{\sqrt{n}} & \text{if sampling with replacement} \end{cases}$$

where N is the population size. If the sample size (n) is small (n≤0.05N) relative to the population size then, $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ in both cases. In this lab we will deal, only, with infinite populations such as normal or binomial populations where the sampling as always assumed to be with replacement. So we will consider always that $\mu_{\overline{X}}$ = μ and $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$. It is important to notice that drawing with replacement is equivalent to drawing from an infinite population and vice versa.

Open the Minitab project *Lab8.mpj* from your floppy disk.

Example 1. Given the following sample: 1, 10, 100 and 1000 then describe this sample.
Session commands:

```
MTB > desc c1
```

Descriptive Statistics: C1

Variable	N	Mean	Median	TrMean	StDev	SE Mean
C1	4	278	55	278	484	242

Variable	Minimum	Maximum	Q1	Q3
C1	1	1000	3	775

Compare the theoretical result of $\sigma_{\bar{x}}$ with **SE Mean** in the last printout!

Central Limit Theorem (CLT)

Suppose a random sample of size n is to be taken from a population with mean μ and standard deviation σ. Then the random variable $\overline{X}$ has approximately the distribution of a normal variable with mean $\mu_{\overline{X}}=\mu$ and standard deviation $\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}$ as n increases without limit.

The standardized version of $\overline{X}$ is $Z=\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}$ which has approximately a standard normal distribution. Minitab can be employed to verify the CLT by drawing different large samples from some distribution, calculating their means and showing that the distribution of the population of these means are approximately normal. Practically speaking we mean by a large sample is that with a size ≥ 30.

Procedure of verification:

1. Generate 50 random samples each of size 5, 15, 30 and 40 from a Binomial distribution with n=10 and p=0.1.
2. Calculate the mean of each sample and store the means in a separate column named Means.
3. Test the normality assumption of the means column.

Note that we chose B(10, 0.1) because it is not Normal, it is a discrete distribution and it is much skewed to the left. Also, note that we can not place each sample in a separate column because we will face a hard difficulty in step 2 above if so, but instead we will place the samples in separate rows which would make it much easier to achieve step 2.

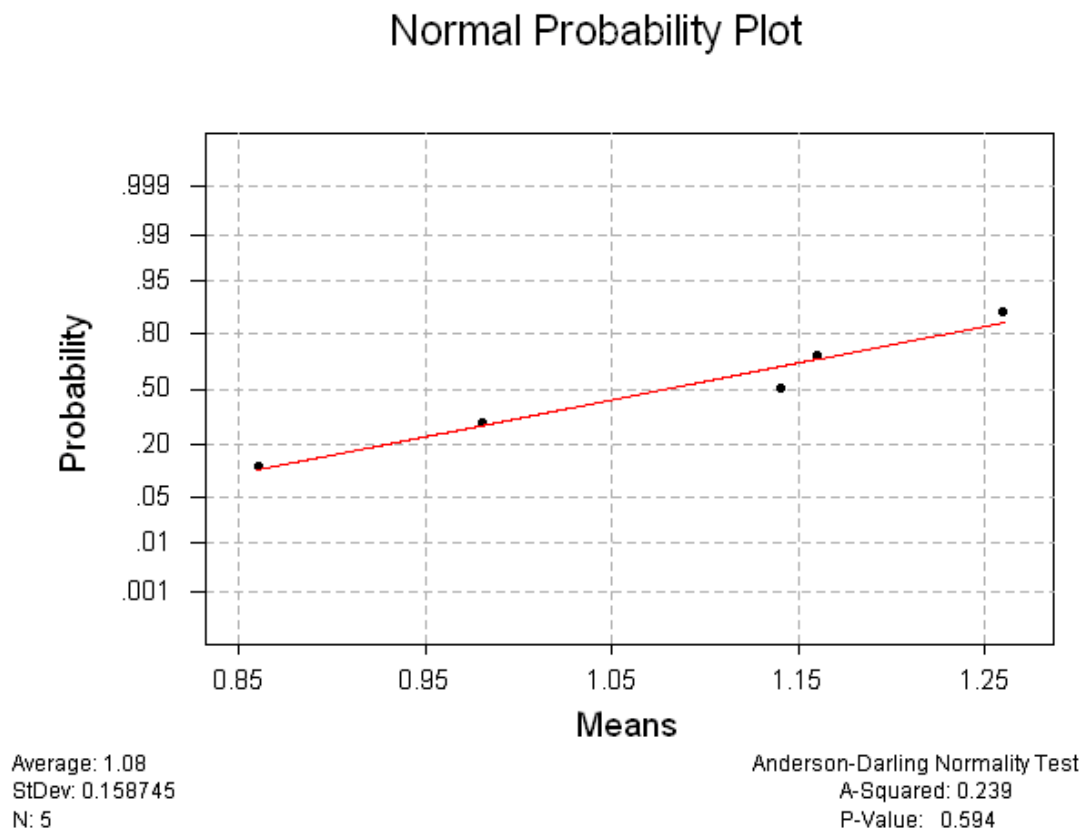
Example 2. The following example verifies the CLT for B(10,0.1) by generating 50 samples each of size n=5.

Session commands:

```
MTB > base 0
MTB > rand 5 c1-c50;
SUBC> bino 10 .1.
MTB > rmean c1-c50 c51
MTB > GPro.
* NOTE  * Professional Graphics are enabled.
         Standard Graphics are disabled.
         Use the GSTD command to enable
        Standard Graphics.
MTB > %NormPlot 'Means'.
```

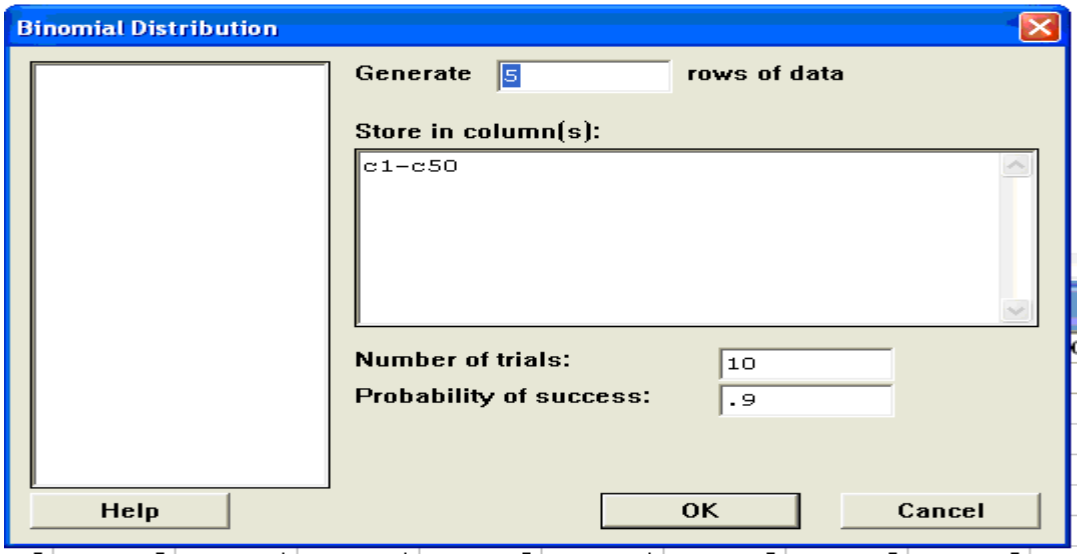
```
Executing from file:
D:\Minitab13\MACROS\NormPlot.MAC
Macro is running ... please wait
```

Normal Prob Plot: Means



Comment: it is obvious that we accept that the means population is Normally distributed.

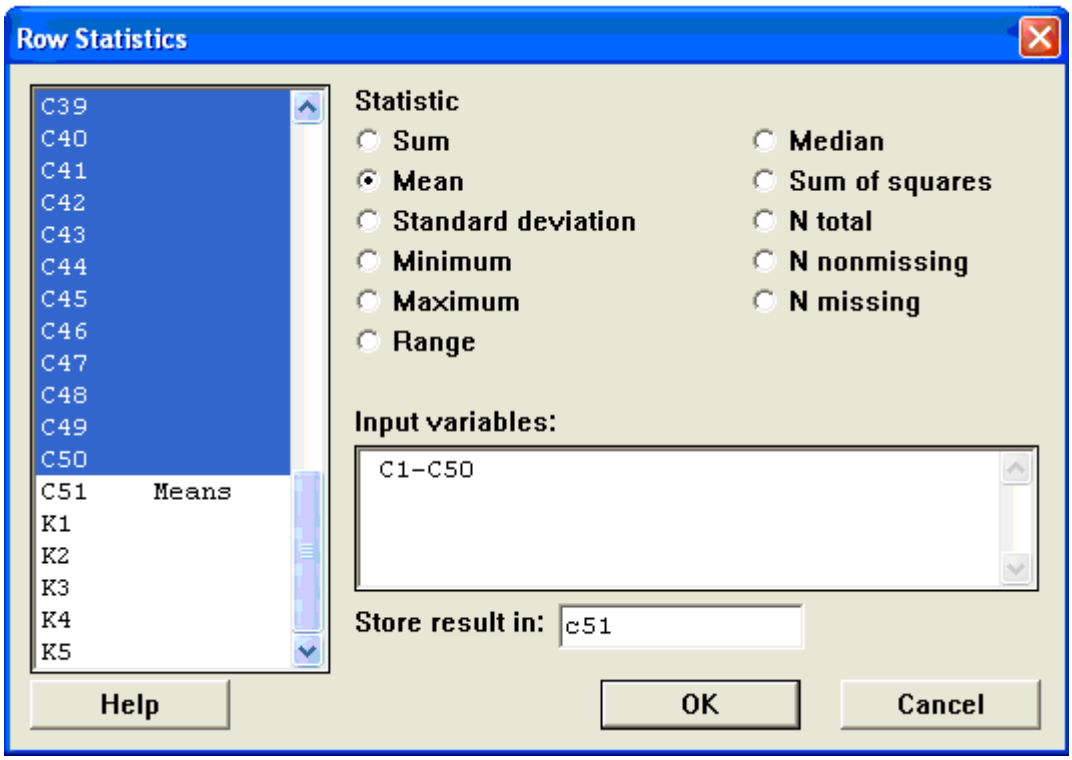
Menu commands: **Calc>Random Data>Binomial…**



Here is the commands for calculating the row mean.

Session commands: `MTB > RMean C1-C50 C51`

Menu commands: **Calc>Row Statistics…**



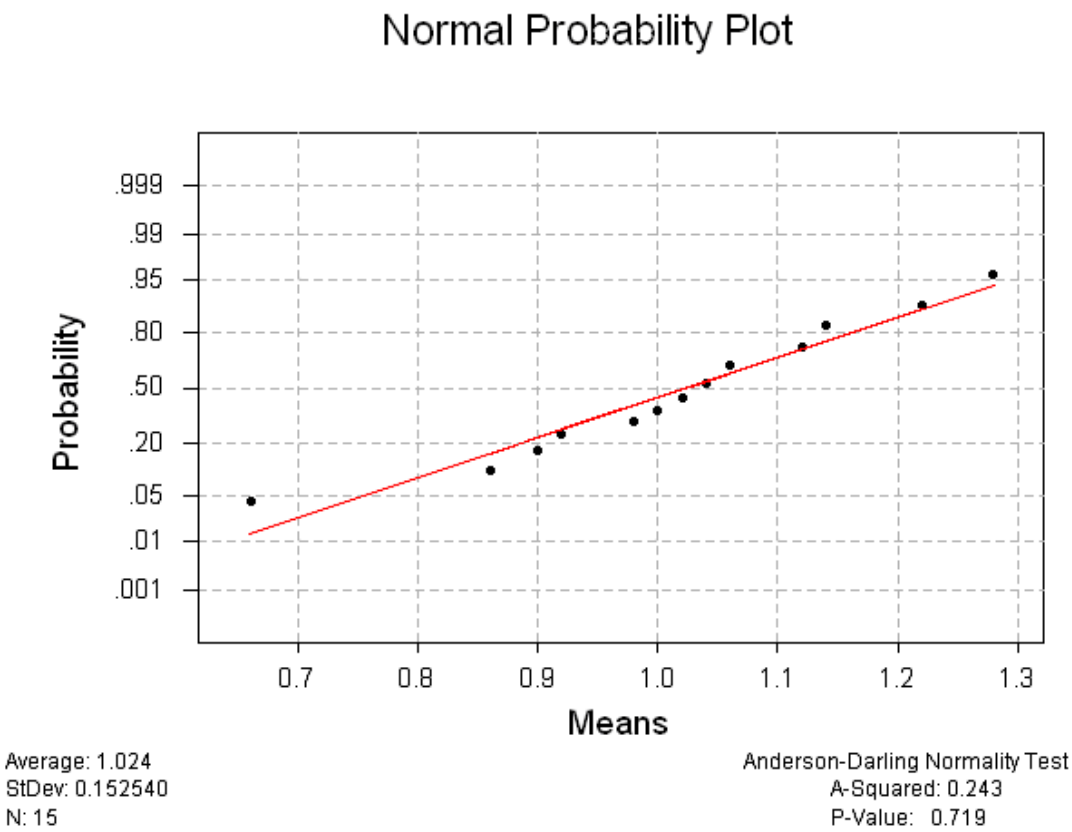
Now repeat Example.2 three times for n=15, 30 and 40. State your conclusion about the three additional results.

Example 3. The following example verifies the CLT for B(10,0.1) by generating 50 samples each of size n=15.

Session commands:

```
MTB > base 0
MTB > rand 15 c1-c50;
SUBC> bino 10 .1.
MTB > rmean c1-c50 c51
MTB > GPro.
* NOTE  * Professional Graphics are enabled.
          Standard Graphics are disabled.
          Use the GSTD command to enable
         Standard Graphics.
MTB > %NormPlot 'Means'.
Executing from file:
D:\Minitab13\MACROS\NormPlot.MAC
Macro is running ... please wait
```

Normal Prob Plot: Means

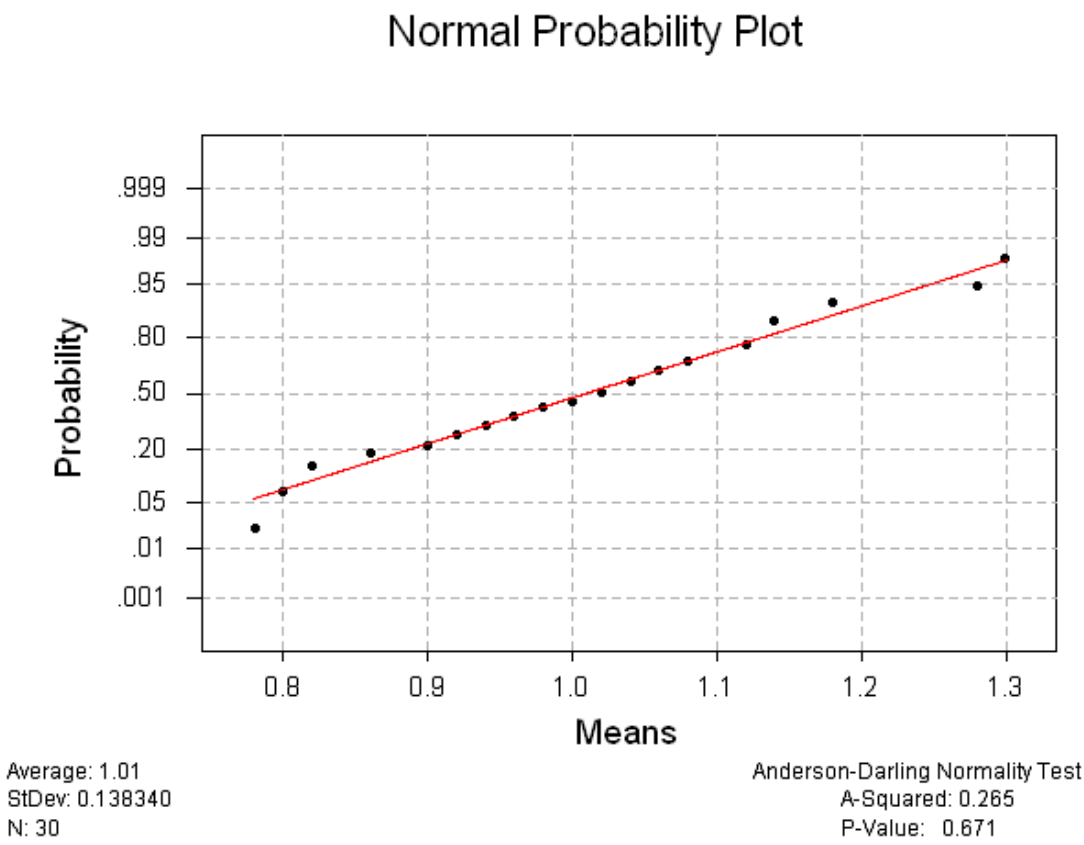


Example 4. The following example verifies the CLT for B(10,.1) by generating 50 samples each of size n=30.

Session commands:

```
MTB > base 0
MTB > rand 30 c1-c50;
SUBC> bino 10 .1.
MTB > rmean c1-c50 c51
MTB > GPro.
* NOTE  * Professional Graphics are enabled.
         Standard Graphics are disabled.
         Use the GSTD command to enable
        Standard Graphics.
MTB > %NormPlot 'Means'.
Executing from file:
D:\Minitab13\MACROS\NormPlot.MAC
Macro is running ... please wait
```

Normal Prob Plot: Means

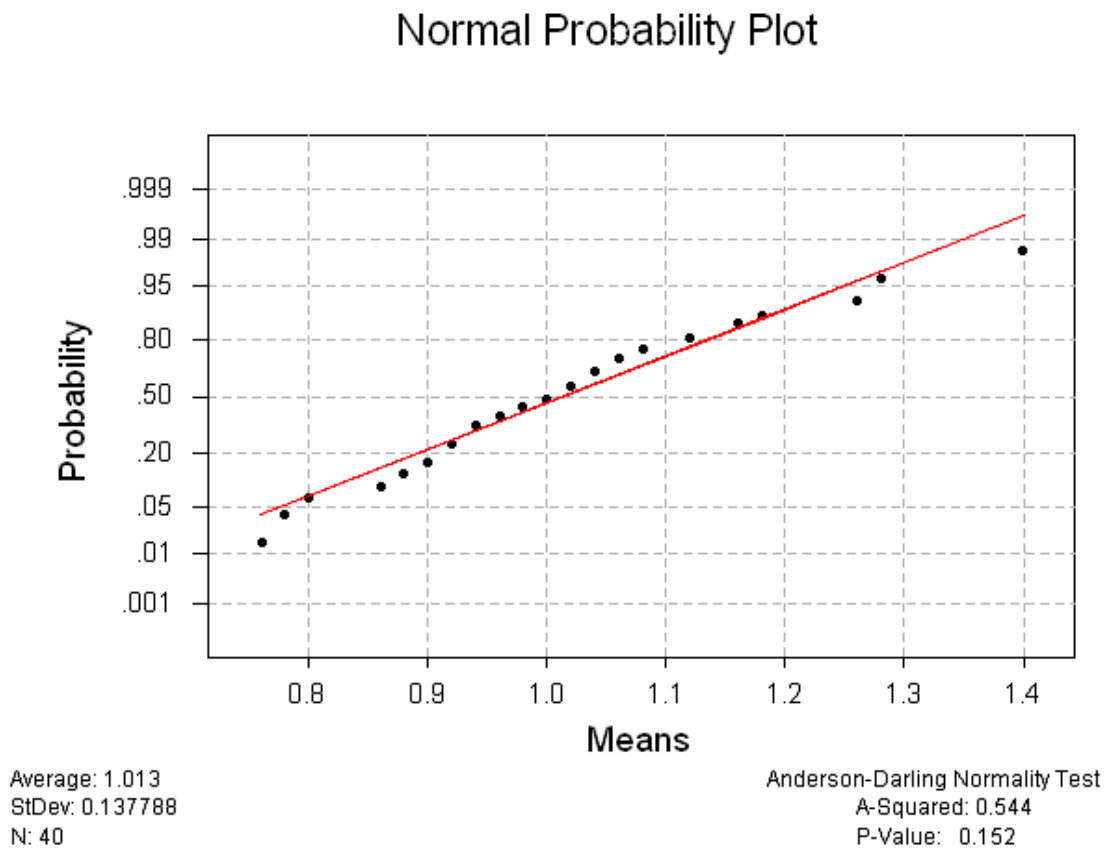


Example 5. The following example verifies the CLT for B(10,.1) by generating 50 samples each of size n=40.

Session commands:

```
MTB > base 0
MTB > rand 40 c1-c50;
SUBC> bino 10 .1.
MTB > rmean c1-c50 c51
MTB > GPro.
* NOTE  * Professional Graphics are enabled.
         Standard Graphics are disabled.
         Use the GSTD command to enable
        Standard Graphics.
MTB > %NormPlot 'Means'.
Executing from file:
D:\Minitab13\MACROS\NormPlot.MAC
Macro is running ... please wait
```

Normal Prob Plot: Means



General comment: From all of the above we saw that the result of the normality test was getting better as the number of the samples getting larger which verifies clearly the CLT.

UNIT IV
ESTIMATION AND TESTING

LAB 9
CONFIDENCE INTERVALS ABOUT THE MEAN

Definitions

Parameters: Are the characteristics of a random variable (or population).
Estimation: Is the approximating value of the population parameter using the sample elements.
Statistic: Is a function based on the sample elements used to estimate the population parameter.

Point Estimation

In most applications, the population under consideration undergoes the Normal Distribution. In addition, the most frequently used measure of central tendency is the population Mean μ. Subsequently, one is interested in estimating the mean of the normal population for small samples or estimating the mean of any population for large samples.

Theorem: Let $X_1, X_2, \ldots\ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2(i.e.N (μ, σ^2)), then $\overline{X}$ is an *unbiased*, that is $E(\overline{X}) = \mu$, estimator of μ.
So, $\overline{X}$ is called a *Point Estimate* of μ, since it gives a single numerical value.

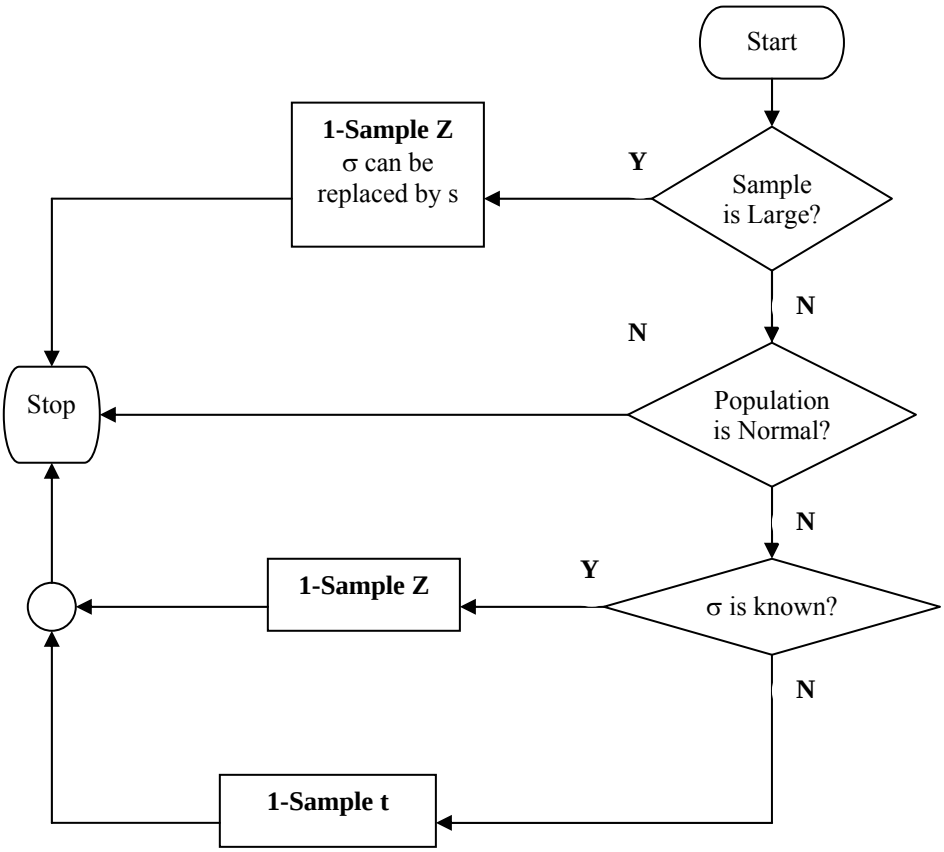
Interval Estimation (Confidence Intervals)

One can compute a point estimate for a population parameter once a random sample is drawn from the population. However, in many applications it is preferable to obtain an interval of values that has a high probability of containing μ. Such intervals are often called Confidence Intervals denoted by C.I.

In general, a C.I., of a confidence level $(1-\alpha)100\%$, about a population parameter μ with a point estimate $\hat{\theta}$, that has a variance $\sigma_{\hat{\mu}}$, and a confidence coefficient $z_{1-\alpha/2}$ is given by the form $\hat{\mu} \pm Z_{1-\alpha/2}\, \sigma_{\hat{\mu}}$.

Consequently, if $\hat{\mu} = \overline{X}$ then $\sigma_{\hat{\mu}} = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

Since there are four cases to be studied depending on the sample size, the distribution of the population and the status of σ, it is very useful to consider this flow chart as a brief figure for the procedure of finding the Confidence Interval;



1. Small sample C.I. for μ of a Normal population

Part I: σ^2 is known

Theorem: If $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, then a $(1-\alpha)100\%$ C.I. for μ is given by $\overline{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$.

A C.I. can be computed in Minitab using the command **OneZ**.

Open the Minitab project *Lab9.mpj* from your floppy disk.

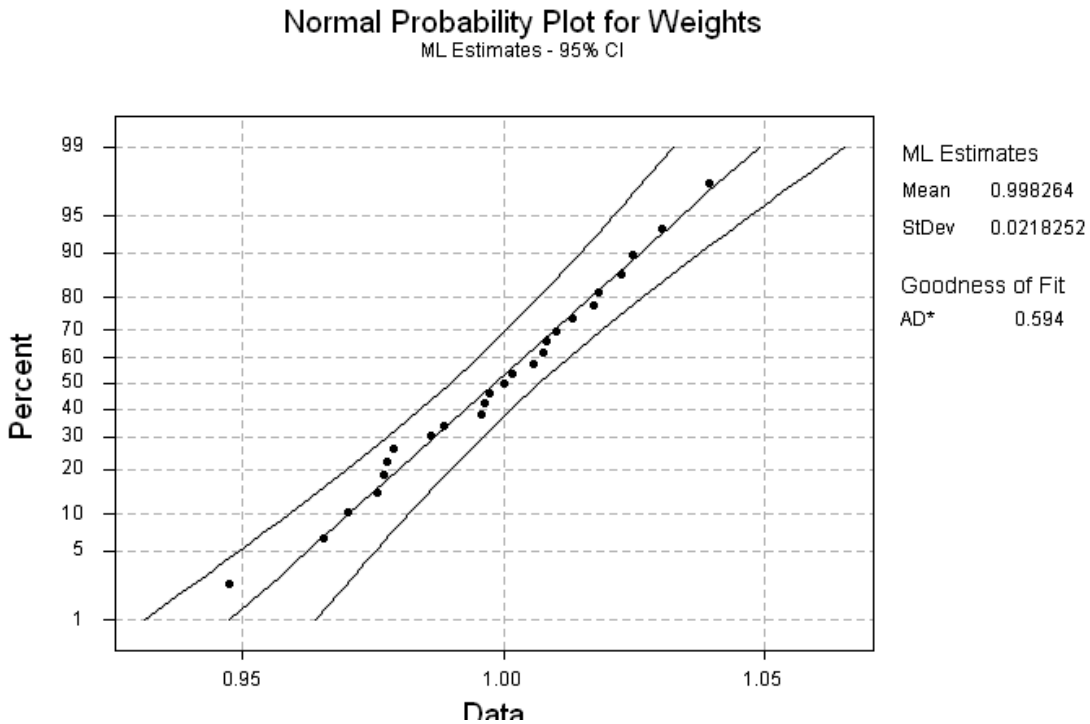
Example 1. Given the sample in C1, if the population is N(μ,0.0004), find a 95% C.I. for the mean weight.
Here, n=25 (small sample), 100(1-α) = 95 and σ is known so we must use 1-Sample Z to find the C.I.

Session commands:

```
MTB > let c2=nsco(c1)
MTB > %Qqplot 'Weights';
SUBC>   Normal;
SUBC>   MLE;
SUBC>   Allpts;
SUBC>   Table;
SUBC>   Conf 95;
SUBC>   Ci.
Executing from file:
D:\Minitab13\MACROS\Qqplot.MAC
```

It is **optional** in this case because it is mentioned in the problem that the population is **Normal**.
But, in general, it is **compulsory** in the case of **small** samples and **unknown** populations.

Prob Plot for Weights



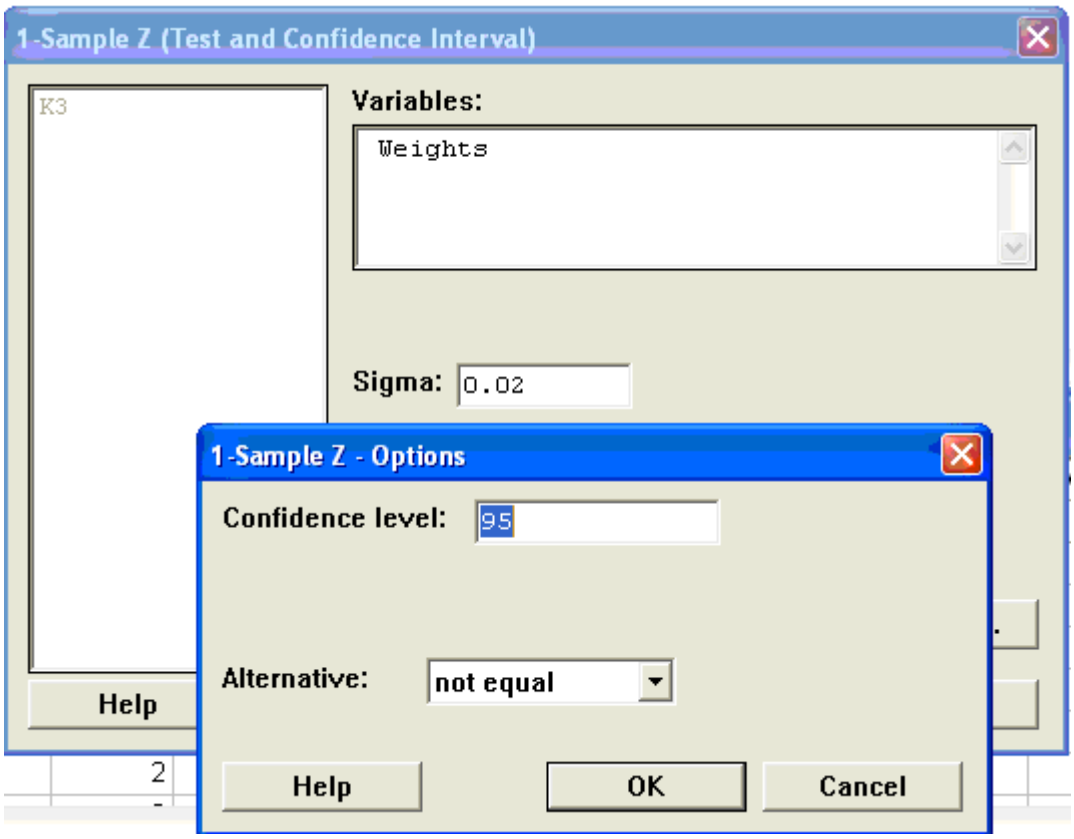
Session commands:

```
MTB > OneZ  'Weights';
SUBC>   Sigma 0.02;
SUBC>   Confidence 95.
```

One-Sample Z: Weights

```
The assumed sigma = 0.02

Variable          N      Mean     StDev   SE Mean          95.0% CI
Weights          25   0.99826   0.02228   0.00400  ( 0.99042, 1.00610)
```

Menu commands: **Stat> Basic Statistics > 1- Sample Z…**
Select : **Options…** and fill in **Confidence level:**



Part II: σ^2 is unknown

Theorem: If $X_1,X_2,\ldots,X_n$ is a random sample from $N(\mu,\sigma^2)$ and σ^2 is unknown, then a $(1-\alpha)100\%$ C.I. for μ is given by $\overline{X} \pm t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}$, Where ***s*** is the sample standard deviation.

A C.I. can be computed in Minitab, when σ^2 is unknown, using the command **OneT**.

Example 2. For the previous example, suppose that σ^2 were unknown. Find a 95% C.I. for the mean weight.
Here, n=25 (small sample), $100(1-\alpha) = 95$ and σ is unknown, s will be used instead, so we must use 1-Sample t.

Session commands:

```
MTB > OneT 'Weights'.
```

One-Sample T: Weights

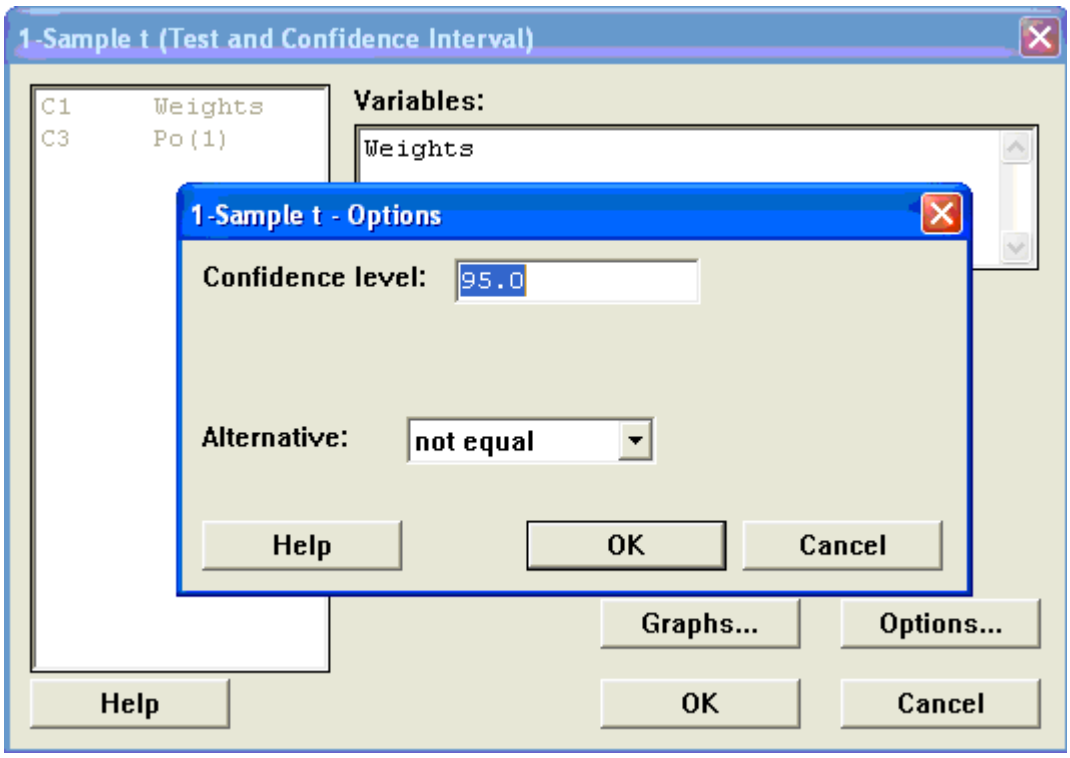
Variable	N	Mean	StDev	SE Mean	95.0% CI
Weights	25	0.99826	0.02228	0.00446	**(0.98907, 1.00746)**

Note: The TC.I. is slightly wider than the ZC.I. because:

1. The curve of the t-distribution is shorter but wider than that of the Normal distribution.
2. The population StDev. In TC.I. is unknown and substituted by s which is less accurate.

Menu commands: **Stat> Basic Statistics> 1-Sample t…**

Select: **Options…** and fill in **Confidence level:**



2. Large Sample C.I. for the mean of a general population

One can employ the CLT to get the following theorem, when the sample size ≥ 30.

Theorem: If $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2, regardless of the distribution of the population, then an approximate $(1-\alpha)100\%$ C.I. for μ is given by $\overline{X} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}}$. Where n is assumed to be large.

A C.I. can be computed in Minitab using the command **OneZ**.

Example 3. Compute a 90% C.I. for mean λ using the sample in C3.

Here, n=100 (large sample), $100(1-\alpha) = 90$ and σ is unknown so we must calculate s for the given sample before we find the C.I.

Session commands:

```
MTB > stan c3 k3
```

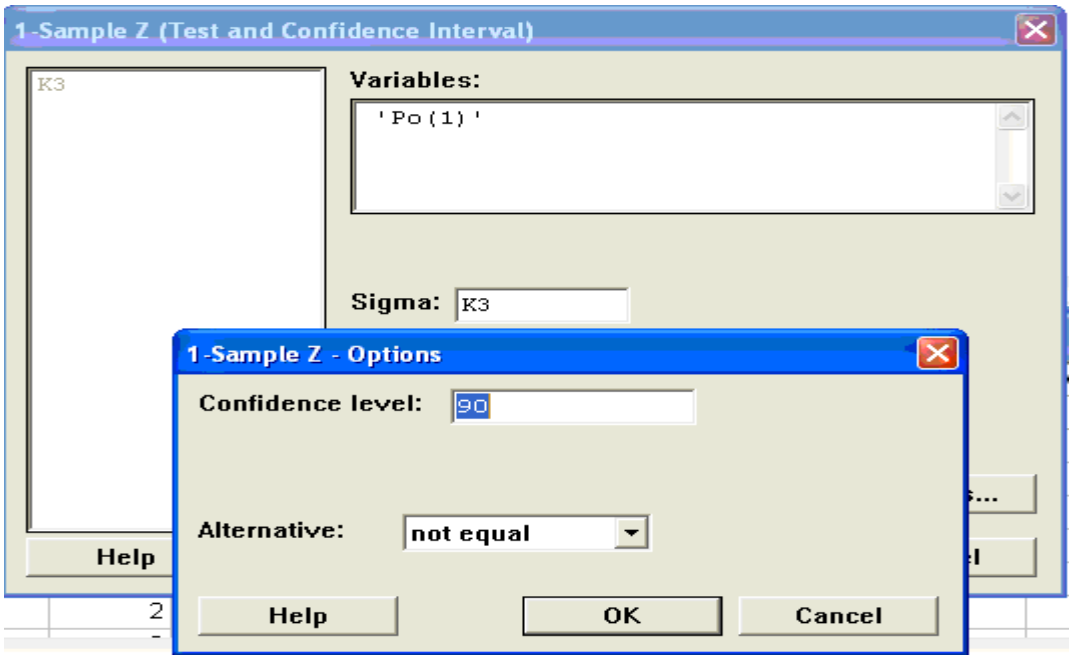
Standard Deviation of Po(1)

```
   Standard deviation of Po(1) = 0.95007
MTB > OneZ  'Po(1)';
SUBC>   Sigma K3;
SUBC>   Confidence 90.
```

One-Sample Z: Po(1)

```
The assumed sigma = 0.950066

Variable          N      Mean     StDev   SE Mean          90.0% CI
Po(1)           100    0.9200    0.9501    0.0950  (  0.7637,  1.0763)
```

Menu commands: **Stat> Basic Statistics > 1- Sample Z…**



Note: If the confidence level is NOT mentioned in the problem, it can be assumed to be 95%.

Note: If the confidence level is 95%, which is the default value in the dialogue box of testing, you may NOT write it in the session commands.

LAB 10

TESTING HYPOTHESES ABOUT THE MEAN

Definitions

Statistical Hypothesis: Is a statement about the value or values of the population parameter(s).

Null Hypothesis H_o: Is that hypothesis that generally declares that no change from the current, or the last, situation exists.

Alternative Hypothesis H_1: Is that hypothesis indicating that a conjecture or new idea of an investigator is true.

Test Statistic: Is a calculation from the sampled data, the value of which will be used to decide whether to accept or reject the null hypothesis, it is sometimes called the Test Ratio (TR) or the Test Function.

Critical Region: Is the set of values of the test statistic that lead to the rejection of H_0, it is also known as the rejection region.

Type I error: Is made if H_0 is rejected while it is actually true.

Significance Level (α): Is the probability that the test statistic will fall within the critical region when H_0 is true; that is, it is equal to the probability of type I error of the test.

P-value (P): Is the area of the region lies beyond the test function value.

1. Small Sample Hypothesis Testing About the Mean μ:

Two cases will be considered, the first when σ^2 is known and the other when σ^2 is unknown, under the assumption that the sample is drawn from a normal population.

Part I: σ^2 is known:

To test $H_0 : \mu = \mu_0$ vs.

$H_1 : \mu > \mu_0$ Reject H_0, at level α, if $Z_0 > z_{1-\alpha}$ (right-tailed test)

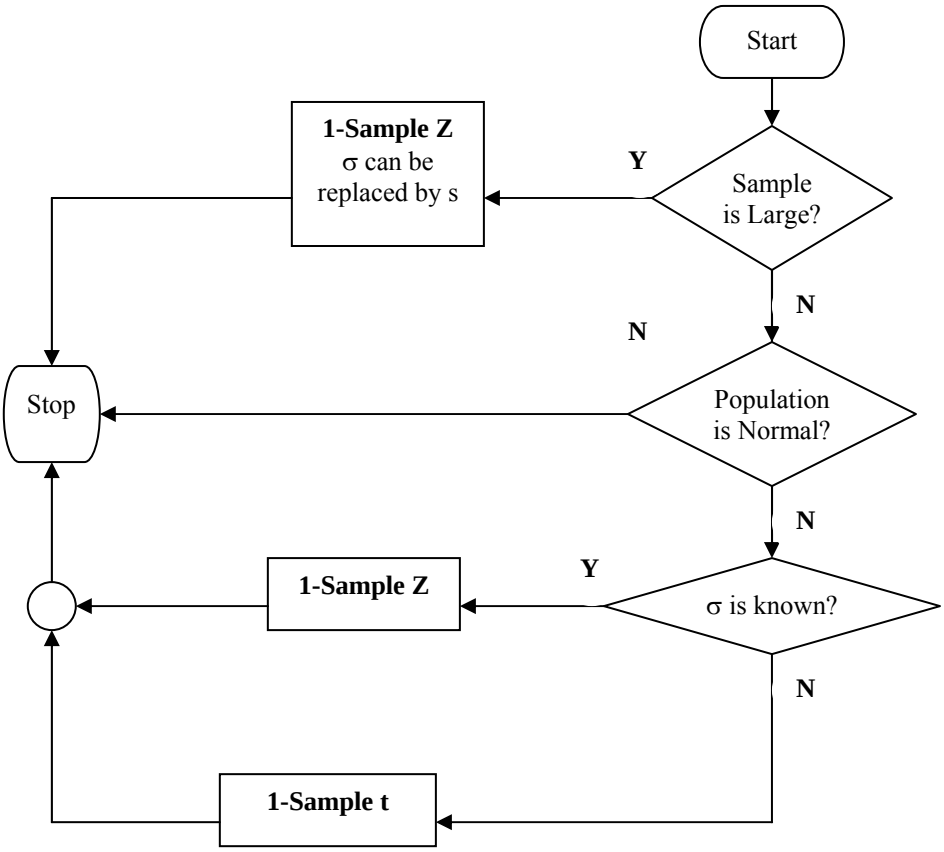
$H_1 : \mu < \mu_0$ Reject H_0, at level α, if $Z_0 < -z_{1-\alpha}$ (left-tailed test)

$H_1 : \mu \neq \mu_0$ Reject H_0, at level α, if $|Z_0| > z_{1-\alpha/2}$ (two-tailed test)

Where $Z_0 = \dfrac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$ is the value of the test function and $z_{1-\alpha}$ is the value of N(0,1) under which an area $1-\alpha$ lies.

A simpler approach of testing hypotheses is using the P-value, following the rule; if $P < \alpha$ then *Reject* H_0, otherwise *Do not Reject* H_0.

Again we refer to the flow chart given in the previous lab.



Example 1. For the sample of weights stored in C1, do the data provide sufficient evidence to conclude that the mean weights, differs from 1 lb.? At 5% sig. level and the population standard deviation is 0.02.

To answer this question is to test the following hypotheses H_0 : $\mu=1$ vs. H_1 : $\mu\neq1$ where $\alpha= 0.05$ and $\sigma = 0.02$, using Minitab;

Session command:

```
MTB > OneZ  'Weights';
SUBC>   Sigma 0.02;
SUBC>   Test 1;
SUBC>   Confidence 95.
```

One-Sample Z: Weights

```
Test of mu = 1 vs mu not = 1
The assumed sigma = 0.02

Variable          N      Mean     StDev   SE Mean
Weights          25   0.99826   0.02228   0.00400
```

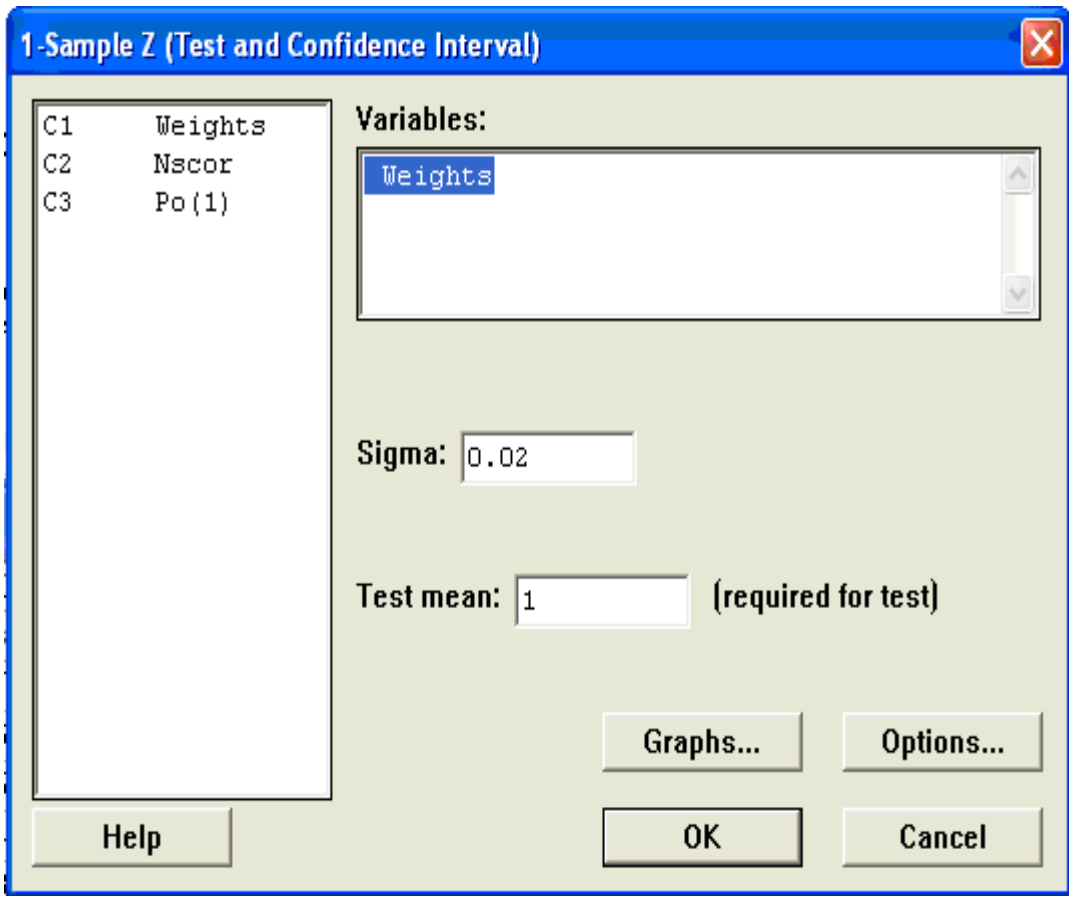
```
Variable              95.0% CI              Z      P
Weights       ( 0.99042, 1.00610)     -0.43  0.664
```

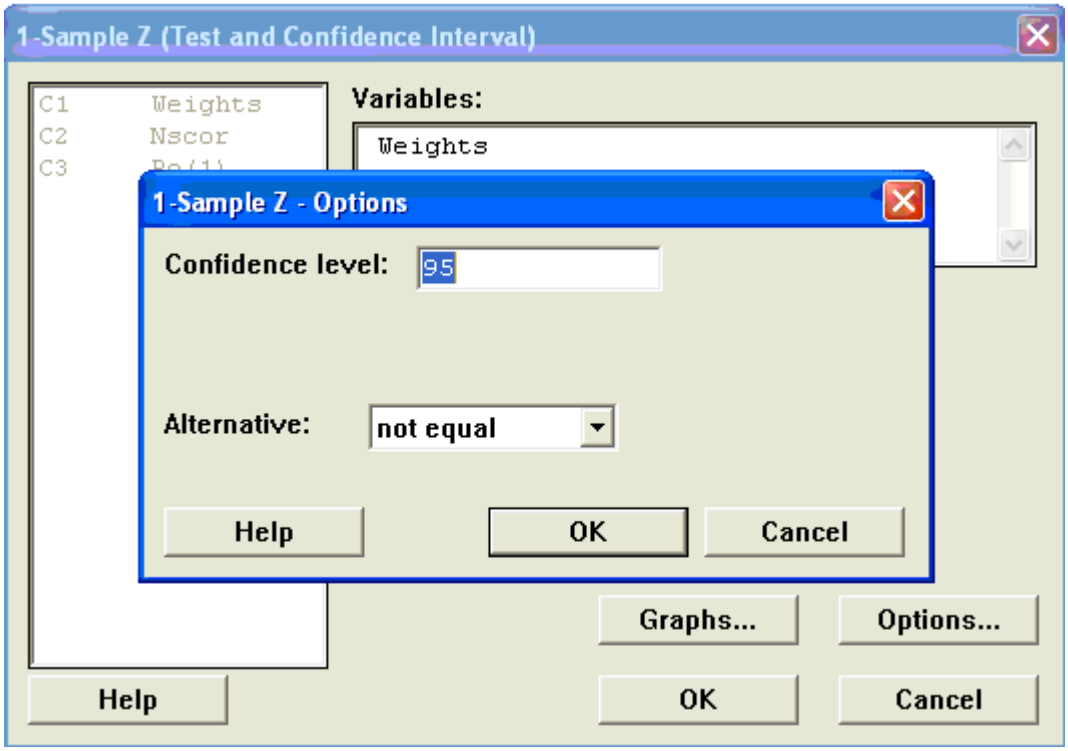
Note: To distinguish among the three cases of the alternative hypothesis, you may identify the following notation;

1. If H_1 has < sign write **alte -1** in SUBC> on the session.
2. If H_1 has ≠ sign write **alte 0** in SUBC> on the session.
3. If H_1 has > sign write **alte 1** in SUBC> on the session.

Since 0.66 > .05, our decision is NOT to Reject H_0. Consequently, the data provide sufficient evidence to conclude that *the mean weight, for all of the one-pound containers,* does not *differ from 1 lb. at 5% sig. level.*

Menu commands: **Stat> Basic Statistics> 1-Sample Z…**
Fill in: **Test mean:** and
Select **Options...** to fill in **Alternative:**





Note: If the significance level is NOT mentioned in the problem, it can be assumed to be 5%.

Note: If the significance level is 5%, which is the default value in the dialogue box of testing, you may NOT write it in the session commands.

Note: If the alternative hypothesis is Not equal, ≠, you may NOT write it in the session commands.

Example 2. For the weights in C1, test that the mean weights exceeds 0.99 lb. Given that the true, or the population, standard deviation is 0.02.

To answer this question is to test the following hypotheses H_0 : $\mu=0.99$ vs. H_1 : $\mu > 0.99$ where $\alpha= 0.05$, since it was not mentioned in the question, and $\sigma = 0.02$, using Minitab;

Session command:

```
MTB > OneZ  'Weights';
SUBC>   Sigma 0.02;
SUBC>   Test 0.99;
SUBC>   Confidence 95;
SUBC>   Alternative 1.
```

One-Sample Z: Weights

```
Test of mu = 0.99 vs mu > 0.99
The assumed sigma = 0.02
```

```
Variable          N      Mean     StDev   SE Mean
Weights          25   0.99826   0.02228   0.00400

Variable      95.0% Lower Bound         Z       P
Weights                 0.99168      2.07   0.019
```

Since 0.02 < .05, our decision is to Reject H_0. Consequently, the data provide sufficient evidence to conclude that *the mean weights, for all of the one-pound containers, exceeds 0.99 lb.*

Menu commands: **Stat> Basic Statistics> 1-Sample Z…**
Fill in: **Test mean:** and
Select **Options...** to fill in **Alternative:**

Part II : σ^2 is unknown:

To test $H_0 : \mu = \mu_0$ vs.
$H_1 : \mu > \mu_0$ Reject H_0, at level α, if $T_0 > t_{1-\alpha,n-1}$ (right-tailed test)
$H_1 : \mu < \mu_0$ Reject H_0, at level α, if $T_0 < -t_{1-\alpha,n-1}$ (left-tailed test)
$H_1 : \mu \neq \mu_0$ Reject H_0, at level α, if $|T_0| > t_{1-\alpha/2,n-1}$ (two-tailed test)

Where $T_0 = \dfrac{\bar{X}-\mu_0}{s/\sqrt{n}}$ is the value of TR and $t_{1-\alpha,n-1}$ is the value of t-distribution, with n-1 degrees of freedom, under which an area 1-α lies.

If σ is assumed to be unknown we use the t-test. From the last printout note that s=0.02228. We will perform t-test for the previous hypotheses using Minitab.

Example 3. For the weights in C1, Test the claim that the mean weights, for all of the one-pound containers, differs from 1 lb.
Session command:

```
MTB > OneT  'Weights';
SUBC>   Test 1.
```

One-Sample T: Weights

```
Test of mu = 1 vs mu not = 1

Variable          N      Mean     StDev   SE Mean
Weights          25   0.99826   0.02228   0.00446

Variable             95.0% CI            T       P
Weights       ( 0.98907, 1.00746)    -0.39   0.700
```

Menu commands: **Stat>Basic Statistics>1-Sample t…**

2. Large Samples Hypothesis Testing about μ (n ≥ 30):

For the testing of hypothesis about a population mean μ, regardless of the population distribution, the Large Sample Z-test can be used to perform such a test.

To test $H_0 : \mu = \mu_0$ vs.

$H_1 : \mu > \mu_0$ Reject H_0, at level α, if $Z_0 > z_{1-\alpha}$ (right-tailed test)

$H_1 : \mu < \mu_0$ Reject H_0, at level α, if $Z_0 < -z_{1-\alpha}$ (left-tailed test)

$H_1 : \mu \neq \mu_0$ Reject H_0, at level α, if $|Z_0| > z_{1-\alpha/2}$ (left-tailed test)

Where $Z_0 = \dfrac{\overline{X} - \mu_0}{\sigma(s)/\sqrt{n}}$ is the value of the Test Statistic and $z_{1-\alpha}$ is the value of N(0,1) under which an area 1-α lies. If σ is unknown it can be replaced by s.

Example 4. For the sample in C3, test the claim that the mean λ is less than 2 at 10% sig. level.

To answer this question is to test the following hypotheses $H_0 : \mu=2$ vs. $H_1 : \mu < 2$ where $\alpha= 0.10$, and σ is unknown, so we have to calculate the sample StDev before testing;

Session commands:

```
MTB > stan c3 k3
```

Standard Deviation of Po(1)

```
   Standard deviation of Po(1) = 0.95007
MTB > OneZ  'Po(1)';
SUBC>   Sigma K3;
SUBC>   Test 2;
SUBC>   Confidence 90;
SUBC>   Alternative -1.
```

One-Sample Z: Po(1)

```
Test of mu = 2 vs mu < 2
The assumed sigma = 0.950066

Variable          N      Mean     StDev   SE Mean
Po(1)           100    0.9200    0.9501    0.0950

Variable      90.0% Upper Bound         Z      P
Po(1)                    1.0418   -11.37  0.000
```

Menu commands: **Stat> Basic Statistics> 1-Sample Z…**
Fill in: **Test mean:** and
Select **Options...** to fill in **Alternative:**