उत्तर प्रदेश राजर्षि टण्डन मुक्त
विश्वविद्यालय, इलाहाबाद

PGSTAT- 21

/

MASTAT-21

Statistical Software

Printed Chandrakala Universal Pvt Ltd, 42/7 J.L.N. Road Allahabad

Data Analysis Using Microsoft Excel

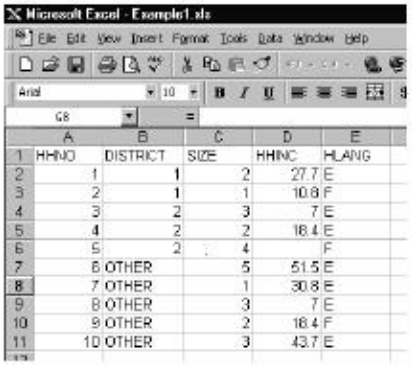
Introduction

The program *SampleCalc* (short for Sample Calculator) calculates unbiased point estimates and approximate confidence intervals of the principal population characteristics for each specified variable and each category of specified attributes. *SampleCalc* is a Microsoft Excel Add-In developed by Peter Tryfos, Professor of Management Science, Faculty of Administrative Studies, at York University, Ontario, Canada. It is assumed that the observations are entered in an Excel worksheet and arranged in the form of a table, the columns of which correspond to the variables and attributes of interest and the rows to the sampled elements. The editing of the data, and the coding of the variables, attributes, and missing values, should take place before *SampleCalc* is used. Before using *SampleCalc*, be aware that the following additional information is required by each sampling method. Make sure that this information is entered on the worksheet containing the observations or is otherwise available before *SampleCalc* is called.

1. Simple Number of elements in the population
   Confidence level
2. Stratified Labels identifying the groups (strata)
   Number of elements in each group (stratum) in the population
   Confidence level
3. Two-Stage Labels identifying the selected groups (strata)
   Number of elements in each selected group (stratum) in the population
   Number of elements in the population
   Number of groups (strata) in the population
   Number of selected groups (strata)
   Confidence level
4. Cluster Labels identifying the selected groups (strata)
   Number of elements in the population
   Number of groups (strata) in the population
   Number of selected groups (strata)
   Confidence level

It should be noted that the confidence intervals generally require large samples; however, *SampleCalc* does not check that this requirement is satisfied. Also, the number of categories of an attribute cannot exceed 50.

The application of SampleCalc will be illustrated with the help of a simple example. It will be assumed that a random sample without replacement of 10 households was selected from the population of households in a city. The observations are entered in an Excel spreadsheet in the manner shown below:



The labels of the variables and attributes are entered in the first row. HHNO stands for household number, DISTRICT for the district identification (1, 2, or OTHER), SIZE for the number of persons in the household, HHINC for household income, and HLANG for language spoken at home (E: English, F: French). There is a missing value (an empty cell) in cell D6.

Common First Steps

Enter the sample observations (data) in the form of a compact table in an Excel worksheet. The rows must correspond to the sampled elements and the columns to the variables and attributes of the study. It is strongly recommended that the first row of this table contain the labels of the variables and attributes. A missing value should be represented either by a blank cell or one containing a period.

1. **Click on any one cell of this compact table.**
2. **In the Tools menu, click on *SampleCalc*. A dialog box entitled Data and Labels will appear.**

- **If the displayed data range and labels are correct, click the OK button and go to Step 3.**
- **If the displayed data range or the labels are not correct, or if the data do not have the recommended format, click the Change button. In the two ensuing dialog boxes, select the proper ranges for the data and the labels. The Data and Labels dialog box will reappear. Click the OK button and go to Step 3.**

3. **The Method dialog box now appears. Select the method by which the sample was selected and click either the OK button to proceed, or the Cancel button to abort *SampleCalc*.**

The subsequent conversation with *SampleCalc* depends on the sampling method selected. Refer to the section corresponding to the method by which the sample was selected.

Method: Simple Random Sampling

To illustrate the application of SampleCalc, it will be assumed that the observations were selected by drawing a simple random sample without replacement of 10 households from among the 2400 households in the city. ***After selecting Simple in the Method box, the Simple random sampling dialog box appears.***

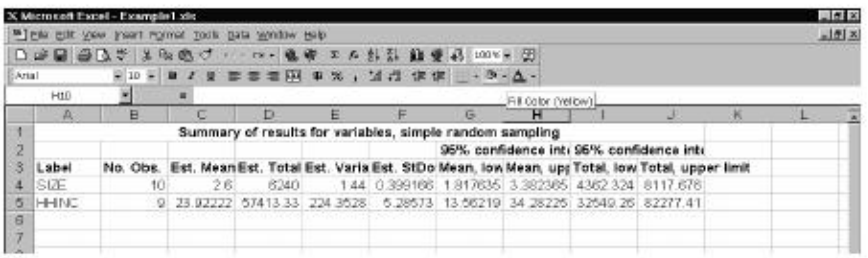
1 In the dialog box entitled **Simple random sampling,** enter the number of elements in the population and select the confidence level. Then, click either the OK button to proceed, or the Cancel button to abort *SampleCalc*.

2 The program will display a message for your information to the effect that it is creating the worksheet VRESULTS; acknowledge by clicking OK. A similar second message to the effect that the program is creating the worksheet ARESULTS should also be acknowledged by clicking OK. (VRESULTS will contain the results of calculations concerning the selected variables, while ARESULTS will contain the results regarding the selected attributes.)

3 The dialog box entitled **Data analysis** will now appear. To analyze one or more variables, select the first option. To analyze one or more attributes, select the second option. To stop, select the third option. Click OK.

If you chose to analyze one or more variables, a dialog box entitled **Selection of variables** will appear. Select the variables to be analyzed by clicking on their labels. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed

4 If you clicked OK, you will find that the VRESULTS worksheet now contains for each selected variable the number of observations with non-missing values, the estimates of the mean and total of the variable, the estimated variance of the variable, the estimated standard deviation of the estimate of the population mean, and the lower and upper limits of the specified confidence intervals for the mean and total of the variable.

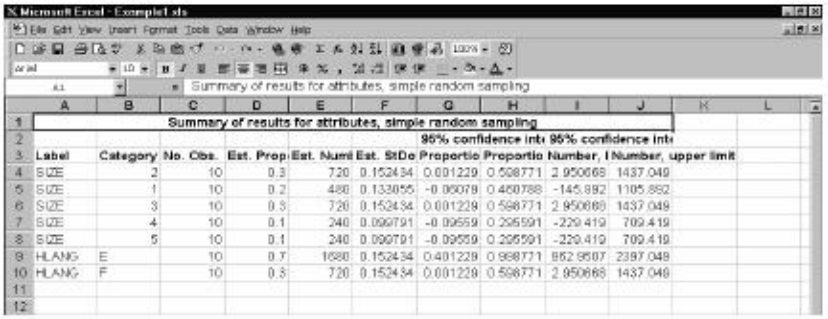


Summary of results for variables, simple random sampling

Label	No. Obs.	Est. Mean	Est. Total	Est. Varia	Est. StDe	95% confidence inte Mean, low	95% confidence inte Mean, upp	Total, low	Total, upper limit
SIZE	10	2.6	6240	1.44	0.399166	1.817635	3.382365	4362.324	8117.676
HHINC	9	23.92222	57413.33	224.3528	5.28573	13.56219	34.28225	32549.26	82277.41

If you chose to analyze one or more attributes, a dialog box entitled **Selection of Attributes** will appear. Select the attributes to be analyzed by clicking on their labels in the first list box. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed

5 If you clicked OK, you will find that the ARESULTS worksheet now contains for each category of each selected attribute the number of observations with non-missing values, the estimates of the proportion and number in the category, the estimated standard deviation of the estimate of the population proportion, and the lower and upper limits of the specified confidence intervals for the proportion and number in the category.



Summary of results for attributes, simple random sampling

	A	B	C	D	E	F	G	H	I	J	K	L
1	Summary of results for attributes, simple random sampling											
2							95% confidence inte		95% confidence inte			
3	Label	Category	No. Obs.	Est. Prop	Est. Numl	Est. StDe	Proportio	Proportio	Number, l	Number, upper limit		
4	SIZE	2	10	0.3	720	0.152434	0.001229	0.598771	2.950868	1437.049		
5	SIZE	1	10	0.2	480	0.133855	-0.06078	0.460788	-145.882	1105.882		
6	SIZE	3	10	0.3	720	0.152434	0.001229	0.598771	2.950868	1437.049		
7	SIZE	4	10	0.1	240	0.099791	-0.09559	0.295591	-229.419	709.419		
8	SIZE	5	10	0.1	240	0.099791	-0.09559	0.295591	-229.419	709.419		
9	HLANG	E	10	0.7	1680	0.152434	0.401229	0.998771	862.9507	2397.049		
10	HLANG	F	10	0.3	720	0.152434	0.001229	0.598771	2.950868	1437.049		
11												
12												

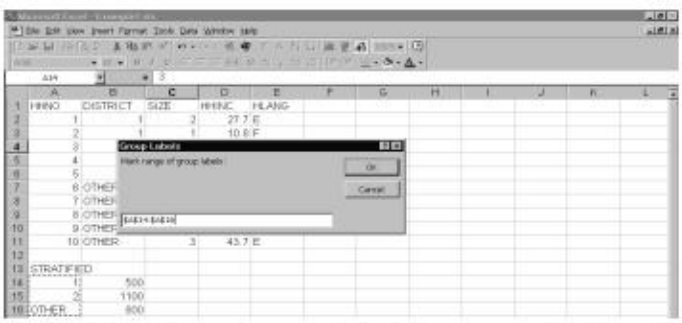
6 If you chose to Stop, a message will appear reminding you to save the calculations before leaving Excel. Click OK to leave *SampleCalc* and return to Excel. You may want to adjust the column widths of VRESULTS and ARESULTS in order to see the results more clearly. (If you wish to delete the worksheets VRESULTS and ARESULTS, right-click on the tab of the worksheet and select Delete from the pop-up menu.)

Method: Stratified Random Sampling

To illustrate the application of SampleCalc, it will be assumed that the city households are grouped into three districts and that the observations were selected by drawing a simple random sample without replacement of 2 households from among the 500 households in District 1, one of 3 households from among the 1100 households in District 2, and one of 5 households from among the 800 households in the district labeled Other. The labels and the number of households in each district are entered in the same worksheet as the sample observations.

After selecting Stratified in the Method box, the Group labels dialog box appears.

1 In the dialog box entitled Group labels, select the range containing the labels of the groups (strata). Click OK to proceed, or Cancel to abort the program.

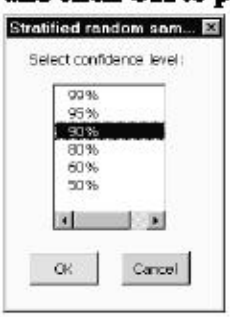


2 In the dialog box entitled Number of Pop. Elements, select the range containing the number of elements in each group (stratum) in the population. Click OK to proceed, or Cancel to abort the program.

3 A dialog box entitled Group Identifier will now appear. Select the label of the column in the table of data that identifies the group (stratum) to which each sampled element belongs. Click OK to proceed, or Cancel to abort.

4 A dialog box entitled First check of population and sample elements will now appear. It shows the program's understanding of the group (stratum) labels, and of the corresponding number of elements in the population and sample. If you observe an anomaly, click Cancel to abort SampleCalc, check the data, and begin anew. If the program's understanding appears correct, click OK to proceed.

5 In the dialog box entitled Stratified random sampling, select the desired confidence level, and click OK to proceed or Cancel to abort.

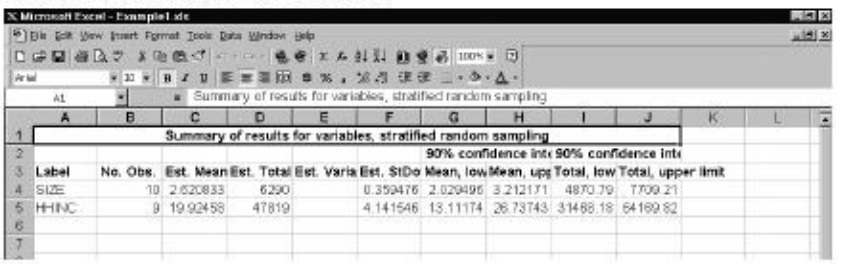


6 information to the effect that it is creating the worksheet VRESULTS; acknowledge by clicking OK. A similar second message to the effect that the program is creating the worksheet ARESULTS should also be acknowledged by clicking OK. (VRESULTS will contain the results of calculations concerning the selected variables, while ARESULTS will contain the results regarding the selected attributes.)

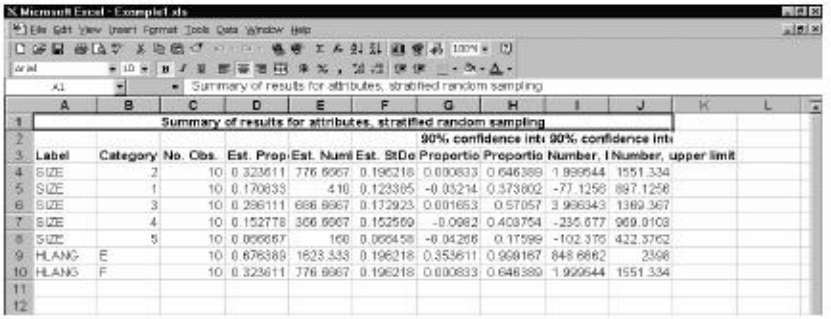
7 The dialog box entitled Data analysis will now appear. To analyze one or more variables, select the first option. To analyze one or more attributes, select the second option. To stop, select the third option. Click OK.

8 If you chose to analyze one or more variables, a dialog box entitled Selection of variables will appear. Select the variables to be analyzed by clicking on their labels. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the Data analysis dialog box.

9 If you clicked OK, you will find that the VRESULTS worksheet now contains for each selected variable the number of observations with non-missing values, the estimates of the mean and total of the variable, the estimated standard deviation of the estimate of the population mean, and the lower and upper limits of the specified confidence intervals for the mean and total of the variable.



10 If you chose to analyze one or more attributes, a dialog box entitled Selection of attributes will appear. Select the attributes to be analyzed by clicking on their labels in the first list box. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the Data analysis dialog box. If you clicked OK, you will find that the ARESULTS worksheet now contains for each category of each selected attribute the number of observations with non-missing values, the estimates of the proportion and number in the category, the estimated standard deviation of the estimate of the population proportion, and the lower and upper limits of the specified confidence intervals for the proportion and number in the category.



Summary of results for attributes, stratified random sampling

	A	B	C	D	E	F	G	H	I	J	K	L
1	Summary of results for attributes, stratified random sampling											
2							90% confidence inte		90% confidence inte			
3	Label	Category	No. Obs.	Est. Prop	Est. Numl	Est. StDe	Proportio	Proportio	Number, l	Number, upper limit		
4	SIZE	2	10	0.323611	776.6667	0.196218	0.000833	0.646389	1.999544	1551.334		
5	SIZE	1	10	0.170833	410	0.123385	-0.03214	0.373802	-77.1256	897.1256		
6	SIZE	3	10	0.296111	666.6667	0.172923	0.001653	0.57057	3.996343	1329.367		
7	SIZE	4	10	0.152778	366.6667	0.152569	-0.0982	0.403754	-235.677	969.0103		
8	SIZE	5	10	0.096667	160	0.066458	-0.04266	0.17599	-102.375	422.3752		
9	HLANG	E	10	0.676389	1623.333	0.196218	0.353611	0.998167	848.6662	2398		
10	HLANG	F	10	0.323611	776.6667	0.196218	0.000833	0.646389	1.999544	1551.334		
11												
12												

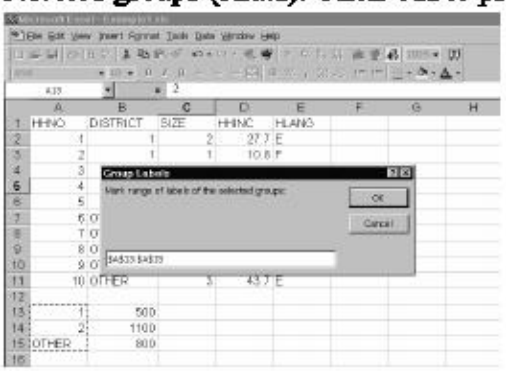
11 If you chose to Stop, a message will appear reminding you to save the calculations before leaving Excel. Click OK to return to Excel. You may want to adjust the column width of VRESULTS and ARESULTS in order to see the results more clearly

Method: Two-Stage Random Sampling

To illustrate the application of SampleCalc, it will be assumed that the 5000 city households are grouped into six districts and that the observations were selected in two stages. In the first, three of the six districts were selected at random and without replacement; these were the districts 1, 2, and other. In the second stage, a simple random sample without replacement of 2 households was drawn from among the 500 households in District 1, one of 3 households from among the 1100 households in District 2, and one of 5 households from among the 800 households in the district labeled Other. The labels and the number of households in each district are entered in the same worksheet as the sample observations.

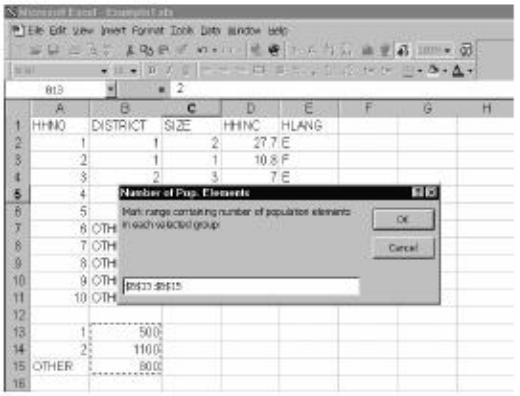
After selecting Two-stage in the **Method** box, the **Group labels** dialog box appears.

1 In the dialog box entitled Group labels, select the range containing the labels of the selected groups (strata). Click OK to proceed, or Cancel to abort the program.

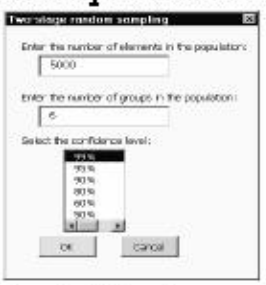


	A	B	C	D	E	F	G	H
1	HHNO	DISTRICT	SIZE		HHINC	HLANG		
2	1		1	2	27.7	E		
3	2		1	1	10.8	F		
4	3							
5	4							
6	5							
7	6	0						
8	7	0						
9	8	0						
10	9	0						
11	10	OTHER		3	43.7	E		
12								
13	1	500						
14	2	1100						
15	OTHER	800						
16								

2 In the dialog box entitled Number of Pop. Elements, select the range containing the number of elements in each selected group (stratum) in the population. Click OK to proceed, or Cancel to abort the program.

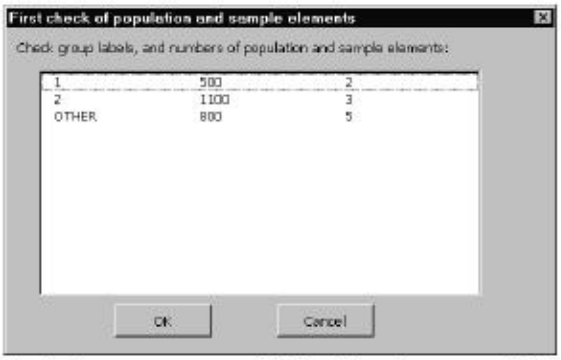


3 In the dialog box entitled Two-stage random sampling, enter the number of elements and the number of groups (strata) in the population, and select the desired confidence level. Click OK to proceed or Cancel to abort.



4 A dialog box entitled Group Identifier will now appear. Select the label of the one column in the table of data that identifies the group (stratum) to which each sampled element belongs. These labels should be consistent with those in Step 5; otherwise, an error message will eventually appear. Click OK to proceed, or Cancel to abort.

5 A dialog box entitled First check of population and sample elements will now appear. It shows the program's understanding of the labels of the selected groups (strata), and of the corresponding number of elements in the population and sample. If you observe an anomaly, click Cancel to abort SampleCalc, check the data, and begin anew. If the program's understanding appears correct, click OK to proceed.

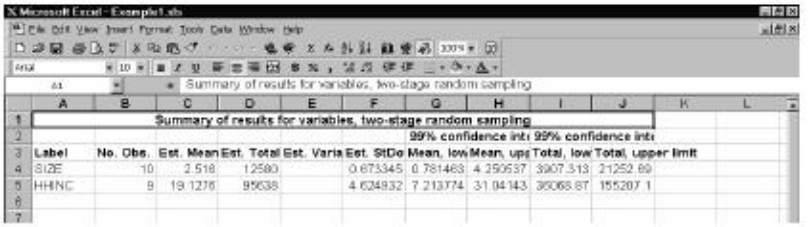


6 The program will display for your information a message to the effect that it is creating the worksheet VRESULTS; acknowledge by clicking OK. A similar second message to the effect that the program is creating the worksheet ARESULTS should also be acknowledged by clicking OK.

7 The dialog box entitled Data analysis will now appear. To analyze one or more variables, select the first option. To analyze one or more attributes, select the second option. To stop, select the third option. You can change the selection at any time before clicking OK. Click OK to confirm your selection and proceed.

8 If you chose to analyze one or more variables, a dialog box entitled Selection of variables will appear. Select the variables to be analyzed by clicking on their labels. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the Data analysis dialog box.

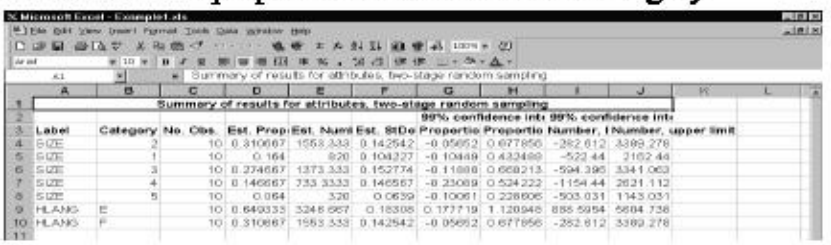
9 If you clicked OK, you will find that the VRESULTS worksheet now contains for each selected variable the number of observations with non-missing values, the estimates of the mean and total of the variable, the estimated standard deviation of the estimate of the population mean, and the lower and upper limits of the specified confidence intervals for the mean and total of the variable.



Summary of results for variables, two-stage random sampling

Label	No. Obs.	Est. Mean	Est. Total	Est. Varia	Est. StDe	99% confidence inte Mean, low	Mean, upp	99% confidence inte Total, low	Total, upper limit
SIZE	10	2.518	12580		0.673345	0.781463	4.250537	3907.313	21252.69
HHINC	9	19.1276	95638		4.624932	7.213774	31.04143	36068.87	155207.1

10 If you chose to analyze one or more attributes, a dialog box entitled Selection of attributes will appear. Select the attributes to be analyzed by clicking on their labels in the first list box. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the Data analysis dialog box. If you clicked OK, you will find that the ARESULTS worksheet now contains for each category of each selected attribute the number of observations with non-missing values, the estimates of the proportion and number in the category, the estimated standard deviation of the estimate of the population proportion, and the lower and upper limits of the specified confidence intervals for the proportion and number in the category.



Summary of results for attributes, two-stage random sampling

Label	Category	No. Obs.	Est. Prop	Est. Numb	Est. StDe	99% confidence inte Proportio	Proportio	99% confidence inte Number, l	Number, upper limit
SIZE	2	10	0.310667	1553.333	0.142542	-0.05952	0.677856	-262.612	3369.278
SIZE	1	10	0.164	820	0.104227	-0.10449	0.432489	-522.44	2162.44
SIZE	3	10	0.274667	1373.333	0.152774	-0.11688	0.666213	-584.395	3331.063
SIZE	4	10	0.146667	733.3333	0.146567	-0.23069	0.524222	-1154.44	2621.112
SIZE	5	10	0.064	320	0.0639	-0.10061	0.228606	-503.031	1143.031
HLANG	E	10	0.649333	3246.667	0.18308	0.177719	1.120946	888.5954	5604.738
HLANG	F	10	0.310667	1553.333	0.142542	-0.05952	0.677856	-262.612	3369.278

11 If you chose to Stop, a message will appear reminding you to save the calculations before leaving Excel. Click OK to return to Excel. You may want to adjust the column widths of VRESULTS and ARESULTS in order to see the results more clearly.

Method: Cluster Random Sampling

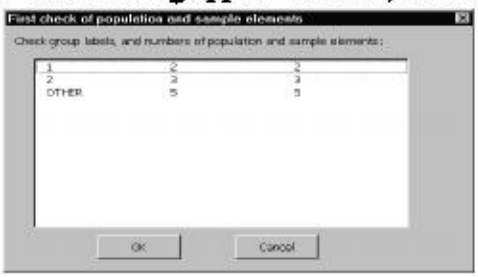
To illustrate the application of SampleCalc, it will be assumed that a population consists of 500 city households grouped into 20 districts, and that the observations were selected by drawing a simple random sample without replacement of three of the six districts; these were the districts 1, 2, and Other. All the households in the selected districts were interviewed.

After selecting Cluster in the Method box, the Cluster random sampling dialog box appears.

1 In the dialog box entitled Cluster random sampling, enter the number of elements in the population, the number of groups (strata) in the population, and the number of selected groups, and select the desired confidence level. Click OK to proceed or Cancel to abort.

2 A dialog box entitled Group Identifier will now appear. Select the label of the one column in the table of data that identifies the group (stratum) to which each sampled element belongs. Click OK to proceed, or Cancel to abort.

3 A dialog box entitled First check of population and sample elements will now appear. It shows the program's understanding of the labels of the selected groups (strata), and of the corresponding number of elements in the population and sample. If you observe an anomaly, click Cancel to abort SampleCalc, check the data, and begin anew. If the program's understanding appears correct, click OK to proceed.



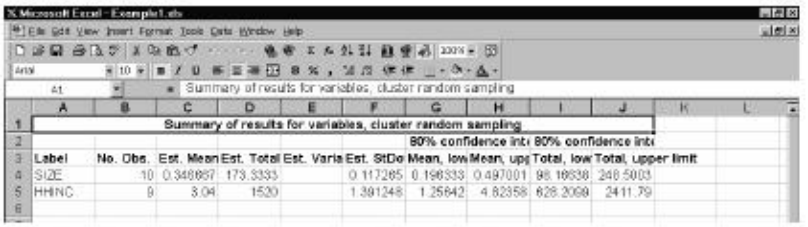
4 The program will display for your information a message to the effect that it is creating the worksheet VRESULTS; acknowledge by clicking OK. A similar second message to the effect that the program is creating the worksheet ARESULTS should also be acknowledged by clicking OK. (VRESULTS will contain the results of calculations concerning the selected variables, while ARESULTS will contain the results regarding the selected attributes.)

5 The dialog box entitled Data analysis will now appear. To analyze one or more variables, select the first option. To analyze one or more attributes, select the second option. To stop, select the third option. You can change the selection at any time before clicking OK. Click OK to confirm your selection and proceed.

6 If you chose to analyze one or more variables, a dialog box entitled **Selection of variables** will appear. Select the variables to be analyzed by clicking on their labels. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the **Data analysis** dialog box.

7 If you clicked OK, you will find that the VRESULTS worksheet now contains for each selected variable the number of observations with non-missing values, the estimates of the mean and total of the variable, the estimated standard deviation of the estimate of the

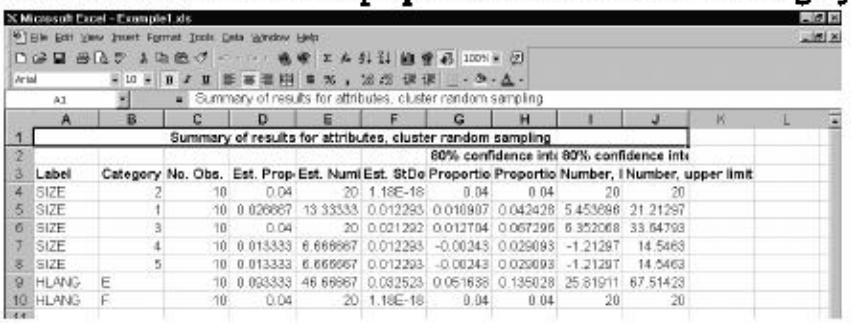
population mean, and the lower and upper limits of the specified confidence intervals for the mean and total of the variable.



Summary of results for variables, cluster random sampling

Label	No. Obs.	Est. Mean	Est. Total	Est. Varia	Est. StDe	80% confidence int: Mean, low	80% confidence int: Mean, up	80% confidence int: Total, low	80% confidence int: Total, upper limit
SIZE	10	0.348667	173.3333		0.117265	0.196333	0.497001	98.18838	248.5003
HHINC	9	3.04	1520		1.391248	1.25642	4.82358	628.2099	2411.79

8 If you chose to analyze one or more attributes, a dialog box entitled **Selection of attributes** will appear. Select the attributes to be analyzed by clicking on their labels in the first list box. (Clicking on a label again cancels the selection. Clicking the Reset button cancels the entire selection.) Click OK to proceed, or Cancel to return to the **Data analysis** dialog box. If you clicked OK, you will find that the ARESULTS worksheet now contains for each category of each selected attribute the number of observations with non-missing values, the estimates of the proportion and number in the category, the estimated standard deviation of the estimate of the population proportion, and the lower and upper limits of the specified confidence intervals for the proportion and number in the category.



Summary of results for attributes, cluster random sampling

Label	Category	No. Obs.	Est. Prop	Est. Numl	Est. StDe	80% confidence int: Proportio	80% confidence int: Proportio	80% confidence int: Number, l	80% confidence int: Number, upper limit
SIZE	2	10	0.04	20	1.18E-18	0.04	0.04	20	20
SIZE	1	10	0.026667	13.33333	0.012293	0.010907	0.042428	5.453696	21.21297
SIZE	3	10	0.04	20	0.021292	0.012704	0.067296	6.352068	33.64793
SIZE	4	10	0.013333	6.666667	0.012293	-0.00243	0.029093	-1.21297	14.5463
SIZE	5	10	0.013333	6.666667	0.012293	-0.00243	0.029093	-1.21297	14.5463
HLANG	E	10	0.093333	46.66667	0.032523	0.051638	0.135028	25.81911	67.51423
HLANG	F	10	0.04	20	1.18E-18	0.04	0.04	20	20

9 If you chose to Stop, a message will appear reminding you to save the calculations before leaving Excel. Click OK to return to Excel. You may want to adjust the column widths of VRESULTS and ARESULTS in order to see the results more clearly

Glossary

Sampling With and Without Replacement: Sampling is said to be with or without replacement according to whether or not an element or group of elements can appear more than once in the sample.

Simple Random Sampling: One method of selecting a simple random sample is by a series of draws, in every one of which each eligible element has the same chance of being selected, and each selection is unrelated to (independent of) other selections. Such a sample is without replacement if the element selected in any draw is not eligible for selection in any subsequent draw.

Stratified random sampling: Requires that the elements of the population be grouped according to one or more criteria. A stratified random sample consists of simple random samples without replacement from each and every group (stratum).

Two-Stage Random Sampling: Requires that the elements of the population be grouped according to one or more criteria. As the name implies, the sample is selected in two stages. In the first, a simple random sample of groups (strata) is selected. Then, in the second stage, a

simple random sample of elements is selected from each group (stratum) that was selected in the first stage.

Cluster Random Sampling: Cluster sampling can be viewed as a special case of two-stage sampling. A cluster sample requires that the elements of the population be grouped according to one or more criteria. In the first stage, a simple random sample of groups (strata) is selected. Then, in the second stage, *all* the elements are selected from each group (stratum) that was selected in the first stage.

Variables and Attributes: A variable is a feature or aspect of an element that lends itself naturally to a numerical description. For example, the age of a person, the income of a household, the maximum temperature in a day, etc. Unlike a variable, an attribute is a feature or aspect of an element that lends itself only to a categorical, qualitative—not numerical—description. For example, a person's gender (male, female), a household's location (in, say, district A, B, or C), the industrial classification of a manufacturing company, etc.

Population Characteristics: The population characteristics of greatest interest in practice are the mean or total of a variable, and the proportion or number in a category of an attribute.

Point and Interval Estimates: Rather than, or in addition to, saying that a population characteristic is estimated to have such and such a value (the point estimate of the characteristic), it may be informative to state that a certain interval is estimated to contain a population characteristic with a certain probability. For example, instead of saying that the population mean of a variable is estimated to be 7.3, it may be informative to state that the population mean of the variable is in the interval from 6.5 to 8.1, where this statement is correct with probability 95%. The probability is often referred to as the confidence level and the interval estimate as a confidence interval.

Unbiased Estimates: An estimate (more precisely, an estimator) of a population characteristic is said to be unbiased if its average value in a very large number of identical samples is equal to the population characteristic.

REFERENCES

- http://www.yorku.ca/ptryfos/index.htm
- http://www.yorku.ca/ptryfos/compprog.htm
- http://office.microsoft.com/en-in/excel-help/about-add-in-programs-HP005238607.aspx

Statistical Computing Using Microsoft Excel

What is a Spreadsheet

A spreadsheet is a grid of rows and columns that helps organize, summarize, and calculate data. Spreadsheets are an everyday part of many professions, including accounting, statistical analysis, and project management. You can use Excel to create business forms, such as invoices and purchase orders, among many other useful documents. Microsoft Excel is a spreadsheet program that is capable of performing these operations along with many more additional functions.

To open Microsoft Excel click on **Start, All Programs,** and **Microsoft Excel.** Users of Windows 7 or above may click on the windows image available on bottom left corner and type "Excel" to find MS Excel (Fig. 1).

Fig. 1: Start button on Desktop of Windows 8

Let's look at the toolbars (Fig. 2).

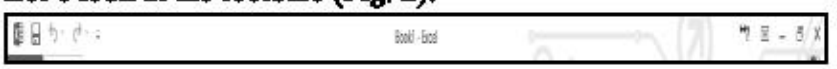


Fig. 2: MS Excel toolbar

The picture above shows the **Title Bar.** In the middle of the toolbar name of the program and the title of the workbook you are using is shown. Since we have just opened up a new workbook and have not saved it with a name, the default title is Book1.

Fig. 3: MS Excel Ribbon

Next we have the **Ribbon.** The **Ribbon** has seven **Tabs** that give instructions to the software. The **Ribbon Tabs** begin with **Home** and continue with **Insert, Page Layout, Formulas, Data, Review,** and **View.** On the right-hand end, there is an icon for the Help Menu, Minimize, Restore Down, and Close. Clicking on one of these Tabs will open the **Group.** The **Group** that belongs to each **Tab** shows related Command items together. You may then choose a Command.

Workbooks and Worksheets

When Excel is opened, a workbook appears with three worksheets. Each worksheet contains columns and rows. There are **1,048,576 rows and 16,384 columns.** The combination of a column coordinate and a row coordinate make up a cell address. For example, the cell located in the upper left corner of the worksheet is cell A1, meaning column Aand row 1. The cell address is visible in the **Name Box.**

Place your cursor in the first cell, A1. The formula bar will display the cell address in the **Name Box** on the left side of the **Formula** bar. Notice that the address changes as you move around the sheet. You can easily move from cell to cell by pressing tab or using the arrow keys.

A cell can contain any of the following:

- A number (and any associated punctuation, such as decimal points, commas, and currency symbols).
- Text (including any combination of letters, numbers, and symbols that aren't number-related).
- A **formula,** which is a math equation.
- A **function,** which is a named equation that shortcuts an otherwise complex operation.

Creating a New Workbook

It is easy to create a new workbook! Simply, click on ***File– New*** and click on ***Blank Workbook*** to create a new workbook.

Creating a New Worksheet

Creating a new worksheet is just as easy. By default, each Excel workbook contains three worksheets. Three tabs displaying ***Sheet 1***, ***Sheet 2***, and ***Sheet 3*** will be displayed at the bottom of the workbook to indicate the separate sheets. Microsoft office 2013 shows only one worksheet with a circled plus sign for adding new worksheets. To add a new worksheet, simply click on the tab after the tab that says Sheet 3.

Navigating and Selecting

Moving around a worksheet is easy! You can easily move from cell to cell by using the arrow keys or pressing tab (will move the cursor to the right) or shift-tab (shift-tab will move you to the left). You can also use your mouse to click within a cell which will select that cell. Sometimes you will want to select a **range** of cells.

A **range** is a group of one or more cells. If you select more than one cell at a time, you can then perform actions on the group of them at once, such as applying formatting or clearing the contents. A range can even be an entire worksheet.

A range is referenced by the diagonal corners i.e. upper left and lower right cells. For example, the range of cells B1, B2, C1, and C2 would be referred to as B1:C2.

To select a range:

- **Using the Mouse:** Drag across the desired cells with the left mouse button held down. Be careful when you're positioning the mouse over the first cell (before pressing the mouse button). Position the pointer over the **center** of the cell, and not over an edge.

If you drag while the pointer is on the edge of the cell, Excel interprets the selection as a move operation and whatever is in the cell(s) is dragged to a different spot.

- **Using the keyboard:** Select the first cell, and then hold down the **Shift** key while you press the arrow keys to expand the selection area.

To select a non-rectangular or noncontiguous range, select the first portion of the range (that is, the first rectangular piece), and then hold down the **Ctrl** key while you select additional cells/ranges with the mouse.

To select an entire column, click the column header (where the letter is). To select an entire row, click the row header (where the number is). You can click one row or column and then drag to select additional columns, or hold down **Ctrl** as you click on the headers for non-contiguous rows and/or columns.

Entering and Editing Data

Let's learn how to enter data into your worksheet. First, you place the cursor in the cell in which you would like to enter data. Then you type the data and press Enter.

You can also edit information in a cell by double-clicking in a cell or by clicking in the formula bar for editing the active cell. You can try these two options for yourself.

Inserting Columns and Rows

If you don't plan your worksheet layout correctly, you might end up with too many or too few rows or columns in a certain area. You can always move data around in the sheet to help with this, but sometimes it's easier to simply insert or remove columns or rows.

Formatting Columns and Rows

Often you will need to change your columns and rows in order for text to fit or for the text to fit on the page correctly. There are a number of different methods one can use to do this. Let's start with columns.

Column Width: The formatting that is unique to columns is **Column Width. Column Width** is measured in characters. A column's width can be from 0 to 255 characters, which is a **really** wide column! Decimal values are allowed. In fact, the default size is 8.43 characters. A width of 12, for example, means the column is wide enough for 12 average characters, using whatever you chose as the Standard font. The default is Calibri 11 pts. To change the font from the default, go to ***Tools-Options-General-Standard font.***

Column Width

Be careful when you set a column's width with ***AutoFit.*** The column may wind up wider than you expected. Any text will be on a ***single line*** in its cell. No matter how long the text is! If you accidentally find you've widened a cell out of sight to the right, use ***Undo.*** (my favorite button!) Then resize the column with another method.

Column Width - Drag

Dragging is a natural method of adjusting column width. But since you can't see the change until you release the mouse button, it may take you several attempts to get a satisfactory width.

Row Height

The only unique formatting for rows is ***Row Height. Row Height*** is measured in points, like font size, from 0 to 409 points. A row height of zero hides the row.

The default setting for **Row Height** is ***AutoFit.*** The row height adjusts to the largest font size in the row.

AutoFit will leave a little white space, called the cell padding, between the text in the cell and the cell edges. When Arial 10 pt. is the Standard Font, the ***Row Height*** is **12.75 points.** You may find that this looks a bit crowded when the gridlines are shown. If you don't print the gridlines, your paper version will look OK.

Moving to a New Worksheet

In Microsoft Excel, each workbook is made up of several worksheets. Before moving to the next topic, let's move to a new worksheet. You can move from worksheet to worksheet by clicking on the tabs at the bottom of the worksheet. Let's move to **Sheet 2.**

Formatting Text and Data

Once information has been entered into a cell, you might want to change or enhance the way the information is displayed. Text can be formatted in the same way that one uses in

Microsoft Word or PowerPoint. Most of the formatting choices can be found in the **Font** grouping under the **Home** tab. There are numerous ways to format data. Let's look at some. First remember to always make sure that the cell you want to format is selected.

Using Formatting Buttons– On the **Ribbon,** make sure the Home tab is selected. In the **Number Group** box, there are several buttons which allow one-click formatting.

Notice how each number changes depending on the formatting.

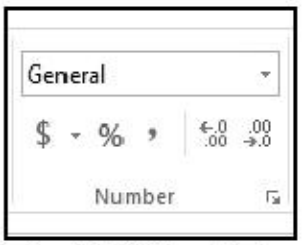


Fig. 4: MS Excel Formatting

Formatting Numbers

Let's look at other formatting options.

After Formatting

Let's change it to a dollar amount.

1. Make sure that the cursor is in cell A5.
2. Right-click again.
3. Click on *Format Cells*.
4. Click Currency in drop down menu.
5. Look at the options available including currency symbols.

Deleting vs Clearing a Cell

Many beginners get confused about clearing versus deleting in Excel, so let's look at this concept briefly. When you clear the content from a cell, the formatting for that cell is still there. It may be helpful to think of an Excel worksheet as a stack of empty cardboard boxes, each one with its open side facing you. You can put something into a cell or take something out. When you take something out of a cell, it's called clearing its content. The cell itself remains in the "stack, "but it's now empty.

Fig. 5: MS Excel Clear options

To clear the content from a cell:

1. Press Delete on the keyboard.
2. Right-click the cell and then select Clear Contents.
3. On the Home tab, in the Editing group, select Clear > Clear Contents.

Unfortunately, clearing a cell's content doesn't clear its formatting.

To clear formatting:

1. On the Home tab, in the Editing group, select Clear > Clear Formats
2. To clear both contents and formats at once, select Clear All.

In contrast, deleting the cell removes the cell itself from the stack and makes the surrounding cells shift. Think about what happens when you pull a box out of a stack of boxes—the boxes above it fall down one position, right? It's the same thing with Excel cells, except it's reverse-gravity (cells fall up rather than down), and you have the choice of making the remaining cells shift up or to the left. Let's look at how this works.

Filling Cells Automatically

You can use Microsoft Excel to fill cells automatically with a series. For example, you can have Excel automatically fill in times, the days of the week or months of the year, years, and other types of series. Days of the week and months of the year fill in a similar fashion.

Merging Cells

Sometimes, rather than having text wrap in a cell, you will actually want the text to run across the width of the data. Usually when making a spreadsheet, you need to create a heading for the sheet. This heading should run across the width of your data. To do this, one must merge the cells across the width of the data.

Performing Calculations

Let's add a column of numbers using the **AutoSum** Button∑. To select the **AutoSum** button choose **Home> Editing>**∑and automatically add a column of numbers.

What's a Formula?

A formula is an equation that performs some type of operation and issues a result. In Excel, formulas always begin with an equal sign. Here are some formula examples:

- =2+6: This formula is strictly math. If you place this formula in a cell, the cell displays 8.
- =A1+6: Same as the preceding, but this time you're adding 6 to whichever value is in cell A1 and displaying the result in the cell into which you enter this formula. This formula does not change A1's contents.
- =A1+A2: Same thing again, but you're adding the contents of cell A1 to the contents of cell A2.
- =A1+A2-A3: In this example, multiple cells are referenced.

Here are the symbols you can use in formulas to indicate mathematical operations:

- +: Addition
- -: Subtraction
- *: Multiplication
- /: Division

More Formula Examples

The math operators in Excel have an order of operation, just like in regular math. The order of operation is the order in which they're processed when multiple operators appear in the same formula. Here are the rules that determine the order:

1. Any operations that are in parentheses, from left to right
2. Multiplication (*) and division (/)
3. Addition (+) and subtraction (-)

Parentheses override everything and go first. So, if you need to execute an operation out of the normal order, you place it in parentheses. Now let's try some formula examples that refer

to cells and use math operations. For this exercise, enter the following values in cells in a blank worksheet:

A1: 12 A2: 6 A3: 4 A4: 9

Printing

Let's prepare to print! If your worksheet is more than one printed page, it is possible to have the heading on each page by going to the **Page Layout tab, in the Page Setup** group and click **Print Titles.**

Fig. 6: MS Excel Print area setting

On the **Sheet tab, under Print Titles,** do one or both of the following:

In the **Rows to repeat at top box,** type the reference of the rows that contain the column labels if you want the heading repeated on each page.

In the **Columns to repeat at left box,** type the reference of the columns that contain the row labels if you want those to show.

We want our sheet to print with no gridlines, and centered horizontally across the page, but not vertically. Let's go the *Page Layout* > *Sheet Options*. There should not be a check under **Print** in the Gridline section.

Make sure that you have checked your spelling and made any necessary corrections. Click on the **Office Button** and **Print>Print Preview** (Always do a print preview in Excel!). Click on Page Setup>Margins and make sure that there is a check under **"Center on Page" >horizontally.** Now let's print!!!

Fig. 7: MS Excel Print options

→→

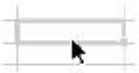
Recognizing Cursor Styles

Click and drag to highlight multiple cells with this cursor, or click

Click and drag with this cursor to fill cell contents into cells

in a cell to select the single cell

below or to the right.



Click and drag the contents of the selected cell to any other cell.

Click to place the cursor into the Formula bar so that you can edit an equation or function.

There are four common cursor styles used in Excel.

Common formula errors

Here are some of the most common mistakes people make when entering formulas and functions:

- Not putting in all the required arguments: If a function is expecting more arguments than you have entered, and you get a dialog box, be sure you've placed commas between the arguments and that you haven't overlooked any.

Fig. 8: Error due to circular reference in MS Excel

- Circular references: If you refer to the cell's own address in a function, you create a circular error, which is like an endless loop. Suppose that you enter =A1+1 into cell A1. You'll get an error message like the one below. If you click OK at this message, a Help window appears to help you find the problem.
- Text in an argument: Most functions require numeric arguments. If you enter text as an argument, for example, =SUM(text), the word #NAME? appears in the cell. This happens because Excel allows you to name ranges of cells using text, so technically =SUM(text) isn't an invalid function. It is invalid only if there's no range that has been assigned the name "text."
- Hash marks (###) in a cell: This happens when the cell isn't wide enough to display its value. Widen the column to fix this.

If you receive an error when copying a formula, don't panic; it happens to everyone. Use the skills you learned earlier in this chapter to display the formulas and then check them for the common errors discussed here.

Using Functions in Excel

Let us see the use of Excel to get sample statistics and how to make a histogram. You will learn how to use Excel formulas and take advantage of the formula lookup function.

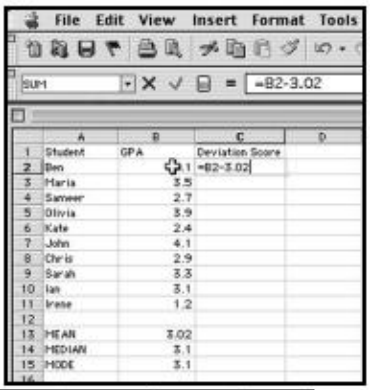
There are two ways to start Excel. The first is to open an Excel file (extension .xls or .xlsx). The second is to open the program directly from the menu bar. Open Excel and type the data below into the spreadsheet. Save the file as the "test data" in a folder.

1. Get the average of these scores by typing this formula into the cell B12: "=average(B2:B11)" and hitting enter.
2. Get the median of the scores. This time use the "Insert" -> "Function" menus to select the median function. These menus can be helpful when you don't remember the exact command for a function.
3. Get the mode. Can you guess the formula?

4. Calculate a column of deviation scores in column C. First, type the formula "=B2-3.02" into cell C2. Then highlight the cell, click "copy", highlight the rest of the column, and click "paste."

5. Calculate a column of squared deviation scores in column D.

	A	B
1	Student	GPA
2	Ben	3.1
3	Maria	3.5
4	Sameer	2.7
5	Olivia	3.9
6	Kate	2.4
7	John	4.1
8	Chris	2.9
9	Sarah	3.3
10	Ian	3.1
11	Irene	1.2
12		



	A	B	C	D
1	Student	GPA	Deviation Score	Squared Deviation
2	Ben	3.1	0.08	=(C2)^2
3	Maria	3.5	0.48	
4	Sameer	2.7	-0.32	
5	Olivia	3.9	0.88	
6	Kate	2.4	-0.62	
7	John	4.1	1.08	
8	Chris	2.9	-0.12	
9	Sarah	3.3	0.28	
10	Ian	3.1	0.08	
11	Irene	1.2	-1.82	

You can then get the sum of squared deviations, the variance, and the standard deviation from column D.

In cell D13 put "=sum(D2:D11)"
In cell D14 put "=D13/9"
In cell D15 put "=sqrt(D14)"

Make sure you understand *why* those formulas yield the SS, variance, and standard deviation.

6. Of course, there is a faster way to calculate the SS, the variance, and the standard deviation directly using Excel commands for these statistics. See if you can do this using the "Insert->Function" menus. (Hint: look for sums of squared deviations under D not S).

7. What if you wanted to get other measures of position for each score? (You've already got deviation scores). Let's get percentiles for each score. In E2, type this formula: "=percentrank($A$2:$A$11, A2)". The range of cells before the comma tells Excel which cells contain the full dataset. The cell after the comma tells Excel the value for which you'd like to get a percentile. The dollar signs tell Excel that when you copy and paste the formula, those cell references should remain absolute.

8. Let's get z-scores for each score. Figure out two different ways to do this on your own (step by step using multiple columns and quickly using a single formula).

9. Show that the z-scores do not change when you add a constant to each score or multiply each score by a constant.

10. What if we wanted to look at the inter quartile range? To do this, we need to identify the 25th and 75th percentiles. In (8) we went from scores to percentiles. Now we want to do the

reverse. Somewhere in your worksheet, type "=percentile(A2:A11, .25)". This will give you Q1 (the 25th percentile). Use the formula to get Q2 (50th percentile or median) and Q3 (75thpercentile).

Of course, in Excel there are multiple ways to accomplish the same thing. Check out the MEDIAN () and QUARTILE () functions.

11. How would the box-plot look for this data?

12. Use Excel to make a histogram

Click on Tools → Data Analysis. A new window should open, scroll down to Histogram, click OK. The dialogue box will ask you for an input range, a bin range, and other stuff. For the input range, highlight the column of data points in column A (or type the cell range directly). Leave the bin range blank for now – Excel will determine the bin sizes for you. Don't forget to check off "Chart output" so that it gives you a chart, that's the whole point. Hit OK and see what happens.

Get rid of the spaces between the bars by highlighting the chart, then double clicking on the bars themselves. A new window should open, use the tabs to choose "Options" and it will ask for overlap and gap width. *Overlap* is usually set to 0 and *Gap width* is usually set to 150. Change *Gap width* to zero. Resize the histogram by clicking on the white area surrounding the chart, then dragging one of the corners down.

How many intervals would be best for this data? A number that neither hides too much information nor presents too much information. Too many intervals produce a histogram that is cumbersome and a poor summary of the data; too few produce a histogram that fails to summarize because it loses too much important detail. Remember: it's a judgment call.

To find the optimal bin width, consider the range of data by sorting (highlight the data, go to the Data menu, and choose Sort). Check out the minimum and the maximum. Think about what intervals could be used for this data set. Take these points into consideration:

- all intervals should be the same width
- the entire range should be covered
- the bottom score of each interval should be a multiple of the interval width (e.g. if the interval width is 5 You should not start at 12 but rather at 10)
- the interval should reflect the conventions of the research area and make sense (especially make sense)

Now, try setting the intervals yourself. To do this, you need to create a column that contains the values you want on your x-axis (the "bin bottoms"). Then find the Histogram dialogue box again, enter the input range, and then move the cursor to "Bin range" and highlight your bin column. Hit return. Did it work? Try again with different bins. What information are you losing? What are you gaining? Find an optimal bin interval and describe why you think it's so good.

REFERENCES

Baycon Group: http://www.baycongroup.com/el0.htm

Cheat Sheet – Excel 2010 Keyboard Shortcuts and Ribbon Commands http://www.dummies.com/how-to/content/excel-2010-allinone-for-dummies-cheat-sheet.html

Computing (Excel statistics/modelling)

Get Started with Excel 2010 (also available through File/Help/Getting Started): http://office.microsoft.com/en-us/excel-help/getting-started-with-excel-2010-HA010370218.aspx?CTT=1

Goodwill Industries: http://www.gcflearnfree.org/

HP Learning Center: http://h30187.www3.hp.com/

http://cloudbase.phy.umist.ac.uk/people/connolly

Internet4Classrooms: http://www.internet4classrooms.com/on-line_excel.htm

Make the Switch to Excel http://office.microsoft.com/en-us/excel-help/make-the-switch-to-excel-2010-RZ101809963.aspx

Menu to Ribbon guide – 2003 to 2010 http://office.microsoft.com/en-us/templates/word-2010-menu-to-ribbon-reference-workbook TC101817139.aspx?CTT=5&origin=ZA101796062

Microsoft: http://office.microsoft.com/en-us/excel/FX100646951033.aspx

Hypothesis Testing in Excel

Hypothesis testing refers to the process of choosing between competing hypotheses about a probability distribution, based on observed data from the distribution. It is a core topic in mathematical statistics, and indeed is a fundamental part of the language of statistics. In this chapter, we study the basics of hypothesis testing, and explore hypothesis tests in some of the popular test using Excel.

Test 1: Z-test for a population mean (variance known)

Object: To investigate the significance of the difference between an assumed population mean μ_0 and a sample mean $\bar{x}$

Method: From a population with assumed mean μ_0 and known variance σ^2, a random sample of size *n* is taken and the sample mean *x* calculated. The test statistics

$$Z = \frac{x - \mu_0}{\sigma / \sqrt{n}}$$

may be compared with the standard normal distribution using either a one or two tailed test, with critical region of size α.

Limitations:

1. It is necessary that the population variance σ^2 is known.
2. The test is accurate if the population is normally distributed. If the population is not normal, the test will still give an approximate guide.

Example:

For a particular range of cosmetics a filling process is set to fill tubs of face powder with 4 gm on average and standard deviation 1 gm. A quality inspector takes a random sample of 9 tubs and weights the powder in each as follows: 3, 5, 4, 3, 4, 7, 5, 6, 5. what can be said about the filling process?

Solution: $H_0 = \mu = \mu_0 = 4, \quad \sigma = 1$

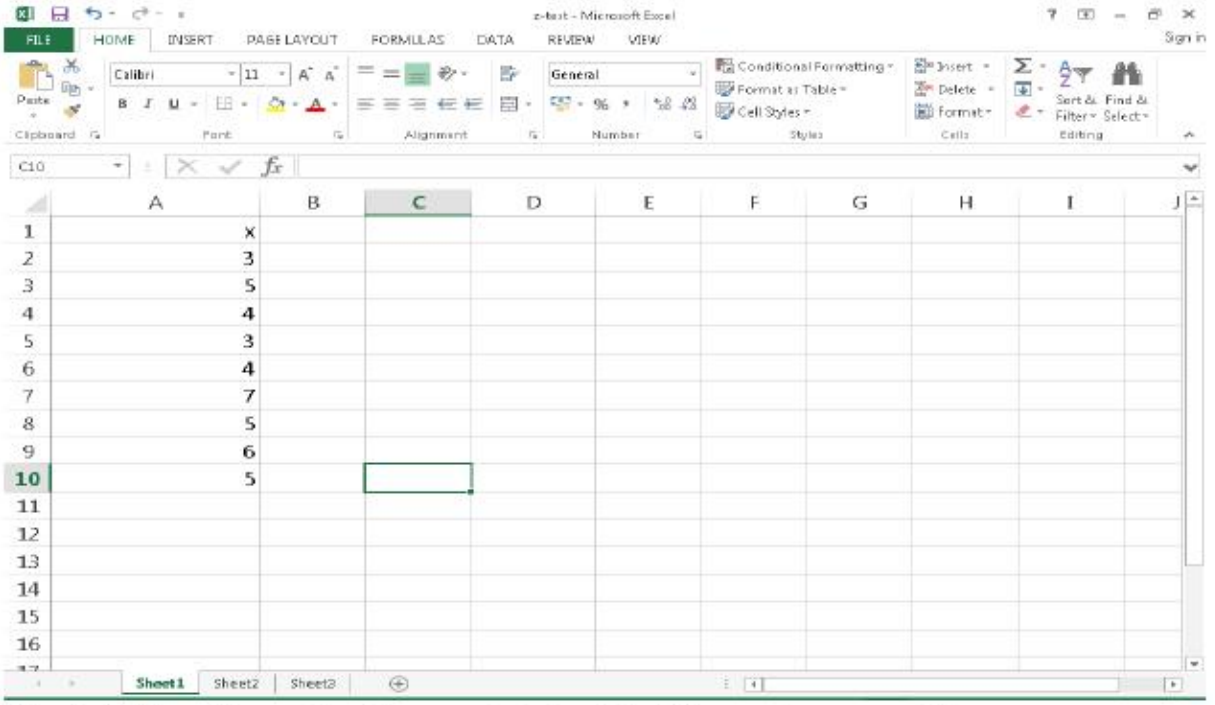
$H_1 = \mu > 4$ (One Tailed)

$H_1 = \mu < 4$ (One Tailed)

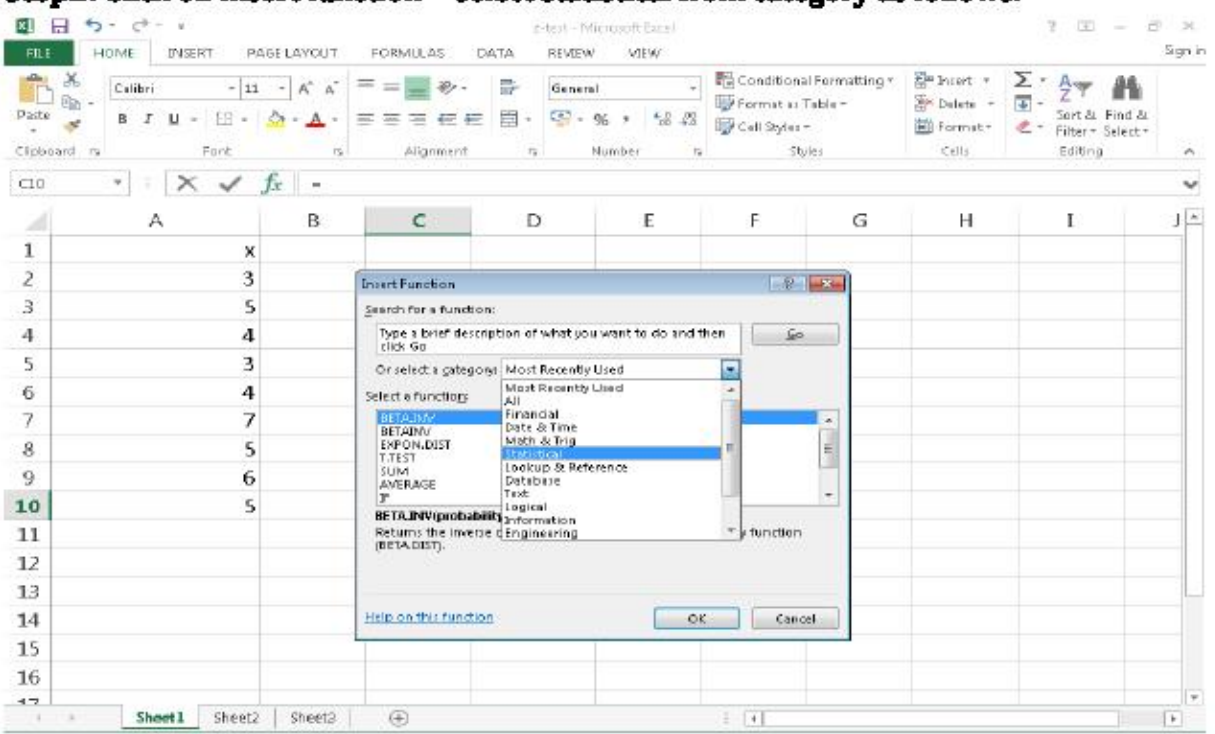
$H_1 = \mu \neq 4$ (Two Tailed)

To apply Z-test for the above, we proceed in the following steps:

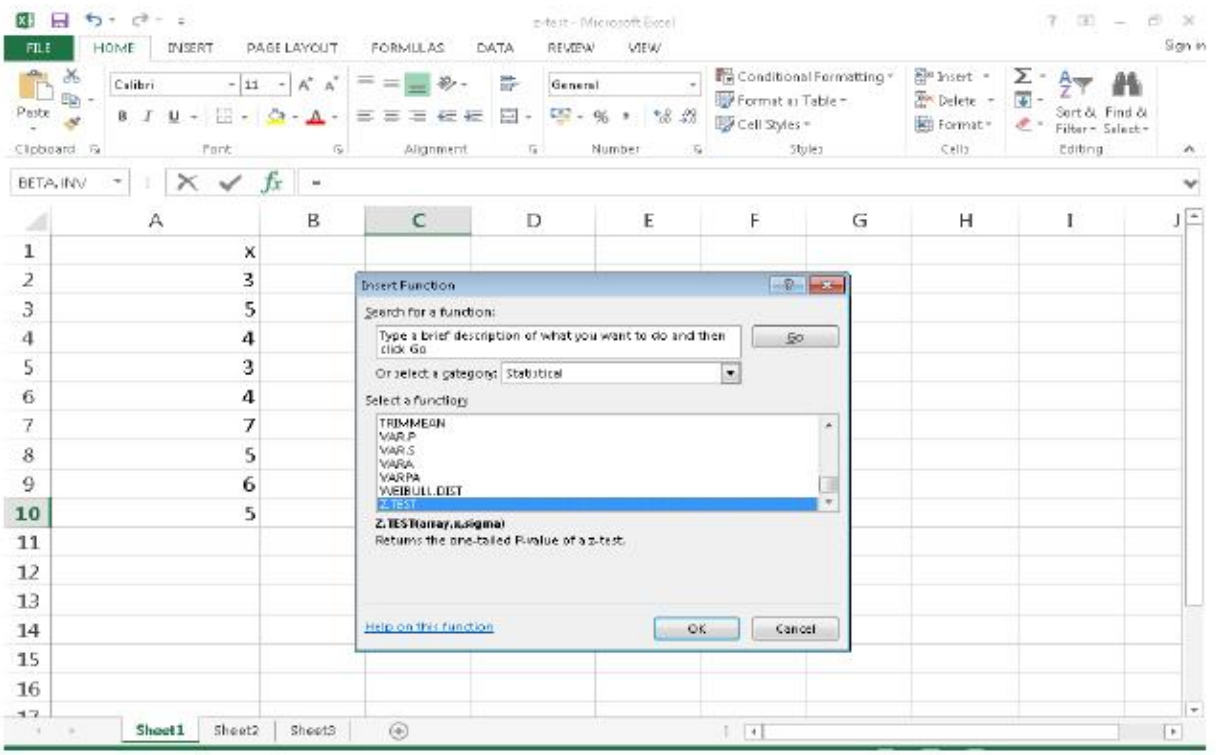
Step1: First of all, enter the data i.e. the weights of 9 tubes as follows:



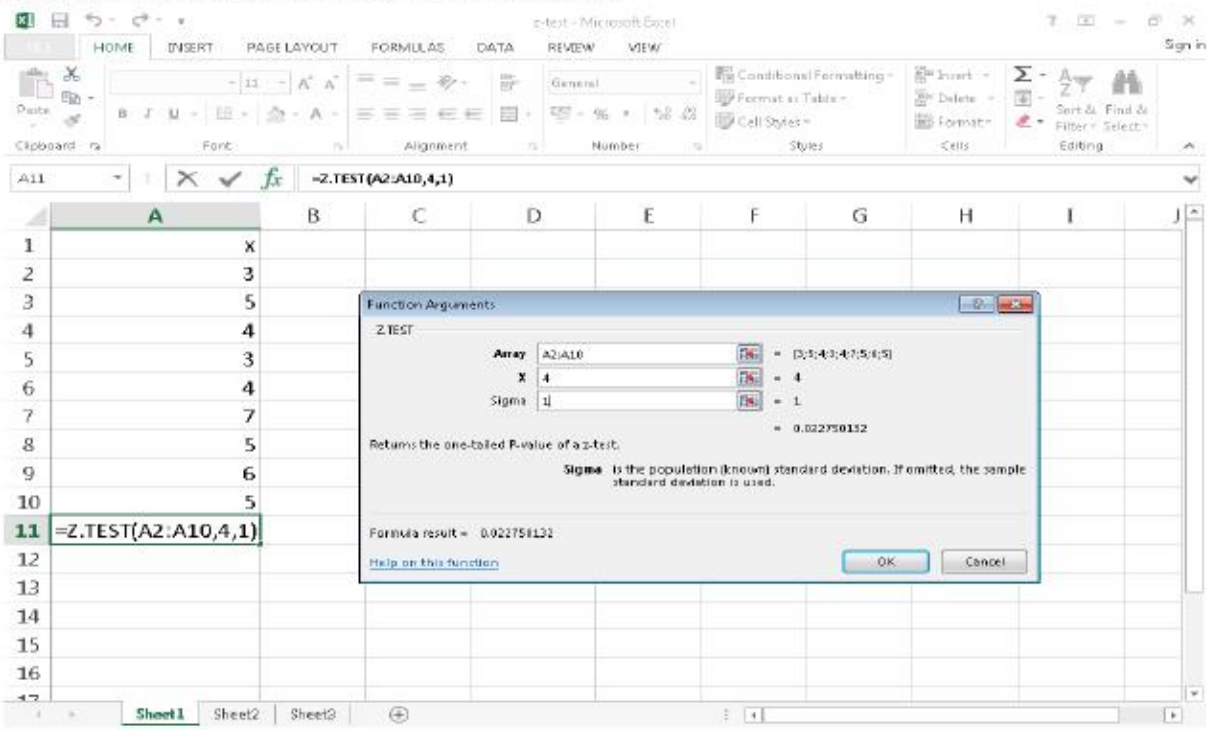
Step2: Click on Insert function →select statistical from category as follows:



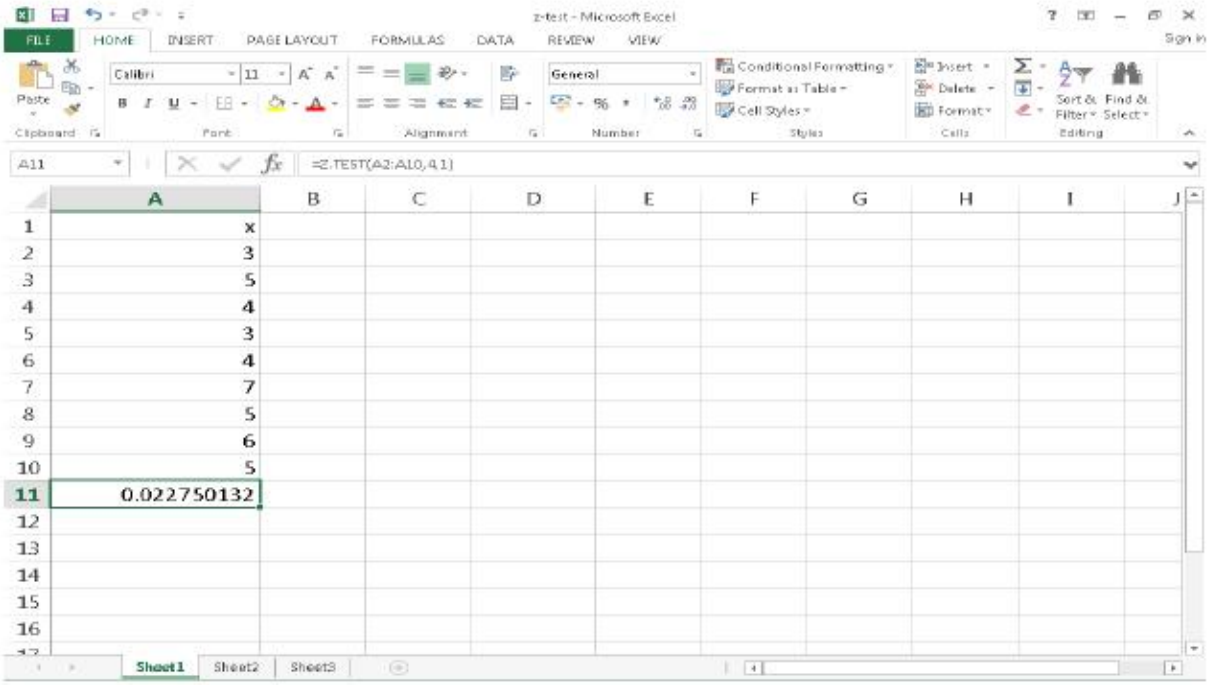
Step3: select a function z-test



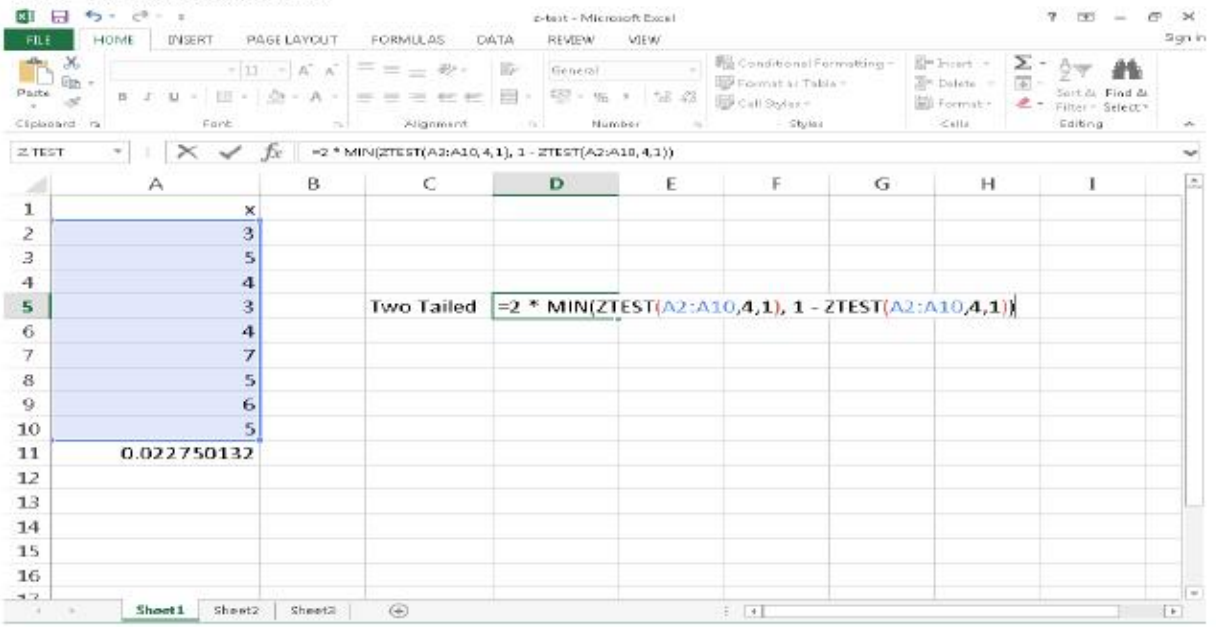
Step 4: In Array, means the data given, so select data from A2:A10, the value of X is the value of population mean at which testing is to be performed, so select 4 and Sigma i.e. the standard deviation which is 1 in this case.



Step 5: click ok, we get the p-value of z test.



P-value for z-test is 0.022750132 (One Tailed). Their is no inbuilt function in excel to calculate the two-tailed p value for z test, however one can calculate manually the p-value for two tailed test as follows: 2 * MIN(ZTEST(A2:A11,4,1), 1 - ZTEST(A2:A11,4,1)) which becomes 0.0455..



Interpretation of output: Since, for one tailed p-value=0.02275<0.05, hence we reject the null hypothesis against one tail alternative as well as for two tailed alternative.

Remark: Note that, in the case variance σ^2 (standard deviation σ) is not known, we can use its estimate i.e. sample variance s^2 (sample standard deviation s) in place of variance σ^2 (standard deviation σ). Also, however z is test is use in the case of large sample (n>=30), in the above example sample size is just 9, so don't bother about this. We have taken small sample for ease only.

Test 2: z-test for two populations means (variances known)

Object: To investigate the significance of the difference between the means of two populations.

Limitations:

1. Both populations must have equal variances and this variance σ^2 must be known. (If σ^2 is not known, apply the *t*-test for two populations means)
2. The test is accurate if the populations are normally distributed. If not normal, the test may be regarded as approximate.

Method: Consider two populations with means μ_1 and μ_2. Independent random samples of size n_1 and n_2 are taken which give sample means $\bar{x}_1$ and $\bar{x}_2$. The test statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

may be compared with the standard normal distribution using either a one or two tailed test.

Example: The values of two samples of 35 and 40 members are

Sample 1 (in inches): 60.37 67.60 64.29 66.98 66.98 67.87 65.44 64.18 66.59 63.57 64.72 70.33 69.53 63.57 66.86 69.41 67.23 67.03 68.46 68.31 64.61 68.79 65.76 71.21 62.35 66.96 66.76 64.24 66.51 65.45 73.36 67.30 66.02 66.97 71.13

Sample 2 (in inches): 60.87 68.10 64.79 67.48 67.48 68.37 65.94 64.68 67.09 64.07 65.22 70.83 70.03 64.07 67.36 69.91 67.73 67.53 68.96 68.81 65.11 69.29 66.26 71.71 62.85 67.46 67.26 64.74 67.01 65.95 73.86 67.80 66.52 67.47 71.63 66.27 67.02 68.38 69.91 69.94

Can the samples be regarded as drawn from the same population of standard deviation 2·5 inches?

Solution: We are given: n_1=35 and n_2=30

Null hypothesis: $H_0 = \mu_1 = \mu_2$ *and* $\sigma = 2.5$ i.e. the samples have been drawn from the same population of standard deviation 2.5 inches.

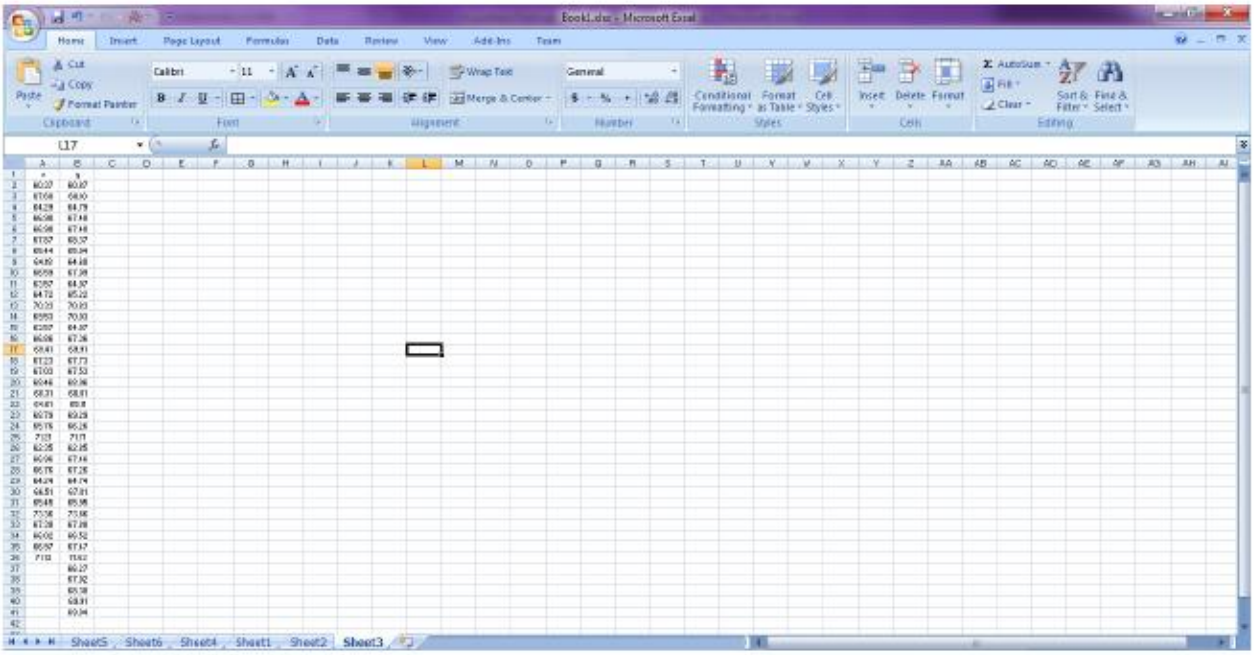
Alternative hypothesis: $H_1 : \mu_1 > \mu_2$ (One Tailed)

$H_1 : \mu_1 < \mu_2$ (One Tailed)

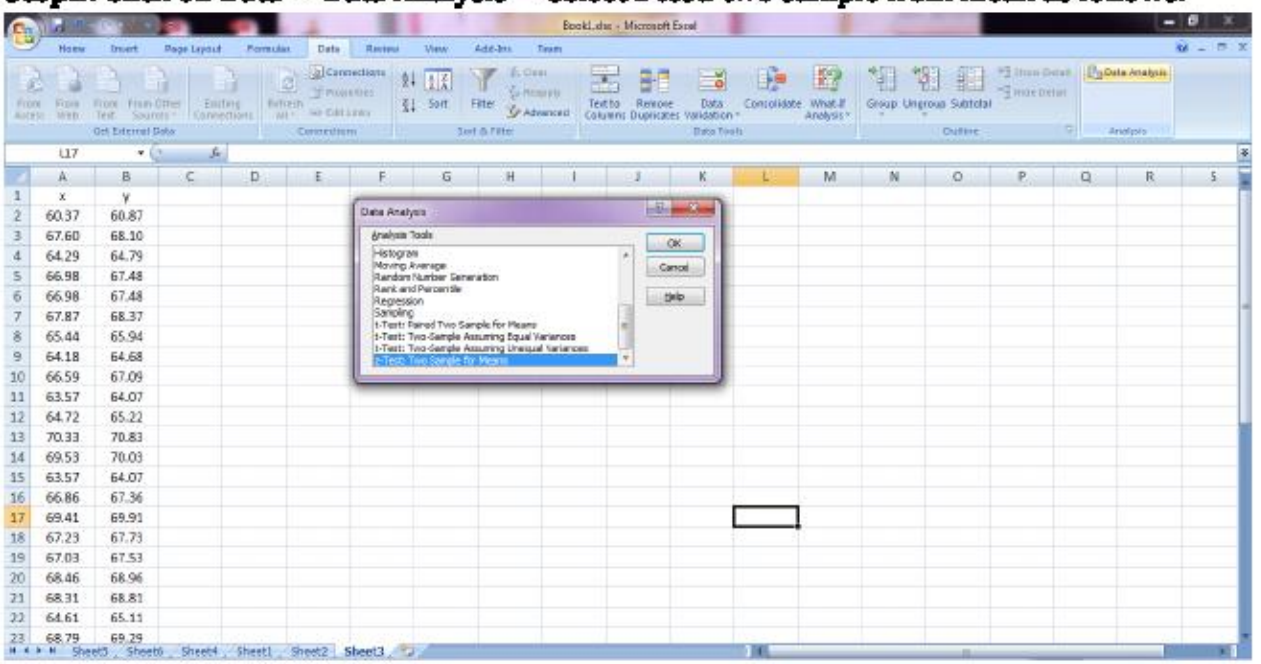
$H_1 : \mu_1 \neq \mu_2$ (Two Tailed)

Now, to apply z-test, we proceed in the following steps:

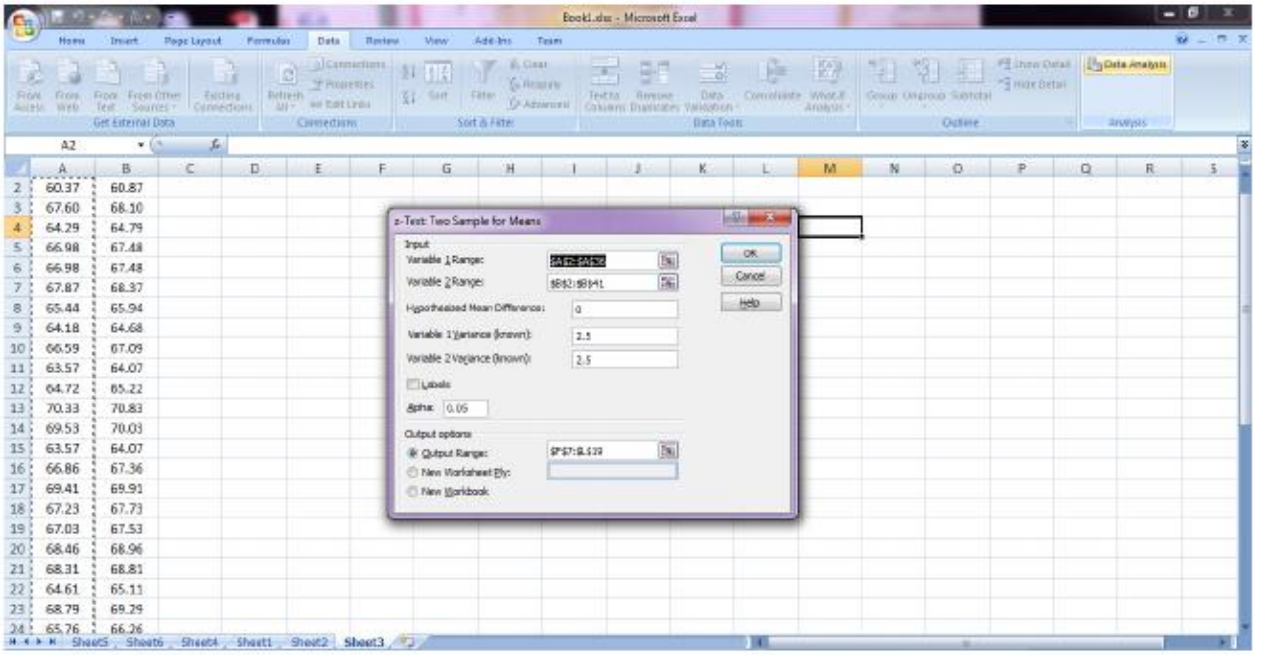
Step1: First of all, enter the data i.e. the values of two samples as follows:



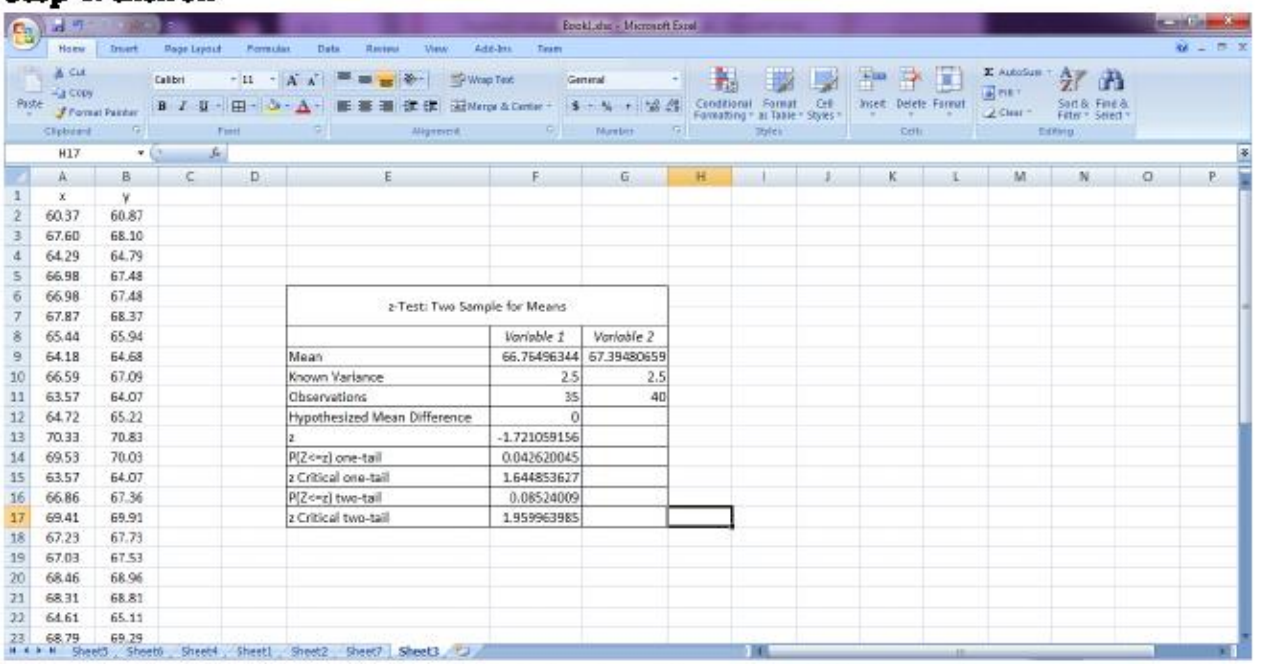
Step2: Click on Data → Data Analysis → select z test: two sample from mean as follows:



Step3: Click ok, Input variable 1 Range, variable 2 Range, Hypothesized mean difference (0 in this example under null hypothesis), variable 1 and variable 2 variance (both are equal to 2.5 our case), Alpha (the level of significance), output option (provide output range if one wants to be output on the same sheet).



Step 4: click ok



z-Test: Two Sample for Means		
	Variable 1	*Variable 2*
Mean	66.76496344	67.39480659
Known Variance	2.5	2.5
Observations	35	40
Hypothesized Mean Difference	0	
z	-1.721059156	
P(Z<=z) one-tail	0.042620045	
z Critical one-tail	1.644853627	
P(Z<=z) two-tail	0.08524009	
z Critical two-tail	1.959963985	

Description of output: Since, calculated value of mode z in 13th row (for one tailed test) is greater than 1.644853627 (i.e. critical value for one tailed test at 5 percent level of significance). Same evidence is also found from p-value=0.042620045<0.05. Hence, null hypothesis is rejected for one tailed test at 5% level of significance. Now, for two tailed test, since calculated value of mode z (-1.721059156) is less than 1.959963985, hence we accept null hypothesis for two tailed test i.e. the two sample can be regarded from same population. Same evidence is also observed from p-value as p=0.08524009>0.05.

Test 3: *t*-test for two populations mean (Independent sample and variances unknown)

Object: To investigate the significance of the difference between the means of two populations.

Method: Consider two populations with means μ_1 and μ_2. Independent random samples of size

n_1 and n_2 are taken from which sample means $\bar{x}_1$ and $\bar{x}_2$ and variances

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 \text{ and } s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_i - \bar{x}_2)^2$$

The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

which may be compared with Student's t-distribution with degrees of freedom $n_1 + n_2 - 2$

Limitations:

1. If the variances of the populations are known, a more powerful test is available: the Z-test for two population means.
2. The test is approximate if the populations are normally distributed or if the sample sizes are sufficiently large.
3. The test should only be used to test the hypothesis $\mu_1 = \mu_2$

Example: Below are given the gain in weights (in lbs.) of pigs fed on two diets x and y.

Gain in weight

Diet x : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet y : 44,34,22,10,47,31,40,30,32,35,18,21,35,29,22

Test if the two diets differ significantly as regards their effect on increase in weight.

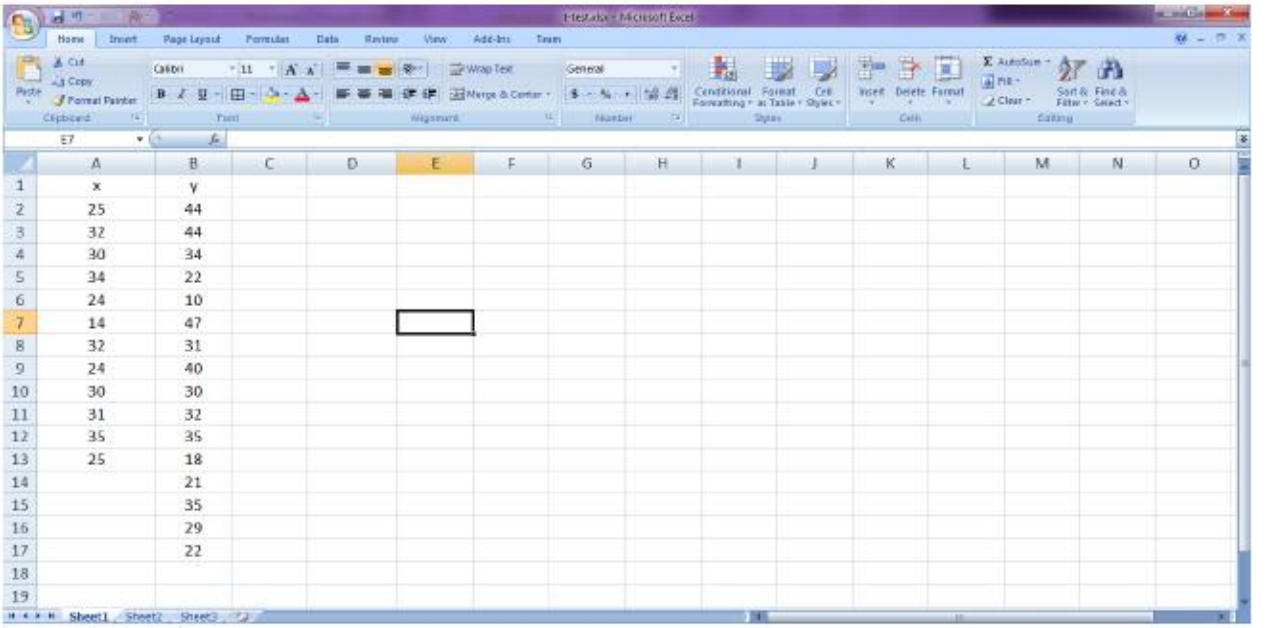
Solution:

Null hypothesis: Ho: $\mu_1 = \mu_2$ i.e. there is no significant difference between the mean increase in weight due to diets A and B.

Alternative hypothesis: $\mu_1 \neq \mu_2$ (two tailed)

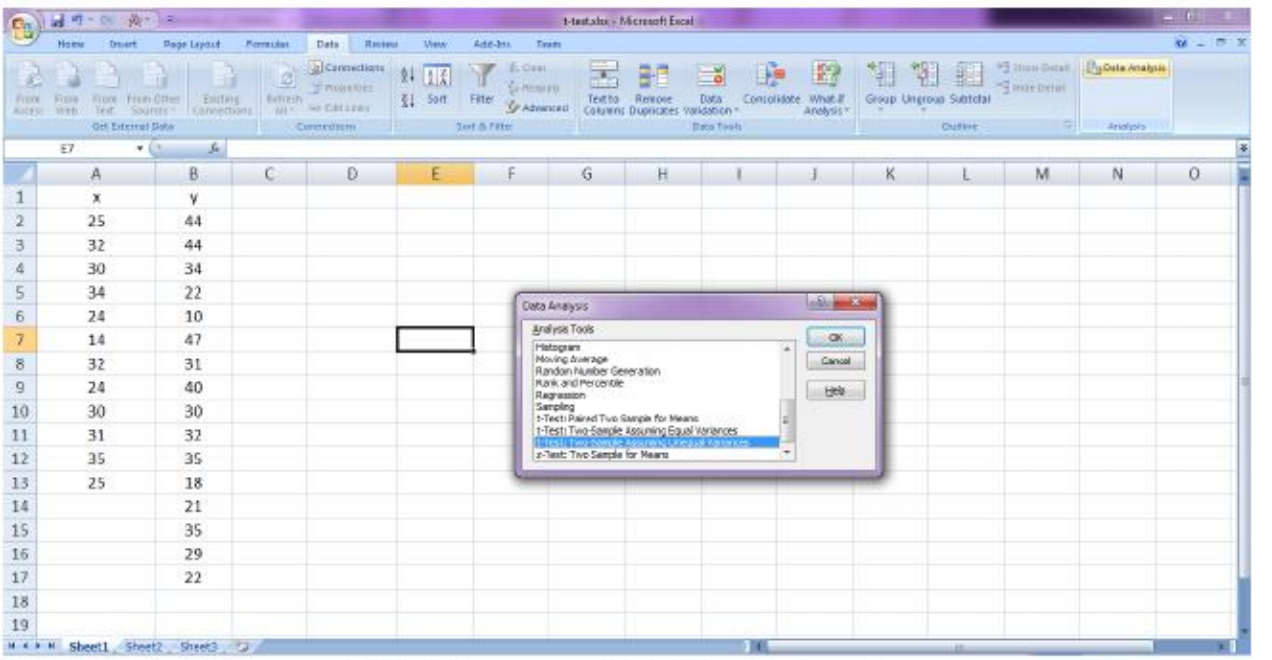
To apply t -test for the above, we proceed in the following steps:

Step1: First of all, enter the data i.e. the value of sample:



	A	B
1	x	y
2	25	44
3	32	44
4	30	34
5	34	22
6	24	10
7	14	47
8	32	31
9	24	40
10	30	30
11	31	32
12	35	35
13	25	18
14		21
15		35
16		29
17		22

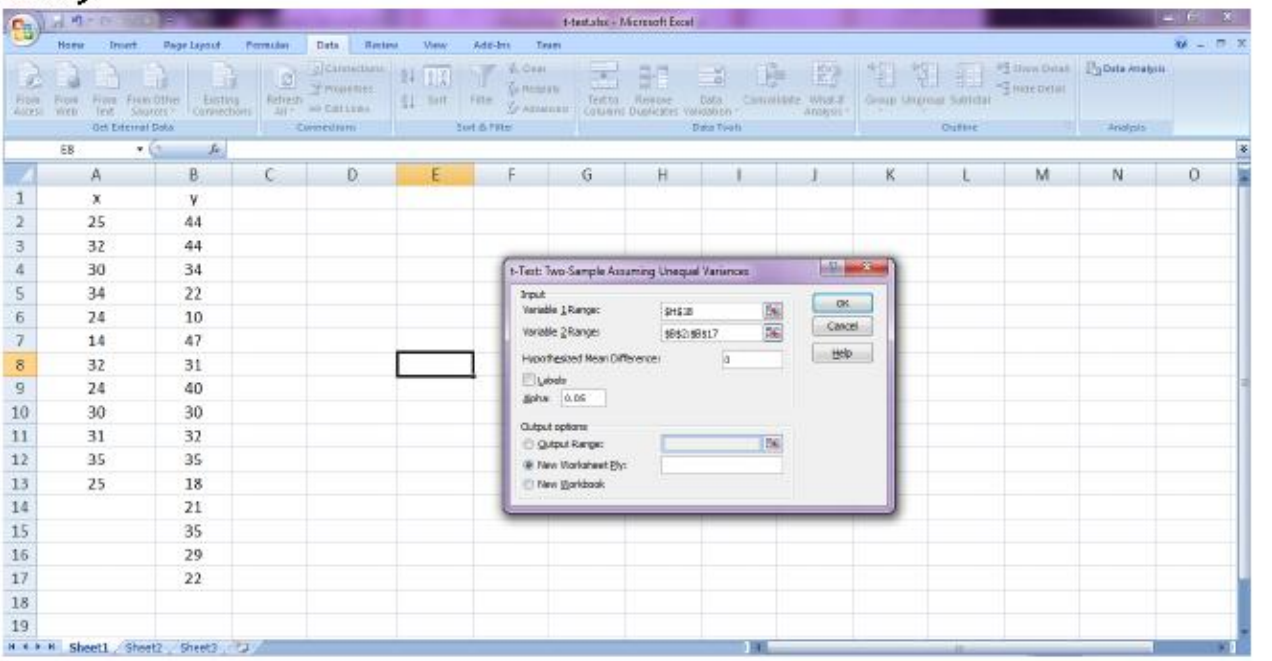
Step2: Click on Data →Data Analysis → select t test: Two Sample Assuming Unequal Variance →click ok



	A	B
1	x	y
2	25	44
3	32	44
4	30	34
5	34	22
6	24	10
7	14	47
8	32	31
9	24	40
10	30	30
11	31	32
12	35	35
13	25	18
14		21
15		35
16		29
17		22

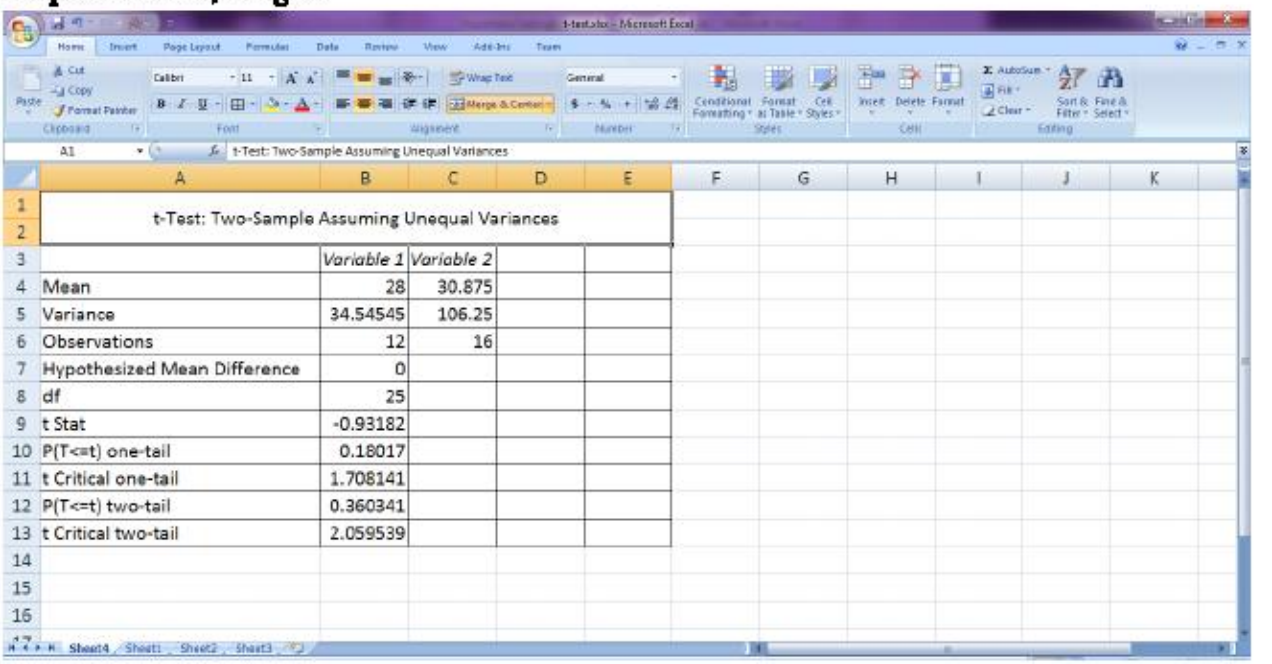
Step3: Click ok, Input variable 1 Range, variable 2 Range, Hypothesized mean difference (0 in this example under null hypothesis), Alpha (the level of significance), output option (provide output range if one wants to be output on the same

sheet).



	A	B
1	x	y
2	25	44
3	32	44
4	30	34
5	34	22
6	24	10
7	14	47
8	32	31
9	24	40
10	30	30
11	31	32
12	35	35
13	25	18
14		21
15		35
16		29
17		22

Step 4: click ok, we get



	A	B	C	D	E
1	t-Test: Two-Sample Assuming Unequal Variances				
2					
3		*Variable 1*	*Variable 2*		
4	Mean	28	30.875		
5	Variance	34.54545	106.25		
6	Observations	12	16		
7	Hypothesized Mean Difference	0			
8	df	25			
9	t Stat	-0.93182			
10	P(T<=t) one-tail	0.18017			
11	t Critical one-tail	1.708141			
12	P(T<=t) two-tail	0.360341			
13	t Critical two-tail	2.059539			

Description of output: Since, calculated value of mode t (0.93182) in 9th row (for one tailed test) is less than 1.708141 (i.e. critical value for one tailed test at 5 percent level of significance). Same evidence is also found from p-value=0.1807>0.05. Hence, null hypothesis is accepted for one tailed test at 5% level of significance. Now, for two tailed test, since calculated value of mode t (0.93182) is less than 2.059539, hence we accept null hypothesis for two tailed test i.e. the two samples can be regarded from same population. Same evidence is also observed from p-value as p=0.360341>0.05.

Test 4: *t*-test for two populations mean (dependent sample) or Paired t-test:

Object: To investigate the significance of the difference between two population (dependent) means, μ_1 and μ_2. No assumption is made about the population variances.

Method: The differences $d_i\ (x_i - y_i)$ are formed for each pair of observations. If there are n such pairs of observations, we can calculate the variance of the differences by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2 \text{ where } \bar{d} = \bar{x}_1 - \bar{x}_2$$

Let the means of the samples from the two populations be denoted by $\bar{x}_1$ and $\bar{x}_2$ then the test statistic becomes

$$t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}$$

which follows Student's t-distribution with $n - 1$ degrees of freedom. The test may be either one-tailed or two-tailed.

Limitations:

1. The observations for the two samples must be obtained in pairs. Apart from population differences, the observations in each pair should be carried out under identical, or almost identical, conditions.
2. The test is accurate if the populations are normally distributed.

Example: Below are given the gain in weights (in lbs.) of pigs fed on two diets x and y.

Gain in weight

Diet x : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet y : 44, 44,34,22,10,47,31,40,30,32,35,18

Test if the two diets differ significantly if same set of 12 pigs were used in both the foods

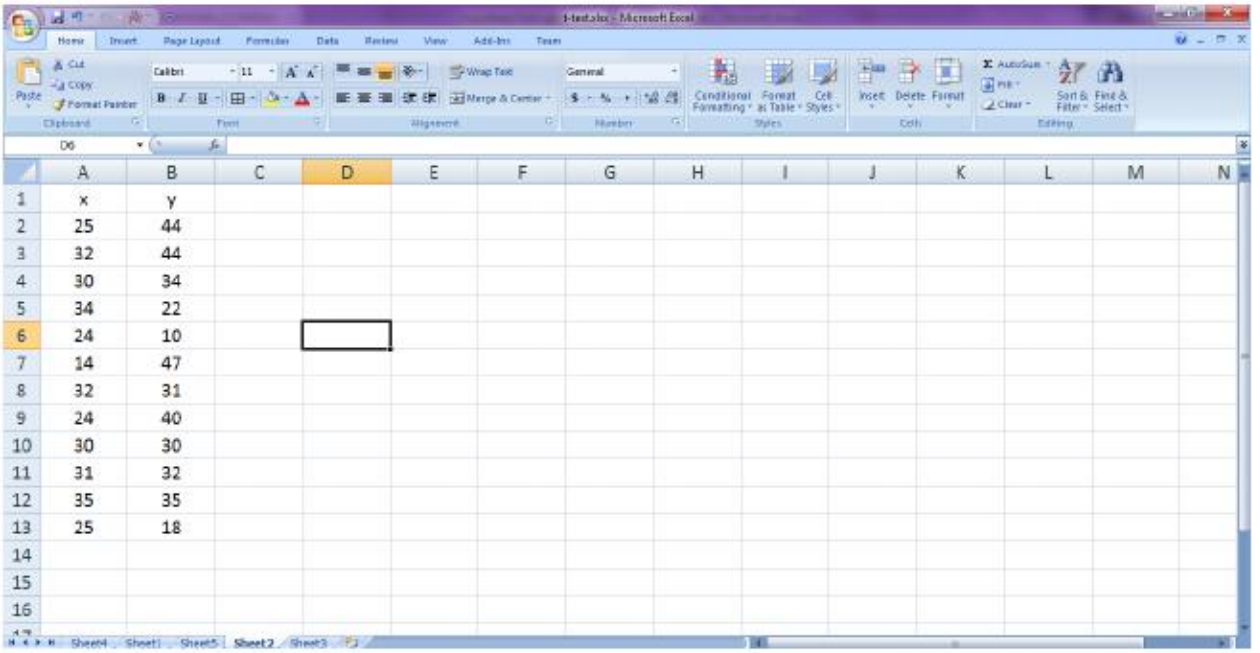
Solution:

Null hypothesis: Ho: $\mu_1 = \mu_2$ i.e. there is no significant difference between the mean increase in weight due to diets A and B.

Alternative hypothesis: $\mu_1 \neq \mu_2$ (two tailed)

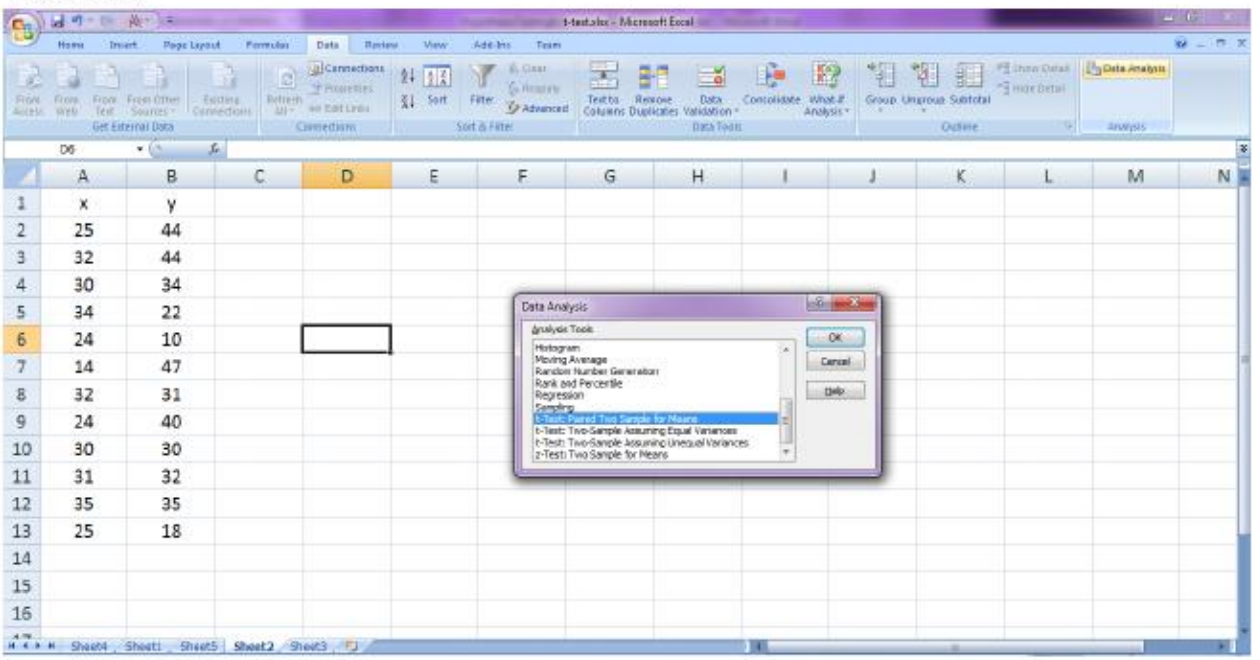
To apply t -test for the above, we proceed in the following steps:

Step1: First of all, enter the data i.e. the value of sample:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	x	y												
2	25	44												
3	32	44												
4	30	34												
5	34	22												
6	24	10												
7	14	47												
8	32	31												
9	24	40												
10	30	30												
11	31	32												
12	35	35												
13	25	18												
14														
15														
16														

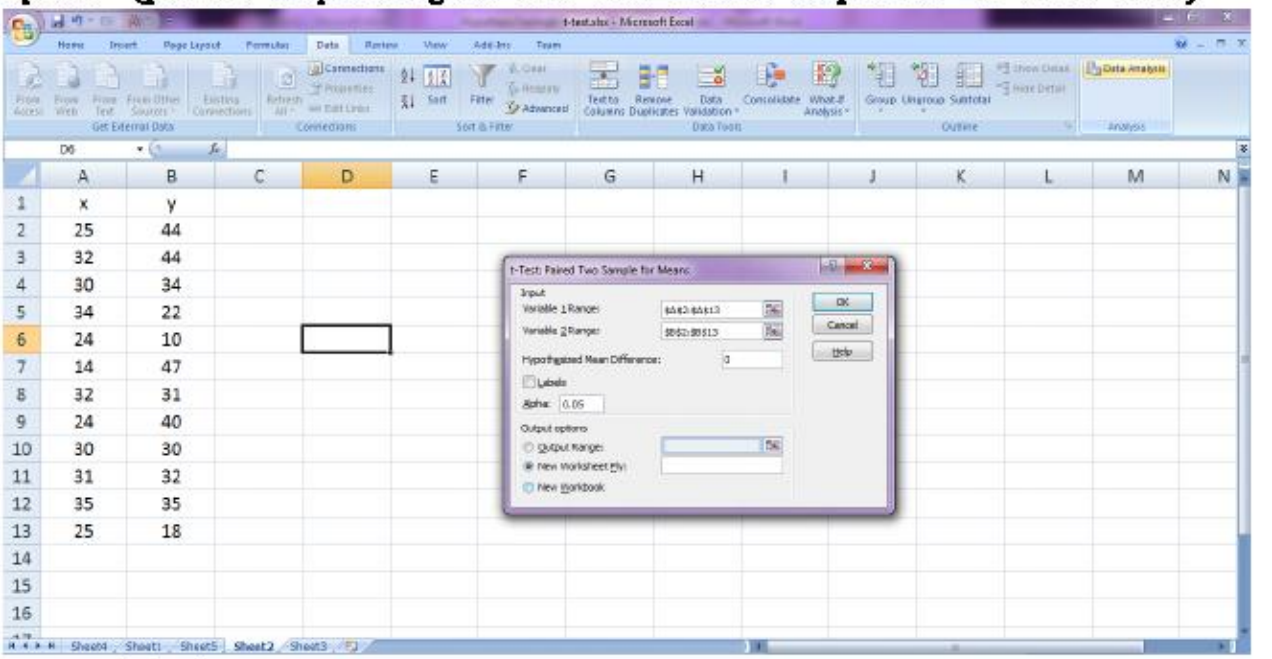
Step2: Click on Data →Data Analysis → select t test: Paired Two Sample for Means →click ok



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	x	y												
2	25	44												
3	32	44												
4	30	34												
5	34	22												
6	24	10												
7	14	47												
8	32	31												
9	24	40												
10	30	30												
11	31	32												
12	35	35												
13	25	18												
14														
15														
16														

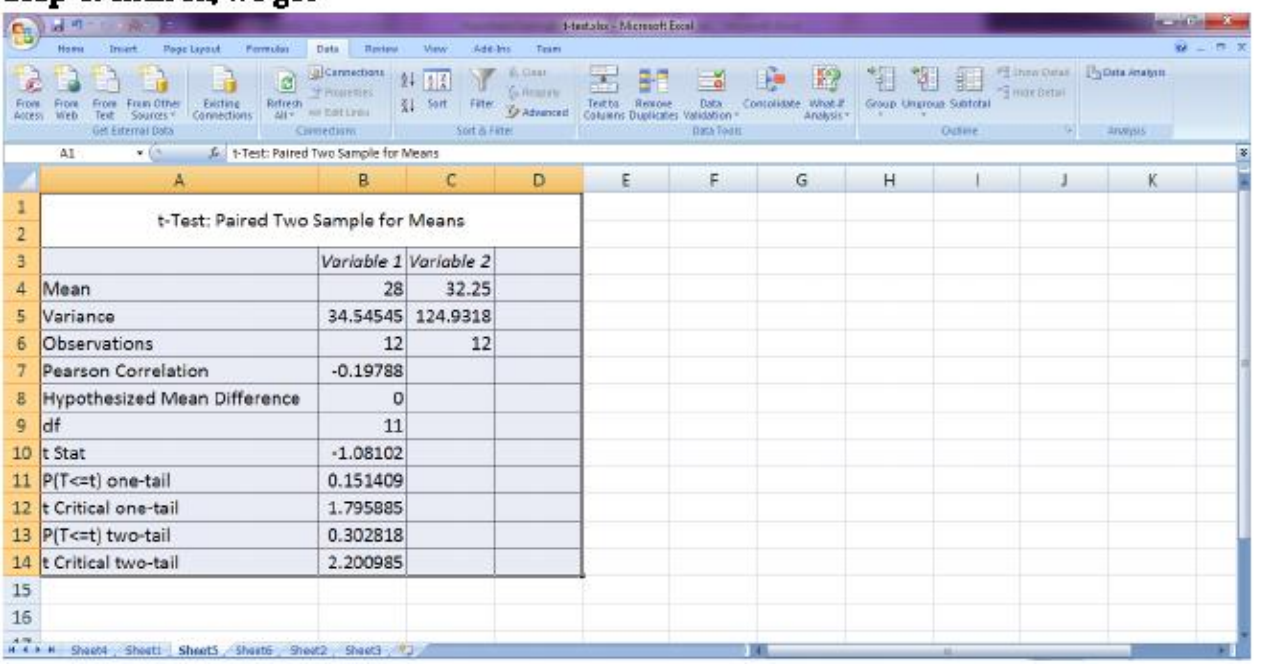
Step3: Click ok, Input variable 1 Range, variable 2 Range, Hypothesized mean difference (0 in this example under null hypothesis), Alpha (the level of significance), output

option (provide output range if one wants to be output on the same sheet).



	A	B
1	x	y
2	25	44
3	32	44
4	30	34
5	34	22
6	24	10
7	14	47
8	32	31
9	24	40
10	30	30
11	31	32
12	35	35
13	25	18

Step 4: click ok, we get



t-Test: Paired Two Sample for Means		
	Variable 1	*Variable 2*
Mean	28	32.25
Variance	34.54545	124.9318
Observations	12	12
Pearson Correlation	-0.19788	
Hypothesized Mean Difference	0	
df	11	
t Stat	-1.08102	
P(T<=t) one-tail	0.151409	
t Critical one-tail	1.795885	
P(T<=t) two-tail	0.302818	
t Critical two-tail	2.200985	

Description of output: Since, calculated value of mode t (1.08102) in 10th row (for one tailed test) is less than 1.795885 (i.e. critical value for one tailed test at 5 percent level of significance). Same evidence is also found from p-value=0.151409>0.05. Hence, null hypothesis is accepted for one tailed test at 5% level of significance. Now, for two tailed test, since calculated value of mode t (1.08102) is less than 2.200985, hence we accept null hypothesis for two tailed test i.e. the two samples can be regarded from same population. Same evidence is also observed from p-value as p=0.302818>0.05.

Test 5: *F*-test for two population variances (variance ratio test)

Object: To investigate the significance of the difference between two population variances.

Method: Given samples of size n_1 with values $(x_1, x_2, \ldots\ldots x_{n_1})$ and size n_2 with values $(y_1, y_2, \ldots\ldots y_{n_2})$ from the two populations, the values of

$$\bar{x}_1 = \frac{1}{n_1}\sum x_i \quad \bar{y}_2 = \frac{1}{n_2}\sum y_i \quad s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2 \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2$$

can be calculated. Under the null hypothesis that the variances of the two populations are equal the test statistic $F = s_1^2 / s_2^2$ follows the F-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom. The test may be either one-tailed or two-tailed.

Limitations: The two populations should both follow normal distributions. (It is not necessary that they should have the same means.)

Example: Two random samples are given as follows:

A: 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

B: 44, 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18

Test whether the samples come from the same normal population at 5% level of significance.

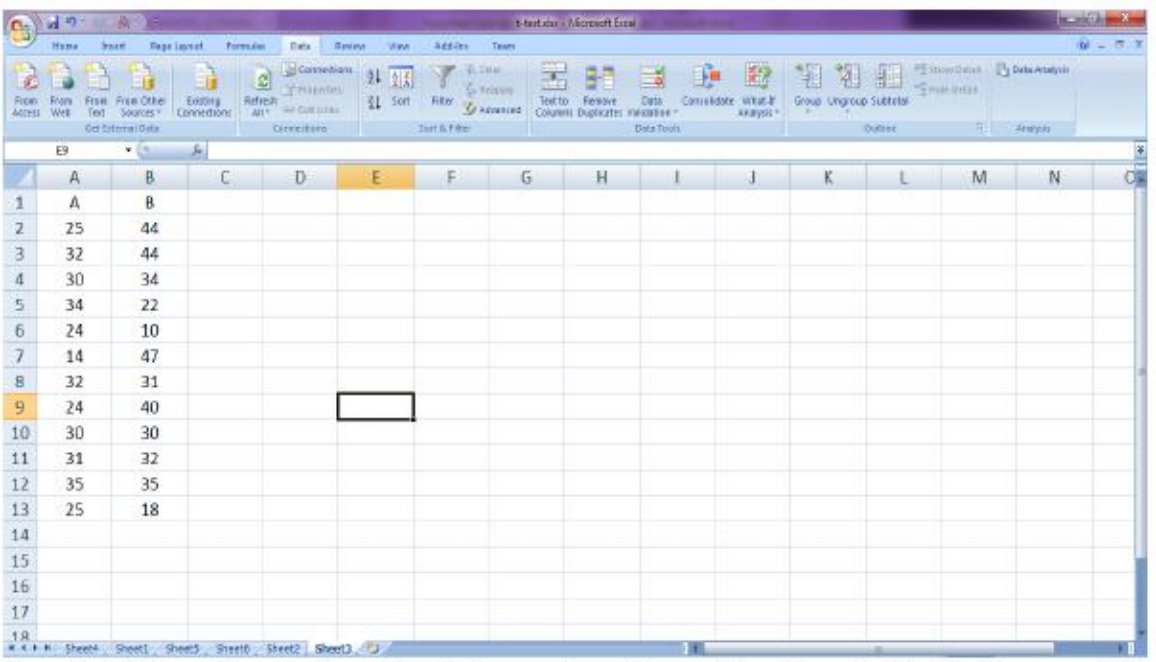
Solution:

Null hypothesis: Ho: $\sigma_1^2 = \sigma_2^2$

Alternative hypothesis: $\sigma_1^2 \neq \sigma_2^2$ (two tailed)

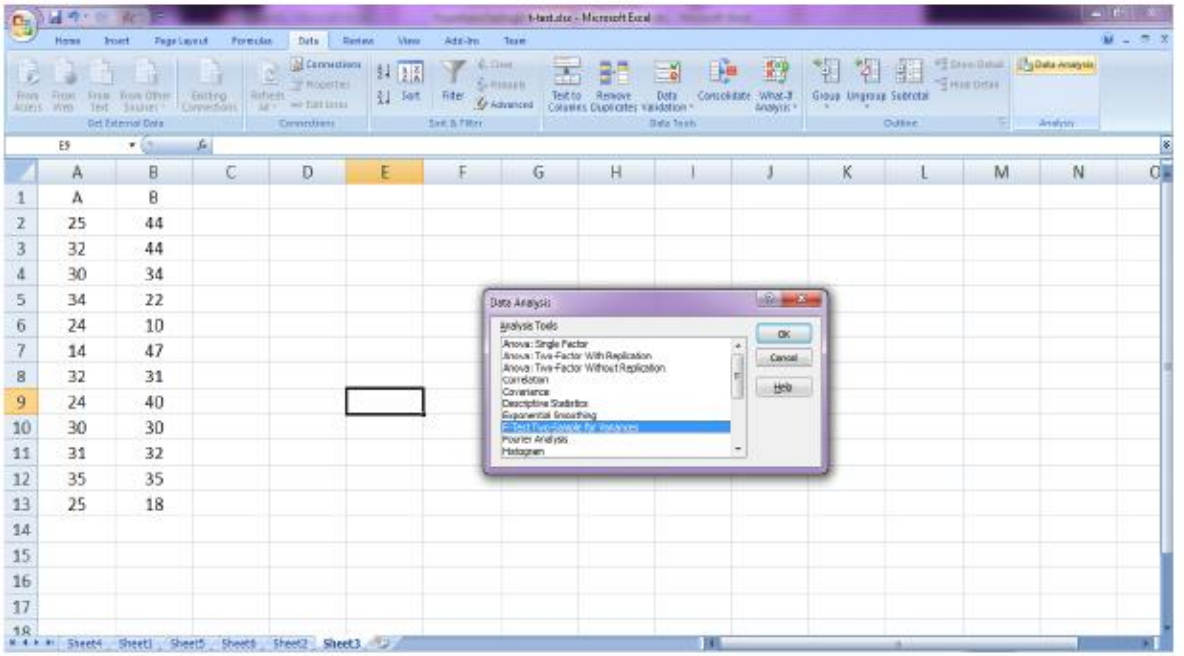
To apply F-test for the above, we proceed in the following steps:

Step1: First of all, enter the data i.e. the values of sample:

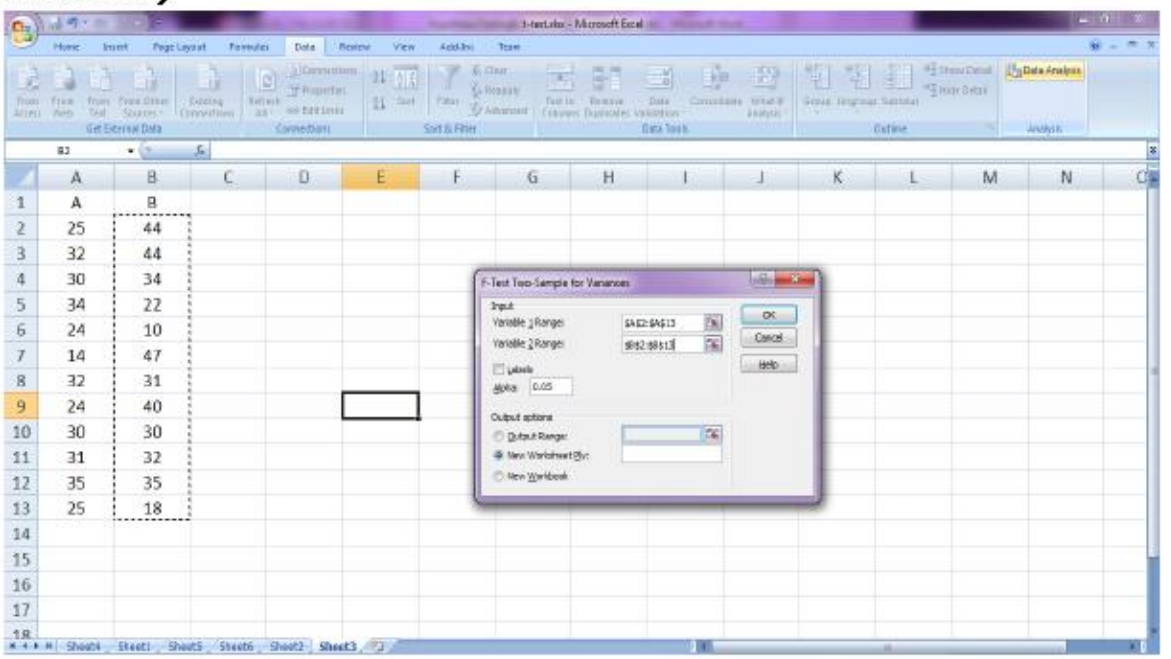


	A	B
1	A	B
2	25	44
3	32	44
4	30	34
5	34	22
6	24	10
7	14	47
8	32	31
9	24	40
10	30	30
11	31	32
12	35	35
13	25	18

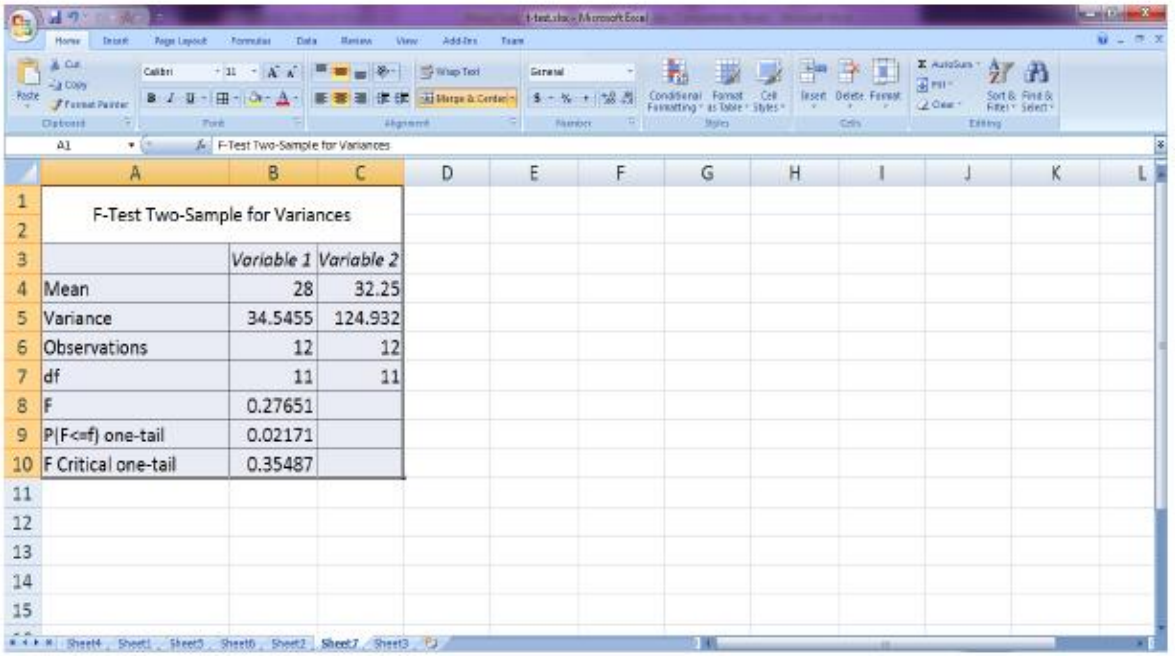
Step2: Click on Data →Data Analysis → select F test: Two Sample for Variance →click ok



Step3: Click ok, Input variable 1 Range, variable 2 Range, Alpha (the level of significance), output option (provide output range if one wants to be output on the same sheet).



Step 4: click ok, we get



F-Test Two-Sample for Variances		
	Variable 1	*Variable 2*
Mean	28	32.25
Variance	34.5455	124.932
Observations	12	12
df	11	11
F	0.27651	
P(F<=f) one-tail	0.02171	
F Critical one-tail	0.35487	

Description of output: Since, observed p-value p=0.2171<0.05, so the null hypothesis may be accepted at 5% level of significance and hence samples cannot be regarded from same population.

SPSS is a comprehensive and flexible statistical analysis and data management tool. It can be used to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and conduct complex statistical analyses. SPSS is used in every field, including telecommunications, banking, finance, insurance, healthcare, manufacturing, retail, consumer packaged goods, higher education, government, and research.

The objective of this presentation is to equip a researcher with the power and feature of SPSS. In view of this, it is necessary to understand the difference between:

- Quantitative and qualitative data
- Primary and Secondary data
- Data and Information

Quantitative data are anything that can be expressed as a number, or quantified. Examples of quantitative data are scores on achievement tests, number of hours of study, or weight of a subject. These data may be represented by interval or ratio scales and lends themselves to most statistical manipulation.

Qualitative data cannot be expressed as a number. It is basically categorical measurement expressed by means of a natural language description. This type of data cannot be measured but it can be observed. For example, gender, socio economic status, religious preference are usually considered to be qualitative data.

Both types of data are valid types of measurement. Only quantitative data can be analyzed statistically.

Primary Data means original data that has been collected specially for the purpose. It means someone collected the data from the original source first hand. Data collected this way is called primary data.

Secondary data is data that has been collected for another purpose. When we use Statistical Method with Primary Data from another purpose for our purpose we refer to it as Secondary Data. It means that one's Primary Data is another's Secondary Data. Secondary data is data that is being reused. Usually in a different context.

Data is raw, unorganized facts that need to be processed. Data can be something simple and seemingly random and useless until it is organized. Each student's test score is one piece of data. When data is processed, organized, structured or presented in a given context so as to make it useful, it is called **information**. The average score of a class or of the entire school is information that can be derived from the given data.

Scope of SPSS

SPSS is suitable for data that comes under the category of Primary and Quantitative data. Qualitative data may also be used, in case; it is converted into quantitative data.

Three basic steps are necessary to explore SPSS:

- Data preparation
- Data Management
- Data Analysis

* Programmer, G B Pant Social Science Institute, Allahabad.

Data Preparation

Considering the factor of Primary data, it is necessary to develop instrument for the survey. The survey instrument consists of four types of questions:

Open ended questions that give rich information but are difficult to quantify. For example, asking the respondent for suggestions in improving the quality of life or feedback of participants regarding this workshop.

Closed ended dichotomous questions offer only two responses from the respondent. For example, asking the respondent for whether he/she is married or whether he/she has accessed his voting rights during election.

Single/Multiple choice questions offer respondent to pick one or more options given to him for his/her answers.

Scaled responses are used for respondents when a range of discrete values are given to him/her for response. For example, asking the respondent about his/her view whether *Crime is evil of society*. The options are strongly disagree, Disagree, Neutral, Agree, Strongly agree.

It is also suggested that while developing instrument following measures should be taken in account:

- Plan a user friendly format
- Gather demographic data
- Guarantee anonymity
- Ensure ease of tabulation
- Ask well-phrased questions that can answered
- Develop for completeness
- Pilot test the instrument

Before moving to SPSS, it is necessary to collect all questionnaires for coding. When coding is complete, one can start working with SPSS.

First step is to create variables (each question is converted into variable) under Variable View of SPSS. Care should be taken while creating variable as one has to fill each attribute of the variable. The basic attributes are:

Name:	A suitable variable name for the question.
Type:	Whether the response is numeric/string/date etc.
Width:	No. of digits required.
Label:	Description of variable.
Value & Value Label:	Code & detail about the options available for responses
Missing:	How to handle the responses which comes under No response or Not Applicable cases.
Column:	Width of the column to display in Data View
Align:	Whether data is to be displayed left/right/centre aligned
Measure:	Nominal/Ordinal/Scale

It is strongly suggested that you should know whether data is nominal or ordinal or scale, because it will help you in selecting appropriate statistical tool for analysis.

Nominal data referred to qualitative data for which numeric codes are given only for the sake of identification.

Ordinal data referred to qualitative data for which numeric codes are given in hierarchal order.

Scale data refers to quantitative data for which numeric values are directly recorded from the responses of the respondent.

Data Management

First step in data management is **cleaning of data** which refers to cross check data from original dataset to ensure that it is error free. Because wrong data takes you to incorrect results.

Second step is to classify the data as per our requirement e.g., frequency, mean, minimum, maximum, standard deviation etc. of variable. After going through these, you may need to generate new variables based on initial variables. For example, you may be interested in creating a group, based upon the values of responses or club some responses into one response.

Third step is to move gradually for more statistics that are required for your assumptions.

REFERENCES

Field, A. (2006); Discovering Statistics Using SPSS; New Delhi, Sage Publication.

Introduction

SPSS is a widely used software package for statistical analysis in social science. The original SPSS manual (Nie *et al.*, 1970) has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis. Originally it is an acronym of *Statistical Package for the Social Science* but now it stands for *Statistical Product and Service Solutions*. The current versions (2015) are officially named IBM SPSS Statistics. Long produced by SPSS Inc., it was acquired by IBM in 2009. During 2009 and 2010 it was called *PASW* (*Predictive Analytics SoftWare*) *Statistics*. It is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with rather simple instructions. This package of programs is available for both personal as well as mainframe computers. SPSS package consists of a set of software tools for data entry, data management, statistical analysis and presentation. SPSS integrates complex data and file management, statistical analysis and reporting functions. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and complex statistical analyses.

Some versions of SPSS released in recent years are:

- SPSS Statistics 17.0.1 - December 2008
- PASW Statistics 17.0.3 - September 2009
- PASW Statistics 18.0, 18.0.1, 18.0.2, 18.0.3
- IBM SPSS Statistics 19.0 - August 2010
- IBM SPSS Statistics 19.0.1, 20.0, 20.0.1, 21.0

Companion products in the same family are used for survey authoring and deployment (IBM SPSS Data Collection), data mining (IBM SPSS Modeler), text analytics, and collaboration and deployment (batch and automated scoring services). Purpose of this chapter is to introduce the basic features of the SPSS and also to provide some basic statistical analysis using this software.

Key Features of SPSS

Some of the key feature of SPSS are:

- It is easy to learn and use with its pull-down menu features
- It includes a full range of data management system
- It provides in-depth statistical analysis capabilities
- It offers comprehensive range of plotting, reporting and presentation features.

In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary stored in the data file) are features of the base software. There are varieties of statistics included in the base software. Some of the important statistics are:

Descriptive statistics: Cross tabulation, Frequencies, Descriptive, Explore, Descriptive Ratio Statistics etc.

Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), Nonparametric tests etc.

Prediction for numerical outcomes: Linear regression, Multiple Regression

Prediction for identifying groups: Factor analysis, Cluster analysis (two-step, K-means, hierarchical), Discriminant analysis etc.

Windows of SPSS

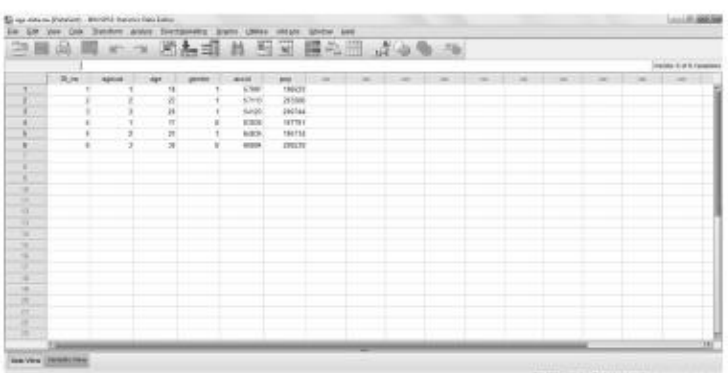
SPSS makes statistical analysis accessible for the naive user and convenient for the experienced user. There are a number of different types of windows in SPSS. The data editor offers a simple and efficient spreadsheet like facility for entering data and browsing the working data file. To invoke SPSS in the windows environment, select the appropriate SPSS icon.

Data Editor: This graphical user interface displays the contents of the data file. One can create new data files or modify existing ones. The Data Editor window opens automatically when an SPSS session is started. One can have only one data file open at a time. This editor has two views which can be toggled by clicking on one of the two tabs in the bottom left of the SPSS window.

✓ **Data view**: Displays the actual data values or defined value labels. The 'Data View' shows a spreadsheet view of the cases (rows) and variables (columns). Unlike spreadsheets, the data cells can only contain numbers or text, and formulas cannot be stored in these cells. One can modify data values in the Data view in many ways like change data values; cut, copy and paste data values; add and delete cases;

✓ **Variable view**: Displays variable definition information contained or metadata dictionary where each row represents a variable and shows the variable name, variable label, value label(s), print width, measurement type, and a variety of other characteristics. One can modify variable properties in the Variable view for example, add and delete variables, change the order of variables etc.

Cells in both views can be manually edited, defining the file structure and allowing data entry without using command syntax. This may be sufficient for small datasets. Larger datasets such as statistical surveys are more often created in data entry software, or entered during computer-assisted personal interviewing, by scanning and using optical character recognition and optical mark recognition software, or by direct capture from online questionnaires. These datasets are then read into SPSS. Extension of the saved data file will be ".sav".



Viewer: All results, tables, and charts performed by different statistical analysis are displayed in the Viewer window. Extension of the saved output file is ".spv". One can use the Viewer to browse results, show or hide selected tables and charts, change the display order of results by moving selected items or move items between the Viewer and other applications. The output presented in Viewer can be edited and saved for later use. A Viewer window opens automatically the first time a procedure is run that generates output. The Viewer is divided into two panes:

- ✓ The left pane contains an outline view of the contents. One can click an item in the outline to go directly to the corresponding table or chart.
- ✓ The right pane contains statistical tables, charts, and text output.

Syntax Editor: The pull-down menu interface generates command syntax. This can be displayed in the output. These command syntax can also be pasted into a syntax file in a syntax window using the "paste" button present in each menu. One can then edit the command syntax totalize special features of SPSS not available through dialog boxes. These commands can be saved in a file for use in subsequent SPSS sessions. Extension of the saved syntax file will be ".sps". Command syntax programming has the benefits of reproducibility, simplifying repetitive tasks, and handling complex data manipulations and analyses. Additionally, some complex applications can only be programmed in syntax that is not accessible through the menu structure.

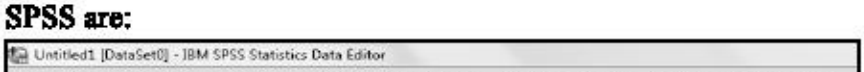
Pivot Table Editor: The results from most statistical procedures are displayed in pivot tables. These pivot tables outputs can be modified in many ways with pivot table editor. One can edit text, swap data in rows and columns, create multidimensional tables, and selectively hide and show results. Changing the layout of the table does not affect the results. Instead, it's a way to display information in a different or more desirable manner.

Text Output Editor: Text output not displayed in pivot tables can be modified with the Text Output Editor. One can edit the output and change font characteristics (type, style, colour, size).

Chart Editor: High-resolution charts and plots can be modified in chart windows. One can change the colours, select different type of fonts and sizes, switch the horizontal and vertical axes, rotate 3-D scatter plots, and even change the chart type.

Script Window: It provides the opportunity to write full-blown programs, in a BASIC-like language. It is a text editor for syntax composition. Extension of the saved script file will be ".sbs"

Many features of SPSS Statistics are accessible via pull-down menus or can be programmed with a proprietary 4GL command syntax language. Many of the tasks that are to be performed with SPSS start with menu selections. Each window has its own menu bar with menu selections appropriate for that window type. The various menu options available in SPSS are:



Most menu selections open dialog boxes. One can use dialog boxes to select variables and options for analysis. Since most procedures provide a great deal of flexibility, not all of the possible choices can be contained in a single dialog box. The main dialog box usually contains the minimum information required to run a procedure. Additional specifications are

made in sub-dialog boxes. All these above mentioned options have further sub-options. To see what applications there are, we simply move the cursor to a particular option and press, when a drop-down menu will appear. To cancel a drop-down menu, place the cursor anywhere outside the option and press the left button.

The three dots after an option term (...) on a drop-down menu, such as **Define Variable...**option in Data option, signify that a dialog box will appear when this option is chosen. To cancel a dialog box, select the **Cancel** button in the dialog box. A right-facing arrow head after an option term indicates that a further submenu will appear to the right of the drop-down menu. An option with neither of these signs means that there is no further drop down menus to select. There are five standard command pushbuttons in most dialog boxes.

OK: It runs the procedure. After the variables and additional specifications are selected, click OK to run the procedure.

Paste: It generates command syntax from the dialog box selections and pastes the syntax into a syntax window.

Reset: It deselects any variables in the selected variable list and resets all specifications in the dialog box.

Cancel: It cancels any changes in the dialog box settings since the last time it was opened and closes the dialog box.

Help: It contains information about the current dialog box.

ENTERING AND EDITING DATA

The easiest way of entering data in SPSS is to type it directly into the matrix of columns and numbered rows in the **Data Editor** window. The columns represent variables and the rows represent cases. The variables can be defined in the variable view. Variable name must be no longer than eight characters and the name must begin with a letter.

Saving data

To be able to retrieve a file, the file must be saved with a proper name. The default extension name for saving files is **sav**. To save this file on a floppy disk, we carry out the following sequence:

→File →Save As... [opens **Save Data As** dialog box]→box under **File Name:** delete the asterisk and type file name **→OK**

The output file can also be printed and saved. The extension name for output file is **.spo**.

Retrieving a Saved File

To retrieve this file at a later stage when it is no longer the current file, use the following procedure:

File → Open → Data ... [opens the **Open Data File** dialog box] →choose drive from options listed →type name under **File Name:** →file name → **OK**

BASIC STEPS IN DATA ANALYSIS

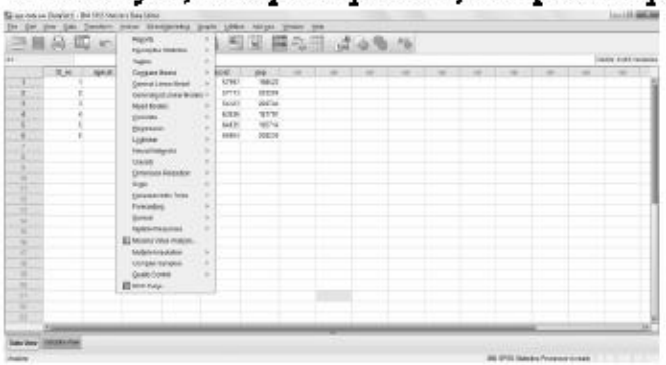
- **Get your data into SPSS.** You can open a previously saved SPSS data file, read a spreadsheet, database, or text data file, or enter your data directly in the Data Editor.
- **Select a procedure.** Select a procedure from the menus to calculate statistics or to create a chart.
- **Select the variables for the analysis.** The variables in the data file are displayed in a dialog box for the procedure.

- **Run the procedure.** Results are displayed in the Viewer.

STATISTICAL PROCEDURES

After entering the data set in **Data Editor** or reading an ASCII data file, we are now ready to analyze it. The **Analyze** option has the following sub options:

Reports, Descriptive Statistics, Tables, Compare means, General Linear model, Mixed Models, Correlate, Regression, Log linear, Neural Networks, Classify, Dimension Reduction, Scale, Non parametric tests, Forecasting, Time Series, Survival, Multiple response, Missing value analysis, Multiple imputation, Complex samples, Quality control, ROC curve.



Reports

This submenu provides techniques for reporting the results. The various sub-sub menus under this are as follows:

Codebook reports the dictionary information such as variable names, variable labels, value labels, missing values and summary statistics for all or specified variables and multiple response sets in the active dataset. For nominal and ordinal variables and multiple response sets, summary statistics include counts and percents. For scale variables, summary statistics include mean, standard deviation, and quartiles.

OLAP (Online Analytical Processing) **Cubes** procedure calculates totals, means, and other univariate statistics for continuous summary variables within categories of one or more categorical grouping variables. A separate layer in the table is created for each category of each grouping variable.

Case Summaries calculates subgroup statistics for variables within categories of one or more grouping variables. All levels of the grouping variable are cross tabulated. One can choose the order in which the statistics are displayed. Summary statistics for each variable across all categories are also displayed. With large datasets, one can choose to list only the first n cases.

Report Summaries in Rows produces reports in which different summary statistics are laid out in rows. Case listings are also available from this command, with or without summary statistics.

Report Summaries in Columns produces reports in which different summary statistics are laid out in separate columns.

Descriptive Statistics

This submenu provides techniques for summarizing data with statistics, charts, and reports. The various sub-sub menus under this are as follows:

Frequencies provide information about the relative frequency of the occurrence of each category of a variable. This can be used it to obtain summary statistics that describe the

typical value and the spread of the observations. To compute summary statistics for each of several groups of cases, Means procedure or the Explore procedure can be used.

Descriptive is used to calculate statistics that summarize the values of a variable like the measures of central tendency, measures of dispersion, skewness, kurtosis etc..

Explore produces and displays summary statistics for all cases or separately for groups of cases. Box plots, stem-and leaf plots, histograms, tests of normality, robust estimates of location, frequency tables and other descriptive statistics and plots can also be obtained.

Crosstabs is used to count the number of cases that have different combinations of values of two or more variables, and to calculate summary statistics and tests. The variables you use to form the categories within which the counts are obtained should have a limited number of distinct values.

P-P Plots provides the cumulative proportions of a variable's distribution against the cumulative proportions of the normal distribution.

Q-Q plots provide the quantiles of a variable's distribution against the quantiles of the normal distribution.

Tables

Custom Tables submenu provides attractive, flexible displays of frequency counts, percentages and other statistics.

Compare Means

This submenu provides techniques for testing differences among two or more means for both independent and related samples.

Means computes summary statistics for a variable when the cases are subdivided into groups based on their values for other variables.

One-Sample t-Test procedure tests whether the mean of a single variable differs from a specified constant. For each test variable: mean, standard deviation, and standard error of the mean.

Independent Sample t-Test is used if two unrelated samples come from populations with the same mean. The observations should be from two unrelated groups, and for testing, the mean must be an appropriate summary measure for the variable to be compared in the two groups. For more than two independent groups, the *One-way ANOVA* option could be used.

Paired Sample t-Test is used to compare the means of the same subjects in two conditions or at two points in time i.e. to compare subjects who had been matched to be similar in certain respects and then to test if two related samples come from populations with the same mean. The related, or paired, samples often result from an experiment in which the same person is observed before and after an intervention. If the distribution of the differences of the values between the members of a pair is markedly non-normal you should consider one of the nonparametric tests.

One-Way ANOVA is used to test that several independent groups come from populations with the same mean. To see which groups are significantly different from each other, multiple comparison procedures can be used through *Post Hoc Multiple Comparison option* which consist of the options like *Least-significant difference, Duncan's multiple range test, Scheffe* etc. The contrast analysis can also be performed in order to compare the different groups or

treatments by using the *Contrast* option. The data obtained using completely randomized design can be analyzed through this option.

General Linear Model

This submenu provides techniques for testing univariate and multivariate Analysis-of-Variance models, including repeated measures.

Univariate sub-option could be used to analyze the experimental designs like Completely randomized design, Randomized block design, Latin square design, Designs for factorial experiments etc. The covariance analysis can also be performed and alternate methods for partitioning sums of squares can be selected. If only some of the interactions of a particular order are to be included, the *Custom* procedure should be used. If there is only one factor then One-Way ANOVA procedure should be used.

Multivariate analyses analysis-of-variance and analysis-of-covariance designs when you have two or more correlated dependent variables. Multivariate analysis of variance is used to test hypotheses about the relationship between a set of interrelated dependent variables and one or more factor or grouping variables. For example, you can test whether verbal and mathematical test scores are related to instructional method used sex of the subject and the interaction of method and sex. This procedure should be used only if there are several dependent variables which are related to each other. For a single dependent variable or unrelated dependent variables, the Univariate ANOVA procedures can be adopted. If the same dependent variable is measured on several occasions for each subject, the Repeated Measures procedure is to be used.

Repeated Measures is used to test hypotheses about the means of a dependent variable when the same dependent variable is measured on more than one occasion for each subject. Subjects can also be classified into mutually exclusive groups, such as males or females, or type of job held. Then you can test hypotheses about the effects of the between-subject variables and the within-subject variables, as well as their interactions.

Correlate

This submenu provides measures of association for two or more variables measured at the interval level.

Bivariate Calculates Matrices of Pearson product-moment correlations, and of Kendall and Spearman nonparametric correlations, with significance levels and optional univariate statistics. The correlation coefficient is used to quantify the strength of the linear relationship between two variables. The *Pearson correlation coefficient* should be used only for data measured at the interval or ratio level. Spearman and Kendall correlation coefficients are nonparametric measures which are particularly useful when the data contain outliers or when the distribution of the variables is markedly non-normal. Both the Spearman and Kendall coefficients are based on assigning ranks to the variables.

Partial calculates *partial correlation coefficients* that describe the relationship between two variables, while adjusting for the effects of one or more additional variables. If the values of a dependent variable from a set of independent variables are to be predicted then the Linear Regression procedure may be used. If there are no control variables then the Bivariate Correlations procedure can be adopted. Nominal variables should not be used in the partial correlation procedure.

Distances calculate statistics measuring either similarities or dissimilarities (distances), either between pairs of variables or between pairs of cases. These similarity or distance measures can then be used with other procedures, such as factor analysis, cluster analysis, or multidimensional scaling, to help analyze complex datasets. Dissimilarity (distance) measures for interval data are Euclidean distance, squared Euclidean distance, Chebychev, block, Minkowski, or customized; for count data, chi-square or phi-square; for binary data, Euclidean distance, squared Euclidean distance, size difference, pattern difference, variance, shape, or Lance and Williams. Similarity measures for interval data are Pearson correlation or cosine; for binary data, Russel and Rao, simple matching, Jaccard, etc.

Regression

This submenu provides a variety of regression techniques, including linear, logistic, nonlinear, weighted, and two-stage least-squares regression.

Linear is used to examine the relationship between a dependent variable and a set of independent variables. If the dependent variable is dichotomous, then the logistic regression procedure should be used. If the dependent variable is censored, such as survival time after surgery, use the Life Tables, Kaplan-Meier, or proportional hazards procedure.

Curve Estimation produces curve estimation regression statistics and related plots for 11 different curve estimation regression models. A separate model is produced for each dependent variable. One can also save predicted values, residuals, and prediction intervals as new variables.

Logistic estimates regression models in which the dependent variable is dichotomous. If the dependent variable has more than two categories, use the Discriminant procedure to identify variables which are useful for assigning the cases to the various groups. If the dependent variable is continuous, use the Linear Regression procedure to predict the values of the dependent variable from a set of independent variables. In recent versions there are two options **Binary Logistic** as well as **Multinomial Logistic.**

Probit performs probit analysis which is used to measure the relationship between a response proportion and the strength of a stimulus. For example, the probit procedure can be used to examine the relationship between the proportion of plants dying and the strength of the pesticide applied or to examine the relationship between the proportion of people buying a product and the magnitude of the incentive offered. The Probit procedure should be used only if the response is dichotomous buy/not buy, alive/dead—and several groups of subjects are exposed to different levels of some stimulus. For each stimulus level, the data must contain counts of the totals exposed and the totals responding. If the response variable is dichotomous but you do not have groups of subjects with the same values for the independent variables you should use the Logistic Regression procedure.

Nonlinear estimates nonlinear regression models, including models in which parameters are constrained. The nonlinear regression procedure can be used if one knows the equation whose parameter sare to be estimated, and the equation cannot be written as the sum of parameters times some function of the independent variables. In nonlinear regression the parameter estimates are obtained iteratively. If the function is linear, or can be transformed to a linear function, then the Linear Regression procedure should be used.

Weight Estimation estimates a linear regression model with differential weights representing the precision of observations. This command is in the Professional Statistics option. If the

variance of the dependent variable is not constant for all of the values of the independent variable, weights which are inversely proportional to the variance of the dependent variable can be incorporated into the analysis. This results in a better solution. The Weight Estimation procedure can also be used to estimate the weights when the variance of the dependent variable is related to the values of an independent variable. If you know the weights for each case you can use the linear regression procedure to obtain a weighted least squares solution. The linear regression procedure provides a large number of diagnostic statistics which help you evaluate how well the model fits your data.

2-Stage Least Squares performs two-stage least squares regression for models in which the error term is related to the predictors. This command is in the Professional Statistics option. For example, if you want to model the demand for a product as a function of price, advertising expenses, cost of the materials, and some economic indicators, you may find that the error term of the model is correlated with one or more of the independent variables. Two-stage least squares allows you to estimate such a model.

The **Log linear** submenu provides general and hierarchical log-linear analysis and logit analysis.

Classify

This submenu provides cluster and discriminant analysis.

Two Step Cluster performs Two Step Cluster Analysis procedure which is an exploratory data analysis tool designed to reveal natural clustering within a dataset that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques. The Log-likelihood and Euclidean Distance Measures are used as the similarity measure between two clusters.

K-means Cluster performs cluster analysis using an algorithm that can handle large numbers of cases, but that requires you to specify the number of clusters. The goal of cluster analysis is to identify relatively homogeneous groups of cases based on selected characteristics. If the number of clusters to be formed is not known, then Hierarchical Cluster procedure can be used. If the observations are in known groups and one wants to predict group membership based on a set of independent variables, then the Discriminant procedure can be used.

Hierarchical Cluster combines cases into clusters hierarchically, using a memory-intensive algorithm that allows you to examine many different solutions easily.

Discriminant is used to classify cases into one of several known groups on the basis of various characteristics. To use the Discriminant procedure the dependent variable must have a limited number of distinct categories. Independent variables that are nominal must be recoded to dummy or contrast variables. If the dependent variable has two categories, Logistic Regression can be used. If the dependent variable is continuous one may use Linear Regression.

Nearest Neighbor performs Nearest Neighbor Analysis for classifying cases based on their similarity to other cases. In machine learning, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity.

Dimension Reduction

This submenu provides factor analysis, correspondence analysis, and optimal scaling.

Factor is used to identify factors that explain the correlations among a set of variables. Factor analysis is often used to summarize a large number of variables with a smaller number of derived variables, called factors.

Correspondence Analysis analyzes correspondence tables (such as cross-tabulations) to best measure the distances between categories or between variables. This command is in the Categories option.

Distances compute many different measures of similarity, dissimilarity or distance. Many different measures can be used to quantify how much alike or how different two cases or variables are. Similarity measures are constructed so that large values indicate much similarity and small values indicate little similarity. Dissimilarity measures estimate the distance or unlikeness of two cases. A large dissimilarity value tells that two cases or variables are far apart. In order to decide which similarity or dissimilarity measure to use, one must consider the characteristics of the data. Special measures are available for interval data, frequency counts, and binary data. If the cases are to be classified into groups based on similarity or dissimilarity measures, one of the Cluster procedures should be used.

The **Scale** submenu provides reliability analysis and multidimensional scaling.

Nonparametric Tests

This submenu provides nonparametric tests for one sample, or for two and more paired or independent samples. Legacy dialogs sub-submenu consists following tests:

Chi-Square is used to test hypotheses about the relative proportion of cases falling into several mutually exclusive groups. For example, if one wants to test the hypotheses that people are equally likely to buy six different brands of cereals, one can count the number buying each of the six brands. Based on the six observed counts Chi-Square procedure could be used to test the hypothesis that all six cereals are equally likely to be bought. The expected proportions in each of the categories don't have to be equal. The hypothetical proportions to be tested should be specified.

Binomial is used to test the hypothesis that a variable comes from a binomial population with a specified probability of an event occurring. The variable can have only two values. For example, to test that the probability of an item on the assembly line is defective is one out of ten ($p=0.1$), take a sample of 300 items and record whether each is defective or not. Then use the binomial procedure to test the hypothesis of interest.

Runs are used to test whether the two values of a dichotomous variable occur in a random sequence. The runs test is appropriate only when the order of cases in the data file is meaningful.

1. Sample K-S is used to compare the observed frequencies of the values of an ordinal variable, such as rated quality of work, against some specified theoretical distribution. It determines the statistical significance of the largest difference between them. In SPSS, the theoretical distribution can be **Normal, Uniform or Poisson.** Alternative tests for normality are available in the Explore procedure, in the Summarize submenu. The P-P and Q-Q plots in the Graphs menu can also be used to examine the assumption of normality.##

2. Independent Samples is used to compare the distribution of a variable between two nonrelated groups. Only limited assumptions are needed about the distributions from which the sample are selected. The Mann-Whitney U test is an alternative to the two sample t-test. The actual values of the data are replaced by ranks. The Kolmogorov-Smirnov test is based

on the differences between the observed cumulative distributions of the two groups. The Wald-Woflowitz runs tests sorts the data values from smallest to largest and then performs a runs test on the group's numbers. The Moses Test of Extreme Reaction is used to test for differences in range between two groups.

K-Independent Samples is used to compare the distribution of a variable between two or more groups. Only limited assumptions are needed about the distributions from which the samples are selected. The Kruskal-Wallis test is an alternative to one-way analysis of variance, with the actual values of the data replaced by ranks. The Median tests counts the number of cases in each group that are above and below the combined median, and then performs a chi-square test.

2 Related Samples is used to compare the distribution of two related variables. Only limited assumptions are needed about the distributions from which the samples are selected. The Wilcoxon and Sign tests are nonparametric alternative to the paired samples t-test. The Wilcoxon test is more powerful than the Sign test. *McNemar's test* is used to determine changes in proportions for related samples. It is often used for "before and after" experimental designs when the dependent variable is dichotomous. For example, the effect of a campaign speech can be tested by analyzing the number of people whose preference for a candidate changed based on the speech. Using *McNemar's test* you analyze the changes to see if change in both directions is equally likely.

K Related Samples is used to compare the distribution of two or more related variables. Only limited assumptions are needed about the distributions from which the samples are selected. *The Friedman test* is a nonparametric alternative to a single-factor repeated measures analysis of variance. One can use it when the same measurement is obtained on several occasions for a subject. For example, the Friedman test can be used to compare consumer satisfaction of 5 products when each person is asked to rate each of the products on a scale. *Cochran's Q test* can be used to test whether several dichotomous variables have the same mean. For example, if instead of asking each subject to rate their satisfaction with five products, you asked them for a yes/no response about each, you could use Cochran's test to test the hypothesis that all five products have the same proportion of satisfied users. *Kendall's W measures* the agreement among raters. Each of your cases corresponds to a rater; each of the selected variables is an item being rated. For example, if you ask a sample of customers to rank 7 ice-cream flavours from least to most liked, you can use Kendall's W to see how closely the customers agree in their ratings.

Forecasting

This submenu provides create models, seasonal decomposition, spectral analysis, autocorrelations, cross-correlations etc.

Autocorrelations calculates and plots the autocorrelation function (ACF) and partial autocorrelation function of one or more series to any specified number of lags, displaying the Box-Ljung statistic at each lag to test the overall hypothesis that the ACF is zero at all lags.

Cross-correlations calculate and plot the cross-correlation function of two or more series for positive, negative, and zero lags.

Spectral analysis calculates and plots univariate or bivariate periodograms and spectral density functions, which express variation in a time series (or covariation in two time series)

as the sum of a series of sinusoidal components. It can optionally save various components of the frequency analysis as new series.

The **Survival** submenu provides techniques for analyzing the time for some terminal event to occur, including Kaplan-Meier analysis and Cox regression.

Multiple Responses submenu provides facilities to define and analyze multiple-response or multiple-dichotomy sets.

Quality Control submenu provides facilities to for obtaining control charts and Pareto charts.

Complex Samples submenu provides procedures for Sampling from Complex Designs. The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind.

Other than this **Analyze** menu there are several other important menus available in SPSS.

Transform

Compute calculates the values for either a new or an existing variable, for all cases or for cases satisfying a logical criterion.

Random Number Seed sets the seed used by the pseudo-random number generator to a specific value, so that you can reproduce a sequence of pseudo-random numbers.

Count creates a variable that counts the occurrences of the same value(s) in a list of variables for each case.

Recode into Same Variables reassigns the values of existing variables or collapses ranges of existing values into new values.

Recode into Different Variables reassigns the values of existing variables to new variables or collapses ranges of existing values into new variables.

Rank Cases creates new variables containing ranks, normal scores, or similar ranking scores for numeric variables.

Automatic Recode reassigns the values of existing variables to consecutive integers in new variables.

Create Time Series creates a time-series variable as a function of an existing series, for example, lagged or leading values, differences, cumulative sums. This command is in the Trends option.

Replace Missing Values substitutes non-missing values for missing values, using the series mean or one of several time-series functions. This command is in the Trends option.

Run Pending Transforms executes transformation commands that are pending due to the Transformation Options setting in the Preferences dialog.

Utilities

Command Index takes you to the dialog box for a command if you know its name in the SPSS command language.

Fonts let you choose a font, style, and size for SPSS Data Editor, output, and syntax windows.

Variable Information displays the Variables window, which shows information about the variables in your working data file, and allows you to scroll the data editor to a specific variable, or copy variable names to the designated syntax window.

File Information displays information about the working data file in the output window.

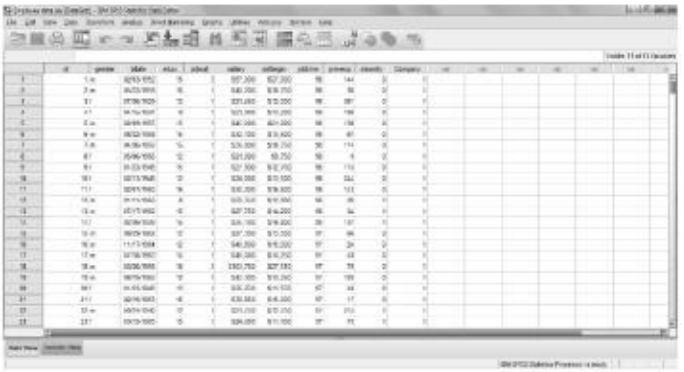
Output Page Titles lets you specify a title and subtitle for output from SPSS. They appear in the page header, if it is displayed. (Preferences in the Edit menu controls the page header.)
Define Sets defines sets of variables for use in other dialog boxes.
Use Sets lets you select which defined sets of variables should appear in the source-variable lists of other dialog boxes.
Grid Lines turns grid lines on and off in the Data Editor window. This command is available when the Data Editor is active.
Value Labels turns on and off the display of Value Labels (instead of actual values) in the Data Editor window. When Value Labels are displayed you can edit data with a pop-up menu of labels. This command is available when the Data Editor is active.
Auto New Case turns on and off the automatic creation of new cases by cursor movement below the last case in the Data Editor window. This command is available when the Data Editor is active.
Designate Window designates the active window to receive output from SPSS commands (if it is an output window); or to receive commands pasted from dialog boxes (if it is a syntax window). You can also designate a window by clicking the! Button on its icon bar. This command is available when an output or syntax window is active.

Graphs

The Chart Builder available in Graph menu allows to build charts from predefined gallery charts or from the individual parts (for example, axes and bars). You build a chart by dragging and dropping the gallery charts or basic elements onto the canvas, which is the large area to the right of the Variables list in the Chart Builder dialog box.
Legacy Dialogs submenu provides following graph submenus
Bar generates a simple, clustered, or stacked bar chart of the data.
3-D Bar Charts allows to generating bar graph in 3-dimensional axis.
Line generates a simple or multiple line charts of the data.
Area generates a simple or stacked area chart of the data.
Pie generates a simple pie chart or a composite bar chart from the data.
High-Low plots pairs or triples of values, for example high, low, and closing prices.
Box plot generates box plots showing the median; inter quartile range, outliers, and extreme cases of individual variables.
Error Bar Charts plot the confidence intervals, standard errors, or standard deviations of individual variables.
Scatter/dot generates a simple or overlay scatter plot, a scatter plot matrix, or a 3-D scatter plot from the data.
Histogram generates a histogram showing the distribution of an individual variable.

PRACTICAL EXERCISES

Exercise 1: Here, a practical exercise is provided in which SPSS has been used for Survey Data Analysis. For illustration purpose, we are going to use "Employee data" from the sample folder of SPSS available at Samples folder in the Program Files folder in C drive. In addition a new variable "Company" has been added to the dataset which is having values from 1,2,…,10. Finally, there are 400 data-points clustered into 10 clusters each of size 40 units. This dataset has been considered as population used for further illustration (available at https://www.dropbox.com/s/rxxccpuk3jieepa/Employee%20data.sav?dl=0).

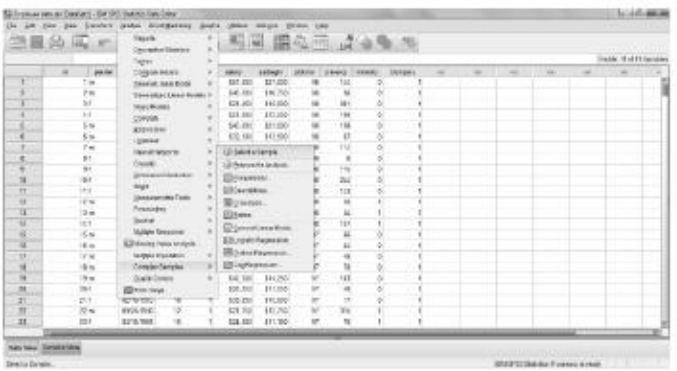


The Sampling Wizard guides through the steps for creating, modifying, or executing a sampling plan file. Before using the Wizard, one should have a well-defined target population, a list of sampling units, and an appropriate sample design in mind. The Complex Samples option allows to select a sample according to a complex design and incorporate the design specifications into the data analysis

Creating a New Sample Plan

1. From the menus choose

Analyze → Complex Samples → Select a Sample....



2. Select **Design a sample** and choose a plan filename to save the sample plan and click **Next.**

3. Optionally, in the Design Variables step, one can define strata, clusters, and input sample weights.

In our current example we wish to select samples employing Cluster and Multistage Sampling designs. These sampling designs are most popular in survey sampling. Thus, select the variable "Company" as **cluster**. Then, click **Next**.

4. In the **Sampling Method** step, one can choose a method for selecting items.

- If one select PPS Brewer or PPS Murthy, one can click Finish to draw the sample. Otherwise, click **Next.**

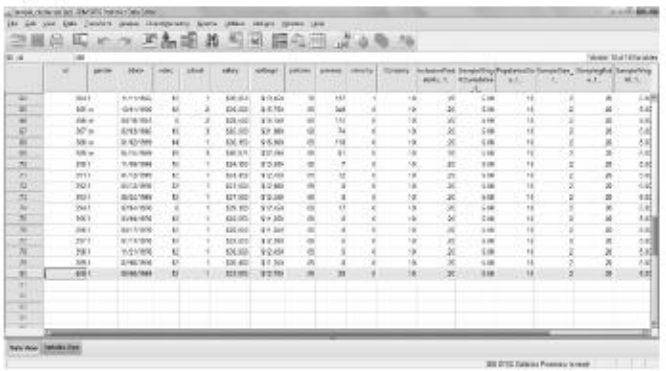
5. In the **Sample Size** step, specify the number or proportion of units to sample.
6. Optionally, in further steps one can:

- Choose output variables to save.
- Add a second or third stage to the design.

- Set various selection options, including which stages to draw samples from, the random number seed, and whether to treat user-missing values as valid values of design variables.
- Choose where to save output data.

7. Now click **Finish** to draw the sample. Following selected sample contains 80 units (2 clusters containing 40 units each).

→→



Developed Sample Plan can be used for furthermore random sample selection as follows:

Analyze → Complex Samples → Draw the Sample...

After selection of Sample next step is to prepare the selected sample for analysis. The Analysis Preparation Wizard guides through the steps for creating or modifying an analysis plan for use with the various Complex Samples analysis procedures. Before using the Wizard, one should have a sample drawn according to a complex design.

Creating a New Analysis Plan

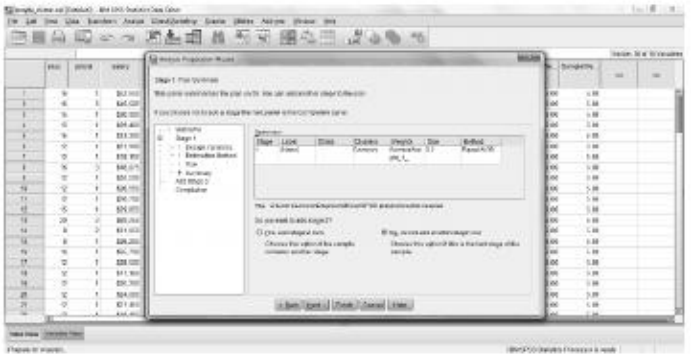
1. From the menus choose:

Analyze → Complex Samples → Prepare for Analysis...

2. Select **Create a plan file**, choose a filename to save the analysis plan and click **Next**.
3. Specify the variable containing sample weights in the **Design Variables** step, optionally defining strata and clusters.
4. Optionally, in further steps one can:

- Select the method for estimating standard errors in the Estimation Method step.
- Specify the number of units sampled or the inclusion probability per unit in the Size step.
- Add a second or third stage to the design.

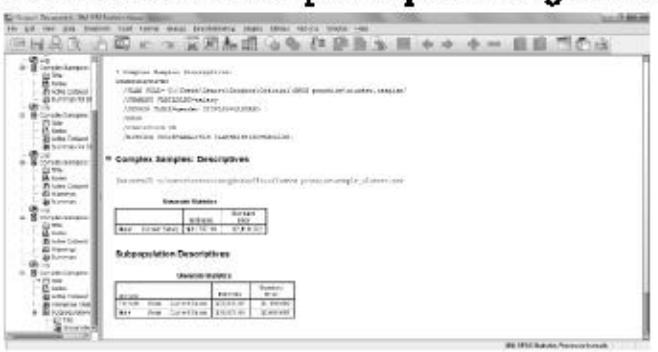
5. Now click **Finish** to save the plan.



Now using this Analysis Plan file one generates several types of outputs available in the Complex Samples option like

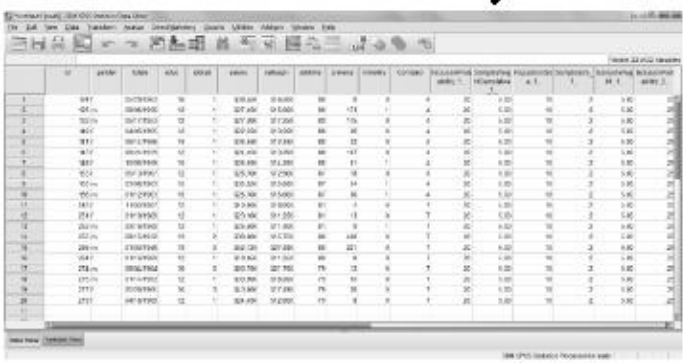
- **Frequencies**
- **Descriptives**
- **Crosstabs**
- **Ratios**
- **General Linear Model**
- **Logistic Regression**
- **Ordinal Regression**
- **Cox Regression**

Results from the Descriptives options using the "Current Salary" is given by



For selection of samples by Multistage sampling design one can edit the existing Sample Plan for cluster sampling or prepare new sampling plan according to Multistage sampling.
At the seventh step of the earlier shown "Creating a New Sample Plan", one should select "Yes, add stage 2 now" when the question "Do you want to add Stage 2" pops up in the sampling wizard as shown below:

Then define **"sample size"** for the stage 2 and give path where to save the output file. An output file is given below. First, 2 clusters are selected by SRSWOR and, the, within each selected cluster 10 units are selected by SRSWOR.



For analysis as per two stage sampling design, New Analysis Plan shall be created and further desired analysis of the sample shall be carried out.

Exercise 2: The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot).

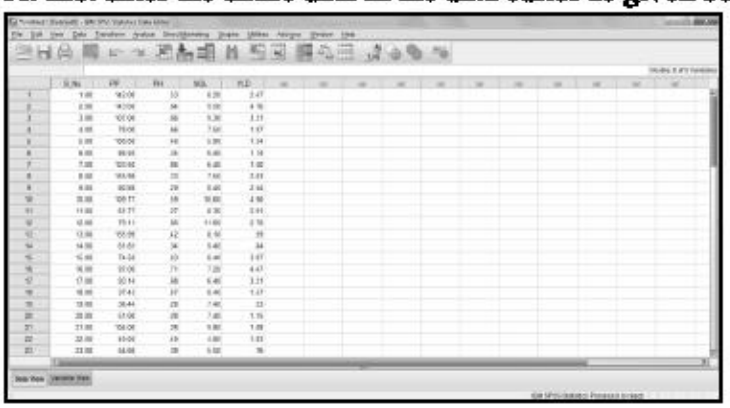
S.No.	PP	PH	NGL	Yield	S.No.	PP	PH	NGL	Yield
1	142.00	0.525	8.2	2.470	24	55.55	0.265	5.0	0.430
2	143.00	0.640	9.5	4.760	25	88.44	0.980	5.0	4.080
3	107.00	0.660	9.3	3.310	26	99.55	0.645	9.6	2.830
4	78.00	0.660	7.5	1.970	27	63.99	0.635	5.6	2.570
5	100.00	0.460	5.9	1.340	28	101.77	0.290	8.2	7.420
6	86.50	0.345	6.4	1.140	29	138.66	0.720	9.9	2.620
7	103.50	0.860	6.4	1.500	30	90.22	0.630	8.4	2.000
8	155.99	0.330	7.5	2.030	31	76.92	1.250	7.3	1.990
9	80.88	0.285	8.4	2.540	32	126.22	0.580	6.9	1.360
10	109.77	0.590	10.6	4.900	33	80.36	0.605	6.8	0.680
11	61.77	0.265	8.3	2.910	34	150.23	1.190	8.8	5.360
12	79.11	0.660	11.6	2.760	35	56.50	0.355	9.7	2.120
13	155.99	0.420	8.1	0.590	36	136.00	0.590	10.2	4.160
14	61.81	0.340	9.4	0.840	37	144.50	0.610	9.8	3.120
15	74.50	0.630	8.4	3.870	38	157.33	0.605	8.8	2.070
16	97.00	0.705	7.2	4.470	39	91.99	0.380	7.7	1.170
17	93.14	0.680	6.4	3.310	40	121.50	0.550	7.7	3.620

18	37.43	0.665	8.4	1.570	41	64.50	0.320	5.7	0.670
19	36.44	0.275	7.4	0.530	42	116.00	0.455	6.8	3.050
20	51.00	0.280	7.4	1.150	43	77.50	0.720	11.8	1.700
21	104.00	0.280	9.8	1.080	44	70.43	0.625	10.0	1.550
22	49.00	0.490	4.8	1.830	45	133.77	0.535	9.3	3.280
23	54.66	0.385	5.5	0.760	46	89.99	0.490	9.8	2.690

Source: Design Resources Server. ICAR - Indian Agricultural Statistics Research Institute, New Delhi 110 012, India. www.iasri.res.in/design (accessed lastly on <05-05-2015>).

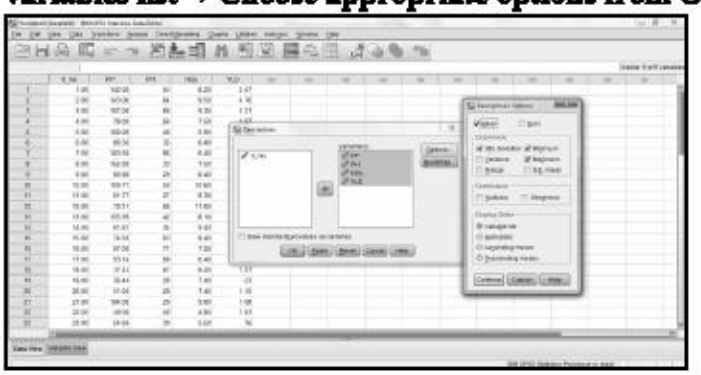
1. Find mean, standard deviation, minimum and maximum values of all the characters.
2. Find correlation coefficient between each pair of the variables.
3. Give a scatter plot of the variable PP with dependent variable yield.
4. Fit a multiple linear regression equation where yield is dependent variable whereas all other characters as independent variables.

At first enter the entire data in the data editor as given below.

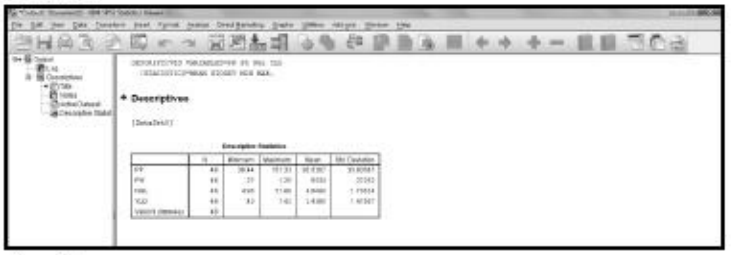


There are several ways to answer Q no. 1 in SPSS. Commands following first way are as follows

Analyze → Descriptive Statistics → Descriptives...→ Put PP, PH, NGL, YLD in the variables list→ Choose appropriate options from Options tab → Press Continue → Ok

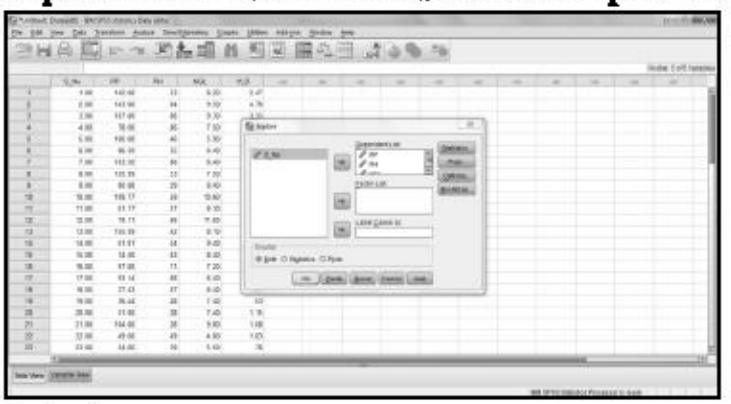


Output

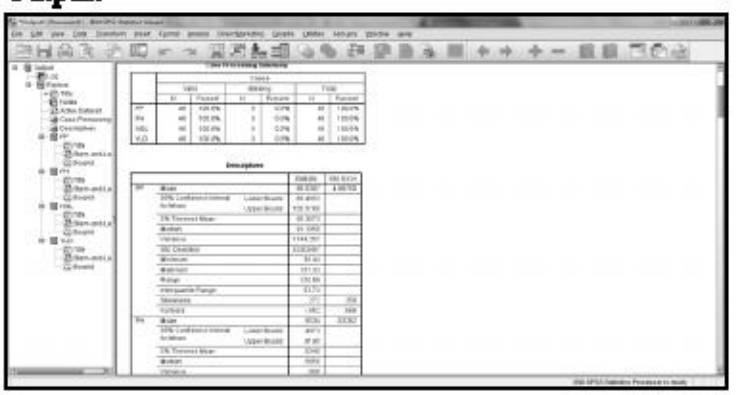


Another way:

Analyze → Descriptive Statistics → Explore...→ Put PP, PH, NGL, YLD in the Dependent list→ Choose both Statistics and plot→ Press Ok

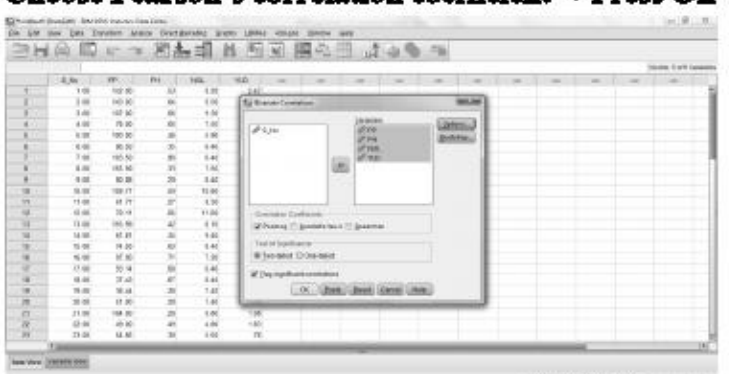


Output:

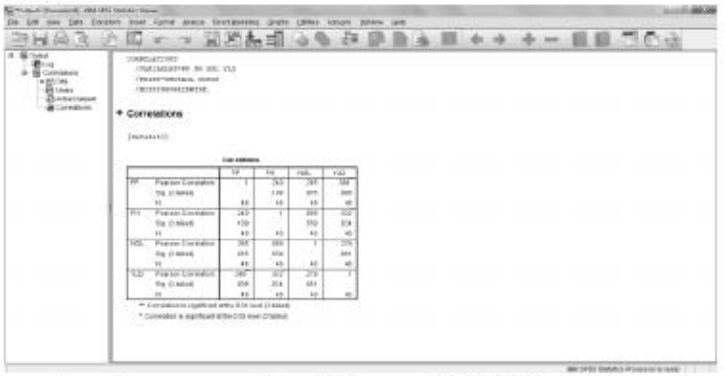


To answer Q no 2 follow the following steps

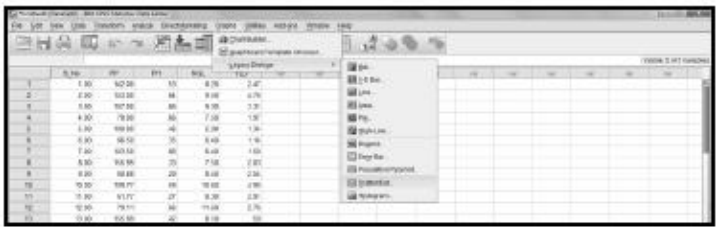
Analyze → Correlate → Bivariate→ Put PP, PH, NGL, YLD in the Valiables list→ Choose Pearson's correlation coefficient→ Press Ok

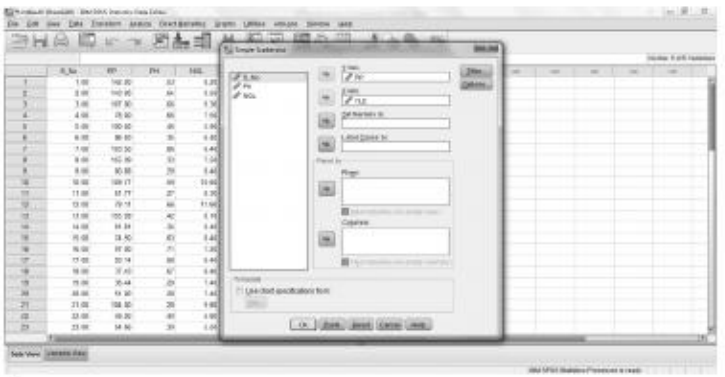


Output:

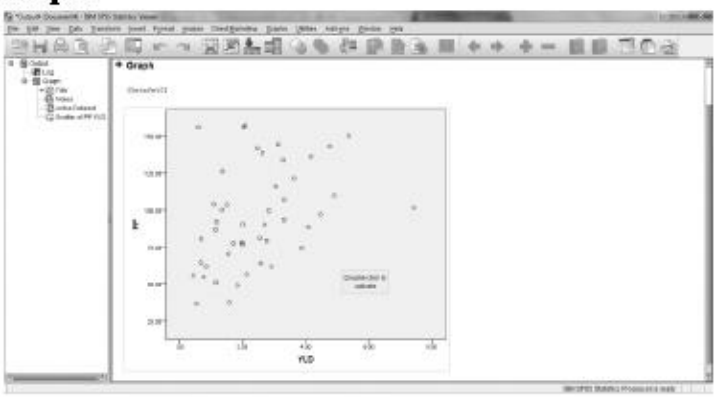


To give the scatter plot of the variable PP with dependent variable yield use following steps:
Graphs → Legacy dialogs→ Scatter plot→ Put PP at Y axis and YLD at X axis→ Press Ok



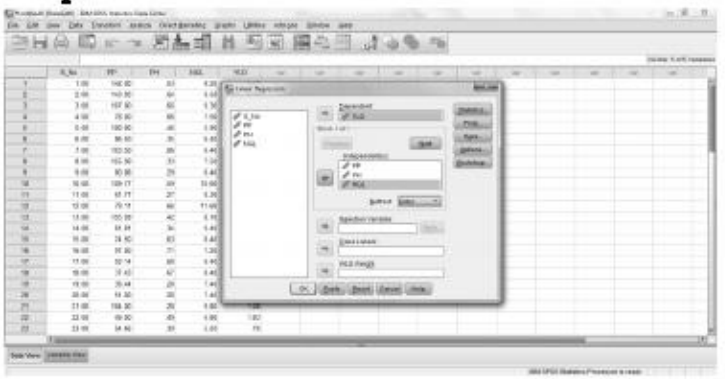


Output:

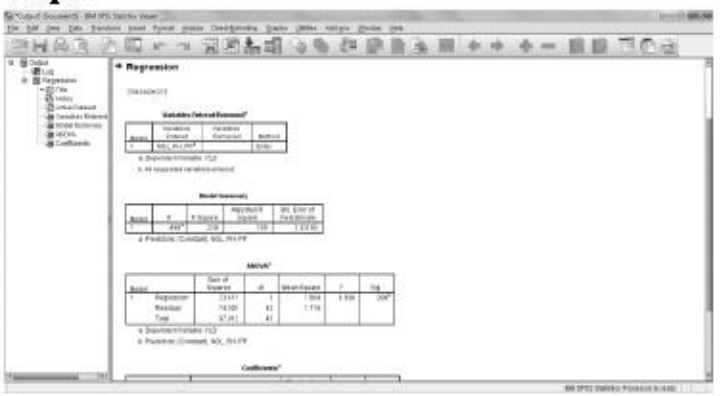


To fit a multiple linear regression equation taking yield as dependent variable and all other characters as independent variables perform following steps

Analyze → Regression → Linear → Put Yld in Dependent variable and PP, PH, NGL in independent variable list → Press Ok



Output:



REFERENCES

Design Resources Server. Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India. www.iasri.res.in/design (accessed lastly on 05 May, 2015).

Morgan, G.A., Leech, N.L., Gloeckner G.W. and Barrett, K.C. (2012).*IBM SPSS for Introductory Statistics: Use and Interpretation*. Fifth Edition, Routledge.

Nie, N.H., Bent, D.H. and Hull, C.H. (1970). *SPSS: Statistical Package for the Social Sciences*. New York: McGraw-Hill.

General Introduction to STATA and its Environment

STATA is statistical software, which have characteristic of a package and language for statistical computations. Although STATA is meant for statistical computations, it has strong capability to handle large databases (for off line work). It works as menu based and command based system. STATA commands are often simple enough that it is faster to use them directly. This will be especially true once you become familiar with the commands you will use the most in your regular use of STATA.

STATA User Interface

When you start STATA the five main windows will be shown: Review, Variables, Results, Command and Properties. These are initially contained in the main STATA window. These five windows are typically in use the whole time STATA is open. There are other more specialized windows, such as the Viewer, Data editor, Data browser, Do-file editor, Graph, Help and Graph editor windows. The functions of important windows are:

(1) Data Browser: to see the contents of data loaded in memory;

(2) Data Editor: to modify the data loaded in memory;

(3) Do Editor: for writing and executing STATA program;

(4) Viewer: to see log files and other text type files;

(5) Variable window: Contains list of variables. Variables from this window can be exported to command window (by mouse click);

(6) Review window: Previous commands of current session are kept in this window and can be exported to command window (by mouse click);

(7) Result window: Results of command are displayed in this window. Results can be exported to text file or spreadsheet;

(8) Command window: Instruction in form of command is given in this window (they can be recalled from the Review window or by <Page UP> option of Key Board);

(9) Help: To see on-line help.

Basic Unit of Data

Data in STATA is organised in tabular (rectangular) form i.e. in the form of rows and columns. Each row in STATA data set (.dta file) is called an observation (record), which corresponds to the object (we can call it identifier) about which data pertain. To recognise identifiers, is most fundamental and essential aspect of understanding the provided data. Basic unit of STATA data set is a column (variable). Each column (variable) has its own name, type and characteristics and contains particular aspect of observation. Name of variables in STATA may be 32 characters long, but for compatibility with earlier versions and other packages it is better to use 8 character long variable names. Since STATA is case sensitive, use of different cases in name of variable will be treated as different variable. We can divide columns in two parts according to characteristics:

(a) column/columns as part of identifiers (b) column/columns for explaining the characteristics of the identifiers. For example columns related with household number and individual id number may be part of identifiers and column related with age and sex may be used for explaining the identifiers in certain observations. These columns can contain

information in both forms i.e. in string as well as numeric form. In fact how data, belonging to particular column will be stored in memory and which type of operation can be performed on it, depends on type of column. Types for numeric values in STATA are:

Type	Bytes	Minimum	Maximum	Closest value to 0
byte	1	-127	126	+/-1
int	2	-32767	32,766	+/-1
long	4	-2147483647	2,147,483,646	+/-1
float	4	$-1.70141173319 \times 10^{38}$	$1.70141173319 \times 10^{38}$	$+/-10^{-38}$
double	8	$-8.9884656743 \times 10^{307}$	$8.9884656743 \times 10^{307}$	$+/-10^{-323}$

Numbers are stored as byte, int (integer), long, float, or double, with the default being float. byte, int, and long are said to be of integer type, as they can hold only integers.

Stata keeps data in its memory and it should be recorded as parsimoniously as possible. If you have a string variable that has maximum length 6, it would waste of memory to store it as a str20 (string with 20 characters). Similarly, if you have an integer variable, it would be a waste to store it as a double.

Remark: Stata puts a value max (according to type)+1 in numeric type columns for blank entry. Hence blank is not really blank in real data set. For example Stata puts 32,767 for blank in column of an integer type.

Strings are stored as str# (# indicates the 'place of number' in Stata), for instance, str1, str2, str3, ..., str80. The number after the 'str' indicates the maximum length of the string. A column of type str5 could hold a word of maximum length 5 and hence it could store word "male", but not the word "female" because "female" contains six characters.

Unlike many databases, STATA does not have a separate data type for date. Dates are stored as number in database but they can be viewed and seen with the available date formats. Manipulation on dates is very easy through its functions. Apart from data columns, data may be kept in memory through macro and matrix. Details of manipulation of date, macro and matrix will be discussed later on.

Functions: At most upper layer, STATA assists you through its command and function. Each STATA functions could be categorised according to following classes:

- Mathematical functions
- Statistical functions
- Random numbers
- String functions
- Special functions
- Date functions
- Time-series functions
- Matrix functions returning scalar values

In STATA there are two types of functions - (1) functions used in expressions (2) specific functions used in particular commands like **egen** (to generate a new column). The first category of functions are the free (independent of specific command) functions, which can be used in creation of any expression. Under second category, functions tied with any particular command cannot be used in expressions. Each function may be characterised by its domain (input or argument) and range (output). General format for STATA function is

Function Name (list of argument separated by space or comma)

Generally, arguments (options) of independent functions are separated by comma while functions tied with particular command are separated by space. For example see the **recode** function (which is independent function) for generating the categorical data from continuous data. Here arguments are separated by comma while for **'rsum'** function, tied with **egen** command, arguments are separated by space.

Expressions in STATA: An expression in STATA is composed through combination of constants, variables, operators and functions. String Constants are given within double quotes (" ").

Operators: There are three types of operators in STATA: (1) Arithmetic (2) Logical (3) Relational. Following are the list of operators with their symbols.

Arithmetic	**Logical**	**Relational (numeric and string)**
+ addition	~ not	> greater than
- subtraction	| or	< less than
* multiplication	& and	>=> or equal
/ division		<=< or equal
^ power		== equal
		~= not equal
+ string concatenation		

Note that a double equal sign (==) is used for equality testing while = sign is used for assignment.

Component of STATA Command: Components of STATA commands are: (1) command name (2) arguments (3) by (4) if (5) in (6) weight (7) using (8) options. We can call **by, if, in, weight, using, option** as clauses for STATA command sentence.

General command syntax is

[by varlist:] command [varlist] [=exp] [if exp] [in range] [weight][using filename] [, options]

Here 'varlist' is used for list of variables and 'exp' is used for any expression. We can specify 'varlist' using wild cards like -,*,? (var11, var12 can be called as var?? or var*). In fact we get two terms - varlist and numlist, frequently used in STATA help. 'varlist' is list of variables and 'numlist' is list of numbers. There are certain rules (for short-cut) for defining these rules. Anything written in [ ] is optional (command may run without this. Components following the command name before any clause (if present) are called arguments. Sequence of if, in, weight might be changed in STATA command sentence. From general command syntax, we can see that only command name is essential part. But it is not true for all commands. For any specific command, which other components are essential; see help for that particular command. For using the help from command window syntax it can be given as

help [command name]

Alternatively, one can use the search command. Syntax of this command is
search [command name]
For practice, see the help for 'use' 'log' and 'tabulate' commands and list the components of these commands. Find the components those are essential.

Remark: STATA is case sensitive. All command sentences are written in lower case only. Name of variables or part of variable name may be in upper case up to width 32.

Help mechanism: Components of help are:

(1) General command syntax: Here general command syntax with abbreviations (if used) is placed. Underlined letters identifies abbreviations for a command which can be used to run the command instead of typing the full spelling. For example, command tabulate can be run by typing only tab.

(2) Description: Here purpose of command is explained.

(3) Options: Gives the detail about behaviour of command, which can be changed through different options.

(4) Example: This part of help explains how command may be used.

(5) See also: Gives the list of other related commands.

Broad classification of STATA commands for tabulation On the basis of purpose, STATA commands may be classified in eight parts

(1) Commands for loading (or importing) and saving data in main memory: Commands like **use, infile, insheet, infix, save, outfile, outsheet** belong to this class. In recent versions of STATA **import** command has been added for importing data from other sources.

(2) Commands related to data manipulation (changing existing data): Commands like **generate, egen, edit, sort, recode, xtile, pctile** belong to this class.

(3) Commands related to tabulation (or summarising): Commands like **tabulate, summarize, table, tabstat** belong to this class.

(4) Command related to combining data contained in two files: Commands like **append** and merge (mmerge and xmerge are also available from internet) belong to this class.

(5) Commands for reshaping (re structuring) the data sets: Commands like reshape, **compress,** collapse, **separate** belongs to this class. Commands which have 'replace' option, behave as member of this class.

(6) Commands for controlling working environment of STATA: Commands like **set** (with different arguments), **log, cmdlog, more, for, cd, dir, type, shell, mkdir, copy, erase, help, search, view** belongs to this class.

(7) Commands for adding auxiliary information: Commands like **label, notes, rename** belongs to this class.

(8) Commands for displaying status of data: Commands like **describe, inspect, cf, compare, browse, list, count** belongs to this class. Command related with tabulation may be used for this purpose.

Role of Log Files

Different packages provide facility of log. These log files keep the history of work, performed on computer. Sometimes these log files are created and maintained automatically and sometime by user. In STATA, these log files are used for keeping history of STATA commands for preparing program files and recording the results for preparing reports. Main characteristics of STATA log files are following:

1. Two types of log files may be created simultaneously (but not necessarily) with two different names of files in STATA - one for recording history of command only and other for command and results both.
2. Command for opening log files for keeping history of command only is **'cmdlog'**. For purpose of recording results and command both, command is **'log'**
3. Both types of log files can be opened, closed and suspended.
4. Currently opened log file can be viewed through viewer without closing it.

Remark: Although not closing STATA log files before ending the STATA session cause any problem but sudden termination of STATA program (due to power failure) may disturb log file.

Some Important Concepts - Tabulation

Following concepts are not tied with STATA command. These concepts are more or less needed in any tabulation, although implementation of these concepts may be different in different relational databases.

Scope of Command

Whenever a command issued in STATA, it is important to think about scope of command. It is the scope of command, which specifies whether command should be applied on whole dataset or a part of it. 'if' and 'in' clauses reduce the scope but 'by' clause classify the scope of command. Commands like tabulate or table also classify scope of command. For example consider the following data

Data - 1

hhno	ocpcode	dist	caste	income
1	1	1	1	100
1	2	1	1	200
2	1	1	1	200
2	2	1	1	100
3	1	1	2	0
3	2	1	2	100
4	1	1	1	100
4	2	1	1	500
5	1	1	2	300
5	2	1	2	0
6	1	1	2	100
6	2	1	2	0
7	1	1	1	500
7	2	1	1	0
8	1	1	1	0
8	2	1	1	1000
9	1	2	2	100
9	2	2	2	200
10	1	2	1	0
10	2	2	1	100

11	1	2	2	100
11	2	2	2	500
12	1	2	1	100
12	2	2	1	600

The command **'summarize income'**, summarize has scope on total data. In command **'tab caste ocpcode if dist==1, summ(income)'**, scope of summarize has reduced to district with code 1. Again scope of summarize is partitioned by caste and occupation group. Total 9 groups (by including total, caste and occupation has 3 groups each) are formed and for each group means are calculated within the scope.

Analysis of Base Statement

Very often people start preparation of table before analysing the basic statement (title) of table. Some time titles are given and some time it has to be created. Basic problem arises when basic statement for table is not created in structured manner. There are four components of structured statement for table: (1) summary (2) summary variables (3) base (4) by variables. 'Summary' part contains statistical procedure, like mean, sum, min, max, percent etc. on data. 'Summary variable' is such variable on which statistical procedure has been applied. Identifying **base** for tabulation is most fundamental part but sometimes it may be hidden in the title statement (it generally follow 'of'). By variables are variables through which scope of statistical procedure is classified. A structured statement is written in following manner

summary: summary variable: base: [by]

It is pertinent to mention that 'by' is optional.

Now consider the basic statement for creating a table (based on Data-1):

"mean of income by caste". Sometimes, for same purpose title may be given as: "mean income of caste". There may be other ways to write the title. But above-mentioned titles are not structured. Second statement is more nearer to structured statement. If we follow the rule for structured statement, above statements may be written as (a) and (b) respectively.

(a) mean: income: missing: caste

(b) mean: income: caste: missing

In first statement 'base' part is missing and hence it is not clear whether income should be calculated for individuals or households (as both are given). In second statement 'by' part is missing which is not necessary but use of 'by' part gives clear picture of relationship. Real structured statement may be obtained by filling missing part in (a).

After writing structured statement, most important aspect is to check layout data on which tables should be prepared. Here fundamental rule is **data should be uniquely identified on base.** For example if missing part is filled as 'household' (and hence real statement is 'mean income of household by caste') and data is supplied by Data-1. This data cannot be used as such for the purpose of making above mentioned table directly, because data is not uniquely identified on household. This data can be directly used for statement 'mean income of occupation of household by caste'. In this case parts of statement will be as follows:

mean: income: occupation of household: caste

Before tabulation always check whether any indication of reduction of scope. For example if statement is given as "mean income of household by caste for district 1" or "mean income of household by caste for occupation 1 and district 1"

Categorical Vs. Continuous Data

Information contained in variable and represented by number may be categorised in two types - quantitative or qualitative. Information about quantities or value like quantity or value of some serial number is example of quantitative data. Sometime numbers may represent different categories like caste or land group. Such variables are called qualitative or categorical data. These variables are used for categorising the scope. In general there is no meaning of applying operation like mean or sum on such variables. On the basis of quantitative data, categorical data may be generated through **recode** function through generate command or **cut** function in **egen** command.

Composition and Incidence

These two concepts are frequently used in preparation of table. These are represented in terms of percent. Term composition is alternatively used as Distribution. This distribution is represented in 100 and generally 'tabulates' command is used in STATA for this purpose. Suppose data is given as follows

Data - 2

hhno	idcode	sex	edu	caste
1	1	1	Illiterate	2
1	2	2	Illiterate	2
1	3	2	Illiterate	2
1	4	1	Below prim.	2
1	5	2	Un formal	2
2	1	1	Below prim.	1
2	2	2	Middle	1
2	3	1	High	1
2	4	1	Un formal	1
2	5	2	Illiterate	1

For sex, 1 represents male and 2 represents female

For caste, 1 represents General caste and 2 represent SC

Now if we want to know the composition of sex on the basis of caste (i.e. by caste). Structured statement may be written as:

composition: sex: individual of household: caste

Since data is uniquely identified on individual of household direct data may be used without any reshaping. Total number of persons by caste and sex is

caste	sex		Total
	male	female	
1	3	2	5
2	2	3	5
Total	5	5	10

For getting composition of caste, total number of male and female in each caste may be represented as percent of total number of persons in respective caste. Command used for purpose is

tab caste sex, row nof

caste	sex		Total
	male	female	
1	60	40	100
2	40	60	100
Total	50	50	100

Position of 100 gives information about variable for which composition is obtained. Since 100 is placed in line of caste, table gives information about composition of caste.

If we use the command

tab caste sex, col nof

caste	sex		Total
	male	female	
1	60	40	50
2	40	60	50
Total	100	100	100

Since position of 100 is placed in line of sex, this table gives composition of sex.

STATA can represent both compositions in single table. Through place of 100, we can identify composition of respective variable. For example if we use command

tab caste sex, row col nof

the result will be like this

caste	sex		Total
	male	female	
1	60	40	100
	60	40	50
2	40	60	100
	40	60	50
Total	50	50	100
	100	100	100

Here collection of upper figure in each cell (as 100 is placed in upper half of cell) gives composition of sex. Similarly collections of lower figure in each cell give composition of caste.

Incidence is third factor represented in data. In general other factors effect the table through reducing or further classifying the tables (through 'by' clause or as supper column variable). Incidence does not affect the scope. Incidences in certain groups give percentage of number of observations, satisfying the condition of incidence factor w.r.t. total number of observations belonging to that group. In other way incidence is some sort of rate. For example suppose we want to get the incidence of literacy by caste and sex on the basis of Data-2.

We get total number of person by caste and sex as follows:

tab caste sex, nolabel

caste	sex		Total
	1	2	
1	3	2	5
2	2	3	5
Total	5	5	10

Total number of literate person by caste and sex is as follows:
tab caste sex if edu~= "Illiterate", nolabel

caste	sex		Total
	1	2	
1	3	1	4
2	1	1	2
Total	4	2	6

Incidence of literacy may be obtained on the basis of above two tables as :

caste	sex	
	1	2
1	100.00	50.00
2	50.00	33.33
Total	80.00	40.00

Here we can see total number of male in caste 1 is 3 and out of these 3 all are literate. Hence incidence of literacy for male of caste 1 is 100. Similarly total number of female in caste 1 is 2 and out of 2 only 1 is literate hence incidence of literacy is 50.
For getting such incidences it is not necessary to generate two separate tables. First write the statement in structure form in following manner:
incidence: literacy: individual: caste and sex
Thus we see here literacy is new variable, which may be created on the basis of education. This variable contains two type of values 0 for all ids which are not literate and 100 which are literate. Now modify the structured form as
mean: literacy: individual: caste and sex
It is left as exercise, how this mean represents percent of literate people in each group.

Wide vs. long data

Many times it is required for computation to create separate records for information belonging to different columns or variables. For example

Data - 3

hhno	idgroup1	incom e1	idgroup2	income2	idgroup3	income 3
1	01	1000	020305	2000		
2	01	2000	0203	2000	0206	1000
3			0104	1000	0402	2000
4					01020304	4500
5			010203	2000	0304	2800
6	0102	2000			0104	500

7	03	4000				
8			010203	2000	02	3000
9					01	2500
10			0102030 4	2000		
11			01	1500	03	1500
12			01	500		

In this data three type of incomes, income from agriculture (income1), income from wage (income2) and income from business (income3) are given for 12 households. Similarly idgroup1, idgroup2, idgroup3 represent list of idcode involved in respective occupations. Other form of database for similar information may be as follows:

Data - 4

hhno	ocpcode	idgroup	income
1	1	01	1000
1	2	020305	2000
2	1	01	2000
2	2	0203	2000
2	3	0206	1000
3	2	0104	1000
3	3	0402	2000
4	3	01020304	4500
5	2	010203	2000
5	3	0304	2800
6	1	0102	2000
6	3	0104	500
7	1	03	4000
8	2	010203	2000
8	3	02	3000
9	3	01	2500
10	2	01020304	2000
11	2	01	1500
11	3	03	1500
12	2	01	500

Here additional field 'ocpcode' (occupation code) gives information about occupation of household. Occupation code 1, 2, 3 represents agriculture, wage and business respectively. Format of Data-3 is wide type of data, while format of Data-4 is called long type of data. We can see from both layouts that in wide type of data more columns are needed while in long type of data more records are needed to accommodate same information. Both layouts have their benefits. In Data-3 we can identify that households very easily, which have more income from wage than agriculture, while it is not as straight forward in long type of data (Data-4) (may be possible through use of subscripts in STATA). In other way it is not possible to get average income by occupation in single table for Data-3 which is possible in Data-4. Many times for merging two files it is necessary to change layout of data. For

changing layouts of data commands like collapse or reshape are used in STATA. Many commands like table, which has replace option, may be used for changing layout of data.

Merging Vs. Appending Data Files

For practical purposes it is general practice to break total information in separate files. Since STATA can open only single data file at a time, at time of tabulation, it is required to bring different information from different files at single place. In different paradigm of databases, where many tables can be opened (directly or indirectly) in different work areas at the same time, same purpose is achieved through linking tables. In STATA, such type of work may be done by merge command. In STATA only two files can be merged with single merge command. In STATA, two files which are merged are called as master file (which is currently opened) and using file which resides in hard disk. After merging, all variables of using files are added to variables of master file. Hence it is suggested to keep only necessary variables and save this with different name and then merge with master file. Before merging, it is necessary to decide key fields (common in both file) through which records of both files can be linked. These fields should be such that, records of using files (not necessarily master file) should be uniquely identified on the basis of key field. Steps for merging two files can be summarised as follows:

(1) Decide which two files should be merged;

(2) Decide which fields are key fields;

(3) Decide which file should be 'master' file and which should be 'using' file. 'Using' file should be uniquely identified on key field(s);

(4) Keep necessary fields in 'using' file, sort on key fields and save with other name;

(5) Open 'master' file and sort it on key field(s);

(6) Use merge command.

Merge command generate a system variable with name '_merge' (if other name is not specified through _merge () option). This variable shows the link status of records of both files. For detail see help of 'merge'.

Concept of Blanks

In STATA blank is treated in very special manner. In most of the tabulation commands blank is not included unless it is specified by given option (if it is given) in command. A special care is needed in calculation of mean.

Data - 5

hhno	caste	land	size
1	2	.05	5
2	1	0	8
3	3	0	4
4	3	1.5	6
5	2	2.5	8
6	4	3	7
7	4	.05	6
8	1	.09	5
9	2	0	4
10	2	0	6

hhno	caste	land	size
11	2	.25	3
12	2	.25	2

Suppose we want to calculate average land size by caste. Two different result could be obtained if landless households are represented by '.' (blank) or '0'. In first case average size will represent average of those households who possess land while in other case it represent average of all households whether it possess or not any land. Before obtaining average it is better to check the summary variable and convert it to '.' or '0' according to need.

Free vs. fixed format

Although there are separate utilities for importing data from other type of databases, most common technique is convert data of other sources in text format (flat file). With most of the databases and spreadsheet such utilities are available. Import of such text file can be done through **'insheet'** or **'infile'** command (in recent versions of STATA, **import** command has been used for importing data). These two types of commands work with two different formats of text files - free format and fixed format. Free formats (also called as delimited files) are those files in which information contained in different columns (or fields) is separated by some separator like tab or comma. In the fixed formats, information is not separated by separator but is separated according to defined position in the record. In STATA, such type of definition (byte position for a variable) is given in form of dictionary or it can be directly used in command. Command **'infile'** or **'infix'** uses dictionary for importing data in STATA. STATA exports data only in fixed format through **'outfile'** or **'outsheet'** command.

Weight

Due to sampling design or due to some logical reason, it is some time needed to provide weight to each observation in database. Information contained in some variable may be used as weight or additional variable may be created with that type of information. For attaching weight to observations, variable which contains weight, is used directly. According to way through which these weights are used STATA provides four types of weights namely- fweight (frequency weight), aweight (analytical weight), pweight (inverse probability weight), iweight (importance weight). Most frequently used weight is fweight. For getting a feeling how weight effects results, for Data-5, use command *tab caste*. Result will be as followed:

caste	Freq.	Percent	Cum.
1	2	16.67	16.67
2	6	50.00	66.67
3	2	16.67	83.33
4	2	16.67	100.00
Total	12	100.00	

Now using the command

tab caste [w=size]

we get following result with additional line

(frequency weights assumed)

caste	Freq.	Percent	Cum.
1	13	20.31	20.31

2	28	43.75	64.06
3	10	15.63	79.69
4	13	20.31	100
Total	64	100	

Generate a new variable with

gen temp=1* size

table caste, c(sum temp) row

caste	sum temp
1	13
2	28
3	10
4	13
Total	64

Comparing this table with just above it, we get justification of 'fweight'. 'fweight' create duplicate records (hypothetically) those many number of times given by weight variable (here size). Thus for caste 1 there are two observations related with house no 2 and 8 with weight 8 and 5. Hence on the basis of 8 duplicate records of house 2 and 5 duplicate records of house no 8 total numbers of observations may be obtained 13. Similarly by using command table caste [w=size], c(mean land) format(%5.2f) row

(frequency weights assumed)

caste	mean land
1	0.03
2	0.77
3	0.9
4	1.64
Total	0.82

Using the above mentioned logic of duplicate records, we can check average land size for caste 1 is (0*8+.09*5)/13 (.03). For details of aweight, pweight and iweight see STATA manual.

Labels

Stata provides facility of attaching much auxiliary information with data. Although from Stata7.0 onwards, we can keep variable name up to 32 lengths and it is sufficient to fix informative name to variable. Apart from this we can attach a label to variable for providing more information about variable. Some of the commands use this label in preparation of table and makes table more readable. Another type of label may be attached to value of categorical data with the purpose of making it more readable. Such type of labels is called value label. All type of operation like defining labels, dropping labels; modifying labels are governed by **'label'** command.

Preparation of Report

(1) Locate the files in which necessary information (variables or columns are contained).

(2) Bring all necessary information in one file using merge and append (in STATA only one file can be opened at one time).

(3) Clean the logical inconsistencies of data contained in variables related with report.

(4) Divide the whole table into parts according to limits of STATA commands (**tabulate, table** or **tabstat**). Number of required commands (without 'for' command) depends upon columns used in report. If all columns of report belong to single variable then one tab (or table) may be sufficient. Sometimes it is possible to bring data of different variable in one variable though '**reshape**' command. In such case many **tab** (or **table**) command may be replaced by single tab (or **tabulate**) command.

(5) Check whether any reshaping (use of **reshape** or **collapse**) is needed for particular part. If yes, then save data as some temporary file and proceed further. This temporary file may be reopened if data before reshaping are needed.

(6) Generate variables for respective parts of report (if needed).

(7) Generate tables for respective part.

(8) Bring all tables in EXCEL sheet and prepare final report.

R Software for Statistical Analysis

Introduction

R is a free software environment for statistical computing and graphics. It is almost perfectly compatible with S-plus. The only thing you need to do is download the software from the internet and use an editor to write your program (e.g. Notepad). It contains most standard methods of statistics as well as lot of less commonly used methods and can be used for programming and to construct your own functions. It is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. It is available for down load from **http://www.r-project.org/**. The primary purpose of this chapter is to introduce R software for beginners.

To Download R Software

In any web browser (e.g. Microsoft Internet Explorer), go to: **http://www.r-project.org**. You will see a page like

→→

- Downloads: CRAN (On your left hand side you will see CRAN).
- Set your Mirror: Anyone in the **India or any other country** is fine.

- **On your right hand side you will see Download R for Windows. Click there**

- **Click on <u>base</u>**

- **Click on <u>R-3.2.0.exe</u> and save it to your hard disc.**

- **This is the latest available version of the software. It is an '.exe' file, which you can save in your hard disc. By double clicking on the name of this file, R is automatically installed. All you need to do is follow the installation process.**

To Open R Software

The installation process automatically creates a shortcut for R. Double click this icon to open the R environment. R will open up with the appearance of a standard Windows implementation (i.e. various windows and pull-down menus). Note that R is an interpreted language and processes commands on a line by line basis. Consequently it is necessary to hit **ENTER** after typing in (or pasting) a line of R code in order to get R to implement it.

To Run R Program Code

The main active window within the R environment is the **R Console**. This is a line editor and output viewer combined into one window. Here at the command **prompt** (the symbol >), we can enter R commands which run instantly upon pressing the carriage return key. This sign (>) is called prompt, since it prompts the user to write something, see below.

We can also run blocks of code which we have copied into the paste buffer from another source. In this session we shall use the Windows-supplied editor Notepad to display and edit our R program code. If we were to write some R of code, then simply copy it from the editor and paste it into the R Console, then the code would run in real time.

To Open the Editor

Here we are using the Windows-supplied editor **Notepad** to display and edit our R program code, although any general-purpose editor will suffice. Open Notepad by going to the Start button and clicking on:

Start > All Programs > Accessories > Notepad

Having opened Notepad, open the file, for example, **Intro_to_R.txt** (containing the program code, assume that it is copied in C: / derive) by selecting the following option from the pull-down menu:

File > Open

Click on the down-arrow at the top of the "Open" dialog box and change the selection to "Look in" C:\. You should now see the filename **Intro_to_R.txt** among a list of files. Double-click on the filename to open it.

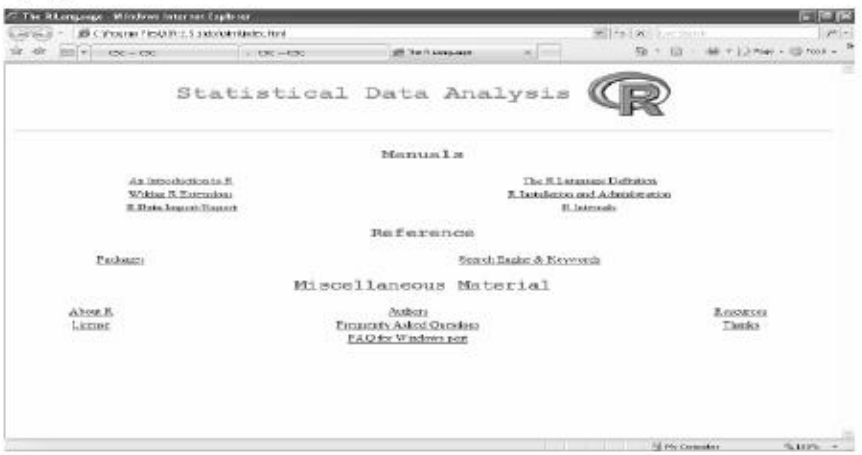
A Couple of Other Useful Things

- Please remember that R is **case-sensitive** so we need to be consistent in our use of lower and upper case letters, both for commands and for objects.
- When the program has finished, we should see the red command prompt (>) pop up in the R Console window. This indicates that control is returned to the user, so that you can now type more R commands if you wish.
- A comment in R code begins with a hash symbol (#). Whole lines may be commented or just the tail-end of a line. Examples are:

Help

Html-help can be invoked from the Help-menu. From the opening webpage, you can access manuals, frequently asked questions, references to help for individual packages, and most importantly, Search Engine. Help is the best place to find out new functions, and get descriptions on how to use them.

→→



Getting Started With R

Commands in R are given at the command prompt.

Simple Calculations, Vectors and Graphics

To begin with, we'll use R as a calculator. Try the following commands:

```
2+7
2/(3+5)
sqrt(9)+5^2
sin(pi/2)-log(exp(1))
```

Help about a specific command can be had by writing a question mark before the command, for instance:

```
?log
```

As an alternative, help can be used; in this case, help (log). The help files are a great resource and you will soon find yourself using them frequently.

Comments can be written using the #-symbol as follows:

```
2+3    # The answer should be 5
```

Vectors and Matrices

Vectors and matrices are of great importance in many numerical problems. To create a vector named my data and assign the values 7, −2, 5 to it, we write as follows:

```
mydata <- c(7,-2,5)
```

The symbol <- (or alternatively use =) should be read as "assigns". The command c can be interpreted (by you, the user) as column or combine. The second element of the vector can be referred to by the command

```
mydata[2]
```

and elements between 2 and 3 (i.e. elements 2 and 3) by

```
mydata[2:3]
```

Vectors can be manipulated, for instance by adding a constant to all elements, as follows.

```
myconst <- 100; mydata + myconst
```

Using the semicolon allows us to write multiple commands on a single line.

A vector x consisting of the integers between 1 and 10; 1, 2, . . . , 10; can be created by writing

```
x <- c(1:10)
```

Vectors with sequences of numbers with particular increments can be created with the seq command:

```
mydata1 <- seq(0,10,2)          # integers between 0 and 10, with increment 2
```

Read x and y

```
x<- c(2,3,1,5,4,6,5,7,6,8)
y<- c(10, 12, 14, 13, 34, 23, 12, 34, 25, 43)
```

Read two vectors

```
weight<- c(60, 72, 57,90)
height<-c(1.75, 1.80, 1.65, 1.90)
bmi<- weight/height^2           # Compute body mass index (BMI)
```

Functions on Vectors

```
length(x)                       #To compute length of data in x.
[1] 10
sum(x)                          #To compute sum of data in x.
[1] 47
sum(x^2)
[1] 265
mean(x)                         #To compute mean of data in x.
[1] 4.7
mean(y)
[1] 22
var(x)                          #To compute variance of x.
[1] 4.9
sqrt(var(x))                    # To compute standard deviation of x.
[1] 2.213594
sum((x-mean(x))^2)
[1] 44.1
sqrt(var(x))/mean(x)*100        #To compute coefficient of variation
```

To compute summary features of data in x

summary(x)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	3.25	5.00	4.70	6.00	8.00

To compute summary features of data in x^2

summary(x^2)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	10.75	25.00	26.50	36.00	64.00

Some Calculations

sum(weight)

mean(weight) or sum(weight)/ length(weight)

Denote by xbar= mean(weight) then

sqrt(sum((weight- xbar)^2))/ length(weight))

sd(weight)

cor(x,y) #To compute correlation coefficient between x and y.

var(x,y) #To compute covariance between x and y.

Slightly more complicated example …

The rule of thumb is that the BMI for a normal weight individual should be between 20 and 25, and we want to know if our data deviate systematically from that.

- We can use a one sample t test to assess whether the 6 persons' BMI can be assumed to have mean 22.5 given that they come from a normal distribution.
- We can use function t.test
- Although you might not be knowing about t test but example is just to give some indication of what real statistical output look like

t test (see ? t.test)

t.test (bmi, mu=22.5)

One Sample t-test

data: bmi

t = -0.5093, df = 3, p-value = 0.6456

alternative hypothesis: true mean is not equal to 22.5

95 percent confidence interval:18.29842 25.54231

sample estimates:

mean of x

21.92036

If mu is not given then t.test would use default mu=0

The p value is not small, indicating that it is not at all unlikely to get data like those observed if the mean were in fact 22.5

Packages

The base distribution already comes with some high priority add on packages, for example, boot, nlme, stats, grid, foreign, MASS, spatial etc. The packages included as default in base distribution implement standard statistical functionality, for example, linear models,

classical tests, a huge collection of high level plotting functions etc. Packages not included in the base distribution can be installed directly from R prompt.

Classical Tests

To load the library of classical tests statistics available with R software use library(stats)

#To get results of t-test for comparing population means of x and y when variances are not equal.

t.test(x,y)

To get results for usual t-test when variances are equal. If T is replaced by F then it is equal to t.test(x, y)

t.test(x,y,var.equal=T)

?t.test

library(stats)

x<- c(2,3,1,5,4,6,5,7,6,8)

y <- c(10, 12, 14, 13, 34, 23, 12, 34, 25, 43)

mean(x)

mean(y)

var(x,y)

cor(x,y)

t.test(x)

t.test(x,y)

t.test(x,y,var.equal=T)

var.test(x,y) #To compare variances of x and y.

The commands **rbind** and **cbind** can be used to merge row or column vectors to matrices. Try the following:

x <- c(1,2,3)

y <- c(4,5,6)

A = cbind(x,y)

B = rbind(x,y)

C = t(B)

The last command gives the matrix transpose of B. Now type A, B or C to see what the different matrices look like.

Simple Graphics

Graphics- one of the most important aspects of presentation and analysis of data is generation of proper graphics

- Graphic features of a data can be viewed very effectively using R-software
- Graphs of functions can be drawn by constructing suitable vectors and using the plot command.
- **plot**: both 1D and 2D plots (see ?plot)

Scatter plots: are useful for studying dependencies between variables. Try writing

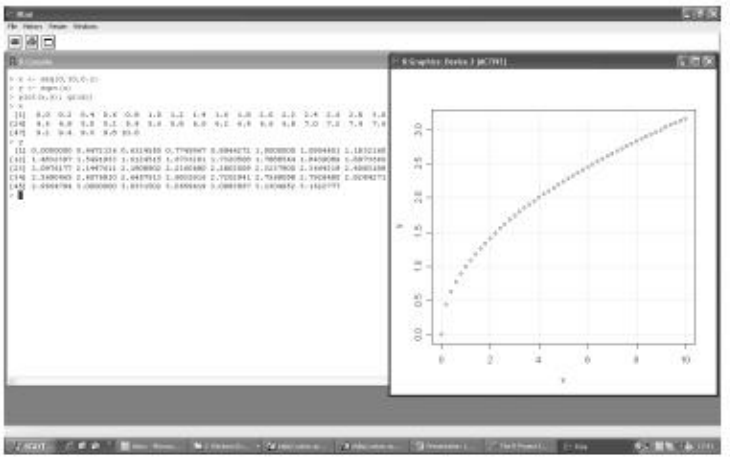
- Using the **plot command.**

x <- seq(0,10,0.2)

y <- sqrt(x)

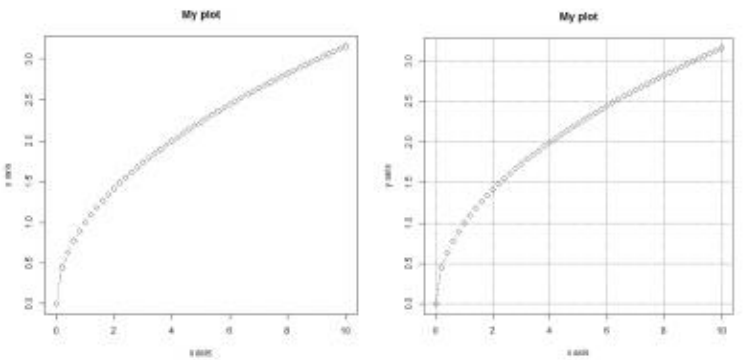
plot(x,y); grid()

- As one might guess, the last command adds a grid to the plot.

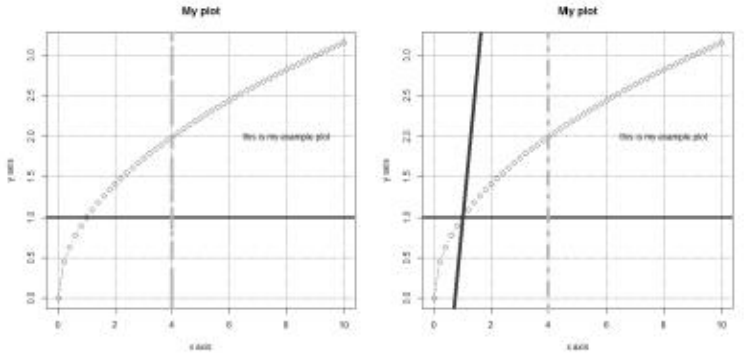


plot(x,y,type="b",col="blue",lwd=1,lty=4,pch=5, main="My plot", xlab="x axis", ylab="y axis")
grid(col="red")

→ →



plot(x,y,type="b",col="blue",lwd=1,lty=4,pch=5, main="My plot", xlab="x axis", ylab="y axis")
grid(col="red")
text(8,2,"this is my example plot")
abline(h=1,v=4, col=c("darkred","green"), lty=c(1,4),lwd=c(4,6))
reg.lm=lm(x~y)
abline(reg.lm, col="red",lwd=6) #To add the regression line



Save graphics by choosing File -> Save as

Bar Plot

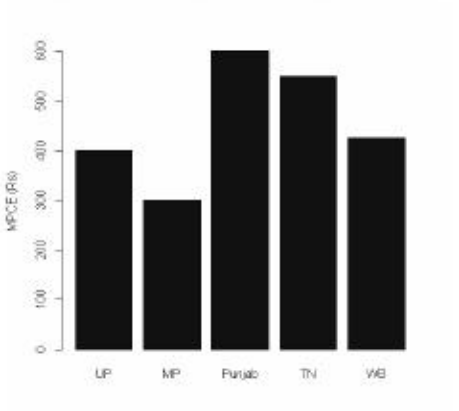
x1 <- c(400, 300,600,550,425)

Suppose data in x1 are average MPCE of some states whose names are to be assigned against their value. Following commands are required:

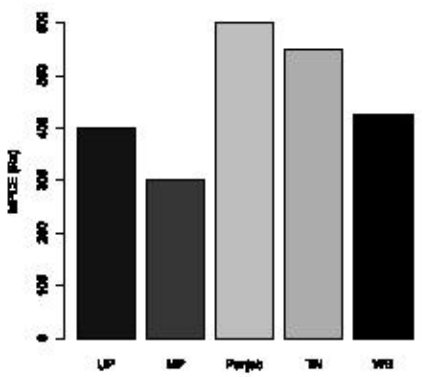
names(x1)<-c("UP","MP","Punjab","TN","WB")

To assign names of states. Double quotation mark " " means that names are characters not numeric.

barplot(x1, names=names(x1), ylab="MPCE (Rs)",col="blue")



barplot (x1,names = names (x1), ylab = "MPCE (Rs)", col = c ("blue", "red", "gray", "orange", "black"))

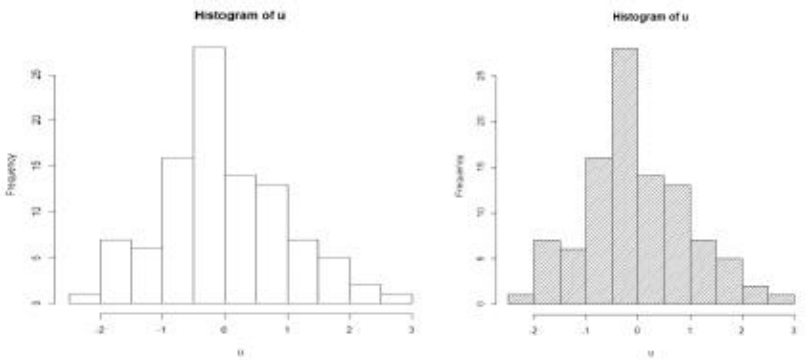


?barplot

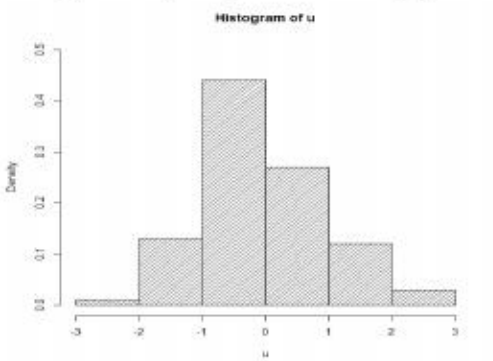
Histograms

A histogram can be used to study the distribution of continuous data. Unless they are explicitly stated, R chooses the numbers of classes and class width when the command hist is used.

```
#generate 100 random numbers from standard normal distribution
u<- rnorm(100)
hist(u)                          #default histogram
#with shading
hist(u, density=20)
```



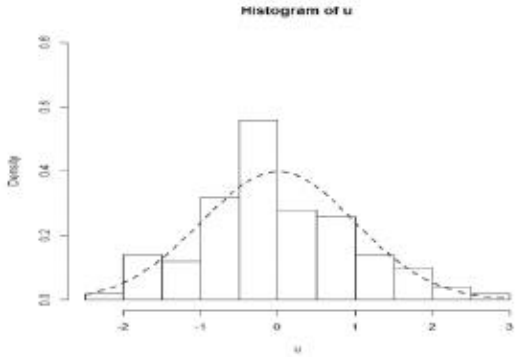

```
# proportion, instead of frequency also specifying y-axis

hist(u, density=20, breaks=-3:3, ylim=c(0,.5), prob=TRUE)
```



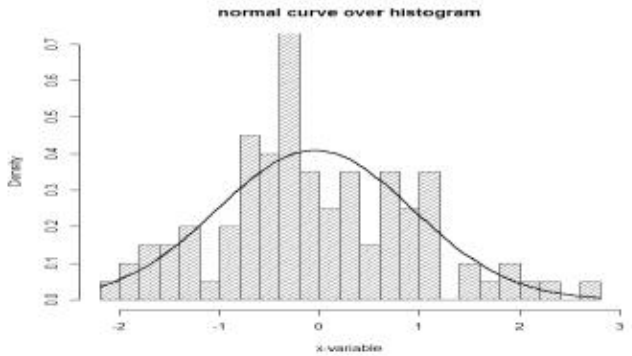

```
hist(u,freq=F,ylim = c(0,0.8))
curve(dnorm(x), col = 2, lty = 2, lwd = 2, add = TRUE)
```

The freq=F argument to hist ensures that the histogram is in terms of densities rather than absolute counts

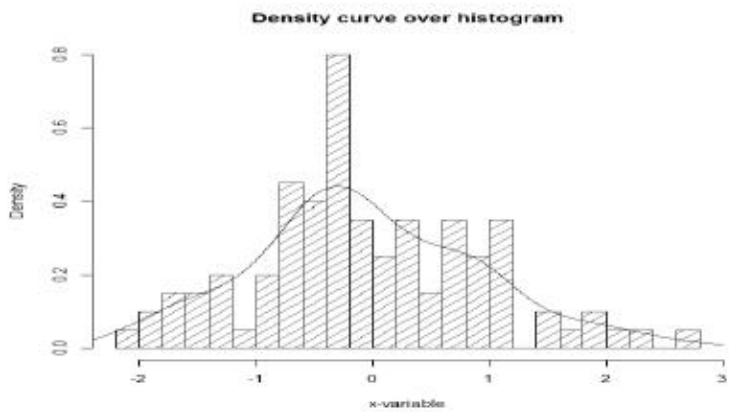
→ →



 # overlay normal curve with x-lab and ylim
 # colored normal curve
m<-mean(u) ;std<-sqrt(var(u))
hist(u, density=20, breaks=20, prob=TRUE, xlab="x-variable", col="red", ylim=c(0, 0.7), main="normal curve over histogram")
curve(dnorm(x, mean=m, sd=std), col="darkblue", lwd=2, add=TRUE)



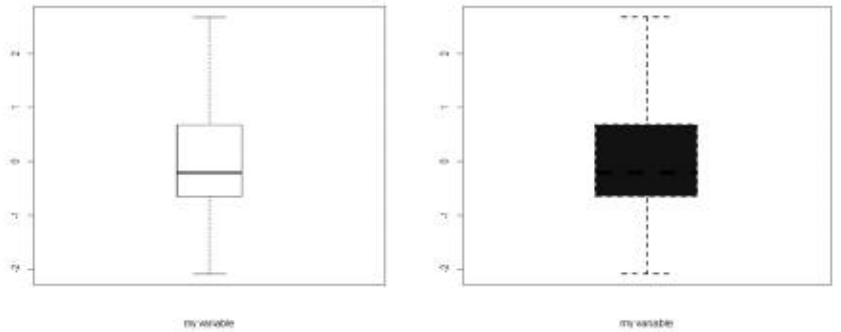
hist(u, density=10, breaks=20, col="red", prob=TRUE, xlab="x-variable", ylim=c(0,0.8),main="Density curve over histogram")
lines(density(u),col = "blue")
→ →



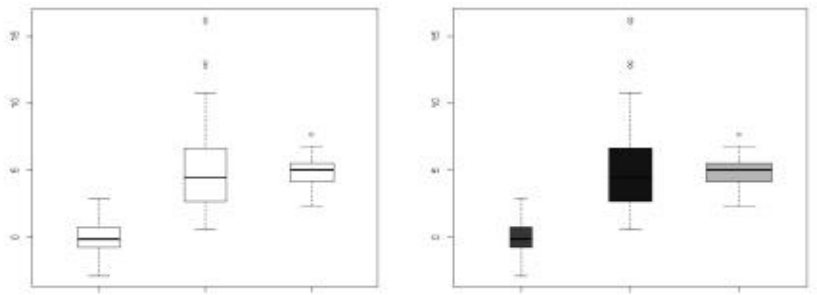
Box Plots

Box plots are also a useful tool for studying data. It shows the median, quartiles and possible outliers. The R command is **boxplot**, which we use on the same variables as the histogram.

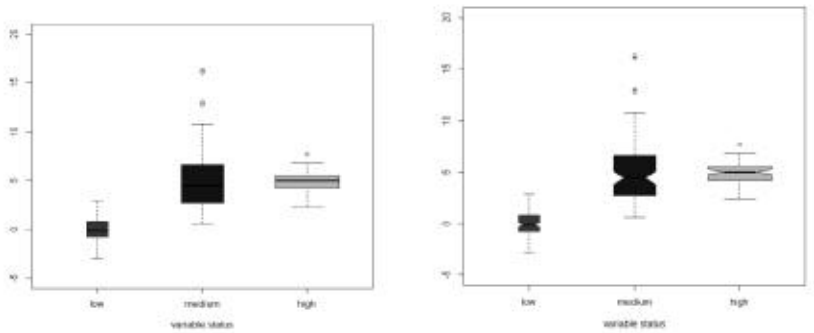
```
boxplot(u, xlab="my variable", boxwex=.4)
boxplot(u, xlab="my variable", boxwex=.6,col="blue", lty=2,lwd=2)
```



```
## we creat data: three variables
u1<- rnorm(100) # 100 random number from standard normal distribution
u2<- rchisq(100,5) # 100 random number from chisq distribution with mean 5
u3<- rnorm(100,5,1) # 100 random number from normal distribution with mean 5, sd 1

boxplot(u1,u2,u3, boxwex=.4)
boxplot(u1,u2,u3, boxwex=c(.2,.4,.6),col=c("red","blue","green"))
```



```
variablename<-c("low","medium", "high")
boxplot(u1,u2,u3,names=variablename,boxwex=c(.2,.4,.6), col=c("red","blue","green"),
ylim=c(-5, 20), xlab="variable status")
boxplot(u1,u2,u3,names=variablename,
boxwex=c(.2,.4,.6),col=c("red","blue","green"),ylim=c(-5, 20),xlab="variable status", notch
= TRUE)
```



try

boxplot(u, xlab="my variable", pars = list(boxwex = 0.5, staplewex = .5, outwex = 0.5),plot = F)

boxplot(u, xlab="my variable", pars = list(boxwex = 0.5, staplewex = .5, outwex = 0.5),plot = T)

7boxplot

Handling Data

Creating Data Frames

The command data.frame can be used to organize data of different kinds and to extract subsets of said data. Assume that we have data about three persons and that we store it as follows:

```
length <- c(180,175,190)
weight <- c(75,82,88);
name <- c("Anil","Ankit","Sunil")
friends <- data.frame(name,length,weight)
```

friends is now a data frame containing the data for the three persons. Data can easily be extracted:

```
my.names <- friends$name
length1 <- friends$length[1]
```

Reading Data from Files

It is common that data is stored in a text file and that we wish to import the data to R. We will study two cases; one with purely numerical data and one with numerical data with headers. In the file **coins.dat** data about the amount of silver in 27 silver coins from different epochs is stored. When we use the command

```
mynt1 <- read.table("coins.dat")
```

R creates a structure with two columns with headers V1 (amount of silver) and V2 (epoch). To see the properties of the tenth observations we write

```
mynt1[10,]                    # [10,] means "everything on row 10"
```

If we only want to see the amount of silver in the tenth coin we write

```
mynt1$V1[10] # vector V1 (the first vector), row 10
```

or

```
mynt1[10,1] # row 10, column 1
```

Clearly, it's important to know which data is stored in which column. To make this clear one can give the columns headers when importing the data:

```
mynt2 <- read.table("coins.dat",col.names=c("Silver","Epoch"))
```

so that the strings Silver and Epoch will be used instead of V1 and V2. Let us know study the same data set, but with the data stored in a slightly different way. The file coins.txt has the same data stored, but with headers stored on the first row of the file. To import the data with headers we write

```
mynt2 <- read.table ("coins.txt", header=TRUE)
```

Try to access information about particular observations as before!

There are functions for importing data from, for instance, databases or Excel spreadsheets as well, but these are more advanced and not covered in this lecture. Usually, however, one can copy the data from the database or the spreadsheet to a **text file** and then import it, so that the **read.table** command can be used.

```
# clean out the workspace
        rm(list=ls())
#List objects in workspace
        ls()
#File path is relative to working directory
#Get or Set Working Directory
getwd()
setwd()
# e.g. setwd("C:/Documents and Settings/Myfiles")
```

Writing data from files

```
x <- matrix(1:20,ncol=5)  # generate data in matrix form
write(x, "C:/ xm.txt")
write (x, " C:/ xm.csv")
write (x, "C:/ xm.csv",sep=",")
write.table(x," C:/ xm.xls",sep="\t")
```

Analysis of a Data Set

We will study a data set from the early 70's, with data about different cars (Cars data set). Load the data set by writing

```
data(mtcars)
```

it can read more about the data by looking at the help file:

```
?mtcars
```

mtcars package:datasets R Documentation

Motor Trend Car Road Tests

Description:

The data was extracted from the 1974 _Motor Trend_ US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Usage:

```
mtcars
```

Format:

A data frame with 32 observations on 11 variables.

[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (lb/1000)
[, 7] qsec 1/4 mile time
[, 8] vs V/S
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

Source:

Henderson and Velleman (1981), Building multiple regression models interactively. _Biometrics_, *37*, 391-411.

Examples

```
pairs(mtcars, main = "mtcars data")
coplot(mpg ~ disp | as.factor(cyl), data = mtcars, panel = panel.smooth, rows = 1)
```

Exercise. Answer the following questions using the help file:

1. How many cars are included in the data set?
2. Which years are the models from?
3. What does the mpg value describe?

To see the entire data set, simply write

```
mtcars
```

Exercise. To get familiar with the data set, answer the following non-statistical questions.

1. Are there any cars that weigh more than 5000 (lb/1000)?
2. How many cylinder has the motor of the Volvo 142E?
3. Are there any cars with 5 forward gears? Do they have automatic or manual transmisson?

Descriptive Statistics

Data can be summarized using simple measures such as mean, median, standard deviation, maximum and minimum and so on. A summary of a few such measures for the mtcars data set is obtained by writing

```
summary(mtcars)
```

Measures can also be studied one at a time:

```
mean(mtcars$hp); median(mtcars$hp); quantile(mtcars$wt); max(mtcars$mpg)
sd(mtcars$mpg)        # standard deviation
var(mtcars$mpg)       # variance
sd(mtcars$mpg)^2          # sd*sd=var?
```

The command attach is very useful when dealing with data frames. By writing attach(mtcars) the references to the variables in mtcars can be shortened; instead of the long references above we can write:

```
mean(hp); median(hp); quantile(wt); max(mpg)
```

The sequence of commands below plots two histograms in one window, the first being the histogram for mpg and the second for wt.

par(mfrow=c(1,2)); hist(mtcars$mpg); hist(mtcars$wt)

par(mfrow=c(a,b)) gives a rows with b plots on each row. Using the parameters freq when calling hist we can plot a histogram with relative frequencies instead of frequencies. Such histograms can be viewed as estimates of the density function of the data. Read in the help file about what hist(mtcars$mpg,freq=FALSE) means and then see for yourself by typing the command.

boxplot(mtcars$mpg); x11(); boxplot(mtcars$wt)

The x11 command opens a new window which the next figure will be plotted in.

plot(mtcars$wt,mtcars$mpg)

Does the slope of the cluster seem reasonable? The correlation (which measures linear dependence) can be calculated using the command cor (use to help file to see how). What is the correlation in this case? Does it agree with the slope?

cor(mtcars$wt,mtcars$mpg)

Linear regression

lm(mtcars$wt~mtcars$mpg)

Try to see help (lm)

Quitting R

R can be closed with the command q( ). After issuing the quit command, R asks whether to save the workspace or not:

It is usually a good idea to save the workspace, since this creates a special file that can be directly read into R, and one can commence working with the same datasets and results already generated without a need to start from the scratch again. Saved workspace is in a file called .RData, and all the commands given during the same R session are saved in a file called .Rhistory. To load the workspace into R again, one can simply double-click on the file .Rdata, and R should open automatically with all the data and results loaded. Note however that libraries are not loaded automatically, and these should be loaded (if needed) before commencing the work.

Strengths and Weaknesses of R

Strengths

- free and open source, supported by a strong user community
- highly extensible and flexible
- implementation of modern statistical methods
- moderately flexible graphics with intelligent defaults

Weaknesses

- slow or impossible with large data sets
- non-standard programming paradigms

REFERENCES

http://cran.r-project.org/doc/contrib/Short-refcard.pdf

http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/index.html

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org.

www.r-project.org

An Introduction to MATLAB

MATLAB is a programming language developed by MathWorks. It started out as a matrix programming language where linear algebra programming was simple. It can be run both under interactive sessions and as a batch job. MATLAB is an interpreted language for numerical computation. It allows one to perform numerical calculations, and visualize the results without the need for complicated and time consuming programming. It has statistics toolbox which is very useful. MATLAB allows its users to accurately solve problems, produce graphics easily and produce code efficiently. Because MATLAB is an interpreted language, it can be slow, and poor programming practices can make it unacceptably slow.

Introduction

MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Typical uses include:

- Math and computation
- Algorithm development
- Modeling, simulation, and prototyping
- Data analysis, exploration, and visualization
- Scientific and engineering graphics
- Application development, including Graphical User Interface building

MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. This allows you to solve many technical computing problems, especially those with matrix and vector formulations, in a fraction of the time it would take to write a program in a scalar noninteractive language such as C or Fortran.

The name MATLAB stands for matrix laboratory. MATLAB was originally written to provide easy access to matrix software developed by the LINPACK and EISPACK projects, which together represent the state-of-the-art in software for matrix computation.

MATLAB has evolved over a period of years with input from many users. In university environments, it is the standard instructional tool for introductory and advanced courses in mathematics, engineering, and science. In industry, MATLAB is the tool of choice for high-productivity research, development, and analysis.

MATLAB features a family of application-specific solutions called toolboxes. Very important to most users of MATLAB, toolboxes allow you to learn and apply specialized technology. Toolboxes are comprehensive collections of MATLAB functions (M-files) that extend the MATLAB environment to solve particular classes of problems. Areas in which toolboxes are available include signal processing, control systems, neural networks, fuzzy logic, wavelets, simulation, and many others.

The MATLAB System

The MATLAB system consists of five main parts:

The MATLAB language.

This is a high-level matrix/array language with control flow statements, functions, data structures, input/output, and object-oriented programming features. It allows both

"programming in the small" to rapidly create quick and dirty throw-away programs, and "programming in the large" to create complete large and complex application programs.

The MATLAB working environment.

This is the set of tools and facilities that you work with as the MATLAB user or programmer. It includes facilities for managing the variables in your workspace and importing and exporting data. It also includes tools for developing, managing, debugging, and profiling M-files, MATLAB's applications.

Handle Graphics.

This is the MATLAB graphics system. It includes high-level commands for two-dimensional and three-dimensional data visualization, image processing, animation, and presentation graphics. It also includes low-level commands that allow you to fully customize the appearance of graphics as well as to build complete Graphical User Interfaces on your MATLAB applications.

The MATLAB mathematical function library.

This is a vast collection of computational algorithms ranging from elementary functions like sum, sine, cosine, and complex arithmetic, to more sophisticated functions like matrix inverse, matrix eigenvalues, Bessel functions, and fast Fourier transforms.

The MATLAB Application Program Interface (API).

This is a library that allows you to write C and Fortran programs that interact with MATLAB. It include facilities for calling routines from MATLAB (dynamic linking), calling MATLAB as a computational engine, and for reading and writing MAT-files.

Basic Operations

Assignment to variable

```
>> x=3
x =  3
>> y=7
y =  7
```

Addition of variables

```
>> z=x+y
z = 10
>> z
z = 10
```

Subtraction of variables

```
>> y-x
ans =  4
```

Multiplication and Power of variables

```
>> x*y
ans =  21
>> x^2
ans =  9
>> y^2
ans =  49
>> y^x
ans = 343
```

Creating matrices

The basic data element in MATLAB is a matrix. A scalar in MATLAB is a 1x1 matrix, and a vector is a 1xn (or nx1) matrix.

Example: Create a 3x3 matrix A that has 1's in the first row, 2's in the second row, and 3's in the third row:

```
>> A = [1 1 1; 2 2 2; 3 3 3]
```

The semicolon is used here to separate rows in the matrix. MATLAB gives you:

```
A=
   1   1   1
   2   2   2
   3   3   3
```

If you don't want MATLAB to display the result of a command, put a semicolon at the end:

```
>> A = [1 1 1; 2 2 2; 3 3 3];  Matrix A has been created but MATLAB doesn't display it.
```

The semicolon is necessary when you're running long scripts and don't want everything written out to the screen! Suppose you want to access a particular element of matrix A:

```
>> A(1,2)
ans = 1
```

Suppose you want to access a particular row of A:

```
>> A(2,:)
ans = 2  2  2
```

MATLAB has several built-in matrices that can be useful. For example, zeros(n,n) makes an nxn matrix of zeros.

```
>> B = zeros(2,2)
B =
   0   0
   0   0
```

A few other useful matrices are:

zeros – create a matrix of zeros
ones – create a matrix of ones
rand – create a matrix of random numbers
eye – create an identity matrix

Matrix operations An important thing to remember is that since MATLAB is matrix-based, the multiplication operator "*" denotes matrix multiplication. Therefore, A*B is not the same as multiplying each of the elements of A times the elements of B. However, you'll probably find that at some point you want to do element-wise operations (array operations). In MATLAB you denote an array operator by playing a period in front of the operator. The difference between "*" and ".*" is demonstrated in this example:

```
>> A = [1 1 1; 2 2 2; 3 3 3];
B = ones(3,3);
A*B
ans =
   3   3   3
   6   6   6
```

```
   9   9   9
>> A.*B

ans =
    1   1   1
   2   2   2
   3   3   3
```

Other than the bit about matrix vs. array multiplication, the basic arithmetic operators in MATLAB work pretty much as you'd expect. You can add (+), subtract (-), multiply (*), divide (/), and raise to some power (^).

MATLAB provides many useful functions for working with matrices. It also has many scalar functions that will work element-wise on matrices (e.g., the function sqrt(x) will take the square root of each element of the matrix x). Below is a brief list of useful functions. You'll find many, many more in the MATLAB help index, and also in the "Other Resources" listed at the end of this handout.

Useful matrix functions

A' – transpose of matrix A. Also transpose(A).
det(A) – determinant of A
eig(A) – eigenvalues and eigenvectors
inv(A) – inverse of A
svd(A) – singular value decomposition
norm(A) – matrix or vector norm
find(A) – find indices of elements that are nonzero. Can also pass an expression to this function, e.g. find(A > 1) finds the indices of elements of A greater than 1.

A few useful math functions:

sqrt(x) – square root
sin(x) – sine function. See also cos(x), tan(x), etc.
exp(x) – exponential
log(x) – natural log log10(x) – common log
abs(x) – absolute value
mod(x) – modulus
factorial(x) – factorial function
floor(x) – round down. See also ceil(x), round(x).
min(x) – minimum elements of an array. See also max(x).
besselj(x) – Bessel functions of first kind

MATLAB also has a few built-in constants, such as pi (π) and i (imaginary number).

Symbolic math

Although MATLAB is primarily used for numerical computations, you can also do symbolic math with MATLAB. Symbolic variables are created using the command "sym."

```
>> x = sym('x');
```

Here we have created the symbolic variable x. If it seems kind of lame to you to have to type in all this just to create "x", you're in luck—MATLAB provides a shortcut.

```
>> syms x
```

This is a shortcut for x = sym('x').

Symbolic variables can be used for solving algebraic equations. For example, suppose we want to solve the equation "x^4 + 3*x^2 + 3 = 5":

```
>> y = solve('x^4 + 3*x^2 + 3 = 5',x)
y =
 (- 17^(1/2)/2 - 3/2)^(1/2)
  (17^(1/2)/2 - 3/2)^(1/2)
-(- 17^(1/2)/2 - 3/2)^(1/2)
  -(17^(1/2)/2 - 3/2)^(1/2)
```

Descriptive Statistics

MATLAB provides a number of commands that you can use to perform basic statistics tasks. When working with *descriptive statistics,* the math quantitatively describes the characteristics of a data collection, such as the largest and smallest values, the mean value of the items, and the average. This form of statistics is commonly used to summarize the data, thus making it easier to understand.

The following steps help you work through some of these tasks:

1. **Type w = 100 * rand(1, 100); and press Enter.**
   This command produces 100 pseudo-random numbers that are uniformly distributed between the values 0 and 1. The numbers are then multiplied by 100 to bring them up to the integer values used in Steps 4 and 5.
2. **Type x = 100 * randn(1, 100); and press Enter.**
   This command produces 100 pseudo-random numbers that are normally distributed. The numbers can be positive or negative, and multiplying by 100 doesn't necessarily ensure that the numbers are between –100 and 100 (as you see later in the procedure).
3. **Type y = randi(100, 1, 100); and press Enter.**
   This command produces 100 pseudo-random integers that are uniformly distributed between the values of 0 and 100.

Of course, you can interact with the vectors in other ways. For example, you can use standard **statistical functions** on them. Here is a list of the functions.

Function	Usage	Example
corrcoef()	Determines the correlation coefficients between members of a matrix.	corrcoef(AllVals)
cov()	Determines the covariance matrix for either a vector or a matrix.	cov(AllVals)
max()	Specifies the largest element in a vector. When working with a matrix, you see the largest element in each row.	max(w)
mean()	Calculates the average or mean value of a vector. When working with a matrix, you see the mean for each row.	mean(w)

median()	Calculates the median value of a vector. When working with a matrix, you see the median for each row.	median(w)
min()	Specifies the smallest element in a vector. When working with a matrix, you see the smallest element in each row.	min(w)
mode()	Determines the most frequent value in a vector. When working with a matrix, you see the most frequent value for each row.	mode(w)
std()	Calculates the standard deviation for a vector. When working with a matrix, you see the standard deviation for each row.	std(w)
var()	Determines the variance of a vector. When working with a matrix, you see the variance for each row.	var(w)

Example 1 — Calculating Maximum, Mean, and Standard Deviation

This example shows how to use MATLAB functions to calculate the maximum, mean, and standard deviation values for a 24-by-3 matrix called count. MATLAB computes these statistics independently for each column in the matrix.

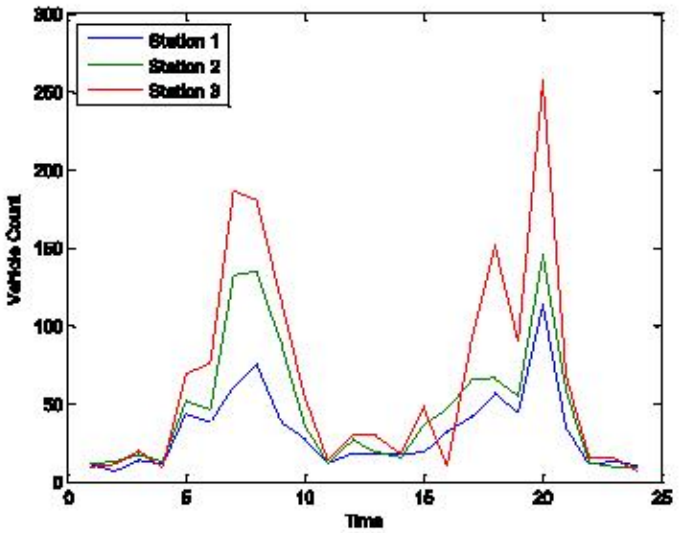
```
>> % Load the sample data
load count.dat
% Find the maximum value in each column
mx = max(count)
% Calculate the mean of each column
mu = mean(count)
% Calculate the standard deviation of each column
sigma = std(count)
mx =
  114   145   257
mu =
  32.0000   46.5417   65.5833
sigma =
  25.3703   41.4057   68.0281
```

Calculating and Plotting Descriptive Statistics

1. Load and plot the data:

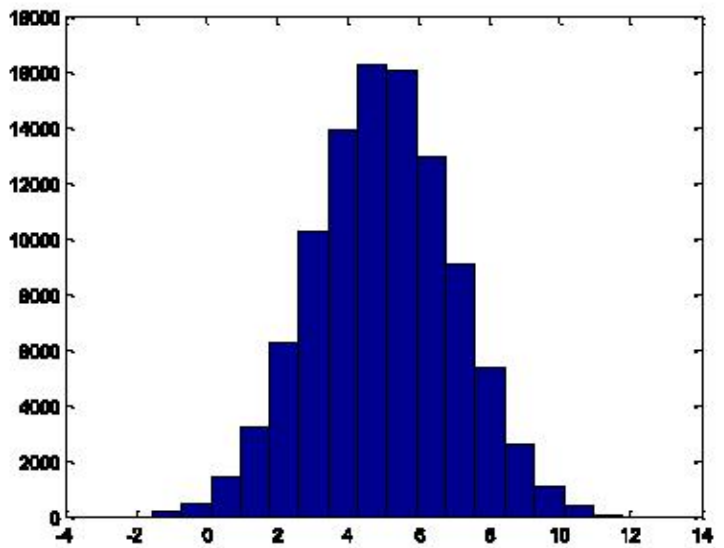
```
>> load count.dat
[n,p] = size(count);
% Define the x-values
t = 1:n;
% Plot the data and annotate the graph
plot(t,count)
legend('Station 1','Station 2','Station 3','Location','northwest')
```

xlabel('Time')
ylabel('Vehicle Count')



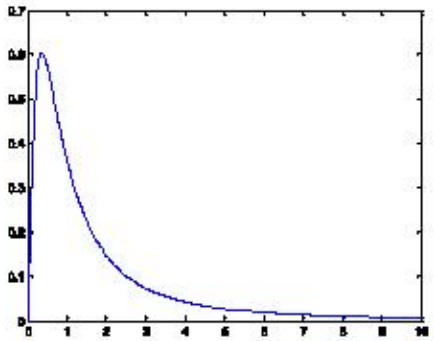
Ex.

```
>> k=100000;
sigma =2;
mu=5;
y = normrnd( mu , sigma , k , 1 );
hist( y , 20 );
```



Example and Plot. The most common application of the F distribution is in standard tests of hypotheses in analysis of variance and regression. The plot shows that the F distribution exists on the positive real numbers and is skewed to the right.

```
x = 0:0.01:10;
y = fpdf(x,5,3);
plot(x,y)
```



Example and Plot. The following commands generate a plot of the noncentral F pdf.

```
x = (0.01:0.1:10.01)';
p1 = ncfpdf(x,5,20,10);
p = fpdf(x,5,20);
plot(x,p,'- -',x,p1,'-')
```

Correlation Coefficients

Compute the correlation coefficients for a matrix with two normally distributed, random columns and one column that is defined in terms of another. Since the third column of A is a multiple of the second, these two variables are directly correlated, thus the correlation coefficient in the (2,3) and (3,2) entries of R is 1.

Ex.

```
>> x = randn(6,1);
y = randn(6,1);
A = [x y 2*y+3];
R = corrcoef(A)
R =
   1.0000   0.4973   0.4973
   0.4973   1.0000   1.0000
   0.4973   1.0000   1.0000
```

Compute the correlation coefficient matrix between two normally distributed, random vectors of 10 observations each.

```
>> A = randn(10,1);
B = randn(10,1);
R = corrcoef(A,B)
R =
   1.0000   0.3907
   0.3907   1.0000
```

Regression

Linear regression

Syntax

```
[r,m,b] = regression(t,y)
[r,m,b] = regression(t,y,'one')
```

Description

[r,m,b] = regression(t,y) takes these arguments,

t Target matrix or cell array data with a total of N matrix rows

y Output matrix or cell array data of the same size

and returns these outputs,

r Regression values for each of the N matrix rows

m Slope of regression fit for each of the N matrix rows

b Offset of regression fit for each of the N matrix rows

[r,m,b] = regression(t,y,'one') combines all matrix rows before regressing, and returns single scalar regression, slope, and offset values.

Examples

Train a feedforward network, then calculate and plot the regression between its targets and outputs.

```
>> [x,t] = simplefit_dataset;
net = feedforwardnet(20);
net = train(net,x,t);
y = net(x);
[r,m,b] = regression(t,y)
plotregression(t,y)
```

Command Summary

The command

```
>> help
```

will give a list of categories for which help is available (e.g.matlab/generalcovers the topics listed in Table 3. Further information regarding the commands listed in this section may then be obtained by using:

```
>> help topic
```

try, for example,

```
>> help help
```

Managing commands and functions.	
help	On-line documentation.
doc	Load hypertext documentation.
what	Directory listing of M-, MAT-and MEX-Files.
type	List M-File.
lookfor	Keyword search through the
demo	Run demos.
Working with les and the operating system.	
cd	Change current working directory.
dir	Directory listing.
delete	Delete File.
!	Execute operating system command.
unix	Execute operating system command & return result.

diary	Save text of MATLAB session.
Controlling the command window.	
cedit	Set command line edit/recall facility parameters.
clc	Clear command window.
home	Send cursor home.
format	Set output format.
echo	Echo commands inside script files.
more	Control paged output in command window.
Quitting from MATLAB.	
quit	Terminate MATLAB.
Matrix analysis.	
cond	Matrix condition number.
norm	Matrix or vector norm.
rcond	LINPACK reciprocal condition estimator.
rank	Number of linearly independent rows or columns.
det	Determinant.
trace	Sum of diagonal elements.
null	Null space.
orth	Orthogonalization.
rref	Reduced row echelon form.
Linear equations.	
nand/	Linear equation solution; use \help slash".
chol	Cholesky factorization.
lu	Factors from Gaussian elimination.
inv	Matrix inverse.
qr	Orthogonal- triangular decomposition.
qrdelete	Delete a column from the QR factorization.
qrinsert	Insert a column in the QR factorization.
nnls	Non{negative least- squares.
pinv	Pseudoinverse.
lscov	Least squares in the presence of known covariance.
Eigenvalues and singular values.	
eig	Eigenvalnes and eigenvectors.
poly	Characteristic polynomial.
polyeig	Polynomial eigenvalue problem.
hess	Hessenberg form.
qz	Generalized eigenvalues.
rsf2csf	Real block diagonal form to complex diagonal form.
cdf2rdf	Complex diagonal form to real block diagonal form.
schur	Schur decomposition.
balance	Diagonal scaling to improve eigenvalue accuracy.
svd	Singular value decomposition.
Matrix functions.	

expm	Matrix exponential.
expm1	M- File implementation of expm.
expm2	Matrix exponential via Taylor se-ries.
expm3	Matrix exponential via eigenval-ues and eigenvectors.
logm	Matrix logarithm.
sqrtm	Matrix square root.
funm	Evaluate general matrix function.
Graphics & plotting.	
figure	Create Figure (graph window).
clf	Clear current gure.
close	Close gure.
subplot	Create axes in tiled positions.
Axis	Control axis scaling and appear-ance.
hold	Hold current graph.
figure	Create gure window.
text	Create text.
print	Save graph to le.
plot	Linear plot.
loglog	Log-log scale plot.
semilogx	Semi-log scale plot
semilogy	Semi-log scale plot.
Specialized X-Y graphs.	
polar	Polar coordinate plot.
bar	Bar graph.
stem	Discrete sequence or "stem" plot.
stairs	Stairstep plot.
errorbar	Error bar plot.
hist	Histogram plot.
rose	Angle histogram plot.
compass	Compass plot.
feather	Feather plot.
fplot	Plot function.
comet	Comet-like trajectory.
Graph annotation.	
title	Graph title.
xlabel	X-axis label.
ylabel	Y-axis label.
text	Text annotation.
gtext	Mouse placement of text.
grid	Grid lines.
contour	Contour plot.
mesh	3-D mesh surface.
surf	3-D shaded surface.

waterfall	Waterfall plot.
view	3-D graph viewpoint specica-tion.
zlabel	Z-axis label for 3-D plots.
gtext	Mouse placement of text.
grid	Grid lines.

This chapter is an introduction to Matlab for people without much programming, mathematical or Unix background. Matlab is an easy software package to use even without much knowledge. That is what makes it so convenient. In this tutorial, some basic and very useful functions are described. These should get you started. Having understood a few basics, it will be pretty easy to expand your knowledge using Help, the internet and manuals.

References:

[1] M. Barnsley, Fractals Everywhere, Academic Press, Boston, 1993.

[2] D. C. Hanselman and B. Littlefield, Mastering MATLAB6, A Compre-hensive Tutorial and Reference, Prentice{Hall, Upper Saddle River, NJ, 2000.

[3] D. J. Higham and N. J. Higham, MATLAB Guide, SIAM, Philadelphia, 2000.

[4] N. J. Higham, Accuracy and Stability of Numerical Algorithms, SIAM, Philadelphia, 2002.

[5] J. Lagarias, The3x+ 1problem and its generalizations, American Mathemat-ical Monthly, 92 (1985), pp. 3{23.http://www.cecm.sfu.ca/organics/papers/lagarias}

[6] M. Overton, Numerical Computing with IEEE Floating Point Arithmetic, SIAM, Philadelphia, 2001.

[7] I. Peterson, Prime Spirals, Science News Online, 161 (2002). http://www.sciencenews.org/20020504/mathtrek.asp

[8] K. Sigmon and T. A. Davis, MATLAB Primer, Sixth Edition, Chapman and Hall/CRC, Boca Raton, FL, 2002.

[9] E. Weisstein, World of Mathematics, Prime Spiral, http://mathworld.wolfram.com/PrimeSpiral.html

[10] The MathWorks, Inc., Getting Started with MATLAB. http://www.mathworks.com/access/helpdesk/help/techdoc/learn_matlab/learn_matlab.shtml

[11] The MathWorks, Inc., List of Matlab-based books. http://www.mathworks.com/support/books/index.jsp

Hypothesis Testing in R

Hypothesis testing refers to the process of choosing between competing hypotheses about a probability distribution, based on observed data from the distribution. It is a core topic in mathematical statistics, and indeed is a fundamental part of the language of statistics. In this chapter, we study the basics of hypothesis testing, and explore hypothesis tests in some of the popular test using R.

Test 1: *t*-test for two populations means (Independent sample and variances unknown)

Object:To investigate the significance of the difference between the means of two populations.

Method: Consider two populations with means μ_1 and μ_2. Independent random samples of size

n_1 and n_2 are taken from which sample means $\bar{x}_1$ and $\bar{x}_2$ and variances

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 \text{ and } s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_i - \bar{x}_2)^2$$

The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

which may be compared with Student's *t*-distribution with degrees of freedom

$n_1 + n_2 - 2$

Limitations:

1. If the variances of the populations are known, a more powerful test is available: the *Z*-test for two population means.
2. The test is approximate if the populations are normally distributed or if the sample sizes are sufficiently large.
3. The test should only be used to test the hypothesis $\mu_1 = \mu_2$

Example: Below are given the gain in weights (in lbs.) ofpigs fed on two diets x and y.

Gain in weight

Diet x : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet y : 44,34,22,10,47,31,40,30,32,35,18,21,35,29,22

Test if the two diets differ significantly as regards their effect on increase inweight.

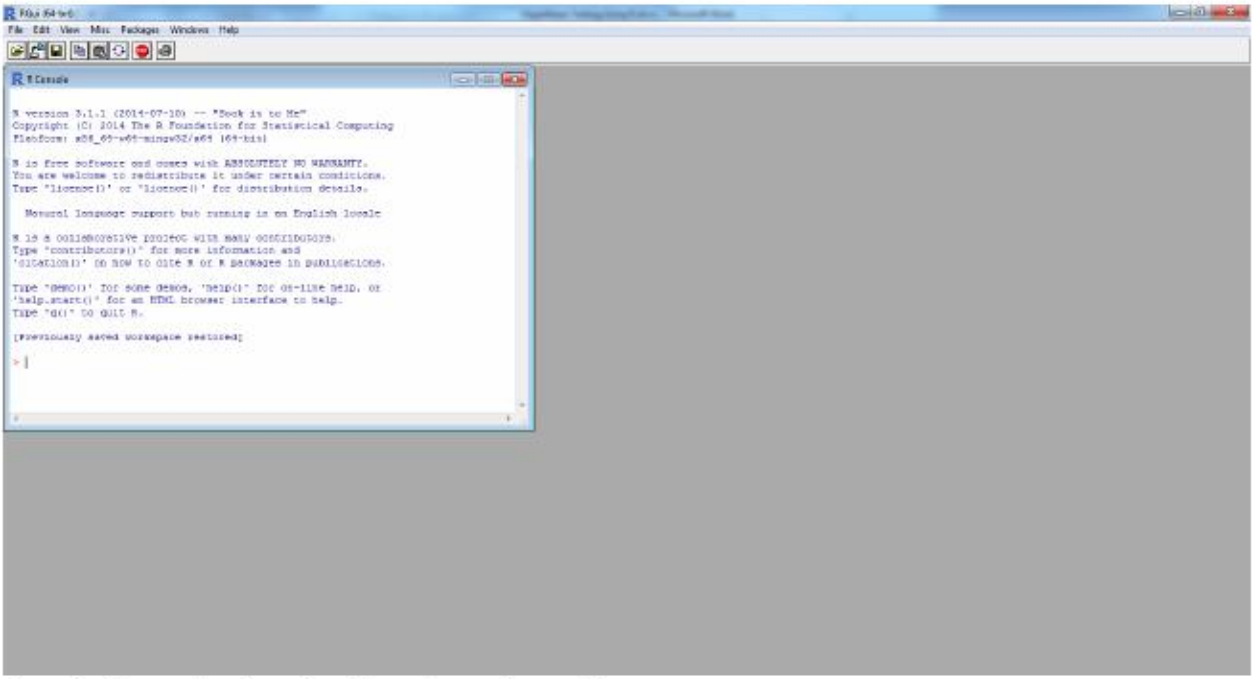
Solution:

Null hypothesis: Ho: $\mu_1 = \mu_2$ i.e. there is no significant difference between the mean increase in weight due to diets A and B.

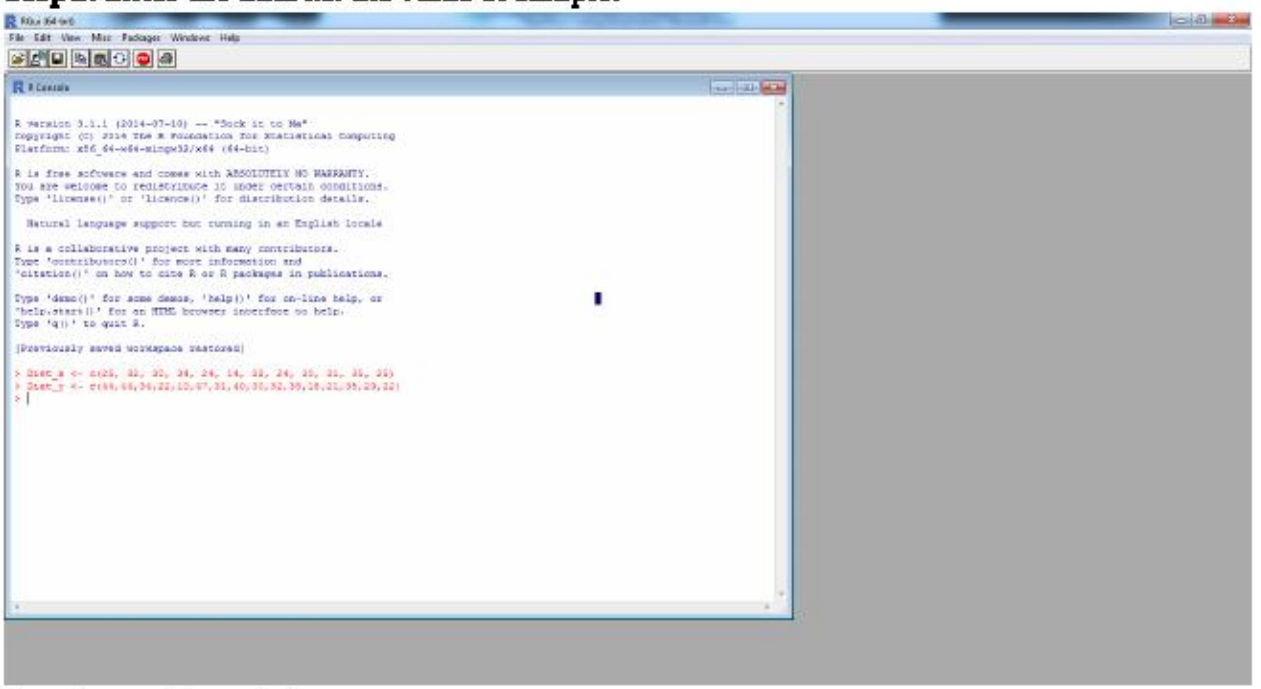
Alternative hypothesis: $\mu_1 \neq \mu_2$ (two tailed)

To apply t -test for the above, we proceed in the following steps:

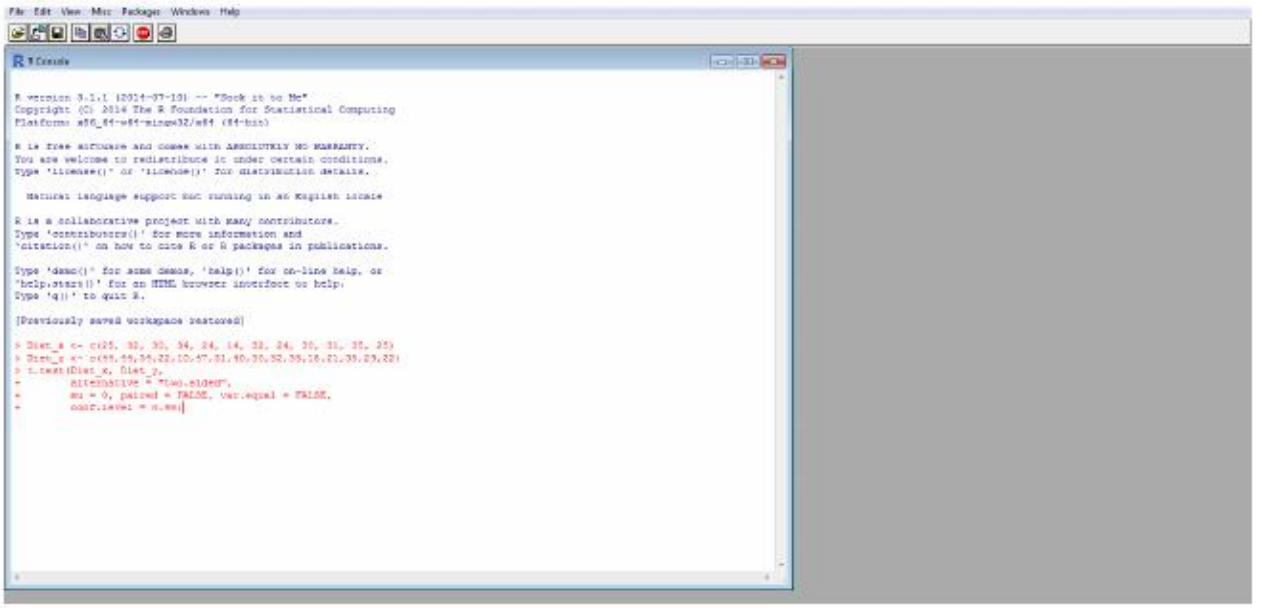
Step 1: First of all, open R; window will appear just as bellow:



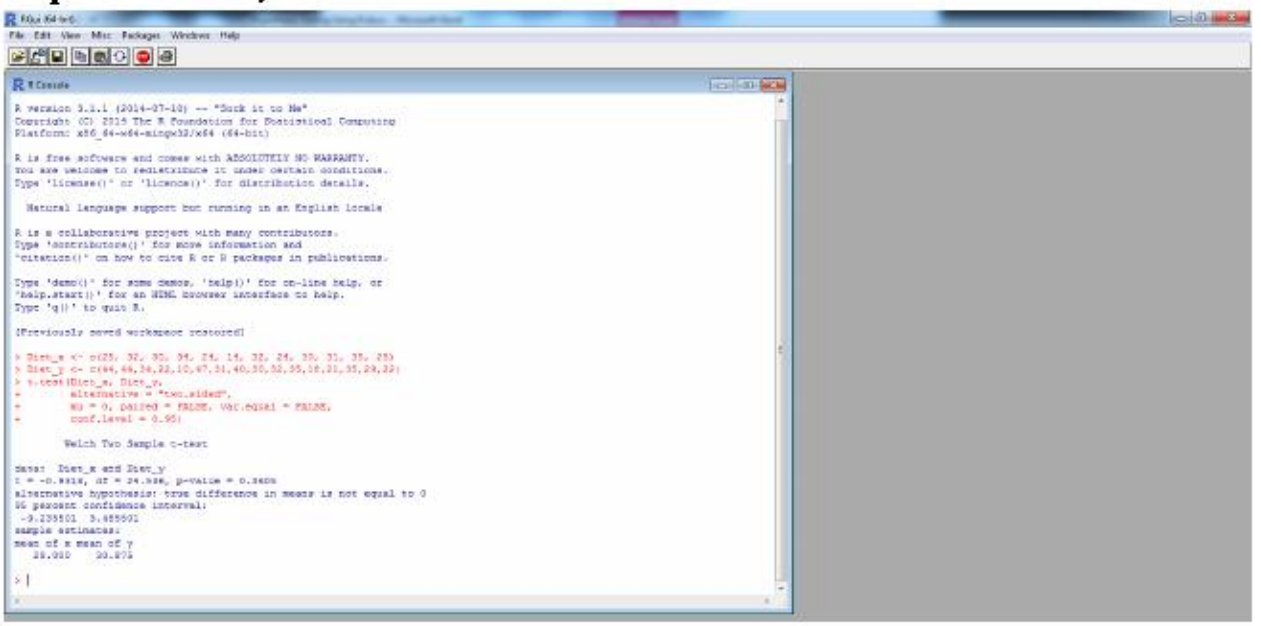
Step1: Enter the data i.e. the value of sample:



Step2:type the code for t-test as:



Step3: Press enter; the result will look like as



Description of output: Since, calculated value of mode t (0.93182)for two tailed test is less than 1.708141 (i.e. critical value for one tailed test at 5 percent level of significance). Same evidence is also found from p-value=0.3605>0.05. Hence, null hypothesis is accepted for two tailed test at 5% level of significance. Now, for one tailed test, one can simply change in the R-code as: type "less" or "greater" in place of two.sided.

##

R-Code:

```
Diet_x <- c(25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25)
Diet_y <- c(44,44,34,22,10,47,31,40,30,32,35,18,21,35,29,22)
t.test(Diet_x, Diet_y,
    alternative = "two.sided",
    mu = 0, paired = FALSE, var.equal = FALSE,
    conf.level = 0.95)
```

Output:

```
Welch Two Sample t-test
data:  Diet_x and Diet_y
t = -0.93182, df = 24.536, p-value = 0.3605
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.235501  3.485501
sample estimates:
mean of x mean of y
   28.000    30.875
```

##

Test 2: *t*-test for two populations mean (dependent sample) or Paired t-test:

Object: To investigate the significance of the difference between two population (dependent) means, μ_1 and μ_2. No assumption is made about the population variances.

Method: The differences $d_i (x_i - y_i)$ are formed for each pair of observations. If there are n such pairs of observations, we can calculate the variance of the differences by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2 \text{ where } \bar{d} = \bar{x}_1 - \bar{x}_2$$

Let the means of the samples from the two populations be denoted by $\bar{x}_1$ and $\bar{x}_2$ then the test statistic becomes

$$t = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{n-1}$$

which follows Student's t-distribution with $n - 1$ degrees of freedom. The test may be either one-tailed or two-tailed.

Limitations:

1. The observations for the two samples must be obtained in pairs. Apart from population differences, the observations in each pair should be carried out under identical, or almost identical, conditions.
2. The test is accurate if the populations are normally distributed.

Example: Below are given the gain in weights (in lbs.) of pigs fed on two diets x and y.

Gain in weight

Diet x : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet y : 44, 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18

Test if the two diets differ significantly if same set of 12 pigs were used in both the foods

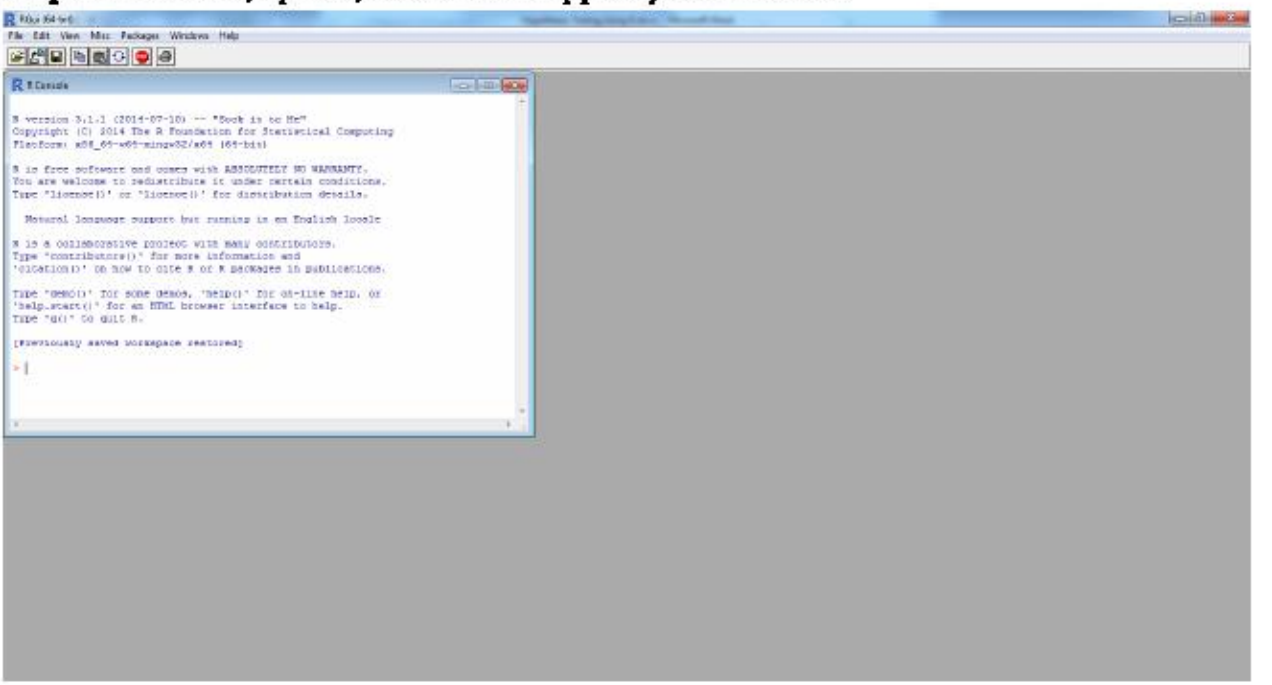
Solution:

Null hypothesis: Ho: $\mu_1 = \mu_2$ i.e. there is no significant difference between the mean increase in weight due to diets A and B.

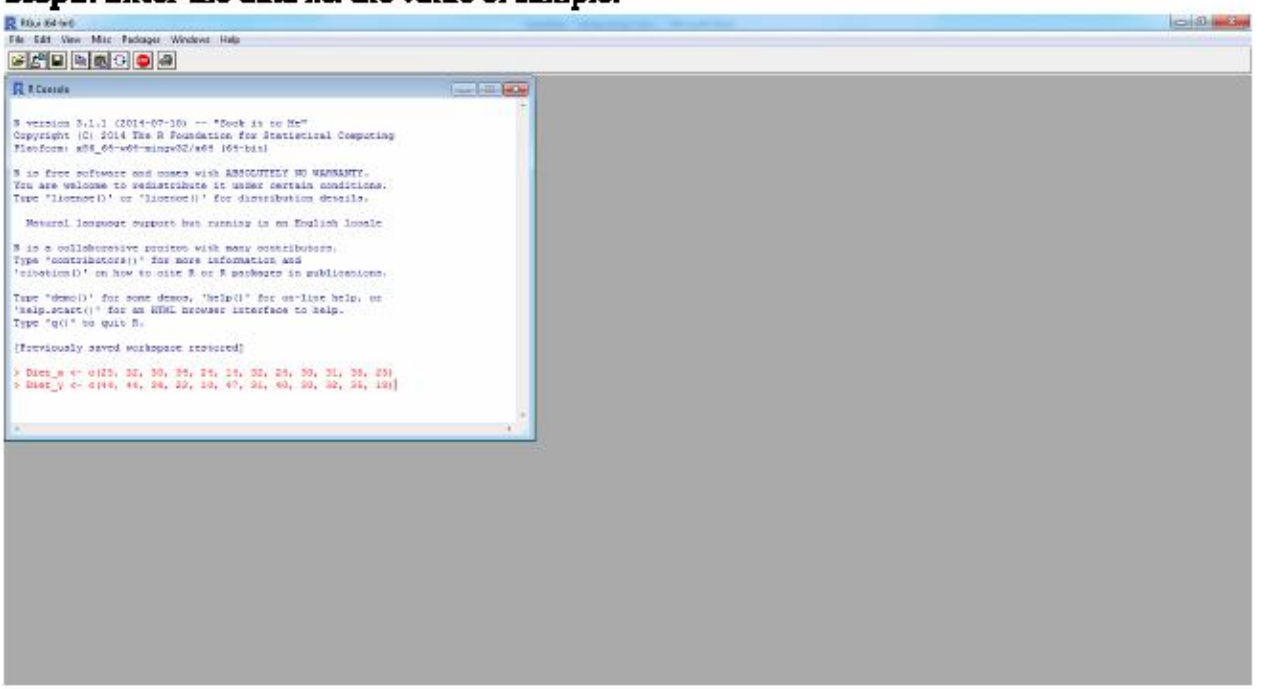
Alternative hypothesis: $\mu_1 \neq \mu_2$ (two tailed)

To apply t -test for the above, we proceed in the following steps:

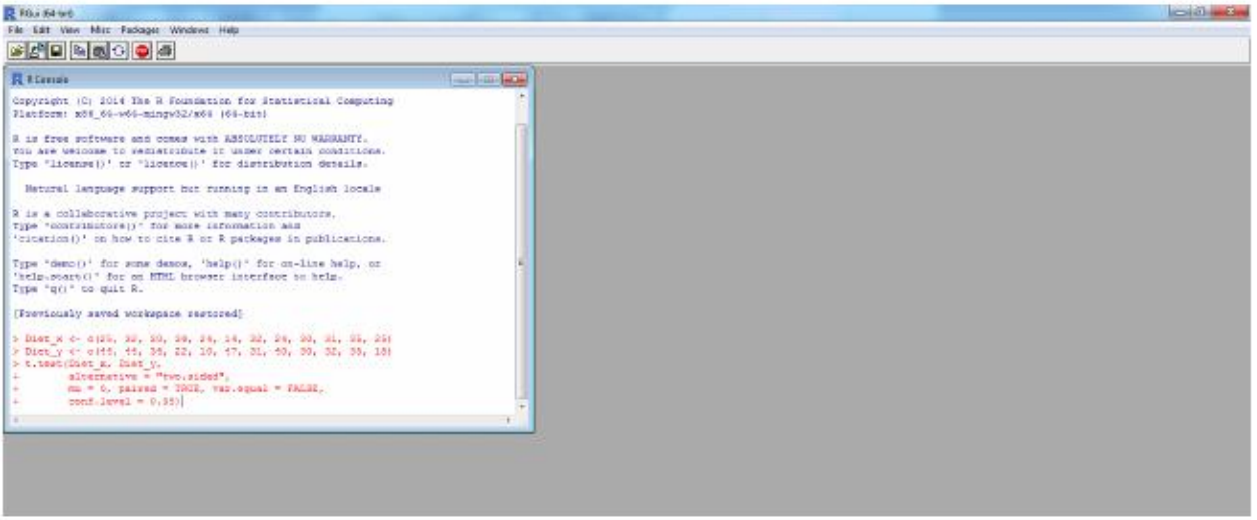
Step 1: First of all, open R; window will appear just as bellow:



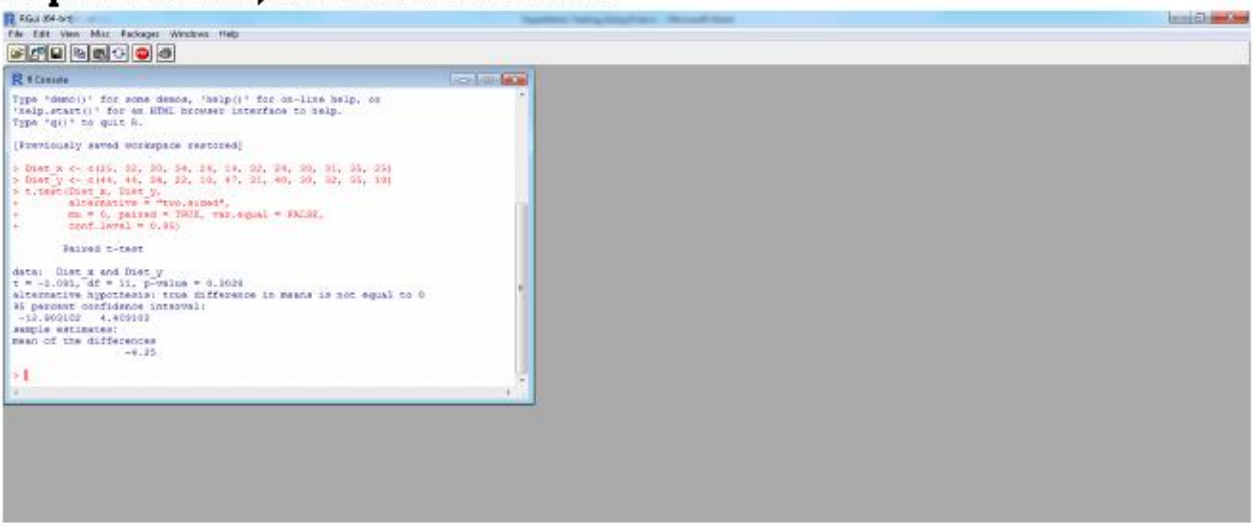
Step1: Enter the data i.e. the value of sample:



Step2:type the code for t-test as:



Step3: Press enter; the result will look like as



Description of output: Since, calculated value of mode t (1.081) for two tailed test is less than 2.2009 (i.e. critical value for one tailed test at 5 % level of significance). Same evidence is also found from p-value=0.3028>0.05. Hence, null hypothesis is accepted for two tailed test at 5% level of significance. . Now, for one tailed test, one can simply change in the R-code as: type "less" or "greater" in place of two.sided.

##

R-Code:

```
Diet_x <- c(25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25)
Diet_y <- c(44,44,34,22,10,47,31,40,30,32,35,18)
t.test(Diet_x, Diet_y,
    alternative = "two.sided",
    mu = 0, paired = TRUE, var.equal = FALSE,
    conf.level = 0.95)
```

Output:

```
Paired t-test
data:  Diet_x and Diet_y
t = -1.081, df = 11, p-value = 0.3028
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.903102   4.403102
sample estimates:
mean of the differences
             -4.25
```

##########

Test 3: *F*-test for two population variances (variance ratio test)

Object: To investigate the significance of the difference between two population variances.

Method: Given samples of size n_1 with values $(x_1, x_2, \ldots\ldots x_{n_1})$ and size n_2 with values $(y_1, y_2, \ldots\ldots y_{n_2})$ from the two populations, the values of

$$\bar{x}_1 = \frac{1}{n_1}\sum x_i \quad \bar{y}_2 = \frac{1}{n_2}\sum y_i \qquad s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2 \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2$$

can be calculated. Under the null hypothesis that the variances of the two populationsare equal the test statistic $F = s_1^2 / s_2^2$ follows the *F*-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom. The test may be either one-tailed or two-tailed.

Limitations: The two populations should both follow normal distributions. (It is not necessary that they should have the same means.)

Example: Two random samples are given as follows:

A: 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

B: 44, 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18

Test whether the samples come from the same normal population at 5% level of significance.

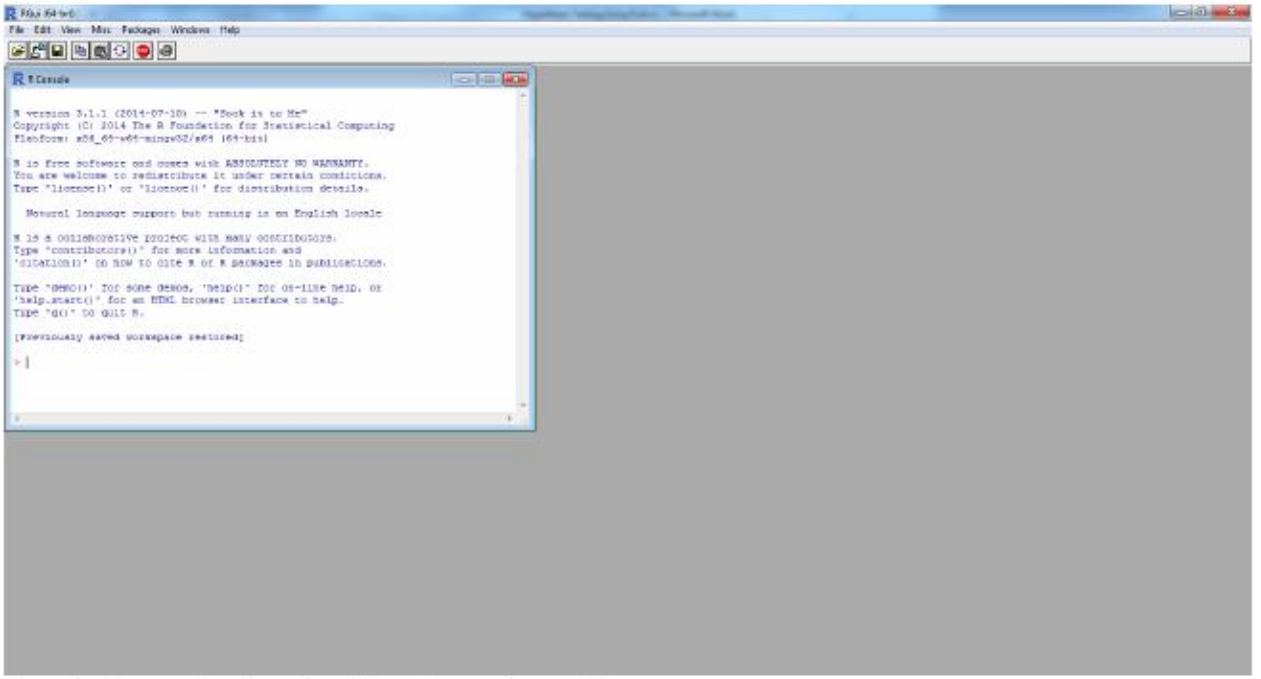
Solution:

Null hypothesis: Ho: $\sigma_1^2 = \sigma_2^2$

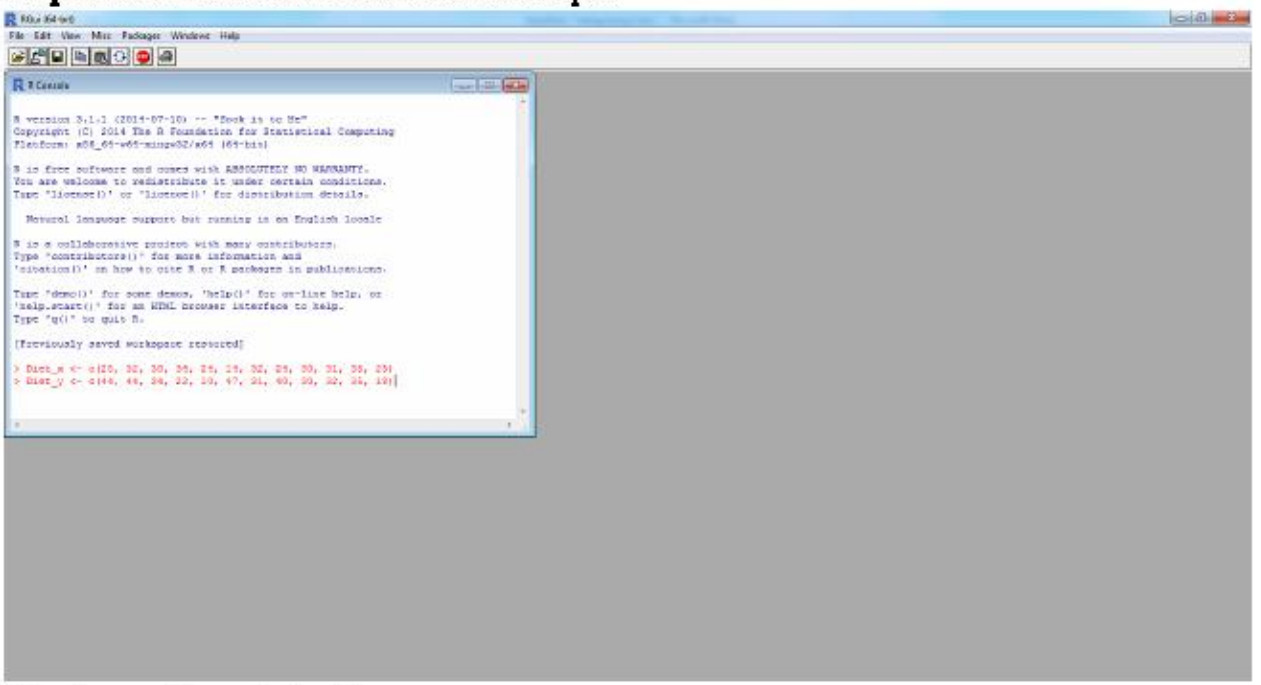
Alternative hypothesis: $\sigma_1^2 \neq \sigma_2^2$ (two tailed)

To apply F-test for the above, we proceed in the following steps:

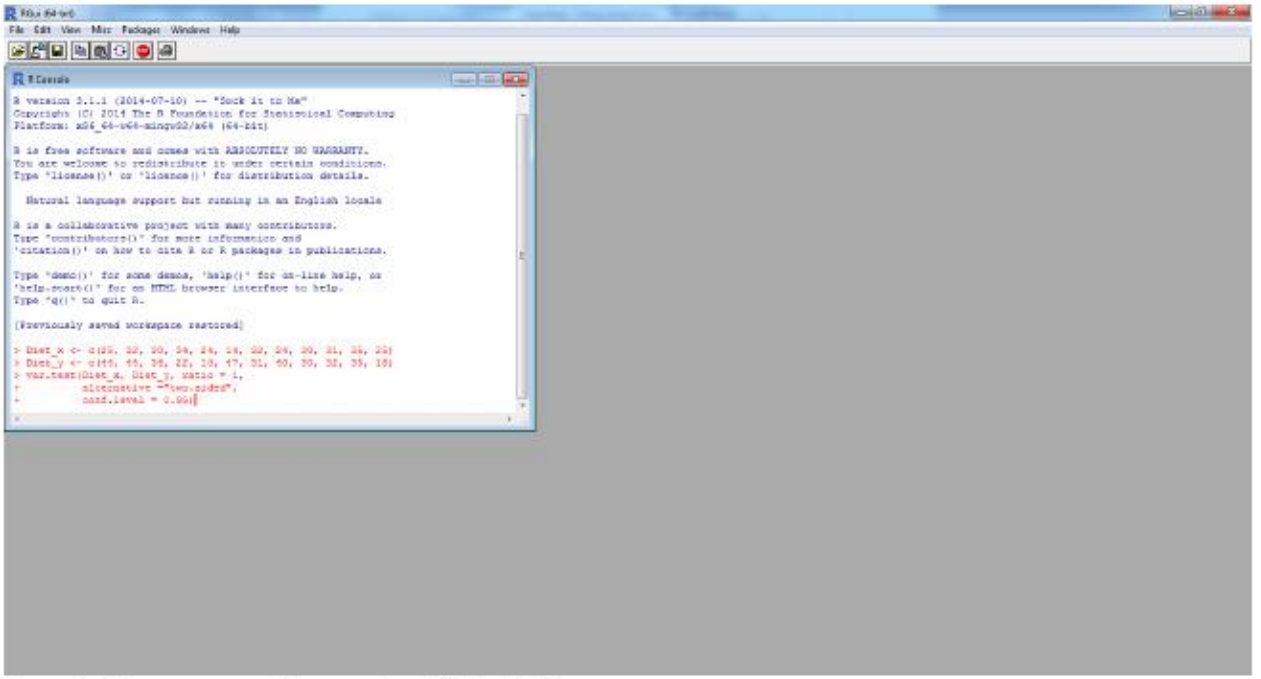
Step 1: First of all, open R; window will appear just as bellow:



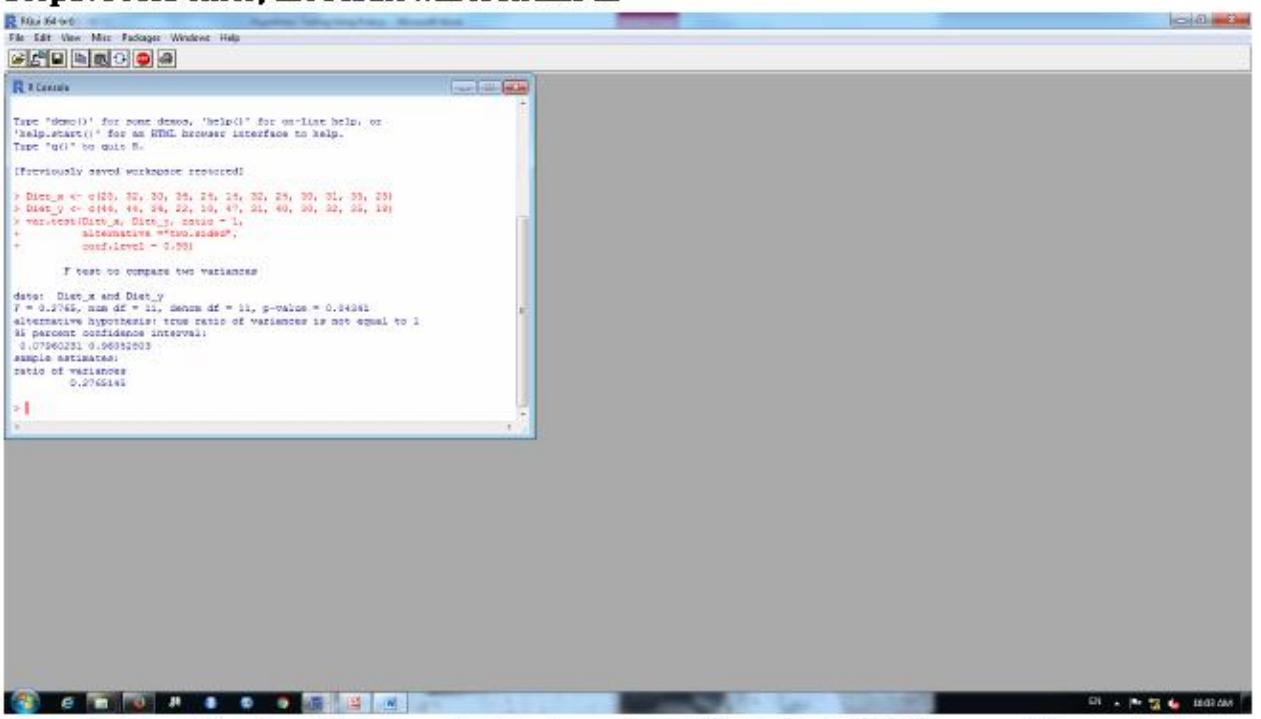
Step1: Enter the data i.e. the value of sample:



Step2:type the code for F-test as:



Step3: Press enter; the result will look like as



Description of output: Since, calculated p-value=0.04341<0.05. Hence, null hypothesis is rejected for two tailed test at 5% level of significance. . Now, for one tailed test, one can simply change in the R-code as: type "less" or "greater" in place of two.sided.

R-Code:

```
Diet_x <- c(25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25)
Diet_y <- c(44,44,34,22,10,47,31,40,30,32,35,18)
var.test(Diet_x, Diet_y, ratio = 1,
      alternative = c("two.sided"),
      conf.level = 0.95)
```

Output:

```
F test to compare two variances
data:  Diet_x and Diet_y
F = 0.27651, num df = 11, denom df = 11, p-value = 0.04341
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.07960231 0.96052803
sample estimates:
ratio of variances
      0.2765145
```

Test 4: χ^2-test for testing the goodness of fit of the data.

Object: To investigate the significance of the differences between observed data arranged in K classes, and the theoretically expected frequencies in the K classes.

Limitation:

1. The observed and theoretical distributions should contain the same number of elements.
2. The division into classes must be the same for both distributions.
3. The expected frequency in each class should be at least 5.
4. The observed frequencies are assumed to be obtained by random sampling.

Method: The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i represent the observed and theoretical frequencies respectively for each of the K classes. This statistic is compared with a value obtained from $\chi 2$ tables with v degrees of freedom. In general, $v = K - 1$. If $\chi 2$ is greater than the critical value we reject the null hypothesis that the observed and theoretical distributions agree.

Example: Test the hypothesis whether the students smoking habit is independent of their exercise level at 0.05 significance level.

Freq	None	Some
Heavy 7	1	3
Never 87	18	84
Occas 12	3	4
Regul 9	1	7

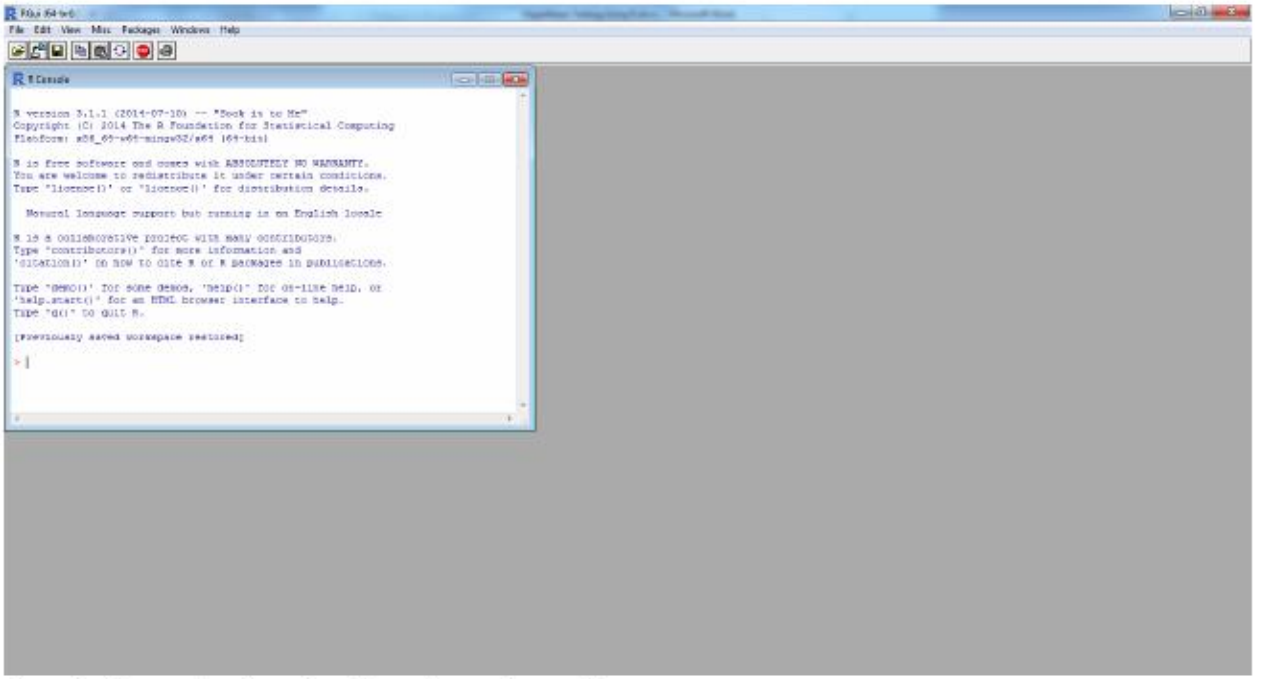
Solution:

Null hypothesis: Ho: The students smoking habit is independent of their exercise.

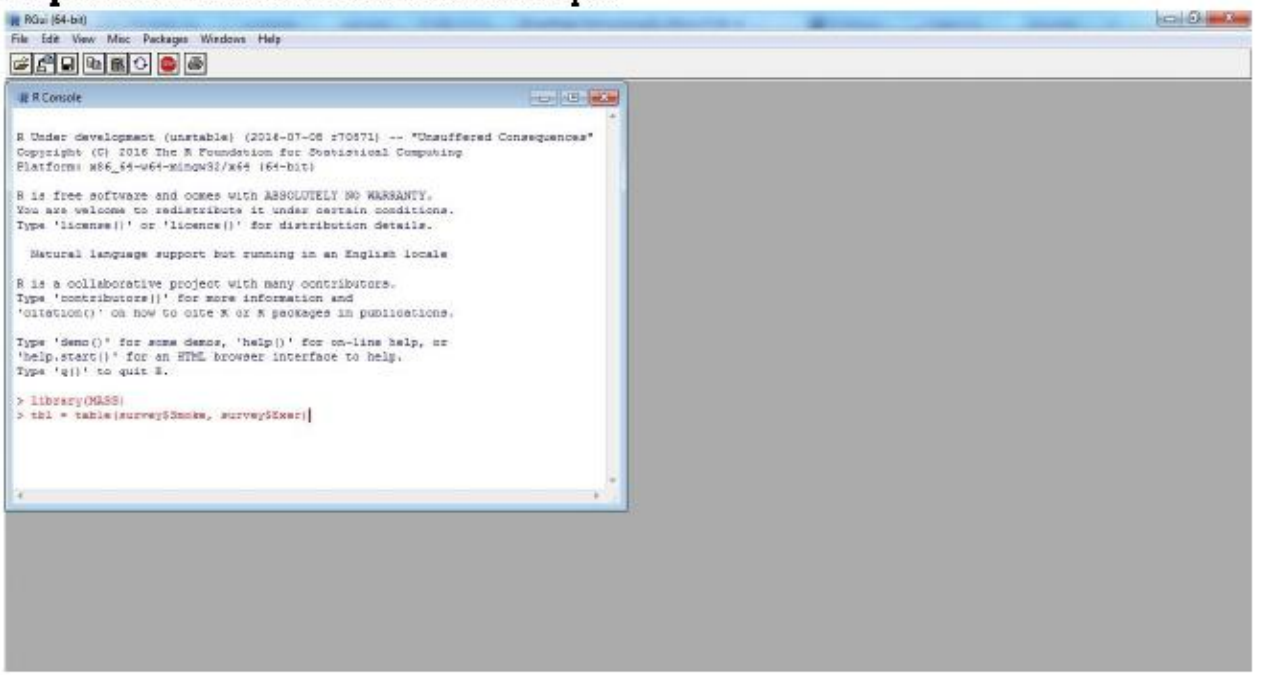
Alternative hypothesis:H1: The students smoking habit is dependent of their exercise.

To apply χ^2 test for the above, we proceed in the following steps:

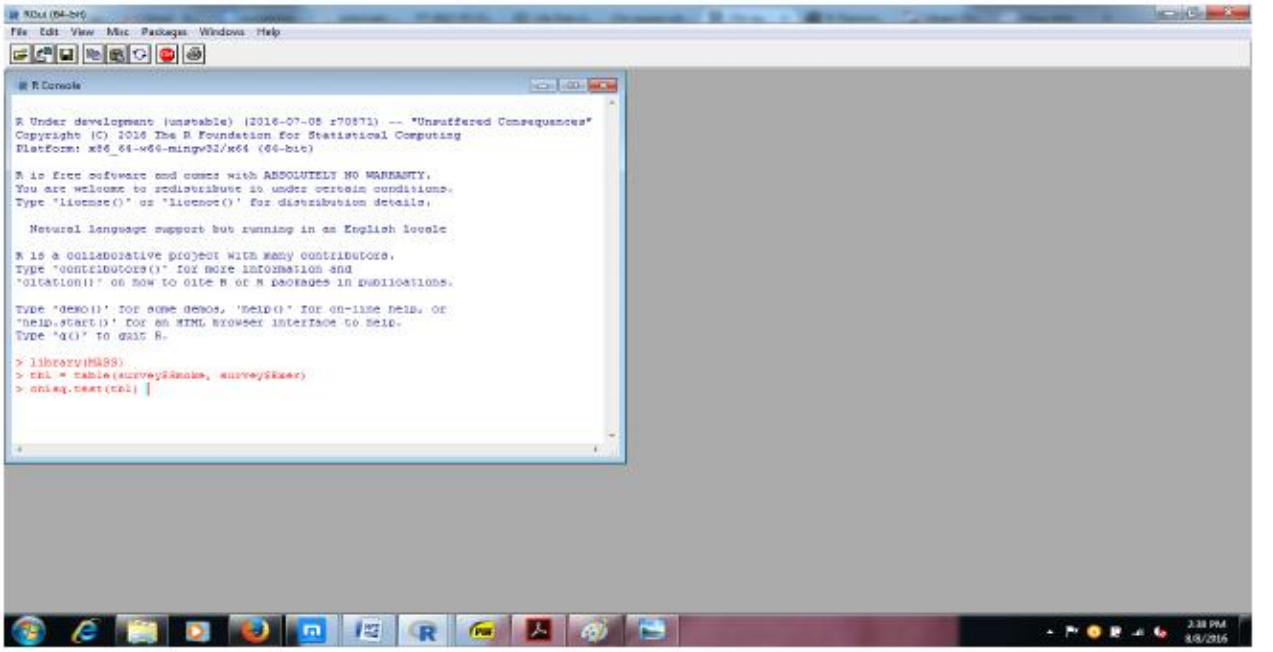
Step 1: First of all, open R; window will appear just as bellow:



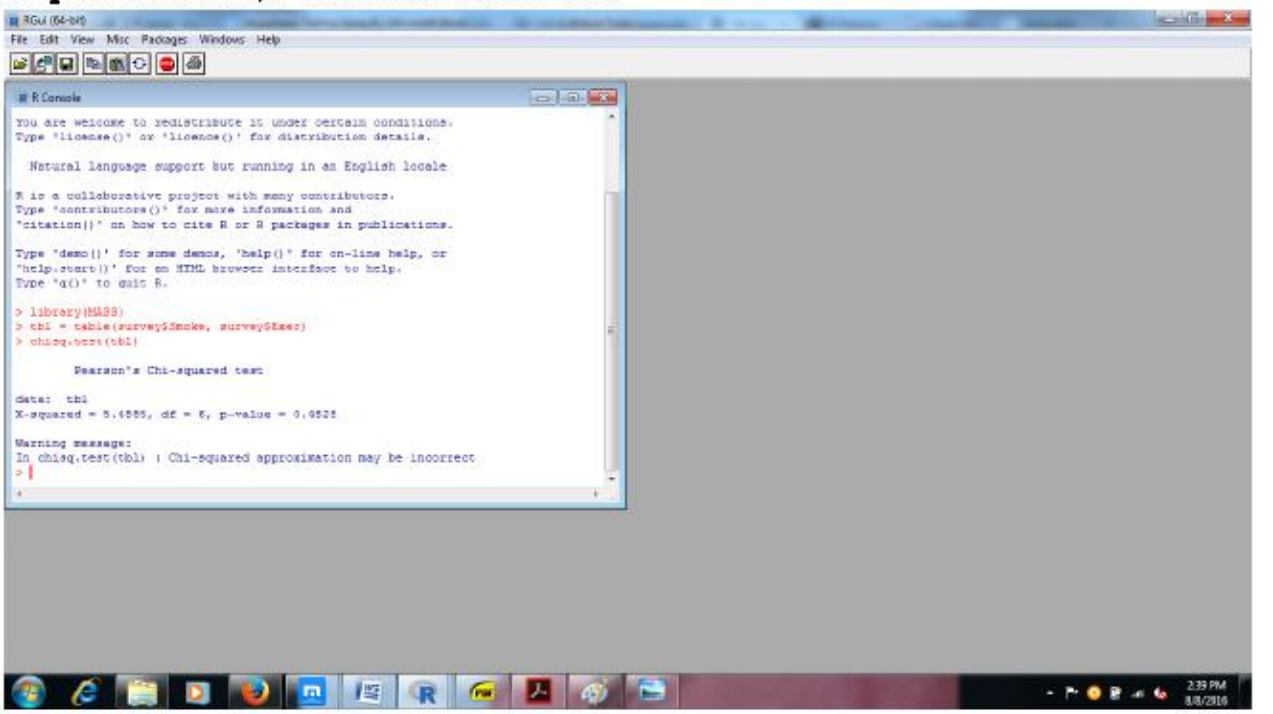
Step1: Enter the data i.e. the value of sample:



Step2:type the code for chi-square test as:



Step3: Press enter; the result will look like as



Description of output: Since, calculated p-value=0.4828>0.05. Hence, null hypothesis is accepted at 5% level of significance. and hence we may conclude that The students smoking habit is independent of their exercise.

R-Code:

```
library(MASS)
tbl = table(survey$Smoke, survey$Exer)
chisq.test(tbl)
```

Output:

```
Pearson's Chi-squared test
data: tbl
X-squared = 5.4885, df = 6, p-value = 0.4828
```