SECTION 1: DESCRIPTIVE STATISTICS AND STATISTICAL INFERENCES

Structure **Page Nos.**

1.0 INTRODUCTION

Descriptive Statistics deals with the process of identification of characteristics of collected data. The data collected from many sources may be of different type. You may like to arrange such data so that you may be able to draw logical conclusion from it. You may use several techniques to describe characteristics of data. These techniques may involve graphical representation of data, measures of finding central tendencies of data and summarisation of data by finding the data dispersion and variability. Once the data has been described, the next step would be to draw reliable conclusions using the sample data for the entire population. This section elaborates use of some of the data distributions that are used in statistical analysis – t, chi-square and F distributions. This section also deals with hypothesis testing and test of significance. You are advised to go through BCS040/Block1 and Block 2, before going through this unit.

All the spreadsheet figures used in this file has been put in a file. You can download this file from the BCA pages of IGNOU's website.

1.1 OBJECTIVES

After performing the activities of this section, you should be able to:

- Use functions that are used for descriptive statistics
- Draw charts using spreadsheet software for describing data
- Use functions for statistical inferences.
- Appreciate the steps involved in some of the statistical calculations.

1.2 FREQUENCY DISTRIBUTION OF A VARIABLE

A statistical study is conducted to generate valid conclusions about the problem under investigation. Recall from BCS040: Block 1 Unit 1 to appreciate the fact that relevant data is necessary for this purpose. Further, data can be collected either for the first time (*primary data*) directly by the investigator or from other sources already available (*secondary data*) in the form of some published work or from Government data resources etc. Broadly, statistics deals with data collection, data analysis and interpretation of results, with Statistical inference playing a role. Some of the key definitions used from this view point are given below: (Please refer to BCS040: Block 1 Unit 1, for details):

Population and Sample: The set of all the observations relating to the problem under investigation consists of the population. The sample on the other hand, consists of data actually collected from the few units selected from the population. For example, if we want to find the average income of adults in our country, the population consists of the data on income of every adult in the country. However, collecting and analyzing such huge data is difficult or impossible. Thus, we may take only a small part (See Book 3 Chapter 5) of the observations in the population, which is called a sample.

Discrete or discontinuous Variables: Consider an example of marks scored by students in statistics out of 100 that are awarded as whole numbers. A variable of this type that can take distinct, finite or countably infinite values is called a discrete variable.

Continuous Variables: The continuous variables on the other hand can take any value between a low and high value. Measures of height, weight etc. are examples of this type.

Frequency Distribution: is the most common method for summarizing and presentation of data, which enables us to quickly assess how frequently any value occurs in the given data set.

You must read Unit 1 of Block 1 of BCS-040 for more definitions and examples. In the following example the use of spreadsheet package is demonstrated for the purpose of generating a frequency distribution.

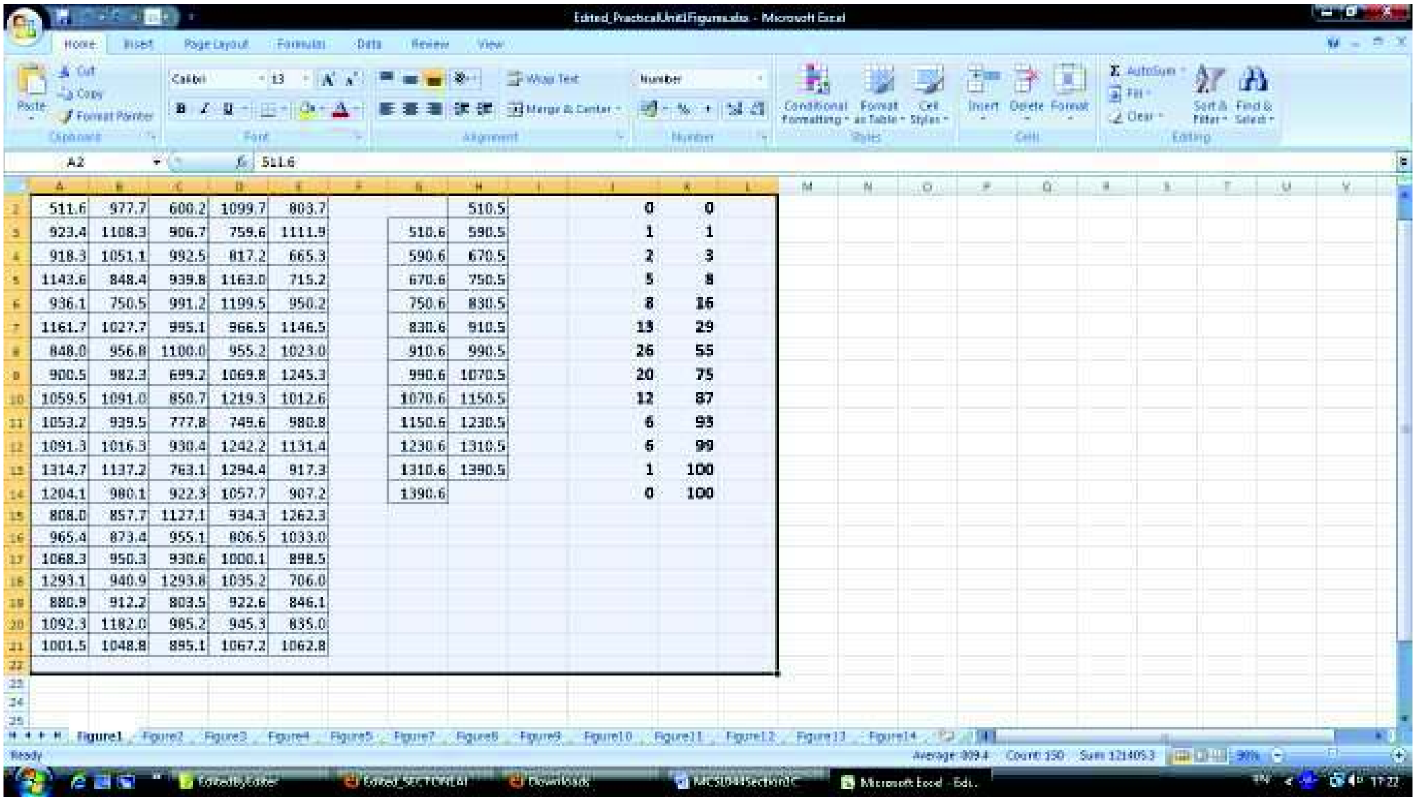
Example 1: (Data used for this example has been taken from Block 1 Unit 1 of BCS 040 example 7, table 3 with five modifications) Table 3 shows the lives of 100 electric bulbs. You are required to construct a frequency distribution and create histogram from this data using Spreadsheet package.

Solution: Figure 1 shows the spreadsheet (for the example, MS-Excel has been used, but you are free to use any spreadsheet package). Enter the given data in the cells A2..E21, that needs to be used to construct the frequency distribution. (Read page 15, Unit 1 of Block 1 of BCS-040). In the next step, you are required to specify class intervals to construct a frequency distribution. In this context, observe that the minimum and maximum values for the raw data are 511.6 and 1314.7 respectively (use ***min(A2:E21)***, ***max(A2:E21)***), as these are useful information in deciding about the classes to be formed. Here, the data under bins shows the class categories and they can be read as 0 to 510.5 (included), 510.6 to 590.5 (included) and so on. The last category is 1390.6 and above. Please note that for the purpose of calculation of frequencies, the upper limits of the classes (H2..H13) alone are required.

Once the basic data is entered you need to calculate the Frequency distribution for these class intervals. For this you need to enter an array formula in the spreadsheet. An array formula performs calculations on an array of data and may produce an array of results. To enter array formula you may need to press CTRL+SHIFT+ENTER into a worksheet. The formula for calculating frequency is also an array formula. To enter this formula, perform the following steps:

- Select the cells in which result is expected in our case these are J2..J14.
- Enter the formula: ***=frequency(A2:E21,H2:H13)*** and press CTRL+SHIFT+ENTER keys.
- You can also find the Cumulative frequencies, use simple addition formula for this purpose in cells K2..K14. Write your own formula for this purpose.

On completion of these operations the screen will look like as that of Figure 1.



A2 511.6

	A	B	C	D	E	F	G	H	I	J	K
2	511.6	977.7	600.2	1099.7	803.7			510.5		0	0
3	923.4	1108.3	906.7	759.6	1111.9		510.6	590.5		1	1
4	918.3	1051.1	992.5	817.2	665.3		590.6	670.5		2	3
5	1143.6	848.4	939.8	1163.0	715.2		670.6	750.5		5	8
6	936.1	750.5	991.2	1199.5	950.2		750.6	830.5		8	16
7	1161.7	1027.7	995.1	966.5	1146.5		830.6	910.5		13	29
8	848.0	956.8	1100.0	955.2	1023.0		910.6	990.5		26	55
9	900.5	982.3	699.2	1069.8	1245.3		990.6	1070.5		20	75
10	1059.5	1091.0	850.7	1219.3	1012.6		1070.6	1150.5		12	87
11	1053.2	939.5	777.8	749.6	980.8		1150.6	1230.5		6	93
12	1091.3	1016.3	930.4	1242.2	1131.4		1230.6	1310.5		6	99
13	1314.7	1137.2	763.1	1294.4	917.3		1310.6	1390.5		1	100
14	1204.1	980.1	922.3	1057.7	907.2		1390.6			0	100
15	808.0	857.7	1127.1	934.3	1262.3						
16	965.4	873.4	955.1	806.5	1033.0						
17	1068.3	950.3	930.6	1000.1	898.5						
18	1293.1	940.9	1293.8	1035.2	706.0						
19	880.9	912.2	803.5	922.6	846.1						
20	1092.3	1182.0	985.2	945.3	835.0						
21	1001.5	1048.8	895.1	1067.2	1062.8						

Figure 1: Calculating Frequencies in Spreadsheet

- Compare these results to those given in the Unit 1 of Block 1 of BCS-040. Do you find any difference? If yes, then discuss the reason for this difference with other students or counsellors.

- Please note that you may also use the Struges' rule that enables you to calculate the number of classes (k) in terms of total number of observations (N) as $k = [1 + log_2N]$, where [….] is the celling operator (see help on function ***CEILING(…,…)***). Can you redo the task above using this method? Further, discuss with other students or counselors how would it help in developing a C-program for construction of frequency distribution without user intervention.

To create a histogram as well as frequency distribution, you may also use *Data* → *Data Analysis* Menu. In some of your worksheets this option may not be displayed. In such case take the help of the software for "*Load the Analysis ToolPak*". On loading this, you will be able to see the *Data Analysis* option in the *Data* Menu option.
In order to make the histogram, you just need to use the original raw data and bins, so delete all other data in Figure 1 except keep the columns A to E and H. Moreover,

transfer the column H data to G by incrementing previous G data by 0.1. Now to make the histogram perform the following steps:

1. Select *Data* → *Data Analysis* and then select *Histogram* in the resulting dialog box and press Ok button.
2. In the resulting dialog box, set the *Input Range* to A2..E21, *Bin Range* to G2..G13 and *Output Range* to I1..J14. Please do not forget to check the *Chart Output* check box.
3. Click Ok

The Histogram will be displayed as shown in the Figure 2. You need to modify the format of the histogram to make it look as per your need. For example, to remove gaps in between bars:

1. Click on Bar →*Right Click* →*Format Data Series…*
2. *Series Options* →*Gap Width 0%*
3. *Fill* →*Pattern Fill*
4. *Border Colour* →*Solid Line etc..*

Notice that horizontal axis labels now depict the upper class boundaries having class width of 80 hours in Figure 2. Once again compare this histogram to the figure that has been made in the Unit. Can you now plot the frequency polygon and the less than type ogive for the same frequency distribution (refer to Block 1, unit 1, section 1.2.3 page no. 17)? You can also plot Bar diagram for appropriate data as shown in the Unit.

You can also explore the fact that all spreadsheet packages allow you to modify the Title of the chart, Axis titles, colour of bars, size of bars, size of chart and many other formatting features. For this you should refer to documentation of the spreadsheet package that you are using.

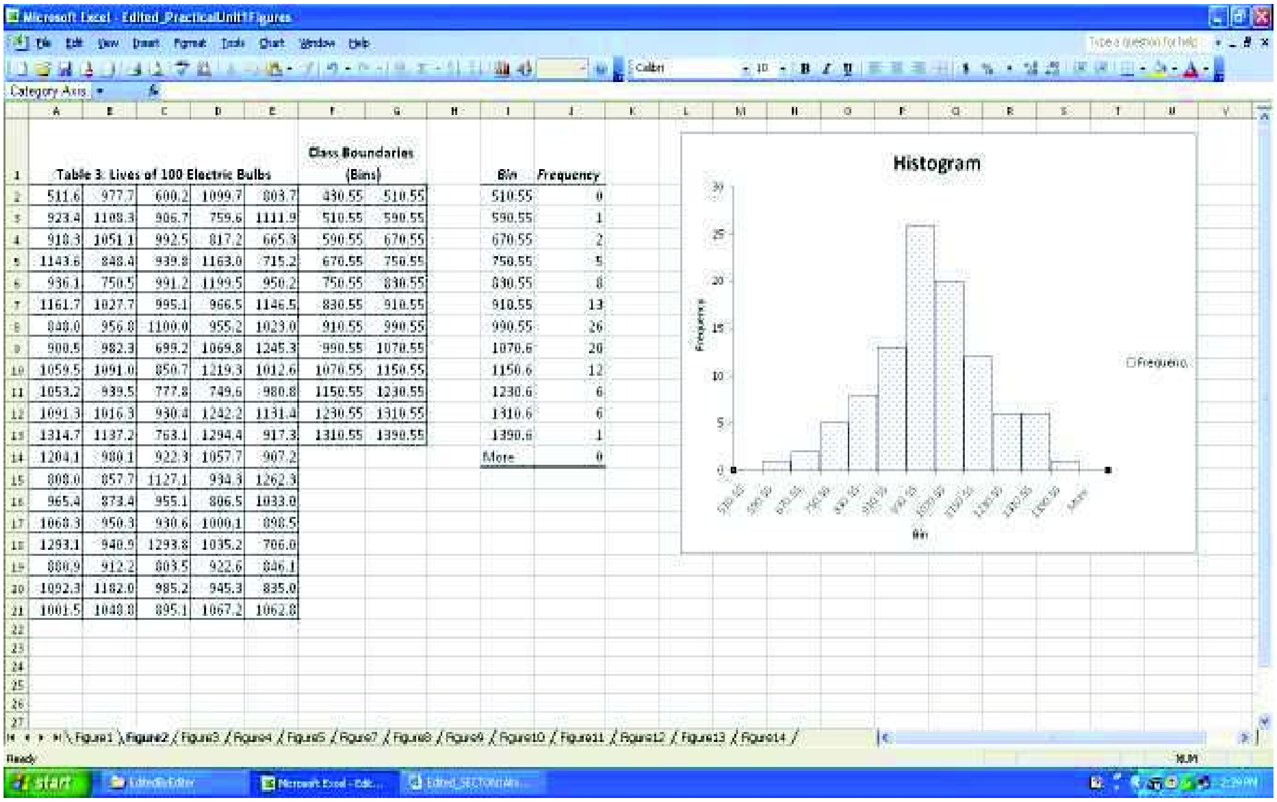

Microsoft Excel - Edited_PracticalUnit1Figures
File Edit View Insert Format Tools Chart Window Help
Calibri
10
Category Axis
Histogram
Frequency
0
5
10
15
20
25
30
510.55
590.55
670.55
750.55
830.55
910.55
990.55
1070.55
1150.55
1230.55
1310.55
1390.55
More
Bin
□Frequency
Figure1 Figure2 Figure3 Figure4 Figure5 Figure7 Figure8 Figure9 Figure10 Figure11 Figure12 Figure13 Figure14
Ready
NUM
start

Table 3: Lives of 100 Electric Bulbs					Class Boundaries (Bins)		Bin	Frequency
511.6	977.7	600.2	1099.7	803.7	430.55	510.55	510.55	0
923.4	1108.3	906.7	759.6	1111.9	510.55	590.55	590.55	1
918.3	1051.1	992.5	817.2	665.3	590.55	670.55	670.55	2
1143.6	848.4	939.8	1163.0	715.2	670.55	750.55	750.55	5
936.1	750.5	991.2	1199.5	950.2	750.55	830.55	830.55	8
1161.7	1027.7	995.1	966.5	1146.5	830.55	910.55	910.55	13
848.0	956.8	1100.0	955.2	1023.0	910.55	990.55	990.55	26
900.5	982.3	699.2	1069.8	1245.3	990.55	1070.55	1070.6	20
1059.5	1091.0	850.7	1219.3	1012.6	1070.55	1150.55	1150.6	12
1053.2	939.5	777.8	749.6	980.8	1150.55	1230.55	1230.6	6
1091.3	1016.3	930.4	1242.2	1131.4	1230.55	1310.55	1310.6	6
1314.7	1137.2	763.1	1294.4	917.3	1310.55	1390.55	1390.6	1
1204.1	980.1	922.3	1057.7	907.2			More	0
808.0	857.7	1127.1	934.3	1262.3				
965.4	873.4	955.1	806.5	1033.0				
1068.3	950.3	930.6	1000.1	898.5				
1293.1	940.9	1293.8	1035.2	706.0				
880.9	912.2	803.5	922.6	846.1				
1092.3	1182.0	985.2	945.3	835.0				
1001.5	1048.8	895.1	1067.2	1062.8				

Figure 2: Histogram using the Data → Data Analysis options.

1.3 SUMMARISATION OF DATA

Two important statistical measures that are used to summarise of data are:

- measures of central tendency
- measure of dispersion

Let us discuss them with the help of the example given in BCS-040, Block 1, Unit 1 page 30-31. The following figure shows an implementation similar to Table 14 of the said block/unit.

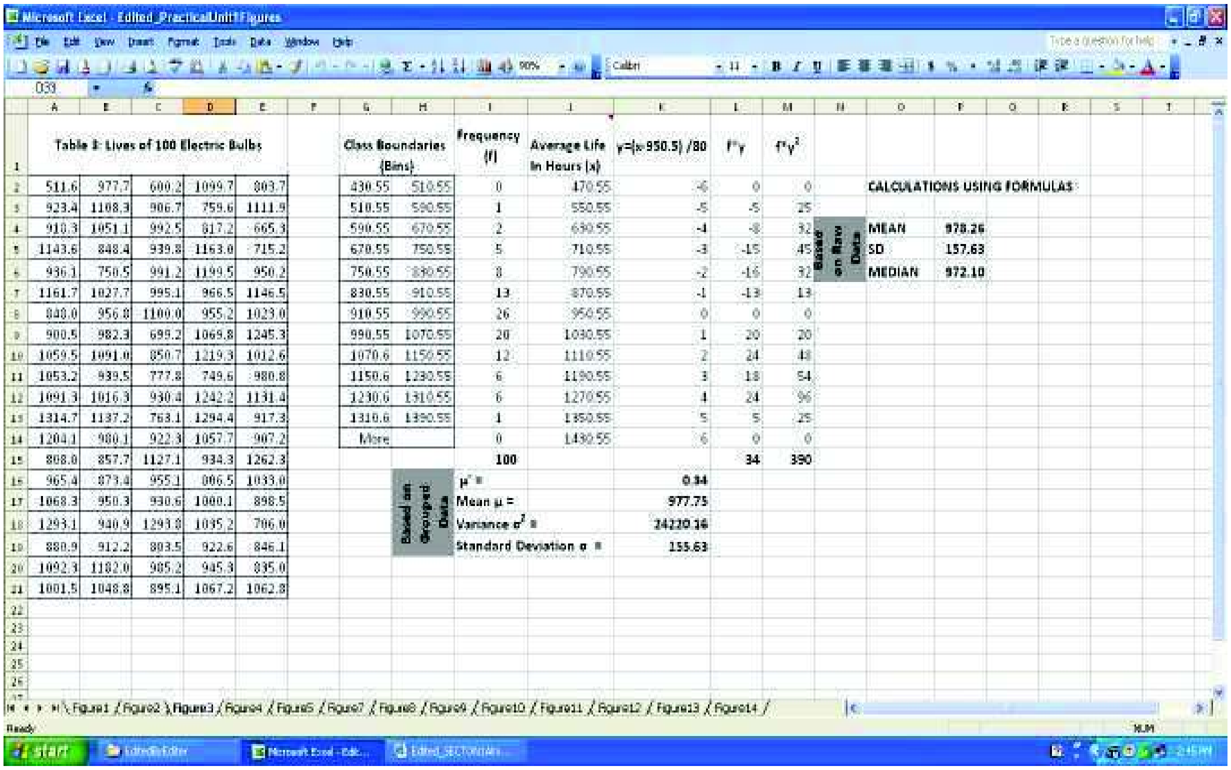

Microsoft Excel - Edited_PracticalUnit1Figures
File Edit View Insert Format Tools Data Window Help
Type a question for help
90%
Calibri
11
O33
Figure1 / Figure2 / Figure3 / Figure4 / Figure5 / Figure7 / Figure8 / Figure9 / Figure10 / Figure11 / Figure12 / Figure13 / Figure14 /
Ready
NUM
start
EditedByEditor
Microsoft Excel - Edit...
Edited_SECTION1An...

Table 3: Lives of 100 Electric Bulbs

A	B	C	D	E
511.6	977.7	600.2	1099.7	803.7
923.4	1108.3	906.7	759.6	1111.9
910.3	1051.1	992.5	817.2	665.3
1143.6	848.4	939.8	1163.0	715.2
936.1	750.5	991.2	1199.5	950.2
1161.7	1027.7	995.1	966.5	1146.5
840.0	956.8	1100.0	955.2	1023.0
900.5	982.3	699.2	1069.8	1245.3
1059.5	1091.0	850.7	1219.3	1012.6
1053.2	939.5	777.8	749.6	980.8
1091.3	1016.3	930.4	1242.2	1131.4
1314.7	1137.2	763.1	1294.4	917.3
1204.1	980.1	922.3	1057.7	907.2
808.0	857.7	1127.1	934.3	1262.3
965.4	873.4	955.1	906.5	1033.0
1068.3	950.3	930.6	1000.1	898.5
1293.1	940.9	1293.8	1095.2	706.0
880.9	912.2	803.5	922.6	846.1
1092.3	1182.0	985.2	945.3	835.0
1001.5	1048.8	895.1	1067.2	1062.8

Class Boundaries (Bins)		Frequency (f)	Average Life In Hours (x)	y=(x-950.5) /80	f*y	f*y^2
430.55	510.55	0	470.55	-6	0	0
510.55	590.55	1	550.55	-5	-5	25
590.55	670.55	2	630.55	-4	-8	32
670.55	750.55	5	710.55	-3	-15	45
750.55	830.55	8	790.55	-2	-16	32
830.55	910.55	13	870.55	-1	-13	13
910.55	990.55	26	950.55	0	0	0
990.55	1070.55	20	1030.55	1	20	20
1070.6	1150.55	12	1110.55	2	24	48
1150.6	1230.55	6	1190.55	3	18	54
1230.6	1310.55	6	1270.55	4	24	96
1310.6	1390.55	1	1350.55	5	5	25
More		0	1430.55	6	0	0
		100			34	390

Based on Grouped Data	
μ' =	0.34
Mean μ =	977.75
Variance σ^2 =	24220.16
Standard Deviation σ =	155.63

CALCULATIONS USING FORMULAS

Based on Raw Data	
MEAN	978.26
SD	157.63
MEDIAN	972.10

Figure 3: Applying the measures of Central tendency and dispersion

Please note that we have used the following procedure in the worksheet to arrive at the result shown in Figure 3:

- In the cell I2 to I14, you can use the array formula as explained in the last section to calculate the frequencies of each class.
- For calculating the Average life in hours (x_j) or the class mid-points, you may enter in the cell J2 the formula *=(G2+H2)/2.*
- Copy this formula from cell J3 to J13. Notice that you need to find the value for cell J14 using different formula.
- Calculate the value of y_j in the cell K2 as per the formula *=(J2-$J$8)/80.* Copy this formula to cells K3 to K14.
- Now enter the formula for calculating $y_j f_j$ and (you can do it yourself).
- Enter the formula for the μ′ (see page 30, Unit 1) - the mean for x-observation, variance for x-observations and Standard deviation for x - observation.

You can compare the results so obtained with the results that you can obtain by directly applying the formula of mean (AVERAGE), median (MEDIAN) and standard deviation (STDEV) on the cell range A2:E21 containing the original data. Discuss with other students or counselors, why are there differences in the values of mean and standard deviations. Please note that the values so calculated are different than the values as shown in the Table 14 of the course BCS040.

☞ Lab Session 1

1) Construct a frequency distribution, plot bar graph, and find the mean and standard deviation of the data on marks of the 100 students as given in BCS040/Block 1/Unit 1 Table 2 (page 11) using spreadsheet package.

2) Develop a program in C to construct a frequency distribution and hence calculate mean & standard deviation for the data mentioned in problem 1. Your program should make provision to input data from a file and output results on the screen (you may also use Struges' rule).

3) Using spreadsheet package generate hypothetical data on marks of 10 different students in 5 different subjects. You may use the spreadsheet package function ***RANDBETWEEN(bottom,top)*** for this purpose. Find the average (μ) standard deviation of marks (σ) and calculate grades of the students on that basis of the following: A student who receives marks in the range of $\mu \pm \sigma$ is awarded grade *B*. A student who gets marks above $\mu + \sigma$ is awarded an *A* grade. The student getting marks below $\mu - \sigma$ is awarded a *C* grade. Make suitable assumptions, if any.

1.4 SAMPLING DISTRIBUTIONS

This section discusses about the basis of various tests that you can perform for hypothesis testing. You should go through BCS-040 Block 2 before going through this section. Let us first revise some of the terms used in that Block.

Population and Sample: In statistics the term population refers to a set or collection of observations relating to the phenomenon under investigation. Thus, the statistical population or simply population comprises all the observations or measurements, relating to the phenomenon under investigation, which can be collected. The population can be finite or infinite. In an infinite population it is not possible to observe the measurements on all the units; even in the case of a finite population it may not be economical or feasible to observe the values from all the units of the population. Thus, *a representative set of units* from the population are chosen *using a statistically valid procedure* and measurements or observations are made from these selected units. This subset of observations comprises the *sample* that are selected for performing statistical analysis (please refer to Part II of Book 3).

Notation: Consistent with the BCS-040/Block 2, Unit 4, the population mean is denoted by μ and the sample mean is denoted by $\bar{x}$. Similarly, you can define the standard deviation for the population and the sample.

Statistic: A function of sample of observations, which does not contain any unknown parameter and whose values can be observed, is called a *statistic*. It is denoted by . For example, $T = \bar{X}$ is a statistic, as its values can be observed and it does not contain any unknown parameter; $T = \bar{X} + \mu$ is not a statistic when μ is unknown.

The *Sampling distribution* of a Statistic refers to the list of all possible values of the statistic and its associated probability distribution.

Let us try to show you the use of Spreadsheet for calculating the sampling distribution. Consider the BCS-040/Block 2, Unit 4, Page 9 Example 1, using which the notion of population mean, sample, sample mean, list of all possible samples and mean of sample mean or grand mean are presented.

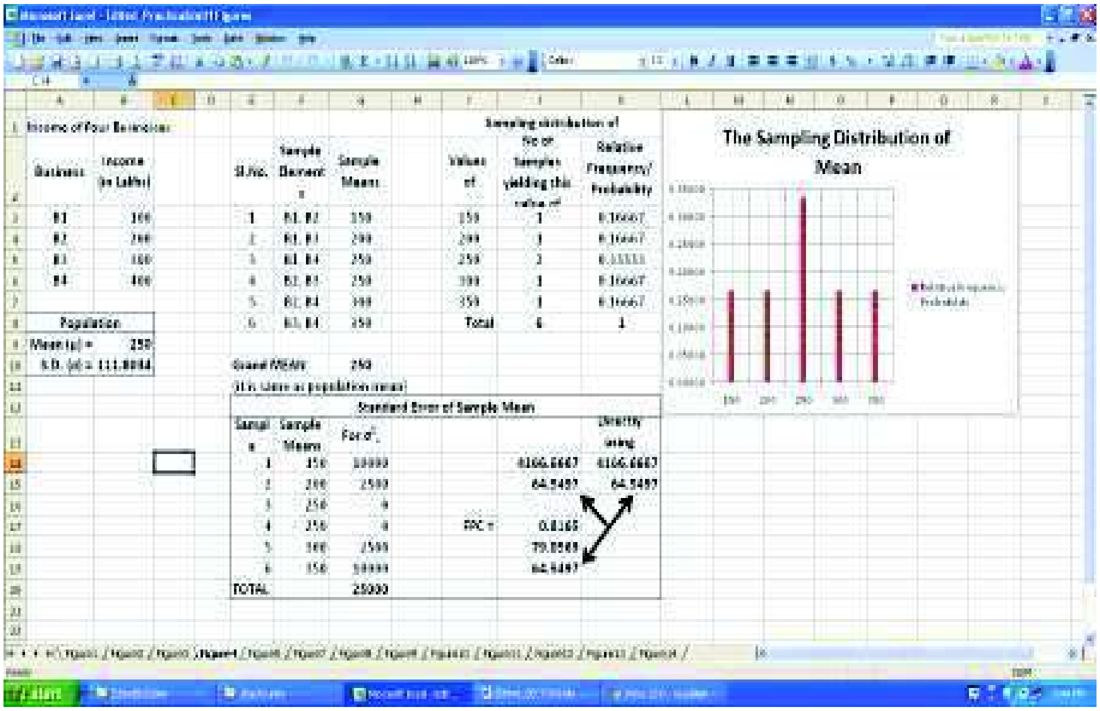
Example 1: Suppose we have a population of incomes of N = 4 Business firms viz., 100, 200, 300, 400 (in Lakhs).

Taks:

- List all possible samples of size n = 2 drawn from the population *without replacement*
- Calculate the mean and variance of each sample
- Construct the sampling distribution of the sample mean
- Calculate the mean and variance of this distribution

Note that in sampling without replacement, a total of ${}^4C_2 = 6$ possible samples can be drawn from the population. The results related to the tasks stated above are shown in Figure 4. In order to generate the content of Figure 4, you can use your own formulas by following the steps below:

- Create the column F of sample elements
- Calculate the sample means in columns G
- Hence construct the list of all possible values of sample mean $\overline{X}$ in column I.
- Use the array formula *=FREQUENCY(G3:G8,I3:I7)* in column J to get the frequency distribution of sample mean $\overline{X}$. From this frequency distribution calculate the Relative frequency / Probability distribution $P(\overline{X})$ of sample mean. Note that here $\overline{X}$ denotes a random variable and $\overline{x}$ its value.
- Now, you can insert the chart of Sample Mean versus relative Frequency. The Chart type used should be decided keeping in view the discrete or continuous nature of the distribution. This is the graph showing the sampling distribution of $\overline{X}$.
- Also note that you can calculate the Grand mean $\overline{\overline{x}}$ (see ***E10..G10***) as well as standard error $SE\left(\overline{X}\right)$ as shown in the figure. Their direct computations have also been shown using worksheet functions (see ***I14..K15***).
- Computation of $SE\left(\overline{X}\right)$ using $\sqrt{\frac{N-n}{N-1}}$ the Finite Population Correction Factor (FPC) has been done in ***I17..J19*** using formula $SE\left(\overline{X}\right) = \sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}}$ (see BCS-040/Block 2, Unit 4, Page 12 or Book 3, Page 5.6).



Income of Four Businesses

Business	Income (in Lakhs)
B1	100
B2	200
B3	300
B4	400

Population	
Mean (μ) =	250
S.D. (σ) =	111.8034

Sl.No.	Sample Element s	Sample Means
1	B1, B2	150
2	B1, B3	200
3	B1, B4	250
4	B2, B3	250
5	B2, B4	300
6	B3, B4	350

Grand MEAN 250
(i.e. same as population mean)

Sampling distribution of

Values of	No of Samples yielding this value of	Relative Frequency/ Probability
150	1	0.16667
200	1	0.16667
250	2	0.33333
300	1	0.16667
350	1	0.16667
Total	6	1

Standard Error of Sample Mean

Sample s	Sample Means	For σ²			Directly using
1	150	10000		4166.6667	4166.6667
2	200	2500		64.5497	64.5497
3	250	0			
4	250	0	FPC =	0.8165	
5	300	2500		79.0569	
6	350	10000		64.5497	
TOTAL		25000			

Figure 4: The sampling distribution of Example 1

Some important exact sampling distributions that are widely used in statistical data analysis are Chi-Square distribution (χ^2), Student's t-distribution, F-distributions and these are briefly discussed subsequently.

1.5 SOME IMPORTANT DEFINITIONS

An important objective of statistical data analysis is to draw inferences and conclusions about the aspect being investigated. In this context, *statistical inference* provides us methodology for doing the same. **The test of significance is a formal procedure that is aimed at assessing evidence that is provided by sample(s) data in favor of some claim/inference about the entire population.** Before going through this section, you may go through the Complete Block2 of BCS-040.

Some of the key terms used here are defined below for your recapitulation:

Normal Distribution: It is an important probability distribution of a *continuous random variable*, which you will use in solving many problems subsequently. The probability density function (PDF) $f(x) = f(x; \mu, \sigma^2)$ of a normal distribution having parameters μ (Mean) and σ^2 (Variance) is below:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), -\infty < x < \infty$$

Further,

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

is a Standard Normal random variable that has PDF

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left(-\frac{z^2}{2}\right), -\infty < z < \infty$$

with parameters $\mu = 0$ (Mean) and $\sigma^2 = 1$ (Variance). BCS-040/Block 1/Unit 3/Section 3.6 explains normal distribution in details.

In general, the expression $z = \frac{x - \mu}{\sigma}$ gives us the standardize values for the values of random variable X having Mean μ and Variance σ^2. What are the mean and variance of Z ?

Task: Try your hand in calculating the following in cells ***Q22..T24*** by altering the values of U and V in cells ***Q23..Q24***.

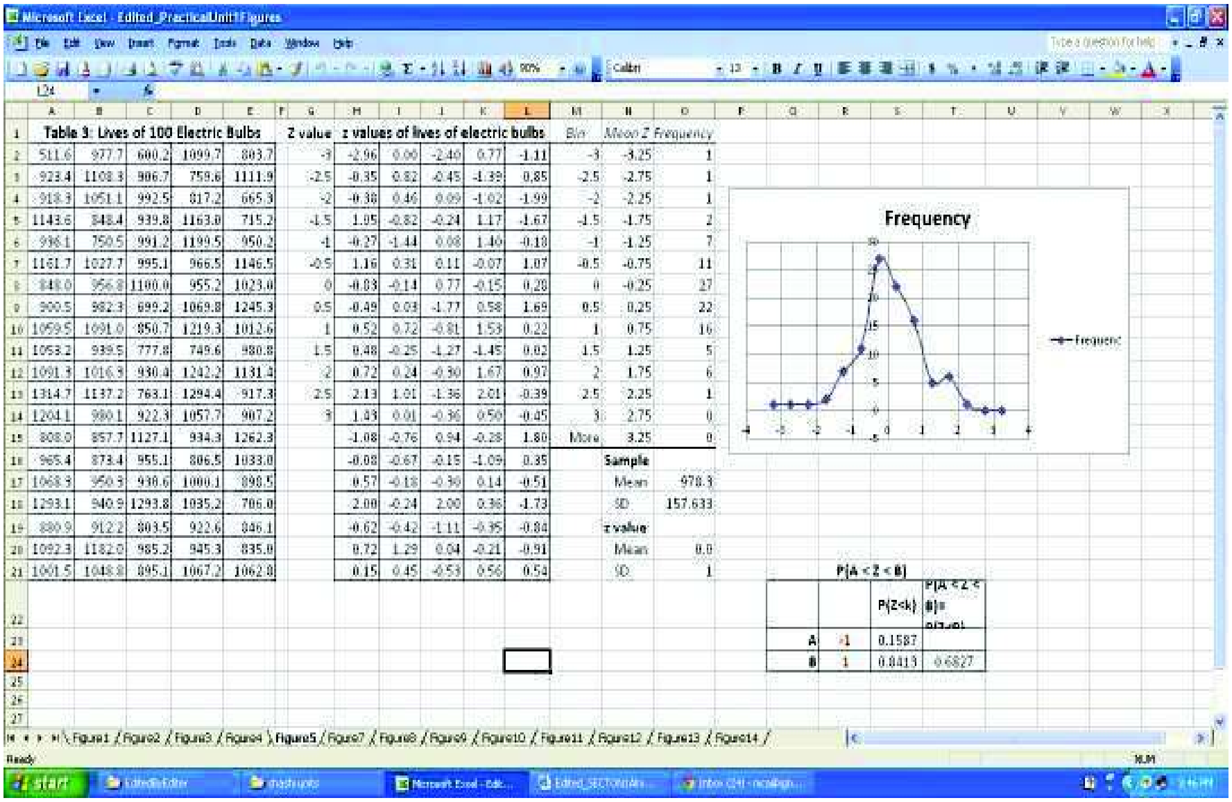
$P(-1 < Z < 1) = 0.6827$
$P(-2 < Z < 2) = 0.9545$
$P(-3 < Z < 3) = 0.9973$

Can you now calculate probabilities $P(-1 < Z < 0)$, $P(-2 < Z < 3)$?

Task: Construct the frequency distribution of the standardized values for the data on life of light bulbs used in Figure 1 & 2 (see Figure 5 for the results).

Following steps will enable you to get the content of Figure 5.

- The standardized values Z of the random variable X are calculated in the range ***H2..L21***.
- The mean of Z -values and standard deviations are given in ***O20..O21***.
- The *Data→Data_Analysis→Histogram* option was used to generate summary statistics (bin, frequency).


Microsoft Excel - Edited_PracticalUnit1Figures
File Edit View Insert Format Tools Data Window Help
Type a question for help
90%
Calibri
12
L24
Frequency
Frequenc
Figure1 / Figure2 / Figure3 / Figure4 / Figure5 / Figure7 / Figure8 / Figure9 / Figure10 / Figure11 / Figure12 / Figure13 / Figure14
Ready
NUM
start

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Table 3: Lives of 100 Electric Bulbs						Z value	z values of lives of electric bulbs					Bin	Mean Z	Frequency
2	511.6	977.7	600.2	1099.7	803.7		-3	-2.96	0.00	-2.40	0.77	-1.11	-3	-3.25	1
3	923.4	1108.3	906.7	759.6	1111.9		-2.5	-0.35	0.82	-0.45	-1.39	0.85	-2.5	-2.75	1
4	918.3	1051.1	992.5	817.2	665.3		-2	-0.38	0.46	0.09	-1.02	-1.99	-2	-2.25	1
5	1143.6	848.4	939.8	1163.0	715.2		-1.5	1.05	-0.82	-0.24	1.17	-1.67	-1.5	-1.75	2
6	936.1	750.5	991.2	1199.5	950.2		-1	-0.27	-1.44	0.08	1.40	-0.18	-1	-1.25	7
7	1161.7	1027.7	995.1	966.5	1146.5		-0.5	1.16	0.31	0.11	-0.07	1.07	-0.5	-0.75	11
8	848.0	956.8	1100.0	955.2	1023.0		0	-0.03	-0.14	0.77	-0.15	0.28	0	-0.25	27
9	900.5	982.3	699.2	1069.8	1245.3		0.5	-0.49	0.03	-1.77	0.58	1.69	0.5	0.25	22
10	1059.5	1091.0	850.7	1219.3	1012.6		1	0.52	0.72	-0.81	1.53	0.22	1	0.75	16
11	1053.2	939.5	777.8	749.6	980.8		1.5	0.48	-0.25	-1.27	-1.45	0.02	1.5	1.25	5
12	1091.3	1016.3	930.4	1242.2	1131.4		2	0.72	0.24	-0.30	1.67	0.97	2	1.75	6
13	1314.7	1137.2	763.1	1294.4	917.3		2.5	2.13	1.01	-1.36	2.01	-0.39	2.5	2.25	1
14	1204.1	980.1	922.3	1057.7	907.2		3	1.43	0.01	-0.36	0.50	-0.45	3	2.75	0
15	808.0	857.7	1127.1	934.3	1262.3			-1.08	-0.76	0.94	-0.28	1.80	More	3.25	0
16	965.4	873.4	955.1	806.5	1033.0			-0.08	-0.67	-0.15	-1.09	0.35		Sample	
17	1068.3	950.3	930.6	1000.1	890.5			0.57	-0.18	-0.30	0.14	-0.51		Mean	978.3
18	1293.1	940.9	1293.8	1035.2	706.0			2.00	-0.24	2.00	0.36	-1.73		SD	157.633
19	880.9	912.2	803.5	922.6	846.1			-0.62	-0.42	-1.11	-0.35	-0.84		z value	
20	1092.3	1182.0	985.2	945.3	835.0			0.72	1.29	0.04	-0.21	-0.91		Mean	0.0
21	1001.5	1048.8	895.1	1067.2	1062.0			0.15	0.45	-0.53	0.56	0.54		SD	1

P(A < Z < B)

		P(Z<k)	P(A < Z < B)= P(Z<B)
A	-1	0.1587	
B	1	0.8413	0.6827

Figure 5: A Sample distribution – almost normal distribution

Point Estimates and Interval Estimates: An estimate of a population parameter that is a single number is called a point estimate. For example, sample mean $\bar{X}$ is an estimate of the unknown population mean μ . An interval estimate on the other hand provides a set of two numbers, within which the unknown parameter is likely to belong (see Book 3, Chapter 6).

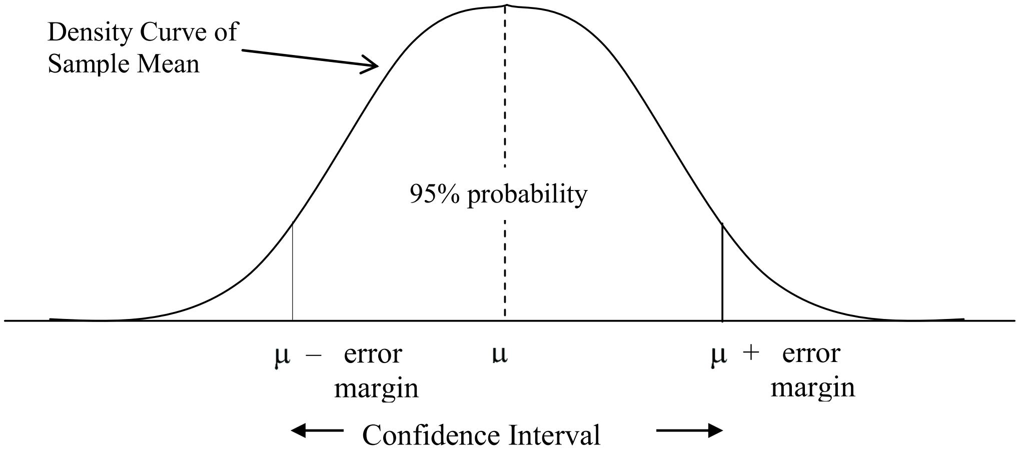
Let μ_T and σ_T be the mean and standard deviation (*standard error*) of the sampling distribution of a static T. Then, if the sampling distribution of T is approximately normal (for CLT see BCS-040/Block 2, Unit 4, page 15), than the values of probability $P\left[\left|\frac{T-\mu_T}{\sigma_T}\right| < z_c\right]$ for $z_c = 1,2,3$ are given in the table below. It follows from the table values that the end numbers $T \pm \sigma_S$, $\pm 2\sigma_S$, $T \pm 3\sigma_S$ are the 68.27%, 95.45%, 99.73% confidence limits for μ_T respectively.

For example, the confidence interval for mean (μ)
Large sample ($n \geq 30$): (Ref. BCS-040/Block 1/Unit 4/page 12-13) using standard error expressions, the confidence interval for the population mean is given by $\overline{X} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$ or $\overline{X} \pm z_c \frac{\sigma}{\sqrt{n}}$. These are of the form $Estimate \pm error\ margin$ (please refer to figure 6(a)).

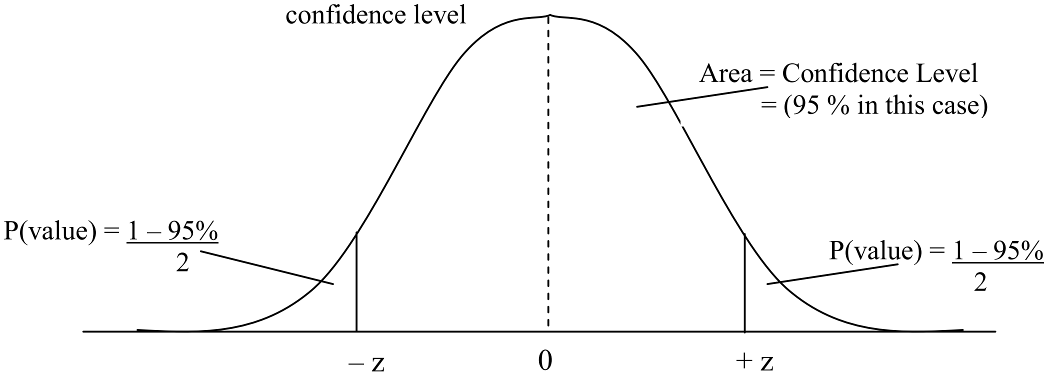
Confidence Level: The probability $P[T - z_c\, \sigma_T < \mu_T < T + z_c\, \sigma_T]$ enables us to state the level of confidence for unknown parameter μ_T to belong to the said range. For different values z_c the table below lists the confidence level..

Confidence level	**68.27%**	90%	95%	**95.45%**	99%	**99.73%**
z_c	**1**	1.645	1.96	**2**	2.58	**3**

In the Figure 6(b), the z value for 95% confidence level will be 1.96.



(a) The density curve for mean of various sample, the error margin here is for 95% confidence level



(a) Standard Normal Curve for 95% confidence level

Figure 6: Confidence Interval, Confidence level and Normal Distribution using density and standard normal curves

Please note that Figure 5(a) shows the density of the mean for various possible samples of defined number of variables. The idea in this figure is that for a confidence level of 95% there will be a range of possible sample mean values that will range within a confidence interval. However, if you convert the mean score to z score, you get a normal distribution as shown in Figure 5(b). This figure shows that there will be a standard confidence values which will be equal to the shown area on the curve (95% in the case of Figure 5). The remaining area called α (5% or 0.05) will be in two tails (2.5% or 0.025 each).

Null Hypothesis: The Null Hypothesis is the statement or the claim that there is no difference in true means or proportions of groups that are being compared. It is observed that a Null hypothesis many a times includes phrase like ***no effect*** or ***no difference***.

For example, take the case of test of the hypothesis that the average height of men in North India (μ_N) is greater than the average height of men in East India (μ_E). Notice that the claim in this case is $\mu_N > \mu_E$ and this forms the *alternative hypothesis* H_1. Having no predetermined reason for the heights to differ, the null hypothesis H_0 is that they are the same.

H_0: $\mu_N = \mu_E$ Average heights of men in North India and East India are same.
H_1: $\mu_N > \mu_E$ Average heights of men in North India is greater than East India

.

In general, the H_0 and H_1 can be one sided or two sided. For example, alternative hypothesis may be H_1: $\mu_N \neq \mu_E$.

Two common types of errors related to the testing of hypothesis are:

Type I error: this occurs when the null hypothesis H_0 is rejected when in fact, H_0 true.
Type II error: this occurs when the Hypothesis H_0 is false and it is not rejected.

p-Values: It is defined as the probability under H_0 of observing an equal or more *extreme* value of the test statistic, where extreme denotes departure from the null case, in the direction of the alternative hypothesis H_1. If the p-value is small, then it may lead to evidence that the sample data *does not* support H_0. For example, if $p = 0.002 \leq 0.05$, the value of statistic belongs to the region of rejection (please refer to Figure 6). Hence, the evidence against null hypothesis H_0 results in its rejection and consequently in the acceptance of alternate hypothesis H_1.

α **-Value:** In order to test the null hypothesis, it is required to fix a cutoff value for p-value. An α value of 0.05 means that the evidence provided by data against H_0 should be so strong that it would not happen more than 5% of the time. In other words, if we carry out the test 100 times, only 5 out of these values of mean or proportion, etc. (statistic T) calculated from samples, is not in the region of acceptance. Further, it is also termed the level of significance of the test.

1.6 t-DISTRIBUTION

Please read the content of BCS-040/Block 2/Unit 4/ Section 4.4 on page 18 on the t-Distribution. We reproduced here the relevant portion below:

Suppose $X_1, X_2, \cdots X_n$ be a random sample drawn from the normal distribution $N(\mu, \sigma^2)$, where the population mean μ and the variance σ^2 is unknown (i.i.d.). Then the expression given below has Student's t-distribution or simply t-distribution with a parameter ν = n – 1

$$t = \frac{\overline{X} - \mu}{S/_{\sqrt{n}}}$$

Here, $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ is the variance of the sample

$\overline{X}$ is the mean of the sample, and

$v = n - 1$ is the degree of freedom (d.f.) of the distribution.

The probability distribution for different values of d.f. are shown in Figure 3, BCS040/Block2/Unit4/Section 4.4/page 19. The probability $P(t > t_\alpha) = \alpha$ for different values of v, α is tabulated in Page 93 (Right-Tail) are shown in Figure 4 (page 19) of the unit.

> **Important Property:** t-distribution is symmetrical about the value $t = 0$. This is useful in calculation of probabilities under the curve using the table mentioned above. Mathematically, $P(t < -t_\alpha) = P(t > t_\alpha)$.
> Thus, $P(|t| > t_\alpha) = P(t < -t_\alpha) + P(t > t_\alpha) = 2P(t > t_\alpha)$ and $-t_\alpha = t_{1-\alpha}$ (Refer to Book 3, Chapter 4).

Please note that the values given in the table are only for the Right-Tail of the t-distribution. For example, if α =0.05 and ν=10, one tailed value $t_{0.05} = 1.812$. However, if we want to find the two tailed value for the same significance level, we have to use $t_{0.05/_2} = t_{0.025}$ to read values from the columns of the table and the value of $v = 10$ in the row to get $t_{0.025} = 2.228$.

Example: Let us solve the problem of t -distribution given in Example 3 of BCS040/Block2/Unit4/Section 4.4/Page 19 using a spreadsheet package.

Given: Degree of freedom $v = 9$, find t, for which

(i) shaded portion of the right = 0.05
(ii) the total shaded area = 0.05
(iii) the total un-shaded area = 0.99
(iv) the shaded area on the left = 0.01

Solution:
The solution using the Table of t-distribution is given in the said Block. To solve this problem using spreadsheet, you need to know the following function (this function may have different name in different spreadsheet package):

TINV(probability, degreesOFfreedom)

This function returns the t -value of the Student t-distribution. It takes two parameters:

> Probability –probability for which you want to determine the t-value. Please note that the function return the value of t for two-tails probability of the curve denoted as $P(|t| > t_\alpha) = 2P(t > t_\alpha)$. What are two-tails? Please discuss it with other students. Can you appraise the values of t in the table mentioned above and the values returned by this function?

Degrees of freedom as defined earlier is defined as $\nu = n - 1$, where n is the number of observations under consideration.

For more details on this function, you may refer to the spreadsheet package that you are using.

Now, you are ready to solve the problem.

(i) Shaded portion of the right (α) = 0.05.
Since the spreadsheet function takes probability in the form of two tails, the total shaded portion will be twice the shaded portion on the right. Thus, probability = 2 × 0.05 = 0.1 (it is calculated in cell F6 of Figure 7), the degree of freedom is given to be 9 (Cell F7 of the Figure 7). Thus, you insert the formula ***=TINV(F6, F7)*** in cell F9. Verify the result with the Example's result as-well-as the table value.

(ii) The total shaded area = 0.05.

The total shaded area is same as the probability, so you just need to insert this probability in cell G6, put 9 in G7 and formula ***=TINV(G6, G7)*** in cell G9

(iii) The total un-shaded area = 0.99.

Calculate the total shaded area and insert the formula for t in cell H9.

(iv) The shaded area on the left = 0.01.

This area is same as area in the right. So insert the required formulas.

Figure 7 shows these results. Now, let us solve the problem 4 of the same section using the spreadsheet

Problem 4: of the same Unit stated above

Given: Sample size and t-value; find the probability.

(i) n = 26, t = 2.485

(ii) n = 14, t = 1.771

Solution:
To solve this problem, you need to use a function from the spreadsheet package:

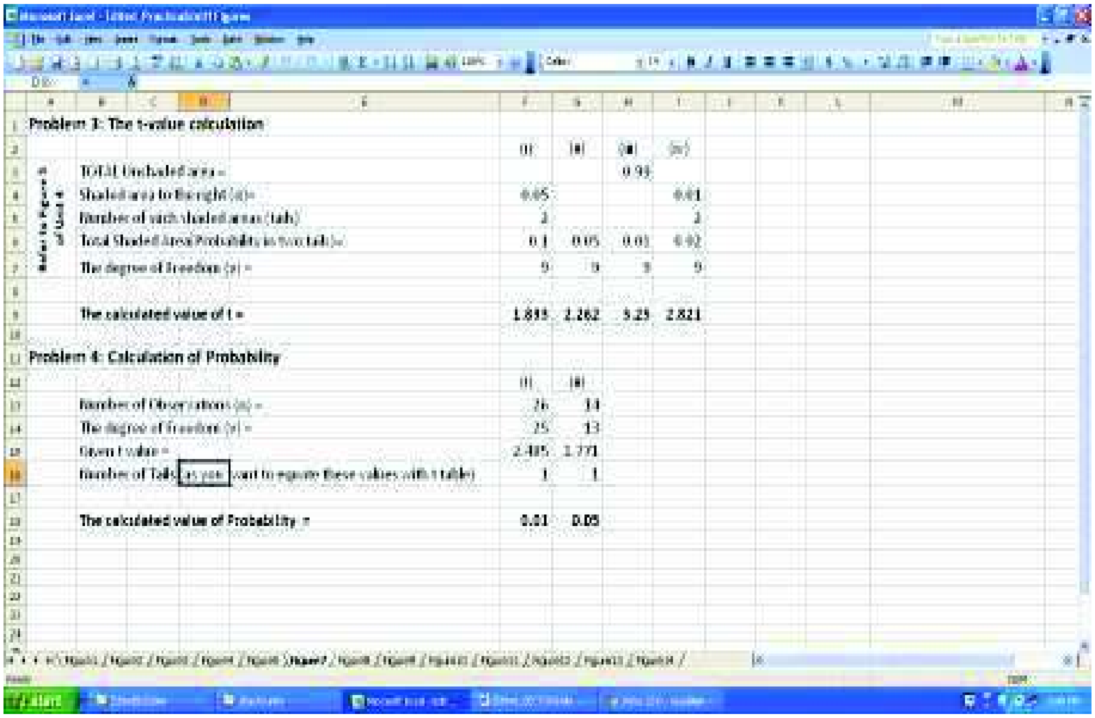
TDIST(x, degreesOFfreedom, tails)

This function returns the percentage points (Probability) of the student t-distribution at a t-value specified by x. You may be able to define the other two parameters. You may refer to spreadsheet package help for more details.

(i) n = 26, t = 2.485 and

(ii) n = 14, t = 1.771

Please check in Figure 7, cells F13 to F16 and G13 to G16, the values have been entered respectively. You can Insert the formula ***=TDIST(F15, F14, F16)*** in cell F18 and copy it to G18. You will get the required probabilities.



		(i)	(ii)	(iii)	(iv)
Problem 3: The t-value calculation					
Refer to Figure 3 of Unit 4	TOTAL Unshaded area =			0.99	
	Shaded area to the right (α)=	0.05			0.01
	Number of such shaded areas (tails)	2			2
	Total Shaded Area/Probability in two tails =	0.1	0.05	0.01	0.02
	The degree of freedom (ν) =	9	9	9	9
	The calculated value of t =	1.833	2.262	3.25	2.821
Problem 4: Calculation of Probability					
		(i)	(ii)		
	Number of Observations (n) =	26	14		
	The degree of freedom (ν) =	25	13		
	Given t value =	2.485	1.771		
	Number of Tails (as you want to equate these values with t table)	1	1		
	The calculated value of Probability =	0.01	0.05		

Figure 7: t-distribution (Problem 3 and Problem 4)

Hypothesis testing using t –statistic

Procedure for Mean: (See Book 3 Chapter 7)
To test for mean $H_0: \mu = \mu_0$ against an alternative

$H_1: \mu > \mu_0$, Calculate t_{cal} and Reject H_0 if $t_{cal} > t_{tab}$
$H_1: \mu < \mu_0$, Calculate t_{cal} and Reject H_0 if $t_{cal} < t_{tab}$
$H_1: \mu \neq \mu_0$, Calculate $|t_{cal}|$ and Reject H_0, if $|t_{cal}| > t_{tab}$.

Example 4 of Course BCS-040/Block2/Unit4/Section4.4/page 20: We briefly describe below the procedure:

Given:
Sample size of fuses $n = 20$, this sample is subjected to 20% overload
The expected average time of blowing of fuses (μ) 12.40
The average time of the sample = 10.63 minutes with Standard deviation 2.48 minutes.

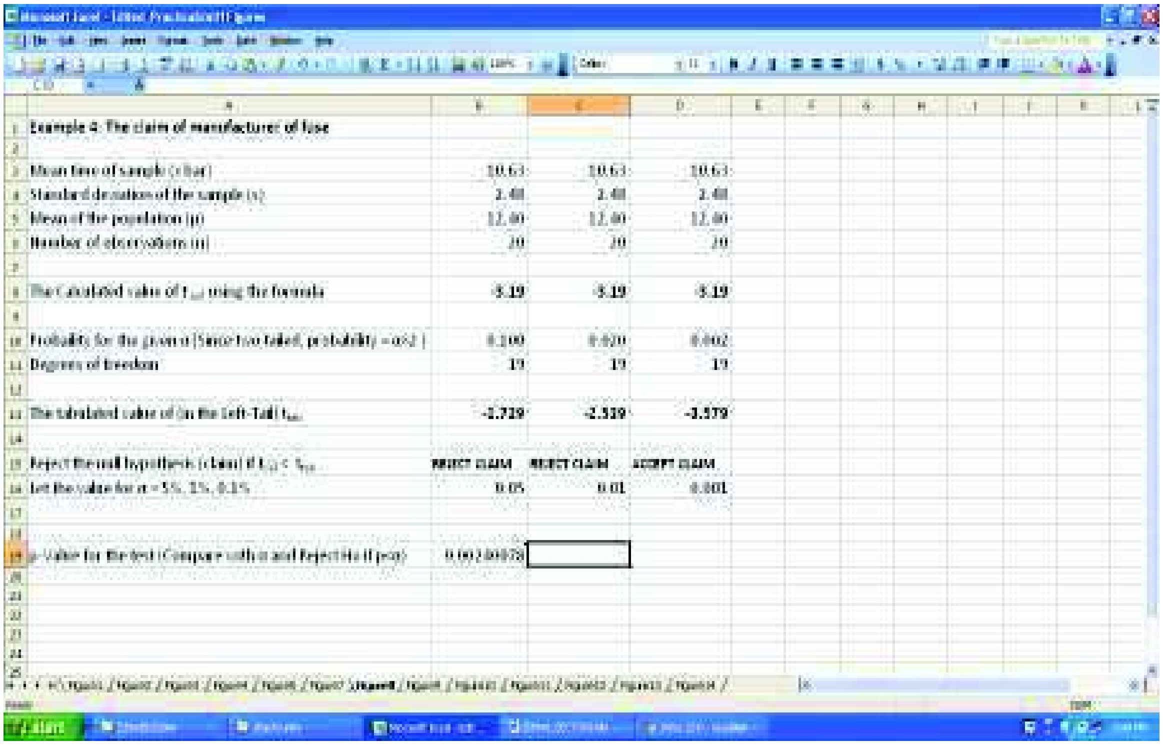
Claim: The fuse *will blow in 12.40 minutes on an average* with 20% overload.
In other words, you can state null hypothesis as:

H_0: The average time for the fuse to blow with 20% overload is 12.40 minutes.
H_1: The average time for the fuse to blow with 20% overload is below 12.40 minutes.
or simply stated as
$H_0: \mu = 12.40$ minutes
$H_1: \mu < 12.40$ minutes

The solution to the problem is discussed in the Unit. We can also use spreadsheet to solve this problem as shown in Figure 8. Insert all the values in the spreadsheet and calculate the t-value. By now you are familiar about the functions that are used in the spreadsheet.

In the Figure 8, you may notice that the tabulated value of t_{tab} is calculated (in cells B13, C13 and D13) using the spreadsheet function:

TINV(probability, degreesOFfreedom)



	A	B	C	D
1	Example 4: The claim of manufacturer of fuse			
2				
3	Mean time of sample (x bar)	10.63	10.63	10.63
4	Standard deviation of the sample (s)	2.40	2.40	2.40
5	Mean of the population (μ)	12.40	12.40	12.40
6	Number of observations (n)	20	20	20
7				
8	The Calculated value of t_{cal} using the formula	-3.19	-3.19	-3.19
9				
10	Probability for the given α (Since two tailed, probability = α×2)	0.100	0.020	0.002
11	Degrees of freedom	19	19	19
12				
13	The tabulated value of (in the Left-Tail) t_{tab}	-1.729	-2.539	-3.579
14				
15	Reject the null hypothesis (claim) if $t_{cal} < t_{tab}$	REJECT CLAIM	REJECT CLAIM	ACCEPT CLAIM
16	Let the value for α = 5%, 1%, 0.1%	0.05	0.01	0.001
17				
18				
19	p-Value for the test (Compare with α and Reject H0 if p<α)	0.002400778		

Figure 8: A sample Hypothesis Testing using t values

The calculation of statistic t_{cal} is done using the formula:

$$t = \frac{\overline{X} - \mu}{S/_{\sqrt{n}}}$$

Please note that the inference in this case is drawn for various α values. You should also notice the Standard error mentioned in the diagram. Please relate the standard error to Figure 6.

The claim of the company has been Rejected as you can notice that $t_{cal} \leq -t_{tab}$, which is the condition to Reject the Null Hypothesis.

Further, the test can also be carried out directly using p -value of the test, which in this case is $p = Pr(t \leq t_{cal}|H_0{:}\mu = 12.40) = 0.002400778$, which is mentioned in the last line of paragraph above Figure 6 of BCS-040/Block2/Unit4. Note that the typical values of level of significance used in practice are $\alpha = 5\%, 1\%, 0.1\%$, and the decision in the last case being different (Acceptance) in the present case.

1.7 CHI-SQUARE DISTRIBUTION

The Chi-square Distribution is defined in the BCS-040/Block 2/Unit 4/ Section 4.5 on page 22.

Suppose $X_1, X_2, \cdots X_n$ be a random sample drawn from the normal distribution $N(\mu, \sigma^2)$. Then distribution of the statistic

$$\chi^2 = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

is called chi-square (χ^2) distribution with $\nu = n - 1$ degree of freedom.

- Chi-square is always positive
- Chi-square distribution is not symmetrical and is skewed
- As ν increases, the shape of distribution approaches the shape of the normal curve
- Table 3 in Unit 7 gives probability $Pr(\chi^2 > \chi^2_\alpha)$, where χ^2 is distributed as chi-square with ν d.f.

Example 5: (*Refer page 23 of the Unit* stated above) you need to find the value of χ_α^2 of the (χ^2) distribution for α = 0.05 and 0.01 for the degree of freedom 5.

Solution:

In order to solve the problem using spreadsheet package, use the function ***CHIINV***(Probability, DegreeOfFreedom), that returns the value of χ^2 for given value of probability and d.f. ν . The solution of this problem is shown in Figure 9. The value of probability and degree of freedom are in cells F3 and F4 respectively for the first case, and G3 and G4 for the second case. The formula entered for calculating chi-square value in cell F6 is: *=CHIINV(F3, F4).* This formula is then copied to cell G6.

Problem 5: (*Refer page 23 of the Unit* stated above)

Variance of the refractive index of glass (expected) = 1.26 × 10^{-4}
Sample size = 20, so degree of freedom = 20-1 = 19
The firm rejects any sample having a variance higher than 2.0 × 10^{-4}

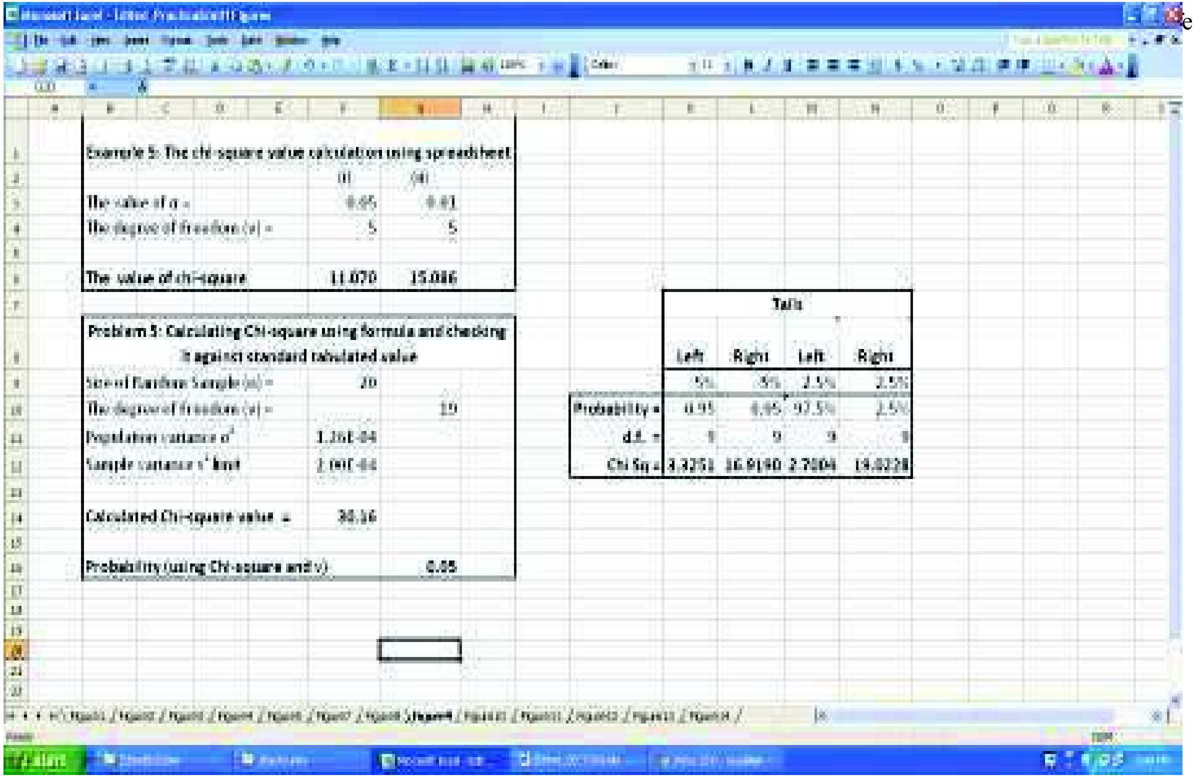
What is the probability that a shipment be rejected even through the variance in the population is 1.26 × 10^{-4}?

Solution:
First please note (as stated in the solution given in the Unit) that the sample is taken from a normal population having the variance σ^2 . The sample has a variance of S^2 and the size of sample is small ($n \le 30$), then chi-square value can be calculated using

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

You can now use this formula to calculate the value of chi-square and check it against the tabulated values (you can use spreadsheet CHIINV function to calculate the tabulated values). Figure 9 shows the details of the calculations. In the cells F9, F11 and F12 size of the sample, the values of variance of population and value of variance of sample has been entered. The cell F14 contains the formula *=((F9-1)*F12)/F11* that calculates the value of chi-square to be $\chi^2 = 30.16$. To calculate the probability, we first calculated the degree of freedom in cell G10 and using the function CHIDIST(x,DegreeOfFreedom), where x is the value of chi-square on which probability is to be calculated, you can calculate the probability of rejecting a valid sample as

$Pr(\chi^2 > 30.16|\sigma^2 = 1.26 \times 10^{-4})$. Thus, you can enter the formula ***=CHIDIST(F14,G10)*** in the cell G16. The calculated probability is 0.05 which means there are 5% chances that a due to our method of sampling, we reject a shipment that has variance of the refractive index of 1.26 × 10^{-4}.



Example 5: The chi-square value calculation using spreadsheet		
	(i)	(ii)
The value of α =	0.05	0.01
The degree of freedom (ν) =	5	5
The value of chi-square	11.070	15.086

Problem 5: Calculating Chi-square using formula and checking it against standard tabulated value		
Size of Random Sample (n) =	20	
The degree of freedom (ν) =		19
Population variance σ²	1.26E-04	
Sample variance s² kmt	2.00E-04	
Calculated Chi-square value =	30.16	
Probability (using Chi-square and ν)		0.05

	Tails			
	Left	Right	Left	Right
	5%	5%	2.5%	2.5%
Probability =	0.95	0.05	97.5%	2.5%
d.f. =	9	9	9	9
Chi Sq =	3.3251	16.9190	2.7004	19.0228

Figure 9: Using chi-square distribution

Hypothesis testing using χ^2 –statistic

For Variance: (See Unit 6/Page 66, Book 3/Chapter 7)
Statistic:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

To test for variance $H_0:\sigma^2 = \sigma_0^2$ against an alternative
$H_1:\sigma^2 > \sigma_0^2$, Calculate χ^2_{cal} and Reject H_0 if $\chi^2_{cal} > \chi^2_{\alpha,n-1}$
$H_1:\sigma^2 < \sigma_0^2$, Calculate χ^2_{cal} and Reject H_0 if $\chi^2_{cal} < \chi^2_{1-\alpha,n-1}$
$H_1:\sigma^2 \neq \sigma_0^2$, Calculate χ^2_{cal} and Reject H_0, if $\chi^2_{cal} < \chi^2_{1-\alpha/2,n-1}$ or $\chi^2_{cal} > \chi^2_{\alpha/2,n-1}$.

Example: Let $n = 10$, $\alpha = 0.05$, $H_0:\sigma^2 = 16$ against an alternative

- $H_1:\sigma^2 < 16$, Reject H_0 if $\chi^2_{cal} < \chi^2_{0.95,9} = 3.3251$
- $H_1:\sigma^2 > 16$, Reject H_0 if $\chi^2_{cal} > \chi^2_{0.05,9} = 16.9190$
- $H_1:\sigma^2 \neq 16$, Reject H_0, if $\chi^2_{cal} < \chi^2_{0.975,9} = 2.7004$ or $\chi^2_{cal} > \chi^2_{0.025,19} = 19.0228$

Computation of these values are shown in cells ***K9..N12*** of Figure 9 using worksheet function ***CHIINV***.

For Goodness of Fit: (See Unit 7/Page 76, Book 3/Chapter 7)
Statistic:

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i}$$

where O_i 's are the observed number of values in the i th class and E_i 's are expected number of values in the same class. Then the approximate distribution of χ^2 is Chi-square with $k-1$ degrees of freedom. Example based on this will be discussed in section 1.10.

1.8 F- DISTRIBUTION

The F- Distribution is defined in the BCS-040/Block 2/Unit 4/ Section 4.6 on page 24. Before starting with this section, please read BCS-040/Block2/Unit6/Section 6.3.2/page 59, in particular page 64.

Recall that in applying t -test to test the hypothesis of the type $H_0: \mu_1 = \mu_2$, against $H_1: \mu_1 \neq \mu_2$ etc., it was assumed that the samples were drawn from normal populations with means μ_1, μ_2 and equal variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$, which is unknown. Note that this equality of variances, which is required can be tested using F -statistics/distribution.

Suppose $X_{11}, X_{12}, \cdots X_{1n_1}$ and $X_{21}, X_{22}, \cdots X_{2n_2}$ be two independent random samples drawn from normal distribution $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively. Then the statistic

$$F = \frac{\chi_1^2/v_1}{\chi_2^2/v_2}$$

has F -distribution with v_1 and v_2 degrees of freedom it is denoted as F_{v_1,v_2}. The random variable F can also be defined as the ratio of two independent chi-square random variables, each divided by their respective number of degree of freedom. Here $v_1 = n_1 - 1$, $v_2 = n_2 - 1$ and

$$\chi_1^2 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \quad \text{and} \quad \chi_2^2 = \frac{(n_2-1)S_2^2}{\sigma_2^2}$$

- F is always positive
- F -distribution is not symmetrical and is skewed
- In order to decide which of the two samples is first or the second, we *take larger of the two quantities in the numerator* and the smaller one in the denominator.
- Table 4&5 in Unit 7 gives probability $Pr(F > F_{\alpha,v_1,v_2})$, and the values so tabulated are greater than unity. Worksheet function to be used $FDIST(x,,v_1,v_2)$.
- $F_{1-\alpha,v_2,v_1} = \frac{1}{F_{\alpha,v_1,v_2}}$, in order to get value of F for given α, v_1, v_2, use worksheet function $FINV(\alpha, v_1, v_2)$

You may also refer to Book 3/Chapter 4.

Example 6: (Refer to page 24 of the Unit stated above) A random variable has F distribution with 40 and 30 degrees of freedom, Find the probability that it will exceed (i) 1.79 and (b) 2.30.

Solution:

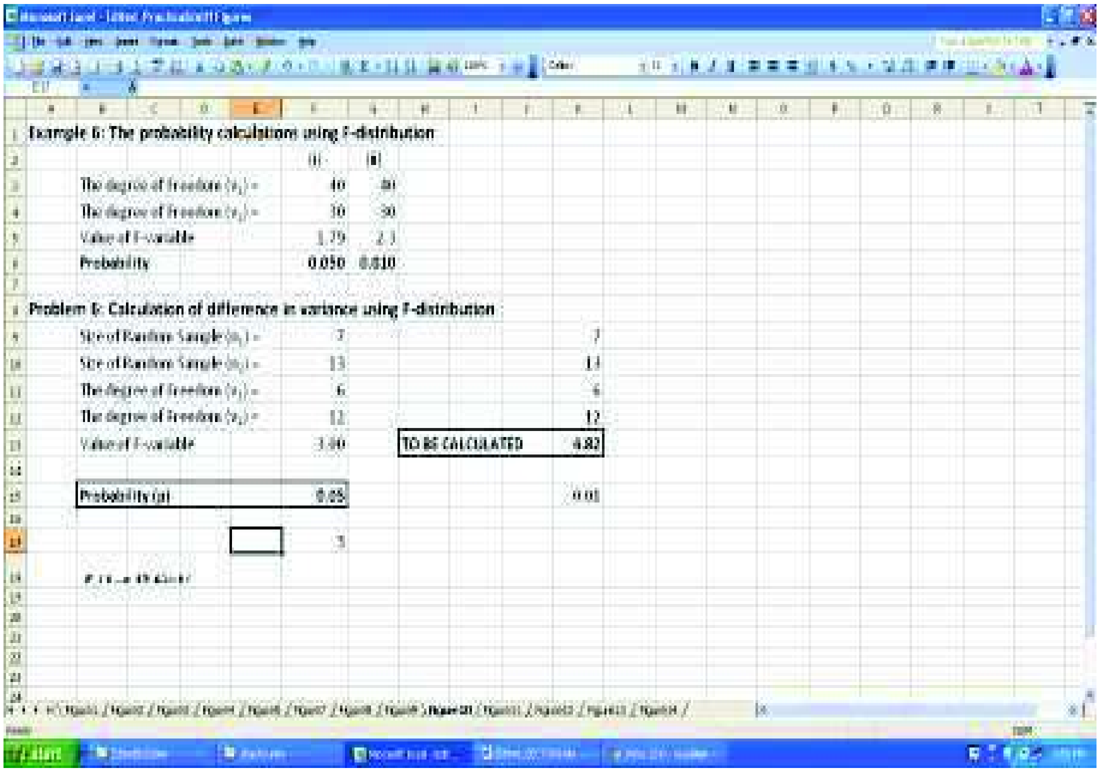
To find the probability for the given value of random variable that has F -distribution, you need to use the function FDIST(x, DegreeOfFreedom1, DegreeOfFreedom2), where x is the value of random variable on which probability is to be calculated. You can insert the data as shown in Figure 10, and enter the formula *=FDIST(F5,F3,F4)* in the cell F6. Copy this formula to G6 to get $Pr(F > 1.79) = 0.05$ and $Pr(F > 2.30) = 0.10$. You can compare the results that you have obtained using the spreadsheet function and the example of the Block or the table lookup value.

Problem 6: (Refer to page 24 of the Unit stated above) If two independent random sample of size $n_1 = 7$ and $n_1 = 13$ are taken from a normal population. What is the probability that the variance of the first sample will be at least three times as large as that of second sample?

Solution:

You can easily find the number of degree of freedom for the value of random variable and calculate the value of random variable under F-distribution as 3.00 (variance of first sample is triple of the second). You can once again use the function FDIST as explained above.
However, in the figure 9 you will find that we have also used a formula of *=FINV(K15,K11,K12)* in cell K13. The purpose of this function is to calculate the value of F for given probability and degrees of freedom. The function in spreadsheet is FINV(probability, DegreesOfFreedom1, DegreesOfFreedom2).

Further, notice that the expression $F_{1-\alpha,12,6} = \dfrac{1}{F_{\alpha,6,12}}$ for $\alpha = 0.05$ (in cell F15) is verified in cell F17.



Row	Label	F	G	H	K
1	Example 6: The probability calculations using F-distribution				
2		(i)	(ii)		
3	The degree of freedom (v_1) =	40	40		
4	The degree of freedom (v_2) =	30	30		
5	Value of F-variable	1.79	2.3		
6	Probability	0.050	0.010		
7					
8	Problem 6: Calculation of difference in variance using F-distribution				
9	Size of Random Sample (n_1) =	7			7
10	Size of Random Sample (n_2) =	13			13
11	The degree of freedom (v_1) =	6			6
12	The degree of freedom (v_2) =	12			12
13	Value of F-variable	3.00		TO BE CALCULATED	4.82
14					
15	Probability (α)	0.05			0.01
16					
17		3			
18	[illegible]				

Figure 10: Using F-Distribution

1.9 TEST OF SIGNIFICANCE

As discussed in the section 1.5, an important aspect of statistical data analysis is to draw inferences and conclusion concerning the problem under investigation based on sample being drawn from the population. Let us reiterate the definition of the test of significance here: **The test of significance is a formal procedure that is aimed at assessing evidence that is provided by sample(s) data in favor of some claim/inference about the entire population.** For details on test of significance, use of chi-square testing you must go through the BCS040/Block2/Unit 6. The following are the steps, in general, in performing the test of significance:

Step 1: State H_0 and H_1 and select a significance level α
Step 2: Specify and select a sample
Step 3: Calculate desired Statistics
Step 4: Calculate the p -value
Step 5: If value of $p \leq \alpha$, Reject Null hypothesis and conclude H_1 may be true.
Otherwise, we conclude that sample does not provide sufficient evidence against H_0, thus, we do not reject H_0

Recapitulation: Based on what has been covered above, the following should be useful in deciding a test statistic:

- Based on Large sample:
  - Use Z -test in testing a hypothesis concerning means such as $H_0: \mu = \mu_0$ or $H_0: \mu_1 = \mu_2$ (Unit 6/page 53/Section 6.3, Book 3/Chapter 7) and
  - Use Z -test in testing a hypothesis concerning proportions such as $H_0: \pi = \pi_0$ or $H_0: \pi_1 = \pi_2$, (Unit 6/page 67/Section 6.5, Book 3/Chapter 7)

- Based on Small sample$(n < 30)$: (see Unit 6)
  - Use t -test in testing a hypothesis concerning means such as $H_0: \mu = \mu_0$ or $H_0: \mu_1 = \mu_2$ (in both cases of two independent samples; paired and not independent samples)
  - Use t -test in testing a hypothesis concerning proportions such as $H_0: \pi = \pi_0$ or $H_0: \pi_1 = \pi_2$, (Unit 6/page 67/Section 6.5, Book 3/Chapter 7)
  - Use χ^2 -test in testing a hypothesis concerning variance such as $H_0: \sigma^2 = \sigma_0^2$.
  - Use F -test in testing a hypothesis concerning variance such as $H_0: \sigma_1^2 = \sigma_2^2$.

The following two examples describe the use of these tests using a spreadsheet package:

Problem 9: (Refer Unit 6, page 65):

The table givens the data on the reading speed of 10 students

Before	9.4	10.3	8.4	6.8	7.8	9.8	9.2	11.2	9.4	9.0
After	9.3	10.6	8.8	7.0	7.7	10.0	9.8	11.7	9.7	9.0

Can you say that the reading course is a success at 0.05 level of significance?

Solution:
In order to judge effectiveness of the reading course, we denote reading speed "before" by X_{1i}, "after" by X_{2i} and the difference by $D_i = X_{1i} - X_{2i}(\text{i.e., Before} - \text{After})$. Then apply t -statistic to test the null hypothesis.

Note:

1. Since in this example, the question being investigated through statistical test of hypothesis is: "Can you say that the reading course is a success?", and it can be considered to be a success in the case of higher average reading speed among students after undergoing the course. In other words, we are effectively testing $\overline{D} < 0$.
2. Consequently, we will apply a one-tail (left-tail) t -test to test the hypothesis.
3. If we start with a definition $D_i = X_{2i} - X_{1i}(\text{i.e., After} - \text{Before})$, we can reproduce the solution discussed in Unit 6, page 65-66 with $H_1 : \overline{D} > 0$.

Here, the hypothesis to be tested is:

H_0: No difference in the mean reading speed before and after undergoing the course i.e., there is no difference in students reading abilities before and after the course, or $\overline{D} = 0$

H_1: The mean reading speed after the course is greater than before undergoing the course i.e., the students reading course is a success, or $\overline{D} < 0$.

You can apply the Left-Tail t -test to test the hypothesis using spreadsheet package. Figure 11 shows the application of "t-test: Paired Two Sample for Means" test on the data. You can apply this test in two ways in the spreadsheet.

- *Select Data→Data Analysis→ t-test*: *Paired Two Sample for Means*
- In the resulting dialog box enter the two ranges of data and output range as shown in Figure 11.
- Here we have used ranges as:
  - Variable 1 Range: $B$2:$K$2
  - Variable 2 Range: $B$3:$K$3
  - Output Range: $A$5
  - Leave the alpha (significance level) as 0.05

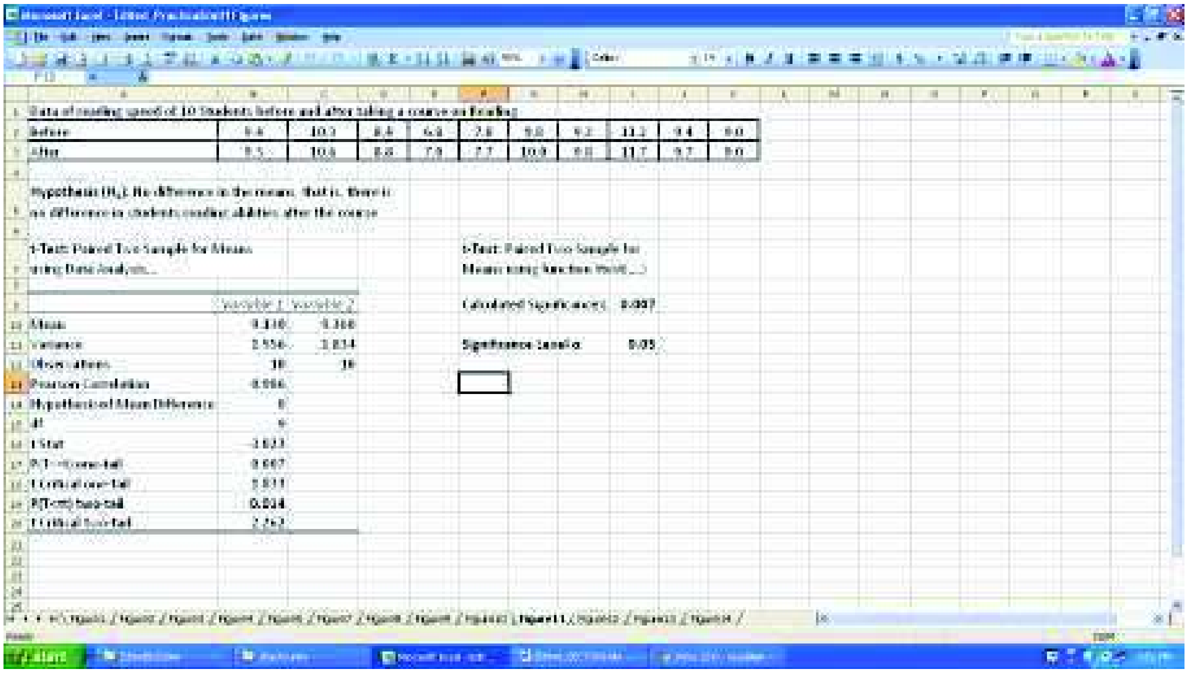
On applying the test, the results so generated are shown in Figure 11. Some of the terms that are of importance in the present case are: *mean, variance and number of observations* for both the samples, *number of degrees of freedom* (d.f.) and p *-value* and t *-value* for one or two tails.

Decision: Here, $t_{cal} = -3.023$ and $t_{1-0.05} = -t_{0.05} = -1.833$.
Since $t_{cal} < -t_{0.05}$, we Reject the Null Hypothesis and conclude that the reading course is a success. Here, p -value for the test is $p = Pr(t \leq t_{cal} | H_0) = 0.007 < 0.05$.

The second way is by using the worksheet function:

TTEST(array1,array2, tails, type); type is the type of t-test, help on which can be found from spreadsheet help.

The result shown in Figure 11 has been generated using the formula **=*TTEST(B2:K2,B3:K3,1,1)*** in cell H9, which returns $p = 0.007$ in cell I9 (*Please Read Note above for explanation on arguments "1,1"*).



Data of reading speed of 10 Students before and after taking a course on Reading

Before	9.4	10.3	8.4	6.9	7.8	9.0	9.2	11.2	9.4	8.0
After	9.5	10.6	8.8	7.8	7.7	10.9	9.0	11.7	9.7	8.0

Hypothesis (H_0): No difference in the means, that is, there is no difference in students reading abilities after the course

t-Test: Paired Two Sample for Means using Data Analysis...

	Variable 1	Variable 2
Mean	9.110	9.360
Variance	2.556	2.834
Observations	10	10
Pearson Correlation	0.986	
Hypothesized Mean Difference	0	
df	9	
t Stat	-3.023	
P(T<=t) one-tail	0.007	
t Critical one-tail	1.833	
P(T<=t) two-tail	0.014	
t Critical two-tail	2.262	

t-Test: Paired Two Sample for Means using function TTEST(...)

Calculated Significance	0.007
Significance Level α	0.05

Figure 11: An example of use of t-test

Note:

- In both the cases you have got p -value (Left-tailed) as 0.007, which is less than the significance level of 0.05. Hence hypothesis H_0 is REJECTED. This implies that there is a significant difference/improvement in the average reading ability of the students and hence the reading course is a success.
- Extreme caution should be exercised in deciding the direction of the rejection region (left, right or 2-tail) and this has a bearing from the basic question being investigated.
- Conclusion of the test should pertain to the basic question being investigated.

1.10 APPLICATION OF CHI-SQUARE TEST

Chi-square testing is used in two important types of test: finding the goodness of fit and performing the test of independence/ For more details on these tests using chi-square testing you must go through the BCS040/Block2/Unit 7. In this section, we present few examples of these tests using spreadsheet.

Test for Goodness of fit

Example 1: (Refer to Unit 7 page 77) Jaswant is interested in breeding flowers of a certain species. The experimental breeding can result in four possible types of flowers:

a) Magenta flowers with a green stigma (MG)
b) Magenta flowers with a red stigma (MR)
c) Red flowers with a green stigma (RG)
d) Red flowers with a red stigma (RR)

As per Mendel's law, the ratios of these flowers are 9:3:3:1.

Jaswant found that under her experiment, out of the 160 flowers that bloomed, number of flowers of the four types are:

MG has 84
MR has 35
RG has 28
RR has 13

Check if this data is compatible with Mendel's law or not.

Solution: The hypothesis to be tested is
H_0: The distribution of the flower types is multinomial with ratio 9:3:3:1
H_1: The distribution is not as per the given ratio.

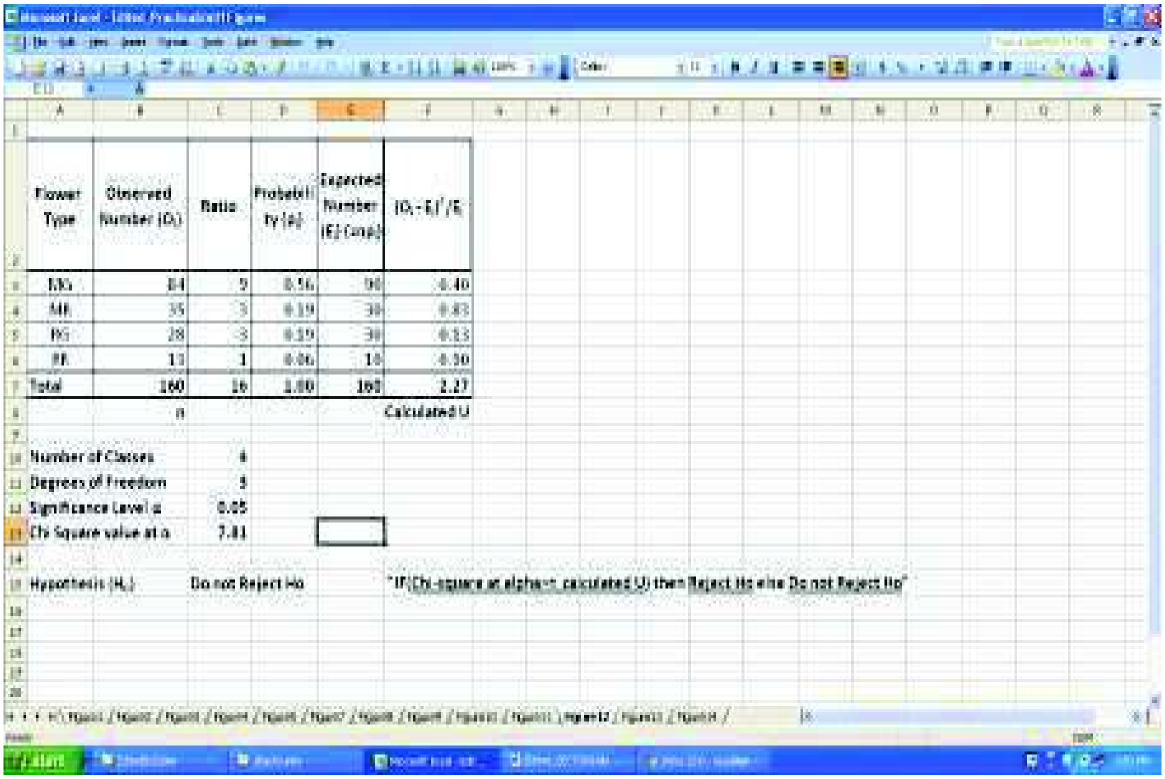
Figure 12 shows the goodness of fit test carried out for the data and the expected theoretical model. In the cell C13, you can enter the formula for calculating the tabulated value of chi-square using the function ***CHIINV***(Probability,DegreeOfFreedom) which is: =***CHIINV(C12,C11)***

The hypotheses of goodness of fit can be tested using the statement in cell C15:
IF(Chi-square at alpha >= calculated U) then Reject H0 ***else*** do not reject H0

You can enter the related formula to check if hypothesis is rejected or not ***Result*:**

Calculated Value of	**U**	**= 2.27,**
Critical value (at 5%)		**= 7.81**
Decision		**Do not Reject Ho**

Let us discuss another example for testing goodness of fit using chi-square statistic in the case of fitting of normal distribution to the given data, which is based on Problem 1/Unit 7/page 80 (also see Book 3/chapter 7,page 7.33 and page P.12 & P.17).



Flower Type	Observed Number (O_i)	Ratio	Probability (p_i)	Expected Number (E_i) (=np)	$(O_i - E_i)^2/E_i$
MG	84	9	0.56	90	0.40
MR	35	3	0.19	30	0.83
RG	28	3	0.19	30	0.13
RR	13	1	0.06	10	0.90
Total	160	16	1.00	160	2.27
	n				Calculated U

Number of Classes	4
Degrees of Freedom	3
Significance Level α	0.05
Chi Square value at α	7.81

Hypothesis (H_0) Do not Reject Ho "IF(Chi-square at alpha>= calculated U) then Reject Ho else Do not Reject Ho"

Figure 12: Chi-square test for goodness of fit

Problem 1: (Refer to Unit 7 page 80) A chemical company wants to know if its sales of a liquid chemical are normally distributed. This information will help them in planning and controlling the inventory. The sales record for a random sample of 200 days is given in the following Table:

Sales (in 1000 litres)	Number of Days
Less than 34.0	0
34.0-35.5	13
35.5-37.0	20
37.0-38.5	35
38.5-40.0	43
40.0-41.5	51
41.5-43.0	27
43.0-44.5	10
44.5-46.0	1
46.0 or more	0

Assume that the upper limit of a class shows that quantities less than that limit are in the class. So, for example, 35.5 will be included in the third class interval, not the second one.

At the 5% level of significance, test the hypothesis that the company's sales are normally distributed.

Solution : H_0 : The company's sales are normally distributed.
H_1 : The company's sales are **not** normally distributed.

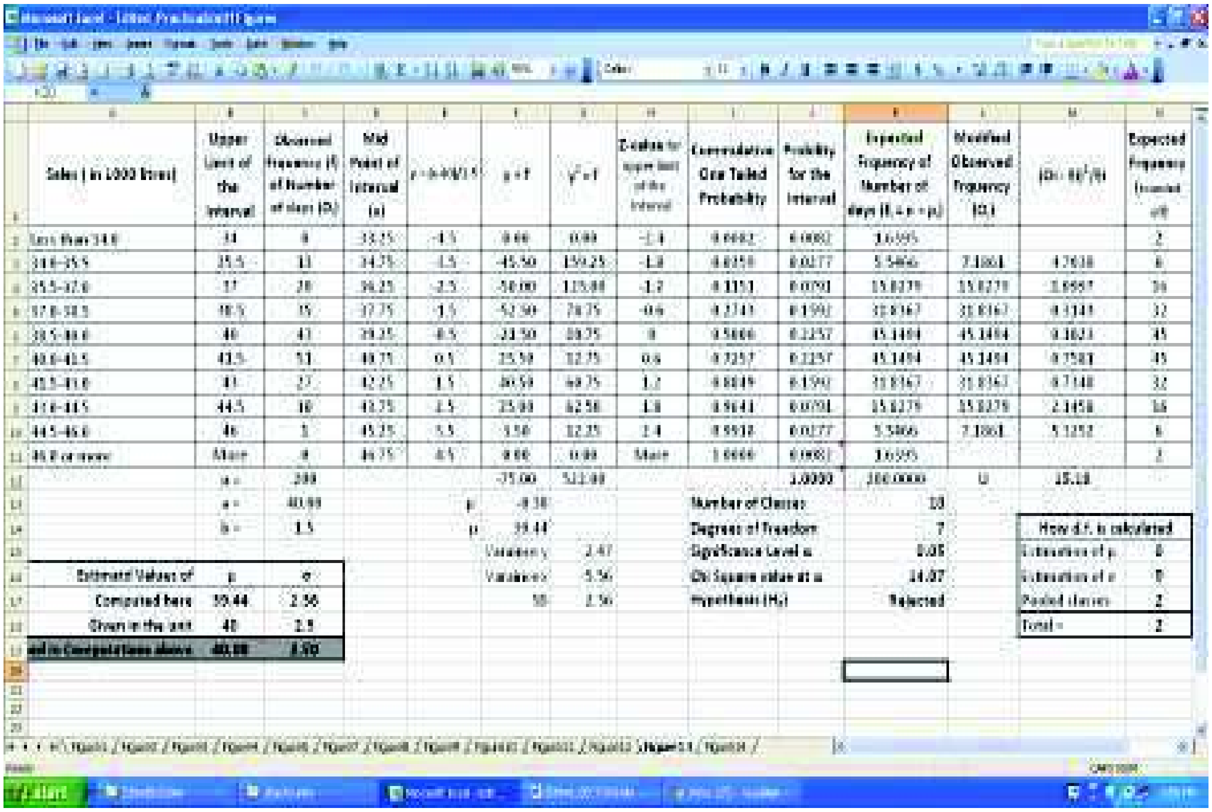
Figure 13 gives the worksheet dealing with the two problems viz.,(i) fitting of normal distribution and hence (ii) test for goodness of fit. The first step in this exercise is to locate if the values of parameters mean (μ) and standard deviation (σ) have been specified in the problem. If they are not specified, we proceed to calculated/estimate their values from the given data. This step will enable us to decide the degrees of freedom for the Chi-square test, as it has to be reduced by 1 for each parameter being estimated and otherwise not. Here, for the purpose of demonstration we have used the estimate value of Unit 7 as the specified values of population mean 40 thousand litters (in cell B19) and standard deviation 2.5 thousand litters (in cell C19). So, the d.f. is not reduced. Subsequent calculations related to the problem shown in the worksheet are self-explanatory.

The spreadsheet calculates the mean and standard deviation based on the formulas given on BCS040/Block 1/Unit 1 Page 30and 31. Please enter the formulas yourself. You can also enter the formulas for calculating U as shown in Figure 13.

Conclusion:

- Since the calculated value of U is greater than the chi-square tabulated value, we Reject H_0 and conclude that normal distribution does not provide a good fit for the data.
- Further, note that if we use the estimated values of μ and σ (enter values in cells B19 and C19).
    - The d.f. has to be reduced by 1 for each parameter (enter 1 in cells N15 and N16).

- Notice the expected frequencies and pool the classes appropriately (for frequency less than 5). Hence, enter the appropriate adjustment to d.f. in cell N17. Is there any change to the d.f.?
- What is the conclusion now?



Sales (in 1000 litres)	Upper Limit of the Interval	Observed Frequency (f) of Number of days (Oi)	Mid Point of Interval (x)	u = (x-40)/1.5	u * f	$u^2 * f$	Z-value for upper limit of the interval	Cummulative One Tailed Probability	Probability for the interval	Expected Frequency of Number of days ($E_i = n * p_i$)	Modified Observed Frequency (O_i)	$(O_i - E_i)^2/E_i$	Expected Frequency (rounded off)
Less than 34.0	34	0	33.25	-4.5	0.00	0.00	-2.4	0.0082	0.0082	1.6395			2
34.0-35.5	35.5	13	34.75	-3.5	-45.50	159.25	-1.8	0.0359	0.0277	5.5466	7.1861	4.7038	6
35.5-37.0	37	20	36.25	-2.5	-50.00	125.00	-1.2	0.1151	0.0791	15.8279	15.8279	1.0997	16
37.0-38.5	38.5	35	37.75	-1.5	-52.50	78.75	-0.6	0.2743	0.1592	31.8367	31.8367	0.3143	32
38.5-40.0	40	43	39.25	-0.5	-21.50	10.75	0	0.5000	0.2257	45.1494	45.1494	0.1023	45
40.0-41.5	41.5	51	40.75	0.5	25.50	12.75	0.6	0.7257	0.2257	45.1494	45.1494	0.7581	45
41.5-43.0	43	27	42.25	1.5	40.50	60.75	1.2	0.8849	0.1592	31.8367	31.8367	0.7348	32
43.0-44.5	44.5	10	43.75	2.5	25.00	62.50	1.8	0.9641	0.0791	15.8279	15.8279	2.1458	16
44.5-46.0	46	1	45.25	3.5	3.50	12.25	2.4	0.9918	0.0277	5.5466	7.1861	5.3252	6
46.0 or more	More	0	46.75	4.5	0.00	0.00	More	1.0000	0.0082	1.6395			2
	n =	200			-75.00	522.00			1.0000	200.0000	0	15.18	

	a =	40.00
	h =	1.5

ū	-0.38
μ	39.44
Variance u	2.47
Variance x	5.56
SD	2.36

Estimated Values of	μ	σ
Computed here	39.44	2.36
Given in the unit	40	2.5
Used in Computations above	40.00	2.50

Number of Classes	10
Degrees of Freedom	7
Significance Level α	0.05
Chi Square value at α	14.07
Hypothesis (H0)	Rejected

How d.f. is calculated	
Estimation of μ	0
Estimation of σ	0
Pooled classes	2
Total =	2

Figure 13: Test of goodness of fit

Test of Independence

Chi-square test can also be used to investigate if there is an agreement between the observed frequencies and the expected frequencies or independence of attributes. See Unit 7/section 7.3 or Book 3/Chapter 7. The following example explains this in detail.

Example 4: (For more details refer to BCS040/Block2/Unit 7/Page 84)
The Glorious Watch Company wants to find out if there is any relationship between the income of a person and the importance she attaches to the price of a brand name. Mr.Zafar, the Chief of the Marketing Division, wants to test the hypothesis:

H_0 : Income of a person and importance to her of price attached are independent. against

H_1 : H_0 is not true or there is some dependence.

Solution: The Figure 14 gives the results computation relating to the test of hypothesis. The test has been done using two approaches –

- Calculations using the formula as given in the Unit 7
- Calculations using the ***CHITEST(…)*** function of the spreadsheet.

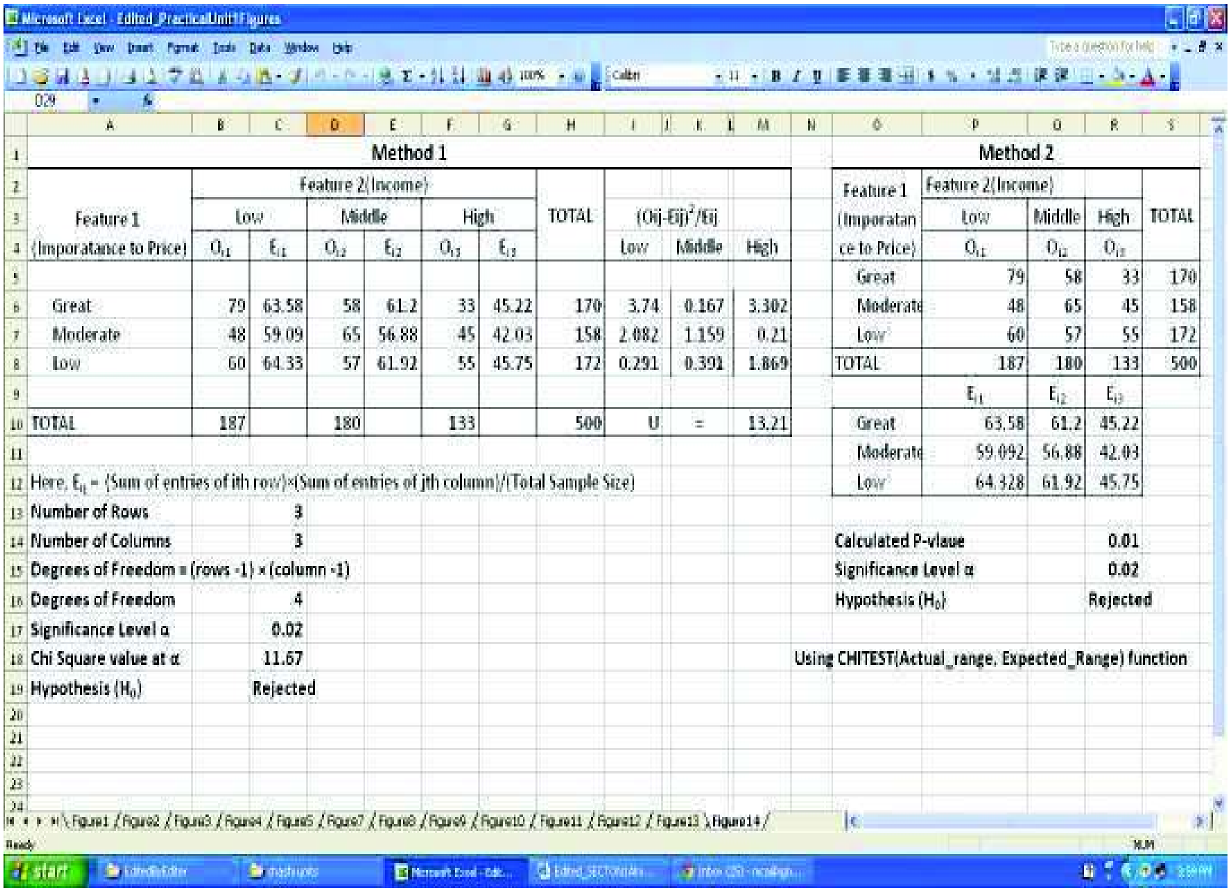
Remark:

- Details of all related computations for the two methods can be seen in the attached spreadsheet. You can enter all the formulas yourself.
- Notice that the notion of "range naming" in worksheets has been used here for the purpose of demonstration and it is explained in the steps below.
    - To name a range Click on *Formulas* →*Name Manager* →*New* →write *Name* →Specify *Refers to*

- ○ Then while writing a function in place of the range, you can write the named range from the list.
- ○ Here, the range has been named as "ObservedFREQ".

- Can you identify which other places in earlier worksheets could you have used the same?
- The "hypothesis is rejected" in both the methods and we conclude that *the income of a person is related to the importance she attaches to the price of a brand name.*
- For Method 1, please refer to Unit 7/page 83.

The result so generated helps us in the process of managerial decision making.


Microsoft Excel - Edited_PracticalUnit1Figures
File Edit View Insert Format Tools Data Window Help
Type a question for help
O29
Figure1 / Figure2 / Figure3 / Figure4 / Figure5 / Figure7 / Figure8 / Figure9 / Figure10 / Figure11 / Figure12 / Figure13 \ Figure14 /
Ready
NUM
start

Method 1

Feature 1 (Imporatance to Price)	Feature 2(Income)						TOTAL	$(Oij-Eij)^2/Eij$		
	Low		Middle		High					
	O_{i1}	E_{i1}	O_{i2}	E_{i2}	O_{i3}	E_{i3}		Low	Middle	High
Great	79	63.58	58	61.2	33	45.22	170	3.74	0.167	3.302
Moderate	48	59.09	65	56.88	45	42.03	158	2.082	1.159	0.21
Low	60	64.33	57	61.92	55	45.75	172	0.291	0.391	1.869
TOTAL	187		180		133		500	U	=	13.21

Here, E_{ij} = (Sum of entries of ith row)×(Sum of entries of jth column)/(Total Sample Size)

Number of Rows	3
Number of Columns	3
Degrees of Freedom = (rows -1) × (column -1)	
Degrees of Freedom	4
Significance Level α	0.02
Chi Square value at α	11.67
Hypothesis (H_0)	Rejected

Method 2

Feature 1 (Imporatance to Price)	Feature 2(Income)			TOTAL
	Low	Middle	High	
	O_{i1}	O_{i2}	O_{i3}	
Great	79	58	33	170
Moderate	48	65	45	158
Low	60	57	55	172
TOTAL	187	180	133	500
	E_{i1}	E_{i2}	E_{i3}	
Great	63.58	61.2	45.22	
Moderate	59.092	56.88	42.03	
Low	64.328	61.92	45.75	

Calculated P-vlaue	0.01
Significance Level α	0.02
Hypothesis (H_0)	Rejected

Using CHITEST(Actual_range, Expected_Range) function

Figure 14: Test of independence

☞ Lab Sessions 2, 3, 4

1) Develop all the spreadsheets as shown in various diagrams in these previous sections. Can you now make effective use of names for ranges instead of cell references?
2) Perform all the examples given in the Unit 4, 5, and 6 using spreadsheet package.
3) Perform the following Exercises of BCS040/Blockl2/Unit 4/ E4 (Page 10), E5 (Page 11), E6 (Page 13), E7 (Page 18), E8, E9, E10 (Page21), E11, E12 (Page 23) and E13, E14, E15, E16 (Page 25) using spreadsheet.
4) Write simple C functions to calculate the t and chi-square values from the formula that are given for their calculations.
5) Perform the exercises of BCS040/Block 2/Unit6 and Unit7, which can be implemented using spreadsheet. If not generate sample data for the tests.

1.11 SUMMARY

This section is an attempt to provide you details on some of the basic concepts of statistics along with how their computations can be carried out in a spreadsheet package. The attempt here is not only to introduce you to various steps that are used to perform the said statistical data analysis, but also to the spreadsheet functions that can be used in performing the same. The unit begins with frequency distribution, summarization of data using central tendency and dispersion and subsequently covers nature of sampling distribution and some important concepts such as p -value, significance level, normal distribution, array formula and finally range naming. The unit then explains how you can use spreadsheet to solve problems using t -distribution, chi-square distribution, F -distribution etc. Finally the unit describes the use of spreadsheet package for performing test of significance. The test of significance is explained with the help of an example on t -test. The last section of the unit is devoted to application of chi-square testing on testing *goodness of fit* and *independence of attribute*.

1.12 FURTHER READINGS

Books

1. Introduction to the Practice of Statistics, Fifith edition, 2004, David S. Moore, George P. McCabe, W.H. Freeman and Company, Newyork
2. BCS-040: Statistical Techniques, IGNOU Material.
3. Probability and Statistics (Schaum's Outlines Series), SIE, 2010, Murray R. Spiegel, John J. Schiller, R. Alu Srinivasan, Debasree Goswami, Tata McGraw-Hill Publishing Co. Ltd., New Delhi

Web link:

- www.wikipedia.org

SECTION 2 CORRELATION& REGRESSION

Structure **Page Nos.**

2.0 INTRODUCTION

The objective of this unit is to enable you to investigate the interdependence of variables in terms of Correlation and Regression analysis through hands-on experience in using MS-Excel and its tools viz., Data Analysis Tool pack. However, prior to this you should go through *BCS 040 Block 3 Unit 9, Regression Analysis*, which is a prerequisite.

Whenever you are going to conduct a study or experiment or research, irrespective of the discipline, it is desired to analyze the dependence of one variable over other(s). Correlation and Regression are used to analyse collected sample data to investigate the relationship between the variables to answer associated questions such as:

- Is there any relationship between the variables under study?
- How strongly the variables are related to each other?
- Can we use the relationship to estimate or forecast the value of one of the variables (dependent variable)?

This section provides a practical orientation in the light of your understanding of BCS 040.

2.1 OBJECTIVES

After going through this unit you will be able to perform:

- Correlation analysis through Excel;
- Multiple Correlation analysis through Excel;
- Linear Regression analysis through Excel; and
- Multiple Regression analysis through Excel.

2.2 CORRELATION

Let's start by describing the following simple example: In any computer system there are various components like Memory, Processor, Motherboard etc, and say you want to study how the performance of any computer system varies with different permutation and combination of some constituent components. For example, you may want to know "is there any relationship between the size of Random Access Memory - RAM and the performance of Computer system" OR "is there any relationship between Hard disk storage capacity and the performance of Computer system" etc. in order to address such queries, you are required to conduct a statistical experiment that entails in the collection and tabulate the data as discussed earlier. So, collect and tabulate data under heading say “RAM-Size” and “Performance-Status” OR ”Hard Disk-Capacity” and “Performance-Status”.

Thus, it is required to carry out a statistical investigation of the fact that a Computer System’s performance improves on increasing the RAM size. Since it is known that there is non-linear relationship between the two variables, a linear regression (which involves linear relationship between variables), will “not be a good approximation” in the study of relationship between system performance and RAM size. Consequently, a non-linear relationship of appropriate type, between computer system performance and RAM size will only give a better approximation (read chapter 8, Book 3). Since, Non Linear regression is not within the scope of this course, we will restrict our discussion only to Correlation and Linear Regression. Recall that a linear relationship between variables (independent variable) and (dependent variable) can be put in the form

.

The objective of our study is to investigate the question

> “Is there any relation between the size of Random Access Memory (RAM) and the Performance of Computer System?”

Steps:

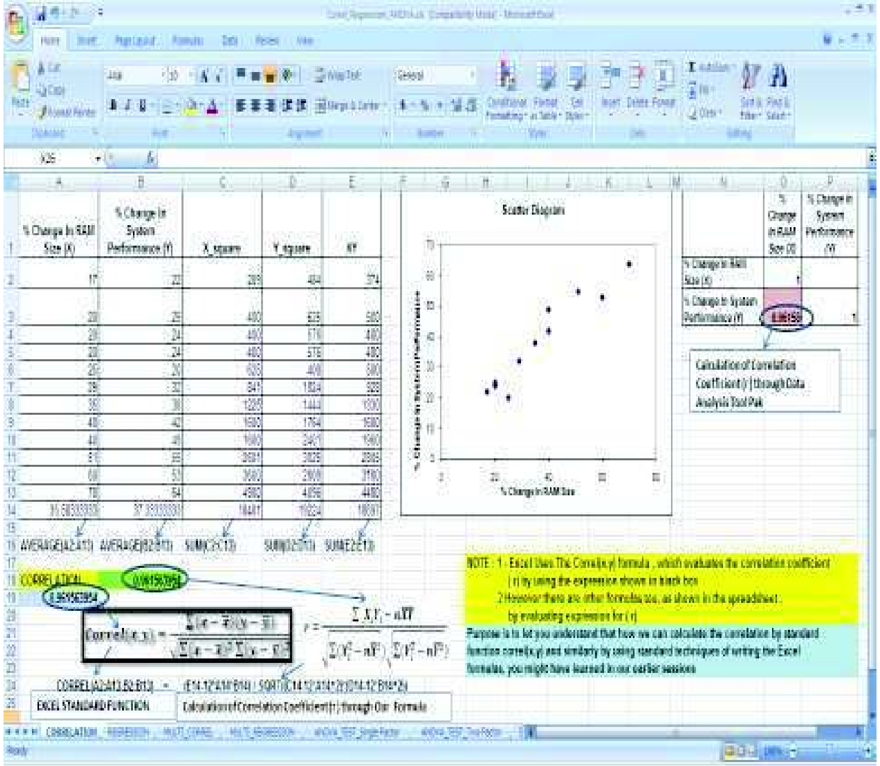
- From the objective of the problem, identify the two variables, which in this case are “size of Random Access Memory (RAM)” and the “performance of Computer system”.
- In order to proceed with our study, we are required to collect data, for which we are require to have RAM of different sizes. Further, for the sake of simplicity, let us restrict the study to the processor existing in the system. It is an exhaustive exercise, in which we have to mount different RAM on the Motherboard slot repeatedly, restart the system and monitor the system performance parameters. Say, we record the Percent (%) variation in RAM size and the percent (%) variation in the System performance in the Excel spread sheet.
- Since we are investigating the effect on System performance, the influence of variation in RAM size, i.e., the % variation in RAM is the independent variable and % variation in system performance is the dependent variable.
- Notice that the importance of identifying the two variables rests on the fact that there are *two different lines of regression* viz., on and on . (see Chapter 8, Book 3)

Now to calculate the correlation between the two continuous variables in excel, tabulate the recorded observations into the excel worksheet. Excel enables us to investigate the question stated above in three different ways, viz., through

1.Excel Formula 2. Standard formula - Correl(x,y)3. Data Analysis ToolPak

As an exercise you are required to observe the consistency of the results obtained through all three methods. This will help you to identify the mistakes, which might have occurred while writing your own formula.

In the screen shot given below you can identify that all three options leads to same result.



	A	B	C	D	E
1	% Change In RAM Size (X)	% Change In System Performance (Y)	X_square	Y_square	XY
2	17	22	289	484	374
3	20	25	400	625	500
4	20	24	400	576	480
5	20	24	400	576	480
6	25	20	625	400	500
7	29	32	841	1024	928
8	35	38	1225	1444	1330
9	40	42	1600	1764	1680
10	40	49	1600	2401	1960
11	51	55	2601	3025	2805
12	60	53	3600	2809	3180
13	70	64	4900	4096	4480
14	35.58333333	37.33333333	18481	19224	18697
16	AVERAGE(A2:A13)	AVERAGE(B2:B13)	SUM(C2:C13)	SUM(D2:D13)	SUM(E2:E13)

CORRELATION 0.96156295

0.96156295

$$Correl(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

$$r = \frac{\sum X_iY_i - n\bar{X}\bar{Y}}{\sqrt{\sum(X_i^2 - n\bar{X}^2)}\sqrt{\sum(Y_i^2 - n\bar{Y}^2)}}$$

CORREL(A2:A13,B2:B13) = (E14-12*A14*B14) / SQRT((C14-12*A14^2)*(D14-12*B14^2))

EXCEL STANDARD FUNCTION

Calculation of Correlation Coefficient (r) through Our Formula

	% Change in RAM Size (X)	% Change in System Performance (Y)
% Change in RAM Size (X)	1	
% Change in System Performance (Y)	0.96156	1

Calculation of Correlation Coefficient (r) through Data Analysis Tool Pak

NOTE : 1 - Excel Uses The Correl(x,y) formula , which evaluates the correlation coefficient (r) by using the expression shown in black box
2 However there are other formulas too, as shown in the spreadsheet ; by evaluating expression for (r)

Purpose is to let you understand that how we can calculate the correlation by standard function correl(x,y) and similarly by using standard techniques of writing the Excel formulas, you might have learned in our earlier sessions

Figure 1

Now, let us explore each of the option to calculate the correlation coefficient( ) i.e., using own Excel Formula, Standard formula - ***correl(x,y)*** and Data Analysis ToolPak

1. **Own Excel Formula**: There are different forms of the formula to calculate correlation coefficient ( ).Standard function described in excel i.e. ***correl(x,y)*** uses the formula in the form

   ——————.

   However, we will calculate the same using the form

   ———— ———— ,

   in our excel implementation.

   To implement it, tabulate the data as shown in the above screen shot and write the formula **(E14-12*A14*B14)/SQRT((C14-12*A14^2)*(D14-12*B14^2))** to calculate correlation coefficient ( ). You might have learned the method of formula writing in excel in earlier courses. Another form useful in computations is

   —————— —————.

2. **Standard formula - correl(x,y)** :

   a. Tabulate the collected data in column A & B as shown in above screen shot

   b. Select any cell in which you wish to have result of correlation coefficient ( )

c. In that cell write CORREL(A2:A13,B2:B13), where A2:A13 and B2:B13 are the columns related to the tabulated/recorded data, i.e., the values of variables and .

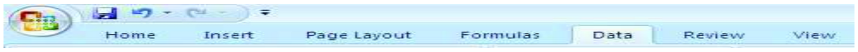
You can also use Excel Help by pressing function F1 key and type the function or formula you want to understand. Say, press F1 and type **correl** in search option you will get all related details of that formula. CORREL(X,Y) is the Excel formula to calculate correlation coefficient . Here,

where & are average values of variable & respectively.

3. **Using Data Analysis ToolPak** : You may calculate the correlation coefficient ( ) by using the Data Analysis Toolpak utility of Excel, which provides a set the tools necessary for statistical Data Analysis.

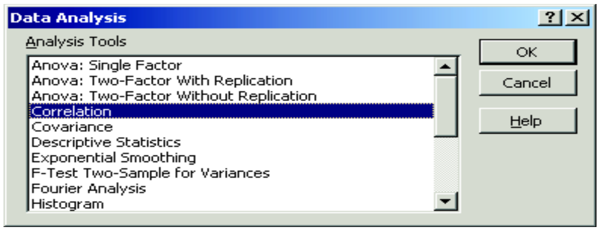
Steps to find Correlation coefficient ( ) using Data Analysis ToolPak:

a. Click the Data tab



b. Click the Data Analysis tab

c. Select correlation option and click OK



d. Select the input range,

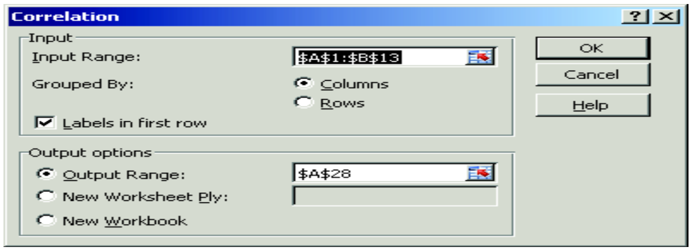
i. Which in our case is A1 to B13, relating to % variation in RAM and % variation in system performance.

ii. Choose the columns option as the data is tabulated in the columnar manner.

iii. Check the Option Labels in first row, as in A1 and B1 we are having headings for respective columns.

iv. Finally, choose the cell location where you want to have the output result.

v. press OK


Correlation

Input

Input Range: $A$1:$B$13

Grouped By: ● Columns ○ Rows

☑ Labels in first row

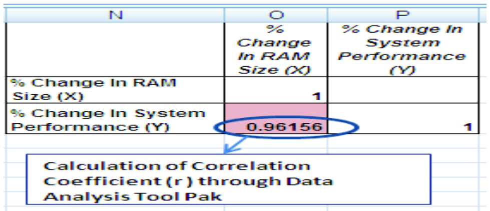
Output options

● Output Range: $A$28

○ New Worksheet Ply:

○ New Workbook

e. Output:

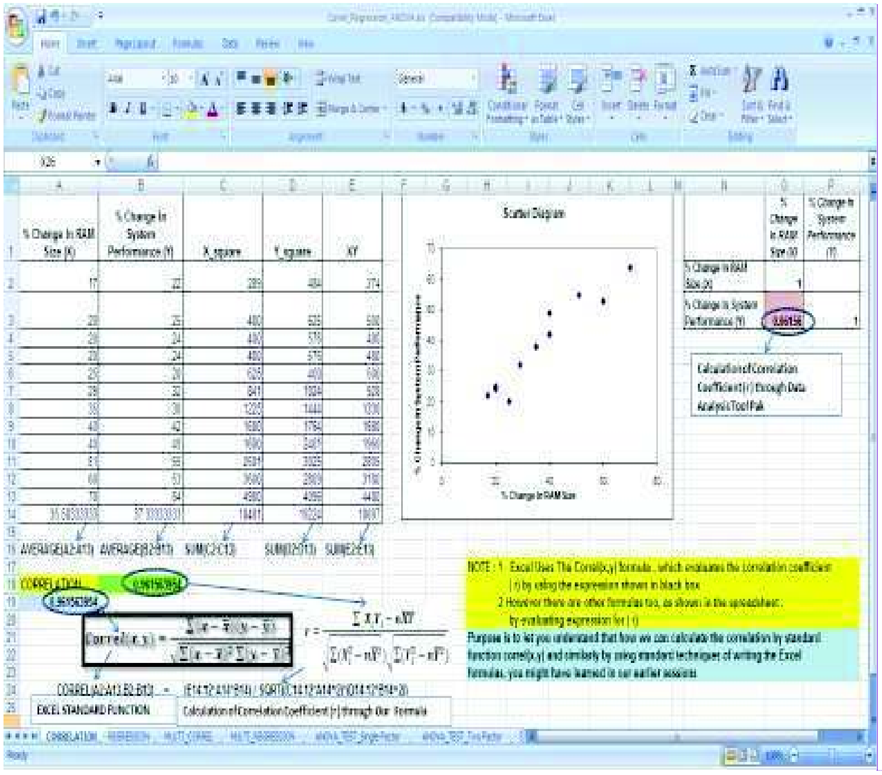


N	O	P
	% Change In RAM Size (X)	% Change In System Performance (Y)
% Change In RAM Size (X)	1	
% Change In System Performance (Y)	0.96156	1

Calculation of Correlation Coefficient (r) through Data Analysis Tool Pak

NOTE : We will prefer to use Data Analysis Toolpak for other applications subsequently and would like to put more emphasis on data interpretation.

Screen Shot of Final Outcome



	A	B	C	D	E
1	% Change In RAM Size (X)	% Change In System Performance (Y)	X_square	Y_square	XY
2	17	22	289	484	374
3	20	25	400	625	500
4	20	24	400	576	480
5	20	24	400	576	480
6	25	20	625	400	500
7	29	32	841	1024	928
8	35	38	1225	1444	1330
9	40	42	1600	1764	1680
10	40	49	1600	2401	1960
11	51	55	2601	3025	2805
12	60	53	3600	2809	3180
13	70	64	4900	4096	4480
14	35.58333333	37.33333333	18481	19224	18697
16	AVERAGE(A2:A13)	AVERAGE(B2:B13)	SUM(C2:C13)	SUM(D2:D13)	SUM(E2:E13)

CORRELATION 0.96156954

0.96156954

$$Correl(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

$$r = \frac{\sum X_iY_i - n\bar{X}\bar{Y}}{\sqrt{\sum(X_i^2 - n\bar{X}^2)}\sqrt{\sum(Y_i^2 - n\bar{Y}^2)}}$$

CORREL(A2:A13,B2:B13) = (E14-12*A14*B14) / SQRT((C14-12*A14*A14)*(D14-12*B14*B14))

EXCEL STANDARD FUNCTION

Calculation of Correlation Coefficient (r) through Our Formula

NOTE : 1 Excel Uses The Correl(x,y) formula, which evaluates the correlation coefficient (r) by using the expression shown in black box
2 However there are other formulas too, as shown in the spreadsheet: by evaluating expression for (r)

Purpose is to let you understand that how we can calculate the correlation by standard function correl(x,y) and similarly by using standard techniques of writing the Excel formulas, you might have learned in our earlier sessions

	N	O	P
		% Change In RAM Size (X)	% Change In System Performance (Y)
	% Change In RAM Size (X)	1	
	% Change In System Performance (Y)	0.96156	1

Calculation of Correlation Coefficient (r) through Data Analysis Tool Pak

Figure 2

Comment [
Attn. Mr.A.Kumar: Please modify the 2nd formula of r in the image. To remove brackets. Similar to earlier one

Sudhansh : Sir I made the desired corrections hope its OK Now

Remark on Data Interpretation

- **Which curve to fit?**
  - ➢ Between two continuous variables the relationship can be ascertained by first plotting a scatter diagram.
  - ➢ Then, simply looking at the trend of the resultant plot, it could be linear or curvilinear.
  - ➢ You are required to apply your understanding of our earlier sessions to plot scatter diagram etc. in Excel.

- **Why ?**

  - If the resulting scatter diagram exhibits a linear relationship between and , the extent can be quantified using correlation coefficient ( ),which "measures the extent of linearity" present in the data.
  - Value of always lies between and .
  - Closer the value of towards , the stronger is the linear relationship between the variables.
  - You can verify in the subsequent section on simple linear regression that depending on the value of being or , the slope of the line of regression will also be or − respectively. In other words, when is or an increase in the value of "independent variable ( )" will result in an increase or decrease in the value of the "dependent variable ( )".

  - If the relationship is found to be significantly strong, the line of "best fit "passing through the bi-variate data can be obtained and it is discussed in section 2.4.

☞ **Check your progress 1**

Try this: Analyze the screen shot and use the data interpretation tips given above, to answer following questions:

a) Study the scatter diagram only, what can we conclude about the data?

b) Analyze the correlation coefficient ( ) and comment on the relationship between RAM size variation and System performance.

c) Can we use the collected data for the purpose of forecasting?

2.3 MULTIPLE CORRELATION

In this section we extend our discussion to the case of coefficient of multiple correlation, which measures the extent to which a dependent variable can be predicted using the *linear function of a set of independent variables*. To explain the concept let us take another example from the IT sector, companies generally have a monitoring index called the technical index(TECH INDEX).The monitoring of this index goes on monthly basis, and the indexing is performed to identify growth of an independent company with respect to the industry requirements. The data related to the variation between industry parameter called TECH INDEX and the monthly rate of return is gathered for few companies viz. Google, Yahoo, Microsoft, Apple; the same is to be analyzed. Since variation of multiple companies in coordination with one parameter is to be analyzed, we take recourse to the concept of Multiple Correlation which you studied in BCS 040 Block 3. Denoting the companies by and industry index by , a multiple regression model that describes this relationship can be put in the form

Comment [AU2]:
Please Check this text is OK in terms of context. The word **synchronization** is NOT clear to me

Sudhansh : Thanks for your valuable comments. I m,ade the correction, hope the same is acceptable now

.

You are required to perform the following:

1. Plot the graph for all collected data.

   This will help you to visually identify the variation in variable synchronization.

2. Run Correlation analysis on all the variables simultaneously.

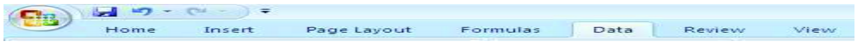
Given below is the data on various IT sector companies stated above:

MONTHLY RATE OF RETURNS FOR IT SECTOR COMPANIES					
		COMPANY			
DATE	TECH INDEX	GOOGLE	YAHOO	MICROSOFT	APPLE
1-Apr-11	0.8799	0.7541	2.1407	-4.6296	-18.8406
1-May-11	7.5187	14.9701	-2.5948	18.986	6.6964
1-Jun-11	5.558	11.9792	7.7869	-1.7226	-3.3473
1-Jul-11	1.3716	7.907	-8.5551	-0.5535	5.8442
1-Aug-11	-1.6289	-5.1724	1.2474	6.679	1.9427
1-Sep-11	2.4171	3.4091	0.8214	1.8261	2.1063

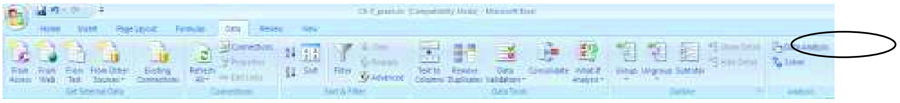
As the first step of analysis, we are interested in visualizing the data graphically and to this end obtaining a scatter plot is simple as well as most appropriate. Notice that in the case of multiple correlation & regression with only *two independent variables*, scatter diagram can be plotted in the three dimensions. Thus, for the sake of completeness, we can generate a plot with date along the -axis and the series values along the same -axis to get an insight into the given data. (see Output Screen Shot given below)

Steps for Multiple Correlation using Data Analysis ToolPak

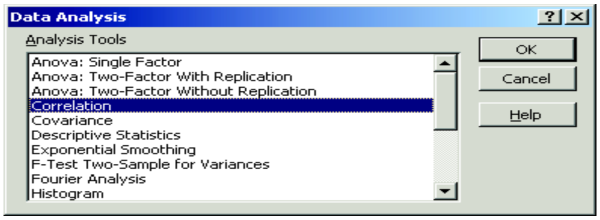
1. Click the Data tab



2. Click the Data Analysis tab



3. Select correlation option and click OK



4. Select the input range,

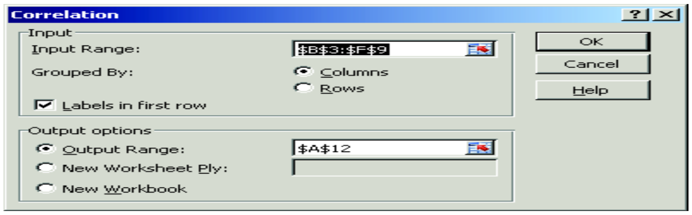
   a. Which in our case is $B$3:$F$9, related to TECH Index and Companies data understudy.

   b. Choose the columns option as the data is tabulated in the columnar manner.

   c. Check the Option Labels in first row, as from B3 to F3 we are heaving headings for respective columns.

   d. Finally, opt for the cell location where you want to have the output result

   e. press OK



Correlation

Input

Input Range: $B$3:$F$9

Grouped By: ● Columns ○ Rows

☑ Labels in first row

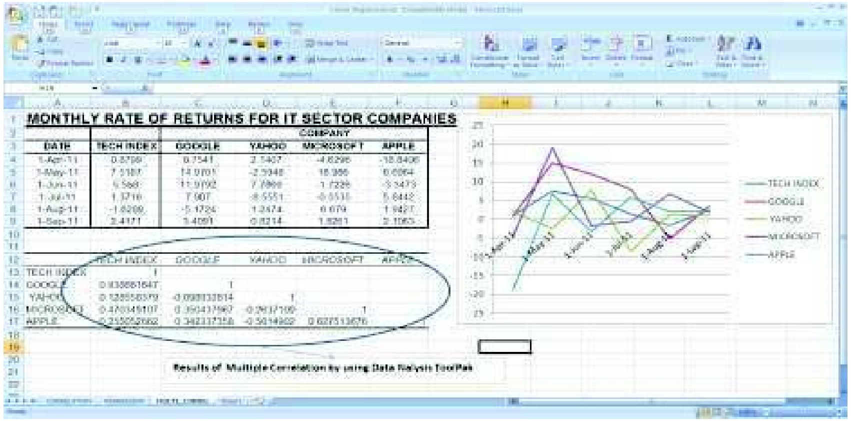
Output options

● Output Range: $A$12

○ New Worksheet Ply:

○ New Workbook

5. Output



MONTHLY RATE OF RETURNS FOR IT SECTOR COMPANIES

DATE	TECH INDEX	COMPANY GOOGLE	YAHOO	MICROSOFT	APPLE
1-Apr-11	0.8799	0.7541	2.1407	-4.6296	-18.8406
1-May-11	7.5187	14.9701	-2.5948	18.386	6.6964
1-Jun-11	5.568	11.9792	7.7869	-1.7228	-3.3473
1-Jul-11	1.3716	7.907	-8.5551	-0.5535	5.6442
1-Aug-11	-1.6209	-5.1724	1.2474	0.679	1.9427
1-Sep-11	2.4171	3.4091	0.8214	1.8261	2.1063

	TECH INDEX	GOOGLE	YAHOO	MICROSOFT	APPLE
TECH INDEX	1				
GOOGLE	0.938661647	1			
YAHOO	0.128558379	-0.098932814	1		
MICROSOFT	0.470349107	0.350437967	-0.2637109	1	
APPLE	0.255052662	0.342337358	-0.5014902	0.627513676	

Results of Multiple Correlation by using Data Nalysis ToolPak

DATA INTERPRETATION

	TECH INDEX	*GOOGLE*	*YAHOO*	*MICROSOFT*	*APPLE*
TECH INDEX	1				
GOOGLE	0.938661647	1			
YAHOO	0.128558379	-0.098932814	1		
MICROSOFT	0.470349107	0.350437967	-0.2637109	1	
APPLE	0.255052662	0.342337358	-0.5014902	0.627513676	1

The table shown above contains correlation coefficients discussed in the previous section, computed between pairs of variables for the data under study. Notice that it is in the form of a matrix, which is known as *matrix of correlation coefficients* and it is symmetric about the diagonal (why?). Clearly the elements along the diagonal are correlation coefficients computed between values of a variable with itself, which is *unity* (refer to section 2.2 and state why?).Further, our study requires us to analyze the relation between industry parameter TECH INDEX and Monthly rate of return of various companies. Clearly,

1. There is high correlation (0.938661647) between Google's monthly rate of return data and the industry TECH INDEX, which is on the expected lines.
2. Similarly, for the Microsoft's monthly rate of return data the correlation coefficient is0.470349107 and it is reasonably high.
3. In contrast to these,the respective correlation coefficients in the case of Yahoo and Apple are 0.128558379 and 0.255052662 respectively, which is low.

Thus, we learned how to interpret the results obtained by using the correlation coefficient in Data analysis ToolPak. However, we will subsequently in section 2.4.2 learn how to compute the value of multiple correlation coefficient (denoted by) and to this end we will extend your knowledge of section 9.3/Unit 9.

2.4 REGRESSION

In section 2.2 above you learned about correlation and the related excel tools and now we extend our discussion towards regression. Regression analysis enables us in estimation and forecasting of the value of dependent variable for given value(s) of the independent variable(s) and is extensively used in various disciplines. Clearly, while in simple linear regression there is a single independent variable; in multiple regression there are more than one independent variable. In other words, the former is confined to the study of two variables only; the latter is concerned with the study of more than two variables. Further, in case of simple regression if the relationship between the dependent and independent variables follow a straight line pattern, it is called *linear regression*. On the other hand, if the relation is expressed in the form of a curve it is called curvilinear regression.

Task: Discuss with other students or counselors

- Example of curves.
- Example of curves that *can be* transformed into linear form (called tractable linear form). What about ?
- Example of curves that *cannot be* transformed into linear form (called non-tractable form). What about ?
- Importance of scatter diagram in deciding a specific curve.

We will restrict our discussion to the simple linear regression and later extend it to the case of Multiple regression. The general problem of finding the equation that fits a given data set is called *curve fitting*.

Before starting with the practical problem solving through excel, we present below a discussion on regression.

2.4.1 Linear Regression

Here we are going to study the case of two variables & . Thus,there are two lines of regression viz., on and on . The regression line on gives the most probable values of for a given value of and the regression line on gives the most probable values of for a given value of . However, when there is a perfect correlation between & i.e., , the two regression lines on and on coincides i.e. we will have one regression line(see chapter 8, Book 3). Otherwise, there are two lines of regression which coincide at ( ).

The regression line on is given by

,

which simplifies to,

Similarly, the regression line on is given by

,

which simplifies to,

.

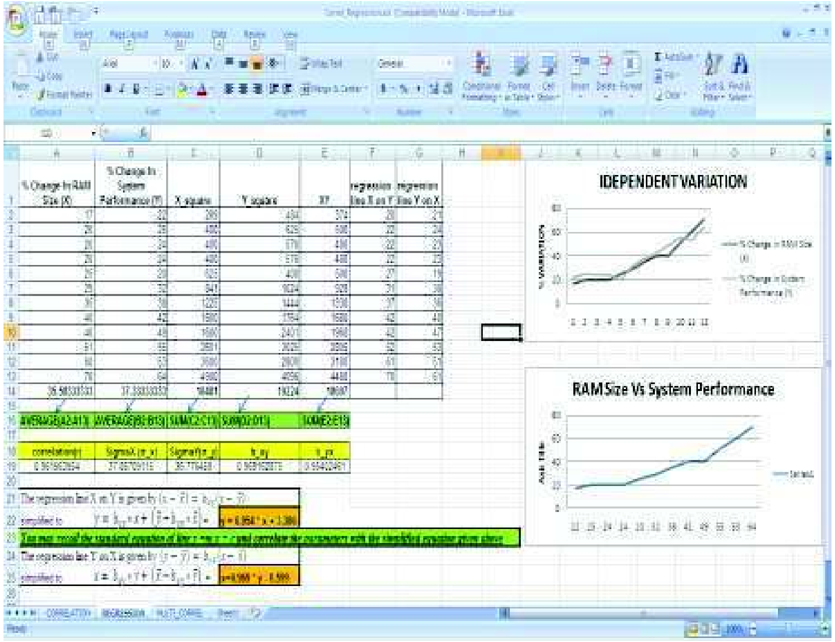
Where,

- and are arithmetic mean and standard deviation of variables and respectively,
- —is the regression coefficient of on ,
- —is the regression coefficient of on ,
- is correlation coefficient.

Note:

1. The two equations above can re-written in a simple way in terms of standardized values in the form given below:
   ── ──and── ── .
2. The two forms written above are the least-squares lines of regressions of on and on .
3. You may recall the equation of a straight line and compare the parameters with the least-squares line of regressions given above to get and .
4. . (you may refer to chapter 8, Book 3 for details)
5. We recommend you to apply the skill you developed in earlier sessions and perform calculations by writing your own formula for correlation and regression.

However, we exhibit the utilization of Excel Data Analysis ToolPak utility, in our subsequent section on multiple regression.



	A	B	C	D	E	F	G
1	% Change In RAM Size (X)	% Change In System Performance (Y)	X_square	Y_square	XY	regression line X on Y	regression line Y on X
2	17	22	289	484	374	20	21
3	20	25	400	625	500	22	24
4	20	24	400	576	480	22	23
5	20	24	400	576	480	22	23
6	25	20	625	400	500	27	19
7	29	32	841	1024	928	31	30
8	35	38	1225	1444	1330	37	36
9	40	42	1600	1764	1680	42	40
10	40	49	1600	2401	1960	42	47
11	51	55	2601	3025	2805	52	53
12	60	53	3600	2809	3180	61	51
13	70	64	4900	4096	4480	70	61
14	35.58333333	37.33333333	18481	19224	18697		
15							
16	AVERAGE(A2:A13)	AVERAGE(B2:B13)	SUM(C2:C13)	SUM(D2:D13)	SUM(E2:E13)		
17							
18	correlation(r)	SigmaX (σ_x)	SigmaY(σ_y)	b_xy	b_yx		
19	0.961962954	[illegible]	[illegible]	[illegible]	[illegible]		

21 The regression line X on Y is given by $(x-\bar{x}) = b_{xy}(y-\bar{y})$

22 simplified to $x = b_{xy} * y + (\bar{x} - b_{xy} * \bar{y})$ = x = 0.964 * y + [illegible]

23 You may recall the standard equation of line y = mx + c and correlate the parameters with the simplified equation given above

24 The regression line Y on X is given by $(y-\bar{y}) = b_{yx}(x-\bar{x})$

25 simplified to $y = b_{yx} * x + (\bar{y} - b_{yx} * \bar{x})$ = y = [illegible] * x - [illegible]

Figure 3

Remark on Data Interpretation:

1. **Correlation coefficient** ( )
   From the above screen shot one can observe that the value of correlation coefficient (r) is quite high, thus we may conclude that the X and Y are directly proportional
2. **Regression coefficients** and
   implies that with unit increase in , the value of increases by times. Similarly implies that with unit increase in , the value of increases by times

☞ Check your progress 2

Try this: Analyze the screen shot & use the data interpretation tips given above, to attempt following questions:

a) Comment on the interdependence of RAM size variation and System performance.

b) With an unit increase in RAM size, how much you expect the system performance to improve?

c) For unit improvement in system performance how much RAM size is expected to be altered?

d) Can you forecast the system performance, if change in RAM size is 25%?

2.4.2 Multiple Regression

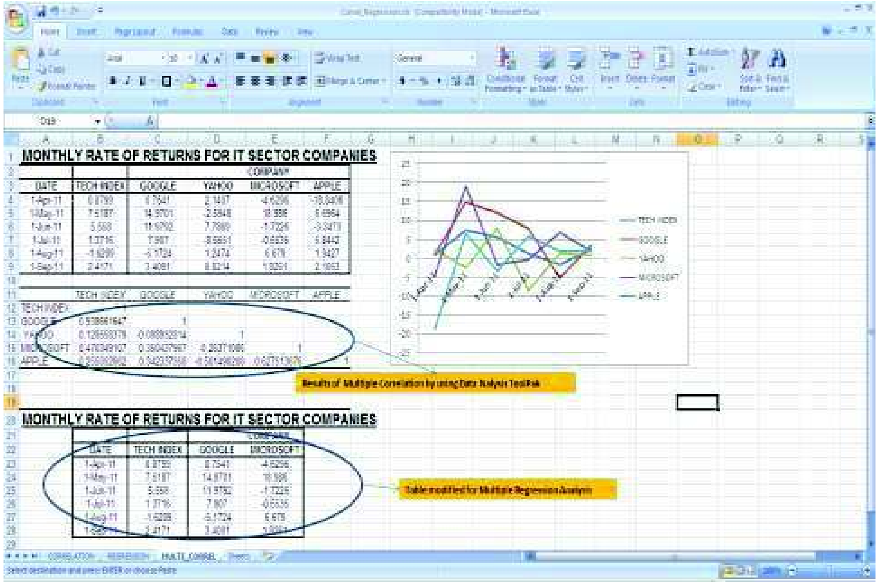
Did you complete the sections 2.3 and 2.4.1 before starting with this section? It is advised that you should complete these sections prior to starting with multiple regression analysis. You may ask why? There as on being it will enable you to identify the continuity and extensions in the topics you are working on. As in the case of TECH INDEX parameter you studied in section 1.2.1 above, we firstly identified the companies which influence (or contributes to)the composite industry index. Thus, we ought to include variables that can explain the variation in the composite industry index, viz. GOOGLE & MICROSOFT and not YAHOO & APPLE. Data on the first two companies considered for multiple regression analysis are listed below:

Commer The word “influence” is more appropriate in place of “harmony”. Can we edit the line accordingly?

Sudhansh : Correction made , harmony is replaced by influence, as desired, and the text is also changed accordingly

		COMPANY	
DATE	TECH INDEX	GOOGLE	MICROSOFT
1-Apr-11	0.8799	0.7541	-4.6296
1-May-11	7.5187	14.9701	18.986
1-Jun-11	5.558	11.9792	-1.7226
1-Jul-11	1.3716	7.907	-0.5535
1-Aug-11	-1.6289	-5.1724	6.679
1-Sep-11	2.4171	3.4091	1.8261

Screen shot is as follows



MONTHLY RATE OF RETURNS FOR IT SECTOR COMPANIES

		COMPANY			
DATE	TECH INDEX	GOOGLE	YAHOO	MICROSOFT	APPLE
1-Apr-11	0.8799	0.7541	2.1497	-4.6296	-10.8408
1-May-11	7.5187	14.9701	-2.5848	18.986	5.6964
1-Jun-11	5.558	11.9792	7.7989	-1.7226	-3.3473
1-Jul-11	1.3716	7.907	-8.5651	-0.5535	5.8442
1-Aug-11	-1.6289	-5.1724	1.2474	6.679	1.9427
1-Sep-11	2.4171	3.4091	8.8214	1.8261	2.1052

	TECH INDEX	GOOGLE	YAHOO	MICROSOFT	APPLE
TECH INDEX	1				
GOOGLE	0.538661647	1			
YAHOO	0.128658079	-0.088932874	1		
MICROSOFT	0.470349107	0.380437967	-0.26371086	1	
APPLE	0.356092862	0.342337358	-0.501498289	0.627513676	1

Results of Multiple Correlation by using Data Nalysis TooPak

MONTHLY RATE OF RETURNS FOR IT SECTOR COMPANIES

		COMPANY	
DATE	TECH INDEX	GOOGLE	MICROSOFT
1-Apr-11	0.8799	0.7541	-4.6296
1-May-11	7.5187	14.9701	18.986
1-Jun-11	5.558	11.9792	-1.7226
1-Jul-11	1.3716	7.907	-0.5535
1-Aug-11	-1.6289	-5.1724	6.679
1-Sep-11	2.4171	3.4091	1.8261

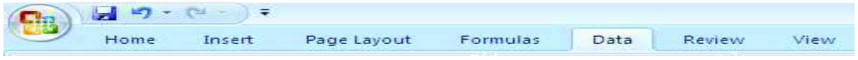
Table modified for Multiple Regression Analysis

Figure 4

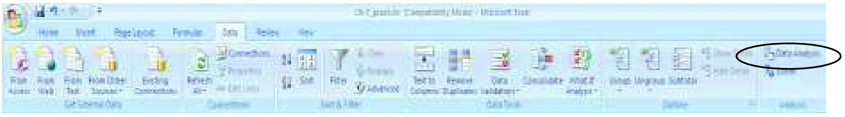
Now let us use the tool i.e., Data Analysis ToolPak's Regression facility to perform multiple regression analysis:

Steps for Multiple Regression using Data Analysis Tool Pak:

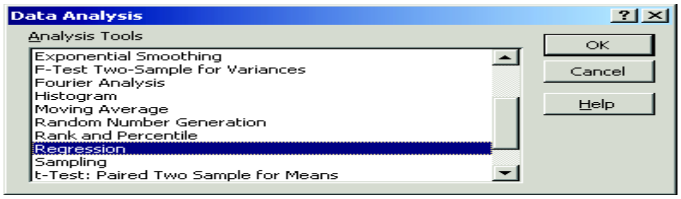
1. Click the Data tab



2. Click the Data Analysis tab



3. Select Regression option and click OK



Analysis Tools

- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means

4. Select the respective data range,

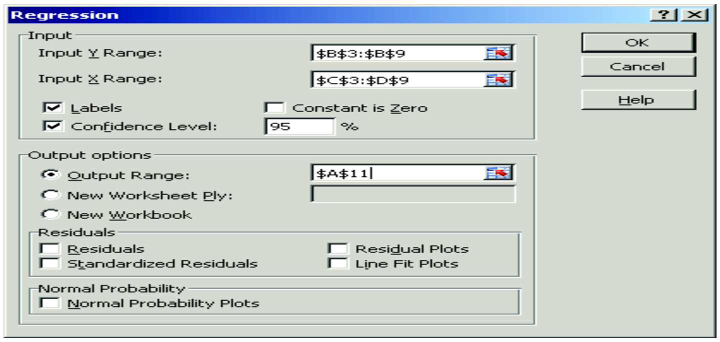
   a. Which in our case is $B$3:$B$9 as Input Y- Range, related to TECH Index and $C$3:$D$9 as Input X-Range, related to shortlisted Companies viz. Google and Microsoft.

   b. Choose the Output range, we opted for cell location $A$11

   c. Check the Option Labels and confidence level at which we wish to fix and this in our case is 95%.

   d. Finally, opt for the cell location where you want to have the output result

   e. Press OK.



Input

Input Y Range: $B$3:$B$9

Input X Range: $C$3:$D$9

☑ Labels ☐ Constant is Zero

☑ Confidence Level: 95 %

Output options

● Output Range: $A$11

○ New Worksheet Ply:

○ New Workbook

Residuals

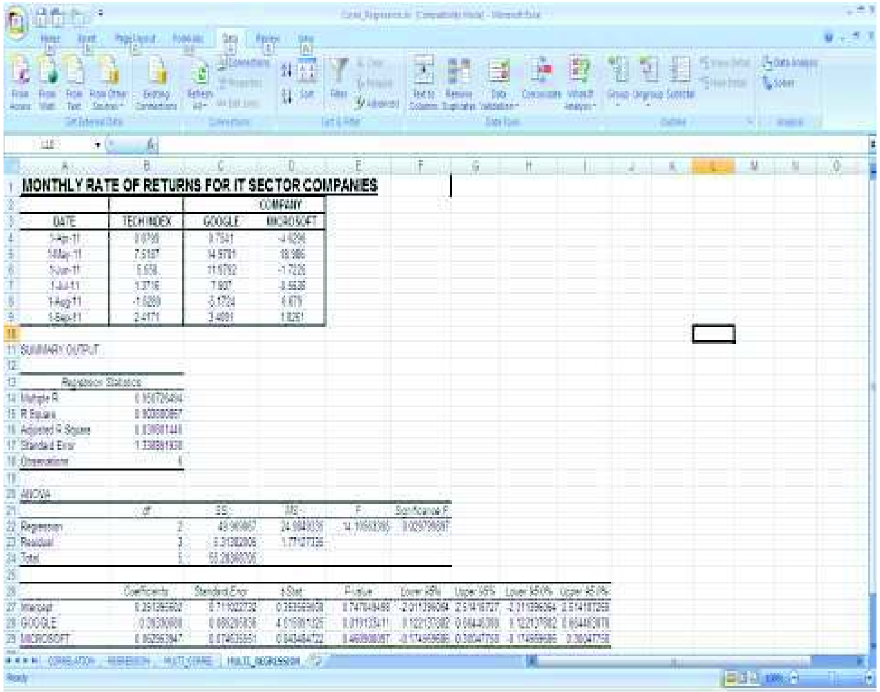
☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

5. Output



MONTHLY RATE OF RETURNS FOR IT SECTOR COMPANIES

		COMPANY	
DATE	TECH INDEX	GOOGLE	MICROSOFT
1-Apr-11	0.9798	0.7541	-4.4296
1-May-11	7.5187	14.5781	18.986
1-Jun-11	5.678	11.9792	-1.7226
1-Jul-11	1.3716	7.937	-8.5538
1-Aug-11	-1.0289	-3.1724	6.679
1-Sep-11	2.4171	3.4891	1.8251

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950726494
R Square	0.903880867
Adjusted R Square	0.839801446
Standard Error	1.330891938
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	2	49.969867	24.9849336	14.10568395	0.029798687
Residual	3	5.31382806	1.77127335		
Total	5	55.28369506			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.251395602	0.711022732	0.353569608	0.747049489	-2.011396084	2.51416727	-2.011396084	2.514187258
GOOGLE	0.38309988	0.09528936	4.01581025	0.01913411	0.12213782	0.6446389	0.12213782	0.64463878
MICROSOFT	0.05296394	0.07463551	0.84348472	0.46050805	-0.17468596	0.30047758	-0.17468596	0.30047758

The Data Analysis ToolPak, gave exhaustive summary for the Data, out of which we will focus on the Regression statistics and subsequent sessions we will extend our discussion to ANOVA.

Interpretation of Results

1. Following is the summary output obtained using Data analysis ToolPak, which we seek to interpret.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950726494
R Square	0.903880867
Adjusted R Square	0.839801446
Standard Error	1.330891938
Observations	6

2. The Multiple is the multiple correlation coefficient and it measures the strength of the association. The range of Multiple R is -1 to +1, which in the present case yields to the value .This implies that the dependent variable which in our case is TECH INDEX has a high positive correlation with the combined effect of Google ( ) and Microsoft ( ).

3. The Square value ( ) is the square of the multiple correlation coefficient . It measures the strength of the regression prediction compared with predicting solely by the response mean. Alternatively, the value of ,can also be interpreted as that of the variation in can be explained by variables & . It may further be noted the closer value of to 1, the stronger is the regression prediction.

4. Defining multiple correlation as ———— which is in terms of sums of squares (SS)can be expressed as ——— ——— (for values see screen shot above).

5. Discuss with other students or counselors that in general, the value of *will always increase* when an additional regressor / independent variable is added (say YAHOO).Details of which is beyond the scope of current content.

6. We present below the Regression coefficients from the screen shot along with summary statistics.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.251395602	0.711022732	0.353569008	0.747049498
GOOGLE	0.39330088	0.085205836	4.615891325	0.019133411
MICROSOFT	0.062953947	0.074635551	0.843484722	0.460900097

The column of coefficients contain the values of intercept ( ) and the regression coefficients and of the regression equation

,

which in the present case is

The value givesus how much will changefor each unit change in , and is held constant. Similarly for .To understand this, consider the regression equation − − . Here since the coefficient of is negative, keeping constant, an unit change in leads to a decrease in by 0.5403. Again, since the value of regression coefficient is 13.4852, keeping constant, an unit change in leads to an increase in by 13.4852.

Read section 9.4.2 of Unit 9

☞ Check your progress 3

1) Analyze the Multiple Regression Summary output of Two independent variable & and dependent variable given below, and determine what percentage of variation in is explained by & .

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.840726494
R Square	0.705680867
Adjusted R Square	0.639801446
Standard Error	1.330891938
Observations	5

....

....

....

....

....

....

2) Consider the following multiple regression analysis results for two independent variable & and dependent variable ,

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	- 5.51395602	0.711022732	0.353569008	0.747049498
	- 39.330088	0.085205836	4.615891325	0.019133411
	6.2953947	0.074635551	0.843484722	0.460900097

Determine the Multiple Regression equation and explain how & are related to .

....

....

....

....

2.5 SUMMARY

The practical sessions covered in this unit enabled you to enrich your understanding of correlation and regression, which you gained in BCS 040, through practical implementation using MS EXCEL Data Analysis ToolPak. It is important to understand that mere usage of MS Excel or any other software will not serve the purpose unless and until you possess an understanding of the subject. It worth noting that regression analysis finds extensive application in a wide range of field of research.

2.6 EXERCISES

Exercise 1:

Is there a relationship between moderate milk consumption and heart disease rate? The table underneath provides data from 6 developed countries from various cultures.

Country	A	B	C	D	E	F
Liters of milk per year per capita (x)	25	24	8	79	18	65
Deaths from hearth disease per 100,000 people per year (y)	211	191	297	107	167	86

Exercise 2:

The data in the table below show the percent of people who purchase their music from the internet (with 1997 corresponding to t = 0). Calculate the equation of the regression line and predict what percent of people will purchase their music from the internet in 2007 if this model is correct.

Year	0	1	2	3	4	5
Percent	0.3	1.1	2.4	3.2	2.9	3.4

1.7 SOLUTIONS

Check Your Progress-1

Q1

a) Scatter plot shows that data is highly positively correlated. Recall that the value is approximately close to 0.9

b) Since , thus variation in one variable directly affects the other. In other words, increase or decrease in RAM size will result in a corresponding increase or decrease in system performance.

c) Yes, we can use the collected data to construct a linear regression model and hence use for the purpose of forecasting. This is justified from the high value of correlation coefficient .

Comment [AU4]:
I have NOT computed the value.
Please write the exact value & delete "or so"

Sudhansh : text "or so " deleted, however the value is close to 0.9 is just to give hint and not any calculation

Check Your Progress-2

Q1

a) Both factors are seen to be highly interdependent.

b) Unit increase in RAM size will lead to 0.69 times increase in systems performance.

c) Unit improvement in systems performance requires 0.54 times of alteration in RAM

d) 25 times of RAM change will lead to improve system performance by 0.69 X 25 times

Check Your Progress-3

Q1

a. of variation in is explained by & .

b. –

Exercise 1:

The coefficient of correlation is given by

so that in our case we get

Exercise 2 :

Let us calculate the coefficient of correlation using the formula in Exercise 1:

Since there is high positive correlation between the variables, we calculate the coefficients of the least squares regression line as below:

Slope and

Intercept – – .

The equation of the least squares regression line thus becomes .
We calculate the percentage of people that purchase their music on the internet in the year 2007 by substituting into the least squares regression line. We get .

SECTION 3 ANALYSIS OF VARIANCE

Structure **Page Nos.**

3.0 INTRODUCTION

By now you must have become familiar with hypothesis testing based on test statistic -test, -test and F-test in the earlier sessions. (Please refer to Section 2, Book 3 or BCS 040 for details). Recall that you learned to test the significance of differences between two sample means earlier. In addition to this, there are situations in which we are interested in testing the significance of difference among two or more means or equivalently equality of more than two means.

For example, an industrial manufacturing unit may be interested in testing the quality of wielding done by workers who works in three different shifts viz., morning, evening and night. In order to assess the quality of welding carried out by these workers, data is collected by floor managers using an advanced imaging technique. The goal is to test for difference in the average welding quality standards. In other words, seek an answer to the query: Is there a significant difference in the average welding quality of the workers who works in the three shifts? Notice that whereas using -test, equality of only two means at a time can be carried out, ANOVA tests the hypothesis concerning differences between two or more means. An advantage in using ANOVA rather than multiple -tests is that it reduces the probability of error. ANOVA is a technique that works by partitioning the total sums of squares into components used in the model under consideration. It may further be noted that ANOVA is "*concerned not with analyzing the variances, but with analyzing the variation in means*."It is recommended that you revise BCS 040 unit 8 prior to starting with the sections below, as we choose to analyse data by exploring the excel tool Data Analysis ToolPak.

3.1 OBJECTIVES

After going through this unit you will be able to:

- Bring out an appraisal of any physical problem
- Identify suitability of using ANOVA
- Use Data Analysis Toolpak for ANOVA, in particular
  - Perform One Way / Single Factor ANOVA Test
  - Perform Two Way / Two Factor ANOVA Test

3.2 ANOVA

Analysis of variance is a technique due to Sir Ronald Fisher which can address questions such as the one mentioned in the example above. It makes use of the statistic you learned earlier.

3.2.1 One-Way Classification

The one-way classified data obtained in an experiment has the following layout with unequal number of observations in each treatment:

Treatment 1	...
Treatment 2	...
Treatment k	...

Recall the linear mathematical model for ANOVA you studied in section 8.3, Unit 8 in the form

, and

and the hypothesis to be tested is .

Assumptions:

- Errors ( )are identically and independently distributed with mean and variance (Homoscedastic).
- Errors ( ) have normal distribution.

Application CASE STUDY

FACTORS INFLUENCING SALES OF STREAMERS

Consider a company is producing Streamers and it is in the process of developing its dealer-distributor network. In order to accomplish the same, they recruited four dealers, having fixed shops in four different parts of a town. The amount of units sold by the dealers is tabulated below:

DEALER SALES RECORD						
	DAY 1	DAY 2	DAY 3	DAY 4	DAY 5	DAY 6
DEALER -1	22	46	62	43	36	
DEALER -2	25	35	42	55		
DEALER -3	22	44	66	33	13	50
DEALER -4	25	34	40	28	40	

- Based on the tabulated data, the company desires to investigate, is there any significant differences in the average number of streamers sold by the dealers.
- Some of the dealers are making efforts to promote their sales. Thus to promote the sales of the streamers, one of the dealer has appointed four salesmen. These

salesmen are guided to visit five localities of the same town randomly in a month and sell the product, whose day wise details are tabulated below.

- The locality wise sales record of each salesman is tabulated below:

Commen
Please see if clarity has been achieved, as desired in the CASE STUDY.

Sudhansh: Sir thanks for your valuable comment, I redrafted the text, hope the same is now OK

DEALER - 1 : LOCALITYWISE SALES RECORD OF SALESMEN

	LOCALITY 1	LOCALITY 2	LOCALITY 3	LOCALITY 4	LOCALITY 5
SALESMAN-1	22	33	9	31	18
SALESMAN-2	13	23	13	11	8
SALESMAN-3	7	15	4	24	15
SALESMAN-4	31	44	13	31	23

As a study, the dealer wants to test whether the salesmen differ in their ability of salesmanship and he wants to test that whether the locality has any influence on the sales of streamers.

ANALYSIS

Based on the case study, following objectives are identified with respect to the company and the dealer.

Company Objective

In order to test "whether the average numbers of streamers sold by the dealers differ significantly or not", we consider the model as below.

, and dealers

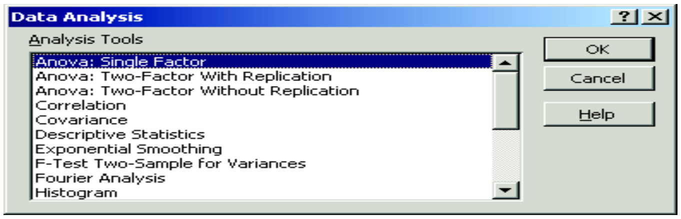
Where the additive model has termed as the additional effect of the th treatment and the hypothesis to be tested is , or equivalently , . Now, we demonstrate how to use the excel tool to perform a test of this hypothesis.

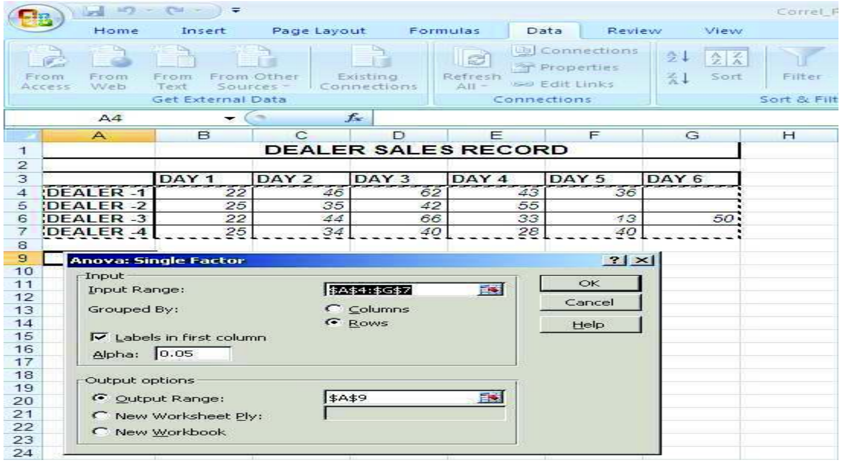
Steps:

1. Tabulate the DEALER SALES RECORD as given above in Excel Spreadsheet screen shot below.
2. Click DATA TAB → DATA ANALYSIS → ANOVA: Single Factor → OK

   "For activation and usage of Data Analysis Toolpak, refer to the earlier unit of Correlation and Regression - The snap shots are readily available there"

   However we are giving some of the relevant screenshots here





	A	B	C	D	E	F	G	H
1		DEALER SALES RECORD						
2								
3		DAY 1	DAY 2	DAY 3	DAY 4	DAY 5	DAY 6	
4	DEALER -1	22	46	62	43	36		
5	DEALER -2	25	35	42	55			
6	DEALER -3	22	44	66	33	13	50	
7	DEALER -4	25	34	40	28	40		

Anova: Single Factor

Input
Input Range: $A$4:$G$7
Grouped By: ○ Columns ● Rows
☑ Labels in first column
Alpha: 0.05

Output options
● Output Range: $A$9
○ New Worksheet Ply:
○ New Workbook

OK | Cancel | Help

Notice from the screen shot above that while selecting the cells only row labels are included (columns excluded). This is because incase of one way classification, the data is required to be classified by only one factor viz., the Dealers in the present case. Data on the sales volume of different dealers is recorded row wise, where the dealer names are entered in first column. Thus, we check the option "Levels in First Column". Further, the level of significance *Alpha* for the test is by default set at or , which can be altered to or etc. as per requirement. We have to identify the output cell address where the results are desired to be placed, which is chosen as . Following is the result of the procedure discussed above.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
DEALER -1	5	209	41.8	213.2
DEALER -2	4	157	39.25	158.9167
DEALER -3	6	228	38	374
DEALER -4	5	167	33.4	46.8

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	184.2	3	61.4	0.290072	0.831921	3.238872
Within Groups	3386.75	16	211.6719			
Total	3570.95	19				

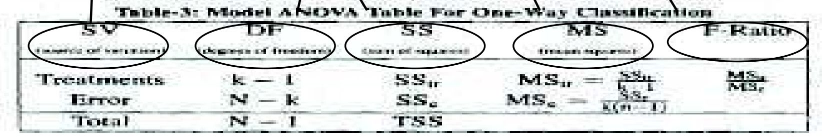


Table-3: Model ANOVA Table For One-Way Classification

SV (source of variation)	DF (degrees of freedom)	SS (sum of squares)	MS (mean squares)	F-Ratio
Treatments	$k-1$	SS_{tr}	$MS_{tr} = \frac{SS_{tr}}{k-1}$	$\frac{MS_{tr}}{MS_e}$
Error	$N-k$	SS_e	$MS_e = \frac{SS_e}{k(n-1)}$	
Total	$N-1$	TSS		

Now refer to Table 3 given at page 12 of BCS 040 Block 3 unit 8 i.e. ANOVA for a comparison of the results. Notice the forms of the tables marked with corresponding column heading shown above.

"F crit" in the Excel output is the critical value of - distribution at the stated level of significance, which can be obtained from the table. -Value for the calculated value of - statistic is also generated in the Excel output.

Data Interpretation

A test of the hypothesis, in the present case, can be carried out based on either of the following two approaches (see chapter 7, Book 3).

- Calculated value of - statistic
  Based Calculated value of - statistic the rule of Thumb is *"if the calculated value of -statistic is less than the critical value of i.e.* **crit** *at the desired level of significance, do not reject the null hypothesis, else reject the null hypothesis"*

- -Value
  Based on -Value the thumb rule is *"if -Value is less than the desired level of significance,reject the null hypothesis, else do not reject the null hypothesis"*

☞ Check your progress 1

Analyze the summary statistics of the ANOVA: Single Factor table given above and Answer the following:

1) What is the level of significance at which ANOVA test is performed?

   ....

   ....

   ....

2) What is the critical value of ? Explain by looking up a table of -distribution given in Appendix, Table 1 of Unit 11.

   ....

   ....

   ....

3) Compare the value of statistic with the critical value of . Use the corresponding thumb rule and comment on the Null Hypothesis constructed to study the company objective i.e., to test "whether there is significant difference between the average number of streamers sold by the dealers."

   ....

   ....

   ....

   ....

4) Use the thumb rule for P Value and comment on the Acceptance or Non Acceptance of Null Hypothesis laid for the study of company objective.

   ....

   ....

   ....

   ....

   ....

3.2.2 Two-way Classification

The two-way classified data obtained in an experiment has the following layout with one observation in each treatment:

	Block Block ... Block
Treatment 1	...
Treatment 2	...
Treatment k	...

The linear mathematical model for ANOVA in this case is of the form

, and

and the hypothesis to be tested are (i)For Treatments , and (ii)For Blocks , . Here, is the grand mean, is the treatment effect and is the block effect.

Now, let us extend our discussion for Two-way or Two factor ANOVA test. Based on case analysis, the company has single objective to study and dealer has two, which are to be tested simultaneously. Thus, two factor ANOVA test is desired to be performed for Dealers.

Dealer Objective

1. To Test “whether the salesmen differ in their ability of salesmanship”
2. To Test “whether the locality has any influence on the sales of streamers.”

From the objectives, we identified, that to apply the ANOVA test we need to establish Two Null hypothesis, and , which are to be tested simultaneously. For ,let be the grand mean, be that part of due to the ith salesman and for , let be that part of due to thejth locality. Thus, the Null Hypothesis to be tested are

and .

Now, we will learn how we use excel to perform testing for these hypothesis.

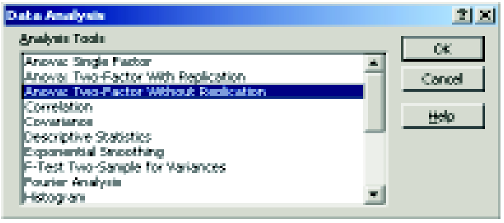
Data analysis through Excel

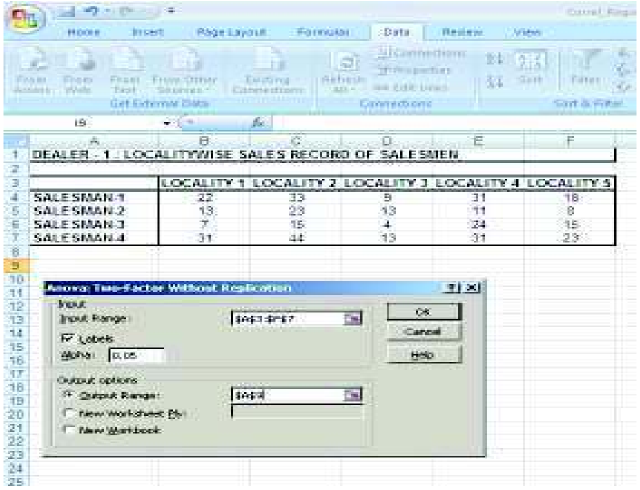
Perform Following Steps:

1. Tabulate the DEALER SALES RECORD as given above in Excel Spreadsheet.
2. Click DATA TAB → DATA ANALYSIS → ANOVA : Single Factor → OK

   “For activation and usage of Data Analysis Toolpak, refer to the earlier unit of Correlation and Regression - The snap shots are readily available there”

   However we are giving some of the relevant screenshots here





DEALER - 1 : LOCALITYWISE SALES RECORD OF SALESMEN

	LOCALITY 1	LOCALITY 2	LOCALITY 3	LOCALITY 4	LOCALITY 5
SALESMAN-1	22	33	9	31	18
SALESMAN-2	13	23	13	11	8
SALESMAN-3	7	15	4	24	15
SALESMAN-4	31	44	13	31	23

Anova: Two-Factor Without Replication

Input

Input Range: $A$3:$F$7

☑ Labels

Alpha: 0.05

Output options

● Output Range: $A$9

○ New Worksheet Ply:

○ New Workbook

Recall that concerned details about *alpha*, *labels*, *output range* are already discussed in single factor ANOVA Test. We now proceed to analyze the results which you will get when OK is clicked, and conclude whether to reject or accept the formulated hypothesis.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
SALESMAN-1	5	113	22.6	96.3
SALESMAN-2	5	68	13.6	31.8
SALESMAN-3	5	65	13	61.5
SALESMAN-4	5	142	28.4	130.8
LOCALITY 1	4	73	18.25	110.25
LOCALITY 2	4	115	28.75	157.5833333
LOCALITY 3	4	39	9.75	18.25
LOCALITY 4	4	97	24.25	88.91666667
LOCALITY 5	4	64	16	39.33333333

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	829.2	3	276.4	8.015466409	0.003373221	3.490295
Columns	867.8	4	216.95	6.291445143	0.005736052	3.259167
Error	413.8	12	34.48333333			
Total	2110.8	19				

Note:

- **F Crit,** is the critical value of -distribution at respective level of significance, which you can get from the table mentioned earlier (Appendix, Unit 11). In addition, -value is generated by Excel for additional data interpretation.
- Notice that the column headings of the ANOVA table are same as earlier, but for the additional row for the columns, which in the present case is for the Localities.
- We can interpret the result based on value of or the -value as stated earlier for both the hypotheses.

☞ Check your progress 2

Analyze the summary statistics of the Anova: Two-Factor Without Replication table given above and Answer the following:

1) At what level of significance at which ANOVA test is performed?
....
....
....

2) What is the critical value of Factor F?
....
....
....

3) Compare the F value with the Critical value of F. Use the thumb rule for F Value and comment on the Acceptance or Rejection of Null Hypothesis H_{01} laid for the study of dealer objective i.e. to Test "whether the salesmen differ in their ability of salesmanship."
....
....
....

4) Compare the F value with the Critical value of F. Use the thumb rule for F Value and comment on the Acceptance or Rejection of Null Hypothesis H_{02}laid for the study of dealer objective i.e. to Test "whether the locality has any influence on the sales of streamers."
....
....
....

5) Use the thumb rule for -Value and comment on the Acceptance or Rejection of Null Hypothesis laid for the study of dealer objective.
....
....
....

3.3 SUMMARY

The practical sessions covered in this unit enabled you to utilize the facility of Data Analysis ToolPak for ANOVA test. Further, it also enriched your understanding by correlating the concepts you studied in BCS 040 with the practical implementation through MS EXCEL. It is important to understand that mere usage of MS Excel or any other software will enable the user to get the standard results/tables etc. without getting into actual act of formula writing, which will require complete knowledge of the mathematical expressions required for computing. The interpretation of results generated through any such software is the sole tasks of the user, for which a complete appraisal of the problem is necessary and hence unavoidable. It has been our effort in this unit (and earlier) to explain the concept of data analysis through suitable example, computation and hence interpretation, which by no means is the end. The journey of data analysis and interpretation is yet to begin under this background.

3.4 EXERCISES

Exercise 1

One important factor in selecting software for word processing and database management systems is the time required to learn how to use a particular system. In order to evaluate three database management systems, a firm devised a test to see how many training hours were needed for five of its word processing operators to become proficient in each of the three systems.

System A	16	19	14	13	18	hours
System B	16	17	13	12	17	hours
System C	24	22	19	18	22	hours

Using a 5% significance level, investigate if there are any differences between the training time needed for the three systems?

Perform the analysis using EXCEL concepts you learned in this course

3.5 SOLUTIONS

Commer
I have NOT checked the results through actual computations.

Sudhansh : the results are displayed after proper calculation

Check Your Progress -1

1) Alpha i.e., , so level of confidence is
2) Critical value of Factor(Fcrit)
3) Because and .Since , we do not reject the NULL Hypothesis.
4) , which is more than the desired significance level, so we do not reject the null hypothesis

Check Your Progress -2

1) Alpha i.e., , thus level of confidence
2) Critical value of (Fcrit) and respectively
3) ; ; Since , we Reject the Null Hypothesis
4) ; ; Since , we Reject the Null Hypothesis
5) Since value greater than Alpha, we Reject the Null Hypothesis

Exercise 1

Here (k = 3 systems, N = 15 values)

Source	S.S.	d.f.	M.S.S.	F
Between systems	103.3	3 - 1 = 2	103.3/2 = 51.65	51.65/6.00 = 8.61
Errors	72.0	14 - 2 = 12	72.0/12 = 6.00	
Total	175.3	15 - 1 = 14		

H_0: $\mu_A = \mu_B = \mu_C$ **H_1**: At least two of the means are different.

Critical value: $F_{0.05}$ (2,12) = 3.89 (Deg. of free. from 'between systems' and 'errors'.)

Test statistic: 8.61

Conclusion: T.S. > C.V. so reject H_0. There is a difference between the mean learning times for at least two of the three database management systems.

SECTION 4 TME SERIES ANAYSIS AND CONTROL CHARTS

Structure **Page Nos.**

4.0 INTRODUCTION

In the previous section, you were introduced to regression analysis, where it was desired to fit a linear (in case of linear regression) or a curve between the dependent and the independent variables. A typical decision making through use of regression analysis may involve plan for future growth of infrastructure of a city based on the growth pattern in population. Thus, prediction primarily focuses on estimating a future value based on the presented study data.

This section introduces the concept of Time series, which you learned in BCS040/Block 3, and how they can be used for forecasting. The Unit begins with an introduction to time series and then shows how you can use a worksheet to forecast using time series data.

Further, the unit introduces the control charts that are used in determining the statistical quality control. You must go through the following two units of BCS040/Block 3 before performing various tasks/steps/activities listed in this section.

Unit 10: Forecasting and Time Series Analysis

Unit 11: Statistical Quality Control

4.1 OBJECTIVES

After performing the activities of this section, you should be able to:

- define time series data
- use a spreadsheet to create time series data
- use spreadsheet to fit the linear trends using the method of least squares
- use spreadsheet for forecasting models

4.2 TIME SERIES AND FORCASTING

In many situations data may be collected over a period of time. This time period may be on days, for example, for shopping analysis you may be interested in recording sales on Mondays, Tuesdays, etc.; a week when you want to plan weekly, a month, a year and so on. Such periodic data is called times series data. Block 3/Unit10/ Page44/ Table 3 which is reproduced in the Figure 1 shows the time series data of crop yields. This data is recorded for a time period of a year over 41 years. This data was collected by Government of AP for studying the changes in the cropping pattern. Such data may be used for first predicting a trend. In this example, it is expected to predict the future economic needs of the Agricultural sector.

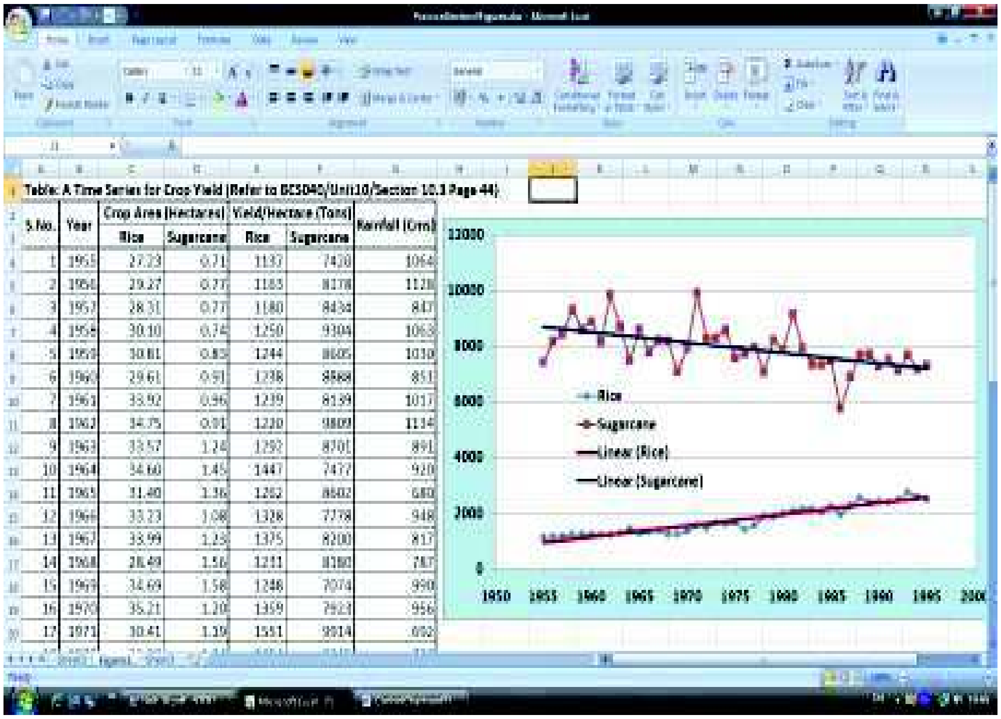
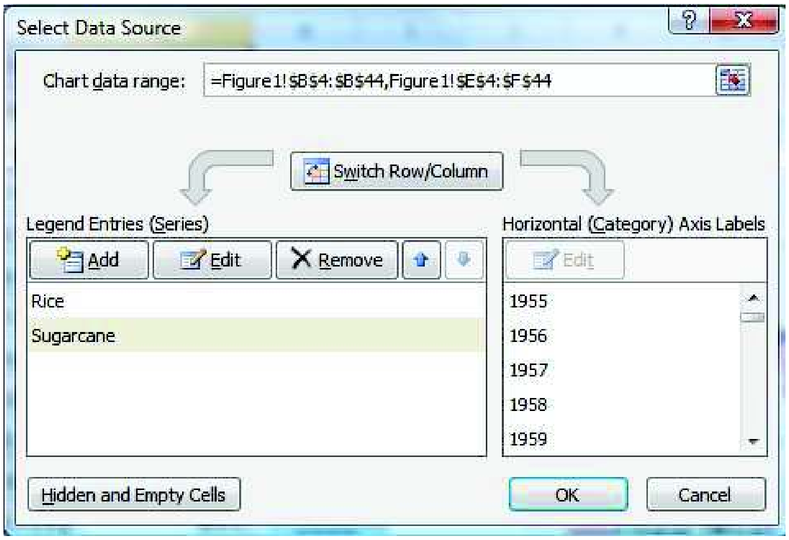


Table: A Time Series for Crop Yield (Refer to BCS040/Unit10/Section 10.3 Page 44)

S.No.	Year	Crop Area (Hectares)		Yield/Hectare (Tons)		Rainfall (Cms)
		Rice	Sugarcane	Rice	Sugarcane	
1	1955	27.23	0.71	1137	7420	1064
2	1956	29.27	0.77	1163	8178	1128
3	1957	28.31	0.77	1180	8434	847
4	1958	30.10	0.74	1250	9304	1063
5	1959	30.81	0.83	1244	8605	1030
6	1960	29.61	0.91	1238	8568	851
7	1961	33.92	0.96	1239	8139	1017
8	1962	34.75	0.91	1220	9809	1134
9	1963	33.57	1.24	1292	8701	891
10	1964	34.60	1.45	1447	7477	920
11	1965	31.40	1.36	1262	8602	680
12	1966	33.23	1.08	1328	7778	948
13	1967	33.99	1.23	1375	8200	817
14	1968	28.49	1.56	1211	8180	787
15	1969	34.69	1.58	1248	7074	990
16	1970	35.21	1.20	1359	7923	956
17	1971	30.41	1.19	1551	9914	692

Figure 1: Long Term Trends for Crop yield

The figure shows the crop yield data and the long term trend for two crops Rice and Sugarcane over a period of 41 years. The Figure also shows the trend line of the two crop yields. To draw the graph we have selected *a linear Scatter plot*. In the *Select Data*… option you can name the Series viz. "Rice" and "Sugarcane. For checking the details of series in a spreadsheet package, right click on the chart, and select *Select Data*… option. You will see the following window:



Select Data Source

Chart data range: =Figure1!$B$4:$B$44,Figure1!$E$4:$F$44

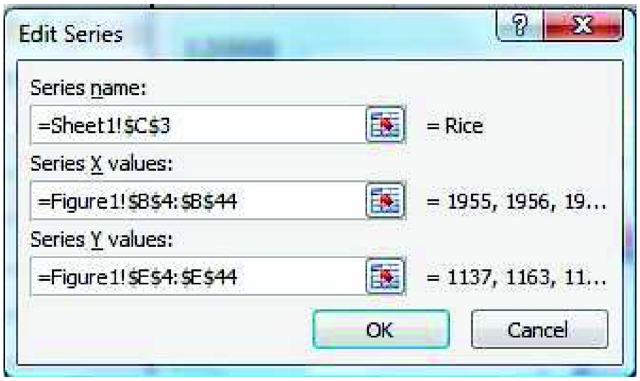
Legend Entries (Series)

Rice
Sugarcane

Horizontal (Category) Axis Labels

1955
1956
1957
1958
1959

You may edit any series, for example, if you select *Rice* Data Series and click on *Edit* Button, you will see the following Range for the series:



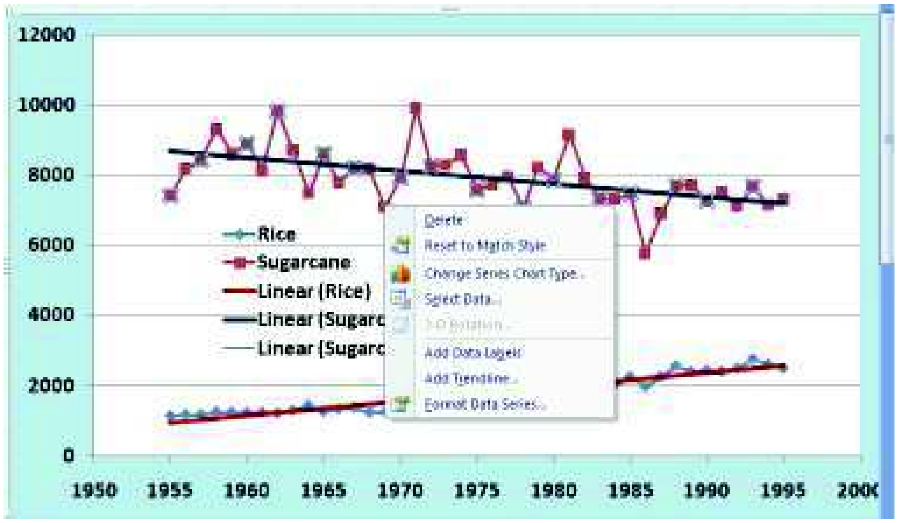
Edit Series

Series name: =Sheet1!$C$3 = Rice

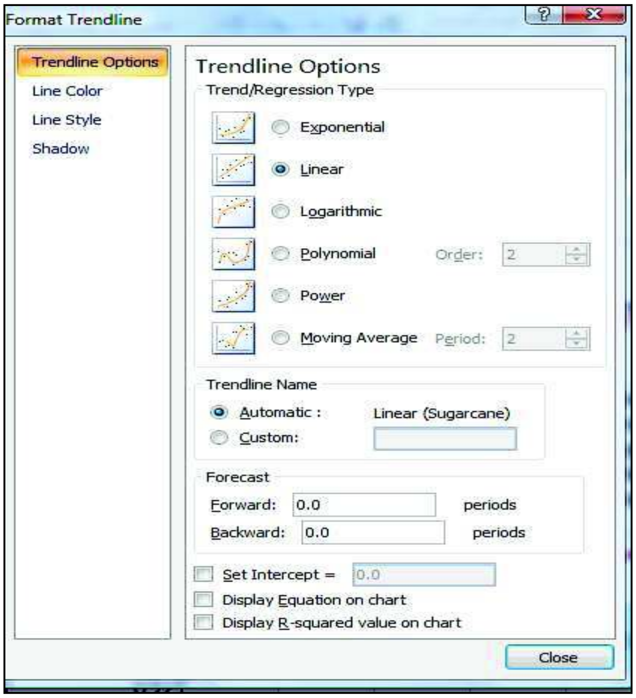
Series X values: =Figure1!$B$4:$B$44 = 1955, 1956, 19...

Series Y values: =Figure1!$E$4:$E$44 = 1137, 1163, 11...

Thus, you can modify the series as per your requirements. Now select any series on the chart and press right mouse button, the following Menu will be displayed:



Using this Menu, you can change series or chart type, select data, add data labels and format data series. You may explore these options on your own. One option which requires a mention due to its usefulness here is *Add Trendline*... On selecting this option you will get the following Menu:



Format Trendline

Trendline Options

Trend/Regression Type

- ○ Exponential
- ● Linear
- ○ Logarithmic
- ○ Polynomial Order: 2
- ○ Power
- ○ Moving Average Period: 2

Trendline Name

- ● Automatic : Linear (Sugarcane)
- ○ Custom:

Forecast

Forward: 0.0 periods

Backward: 0.0 periods

- □ Set Intercept = 0.0
- □ Display Equation on chart
- □ Display R-squared value on chart

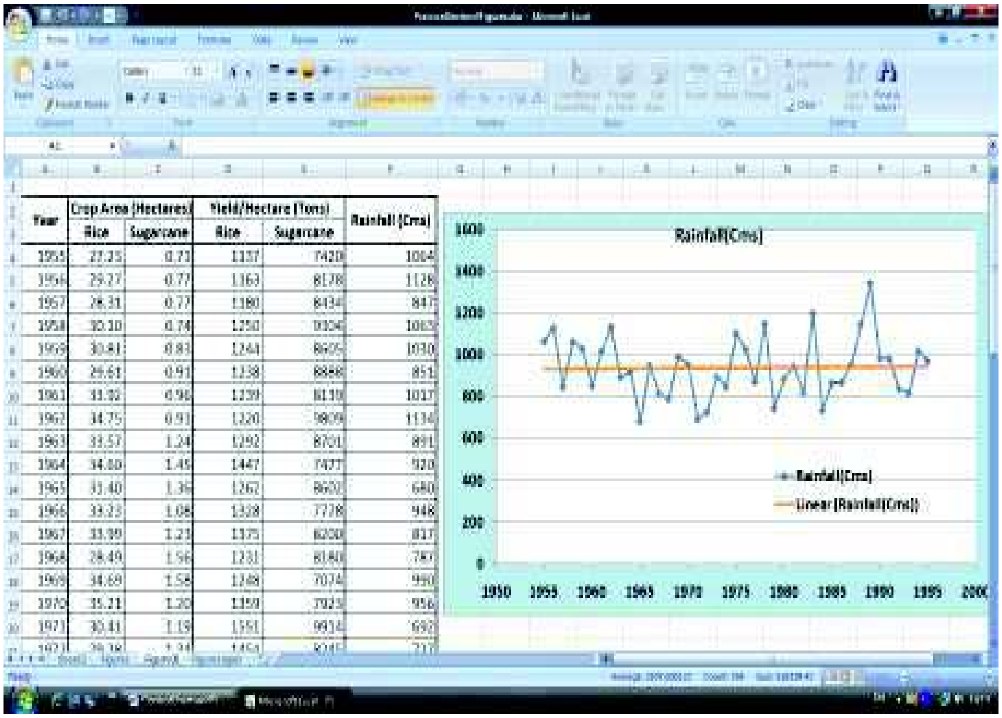
Please note that for Figure 1, you can select Trendline option as Linear. You can also change the Line Color and Line Style using options shown after arrow.

Figure 1(a): Trendline Options

Now observe the data series in Figure 1. There are fluctuations in the time series data. You can attribute these fluctuations on various factors that may include availability and quality of fertilizer, weather conditions etc. You may refer to page 44/Unit 10 for a complete discussion. We briefly present below the following:

Long term trends: From the Figure 1, the long term trend that can be drawn is that average the yield of Rice crop per hectare has increased. On the other hand the yield of Sugarcane per hectare has steadily dropped. These trends can be further analysed to ascertain the reasons of such the declining trends.

Another long term trend about rainfall is shown in Figure 2. You can see that average rainfall over the period of 41 years is almost unchanged.



Year	Crop Area (Hectares)		Yield/Hectare (Tons)		Rainfall (Cms)
	Rice	Sugarcane	Rice	Sugarcane	
1955	27.25	0.71	1137	7420	1004
1956	29.27	0.77	1163	8178	1128
1957	28.31	0.77	1180	8434	847
1958	30.10	0.74	1250	9304	1063
1959	30.81	0.83	1244	8605	1030
1960	29.61	0.91	1238	8888	851
1961	31.92	0.96	1239	8119	1017
1962	34.75	0.93	1220	9809	1134
1963	33.57	1.24	1292	8701	891
1964	34.00	1.45	1447	7477	920
1965	31.40	1.36	1262	8602	680
1966	33.23	1.08	1328	7778	948
1967	33.99	1.23	1175	8200	817
1968	28.49	1.56	1231	8180	787
1969	34.69	1.58	1248	7074	990
1970	35.21	1.20	1359	7923	956
1971	30.41	1.19	1551	9914	692
1972	[illegible]	1.34	1454	[illegible]	717

Figure 2: Trendline for Average Rainfall

Seasonal Variations: As a task, to start with, create the data for the seasonal variations given in the BCS040: Block 3/Unit 10/Page 46 and plot the time series in two possible ways. (i) Plot the figure akin to Fig.5/Unit 10 for all the six years, instead of three shown, with only four quarters along the x-axis. (ii) Plot the time series with years along the x-axis, showing quarters within each of the years. Note the variations that you observe in the curve drawn by the spreadsheet. Discuss with other students or counselors the "relevance" of seasonal variation in data besides commonly encountered examples in real life.

Cyclic variations: Using the data on Sugarcane Area the cyclic variations are plotted in Figure 3 using spreadsheet package. You can observe the cyclic pattern in the plot. Did you understand the distinction between "a season" and "a cycle", which is discussed in section 10.3.3/page 48/Unit 10? Discuss the key points from there with other students or counselors.

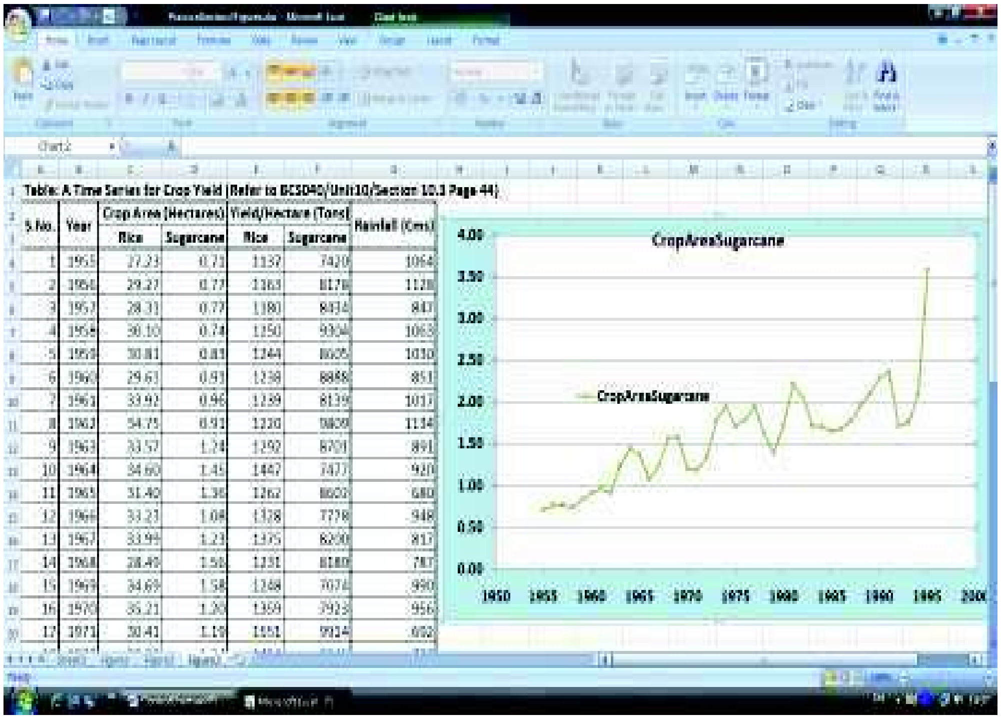
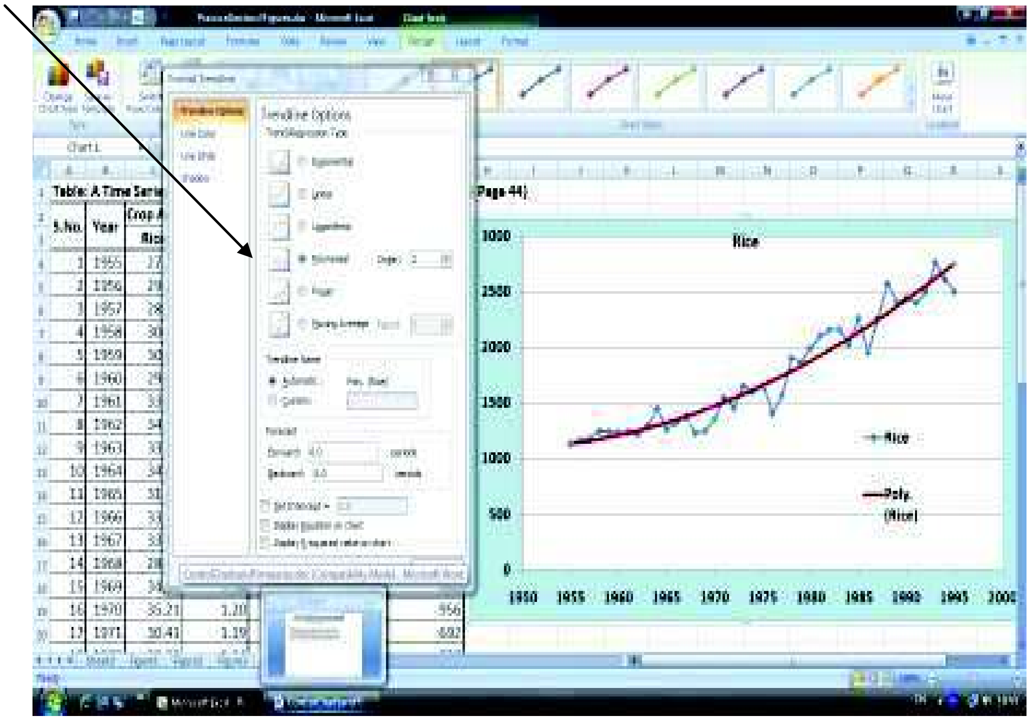


Table: A Time Series for Crop Yield (Refer to BCS040/Unit10/Section 10.3 Page 44)

S.No.	Year	Crop Area (Hectares)		Yield/Hectare (Tons)		Rainfall (Cms)
		Rice	Sugarcane	Rice	Sugarcane	
1	1955	27.23	0.71	1137	7420	1064
2	1956	29.27	0.77	1163	8178	1128
3	1957	28.31	0.77	1180	8434	847
4	1958	30.10	0.74	1250	9304	1063
5	1959	30.81	0.83	1244	8605	1030
6	1960	29.63	0.93	1238	8888	851
7	1961	33.92	0.96	1239	8139	1017
8	1962	34.75	0.91	1220	9809	1134
9	1963	33.57	1.24	1292	8701	891
10	1964	34.60	1.45	1442	7477	920
11	1965	31.40	1.36	1262	8603	680
12	1966	33.23	1.08	1328	7778	948
13	1967	33.99	1.23	1375	8290	817
14	1968	28.49	1.56	1231	8180	787
15	1969	34.69	1.58	1248	7074	990
16	1970	35.21	1.20	1359	7923	956
17	1971	30.41	1.19	1551	9914	692

Figure 3: Cyclic variations

Forecasting Models: Please go through various forecasting models from Section 10.4/page 49/MCS040/Block3/Unit 10, viz., *additive* and *multiplicative* models. You may observe that Trendline actually is a forecasting curve line. In Figure 1 and Figure 2, you have used a linear fit. You can revise from Unit 9 or MCS040/Block3/Unit 10.5.1 the method of least squares method. Further, you can select the method of moving averages (MCS040/Block3/Unit 10.5.2) by simply selecting the Trendline Options as shown in Figure 1(a) or you can fit the Polynomial of second degree using least square method, by simple selecting Polynomial Options and keeping the order as 2 (see Figure 4).



Format Trendline

Trendline Options
Trend/Regression Type
○ Exponential
○ Linear
○ Logarithmic
● Polynomial Order: 2
○ Power
○ Moving Average Period:

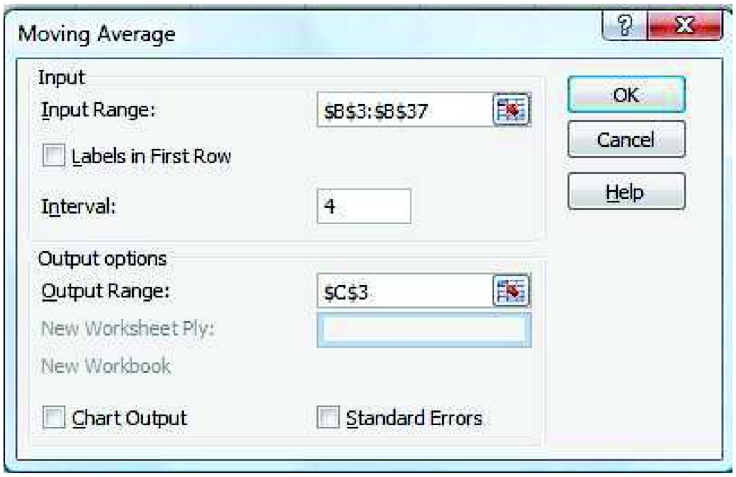
Trendline Name
● Automatic: Poly. (Rice)
○ Custom:

Forecast
Forward: 0.0 periods
Backward: 0.0 periods

☐ Set Intercept = 0.0
☐ Display Equation on chart
☐ Display R-squared value on chart

Figure 4: The Polynomial Trendline of Order 2

You can also select the Moving Average Trendline option from the same window. Another alternative approach for calculating moving averages is that you can use *Data* → *Data Analysis* → *Moving Average* option. On selection of this option you will be presented the following Moving Average Window:



Moving Average

Input

Input Range: $B$3:$B$37

□ Labels in First Row

Interval: 4

Output options

Output Range: $C$3

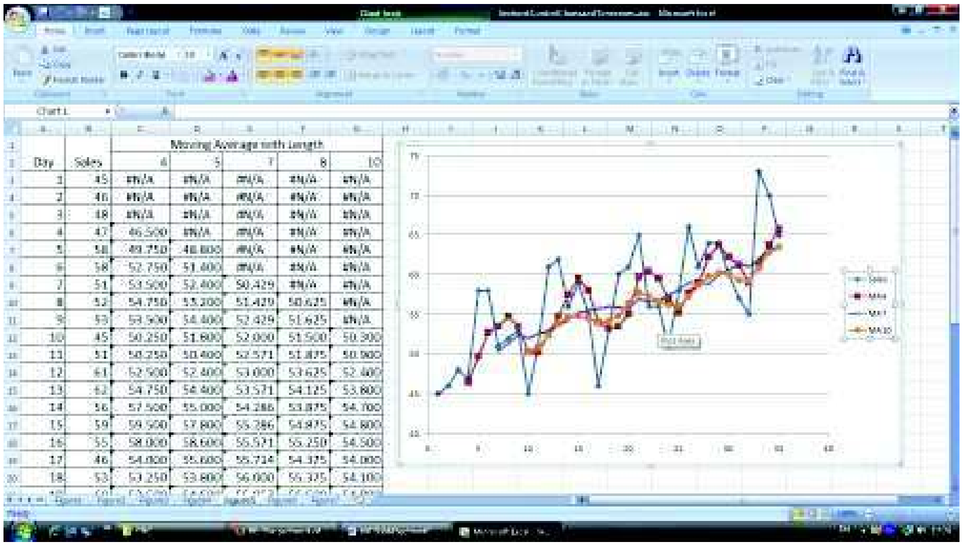
New Worksheet Ply:

New Workbook

□ Chart Output

□ Standard Errors

Select the input range, interval and output range. If you want a chart of moving averages then you can select Chart Output option box. Figure 5 shows moving averages with various intervals using repeated application of this command.



Day	Sales	Moving Average with Length				
		4	5	7	8	10
1	45	#N/A	#N/A	#N/A	#N/A	#N/A
2	46	#N/A	#N/A	#N/A	#N/A	#N/A
3	48	#N/A	#N/A	#N/A	#N/A	#N/A
4	47	46.500	#N/A	#N/A	#N/A	#N/A
5	58	49.750	48.800	#N/A	#N/A	#N/A
6	58	52.750	51.400	#N/A	#N/A	#N/A
7	51	53.500	52.400	50.429	#N/A	#N/A
8	52	54.750	53.200	51.429	50.625	#N/A
9	53	53.500	54.400	52.429	51.625	#N/A
10	45	50.250	51.800	52.000	51.500	50.300
11	51	50.250	50.400	52.571	51.875	50.900
12	61	52.500	52.400	53.000	53.625	52.400
13	62	54.750	54.400	53.571	54.125	53.800
14	56	57.500	55.000	54.286	53.875	54.700
15	59	59.500	57.800	55.286	54.875	54.800
16	55	58.000	58.600	55.571	55.250	54.500
17	46	54.000	55.600	55.714	54.375	54.000
18	53	53.250	53.800	56.000	55.375	54.100

Figure 5: Finding Moving Averages

Thus, you can perform the long term trend analysis using spreadsheet packages.

☞ Lab Sessions 7

1) Develop all the spreadsheets as shown in various diagrams in these previous sections.
2) Workout all the Examples given in the Unit 10 using spreadsheet package.
3) Workout the Exercises of BCS040/Blockl2/Unit 10 using spreadsheet wherever required, or generate sample data wherever possible using spreadsheet functions of random number generation. Alternatively, look for other examples from references with sample data.

4.3 CONTROL CHARTS

This section discusses about the use of a spreadsheet package in constructing a control chart. You should go through BCS-040 Block 3/ Unit 11 before going through this section. Let us first revise some of the terms used in that Block.

Quality and Quality characteristics: The term Quality may be broadly be defined as conformance of a product to certain standards. Quality depends on certain characteristics related to the product. When you design a product certain specifications or levels of tolerances are established on all important quality characteristics of the product. These characteristics are called the quality For example, in case of a ball pen, the specifications on refill length may be that it should lie between 9.90 crns and 10.10 crns.

Controlling a Process: The control operations of a process has to be such that the quality characteristics of output product are maintained at the desired levels.

Statistical Process Control: is a methodology that collects and analyses data on quality characteristics periodically from the process, to take appropriate actions in respect of quality monitoring of a process. Control charts are a most widely used technique for Statistical Process Control.

Variations and Stable/unstable process: A certain amount of variation in quality characteristics (for example, length of refill) is unavoidable despite stable conditions (for example, same machine settings, almost identical material quality and experience of operators). The small changes in quality characteristics may be attributed to small conditional changes. These small causes are inevitable and are called chance causes. The variation caused by such changes is called *chance cause* variation. However, some major chances, such as change in the machine settings, major drop in the quality of raw material etc., disturb the stable process. Such variations must be identified and corrected. Since, you can attribute change to one or more particular reasons, they are called assignable causes. Such causes may make a process unstable.

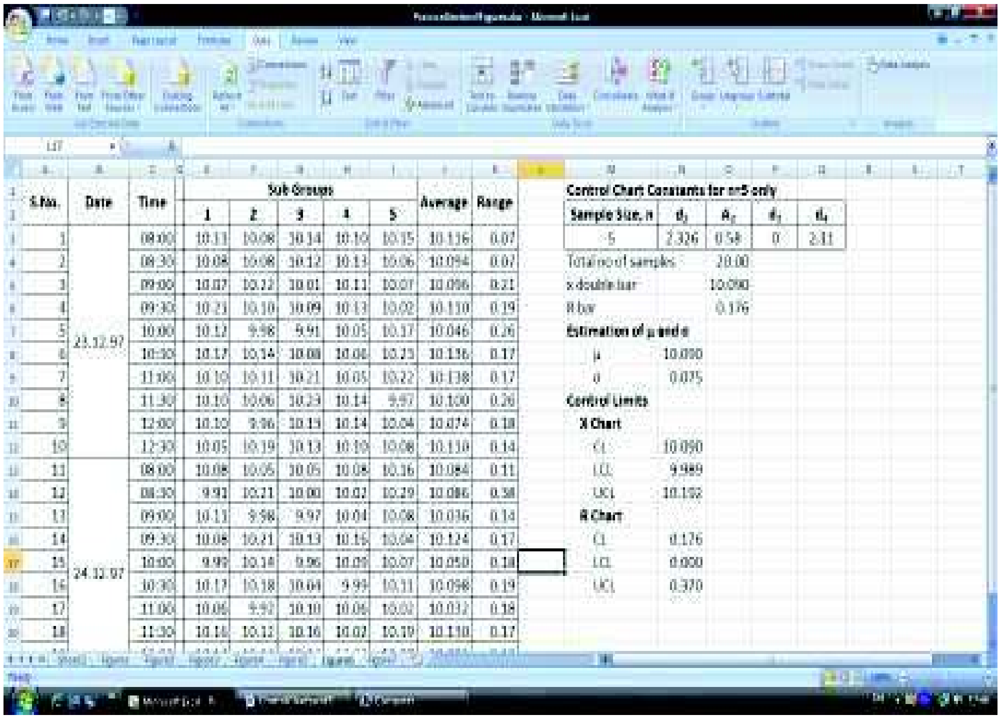
Control Chart Construction steps:

- Collect sample periodically
- Compute quality characteristics for the collected sample
- Plot sample number (-axis) vs quality characteristic (-axis)
- Check if the value of quality characteristics is between the *upper* and *lower control limits*.
- In general, the quality characteristics follow normal distribution. Therefore, the upper and lower limits may be determined by using "*Three Sigma Limits*" as: lower limit and upper limit .
- If quality characteristics are a variable, then control in terms of both the mean and the variability is sought.

Control chart can tell us when process has become unstable, but not that why did it happen?

Problem 1: (Reference MCS040/Block 3/Unit 11/Page 83), Data for the length of refills and control chart constants (only for sample size ( ) are shown in Figure 6. This data is drawn from the Table 1 and Table 2 of the said Unit. You may now estimate and from the refill length data and calculate the control limits for and charts.

The said Unit shows the calculations that are performed with the help of spreadsheet and shown in Figure 6.



S.No.	Date	Time	Sub Groups					Average	Range
			1	2	3	4	5		
1	23.12.97	08:00	10.13	10.08	10.14	10.10	10.15	10.116	0.07
2		08:30	10.08	10.08	10.12	10.13	10.06	10.094	0.07
3		09:00	10.07	10.22	10.01	10.11	10.07	10.096	0.21
4		09:30	10.21	10.10	10.09	10.13	10.02	10.110	0.19
5		10:00	10.12	9.98	9.91	10.05	10.17	10.046	0.26
6		10:30	10.17	10.14	10.08	10.06	10.23	10.136	0.17
7		11:00	10.10	10.11	10.21	10.05	10.22	10.138	0.17
8		11:30	10.10	10.06	10.23	10.14	9.97	10.100	0.26
9		12:00	10.10	9.96	10.13	10.14	10.04	10.074	0.18
10		12:30	10.05	10.19	10.13	10.10	10.08	10.110	0.14
11	24.12.97	08:00	10.08	10.05	10.05	10.08	10.16	10.084	0.11
12		08:30	9.93	10.21	10.00	10.02	10.29	10.086	0.38
13		09:00	10.11	9.98	9.97	10.04	10.08	10.036	0.14
14		09:30	10.08	10.21	10.13	10.16	10.04	10.124	0.17
15		10:00	9.99	10.14	9.96	10.09	10.07	10.050	0.18
16		10:30	10.17	10.18	10.04	9.99	10.11	10.098	0.19
17		11:00	10.06	9.92	10.10	10.06	10.02	10.032	0.18
18		11:30	10.16	10.12	10.16	10.02	10.19	10.130	0.17

Control Chart Constants for n=5 only

Sample Size, n	d_2	A_2	d_3	d_4
5	2.326	0.58	0	2.11

Total no of samples		20.00
x double bar		10.090
R bar		0.176
Estimation of µ and σ		
µ	10.090	
σ	0.075	
Control Limits		
X Chart		
CL	10.090	
LCL	9.989	
UCL	10.192	
R Chart		
CL	0.176	
LCL	0.000	
UCL	0.370	

Figure 6: Calculation of µ and σ for refill length

For creating Figure 6, you can use your own formulas.

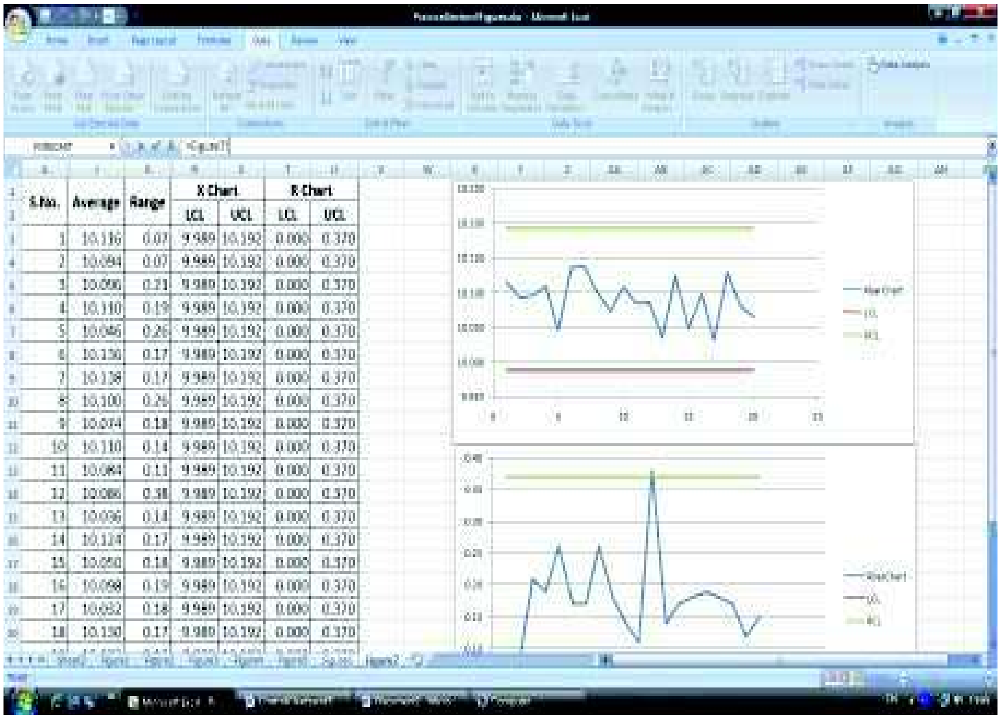
Steps

- You first create the sample data.
- The mean of the sample subgroup at a particular time is found in cell J3 using the formula and similarly
- the Range for this subgroup is calculated at cell K3 using the formula . For .
- On the basis of these mean and range values and is calculated in the cell O5 using the formula *=SUM(J3:J22)/O4* and cell O6 using the formula *=SUM(K3:K22)/O4* respectively.
- On the basis of these values and control chart constants CL, LCL and UCL values are calculated for the and chart. Write the formula for these calculations and compare the values with the values calculated in the Unit.

Further,

- on the basis of these calculated values, you can create data as shown in Figure 7.
- This data is used to draw the control chart.
- You may use Scatter Plot to draw the charts as shown in Figure 7.
- You may keep the x-axis same for all the Ranges that are being plotted as Y axis.

Please note that in Figure 7, that the 12th point on the R-chart crosses the UCL line, which is indicative of “process out of control” situation with respect to variability (as this problem has been detected in R-chart). Consequently, we conclude that during this time period the process may not be stable.



S.No.	Average	Range	X Chart		R Chart	
			LCL	UCL	LCL	UCL
1	10.116	0.07	9.989	10.192	0.000	0.370
2	10.094	0.07	9.989	10.192	0.000	0.370
3	10.090	0.21	9.989	10.192	0.000	0.370
4	10.110	0.19	9.989	10.192	0.000	0.370
5	10.046	0.26	9.989	10.192	0.000	0.370
6	10.130	0.17	9.989	10.192	0.000	0.370
7	10.138	0.17	9.989	10.192	0.000	0.370
8	10.100	0.26	9.989	10.192	0.000	0.370
9	10.074	0.18	9.989	10.192	0.000	0.370
10	10.110	0.14	9.989	10.192	0.000	0.370
11	10.084	0.11	9.989	10.192	0.000	0.370
12	10.086	0.38	9.989	10.192	0.000	0.370
13	10.036	0.14	9.989	10.192	0.000	0.370
14	10.124	0.17	9.989	10.192	0.000	0.370
15	10.050	0.18	9.989	10.192	0.000	0.370
16	10.098	0.19	9.989	10.192	0.000	0.370
17	10.032	0.18	9.989	10.192	0.000	0.370
18	10.130	0.17	9.989	10.192	0.000	0.370

Figure 7: Control Charts (x-chart and R-chart) for the refill length data

Now you should try the following lab session.

☞ **Lab Sessions 8**

1) Develop all the spreadsheets as shown in various diagrams in the previous section.

2) Perform all the examples given in the Unit 11 using spreadsheet package including example given in section 11.3.5 on control charts for attributes.

3) Perform all Exercises of BCS040/Block3/Unit 11 that you can perform using spreadsheet.

4) Write simple C functions to calculate various values required for drawing control charts.

4.4 SUMMARY

This section is an attempt to provide you details on some of the elementary methods of Unit 10 based on which you can perform time series analysis and forecasting using the spreadsheet. Time series analysis help in performing analysis of data in terms of the components and the models discussed in the said Unit 10. The example demonstrated the use of spreadsheet package functions in evaluating trend.

This section also covered computational aspects relating to control charts in statistics using spreadsheet package. The purpose of these charts is to determine if process is under statistical control.

4.5 FURTHER READINGS

Books

1. BCS-040: Statistical Techniques, IGNOU Material.

Web link:

- www.wikipedia.org

SECTION 5 SAMPLING – CASE STUDY

Structure **Page Nos.**

5.0 INTRODUCTION

In our previous sessions you learned about various tools, techniques and practices required to perform the statistical analysis of data. In general, it is observed that the outcome of such analysis (sample estimates)varies from the population values. A common reason for this lies in the size of sample data collected. Thus, the question is "How much data is sufficient?" and the answer relates to sampling. Recall earlier discussion on the population and the sample in Section 1.

It is impractical to study the entire population, so it is devised to study proportionate part of the population, for reliable interpretation of the population behavior. Thus the need of sampling techniques was identified. There are various methodologies devised to perform sampling of identified population, we will try to put light on most of the methodologies, which are relevant at your level of study. Further, you should keep following facts in your mind, that sampling is simply not a formula but in practice it is the Science combined with the Art of calculations and presentation.

You are advised to refer to Block 4 of BCS 040, and study the details about sampling and its related techniques. However, we will discuss about its practical aspects and how to attempt it as a practice, by using excel and its tools.

5.1 OBJECTIVES

After going through this unit you will be able to :

- Use Data Analysis Toolpak for Descriptive statistics of sampled data
- Use Descriptive statistics to calculate sample size through Excel
- Use Excel functions for Sampling

5.2 SAMPLING – SAMPLE SIZE

Since computer is a discipline, that facilitates the working of our day to day life, we choose an example/Case from our daily life in this context and use it to interpret the respective results. Through the case study we would like to emphasize that there are situations when the basic assumptions made while computing the sample size by using the formulas are not met in practice. At other times there are factors which are influential in increasing or decreasing the sample size obtained through the use of formula. It should be kept in mind that sample size determination is the blend of using formulae, experience of similar studies, time and budget constraints and a few other elements such as output or analysis requirements etc.

Through the case, we are going to concentrate on the utilization and implementation of the sample size calculation formula for continuous or interval-scaled variables.

However, for other details you are advised to refer to Block 4 of BCS 040, and apply the skills learned in these practical sessions.

CASE DESCRIPTION :

- As a pilot project, a private company involved in hardware and maintenance related activity, planned to launch its own brand of assembled computers. The company identified a small colony having 12 families in all. Each of the family is having members of different age group. For the sake of their brand performance related study, they decided to issue one computer system to each family as a single unit. They got the system installed with several applications suitable for each age group viz., Computer Games, Movies, Application Software's, E-Books, Internet, Security related software. The company wants to study the application usage pattern and the satisfaction level of the families, so that they can customize their product accordingly. In order to monitor the system usage time, they installed an automatic clock with the system, which records the usage time. This will help them to analyze the questionnaire feedback, which they wish to get filled, at the end of their study.

On the basis of an appraisal of the case cited above, you may identify the objectives of the study to be as under:

- o What is the customer satisfaction level?
- o Which software utility is performed exhaustively in most of the families?

In view of these objectives, you are required to draft a Questionnaire, where the responses to the questions are either direct or indirect, as desired. They can be categorized into one among the following scales, viz.,

- Nominal
- Ordinal
- Quantitative.

A sample questionnaire is given below, this will let you to understand the excel operations, to be performed at the later stages

QUESTIONNAIRE

Family No.:
What is the Average System Utilization time(minutes) ? :

15 30 45 60 75 90 105 120 135 150 165 180

What are the Activities Performed by your family while utilizing the system?

a) Computer Games, b) Internet, c) E-Books,

d) Application Software's, e)Movies f) Security related software

What is the level of satisfaction, achieved by your family ?

-2 = Very dissatisfied, 1 = dissatisfied, 0 = indifference,

1 = satisfied, 2 = Very satisfied.

Would your family like to purchase the product?

Yes No

We got the questionnaire filled from respective families, finally we have the following data from the computer savvy families. The collected data is through following variables

- *Family* (Nominal scale)is the observation/serial number of the questionnaire. In the present case, one questionnaire per household is filled by each family.
- *Time* (Quantitative scale)is a quantitative data type which is measured in minutes. It is the software utilization time for a family.
- *System Utilization Activity* (Nominal scale)is a nominal data type consisting of 6 choices of activity in the park:

  1. Computer Games,
  2. Internet,
  3. E-Books,
  4. Application Software's,
  5. Movies
  6. Security related software

  To this question, there may be multiple choices that for each family, one corresponding to each of the several activities they perform.

- *Product Satisfaction* (Ordinal scale) It measures family satisfaction toward the Product. It is measured in the ordinal scale with 5 values as options listed below:

  -2 = Very dissatisfied,
  1 = dissatisfied,
  0 = indifference,
  1 = satisfied,
  2 = Very satisfied.

- *Product Purchase* (Nominal scale)is measured in the nominal scale (Yes or No) regarding the final decision about purchase of the product.

Based on the responses to the questions framed in the questionnaire, collected data is presented in the following table:

Raw Data

Family	*System Utilization Time*	*System Utilization Activity*	*Product Satisfaction*	*Product Purchase*
1	30	1, 2, 3	0	N
2	30	4,6	1	Y
3	60	1, 2	2	Y
4	45	5	-1	N
5	30	6	1	N
6	60	2	2	Y
7	30	4	1	N
8	45	3, 4	-1	N
9	15	6	1	Y
10	60	2	2	Y
11	180	1, 2, 3, 4	2	Y
12	120	1,2,4	2	Y

Raw data presented in the table above is coded into the following form, which is amenable for further statistical analysis using spreadsheet software.

Raw Data coded for analysis in Formatted Excel Sheet

Family	*System Utilization Time*	*System Utilization Activity*						*Product Satisfaction*	*Product Purchase*
		1	2	3	4	5	6		
1	30	1	1	1	0	0	0	0	0
2	30	0	0	0	1	0	1	1	1
3	60	1	1	0	0	0	0	2	1
4	45	0	0	0	0	1	0	-1	0
5	30	0	0	0	0	0	1	1	0
6	60	0	1	0	0	0	0	2	1
7	30	0	0	0	1	0	0	1	0
8	45	0	0	1	1	0	0	-1	0
9	15	0	0	0	0	0	1	1	1
10	60	0	1	0	0	0	0	2	1
11	180	1	1	1	1	0	0	2	1
12	120	1	1	0	1	0	0	2	1

After the excel file with data in the format shown above has been created from the Raw Data table, notice that the responses under the head "*Product Purchase*" are marked 0 and 1, which represents the codes for responses "Y" and "N", respectively. The same also applies to the responses under "*System Utilization Activity*" head of Raw Data table.

Please note that Family, System Utilization Activity, Product Purchase, are variables measured in NOMINAL SCALE. So, we should NOT calculate sum, average, variance etc. Instead the best is to calculate frequency, percentage, plot graphs only. But here we are going to demonstrate that how you can use the tool of data analysis tool pack to get the desired answers, related to the marked objectives of your study. Now to get the answer for the stated objectives, we use the Descriptive Statistics Tool of Data Analysis ToolPak.

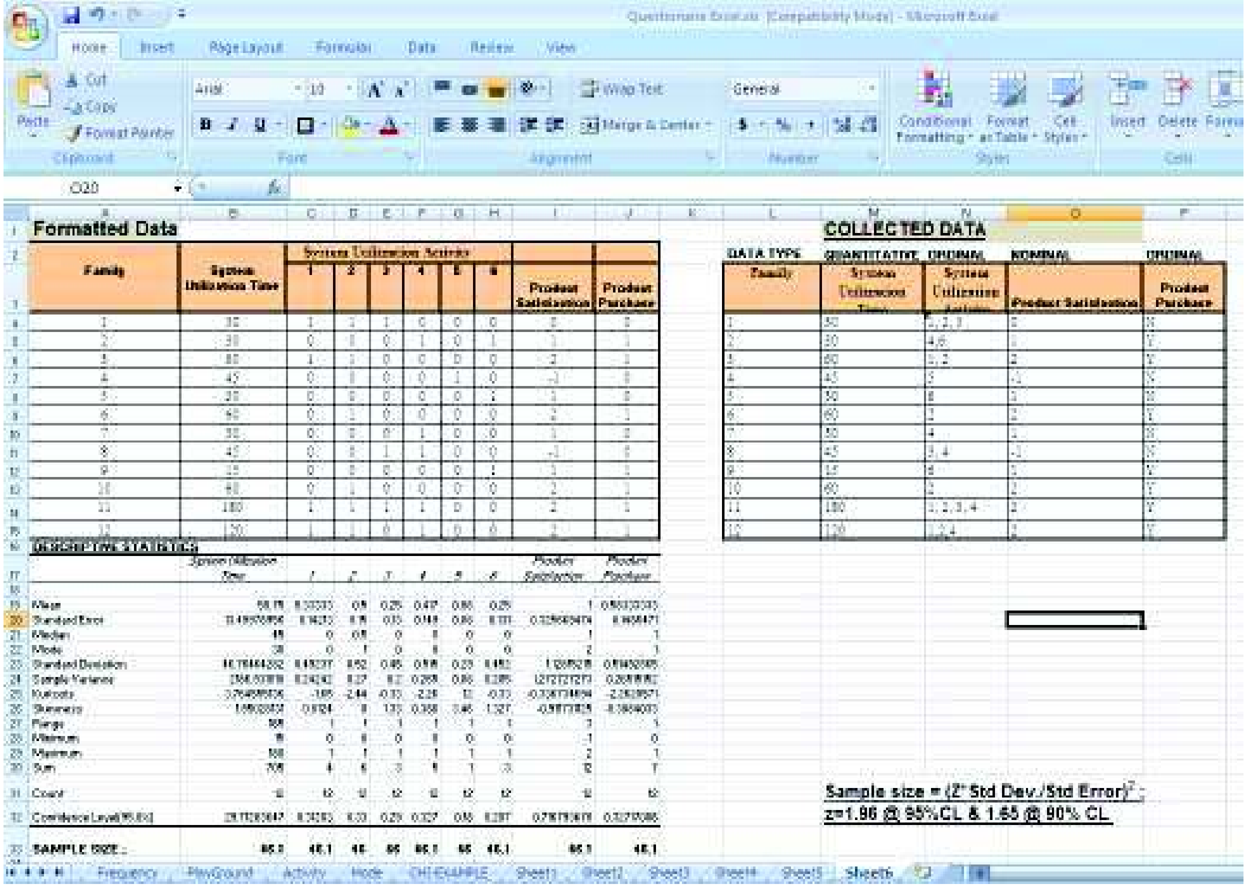
Data analysis through Excel:

Perform Following Steps:

1. Tabulate the collected data as given above in Excel Spreadsheet.
2. Click DATA TAB → DATA ANALYSIS → DESCRIPTIVE STATISTICS → OK

Note:"For activation and usage of Data Analysis Toolpak, refer to the earlier section1 & section Correlation and Regression - The snap shots are readily available there" .

We present below some of the relevant screenshots in the present case.



Formatted Data

Family	System Utilization Time	System Utilization Activity 1	2	3	4	5	6	Product Satisfaction	Product Purchase
1	30	1	1	1	0	0	0	2	0
2	30	0	0	0	1	0	1	1	1
3	60	1	1	0	0	0	0	2	1
4	45	0	0	0	0	1	0	-1	0
5	30	0	0	0	0	0	1	1	0
6	60	0	1	0	0	0	0	2	1
7	30	0	0	0	1	0	0	1	0
8	45	0	0	1	1	0	0	-1	0
9	15	0	0	0	0	0	1	1	1
10	60	0	1	0	0	0	0	2	1
11	180	1	1	1	1	0	0	2	1
12	120	1	1	0	1	0	0	2	1

DESCRIPTIVE STATISTICS

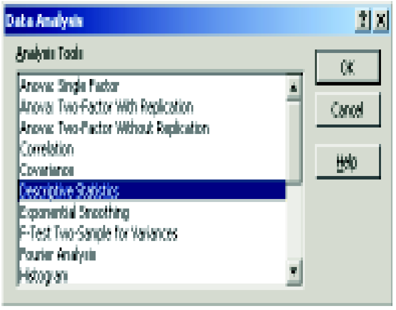
	System Utilization Time	1	2	3	4	5	6	Product Satisfaction	Product Purchase
Mean	58.75	0.333333	0.5	0.25	0.417	0.08	0.25	1	0.58333333
Standard Error	13.49970856	0.14213	0.15	0.13	0.149	0.08	0.131	0.32566947	0.1486471
Median	45	0	0.5	0	0	0	0	1	1
Mode	30	0	1	0	0	0	0	2	1
Standard Deviation	46.76464282	0.49237	0.52	0.45	0.515	0.29	0.452	1.1281521	0.51492865
Sample Variance	2186.93	0.24242	0.27	0.2	0.269	0.08	0.205	1.27272727	0.26515152
Kurtosis	3.76499	-1.65	-2.44	-0.33	-2.26	12	-0.33	-0.33873494	-2.2628571
Skewness	1.99	0.8124	0	1.33	0.388	3.46	1.327	-0.9177825	-0.3984083
Range	165	1	1	1	1	1	1	3	1
Minimum	15	0	0	0	0	0	0	-1	0
Maximum	180	1	1	1	1	1	1	2	1
Sum	705	4	6	3	5	1	3	12	7
Count	12	12	12	12	12	12	12	12	12
Confidence Level(95.0%)	29.71283647	0.31283	0.33	0.29	0.327	0.18	0.287	0.71679381	0.3271086
SAMPLE SIZE :	46.1	46.1	46	46	46.1	46	46.1	46.1	46.1

COLLECTED DATA

DATA TYPE	QUANTITATIVE	ORDINAL	NOMINAL	ORDINAL
Family	System Utilization Time	System Utilization Activity	Product Satisfaction	Product Purchase
1	30	1, 2, 3	2	N
2	30	4,6	1	Y
3	60	1, 2	2	Y
4	45	5	-1	N
5	30	6	1	N
6	60	2	2	Y
7	30	4	1	N
8	45	3, 4	-1	N
9	15	6	1	Y
10	60	2	2	Y
11	180	1, 2, 3, 4	2	Y
12	120	1,2,4	2	Y

Sample size = (Z* Std Dev./Std Error)^2 ;
z=1.96 @ 95%CL & 1.65 @ 90% CL

Results of DESCRIPTIVE STATISTICS Procedure



Statistics	System Utilization Time	System Utilization Activity 1	2	3	4	5	6	Product Satisfaction	Product Purchase
Mean	58.75	0.33	0.50	0.25	0.42	0.08	0.25	1	0.58
Standard Error	13.50	0.14	0.15	0.13	0.15	0.08	0.13	0.33	0.15
Median	45	0	0.5	0	0	0	0	1	1
Mode	30	0	1	0	0	0	0	2	1
Standard Deviation	46.76	0.49	0.52	0.45	0.51	0.29	0.45	1.13	0.51
Sample Variance	2186.93	0.24	0.27	0.20	0.27	0.08	0.20	1.27	0.27
Kurtosis	3.8	-1.65	-2.44	-0.33	-2.26	12	-0.33	-0.34	-2.26
Skewness	2.0	0.81	0	1.33	0.39	3.46	1.33	-0.91	-0.39
Range	165	1	1	1	1	1	1	3	1
Minimum	15	0	0	0	0	0	0	-1	0
Maximum	180	1	1	1	1	1	1	2	1
Sum	705	4	6	3	5	1	3	12	7
Count	12	12	12	12	12	12	12	12	12
Sample size = (Z*Std Dev./Std Error)^2	46.0992	46.099	46.099	46.0999	46.0999	46.099	46.099	46.099	46.0992
Z=1.96 @ 95%CL & 1.65 @ 90% CL						46.099	46.099	46.0992	46.0992

DATA INTERPRETATION

Now, let's interpret the results of descriptive statistics and analyse, what they say about our objectives viz.

- What is the customer satisfaction level?
- Which software utility is performed exhaustively in most of the families?

So far as *customer satisfaction level is concerned, the tabulated data under head of Mean and Mode gives the result as 1 and 2 respectively* i.e.,on an average the customers are satisfied and in fact they are "most frequently" - very satisfied. To interpret we are to refer our data coding standards. As in this case they are :

- *Product Satisfaction* is an ordinal scale with 5 values:

  -2 = Very dissatisfied,
  1 = dissatisfied,
  0 = indifference,
  1 = satisfied,
  2 = Very satisfied.

Similarly, in order to study the objective "*Which software utility is performed exhaustively in most of the families? Refer to the results under the head SUM for the* ***System Utilization Activity****. Here, the activity 2 is identified to be performed the most* i.e. *Families are mostly using the Computer system for internet related activities.*

To interpret we are to refer our data coding standards. As in this case they are:

- *System Utilization Activity* is a nominal data type consist of 6 choices of activity in the park:

  1. Computer Games,
  2. Internet,
  3. E-Books,
  4. Application Software's,
  5. Movies
  6. Security related software

Now, you might be wondering that the purpose of this session is to talk about sampling and we are discussing about the descriptive statistics. This is intentional, we want you to realize that sampling is not simply a formula but it is related to various factors, as in this case the number of families are limited in the target area i.e. target population is limited. By using the formula for Sample size calculation i.e. *Sample size = (Z*Std Dev. / Std Error)^2 ; we can calculate the respective sample size as*

Sample size = (Z*Std Dev. / Std Error)^2

You might have not located this expression *Sample size=(Z*StdDev./Std Error)^2* in the Units12,13,14 of BCS 040. But, for the sake of understanding its just another formula for calculating the sample size, when standard deviation is known. However when standard deviation is not known we use another formula,we are not going to discuss that, here we are going to concentrate on *Sample size=(Z*StdDev./Std Error)^2*

We calculated the sample size to be 46 families but our study is considering only 12. So for the reliability of the study the collected data is insufficient and thus we cannot conclude that the interpretation of the results related to our objectives is fine enough.

You might be in the position to realize the importance of sampling now, because from the above case you identified that by limited data you might find that the objectives of the study are fulfilled but which might not be the fact is the sample size is inappropriate.

After Sample size calculation lets understand how to use excel for Random Sampling. You might have studied various sampling methods in block 4 of BCS 040, but we are confining our discussion to only Random sampling, you can implement the rest of the sampling techniques by applying the understanding of BCS 040 and Excel concepts, learned in these sessions.

5.3 RANDOM SAMPLING

Let's extend our discussion from past session, where we considered the sample size of 12 families and based on the descriptive statistics results we analyzed that the sample size is not large enough for the study. But if we assume that the population itself is only 12 families and sample size turns out to be only 4 families, and we are going to opt for random sampling. Then among the 12 families, how to choose 4 families at random, this is the task we are going to demonstrate in this session. Among these 12 families 4 families are to be chosen at Random so we firstly Randomize the entire population data by using Rand() function. We, simply write Rand() in the subsequent column and apply it for entire population data. But, prior to that we should understand that the sampling mechanism is random sampling, so, to draw the sample by Simple Ransom Sample discussed in section 12.4/ Unit-12 of BCS 040, these probabilities have to be all equal, and we have to draw only 4 random numbers to select the required 4 sample units. Thus, we need to determine the equal probabilities of being chosen, which turns out to be (1/12) where 12 is total number of families, the result 0.08 is the equal probability assigned to each family, as shown below.

You understood the theoretical way of doing it, in our block 4 of BCS 040. Here, we are going to use Excel as the tool for the mentioned purpose. We are going to explore, following functions to perform the task RAND(), SMALL(), MATCH() and INDEX() functions to perform the above framed task i.e. to do Random Sampling.

Say the population data of 12 families under consideration is as follows

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	*Family*	1	2	3	4	5	6	7	8	9	10	11	12
2	*Surname*	Sharma	kumar	Sastry	Venugopal	Gupta	Nagpal	Dhiman	Kaushik	Bhardwaj	Garg	Giri	Iyengar
3	*Probability* ***(1/12=0.08)***	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
4	*Random#*	0.88	0.03	0.43	0.22	0.40	0.82	0.96	0.44	0.55	0.66	0.04	0.86

Task:
From among these 12 families, 4 families are to be chosen at Random.

Steps:

1. At ever draw equal probabilities have to be assigned to the units (see 3rd paragraph/section 12.4/Unit-12).These probabilities are— as shown in C3..C14
2. Use excel function RAND( ) to generate random numbers in the range .
3. Calculate cumulative probabilities as shown in D3..D14.
    a. Value 0 in D2 is marked for the calculation of cumulative frequency
4. Consider a discrete random variable having probability mass function , and . Then we generate an uniform random number and the corresponding observation , if

5. Repeat the above step times to draw the required random sample. This is shown in columns E, F, G and H.

	A	B	C	D
1	*Family*	*Surname*	*Probability*	*Cumulative Probability*
2				0
3	1	Sharma	0.08	0.08
4	2	Kumar	0.08	0.17
5	3	Sastry	0.08	0.25
6	4	Venugopal	0.08	0.33
7	5	Gupta	0.08	0.42
8	6	Nagpal	0.08	0.50
9	7	Dhiman	0.08	0.58
10	8	Kaushik	0.08	0.67
11	9	Bhardwaj	0.08	0.75
12	10	Garg	0.08	0.83
13	11	Giri	0.08	0.92
14	12	Iyengar	0.08	1.00

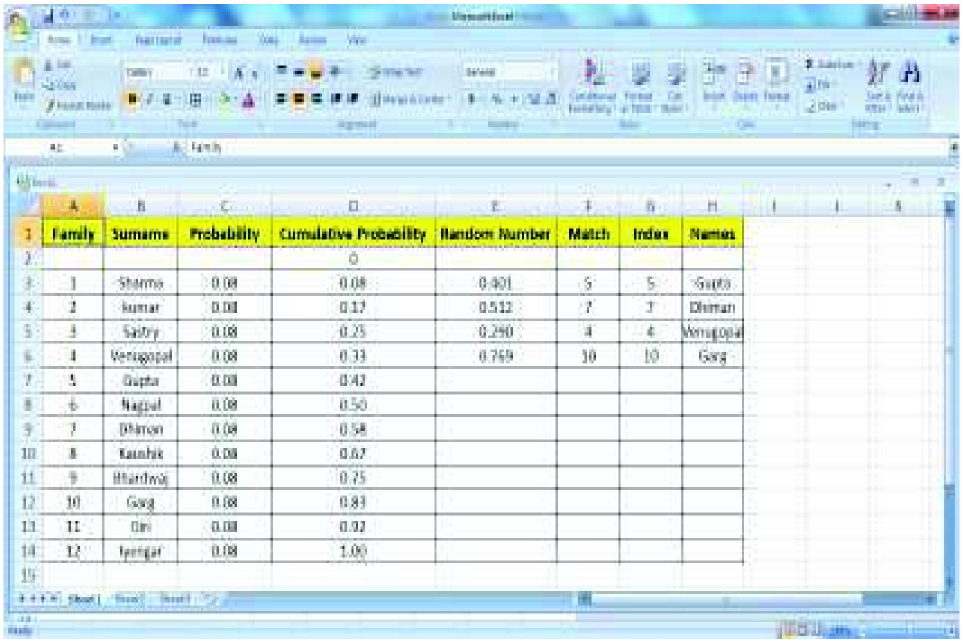
Formula :

The function Rand() generates the random numbers from 0 to 1, from the above results it is observed that $B$3:$M$3 is the range of generated random numbers.

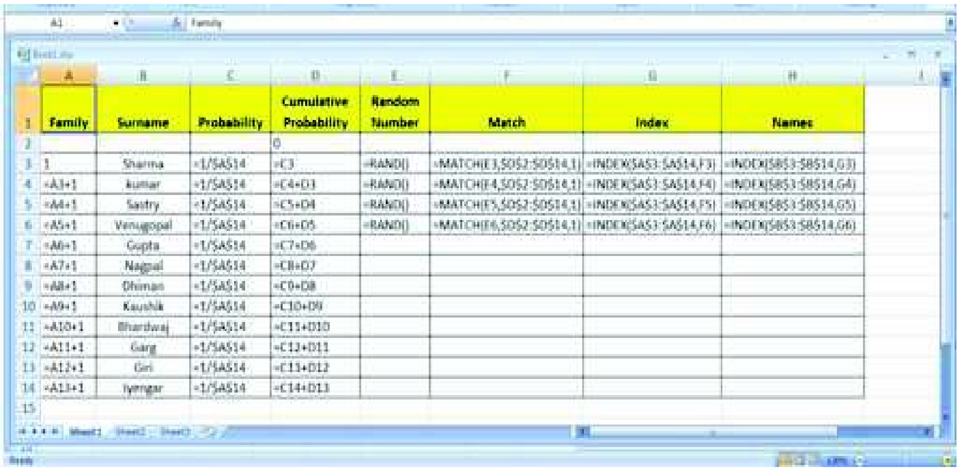
match(lookup value, lookup array), i.e. =MATCH(E3,$D$2:$D$14) :returns the relative position of an item in an array that matches a specified value in a specified order.

To be specific to the application, by using match we can determine the Index value (which is the Family number in our case). Everytime you refresh the Random# column the entire status of Match column will change randomly. Now our sample is of size 4 families, thus either we can directly refer to first 4 entries under match# as the random families to be opted OR we can use the formula Index(). Lets use index to clearly identify the Number and Surname of the Randomly sampled families.

index(reference, row num[column num]), i.e. =INDEX($A$3:$A$14,F3) returns the Randomly sampled family number and =INDEX($B$3:$B$14,G3) returns the Randomly sampled family surname. One of the Randomly Sampled Data is shown in screen shot below.



Family	Surname	Probability	Cumulative Probability	Random Number	Match	Index	Names
			0				
1	Sharma	0.08	0.08	0.401	5	5	Gupta
2	kumar	0.08	0.17	0.512	7	7	Dhiman
3	Sastry	0.08	0.25	0.290	4	4	Venugopal
4	Venugopal	0.08	0.33	0.769	10	10	Garg
5	Gupta	0.08	0.42				
6	Nagpal	0.08	0.50				
7	Dhiman	0.08	0.58				
8	Kaushik	0.08	0.67				
9	Bhardwaj	0.08	0.75				
10	Garg	0.08	0.83				
11	Giri	0.08	0.92				
12	Iyengar	0.08	1.00				



Family	Surname	Probability	Cumulative Probability	Random Number	Match	Index	Names
			0				
1	Sharma	=1/$A$14	=C3	=RAND()	=MATCH(E3,$D$2:$D$14,1)	=INDEX($A$3:$A$14,F3)	=INDEX($B$3:$B$14,G3)
=A3+1	kumar	=1/$A$14	=C4+D3	=RAND()	=MATCH(E4,$D$2:$D$14,1)	=INDEX($A$3:$A$14,F4)	=INDEX($B$3:$B$14,G4)
=A4+1	Sastry	=1/$A$14	=C5+D4	=RAND()	=MATCH(E5,$D$2:$D$14,1)	=INDEX($A$3:$A$14,F5)	=INDEX($B$3:$B$14,G5)
=A5+1	Venugopal	=1/$A$14	=C6+D5	=RAND()	=MATCH(E6,$D$2:$D$14,1)	=INDEX($A$3:$A$14,F6)	=INDEX($B$3:$B$14,G6)
=A6+1	Gupta	=1/$A$14	=C7+D6				
=A7+1	Nagpal	=1/$A$14	=C8+D7				
=A8+1	Dhiman	=1/$A$14	=C9+D8				
=A9+1	Kaushik	=1/$A$14	=C10+D9				
=A10+1	Bhardwaj	=1/$A$14	=C11+D10				
=A11+1	Garg	=1/$A$14	=C12+D11				
=A12+1	Giri	=1/$A$14	=C13+D12				
=A13+1	Iyengar	=1/$A$14	=C14+D13				

5.4 SUMMARY

In this session you learned about the basic formulation of questionnaire for an assigned task, and learned about the various type of scaling viz. nominal, ordinal etc, for data collection. Further, you learned about the tabulation of the data gathered through questionnaire, and its digitization in excel. We also demonstrated the use of data analysis tool pack for data interpretation. The session, also enlightened the concept of random sampling through spread sheets, where you learned to use various formulas viz. rand(), match(), index() etc. After going through this session you might be in the position to practically apply the statistical concepts at some elementary level.

5.5 EXERCISES

EXERCISE – 1 Analyze the Case given below and answer the subsequent questions
“ *A park, maintained by an NGO, the management of the Park wants to make a study which will help them to identify the satisfaction level of the visitor families, the park*

has facility for every family member as if the youngsters may use playground for sports and others may use it for picnic or reading or meditation or walk or jogging, but they should adhere to the norms laid by the park management. As usual the existence of playground is always an issue(generation gap), so the management want to decide that the availability of playground leads to family satisfaction or it should be removed or transformed for some other activity. Further, the management of conveyance and over crowdedness in the park are also to be addressed. The management identified that different activities are performed for different duration of time, they wish to identify the activity performed the most and its respective duration, and the mode or vehicles used by families to visit the park, based on the outcome they may decide that which activities should be commercialize and do they need to focus on parking"

1. Identify the objectives of the study.
2. Draft a Questionnaire to study the identified objectives.
3. Tabulate the collected data and digitize it in excel spreadsheet.
4. Use the Descriptive statistics option of Data analysis toolpak, and generate the results for collected data.
5. Analyze the generated results for the identified objectives, and comment accordingly.
6. Use the results from descriptive statistics to calculate the sample size, comment whether the sample size is sufficient/ insufficient for the conducted study.

EXERCISE – 2 Develop an Excel Application that can perform Random sampling for the data tabulated below, and choose 3 villages randomly for study.

VILLAGE	**A**	**B**	**C**	**D**	**E**	**F**	**G**
IT SERVICES	115	267	98	155	321	144	231

CORRELATION and REGRESSION

Solution :
Correlation CYP-1(P6)
Q1

a) Scatter plot shows that data is quite correlated with positive magnitude and value or r is approximately close to 0.9 or so
b) r= 0.96 thus the variables are synchronized , thus variation in one variable directly affects the other.
c) Yes, we can use the collected data for the forecasting purpose because r~0.96 i.e. good correlation

Regression CYP-2
Q1

a) Both factors are identified to be highly interdependent
b) Unit increase in RAM size will lead to 0.69 times increase in systems performance
c) Unit improvement in systems performance requires 0.54 times of alteration in RAM
d) 25 times of RAM change will lead to improve system performance by 0.69 X 25 times

Multiple Regression CYP-3
Q1
a)~70.6% of variation of Y is explained by X1 and X2
b)-5.51 – 39.33 X1 + 6.295 X2

ANOVA
CYP-1

1. Alpha =0.05 i.e.5% , so level of confidence is 95%
2. Critical value of Factor F (Fcrit) = 3.238872
3. Because F= 0.290072 and Fcrit = 3.238872 i.e. F <Fcrit , so accept the NULL Hypothesis.
4. P=0.831921, which is more than the desired significance level, so accept the null hypothesis

CYP-2

1. Alpha =0.05 i.e. 5%, thus level of confidence = 95%
2. Critical value of F(Fcrit) = 3.49 and 3.25 respectively
3. F=8.015 ; Fcrit = 3.49 ; F >Fcrit , so Non Acceptance of Null Hypothesis
4. F=6.291 ; Fcrit = 3.259 ; F >Fcrit , so Non Acceptance of Null Hypothesis
5. P value > significant value, thus , Non Acceptance of Null Hypothesis