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 conclusions 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 BCS-040/ Block1 and 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; and
- 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 BCS-040: 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: It 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.

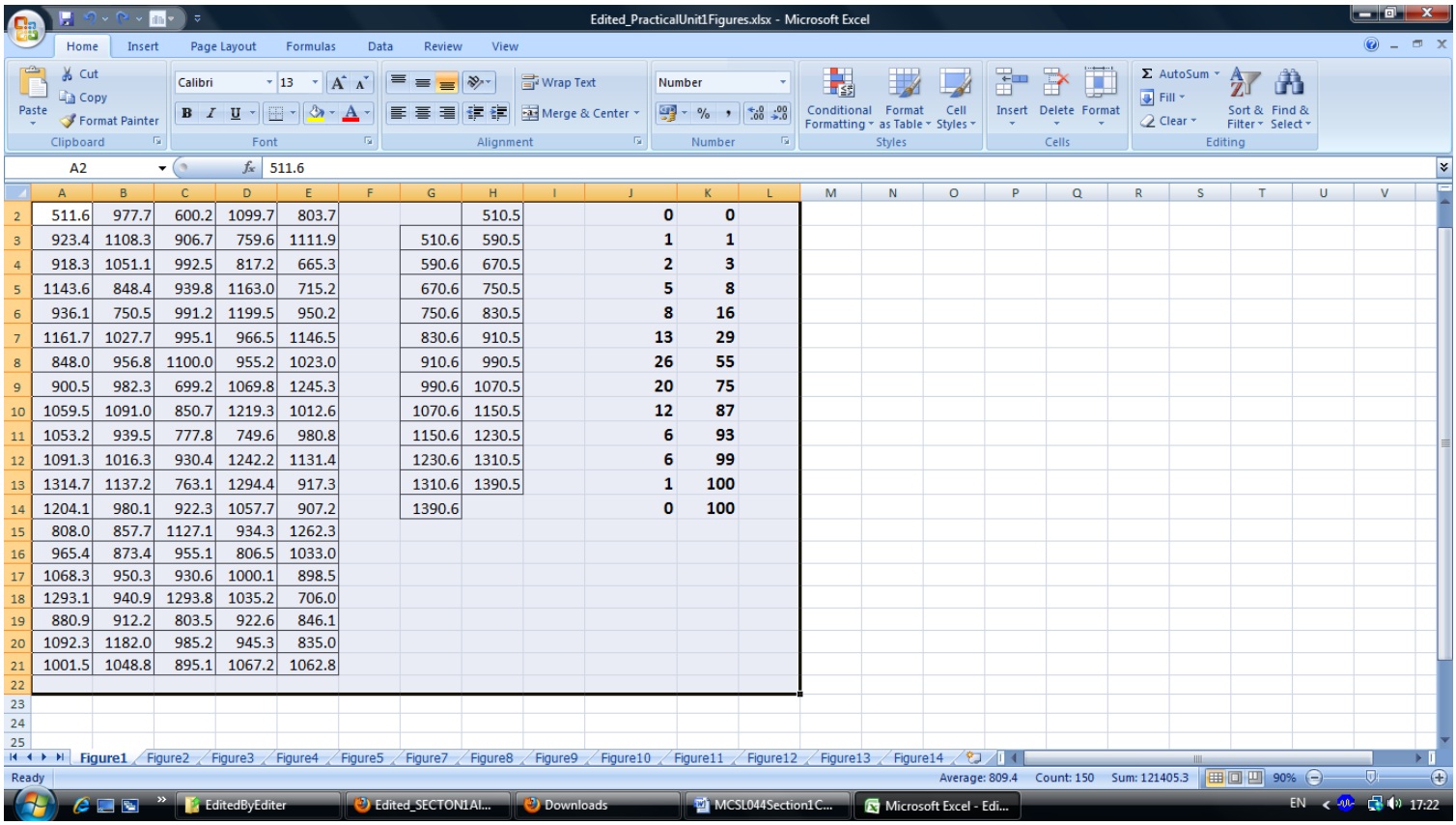
Solution: Figure 1 shows the spreadsheet (for the example, MS-Excel has been used, but you are free to use any spreadsheet package). Enter the given data in the cells *A2….E21*, that needs to be used to construct the frequency distribution.
(Read page 15, Unit 1 of Block 1 of BCS-040). In the next step, 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.



	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 $k = [1 + log_2 N]$, 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*". (Please refer to Section 2 for detailed steps of loading 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:

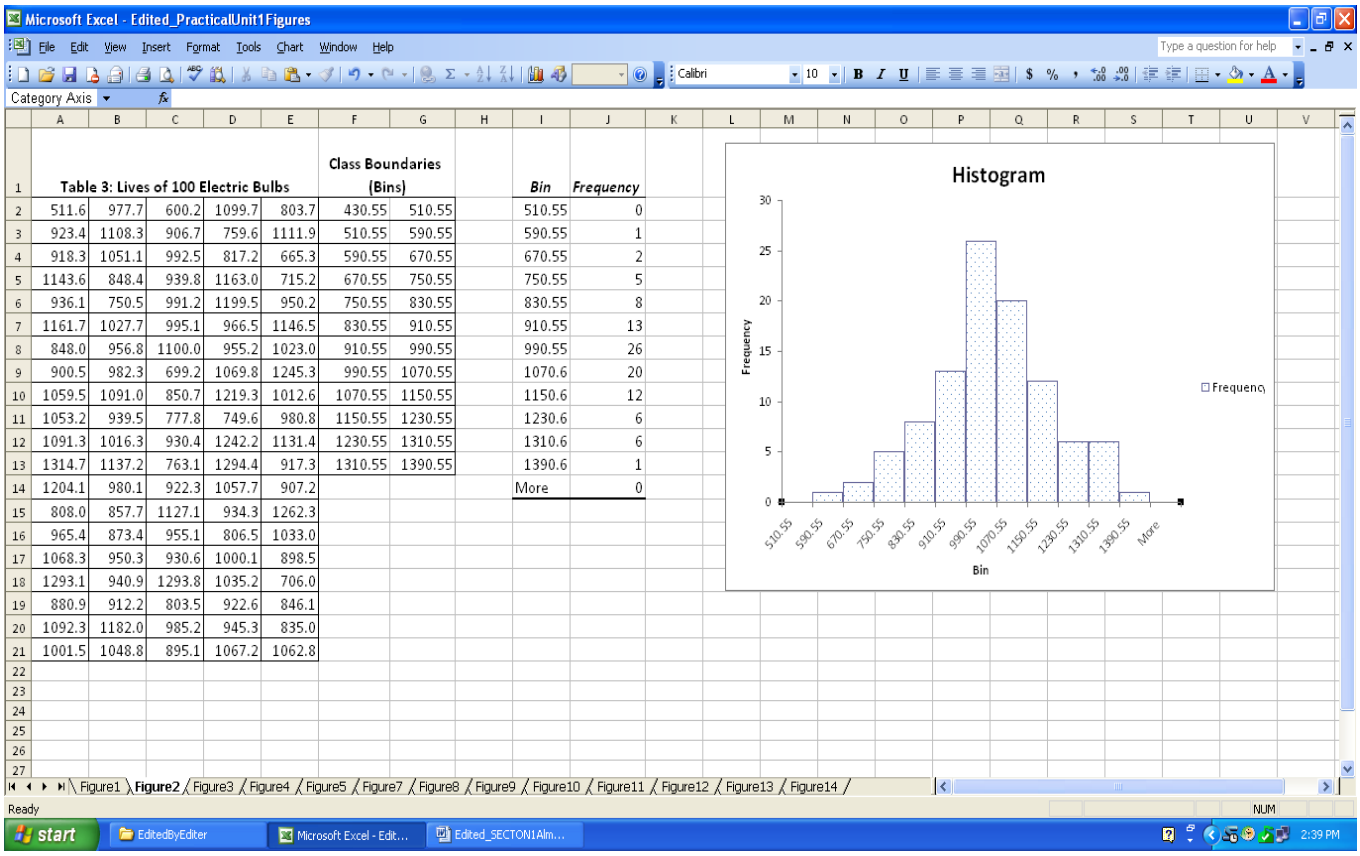
1. Select *Data* → *Data Analysis* and then select *Histogram* in the resulting dialog box and press Ok button.
2. In the resulting dialog box, set the *Input Range* to A2..E21, *Bin Range* to G2..G13 and *Output Range* to I1..J14. Please do not forget to check the *Chart Output* check box.
3. Click 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 to give for the same frequency distribution (refer to Block 1, unit 1, section 1.2.3, page 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.



	A	B	C	D	E	F	G	H	I	J
1	Table 3: Lives of 100 Electric Bulbs					Class Boundaries (Bins)			*Bin*	*Frequency*
2	511.6	977.7	600.2	1099.7	803.7	430.55	510.55		510.55	0
3	923.4	1108.3	906.7	759.6	1111.9	510.55	590.55		590.55	1
4	918.3	1051.1	992.5	817.2	665.3	590.55	670.55		670.55	2
5	1143.6	848.4	939.8	1163.0	715.2	670.55	750.55		750.55	5
6	936.1	750.5	991.2	1199.5	950.2	750.55	830.55		830.55	8
7	1161.7	1027.7	995.1	966.5	1146.5	830.55	910.55		910.55	13
8	848.0	956.8	1100.0	955.2	1023.0	910.55	990.55		990.55	26
9	900.5	982.3	699.2	1069.8	1245.3	990.55	1070.55		1070.6	20
10	1059.5	1091.0	850.7	1219.3	1012.6	1070.55	1150.55		1150.6	12
11	1053.2	939.5	777.8	749.6	980.8	1150.55	1230.55		1230.6	6
12	1091.3	1016.3	930.4	1242.2	1131.4	1230.55	1310.55		1310.6	6
13	1314.7	1137.2	763.1	1294.4	917.3	1310.55	1390.55		1390.6	1
14	1204.1	980.1	922.3	1057.7	907.2				More	0
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					

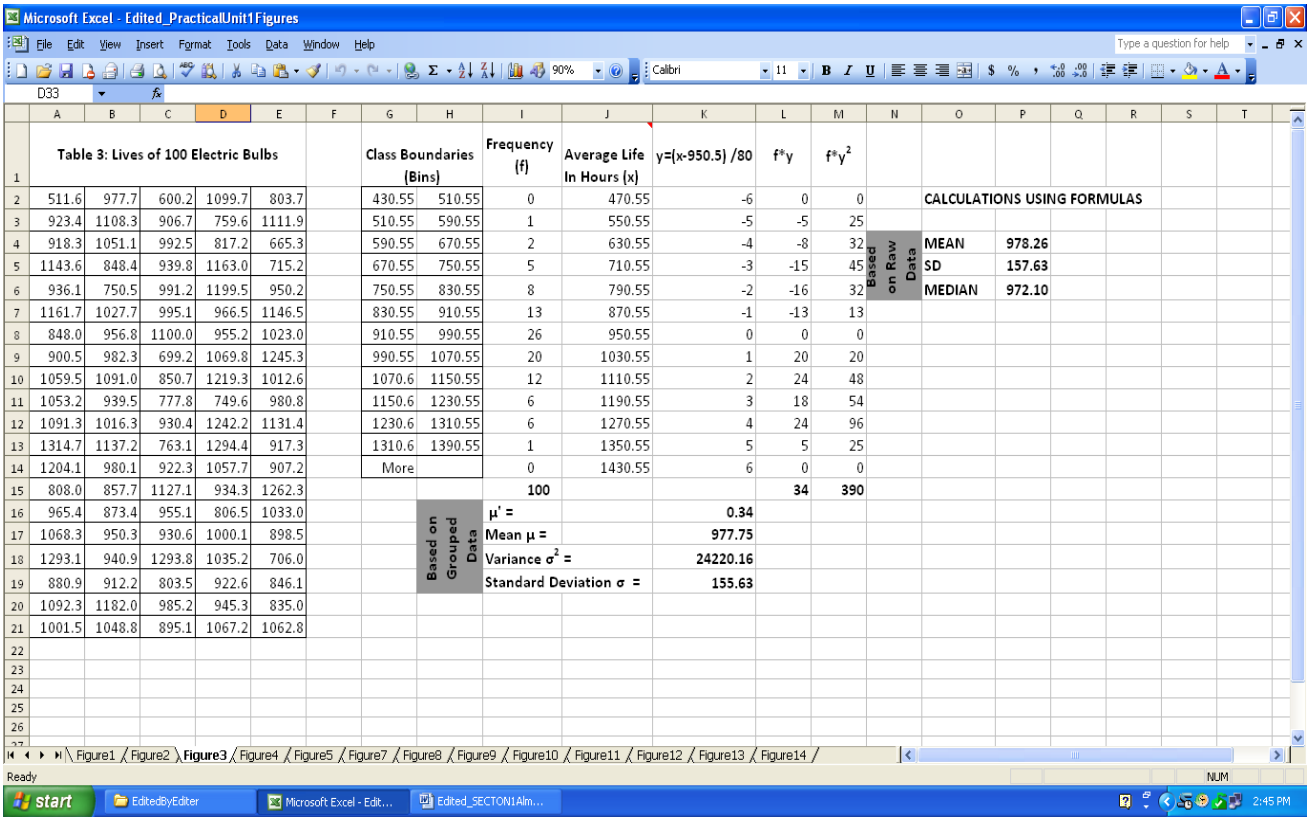
Figure 2: Histogram using the Data → Data Analysis options.

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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Table 3: Lives of 100 Electric Bulbs						Class Boundaries (Bins)		Frequency (f)	Average Life In Hours (x)	y=(x-950.5) /80	f*y	f*y²			
2	511.6	977.7	600.2	1099.7	803.7		430.55	510.55	0	470.55	-6	0	0		CALCULATIONS USING FORMULAS	
3	923.4	1108.3	906.7	759.6	1111.9		510.55	590.55	1	550.55	-5	-5	25			
4	918.3	1051.1	992.5	817.2	665.3		590.55	670.55	2	630.55	-4	-8	32	Based on Raw Data	MEAN	978.26
5	1143.6	848.4	939.8	1163.0	715.2		670.55	750.55	5	710.55	-3	-15	45		SD	157.63
6	936.1	750.5	991.2	1199.5	950.2		750.55	830.55	8	790.55	-2	-16	32		MEDIAN	972.10
7	1161.7	1027.7	995.1	966.5	1146.5		830.55	910.55	13	870.55	-1	-13	13			
8	848.0	956.8	1100.0	955.2	1023.0		910.55	990.55	26	950.55	0	0	0			
9	900.5	982.3	699.2	1069.8	1245.3		990.55	1070.55	20	1030.55	1	20	20			
10	1059.5	1091.0	850.7	1219.3	1012.6		1070.6	1150.55	12	1110.55	2	24	48			
11	1053.2	939.5	777.8	749.6	980.8		1150.6	1230.55	6	1190.55	3	18	54			
12	1091.3	1016.3	930.4	1242.2	1131.4		1230.6	1310.55	6	1270.55	4	24	96			
13	1314.7	1137.2	763.1	1294.4	917.3		1310.6	1390.55	1	1350.55	5	5	25			
14	1204.1	980.1	922.3	1057.7	907.2		More		0	1430.55	6	0	0			
15	808.0	857.7	1127.1	934.3	1262.3				100			34	390			
16	965.4	873.4	955.1	806.5	1033.0			Based on Grouped Data	μ' =		0.34					
17	1068.3	950.3	930.6	1000.1	898.5				Mean μ =		977.75					
18	1293.1	940.9	1293.8	1035.2	706.0				Variance σ^2 =		24220.16					
19	880.9	912.2	803.5	922.6	846.1				Standard Deviation σ =		155.63					
20	1092.3	1182.0	985.2	945.3	835.0											
21	1001.5	1048.8	895.1	1067.2	1062.8											

Figure1 / Figure2 / Figure3 / Figure4 / Figure5 / Figure7 / Figure8 / Figure9 / Figure10 / Figure11 / Figure12 / Figure13 / Figure14

Ready NUM

start | EditedByEditer | Microsoft Excel - Edit... | Edited_SECTON1Alm... | 2:45 PM

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 $y_j^2 f_j$ (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.

☞ Check Your Progress 1

1) From the data as given in this Example 7, Table 3 of Block 1 Unit 1 of BCS-040) construct a frequency distribution using spreadsheet package.

.....

.....

.....

.....

.....

2) Create the histogram from this data using Spreadsheet package.

.....

.....

.....

.....

3) Create the measures of central tendencies and measure of dispersion for this data.

.....

.....

.....

.....

.....

☞ 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 $T = T(X_1, X_2, \ldots\ldots, X_n)$. 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).

Tasks:

- 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 $\bar{X}$ in column I.

- Use the array formula *=FREQUENCY(G3:G8,I3:I7)* in column J to get the frequency distribution of sample mean $\bar{X}$. From this frequency distribution calculate the Relative frequency / Probability distribution $P(\bar{X})$ of sample mean. Note that here $\bar{X}$ denotes a random variable and $\bar{x}$ 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 $\bar{X}$.
- Also note that you can calculate the Grand mean $\bar{\bar{x}}$ (see ***E10..G10***) as well as standard error $SE(\bar{X})$ as shown in the figure. Their direct computations have also been shown using worksheet functions (see ***I14..K15***).
- Computation of $SE(\bar{X})$ using $\sqrt{\frac{N-n}{N-1}}$ the Finite Population Correction Factor (FPC) has been done in ***I17..J19*** using formula $SE(\bar{X}) = \sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}}$ (see BCS-040/Block 2, Unit 4 page 12 or Book 3 page 5.6).

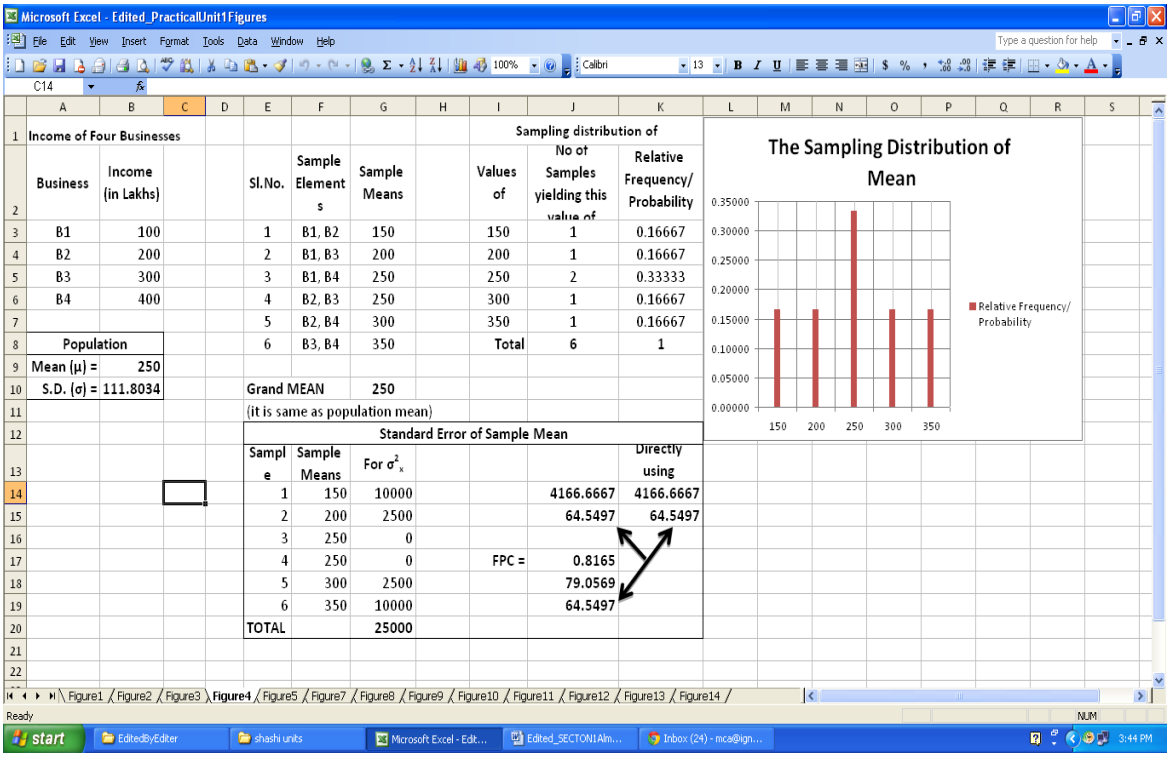


Figure 4: The sampling distribution of Example 1

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 Block 2 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 $\mu^2 = 1$ (Variance).

BCS-040/Block 1/Unit 3 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 μ^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).

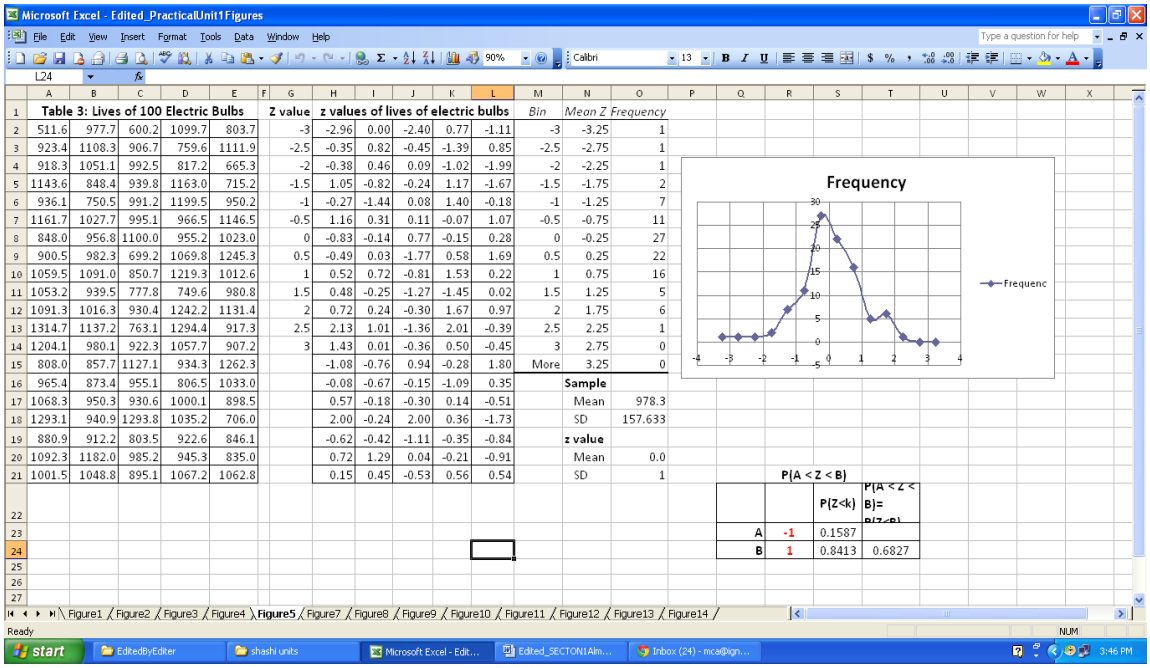


Figure 5: A Sample distribution – almost normal distribution

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).

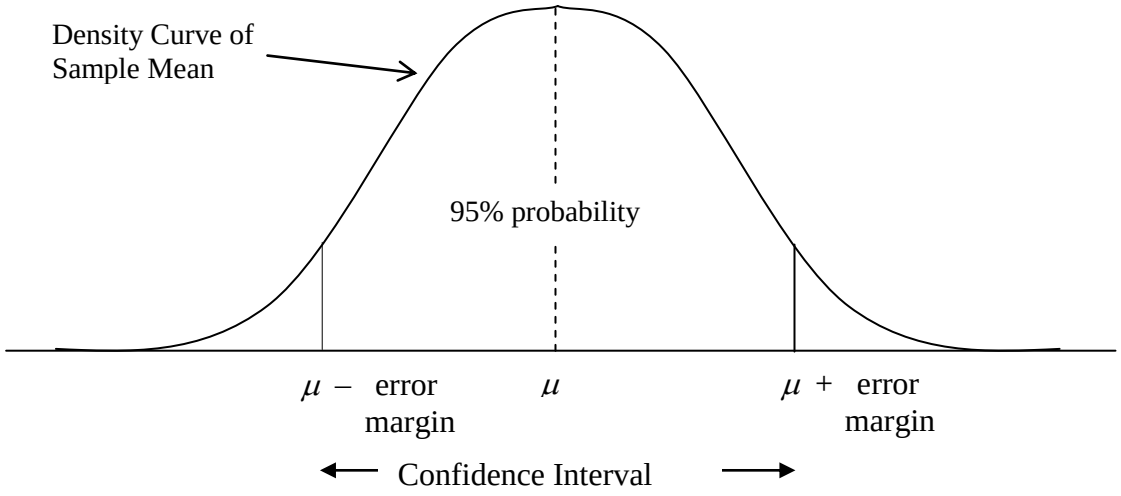
Let μ_T and σ_T be the mean and standard deviation (*standard error*) of the sampling distribution of a static *T*. Then, if the sampling distribution of *T* is approximately normal (for CLT see BCS-040/Block 2, Unit 4 page 15), than the values of probability $P\left[\left|\frac{T-\mu_T}{\sigma_T}\right| < z_c\right]$ for $z_c = 1, 2, 3$ are given in the table below. It follows from the table values that the end numbers T ± σ_S , ± 2σ_S , T ± 3σ_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 ≥ 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 $\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$ or $\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}}$. These are of the form *Estimate* ± *error margin* (please refer to figure 6(a)).

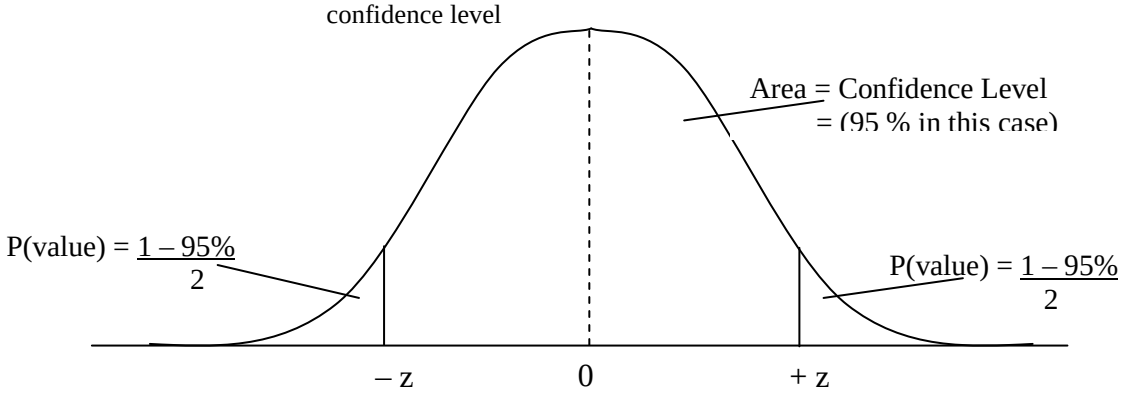
Confidence Level: The probability $P[T - z_C\,\sigma_T < \mu_T < T + z_C\,\sigma_T$ enables us to state the level of confidence for unknown parameter μ_T to belong to the said range. For different values z_C the table below lists the confidence level.

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



(b) 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 $X_1, X_2, \ldots, X_n$ 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 $v = n - 1$.

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

Here, $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the variance of the sample
$\bar{X}$ 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) of BCS-040 / Block 2.

Important Property: *t*-distribution is symmetrical about the value $t = 0$. This is useful in calculation of probabilities under the curve using the above mentioned table. 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 v=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 BCS-040/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 $v = n - 1$, 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.

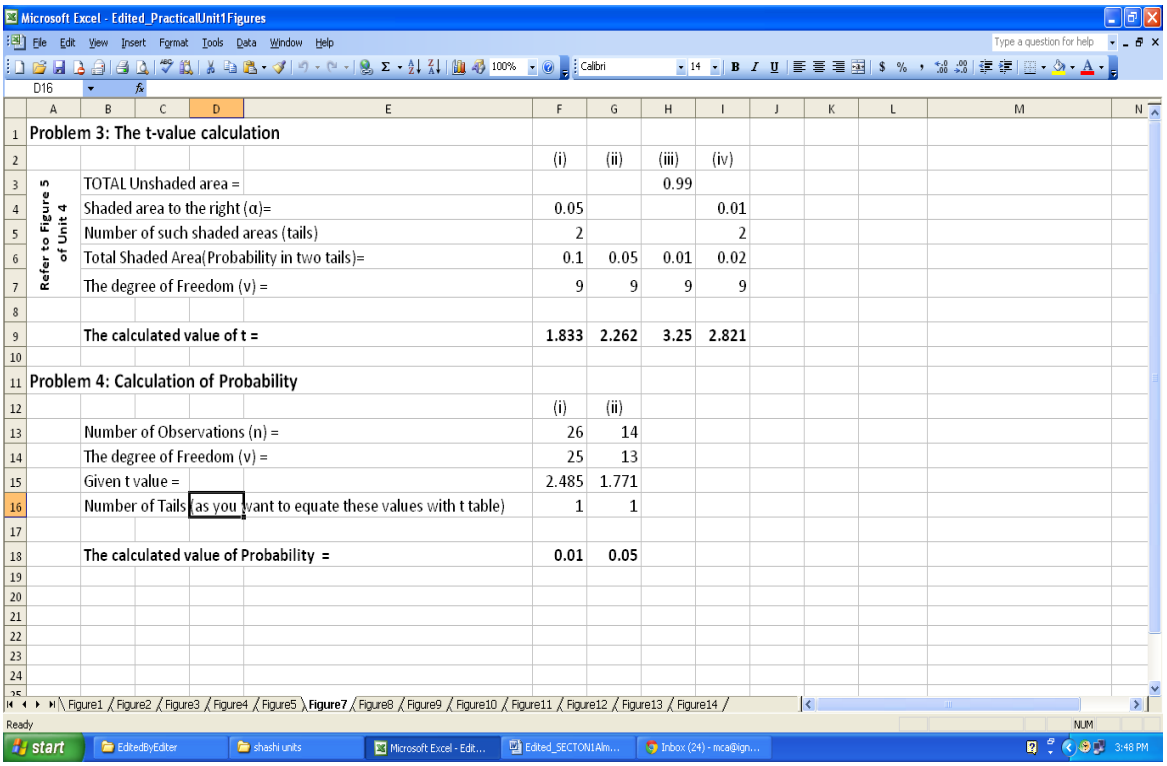


Figure 7: *t*-distribution (Problem 3 and Problem 4)

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 $t_{\bar{a}\,\text{cal}}$ and Reject H_0 , if $t_{\bar{a}\,\text{cal}} > t_{\bar{a}\,\text{tab}}$

$H_1 : \mu < \mu_0$, Calculate $t_{\bar{a}\,\text{cal}}$ and Reject H_0 , if $t_{\bar{a}\,\text{cal}} < t_{\bar{a}\,\text{tab}}$

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

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:

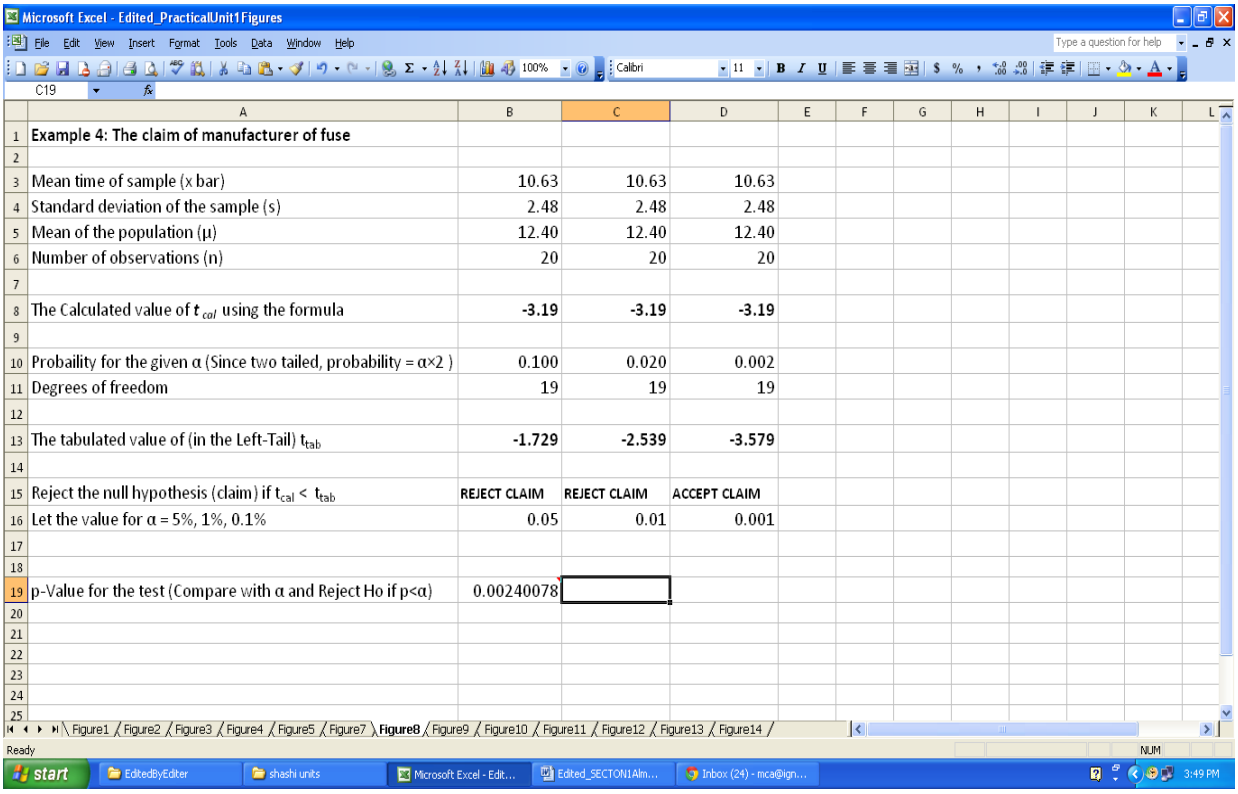
H_0: The average time for the fuse to blow with 20% overload is 12.40 minutes.
H_1: The average time for the fuse to blow with 20% overload is below 12.40 minutes.
or simply stated as

$H_0 : \mu = 12.40$ 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)



Microsoft Excel - Edited_PracticalUnit1Figures

	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)	2.48	2.48	2.48
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	Probaility 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 Ho if p<α)	0.00240078		
20				
21				
22				
23				
24				
25				

Figure1 / Figure2 / Figure3 / Figure4 / Figure5 / Figure7 / Figure8 / Figure9 / Figure10 / Figure11 / Figure12 / Figure13 / Figure14 /

Ready NUM

start EditedByEditer shashi units Microsoft Excel - Edit... Edited_SECTON1Alm... Inbox (24) - mca@ign... 3:49 PM

Figure 8: A sample Hypothesis Testing using t values

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

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

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 $t_{tab} \leq -t_{tab}$, 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 $p = P_r(t \leq P_r(t \leq t_{cal} \mid H_0 : \mu = 12.40) = 0.002400778$, , 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 $X_1, X_2, \ldots, X_n$ be a random sample drawn from the normal distribution $N(\mu, \sigma^2)$. Then distribution of the statistic

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

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

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

Example 5: (*Refer page 23 of the Block 2, Unit-4* stated above) you need to find the value of χ^2_α 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.*v*. 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}

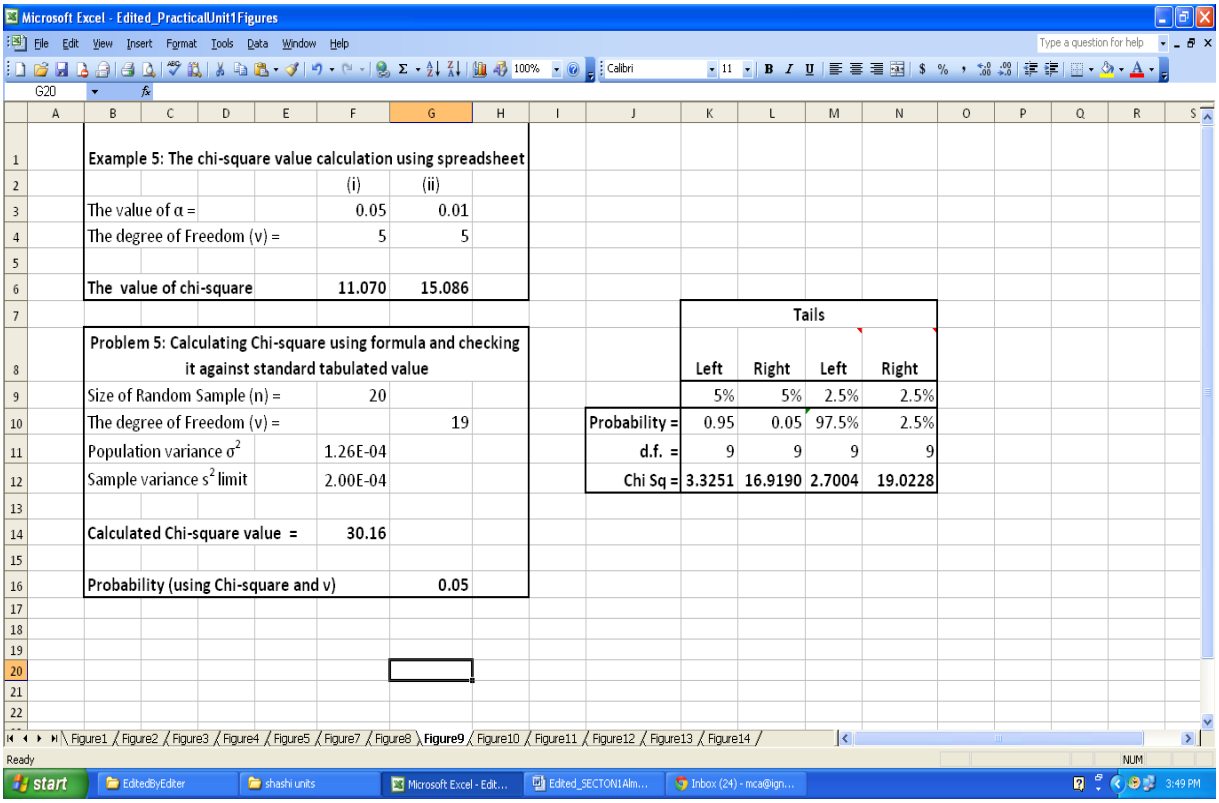
What is the probability that a shipment be rejected even through the variance in the population is 1.26×10^{-4}?

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 ≤ 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 $Pr(\chi^2 > 30.16 \,|\sigma^2 = 1.26\, x\, 10^{-4}$. 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}.



Microsoft Excel - Edited_PracticalUnit1Figures

Example 5: The chi-square value calculation using spreadsheet

	(i)	(ii)
The value of α =	0.05	0.01
The degree of Freedom (v) =	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 (v) =		19
Population variance σ^2	1.26E-04	
Sample variance s^2 limit	2.00E-04	
Calculated Chi-square value =	30.16	
Probability (using Chi-square and v)		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

Figure1 / Figure2 / Figure3 / Figure4 / Figure5 / Figure7 / Figure8 / Figure9 / Figure10 / Figure11 / Figure12 / Figure13 / Figure14

Ready

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 $\chi^2_{cal} > \chi^2_{0.95,9} = 16.9190$
- $H_1: \sigma^2 \neq 16$, 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:

$$\chi^2 = \sum_{i=}^{k} \frac{(0_i - E_i)^2}{E_i}$$

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 $X_{11}, X_{12,\ldots,}X_{1n_1}$ and $X_{21}, X_{22,\ldots,}X_{2n_2}$ 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 Fv_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 $\chi_1^2 = \frac{(n_1-1)S_1^2}{\sigma_1^2}$ and $\chi_2^2 = \frac{(n_2-1)S_2^2}{\sigma_2^2}$

- *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 $P_r(F > F_{\alpha,v_1,v_2})$, 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*(α, 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 $F_{1-\alpha,12,8} = \frac{1}{F_{\alpha,12,8}}$ for $\alpha = 0.05$ (in cell F15) is verified in cell F17.

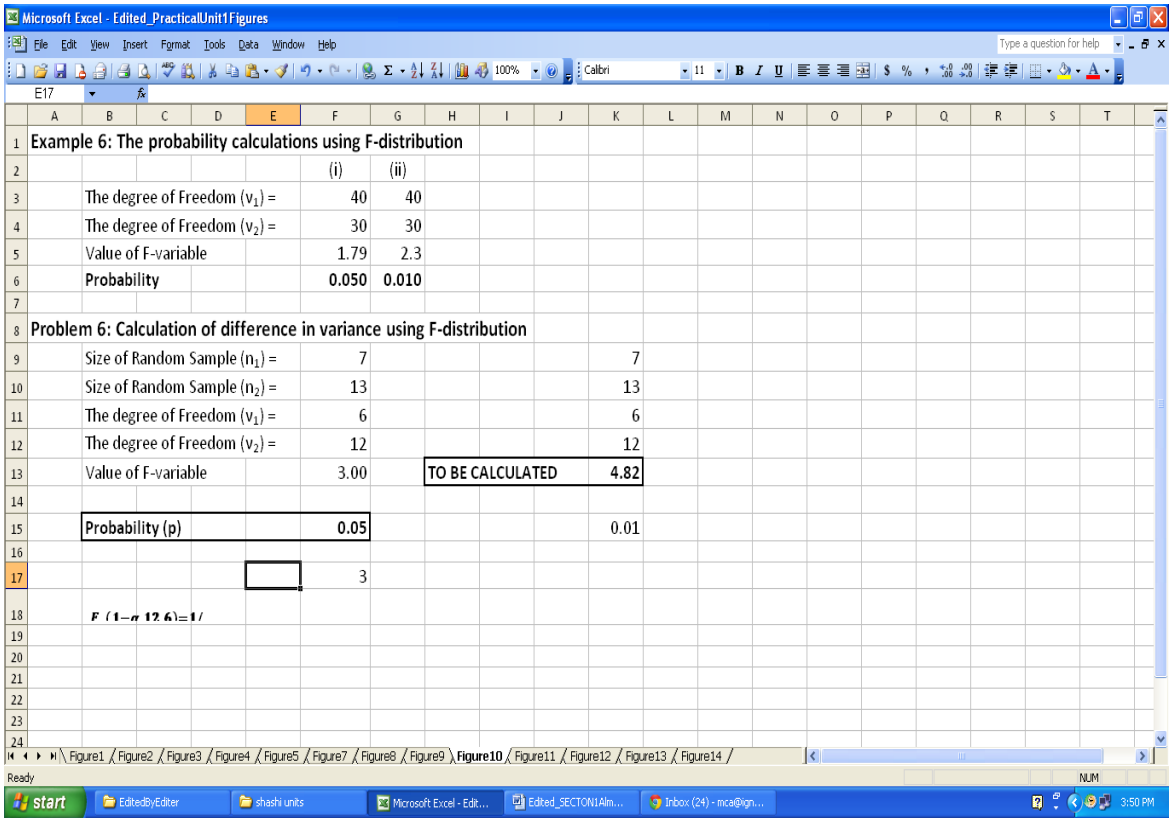


Figure 10: Using *F*-Distribution

☞ Check Your Progress 2

1) Create *t*-distribution for the data given in Figure 7 and Figure 8 of this section using a spreadsheet package.

.....

.....

.....

2) Create chi square-distribution for the data given in Figure 9 of this section using a spreadsheet package.

.....

.....

.....

3) Create *F*-distribution for the data given in Figure 10 of this section using a spreadsheet package

.....

.....

.....

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):

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 $D_i = X_{1i} - X_{2i}$ (i.e., Before – 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 $D_i = X_{2i} - X_{1i}$ (i.e., After –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 $= \Pr(t \leq t_{cal} | H_0 = 0.007 < 0.05$.

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"*).

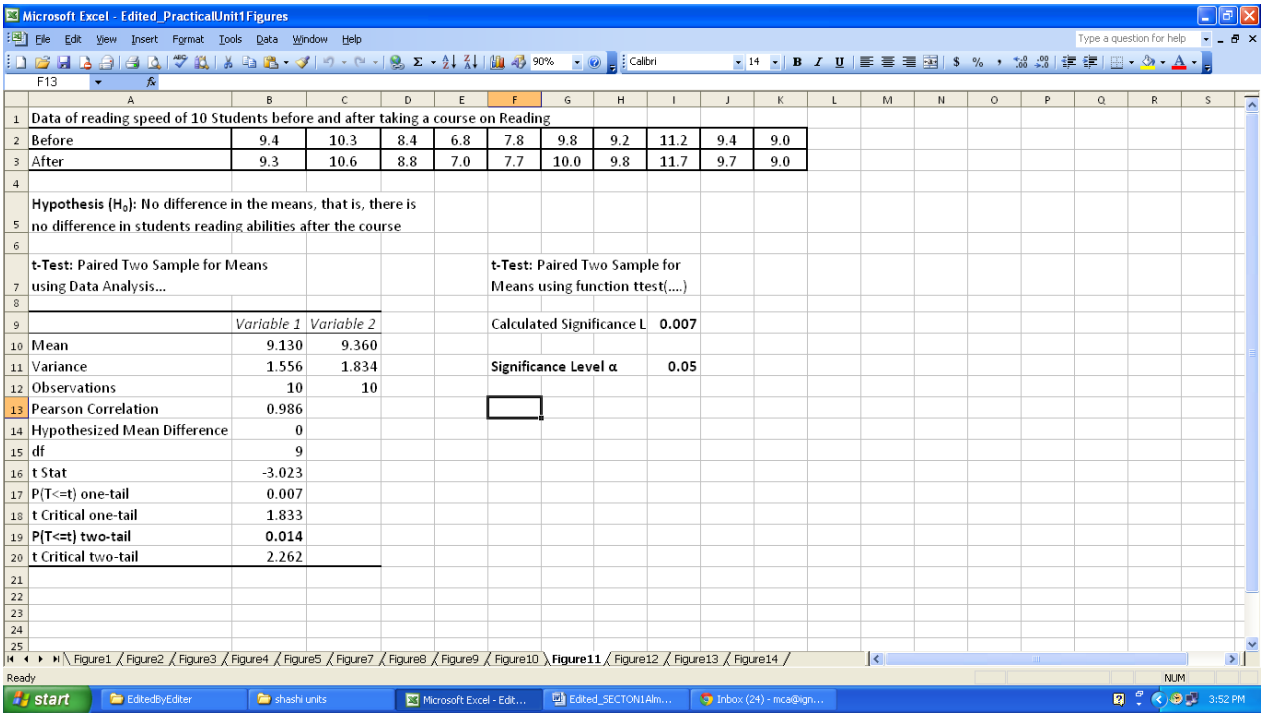


Figure 11: An example of use of *t*-test

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 BCS-040/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) 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 (also see Book 3/chapter 7).

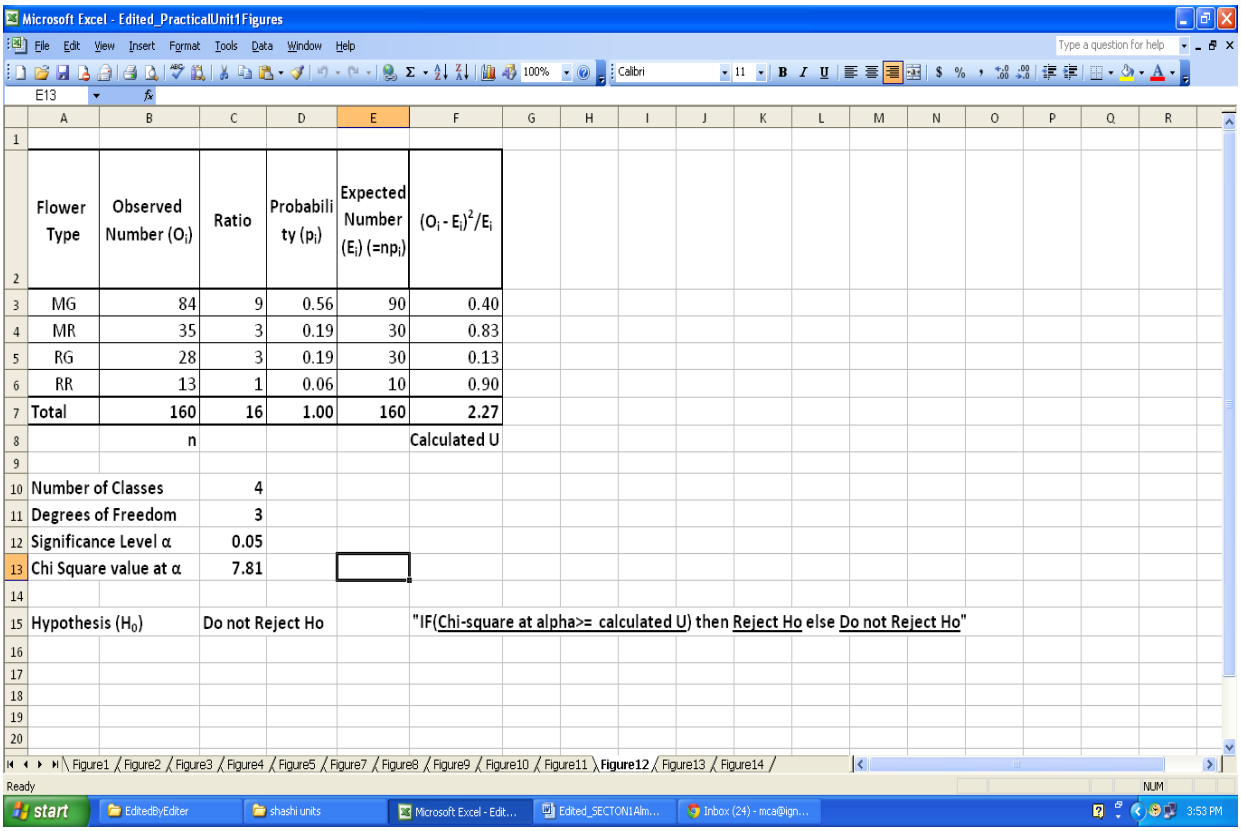


Figure 12: Chi-square test for goodness of fit

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.

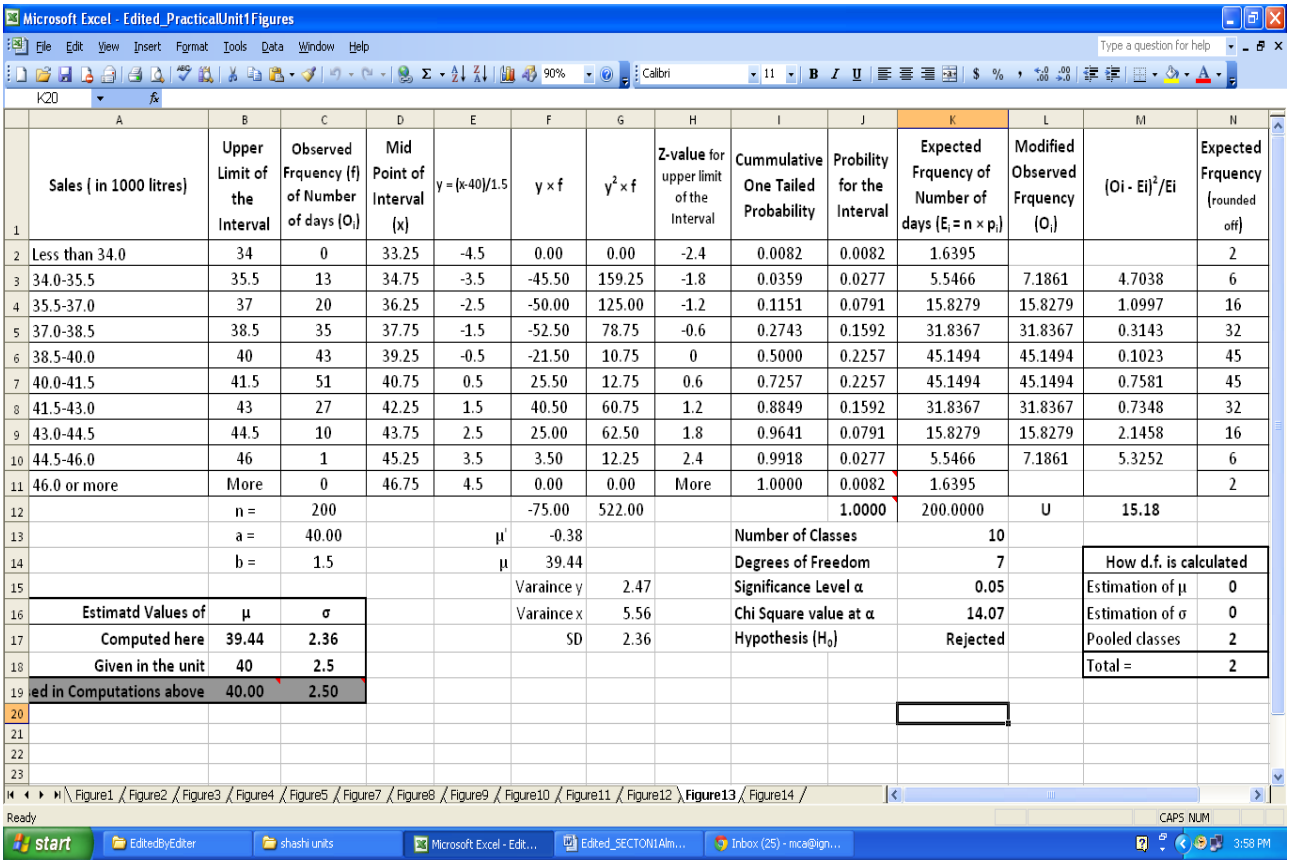
Solution : H_0 : The company's sales are normally distributed.
H_1 : The company's sales are **not** normally distributed.

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?



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Sales (in 1000 litres)	Upper Limit of the Interval	Observed Frquency (f) of Number of days (O_i)	Mid Point of Interval (x)	y = (x-40)/1.5	y × f	$y^2 \times f$	Z-value for upper limit of the Interval	Cummulative One Tailed Probability	Probility for the Interval	Expected Frquency of Number of days ($E_i = n \times p_i$)	Modified Observed Frquency (O_i)	$(Oi - Ei)^2/Ei$	Expected Frquency (rounded off)
2	Less than 34.0	34	0	33.25	-4.5	0.00	0.00	-2.4	0.0082	0.0082	1.6395			2
3	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
4	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
5	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
6	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
7	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
8	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
9	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
10	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
11	46.0 or more	More	0	46.75	4.5	0.00	0.00	More	1.0000	0.0082	1.6395			2
12		n =	200			-75.00	522.00			1.0000	200.0000	U	15.18	
13		a =	40.00		μ'	-0.38			Number of Classes		10			
14		b =	1.5		μ	39.44			Degrees of Freedom		7		How d.f. is calculated	
15						Varaince y	2.47		Significance Level α		0.05		Estimation of μ	0
16	Estimatd Values of	μ	σ			Varaince x	5.56		Chi Square value at α		14.07		Estimation of σ	0
17	Computed here	39.44	2.36			SD	2.36		Hypothesis (H_0)		Rejected		Pooled classes	2
18	Given in the unit	40	2.5										Total =	2
19	ed in Computations above	40.00	2.50											
20														
21														
22														
23														

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)
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.

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

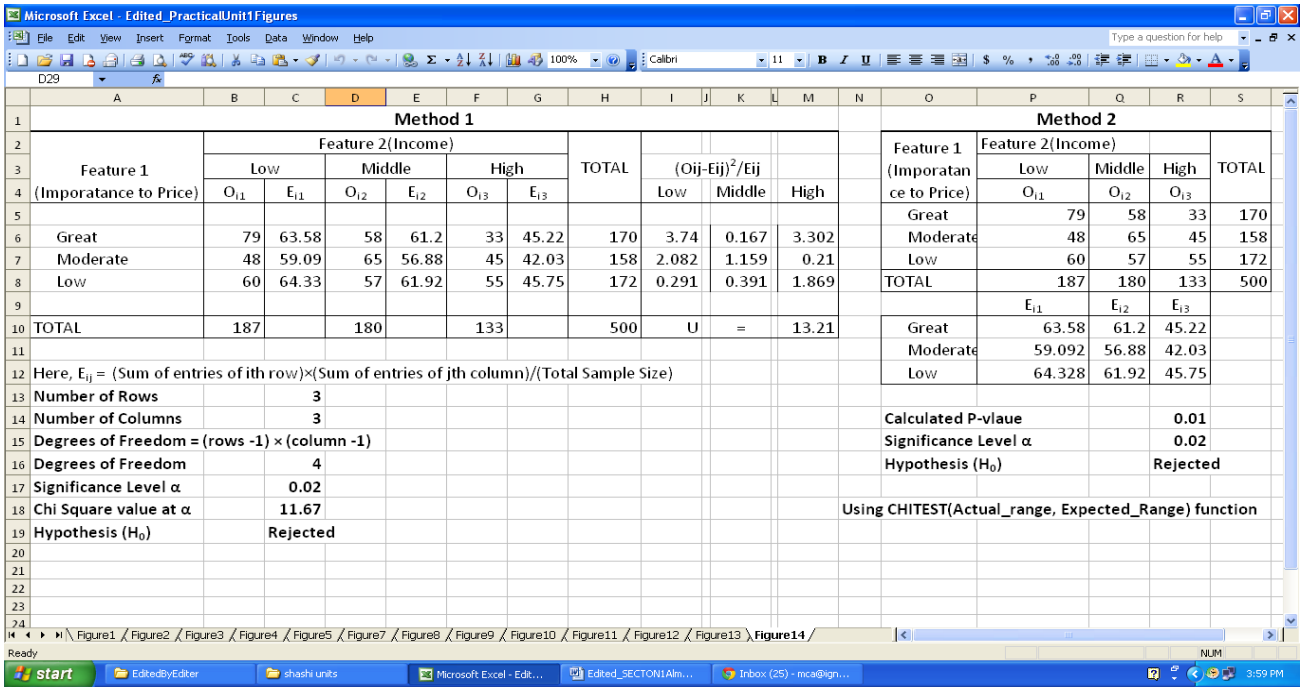


Figure 14: Test of independence

☞ **Check Your Progress 3**

1) Develop all the spreadsheets as shown in Figure 11 to Figure 14. Can you now make effective use of names for ranges instead of cell references?

.....

.....

.....

☞ **Lab Sessions 2, 3, 4**

1) Perform all the examples given in the Unit 4, 5, and 6 using spreadsheet package.

2) 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.

3) Write simple C functions to calculate the *t* and chi-square values from the formula that are given for their calculations.

4) 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 has been 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 was not only to introduce you to various steps that were used to perform the said statistical data analysis, but also to the spreadsheet functions that can be used in performing the same. The unit begin 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 was devoted to application of chi-square testing on testing *goodness of fit* and *independence of attribute*.

1.12 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress 1:

Please use any spreadsheet package and enter all the data as shown in Figure 1 (for question 1), Figure 2 (for question 2) and Figure 3 (for question 3). You may download the file containing all these figures from the website www.ignou.ac.in from BCA page under MCSL044. Find out what formulas have been entered. Also find the errors in the entry of formula.

Check Your Progress 2:

Please use any spreadsheet package and enter all the data as shown in Figure 7, 8 (for question 1), Figure 9 (for question 2) and Figure 10 (for question 3). You may download the file containing all these figures. Find out what formulas have been entered. Also find the errors in the entry of formula.

Check Your Progress 3:

Please use any spreadsheet package and enter all the data as shown in Figure 11 to Figure 14. You may download the file containing all these figures. Find out what formulas have been entered. Also find the errors in the entry of formula. You may define cell ranges and use in formulas.

1.13 FURTHER READINGS

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