PART B

Descriptive Statistics

DESCRIPTIVE STATISTICS

The second part of this lab course comprises Sessions 6 to 8 which will enable you to develop practical skills for computing measures of central tendency, dispersion, moments, skewness and kurtosis. You will also be able to determine the linear relationship between two or more variables in MS Excel 2007. In lab sessions of Part A, you have learnt how to classify qualitative as well as quantitative data and to represent data diagrammatically or graphically using MS Excel 2007.

In Block 1 of MST-002 (Descriptive Statistics), you have learnt that the measures of central tendency provide a single representative value around which data tend to cluster.

In **Session 6**, you will learn how to compute different measures of central tendency and partition values such as arithmetic mean, geometric mean, harmonic mean, mode, median, quartiles, deciles and percentiles using Excel 2007.

The measures of central tendency describe entire data and give a general idea about data. But these values do not tell us about the scatter of data around the central value, which is necessary to get a better description of data. We use measures of dispersion, which study the extent or degree to which data tend to spread or scatter around an average or central value. You have also learnt in Block 1 of MST-002 that we determine the shape of a distribution with the help of skewness and kurtosis, which can be measured using moments. **Session 7** deals with measures of dispersion to study the scatteredness of data while moments, skewness and kurtosis give us an idea about the symmetry and flatness (or peakedness) of a frequency curve, respectively.

In most studies, we come across more than one variable of which two variables may be dependent as discussed in Block 2 of MST-002. In such situations, we can measure the degree of relationship between the variables using Karl Pearson's correlation coefficient. We use the Spearman's rank correlation coefficient to determine the relationship when we have the qualitative variables and we assign ranks to the values of both variables in order of merit. In fact, Spearman's rank correlation coefficient is the Karl Pearson's correlation coefficient calculated for the two sets of ranks. You have learnt in Units 11 and 12 about the multiple and partial correlation coefficients, respectively, when we have more than two interrelated variables.

Whenever we are interested in studying the joint effect of two or more variables on a single variable, we compute multiple correlation coefficient while the partial correlation coefficient helps us study the relationship between two variables after eliminating the effect of the other variables. In **Session 8**, you will learn how to study the relationship between two or more variables using correlation coefficients such as Karl Pearson, Spearman's rank, multiple and partial correlation coefficients using Excel 2007.

All sessions of this part have been designed in such a way that after successfully completing them, you will be able to compute these measures of descriptive statistics using MS Excel 2007. You will also be able to use suitable measures for any given data.

Before studying the lab sessions of Part B, we expect that you have studied Blocks 1 to 3 of the course MST-002 entitled Descriptive Statistics thoroughly. All concepts and formulae used there will be applied here.

Measures of Central Tendency

SESSION 6

Structure

6.1 Introduction
Objectives
6.2 Problem Description
6.3 Arithmetic Mean
Arithmetic Mean for Raw Data
Arithmetic Mean for Discrete Frequency Distribution
Arithmetic Mean for Continuous Frequency Distribution
6.4 Geometric Mean
Geometric Mean for Raw Data
Geometric Mean for Continuous Frequency Distribution
6.5 Harmonic Mean
Harmonic Mean for Raw Data
Harmonic Mean for Discrete Frequency Distribution
Harmonic Mean for Continuous Frequency Distribution
6.6 Mode
Mode for Raw Data
Mode for Discrete Frequency Distribution
Mode for Continuous Frequency Distribution
6.7 Partition Values
Partition Values for Raw Data
Partition Values for Discrete Frequency Distribution
Partition Values for Continuous Frequency Distribution

6.1 INTRODUCTION

In Lab Session 2, you have learnt how to classify raw data and obtain discrete and continuous frequency distributions in order to extract maximum information from the data. In Unit 1 of MST-002 (Descriptive Statistics), we have explained various measures of central tendency to represent the entire data in a single average value.

In this lab session, you will learn how to compute different measures of central tendency, such as arithmetic mean, geometric mean, harmonic mean, mode and median for raw data as well as discrete and continuous frequency distributions using MS Excel 2007. You will also learn how to compute partition values, viz. quartiles, deciles and percentiles for raw data as well as discrete and continuous frequency distributions. These measures give a value around which the data are concentrated. It also helps us in the assessment of a distribution as well as in comparing it with other distributions.

Prerequisite

- Lab Sessions 1 and 2 of MSTL-001 (Basic Statistics Lab).
- Unit 1 of MST-002 (Descriptive Statistics).

In the next lab session, we shall explain various measures of dispersion, which give us an idea about the scatter or spread of data.

Objectives

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

- prepare the spreadsheet in MS Excel 2007;
- determine the arithmetic mean, geometric mean, harmonic mean and mode; and
- compute the median, quartiles, deciles and percentiles.

6.2 PROBLEM DESCRIPTION

In this lab session, we state five problems to illustrate how to compute various measures of central tendency.

For this lab session, we consider

1. The data of Problem 2 given in Lab Session 2 to compute the arithmetic mean, median, quartiles, 7th decile, 15th percentile for a discrete frequency distribution.
2. The data of Problem 3 given in Lab Session 2 to compute the arithmetic mean, median, quartiles, 7th decile, 15th percentile for raw data and continuous frequency distribution.
3. A researcher is studying the growth of a particular strain of bacteria. The daily percentage growth of the bacteria is recorded for 60 days. The data are given in Table 1.

Table 1: Growth of bacteria for 60 days

Day	Growth of Bacteria (%)	Day	Growth of Bacteria (%)	Day	Growth of Bacteria (%)
1	1.7	21	20.1	41	55.3
2	1.9	22	27.4	42	55.8
3	2.4	23	28.3	43	56.8
4	2.6	24	28.7	44	57.6
5	3.5	25	33.2	45	59.4
6	3.8	26	33.4	46	61.2
7	4.3	27	34.6	47	62.2
8	4.3	28	37.4	48	62.3
9	4.8	29	38.8	49	62.4
10	9.1	30	39.7	50	64.7
11	9.6	31	40.3	51	65.1
12	10.9	32	40.9	52	65.3
13	11.5	33	43.7	53	65.9
14	11.8	34	43.2	54	66.2
15	14.3	35	44.6	55	66.9
16	14.7	36	45.5	56	67.2
17	16.3	37	49.6	57	67.9
18	16.7	38	50.4	58	68.5
19	17.8	39	53.3	59	68.6
20	19.2	40	55.1	60	68.9

For this data,

- compute the average percentage growth of bacteria,
- construct the frequency distribution, and
- compute the average percentage growth of bacteria for the frequency distribution.

4. A person travels from city A to city B by car. The average speed of the car per kilometre distance travelled from A to B is recorded in Table 2.

Table 2: Average speed of the car per km from cities A to B

S. No.	Average Speed per km (km/h)	S. No.	Average Speed per km (km/h)	S. No.	Average Speed per km (km/h)
1	45	18	50	35	60
2	25	19	60	36	55
3	30	20	45	37	50
4	30	21	55	38	30
5	40	22	35	39	45
6	45	23	30	40	55
7	50	24	25	41	40
8	50	25	25	42	50
9	60	26	35	43	30
10	65	27	40	44	35
11	80	28	45	45	35
12	80	29	45	46	40
13	65	30	55	47	30
14	65	31	60	48	35
15	60	32	75	49	40
16	55	33	75	50	40
17	35	34	60		

For this data,

- compute the average speed of the car during the entire journey from city A to city B,
- construct the discrete and continuous frequency distributions, and
- compute the average speed of the car during the entire journey for both frequency distributions.

5. A shirt manufacturing company conducted a survey among 100 men to know their shirt size. The data are recorded in Table 3.

Table 3: Data for shirt size

S. No.	Size of Shirt	S. No.	Size of Shirt	S. No.	Size of Shirt
1	38	35	42	69	42
2	39	36	40	70	36
3	36	37	41	71	42
4	41	38	43	72	37
5	38	39	39	73	36
6	40	40	43	74	45
7	38	41	37	75	44
8	40	42	42	76	39
9	39	43	34	77	41
10	41	44	40	78	40
11	46	45	41	79	37
12	42	46	42	80	37
13	40	47	41	81	41
14	36	48	39	82	39
15	40	49	40	83	42
16	35	50	35	84	37
17	41	51	45	85	38
18	43	52	34	86	40

S. No.	Size of Shirt	S. No.	Size of Shirt	S. No.	Size of Shirt
19	39	53	40	87	39
20	38	54	38	88	37
21	43	55	37	89	40
22	42	56	42	90	35
23	43	57	40	91	42
24	40	58	43	92	39
25	39	59	41	93	40
26	41	60	40	94	39
27	41	61	38	95	40
28	40	62	36	96	38
29	43	63	35	97	40
30	44	64	39	98	39
31	36	65	44	99	41
32	38	66	38	100	38
33	42	67	37		
34	44	68	41		

For this data,

- compute the most preferable average size of the shirt,
- construct the discrete and continuous frequency distributions, and
- compute the most preferable average shirt size for frequency distributions.

6.3 ARITHMETIC MEAN

In Unit 1 of MST-002, you have learnt how to compute the arithmetic mean for ungrouped (raw) and grouped data. Arithmetic mean is the most commonly used measure of central tendency and often simply called the average or mean. It is calculated taking all observations of the data into account. We use arithmetic mean when data do not show wide fluctuation, i.e., do not have extreme values. We briefly mention the main points and formulae as follows:

The arithmetic mean is defined as the sum of all observations divided by the number of observations.

- **For Ungrouped (Raw) Data**

  The arithmetic mean of n observations, $x_1, x_2, ..., x_n$ is given by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \ldots (1)$$

- **For Discrete Frequency Distribution**

  If the observations $x_1, x_2, ..., x_k$ occur with frequencies $f_1, f_2, ..., f_k$, respectively, the arithmetic mean is given by

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{N} \quad \text{where } N = \sum_{i=1}^{k} f_i. \qquad \ldots (2)$$

- **For Continuous Frequency Distribution**

  If the observations are arranged in the form of class intervals, we take x_i in equation (2) as the mid value of the ith class interval. The formula for computation of the arithmetic mean is the same as given in equation (2).

Let us now describe the computation of arithmetic mean in Excel 2007 for raw data, discrete and continuous frequency distributions, one at a time.

6.3.1 Arithmetic Mean for Raw Data

We first consider Problem 2 and explain how to compute arithmetic mean for the raw data using MS Excel 2007. In order to compute the average life of the electric bulb for the raw data, we follow the steps given below :

Step 1: We enter the data given in Table 3 of Lab Session 2, in Excel 2007 spreadsheet as shown in Fig. 6.1.

	A	B	C
1	S.No.	Life of Bulb (in hours)	
2	1	1087	
3	2	1289	
4	3	876	
5	4	725	
6	5	900	
7	6	1080	
8	7	952	
9	8	741	
10	9	1000	
11	10	900	

Mean_Raw Data

Fig. 6.1: Partial screenshot of the spreadsheet for the given data.

Step 2: We now compute the arithmetic mean, i.e., the average life of the bulbs. For the given raw data, we can use *Average* function as shown in Fig. 6.2. For this, we

1. select Cell B152,
2. click on the *Formulas* tab, and
3. click on *More Functions* → *Statistical* → *Average*.

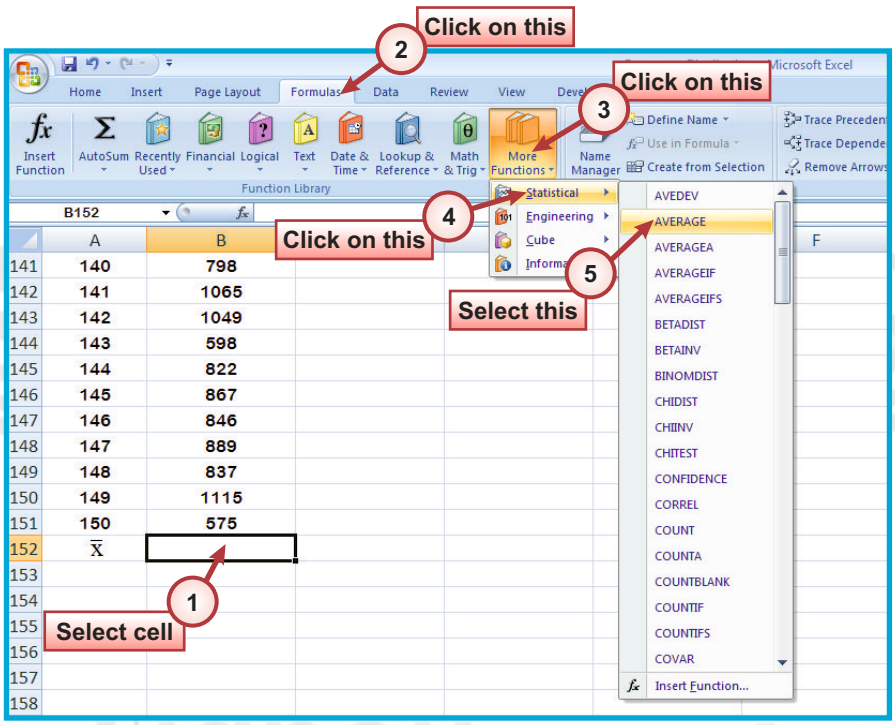


Fig. 6.2

Step 3: A new dialog box opens. We select Cells B2:B151 and click on *OK* as shown in Fig. 6.3.

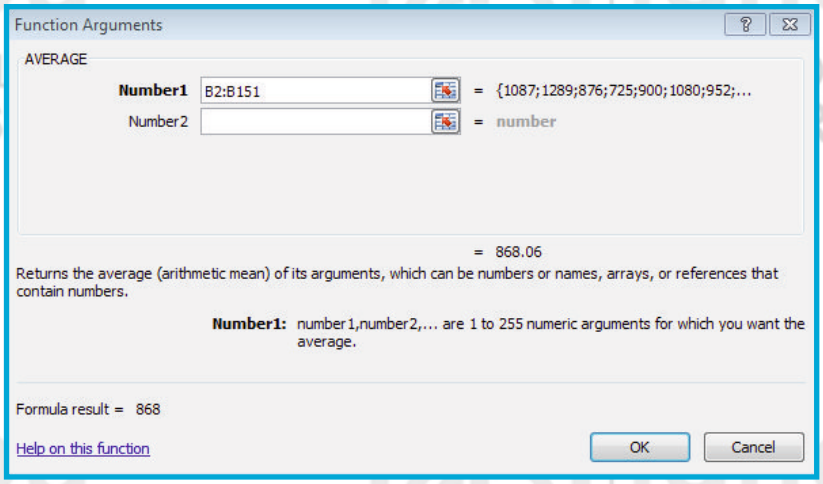


Fig. 6.3

Step 4: We obtain the value of arithmetic mean in Cell B152 as shown in Fig. 6.4.

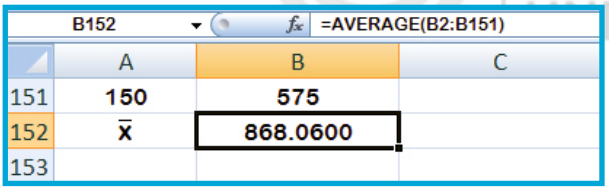


Fig. 6.4

You can also try the short-cut method for computing arithmetic mean as explained in Unit 1 of MST-002.

From Fig. 6.4, the arithmetic mean is given by

$$AM = \bar{x} = 868.0600 \text{ hrs}$$

We can say that the average life of the bulb is 868.06 hrs for the given data.

We now explain how to compute the arithmetic mean for discrete frequency distribution using Excel 2007. For this, we consider Problem 1. Note that Excel does not have any built-in function to compute the arithmetic mean for frequency distributions.

6.3.2 Arithmetic Mean for Discrete Frequency Distribution

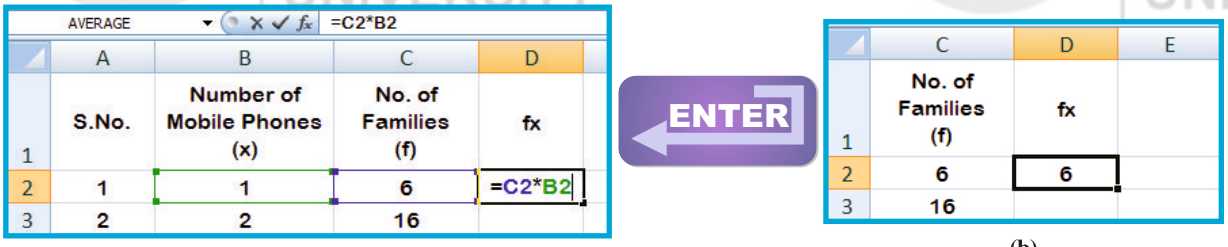
In order to compute the average number of mobile phones per family for the discrete frequency distribution for Problem 1, we follow the steps given below:

Step 1: For the data given in Table 2 of Lab Session 2, we directly consider the discrete frequency distribution as obtained in Sec. 2.4 of Lab Session 2 (Fig. 6.5).

	A	B	C	D
1	S.No.	Number of Mobile Phones (x)	No. of Families (f)	
2	1	1	6	
3	2	2	16	
4	3	3	22	
5	4	4	24	
6	5	5	22	
7	6	6	16	
8	7	7	9	
9	8	8	5	

Fig. 6.5

Step 2: We type "fx" in Cell D1. We now multiply the first value of the number of mobile phones (x), i.e., 1 given in Cell B2 with the corresponding number of families (f), i.e., 6 given in Cell C2 to obtain the first value of fx in Cell D2. We type "=C2*B2" in Cell D2 as shown in Fig. 6.6a and press ***Enter*** to obtain the output (Fig. 6.6b).



(a) (b)

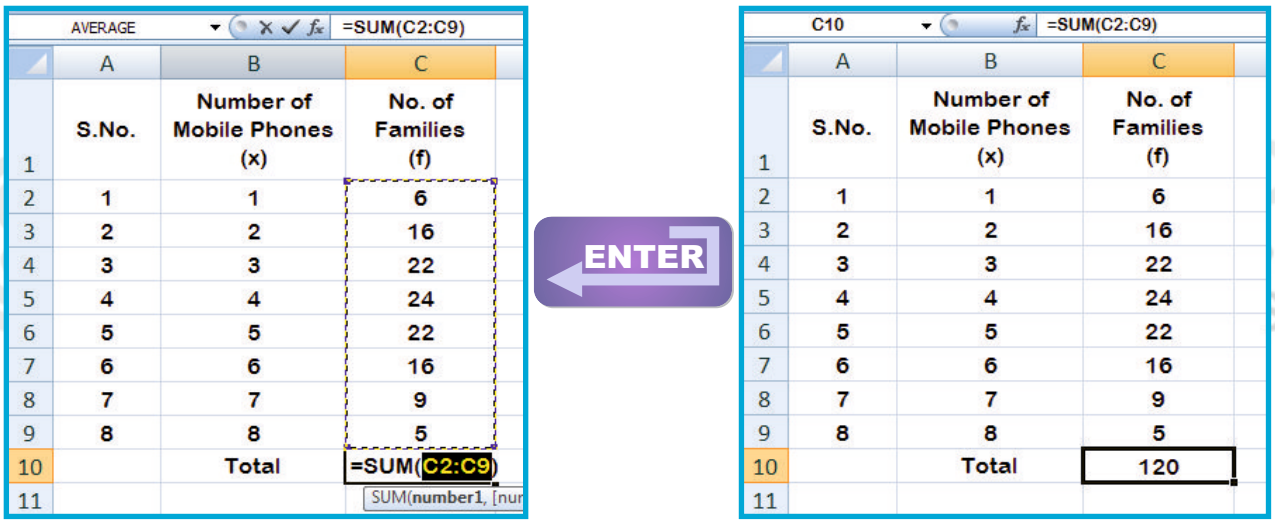
Fig. 6.6

Step 3: We drag Cell D2 down up to Cell D9 to obtain the remaining values of the product of the number of mobile phones (x) and the number of families (f). The result is as shown in Fig. 6.7.

	A	B	C	D	E
1	S.No.	Number of Mobile Phones (x)	No. of Families (f)	fx	
2	1	1	6	6	
3	2	2	16	32	
4	3	3	22	66	
5	4	4	24	96	
6	5	5	22	110	
7	6	6	16	96	
8	7	7	9	63	
9	8	8	5	40	
10					

Fig. 6.7

Step 4: In Cell C10, we compute the total number of families ($\sum f$) using "=Sum(C2:C9)" function as explained in Step 7 of Sec. 2.3, Lab Session 2 and press ***Enter*** (see Fig. 6.8).



(a) (b)

Fig. 6.8

Step 5: In the same way, we obtain $\sum fx$ in Cell D10 as shown in Fig. 6.9.

D10 =SUM(D2:D9)

	B	C	D	E
9	8	5	40	
10	Total	120	509	
11				

Fig. 6.9

Step 6: From equation (2), we compute the arithmetic mean ($\bar{x}$) by typing "=D10/C10" in Cell B11 (Fig. 6.10a) and pressing ***Enter.*** We get the result shown in Fig.6.10b.

MINVERSE =D10/C10

	A	B	C	D
9	8	8	5	40
10		Total	120	509
11	$\bar{x}$	=D10/C10		
12				

(a)

ENTER

B11 =D10/C10

	A	B	C
9	8	8	5
10		Total	120
11	$\bar{x}$	4.2417	
12			

(b)

Fig. 6.10

Step 7: Since the average number of mobile phones cannot be a fraction, we round off its value. The rounded off average number of mobile phones is 4 (Cell C11, Fig. 6.11).

	A	B	C	D
10		Total	120	509
11	$\bar{x}$	4.2417	≈ 4	
12				

Fig. 6.11

From Fig. 6.11, the arithmetic mean is given by

$$AM = \bar{x} = 4$$

We can say that the average number of mobile phones per family is 4.

6.3.3 Arithmetic Mean for Continuous Frequency Distribution

We now consider Problem 2 and explain how to compute the arithmetic mean of a continuous frequency distribution using MS Excel 2007. In order to compute the average life of bulbs for the continuous frequency distribution, we follow the steps given below:

Step 1: For the data given in Table 3 of Lab Session 2, we take the continuous frequency distribution formed in Sec. 2.5 of Lab Session 2 (Fig. 6.12).

	A	B	C
1	Class Interval (x)	No. of Bulbs (f)	
2	500-600	11	
3	600-700	20	
4	700-800	25	
5	800-900	32	
6	900-1000	27	
7	1000-1100	19	
8	1100-1200	11	
9	1200-1300	5	
10			

Fig. 6.12

Step 2: We compute the mid values (Fig. 6.13) as explained in Steps 4 to 6 of Sec. 4.6 of Lab Session 4.

	A	B	C	D	E
1	Class Interval	No. of Bulbs (f)	Lower Limit	Upper Limit	Mid Value (x)
2	500-600	11	500	600	550
3	600-700	20	600	700	650
4	700-800	25	700	800	750
5	800-900	32	800	900	850
6	900-1000	27	900	1000	950
7	1000-1100	19	1000	1100	1050
8	1100-1200	11	1100	1200	1150
9	1200-1300	5	1200	1300	1250

Fig. 6.13

Step 3: To multiply the mid value (x) and frequency (f) of the first class, we type "=B2*E2" in Cell F2 as shown in Fig. 6.14a and press ***Enter*** to obtain the output (Fig. 6.14b).

AVERAGE =B2*E2

	B	C	D	E	F	G
1	No. of Bulbs (f)	Lower Limit	Upper Limit	Mid Value (x)	fx	
2	11	500	600	550	=B2*E2	
3	20	600	700	650		
4	25	700	800	750		

(a)

ENTER

F2 =B2*E2

	B	C	D	E	F	G
1	No. of Bulbs (f)	Lower Limit	Upper Limit	Mid Value (x)	fx	
2	11	500	600	550	6050	
3	20	600	700	650		
4	25	700	800	750		

(b)

Fig. 6.14

Step 4: We drag Cell F2 down up to Cell F9 to obtain the product of the mid value and frequency for the remaining classes as shown in Fig. 6.15.

	B	C	D	E	F	G
1	No. of Bulbs (f)	Lower Limit	Upper Limit	Mid Value (x)	fx	
2	11	500	600	550	6050	
3	20	600	700	650	13000	
4	25	700	800	750	18750	
5	32	800	900	850	27200	
6	27	900	1000	950	25650	
7	19	1000	1100	1050	19950	
8	11	1100	1200	1150	12650	
9	5	1200	1300	1250	6250	
10						

Fig. 6.15

Step 5: We compute the total number of bulbs, i.e., the sum of the frequency ($\sum f$) in Cell B10 as shown in Fig. 6.16.

AVERAGE =SUM(B2:B9)

	A	B	C
1	Class Interval	No. of Bulbs (f)	Lower Limit
2	500-600	11	500
3	600-700	20	600
4	700-800	25	700
5	800-900	32	800
6	900-1000	27	900
7	1000-1100	19	1000
8	1100-1200	11	1100
9	1200-1300	5	1200
10	$N = \sum f$	=SUM(B2:B9)	
11		SUM(number1, [number2], ...)	

(a)

ENTER

B10 =SUM(B2:B9)

	A	B	C
1	Class Interval	No. of Bulbs (f)	Lower Limit
2	500-600	11	500
3	600-700	20	600
4	700-800	25	700
5	800-900	32	800
6	900-1000	27	900
7	1000-1100	19	1000
8	1100-1200	11	1100
9	1200-1300	5	1200
10	$N = \sum f$	150	
11			

(b)

Fig. 6.16

Step 6: In the same way, we calculate $\sum fx$ in Cell F10 as shown in Fig. 6.17.

F10 =SUM(F2:F9)

	E	F	G	H
9	1250	6250		
10	$\sum fx$	129500		
11				

Fig. 6.17

Step 7: To calculate the arithmetic mean, we type "=F10/B10" in Cell F11 (Fig. 6.18a) and press ***Enter*** to obtain the output shown in Fig. 6.18b.

(a)

MINVERSE =F10/B10

	A	B	C	D	E	F	G
9	1200-1300	5	1200	1300	1250	6250	
10	$N = \sum f$	150			$\sum fx$	129500	
11					$\bar{x}$	=F10/B10	
12							

ENTER

(b)

F11 =F10/B10

Name Box

	A	B	C	D	E	F	G
9	1200-1300	5	1200	1300	1250	6250	
10	$N = \sum f$	150			$\sum fx$	129500	
11					$\bar{x}$	863.3333	
12							

Fig. 6.18

From Fig. 6.18b, the arithmetic mean is given by

$$AM = \bar{x} = 863.3333 \text{ hrs}$$

We can say that the average life of the bulb is 863.33 hrs for the given data.

6.4 GEOMETRIC MEAN

You have learnt in Unit 1 of MST-002 that the geometric mean is commonly used in the computation of average rate or percentage growth. We also use geometric mean where we assign larger weights to small items and smaller weights to large items. We briefly mention the main points for calculating geometric mean as follows:

- **For Ungrouped (Raw) Data**

  If $x_1, x_2, ..., x_n$ are n observations of a variable X, then their geometric mean is given by

$$GM = \sqrt[n]{x_1.x_2...x_n} = (x_1.x_2...x_n)^{\frac{1}{n}}$$

or $$GM = \text{Antilog}\left(\frac{1}{n}\sum_{i=1}^{n}\log x_i\right) \quad ... (3)$$

- **For Discrete Frequency Distribution**

  If $x_1, x_2, ..., x_k$ are k values of a variable X with frequencies $f_1, f_2, ..., f_k$, respectively, then

$$GM = \left(x_1^{f_1}.x_2^{f_2}....x_k^{f_k}\right)^{\frac{1}{N}}$$

  where $N = f_1 + f_2 + ... + f_k$

or $$GM = \text{Antilog}\left(\frac{1}{N}\sum_{i=1}^{k} f_i \log x_i\right) \quad ... (4)$$

- **For Continuous Frequency Distribution**

  If the observations are arranged in the form of class intervals, we consider x_i as the mid values of the ith class interval and use equation (4) to compute the geometric mean.

Here we describe how to compute the geometric mean using Excel 2007 for raw data and continuous frequency distribution, one at a time.

6.4.1 Geometric Mean for Raw Data

We consider Problem 3 to explain the computation of geometric mean for raw data. In order to compute the average growth of bacteria for the given data, we apply the geometric mean (not the arithmetic mean) and follow the steps given below:

Step 1: We enter the data given in Table 1 in an Excel sheet as shown in Fig. 6.19 and name it "**GM**".

	A	B	C
1	Day	Bacteria Growth %	
2	1	1.7	
3	2	1.9	
4	3	2.4	
5	4	2.6	
6	5	3.5	
7	6	3.8	
8	7	4.3	
9	8	4.3	
10	9	4.8	
11	10	9.1	

GM

Fig. 6.19

Step 2: Since the given data shows the percentage growth of bacteria, we consider the initial growth as 100. We compute the growth of the first

day by typing "=100+B2" in Cell C2 (Fig. 6.20a) and dragging it down up to Cell C61 to compute the growth of the remaining days as shown in Fig. 6.20b.

(a)

C2 =100+B2

	A	B	C	D
1	Day	Bacteria Growth %	Bacteria Growth (assuming the intial growth as 100)	
2	1	1.7	101.7	
3	2	1.9		
4	3	2.4		
5	4	2.6		
6	5	3.5		

DRAG IT DOWN

(b)

	A	B	C	D
1	Day	Bacteria Growth %	Bacteria Growth (assuming the intial growth as 100)	
2	1	1.7	101.7	
3	2	1.9	101.9	
4	3	2.4	102.4	
5	4	2.6	102.6	
6	5	3.5	103.5	
7	6	3.8	103.8	
8	7	4.3	104.3	
9	8	4.3	104.3	
10	9	4.8	104.8	
11	10	9.1	109.1	

Fig. 6.20

Step 3: To compute the geometric mean for the given raw data, we refer to Fig. 6.21 and

1. select Cell C62,
2. click on the ***Formulas*** tab, and
3. click on ***More Functions*** → ***Statistical*** → ***Geomean***.

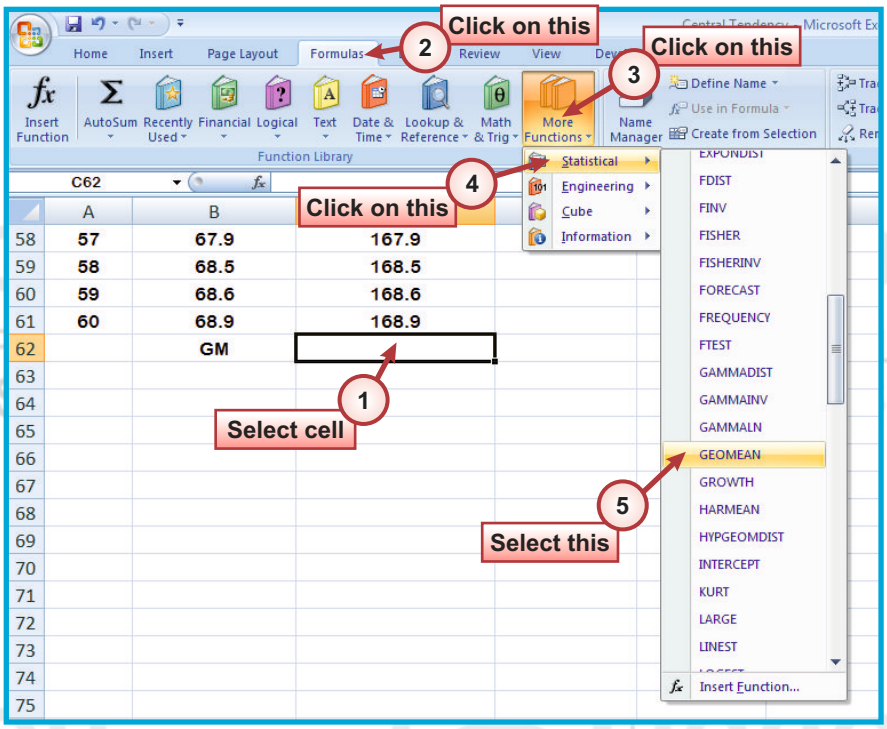


Fig. 6.21

Step 4: The dialog box shown in Fig. 6.22 appears. We select Cell C2:C61 as shown in Fig. 6.22 and click on *OK*.

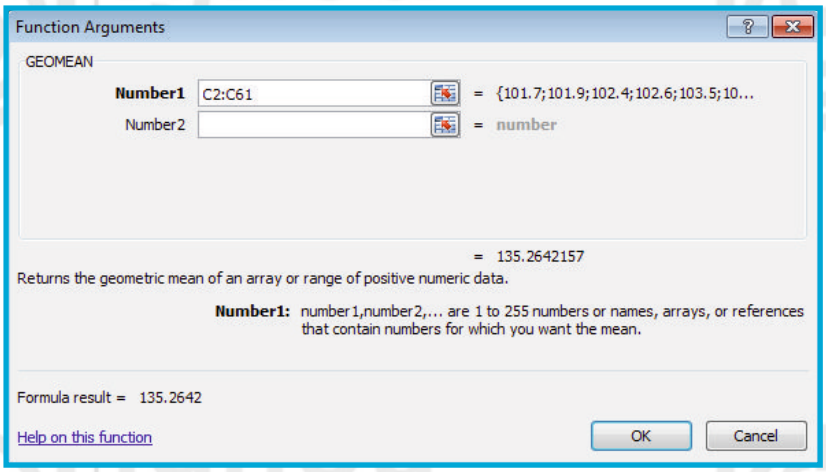


Fig. 6.22

Step 5: The value of the geometric mean is given in Cell C62 (Fig. 6.23).

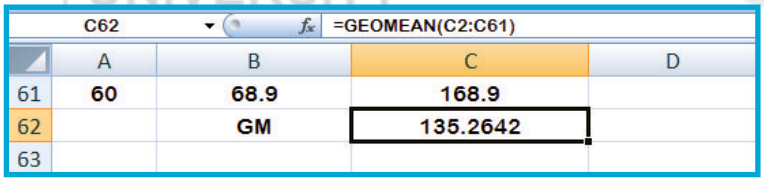


Fig. 6.23

From Fig. 6.23, the geometric mean is 135.2642. Therefore, the average percentage growth from day 1 to 60 = 135.2642 – 100 = 35.2642 or approximately 35%.

To understand the computation of the geometric mean for the grouped frequency distribution, we shall construct the frequency distribution for the data given in Table 1. Since this data has distinct values, we cannot form the discrete frequency distribution. So we shall form the continuous frequency distribution. Here we are continuing with the same Excel sheet. Note that Excel does not have any built-in function to compute the geometric mean for frequency distributions.

6.4.2 Geometric Mean for Continuous Frequency Distribution

In order to compute the average bacteria growth for the continuous frequency distribution, we follow the steps given below:

Step 1: We construct the continuous frequency distribution as explained in Sec. 2.5 of Lab Session 2 (Fig. 6.24).

	D	E	F	G
1		Class Interval	Bin	No. of Days (Frequency) (f)
2	Min. =	100-110	109.99	11
3	101.7	110-120	119.99	9
4	Max. =	120-130	129.99	4
5	168.9	130-140	139.99	6
6	Class Width	140-150	149.99	7
7	9.7292	150-160	159.99	8
8		160-170	169.99	15
9				

Fig. 6.24

Step 2: We compute the mid values as explained in Steps 4 to 6 of Sec. 4.6 of Lab Session 4. The output is shown in Fig. 6.25.

	E	F	G	H	I	J
1	Class Interval	Bin	No. of Days (Frequency) (f)	Lower Limit	Upper Limit	Mid Value (x)
2	100-110	109.99	11	100	110	105
3	110-120	119.99	9	110	120	115
4	120-130	129.99	4	120	130	125
5	130-140	139.99	6	130	140	135
6	140-150	149.99	7	140	150	145
7	150-160	159.99	8	150	160	155
8	160-170	169.99	15	160	170	165
9						

Fig. 6.25

Step 3: To obtain the logarithms with base 10 of the mid values (x), i.e., $\log_{10}(x)$, we refer to Fig. 6.26 and

1. select Cell K2,
2. click on the ***Formulas*** tab, and
3. click on ***Math & Trig*** → ***Log10***

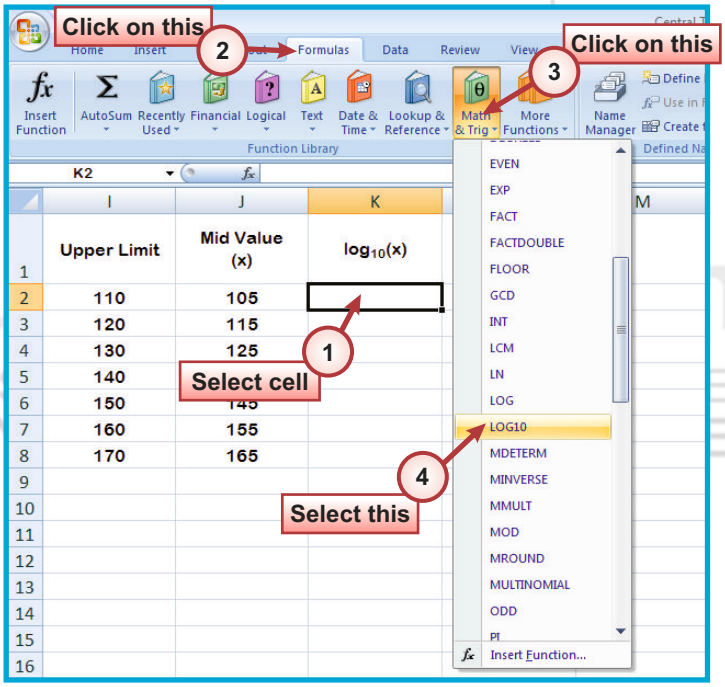


Fig. 6.26

Step 4: A new dialog box appears. We select Cell J2 as shown in Fig. 6.27 and click on ***OK***.

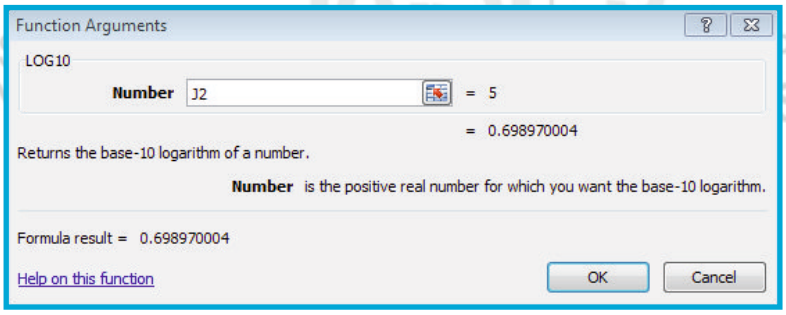


Fig. 6.27

Step 5: The value of $\log_{10}(105)$ is obtained in Cell K2 as shown in Fig. 6.28a. We drag Cell K2 down up to Cell K8 to compute the logarithm of the remaining values as shown in Fig. 6.28b.

K2 =LOG10(J2)

	I	J	K	L
1	Upper Limit	Mid Value (x)	$\log_{10}(x)$	
2	110	105	2.0212	
3	120	115		
4	130	125		

(a)

DRAG IT DOWN

	I	J	K	L
1	Upper Limit	Mid Value (x)	$\log_{10}(x)$	
2	110	105	2.0212	
3	120	115	2.0607	
4	130	125	2.0969	
5	140	135	2.1303	
6	150	145	2.1614	
7	160	155	2.1903	
8	170	165	2.2175	
9				

(b)

Fig. 6.28

Step 6: To multiply $\log_{10}(x)$ by the corresponding frequency (f), we type "=G2*K2" in Cell L2 as shown in Fig. 6.29a and press ***Enter*** to obtain the result (Fig. 6.29b).

LOG10 =G2*K2

	G	H	I	J	K	L	M
1	No. of Days (Frequency) (f)	Lower Limit	Upper Limit	Mid Value (x)	$\log_{10}(x)$	f.$\log_{10}(x)$	
2	11	100	110	105	2.0212	=G2*K2	
3	9	110	120	115	2.0607		

(a)

ENTER

	G	H	I	J	K	L	M
1	No. of Days (Frequency) (f)	Lower Limit	Upper Limit	Mid Value (x)	$\log_{10}(x)$	f.$\log_{10}(x)$	
2	11	100	110	105	2.0212	22.2331	
3	9	110	120	115	2.0607		

(b)

Fig. 6.29

Step 7: We drag down Cell L2 up to Cell L8 to obtain the remaining values as shown in Fig. 6.30.

L2 =G2*K2

	K	L	M
1	$\log_{10}(x)$	f.$\log_{10}(x)$	
2	2.0212	22.2331	
3	2.0607		

(a)

DRAG IT DOWN

	K	L	M
1	$\log_{10}(x)$	f.$\log_{10}(x)$	
2	2.0212	22.2331	
3	2.0607	18.5463	
4	2.0969	8.3876	
5	2.1303	12.7820	
6	2.1614	15.1296	
7	2.1903	17.5227	
8	2.2175	33.2623	
9			

(b)

Fig. 6.30

Step 8: To compute the total frequency, i.e., $\sum f$ in Cell G9, we use ***AutoSum*** function under the ***Home*** tab and press ***Enter***. The output is shown in Fig. 6.31.

G9 =SUM(G2:G8)

	F	G	H
7	159.99	8	150
8	169.99	15	160
9	N = $\sum f$	60	
10			

Fig. 6.31

Step 9: In the same way, we compute the sum of the product, i.e., $\sum f \log_{10}(x)$ in Cell L9 as shown in Fig. 6.32.

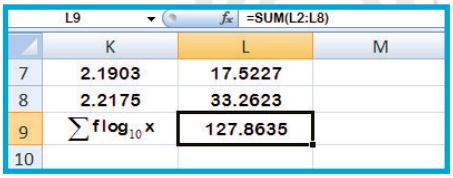
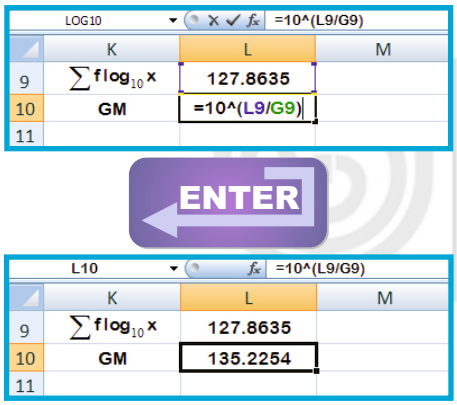


Fig. 6.32

Step 10: We use equation (4) to compute the geometric mean. For this, we type "=10^(L9/G9)" in Cell L10 and press ***Enter*** as shown in Fig. 6.33a. The result shown in Fig. 6.33b.



(a)

(b)

Fig. 6.33

From Fig. 6.33b, the value of geometric mean is given by

$$GM = 135.2254$$

Therefore, the average percentage growth of bacteria from day 1 to 60 = 135.2254 – 100 = 35.2254% or approximately 35%.

6.5 HARMONIC MEAN

You have learnt in Unit 1 of MST-002 that the harmonic mean is commonly used in the computation of average speed. We briefly mention the main points as follows:

- **For Ungrouped (Raw) Data**

  If $x_1, x_2, \ldots, x_n$ are n observations of a variable X, then their harmonic mean is given by

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \qquad \ldots (5)$$

- **For Discrete Frequency Distribution**

  If $x_1, x_2, \ldots, x_k$ are k values of a variable X with their corresponding frequencies $f_1, f_2, \ldots, f_k$, respectively, then

$$HM = \frac{N}{\sum_{i=1}^{k} \frac{f_i}{x_i}} \qquad \ldots (6)$$

  where $N = \sum_{i=1}^{k} f_i$.

Note: HM cannot be calculated if any observation has value zero.

- **For Continuous Frequency Distribution**

  If the observations are arranged in the form of class intervals, we take x_i as the mid value of the ith class interval and use equation (6).

Here we explain how to compute harmonic mean using Excel 2007 for raw data, discrete and continuous frequency distribution, one at a time for Problem 4.

6.5.1 Harmonic Mean for Raw Data

In order to compute the average speed of the car for the given data, we apply the harmonic mean and follow the steps given below:

Step 1: We enter the data of Table 2 in an Excel sheet as shown in Fig. 6.34 and name it "**HM**".

	A	B	C
1	S. No.	Speed (km/h)	
2	1	45	
3	2	25	
4	3	30	
5	4	30	
6	5	40	
7	6	45	
8	7	50	
9	8	50	
10	9	60	
11	10	65	

HM

Fig. 6.34

Step 2: To compute the harmonic mean, we refer to Fig. 6.35 and

1. select Cell B52,
2. click on the *Formulas* tab, and
3. click on *More Functions* → *Statistical* → *Harmean*.

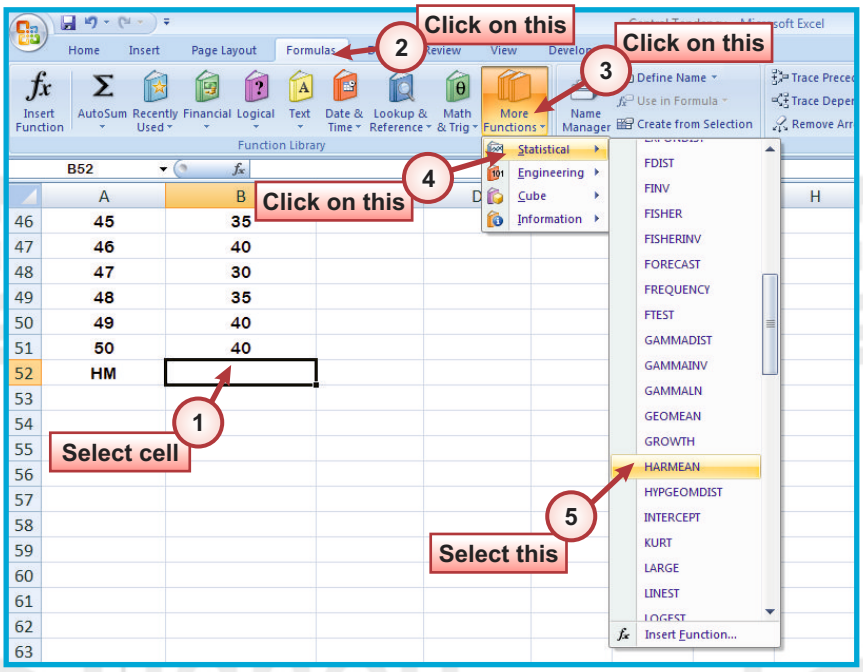


Fig. 6.35

Step 3: A new dialog box appears. We select Cells B2:B51 as shown in Fig. 6.36 and click on ***OK***.

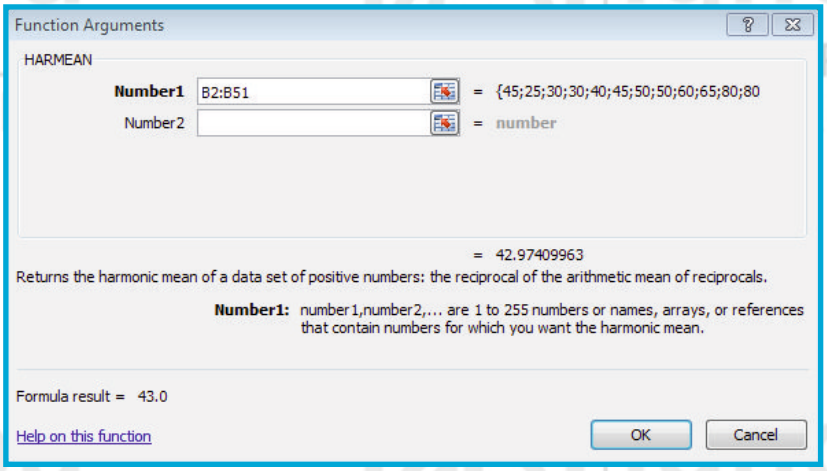


Fig. 6.36

Step 4: The value of the harmonic mean is shown in Cell B52 (Fig. 6.37).

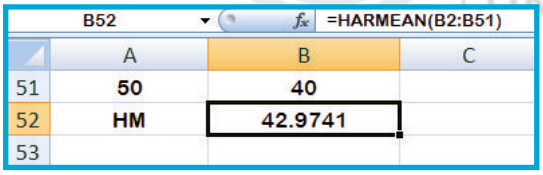


Fig. 6.37

From Fig. 6.37, the harmonic mean is 42.9741. Therefore, the average speed of the car is approximately 42.97 km per hour.

To explain the computation of the harmonic mean for the grouped frequency distribution, we shall first construct the discrete frequency distribution for the data given in Table 2. Here we are continuing with the same Excel sheet. Note that Excel does not have any built-in function for computing the harmonic mean for the frequency distributions.

6.5.2 Harmonic Mean for Discrete Frequency Distribution

In order to compute the average speed of the car for the discrete frequency distribution, we follow the steps given below:

Step 1: We construct the discrete frequency distribution for the data of Table 2 as explained in Sec. 2.4 of Lab Session 2. The output is shown in Fig. 6.38.

	D	E
1	Speed (km/h) (x)	Distance (km) (f)
2	25	3
3	30	6
4	35	6
5	40	6
6	45	6
7	50	5
8	55	5
9	60	6
10	65	3
11	75	2
12	80	2

Fig. 6.38

Step 2: To compute f/x for the first value, we type "=E2/D2" in Cell F2 as shown in Fig. 6.39a and press ***Enter***. We obtain the result shown in Fig. 6.39b.

F2 | =E2/D2

	D	E	F
1	Speed (km/h) (x)	Distance (km) (f)	f / x
2	25	3	0.1200
3	30	6	

(a)

ENTER

HARMEAN | =E2/D2

	D	E	F
1	Speed (km/h) (x)	Distance (km) (f)	f / x
2	25	3	=E2/D2
3	30	6	

(b)

Fig. 6.39

Step 3: We drag down Cell F2 up to Cell F12 to compute (f/x) for the remaining values as shown in Fig. 6.40.

F2 | =E2/D2

	D	E	F	G
1	Speed (km/h) (x)	Distance (km) (f)	f / x	
2	25	3	0.1200	
3	30	6		
4	35	6		

(a)

DRAG IT DOWN

	D	E	F	G
1	Speed (km/h) (x)	Distance (km) (f)	f / x	
2	25	3	0.1200	
3	30	6	0.2000	
4	35	6	0.1714	
5	40	6	0.1500	
6	45	6	0.1333	
7	50	5	0.1000	
8	55	5	0.0909	
9	60	6	0.1000	
10	65	3	0.0462	
11	75	2	0.0267	
12	80	2	0.0250	

(b)

Fig. 6.40

Step 4: We determine the values of $\sum f$ and $\sum f/x$ in Cells E13 and F13, using the ***Sum*** function as shown in Figs. 6.41a and b, respectively.

E13 | =SUM(E2:E12)

	D	E	F
12	80	2	0.0250
13	Total	50	
14			

(a)

F13 | =SUM(F2:F12)

	D	E	F
12	80	2	0.0250
13	Total	50	1.1635
14			

(b)

Fig. 6.41

Step 5: We type "=E13/F13" in Cell F14 to compute the harmonic mean as shown in Fig. 6.42a and press ***Enter***. We get the result shown in Fig. 6.42b.

HARMEAN =E13/F13

	D	E	F
12	80	2	0.0250
13	Total	50	1.1635
14		HM	=E13/F13
15			

(a)

ENTER

F14 =E13/F13

	D	E	F
12	80	2	0.0250
13	Total	50	1.1635
14		HM	42.9741
15			

(b)

Fig. 6.42

From Fig. 6.42b, the value for harmonic mean is given by

$$HM = 42.9741$$

Thus, the average speed of the car is 42.9741 km/hr for the entire journey on the basis of the discrete frequency distribution.

6.5.3 Harmonic Mean for Continuous Frequency Distribution

We now explain how to compute the harmonic mean for a continuous frequency distribution.

Step 1: We construct the continuous frequency distribution as explained in Sec. 2.5 of Lab Session 2 and also compute the mid values as explained in Steps 4 to 6 of Sec. 4.6 of Lab Session 4. The output is shown in Fig. 6.43.

	H	I	J	K	L	M	N
1		Class Interval	Bin	Distance (km) (Frequency) (f)	Lower Limit	Upper Limit	Mid Value (x)
2	Min. =	20-30	29	3	20	30	25
3	25	30-40	39	12	30	40	35
4	Max. =	40-50	49	12	40	50	45
5	80	50-60	59	10	50	60	55
6	Class Width	60-70	69	9	60	70	65
7	8.2782	70-80	79	2	70	80	75

Fig. 6.43

Step 2: We type "=K2/N2" in Cell O2 and drag it down up to Cell O7 to compute f/x for all classes as shown in Fig. 6.44a and b.

O2 =K2/N2

(a)

	K	L	M	N	O	P
1	Distance (km) (Frequency) (f)	Lower Limit	Upper Limit	Mid Value (x)	f / x	
2	3	20	30	25	0.1200	
3	12	30	40	35		
4	12	40	50	45		

DRAG IT DOWN

(b)

	K	L	M	N	O	P
1	Distance (km) (Frequency) (f)	Lower Limit	Upper Limit	Mid Value (x)	f / x	
2	3	20	30	25	0.1200	
3	12	30	40	35	0.3429	
4	12	40	50	45	0.2667	
5	10	50	60	55	0.1818	
6	9	60	70	65	0.1385	
7	2	70	80	75	0.0267	

Fig. 6.44

Step 3: We compute the values of $\sum f$ and $\sum f/x$ in Cells K8 and O8, respectively, using the *Sum* function as shown in Fig. 6.45.

K8 =SUM(K2:K7)

	J	K	L
1	Bin	Distance (km) (Frequency) (f)	Lower Limit
2	29	3	20
3	39	12	30
4	49	12	40
5	59	10	50
6	69	9	60
7	79	2	70
8	N = ∑f	48	
9			

(a)

O8 =SUM(O2:O7)

	N	O	P	Q
1	Mid Value (x)	f / x		
2	25	0.1200		
3	35	0.3429		
4	45	0.2667		
5	55	0.1818		
6	65	0.1385		
7	75	0.0267		
8	∑f/x	1.0765		
9				

(b)

Fig. 6.45

Step 4: To determine the harmonic mean using equation (6), we type "=K8/O8" in Cell O9 as shown in Fig. 6.46a and press ***Enter***. We get the result shown in Fig. 6.46b.

HARMEAN =K8/O8

	N	O	P
7	75	0.0267	
8	∑f/x	1.0765	
9	HM	=K8/O8	
10			

(a)

ENTER

O9 =K8/O8

	N	O	P
7	75	0.0267	
8	∑f/x	1.0765	
9	HM	44.5902	
10			

(b)

Fig. 6.46

From Fig. 6.46b, the harmonic mean is given by

$$HM = 44.5902$$

Thus, the average speed is 44.5902 km/hr for the entire journey on the basis of the continuous frequency distribution.

6.6 MODE

You have learnt in Unit 1 of MST-002 that the mode is defined as the value that is repeated most often in the data. Mode is widely used in production, i.e., to determine shirt size, shoe size, etc. Here we briefly mention the important points as follows:

- **Mode for Ungrouped (Raw) Data**

  Let $x_1, x_2, \ldots, x_n$ be n observations. If some observations are repeated in the data, we determine the observation which is repeated the highest number of times. This value is known as mode.

- **Mode for Discrete Frequency Distribution**

  If observations $x_1, x_2, \ldots, x_k$ occur with frequencies $f_1, f_2, \ldots, f_k$, respectively, the mode will be the value corresponding to the highest frequency.

- **Mode for Continuous Frequency Distribution**

  For continuous frequency distribution, we first determine the modal class. Then we use the formula given below to compute the mode:

$$M_0 = L + \frac{f_1 - f_0}{(2f_1 - f_0 - f_2)} \times h \qquad \ldots (7)$$

where L = lower class limit of the modal class,

f_1 = frequency of the modal class,

f_0 = frequency of the pre-modal class,
f_2 = frequency of the post-modal class, and
h = width of the modal class.

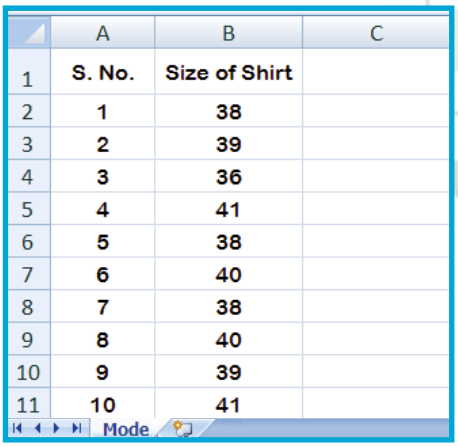
Note that modal class is the class that has the maximum frequency.

Here we explain how to compute mode using Excel 2007 for raw data, discrete and continuous frequency distributions, one at a time for Problem 5.

6.6.1 Mode for Raw Data

In order to compute mode for raw data, we follow the steps given below:

Step 1: We enter the data of Table 3 in Excel sheet as shown in Fig. 6.47 and name it "**Mode**".



	A	B	C
1	S. No.	Size of Shirt	
2	1	38	
3	2	39	
4	3	36	
5	4	41	
6	5	38	
7	6	40	
8	7	38	
9	8	40	
10	9	39	
11	10	41	

Mode

Fig. 6.47 : Partial screenshot of the spreadsheet for the given data.

Step 2: We now compute the mode for the given data. For this, we refer to Fig. 6.48 and

1. select Cell B102,
2. click on the *Formulas* tab, and
3. click on *More Functions* → *Statistical* → *Mode*.

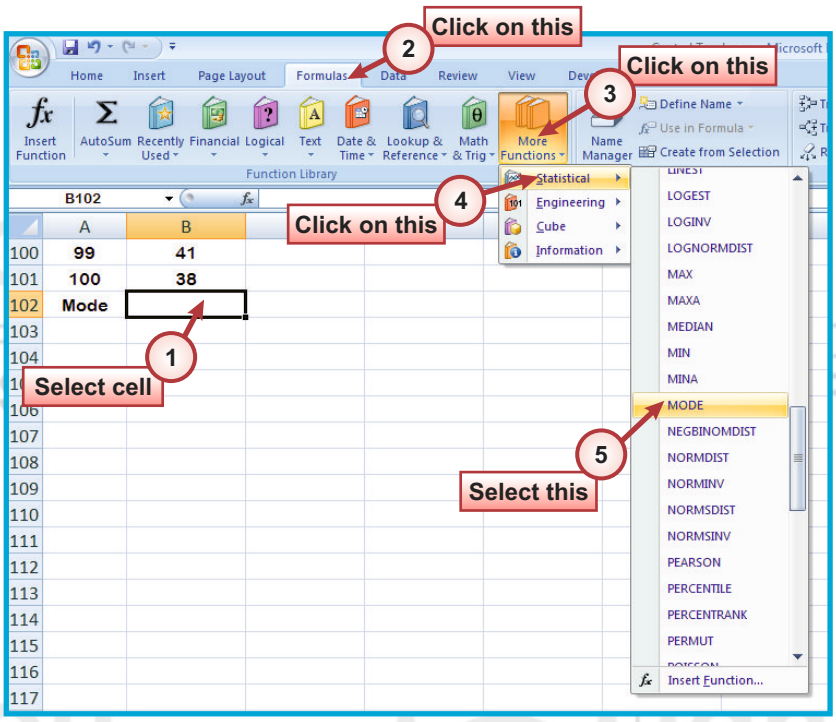


Fig. 6.48

Step 3: A new dialog box appears and we select Cell B2:B101 and click on ***OK*** as shown in Fig. 6.49.

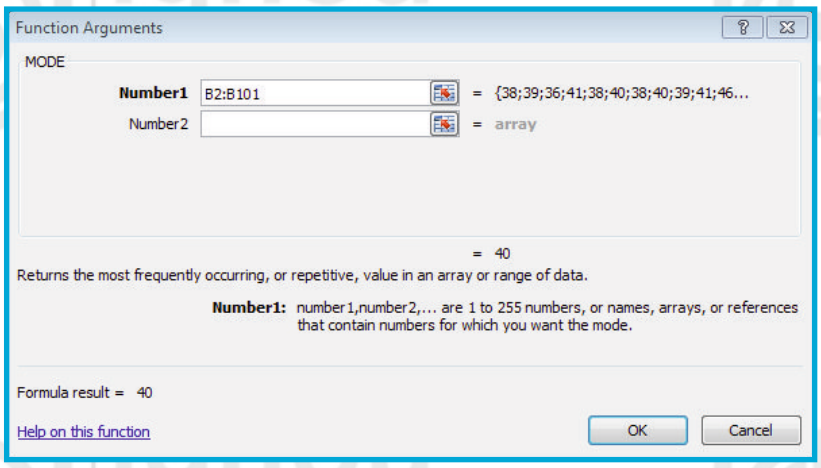


Fig. 6.49

Step 4: The value of the mode is obtained in Cell B102 as shown in Fig. 6.50.

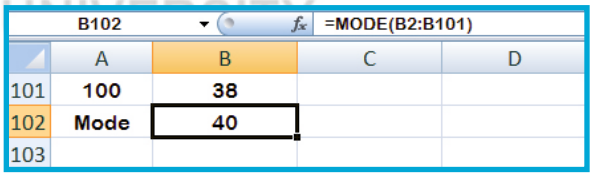


Fig. 6.50

From Fig. 6.50, the mode is given by

Mode = 40.

Therefore, the modal size of the shirt is 40.

To explain the computation of mode for the grouped frequency distribution, we shall first construct the discrete frequency distribution for the data given in Table 3. Here we are continuing with the same Excel sheet. Note that Excel does not have any built-in function to compute the mode for frequency distributions.

6.6.2 Mode for Discrete Frequency Distribution

In order to compute the average size of the shirt for the discrete frequency distribution, we follow the steps given below:

Step 1: We now construct the discrete frequency distribution for the data of Table 3 as explained in Sec. 2.4 of Lab Session 2. The output is shown in Fig. 6.51.

	D	E	F
1	Size of Shirt	Number of Person (Frequency)	
2	34	2	
3	35	4	
4	36	6	
5	37	8	
6	38	11	
7	39	13	
8	40	18	
9	41	13	
10	42	11	
11	43	7	
12	44	4	
13	45	2	
14	46	1	

Fig. 6.51

Step 2: Since there is no built-in function in Excel to compute the mode for the discrete frequency distribution, we visually identify the mode after forming the discrete frequency distribution. Notice from Fig. 6.52 that the highest frequency, i.e., 18, corresponds to the shirt size 40. We have highlighted the corresponding cells with purple colour as shown in Fig. 6.52.

	D	E	F
1	Size of Shirt	Number of Person (Frequency)	
2	34	2	
3	35	4	
4	36	6	
5	37	8	
6	38	11	
7	39	13	
8	40	18	
9	41	13	
10	42	11	
11	43	7	
12	44	4	
13	45	2	
14	46	1	
15	Mode	40	

Fig. 6.52

Therefore, the modal size of the shirt is 40.

6.6.3 Mode for Continuous Frequency Distribution

We now explain how to compute mode for a continuous frequency distribution in the steps given below:

Step 1: We construct the continuous frequency distribution for the data of Table 3 as explained in Sec. 2.5 of Lab Session 2 (see Fig. 6.53).

	G	H	I	J	K
1		Class Interval	Bin	Number of Person (Frequency)	
2	Min. =	34-36	35	6	
3	34.0	36-38	37	14	
4	Max. =	38-40	39	24	
5	46.0	40-42	41	31	
6	Class Width	42-44	43	18	
7	1.5699	44-46	45	6	
8		46-48	47	1	

Fig. 6.53

Step 2: The highest frequency, i.e., 31, corresponds to the class 40-42. We highlight the highest frequency and its corresponding class interval with purple colour as shown in Fig. 6.54.

	H	I	J	K
1	Class Interval	Bin	Number of Person (Frequency)	
2	34-36	35	6	
3	36-38	37	14	
4	38-40	39	24	
5	40-42	41	31	
6	42-44	43	18	
7	44-46	45	6	
8	46-48	47	1	

Fig. 6.54

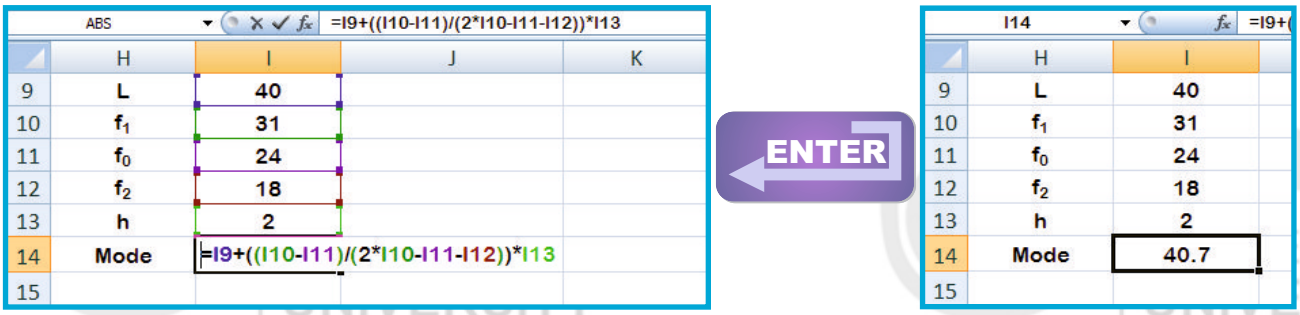
Step 3: We type the values of L, f_1, f_0, f_2 and h in Cells I9, I10, I11, I12 and I13, respectively, as shown in Fig. 6.55.

	H	I	J
1	Class Interval	Bin	Number of Person (Frequency)
2	34-36	35	6
3	36-38	37	14
4	38-40	39	24
5	40-42	41	31
6	42-44	43	18
7	44-46	45	6
8	46-48	47	1
9	L	40	
10	f_1	31	
11	f_0	24	
12	f_2	18	
13	h	2	

Fig. 6.55

In the same way as for discrete frequency distribution, we can compute the weighted average by considering weights (w) instead of frequency (f).

Step 4: To compute mode for the continuous frequency distribution, we use equation (7). The values of L, f_1, f_0, f_2 and h are given in Cells I9, I10, I11, I12 and I13, respectively. We type "= I9+((I10-I11)/(2*I10-I11-I12))*I13" in Cell I14 as shown in Fig. 6.56a and press ***Enter***. The output is shown in Fig. 6.56b.



(a) (b)

Fig. 6.56

From Fig. 6.56b, the mode is given by

$$\text{Mode} = 40.7 \approx 41$$

Thus, the modal size of the shirt is approximately 41.

6.7 PARTITION VALUES

In Unit 1 of MST-002, you have learnt that the partition values are the measures which divide the data into several equal parts. The median divides the data into 2 equal parts, quartiles divide the data into 4 equal parts, deciles divide data into 10 equal parts, and percentiles divide data into 100 equal parts. Here we briefly mention the important points and formulae used to compute these partition values as follows:

Median is the middle most value of the variable which divides the entire data into two equal parts.

For Ungrouped (Raw) Data

If $x_1, x_2, \ldots, x_n$ are n observations, we arrange the given data in ascending or descending order and compute the partition values as follows:

- The **median** is computed from

$$M_d = \left(\frac{n+1}{2}\right)^{th} \text{ observation} \quad \text{(when n is odd)} \qquad \ldots (8)$$

$$M_d = \frac{\left(\frac{n}{2}\right)^{th} \text{observation} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2} \quad \text{(when n is even)} \quad \ldots (9)$$

- The **k^{th} quartile** is

$$Q_k = k\left(\frac{n+1}{4}\right)^{th} \text{observation} \qquad (k = 1, 2, 3) \qquad \ldots (10)$$

- The **k^{th} decile** is

$$D_k = k\left(\frac{n+1}{10}\right)^{th} \text{observation} \qquad (k = 1, 2, \ldots, 9) \qquad \ldots (11)$$

- The **k^{th} percentile** is

$$P_k = k\left(\frac{n+1}{100}\right)^{th} \text{observation} \qquad (k = 1, 2, \ldots, 99) \qquad \ldots (12)$$

For Discrete Frequency Distribution

- If observations $x_1, x_2, \ldots, x_k$ occur with frequencies $f_1, f_2, \ldots, f_k$, respectively, we compute the less than type cumulative frequencies.

- If $N = \sum_{i=1}^{k} f_i$ is the total frequency, the formula used for discrete frequency distribution can be written in the same way as equations (8), (10), (11) and (12) by considering N instead of n.

  Note that if N/2 is not the exact cumulative frequency, we consider the observation corresponding to the next cumulative frequency as the required partition value.

For Continuous Frequency Distribution

We compute the less than type cumulative frequency for the given continuous frequency distribution and then calculate the partition values as follows:

- The kth quartile is

$$Q_k = L + \frac{k\left(\frac{N}{4}\right) - C}{f} \times h \; ; \qquad (k = 1, 2, 3) \qquad \ldots (13)$$

- The kth d1ecile is

$$D_k = L + \frac{k\left(\frac{N}{10}\right) - C}{f} \times h \; ; \qquad (k = 1, 2, \ldots, 9) \qquad \ldots (14)$$

- The kth percentile is

$$P_k = L + \frac{k\left(\frac{N}{100}\right) - C}{f} \times h \; ; \qquad (k = 1, 2, \ldots, 99) \qquad \ldots (15)$$

where L – lower class limit of the corresponding partition value class,
h – width of the partition value class,

N – total frequency,

C – cumulative frequency of the class preceding the partition value class, and

f – frequency of the partition value class.

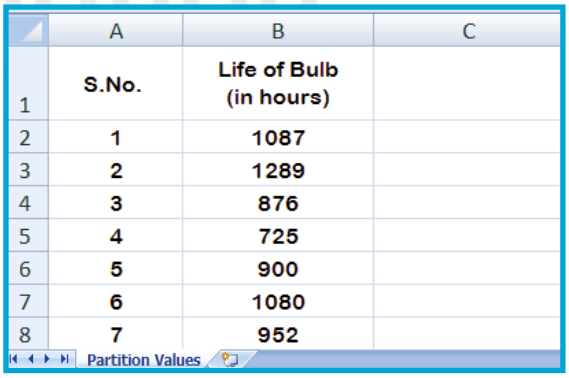
Here we describe the computation of the partition values in Excel 2007 for the raw data, discrete and continuous frequency distribution, one at a time for Problem 2.

6.7.1 Partition Values for Raw Data

In order to compute the median, quartiles, deciles and percentiles for the given data, we follow the steps given below:

Median is defined as the middle or central value of the data.

Step 1: We enter the data of Table 3 of Lab Session 2 in an Excel sheet as shown in Fig. 6.57.



	A	B	C
1	S.No.	Life of Bulb (in hours)	
2	1	1087	
3	2	1289	
4	3	876	
5	4	725	
6	5	900	
7	6	1080	
8	7	952	

Partition Values

Fig. 6.57: Partial screenshot of the spreadsheet for the given data.

Step 2: To compute the median for the raw data, Excel has a built-in function *Median*. For this, we refer to Fig. 6.58 and

1. select Cell B152,
2. click on the *Formulas* tab, and
3. click on *More Functions* → *Statistical* → *Median*.

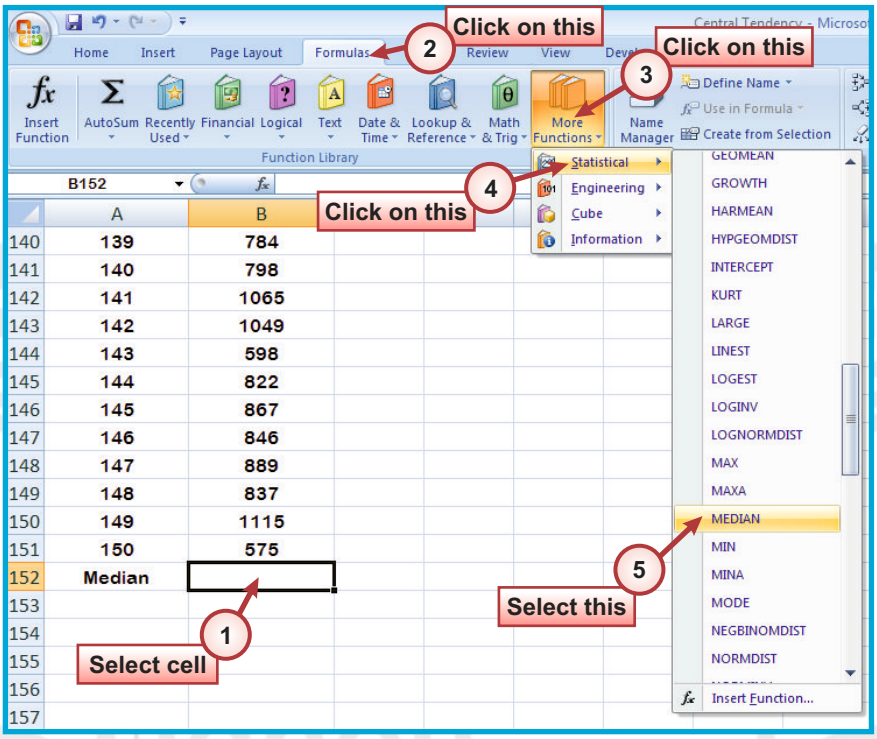
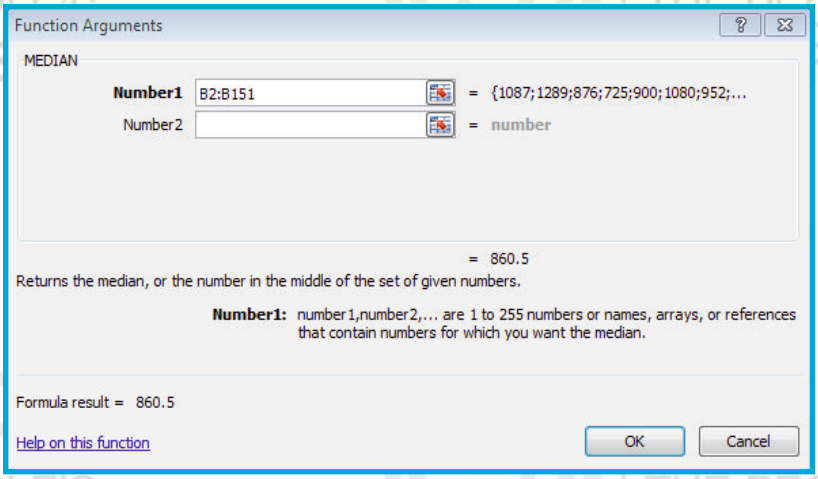


Fig. 6.58

Step 3: A new dialog box appears in which we select Cells B2:B151 and click on ***OK*** as shown in Fig. 6.59a. The value of the median is shown in Fig. 6.59b.

(a)



(b)

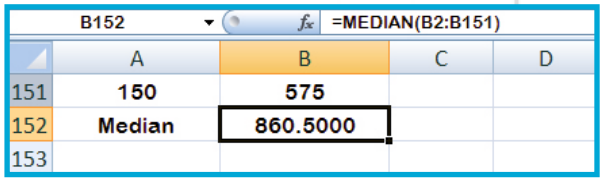


Fig. 6.59

Step 4: To compute the quartile for the raw data, Excel has the built-in function ***Quartile***. To use it, we refer to Fig. 6.60 and

1. select Cell B153,
2. click on the ***Formulas*** tab, and
3. click on ***More Functions*** → ***Statistical*** → ***Quartile***.

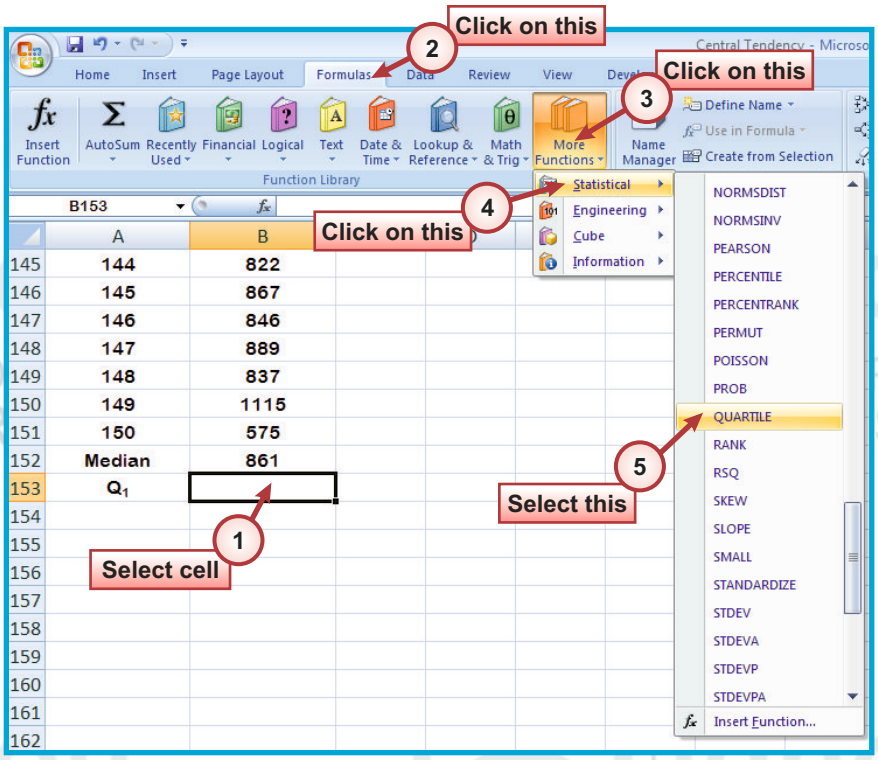
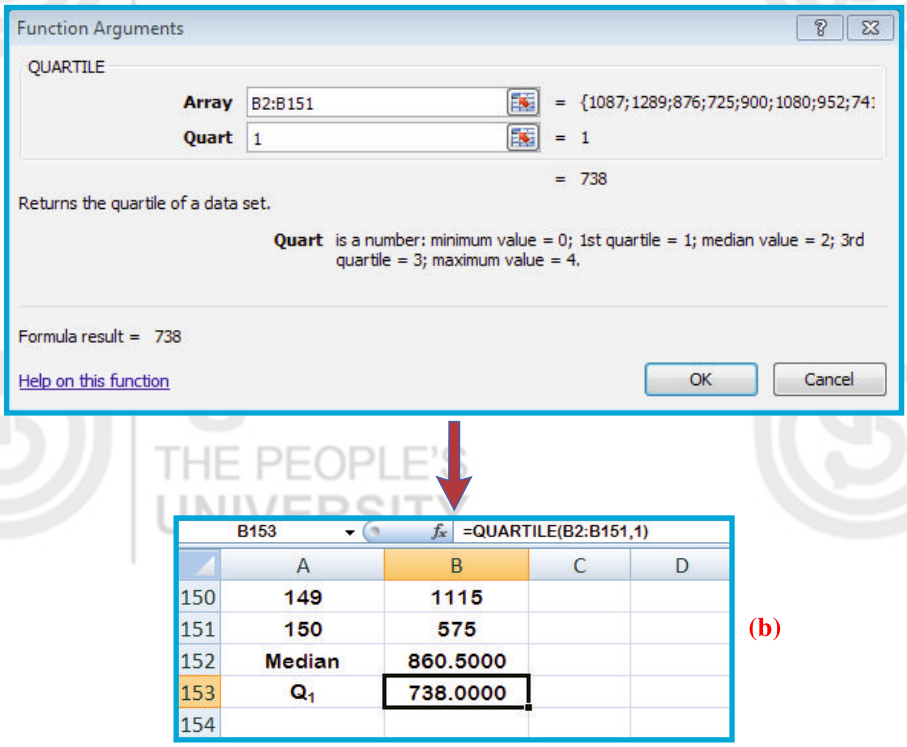


Fig. 6.60

Step 5: A new dialog box appears in which we select Cells B2:B151 in *Array* and type "1" in *Quart* to compute the first quartile and click on *OK* as shown in Fig. 6.61a. The value of the first quartile is shown in Fig. 6.61b.

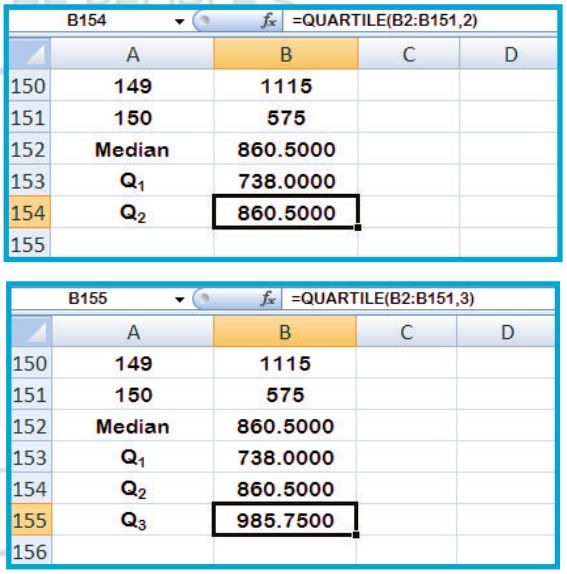


(a)

(b)

Fig. 6.61

Step 6: We can also compute the values of second and third quartiles in Cells B154 and B155 by typing "2" and "3" in *Quart* (Fig. 6.62), respectively, in the same way as explained in Steps 4 and 5.



(a)

(b)

Fig. 6.62

Step 7: Before computing deciles, we explain how to determine the percentile for the raw data. To compute the quartile for the raw data, Excel has a built-in function *Percentile*. To use it, we refer to Fig. 6.63 and

1. select Cell B156,
2. click on the *Formulas* tab, and
3. click on *More Functions* → *Statistical* → *Percentile*.

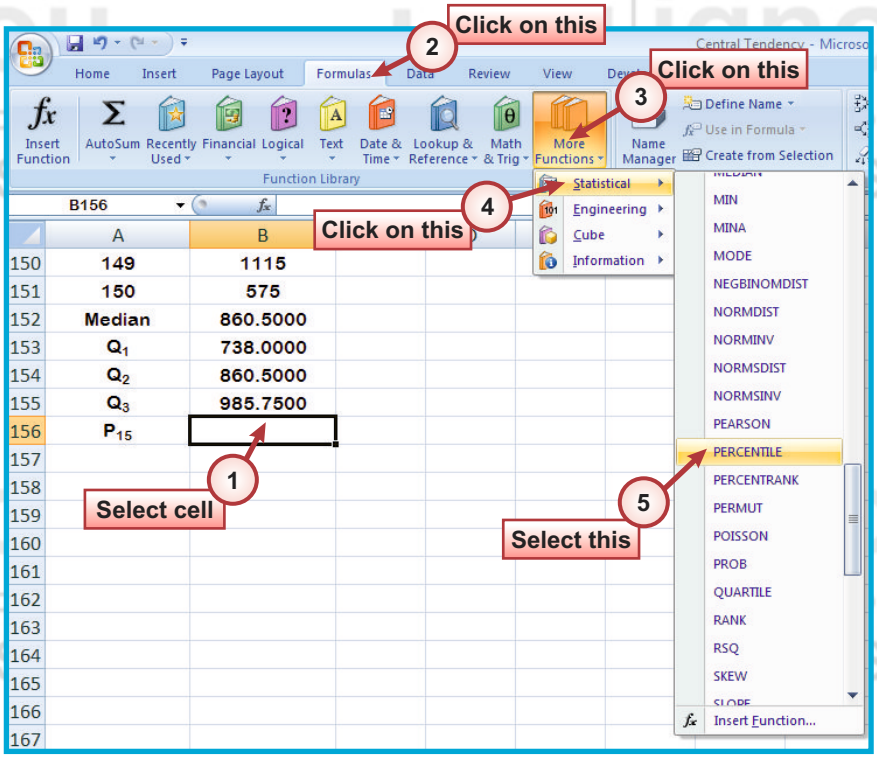


Fig. 6.63

For calculating the kth percentile, we type the value (k/100) in *K* as shown in dialog box (Fig. 6.64a).

Step 8: A new dialog box appears. To compute the 15th percentile, we select Cells B2:B151 in *Array* and type "0.15" in *K* and click on *OK* as shown in Fig. 6.64a. The value of the 15th percentile is shown in Fig. 6.64b.

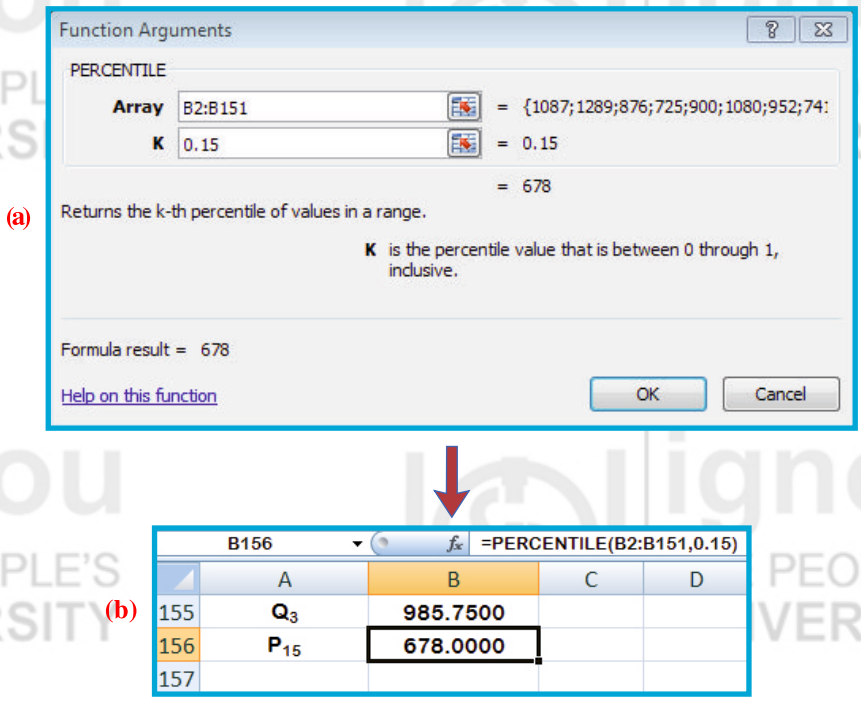
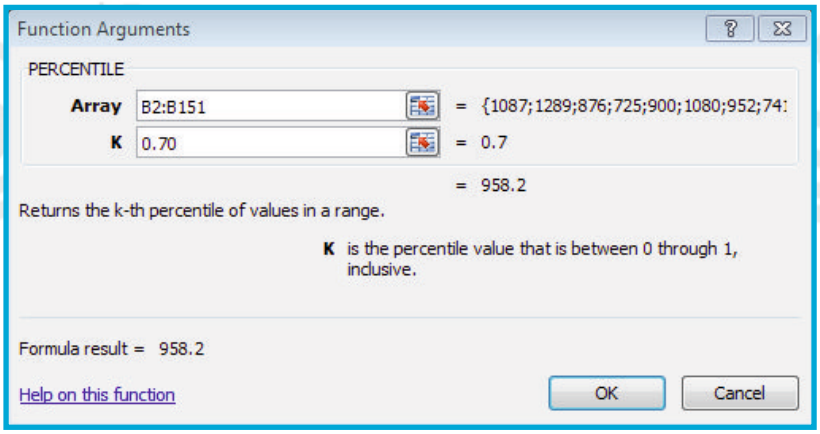
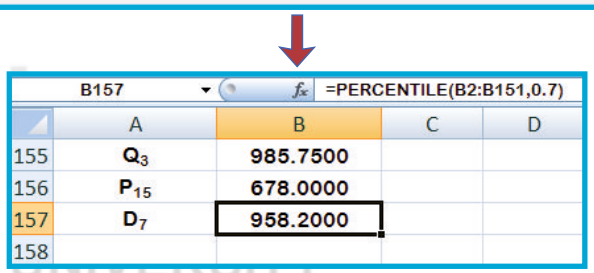


Fig. 6.64

Step 9: A built-in function to determine the deciles is not available in Excel 2007. So we calculate deciles using *Percentile* function. We repeat Step 7 by selecting Cell B157. To compute the 7th decile which corresponds to the 70th percentile, we select Cells B2:B151 in *Array* and type "0.70" in *K* and click on *OK* as shown in Fig. 6.65a. The value of the 7th decile is shown in Fig. 6.65b.



(a)



(b)

The 1st, 2nd, …, 10th deciles correspond to 10th, 20th, …, 100th percentiles.

Fig. 6.65

6.7.2 Partition Values for Discrete Frequency Distribution

Excel does not have any built-in function to compute the Partition values for the frequency distributions. In order to compute the median, quartiles, deciles and percentiles for the discrete frequency distribution, we consider Problem 1 and follow the steps given below:

Step 1: We consider the discrete frequency distribution determined in Sec. 2.4 of Lab Session 2 and compute the less than type cumulative frequencies as shown in Fig. 6.66.

	A	B	C
1	Number of Mobile Phones (x)	No. of Families (f)	No. of Bulbs (Less Than Cumulative Frequency)
2	1	6	6
3	2	16	22
4	3	22	44
5	4	24	68
6	5	22	90
7	6	16	106
8	7	9	115
9	8	5	120
10	$N = \sum f$	120	

Fig. 6.66

Step 2: We type "=B10/2" in Cell B12 to compute the value of N/2 as shown in Fig. 6.67. The value of N/2 = 60 corresponds to the cumulative frequency 68 (Cell C5). The value of the variable (number of mobile phones) corresponding to the cumulative frequency (68) is 4 (Cell A5). Thus, the median is 4.

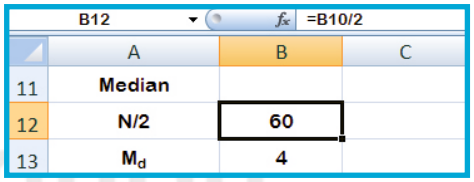


Fig. 6.67

Step 3: To compute the first quartile, we type "=B10/4" in Cell B16 and obtain the value of N/4 as 30, which corresponds to the cumulative frequency 44 (Cell C4). Thus, the number of mobile phones corresponding to the cumulative frequency (44) is 3 (Fig. 6.68). The value of the first quartile is 3 as shown in Fig. 6.68.

B16 =B10/4

	A	B	C
15	1st Quartile		
16	N/4	30	
17	Q_1	3	
18			

Fig. 6.68

Step 4: In the same way as explained in Steps 2 and 3, we can compute the 2nd quartile, 3rd quartile, 7th decile and 15th percentile as shown in Figs. 6.69a to d.

(a) B20 =2*B10/4

	A	B	C
19	2nd Quartile		
20	2N/4	60	
21	Q_2	4	

(c) B24 =3*B10/4

	A	B	C
23	3rd Quartile		
24	3N/4	90	
25	Q_3	5	

(b) B28 =7*B10/10

	A	B	C
27	7th Decile		
28	7N/10	84	
29	D_7	5	

(d) B32 =15*B10/100

	A	B	C
31	15th Percentile		
32	15N/100	18	
33	P_{15}	2	

Fig. 6.69

6.7.3 Partition Values for Continuous Frequency Distribution

We now consider Problem 2 and explain how to compute the partition values, i.e., median, quartiles, deciles and percentiles for continuous frequency distribution in the steps given below:

Step 1: We consider the continuous frequency distribution determined in Sec. 2.5 of Lab Session 2. We compute the less than type cumulative frequencies and type the value of h in Cell B11 as shown in Fig. 6.70.

	A	B	C
1	Class Interval	No. of Bulbs (f)	No. of Bulbs (Less Than Cumulative Frequency)
2	500-600	11	11
3	600-700	20	31
4	700-800	25	56
5	800-900	32	88
6	900-1000	27	115
7	1000-1100	19	134
8	1100-1200	11	145
9	1200-1300	5	150
10	$N = \sum f$	150	
11	h	100	

Fig. 6.70

Step 2: To compute the median, we refer to Fig. 6.71 and

1. compute the value of N/2 by typing "=B10/2" in Cell B14 and highlight the corresponding median class with purple colour (Fig. 6.71a),
2. type the corresponding values of L, C and f in Cells B15, B16 and B17, respectively, and

3. type "=B15+((B14-B16)/B17)*B11" in Cell B18 and get the value of median as shown in Fig.6.71b.

	A	B	C
1	Class Interval	No. of Bulbs (f)	No. of Bulbs (Less Than Cumulative
2	500-600	11	11
3	600-700	20	31
4	700-800	25	56
5	800-900	32	88
6	900-1000	27	115
7	1000-1100	19	134
8	1100-1200	11	145
9	1200-1300	5	150

(a)

B18 f_x =B15+((B14-B16)/B17)*B11

	A	B	C
13	Median		
14	N/2	75	
15	L	800	
16	C	56	
17	f	32	
18	M_d	859.3750	
19			

(b)

Fig. 6.71

Step 3: To compute the 1st quartile, we refer to Fig. 6.72 and

1. compute the value of N/4 by typing "=B10/4" in Cell B21 and highlight the corresponding 1st quartile class with purple colour (Fig.6.72a),
2. type the corresponding values of L, C and f in Cells B22, B23 and B24, respectively, and
3. type "=B22+((B21-B23)/B24)*B11" in Cell B25 and get the value as shown in Fig.6.72b.

	A	B	C
1	Class Interval	No. of Bulbs (f)	No. of Bulbs (Less Than Cumulative
2	500-600	11	11
3	600-700	20	31
4	700-800	25	56
5	800-900	32	88
6	900-1000	27	115
7	1000-1100	19	134
8	1100-1200	11	145
9	1200-1300	5	150

(a)

B25 f_x =B22+((B21-B23)/B24)*B11

	A	B	C
20	1st Quartile		
21	N/4	37.5	
22	L	700	
23	C	31	
24	f	25	
25	Q_1	726.0000	
26			

(b)

Fig. 6.72

Step 4: In the same way as explained in Steps 2 and 3, we can compute the 2nd quartile, 3rd quartile, 7th decile and 15th percentile as shown in Figs. 6.73a to d.

(a)

B32 | =B29+((B28-B30)/B31)*B11

	A	B	C
27	2nd Quartile		
28	2N/4	75	
29	L	800	
30	C	56	
31	f	32	
32	Q_2	859.3750	
33			

(c)

B39 | =B36+((B35-B37)/B38)*B11

	A	B	C
34	3rd Quartile		
35	3N/4	112.5	
36	L	900	
37	C	88	
38	f	27	
39	Q_3	990.7407	
40			

(b)

B46 | =B43+((B42-B44)/B45)*B11

	A	B	C
41	7th Decile		
42	7N/10	105	
43	L	900	
44	C	88	
45	f	27	
46	D_7	962.9630	
47			

(d)

B53 | =B50+((B49-B51)/B52)*B11

	A	B	C
48	15th Percentile		
49	15N/100	22.5	
50	L	600	
51	C	11	
52	f	20	
53	P_{15}	657.5000	
54			

Fig. 6.73

You should now apply this method to other problems for practice.

Activity

Apply suitable measures of central tendency with the help of MS Excel 2007 to :

A1) Examples 1 to 15 given in Unit 1 of MST-002.

A2) Exercises E1 to E16 given in Unit 1 of MST-002.

Match the results with the manual computation of data carried out in Unit 1 of MST-002.

Continuous Assessment 6

1. Consider the data given in Table 5 of Lab Session 2 to compute the arithmetic mean, median, quartiles, 4th decile, 37th percentile for the discrete frequency distribution.
2. Consider the data given in Table 6 of Lab Session 2 to compute the arithmetic mean, median, quartiles, 4th decile, 37th percentile for raw data and the continuous frequency distribution.
3. The price of a particular product is noted for 20 years. The yearly rise in the price is recorded in Table 4.

Table 4: Yearly rise in the price of a product

Year	Rise in Price (%)	Year	Rise in Price (%)
1994	11.8	2004	48.9
1995	12.7	2005	50.4
1996	14.4	2006	54.7
1997	19.7	2007	59.7
1998	21.6	2008	65.2
1999	26.4	2009	67.7
2000	30.2	2010	72.4
2001	38.8	2011	80.1
2002	44.7	2012	90.2
2003	47.5	2013	98.2

Compute the average percentage increase in the price of the product.

4. A student travels a total distance of 35 km from home to college by a two-wheeler. The average speed of the two-wheeler for each kilometre of distance travelled is recorded in Table 5.

Table 5: Average speed of the two-wheeler from home to college

S. No.	Speed (km/h)	S. No.	Speed (km/h)	S. No.	Speed (km/h)
1	30	13	35	25	45
2	25	14	45	26	40
3	30	15	30	27	35
4	35	16	40	28	30
5	35	17	20	29	40
6	45	18	20	30	25
7	30	19	25	31	35
8	35	20	30	32	20
9	45	21	30	33	20
10	45	22	40	34	25
11	40	23	45	35	20
12	20	24	45		

Compute the average speed of the two-wheeler during the entire journey.

5. A shoe manufacturing company conducted a survey among 50 women to determine the size of their shoes. The data is recorded in Table 6.

Table 6: Size of shoes

S. No.	Size of Shoe	S. No.	Size of Shoe	S. No.	Size of Shoe
1	5	18	5	35	8
2	6	19	6	36	5
3	7	20	5	37	7
4	8	21	10	38	6
5	5	22	9	39	8
6	7	23	8	40	7
7	5	24	7	41	4
8	7	25	6	42	9
9	6	26	5	43	7
10	8	27	6	44	6
11	9	28	8	45	7
12	9	29	7	46	5
13	7	30	6	47	7
14	7	31	7	48	6
15	7	32	8	49	8
16	4	33	6	50	5
17	8	34	9		

For this data,

- compute the most preferable average size of the shoes.
- construct the discrete and continuous frequency distributions.
- compute the most preferable average size for the frequency distributions.

Home Work: Do It Yourself

1) Follow the steps explained in Secs. 6.3 to 6.7 to compute the various measures of the central tendency. Take the final screenshots and keep them in your record book.

2) Develop the spreadsheets for the exercises of "Continuous Assessment 6" as explained in this lab session. Take screenshots of the final spreadsheets.

3) **Do not forget** to keep all screenshots in your record book as these will contribute to your continuous assessment in the Laboratory.

Measures of Dispersion, Moments, Skewness and Kurtosis

SESSION 7

Structure

7.1 Introduction
Objectives
7.2 Problem Description
7.3 Range and Coefficient of Range
For Raw Data
7.4 Quartile Deviation and Coefficient of Quartile Deviation
For Raw Data
7.5 Mean Deviation
For Raw Data
For Discrete Frequency Distribution
7.6 Variance and Standard Deviation
For Raw Data
For Discrete Frequency Distribution
7.7 Moments, Skewness and Kurtosis
For Raw Data
For Discrete Frequency Distribution
7.8 Descriptive Statistics using Data Analysis ToolPak

7.1 INTRODUCTION

In Lab Session 6, you have learnt how to compute different measures of central tendency, such as, arithmetic mean, geometric mean, harmonic mean, mode and median for raw data as well as discrete and continuous frequency distributions using MS Excel 2007. You have also learnt computation of the partition values, viz. quartiles, deciles and percentiles for raw data as well as discrete and continuous frequency distributions. These measures of central tendency provide a value around which the data is concentrated. But it does not give us any idea about the nature of scatter or spread. In Unit 2 of MST-002 (Descriptive Statistics), we have explained that the degree to which numerical data tend to spread about an average value is called the variation or dispersion of data. We have also described in Units 3 and 4, the concepts of moments, skewness and kurtosis which are useful for getting an idea about the shape of the distribution.

Prerequisite

- Lab Sessions 2 and 6 of MSTL-001 (Basic Statistics Lab).
- Units 2, 3 and 4 of MST-002 (Descriptive Statistics).

In this lab session, you will learn how to compute different measures of dispersion, such as, range, quartile deviation, and its coefficients for raw data as

well as discrete and continuous frequency distributions using MS Excel 2007. You will also learn the method of computation of mean deviation, variance, standard deviation, coefficient of variation, moments, skewness and kurtosis for raw data as well as discrete and continuous frequency distributions using MS Excel 2007.

In the next lab session, we shall explain correlation analysis for studying the linear relationship between two or more variables using Excel 2007.

Objectives

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

- prepare the spreadsheet in MS Excel 2007;
- determine the range, quartile deviation and its coefficients;
- compute the mean deviation, variance, standard deviation and coefficient of variation; and
- calculate the moments, skewness and kurtosis.

7.2 PROBLEM DESCRIPTION

In this lab session, we state two problems to illustrate the measures of dispersion, moments, skewness and kurtosis of the given data:

1. We consider the discrete frequency distribution constructed in Sec. 2.4 of Lab Session 2 and compute
   - i) the range, quartile deviation, and its coefficients,
   - ii) the mean deviation, variance, standard deviation and coefficient of variation, and
   - iii) the moments, skewness and kurtosis.
2. We consider the raw data of Table 3 of Lab Session 2 as well as continuous frequency distribution constructed in Sec. 2.5 of Lab Session 2 and compute
   - i) the range, quartile deviation, and its coefficients,
   - ii) the mean deviation, variance, standard deviation and coefficient of variation, and
   - iii) the moments, skewness and kurtosis.

7.3 RANGE

You have learnt in Unit 2 of MST-002 that range is the absolute measure of dispersion and coefficient of range is a relative measure of dispersion. In this section, we briefly mention the main points for calculating range as follows:

- We determine the minimum and maximum value of the variable for the given data.
- Range is the difference between the maximum and the minimum value of the variable. It can be defined as

$$R = X_{Max} - X_{Min} \qquad \ldots (1)$$

where X_{Max} – maximum value of the variable, and

X_{Min} – minimum value of the variable.

- The coefficient of range is given by

$$\text{Cofficient of range} = \frac{X_{Max} - X_{Min}}{X_{Max} + X_{Min}} \quad \ldots (2)$$

Steps in Excel

In order to compute the range and the coefficient of range for the raw data in MS Excel 2007, we consider Problem 2 and follow the steps given below.

7.3.1 For Raw Data

Step 1: We enter the data given in Table 3 of Lab Session 2, in Excel 2007 spreadsheet as shown in Fig. 7.1.

	A	B	C
1	S.No.	Life of Bulb (in hours)	
2	1	1087	
3	2	1289	
4	3	876	
5	4	725	
6	5	900	
7	6	1080	
8	7	952	
9	8	741	
10	9	1000	
11	10	900	
12	11	957	
13	12	745	

Fig. 7.1: Partial screenshot of the spreadsheet for the given data.

Step 2: We compute the maximum value of the given data in Cell B152 using the ***Max*** function as explained in Sec. 2.5 of Lab Session 2 (Fig. 7.2).

B152 =MAX(B2:B151)

	A	B	C
150	149	1115	
151	150	575	
152	Max	1298	
153			

Fig. 7.2

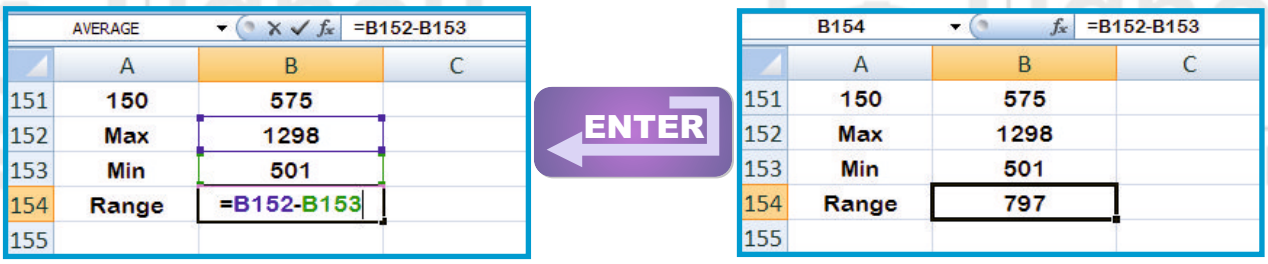
Step 3: We compute the minimum value of the given data in Cell B153 using the ***Min*** function as explained in Sec. 2.5 of Lab Session 2 (Fig. 7.3).

B153 =MIN(B2:B151)

	A	B	C
151	150	575	
152	Max	1298	
153	Min	501	
154			

Fig. 7.3

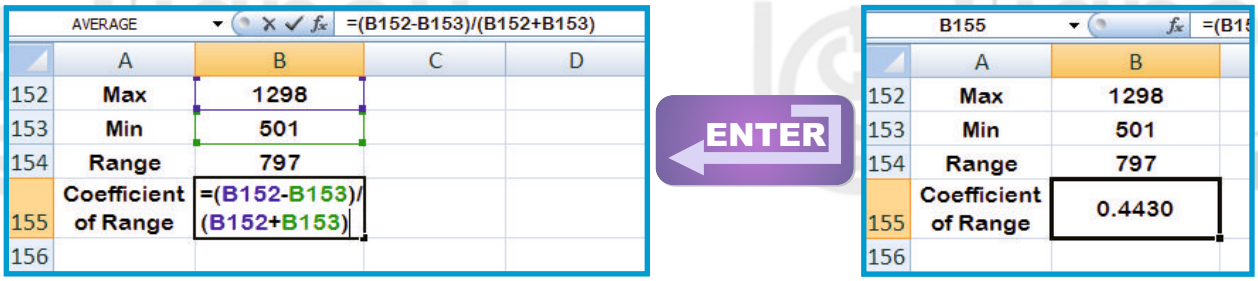
Step 4: To compute the range of the given data, we use equation (1), type "=B152-B153" in Cell B154 and press ***Enter*** as shown in Fig. 7.4.



(a) (b)

Fig. 7.4

Step 5: To compute the coefficient of range of the given data using equation (2), we type "=(B152-B153)/(B152+B153)" in Cell B155 and press ***Enter*** as shown in Fig. 7.5.



(a) (b)

Fig. 7.5

You have learnt how to compute the range and its coefficient for the raw data. You should try to compute the range and its coefficients for discrete as well as continuous frequency distributions before going to the next section.

Note: 1. For discrete frequency distribution, the minimum and maximum value will be the lowest and largest values, respectively, taken by the variable.

2. The minimum and maximum value for the continuous frequency distribution will be the lower limit of the first class interval and the upper limit of the last class interval, respectively.

Activity 1

Compute the range and coefficient of range for the discrete and continuous frequency distributions constructed in Secs. 2.4 and 2.5, respectively, of Lab Session 2 with the help of MS Excel 2007.

7.4 QUARTILE DEVIATION

In Unit 2 of MST-002, you have studied the absolute and relative measures of dispersion which are based on the first and third quartiles. In this section, we mention the main points as follows:

- If Q_1 and Q_3 are the first and third quartiles, respectively, the inter quartile range is calculated as

$$\text{Inter quartile range} = (Q_3 - Q_1) \qquad \ldots (3)$$

- The quartile deviation (QD) or the semi inter quartile range is an absolute measure of dispersion and is calculated as

$$\text{Quartile deviation (QD)} = (Q_3 - Q_1)/2 \qquad \ldots (4)$$

- The relative measure of QD is known as the coefficient of QD and is defined as

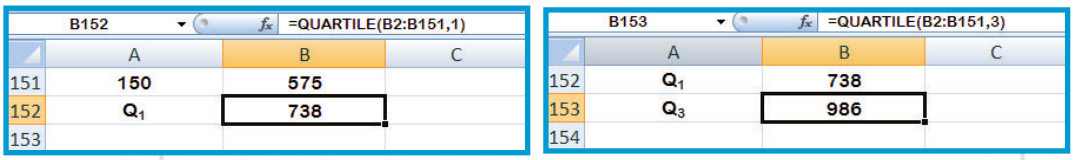
$$\text{Cofficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad \ldots (5)$$

Here we describe the computation of the quartile deviation and its coefficient in Excel 2007 for raw data.

7.4.1 For Raw Data

We consider Problem 2 and explain the quartile deviation and its coefficient for raw data in MS Excel 2007. In order to compute the quartile deviation and its coefficient, we follow the steps given below:

Step 1: We compute the 1st and 3rd quartiles in Cells B152 and B153, respectively, as explained in Sec. 6.7 of Lab Session 6 (see Fig. 7.6).



(a) (b)

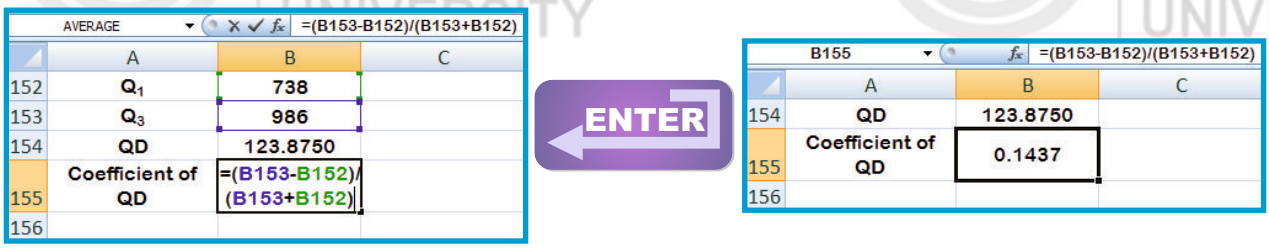
Fig. 7.6

Step 2: The formula for quartile deviation is given in equation (4). The values of Q_1 and Q_3 are calculated in Cells B152 and B153, respectively. So we type "=(B153-B152)/2" in Cell B154 and press ***Enter*** to compute the quartile deviation as shown in Fig. 7.7.

(a) (b)

Fig. 7.7

Step 3: To compute the coefficient of quartile deviation, we type "=(B153-B152)/(B153+B152)" in Cell B155 and press ***Enter*** as shown in Fig. 7.8.



(a) (b)

Fig. 7.8

You have learnt how to compute the quartile deviation and its coefficient for the raw data. You should try to compute the quartile deviation and its coefficients for

discrete as well as continuous frequency distributions before going to the next section.

Activity 2

Compute the quartile deviation and coefficient of quartile deviation for the discrete as well as continuous frequency distributions constructed in Secs. 2.4 and 2.5, respectively, of Lab Session 2 with the help of MS Excel 2007.

7.5 MEAN DEVIATION

You have learnt in Unit 2 of MST-002 that the range and quartile deviation are the positional measures of dispersion and these measures are not based on all observations. Also these do not measure the scatter in observations around an average. To overcome this drawback, we use mean deviation. Here we briefly mention the main points to calculate the mean deviation as follows:

- We consider any arbitrary value around which the mean deviation is to be computed since the mean deviation is defined as the average of the sum of the absolute values of deviation from any arbitrary value, viz., mean, median, mode, etc.
- **For Raw Data**

  The deviation of an observation x_i from the assumed arbitrary value, say, A, is defined as $(x_i - A)$. Therefore, the mean deviation of n observations, say $x_1, x_2, \ldots, x_n$ is given by

  $$MD = \frac{1}{n}\sum_{i=1}^{k}|x_i - A| \qquad \ldots (6)$$

  where A – any arbitrary value, e.g., mean, median, mode, etc.
- **For Frequency Distributions**

  If $x_1, x_2, \ldots, x_k$ are k observations (or mid values in case of class intervals) with corresponding frequencies $f_1, f_2, \ldots, f_k$, respectively, the mean deviation is defined as

  $$MD = \frac{\sum_{i=1}^{k} f_i|x_i - A|}{N} \quad \text{where } N = \sum_{i=1}^{k} f_i. \qquad \ldots (7)$$

It is often suggested that mean deviation be calculated from the median because it has the least value when measured from the median.

Here we describe the computation of the mean deviation about mean in Excel 2007 for raw data, discrete and continuous frequency distribution.

7.5.1 For Raw Data

We consider Problem 2 and explain the mean deviation in case of raw data in Excel 2007. In order to compute the mean deviation about mean, we follow the steps given ahead:

Step 1: We enter the data of Table 3 of Lab Session 2 in an Excel spreadsheet and compute the arithmetic mean in Cell B152 as

explained in Sec. 6.3 of Lab Session 6. The output is shown in Fig. 7.9.

B152 =AVERAGE(B2:B151)

	A	B	C
151	150	575	
152	$\bar{x}$	868.0600	
153			

Fig. 7.9

The formula with dollar sign ($) is used for an absolute reference.

Step 2: We compute the value of the deviation of the first sample from the mean, i.e., $(x-\bar{x})$ by typing "=B2-$B$152" in Cell C2 as shown in Fig. 7.10a. Note that here we are using absolute reference, i.e., BB152 to fix Cell B152 as it contains the value of the arithmetic mean. We drag down Cell C2 up to Cell C151 as shown in Fig. 7.10b.

C2 =B2-$B$152

	A	B	C	D
1	S.No.	Life of Bulb (in hours) (x)	$(x-\bar{x})$	
2	1	1087	218.9400	
3	2	1289		

(a)

DRAG IT DOWN

C2 =B2-$B$152

	A	B	C	D
1	S.No.	Life of Bulb (in hours) (x)	$(x-\bar{x})$	
2	1	1087	218.9400	
3	2	1289	420.9400	
4	3	876	7.9400	
5	4	725	-143.0600	
6	5	900	31.9400	
7	6	1080	211.9400	
8	7	952	83.9400	
9	8	741	-127.0600	
10	9	1000	131.9400	
11	10	900	31.9400	

(b)

Fig. 7.10

Step 3: To compute the absolute deviation from the mean for the first sample, we refer to Fig. 7.11 and

1. select Cell D2,
2. click on the ***Formulas*** tab, and
3. click on the ***Math & Trig*** → ***Abs***.

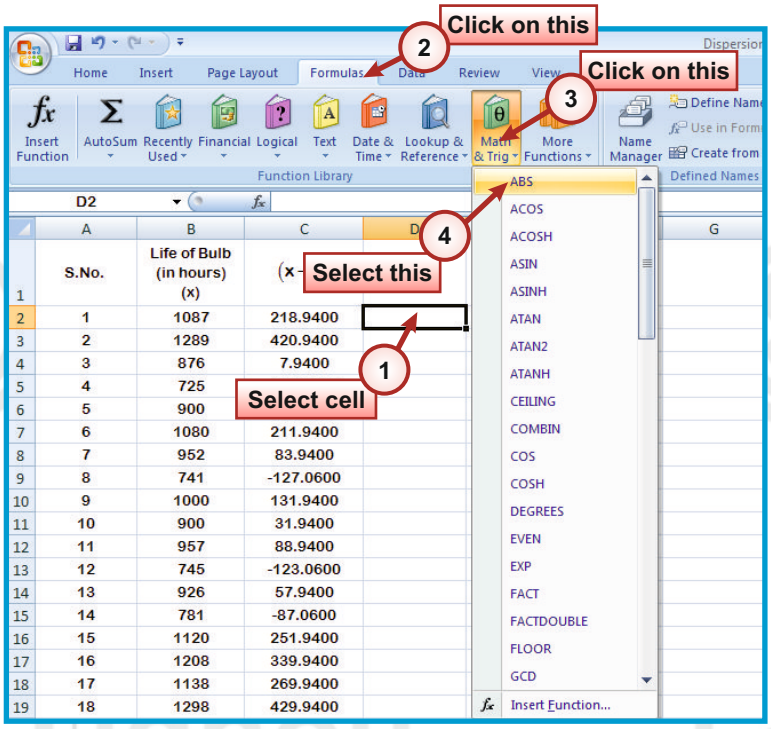
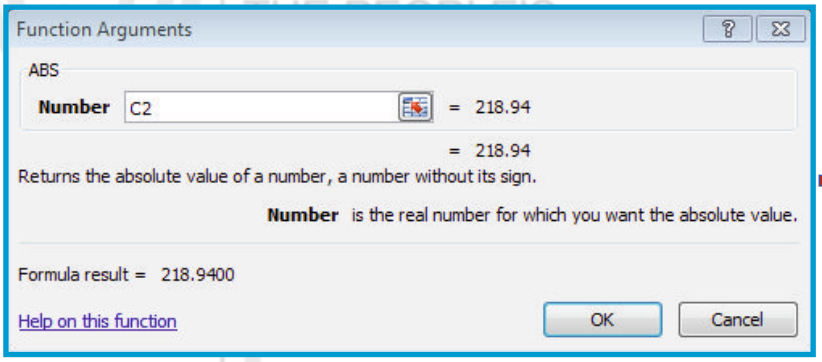
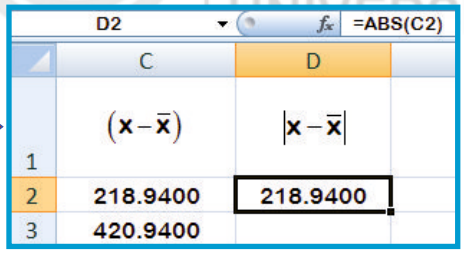


Fig. 7.11

Step 4: A new dialog box opens. We select Cell C2 and click on ***OK*** as shown in Fig. 7.12a. The value of the absolute deviation from the mean for the first sample is obtained in Cell D2 as shown in Fig. 7.12b.



(a)



(b)

Fig. 7.12

Step 5: We drag down Cell D2 up to Cell D151 to determine the absolute deviation from the mean for the remaining samples (Fig. 7.13).

	C	D	E
1	$(x-\bar{x})$	$\lvert x-\bar{x}\rvert$	
2	218.9400	218.9400	
3	420.9400	420.9400	
4	7.9400	7.9400	
5	-143.0600	143.0600	
6	31.9400	31.9400	
7	211.9400	211.9400	
8	83.9400	83.9400	
9	-127.0600	127.0600	
10	131.9400	131.9400	
11	31.9400	31.9400	
12	88.9400	88.9400	
13	-123.0600	123.0600	
14	57.9400	57.9400	

Fig. 7.13

Step 6: The mean deviation is the average of all absolute deviations from the mean obtained in Cells D2:D151 [equation (6)]. So we use the ***Average*** function in Cell B153 and select Cells D2:D151 as explained in Sec. 6.3 of Lab Session 6 to compute the mean deviation about the mean as shown in Fig. 7.14.

B153 =AVERAGE(D2:D151)

	A	B	C	D
151	150	575	-293.0600	293.0600
152	$\bar{x}$	868.0600		
153	MD	144.0373		
154				

Fig. 7.14

Step 7: We can also directly compute the mean deviation about mean using ***Avedev*** function as shown in Fig. 7.15. For this, we

1. select Cell B154,
2. click on the ***Formulas*** tab, and
3. click on ***More Functions*** → ***Statistical*** → ***Avedev***.

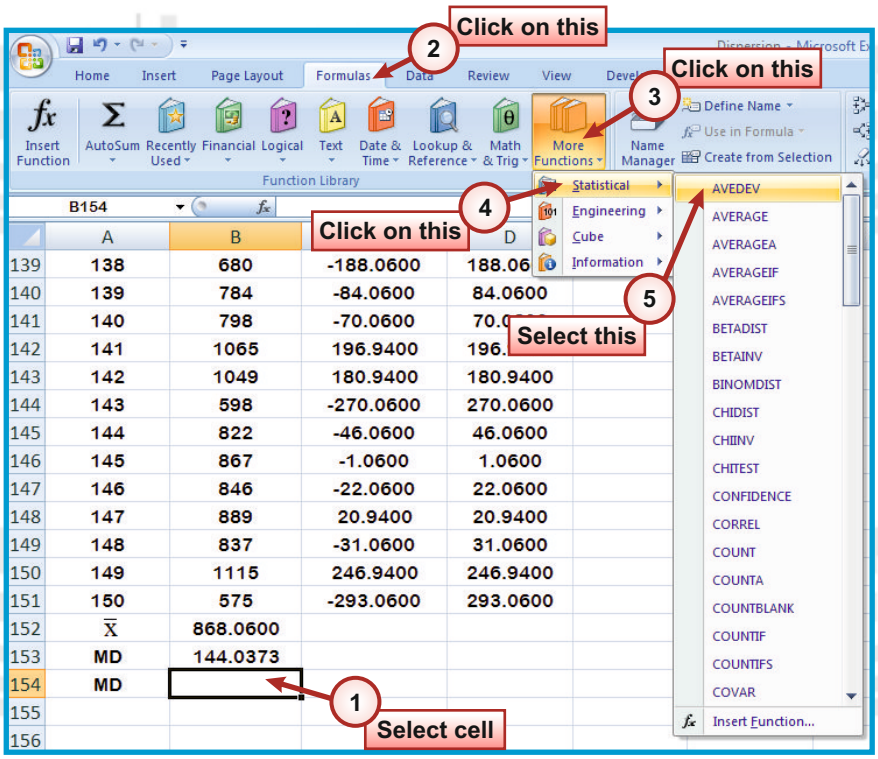
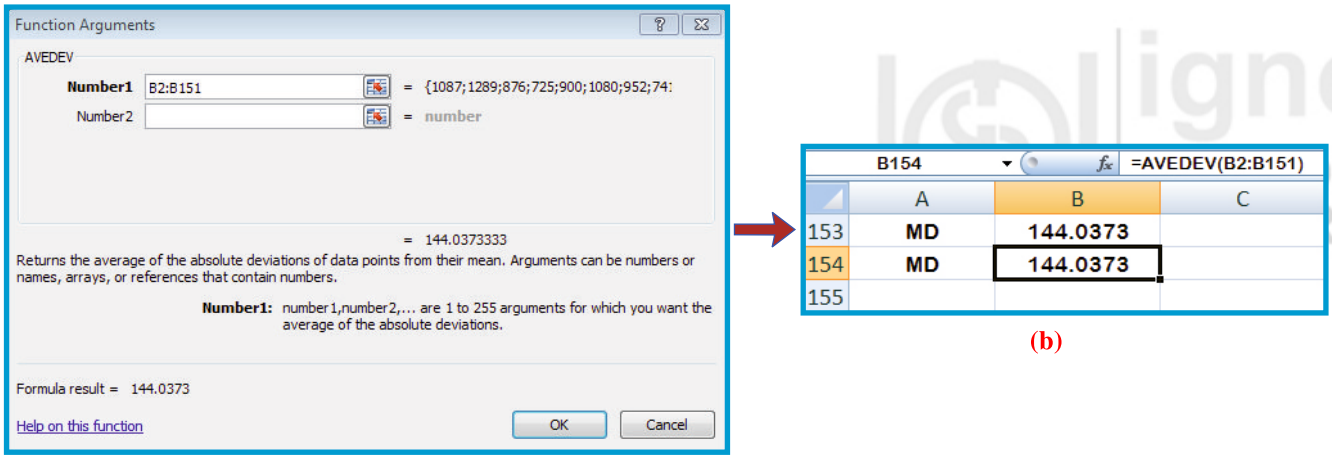


Fig. 7.15

Step 8: A new dialog box opens. We select Cells B2:B151 and click on ***OK*** as shown in Fig. 7.16a. We get the mean deviation about the mean in Cells B154 as shown in Fig. 7.16b.



(a) (b)

Fig. 7.16

Note from Fig. 7.16b that the value of mean deviation about the mean computed in Cell B153 is the same as that computed using ***Avedev*** function in Cell B154.

We now explain the computation of the mean deviation for a discrete frequency distribution.

7.5.2 For Discrete Frequency Distribution

We consider Problem 1 and explain the mean deviation for a discrete frequency distribution in MS Excel 2007. Note that Excel does not have any built-in function to compute the mean deviation for the frequency distributions. In order to compute the mean deviation, we follow the steps given below:

Step 1: To compute the mean deviation about the mean for Problem 1, we repeat Steps 1 to 6 of Sec. 6.3.2 of Lab Session 6. The output is shown in Fig. 7.17.

	A	B	C	D	E
1	S.No.	Number of Mobile Phones (x)	No. of Families (f)	fx	
2	1	1	6	6	
3	2	2	16	32	
4	3	3	22	66	
5	4	4	24	96	
6	5	5	22	110	
7	6	6	16	96	
8	7	7	9	63	
9	8	8	5	40	
10	Total	36	120	509	
11	$\bar{x}$	4.2417			
12					

Fig. 7.17

The formula with dollar sign ($) is used for an absolute reference.

Step 2: We compute the deviation in the values from the mean, i.e., $(x-\bar{x})$ by typing "=B2-$B$11" in Cell E2 and dragging it down up to Cell E9. We also compute the absolute deviation from the mean, i.e., $|x-\bar{x}|$ as explained in Steps 3 to 5 of Sec. 7.5.1. Note that here we are using absolute reference, i.e., $B$11 to fix Cell B11 as it contains the value of arithmetic mean. The output is shown in Fig. 7.18.

	A	B	C	D	E	F
1	S.No.	Number of Mobile Phones (x)	No. of Families (f)	fx	$(x-\bar{x})$	$\lvert x-\bar{x}\rvert$
2	1	1	6	6	-3.2417	3.2417
3	2	2	16	32	-2.2417	2.2417
4	3	3	22	66	-1.2417	1.2417
5	4	4	24	96	-0.2417	0.2417
6	5	5	22	110	0.7583	0.7583
7	6	6	16	96	1.7583	1.7583
8	7	7	9	63	2.7583	2.7583
9	8	8	5	40	3.7583	3.7583

Fig. 7.18

Step 3: We now type "=C2*F2" in Cell G2 to determine the multiplication of f and $|x-\bar{x}|$ as shown in Fig. 7.19a. We drag Cell G2 down up to Cell G9 as shown in Fig. 7.19b.

(a)

G2 =C2*F2

	C	D	E	F	G	H
1	No. of Families (f)	fx	$(x-\bar{x})$	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$	
2	6	6	-3.2417	3.2417	19.4500	
3	16	32	-2.2417	2.2417		
4	22	66	-1.2417	1.2417		
5	24	96	-0.2417	0.2417		
6	22	110	0.7583	0.7583		

DRAG IT DOWN

(b)

	C	D	E	F	G	H
1	No. of Families (f)	fx	$(x-\bar{x})$	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$	
2	6	6	-3.2417	3.2417	19.4500	
3	16	32	-2.2417	2.2417	35.8667	
4	22	66	-1.2417	1.2417	27.3167	
5	24	96	-0.2417	0.2417	5.8000	
6	22	110	0.7583	0.7583	16.6833	
7	16	96	1.7583	1.7583	28.1333	
8	9	63	2.7583	2.7583	24.8250	
9	5	40	3.7583	3.7583	18.7917	

Fig. 7.19

Step 4: In Cell G10, we compute $\sum f|x-\bar{x}|$ using "=Sum(G2:G9)" function as explained in Step 7 of Sec. 2.3 of Lab Session 2 (see Fig. 7.20).

G10 =SUM(G2:G9)

	D	E	F	G	H
1	fx	$(x-\bar{x})$	$\lvert x-\bar{x}\rvert$	$f\lvert x-\bar{x}\rvert$	
2	6	-3.2417	3.2417	19.4500	
3	32	-2.2417	2.2417	35.8667	
4	66	-1.2417	1.2417	27.3167	
5	96	-0.2417	0.2417	5.8000	
6	110	0.7583	0.7583	16.6833	
7	96	1.7583	1.7583	28.1333	
8	63	2.7583	2.7583	24.8250	
9	40	3.7583	3.7583	18.7917	
10	509		$\sum f\lvert x-\bar{x}\rvert$	176.8667	
11					

Fig. 7.20

Step 5: We use equation (7) to compute the mean deviation about the mean, type "=G10/C10" in Cell B12 (Fig. 7.21a) and press ***Enter*** to get the result shown in Fig. 7.21b.

MINVERSE =G10/C10

	A	B	C	D	E	F	G
9	8	8	5	40	3.7583	3.7583	18.7917
10	Total	36	120	509		$\sum f\lvert x-\bar{x}\rvert$	176.8667
11	$\bar{x}$	4.2417					
12	MD	=G10/C10					
13							

(a)

ENTER

	A	B	C	D	E	F	G
9	8	8	5	40	3.7583	3.7583	18.7917
10	Total	36	120	509		$\sum f\lvert x-\bar{x}\rvert$	176.8667
11	$\bar{x}$	4.2417					
12	MD	1.4739					
13							

(b)

Fig. 7.21

We have explained the computation of mean deviation about the mean for discrete frequency distribution. You can now try to compute it for continuous frequency distribution and also about other arbitrary values such as median and mode before going to the next section.

Activity 3

1) Consider the continuous frequency distribution of Sec. 2.5 of Lab Session 2 and compute mean deviation about the mean in the same way as explained in Sec. 7.5.2.

2) Compute the median and mode as explained in Lab Session 6 for Problems 1 and 2 and also determine the mean deviation about the median and mode.

7.6 VARIANCE, STANDARD DEVIATION AND COEFFICIENT OF VARIATION

In Unit 2 of MST-002, you have learnt that variance and standard deviation are the most popular measures of dispersion. Here we briefly mention the main points for calculating these measures as follows:

For Raw Data

- Variance is the average of the square of deviations of the values from mean. If we have n observations $x_1, x_2, ..., x_n$, variance can be calculated as

$$\text{Var}(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \ldots (8)$$

- Standard deviation (SD) is defined as the positive square root of the variance and is given by

$$\text{SD} = \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \qquad \ldots (9)$$

For Frequency Distributions

- If we have k observations (or mid values in case of continuous frequency distribution) $x_1, x_2, ..., x_k$ with corresponding frequencies $f_1, f_2, \ldots, f_k$, respectively, the variance can be computed as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2 \quad \text{where } N = \sum_{i=1}^{k} f_i. \qquad \ldots (10)$$

- The formula for the standard deviation is given by

$$\text{SD} = \sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{N}} \qquad \ldots (11)$$

Coefficient of Variation

- We use coefficient of variation as it is a relative measure of variability to compare two data series. We define the coefficient of variation (CV) as

$$\text{CV} = \frac{\sigma}{\bar{x}} \times 100 \qquad \ldots (12)$$

Here we describe the computation of variance and standard deviation in Excel 2007 for raw data, discrete and continuous frequency distributions. We first consider Problem 2 and explain how to compute variance and standard deviation for raw data in MS Excel 2007.

7.6.1 For Raw Data

In order to compute the variance and standard deviation for raw data, we follow the steps given ahead:

Step 1: We enter the data of Table 3 of Lab Session 2 in an Excel spreadsheet. To compute the variance for the given data, we refer to Fig. 7.22 and

1. select Cell B152,
2. click on the ***Formulas*** tab, and
3. click on ***More Functions*** → ***Statistical*** → ***Varp***.

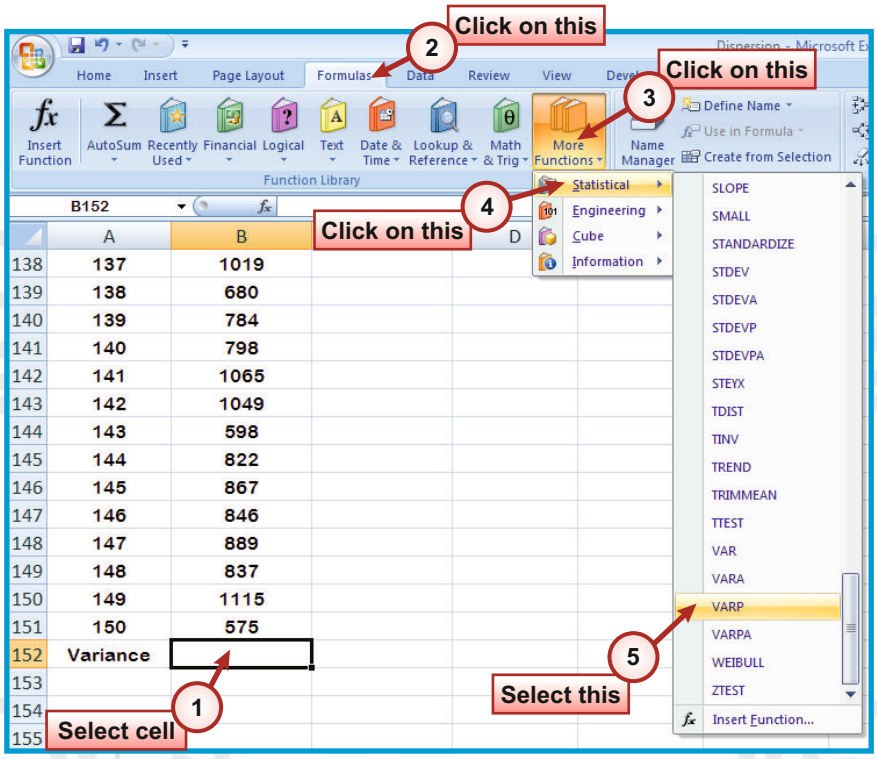
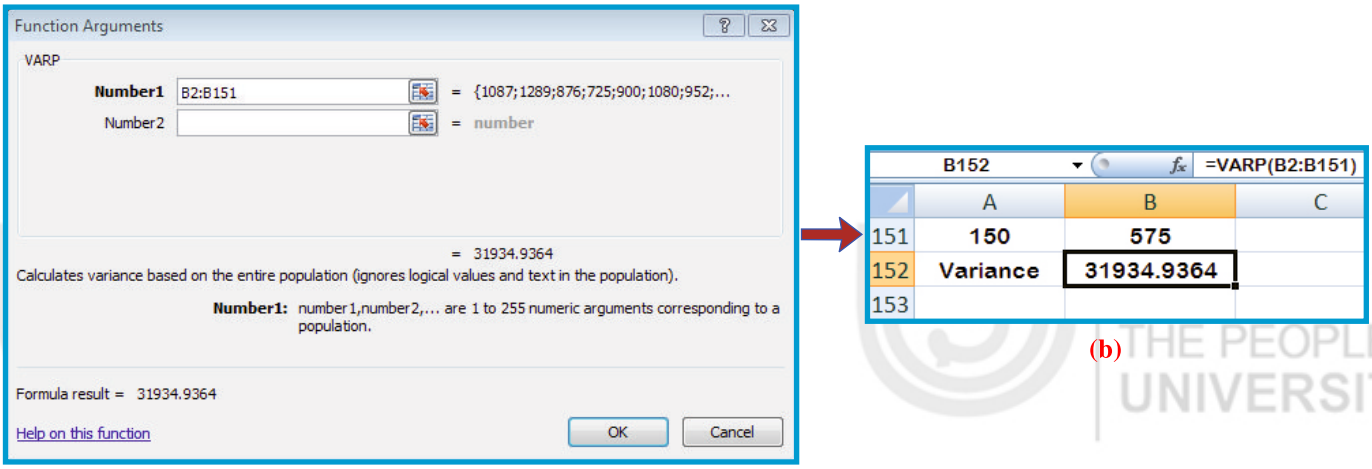


Fig. 7.22

Here we are using n in the denominator to calculate variance, so we use ***Varp*** function in Excel. If we have (n – 1) in the denominator for a sample as discussed in MST-004, we use ***Var*** function in Excel.

Step 2 : A new dialog box appears. We select Cells B2:B151and click on ***OK*** as shown in Fig. 7.23a. We obtain the value of variance in Cell B152 as shown in Fig. 7.23b.



(a) (b)

Fig. 7.23

Step 3: We can compute the standard deviation of the given data directly by taking the square root of variance. For this, we refer to Fig. 7.24 and

1. select Cell B153,
2. click on the ***Formulas*** tab, and
3. click on the ***Math & Trig*** → ***Sqrt***.

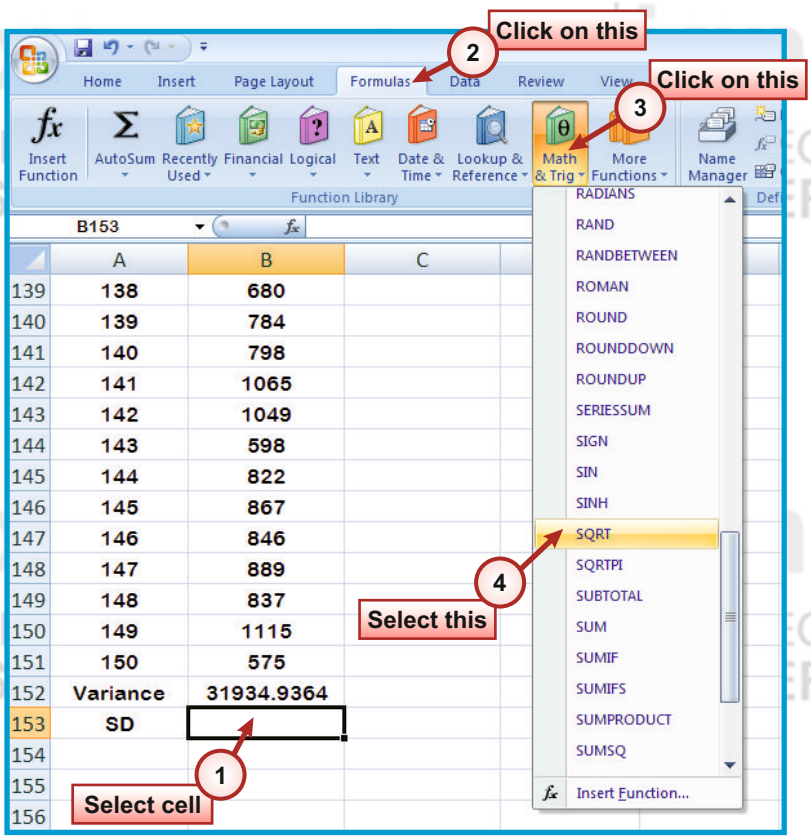


Fig. 7.24

Step 4: A new dialog box opens. We select Cells B152 and click on ***OK*** as shown in Fig. 7.25a. The value of the standard deviation is shown in Fig. 7.25b.

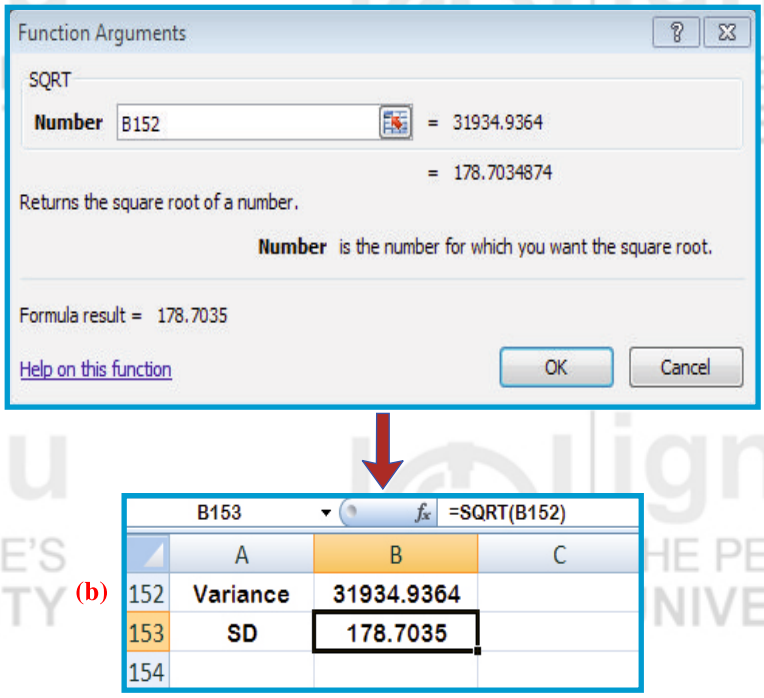


Fig. 7.25

Step 5: We can also directly compute the standard deviation using ***Stdevp*** function as shown in Fig. 7.26. For this, we

1. select Cells B154,
2. click on the ***Formulas*** tab, and
3. click on ***More Functions*** → ***Statistical*** → ***Stdevp***.

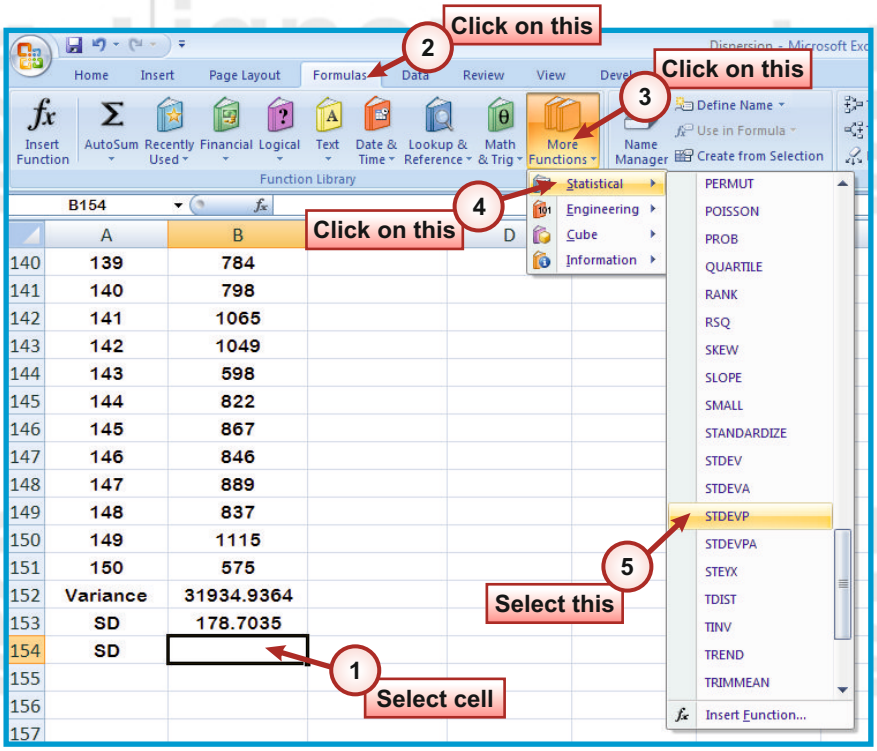
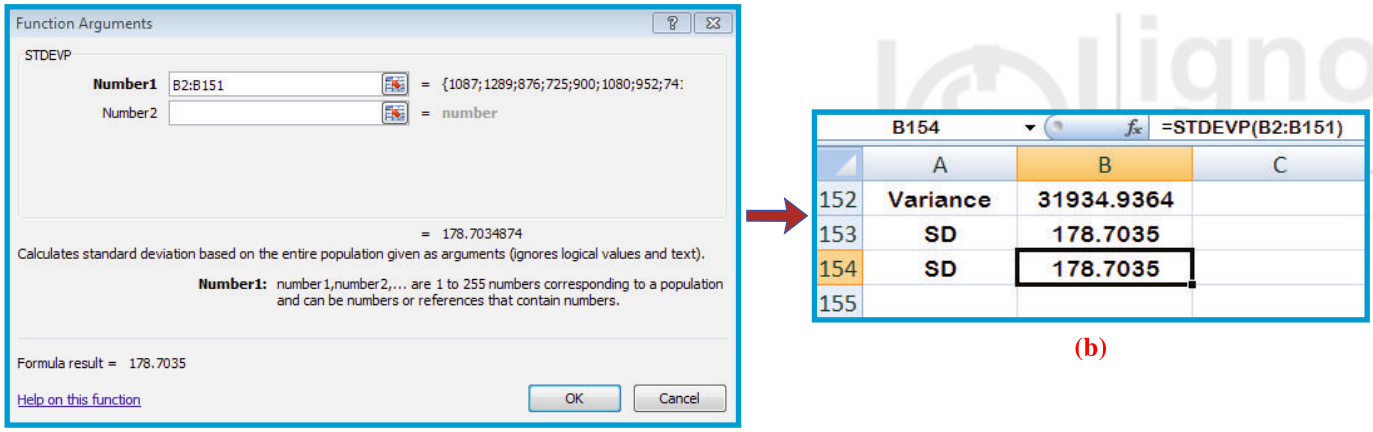


Fig. 7.26

It should be noted that sum of squares of deviations is least when deviations are measured from the mean. This means $\Sigma(x_i - A)^2$ is least when A = Mean.

Step 6: A new dialog box appears. We select Cells B2:B151and click on ***OK*** as shown in Fig. 7.27a. We get the standard deviation in Cell B154 as shown in Fig. 7.27b.



(a) (b)

Fig. 7.27

Note from Fig. 7.27b that the value of standard deviation computed manually in Cell B153 is the same as that computed using ***Stdevp*** function in Cell B154.

The values of SD, MD and variance cannot be negative.

Step 7: To determine the coefficient of variation of the given data, we need to compute the arithmetic mean as explained in Sec.6.3 of Lab session 6. We calculate the arithmetic mean in Cells B155 as shown in Fig. 7.28.

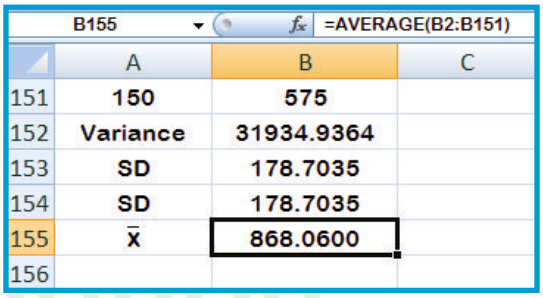


Fig. 7.28

Step 8: To determine the coefficient of variation of the given data, we use equation (12), type "=(B154/B155)*100" in Cell B156 (Fig. 7.29a) and press ***Enter*** as shown in Fig. 7.29b.

(a)

MINVERSE =(B154/B155)*100

	A	B	C
153	SD	178.7035	
154	SD	178.7035	
155	$\bar{x}$	868.0600	
156	CV	=(B154/B155)*100	
157			

ENTER

(b)

B156 =(B154/B155)*100

	A	B	C
154	SD	178.7035	
155	$\bar{x}$	868.0600	
156	CV	21	
157			

Fig. 7.29

We now consider the computation of variance and standard deviation for a discrete frequency distribution.

7.6.2 For Discrete Frequency Distribution

We consider Problem 1 and explain the variance and standard deviation for a discrete frequency distribution in MS Excel 2007. Note that Excel does not have any built-in function to compute variance and standard deviation for frequency distributions. In order to compute variance and standard deviation for discrete frequency distribution, we follow the steps given below:

Step 1: To compute variance or standard deviation for Problem 1, we first calculate the mean of the data by repeating Steps 1 to 6 of Sec. 6.3.2 of Lab Session 6. The output is shown in Fig. 7.30.

	A	B	C	D	E
1	S.No.	Number of Mobile Phones (x)	No. of Families (f)	fx	
2	1	1	6	6	
3	2	2	16	32	
4	3	3	22	66	
5	4	4	24	96	
6	5	5	22	110	
7	6	6	16	96	
8	7	7	9	63	
9	8	8	5	40	
10	Total	36	120	509	
11	$\bar{x}$	4.2417			

Fig. 7.30

The formula with dollar sign ($) is used for an absolute reference.

Step 2: We compute the square of the deviation in the values from the mean, i.e., $(x-\bar{x})^2$ by typing "=(B2-$B$11)^2" in Cell E2 and pressing ***Enter*** (Fig. 7.31a). Note that here we are using absolute reference, i.e., $B$11 to fix Cell B11 as it contains the value of the arithmetic mean.

STDEVP | =(B2-$B$11)^2

	B	C	D	E	F
1	Number of Mobile Phones (x)	No. of Families (f)	fx	$(x-\bar{x})^2$	
2	1	6	6	=(B2-$B$11)^2	
3	2	16	32		

(a)

ENTER

	B	C	D	E	F
1	Number of Mobile Phones (x)	No. of Families (f)	fx	$(x-\bar{x})^2$	
2	1	6	6	10.5084	
3	2	16	32		

(b)

Fig. 7.31

Step 3: We drag Cell E2 down up to Cell E9 to determine the square of the deviation from the mean for the remaining values as shown in Fig. 7.32.

	B	C	D	E	F
1	Number of Mobile Phones (x)	No. of Families (f)	fx	$(x-\bar{x})^2$	
2	1	6	6	10.5084	
3	2	16	32	5.0251	
4	3	22	66	1.5417	
5	4	24	96	0.0584	
6	5	22	110	0.5751	
7	6	16	96	3.0917	
8	7	9	63	7.6084	
9	8	5	40	14.1251	

Fig. 7.32

Step 4: We now type "=C2*E2" in Cell F2 to determine the product of f and $(x-\bar{x})^2$ as shown in Fig. 7.33a and drag Cell F2 down up to Cell F9 as shown in Fig. 7.33b.

F2 | =C2*E2

	C	D	E	F	G
1	No. of Families (f)	fx	$(x-\bar{x})^2$	$f(x-\bar{x})^2$	
2	6	6	10.5084	63.0504	
3	16	32	5.0251		
4	22	66	1.5417		
5	24	96	0.0584		

(a)

DRAG IT DOWN

	C	D	E	F	G
1	No. of Families (f)	fx	$(x-\bar{x})^2$	$f(x-\bar{x})^2$	
2	6	6	10.5084	63.0504	
3	16	32	5.0251	80.4011	
4	22	66	1.5417	33.9182	
5	24	96	0.0584	1.4017	
6	22	110	0.5751	12.6515	
7	16	96	3.0917	49.4678	
8	9	63	7.6084	68.4756	
9	5	40	14.1251	70.6253	

(b)

Fig. 7.33

Step 5: In Cell F10, we compute $\sum f(x-\bar{x})^2$ using "=Sum(F2:F9)" function as explained in Step 7 in Sec. 2.3 of Lab Session 2 (see Fig. 7.34).

F10 =SUM(F2:F9)

	E	F	G
9	14.1251	70.6253	
10	$\sum f(x-\bar{x})^2$	379.9917	
11			

Fig. 7.34

Step 6: We use equation (10) to compute the variance, type "=F10/C10" in Cell B12 (Fig. 7.35a) and press ***Enter*** to get the result (Fig. 7.35b).

(a)

MINVERSE =F10/C10

	A	B	C	D	E	F	G
9	8	8	5	40	14.1251	70.6253	
10	Total	36	120	509	$\sum f(x-\bar{x})^2$	379.9917	
11	$\bar{x}$	4.2417					
12	Variance	=F10/C10					
13							

ENTER

(b)

B12 =F10/C10

	A	B	C
9	8	8	5
10	Total	36	120
11	$\bar{x}$	4.2417	
12	Variance	3.1666	
13			

Fig. 7.35

Step 7: We can compute the standard deviation of the given data directly by taking square root of the variance (see Fig. 7.36) as explained in Steps 3 and 4 of Sec. 7.6.1.

B13 =SQRT(B12)

	A	B	C
10	Total	36	120
11	$\bar{x}$	4.2417	
12	Variance	3.1666	
13	SD	1.7795	
14			

Fig. 7.36

Step 8: To determine the coefficient of variation of the given data, we type "=(B13/B11)*100" in Cell B14 as shown in Fig. 7.37.

B14 =(B13/B11)*100

	A	B	C
11	$\bar{x}$	4.2417	
12	Variance	3.1666	
13	SD	1.7795	
14	CV	42	
15			

Fig. 7.37

You have learnt how to compute the variance, standard deviation and coefficient of variation for raw data as well as discrete frequency distributions. You should try to compute the variance, standard deviation and coefficient of variation for continuous frequency distributions before going to the next section.

Activity 4

Compute the variance, standard deviation and coefficient of variation for the continuous frequency distributions formed in Sec. 2.5 of Lab Session 2 with the help of MS Excel 2007.

7.7 MOMENTS, SKEWNESS AND KURTOSIS

In Units 3 and 4 of MST-002, you have learnt the computation of moments, skewness and kurtosis. Here we briefly mention the main points as follows:

Moments about the origin

- If we have n observations $x_1, x_2, ..., x_n$, we define the rth order moment about the origin (also known as raw moments) for ungrouped or raw data as

$$\mu'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r \qquad \ldots (13)$$

- If we have k observations (or mid values in case of class intervals) $x_1, x_2, ..., x_k$, with corresponding frequencies $f_1, f_2, ..., f_k$, respectively, the rth order moment about the origin for discrete or continuous frequency distributions is given by

$$\mu'_r = \frac{1}{k}\sum_{i=1}^{n} f_i x_i^r \qquad \ldots (14)$$

Moments about the mean

- The first four moments about the mean are also known as central moments and can be obtained with the help of the moments about the origin as follows :

$$\mu = 0 \qquad \ldots (15a)$$

$$\mu_2 = \mu'_2 - \mu'^2_1 \qquad \ldots (15b)$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 \qquad \ldots (15c)$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1 \qquad \ldots (15d)$$

Coefficient of Skewness

- Karl Pearson defined two coefficients of skewness, β_1 and γ_1, based upon the central moments. These are given as follows:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad \ldots (16)$$

Since β_1 does not give the direction of skewness, i.e., whether it is positive or negative, another measure of skewness (γ_1) is defined as

$$\gamma_1 = \pm\sqrt{\beta_1} \qquad \ldots (17)$$

The sign of skewness would depend upon whether μ_3 is positive or negative.

- **Description:**
  1. If $\gamma_1 = 0$, the distribution is symmetric.
  2. If $\gamma_1 < 0$, the distribution is negatively skewed.
  3. If $\gamma_1 > 0$, the distribution is positively skewed.

Coefficient of Kurtosis

- Karl Pearson also defined two coefficients of kurtosis β_2 and γ_2 based upon the central moments:

  β_2 measures kurtosis and is defined as

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \qquad \ldots (18)$$

- The other measure of kurtosis (γ_2) is defined as

$$\gamma_2 = \beta_2 - 3 \qquad \ldots (19)$$

- **Description:**

  If $\beta_2 = 3$ or $\gamma_2 = 0$, the distribution is mesokurtic.

  If $\beta_2 < 3$ or $\gamma_2 < 0$, the distribution is platykurtic.

  If $\beta_2 > 3$ or $\gamma_2 > 0$, the distribution is leptokurtic.

Here we describe the computation of moments, skewness and kurtosis using Excel 2007 for raw data, discrete and continuous frequency distributions.

7.7.1 For Raw Data

We consider Problem 2 and explain the computation of moments, skewness and kurtosis for raw data using MS Excel 2007. In order to compute the moments, skewness and kurtosis, we follow the steps given below:

Step 1: We enter the data of Table 3 of Lab Session 2 in an Excel spreadsheet. We type "=B2^2" in Cell C2 (Fig. 7.38a), "=B2^3" in Cell D2 (Fig. 7.38b) and "=B2^4" in Cell E2 (Fig. 7.38c) to compute the values of x^2, x^3 and x^4, respectively, for the first sample.

(a)

C2 | =B2^2

	B	C	D
1	Life of Bulb (in hours) (x)	x^2	
2	1087	1181569	
3	1289		

(b)

D2 | =B2^3

	C	D	E
1	x^2	x^3	
2	1181569	1284365503	
3			

(c)

E2 | =B2^4

	D	E	F
1	x^3	x^4	
2	1284365503	1396105301761	
3			

Fig. 7.38

Step 2: We select Cell C2:E2 as shown in Fig. 7.39a and drag down these cells up to Row 151 to determine the values of x^2, x^3 and x^4 for the remaining samples as shown in Fig. 7.39b.

	B	C	D	E	F
1	Life of Bulb (in hours) (x)	x^2	x^3	x^4	
2	1087	1181569	1284365503	1396105301761	
3	1289				
4	876				

(a)

DRAG THEM DOWN

	B	C	D	E	F
1	Life of Bulb (in hours) (x)	x^2	x^3	x^4	
2	1087	1181569	1284365503	1396105301761	
3	1289	1661521	2141700569	2760652033441	
4	876	767376	672221376	588865925376	
5	725	525625	381078125	276281640625	
6	900	810000	729000000	656100000000	
7	1080	1166400	1259712000	1360488960000	
8	952	906304	862801408	821386940416	
9	741	549081	406869021	301489944561	
10	1000	1000000	1000000000	1000000000000	
11	900	810000	729000000	656100000000	
12	957	915849	876467493	838779390801	
13	745	555025	413493625	308052750625	
14	926	857476	794022776	735265090576	

(b)

Fig. 7.39

Step 3: To get the first four moments about the origin (μ'_1, μ'_2, μ'_3 and μ'_4) for the raw data, we use equation (13) and compute the average of the series x, x^2, x^3 and x^4 given in Cells B2:B151, Cells C2:C151, Cells D2:D151 and Cells E2:E151, respectively, using ***Average*** function in Cells B153 to B156, as explained in Sec. 6.3 of Lab Session 6 (Fig. 7.40).

B153 =AVERAGE(B2:B151)

	A	B	C
152			
153	μ'_1	868.0600	
154			

(a)

B154 =AVERAGE(C2:C151)

	A	B	C
153	μ'_1	868.0600	
154	μ'_2	785463.1000	
155			

(b)

B155 =AVERAGE(D2:D151)

	A	B	C
154	μ'_2	785463.1000	
155	μ'_3	738499562.6200	
156			

(c)

B156 =AVERAGE(E2:E151)

	A	B	C
155	μ'_3	738499562.6200	
156	μ'_4	719029405462.1400	
157			

(d)

Fig. 7.40

Step 4: You have studied that the first moment about the mean is always zero [equation (15a)]. So we type "0" in Cell B157 (Fig. 7.41).

B157 0

	A	B	C
156	μ'_4	719029405462.1400	
157	μ_1	0	
158			

Fig. 7.41

Step 5: We use equation (15b) and type "=B154-B153^2" in Cell B158 to determine the second moment about the mean as shown in Fig. 7.42.

B158 f_x =B154-B153^2

	A	B	C
157	μ_1	0	
158	μ_2	31934.9364	
159			

Fig. 7.42

Step 6: We use equation (15c) and type "=B155-3*B154*B153+2*B153^3" in Cell B159 to determine the third moment about the mean as shown in Fig. 7.43.

B159 f_x =B155-3*B154*B153+2*B153^3

	A	B	C
158	μ_2	31934.9364	
159	μ_3	1227582.2512	
160			

Fig. 7.43

Step 7: We use equation (15d) and type "=B156-4*B155*B153+6*B154*B153^2-3*B153^4" in Cell B160 to determine the fourth moment about mean (Fig. 7.44).

B160 f_x =B156-4*B155*B153+6*B154*B153^2-3*B153^4

	A	B	C
159	μ_3	1227582.2512	
160	μ_4	2579008046.6851	
161			

Fig. 7.44

Step 8: We use equations (16) and (17) and type "=(B159^2)/(B158^3)" and "=Sqrt(B161)" in Cells B161 and B162 to determine the values of β_1 and γ_1, respectively, as shown in Figs. 7.45a and b.

B161 f_x =(B159^2)/(B158^3)

	A	B	C
160	μ_4	2579008046.6851	
161	β_1	0.0463	
162			

(a)

B162 f_x =SQRT(B161)

	A	B	C
161	β_1	0.0463	
162	γ_1	0.2151	
163			

(b)

Fig. 7.45

Notice from Fig. 7.45 that the value of γ_1 is 0.2151 and so the data for the life of bulbs are slightly positively skewed.

Step 9: We use equations (18) and (19) and type "=(B160)/(B158^2)" and "=B163-3" in Cells B163 and B164 to determine the values of β_2 and γ_2, respectively, as shown in Figs. 7.46a and b.

B163 f_x =(B160)/(B158^2)

	A	B	C
162	γ_1	0.2151	
163	β_2	2.5288	
164			

(a)

B164 f_x =B163-3

	A	B	C
163	β_2	2.5288	
164	γ_2	-0.4712	
165			

(b)

Fig. 7.46

Notice from Fig. 7.46b that the value of γ_2 is – 0.4712 and so the data for the life of bulbs are platykurtic.

We now consider the computation of moments, skewness and kurtosis for discrete frequency distribution.

7.7.2 For Discrete Frequency Distribution

We consider Problem 1 and explain how to compute the moments, skewness and kurtosis for discrete frequency distribution using MS Excel 2007. In order to compute the moments, skewness and kurtosis, we follow the steps given below:

Step 1: We consider the discrete frequency distribution formed in Sec. 2.4 of Lab Session 2. We type "=C2*B2" in Cell D2 (Fig. 7.47a), "=C2*(B2^2)" in Cell E2 (Fig. 7.47b), "=C2*(B2^3)" in Cell F2 (Fig. 7.47c) and "=C2*(B2^4)" in Cell G2 (Fig. 7.47d) to compute the values of fx, fx^2, fx^3 and fx^4, respectively, for the first sample.

D2 =C2*B2

	C	D	E
1	No. of Families (f)	fx	
2	6	6	
3	16		

(a)

E2 =C2*(B2^2)

	D	E	F
1	fx	fx^2	
2	6	6	
3			

(b)

F2 =C2*(B2^3)

	E	F	G
1	fx^2	fx^3	
2	6	6	
3			

(c)

G2 =C2*(B2^4)

	F	G	H
1	fx^3	fx^4	
2	6	6	
3			

(d)

Fig. 7.47

Step 2: We select Cell D2:G2 as shown in Fig. 7.48a and drag down these cells up to Row 9 to determine the values of fx, fx^2, fx^3 and fx^4 for the remaining samples as shown in Fig. 7.48b.

	C	D	E	F	G	H
1	No. of Families (f)	fx	fx^2	fx^3	fx^4	
2	6	6	6	6	6	
3	16					

(a)

DRAG THEM DOWN

	C	D	E	F	G	H
1	No. of Families (f)	fx	fx^2	fx^3	fx^4	
2	6	6	6	6	6	
3	16	32	64	128	256	
4	22	66	198	594	1782	
5	24	96	384	1536	6144	
6	22	110	550	2750	13750	
7	16	96	576	3456	20736	
8	9	63	441	3087	21609	
9	5	40	320	2560	20480	
10						

(b)

Fig. 7.48

Step 3: We determine the values of $\sum f$ in Cells C10 using the ***Sum*** function as shown in Fig. 7.49a and drag right Cell C10 up to Cell G10 to compute the values of $\sum fx$, $\sum fx^2$, $\sum fx^3$ and $\sum fx^4$, respectively, as shown in Fig. 7.49b.

(a)

C10 | =SUM(C2:C9)

	B	C	D
9	8	5	40
10	Total	120	
11			

DRAG IT RIGHT

(b)

C10 | =SUM(C2:C9)

	B	C	D	E	F	G	H
9	8	5	40	320	2560	20480	
10	Total	120	509	2539	14117	84763	
11							

Fig. 7.49

Step 4: We use equation (14) for r = 1 and type "=D10/C10" in Cell B12 to determine the first moment about the origin (Fig. 7.50).

B12 | =D10/C10

	A	B	C
11			
12	μ_1'	4.2417	
13			

Fig. 7.50

Step 5: We use equation (14) for r = 2 and type "=E10/C10" in Cell B13 to determine the second moment about the origin (Fig. 7.51).

B13 | =E10/C10

	A	B	C
12	μ_1'	4.2417	
13	μ_2'	21.1583	
14			

Fig. 7.51

Step 6: We use equation (14) for r = 3 and type "=F10/C10" in Cell B14 to determine the third moment about the origin (Fig. 7.52).

B14 | =F10/C10

	A	B	C
13	μ_2'	21.1583	
14	μ_3'	117.6417	
15			

Fig. 7.52

Step 7: We use equation (14) for r = 4 and type "=G10/C10" in Cell B15 to determine the fourth moment about the origin (Fig. 7.53).

B15 | =G10/C10

	A	B	C
14	μ_3'	117.6417	
15	μ_4'	706.3583	
16			

Fig. 7.53

Step 8: We compute the first four moments about the mean, i.e., μ_1, μ_2, μ_3 and μ_4 in the same way as explained in Steps 4 to 7 of Sec. 7.7.1.

We also determine the values of β_1, γ_1 , β_2, and γ_2 as explained in Steps 8 and 9 of Sec.7.7.1 (see Fig. 7.54).

	A	B	C
16	μ_1	0.0000	
17	μ_2	3.1666	
18	μ_3	1.0318	
19	μ_4	23.3146	
20	β_1	0.0335	
21	γ_1	2.3251	
22	β_2	0.1831	
23	γ_2	-0.6749	
24			

Fig. 7.54

Notice from Fig. 7.54 that the value of γ_1 is 2.3251 and so the data of mobile phones are positively skewed. Also the value of γ_2 is −0.6749 and so the data of mobile phones are platykurtic.

You have learnt how to compute the moments, skewness and kurtosis for raw data as well as discrete frequency distributions. You should try to compute the moments, skewness and kurtosis for continuous frequency distributions before going to the next section.

Activity 5

Compute the first four moments about the origin, moments about the mean, β and γ coefficients of skewness and kurtosis for the continuous frequency distributions formed in Sec. 2.5 of Lab Session 2 with the help of MS Excel 2007.

7.8 DESCRIPTIVE STATISTICS USING DATA ANALYSIS TOOLPAK

We can also use the ***Data Analysis ToolPak*** for determining descriptive statistics in Excel. Here we consider Problem 2, i.e., data of Table 3 of Lab Session 2 to explain the procedure. We follow the steps given below:

Step 1: We enter the data of Table 3 of Lab Session 3 in an Excel spreadsheet. To compute descriptive statistics using ***Data Analysis ToolPak*** in Excel 2007, we

1. click on the ***Data*** tab,
2. select ***Data Analysis*** as shown in Fig. 7.55a,
3. select the ***Descriptive Statistics*** option,
4. click on ***OK*** as shown in Fig. 7.55b, and
5. get a new dialog box as shown in Fig. 7.55c.

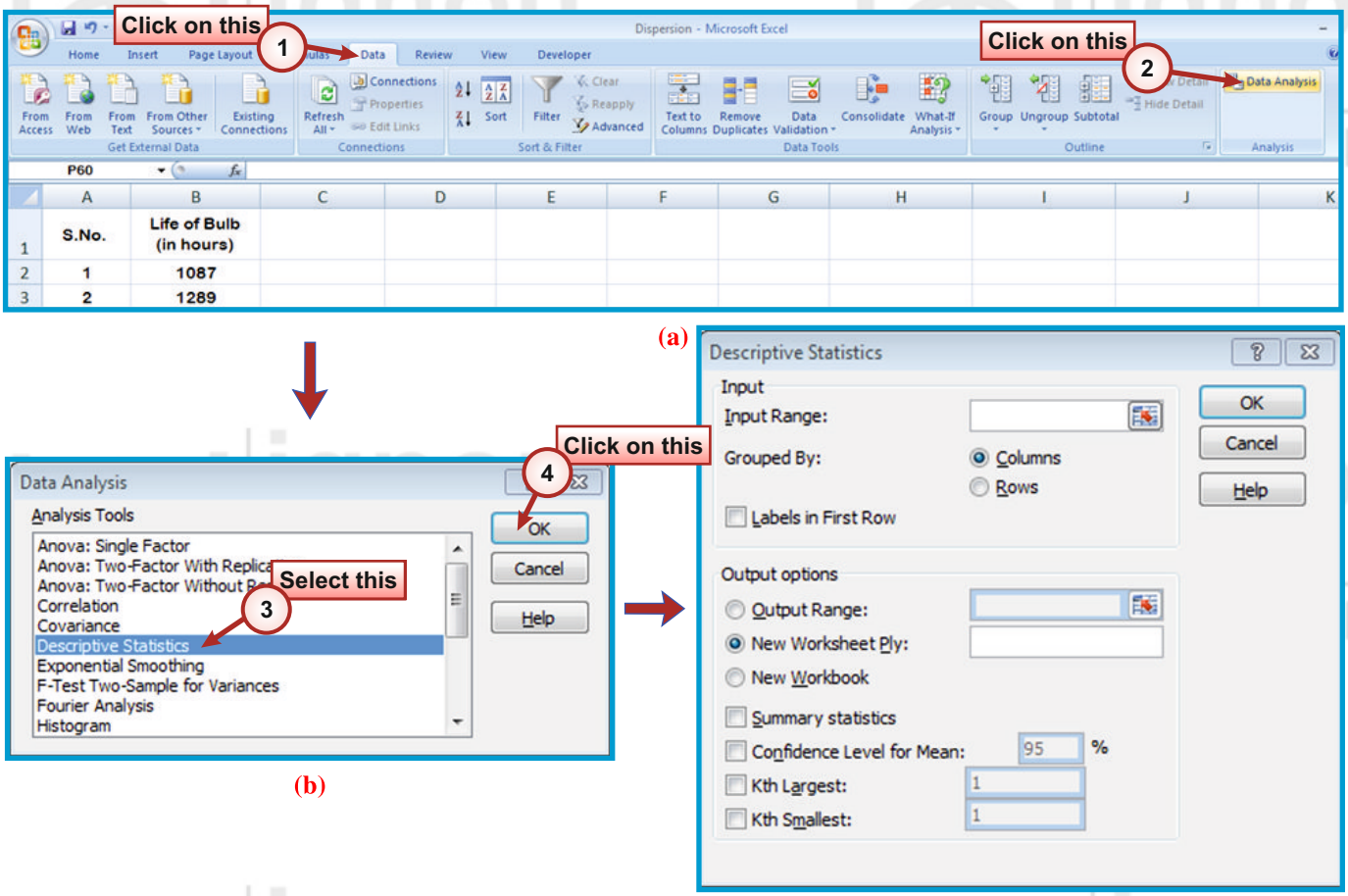


Fig. 7.55

Step 2: As shown in Fig. 7.56, we

1. specify the data with the label **Life of Bulb** in ***Input Range***, i.e., Cells B1:B151,
2. tick on the ***Labels*** box since we have included data labels in the first row, i.e., **Life of Bulb (in hours)** given in Cell B1 in the input range,

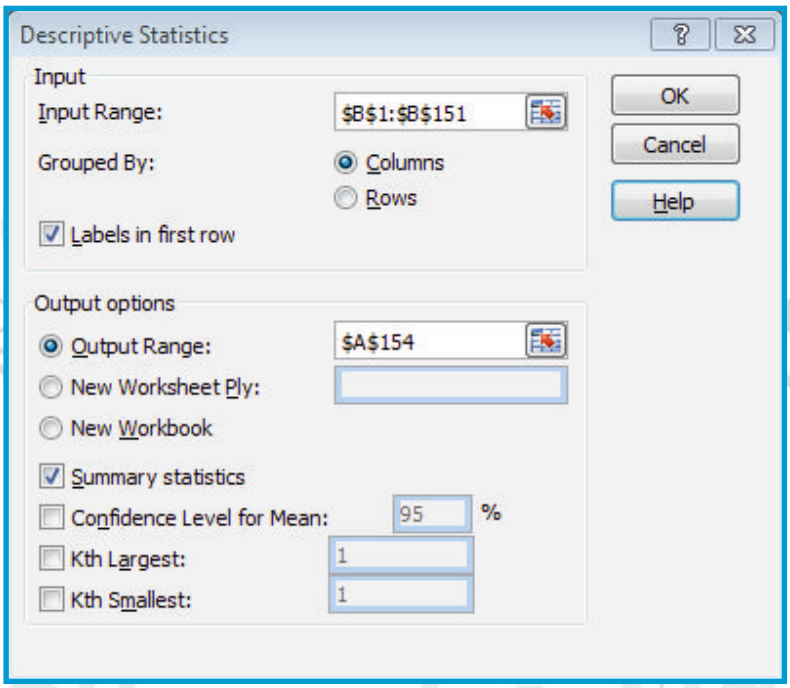


Fig. 7.56

3. provide the ***Output Range*** under the ***Output Options***. Here we select Cell A154 for the output, and
4. tick on the ***Summary statistics*** box and then click on ***OK***.

Step 3: After completing Step 2, we obtain the output shown in Fig. 7.57.

	A	B	C
153			
154	*Life of Bulb (in hours)*		
155			
156	Mean	868.06	
157	Standard Error	14.6400	
158	Median	860.5000	
159	Mode	754	
160	Standard Deviation	179.3022	
161	Sample Variance	32149.2648	
162	Kurtosis	-0.4462	
163	Skewness	0.2173	
164	Range	797	
165	Minimum	501	
166	Maximum	1298	
167	Sum	130209	
168	Count	150	
169			

Fig. 7.57

Notice from Fig. 7.57 that the ***Descriptive Statistics*** option under ***Data Analysis ToolPak*** provides the following information:

- ✓ Arithmetic mean in Cell B156, which is the same as shown in Fig. 6.4 of Lab Session 6.
- ✓ Standard error in Cell B157, which you have studied in Unit 1 of MST-004 (Statistical Inference).
- ✓ Median and mode in Cells B158 and B159, respectively.
- ✓ Standard deviation and variance in Cells B160 and B161, respectively. Note that the ***Data Analysis ToolPak*** determines the sample standard deviation and variance, which has (n – 1) in the denominator instead of n, as explained in Unit 1 of MST-004.
- ✓ Kutosis and skewness in Cells B162 and B163, respectively. These values differ from the values shown in Figs. 7.45 and 7.46. This is because Excel as well as major statistical software packages include an adjustment for the sample size and provide the adjusted coefficient of skewness and kurtosis. Since it is beyond the scope of this course, we shall not discuss this aspect.
- ✓ The range, minimum and maximum values of the data, which are given in Cells B164, B165 and B166, respectively.
- ✓ The sum, i.e., total of the given values and count, i.e., the number of observations in the given data in Cells B167 and B168, respectively.

You can now try the following exercises for practice.

Activity 6

Work out the following exercises with the help of MS Excel 2007:

A1) Examples 1 to 10 given in Unit 2 of MST-002.

A2) Exercises E1 to E8 given in Unit 2 of MST-002.

A3) Example 1 given in Unit 3 of MST-002.

A4) Exercises E1 to E3 given in Unit 3 of MST-002.

A5) Exercise E4 given in Unit 4 of MST-002.

Match the results with the manual computation of the data carried out in Units 2, 3 and 4 of MST-002.

Continuous Assessment 7

1. Consider the discrete frequency distribution constructed for the data of Table 5 of Lab Session 2 and compute

   i) the range, quartile deviation and its coefficients,

   ii) the mean deviation, variance, standard deviation and coefficient of variation, and

   iii) moments, skewness and kurtosis.

2. Consider the raw data as well as continuous frequency distribution of Table 6 of Lab Session 2 and compute

   i) the range, quartile deviation and its coefficients,

   ii) the mean deviation, variance, standard deviation and coefficient of variation, and

   iii) moments, skewness and kurtosis.

Home Work: Do It Yourself

1) Follow the steps explained in Secs. 7.3 to 7.8 to compute the measures of dispersion, moments, skewness and kurtosis. Take the screenshots and keep them in your record book.

2) Develop the spreadsheets for the exercises given in "Continuous Assessment 7" as explained in this lab session. Take the screenshots of the final spreadsheets.

3) **Do not forget** to keep these screenshots in your record book as these will contribute to your continuous assessment in the Laboratory.

Correlation Analysis

SESSION 8

Structure

8.1 Introduction
Objectives
8.2 Problem Description
8.3 Simple Correlation Coefficient
8.4 Multiple Correlation Coefficients
8.5 Partial Correlation Coefficients
8.6 Correlation Coefficients using Data Analysis ToolPak
8.7 Rank Correlation Coefficient
8.8 Rank Correlation Coefficient for Tied or Repeated Ranks

8.1 INTRODUCTION

In Lab Sessions 6 and 7, you have learnt how to compute measures of central tendency, measures of dispersion, moments, skewness and kurtosis, which analyse data involving a single variable. But in many situations, we are interested in studying the relationship between two or more variables. In Unit 6 of MST-002 (Descriptive Statistics), we have explained that the coefficient of correlation measures the strength and direction of the linear relationship between two variables. We use the Karl Pearson's correlation coefficient to measure the extent or degree of relationship between two variables. When we have more than two variables, we use the multiple correlation coefficient to determine the joint effect of two or more variables on a single variable. We use partial correlation to obtain the relationship between two variables, eliminating the effects of other variables, as explained in Units 11 and 12 of MST-002.

Prerequisite

- Lab Sessions 6 and 7 of MSTL-001 (Basic Statistics Lab).
- Units 6, 7, 11 and 12 of MST-002 (Descriptive Statistics).

You have also learnt in Unit 7 of MST-002 that the rank correlation is applied to measure the association between two ordinal (or rank) variables.

In this lab session, you will learn how to determine the simple and rank correlation coefficients for two variables using MS Excel 2007. You will also learn how to compute multiple and partial correlation coefficients for more than two variables using Excel 2007.

Objectives

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

- prepare the spreadsheet in MS Excel 2007;
- compute the simple correlation coefficient;
- determine the multiple and partial correlation coefficients; and
- obtain the rank correlation coefficient.

8.2 PROBLEM DESCRIPTION

1. Suppose we are interested in investigating the relationship between the harvested yield of a particular crop (kg/ha), the amount of rainfall (mm) and the amount of fertiliser (kg/ha). The data for 35 years for the three variables are recorded in Table 1.

Table 1: Crop yield data

Yield of a Crop (kg/h)	Rainfall (mm)	Fertiliser (kg/h)
133	304	168
72	247	45
91	270	80
59	70	10
90	264	80
84	112	66
115	167	123
112	257	124
96	117	91
61	131	17
168	297	237
92	281	83
97	130	102
74	143	38
105	176	119
136	158	171
95	201	86
143	266	186
113	193	118
151	283	202
144	228	191
210	378	320
220	259	337
157	284	215
219	233	336
200	249	298
218	275	339
287	287	477
203	232	307
341	503	158
237	393	214
290	305	480
323	418	542
397	501	698
334	497	574

For investigating the relationship between any two or all three of these variables, we need to

i) compute the Pearson's correlation coefficients between (a) the yield of the crop and the amount of rainfall, (b) the yield of the crop and the amount of fertiliser and (c) the amount of rainfall and fertiliser for the given data;

ii) determine the multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$; and

iii) obtain the partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.

2. Twenty five students participated in a science talent contest and were assessed in two parts: A and B. Part A was based on written examination while Part B was based on oral quiz. Scores between 0 and 100 were awarded separately for both parts. The scores of each contestant for both parts are recorded in Table 2.

Table 2: Scores of students in the science talent contest

Part A	51	78	68	58	62	65	96	60	76	66	85	95	91	55	81
	99	61	53	98	73	83	84	94	63	80					
Part B	55	67	69	57	60	62	86	73	68	75	81	100	72	59	79
	65	64	53	97	96	85	76	95	50	83					

For investigating the relationship between the students' performance in both parts for the given data, we need to compute the Spearman's rank correlation coefficient between the scores obtained.

3. A survey was conducted to investigate the impact of the average number of hours spent in watching television per day by students on the marks obtained by them in examinations. A sample of fifteen students was taken. The number of hours spent in watching TV and the marks obtained are recorded in Table 3.

Table 3: Time spent in watching TV and marks obtained

S. No.	Marks	TV Viewing Hours
1	60	6
2	72	1
3	64	2
4	71	2
5	77	3
6	65	2
7	62	1
8	78	2
9	65	4
10	72	2
11	73	1
12	79	1
13	67	3
14	65	3
15	61	5

For investigating the impact of viewing TV on marks obtained, we need to compute the Spearman's rank correlation coefficient between the number of hours spent in watching TV and marks obtained in the examination for the given data.

8.3 SIMPLE CORRELATION COEFFICIENT

You have learnt how to compute Karl Pearson's correlation coefficient in Unit 6 of MST-002. This is also known as simple or product moment correlation coefficient. It is the most widely used statistical method for determining the degree of linear relationship between two variables. Here we briefly mention the main steps as follows:

Step 1: If X and Y are two random variables considering n observations $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, the covariance between X and Y is given by

$$\text{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad \ldots (1)$$

Step 2: We compute the variances of the variables X and Y as

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ and} \qquad \ldots (2)$$

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \ldots (3)$$

Step 3: The correlation coefficient (r) between X and Y is given by

$$r = \frac{\text{Cov}(x, y)}{\sqrt{V(x)V(y)}} \qquad \ldots (4)$$

Note that the correlation coefficient lies between –1 and +1.

We consider Problem 1 and explain the computation of the correlation coefficient between two variables in Excel 2007. In order to compute Karl Pearson's correlation coefficient, we follow the steps given below:

Step 1: We enter the data of Table 1 in an Excel spreadsheet as shown in Fig. 8.1.

	A	B	C
1	Yield of a crop (X_1) (kg/h)	Rainfall (X_2) (mm)	Fertiliser (X_3) (kg/h)
2	133	304	168
3	72	247	45
4	91	270	80
5	59	70	10
6	90	264	80
7	84	112	66
8	115	167	123
9	112	257	124
10	96	117	91

Correlation

Fig. 8.1: Partial screenshot of the given data in Excel spreadsheet.

Step 2: We determine the correlation coefficient between the yield of the crop and the amount of rainfall, and denote it by r_{12}. To determine Karl Pearson's correlation coefficient, we use the built-in function *Correl* of Excel. As shown in Fig. 8.2, we

1. select Cell B38,
2. click on the *Formulas* tab, and

3. click on *More Functions* → *Statistical* → *Correl*.

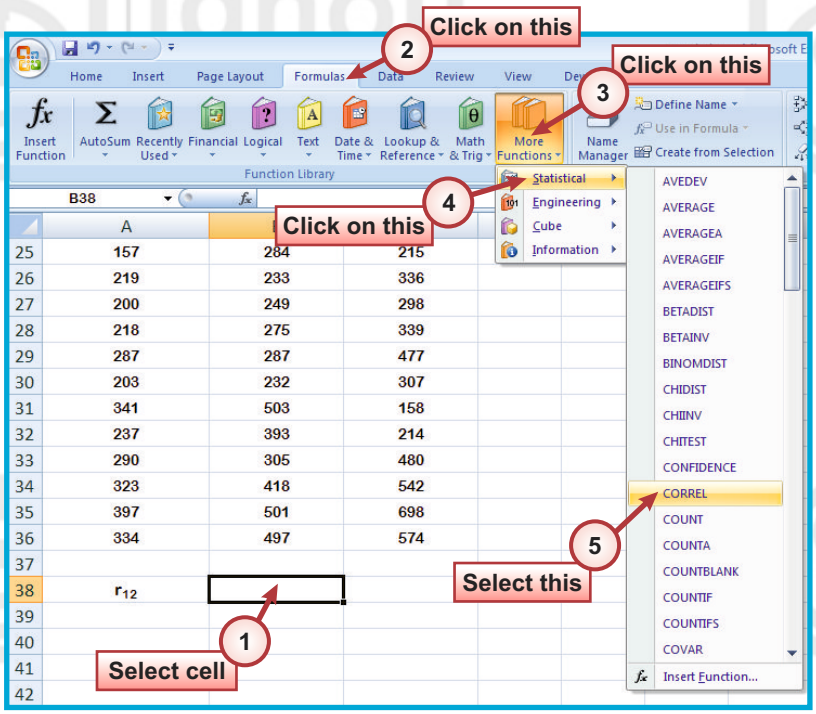
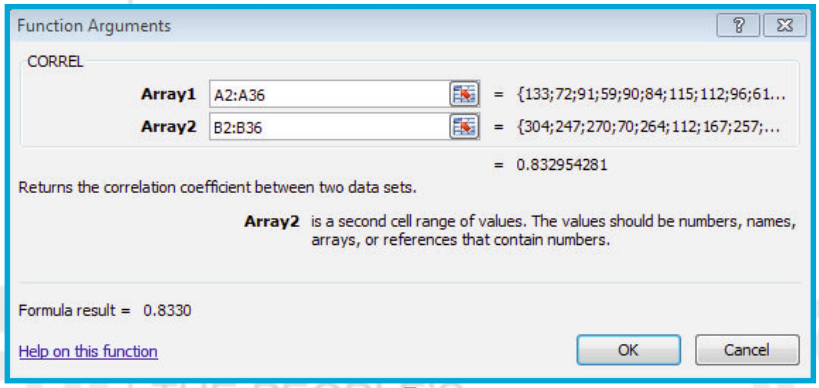


Fig. 8.2

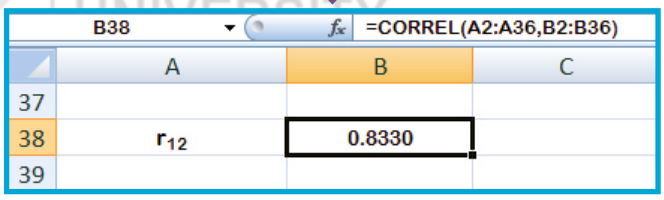
Step 3: A new dialog box opens, which requires the values of the two variables to compute the correlation between them. As shown in Fig. 8.3, we

1. select Cells A2:A36 as *Array 1*,
2. select Cells B2:B36 as *Array 2*,
3. click on *OK* (Fig. 8.3a), and
4. obtain the value of the correlation coefficient in Cell B38 (Fig. 8.3b).



(a)

You can directly type "=Correl(A2 : A36, B2: B36)" in Cell B38 to obtain the correlation coefficient between the yield of the crop and the amount of rainfall.



(b)

Fig. 8.3

Notice from Fig. 8.3b that the correlation coefficient between the yield of the crop and the amount of rainfall is 0.8330. We can conclude that there is high positive correlation between the yield of the crop and the amount of rainfall for the given data.

Step 4: We now compute the correlation coefficients between (i) the yield of the crop and the amount of fertiliser (r_{13}) and (ii) the amounts of fertiliser and rainfall (r_{23}) in Cells B39 and B40 in the same way as explained in Steps 2 and 3. The values of r_{13} and r_{23} are shown in Figs. 8.4a and 8.4b, respectively.

B39 =CORREL(A2:A36,C2:C36)

	A	B	C
38	r_{12}	0.8330	
39	r_{13}	0.9084	
40			

(a)

B40 =CORREL(B2:B36,C2:C36)

	A	B	C
39	r_{13}	0.9084	
40	r_{23}	0.6958	
41			

(b)

Fig. 8.4

Notice from Fig. 8.4a that the correlation coefficient between the yield of the crop and the amount of fertiliser is 0.9084. It means that the crop yield and the amount of fertiliser are highly positively correlated. However, the correlation coefficient between the amount of fertiliser and rainfall is 0 6958 (Fig 8 4b) It means that the amounts of fertiliser and rainfall are moderately positively correlated for the given data.

8.4 MULTIPLE CORRELATION COEFFICIENTS

You have learnt in Unit 11 of MST-002 that multiple correlation determines the combined influence of two or more variables on a single variable. Here we briefly mention the main steps as under:

Step 1: Suppose, X_1, X_2 and X_3 are three variables having n observations or units. We compute the correlation coefficient using equation (4) between (i) X_1 and X_2 denoted by r_{12}, (ii) X_1 and X_3 denoted by r_{13}, and (iii) X_2 and X_3 denoted by r_{23}.

Step 2: The multiple correlation coefficient of X_1 on X_2 and X_3 is the simple correlation coefficient between X_1 and the joint effect of X_2 and X_3. It is denoted by $R_{1.23}$ and given by

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \quad \ldots (5)$$

where r_{12} – correlation coefficient between variables X_1 and X_2,

r_{23} – correlation coefficient between variables X_2 and X_3, and

r_{13} – correlation coefficient between variables X_1 and X_3.

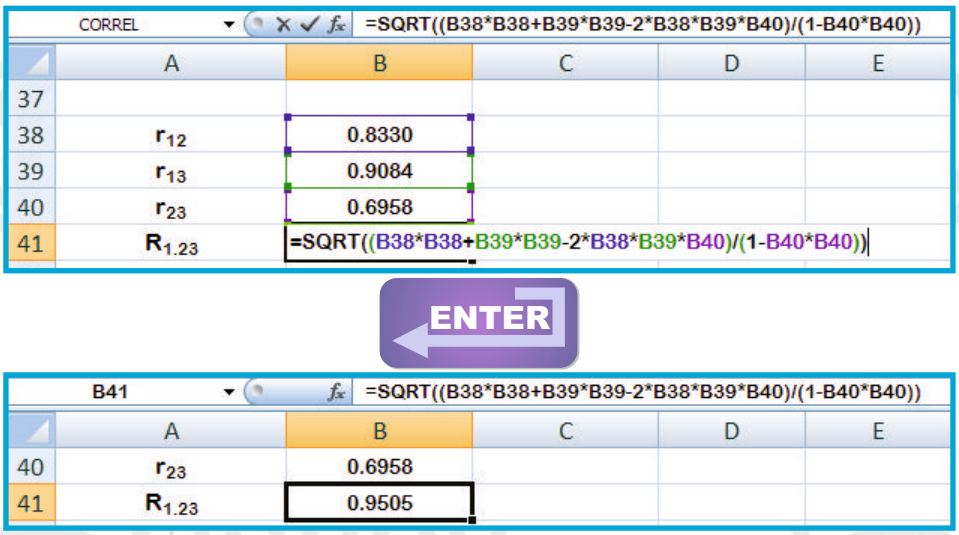
Step 3: The multiple correlation coefficients $R_{2.13}$ and $R_{3.12}$ are given by

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \quad \ldots (6)$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}} \quad \ldots (7)$$

Here we describe how to compute multiple correlation coefficients using Excel 2007. We continue with the same Excel sheet prepared in Sec. 8.3 to compute the multiple correlation coefficients and follow the steps given below:

Step 1: The multiple correlation coefficient of the crop yield (X_1) on rainfall (X_2) and fertiliser (X_3) is the simple correlation coefficient between the crop yield and the joint effect of rainfall and fertiliser. It is denoted by $R_{1.23}$. We use equation (5) to compute the multiple correlation coefficient and type "=Sqrt((B38*B38+B39*B39-2*B38*B39*B40)/(1-B40*B40))" in Cell B41 as shown in Fig. 8.5a. When we press ***Enter***, we get the output shown in Fig. 8.5b.



(a)

(b)

Fig. 8.5

Notice from Fig. 8.5b that the multiple correlation coefficient of the crop yield on the amounts of rainfall and fertiliser is 0.9505. It means that the amounts of rainfall and fertiliser are highly positively correlated with the crop yield for the given data.

Step 2: We follow Step 1 to compute the multiple correlation coefficient of amount of rainfall (X_2) on the crop yield (X_1) and the amount of fertiliser (X_3), i.e., $R_{2.13}$. We use equation (6) and type "=Sqrt((B38*B38+B40* B40-2*B38*B39*B40)/(1-B39*B39))" in Cell B42 as shown in Fig. 8.6.

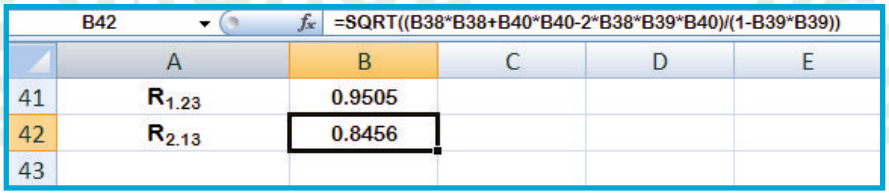


Fig. 8.6

Fig. 8.6 reveals that the multiple correlation coefficient of the amount of rainfall on the crop yield and the amount of fertiliser is 0.8456. It means that the yield of the crop and the amount of fertiliser are highly positively correlated with the amount of rainfall for the given data.

Step 3: We compute the multiple correlation coefficient of the amount of

fertiliser (X_3) on the crop yield (X_1) and the amount of rainfall (X_2), i.e., $R_{3.12}$. We use equation (7) and type "=Sqrt((B39*B39+B40*B40-2*B38*B39*B40)/(1-B38*B38))" in Cell B43 as shown in Fig. 8.7.

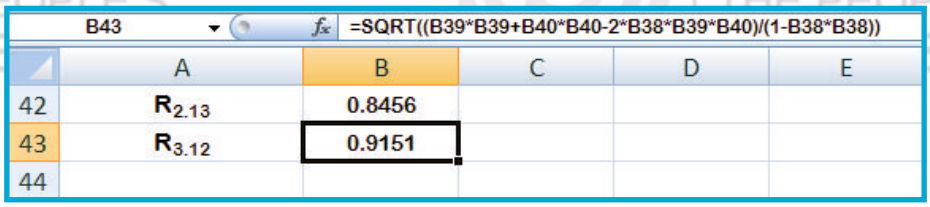


Fig. 8.7

Notice from Fig. 8.7 that the multiple correlation coefficient of the amount of fertiliser on the yield of the crop and the amount of rainfall is 0.9151. It means that the yield of the crop and the amount of rainfall are highly positively correlated with the amount of fertiliser for the given data.

8.5 PARTIAL CORRELATION COEFFICIENTS

You have learnt in Unit 12 of MST-002 that many-a-times, the correlation between two variables is only partly due to the third variable. In such situations, we obtain the relationship between two variables ignoring the effect of the third variable. This is known as partial correlation. Here we briefly mention the main steps as follows:

Step 1: Suppose, X_1, X_2 and X_3 are three variables having n observations or units. We compute the correlation coefficient using equation (4) between (i) X_1 and X_2 denoted by r_{12}, (ii) X_1 and X_3 denoted by r_{13}, and (iii) X_2 and X_3 denoted by r_{23}.

Step 2: The partial correlation is the correlation between two variables, after eliminating the linear effects of other variables on them. The correlation coefficient between X_1 and X_2 after eliminating the linear effect of X_3 on X_1 and X_2 is called the partial correlation coefficient. It is denoted by $r_{12.3}$ and given by

$$r_{12.3} = \frac{(r_{12} - r_{13}r_{23})}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \quad \ldots (8)$$

Step 3: The partial correlation coefficients $r_{13.2}$ and $r_{23.1}$ are given by

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \quad \ldots (9)$$

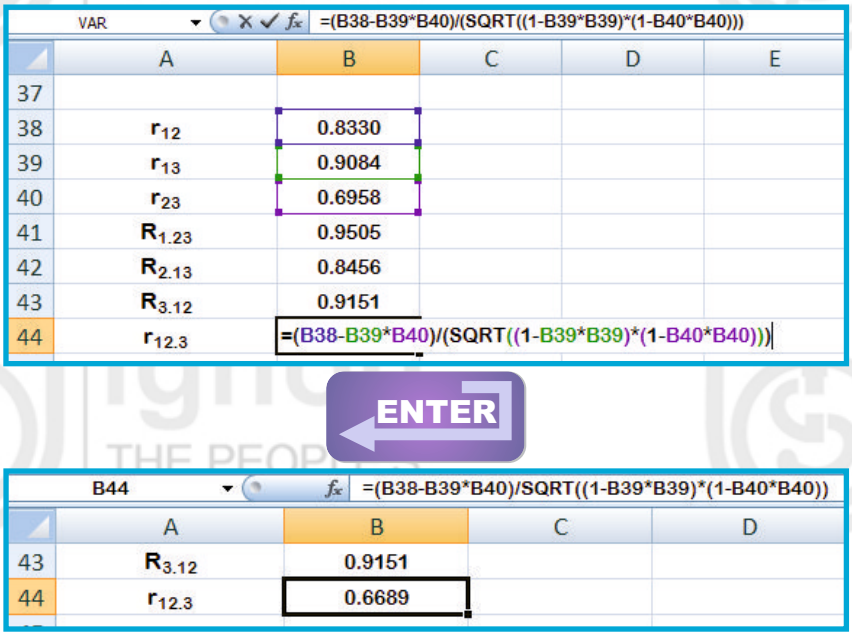
and

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}} \quad \ldots (10)$$

We now describe how to compute the partial correlation coefficients using Excel 2007 and continue with the Excel sheet prepared in Sec. 8.4. In order to compute the partial correlation coefficients, we follow the steps given below:

Step 1: The partial correlation coefficient of the yield of the crop (X_1) and the amount of rainfall (X_2) after eliminating the linear effect of the amount of fertiliser (X_3) is denoted by $r_{12.3}$. We use equation (8) to

compute the partial correlation coefficient and type "=(B38-B39*B40)/(Sqrt((1-B39*B39)*(1-B40*B40)))" in Cell B44 as shown in Fig. 8.8a. When we press ***Enter***, we get the output shown in Fig. 8.8b.



(a)

(b)

Fig. 8.8

Notice from Fig. 8.8b that the partial correlation coefficient between the yield of the crop and the amount of rainfall after eliminating the effect of the amount of fertiliser is 0.6689. It means that the yield of the crop and the amount of rainfall are moderately positively correlated after eliminating the effect of the amount of fertiliser for the given data.

Step 2: We follow Step 1 to compute the partial correlation coefficient of the yield of the crop (X_1) and the amount of fertiliser (X_3) after eliminating the linear effect of the amount of rainfall (X_2) denoted by $r_{13.2}$. We use equation (9) and type "=(B39-B38*B40)/(Sqrt((1-B38*B38)*(1-B40*B40)))" in Cell B45 (Fig. 8.9).

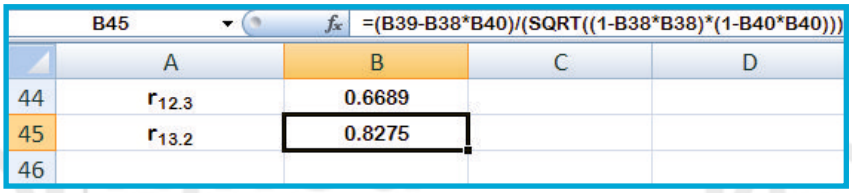


Fig. 8.9

Fig. 8.9 reveals that the partial correlation coefficient between the yield of the crop and the amount of fertiliser after eliminating the effect of the amount of rainfall is 0.8275. It means that the yield of the crop and the amount of the fertiliser are highly positively correlated after eliminating the effect of the amount of rainfall for the given data.

Step 3: We compute the partial correlation coefficient $r_{23.1}$ of the amounts of rainfall (X_2) and fertiliser (X_3) after eliminating the linear effect of the yield of the crop (X_1). We use equation (10) and type "=(B40-B38*B39)/(Sqrt((1-B38*B38)*(1-B39*B39)))" in Cell B46 as shown in Fig. 8.10.

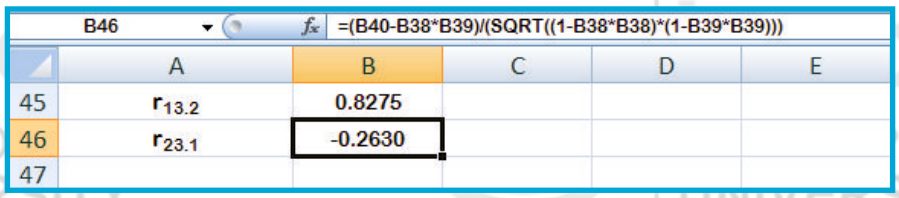


Fig. 8.10

Notice from Fig. 8.10 that the partial correlation coefficient between the amounts of fertiliser and rainfall after eliminating the effect of the yield of the crop is – 0.2630. It means that the amounts of fertiliser and rainfall are poorly negatively correlated after eliminating the effect of the yield of the crop for the given data.

8.6 CORRELATION COEFFICIENTS USING DATA ANALYSIS TOOLPAK

We can also use the *Data Analysis ToolPak* for determining Karl Pearson's correlation coefficient in Excel. It provides the correlation matrix among the variables. To obtain the correlation matrix for Problem 1, we use data on the Excel spreadsheet entered in Step 1 of Sec. 8.3. We follow the steps given below:

Step 1: As shown in Fig. 8.11, we

1. click on the *Data* tab,
2. select *Data Analysis* (Fig. 8.11a),
3. select the *Correlation* option,
4. click on *OK* (Fig. 8.11b), and
5. get a new dialog box (Fig. 8.11c).

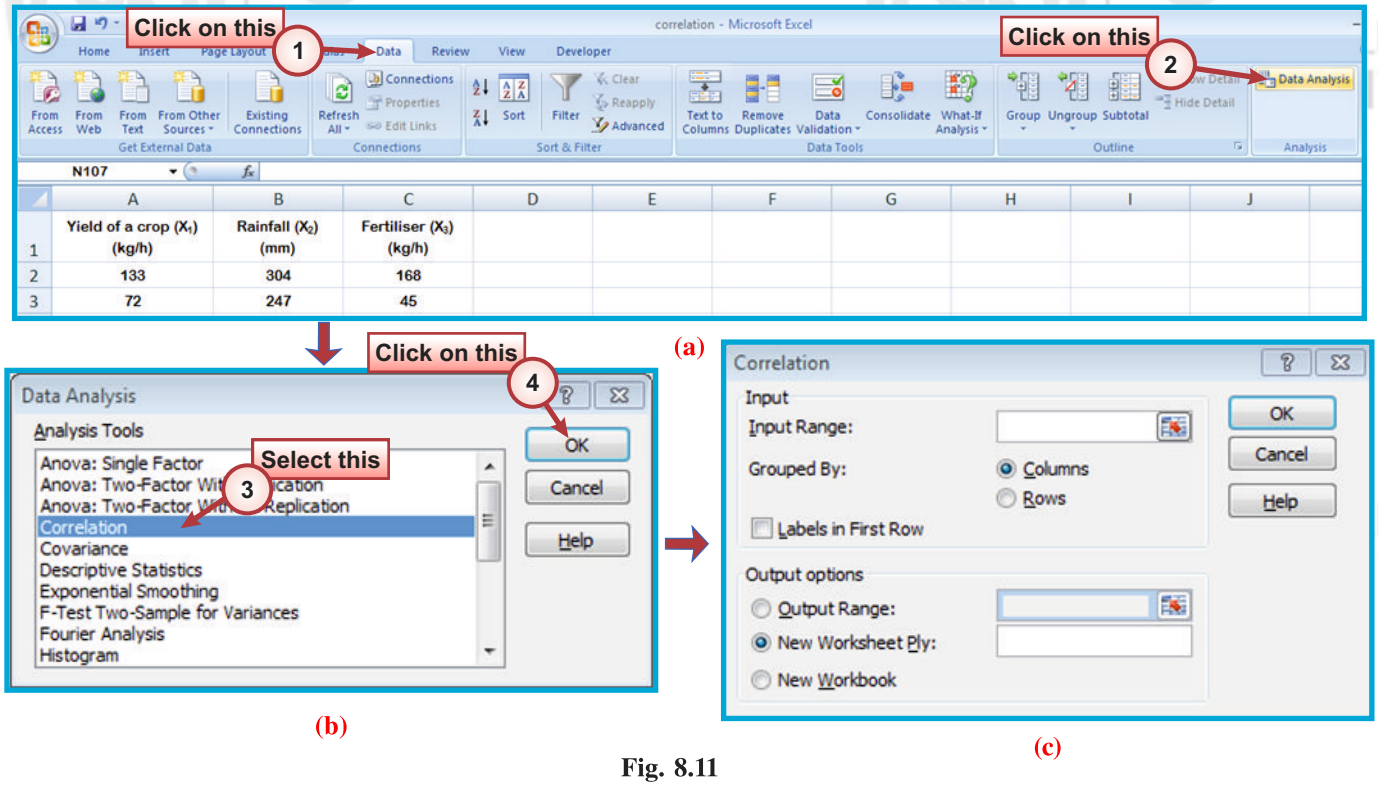


Fig. 8.11

Step 2: As shown in Fig. 8.12, we

1. specify the entire data with labels in *Input Range*, i.e., Cells A1:C36,
2. tick on the box *Labels in first row* since we have included data labels given in Cells A1:C1 in the *Input Range*,

3. provide the *Output Range* under *Output Options*. Here we select Cell A48 for the output, and
4. tick on the *Columns* box under the *Grouped by* option as we have different variables in Columns A, B and C and then click on *OK*.

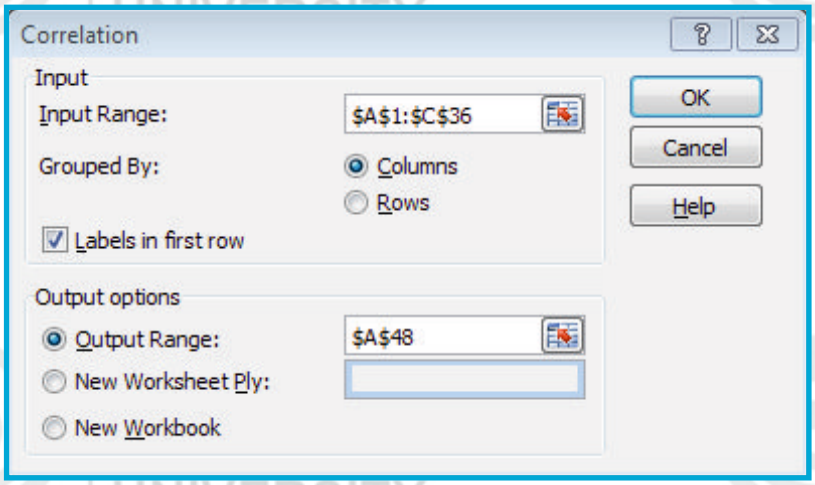


Fig. 8.12

Step 3: After completing Step 2, we obtain the output shown in Fig. 8.13.

	A	B	C	D	E
38	r_{12}	0.8330			
39	r_{13}	0.9084			
40	r_{23}	0.6958			
41	$R_{1.23}$	0.9505			
42	$R_{2.13}$	0.8456			
43	$R_{3.12}$	0.9151			
44	$r_{12.3}$	0.6689			
45	$r_{13.2}$	0.8275			
46	$r_{23.1}$	-0.2630			
47					
48		Yield of a crop (X1)	Rainfall (X2)	Fertiliser (X3)	
49	Yield of a crop (X1)	1			
50	Rainfall (X2)	0.8330	1		
51	Fertiliser (X3)	0.9084	0.6958	1	
52					

Fig. 8.13

Notice from Fig. 8.13 that the *Correlation* option under the *Data Analysis ToolPak* provides the following information:

- ✓ the correlation coefficient between the yield of the crop and the amount of rainfall is 0.8330 (Cell B50).
- ✓ the correlation coefficient between the yield of the crop and the amount of fertiliser is 0.9084 (Cell B51).
- ✓ the correlation coefficient between the amount of fertiliser and rainfall is 0.6958 (Cell C51).

Note that the correlation coefficients computed by *Correlation* under the *Data Analysis ToolPak* in Cells B50, B51 and C51 are the same as computed by the *Correl* function in Cells B38, B39 and B40 (Fig. 8.13).

8.7 RANK CORRELATION COEFFICIENT

You have studied in Unit 7 of MST-002 that the rank correlation coefficient was given by Spearman. Hence, it is known as Spearman's rank correlation coefficient. We denote the rank correlation coefficient by r_s. In this section, we assume that no two (or more) values of the variables have the same rank for either one or both variables. Here we mention the main steps as follows:

Step 1: Suppose we have n observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ corresponding to the variables X and Y, respectively. We compute the ranks of x_i and y_i denoted by R_{xi} and R_{yi}, respectively.

Step 2: We calculate the difference between the ranks of the observation of the variables as

$$d_i = R_{xi} - R_{yi} \qquad \ldots (11)$$

Step 3: Spearman's rank correlation coefficient when two or more values of the variables do not have the same rank in either one or both the variables is given by

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \ldots (12)$$

Here we describe how to compute the rank correlation coefficient using Excel 2007. We first consider Problem 2 and follow the steps given below:

Step 1: We enter the data of Table 2 in an Excel spreadsheet as shown in Fig. 8.14.

	A	B	C
1	S. No.	Part A	Part B
2	1	51	55
3	2	78	67
4	3	68	69
5	4	58	57
6	5	62	60
7	6	65	62
8	7	96	86
9	8	60	73
10	9	76	68
11	10	66	75
12	11	85	81
13	12	95	100

Rank Correlation

Fig. 8.14 : Partial screenshot of the data entered in Excel spreadsheet.

Step 2: We compute the rank for the marks given in Cell B2 for Part A of the first student. As shown in Fig. 8.15, we

1. select Cell D2,
2. click on the *Formulas* tab, and
3. click on *More Functions* → *Statistical* → *Rank*.

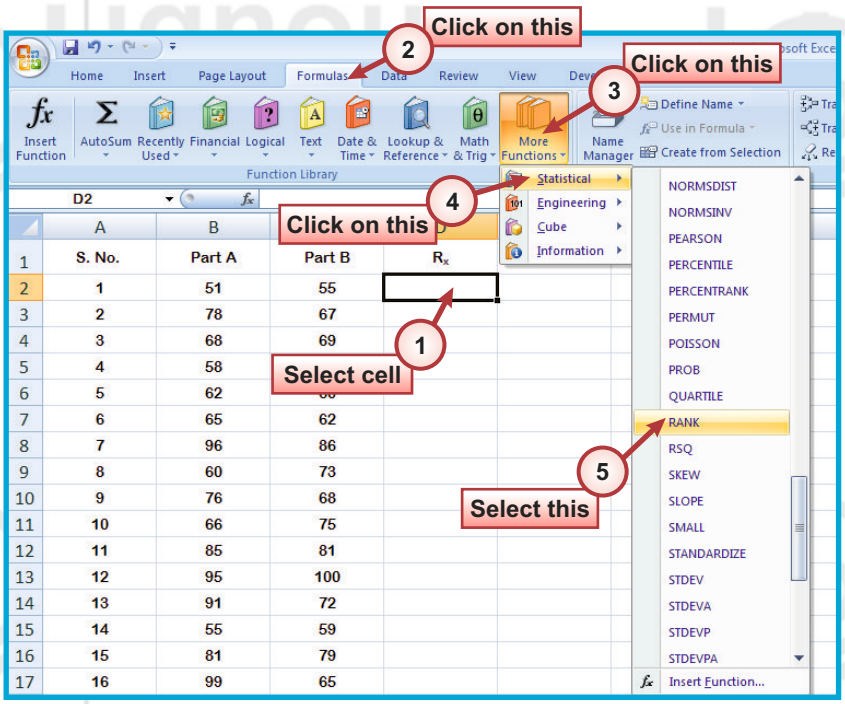


Fig. 8.15

Step 3: A new dialog box opens. As shown in Fig. 8.16a, we

- ✓ select Cell B2 as ***Number*** to compute the rank for the marks of the first student in Part A entered in Cell B2,
- ✓ select Cells B2:B26 as ***Ref***. It means the reference of the data, i.e., the marks of the Part A. It assigns ranks on that basis,
- ✓ type "1" in ***Order*** since we want ranks in ascending order (if we want the ranks in descending order, we type "0" or leave it blank), and
- ✓ click on ***OK*** (Fig. 8.16a). The value of rank for the marks given for part A to the first student is shown in Fig. 8.16b.

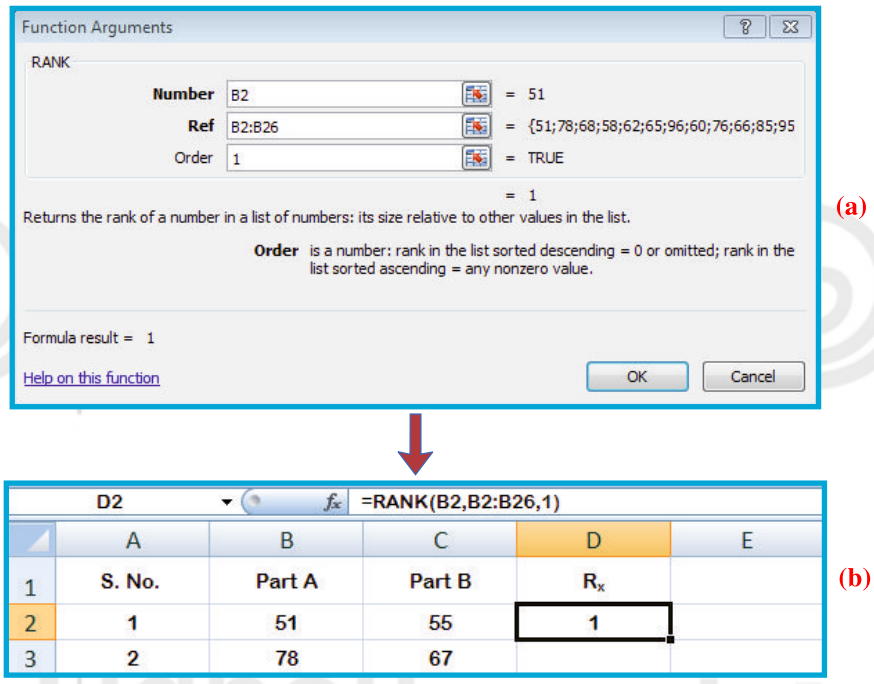


Fig. 8.16

Step 4: We now fix the range given in ***Ref***, i.e., B2:B26 by putting the "$" sign, i.e., $B$2:$B$26 as shown in Fig. 8.17. Note that it helps us in copying the formula in Cells D3:D26 as the reference range is fixed for Part A.

D2 | =RANK(B2,$B$2:$B$26,1)

	A	B	C	D	E
1	S. No.	Part A	Part B	R_x	
2	1	51	55	1	
3	2	78	67		
4	3	68	69		
5	4	58	57		

Fig. 8.17

Step 5: We follow Steps 2 and 3 to compute the rank for the marks given for Part B (Cell C2) to the first student in Cell E2. We also fix the range as explained in Step 4. The output is shown in Fig. 8.18.

E2 | =RANK(C2,$C$2:$C$26,1)

	B	C	D	E	F
1	Part A	Part B	R_x	R_y	
2	51	55	1	3	
3	78	67			
4	68	69			
5	58	57			

Fig. 8.18

Step 6: We compute the value of the difference between the ranks obtained for Parts A and B in Cells D2 and E2, respectively, i.e., we compute $d = R_x - R_y$ by typing "=D2-E2" in Cell F2 as shown in Fig. 8.19a. We get the output when we press ***Enter*** (Fig. 8.19b).

(a)

MINVERSE | =D2-E2

	C	D	E	F	G
1	Part B	R_x	R_y	$d = R_x - R_y$	
2	55	1	3	=D2-E2	
3	67				
4	69				
5	57				

ENTER

(b)

F2 | =D2-E2

	C	D	E	F	G
1	Part B	R_x	R_y	$d = R_x - R_y$	
2	55	1	3	-2	
3	67				
4	69				
5	57				

Fig. 8.19

Step 7: We compute the square of the difference between the two ranks, i.e., d^2, by typing "=F2*F2" in Cell G2 and pressing ***Enter*** as shown in Fig. 8.20.

RANK =F2*F2

	E	F	G	H
1	R_y	$d = R_x - R_y$	d^2	
2	3	-2	=F2*F2	
3				

(a)

ENTER

G2 =F2*F2

	F	G	H
1	$d = R_x - R_y$	d^2	
2	-2	4	
3			

(b)

Fig. 8.20

Step 8: We now select Cells D2:G2 and drag them down up to Row 26 to determine the required values for the remaining students as shown in Fig. 8.21.

	C	D	E	F	G	H
1	Part B	R_x	R_y	$d = R_x - R_y$	d^2	
2	55	1	3	-2	4	
3	67					
4	69					
5	57					

(a)

DRAG THEM DOWN

	C	D	E	F	G	H
1	Part B	R_x	R_y	$d = R_x - R_y$	d^2	
2	55	1	3	-2	4	
3	67	14	10	4	16	
4	69	11	12	-1	1	
5	57	4	4	0	0	
6	60	7	6	1	1	
7	62	9	7	2	4	
8	86	23	21	2	4	
9	73	5	14	-9	81	
10	68	13	11	2	4	
11	75	10	15	-5	25	
12	81	19	18	1	1	
13	100	22	25	-3	9	

(b)

Fig. 8.21

Step 9: We now compute the value of $\sum d^2$ using "=Sum(G2:G26)" function in Cell G27 as shown in Fig. 8.22.

G27 =SUM(G2:G26)

	E	F	G	H
26	19	-4	16	
27		Σd^2	664	
28				

Fig. 8.22

Step 10: We type the value of n (= 25), i.e., the number of students who participated in the contest in Cell G28 as shown in Fig. 8.23.

	F	G	H
27	Σd^2	664	
28	n	25	
29			

Fig. 8.23

Step 11: The formula for computing the rank correlation coefficient is given by equation (12). We type "=1-((6*G27)/(G28*(G28*G28-1)))" in Cell G29 as shown in Fig. 8.24a and press ***Enter***. The output is shown in Fig. 8.24b.

(a)

RANK | =1-((6*G27)/(G28*(G28*G28-1)))

	E	F	G	H	I
26	19	-4	16		
27		Σd^2	664		
28		n	25		
29		r_s	=1-((6*G27)/(G28*(G28*G28-1)))		
30					

ENTER

(b)

G29 | =1-((6*G27)/(G28*(G28*G28-1)))

	E	F	G	H	I
27		Σd^2	664		
28		n	25		
29		r_s	0.7446		
30					

Fig. 8.24

Notice from Fig. 8.24 that the rank correlation coefficient between the marks of Parts A and B is 0.7446. We can conclude that there is high positive correlation between the marks obtained in both parts for the given data.

Step 12: Note that we can also compute Spearman's rank correlation coefficient by using ***Correl*** function since Spearman's rank correlation coefficient [equation (12)] is the same as Karl Pearson's correlation coefficient [equation (4)] computed for the ranks of both the variables instead of their values. We use ***Correl*** function (as explained in Sec. 8.3) in Cell G30. We select Cells D2:D26 as ***Array 1***, Cells E2:E26 as ***Array 2***, and click on ***OK*** to obtain the value of the correlation coefficient in Cell G30 (Fig. 8.25).

G30 | =CORREL(D2:D26,E2:E26)

	E	F	G	H
29		r_s	0.7446	
30	Karl Pearson	r_s	0.7446	
31				

Fig. 8.25

Notice from Fig. 8.25 that Karl Pearson's correlation coefficient given in Cell G30 computed from equation (4) is the same as the Spearman's rank correlation coefficient given in Cell G29 computed from equation (12). Hence we can apply ***Correl*** function to compute the Spearman's rank correlation coefficient directly.

In this section, we have explained the computation of the rank correlation coefficient when two or more values of the variables do not have the same rank. But there might be a situation when two or more values have the same rank in either one or both variables. Then the rank is said to be tied. We consider this situation in the next section.

8.8 RANK CORRELATION COEFFICIENT FOR TIED OR REPEATED RANKS

You have studied in Unit 7 of MST-002 that to compute Spearman's rank correlation coefficient, we use the correction factor when two or more values of the variables are the same. Here we mention the main steps as follows:

Step 1: If two or more values of the variables are the same, we assign common ranks to the repeated values. This common rank is the average of ranks that these repeated values would receive if there were no repetition.

Step 2: We use the correction factor $\frac{m(m^2-1)}{12}$ when there is a repetition of ranks in Spearman's rank correlation coefficient given in equation (12).

Step 3: If $(R_{xi})_c$ and $(R_{yi})_c$ are the corrected ranks of x_i and y_i, for $i = 1, 2, \ldots, n$, respectively, the corrected Spearman's rank correlation coefficient is given by

$$(r_s)_c = \frac{\frac{(n^3-n)}{6} - \left(\sum_{i=1}^{n} d_{ci}^2 + T_x + T_y\right)}{\sqrt{\left[\left(\frac{n^3-n}{6} - 2T_x\right)\left(\frac{n^3-n}{6} - 2T_y\right)\right]}} \quad \ldots (13)$$

where

$$T_x = \sum_{i=1}^{p} \frac{(m_{xi}^3 - m_{xi})}{12}, \quad \ldots (14)$$

$$T_y = \sum_{i=1}^{q} \frac{(m_{yi}^3 - m_{yi})}{12}, \quad \ldots (15)$$

m_{xi} – number of times the rank of x_i is repeated,

m_{yi} – number of times the rank of y_i is repeated,

p – number of ranks of variable X, which have repeated values, and

q – number of ranks of variable Y, which have repeated values.

Note that this correction factor is added for every repetition of rank in both variables.

Step 4: In Unit 7 of MST-002, we have used the correction factors only in the numerator. So equation (13) becomes

$$(r_s)_c = 1 - \frac{6\left(\sum_{i=1}^{n} d_{ci}^2 + T_x + T_y\right)}{n(n^2-1)} \quad \ldots (16)$$

Note that here we are using equation (13) because most of the statistical software including Excel uses this formula.

Here, we describe how to compute the rank correlation coefficient for tied or repeated ranks using Excel 2007. We first consider Problem 3. In order to compute the rank correlation coefficient, we follow the steps given below:

Step 1: We enter the data of Table 3 in an Excel spreadsheet and follow Steps 1 to 5 of Sec. 8.7 and drag down Cells D2:E2 up to Row 16 to compute the ranks for the given data. The output is shown in Fig. 8.26.

The value 65 is repeated 3 times and its position will be 5th, 6th and 7th in the data given in Cells B2:B16. The *Rank* function of Excel assigns the lowest rank, i.e., 5 to each 65. But we assign the ranks as an average of 5, 6 and 7, i.e., 6 to each 65. So we use a correction factor in Excel to modify the ranks.

	A	B	C	D	E
1	S. No.	Marks	TV Viewing Hour	R_x	R_y
2	1	60	6	1	15
3	2	72	1	10	1
4	3	64	2	4	5
5	4	71	2	9	5
6	5	77	3	13	10
7	6	65	2	5	5
8	7	62	1	3	1
9	8	78	2	14	5
10	9	65	4	5	13
11	10	72	2	10	5
12	11	73	1	12	1
13	12	79	1	15	1
14	13	67	3	8	10
15	14	65	3	5	10
16	15	61	5	2	14

Fig. 8.26

Notice from Fig. 8.26 that the ranks of the first variable (marks), i.e., 5 (highlighted with orange colour) and 10 (highlighted with purple colour) are repeating 3 and 2 times, respectively. Also, the ranks of the second variable (number of hours spent in watching TV), i.e., 1 (highlighted with green colour), 5 (highlighted with yellow colour) and 10 (highlighted with blue colour) are repeating 4, 5 and 3 times, respectively.

Step 2: Excel assigns the lowest rank for all repeated values of the variable. But we assign the average of the ranks to all repeated values. So we add a correction factor to these ranks, i.e., (Count(reference) + 1 – Rank(number, reference, 0) – Rank(number, reference, 1))/2.

Here, number is the value of the variable for which we need to compute rank and reference is all given values of that variable. To compute the corrected rank for the first variable, we type "=D2+(Count($B$2:$B$16)+1-Rank(B2, $B$2:$B$16,0)-Rank(B2,$B$2:$B$16,1))/2" in Cell F2 as shown in Fig. 8.27.

F2 | =D2+(COUNT($B$2:$B$16) + 1 - RANK(B2, $B$2:$B$16, 0)- RANK(B2, $B$2:$B$16, 1))/2

	A	B	C	D	E	F	G	H
1	S. No.	Marks	TV Viewing Hour	R_x	R_y	$(R_x)_c$		
2	1	60	6	1	15	1		
3	2	72	1	10	1			

Fig. 8.27

Step 3: To compute the corrected rank for the second variable, we type "=E2+(Count($C$2:$C$16)+1-Rank(C2,$C$2:$C$16,0)-Rank(C2,$C$2:$C$16,1))/2" in Cell G2 as shown in Fig. 8.28.

G2 | =E2+(COUNT($C$2:$C$16) + 1 - RANK(C2, $C$2:$C$16, 0)- RANK(C2, $C$2:$C$16, 1))/2

	A	B	C	D	E	F	G	H
1	S. No.	Marks	TV Viewing Hour	R_x	R_y	$(R_x)_c$	$(R_y)_c$	
2	1	60	6	1	15	1	15	
3	2	72	1	10	1			
4	3	64	2	4	5			

Fig. 8.28

Step 4: We compute the value of the difference between the ranks, i.e., d_c and the square of the difference between the two ranks, i.e., d_c^2 by typing "=F2-G2" and "=H2*H2" in Cells H2 and I2, respectively, as shown in Fig. 8.29.

	F	G	H	I
1	$(R_x)_c$	$(R_y)_c$	$d_c = (R_x)_c - (R_y)_c$	d_c^2
2	1	15	-14	196

Type "=F2-G2" here | Type "=H2*H2" here

Fig. 8.29

Step 5: We now select Cells F2:I2 and drag them down up to Row 16 to determine the required values for the remaining students as shown in Fig. 8.30.

(a)

	F	G	H	I	J
1	$(R_x)_c$	$(R_y)_c$	$d_c = (R_x)_c - (R_y)_c$	d_c^2	
2	1	15	-14	196	
3					
4					

DRAG THEM DOWN

(b)

	F	G	H	I	J
1	$(R_x)_c$	$(R_y)_c$	$d_c = (R_x)_c - (R_y)_c$	d_c^2	
2	1	15	-14	196	
3	10.5	2.5	8	64	
4	4	7	-3	9	
5	9	7	2	4	
6	13	11	2	4	
7	6	7	-1	1	
8	3	2.5	0.5	0.25	
9	14	7	7	49	
10	6	13	-7	49	
11	10.5	7	3.5	12.25	
12	12	2.5	9.5	90.25	
13	15	2.5	12.5	156.25	
14	8	11	-3	9	
15	6	11	-5	25	
16	2	14	-12	144	
17					

Fig. 8.30

Step 6: We compute the value of $\sum d_c^2$ using "=Sum(I2:I16)" function in Cell I17 as shown in Fig. 8.31.

I17 =SUM(I2:I16)

	H	I	J
16	-12	144	
17	Σd_c^2	813	
18			

Fig. 8.31

Step 7: We type the value of n (= 15), i.e., the number of students in Cell I18 (Fig. 8.32). We also type "=(I18^3-I18)/6" in Cell I19 to compute the value of $\frac{(n^3 - n)}{6}$ as shown in Fig. 8.32.

I19 =(I18^3-I18)/6

	H	I	J
17	Σd_c^2	813	
18	n	15	
19	$(n^3-n)/6$	560	
20			

Fig. 8.32

Step 8: We type the values of the corrected tie ranks in Cells A19 and A20 for the first variable and in Cells D19 to D21 for the second variable. We also type the number of times the particular tied rank is repeated in Cells B19 to B20 for the first variable (m_x) and in Cells E19 to E21 for the second variable (m_y) as shown in Fig. 8.33.

	A	B	C	D	E
17					
18	Tie $(R_x)_c$	m_x		Tie $(R_y)_c$	m_y
19	6	3		2.5	4
20	10.5	2		7	5
21				11	3

Fig. 8.33

Step 9: We type "=(B19^3-B19)/12" in Cell C19 to compute the value of $\frac{(m_x^3 - m_x)}{12}$ and drag down Cell C19 up to Cell C20 (Fig. 8.34).

(a)

C19 =(B19^3-B19)/12

	A	B	C	D
18	Tie $(R_x)_c$	m_x	$(m_x^3-m_x)/12$	
19	6	3	2	
20	10.5	2		

DRAG IT DOWN

(b)

	A	B	C	D
18	Tie $(R_x)_c$	m_x	$(m_x^3-m_x)/12$	
19	6	3	2	
20	10.5	2	0.5	
21				

Fig. 8.34

Step 10: We type "=(E19^3-E19)/12" in Cell F19 to compute the value of $\frac{(m_y^3 - m_y)}{12}$ and drag Cell F19 down up to Cell F21 (Fig. 8.35).

F19 =(E19^3-E19)/12

	D	E	F	G
18	Tie $(R_y)_c$	m_y	$(m_y^3-m_y)/12$	
19	2.5	4	5	
20	7	5		

(a)

DRAG IT DOWN

	D	E	F	G
18	Tie $(R_y)_c$	m_y	$(m_y^3-m_y)/12$	
19	2.5	4	5	
20	7	5	10	
21	11	3	2	
22				

(b)

Fig. 8.35

Step 11: We use equation (14) and compute the value of T_x using "=Sum(C19:C20)" function in Cell C22 as shown in Fig. 8.36.

C22 =SUM(C19:C20)

	A	B	C	D
18	Tie $(R_x)_c$	m_x	$(m_x^3-m_x)/12$	Tie $(R_y)_c$
19	6	3	2	2.5
20	10.5	2	0.5	7
21				11
22		T_x	2.5	
23				

Fig. 8.36

Step 12: We use equation (15) and compute the value of T_y using "=Sum(F19:F21)" function in Cell F22 as shown in Fig. 8.37.

F22 =SUM(F19:F21)

	D	E	F	G
18	Tie $(R_y)_c$	m_y	$(m_y^3-m_y)/12$	
19	2.5	4	5	
20	7	5	10	
21	11	3	2	
22		T_y	17	
23				

Fig. 8.37

Step 13: The formula for computing the corrected rank correlation coefficient is given by equation (13). We type "=(I19-(I17+C22+F22))/((Sqrt(I19-2*C22))*(Sqrt(I19-2*F22)))" in Cell I20 as shown in Fig. 8.38.

I20 =(I19-(I17+C22+F22))/((SQRT(I19-2*C22))*(SQRT(I19-2*F22)))

	B	C	D	E	F	G	H	I	J
17							Σd_c^2	813	
18	m_x	$(m_x^3-m_x)/12$	Tie $(R_y)_c$	m_y	$(m_y^3-m_y)/12$		n	15	
19	3	2	2.5	4	5		$(n^3-n)/6$	560	
20	2	0.5	7	5	10		r_s	-0.5043	
21			11	3	2				
22	T_x	2.5		T_y	17				

Fig. 8.38

Notice from Fig. 8.38 that the rank correlation coefficient between the marks of Parts A and B is – 0.5043. We can conclude that there is

negative correlation between the number of hours spent in watching TV and the marks obtained for the given data, i.e., an increase in the number of hours spent in watching TV is correlated with a decrease in the marks obtained in the examination.

Step 14: We can also compute Spearman's rank correlation coefficient using the *Correl* function as explained in Sec. 8.7. We use the *Correl* function, as explained in Sec. 8.3, in Cell I21. We select Cells F2:F16 as ***Array 1***, Cells G2:G16 as ***Array 2***, and click on *OK*. We obtain the value of the correlation coefficient in Cell I21 (Fig. 8.39).

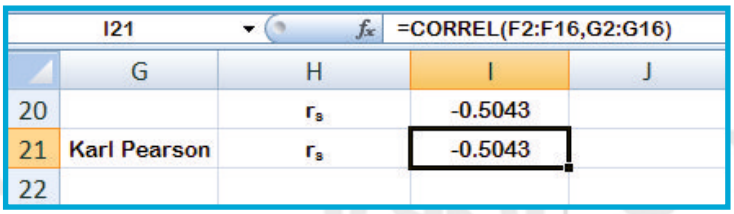


Fig. 8.39

Notice from Fig. 8.39 that the Karl Pearson's correlation coefficient computed from equation (4) is the same as the rank correlation coefficient computed from equation (13).

You can now try the following exercises for practice.

Activity

Work out the following exercises with the help of MS Excel 2007 and interpret the results:

A1) Examples 1 to 3 given in Unit 6 of MST-002.

A2) Exercises E4 to E6 given in Unit 6 of MST-002.

A3) Examples 1 to 4 given in Unit 7 of MST-002.

A4) Exercises E1 to E3 given in Unit 7 of MST-002.

A5) Examples 1 to 3 given in Unit 11 of MST-002.

A6) Exercises E1 to E5 given in Unit 11 of MST-002.

A7) Examples 1 to 3 given in Unit 12 of MST-002.

A8) Exercises E1 to E4 given in Unit 12 of MST-002.

Match the results with the manual computation of the correlation coefficients done in Units 6, 7, 11 and 12 of MST-002.

Continuous Assessment 8

1. Suppose we are interested in determining the correlation coefficients between the electricity consumption, size of the house and the number of hours an AC is used in a household during summers. For this purpose, a

sample of 40 houses having one AC was selected. We have recorded the electricity consumption (in kWh), size of the house (in square feet) and number of hours of AC use for one month during summers in Table 4.

Table 4: Electricity consumption data

S. No.	Unit (in kWh)	Area (in sq ft)	AC (in hours)
1	1060	1316	5
2	1150	1420	7
3	1365	1556	12
4	1275	1488	9
5	1425	1612	13
6	1310	1516	10
7	1365	1556	12
8	1075	1352	6
9	925	1168	4
10	1340	1540	11
11	1425	1612	13
12	1150	1420	8
13	1060	1316	5
14	1545	1680	15
15	1140	1388	7
16	1075	1352	6
17	1620	1736	16
18	1050	1296	5
19	1310	1516	10
20	1645	1760	16
21	1565	1696	15
22	1215	1464	9
23	1275	1488	10
24	1465	1632	13
25	1080	1356	7
26	975	1196	4
27	1040	1256	5
28	1340	1540	11
29	865	1144	4
30	1175	1440	8
31	1080	1356	7
32	1500	1652	15
33	1175	1440	9
34	1050	1296	5
35	1365	1580	12
36	1465	1632	15
37	1215	1464	9
38	1365	1580	12
39	1140	1388	7
40	1005	1224	4

i) Compute Pearson's correlation coefficients between (a) electricity consumption and size of the house, (b) electricity consumption and number of hours the AC is used and (c) size of the house and number of hours the AC is used for the given data.

ii) Determine the multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$.

iii) Obtain the partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.

2. 25 singers participated in a singing contest and were evaluated by two judges A and B. The judges awarded scores between 0 and 100 to each contestant, which are recorded in Table 5.

Table 5: Scores of the singing contest

Judge A	48	75	65	55	59	62	93	57	73	63	82	99	88
	52	78	61	58	50	95	70	80	81	97	47	77	
Judge B	51	76	65	53	56	58	94	69	64	71	77	98	87
	55	75	61	60	49	93	92	81	72	96	46	79	

Compute Spearman's rank correlation coefficient between the scores awarded by both judges for the given data.

3. Suppose we are interested in determining the relationship between the ranking of 15 universities of the country and their success rates. The rank from 0 to 10 is assigned to a university on the basis of different factors and the success rate is based on the percentage of students passing out. The ranks and success rates of 15 universities are recorded in Table 6.

Table 6: Ranks and success rates of 15 universities

S. No.	Rank of the University	Success Rate (in %)
1	2	57
2	9	98
3	5	70
4	9	95
5	6	75
6	4	70
7	8	96
8	8	92
9	7	87
10	8	75
11	7	85
12	3	76
13	5	64
14	7	82
15	5	58

Compute Spearman's rank correlation coefficient between the ranks and success rates of the universities for the given data.

Home Work: Do It Yourself

1) Follow the steps explained in Secs. 8.3 to 8.8 to compute the various correlation coefficients for the data of Tables 1, 2 and 3. Take the screenshots and keep them in your record book.
2) Develop the spreadsheets for the exercises given in "Continuous Assessment 8" as explained in this lab session. Take screenshots of the final spreadsheets.
3) **Do not forget** to keep the screenshots in your record book as these will contribute to your continuous assessment in the Laboratory.