Two-Sample Tests

SESSION 10

Structure

10.1 INTRODUCTION

Prerequisite

- Lab Sessions 6, 7 and 9 of MSTL-001 (Basic Statistics Lab).
- Block 3 of MST-004 (Statistical Inference).

In Lab Session 9, you have learnt how to apply the Z-test, t-test and chi-square test for one sample using MS Excel 2007.

In Units 10, 11 and 12 of MST-004, you have studied two-sample parametric tests, namely, Z-test for the difference of two population means and difference of two population proportions, t-test for the difference of two population means, paired t-test and F-test for two population variances. You have also learnt which of these tests should be applied in a given situation.

In this lab session, you will learn how to apply the Z-test, t-test and F-test for two samples using MS Excel 2007. In the next lab session, you will learn how to apply the analysis of variance (ANOVA) using MS Excel 2007.

Objectives

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

- prepare the spreadsheet in MS Excel 2007;
- apply the Z-test for the difference of two population means;
- apply the t-test for the difference of two population means;
- apply the paired t-test;
- apply the Z-test for the difference of two population proportions; and
- apply the F-test for two population variances.

10.2 PROBLEM DESCRIPTION

In this lab session, we state five problems to illustrate the applications of different kinds of two-sample parametric tests:

1. A medical researcher wishes to investigate whether the pulse rates of smokers are higher than the pulse rates of non-smokers. She takes samples of 50 smokers and 60 non-smokers and measures their pulse rates. The results are shown in Table 1. Can the researcher conclude that smokers have higher pulse rates than the non-smokers, at 1% level of significance?

bpm stands for beats per minute.

Table 1: Pulse rates of smokers and non-smokers

S. No.	Pulse Rate (in bpm)		S. No.	Pulse Rate (in bpm)	
	Smoker	Non-smoker		Smoker	Non-smoker
1	80	90	31	70	90
2	82	70	32	72	98
3	110	78	33	84	80
4	95	80	34	100	76
5	100	88	35	98	106
6	98	70	36	110	90
7	100	90	37	105	85
8	114	100	38	90	87
9	120	110	39	75	88
10	90	92	40	70	90
11	95	95	41	87	78
12	110	88	42	88	95
13	95	85	43	90	90
14	120	75	44	100	88
15	98	60	45	88	74
16	100	100	46	90	100
17	105	94	47	70	98
18	90	88	48	88	78
19	99	92	49	75	88
20	110	75	50	90	90
21	105	74	51		100
22	95	94	52		86
23	98	68	53		85
24	100	90	54		98
25	95	102	55		90
26	90	85	56		88
27	98	100	57		90
28	100	75	58		94
29	88	94	59		90
30	90	88	60		102

2. A researcher wishes to find out whether the waiting time for a patient to meet a doctor in the emergency room at a government hospital is more than the corresponding waiting time at a private hospital. She records the waiting time for 25 patients in both hospitals. The data are given in Table 2.

Table 2 : Waiting time

S. No.	Waiting Time (in minutes)		S. No.	Waiting Time (in minutes)	
	Government	Private		Government	Private
1	30	12	14	20	30
2	20	10	15	35	15
3	15	20	16	20	10
4	20	15	17	15	10
5	24	10	18	40	12
6	20	8	19	20	30
7	15	10	20	15	15

S. No.	Waiting Time (in minutes)		S. No.	Waiting Time (in minutes)	
	Government	Private		Government	Private
8	20	18	21	25	
9	25	15	22	30	
10	15	10	23	14	
11	22	15	24	20	
12	20	20	25	25	
13	34	15			

Assuming that the waiting time is normally distributed and the variances of the distributions of waiting time are equal in both cases:

- Formulate the null and alternative hypotheses.
- Is there enough evidence that the average waiting time for a patient to see a doctor in the emergency room at a government hospital is more than the average waiting time at a private hospital at 2% level of significance?

3. A group of 20 children was tested to find out how many digits they would repeat from memory after hearing them once. Then they were given practice for the test and retested after one week. The results obtained before and after the practice were as follows:

Table 3: Results of the tests before and after the practice

Child Number	Recall Before (X)	Recall After (Y)	Child Number	Recall Before (X)	Recall After (Y)
1	6	8	11	8	8
2	4	4	12	5	6
3	5	8	13	5	7
4	8	8	14	4	4
5	4	4	15	6	8
6	2	5	16	4	4
7	3	5	17	2	3
8	5	5	18	2	4
9	2	6	19	6	6
10	7	8	20	4	5

Assuming that the memories of the children before and after practice follow normal distributions, does the practice improve the performance of the children at 5% level of significance?

4. A textile company launched 100% cotton cloth for both male and female customers. The company conducted a survey to understand the perception of customers about 100% cotton cloth. The company took a random sample of 120 male and 100 female customers. Out of 120 males, 80 responded that 100% cotton cloth matched their lifestyle. Out of 100 females, 60 females responded that 100% cotton cloth matched their lifestyle. Does this indicate that there is a significant difference in the proportion of male and female customers in the population stating that 100% cotton cloth matches with their lifestyle at 5% level of significance?

5. In Problem 2, if the researcher also wishes to investigate whether the variances of both distributions of waiting time are equal, then

- formulate the null and alternative hypotheses.
- are the variances of distributions of waiting time for a patient to see a doctor in the emergency room at the government hospital and the private hospital equal at 2% level of significance?

10.3 Z-TEST FOR THE DIFFERENCE OF TWO POPULATION MEANS

In Unit 10 of MST-004, you have learnt that the Z-test is used for testing the difference of two population means when the population standard deviations (σ_1 and σ_2) are known or unknown and the population under study is normal or non-normal for the large samples. But when the samples are of small size, we apply the Z-test only when the populations under study are normal and the population standard deviations (σ_1 and σ_2) are known.

The procedure of the Z-test for the difference of two population means has been described in Unit 10 of MST-004. We briefly mention the main steps as follows:

Step 1: We first formulate the null hypothesis (H_0) and alternative hypothesis (H_1). If σ_1 and σ_2 are the means of population-I and population-II, respectively, we can take the null and alternative hypotheses as follows:

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 \neq \mu_2 \qquad [\text{for two-tailed test}]$$

or

$$\left.\begin{array}{l} H_0 : \mu_1 \leq \mu_2 \text{ and } H_1 : \mu_1 > \mu_2 \\ H_0 : \mu_1 \geq \mu_2 \text{ and } H_1 : \mu_1 < \mu_2 \end{array}\right\} \qquad [\text{for one-tailed test}]$$

Step 2: We calculate the value of the test statistic Z using the formula given below:

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \text{... (1)}$$

where $\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ and $\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ are the sample means selected from population-I and population-II, respectively.

If σ_1^2 and σ_2^2 are unknown and the sample sizes n_1 and n_2 are large (> 30), the test statistic Z is given by

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad \text{... (2)}$$

where S_1^2 and S_2^2 are the variances of the first and second samples, respectively, and are given below:

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_i - \bar{X}\right)^2 \qquad \text{... (3)}$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(Y_i - \bar{Y}\right)^2 \qquad \text{... (4)}$$

Step 3: The other steps are the same as explained in the procedure for the Z-test in Sec. 9.3 of Lab Session 9.

Steps in Excel

In the given Problem 1, we have to test whether the pulse rates of smokers are higher than the pulse rates of non-smokers. Here the population variances (σ_1^2 and σ_2^2) are unknown. But the sample sizes $n_1 = 50$ (> 30) and $n_2 = 60$ (>30) are large. So we may use the Z-test for the difference of two population means.

For applying the Z-test, we first have to set up the null and alternative hypotheses. Here we have to find out whether the pulse rates of smokers are higher than the pulse rates of non-smokers. If μ_1 and μ_2 denote the mean pulse rates of smokers and non-smokers, respectively, we can formulate the null and alternative hypotheses as follows:

We can also take the null and alternative hypotheses as follows:
$H_0 : \mu_1 - \mu_2 \le 0$ and
$H_1 : \mu_1 - \mu_2 > 0$

$$H_0 : \mu_1 \le \mu_2 \text{ and } H_1 : \mu_1 > \mu_2 \text{ (claim)} \qquad [\text{right-tailed test}]$$

The following steps are used to apply the Z-test for the difference of two population means in Excel 2007:

Step 1: We enter the data (given in Table 1) in an Excel 2007 spreadsheet. We start by entering the headings of the data in Rows 1 and 2 in Excel sheet and the data from Row 3 onwards. The entries will go up to Row 52 for the first sample (smokers) and Row 62 for the second sample (non-smokers). For the given data, the spreadsheet will look like the one shown in Fig. 10.1.

	A	B	C
1	S. No.	Pulse Rate(in bpm)	
2		Smoker	Non-smoker
3	1	80	90
4	2	82	70
5	3	110	78
6	4	95	80
7	5	100	88
8	6	98	70
9	7	100	90
10	8	114	100
11	9	120	110
12	10	90	92

Step 2: We calculate the sample mean pulse rate of smokers and non-smokers as explained in Step 2 under the heading 'Steps in Excel' of Sec. 9.3 of Lab Session 9. We use Cells B63 and C63 for computing the mean pulse rates of the samples of smokers and non-smokers, respectively. The outputs are shown in Figs. 10.2a and 10.2b, respectively.

B63 =AVERAGE(B3:B52)

	A	B	C
62	60		102
63	Sample mean	94.2000	
64			

(a)

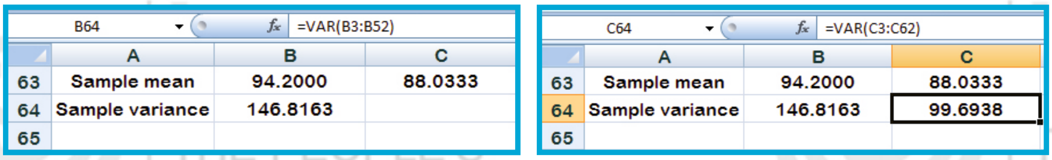
C63 =AVERAGE(C3:C62)

	A	B	C
62	60		102
63	Sample mean	94.2000	88.0333
64			

(b)

Fig. 10.2

Step 3: We calculate the sample variances of the samples of smokers and non-smokers as explained in Step 2 under the heading 'Steps in Excel' of Sec. 9.6 of Lab Session 9. We use Cells B64 and C64 for computing the sample variances of the samples of smokers and non-smokers, respectively. The outputs are shown in Figs. 10.3a and 10.3b, respectively.



(a)

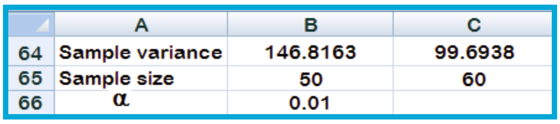
	B64	f_x =VAR(B3:B52)
	A	B	C
63	Sample mean	94.2000	88.0333
64	Sample variance	146.8163	
65			

(b)

	C64	f_x =VAR(C3:C62)
	A	B	C
63	Sample mean	94.2000	88.0333
64	Sample variance	146.8163	99.6938
65			

Fig. 10.3

Step 4: We now type the values of the sample sizes n_1, n_2 and level of significance α in Cells B65, C65 and B66, respectively (Fig. 10.4).

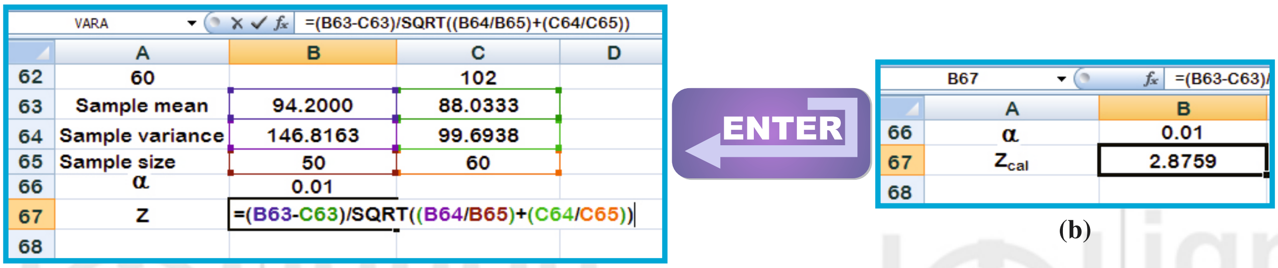


	A	B	C
64	Sample variance	146.8163	99.6938
65	Sample size	50	60
66	α	0.01	

Fig. 10.4

We use the test statistic $Z = \dfrac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}}$ because σ_1^2 and σ_2^2 are unknown and sample sizes are large.

Step 5: We compute the value of the test statistic Z using equation (2). Here we shall use Cell B67 for putting the value of Z_{cal}. Since the values of $\bar{X}$, $\bar{Y}$, S_1^2, S_2^2, n_1 and n_2 are given in Cells B63, C63, B64, C64, B65 and C65, respectively, we type "=(B63-C63)/Sqrt((B64/B65)+(C64/C65))" in Cell B67 and press ***Enter***. The value of Z_{cal} in Cell B67 is shown in Fig. 10.5b.



(a)

	VARA	f_x =(B63-C63)/SQRT((B64/B65)+(C64/C65))		
	A	B	C	D
62	60		102	
63	Sample mean	94.2000	88.0333	
64	Sample variance	146.8163	99.6938	
65	Sample size	50	60	
66	α	0.01		
67	Z	=(B63-C63)/SQRT((B64/B65)+(C64/C65))		
68				

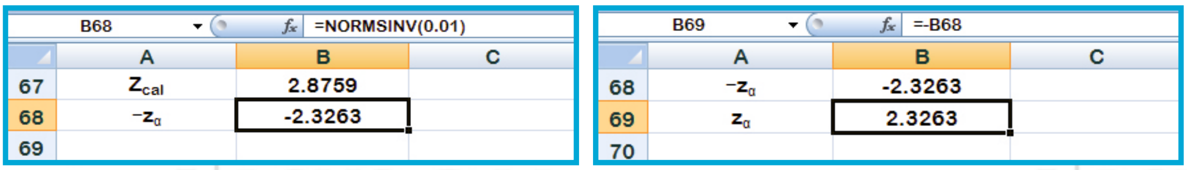
(b)

	B67	f_x =(B63-C63)/
	A	B
66	α	0.01
67	Z_{cal}	2.8759
68		

Fig. 10.5

Step 6: Decision using the critical region approach

We obtain the critical (tabulated) value for the test statistic Z at α = 0.01 in Cell B68 as explained in Step 6 under the heading 'Steps in Excel' of Sec. 9.3 of Lab Session 9. The outputs are shown in Figs. 10.6a and 10.6b.



(a)

	B68	f_x =NORMSINV(0.01)	
	A	B	C
67	Z_{cal}	2.8759	
68	$-z_\alpha$	-2.3263	
69			

(b)

	B69	f_x =-B68	
	A	B	C
68	$-z_\alpha$	-2.3263	
69	z_α	2.3263	
70			

Fig. 10.6

Conclusion

Since Z_{cal} (= 2.8759) > z_α (= 2.3263), it means that Z_{cal} lies in the rejection region as shown in Fig. 10.7. So we reject the null hypothesis. Since the alternative hypothesis is the claim, we do not reject the claim. Hence, we conclude that the samples do not provide us sufficient evidence against the claim. Thus, we may assume that the pulse rates of smokers are higher than the pulse rates of non-smokers at 1% level of significance.

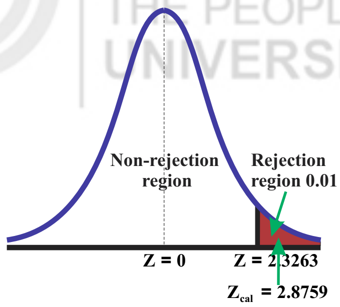


Fig. 10.7

Step 7: Decision using the p-value approach

Since the test is right-tailed, the p-value $= P[Z \geq Z_{cal}] = P[Z \geq 2.8759]$.

We calculate this in Cell B70 as explained in Step 8 under the heading

'Steps in Excel' of Sec. 9.5 of Lab Session 9. The outputs are shown in Figs. 10.8a and 10.8b.

B70 f_x =NORMSDIST(B67)

	A	B	C
69	z_α	2.3263	
70	$P[Z \le Z_{cal}]$	0.9980	
71			

(a)

B71 f_x =1-B70

	A	B	C
70	$P[Z \le Z_{cal}]$	0.9980	
71	p-value	0.0020	
72			

(b)

Fig. 10.8

Conclusion

We now compare the calculated p-value with the level of significance α. Since the p-value (= 0.0020) is less than (α = 0.01), we reject the null hypothesis.

10.3.1 Z-test using the Data Analysis ToolPak

We can also apply the Z-test for testing the difference of two population means with the help of the ***Data Analysis ToolPak*** using MS Excel 2007.

Step 1: We enter the data (given in Table 1) in Excel 2007 spreadsheet as explained in Step 1 of Sec. 10.3.

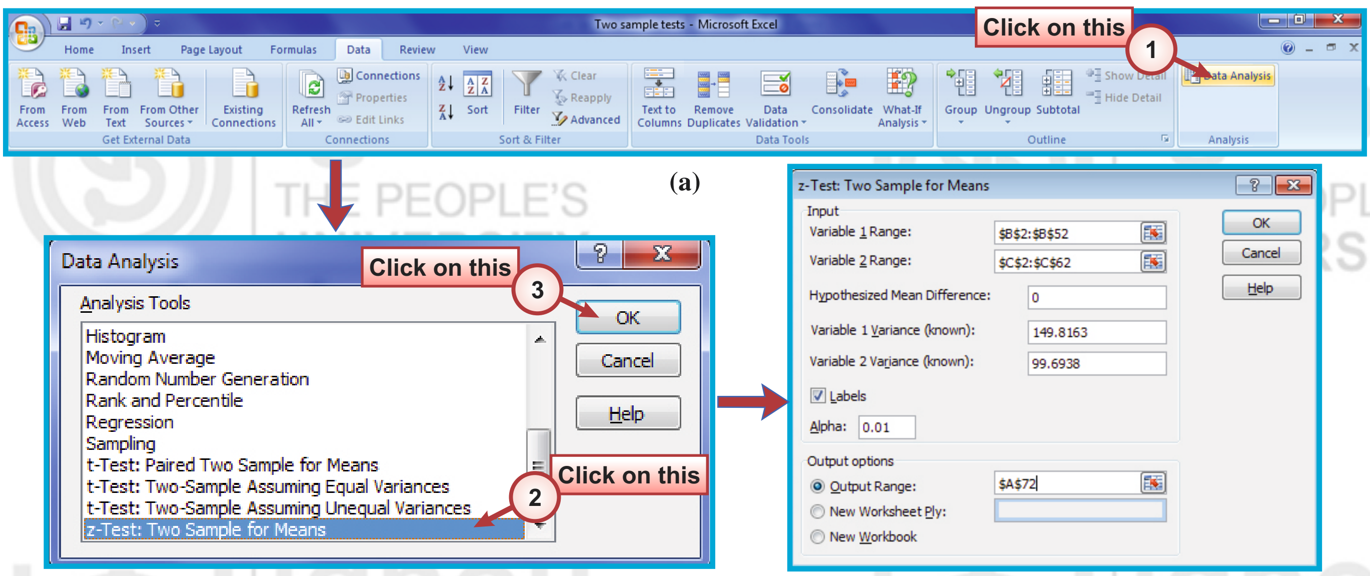
Step 2: In our problem, the population variances are unknown. So we first calculate sample variances in Cells B63 and C63 as explained in Step 3 of Sec. 10.3. The output is shown in Fig. 10.9.

	A	B	C
62	60		102
63	Sample variance	146.8163	99.6938
64			

Fig. 10.9

Step 3: We now apply the ***Data Analysis ToolPak*** for the Z-test as follows:

1. We click on the ***Data*** tab on the menu ribbon and then select the ***Data Analysis ToolPak*** as shown in Fig. 10.10a.
2. We select the ***z-Test: Two Sample for Means*** and then click on ***OK*** as shown in Fig. 10.10b.
3. We get a new dialog box as shown in Fig. 10.10c.



(a)

(b)

Fig. 10.10

(c)

4. We specify the data with label for smokers in ***Variable 1 Range***, i.e., Cells B2:B52.
5. We specify the data with label for non-smokers in ***Variable 2 Range***, i.e., Cells C2:C62.
6. We type 0 in ***Hypothesized Mean Difference***.
7. We type the values of sample variances for smokers and non-smokers in ***Variable 1 Variance (known)*** and ***Variable 2 Variance (known)***, respectively.
8. We click on the ***Labels*** box since we have included data labels, i.e., Smoker in Cell B2 and Non-smoker in Cell C2 in the variable ranges.
9. We type the value of level of significance (α) in ***Alpha***, i.e., 0.01.
10. We click on the ***Output Range*** and then select the cell on the Excel sheet where we wish to put the results. Here we select Cell A72.

Step 4: After completing Step 3, we obtain the results shown in Fig. 10.11.

	A	B	C
72	z-Test: Two Sample for Means		
73			
74		*Smoker*	*Non-smoker*
75	Mean	94.2000	88.0333
76	Known Variance	149.8163	99.6938
77	Observations	50	60
78	Hypothesized Me	0	
79	z	2.8573	
80	P(Z<=z) one-tail	0.0021	
81	z Critical one-tail	2.3263	
82	P(Z<=z) two-tail	0.0043	
83	z Critical two-tail	2.5758	

Fig. 10.11

It gives:

1. The mean, known variance and the number of observations for both samples of smokers and non-smokers in Cells B75, C75; B76, C76 and B77, C77, respectively.
2. The hypothesized mean in Cell B78 under which it gives the difference of population means. Here we take it as 0.
3. The value of Z_{cal} in Cell B79.
4. The critical values for one and two tails in Cells B81 and B82, respectively.
5. The p-values for one and two tails in Cells B80 and B82.

Step 5: Decision using the critical region approach

Here we use the one-tailed test. So we consider the critical value 2.3263 for one tail and compare it with the value of Z_{cal} (= 2.8759) for one tail. We draw the conclusion as explained in Step 6 of Sec. 10.3.

Step 6: Decision using the p-value approach

Here we use the one-tailed test. So we consider the p-value (= 0.0020) for one tail and compare it with the value of the level of significance α. We draw the conclusion as explained in Step 7 of Sec. 10.3.

10.4 t-TEST FOR THE DIFFERENCE OF TWO POPULATION MEANS

In Unit 11 of MST-004, you have learnt that the t-test for the difference of two population means is used when the population standard deviations (σ_1 and σ_2) are unknown, the populations under study are normal and the samples are independent and small. The procedure for this test has been described in Unit 11. We briefly mention the main steps as follows:

Step 1: We first formulate the null hypothesis (H_0) and alternative hypothesis (H_1) in the same way as discussed in Step 1 of the procedure for the Z-test in Sec. 10.3.

Step 2: We calculate the value of the test statistic t using the formula given below:

$$t = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{... (5)}$$

where S_p is the pooled sample variance given by

$$S_p^2 = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right] \qquad \text{... (6)}$$

Step 3: The other steps are the same as explained in the procedure for the t-test in Sec. 9.4 of Lab Session 9.

Steps in Excel

In Problem 2, the researcher wishes to find out whether the waiting time for a patient to meet a doctor in the emergency room at a government hospital is more than that at a private hospital. The waiting time of both hospitals is normally distributed and the variances of both waiting time distributions are unknown and equal. So we can use the t-test for the difference of two population means.

For applying the t-test, we first have to set up the null and alternative hypotheses. Let μ_1 and μ_2 denote the average (mean) waiting time for a patient to meet a doctor in the emergency room at the government and private hospitals, respectively. Then we formulate the null and alternative hypotheses as follows:

$$H_0 : \mu_1 \le \mu_2 \text{ and } H_1 : \mu_1 > \mu_2 \text{ (claim)} \qquad [\text{right-tailed test}]$$

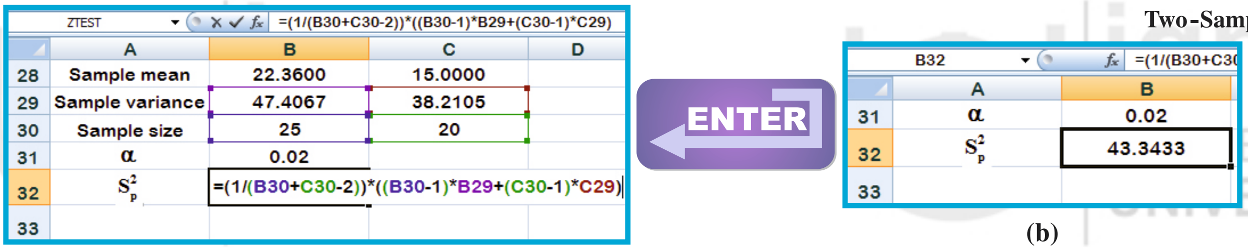
We follow the steps given below to apply the t-test for testing the difference of two population means using Excel 2007:

Step 1: We repeat Steps 1-4 under the heading 'Steps in Excel' of Sec.10.3 to enter the data given in Table 2, calculate the sample means and variances, and type the values of sample sizes and level of significance α. We get the output shown in Fig. 10.12.

	A	B	C
27	25	25	
28	Sample mean	22.3600	15.0000
29	Sample variance	47.4067	38.2105
30	Sample size	25	20
31	α	0.02	
32			

Fig. 10.12

Step 2: We now compute the value of the pooled variance S_p^2 using equation (6). Here we shall use Cell B32 for putting the value of S_p^2. Since the values of n_1, n_2, S_1^2 and S_2^2 are given in Cells B30, C30, B29 and C29, respectively (see Fig. 10.12), we type "=(1/(B30+C30-2))*((B30-1)*B29+(C30-1)*C29)" in Cell B32 (see Fig. 10.13a) and press ***Enter***. Then we get the value of S_p^2 in Cell B32 (Fig. 10.13b).



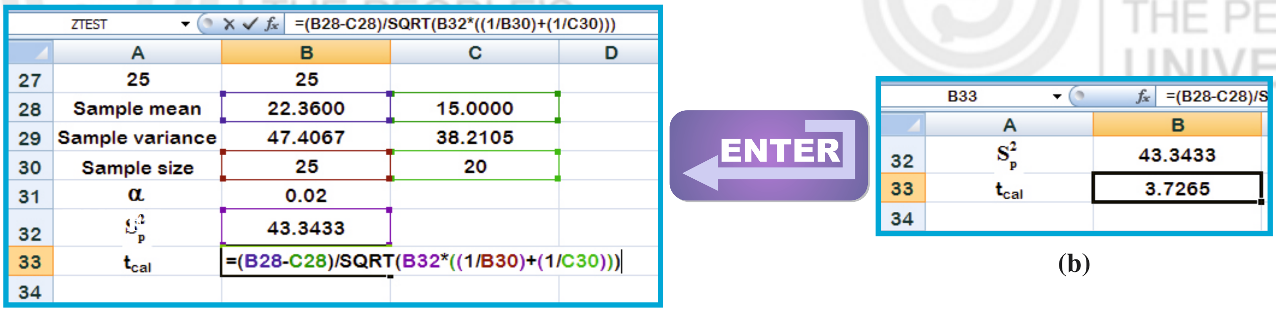
(a) (b)

Fig.10.13

Step 3: We compute the value of the test statistic t using equation (5). Here we shall use Cell B33 for putting the value of the test statistic. Since the values of $\bar{X}$, $\bar{Y}$, S_p^2, n_1 and n_2 are given in Cells B28, C28, B32, B30 and C30, respectively, we type "=(B28-C28)/Sqrt(B32*((1/B30)+(1/C30)))" in Cell B33 (see Fig. 10.14a) and press ***Enter***. Then we get the value of t_{cal} in Cell B33 (Fig. 10.14b).

The test statistic is

$$t = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



(a) (b)

Fig. 10.14

Step 4: Decision using the critical region approach

We now obtain the critical (tabulated) value with the help of Excel as follows:

1. We first compute the value of 2α (since Excel calculates the critical value on two-tails) in Cell B34 by typing "=2*B31" and press ***Enter***. We get the output shown in Fig. 10.15.

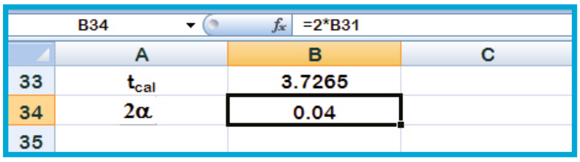


Fig. 10.15

2. We compute the degrees of freedom $(n_1 + n_2 - 2)$ in Cell B35 by typing "=B30+C30-2" and pressing ***Enter***. We get the df in Cell B35 (Fig. 10.16).

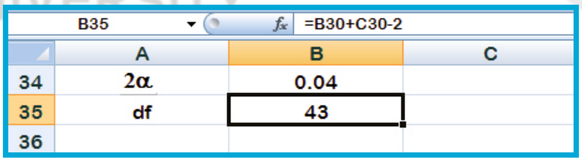


Fig. 10.16

3. We now compute the critical (tabulated) value in Cell B36 as explained in Step 4 under the heading 'Steps in Excel' of Sec. 9.4 of Lab Session 9. We take $2\alpha = 0.04$ instead of α (since here the test is one-tailed) and get the value of t_α (Fig. 10.17).

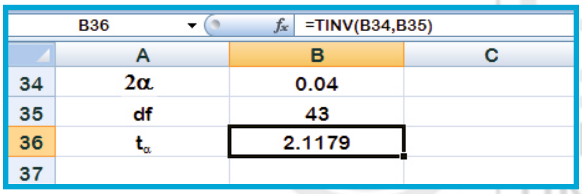


Fig. 10.17

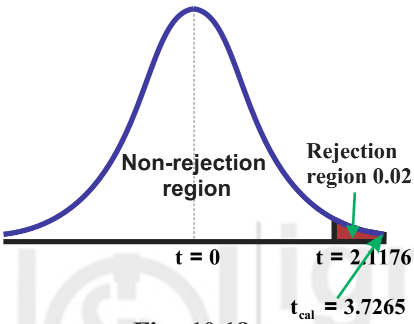


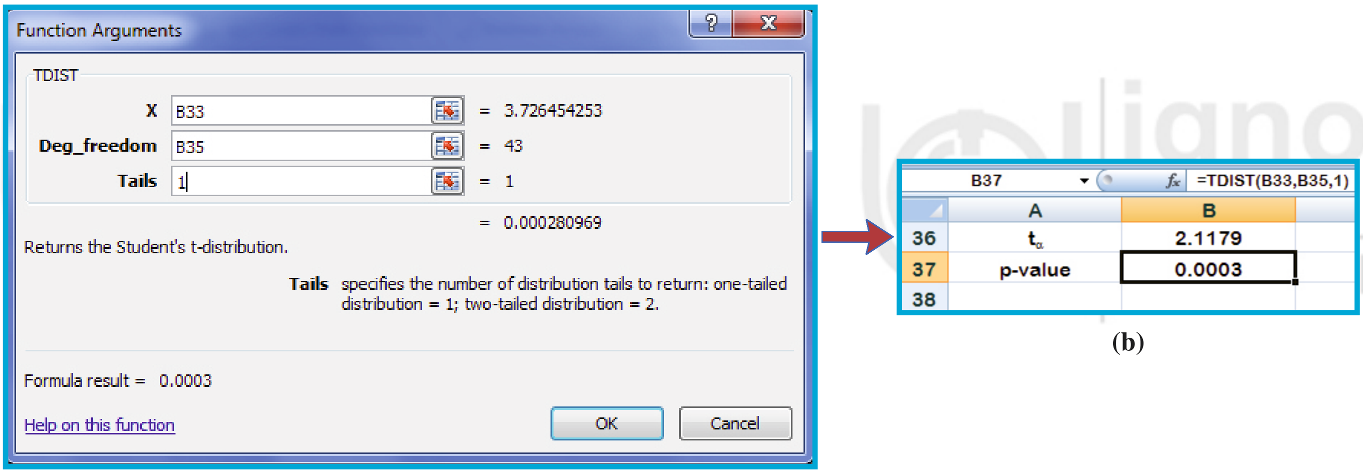
Fig. 10.18

Conclusion

We now take decision about the null hypothesis. Since t_{cal} (= 3.7265) > t_{α} (= 2.1179), it means that t_{cal} lies in the rejection region as shown in Fig. 10.18. So we reject the null hypothesis. Since the alternative hypothesis is the claim, we do not reject the claim. Hence, we conclude that the samples do not provide us sufficient evidence against the claim. So we may assume that the average waiting time for a patient to meet a doctor in the emergency room at a government hospital is more than that at a private hospital at 2% level of significance.

Step 5: Decision using the p-value approach

Since the test is right one-tailed, the p-value = $P[t \geq t_{cal}] = P[t \geq 3.7265]$. We calculate the p-value as explained in Step 5 under the heading 'Steps in Excel' of Sec. 9.4 of Lab Session 9. The output obtained is shown in Figs. 10.19a and 10.19b.



(b)

(a)

Fig. 10.19

Conclusion

We now take the decision about the null hypothesis on the basis of the p-value. Since the p-value (= 0.0003) is less than α (= 0.02), we reject the null hypothesis.

10.4.1 t-test using Data Analysis ToolPak

We can also apply the t-test for testing the difference of two population means with the help of the ***Data Analysis ToolPak*** in MS Excel 2007.

Step 1: We enter the data (given in Table 2) in Excel 2007 spreadsheet as explained in Step 1 under the heading 'Steps in Excel' of Sec. 10.3.

Step 2: We now apply the ***Data Analysis ToolPak*** for the t-test as follows:

1. We click on ***Data*** → ***Data Analysis*** → ***t-Test: Two-sample Assuming Equal Variances*** → ***OK***. Then we get a new dialog box as shown in Fig. 10.20.

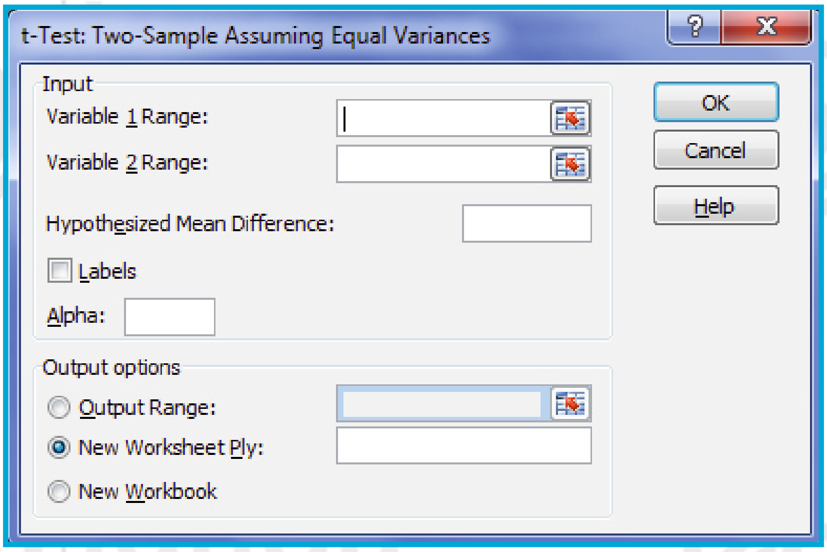
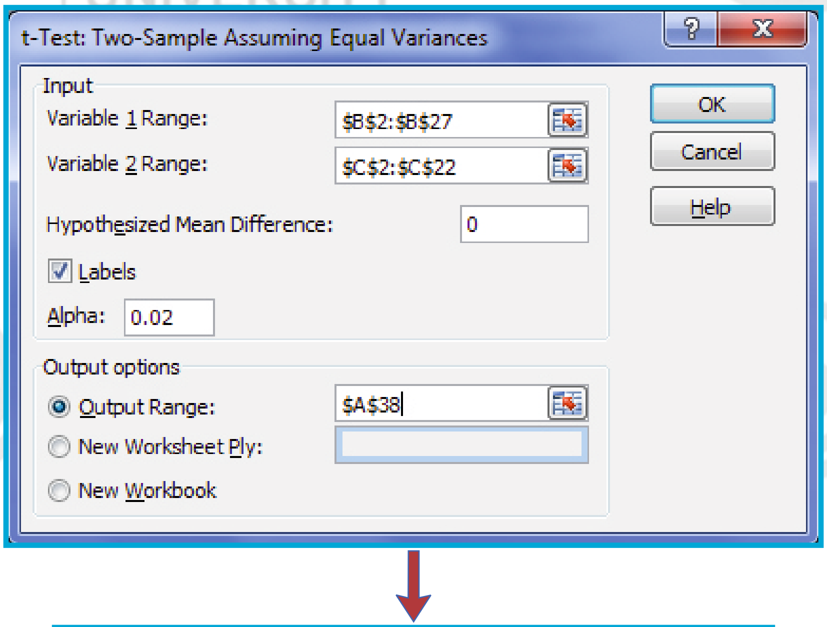


Fig. 10.20

2. We put the values as explained in Sec 10.3.1 (see Fig. 10.21a) and press ***OK***. We get the output shown in Fig. 10.21b.



(a)

38	t-Test: Two-Sample Assuming Equal Variances		
39			
40		*Goverentment*	*Private*
41	Mean	22.3600	15.0000
42	Variance	47.4067	38.2105
43	Observations	25	20
44	Pooled Variance	43.3433	
45	Hypothesized Mean Difference	0.0000	
46	df	43.0000	
47	t Stat	3.7265	
48	P(T<=t) one-tail	0.0003	
49	t Critical one-tail	2.1179	
50	P(T<=t) two-tail	0.0006	
51	t Critical two-tail	2.4163	

(b)

Fig. 10.21

Step 3: The interpretation of the results of the ***Data Analysis ToolPak*** is the same as given in Sec. 10.3.1.

Step 4: We take the decision about the null hypothesis/claim using the critical region approach and the p-value approach as explained in Steps 4 and 5 under the heading ‘Steps in Excel’ of Sec. 10.4, respectively.

10.5 PAIRED t-TEST

In Sec. 10.4, you have learnt how to apply the t-test for the difference of two population means in MS Excel 2007. This t-test is applied when the samples are independent. In Unit 11 of MST-004, you have learnt that in many situations, the samples are not independent while the observations are recorded on the same individuals or items. Generally, the observations are recorded **before and after** the insertion of training, treatment, etc. For example, if a dietician wishes to test a new diet on some individuals, then the weight of the individuals before and after administering the diet will form two different samples in which observations will be paired for each individual. Similarly, in the test of blood-sugar in a patient, both the fasting sugar level and the sugar level after a meal are recorded. These are paired observations. In such situations, the paired t-test is used for testing the difference of the two population means when the samples are not independent.

The procedure for the paired t-test has been described in Unit 11 of MST-004. We briefly mention the main steps as follows:

Step 1: We first formulate the null hypothesis (H_0) and alternative hypothesis (H_1). If μ_1 and μ_2 are the means before and after the training, treatment, etc., respectively, we can formulate the null and alternative hypotheses as follows:

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 \neq \mu_2 \qquad [\text{for two-tailed test}]$$

$$\text{or} \quad \left.\begin{array}{l} H_0 : \mu_1 \leq \mu_2 \text{ and } H_1 : \mu_1 > \mu_2 \\ H_0 : \mu_1 \geq \mu_2 \text{ and } H_1 : \mu_1 < \mu_2 \end{array}\right\} \qquad [\text{for one-tailed test}]$$

We can also formulate the null and alternative hypotheses as follows:

$$H_0 : \mu_D = 0 \text{ and } H_1 : \mu_D \neq 0 \qquad [\text{for two-tailed test}]$$

$$\text{or} \quad \left.\begin{array}{l} H_0 : \mu_D \leq 0 \text{ and } H_1 : \mu_D > 0 \\ H_0 : \mu_D \geq 0 \text{ and } H_1 : \mu_D < 0 \end{array}\right\} \qquad [\text{for one-tailed test}]$$

where $\mu_D = \mu_1 - \mu_2$.

Step 2: We calculate the value of the test statistic t using the formula given below:

$$t = \frac{\bar{D}}{S_D / \sqrt{n}} \qquad \ldots (7)$$

where $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ and $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2$

Step 3: The other steps are the same as explained in the procedure for the t-test in Sec. 9.4 of Lab Session 9.

Steps in Excel

In Problem 3, we have to test whether practice improves the memory and performance of the children. Here the observations are recorded before and after practice and the memories of the children before and after practice follow normal distributions. So we can apply the paired t-test for this problem.

For applying the paired t-test, we first have to set up the null and alternative hypotheses. If μ_1 and μ_2 denote the average (mean) of digits repetition before and after practice, respectively, we can take the null and alternative hypotheses as follows:

$$H_0: \mu_1 \geq \mu_2 \text{ and } H_1: \mu_1 < \mu_2 \text{ (claim)} \quad [\text{left-tailed test}]$$

or $$H_0: \mu_D \geq 0 \text{ and } H_1: \mu_D < 0 \text{ where } \mu_D = \mu_1 - \mu_2$$

We apply the paired t-test in Excel 2007 as follows:

Step 1: We enter the data (given in Table 3) in Excel 2007 spreadsheet as shown in Fig. 10.22.

	A	B	C	D
1	Child Number	Recall Before (X)	Recall After (Y)	
2	1	6	8	
3	2	4	4	
4	3	5	8	
5	4	8	8	
6	5	4	4	
7	6	2	5	
8	7	3	5	
9	8	5	5	
10	9	2	6	
11	10	7	8	

Fig. 10.22: Partial screenshot of the spreadsheet for the given data.

Step 2: We compute the difference D = X – Y in Cell D2 by typing "=B2-C2" and then drag it down up to the last sample, i.e., Cell D21. In this way, we get the difference up to the last sample observation as shown in Fig. 10.23.

D2 fx =B2-C2

	A	B	C	D	E
1	Child Number	Recall Before (X)	Recall After (Y)	Difference (D = X-Y)	
2	1	6	8	-2	
3	2	4	4		
4	3	5	8		
5	4	8	8		

(a)

DRAG IT DOWN

	A	B	C	D	E
1	Child Number	Recall Before (X)	Recall After (Y)	Difference (D = X-Y)	
2	1	6	8	-2	
3	2	4	4	0	
4	3	5	8	-3	
5	4	8	8	0	
6	5	4	4	0	
7	6	2	5	-3	
8	7	3	5	-2	
9	8	5	5	0	
10	9	2	6	-4	
11	10	7	8	-1	
12	11	8	8	0	
13	12	5	6	-1	
14	13	5	7	-2	
15	14	4	4	0	
16	15	6	8	-2	
17	16	4	4	0	
18	17	2	3	-1	
19	18	2	4	-2	
20	19	6	6	0	

(b)

Fig. 10.23

Step 3: We compute the mean ($\bar{D}$) and standard deviation (S_D) of D as explained in Step 2 of Sec. 9.3 and Step 1 of Sec. 9.4 of Lab Session 9 in Cell D22 and D23, respectively (Figs. 10.24a and b).

(a)

D22 | f_x =AVERAGE(D2:D21)

	A	B	C	D
21	20	4	5	-1
22	$\bar{D}$			-1.2000
23				

(b)

D23 | f_x =STDEV(D2:D21)

	A	B	C	D
22	$\bar{D}$			-1.2000
23	s_D			1.2397
24				

Fig. 10.24

Step 4: We now type the values of the sample size n and the level of significance α in Cells D24 and D25, respectively (Fig. 10.25).

	A	B	C	D
23	s_D			1.2397
24	n			20
25	α			0.05
26				

Fig. 10.25

Step 5: We compute the value of the test statistic t using equation (7) in Cell D26. Since the values of $\bar{D}$, S_D and n are given in Cells D22, D23 and D24, respectively, we type "=D22/(D23/Sqrt(D24))" in Cell D26 and press ***Enter*** (see Fig. 10.26a). Then we get the value of t_{cal} in Cell D26 (Fig. 10.26b).

The test statistic for the paired t-test is

$$t = \frac{\bar{D}}{S_D / \sqrt{n}}$$

(a)

IF | f_x =D22/(D23/SQRT(D24))

	A	B	C	D
20	19	6	6	0
21	20	4	5	-1
22	$\bar{D}$			-1.2000
23	s_D			1.2397
24	n			20
25	α			0.05
26	t_{cal}			=D22/(D23/ SQRT(D24))
27				
28				

(b)

D26 | f_x =D22/(D23/SQRT(D24))

	A	B	C	D
24	n			20
25	α			0.05
26	t_{cal}			-4.3289
27				

Fig. 10.26

Step 6: Decision using the critical region approach

We now obtain the critical (tabulated) value with the help of Excel as follows:

1. We first compute the value of 2α (since Excel calculates the critical value on two-tails) in Cell D27 by typing "=2*D25" (see Fig. 10.27a) and then the degree of freedom (n – 1) in Cell D28 by typing "=D24-1". We get the output shown in Fig. 10.27b.

D27 | f_x =2*D25

	A	B	C	D
26	t_{cal}			-4.3289
27	2α			0.10
28				

(a)

D28 | f_x =D24-1

	A	B	C	D
27	2α			0.10
28	df			19
29				

(b)

Fig. 10.27

2. We now compute the critical (tabulated) value in Cell D29 as explained in Step 4 under the heading 'Steps in Excel' of Sec. 9.4 of Lab Session 9 by taking 2α instead of α since here the test is one tailed. We get the value of t_α as shown in Fig. 10.28.

D29 | f_x =TINV(D27,D28)

	A	B	C	D
28	df			19
29	$t_{(19),0.05}$			1.7291
30				

Fig. 10.28

3. Since our test is left one-tailed, we calculate the critical value on the left-tail by typing "=-D29" in Cell D30 (Fig. 10.29).

D30 | f_x =-D29

	A	B	C	D
29	$t_{(19),0.05}$			1.7291
30	$-t_{(19),0.05}$			-1.7291
31				

Fig. 10.29

Conclusion

We now take decision about the null hypothesis. Since, t_{cal} (= – 4.3289) < $-t_{(19),0.05}$ (= – 1.7291), it means that it lies in the rejection region as shown in Fig. 10.30. So we reject the null hypothesis. Since the alternative hypothesis is the claim, we do not reject the claim. Hence, we conclude that the samples do not provide us sufficient evidence against the claim. So we may assume that practice improves the memory and performance of children at 5% level of significance.

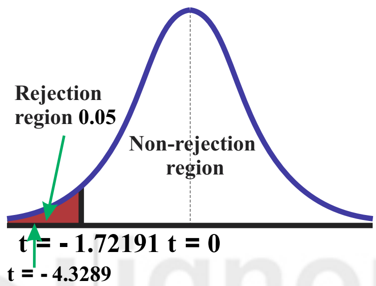
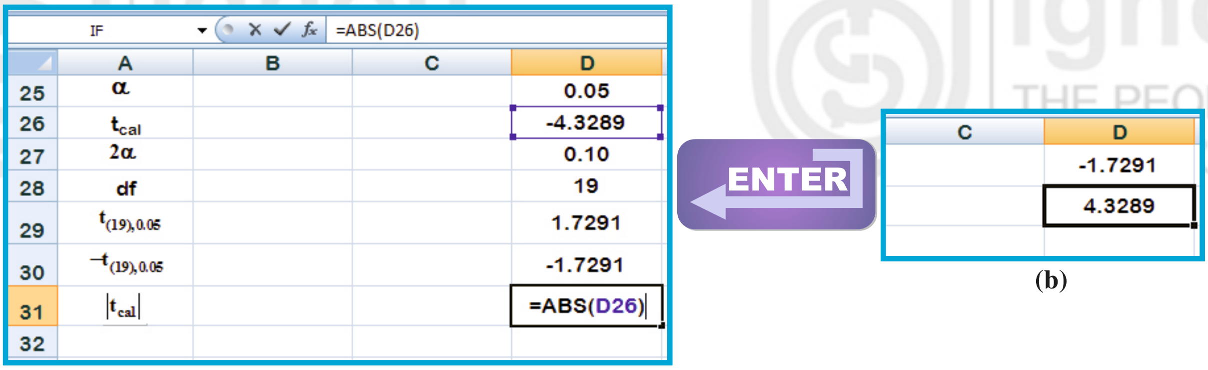


Fig. 10.30

Step 7: Decision using the p-value approach

Since the test is one-tailed, the p-value = $P[t \le t_{cal}] = P[t \le -4.3289]$. Since Excel calculates $P[t \ge t_{cal}]$ when t_{cal} is positive and we know that the t-distribution is symmetrical about $t = 0$, the probability $P[t \le -a] = P[t \ge a]$. Therefore, p-value $= P\left[t \ge |t_{cal}|\right]$. So we first calculate $|t_{cal}|$ and then the p-value as follows:

1. We calculate $|t_{cal}|$ in Cell 31 by typing "=Abs(B26)" and pressing ***Enter***. We get the output shown in Fig. 10.31.

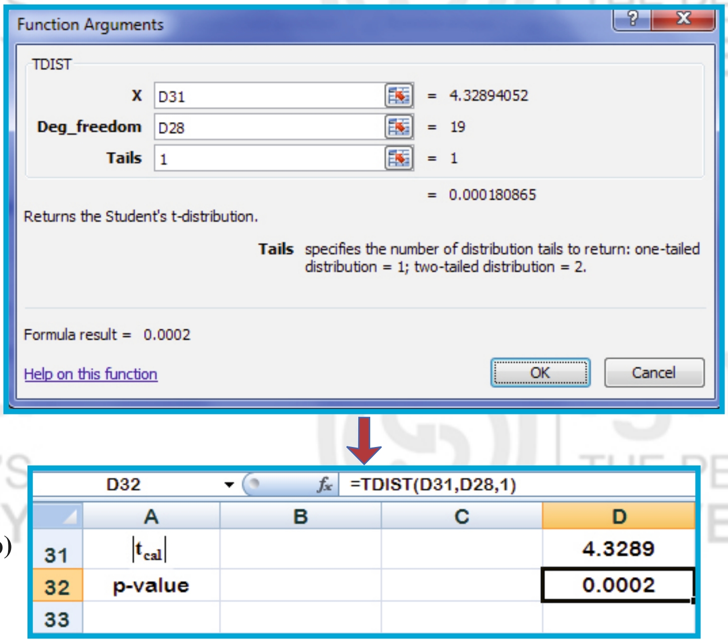


(a)

(b)

Fig. 10.31

2. We now calculate the p-value ($= P\left[t \geq \left|t_{cal}\right|\right]$) as explained in Step 5 under the heading 'Steps in Excel' of Sec. 9.4 of Lab Session 9 (see Fig. 10.32a). The output is shown in Fig. 10.32b.



(a)

(b)

Fig. 10.32

Conclusion

We now take decision about the null hypothesis on the basis of the p-value. Since the p-value (= 0.0002) is less than α (= 0.05), we reject the null hypothesis.

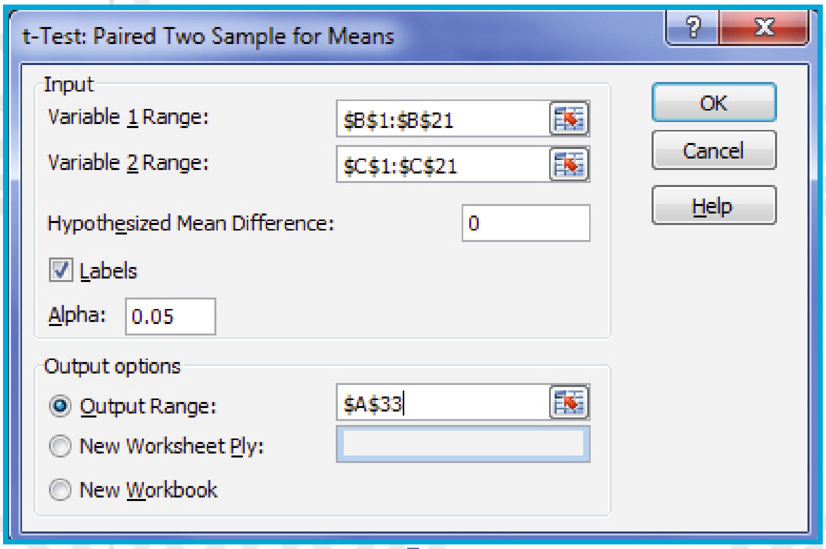
10.5.1 Paired t-test using the Data Analysis ToolPak

We can also apply the paired t-test with the help of the ***Data Analysis ToolPak*** in MS Excel 2007.

Step 1: We enter the data (given in Table 3) in Excel 2007 spreadsheet as explained in Step 1 under the heading 'Steps in Excel' of Sec. 10.5.

Step 2: We now apply the ***Data Analysis ToolPak*** for paired t-test as follows:

1. We click on ***Data*** → ***Data Analysis*** → ***t-Test: Two-sample Assuming Equal Variances*** → ***OK***. Then we get a new dialog box. We put the values as explained in Step 3 under the heading 'Steps in Excel' of Sec. 10.3.1 (Fig. 10.33a) and press ***OK***. We get the output shown in Fig. 10.33b.



(a)

33	t-Test: Paired Two Sample for Means		
34			
35		*Recall Before (X)*	*Recall After (Y)*
36	Mean	4.6000	5.8000
37	Variance	3.5158	3.0105
38	Observations	20	20
39	Pearson Correlation	0.7668	
40	Hypothesized Mean Difference	0	
41	df	19	
42	t Stat	-4.3289	
43	P(T<=t) one-tail	0.0002	
44	t Critical one-tail	1.7291	
45	P(T<=t) two-tail	0.0004	
46	t Critical two-tail	2.0930	

(b)

Fig. 10.33

Step 3: The interpretation of the results of the ***Data Analysis ToolPak*** is the same as explained in Sec. 10.3.1.

Step 4: We take decision about the null hypothesis/claim using the critical region approach and the p-value approach as explained in Steps 6 and 7, respectively, of Sec. 10.5.

Activity 1

Apply the t-test on Problem 1 as explained in Sec. 10.4.

10.6 Z-TEST FOR THE DIFFERENCE OF TWO POPULATION PROPORTIONS

In Unit 10 of MST-004, you have learnt that in many situations we are interested in testing the hypothesis about the difference of the proportions of an attribute in two different populations or groups. For example, we may wish to test whether the proportion of teachers in the populations of two cities is the same, or whether the proportion of literates in a village is greater than that in another village, and so on. In such situations, we apply the Z-test for the difference of two population proportions.

The procedure for the Z-test for the difference of two population proportions has been described in Unit 10 of MST-004. We briefly mention the main steps as follows:

Step 1: We first formulate the null hypothesis (H_0) and alternative hypothesis (H_1). If P_1 and P_2 are the proportions of an attribute in population-I and population-II, respectively, we can formulate the null and alternative hypotheses as follows:

$$H_0 : P_1 = P_2 = P \text{ and } H_1 : P_1 \neq P_2 \qquad [\text{for two-tailed test}]$$

or

$$\left. \begin{array}{l} H_0 : P_1 \leq P_2 \text{ and } H_1 : P_1 > P_2 \\ H_0 : P_1 \geq P_2 \text{ and } H_1 : P_1 < P_2 \end{array} \right\} \qquad [\text{for one-tailed test}]$$

Step 2: We calculate the value of the test statistic Z using the formula given below:

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \ldots (8)$$

where

$p_1 = \frac{X_1}{n_1}$ – the first sample proportion drawn from population-I,

$p_2 = \frac{X_2}{n_2}$ – the second sample proportion drawn from population-II,

X_1 and X_2 – the number of observations/individuals/items/units that possess the given attribute in the samples of sizes n_1 and n_2, respectively.

If P is unknown, the test statistic Z is given as

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \ldots (9)$$

where $\hat{P}$ is the pooled proportion given as

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} \text{ and } \hat{Q} = 1 - \hat{P} \qquad \ldots (10)$$

Step 3: The other steps are the same as explained for the procedure of the Z-test in Sec. 9.3 of Lab Session 9.

Steps in Excel

We apply the Z-test for the difference of two population proportions for Problem 4. We have to set up the null hypothesis (H_0) and alternative hypothesis (H_1) first. Here we have to test whether there is a significant difference in the proportions of male and female customers in the population stating that 100% cotton cloth matches their life styles. If P_1 and P_2 denote the proportions of male and female customers in the population stating that 100% cotton cloth matches their life styles, respectively, we can formulate the null and alternative hypotheses as follows:

$$H_0 : P_1 = P_2 \text{ and } H_1 : P_1 \neq P_2 \text{ (claim)} \qquad [\text{two-tailed test}]$$

The steps for applying the Z-test for testing the difference of two population proportions in Excel 2007 are as follows:

Step 1: In this problem, the attribute is that 100% cotton cloth matches the customers' lifestyle. Since the values of X_1 and X_2 are given, we type the values of n_1, n_2, X_1, X_2 and α in Cells B2, C2, B3, C3 and B4, respectively, as shown in Fig. 10.34.

	A	B	C
1		Male	Female
2	Sample size	120	100
3	Responded 100% cotton	80	60
4	α	0.05	
5			

Fig. 10.34

Step 2: We now calculate the values of p_1 $(= X_1/n_1)$ and p_2 $(= X_2/n_2)$ in Cells B5 and C5 by typing "=B3/B2" in Cell B5 and "=C3/C2" in Cell C5, respectively. The outputs are shown in Figs. 10.35a and b.

B5 f_x =B3/B2

	A	B
4	α	0.05
5	Sample proportion	0.6667
6		

(a)

C5 f_x =C3/C2

	A	B	C
4	α	0.05	
5	Sample proportion	0.6667	0.6000
6			

(b)

Fig. 10.35

Step 3: We calculate the value of $\hat{P}\left(=\frac{X_1+X_2}{n_1+n_2}\right)$ in Cell B6 by typing "=(B3+C3)/(B2+C2)" and the value of $\hat{Q}\,(=1-\hat{P})$ in Cell B7 by typing "=1-B6". We get the outputs shown in Figs. 10.36a and b.

B6 f_x =(B3+C3)/(B2+C2)

	A	B
5	Sample proportion	0.6667
6	$\hat{P}$	0.6364
7		

(a)

B7 f_x =1-B6

	A	B
5	Sample proportion	0.6667
6	$\hat{P}$	0.6364
7	$\hat{Q}$	0.3636

(b)

Fig. 10.36

Step 4: We now calculate the value of the test statistic Z using equation (9). Here we shall use Cell B8 for putting the value of the test statistic. Since the values of p_1, p_2, $\hat{P}$, $\hat{Q}$, n_1 and n_2 are given in Cells B5, C5, B6, B7, B2 and C2, respectively, we type "=(B5-C5)/Sqrt(B6*B7*((1/B2)+(1/C2)))" in Cell B8 (see Fig. 10.37a) and press ***Enter***. The result is shown in Fig. 10.37b.

TDIST f_x =(B5-C5)/SQRT(B6*B7*((1/B2)+(1/C2)))

	A	B	C	D
1		Male	Female	
2	Sample size	120	100	
3	Responded 100% cotton	80	60	
4	α	0.05		
5	Sample proportion	0.6667	0.6000	
6	$\hat{P}$	0.6364		
7	$\hat{Q}$	0.3636		
8	Z_{cal}	=(B5-C5)/SQRT(B6*B7*((1/B2)+(1/C2)))		

(a)

	A	B
7	$\hat{Q}$	0.3636
8	Z_{cal}	1.0235

(b)

Fig. 10.37

Step 5: Decision using the critical region approach

We now obtain the critical values with the help of Excel as follows:

1. Since the test is two-tailed, we first compute the value of $\alpha/2$ in Cell B9 by typing "=B4/2" (Fig. 10.38).

B9 | f_x =B4/2

	A	B	C
8	Z_{cal}	1.0235	
9	$\alpha/2$	0.0250	
10			

Fig. 10.38

2. We calculate the critical values $-z_{\alpha/2}$ and $z_{\alpha/2}$ as explained in Step 6 under the heading 'Steps in Excel' in Sec. 9.3 of Lab Session 9. We get the output shown in Fig. 10.39.

(a)

B109 | f_x =NORMSINV(B105)

	A	B	C
108	Z_{cal}	0.75	
109	$-z_\alpha$	-2.3263	
110			

(b)

B110 | f_x =-B109

	A	B	C
109	$-z_\alpha$	-2.3263	
110	z_α	2.3263	
111			

Fig. 10.39

Conclusion

Since $-z_{\alpha/2}$ (= – 1.96) < Z_{cal} (= 1.0235) < $z_{\alpha/2}$ (= 1.96), it means that Z_{cal} lies in the non-rejection region as shown in Fig. 10.40. So we do not reject the null hypothesis. Since our claim is the alternative hypothesis, we reject the claim. Hence, we conclude that the samples provide us sufficient evidence against the claim. So we may assume that there is a significant difference in the proportion of male and female customers in the population stating that 100% cotton cloth matches their lifestyles at 5% level of significance.

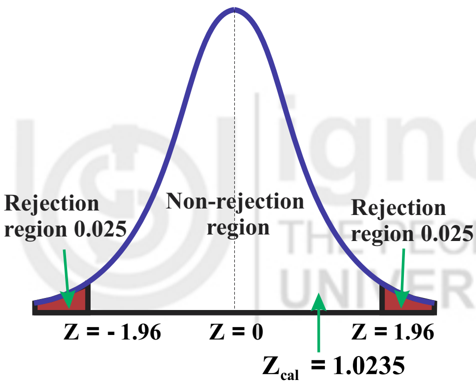


Fig. 10.40

Step 6: Decision using the p-value approach

Since the test is two-tailed, p-value $= 2P\left[Z \geq |Z_{cal}|\right] = 2P\left[Z \geq |-1.0235|\right]$.

We calculate the p-value as explained in Step 7 under the heading 'Steps in Excel' in Sec. 9.3 of Lab Session 9. We get the output shown in Figs. 10.41a and b.

B12 | f_x =NORMSDIST(B8)

	A	B
11	$z_{\alpha/2}$	1.9600
12	$P\left[Z \leq |Z_{cal}|\right]$	0.8470
13		

(a)

B13 | f_x =2*(1-B12)

	A	B
12	$P\left[Z \leq |Z_{cal}|\right]$	0.8470
13	p-value	0.3061
14		

(b)

Fig. 10.41

Conclusion

We now compare the calculated p-value with the level of significance α. Since the p-value (= 0.3061) is greater than α = (0.05), we do not reject the null hypothesis.

10.7 F-TEST FOR TWO POPULATION VARIANCES

In Unit 12 of MST-004, you have learnt that the t-test is based on the assumption that the variances of the two populations are equal. To check this assumption, we use the F-test for two population variances. The F-test is also important in a number of contexts. For example, an economist may wish to test whether the variability in income differs in two populations, a quality controller may wish to test whether the quality of the product is changing over time, etc.

The F-test is used when the populations under study are normal. The procedure for this test is described in Unit 12. We briefly mention the main steps as follows:

Step 1: We first formulate the null hypothesis (H_0) and the alternative hypothesis (H_1). If σ_1^2 and σ_2^2 are the variances of population-I and population-II, respectively, we can formulate the null and alternative hypotheses as follows:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ and } H_1: \sigma_1^2 \neq \sigma_2^2 \quad [\text{for two-tailed test}]$$

or

$$\left.\begin{array}{l} H_0: \sigma_1^2 \leq \sigma_2^2 \text{ and } H_1: \sigma_1^2 > \sigma_2^2 \\ H_0: \sigma_1^2 \geq \sigma_2^2 \text{ and } H_1: \sigma_1^2 < \sigma_2^2 \end{array}\right\} \quad [\text{for one-tailed test}]$$

Step 2: We calculate the value of the test statistic F using the formula given below:

$$F = \frac{S_1^2}{S_2^2} \qquad \ldots (11)$$

where S_1^2 and S_2^2 are the variances of the samples selected from population-I and population-II, respectively, and are given by

$$S_1^2 = \frac{1}{n_1 - 1}\sum\left(X - \bar{X}\right)^2 \text{ and } S_2^2 = \frac{1}{n_2 - 1}\sum\left(Y - \bar{Y}\right)^2 \qquad \ldots (12)$$

Step 3: We obtain the critical (cut-off or tabulated) value(s) of the test statistic F corresponding to the given level of significance (α).

Step 4: We take the decision about the null hypothesis as follows:

i) Using the critical region approach

We compare the calculated value of the test statistic (F_{cal}) with the critical values obtained in Steps 2 and 3, respectively. Since critical values depend upon the nature of the test (that is, whether it is one-tailed test or two-tailed), the following cases arise:

Case of two-tailed test:

$H_0: \sigma_1^2 = \sigma_2^2$ and $H_1: \sigma_1^2 \neq \sigma_2^2$

Let $F_{(\nu_1, \nu_2), (1-\alpha/2)}$ and $F_{(\nu_1, \nu_2), \alpha/2}$ be the two critical values on the left-tail and right-tail, respectively, on a pre-fixed level of significance (see Fig. 10.42). If $F_{cal} \geq F_{(\nu_1, \nu_2), \alpha/2}$ or $F_{cal} \leq F_{(\nu_1, \nu_2), (1-\alpha/2)}$, we reject H_0 and if $F_{(\nu_1, \nu_2), (1-\alpha/2)} < F_{cal} < F_{(\nu_1, \nu_2), \alpha/2}$, we do not reject H_0.

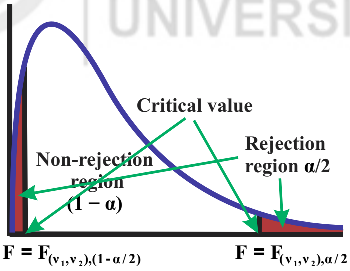


Fig. 10.42

Case of right-tailed test:

$H_0: \sigma_1^2 \leq \sigma_2^2$ and $H_1: \sigma_1^2 > \sigma_2^2$

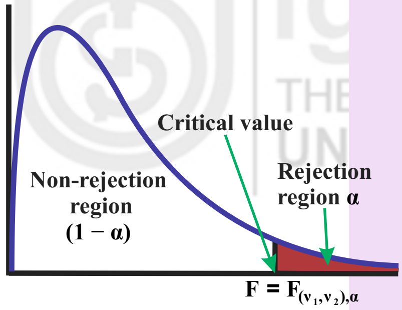


Fig. 10.43

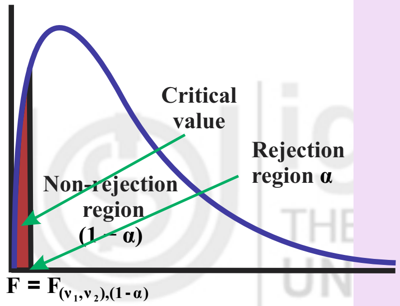


Fig. 10.44

Let $F_{(\nu_1,\nu_2),\alpha}$ be the critical value at α % level of significance which lies in the right-tail in this case (see Fig. 10.43). If $F_{cal} \geq F_{(\nu_1,\nu_2),\alpha}$, we reject H_0 and if $F_{cal} < F_{(\nu_1,\nu_2),\alpha}$, we do not reject H_0.

Case of left-tailed test:

$H_0 : \sigma_1^2 \geq \sigma_2^2$ and $H_1 : \sigma_1^2 < \sigma_2^2$

Let $F_{(\nu_1,\nu_2),(1-\alpha)}$ be the critical value at α% level of significance which lies in the left tail in this case (see Fig. 10.44). If $F_{cal} \leq F_{(\nu_1,\nu_2),(1-\alpha)}$, we reject H_0 and if $F_{cal} > F_{(\nu_1,\nu_2),(1-\alpha)}$, we do not reject H_0.

ii) Using the p-value value approach

We calculate the p-value using the following formulae as required:

$$\text{p-value} = 2P[F \geq F_{cal}] \qquad [\text{for two-tailed test}] \qquad \ldots (13)$$

$$\text{p-value} = P[F \geq F_{cal}] \qquad [\text{for right-tailed test}] \qquad \ldots (14)$$

$$\text{p-value} = P[F \leq F_{cal}] \qquad [\text{for left-tailed test}] \qquad \ldots (15)$$

We compare the calculated p-value with given level of significance (α). If p-value is less than or equal to α, we reject the null hypothesis and if it is greater than α, we do not reject the null hypothesis.

Step 5: Conclusion

We draw the conclusion as discussed in Step 5 of the procedure for the Z-test in Sec. 9.3 of Lab Session 9.

Steps in Excel

In Problem 5, we have to test whether the variances of waiting time distributions for patients to meet a doctor in the emergency room at a government hospital and a private hospital are equal. The waiting times in both the hospitals are normally distributed. So we can use the F-test for the two population variances.

For applying the F-test, we first have to set up the null and alternative hypotheses. If σ_1^2 and σ_2^2 denote the variances of waiting time distributions for a patient to meet a doctor in the emergency room at the government and the private hospitals, respectively, we can take the null and alternative hypotheses as follows:

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ (claim) and } H_1 : \sigma_1^2 \neq \sigma_2^2 \qquad [\text{two-tailed test}]$$

We apply the F-test for testing the hypothesis for two population variances in Excel 2007 as follows:

Step 1: We repeat Steps 1 to 4 under the heading 'Steps in Excel' of Sec.10.3 to enter the data, calculate the sample variances, type the values of sample sizes and level of significance α. We get the output shown in Fig.10.45.

	A	B	C
27	25	25	
28	Sample variance	47.4067	38.2105
29	Sample size	25	20
30	α	0.02	
31			

Fig. 10.45

Step 2: We now calculate the value of the test statistic F using equation (11) in Cell B31. For calculating F_{cal}, we type "=B28/C28" in Cell B31 (see Fig. 10.46a) and press ***Enter***. Then we get the value of F_{cal} in Cell B31 (Fig. 10.46b).

FDIST | f_x =B28/C28

	A	B	C
27	25	25	
28	Sample variance	47.4067	38.2105
29	Sample size	25	20
30	α	0.02	
31	F_{cal}	=B28/C28	
32			

(a)

B31 | f_x =B28/C28

	A	B
30	α	0.02
31	F_{cal}	1.2407
32		

(b)

Fig. 10.46

Step 3: Decision using the critical region approach

We now obtain the critical (tabulated) value with the help of Excel as follows:

1. Since Excel calculates the critical value of the F-test on the right-tail, we first compute the values of α/2 and 1– α/2 in Cells B32 and B33 by typing "=B30/2" in Cell B32 and "=1-B32" in Cell B33 and pressing ***Enter***. We get the outputs shown in Figs. 10.47a and b.

B32 | f_x =B30/2

	A	B
31	F_{cal}	1.2407
32	α/2	0.01
33		

(a)

B33 | f_x =1-B32

	A	B
32	α/2	0.01
33	1–α/2	0.99
34		

(b)

Fig. 10.47

2. The F-test requires two degrees of freedom: one for the numerator $(n_1 – 1)$ and the other for the denominator $(n_2 – 1)$. So we calculate these dfs in Cells B34 and C34 by typing "=B29-1" in Cell B34 and "=C29-1" in Cell C34, respectively. The outputs are shown in Figs. 10.48a and b.

B34 | f_x =B29-1

	A	B	C
33	1–α/2	0.99	
34	df	24	
35			

(a)

C34 | f_x =C29-1

	A	B	C
33	1–α/2	0.99	
34	df	24	19
35			

(b)

Fig. 10.48

3. We now compute the critical (tabulated) values. Since Excel gives the right-tail critical value for the F-test, we first compute the right-tail critical value in Cell B35 by clicking on ***Formulas*** → ***More Functions***→***Statistical***→***Finv*** function as shown in Fig. 10.49.

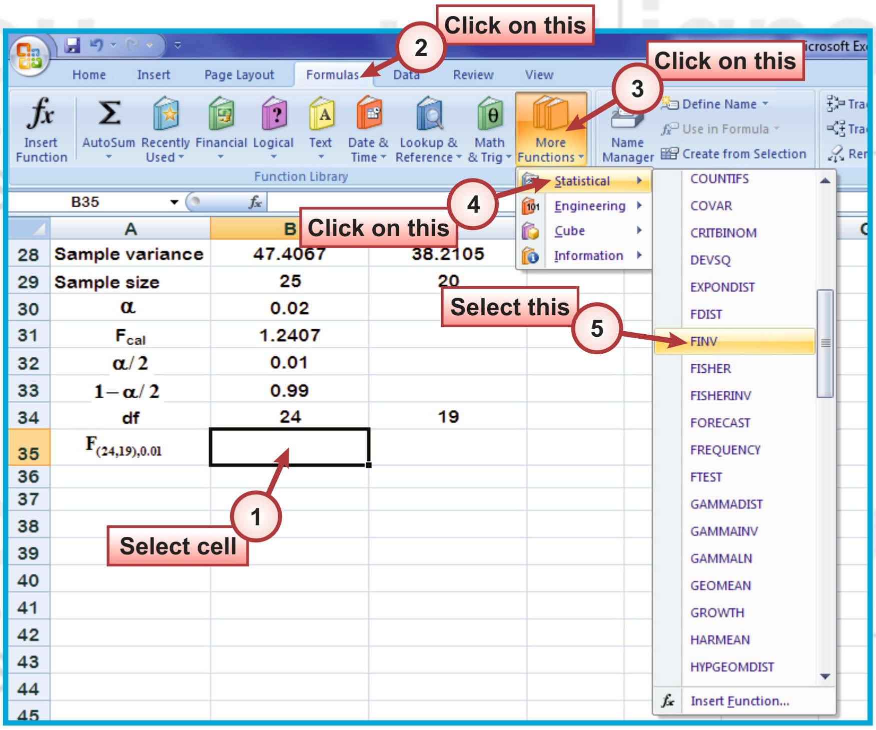


Fig. 10.49

4. We get a new dialog box. This dialog box requires the value of ***probability*** (level of significance α), ***Deg_freedom1*** and ***Deg_freedom2***. Since Excel gives the critical value for the right-tailed F-test and here the test is two-tailed, we consider $\alpha/2 = 0.01$ which is put in Cell B32. So we select Cell B32 in ***Probability***, Cell B34 in ***Deg_freedom1***, Cell C34 in ***Deg_freedom2*** and then click on ***OK*** (see Fig. 10.50a). This function gives the value of $F_{(24,19),0.01}$ as shown in Fig. 10.50b.

(a)

(b)

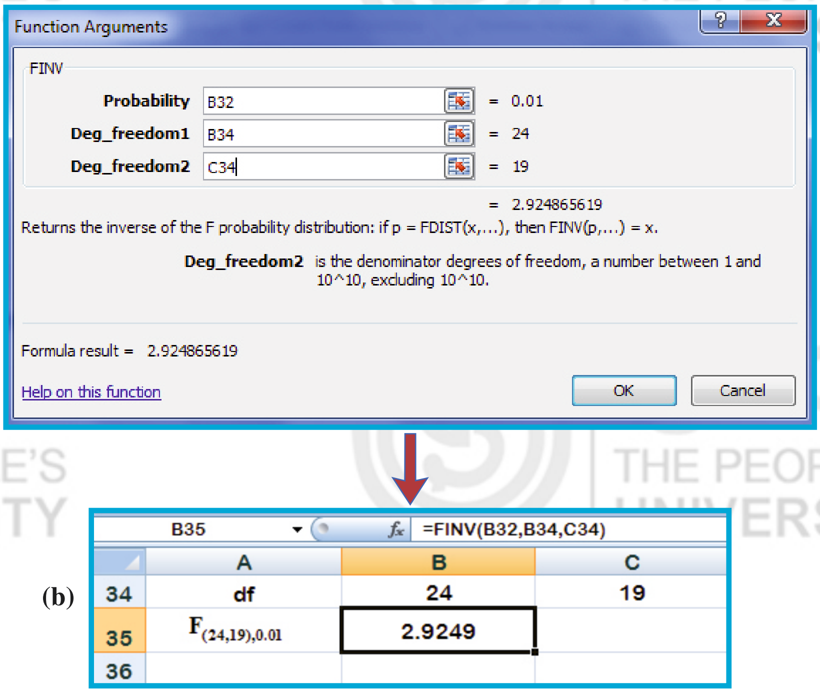


Fig. 10.50

5. We calculate the left-tail critical value $F_{(24,19),0.99}$ in Cell B36 as explained in points 3 and 4 by taking Cell B33 in ***Probability***, Cell B34 in ***Deg_freedom1***, Cell C34 in ***Deg_freedom 2*** and then clicking on ***OK***. We get the output shown in Fig. 10.51.

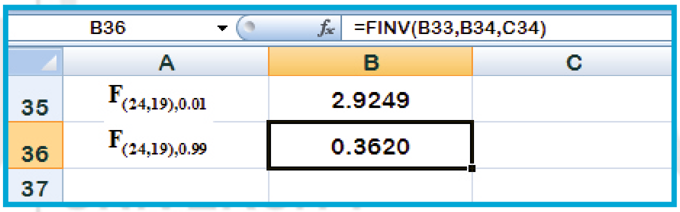


Fig. 10.51

Conclusion

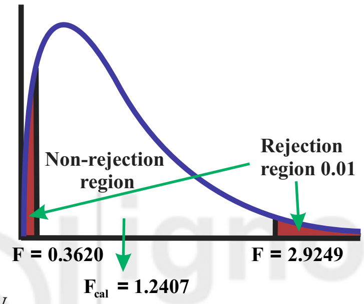


Fig. 10.52

Since $F_{(24,19),0.99}$ $(= 0.3620) < F_{cal}$ $(= 1.2407) < F_{(24,19),0.01}$ $(= 2.9249)$, it means that the calculated value lies in the non-rejection region as shown in Fig. 10.52. So we do not reject the null hypothesis. Since our claim is under the null hypothesis, we do not reject the claim. Hence, we conclude that the samples do not provide us sufficient evidence against the claim. Thus, we may assume that the variances of waiting time distributions for patients to meet a doctor in the emergency room at the government hospital and the private hospital are equal at 2% level of significance.

Step 4: Decision using the p-value approach

Since our problem is two-tailed, the p-value = $2P\left[F \geq F_{cal}\right]$.

Excel gives $P\left[F \geq F_{cal}\right]$. So we calculate $P\left[F \geq F_{cal}\right]$ first and then the p-value as follows:

1. We select Cell B37.
2. Then we click on ***Formulas*** → ***More Functions*** → ***Statistical*** → ***Fdist*** function as shown in Fig.10.53.

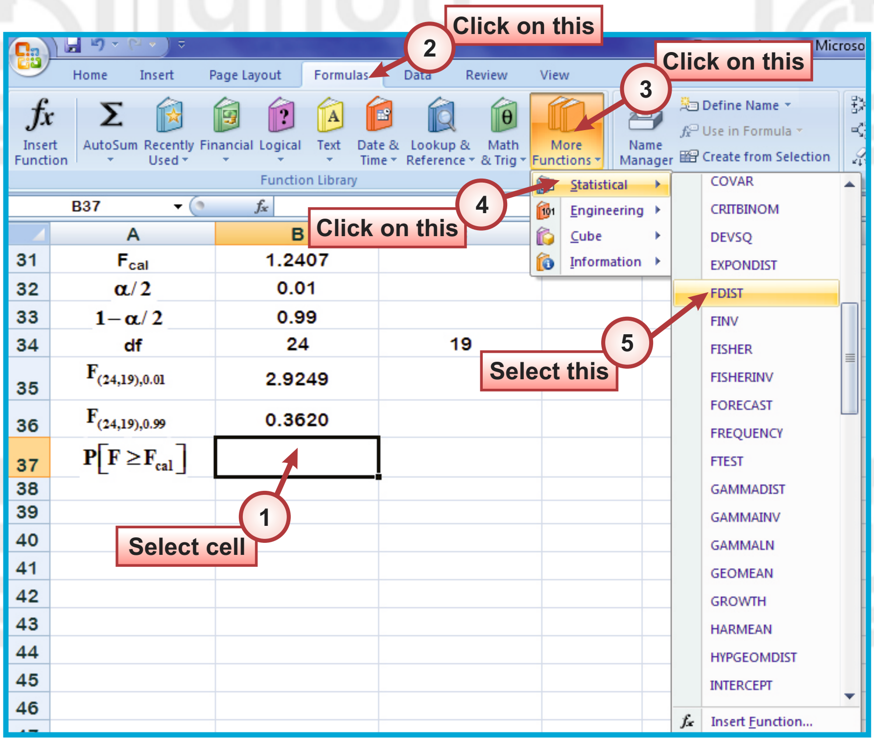


Fig. 10.53

3. When we click on ***Fdist***, a new dialog box is opened. The dialog box requires the value of ***X*** (F_{cal}), ***Deg_freedom1*** and ***Deg_freedom2***. We select Cell B31 in ***X***, Cell B34 in ***Deg_freedom1***, Cell C34 in ***Deg_freedom 2*** and then click on ***OK*** (see Fig. 10.54a). The output is shown in Fig. 10.54b.

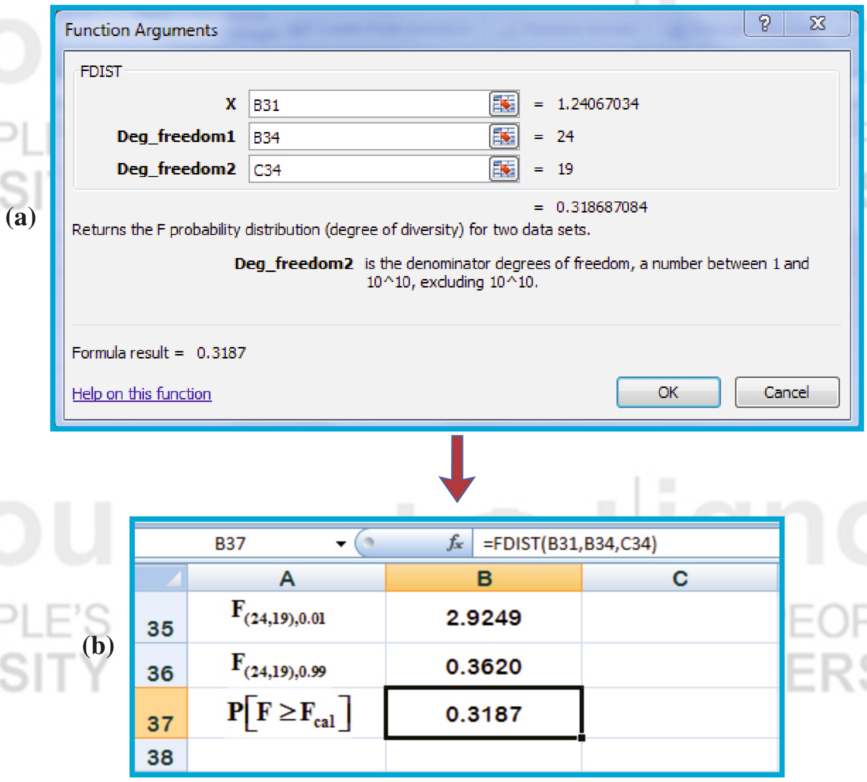


Fig. 10.54

4. We now calculate the p-value, $2P[F \geq F_{cal}]$, in Cell B38 by typing "=2*B37" in Cell B38 and pressing ***Enter***. We get the p-value as shown in Fig. 10.55.

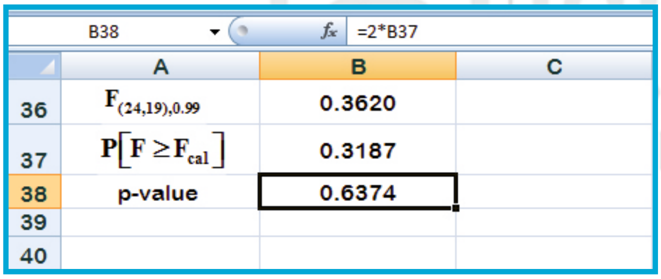


Fig. 10.55

Conclusion

We compare the calculated p-value with the level of significance $\alpha = 0.02$. Since the p-value (= 0.6374) is greater than 0.02, we do not reject the null hypothesis.

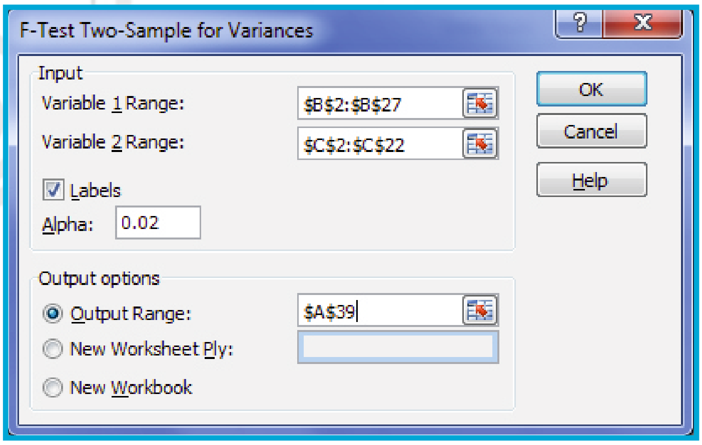
10.7.1 F-test using the Data Analysis ToolPak

We can also apply the F-test with the help of the ***Data Analysis ToolPak*** in MS Excel 2007.

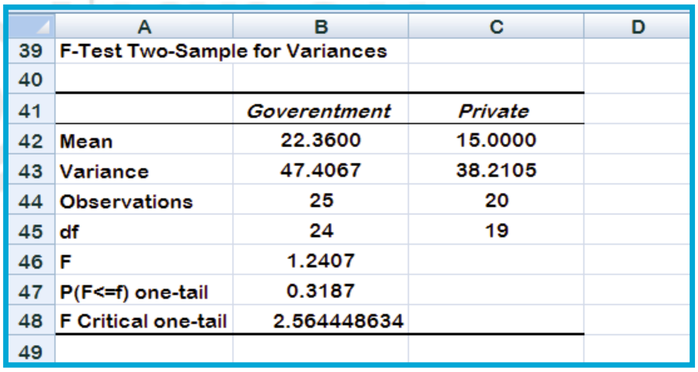
Step 1: We enter the data (given in Table 2) in Excel 2007 spreadsheet as explained in Step 1 under the heading 'Steps in Excel' of Sec. 10.3.

Step 2: We now apply the ***Data Analysis ToolPak*** for the F-test as follows:

We click on ***Data*** → ***Data Analysis*** → ***F-Test Two-sample for Variances*** → ***OK***. A new dialog box opens. We put the values as explained in Step 3 under the heading 'Steps in Excel' of Sec. 10.3 (see Fig. 10.56a) and then press ***OK***. We get the output shown in Fig. 10.56b.



(a)



	A	B	C	D
39	F-Test Two-Sample for Variances			
40				
41		*Goverentment*	*Private*	
42	Mean	22.3600	15.0000	
43	Variance	47.4067	38.2105	
44	Observations	25	20	
45	df	24	19	
46	F	1.2407		
47	P(F<=f) one-tail	0.3187		
48	F Critical one-tail	2.564448634		
49				

(b)

Fig. 10.56

Step 3: The ***Data Analysis ToolPak*** of MS Excel 2007 gives the results of the F-test only for the right tail. We repeat Steps 3 and 4 under the heading 'Steps in Excel' of Sec. 10.7, respectively, for two-tail critical value and two-tail p-value.

Step 4: We take the decision about the null hypothesis/claim using the critical region approach and the p-value approach as explained in Steps 3 and 4 under the heading 'Steps in Excel' of Sec. 10.7, respectively.

You should now apply these tests to other problems for practice.

Activity 2

Apply the suitable test on the data with the help of MS Excel 2007 for the following exercises and interpret the results:

A1) Examples 3, 4, 7 and 8 given in Unit 10 of MST-004.

A2) Exercises E6, E7, E10 and E11 given in Unit 10 of MST-004.

A3) Examples 3, 4, 5 and 6 given in Unit 11 of MST-004.

A4) Exercises E4, E5, E6 and E7 given in Unit 11 of MST-004.

A5) Examples 3 and 4 given in Unit 12 of MST-004.

A6) Exercises E3 and E4 given in Unit 12 of MST-004.

Match the results with the manual computation of data carried out in Units 10, 11 and 12 of MST-004.

Continuous Assessment 10

1. A local pizza restaurant and a branch of a branded pizza chain are located across the street from a university campus. The local pizza restaurant advertises that it delivers pizzas to the dormitories faster than the branded restaurant. In order to determine whether this advertisement is valid, some students of Statistics decide to order 15 pizzas from the local pizza restaurant and 15 pizzas from the branch of the branded pizza chain, at different times and record the delivery times (in minutes). The data are given in Table 4.

Table 4: Delivery time

S. No.	Delivery Time (in minutes)	
	Local Pizza Restaurant	Branch of Branded Pizza Chain
1	16.4	20.4
2	15.0	16.2
3	17.5	15.0
4	14.2	18.2
5	20.0	22.5
6	15.4	16.2
7	17.5	20.0
8	14.1	15.0
9	13.4	15.0
10	20.7	24.2
11	18.6	20.0
12	20.0	20.0
13	15.3	18.2
14	15.0	19.5
15	12.8	20.0

Assuming that the delivery time is normally distributed in both local and branded restaurants, is there evidence that the mean delivery time for the local restaurant is less than the mean delivery time for the branded restaurant at 5% level of significance if:

i) the standard deviations of the delivery time of the local pizza restaurant and the branded restaurant are known to be 2.5 and 3.0, respectively,

ii) the standard deviations of the delivery time of the local pizza restaurant and the branded restaurant are equal and not known, and

iii) the students matched the samples for each of the 15 times that the pizzas were ordered and have one measurement from the local pizza restaurant and one from the of branded restaurant.

2. A researcher would like to test whether there is any significant difference between safety-consciousness of men and women while driving a car. In a sample of 300 men, 130 said that they used seat belts. In a sample of 200 women, 90 said that they used seat belts. Test the claim that there is no significant difference between safety-consciousness of men and women while driving a car at 5% level of significance.

3. In Exercise 1, the students of Statistics also wish to determine whether the variance of delivery time of the local pizza restaurant is less than that of the branded restaurant. Then
   - formulate the null and alternative hypotheses.
   - is the variance of delivery time of the local pizza restaurant less than that of the branded restaurant at 1% level of significance?

Home Work: Do It Yourself

1) Follow the steps explained in Secs.10.3, 10.4, 10.5, 10.6 and 10.7 to apply the tests on the data of Tables 1, 2 and 3. Take the final screenshots and keep them in your record book.
2) Develop the spreadsheets for the exercises of "Continuous Assessment 10" as explained in this lab session. Take a screenshot of the final spreadsheet.
3) **Do not forget** to keep all screenshots in your record book as these will contribute to your continuous assessment in the Laboratory.