SECTION 3 ANALYSIS OF VARIANCE

Structure **Page Nos.**

3.0 INTRODUCTION

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

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

3.1 OBJECTIVES

After going through this unit you will be able to:

- bring out an appraisal of any physical problem;
- identify suitability of using ANOVA;
- use Data Analysis Toolpak for ANOVA, in particular;
  - Perform One Way / Single Factor ANOVA Test; and
  - Perform Two Way / Two Factor ANOVA Test.

3.2 ANOVA TEST

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

3.2.1 One-Way Classification

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

Treatment 1	y_{11}	y_{12}	⋯	y_{1,n_1}
Treatment 2	y_{21}	y_{22}	⋯	y_{2,n_2}
⋮	⋮	⋮	⋮	⋮
Treatment k	y_{k1}	y_{k2}	⋯	y_{k,n_k}

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

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, j = 1,2, \dots, n_i \text{ and } i = 1,2, \dots k$$

and the hypothesis to be tested is $H_0: \tau_1 = \tau_2 = \cdots = \tau_k = 0$.

Assumptions:

- Errors (ϵ_{ij})are identically and independently distributed with mean 0 and variance σ^2 (Homoscedastic).
- Errors (ϵ_{ij}) have normal distribution.

Application CASE STUDY

FACTORS INFLUENCING SALES OF STREAMERS

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

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

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

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

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

DEALER - 1 : LOCALITYWISE SALES RECORD OF SALESMEN

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

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

ANALYSIS

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

Company Objective

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

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, j = 1,2, \dots, n_i \text{ and } i = 1,2, \dots 4 \text{ dealers}$$

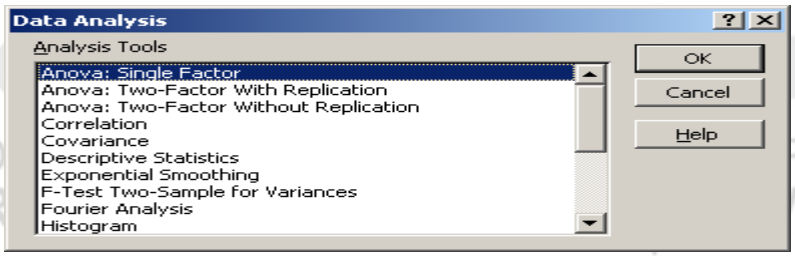
Where the additive model has $\tau_i = \mu_i - \mu$ termed as the additional effect of the ith treatment and the hypothesis to be tested is$H_0: \tau_i = 0$, or equivalently$H_0: \mu_i = \mu$, $i = 1,2,3,4$. Now, we demonstrate how to use the excel tool to perform a test of this hypothesis.

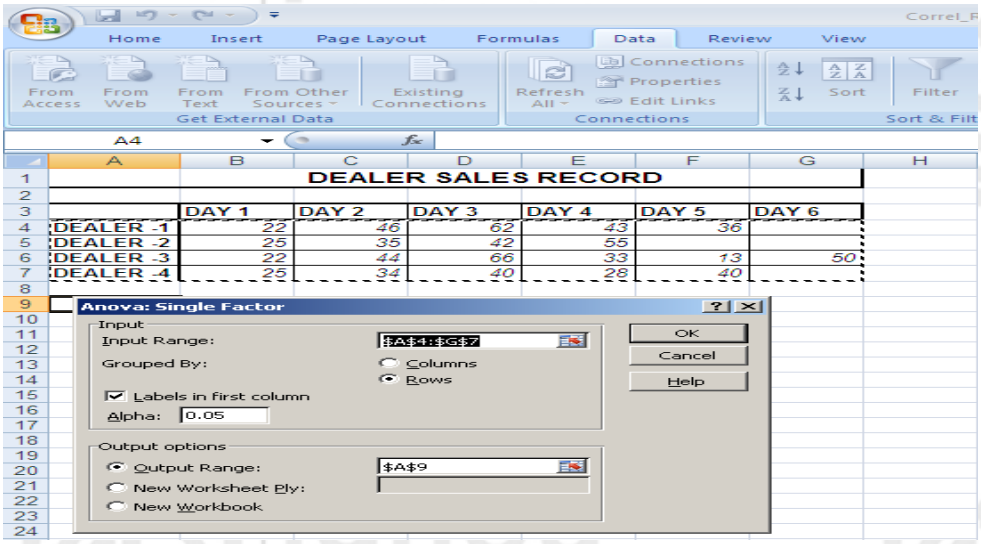
Steps:

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

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

   However we are giving some of the relevant screenshots here





Notice from the screen shot above that while selecting the cells, only row labels are included (columns excluded). This is because incase of one way classification, the data is required to be classified by only one factor viz., the Dealers in the present case. Data on the sales volume of different dealers is recorded row wise, where the dealer names are entered in first column. Thus, we check the option "Levels in First Column". Further, the level of significance *Alpha* for the test is by default set at 5% or 0.05, which can be altered to 1% or 0.01 etc. as per requirement. We have to identify the output cell address where the results are desired to be placed, which is chosen as A10. Following is the result of the procedure discussed above.

Anova: Single Factor

SUMMARY

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

ANOVA

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

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

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

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

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

Data Interpretation

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

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

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

☞ Check your progress 1

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

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

   .....

   .....

   .....

   .....

   .....

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

   .....

   .....

   .....

   .....

   .....

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

   .....

   .....

   .....

   .....

   .....

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

   .....

   .....

   .....

   .....

   .....

3.2.2 Two-way Classification

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

	Block 1	Block 2	…	Block n
Treatment 1	y_{11}	y_{12}	…	$y_{1,n}$
Treatment 2	y_{21}	y_{22}	…	$y_{2,n}$
⋮	⋮	⋮	⋮	⋮
Treatment k	y_{k1}	y_{k2}	…	$y_{k,n}$

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

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, j = 1,2, \dots, n \text{ and } i = 1,2, \dots k$$

and the hypothesis to be tested are (i)For Treatments$H_{01}: \alpha_i = 0, i = 1,2, \dots k$ and (ii)For Blocks$H_{02}: \beta_j = 0, j = 1,2, \dots, n$. Here, μ is the grand mean, α_i is the treatment effect and β_j is the block effect.

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

Dealer Objective

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

From the objectives, we identified, that to apply the ANOVA test we need to establish Two Null hypothesis, H_{01} andH_{02}, which are to be tested simultaneously. ForH_{01},letμ be the grand mean,α_i be that part of y_{ij} due to the i[th] salesman and forH_{02}, let $\beta_j$be that part of y_{ij} due to thej[th] locality. Thus, the Null Hypothesis to be tested are

$$H_{01}: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0 \text{and} H_{02}: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.$$

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

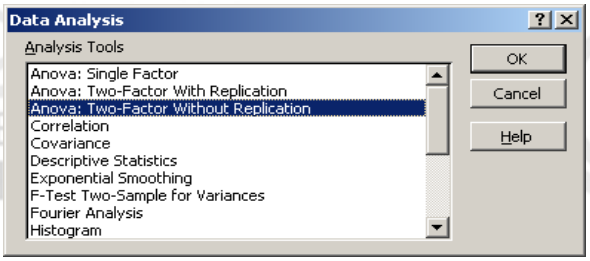
Data analysis through Excel

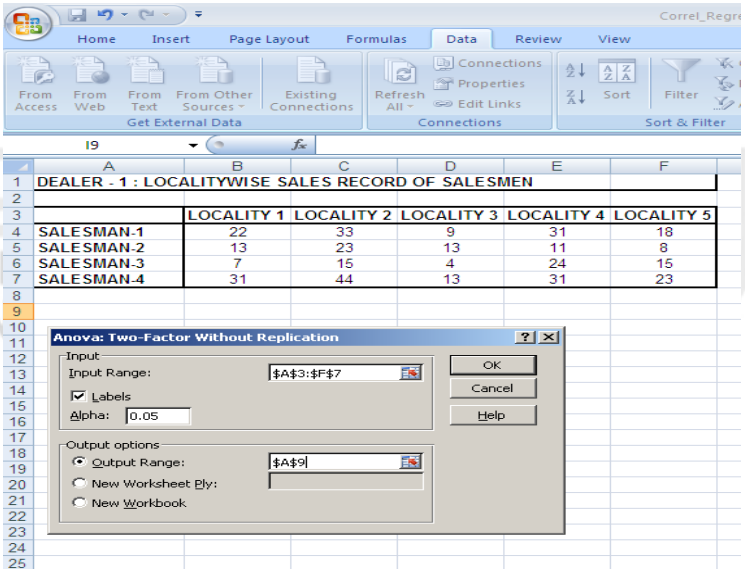
Perform Following Steps:

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

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

   However we are giving some of the relevant screenshots here





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

Anova: Two-Factor Without Replication

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

ANOVA

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

Note:

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

☞ Check Your Progress 2

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

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

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

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

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

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

☞ **Lab Sessions 6**

Perform the following using spreadsheet package.

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

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

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

2) Implement questions of Check Your Progress using spreadsheet, wherever possible.

3.3 SUMMARY

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

3.4 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress -1

1) Alpha $\alpha = 0.05$ i.e., 5% , so level of confidence is 95%
2) Critical value of Factor(Fcrit)$F = 3.238872$
3) Because $F = 0.290072$and $Fcrit = 3.238872$.Since$< Fcrit$, we do not reject the NULL Hypothesis.
4) $p = 0.831921$, which is more than the desired significance level, so we do not reject the null hypothesis

Check Your Progress -2

1) Alpha $\alpha 0.05 =$ i.e.,5%, thus level of confidence 95%
2) Critical value of (Fcrit) $F = 3.49$ and 3.25 respectively
3) $F = 8.015$; $Fcrit = 3.49$; Since$F > Fcrit$, we Reject the Null Hypothesis
4) $F = 6.291$; $Fcrit = 3.259$; Since$F > Fcrit$, we Reject the Null Hypothesis
5) Since p value greater than Alpha, we Reject the Null Hypothesis

Tips for solving Session 6

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

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

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

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

Test statistic: 8.61

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