
UNIT 11 ANALYSIS OF VARIANCE

Structure

- 11.0 Introduction
- 11.1 Objectives
- 11.2 Analysis of Variance (ANOVA)
 - 11.2.1 Significance of Analysis of Variance
- 11.3 One-way Analysis of Variance (ANOVA)
 - 11.3.1 Procedure of Testing of Hypothesis of One-Way ANOVA
- 11.4 Two-way Analysis of Variance (ANOVA)
 - 11.4.1 Procedure of Testing of Hypothesis of Two-way ANOVA
- 11.5 Summary
- 11.6 Solutions /Answers
- References

11.0 INTRODUCTION

In block 3 of this course, we tested the equality of means of two independent groups by using a t-test. Sometimes, situations may arise where testing of more than two means is required. As an example, in crop-cutting experiments, it may be required to test whether under similar conditions the average yield of some types of crops (more than two) in a number of fields is same or not. For obvious reasons, in such cases, t-test cannot be applied. Generally, for such situations, the technique of Analysis of Variance (ANOVA) is used, in which the testing of equality of several means is done by dividing the population variability into different components. The usual F-test is used to test the equality of means of several groups.

As its name suggests, the analysis of variance focuses on variability. It involves the calculation of several measures of variability, when the total variability of the population is divided into many components, like, variability within the sub-groups, variability between the sub-groups, etc. In other words, ANOVA is a technique which split up the total variation of data which may be attributed to various “sources” or “causes” of variation. There may be variation between variables and also within different levels of variables. In this way, Analysis of Variance is used to test the homogeneity (equality) of several population means by comparing the variances between the sample and within the sample.

In this unit, we shall discuss the one-way as well as two-way analysis of variance. One-way analysis of variance is a technique where only one independent variable at different levels is considered which affects the response variable whereas in two-way analysis of variance technique, we will consider two independent variables at different levels which affect the response variable.

11.1 OBJECTIVES

After reading this unit, you would be able to:

- explain the analysis of variance technique;

- explain the significance of analysis of variance technique;
- test the hypothesis under one-way analysis of variance; and
- use the method of performing two-way ANOVA test.

11.2 ANALYSIS OF VARIANCE (ANOVA)

The statistical technique known as "Analysis of Variance"(abbreviated as ANOVA), was propounded by Professor R.A. Fisher in 1920's, in which he developed the method of testing equality of means of several sub-populations by dividing the total variability into different components. Variation is inherent in nature, so analysis of variance means examining the variation present in data or parts of data. In other words, analysis of variance means to find out the cause of variation in the data. The total variation in any set of numerical data of an experiment is due to number of causes which may be **assignable causes** and **chance causes**.

The variation in the data due to assignable causes can be detected, measured and controlled whereas the variation due to chance causes is not in the control of human beings and cannot be traced or found separately.

According to Professor R.A. Fisher, Analysis of Variance (ANOVA) is the "separation of variance ascribable to one group of causes from the variance ascribable to other groups". So, by this technique, the total variation present in the data is divided into two components of variation: one due to assignable causes (between the groups variability) and the other due to chance causes (within Group variability).

If we consider only one independent variable which affects the response / dependent variable then it is called one-way ANOVA. If the independent variables / explanatory variables are more than one, say n , then it is called n -way ANOVA. If n is equal to two then the ANOVA is called two-way ANOVA.

11.2.1 Significance of Analysis of Variance

One obvious question may arise that why test of equality of several means (more than two) is called analysis of variance test instead of the analysis of means? How do test for equality of means for more than two populations propose by analysis the variances? As a matter of fact, in order to determine that means of several populations are equal or not, the measure of variance is considered.

The amount of the total variation in the population σ^2 is computed by two different approaches. The first approach is to compute the variation in the population in such a manner that even if the population means are not equal it will have no effect on it. This means that the differences in the values of the population means does not affect on the population variation σ^2 as calculated by a given method. This amount of variation in the population σ^2 is the average of the variation found within each of the samples. For example, if we take k samples of size n , then each sample will have a mean and a variance. Then the mean of the amount of k variations would be considered as an unbiased estimate of σ^2 , the population variance, and its value

remains appropriate irrespective of whether the population means are equal or not. This is really done by pooling all the variations within the samples to estimate a common population variation, which is the average of all sample variances. This common variance is known as variation within samples.

The second approach to calculate the estimate of total variation of the population σ^2 is based upon the Central Limit theorem that we have applied in Large sample and Small sample tests. This means that, in fact, if there are no differences among the k population means, then the total amount of variation in the population determined by the second approach should not differ significantly from that computed by the first approach. Hence, if these two values of population variance are approximately the same, then we can accept the null hypothesis of equality of means.

Based upon the Central Limit Theorem, we have previously found that the standard error of the sample means is calculated by:

$$SE(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

or, the variance would be:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \Rightarrow \sigma^2 = n\sigma_{\bar{x}}^2$$

Thus, by knowing the square of the standard error of the mean $(\sigma_{\bar{x}})^2$, we multiply it by n and obtain a precise estimate of σ^2 . This approach of determining the estimate of variation in the population is known as population variation between samples $\sigma_{\text{between}}^2$. Now, if all population means are equal, then population variation between samples $\sigma_{\text{between}}^2$ should be approximately the same as population variation within samples σ_{within}^2 . A significant difference between these two values would lead us to conclude that this difference is the result of difference between the population means.

But, how do any difference between these two variations is significant or not be known? How do we determine that whether this difference, if any, is simply due to random sampling error or due to actual differences among the population means?

R. A. Fisher developed a Fisher test or F-test to answer the above question. He determined that the difference between the population variation between samples and population variation within samples could be expressed as a ratio to be designated as the F-value, so that

$$F = \frac{\text{Total Variation Between Samples}}{\text{Total Variation Within Samples}} = \frac{\sigma_{\text{between}}^2}{\sigma_{\text{within}}^2}$$

In the above case, if the population means are exactly the same, then variation between samples will be equal to the variation within samples and the value of F will be equal to 1.

However, because of sampling errors and other variations, some disparity between these two values will be there, even when the null hypothesis is true, meaning that all population means are equal. The extent of disparity between the two variances

and consequently, the value of F , will influence our decision on whether to accept or reject the null hypothesis. It is logical to conclude that if the population means are not equal then their sample means will also vary greatly from one another, resulting in a large value of variation between samples, and hence a larger value of F (variation within samples is based only on sample variances and not on sample means and hence is not affected by differences in sample means). Accordingly, larger the value of F statistic, the more likely the decision is to reject the null hypothesis. It means that the null hypothesis will be rejected if the computed value of F must be larger than the critical value of F , given in Table VII of the Appendix, for a given level of significance and calculated number of degrees of freedom.

Let us now answer the following Check Your Progress questions.

Check Your Progress 1

- 1) Describe the analysis of variance and differentiate between one-way and two-way ANOVA.

.....
.....
.....

- 2) Explain the significance of analysis of variance briefly.

.....
.....
.....

11.3 ONE-WAY ANALYSIS OF VARIANCE

If we consider only one independent variable which affects the response / dependent variable then it is called One-way ANOVA. It is used to test the equality of more than two means when the observations are classified according to one factor/treatment at different levels. For example, we may wish to study the simultaneous effects of five varieties of wheat (independent variable) on the yield (dependent variable) or test the stress level of employees in three different organisations, and so on. In such situations, we can also apply one-way ANOVA for each treatment/factor. As we have studied that two samples t-test is used to decide whether two groups (two levels) of a factor have the same mean. So one-way ANOVA generalises this problem to k levels (greater than two) of a factor.

Suppose there are k normal populations or k levels of a factor/treatment with mean/average effect $\mu_1, \mu_2, \dots, \mu_k$ with common variance σ^2 . Also let us draw k random samples one from each population of size n_i ($i = 1, 2, \dots, k$). The observations of different samples or on different levels of a factor can be exhibited as shown on the next page. In this table, y_{ij} is the j th observation of the i th sample/level of a factor. For example, y_{23} refers to third observation in the second level of a factor.

Level of Factor/ Treatment	1	2	...	k
Observations	y_{11}	y_{21}	...	y_{k1}
	y_{12}	y_{22}	...	y_{k2}
	y_{13}	y_{23}	...	y_{k3}
	...	...	...	...
	y_{1n_1}	y_{2n_2}	...	y_{kn_k}
Total	T_1	T_2	...	T_k
Mean	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_k$

11.3.1 Procedure of Testing of Hypothesis of One-Way ANOVA

In this section, we discuss the step-by-step computation procedure for one-way analysis of variance for k independent samples as

Step 1: We first formulate the null hypothesis (H_0) and the alternative hypothesis (H_1). We want to test the equality of the population means $\mu_1, \mu_2, \dots, \mu_k$ or test the homogeneity of the effect of different levels of a factor. Hence, the null and alternative hypotheses are

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$
the alternative hypothesis
 $H_1: \text{At least two are not equal}$

Step 2: We calculate the total sum of squares of variation (TSS) as

$$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 \text{ where } \bar{\bar{y}} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}$$

Or

$$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - CF$$

where $\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2$ is called raw sum of squares and is calculated as

$$\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 = (y_{11}^2 + y_{12}^2 + \dots + y_{1n_1}^2) + (y_{21}^2 + y_{22}^2 + \dots + y_{2n_2}^2) + \dots + (y_{k1}^2 + y_{k2}^2 + \dots + y_{kn_k}^2)$$

And CF is called correction factor and is calculated as

$$CF = \frac{(\text{Grand Total})^2}{\text{Total number of observations}} = \frac{G^2}{N}$$

Step 3: We calculate the sum of squares of variation between the samples (SSB) or sum of squares of variation due to different levels of a factor as

$$SSB = \sum_{i=1}^k n_i (\bar{y}_i - \bar{\bar{y}})^2 \text{ or } SSB = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \dots + \frac{T_k^2}{n_k} - CF$$

Step 4: After that, we calculate the sum of squares variation within samples (SSW) or sum of squares of variation due to error (SSE) as

$$SSE(\text{or } SSW) = TSS - SSB$$

Step 5: We determine the degrees of freedom (df) as

The degrees of freedom (df) for total = $N - 1$

The degrees of freedom (df) for different samples = $k - 1$

The degrees of freedom (df) for error = $N - k$

Step 6: We obtain the various mean sums of squares of variation as follows:

$$\text{Mean sum of squares of variation due to factor (MSSB)} = \frac{SSB}{k - 1}$$

$$\text{Mean sum of squares of variation due to error (MSSE)} = \frac{SSE}{N - k}$$

Step 7: We calculate the value of the test statistics using the formulae given below:

$$F = \frac{MSSB}{MSSE}$$

After the various calculations for SSB, SSE and the degrees of freedom, we present these figures in a simple table called Analysis of Variance table or simply ANOVA table, as follows;

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Sum of Squares	Variance Ratio F
Between samples	SSB	$(k - 1)$	$MSSB = \frac{SSB}{(k - 1)}$	$F = \frac{MSSB}{MSSE}$
Within samples (Error)	SSE	$(N - k)$	$MSSE = \frac{SSE}{(N - k)}$	
Total	TSS			

Step 8: For taking the decision about the null hypothesis, we compare the calculated value of F with tabulated value of F obtained from the F-table at $(k - 1, N - k)$ df for α level of significance. If calculated value is greater than the tabulated value then we reject the hypothesis H_0 , otherwise, it may be accepted.

Note: We can use any formula mentioned in Steps 2 and 3 to calculate TSS and

SSB. But the formulae $TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$ and $SSB = \sum_{i=1}^k n_i (\bar{y}_i - \bar{\bar{y}})^2$ makes

calculation easy only when we get $\bar{y}_i$'s and $\bar{\bar{y}}$ as a whole number. We have used both the formulae for your understanding.

Now, after discussing the procedure of one-way ANOVA test let us practice and solve some examples.

Example 1: To test whether the performance of the students of different sections, who took the class of the introductory statistics of the same material by different professors, is same or not, 4 sections of the same course were selected and a common test was administered to 5 students selected at random from each section. The scores for each student from each section were noted and are given below. Test for any difference in learning, as reflected in the average scores for each section at 5% level of significance.

Student	Section 1 Scores	Section 2 Scores	Section 3 Scores	Section 4 Scores
1	8	12	10	12
2	10	12	13	15
3	12	10	11	13
4	10	8	12	10
5	5	13	14	10
Totals	$T_1 = 45$	$T_2 = 55$	$T_3 = 60$	$T_4 = 60$

Solution: The method is as follows:

We are assuming that there is no significant difference among the average scores of students from these 4 sections and hence all professors are teaching the same material with the same effectiveness, i.e.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

H_1 : At least two means differ from each other

To calculate TSS and SSB, we make the following calculations as follows:

Student	y_1	y_2	y_3	y_4
1	8	12	10	12
2	10	12	13	15
3	12	10	11	13
4	10	8	12	10
5	5	13	14	10
Total	$T_1 = 45$	$T_2 = 55$	$T_3 = 60$	$T_4 = 60$
Mean	$\bar{y}_1 = 9$	$\bar{y}_2 = 11$	$\bar{y}_3 = 12$	$\bar{y}_4 = 12$

The grand mean $\bar{\bar{y}}$ is

$$\bar{\bar{y}} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} = \frac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3 + \bar{y}_4}{4} = \frac{9+11+12+12}{4} = 11$$

From the above calculations, we observed that $\bar{y}$'s and $\bar{\bar{y}}$ are whole numbers so we use $TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$ and $SSB = \sum_{i=1}^k n_i (\bar{y}_i - \bar{\bar{y}})^2$ formulae.

$$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 = (8-11)^2 + (10-11)^2 + (12-11)^2 + \dots + (10-11)^2 = 102$$

We can calculate the value of SSB as

$$\begin{aligned} SSB &= \sum_{i=1}^k n_i (\bar{y}_i - \bar{\bar{y}})^2 = n_1 (\bar{y}_1 - \bar{\bar{y}})^2 + n_2 (\bar{y}_2 - \bar{\bar{y}})^2 + n_3 (\bar{y}_3 - \bar{\bar{y}})^2 + n_4 (\bar{y}_4 - \bar{\bar{y}})^2 \\ &= 5(9-11)^2 + 5(11-11)^2 + 5(12-11)^2 + 5(12-11)^2 \\ &= 20 + 0 + 5 + 5 = 30 \end{aligned}$$

$$SSE = TSS - SSB = 102 - 30 = 72$$

We can construct an ANOVA table for the problem solved above as follows:

ANOVA Table				
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio F
Treatment	SSB = 30	(k - 1) = 3	$MSSB = \frac{SSB}{(k-1)}$ $MSSB = \frac{30}{3} = 10$	$\frac{MSSB}{MSSE}$ $= 2.22$
Within (or error)	SSE = 72	(N - k) = 16	$MSSE = \frac{SSE}{(N-k)}$ $MSSE = \frac{72}{16} = 4.5$	
Total	SST = 102			

Now, we check for the critical value of F from the table for $\alpha = 0.05$ and degrees of freedom as follows:

$$df(\text{numerator}) = (k - 1) = (4 - 1) = 3$$

$$df(\text{denominator}) = (N - k) = (20 - 4) = 16$$

This value of F from Table VII in the Appendix is given as 3.24. Now, since our calculated value of F = 2.22 is less than the critical value of F = 3.24, we cannot reject the null hypothesis.

Example 2: An investigator is interested to know the level of knowledge about the history of India of 4 different schools in a city. A test is given to 5, 6, 7, 6 students of 8th class of 4 schools. Their scores out of 10 are given below:

School I (y_1)	8	6	7	5	9		
School II (y_2)	6	4	6	5	6	7	
School III (y_3)	6	5	5	6	7	8	5
School IV (y_4)	5	6	6	7	6	7	

Solution: If $\mu_1, \mu_2, \mu_3, \mu_4$ denote the average score of students of 8th class of schools I, II, III, IV respectively. Then

Null Hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Alternative hypothesis H_1 : Difference among $\mu_1, \mu_2, \mu_3, \mu_4$ are significant.

S. No.	y_1	y_2	y_3	y_4	y_1^2	y_2^2	y_3^2	y_4^2
1	8	6	6	5	64	36	36	25
2	6	4	5	6	36	16	25	36
3	7	6	5	6	49	36	25	36
4	5	5	6	7	25	25	36	49
5	9	6	7	6	81	36	49	36
6		7	8	7		49	64	49
7			5				25	
Total	35	34	42	37	255	198	260	231

Grand Total $G = 35 + 34 + 42 + 37 = 148$

Correction Factor (CF) = $\frac{G^2}{N} = \frac{148^2}{24} = 912.6667$ Since $N = n_1 + n_2 + n_3 + n_4$

Raw Sum of Square (RSS) = $\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 = 255 + 198 + 260 + 231 = 944$

Total Sum of Square (TSS) = $RSS - CF = 944 - 912.6667 = 31.3333$

Sum of Squares due to Treatments (SST) = $\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} - CF$

$$= \frac{35^2}{5} + \frac{34^2}{6} + \frac{42^2}{7} + \frac{37^2}{6} - 912.6667$$

$$= 245 + 192.6667 + 252 + 228.1667 - 912.6667$$

$$= 5.1667$$

Sum of Squares due to Errors (SSE) = $TSS - SST$

$$= 31.3333 - 5.1667 = 26.1666$$

Now,

$$MSST = \frac{SST}{k-1} = \frac{5.1667}{3} = 1.7222$$

$$MSSE = \frac{SSE}{N-k} = \frac{26.1667}{20} = 1.3083$$

ANOVA Table

Source of Variation	SS	df	MSS	F
Between schools	5.1667	3	1.7222	$F = \frac{1.7222}{1.3083} = 1.3164$
Within schools	26.1666	20	1.3083	

Calculated $F = 1.3164$

Tabulated F (Refer to Table VII of Appendedx) at 5% level of significance with (3, 20) degree of freedom is 3.10.

Conclusion: Since Calculated $F < \text{Tabulated } F$, so we may accept H_0 and conclude that level of knowledge of schools I, II, III and IV do not differ significantly.

Check Your Progress 2

- There are three sections of an introductory course in Statistics. Each section is being taught by a different professor. There are some complaints that at least one of the professors does not cover the necessary material. To make sure that all the students receive the same level of material in a similar manner, the chairperson of the department has prepared a common test to be given to students of the three sections. A random sample of seven students is selected from each class and their test scores out of a total of 20 points are tabulated as follows:

Students	Section (1)	Section (2)	Section (3)
1	20	12	16
2	18	11	15
3	18	10	18
4	16	14	16
5	14	15	16
6	18	12	17
7	15	10	14
Total	119	84	112

Do you think that at 95% confidence level, there is significant difference in the average test scores of students taught by the different professors?

.....

.....

.....

- 2) Able Insurance Company wants to test whether three of its salesmen, A, B, and C, in a territory make a similar number of appointments with prospective customers during a given period of time. A record of the previous four months showed the following results for the number of appointments made by each salesman for each month.

Month	Salesman		
	A	B	C
1	8	6	14
2	9	8	12
3	11	10	18
4	12	4	8
Totals	40	28	52

Do you think that at 95% confidence level, there is significant difference in the average number of appointments made by the salesmen per month?

.....

.....

.....

11.4 TWO-WAY ANALYSIS OF VARIANCE (ANOVA)

In the previous section, we considered the case where only one predictor/independent/ explanatory variable was categorised at different levels. If we are interested in studying the simultaneous effect of two independent factors each at different levels on the dependent variable, we use two-way ANOVA. For example, we may wish to study the simultaneous effects of five varieties of wheat (first criterion) and four different types of fertilisers (second criterion) on the yield (dependent variable) or test the stress level of employees in three different organisations in different regions, and so on. In such situations, we can also apply two separate one-way ANOVA for each treatment/factor. However, it is more advantageous to use two-way ANOVA because the variance can be reduced by introducing the second factor.

In such an ANOVA, generally, we have an experiment in which we simultaneously study the effect of two factors in the same experiment. For each factor, there will be a number of classes/groups or levels. Let the factors be A and B and the respective levels be $A_1, A_2, \dots, A_r$ and $B_1, B_2, \dots, B_s$. Let y_{ij} be the observation/ response/ dependent variable under the i^{th} level of factor A and j^{th} level of factor B. The observations then can be represented in a table as follows:

Two-way Classified Data

A/B	B_1	B_2	...	B_j	...	B_s	TOTAL	MEAN
A_1	y_{11}	y_{12}	...	y_{1j}	...	y_{1s}	$T_{1.}$	$\bar{y}_{1.}$
A_2	y_{21}	y_{22}	...	y_{2j}	...	y_{2s}	$T_{2.}$	$\bar{y}_{2.}$
.	.	.	.	.	.	.	.	.

.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
A_i	y _{i1}	y _{i2}	...	y _{ij}	...	y _{is}	T _{i.}	$\bar{y}_{i.}$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
A_r	y _{r1}	y _{r2}	...	y _{ri}	...	y _{rs}	T _{r.}	$\bar{y}_{r.}$
Total	T _{.1}	T _{.2}	...	T _{.j}	...	T _{.s}	G (grand total)	
Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	...	$\bar{y}_{.j}$	...	$\bar{y}_{.s}$	-	$\bar{y}_{..}$ (grand mean)

11.4.1 Procedure of Testing of Hypothesis of Two-way ANOVA

In two-way ANOVA, the total variation in the data is divided into three components: variation due to the first criterion (factor), variation due to the second criterion (factor) and variation due to error.

Now, let us discuss the testing procedure of two-way ANOVA briefly mentioning the main steps and formulae as follows:

Step 1: We first formulate the null hypothesis (H_0) and the alternative hypothesis (H_1). In two-way ANOVA, we can test two hypotheses simultaneously: one for different levels of factor A and the other for different levels of factor B. If factor A has r levels and α_i is the average effect of i^{th} level of factor A then we can set up the null and alternative hypotheses as follows:

$$H_{0A} : \alpha_1 = \alpha_2 = \dots = \alpha_r$$

$$H_{1A} = \text{At least one } \alpha_i \neq \alpha_j (i \neq j = 1, 2, \dots, r)$$

Similarly, if factor B has s levels and β_j is the average effect of j^{th} level of factor B then we can set up the null and alternative hypotheses as follows:

$$H_{0B} : \beta_1 = \beta_2 = \dots = \beta_s$$

$$H_{1B} = \text{At least one } \beta_i \neq \beta_j (i \neq j = 1, 2, \dots, s)$$

Step 2: In this step, we calculate the total sum of squares of variation (TSS) as calculate in the case of one-way ANOVA as

$$TSS = \sum_{i=1}^r \sum_{j=1}^s (y_{ij} - \bar{\bar{y}})^2 \text{ where } \bar{\bar{y}} = \frac{1}{N} \sum_{i=1}^r \sum_{j=1}^s y_{ij} \text{ grand mean}$$

Or

$$TSS = \sum_{i=1}^r \sum_{j=1}^s y_{ij}^2 - CF$$

where $\sum_{i=1}^r \sum_{j=1}^s y_{ij}^2$ it is called raw sum of squares and calculated as

$$\sum_{i=1}^r \sum_{j=1}^s y_{ij}^2 = (y_{11}^2 + y_{12}^2 + \dots + y_{1s}^2) + (y_{21}^2 + y_{22}^2 + \dots + y_{2s}^2) + \dots + (y_{r1}^2 + y_{r2}^2 + \dots + y_{rs}^2)$$

and CF is called correction factor and calculate as

$$CF = \frac{(\text{Grand total})^2}{\text{Total number of observations}} = \frac{G^2}{N}, N = r \times s$$

Step 3: We calculate the sum of squares of variation between rows or sum of squares of variation due to factor A (SSA) as

$$SSA = s \sum_{i=1}^r (\bar{y}_{i.} - \bar{\bar{y}})^2 = \frac{T_{1.}^2}{s} + \frac{T_{2.}^2}{s} + \dots + \frac{T_{r.}^2}{s} - CF$$

Step 4: We calculate the sum of squares of variation between columns or sum of squares of variation due to factor B (SSB) as

$$SSB = r \sum_{j=1}^s (\bar{y}_{.j} - \bar{\bar{y}})^2 = \frac{T_{.1}^2}{r} + \frac{T_{.2}^2}{r} + \dots + \frac{T_{.s}^2}{r} - CF$$

Step 5: After that we calculate the sum of squares of variation due to error (SSE) as

$$SSE = TSS - SSA - SSB$$

Step 6: We determine the degrees of freedom (df) as

The degrees of freedom (df) for factor A = $r - 1$

The degrees of freedom (df) for factor B = $s - 1$

The degrees of freedom (df) for error = $(r - 1)(s - 1)$

Step 7: We obtain the various mean sums of squares of variation as follows:

$$\text{Mean sum of squares of variation due to factor A (MSSA)} = \frac{SSA}{r - 1}$$

$$\text{Mean sum of squares of variation due to factor B (MSSB)} = \frac{SSB}{s - 1}$$

$$\text{Mean sum of squares of variation due to error (MSSE)} = \frac{SSE}{(r - 1)(s - 1)}$$

Step 8: We calculate the value of the test statistics using the formulae given below:

$$F_A = \frac{MSSA}{MSSE} \quad \text{and} \quad F_B = \frac{MSSB}{MSSE}$$

ANOVA Table

Sources of Variation	DF	SS	MSS	F-Test
Variation Due to Factor A	$r - 1$	SSA	$MSSA = \frac{SSA}{r - 1}$	$F_A = \frac{MSSA}{MSSE}$ $F_B = \frac{MSSB}{MSSE}$
Variation Due to Factor B	$s - 1$	SSB	$MSSB = \frac{SSB}{s - 1}$	
Variation Due to Error	$(r - 1)(s - 1)$	SSE	$MSSE = \frac{SSE}{(r - 1)(s - 1)}$	
Total	$rs - 1$	TSS		

Step 9: We take decisions about the null hypotheses for factor A and factor B as explained below:

- Compare the calculated value of F_A with tabulated value of F_A at $[(r - 1), (r - 1)(s - 1)]$ df. If calculated value is greater than the tabulated value then reject the hypothesis H_{0A} , otherwise it may be accepted.
- Compare the calculated value of F_B with tabulated value of F_B at $[(s - 1), (r - 1)(s - 1)]$ df. If calculated value is greater than the tabulated value then reject the hypothesis H_{0B} , otherwise, it may be accepted.

Now after discussing the procedure of two-way ANOVA test, let us practice and solve an example:

Example 3: An experiment was conducted to determine the effect of different dates of planting and different methods of planting of sugar-cane on the field. The data below show the fields of sugar-cane for four different data and the different methods of planting:

Method of Planting	Dates of Planting			
	October	November	February	March
I	7.1	3.69	4.7	1.9
II	10.29	4.79	4.58	2.65
III	8.3	3.58	4.9	1.8

Carry out an analysis of the above data.

Solution: Here, we can formulate the hypotheses as

H_{0A} : There is no difference in the average yield due to different methods of planting, that is,

$H_{0A} : \alpha_1 = \alpha_2 = \alpha_3$ **against,** H_{1A} = At least two are different H_{1A}

H_{0B} : There is no difference in the average yield due to different dates of planting, that is,

$H_{0B}: \beta_1 = \beta_2 = \beta_3 = \beta_4$ **against** $H_{1B} =$ At least two are different

$N =$ No. of observations $= 12$

We make some calculations as

Method of Planting	Dates of Planting				Total (M)
	October	November	February	March	
I	7.1	3.69	4.7	1.9	17.39
II	10.29	4.79	4.58	2.65	22.310
III	8.3	3.58	4.9	1.8	18.58
Total (D)	25.69	12.06	14.18	6.354	$G = 58.28$

$$G = \sum_{i=1}^r \sum_{j=1}^s y_{ij} = \text{Total of all observations}$$

$$= 7.10 + 3.69 + 4.70 + 1.90 + 10.29 + 4.79 + 4.58 + 2.64 + 8.30 + 3.58 + 4.90 + 1.80 = 58.28$$

$$\text{Correction Factor} = CF = \frac{G^2}{N} = \frac{58.28 \times 58.28}{12} = 283.0465$$

Raw Sum of Square (RSS) =

$$(7.10)^2 + (3.69)^2 + (4.70)^2 + (1.90)^2 + (10.29)^2 + (4.79)^2 + (4.58)^2 + (2.64)^2 + (8.30)^2 + (3.58)^2 + (4.90)^2 + (1.80)^2 = 355.5096$$

$$\text{Total Sum of Square (TSS)} = \text{RSS} - CF = 355.5096 - 283.0465 = 72.4631$$

$$\begin{aligned} \text{Sum of Squares due to Dates of Planting (SSD)} &= \frac{D_1^2}{3} + \frac{D_2^2}{3} + \frac{D_3^2}{3} + \frac{D_4^2}{3} - CF \\ &= \frac{(25.69)^2}{3} + \frac{(12.06)^2}{3} + \frac{(14.18)^2}{3} + \frac{(6.35)^2}{3} - 283.0465 \end{aligned}$$

$$\begin{aligned} \text{Sum of Squares to Method of Planting (SSM)} &= \frac{M_1^2}{4} + \frac{M_2^2}{4} + \frac{M_3^2}{4} - CF \\ &= \frac{(17.39)^2}{4} + \frac{(22.31)^2}{4} + \frac{(15.58)^2}{4} - 283.0465 \\ &= 286.3412 - 283.0465 = 3.2947 \end{aligned}$$

$$\begin{aligned} \text{Sum of Squares due to Error (SSE)} &= \text{TSS} - \text{SSD} - \text{SSM} \\ &= 72.4631 - 3.2947 - 65.8917 = 3.2767 \end{aligned}$$

$$\begin{aligned} \text{MSSD} &= \frac{65.8917}{3} = 21.9639 & \text{MSSM} &= \frac{3.2947}{2} = 1.6473 \\ \text{MSSE} &= \frac{3.2767}{6} = 0.5461 \end{aligned}$$

For testing $H_{0A} : F_M$ is $\frac{\text{MSSM}}{\text{MSSE}} = \frac{1.6473}{0.5461} = 3.02$

For testing $H_{0B} : F_D$ is $\frac{\text{MSSD}}{\text{MSSE}} = \frac{21.9639}{0.5461} = 40.22$

The tabulated value of $F_{2,6}$ at 5% level of significance is 5.14 which is greater than the calculated value of F_M (3.02) so H_{0A} is accepted. So, we conclude that there is no significant difference between the different methods of planting.

The tabulated value of $F_{3,6}$ at 5% level of significance is 4.76 which is less than calculated value of F_D (40.22). So, we reject the null hypothesis H_{0B} . Hence there is a significant difference among the dates of planting.

Let us now try to solve the questions of the following check your progress.

Check Your Progress 3

- 1) A researcher wants to test four diets A, B, C, D on growth rate in mice. These animals are divided into 3 groups according to their weights. Heaviest 4, next 4 and lightest 4 are put in Block I, Block II and Block III respectively. Within each block, one of the diets is given at random to the animals. After 15 days increase in weight is noted and given in the following table:

Block	Treatment/Diet			
	A	B	C	D
I	12	8	6	5
II	15	12	9	6
III	14	10	8	5

Perform a two-way ANOVA to test whether the data indicate any significant difference between the four diets due to different blocks.

.....

.....

.....

- 2) Future Group wishes to enter the frozen shrimp market. They contract a researcher to investigate various methods of groups of shrimp in large tanks. The researcher suspects that temperature and salinity are important factors influencing shrimp yield and conducts a two-way analysis of variance with their levels of temperature and salinity. That is each combination of yield for each (for identical gallon tanks) is measured. The recorded yields are given in the following table:

	Salinity (in pp)				
Temperature	700	1400	2100	Total	Mean
60° F	3	5	4	12	4
70° F	11	10	12	33	11
80° F	16	21	17	54	18
Total	30	36	33	99	11

Carry out an analysis of the above data.

.....

11.5 SUMMARY

In this unit, we have discussed:

1. Basic ideas and concepts related to the technique of analysis of variance as applied to test the equality of several population means simultaneously;
2. Types of data which can be analysed through ANOVA technique and the meaning of one-way and two-way classified data as well as the concept of one-way and two-way ANOVA;
3. Formation of null and alternative hypotheses to be tested and different steps of one-way ANOVA for testing the hypothesis under consideration;
4. Null and alternative hypotheses to be tested and different steps of two-way ANOVA for testing the hypothesis under consideration.

11.6 SOLUTIONS /ANSWERS

Check Your Progress 1

- 1) Referred to Section 11.2.
- 2) Referred to sub-section 11.2.1.

Check Your Progress 2

- 1) The null hypothesis states that there are no differences among the mean scores of the three sections (denoted by μ_i , $i = 1, 2, 3$), so that we have

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \text{At least two means differ.}$$

As we know, the sum squares between the samples is given by

$$SSB = \sum_{i=1}^k n_i (\bar{X}_i - \bar{\bar{X}})^2$$

In our case, we have 3 samples, therefore,

$$\bar{y}_1 = \frac{119}{7} = 17, \bar{y}_2 = \frac{84}{7} = 12, \bar{y}_3 = \frac{112}{7} = 16$$

$$\text{and hence, } \bar{\bar{y}} = \frac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3}{3} = \frac{17+12+16}{3} = 15$$

We can calculate the total sum of squares (TSS) as

$$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 = (20-15)^2 + (18-15)^2 + (18-15)^2 + \dots + (14-15)^2 = 156$$

and the sum of squares between the samples (SSB) as

$$\begin{aligned} SSB &= \sum_{i=1}^k n_i (\bar{y}_i - \bar{\bar{y}})^2 = n_1 (\bar{y}_1 - \bar{\bar{y}})^2 + n_2 (\bar{y}_2 - \bar{\bar{y}})^2 + n_3 (\bar{y}_3 - \bar{\bar{y}})^2 \\ &= 7(17-15)^2 + 7(12-15)^2 + 7(16-15)^2 \\ &= 28 + 63 + 7 = 98 \end{aligned}$$

$$SSE = TSS - SSB = 156 - 98 = 58$$

We can construct an ANOVA table for the problem solved above as follows:

ANOVA Table				
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio F
Treatment	SSB = 98	$(k - 1) = 2$	$MSSB = \frac{SSB}{(k-1)}$ $= \frac{98}{2} = 49$	$\frac{MSSB}{MSSE}$ $= 15.22$
Within (or error)	SSE = 58	$(N - k) = 18$	$MSSE = \frac{SSE}{(N-k)}$ $= \frac{58}{18} = 3.22$	
Total	SST = 156			

Now, we check for the critical value of F from the table for $\alpha = 0.05$ and degrees of freedom as follows:

$$df(\text{numerator}) = (k - 1) = (3 - 1) = 2$$

$$df(\text{denominator}) = (N - k) = (21 - 3) = 18$$

The tabulated (see Appendix) $F_{2,18}$ at 5% level of significance is 3.55. Hence, since $F_{\text{tab}} < F_{\text{cal}}$; H_0 is rejected implying that mean scores of three sections are not same.

2) In the usual notations, our null hypothesis is

$$H_0: \mu_1 = \mu_2 = \mu_3, \text{ and } H_1 \text{ at least two means are unequal.}$$

$$\text{Here } \bar{y}_1 = \frac{40}{4} = 10 \quad \bar{y}_2 = \frac{28}{4} = 7 \quad \bar{y}_3 = \frac{52}{4} = 13$$

$$\text{so that, } \bar{\bar{y}} = \frac{10+7+13}{3} = \frac{30}{3} = 10.$$

Then, we calculate TSS as

$$\text{TSS} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2 = (8-10)^2 + (9-10)^2 + (11-10)^2 + \dots + (8-10)^2 = 154$$

We can calculate the value of SSB as

$$\begin{aligned} \text{SSB} &= \sum_{i=1}^k n_i (\bar{y}_i - \bar{\bar{y}})^2 = n_1 (\bar{y}_1 - \bar{\bar{y}})^2 + n_2 (\bar{y}_2 - \bar{\bar{y}})^2 + n_3 (\bar{y}_3 - \bar{\bar{y}})^2 + n_4 (\bar{y}_4 - \bar{\bar{y}})^2 \\ &= 4(10-10)^2 + 4(7-10)^2 + 4(13-10)^2 \\ &= 0 + 36 + 36 = 72 \end{aligned}$$

$$\text{SSE} = \text{TSS} - \text{SSB} = 154 - 72 = 82$$

We can construct an ANOVA table for the problem solved above as follows:

ANOVA Table				
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio F
Treatment	SSB = 72	(k - 1) = 2	$\text{MSSB} = \frac{\text{SSB}}{(k-1)}$ $\text{MSSB} = \frac{72}{2} = 36$	$\frac{\text{MSSB}}{\text{MSSE}}$ $= 3.95$
Within (or error)	SSE = 82	(N - k) = 9	$\text{MSSE} = \frac{\text{SSE}}{(N-k)}$ $\text{MSSE} = \frac{82}{9} = 9.1$	
Total	SST = 154			

The value of F from the table (Table VII of Appendix) at 5% level of significance and degrees of freedom (2, 9) is given as 4.26.

Since our calculated value of F is less than the tabulated value of F, we cannot reject the null hypothesis.

Check Your Progress 3

- 1) We can formulate the null Hypotheses and the alternative hypotheses as:
 H_{01} : There is no significant difference between mean effects of diets.
 H_{11} : There is significant difference between mean effects of diets
 H_{02} : There is no significant difference between mean effects of different blocks.
 H_{12} : There is significant difference between mean effects of different blocks.

Blocks	Treatments/Diets				Totals
	A	B	C	D	
I	12	8	6	5	$B_1 = 31$
II	15	12	9	6	$B_2 = 42$
III	14	10	8	5	$B_3 = 37$
Totals	$T_1=41$	$T_2=30$	$T_3=23$	$T_4=16$	$B_4 = 110$

Squares of observations

Blocks	Treatments/Diets				Totals
	A	B	C	D	
I	144	64	36	25	269
II	225	144	81	36	486
III	196	100	64	25	385
Totals	565	308	181	86	1140

$$\text{Grand Total} = G = \sum_{i=1}^r \sum_{j=1}^s y_{ij} = 110$$

$$\text{Correction Factor (CF)} = \frac{G^2}{N} = \frac{(110)^2}{12} = 1008.3333$$

$$\text{Raw Sum of Squares (RSS)} = \sum_{i=1}^r \sum_{j=1}^s y_{ij}^2 = 1140$$

$$\text{Total Sum Squares (TSS)} = \text{RSS} - \text{CF} = 1140 - 1008.3333 = 131.6667$$

Sum of Squares due to Treatments/ Diets (SST)

$$= \frac{T_1^2}{3} + \frac{T_2^2}{3} + \frac{T_3^2}{3} + \frac{T_4^2}{3} - \text{CF}$$

$$= \frac{1}{3}[(41)^2 + (30)^2 + (23)^2 + (16)^2] - 1008.3333$$

$$= \frac{1}{3}[1681 + 900 + 529 + 256] - 1008.3333$$

$$= 1122 - 1008.3333 = 113.6667$$

$$\text{Sum of squares due to block (SSB)} = \frac{B_1^2}{4} + \frac{B_2^2}{4} + \frac{B_3^2}{4} - \text{CF}$$

$$= \frac{1}{4}[(31)^2 + (42)^2 + (37)^2] - 1008.3333$$

$$= 1023.5 - 1008.3333 = 15.1667$$

Sum of Squares due to Errors (SSE) = TSS – SST – SSB

$$= 131.6667 - 113.6667 - 15.1667 = 2.8333$$

$$\text{Mean Sum of Squares due to Treatments (MSST)} = \frac{\text{SST}}{\text{df}} = \frac{113.6667}{3} = 37.8889$$

$$\text{Mean Sum of Squares due to Blocks (MSSB)} = \frac{\text{SSB}}{\text{df}} = \frac{15.1667}{2} = 7.58335$$

$$\text{Mean Sum of Squares due to Errors (MSSE)} = \frac{\text{SSE}}{\text{df}} = \frac{2.8333}{6} = 0.4722$$

ANOVA TABLE

Source of Variation	SS	df	MSS	F
Between Treatments/ Diets	113.6667	3	$\frac{113.6667}{3} = 37.8889$	$F_1 = \frac{37.8889}{0.4722} = 80.2353$
Between Blocks	15.1667	2	$\frac{15.1667}{2} = 7.58335$	$F_2 = \frac{7.58335}{0.4722} = 16.0596$
Due to Errors	2.8333	6	$\frac{2.8333}{6} = 0.4722$	
Total	131.6667	11		

Tabulated F at 5% level of significance with (3, 6) degree of freedom is $F_1 = 4.76$ (See Appendix) & with (2, 6) degree of freedom is $F_2 = 5.14$ (See Appendix).

Conclusion: Since calculated $F_1 >$ tabulated F_1 (4.76), so we reject H_{01} and conclude that there is significant difference between mean effect of diets.

Also calculated F_2 is greater than tabulated F_2 (5.14), so we reject H_{02} and conclude that there is significant difference between mean effect of different blocks.

- 2) If α_i is the average effect of i^{th} level of temperature then we can set up the null and alternative hypotheses as follows:

$$H_{0T} : \alpha_1 = \alpha_2 = \alpha_3$$

$$H_{1T} = \text{At least two are different}$$

Similarly, if β_j is the average effect of j^{th} level of salinity then we can set up the null and alternative hypotheses as follows:

$$H_{0S} : \beta_1 = \beta_2 = \beta_3$$

$$H_{1S} = \text{At least two are different}$$

The computations are as follows:

Grand Total (G) = 99

No. of observations (N) = 9

Correction Factor CF = $(99 \times 99) / 9 = 1089$

Raw Sum of Square (RSS) = 1401

Total Sum of Square (TSS) = RSS - CF = $1401 - 1089 = 312$

Sum of Square due to Temperature (SST)

$$= (12)^2/3 + (33)^2/3 + (54)^2/3 - 1089 = 294$$

Sum of Square due to Salinity (SSS)

$$= (30)^2/3 + (36)^2/3 + (33)^2/3 - 1089 = 6$$

Sum of Square due to Error = TSS - SST - SSS = $312 - 294 - 6 = 12$

ANOVA Table

Sources of Variation	DF	SS	MSS	F-Test
Due to Temperature	2	294	147	$F_T = MSST/MSSE$ $= 147/3 = 49$
Due to Salinity	2	6	3	$F_S = 3/3 = 1$
Due to error	4	12	3	
Total	8	312		

Since the tabulated value of $F_{2,4}$ at 5% level of significance is 6.94 (See Appendix) which is less than the calculated F_T for testing the significant difference in shrimp yield due to differences in levels of temperature (49). So, H_{0T} is rejected. Hence, there are differences in shrimp yield due to temperature at 5% level of significance.

Since the tabulated value of $F_{2,4}$ at 5% level of significance is 6.94 (See Appendix) which is greater than the calculated F_S for testing the significant difference in shrimp yield due to difference in the level of salinity (1). So, H_{0S} is not rejected. Hence there is no any difference in shrimp yield due the salinity level of 5% level of significance.

11.7 REFERENCES

IGNOU Course Material

Master of Science (Environmental Science) (MSCENV), MEV 019: Research Methodology for Environmental Science, Block 3: Statistical Analysis UNIT 12: Analysis of Variance Tests (Entire Unit)