UNIT 11: F-DISTRIBUTION AND ANALYSIS OF VARIANCE (ANOVA)

UNIT STRUCTURE

11.1 Learning Objectives

11.2 Introduction

11.3 The F Distribution

11.4 Assumptions For F-Test

11.5 Properties Of F-Distribution

11.6 Applications Of F-Distribution

11.6.1 Hypothesis Testing for Difference between Two Population variances

11.7 Analysis Of Variance

11.8 Assumptions Of Analysis Of Variance

11.9 Classification Of Analysis Of Variance

11.9.1 One -Way Classification

11.9.2 Two-Way Classification

11.10 Let Us Sum Up

11.11 Further Readings

11.12 Answer To Check Your Progress

11.13 Model Questions

11.1 LEARNING OBJECTIVES

After going through this unit you will be able to

- introduce F distribution and learn how to use it in statistical inferences
- understand the assumptions involved in the use of analysis of variance
- to compare more than two population means using analysis of variance

11.2 INTRODUCTION

In several statistical applications a decision maker might want to compare the difference in two population variances. For example, (i) Quality of products, as measured by some indicators like diameter, length, etc. manufactured by two different production processes. The process with lower variance, implying more consistent quality, is preferred. (ii) Price or Earning

Per Share (EPS) of two shares on day-to-day basis, say for a week or month, etc. The share with lower variance implying lesser volatility could be considered less risky and as a result preferred by some investors. (iii) Service time taken by two systems or agencies—the one having lower variance is preferred and so on are some of the areas where comparison of variances is needed. In testing the hypotheses of the difference in two population variances, sample variances are used. Thus the ratio of the two sample variances $\frac{S_1^2}{S_2^2}$ is termed as F test. Therefore the test statistic used for the difference in two population variances

$$F = \frac{S_1^2}{S_2^2}$$

follows F distribution with $d.f. = n_1 - 1$ and $d.f. = n_2 - 1$, where S_1^2 and S_2^2 are defined by

$$S_1^2 = \frac{\sum\left(x_1 - \overline{x_1}\right)^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\sum\left(x_2 - \overline{x_2}\right)^2}{n_2 - 1}$$

$$f(F) = y_0 \left(F\right)^{\left(\frac{v_1}{2}\right)-1} \left(1 + \frac{v_1}{v_2} F\right)^{-(v_1+v_2)/2}$$

11.3 THE F DISTRIBUTION

F distribution is based on the assumption that the populations from which samples are drawn are normally distributed. This distribution was first originated by G.W.Snedecor and the distribution was named by Snedecor as F distribution in honour of the famous English statistician Sir Ronald A Fisher. However, this distribution is also called Snsdecore's F distribution. As in the case with and Chi-square distributions, the shape of a particular F distribution curve depends on the number of degrees of freedom. A major difference between Chi-square distributions and F distribution is that the former two distributions have only one number of degrees of freedom whereas the latter has two numbers of degrees of freedom. This is because in the F distribution, the degrees of freedoms are attached to both numerator and denominator. It is to be noted that the F distribution is a continuous probability distribution and is positively skewed. The probability density function of F distribution is given by

where $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ and y_0 is a constant which is so determined that the total area under the curve is unity.

11.4 ASSUMPTIONS FOR F-TEST

Few assumptions for the ratio $\frac{S_1^2}{S_2^2}$ to have an F distribution are given below:

(i) Independent random samples are drawn from each of two normal populations.

(ii) The variability of the measurements in the two populations is same and can be measured by a common variance σ^2, *i.e.*, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. The assumption is needed in order to combine the variances within the populations into a single source of variation 'within populations'.

11.5 PROPERTIES OF F DISTRIBUTION

Some important properties of F distribution are given below:

(i) The total area or probability under the curve is unity.

(ii) The F distribution is positively skewed with a range 0 to ∞ and its degree of skewness decreases with the increase in degrees of freedom for numerator and $v_2 = n_2 - 1$ for denominator. For $v_2 \geq 30,$ F distribution is approximately normal.

(iii) The mean of the F distribution is

$$\text{Mean} = \frac{v_2}{v_2 - 2}; v_2 > 2$$

Thus mean of F distribution is independent of the first degree of freedom . Obviously mean is greater than unity.

(iv) Variance of F distribution is

Variance =

(v) Mode of the distribution is

Mode = $\frac{v_2(v_1 - 1)}{v_1(v_2 + 2)}$

Since F>0, mode exists only if $\nu_1 > 2.$ Clearly mode is less than unity.

(vi) The curve of the F distribution rises steadily from 0 till the model value and then decreases slowly.

Critical values of F distribution: Critical values of F for different pairs of degrees of freedom at 1% level and 5% level of significance for the right-tailed test are available in F-table. The critical region corresponding to a specified level of significance and degrees of freedom is determined by the right-tail area. The significant value (i.e., critical value) at level of significance α and is determined by the equation

as drawn in the following figure.

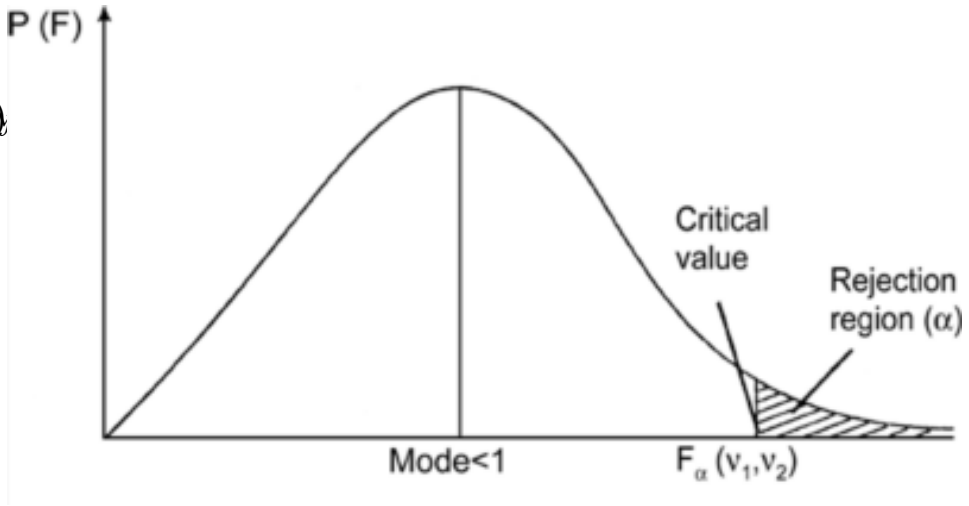


Figure 11.1



CHECK YOUR PROGRESS

Q 1. State whether the following statements are true or false:

(i) The F -distribution is positively skewed.

(ii) The -distribution is based on only one degrees of freedom.

(iii) The test statistic $F = S_1^2 / S_2^2$ is used to test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$.

11.6 APPLICATIONS OF F DISTRIBUTION

F distribution is widely used in business environment by decision maker. Some of the applications are outlined below:

(i) F test is applied to test the equality of population variances.

(ii) F test is used to test the significance of an observed sample multiple correlation.

(iii) It is used to test the significance of an observed sample correlation ratio.

(iv) It is applied for testing the linearity of regression.

(v) It is used to test the equality of several population means.

Hereunder we shall, of course, discuss only the application of testing the equality of population variances.

11.6.1 HYPOTHESIS TESTING FOR DIFFERENCE BETWEEN TWO POPULATION VARIANCES

Let us consider two independent random samples of sizes n_1 and n_2 drawn from two normal populations with variances σ_1^2 and σ_2^2. We want to test the hypothesis whether the variances of the two populations are equal or if there exists any significant difference between the population variances.

Thus the null hypothesis is $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$

The alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Under the null hypothesis, the test statistic to be constructed is

$$F = \frac{S_1^2}{S_2^2}$$

follows F distribution with $d.f. = \nu_1 = n_1 - 1$ and ,
where and S_2^2 are unbiased estimates of the common population variance σ^2 defined by

$$S_1^2 = \frac{\sum\left(x_1 - \overline{x_1}\right)^2}{n_1 - 1} \text{ and } S_2^2 = \frac{\sum\left(x_2 - \overline{x_2}\right)^2}{n_2 - 1}$$

x_1 and x_2 denote respectively the sample observations of the two random samples drawn from the two normal populations.

Since F test is based on the ratio of two variances it is also termed as Variance Ratio Test.

The criteria of acceptance or rejection of null hypothesis H_0 are as under:

Accept H_0 if the calculated value of the test statistic F is less than its critical value $F_\alpha(\nu_1, \nu_2)$ and otherwise reject the null hypothesis.

An Important Note: 1. Since the available tables of the significant values of F at different levels of significance are for the right-tail test, therefore in numerical problems we will take greater of the variances S_1^2 and S_2^2 in the numerator and adjust for the degrees of freedom accordingly. Thus if $F \sim$ then refers to the degrees of freedom of the larger variance, which must be taken in the numerator while calculating

$\phi_1^2 f. = \nu_2 = n_2 - 1$

2. In numerical problems, usually sample varianceis given from which can be obtained on using the relation

$$ns^2 = (n-1)S^2$$

where $s^2 = \frac{\sum\left(x - \overline{x}\right)^2}{n}$

Relation between t, F and χ^2 Distributions:

We give below (without proof), the relationship between and Distributions:

Relation between and distributions:

If a statistic t follows Student's t distribution with then its square follows

Snedecores's F distribution with (1, n) degrees of freedom. Symbolically,

$$t \sim t_n \quad \Rightarrow \quad t^2 \sim F(1, n)$$

with the help of this relation, all the applications of t distribution may be taken as applications of F distribution. For example, for testing the significance of sample mean, we may consider the test statistic

$$F = t^2 = \left[\frac{\bar{x} - \mu}{S/\sqrt{n}}\right]^2 \sim F(1, n)$$

and apply F-test.

Relation between F and χ^2 Distributions:

In F distribution with $(n_1, n_2)d.f.$ if we put

$$n_1 F = \chi^2 \Rightarrow F = \chi^2 / n_1$$

and let $n_2 \to \infty$, then we get Chi-square distribution with $n_1\, d.f.$

Example 11.1: The time taken by workers in performing a job by method I and method II is given below:

Method I: 19 22 24 27 24 18 20 19 25

Method II: 26 37 40 35 30 30 40 26 30 35 45

Do the data show that the variance of time distributions for populations from which these samples are drawn do not differ significantly?

Solution: Let us set up null and alternative hypotheses as

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ i.e., there is no significant difference between the variances of the time distribution by the workers in performing a job by Method I and Method II.

Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Under the null hypothesis, the test statistic to be constructed is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) \quad provided\ S_1^2 > S_2^2$$

where S_1^2 and S_2^2 are unbiased estimates of the common population variance σ^2 defined by

$$S_1^2 = \frac{\sum\left(x_1 - \overline{x_1}\right)^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\sum\left(x_2 - \overline{x_2}\right)^2}{n_2 - 1}$$

We now calculate sample variances

Method I Method II

x_1	$x_1 - 22$	$(x_1 - 22)^2$	x_2	$x_2 - 34$	$(x_2 - 34)^2$
19	-3	9	26	-8	64
22	0	0	37	3	9
24	2	4	40	6	36
27	5	25	35	1	1
24	2	4	30	-4	16
18	-4	16	30	-4	16
20	-2	4	40	6	36
19	-3	9	26	-8	64
25	3	9	30	-4	16
			35	1	1
			45	11	121
198	0	80	374	0	380

$$\overline{x_1} = \frac{\sum x_1}{n_1} = \frac{198}{9} = 22; \qquad S_1^2 = \frac{\sum\left(x_1 - \overline{x_1}\right)^2}{n_1 - 1} = \frac{80}{8} = 10;$$

$$S_2^2 = \frac{\sum\left(x_2 - \overline{x_2}\right)^2}{n_2 - 1} = \frac{380}{10} = 38$$

Since $S_1^2 > S_2^2$, under H_0, the test statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{38}{10} = 3.8$$

The critical value $F_{0.05}(10,8) = 3.35$

Since the calculated value of F is greater than the critical value therefore we reject the null hypothesis at 5% level and conclude that the variability of the time distribution in the two populations is not same.

Example 11.2: A plant has installed two machines producing polythene bags. During the installation, the manufacturer of the machine has stated that the capacity of the machine is to produce 20

bags in a day. Owing to various factors such as different operators working on these machines, raw material, etc. there is a variation in the number of bags produced at the end of the day. The company researcher has taken a random sample bags produced in 10 days for machine 1 and 13 days for machine 2, respectively. The following data gives the number of units of an item produced on a sampled day by the two machines:

Machine I: 18 19 19 18 17 19 18 19 18 19

Machine II: 18 17 17 17 16 18 18 16 18 17 16 16 17

Solution: Let us set up null and alternative hypotheses as

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ i.e., there is no significant difference between the variances of the time distribution by the workers in performing a job by Method I and Method II.

Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Under the null hypothesis, the test statistic to be constructed is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) \quad provided\ S_1^2 > S_2^2$$

where S_1^2 and S_2^2 are unbiased estimates of the common population variance σ^2 defined by

$$S_1^2 = \frac{\sum (x_1 - \overline{x_1})^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\sum (x_2 - \overline{x_2})^2}{n_2 - 1}$$

We now calculate sample variances

x_1	$x_1 - 18.4$	$(x_1 - 18.4)^2$	x_2	$x_2 - 17$	$(x_2 - 17)^2$
18	-0.4	0.16	18	1	1
19	0.6	0.36	17	0	0
19	0.6	0.36	17	0	0
18	-0.4	0.16	17	0	0
17	-1.4	1.96	16	-1	1
19	0.6	0.36	18	1	1
18	-0.4	0.16	18	1	1
19	0.6	0.36	16	-1	1
18	-0.4	0.16	18	1	1
19	0.6	0.36	17	0	0
			16	-1	1
			16	-1	1
			17	0	0
184	0	4.4	221	0	8

$$\overline{x_1} = \frac{\sum x_1}{n_1} = \frac{184}{10} = 18.4;$$

$$S_2^2 = \frac{\sum \left(x_2 - \overline{x_2}\right)^2}{n_2 - 1} = \frac{8}{12} = 0.66667$$

Since $S_2^2 > S_1^2$, under H_0, the test statistic is

$$F = \frac{S_2^2}{S_1^2}$$
$$= \frac{0.66667}{0.488889} = 1.363$$

which follows F-distribution with (12, 9) degrees of freedom.

The critical value $F_{0.05}(12,9) = 3.07$

Since the calculated value of F is less than the critical value therefore we may accept the null hypothesis at 5% level and conclude that there is no significant difference between the production capacity of the two machines.

$$S_1^2 = \frac{\sum (x_1 - \overline{x_1})^2}{n_2 n_1 - 1} = \frac{174}{13} = 0.488889;$$

Example 11.3: It is known that the mean diameters of rivets produced by two firms A and B are practically the same but the standard deviations may differ. For 22 rivets produced by firm A, the standard deviation is 2.9 mm, while for 16 rivets manufactured by firm B, the standard deviation is 3.8 mm. Compute the statistic you would use to test whether the products of firm A have the same variability as those of firm B and test its significance.

Solution: Given that

$$n_1 = 22, \qquad s_1 = 2.9\,mm.$$

$$n_2 = 16, \qquad s_2 = 3.8\,mm.$$

We state null and alternative hypotheses as

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ i.e., the products of both the firms A and B have the same variability.

Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Now, we have

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{22 \times (2.9)^2}{21} = 8.805$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{16 \times (3.8)^2}{15} = 15.393$$

Since $S_2^2 > S_1^2$, under H_0, the test statistic is

$$F = \frac{S_2^2}{S_1^2}$$
$$= \frac{15.393}{8.805} = 1.74821$$

which follows F-distribution with (15, 21) degrees of freedom.

The critical value $F_{0.05}(15, 21) = 2.20$

Since the calculated value of F is less than the critical value therefore we may accept the null hypothesis at 5% level and conclude that the hypothesis of equal variability may be accepted.

Example 11.4: Suppose a company manufacturing light bulbs is using two different processes A and B. The life of the light bulbs of process A has a normal distribution with mean μ_1 and standard deviation σ_1. Similarly the process B has also normal distribution with mean μ_2 and standard deviation σ_2. The data pertaining to the two processes are given below:

Sample A	Sample B
$n_1 = 16$	$n_2 = 21$
$\overline{x}_1 = 1200 hr.$	$\overline{x}_2 = 1300 hr.$
$s_1 = 60 hr.$	$s_2 = 50 hr.$

Test the hypothesis that the variability of the two processes is the same.

Solution: We state null and alternative hypotheses as

Null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ i.e., there is no significant difference in

the variability of the two processes..

Alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$

Now, we have

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{16 \times (60)^2}{16 - 1} = 3840$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{21 \times (50)^2}{21 - 1} = 2625$$

Since $S_1^2 > S_2^2$, under H_0, the test statistic is

$$F = \frac{S_1^2}{S_2^2}$$
$$= \frac{3840}{2625} = 1.46$$

which follows F-distribution with (15, 21) degrees of freedom.

σ.

The critical value $F_{0.05}(15, 20) = 2.20$

Since the calculated value of F is less than the critical value therefore we may accept the null hypothesis at 5% level and conclude that there is no significant difference in the variability of the two samples.

Example 11.5: Can the following two samples be regarded as coming from the same normal population?

Sample	Size	Sample mean	Sum of squares of deviations from the mean
1	10	12	120
2	12	15	314

Solution: A normal population has two parameters, viz., mean μ and standard deviation To test whether two independent samples have been drawn from the same normal population, we have to test:

(i) the equality of population means

and (ii) the equality of population variances.

We set up the null hypothesis i.e., the two samples can be regarded as drawn from the same normal population.

To test the equality of population means we apply the $t-$test (sizes of both samples are small) and equality of populations variances can be performed with the help of F-test. Since test assumes, therefore first we apply F-test and then apply test.

F-test:

Null hypothesis

Given

$$n_1 = 10, \quad n_2 = 12; \quad \overline{x_1} = 12, \quad \overline{x_2} = 15$$

$$\sum \left(x_1 - \overline{x_1}\right)^2 = 120, \qquad \sum \left(x_2 - \overline{x_2}\right)^2 = 314$$

$$\therefore S_1^2 = \frac{\sum \left(x_1 - \overline{x_1}\right)^2}{n_1 - 1} \qquad \therefore S_2^2 = \frac{\sum \left(x_2 - \overline{x_2}\right)^2}{n_2 - 1}$$

$$= \frac{120}{9} \qquad = \frac{314}{11}$$

$$= 13.33$$

Since , under H_0, the test statistic is

$$F = \frac{S_2^2}{S_1^2}$$

$$= \frac{28.55}{13.33} = 2.14$$

which follows F-distribution with (10, 9) degrees of freedom.

The critical value $F_{0.05}(11,9) = 3.10$

Since the calculated value of F is less than the critical value therefore we may accept the null hypothesis at 5% level and the assumption of equality of population variances is valid.

Since $\sigma_1^{\,2} = \sigma_2^{\,2}$, we can now apply $t-$test to test

$t-$Test:

Under the null hypothesis against the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$, the test statistic is

$$t = \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum\left(x_1 - \overline{x_1}\right)^2 + \sum\left(x_2 - \overline{x_2}\right)^2\right]$$

$$= \frac{1}{20}[120 + 314]$$

$$= 21.7$$

$$= -1.51$$

The degrees of freedom = $n_1+n_2-2=20$

$H_0 : \mu_1 = \mu_2$ $t = \frac{-15}{\sqrt{21.7\left(\frac{1}{10} + \frac{1}{12}\right)}} = \frac{-15}{\sqrt{21.7 \times \frac{11}{60}}}$

The critical value $t_{0.05}(20)$ = 2.086.

Since the calculated value of $|t|$ is less than the critical value of t at 5% level for 20 d.f., therefore we may accept the null hypothesis at 5% level of significance.

Since both the hypotheses and $H_0 : \sigma_1^2 = \sigma_2^2$ are accepted, therefore we may regard the given samples to have been drawn from the same normal population.



CHECK YOUR PROGRESS

Q 2: Determine the degrees of freedom of F - value for a sample of six observations in the numerator and four in the denominator? What is the critical value of for two-tailed at 5% level of significance?

Q 3: A random sample of eight observations from the first population resulted in a standard deviation of 10. A random sample of six observations from the second population resulted in a standard deviation of 7. At the 0.01 significance level, is there a significant difference in the variation of the two populations?

(a) State the null hypothesis and alternative hypothesis, (b) State the decision rule, (c) Compute the test statistic. (d) What is your decision regarding?

11.7 ANALYSIS OF VARIANCE

In unit 9, we have introduced hypothesis testing procedures to test the significance of difference between two sample means or equivalently to test if there is any significant difference between the two population means from which the samples have been drawn. For this, one test statistic employed was the t-test where we assumed that the two populations from which the samples were drawn had the same variance. But in real life, there may be situations when instead of comparing two sample means, a researcher has to compare three or more than three sample means (specifically, more than two). A researcher may have to test whether the three or more sample means computed from the three populations are equal. In other words, the null hypothesis can be that three or more population means are equal as against the alternative hypothesis that these population means are not equal. For example, suppose that a researcher wants to measure work attitude of the employees in four organizations. The researcher has prepared a questionnaire consisting of 10 questions for measuring the work attitude of employees. A five-point rating scale is used with 1 being the lowest score and 5 being the highest score. So, an employee can score 10 as the minimum score and 50 as the maximum score. The

null hypothesis can be set as all the means are equal (i.e., there is no difference in the degree of work attitude of the employees) as against the alternative hypothesis that at least one of the means is different from the others (there is a significant difference in the degree of work attitude of the employees). In this situation, analysis of variance technique is used. "In its simplest form analysis of variance may be regarded as an extension or development of the $t-test$." Analysis of variance technique makes use of F distribution (F-statistic). Some more examples are presented below where we are required to test the equality of means of three or more populations. For example, whether:

(1) The average life of light bulbs being produced in three different plants is the same.

(2) All the three varieties of fertilizers have the same impact on the yield of rice.

(3) The level of satisfaction among the participants in all IIMs is the same.

(4) The impact of training on salesmen trained in three institutes is the same.

(5) The service time of a transaction is the same on four different counters in a service unit.

(6) The average price of different commodities in four different retail outlets is the same.

(7) Performance of salesmen in four zones is the same.

The word 'analysis of variance' is used since the technique involves first finding out the total variation among the observations in the collected data, then assigning causes or components of variation to various factors and finally drawing conclusions about the equality of means. Thus Analysis of variance or ANOVA can be defined as a technique of testing hypotheses about the significant difference in several population means. This statistical technique was developed by R.A.Fisher. The main purpose of analysis of variance is to detect the difference among various population means based on the information gathered from the samples (sample means) of the respective populations.

11.8 ASSUMPTIONS OF ANALYSIS OF VARIANCE

Analysis of variance is based on some assumptions.

(i) Each sample is a simple random sample.

(ii) Populations from which the samples are selected are normally distributed. However, in case of large samples this assumption is not required.

(iii) Samples are independent.

(iv) The population variances are identical, i.e.,

Computation of Test Statistic:

The beauty of the technique of analysis of variance is that it performs the test of equality of more than two population means by actually analyzing the variance. In simple terms, ANOVA decomposes the total variation into two components of variation, namely, variation between the samples known as the mean square between samples and variation within the samples known as the mean square within samples. The variance ratio denoted by F is given by:

$$F = \frac{Mean\ square\ between\ samples\ or\ groups}{Mean\ square\ within\ samples\ or\ groups}$$

If the calculated value of F is greater than the critical value of, we must reject the null hypothesis. In case the calculated value of is less than the critical value of, we are to retain or accept the null hypothesis.

Analysis of variance table (ANOVA):

The table showing the sources of variation, the sum of squares, degrees of freedom, mean squares and the formula for the F statistic is known as ANOVA table.

11.9 CLASSIFICATION OF ANALYSIS OF VARIANCE

ANOVA is mainly carried on under the following two classifications:

(i) One-way classification

(ii) Two- way classification

Variance and its different components may be obtained in each of the two types of classification by:

(a) Direct Method, (b) Short-cut Method and (c) Coding Method

11.9.1 ONE-WAY CLASSIFICATION

Many business applications involve experiments in which different populations (or groups) are classified with respect to only one attribute of interest such as (i) percentage of marks secured by students in a course, (ii) flavor preference of ice-cream by customers, (iii) yield of crop due to varieties of seeds, and so on. In all such cases observations in the sample data are classified into several groups based on a single attribute (i.e., criterion) and are termed one-way classification of sample data.

Under this one-way classification we set up the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ where $\mu_1, \mu_2, ..., \mu_k$ are the arithmetic means of the populations from which the k samples are drawn.

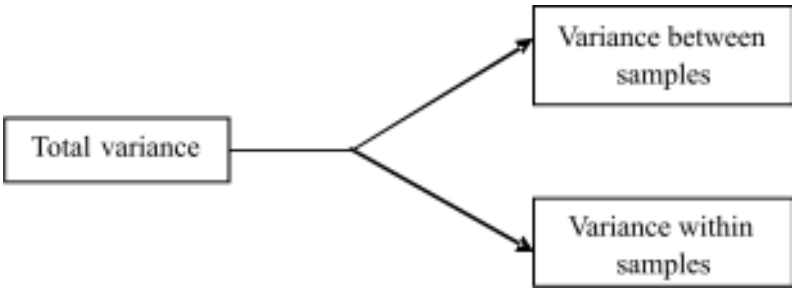
The alternative hypothesis is $H_1 : \mu_1 \neq \mu_2 \neq ..... \neq \mu_k$

After formulation of the null hypothesis and alternative hypothesis one needs to calculate the following by using any one of the above mentioned methods.

(i) Variance between the samples.
(ii) Variance within the samples.
(iii) Total variance by summing (i) and (ii) which may also be calculated directly for verification of calculations.

(iv) F-ratio (or F-statistic)

Thus in this process, the total variance can be divided into two additive and independent parts as shown:



After calculating the test statistic F, it should be compared with the critical value of F at a specified level of significance α for degrees of freedom and on the basis of this comparison; accordingly the decision to accept or reject the null hypothesis is taken.

(a) Direct Method

I. Calculation of variance between samples

It is the sum of squares of the deviations of the means of various samples from the grand mean. The procedure of calculating the variance between the samples is as shown below:

Observations	Number of samples				
	1	2	……j	……	k
1	x_{11}	x_{12}	…… x_{1j}	……	x_{1k}
2	x_{21}	x_{22}	……x_{2j}	……	x_{2k}
.	.	.	.		.
i	x_{i1}	x_{i2}	……x_{ij}	……	x_{ik}
.	.	.			.
N	x_{n1}	x_{n2}	……x_{nj}	……	x_{nk}
Total	T_1	T_2	……T_j	……	T_k
A.M.	$\overline{x_1}$	$\overline{x_2}$	……$\overline{x_j}$	……	$\overline{x_k}$

Where

$$T_j = \sum_{i=1}^{n} x_{ij}\ , \qquad T = \sum_{j=1}^{k} T_j\ , \qquad \overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad \text{and } \bar{x} = \frac{1}{nk}\sum_{j=1}^{k} \overline{x_j}$$

The steps in calculating variance between samples are given by:

(i) In the first step, we need to calculate the mean of each sample i.e.,

of all the k samples.

(ii) Next, the grand mean is calculated by using the formula

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \ldots\ldots\ldots\ldots + \bar{x}_k}{k} = \frac{T}{N}$$

where T =Grand total of all observations.

N =Total number of observations in all k samples.

(iii) In step 3, the difference between the mean of each sample and grand mean is calculated, that is we calculate $\bar{x}_1 - \bar{\bar{x}}, \bar{x}_2 - \bar{\bar{x}}, \ldots\ldots\ldots, \bar{x}_k - \bar{\bar{x}}$

(iv) In step 4, we square the deviations obtained in step (iii) and multiply by the number of items in the corresponding sample and then add the total. This total gives the sum of the squares of the deviations between the samples (or between the columns) and it is denoted by SSB or SSC

$\bar{x}_1, \bar{x}_2, \ldots\ldots\ldots, \bar{x}_k$

$$\text{Thus} \quad \text{MSB} = \frac{SSB}{k-1}$$

$$\text{Thus} \quad \text{SSB} = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

where is the number of groups or samples being compared, n_j the number of observations in group j, the sample mean of group , and the grand mean.

(v) In the last step, the total obtained in step (iv) is divided by the degrees of freedom. The degrees of freedom is one less than the total number of samples. If there are k samples, the degrees of freedom will be $\nu = k - 1.$ When the sum of squares obtained in step (iv) is divided by the number of degrees of freedom, the result is called mean square which is denoted by MSB or MSC. MSB indicates the degree of explained variance due to sampling fluctuations.

II. Calculation of Variance within the samples:

This is usually referred to as the sum of squares within samples. The variance within samples measures the difference within the samples due to chance. This is usually denoted by SSE. The steps involved are the following:

(i) The first step is to calculate the mean values of all k samples,

(ii) Calculate the deviations of the various observations of k samples from the mean values of the respective samples,

(iii) Square all the deviations obtained in (ii) and find the total of these squared deviations. This total gives the sum of the squares of deviations within the samples or the sum of squares due to error. It is denoted by SSW or SSE.

Thus $$\text{SSW or SSE}=\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij}-\overline{x_j}\right)$$

where is the ith observation in the jth group, $\overline{x_j}$ the sample mean of jth group, k the number of groups being compared, and N the total number of observations in all the groups.

(iv) As the last step, divide the total squared deviations obtained in step (iii) by the degrees of freedom and obtain the mean square. The number of degrees of freedom can be calculated as the difference between the total number of observations and the number of samples. If there are N observations and k samples then the degrees of freedom is $\nu = N - k.$

III. Calculate total sum of squares: The total variation is equal to the sum of the squared difference between each observation (sample value) and the grand mean and is often referred to as SST.

Thus SST can be calculated as:

SST=SSB+SSE

III. Calculation of the Test Statistic F:

When the null hypothesis is true, both mean squares MSB (MSC) and MSW (MSE) are the independent unbiased estimates of the same population variance σ^2. Hence, the test statistic is

$$F = MSB/MSW \text{ Or, } F = MSC/MSE$$

which follows F-distribution with degrees of freedom $(k-1, N-k)$

F is the ratio between the greater variance to the smaller variance. Generally, variance between the samples (MSB or MSC) is greater than the variance within the samples (MSW or MSE). But if MSW>MSB then the reverse ratio of F will be used i.e.,

$$F = MSW/MSB \quad or, F = MSE/MSC$$

IV. Conclusion:

To compare the calculated value of F with the critical value (tabulated) of F for $(k-1, N-k)$ degrees of freedom at the specified level of significance, usually 5% or 1% level. If the calculated value is greater than the tabulated value, the null hypothesis H_0 is rejected and conclude that all population means are not equal. Otherwise, the null hypothesis is accepted.

For analyzing variance in case of one way classification the following table known as the Analysis of Variance Table (or ANOVA Table) is constructed.

ANOVA Table

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)	Test Statistic
Between samples	SSB	$k-1$	MSB=$\frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within samples	SSW	$N-k$	MSW=$\frac{SSW}{N-k}$	
Total	SST	$N-1$		

(b) Short-cut Method: The calculation of F-statistic (variance ratio) by using the direct Method is very time consuming. In practice, a short-cut Method based on sum of squares of the individual values (observations) is usually used. The computational work is much minimized in this method. The method is more convenient when some or all the sample means and the grand mean are fractional. The various steps involved in the calculations of variance ratio are the following:

(i) Calculate the grand total of all observations in samples, T

$$T = \sum x_1 + \sum x_2 + \ldots\ldots\ldots + \sum x_k$$

(ii) Calculate the correction factor $CF = \frac{T^2}{N}$; N= Total observations in samples.

(iii) Find the sum of the squares of all observations in samples from each of k samples and subtract CF from this sum to obtain the total sum of the squares of deviations SST:

$$SST = \left(\sum x_1^2 + \sum x_2^2 + \ldots\ldots\ldots\ldots + \sum x_k^2\right) - CF$$

where n_j is the size of the j[th] sample.

And $SSE = SST - SSB$

(iv) MSB, MSW and F are obtained as in the Direct method and decision either to accept or to reject is exactly same as taken in case of Direct method.

(c) Coding Method:

When the magnitudes of the observations (values) in various samples are large then Short-cut-Method also becomes tedious. The variance ratio F (or the F statistic) has the property that it remains unchanged if all the values are multiplied or divided by a common factor or if a common factor is added to or subtracted from each value. Owing to this property big figures (values or observations) are reduced in magnitude by subtraction or division and the computational work is

much simplified without altering the value of F. When the magnitude of data is reduced by subtractions or divisions, it is said that data have been coded to simplify calculations.

Example 11.6: The following data give the yield on 12 plots of land in three samples of 4 plots each, under three varieties of fertilizers A, B and C

A	B	C
25	20	24
22	17	26
24	16	30
21	16	20

Test whether there is any significant difference in the average yields of land under three varieties of fertilizers.

Solution: Here first we set up null and alternative hypotheses

Null hypothesis There is no significant difference in the average yields under the three varieties.

$H_0 : \overline{X_3} = \frac{\sum X_3}{n_3} = \frac{100}{4} = 25$

Alternative hypothesis H_1 : There is significant difference in the average yield under the three varieties.

We calculate sample means, the variance between the samples and the variance within the samples by using the Direct Method.

	Sample I X_1	Sample II X_2	Sample III X_3
	25	20	24
	22	17	26
	24	16	30
	21	16	20
Total	92	72	100

$$\overline{X_1} = \frac{\sum X_1}{n_1} = \frac{92}{4} = 23$$

Grand mean ( ) $= \dfrac{\overline{X_1} + \overline{X_2} + \overline{X_3}}{3}$

$$= \frac{23+18+25}{3} = \frac{66}{3} = 22$$

=104

Degrees of freedom,

Now MSW=Mean square between the samples

Calculation for SSW

Sample I		Sample II		Sample III	
X_1	$(X_1 - \overline{X_1})^2$	X_2	$(X_2 - \overline{X_2})^2$	X_3	$(X_3 - \overline{X_3})^2$
25	4	20	4	24	1
22	1	17	1	26	1
24	1	16	4	30	25
21	4	19	1	20	25
	10		10		52

SSW = Sum of squares within the samples

= 10+10+52=72

The degrees of freedom, $\nu_2 = N - k = 9$

MSW=72/9=8

Now we prepare the Analysis of Variance Table

Analysis of Variance Table (ANOVA)

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)	Test Statistic
Between samples	104	2	52	F=52/8=6.5
Within samples	72	9	8	
Total	SST=176	N-1=11		

The critical value of F for $\nu_1 = 2$ and at 5% level i.e., is 4.26.

Since the calculated value of F (i.e., 6.5) is greater than the critical value $F_{0.05}(2,9)$, herefore we reject the null hypothesis at 5% level and conclude that the difference in the average yields under the three varieties is significant.

Solution of the above problem by Short cut Method:

After formulating the null and alternative hypotheses as in the Direct Method one can proceed with the Short cut Method as follows:

$\frac{T^2}{N} = \frac{(264)^2}{12} = 5808$ Calculation of T, SST, SSB, SSW

Sample		Sample I			Sample I
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
25	625	20	400	24	576
22	484	17	289	26	676
24	576	16	256	30	900
21	441	19	361	20	400
92	2126	72	1306	100	2552

T = Sum of all the values in the three samples

$= \sum X_1 + \sum X_2 + \sum X_3 = 92 + 72 + 100 = 264$

Correction factor =

SST = Total sum of squares

$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2\right) - \frac{T^2}{N}$

= (2126+1306+2552)-5808 = 176

SSB = Sum of squares between the samples

= 2116+1296+2500 - 5808

= 104

Degrees of freedom = k-1=2

SSW = Sum of squares within the samples.

= SST-SSB = 176-104 = 72

Degrees of freedom = N – k=12-3=9

Now proceeding with the ANOVA table as in the Direct Method, one can arrive at the same conclusion.

Example 11.7: A study investigated the perception of corporate ethical values among individuals specializing in marketing. Using $\alpha = 0.05$ and the following data (higher scores indicate higher ethical values), test for significant differences in perception among three groups

Marketing Manager	Marketing Research	Advertising
6	5	6
5	5	7
4	4	6
5	4	5
6	5	6
4	4	6

Solution: We set up the null hypothesis that there is no significant difference in ethical values among individuals specializing in marketing.

Now we calculate sample means, the variance between the samples and the variance within the samples by using the Short cut Method.

Marketing Manager		Marketing Research		Advertising	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
6	36	5	25	6	36
5	25	5	25	7	49
4	16	4	16	6	36
5	25	4	16	5	25
6	36	5	25	6	36
4	16	4	16	6	36
30	154	27	123	36	218

T = Sum of all the values in the three samples

$$= \sum X_1 + \sum X_2 + \sum X_3 = 30 + 27 + 36 = 93$$

CF = Correction factor =

SST = Total sum of squares

$$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2\right) - \frac{T^2}{N}$$

= (154+123+218)-480.50 = 14.50

SSB = Sum of squares between the samples

$$= \frac{(30)^2}{6} + \frac{(27)^2}{6} + \frac{(36)^2}{6} - 480.50$$

= 150+121.5+216 -480.50

$$\frac{T^2}{N} = \frac{(93)^2}{18} = 480.50 \qquad \frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_3\right)^2}{n_3} - \frac{T^2}{N}$$

= 7

Degrees of freedom = k-1=2

SSW = Sum of squares within the samples.

= SST-SSB = 14.50-7 = 7.50

Degrees of freedom = N – k=18-3=15

Now we prepare the Analysis of Variance Table

Analysis of Variance Table (ANOVA)

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)	Test Statistic
Between samples	7	2	7/2=3.5	F=3.5/0.5=7
Within samples	7.5	15	7.5/15=0.5	
Total	14.5	N-1=17		

The critical value of F for $\nu_1 = 2$ and $\nu_2 = 15$ at 5% level i.e., $F_{0.05}(2,15)$ is 3.68.

Since the calculated value of F (i.e., 7) is greater than the critical

value $F_{0.05}(2,15)$, therefore we reject the null hypothesis at 5% level and conclude that there is significant difference in ethical values among individuals specializing in marketing.

Example 11.8: Three samples of five, four and five scooters are drawn respectively from brands A, B and C manufactured by three companies. The mileage of these scooters (in kilometer) is given below:

Test whether there is any significant difference between the mean mileages of the three brands of scooters. Use coding method.

A	B	C
45	41	44
42	40	42
43	42	38
44	43	43
42		39

Solution: We set up the null hypothesis H_0 and the alternative hypothesis H_1 as

H_0 : There is no significant difference between the mean mileages of the three brands of scooters.

H_1 : There is significant difference between the mean mileages of the three brands of scooters.

Let us subtract 40 from each of the sample values and let X_1, X_2 and X_3 denote respectively the coded sample values in respect of brand A, B and C respectively. We make the following table:

Calculation of T, SST, SSB, SSW

Sample		Sample II		Sample III	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
5	25	1	1	4	16
2	4	0	0	2	4
3	9	2	4	-2	4
4	16	3	9	3	9
2	4	..	..	-1	1
16	58	6	14	6	34

Now T = Sum of all the values in the three samples.

$= \sum X_1 + \sum X_2 + \sum X_3 = 16 + 6 + 6 = 28$

CF = Correction factor =

SST = Total sum of squares

$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2\right) - \frac{T^2}{N}$

= (58+14+34)-56 =50

SSB = Sum of squares between the samples

$= \frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_3\right)^2}{n_3} - \frac{T^2}{N}$

$= \frac{(16)^2}{5} + \frac{(6)^2}{4} + \frac{(6)^2}{5} - 56$

= 51.2+9+7.2 -56

= 11.4

$\frac{T^2}{N} = \frac{(28)^2}{14} = 56$

Degrees of freedom, = k-1=2

SSW = Sum of squares within the samples.

= SST-SSB = 50-11.4 =38.6

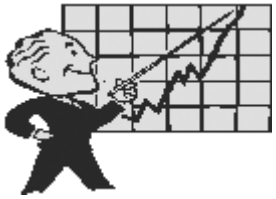
Degrees of freedom, = N – k=14-3=11

Analysis of Variance Table (ANOVA)

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)	Test Statistic
Between samples	11.4	2	11.4/2=5.7	F=5.7/3.51=1.624
Within samples	38.6	11	38.6/11=3.51	
Total	50	N-1=13		

The critical value of F for and $\nu_2 = 11$ at 5% level i.e., $F_{0.05}(2,11)$ is 3.98.

Since the calculated value of F (i.e., 1.624) is less than the critical value $F_{0.05}(2,11)$, therefore we may accept the null hypothesis at 5% level and conclude that there is no significant difference between the mean mileage of the three brands of scooters.



CHECK YOUR PROGRESS

Q 4. State whether the following statements are true or false:

(i) Analysis of variance is used to test the hypothesis that the means of several populations do not differ.

(ii) For analysis of variance, samples drawn from populations need not be independent.

(iii) The number of parts in which total variance in a one-way analysis of variance partitioned is 2.

(iv) The sum of squares within a one-way analysis of variance is SST+SSB.

(v) The test statistic used for equality of k population means is MSB/MSE.

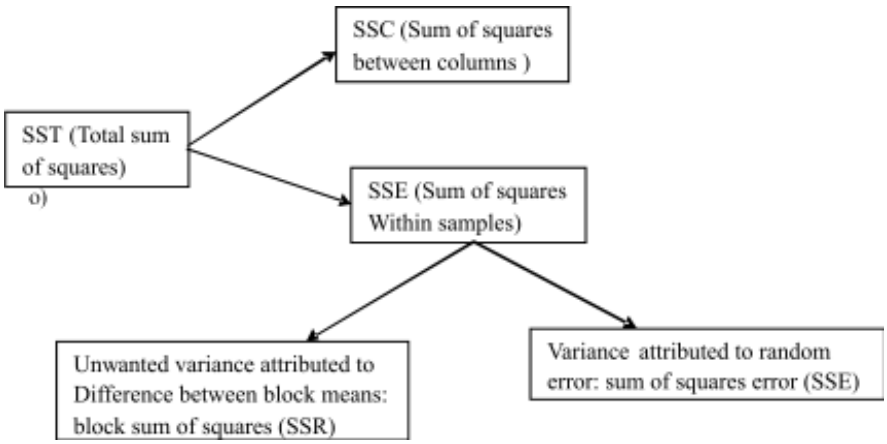
11.9.2 TWO-WAY CLASSIFICATION

In one-way ANOVA the total variation is partitioned into two components: variations between the samples or columns and variation within the samples due to random error. However, there might be a possibility that some of the variation left in the random error from one-way analysis of variation was not due to random error or chance but due to some other measurable factors. If this measurable factor is included in the MSE, it will result in an increase in the MSE. Any increase in the MSE would result in a small value of F (MSE being a denominator in the F statistic formula), which would ultimately lead to the acceptance of the null hypothesis.

The two-way classification is a mere extension of the one-way ANOVA. It can be used to investigate two criteria (factors) of interest for testing the difference between sample means. For example, we can study the effect of the two factors namely different varieties of seeds and different varieties of fertilizers on the yield of a crop in several plots of land.

The total variance or the sum of squares of variation in a two-way classification consists of three parts: (i) SSC i.e., the sum of

squares between the columns or variance between the columns, (ii) SSR i.e., the sum of squares between the rows or variance between the rows and (iii) SSE i.e., the sum of squares for the residuals or residual variance or the sum of squares due to error. Thus the partitioning of the total variation in the sample data is shown as



The total number of degrees of freedom = cr-1,

where c and r are the number of columns and the number of rows respectively.

Degrees of freedom for columns = c-1

Degrees of freedom for rows = r-1

Degrees of freedom for residuals= (cr-1)-(c-1)-(r-1)

= (c-1)(r-1)

The general ANOVA Table in case of two-way classification is given below:

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)
Between columns	SSC	$c-1$	MSC=SSC/(c-1)
Between rows	SSR	$r-1$	MSR=SSR/(r-1)
Residual	SSE	$(c-1)(r-1)$	MSE=SSE/(c-1)(r-1)
Total	SST	$N-1$	

The test statistic F for variance between columns is given by

$F = \frac{MSC}{MSE} = \frac{Variance\ between\ columns}{Residual\ variance}$, provided MSC > MSE

If the calculated value of F is greater than the critical value $F_{0.05}((c-1),(c-1)(r-1))$ then we reject the null hypothesis at 5% level and conclude that the difference between the column values is significant.

If MSE > MSC then we shall take $F = \frac{MSE}{MSC}$

Again the test statistic F for variance between rows is given by

$F = \frac{MSR}{MSE} = \frac{Variance\ between\ rows}{Residual\ variance}$ provided MSR > MSE

If MSE > MSR then we shall take $F = \frac{MSE}{MSR}$

If the calculated value of F is greater than the critical value $F_{0.05}((r-1),(c-1)(r-1))$ then we reject the null hypothesis at 5% level and conclude that the difference between the row values is significant.

Example 11.9: The following table gives the number of refrigerators sold by 4 salesmen in three months May, June and July:

Month	Salesmen			
	A	B	C	D
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

Is there any significant difference in the sales made by the four salesmen? Is there a significant difference in the sales made during different months?

Solution: Let us take the null hypothesis that (a) there is no significant difference between sales made by the four salesmen and (b) there is no significant difference in the sales made during different months. The given data are coded by subtracting 40 from each observation. The coded data are given below:

Month	Salesman								Row sum
	$A(x_1)$	x_1^2	$B(x_2)$	x_2^2	$C(x_3)$	x_3^2	$D(x_4)$	x_4^2	
May	10	100	0	0	8	64	-1	1	17
June	6	36	8	64	10	100	5	25	29
July	-1	1	4	16	0	0	-1	1	2
Column sum	15	137	12	80	18	164	3	27	48

From the above table, we work out the requisite sums of squares for preparing the ANOVA Table as follows:

T = Sum of all observations in three samples of months = 48

CF = Correction factor = $\frac{T^2}{N} = \frac{(48)^2}{12} = 192$

SSC = Sum of squares between salesmen (Columns)

$$= \frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_3\right)^2}{n_3} + \frac{\left(\sum X_4\right)}{n_4} - \frac{T^2}{N}$$

$$= \frac{(15)^2}{3} + \frac{(12)^2}{3} + \frac{(18)^2}{3} + \frac{(3)^2}{3} - 192$$

= (75+48+108+3) - 192

= 42

Degrees of freedom, $df_c = c - 1 = 4 - 1 = 3,$

SSR = Sum of squares between months (rows)

$$= \frac{(17)^2}{4} + \frac{(29)^2}{4} + \frac{(2)^2}{4} - 192$$

= (72.25+210.25+1) - 192

= 91.5

Degrees of freedom, $df_r = r - 1 = 3 - 1 = 2$

SST = Total sum of squares

$$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2\right) - \frac{T^2}{N}$$

= (137+80+164+27) –192

= 216

Degrees of freedom, $df = N - 1 = 12 - 1 = 11$

Residual or Error Sum of Squares, i.e., $SSE = SST - (SSC + SSR)$

= 216 – (42+91.5)

= 82.5

Degrees of freedom,

Thus $$MSC = \frac{SSC}{(c-1)} = \frac{42}{3} = 14$$

$$MSR = \frac{SSR}{(r-1)} = \frac{91.5}{2} = 45.75$$

$$MSE = \frac{SSE}{(c-1)(r-1)} = \frac{82.5}{6} = 13.75$$

The above summarized results are presented in the following ANOVA Table as given below:

Sources of Variation	Sum of squares (SS)	Degrees of Freedom ($d.f.$)	Mean Squares(MS)	Test statistic
Between salesmen	42.0	3	14.00	$F_{treatment} = 14/13.75 = 1.018$
Between months	91.5	2	45.75	$F_{block} = 45.75/13.75 = 3.327$
Residual error	82.5	6	13.75	
Total	216	11		

(a) The critical value of F at 5% level for (3, 6) degrees of freedom is 4.75.

Since the calculated value of $F_{treatment} = 1.018$ is less than the critical value, therefore we may accept the null hypothesis and conclude that sales made by the salesmen do not differ significantly.

(b) The critical value of F at 5% level for (2, 6) degrees of freedom is 5.14.

Since the calculated value of $F_{block} = 3.327$ is less than the critical value, hence we may accept the null hypothesis and conclude that sales made during different months do not differ significantly.

Example 11.10: To study the performance of three detergents and three different water temperatures, the following 'whiteness' readings

were obtained with specially designed equipment:

Water Temperature	Detergent A	Detergent B	Detergent C
Cold water	57	55	67
Warm water	49	52	68
Hot water	54	46	58

Perform a two-way s analysis of variance, using 5% level of significance.

Solution: Let us state the null hypothesis that there is no significant difference in the performance of three detergents due to water temperature and vice-versa.

The data are coded by subtracting 50 from each observation. The data in coded form are given below:

Month	Detergent						Row sum
	$A(x_1)$	x_1^2	$B(x_2)$	x_2^2	$C(x_3)$	x_3^2	
Cold water	7	49	5	25	17	289	29
Warm water	-1	1	2	4	18	324	19
Hot water	4	16	-4	16	8	64	8
Column sum	10	66	3	45	43	677	56

$$= \frac{(10)^2}{3} + \frac{(3)^2}{3} + \frac{(43)^2}{3} - 348.44$$

T = Sum of all observations in three samples of months = 56

CF = Correction factor = $\frac{T^2}{N} = \frac{(56)^2}{9} = 348.44$

SSC = Sum of squares between detergents (Columns)

$$= \frac{\left(\sum X_1\right)^2}{n_1} + \frac{\left(\sum X_2\right)^2}{n_2} + \frac{\left(\sum X_3\right)^2}{n_3} - \frac{T^2}{N}$$

= (33.33+3+616.33) – 348.44

= 304.22

Degrees of freedom, $df_c = c - 1 = 3 - 1 = 2,$

SSR = Sum of squares between water temperature (rows)

$$= \frac{(29)^2}{3} + \frac{(19)^2}{3} + \frac{(8)^2}{3} - 348.44$$

= (280.33+120.33+21.33) – 348.44

= 73.55

Degrees of freedom, $df_r = r - 1 = 3 - 1 = 2$

SST = Total sum of squares

$$= \left(\sum X_1^2 + \sum X_2^2 + \sum X_3^2\right) - \frac{T^2}{N}$$

= (66+45+677) – 348.44

= 439.56

Degrees of freedom,

$$\therefore\ SSE = SST - \left(SSC + SSR\right)$$

= 439.56 – (304.22+73.55)

= 61.79

Degrees of freedom,

Thus

$$MSC = \frac{SSC}{(c-1)} = \frac{304.22}{2} = 152.11$$

$$MSR = \frac{SSR}{(r-1)} = \frac{73.55}{2} = 36.775$$

$$MSE = \frac{SSE}{(c-1)(r-1)} = \frac{61.79}{4} = 15.447$$

The ANOVA Table is given below:

Sources of Variation	Sum of squares (SS)	Degrees of Freedom (d.f.)	Mean Squares(MS)	Test statistic
Between detergents(columns)	304.22	2	152.110	$F_{treatment} = 152.11/15.447 = 9.847$
Between temperature (rows)	73.55	2	36.775	$F_{block} = 36.775/15.447 = 2.380$
Residual error	61.79	4	15.447	
Total	439.56	8		

The critical value of F at 5% level for (2, 4) degrees of freedom is 6.94.

(a) Since the calculated value of $F_{treatment} = 9.847$ is greater than the critical value $F_{0.05}(2,4) = 6.94$, therefore we reject the null hypothesis and conclude that there is significant difference between the

performance of the three detergents.

(b) Since the calculated value of $F_{block} = 2.380$ is less than the critical value, hence we may accept the null hypothesis and conclude that the water temperature do not make a significant difference in the performance of the detergent.

Example 11.11: The following table gives the number of subscribers added by four major telecom players in India, in the months of August, September, October and November 2005. The data are given in 000's and are rounded off to the nearest 100, and are thus in lakhs.

Months	Company			
	Bharati	BSNL	Tata Indicom	Reliance
August	6	6	2	5
September	7	6	2	3
October	7	6	6	4
November	7	8	7	4

(Source: Indianfoline.com on India Mobile Industry)

Find out (use 5% level of significance)

(i) If the four companies significantly differ in their performance.

(ii) Is there significant difference between the months?

Solution: First of all, we set up the null hypothesis that (a) there is no significant difference between the companies in their performance and (b) there is no significant difference between the months.

Now we prepare the following table for calculation:

Months	Company					
	Bharati	BSNL	Tata Indicom	Reliance	Total	Average
August	6	6	2	5	19	4.75
September	7	6	2	3	18	4.50
October	7	6	6	4	23	5.75
November	7	8	7	4	26	6.50
Total	27	26	17	16	Grand total=86	
Average	6.75	6.5	4.25	4	Grand mean=5.375	

From the above table, we work out the requisite sums of squares for preparing the ANOVA Table as follows:

T = Sum of all observations in three samples of months = $\sum\sum x_{ij} = 86$

CF = Correction factor = $\frac{T^2}{N} = \frac{\left(\sum\sum x_{ij}\right)^2}{16} = \frac{(86)^2}{16} = 462.25$

SST = Total sum of squares

= $\sum\sum x_{ij}^{\ 2} - CF$

= (6^2 +6^2+……..+4^2) –462.25

= 514-462.25

= 51.75

Degrees of freedom, $df = N - 1 = 16 - 1 = 15$

SSR = Sum of squares between months (rows)

$$= \frac{(19)^2}{4} + \frac{(18)^2}{4} + \frac{(23)^2}{4} + \frac{(26)^2}{4} - 462.25$$

= 472.5-462.25

= 10.25

Degrees of freedom, $df_r = r - 1 = 4 - 1 = 3$

SSC = Sum of squares between Companies (Columns)

$$= \frac{(27)^2}{4} + \frac{(26)^2}{4} + \frac{(17)^2}{4} + \frac{(16)^2}{4} - 462.25$$

= 487.5-462.25

= 25.25

Degrees of freedom, $df_c = c - 1 = 4 - 1 = 3,$

SSE=Residual or Error Sum of Squares

i.e., $SSE = SST - (SSC + SSR)$

= 51.75 – (10.25+25.25)

= 16.25

Degrees of freedom,

The above obtained summarized results are presented in the following ANOVA able:

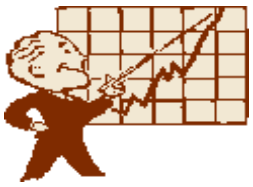
Source of Variation	Sum of Squares	Degrees of Freedom(d.f.)	Mean Squares(M.S.)	F (Calculated)	$\alpha = 5\%$ (Tabulated)
Between Months	10.25	3	3.42	3.42/1.81 = 1.89	3.86
Between Companies	25.25	3	8.42	8.42/1.81 =4.65	3.86
Error	16.25	9	1.81		
Total	51.75	15			

With calculated and tabulated values of F we can derive following conclusions:

(a) Since the calculated value of $F_{Months} = 1.89$ for months is less than the critical value $F_{0.05}(3,9) = 3.86$, therefore we may accept the null hypothesis and conclude that there is no significant difference between the months in terms of total additional consumers.

(b) Since the calculated value of $F_{Companies} = 4.65$ for companies is greater than the critical value $F_{0.05}(3,9) = 3.86$, therefore we reject the null hypothesis and conclude that there is a significant difference between the companies in terms of adding subscribers.

H_{02} :



CHECK YOUR PROGRESS

Q 5: The two null hypotheses to the following two-way ANOVA table are

H_{01} : The treatments means are equal i.e., $\mu_1 = \mu_2$

The block means are equal i.e., $\mu_1 = \mu_2 = \mu_3$

Source	SS	df	MS	F
Treatment	204.167			
Blocks	76.000	2	38.00	
Error		2		
Total	289.500	5		

(a) Complete the ANOVA table.

(b) Give your decision regarding the two sets of hypothese.



11.10 LET US SUM UP

- The major characteristics of the F distribution are:
  - (1) It is never negative.
  - (2) It is continuous.
  - (3) It is positively skewed.
  - (4) It is based on two sets of degrees of freedom.

- The distribution is used to test whether two population variances are the same. In this situation, the sampled populations must follow the normal distribution.

- The distribution can also be used as given below:
  - (ii) F test is used to test the significance of an observed sample multiple correlation.
  - (iii) It is used to test the significance of an observed sample correlation ratio.
  - (iv) It is applied for testing the linearity of regression.
  - (v) It is used to test the equality of several population means.

- This unit also focuses on a procedure called the analysis of variance, or ANOVA. It is a technique used to test simultaneously whether the means of several populations are equal. It uses the distribution as the distribution of the test statistic.



11.11 FURTHER READINGS

1) Srivastava, T.N., Rego, S. (2008). Statistics for Management. New Delhi. Tata McGraw Hill Education Private Limited.

2) Sharma, J.K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.

3) Hazarika, P.L. (2016). Essential Statistics For Economics And Business Studies. New Delhi. Akansha Publishing House.

4) Lind, D.A., Marshal, W.G., Wathen, S.A. (2009) Statistical Techniques in Business and Economics. New Delhi. Tata McGraw Hill Education Private

Limited.

5) Bajpai, N. (2014). Business Statistics. New Delhi. Pearson Education Ltd.

11.12 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q No 1: (i) True, (ii) False, (iii) True.

Ans. to Q No 2: The degrees of freedom of F -value are 5 in the numerator and 3 in the denominator.

The critical value of with (5, 3) degrees of freedom for two-tailed at 5% level of significance = 9.01

Ans. to Q No 3: (a) Null hypothesis is: There is no significant difference in the variation of the two populations i.e.,.

Alternative hypothesis is: There is significant difference in the variation of the two populations i.e.,

(b) Reject if 10.5, the critical value of with (7, 5) degrees of freedom for two-tailed at 5% level of significance.

(c) The test statistic $F = \frac{S_1^2}{S_2^2} = \frac{10^2}{7^2} = 2.04$

(d) H_0 may be accepted.

Ans. to Q No 4: (i) True, (ii) False, (iii) True, (iv) False, (v) True.

Ans. to Q No 5: (a) With the given information the ANOVA table is completed as follows:

Source	SS	df	MS	F
Treatment	204.167	1	$\frac{204.167}{1} = 204.167$	$\frac{204.167}{4.667} = 43.75$
Blocks	76	2	38	$\frac{38}{4.667} = 8.14$
Error	289.5-(204.167+76) =9.333	2	$\frac{9.333}{2} = 4.667$	
Total	289.5	5		

(b) Now, since $F_{Treatment} > F_{0.05}(1,2) = 18.5$, the critical value of F therefore reject H_{01}.

Again, since $F_{blocks} < F_{0.05}(2,2) = 19.0$, the critical value of thereforemay be accepted.



11.13 MODEL QUESTIONS

1. Define F-distribution.
2. Write a note on properties and applications of F-distribution.
3. What is Analysis of variance? What are the chief assumptions of ANOVA?
4. Tests for breaking strength were carried out on two lots of 5 and 9 steel wires respectively. The variance of one lot was 230 and that of the other was 492. Is there a significant difference in their variability?
5. Two random samples drawn from normal population are:

Sample I : 20 16 26 27 23 22 18 24 25 19

Sample II : 27 33 42 35 32 34 38 28 41 43 30 37

Obtain estimates of the variances of the population and test whether the two population have the same variance?

6. An agriculture research organization wants to study the effect of four types of fertilizers on the yield of a crop. It divided the entire field into 24 plots of land and used fertilizer at random in 6 plots of land. Part of the calculations are given below:

Source	SS	D.F.	MS	F
Fertilizers	2940	3	.....	5.99
Within groups	....	.....	.....	
Total	6212	24		

(a) Fill in the blanks in the ANOVA table.

(b) Test at 5% level of significance, whether the fertilizers differ significantly.

7. There are three main brands of a certain powder. A sample of 120 packets sold is examined and found to be located among four groups A, B, C and D, and brands I, II and III as shown below:

Brand	Group			
	A		B	
I	0	4	8	15
II	5	8	13	6
III	18	19	11	13

Is there any significant difference in brand preferences?

8. A tea company appoints four salesmen A, B, C and D, and observes their sales in three seasons—summer, winter and monsoon. The figures (in lakhs) are given below:

Season	Salesmen				
	A	B	C	D	Total
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Total	90	93	81	96	360

(a) Do the salesmen significantly differ in performance?

(b) Is there significant difference between the seasons?

*** ***** ***