4

F-Test and Analysis of Variance

F-Distribution

The Fisher's F-distribution may be defined as :—

$$F = \frac{\chi_1^2/v_1}{\chi_2^2/v_2} \text{ and } v_1 > v_2$$

i.e. F-distribution is a distribution of the ratio of two independent Chi-square variates (χ_1^2 and χ_2^2) each divided by the corresponding degrees of freedom (v_1 and v_2).

Characteristics of F-Distribution—

1. The value of F cannot be negative since both terms of F ratio are squared values.
2. The range of the values of F is from 0 to infinity (∞).
3. The shape of F-distribution curve depends upon the number of degrees of freedom for the first term and the second term of the F ratio.

F-Test

The F-test based on F-distribution is called so in honour of a great statistician ***R.A. Fisher***. It is also known as **Fisher's F-test** or **Variance Ratio test**. This technique of evaluation of variance ratios was developed by the famous statistician ***Ronald Aylmer Fisher*** in 1920 in collaboration of his colleague ***George W. Snedecor***, who prepared a table of the estimated, or critical values of F for different degrees of freedom of the different variances.

The F-test refers to a test of hypothesis concerning two variances derived from two samples. This test is used specially when it is to be decided whether two samples may be regarded as drawn from the normal population having the same variance or not. Since F-test is based on the ratio of two varainces, it is also known as ***variance ratio test***.

The various tests of significance like Z-test, *t*-test etc. can be applied in case of two samples only, hence, are not suitable for test of significance of more than two number of samples. Hence, the need for a separate test arises. F-test can be used advantageouly in the analysis of variances involving two and more number of samples.

Moreover, whereas the mean of samples drawn randomly from a normal population are distributed normally, the variances of random samples drawn from such a population are not normally distributed, they are skewed positively, hence the need for a separate test also arose.

Assumptions of F-test—

The theoretical assumptions on which F-test is based are :—

(1) **Normality**—The populations for each sample must be normally distributed with identical mean and variance.

(2) **Random method and independence**—All sample observations must be randomly selected and independent.

(3) **Variance ratio must be 1 or greater**—The ratio of σ_1^2 to σ_2^2 should be equal to or greater than 1. In this context, larger estimate of variance is divided by smaller estimate of variance.

(4) **Always positive**—Since, the F-distribution is always formed by a ratio of squared values, it can never be a negative number.

(5) **Positively skewed**—All F-distributions are uni-modal and are skewed to the right. They tend to become more nearly symmetrical as v_1 and v_2 increase.

(6) **Additive property**—Total of different components of variance is equal to total variance *i.e.*

Total variance = Variance between samples + Variance within samples

Calculation of F-Test (two samples)—

The computation of F-Test involves the following steps :—

Step 1. Null hypothesis—$H_o : \sigma_1^2 = \sigma_2^2$

The variance of populations corresponding to both samples are equal.

or

The two samples are drawn from the same population *i.e.* there is no significant difference between the variances of two samples.

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

The variance of populations corresponding to both samples are not equal.

or

The two samples are drawn from different populations *i.e.* there is a significant difference between the variances of two samples.

Step 2. Computation of test statistic—Under H_o, the test statistic is—

$$F = \frac{\text{Larger Variance Estimate}}{\text{Smaller Variance Estimate}}$$

i.e. $$\boxed{F = \frac{\hat{s}_1^2}{\hat{s}_2^2}}; \qquad \hat{s}_1^2 > \hat{s}_2^2$$

where, $\hat{s}_1^2 = n_1 s_1^2/(n_1 - 1)$

$\hat{s}_2^2 = n_2 s_2^2/(n_2 - 1)$

s_1^2 and s_2^2 = Sample variances

n_1 and n_2 = Sample sizes

Note : In numerical problems, we generally take greater of the variances $\hat{s}_1^2$ or $\hat{s}_2^2$ in the numerator and adjust it for the degree of freedom accordingly. It means that if :—

$\hat{s}_1^2 < \hat{s}_2^2$ then the formula will be—

$$F = \frac{\hat{s}_2^2}{\hat{s}_1^2}$$

***Step 3.* Degrees of freedom**—Degrees of freedom follows Snedecor's F-distribution with—

$v_1 = n_1 - 1, v_2 = n_2 - 1$ when $\hat{s}_1^2 > \hat{s}_2^2$

$v_1 = n_2 - 1, v_2 = n_1 - 1$ when $\hat{s}_1^2 < \hat{s}_2^2$

or

Step 4. Level of significance—Generally 5% (or 1%)

Step 5. Critical value—The critical value of F is obtained from F-table at specified level of significance for v_1 and v_2 degrees of freedom in the manner as shown in the table below:—

v_2 \ v_1	*d.f.* for numerator →				
d.f. for denominator ↓	.....	.....	.....	.....	.....
	.....	.....	.....	.....	.....
	.....	.....	.....	.....	.....
	.....	.....	.....	.....	.....
	.....	.....	.....	.....	.....
	.....	.....	.....	.....	.....

Step 6. Decision—If computed value of F (step 2) is less than the critical value of F (step 5), null hypothesis will be accepted and it will be concluded that ***both the samples have been drawn from same population*** (*i.e.* variance of first sample is not significantly different from variance of second sample).

If computed value of F (step 2) is greater than the critical value of F (step 5), null hypothesis will be rejected and it will be concluded that ***both the samples have not been drawn from same population*** (*i.e.* variance of first sample is significantly different from variance of second sample).

Illustration 1.

Following information about two samples selected from two normal population is available :—

$n_1 = 9$ and $s_1 = 2{\cdot}9$

$n_2 = 7$ and $s_2 = 6{\cdot}3$

Test whether the two samples have come from the population having the same variance. [*C.A. Foundation, May, 2005*]

[Given : $F_{6,8\,(0{\cdot}05)} = 3{\cdot}58$, $F_{8,6\,(0{\cdot}05)} = 4{\cdot}15$]

Solution :

Unbiased estimate for the first population variance—

$$= \hat{s}_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$$

$$= \frac{9 \times (2{\cdot}9)^2}{9 - 1} = \frac{75{\cdot}69}{8} = 9{\cdot}46$$

Unbiased estimate for the second population variance—

$$= \hat{s}_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{7 \times (6{\cdot}3)^2}{7 - 1} = \frac{277{\cdot}83}{6} = 46{\cdot}31$$

Step 1. Null hypothesis—$H_o : \sigma_1^2 = \sigma_2^2$ *i.e.* the two samples have come from the population having the same variance.

Alternative hypothesis—$H_1 : \sigma_1^2 \neq \sigma_2^2$ *i.e.* the two samples have come from the population having different variance.

Step 2. Test statistic—Under, H_o, the test statistic is—

$$F = \frac{\hat{s}_2^2}{\hat{s}_1^2} \qquad (\text{since } \hat{s}_2^2 > \hat{s}_1^2)$$

$$= \frac{46{\cdot}31}{9{\cdot}46} = \mathbf{4{\cdot}89}$$

Step 3. Degrees of freedom—Since $\hat{s}_2^2 > \hat{s}_1^2$, then—

$$v_1 = n_2 - 1 = 7 - 1 = 6$$

$$v_2 = n_1 - 1 = 9 - 1 = 8$$

Step 4. Level of significance—$\alpha = 0{\cdot}05$ (5%)

Step 5. Critical value—At 5% level of significance and (6, 8) degrees of freedom, the critical value of F is—

$$F_{6,8\,(0{\cdot}05)} = 3{\cdot}58 \text{ (given)}$$

Step 6. Decision—Since, the computed value of F = 4·89 is greater than the critical value of F = 3·58, it falls in the rejection region. Hence, null hypothesis is rejected and it may be concluded that *the two samples have not come from the population having the same variance.*

Illustration 2.

It is known that the mean diameters of bolts produced by two firms A and B are practically the same but the standard deviations differ. For 16 bolts produced by firm A, the standard deviation is 3·8 mm, while for 22 bolts manufactured by firm B, the standard deviation is 2·9 mm. Compute the statistic you would use to test whether the product of firm A has the same variability as that of firm B.

Solution :

Here,

For Firm A : $n_1 = 16$ $s_1 = 3{\cdot}8$

For Firm B : $n_2 = 22$ $s_2 = 2{\cdot}9$

Now, estimated population variance in Firm A—

$$= \hat{s}_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{16 \times (3{\cdot}8)^2}{16 - 1} = 15{\cdot}40$$

Estimated population variance in Firm B—

$$= \hat{s}_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{22 \times (2{\cdot}9)^2}{22 - 1} = 8{\cdot}81$$

Step 1. Null hypothesis—$H_o : \sigma_1^2 = \sigma_2^2$ *i.e.* the product of firm A has the same variability as that of firm B.

Alternative hypothesis—$H_1 : \sigma_1^2 \neq \sigma_2^2$, *i.e.* the product of firm A does not have the same variability as that of firm B.

Step 2. Test statistic—Under H_o, the test statistic is—

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2} \qquad (\text{since } \hat{s}_1^2 > \hat{s}_2^2)$$

$$= \frac{15{\cdot}40}{8{\cdot}81} = \mathbf{1{\cdot}75}$$

Step 3. Degrees of freedom—Since $\hat{s}_1^2 > \hat{s}_2^2$, then—

$$v_1 = n_1 - 1 = 16 - 1 = 15$$

$$v_2 = n_2 - 1 = 22 - 1 = 21$$

Step 4. Level of significance—$\alpha = 0{\cdot}05$ (5%)

Step 5. Critical value—At 5% level of significance and (15, 21) degrees of freedom, the critical value of F is—

$$F_{15,\,21\,(0{\cdot}05)} = 2{\cdot}18 \text{ (from table)}$$

Step 6. Decision—Since, the computed value of F = 1·75 is less than the critical value of F = 2·18, it falls in the acceptance region. Hence, null hypothesis is accepted and it may be concluded ***that the product of firm A has the same variability as that of firm B.***

Illustration 3.

In a laboratory experiment, two random samples gave the following results :—

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

Test the equality of sample variances at 5% level of significance.

Solution :

Here,

For Sample 1 : $n_1 = 10, \overline{X}_1 = 12, \Sigma(X_1 - \overline{X}_1)^2 = 120$

For Sample 2 : $n_2 = 12, \overline{X}_2 = 15, \Sigma(X_2 - \overline{X}_2)^2 = 314$

Now, ***estimated population variance for the first population***—

$$= \hat{s}_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{\Sigma(X_1 - \overline{X}_1)^2}{n_1 - 1}$$

$$= \frac{120}{9} = 13{\cdot}33$$

Estimated population variance for the second population—

$$= \hat{s}_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{\Sigma(X_2 - \overline{X}_2)^2}{n_2 - 1}$$

$$= \frac{314}{11} = 28{\cdot}55$$

Step 1. Null hypothesis—$H_o : \sigma_1^2 = \sigma_2^2$, *i.e.* the two population variances do not differ significantly.

Alternative hypothesis—$H_1 : \sigma_1^2 \neq \sigma_2^2$ *i.e.* the two population variances differ significantly.

Step 2. Test statistic—Under H_o, the test statistic is— (since $\hat{s}_2^2 > \hat{s}_1^2$)

$$F = \frac{\hat{s}_2^2}{\hat{s}_1^2}$$

$$\frac{28{\cdot}55}{} = 2{\cdot}14$$

Estimated population variance in worker A

$$= \hat{s}_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{\sum (X_1 - \bar{X}_1)^2}{n_1 - 1}$$

$$= \frac{80}{5} = 16$$

Estimate population variance in worker B

$$= \hat{s}_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{\sum (X_2 - \bar{X}_2)^2}{n_2 - 1}$$

$$= \frac{84}{7} = 12$$

Step 2. Test statistic—Under H_o, the test statistic is—

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2} \qquad (\text{since } \hat{s}_1^2 > \hat{s}_2^2)$$

$$= \frac{16}{12} = \mathbf{1{\cdot}33}$$

Step 3. Degrees of freedom—Since $\hat{s}_1^2 > \hat{s}_2^2$, then—

$$\nu_1 = n_1 - 1 = 6 - 1 = 5$$

$$\nu_2 = n_2 - 1 = 8 - 1 = 7$$

Step 4. Level of significance—$\alpha = 0{\cdot}05$ (5%)

Step 5. Critical value—At 5% level of significance and (5, 7) degrees of freedom, the critical value of F is—

$$F_{5,\,7\,(0{\cdot}05)} = 3{\cdot}97 \text{ (from table)}$$

Step 6. Decision—Since, the computed value of F = 1·33 is less than the critical value of F = 3·97, it falls in the acceptance region. Hence, null hypothesis is accepted and it may be concluded that ***worker B is not a more stable worker.***

ANALYSIS OF VARIANCE

To test the significance of mean of one sample or significance of difference between means of two samples *t*-test or Chi-square test are very useful. However, if there are more than two samples, then these tests are not appropriate. To solve such problem, the method of analysis of variance is used, which was developed by ***R.A. Fisher*** in the year 1923. A test so developed by him is known as the Fisher's test or more commonly as F-test. Now a days, F-test is widely used in the analysis of variance. It is mainly used to test the hypothesis of equality between two variances. But the main objective of analysis of variance is to test the significance of means of more than two samples only by one test.

The technique of analysis of variance is also referred to as **'ANOVA'**. This test is particularly suitable for experimental work, as no assumption of equality of variance is required. The analysis of variance is mainly carried on under :—

(i) One-way classification (or one-fold or one factor classification).

(ii) Two-way classification (or two-fold or two factor classification).

Definitions—

According to ***R.A. Fisher***, "It is a process of separation of the variance ascribable to one group of causes from the variance as ascribable to other groups."

According to ***Davis***, "The analysis of variance is essentially a method of analysing

...groups of data for homogeneity." ...ually a procedure

The actual analysis of variance is carried out on the basis of variance ratio, which is obtained by dividing the variance between the samples by the variance within the samples. This ratio forms the F-statistic. Hence, the test statistic—

$$F = \frac{\text{Variance between the samples}}{\text{Variance within the samples}}$$

Generally, the variance between the samples happens to be greater than the variance within the samples. So it is placed as numerator. However, if on account of any reason, it happens to be smaller, the numerator and denominator should be interchanged so that the value of F is always greater than or equal to 1.

In general,

$$F = \frac{\text{Greater Variance}}{\text{Smaller Variance}}$$

Uses or Applications or Importance of Analysis of Variance—

Analysis of variance is useful in nearly every type of experimental work. The main applications are:—

1. To test the significance of differences between means of more than two samples.
2. To test the significance of differences between variances of different samples and for this purpose variance ratio or F-coefficient is calculated.
3. Useful in two-way classification.
4. To test the significance of regression, coefficient of correlation and significance of multiple correlation.

ANALYSIS OF VARIANCE — ONE WAY CLASSIFICATION

Under one-way classification, the influence of only one attribute or factor is considered. For example—if we collect data of agricultural production on the basis of use of different type of fertilizers, it will be one-way classification because all other factors have been ignored except the use of fertilizer. Normally, the columns represent the various types of fertilizers, the production of sample plots under each type of fertilizer will be in the respective columns. The following three methods can be used for analysis of variance in one-way classification.

1. Direct Method;
2. Short-cut method; and
3. Coding Method.

(1) Direct Method—

The following steps are required for analysis of variance under direct method :—

Step 1. Null hypothesis—$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_k$ *i.e.* the arithmetic means of the populations from which k samples have been randomly drawn are equal to one another. In other words, there is no significant difference between these means.

Alternative hypothesis—$H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \ldots\ldots \neq \mu_k$ *i.e.* the arithmetic means of the populations from which k samples have been randomly drawn are not equal to one another. In other words. there is a significant difference between these means.

Step 2. **Sum of the squares of variations amongst (or between) the samples (columns) (SSC)**—It is calculated as follows :—

(i) Calculate the mean of each sample (column) *i.e.* obtain $\overline{X}_1, \overline{X}_2 \ldots\ldots \overline{X}_k$

(ii) Calculate grand mean (or combined mean) on the basis of sample means, which would be :—

$$\overline{\overline{X}} = \frac{\overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \ldots\ldots + \overline{X}_k}{k}$$

where $\overline{\overline{X}}$ = Grand mean

k = No. of samples

(iii) Calculate the deviation of sample means from the grand mean *i.e.* $(\overline{X}_1 - \overline{\overline{X}})$, $(\overline{X}_2 - \overline{\overline{X}}) \ldots\ldots, (\overline{X}_k - \overline{\overline{X}})$

(iv) Square the deviations obtained in (iii) and multiply by the number of items in the relevant sample *i.e.*

$$n_1(\overline{X}_1 - \overline{\overline{X}})^2, n_2(\overline{X}_2 - \overline{\overline{X}})^2 \ldots\ldots, n_k(\overline{X}_k - \overline{\overline{X}})^2$$

(v) Calculate the sum of these figures. This will be SSC *i.e.*

$$SSC = n_1(\overline{X}_1 - \overline{\overline{X}})^2 + n_2(\overline{X}_2 - \overline{\overline{X}})^2 + \ldots + n_k(\overline{X}_k - \overline{\overline{X}})^2$$

Step 3. **Mean Sum of Squares of Columns** (MSC)—Divide the sum of squares of deviation (SSC) by the degrees of freedom which is $(k-1)$ where k denotes the number of samples *i.e.*

$$MSC = \frac{SSC}{k-1}$$

Step 4. **Sum of Squares of Variations within Samples** (SSE)—The computation of variance within samples involves the following steps :—

(i) Calculate the mean of each sample *i.e.*

$$\overline{X}_1, \overline{X}_2, \ldots\ldots \overline{X}_k$$

(ii) Calculate the deviation of various items in a sample from the mean value of the relevant sample *i.e.*

$$(X_1 - \overline{X}_1), (X_2 - \overline{X}_2) \ldots\ldots (X_k - \overline{X}_k)$$

(iii) Square these deviations and obtain their total *i.e.*

$$\Sigma(X_1 - \overline{X}_1)^2, \Sigma(X_2 - \overline{X}_2)^2 \ldots\ldots \Sigma(X_k - \overline{X}_k)^2$$

(iv) Total the sum of the squares of the deviations of the various samples from their respective means. This would be the value of SSE *i.e.*

$$SSE = \Sigma(X_1 - \overline{X}_1)^2 + \Sigma(X_2 - \overline{X}_2)^2 + \ldots\ldots + \Sigma(X_k - \overline{X}_k)^2$$

Step 5. **Mean Sum of Squares of the Error** (MSE)—Divide SSE by the number of degrees of freedom which would be—

$$(n_1 - 1) + (n_2 - 1) + \ldots\ldots + (n_k - 1)$$

or $N - k$

where N = Total number of items in all the samples

k = Number of samples

Hence, $\text{MSE} = \frac{\text{SSE}}{N - k}$ or $\frac{\text{SSE}}{N - C}$

as number of columns, (C) would be equal to the number of samples.

Step 6. Total Sum of Squares of Variations (SST)—It is computed by adding the sum of squares of deviations between the samples (SSC) and the sum of squares of deviations within the samples (SSE) *i.e.*

$$\text{SST} = \text{SSC} + \text{SSE}$$

Step 7. Variance Ratio—Variance ratio or F-ratio is the ratio between greater variance and smaller variance. Generally the variance between the samples happens to be greater than variance within the samples. Hence,

$$\text{F} = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{\text{MSC}}{\text{MSE}}$$

However if, on account of any reason, it happens that variance within samples is greater than the variance between samples, then the numerator and denominator should be interchanged and degrees of freedom adjusted accordingly.

ANOVA TABLE (ONE-WAY CLASSIFICATION)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-Ratio
Between samples	SSC	$\nu_1 = k - 1$	$\text{MSC} = \frac{\text{SSC}}{\nu_1}$	$\text{F} = \frac{\text{MSC}}{\text{MSE}}$
Within samples	SSE	$\nu_2 = N - k$	$\text{MSE} = \frac{\text{SSE}}{\nu_2}$	
Total	SST	$N - 1$		

Step 8. Level of significance—Generally 5% (or 1%)

Step 9. Critical value—The critical value of F is obtained from F-table at a certain level of significance and on the basis of degrees of freedom (ν_1, ν_2), where

ν_1 = Degree of freedom of greater variance

ν_2 = Degree of freedom of smaller variance

Step 10. Decision—If the computed value of F is less than the critical value of F then null hypothesis is accepted and it will be ***concluded that the differences between the means is not significant and has arisen due to sampling fluctuations.***

If the computed value of F is greater than the critical value of F then null hypothesis is rejected and it will be ***concluded that the difference between the means is significant*** *i.e.* it could not have arisen due to sampling fluctuations but it is due to samples have not been drawn from the same population.

Illustration 5.

The following table gives the yields on 12 sample plots under three varieties of seed A, B and C :—

A	B	C
10	9	4
6	7	8
7	7	6
9	5	6

Set up a table of analysis of varaince and find out whether there is a significant difference between the mean yields of three varieties.

Solution :

Step 1. Null hypothesis—$H_0 : \mu_1 = \mu_2 = \mu_3$, *i.e.* there is no significant difference between the yields of three varieties.

Alternative hypothesis—$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$, *i.e.* there is a significant difference between the mean yields of three varieties.

Step 2. Sum of the squares of variations between the varieties (SSC)—

(i) $$\overline{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{10+6+7+9}{4} = 8$$

$$\overline{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{9+7+7+5}{4} = 7$$

$$\overline{X}_3 = \frac{\Sigma X_3}{n_3} = \frac{4+8+6+6}{4} = 6$$

(ii) $$\overline{\overline{X}} = \frac{\overline{X}_1 + \overline{X}_2 + \overline{X}_3}{3} = \frac{8+7+6}{3} = 7$$

(iii) $$SSC = n_1(\overline{X}_1 - \overline{\overline{X}})^2 + n_2(\overline{X}_2 - \overline{\overline{X}})^2 + n_3(\overline{X}_3 - \overline{\overline{X}})^2$$
$$= 4(8-7)^2 + 4(7-7)^2 + 4(6-7)^2$$
$$= 4 + 0 + 4 = 8$$

Step 3. Mean sum of squares of columns (MSC)—

$$MSC = \frac{SSC}{k-1} = \frac{8}{3-1} = 4$$

Step 4. Sum of Squares of Variations within Samples (SSE)—

A			B			C		
X_1	$(X_1 - \overline{X}_1)$	$(X_1 - \overline{X}_1)^2$	X_2	$(X_2 - \overline{X}_2)$	$(X_2 - \overline{X}_2)^2$	X_3	$(X_3 - \overline{X}_3)$	$(X_3 - \overline{X}_3)^2$
10	2	4	9	2	4	4	−2	4
6	−2	4	7	0	0	8	2	4
7	−1	1	7	0	0	6	0	0
9	1	1	5	−2	4	6	0	0
		10			8			8

$$SSE = \Sigma(X_1 - \overline{X}_1)^2 + \Sigma(X_2 - \overline{X}_2)^2 + \Sigma(X_3 - \overline{X}_3)^2$$
$$= 10 + 8 + 8 = 26$$

Step 5. Mean sum of squares of the error (MSE)—

$$MSE = \frac{SSE}{N-k} = \frac{26}{12-3} = \frac{26}{9} = 2{\cdot}89$$

Step 6. Total sum of squares of Variations (SST)—

$$SST = SSC + SSE$$
$$= 8 + 26 = 34$$

Step 7. Variance ratio—

$$F = \frac{MSC}{MSE} = \frac{4}{2{\cdot}89} = \mathbf{1{\cdot}38}$$

with the help of above data, the table of analysis of variance (ANOVA) is set-up as follows :—

ANOVA TABLE

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	Variance Ratio (F)
Between samples	SSC = 8	$\nu_1 = k - 1 = 3 - 1 = 2$	MSC = 4	$F = \frac{4}{2 \cdot 89} = 1 \cdot 38$
Within samples	SSE = 26	$\nu_2 = N - k = 12 - 3 = 9$	MSE = 2·89	
Total	SST = 34	$N - 1 = 11$		

Step 8. Level of significance—$\alpha = 0 \cdot 05$ (5%)

Step 9. Critical value—At 5% level of significance and (2, 9) degrees of freedom, the critical value of F is :—

$$F_{2, 9\,(0 \cdot 05)} = 4 \cdot 26$$

Step 10. Decision—Since the computed value of F = 1·38 is less than the critical value of F = 4·26, it falls in the acceptance region. Hence, null hypothesis is accepted and it may be concluded that ***there is no significant difference between the mean yields of three varieties.***

(II) Short-cut Method—

Direct method of calculating the F-ratio is very tedious and time consuming. Moreover, if the means of samples are in fractions and not in whole number, calculation process becomes more complicated. Hence, it is advisable to use an easier method known as short-cut method. The following steps are required for calculation of variance ratio by short-cut method:—

1. Total of sample items—Obtain the sum of the values of all the items of all the samples and denote it by T *i.e.*

$$T = \Sigma X_1 + \Sigma X_2 + \ldots\ldots + \Sigma X_k$$

2. Correction factor—Calculate the correction factor which is obtained as—

$$\text{Correction factor} = \frac{T^2}{N}$$

where N = Total number of items in all the samples.

3. Total sum of squares of items—Calculate the square of all the items of all the samples and add them together *i.e.* $\Sigma X^2 + \Sigma X_2^2 + \ldots\ldots + \Sigma X_k^2$

4. Total sum of squares (SST)—It is obtained by subtracting the correction factor from the total sum of squares of all the items of the samples *i.e.*

$$SST = [\Sigma X_1^2 + \Sigma X_2^2 + \ldots + \Sigma X_k^2] - \frac{T^2}{N}$$

5. Sum of squares between samples (SSC)—It is obtained by the following steps:—

(i) Square the totals of samples *i.e.* $(\Sigma X_1)^2, (\Sigma X_2)^2, \ldots\ldots (\Sigma X_k)^2$

(ii) Divide each square of totals by respective number of items in each sample and add these figures *i.e.* $\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \ldots\ldots + \frac{(\Sigma X_k)^2}{n_k}$

(iii) Subtract the correction factor from (ii) to obtain SSC:—

$$SSC = \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \ldots\ldots + \frac{(\Sigma X_k)^2}{n_k}\right] - \frac{T^2}{N}$$

Step 3. Degrees of freedom—$\nu_1 = 1$, $\nu_2 = 88$

Step 4. Level of significance—$\alpha = 0{\cdot}05$ (5%)

Step 5. Critical value—At 5% level of significance and (1, 88) degrees of freedom, the critical value of F is—

$$F_{1,\,88\,(0{\cdot}05)} = 3{\cdot}96 \text{ (approx.) from table}$$

Step 6. Decision—Since, computed value of F = 23·88 is greater than the critical value of F = 3·96, it falls in the rejection region. Hence, null hypothesis is rejected and it may be concluded ***that there is a significant difference in the mean level of knowledge of students of these two groups, and it has not arisen just due to fluctuations in sampling***.

ANALYSIS OF VARIANCE—TWO WAY CLASSIFICATION

Under two-way classification, the influence of two attributes or factors are considered. *For example,*

(i) Influence of different salesmen and various seasons on sales.

(ii) Influence of different types of fertilizers and different soil textures on yield.

(iii) Influence of advertisement and price level on sales etc.

But there may be sampling variations besides the two factors considered which we characterise as residual variations.

Such analysis provides test of two sets of hypothesis by the same set of data. In two-way classification, one factor (or attribute) is represented in columns (like different salesmen) and other factor is represented in rows (like different seasons). The following steps are required for analysis of variance in two-way classification—

Step 1. Set-up of hypothesis—Null hypothesis (H_o) and alternative hypothesis (H_1) relating to both factors are formulated.

Step 2. Coding—If the given values are large, coding method may be followed in order to reduce the complications of calculation work.

Step 3. Total of sample items—Obtain the sum of the values of all the items of all the samples and denote it by T *i.e.*

$$T = \text{Grand Total}$$

Step 4. Correction factor—Calculate the correction factor which is obtained as—

$$\text{Correction factor} = \frac{T^2}{N}$$

where N = Total number of items in all samples

Step 5. Total sum of squares of items—Calculate the square of all the items of all the samples in a table and add them together to obtain their grand total.

Step 6. Total Sum of Squares (SST)—It is obtained by subtracting the correction factor from the total sum of squares of all the items of all samples (obtained in step 5).

Step 7. Sum of Squares between Columns (SSC)—It is obtained by the following steps :—

(i) Square the totals of each columns *i.e.* $(\Sigma X_{c_1})^2, (\Sigma X_{c_2})^2 \ldots\ldots$

(ii) Divide each square of totals by respective number of items in each column and add these figures *i.e.*

$$\frac{(\Sigma X_{c_1})^2}{n_{c_1}} + \frac{(\Sigma X c_2)^2}{n_{c2}} + \ldots\ldots$$

(iii) Subtract the correction factor from (ii) to obtain SSC—

$$SSC = \left[\frac{(\Sigma X_{c_1})^2}{n_{c_1}} + \frac{(\Sigma X c_2)^2}{n_{c_2}} + \ldots\ldots\right] - \frac{T^2}{N}$$

Step 8. Sum of Squares between rows (SSR)—It is obtained by the following steps—

(i) Square the totals of each rows *i.e.* $(\Sigma X_{r_1})^2, (\Sigma X_{r_2})^2 \ldots\ldots$

(ii) Divide each square of totals by respective number of items in each row and add these figures *i.e.*

$$\frac{(\Sigma X_{r_1})^2}{n_{r_1}} + \frac{(\Sigma X r_2)^2}{n_{r_2}} + \ldots\ldots$$

(iii) Subtract the correction factor from (ii) to obtain SSR

$$SSR = \left[\frac{(\Sigma X_{r_1})^2}{n_{r_1}} + \frac{(\Sigma X r_2)^2}{n_{r_2}} + \ldots\ldots\right] - \frac{T^2}{N}$$

Step 9. Sum of Squares due to error (or Residual sum of squares) (SSE)—It is obtained as—

$$SSE = SST - (SSC + SSR)$$

Remark—

Total sum of squares (SST) = Sum of Squares between columns (SSC) + Sum of squares between rows (SSR) + Residual sum of squares (SSE)

i.e. $SST = SSC + SSR + SSE$

Step 10. Degrees of freedom—

Let r = Number of rows

c = Number of columns, then

Number of degrees of freedom between columns = $c - 1$

Number of degrees of freedom between rows = $r - 1$

Number of degrees of freedom for residual = $(r - 1)(c - 1)$

Total number of degrees of freedom = $N - 1$

Step 11. Mean Sum of Squares of Columns (MSC)—It is obtained by dividing the sum of squares between columns (SSC) by the degrees of freedom between column *i.e.*

$$MSC = \frac{SSC}{c - 1}$$

Step 12. Mean Sum of Squares of Rows (MSR)—It is obtained by dividing the sum of squares between rows (SSR) by the degrees of freedom between rows *i.e.*

$$MSR = \frac{SSR}{r - 1}$$

Step 13. Mean Sum of Squares of Residual (MSE)—It is obtained by dividing the sum of squares due to error (or Residual sum of squares) by the degrees of freedom for residual *i.e.*

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

Step 14. Variance ratio—

(i) **Between Columns—**

$$F_C = \frac{MSC}{MSE} \quad \text{where MSC} > \text{MSE}$$

(ii) **Between Rows—**

$$F_R = \frac{MSR}{MSE} \quad \text{where MSR} > \text{MSE}$$

However, if on account of any reason, it happens that Residual sum of squares (MSE) is greater than MSC or MSR, then the numerator and denominator should be interchanged and degrees of freedom adjusted accordingly.

Step 15. Level of significance—Generally 5% (or 1%)

Step 16. Critical value—The critical value of F is obtained from F-table at a certain level of significance and on the basis of degrees of freedom (v_1, v_2) where—

v_1 = Degree of freedom of greater variance

v_2 = Degree of freedom of smaller variance

The ANOVA table now takes the following form :—

ANOVA TABLE FOR TWO-WAY CLASSIFICATION

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	Variance Ratio (F)
Between columns	SSC	$c-1$	$MSC = \frac{SSC}{c-1}$	$F_C = \frac{MSC}{MSE}$
Between rows	SSR	$r-1$	$MSR = \frac{SSR}{r-1}$	
Residual	SSE	$(r-1)(c-1)$	$MSE = \frac{SSE}{(r-1)(c-1)}$	$F_R = \frac{MSR}{MSE}$
	SST	$N-1$		

Step 17. Decision—If the computed value of F is less than the critical value of F then null hypothesis is accepted otherwise rejected.

Illustration 10.

The following data represent the number of units of commodity produced by 3 different workers using 3 different types of machines :—

Workers	Machines		
	A	B	C
X	16	64	40
Y	56	72	56
Z	12	56	28

Test—(i) Whether the mean productivity is the same for the different machine types, and

(ii) Whether the three workers differ with respect to mean productivity.

Solution :

Here, $N = 9$, $r = 3$, $c = 3$

Step 1. Null hypothesis (H_o)—

(i) Mean productivity is the same for the different machine types.

(ii) There is no difference in the mean productivity of three workers.

Alternative hypothesis (H_1)—

(i) Mean productivity is not same for the different machine types.

(ii) There is a significant difference in the mean productivity of three workers.

Step 2. Coding of data may be done by subtracting 36 from each item—

CODED DATA

Workers	Machines			Row Total
	A	B	C	
X	− 20	+ 28	+ 4	+ 12
Y	+ 20	+ 36	+ 20	+ 76
Z	− 24	+ 20	− 8	− 12
Column Total	− 24	+ 84	+ 16	T = + 76

Step 3.

SUM OF SQUARES

Workers	Machines			Row Total
	A	B	C	
X	400	784	16	1200
Y	400	1296	400	2096
Z	576	400	64	1040
Column Total	1376	2480	480	4336

Step 4. Correction factor— $\frac{T^2}{N} = \frac{(76)^2}{9} = 641{\cdot}78$

Step 5. Total Sum of Squares (SST)—

$$SSR = 4336 - \frac{T^2}{N}$$

$$= 4336 - 641{\cdot}78 = 3694{\cdot}22$$

Step 6. Sum of Squares between Columns (machines) (SSC)—

$$SSC = \left[\frac{(-24)^2}{3} + \frac{(84)^2}{3} + \frac{(16)^2}{3}\right] - \frac{T^2}{N}$$

$$= [192 + 2352 + 85{\cdot}33] - 641{\cdot}78$$

$$= 1987{\cdot}55$$

Step 7. Sum of Squares between rows (workers) (SSR)—

$$SSR = \left[\frac{(12)^2}{3} + \frac{(76)^2}{3} + \frac{(-12)^2}{3}\right] - \frac{T^2}{N}$$

$$= [48 + 1925{\cdot}33 + 48] - 641{\cdot}78$$

$$= 1379{\cdot}55$$

Step 8. Mean Sum of Squares of Columns (machines) (MSC)—

$$MSC = \frac{SSC}{c-1} = \frac{1987{\cdot}55}{3-1} = 993{\cdot}78$$

Step 9. Mean Sum of Squars of Rows (workers) (MSR)—

$$MSR = \frac{SSR}{r-1}$$

$$= \frac{1379{\cdot}55}{3-1} = 689{\cdot}78$$

Step 10. Residual Sum of Squares (SSE)—

$$SSE = SST - (SSC + SSR)$$

$$= 3694{\cdot}22 - (1987{\cdot}55 + 1379{\cdot}55)$$

$$= 327{\cdot}12$$

Step 11. Mean Sum of Squares of Residual (MSE)—

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

$$= \frac{327{\cdot}12}{(3-1)(3-1)} = \frac{327{\cdot}12}{4} = 81{\cdot}78$$

Step 12. Variance Ratio—

$$F_C = \frac{MSC}{MSE} \qquad \text{(Since MSC > MSE)}$$

$$= \frac{993{\cdot}78}{81{\cdot}78} = 12{\cdot}15$$

$$F_R = \frac{MSR}{MSE} \qquad \text{(Since MSR > MSE)}$$

$$= \frac{689{\cdot}78}{81{\cdot}78} = 8{\cdot}43$$

The above data can be represented in two-way ANOVA table as—

ANOVA TABLE

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	Variance Ratio (F)
Between columns (machines)	1987·55	2	993·78	$F_C = 12{\cdot}15$
Between rows (workers)	1379·55	2	689·78	$F_R = 8{\cdot}43$
Residual	327·12	4	81·78	
Total	3694·22	8		

Step 13. Degrees of freedom—

$$\text{Between columns} = c - 1 = 3 - 1 = 2$$

$$\text{Between rows} = r - 1 = 3 - 1 = 2$$

$$\text{For residual} = (r-1)(c-1)$$

$$= 2 \times 2 = 4$$

Step 14. Level of significane—$\alpha = 0{\cdot}05$ (5%)

Step 15. Critical value—At 5% level of significance and (2, 4) degrees of freedom, the critical value of F is—

$$F_{2,\,4\,(0{\cdot}05)} = 6{\cdot}94 \text{ (from table)}$$

(for both—between columns and between rows)

Step 16. **Decision**—

(*i*) Since computed value of $F_C = 12{\cdot}15$ is greater than the critical value of F = 6·94, it falls in the rejection region. Hence, null hypothesis is rejected and it may be concluded that ***mean productivity is not same for the different machine types.***

(*ii*) Since computed values of $F_R = 8{\cdot}43$ is greater than the critical value of F = 6·94, it also falls in the rejection region. Hence, null hypothesis is rejected and it may be concluded that ***there is a significant difference in the mean productivity of three workers.***

***Illustration* 11.**

You are given the following data relating to the number of table fans sold by

	Salesmen	
		Krishan

months April, May and June :—