18

Analysis of Variance

INTRODUCTION

Earlier in Chapter 17 on chi-square test, we used the chi-square test to examine the difference between more than two proportions and to make inferences about whether such samples are drawn from populations each having the same proportion. In this chapter, we shall learn a technique known as *analysis of variance* to test for the significance of the difference between more than two sample means and to make inferences about whether our samples are drawn from the populations having the same mean. The "analysis of variance" procedure or "F-test" is used in such problems, where we want to test for the significance of the difference among more than two sample means. In fact, the technique of analysis of variance is one of the most powerful statistical methods developed by R.A. Fisher.

The analysis of variance originated in agrarian research and its language is thus loaded with such agricultural terms as *blocks* (referring to land) and *treatments* (referring to populations or samples which are differentiated in terms of varieties of seeds, fertilisers or cultivation methods). Today, analysis of variance finds application in every type of experimental design, in natural sciences as well as social sciences and has become a very broad and technical subject. The methods of analysis of variance are a fundamental part of planned research and the design of experiment, comparative studies are essential in judging the effects of new technology, procedures and policies. Though analysis of variance can be used in a number of ways, in this chapter, an attempt would be made to illustrate some business applications of this highly useful tool.

Assumptions in Analysis of Variance

The analysis of variance technique is based on the following assumptions :

(1) Each sample is drawn randomly from a normal population and the sample statistics tend to reflect the characteristics of the population.

(2) The populations from which the samples are drawn have identical means and variances, *i.e.*,

$$\mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_n$$
$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots\ldots = \sigma_n^2$$

In case we are not in a position to make these assumptions in a particular problem, the analysis of variance technique should not be used. In such cases, we should consider using a "Non-parametric (distribution-free) technique."

Computation of Analysis of Variance

The null hypothesis taken while applying analysis of variance technique is that the means of different samples do not differ significantly. The procedure followed in the analysis of variance would be explained separately for

(1) One-way classification, and

(2) Two-way classification.

However, irrespective of the type of classification, the analysis of variance is a technique of partitioning the total sum of squared deviations of all sample values from the grand mean and is divided into two parts—sum of squares between the samples and sum of squares within the samples. Individual observation in the same treatment samples, however, can differ from each other only because of chance variation, since each individual within the group receives exactly the same treatment.

ONE-WAY CLASSIFICATION

The term 'one-factor analysis of variance' refers to the fact that a single variable or factor of interest is controlled and its effect on the elementary units is observed. In other words, in one-way classification, the data are classified according to only one criterion. Suppose we have k independent random samples of $n_1, n_2, .., n_k$ observations from k populations. The population means are denoted by $\mu_1, \mu_2, ..., \mu_k$. The one-way analysis of variance is designed to test the null hypothesis :

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

i.e., the arithmetic means of the population from which the k samples are randomly drawn are equal to one another. The steps involved in carrying out the analysis are :

(1) Calculate the variance between the samples

The variance (sum of squares) between samples reflects the contribution of both different treatments and chance to inter-sample variability. Sum of squares is a measure of variation. The sum of squares between samples is denoted by SSB. For calculating variance between sample, we take the total of the square of the variations of the means of various samples from the grand mean and divide this total by the degrees of freedom. Thus, the steps in calculating variance between samples will be :

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

(*b*) Calculate the grand mean $\overline{\overline{X}}$. Its value is obtained as follows :

$$\overline{\overline{X}} = \frac{\overline{X}_1, + \overline{X}_2 + ... + \overline{X}_k}{n_1 + n_2 + ... + n_k}$$

(*c*) Take the difference between the means of the various samples and the grand mean ;

(*d*) Square these deviations and obtain the total which will give sum of squares between the samples; and

(*e*) Divide the total obtained in step (*d*) by the degrees of freedom. The degrees of freedom will be one less than the number of samples, *i.e.*, if there are 4 samples then the degrees of freedom will be $4 - 1 = 3$ or in general $v = k - 1$, where k = number of samples.

(2) Calculate the variance within the samples

The variance (sum of squares) within samples measures those inter-sample differences that arise due to chance only. It is denoted by SSW. For calculating the variance within the samples, we take the total of the sum of squares of the deviation of various items from the mean values of the respective samples and divide this total by the degrees of freedom. Thus, the steps in calculating variance within the samples will be :

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

(*b*) Take the deviations of the various observations in a sample from the mean values of the respective samples.

(*c*) Square these deviations and obtain the total which gives the sum of squares within the samples.

(*d*) Divide this total obtained in step (*c*) by the degrees of freedom, the degrees of freedom is obtained by deducting from the total number of observations, the number of samples, *i.e.*, $v = n - k$, where k refers to the number of samples and n refers to the total number of all the observations.

(3) Calculate the *F*-ratio

Calculate the *F*-ratio as follows :

$$F^* = \frac{\text{Variance between the samples}}{\text{Variance within the samples}} \text{ or } F = \frac{S_1^2}{S_2^2}.$$

F is always computed with the variance between the sample means as the numerator and the variance within the sample means as the denominator. The denominator is computed by combining the variance within the *k* samples into single measures.

(4) Compare the calculated value of *F*

Compare the calculated value of *F* with the table value of *F* for the given degrees of freedom at a certain critical level (generally we take 5% level of significance). If the calculated value of *F* is greater than the table value of *F*, it indicates that the difference in sample means is significant, *i.e.*, it could not have arisen due to fluctuations of random sampling or, in other words, the *samples do not come from the same population*. On the other hand, if the calculated value of *F* is less than the table value, the difference is not significant and hence could have arisen due to fluctuations of random sampling.

Illustration 1. As head of a department of a consumer's research organisation, you have the responsibility for testing and comparing lifetimes of four brands of electric bulbs. Suppose you test the lifetime of three electric bulbs of each of the four brands. The data is shown below, each entry representing the lifetime of an electric bulb, measured in hundreds of hours :

	Brand		
A	*B*	*C*	*D*
20	25	24	23
19	23	20	20
21	21	22	20

Can we infer that the mean lifetime of the four brands of electric bulbs are equal ? *(MBA, Univ. of Roorkee, 2000)*

Solution. The null hypothesis is that the mean lifetime of the four brands of electric bulbs are equal, *i.e.*,

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

Let $\bar{X}_1, \bar{X}_2, \bar{X}_3$ and $\bar{X}_4$ denote the mean lifetime of Brand *A*, *B*, *C* and *D* respectively and be the overall grand mean.

Then,

X_1	X_2	X_3	X_4
20	25	24	23
19	23	20	20
21	21	22	20
$\bar{X}_1 = 20$	$\bar{X}_2 = 23$	$\bar{X}_3 = 22$	$\bar{X}_4 = 21$

and $\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{4} = \frac{20 + 23 + 22 + 21}{4} = \frac{86}{4} = 21.5.$

The variance between samples can be computed as follows :

$\bar{X}$	$\bar{\bar{X}}$	$(\bar{X} - \bar{\bar{X}})$	$(\bar{X} - \bar{\bar{X}})^2$
20	21.5	−1.5	2.25
23	21.5	+1.5	2.25
22	21.5	+0.5	0.25
21	21.5	−0.5	0.25
			$\Sigma(\bar{X} - \bar{\bar{X}})^2 = 5.0$

* If this ratio is close to 1, there would be little cause to doubt the null hypothesis of equality to population means. On the other hand, if the variability between groups is large compared to the variability within groups, we would suspect, the null hypothesis to be false.

$$s_{\bar{x}}^2 = \frac{\Sigma\left(\bar{X} - \bar{\bar{X}}\right)^2}{k-1} = \frac{5.0}{4-1} = \frac{5}{3}$$

Put $\sigma_{\bar{x}} = \frac{s_{\bar{x}}}{\sqrt{n}}$ or $s_{\bar{x}}^2 = n\,\sigma_{\bar{x}}^2 = 3 \times \frac{5}{3} = 5.$

[Here, n represents the sample size and not the number of samples (k).]

Therefore, our first estimate of the population variance is based on the variance between the sample means and is given by

$$s_1^{\,2} = 5.$$

The variance within samples can be computed as follows :

Brand A		*Brand B*		*Brand C*		*Brand D*	
X	$(X-\bar{X})^2$	X	$(X-\bar{X})^2$	X	$(X-\bar{X})^2$	X	$(X-\bar{X})^2$
20	0	25	4	24	4	23	4
19	1	23	0	20	4	20	1
21	1	21	4	22	0	20	1
$\bar{X} = 20$	$\Sigma(X-\bar{X})^2 = 2$	$\bar{X} = 23$	$\Sigma(X-\bar{X})^2 = 8$	$\bar{X} = 22$	$\Sigma(X-\bar{X})^2 = 8$	$\bar{X} = 21$	$\Sigma(X-\bar{X})^2 = 6$

Therefore,

Sample variance, $s_1^{\,2} = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{2}{2} = 1$; Sample variance, $s_2^{\,2} = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{8}{2} = 4$

Sample variance, $s_3^{\,2} = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{8}{2} = 4$; Sample variance, $s_4^{\,2} = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{6}{2} = 3.$

Therefore, the pooled estimate s^2 is given by

$$s^2 = \frac{s_1^{\,2} + s_2^{\,2} + s_3^{\,2} + s_4^{\,2}}{4} = \frac{1+4+4+3}{4} = 3.$$

Thus, the second estimate of the population variance based on within the samples is given by

$$s_2^{\,2} = 3$$

Therefore,

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{s_1^{\,2}}{s_2^{\,2}} = \frac{5}{3} = 1.67.$$

From the F-table in the appendix, the table value of F for (3, 8) d.f. and at 5% level of significance is 4.07. Since the computed value of $F = 1.67$ is less than the table value of $F = 4.07$, therefore, we accept our null hypothesis. Hence, the difference is insignificant and we can infer that the average lifetime of different brands of bulbs are equal.

The Analysis of Variance Table

Since there are several steps involved in the computation of both the between and within sample variances, the entire set of results may be organised into an analysis of variance (ANOVA) table. This table is summarised and shown below :

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom d.f.*	*Mean Square MS*	*Variance Ratio F*
Between samples	SSB	$c-1$	$MSB = \frac{SSB}{c-1}$	$F = \frac{MSB}{MSW}$
Within samples	SSW	$n-c$	$MSW = \frac{SSB}{n-c}$	
Total	SST	$n-1$		

To use ANOVA table, it is convenient to use the following short-cut computational formulas :

$$\text{Between samples sum of squares} = SSB = \sum_{j=1}^{c} \frac{T_j^2}{n_j} = \frac{T^2}{N}$$

$$\text{Within samples sum of squares} = SSW = \sum_{j=1}^{c} \sum_{j=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{c} \frac{T_j^2}{n_j}$$

$$\text{Total sum of squares} = SST = \sum_{j=1}^{c} \sum_{j=1}^{n_j} X_{ij}^2 - \frac{T^2}{N}.$$

The format for the ANOVA table using the computational formulas is shown below :

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom d.f.*	*Mean Square MS*	*Variance Ratio F*
Between Samples	$SSB = \sum_{j=1}^{c} \frac{T_j^2}{n_j} - \frac{T^2}{N}$	$c-1$	$MSB = \frac{SSB}{c-1}$	$F = \frac{MSB}{MSW}$
Within Samples	$SSW = \sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{k} \frac{T_j^2}{N_j}$	$n-c$	$MSW = \frac{SSW}{n-c}$	
Total	$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}^2 - \frac{T^2}{N}$	$n-1$		

To use ANOVA table, let us consider Illustration 1 again and see how it helps in computation.

In order to use the computational formulas, the following four quantities must be computed :

$$\sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}^2, T_j, \sum_{j=1}^{c} \frac{T_j^2}{n_j}, \text{ and } \frac{T^2}{N}.$$

To obtain these quantities, let us make the following table :

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
20	400	25	625	24	576	23	529
19	361	23	529	20	400	20	400
21	441	21	441	22	484	20	400
$\Sigma X_1 = 60$	$\Sigma X_1^2 = 1202$	$\Sigma X_2 = 69$	$\Sigma X_2^2 = 1595$	$\Sigma X_3 = 66$	$\Sigma X_3^2 = 1460$	$\Sigma X_4 = 63$	$\Sigma X_4^2 = 1329$

T = sum of all the observations = $SX_1 + SX_2 + SX_3 + SX_4$

$= 60 + 69 + 66 + 63 = 258$

or $$\frac{T^2}{N} = \frac{(258)^2}{12} = \frac{258 \times 258}{12} = 5547.$$

$$\sum_{j=1}^{4} \sum_{i=1}^{3} X_{ij}^2 = X_1^2 + X_2^2 + X_3^2 + X_4^2$$

$$= 1202 + 1595 + 1460 + 1329 = 5586$$

$$\sum_{j=1}^{4} \frac{T_j^2}{n_j} = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} = \frac{(60)^2}{3} + \frac{(69)^2}{3} + \frac{(66)^2}{3} + \frac{(63)^2}{3}$$

$$= \frac{1}{3}[3600 + 4761 + 4356 + 3969] = \frac{16686}{3} = 5562.$$

With this information, the ANOVA table for the electric bulb problem can be set as given below :

ANOVA TABLE : ONE-WAY CLASSIFICATION

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom df*	*Mean Square MS*	*Variance Ratio F*
Between Samples	$SSB = \sum_{j=1}^{c} \frac{T_j^2}{n_j} - \frac{T^2}{N}$ $= 5562 - 5547 = 15$	$4 - 1 = 3$	$MSB = \frac{15}{3}$ $= 5$	$F = \frac{MSB}{MSW}$ $= \frac{5}{3}$ $= 1.67$
Within Samples	$SSW = \sum_{j=1}^{c} \sum_{j=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{c} \frac{T_j^2}{n_j}$ $= 5586 - 5562 = 24$	$12 - 4 = 8$	$MSW = \frac{24}{8}$ $= 3$	
Total	$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}^2 - \frac{T^2}{N}$ $= 5586 - 5547 = 39$	$12 - 1 = 11$		

If we compare the F-ratio with the previous F-ratio obtained, we can see, of course, that the two methods (conceptual and computational) have produced exactly the same results. Therefore, it is recommended that the computational method be utilized for ANOVA table since these computations are generally less tedious and easy to perform.

Coding of data

If we add, subtract, multiply or divide the given data, the solution will not change. To show this, let us subtract all the observations from 20 in the electric bulb illustration. The coded data are given below.

X_1	X_2	X_3	X_4
0	+ 5	+ 4	+ 3
− 1	+ 3	0	0
+ 1	+ 1	+ 2	0

To compute different quantities, let us make the following table :

X_1	X^2_1	X_2	X^2_2	X_3	X^2_3	X_4	X^2_4
0	0	+ 5	25	+ 4	16	+ 3	9
− 1	1	+ 3	9	0	0	0	0
+ 1	1	+ 1	1	+ 2	4	0	0
$\Sigma X_1 = 0$	$\Sigma X^2_1 = 2$	$\Sigma X_2 = 9$	$\Sigma X^2_2 = 35$	$\Sigma X_3 = 6$	$\Sigma X^2_3 = 20$	$\Sigma X_4 = 3$	$\Sigma X^2_4 = 9$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = 0 + 9 + 6 + 3 = 18$$

$$\frac{T^2}{N} = \frac{18 \times 18}{12} = 27$$

$$\Sigma X_{ij}^2 = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 = 2 + 35 + 20 + 9 = 66$$

$$\Sigma \frac{T_j^2}{n_j} = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4} = \frac{(0)^2}{3} + \frac{(9)^2}{3} + \frac{(6)^2}{3} + \frac{(3)^2}{3}$$

$$= [0 + 81 + 36 + 9]/3 = 42.$$

ANOVA TABLE

Source of Variation	*Sum of Squares SS*	*Degrees of Freedom d.f.*	*Mean Square MS*	*Variance Ratio F*
Between Samples	42 – 27 = 15	4 – 1 = 3	$MSB = \frac{15}{3} = 5$	$F = \frac{5}{3}$
Within Samples	66 – 42 = 24	12 – 4 = 8	$MSW = \frac{24}{8} = 3$	= 1.67
Total	66 – 27 = 39	12 – 1 = 11		

It may be noted that, we again get the value of *F*-ratio.

Readers are advised to use the following relation to further simplify calculations :

Total Variance = Variance between samples + Variance within samples.

Therefore, if we know any two values, the third can be automatically obtained. For example, variance within samples = Total Variance – Variance between samples.

Illustration 2. The Amit Merchandising Company wishes to test whether its three salesmen *A*, *B* and *C* tend to make sales of the same size or whether they differ in their selling ability as measured by the average size of their sales. During the last week, out of 14 sales, *A* made 5, *B* made 4 and *C* made 5 calls. The following are the weekly sales (in Rs. Thousand) record of three salesmen :

A	*B*	*C*
300	600	700
400	300	300
300	300	400
500	400	600
0	—	500

Test, whether the three salesmen's average sales differ in size. *(MBA, Bharathidasan Univ., 2000)*

Solution. Let us take the null hypothesis that there is no significant difference in the average sales volume of the three salesmen, *i.e.*, $H_0 : \mu_1 = \mu_2 = \mu_3$. In order to simplify calculations, let us divide each observation by 100 so that the coded data are :

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
3	9	6	36	7	49
4	16	3	9	3	9
3	9	3	9	4	16
5	25	4	16	6	36
0	0			5	25
$\Sigma X_1 = 15$	$\Sigma X_1^2 = 59$	$\Sigma X_2 = 16$	$\Sigma X_2^2 = 70$	$\Sigma X_3 = 25$	$\Sigma X_3^2 = 135$

The sum of the sales of various samples

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 15 + 16 + 25 = 56$$

Correction factor $\frac{T^2}{N} = \frac{(56)^2}{14} = 224$

Total sum of squares $(SST) = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - \frac{T^2}{N}$

$= (59 + 70 + 135) - 224 = 264 - 224 = 40$

Sum of squares between samples $(SSB) = \Sigma \frac{T_j^2}{n_j} - \frac{T^2}{N}$

$= \frac{(15)^2}{5} + \frac{(16)^2}{4} + \frac{(25)^2}{5} - 224 = 45 + 64 + 125 - 224 = 10$

Sum of squares within samples $(SSW) = SST - SSB = 40 - 10 = 30$

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	10	2	5
Within samples	30	11	2.73
Total	40	13	

$F_{(2, 11)} = \frac{5}{2.73} = 1.83$

The table value for $F_{(2, 11)}$ at 5% level of significance = 3.98. The calculated value of F is less than the table value*. Hence, we accept the null hypothesis and conclude that three salesmen do not differ significantly in their selling ability as measured by the average size of their sales.

Illustration 3. Four machines *A, B, C* and *D* are used to produce a certain kind of cotton fabrics. Samples of size 4 with each unit as 100 square metres are selected from the outputs of the machines at random, and the number of flaws in each 100 square metres are counted, with the following result.

A	*B*	*C*	*D*
8	6	14	20
9	8	12	22
11	10	18	25
12	4	9	23

Do you think that there is a significant difference in the performance of the four machines ?

(*MBA, Kumaun Univ., 2006*)

Solution. Let us take the null hypothesis that the machines do not differ significantly in performance, i.e., $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
8	64	6	36	14	196	20	400
9	81	8	64	12	144	22	484
11	121	10	100	18	324	25	625
12	144	4	16	9	81	23	529
$\Sigma X_1 = 40$	$\Sigma X_1^2 = 410$	$\Sigma X_2 = 28$	$\Sigma X_2^2 = 216$	$\Sigma X_3 = 53$	$\Sigma X_3^2 = 745$	$\Sigma X_4 = 90$	$\Sigma X_4^2 = 2038$

$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = 40 + 28 + 53 + 90 = 211$

Correction factor $= \frac{T^2}{N} = \frac{(211)^2}{N} = \frac{44521}{16} = 2782.56$

*It may be pointed out that computer software package include programs for performing analysis of variance calculations.

$$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \frac{T^2}{N}$$

$$= 410 + 216 + 745 + 2038 - 2782.56 = 3409 - 2782.56 = 626.44$$

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

$$= \frac{(40)^2}{4} + \frac{(28)^2}{4} + \frac{(53)^2}{4} + \frac{(90)^2}{4} - 2782.56$$

$$= 400 + 196 + 702.25 + 2025 - 2782.56 = 540.69$$

$$SSW = SST - SSB = 626.44 - 540.69 = 85.75$$

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	540.69	3	180.23
Within samples	85.75	12	7.15
Total	626.44	15	

$$F_{(3,12)} = \frac{180.23}{7.15} = 25.207.$$

The table value for $F_{(3,12)}$ at 1% level of significance is 5.95. The calculated value of F is greater than the table value. Hence, we reject the null hypothesis and conclude that there is a significant difference in the performance of the four machines.

Illustration 4. A random sample is selected from each of three makes of rope and their breaking strength (in pounds) are measured, with the following results :

I	*II*	*III*
70	100	60
72	110	65
75	108	57
80	112	84
83	113	87
	120	73
	107	

Test, whether the breaking strength of the ropes differ significantly.

Solution. Let us take the null hypothesis that the breaking strength of the ropes does not differ significantly, *i.e.*, $H_0 : \mu_1 = \mu_2 = \mu_3$. For simplifying calculations, let us take 80 as common. The coded data are given below :

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
–10	100	20	400	–20	400
–8	64	30	900	–15	225
–5	25	28	784	–23	529
0	0	32	1024	+4	16
+3	9	33	1089	+7	49
		40	1600	–7	49
		27	729		
$\Sigma X_1 = -20$	$\Sigma X_1^2 = 198$	$\Sigma X_2 = 210$	$\Sigma X_2^2 = 6526$	$\Sigma X_3 = -54$	$\Sigma X_3^2 = 1268$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = -20 + 210 - 54 = 136.$$

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(136)^2}{18} = \frac{18496}{18} = 1027.56$$

$$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - \frac{T^2}{N} = 198 + 6526 + 1268 - 1027.56 = 6964.44$$

$$SSB = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} - \frac{T^2}{N} = \frac{(-20)^2}{5} + \frac{(210)^2}{7} + \frac{(-54)^2}{6} - 1027.56$$

$$= 80 + 6300 + 486 - 1027.56 = 5838.44$$

$SSW = SST - SSB = 6964.44 - 5838.44 = 1126$

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	5838.44	2	2919.22
Within samples	1126	15	75.07
Total	6964.44	17	

$$F_{(2,15)} = \frac{2919.22}{75.07} = 38.89$$

The table value for $F_{(2,15)}$ at 5% level of significance is 3.68. The calculated value of F is greater than the table value. The null hypothesis stands rejected. We, therefore, conclude that the breaking strength of the ropes differ significantly.

TWO-WAY CLASSIFICATION

In a one-factor analysis of variance explained earlier, the treatments constitute different levels of a single factor which is controlled in one experiment. There are, however, many situations in which the response variable of interest may be affected by more than one factor. For example, sales of a particular brand of cosmetics, in addition to being affected by the point of sale display, might also be affected by the price charged, the size and or location of the store or the number of competitive products sold by the store. Similarly, petrol mileage may be affected by the type of car driven, the way it is driven, road conditions and other factors in addition to the brand of petrol used.

When it is believed that two independent factors might have an effect on the response variable of interest, it is possible to design the test so that an analysis of variance can be used to test for the effects of the two factors simultaneously. Such a test is called two-factor (way) analysis of variance.

Thus, with the two-factor analysis of variance, we can test two sets of hypothesis with the same data at the same time.

We can plan an experiment in such a way as to study the effects of two factors in the same experiment. For each factor, there will be a number of classes or levels.

The procedure for analysis of variance is somewhat different from the one followed while dealing with problems of one-way classification. In a two-way classification, the analysis of variance table takes the following form :

ANOVA TABLE : TWO-WAY CLASSIFICATION

Source of Variation	*Sum of Squares*	*d.f.*	*Mean Square*
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	$rc - 1$	

SSC = Sum of squares between columns

SSR = Sum of squares between rows

SSE = Sum of squares for the residual

SST = Total sum of squares.

The sum of squares for the source "Residual"* is obtained by subtracting from the total sum of squares, the sum of squares between columns and rows.

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

where, c refers to columns and r refers to rows.

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 = $(c - 1)(r - 1)$

The total sum of squares, sum of squares between columns and sum of squares between rows are obtained in the same way as before.

Residual = Total sum of squares – Sum of squares between columns – Sum of squares between rows.

Illustration 5. A 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 in the following table :

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

Carry out an analysis of variance.

Solution. Let us take the null hypothesis that there is no significant difference between the sales of salesmen and that of seasons. The above data are classified according to criteria (*i*) salesman and (*ii*) season. In order to simplify calculations, we code the data by subtracting 30 from each figure. The data in the coded form are given below :

	Salesman				*Seasons Total*
Season	*A*	*B*	*C*	*D*	
Summer	+6	+6	–9	+5	+8
Winter	–2	–1	+1	+2	0
Monsoon	–4	–2	–1	–1	–8
	0	3	–9	6	Grand Total $T = 0$

Correction factor = $\frac{T^2}{N} = \frac{(0)^2}{12} = 0$ (number of observations or N is 12).

Sum of squares between columns (salesmen)

This will be obtained by squaring the salesmen totals, dividing each total by the number of observations included in it, adding these figures and then subtracting the correction factor from them. Thus, sum of squares between salesmen is

$$= \frac{(0)^2}{3} + \frac{(3)^2}{3} + \frac{(-9)^2}{3} + \frac{(6)^2}{3} - \frac{T^2}{N} = 0 + 3 + 27 + 12 - 0 = 42$$

Degrees of freedom = $v = (4 - 1) = 3$.

Sum of squares between rows (seasons)

This will be obtained by dividing the squares of the season totals by the number of observations that make up each total, adding all such figures and subtracting these from the correction factor. Thus, sum of squares between seasons is

$$= \frac{(8)^2}{4} + \frac{(0)^2}{4} + \frac{(-8)^2}{4} - \frac{T^2}{N} = 16 + 0 + 16 - 0 = 32$$

Degrees of freedom = $v = (3 - 1) = 2$.

* In these types of problems involving two-way classification, "Residual" is the measuring rod testing significance. It represents the magnitude of variations due to forces called 'chance'.

Total sum of squares

This will be obtained by adding the squares of all the observations in the table and subtracting the correction factor therefrom. Thus, total sum of squares is

$$= (6)^2 + (-2)^2 + (-4)^2 + (6)^2 + (-1)^2 + (-2)^2 + (-9)^2 + (1)^2 + (-1)^2 + (5)^2 + (2)^2 + (-1)^2 - \frac{T^2}{N}$$

$$= 210 - 0 = 210$$

$$v = (12 - 1) = 11$$

The above information is presented in the following table :

ANOVA TABLE

Source of Variation	*Sum of Squares*	*d.f.*	*Mean Square*
Between columns (salesmen)	42	3	14
Between rows (seasons)	32	2	16
Residual	136	6	22.67
Total	210	11	

To test the hypothesis that there is no significant difference between the sales of salesmen and of seasons or, in other words, the three independent estimates of variance are the estimates of variance of a common population.

Now, first compare the salesmen variance estimate with the residual variance estimate, thus,

$$F = \frac{\text{Greater Variance}}{\text{Smaller Variance}} = \frac{22.67}{14} = 1.62.$$

The table value of F for 3 and 6 degrees of freedom at 5% level of significance is 4.76. The calculated value is less than this and we conclude that the sales of salesmen do not differ significantly.

Now, let us compare the season variance estimate with the residual variance estimate, thus,

$$F = \frac{\text{Greater Variance}}{\text{Smaller Variance}} = \frac{22.67}{16} = 1.42.$$

The critical value of F for 2 and 6 degrees of freedom at 5% level of significance is 5.14. The calculated value is less than this and hence there is no significant difference in the seasons as far as the sales are concerned. Thus, the test shows that the salesmen and the seasons are alike so far as the sales are concerned.

Illustration 6. The following data represent the number of units of production per day turned out by 5 different workers using 4 different types of machines :

	Machine Type			
	A	*B*	*C*	*D*
Workers				
1	44	36	48	38
2	48	40	50	44
3	37	38	40	36
4	45	34	45	32
5	40	44	50	40

Test (*a*) Whether the mean productivity is the same for 4 different machine types.

(*b*) Whether the 5 workers differ with respect to mean productivity.

Solution. Let us take the null hypothesis that (*a*) the mean productivity is the same for four different machines and (*b*) the 5 workers do not differ with respect to mean productivity. To simplify calculations, let us deduct 40 from each value.

Workers	Machine Type				Total
	A	B	C	D	
1	+ 4	– 4	+ 8	– 2	+ 6
2	+ 8	0	+ 10	+ 4	+ 22
3	– 3	– 2	0	– 4	– 9
4	+ 5	– 6	+ 5	– 8	– 4
5	0	+ 4	+ 10	0	+ 14
	+ 14	– 8	+ 33	– 10	$T = 29$

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(29)^2}{20} = 42.05$$

Sum of squares between machines (Column)

$$= \frac{(14)^2}{5} + \frac{(-8)^2}{5} + \frac{(33)^2}{5} + \frac{(-10)^2}{5} - \frac{T^2}{N}$$

$$= 39.2 + 12.8 + 217.8 + 20 - 42.05 = 289.8 - 42.05 = 247.75$$

$$v = (c - 1) = (4 - 1) = 3$$

Sum of squares between workers (rows)

$$= \frac{(6)^2}{4} + \frac{(22)^2}{4} + \frac{(-9)^2}{4} + \frac{(-4)^2}{4} + \frac{(14)^2}{4} - \frac{T^2}{N}$$

$$= 9 + 121 + 20.25 + 4 + 49 - 42.05 = 161.20$$

$$v = (r - 1) = (5 - 1) = 4$$

Total sum of squares

$$= (4)^2 + (8)^2 + (-3)^2 + (5)^2 + (-4)^2 + (-2)^2 + (-6)^2 + (4)^2 + (8)^2 + (10)^2 + (5)^2 + (10)^2 + (-2)^2 + (4)^2 + (-4)^2 + (-8)^2 - \frac{T^2}{N}$$

$$= 16 + 64 + 9 + 25 + 16 + 4 + 36 + 16 + 64 + 100 + 25 + 100 + 4 + 16 + 16 + 64 - 42.05 = 532.95$$

Residual = Total sum of squares – Sum of squares between machines – Sum of squares between workers

$$= 532.95 - 247.75 - 161.20 = 124$$

$$v = (c - 1)(r - 1) = 3 \times 4 = 12$$

ANOVA TABLE

Source of Variation	S.S.	d.f.	MS	Variation Ratio or F
Between Machines	247.75	3	82.583	$\frac{82.583}{13.78} = 5.99$
Between Workers	161.20	4	40.30	$\frac{40.30}{13.78} = 2.92$
Residual	124	12	10.33	
Total	532.95	19		

(*a*) For $v_{2, 12}$, $F_{0.05} = 3.49$

Since the calculated value (5.99) is greater than the table value (3.49), the null hypothesis is rejected. Hence, the mean productivity is not the same for four different types of machines.

(*b*) For $v_{4, 12}$, $F_{0.05} = 3.26$

The calculated value (2.92) is less than the table value (3.26). The null hypothesis holds true. Hence, the 5 workers do not differ with respect to mean productivity.

MISCELLANEOUS ILLUSTRATIONS

Illustration 7. We wish to test whether there are any differences in the performance of 3 brands of television sets. Samples of size $n = 5$ are selected from each brand and the frequency of repair during the first year of purchase is observed. The results are given below :

	T.V. Brand	
A_1	A_2	A_3
4	7	4
6	4	6
7	3	6
5	6	3
8	5	1

In view of the above data, can it be concluded that there is a significant difference between the three brands ?

Solution. Let us take the null hypothesis that there is no significant difference in the three brands of television sets with regard to their performance, *i.e.*, $H_0 : \mu_1 = \mu_2 = \mu_3$. Carrying out analysis of variance :

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
4	16	7	49	4	16
6	36	4	16	6	36
7	49	3	9	6	36
5	25	6	36	3	9
8	64	5	25	1	1
$\Sigma X_1 = 30$	$\Sigma X_1^2 = 190$	$\Sigma X_2 = 25$	$\Sigma X_2^2 = 135$	$\Sigma X_3 = 20$	$\Sigma X_3^2 = 98$

$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 30 + 25 + 20 = 75$

$$\text{Correction Factor} = \frac{T^2}{N} = \frac{(75)^2}{15} = 375.$$

$$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 - \frac{T^2}{N} = 190 + 135 + 98 - 375 = 48$$

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

$$= \frac{(30)^2}{5} + \frac{(25)^2}{5} + \frac{(20)^2}{5} - 375 = 180 + 125 + 80 - 375 = 10$$

$$SSW = SST - SSB = 48 - 10 = 38.$$

ANOVA TABLE

Source of variation	*Sum of square*	*Degrees of freedom*	*Mean Square*
Between Samples	10	2	5.00
Within Samples	38	12	3.17
Total	48	14	

$$F_{(2, 12)} = \frac{5.0}{3.17} = 1.58.$$

The table value for $F_{(2, 12)}$ at 5% level of significance = 3.89. The calculated value of F is less than the table value. Hence, we accept the null hypothesis and conclude that there is no significant difference between the three brands of television sets.

Illustration 8. The following represent the number of units of production per day turned out by 4 different workers using 5 different types of machines :

MACHINE TYPES

Worker	*A*	*B*	*C*	*D*	*E*	*Total*
1	4	5	3	7	6	25
2	6	8	6	5	4	29
3	7	6	7	8	8	36
4	3	5	4	8	2	22
Total	20	24	20	28	20	112

On the basis of this information, can it be concluded that (*a*) the mean productivity is the same for different machines, (*b*) the workers don't differ with regard to productivity ?

Solution. Let us take the null hypothesis that (*a*) the mean productivity of different machines is same, *i.e.*, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ and that (*b*) the 4 workers don't differ in respect of productivity. Carrying out analysis of variance :

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(112)^2}{20} = 627.2.$$

Sum of squares between machines :

$$= \frac{(20)^2}{4} + \frac{(24)^2}{4} + \frac{(20)^2}{4} + \frac{(28)^2}{4} + \frac{(20)^2}{4} - \text{C.F.}$$

$$= 100 + 144 + 100 + 196 + 100 - 627.2 = 640 - 627.2 = 12.8$$

$$v = (c - 1) = (5 - 1) = 4.$$

Sum of squares between workers :

$$= \frac{(25)^2}{5} + \frac{(29)^2}{5} + \frac{(36)^2}{5} + \frac{(22)^2}{5} - \text{C.F.}$$

$$= 125 + 168.2 + 259.2 + 96.8 - 627.2 = 649.2 - 627.2 = 22$$

$$v = (r - 1) = (4 - 1) = 3.$$

Total sum of squares :

$$= (4)^2 + (6)^2 + (7)^2 + (3)^2 + (5)^2 + (8)^2 + (6)^2 + (5)^2 + (3)^2 + (6)^2 + (7)^2 + (4)^2 + (7)^2 + (5)^2 + (8)^2 + (8)^2 + (6)^2 + (4)^2 + (8)^2 + (2)^2 - 627.2$$

$$= 692 - 627.2 = 64.8.$$

Residual = Total *SS* – *SS* between machines –*SS* between workers

$$= 64.8 - 12.8 - 22.0 = 30$$

$$v = (c - 1)(r - 1) = (5 - 1)(4 - 1) = 12$$

ANOVA TABLE

Source of variation	*SS*	*d.f.*	*Mean Square*	*Variance ratio 'F'*
Between machines	12.8	4	3.20	$\frac{3.2}{2.5} = 1.28$
Between workers	22.0	3	7.33	$\frac{7.33}{2.5} = 2.93$
Residual	30.0	12	2.50	
Total	64.8	19		

For $v_{(4, 12)}$, $F_{0.05} = 3.26$

The calculated value of *F* is less than the table value. Our null hypothesis is true. Hence, there is no significant difference in the mean productivity of five different machines.

For $v_{(3, 12)}$, $F_{0.05} = 3.49$

The calculated value of *F* is less than the table value. Our null hypothesis is true. Hence, there is no significant difference in the mean productivity of four different workers.

Illustration 9. Four salesmen were posted in different areas by a company. The number of units sold by them is given below :

A	20	23	28	29
B	25	32	30	21
C	23	28	35	18
D	15	21	19	25

On the basis of this information, can it be concluded that there is a significant difference in the performance of the salesmen.

Solution. Let us take the null hypothesis that there is no significant difference in the performance of the four salesmen.

Sample I *A*	*Sample II* *B*	*Sample III* *C*	*Sample IV* *D*
20	25	23	15
23	32	28	21
28	30	35	19
29	21	18	25
Total : 100	108	104	80
$\bar{X}$: 25	27	26	20

$$\bar{\bar{X}} = \frac{25 + 27 + 26 + 20}{4} = \frac{98}{4} = 24.5$$

VARIANCE BETWEEN SAMPLES

$(\bar{X}_1 - \bar{\bar{X}})^2$	$(\bar{X}_2 - \bar{\bar{X}})^2$	$(\bar{X}_3 - \bar{\bar{X}})^2$	$(\bar{X}_4 - \bar{\bar{X}})^2$
0.25	6.25	2.25	20.25
0.25	6.25	2.25	20.25
0.25	6.25	2.25	20.25
0.25	6.25	2.25	20.25
1.00	25.00	9.00	81.00

Sum of squares between samples = 1 + 25 + 9 + 81 = 116.

VARIANCE WITHIN SAMPLES

$(X_1 - \bar{X}_1)^2$	$(X_2 - \bar{X}_2)^2$	$(X_3 - \bar{X}_3)^2$	$(X_4 - \bar{X}_4)^2$
25	4	9	25
4	25	4	1
9	9	81	1
16	36	64	25
54	74	158	52

Sum of squares within samples = 54 + 74 + 158 + 52 = 338

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	116	3	38.67
Within samples	338	12	28.17

$$F = \frac{38.67}{28.17} = 1.37$$

For $v_1 = 3$ and $v_2 = 12$, $F_{0.05} = 3.24$. The calculated value of F is less than the table value. The null hypothesis holds true. Hence, it cannot be concluded that there is a significant difference in the performance of the four salesmen.

Illustration 10. The following table gives the yields on 15 sample plots under three varieties of seed :

A :	20	21	23	16	20
B :	18	20	17	15	25
C :	25	28	22	18	32

Find out whether the average yield of land under different varieties of seed show significant differences.

Solution. Let us take the null hypothesis that the average yield of land under different varieties of seed do not differ significantly. Applying analysis of variance technique :

X_1	X_2	X_3
20	18	25
21	20	28
23	17	22
16	15	28
20	25	32
Total : 100	95	135
$\bar{X}$: 20	19	27

$$\bar{\bar{X}} = \frac{20 + 19 + 27}{3} = \frac{66}{3} = 22$$

VARIANCE BETWEEN SAMPLES

$(X_1 - \bar{\bar{X}})^2$	$(X_2 - \bar{\bar{X}})^2$	$(X_3 - \bar{\bar{X}})^2$
4	9	25
4	9	25
4	9	25
4	9	25
4	9	25
20	45	125

Sum of squares between samples = 20 + 45 + 125 = 190

VARIANCE WITHIN SAMPLE

$(X_1 - \bar{X}_1)^2$	$(X_2 - \bar{X}_2)^2$	$(X_3 - \bar{X}_3)^2$
0	1	4
1	1	1
9	4	25
16	16	1
0	36	25
26	58	56

Sum of squares within samples = 26 + 58 + 56 = 140

ANOVA TABLE

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean square*
Between samples	190	2	95
Within samples	140	12	11.67

$$F = \frac{95}{11.67} = 8.14$$

For v_1=2 and v_2=12, the table value of F is 3.88. Since the calculated value is more than the table value, the null hypothesis is rejected. Hence, the average yield of land under different varieties of seed differ significantly.

Illustration 11. Three varieties of coal were analysed by five chemists and the ash content in the varieties was found to be as under :

	Chemist				
Variety	*I*	*II*	*III*	*IV*	*V*
A	9	7	6	5	8
B	7	4	5	4	5
C	6	5	6	7	6

Do the varieties differ significantly in their ash content ?

Solution. Let us take the null hypothesis that there is no significant difference in the varieties with regard to the ash conten
Applying analysis of variance technique :

Sample I	*Sample II*	*Sample III*	*Sample IV*	*Sample V*
X_1	X_2	X_3	X_4	X_5
9	7	6	5	8
7	4	5	4	5
6	5	6	7	6
Total : 22	16	17	16	19
$\bar{X}$: 7.33	5.33	5.66	5.33	6.33

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 + \bar{X}_5}{N} = \frac{7.33 + 5.33 + 5.66 + 5.33 + 6.33}{5} = \frac{29.98}{5} = 5.996 \text{ or } 6$$

VARIANCE BETWEEN SAMPLES

$(\bar{X}_1 - \bar{\bar{X}})^2$	$(\bar{X}_2 - \bar{\bar{X}})^2$	$(\bar{X}_3 - \bar{\bar{X}})^2$	$(\bar{X}_4 - \bar{\bar{X}})^2$	$(\bar{X}_5 - \bar{\bar{X}})^2$
1.7689	0.4489	0.1156	0.4489	0.1089
1.7689	0.4489	0.1156	0.4489	0.1089
1.7689	0.4489	0.1156	0.4489	0.1089
$\Sigma(\bar{X}_1 - \bar{\bar{X}})^2$	$\Sigma(\bar{X}_2 - \bar{\bar{X}})^2$	$\Sigma(\bar{X}_3 - \bar{\bar{X}})^2$	$\Sigma(\bar{X}_4 - \bar{\bar{X}})^2$	$\Sigma(\bar{X}_5 - \bar{\bar{X}})^2$
= 5.3067	= 1.3467	= 0.3468	= 1.3467	= 0.3267

Sum of squares between samples

$$= 5.3067 + 1.3467 + 0.3468 + 1.3467 + 0.3267 = 8.6736.$$

VARIANCE WITHIN SAMPLES

X_1	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)^2$	X_3	$(X_3 - \bar{X}_3)^2$	X_4	$(X_4 - \bar{X}_4)^2$	X_5	$(X_5 - \bar{X}_5)^2$
9	2.789	7	2.789	6	0.116	5	0.109	8	2.789
7	0.109	4	1.769	5	0.436	4	1.769	5	1.769
6	1.769	5	0.109	6	0.116	7	2.789	6	0.109
	$\Sigma(X_1 - \bar{X}_1)^2$		$\Sigma(X_2 - \bar{X}_2)^2$		$\Sigma(X_3 - \bar{X}_3)^2$		$\Sigma(X_4 - \bar{X}_4)^2$		$\Sigma(X_5 - \bar{X}_5)^2$
	= 4.667		= 4.667		= 0.668		= 4.667		= 4.667

Sum of squares within samples

$$= 4.667 + 4.667 + 0.668 + 4.667 + 4.667 = 19.336.$$

TOTAL SUM OF SQUARES

X_1	$(X_1 - \bar{\bar{X}})^2$	X_2	$(X_2 - \bar{\bar{X}})^2$	X_3	$(X_3 - \bar{\bar{X}})^2$	X_4	$(X_4 - \bar{\bar{X}})^2$	X_5	$(X_5 - \bar{\bar{X}})^2$
9	9	7	1	6	0	5	1	8	4
7	1	4	4	5	1	4	4	5	1
6	0	5	1	6	0	7	1	6	0
	$\Sigma(X_1 - \bar{\bar{X}})^2$		$\Sigma(X_2 - \bar{\bar{X}})^2$		$\Sigma(X_3 - \bar{\bar{X}})^2$		$\Sigma(X_4 - \bar{\bar{X}})^2$		$\Sigma(X_5 - \bar{\bar{X}})^2$
	= 10		= 6		= 1		= 6		= 5

Total sum of squares = 10 + 6 + 1 + 6 + 5 = 28.

Total sum of squares = Sum of squares between samples + Sum of squares within samples

= 8.6736 + 19.336 = 28 01.

Hence, our calculations are correct.

ANOVA TABLE

Source of Variation	*SS*	*d.f.*	*Mean Square*
Between samples	8.6736	2	4.337
Within samples	19.336	12	1.6113
Total	28.01		

$$F = \frac{4.327}{1.6613} = 2.60$$

For $v_1 = 2$ and $v_2 = 12$, $F_{0.05} = 3.88$. The calculated value of F is less than the table value. Our hypothesis holds true. Hence, we conclude that there is no significant difference in the ash content of five different varieties.

Illustration 12. The three samples have been obtained from normal populations with equal variances. Test the hypothesis that the population means are equal.

SAMPLE

I	*II*	*III*
8	7	12
10	5	13
7	10	13
14	9	12
11	9	14

Solution. Let us take the null hypothesis that there is no significant difference in the means of three samples.

Sample	*I*	*II*	*III*
	8	7	12
	10	5	9
	7	10	13
	14	9	12
	11	9	14
Total	50	40	60
$\bar{\bar{X}}$	10	8	12

$$\bar{\bar{X}} = \frac{10 + 8 + 12}{3} = \frac{30}{3} = 10.$$

VARIANCE BETWEEN SAMPLES

$(X_1 - \bar{\bar{X}})^2$	$(X_2 - \bar{\bar{X}})^2$	$(X_3 - \bar{\bar{X}})^2$
0	4	4
0	4	4
0	4	4
0	4	4
0	4	4
0	20	20

Sum of squares between samples = 0 + 20 + 20 = 40

VARIANCE WITHIN SAMPLES

$(X_1 - \bar{X})^2$	$(X_2 - \bar{X})^2$	$(X_3 - \bar{X})^2$
4	1	0
0	9	9
9	4	1
16	1	0
1	1	4
30	16	14

Sum of squares within samples = 30 + 16 + 14 = 60

ANOVA TABLE

Source of Variation	*SS*	*d.f.*	*MS*	*F*
Between samples	40	2	20	$\frac{20}{5} = 4$
Within samples	60	12	5	
Total	100	14		

For $v_1 = 2$ and $v_2 = 12$, the table value of F at 5% level of significance is 3.38. The calculated value of F is more than the table value. The hypothesis is rejected. Hence, the population means are not equal.

Caution while Applying Analysis of Variance Technique

The analysis of variance has been developed under a set of rigid assumptions as pointed out in the beginning of the chapter. Whenever, any of these assumptions is not met, the F-test cannot be employed to yield valid inferences. It is indeed fortunate that many economic and business experiments do conform to these assumptions. However, where departure from the premises exist, the analysis of variance may still be applied by way of *transformation.* Transformation refers to a process of transforming the original data into some other form, such as square roots, inverse sines of logarithms, before the analysis is made.

PROBLEMS

Answer the following questions, each question carries **one** mark:

(*i*) What is Analysis of Variance ?

(*ii*) The technique of analysis of variance was developed by .....

(*iii*) Define F-test. *(M. Com., M.K. Univ., 2001)*

(*iv*) Give two applications of analysis of variance.

(*v*) On what assumptions, analysis of variance is based ?

(*vi*) What do you understand by one-way analysis of variance ?

(*vii*) Give the format of ANOVA table in one-way classification.

(*viii*) What is two-way classification in analysis of variance ?

(*ix*) Give the components of source of variation in one-way classification.

(*x*) What is coding method in analysis of variance ?

Answer the following questions, each question carries **four** marks:

(*i*) Explain one-way classification technique in analysis of variance. *(M. Com., M.K. Univ., 2002)*

(*ii*) Tabulate the ANOVA table in one-way classification. *(M. Com., M.K. Univ., 2001)*

(*iii*) Explain the F-test. What are the assumptions of F-test ? *(M. Com., M.K. Univ., 2001)*

(*iv*) Differentiate between one-way and two-way classification by giving suitable example.

(*v*) Explain the procedure involved in ANOVA for testing of a hypothesis.

(*vi*) What is ANOVA? *(MBA, Madras Univ., 2002)*

What is 'analysis of variance' and where it is used ? Give two suitable examples.

How is analysis of variance technique helpful in solving business problems? Illustrate your answer with suitable examples. *(MBA, Kumaun Univ., 2000)*

Briefly describe the procedure followed in analysis of variance.

What are the basic and common assumptions made for analysis of variance ?

Distinguish between one-way and two-way classification models and explain the procedure followed for carrying out analysis of variance.

(*a*) Explain the meaning and significance of Analysis of Variance.

(*b*) State some applications of the analysis of variance.

(*c*) Explain the use of Analysis of variance (ANOVA) to check how good is the regression.

How is the F-distribution related to the Student's t-distribution and the chi-square distribution? What important hypothesis can be tested by the F-distribution?

9. In order to determine, whether there are significant differences in the durability of three makes of computer, samples of size $n = 5$ are selected from each make and the frequency of repair during the first year of purchase is observed. The results are as follow :

	Make	
A	*B*	*C*
5	8	7
6	10	3
8	11	5
9	12	4
7	4	1

In view of the above data, what conclusion can you draw ?

[$F = 5.34$, $F_{2, 12}$ at 5% level = 3.89]

10. A plastic manufacturer tests the tensile strength of different types of polythene material. A sample of three measurements is taken for each material type and data in pounds per square inch are as follows :

Type I	*Type II*	*Type III*
200	260	245
215	255	248
218	277	272

Determine, if the mean tensile strength of the three different types of materials differ significantly.

[$F = 16.30$, $F_{2, 6}$ at 5% level = 5.14, yes] *(MBA, Hyderabad Univ., 2005)*

11. The number of automobiles arriving at four toll stations were recorded for 2 hours time period (10 A.M. to 12 P.M.) for each of six different days. The data are as follows :

Day	*Gate I*	*Gate 2*	*Gate 3*	*Gate 4*
Monday	200	228	212	301
Tuesday	208	230	215	305
Wednesday	225	240	228	288
Thursday	223	242	224	212
Friday	228	210	235	215
Saturday	220	208	245	200

(*a*) Determine, whether the rate of arrival is essentially the same at each toll station.

(*b*) Determine, whether the rate of arrival differs significantly during the six different days of the week or not.

[(*a*) $F = 1.78$, No ; (*b*) $F = 0.56$, No]

12. Following table gives the number of refrigerators sold by 4 salesmen in three months:

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

(*a*) Determine, whether there is any significant difference in the average sales made by four salesmen.

(*b*) Determine, whether the sales differ with respect to different months.

[(*a*) $F = 1.01$, No ; (*b*) F=3.29, No]

13. Miss Neena, a supervisor has 3 typists working under her supervision. She is concerned with the time they spend on the tea in addition to the normal lunch tea break. Her observations recorded in minutes for each typist are as follows :

	Average time (minutes)									
A	25	18	30	32	35	37	19			
B	24	22	26	28	30	32	28	26		
C	28	20	27	19	29	35	30	23	27	32

Can the differences in average time that the three typists spend on tea break be explained by chance variation ?

14. Five different brands of tyres used by a car rental agency in the process of deciding the brand of tyre to purchase as standard equipment for their fleet, find that each of five tyres of each brand last the following number of kilometres (in '000s) :

Tyre Brand				
A	*B*	*C*	*D*	*E*
36	46	35	45	41
37	39	42	36	39
42	35	37	39	37
48	37	43	35	35
47	48	38	32	38

Test the hypothesis that the five different brands of tyres have identical average life.

15. It is suspected that four machines, each in a canning operation fills cans to different levels on the average. Random samples of cans produced from each machine were taken and the fill in ounces was measured. The results are tabulated below :

Machine			
A	*B*	*C*	*D*
10.20	10.22	10.17	10.15
10.18	10.27	10.22	10.37
10.36	10.26	10.34	10.28
10.21	10.25	10.27	10.40
10.25			10.30

Do the machines appear to be filling the cans at different average levels ?

16. During the last week, there were 14 sales calls. *A* made 5 calls, *B* made 4 calls and *C* made 5 calls. Following are the weekly sales (in 000's Rs.) record of the three salesmen :

	Salesmen		
	A	*B*	*C*
Calls	3	6	7
	4	3	3
	3	3	4
	5	4	6
	0	–	5

With the help of analysis of variance, test the selling ability of the three salesmen. *(MBE, Delhi Univ., 2002)*

17. Suppose that we are interested in establishing the yield producing ability of four types of soyabeans, *A, B,* C and *D*. We have three blocks of land *X, Y* and *Z* which may be different in fertility. Each block of land is divided into four plots, and the different types of soyabeans are assigned to the plots in each block by a random procedure. The following results are obtained :

	Type			
Block	*A*	*B*	*C*	*D*
X	5	9	11	10
Y	4	7	8	10
Z	3	5	8	9

Test whether *A, B, C* and *D* are significantly different.

18. The chairman of a large chain of supermarkets was prepared to order a large number of frozen food display cases for use in the markets. Before placing the order, he decided to test the products by storing half-litre containers of milk in the case made by different manufacturers and observing the spoilage time of each types of case. The display case came from three different manufacturers designated, *A,B* and *C*. Nine half-litre containers of milk were randomly selected and assigned, three to each case. The response variable observed was the spoilage time in day. The data for this test is provided below :

Treatment		
A	*B*	*C*
7	8	7
5	4	8
9	7	10

Test for a significant difference in the effect of the display cases at 5% level of significance.

19. The following table gives the monthly sales (in thousand rupees) of a certain firm in three different States by your different salesmen :

		Salesmen		
States	*W*	*X*	*Y*	*Z*
A	10	8	8	14
B	14	16	10	8
C	18	12	12	14

State, whether the difference between sales affected by the four salesmen and difference between sales affected in three States are significant.

20. Four brands of tyres were tested for durability and wear on specially designed machines which simulate road conditions. Four tyres of each brand were subjected to the same test and the number of kilometres until wear out was noted for each tyre. The data in thousands of kilometres is provided below :

	Tyre Brand		
A	*B*	*C*	*D*
24	26	28	12
18	16	17	18
23	19	26	30
13	30	19	20

Test for a significant difference tyre mileage at the 5% level of significance.

21. Four different drugs have been developed for the cure of a certain disease. These drugs are tried on patients in three different hospitals. The results given below show the number of cases of recovery from the disease per 100 people who have taken the drugs. The randomized blocks design has been employed to eliminate the effects of the hospital.

	A	*B*	*C*	*D*
H_1	32	18	20	21
H_2	15	23	26	13
H_3	26	10	17	17

Carry out an analysis of variance and interpret your results.

22. A manufacturer of footballs wants to introduce two additional styles of footballs to accompany the plastic version he already produces. The new footballs will be made of leather and rubber. All three styles were test marketed in five different stores. The manufacturer wants to concentrate on producing the type that promises the most sales. Is there a difference in sales of three types of footballs in the five different stores ?

Store	*Rubber*	*Plastic*	*Leather*
1	550	600	450
2	720	700	300
3	680	750	520
4	600	800	380
5	650	550	250

What should the manufacturer do about marketing his footballs?

23. A manufacturer has just introduced a new product that will be sold in sizes : small, medium and large. Five salesmen are randomly selected from the sales force and given each of the three products to sell. The sales figure for one month are used to find out whether there is a difference in sales volume for the different sizes. The amounts sold by the five salesmen are as follows :

	Size		
Salesman	*Small*	*Medium*	*Large*
1	850	900	880
2	720	880	760
3	880	970	930
4	900	890	670
5	750	960	880

Using a 0.05 level of significance, determine whether there is a significantly difference in the amount sold by size. What is your marketing decision ?

24. An economist wishes to assess the effects of Factor *A* (education) with five levels and Factor *B* (occupation) with four levels upon a person's annual earnings. The following data have been obtained for 20 randomly chosen person :
SSW = 8,00,000; *SSB* = 9,00,000 and Total *SS* = 20,00,000.
(*a*) Construct a one factor ANOVA table, using education as the only treatment. At 5 per cent significance level, can you conclude that the treatment means differ? (*b*) Construct a two factor ANOVA table, using education for occupation. What can you conclude about the respective null hypothesis for identical mean incomes for education levels and occupation ?

25. Mr. Ram wants to build a service station on one of three locations. He measures the traffic passing each location for six days. The following are the average amounts of traffic per hour passing each location for each of the six days :

Day	*Location A*	*Location B*	*Location C*
1	75	85	90
2	78	94	118
3	65	90	125
4	76	68	70
5	88	74	81
6	98	87	80

Is there any significant difference in the amount of traffic passing the three locations ? Where would you advise Mr. Ram to build his service station?

26. A company selling coffee appoints four salesmen *A,B,C* and *D*. Observe their sales in 3 seasons; summer, winter and monsoon.

The figures (in lakhs of rupees) are given below :

	Salesman			
Season	*A*	*B*	*C*	*D*
Summer	30	25	33	20
Winter	28	26	31	35
Monsoon	32	30	32	32

Carry out an analysis of variance and comment on your results.

27. The numbers of automobiles arriving at four gasoline stations were recorded for four-hour period from 8 A.M. to 12 Noon, from Monday through Saturday. Determine, whether the rate of arrival is essentially same at all stations.

Day	*Stations*			
	1	*2*	*3*	*4*
Monday	49	53	48	53
Tuesday	45	51	46	51
Wednesday	51	47	53	49
Thursday	48	53	42	51
Friday	50	50	50	53
Saturday	48	51	47	54

28. Three machines in a workshop are equally efficient. To measure the efficiency of four operators, the data on the number of units produced per shift by each operator on different machines on randomly selected shifts has been collected as follows :

	Machine		
Operator	*A*	*B*	*C*
I	22	20	19
II	*24*	*19*	*17*
III	27	23	21
IV	23	24	18

Test at 5% level of significance, whether machine operators are equally efficient.

29. A large retailer must make a choice between three sales locations within a shopping complex.

The following data are traffic counts for a 7-day period :

Location *X* :	643	542	569	552	607	514	576
Location *Y* :	249	404	378	337	426	298	345
Location *Z* :	458	513	485	482	539	491	368

Is there is significant difference in the average traffic count at the three locations ?

30. The following table gives monthly sales (in thousand rupees) of a certain firm in three States by its four salesmen :

	Salesmen			
States	*I*	*II*	*III*	*IV*
A	6	5	5	8
B	8	9	6	5
C	10	7	8	7

Test, whether there is any significant difference (*i*) between sales by the firm salesmen, and (*ii*) between sales by the four salesmen, and (*iii*) between sales in the three States.

31. The following are the defective pieces produced by four operators working, in turn, on four different machines :

	Operator			
Machine	B_1	B_2	B_3	B_4
A_1	34	28	33	29
A_2	31	24	35	23
A_3	27	20	43	72
A_4	28	28	29	26

Perform analysis of variance at 0.05 level of significance to ascertain whether variability in production is due to variability in operators' performance or variability in machines' performance.

32. 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 analysis of variance, using 5% level of significance.

33. Agricultural engineers conducted an experiment to assess the effects of three different fertilisers on the yields of mango trees. They planted 15 plots of equal size and treated them alike except for the type of fertiliser applied, Fertiliser *A* was applied to 4 plots, fertiliser *B* to 5 plots and fertiliser *C* to 6 plots. The following table shows the yields, in quintals, per plot. Do these data provide sufficient evidence to indicate a difference in the treatment effect ? Use 5% level of significance.

Fertiliser						
A	7	5	6	4		
B	8	5	5	5	2	
C	5	6	4	1	2	3

34. The performances of a class of 300 students in the subjects of Statistics and Finance were graded into four classes *A, B,C* and *D*. The table below gives the cross tabulation of the number of students by grades in each of the two subjects :

	Statistics			
Finance	*A*	*B*	*C*	*D*
A	12	12	10	6
B	16	25	12	7
C	18	21	14	17
D	4	12	9	5

Test at significance of 5% and 1%, whether the performance can be inferred as independent.

35. The following table gives the number of units of production per day turned out by four different employees, using four different types of machines :

	Type of machines			
Employee	M_1	M_2	M_3	M_4
E_1	40	36	45	30
E_2	38	42	50	41
E_3	36	30	48	35
E_4	46	47	52	44

Using analysis of variance (*i*) test the hypothesis that the mean production is the same for the four machines and (*ii*) test the hypothesis that the four employees do not differ with respect to mean productivity.

36. In order to evaluate four comparable typewriters of different brands, five typists are randomly assigned to each machine and asked to type the same copy matter for 10 minutes. At the end of the period, the words per minute (wpm) are recorded. The data are presented in the table below.

	Output from typewriter (wpm)				
Typewriter	*A*	*B*	*C*	*D*	*E*
A Brand	69	62	70	57	62
B Brand	67	72	76	69	71
C Brand	76	70	71	66	77
D Brand	60	64	67	58	66

Carry out an analysis of variance to assess whether the mean wpm on the different brands of typewriters may be assumed to be the same, or are different.

37. Experiments were performed to determine whether the yield from a chemical process is influenced by the concentration of the catalyst and the temperature of the reaction. Five different concentration levels C_1 to C_5 were combined with three levels of temperature T_1 to T_3.

	Concentration levels of Catalyst				
Temperature Levels	C_1	C_2	C_3	C_4	C_5
T_1	66	72	59	74	68
T_2	64	70	62	73	72
T_3	68	74	64	70	70

Test at 5 per cent significance level, whether the mean yields are influenced by concentration of catalyst or by temperature of reaction.

38. In a certain factory, production can be accomplished by four different types of machines. A sample study, in context of a two-way design without repeated values, is being made with two-fold objectives of examining whether the four workers differ with respect to mean productivity and whether the mean productivity is the same for the five different machines. The researcher involved in this study reports while analysing the gathered data as under:
   (i) Sum of squares for variance between machines = 35.2
   (ii) Sum of squares for variance between workmen = 53.4
   (iii) Sum of squares for total variance = 174.2
   Set up ANOVA table for the given information and draw the inference about variance at 5 per cent level of significance.

39. Three training methods were compared to see if they led to greater productivity after training. Below are productivity measures for individuals trained by each method :

Method 1 :	10	6	8	12	6					
Method 2 :	6	6	7	9	4	6	10	5	6	8
Method 3 :	11	8	13	10	10	12				

   At 0.05% level of significance, do the three training methods lead to different levels of productivity ?

40. Perform a two-way ANOVA on the data given below :

Plots of Land	*Treatment*			
	A	*B*	*C*	*D*
I	38	40	41	39
II	45	42	49	36
III	40	38	42	42

41. The following data pertain to the number of units of a product manufactured per day by five workmen from four different brands of machines.

	Machine Brands			
Workmen	*A*	*B*	*C*	*D*
1	46	40	49	38
2	48	42	54	45
3	36	38	46	34
4	35	40	48	35
5	40	44	51	41

   (i) Test, whether the mean productivity is the same for the four brands of machine type.
   (ii) Test, whether five different workmen differ with respect to productivity. *(M.Com., DU,1999)*

42. The following data represent the number of units produced by 4 operators during 3 different shifts :

	Operator			
Shifts	*A*	*B*	*C*	*D*
I	10	8	12	13
II	10	12	14	15
III	12	10	11	14

   Perform a two-way analysis of variance and interpret the result. *(MBA, Madras Univ., 2005)*

43. What is 'Analysis of variance' and where it is used ? Given below are the lives (in hours) of three randomly selected batches of electric lamps.

Batch 1	1610	1615	1625	1630	
" 2	1590	1605	1620		
" 3	1580	1585	1600	1610	1625

   Analyse the data and draw your conclusions.

   For $\alpha = 0.5$, $F_{2,9} = 4.26$, $F_{2,10} = 4.10$, $F_{2,11} = 3.98$
   $F_{3,8} = 4.07$, $F_{3,9} = 3.87$, $F_{4,7} = 4.12$ *(M. Com., A.M.U., 2001)*

44. As part of the investigation of the collapse of the roof of a building, a testing laboratory is given all the available bolts that connected the steel structure at three different positions on the roof. The forces required to shear each of these bolts (coded values) are as follows :

Position 1	:	90	82	79	98	83	91	
Position 2	:	105	89	93	104	89	95	86
Position 3	:	83	89	80	94			

   Perform an analysis of variance to test at the 0.05 level of significance, whether the differences among the sample means at the three positions are significant. *(B.E./B.Tech, Madras Univ., 2003)*

45. The R & D manager of an automobile company wishes to study the effect of "Tyre Brand" on the tread loss (in millimetre) of tires. Four tyres from each of four different brands (*A, B, C* and *D*) are fitted to four different cars using the completely randomized design. The data as per this design are presented below :

Tyre Brand			
A	*B*	*C*	*D*
6	3	8	4
7	6	6	2
10	2	7	1
9	3	2	4

(*i*) Write the corresponding model.

(*ii*) Check whether the tyre brand has effect on the tread loss of tyres at a significant level of 5%.

(*MBA, Bharathidasan Univ., 2002*)

46. There are three main brands of a certain powder. A set of 120 sales is examined and found to be allocated among four groups (*A, B, C* and *D*) and brands (*I, II* and *III*) as shown below :

		Replications			
		Groups			
	Brands	*A*	*B*	*C*	*D*
	I	0	4	8	15
Factor	II	5	8	13	6
	III	18	19	11	13

Check whether the factor "Brand" has significant effect on the sales at $\alpha = 0.05$ using one way ANOVA.

(*MBA, Bharathidasan Univ., 2006*)

47. The following are the number of mistakes made in 5 successive days by 4 technicians working for a photographic laboratory. Test at a level of significance $\alpha = 0.01$, whether the differences among the four sample means can be attributed to chance.

Mistakes	*Technician I*	*Technician II*	*Technician III*	*Technician IV*
Day 1	6	14	10	9
Day 2	14	9	12	12
Day 3	10	12	7	8
Day 4	8	10	15	10
Day 5	11	14	11	11

(*MBA, Anna Univ., 2007*)

*****