11

Analysis of Variance and Co-variance

ANALYSIS OF VARIANCE (ANOVA)

Analysis of variance (abbreviated as ANOVA) is an extremely useful technique concerning researches in the fields of economics, biology, education, psychology, sociology, business/industry and in researches of several other disciplines. This technique is used when multiple sample cases are involved. As stated earlier, the significance of the difference between the means of two samples can be judged through either z-test or the t-test, but the difficulty arises when we happen to examine the significance of the difference amongst more than two sample means at the same time. The ANOVA technique enables us to perform this simultaneous test and as such is considered to be an important tool of analysis in the hands of a researcher. Using this technique, one can draw inferences about whether the samples have been drawn from populations having the same mean.

The ANOVA technique is important in the context of all those situations where we want to compare more than two populations such as in comparing the yield of crop from several varieties of seeds, the gasoline mileage of four automobiles, the smoking habits of five groups of university students and so on. In such circumstances one generally does not want to consider all possible combinations of two populations at a time for that would require a great number of tests before we would be able to arrive at a decision. This would also consume lot of time and money, and even then certain relationships may be left unidentified (particularly the interaction effects). Therefore, one quite often utilizes the ANOVA technique and through it investigates the differences among the means of all the populations simultaneously.

WHAT IS ANOVA?

Professor R.A. Fisher was the first man to use the term 'Variance'* and, in fact, it was he who developed a very elaborate theory concerning ANOVA, explaining its usefulness in practical field.

* Variance is an important statistical measure and is described as the mean of the squares of deviations taken from the mean of the given series of data. It is a frequently used measure of variation. Its squareroot is known as standard deviation, i.e., Standard deviation = $\sqrt{\text{Variance.}}$

Later on Professor Snedecor and many others contributed to the development of this technique. ANOVA is essentially a procedure for testing the difference among different groups of data for homogeneity. "The essence of ANOVA is that the total amount of variation in a set of data is broken down into two types, that amount which can be attributed to chance and that amount which can be attributed to specified causes."[1] There may be variation between samples and also within sample items. ANOVA consists in splitting the variance for analytical purposes. Hence, it is a method of analysing the variance to which a response is subject into its various components corresponding to various sources of variation. Through this technique one can explain whether various varieties of seeds or fertilizers or soils differ significantly so that a policy decision could be taken accordingly, concerning a particular variety in the context of agriculture researches. Similarly, the differences in various types of feed prepared for a particular class of animal or various types of drugs manufactured for curing a specific disease may be studied and judged to be significant or not through the application of ANOVA technique. Likewise, a manager of a big concern can analyse the performance of various salesmen of his concern in order to know whether their performances differ significantly.

Thus, through ANOVA technique one can, in general, investigate any number of factors which are hypothesized or said to influence the dependent variable. One may as well investigate the differences amongst various categories within each of these factors which may have a large number of possible values. If we take only one factor and investigate the differences amongst its various categories having numerous possible values, we are said to use one-way ANOVA and in case we investigate two factors at the same time, then we use two-way ANOVA. In a two or more way ANOVA, the interaction (i.e., inter-relation between two independent variables/factors), if any, between two independent variables affecting a dependent variable can as well be studied for better decisions.

THE BASIC PRINCIPLE OF ANOVA

The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples. In terms of variation within the given population, it is assumed that the values of (X_{ij}) differ from the mean of this population only because of random effects i.e., there are influences on (X_{ij}) which are unexplainable, whereas in examining differences between populations we assume that the difference between the mean of the *j*th population and the grand mean is attributable to what is called a 'specific factor' or what is technically described as treatment effect. Thus while using ANOVA, we assume that each of the samples is drawn from a normal population and that each of these populations has the same variance. We also assume that all factors other than the one or more being tested are effectively controlled. This, in other words, means that we assume the absence of many factors that might affect our conclusions concerning the factor(s) to be studied.

In short, we have to make two estimates of population variance viz., one based on between samples variance and the other based on within samples variance. Then the said two estimates of population variance are compared with *F*-test, wherein we work out.

$$F = \frac{\text{Estimate of population variance based on between samples variance}}{\text{Estimate of population variance based on within samples variance}}$$

[1] Donald L. Harnett and James L. Murphy, *Introductory Statistical Analysis*, p. 376.

This value of F is to be compared to the F-limit for given degrees of freedom. If the F value we work out is equal or exceeds* the F-limit value (to be seen from F tables No. 4(a) and 4(b) given in appendix), we may say that there are significant differences between the sample means.

ANOVA TECHNIQUE

One-way (*or single factor*) *ANOVA:* Under the one-way ANOVA, we consider only one factor and then observe that the reason for said factor to be important is that several possible types of samples can occur within that factor. We then determine if there are differences within that factor. The technique involves the following steps:

(i) Obtain the mean of each sample i.e., obtain

$$\overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots, \overline{X}_k$$

when there are k samples.

(ii) Work out the mean of the sample means as follows:

$$\overline{\overline{X}} = \frac{\overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \ldots + \overline{X}_k}{\text{No. of samples } (k)}$$

(iii) Take the deviations of the sample means from the mean of the sample means and calculate the square of such deviations which may be multiplied by the number of items in the corresponding sample, and then obtain their total. This is known as the sum of squares for variance between the samples (or *SS* between). Symbolically, this can be written:

$$SS \text{ between} = n_1\left(\overline{X}_1 - \overline{\overline{X}}\right)^2 + n_2\left(\overline{X}_2 - \overline{\overline{X}}\right)^2 + \ldots + n_k\left(\overline{X}_k - \overline{\overline{X}}\right)^2$$

(iv) Divide the result of the (iii) step by the degrees of freedom between the samples to obtain variance or mean square (*MS*) between samples. Symbolically, this can be written:

$$MS \text{ between} = \frac{SS \text{ between}}{(k-1)}$$

where $(k - 1)$ represents degrees of freedom (d.f.) between samples.

(v) Obtain the deviations of the values of the sample items for all the samples from corresponding means of the samples and calculate the squares of such deviations and then obtain their total. This total is known as the sum of squares for variance within samples (or *SS* within). Symbolically this can be written:

$$SS \text{ within} = \Sigma\left(X_{1i} - \overline{X}_1\right)^2 + \Sigma\left(X_{2i} - \overline{X}_2\right)^2 + \ldots + \Sigma\left(X_{ki} - \overline{X}_k\right)^2$$
$$i = 1, 2, 3, \ldots$$

(vi) Divide the result of (v) step by the degrees of freedom within samples to obtain the variance or mean square (*MS*) within samples. Symbolically, this can be written:

* It should be remembered that ANOVA test is always a one-tailed test, since a low calculated value of F from the sample data would mean that the fit of the sample means to the null hypothesis (viz., $\overline{X}_1 = \overline{X}_2 \ldots = \overline{X}_k$) is a very good fit.

$$MS \text{ within} = \frac{SS \text{ within}}{(n-k)}$$

where $(n - k)$ represents degrees of freedom within samples,
n = total number of items in all the samples i.e., $n_1 + n_2 + \ldots + n_k$
k = number of samples.

(vii) For a check, the sum of squares of deviations for total variance can also be worked out by adding the squares of deviations when the deviations for the individual items in all the samples have been taken from the mean of the sample means. Symbolically, this can be written:

$$SS \text{ for total variance } = \Sigma\left(X_{ij} - \overline{\overline{X}}\right)^2 \qquad i = 1, 2, 3, \ldots$$
$$j = 1, 2, 3, \ldots$$

This total should be equal to the total of the result of the (iii) and (v) steps explained above i.e.,

$$SS \text{ for total variance} = SS \text{ between} + SS \text{ within.}$$

The degrees of freedom for total variance will be equal to the number of items in all samples minus one i.e., $(n - 1)$. The degrees of freedom for between and within must add up to the degrees of freedom for total variance i.e.,

$$(n-1) = (k-1) + (n-k)$$

This fact explains the additive property of the ANOVA technique.

(viii) Finally, F-ratio may be worked out as under:

$$F\text{-ratio} = \frac{MS \text{ between}}{MS \text{ within}}$$

This ratio is used to judge whether the difference among several sample means is significant or is just a matter of sampling fluctuations. For this purpose we look into the table*, giving the values of F for given degrees of freedom at different levels of significance. If the worked out value of F, as stated above, is less than the table value of F, the difference is taken as insignificant i.e., due to chance and the null-hypothesis of no difference between sample means stands. In case the calculated value of F happens to be either equal or more than its table value, the difference is considered as significant (which means the samples could not have come from the same universe) and accordingly the conclusion may be drawn. The higher the calculated value of F is above the table value, the more definite and sure one can be about his conclusions.

SETTING UP ANALYSIS OF VARIANCE TABLE

For the sake of convenience the information obtained through various steps stated above can be put as under:

* An extract of table giving F-values has been given in Appendix at the end of the book in Tables 4 (a) and 4 (b).

Table 11.1: Analysis of Variance Table for One-way Anova
(There are k samples having in all n items)

Source of variation	*Sum of squares (SS)*	*Degrees of freedom (d.f.)*	*Mean Square (MS) (This is SS divided by d.f.) and is an estimation of variance to be used in F-ratio*	*F-ratio*
Between samples or categories	$n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + \ldots + n_k\left(\bar{X}_k - \bar{\bar{X}}\right)^2$	$(k-1)$	$\dfrac{SS \text{ between}}{(k-1)}$	$\dfrac{MS \text{ between}}{MS \text{ within}}$
Within samples or categories	$\Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \ldots + \Sigma\left(X_{ki} - \bar{X}_k\right)^2$ $i = 1, 2, 3, \ldots$	$(n-k)$	$\dfrac{SS \text{ within}}{(n-k)}$	
Total	$\Sigma\left(X_{ij} - \bar{\bar{X}}\right)^2$ $i = 1, 2, \ldots$ $j = 1, 2, \ldots$	$(n-1)$		

SHORT-CUT METHOD FOR ONE-WAY ANOVA

ANOVA can be performed by following the short-cut method which is usually used in practice since the same happens to be a very convenient method, particularly when means of the samples and/or mean of the sample means happen to be non-integer values. The various steps involved in the short-cut method are as under:

(i) Take the total of the values of individual items in all the samples i.e., work out ΣX_{ij}
$i = 1, 2, 3, \ldots$
$j = 1, 2, 3, \ldots$
and call it as T.

(ii) Work out the correction factor as under:

$$\text{Correction factor} = \frac{(T)^2}{n}$$

(iii) Find out the square of all the item values one by one and then take its total. Subtract the correction factor from this total and the result is the sum of squares for total variance. Symbolically, we can write:

$$\text{Total } SS = \Sigma X_{ij}^2 - \frac{(T)^2}{n} \qquad i = 1, 2, 3, \ldots$$
$$j = 1, 2, 3, \ldots$$

(iv) Obtain the square of each sample total $(T_j)^2$ and divide such square value of each sample by the number of items in the concerning sample and take the total of the result thus obtained. Subtract the correction factor from this total and the result is the sum of squares for variance between the samples. Symbolically, we can write:

$$SS \text{ between} = \Sigma\frac{(T_j)^2}{n_j} - \frac{(T)^2}{n} \qquad j = 1, 2, 3, \ldots$$

where subscript j represents different samples or categories.

(v) The sum of squares within the samples can be found out by subtracting the result of (iv) step from the result of (iii) step stated above and can be written as under:

$$SS \text{ within} = \left\{\Sigma X_{ij}^2 - \frac{(T)^2}{n}\right\} - \left\{\Sigma\frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}\right\}$$

$$= \Sigma X_{ij}^2 - \Sigma\frac{(T_j)^2}{n_j}$$

After doing all this, the table of ANOVA can be set up in the same way as explained earlier.

CODING METHOD

Coding method is furtherance of the short-cut method. This is based on an important property of F-ratio that its value does not change if all the n item values are either multiplied or divided by a common figure or if a common figure is either added or subtracted from each of the given n item values. Through this method big figures are reduced in magnitude by division or subtraction and computation work is simplified without any disturbance on the F-ratio. This method should be used specially when given figures are big or otherwise inconvenient. Once the given figures are converted with the help of some common figure, then all the steps of the short-cut method stated above can be adopted for obtaining and interpreting F-ratio.

Illustration 1

Set up an analysis of variance table for the following per acre production data for three varieties of wheat, each grown on 4 plots and state if the variety differences are significant.

Plot of land	*Per acre production data*		
	Variety of wheat		
	A	*B*	*C*
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Solution: We can solve the problem by the direct method or by short-cut method, but in each case we shall get the same result. We try below both the methods.

Solution through direct method: First we calculate the mean of each of these samples:

$$\bar{X}_1 = \frac{6+7+3+8}{4} = 6$$

$$\bar{X}_2 = \frac{5+5+3+7}{4} = 5$$

$$\bar{X}_3 = \frac{5+4+3+4}{4} = 4$$

Mean of the sample means or $\bar{\bar{X}} = \dfrac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{k}$

$$= \frac{6+5+4}{3} = 5$$

Now we work out *SS* between and *SS* within samples:

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + n_3\left(\bar{X}_3 - \bar{\bar{X}}\right)^2$$

$$= 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2$$

$$= 4 + 0 + 4$$

$$= 8$$

$$SS \text{ within} = \Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2 + \Sigma\left(X_{3i} - \bar{X}_3\right)^2, \qquad i = 1, 2, 3, 4$$

$$= \{(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2\}$$

$$+ \{(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2\}$$

$$+ \{(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2\}$$

$$= \{0 + 1 + 9 + 4\} + \{0 + 0 + 4 + 4\} + \{1 + 0 + 1 + 0\}$$

$$= 14 + 8 + 2$$

$$= 24$$

$$SS \text{ for total variance } = \Sigma\left(X_{ij} - \overline{\overline{X}}\right)^2 \qquad i = 1, 2, 3\ldots$$
$$j = 1, 2, 3\ldots$$
$$= (6-5)^2 + (7-5)^2 + (3-5)^2 + (8-5)^2$$
$$+ (5-5)^2 + (5-5)^2 + (3-5)^2$$
$$+ (7-5)^2 + (5-5)^2 + (4-5)^2$$
$$+ (3-5)^2 + (4-5)^2$$
$$= 1 + 4 + 4 + 9 + 0 + 0 + 4 + 4 + 0 + 1 + 4 + 1$$
$$= 32$$

Alternatively, it (*SS* for total variance) can also be worked out thus:

SS for total = *SS* between + *SS* within

$$= 8 + 24$$
$$= 32$$

We can now set up the ANOVA table for this problem:

Table 11.2

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	(3 – 1) = 2	8/2 = 4.00	4.00/2.67 = 1.5	*F*(2, 9) = 4.26
Within sample	24	(12 – 3) = 9	24/9 = 2.67		
Total	32	(12 – 1) = 11			

The above table shows that the calculated value of *F* is 1.5 which is less than the table value of 4.26 at 5% level with d.f. being $v_1 = 2$ and $v_2 = 9$ and hence could have arisen due to chance. This analysis supports the null-hypothesis of no difference is sample means. We may, therefore, conclude that the difference in wheat output due to varieties is insignificant and is just a matter of chance.

Solution through short-cut method: In this case we first take the total of all the individual values of *n* items and call it as *T*.

T in the given case = 60

and $n = 12$

Hence, the correction factor = $(T)^2/n = 60 \times 60/12 = 300$. Now total *SS*, *SS* between and *SS* within can be worked out as under:

$$\text{Total } SS = \Sigma X_{ij}^2 - \frac{(T)^2}{n} \qquad i = 1, 2, 3, \ldots$$
$$j = 1, 2, 3, \ldots$$

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

$$+ (7)^2 + (5)^2 + (4)^2 + (3)^2 + (4)^2 - \left(\frac{60 \times 60}{12}\right)$$

$$= 332 - 300 = 32$$

$$SS \text{ between} = \sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$$

$$= \left(\frac{24 \times 24}{4}\right) + \left(\frac{20 \times 20}{4}\right) + \left(\frac{16 \times 16}{4}\right) - \left(\frac{60 \times 60}{12}\right)$$

$$= 144 + 100 + 64 - 300$$

$$= 8$$

$$SS \text{ within} = \sum X_{ij}^2 - \sum \frac{(T_j)^2}{n_j}$$

$$= 332 - 308$$

$$= 24$$

It may be noted that we get exactly the same result as we had obtained in the case of direct method. From now onwards we can set up ANOVA table and interpret *F*-ratio in the same manner as we have already done under the direct method.

TWO-WAY ANOVA

Two-way ANOVA technique is used when the data are classified on the basis of two factors. For example, the agricultural output may be classified on the basis of different varieties of seeds and also on the basis of different varieties of fertilizers used. A business firm may have its sales data classified on the basis of different salesmen and also on the basis of sales in different regions. In a factory, the various units of a product produced during a certain period may be classified on the basis of different varieties of machines used and also on the basis of different grades of labour. Such a two-way design may have repeated measurements of each factor or may not have repeated values. The ANOVA technique is little different in case of repeated measurements where we also compute the interaction variation. We shall now explain the two-way ANOVA technique in the context of both the said designs with the help of examples.

(a) *ANOVA technique in context of two-way design when repeated values are not there:* As we do not have repeated values, we cannot directly compute the sum of squares within samples as we had done in the case of one-way ANOVA. Therefore, we have to calculate this residual or error variation by subtraction, once we have calculated (just on the same lines as we did in the case of one-way ANOVA) the sum of squares for total variance and for variance between varieties of one treatment as also for variance between varieties of the other treatment.

The various steps involved are as follows:

(i) Use the coding device, if the same simplifies the task.

(ii) Take the total of the values of individual items (or their coded values as the case may be) in all the samples and call it T.

(iii) Work out the correction factor as under:

$$\text{Correction factor} = \frac{(T)^2}{n}$$

(iv) Find out the square of all the item values (or their coded values as the case may be) one by one and then take its total. Subtract the correction factor from this total to obtain the sum of squares of deviations for total variance. Symbolically, we can write it as:

Sum of squares of deviations for total variance or total SS

$$= \Sigma X_{ij}^2 - \frac{(T)^2}{n}$$

(v) Take the total of different columns and then obtain the square of each column total and divide such squared values of each column by the number of items in the concerning column and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between columns or (SS between columns).

(vi) Take the total of different rows and then obtain the square of each row total and divide such squared values of each row by the number of items in the corresponding row and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between rows (or SS between rows).

(vii) Sum of squares of deviations for residual or error variance can be worked out by subtracting the result of the sum of (v)th and (vi)th steps from the result of (iv)th step stated above. In other words,

Total SS – (SS between columns + SS between rows)

= SS for residual or error variance.

(viii) Degrees of freedom (d.f.) can be worked out as under:

d.f. for total variance	$= (c \cdot r - 1)$
d.f. for variance between columns	$= (c - 1)$
d.f. for variance between rows	$= (r - 1)$
d.f. for residual variance	$= (c - 1)\ (r - 1)$

where c = number of columns

r = number of rows

(ix) ANOVA table can be set up in the usual fashion as shown below:

Table 11.3: Analysis of Variance Table for Two-way Anova

Source of variation	*Sum of squares (SS)*	*Degrees of freedom (d.f.)*	*Mean square (MS)*	*F-ratio*
Between columns treatment	$\sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$	$(c-1)$	$\frac{SS \text{ between columns}}{(c-1)}$	$\frac{MS \text{ between columns}}{MS \text{ residual}}$
Between rows treatment	$\sum \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$	$(r-1)$	$\frac{SS \text{ between rows}}{(r-1)}$	$\frac{MS \text{ between rows}}{MS \text{ residual}}$
Residual or error	Total *SS* – (*SS* between columns + *SS* between rows)	$(c-1)(r-1)$	$\frac{SS \text{ residual}}{(c-1)(r-1)}$	
Total	$\sum X_{ij}^2 - \frac{(T)^2}{n}$	$(c.r-1)$		

In the table c = number of columns

r = number of rows

SS residual = Total *SS* – (*SS* between columns + *SS* between rows).

Thus, *MS* residual or the residual variance provides the basis for the *F*-ratios concerning variation between columns treatment and between rows treatment. *MS* residual is always due to the fluctuations of sampling, and hence serves as the basis for the significance test. Both the *F*-ratios are compared with their corresponding table values, for given degrees of freedom at a specified level of significance, as usual and if it is found that the calculated *F*-ratio concerning variation between columns is equal to or greater than its table value, then the difference among columns means is considered significant. Similarly, the *F*-ratio concerning variation between rows can be interpreted.

Illustration 2

Set up an analysis of variance table for the following two-way design results:

Per Acre Production Data of Wheat

(*in metric tonnes*)

Varieties of seeds	*A*	*B*	*C*
Varieties of fertilizers			
W	6	5	5
X	7	5	4
Y	3	3	3
Z	8	7	4

Also state whether variety differences are significant at 5% level.

Solution: As the given problem is a two-way design of experiment without repeated values, we shall adopt all the above stated steps while setting up the ANOVA table as is illustrated on the following page.

ANOVA table can be set up for the given problem as shown in Table 11.5.

From the said ANOVA table, we find that differences concerning varieties of seeds are insignificant at 5% level as the calculated F-ratio of 4 is less than the table value of 5.14, but the variety differences concerning fertilizers are significant as the calculated F-ratio of 6 is more than its table value of 4.76.

(b) *ANOVA technique in context of two-way design when repeated values are there:* In case of a two-way design with repeated measurements for all of the categories, we can obtain a separate independent measure of inherent or smallest variations. For this measure we can calculate the sum of squares and degrees of freedom in the same way as we had worked out the sum of squares for variance within samples in the case of one-way ANOVA. Total *SS*, *SS* between columns and *SS* between rows can also be worked out as stated above. We then find left-over sums of squares and left-over degrees of freedom which are used for what is known as '*interaction variation*' (Interaction is the measure of inter relationship among the two different classifications). After making all these computations, ANOVA table can be set up for drawing inferences. We illustrate the same with an example.

Table 11.4: Computations for Two-way Anova (in a design without repeated values)

Step (i) $T = 60, n = 12, \therefore$ Correction factor $= \dfrac{(T)^2}{n} = \dfrac{60 \times 60}{12} = 300$

Step (ii) Total $SS = (36 + 25 + 25 + 49 + 25 + 16 + 9 + 9 + 9 + 64 + 49 + 16) - \left(\dfrac{60 \times 60}{12}\right)$

$= 332 - 300$

$= 32$

Step (iii) SS between columns treatment $= \left[\dfrac{24 \times 24}{4} + \dfrac{20 \times 20}{4} + \dfrac{16 \times 16}{4}\right] - \left[\dfrac{60 \times 60}{12}\right]$

$= 144 + 100 + 64 - 300$

$= 8$

Step (iv) SS between rows treatment $= \left[\dfrac{16 \times 16}{3} + \dfrac{16 \times 16}{3} + \dfrac{9 \times 9}{3} + \dfrac{19 \times 19}{3}\right] - \left[\dfrac{60 \times 60}{12}\right]$

$= 85.33 + 85.33 + 27.00 + 120.33 - 300$

$= 18$

Step (v) SS residual or error = Total SS – (SS between columns + SS between rows)

$= 32 - (8 + 18)$

$= 6$

Table 11.5: The Anova Table

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (or the tables values)*
Between columns (i.e., between varieties of seeds)	8	$(3 - 1) = 2$	$8/2 = 4$	$4/1 = 4$	$F(2, 6) = 5.14$
Between rows (i.e., between varieties of fertilizers)	18	$(4 - 1) = 3$	$18/3 = 6$	$6/1 = 6$	$F(3, 6) = 4.76$
Residual or error	6	$(3 - 1) \times (4 - 1) = 6$	$6/6=1$		
Total	32	$(3 \times 4) - 1 = 11$			

Illustration 3

Set up ANOVA table for the following information relating to three drugs testing to judge the effectiveness in reducing blood pressure for three different groups of people:

Amount of Blood Pressure Reduction in Millimeters of Mercury

	Drug		
	X	*Y*	*Z*
Group of People *A*	14	10	11
	15	9	11
B	12	7	10
	11	8	11
C	10	11	8
	11	11	7

Do the drugs act differently?

Are the different groups of people affected differently?

Is the interaction term significant?

Answer the above questions taking a significant level of 5%.

Solution: We first make all the required computations as shown below:

We can set up ANOVA table shown in Table 11.7 (Page 269).

Table 11.6: Computations for Two-way Anova (in design with repeated values)

Step (i) $T = 187$, $n = 18$, thus, the correction factor $= \dfrac{187 \times 187}{18} = 1942.72$

Step (ii) Total $SS = [(14)^2 + (15)^2 + (12)^2 + (11)^2 + (10)^2 + (11)^2 + (10)^2 + (9)^2 + (7)^2 + (8)^2 + (11)^2 + (11)^2 + (11)^2$

$$+ (11)^2 + (10)^2 + (11)^2 + (8)^2 + (7)^2] - \left[\frac{(187)^2}{18}\right]$$

$= (2019 - 1942.72)$

$= 76.28$

Step (iii) SS between columns (i.e., between drugs) $= \left[\dfrac{73 \times 73}{6} + \dfrac{56 \times 56}{6} + \dfrac{58 \times 58}{6}\right] - \left[\dfrac{(187)^2}{18}\right]$

$= 888.16 + 522.66 + 560.67 - 1942.72$

$= 28.77$

Step (iv) SS between rows (i.e., between people) $= \left[\dfrac{70 \times 70}{6} + \dfrac{59 \times 59}{6} + \dfrac{58 \times 58}{6}\right] - \left[\dfrac{(187)^2}{18}\right]$

$= 816.67 + 580.16 + 560.67 - 1942.72$

$= 14.78$

Step (v) SS within samples $= (14 - 14.5)^2 + (15 - 14.5)^2 + (10 - 9.5)^2 + (9 - 9.5)^2 + (11 - 11)^2 + (11 - 11)^2$

$+ (12 - 11.5)^2 + (11 - 11.5)^2 + (7 - 7.5)^2 + (8 - 7.5)^2$

$+ (10 - 10.5)^2 + (11 - 10.5)^2 + (10 - 10.5)^2 + (11 - 10.5)^2$

$+ (11 - 11)^2 + (11 - 11)^2 + (8 - 7.5)^2 + (7 - 7.5)^2$

$= 3.50$

Step (vi) SS for interaction variation $= 76.28 - [28.77 + 14.78 + 3.50]$

$= 29.23$

Table 11.7: The Anova Table

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit*
Between columns (i.e., between drugs)	28.77	$(3 - 1) = 2$	$\dfrac{28.77}{2} = 14.385$	$\dfrac{14.385}{0.389} = 36.9$	$F(2, 9) = 4.26$
Between rows (i.e., between people)	14.78	$(3 - 1) = 2$	$\dfrac{14.78}{2} = 7.390$	$\dfrac{7.390}{0.389} = 19.0$	$F(2, 9) = 4.26$

Contd.

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit*
Interaction	$\boxed{29.23^*}$	$\boxed{4^*}$	$\frac{29.23}{4}$	$\frac{7.308}{0.389}$	$F(4, 9) = 3.63$
Within samples (Error)	3.50	(18 – 9) = 9	$\frac{3.50}{9}$ = 0.389		
Total	76.28	(18 – 1) = 17			

*These figures are left-over figures and have been obtained by subtracting from the column total the total of all other value in the said column. Thus, interaction *SS* = (76.28) – (28.77 + 14.78 + 3.50) = 29.23 and interaction degrees of freedom = (17) – (2 + 2 + 9) = 4.

The above table shows that all the three *F*-ratios are significant of 5% level which means that the drugs act differently, different groups of people are affected differently and the interaction term is significant. In fact, if the interaction term happens to be significant, it is pointless to talk about the differences between various treatments i.e., differences between drugs or differences between groups of people in the given case.

Graphic method of studying interaction in a two-way design: Interaction can be studied in a two-way design with repeated measurements through graphic method also. For such a graph we shall select one of the factors to be used as the *X*-axis. Then we plot the averages for all the samples on the graph and connect the averages for each variety of the other factor by a distinct mark (or a coloured line). If the connecting lines do not cross over each other, then the graph indicates that there is no interaction, but if the lines do cross, they indicate definite interaction or inter-relation between the two factors. Let us draw such a graph for the data of illustration 3 of this chapter to see whether there is any interaction between the two factors viz., the drugs and the groups of people.

Graph of the averages for amount of blood pressure reduction in millimeters of mercury for different drugs and different groups of people.*

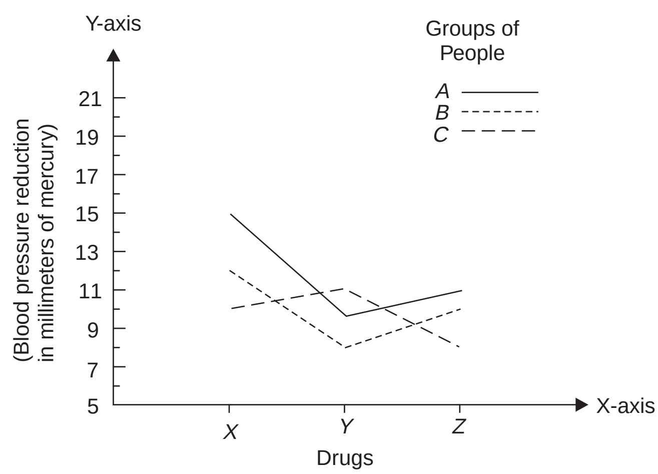

Fig. 11.1

*Alternatively, the graph can be drawn by taking different group of people on *X*-axis and drawing lines for various drugs through the averages.

The graph indicates that there is a significant interaction because the different connecting lines for groups of people do cross over each other. We find that *A* and *B* are affected very similarly, but *C* is affected differently. The highest reduction in blood pressure in case of *C* is with drug *Y* and the lowest reduction is with drug *Z*, whereas the highest reduction in blood pressure in case of *A* and *B* is with drug *X* and the lowest reduction is with drug *Y*. Thus, there is definite inter-relation between the drugs and the groups of people and one cannot make any strong statements about drugs unless he also qualifies his conclusions by stating which group of people he is dealing with. In such a situation, performing *F*-tests is meaningless. But if the lines do not cross over each other (and remain more or less identical), then there is no interaction or the interaction is not considered a significantly large value, in which case the researcher should proceed to test the main effects, drugs and people in the given case, as stated earlier.

ANOVA IN LATIN-SQUARE DESIGN

Latin-square design is an experimental design used frequently in agricultural research. In such a design the treatments are so allocated among the plots that no treatment occurs, more than once in any one row or any one column. The ANOVA technique in case of Latin-square design remains more or less the same as we have already stated in case of a two-way design, excepting the fact that the variance is splitted into four parts as under:

(i) variance between columns;
(ii) variance between rows;
(iii) variance between varieties;
(iv) residual variance.

All these above stated variances are worked out as under:

Table 11.8

Variance between columns or *MS* between columns	$= \dfrac{\sum \dfrac{(T_j)^2}{n_j} - \dfrac{(T)^2}{n}}{(c-1)} = \dfrac{SS \text{ between columns}}{\text{d.f.}}$
Variance between rows or *MS* between rows	$= \dfrac{\sum \dfrac{(T_i)^2}{n_i} - \dfrac{(T)^2}{n}}{(r-1)} = \dfrac{SS \text{ between rows}}{\text{d.f.}}$
Variance between varieties or *MS* between varieties	$= \dfrac{\sum \dfrac{(T_v)^2}{n_v} - \dfrac{(T)^2}{n}}{(v-1)} = \dfrac{SS \text{ between varieties}}{\text{d.f.}}$
Residual or error variance or *MS* residual	$= \dfrac{\text{Total } SS - (SS \text{ between columns} + SS \text{ between rows} + SS \text{ between varieties})}{(c-1)(c-2)^{*}}$

Contd.

* In place of *c* we can as well write *r* or *v* since in Latin-square design $c = r = v$.

where total $SS = \Sigma(x_{ij})^2 - \frac{(T)^2}{n}$

c = number of columns

r = number of rows

v = number of varieties

Illustration 4

Analyse and interpret the following statistics concerning output of wheat per field obtained as a result of experiment conducted to test four varieties of wheat viz., A, B, C and D under a Latin-square design.

C 25	B 23	A 20	D 20
A 19	D 19	C 21	B 18
B 19	A 14	D 17	C 20
D 17	C 20	B 21	A 15

Solution: Using the coding method, we subtract 20 from the figures given in each of the small squares and obtain the coded figures as under:

		Columns				Row totals
		1	2	3	4	
	1	C 5	B 3	A 0	D 0	8
Rows	2	A –1	D –1	C 1	B –2	–2
	3	B –1	A –6	D –3	C 0	–10
	4	D –3	C 0	B 1	A –5	–7
Column totals		0	–4	–1	–7	T = –12

Fig. 11.2 (a)

Squaring these coded figures in various columns and rows we have:

		Squares of coded figures				Sum of squares
		Columns				
		1	2	3	4	
	1	*C* 25	*B* 9	*A* 0	*D* 0	34
Rows	2	*A* 1	*D* 1	*C* 1	*B* 4	7
	3	*B* 1	*A* 36	*D* 9	*C* 0	46
	4	*D* 9	*C* 0	*B* 1	*A* 25	35
Sum of squares		36	46	11	29	*T* = 122

Fig. 11.2 (b)

$$\text{Correction factor} = \frac{(T)^2}{n} = \frac{(-12)\,(-12)}{16} = 9$$

$$SS \text{ for total variance} = \Sigma\left(X_{ij}\right)^2 - \frac{(T)^2}{n} = 122 - 9 = 113$$

$$SS \text{ for variance between columns} = \Sigma\frac{\left(T_j\right)^2}{n_j} - \frac{(T)^2}{n}$$

$$= \left\{\frac{(0)^2}{4} + \frac{(-4)^2}{4} + \frac{(-1)^2}{4} + \frac{(-7)^2}{4}\right\} - 9$$

$$= \frac{66}{4} - 9 = 7.5$$

SS for variance between rows

$$= \Sigma\frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}\left\{\frac{(8)^2}{4} + \frac{(-3)^2}{4} + \frac{(-10)^2}{4} + \frac{(-7)^2}{4}\right\} - 9$$

$$= \frac{222}{4} - 9 = 46.5$$

SS for variance between varieties would be worked out as under:

For finding *SS* for variance between varieties, we would first rearrange the coded data in the following form:

Table 11. 9

Varieties of wheat	*Yield in different parts of field*				*Total* (T_v)
	I	II	III	IV	
A	–1	–6	0	–5	–12
B	–1	3	1	–2	1
C	5	0	1	0	6
D	–3	–1	–3	0	–7

Now we can work out *SS* for variance between varieties as under:

$$SS \text{ for variance between varieties } = \sum \frac{(T_v)^2}{n_v} - \frac{(T)^2}{n}$$

$$= \left\{ \frac{(-12)^2}{4} + \frac{(1)^2}{4} + \frac{(6)^2}{4} + \frac{(-7)^2}{4} \right\} - 9$$

$$= \frac{230}{4} - 9 = 48.5$$

∴ Sum of squares for residual variance will work out to

$$113 - (7.5 + 46.5 + 48.5) = 10.50$$

d.f. for variance between columns $= (c - 1) = (4 - 1) = 3$
d.f. for variance between rows $= (r - 1) = (4 - 1) = 3$
d.f. for variance between varieties $= (v - 1) = (4 - 1) = 3$
d.f. for total variance $= (n - 1) = (16 - 1) = 15$
d.f. for residual variance $= (c - 1)\,(c - 2) = (4 - 1)\,(4 - 2) = 6$

ANOVA table can now be set up as shown below:

Table 11. 10: The Anova Table in Latin-square Design

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit*
Between columns	7.50	3	$\frac{7.50}{3} = 2.50$	$\frac{2.50}{1.75} = 1.43$	$F\,(3, 6) = 4.76$
Between rows	46.50	3	$\frac{46.50}{3} = 15.50$	$\frac{15.50}{1.75} = 8.85$	$F\,(3, 6) = 4.76$

contd.

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit*
Between varieties	48.50	3	$\frac{48.50}{3} = 16.17$	$\frac{16.17}{1.75} = 9.24$	$F(3, 6) = 4.76$
Residual or error	10.50	6	$\frac{10.50}{6} = 1.75$		
Total	113.00	15			

The above table shows that variance between rows and variance between varieties are significant and not due to chance factor at 5% level of significance as the calculated values of the said two variances are 8.85 and 9.24 respectively which are greater than the table value of 4.76. But variance between columns is insignificant and is due to chance because the calculated value of 1.43 is less than the table value of 4.76.

ANALYSIS OF CO-VARIANCE (ANOCOVA)
WHY ANOCOVA?

The object of experimental design in general happens to be to ensure that the results observed may be attributed to the treatment variable and to no other causal circumstances. For instance, the researcher studying one independent variable, X, may wish to control the influence of some uncontrolled variable (sometimes called the covariate or the concomitant variables), Z, which is known to be correlated with the dependent variable, Y, then he should use the technique of analysis of covariance for a valid evaluation of the outcome of the experiment. "In psychology and education primary interest in the analysis of covariance rests in its use as a procedure for the statistical control of an uncontrolled variable."[2]

ANOCOVA TECHNIQUE

While applying the ANOCOVA technique, the influence of uncontrolled variable is usually removed by simple linear regression method and the residual sums of squares are used to provide variance estimates which in turn are used to make tests of significance. In other words, covariance analysis consists in subtracting from each individual score (Y_i) that portion of it Y_i' that is predictable from uncontrolled variable (Z_i) and then computing the usual analysis of variance on the resulting $(Y - Y')$'s, of course making the due adjustment to the degrees of freedom because of the fact that estimation using regression method required loss of degrees of freedom.*

[2] George A-Ferguson, *Statistical Analysis in Psychology and Education*, 4th ed., p. 347.

* Degrees of freedom associated with adjusted sums of squares will be as under:

Between	$k - 1$
within	$N - k - 1$
Total	$N - 2$

ASSUMPTIONS IN ANOCOVA

The ANOCOVA technique requires one to assume that there is some sort of relationship between the dependent variable and the uncontrolled variable. We also assume that this form of relationship is the same in the various treatment groups. Other assumptions are:

(i) Various treatment groups are selected at random from the population.

(ii) The groups are homogeneous in variability.

(iii) The regression is linear and is same from group to group.

The *short-cut method for ANOCOVA* can be explained by means of an example as shown below:

Illustration 5

The following are paired observations for three experimental groups:

Group I		*Group II*		*Group III*	
X	*Y*	*X*	*Y*	*X*	*Y*
7	2	15	8	30	15
6	5	24	12	35	16
9	7	25	15	32	20
15	9	19	18	38	24
12	10	31	19	40	30

Y is the covariate (or concomitant) variable. Calculate the adjusted total, within groups and between groups, sums of squares on *X* and test the significance of differences between the adjusted means on *X* by using the appropriate *F*-ratio. Also calculate the adjusted means on *X*.

Solution: We apply the technique of analysis of covariance and work out the related measures as under:

Table 11.11

	Group I		*Group II*		*Group III*	
	X	*Y*	*X*	*Y*	*X*	*Y*
	7	2	15	8	30	15
	6	5	24	12	35	16
	9	7	25	15	32	20
	15	9	19	18	38	24
	12	10	31	19	40	30
Total	49	33	114	72	175	105
Mean	9.80	6.60	22.80	14.40	35.00	21.00

$$\sum X = 49 + 114 + 175 = 338$$