Chapter 5

ANOVA (Analysis of Variance)

Introduction

Analysis of Variance (ANOVA) is a statistical method used to determine whether several population means can be considered equal. Unlike procedures limited to two samples, ANOVA allows the comparison of three or more means by examining how data vary both *within* groups and *between* groups. The technique relies on the F-distribution, which arises from the ratio of two variance estimates.

The hypotheses tested in ANOVA are :

$$H_0 : \quad \mu_1 = \mu_2 = \cdots = \mu_k,$$
$$H_1 : \quad \text{At least one mean differs from the others.}$$

The purpose of the ANOVA statistic is twofold :

1. Determine whether several population means are identical.
2. Compare the variability within samples (unexplained variation) with the variability between sample means (explained variation). The between-group variation estimates σ^2 only when H_0 is true.

If the null hypothesis is false, the between-group variation becomes noticeably larger than the within-group variation. By analyzing this difference through the F-statistic, we can decide whether H_0 should be rejected.

ANOVA is efficient because it tests all groups simultaneously, avoiding the increase in error risk that occurs when performing multiple two-sample tests. The fundamental idea is to decompose the total variation of the data into :

— variation explained by differences between groups, and
— random variation within groups.

The measure of variability used in ANOVA is the *mean square*, defined by :

$$\text{Mean Square} = \frac{\text{Sum of squared deviations from the mean}}{\text{Degrees of freedom}}.$$

The ANOVA table summarizes these quantities and forms the basis for the F-test.

Assumptions

ANOVA requires the following conditions to be satisfied :

1. Samples are drawn independently.
2. Each population follows a normal distribution with identical variance.
3. Observations within each group occur randomly and independently.
4. The effects of different factors combine additively.
5. Population variances are equal (homoscedasticity).

5.1 One-Way ANOVA

One-way ANOVA, also known as single-factor ANOVA, is used when the study involves only one independent variable (or factor) with several groups or levels. The goal is to determine whether the population means corresponding to these groups are all equal.

This method compares the sample means from the different groups to draw conclusions about the underlying population means. Because there is only one factor influencing the grouping, the procedure is called "one-way."

The validity of the F-statistic obtained from one-way ANOVA relies on several assumptions :

— Each group consists of observations that are normally distributed.

— The groups may have different true means (although the null hypothesis assumes they are identical), but they must share a common population standard deviation.

The normality requirement is similar to that found in many significance tests, but here it is applied simultaneously to multiple groups, making it slightly more restrictive. Since real data do not always follow the normal curve perfectly, this assumption should be considered carefully before applying ANOVA.

Allowing populations to have different means is not a limitation, because the null hypothesis of equal means is only used to compute the F-statistic.

The assumption of equal standard deviations (homogeneity of variance) is essential. Although it cannot be verified directly, a practical rule is to compare the sample standard deviations from each group. Statisticians generally consider ANOVA reliable if the largest sample standard deviation does not exceed twice the smallest one.

A completely randomized design is used when we want to compare the means of two or more independent groups. In this setting, there is one categorical factor (with k groups) and one numerical response variable. Below are the steps for performing a one-way ANOVA in a simplified and practical manner.

Step 1 : State the hypotheses.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \qquad H_1 : \text{At least one population mean is different.}$$

Step 2 : Compute the total of all observations.

If the samples contain $X_1, X_2, \dots, X_k$ as group sums, then the overall total is

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij},$$

Step 3 : Compute the correction factor.

Let

$$N = n_1 + n_2 + \cdots + n_k,$$

then the correction factor is

$$CF = \frac{T^2}{N}.$$

Step 4 : Compute the total sum of squares (SST).

This measures the total variability in the data :

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - CF.$$

Step 5 : Compute the sum of squares between groups (SSB).

This represents the variation explained by the differences between group means :

$$SSB = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i} - CF.$$

Step 6 : Compute the sum of squares within groups (SSW).

This measures the unexplained (random) variation :

$$\mathrm{SSW} = \mathrm{SST} - \mathrm{SSB}.$$

Step 7 : Determine the degrees of freedom.

$$v_1 = k - 1, \qquad v_2 = N - k.$$

Step 8 : Compute the mean squares.

$$\mathrm{MSB} = \frac{\mathrm{SSB}}{v_1}, \qquad \mathrm{MSW} = \frac{\mathrm{SSW}}{v_2}.$$

Step 9 : Calculate the F-statistic.

$$F = \frac{\mathrm{MSB}}{\mathrm{MSW}}.$$

Step 10 : Make a decision.

Compare the computed F value with the critical value from the F-table with (v_1, v_2) degrees of freedom :

- If F is **less** than the table value → fail to reject H_0. - If F is **greater** than the table value → reject H_0.

Thus, ANOVA allows us to determine whether the group means differ significantly.

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Between groups	SSB	$df_1 = k - 1$	$MSB = \frac{SSB}{df_1}$	$\frac{MSB}{MSW}$
Within groups	SSW	$df_2 = N - k$	$MSW = \frac{SSW}{df_2}$	
Total	SST	$N - 1$		

Example 5.1.1. *Three wheat varieties (A, B, and C) were planted on four plots each. The yields (in tonnes per acre) were recorded as follows :*

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

We want to test whether the three varieties differ in their average yields. The table value of F at the 5% *level for* $(2, 9)$ *degrees of freedom is* 4.26.

Step 1 : State the hypotheses

$$H_0 : \mu_A = \mu_B = \mu_C$$

$$H_1 : \textit{At least one mean yield is different.}$$

Step 2 : Compute totals and sum of squares

Each sample contains 4 observations. Thus,

$$n_1 = n_2 = n_3 = 4, \qquad N = 12, \qquad k = 3.$$

***Sample totals** :*

$$T_A = 25, \qquad T_B = 20, \qquad T_C = 15.$$

***Sum of squares of individual values** :*

$$SS_A = 165, \qquad SS_B = 108, \qquad SS_C = 61.$$

Total of all observations :

$$T = T_A + T_B + T_C = 60.$$

Correction factor :

$$CF = \frac{T^2}{N} = \frac{3600}{12} = 300.$$

Step 3 : Total sum of squares (SST)

$$SST = (SS_A + SS_B + SS_C) - CF$$

$$SST = 334 - 300 = 34.$$

Step 4 : Sum of squares between groups (SSB)

Using group totals :

$$SSB = \left(\frac{T_A^2}{n_1} + \frac{T_B^2}{n_2} + \frac{T_C^2}{n_3}\right) - CF$$

$$SSB = \left(\frac{25^2}{4} + \frac{20^2}{4} + \frac{15^2}{4}\right) - 300$$

$$SSB = (156.25 + 100 + 56.25) - 300 = 12.5.$$

Step 5 : Sum of squares within groups (SSW)

$$SSW = SST - SSB = 34 - 12.5 = 21.5.$$

Step 6 : Degrees of freedom

$$df_1 = k - 1 = 2, \qquad df_2 = N - k = 9.$$

Step 7 : Mean squares

$$MSB = \frac{SSB}{df_1} = \frac{12.5}{2} = 6.25.$$

$$MSW = \frac{SSW}{df_2} = \frac{21.5}{9} \approx 2.39.$$

Step 8 : F–statistic

$$F = \frac{MSB}{MSW} = \frac{6.25}{2.39} \approx 2.615.$$

ANOVA Table

Source	***Sum of Squares***	***df***	***Mean Square***	***F***
Between varieties	*12.5*	*2*	*6.25*	*2.615*
Within varieties	*21.5*	*9*	*2.39*	
Total	*34.0*	*11*		

Conclusion

Since the calculated value,

$$F_{calculated} = 2.615,$$

is less than the table value,

$$F_{table}(2, 9) = 4.26,$$

we fail to reject the null hypothesis. Thus, at the 5% level of significance, there is no sufficient statistical evidence to conclude that the mean yields of the three wheat varieties are different.

5.2 Two-Way ANOVA

A one-way analysis of variance compares several populations with respect to a single factor. A two-way ANOVA extends this idea by studying the influence of *two* categorical factors on one numerical (dependent) variable. The name "two-way" comes from the fact that every observation is classified according to two criteria rather than one.

This technique is especially useful when we want to examine :

— how each factor affects the response variable, and

— whether the two factors interact with each other.

Compared to the one-way ANOVA, this design is often more efficient because it separates the total variability into more meaningful sources. To perform a two-way ANOVA properly, the design must include the same number of observations in each cell of the factor-level combinations.

Assumptions

The two-way ANOVA relies on the same basic assumptions as the one-way ANOVA :

- The populations corresponding to each cell are normally distributed.
- All observations are independent of one another.
- The population variances are equal across all cells.
- Each factor–level combination contains the same number of observations.

Hypotheses

A two-way ANOVA evaluates three separate questions. The corresponding null hypotheses are :

1. The mean responses for the levels of **Factor 1** are equal. (This is analogous to a one-way ANOVA applied to the row factor.)
2. The mean responses for the levels of **Factor 2** are equal. (This mirrors a one-way ANOVA for the column factor.)
3. There is **no interaction** between the two factors. In other words, the effect of one factor does not depend on the level of the other factor.

Each factor must have at least two levels, and the degrees of freedom for a factor equal the number of its levels minus one.

Working Rule for Two–Way ANOVA

Consider a two–way ANOVA table with :
— r rows (first factor),
— c columns (second factor),
— one observation in each cell : x_{ij} for row i and column j.
The total number of observations is

$$N = rc.$$

The following steps describe a simplified procedure for carrying out an ANOVA calculation using the shortcut method : **Step 1 : State the hypotheses**
— H_0 (Row factor) : All row means are equal.
— H_0 (Column factor) : All column means are equal.
— H_0 (Interaction) : There is no interaction between row and column factors.

Step 2 : Compute basic totals Row sums :

$$R_i = \sum_{j=1}^{c} x_{ij}, \qquad i = 1, \ldots, r$$

Column sums :

$$C_j = \sum_{i=1}^{r} x_{ij}, \qquad j = 1, \ldots, c$$

Grand total :

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij}.$$

Step 3 : Correction factor

$$\text{CF} = \frac{T^2}{N}.$$

Step 4 : Total sum of squares (SST)

$$\text{SST} = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij}^2 - \text{CF}.$$

Step 5 : Sum of squares for columns (SSC)

$$\text{SSC} = \left(\sum_{j=1}^{c} \frac{C_j^2}{r} \right) - \text{CF}.$$

Step 6 : Sum of squares for rows (SSR)

$$\text{SSR} = \left(\sum_{i=1}^{r} \frac{R_i^2}{c} \right) - \text{CF}.$$

Step 7 : Error sum of squares (SSE)

$$\text{SSE} = \text{SST} - \text{SSC} - \text{SSR}.$$

Step 8 : Degrees of freedom

$$df_{\text{columns}} = c - 1, \qquad df_{\text{rows}} = r - 1, \qquad df_{\text{error}} = (r-1)(c-1).$$

Step 9 : Compute mean squares

$$MSC = \frac{\text{SSC}}{c-1}, \qquad MSR = \frac{\text{SSR}}{r-1}, \qquad MSE = \frac{\text{SSE}}{(r-1)(c-1)}.$$

Step 10 : Compute the F ratios

$$F_C = \frac{MSC}{MSE}, \qquad F_R = \frac{MSR}{MSE}.$$

ANOVA Summary Table

Source of Variation	Sum of Squares	df	Mean Square	F
Columns	SSC	$c-1$	MSC	$F_C = \dfrac{MSC}{MSE}$
Rows	SSR	$r-1$	MSR	$F_R = \dfrac{MSR}{MSE}$
Error	SSE	$(r-1)(c-1)$	MSE	
Total	SST	$rc-1$		

Example 5.2.1. *We consider the following table*

Plot of land	*A*	*B*	*C*	*D*
I	3	4	6	6
II	6	4	5	3
III	6	6	4	7

(For $(3,6)$ *df,* $F_{0.05} = 4.76$ *and for* $(2,6)$ *df,* $F_{0.05} = 5.14$*)*

Solution

Null hypotheses

1. H_{0C} : All four wheat varieties have the same mean yield.
2. H_{0R} : All three plots of land have the same mean yield.

Step 1 : Compute totals

Column totals (varieties) :

$$C_1 = 3+6+6 = 15, \qquad C_2 = 4+4+6 = 14,$$
$$C_3 = 6+5+4 = 15, \qquad C_4 = 6+3+7 = 16.$$

Row totals (plots) :

$R_1 = 3+4+6+6 = 19,\qquad R_2 = 6+4+5+3 = 18,\qquad R_3 = 6+6+4+7 = 23.$

Grand total :

$$T = \sum_{i=1}^{3}\sum_{j=1}^{4} x_{ij} = 60.$$

Total number of observations :

$$N = 12.$$

Correction factor :

$$\text{CF} = \frac{T^2}{N} = \frac{60^2}{12} = 300.$$

Step 2 : Total sum of squares

$$\text{SST} = \sum_{i=1}^{3}\sum_{j=1}^{4} x_{ij}^2 - \text{CF} = 320 - 300 = 20.$$

Step 3 : Sum of squares for columns (varieties)

$$\text{SSC} = \sum_{j=1}^{4} \frac{C_j^2}{3} - \text{CF} = \left[\frac{15^2}{3} + \frac{14^2}{3} + \frac{15^2}{3} + \frac{16^2}{3}\right] - 300.$$

$$\text{SSC} = \frac{902}{3} - 300 = 300.67 - 300 = 0.67.$$

Step 4 : Sum of squares for rows (plots)

$$\text{SSR} = \sum_{i=1}^{3} \frac{R_i^2}{4} - \text{CF} = \left[\frac{19^2}{4} + \frac{18^2}{4} + \frac{23^2}{4}\right] - 300.$$

$$\text{SSR} = 3.50.$$

Step 5 : Error sum of squares

$$\text{SSE} = \text{SST} - \text{SSC} - \text{SSR} = 20 - (0.67 + 3.50) = 15.83.$$

Step 6 : Degrees of freedom

$$df_C = 4 - 1 = 3, \qquad df_R = 3 - 1 = 2, \qquad df_E = (4-1)(3-1) = 6.$$

Step 7 : Mean squares

$$MSC = \frac{0.67}{3} = 0.223, \qquad MSR = \frac{3.50}{2} = 1.75, \qquad MSE = \frac{15.83}{6} = 2.930.$$

Step 8 : F–ratios

$$F_C = \frac{MSC}{MSE} = \frac{0.223}{2.930} = 0.076,$$
$$F_R = \frac{MSR}{MSE} = \frac{1.75}{2.930} = 0.597.$$

ANOVA Table

Source	SS	df	MS	F
Columns (varieties)	0.67	3	0.223	0.076
Rows (plots)	3.50	2	1.75	0.597
Error	15.83	6	2.930	
Total	20	11		

Conclusion

Since both calculated values

$$F_C = 0.076 < 4.76, \qquad F_R = 0.597 < 5.14,$$

we **fail to reject** both null hypotheses.
— The four wheat varieties do not differ significantly in yield.
— The three plots of land do not differ significantly in yield.