• MODULE→5

Analysis of Variance (ANOVA)

In the simplest form, the **ANOVA** can be viewed as a hypothesis test of group means.

* **Steps in ANOVA:**

Step→1: Formulation of hypothesis — If a problem involves k groups, the following hypotheses are appropriate for comparing k group means:

$H_0: m_1 = m_2 = \cdots = m_k$

H_1: at least one pair of group means are not equal

The test compares the means, but if the null hypothesis is rejected, the following five steps do not identify which pair or pairs of means are not equal; this problem is covered later. The group means are expressed as population values, not sample values.

Step→2: Define the test statistic and its distribution — The hypotheses of Step-1 can be tested using the following test statistic:

$$F = \frac{MS_b}{MS_w} \quad \text{——} ①$$

in which MS_b and MS_w are mean squares between and within variations, respectively, and F is the value

of a random variable having an F distribution with degrees of freedom of $(k-1, N-k)$. The mean squares are computed as shown below:

* Computation Table for Analysis of Variance (ANOVA) Test

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Between	$SS_b = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$	$k-1$	$MS_b = \frac{SS_b}{k-1}$
Within	$SS_w = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$	$N-k$	$MS_w = \frac{SS_w}{N-k}$
Total	$SS_t = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$	$N-1$	—

Step→3: Specify the level of significance.

Step→4: Collect data and compute test statistic — The data should be collected and used to compute the value of the test statistic (F) of ①. The data are least portrayed as a matrix as shown below:

Data Matrix for the ANOVA:

Observations for Group				
1	2	3	------	k
X_{11}	X_{12}	X_{13}	------	X_{1k}
X_{21}	X_{22}	X_{23}	------	X_{2k}
X_{31}	X_{32}	X_{33}	------	X_{3k}
⋮	⋮	⋮	------	⋮
$X_{n_1 1}$	$X_{n_2 2}$	$X_{n_3 3}$	------	$X_{n_k k}$
Mean $\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	------	$\bar{X}_k$

There are k columns and the jth column contains n_j values. The total number of observations N is given by

$$N = \sum_{j=1}^{k} n_j$$

Step→5: Determine the critical value of the test statistic.

Step→6: Make a decision.

e.g 1. The students in a course on introductory statistics are divided into five groups, each of which is assigned a different textbook. The students attend the same lectures and are tested using the same homework assignments and examinations. The average final grade, from 0 to 100, for each of the five groups is computed. If the average for one group is considerably higher than that of the other four groups, can we conclude that the textbook used by this group is a better learning tool?

Grades for Group

1	2	3	4	5
82	67	91	66	82
75	79	82	73	71
87	77	76	89	67
76	81	79	84	76
Mean: $\bar{x}_1 = 80$	$\bar{x}_2 = 76$	$\bar{x}_3 = 82$	$\bar{x}_4 = 78$	$\bar{x}_5 = 74$

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_1: at least one pair of group means are not equal

* **ANOVA Table:**

$k = 5$. $n_1 = n_2 = n_3 = n_4 = n_5 = 4$ $N = \sum_{j=1}^{5} n_j = 20$.

$\bar{x}_1 = 80$, $\bar{x}_2 = 76$, $\bar{x}_3 = 82$, $\bar{x}_4 = 78$, $\bar{x}_5 = 75$.

$\therefore \bar{x} = 78$.

$\therefore SS_b = 4[(80-78)^2 + (76-78)^2 + (82-78)^2 + (78-78)^2 + (74-78)^2]$

$\Rightarrow SS_b = 4[4+4+16+0+16]$

$\Rightarrow \boxed{SS_b = 160}$ $\therefore \boxed{MS_b = \frac{SS_b}{k-1} = 40}$

$\therefore SS_w = [(82-80)^2 + (75-80)^2 + (87-80)^2 + (76-80)^2$
$+ (67-76)^2 + (79-76)^2 + (77-76)^2 + (81-76)^2$
$+ (91-82)^2 + (82-82)^2 + (76-82)^2 + (79-82)^2$
$+ (66-78)^2 + (73-78)^2 + (89-78)^2 + (84-78)^2 + (82-74)^2$
$+ (71-74)^2 + (67-74)^2 + (76-74)^2]$

$\Rightarrow SS_W = [4+25+49+16+81+9+1+25+81+0+36+$
$9+144+25+121+36+64+9+49+4]$

$\Rightarrow \boxed{SS_W = 788}$

$\therefore \boxed{MS_W = \frac{788}{15} = 52.53}$

Source of Variation	Sum of Square	Degrees of Freedom	Mean Square
Between	160	4	40
Within	788	15	52.53
Total	948	19	—

$\therefore F = \frac{40}{52.53} = 0.76 < 3.06$. Hence, H_0 is accepted.

$F_{0.05} = 3.06$

Though the group means are not equal, but the evidence is not enough to reject H_0. This means that the textbook is not responsible for the differences in the group means. It can be concluded that the difference are due to random factors.

2. Equality of Group Means:

Scores for Group

1	2	3
66	72	80
73	81	85
89	84	70
84	75	77
Means: $\bar{x}_1 = 78$	$\bar{x}_2 = 78$	$\bar{x}_3 = 78$

* **ANOVA Table:**

$n_1 = n_2 = n_3 = 4$ $\quad N = n_1 + n_2 + n_3 = 12$.

$k = 3$ $\quad \bar{x} = 78$.

$H_0: \bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{x}$

H_1: at least one pair of group means are not equal

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square
Between	$SS_b = 0$	2	$MS_b = 0$
Within	$SS_w = 534$	9	$MS_w = 59.33$
Total	$SS_t = 534$	11	—

$\therefore F = 0$ $\quad F_{0.05} = 4.26$.

$\therefore$ H_0 is accepted.

3. Inequality Group Means:

Scores for Group

1	2	3
78	82	74
78	82	74
78	82	74
78	82	74
Means: $\bar{X}_1 = 78$	$\bar{X}_2 = 82$	$\bar{X}_3 = 74$

* ANOVA Table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Between	$SS_b = 128$	2	$MS_b = 64$
Within	$SS_w = 0$	9	$MS_w = 0$
Total	$SS_t = 128$	11	—

$$\therefore F = \frac{MS_b}{MS_w} \to \infty$$

Hence, H_0 is rejected.

- **Computational Equations:**

$$SS_t = \left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2\right) - \frac{T^2}{N}$$

$$SS_w = \left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2\right) - \sum_{j=1}^{k}\frac{T_j^2}{n_j}$$

$$SS_b = \left(\sum_{j=1}^{k}\frac{T_j^2}{n_j}\right) - \frac{T^2}{N}$$

where T_j is the total of the scores in group j and T is the sum of the T_j values.

e.g. 1. A 20-acre field is divided into 20 1-acre plots that are randomly assigned to four groups, with five plots in each group. Plots within each group are treated with a specific mix of fertilizers. The crop yield from each plot is computed and given as follows:

Crop Yields (bushels/acre) for 20 Plots:

Plot	1	2	3	4	Total
1	75	103	67	87	
2	93	89	84	79	
3	82	97	73	86	
4	104	108	82	73	
5	96	98	74	80	
Mean	$\bar{x}_1 = 90$	$\bar{x}_2 = 98$	$\bar{x}_3 = 76$	$\bar{x}_4 = 81$	
S.D.	11.51	7.11	6.96	5.70	
T_j	450	495	380	405	1730
$\sum_{i=1}^{5} x_{ij}^2$	41030	49207	29074	32935	152246

Test whether the mix have effect on the crop yield.

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

H_1: at least one pair of means are not equal.

Source of Variation	Sum of Square	Degrees of freedom	Mean Square
Within	$SS_W = 152246 - \left(\frac{450^2 + 495^2 + 380^2 + 405^2}{5}\right)$ $= 1056$	16	$MS_W = \frac{SS_b}{16}$ $= 66$
Between	$SS_b = \left(\frac{450^2 + 495^2 + 380^2 + 405^2}{5}\right) - \frac{1730^2}{20}$ $= 1545$	3	$MS_b = \frac{1545}{3}$ $= 515$
Total	$SS_t = 2601$	19	—

$$F = \frac{515}{66} = 7.80 > 3.24. \quad F_{(3,16)\,0.05} = 3.24$$

$\therefore$ Ho is rejected

ANOVA 2. Model:

ANOVA 2 analysis involves three null hypotheses, one for for the each of the independent variables and one for the interaction between the two independent variables; both null hypothesis for the independent variables test for equality of means:

H_0: column means are not significantly different- $(m_1' = m_2' = \cdots = m_{nc}')$ —(I)

H_1: at least one pair of column means is different-

and

H_0: row means are not significantly different- $(m_1 = m_2 = \cdots = m_{nr})$ —(II)

H_1: at least one pair of row means is different

in which the subscripts nc and nr are the numbers of columns and rows, respectively. The hypothesis for testing for a significant interaction between the two independent variables are

H_0: the interaction between the independent variables is not significant- —(III)

H_1: the interaction between the independent variables is significant

The hypotheses of (III) must be tested before those of (I) and (II).

For ANOVA2 model, the sum of squares can be computed using the following equations:

1. The between-columns sum of squares:

$$SS_C = \sum_{K=1}^{nc} \frac{T_{\cdot K}^2}{n_K} - \frac{T^2}{N}$$

2. The between-rows sum of squares:

$$SS_r = \sum_{j=1}^{nr} n_j^\circ (\bar{x}_{j\cdot} - \bar{x})^2$$

$$= \sum_{j=1}^{nr} \frac{T_{j\cdot}^2}{n_j^\circ} - \frac{T^2}{N}$$

3. The between-cells sum of squares

$$SS_b = \sum_{j=1}^{k} \sum_{K=1}^{nc} \frac{T_{jK}^2}{n_{jK}^\circ} - \frac{T^2}{N}$$

4. The interaction sum of squares

$$SS_{rc} = SS_b - SS_c - SS_r$$

5. The total sum of squares

$$SS_t = \sum_{j=1}^{nr} \sum_{k=1}^{nc} \sum_{i=1}^{n_{jk}} X_{ijk}^2 - \frac{T^2}{N}$$

6. The within-cells sum of squares

$$SS_w = SS_t - SS_b$$

where $T_{j.}$ and $T_{.k}$ are the total scores in row j and column k, respectively; n_j and n_k are the number of scores in row j and column k, respectively; n_{jk} is the number of scores in cell jk; T is the total of all scores in the double-entry table; and N the total number of scores and is equal to

$$N = \sum_{j=1}^{nr} n_j = \sum_{k=1}^{nc} n_k = \sum_{j=1}^{nr} \sum_{k=1}^{nc} n_{jk}$$

* ANOVA 2 Table:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Between cells	—	SS_b	—	—
Between columns	$v_c = nc - 1$	SS_c	MS_c	$F_c = \frac{MS_c}{MS_w}$
Between rows	$v_r = nr - 1$	SS_r	MS_r	$F_r = \frac{MS_r}{MS_w}$
Interaction	$v_{rc} = (nc-1)(nr-1)$	SS_{rc}	MS_{rc}	$F_{rc} = \frac{MS_{rc}}{MS_w}$
Within cells	$v_w = N - nc(nr)$	SS_w	MS_w	—
Total	$v_t = N - 1$	SS_t	—	—

The decision making procedure is as follows:

1. Compute all entries to the ANOVA2 table.
2. Using the level of significance and degrees of freedom $(\upsilon_{rc}, \upsilon_w)$, find the critical F value (say, F_α).
3. If $F_{rc} > F_\alpha$, reject Ho of (III) and perform one-way ANOVA for the row and column variables.
4. If $F_{rc} < F_\alpha$, use F_c and F_r to test Ho of (I) and (II), respectively; use degrees of freedom (υ_c, υ_w) and (υ_r, υ_w) to test for significant column and row effects, respectively.

e.g. 1. A series of experiments is conducted in which the coefficient of friction is measured between a shaft and the bearing. Two factors that potentially affect the coefficient are the degree of machining of the bearing (highly polished, moderately polished, and no machining) and the percentage of lead and antimony in the shaft (high, moderate, low).

* Coefficient of Friction, with Two Measurements for Each Case of High, Medium, and Low:

Percentage of Lead and Antimony

Machining	High		Medium		Low		Row Means
Highly polished	0.0005	0.0011	0.0017	0.0019	0.0031	0.0031	0.0019
Moderately polished	0.0022	0.0022	0.0027	0.0031	0.0023	0.0025	0.0025
No machining	0.0039	0.0043	0.0036	0.0038	0.0033	0.0037	0.0038
Column Means:	0.0024		0.0028		0.0030		

* ANOVA 2 Table:

Source of Variation	D.F.	Sum of Squares	Mean Square	Computed F
Between cells	—	$SS_b = \sum_{j=1}^{k} \sum_{k=1}^{nc} \frac{T_{jk}^2}{n_{jk}} - \frac{T^2}{N}$ $\therefore N = 18,$ $\therefore T = \frac{0.00000256 + 0.00001296 + 0.00003844}{2} + \frac{0.00001936 + 0.00003364}{2} + \frac{0.00002304 + 0.00006724}{2} + \frac{0.00005476 + 0.00009409}{2} - \frac{0.002101}{18}$	—	—

Source of Variation	Sum of Square	Degrees of Freedom	Mean Square	Computed F
Between cells	$SS_b = 17.1111 \times 10^{-6}$	—	—	—
Between columns	$V_c = 3-1=2$	$SS_c = 1.2578 \times 10^{-6}$	$MS_c = 0.6289 \times 10^{-6}$	$F_c = 11.80$
Between rows	$V_r = 3-1=2$	$SS_r = 10.8978 \times 10^{-6}$	$MS_r = 5.4489 \times 10^{-6}$	$F_r = 102.23$
Interaction	$V_{rc} = 2.2 = 4$	$SS_{rc} = 4.9555 \times 10^{-6}$	$MS_{rc} = 1.2389 \times 10^{-6}$	$F_{rc} = 23.24$
Within	$V_w = 18-9=9$	$SS_w = 0.4800 \times 10^{-6}$	$MS_w = 0.0533 \times 10^{-6}$	—
Total	$V_t = 18-1=17$	$SS_t = 17.5911 \times 10^{-6}$	—	—

$(4, 9) : F_{0.05} = 3.63 \quad \therefore F_{rc} > F_{0.05}$

$\therefore$ H_0 is rejected, consequently, the interaction between Machining and Percentage of lead and antimony is significant.

Then we go for one-way ANOVA of row and column effect.

* One-Way ANOVA Table for Row Effect:

Source of Variation	D.F.	Sum of Squares	Mean Square	F
Between	2	10.8978×10^{-6}	5.4489×10^{-6}	12.21
Within	15	6.6933×10^{-6}	0.4462×10^{-6}	—
Total	17	17.5911×10^{-6}	—	—

$F = 12.21 > F_{0.05} = 3.68$

$\therefore$ H_0 is rejected.

Hence, row effect is significant. This indicates that the level of machining influences the coefficient of friction.

* One-Way ANOVA Table for Column Effect:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Computed F
Between	2	1.2578×10^{-6}	0.6289×10^{-6}	0.58
Within	15	16.3333×10^{-6}	1.0889×10^{-6}	—
Total	17	17.5911×10^{-6}	—	—

$F = 0.58 < 3.68$.

$\therefore$ H_0 is accepted.

Hence, the percentage of lead and antimony in the shaft does not influence the coefficient of friction as a separate variable; it does, however, interact with the degree of machining to affect the coefficient of friction.