MODULE-5

ANOVA

Anova: For $k > 2$, if there are k samples from k populations, to test the population means, one common procedure is called Analysis-of- variance or Anova.

One- way Anova: Random samples of size n are selected from each of k populations, and k -populations are independent and normally distributed with means $\mu_1, \ \mu_2, \ \mu_3, \ \cdots \mu_k$ and common variance σ^2.

One- way Anova methods for testing the hypothesis

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

H_1: At least two of the means are not equal,

is given below.

Let the grand mean is $\mu = \frac{\sum \mu_i}{k}$.

Population number	1	2	$\cdots$	i	$\cdots$	k	
	x_{11}	x_{21}	$\cdots$	x_{i1}	$\cdots$	x_{k1}	
	x_{12}	x_{22}	$\cdots$	x_{i2}	$\cdots$	x_{k2}	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	x_{1n}	x_{2n}	$\cdots$	x_{in}	$\cdots$	x_{kn}	
Total	$\sum x_{1j}$	$\sum x_{2j}$	$\cdots$	$\sum x_{ij}$		$\sum x_{kj}$	$T = \sum x$
Mean	$\overline{x_1} = \frac{\sum x_{1j}}{n}$	$\overline{x_2} = \frac{\sum x_{1j}}{n}$		$\overline{x_i} = \frac{\sum x_{ij}}{n}$		$\overline{x_k} = \frac{\sum x_{kj}}{n}$	$\overline{x} = \frac{T}{N}$
Total sum of squares $= SST = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ij} - \overline{x}\right)^2$							
Population sums of squares $= SSA = n \sum_{i=1}^{k} (\overline{x_i} - \overline{x})^2$							
Error sum of squares $= SSE = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ij} - \overline{x_i}\right)^2$							

Clearly, $SST = SSA + SSE$. Or $SSE = SST - SSA$

$S_1^2 = \frac{SSA}{k-1}$ and $S^2 = \frac{SSE}{k(n-1)} = \frac{SSE}{N-k}$.

$F = \frac{S_1^2}{S^2}$.

The null hypothesis H_0 is rejected at the α-level of significance, if $F > F_{\propto}[k-1, \ N-k]$ otherwise accept.

Examples:

1. An experimenter wished to study the effect of four fertilizers on the yield of crop. He divided the field into 24 plots and assigned each fertilizer at random of 6 plots. Part of his calculations are given bellow.

Source	d.f.	Sum of Squares	Mean S	F	$F_{0.05}[3, \ 20]$ $= 3.10$
Fertilizers	–	2940	–	–	
Within the group	–	–	–		
Total	–	6212			

Complete the table and test at 5% level to see whether the fertilizers differ significantly?

Solution:

Anova table:

Source	d.f.	Sum of Squares	Mean S	F	$F_{0.05}[3,\ 20]$
Fertilizers	$k-1=3$	SSA =2940	$S_1^2 = 980$	$F = \frac{S_1^2}{S^2}$	$= 3.10$
Within the group	$N-k=20$	$SST - SSA = 3272$	$S^2 = 163.6$	$= 5.99$	$F > F_{0.05}$
Total	23	SST =6212	Conclusion: Reject the null hypothesis.		

Therefore fertilizers differ significantly.

2. Test the hypothesis $\mu_1 = \mu_2 = \cdots = \mu_5$ at the 0.05 level of significance for the data given below on absorption of moisture by various types of cement aggregates.

Aggregate	1	2	3	4	5
	551	595	639	417	563
	457	580	615	449	631
	450	508	511	517	522
	731	583	573	438	613
	499	633	648	415	656
	632	517	677	555	679

Solution: Given that, $k = 5,\ n = 6$.

Aggregate	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.33	569.33	610.50	465.17	610.67	561.80

Population sums of squares $= SSA = n\sum_{i=1}^{k}(\overline{x_i} - \overline{x})^2$

$$= 6(71.7409 + 56.7009 + 2371.69 + 9337.3569 + 2388.2769)$$
$$= 85354.5936$$

Total sum of squares $= SST = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \overline{x}\right)^2$

$$= 116.64 + 1102.24 + 5959.84 + 20967.04 + 1.44$$
$$+10983.04 + 331.24 + 2830.24 + 12723.84 + 4788.64$$
$$+12499.24 + 2894.44 + 2580.64 + 2007.04 + 1584.04$$
$$+28628.64 + 449.44 + 125.44 + 15326.44 + 2621.44$$
$$+3943.84 + 5069.44 + 7430.44 + 21550.24 + 8873.64$$
$$+4928.04 + 2007.04 + 13271.04 + 46.24 + 13735.84$$
$$= 209377.16.$$

$$SSE = SST - SSA = 209377.16 - 85354.5936 = 124022.5664$$

$S_1^2 = \frac{SSA}{k-1} = \frac{85354.5936}{4} = 21338.6484$ and $S^2 = \frac{SSE}{k(n-1)} = \frac{124022.5664}{25} = 4960.9027$.

$$F = \frac{S_1^2}{S^2} = \frac{21338.6484}{4960.9027} = 4.3014 \ .$$

Anova table:

Source	d.f.	Sum of Squares	Mean S	F	$F_{0.05}[4,\ 25]$
Aggregate	$k-1=4$	$SSA = 85354.5936$	$S_1^2 = 21338.6484$	$F = \frac{S_1^2}{S^2} = 4.3014$	$= 2.76$
Within the group	$N-k=25$	$SSE = 124022.5664$	$S^2 = 4960.9027$		$F > F_{0.05}$
Total	29	$SST = 209377.16$	Conclusion: Reject the null hypothesis.		

Clearly $F > F_{0.05}[4,\ 25]$, since $F_{0.05}[4,\ 25] = 2.76$.

Hence, reject the hypothesis $\mu_1 = \mu_2 = \cdots = \mu_5$, and conclude that the aggregates do not have the same mean absorption.

Shortcut method:

Aggregate	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854

Correction factor$= \frac{T^2}{N} = \frac{16{,}854^2}{30} = 9468577.2$

Total sum of squares $= SST = \sum X^2 - \frac{T^2}{N}$

$$= 1897736 + 1956356 + 2254229 + 1314873 + 2254760 - 9468577.2 = 209376.8$$

Population sums of squares $= SSA = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} + \frac{(\sum X_4)^2}{n_4} + \frac{(\sum X_5)^2}{n_5} - \frac{T^2}{N}$

$= 9553933.667 - 9468577.2$

$= 85356.46667$

$SSE = SST - SSA = 209376.8 - 85356.46667 = 124020.3333$

$S_1^2 = \frac{SSA}{k-1} = \frac{85356.46667}{4} = 21339.11667$ and $S^2 = \frac{SSE}{k(n-1)} = \frac{124020.3333}{25} = 4960.81333$.

$$F = \frac{S_1^2}{S^2} = \frac{21339.11667}{4960.81333} = 4.301536 \ .$$

Anova table:

Source	d.f.	Sum of Squares	Mean S	F	$F_{0.05}[4, 25]$
Aggregate	$k-1=4$	$SSA = 85356.46667$	$S_1^2 = 21339.11667$	$F = \frac{S_1^2}{S^2} = 4.301536$	$= 2.76$
Within the group	$N-k=25$	$SSE = 124020.3333$	$S^2 = 4960.81333$		$F > F_{0.05}$
Total	29	$SST = 209376.8$	Conclusion: Reject the null hypothesis.		

Clearly $F > F_{0.05}[4, 25]$, since $F_{0.05}[4, 25] = 2.76$.

Hence, reject the hypothesis $\mu_1 = \mu_2 = \cdots = \mu_5$, and conclude that the aggregates do not have the same mean absorption.

3. The verities of wheat A, B, C were shown in four plots and yields in quintals per acre is as follows.

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

Test the significance difference between the yields of verities given that $F_{0.05}[2, 9] = 4.26$.

verities of wheat	A	B	C	
	8	7	2	
	4	6	5	
	6	5	4	
	7	3	4	
Total	25	21	15	61

Correction factor$= \frac{T^2}{N} = \frac{61^2}{12}$

Total sum of squares $= SST = \sum X^2 - \frac{T^2}{N} = 34.916667$

Population sums of squares $= SSA = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} - \frac{T^2}{N} = 12.666667$

$SSE = SST - SSA = 34.916667 - 12.666667 = 22.25$

$S_1^2 = \frac{SSA}{k-1} = \frac{12.66667}{2} = 6.3333$ and $S^2 = \frac{SSE}{k(n-1)} = \frac{22.25}{9} = 2.4722$.

$$F = \frac{S_1^2}{S^2} = \frac{6.3333}{2.4722} = 2.56518\ .$$

Anova table:

Source	d.f.	Sum of Squares	Mean S	F	
Aggregate	$k-1=2$	$SSA = 12.666667$	$S_1^2 = 6.3333$	$F = \frac{S_1^2}{S^2} = 2.56518$	$F_{0.05}[2,\ 9] = 4.26$
Within the group	$N-k=9$	$SSE = 22.25$	$S^2 = 2.4722$		$F < F_{0.05}$
Total	11	$SST = 34.916667$	Conclusion: Accept the null hypothesis.		

Clearly $F < F_{0.05}[2,\ 9]$, since $F_{0.05}[2,\ 9] = 4.26.$

Hence, accept the hypothesis $\mu_1 = \mu_2 = \mu_3$, and conclude that there is no difference between the yields of verities.

Coding method: Since F is the variance ratio, its value remains unaffected if all the given values are codded. That is all the values are subtracted, added, divided or multiplied by a fixed number.

Example: A 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 one week there have been 14 sale calls; A made 5 calls, B made 4 calls and C make 5 calls. Following are the sales record.

A	300	400	300	500	0
B	600	300	300	400	-
C	700	300	400	600	500

Perform the Anova by codding method. (dividing by 100) and draw your conclusion.

salesman	A	B	C	
	3	6	7	
	4	3	3	
	3	3	4	
	5	4	6	
	0	-	5	
Total	15	16	25	56

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

Total sum of squares $= SST = \sum X^2 - \frac{T^2}{N} = 40$

Population sums of squares $= SSA = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \frac{(\sum X_3)^2}{n_3} - \frac{T^2}{N} = 10$

$SSE = SST - SSA = 40 - 10 = 30$

$S_1^2 = \frac{SSA}{k-1} = \frac{10}{2} = 5$ and $S^2 = \frac{SSE}{N-k} = \frac{30}{11} = 2.7273$.

$$F = \frac{S_1^2}{S^2} = \frac{5}{2.7273} = 1.8333 \ .$$

Anova table:

Source	d.f.	Sum of Squares	Mean S	F	
Aggregate	$k-1=2$	$SSA = 10$	$S_1^2 = 5$	$F = \frac{S_1^2}{S^2} = 1.8333$	$F_{0.05}[2,\ 11] = 3.98$
Within the group	$N-k = 11$	$SSE = 30$	$S^2 = 2.7273$		$F < F_{0.05}$
Total	13	$SST = 40$	Conclusion: Accept the null hypothesis.		

Clearly $F < F_{0.05}[2,\ 11]$, since $F_{0.05}[2,\ 11] = 3.98$.

Hence, accept the null hypothesis, and conclude that three salesmen A, B and C make sales of the same size.

Two - way Anova: If the data are classified according to two different factors. Then total variance comprises of three factors. $SST = SSC + SSR + SSE$.

SST is the total sum of squares.

SSC is the sum of squares between the columns. (SSA in one way Anova)

SSR is the sum of squares between the rows.

SSE is the error sum of squares.

Two-way Anova table:

Source	d.f.	Sum of Squares	Mean S	F	
Between the Columns	$v_1 = c-1$	SSC	$S_1^2 = \frac{SSC}{v_1}$	$F_C = \frac{S_1^2}{S^2}$	$F_\propto[v_1,\ v_3]$
Between the Rows	$v_2 = r-1$	SSR	$S_2^2 = \frac{SSR}{v_2}$	$F_R = \frac{S_2^2}{S^2}$	$F_\propto[v_2,\ v_3]$
Error	$v_3 = (c-1)(r-1)$	SSE	$S^2 = \frac{SSE}{v_3}$		
Total	$cr-1$	SST			

Examples:

1. Yield of Four varieties of wheat in 3 different plots per hectare is given bellow.

Set up the two-way Anova, and test whether varieties differ significantly, whether Plots differ significantly.

Plot \ Wheat	A	B	C	D
I	3	4	6	6
II	6	4	5	3
III	6	6	4	7

Solution:

Plot \ Wheat	A	B	C	D	Total
I	3	4	6	6	19
II	6	4	5	3	18
III	6	6	4	7	23
Total	15	14	15	16	$T = 60$

Correction factor$=\frac{T^2}{N} = \frac{60^2}{12} = 300.$

$$SST = \sum x^2 - \frac{T^2}{N} = 320 - 300 = 20.$$

$$SSC = \frac{\left(\sum X_{1j}\right)^2}{r} + \frac{\left(\sum X_{2j}\right)^2}{r} + \frac{\left(\sum X_{3j}\right)^2}{r} + \frac{\left(\sum X_{4j}\right)^2}{r} - \frac{T^2}{N} = \frac{902}{3} - 300 = 0.667$$

$$SSR = \frac{(\sum X_{i1})^2}{c} + \frac{(\sum X_{i2})^2}{c} + \frac{(\sum X_{i3})^2}{c} - \frac{T^2}{N} = \frac{1214}{4} - 300 = 3.5.$$

$$SSE = SST - SSC - SSR = 15.833.$$

Two-way Anova table:

Source	d.f.	Sum of Squares	Mean S	F	
Between the Columns	$v_1 = 3$	$SSC = 0.667$	$S_1^2 = \frac{SSC}{v_1}$ $= 0.222$	$F_C =$ 11.887	$F_{0.05}[6,\ 3]$ $= 8.93$
Between the Rows	$v_2 = 2$	$SSR = 3.5$	$S_2^2 = \frac{SSR}{v_2}$ $= 1.75$	$F_R =$ 1.508	$F_{0.05}[6,\ 2]$ $= 19.33$
Error	$v_3 = 6$	$SSE = 15.833$	$S^2 = \frac{SSE}{v_3} =$ 2.639	Conclusion: Reject H_{0C} . Accept H_{0R}	
Total	11	$SST = 20$			

Since $F_C > F_{0.05}[6,\ 3]$,varieties of wheat differs significantly, and $F_R < F_{0.05}[6,\ 2]$ no difference between the plots.

Two - way Anova if there is more than one entry in each factor: If the data are classified according to two different factors and one of the factors has multiple k entry. Then total variance comprises of four factors.
$SST = SSC + SSR + SSE + SSI.$

SST is the total sum of squares.

SSC is the sum of squares between the columns. (SSA in one way Anova)

SSR is the sum of squares between the rows.

SSE is the error sum of squares.

SSI is the interaction sum of squares

$$\therefore SSI = SST - (SSC + SSR + SSE)$$

Source	d.f.	Sum of Squares	Mean S	F	
Between the Columns	$v_1 = c - 1$	SSC	$S_1^2 = \frac{SSC}{v_1}$	$F_C = \frac{S_1^2}{S^2}$	$F_\propto[v_1,\ v_3]$
Between the Rows	$v_2 = r - 1$	SSR	$S_2^2 = \frac{SSR}{v_2}$	$F_R = \frac{S_2^2}{S^2}$	$F_\propto[v_2,\ v_3]$
Interaction	$v_3 = (c-1)(r-1)$	SSE	$S_3^2 = \frac{SSI}{v_3}$	$F_I = \frac{S_3^2}{S^2}$	
Error	$v_4 = (kcr - cr)$	SSI	$S^2 = \frac{SSE}{v_3}$		
Total	$kcr - 1$	SST			

Example:
Effectiveness in reducing blood pressure for three different groups of people; amount of blood pressure reduction in millimeters of mercury is as follows.

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

Answer the following in 5% level of significance.

1. Do the drugs act differently?
2. Are the different groups of people affected differently?
3. Is the interaction term significant?

Solution:

Drug / Group of people	X	Y	Z	Total
A	14 15	10 9	11 11	70
B	12 11	7 8	10 11	59
C	10 11	11 11	8 7	58
Total	73	56	58	187

Correction factor$=\frac{T^2}{N}=\frac{187^2}{18}=1942.72.$

$$SST=\sum x^2-\frac{T^2}{N}=2019-1942.72=76.28.$$

$$SSC=\frac{\left(\sum X_{1j}\right)^2}{kr}+\frac{\left(\sum X_{2j}\right)^2}{kr}+\frac{\left(\sum X_{3j}\right)^2}{kr}-\frac{T^2}{N}=\frac{11829}{6}-1942.72=28.78.$$

$$SSR=\frac{(\sum X_{i1})^2}{kc}+\frac{(\sum X_{i2})^2}{kc}+\frac{(\sum X_{i3})^2}{kc}-\frac{T^2}{N}=\frac{11745}{6}-1942.72=14.78.$$

$$\begin{aligned}SSE&=(14-14.5)^2+(15-14.5)^2+(10-9.5)^2+(9-9.5)^2+(11-11)^2+(11-11)^2\\&\quad+(12-11.5)^2+(11-11.5)^2+(7-7.5)^2+(8-7.5)^2+(10-10.5)^2+(11-10.5)^2\\&\quad+(10-10.5)^2+(11-10.5)^2+(11-11)^2+(11-11)^2+(8-7.5)^2+(7-7.5)^2\\&=3.5\,.\end{aligned}$$

$$SSI=SST-(SSC+SSR+SSE)=29.22\,.$$

Source	d.f.	Sum of Squares	Mean S	F	
Between the Columns	$v_1=2$	$SSC=28.78$	$S_1^2=14.39$	$F_C=36.99$	$F_{0.05}[2,\ 9]=4.26$
Between the Rows	$v_2=2$	$SSR=14.78$	$S_2^2=7.39$	$F_R=19$	$F_{0.05}[2,\ 9]=4.26$
Interaction	$v_3=4$	$SSI=29.22$	$S_3^2=7.31$	$F_I=18.79$	$F_{0.05}[4,\ 9]=3.63$
Error	$v_4=9$	$SSE=3.5$	$S^2=0.389$		
Total	17	SST			

The above table shows that all the three F ratios are significant of 5% level. Therefore

1. The drugs act differently.
2. The different groups of people affected differently.
3. The interaction term significant.

Since the interaction term significant it is pointless to talk about the difference between the groups of people of drugs.

Latin- Square: Latin-Square is an arrangements of n distinct symbols in a $n \times n$ matrix in which each row and each column must contain all n distinct symbols.

Examples: $\begin{bmatrix} A & B \\ B & A \end{bmatrix}$ $\begin{bmatrix} A & B & C \\ B & C & A \\ C & A & B \end{bmatrix}$

Latin- Square Design: If the data are arranged according to two different factors of n type each and n treatments as Latin- Square. Such design is called Latin- Square Design.

Two- way Anova for Latin- Square Design: If the data are classified according to two different factors and a treatment. Then total variance comprises of three factors. $SST = SSC + SSR + SSE + SSK.$

SST is the total sum of squares.

SSC is the sum of squares between the columns. (SSA in one way Anova)

SSR is the sum of squares between the rows.

SSE is the error sum of squares.

SSK is the sum of squares between the treatments.

Two-way Anova table:

Source	d.f.	Sum of Squares	Mean S	F	
Between the Columns	$v_1 = n-1$	SSC	$S_1^2 = \frac{SSC}{v_1}$	$F_C = \frac{S_1^2}{S^2}$	$F_\propto[v_1, v_4]$
Between the Rows	$v_2 = n-1$	SSR	$S_2^2 = \frac{SSR}{v_2}$	$F_R = \frac{S_2^2}{S^2}$	$F_\propto[v_2, v_4]$
between the treatments	$v_3 = n-1$	SSK	$S_3^2 = \frac{SSK}{v_2}$	$F_K = \frac{S_3^2}{S^2}$	$F_\propto[v_3, v_4]$
Error	$v_4 = (n-1)(n-2)$	SSE	$S^2 = \frac{SSE}{v_3}$		
Total	$n^2 - 1$	SST			

Examples:

1. Analyze and interpret the following statistics concerning output of wheat per field obtained as result of experiment conducted to test 4 varieties of wheat A, B , C and D under the 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:

	C 25	B 23	A 20	D 20	88
	A 19	D 19	C 21	B 18	77
	B 19	A 14	D 17	C 20	70
	D 17	C 20	B 21	A 15	73
Total	80	76	79	73	308
	$\sum A = 68$	$\sum B = 81$	$\sum C = 86$	$\sum D = 73$	

Correction factor$=\frac{T^2}{n^2} = \frac{308^2}{16} = 5929.$

$$SST = \sum x^2 - \frac{T^2}{n^2} = 6042 - 5929 = 113.$$

$$SSC = \frac{\left(\sum X_{1j}\right)^2}{n} + \frac{\left(\sum X_{2j}\right)^2}{n} + \frac{\left(\sum X_{3j}\right)^2}{n} + \frac{\left(\sum X_{4j}\right)^2}{n} - \frac{T^2}{n^2} = \frac{23746}{4} - 5929 = 7.5$$

$$SSR = \frac{(\sum X_{i1})^2}{n} + \frac{(\sum X_{i2})^2}{n} + \frac{(\sum X_{i3})^2}{n} - \frac{T^2}{n^2} = \frac{23902}{4} - 5929 = 46.5.$$

$$SSK = \frac{(\sum A)^2}{n} + \frac{(\sum B)^2}{n} + \frac{(\sum C)^2}{n} + \frac{(\sum D)^2}{n} - \frac{T^2}{n^2} = \frac{23910}{4} - 5929 = 48.9$$

$$SSE = SST - SSC - SSR - SSK = 10.1.$$

Anova table:

Source	d.f.	Sum of Squares	Mean S	F	$F_{0.05}[3,\ 6]$
Between the Columns	$v_1 = 3$	7.5	$S_1^2 = 2.5$	$F_C = 1.485$	4.76
Between the Rows	$v_2 = 3$	46.5	$S_2^2 = 15.5$	$F_R = 9.21$	
between the treatments	$v_3 = 3$	48.9	$S_3^2 = 16.3$	$F_K = 9.685$	$F_K > F_{0.05}$
Error	$v_4 = 6$	10.1	$S^2 = 1.683$		
Total	15	113	Conclusion: Accept H_{0C}; Reject H_{0R} ; RejectH_{0k}.		

Wheat varieties are differ significantly for any factor.