One-Way ANOVA

One-Way Analysis of Variance

The **ANALYSIS OF VARIANCE (ANOVA)** is used to determine whether there is any statistical significant difference between the means of three or more independent (unrelated) groups.

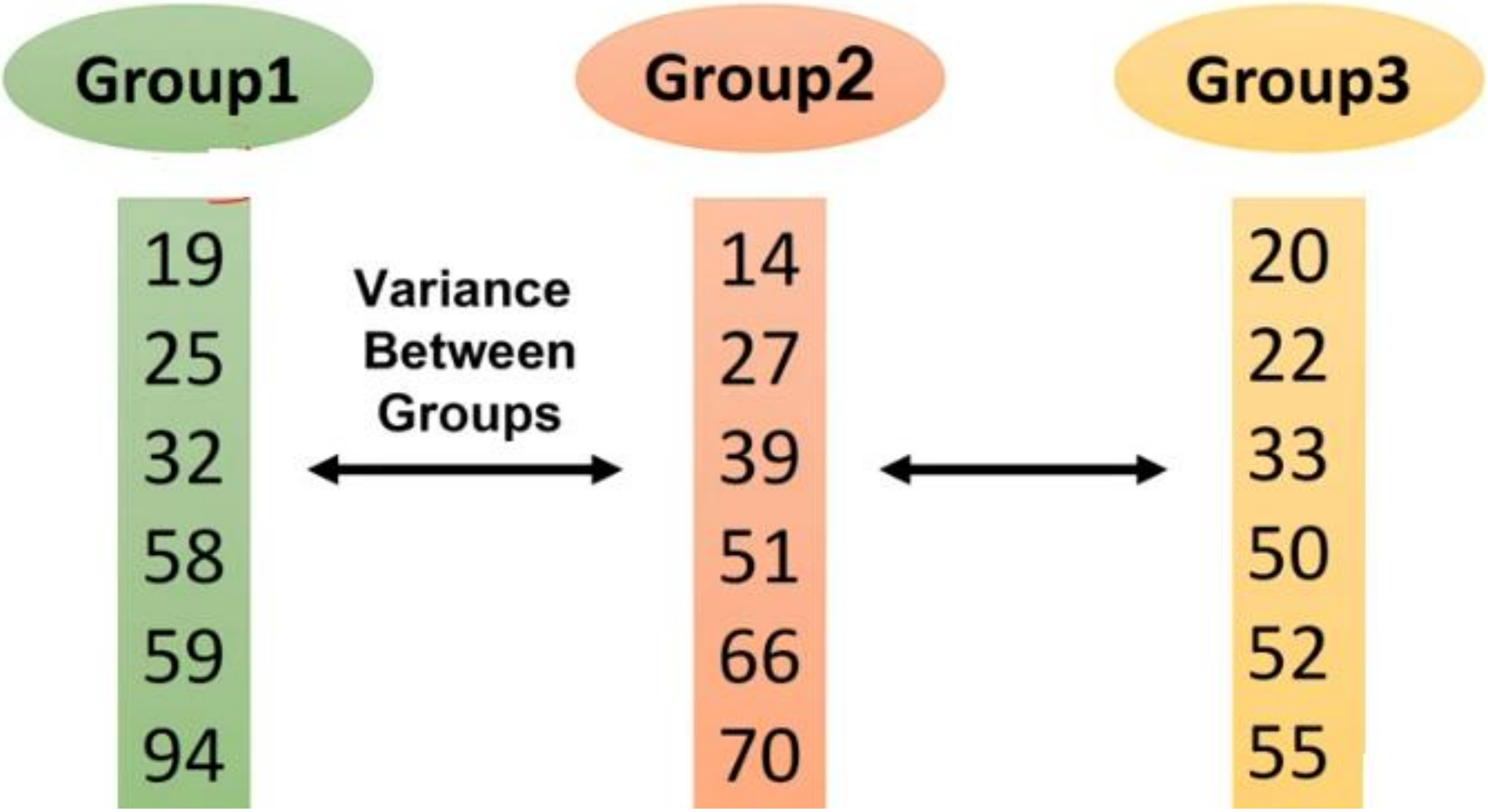
ANOVA

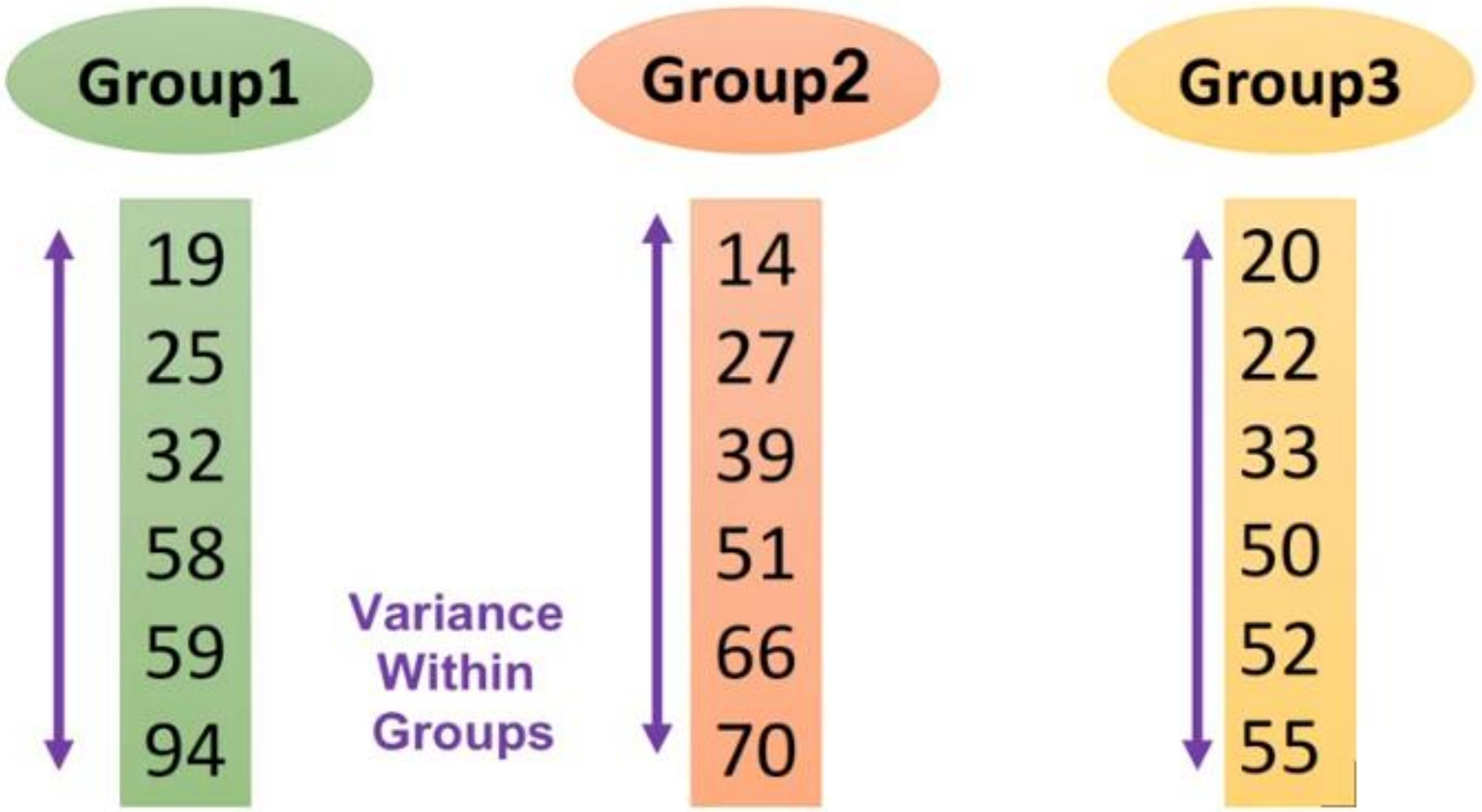
- Developed by **R.A. Fisher** in **1920**.
- Also known as **f-test** which is based on f distribution (G.W. Snedecor).
- It compares the means between the groups and determines whether any of those means are statistically significantly different from each other.
- It determines whether all groups are taken from common population or not.

- ANOVA is a ratio between **"Mean Sum of Squares between (MSS_B)"** and **"Mean Sum of Squares Within (MSS_W)"**.

$$F = \frac{MSS_B}{MSS_W} = \frac{\text{Between Variance}}{\text{Within Variance}}$$

- The variation among the observations of each specific group is called its **internal variation** and the totality of the internal variations is called variability within groups.
- The totality of variations from one group to another, i.e. variation due to groups is called **variability between Groups**.


Group1
Group2
Group3
19
25
32
58
59
94
14
27
39
51
66
70
20
22
33
50
52
55
Variance Between Groups


Group1
Group2
Group3
19
25
32
58
59
94
14
27
39
51
66
70
20
22
33
50
52
55
Variance
Within
Groups

It tests the null hypothesis

H_0*: *There is no significant difference between the means of all groups.

(All groups are same)

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

where μ = group mean k = number of groups

Alternative Hypothesis

H_A*: *There are at least two group means that are statistically significantly different from each other.

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

Assumptions

1. **Random Selection:** Samples are randomly selected.
2. **Normal Distribution:** Independent variable should be normally distributed.
3. **Homogeneity of Variances:** All sub populations have the same variance ·

$$\sigma_1 = \sigma_2 = \sigma_3 = ... = \sigma_k$$

4. **Additivity of Variances:** Total variance should be equal to sum of between variance & within variance.

Analysis of variance Summary table
(for one way classification)

Summary table for one way ANOVA

Source of Variation	*df*	SS	MSS	F calculated	F Tabulated at 5% and 1% level
Between (factors)	K-1	SSB	MSSB =SSB/(K-1)	$= \frac{MSSB}{MSSW}$	
Within (Error)	N-K	SSW	MSSW =SSW/(N-K)		
Total	N-1	TSS			

ANOVA Formula

ANOVA Test Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma n_j(\overline{X}_j - \overline{X})^2$	$df_1 = k - 1$	MSB = SSB / (k - 1)	f = MSB / MSE
Error	$SSE = \Sigma\Sigma(X - \overline{X}_j)^2$	$df_2 = N - k$	MSE = SSE / (N - k)	
Total	SST = SSB + SSE	$df_3 = N - 1$		

There are several components to the ANOVA formula. The best way to solve a problem on an ANOVA test is by organizing the formulas into an ANOVA table. The ANOVA formulas are given below.

Sum of squares between groups, $SSB = \sum n_j(\overline{X}_j - \overline{X})^2$. Here, $\overline{X}_j$ is the mean of the j^{th} group, $\overline{X}$ is the overall mean and n_j is the sample size of the j^{th} group.

$$\overline{X} = \frac{\overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \ldots + \overline{X}_j}{j}$$

Sum of squares of errors, SSE = $\sum\sum(X - \overline{X}_j)^2$. Here, X refers to each data point in the j^{th} group.

Total sum of squares, SST = SSB + SSE

Degrees of freedom between groups, $df_1 = k - 1$. Here, k denotes the number of groups.

Degrees of freedom of errors, $df_2 = N - k$, where N denotes the total number of observations across k groups.

Total degrees of freedom, $df_3 = N - 1$.

Mean squares between groups, MSB = SSB / (k - 1)

Mean squares of errors, MSE = SSE / (N - k)

ANOVA test statistic, f = MSB / MSE

Critical Value at $\alpha = F(\alpha, k - 1, N - k)$

The steps to perform the one way ANOVA test are given below:

- **Step 1:** Calculate the mean for each group.
- **Step 2:** Calculate the total mean. This is done by adding all the means and dividing it by the total number of means.
- **Step 3:** Calculate the SSB.
- **Step 4:** Calculate the between groups degrees of freedom.
- **Step 5:** Calculate the SSE.
- **Step 6:** Calculate the degrees of freedom of errors.
- **Step 7:** Determine the MSB and the MSE.
- **Step 8:** Find the f test statistic.
- **Step 9:** Using the f table for the specified level of significance, α, find the critical value. This is given by $F(\alpha, df_1. df_2)$.
- **Step 10:** If $f > F$ then reject the null hypothesis.

Production of Three Varieties of Wheat (Metric Tonnes/Acre)			
Fields	A	B	C
F1	6	5	5
F2	7	5	4
F3	3	3	3
F4	8	7	4

- Outcome (result): Production of Three varieties of wheat
- Variety of Wheat: 3
- Factors affecting the result: Variety of seed (wheat)
- Null Hypothesis: Production of all three varieties are equal
- Significance level: 5%

Production of Three Varieties of Wheat (Metric Tonnes/Acre)		
A	B	C
6	5	5
7	5	4
3	3	3
8	7	4

Step#1: Null and Alternative Hypothesis

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

(Production of variety A , B and C is same, it implies the variety of wheat doesn't affect the production of wheat significantly)

$$H_a: \text{Not all three are equal}$$

(Production of variety A , B and C not all three are equal, at least one is different significantly, it implies the variety of wheat affects the production of wheat significantly)

Production of Three Varieties of Wheat (Metric Tonnes/Acre)		
A	B	C
6	5	5
7	5	4
3	3	3
8	7	4

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

H_a: Not all three are equal

Step#2: SS between

A: $n_1 = 4, \quad \overline{X_1} = \frac{6+7+3+8}{4} = \frac{24}{4} = 6$

B: $n_2 = 4, \quad \overline{X_2} = \frac{5+5+3+7}{4} = \frac{20}{4} = 5$

C: $n_3 = 4, \quad \overline{X_3} = \frac{5+4+3+4}{4} = \frac{16}{4} = 4$

Mean of sample means or Total mean value $(\bar{\bar{X}}) = \frac{6+5+4}{3} = \frac{15}{3} = 5$

$$SS\ between = n_1(\overline{X_1} - \bar{\bar{X}})^2 + n_2(\overline{X_2} - \bar{\bar{X}})^2 + n_3(\overline{X_3} - \bar{\bar{X}})^2$$

$$\Rightarrow SS\ between = 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2$$

$$\Rightarrow SS\ between = 4 + 0 + 4 = 8$$

Production of Three Varieties of Wheat (Metric Tonnes/Acre)		
A	B	C
6	5	5
7	5	4
3	3	3
8	7	4

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

H_a: Not all three are equal

Step#3: SS within

A: $n_1 = 4$, $\overline{X_1} = 6$

B: $n_2 = 4$, $\overline{X_2} = 5$

C: $n_3 = 4$, $\overline{X_3} = 4$

$(\overline{\overline{X}}) = 5$

- SS between = 8

$$A \Rightarrow \Sigma(X_{1i} - \overline{X_1})^2 = (6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2 = 0+1+9+4 = 14$$

$$B \Rightarrow \Sigma(X_{2i} - \overline{X_2})^2 = (5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2 = 0+0+4+4 = 8$$

$$C \Rightarrow \Sigma(X_{3i} - \overline{X_3})^2 = (5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2 = 1+0+1+0 = 2$$

$$SS\ within = \Sigma(X_{1i} - \overline{X_1})^2 + \Sigma(X_{2i} - \overline{X_2})^2 + \Sigma(X_{3i} - \overline{X_3})^2$$

$$\Rightarrow SS\ within = 14 + 8 + 2 = 24$$

Production of Three Varieties of Wheat (Metric Tonnes/Acre)		
A	B	C
6	5	5
7	5	4
3	3	3
8	7	4

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

H_a: Not all three are equal

Step#4: SS Total

A: $n_1 = 4,\quad \overline{X_1} = 6$

B: $n_2 = 4,\quad \overline{X_2} = 5$

C: $n_3 = 4,\quad \overline{X_3} = 4$

$(\bar{\bar{X}}) = 5$

- SS between = 8
- SS within = 24

SS for total variance $= \Sigma\left(X_{ij} - \bar{\bar{X}}\right)^2$

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

$= 1 + 4 + 4 + 9 + 0 + 0 + 4 + 4 + 0 + 1 + 4 + 1 = 32$

Alternatively $\quad SS\ Total = SS\ between + SS\ within$

$= 8 + 24 = 32$

Production of Three Varieties of Wheat (Metric Tonnes/Acre)		
A	B	C
6	5	5
7	5	4
3	3	3
8	7	4

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

H_a: Not all three are equal

Step#5: Prepare the ANOVA Table

A: $n_1 = 4$, $\overline{X_1} = 6$

B: $n_2 = 4$, $\overline{X_2} = 5$

C: $n_3 = 4$, $\overline{X_3} = 4$

$(\overline{\overline{X}}) = 5$

- SS between = 8
- SS within = 24
- SS Total = 32
- Number of samples (m) = 3
- Number of elements in all samples (n) = 12

Table for One-Way ANOVA					
Source of variation	SS	d.f.	MS	F-ratio	F-limit 5% from F-distribution table
Between Sample	8	$(m-1) = (3-1) = 2$	$\frac{8}{2} = 4.00$	$\frac{4.00}{2.67} = 1.5$	F(2,9)=4.26
Within sample	24	$(n-m) = 12 - 3 = 9$	$\frac{24}{9} = 2.67$		
Total	32	$(n-1) = (12-1) = 11$			

Critical Values of F-Distribution @5% Significance level											
		Numerator									
	d.f.	1	2	3	4	5	6	8	12	24	30
D	1	161.45	199.50	215.71	224.58	230.16	233.99	238.88	243.91	249.05	250.10
e	2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.46
n	3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74	8.64	8.62
o	4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.75
m	5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.50
i	6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.81
n	7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.38
a	8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	3.08
t	9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.86
o	10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.70
r	11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.57

Production of Three Varieties of Wheat (Metric Tonnes/Acre)		
A	B	C
6	5	5
7	5	4
3	3	3
8	7	4

Step#6: Conclusion

Table for One-Way ANOVA

Source of variation	SS	d.f.	MS	F-ratio	F-limit 5% from F-distribution table
Between Sample	8	$(m-1)=(3-1)=2$	$\frac{8}{2}=4.00$	$\frac{4.00}{2.67}=1.5$	F(2,9)=4.26
Within sample	24	$(n-m)=12-3=9$	$\frac{24}{9}=2.67$		
Total	32	$(n-1)=(12-1)=11$			

H_0: $\mu_A = \mu_B = \mu_C$ True and can't be rejected

- Production of variety of wheat A , B and C are same,