ANALYSIS OF VARIANCE
(ANOVA)

PG- STAT-511

DEPARTMENT OF AGRICULTURAL ECONOMICS & STATISTICS

CHANDRA SHEKHAR AZAD UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

PRESENTED BY: PREETI TIWARI

HISTORY OF ANALYSIS OF COVARIANCE (ANOVA)

- **ANOVA was developed by the statistician Ronald Fisher in 1921 .**

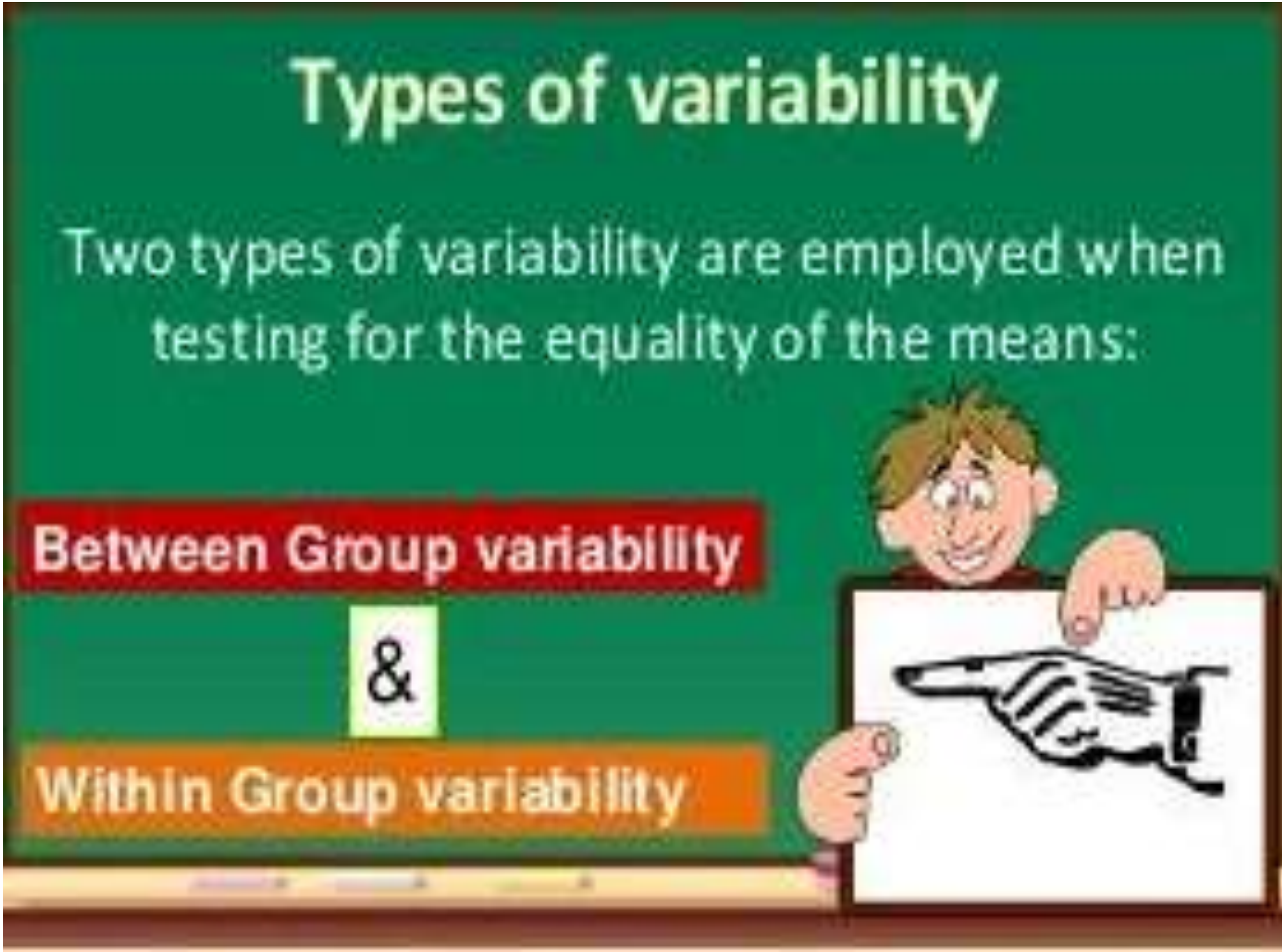
- **His first application of the analysis of variance was published in 1921. Analysis of variance became widely known after being included in Fisher's 1925 book Statistical Methods for Research Workers.**

- **ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation.**

MEANING OF ANALYSIS OF VARIANCE

- **Analysis of Variance** (ANOVA) is **a statistical formula used to compare variances across the means (or average) of different groups**.

- **ANOVA is a statistical method used to test differences between Means of two or more groups.**


Types of variability
Two types of variability are employed when testing for the equality of the means:
Between Group variability
&
Within Group variability

- ANOVA measures two Sources of Variation in the Data and compares their relatives Sizes.

- Variation BETWEEN Groups:

  For each data value look at the difference between its groups MEAN and overall Mean.

- Variation WITHIN Groups :

  For each data value we look at the difference between that value and the mean of its Group.

The ANOVA F- statistic is a ratio of the Between Group variation by the Within Group variation:

$$F = \frac{\text{variance between the sample}}{\text{Variance within the sample}}$$

CLASSIFICATION OF ANOVA

- The Analysis of variance is classified into two ways:

I. ONE-WAY CLASSIFICATION

II. TWO-WAY CLASSIFICATION

HOW THEY DIFFER?

ANOVA

One-Way ANOVA

- One independent variable
- Only one ‘p’ value is obtained.

Two way ANOVA

- Two independent Variables.
- Three different ‘p’ values are obtained.
- Outcome of factorial Design.

ONE-WAY ANOVA

- The one-way analysis of variance is used to test the claim that three or more population means are equal.
- This is an extension of the two independent samples t-test.

Data representation for one-way ANOVA

Treatments					Total
Treatment 1	x_{11}	x_{12}	…	x_{1n_1}	$x_{1.}$
Treatment 2	x_{21}	x_{22}	…	x_{2n_2}	$x_{2.}$
⋮	⋮	⋮	…	⋮	⋮
Treatment k	x_{k1}	x_{k2}	…	x_{kn_k}	$x_{k.}$

- **x_{ij} - the j th sample value from the ith treatment, j 2,…,k**
- **k - number of treatments compared.**
- **$x_{i.}$ - the sample total of i^{th} treatment.**
- **n_i - the number of observations in the i^{th} treatment.**

$$\sum_{i=1}^{k} n_i = n$$

The total variation in the observations x_{ij} can be split into the following two components

(i) variation between the levels or the variation due to different bases of classification, commonly known as treatments.

(ii) The variation within the treatments *i.e.*
inherent variation among the observations within levels.

State the null hypothesis and alpha level

The null hypothesis is that all the groups have equal means.

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

The alternative hypothesis is that there is at least one significant difference between the means

Level of significance α is selected as 0.05

- General Formulas : Total Variation

$$SST = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (X_{ij} - \bar{\bar{X}})^2$$

where $\bar{\bar{X}} = \dfrac{\sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}}{n}$ = grand mean

X_{ij} = ith value in group j

n_j = number of values in group j

n = total number of values in all groups combined

c = number of groups

- General Formulas : Variation Among the groups

$$SSA = \sum_{j=1}^{c} n_j (\bar{X}_j - \bar{\bar{X}})^2$$

where $\bar{\bar{X}}$ = grand mean

$\bar{X}_j$ = sample mean of group j

n_j = number of values in group j

c = number of groups

•General Formulas : Variation within the Groups

$$SSW = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2$$

where X_{ij} = ith value in group j

$\overline{X}_j$ = sample mean of group j

Degrees of Freedom (d.f)

Degrees of freedom (d.f.)		d.f.
Total Sum of Squares	Total no. of observations -1	n-1
Treatment Sum of Squares	Total no. of observations -1	k-1
Error of Sum Squares	Total no. of observations -1	n-k

ONE-WAY ANOVA TABLE

Source of Variance	Degree of Freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups (Treatment)	k-1	$SSB = \sum_{j=1}^{k} \left(\frac{T_j^2}{n_j}\right) - \frac{T^2}{n}$ $\quad SSB = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X}_t)^2$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups (Error)	n-k	$SSW = \sum_{j=1}^{K} \sum_{i=1}^{n} X_{ij}^2 - \sum_{j=1}^{K} \left(\frac{T_j^2}{n_j}\right)$ $\quad SSW = \sum_{j=1}^{k} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2$	$MSW = \frac{SSW}{n-k}$	
Total	n-1	$SST = \sum_{j=1}^{K} \sum_{i=1}^{n} X_{ij}^{2} - \frac{T^2}{n}$ $\quad SST = \sum_{j=1}^{k} \sum_{i=1}^{n} (X_{ij} - \bar{X}_t)^2$		

- SST = SSB + SSW

k: number of groups n: number of samples

df: degree of freedom

PROBLEM : **Three different techniques namely medication, exercises and special diet are randomly assigned to (individuals diagnosed with high blood pressure) lower the blood pressure. After four weeks the reduction in each person's blood pressure is recorded. Test at 5% level, whether there is significant difference in mean reduction of blood pressure among the three techniques.**

Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	12	8	4

Step 1 : Hypotheses

Null Hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3$

That is, there is no significant difference among the three groups on the average reduction in blood pressure.

Alternative Hypothesis: H_1: $\mu_i \neq \mu_j$ for atleast one pair (i, j); $i, j = 1, 2, 3$; $i \neq j$.

						Total	Square
Medication	10	12	9	15	13	59	3481
Exercise	6	8	3	0	2	19	361
Diet	5	9	12	8	4	38	1444
						G = 116	5286

Individual squares

Medication	100	144	81	225	169
Exercise	36	64	9	0	4
Diet	25	81	144	64	16

$$\sum\sum x_{ij}^2 = 1162$$

1. Correction Factor: $CF = \frac{G^2}{n} = \frac{(116)^2}{15} = \frac{13456}{15} = 897.06$

2. Total Sum of Squares: $TSS = = \sum\sum x_{ij}^2 - C.F$

$$= 1162 - 897.06 = 264.94$$

3. Sum of Squares between Treatments: $SST = \frac{\sum x_i^2}{n_i} - C.F$

$$= \frac{5286}{5} - 897.06$$

$$= 1057.2 - 897.06$$

$$= 160.14$$

4. Sum of Squares due to Error: $SSE = TSS - SST$

$$= 264.94 - 160.14 = 104.8$$

ANOVA Table (one-way)

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Between treatments	160.14	$3 - 1 = 2$	80.07	$F_0 = \frac{80.07}{8.73} = 9.17$
Error	104.8	12	8.73	
Total	264.94	$n - 1 = 15 - 1 = 14$		

Step 6 : Critical value

$f_{(2,\ 12),0.05} = 3.8853$.

Step 7 : Decision

As $F_0 = 9.17 > f_{(2,\ 12),0.05} = 3.8853$, the null hypothesis is rejected. Hence, we conclude that there exists significant difference in the reduction of the average blood pressure in atleast one pair of techniques.

TWO-WAY ANOVA

THE DATA CAN BE REPRESENTED IN THE FOLLOWING TABULAR FORM.

Blocks

1	2	3	...
x_{11}	x_{12}	$x_{13.}$	...
x_{21}	x_{22}	$x_{2.}$	...
x_{31}	x_{32}	$x_{3.}$	...
⋮	⋮	⋮	⋮
x_{k1}	x_{k2}	x_{k3}	...
$x_{.1}$	$x_{.2}$	$x_{.3}$	...

We use the following notations.

x_{ij} - i^{th} treatment value from the j^{th} block, i = 1,2, ..., k; j = 1,2, ..., m.

The i^{th} treatment total - x_i = , i 1, 2, ..., k

The j^{th} block total - $x_{.j}$ x_{ij} , j 1, 2, ..., m

Note that, $k \times m = n$, where m = number of blocks, and k = number of treatments (groups) and n is the total number of observations in the study.

- The total variation present in the observations x_{ij} can be split into the following three components:

- (i) The variation between treatments (groups)
- (ii) The variation between blocks.
- (ii) The variation inherent within a particular setting or combination of treatment and block.

❑ Test Procedure

- Steps involved in two-way ANOVA are:
- **Step 1 :** In two-way ANOVA we have two pairs of hypotheses, one for treatments and one for the blocks.
- **Framing Hypotheses**
- **Null Hypotheses**
- H_{01}: There is no significant difference among the population means of different groups (Treatments)
- H_{02}: There is no significant difference among the population means of different Blocks
- **Alternative Hypotheses**
- H_{11}: Atleast one pair of treatment means differs significantly
- H_{12}: Atleast one pair of block means differs significantly

- **Step 2 : Data** is presented in a rectangular table form as described in the previous section.

- **Step 3 : Level of significance** α.

- **Step 4 : Test Statistic**

- F_{0t} (treatments) = MST / MSE
- F_{0b} (block) = MSB / MSE
- To find the test statistic we have to find the following intermediate values.
-

❑ To find the test statistic we have to find the following intermediate values.

i) Correction Factor: $C.F = \frac{G^2}{n}$ where $G = \sum_{j=1}^{m}\sum_{i=1}^{k} x_{ij}$

ii) Total Sum of Squares: $TSS = \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}^2 - C.F$

iii) Sum of Squares between Treatments: $SST = \sum_{i=1}^{k} \frac{x_{i.}^2}{m} - C.F$

iv) Sum of squares between blocks: $SSB = \sum_{j=1}^{m} \frac{x_{.j}^2}{k} - C.F$

v) Sum of Squares due to Error:

$$SSE = TSS - SST - SSB$$

vi) Degrees of freedom

Degrees of freedom (d.f.)	d.f.
Total Sum of Squares	n-1
Treatment Sum of Squares	k-1
Block Sum of Squares	m-1
Error of Sum Squares	$(m - 1)(k - 1)$

❑ Mean Sum of Squares

Mean sum of Squares due to Treatments: $MST = \dfrac{SST}{k-1}$

Mean sum of Squares due to Blocks: $MSB = \dfrac{SSB}{m-1}$

Mean sum of Squares due to Error: $MSE = \dfrac{SSE}{(k-1)(m-1)}$

- Step 5 : Calculation of the Test Statistic

TWO-WAY ANOVA TABLE

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Treatments	SST	k-1	MST	$F_{0t} = \frac{MST}{MSE}$
Blocks	SSB	m-1	MSB	$F_{0b} = \frac{MSB}{MSE}$
Error	SSE	$(k$-1$)(m$-1$)$	MSE	
Total	TSS	n-1		

PROBLEM : A reputed marketing agency in India has three different training programs for its salesmen. The three programs are Method – A, B, C. To assess the success of the programs, 4 salesmen from each of the programs were sent to the field. Their performances in terms of sales are given in the following table. Test whether there is significant difference among methods and among salesmen.

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Step 1 : Hypotheses

Null Hypotheses: $H_{01} : \mu_{M1} = \mu_{M2} = \mu_{M3}$ (for treatments)
That is, there is no significant difference among the three programs in their mean sales.
H_{02} **:** $\boldsymbol{\mu_{S1} = \mu_{S2} = \mu_{S3} = \mu_{S4}}$ **(for blocks)**

Alternative Hypotheses:
H_{11} : At least one average is different from the other, among the three programs.
H_{12} : At least one average is different from the other, among the four salesmen.

Step 2 : Data

Salesmen	Methods		
	A	B	C
1	4	6	2
2	6	10	6
3	5	7	4
4	7	5	4

Calculation of the Test Statistic

	Methods			Total $x_{i.}$	$x_{i.}^2$
	A	B	C		
1	4	6	2	12	144
2	6	10	6	22	484
3	5	7	4	16	256
4	7	5	4	16	256
x_{i}	22	28	16	66	1140
$x_{i.}^2$	484	784	256	1524	

Squares

16	36	4
36	100	36
25	49	16
49	25	16
		$\sum\sum x_{ij}^2 = 408$

Correction Factor: $$CF = \frac{G^2}{n} = \frac{(66)^2}{12} = \frac{4356}{12} = 363$$

Total Sum of Squares: $$TSS = \sum\sum x_{ij}^2 - C.F$$
$$= 408 - 363 = 45$$

Sum of Squares due to Treatments: $$SST = \frac{\sum_{i=1}^{k} x_{.j}^2}{k} - C.F$$
$$= \frac{1140}{3} - 363$$
$$= 380 - 363 = 17$$

Sum of Squares due to Blocks: $$SSB = \frac{\sum_{i=1}^{k} x_{.j}^2}{k} - C.F$$
$$= \frac{1524}{4} - 363$$
$$= 381 - 363$$
$$= 18$$

Sum of Squares due to Error: $$SSE = TSS - SST - SSB$$
$$= 45 - 17 - 18 = 10$$

Mean sum of squares: $$MST = \frac{SST}{k-1} = \frac{17}{3} = 5.67$$
$$MSB = \frac{SSB}{m-1} = \frac{18}{2} = 9$$
$$MSE = \frac{SSE}{(k-1)(m-1)} = \frac{10}{6} = 1.67$$

Step 6 : Critical values

$f_{(3, 6),0.05}$ = 4.7571 (for treatments)
$f_{(2, 6),0.05}$ = 5.1456 (for blocks)

Step 7 : Decision

(i) Calculated F_{0t} = 3.40 < $f_{(3, 6),0.05}$ = 4.7571, the null hypothesis is not rejected and we conclude that there is significant difference in the mean sales among the three programs.

(ii) Calculate F_{0b} = 5.39 > $f_{(2, 6),0.05}$ = 5.1456, the null hypothesis is rejected and conclude that there does not exist significant difference in the mean sales among the four salesmen.

Sources of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F-ratio
Between treatments (Programs)	17	3	5.67	$F_{0t} = \frac{5.67}{1.67} = 3.40$
Between blocks (Salesmen)	18	2	9	$F_{0b} = \frac{9}{1.67} = 5.39$
Error	10	6	1.67	
Total		11		

Thank You