# Analysis of Variance (ANOVA)
# one way

**ANOVA General ANOVA Setting** "Slide 43-45)

- Investigator controls one or more factors of interest
    - Each factor contains two or more levels
    - Levels can be numerical or categorical
    - Different levels produce different groups
    - Think of each group as a sample from a different population
    - Observe effects on the dependent variable, Are the groups the same?
- Experimental design: the plan used to collect the data
- Experimental units (subjects) are assigned randomly to groups , Subjects are assumed homogeneous
- Only one factor or independent variable ,With two or more levels.
- Analyzed by one-factor analysis of variance (ANOVA)

**One-Way Analysis of Variance**

- Evaluate the difference among the means of three or more groups

Examples:  Number of accidents for $1^{st}$, $2^{nd}$, and $3^{rd}$ shift            Expected mileage for five brands of tires

- Assumptions
    - Populations are normally distributed
    - Populations have equal variances
    - Samples are randomly and independently drawn

**(Textbook: P374)**

**Data frame:**

| Observations | Groups | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | …. …. …… | C | Total |
| 1 | $X_{11}$ | $X_{21}$ | …..$X_{i1}$ ……. | $X_{c1}$ | |
| 2 | $X_{12}$ | $X_{22}$ | …..$X_{i2}$…….. | $X_{c2}$ | |
| …. | … | …. | …….………. | …. | |
| j | $X_{1j}$ | $X_{2j}$ | ……$X_{ij}$……. | $X_{cj}$ | |
| Sum | $X_{1.}$ | $X_{2.}$ | ……$X_{j.}$….. | $X_{c.}$ | X (Grand Total) |
| Samples | $n_1$ | $n_2$ | …..$n_j$….. | $n_c$ | n The Total sample |
| Mean | $\bar{X}_1$ | $\bar{X}_2$ | …..$\bar{X}_i$…… | $\bar{X}_c$ | $\bar{\bar{X}}$ (Grand Mean) |

c : number of groups or levels

$n_j$ : number of values in group j

$X_{ij}$ : $i^{th}$ observation from group j

X : Total mean (Total of all data values)

n: The Total of all samples ($n = n_1 + n_2 + …… + n_j$)

$\bar{\bar{X}}$ : grand mean (mean of all data values)

Analysis of variance is a general method for studying sampled-data relationships. The method enables the difference between two or more sample means to be analyzed, achieved by subdividing the total sum of squares.
Basic idea is to partition total variation of the data into two sources:
   1- Variation Within Groups (variation due to factor) .
   2- Variation Among Groups (variation due to random error)

Total Variation : the aggregate variation of the individual data values across the various factor levels (SST)

Among-Group Variation : variation among the factor sample means (SSA)

Within-Group Variation : variation that exists among the data values within a particular factor level (SSW)
(Slide 51-52)

**The equations used to calculate these totals are:**

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

$$SST = (X_{11} - \bar{\bar{X}})^2 + (X_{12} - \bar{\bar{X}})^2 + \cdots + (X_{cn_c} - \bar{\bar{X}})^2$$

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

$$SSA = n_1 (\bar{X}_1 - \bar{\bar{X}})^2 + n_2 (\bar{X}_2 - \bar{\bar{X}})^2 + \cdots + n_c (\bar{X}_c - \bar{\bar{X}})^2$$

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

$$SSW = (X_{11} - \bar{X}_1)^2 + (X_{12} - \bar{X}_2)^2 + \cdots + (X_{cn_c} - \bar{X}_c)^2$$

**Obtaining the Mean Squares**

The Mean Squares are obtained by dividing the various sum of squares by their associated degrees of freedom

Mean Square Among (d.f. = c-1)　　　:　　　$MSA = \dfrac{SSA}{c-1}$

Mean Square Within (d.f. = n-c)　　:　　　$MSW = \dfrac{SSW}{n-c}$

Mean Square Total (d.f. = n-1)　　:　　　$MST = \dfrac{SST}{n-1}$

(c-1) : The degrees of freedom for the Among group
(n-c) : The degrees of freedom for the within group

**F Test for differences among more than two means** **Slide (63-63)**

Step (1) : State the null and alternate hypotheses :

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots.. = \mu_c$

$H_1$ : At least two population means are different.

Step (2): Select the level of significance (α)

Step (3): The test statistic : Because we are comparing means of more than two groups, use the F statistic.

$$F_{STAT} = \frac{MSA}{MSW} = \frac{SSA/c - 1}{SSW/n - c}$$

Step (4): The critical value:

The degrees of freedom for the numerator are the degrees of freedom for the Among group (c-1)

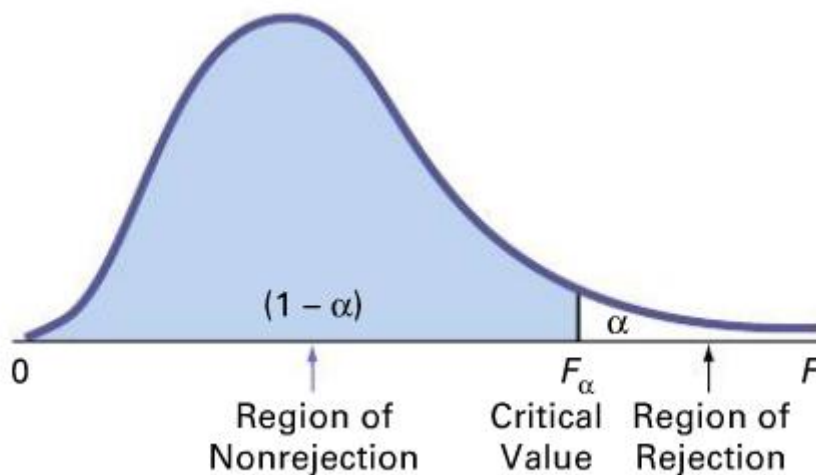The degrees of freedom for the denominator are the degrees of freedom for the within group (n-c).

$$F_{(\alpha , c-1 , n-c)}$$

**(Textbook: Table E.5 P548-553)**

Step (5) : Formulate the decision Rule and make a decision

Reject $H_o$

If $F_c > F_{(\alpha , c-1, n-c)}$

**ANOVA TABLE**

It is convenient to summarize the calculation of the F statistic in (ANOVA Table)
(Slide 62)

| Source of variation (S.V) | Degrees of freedom | Sum of Squares (S.S) | Mean Squares (MS) | F- ratio |
|---|---|---|---|---|
| Among groups | c-1 | SSA | MSA =SSA/c-1 | $F_{STAT}$ = MSA / MSW |
| Within groups | n-c | SSW | MSW = SSW/n-c | |
| Total | n-1 | SST | | |

**Example** (Slide 65-68)

You want to see if three different golf clubs yield different distances. You randomly select five measurements from trials on an automated driving machine for each club. At the 0.05 significance level, is there a difference in mean distance?

| | Club 1 | Club 2 | Club 3 |
|---|---|---|---|
| | 254 | 234 | 200 |
| | 263 | 218 | 222 |
| | 241 | 235 | 197 |
| | 237 | 227 | 206 |
| | 251 | 216 | 204 |
| Total | 1246 | 1130 | 1029 |
| Mean | 249.2 | 226 | 205.8 |
| $\bar{\bar{X}} = 227$ | | | |

C=3      $n_1=n_2=n_3= 5$      n=15

$$SSA = 5(249.2 - 227)^2 + 5(226 - 227)^2 + 5(205.8 - 227)^2 = 4716.4$$

$$SSW = (254 - 249.2)^2 + (263 - 249.2)^2 + \cdots + (204 - 205.8)^2 = 1119.6$$

Step (1) :  State the null and alternate hypotheses :
$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1$ : At least two population means are different.

Step (2): Select the level of significance ($\alpha$ =0.05)

Step (3): The test statistic  :
$$F_{STAT} = \frac{MSA}{MSW} = \frac{SSA/c - 1}{SSW/n - c} = \frac{4716.4/3 - 1}{1119.6/15 - 3} = 25.275$$
Step (4): The critical value:

The degrees of freedom for <mark>the numerator</mark> $(c-1) = 3-1 = 2$

The degrees of freedom for the <mark>denominator</mark> $(n-c) = 15-3 = 12$
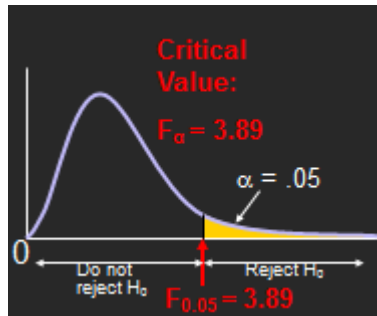
$$F_{(0.05, \ 2, 12)} = 3.89$$

<mark>Step (5)</mark> : Formulate the decision Rule and make a decision

$$F_{STAT} \ (25.275) > F_{(0.05, 2, 12)} (3.89)$$

Reject Ho at $\alpha = 0.05$

Conclusion: There is evidence that at least one $\mu j$ differs from the rest



(Textbook P" 378-381 "Starting from paragraph 3, there is an illustrative example)