8.1 Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: **systematic factors and random factors.** The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The Formula for ANOVA is:

$$F = \frac{MST}{MSE}$$

where:

F = ANOVA coefficient

MST = Mean sum of squares due to treatment

MSE = Mean sum of squares due to error

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

1- **H_0:μ_1=μ_2=μ_3(**H_0=there are no difference between means of groups

2- **H_1:$\mu1\neq\mu2\neq\mu3$(H_1=there are difference between means of groups**. We need a test to tell which means are different.

Post Hoc Tests:

-Tukey HSD

HSD (Honestly Significant Difference) use with equal sample size per cell.

$$HSD = q\sqrt{\frac{MSw}{NA}}$$

q=tables Tukey value

NA=number sample within any groups.

MSw=square mean within groups.

Example1: We have data for 15 students. They were divided into 3 groups. Does the mean differ between the groups or not?

G1	G2	G3
75	80	70
77	82	72
79	84	74
81	86	76
83	88	78

Sol:

Mean(within groups)		
G1	G2	G3
395/5=79	420/5=84	370/5=74

Variance		
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 ,$$		
$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2$
$(75-79)^2$	$(80-84)^2$	$(70-74)^2$
$(77-79)^2$	$(82-84)^2$	$(72-74)^2$
$(79-79)^2$	$(84-84)^2$	$(74-74)^2$
$(81-79)^2$	$(86-84)^2$	$(76-74)^2$
$(83-79)^2$	$(88-84)^2$	$(78-74)^2$
(square variance)SS total = $(x_i - \bar{x})^2$ **=120**		
Standard Deviation $$s = \sqrt{\text{variance}}$$		
$\sqrt{\frac{40}{4}}$**=3.16**	$\sqrt{\frac{40}{4}}$**=3.16**	$\sqrt{\frac{40}{4}}$**=3.16**
Mean(between groups)		
G1	**G2**	**G3**
(79+84+74)/3=79		
$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2$
$(75-79)^2$	$(80-79)^2$	$(70-79)^2$
$(77-79)^2$	$(82-79)^2$	$(72-79)^2$
$(79-79)^2$	$(84-79)^2$	$(74-79)^2$
$(81-79)^2$	$(86-79)^2$	$(76-79)^2$

$(83-79)^2$	$(88-79)^2$	$(78-79)^2$
(square variance)SS total = $(x_i - \bar{x})^2$ **=250**		

ANOVA Tables

Source	SS	Degree Freedom df	MS	$F=MS_B/MS_W$
Between groups	**250**	**(Groups)k-1=2**	**SS/DF 250/2=125** MS_B	**125/10=12.5**
Within groups	**120**	**N-K=15-3=12**	**120/12=10** MS_W	
Total	**370**	**14**		

F critical value=3.89

F value=12.5

12.5>3.89 (Reject H_O)

H1: μ1≠μ2≠μ3(H1=there are difference between means of groups.

We need a test to tell which means are different.

Post Hoc Tests:

-Tukey HSD

$$HSD = q\sqrt{\frac{MSw}{NA}}$$

q=tables Tukey value =3.77 , NA=5 .MSw=10

$$HSD = 3.77\sqrt{\frac{10}{5}} = 5.33$$

Difference between mean G1 and G2(79-84=-5)

Difference between mean G1 and G3(79-74=5)

Difference between mean G2 and G3(84-74=10)

10>5.33 (function difference because difference between mean G2 and G3)