ANALYSIS OF VARIANCE

The analysis of variance is a powerful tool for tests of significance. The test of significance of based on t- distribution is an adequate test procedure only for testing the significance of the difference between two sample means.

In a situation when we have a three or more samples to consider at a time an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population. i.e. they have the same mean.

Use a one-way ANOVA when you have collected data about one categorical independent variable and one quantitative dependent variable. The independent variable should have at least three levels (i.e. at least three different groups or categories).

ANOVA tells you if the dependent variable changes according to the level of the independent variable. For example:

- Your independent variable is social media use, and you assign groups to low, medium, and high levels of social media use to find out if there is a difference in hours of sleep per night.
- Your independent variable is brand of soda, and you collect data on Coke, Pepsi, Sprite, and Fanta to find out if there is a difference in the price per 100ml.
- You independent variable is type of fertilizer, and you treat crop fields with mixtures 1, 2 and 3 to find out if there is a difference in crop yield.

For e.g. Five fertilizers are applied to four plots each of wheat and yield of wheat on each of the plot is given. We may be interested in finding out whether the effect of these fertilizers on the yield is significantly different or in other words the samples are drawn from the same population.

The answer to this problem is provided by the technique of analysis of variance. The basic purpose of the analysis of variance is to test the homogeneity of several means.

The term 'Analysis of Variance' was introduced by Prof. R. A. Fisher in 1920's to deal with agronomical data. Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as:

(i) Assignable causes (ii) Chance causes

The variation due to assignable causes can be detected and measured whereas the variation due to chance causes is beyond the control of human hand and cannot be traced separately.

- **DEFINITION**: According to R.A. Fisher, Analysis of Variance (ANOVA) is the "separation of variance ascribable to one group of causes from the variance ascribable to other group."

By this technique the total variation in the sample data is expressed as the sum of its non-negative components where each of these components is a measure of variation due to some specific independent cause. The ANOVA consists in the estimation of the amount of variation due to each of the independent factors with the estimate due to chance factor (commonly known as experimental error)

- ASSUMPTIONS for ANOVA test:

ANOVA test is based on the test statistic F (or variance ratio).

For the validity of F- test in ANOVA the following assumptions are made:

I) the observations add are independent.

II) Parent population from which observations are taken is normal.

III) Various treatment and environmental effects are additive in nature.

NOTE: The test statistic F assumes equal variability in the k populations (i.e., the population variances are equal, or $s_1^2 = s_2^2 = ... = s_k^2$). This means that the outcome is equally variable in each of the comparison populations. This assumption is the same as that assumed for appropriate use of the test statistic to test equality of two independent means. It is possible to assess the likelihood that the assumption of equal variances is true and the test can be conducted in most statistical computing packages. If the variability in the k comparison groups is not similar, then alternative techniques must be used.

The F statistic is computed by taking the ratio of what is called the "between treatments" variability to the "residual or error" variability. This is where the name of the procedure originates. In analysis of variance we are testing for a difference in means (H_0: means are all equal versus H_1: means are not all equal) by evaluating variability in the data. The numerator captures between treatment variability (i.e., differences among the sample means) and the denominator contains an estimate of the variability in the outcome. The test statistic is a measure that allows us to assess whether the differences among the sample means (numerator) are more than would be expected by chance if the null hypothesis is true.

The decision rule for the F test in ANOVA is set up in a similar way as we established for other tests. The decision rule again depends on the level of significance and the degrees of freedom. The F statistic has two degrees of freedom. These are denoted df_1 and df_2, and called the numerator and denominator degrees of freedom, respectively.

If the null hypothesis is true, the between treatment variation (numerator) will not exceed the residual or error variation (denominator) and the F statistic will small. If the null hypothesis is false, then the F statistic will be large. The rejection region for the F test is always in the upper (right-hand) tail of the distribution.