Analysis of Variance (ANOVA) using R

Asst. Prof. Ashwini Mathur

ANOVA (Analysis of Variance)

ANOVA stands for Analysis Of Variance. ANOVA was founded by Ronald Fisher in the year 1918.

The name Analysis Of Variance was derived based on the approach in which the method uses the variance to determine the means whether they are different or equal.

Overview

It is a statistical method used to test the differences between two or more means. It is used to test general differences rather than specific differences among means. It assesses the significance of one or more factors by comparing the response variable means at different factor levels.

Null hypothesis states that all population means are equal. The alternative hypothesis proves that at least one population mean is different

General purpose of ANOVA

The reason for performing ANOVA is to see whether any difference exists between the groups on some variable.

You can use t-test to compare the means of two samples but when there are more than two samples to be compared then ANOVA is the best method to be used.

Assumptions of ANOVA

There are four main assumptions

- The expected values of the errors are zero
- The variances of all the errors are equal to each other
- The errors are independent
- They are normally distributed

ANOVA Types

One Way ANOVA is used to check whether there is any significant difference between the means of three or more unrelated groups. It mainly tests the null hypothesis.

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

Where μ means group mean and x means number of groups. One Way ANOVA gives a significant result.

Example of One Way ANOVA

20 people are selected to test the effect of five different exercises. 20 people are divided into 4 groups with 5 members each. Their weights are recorded after a few days. The effect of the exercises on the 5 group of men are compared. Here weight is the only one factor.

Assumptions

The dependent variable is normally distributed in each group

There is homogeneity of variances

Independence of observations

Two way between groups

The two way ANOVA compares the mean difference between groups that have been split on two factors.

The main objective of a two way ANOVA is to find out if there is any interaction between the two independent variables on the dependent variables.

It also lets you know whether the effect of one of your independent variables on the dependent variable is same for all the values of your other independent variable.

Example

The research of the effect of fertilizers on yield of rice. You apply five fertilizers of different quality on five plots of land each cultivating rice.

The yield from each plot of land is recorded and the difference between each plot is observed. Here the effect of the fertility of the plots can also be studied. Thus there are two factors, Fertilizer and Fertility.

Assumptions

Before starting with your two way ANOVA your data should pass through six assumptions to make sure that the data you have is sufficient for performing two way ANOVA. The six assumptions are listed below

- Your dependent variable should be measured at the continuous level
- Your two independent variable should contain two or more categorical independent groups for each
- You should have independence of observations
- Avoid any outliers
- Your dependent variable should be normally distributed for each combination of the groups of the two independent variable
- Homogeneity of variances

Demonstration of ANOVA Test Using R

Imagine that you are interested in understanding whether knowing the brand of car tire can help you predict whether you will get more or less mileage before you need to replace them.

We'll draw what is hopefully a random sample of 60 tires from four different manufacturers and use the mean mileage by brand to help inform our thinking. While we expect variation across our sample we're interested in whether the differences between the tire brands (the groups) is significantly different than what we would expect in random variation within the groups.

Our research or testable hypothesis is then described

$$\mu Apollo \neq \mu Bridgestone \neq \mu CEAT \neq \mu Falken$$

as at least one of the tire brand populations is different than the other three. Our null is basically “nope, brand doesn’t matter in predicting tire mileage – all brands are the same”.

Compute one-way ANOVA test

In [18]:
```
# Compute the analysis of variance
res.aov <- aov(weight ~ group, data = my_data)
# Summary of the analysis
summary(res.aov)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
group        2  3.766  1.8832   4.846 0.0159 *
Residuals   27 10.492  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpret the result of one-way ANOVA tests As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the groups highlighted with “*" in the model summary.