ANOVA Test

The one-way analysis of variance (ANOVA), also known as one-factor ANOVA, is an extension of independent two-samples t-test for comparing means in a situation where there are more than two groups. In one-way ANOVA, the data is organized into several groups base on one single grouping variable (also called factor variable). This tutorial describes the basic principle of the one-way ANOVA test and provides practical anova test examples in R software.

Assumptions of ANOVA test

- The observations are obtained independently and randomly from the population defined by the factor levels
- The data of each factor level are normally distributed.
- These normal populations have a common variance. (Levene's test can be used to check this.)

ANOVA test hypotheses:

- Null hypothesis, (H_0): the means of the different groups are the same
- Alternative hypothesis (H_1): At least one sample mean is not equal to the others.

Note that, if you have only two groups, you can use t-test. In this case the F-test and the t-test are equivalent.

How one-way ANOVA test works?

Assume that we have 3 groups ($ A, B, C$) to compare:

1. Compute the common variance, which is called variance within samples (S^2_{within}) or residual variance.
2. Compute the variance between sample means as follow:

- Compute the mean of each group
- Compute the variance between sample means ($S^2_{between}$)

3. Produce F-statistic as the ratio of $S^2_{between}/S^2_{within}$.

Note that, a lower ratio (ratio < 1) indicates that there are no significant difference between the means of the samples being compared. However, a higher ratio implies that the variation among group means are significant.

Packges used

- dplyr
- ggplot2
- ggpubr
- car

Import data into R

Here, we'll use the built-in R data set named PlantGrowth. It contains the weight of plants obtained under a control and two different treatment conditions. There are still various methods to import data into R.

{r, import data} my_data <- PlantGrowth

Creating a summary based on group

{r} library(dplyr) group_by(my_data, group) %>% summarise(count = n(), mean = mean(weight, na.rm = TRUE), sd = sd(weight, na.rm = TRUE))

Visualizing Data

{r} library(“ggpubr”) ggboxplot(my_data, x = “group”, y = “weight”, color = “group”, palette = c(“red”, “green”, “blue”), order = c(“ctrl”, “trt1”, “trt2”), ylab = “Weight”, xlab = “Treatment”)

Compute one-way ANOVA test

** Objective** : We want to know if there is any significant difference between the average weights of plants in the 3 experimental conditions.

The R function aov() can be used to answer to this question. The function summary.aov() is used to summarize the analysis of variance model.

{r} # Compute the analysis of variance res.aov <- aov(weight ~ group, data = my_data) # Summary of the analysis summary(res.aov)

The output includes the columns F value and Pr(>F) corresponding to the p-value of the test.

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.

Multiple pairwise-comparison between the means of groups

n one-way ANOVA test, a significant p-value indicates that some of the group means are different, but we don’t know which pairs of groups are different.

It’s possible to perform multiple pairwise-comparison, to determine if the mean difference between specific pairs of group are statistically significant.

Tukey multiple pairwise-comparisons

As the ANOVA test is significant, we can compute Tukey HSD (Tukey Honest Significant Differences, R function: TukeyHSD()) for performing multiple pairwise-comparison between the means of groups.

The function TukeyHD() takes the fitted ANOVA as an argument.

{r} TukeyHSD(res.aov)

- diff: difference between means of the two groups
- lwr, upr: the lower and the upper end point of the confidence interval at 95% (default)
- p adj: p-value after adjustment for the multiple comparisons.

It can be seen from the output, that only the difference between trt2 and trt1 is significant with an adjusted p-value of 0.012.

Check ANOVA assumptions: test validity?

The ANOVA test assumes that, the data are normally distributed and the variance across groups are homogeneous. We can check that with some diagnostic plots.

Check the homogeneity of variance assumption

The residuals versus fits plot can be used to check the homogeneity of variances.

In the plot , there is no evident relationships between residuals and fitted values (the mean of each groups), which is good. So, we can assume the homogeneity of variances.

Residual plot

{r} # 1. Homogeneity of variances plot(res.aov, 1)

Points 17, 15, 4 are detected as outliers, which can severely affect normality and homogeneity of variance. It can be useful to remove outliers to meet the test assumptions.

Check for normality

I recommend Levene's test, which is less sensitive to departures from normal distribution. The function leveneTest() [in car package] will be used:

{r,warning=FALSE} library(car) leveneTest(weight ~ group, data = my_data)

From the output above we can see that the p-value is not less than the significance level of 0.05. This means that there is no evidence to suggest that the variance across groups is statistically significantly different. Therefore, we can assume the homogeneity of variances in the different treatment groups.

Relaxing the homogeneity of variance assumption

The classical one-way ANOVA test requires an assumption of equal variances for all groups. In our example, the homogeneity of variance assumption turned out to be fine: the Levene test is not significant.

Welch one-way test

An alternative procedure (i.e.: Welch one-way test), that does not require that assumption have been implemented in the function oneway.test(). This model is given in the following code chunk/

{r} oneway.test(weight ~ group, data = my_data)

Pairwise t-tests with no assumption of equal variances

{r} pairwise.t.test(my_data$weight, my_data$group, p.adjust.method = "BH", pool.sd = FALSE)

Check the normality assumption

Normality plot of residuals. In the plot below, the quantiles of the residuals are plotted against the quantiles of the normal distribution. A 45-degree reference line is also plotted.

The normal probability plot of residuals is used to check the assumption that the residuals are normally distributed. It should approximately follow a straight line.

{r} # 2. Normality plot(res.aov, 2)

As all the points fall approximately along this reference line, we can assume normality.

Shapiro-Wilk test on the ANOVA residuals

The conclusion above, is supported by the Shapiro-Wilk test on the ANOVA residuals (W = 0.96, p = 0.6) which finds no indication that normality is violated.

{r} # Extract the residuals aov_residuals <- residuals(object = res.aov) # Run Shapiro-Wilk test shapiro.test(x = aov_residuals)

What to do if ANNOVA assumptions are not met?

Note that, a non-parametric alternative to one-way ANOVA is Kruskal-Wallis rank sum test, which can be used when ANNOVA assumptions are not met.

{r} kruskal.test(weight ~ group, data = my_data)

{r} # library(xtable) # print(xtable(res.aov), type="latex")

Take aways from this session

- Import your data from a .txt tab file: my_data <- read.delim(file.choose()). Here, I used my_data <- PlantGrowth.
- Visualize your data: ggpubr::ggboxplot(my_data, x = "group", y = "weight", color = "group")
- Compute one-way ANOVA test: summary(aov(weight ~ group, data = my_data))
- Tukey multiple pairwise-comparisons: TukeyHSD(res.aov)
- non-parametric test alternative Kruskal-Wallis-rank-sum-test