ANOVA in R

Kaloyan Ganev

2022/2023

Lecture contents

1. Introductory notes

2. Basics of One-way ANOVA

3. Basics of Non-parametric ANOVA

Introductory notes

What is ANOVA?

- Decrypted as “Analysis of Variance”
- A class of statistical models designed to study the existence of significant differences between means of groups of data
- In this respect it is analogical to the t-test
- Also, it generalizes the essence of the t-test to more than two groups
- However, it differs in an important aspect: to test the significance of this difference, a comparison of variances is made

What is ANOVA? (2)

- Main idea: to model a random variable
- However, modelling is not carried out through estimating the parameters of its conditional distribution
- Instead, it is carried out through using known covariates
- Three or more groups of data are needed to perform the analysis
- Usually populations are assumed to be normally distributed but this is not mandatory
- An F-statistic is used to make inference

Basics of One-way ANOVA

The ANOVA Null Hypothesis

- Assume there is a single quantitative (i.e. non-qualitative) variable
- Also called *response* variable
- There are many samples on this variable, say k
- Each sample either possesses a quality/is subject to a *treatment* or not[1]
- The goal is to test the hypothesis that the populations from which the samples are drawn have equal means
- Formally written,

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \ldots = \mu_k \\ H_1: &\quad \text{At least one } \mu \text{ is different} \end{aligned}$$

[1] This implies a quality, or factor variable.

Within and Between Estimation

- Key requirement for the validity of the test: All populations have the same variance!
- This variance is unknown and has to be estimated first before used in hypothesis testing
- Two methods are used for the purpose: *within* and *between* estimation
- Within estimation is used to produce a valid estimate of the unknown common variance irrespective of whether the population means are equal or not
- Between estimation explicitly assumes equality of population means and also aims at such an estimate

Within and Between Estimation (2)

- After the two types of estimates are produced, their ratio is calculated

$$Ratio = \frac{\text{Between estimate}}{\text{Within estimate}}$$

- Under the null hypothesis, this ratio has the F distribution
- When the means are not equal, the ratio will have a large value
- As a consequence, the null will be rejected
- Since the ratio is always positive (why?), this is a one-tailed test

The Within Method

- In this method, each observation is compared with its own sample's mean
- To formalize, denote by n the number of observations in each sample
- Denote by X_{ij} the ith observation ($i = 1, 2, \ldots, n$) from sample j, $j = 1, 2, \ldots, k$
- Denote by $\overline{X}_j$ the mean of sample j, $j = 1, 2, \ldots, k$
- Then, the within estimator of the variance is

$$s_w^2 = \frac{\sum_{j=1}^{k}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2\right]}{k}$$

The Within Method (2)

- Note that I wrote the formula in an a bit awkward way so that you grasp the essence better
- Usually it is written as

$$s_w^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2}{k(n-1)}$$

- The sum in the numerator is called the *sum of squares within*, i.e.

$$SS_w = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$$

- The denominator contains the degrees of freedom (k samples/groups times $n-1$ degrees of freedom in each of them)

The Between Method

- Recall the Central Limit Theorem
- It states that the distribution of sample means tends to the normal distribution as sample size grows
- The parameters of the normal distribution are respectively μ and $\frac{\sigma^2}{n}$
- The variance, $\frac{\sigma^2}{n}$, measures the variability of values of the means of the samples around the true population mean

The Between Method (2)

- The first step in the between method is to estimate the population mean; denote it by $\overline{\overline{X}}$
- The latter is also called *the grand mean*
- This is done by using all observations from all samples/groups
- The estimator of the variance of the sample mean is

$$s^2_{\overline{X}} = \frac{\sum\limits_{j=1}^{k} (\overline{X}_j - \overline{\overline{X}})^2}{k - 1}$$

- The sum in the numerator is called *the sum of squares between*, i.e.

$$SS_b = \sum_{j=1}^{k} (\overline{X}_j - \overline{\overline{X}})^2$$

The Between Method (3)

- We know that the true $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$
- From this follows

$$\sigma^2 = n\sigma^2_{\overline{X}}$$

- In the sample context, this becomes

$$s^2 = ns^2_{\overline{X}}$$

- Thus, the between estimator of the variance is

$$s^2_b = \frac{n\sum\limits_{j=1}^{k}(\overline{X}_j - \overline{\overline{X}})^2}{k-1}$$

The ANOVA F test

- Formed as:

$$F = \frac{s_b^2}{s_w^2} \sim F(k-1, k(n-1))$$

- The computed value of the statistic (the estimate) is then compared to the critical value from the F table
- If the critical value is exceeded, the null hypothesis is rejected
- We will take an example in R

Basics of Non-parametric ANOVA

Non-parametric ANOVA: The Kruskal-Wallis Test

- Assume that there are k groups of data (treatment levels)
- Assume also that in each group the data have (approximately) equal variances
- However, the distribution of the data is *not* normal
- Suppose that those data cannot be transformed in a sensible way so that they become normal
- In such a situation, classical ANOVA would not be appropriate
- The Kruskal-Wallis test is the alternative that can be used
- (The test extends the two-sample Wilcoxon test)

Non-parametric ANOVA: The Kruskal-Wallis Test (2)

- Let the total number of observations be n
- There are k groups, and n_j, $j = 1, 2, \ldots, k$, observations in each of them
- Obviously, $n = \sum_j n_j$
- The Kruskal-Wallis test statistic is[2]

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

where R_j is the sum of the ranks for group j

- The statistic is valid only if less than 1/4 of the observations are ties.[3]

[2] A correction to this formula is applied if there are ties. See https://en.wikipedia.org/wiki/Kruskal-Wallis_one-way_analysis_of_variance.

[3] Explanation of ties will follow.

Non-parametric ANOVA: The Kruskal-Wallis Test (3)

- To determine ranks, all observations are sorted in an ascending order; ranks match orderings
- If two values are equal, the ranks they occupy are summed and then the sum is divided by the number of values
- For example, if the smallest two values are 2 and 2, the sum of their ranks is 3, therefore each receives a rank of 1.5
- The null hypothesis is the same as in standard ANOVA
- If $n_j \geq 5, \forall j$, under the null, the distribution of the test statistic is

$$H \sim \chi^2(k-1)$$

- If the computed statistic exceeds the critical value under the chosen level of significance, the null is rejected
- An example in R follows