Analysis of Variance (ANOVA)

XDASI Fall 2021

November 11, 2021

Contents

Suggested reading

- Whitlock & Schluter, Ch 15 and online lab
- Online tutorials
  - Antoine Soetewey (UCLouvain, Belgium) - Stats and R blog: ANOVA (2020-10-02)
  - (Steven Doogue, 2019-07-09) - Chapter 7.1: One-way ANOVA

Introduction

We've previously used t-tests and non-parametric methods to compare two samples. What if we need to compare more than two samples? The problem with just performing multiple t-tests is that each test has a certain Type I error rate, and when we perform multiple tests this error simply compounds. This issue can be addressed by ***correction for multiple hypothesis testing***, which we will cover later.

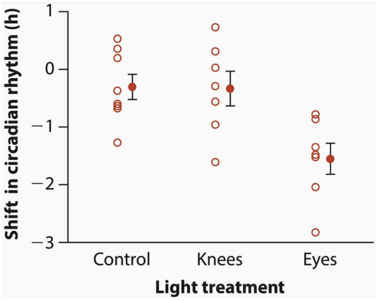
If doing all pairwise t-tests between groups is not a good approach, what can we do instead? ANOVA to the rescue! ANOVA extends t-tests to more than two groups by allowing the comparison of the means between multiple groups. However, instead of simply comparing the groups using the difference in their means and their SD, ANOVA compares the ***overall variation between groups*** to the ***variation within each group***. If the overall variation is significantly greater than the individual variation, then we consider the groups to be different.

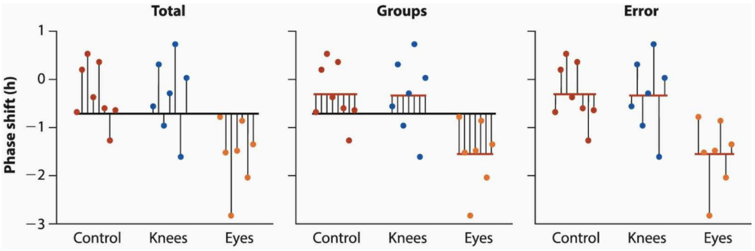
The simplest form of ANOVA is ***one-way ANOVA***, which we will discuss here.

Assumptions

Like t-tests, ANOVA assumes that the distributions of the samples are relatively normal, and also that their variances are similar. When these do not hold, a non-parametric analog ANOVA can be performed called the ***Kruskall-Wallace test***.

The basic idea is illustrated in Figures 15.1-1 and 15.1-2:





To determine whether there is a significant difference between groups, the variation is decomposed into three parts:

- Total variation with respect to the grand mean
- Variation between group means and the grand mean
- Variation within each group w.r.t. its group mean

These differences are calculated using the ***sum of squares*** of the difference between each data point and the mean used for comparison:

- Total variation: $SST = \sum_{i=1}^{n}$

The calculations that are involved are summarized in Table 10.1 from Ken Aho's book:

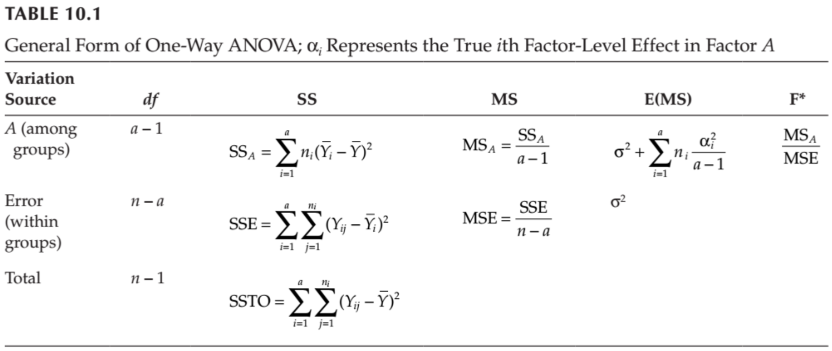


TABLE 10.1

General Form of One-Way ANOVA; α_i Represents the True *i*th Factor-Level Effect in Factor *A*

Variation Source	*df*	SS	MS	E(MS)	F*
A (among groups)	$a-1$	$SS_A = \sum_{i=1}^{a} n_i(\bar{Y}_i - \bar{Y})^2$	$MS_A = \frac{SS_A}{a-1}$	$\sigma^2 + \sum_{i=1}^{a} n_i \frac{\alpha_i^2}{a-1}$	$\frac{MS_A}{MSE}$
Error (within groups)	$n-a$	$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$	$MSE = \frac{SSE}{n-a}$	σ^2	
Total	$n-1$	$SSTO = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2$			

The ***total sum of squares (TSS)*** equals the ***within-group SS (SSW)***, also called the ***error SS (SSE)***, plus the ***between-group SS (SSB or SSG)***:

$$SST = SSW + SSB$$

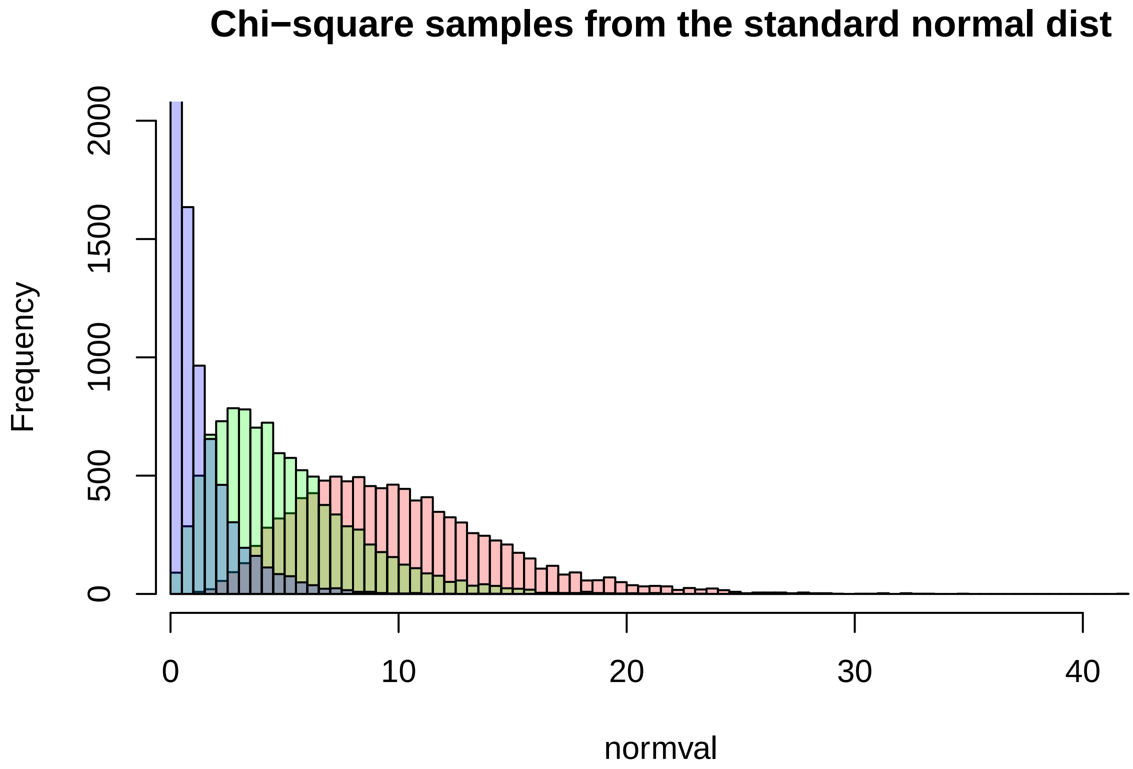
As usual, we need to take into account the ***degrees of freedom***, which is just the number of data points in each equation minus one (the mean used for comparison).

The χ^2 distribution

Since we are measuring differences using ***sums of squares***, the differences will follow a χ^2 distribution, which represents the **sum of squared random values** selected from a **normal distribution**. The degrees of freedom, k is simply the number of random values.

$$Q = \sum_{i=1}^{k} Z_i^2$$

χ^2 with different sample sizes Let's simulate some data to see what it looks like when k = 10, 5, and 1 and the values are retrieved from a standard normal distribution. (We have also done this in a previous class.)



It is clear to see that as k increases, the distribution begins to look like a **normal distribution**.

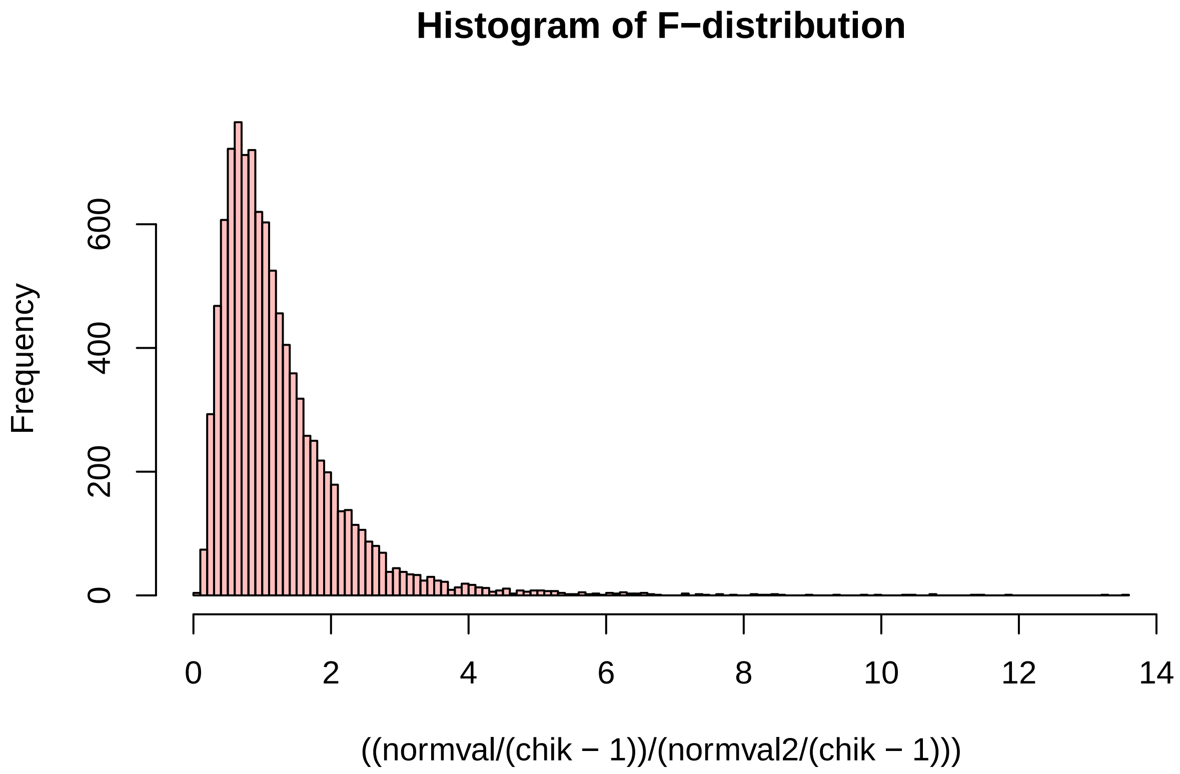
This only happens when the **sample size is at least 10**, which is why it is not recommended to use the χ^2 test for small values of k (<10).

The F-statistic

To ***compare two*** χ^2 ***distributions***, we can simply take a ***ratio*** of them (taking into account their respective degrees of freedom).

This distribution is called an F**-distribution** and the ***ratio*** is the F**-statistic**. (The F-distribution is named after the statistician Ron Fisher.)

F-distribution for samples from the same population Let's first see what it would look like if the the means of the two populations that are being sampled are **equal**.

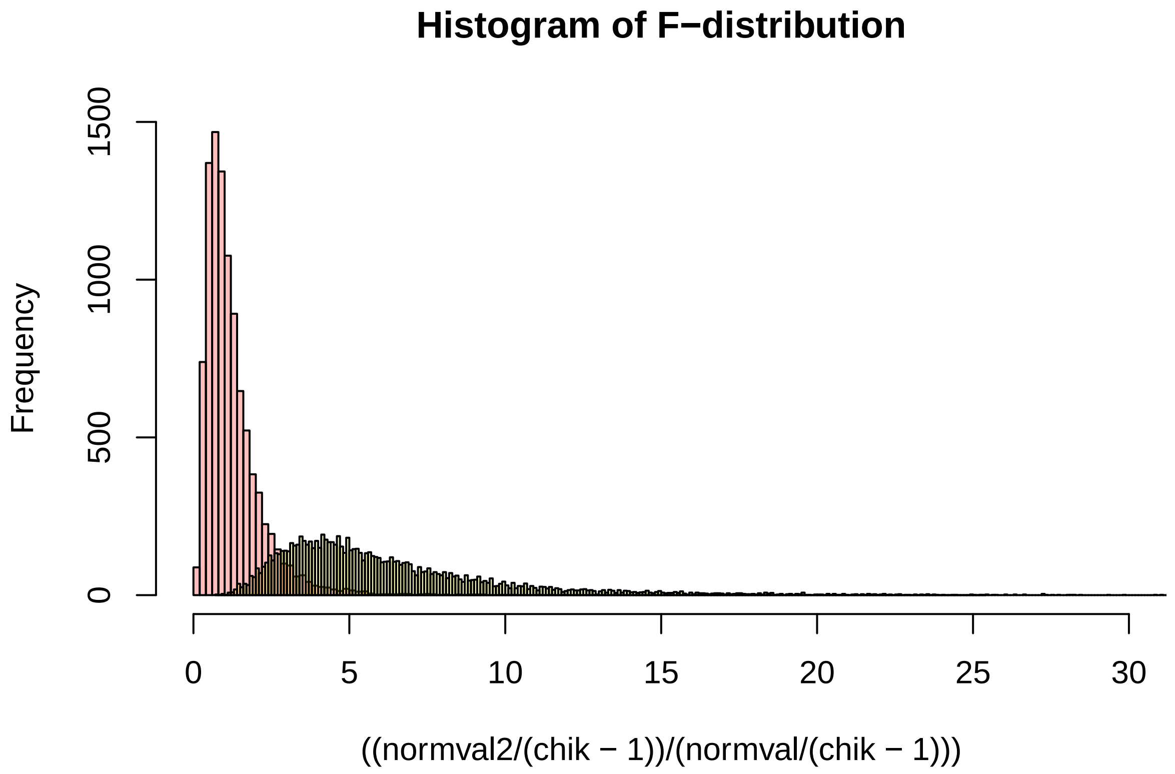


Note that depending on how close most of the density is to the mean, the heights of two histograms will vary, even if they are generated using exactly the same procedure. Run the above code several times and see how the results change each time.

F-distribution for populations with different means What if the means of our normal distributions are different?

We can make a second histogram showing the same ratio for data sampled from two normal distributions with different means: the standard normal and a normal distribution with mean = 2 and sd = 1.

Now, the ratio of the sums of the two samples will look quite different. Let's try this and superimpose the two histograms for comparison.



Remember that ***variance*** is essentially a sum of squares as discussed above. So now we have the ability to compare two different variances and use a statistic to determine if they are significantly different.