Statistical Inference with ANOVA

Dr. Marko Mitic

The one-way analysis of variance (ANOVA) test checks to see if there is a difference between the values of one quantitative variable that you collect from more than two groups. It's like a three-or-more-sample t-test, and it's particularly useful for analyzing data from experiments where you're evaluating changes in a response variable related to a particular treatment. Our problem can be described as follows: you are a trainer or instructor for a certification program and you want to compare three different textbooks to see which one you should use. You randomly assign each of your students to a textbook, and you keep track of the scores they get on the last practice certification exam they take before they try the real thing. Is one of the textbooks more effective than the others? (A two-way ANOVA could explore the impact of each of the three textbooks and two different instructors.)

Before we start our analysis, we need to check initial assumptions:

- Random sample
- Observations are independent
- Response variable is normal or nearly normal
- The sample size for each treatment group is the same
- Homogeneity of variances

If your assumptions check out, we can set the null hypothesis which indicate that all means of the treatmens groups are equal. Alternative hypothesis suggest that at least one of the means is different. We will set more rigorous significance level - alpha = 0.01, because we want to be almost apsolutely positive in the choice of appropriate textbook; e.g. we want our students to have the best studing source out there.

For ANOVA, because we are slicing and dicing variances within and between our groups, our test statistic will be a F. It's calculated by taking the mean square error between the groups and dividing by the mean square error within the groups.

Once we compute SSB (sum of squares between the groups), SSW (sum of squares within the groups), and the SST (the sum of SSB and SSW), the remainder of the calculations are straightforward. For this example, the data is obtained from Prof. Nicole Radziwill github profile.

```r
# install.packages("RCurl")
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 3.2.1
```

```
## Loading required package: bitops
```

```r
url = "https://raw.githubusercontent.com/NicoleRadziwill/Data-for-R-Examples/master/anova-textbooks.txt
score.data = getURL(url,ssl.verifypeer=FALSE)
scores = read.table(text=score.data,header=T)
```

Let's start with SST. Firstly, take the difference between each data point and the grand mean (the mean of all the points in your dataset), and square it. Secondly, add all those up.

```r
mean(scores$score)
```

```
## [1] 85.64583
```

```
dev.from.grand.mean = (scores$score - mean(scores$score))^2

dev.from.grand.mean
```

```
##  [1]  58.4587674  13.2921007 178.3337674  13.2921007 135.6254340
##  [6]  28.6671007  40.3754340  18.9587674   5.5421007   2.7087674
## [11] 186.2087674   7.0004340   5.5421007   1.8337674 113.3337674
## [16]  13.2921007   0.1254340 186.2087674   0.4171007  18.9587674
## [21]  28.6671007  13.2921007   7.0004340  21.5837674  93.0421007
## [26] 113.3337674  21.5837674  54.0837674   0.1254340 186.2087674
## [31]   1.8337674  18.9587674  11.2504340  74.7504340   5.5421007
## [36] 152.6254340  54.0837674 107.2087674   5.5421007  28.6671007
## [41] 113.3337674   5.5421007  11.2504340  54.0837674  44.1671007
## [46] 128.9171007 178.3337674  69.7921007
```

```
sst = sum(dev.from.grand.mean) # This is your SST!!!

sst
```

```
## [1] 2632.979
```

Calculating the SSB value is little more complicated. Firstly, we find means of each of the groups.

```
group.means = aggregate(scores$score, by=list(scores$textbook), FUN=mean)
group.means
```

```
##   Group.1       x
## 1       1 84.1875
## 2       2 83.1250
## 3       3 89.6250
```

```
#or

tapply(scores$score, scores$textbook, FUN=mean)
```

```
##       1       2       3
## 84.1875 83.1250 89.6250
```

Now we can calculate all the deviations between the group means and the grand means, and go through the squaring and summation process:

```
group.dev.from.grand.mean = ((group.means$x - mean(scores$score))^2)*16
```

```
ssb <- sum(group.dev.from.grand.mean)
ssb
```

```
## [1] 389.0417
```

The next step on our path is for us to use SSB, SSW, and SST to calculate the mean squares: MSB and MSW. COnsequently, we calculate F score:

```r
msb = ssb / 2 # The denominator is (# of groups - 1)
ssw = sst - ssb
msw = ssw / (48-3)

msb
```

```
## [1] 194.5208
```

```r
msw
```

```
## [1] 49.86528
```

```r
F = msb/msw
F
```

```
## [1] 3.900927
```

Finally, the p-value is calculated as:

```r
1-pf(3.9,2,45)
```

```
## [1] 0.0274169
```