# ANOVA Examples

Luis Cárceles

2024-04-11

Examples taken from "Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences" By J. Susan Milton and Jesse Arnold, chapter 13.

## Introduction

To test that $k$ samples belong to he same population or come from different populations, we build the following hypothesis test:

- $H_0$: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$
- $H_1$: $\mu_i \neq \mu_j$ for some $i$ and $j$

### Statistic to be used in the test:

We examine first the vulnerabilities in the next way. First, we define the total variability as:

$$SS_{tot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y_{..}})^2$$

The mean of the variability of the data due the fact of different treatments, or samples, is defined as:

$$SS_{tr} = \sum_{i=1}^{k} n_i (\bar{Y_{i.}} - \bar{Y_{..}})^2$$

And the variability of the data inside the same treatment:

$$SS_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y_{i.}})^2$$

The next equation is verified:

$$SS_{tot} = SS_{tr} + SS_E$$

From these relations we can define the mean square error of the treatment:

$$MSS_{tr} = SS_{tr}/(k-1)$$

And the mean square error:

$$MSS_E = SS_E/(N-k)$$

To the null hypothesis: $H_0$: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$, we use the statistic:

$$F_{k-1,N-k} = MSS_{tr}/MSS_E$$

Which has a $F$ probability distribution with $k-1$ and $N-k$ degrees of freedom.

### Shortcuts to calculate the square sums

The square sums can also be calculated with the following formulas:

$$SS_{tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{T_{..}^2}{N}$$

$$SS_{tr} = \sum_{i=1}^{k} \frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{N}$$

$$SS_E = SS_{tot} - SS_{tr}$$

### Example 1

From five different veins of coal extracted from the same geographic area, we are given with the attached data of the sulfur contents in the samples. Can we distinguish among all those five veins or we must say that they are identical?

**Manual solution** Data collection:

```
Y1=c(1.51,1.92,1.08,2.04,2.14,1.76,1.17)
Y2=c(1.69,0.64,0.90,1.41,1.01,0.84,1.28,1.59)
Y3=c(1.56,1.22,1.32,1.39,1.33,1.54,1.04,2.25,1.49)
Y4=c(1.30,0.75,1.26,0.69,0.62,0.90,1.20,0.32)
Y5=c(0.73,0.80,0.90,1.24,0.82,0.72,0.57,1.18,0.54,1.3)
Y=c(Y1,Y2,Y3,Y4,Y5)
```

Preliminary Calculations

```
k=5 #Number of treatments
N=length(Y) #Total number of data
#number of data in each treatment
n1<-length(Y1)
n2<-length(Y2)
n3<-length(Y3)
n4<-length(Y4)
n5<-length(Y5)
Tot<-sum(Y)
SStot<-sum(Y*Y)-Tot^2/N
invn<-c(1/n1, 1/n2, 1/n3, 1/n4, 1/n5)
Tots<-c(sum(Y1), sum(Y2), sum(Y3), sum(Y4), sum(Y5))
Tsqrd<-Tots*Tots
SStr<-sum(Tsqrd*invn) - Tot^2/N
SSE=SStot-SStr
```

Statistic calculations:

```
MSStr<-SStr/(k-1)
MSSE<-SSE/(N-k)
#Statistic F, with k-1 and N-k degrees of freedom
Fs<-MSStr/MSSE
sprintf("The value of the F statistic is: %f", Fs)
```

```
## [1] "The value of the F statistic is: 8.094810"
```

P-value calculation

```
p_value<-1-pf(Fs, df1=k-1, df2=N-k, lower.tail = TRUE, log.p = FALSE)
sprintf("The p-value of the statistic Fs is: %f", p_value)
```

```
## [1] "The p-value of the statistic Fs is: 0.000086"
```

In this case, we must reject the null hypothesis (all means are equal). There are sufficient evidence to support alternative hypothesis, our claim. We are dealing with samples of different populations.

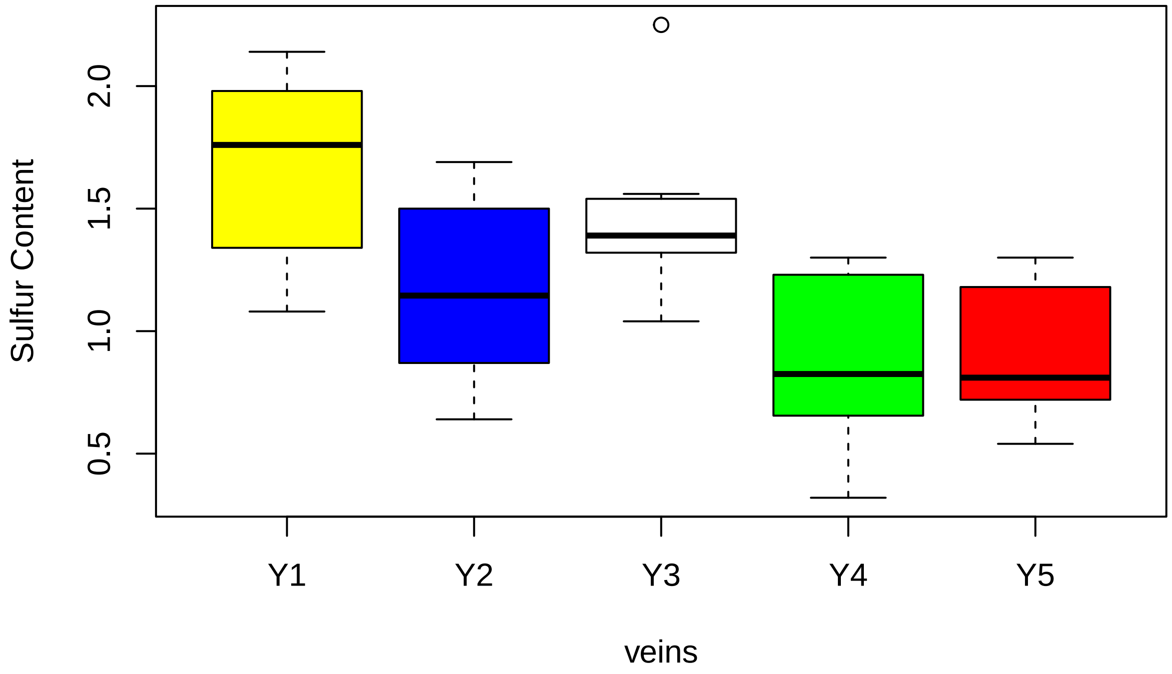
**Solution using R built in functions** We classify every data point into a factor (vein sample that it belongs to):

```
veins<-as.factor(rep(c("Y1", "Y2", "Y3", "Y4", "Y5"), c(n1, n2, n3, n4, n5)))
veins
```

```
##  [1] Y1 Y1 Y1 Y1 Y1 Y1 Y1 Y2 Y2 Y2 Y2 Y2 Y2 Y2 Y2 Y3 Y3 Y3 Y3 Y3 Y3 Y3 Y3 Y3 Y4
## [26] Y4 Y4 Y4 Y4 Y4 Y4 Y4 Y5 Y5 Y5 Y5 Y5 Y5 Y5 Y5 Y5 Y5
## Levels: Y1 Y2 Y3 Y4 Y5
```

Make a box plot to observe the differences:

```
boxplot(Y ~ veins, col = c("yellow", "blue", "white","green", "red"), ylab = "Sulfur Content")
```



Do the analysis of the variance (ANOVA)

```r
aovtest=aov(lm(Y ~ veins))
summary(aovtest)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## veins        4  3.935  0.9838   8.095 8.57e-05 ***
## Residuals   37  4.497  0.1215                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```