One factor ANOVA using R

Pramudita Satria Palar, Ph.D.

23/04/2022

One factor ANOVA

Analysis of variance (ANOVA) is one of the most important statistical tool that is widely used in industry and research. In this tutorial, we will learn how to use the R built-in `aov()` function to perform ANOVA. Although we will demonstrate the usefulness of `aov()` only for a single factor experiment, the same function can be used for many types of ANOVA. Furthermore, once you learn how to use `aov()`, it is actually pretty easy to use it for other types of ANOVA.

The single factor ANOVA assumes that an observation Y is generated from the following linear statistical model:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{1}$$

for $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, n$, where μ is the overall mean, τ_i is the $i-$th treatment effect, and ε_{ij} is a random error. Notice that we assume each treatment i (where a is the number of treatments) has n observations. The random error ε_{ij} is assumed to be independent and normally distributed with $\mathcal{N}(0, \sigma^2)$ (that is, zero mean and variance described by σ^2)

With single-factor ANOVA, assuming that each treatment has the same number of observations, we are testing the following hypotheses:

- $\mathcal{H}_0 : \tau_1, \tau_2, \ldots, \tau_a = 0$
- $\mathcal{H}_a : \tau_i \neq 0$ for at least one i.

The null hypothesis simply states that the treatment does not affect the dependent variable. If we reject the null hypothesis, then it means we accept the null hypothesis that the treatment yields an observable effect for at least one treatment.

The ANOVA identity states that the total sum of squares (SS_T) can be decomposed into the treatment sum of squares ($SS_{\text{treatments}}$) and error sum of squares (SS_{error}) as follows:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2, \tag{2}$$

$$SS_T = SS_{\text{treatments}} + SS_{\text{error}} \tag{3}$$

where

- $y_{i.} = \sum_{j=1}^{n} y_{ij}$ (total of the observations under the i-th treatment)
- $y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij})$ (Grand total of all observations)
- $\bar{y}_{i.} = y_{i.}/n$ (average of the observations under the i-th treatment)
- $\bar{y}_{ii} = y_{..}/N$ (grand mean of all observations)

The degrees of freedom can also be partitioned as

$$an - 1 = a - 1 + a(n-1) \tag{4}$$
$$df_{\text{total}} = df_{\text{treatments}} + df_{\text{error}} \tag{5}$$

We can then calculate the mean square for treatmens and mean square for error, respectively, as follows:

$$MS_{\text{treatments}} = \frac{SS_{\text{treatments}}}{a-1} \tag{6}$$
$$MS_{\text{error}} = \frac{SS_{\text{error}}}{a(n-1)} \tag{7}$$

The expected value for $MS_{\text{treatments}}$ and MS_{error} are, respectively, as follows:

$$\mathbb{E}(MS_{\text{treatments}}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1} \tag{8}$$
$$\mathbb{E}(MS_{\text{error}}) = \sigma^2 \tag{9}$$

We can show if the $\mathcal{H}_0$ is true by ANOVA F-test. The statistic that we test is

$$F_0 = \frac{SS_{\text{treatments}}/(a-1)}{SS_{\text{error}}/[a(n-1)]} = \frac{MS_{\text{treatments}}}{MS_{\text{error}}}. \tag{10}$$

We reject $\mathcal{H}_0$ if $f_0 > f_{\alpha,a-1,a(n-1)}$ (where $a-1$ and $a(n-1)$ are the degrees of freedom for the F-distribution, i.e., ν_1 and ν_2, respectively), where f_0 is the computed value of F_0.

Just in case you forget the F-distribution, the following figure shows the F-distribution for various degrees of freedom:

Let's do the ANOVA now.

Using synthetic data

Creating the data

We will begin with a synthetic data set. Imagine that we want to investigate whether the daily duration of self-study affects the examination score or not. For that, we collected the data of five observations per group, with the treatments including 2, 4, and 6 hours of study duration (so that we have a total of 15 data). We will need two variables, so we save them in two R variables (namely `score` and `duration`):

score	duration
60	2
70	2
65	2
66	2
55	2
70	4
74	4
75	4
68	4
69	4
79	6

score	duration
82	6
84	6
86	6
91	6

```
score <- c(60,70,65,66,55,70,74,75,68,69,79,82,84,86,91)
duration <-c(2,2,2,2,2,4,4,4,4,4,6,6,6,6,6)
```

It is always recommended to plot the data using boxplot first to see the distribution of our data. We will use `boxplot()` for that:

```
boxplot(score~duration, xlab="Duration (hours)", ylab = "Score")
```

The `boxplot()` function is used with the formula written such that the data values (i.e., `score`) are splitted according to the grouping variable (i.e., `duration`)

From the visual observation, we can see that changing the amount of treatment yields a notable effect on the score. First, however, we need to apply a statistical hypothesis test to answer the question formally and statistically. Let's use ANOVA to do just that.

Performing ANOVA on synthetic data

Once the data is ready, using ANOVA in R is surprisingly easy. We will use `aov()` with the detail is as follows:

```
aov_result <- aov(formula = score~factor(duration))
```

The `aov()` function takes at least one input, namely the `formula`. The syntax for the formula should be written such that `xx~factor(yy)`, where `yy` is the treatment variable (duration) and `x` is the dependent variable (score).

Try to print the summary of result:

```
summary(aov_result)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## factor(duration)  2 1146.1   573.1   26.99 3.62e-05 ***
## Residuals        12  254.8    21.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Shown in the summary are as follows:

- **Df** shows the degrees of freedom for the independent (treatment) and dependent variable.
- **Sum Sq** shows the sum of squares.
- **Mean Sq** shows the mean of the sum of squares, that is, the sum of squares divided by the corresponding degree of freedom.

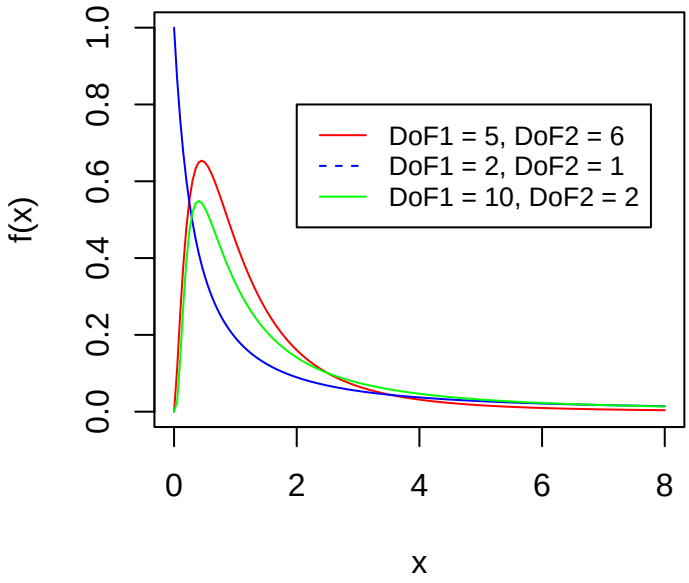

Figure 1: Examples of F-distribution

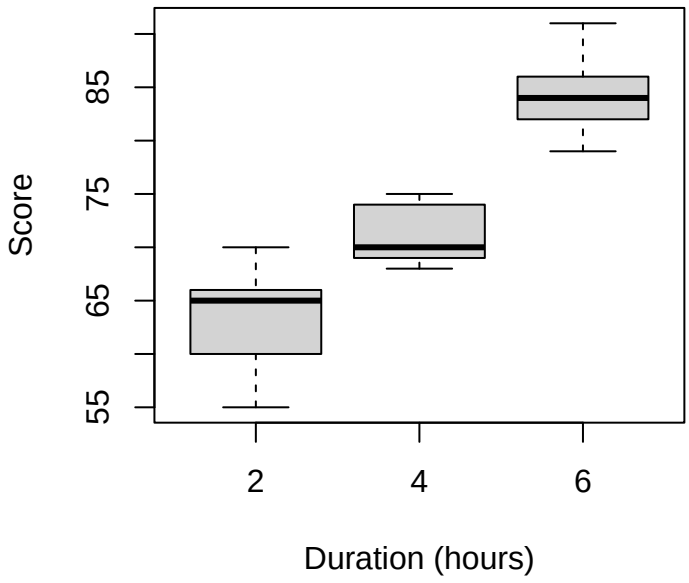

Figure 2: Distribution of the score-duration data in the form of boxplot

- **F value** is the test statistic. as discussed earlier.
- **Pr(>F)** is the p-value of the computed F-statistic.
- **Signif. codes** shows how significant is the result from the hypothesis test.

In our current example, the **Pr(>F)** is extremely low (i.e., $Pr(> F) = 3.62 \times 10^{-5}$) with the significant code is ***, which means that the p-value is very small. In other words, we accept the alternative hypothesis that there is any group that differs significantly from the overall group mean (in plain English, the treatment yields observable effects!)

Using data in the dataframe format

Importing the data from CSV

Sometimes it is convenient to import the data from other sources (e.g., CSV or Microsoft Excel format). We will import our data from a CSV file for this tutorial.

Let us begin with uploading the CSV file. We can use `read.csv()` to do that. The following snippet import the data from `hardwood_concentration_two_col.csv` and automatically save into a data frame format (let us name it `DF`):

```
DF <- read.csv('hardwood_concentration_two_col.csv')
```

As usual, you can see the inside of `DF` by using the dollar sign (`$`). For example, to see the Tensile Strength:

```
DF$TS
```

```
##  [1]  5  5  5  5  5  5 10 10 10 10 10 10 15 15 15 15 15 15 20 20 20 20 20 20
```

and to see the Hardwood concentration

```
DF$HC
```

```
##  [1]  7  8 15 11  9 10 12 17 13 18 19 15 14 18 19 17 16 18 19 25 22 23 18 20
```

Alternatively, you can type `View(DF)` in the console to view `DF` in a spreadsheet-like display.

As usual, let's plot our data in the form of boxplot:

```
boxplot(HC~TS,data=DF)
```

The `data()` argument means that we take the data for plotting from `DF`.

Performing the ANOVA

Just like our previous example, we will now use the `aov()` function with some tweaks:

```
aov_df <- aov(formula=HC~factor(TS),data=DF)
```

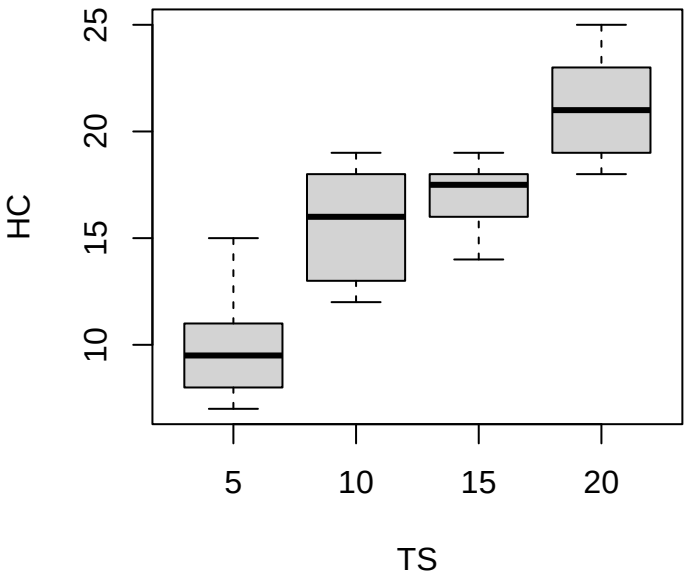

Figure 3: Distribution of the tensile strength-hardwood concentration data in the form of boxplot

See the snippet above. The `data` argument defines the variable (in data frame format), that will we analyze using ANOVA. On the other hand, the `formula` argument defines the formula specifying the model. For example, the independent and dependent variables for our hardwood concentration data are tensile strength and hardwood concentration itself; our formula is then `HC~factor(TS)`.

Finally, let's see the result by using the `summary()` function on `aov_df`:

```
summary(aov_df)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## factor(TS)   3  382.8  127.60   19.61 3.59e-06 ***
## Residuals   20  130.2    6.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Surely now you know how to interpret the result from ANOVA. In essence, for this problem, we reject the null hypothesis that there is no difference in means. Instead, we accept the alternative hypothesis that the treatment (tensile strength) affects the hardwood concentration.

The non-significant difference case

Lastly, let us try again applying ANOVA but for a case in which ANOVA yields a non-significant difference. We will use another synthetic data set for this purpose:

```r
treatment <- c("a","a","a","a","a","b","b","b","b","b","c","c","c","c","c")

target <- c(8,9,8,10,7,10,11,8,7,6,13,9,7,8,10)
```

with `treatment` and `target` are just generic names for our variables. We visualize the data in the form of boxplot as follows:

```r
boxplot(target~treatment)
```

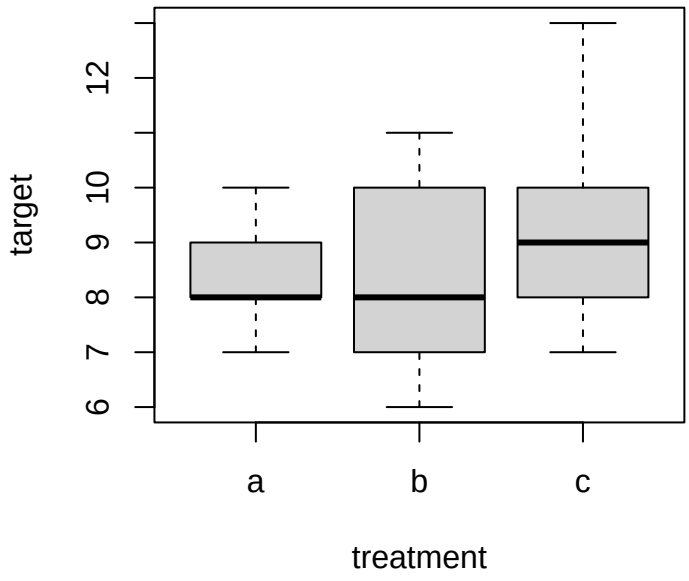

Figure 4: Distribution of the second synthetic data in the form of boxplot

Now let us try using `aov()` on this data:

```r
aov_result_2 <- aov(target~factor(treatment))
summary(aov_result_2)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## factor(treatment)  2   3.33   1.667   0.459  0.643
## Residuals         12  43.60   3.633
```

Please pay attention that the value of **Pr(>F)** is high, indicating that the evidence is not sufficient for us to reject the null hypothesis. Thus, we fail to reject the null hypothesis that there is no difference in means.