Using ANOVA in R

Mike Levine

June 27, 2019

Welcome to Using ANOVA in R! In this installment, we'll discuss what ANOVA is and how it's useful for analysts in the data analytics or data science space.

Analysis of Variance (ANOVA) is a statistical method used to determine the differences among group means of a sample. It's essentially a way to look at different groups and see if there's a statistically significant difference between them.

For example, ANOVA could be useful when we want to determine whether a product's brand is outperforming various other brands (and by how much). Personally, before understanding ANOVA I would have tackled this by creating a simple table comparing the means of each brand (in statistics, we call this a series of t-tests). Little did I know this would have lacked the statistical rigor that ANOVA provides us. A bunch of t-tests, for example, wouldn't consider whether we have a sufficient sample size from each brand–or whether the differences are great enough. It also has the potential to inflate what we 'stat nerds' call Type I Error (essentially, making an interpretation that's wrong). Brand A may do five hundred dollars in sales per store to Brand B through F's measly one hundred dollars per store, for example. If our dataset only has three stores for Brand A but has hundreds of stores for brands B through F, then we could be committing Type I Error by saying Brand A outperforms its competition.

That's not to say that you can never use simpler methods again, like an Excel table of means. ANOVA is just a more statistically sound way of looking at differences between means. Now let's look at applying ANOVA in the real world.

Note: *the following was completed through the Fox School of Business at Temple University's Master of Science in Business Analytics program. It was completed as an assignment for STAT 5607-308 (Advanced Business Statistics) in spring 2019.*

Let's consider an antiperspirant company. Their leadership is looking to purchase one of four brands on the market. Before making their decision, they want to study how effective each brand is. Specifically, they want to determine if there is a difference in sweat reduction between each brand. They ask 24 people to use each brand of deodorant for an amount of time and record as a percentage the sweat reduction of each brand. Our dataset includes the following variables:

- **Person**: the unique person
- **Brand**: the antiperspirant brand (1, 2, 3, 4)
- **Sweat**: the sweat reduction as a percentage

First, let's set our working directory in R and look at the first five rows of our dataset.

```
setwd("C:/Users/mdlev/OneDrive/Documents/Education/Graduate - Temple
University/2nd Semester/Advanced Business Statistics/R Datasets")
dat <- read.csv("Antipersp.csv")
head(dat)
##   Person Brand Sweat
## 1      1     1    40
## 2      1     2    52
## 3      1     3    53
## 4      1     4    46
## 5      2     1    20
## 6      2     2     0
```

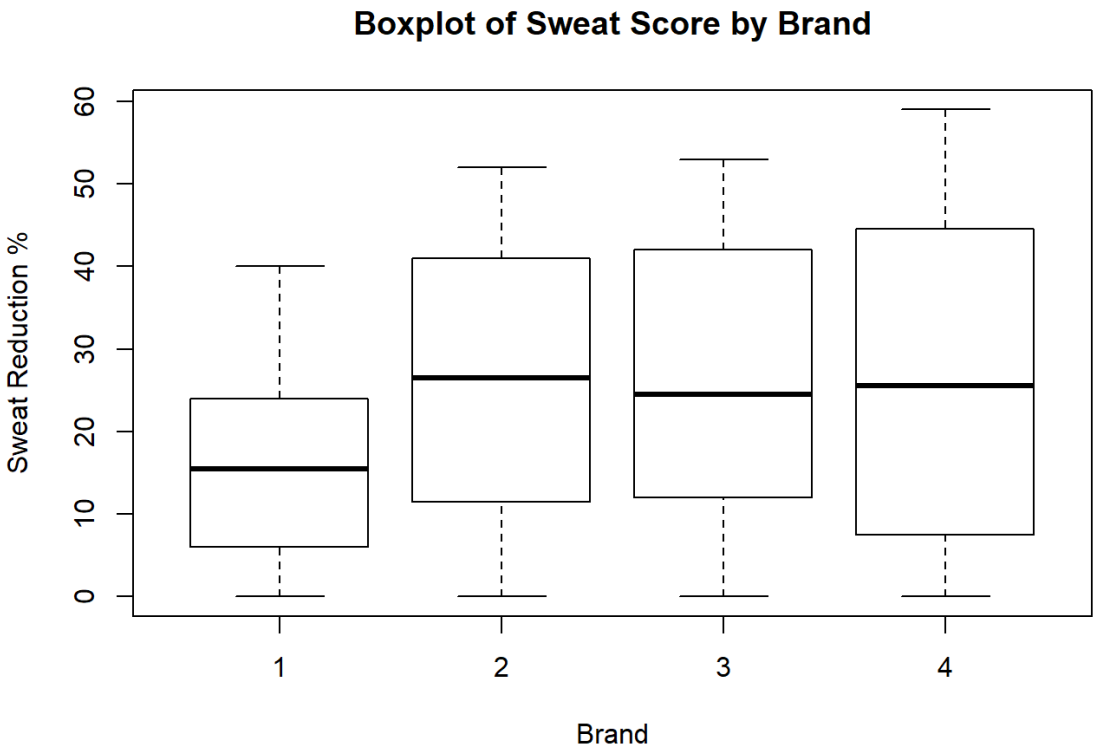
We can see that our data is easy to understand. The 'Person' field contains a unique identifier of each study participant, 'Brand' contains the unique brand, and 'Sweat' contains how effective each deodorant is for each individual.

Next, let's make sure that 'Person' and 'Brand' are both being treated as factors in R. (For any Excel nerds out there, this is like formatting our data as either a number or text. Only in this case, a factor is more like a categorical variable.) Our code demonstrates that they're not factors (and then converts them to factors).

```
is.factor(dat$Brand)
## [1] FALSE
dat$Brand <- as.factor(dat$Brand)
is.factor(dat$Brand)
## [1] TRUE
levels(dat$Brand)
## [1] "1" "2" "3" "4"
is.factor(dat$Person)
## [1] FALSE
dat$Person <- as.factor(dat$Person)
is.factor(dat$Person)
## [1] TRUE
levels(dat$Person)
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
## [15] "15" "16" "17" "18" "19" "20" "21" "22" "23" "24"
```

Next, let's create a series of boxplots for the distribution of sweat reduction (broken down by brand). This will allow us to visually see whether or not we suspect one brand performs better than the other.

```
boxplot(dat$Sweat~dat$Brand, ylab="Sweat Reduction %", xlab="Brand",
main="Boxplot of Sweat Score by Brand")
```



Looking at the boxplot, it visually looks like brands 2-4 do better than brand 1. Without further analysis, it's easy to conclude this and make that recommendation to a brand manager. However, what we don't know is whether this difference is significant. With ANOVA, we can see whether this difference does in fact mean that one brand performs better than the other.

Let's look at our mean scores across each brand:

```
tapply(dat$Sweat, dat$Brand, mean)
##        1        2        3        4
## 15.58333 25.00000 26.45833 26.45833
```

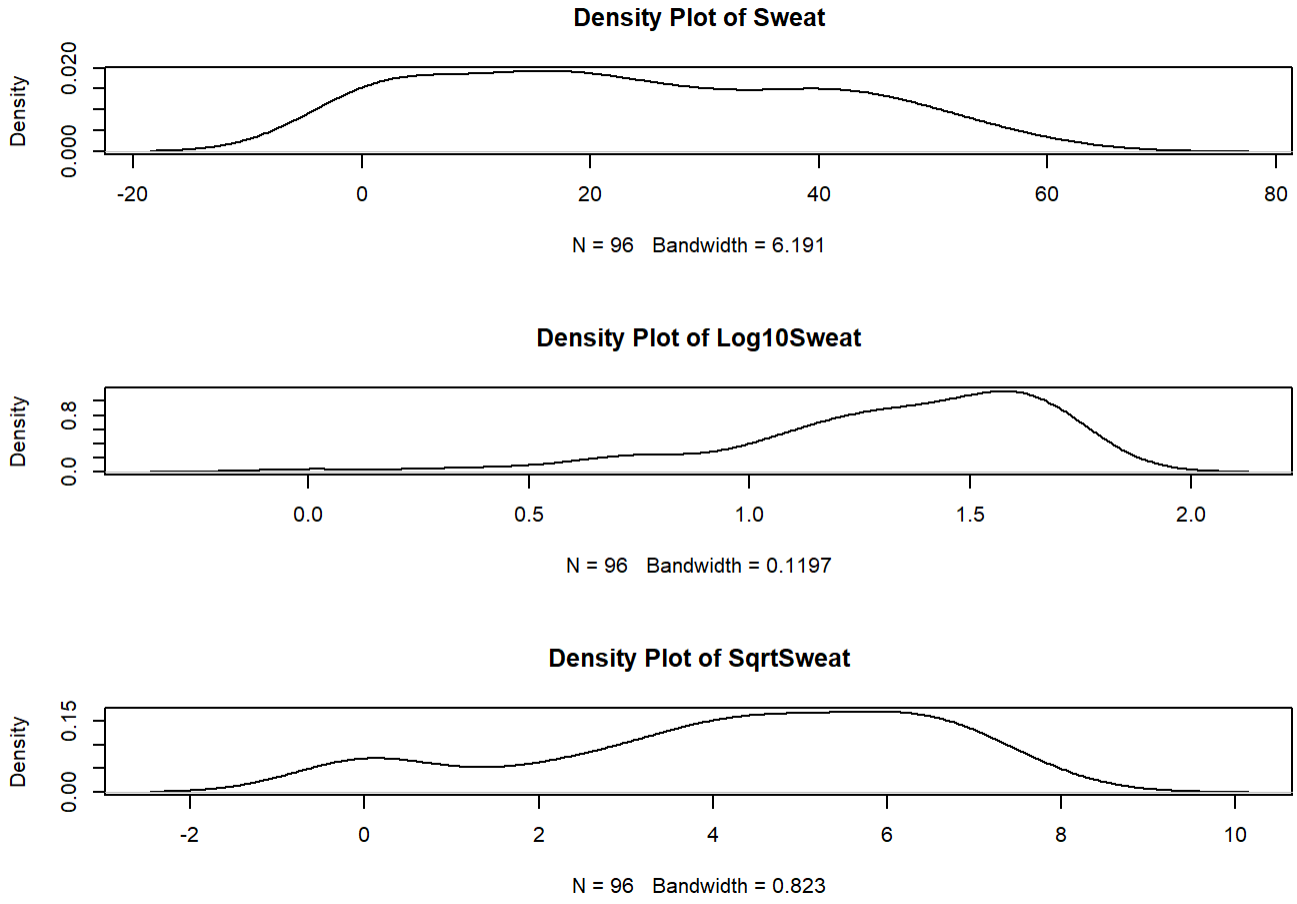
Now, let's look at our standard deviation scores across each brand:

```
tapply(dat$Sweat, dat$Brand, sd)
##        1        2        3        4
## 11.35941 16.67855 18.33500 19.58921
```

Looking at both tables above, we can make a couple inferences. First, our means and standard deviations match up with our box plots. Brand 1 has a lower mean and a lower standard deviation. Brands 2-4 are homogeneous in that they have similar means and standard deviations. We can also say that our standard deviations are (somewhat) similar. Let's dig a little deeper.

First, let's make sure that our 'Sweat' variable is normally distributed. We'll use a density plot to make this determination. Looking at our code and output below, we have three versions of 'Sweat' in density plots: the 'Sweat' variable with no transformation, log10 of the 'Sweat' variable, and the square root of the 'Sweat' variable. Looking at these three, we can say that the 'Sweat' varaible is normally distributed and that there's no reason to transform it for our ANOVA.

```
log10sweat <- log10(dat$Sweat)
sqrtsweat <- sqrt(dat$Sweat)
par(mfrow=c(3,1))
plot(density(dat$Sweat),main="Density Plot of Sweat")
plot(density(log10sweat),main="Density Plot of Log10Sweat")
plot(density(sqrtsweat),main="Density Plot of SqrtSweat")
```



Now, let's run our ANOVA.

```
library(ez)
## Warning: package 'ez' was built under R version 3.5.3
mod1 <-ezANOVA(data = dat, dv = .(Sweat), wid = .(Person),  within =
.(Brand), type = 3, detailed = TRUE)
mod1
```

```
## $ANOVA
##        Effect DFn DFd      SSn      SSd         F            p p<.05
## 1 (Intercept)   1  23 52453.50 14183.50 85.058730 3.439717e-09     *
## 2       Brand   3  69  1976.75 11740.25  3.872596 1.277762e-02     *
##          ges
## 1 0.66924395
## 2 0.07084998
##
## $`Mauchly's Test for Sphericity`
##   Effect         W          p p<.05
## 2  Brand 0.4288349 0.00252169     *
##
## $`Sphericity Corrections`
##   Effect       GGe      p[GG] p[GG]<.05       HFe      p[HF] p[HF]<.05
## 2  Brand 0.7677855 0.02198047         * 0.8581191 0.01778283         *
```

First, sphericity is violated since the p-value for that test in the results is less than 0.05. Using the GG and HF corrected values, we see that both are less than 0.05. As such, we average them (0.022 + 0.018 / 2) = 0.02. Based on this, we conclude that our difference is significant. Moreover, the appropriate p-value for this test is 0.05. The p-value is less than 0.05, meaning at least one set of means between groups is different.

Lastly, let's run our Bonferroni post hoc tests.

```
pairwise.t.test(dat$Sweat, dat$Brand, paired = TRUE, p.adjust.method =
"bonferroni")
##
##  Pairwise comparisons using paired t tests
##
## data:  dat$Sweat and dat$Brand
##
##   1     2     3
## 2 0.136 -     -
## 3 0.075 1.000 -
## 4 0.135 1.000 1.000
##
## P value adjustment method: bonferroni
options(digits = 7)
```

According to our post-hoc tests, there aren't any significant differences in means between the four brands. We know this because our p-values in the output above are all above 0.05 (our threshold).

Conclusion: our ANOVA test initially indicated that at least one grouping of means was statistically different from the others. However, when we ran our post-hoc test it was discovered that this wasn't the case. Overall, the takeaway for our antiperspirant company is that there aren't significant differences in means between the four brands. While Brands 2 through 4 visually appear to do better than Brand 1 (and our ANOVA suggests that may be the case), it still isn't statistically significant.