Introduction to ANOVA

Aleksandra Ma

October 30, 2017

```
library(knitr)
```

INTRODUCTION

The first time I ran into “ANOVA”, which stands for Analysis of Variance, was when I was taking a business class on cost behavior. I was curious about what this is because ANOVA analysis seems to be everywhere, in science, business, and even election. In this post, I will give a brief background introduction on ANOVA, and then go into details on how to do a one-way ANOVA analysis in R.

Background Information on ANOVA

What is ANOVA?

ANOVA, also known as “Analysis of Variance”, is a statistical technique that is used to analyze the differences among group means and their associated procedures with two and more categories, developed by statistician Ronald Fisher. For example, an ANOVA can examine the potential difference in the voting breakdown in different states(California, New York, Ohio, Maryland, etc.). It basically generalizes the t-test, which compares whether the average difference between two groups is really significant or just due to random chance, to more than two groups. Conceptually similar to multiple two-sample t-tests, ANOVA results in less type I error and is therefore more accurate.

Kinds of ANOVA

One-Way ANOVA:A one-way ANOVA has only one independent variable. In the previous example, the party people in different states support is the dependent variable, and the state that they are in is the independent variable.

Two-way ANOVA(Factorial ANOVA): A two-way ANOVA has two independent variables. Going back to the election example, we can examine the supporting party difference by both states and gender. Now both states and gender are independent variables, and both of them can have two or more categories. Two-way ANOVA is used to examine the interaction between the two independent variables. For example, females in all states tend to be more likely to be a democrat, but this difference could be greater in states like California than in states like Texas.

N-Way ANOVA: We can use an n-way ANOVA to examine the difference by more than two independent variables. For example, besides country and gender, we can also have ethnicity, and even age groups. N is the number of independent variables we have.

Terminology that may turn up

- Variance: the average of the squared differences from the mean, which measures how far a data set is spread out.
- Type I error: this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance.
- Type II error: this is the error of failing to accept an alternative hypothesis when you don’t have adequate power.
- F-test: An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.
- Df: the number of degrees of freedom is the number of independent ways by which a dynamic system can move, without violating any constraint imposed on it.
- Sum Sq: sum of squares, which measures how far individual measurements are from the mean.
- Mean Sq: Mean squares are estimates of variance across groups. It is calculated as sum of squares divided by its appropriate degrees of freedom.
- F value: a value you get when you run an ANOVA test or a regression analysis to find out if the means between two populations are significantly different.
- Pr(>F): P value, which is the evidence against a null hypothesis. The smaller the p-value, the strong the evidence that you should reject the null hypothesis.

Assumptions of ANOVA

1. Normality of the dependent variable distribution: The data in each cell should be approximately normally distributed.
2. Homogeneity of variance: The variance in each cell should be similar.
3. Sample size: per cell > 20 is preferred; aids robustness to violation of the first two assumptions, and a larger sample size increases power.
4. Independent observations: scores on one variable or for one group should not be dependent on another variable or group.

Example of Dog Show (Why ANOVA?)

ANOVA can be used as a tool to explain observations. Below I will use an example of the dog weights at a dog show to explain how ANOVA can be used to prove hypothesis. A dog show is not a random sampling of the breed: the dogs there are mostly adult, pure-bred, and exemplary. In the three illusions below, the yellow distribution on the left is a histogram of dog weight from a show. If we want to predict the weight of a dog at the show, what characteristic should we depend on? A successful grouping will split dogs so that: (1) for each grouping the variance of dog weights will be low, meaning that the weight in each grouping is not very spread out, and (2) the mean of each group is distinct because otherwise it would not be reasonable to say that each group is separate.

First, we can choose the characteristic to be young vs old, short-haired vs long-haired. As the illustration below has shown, this grouping has failed to fit the yellow distribution that we are trying to explain. Because each blue group on the right has a very big variance, and their means are very close, it is not a very effective characteristic to predict a dog’s weight with.

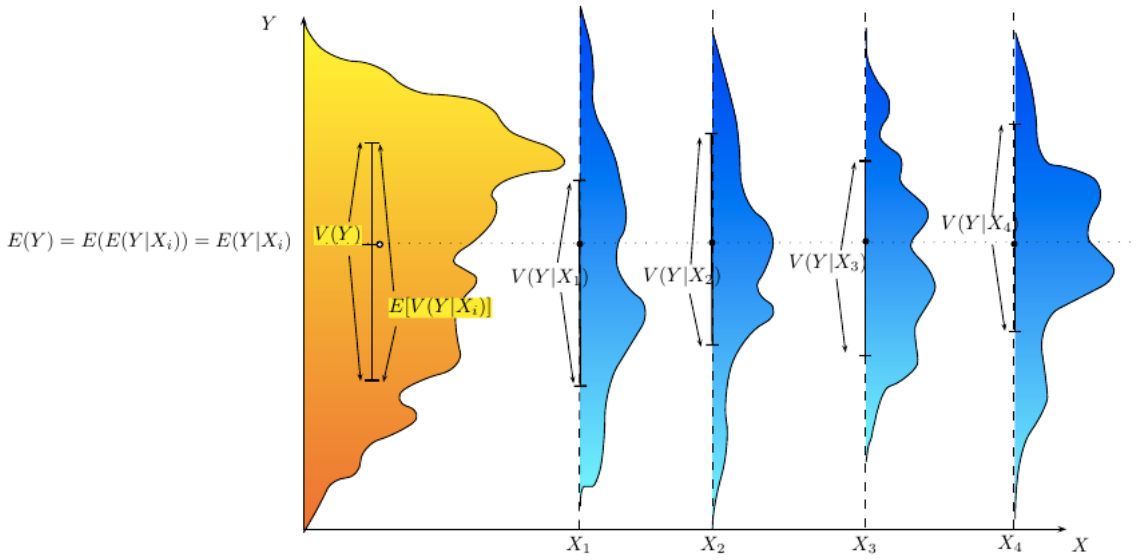


Figure 2: ANOVA : No fit

Now let's try using another characteristic and grouping the dogs as pet vs working breed and less athletic vs more athletic. Since the variance is smaller and the mean is more spread out, this grouping is to some extent more successful (fair fit). And if we think about it, the working breed dogs indeed tend to be heavier than the smaller and lighter dogs that are kept as pets. However, as we have seen below in the illustration, there are still significant overlaps in each grouping, which means we cannot say that each group is truly distinct. That is also why this way of grouping only gives us fair fit.

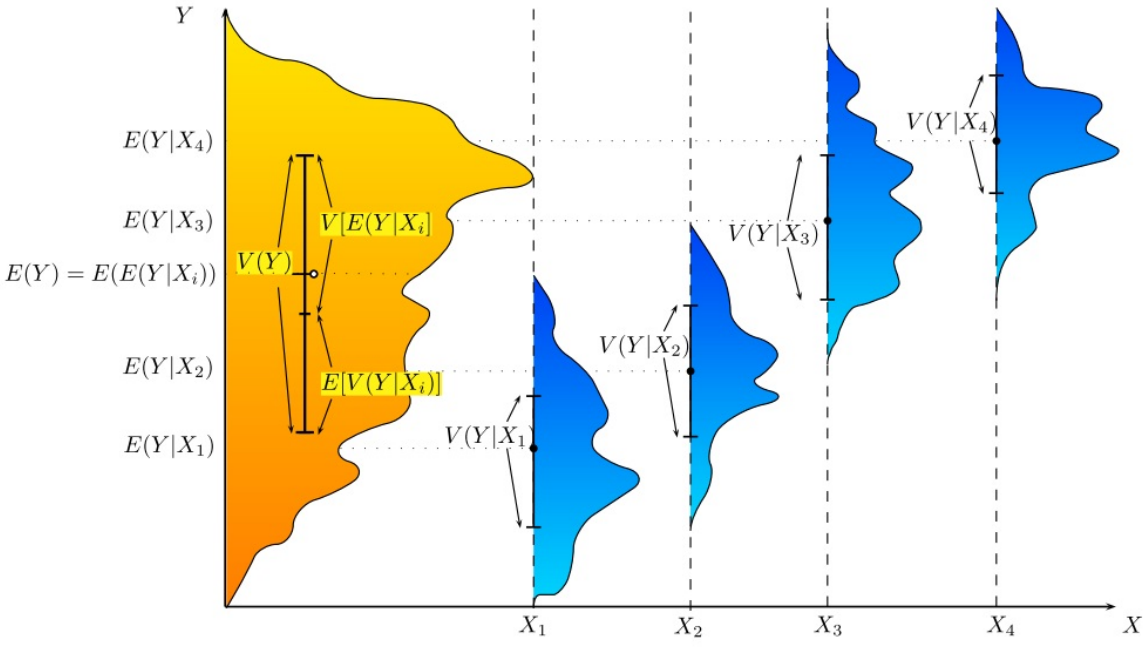


Figure 1: ANOVA : Fair fit

The last attempt is to explain weight by breed, which produces a very good fit according to the illustration below. All Chihuahuas are light and all St Bernards are heavy. The difference between Setters and Pointers does not justify separate breeds. One advantage ANOVA has over correlation is that not all data must be numeric.

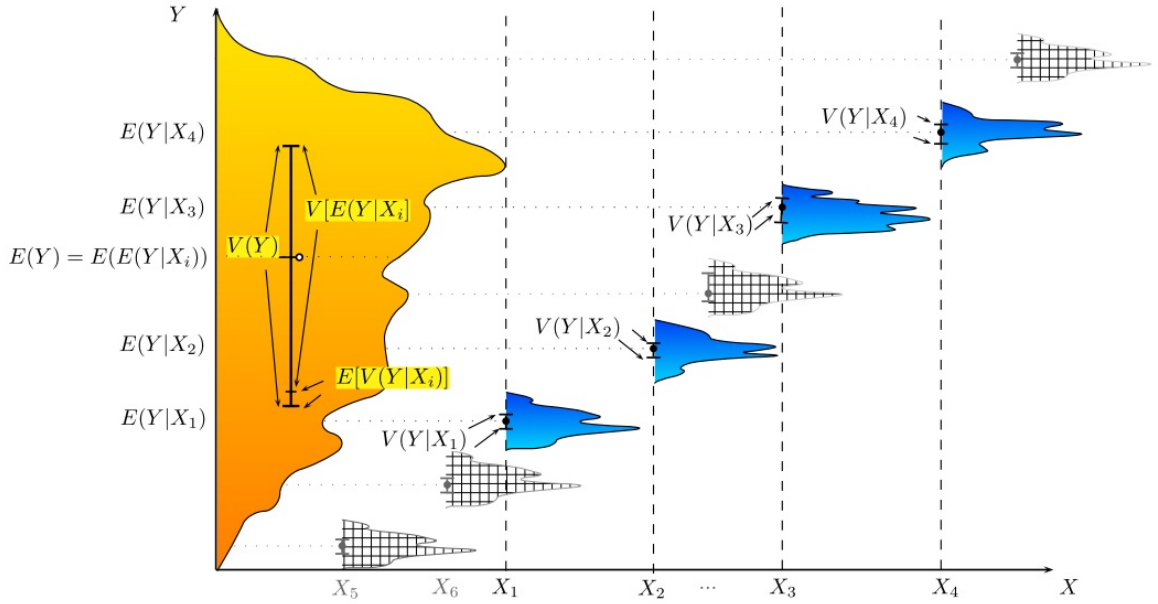


Figure 3: ANOVA : very good fit

One-way ANOVA Using R

Having learned about one-way ANOVA and its purpose, now let's do a simple one-way analysis of variance(ANOVA) using the R function aov().

The first step is to compare the means of our independent variables graphically. We can do this by creating side-by-side boxplots of measurements organized in groups using *plot()*:

plot(response ~ factor, data = data.name)

where *response* is the name of the dependent variable, and *factor* is the independent variable. Both variables should be contained in a data frame called *data.name*.

Example of a drug company

A drug company tested three formulations of a pain relief medicine for migraine headache sufferers. For the experiment 27 volunteers were selected and 9 were randomly assigned to one of three drug formulations. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 to 10 (10 being most pain).

- Drug A 4 5 4 3 2 4 3 4 4
- Drug B 6 8 4 5 4 6 5 8 6
- Drug C 6 7 6 6 7 5 6 5 5

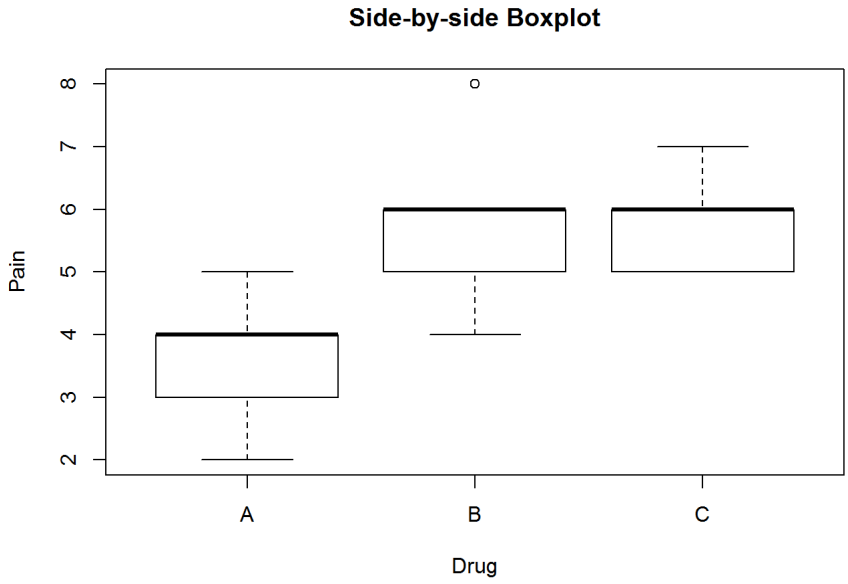
```
pain = c(4, 5, 4, 3, 2, 4, 3, 4, 4, 6, 8, 4, 5, 4, 6, 5, 8, 6, 6, 7, 6, 6, 7, 5, 6, 5, 5)
drug = c(rep("A",9), rep("B",9), rep("C",9))
migraine = data.frame(pain,drug)
migraine
```

```
##    pain drug
## 1     4    A
## 2     5    A
## 3     4    A
## 4     3    A
## 5     2    A
## 6     4    A
## 7     3    A
## 8     4    A
## 9     4    A
## 10    6    B
## 11    8    B
## 12    4    B
## 13    5    B
## 14    4    B
## 15    6    B
## 16    5    B
## 17    8    B
## 18    6    B
## 19    6    C
## 20    7    C
## 21    6    C
## 22    6    C
## 23    7    C
## 24    5    C
## 25    6    C
## 26    5    C
## 27    5    C
```

The way we construct this data frame is to make side-by-side boxplots and perform ANOVA.

And now according to the plot() function above, we can make the boxplot:

```
plot(pain ~ drug, data=migraine,main="Side-by-side Boxplot",xlab="Drug",ylab="Pain")
```



From the boxplots, it appears that the mean of the pain for drug A is lower than that for drugs B and C.

Next we can use the R function *aov()* to fit ANOVA models. The general form is similar to plot:

aov(response ~ factor, data=data.name)

where *response* is the name of the dependent variable, and *factor* is the independent variable. Both variables should be contained in a data frame called *data.name*. Once our ANOVA model is fit, we can look at the results by using the *summary()* function that we've learned in class, which gives us the standard ANOVA table.

```
results = aov(pain ~ drug, data = migraine)
summary(results)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## drug         2  28.22  14.111   11.91 0.000256 ***
## Residuals   24  28.44   1.185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Like t-test, f-test also has a f-value and its corresponding p-value. In our case, the f-value is 11.91 with a p-value of 0.0003, which is even smaller than the last tier 0.001 according to the significant codes given by the summary. We clearly reject the null hypothesis of equal means for all three drug groups. This tells us that the difference in the mean pain for three drugs is not just due to random chance. In order to understand the summary result even better, refer back to **Terminology that may turn up** for more information.

Take-home Message

Now I've finally understood what ANOVA means in my cost behavior class. It is to see if the difference between the cost among different factories is due to chance or actually significant. I hope after reading this post, you have an idea of what it is, too. Analysis of Variance is very common nowadays in almost every field, because as the old saying goes, "everything connects to everything else". With ANOVA, you can check if everything is indeed connected to everything else, or if it is just due to random chance. And you don't even need to worry about turning your independent variables into numbers!

Reference

- http://www.statisticssolutions.com/manova-analysis-anova/
- https://en.wikipedia.org/wiki/One-way_analysis_of_variance
- https://www.stat.berkeley.edu/~hhuang/STAT141/Lecture-FDR.pdf
- https://en.wikiversity.org/wiki/Analysis_of_variance/Assumptions
- https://en.wikipedia.org/wiki/Analysis_of_variance
- http://www.stat.columbia.edu/~martin/W2024/R3.pdf
- https://en.wikipedia.org/wiki/F-test