ANOVA Example

D2K Course Staff

January 23 2019

Data

Below, we show an example of how to run ANOVA tests. This is done on the `iris` data set, which contains data on 3 different species of iris flowers.

```
data(iris)
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

In this data set, `Species` is a categorical variable, while the sepal and petal lengths and widths are continuous variables.

One-Way ANOVA

The one-way ANOVA is used to test the equivalence of means of a response for different categories of a single predictor variable. The response variable must be continuous, and the predictor must be categorical. The assumptions of all ANOVA tests, including the variants listed below:

- The population follows a normal distribution.
- Samples are independent.
- The variance of the response variable is the same between groups.

However, mild violations of these assumptions are ok.

In the code below, we use a one-way ANOVA to test if there is a difference between the means of the sepal length between the three different species. (Checking the assumptions of the model is skipped.) In this case, we would be running the hypothesis test:

H_0: The mean sepal length for all 3 species is the same.

H_a: The mean sepal lengths are not all the same for the 3 species.

```r
# Run ANOVA
an <- anova(lm(Sepal.Length ~ Species, data = iris))

an
```

```
Analysis of Variance Table

Response: Sepal.Length
           Df Sum Sq Mean Sq F value    Pr(>F)
Species     2 63.212  31.606  119.26 < 2.2e-16 ***
Residuals 147 38.956   0.265
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we see in the summary, the p-value for the `Species` variable is extremely small, meaning that there is a statistically significant difference between the sepal lengths for the 3 different iris species. (Note that this does not tell us anything about how they are different, only that a difference exists!)

Two-Way ANOVA

The two-way ANOVA is used to test the equivalence of means for different categories of two predictor variable. For the following example, we first create a binary category variable based on petal width.

```
# New categorical variable
iris$Petal.Width.Bin <- ifelse(iris$Petal.Width > 1, 0, 1)
```

In the code below, we use a two-way ANOVA to test if there is a difference between the means of the sepal length between the three different species and two different petal width categories. (Checking the assumptions of the model is skipped.) In this case, we would be running the hypothesis test:

H_0: The mean sepal length for all 3 species and two petal width categories is the same.

H_a: The mean sepal lengths are not all the same for the 3 species and/or two petal width categories.

```
# Run ANOVA
an <- anova(lm(Sepal.Length ~ Species + Petal.Width.Bin, data = iris))

an
```

```
Analysis of Variance Table

Response: Sepal.Length
                 Df Sum Sq Mean Sq  F value    Pr(>F)
Species           2 63.212 31.6061 125.5959 < 2.2e-16 ***
Petal.Width.Bin   1  2.215  2.2155   8.8038  0.003514 **
Residuals       146 36.741  0.2516
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we see in the summary, the p-values for the `Species` variable and the `Petal.Width.Bin` variable are extremely small, meaning that there is a statistically significant difference between the sepal lengths for the 3 different iris species and the two petal width categories.

MANOVA

A MANOVA model simply runs simultaneous ANOVA tests on multiple continuous response variables. In the code below, we use aa MANOVA test to see if there is a difference between the means of the sepal length and/or sepal widths between the three different species. (Checking the assumptions of the model is skipped.) In this case, we would be running the hypothesis test:

H_0: The mean sepal lengths and widths for all 3 species are the same.

H_a: The mean sepal lengths and/or widths are not all the same for the 3 species.

```r
# Run MANOVA
man <- manova(cbind(Sepal.Length, Sepal.Width) ~ Species, data = iris)

summary(man)
```

```
           Df  Pillai approx F num Df den Df    Pr(>F)    
Species     2 0.94531   65.878      4    294 < 2.2e-16 ***
Residuals 147                                             
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we see in the summary, the p-value for the `Species` variable is extremely small, meaning that there is a statistically significant difference between the sepal lengths and/or for the 3 different iris species. (Note that this does not tell us anything about whether sepal lengths, sepal widths, or both are different!)

ANCOVA

An ANCOVA is used to test if the equivalence of means of a response for different categories of a predictor variable, conditional on other quantitative variables. The response and covariates are continuous, and the predictor is categorical.

The code below tests the hypothesis:

H_0: The mean sepal length for all 3 species is the same given the petal lengths.

H_a: The mean sepal lengths are not all the same for the 3 species given the petal lengths.

Note that it is important for the variable of interest to be the last one in the formula.

```
# Run ANCOVA
anc <- anova(lm(Sepal.Length ~ Petal.Length + Species, data = iris))

anc
```

```
Analysis of Variance Table

Response: Sepal.Length
              Df Sum Sq Mean Sq F value    Pr(>F)
Petal.Length   1 77.643  77.643 679.544 < 2.2e-16 ***
Species        2  7.843   3.922  34.323 6.053e-13 ***
Residuals    146 16.682   0.114
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we see in the summary, the p-value for the `Species` variable is extremely small, meaning that there is a statistically significant difference between the sepal lengths for the 3 different iris species once we condition on petal length.

What Next?

Say we have run an ANOVA and rejected the null that the means between categories are the same. We still do not know which ones are different. The next step should be running pairwise t-tests, remembering to correct for multiple comparisions, to identify specifically which means are different.

What If My Data Are Non-Gaussian?

In this case, we would want to use a nonparametric hypothesis test, such as the Mann-Whitney U test or the Kruskal-Wallis test. This is covered in a separate example.