15.6.5 Use the ANOVA Test

Woah! Two-sample t-tests can get complex pretty fast. Thankfully, Jeremy has one more trick up his sleeve—the analysis of variance (ANOVA) test.

When dealing with large real-world numerical data, we're often interested in comparing the means across more than two samples or groups. The most straightforward way to do this is to use the **analysis of variance (ANOVA) test,** which is used to compare the means of a continuous numerical variable across a number of groups (or factors in R).

Depending on your dataset and questions you wish to answer, an ANOVA can be used in multiple ways. For the purposes of this course, we'll concentrate on two different types of ANOVA tests:

- A **one-way ANOVA** is used to test the means of a single dependent variable across a single independent variable with multiple groups. (e.g., fuel efficiency of different cars based on vehicle class).
- A **two-way ANOVA** does the same thing, but for two different independent variables (e.g., vehicle braking distance based on weather conditions and transmission type).

Regardless of whichever type of ANOVA test we use, the statistical hypotheses of an ANOVA test are the same:

H_0 : The means of all groups are equal, or $\mu_1 = \mu_2 = \ldots = \mu_n$.

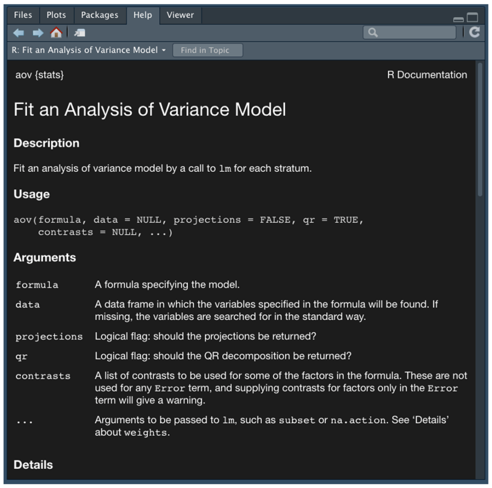
H_a : At least one of the means is different from all other groups.

Additionally, both ANOVA tests have assumptions about the input data that must be validated prior to using the statistical test:

1. The dependent variable is numerical and continuous, and the independent variables are categorical.
2. The dependent variable is considered to be normally distributed.
3. The variance among each group should be very similar.

In R, we can use the `aov()` function to perform both the one-way and two-way ANOVA test. Type the following code into the R console to look at the `aov()` documentation in the Help pane:

```
>?aov()
```



aov {stats} R Documentation

Fit an Analysis of Variance Model

Description

Fit an analysis of variance model by a call to `lm` for each stratum.

Usage

```
aov(formula, data = NULL, projections = FALSE, qr = TRUE,
    contrasts = NULL, ...)
```

Arguments

`formula`	A formula specifying the model.
`data`	A data frame in which the variables specified in the formula will be found. If missing, the variables are searched for in the standard way.
`projections`	Logical flag: should the projections be returned?
`qr`	Logical flag: should the QR decomposition be returned?
`contrasts`	A list of contrasts to be used for some of the factors in the formula. These are not used for any `Error` term, and supplying contrasts for factors only in the `Error` term will give a warning.
`...`	Arguments to be passed to `lm`, such as `subset` or `na.action`. See 'Details' about `weights`.

Details

To perform an ANOVA test in R, we have to provide the `aov()`function two arguments:

- **formula**
- **data**

Unlike the `t.test()` function, where each group was a separate numeric vector, the `aov()` function expects that all of the observations and grouping information are contained within a single data frame. Once we have our cleaned and labeled data frame, we're ready to perform our ANOVA test using the `aov()`function.

To practice our one-way ANOVA, return to the mtcars dataset. For this statistical test, we'll answer the question, "Is there any statistical difference in the horsepower of a vehicle based on its engine type?"

In this case, we will use the "hp" and "cyl" columns from our mtcars dataset:

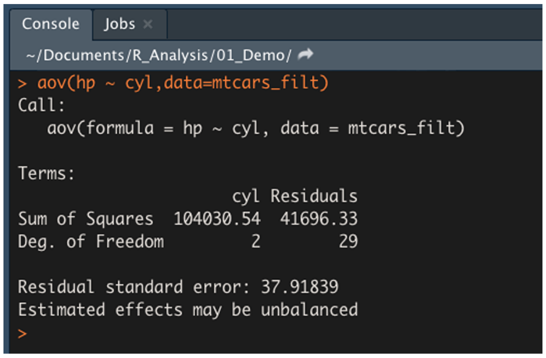
- horsepower (the "hp" column) will be our dependent, measured variable
- number of cylinders (the "cyl" column) will be our independent, categorical variable.

However, in the mtcars dataset, the `cyl` is considered a numerical interval vector, not a categorical vector. Therefore, we must clean our data before we begin, using the following code:

```
> mtcars_filt <- mtcars[,c("hp","cyl")] #filter columns from mtcars dataset
> mtcars_filt$cyl <- factor(mtcars_filt$cyl) #convert numeric column to fact
```

Now that we have our cleaned dataset, we can use our `aov()` function as follows:

```
> aov(hp ~ cyl,data=mtcars_filt) #compare means across multiple levels
```



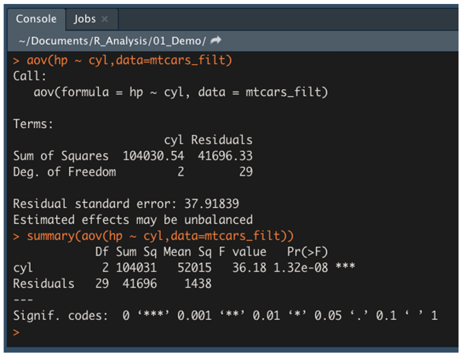
```
> aov(hp ~ cyl,data=mtcars_filt)
Call:
   aov(formula = hp ~ cyl, data = mtcars_filt)

Terms:
                      cyl Residuals
Sum of Squares  104030.54  41696.33
Deg. of Freedom         2        29

Residual standard error: 37.91839
Estimated effects may be unbalanced
>
```

Due to the fact that the ANOVA model is used in many forms, the initial output of our `aov()` function does not contain our p-values. To retrieve our p-values, we have to wrap our `aov()` function in a `summary()` function as follows:

```
> summary(aov(hp ~ cyl,data=mtcars_filt))
```



```
> aov(hp ~ cyl,data=mtcars_filt)
Call:
   aov(formula = hp ~ cyl, data = mtcars_filt)

Terms:
                      cyl Residuals
Sum of Squares  104030.54  41696.33
Deg. of Freedom         2        29

Residual standard error: 37.91839
Estimated effects may be unbalanced
> summary(aov(hp ~ cyl,data=mtcars_filt))
            Df Sum Sq Mean Sq F value   Pr(>F)    
cyl          2 104031   52015   36.18 1.32e-08 ***
Residuals   29  41696    1438                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

When using the formula statement, each independent variable will be shown as a separate row, with an additional "Residuals" row that tells us what the residual error is for our ANOVA model. For our purposes, we are only concerned with the "Pr(>F)" column, which is the same as our p-value statistic.

Depending on how small our p-value is, there may be symbols on the right side that indicate which significance level the p-value is below. In this case, our p-value is 1.32 × 10^-8^, which is much smaller than our assumed 0.05 percent significance level. Therefore, we would state that there is sufficient evidence to reject the null hypothesis and accept that there is a significant difference in horsepower between at least one engine type and the others.

Now that you have learned the differences between our t-tests and ANOVA tests, you're ready to analyze data and perform statistical tests when comparing means. Feel free to explore more datasets and practice implementing analysis of means on your own.