# Using ANOVA in R

## ANOVA in R

Open RStudio and highlight the open a new R Notebook.

Before we get started, we should load a library with a lot of helper functions that will be useful to us. The library is called `tidyverse`. You will want to load this every time you start working in R.

```
library(tidyverse)
```

```
-- Attaching packages --------

v ggplot2 3.2.0     v purrr   0.3.2
v tibble  2.1.3     v dplyr   0.8.3
v tidyr   0.8.3     v stringr 1.4.0
v readr   1.3.1     v forcats 0.4.0

-- Conflicts -----------------
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

### The setup

Recall the scenario from the previous screencast. You have stores in three locations and want to see how their customer satisfaction compares. You randomly survey customers from each store. Is there a difference in customer satisfaction between the stores?

To answer this we'll explore the data a little, then use the ANOVA procedure to see if we have strong evidence.

First load the data. Download the csv from Canvas. In the lower right panel of Rstudio find the Files tab, find where you downloaded the file and import. This creates a data frame holding the data from the csv file. You should see output something like below:

```
SurveyData <- read_csv("SurveyData.csv")
```

```
Parsed with column specification:
cols(
  Person = col_double(),
  Location = col_character(),
  Rating = col_double()
)
```

We will use this dataset throughout the rest of this problem, so let's make it the default dataset using the command:

```
attach(SurveyData)
```

### Understand the data

What does the data look like? To get a quick idea we can just enter the name of the data frame we just created and R will print the first 10 lines of the data.

```
SurveyData
```

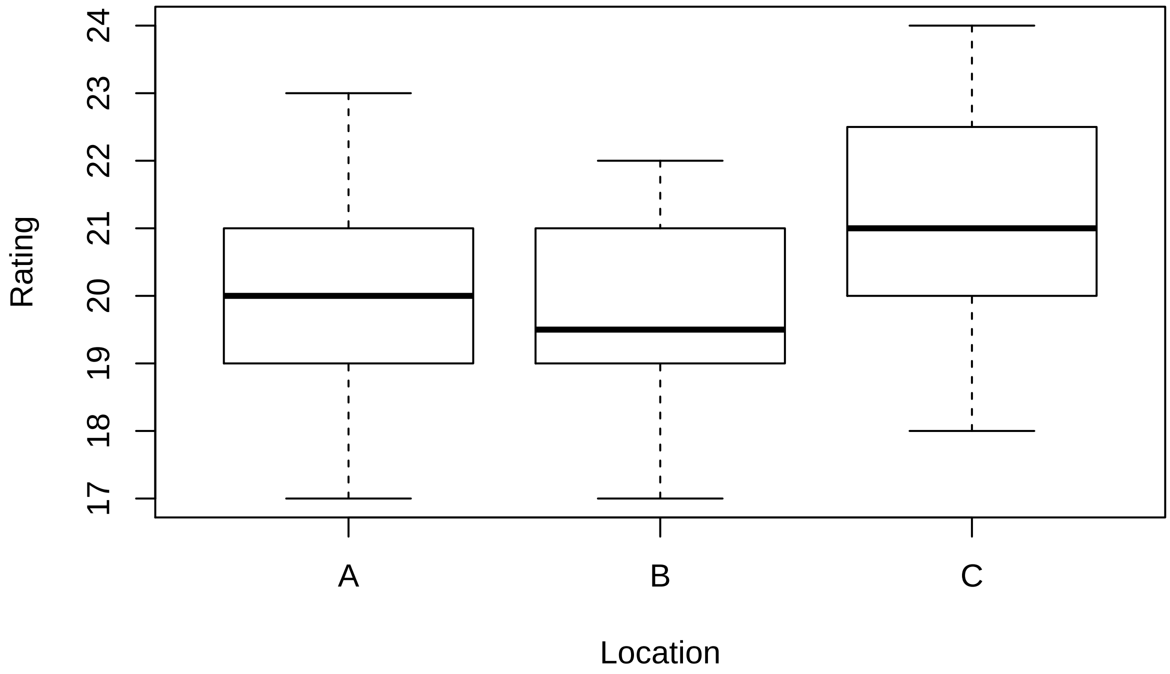
```
# A tibble: 58 x 3
   Person Location Rating
    <dbl> <chr>     <dbl>
 1      1 A            23
 2      2 A            21
 3      3 A            21
 4      4 C            21
 5      5 C            21
 6      6 A            17
 7      7 B            20
 8      8 A            19
 9      9 A            19
10     10 A            21
# ... with 48 more rows
```

So we have a column labelled "Person" with a number, a "Location" column with three choices, and a "Rating" column with the satisfaction ratiing from a customer at that Location.

Next a boxplot can be helpful to get a sense of the data. Note the command below uses numbers from the Rating column, but groups the data by Location.

Can you see where I told the function where to find the data? Can you see where I told it to group the data by Location?

```
boxplot(Rating ~ Location)
```



The centers for each region seem different, but are they really? All the boxplots do overlap to some extent.

## Running an ANOVA test

Let's use an ANOVA test to decide if we have evidence that there is a difference in the means. Before we run the test we should be sure all the necessary conditions are met.

Requirements:

3. **Independent samples**: Since these are different stores where customers were sampled randomly we are ok here.
4. **Normality**: the sample sizes are pretty large, but we can still think about normality. Based on the boxplots, there is not a significant skew, so we are safe here.
5. **Standard deviations**: remember, the largest should be less than twice the smallest. Let's check that. The `tapply` command applies a command to data in a table column, grouping by another column. So with the one command below we compute the standard deviation for each region.

```
tapply(Rating, Location, sd)
```

```
       A        B        C 
1.596126 1.274434 1.627613 
```

So the largest SD is store 3 at 137.97 and the smallest is store 1 at 100.57. The ratio between these is less than 2 so this requirement is met.

## Running the test

The assumptions of ANOVA are met, so let's actually run the test. In the first line we run the test with the `aov` command and store with the name `analysis`. The next line asks for a summary of the analysis. R stores a lot more information about the test, but this is all we really need.

```
analysis <- aov(Rating ~ Location)
summary(analysis)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)   
Location     2  24.86  12.428   5.414 0.00714 **
Residuals   55 126.25   2.295                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Recall ANOVA uses the F distribution. We find a p-value of 0.00714 which is quite small. So we reject the null hypothesis.

In other words, we have strong evidence that at least two of the stores have different means. But which ones are different?

We'll use the code below to generate a confidence interval for the mean of each store. Remember the `tapply` command from above? The `t.test` command is a convenient way to generate a confidence interval.

```
tapply(Rating, Location, t.test)
```

```
$A

    One Sample t-test

data:  X[[i]]
t = 57.558, df = 20, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 19.32107 20.77417
sample estimates:
mean of x 
 20.04762 


$B
```

```
    One Sample t-test

data:  X[[i]]
t = 65.656, df = 17, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 19.08846 20.35598
sample estimates:
mean of x
 19.72222


$C

    One Sample t-test

data:  X[[i]]
t = 56.945, df = 18, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 20.47867 22.04764
sample estimates:
mean of x
 21.26316
```

We see Stores A and B have significant overlap in their confidence intervals, so I'm not convinced their means are different. The only stores whose intervals do not overlap are B and C. Since our ANOVA told us at least two means must be different, we can say we have strong evidence that the mean for Store B is less than the mean for Store A.