15.6.4

Use the Two-Sample t-Test to Compare Samples

Jeremy can already tell that the two-sample t-test will be a standard part of his analytical procedure—especially when Colleen stops by and tells him you can actually use this test to compare samples from different populations!

In many cases, the two-sample t-test will be used to compare two samples from a single population dataset. However, two-sample t-tests are flexible and can be used for another purpose: to compare two samples, each from a different population. This is known as a **pair t-test**, because we pair observations in one dataset with observations in another. We use the pair t-test when:

- Comparing measurements on the same subjects across a single span of time (e.g., fuel efficiency before and after an oil change)
- Comparing different methods of measurement (e.g., testing tire pressure using two different tire pressure gauges)

The biggest difference between paired and unpaired t-tests is how the means are calculated. In an unpaired t-test, the means are calculated by adding up all observations in a dataset, and dividing by the number of data points. In a paired t-test, the means are determined from the difference between each paired observation. As a result of the new mean calculations, our paired t-test hypotheses will be slightly different:

- H_0 : The **difference** between our paired observations (the true mean difference, or "μ_d") is **equal to zero**.
- H_a : The **difference** between our paired observations (the true mean difference, or "μ_d") is **not equal to zero**.

When it comes to implementing a paired t-test in R, we'll use the `t.test()` function.

Retake

The required arguments are slightly different from the unpaired two-sample t-test:

- **x** is the first numeric vector of sample data.
- **y** is the second numeric vector of sample data.
- **paired** tells the `t.test()` function to perform a paired t-test. This value must be set to TRUE.
- **alternative** tells the `t.test()` function if the hypothesis is one-sided (one-tailed) or two-sided (two-tailed). The options for the alternative argument are "two.sided," "less," or "greater." By default, the `t.test()` function assumes a two-sided t-test.

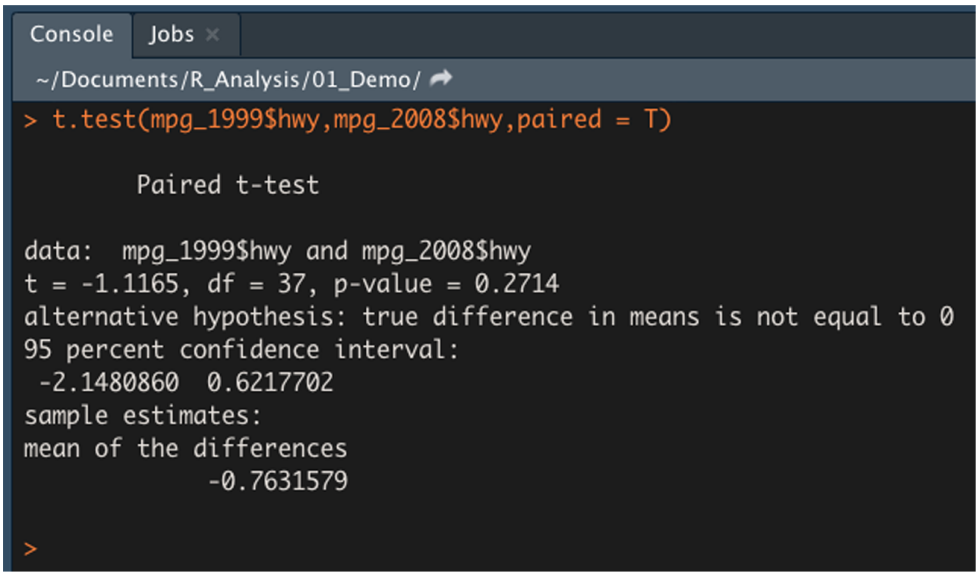
To practice calculating a paired t-test in R, download the modified **mpg dataset** (https://2u-data-curriculum-team.s3.amazonaws.com/dataviz-online/module_15/mpg_modified.csv) data file contains a modified version of R's built-in mpg dataset, where each 1999 vehicle was paired with a corresponding 2008 vehicle.

First, let's generate our two data samples using the following code:

```
> mpg_data <- read.csv('mpg_modified.csv') #import dataset
> mpg_1999 <- mpg_data %>% filter(year==1999) #select only data points where the year is 1999
> mpg_2008 <- mpg_data %>% filter(year==2008) #select only data points where the year is 2008
```

Now that we have our paired datasets, we can use a paired t-test to determine if there is a statistical difference in overall highway fuel efficiency between vehicles manufactured in 1999 versus 2008. In other words, we are testing our null hypothesis—that the overall difference is zero. Using our `t.test()` function in R, our code would be as follows:

```
> t.test(mpg_1999$hwy,mpg_2008$hwy,paired = T) #compare the mean difference between two samples
```



Calculating score. This might take a while. Please wait...

Your assignment has been successfully submitted.