The second main form of the t-Test is a two-sample t-Test. Instead of testing whether a sample mean is statistically different from its population mean, the two-sample t-Test determines whether the means of two samples are statistically different. In other words, a two-sample t-Test is used to test the following hypotheses:

* H0 : There is **no statistical difference** between the two observed sample means.
* Ha : There is **a statistical difference** between the two observed sample means.

There are also five assumptions regarding our input data when using the two-sample t-Test, which are the same as the one-sample t-Test:

1. The input data is numerical and continuous.
2. Each sample data was selected randomly from the population data.
3. The input data is considered to be normally distributed.
4. Each sample size is reasonably large. Generally speaking, this means that the sample data distribution should be similar to its population data distribution.
5. The variance of the input data should be very similar.

In R, we use the same t.test() function to calculate both a one-sample t-Test and two-sample t-Test. However, the two-sample t-Test arguments are slightly different:

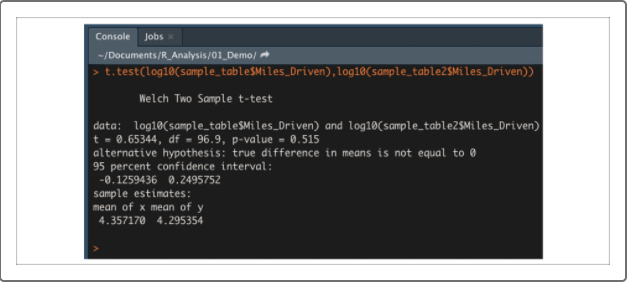
* ***x*** is the first numeric vector of sample data.
* ***y*** is the second numeric vector of sample data.
* **alternative** tells the t.test() function if the hypothesis is one-sided (one-tailed) or two-sided (two-tailed). The options for the alternative argument are "two.sided,""less," or "greater." By default, the t.test() function assumes a two-sided t-Test.

Once we have provided the necessary numeric vectors for each sample, the t.test() function will calculate our two-sample t-Test and return the same output as before. As practice, let's test whether the mean miles driven of two samples from our used car dataset are statistically different.

First, we produce our two samples using the following R statements:

> sample\_table <- population\_table %>% sample\_n(50) #generate 50 randomly sampled data points

> sample\_table2 <- population\_table %>% sample\_n(50) #generate another 50 randomly sampled data points



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:

* H0 : The **difference** between our paired observations (the true mean difference, or "μd") is **equal to** **zero**.
* Ha : The **difference** between our paired observations (the true mean difference, or "μd") is **not equal to zero**.

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

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

