Module 6 Lab

The first part of this lab is in DataCamp. Complete Chapter 5, on **data frames**, in the "Introduction to R" course. Data frames are integral to data analysis in R; pay close attention.

The next two sections will cover **two sample t-tests** and **paired t-tests**.

Two Sample t-test

Suppose a marketing firm is testing the public approval of two new pairs of basketball shoes. They obtain two random samples of numeric consumer ratings of the shoes. We will randomly generate some make-believe samples, but let's set a seed first so we get the same results.

```r
set.seed(1964) # Nike was founded in 1964
shoeA <- rnorm(25, mean = 50, sd = 20)
shoeB <- rnorm(30, mean = 60, sd = 15)
```

Notice we have a different number of samples for each shoe; this is okay because it is **not** a paired t-test. We know the answer becasue we simulted the data, but in practice we wouldn't. The shoe company wants to know whether or not the two shoes have the same average consumer rating.

$$H_0 : \mu_A - \mu_B = 0$$

$$H_A : \mu_A - \mu_B \neq 0$$

We will perform a t-test to find out. Like the one sample t-test, a two sample t-test is conducted using the `t.test()` function in R:

```r
t.test(shoeA, shoeB,
  mu = 0, alternative = "two.sided", paired = FALSE, var.equal = FALSE)
```

The first two arguments are the sample observations, now provided as two vectors, one for each sample. The final four arguments in `t.test()` here are the defaults (so they aren't required), but I include them to highlight the details of our test. The third argument, `mu = 0`, declares that we want to test if the differnce of group means is equal to 0. Next, `alternative = "two.sided"` indicates that the alternative hypothesis is $H_A : \mu_A - \mu_B \neq 0$, rather than a one sided alernative. This is **not** a paired t-test, so we specify `paired = FALSE`. Finally, `var.equal = FALSE` ensures we get the Welch's version of the t-test, which does not make an additional assumption of equal population variances. In general, we do not know the population variances, and it is safer to assume they are not the same. In this context "safer" means that assuming group variances are the same when they are not, could significantly affect results, whereas assuming they are different when they are in fact the same, has minimal impact.

```
##
##  Welch Two Sample t-test
##
## data:  shoeA and shoeB
## t = -2.6985, df = 45.664, p-value = 0.009729
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -25.158439  -3.658598
## sample estimates:
## mean of x mean of y
##  49.52960  63.93812
```

The output is very similar to what we have seen before. The data is listed in the first line, followed by the test statistic, degrees of freedom, and p-value in the second line. The alternative hypothesis is stated next, followed by a 95% confidence interval for $\mu_A - \mu_B$, and point estimates for the individual group means.

We might summarise this particular result as follows:

There is convincing evidence the mean approval rating for shoe A and shoe B are not the same (Welch's t-test, d.f. = 45.6, t = -2.70, p-value = 0.01). We estimate that shoe A has a mean approval rating of 49.5 and shoe B has a mean approval rating of 63.9. With 95% confidence we estimate that shoe B has a mean approval rating between 3.7 and 25.2 units higher than shoe A.

Paired t-test

The shoe marketing executives also want to know if married husbands and wives feel the same way, on average, about shoeA. For this, they conduct a **paired t-test** on a sample of 32 randomly chosen couples.

$$H_0 : \mu_{H-W} = 0$$

$$H_A : \mu_{H-W} \neq 0$$

Let's generate some make-believe data, pretending that there is no difference between paired husband and wife ratings, on average.

```
set.seed(1809) # Carl Friedrich Gauss helps establish the Normal distribution in 1809
means <- rnorm(32, 50, 25) # Create 32 distinct means; one for each couple
wife <- rnorm(32, mean = means, sd = 15)
husband <- rnorm(32, mean = means, sd = 15) # Give each married couple same mean
```

Notice that each married couple has their own distinct mean *shoeA* rating that they share. This means that, though the executives don't know it, the average difference in husband-wife rating is 0. Let's see what the test says. This time, we specifiy `paired = TRUE`, because we perform a paired t-test.

```
t.test(wife, husband, mu = 0, alternative = "two.sided", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  wife and husband
## t = -0.84167, df = 31, p-value = 0.4064
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.305537   4.284485
## sample estimates:
## mean of the differences
##               -3.010526
```

By now the output should look familiar, so let's give a summary.

There is no evidence against the hypothesis that the mean difference in rating between husbands and wives is zero (Paired t-test, t = -0.8417, d.f. = 31, p-value = 0.4064). With 95% confidence, the rating given by husbands is on average between 4.28 units lower and 10.31 units higher than that given by their wives.