Lecture 8: t-tests in R

Contents

Learning Objectives

- Learn how to use R to test for a difference in population means across two groups.
- Reproduce the results that R gives you "by hand" to ensure that you understand where R's results come from.
- Get mini-quiz credit for attendance and effort/progress.

Loading the data

We will work with the `nycflights` dataset from the `openintro` package. Please start by loading the required packages and the dataset.

```
library(tidyverse)
library(openintro)
data(nycflights)
```

Even though this dataset is massive, the `openintro` documentation says that this is a random sample of flights from the population of all flights that departed from NYC in 2013. To see the documentation, you can type `?nycflights` in your console.

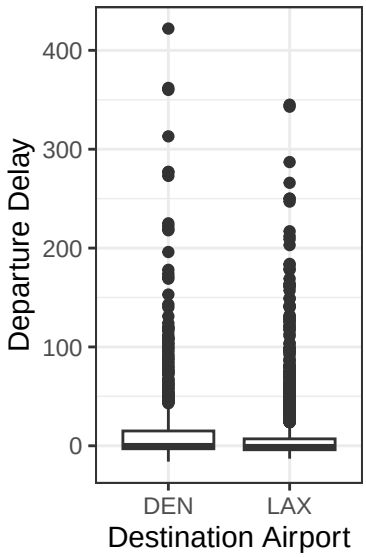
Setting up a research question

I would like to know if flights headed to Denver are more delayed, on average, than flights headed to Los Angeles.

Since my research question relates only to a certain subset of flights in my dataset, my first step is to subset my dataset.

```
nycflights_subsamp <- nycflights %>%
  filter(dest=="LAX" | dest == "DEN")
```

Next, I can either visually or numerically compare the distribution of flights headed to Los Angeles vs. Denver in my sample. I decided to do this with boxplots, but you could do this however you choose. I came up with the following plot:



Based on the plot, I can tell that there is a difference in means between flights headed to `LAX` and `DEN` in my sample. However, I cannot tell if this is driven by a true underlying population difference, or just a few unlucky outliers in my sample.

We will use a t-test to get to the bottom of whether or not there is evidence of an underlying population difference here.

A hypothesis test for a difference in means

Let μ_1 be the population average of delay time for flights headed from NYC to Los Angeles. Let μ_2 be the population average of delay time for flights headed from NYC to Denver. The null hypothesis is

$$H_0 : \mu_1 - \mu_2 = 0.$$

We want to see if our sample provides evidence that $\mu_1 - \mu_2 \neq 0$. We can actually get R to do all of this work for us. The following line of code will conduct a t-test to see if `dep_delay` differs by destination airport in our subsetted dataset.

```
t.test(dep_delay~dest, data=nycflights_subsamp)
```

```
##
##  Welch Two Sample t-test
##
## data:  dep_delay by dest
## t = 3.4621, df = 1118.2, p-value = 0.0005562
## alternative hypothesis: true difference in means between group DEN and group LAX is not equal to 0
## 95 percent confidence interval:
##   2.816468 10.184631
## sample estimates:
## mean in group DEN mean in group LAX
##         16.282609          9.782059
```

When you run this code, it prints out a lot of information for you! You may or may not be able to understand the output.

Your task in groups

Your task is to reproduce all of the numbers in the `t.test()` output using the dataset and the theory that we learned in class.

1. Find the number 16.28 in your dataset, and save this to a variable called `xbar1`.

Hint: you are not supposed to write:

```
xbar1 <- 16.28
```

You are supposed to show how you would get this from your dataset.

2. Find the number 9.78 in your dataset, and save this to a variable called `xbar2`.
3. R is basing this t-test off of a sampling distribution, whose standard deviation is supposed to be

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

   However, σ_1 and σ_2 are both unknown. Obtain estimates from them using your data, and save the result to variables valled s_1 and s_2 in R.

Next, you need to figure out what *standard error*