

Lecture 8: t-tests in R

Contents

Loading the data	1
Setting up a research question	1
A hypothesis test for a difference in means	2
One more airport	2

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.

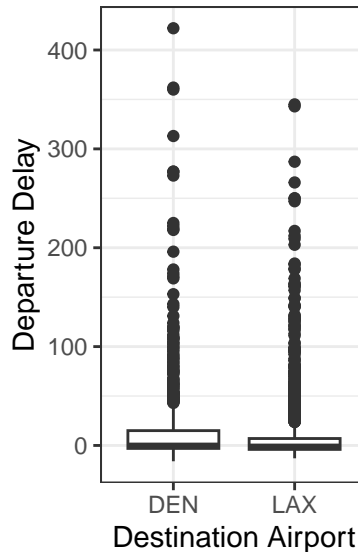
Setting up a research question

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.

1. Are the conditions met to do a t-test for a difference in means?

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 LA/SF-only dataset.

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

This prints a lot of information for you! Your task is to reproduce all of the numbers in this output. The task is broken

One more airport

Do flights going to Denver tend to be more delayed, on average, than those going to LAX? The following visual seems to suggest a difference.

```
nycflights_subsamp <- nycflights %>% filter(dest=="LAX" | dest == "DEN")
ggplot(nycflights_subsamp, aes(x=dest, y=dep_delay))+geom_boxplot()+ylab("Departure Delay")+xlab("Destin
```

Using R, conduct a t-test to see if the difference in means that we see in our sample provides convincing evidence of a true difference in population means. Write down your answer and interpret your results. Compare your results here to your results for Denver vs. LA. What could explain the difference?