# Lab 7 - Math 58 / 58b: chi-sq tests

*your name here*

*due Mar 10, 2020*

## Lab Goals

- Run a chi-sq goodness-of-fit test using:
    - `infer` with simulation
    - `chi.test` as a function
- Run a chi-sq test of independence using:
    - `infer` with simulation
    - `chi.test` as a function

## Getting started

### Load packages

For the lab, we'll use the `infer` syntax and the `flights` data.

```
library(tidyverse)
library(nycflights13)
library(infer)
```

### Load the data

Recall the `flights` data from a previous lab. The Bureau of Transportation Statistics (BTS) is a statistical agency that is a part of the Research and Innovative Technology Administration (RITA). As its name implies, BTS collects and makes available transportation data, such as the flights data we will be working with in this lab.

If you have forgotten what the `flights` data contains, you should look at it!

```
library(nycflights13)
data(flights)
```

## To turn in

The four biggest airline carriers in the US (percent of domestic passengers) are Southwest (WN 32.6%), Delta (DL, 25.6%), American (AA, 24.7%), and United (UA, 17.1%).[1] (Where the percent is measured out of those four airlines, not out of the total.)

First, filter the dataset (using `filter()`) to include only those flights whose carrier was one of the four biggest airlines. Be sure to save the new dataset.

Note how `filter()` keeps only the observational units where the variable (`carrier`) has categories `%in%` a column representing the 4 carriers of interest.

Also, I took a random sample of the flights, just so that we'd all be working with the same slightly smaller dataset.

```
set.seed(4747)
flights4 <- flights %>%
  filter(carrier %in% c("AA", "DL", "WN", "UA")) %>%
  sample_n(1000)
```

[1] 2017 https://www.bts.dot.gov/newsroom/2018-traffic-data-us-airlines-and-foreign-airlines-us-flights

### One categorical (>2 level) - GoF

https://infer.netlify.com/articles/observed_stat_examples.html#one-categorical-2-level---gof

1. Test whether the flights out of NYC (that is, the dataset at hand, do not do any additional filtering) have the same proportions of carriers with market share as specified by domestic passengers using:

(a) Use the `table()` function to guess whether or not the data will be consistent with your hypothesis. Explain your guess in a sentence or two.
(b) `infer`
(c) `chisq.test` [not `infer`!]. The steps for using `chisq.test` are:

- create a column with the relevant proportions to test. Call that column `testprops`.
- using the `flights4` select only the `carrier` column … AND THEN
- use `table()` to tabulate the levels AND THEN
- use `chisq.test(p = testprops)` (where the very first argument has been piped in as above, check the output of table to make sure your testprops are in the right order!)

**Hint:** do you know what the `table()` function does? If not, try it out before you pipe it into the `chisq.test()` function. Test out your code line by line.

2. Are the two conclusions in #1 the same? Provide the conclusions to your hypothesis test in the words of the problem.

### Two categorical (>2 level) variables

https://infer.netlify.com/articles/observed_stat_examples.html#two-categorical-2-level-variables-1

3. Given the flights out of NYC (that is, the dataset at hand, do not do any additional filtering) test whether the distribution of carrier (from the 4 major carriers) the same across the 3 NYC airports.

(a) Use the `table()` function to guess whether or not the data will be consistent with your hypothesis. Explain your guess in a sentence or two.
(b) `infer`
(c) `chisq.test` [not `infer`!]. The steps for using `chisq.test` are:

- using the `flights4` select the `carrier` and `origin` columns … AND THEN
- use `table()` to tabulate the levels AND THEN
- use `chisq.test()` (where the very first argument has been piped in as above!)

4. Are the two conclusions in #3 the same? Provide the conclusions to your hypothesis test in the words of the problem.