Homework 1

Scott Girten

All homework will be submitted via R Markdown files.

### Question 1

Complete problem 3.5 (page 71) in the text. The two requested hypothesis tests should follow the format below:

* Define the parameters of interest and state the hypotheses to be tested.
* Define the test statistic used to test the claim.
* Simulate the null distribution of the test statistic, including a histogram of the distribution with the observed test statistic identified.
* Caclulate the p-value of the test.
* Report a conclusion for the test.

= mean delay time for American Airlines

= mean delay time for United Airlines

**Test Statistic:** if the mean delay times for American Airlines and United Airlines is statistically significant, the test statistic can be used.

Mean of = 1.1681562

Standard Deviation of = 0.8844607

**p-value Calculation**

The p-value for is 210^{-4}

### Question 2

Complete problem 3.7 (page 71). The problem asks you to repeat part a of question 3.5 using 3 different test statistics. Please make sure the caclulation of the 3 test statistics all occurs within the same FOR loop.

### Question 3

The Department of Tourism in Illinois wants to determine if there is a difference in the mean cost of a taxi ride for customers paying by cash and those paying by credit. A sample of 150 taxi rides is selected, and the total fare and method of payment is recorded. The data for 78 cash payments and 72 credit payment is below:

Suppose a permutation test was conducted on this data to determine if there is a difference in the mean taxi fare for the two payment methods.

* If the *exact* null distribution were constructed, how many unique resamples exist?
* With a much larger pool of resamples, let’s investigate the effect number of resamples has on the mean and varaibility of p-values from the permutation test. Write code to generate output that will allow the table below to be filled in based on the simulation of 1000 p-values (note to save computing time, I’ve provided the values for 20,000 resamples):

|  |  |  |
| --- | --- | --- |
| Number of Resamples | Mean of the p-values | Std Dev of the p-values |
| 1000 |  |  |
| 5000 |  |  |
| 10,000 |  |  |
| 15,000 |  |  |
| 20,000 | 0.8837033 | 0.002207726 |

* Breifly comment on the pattern observed in the means and standard deviations of the p-values.

The data refereneced in the above questions is accessible at this [link](https://sites.google.com/site/ChiharaHesterberg). Individual files can be downloaded from the [Data sets](https://sites.google.com/site/chiharahesterberg/data2) link. Or, an R package has been created with all the data from the text included. The package can be installed with the following code:

The data you’ll be using contains flight delay information for United Airlines and American Airlines departures out of LaGuardia. Information on the variables contained in the datset is given in Section 1.1 of the textbook.

The hypothesis tests you’ll be conducting will required you to look at subsets of the dataset. For example, suppose we wanted to create a vector of delay times for all United Airlines flights. Assuming the data has been loaded as a Data Frame with the name FlightDelays, the following code would accomplish this: