Intro to Data Science
Homework 4: Due Wednesday October 23 at 2:00pm

Exercises:

Many of the exercises for this assignment ask you to use the resampling methods (bootstrap and permutation tests) introduced in class in the R notebook basicStatsNotebook.

1. This problem concerns permutation testing. Suppose you conduct an experiment and inject a drug into three mice. Their times for running a mze are 8, 10, and 15 seconds; the times for two control mice are 5 and 9 seconds.

   (a) Compute the difference in mean times between the treatment group and the control group.

   (b) The following command will produce all possible permutations (note that there are 5! =120) of the values 8, 10, 15, 5, and 9:

   ```
   perm_res <- permutations(5,5,c(8,10,15,5,9))
   ```

   Note that you must load the library gtools for this to work. Use this to compute the difference in means between the first three and last two values in each of the possible permutations. Here is one way to do this:

   ```
   perm_res <- as.data.frame(perm_res)
   perm_res <- perm_res %>%
     mutate(mean1=(V1+V2+V3)/3,mean2=(V4+V5)/2,mean_diff=mean1-mean2)
   ```

   Note that this adds three columns, the last of which is the difference in means:

   ```
   ##   V1 V2 V3 V4 V5    mean1 mean2  mean_diff
   ## 1  5  8  9 10 15 7.333333  12.5 -5.1666667
   ## 2  5  8  9 15 10 7.333333  12.5 -5.1666667
   ## 3  5  8 10  9 15 7.666667  12.0 -4.3333333
   ## 4  5  8 10 15  9 7.666667  12.0 -4.3333333
   ## 5  5  8 15  9 10 9.333333   9.5 -0.1666667
   ## 6  5  8 15 10  9 9.333333   9.5 -0.1666667
   ```

   (c) What proportion of the differences are as large or larger than the observed difference in mean times?

   (d) Make a histogram (choose your binwidth wisely) of the difference in means and add a vertical line to the plot corresponding to the value for the observed difference in mean times

2. In this exercise you will conduct a permutation test on the difference of mean delay times for the FlightDelays dataset found in the resampledata package (make sure you install and load this package). The FlightDelays data contain flight delays for two airlines, American Airlines and United Airlines.

(a) Conduct a two-sided permutation test to see if the difference in mean delay times between the two carriers are statistically significant. Plot a histogram to illustrate your results. (Note: You do not need to consider all possibe permutations, just a large number of samples with replacement as we did in class.)

(b) The flights took place in May and June 2009. Conduct a two-sided permutation test to see if the difference in mean delay times between the two months is statistically significant.

3. This exercise uses the Bangladesh dataset from the resampledata package.

(a) Conduct and EDA on the chlorine concentrations (recordedin the Chlorine variable) and describe the salient features.

(b) Bootstrap the mean chlorine concentration.

(c) Plot the bootstrap distribution for the mean.

(d) Find and interpret the 95% bootstrap percentile confidence interval.

4. What is normal body temperature? The standard has been 98.6 degrees Fahrenheit. Suppose a medical worker suspects that body temperatures in children in Sodor are higher than the norm. She obtains measurements from a random sample of 18 children and finds the following: 98, 98.9, 99, 98.9, 98.8, 98.6, 99.1, 98.9, 98.5, 98.9, 98.9, 98.4, 99, 99.2, 98.6, 98.8, 98.9, 98.7.

(a) What are the hypotheses to test?

(b) Carry out the test, and state a conclusion in a complete sentence.

(c) Make relevant plots to display your results.

5. The data set Walleye in the resampledata package contains data from the Minnesota Pollution Agency recording the lengths in inches and weights in pounds for 60 walleye caught in the Minnesota lakes during the 1990's. There is some suspicion that on average, walleye are smaller now than in the past. Suppose historical records from the early 1900's indicate that the average weight of walleye caught by fishermen was 2.5lb. Assume the data are representative of caught walleye; do they indicate that walleye from the 1990's weight less?

(a) Create a histogram of walleye weight and describe the distribution.

(b) Use the bootstrap t-test to carry out an appropriate hypothesis test. What do you conclude?

(c) Plot a histogram of the bootstrap t-distribution and indicate with a vertical line the observed t-statistic value.