Stat 311 Homework 5

Most questions in this assignment require R, so do this assignment in Rmarkdown and uploaded as a knitted html file to Canvas.

1. Your local grocery store sells 1 lb. bags of potatoes. However, the 1 lb. bags do not weigh exactly 1 lb. If we let X_i be the weight of a randomly selected 1 lb. bag of potatoes, historical data indicates that $X_i \sim N(1.09, 0.10)$. A local cash and carry store sell 5 lb. bags of potatoes, which also do not weigh exactly 5 lbs. If Y is the weight of a randomly selected 5 lb. bag, historical data indicates that $Y \sim N(5.12, 0.18)$. If we randomly select five 1 lb bags of potatoes and one 5 lb bag of potatoes, what is the probability that the sum of the weights of the five 1 lb. bags exceeds the weight of one 5 lb. bag? **For this problem you can use R to get the probability, but you must show some work to convince us you know what you are doing to solve this problem.**

2. Recall the zone out duration (ZOD) data we looked at in one of the regression lectures from Lesson 3. An additional experiment was conducted to look at the impact of sugary desserts eaten at lunch, two hours before class, and ZOD. Twelve students volunteered to participate in the experiment. Students were randomly assigned to eat a large slice of apple or cherry pie, with six participants randomized in each group. Two hours later, their ZODs (in minutes) were recorded during a 50-minute lecture. The data are in the file ZODTwoGroups.csv.

   a) Make a comparative boxplot for ZOD by pie type. Describe what you can get from the boxplots regarding the two distributions. Does there appear to be a difference between the ZODs for the two groups?

   b) Use `set.seed(25)` and then create 1000 permutations for the difference of mean ZOD for cherry pie minus the mean ZOD for apple pie. What is the observed difference in means for the sample data?

   c) Write out the statistical hypotheses, using symbols, for testing that mean ZOD for cherry pie is greater than the mean ZOD for apple pie.

   d) Make a histogram of the null distribution and add a vertical line for the observed sample difference. **Set the number of bins to 13 if you use ggplot2 for the histogram**. Describe the shape of the null distribution and how the observed sample difference compares with the overall distribution.

   e) Calculate the *p*-value for this permutation test. If you set up your code correctly, you should get a small *p*-value in the range of ≤ 0.005 or so. What is the meaning of this *p*-value as a probability?

   f) What do you conclude for this hypothesis test in the context of the problem?

Stat 311 Homework 5

3. Recall the popular diets data set (PopularDiets.csv) that we have used before. We created a new data set called PopularDietsCombined.csv. It only contains observations for participants that completed the study, and it only contains the two variables WTLossKG (weight loss in kilograms) and Diet.

   a) Read in the data and make a quick boxplot to refresh your memory that there did not seem to be much difference in weight loss by diet. We provide the read statement in Homework5Template.Rmd. Use the boxplot function in base R to get a quick comparative boxplot of weight loss by diet. No need to write anything here.

   b) Since there does not seem to be too much difference by diet type, we will only work WTLossKG. What is the point estimate for mean weight loss across all diets?

   c) Using all 93 observations across all diets, create 1000 bootstrapped samples. Display a histogram of the bootstrapped distribution for mean weight loss. **Set the number of bins to 10 if you use ggplot2 for the histogram**. Describe the shape of the distribution.

   d) Report the 95% bootstrap confidence interval for mean weight loss and provide an interpretation of this interval in the context of the problem. Use `set.seed(25)`.

   e) Report the 90% and 99% bootstrap confidence intervals for mean weight loss from the same bootstrap sample. How do these intervals compare with the 95% interval reported in part (d)? If you redraw your bootstrap samples, use `set.seed(25)` before each set of draws.

   f) If you were to draw a new set of 1000 bootstrap samples and find the 95% bootstrap confidence interval, how do you think the new interval will compare with the interval reported in part (d)? Answer this before doing part (g) below. This question is effort only.

   g) Draw two new sets of 1000 bootstrap samples using the 93 observations and report the 95% bootstrap confidence interval for mean weight loss for each (no need to plot the two new bootstrap distributions and **do NOT use `set.seed`**). Do these intervals support what you wrote for part (f)? Explain.

   h) If you drew smaller sized bootstrap samples and calculated the 95% bootstrap confidence interval, how do you think the new intervals would compare with the interval reported in part (d)? Answer this before doing part (i) below. This question is effort only.

   i) Use `set.seed(25)` and draw three smaller sets ($n = 500, 100, \text{and } 10$) of bootstrap samples using the 93 observations (be sure to reset the seed before each bootstrap draw). Report the 95% bootstrap confidence intervals for mean weight loss for each of these new sets. Do these intervals support what you wrote for part (h)? Explain.

   j) Now make two subsets of the data, one for the Ornish diet and one for the Weight Watchers diet. Use `set.seed(25)` and draw 1000 bootstrap samples to report a 95% bootstrap confidence interval for mean weight loss for each diet (so two CIs). Reset `set.seed(25)` for each diet type. Compare the intervals. What can you say about mean weight loss for the Ornish diet compared with the Weight Watchers diet?