Homework 5

Dr. Purna Gamage

Problem 1 (20 points, 5 for each)

Problem 5.10 #12 in Chihara/Hesterberg.

The data set FishMercury contains mercury levels (parts per million) for 30 fish caught in lakes in Minnesota.

(a) Create a histogram or boxplot of the data. What do you observe?

(b) Find the Bootstrap sampling mean and record the bootstrap standard error and the 95% bootstrap percentile interval.

(c) Remove the outlier and find bootstrap sampling mean of the remaining data. Record the bootstrap standard error and the 95% bootstrap percentile interval. Comment on your results.

(d) What effect did removing the outlier have on the bootstrap distribution, in particular, the standard error?

Problem 2 (20 points, extra bonus 5 points inside the question)

Problem 3.9 #12abc in Chihara/Hesterberg.

Two students went to a local supermarket and collected data on cereals; they classified cereals by their target consumer (children versus adults) and the placement of the cereal on the shelf (bottom, middle, and top). The data are given in *Cereals*.

(a) (*2 points*) Create a table (Two-way) to summarize the relationship between age of target consumer and shelf location.

(b) (*3 points*) Conduct a chi-square test using R's chisq.test() command. Write your null and alternative hypthesis. What is your conclusion based on the results of your test?

(c) (*2 points*) R returns a warning message. Compute the expected counts for each cell to see why.

(d) (*3 points*) Use a Yate's continuity correction and do the test again. What is your conclusion?

(e) (*5 points*) (self-learn question) Use a Fisher's Exact Test. What is your conclusion?. (We use a Fisher's Exact Test when the sample sizes are small and the expected cell counts are less than 5 : Example to refer:https://statsandr.com/blog/fisher-s-exact-test-in-r-independence-test-for-a-small-sample/).

(f) (*5 points*) Compare your results of part (b),(d),(f). Explain/Compare in few sentences where/when/what situations should we use Yate's Continuity correction and Fisher's exact test.

Problem 3 (15 points)

Distribution A is a standard normal distribution and distribution B is a $N(1, 2^2)$ *distribution. Generate 20 random numbers from distribution A and 30 random numbers form distribution B and record these in a suitable data frame.*

Examine the null hypothesis that the means of A and B are the same against the alternative that the mean of B is larger, using a permutation test. Report the p-value and state your conclusion.

Problem 4 (20 points)

This problem is similar to what we have done in the lab using the spotify data. Please use the "Artists.csv" data set.

Data Science Question: Does the average "liveness" is larger for Beyoncé than that of Taylor Swift?

Liveness: This value describes the probability that the song was recorded with a live audience. According to the official documentation "a value above 0.8 provides strong likelihood that the track is live".

a. Perform meaningful EDA (Exploratory Data Analysis) using some Data Visualizations (relevant to this data science question).

b. Write the null and alternative hypothesis for this test.

c. Perform a t-test and state your results and non-technical conclusion.

d. What can you say about the confidence interval? (Interpret)

e. Perform a bootstrap test for ratio of means of "liveness", Find the 95% bootstrap percentile interval for the ratio of means and write your conclusion.

f. What is the bootstrap estimate of the bias for the mean ratio?

g. Compare your results from part c) and part e).

Problem 5 (15 points)

Write an R function that computes the t-formula confidence interval in (7.8) from sample mean, sample standard deviation, sample size, and confidence level, and use it to do exercise 7.6 #6 in Chihara/Hesterberg.

Q: Julie is interested in the sugar content of vanilla ice cream. She obtains a random sample of $n = 20$ brands and finds an average of $18.05g$ with standard deviation $5g$ (per half cup serving). Assuming that the data come from a normal distribution, find a 90% confidence interval for the mean amount of sugar in a half cup serving of vanilla ice cream.

Problem 6 (15 points, 5 for each)

Exercise 7.6 #12 in Chihara/Hasterberg.

Q: Consider the data set *Girls2004* (see Case Study in Section 1.2).

(a) Create exploratory plots and compare the distribution of weights between babies born to nonsmokers and babies born to smokers.

(b) Find a 95% one-sided lower t confidence bound for the mean difference in weights between babies born to nonsmokers and smokers. Give a sentence interpreting the interval.

(c) What is your conclusion?

BONUS: Submit *ONE* of the Extra Problems:

Ex. Problem 1 (10 Points)

Exercise 6.4 #1 in Chihara/Hesterberg.

Let X be a binomial random variable, $X \sim Binom(n, p)$. Show that the MLE of p is $\hat{p} = X/n$.

Ex. Problem 2 (10 Points)

Exercise 6.4 #14 in Chihara/Hesterberg.

Let the five numbers 2, 3, 5, 9, 10 come from the uniform distribution on $[\alpha, \beta]$. Find the method of moments estimates of α and β.