Homework Assignment: Week 2

Assignment Due: Sunday, January 16, 2022 by Midnight CT

Submit electronic solutions to drop box in BrightSpace

1. Define selection bias and list two or three ways in which selection bias can influence the outcome of an analysis.

Selection bias is when selection takes place in such a way that true randomization is not achieved. Therefore, it fails to actually represent the larger population intended to be analyzed.

One way that selection bias can influence the outcome of an analysis is when a sample of the data is being selected for analysis. If the selection process is done in a way that causes some observations to be less likely to be selected, then there is bias. The sampling data will not reflect the true distribution of the larger set. Another way selection bias can influence an analysis is called volunteer bias. Volunteer bias happens when a study is being conducted with participants who volunteer to participate. It’s been proven that those who volunteer for studies comes from higher socioeconomic backgrounds. In addition, women are more likely to volunteer than men. If studies are conducted purely off volunteers, then their sample data will not reflect the larger set. This will lead to inaccurate results.

2. Briefly explain the concept of the data lake.

A data lake is large storage destination that stores an entity’s data. Often times, data is extracted from the lake, cleaned (transformed), and loaded into a data warehouse to be used. A data lake is named rightly so because data is being sent to the lake from many different sources and is susceptible to losing some of that data stored in the lake.

3. Look up the definition of Frequentist and Bayesian Statistics and rewrite them in you own words.

Frequentist Statistics is statistics centered on being able to repeat experiments given an underlying assumption.

Bayesian Statistics is statistics that predict future outcomes based on what has already happened.

4. Use the data set state.area in R and determine the mean, median, max, min, range, and standard deviation of the data set. Determine the skewness and kurtosis values and using these, describe the distribution. Turn in the R files with the code for this.

Based on finding the above values of the data, we can determine that the distribution of the data is very similar to exponential decay. With a relatively large kurtosis value and a skewness of ~4.19, we can determine that a histogram of the data will have a large value at the beginning with a steep drop right after – similar to an exponential decay graph.

5. Modify the R code for bootstrapping to determine the 90% confidence limits for the state.area data set. Report the confidence limits. Why would bootstrapping be the best method for determining the confidence limits for this data?

After bootstrapping the data, we have a 90% confidence interval of 45,009.64 – 111,223. Bootstrapping is the best method for determining the confidence limits because state.area data is not a normal distribution. By using bootstrapping, we can easily resample our data and create a normal distribution for future analyses.

6. Grade Point Average (GPA) data was collected from 12 students (see below). Using R, preform a one – sample t.test and report the confidence interval and the p-value. Explain why or why not the null hypothesis could be rejected.

GPA data: 2.01, 3.96, 1.92, 2.71, 3.03, 3.15, 3.22, 3.17, 3.99, 2.63, 2.51, 2.63

Running a one-sample t test with a 95% confidence interval returns the interval of ~2.50 - ~3.32. The p-value returned is 8.032e-09. With this information, we can reject the null hypothesis because the p-value is lesser than 0.05.

7. Explain Type I and Type II errors in hypothesis testing and why they are inversely proportional to one another.

Type 1 errors are when we reject the null hypothesis incorrectly. For example, if alpha=0.05, there’s a 5% probability that we will incorrectly reject the null hypothesis. Type II errors occur when we retain the null hypothesis even when we should reject. This is the beta value, and it is difficult to calculate or estimate.

Type I and Type II errors are inversely proportional because if we make one error smaller, we’ll make the other one bigger. This means that we cannot minimize both types of errors.

8. Data were collected on two athletes 100 meter dash times (see below) during five recent major competitions. Neither athlete has ever run in a head-to-head competition. You would like to predict which one is expected to be faster. Perform a hypothesis test, assuming equal variance, to determine if you can state that one athlete is faster than the other and if so, which one.

Times of Athlete 1 in seconds. 9.79, 10.11, 9.99, 10.08, 10.22

Times of Athlete 2 in seconds: 9.61, 10.31, 10.02, 10.38, 9.81

Explain the results and justify the acceptance or rejection of the null hypothesis.

After running the hypothesis test, we receive a p-value of 0.943. This means that we should accept the null hypothesis, or in other words, accept the hypothesis that the means of these two vectors are the same. In addition, we cannot definitively decide which athlete is faster of the two.

9. Briefly describe the pitfalls of hypothesis testing.

Hypothesis testing brings a variety of pitfalls. To start, we must assume that the null hypothesis is true. If we don’t end up rejecting the null hypothesis, this means we haven’t found sufficient evidence, and we’ve made a non-decision. Another pitfall is never knowing the population. Since we do not know the population, we cannot classify when we make Type I or II errors. Hypothesis testing is useful is gaining confidence in parameters we are testing, but it falls short when are trying to accept or reject the null hypothesis.

10. Use the CLTLLN code file to generate an iid distribution of a 5 – sided die with a size of 10,000. Find the skewness and kurtosis of this distribution. Do the values calculated align with your expectations for this distribution? Explain.

Yes, the values calculated for skewness and kurtosis align with my expectations for this distribution. Generating a distribution of a 5-sided die with a size of 10,000 means that we’ll have a flat distribution across all the values. The probability of rolling each value is constant for each value.

After testing this, I receive a skewness value of -0.022 and a kurtosis value of -1.286. This makes sense because the data is very slightly skewed in a negative direction. It appears the values of 1 and 2 were rolled very slightly less than 3, 4, or 5. In addition, the kurtosis value makes sense for the same reason as the skewness value. There is a slight taper in the tails for the values of 1 and 5.