INFO 5502: Assignment 6

Shristi Dhakal

# 1. The Bootstrap and The Normal Curve

In this exercise, we will explore a dataset that includes the safety inspection scores for restaurants in the city of Austin, Texas. We will be interested in determining the average restaurant score for the city from a random sample of the scores; the average restaurant score is out of 100. We'll compare two methods for computing a confidence interval for that quantity: the bootstrap resampling method, and an approximation based on the Central Limit Theorem.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

# 1.1 Question 1

Complete the function one\_resampled\_mean below. It should take in an original table data, with a column Score, and return the mean score of one resampling from data.

A screenshot of a cell phone

Description automatically generated

# 1.2 Question 2

Complete the function bootstrap\_scores below. It should take no arguments. It should simulate drawing 5000 resamples from restaurant\_sample and compute the mean restaurant score in each resample. It should return an array of those 5000 resample means.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

# 1.3 Question 3

Compute a 95 percent confidence interval for the average restaurant score using the array resampled\_means.

A screenshot of a cell phone

Description automatically generated

# 1.4 Question 4

What distribution is the histogram between question 2 and 3 displaying (that is, what data are plotted), and why does it have that shape?

>> The histogram is displaying the means of 5000 bootstrapped resamples drawn with replacement from our original restaurant samples of 100, which is also drawn with replacement. It is roughly normally distributed because of the Central Limit Theorem i.e. if we repeatedly calculate the arithmetic average of samples drawn randomly from a population with replacement, the probability distribution of those averages tends to follow a normal distribution even if the population's distribution is not normal.

# 1.5 Question 5

Does the distribution of the sampled scores look normally distributed? State “yes” or “no” and describe in one sentence why you should expect this result.

>>No, the distribution of the sampled scores does not look normally distributed, instead it is left-tailed with most observations centered towards the right (close to the value of 100). This is because the distribution of samples does not necessarily equate to the distribution of mean of the samples, hence in this case, the Central Limit Theorem does not apply.

# 1.6 Question 6

Without referencing the array resampled\_means or performing any new simulations, calculate an interval around the sample\_mean that covers approximately 95% of the numbers in the resampled\_means array. This confidence interval should look very similar to the one you computed in Question 3.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

# 2 Testing the Central Limit Theorem

To recap the properties, we just saw: The Central Limit Theorem tells us that the probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, regardless of the distribution of the population from which the sample is drawn.

# 2.1 Question 1

Define the function one\_statistic\_prop\_heads which should return exactly one simulated statistic of the proportion of heads from n coin flips.

A screenshot of a cell phone

Description automatically generated

# 2.2 Question 2

The CLT only applies when sample sizes are "sufficiently large." This isn't a very precise statement. Is 10 large? How about 50? The truth is that it depends both on the original population distribution and just how "normal" you want the result to look. Let's use a simulation to get a feel for how the distribution of the sample mean changes as sample size goes up.

Consider a coin flip. If we say Heads is 11 and Tails is 00, then there's a 50% chance of getting a 1 and a 50% chance of getting a 0, which definitely doesn't match our definition of a normal distribution. The average of several coin tosses, where Heads is 1 and Tails is 0, is equal to the proportion of heads in those coin tosses (which is equivalent to the mean value of the coin tosses), so the CLT should hold true if we compute the sample proportion of heads many times.

Write a function called sample\_size\_n that takes in a sample size 𝑛n. It should return an array that contains 5000 sample proportions of heads, each from 𝑛n coin flips.

A screenshot of a cell phone

Description automatically generated

# 2.3 Question 3

Write a function called empirical\_sample\_mean\_sd that takes a sample size n as its argument. The function should simulate 500 samples with replacement of size n from the flight delays dataset, and it should return the standard deviation of the means of those 500 samples.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

# 2.4 Question 4

Now, write a function called predict\_sample\_mean\_sd to find the predicted value of the standard deviation of means according to the relationship between the standard deviation of the sample mean and sample size that is discussed here in the textbook. It takes a sample size n (a number) as its argument. It returns the predicted value of the standard deviation of the mean delay time for samples of size n from the flight delays (represented in the table united).

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

>>The predicted value of the standard deviation of the mean delay time for samples of size 10 was around 12.29 whereas the predicted valued of the same for sample size 500 was around 1.77. This reinforces the core idea of the standard deviation that as the sample size increases, the standard deviation tends to decrease.