Lab 3 – Simulation Study with Diamonds

NAME 1 – NETID

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Formatting Instructions

- Please number your answers (leaving the original question text is optional), save your work as a pdf, and submit to Gradescope.
- If working in a group, be sure that all **group members** are **added** in your submission to Gradescope.
- When you upload to Gradescope, please **match pages** with the **question number**.

Assignment Overview

- For this lab, we're going to explore distributions and the variability of a sample statistic through simulation!
- To do this, we will explore the diamonds dataset saved in the ggplot2 package. This dataset has over 59,000 diamonds catalogued, and we will treat this dataset like it's a population.
- Let's see how much variation we see from sample to sample and how reliable our sample statistics are in different situations!

Step 0

- Open RStudio (or RStudio Cloud) to get started
- We will be using the `diamonds` dataset stored in the tidyverse package. So start by running `library(tidyverse)`
- Open the `diamonds` data by running the code: `View(diamonds)`. Each row represents one diamond from a collection of over 59,000.
- Take a look at the documentation for `diamonds` by running the code: `?diamonds`

Question 1 (5pts): Create a histogram of the `price` variable (*For all histograms in this assignment, use the base R function `hist`).* Also calculate the mean and standard deviation of this variable.

Include the image of your histogram in your report *(you may either save it to your computer and upload it, or include a properly cropped screenshot).*

Include the mean and standard deviation values

Briefly describe the shape of this variable. Is this a symmetric distribution or would you say it's skewed?

Question 2 (5pts): Take a random sample of 50 diamond prices from this dataset and name this vector `fifty_diam` *(If saved properly, you will see this vector of length 50 saved in your global environment!).* Sample without replacement (this will be the default option). Create a histogram of your sample, and then calculate the mean and standard deviation of this sample.

Include the image of your histogram in your report

Include the mean and standard deviation values

What is the *absolute* error of your sample mean as an estimate of the true mean? *(absolute error is just distance between your measurement and the true value: https://www.statisticshowto.com/absolute-error/)*

What is the *absolute* error of your sample standard deviation (SD) as an estimate of the true SD?

Question 3 (5pts): Next, set up a `for` loop to simulate taking a sample of size 50 at least 10,000 times. Inside your loop, calculate the mean price and save it to a vector called `means`. Here are two tips:

- Remember before the loop to define `means = NULL` so that your loop knows where to save the means.
- Remember inside the loop to include an index indicator with your means vector so that the vector fills iteratively for each iteration of the loop.
- Try running the loop 10 times to ensure it works. This should be instantaneous. Then try running it 10,000 times. The loop should only take a few seconds to complete at 10,000 simulations, so if you wait more than a minute, click the stop button and see if something is defined incorrectly.

After successfully running your simulation, create a histogram of your `means` vector. Again, continue to use the `hist()` function for all histograms in this lab.

Include the image of your histogram in your report

Include the R code you used to generate this loop

Briefly describe the shape of your histogram. Is this a symmetric distribution or would you say it's skewed? How does this relate to the Central Limit Theorem we learned in class?

Question 4 (5pts): As you should notice from your histogram, our sample means will vary with each sample we take. Calculate the standard deviation of the `means` vector.

Report the standard deviation of the simulated means

Notice that every time you run your loop again, your vector of sample means will change, and so will your standard deviation of those simulated sample means. What measure is the standard deviation of the simulated means approximating? **Report the <u>name</u> of this measure and calculate the <u>true value</u> for this measure too** *(hint: check the "Distribution of a Sample Statistic" notes!)*

Question 5 (5pts): Repeat question 3, but with a sample size of 10 instead of 50. Call your vector of sample means `means_ten`. After successfully running your simulation, create a histogram of your `means_ten` vector using the `hist` function again.

Include the image of your histogram in your report

Include the R code you used to generate this loop

Briefly describe the shape of your histogram. Is this a symmetric distribution or would you say it's skewed? How does this relate to the Central Limit Theorem we learned in class?

Is the standard deviation of the simulated means higher or lower than it was for n = 50?

Question 6 (5pts): We spent some time exploring the behavior of the sample mean, but now let's look at the **sample median**! Redo question 3 with a sample size of 50, but now calculate the sample median inside your loop. Call your vector of sample medians `medians_fifty`. After successfully running your simulation, create a histogram of your `medians_fifty` vector using the `hist` function again.

Include the image of your histogram in your report

Include the R code you used to generate this loop

Briefly describe the shape of your histogram. Is this a symmetric distribution or would you say it's skewed? Do you have any predictions for what would happen if we repeated this simulation again, but with a much larger sample size?

Calculate and report the standard deviation of the `medians_fifty` vector. *This is the expected error in a randomly generated sample median as an estimate of the true median.*

Question 7 (5pts): Repeat question 6, but with a sample size of 500.

Include the image of your histogram in your report

Include the R code you used to generate this loop

Briefly describe the shape of your histogram. How has the shape changed in comparison to the distribution of sample medians when we took samples of size 50?

Calculate and report the standard deviation of your newest vector of medians. **How does this expected error compare to when we had samples of size 50? Is this expected or surprising to you?**