Lab 1 – Simulation Study with Diamonds

NAME 1 – NETID

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Formatting Instructions

- Please include all requested responses in a document, then save it as a **pdf** when done.
  - You may use this instructions document, or you may create a new document.
  - All responses should be numbered (leaving the original question text is optional!)
- Upload your pdf to **Gradescope** and please **match pages** with the **question number** when prompted to.
- If working with one or two **partners**, be sure to do **both** of these things:
  - Please put all **names and netIDs** at the top of your document (like shown above).
  - Have one person upload the pdf and then ensure **group members** are **added** in your submission to Gradescope (click view/edit group on the top right of the page once shown your final submission after matching pages).

Assignment Overview

- For this lab, we will explore the diamonds dataset stored in the tidyverse 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 – *Come to Lab day, or ask for help if you get stuck somewhere in Step 0!*

- **Pre-lab work**
  - Complete the pre-lab tutorials for Lab 1: https://stat212-learnr.stat.illinois.edu/
  - Watch videos 1 (or 2), 3, 4, and 5 in this playlist: https://www.youtube.com/playlist?list=PLTE0IJCCTM9ILfW8OaLqZd37G7X4WDtl-
- **Open RStudio** (or RStudio Cloud) to get started
  - Be careful **not** to open up **R** (this icon with just R and a swirly thing on the left).
  - Open up **RStudio** (this icon with the blue circle on the right!).

- **Open the starter script** linked in the assignment description.
  - I don't recommend coding directly into the console (command line). Coding in your script is much easier for editing your code, saving your code, and making comments for what each code does (video 3!)
- **Install and library tidyverse**
  - Write and run the following code: `install.packages("tidyverse")`
  - This will take a minute or two! Wait until the little stop sign disappears to proceed.
  - Next, you will want to run the following code: `library(tidyverse)`
- **Open the Data**
  - We will be using the `diamonds` dataset stored in the tidyverse package.
  - After librarying tidyverse, run the code: `View(diamonds)`. Each row represents one diamond from a collection of over 59,000.

Question 1 (5pts): Create a histogram of the `price` variable. 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

Would you describe this distribution as symmetric or 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 data, 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 of your sample data

What is the *absolute* error of *your* sample mean as an estimate of the true mean? *(for example: if your estimate was 85 and the true value was 100, that would be an absolute error of 15).*

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 three 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

Would you describe this distribution as symmetric or 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

Try running your loop again and calculating your standard deviation of sample means again. You'll likely find that the number changed a little bit! What is the standard deviation of the simulated means approximating? **Report the name of this measure and calculate the true value for this measure too** *(hint: check the "Testing a Mean" Chapter: pages 2-3 and page 9 of the chapter).*

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. Give this vector a new name as well!

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?**