CLT Funsheet

STA 209 Spring 25

Introduction

This worksheet is intended to help illustrate some of the concepts associated with sampling distributions, confidence intervals, and the central limit theorem. The questions posed here will indicate using functions in R to solve the problems. The necessary R tools will be provided on the course website under the Lab 7 link. In particular it will show:

- The functions to use
- How to use them

You do not need to include any R code or plots for this worksheet, but you may find it helpful to record them somewhere for later reference.

Note that I will use the term "xbar" to refer to a column in a dataset titled `xbar`. This is meant to represent the sample mean, denoted symbolically as $\overline{x}$.

Sampling Distributions and CLT

Sampling from Normal Distribution For these problems, we will be using the `sampleNormalData()` function from the lab.

1. Using the `sampleNormalData()` function, run a simulation that has $n = 15$ observations in each sample, with a population mean of 100 and standard deviation of 15.

   (a) Create a histogram of the simulated xbar values. What distribution does xbar seem to follow? How can you tell?

   (b) In what range do the majority of values tend to fall? How big is this range? (Note: I'm not looking for exact values. Give me a ballpark estimate of where approx 90% of the data seems to fall. Round numbers are easiest)

   (c) Repeat this simulation, this time setting the population mean to be 200 while keeping everything else the same. In what range do the majority of observations tend to fall? How big is this range? How does this compare to what you found in (b)?

   (d) Based on this, how does changing the mean of the population seem to impact the distribution of $\overline{x}$?

2. Using the `sampleNormalData()` function, run a simulation that has $n = 50$ observations in each sample, with a population mean of 100 and standard deviation of 15.

   (a) Again create a histogram of xbar and indicate in what interval do the majority of statistics tend to fall. How big is this range?

   (b) Repeat this process, this time setting $n = 5$ and create a histogram. Where do most of the statistic fall? How big is this range?

   (c) Compare what you found in (a) and (b) with what you found in Question 1. What seems to be driving the difference between these ranges?

3. Using the `sampleNormalData()` function, run a simulation that has $n = 15$ observations in each sample, with a population mean of 100 and *standard deviation of 5*.

   (a) Again create a histogram and note the size of the interval. How does this compare to what was found in Question 1 (a)?

   (b) Repeat this simulation, this time setting the standard deviation to be 30. How does the size of the interval change?

   (c) Based on what you have seen, how does changing the standard deviation of the population impact the shape of the sampling distribution? How does this compare to what you found in (1c)?

   (d) If I know my population has a large standard deviation, what should I do to get a more precise estimate of $\overline{x}$?

4. Suppose that we have a population with $\mu = 100$ and $\sigma = 15$. According to the CLT, if I have a sample of size $n = 20$, what should the distribution of $\overline{x}$ be? Write your answer in the form $\overline{x} \sim N(\cdot, \cdot)$ where you should replace $\cdot$ with numerical values

5. Use `sampleNormalData()` to run a simulation to match the conditions of Question 4. Then, using `summarize()` from dplyr, find the mean and standard deviation (`sd()` in R) of the column `xbar`. In other words, instead of relying on the CLT, use the simulated sampling distribution to derive its mean and variance. How do these values compare with what you found in Question 4. Is this what you would expect? Explain.

Sampling from `college` dataset

6. First, using the college dataset, create a histogram of the variable `Enrollment`. Is this distribution normal? If not, how would you describe it?

7. Using the function `getSampleMean()`, collect 1000 samples with $n = 5$ observations. Create a histogram of xbar. What do you see? Is this normally distributed?

8. Repeat Question 7 several times, using the values $n = 15, 25, 100$ for your sample size. How does the distribution of `xbar` change as n increases?

9. Using `summarize()` from dplyr, find the mean and standard deviation of the variable `Enrollment`. Based on this, with a sample size of $n = 100$, what would you expect the sampling distribution of $\overline{x}$ to be?

10. Using your simulation from Question 8 with $n = 100$, use summarize to find the mean and standard deviation of the variable `xbar`. Is this what you would expect, based on what you found in Question 9?

11. Based on your solutions here with the `college` dataset, answer the following:
    (a) Does a population have to be normally distributed for the central limit theorem to apply?
    (b) If a variable is highly skewed, what do we need for our normal approximation to hold?
    (c) What if our population is normally distributed, as it was in Question 1-5?
    (d) Based on these, what appears to be the relationship between sample size, the distribution of my population, and the usefulness of the normal approximation provided by the CLT?

Confidence Intervals

This last section is going to explore the relationship between sample size, the standard deviation of the population, and the amount of "confidence" we put into our plots, also known as the *critical value*. We will go into more detail later in the semester, but for now, we think of the amount of confidence we are putting into our lab as a multiplier, C, that determines how many standard standard errors away from our sample mean we wish to construct our interval:

$$\overline{x} \pm C \times SE$$

Remember: just like the sample mean, we find our estimate of the standard error using our sample.

12. Use the `simulateConfInt()` function to generate a sample with `n = 15`, `C = 1`, and `sd = 15`.

    (a) For the first simulation you ran, how many intervals do not contain the population mean, indicated by the black horizontal line?

    (b) Run this function several more times with the same arguments. Is the number of confidence intervals that fail to contain the mean the same? Why do you think this is?

    (c) For the last interation you ran, look closely at the length of the error bars for each simulation. Are these the same length? Do you think it is possible for two samples to have the exact same sample mean, with the confidence interval from one sample containing μ while the other does not? What would cause these to be different?

13. Using the `simulateConfInt()` function, set `C = 1.5` and `sd = 5`. Then, run the function a few times each with the arguments `n = 5`, `n = 15`, `n = 50`, and `n = 100`.

    (a) What is happening to the length of error bars as n increases?

    (b) On average, does the number of confidence intervals containing μ seem to change as n increases?

    (c) Based on this, assuming that C and σ are fixed, what seems to change about our CI when n changes? What doesn't change?

14. Using the `simulateConfInt()` function, set `n = 25` and `C = 1.5`. Then, run the function a few times each with the arguments `sd = 10`, `sd = 5`, `sd = 2`, and `sd = 1`.

    (a) What is happening to the length of error bars as the population standard deviation decreases?

    (b) On average, does the number of confidence intervals containing μ seem to change much as `sd` decreases?

    (c) Based on this, if everything else is fixed, what seems to change about our CI when `sd` changes? What doesn't change?

15. Write out the distribution of $\overline{x}$ from the Central Limit Theorem using μ, σ, and n (that is, without using actual numbers). In light of this, comment on what you found in Question 13 and 14. Specifically comment on:

- How does changing n impact the size of our confidence intervals?
- How does changing σ impact the size of our confidence intervals?
- Why does changing n and σ *not* impact the proportion of intervals that cover μ?

16. Using the `simulateConfInt()` function, set `n = 15` and `sd = 5`. Then, run the function a few times each with the arguments `C = .5`, `C = 1`, `C = 1.5`, and `C = 2.5`.

    (a) What is happening to the length of error bars as the standard error multiplier?

    (b) On average, does the number of confidence intervals containing μ seem to change as `C` increases? Is this different than what we saw in Question 13 and 14?

    (c) Using everything you have seen in this lab, explain what impact the values `n`, `sd`, and `C` have on (i) the size of our confidence interavls and (ii) the proportion of times we can expect the interval to contain μ. Does having larger error bars necessarily mean better coverage? Explain your answer

17. In one paragraph (4-5 sentences), reflect on the main points of this worksheet. What are the most important concepts you took away from it? What is something you feel like you still need more practice with?