CLT Funsheet

STA 209 Spring 26

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 5 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()` simulation function from the CLT lab. The default values for this function are `mu = 100`, `n = 15`, and `sig = 15` unless otherwise stated.

1. Using the `sampleNormalData()` function, I will specify a number of changes to the default arguments I want you to make. For each change, I want you to specify:

(a) How did changing these values impact the distribution of $\overline{X}$?

(b) Identify the range in which the majority of the values tend to fall (Note: I'm not looking for exact values, give me a ballpark estimate of where the middle 90% fall. Round numbers are easiest). How does changing the values change the size of this interval?

Answer these for each of the collection of settings:

(1) Run a simulation with `mu = 100` and `mu = 200`

(2) Run simulations with `n = 5, 15, 50`

(3) Run simulations with `sig = 5, 15, 30`

2. Considering what you found in Problem 1, which parameters impact the location of the distribution of $\overline{X}$? Which parameters dictate its spread?

3. Suppose I have a population with a large standard deviation, σ. What can I do to help 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 $\overline{x}$ approximately 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, with `mu = 100`, `n = 20`, and `sig = 15`. 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 standard error 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 of the `Enrollment` variable 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? Answer in 2-3 sentences

Confidence Intervals

This last section is going to explore the relationship between sample size, standard deviation and the amount of "confidence" we have in our constructed intervals. Recall that confidence is mediated only through our choice of C, also called the *critical value*, which tells us how many standard errors away from the 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 iteration you ran, look closely at the length of the error bars for each simulation. Are the lengths the same? If two samples had identical sample means, sample size, and confidence, is it possible that the length of their error bars could be different? Why or why not?

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. Recall from the CLT we have that $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$. In light of this, comment on what you found in Questions 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?