fun CLT ~~Work~~sheet

STA 209 Fall 24

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 indicating using particular 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?
   (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 $\overline{x}$?

a) The xbar seems to follow normal distribution. I know this because it's bell shaped and symmetric about the mean.

b) Most fall between 95 and 105 (size 10)

c) Most fall between 195 and 205 (size 10). The location is different in terms of values, but the size of the range is the same.

d) Changing the mean changes the location of confidence intervals, not the width.

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?

a) Most fall between 96 and 104 (approx.). Size is 8.

b) Most fall between 90 and 110. Size is 20.

c) From the examples, I concluded that increasing the number of observations can impact the size of the range. The more observations, the narrower the range.

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) What is different from this simulation compared to what we did in Question (1a)?

   (b) Again create a histogram and note the size of the interval. How does this compare to (1a)?

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

   (d) 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)?

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

a) We're changing the SD and leaving the mean alone.

b) The size of the interval is about 4. This is much smaller than the size in 1a (which was 10).

c) The size of the interval is about 30, which is much larger than either of the previous

d) Changing the standard deviation changes the size of the confidence intervals (but the Normal bell shape stays the same).

e) Based on my conclusion for question 2, I would have to increase my n (number of observations) to get a more precise estimate of the mean.

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 $\overline{x} \sim N(\cdot, \cdot)$

$\overline{x} \sim N(100, 15/\sqrt{20})$

$\mu \quad \sigma \quad \sqrt{n}$

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`. How do these values compare with what you found in Question 4. Is this what you would expect? Explain.

The SD appears to be 3.39, which comes very close to $15/\sqrt{20}$ from the previous question. The mean is 100.1, matching the previous result of 100. So yes, this is what I expected to find.

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?

It does not seem to be normal - it looks skewed right or exponential according to the lecture.

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?

No, it isn't. It again seems skewed right and exponential (excluding a small amount of data on the left).

8. Repeat Question 7 several times, each time using $n = 15, 25, 100$. How does the distribution of `xbar` change as n increases?

The distribution becomes more and more normal as we increase the number of observations.

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?

From R: SD is 8192, mean is 6241

$\bar{x} \sim N(6241, 8192/\sqrt{100})$

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?

From R: SD is 766.96, mean is 6232. I did expect this - mean stayed the same but a higher number of observations lowered the standard deviation.

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?

a) No, as we saw in Q6, we saw the mean was normal despite the distribution not being normal.

b) We need to have a larger n number of observations. As we increase that, the mean becomes more and more normal.

c) The mean will always be normal despite a potentially small number of observations.

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. 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, m, that determines how many standard standard errors away from our sample mean we wish to construct our interval:

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

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`, `m = 1`, and `sd = 15`.
    (a) 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 fails 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? Explain what is happening.

a) 8 intervals do not contain the population mean

b) Running it again, I got 9, 7, 9, 10, and 9 intervals. The number is not the same (I believe due to randomness).

c) No, the error bars are not the same length. I think it would be possible for them to have the same sample mean, because the different lengths may represent variability - the one containing μ may have more variability

13. Using the `simulateConfInt()` function, set `m = 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, if everything else is fixed, what seems to change about our CI when n changes? What doesn't change?

a) As n increases, the length of the error bars gets smaller.

b) Not really - in the first and last function, there seem to be less confidence intervals than in the middle two which doesn't really indicate a pattern.

c) The length of the error bars changes while the number of error bars (on average) stays the same.

14. Using the `simulateConfInt()` function, set `n = 25` and `m = 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 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?

(d) Write out the distribution of $\bar{x}$ from the Central Limit Theorem. In light of this, comment on what you found in Question 13 and 14

a) The length of the error bars decreases as SD decreases.

b) Not really - once again there doesn't seem to be a pattern in CI number no matter how many times I run it.

c) The length of the error bars changes but the CI number / number of error bars stays the same (on average) as SD decreases

d) $\bar{X} = N \sim (\mu, \delta/\sqrt{n})$

↳ as n increases, $\delta/\sqrt{n}$ decreases (CI intervals are shorter). As SD decreases, $\delta/\sqrt{n}$ decreases. (CI intervals are shorter)

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

a) The length of error bars is increasing

b) yes, the number of confidence intervals now seems to increase (as m increases), as opposed to what we saw in Q13 and Q14.

c) Increasing n and m leads to shorter confidence intervals while increasing SD leads to longer confidence intervals. The only factor that affects the proportion seems to be m. And larger error bars don't always mean better coverage because due to the variability, their mean could be way off despite containing the population μ. This makes our calculations less accurate.