

CLT Funsheet

STA 209 Fall 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 $\bar{x}$.

Note: Because we are running simulations here, you should expect your values to change a little bit every time you run them

Sampling Distributions and CLT

Sampling from Normal Distribution For these problems, we will be using the `sampleNormalData()` simulation function from the CLT lab. Specify all three arguments, `mu`, `n`, and `sig`, each time you use this function. Each simulation collects 1000 samples and returns their sample means.

1. Using the `sampleNormalData()` function, run a simulation with `mu = 100`, `n = 15`, and `sig = 15`. Create a histogram of `xbar`. Then, run three more simulations, changing only one argument at a time as indicated below:

Simulation	mu	n	sig	Middle 90% of <code>xbar</code>
Starting values	100	15	15	
Change population mean	200	15	15	
Change sample size	100	50	15	
Change population standard deviation	100	15	30	

- (a) For each simulation, identify the range in which the majority of the sample means tend to fall. Note: I'm not looking for exact values, give me a ballpark estimate of where the middle 90% fall and include them in table above. Round numbers are easiest.
- (b) How did changing each argument impact the location and spread of the distribution of $\bar{X}$? Did its shape appear to change?

2. Considering what you found in Problem 1, answer the following:
- (a) Which parameter impacts the location of the distribution of $\bar{X}$? Which parameters dictate its spread?
 - (b) Suppose I have a population with a large standard deviation, σ . What can I do to help get a more precise estimate of the population mean, μ ?
3. Suppose that we have a *normally distributed* population with $\mu = 100$ and $\sigma = 15$, and collect samples of size $n = 20$.
- (a) What should the distribution of $\bar{X}$ be? Write your answer in the form $\bar{X} \sim N(\cdot, \cdot)$, where the second value represents the **standard deviation**, not the variance. For a normal population, this distribution is exactly normal, even for small samples.
 - (b) Use `sampleNormalData()` to run a simulation 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`. The standard deviation of these simulated sample means estimates the standard error of $\bar{X}$.
 - (c) How do these values compare with what you found in (a)? Would you expect them to match exactly? Explain.
 - (d) When the population is normally distributed, the distribution for $\bar{X}$ is exactly normal, rather than approximate, even for small sample sizes. Why? What should the distribution look like when $n = 1$?

Sampling from college dataset For these problems, we will treat the colleges in the dataset as our population. The `getSampleMean()` function samples colleges *without replacement* within each sample. It starts over with the full dataset for the next sample.

4. First, using the `college` dataset, create a histogram of the variable `Enrollment`.
 - (a) Is this distribution normal? If not, how would you describe it?
 - (b) Using the function `getSampleMean()`, collect 1000 samples of the `Enrollment` variable with $n = 5$ observations in each sample. Create a histogram of `xbar`. How does this compare to the histogram of `Enrollment`?
 - (c) Repeat (b), using $n = 25$ and $n = 100$. How do the shape and spread of the distribution of `xbar` change as n increases? Does its center appear to change? Verify by finding the mean of the sample means from each simulation.

5. Following up on Question 4, we will compare the sampling distribution you simulated with what we would expect. Using `summarize()` from `dplyr`, find the mean and standard deviation of the variable `Enrollment`.
- (a) Based on this, with a sample size of $n = 100$, what would you expect the mean and standard error of $\bar{X}$ to approximately be?
 - (b) Using your simulation from Question 4 with $n = 100$, use `summarize()` to find the mean and standard deviation of `xbar`. How do these compare with your predictions?
6. Based on your solutions here with the `college` dataset and the normal population in Questions 1-3, explain the relationship between sample size, the distribution of my population, and the usefulness of a normal approximation for $\bar{X}$. Answer in 2-3 sentences. In particular, does a population have to be normally distributed? If it is highly skewed, what do we generally need? What if it is already normally distributed?

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 our choice of C , also called the *critical value*, tells us how many standard errors away from the sample mean we wish to construct our interval:

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

For these problems, we will use the population standard error, $SE = \sigma/\sqrt{n}$. The argument `sd` sets the population standard deviation, σ .

7. Use the `simulateConfInt()` function with `n = 15`, `C = 1`, and `sd = 15` to construct 25 confidence intervals.
 - (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 a few 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?

8. Using the `simulateConfInt()` function, set $C = 1.5$. Run the function a few times for each collection of settings below. For the last simulation at each setting, record the percentage of intervals that contain μ in the table.

Simulation	n	sd	Estimated coverage (%)
Starting values	15	5	
Smaller sample	5	5	
Larger sample	100	5	
Smaller population σ	15	1	
Larger population σ	15	10	

Remember that the percentage will vary from one simulation to the next. When comparing the length of error bars, keep in mind that the plots may use different vertical scales.

- What is happening to the length of error bars as n increases? On average, does the percentage of confidence intervals containing μ seem to change?
- What is happening to the length of error bars as the population standard deviation decreases? On average, does the percentage of confidence intervals containing μ seem to change much?
- Recall that the standard error is $\sigma/\sqrt{n}$. In light of this, explain why changing n and σ changes the size of our confidence intervals, but does not change the proportion of intervals we expect to cover μ when C is fixed.

9. Using the `simulateConfInt()` function, set `n = 15` and `sd = 5`. Then, run the function a few times each with `C = .5`, `C = 1.5`, and `C = 2.5`.
- (a) What is happening to the length of error bars as the standard error multiplier, `C`, increases?
 - (b) On average, does the percentage of confidence intervals containing μ seem to change as `C` increases? Is this different than what we saw in Question 8?
 - (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 intervals 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.
10. In 2-3 sentences, reflect on the main points of this worksheet. What is the most important concept you took away from it? What is something you feel like you still need more practice with?