Problem Set 1

SOC-GA 2332 Intro to Stats (Fall 2025)

Due: Friday, Sep 26th, 11:59 pm

Contents

Instructions

1. Submit two files for each problem set. The first is a **R Markdown** (`.Rmd`) file that can be run without error from start to end. The second is a **PDF** rendered from your R Markdown file or created using LaTeX.
2. Name your files following this convention: `[Last Name]_ps1.Rmd` and `[Last Name]_ps1.pdf` (for example, `Gelles-Watnick_ps1.Rmd`). Please set your code chunks to show in your PDF.
3. Both files should be submitted to the TA via e-mail (rg4895@nyu.edu) before the time specified above. Please email the TA in advance if you need extensions with justified reasons. Please plan ahead and start early.
4. You are encouraged to discuss the problems with your classmates. But **the R Markdown and PDF files that you submit have to be created on your own**. Please do not ask for solutions from students in earlier cohorts.
5. This assignment will be graded on effort, but I have left in point values in case it's helpful for figuring out how challenging/time consuming different parts will be.
6. Comment on your code wherever possible and explain your ideas in detail. This is good practice for future coding and will also help me give you feedback.

Prerequisite

Start by loading the `tidyverse`, `kableExtra`, and `gridExtra` packages into your environment. Feel free to load additional packages if needed for your code.

```
# load packages
```

Part 1. Functions

Recall the formulas for population mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

and variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{2}$$

where N is the population size.

1.[5pts] Write a function that calculates the **population mean** according to Equation 1 without using any R functions that directly calculate the mean. For example, you cannot use `mean()` from base R, or `summarise(., mean = mean())` from `tidyverse`.

2.[5pts] Write a function that calculates the **population variance** according to Equation 2 without using any R functions that directly calculate the variance. For example, you cannot use `var()` from base R, or `summarise(., var = var())` from `tidyverse`. You cannot use `mean()` from base R, either.

3.[10pts] Import `gapminder.csv` to your R environment.

- Apply the two functions you just created to the `lifeExp` variable in `gapminder`.
- Use base R functions that directly calculate mean and variance to the same `lifeExp` variable vector.
- Report your results of the above two steps either in text or in a table. The results for the mean should be equal, but the results for variance should be different. Find out and explain why the results in variance differ.

Part 2. Population, sample, and sampling distribution

1.[3pts] Create a population that follows a normal distribution with population mean $\mu = 5$ and population variance $\sigma^2 = 1$ with $100,000$ observations.

2.[3pts] Create a histogram of the population with appropriate title and labels. Add a vertical line at the population mean.

3.Draw a random sample from the population (without replacement), with sample size $n = 50$.

4.[3pts] Plot a histogram of the sample with appropriate title and labels. Add a vertical line at your point estimate of the population mean. How does this histogram compare to the one you created in question 2?

5.[6pts] Based on your sample, report your point estimate, $\hat{\mu}$, of the population mean and the standard error of this estimate.

6.[6pts] Simulate the sampling distribution of the sample mean ($n = 50$) using 1000 draws. That is, repeat the action you took for question 3 for 1000 times and save the mean you get for each repetition to a numeric vector.

7.[3pts] Create a histogram of the sampling distribution of the sample mean you simulated in question 6 with appropriate title and labels. Add a vertical line at your point estimate of the population mean.

8.[6pts] Using the sampling distribution you obtained in question 6, report your point estimate of the population mean $\hat{\mu}$ and the standard error of this estimate.

9.[20pts] Repeat questions 3 to 8 increasing the size of your sample to $n = 1000$. Plot and report your results. Then, using the concepts that we learned in class, summarize the differences with respect to what you obtained with a sample of 50 (Hint: there are two related concepts, one related to the sample estimate of the population mean, and the other related to the uncertainty of the standard error of this estimate).

Part 3. Simulate confidence interval

1.[10pts] In your own words, explain what is a 95% confidenece interval and what is a p-value.

2.[5pts] Write down the mathematical formulas you use to calculate the 95% confidence interval of the sample mean for a given sample (assume the sample size is larger than 50 so you can use the z instead of the t-statistics). Make sure you explain the meanings of the notations you use in your formulas.

3.[15pts] Replicate the plot below and interpret the plot in one or two sentences.

Hint: The plot shows the simulation result when a sample (size n) is randomly drawn 100 times without replacement from a hypothetical population with mean μ and standard deviation σ. μ has a correct, fixed answer, but n and σ depend on your choice, as long as the simulation results look similar to the plot.

Note: As there is randomness involved in the simulation process, your plot does not need to replicate the exact same result as the example. Also, try your best to tweak the layout of your plot, but it does not need to look exactly the same as the example. You will get credits as long as you generate a similar enough graph that convey the conceptual points clearly and correctly. Remember to use `set.seed()` for any random process.