Problem Set 1

SOC-GA 2332 Intro to Stats (Spring 2021)

Due: Saturday, Feb. 27th, 11:59 pm

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`.
3. Both files should be submitted to the TA via e-mail (di.zhou@nyu.edu) before the time specified above.
4. You are given plenty of time to work on the problem set. Please plan ahead and start early. **Except for special circumstances, the TA will not accept last-minute questions asked on the day when the problem set is due**.
5. You are encouraged to discuss the problems with your classmates. Notice as well that we have students in this class who are not in your cohort. It would be great if you could reach out to them and work together. But **the R Markdown and PDF files that you submit have to be created on your own**.
6. Comment on your code wherever possible and explain your ideas in detail. You will get credits for showing the steps you take and for explaining your reasoning, even if you do not get the correct final result.

---

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. Write a function in R 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`.
   - Name your function `pop_mean`.
   - The function should take a numeric vector as its input.
   - The function should return a numeric variable that is the population mean calculated based on the vector input.

2. Write a function in R 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`.
    - Name your function `pop_var`.
    - The function should take a numeric vector as its input.
    - The function should return a numeric variable that is the population variance calculated based on the vector input.
    - You can use the `pop_mean()` function you just created for your `pop_var()` function.
3. Import `gapminder.csv` to your R environment.
    - Apply the two functions you just created to the `lifeExp` variable in `gapminder`.
    - Use 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.

Note: For this exercise, we will assume that there is no missing values (i.e. no NAs) in the vector, so you don't need to consider how to deal with NA values. *Hint*: You can format tables using `kbl()` from the `kableExtra` package.

2 Data transformation using `tidyverse`

Import `parent_inc.csv` to your R environment. The data frame looks like this:

famid	father_name	mother_name	father_income	mother_income
1	Arthur	Jess	42000	45000
2	Harry	Pam	35000	24000
3	Matt	Mary	78000	55000

Use `tidyverse` functions and the piping syntax to transform the data frame to the following structure:

famid	type	name	income
1	father	Arthur	42000
1	mother	Jess	45000
2	father	Harry	35000
2	mother	Pam	24000
3	father	Matt	78000
3	mother	Mary	55000

Make sure to document the steps you take in your code and display the tidied data frame in your PDF document.

Hints:

- You can review how to use the `pivoting` functions here.
- You can use `str_remove()` or `str_extract()` functions for mutating a new variable that extracts part of the text from a string, for example extracting "father" from "father_name".
- You can separate the original data frame into parts and then combine them if you cannot figure out how to transform it altogether.
- You can format tables using `kbl()` from the `kableExtra` package.

3 Population, sample, and sampling distribution

To make your code reproducible, use the `set.seed()` function whenever you are generating random numbers or sampling randomly. Read the documentation of this function in R if you do not know how it works.

1. Create a population data frame that has one variable called "value", whose value follows a normal distribution with population mean $\mu = 5$ and population variance $\sigma^2 = 1$ with 100,000 observations.
2. 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, with sample size n = 50.
4. 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. Based on your sample, report your point estimate of the population mean $\hat{\mu}$, the standard error of this estimate, and its 95% confidence interval. Show the formulas you used for calculating these statistics.
6. Simulate the sampling distribution of the sample mean (n = 50) using 1,000 draws. That is, repeat the action you took for question 3 for 1,000 times and save the mean you get for each repetition to a data object. *Hint:* Use `for` loop.
7. 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. Using the sampling distribution you obtained in question 6, report your point estimate of the population mean $\hat{\mu}$, the standard error of this estimate, and the 95% confidence interval of this estimate. Show the definitions or formulas you used for calculating these statistics. *Hint:* The standard error in this question should be worked out based on the properties of the sampling distribution.
9. Repeat questions 3 to 8 increasing the size of your sample to n = 1,000. 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:* Which law or theorem that we learned in class is being demonstrated here?