Introduction to Statistics (SOC-GA 2332)

Assignment I

Due Date: Sep. 25, 11:59pm

Instructions

Submit a pdf-file that answers the following problems together with the `R` script that is used in your analysis. Make sure that the script runs without error from start to end before submitting it.

The files you submit should be named as follows: `[Last Name]_Assignment1.pdf` and `[Last Name]_Assignment1.R`. If you are Bob Smith, your document should be named `Smith_Assignment1.pdf` and your `R` script `Smith_Assignment1.R`. The pdf-file should be created with LaTeX or Rmarkdown.

You are encouraged to work with others on the problem sets. However, the documents and `R` scripts that you submit have to be created on your own. Both the pdf-file and the `R` script must be submitted to me via e-mail **before September 25, 11:59pm**.

Important Notes:

1. These problems are designed to help you figure out what to do when you are confronted with a statistical problem that is not obvious at first sight, rather than to test your knowledge. You've learned all the relevant concepts to answer these questions, either in the lecture or in the math camp. The difficult part is putting these things together.
2. Some of these problems are easy, while others might be rather challenging. In either case, you probably won't be able to solve all of the problems in one day. So, please **plan ahead**. Read through the problems and take your time thinking about them.
3. To reiterate, you are encouraged to **discuss the problems with your classmates**. Some of you will be better in dealing with probability, others will be more advanced in coding; so I hope that you will learn from each other. 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.
4. When answering the questions, try to be as detailed as possible. Otherwise, it is difficult to give you feedback.
5. If there are one or more problems for which you remain completely clueless by **Sep. 23rd**, send me an e-mail (`bp1094@nyu.edu`). In this e-mail, **please tell me what you have tried so far**, so I can understand exactly why you are struggling and help you efficiently. If you forget to include your prior trials, my first response will be most likely "can you tell me what you have tried so far?"

6. On September 26, I will share a solution manual that explains the answers to the problems. In the lab on the 27th, we will have some time to discuss the problems. So, **please read the solution manual before you come to class**.

Less Important:

1. In all numerical answers, round the numbers to three decimal digits. That is, if you get 2.2014542, you should report 2.201.
2. When you include figures in your document, include them within the text rather than pushing them to the end of the document.
3. When creating your `R` script, please keep in mind the **Style Suggestions** that I've provided throughout the Math Refresher and the labs.
4. Make sure that your `R` script is well-documented with comments. In addition, make clear which parts of your code are used for what problems in the problem set. I have numbered all problems, so this should not be difficult.

Problems

Problem 1

Load the `simdat.csv` data into your `R` environment using the `fread` function from the `data.table` package. Call the data `pop`. Recall that you'll have to load the packages first before you can use their functionalities.

1. Report the mean of `x2` using the `mean` function and by hand-coding the formula

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

   where N is the population size. Here, by hand-coding, I mean that you can make use any other function except those which directly calculate the mean of `x2` for you.

2. Report the variance of `x2` using the `var` function and by hand-coding the formula

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

3. You'll see that the results from the `R` function and the hand-coded one will be equal, but that the results for the variance will be different. If they differ in their fourth decimal digit, you did everything right. Find out and explain why the results differ.

Problem 2

Using the `pop` dataset, do the following:

1. Plot a histogram of `x1`. Do the following in order, by adding new features to the plot with the `+` operator, and examine how the plot changes at every step. Do not include these plots in your final document.

    (a) Label the y-axis as "Frequency"

    (b) Change the border colors of the bars to "white"

    (c) Limit the range of the x-axis to $[-1, 3]$.

    (d) Apply the `theme_bw` theme.

    (e) Find out how to hide the "ticks," the "title," as well as the numbers/labels on the x-axis; hide all of them

    (f) Assign the final plot to an object named `plot1`.

2. Plot a box plot of `x1`. Do the following in order, by adding new features to the plot with the `+` operator, and examine how the plot changes at every step. Do not include these plots in your final document.

    (a) Label the y-axis of this plot as "Variable X1".

    (b) Limit the range of the y-axis to $[-1, 3]$.

    (c) Apply the `theme_bw` theme.

    (d) Hide the "title," "ticks," and labels/numbers from the y-axis of the plot.

    (e) Flip the coordinates so that the x-axis of the resulting plot is the same as the x-axis of `plot1`

    (f) Assign the final plot to an object named `plot2`

3. Using the `cowplot::plot_grid` function, combine `plot1` and `plot2`.

    (a) Place `plot1` on top of `plot2`.

    (b) Make sure that the x-axis of `plot1` aligns with the (flipped) x-axis of `plot2`

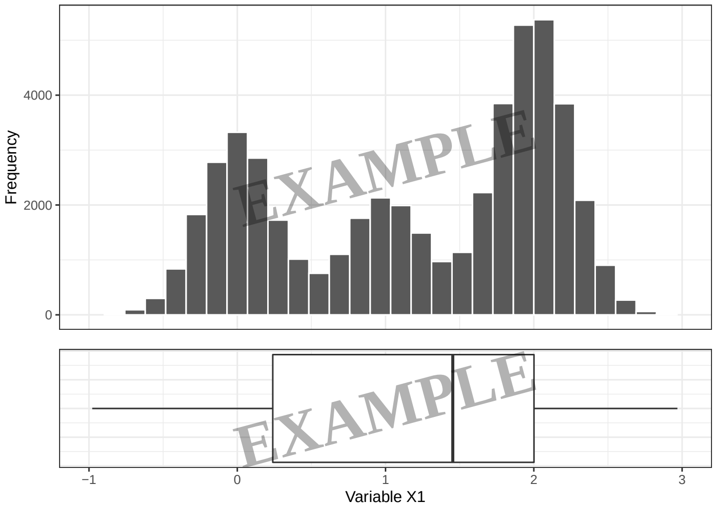
    (c) Search how to adjust the relative height of the two plots. Then, make sure that the height of the histogram is two times that of the box plot. Here the "height of the plot" *does not mean* the the plotting area, but the total size of the plot (including labels, axes, and so on). The final plot that you must produce should look like the one in Figure 1.

4. Add the combined plot to your document that you return. Compare the two plots. Do they seem to represent the same distribution? Which one would you prefer, when examining the distribution of `x1`? Explain why you'd do so. Also, why are these two plots giving such a different impression regarding the distribution of `x1`? Explain.

Problem 3

Let N be the number of rows in the `pop` dataset. By "sampling" from `pop`, I mean sampling from the rows of the dataset (which we will treat as "individuals"). In the

Figure 1: Example Figure



Note: please make sure that the vertical grid lines in your plot are aligned as in this example plot.

following questions, **do not provide any numbers**. State the answer as an equation, and explain, in detail, how you came to that conclusion.

(A small hint: Among all problems in the whole assignment, Problem 3.2 might be the most difficult one. So, it might be a good strategy to solve other problems first and return to this question later if you are not sure how to solve it. It is always a good idea to check your answer with some concrete numbers. For example, you might check whether your equation gives the right answer if the population size is 5 and the sample size is 3. By the way, if you came to the conclusion that $N^n/n!$ is the answer, I can assure you that you are wrong. If you put $N = 5$ and $n = 3$ and calculate $N^n/n!$ you'll get something like 20.833; but there cannot be 20.833 ways to draw a sample!)

1. How many ways are there to create a sample of size n from this datatset when the order in which we sample is unimportant and when we are sampling *without replacement*? The order being unimportant means that the sampling first the 1st row of `pop` and thereafter the 12th row is treated as the same sample as one in which we first sample the 12th row and thereafter the 1st row, given that the rest of the sample contains the same individuals. Hence, only the sampled individuals matter, not the order in which we choose them.

2. How many ways are there to create a sample of size n from this dataset when the order of the sample is unimportant and when we are sampling *with replacement*?

3. Solve the “taxi driver a the grid” exercise that is provided at the end of this webpage.

Problem 4

Consider the variable `x3` in the `pop` dataset. Let us call this variable “motivation to learn statistics,” where the values correspond to “1” not interested at all, “2” curious about stats, “3” loving stats above all else.

1. Use the `table` function to examine the distribution of motivation in the population.

2. Suppose I sample $n = 1$ individual from the dataset, what is the probability that this individual will be "not interested at all" in statistics?

3. If I would sample $n = 5$ individuals from this population *independently* and *with replacement*, what is the probability that at least one of them will be "curious about stats" or "love statistics above all else" (i.e., not "not interested")?

Problem 5

This will be a small simulation.

1. Set the random seed to a integer that you like

2. Use a total of `n_sim = 10,000` simulation runs.

3. Create an empty numeric vector of length `n_sim`

4. Begin the simulation:

   (a) For the `ith` iteration of the simulation, sample 5 rows from the `pop$x3` *with replacement.*

   (b) Count the number of sampled individuals who are "curious about" or "love" stats

   (c) If this number is (strictly) greater than zero, store a `0` in the `ith` element of the empty vector that you have prepared, otherwise store a `1`.

5. Calculate the mean of the vector in which the results are stored.

6. What number do you get? How is this number related to the probability we were seeking in Problem 4.3? Explain.