JSC270 Winter 2023 Assignment 3: Simulation

Submission Deadline: Saturday, March 11th at 11:59PM EST

What to submit:

1. A **PDF report** containing your written responses and any figures that are asked for in the assignment.
2. A Google Colab **notebook** containing relevant python code. Provide a link to this notebook at the top of your PDF report. Make sure you change the share settings (top right of the colab you'll see a "Share" icon) to "Anyone with a Link" and "Commenter."
    - Alternatively, you may place the code on GitHub, and instead provide a link to that repo following the instructions provided in Part I of Assignment 2.

Use section headers or comments in your notebook to indicate which pieces of code correspond to which question.

Where to submit: The PDF report is to be submitted through Quercus by the due date.

Grading Scheme: This assignment contains a total of 45 marks. *This assignment is to be completed individually.*

Part 1: Approximating pi

Pi day is coming up so let's have an early celebration with this question! 🎉🥧

A. (4 pts) Suppose you can only generate pairs of uniform random numbers between 0 and 1 (ie. points within a unit square centered at (½, ½)).

Describe a method to approximate π by generating many pairs of these uniform random numbers. Implement your method in your notebook to obtain an estimate of π.

Hint: Recalling the formula for the area of a circle might come in handy here.

B. (4 pts) How many pairs of uniform numbers did you generate in your implementation from (A) and why? How close is your estimate to π?

C. (2 pts) Suppose you generated a 1 million estimates of π using your proposed method from (A) and plotted a histogram of the 1 million estimates (you are not required to implement this). Would you expect the distribution of the estimates to be symmetric? Why or why not?

Bonus (1 pt): Describe how you could determine the number of pairs of uniform random numbers required to estimate π with a specified amount of error.

Part 2: Understanding bias

Let $\hat{\theta}$ be an estimator for some unknown parameter θ. The bias of an estimator is defined as the expected difference between $\hat{\theta}$ and θ, ie.

$$\text{bias}(\hat{\theta}) = E(\hat{\theta} - \theta)$$

An estimator is said to be unbiased for θ if $\text{bias}(\hat{\theta}) = 0$. Early in the semester, I mentioned that the sample variance

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is unbiased for the population variance σ^2 when the data is independent and identically distributed. Another reasonable estimator for σ^2 is

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Let's use simulation to compare these two estimators and gain a better understanding of bias.

A. (3 pts) Run a simulation to compare the bias of the above two estimators for σ^2 in your notebook. Specifically, consider sample sizes of n = 10, 25, 50, 100, 250, 500. For each sample size, generate the n samples independently from a normal distribution with mean 2 and standard deviation 2 and compute both estimators. Repeat this process 1000 times for each sample size. With the 1000 estimates, compute the bias of both estimators for σ^2 for each sample size.

B. (2 pts) Make a plot of bias vs sample size for the two estimators. What do you observe? Is this behavior expected?

C. (2 pts) Which of the two estimators for σ^2 do you prefer? Why?

D. (3 pts) Write out the steps you would need to take to evaluate the bias of the slope parameter in a simple linear regression model with simulation. You do not need to implement the steps in your notebook, just write them out clearly so that someone could easily implement them if they wanted to.

E. (2 pts) For part (D), what parameters do you need to specify to run your simulation? How would you go about specifying them?

Part 3: Simulation IRL

A. (4 pts) Read through this tutorial. Can you identify any weaknesses in the author's suggestion for how to generate data from a time series?

B. (4 pts) Read the following article. Describe some advantages and disadvantages of Meta's simulator to detect harmful behavior.

C. (4 pts) Watch this video on IBM's Project Photoresist. Describe the problem the team at IBM wanted to solve and the role of simulation in this project. Can you think of any weakness in the team's approach? What do you like about their approach?

Part 4: Asymptotic behavior

A. (3 pts) Let's run a simulation to investigate the behavior of the sample mean. Consider sample sizes of n = 10, 25, 50, 100, 250, 500, 2000, 5000. For each sample size, generate the n samples independently from an exponential distribution with a mean (or expected value) of 2. Compute the empirical mean in each sample.

   Plot the value of the empirical mean vs the sample size with a horizontal line at 2. What pattern do you observe? Is the pattern what you would expect? Why or why not? Would you expect the same pattern if you simulated from an exponential distribution with a different mean?

B. (3 pts) Again consider sample sizes of n = 10, 25, 50, 100, 250, 500, 2000, 5000. For each sample size, generate the n samples independently from a standard Cauchy distribution (location and scale parameters equal to 0 and 1, respectively) and compute the empirical mean. Repeat this process 1000 times for each sample size.

   Make a histogram of the 1000 empirical means for each sample size. What pattern do you observe? Is the pattern what you would expect? Why or why not?

Part 5: Logistic Regression

Consider the simple logistic regression model covered in class, ie.

$$P(Y = 1 \mid X) = g(\beta_0 + \beta_1 X)$$

where g() is the expit function.

a.) (2 pts) Provide an interpretation of the slope parameter, β_1. Justify your interpretation. You can use a similar strategy to what was discussed in lecture for the interpretation of linear regression coefficients.

b.) (1 pts) Provide an interpretation of e^{β_1}.

c.) (2 pts) Suppose you are presenting the results of a simple logistic regression model to a collaborator who is not extremely familiar with data science (eg. a clinician, product manager, etc). Would you present the estimate of β_1 or e^{β_1} to explain the results of your model? Explain your reasoning.