JSC270 Assignment 3

Yuwei (Johnny) Meng

11 March 2023

Link to Notebook:

Link to GitHub Repo:

Part 1: Approximating π

a) Suppose you can only generate pairs of uniform random numbers between 0 and 1 (i.e. points within a unit square centered at (0.5, 0.5)). 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.

Solution. Given a unit square centered at (0.5, 0.5), the largest circle that can be fitted into the square is a circle centered at (0.5, 0.5) with radius 0.5. Hence, we can simulate many pairs, say 1000000, of uniform random numbers between 0 and 1, and compute the Euclidean distance between the generated point and the origin of the circle for each of them. If the distance is greater than 0.5, then the point is outside the circle. If the distance is within 0.5, then the point is inside the circle. We count the number of generated points that are inside the circle and divide by 1000000. Hopefully, the ratio approaches $\pi/4$, which is the ratio between the area of the circle and the area of the unit square. Then we can multiply the ratio by 4 to get an approximation of π.

The approximated π is 3.143524, which is not too far from the real value of $\pi \approx 3.1415926$.

b) How many pairs of uniform numbers did you generate in your implementation from (a) and why? How close is your estimate to π?

Solution. I generated 1000000 pairs of uniform numbers. I chose this number because it's a large round number so I could get an accurate approximation for π. The difference between my approximation and the true value is smaller than 0.002, which means the approximation is very close.

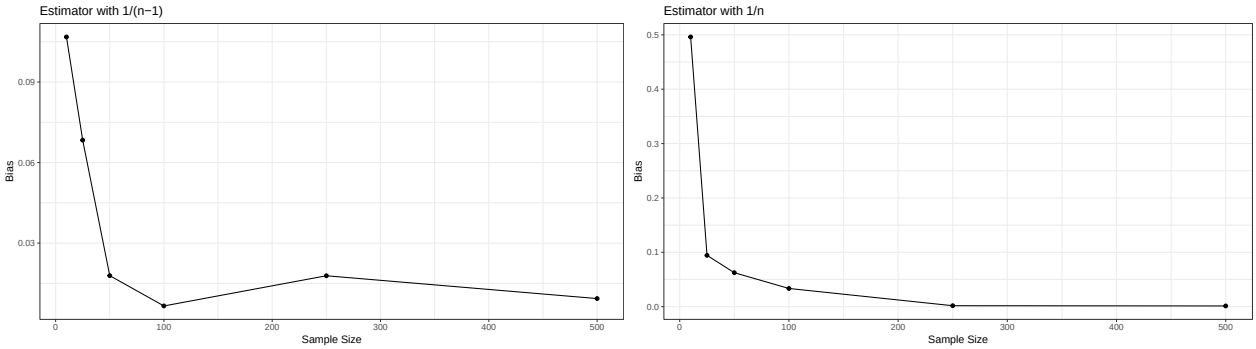
c) Suppose you generated a 1 million estimates of π using your proposed method from (a) and plotted a histogram of the 1 million estimates. Would you expect the distribution of the estimates to be symmetric? Why or why not?

Solution. Yes, I would expect the distribution of the estimates to be symmetric. In this case, the largest possible distance from the origin is $\sqrt{2}/2$, which is approximately 0.7. Hence, by the *Central Limit Theorem*, I would expect the distribution to be centered at around 0.35, because this is the average (the middle) value of the distance. The farther from the center, the fewer the observations.

Part 2: Understand Bias

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

Solution.



Since there were negative biases and those were not comparable with the positive ones, the above plots were generated after taking the absolute values of the biases. It was expected that the absolute value of the bias approaches 0 as the sample size increases, because the estimates would be closer to the mean (i.e. the true variance) with a smaller standard deviation by the *Central Limit Theorem*. The above plots confirmed this expectation. As shown in the plots, the absolute value of the bias decreased as the sample size increased. Despite an unexpected bump in the bias for variance estimator 1, the general trend was decreasing.

c) Which of the two estimators for σ^2 do you prefer? Why?

Solution. I think both estimators have their advantages, depending on the situation. It seems like for smaller samples (like 10, 25, 50, and 100), estimator 1 is doing a greater job than estimator 2 because the bias for estimator 1 is smaller than the bias for estimator 2. However, estimator 2 has smaller bias than estimator 1 when the sample size is larger (like 250 and 500). In other words, the bias for estimator 2 has a higher starting value but approaches 0 more quickly than estimator 1. Thus, it is important to determine the situation before deciding which estimator to use. In reality, I prefer estimator 2 because most of the time, we would have a large sample and in which case estimator 2 has a smaller bias than estimator 1.

d) 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.

Solution.

1. Since we don't usually know the true value of β_1, we need to regard the slope estimate $\hat{\beta}_1$ as the true value, given a simple linear model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
2. Know the distributions of x and y.
3. Set $n = 10$.
4. Randomly generate n pairs of data points (x, y) according to their distributions from step 2.
5. Fit a linear model on these data points.
6. Record the slope estimate $\hat{\beta}_1$.
7. Repeat steps 4-6 for 1000 times to get 1000 estimates of $\hat{\beta}_1$.
8. Compute and record the bias by taking the differences between the mean of the 1000 slope estimates and the "true" value from step 1.
9. Repeat steps 3-8 for other sample sizes n, such as 25, 50, 100, 250, 500.

10. When we have multiple estimates of the bias, plot them against the sample sizes to see if the bias approaches 0 as sample size increases.

e) For part (d), what parameters do you need to specify to run your simulation? How would you go about specifying them?

Solution. To run the simulation in part (d), we need to know the "true" value of $\hat{\beta}_1$, which should be known if we want to simulate for this parameter. We also need to know the distributions of x and y, which is the most difficult part. To specify these parameters, we can plot the distributions of x and y respectively and then evaluate the underlying distributions. We can compute the estimates for the parameters of those distributions using the given data. For example, if we believe x follows a normal distribution, then we can use the sample mean and sample variance as estimates to the true parameters μ and σ^2. The thing to be careful about is that if we specify the distributions wrong, then the bias might not be computed correctly.

Part 3: Simulation IRL

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

Solution. The biggest weakness I would say is related to the dependency of the data. In reality, the reason we study time series data is to discover how time affects a particular variable. Hence, there must be some trend, if any, between the target variable and time. For example, as time moves forward, we would expect the height of a child increases because they grow. Nevertheless, the way that the author generates time series data is by using the `randint` function in `np.random`, which does not make sense since this function generates independent random integers. Plotting independent data against time is meaningless. Overall, generating dependent data is difficult because there are many unknown factors in real life that might affect the target variable other than time.

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

Advantages:

- Able to automate interactions between millions of bots. This can model the complex social networks for achieving an accurate representation of reality.
- Bots cannot interact with real people. This ensures that the simulation reflects actual use of Facebook while refraining from affecting real people.
- This simulator integrates a variety of topics including software engineering, machine learning, programming languages, etc., so that it can model real life in an almost accurate way.

Disadvantages:

- Some complex interactions are impossible to predict. Thus, not all real-life situations can be simulated by this model.
- Right now it's still at the research-only stage. It might take a long time before this simulator is of actual use in real life.

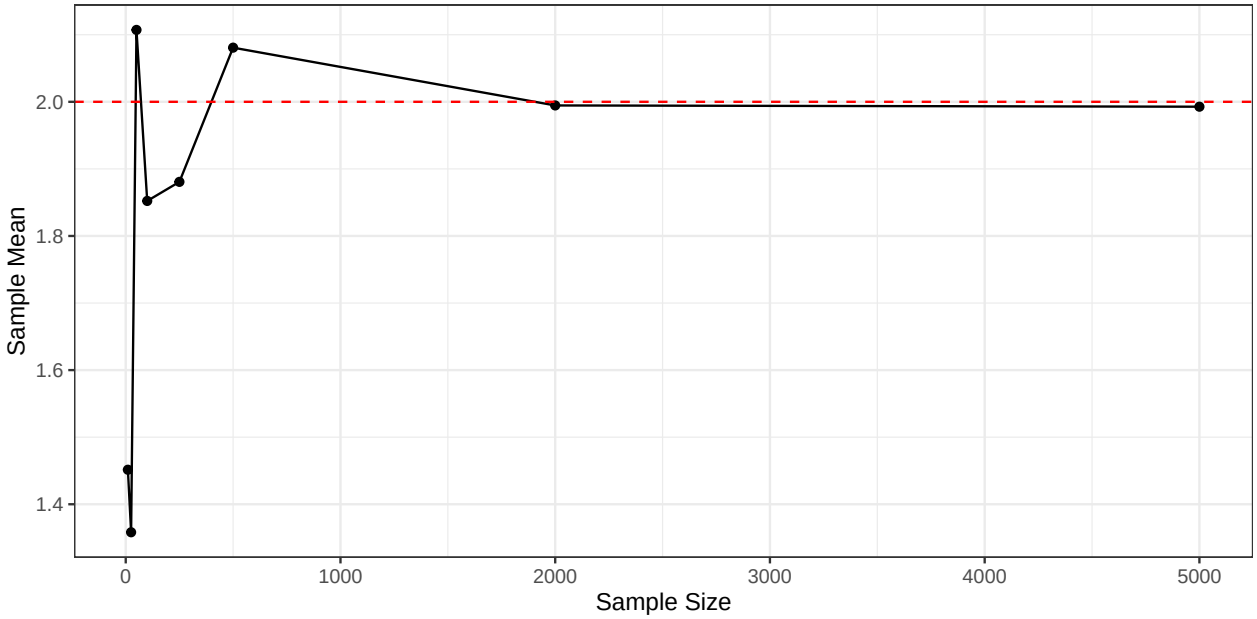
c) Watch this video on IBM's Project Photoresist. Describe the problem the team at IBM wanted to solve and the role of simulation in this project. Can you think of any weakness in the team's approach? What do you like about their approach?

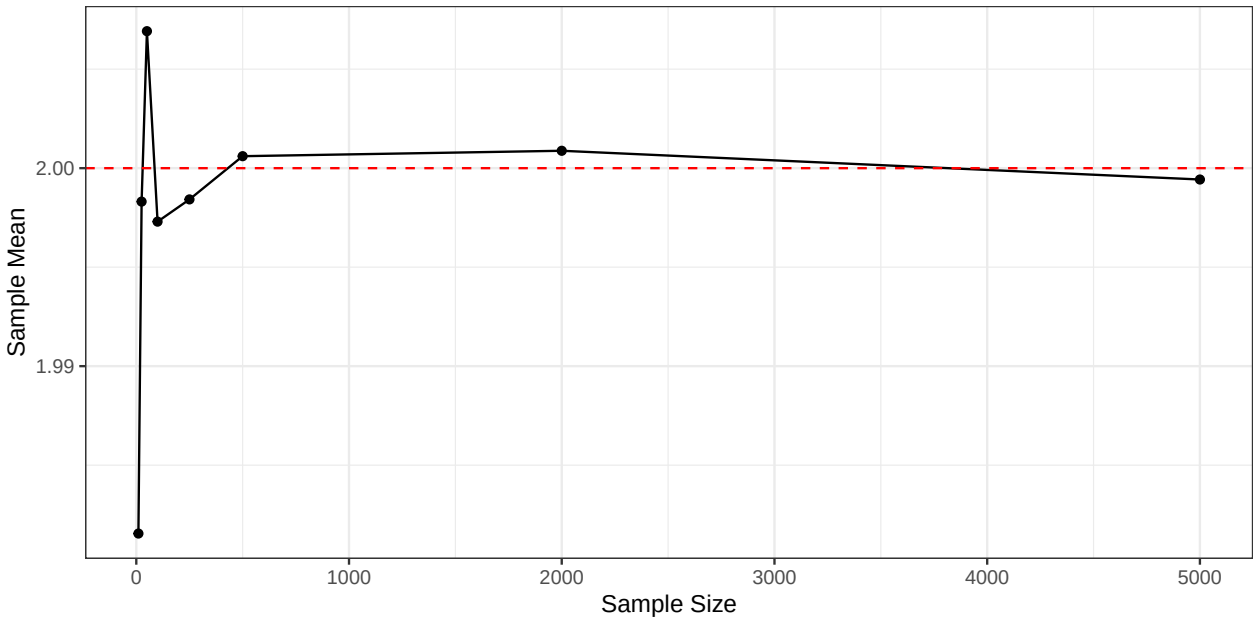
Solution. The problem that the team at IBM wanted to solve was to discover a more sustainable photoacid generator. They found that the most common language among researchers is data, which led them to use data simulation as a method of research. While reviewing the literature, they found that many properties of the materials being studied were not publicly available in the literature, so they turned to data simulation for augmenting their data set in order to discover the unknown properties of the materials, such as the peg molecules. For the weakness, I was afraid that if they didn't know the properties of the materials, then the simulation might be run incorrectly. Probably this was not a problem for them because they definitely knew the underlying mechanisms of the materials, but this was my curiosity. I like how they used simulation to generate data when the public data were not ample, and how they incorporated AI into their research to make the process more efficient.

Part 4: Asymptotic Behavior

a) 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?

Solution.

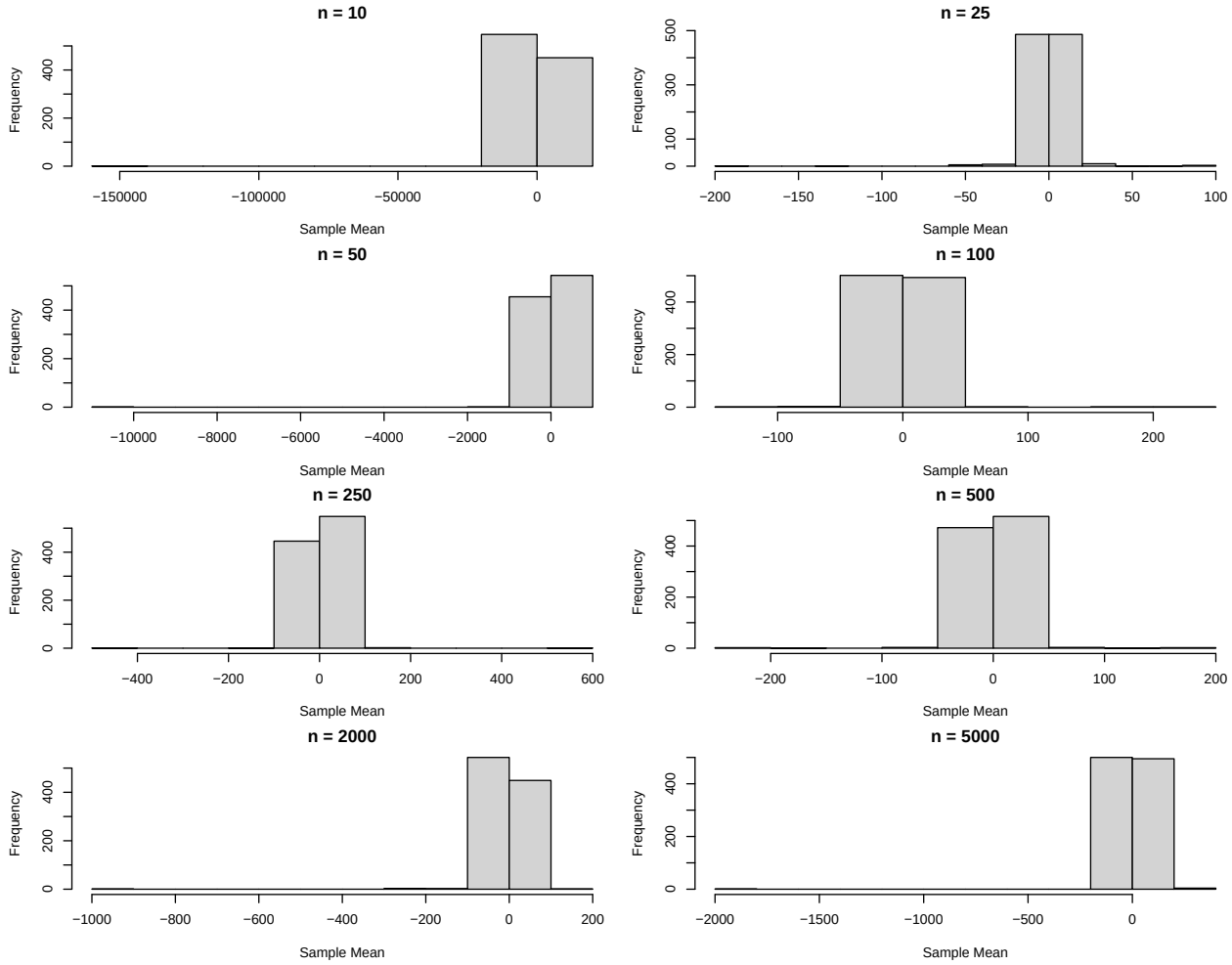




From the above plot, I observe that as the sample size increases, the sample mean approaches the true mean value of 2. This is expected because according to the *Law of Large Number*, the sample mean approaches the true mean as the sample size increases. I would expect the same pattern if I simulate with a different mean. The only thing that would change is the asymptote that the sample mean is approaching to.

b) 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?

Solution.



From the above plots, we can see that all plots have a peak at 0, but other values are unpredictable. For example, for $n = 10$, there is an observation with value at about -160000. This is expected because the expected value and the variance for the Cauchy distribution are both undefined, though the median is 0. That explains why we would get such extreme values even for small sample sizes but most data points centered at 0.

Part 5: Logistic Regression

a) 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.

Solution. On the log-odds scale, one unit increase in the predictor X is expected to result in β_1 increase in the log-odds of the probability of success. This is because the log-odds is computed as

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Treating this like a linear model, one unit increase in X is associated with β_1 units increase in the log-odds.

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

Solution. The odds of the probability of success can be computed as

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 X) = \exp(\beta_0)\exp(\beta_1 X)$$

Thus, one unit increase in X is expected to multiply the odds of the probability of success by e^{β_1}.

c) Suppose you are presenting the results of a simple logistic regression model to a collaborator who is not extremely familiar with data science (e.g. 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.

Solution. I would probably present the estimate of e^{β_1}. Even though β_1 is additive and e^{β_1} is multiplicative, e^{β_1} is a fixed number anyway. Using β_1 to present involves explaining the log-odds to the client, which I think is an even harder topic to comprehend. Therefore, I would rather use e^{β_1} and explain how it has a multiplicative impact on the odds of the probability of success.