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?

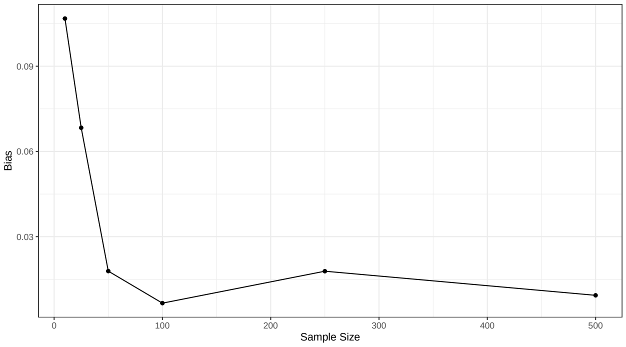
Solution. I generated 1000000 pairs of uniform numbers. I chose this number because it's a large round number so I could get an accurate approximation for π. The difference between my approximation and the true value is smaller than 0.002, which means the approximation is very close.

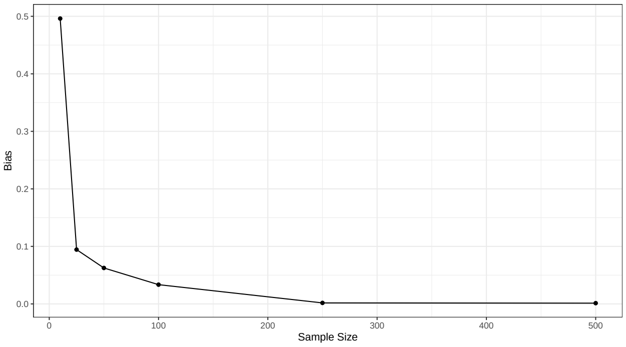
c) Suppose you generated a 1 million estimates of π using your proposed method from (a) and plotted a histogram of the 1 million estimates. Would you expect the distribution of the estimates to be symmetric? Why or why not?

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?





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?

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.

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?

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.