JSC270H1 - Assignment 3

Colab notebook:
https://colab.research.google.com/drive/1uIF4miOgh1T_GxfOlCsycCkluTc3rUjv?usp=sharing

Part 1:

A. (Version with LaTeX in the Colab notebook.)
Set r = ½. The unit square centered at (½, ½) has side length 1 = 2r, so its area is (2r)^2 = (2*½)^2 = 1. The largest circle contained in this square is the circle of radius r centered at (½, ½), and its area is πr^2 = π/4. Notice the ratio of the circle's area to the square's area is π/4.

We can randomly generate points (i.e., pairs of uniform random numbers) in the square to get an estimate of the proportion of those points that also lie inside the circle. This proportion should heuristically equal the ratio of the areas since the larger the area is, the more likely a point will be inside it.

To determine if a point lies inside the circle, we check whether it satisfies the inequality (x-0.5)^2 + (y-0.5)^2 <= 0.5^2, which is the implicit equation of the filled in circle. After calculating this proportion, since it should approximately equal the true ratio of π/4, we multiply it by 4 to get an estimate of π.

After setting a seed, the estimate of π obtained from my estimate_pi function is 3.1422.

B. I generated 100,000 pairs of uniform random numbers for part A instead of 1,000 or 10,000 pairs in order to obtain a more accurate estimate. This follows from the Law of Large Numbers, which states that the sample mean (in this case the estimate of the proportion of areas) approaches the true mean (in this case π/4) as the sample size increases.

My estimate is accurate to 2 decimal places, and its percent error is 0.09%, indicating that it is quite close to π.

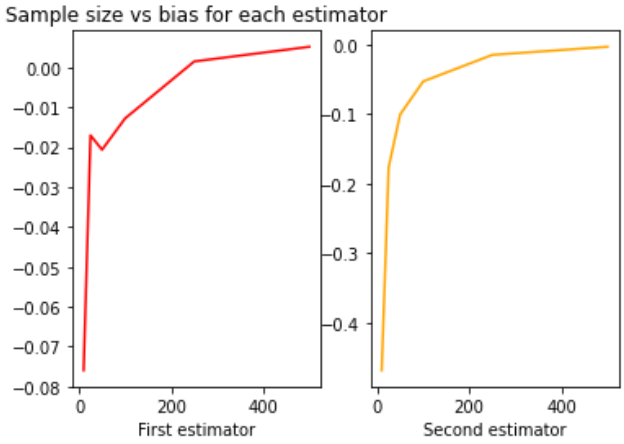
C. Yes. The estimates are random and each coordinate of an estimate is generated following the Uniform(0, 1) distribution, which is symmetric, meaning the pairs of estimates should be symmetric as well.

Part 2:

A. Code is in the Google Colab notebook. The results are shown below:

Sample size	Bias of the first estimator	Bias of the second estimator
10	-0.07594573889978351	-0.46835116500980467
25	-0.0170483427700856	-0.1763664090592818
50	-0.02065598564644544	-0.10024286593351661
100	-0.012797575628888502	-0.05266959987259945
250	0.0014821442894419334	-0.014523784287716168
500	0.005159628472188871	-0.0028506907847556384

B. Below are the generated plots:



I notice that in each plot, the bias of the estimator (indicated by the y-axis) generally tends to 0 as the sample size (indicated by the x-axis) increases. This is expected since by the Law of Large Numbers, as the sample size increases, the sample mean of the estimate of a parameter (the variance here) approaches the true value of that parameter.

However, for the first estimator, the bias passes the y = 0 mark and even increases slightly from x = 250 to x = 500. This is not expected since by the Law of Large Numbers, it is unlikely for the bias to increase as the sample size increases. Nevertheless, this phenomenon can be explained by the fact that the increase is quite small, meaning it is likely to be the result of the randomness in the calculation process.

C. I prefer the first estimator (with the 1/n-1 factor). For every sample size except for 500, its bias is much closer to zero than that of the second estimator (for instance, it has a bias of 0.00148 compared to the second estimator's bias of -0.0145 for n = 250). In

addition, for the sample size of 500, its bias is similar to that of the second estimator. This makes the first estimator ideal for estimating the population variance from a sample since it is likely to be more accurate.

D. Assume that we have a sample of the population modeled by the linear model and that we set this sample slope as the true slope parameter. Then, I would take the bootstrap approach for random sampling, which is to randomly pick a subset of data points from the model with replacement (and this subset has the same size as the original sample). For this subset, I would calculate an estimate of the slope parameter by the method of least squares (or minimizing the residual sum of squares). I would repeat this process 1000 times so that I have 1000 different estimates of the slope. Finally, I would average these estimates by calculating their empirical mean and subtract this by the true slope to get an approximation of its bias.

E. I would need to specify the estimate of the slope and y-intercept for the initial sample and for each of the 1000 subsets, in order to calculate the approximation of the slope's bias. This can be done by the method of least squares, which minimizes the residual sum of squares.

Part 3:

A. A weakness of this approach is that the data generated is too generic. It uses the same approach every time, meaning it is unlikely to represent all possible situations well. Another weakness is that this approach always generates completely random numbers for each time period in the time series. There might be situations in which the data for each time period somehow depends on the data for the previous time period (e.g., a demand and supply cycle), which creates a broader pattern in the data. Since this approach does not take into account these patterns, it would poorly simulate such situations.

B. An advantage of Meta's simulator is that the simulated users can access Facebook but cannot engage with real users, meaning Meta can train the simulator without changing or disrupting how real users use Facebook. Another advantage is the simulator simulates interactions on a large scale (e.g., up to a million bots), meaning the simulations are likely to be accurate and representative of the wider society.

   A disadvantage is that although all real users of Facebook agreed to its Terms of Service, giving Meta the legal basis for using their data, Meta does not have their explicit permission to use their data for the specific purpose of training the simulator, which may be unethical. Another disadvantage is that a breach or glitch of Meta's databases is not impossible, and if one ever occurs, it could allow the simulated users to interact with real users, creating chaos on the platform and in the wider society.

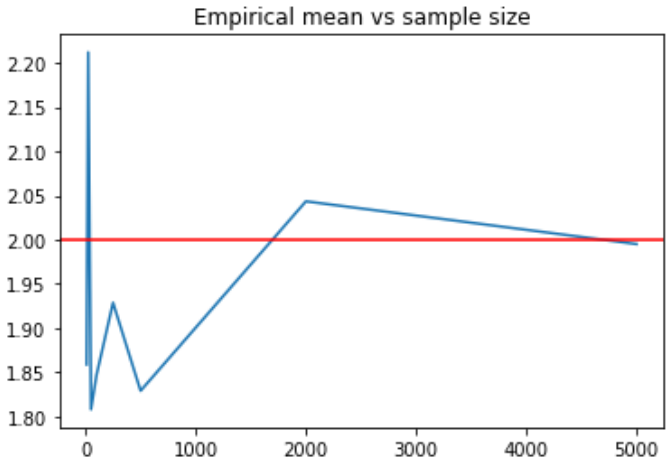
C. The IBM team aimed to develop a method of discovering new chemicals, specifically photoacid generators, for semiconductor production. AI simulation was used to augment the dataset of chemical properties, helping to discover any subsets of a chemical that might improve its suitability for semiconductors. For instance, AI helped to reduce the amount of simulations needed for quantum mechanics and physicochemical models. AI simulation was also used to create models of new chemicals with specific properties.

   A possible weakness is that since the amount of simulations needed was reduced by the AI, not all possible outcomes were generated, meaning that some truly useful chemicals might have been skipped over in the research.

   I like how the research integrated chemistry-specific knowledge into its methodology, such as the properties of PEG, since this allowed it to be more accurate. I also like how the research took into account the environmental impact of chemicals such as their biodegradation rate and toxicity, which can be an issue for sustainable production.

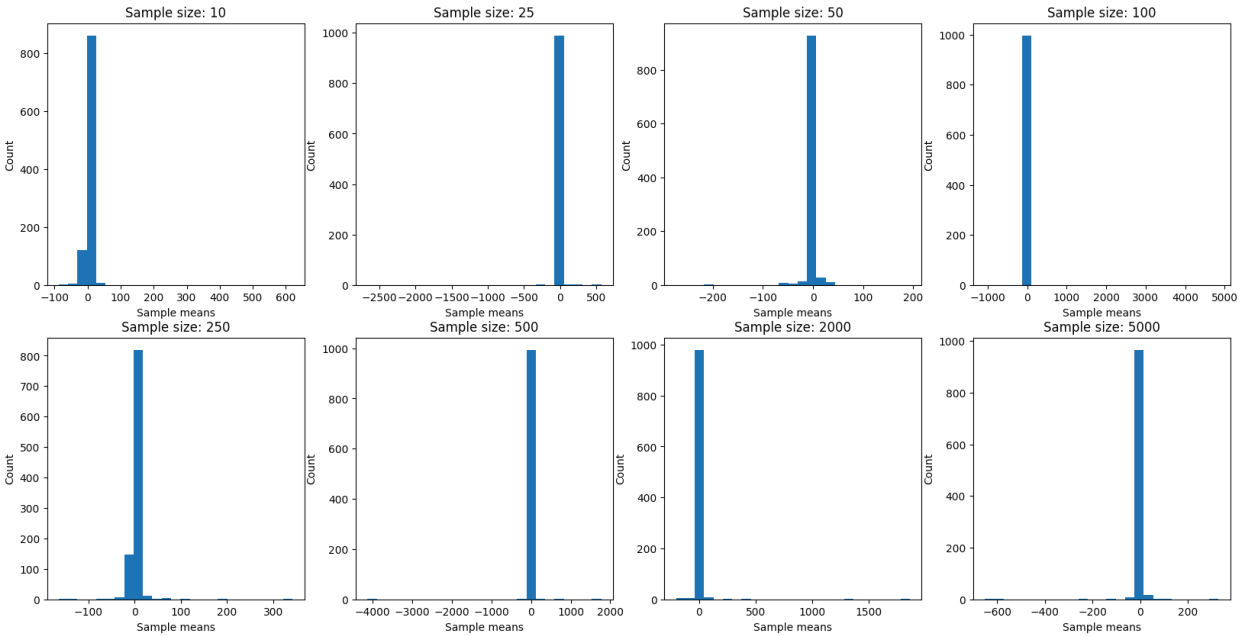
Part 4:

A. Below is my plot:



   I observe that the empirical mean fluctuates greatly for small sample sizes, and as the sample size increases, it approaches 2. This is a pattern that I would expect since by the Law of Large Numbers, the empirical mean gets closer to the actual mean as the sample size increases, which is exactly what I observe. I would expect the same pattern for a simulation with a different exponential distribution since the Law of Large Numbers holds generally, meaning it is true for different distributions as well.

B. Below is my plot:



I notice that for all sample sizes, the vast majority of the empirical means are at 0. This is expected since the mode of the Cauchy distribution is its location factor, which in this case is 0.

I also notice that the shape is slightly different for each histogram. The histograms of sample sizes 10, 100, 250, and 2000 are right-skewed, while the rest are left-skewed. Importantly, there doesn't appear to be any pattern in the change in shape as the size increases. One might not expect this since, in many situations, the distribution should look more like a Normal(0,1) PDF as the size increases due to the Central Limit Theorem. However, the CLT does not apply here since the mean and variance of the Cauchy distribution are undefined.

Part 5:

A. β_1 represents the estimated average additive change in the log-odds of the outcome, or log(p/(1-p)), for every one-unit increase in x. This is because the log-odds equation is:

$$\text{Log}\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

in which β_1 is the coefficient of x. If we treat the log-odds of the outcome as the y parameter in linear regression, the interpretation of β_1 is similar to its interpretation in linear regression.

B. e^{β_1} is the factor (or multiplicative) change in the odds of the outcome for a one-unit increase in x. In other words, as x increases by 1, the odds of the outcome is multiplied by e^{β_1}.

C. I would present e^{β_1} as the estimate for my explanation. This is because if we are interested in the change of the odds of the outcome, the extra logarithm in log(p/(1-p)) can make it confusing for an audience unfamiliar with the function. In addition, a change by a multiplicative factor is no more difficult to visualize or interpret than an additive change, which is the change we are familiar with in linear regression. In this case, raising e to β_1 can benefit my explanation and aid the audience's understanding.