Link to Github & Link to Notebook (Click to open)

Part1: Approximating pi

A. To approximate π, generate n pairs of uniform random numbers between 0 and 1, and treat them as the x, y coordinates of sample points. Notice that these sample points are uniformly distributed in the rectangle $R = [0,1] \times [0,1]$, let C be the circle with radius $\frac{1}{2}$ centered at $(\frac{1}{2}, \frac{1}{2})$, then the probability of a sample point is included in the circle is $\frac{\text{area}(C)}{\text{area}(R)} = \frac{\pi}{4}$. Let X be a Bernoulli random variable such that it equals to 1 if the sample point is included in the circle, 0 otherwise. Then we know that $E(X) = \frac{\pi}{4}$. For each sample point, we determine if it is included in the circle by calculating its distance from $(\frac{1}{2}, \frac{1}{2})$. By Law of Large Numbers, we know that the empirical average ($\overline{x} = \frac{\text{num of sample points in the circle}}{n}$) gets closer and closer to the population average ($\frac{\pi}{4}$) as the sample size increases, so we can use $4 \times \frac{\text{num of sample points in the circle}}{n}$ to approximate π.

B. I use 10000 pairs of uniform numbers generate in my implementation from (A), I use so many pairs since I have mentioned that the accuracy of my estimation is based on the sample size, a larger sample size tends to give a more accurate estimation. The difference between my estimated value and the actual value is less than 0.04.

C. I would expect the distribution of the estimates to be symmetric. Since the population mean and standard deviation of the random variable X I described in (A) exists, by Central Limit Theorem, the distribution of the sample means (which are the estimates) will be approximately normally distributed. Thus the distribution of the estimates will probably be symmetric.

Bonus. We know that the following is true for uniform distribution

$$SE[\overline{X}] = \frac{SD[X_1]}{\sqrt{n}}$$

So we have $n = (\frac{SD[X_1]}{SE[\overline{X}]})^2$ and we know that $SD[X_1] \approx (\frac{1}{12})^2$, so we can calculate the number of sample points we need based on the $SE[\overline{X}]$.

Part2: Understanding bias

A. The printed output is a list of tuples, each tuple contains the bias of both estimators for σ^2. From top to bottom, corresponding to the sample size from 10 to 500. The first element of each tuple is the bias of the first estimator and the second element is the bias of the second estimator.

B. I find that bias for both estimator have roughly the same trend, with larger oscillations when the sample size is small, and both converging to 0 when the sample size is large.

C. I prefer the first estimator. Even though both of them converge to the same value when the sample size is large, since the first one is unbiased, which means the expected value of this estimator is the same as the true variance, independent of sample size.

D. First I need to determine a liner regression model by defining its $\beta_0(=0)$ and $\beta_1(=1)$. Then generate some sample points based on this model, to be specific, randomly choose x in $[0, 100]$, then calculate $y = \beta_0+\beta_1 x+\varepsilon$, where ε is randomly chosen from $[-1, 1]$. Finally fit these sample points in a linear regression model and the slope parameter of this model is the estimated slope parameter $\hat{\beta}_1$, and $\beta_1 - \hat{\beta}_1$ is the bias. We can repeat the above process for several times with sample size equal to 10, 25, 50, 100, 250, 500, to see if the bias converge to 0.

E. I need to specify which β_0 and β_1 I am choosing which range to choose x and what is the ε term. I would choose ε far more less than $\beta_0 + \beta_1 x$, so that the sample has a small variance.

Part3: Simulation IRL

A. There might be bias when we generate random samples and we can only generate uniformly distributed random numbers in the range we specified. What if we want to let the numbers fit a normal distribution?

B.
- Advantages:
  1. The bots cannot interact with real users of Facebook, so that the user experience is not affected.
  2. The bots can operate like a real user, so it can detect some potential hazards on the platform. For example, seek to buy guns and drugs.
- Disadvantages:
  1. The complexity and magnitude of the potential search space and the way the bots interact with each other makes this simulation very challenging, and some small changes in the simulation may produce unexpected results. Some complex interactions can produce emergency properties that are difficult or even impossible to predict.

C. Problem: development of new, more sustainable photo acid generators.

The role of simulation in this project: enrich the dataset and fill in the missing information. They applied the simulations in combination with knowledge from other fields to understand the peg molecules. And iron-rich simulations were used to understand the molecules produced by their model.

Weakness: It is very difficult to generate an accurate and comprehensive model for the molecules since they are so complicated. And the properties we discovered from one kind of molecules are difficult to extend to other kinds of molecules.

I like their approach since the property information the simulation generate is very important and it is not available to public, but we can obtain them by simulation.

D.
- Advantages:
  1. Simulation enables the testing of strategic and operational decisions in the digital world where there is a low cost of failure.
  2. Simulation creates synthetic data to train machine learning models at a scale and resolution that may be impossible to obtain otherwise.
  3. It is cheaper to develop machine learning models than training equivalent models only on real data. And the machine learning models perform better.
- Disadvantages:
  1. Conducting simulations in healthcare is more difficult than other industries due to policies, resources, professional norms, and large differences between clinicians. And there is also the need to consider the protection of privacy in the healthcare industry.
  2. Simulation needs three kinds of investments, collecting data, enabling technology and identifying a near term return on investment, which is a great challenge.
  3. Simulation has been used late in the healthcare industry and the technology is relatively lagged compared to other industry.

Part4: Asymptotic behavior

A. I observe that as the sample size gets larger and larger, the average value of the sample mean is getting closer and closer to 2, which is the population mean. This is the pattern I expect since by (Strong) Law of Large Numbers, the empirical average gets closer and closer to the population average as the sample size increases. If I simulate from an exponential distribution with a different mean, I would expect that the average value of the sample mean will converge to the mean value of the new distribution.

B. I found that for different sample sizes, most of the empirical means of the 1000 processes were concentrated, but there was no pattern in where they were concentrated. At the beginning, I expected that the mode of the distribution of different sample sizes will converge to a certain value as the sample size gets larger and larger, for the same reason of 4A. But the pattern is different from my expectation and I realized that this is reasonable since the Cauchy distribution does not have a mean.

Part5: Logistic Regression

a) β_1 is the expected change in log odds of having the outcome per unit change in X. Which means increasing the predictor by 1 unit multiplies the log odds of having the outcome by e^{β_1}.

Justification:

$$
\begin{aligned}
& P(Y=1\,|\,X) = g(\beta_0+\beta_1 X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \\
\iff & 1+e^{-(\beta_0+\beta_1 X)} = \frac{1}{P(Y=1\,|\,X)} \\
\iff & e^{-(\beta_0+\beta_1 X)} = \frac{1-P(Y=1\,|\,X)}{P(Y=1\,|\,X)} = \frac{P(Y=0\,|\,X)}{P(Y=1\,|\,X)} \\
\iff & e^{(\beta_0+\beta_1 X)} = \frac{P(Y=1\,|\,X)}{P(Y=0\,|\,X)} \qquad (1) \\
\iff & \beta_0+\beta_1 X = \ln\big(\frac{P(Y=1\,|\,X)}{P(Y=0\,|\,X)}\big) \qquad (2)
\end{aligned}
$$

Equation (1) shows that if X increase by 1 unit, then $\frac{P(Y=1\,|\,X)}{P(Y=0\,|\,X)}$ will be multiply by e^{β_1}.

Equation (2) shows that if X increase by 1 unit, then $\ln\big(\frac{P(Y=1\,|\,X)}{P(Y=0\,|\,X)}\big)$ will increase by β_1.

b) e^{β_1} is the multiple of the increase in the log odds of having the out come due to a one unit increase in X.

c) I would present the estimate of e^{β_1} to explain the results of my model. From a) we know that I can explain the result like this: one group has a e^{β_1} times the odds of the other group of having the result. This is much easier to understand than explaining how the log of the log odds will increase or decrease. And The multiplication of a quantity is far more intuitive than the increase in its logarithm.