Unit 5 Homework: Learning from Random Samples

w203: Statistics for Data Science

Comparing Estimators

Say that $\{X_1, ..., X_n\}$ is an i.i.d. sample from an exponetial distribution, with common density,

$$f_X(x) = \lambda e^{-\lambda x}$$

This distribution has mean $1/\lambda$ and variance $1/\lambda^2$. You are considering the following estimators for the mean:

1. $\hat{\theta}_1 = \frac{X_1+X_2+...+X_n}{n}$
2. $\hat{\theta}_2 = \frac{X_1+X_2}{2}$
3. $\hat{\theta}_3 = \frac{X_1+X_2+...+X_n}{n+1}$

a. (3 points) Compute the bias of each estimator, $E[\hat{\theta}_i] - E[X]$.
b. (3 points) Compute the sampling variance of each estimator.
c. (3 points) Compute the MSE of each estimator.

d. (3 points) Explain in your own words, why estimator 3 has the highest bias, but the lowest MSE.

What does this mean mean?

Given an iid sample, $\{X_1, ..., X_n\}$, the geometric mean is defined as,

$$G_n = (X_1 \cdot X_2 \cdot ... \cdot X_n)^{1/n}$$

Another way of writing this is,

$$G_n = \prod_{i=1}^{n} X_i^{1/n}$$

In this exercise, let's show that the geometric mean is not a consistent estimator for E$[X]$.

For an easy counter example, assume that X has a uniform distribution on $[0, 1]$.

a. Think of a common function f such that $f(a + b) = f(a)f(b)$. Rewrite G_n in this form:

$$G_n = f\left(\frac{Y_i + Y_2 + ... + Y_n}{n}\right) = f(\frac{1}{n}\sum_{i=1}^{n} Y_i)$$

Here, each Y_i should be a function of X_i.

b. Apply the WLLN to compute the probability limit $\text{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i$. You can use the fact that $\int_0^1 lnxdx = -1$.
c. Apply the Continuous Mapping Theorem to find the probability limit $\text{plim}_{n\to\infty} G_n$. Is it the same as E$[X]$?

In the World of Competitive Coin Flipping

You have a fair coin and flip it 100 times.

a. Apply the central limit theorem and the R command `pnorm` to estimate the probability of getting between 54 and 60 heads.

b. Write a function that simulates 100 fair coin flips and returns the number of heads. Run this simulation a bunch of times (e.g. 10,000) and compute the fraction of results between 54 and 60 (inclusive of the end points)

c. Do your answers for part a) and b) match perfectly? Explain why there might be some difference.

Note: Maximum score on any homework is 100%