Data Science and Social Inquiry: HW1

R11323006 Bo-Yu, Chen

September 26, 2022

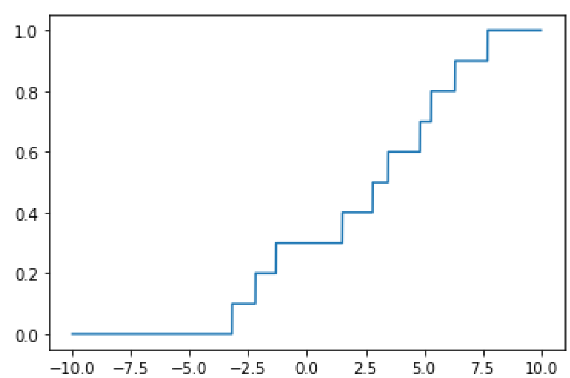
(a) (2pts) What is $\hat{F}_n(4)$? How about $\hat{F}_n(-3)$?

$$\hat{F}_n(4) = 0.6$$
$$\hat{F}_n(-3) = 0.1$$

(b) (1pt) In the previous part, we find $\hat{F}_n(x)$ at two points, namely $x = 4$ and $x = -3$. We can of course keep going and try other values of x, but this is rather repetitive and boring. Luckily, we have computers, and we know how to write program. Write a program to find out $\hat{F}_n(x)$ for $x = -10,\ -9.99,\ -9.98, ...,\ 9.99,\ 10$.

(c) (1pt) Use the result from (b) and plot $\hat{F}_n(\cdot)$. How does it look like? Is it non-decreasing? [1]

Yes, it's non-decreasing.



Now, let's investigate the statistical property of $\hat{F}_n(\cdot)$. For parts (d) - (f), we no longer assume $n = 10$, and we will treat $X_1, X_2, ..., X_n$ as random. For simplicity, we focus on in $F_n(0)$, the probability that X is less than or equal to 0, for the rest of this exercise.

[1] $\hat{F}_n(\cdot)$ is non-decreasing if $\hat{F}_n(x_1) \leq \hat{F}_n(x_2)$ for $x_1 \leq x_2$.

(d) (1pt) What is the expected value of $\hat{F}_n(0)$? Does it depend on n? **Hint**: $\mathbb{1}_{(-\infty,x]}(X_i)$ takes value only in 0 and 1. Which family of random variable only takes value in 0 and 1? What is its expected value? P.S. Your answer can be is related to $F_X(\cdot)$.

let

$$Y_i = \mathbb{1}_{(-\infty,x]}(X_i) = \begin{cases} 1, & X_i \leq 0 \\ 0, & \text{o.w.} \end{cases}$$

We can denote that

$$Y_i \overset{i.i.d.}{\sim} Bernoulli(p) \text{ where } p = P(X_i \leq 0) = F_X(0)$$

so

$$\begin{aligned}
\mathbb{E}(\hat{F}_n(0)) &= \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\mathbb{1}_{(-\infty,x]}(X_i)) \\
&= \mathbb{E}(\mathbb{1}_{(-\infty,x]}(X_i)) \\
&= \mathbb{E}(Y) \\
&= P(Y = 1) = P(X_i \leq x) \\
&= F_X(x)
\end{aligned}$$

(e) (1pt) What is the variance of $\hat{F}_n(0)$? Does it depend on n?

$$\begin{aligned}
Var(\hat{F}_n(0)) &= Var(\frac{1}{n}\sum_{i=1}^{n} Y_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) \\
&= np(1-p) \\
&= \frac{1}{n} F_X(0)[1 - F_X(0)]
\end{aligned}$$

The third equation holds because Y_i is $i.i.d$

(f) (1pt) What happens when $n \to \infty$? Do you think $\hat{F}_n(0)$ is a good estimator of $F_X(0)$?

By WLLN,

$$\bar{X} \to \mathbb{E}(X), \text{ where } \bar{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i$$

and we know

$$\hat{F}_n(0) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x]}(X_i),$$

so

$$\hat{F}_n(0) \to \mathbb{E}(\mathbb{1}_{(-\infty,x]}(X_i)) = F_X(0) \text{ as } n \to \infty$$

Therefore, $\hat{F}_n(0)$ is a good estimator of $F_X(0)$ because it satisfies unbiasedness and consistency.

An alternative way to study the statistical property of $\hat{F}_x(0)$ is through conducting simulation experiments, which are commonly known as ***Monte Carlo simulations***.

A simulation experiment typically contains many rounds. In each round, we will draw a random sample $(X_1, X_2, ..., X_n)$ from a distribution chosen by the researcher and calculate $\hat{F}_n(x)$ given $(X_1, X_2, ..., X_n)$. For example, we can set $n = 100$, generate

$$X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} N(0,1),$$

in each round of the simulation, and calculate the resulting $\hat{F}_n(0)$.

If we run $B = 10,000$ rounds, we will get $10,000$ realizations of $\hat{F}_n(0)$. We then evaluate the performance of $\hat{F}_n(0)$ by comparing 1000 realizations of $\hat{F}_n(0)$ to its true value $F_X(0)$.

(g) (1pt) What is $F_X(0)$, the true value of the parameter of interest, given that $X \sim N(0,1)$?

$$F_X(0) = P(X < 0) = \frac{1}{2}$$

(h) (1pt) Set the seed with `numpy.random.seed(5516)`, use `numpy.random.normal` to generate $X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} N(0,1)$ for $n = 100$, and calculate $\hat{F}_{n,1}(0)$, where the subscript 1 means that $\hat{F}_{n,1}(0)$ is obtained in the first round of simulation. Repeat $10,000$ times and collect the estimates $\hat{F}_{n,1}(0)$, $\hat{F}_{n,2}(0)$, ..., and $\hat{F}_{n,10000}(0)$. Calculate the **mean squared error** (MSE)

$$\frac{1}{10000}\sum_{b=1}^{10000}\left[\hat{F}_{n,b}(0)-F_X(0)\right]^2,$$

which is the average squared distance between the estimator $\hat{F}_{n,b}(0)$ and its true value $F_X(0)$.

(i) (1pt) Repeat (h) with $n = 200$ and $n = 500$. Is MSE larger or smaller when n is larger?

Ans: MSE is smaller when n is larger.

n =	100	200	500
MSE	0.0025	0.00124	0.0005

Table 1: MSE vs. sample size

(j) **(Bonus, 2pts)** the **Central Limit Theorem** (CLT) implies that

$$\sqrt{n}(\hat{F}_n(0)-F_X(0))$$

will converge to the normal distribution. We can verify that CLT holds in our case by plotting the histogram of

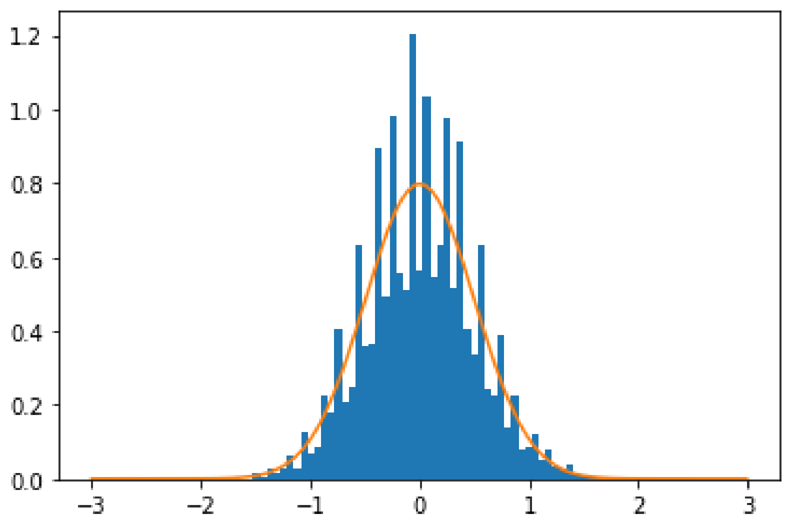
$$\sqrt{n}(\hat{F}_{n,b}(0)-F(0)),\ b=1,2,....,10000$$

for $n = 500$. Does your plot support CLT?

The orange curve is the p.d.f. of $N(0,\frac{1}{4})$ because by CLT, we have

$$\sqrt{n}(\hat{F}_n(0)-F_X(0))\sim N(0,F_X(0)[1-F_X(0)])=N(0,\frac{1}{4})$$

and we can find that the histogram is quite similar to the p.d.f. curve, so I think we have verified that CLT holds



1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3