Homework 4: Mathematical Statistics (MATH-UA 234)

Due 10/20 at the beginning of class on Gradescope

Chebyshev's inequality asserts that, for any random variable Z,

$$\mathbb{P}[|Z-\mathbb{E}[Z]| \geq \epsilon] \leq \frac{\mathbb{V}[Z]}{\epsilon^2}, \qquad \forall \epsilon > 0.$$

This fact will come in useful on this assignment.

Problem 1. *Suppose that for some unknown $\theta \in (0,1)$, $X_1, X_2, \ldots, X_n \sim \mathrm{Unif}(0,\theta)$ are independent and identically distributed. Find L_n and U_n (depending on α but not on θ) such that*

$$\mathbb{P}[\theta \in (L_n, U_n)] \geq 1-\alpha.$$

You can use Chebyshev's inequality and the results from HW3.

Solution.

Note that

$$|Z-\mathbb{E}[Z]| < \epsilon \qquad \Longleftrightarrow \qquad \mathbb{E}[Z] \in (Z-\epsilon, Z+\epsilon)$$

so, from Chebyshev's inequality, we find

$$\begin{aligned}\mathbb{P}[\mathbb{E}[Z] \in (Z-\epsilon, Z+\epsilon)] &= \mathbb{P}[|Z-\mathbb{E}[Z]| < \epsilon] \\ &= 1-\mathbb{P}[|Z-\mathbb{E}[Z]| \geq \epsilon] \\ &\geq 1-\frac{\mathbb{V}[Z]}{\epsilon^2}.\end{aligned}$$

Setting $\alpha = \mathbb{V}[Z]/\epsilon^2$ (or equivalently $\epsilon = \sqrt{\mathbb{V}[Z]/\alpha}$ we find

$$\mathbb{P}[\mathbb{E}[Z] \in (Z-\sqrt{\mathbb{V}[Z]/\alpha}, Z+\sqrt{\mathbb{V}[Z]/\alpha})] \geq 1-\alpha.$$

From homework 3, we know that $\hat{\theta}_n = \frac{2}{n}(X_1+\cdots+X_n)$ is an estimator for θ satisfying

$$\mathbb{E}[\hat{\theta}_n] = \theta, \qquad \mathbb{V}[\hat{\theta}_n] = \frac{\theta^2}{3n}.$$

Thus, we find that

$$\mathbb{P}[\theta \in (\hat{\theta}_n - \sqrt{\theta^2/(3n\alpha)}, \hat{\theta}_n + \sqrt{\theta^2/(3n\alpha)})] \geq 1-\alpha.$$

But our confidence interval should not depend on the unknown parameter theta. However, we know that $\theta \in (0,1)$ so

$$\theta \in (\hat{\theta}_n - \sqrt{\theta^2/(3n\alpha)}, \hat{\theta}_n + \sqrt{\theta^2/(3n\alpha)}) \qquad \Longrightarrow \qquad \theta \in (\hat{\theta}_n - 1/\sqrt{3n\alpha}, \hat{\theta}_n + 1/\sqrt{3n\alpha}).$$

This implies

$$\mathbb{P}[\theta \in (\hat{\theta}_n - 1/\sqrt{3n\alpha}, \hat{\theta}_n + 1/\sqrt{3n\alpha})] \geq 1-\alpha.$$

problems with a textbook reference are based on, but not identical to, the given reference

Thus, we can take our final confidence interval as

$$\left(\frac{2}{n}(X_1+\cdots+X_n)-\frac{1}{\sqrt{3n\alpha}},\frac{2}{n}(X_1+\cdots+X_n)+\frac{1}{\sqrt{3n\alpha}}\right).$$

Problem 2. *Let X be a t-step random walk with parameter p. Then we can write*

$$X=\sum_{i=1}^{t} Y_i, \qquad Y_i=\begin{cases}-1 & w.p.\ p\\ +1 & w.p.\ 1-p\end{cases}$$

where the Y_i are iid. Let F be the distribution function for X. That is, $F(x)=\mathbb{P}[X\leq x]$.

Recall that $\mathbb{E}[X]=t(1-2p)$. Thus,

$$p=\frac{1}{2}\left(1-\frac{\mathbb{E}[X]}{t}\right)=\frac{1}{2}\left(1-\frac{1}{t}\int x\mathrm{d}F(x)\right).$$

(a) Is p a linear functional? Why or why not?

(b) Let $X_1,X_2,\dots,X_n$ be iid copies of X (i.e. each X_i is a different t-step random walk with parameter p) and define

$$\hat{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[X_i\leq x].$$

What is the plug in estimator for the parameter p?

(c) Use this estimator and Chebyshev's inequality to derive a $1-\alpha$ confidence interval for p.

Solution.

(a) Suppose $T(F)$ were linear. Then $T(F)=\int r(x)\mathrm{d}F(x)$ for some $r(x)$ and therefore $T(cF)=cT(f)$ for any constant c.

However, we have

$$\frac{1}{2}\left(1-\frac{1}{t}\int x\mathrm{d}(cF)(x)\right)=\frac{1}{2}\left(1-\frac{c}{t}\int x\mathrm{d}(cF)(x)\right)\neq c\left(\frac{1}{2}\left(1-\frac{1}{t}\int x\mathrm{d}F(x)\right)\right).$$

Thus, p is not a linear functional.

(b) Since $\int x\mathrm{d}F(x)$ is a linear functional, we have

$$\int x\mathrm{d}\hat{F}_n(x)=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

Thus, we have plug in estimator

$$\hat{p}_n=\frac{1}{2}\left(1-\frac{1}{nt}\sum_{i=1}^{n}X_i\right).$$

(c) We have seen this estimator before! In particular, from the in class worksheet on 10/27 we know

$$\mathbb{E}[\hat{p}_n]=p, \qquad \mathbb{V}[\hat{p}_n]=\frac{p(1-p)}{nt}.$$

Since $p \in (0,1)$, we know $p(1-p) \leq 1/4$ (if this is not clear, you should stop and convince yourself why this is the case). Thus, using the same techniques as above, we find a confidence interval

$$\left(\frac{1}{2}\left(1-\frac{1}{nt}\sum_{i=1}^{n} X_i\right)-\frac{1}{2\sqrt{\alpha n t}}, \frac{1}{2}\left(1-\frac{1}{nt}\sum_{i=1}^{n} X_i\right)+\frac{1}{2\sqrt{\alpha n t}}\right).$$

Problem 3 (Wasserman 7.5)**.** *Let x and y be distinct points. What is* $\mathrm{Co}\mathbb{V}[\hat{F}_n(x), \hat{F}_n(y)]$*?*

Solution. Using our rules for covariance of sums,

$$\begin{aligned}\mathrm{Co}\mathbb{V}[\hat{F}_n(x), \hat{F}_n(y)] &= \mathrm{Co}\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \leq x], \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[X_j \leq y]\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Co}\mathbb{V}[\mathbb{1}[X_i \leq x], \mathbb{1}[X_j \leq y]].\end{aligned}$$

Using the independence of X_i and X_j for $i \neq j$ followed by the fact that all of the $\{X_i\}$ have the same distribution, we have

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Co}\mathbb{V}[\mathbb{1}[X_i \leq x], \mathbb{1}[X_j \leq y]] &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Co}\mathbb{V}[\mathbb{1}[X_i \leq x], \mathbb{1}[X_i \leq y]]. \\ &= \frac{1}{n}\mathrm{Co}\mathbb{V}[\mathbb{1}[X_1 \leq x], \mathbb{1}[X_1 \leq y]].\end{aligned}$$

We can now compute

$$\mathbb{E}[\mathbb{1}[X_1 \leq x]\mathbb{1}[X_1 \leq y]] = \int \mathbb{1}[t \leq x]\mathbb{1}[t \leq y]\mathrm{d}F(t) = \int_{-\infty}^{\min\{x,y\}} 1\mathrm{d}F(t) = F(\min\{x,y\}).$$

Likewise,

$$\mathbb{E}[\mathbb{1}[X_1 \leq x]] = F(x), \qquad \mathbb{E}[\mathbb{1}[X_1 \leq y]] = F(y).$$

Thus,

$$\mathrm{Co}\mathbb{V}[\mathbb{1}[X_1 \leq x], \mathbb{1}[X_1 \leq y]]. = \mathbb{E}[\mathbb{1}[X_1 \leq x]\mathbb{1}[X_1 \leq y]] - \mathbb{E}[\mathbb{1}[X_1 \leq x]]\mathbb{E}[\mathbb{1}[X_1 \leq y]] = F(\min\{x,y\}) - F(x)F(y).$$

This means

$$\mathrm{Co}\mathbb{V}[\hat{F}_n(x), \hat{F}_n(y)] = \frac{1}{n}\big(F(\min\{x,y\}) - F(x)F(y)\big).$$

Problem 4 (Wasserman 7.6)**.** *Let $X_1, \dots, X_n \sim F$ (iid) and let $\hat{F}_n$ be the empirical distribution function. Let $a < b$ be fixed numbers and define $\theta = T(F) = F(b) - F(a)$. Let $\hat{\theta}_n = T(\hat{F}_n) = \hat{F}_n(b) - \hat{F}_n(a)$.*

(a) Find the bias and standard error of $\hat{\theta}_n$ as an estimator for θ

(b) Using this result and Chebyshev's inequality, find an $1-\alpha$ confidence interval for θ.

Solution.

(a) Using the linearity of expectation we have

$$\mathbb{E}[\hat{\theta}_n] = \mathbb{E}[\hat{F}_n(b) - \hat{F}_n(a)] = \mathbb{E}[\hat{F}_n(b)] - \mathbb{E}[\hat{F}_n(a)] = F(b) - F(a) = \theta.$$

Thus, the bias is zero.

We have that

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \le b] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \le a] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \in (a,b]].$$

Using the independence of our data,

$$\mathbb{V}[\hat{\theta}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \in (a,b]]\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}[\mathbb{1}[X_i \in (a,b]]] = \frac{1}{n}\mathbb{V}[\mathbb{1}[X_1 \in (a,b]]].$$

Notice that $\mathbb{P}[X_1 \in (a,b]] = F(b) - F(a)$. Thus, $\mathbb{1}[X_1 \in (a,b]]$ is a Bernoulli random variable with parameter $F(b) - F(a)$. This implies

$$\mathbb{V}[\mathbb{1}[X_1 \in (a,b]] = (F(b) - F(a))(1 - (F(b) - F(a))).$$

Putting these together we find $\mathbb{V}[\hat{\theta}_n] = (F(b) - F(a))(1 - (F(b) - F(a)))/n$.

(b) As in problems 1 and 2, we have a confidence interval

$$\left(\hat{F}_n(b) - \hat{F}(a) - \sqrt{\frac{(F(b) - F(a))(1 - (F(b) - F(a)))}{\alpha}}, \hat{F}_n(b) - \hat{F}_n(a) + \sqrt{\frac{(F(b) - F(a))(1 - (F(b) - F(a)))}{\alpha}}\right).$$

This depends on our unknown parameters a and b. However, since $a \le b$, $F(b) - F(a) \in (0,1)$. Thus, using the same fact as in problem 2, we have the confidence interval

$$\left(\hat{F}_n(b) - \hat{F}(a) - \frac{1}{2\sqrt{\alpha}}, \hat{F}_n(b) - \hat{F}_n(a) + \frac{1}{2\sqrt{\alpha}}\right).$$

Problem 5 (Wasserman 8.5). *Let $X_1, \dots, X_n \sim F$ (iid) and $\hat{F}_n$ the empirical CDF. Let $X_1^*, \dots, X_n^* \sim \hat{F}_n$ and define*

$$\bar{X}_n^* = \frac{1}{n}(X_1^* + \cdots + X_n^*).$$

(a) What is $\mathbb{E}[\bar{X}_n^* | X_1, \dots, X_n]$ and $\mathbb{V}[\bar{X}_n^* | X_1, \dots, X_n]$?

(b) What is $\mathbb{E}[\bar{X}_n^*]$ and $\mathbb{V}[\bar{X}_n^*]$?

(c) Suppose we make iid copies $\bar{X}_{n,1}^*, \dots, \bar{X}_{n,B}^*$ of $\bar{X}_n^*$. Let

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(\bar{X}_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}\bar{X}_{n,r}^*\right)^2.$$

What is $\mathbb{E}[v_{\text{boot}}]$?

(d) Suppose we use v_{boot} as an approximation for $\mathbb{V}[\bar{X}_n^*]$. Describe the potential sources of error and when this would be a good/bad approximation.

Solution.

(a) Conditioning on $X_1, \dots, X_n$, $X_1^*, \dots, X_n^*$ are independent, so by the linearity of expectation

$$\mathbb{E}[\bar{X}_n^*|X_1, \dots, X_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i^*|X_1, \dots, X_n\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^*|X_1, \dots, X_n] = \mathbb{E}[X_1^*|X_1, \dots, X_n].$$

Finally (note about repeated values),

$$\mathbb{E}[\bar{X}_n^*|X_1, \dots, X_n] = \mathbb{E}[X_1^*|X_1, \dots, X_n]. = \sum_{j=1}^n X_j \mathbb{P}[X_1^* = X_j] = \sum_{j=1}^n X_j \frac{1}{n} = \bar{X}_n.$$

Again using independence,

$$\mathbb{V}[\bar{X}_n^*|X_1, \dots, X_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n X_i^*|X_1, \dots, X_n\right] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[X_i^*|X_1, \dots, X_n] = \frac{1}{n}\mathbb{V}[X_1^*|X_1, \dots, X_n].$$

Finally,

$$\mathbb{V}[\bar{X}_n^*|X_1, \dots, X_n] = \frac{1}{n}\mathbb{V}[X_1^*|X_1, \dots, X_n] = \frac{1}{n}\sum_{j=1}^n (X_j - \bar{X}_n)^2 \frac{1}{n} = \frac{1}{n^2}\sum_{j=1}^n (X_j - \bar{X}_n)^2$$

(b) Using the law of iterated expectation we know $\mathbb{E}[\bar{X}_n^*] = \mathbb{E}[\mathbb{E}[\bar{X}_n^*|X_1, \dots, X_n]$. Thus,

$$\mathbb{E}[\bar{X}_n^*] = \mathbb{E}[\mathbb{E}[\bar{X}_n^*|X_1, \dots, X_n] = \mathbb{E}[\bar{X}_n] = \mathbb{E}[X_1].$$

Now note that $\mathbb{E}[(\bar{X}_n^*)^2|X_1, \dots, X_n] = n^{-2}\sum_{i=1}^n X_i^2$. Thus,

$$\begin{aligned}
\mathbb{V}[\bar{X}_n^*] &= \mathbb{E}[(\bar{X}_n^*)^2] - \mathbb{E}[\bar{X}_n^*]^2 \\
&= \mathbb{E}[\mathbb{E}[(\bar{X}_n^*)^2|X_1, \dots, X_n]] - \mathbb{E}[\mathbb{E}[\bar{X}_n^*|X_1, \dots, X_n]]^2 \\
&= \left(\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[X_i^2]\right) - \mathbb{E}[X_1]^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n \left(\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}[X_i] \\
&= \frac{1}{n}\mathbb{V}[X_1].
\end{aligned}$$

(c) Write $\bar{X}_{n,i}^* = Y_i$ and $\bar{X}_n^* = Y$. Then we see that Y_i are iid copies of Y (this is just introducing new notation). Using this notation, we can write

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^B \left(Y_b - \frac{1}{B}\sum_{r=1}^B Y_r\right)^2.$$

This is the sample variance of B copies of Y. Thus, using past computations we have done many times

$$\begin{aligned}
\mathbb{E}[v_{\text{boot}}] &= \mathbb{E}\left(\frac{1}{B}\sum_{b=1}^B \left(Y_b - \frac{1}{B}\sum_{r=1}^B Y_r\right)^2\right) \\
&= \frac{B-1}{B}\mathbb{E}\left(\frac{1}{B-1}\sum_{b=1}^B \left(Y_b - \frac{1}{B}\sum_{r=1}^B Y_r\right)^2\right) \\
&= \frac{B-1}{B}\mathbb{V}[Y].
\end{aligned}$$

But recall $\mathbb{V}[Y] = \mathbb{V}[\bar{X}_n^*] = \frac{1}{n}\mathbb{V}[X_1]$ by part (b). Thus,

$$\mathbb{E}[v_{\text{boot}}] = \frac{B-1}{nB}\mathbb{V}[X_1].$$

(d) We know that the sample variance converges in probability to the true variance when the number of samples is large. Thus,

$$\mathbb{V}[v_{\text{boot}}] \xrightarrow{P} \mathbb{V}[Y] = \mathbb{V}[\bar{X}_n^*].$$

However, for finite B, the sample variance might not have convergence.

Thus, we have error due to the finite sample size which will be bigger if B is smaller.

Problem 6 (Wasserman 7.7). *Suppose $X_1, X_2, \ldots, X_n \sim \mathrm{Unif}(0,\theta)$ are independent and identically distributed. Assume that the $X_1, \ldots, X_n$ are distinct (this happens with probability one) and let $\hat{\theta}_n = \max\{X_1, \ldots, X_n\}$. Let $\hat{\theta}_n^* = \max\{X_1^*, \ldots, X_n^*\}$, where $X_1^*, \ldots, X_n^* \sim \hat{F}_n$ (iid).*

(a) Show that for any $t \in \mathbb{R}$, $\mathbb{P}[\hat{\theta}_n = t] = 0$.

(b) Show that there exists $t \in \mathbb{R}$ such that $\mathbb{P}[\hat{\theta}_n^* = t] \neq 0$. In particular, show that $\mathbb{P}[\hat{\theta}_n^* = \hat{\theta}] \approx 0.632$ for n large. (Hint: show that $\mathbb{P}[\hat{\theta}_n^* = \hat{\theta}_n] = 1-(1-1/n)^n$)

(c) Recall that our bootstrap sample $\hat{\theta}_n^*$ is meant to approximate $\hat{\theta}_n$. However, as seen in (b), $\hat{\theta}_n^*$ is far more likely than certain values than $\hat{\theta}_n$. Discuss why the distribution of $\hat{\theta}_n$ and $\hat{\theta}_n^*$ is so different and whether bootstrapping is effective.

Solution.

(a) We have previously computed the distribution function for $\hat{\theta}_n$. In particular,

$$\mathbb{P}[\hat{\theta}_n \leq x] = \begin{cases} x^n/\theta^n & x \in (0,\theta) \\ 0 & x \notin (0,\theta). \end{cases}$$

This is continuous, so for any t, $\mathbb{P}[\hat{\theta}_n = t] = 0$.

(b) We have $\hat{\theta}_n^* = \hat{\theta}_n$ if any of the $\hat{X}_i^* = \hat{\theta}_n$. For any given i, the probability that we pick the $\hat{\theta}_n$ is $1/n$, since there are n possible X_i we could sample, but only one of them is the max. Thus, the probability that we did not pick $\hat{\theta}_n$ is $(1-1/n)$, and the probability none of the X_i are equal to $\hat{\theta}_n$ is $(1-1/n)^n$ (since they are independent). Finally, the probability that at least one of the X_i was equal to $\hat{\theta}_n$ is $1-(1-1/n)^n$.

It's a well known fact that

$$\lim_{n\to\infty} 1-(1-1/n)^n = 1-1/\mathrm{e} \approx 0.632\ldots.$$

(c) Bootstrapping cannot tell us anything that was not already in our data. In particular, the maximum value of bootstrapping will never be bigger than the maximum of the data. Thus $\hat{\theta}_n^*$ can never be bigger than $\hat{\theta}_n$. Moreover, because bootstrapping simply ammounts to resampling from our data, it is fairly likely that at least one of our samples will be of $\hat{\theta}_n$ in which case the maximum possible value of $\hat{\theta}_n^*$ is attained.

This illustrates the fact that the assumption that our bootstrap random variable is a good approximation to the true random variable we have is very important, and that we must take care if the assumption is not satisfied.

Problem 7. *Below is data for the hospitalization time (in hours) of $n = 200$ patients with COVID 19. i* 108, 144, 89, 122, 53, 153, 165, 183, 101, 114, 115, 203, 31, 51, 31, 109, 45, 85, 72, 107, 80, 157, 73, 107, 19, 140, 183, 38, 112, 143, 49, 61, 46, 99, 42, 79, 81, 53, 112, 79, 136, 149, 38, 52, 125, 92, 80, 79, 91, 110, 65, 12, 46, 59, 62, 39, 119, 103, 95, 97, 109, 104, 17, 184, 37, 110, 118, 166, 20, 44, 66, 118, 13, 151, 163, 90, 99, 80, 166, 89, 37, 64, 174, 24, 110, 36, 75, 90, 145, 59, 96, 69, 25, 43, 144, 55, 49, 53, 98, 89, 52, 91, 88, 74, 104, 188, 64, 67, 153, 153, 104, 36, 107, 2, 34, 169, 152, 84, 34, 54, 93, 89, 28, 116, 141, 63, 124, 95, 127, 28, 118, 99, 97, 97, 61, 100, 68, 103, 90, 131, 79, 144, 147, 46, 82, 117, 126, 94, 155, 111, 50, 215, 76, 35, 80, 94, 167, 27, 106, 85, 142, 116, 91, 126, 178, 47, 25, 114, 15, 71, 128, 151, 119, 104, 124, 125, 40, 59, 77, 105, 100, 123, 30, 65, 136, 136, 253, 158, 205, 145, 42, 173, 67, 49, 67, 92, 79, 115, 24, 97

(a) *Plot the empirical CDF $\hat{F}_n$. Make sure the label the horizontal axis*

(b) *Compute the value of the plug-in estimator for the mean hospitalization time $\int x\mathrm{d}F(x)$ for the given data.*

(c) *Suppose we use bootstrapping to sample $X_1^*, \dots, X_n^* \sim \hat{F}_n$ (iid) and get $\bar{X}_n^* = \frac{1}{n}(X_1^* + \cdots + X_n^*)$. We can then compute the estimate the bootstrap variance by*

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(\bar{X}_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}\bar{X}_{n,r}^*\right)^2$$

where $\bar{X}_{n,1}^*, \dots, \bar{X}_{n,B}^*$ are B iid copies of $\bar{X}_n^*$. Compute

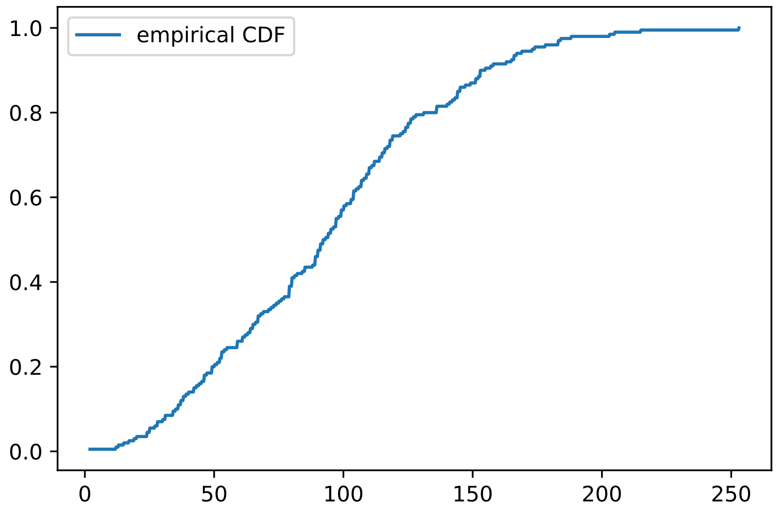
$$\lim_{B\to\infty} v_{\text{boot}}.$$

(d) *What assumptions have you made throughout this process?*

Solution.

(a)
```
import numpy as np
import scipy as sp
from scipy import stats

import matplotlib.pyplot as plt

X= [108, 144, 89, 122, 53, 153, 165, 183, 101, 114, 115, 203, 31, 51, 31, 109,45, 8
73, 107, 19, 140, 183, 38, 112, 143, 49, 61, 46,  99, 42, 79, 81, 53, 112, 79, 136,
39,119, 103, 95, 97, 109, 104, 17, 184, 37, 110, 118, 166, 20, 44, 66, 118, 13, 151
80, 166, 89, 37, 64, 174, 24, 110, 36, 75, 90, 145, 59, 96, 69, 25, 43, 144,55, 49,
89, 52, 91, 88, 74, 104, 188, 64, 67, 153, 153, 104, 36, 107,  2, 34,169, 152, 84,
89, 28, 116, 141, 63, 124, 95, 127, 28, 118, 99,  97, 97, 61, 100, 68, 103, 90, 131
46, 82, 117, 126, 94, 155, 111, 50, 215, 76, 35,  80, 94, 167, 27, 106, 85, 142, 11
47, 25, 114, 15, 71, 128, 151, 119, 104, 124, 125,  40, 59, 77, 105, 100, 123, 30,
24, 97]

n = len(X)

def F_n_hat(x):
  out = 0
```

```
    for i in range(n):
        out += (1/n)*(X[i]<=x)
    return out

xx = np.linspace(np.min(X),np.max(X),1000)
plt.plot(xx,F_n_hat(xx),label='empirical CDF')
plt.legend()
```



(b) We know $\int x d\hat{F}_n(x) = n^{-1}\sum_{i=1}^n X_i$ which we can compute as 93.945.

(c) Similar to 5(c) we have

$$\mathbb{V}[v_{\text{boot}}|X_1,\ldots,X_n] \xrightarrow{P} \mathbb{V}[Y] = \mathbb{V}[\bar{X}_n^*|X_1,\ldots,X_n] = \frac{1}{n^2}\sum_{j=1}^n (X_j - \bar{X}_n)^2.$$

We can easily compute this as 2120.901975.

Problem 8. *What is one thing that is going well in the class and one thing you would like to improve on?*