

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 ,

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

This fact will come in useful on this assignment.

Problem 1. Suppose that for some unknown $\theta \in (0, 1)$, $X_1, X_2, \dots, X_n \sim \text{Unif}(0, \theta)$ 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.

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

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

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,

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

(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 .

Problem 3 (Wasserman 7.5). Let x and y be distinct points. What is $\text{CoV}[\hat{F}_n(x), \hat{F}_n(y)]$?

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 .

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

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

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

- (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

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

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.

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

Problem 7. Below is data for the hospitalization time (in hours) of $n = 200$ patients with COVID 19.

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 $\int x dF(x)$ for the given data.
- (c) Suppose we use bootstrapping to sample $X_1^*, \dots, X_n^* \sim \hat{F}_n$ (iid) and get $\bar{X}_n^* = \frac{1}{n}(X_1^* + \dots + X_n^*)$. 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

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

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

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