STAT 510: Homework 04

David Dalpiaz

Due: Friday, September 25, 11:59 PM

General Directions

This assignment is worth 10 points. For each exercise, you may obtain a score of 0, 0.5, or 1.

- To obtain a score of **1**, your answer must be correct, contain valid supporting work, and be reasonably formatted up to and including boxing your answer when possible.
- A score of **0.5** will be given to solutions which show reasonabe effort, but contain errors. (A score of **1** may be granted to some solutions containing errors if they are extremely minor.)
- A score of **0** will be given to a blank solution or a solution that shows no reasonable progress towards the correct solution. Note that if you do not indicate a page for a problem on Gradescope, it will be considered blank.

Please submit your assignment to Gradescope by the due date listed above. You may submit up to 48 hours late with a two point late penalty. After that, no late work will be accepted.

Any grade disputes must be petitioned through Gradescope within one week of receiving a grade.

Please attempt to submit your work as a single PDF and complete the process of indicating which problem is on which page. You may need to merge together PDF files from various sources and scans. We will keep track of best practice for submitting to Gradescope in this Piazza thread.

Homework assignments are meant to be learning experiences. You may discuss the exercises with other students, but you must write the solutions on your own. Directly sharing or copying any part of a homework solution is an infraction of the University's rules on academic integrity. Any violation will be punished as severely as possible.

For this, and all homework assignments, you may use any computational tools that you wish, such as a statistical computing enviroment or integral solver. The course staff is most familiar with `R`, so we will be able to best support `R` users, but you may use any software that you like.

Graded Exercises

Exercise 1 (Make It So)

Let $X_1, X_2, \ldots, X_n \sim \text{Uniform}(0, \theta)$. Consider the estimator

$$\hat{\theta} = \max\{X_1, X_2, \ldots, X_n\}.$$

Find the bias, variance, and MSE of this estimator. Assuming the estimator is biased, create a new estimator which is a simple function of $\hat{\theta}$ that is unbiased.

Exercise 2 (More Data, Less Problems)

Let $X_1, X_2, \ldots, X_n \sim \text{Uniform}(0, \theta)$. Consider the estimator

$$\hat{\theta} = 2 \cdot \bar{X}_n$$

Find the bias, variance, and MSE of this estimator. Is this estimator consistent? Justify.

Exercise 3 (A Little Bit of Bias Goes a Long Way)

Let Y have a binomial distribution with parameters n and p. Consider two estimators for p:

$$\hat{p}_1 = \frac{Y}{n}$$

and

$$\hat{p}_2 = \frac{Y+1}{n+2}$$

For what values of p does $\hat{p}_2$ achieve a lower mean square error than $\hat{p}_1$?

Exercise 4 (Dependence in the Empirical Distribution)

Let x and y be two distinct points. Find

$$\text{Cov}\left(\hat{F}_n(x), \hat{F}_n(y)\right).$$

Exercise 5 (Empirical Distribution Properties)

For any fixed value of x, show each of the following.

$$\mathbb{E}\left[\hat{F}_n(x)\right] = F(x)$$

$$\mathbb{V}\left[\hat{F}_n(x)\right] = \frac{F(x) \cdot (1 - F(x))}{n}$$

$$\text{MSE}\left[\hat{F}_n(x)\right] = \frac{F(x) \cdot (1 - F(x))}{n} \to 0$$

$$\hat{F}_n(x) \xrightarrow{p} F(x)$$

Exercise 6 (Limiting Distribution of Empirical Distribution)

Let $X_1, X_2, \ldots, X_n \sim F$. Given the empirical distribution function $\hat{F}_n(x)$ and a fixed point x, use the central limit theorem to find the limiting distribution of $\sqrt{n}\left(\hat{F}_n(x) - F(x)\right)$.

Exercise 7 (Using Statistical Functionals)

Let $X_1, X_2, \ldots, X_n \sim F$ and let $\hat{F}_n(x)$ be the empirical distribution function. Let fixed numbers $a < b$ and define

$$\theta = T(F) = F(b) - F(a).$$

Find the estimated standard deviation of

$$\hat{\theta} = T\left(\hat{F}_n(x)\right) = \hat{F}_n(b) - \hat{F}_n(a).$$

Exercise 8 (More Coverage)

Let $X_1, X_2, \ldots X_n \sim \text{Bernoulli}(p)$. Set $n = 100$ and $\alpha = 0.05$. Consider two confidence intervals for p. For both, define

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

First, consider the interval from the previous homework that we justified via Hoeffding's inequality.

$$C_n^H = \left(\hat{p}_n - \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)},\ \ \hat{p}_n + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}\right)$$

Second, consider the "normal" interval,

$$C_n^N = \left(\hat{p}_n - z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}},\ \ \hat{p}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right).$$

Use simulation to check these intervals' coverage and expected length. Report your results using appropriate plots. Consider as many values of p as you can, but at minimum use

$$p \in (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9).$$

Comment on the validity of these intervals and the interval lengths.

Exercise 9 (Empirical Distribution Confidence Bands)

The following code simulates data from three different distributions.

```
set.seed(42)
data_1 = rexp(n = 100)
data_2 = rnorm(n = 25)
data_3 = rt(n = 500, df = 3)
```

For each, plot the empirical distribution with 95% confidence bands. For each, overlay the true cumulative distribution function. Do not use R's `ecdf()` function or anything similar. You may use R's `stepfun()` function.

Exercise 10 (Estimating Functionals with the Empirical Distribution)

The following code simulates data from a Weibull distribution.

```
set.seed(42)
some_data = rweibull(n = 250, shape = 2, scale = 3)
```

Use the empirical distribution function to create plug-in estimates of the following:

- Mean
- Variance
- Skewness
- Median

Compare these results to their true values given the data generating process defined above. Report your results as a table.

Exercise 11 (BYOQ: Bring Your Own Question)

Submit your own question with a solution! If accepted, you will receive one **buffer point**. To be accepted:

- Your question must be reasonably challenging. It should be at least as challenging as the "average" question on this assignment.
- Is must be **original**. If, based on some quick searching we can find the *exact* question (that is, you just copy-pasted a question), you will **lose** a point instead of receiving a buffer point. If we find that it is simply a derivative of another exercise (for example, just changing a constant) you will not receive a point.

The instructor may create videos solving these, or they may be circulated for additional practice.