DATA 1010
PROBLEM SET 12
DUE 17 DECEMBER 2018 AT 12 PM

Problem 1

Label each of the following four estimators as either (i) biased and consistent, (ii) biased and inconsistent, (iii) unbiased and consistent, or (iv) unbiased and inconsistent. The matching will be one-to-one.

(a) $X_1, X_2, \ldots$ are i.i.d. Bernoulli random variables with unknown p and estimator

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

(b) $X_1, X_2, \ldots$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with unknown μ and σ^2 and estimator

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}.$$

(c) $X_1, X_2, \ldots$ are i.i.d. uniform random variables on an unknown bounded interval. For $n \geq 100$ we estimate the mean using

$$\widehat{\mu} = \frac{\sum_{i=1}^{100} X_i}{100}.$$

(d) $X_1, X_2, \ldots$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with unknown μ and σ^2. For $n \geq 100$ we estimate the standard deviation using

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{100}(X_i - \overline{X})^2}{99}}.$$

Problem 2

Suppose that $X_1, \ldots, X_n$ are independent Unif$[0, \theta]$ random variables, where θ is an unknown parameter, and consider the following estimators for θ:

$$\widehat{\theta}_1 = \max(X_1, \ldots, X_n), \qquad \widehat{\theta}_2 = 2 \cdot \frac{X_1 + \cdots + X_n}{n}.$$

(a) Find the CDF of $\widehat{\theta}_1$.

(b) Recall that if $F_{\widehat{\theta}_1}(x)$ and $f_{\widehat{\theta}_1}(x)$ are the CDF and PDF of $\widehat{\theta}_1$ respectively, then $\frac{d}{dx}F_{\widehat{\theta}_1}(x) = f_{\widehat{\theta}_1}(x)$.

Differentiate your answer to (a) to find the PDF of $\widehat{\theta}_1$.

(c) Show that $\widehat{\theta}_1$ is consistent.

(d) Find $\mathbb{E}\left[\widehat{\theta}_1\right]$ and $\mathbb{E}\left[\widehat{\theta}_2\right]$. Which estimator is biased?

(e) Find Var $\left(\widehat{\theta}_1\right)$ and Var $\left(\widehat{\theta}_2\right)$. Which estimator has lower variance?

(f) Show that the mean squared error of $\widehat{\theta}_1$ is less than the mean squared error of $\widehat{\theta}_2$ whenever $n \geq 3$.

Problem 3

(a) **Hoeffding's inequality** says that if $Y_1, Y_2, \ldots$ are independent random variables with the property that $\mathbb{E}[Y_i] = 0$

and $a_i \leq Y_i \leq b_i$ for all i, then for all $\epsilon > 0$ and $t > 0$, we have

$$\mathbb{P}\left(Y_1 + Y_2 + \cdots + Y_n \geq \epsilon\right) \leq \mathrm{e}^{-t\epsilon} \prod_{i=1}^{n} \mathrm{e}^{t^2(b_i - a_i)^2/8}.$$

Use Hoeffding's inequality to show that if $X_1, X_2, X_3, \ldots$ is a sequence of independent Bernoulli(p) random variables, then for all $\alpha > 0$, the interval $\left(\overline{X}_n - \sqrt{\frac{1}{2n}\log(2/\alpha)}, \overline{X}_n + \sqrt{\frac{1}{2n}\log(2/\alpha)}\right)$ is a confidence interval for p with confidence level $1 - \alpha$. Explain what happens to the width of this confidence interval if n gets large, and also what happens to the width if α is made very small.

(b) As above, consider n independent Bernoulli(p)'s. Find the normal-approximation confidence interval for p

(c) As above, consider n independent Bernoulli(p)'s. Find the Chebyshev confidence interval for p.

(d) Find the numerical values of the half-widths for each of the above confidence intervals when $p = \frac{1}{2}$, $n = 1000$, and $\alpha = 0.05$ (approximating $\overline{X}$ as p).

Problem 4

I drew 6 samples from an undisclosed distribution and obtained the following results:

```
c(6.19,7.048,6.143,5.459,4.603,4.335)
```

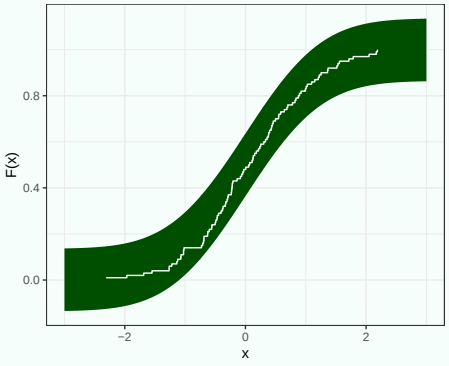
I also drew 8 samples from another undisclosed distribution and got

```
c(8.924,4.698,6.095,4.223,3.643,1.624,1.444,6.309)
```

Determine whether the Wald hypothesis test (with significance $\alpha = 0.05$) rejects the null hypothesis that the mean of the two distributions are equal.

Problem 5

One hundred samples were drawn from the standard normal distribution, and the resulting empirical CDF was plotted in the figure shown. Also plotted are the DKW bounds for $\alpha = 0.05$.



(a) Replicate this figure in `ggplot`. Some tips: do `set.seed(123)` and draw samples with `rnorm`. Evaluate the CDF of the normal distribution with `pnorm`. You can make the CDF graph using the step geom. You'll want to make some tibbles to contain the generated data, and bear in mind that you can specify the data frame on a geom-by-geom basis in ggplot.

(b) Run the code block repeatedly (though comment out the `set.seed` line first!) until you see the empirical CDF graph fall outside the ribbon. How many times did it take? Does your result seem consistent with the claim that the graph of the empirical CDF lies in the ribbon with probability at least $1 - \alpha$?

Problem 6

Consider a distribution ν which is known only via a dozen samples therefrom, the values of which are

```
c(8.924,4.698,6.095,4.223,3.643,1.624,1.444,6.309)
```

(a) Obtain a bootstrap estimate of the standard deviation of the median of five independent samples from ν.

(b) The actual standard deviation of the median of 5 samples from ν is approximately 2.14. How close is the value

you found? Could you have gotten as close as desired to this value by choosing sufficiently many bootstrap re-samplings?

Problem 7

Consider the family of densities

$$\left\{ \frac{1}{2(\beta-\alpha)} \mathbf{1}_{\{\alpha \le x \le \beta\}} + \frac{1}{2(\delta-\gamma)} \mathbf{1}_{\{\gamma \le x \le \delta\}} \, : \, \alpha < \beta < \gamma < \delta \right\}.$$

Show that there is no maximum likelihood estimator for this family of distributions. In other words, consider a set of samples drawn from one of these distributions, and show that arbitrarily large likelihoods may be obtained for those samples by choosing suitable values for the parameters.