Homework 2 Solutions
Statistics 200B
Due Feb. 7, 2019

1. Let $X_1, X_2, \ldots \overset{iid}{\sim} Unif(0,\theta)$. Consider the following two estimators of θ:

$$\begin{aligned}\hat{\theta}_n &= \max\{X_1,\ldots,X_n\}\\ \tilde{\theta}_n &= 2\bar{X}_n\end{aligned}$$

(a) Find the PDF of $\hat{\theta}_n$.

(b) Find the bias, standard error, and MSE of $\hat{\theta}_n$.

(c) Find the bias, standard error, and MSE of $\tilde{\theta}_n$.

(d) Fix $\theta = 1$ and use `R` to make a plot of both MSEs as a function of n. (That is, put two lines on the same plot.) What does the plot tell you about the conditions under which we might prefer $\hat{\theta}_n$ or $\tilde{\theta}_n$?

Solution:

(a) The CDF of $\hat{\theta}_n$ is

$$F_{\hat{\theta}_n}(t) = P(\hat{\theta}_n \le t) = P(X_1 \le t, \ldots, X_n \le t) = \prod_{i=1}^n P(X_i \le t) = \left(\frac{t}{\theta}\right)^n.$$

The PDF of $\hat{\theta}_n$ is

$$f_{\hat{\theta}_n}(t) = \frac{d}{dt}F_{\hat{\theta}_n}(t) = \frac{n}{\theta^n}t^{n-1}.$$

(b)

$$\text{bias}(\hat{\theta}_n) = E[\hat{\theta}_n] - \theta = \int_0^\theta t f_{\hat{\theta}_n}(t)dt - \theta = \int_0^\theta \frac{n}{\theta^n}t^n dt = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta.$$

$$\begin{aligned}V[\hat{\theta}_n] = E[\hat{\theta}_n^2] - (E[\hat{\theta}_n])^2 &= \int_0^\theta t^2 f_{\hat{\theta}_n}(t)dt - \left(\frac{n}{n+1}\theta\right)^2 = \int_0^\theta \frac{n}{\theta^n}t^{n+1}dt - \frac{n^2}{(n+1)^2}\theta^2\\ &= \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \frac{n}{(n+1)^2(n+2)}\theta^2.\end{aligned}$$

$$\text{MSE}(\hat{\theta}_n) = [\text{bias}(\hat{\theta}_n)]^2 + V[\hat{\theta}_n] = \frac{1}{(n+1)^2}\theta^2 + \frac{n}{(n+1)^2(n+2)}\theta^2 = \frac{2}{(n+1)(n+2)}\theta^2.$$

(c) Since $X_1, X_2, \ldots \overset{iid}{\sim} Unif(0, \theta)$,

$$E[\bar{X}_n] = \frac{\theta}{2}, \quad V[\bar{X}_n] = \frac{\theta^2}{12n}.$$

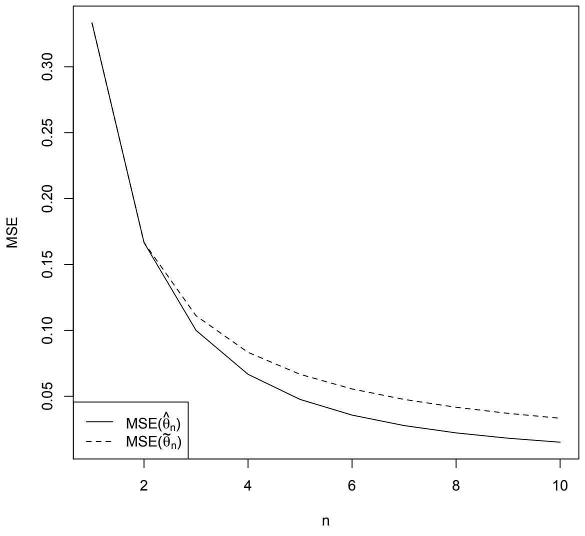
So

$$\begin{aligned}
&\text{bias}(\tilde{\theta}_n) = E[\tilde{\theta}_n] - \theta = 2E[\bar{X}_n] - \theta = \theta - \theta = 0. \\
&V[\tilde{\theta}_n] = 4V[\bar{X}_n] = \frac{\theta^2}{3n}. \\
&\text{MSE}(\tilde{\theta}_n) = [\text{bias}(\tilde{\theta}_n)]^2 + V[\tilde{\theta}_n] = \frac{\theta^2}{3n}.
\end{aligned}$$

(d) `R` code is as follows.

```
n <- 1:10
plot(x=n, y=2/((n+1)*(n+2)), xlab="n", ylab="MSE", type="l")
lines(x=n, y=1/(3*n), lty=2)
legend("bottomleft", legend=c(expression(paste("MSE(",hat(theta)[n],")",
    sep="")), expression(paste("MSE(",tilde(theta)[n],")",sep=""))),
    lty=1:2)
```

The plot is as below.



From the plot we can see that $\hat{\theta}_n$ is preferred, as it has smaller MSE for positive integer values of n.

2. Again let $X_1, X_2, \ldots \overset{iid}{\sim} Unif(0, \theta)$. Consider a confidence interval for θ constructed as $[a\hat{\theta}_n, b\hat{\theta}_n]$, where $\hat{\theta}_n = \max\{X_1, \ldots, X_n\}$. Calculate the coverage of this interval and show that it depends only on a and b. If $a = 1$, what should b be to obtain a coverage of 95%?

   Solution:

$$\begin{aligned} P(a\hat{\theta}_n \le \theta \le b\hat{\theta}_n) &= P(\theta/b \le \hat{\theta}_n \le \theta/a) = P(\hat{\theta}_n \le \theta/a) - P(\hat{\theta}_n \le \theta/b) \\ \text{(from 1(a))} \quad &= (1/a)^n - (1/b)^n, \end{aligned}$$

   so the coverage of interval $[a\hat{\theta}_n, b\hat{\theta}_n]$ depends only on a and b.

   If $a = 1$, to make $P(a\hat{\theta}_n \le \theta \le b\hat{\theta}_n) = 0.95$, we need

$$1 - (1/b)^n = 0.95 \quad \Rightarrow \quad b = 20^{1/n}.$$

3. Let $X_1, \ldots, X_n \sim Bernoulli(p)$ and let $Y_1, \ldots, Y_n \sim Bernoulli(q)$. Find the plug-in estimator and estimated standard error for p. Find an approximate 90% confidence interval for p. Find the plug-in estimator and estimated standard error for $p - q$. Find an approximate 90% confidence interval for $p - q$.

   Solution:

   Since $p = E[X_i] = \int x dF(x)$, the plug-in estimator is $\hat{p} = \int x d\hat{F}(x) = \bar{X}_n$.

   The estimated standard error for p is

$$\hat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}.$$

   An approximate 90% confidence interval for p is

$$\hat{p} \pm z_{0.05}\hat{se}(\hat{p}) = \bar{X}_n \pm 1.64\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}.$$

   The plug-in estimator and estimated standard error for $p - q$ are

$$\hat{p} - \hat{q} = \bar{X}_n - \bar{Y}_n, \quad \hat{se}(\hat{p} - \hat{q}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{q}(1-\hat{q})}{n}} = \sqrt{\frac{\bar{X}_n(1-\bar{X}_n) + \bar{Y}_n(1-\bar{Y}_n)}{n}}.$$

   An approximate 90% confidence interval for $p - q$ is

$$(\hat{p} - \hat{q}) \pm z_{0.05}\hat{se}(\hat{p} - \hat{q}) = (\bar{X}_n - \bar{Y}_n) \pm 1.64\sqrt{\frac{\bar{X}_n(1-\bar{X}_n) + \bar{Y}_n(1-\bar{Y}_n)}{n}}.$$

4. A manufacturer of booklets packages them in boxes of 100. It is known that, on average, the booklets weigh one ounce, with a standard deviation of 0.05 ounce. The manufacturer is interested in calculating

$$P(100 \text{ booklets weigh more than } 100.4 \text{ ounces}),$$

a number that would help detect whether too many booklets are being put into a box. Explain how you would calculate an approximate value of this probability. Mention any relevant theorems or assumptions needed.

Solution:

Define the weight of the ith booklet as X_i. We assume that $X_1, \ldots, X_{100}$ are IID, and we know that $E[X_i] = 1$ and $V[X_i] = 0.05^2$.

From Central Limit Theorem,

$$\sqrt{n}(\bar{X}_n - 1) \overset{D}{\to} N(0, 0.05^2).$$

So $\bar{X}_{100}$ is approximately distributed as $N(1, 0.05^2/100)$ for large n.

Then

$$\begin{aligned}
&P(100 \text{ booklets weigh more than } 100.4 \text{ ounces}) \\
=&P(\sum_{i=1}^{100} X_i \geq 100.4) = P(\bar{X}_{100} \geq 1.004) \\
=&P\left(Z \geq \frac{1.004 - 1}{0.005}\right) = P(Z \geq 0.8) \\
=&0.21.
\end{aligned}$$

5. Let $X_1, \ldots, X_n \sim F$ 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 $\hat{\theta} = T(\hat{F}_n) = \hat{F}_n(b) - \hat{F}_n(a)$. Find the estimated standard error of $\hat{\theta}$. Find an expression for an approximate $1 - \alpha$ confidence interval for θ.

Solution:

$$\hat{\theta} = \hat{F}_n(b) - \hat{F}_n(a) = \frac{\sum_{i=1}^n I(X_i \leq b)}{n} - \frac{\sum_{i=1}^n I(X_i \leq a)}{n} = \frac{\sum_{i=1}^n I(a < X_i \leq b)}{n}$$

Define $Y_i = I(a < X_i \leq b)$. Then

$$E[Y_i] = F(b) - F(a), \quad V[Y_i] = [F(b) - F(a)][1 - F(b) + F(a)].$$

So $\hat{\theta} = \bar{Y}_n$, and

$$se(\hat{\theta}) = \sqrt{\frac{[F(b) - F(a)][1 - F(b) + F(a)]}{n}}.$$

The estimated standard error of $\hat{\theta}$ is

$$\hat{se}(\hat{\theta}) = \sqrt{\frac{[\hat{F}_n(b) - \hat{F}_n(a)][1 - \hat{F}_n(b) + \hat{F}_n(a)]}{n}} = \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}.$$

So an approximate $1 - \alpha$ confidence interval for θ is

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}.$$

6. Data on the magnitudes of earthquakes near Fiji are available at bCourse. Download this data and load it into `R` using

```
quakes <- read.table(file = "fijiquakes.dat", header = TRUE)
```

(You may need to change the file argument depending on where on your computer you saved the file.) Type

```
head(quakes)
```

to see the first several lines. This is a special type of object in `R` called a dataframe. You can extract elements from the dataframe using the dollar sign; for example

```
hist(quakes$mag)
```

makes a histogram of the magnitudes. Estimate the CDF $F(x)$ for the magnitudes and plot it. Compute and lines to show a 95% confidence envelope for F using the Dvoretzky-Kiefer-Wolfowitz inequality. Turn in your code and your plot.

Solutions:

The code is as follows.

```
x <- quakes$mag
xseq <- seq(min(x), max(x), length = 100)
Fhat <- apply(outer(x, xseq, "<"), 2, mean) # compute the empirical CDF
n <- length(x)
epsilon <- sqrt(1/(2*n)*log(2/0.05))

L <- sapply(Fhat, FUN=function(x) max(x-epsilon, 0)) #(1)
U <- sapply(Fhat, FUN=function(x) min(x+epsilon, 1)) #(2)
# Another way to do (1) and (2) is
L <- pmax(Fhat - epsilon, 0)
U <- pmin(Fhat + epsilon, 0)
# The results should be the same

plot(xseq, Fhat, type = "s", xlab = "x", ylab = "Fhat(x)",
    main = "Magnitudes of Earthquakes")
rug(x)
lines(xseq, L, lty=2)
lines(xseq, U, lty=2)
legend("right", legend=c("Fhat(x)", "Confidence band"), lty=1:2)
```

The plot is as below.

7. In 1975, an experiment was conducted to see if cloud seeding produced rainfall. Twenty-six clouds were seeded with silver nitrate and 26 were not. The decision to seed or not was made at random. Download the file `clouds.dat` from bCourse. Let θ be the difference in the mean precipitation from the two groups (seeding minus no seeding). Estimate θ. Estimate the standard error of the estimate and produce a 95% confidence interval. Turn in your code and your results.

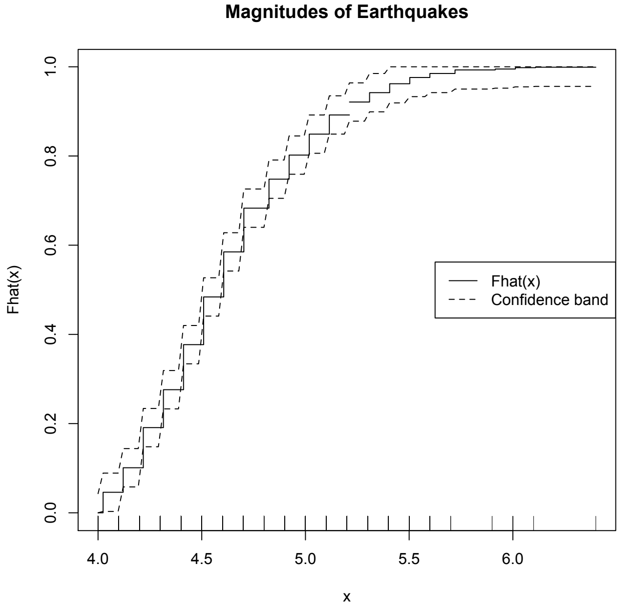
   Solutions:

   Define the seeded as $X_1, \ldots, X_n$, and the unseeded as $Y_1, \ldots, Y_n$. Then

$$\hat{\theta} = \bar{X}_n - \bar{Y}_n$$

$$se(\hat{\theta}) = \sqrt{\frac{V[X] + V[Y]}{n}} \quad \Rightarrow \quad \hat{se}(\hat{\theta}) = \sqrt{\frac{\hat{V}[X] + \hat{V}[Y]}{n}},$$

   where $\hat{V}[X] = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$.



The 95% confidence interval is

$$\hat{\theta} \pm z_{0.025}\hat{se}(\hat{\theta}).$$

The code is as follows.

```
clouds <- read.table("clouds.dat", header=TRUE)
unseeded <- clouds$Unseeded
seeded <- clouds$Seeded
n <- dim(clouds)[1]
theta_hat <- mean(seeded) - mean(unseeded)
se_hat <- sqrt((var(seeded)+var(unseeded))/n)
CI <- c(theta_hat + qnorm(0.025)*se_hat, theta_hat - qnorm(0.025)*se_hat)
print(CI)
```

The result is $[5.314116, 549.478192]$.