Homework 3 Solutions
Statistics 200B
Due Feb. 14, 2019

1. Consider again the cloud seeding data from Homework 2. Let θ be the difference in the median precipitation from the two groups. Find the plug-in estimate of θ. Using the bootstrap, estimate the standard error of the plug-in estimate and produce an approximate 95% Normal confidence interval for θ.

   Solution:

   Define $X_1, \ldots, X_n$ as the unseeded data, and $Y_1, \ldots, Y_n$ as the seeded data. Assume that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are independent.

   Suppose $X_1, \ldots, X_n \overset{IID}{\sim} F_1$, and $Y_1, \ldots, Y_n \overset{IID}{\sim} F_2$. Then $\theta = \text{median}(F_2) - \text{median}(F_1)$. The plug-in estimate of θ is

   $$\hat{\theta} = \text{median}(\hat{F}_2) - \text{median}(\hat{F}_1) = \text{median}(Y_1, \ldots, Y_n) - \text{median}(X_1, \ldots, X_n).$$

   By using bootstrap to estimate the standard error, we can draw B pairs of bootstrap samples from $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively. For the ith $(i = 1, \ldots, B)$ pair of bootstrap samples, $X^*_{1,i}, \ldots, X^*_{n,i}$ and $Y^*_{1,i}, \ldots, Y^*_{n,i}$, the bootstrapped plug-in estimate is

   $$\hat{\theta}^*_i = \text{median}(Y^*_{1,i}, \ldots, Y^*_{n,i}) - \text{median}(X^*_{1,i}, \ldots, X^*_{n,i}).$$

   Then the standard error of $\hat{\theta}$ under F_1 and F_2 can be estimated by the plug-in standard error estimate under $\hat{F}_1$ and $\hat{F}_2$, which can be approximated by the monte carlo integration of the bootstrap estimates

   $$\hat{se}(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}^*_i - \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^*_i\right)^2}$$

   The approximate 95% Normal confidence interval for θ is

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

   The `R` code is as follows.

```
clouds<- read.table("clouds.dat",header=TRUE)
unseeded<- clouds$Unseeded
seeded<- clouds$Seeded
n <- nrow(clouds)
theta_hat<- median(seeded)- median(unseeded)
# Boostrap
B <- 1000
bootstrap_unseeded <- sapply(1:B, FUN=function(i)
    sample(unseeded, n, replace=TRUE))
bootstrap_seeded <- sapply(1:B, FUN=function(i)
    sample(seeded, n, replace=TRUE))
bootstrap_theta_hat <- sapply(1:B, FUN=function(i)
    median(bootstrap_seeded[,i]) - median(bootstrap_unseeded[,i]))
se_hat<- sqrt(var(bootstrap_theta_hat)*(B-1)/B)
# or se_hat <- sd(bootstrap_theta_hat) without correction
print(se_hat)
# CI
CI <-c(theta_hat+qnorm(0.025)*se_hat, theta_hat-qnorm(0.025)*se_hat)
print(CI)
```

The plug-in estimate of θ is 177.4. The estimated standard error of $\hat{\theta}$ is 60.86. The 95% Normal CI is [58.12, 296.68].

Note: For this problem, the results calculated from the bootstrap may differ from the results presented here due to random variation.

2. Let $X_1, \ldots, X_n$ be distinct observations (no ties). Let $X_1^*, \ldots, X_n^*$ denote a bootstrap sample (a sample from the empirical CDF), and let $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$. Find: $E(\bar{X}_n^*|X_1, \ldots, X_n)$, $V(\bar{X}_n^*|X_1, \ldots, X_n)$, $E(\bar{X}_n^*)$, and $V(\bar{X}_n^*)$.

   Solution: Suppose $X_1, \ldots, X_n$ are IID, $E(X_i) = \mu$, and $V(X_i) = \sigma^2$.

$$\begin{aligned} E(\bar{X}_n^*|X_1, \ldots, X_n) &= E\left[\frac{1}{n}\sum_{i=1}^n X_i^*|X_1, \ldots, X_n\right] \\ &= E\left[X_1^*|X_1, \ldots, X_n\right] = E_{\hat{F}_n}[X_1^*] \\ &= \sum_{i=1}^n X_i P(X_1^* = X_i) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n. \end{aligned}$$

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

$$
E(\bar{X}_n^*) = E(E(\bar{X}_n^*|X_1,\ldots,X_n)) = E(\bar{X}_n) = \mu.
$$

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

3. The file `bigcity.dat` on bCourse contains populations in thousands for $n = 49$ U.S. cities in 1920 (labeled u) and 1930 (labeled x). Demographers are interested in estimating $\theta = E_F[X]/E_F[U]$, where F represents the joint distribution of X and U. Calculate the plug-in estimate of θ and use the bootstrap to estimate the standard error and construct a 95% bootstrap pivotal interval. Hint: To sample once from $\hat{F}$, use something like

```
index <- sample(1:n, n, replace = TRUE)
u.star <- cities$u[index]
x.star <- cities$x[index]
```

Solutions:

$E_F[X] = \int\int x dF(u,x)$, so $E_{\hat{F}}[X] = \sum_{i=1}^n X_i P(U = U_i, X = X_i) = \frac{1}{n}\sum_{i=1}^n X_i$, and likewise for $E_F[U]$.

The plug-in estimate of $\theta = E_F[X]/E_F[U]$ is

$$\hat{\theta} = \frac{E_{\hat{F}}[X]}{E_{\hat{F}}[U]} = \frac{\bar{X}_n}{\bar{U}_n}.$$

Since X_i and U_i are about the same city, X and U are paired data: $(X_1, U_1), \ldots, (X_n, U_n)$. Therefore, the bootstrap sampling should be on the paired data. Denote the ith ($i = 1, \ldots, B$) bootstrap sample as

$$(X^*_{1,i}, U^*_{1,i}), \ldots, (X^*_{n,i}, U^*_{n,i}).$$

The plug-in estimate for the ith bootstrap sample is

$$\hat{\theta}^*_i = \frac{\bar{X}^*_{n,i}}{\bar{U}^*_{n,i}}.$$

The standard error of $\hat{\theta}$ can be estimated by

$$\hat{se}(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}^*_i - \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^*_i\right)^2}$$

Let $\hat{\theta}^*_{(0.025)}$ and $\hat{\theta}^*_{(0.975)}$ be the 2.5%th and 97.5%th quantiles of $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$. The 95% bootstrap pivotal interval is

$$[2\hat{\theta} - \hat{\theta}^*_{(0.975)}, 2\hat{\theta} - \hat{\theta}^*_{(0.025)}].$$

The `R` code is as follows.

```
cities <- read.table("bigcity.dat", header=TRUE)
n <- 49
theta_hat <- mean(cities$x)/mean(cities$u)
print(theta_hat)
# Bootstrap
B <- 1000
indices <- sapply(1:B, FUN=function(i) sample(1:n, n, replace = TRUE))
u.star <- apply(indices, MARGIN=2, FUN=function(col) cities$u[col])
x.star <- apply(indices, MARGIN=2, FUN=function(col) cities$x[col])
bootstrap_theta_hat <- colMeans(x.star)/colMeans(u.star)
```

```
se_hat<- sqrt(var(bootstrap_theta_hat)*(B-1)/B)
print(se_hat)
# CI
CI <- 2*theta_hat - rev(quantile(bootstrap_theta_hat, probs=c(0.025, 0.975)))
names(CI) <- c("2.5%", "97.5%")
print(CI)
```

The plug-in estimate of θ is $\hat{\theta} = 1.24$. The estimated standard error of $\hat{\theta}$ is 0.036. The 95% bootstrap pivotal interval is [1.161, 1.299].

4. Let $X_1, \ldots, X_n \overset{iid}{\sim} Unif(a, b)$, where a and b are unknown parameters and $a < b$.

   (a) Find the method of moments estimators for a and b.

   (b) Find the MLE $\hat{a}$ and $\hat{b}$.

   Solutions:

   (a) The PDF of $X \sim Unif(a, b)$ is

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}.$$

   The first and second moments are

$$E[X] = \int_a^b x f(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}. \tag{1}$$

$$E[X^2] = \int_a^b x^2 f(x) dx = \int_a^b x^2 \frac{1}{b-a} dx = \frac{a^2 + ab + b^2}{3}. \tag{2}$$

   By solving (1) and (2), we have

$$a = E[X] - \sqrt{3(E[X^2] - (E[X])^2)};$$
$$b = E[X] + \sqrt{3(E[X^2] - (E[X])^2)}.$$

   Hence, the method of moments estimators for a and b are

$$\hat{a} = \bar{X}_n - \sqrt{3(\bar{X_n^2} - (\bar{X}_n)^2)};$$
$$\hat{b} = \bar{X}_n + \sqrt{3(\bar{X_n^2} - (\bar{X}_n)^2)}.$$

(b) Order $X_1, \ldots, X_n$ as $X_{(1)} \leq \ldots \leq X_{(n)}$. Note that when $X_{(1)} < a$ or $X_{(n)} > b$, the likelihood is 0.

When $a \leq X_{(1)} \leq X_{(n)} \leq b$, the likelihood is

$$L_n(a, b) = \prod_{i=1}^{n} f(X_i) = \frac{1}{(b-a)^n},$$

which is a strictly decreasing function for b, and a strictly increasing function for a.

Therefore, $L_n(a, b)$ is maximized by $a = X_{(1)}$ and $b = X_{(n)}$. The MLEs are

$$\hat{a} = X_{(1)}, \quad \hat{b} = X_{(n)}.$$

5. Let $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$. Find the MLE for λ and an estimated standard error.

Solutions:

Since $X \sim Poisson(\lambda)$, the mass function of X is

$$f(k) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

The likelihood is

$$L_n(\lambda) = \prod_{i=1}^{n} f(X_i) = \frac{\lambda^{\sum_{i=1}^{n} X_i} e^{-n\lambda}}{X_1! \ldots X_n!}.$$

The log-likelihood is

$$\ell_n(\lambda) = \log L_n(\lambda) = (\sum_{i=1}^{n} X_i) \log \lambda - n\lambda - \log(X_1! \ldots X_n!).$$

By taking the first derivative of $\ell_n(\lambda)$ and set it as 0, we have

$$\frac{d}{d\lambda} \ell_n(\lambda) = \frac{\sum_{i=1}^{n} X_i}{\lambda} - n = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}_n.$$

Since $\lambda > 0$ and

$$\frac{d^2}{d\lambda^2} \ell_n(\lambda) = -\frac{\sum_{i=1}^{n} X_i}{\lambda^2} < 0 \quad \text{for all } \lambda > 0,$$

$\hat{\lambda} = \bar{X}_n$ maximizes $\ell_n(\lambda)$ and is the MLE for λ.

Also,

$$V(\hat{\lambda}) = V(\bar{X}_n) = \frac{V(X_1)}{n} = \frac{\lambda}{n} \quad \Rightarrow \quad se(\hat{\lambda}) = \sqrt{\frac{\lambda}{n}}.$$

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

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

6. Let $X_1, \ldots, X_n$ be *iid* with PDF $f(x; \theta) = 1/\theta$ for $0 \leq x \leq \theta$ and $\theta > 0$. Estimate θ using both the method of moments and maximum likelihood. Calculate the mean squared error for each estimator. Which one should be preferred and why?

   Solutions:

   The first moment of X is

$$E[X] = \int_0^\theta x(1/\theta)dx = \frac{\theta}{2}.$$

$$\Rightarrow \theta = 2E[X].$$

   So the method of moments estimator for θ is

$$\hat{\theta}_{MoM} = 2\bar{X}_n.$$

   Order $X_1, \ldots, X_n$ as $X_{(1)} \leq \ldots \leq X_{(n)}$. Note that when $X_{(n)} > \theta$, the likelihood is 0.

   When $X_{(n)} \leq \theta$, the likelihood is

$$L_n(\theta) = \prod_{i=1}^n f(X_i) = \frac{1}{\theta^n},$$

   which is a strictly decreasing function for θ.

   Therefore, $L_n(\theta)$ is maximized by $\theta = X_{(n)}$. The MLE for θ is

$$\hat{\theta}_{MLE} = X_{(n)}.$$

From Problem 1 in Homework 2, we have calculated that

$$MSE(\hat{\theta}_{MoM}) = \frac{\theta^2}{3n}; \quad MSE(\hat{\theta}_{MLE}) = \frac{2\theta^2}{(n+1)(n+2)}.$$

Since $MSE(\hat{\theta}_{MoM}) = MSE(\hat{\theta}_{MLE})$ for $n = 1, 2$, and $MSE(\hat{\theta}_{MoM}) > MSE(\hat{\theta}_{MLE})$ for $n \geq 3$, $\hat{\theta}_{MLE}$ is preferred.

7. Let $X_1, \ldots, X_n$ be *iid* with common distribution

$$P(X_i \leq x | \alpha, \beta) = \begin{cases} 0 & x < 0 \\ (x/\beta)^\alpha & 0 \leq x \leq \beta \\ 1 & x > \beta \end{cases}$$

(a) Find the MLEs for α and β.

(b) The length (in millimeters) of cuckoo's eggs found in hedge sparrow nests can be modeled with this distribution. For the data

$$22.0, 23.9, 20.9, 23.8, 25.0, 24.0, 21.7, 23.8, 22.8, 23.1, 23.1, 23.5, 23.0, 23.0$$

find the MLEs of α and β.

Solution:

(a) The PDF of X_i is

$$f(x) = \frac{d}{dx} P(X_i \leq x) = \frac{\alpha}{\beta^\alpha} x^{\alpha - 1}, \quad 0 \leq x \leq \beta.$$

Since $f(x) > 0, 0 \leq x \leq \beta$, we know that $\alpha > 0$.
The likelihood of α and β is

$$L_n(\alpha, \beta) = \prod_{i=1}^n f(X_i) = \frac{\alpha^n}{\beta^{\alpha n}} (X_1 \ldots X_n)^{\alpha - 1}.$$

The log-likelihood is

$$\ell_n(\alpha, \beta) = n \log \alpha - \alpha n \log \beta + (\alpha - 1) \sum_{i=1}^n \log X_i.$$

Order $X_1, \dots, X_n$ as $X_{(1)} \le \dots \le X_{(n)}$. Note that when $\beta < X_{(n)}$, $L_n(\alpha, \beta) = 0$.

When $\beta \ge X_{(n)}$, $\ell_n(\alpha, \beta)$ is a strictly decreasing function in β. Therefore, the MLE for β is

$$\hat{\beta} = X_{(n)}.$$

The first-order partial derivative with regard to α is

$$\frac{\partial}{\partial \alpha} \ell_n(\alpha, \beta) = \frac{n}{\alpha} - n \log \beta + \sum_{i=1}^{n} \log X_i.$$

Plug in $\hat{\beta} = X_{(n)}$ and solve $\frac{\partial}{\partial \alpha} \ell_n(\alpha, \hat{\beta}) = 0$, we have

$$\hat{\alpha} = \frac{n}{n \log X_{(n)} - \sum_{i=1}^{n} \log X_i} > 0.$$

For all $\alpha > 0$, the second-order partial derivative

$$\frac{\partial^2}{\partial \alpha^2} \ell_n(\alpha, \beta) = -\frac{n}{\alpha^2} < 0.$$

Therefore, $\hat{\alpha}$ maximizes $\ell_n(\alpha, \hat{\beta})$

(b) By using the following `R` code, we can find the MLEs of α and β.

```
X <- c(22.0, 23.9, 20.9, 23.8, 25.0, 24.0, 21.7, 23.8, 22.8, 23.1,
    23.1, 23.5, 23.0, 23.0)
n <- length(X)
beta_hat <- max(X)
print(beta_hat)
alpha_hat <- n/(n*log(max(X)) - sum(log(X)))
print(alpha_hat)
```

The results are $\hat{\alpha} = 12.59$, and $\hat{\beta} = 25$.

8. Consider the Normal linear regression model

$$Y_i \overset{indep}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2), \quad i = 1, \dots, n$$

Find the MLEs for β_0, β_1, and σ^2.

Solutions:

In this Normal linear regression model, we may assume that X_i's are fixed values, and write the PDF of Y_i as

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right).$$

Then the likelihood is

$$L_n(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(Y_i) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right).$$

The log-likelihood is

$$\ell_n(\beta_0, \beta_1, \sigma^2) = -n\log(\sqrt{2\pi}) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}.$$

By taking the first-order partial derivatives with regard to β_0, β_1 and σ^2 and setting them as 0, respectively, we have

$$\frac{\partial}{\partial \beta_0}\ell_n(\beta_0, \beta_1, \sigma^2) = \frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)}{\sigma^2} = 0;$$

$$\frac{\partial}{\partial \beta_1}\ell_n(\beta_0, \beta_1, \sigma^2) = \frac{\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i)}{\sigma^2} = 0;$$

$$\frac{\partial}{\partial \sigma^2}\ell_n(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^4} = 0.$$

By solving the first two equations, we have

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}_n \bar{Y}_n}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2};$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n.$$

By solving the third equation, we have

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n}.$$

By checking the second-order partial derivatives of $\ell_n(\beta_0, \beta_1, \sigma^2)$, we can show that the above are MLEs of β_0, β_1 and σ^2.