Statistical Interference
Set 3

Caltech IDS/ACM/CS 157
April 2022

Dallas Taylor

---

1 Problem 1

Let $X_1, ..., X_n \sim F \in \mathcal{F}$, and let $\hat{F}_n$ be the empirical CDF. Find the covariance between two random variables $\hat{F}_n(x)$ and $\hat{F}_n(y)$ for $x \neq y$.

Solution : *We can assume that all X_i are i.i.d. for all i. We also know that the covariance between two variables is given as the following:*

$$\begin{aligned}
\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) &= \mathbb{E}[\hat{F}_n(x) \cdot \hat{F}_n(y)] - \mathbb{E}[\hat{F}_n(x)]\mathbb{E}[\hat{F}_n(y)] \\
\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) &= \mathbb{E}[\hat{F}_n(x) \cdot \hat{F}_n(y)] - F(x)F(y) \\
\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} H(x - X_i) \cdot \frac{1}{n}\sum_{j=1}^{n} H(y - X_j)\right] - F(x)F(y) \\
\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) &= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} H(x - X_i)H(y - X_j)\right] - F(x)F(y) \\
\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) &= \mathbb{E}\left[\frac{1}{n^2}(\sum_{i=1}^{n}\sum_{j\neq 1}^{n} H(x - X_i)H(y - X_j) + \sum_{i=1}^{n} H(x - X_i)H(y - X_i))\right] - F(x)F(y)
\end{aligned}$$

The above is trivially equal to the following, since the $H(x) = 0$ if $x \leq 0$ or $H(x) = 1$ if $x > 0$. This is also from when we have $i \neq j$, and are thus considering two independent X_i, X_j.

$$\begin{aligned}
\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq 1}^{n} F(x)F(y) + \frac{1}{n^2}\sum_{i=1}^{n} F(\min\{x, y\}) - F(x)F(y) \\
\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) &= \frac{1}{n^2}(n)(n-1)F(x)F(y) + \frac{1}{n^2}(n)F(\min\{x, y\}) - F(x)F(y) \\
\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) &= \frac{n-1}{n}F(x)F(y) + \frac{1}{n}F(\min\{x, y\}) - F(x)F(y) \\
\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) &= \frac{1}{n}\left(F(\min\{x, y\}) - F(x)F(y)\right)
\end{aligned}$$

Dallas Taylor

2 Problem 2

The skewness is a parameter that measures the lack of symmetry of a distribution. It is defined as follows:

$$\kappa_F = \frac{\int (x-\mu_F)^3 dF(x)}{(\int (x-\mu_F)^2 dF(x))^{3/2}}$$

Find the plug-in estimate of κ_F.

Solution : *We want to find $\hat{\kappa}_F = t(\hat{F}_n)$. Thus, we perform the following:*

$$\begin{aligned}
\hat{\kappa}_F &= \frac{\int (x-\mu_F)^3 d\hat{F}_n(x)}{(\int (x-\mu_F)^2 d\hat{F}_n(x))^{3/2}} \\
\hat{\kappa}_F &= \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_F)^3}{(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_F)^2)^{3/2}}\right) \\
\hat{\kappa}_F &= \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^3}{(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2)^{3/2}}\right) \\
\hat{\kappa}_F &= \left(\frac{1}{n}\right)\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^3}{(\hat{\sigma}_n^2)^{3/2}}\right) \\
\hat{\kappa}_F &= \frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^3}{n\hat{\sigma}_n^3}
\end{aligned}$$

Dallas Taylor

3 Problem 3

Solution : *See attached scripts.*

Dallas Taylor

4 Problem 4

Let $X_1, ..., X_n$ be data, and suppose that we know that this is a sample from the uniform distribution $\mathcal{U}[0, \theta]$, but we don't know θ.

Hint: If $X_1, ..., X_n \sim \mathcal{U}[0, \theta]$ and $X_{(k)}$ is the k-th order statistic, then

$$\mathbb{E}[X_{(k)}] = \frac{k\theta}{n+1}.$$

Problem A: Find the plug-in estimate $\hat{\theta}_n$ of θ using the following representation of θ:

$$\theta = \min\{x : F(x) = 1\}.$$

Solution A: *We know that if we have $\theta = t(F)$, that $\hat{\theta}_n = t(\hat{F}_n)$. We also know that $\hat{F}_n = \frac{1}{n}\sum_{i=1}^{n} H(x - X_i)$ and thus, in order for $\hat{F}_n(x)$ to be closest to 1, we must have $x - X_i > 0$ for as many i as possible. Thus, we can clearly determine that $\hat{\theta}_n = X_{(n)}$ where $X_{(n)} = \max\{X_1, ..., X_n\}$ from the given representation of θ. Thus,*

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

Problem B: Find the bias of $\hat{\theta}_n$.

Solution B: *Let us compute the following:*

$$\begin{aligned}
\mathbb{B}[\hat{\theta}_n] &= \mathbb{E}[\hat{\theta}_n] - \theta \\
\mathbb{B}[\hat{\theta}_n] &= \mathbb{E}[X_{(n)}] - \theta \\
\mathbb{B}[\hat{\theta}_n] &= \frac{n\theta}{n+1} - \theta \\
\mathbb{B}[\hat{\theta}_n] &= \frac{-\theta}{n+1}
\end{aligned}$$

Problem C: Find the bias-corrected estimate $\hat{\theta}_n^J$ using the jackknife method.

Solution C: *We know that* $\hat{\theta}_n^J = n\hat{\theta}_n - (n-1)\bar{\theta}_n^J$ *and that* θ_n^{-J} *is constructed by creating* n *samples with one element removed:*

$$\hat{\theta}_n^{(-i)} = X_{(n)}^{(-i)} = \max\{X_1, ..., X_{i-1}, X_{i+1}, ..., X_n\}$$

Thus, it is clear that we have $\bar{\theta}_n^J = \frac{1}{n}\sum_{i=1}^n \hat{\theta}_n^{(-i)} = \frac{1}{n}\left((n-1)X_{(n)} + X_{(n-1)}\right)$. *Therefore, we have:*

$$\begin{aligned}
\hat{\theta}_n^J &= n\hat{\theta}_n - (n-1)\left(\frac{(n-1)X_{(n)}}{n} + \frac{X_{(n-1)}}{n}\right) \\
\hat{\theta}_n^J &= nX_{(n)} - \frac{(n-1)^2 X_{(n)}}{n} - \frac{(n-1)X_{(n-1)}}{n} \\
\hat{\theta}_n^J &= \frac{(2n-1)X_{(n)}}{n} - \frac{(n-1)X_{(n-1)}}{n}
\end{aligned}$$

Problem D: Find the bias of $\hat{\theta}_n^J$.

Solution D: *We know that* $\hat{\mathbb{B}}_J[\hat{\theta}_n] = (n-1)(\bar{\theta}_n^J - \hat{\theta}_n)$ *and* $\bar{\theta}_n^J = \frac{n-1}{n}X_{(n)} + \frac{1}{n}X_{(n-1)}$. *Thus, we have:*

$$\begin{aligned}
\mathbb{B}[\hat{\theta}_n^J] &= \mathbb{E}[\hat{\theta}_n] - \mathbb{E}[\hat{\mathbb{B}}_J[\hat{\theta}_n]] - \theta \\
\mathbb{B}[\hat{\theta}_n^J] &= \mathbb{E}[X_{(n)}] - \mathbb{E}[(n-1)(\bar{\theta}_n^J - \hat{\theta}_n)] - \theta \\
\mathbb{B}[\hat{\theta}_n^J] &= \frac{n\theta}{n+1} - \left((n-1)(\frac{n-1}{n}\mathbb{E}[X_{(n)}] + \frac{1}{n}\mathbb{E}[X_{(n-1)}] - \mathbb{E}[X_{(n)}])\right) - \theta \\
\mathbb{B}[\hat{\theta}_n^J] &= \frac{n\theta}{n+1} - \left((n-1)(\frac{-1}{n}\mathbb{E}[X_{(n)}] + \frac{1}{n}\mathbb{E}[X_{(n-1)}])\right) - \theta \\
\mathbb{B}[\hat{\theta}_n^J] &= \frac{n\theta}{n+1} - \left(\frac{-n}{n}\mathbb{E}[X_{(n)}] + \frac{n}{n}\mathbb{E}[X_{(n-1)}] + \frac{1}{n}\mathbb{E}[X_{(n)}] - \frac{1}{n}\mathbb{E}[X_{(n-1)}]\right) - \theta \\
\mathbb{B}[\hat{\theta}_n^J] &= \frac{n\theta}{n+1} - \left(\frac{1-n}{n}\mathbb{E}[X_{(n)}] + \frac{n-1}{n}\mathbb{E}[X_{(n-1)}]\right) - \theta \\
\mathbb{B}[\hat{\theta}_n^J] &= \frac{n\theta}{n+1} + \frac{n-1}{n}\left(\frac{n\theta}{n+1}\right) - \frac{n-1}{n}\left(\frac{(n-1)\theta}{n+1}\right) - \theta \\
\mathbb{B}[\hat{\theta}_n^J] &= \frac{n\theta}{n+1} + \frac{(n-1)\theta}{n+1} - \frac{(n-1)^2\theta}{n(n+1)} - \theta \\
\mathbb{B}[\hat{\theta}_n^J] &= \frac{(n^2 + n^2 - n - n^2 + 2n - 1 - n^2 - n)\theta}{n(n+1)} \\
\mathbb{B}[\hat{\theta}_n^J] &= \frac{-\theta}{n(n+1)}
\end{aligned}$$

5 Problem 5

Let us now implement the jackknife method. Let $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$, with $\sigma^2 = 1$. Suppose that the parameter of interest is $\theta = e^{\mu}$. The plug-in estimate of θ is $\hat{\theta}_n = e^{\bar{X}_n}$. It is biased. Our goal is to reduce the bias by jackknifing $\hat{\theta}_n$.

Hint: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = e^X$ follows the log-normal distribution, $Y \sim \ln\mathcal{N}(\mu, \sigma^2)$. In particular, $\mathbb{E}[Y] = e^{\mu + \frac{\sigma^2}{2}}$

Problem A: Recall that the jackknife assumes

$$\mathbb{B}[\hat{\theta}_n] = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right) \quad \text{as } n \to \infty$$

Check this assumption for $\hat{\theta}_n = e^{\bar{X}_n}$.

Solution A: *Here, we know that the sample mean $\bar{X}_n$ has mean μ and variance $\frac{\sigma^2}{n}$. Thus, we have the following:*

$$\begin{aligned} \mathbb{B}[\hat{\theta}_n] &= \mathbb{E}[\hat{\theta}_n] - \theta \\ \mathbb{B}[\hat{\theta}_n] &= \mathbb{E}[e^{\bar{X}_n}] - e^{\mu} \\ \mathbb{B}[\hat{\theta}_n] &= e^{\mu}(e^{\frac{1}{2n}} - 1) \end{aligned}$$

We can utilize the Taylor expansion of $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$ to rewrite this as:

$$\begin{aligned} \mathbb{B}[\hat{\theta}_n] &= e^{\mu}(1 + \frac{1}{2n} + \frac{1}{8n^2} + O\left(\frac{1}{n^3}\right) - 1) \\ \mathbb{B}[\hat{\theta}_n] &= \frac{e^{\mu}}{2n} + \frac{e^{\mu}}{8n^2} + O\left(\frac{1}{n^3}\right) \end{aligned}$$

Thus, we clearly see that this assumptions holds when $a = \frac{e^{\mu}}{2}$ and $b = \frac{e^{\mu}}{8}$.

Problem B:

Solution B: *See attached scripts.*

Problem C:

Solution C: *See attached scripts.*

IDS/ACM 157 PS3 MatLab - Problem 3

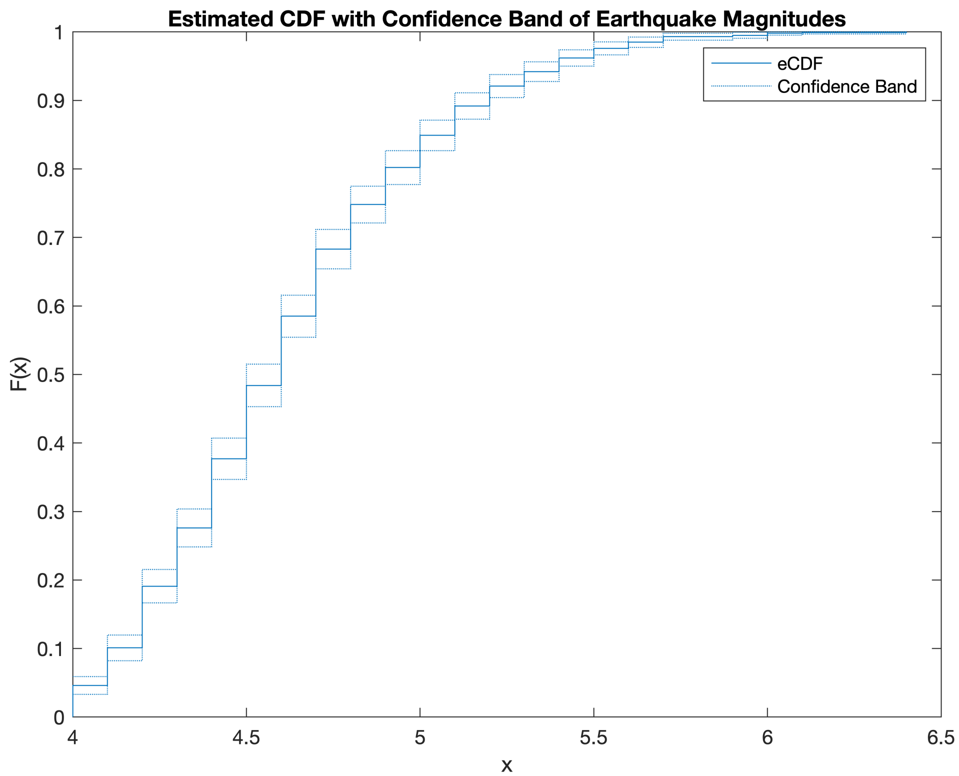
In particular, the data set contains the magnitudes of $n = 1000$ seismic events occurred near Fiji since 1964 . Estimate the CDF F of the earthquake magnitudes and construct a 95% confidence band for F.

```
fiji = readmatrix('./fiji.txt');
magnitudes = fiji(:,5);
```

Part a

Plot both the estimated CDF and the confidence band.

```
ecdf(magnitudes,'Alpha',0.05,'Bounds','on')
title('Estimated CDF with Confidence Band of Earthquake Magnitudes');
legend('eCDF','Confidence Band');
```

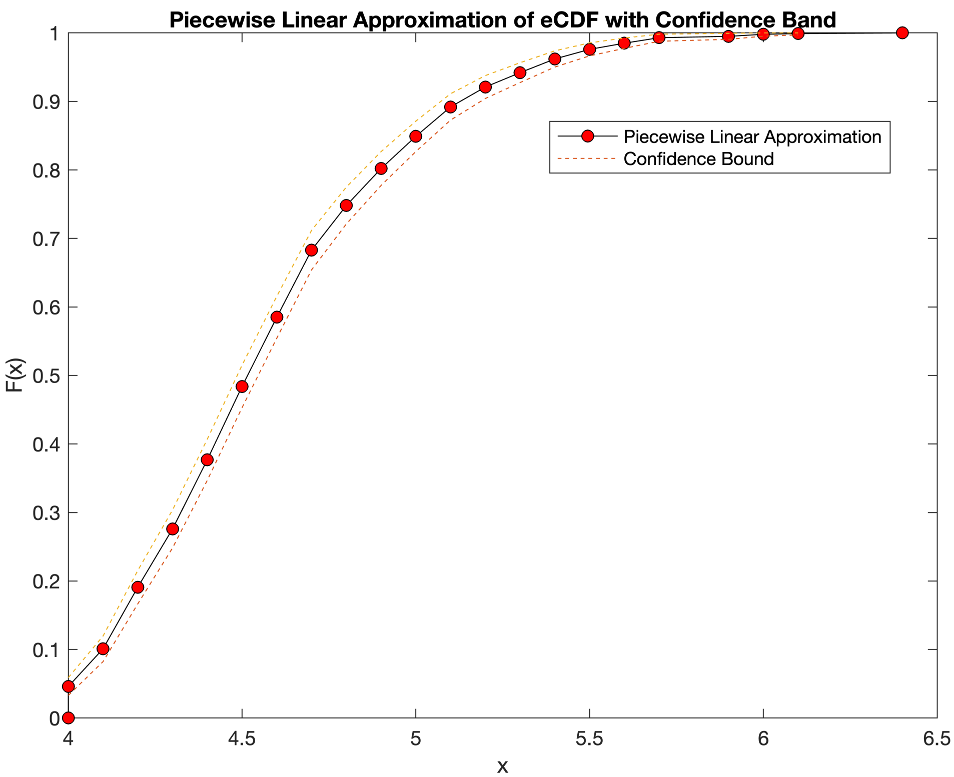


Part b

To make the figure more visually appealing (more "smooth"), instead of piecewise constant functions, use their piecewise linear approximation.

```
[f,x,cu,cl] = ecdf(magnitudes);
figure
plot(x,f,'ko-','MarkerFace','r')
hold on
plot(x,cu,'--','MarkerFace','r')
plot(x,cl,'--','MarkerFace','r')
```

```
title('Piecewise Linear Approximation of eCDF with Confidence Band')
xlabel('x');
ylabel('F(x)');
legend('Piecewise Linear Approximation','Confidence Bound', ...
    'Location','best')
```



IDS/ACM 157 PS3 MatLab - Problem 5

Part b

Generate a data set $X_1, \ldots, X_n$ using $\mu = 5$ and $n = 100$. Estimate the bias of $\hat{\theta}_n$ using the jackknife method and compare the estimated value with the exact value obtained in part (a).

```
mu = 5;
n = 100;
sigma = 1;
data = normrnd(mu,sigma,n,1);
X_bar = mean(data);

% finding actual bias
bias_a = exp(mu) * (exp(1/(2*n)) - 1);

% finding jackknife estimate
theta_j = zeros(n,1);
for i = 1:n
    X_i = data([1:i-1,i+1:end]);
    theta_j(i) = exp(mean(X_i));
end
theta_j_bar = mean(theta_j);
bias_e = (n - 1) * (theta_j_bar - exp(X_bar));

disp('Estimated bias using jackknife method:'); disp(bias_e);
```

```
Estimated bias using jackknife method:
    0.5535
```

```
disp('Actual bias:'); disp(bias_a);
```

```
Actual bias:
    0.7439
```

Part c

In this part, our goal is to experimentally observe the theoretical statement that the bias of the jackknife estimate $\hat{\theta}_n{}^J$ is smaller than the bias of $\hat{\theta}_n$. First, generate $r = 10^4$ data sets $X_1^{(j)}, \ldots, X_n^{(j)}, j = 1, \ldots, r$, as in (b). For each set, compute the estimate $\hat{\theta}_n{}^{(j)}$ and the corresponding bias-corrected estimate $\hat{\theta}_n{}^{(j),J}$. Estimate the biases of $\hat{\theta}_n$ and $\hat{\theta}_n{}^J$ as follows:

$$\mathrm{B}\left[\hat{\theta}_n\right] \approx B_1 = \frac{1}{r} \sum_{j=1} \hat{\theta}_n{}^{(j)} - \theta,$$

$$\mathrm{B}\left[\hat{\theta}_n{}^J\right] \approx B_2 = \frac{1}{r} \sum_{j=1} \hat{\theta}_n{}^{(j),J} - \theta$$

Compute both B_1 and B_2. We expect B_1 to be approximately equal to the exact value given by (4), and to be larger (in absolute value) than B_2.

```
r = 10^4;

theta_j = zeros(r,1);
theta_jJ = zeros(r,1);
for s = 1:r
    data = normrnd(mu,sigma,n,1);
    theta_j(s) = exp(mean(data));
    theta_jJ_s = zeros(n,1);
    for i = 1:n
        X_i = data([1:i-1,i+1:end]);
        theta_jJ_s(i) = exp(mean(X_i));
    end
    theta_jJ(s) = n * exp(mean(data)) - (n-1) * mean(theta_jJ_s);
end

B1 = mean(theta_j) - exp(mu);
B2 = mean(theta_jJ) - exp(mu);
disp('B1:'); disp(B1);
```

```
B1:
    0.7337
```

```
disp('B2:'); disp(B2);
```

```
B2:
   -0.0120
```