Dallas Taylor

1 Problem 1

Solution : *See attached scripts.*

2 Problem 2

Let $X_1, ..., X_n \sim \mathcal{U}[\alpha, \beta]$, where α and β are unknown parameters, $\alpha < \beta$, and $\mathcal{U}[\alpha, \beta]$ is the uniform distribution on $[\alpha, \beta]$.

Problem A: Find the method of moments estimates of α and β.

Solution A: *We want to begin by finding two moments since we have two parameters to solve for:*

$$
\begin{aligned}
m_1(\alpha,\beta) &= \int x f(x;(\alpha,\beta))dx \\
m_1(\alpha,\beta) &= \int \frac{xdx}{\beta-\alpha} = \left[\frac{x^2}{2(\beta-\alpha)}\right]_\alpha^\beta = \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} = \frac{(\beta+\alpha)(\beta-\alpha)}{2(\beta-\alpha)} = \frac{\beta+\alpha}{2} \\
m_2(\alpha,\beta) &= \int x^2 f(x;(\alpha,\beta))dx \\
m_2(\alpha,\beta) &= \int \frac{x^2dx}{\beta-\alpha} = \left[\frac{x^3}{3(\beta-\alpha)}\right]_\alpha^\beta = \frac{\beta^3-\alpha^3}{3(\beta-\alpha)} = \frac{(\beta-\alpha)(\beta^2+\alpha\beta+\alpha^2)}{3(\beta-\alpha)} = \frac{\beta^2+\alpha\beta+\alpha^2}{3}
\end{aligned}
$$

From here, we are able to solve for α, β:

$$
\begin{aligned}
m_1(\alpha,\beta) = \hat{m}_1 &= \frac{\beta+\alpha}{2} \\
2\hat{m}_1 - \beta &= \alpha \\
m_2(\alpha,\beta) = \hat{m}_2 &= \frac{\beta^2+\alpha\beta+\alpha^2}{3} \\
3\hat{m}_2 &= \beta^2+\alpha\beta+\alpha^2 \\
\therefore 3\hat{m}_2 &= \beta^2 + (2\hat{m}_1-\beta)\beta + (2\hat{m}_1-\beta)^2 \\
3\hat{m}_2 &= \beta^2 + 2\beta\hat{m}_1 - \beta^2 + 4\hat{m}_1^2 - 4\beta\hat{m}_1 + \beta^2 \\
3\hat{m}_2 &= \beta^2 + 4\hat{m}_1^2 - 2\beta\hat{m}_1 \\
3\hat{m}_2 - 3\hat{m}_1^2 &= \beta^2 - 2\beta\hat{m}_1 + \hat{m}_1^2 \\
\therefore \beta = \hat{m}_1 + \sqrt{3(\hat{m}_2-\hat{m}_1^2)}, &\qquad \alpha = \hat{m}_1 - \sqrt{3(\hat{m}_2-\hat{m}_1^2)}
\end{aligned}
$$

Dallas Taylor

Problem B: Find the MLEs of α and β.

Solution B: *We can find the MLE of α, β to be the following:*

$$\begin{aligned}
\mathcal{L}_n((\alpha,\beta) \mid X_1, ..., X_n) &= \prod_{i=1}^{n} f(X_i; (\alpha,\beta)) \\
\mathcal{L}_n &= \prod_{i=1}^{n} \frac{1}{\beta-\alpha} = \frac{1}{(\beta-\alpha)^n} \\
\therefore \hat{\alpha}_{MLE} &= \arg\max_{\alpha \in \{X_1,...,X_n\}} \frac{1}{(\beta-\alpha)^n} \\
\therefore \hat{\beta}_{MLE} &= \arg\max_{\beta \in \{X_1,...,X_n\}} \frac{1}{(\beta-\alpha)^n}
\end{aligned}$$

From this, we can clearly determine that the maximizing argument for α is when we have the smallest value in our sample. Similarly, the maximizing argument for β is when we have the largest value in our sample. Thus, we produce the following:

$$\begin{aligned}
\hat{\alpha}_{MLE} &= X_{(1)} \\
\hat{\beta}_{MLE} &= X_{(n)}
\end{aligned}$$

Dallas Taylor

3 Problem 3

Let, as in Problem 2, $X_1, ..., X_n \sim \mathcal{U}[\alpha, \beta]$. Let the parameter of interest be the mean of the distribution, $\mu = \int x dF(x)$.

Problem A: Find the MLE (maximum likelihood estimate) of μ.

Solution A: *It is clear that we have $\mu = \frac{\beta+\alpha}{2}$ since we are considering a uniform distribution. Thus, we can find the MLE of μ to be the following:*

$$\hat{\mu}_{MLE} = \mu(\hat{\alpha}_{MLE}, \hat{\beta}_{MLE})$$
$$\hat{\mu}_{MLE} = \frac{\hat{\beta}_{MLE} + \hat{\alpha}_{MLE}}{2}$$

Problem B: Let $\hat{\mu}_n = \bar{X}_n$ be the plug-in estimate of μ. Set $n = 10$, $\alpha = 1$, and $\beta = 3$. Find the mean squared error (MSE) of $\hat{\mu}_n$ analytically, and find the MSE of $\hat{\mu}_{\text{MLE}}$ by simulation (plain Monte Carlo, not bootstrap). Compare. Write a script that implements the task in part (b). Write you results and conclusions as comments in the script.

Solution B: *We can calculate the MSE of $\hat{\mu}_n$ to be equal to the following:*

$$\begin{aligned}
MSE[\hat{\mu}_n] &= bias[\hat{\mu}_n]^2 + se[\hat{\mu}_n]^2 \\
MSE[\hat{\mu}_n] &= (\mathbb{E}[\hat{\mu}_n] - \mu)^2 + \mathbb{V}[\hat{\mu}_n] \\
MSE[\hat{\mu}_n] &= \mathbb{V}[\hat{\mu}_n] = \frac{1}{n}\left(\mathbb{E}[X^2] - E^2[X]\right) \\
MSE[\hat{\mu}_n] &= \frac{1}{n}\left(\int_\alpha^\beta x^2 f(x)dx - \left(\int_\alpha^\beta x f(x)dx\right)^2\right) \\
MSE[\hat{\mu}_n] &= \frac{1}{n}\left(\int_\alpha^\beta \frac{x^2}{\beta-\alpha}dx - \left(\int_\alpha^\beta \frac{x}{\beta-\alpha}dx\right)^2\right) \\
MSE[\hat{\mu}_n] &= \frac{1}{n}\left(\left[\frac{x^3}{3(\beta-\alpha)}\right]_\alpha^\beta - \left(\left[\frac{x^2}{2(\beta-\alpha)}\right]_\alpha^\beta\right)^2\right) = \frac{1}{n}\left(\frac{\beta^3-\alpha^3}{3(\beta-\alpha)} - \left(\frac{\beta^2-\alpha^2}{2(\beta-\alpha)}\right)^2\right) \\
MSE[\hat{\mu}_n] &= \frac{\beta^2+\alpha\beta+\alpha^2}{3n} - \frac{\beta^2+2\alpha\beta+\alpha^2}{4n} = \frac{(\beta-\alpha)^2}{12n} = \frac{(3-1)^2}{12(10)} = \frac{1}{30}
\end{aligned}$$

See attached scripts for remaining work.

Dallas Taylor

4 Problem 4

Suppose $X_1, ..., X_n \sim \mathcal{N}(\theta, 1)$. Define

$$Y_i = \begin{cases} 1, & \text{if } X_i > 0 \\ 0, & \text{if } X_i \leq 0 \end{cases}$$

Finally, let $\psi = \mathbb{E}[Y_1]$.

Problem A: Express ψ in terms of θ and find the MLE of ψ based on the data $X_1, ..., X_n$.

Solution A: *We can express ψ in terms of θ by completing the following:*

$$\begin{aligned} \psi &= \mathbb{E}[Y_1] = \int_{-\infty}^{\infty} Y_i \mathcal{N}(\theta, 1) dY = \int_{-\infty}^{0} (0)\mathcal{N}(\theta, 1) dY + \int_{0}^{\infty} (1)\mathcal{N}(\theta, 1) dY \\ \psi &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\theta)^2}{2}\right) dx \end{aligned}$$

We know that $\hat{\theta}_{MLE} = \bar{X}_n$ from Maximum Likelihood notes. Therefore, we can find that $\hat{\psi}_{MLE}$ to be:

$$\begin{aligned} \hat{\psi}_{MLE} &= \psi(\hat{\theta}_{MLE}) = \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\theta)^2}{2}\right) dx \\ \hat{\psi}_{MLE} &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\bar{X}_n)^2}{2}\right) dx \end{aligned}$$

Problem B: Find an approximate 95% confidence interval for ψ from the data $X_1, ..., X_n$.

Solution B: *We create a 95% confidence interval for ψ in the following way:*

$$\begin{aligned} \mathcal{I}_n &= \hat{\theta}_n \mp z_{\alpha/2} se[\hat{\theta}_n] \\ \mathcal{I}_n &= \bar{X}_n \mp (-1.96)\frac{1}{\sqrt{n}} \\ \therefore \mathcal{I}_\psi &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x-\bar{X}_n \pm (-1.96)\frac{1}{\sqrt{n}})^2}{2}\right) dx \end{aligned}$$

5 Problem 5

The bootstrap method can be also used in the context of parametric models. In this case, the method is often referred to as the parametric bootstrap. There is only one change: in the nonparametric bootstrap, we generate bootstrap samples from the eCDF constructed from the data $X_1, ..., X_n$; in the parametric bootstrap, we sample from $f(x; \hat{\theta}_n)$, where $\hat{\theta}_n$ is an estimate (MLE, MoM, plug-in, etc) of the model parameter θ. In this problem, you will see the difference in the performance of parametric and nonparametric bootstraps.

Let $X_1, ..., X_n \sim \mathcal{U}[0, \theta]$. We know (Lecture 9) the MLE of θ is $\hat{\theta}_{\text{MLE}} = X_{(n)}$.

Problem A: Find the probability density function of $\hat{\theta}_{\text{MLE}}$ analytically.

Solution A: *We can complete the following to find the probability density $f(x)$ of $\hat{\theta}_{\text{MLE}}$:*

$$\begin{aligned} F(x) &= \mathbb{P}[X_{(n)} \leq x) = \prod_{i=1}^{n} \frac{x}{\theta} = \frac{x^n}{\theta^n} \\ \therefore f(x) &= \frac{d}{dx} \frac{x^n}{\theta^n} = \frac{nx^{n-1}}{\theta^n} \end{aligned}$$

Problem B:

Solution B: *See attached scripts.*

IDS/ACM 157 PS4 MatLab - Problem 1

Here your task is to implement the bootstrap method.

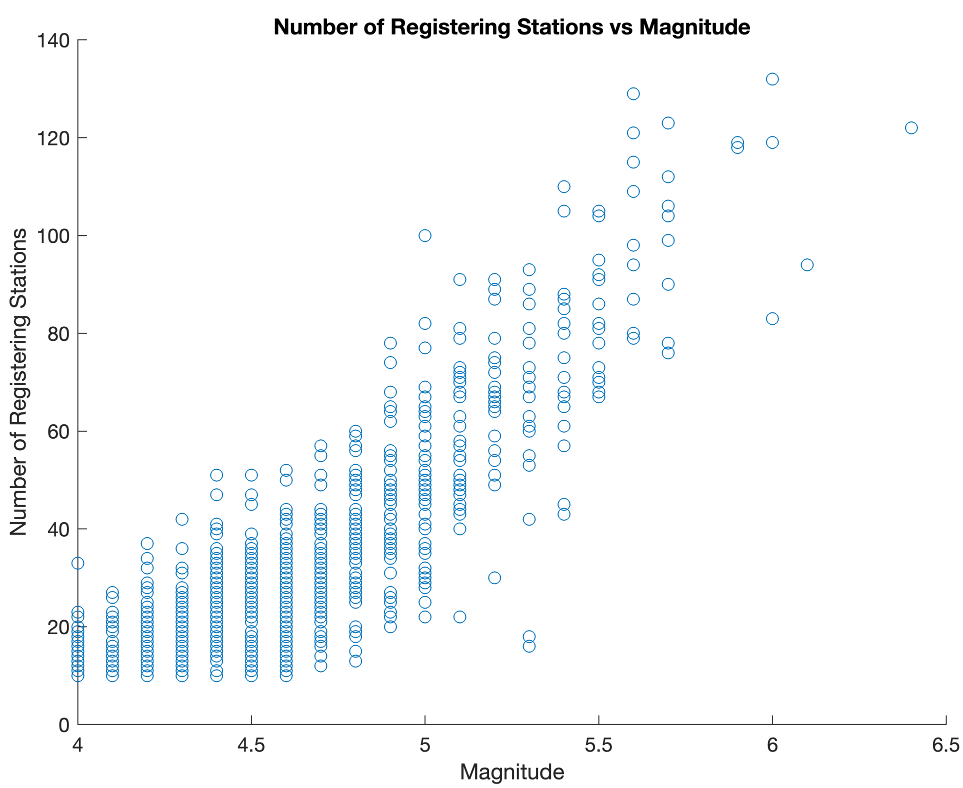
For each of the $n = 10^3$ earthquakes occurred near Fiji , the 5th and 6th columns give the magnitudes X_i and the numbers of stations Y_i registered these events, $i = 1, \ldots, n$.

```
fiji = readmatrix('./fiji.txt');
magnitudes = fiji(:,5);
stations = fiji(:,6);
```

Part a

Draw a scatter plot of Y(numbers of stations) versus X (magnitudes).

```
figure;
scatter(magnitudes,stations);
title('Number of Registering Stations vs Magnitude')
xlabel('Magnitude');
ylabel('Number of Registering Stations');
```



Part b

The scatter plot reveals some positive correlation between X and Y . This is expected since the common sense suggests that the larger the earthquake, the more stations should be able to register it. Let's estimate the correlation. So, our parameter of interest is

$$\theta = \operatorname{corr}\left(X, Y\right) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{V[X]V[Y]}}$$

$$= \frac{\int\int (x-\mu_X)(y-\mu_Y)\mathrm{dF}(x, y)}{\sqrt{\int \left(x-\mu_X\right)^2 \mathrm{dF}\left(x\right) \int \left(y-\mu_Y\right)^2 \mathrm{dF}\left(y\right)}}.$$

Find the plug-in estimate $\hat{\theta}_n$ of θ and compute its value using the data.

```
theta_hat = corr(magnitudes,stations);
disp('Plug in estimate of theta:'); disp(theta_hat);
```

```
Plug in estimate of theta:
    0.8512
```

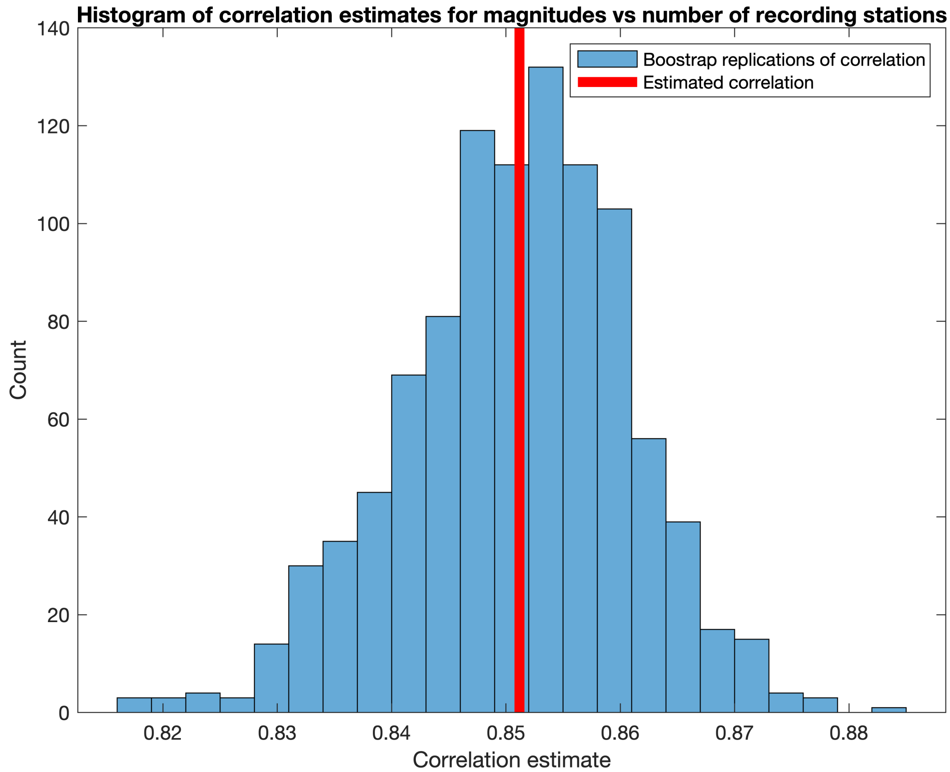
Part c

Find the bootstrap estimate of the standard error of $\hat{\theta}_n$ based on $B = 10^3$ bootstrap samples, and show a histogram of the bootstrap replications $\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(B)}$ together with the estimated value of θ.

```
B = 10^3;
n = 10^3;
theta_hats = zeros(B,1);
for b = 1:B
    data = datasample([magnitudes, stations],n);
    theta_hats(b) = corr(data(:,1),data(:,2));
end
serr = sqrt(var(theta_hats));
disp('Bootstrap estimate of standard error:'); disp(serr);
```

```
Bootstrap estimate of standard error:
    0.0099
```

```
figure;
histogram(theta_hats);
hold on
line([theta_hat, theta_hat], ylim,'LineWidth', 5, 'Color','red');
title('Histogram of correlation estimates for magnitudes vs number of recording statio
xlabel('Correlation estimate');
ylabel('Count');
legend('Boostrap replications of correlation','Estimated correlation');
```



Part d

Find the normal and pivotal 95% confidence intervals for θ.

```
z_25 = -1.96;
normal = [theta_hat + z_25 * serr,theta_hat - z_25 * serr];
pivotal = bootci(B,@corr,magnitudes,stations);
disp('Normal 95% confidence interval:'); disp(normal);
```

```
Normal 95% confidence interval:
    0.8318    0.8706
```

```
disp('Pivotal 95% confidence interval:'); disp(pivotal);
```

```
Pivotal 95% confidence interval:
    0.8298
    0.8691
```

Discussion

We saw that the histogram of theta_hat values displays an almost normal distribution, as discussed in the notes. Additionally, the range of calculated estimates is within ± 0.03, showing a tight range. For part d, we see that the normal and pivotal confidence intervals are extremely close to each other, often varing around $a = 0.83, b = 0.87$.

IDS/ACM 157 PS4 MatLab - Problem 3

Part b

Find the MSE of $\widehat{\mu_{\mathrm{MLE}}}$ by simulation (plain Monte Carlo, not bootstrap). Compare. Write a script that implements the task in part (b). Write you results and conclusions as comments in the script.

```
M = 10^4;
MSE = 0;
n = 10;
alpha = 1;
beta = 3;
mu = (beta+alpha)/2;

for m = 1:M
    data = unifrnd(alpha,beta,n,1);
    mu_m = (max(data) + min(data))/2;
    MSE = MSE + (mu_m - mu)^2;
end
MSE = MSE/M;

disp('MSE found by Monte Carlo simulation:'); disp(MSE);
```

```
MSE found by Monte Carlo simulation:
    0.0148
ans = 0.0333
```

Discussion

Here, we found that the MSE from Monte Carlo simulation was around 0.015. The MSE calculated analytically was found to be 0.0333. Thus, we find that the Monte Carlo MSE predicts an error that is lower by a factor of 2.

IDS/ACM 157 PS4 MatLab - Problem 5

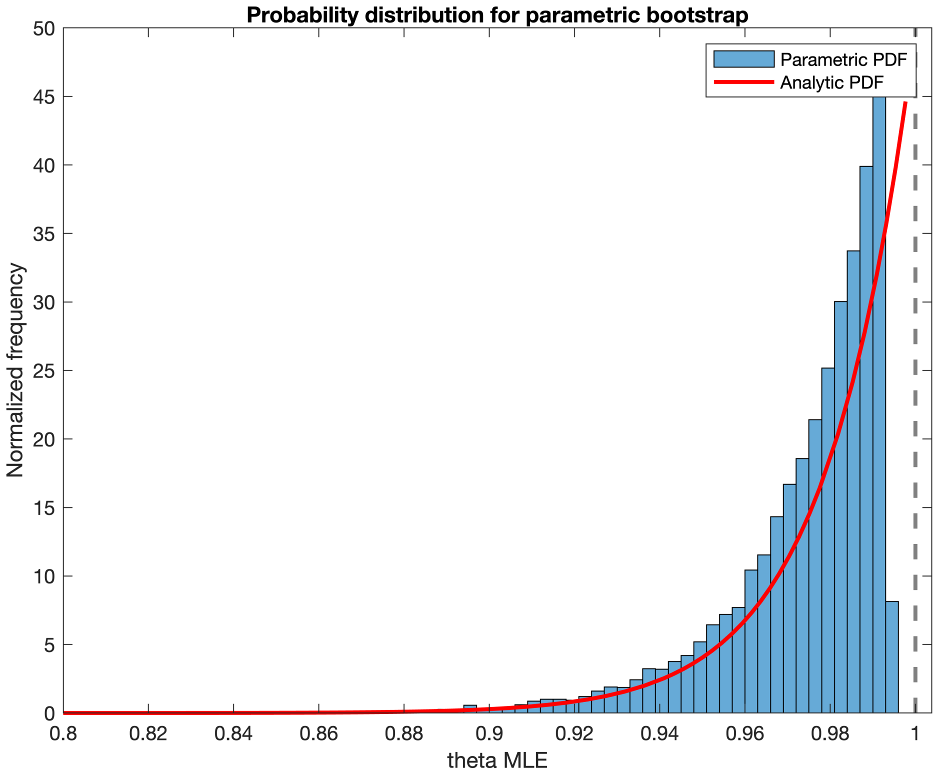
Part b

Create a sample of size $n = 50$ with $\theta = 1$. Compare the true distribution of $\widehat{\theta_{\text{MLE}}}$ from part (a) with the histograms of $B = 10^4$ parametric and nonparametric bootstrap replications of $\widehat{\theta_{\text{MLE}}}$. Write a script that implements the task in part (b). Write your results and conclusions as comments in the script.

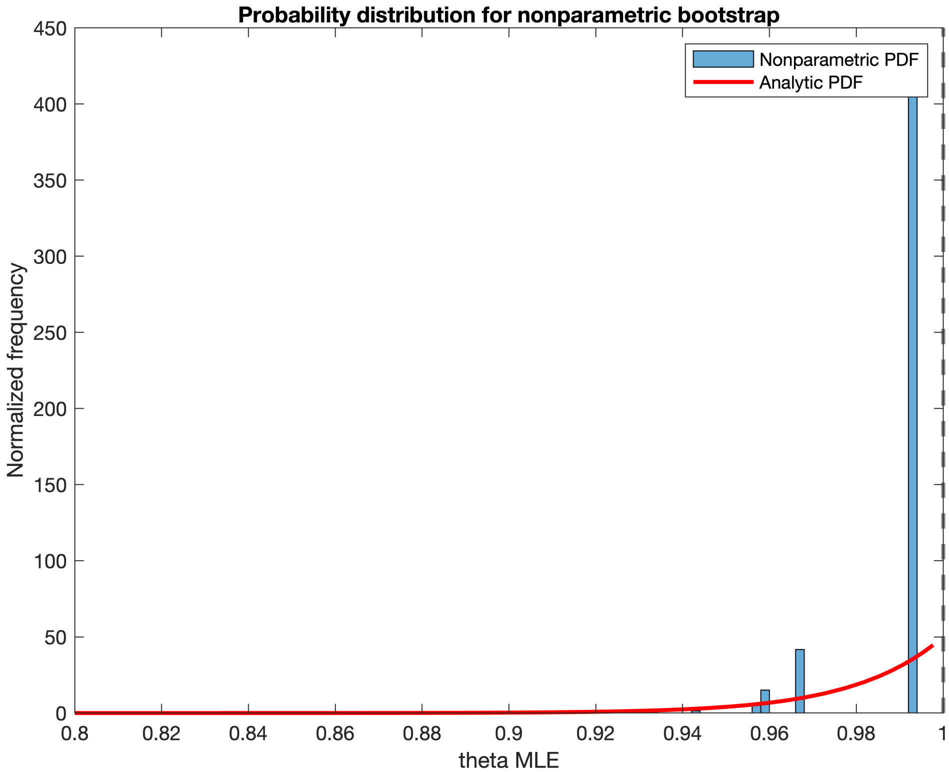
```
n = 50;
theta = 1;
B = 10^4;
sample = unifrnd(0,1,n,1);
theta_hat_MLE = max(sample);

parametric = zeros(B,1);
nonparametric = zeros(B,1);
for b = 1:B
    data1 = unifrnd(0,theta_hat_MLE,n,1);
    parametric(b) = max(data1);
    data2 = datasample(sample,n);
    nonparametric(b) = max(data2);
end

figure;
histogram(parametric,'Normalization','pdf');
hold on
fplot(@(x) (n.*x.^(n-1)),[0.8,1],'LineWidth',2,'Color','red');
title('Probability distribution for parametric bootstrap');
xlabel('theta MLE');
ylabel('Normalized frequency');
legend('Parametric PDF','Analytic PDF');
```




```
figure;
histogram(nonparametric,'Normalization','pdf');
hold on
fplot(@(x) (n.*x.^(n-1)),[0.8,1],'LineWidth',2,'Color','red');
title('Probability distribution for nonparametric bootstrap');
xlabel('theta MLE');
ylabel('Normalized frequency');
legend('Nonparametric PDF','Analytic PDF');
```



Discussion

The distributions of the parametric and nonparametric bootstrap replications both follow a similar upwards, exponential trend. However, it is clear that both bootstrap distributions often trend above the analytic PDF. Sometimes, the parametric PDF very closely matches the analytic PDF, and the general distribution of the parametric PDF is always a closer match to analytic than the nonparametric PDF. The worst variances from the analytic PDF occur when the parametric PDF is shifted left. The nonparametric PDF has very high frequencies for the values close to 1, but only a select few values and thus we have a very sparse histogram. This makes sense considering the nonparametric bootstrap only takes the maximum of our predetermined sample.