APSTA Week 09 Exercises

1. Basic bootstrapping (Theory)

We generate a single bootstrap dataset $x_1^*, \ldots, x_n^*$ from the empirical distribution function of

$$1, 4, 6, 7, 8, 11, 15, 19$$

a. What is the probability that the bootstrap sample mean is equal to 19?
b. What is the probability that the minimum of the bootstrap dataset is 1?
c. What is the probability that in the bootstrap sample exactly two elements are ≤ 6 and all the other are ≥ 15?

Solution:
a. What is the probability that the bootstrap sample mean is equal to 19?
The bootstrapped sample mean is calculated by

$$\bar{X}_n^* = \frac{X_1^* + \cdots + X_n^*}{n}.$$

Since we have 8 elements, we set $n = 8$:

$$\bar{X}_8^* = \frac{X_1^* + \cdots + X_8^*}{8}.$$

Each X_1^* has probability $\frac{1}{8}$ to be any of the 8 values $1, 4, 6, 7, 8, 11, 15, 19$. To calculate the probability that $\bar{X}_8^* = 19$, we need to find divide the total number of situations where that is the case, with the total number of situations.
Since there are 8 variables, each with 8 different cases, the total number of different cases is $8^8 = 16777216$.
The only situation where the fraction equals 19 is if the sum equals 152, as $152/8 = 19$.
The *only* situation where the $X_1^* + \cdots + X_8^* = 152$ is if all $X_i^* = 19$. That can only happen in one of the 16777216 cases, thus

$$\mathrm{P}(\bar{X}_8^* = 19) = \frac{1}{16777216} \approx 0.0000000596$$

a *very* low probability.

b. What is the probability that the minimum of the bootstrap dataset is 1?
For each bootstrap value, there is a $\frac{1}{8}$ probability of it being equal to 1. Repeat this 8 times, and we have a binomial distribution with parameters $n = 8, p = 1/8$. The cumulative distribution function of this distribution is given by:

$$F_X = \mathrm{P}(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}.$$

Inserting $n = 8$ and $p = 1/8$ we get:

$$\begin{aligned}
\mathrm{P}(X \geq 1) &= 1 - \mathrm{P}(X < 1) \\
&= 1 - \mathrm{P}(X \leq 0) \\
&= 1 - F_X(0) \\
&= 1 - \sum_{i=0}^{0} \binom{8}{0} \left(\frac{1}{8}\right)^i \left(1 - \frac{1}{8}\right)^{8-i} \\
&= 1 - \left(1 - \frac{1}{8}\right)^8 \\
&= 1 - \left(\frac{7}{8}\right)^8 \\
&= 1 - \frac{7}{8} \cdot \frac{7}{8} \cdot \frac{7}{8} \cdot \frac{7}{8} \cdot \frac{7}{8} \cdot \frac{7}{8} \cdot \frac{7}{8} \cdot \frac{7}{8} \\
&= 1 - \frac{5764801}{16777216} \\
&\approx 0.656
\end{aligned}$$

c. What is the probability that in the bootstrap sample exactly two elements are ≤ 6 and all the other are ≥ 15?
I'll calculate the total number of cases for this one.
There are 8 bootstrap samples. 2 of them have to be one of $1, 4, 6$, the other 6 have to be one of $15, 19$. This means we have 3^2 combinations for the first one, and 2^6 for the second one, for a total of $3^2 \cdot 2^6 = 576$ combinations.
This, however, assumes that the positions can't change, e.g. the 2 numbers have to be the first 2. We can take into account positions by asking "If we force one of the numbers to be the first one, how many positions does the second number have left?"
$Lo_1, Lo_2, Hi, Hi, Hi, Hi, Hi, Hi$.
As we can see, Lo_2 has a total of 7 positions to be in. If we move Lo_1 one spot to the right, Lo_2 has another 6 positions. Continuing with this logic, we get $7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 = 7!$ different orders. Multiplying that by the previous 576 we get:

$$\frac{576 \cdot 7!}{16777216} = \frac{576 \cdot 5040}{16777216} = \frac{2903040}{16777216} \approx 0.173.$$

2. Bright stars (R)

Consider the `brightness` dataset from the `UsingR` package, which collects the brightness of 966 stars. Using empirical bootstrap, estimate the probability

$$Pr[|\bar{X}_n - \mu| > 0.1]$$

where μ is the true mean of the distribution. *Hint*: as we did in class, you will need to approximate this probability by replacing the sample mean with the bootstrapped mean, and μ with the sample mean.

Solution:

```
library(UsingR)
```

```
## Indlæser krævet pakke: MASS
```

```
## Indlæser krævet pakke: HistData
```

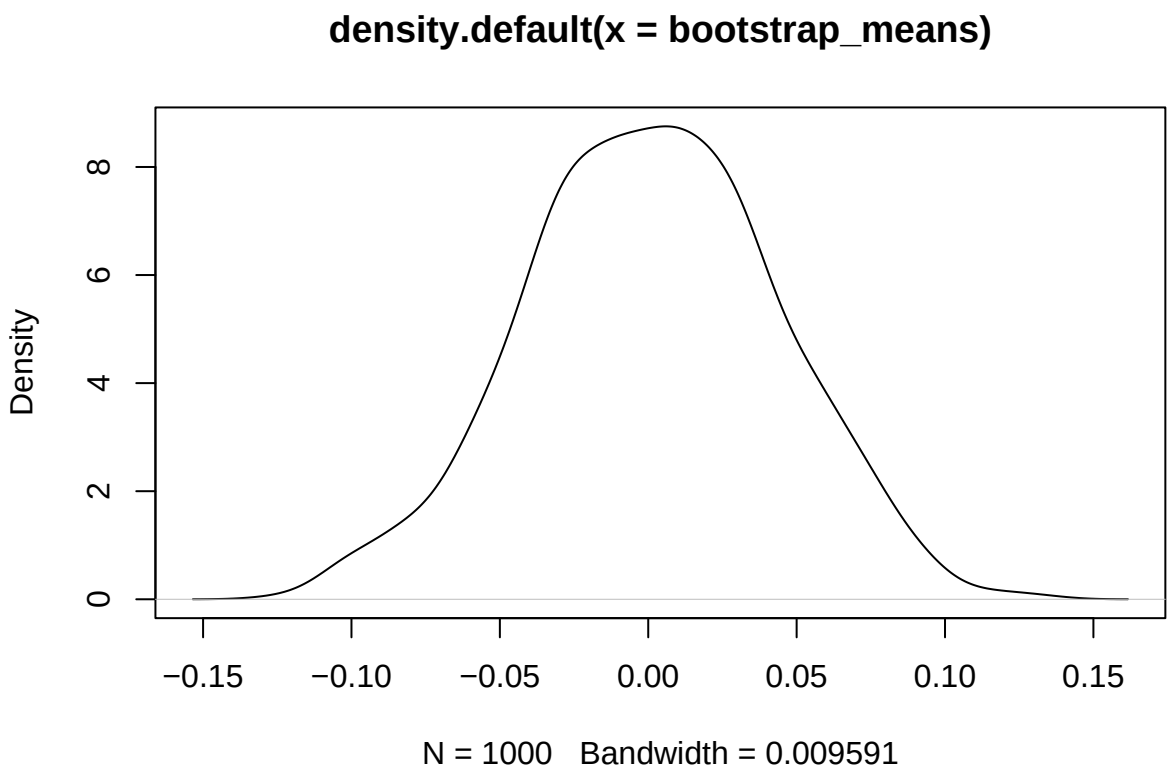
```
## Indlæser krævet pakke: Hmisc

## Indlæser krævet pakke: lattice

## Indlæser krævet pakke: survival

## Indlæser krævet pakke: Formula

## Indlæser krævet pakke: ggplot2

##
## Vedhæfter pakke: 'Hmisc'

## De følgende objekter er maskerede fra 'package:base':
##
##     format.pval, units

##
## Vedhæfter pakke: 'UsingR'

## Det følgende objekt er maskeret fra 'package:survival':
##
##     cancer
```

```
# Take a look at the dataset
head(brightness)
```

```
## [1]  9.10  9.27  6.61  8.06  8.55 12.31
```

$\Pr(|\bar{X}_n - \mu| > 0.1)$ can be estimated by the bootstrapped version, $\Pr(|\bar{X}_n^* - \bar{x}_n| > 0.1)$. We already know $\bar{x}_n$, so we just want to simulate a thousand realisations of $\bar{X}_n^*$.

```
sample_mean <- mean(brightness)
bootstrap_means <- c()
for (i in 1:1000) {
  bootstrap_sample <- sample(brightness, size = length(brightness), replace = TRUE)
  bootstrap_means <- c(bootstrap_means, sample_mean - mean(bootstrap_sample))
}
plot(density(bootstrap_means))
```



We can use this to calculate the probability of this absolute value being greater than 0.1.

```
bootstrap_means_difference_distribution <- ecdf(bootstrap_means)
probability_outside <- 1 - (bootstrap_means_difference_distribution(0.1) - bootstrap_means_difference_di
probability_outside
```

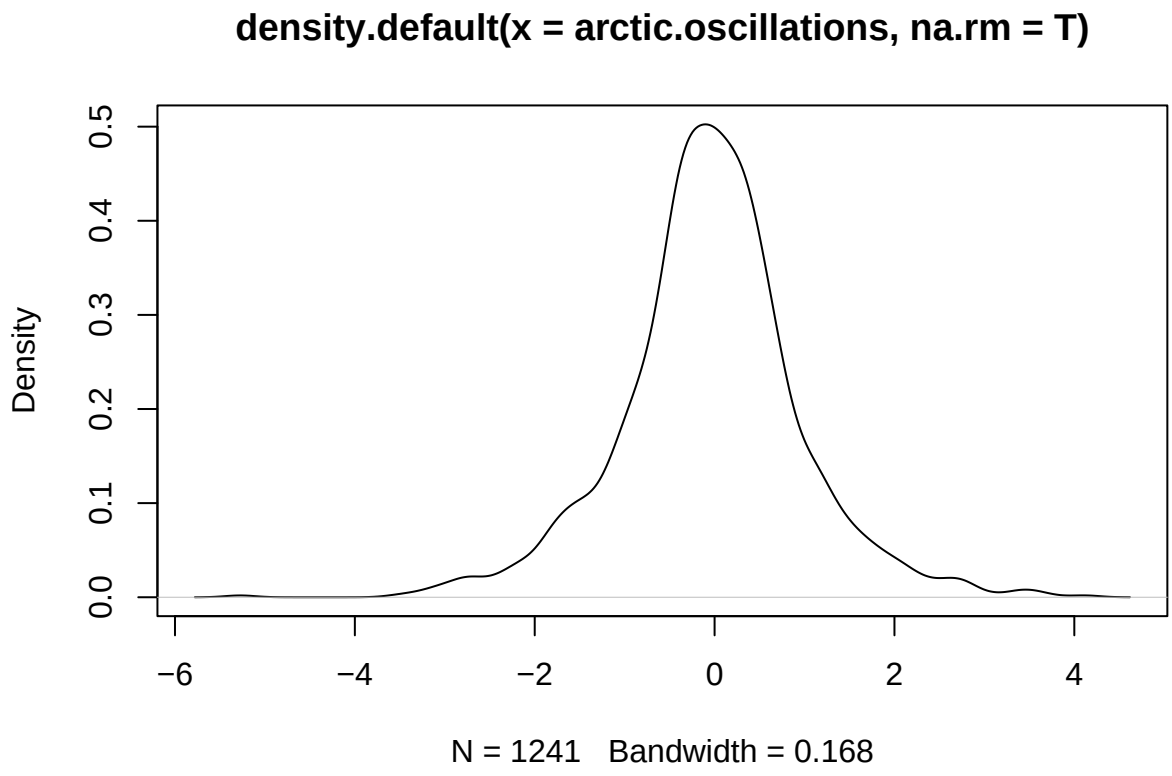
```
## [1] 0.015
```

3. Parametric bootstrap (R)

The dataset `arctic.oscillations` (in package `UsingR`) contains a time series from January to June 2002 of sea-level pressure measurement at the arctic, relative to some base line. Use parametric bootstrap to judge whether it is safe to assume that the measurements are samples from normal distribution or not. *Hint*: use parametric bootstrap in combination with the Kolmogorov-Smirnov distance, as we did in class.

Solution:

```
# start by taking a look at the data
plot(density(arctic.oscillations, na.rm = T))
```

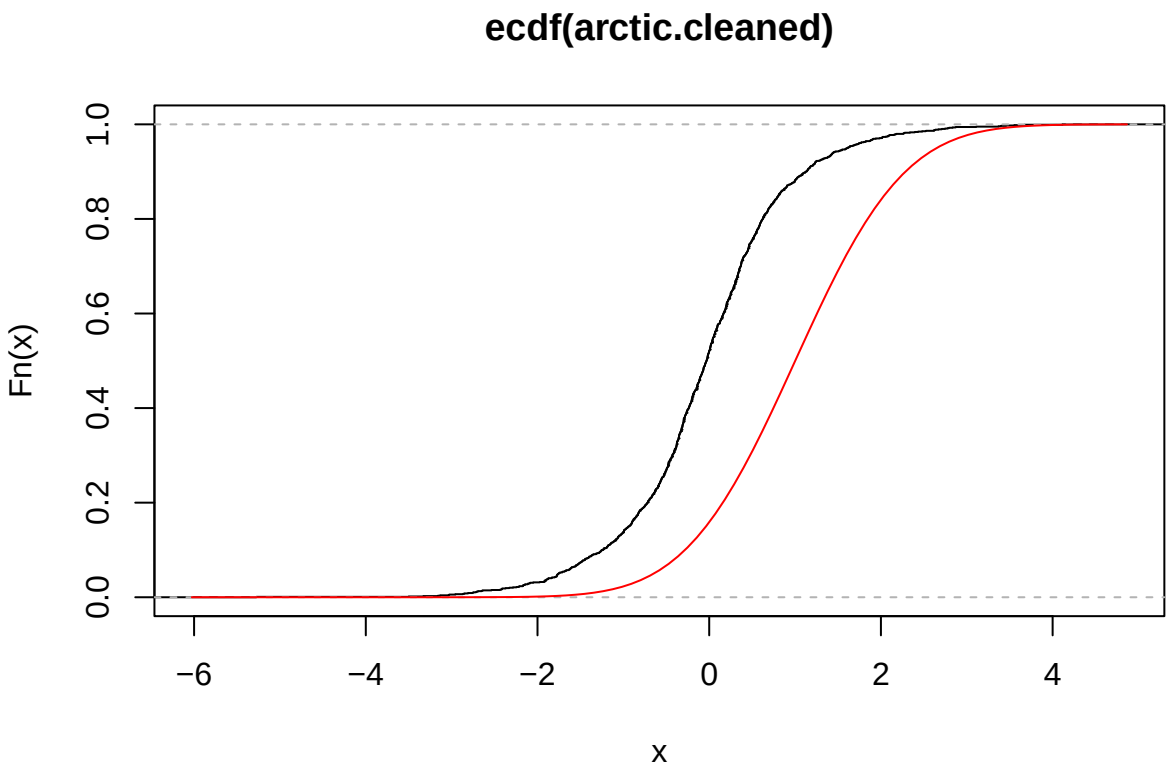


Looks like the dataset has missing values, so let's remove those.

```
arctic.cleaned <- na.omit(arctic.oscillations)
```

To be able to use the parametric bootstrap, we need to estimate the parameters of a normal distribution. The normal distribution is $N(\mu, \sigma^2)$, where μ is the expected value and σ^2 is the variance. We can estimate these using the sample mean $\bar{X}_n$ and the sample variance S_n^2 respectively.

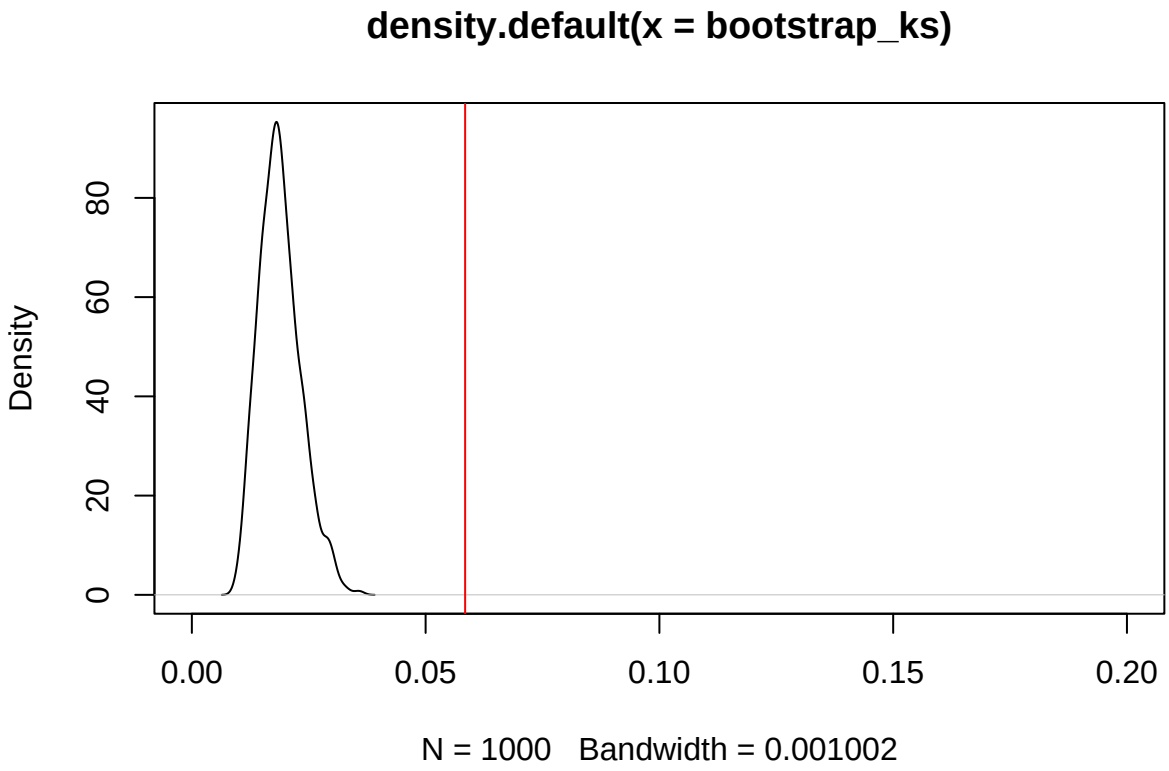
```
expectation_estimate <- mean(arctic.cleaned)
variance_estimate <- var(arctic.cleaned)
plot(ecdf(arctic.cleaned))
curve(pnorm(x, mean=variance_estimate, sd=variance_estimate), col='red', add=T)
```



They look sort of similar, but let's check with the KS distance.

```
ks_distance_norm_distribution <- function(bootstrapped_data, mean, sd) {
  empirical_distribution <- ecdf(bootstrapped_data)
  return(max(
    abs(
      empirical_distribution(bootstrapped_data)
      - pnorm(bootstrapped_data, mean = mean, sd = sd)
    )
  ))
}
ks_estimate <- ks_distance_norm_distribution(
  arctic.cleaned,
  expectation_estimate,
  variance_estimate
)

bootstrap_ks <- c()
for (i in 1:1000) {
  bootstrap_sample <- rnorm(
    length(arctic.cleaned),
    mean = expectation_estimate,
    sd = variance_estimate
  )
  bootstrap_expectation <- mean(bootstrap_sample)
  bootstrap_variance <- var(bootstrap_sample)
  bootstrap_ks <- c(
```

```
    bootstrap_ks,
    ks_distance_norm_distribution(
      bootstrap_sample,
      bootstrap_expectation,
      bootstrap_variance
    )
  )
}

plot(density(bootstrap_ks), xlim=c(0,0.2))
abline(v=ks_estimate, col='red')
```



Looks like our KS distance is wayy higher than we would expect if the distribution was actually normal.

4. Unbiased estimators (Theory)
   Consider a random sample $X_1, \ldots, X_n$ from a uniform distribution in the interval $-\theta, \theta$, where θ is an unknwon parameter. You are interested in estimating the values of θ.

   a. Show that
   $$\hat{\Theta} = \frac{2}{n}(|X_1| + |X_2| + \cdots + |X_n|)$$
   is an unbiased estimator for θ. *Hint*: you may need to use the *change of variable formula* (cfr. Chapter 7 of the book).

b. Consider instead the problem of estimating θ^2. Show that

$$T = \frac{3}{n}(X_1^2 + X_1^2 + \cdots + X_n^2)$$

is an unbiased estimator for θ^2

c. Is $\sqrt{T}$ an unbiased estimator for θ? If not, discuss whether it has positive or negative bias.

Solution:
a.
To check whether $\hat{\Theta}$ is unbiased, we need to find its expectation.

$$\begin{aligned}
\mathrm{E}\left[\hat{\Theta}\right] &= \mathrm{E}\left[\frac{2}{n}(|X_1| + |X_2| + \cdots + |X_n|)\right] \\
&= \mathrm{E}\left[\frac{2}{n}|X_1| + \frac{2}{n}|X_2| + \cdots + \frac{2}{n}|X_n|\right]
\end{aligned}$$

I use linearity of expectations:

$$\begin{aligned}
&= \mathrm{E}\left[\frac{2}{n}|X_1|\right] + \mathrm{E}\left[\frac{2}{n}|X_2|\right] + \cdots + \mathrm{E}\left[\frac{2}{n}|X_2|\right] \\
&= \frac{2}{n}(\mathrm{E}[|X_1|] + \mathrm{E}[|X_2|] + \cdots + \mathrm{E}[|X_n|])
\end{aligned}$$

All the X_i are from the same $U(-\theta, \theta)$ distribution. We want the expectation of $|X_i|$, so we need to use the change of variable formula. The probability density function $f(x)$ of the uniform distribution is $f(x) = 1/(\beta - \alpha)$ where α is the lower bound, here $-\theta$, and β the upper bound, here θ.

$$\begin{aligned}
\mathrm{E}[|X_i|] &= \int_{-\theta}^{\theta} |x| f(x)\ \mathrm{d}x \\
&= \int_{-\theta}^{\theta} |x| \frac{1}{\theta - (-\theta)}\ \mathrm{d}x \\
&= \int_{-\theta}^{\theta} \frac{|x|}{2\theta}\ \mathrm{d}x \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta} |x|\ \mathrm{d}x
\end{aligned}$$

This integral is a bit tricky, since it includes an absolute value. We can split it up into two cases, the case where x is negative, and the case where it isn't:

$$\begin{aligned}
\int |x|\ \mathrm{d}x &= \begin{cases} \int x\ \mathrm{d}x & \text{if } x \geq 0 \\ \int -x\ \mathrm{d}x & \text{if } x < 0 \end{cases} \\
&= \begin{cases} \frac{x^2}{2} + C & \text{if } x \geq 0 \\ \frac{-x^2}{2} + C & \text{if } x < 0 \end{cases} \\
&= \begin{cases} \frac{x \cdot |x|}{2} + C & \text{if } x \geq 0 \\ \frac{x \cdot |x|}{2} + C & \text{if } x < 0 \end{cases} \\
&= \frac{x|x|}{2} + C
\end{aligned}$$

We insert this into the previous integral.

$$\begin{aligned}
\frac{1}{2\theta}\int_{-\theta}^{\theta} |x|\ \mathrm{d}x &= \frac{1}{2\theta}\left[\frac{x|x|}{2}\right]_{x=-\theta}^{x=\theta} \\
&= \frac{1}{2\theta}\left(\left(\frac{\theta|\theta|}{2}\right) - \left(\frac{-\theta|-\theta|}{2}\right)\right)
\end{aligned}$$

θ may be negative, so I can't remove the absolute function, but I can remove the negative sign from the $|-\theta|$, as it will end up positive no matter what.

$$\begin{aligned}
\mathrm{E}[|X_i|] &= \frac{1}{2\theta}\left(\left(\frac{\theta|\theta|}{2}\right) - \left(\frac{-\theta|\theta|}{2}\right)\right) \\
&= \frac{1}{2\theta}\left(\frac{\theta|\theta|}{2} + \frac{\theta|\theta|}{2}\right) \\
&= \frac{1}{2\theta}\left(\frac{2\theta|\theta|}{2}\right) \\
&= \frac{\theta|\theta|}{2\theta} \\
&= \frac{|\theta|}{2}
\end{aligned}$$

Now that we know the expectation of X_i, we can insert it into the expectation for $\hat{\Theta}$ function.

$$\begin{aligned}
\mathrm{E}[\hat{\Theta}] &= \frac{2}{n}(\mathrm{E}[|X_1|] + \mathrm{E}[|X_2|] + \cdots + \mathrm{E}[|X_n|]) \\
&= \frac{2}{n}\left(\frac{|\theta|}{2} + \frac{|\theta|}{2} + \cdots + \frac{|\theta|}{2}\right) \\
&= \frac{2}{n}\cdot\frac{|\theta|}{2} + \frac{2}{n}\cdot\frac{|\theta|}{2} + \cdots + \frac{2}{n}\cdot\frac{|\theta|}{2}) \\
&= \frac{2|\theta|}{2n} + \frac{2|\theta|}{2n} + \cdots + \frac{2|\theta|}{2n} \\
&= \frac{|\theta|}{n} + \frac{|\theta|}{n} + \cdots + \frac{|\theta|}{n} \\
&= n\cdot\frac{|\theta|}{n} \\
&= |\theta|
\end{aligned}$$

Now technically we don't know if it actually gives θ, because if θ is negative, it gives $-\theta$. This however, implies that $-\theta > \theta$, which is not allowed for the uniform distribution, as the lower bound *has* to be lower than the upper bound, meaning θ cannot be negative. Thus we know that $\theta > -\theta$, and $\mathrm{E}[\hat{\Theta}] = \theta$.

b. Now we ant to prove that

$$T = \frac{3}{n}(X_1^2 + X_2^2 + \cdots + X_n^2)$$

is an unbiased estimator for θ^2.

$$\mathrm{E}[T] = \mathrm{E}[\frac{3}{n}(X_1^2 + X_2^2 + \cdots + X_n^2)]$$

Use linearity of expectations to write

$$\begin{aligned}
\mathrm{E}[T] &= \mathrm{E}[\frac{3}{n}(X_1^2 + X_2^2 + \cdots + X_n^2)] \\
&= \frac{3}{n}(\mathrm{E}[X_1^2] + \mathrm{E}[X_2^2] + \cdots + \mathrm{E}[X_n^2])
\end{aligned}$$

Now we find the expectation of X_i^2, again using the change of variable formula.

$$\begin{aligned}
\mathrm{E}[X_i^2] &= \int_{-\theta}^{\theta} x^2 \frac{1}{2\theta}\,\mathrm{d}x \\
&= \frac{1}{2\theta}\int_{-\theta}^{\theta} x^2\,\mathrm{d}x \\
&= \frac{1}{2\theta}\left[\frac{x^3}{3}\right]_{-\theta}^{\theta} \\
&= \frac{1}{2\theta}\left(\frac{\theta^3}{3}+\frac{\theta^3}{3}\right) \\
&= \frac{1}{2\theta}\cdot\frac{2\theta^3}{3} \\
&= \frac{\theta^2}{3}
\end{aligned}$$

Inserting in the previous equation...

$$\begin{aligned}
\mathrm{E}[T] &= \frac{3}{n}(\mathrm{E}[X_1^2]+\mathrm{E}[X_2^2]+\cdots+\mathrm{E}[X_n^2]) \\
&= \frac{3}{n}(\frac{\theta^2}{3}+\frac{\theta^2}{3}+\cdots+\frac{\theta^2}{3}) \\
&= \frac{3}{n}\cdot\frac{n\theta^2}{3} \\
&= \theta^2
\end{aligned}$$

So it is unbiased.

c. $\sqrt{T}$ is biased, as unbiasedness does not carry over in the square root operation, but I'm too tired to prove it right now.
I would assume it has a negative bias.