Applied Statistics 2021 - Exercise 9

1. Basic bootstrapping (Theory)

We generate a single bootstrap dataset $x_1^*, \dots, x_n^*$ 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?

solution: Since the highest value is 19, and there is just a single 19, all elements in our bootstrap sample must consist of that 19, which represents $\frac{1}{8}$ of the dataset.

$$P(\bar{X}_i^* = 19) = P(X_1^* = 19, \dots, X_n^* = 19) = P(X_1^*) \cdots P(X_n^*) = \left(\frac{1}{8}\right)^n = \frac{1}{8^n}$$

b. What is the probability that the minimum of the bootstrap dataset is 1?

solution: since the minimum of the dataset is 1, we want to calculate the probability of picking it at least once:

$$P(min(X_i^*) = 1) = 1 - P(min(X_i^*) > 1) = 1 - P(X_1^* > 1, \dots, X_n^* > 1) = 1 - P(X_1^* > 1) \cdots P(X_n^* > 1) 1 - \left(\frac{7}{8}\right)^n$$

c. What is the probability that in the bootstrap sample exactly two elements are ≤ 6 and all the other are ≥ 15?

solution: Since we want to draw exactly 2 elements less than or equal to 6 out of a sample of n elements, there are $\binom{n}{2} = \frac{(n-1)n}{2}$ ways of drawing these 2 elements, each with a probability of $\left(\frac{3}{8}\right)^2$, because there are 3 elements less than or equal to 6 in our dataset. For the remaining $n-2$ elements, we want to draw elements greater than or equal to 15, of which there are 2, giving us a probability of $\left(\frac{2}{8}\right)^{n-2}$.

Together the probability that our bootstrap sample contains exactly 2 elements less than or equal to 6 and all other greater than or equal to 15 is $\frac{(n-1)n}{2} \left(\frac{3}{8}\right)^2 \left(\frac{1}{4}\right)^{n-2}$.

2. Unbiased estimators (Theory)

Consider a random sample $X_1, \dots, X_n$ 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

$$\hat{\Theta} = \frac{2}{n}(|X_1| + |X_2| + \dots + |X_n|)$$

is an unbiased estimator for θ. *Hint*: you may need to use the *change of variable* formula (cfr. Chapter 7 of the book).

solution: To show that $\hat{\Theta}$ is an unbiased estimator of θ, we have to show $E[\hat{\Theta}] = \theta$ cf. pg. 290 (Dekking et al.).

$$\begin{aligned}
\mathbb{E}[\hat{\Theta}] &= \mathbb{E}\left[\frac{2}{n}(|X_1| + \dots + |X_n|)\right] && (1)\\
&= \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[|X_i|] && (2)\\
&= \frac{2}{n}\sum_{i=1}^{n}\int_{-\theta}^{\theta}|x_i|\frac{1}{2\theta}\mathrm{d}x && (3)\\
&= \frac{2}{n}\sum_{i=1}^{n}\frac{\theta}{2} && (4)\\
&= \theta && (5)
\end{aligned}$$

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

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

is an unbiased estimator for θ^2

$$\begin{aligned}
\mathbb{E}[T] &= \mathbb{E}\left[\frac{3}{n}(X_1^2 + \dots + X_n^2)\right] && (6)\\
&= \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] && (7)\\
&= \frac{3}{n}\sum_{i=1}^{n}Var(X_i) + (\mathbb{E}[X_i])^2 && (8)\\
&= \frac{3}{n}\sum_{i=1}^{n}\frac{(2\theta)^2}{12} && (9)\\
&= \theta^2 && (10)
\end{aligned}$$

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

solution: Instead of trying to find $E[\sqrt{T}]$, we'll use Jensen's inequality (pg. 107 in Dekking et al.). The function $g(x) = -\sqrt{x}$ is a strictly convex function, because its second derivative is positive $\frac{x^{-3/2}}{4} > 0$. By Jensen's inequality we now have

$$\begin{aligned}
-\sqrt{E[T]} &\leq E\left[-\sqrt{T}\right] && (11)\\
-\sqrt{E[T]} &\leq -E\left[\sqrt{T}\right] && (12)\\
\sqrt{E[T]} &> E\left[\sqrt{T}\right] && (13)\\
\theta &> E\left[\sqrt{T}\right]. && (14)
\end{aligned}$$

In other words, $\sqrt{T}$ is a biased estimator for parameter θ with negative bias, because its expected value is less than θ.

3. When the empirical bootstrapp fails (Theory)

The empirical bootstrap is a very powerful tool[1], but there are some situations where its usage is not appropriate.

Consider a dataset $x_1, x_2, \dots, x_n$, which is a realization of the random sample $X_1, X_2, \dots, X_n$ from a $U(0, \theta)$ distribution. Consider the following sample statistic

$$T_n = 1 - \frac{M_n}{\theta}$$

where M_n is the maximum of $X_1, \dots, X_n$. Let m_n be the maximum of the dataset $x_1, x_2, \dots, x_n$, and let $X_1^*, \dots, X_n^*$ be a bootstrap random sample from the empirical distribution function of our dataset. Finally, let M_n^* be the maximum of the bootstrap sample, and consider

$$T_n^* = 1 - \frac{M_n^*}{m_n}$$

a. Compute $Pr[M_n^* < m_n]$.

solution: Since there are $n-1$ elements less than m_n and we have n elements in our bootstrap sample, we get

$$P(M_n^* < m_n) = \left(\frac{n-1}{n}\right)^n = \left(1 - \frac{1}{n}\right)^n \tag{15}$$

b. Argue that $Pr[T_n^* \leq 0] = Pr[M_n^* = m_n]$ and then use the result from the previous point to show that

$$Pr[T_n^* \leq 0] = 1 - \left(1 - \frac{1}{n}\right)^n$$

Furthermore, argue that $Pr[T_n \leq 0] = 0$.

solution: Since the bootstrap random sample is from the original dataset, we always have that $M_n^* \leq m_n$. Hence

$$P(T_n^* \leq 0) = P(M_n^* \geq m_n) = P(M_n^* = m_n) \tag{16}$$

Since $Pr[M_n^* \geq m_n] = 1 - Pr[M_n^* < m_n]$, we have

$$P(T_n^* \leq 0) = 1 - Pr[M_n^* < m_n] = 1 - \left(1 - \frac{1}{n}\right)^n \tag{17}$$

Firstly, if $T_n < 0$, then $M_n > \theta$, which is a contradiction since $M_n \leq \theta$, hence, $P(T_n < 0) = 0$. Secondly, if $T_n = 0$, then $M_n = \theta$, but $P(X_i = \theta) = 0$ cf. pg. 58 (Dekking et al.).

This is equivalent to saying that all X_i are less than θ with probability 1.

c. Let $F_n(t) = Pr[T_n \leq t]$ be the distribution function of T_n, and let $F_n^*(t) = Pr[T_n^* \leq t]$ be the distribution function of the bootstrap statistic T_n^*. Using the result of point b, show that the Kolmogorov-Smirnov distance between the two distributions can be lower bounded as

[1] So much that it has recently been defined “the best statistical pain reliever ever produced”, along with other colorful names, see the acknowledgements section of Efron’s original paper

$$\sup_{t\in\mathbb{R}} |F_n^*(t) - F_n(t)| \geq 1 - \left(1 - \frac{1}{n}\right)^n$$

Hint: consider what happens for $t = 0$ to find the lower bound to the Kolmogorov-Smirnov distance.

solution: At $t = 0$ we have

$$F_n(0) = P(T_n \leq 0) = 1 - P(M < \theta) = 1 - P(X_1 < \theta) \cdots P(X_n < \theta) = 0. \tag{18}$$

The last equation comes from X_i following a $U(0, \theta)$ distribution. Using our result from part b, we get

$$\sup_{t\in\mathbb{R}} |F_n^*(t) - F_n(t)| \geq |F_n^*(0) - F_n(0)| = P(T_n^* \leq 0) = 1 - \left(1 - \frac{1}{n}\right)^n.$$

d. Use the fact that $e^{-x} \geq 1 - x$ to show that

$$1 - \left(1 - \frac{1}{n}\right)^n \geq 1 - e^{-1} \approx 0.632$$

Solution: From the inequality it follows that

$$1 - \left(1 - \frac{1}{n}\right)^n \geq 1 - \left(e^{-1/n}\right)^n = 1 - e^{-1} \tag{19}$$

```r
1-exp(-1)
```

```
## [1] 0.6321206
```

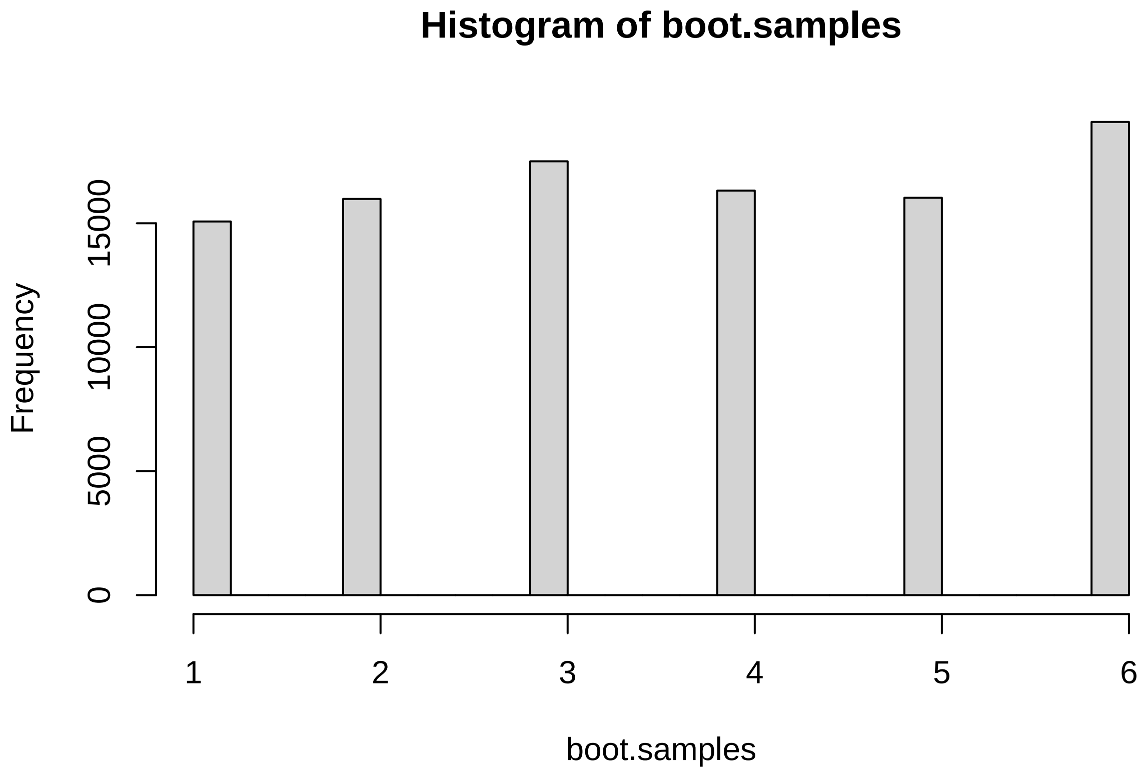
Therefore, you have shown that the Kolmogorov-Smirnov distance between the two distributions is always larger than 0.632, independently of the number of samples n. Discuss, during the exercise session, the consequences of this fact for the bootstrap statistic T_n^*.

4. Is this die fair? (R)

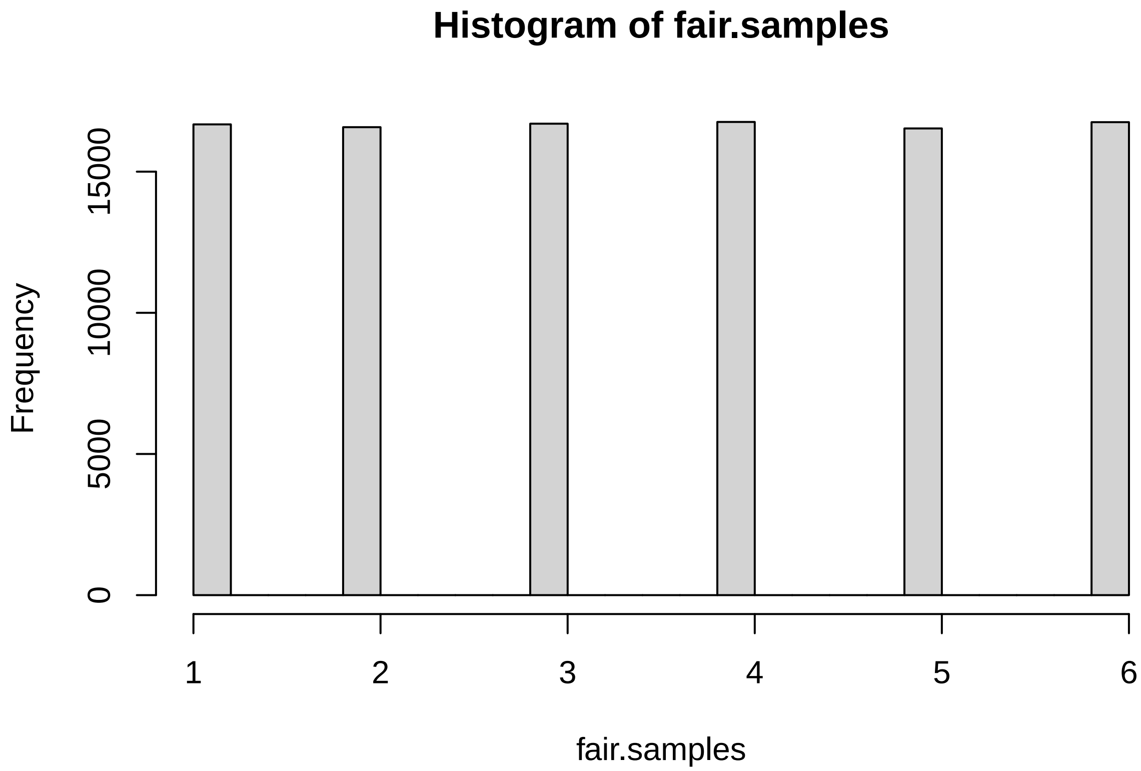
Imagine a situation of purchasing an antique die from the internet for gambling. Before purchasing, you want to make sure that the die is fair. The seller provides 1000 samples of the outcomes available in the file `die_samples.Rdata` (you can access them by the command `load("die_samples.Rdata")`, provided that the file is in the same directory as you RMarkdown file). After loading the file, you can access the dataset under the name `die_samples`. Use bootstrap to determine whether the die is fair or not. Hint: investigate the indicator random variables $I_k = h_k(X), k = 1, 2, ..., 6$, where $I_k = 1 \mathrm{if} X = k$, and $I_k = 0$, otherwise. What would be their expectation if the die was fair? Can you observe a systematic deviation?

```r
load('die_samples.Rdata')
n <- length(die_samples)
B <- 10 ** 5

boot.samples <- sample(die_samples, size = B, replace = TRUE)
hist(boot.samples)
```



```
fair.samples <- sample(1:6, size = B, replace = TRUE)
hist(fair.samples)
```



5. 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.

```
library('UsingR')
```

```
## Loading required package: MASS

## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula
## Loading required package: ggplot2
```

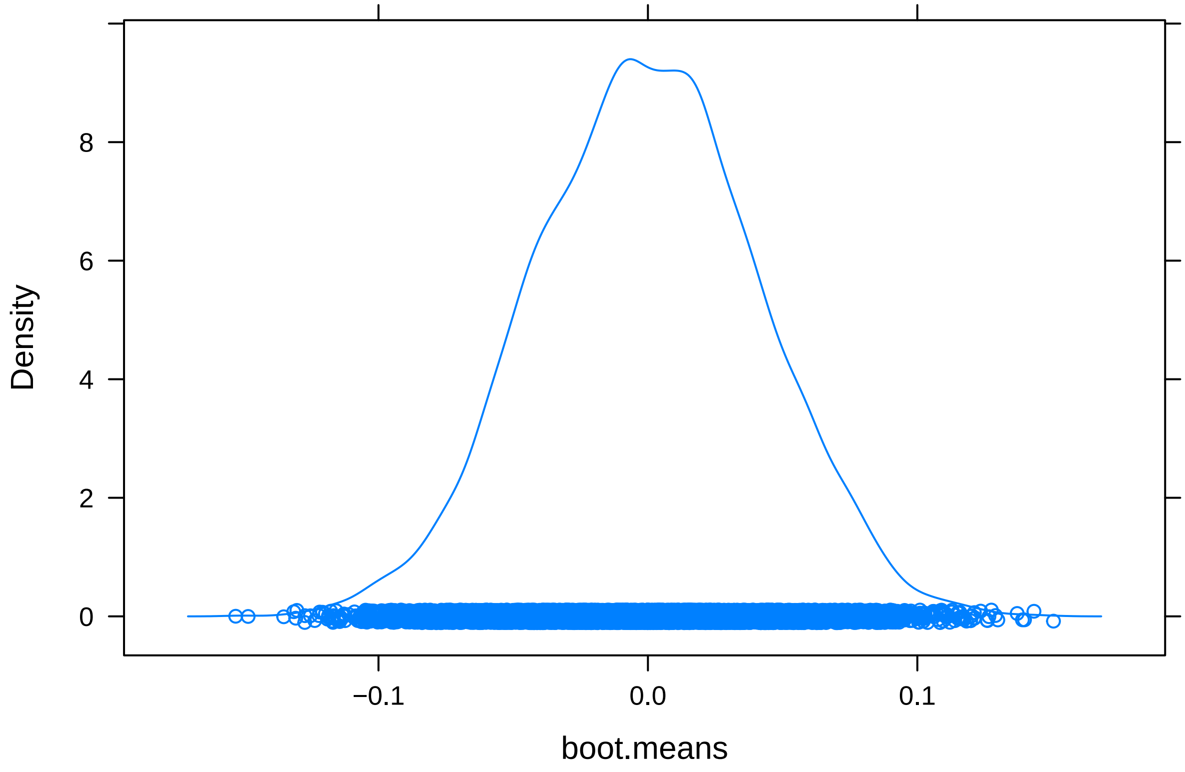
```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
##
## Attaching package: 'UsingR'
```

```
## The following object is masked from 'package:survival':
##
##     cancer
```

```
sample.mean <- mean(brightness)
boot.means <- c()
for (i in 1:10000) {
  boot.sample <- sample(brightness, size = length(brightness), replace = TRUE)
  boot.means <- c(boot.means, sample.mean - mean(boot.sample))
}
densityplot(boot.means)
```



We see what's the probability of being more than 0.1 away from the norm:

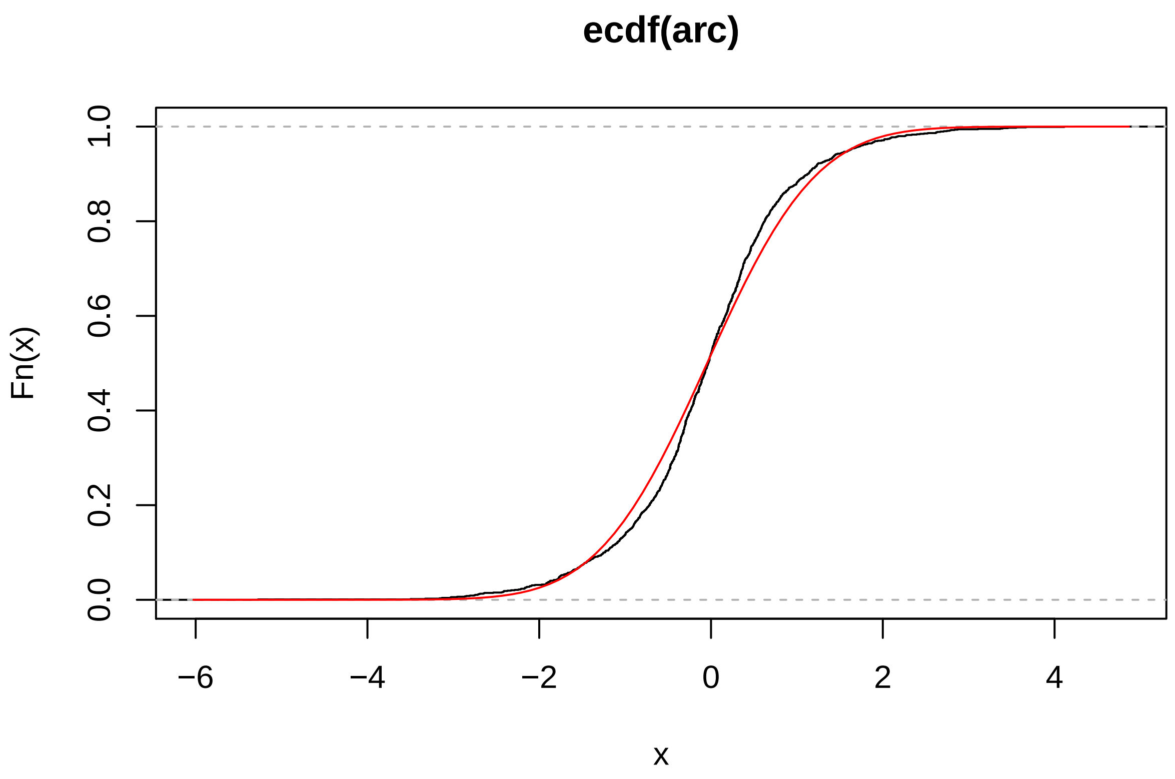
```
boot.means.dist <- ecdf(boot.means)
prob <- (1 - (boot.means.dist(0.1)) + boot.means.dist(-0.1))
prob
```

```
## [1] 0.0158
```

6. 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.

```
arc <- na.omit(arctic.oscillations)
plot(ecdf(arc))
curve(pnorm(x, mean(arc), sd(arc)), col = 'red', add = T)
```

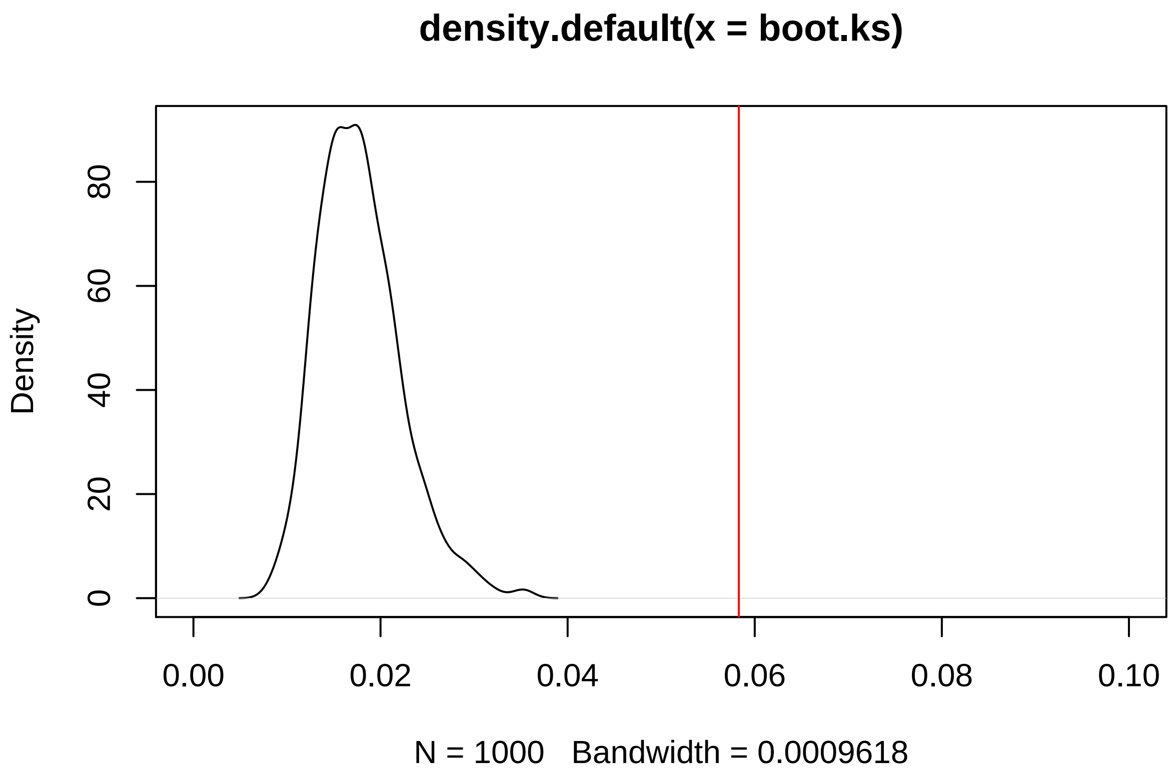


As the two curves look visually similar, cursory inspection does not disprove a normal hypothesis. We thusly use *ks*-distance to *dis*-confirm it:

```
ks.dist.norm <- function(data) {
  emp.dist <- ecdf(data)
  max(abs(emp.dist(data) - pnorm(data, mean(data), sd(data))))
}
```

```r
ks.estimate <- ks.dist.norm(arc)

boot.ks <- c()
for (i in 1:1000) {
  boot.sample <- rnorm(length(arc), mean(arc), sd(arc))
  boot.mean <- mean(boot.sample)
  boot.ks <- c(boot.ks, ks.dist.norm(boot.sample))
}
plot(density(boot.ks), xlim=c(0, 0.1))
abline(v=ks.estimate, col='red')
```


density.default(x = boot.ks)
Density
N = 1000   Bandwidth = 0.0009618