Homework 8

Rylie Fleckenstein

Use set.seed(20218) in each exercise to make results reproducible.

Use 1,000 bootstrap samples where bootstrap is required.

1. Question 5.4.2 pg 197. *Justify your answers and spend some time thinking about the implications of these experiments.*

We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

(a) What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.

The probability that the first bootstrap sample from a set of n observations is $(1 - \frac{1}{n})$. This answer can be justified by simple probability as the probability of selecting a particular sample from a data set is $\frac{1}{n}$.

(b) What is the probability that the second bootstrap observation is not the jth observation from the original sample?

Since the selections are independent of each other, the probability of the second bootstrap observation not being the jth observation from the original sample is the same as the first: $(1 - \frac{1}{n})$.

(c) Argue that the probability that the jth observation is not in the bootstrap sample is $(1 - \frac{1}{n})$.

By using the logic laid out in the previous two questions we have determined that the probability of a single observation in the bootstrap sampling not being the jth observation (so one particular observation) in the original sample to be $(1 - \frac{1}{n})$. Now, we know there are n observations in the original sample and therefore, the probability that the jth observation is not in the bootstrap sample at all is $(1 - \frac{1}{n})^n$.

(d) When n = 5, what is the probability that the jth observation is in the bootstrap sample?

Since the probability that a jth observation is not in the sample is $(1 - \frac{1}{n})^n$. In order to find the probability that the jth observation IS in the sample we must take the compliment of the probability that it is not, which is done in the following manner: $1 - (1 - \frac{1}{n})^n$

```
## [1] 0.36
```

(e) When n = 100, what is the probability that the jth observation is in the bootstrap sample?

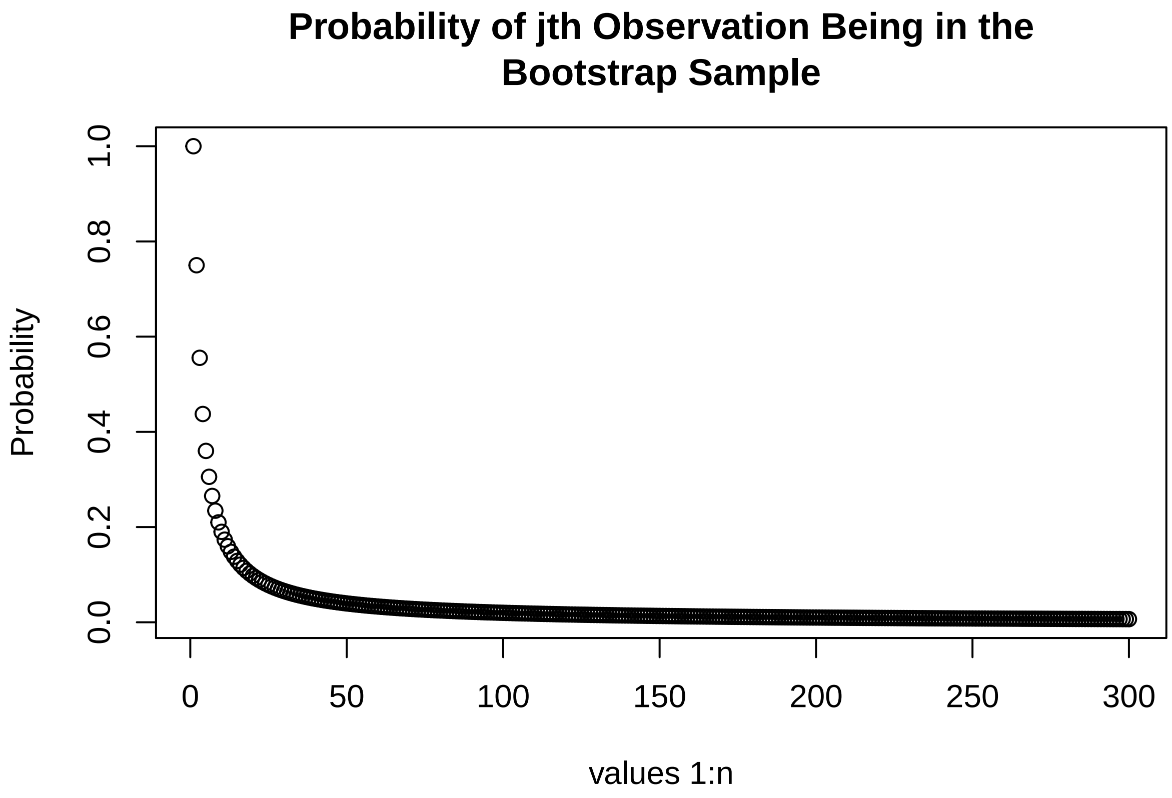
```
## [1] 0.0199
```

(f) When n = 10, 000, what is the probability that the jth observation is in the bootstrap sample?

```
## [1] 0.00019999
```

(g) Create a plot that displays, for each integer value of n from 1 to 100, 000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

Below is a graph displaying the probability that the jth observation is in the bootstrap sample for each integer from 1 to n. The limit of the probability function $((1-\frac{1}{n})^n)$ as n approaches infinity is $1/e$. Therefor, we are going to hit a horizontal asymptote at around 0.368. Because of this limit, I am only graphing the first 300 hundred probabilities so that we can have a closer look into what is going on.



(h) We will now investigate numerically the probability that a bootstrap sample of size n = 100 contains the jth observation. Here j = 4. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

Below we can see that calculated probability is extremely close to what statistical theory tells us it should. The calculated probability that the jth value (4) being in the bootstrap sample is about 0.6316.

```
## [1] 0.6316
```

2. Question 5.4.9 pg 201. *For this question, do not use the* **boot** *library or similar functions. You are expected to code it up in base R with formal annotated code.*

We will now consider the Boston housing data set, from the MASS library.

(a) Based on this data set, provide an estimate for the population mean of medv. Call this estimate $\hat{\mu}$.

Below we can see that our calculations tell us $\hat{\mu} = 22.53$.

```
## [1] 22.53281
```

(b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result. Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.

```
## Standard error of the sample mean: 0.4088611
```

(c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?

As we can see the standard error of $\hat{\mu} = 0.408$ using normal methods of calculation. Below we can see that through the use of bootstrapping we were able to estimate the standard error of $\hat{\mu} = 0.409$ which is well within the margin of error and supports our findings.

```
## Standard error of the sample mean using boostrapping: 0.4093731
```

(d) Based on your bootstrap estimate from (c), provide a 95 % confidence interval for the mean of medv. Compare it to the results obtained using t.test('boston.medv').

We can see that both sets of confidence intervals are extremely close in value with each only differing by a maximum of 0.01.

```
## Bootsrap 95% confidence interval: 21.722 23.35949
```

```
## [1] "t.test results"
```

```
## $conf.int
## [1] 21.72953 23.33608
## attr(,"conf.level")
## [1] 0.95
```

(e) Based on this data set, provide an estimate, $\hat{\mu}_{med}$, for the median value of medv in the population.

```
## Sample median: 21.2
```

(f) We now would like to estimate the standard error of $\hat{\mu}_{med}$. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.

Below I was able to implement a similar bootstrapping method as was used in the previous questions. Through that I was able to estimate the standard error of the sample median to be 0.409.

```
## Bootstrap estimated standard error: 0.4093731
```

(g) Based on this data set, provide an estimate for the tenth percentile of medv in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$. (You can use the quantile() function.)

```
## Tenth percentile esimate for 'medv': 12.75
```

(h) Use the bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.

Again, the bootstrap estimates for the 10th percentile of medv are extremely close in value to the standard calculated estimates.

```
## Bootstrap estimate for the tenth percentile of 'medv' 12.76025
```

Sources:

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer

Rick Wicklin on The DO Loop. (2017, June 28). The average BOOTSTRAP SAMPLE omits 36.8% of the data. Retrieved March 19, 2021, from https://blogs.sas.com/content/iml/2017/06/28/average-bootstrap-sample-omits-data.html#:~:text=A%20bootstrap%20sample%20is%20generated,1%2Fn)%5En.