Modern Applied Statistics exercises from ISLR

Yamuna Dhungana

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.

In a bootstrap, the sample is taken with replacement from the total sample (n) therefore, probability of being jth observation is 1/n hence probability of not being jth observation is 1-1/n

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

For all the samples, data is taken randomly with replacement therefore, the probability of observation that is not jth observation will not be different from the first one. The probability of not being jth observation will be $1 - 1/n$

c. Argue that the probability that the jth observation is not in the bootstrap sample is $(1 - 1/n)^n$.

We know that the n is the total observation and the probability of not being the jth observation is $1 - 1/n$. *The oribability that the jth observation is not in the bootstrap sammple is* $(1 - 1/n)^n$. *Here, we multiplied the n because the sample is taken without replacement and is are an independent sample.* $(1 - 1/n) \cdots (1 - 1/n) = (1 - 1/n)^n$

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

*The probability that the probability that the Jth observation is in the bootstrap sample is $1-(1-1/n)^n$

Now, calculating with n=5 p(j belong in bootstrap sample) $= 1-(1-1/n)^n = 1-(1-1/5)^5 = 0.67232$

Hence, The probabilty that the Jth observation is in boot strap sample is 0.67232*

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

*The probability that the probability that the Jth observation is in the bootstrap sample is $1-(1-1/n)^n$

Now, calculating with n=100 p(j belong in bootstrap sample) $= 1 - (1 - 1/n)^n = 1 - (1 - 1/100)^{100}$ $= 0.6339677$

Hence, The probabilty that the Jth observation is in boot strap sample is 0.634*

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

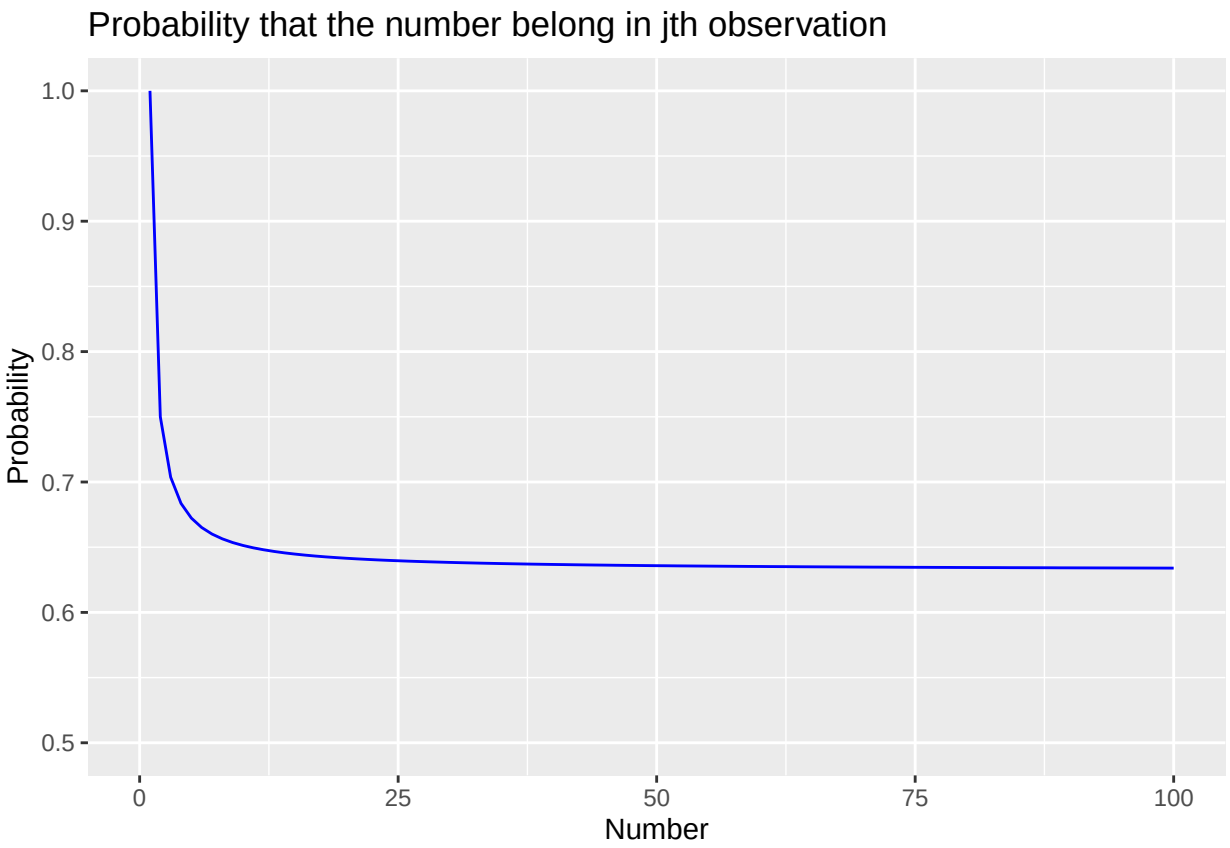
*The probability that the probability that the Jth observation is in the bootstrap sample is $1-(1-1/n)^n$

calculating with n=10000

p(j belong in bootstrap sample) $= 1 - (1 - 1/n)^n = 1 - (1 - 1/10000)^{10000}$ \$ $= 0.632139$\$

Hence, The probabilty that the Jth observation is in boot strap sample is 0.632*

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.



Above is the graphical representation of each integer value of n from 1 to 100,000 and the probability that the jth observation is in the bootstrap sample. Here the line continues to 100000 number from number 12.5. Here, I have limited the x-axis because I wanted to see the curve, which we see in the graphs before 12.5. Without zooming in the graph, it appears a straight line starting from 6.3 on the y-axis. The highest probability of the number is one and, the lowest is near 0.63.

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.

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

Here, We are trying to obtain a bootstrap sample of size n = 100 that contains the jth observation. Here j = 4. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is in the bootstrap sample. We have created a bootstrap sample of 100 samples repeating with a replacement for 10000 times using for loop. The statistic (mean) is performed of the sample that has the fourth observation. The result is reported. Here, The mean of the data that satisfied the condition is 0.6316, which means 63.16% of the samples had the fourth observation. Here I added set.seed in my answer for my convenience.

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 ˆ .

```
## [1] "The standard error of mu is 22.533"
```

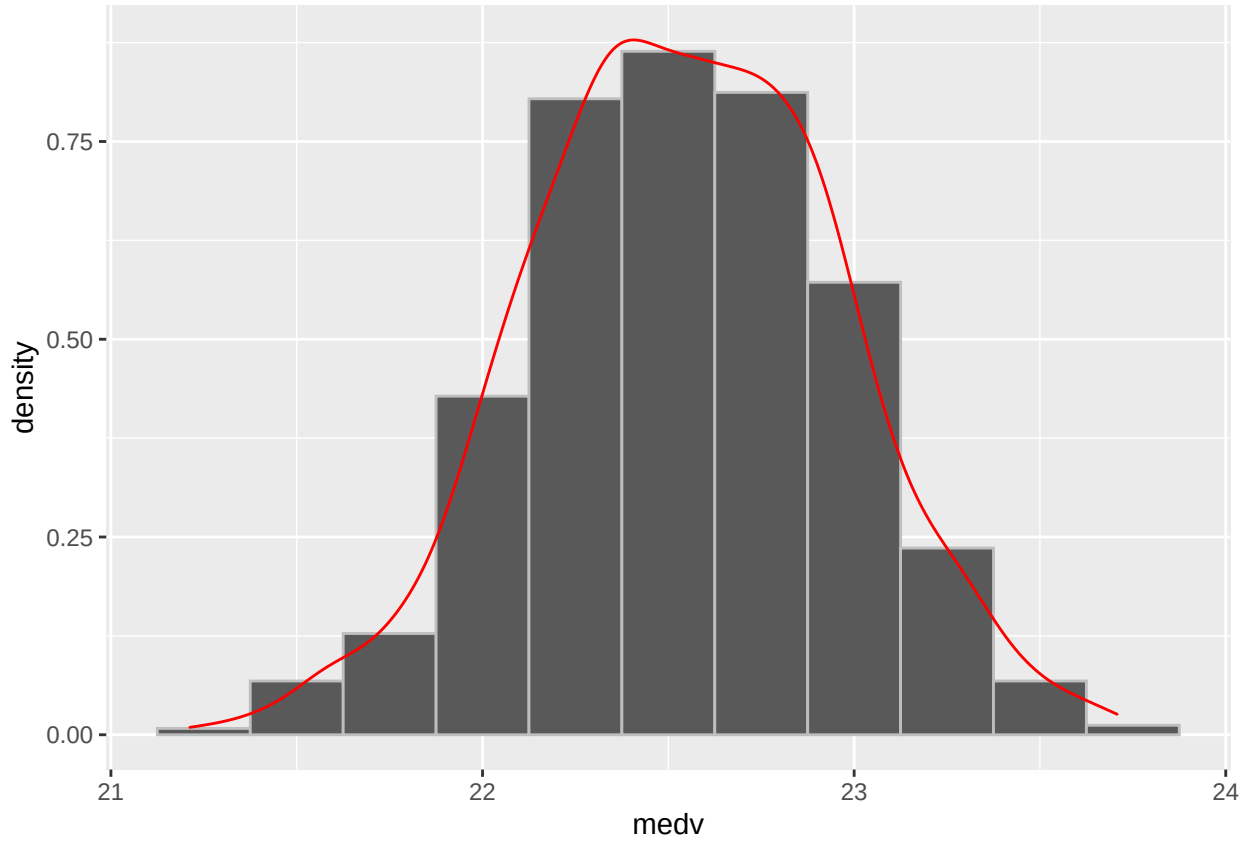
The question is straightforward and is the estimation of the population mean is the sample mean. The estimated population mean of medv was found to be 22.53281.

b. Provide an estimate of the standard error of ˆ . 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.

```
## [1] "The standard error of mu is 0.408861"
```

- By using the hint we found out that the standard error of the mu is 0.408861.*

c. Now estimate the standard error of ˆ using the bootstrap. How does this compare to your answer from (b)?




```
## [1] "The standard error of medv with bootstrap is 0.417868"
```

Here, I have set the seed with 20218 and the bootstrap number of 1000 as mentioned by the question. I have plotted the density plot of the mean of these data. The standard error in bootstrap is calculated by the standard deviation. The standard error of medv with bootstrap is 0.417868. The error obtained by bootstrap is a bit large (about 1 %) than the error obtained without bootstrap.

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). Hint: You can approximate a 95% confidence interval using the formula [ˆ − 2SE(ˆ), ˆ + 2SE(ˆ)].

```
##
##  One Sample t-test
##
## data:  Boston$medv
## t = 55.111, df = 505, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  21.72953 23.33608
## sample estimates:
## mean of x
##  22.53281
```

```
## [1] 21.69707 23.36854
```

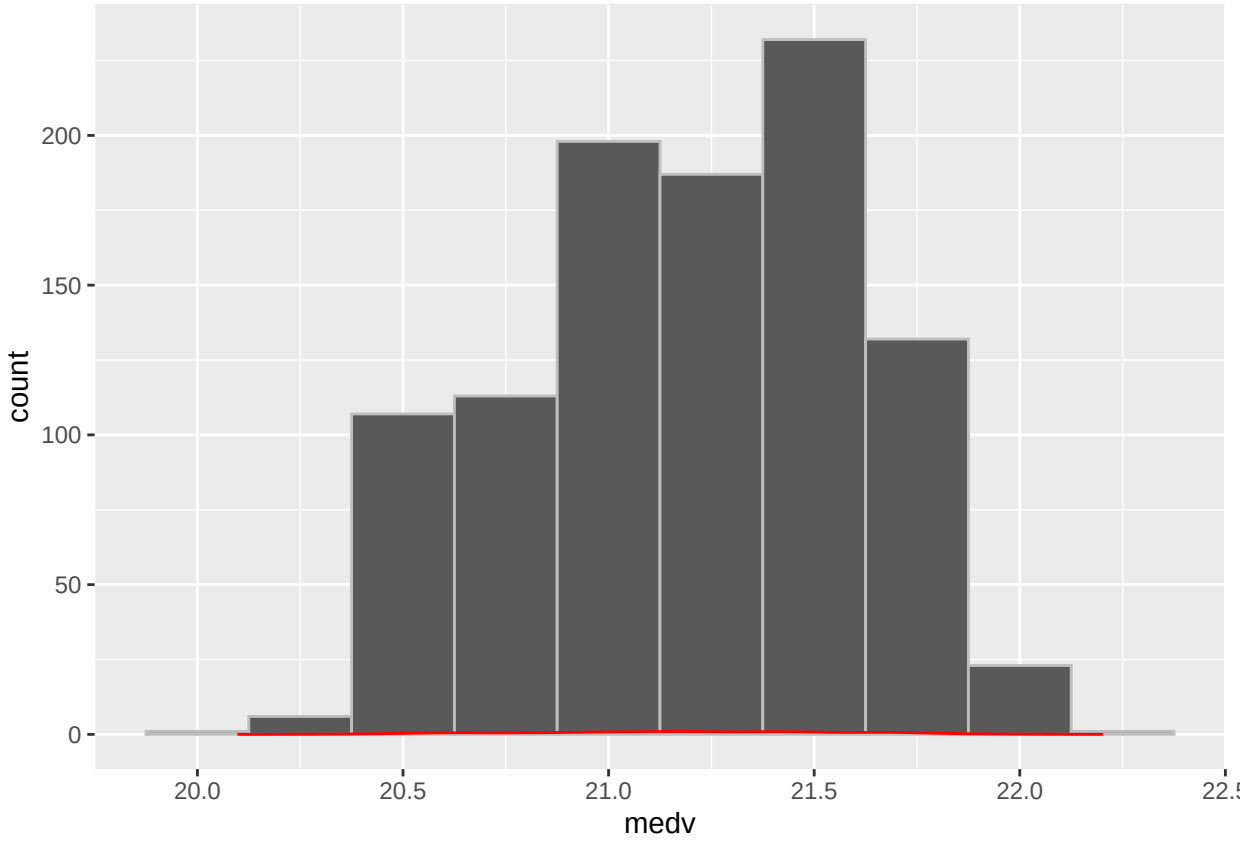
By using the hint given, I have calculated the confidence interval of 95 percentile. The result was obtained by the t-test and the given hint. I have obtained the same confidence with one decimal point of intervals 21.7 and 23.3.

e. Based on this data set, provide an estimate, ˆ med, for the median value of medv in the population.

```
## [1] "The standard error of medv with bootstrap is 21.2"
```

f. We now would like to estimate the standard error of ˆ 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.

```
## [1] 1000  506
```




```
## [1] "The standard error of medv with bootstrap is 0.384951"
```

Like earlier, I have repeated the same process as before the only difference from before is median, is preformed here. Here, I have set the seed with 20218 and the bootstrap number of 1000 as mentioned by the question. I have plotted the bar chat of the median of these data. The standard error in bootstrap is calculated by the standard deviation. The standard error of medv with bootstrap is 0.384951. The error obtained by bootstrap is 0.384951 which is small.

g. Based on this data set, provide an estimate for the tenth percentile of medv in Boston suburbs. Call this quantity ˆ 0.1. (You can use the quantile() function.

```
## [1] "The tenth percentile of medv is 12.75"
```

h. Use the bootstrap to estimate the standard error of ˆ 0.1. Comment on your findings.

```
## [1] "The tenth percentile of medv is 0.83574"
```

Like earlier, I have repeated the same process as before.Here, I have set the seed with 20218 and the bootstrap number of 1000 as mentioned by the question.I have performed with the mean. The confidence interval is performed and The standard error in bootstrap is calculated by the standard deviation. The standard error of medv with bootstrap is 0.384951. The error obtained by bootstrap is 0.835736 which is still small.