Homework 8

Snigdha Peddi

Exercises (ISLR)

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

Use 1,000 bootstrap samples where bootstrap is required.

Question 1 (Question 5.4.2 pg 197), 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.

Answer:

For a total of n observations in original sample, the probability that first bootstrap observation is not the jth observation from original sample is $1 - 1/n$.

For example, if there are 4 observations in original sample, the probability of the first bootstrap observation which is randomly selected (from 1 of the four observations) being the 1st observation is $1/4$ and its probability of not being 1st observation is $(1 - 1/4) = 0.75$.

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

Answer:

For a total of n observations in original sample, the probability that second bootstrap observation is not the jth observation from original sample is $1 - 1/n$. As the bootstrap sampling is done with *replacement*, after the first bootstrap observation is picked the observation is placed back in the original sample.The *second event is independent of 1st event*. When the *second observation* is picked the probability of it being the jth observation is 1/n and *probability of it not being jth observations is* $1 - 1/n$.

For example, if there are 4 observations in original sample, the probability of the first bootstrap observation which is randomly selected (from 1 of the four observations) being the 1st observation is $1/4$ and its probability of not being 1st observation is $(1 - 1/4) = 0.75$. And the probability of the second bootstrap observation being the 1st observation is still $1/4$ as the the resampling is done randomly and is an independent event as resampling is done by replacement. And the second observation not being 1st observation from original sample is $(1 - 1/4) = 0.75$.

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

Answer:

As the bootstrap resampling is done by replacement, the probability of the jth observation not being in the bootstrap sample is product of probability of each of resampling observation not being the jth observation from original sample.

For n observations,

$$(1 - 1/n) * (1 - 1/n) * (1 - 1/n) * \ldots = (1 - 1/n)^n$$

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

Answer:

Probability of jth observation being in bootstrap sample = (1- Probability of jth observation not being in bootstrap sample).

$$n = 5,$$

$$P(jth\ obs\ in\ Bootstrap\ sample) = 1 - (1 - 1/n)^n$$

$$= 1 - (1 - 1/5)^5$$

$$= 1 - 0.32768$$

$$= 0.67232$$

$$= 0.672$$

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

Answer:

Probability of jth observation being in bootstrap sample = (1- Probability of jth observation not being in bootstrap sample).

$$n = 100,$$

$$P(jth\ obs\ in\ Bootstrap\ sample) = 1 - (1 - 1/n)^n$$

$$= 1 - (1 - 1/100)^{100}$$

$$= 1 - 0.3660$$

$$= 0.634$$

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

Answer:

Probability of jth observation being in bootstrap sample = (1- Probability of jth observation not being in bootstrap sample).

$$n = 10000,$$

$$P(jth\ obs\ in\ Bootstrap\ sample) = 1 - (1 - 1/n)^n$$

$$= 1 - (1 - 1/10000)^{10000}$$

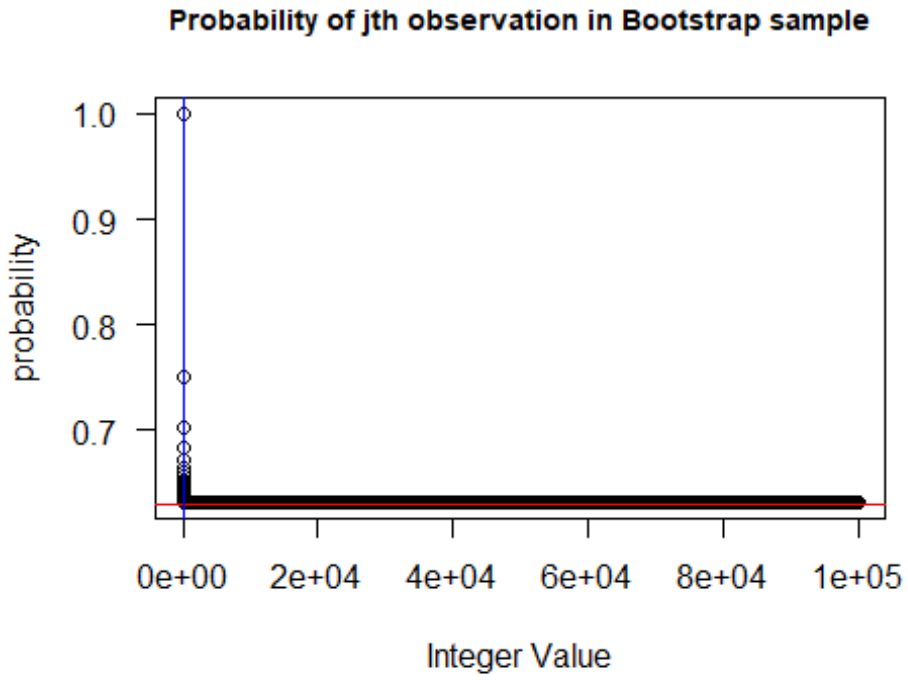
$$= 1 - 0.3679$$

$$= 0.632$$

(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.

Answer:

The plot indicates that approximately when n =240 the probability of jth observation is in the bootstrap sample is around 0.63 and remains consistent for remaining n values.



(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.Comment on the results obtained.

Answer:

As the observations for the bootstrap sample are picked randomly from the original sample with replacement there is a high probability of the jth observation being in most of the bootstrap samples. For a sample of size n=100, the probability that 4th observation being in 10000 bootstrap samples is given by adding the number of times the 4th observation appears in each bootstrap sample. If it is more than 0 then it is saved as True in a vector of length 10000(equal to length of number of bootstrap samples). Mean of this vector gives the probability of the 4th observation being in the bootstrap samples. Executing the code gives a probability of 0.632, meaning there is 63.2% probability that the 4th observation is in the bootstrap samples.

```
## Probability of 4th observation in the Bootstrap sample: 0.6316
```

References

- Lecture by MarinStatsLectures," Bootstrap Hypothesis Testing in Statistics with Example|Statistics Tutorial#35|MarinStatsLectures", Dec 10,2018.
- Lecture by MarinStatsLectures," Bootstrapping and Resampling in Statistics with Example|Statistics Tutorial#12|MarinStatsLectures", Sep 13,2018.
- Lecture by MarinStatsLectures," Bootstrap Hypothesis Testing in R with Example|R Tutorial 4.4 |MarinStatsLectures", Dec 17,2018.

Question 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.

```
## Dimensions of Boston housing dataset: 506 14
```

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

Answer:

```
## Population mean of medv, μ: 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.

Answer:

When number of observations are 506 and standard deviation is 22.53, the variance of the sampling distribution of sample mean (standard error)is given by dividing the variance of the original distribution with square root of number of observations.

```
## Standard Error of μ= 0.4089
```

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

Answer:

The standard error of the mean for Median value of owner occupied homes is 0.4089. Error generated after computing the standard error of mean using 1000 bootstrap samples almost same (0.4093) when compared original error.

```
## Standard error of μ using Bootstrap= 0.4093
```

(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 $[\hat{\mu} - 2SE(\hat{\mu}), \hat{\mu} + 2SE(\hat{\mu})]$.

Answer:

The Bootstrap confidence interval is very close to the confidence interval obtained by t.test.

```
##
##  One Sample t-test
##
## data:  Bos$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
```

95% confidence interval using Bootstrap estimate

Lower CI	Upper CI
21.7142	23.3514

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

Answer:

```
## Population median of medv, μmed: 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.

Answer:

The standard error of median is 0.3803. It is lower than the standard error of mean (0.409) using bootstrap samples. The smaller median value in 1000 bootstrap samples show that there is less sampling fluctuation.

```
## Standard error of μmed using Bootstrap= 0.3803
```

(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.)

Answer:

```
## Tenth percentile of medv,μ0.1: 12.75
```

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

Answer:

The Tenth percentile values of bootstrap sample has an standard error of 0.521. It shows that there is not much sampling bias.

```
## Standard error of μ0.1 using Bootstrap= 0.521
```

References

- Blog post by UCLA Institute for Digital Research & Education, Statistical Consulting, *R Library Introduction to Bootstrapping*.
- Chapter 5,Resampling Methods,*An Introduction to Statistical Learning with Applications in R* by Gareth James.