Assignment23

John Bute

2024-11-07

R Markdown

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

Say we have n observations, where each observation is drawn randomly with replacement. The probability of selecting the jth observation in a single draw is 1/n. Thus, the probability of not selecting the jth observation as the first bootstrap observation is 1 - 1/n.

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

When we pick a bootstrap operation, it is independent. Therefore, the second bootstrap observation would have the same probability as the previous one, thus 1 - 1/n.

3. Argue that the probability that the jth observaion is not in the n bootstrap sample is (1- 1/n)^n

Each bootstrap obersvation is drawn independently. The probability to avoid the jth observation at a single bootstrap observation is 1-1/n. Therefore, to avoid selecting the jth observation in all n draws, we merely just apply the power to the n:

(1-1/n)^n. This expression gives us the probability that the jth observation is not included in the bootstrap sample size n.

4. When n = 5, what is the probability that the jth observation is in the bootstrap sample. We know that to avoid selecting the jth observation in n draws, it is (1 - 1/n)^n. Therefore, if we want to know the probability of having the jth obseravtion in n = 5 draws, then we just need to do the compliment:

1 - (1 - 1/5)^5 = 0.67232.

5.When n=100, what is the probability that the jth observation is in the bootstrap sample? We will do the same logic as before:

1 - (1 - 1/100)^ 100 = 0.63397

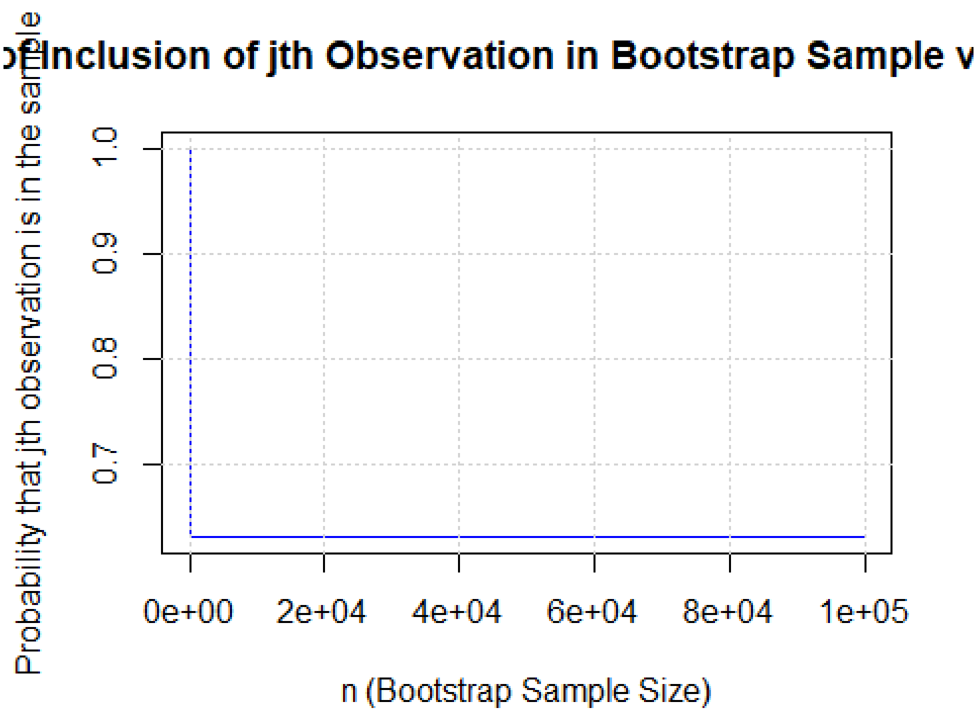
6. When n=10000, what is the probability that the jth observation is in the bootstrap sample? Same as before: 1 - (1 - 1/10000)^10000 = 0.63212

7. Create a plot that displays, for each integer value of n from 1 to 100000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe

```r
n_values <- 1:100000
probabilities <- 1 - (1 - 1 / n_values) ^ n_values

plot(n_values, probabilities, type = "l", col = "blue",
     xlab = "n (Bootstrap Sample Size)",
     ylab = "Probability that jth observation is in the sample",
     main = "Probability of Inclusion of jth Observation in Bootstrap Sample
vs Sample Size n")

grid()
```



It is interesting, as we start from 1, but it seems that as n approaches 100000, the probability that our jth observation is in the bootstraped sample converges to a specific number.

We note that we start with the limit:

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n$$

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$$

$$\text{For } x=-1, \lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = e^{-1} = \frac{1}{e}$$

$$\frac{1}{e} \approx 0.368$$

Which means that there is a 0.368 probability that a specific observation is not included in a bootstrap as n grows large. Therefore, a 0.632 probability that a specific observation is included in a bootstrap sample as n grows large.

8.We 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.

```
store=rep(NA, 10000)


for(i in 1:10000){
   store[i]=sum(sample(1:100, rep=TRUE)==4)>0
}
mean(store)

## [1] 0.6291
```

The probability of the 4th observation being included in a bootstrap of sample size n = 100 is approximately 0.6282, which is awfully close to our formula's answer:

1 - (1 - 1/100)^100 = 0.632