Q23

Rahul Atre

2023-11-01

Q23 [Resampling Methods]

We 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.

1. What is the probability that the first bootstrap observation is not the j^{th} observation from the original sample? Justify your answer.

In a bootstrap sample with n observations, there is an equal probability of choosing any observation in n. So, the probability that a particular j^{th} observation is selected as the first bootstrap observation is $\frac{1}{n}$.

So, by rule of complement, the probability that the **first bootstrap observation is not the j^{th} observation is** $1 - \frac{1}{n}$.

2. What is the probability that the second bootstrap observation is not the j^{th} observation from the original sample?

In bootstrap sampling, the sampling is done with replacement. So, the same sample can be drawn out repeatedly from a data set.

Therefore, the probability that the **second bootstrap observation is not the j^{th} observation is** $1 - \frac{1}{n}$.

3. Argue that the probability that the j^{th} observation is not in the n bootstrap sample is $(1 - \frac{1}{n})^n$.

Since bootstrap sampling is done with replacement, each sample is independent from one another. For the j^{th} probability to occur, there would follows 1 to j independent probabilities beforehand.

Formally, we can say that there is a set of S samples where $S = s_1, s_2, ..., s_n$. The probability of any i sample would be stated as $P(s_i \notin j) = 1 - \frac{1}{n}$. So, the probability that all n samples are not in the j^{th} observation is:

$P(S \notin j) = P(s_1 \notin j) \cap P(s_2 \notin j) \cap ... \cap P(s_n \notin j)$

$= \prod_{i=1}^{n} P(s_i \notin j)$

$= \prod_{i=1}^{n} (1 - \frac{1}{n})$

$= (1 - \frac{1}{n})^n$

Hence, we have proved that the probability of j^{th} observation not being in the n bootstrap sample is $(1 - \frac{1}{n})^n$.

4. When n = 5, what is the probability that the j^{th} observation is in the bootstrap sample?

Modifying the above formula (**in the sample**), we can calculate the probability using an R function:

```r
bootstrap_probabilityInSample_func<- function(n) {
  return(1 - (1 - 1/n)^n) #Since we want it to be in sample
}

bootstrap_probabilityInSample_func(5)
```

```
## [1] 0.67232
```

The probability of the j^{th} observation in the bootstrap sample with n = 5 is **0.67232**.

5. When n = 100, what is the probability that the j^{th} observation is in the bootstrap sample?

```r
bootstrap_probabilityInSample_func(100)
```

```
## [1] 0.6339677
```

The probability of the j^{th} observation in the bootstrap sample with n = 100 is **0.6339677**.

6. When n = 10,000, what is the probability that the j^{th} observation is in the bootstrap sample?

```r
bootstrap_probabilityInSample_func(10000)
```

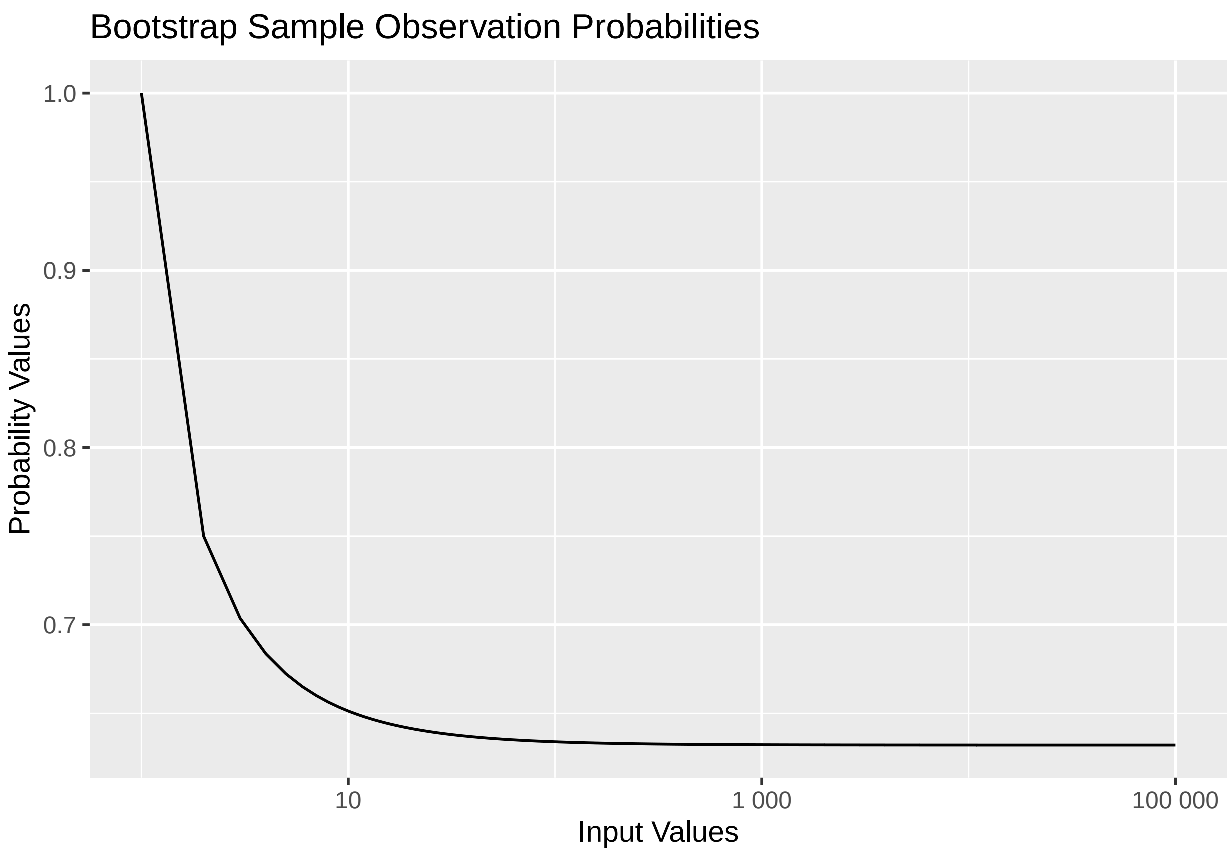
```
## [1] 0.632139
```

The probability of the j^{th} observation in the bootstrap sample with n = 10,000 is **0.632139**.

7. Create a plot that displays, for each integer value of n from 1 to 100,000, the probability that the j^th observation is in the bootstrap sample. Comment on what you observe.

```r
library(ggplot2)
x_values <- 1:100000
y_values <- sapply(x_values, bootstrap_probabilityInSample_func)

ggplot(data.frame(x=x_values, y=y_values), aes(x=x, y=y)) +
  geom_line() +
  scale_x_log10(labels = scales::number_format()) +
  labs(title="Bootstrap Sample Observation Probabilities", x="Input Values", y="Probability Values")
```



From the above function call, we can see that the probabilities are slowly approach an asymptotic line. This is related to the exponential function, because it equates $1 - 1/e = 0.632$. This is often known as the .632 rule in bootstrapping, which is estimating the performance of a classification model using bootstrap samples.

8. We now investigate numerically the probability that a bootstrap sample of size n = 100 contains the j^{th} 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.

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

```
## [1] 0.6247
```

Comment on the results obtained.

For the above results, we generated a set of 100 integers to create 10,000 bootstrap samples to check for the existence of the 4th observation. Since this is a simulation, the probability of expecting the 4th observation lies somewhere between 0.62 and 0.64 most of the time. We know from the previous part that the exponential function converges at 0.632 with a large enough n. Therefore, these simulated outputs are exactly as expected.