Chapter 5 Exercises: Resampling Methods

Statistical Learning with R

Jon Geiger

February 12, 2022

Conceptual Exercise 2

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.

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

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

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

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

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

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.

h. We will not investigate numerically the probability that a boostrap sample of size $n = 100$ contrains 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.

```r
store <- rep(NA, 100000)
for(i in 1:100000) {
    store[i] <- sum(sample(1:100, rep = TRUE) == 4) > 0
mean(store)
}
```

Comment on the results obtained.

Solution

a. With a sample size of n, the probability that the first bootstrap observation *is* the jth observation is $1/n$. This is because there are n possibilities to choose from with equal probabilities, so choosing one out of n possibilities gives a $1/n$ probability.

The complement of this, then, is the probability that the first bootstrap observation is *not* the jth observation, which is $1 - 1/n$.

b. Because we sample with replacement, the probability does not change from the first to the second bootstrap observation. Thus, the probability that the second observation is not the jth observation from the original sample is still $1 - 1/n$.

c. We'll define X_j as being the jth observation from the original data set. We'll call $\Pr(X_j \notin B)$ the probability that the jth observation is not in the bootstrap sample, where $B = \{B_1, B_2, \ldots, B_n\}$ is the set of bootstrap observations from the original sample X. We saw from parts A and B that $\Pr(B_1 \neq X_j) = \Pr(B_2 \neq X_j) = (1 - 1/n)$, and because we sample without replacement, this holds for all n in the bootstrap sample. In other words, for all i, $\Pr(B_i = X_j) = (1 - 1/n)$. So if the bootstrap sample is also of size n, then the probability that the jth observation is not in the bootstrap sample is this probability multiplied together as many times as is the size of the bootstrap sample. So, this would be represented by:

$$\begin{aligned}
\Pr(j \notin B) &= \Pr(B_1 \neq X_j) \times \Pr(B_2 \neq X_j) \times \cdots \times \Pr(B_n \neq X_j) \\
&= \prod_{i=1}^{n} \Pr(B_i \neq X_j) \\
&= \prod_{i=1}^{n} \left(1 - \frac{1}{n}\right) \\
&= \left(1 - \frac{1}{n}\right)^n
\end{aligned}$$

This is the result we wanted to find.

d. The probability that the jth observation *is in* the bootstrap sample is the complement of the probability that it is not, which would be given by:

$$\begin{aligned}
\Pr(X_j \in B) &= 1 - \Pr(X_j \notin B) \\
&= 1 - \left(1 - \frac{1}{n}\right)^n
\end{aligned}$$

```r
prob_Xj_in_B <- function(n) {
    1 - (1 - 1/n)^n
}
prob_Xj_in_B(n=5)
```

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

e.

```r
prob_Xj_in_B(n=100)
```

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

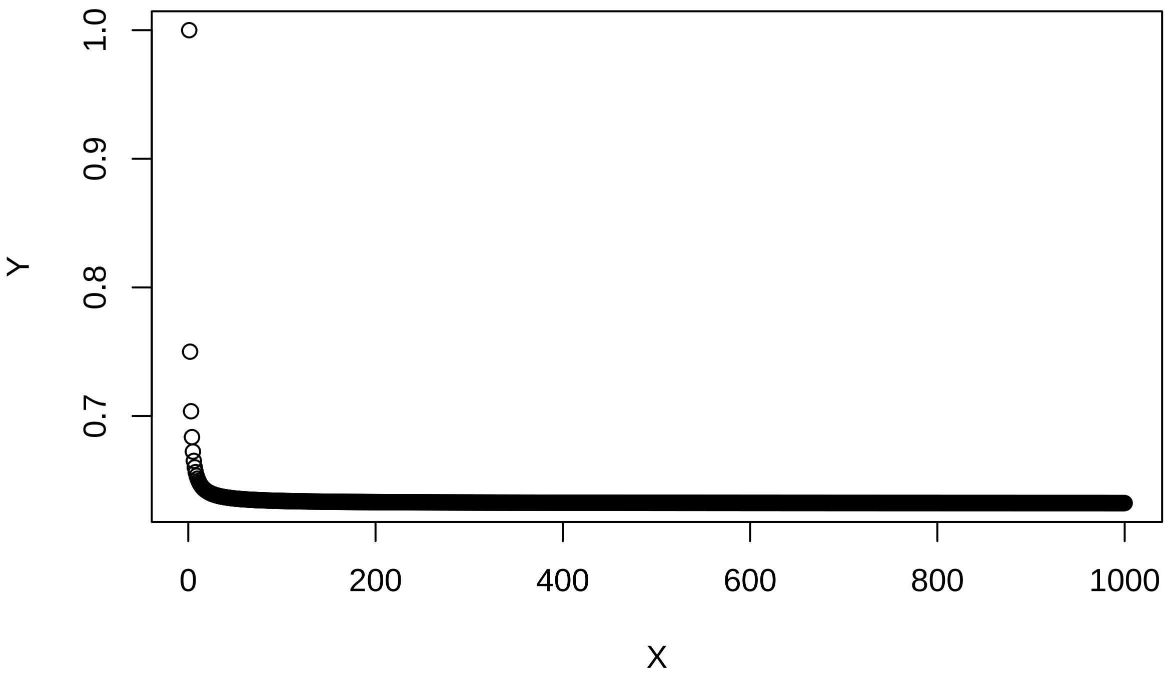
f.

```r
prob_Xj_in_B(n=10000)
```

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

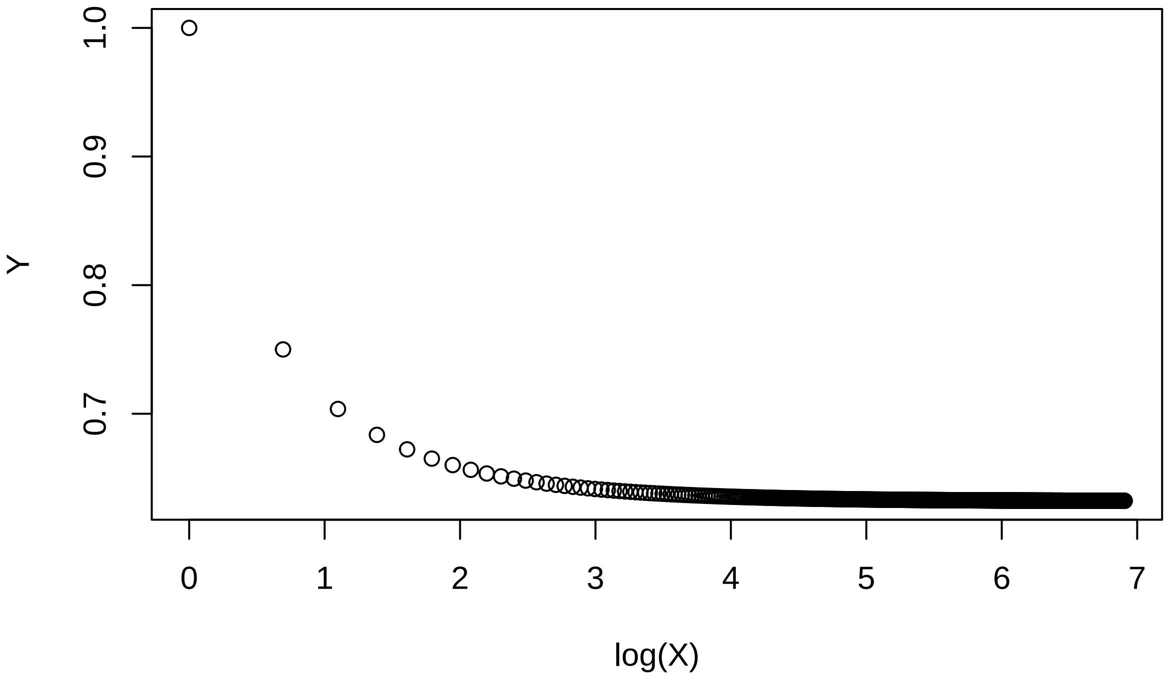
g. To avoid making a gigantic PDF file, we'll just plot a sample of 1000 values for the following plots.

```r
X = c(1:1000)
Y = prob_Xj_in_B(X)
plot(X, Y)
```



It's difficult to see exactly what's going on here, so let's take the log of X and try again:

```
plot(x = log(X), Y)
```



It appears that as n increases, the probability that the jth observation is in the sample tends toward some limiting value. The value of the final data point (a good estimate of the limit) is 0.6321224. It turns out that this value is almost exactly equal to $1 - 1/e$, which is the value you get if you take the limit as $n \to \infty$.

h.
```
store <- rep(NA, 100000)
for(i in 1:100000) {
    store[i] <- sum(sample(1:100, rep = TRUE) == 4) > 0
}
mean(store)
```

```
## [1] 0.63521
```

Across n going from 1 to 10,000, the average proportion of samples for which the fourth observation is contained in the bootstrap sample is roughly equal to $1 - 1/e$

Applied Exercise 9

Solution