# Chapter 5 Exercises: Resampling Methods

Statistical Learning with R

Jon Geiger

February 13, 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 $j$th observation from the original sample? Justify your answer.

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

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

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

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

f. When $n = 10,000$, what is the probability that the $j$th 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 $j$th 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 $j$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.

```
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 $j$th 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 $j$th 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 $j$th observation from the original sample is still $1 - 1/n$.

c. We'll define $X_j$ as being the $j$th observation from the original data set. We'll call $\Pr(X_j \notin B)$ the probability that the $j$th 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 $j$th 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 $j$th observation *is in* the bootstrap sample is the complement of the probability that it is not, which would be given by:

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

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

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

e.
```
prob_Xj_in_B(n=100)
```

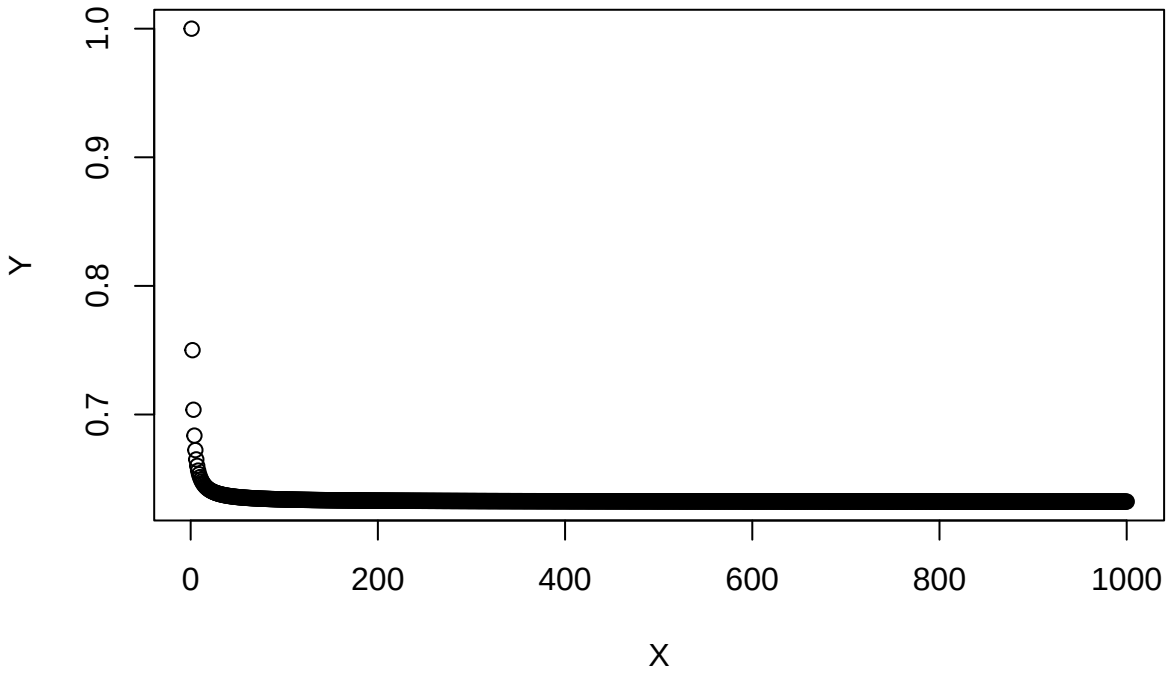
```
## [1] 0.6339677
```

f.
```
prob_Xj_in_B(n=10000)
```

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

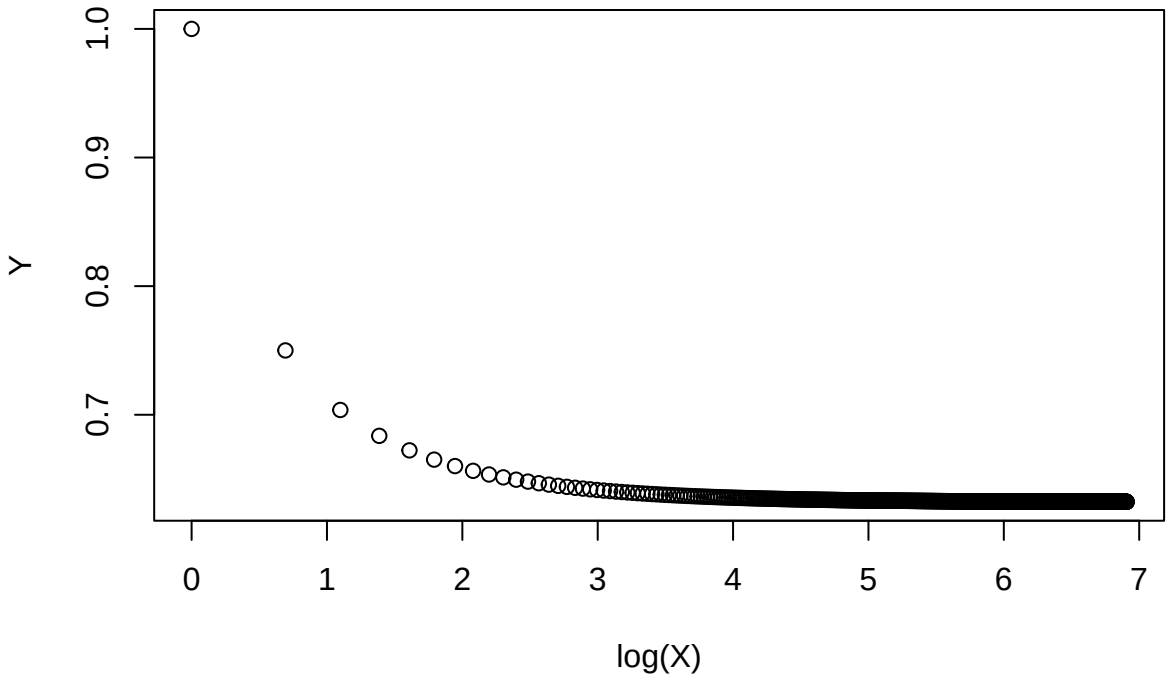
g. To avoid making a gigantic PDF file, we'll just plot a sample of 1000 values for the following plots.

```
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 $j$th 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

We will now consider the `Boston` housing data set, from the ISLR2 library.

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

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.*

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

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

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

f. We would now like to estimate the standard error of $\mu_{med}$. 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.

g. Based on this data set, provide an estimate for the tenth percentile of `medv` in Boston census tracts. Call this quantity $\hat{\mu}_{0.1}$. (You can use the `quantile()` function.)

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

### Solution

```
attach(ISLR2::Boston)
```

a.
```
(mu_hat_medv <- mean(medv))
```

```
## [1] 22.53281
```

b.
```
(se_mu_hat_medv <- sd(medv)/sqrt(length(medv)))
```

```
## [1] 0.4088611
```

c.
```
mu_hat <- function(data, index) {
    X <- data[index]
    return (mean(X))
}
set.seed(1)
boot(medv, mu_hat, R=1000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = medv, statistic = mu_hat, R = 1000)
##
##
## Bootstrap Statistics :
##     original      bias    std. error
## t1* 22.53281 0.007650791   0.4106622
```

So the estimate for our standard error for the mean is 0.410, which is very close to our formula-calculated standard error of 0.408.

d. 95% CI from the bootstrap estimate:

```
cat(mu_hat_medv - 2*0.41066, mu_hat_medv + 2*0.41066)
```

```
## 21.71149 23.35413
```

t-test confidence interval:

```
t.test(medv)$conf.int[1:2]
```

```
## [1] 21.72953 23.33608
```

These confidence intervals are very similar since the standard error estimates are also very similar.

e.
```
median(medv)
```

```
## [1] 21.2
```

f.
```
mu_hat_med <- function(data, index){
    X <- data[index]
    return(median(X))
}
set.seed(1)
boot(medv, mu_hat_med, R=1000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = medv, statistic = mu_hat_med, R = 1000)
##
##
## Bootstrap Statistics :
##     original  bias    std. error
## t1*     21.2 0.02295   0.3778075
```

We can see based on the bootstrap that $\text{SE}(\hat{\mu}_{med}) = 0.3778075$. This is a lower value than we got from the standard error for the mean, implying that it could potentially be a better measure of center than the mean for this distribution.

g.
```
quantile(medv, probs = 0.1)
```

```
##   10%
## 12.75
```

h.
```
mu_hat_0.1 <- function(data, index) {
    X <- data[index]
    return(quantile(X, 0.1))
}
set.seed(1)
boot(medv, mu_hat_0.1, R=1000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = medv, statistic = mu_hat_0.1, R = 1000)
##
##
## Bootstrap Statistics :
##     original  bias    std. error
## t1*    12.75  0.0339   0.4767526
```

The standard error of the tenth percentile for the `medv` variable is 0.4767526. This is on the same order as the other standard errors we found (mean and median), though it is higher. I have no good intuition of what would happen to the standard error as you decreased the quantile either down to zero or up to 100, as you would essentially be estimating the min/max of the bootstrap data set at that point.