Bruce Campbell ST-617 Discussion Group 3

Wed Jul 06 19:09:24 2016

Chapter 5

Problem 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.

Label the data set as $D_j j \in 1, \cdots, n$ and let X_i be the ith boostrap sample.

Since we are independently sampling with replacement the probability of selecting item j for the ith bootsrap observation is

$$P(X_i = D_j) = \frac{1}{n} \quad \forall i, j$$

The probability of the compliment of this event is

$$P(\neg(X_i = D_j)) = (1 - \frac{1}{n}) \quad \forall i, j$$

and we have that $P(\neg(X_1 = D_j)) = (1 - \frac{1}{n})$

b)

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

Again since the samples are independent and with replacement

$P(\neg(X_2 = D_j)) = (1 - \frac{1}{n})$

c)

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

Here we need to calculate

$$P\big(\neg(X_1 = D_j) \bigcap \neg(X_2 = D_j) \cdots \bigcap \neg(X_n = D_j)\big)$$

and by independence of the events we have

$$P(X_i \neq D_j \quad \forall i \quad \in 1, \cdots n) = \prod_{i=1}^{n} P(\neg(X_i = D_j)) = \prod_{i=1}^{n} (1 - \frac{1}{n}) = (1 - \frac{1}{n})^n$$

d)

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

In general

$$P(X_i = D_j for some i) = 1 - P(X_i \neq D_j \forall i \in 1, \cdots n)$$

```
p_5 = 1 - (1 - 1/5)^5
```

Which gives us `0.67232`

e)

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

```
p_100 = 1 - (1 - 1/100)^100
```

Which gives us `0.6339677`

f)

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

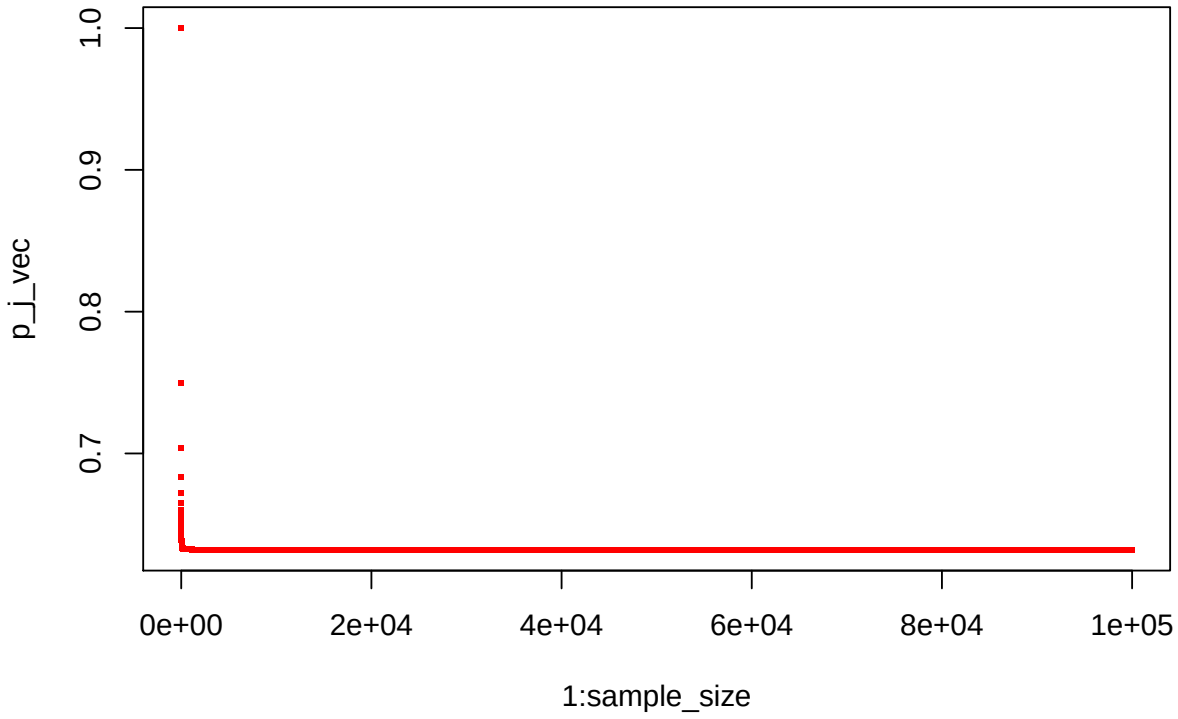
```
p_10000 = 1 - (1 - 1/10000)^10000
```

Which gives us ‘`0.632139`

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.

```
sample_size = 1e+05
p_j_vec = matrix(NA, sample_size, 1)
for (n in 1:sample_size) {
    p_j_vec[n] = 1 - (1 - 1/n)^n
}
plot(1:sample_size, p_j_vec, pch = ".", col = "red", cex = 3)
```



We see rapid convergence. Some calculus yields that the limit is

$$1 - \frac{1}{e}$$

where we've used $\lim_{n\to\infty}(1 - \frac{1}{n})^n = \frac{1}{e}$

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.

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

```
## [1] 0.6363
```

This is close to the limit we have calculated above.

Chapter 5

Problem 4

`Suppose that we use some statistical learning method to make a prediction for the response` Y `for a particular value of the predictor` $X = x$. `Carefully describe how we might estimate the standard deviation of our prediction`

We would use the bootstrap method and equation 5.8 from the text

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{q=1}^{B}\hat{\alpha}^{*q})^2}$$

Here we are usng B bootstrap samples to train the prediction algorithm $Y(X)$ and evaluating it a $X = x$ so $\hat{\alpha}^{*r} = Y_r(X = x)$ where Y_r was trained on the Z^{*r} the rth bootstrap dataset.