HW 04: Resampling

due Thursday, Oct. 13 at 11:59pm

Exercise 1

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. Remember that in bootstrap sampling we repeatedly sample with replacement from the original sample, so each draw is assumed independent of the next.

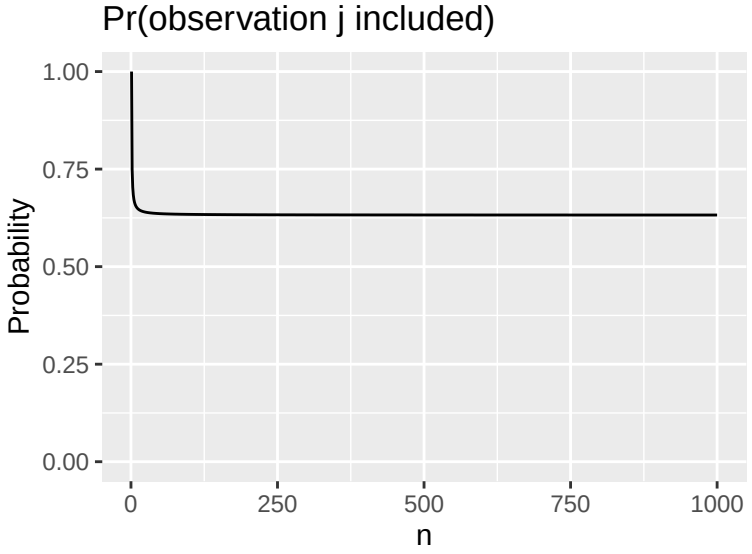
a) What is the probability that the first bootstrap observation is not the j-th observation from the original sample? Justify your answer.

b) What is the probability that the second bootstrap 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 $n = 100$, what is the probability that the j-th observation is in the bootstrap sample?

e) The follow plot displays, for each integer value of n from 1 to 5000, the probability that the j-th observation is in the bootstrap sample. Comment on what you observe.



f) We will 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 for a total of 10000 samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

i) Explain what the commented code in lines 1 and 2 are doing.

ii) Comment on the results obtained.

```r
store <- rep(NA, 10000)
for(i in 1:10000){
  samp <- sample(1:100, rep=TRUE) # line 1
  store[i] <- sum(samp == 4) > 0 # line 2
}
mean(store)
```

```
## [1] 0.6278
```

Exercise 2

We now review k-fold cross-validation.

a) Explain how k-fold cross-validation is implemented.

b) What are the advantages and disadvantages of k-fold cross-validation relative to:

i) The validation set approach?

ii) LOOCV?