2. We will derive the probability that a given observation is part of a bootstrap sample. Suppose we get a bootstrap sample from a set of n observations.

1. What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify.

**Answer:** Probability that any observation is in the sample is 1/n. Therefore, the probability of not choosing the jth observation is 1 - 1/n, or (n-1)/n.

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

**Answer:** Due to the bootstrap method resampling, each observation picked is independent of the last, and therefore has the same probability. Same as 2a.

1. Argue that the probability that the jth observation is not in the bootstrap sample is given by .

**Answer:** Since the probability of not choosing an observation is (1-1/n), and each selection is independent and has the same probability, given there are n selections, the probability of not choosing the jth observation is multiplied n times.

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

**Answer:**

1. D, but n=100.

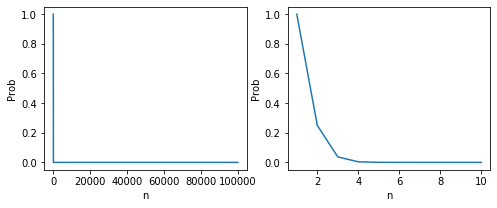
**Answer:**

1. D, but n=10,000

**Answer:**

1. 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. Comments?

**Answer:** The probability of a single observation being in the bootstrap sample decreases rapidly as n increases.



1. Using a program, what is the probability that a bootstrap sample of size n=100 contains the 4th observation. Comment on the results.

**Answer:**

elemList = list(range(0,100))

ioList **=** []

**for** i **in** list(range(0, 100000)):

randomChoices **=** []

**for** i **in** list(range(0, 100)):

elem **=** random.choice(elemList)

randomChoices.append(elem)

​

**if** 4 **in** randomChoices:

io **=** 1

**else**:

io **=** 0

ioList.append(io)

In [61]:

sum(ioList) **/** 100000

Out[61]:

0.63533

Using the above formula derived for the probability that the jth observation will not be in the sample, (0.99)100 = 0.366, the output value makes sense.

3. Reviewing k-fold cross validation

1. Explain how it is implemented

**Answer:** This method is implemented by dividing the data set into k distinct values. Then using all but one of those data sets to train the model. The other data set is used as the test set. This is repeated k times using every data set as the validation set once.

The error values measured with each k test is then averaged to find an accurate error estimate.

1. What are the advantages and disadvantages relative to
   1. Validation set approach

**Answer:** This method requires very little computational effort (comparatively), but is highly variable. This also trains on less data, which implies the model performs worse than other approaches.

* 1. LOOCV (k=n)

**Answer:** This method requires a lot of computational effort, but it has very little variance and much less bias in estimating the error. This estimate tends to be very similar to a 10 or 5 fold CV, but requires many more iterations (n).

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. Carefully describe how we might estimate the standard deviation of our prediction.

**Answer:** We can use a bootstrap method to repeatedly take samples of the data in order to build a model and create estimates/predictions. Each estimate can then be used to find a sample mean and standard deviation.