Question 2

(a) What is the probability that the first bootstrap observation is *not* the j-th observation from the original sample?

In bootstrap sampling, each observation is selected **with replacement** from the original sample of size n.

For any single draw (like the first bootstrap observation), the probability of selecting **any specific** observation (like the j-th) is 1/n.

So, the probability at the **first** draw is **not** the j-th observation is:

$$1 - \frac{1}{n}$$

(b) What is the probability that the second bootstrap observation is not the j-th observation?

Just like part (a), each draw is independent and comes from the same distribution. So the probability is again:

$$1 - \frac{1}{n}$$

(c) What is the probability that the j-th observation is *not in the bootstrap sample*?

A bootstrap sample consists of n draws **with replacement**.

So the j-th observation is **not selected in any of the n draws** if it is not selected in draw 1, and not in draw 2, ..., and not in draw n.

Each draw independently avoids j-th observation with probability $1 - \frac{1}{n}$

, so the probability that **none of the n draws** select the j-th observation is:

$$\left(1 - \frac{1}{n}\right)^n$$

(d) When n=5, what is the probability that the j-th observation *is* in the bootstrap sample?

We already found that the probability the j-th observation is **not** in the bootstrap sample is:

$$\left(1 - \frac{1}{5}\right)^5 = \left(\frac{4}{5}\right)^5 = 0.32768$$

So the probability it **is** included at least once is:

1−0.32768=0.672321

(e) When n=100, what is the probability that the j-th observation is in the bootstrap sample?

$$\left(1 - \frac{1}{100}\right)^{100} \approx e^{-1} \approx 0.3679$$
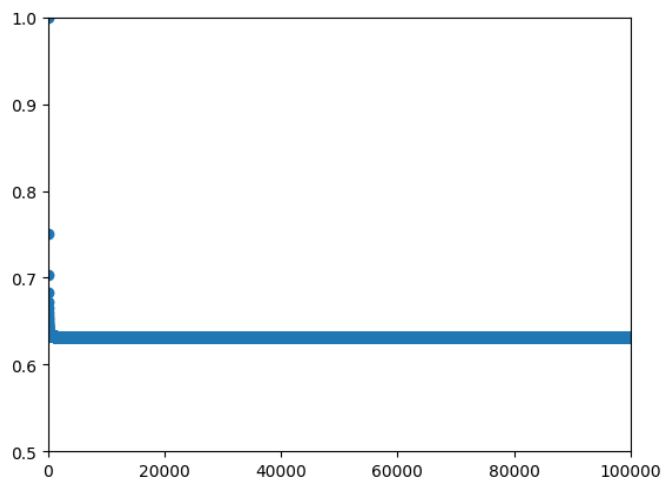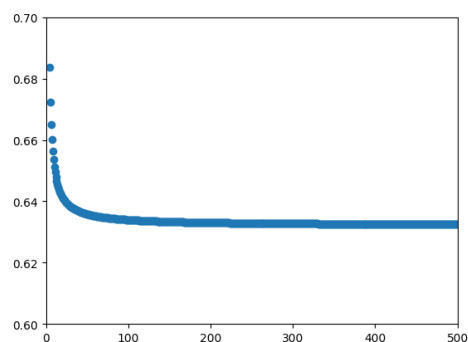
So the probability it *is* included:

1−0.3679≈0.63211

(f)

As n→∞, the probability that a given observation **is included** in the bootstrap sample approaches:

$$1 - \frac{1}{e} \approx 0.63211$$

(g) Plot the probability that the j-th observation is in the bootstrap sample for n=1 to 100000

(h)

0.632, confirming the theoretical result.

Question 3

(a)

In **k-fold cross-validation**, the data is randomly divided into **k approximately equal-sized parts**, called *folds*. The steps are:

1. **Split** the dataset into *k* folds.

2. For each fold:

    o   Use that fold as the **validation set**.

    o   Use the remaining *k-1* folds as the **training set**.

    o   **Train** the model on the training set.

    o   **Evaluate** it on the validation set and record the error.

3. **Repeat** this process *k times*, each time using a different fold as the validation set.

4. **Average the k error estimates** to get the overall cross-validation estimate of test error.

This gives a more stable and reliable estimate of test performance than a single train/test split.

(b)

i. The validation set approach

Advantages:

- **Less variance**: Instead of relying on a single split, k-fold CV averages over multiple splits.

- **More efficient use of data**: Every observation is used for both training and validation (exactly once for validation and k-1 times for training).

- **More reliable test error estimate**: Especially useful when the dataset is small.

Disadvantages:

- **More computationally expensive**: Instead of training one model (as in the validation set approach), you train *k* models.

ii. Leave-One-Out Cross-Validation (LOOCV)

Advantages:

- **Computationally faster**: k-fold CV (with k=5 or 10) requires fewer model fits than LOOCV (which fits the model *n* times for *n* observations).

- **Lower variance** of test error estimate: LOOCV can have high variance because each training set is very similar (just one observation different).

Disadvantages:

- **Slightly more biased** than LOOCV: Because each training set in k-fold CV is smaller than the full dataset (LOOCV uses n−1 examples each time).

- **Less deterministic**: LOOCV gives a fixed result for a given dataset, while k-fold CV depends on the random partition (unless repeated multiple times).

Question 4

To estimate the **standard deviation** (i.e., uncertainty) of our prediction for a specific X, we can use **resampling methods**, especially the **bootstrap**.

Bootstrap Approach to Estimating Prediction Standard Deviation:

1. **Generate many bootstrap samples** from the original dataset:

   o Randomly sample (with replacement) from the dataset to create multiple new datasets of the same size.

2. **Fit the model** on each bootstrap sample.

3. **Make a prediction** for the same value of X using each fitted model.

4. **Collect all the predicted values** for X across the bootstrap samples.

5. **Calculate the standard deviation** of those predictions

Why this works:

The bootstrap mimics the process of sampling from the population. By observing how the prediction varies across different plausible datasets, we can estimate the **variability** (i.e., standard deviation) of the prediction due to sampling variation.

Notes:

- This standard deviation reflects **uncertainty due to sampling variability**, not model bias.

- It **does not account for model misspecification** — it assumes your model form is correct.

- For linear models, you can sometimes compute the prediction variance analytically, but bootstrap works for **any complex model**, including random forests, neural networks, etc.