**Q:** **In k-fold cross-validation, why do we not repeatedly sample the same observation as part of multiple samples?**

A: In cross-validation, we do not sample with replacement because the goal is to systematically partition the data into distinct training and validation sets that collectively cover the entire dataset. This is done for the following reasons:

*1. Accurate Performance Estimation*

Cross-validation aims to provide an unbiased estimate of model performance on unseen data by simulating the process of training the model on a subset of data and then validating it on an independent set. (We will talk about bias in Lecture 9!)

By dividing the dataset into non-overlapping folds (e.g., in k-fold cross-validation), each data point is used exactly once for validation and multiple times for training across different iterations. This ensures that the model's performance is evaluated on every data point in an unbiased manner.

Sampling without replacement allows cross-validation to cover all data points across the validation phases, providing a comprehensive estimate of how well the model generalizes to new data.

*2. No Duplication of Observations*

In cross-validation, using without replacement sampling prevents any duplication of data points within a single fold. This is important because if we sampled with replacement, some observations might appear multiple times in the same training set, while others might be omitted entirely.

This would skew the model's training process, as some data points would have a disproportionate influence on the model, leading to potentially biased performance metrics.

*3. Representativeness of Training and Validation Sets*

Cross-validation (especially k-fold cross-validation) strives to ensure that both the training and validation sets are representative of the overall dataset. When sampling without replacement, we create folds that closely reflect the original distribution of the data, preserving important characteristics such as class balance in classification problems.

This balanced and systematic approach is key for obtaining reliable estimates of model performance, making the results more meaningful when comparing different models or tuning hyperparameters.

*4. Independence Between Training and Validation Sets*

Cross-validation requires that the validation set be independent of the training set to accurately measure the model's ability to generalize. Sampling with replacement would introduce the possibility of overlap between training and validation sets, violating this independence.

For the model's performance estimate to be valid, it must be evaluated on data it has not seen during training. Sampling without replacement ensures this independence, making the evaluation more realistic.

**Q: Is sampling with replacement ever done?**

A: Cross-validation emphasizes performance evaluation across multiple independent subsets. However, sampling with replacement is more aligned with bootstrap validation where the focus is generally on understanding the variability and uncertainty in model predictions.

**Q: What is bootstrap validation?**

A: Bootstrap validation involves creating multiple samples (bootstrap samples) from the original dataset by sampling with replacement. It is called is called "bootstrap" because it draws on the idea of "pulling oneself up by one's bootstraps," which means improving a situation using one's own resources. In the context of statistical methods, the term refers to the technique of using the original data itself to create multiple samples (by sampling *with replacement*) in order to estimate the properties of a statistical model.

Each bootstrap sample is typically the same size as the original dataset but may contain duplicate instances. The model is trained on this bootstrap sample, and the observations that were not included (called "out-of-bag" observations) are used as the validation set. This process is repeated many times (e.g., 100 or 1000 iterations), and the model's performance is averaged over these iterations.

Bootstrap validation is particularly useful when the dataset is small, and the goal is to estimate the uncertainty (variance) in the model's predictions. It gives insights into the model's performance variability because it leverages the randomness in sampling. It's also beneficial for assessing the **bias** and **variance** of a model (something we will talk about in Lecture 9). The bootstrap method allows for the computation of confidence intervals for performance metrics, giving a deeper understanding of model uncertainty.

Bootstrapping can be useful for calculating confidence intervals around model performance metrics, and can be more suitable than cross-validation when dealing with very small datasets where the variability in model performance is high. On the other hand, bootstrapping can be computationally intensive, as it requires repeated training of the model on many bootstrap samples.