交叉验证 **Cross-validation**

金融投资学

Instructor: Haoran LEI

Hunan University

Test-error estimates

- We have discussed model selection methods such as Mallow's C_p, AIC and BIC.
  - These methods try to restore test errors from training errors, by accounting for the degree of model complexity.
- Now we instead consider a class of methods that estimate the test error by *holding out* a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

Validation-set approach

- **Randomly divide** the available dataset into two parts:
  - a *training set* and a *validation set* (or *hold-out set*).
- Use the training set to train (or fit) the model, and the fitted model is used to predict the responses for the observations in the validation set.
- The resulting **validation-set error** provides an estimate of the **test error**. This is typically assessed using **MSE** in the case of a *quantitative response*, and **misclassification rate** in the case of a *qualitative* response.

Drawbacks of validation set approach

1. The validation estimate of the test error can be **highly variable**, depending on how we divide the dataset:

- The validation estimate depends on precisely which observations are included in the training set (and which observations are included in the validation set).

2. The validation set error may tend to **overestimate** the test error for the model fit on the *entire data set.*

K-fold Cross-validation

- Widely used approach for estimating test error.
  - It can be used to *select (the) best model*, and to give an idea of the *test error* of the final chosen model.
- First, Randomly divide the data into K equal-sized parts.
- For each $k \in \{1, ..., K\}$:
  - leave out part k, fit the model to the other $K-1$ parts combined, and then obtain predictions for the left-out k-th part.

K-fold Cross-validation example: $K = 5$

Get MSE_1

1	2	3	4	5
Validation	Train	Train	Train	Train

...

Get MSE_5:

1	2	3	4	5
Train	Train	Train	Train	**Validation**

K-fold Cross-validation: details

- Let the K parts be $C_1, \dots, C_K$ with $|C_k| = n_k$.
- For each $k \in \{1, \dots, K\}$, compute MSE_k by holding out C_k.
- Compute

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$

Or $CV_{(K)} = \sum_k MSE_k / K$ if the dataset is divided equally.

Specialized K-fold Cross-validation: LOOCV

- Setting $K = n$ yields n-fold or **leave-one out cross-validation** (**LOOCV**).
- **Advantage of LOOCV:** with least-squares linear or polynomial regression, an amazing shortcut makes the *computational cost of LOOCV* the same as that of *a single model fit*!
- **Disadvantage of LOOCV:** Typically, LOOCV doesn't *shake up* the data enough.
    - The estimates from each fold are *highly correlated*.

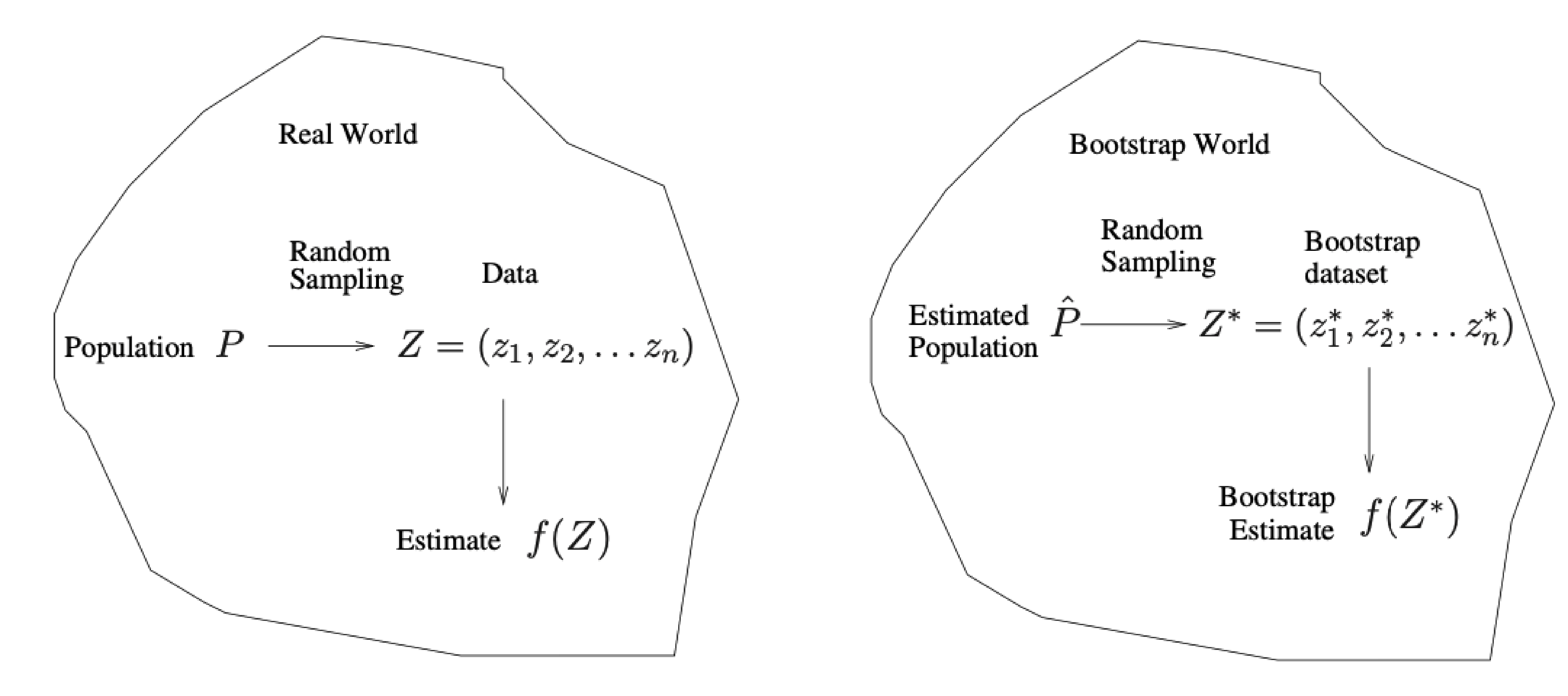
Revisiting bias-variance tradeoff

$K = 5$ or 10 is a better choice for the following reasons:

- Each training set is only $(K-1)/K$ as big as the original training set, the estimates of test error will typically be biased upward.
- This bias is minimized when $K = n$ (LOOCV). However, since the estimates from each fold are *highly correlated*, the estimate of test error has high variance. (**Why?**)
- $K = 5$ or 10 provides a good compromise for this bias-variance tradeoff.

Resampling method

- (K-fold) Cross-validation is a specilized **resampling method**.
  - For example, when $K = 5$, we randomly divide the dataset into two parts (4/5 for training and 1/5 for validation), and we repeat that random division for five times.
  - Each random division is a resample of the dataset.
- Another popular resampling method is **bootstrapping** (自助法). It uses random sampling *with replacement*, and is used to measure **accuracies of an estimate** (e.g., bias, variance, confidence intervals, etc.)



Idea of Bootstrap

Some history: Bootstrap and Jackknife

- The idea of Bootstrap is developed by **Brad Efron** (and Tibshirani), as an improvement of the *Jackknife resampling method*.
- The *Jackknife method* works by **sequentially deleting one observation** in the data set, then **recomputing the desired statistic.**
    - It is both *computationally and conceptually simpler* than bootstrapping. Jackknife **allows exact algebra analysis and more orderly** (i.e. the procedural steps are the same over and over again).

A recent *econometric paper* advocating the usage of Jackknife:

"Jackknife Standard Errors for Clustered Regression," by **Bruce Hansen**, 2022.

> *This paper presents a theoretical case for replacement of conventional heteroskedasticity-consistent and cluster-robust variance estimators with jackknife variance estimators, in the context of linear regression with heteroskedastic and/or cluster-dependent observations. We examine the bias of variance estimation, …*

See 统计学实验: 自助法 for an gentle introduction to bootstrapping.