Resampling Methods

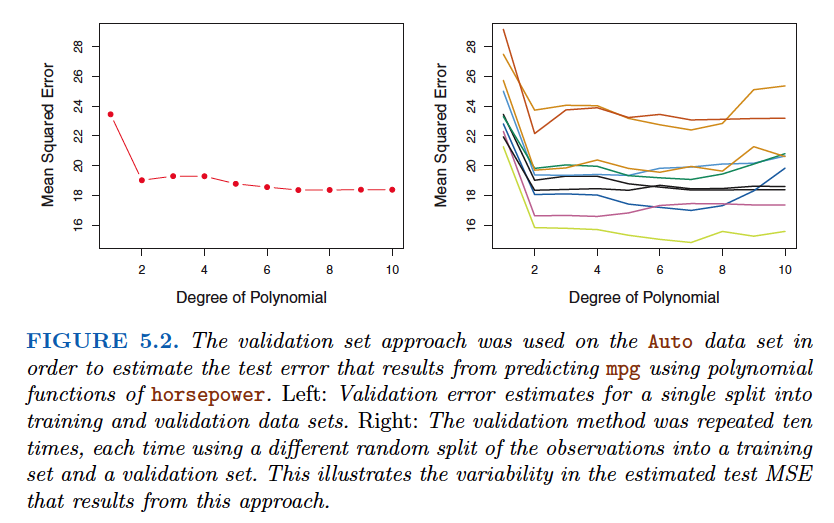
Are tools, involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model. ***For example, in order to estimate the variability of a linear regression fit.***

The process of evaluating a model’s performance is known as ***model assessment*** whereas the process of selecting the proper level of flexibility for a model is known as ***model selection*** .

Cross-Validation:

The Validation Set Approach: It involves randomly dividing the available set of observations into two parts, ***a training set and a validation set or hold-out set*** . The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set. The resulting validation set error rate—typically assessed using MSE in the case of a quantitative response—provides an estimate of the test error rate.

We can assess the model using p-values, but we could also assess it using the validation method. We randomly split the available set of observations into trainings set and validation set. We can fit the model on training set and can get validation set error rates and can compare this error rate for different model. For example, the validation set MSE for the quadratic fit is considerably smaller than for the linear fit. However, the validation set MSE for the cubic fit is actually slightly larger than for the quadratic fit. This implies that including a cubic term in the regression does not lead to better prediction than simply using a quadratic term.



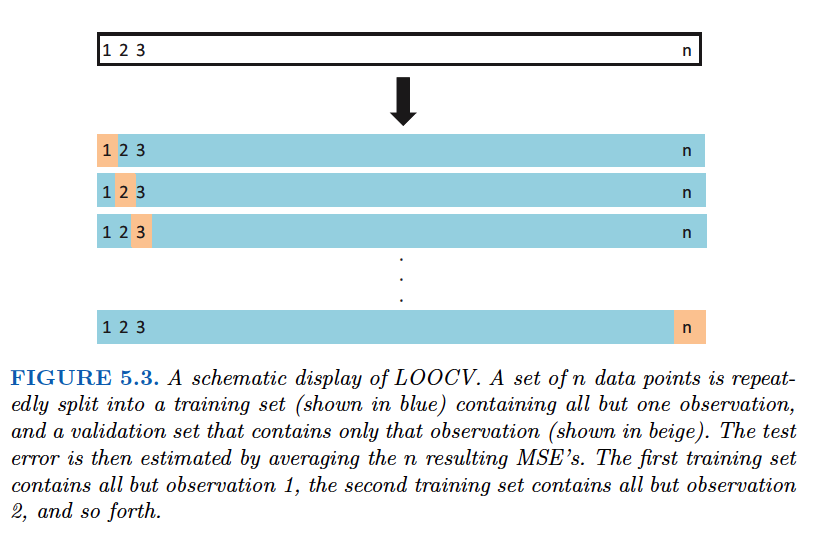
it has two potential drawbacks:

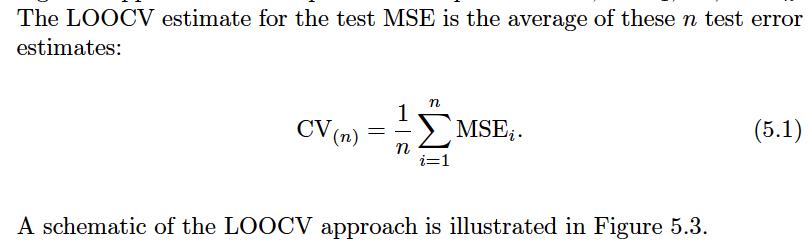
1. As is shown in the right-hand panel of Figure 5.2, the validation estimate of the test error rate can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.

2. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set

Leave-One-Out Cross-Validation:

Instead of creating two subsets of comparable size, a single observation (x1, y1 ) is used for the validation set, and the remaining observations { (x2, y2 ), . . . , (xn, yn )} make up the training set.





***Advantages of LOOCV:***

1. It has far less bias. The LOOCV approach tends not to overestimate the test error rate as much as the validation set approach does. As it is using n-1 observations.
2. In contrast to the validation approach which will yield different results when applied repeatedly due to randomness in the training/validation set splits, performing LOOCV multiple times will always yield the same results: there is no randomness in the training/validation set splits.

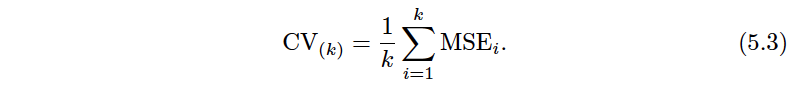
**Disadvantage:**

1.LOOCV has the potential to be expensive to implement, since the model

has to be fit n times.

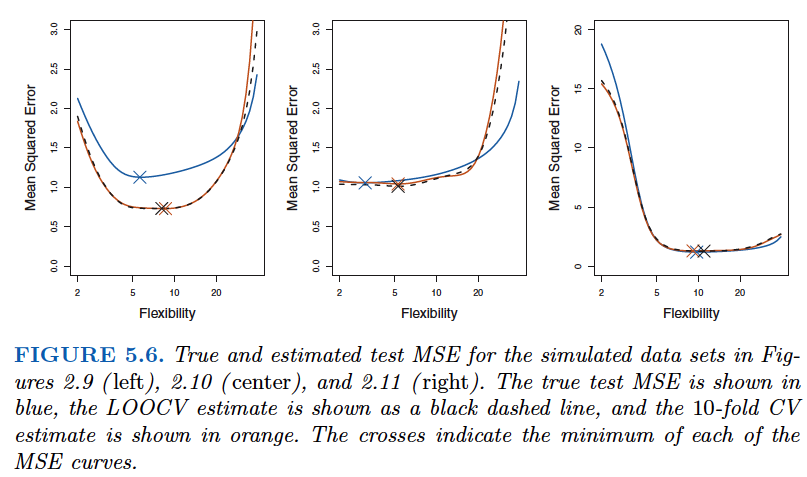
k-Fold Cross-Validation:

An alternative to LOOCV is k-fold CV . This approach involves randomly dividing the set of observations into k groups, or folds , of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds. The mean squared error, MSE1 , is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error, MSE1, MSE2, . . . , MSEk . The k -fold CV estimate is computed by averaging these values,



***Purpose of cross validation:***

* How well a given statistical learning procedure can be expected to perform on independent data; in this case, the actual estimate of the test MSE isof interest.
* But at other times we are interested only in the location of the minimum point in the estimated test MSE curve . This is because we might be performing cross-validation on a number of statistical learning methods, or on a single method using different levels of flexibility, in order to identify the method that results in the lowest test error. For this purpose,the location of the minimum point in the estimated test MSE curve is important, but the actual value of the estimated test MSE is not.



Bias-Variance Trade-Off for k-Fold Cross-Validation:

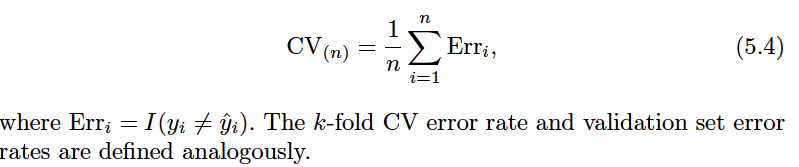
We mentioned in Section 5.1.3 that k -fold CV with k < n has a computational advantage to LOOCV. But putting computational issues aside***, a less obvious but potentially more important advantage of k -fold CV is that it often gives more accurate estimates of the test error rate than does LOOCV.*** This has to do with a bias-variance trade-off.

LOOCV has low bias and high variance, But k-fold CV has high bias and low variance.

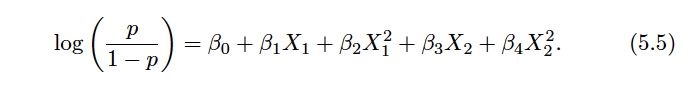
***To summarize, there is a bias-variance trade-off associated with the choice of k in k -fold cross validation.***

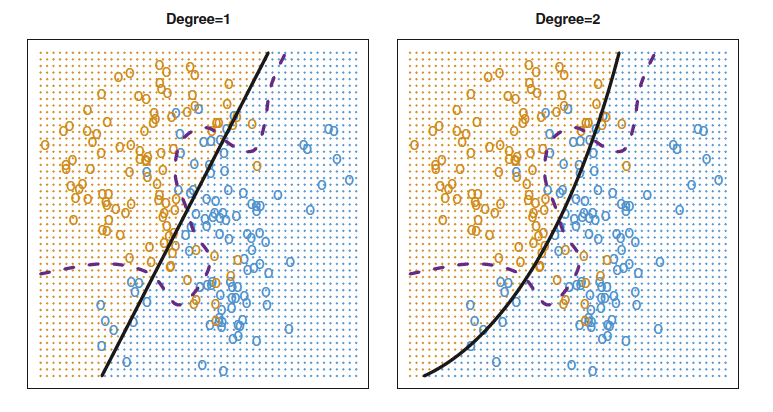
Cross-Validation on Classification Problems

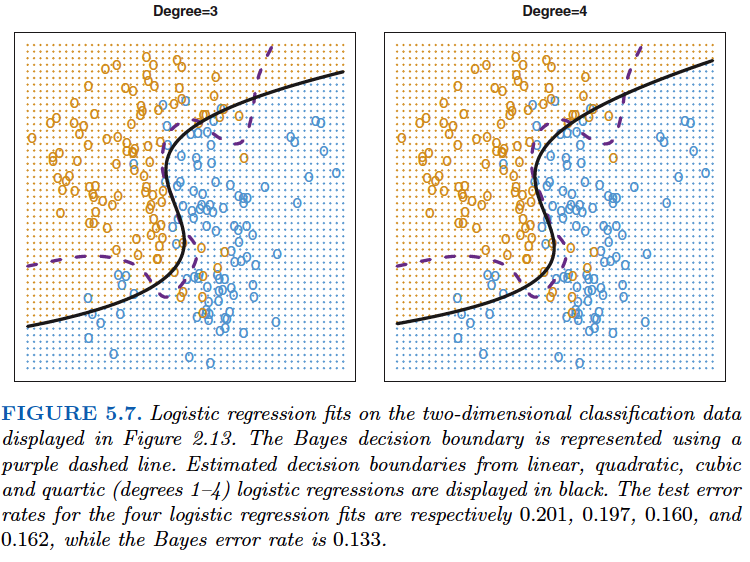
For instance, in the classification setting, the LOOCV error rate takes the form

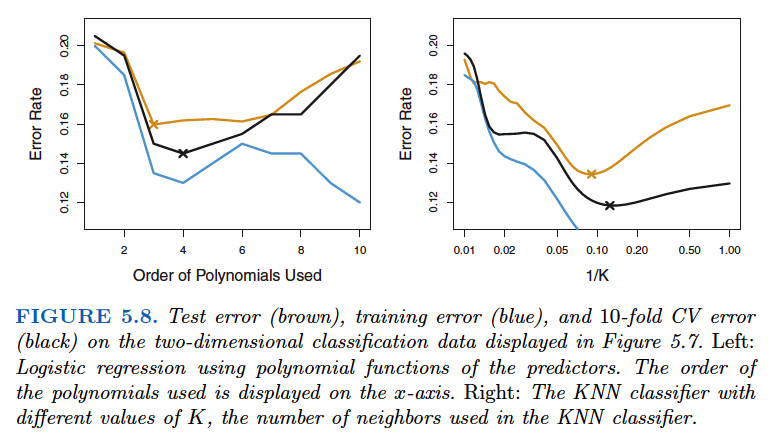


As an example, we fit various logistic regression models on the two dimensional classification data displayed in Figure 2.13. we can fit a quadratic logistic regression model, given by







******

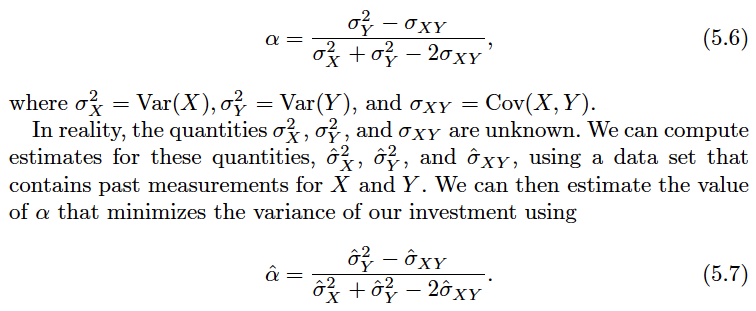
***Above curve for test-error for cross validation, can be used to choose required flexibility for which we will get minimum test error.***

The Bootstrap:

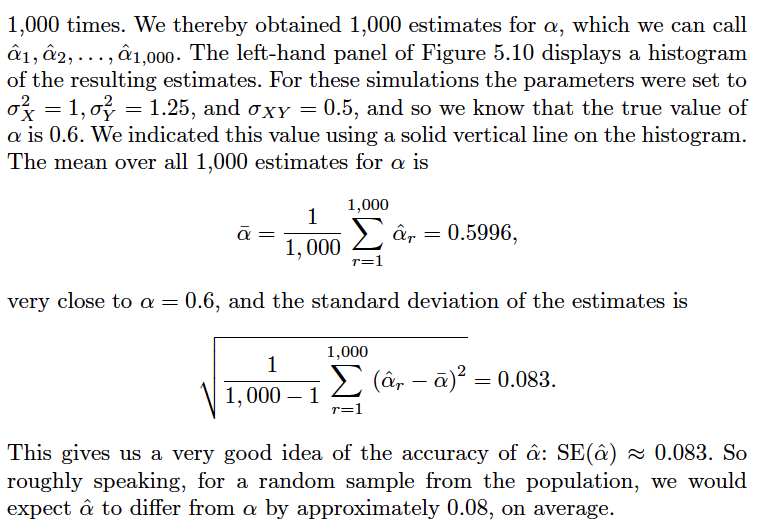
The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to ***quantify the uncertainty associated with a given estimator or statistical learning method.*** As a simple example, the bootstrap can be used to estimate the standard errors of the coefficients from a linear regression fit. the power of the bootstrap lies in the fact that it can be easily applied to a wide range of statistical learning methods, including some for which a measure of variability is otherwise difficult to obtain and is not automatically output by statistical software.

For example: \*\*Don’t go much into how the things are getting calculated but see how bootstrap is working

Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y , respectively, where X and Y are random quantities. We will invest a fraction α of our money in X , and willinvest the remaining 1 − α in Y . Since there is variability associated with the returns on these two assets, we wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize Var(αX +(1 −α )Y ). One can show that the value that minimizes the risk is given by

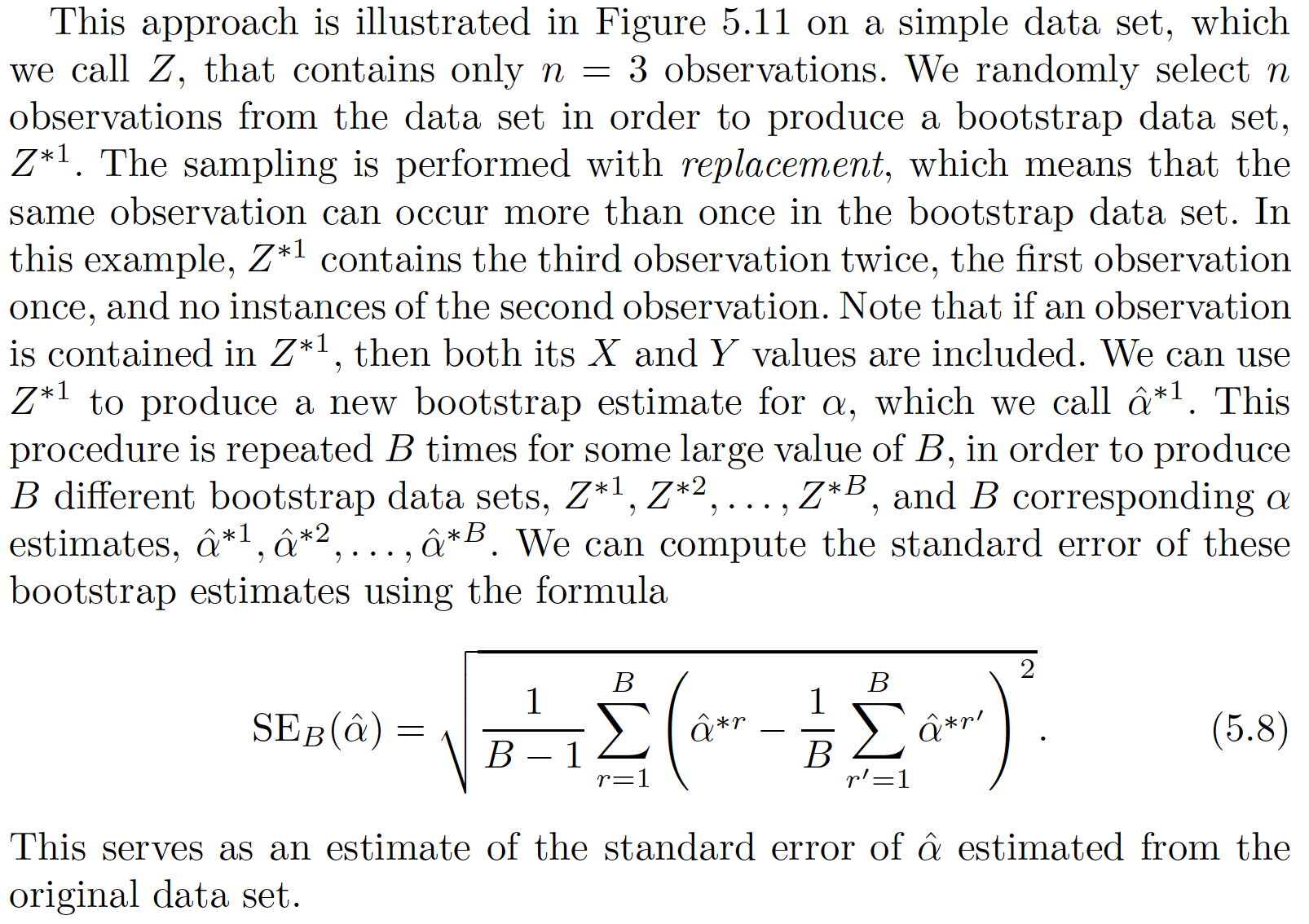


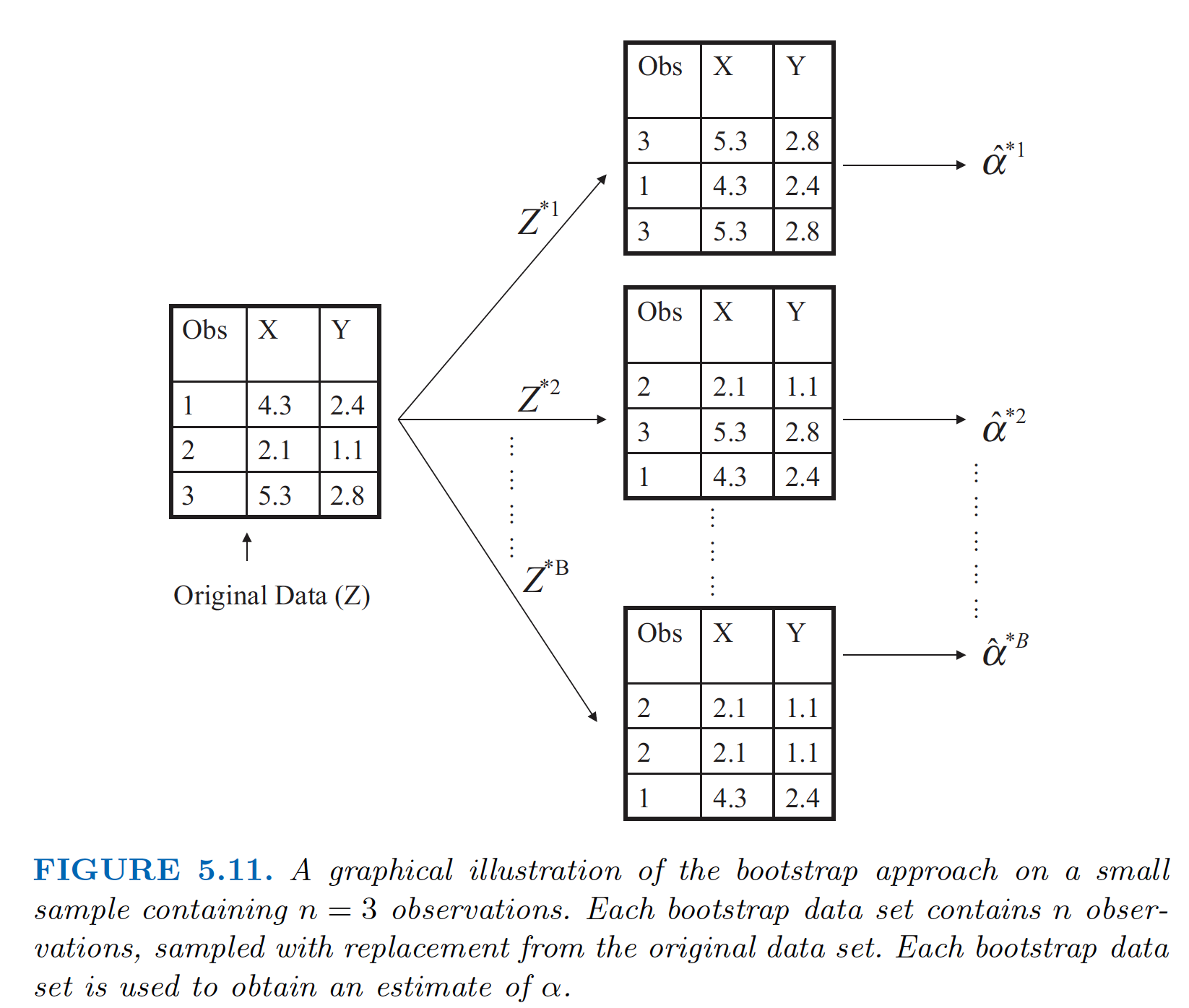
It is natural to wish to quantify the accuracy of our estimate of α . To estimate the standard deviation of ˆα , we repeated the process of simulating 100 paired observations of X and Y , and estimating α using (5.7),

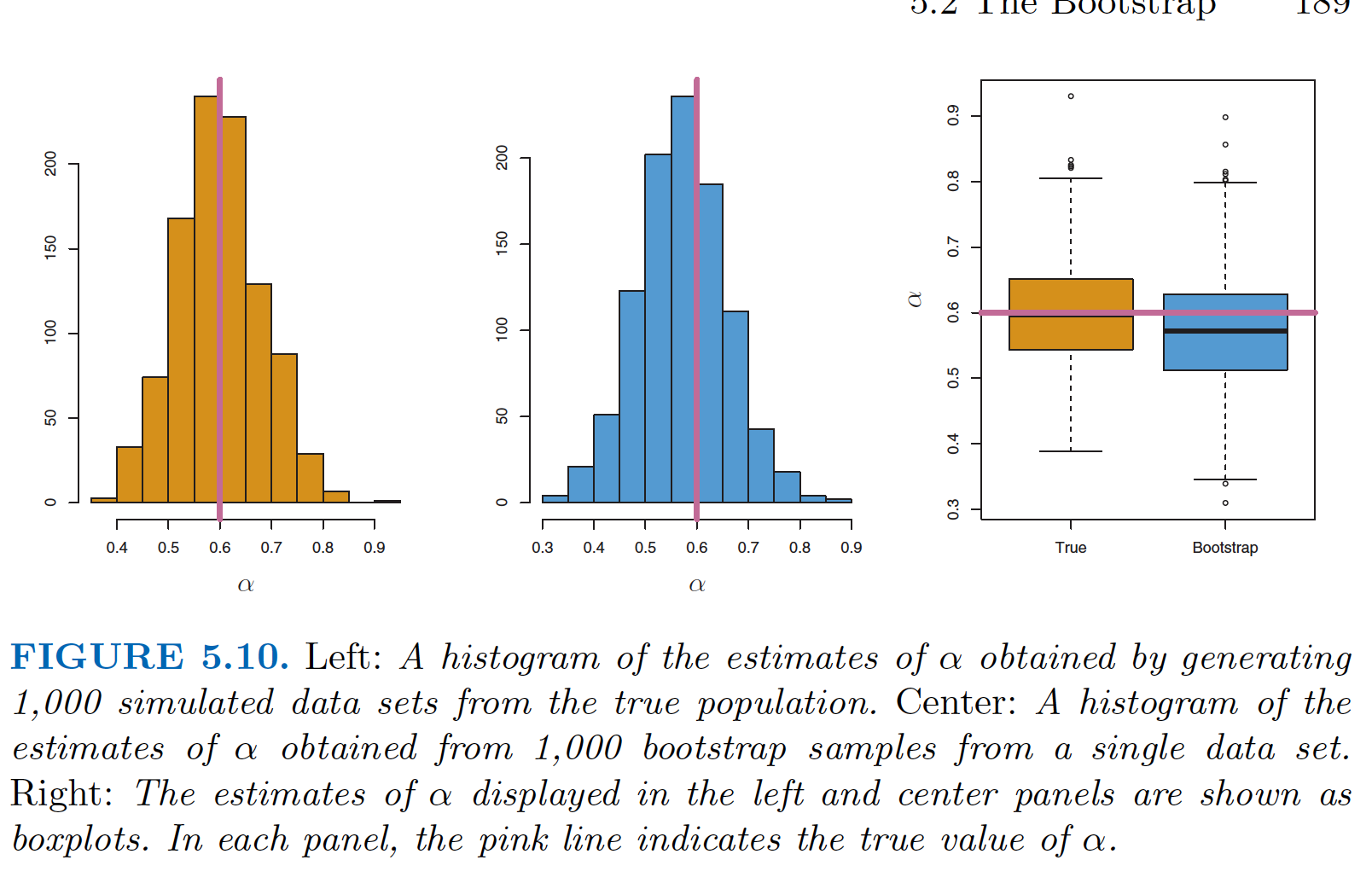


But problem is, for real data we cannot generate new samples from the original population. However, the bootstrap approach allows us to use a computer to emulate the process of obtaining new sample sets,

Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set .







The bootstrap approach is illustrated in the center panel of Figure 5.10, which displays a histogram of 1,000 bootstrap estimates of α , each computed using a distinct bootstrap data set. This panel was constructed on the basis of a single data set, and hence could be created using real data. Note that the histogram looks very similar to the left-hand panel which displays the idealized histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population. In particular the bootstrap estimate SE(ˆα ) from (5.8) is 0. 087, very close to the estimate of 0. 083 obtained using 1,000 simulated data sets. The right-hand panel displays the information in the center and left panels in a different way, via boxplots of the estimates for α obtained by generating 1,000 simulated data sets from the true population and using the bootstrap approach. Again, the boxplots are quite similar to each other, indicating that the bootstrap approach can be used to effectively estimate the variability associated with ˆα .

