Introduction to Machine Learning – Generative Models

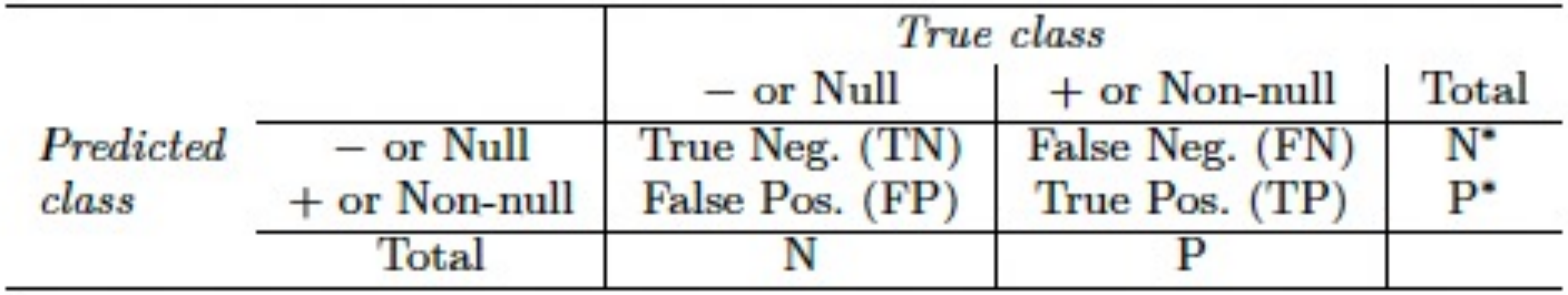
Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

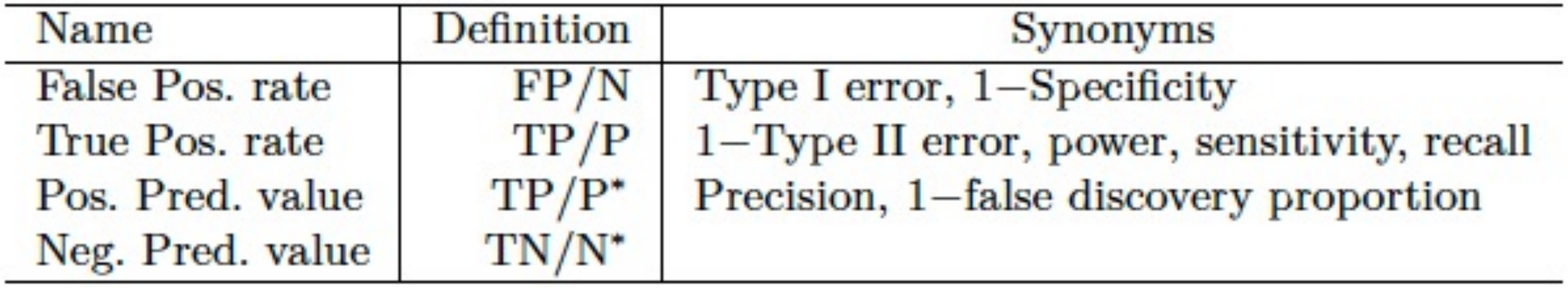
Plan for Today

- Cross Validation
- Bootstrap

Warm Up: Classification Errors



		True class		
		− or Null	+ or Non-null	Total
Predicted class	− or Null	True Neg. (TN)	False Neg. (FN)	N*
	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P*
	Total	N	P	



Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

		Predicted	
		0	1
Actual	0	34	14
	1	85	42

Calculate specificity, sensitivity, and precision for the model that produced this confusion matrix.

Resampling

Resampling involves repeatedly drawing samples from a training dataset and refitting the model of interest on each sample to obtain additional information about the fitted model

What exactly do you think this method allows us to measure?

Resampling

Resampling involves repeatedly drawing samples from a training dataset and refitting the model of interest on each sample to obtain additional information about the fitted model

- ***Model assessment*** is the process of evaluating a model's performance
- ***Model selection*** is the process of selecting the proper level of flexibility for a model

Resampling

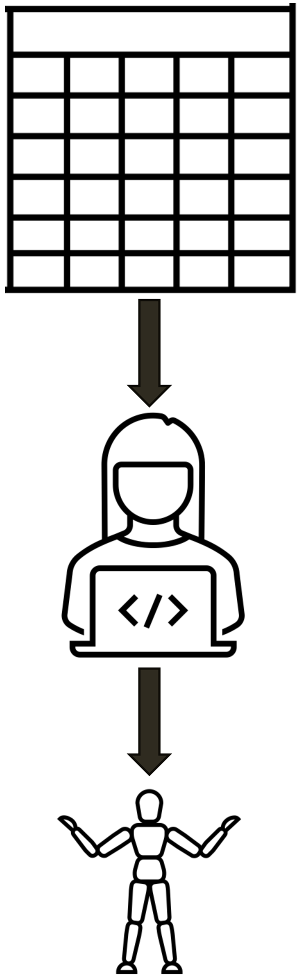
Resampling involves repeatedly drawing samples from a training dataset and refitting the model of interest on each sample to obtain additional information about the fitted model

- ***Model assessment*** is the process of evaluating a model's performance
- ***Model selection*** is the process of selecting the proper level of flexibility for a model

Cross validation is a resampling method used for model assessment and selection.

Bootstrap is a resampling method commonly used for measuring accuracy of a parameter estimate.

Fitting a Supervised ML Model



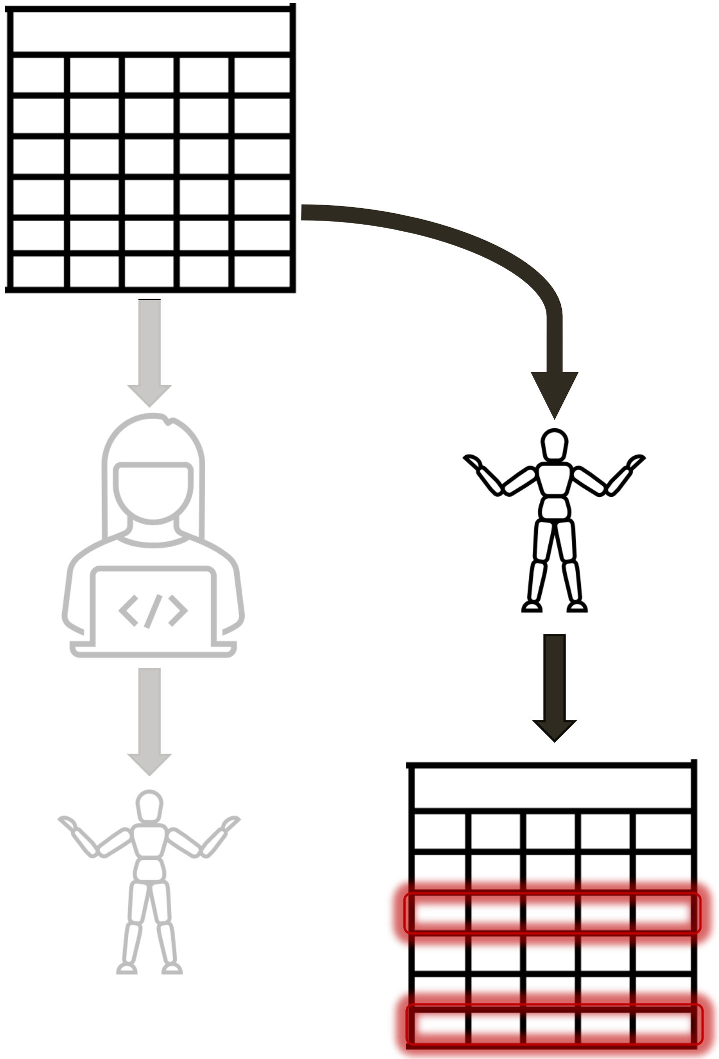
Recall: ***Supervised*** means we fit our model using data that includes the responses we are trying to predict

Therefore, we can see how well our model performs (via error) on the data we used to train it. This is called the model's ***training error***.

Ideally, we would also have new, unseen data on which to test our model. The model's performance on this data is called its ***test error***.

Created by RocketDiction
from Noun Project

Fitting a Supervised ML Model



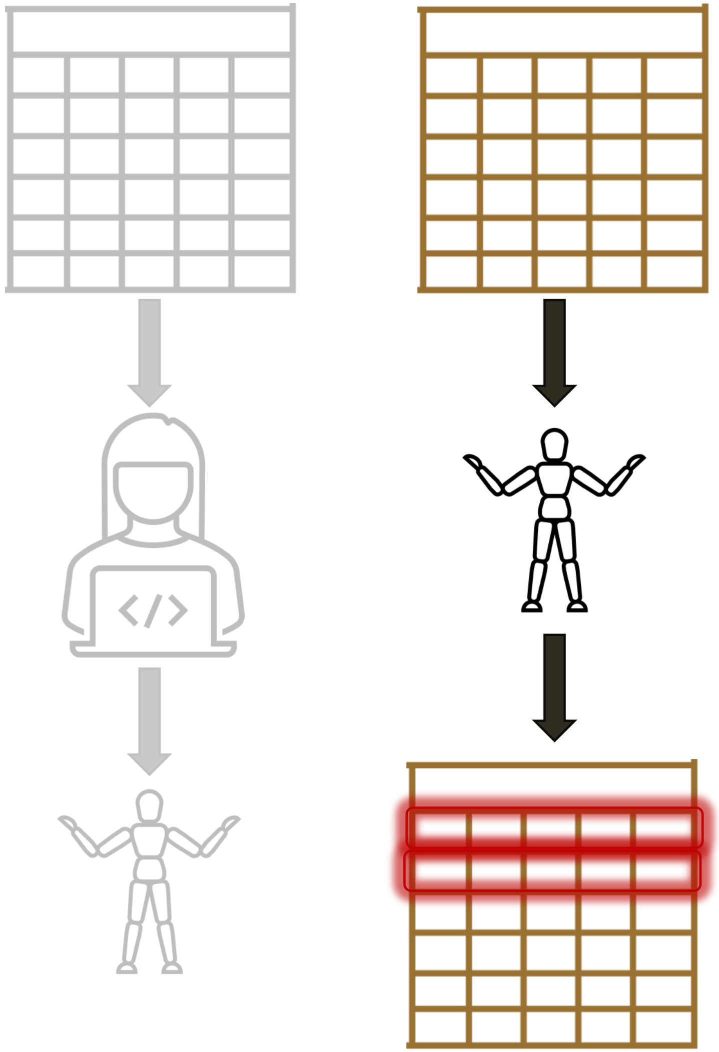
Recall: ***Supervised*** means we fit our model using data that includes the responses we are trying to predict

Therefore, we can see how well our model performs (via error) on the data we used to train it. This is called the model's ***training error***.

Ideally, we would also have new, unseen data on which to test our model. The model's performance on this data is called its ***test error***.

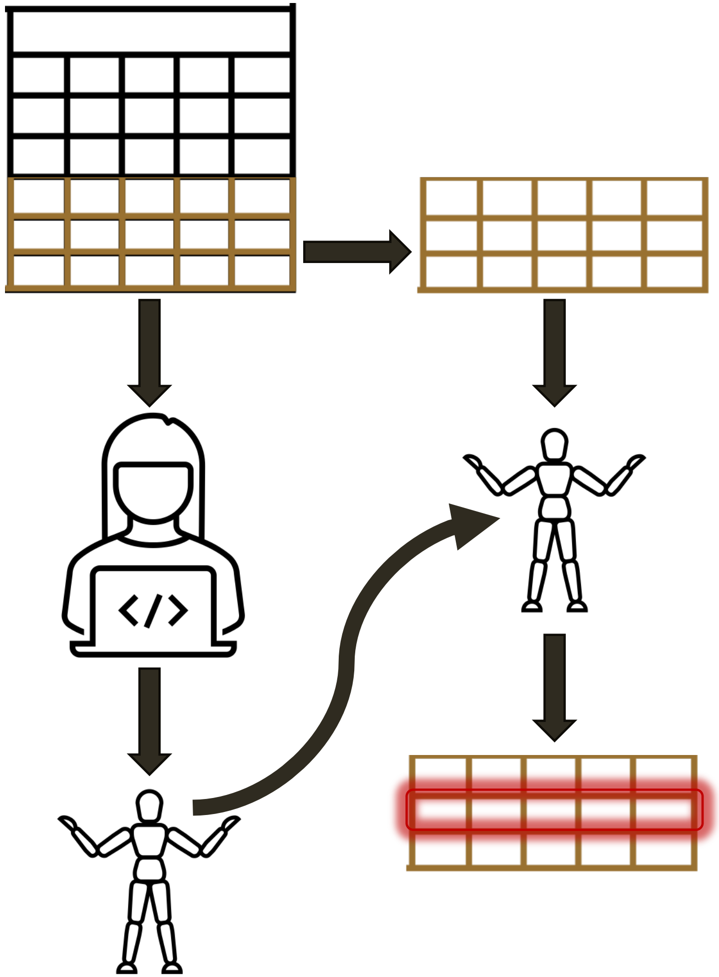
Created by RocketDiction
from Noun Project

Fitting a Supervised ML Model



Ideally, we would also have new, unseen data on which to test our model. The model's performance on this data is called its ***test error***.

Validation Set

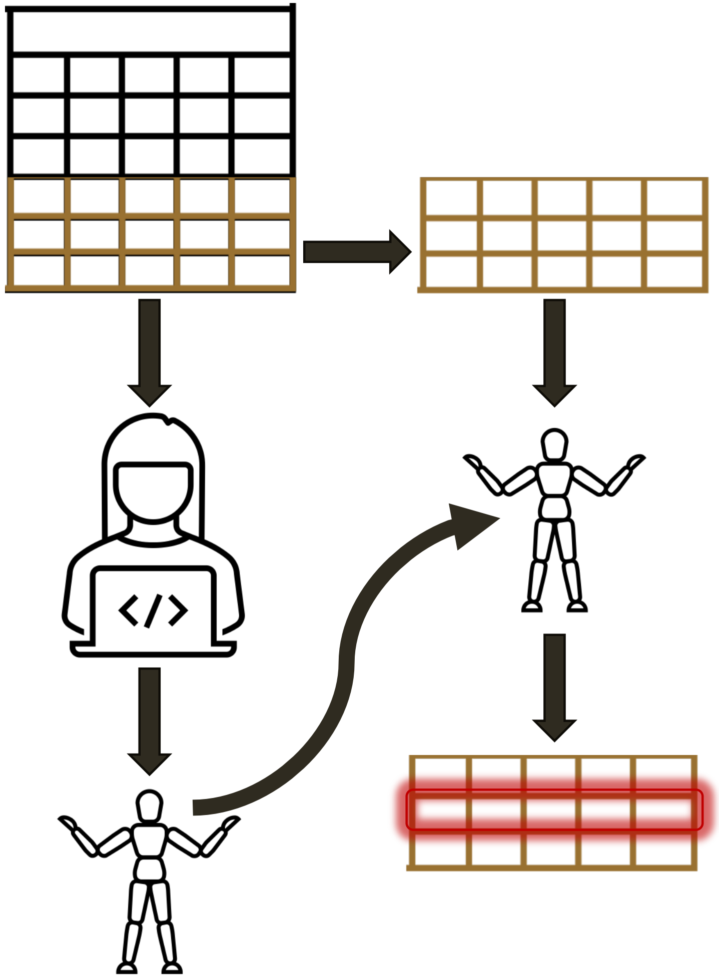


Ideally, we would also have new, unseen data on which to test our model. The model's performance on this data is called its ***test error***.

This usually isn't the case, so instead we ***hold out*** part of our training data to act as testing data.

The hold out is also called a ***validation set***, and is used to estimate test error.

Validation Set



Ideally, we would also have new, unseen data on which to test our model. The model's performance on this data is called its ***test error***.

This usually isn't the case, so instead we ***hold out*** part of our training data to act as testing data.

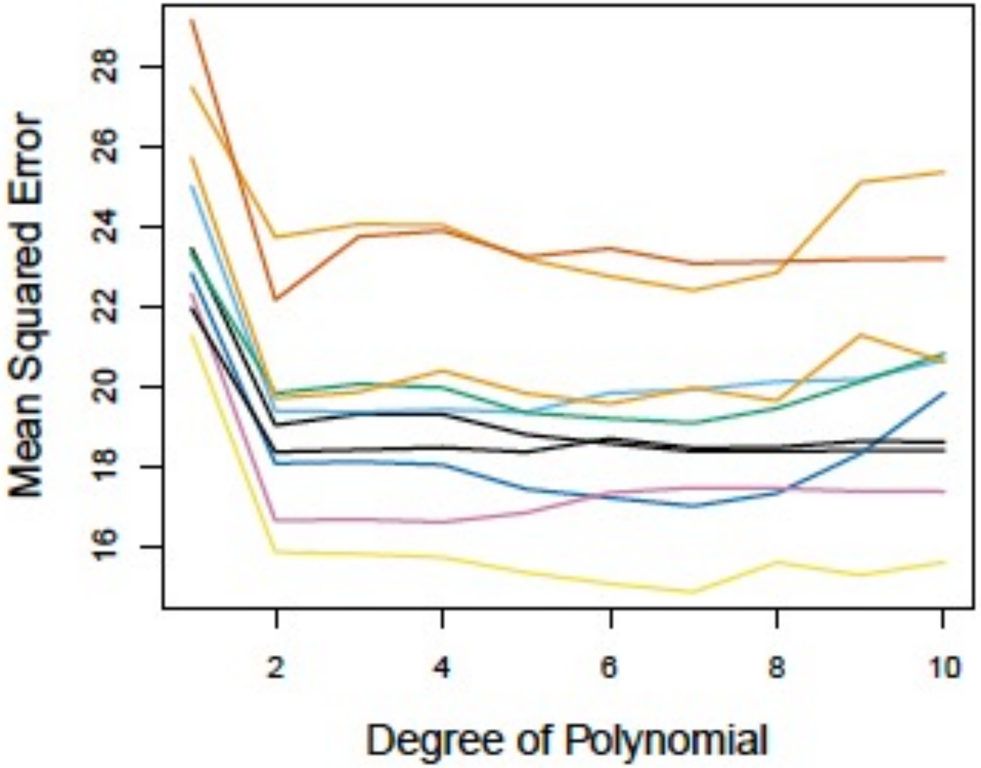
The hold out is also called a ***validation set***, and is used to estimate test error.

Would you expect the same of different error given a different validation set?

Validation Set

Different validation sets will yield different error rates.

Ex. Below are curves plotting MSE for different regression models predicting mpg on horsepower from the Auto dataset (ranging from linear to x^{10}). Each line represents the models with a different validation set.



What do these curves tell you about the appropriate regression model for this data?

Validation Set

Drawbacks of the validation set approach:

- Validation set error rate can be highly variable depending on what observations were included in the validation and training sets
- Only a subset of observations are used to fit our model. Statistical methods tend to perform worse when trained on fewer observations. Therefor, validation set error tends to overestimate test error rate for a model fit on the entire dataset.

Validation Set

Drawbacks of the validation set approach:

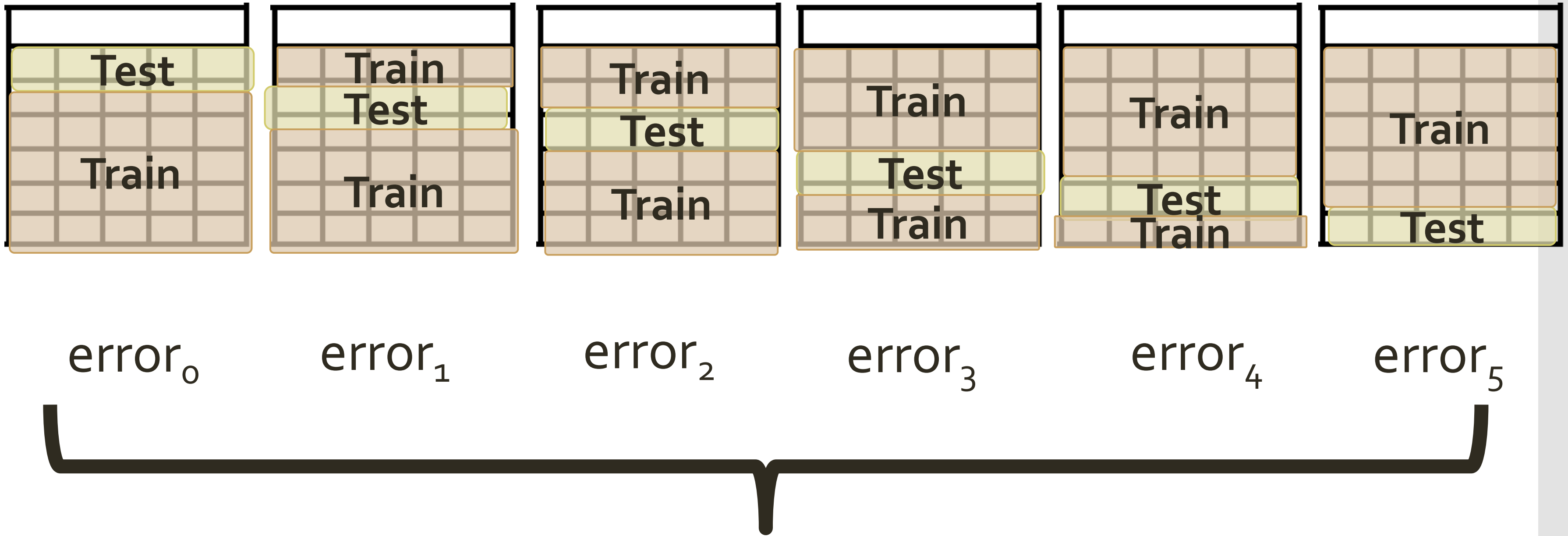
- Validation set error rate can be highly variable depending on what observations were included in the validation and training sets
- Only a subset of observations are used to fit our model. Statistical methods tend to perform worse when trained on fewer observations. Therefor, validation set error tends to overestimate test error rate for a model fit on the entire dataset.

Cross-validation refines this approach to address these issues.

Leave-One-Out Cross-Validation

LOOCV

- Uses one observation for the validation set and all remaining observations for the training set
- Repeats this approach for every observation in the dataset
- Error across all validation sets is averaged to estimate test error



$$CV_n = \frac{1}{n}\sum_{i=1}^{n} e_i$$

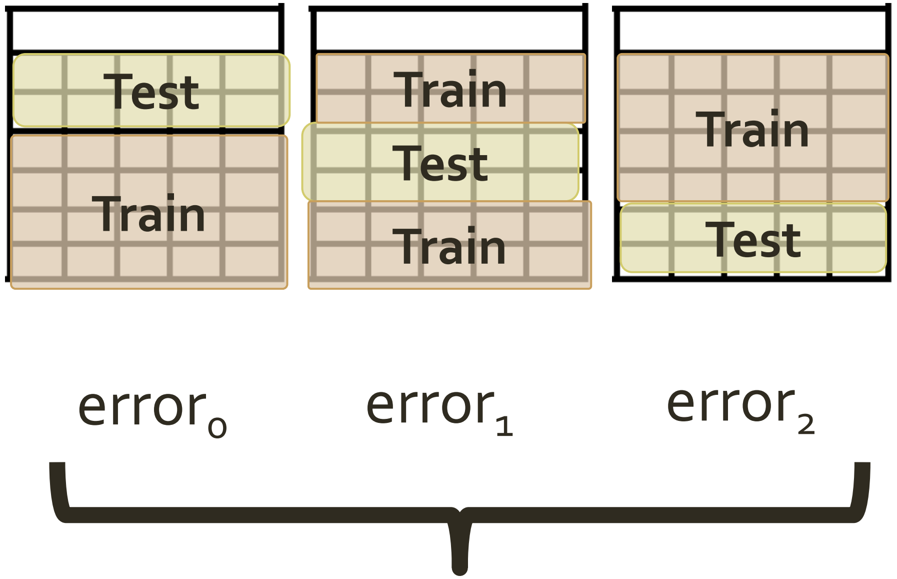
Leave-One-Out Cross-Validation

Benefits of LOOCV

Tends to not overestimate test error rate as much as the validation set approach.

There is no randomness in the training/validation splits, so we will always get the same estimated test error rate (unlike with the validation set approach)

k-Fold Cross-Validation



Test

Train

Train

Test

Train

Train

Test

error$_0$ error$_1$ error$_2$

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} e_i$$

k-Fold CV (an alternative to LOOCV)

- Randomly divides the dataset into k groups, or folds of about equal size
- Uses one fold for the validation set and all remaining folds for the training set
- Repeats this for every fold
- Error across all validation sets is averaged to estimate test error

k-Fold Cross-Validation

Benefits of k-Fold CV

As long as k is not 1, less computationally expensive than LOOCV (usually k is 5 or 10)

Estimated test error rate will vary (unlike LOOCV), but only slightly (unlike the validation set method).

Comparing LOOCV and 10-fold CV

Below are curves plotting MSE against flexibility for smoothing splines. Each chart represents a different dataset.

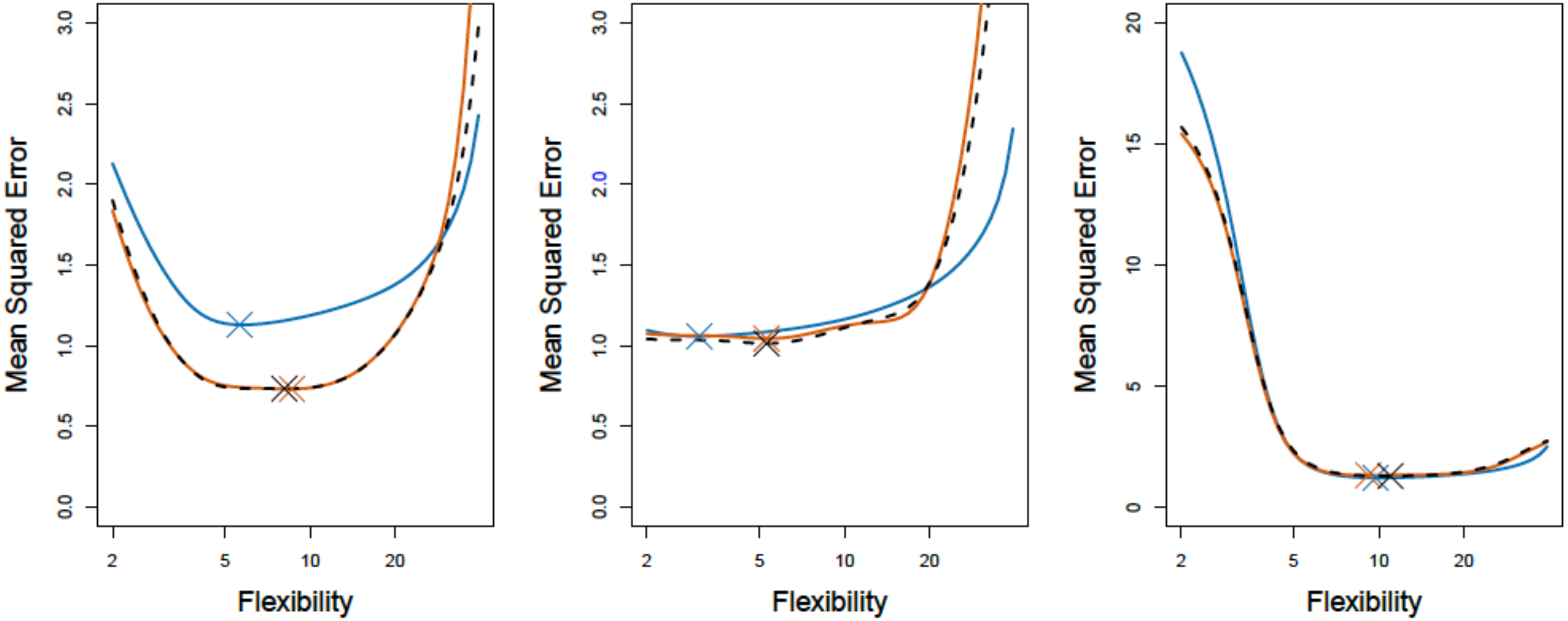
The true test MSE (from a true test set) is shown in blue.

The LOOCV estimate is shown in dashed black.

The 10-fold CV estimate is shown in orange.

The X's represent minimum MSE.

What do you notice across these charts?



Bias-Variance Tradeoff

- The validation set approach will train on about half of observation, $\frac{n}{2}$
- The LOOCV approach will train on almost all observations, $n - 1$
- The k-fold CV approach will train on about $\frac{(k-1)n}{k}$ observations

Which of these trains on the most observations? The least?

As a result, which model would you expect to have the most bias? The least?

Bias-Variance Tradeoff

- The LOOCV approach will train on almost all observations, $n - 1$
- The k-fold CV approach will train on about $\frac{(k-1)n}{k}$ observations

Which of these will produce the most correlated error estimates (on each run)? The least?

As a result, which model would you expect to have the most variance? The least?

Bias-Variance Tradeoff

k-fold CV is usually performed with 5 or 10 folds because empirically these values have shown to yield test error rate estimates that balance the bias-variance tradeoff.

Bootstrap

Bootstrap is a statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.

Bootstrap

Bootstrap is a statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.

Motivation:

Suppose we want to invest a fixed sum of money into two financial assets that yield (variable) returns of X and Y, respectively. We will invest a fraction, α, of our money in X and the rest in Y.

We want to choose α to minimize the total variance of our investment.

Bootstrap

Bootstrap is a statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.

Motivation:

Suppose we want to invest a fixed sum of money into two financial assets that yield (variable) returns of X and Y, respectively. We will invest a fraction, α, of our money in X and the rest in Y.

We want to choose α to minimize the total variance of our investment.

In other words, we want to minimize:

$$Var(\alpha X + (1 - \alpha)Y)$$

Bootstrap

Suppose we want to invest a fixed sum of money into two financial assets that yield (variable) returns of X and Y, respectively. We will invest a fraction, α, of our money in X and the rest in Y.

We want to choose α to minimize the total variance of our investment.

In other words, we want to minimize:

$$Var(\alpha X + (1 - \alpha)Y)$$

To do that we want

$$\alpha = \frac{\sigma_Y^2 \; - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

Bootstrap

Suppose we want to invest a fixed sum of money into two financial assets that yield (variable) returns of X and Y, respectively. We will invest a fraction, α, of our money in X and the rest in Y.

We want to choose α to minimize the total variance of our investment.

In other words, we want to minimize:

$$Var(\alpha X + (1-\alpha)Y)$$

To do that we want

$$\alpha = \frac{\sigma_Y^2 \ - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

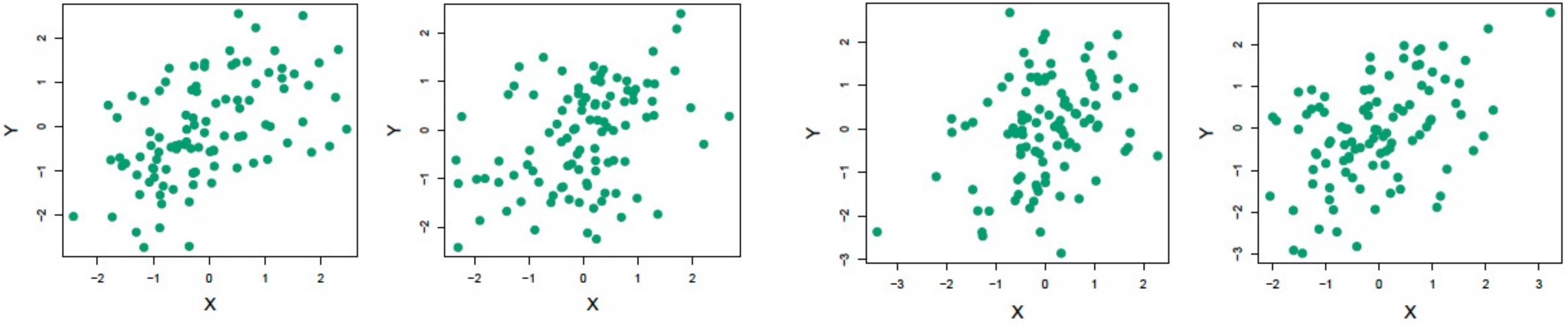
Using past data to estimate, we get an ideal α of

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 \ - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

Bootstrap

What if we want to quantify the accuracy of our estimate for α?

Let’s use simulated data (based on the true data) to get additional estimates for α. In this simulation, $\sigma_X^2 = 1, \sigma_Y^2 = 1.25, \sigma_{XY} = 0.5$. (So the true $\alpha = 0.6$)



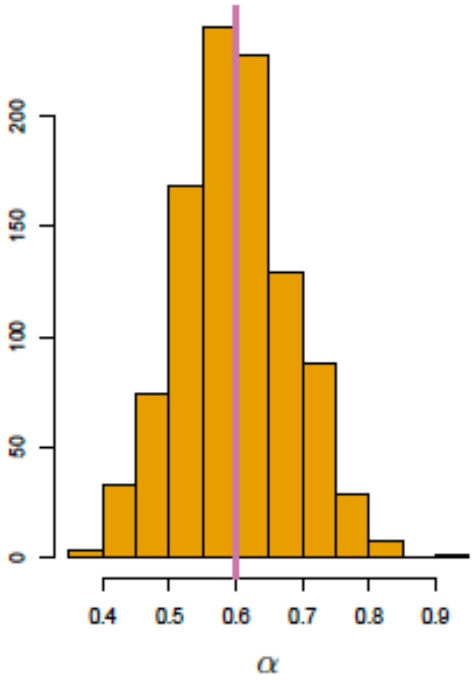
$\alpha = 0.576$ $\alpha = 0.576$ $\alpha = 0.576$ $\alpha = 0.576$

Each plot shows 100 simulated data points.

Bootstrap

If we repeat the simulation 1,000 times we get an average α of 0.5996, with a standard deviation of 0.083, which equates to $SE(\hat{\alpha}) \approx 0.083$.

On average, we would expect $\hat{\alpha}$ to differ from α by about 0.08.

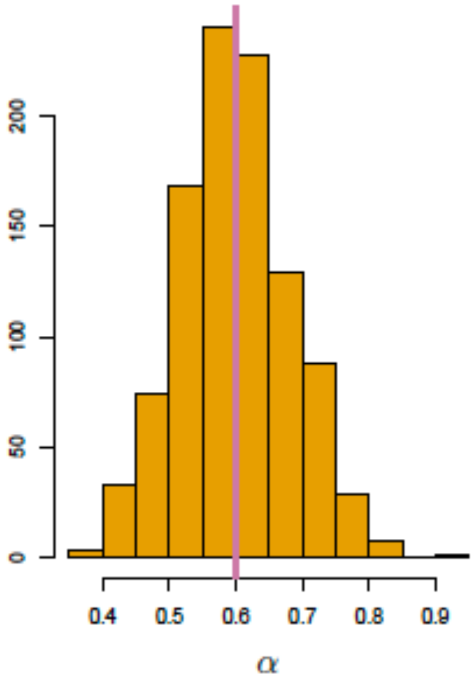


histogram from
1000 simulations

Bootstrap

If we repeat the simulation 1,000 times we get an average α of 0.5996, with a standard deviation of 0.083, which equates to $SE(\hat{\alpha}) \approx 0.083$.

On average, we would expect $\hat{\alpha}$ to differ from α by about 0.08.



histogram from 1000 simulations

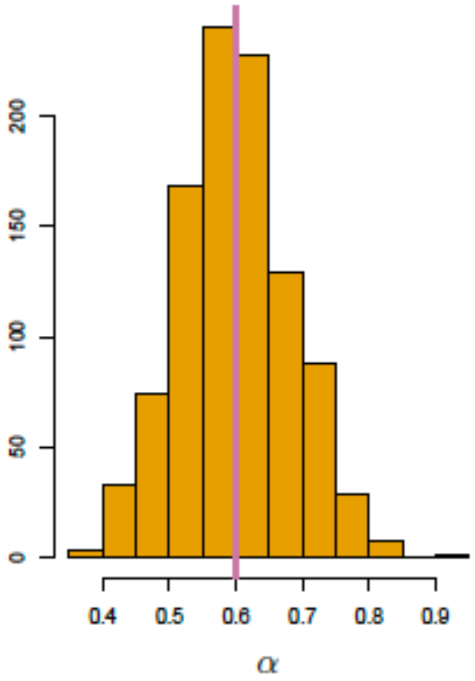
In practice, we cannot use this procedure for estimating SE, because for real data we cannot generate new samples from the original population.

Bootstrap allows us to leverage computing to emulate this process of obtaining new sample sets so that we can estimate variability of $\hat{\alpha}$.

Bootstrap

If we repeat the simulation 1,000 times we get an average α of 0.5996, with a standard deviation of 0.083, which equates to $SE(\hat{\alpha}) \approx 0.083$.

On average, we would expect $\hat{\alpha}$ to differ from α by about 0.08.



histogram from 1000 simulations

In practice, we cannot use this procedure for estimating SE, because for real data we cannot generate new samples from the original population.

Bootstrap allows us to leverage computing to emulate this process of obtaining new sample sets so that we can estimate variability of $\hat{\alpha}$.

Instead of obtaining independent data sets from the population, we obtain distinct sets by repeatedly sampling (with replacement) observations from the original dataset.

Bootstrap

Procedure:

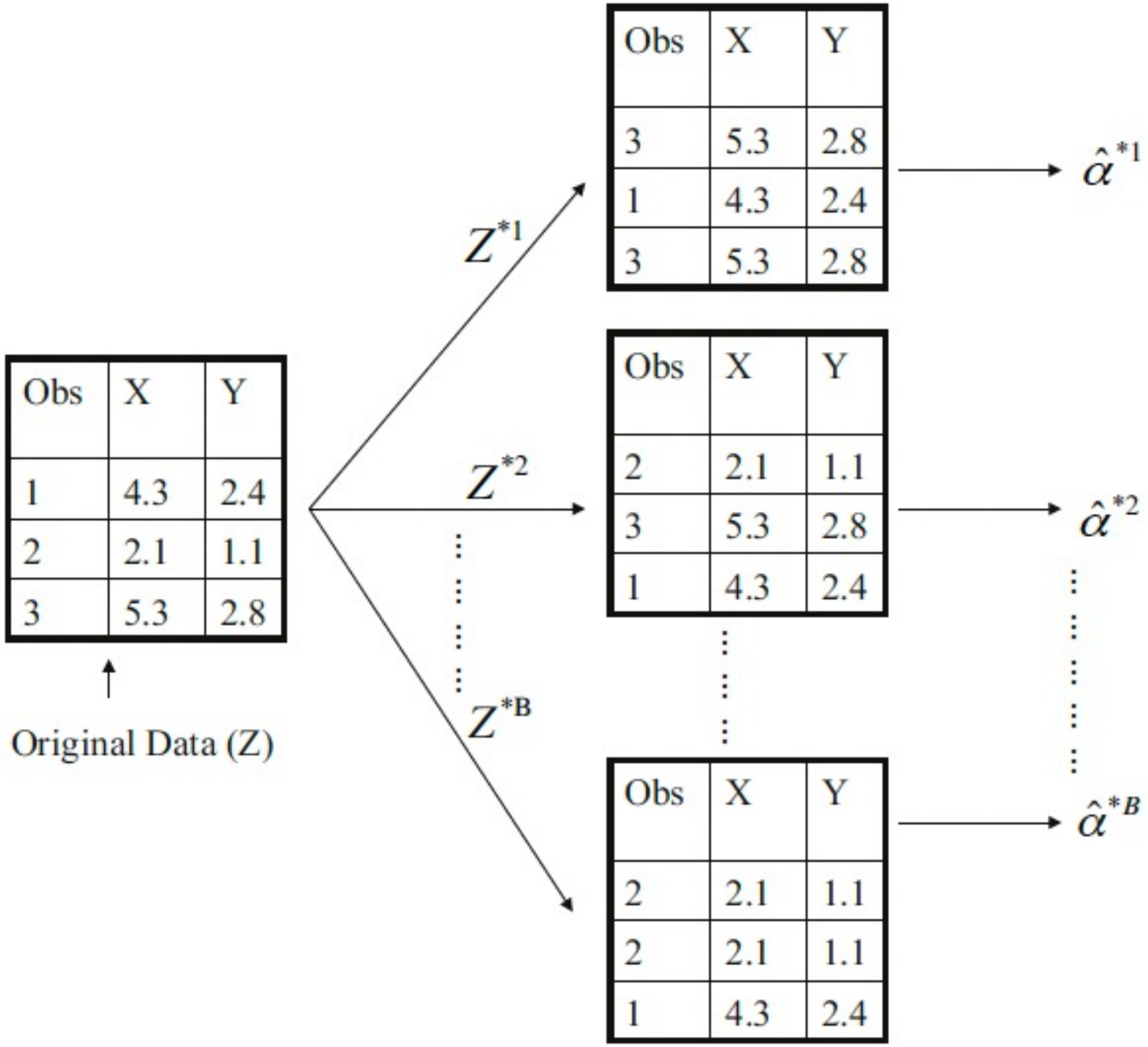
Given a dataset, Z, with n observations

- Randomly sample with replacement n observations to get a bootstrap dataset, Z^{*1}
- Use Z^{*1} to produce a bootstrap estimate, α^{*1}
- Repeat these steps B times (for some large value of B)
- Then

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}$$

This is an estimate of the standard error of α estimated from the original dataset

Bootstrap



Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Original Data (Z)

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

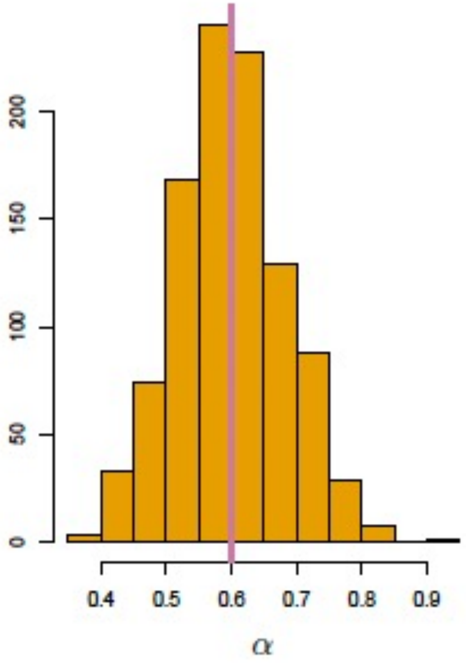
Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

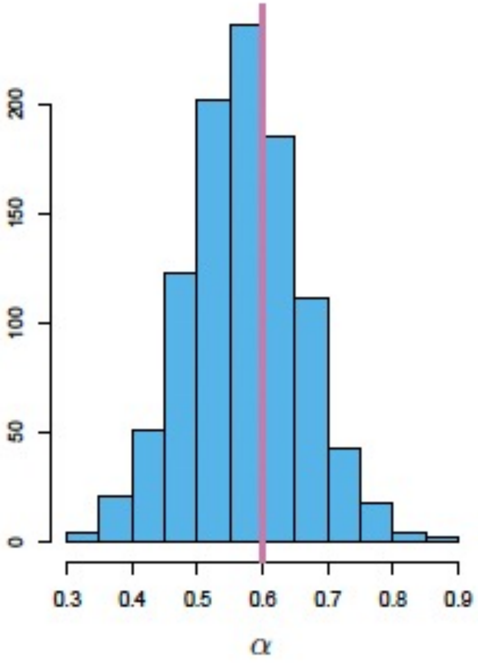
Bootstrap

If we repeat the simulation 1,000 times we get an average α of 0.5996, with a standard deviation of 0.083, which equates to $SE(\hat{\alpha}) \approx 0.083$.

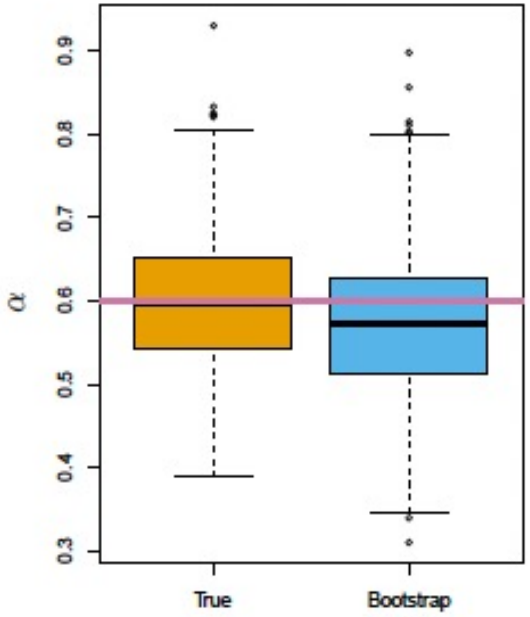
On average, we would expect $\hat{\alpha}$ to differ from α by about 0.08.



histogram from 1000 simulations



histogram from 1000 bootstrap samples from a single dataset


α
True
Bootstrap