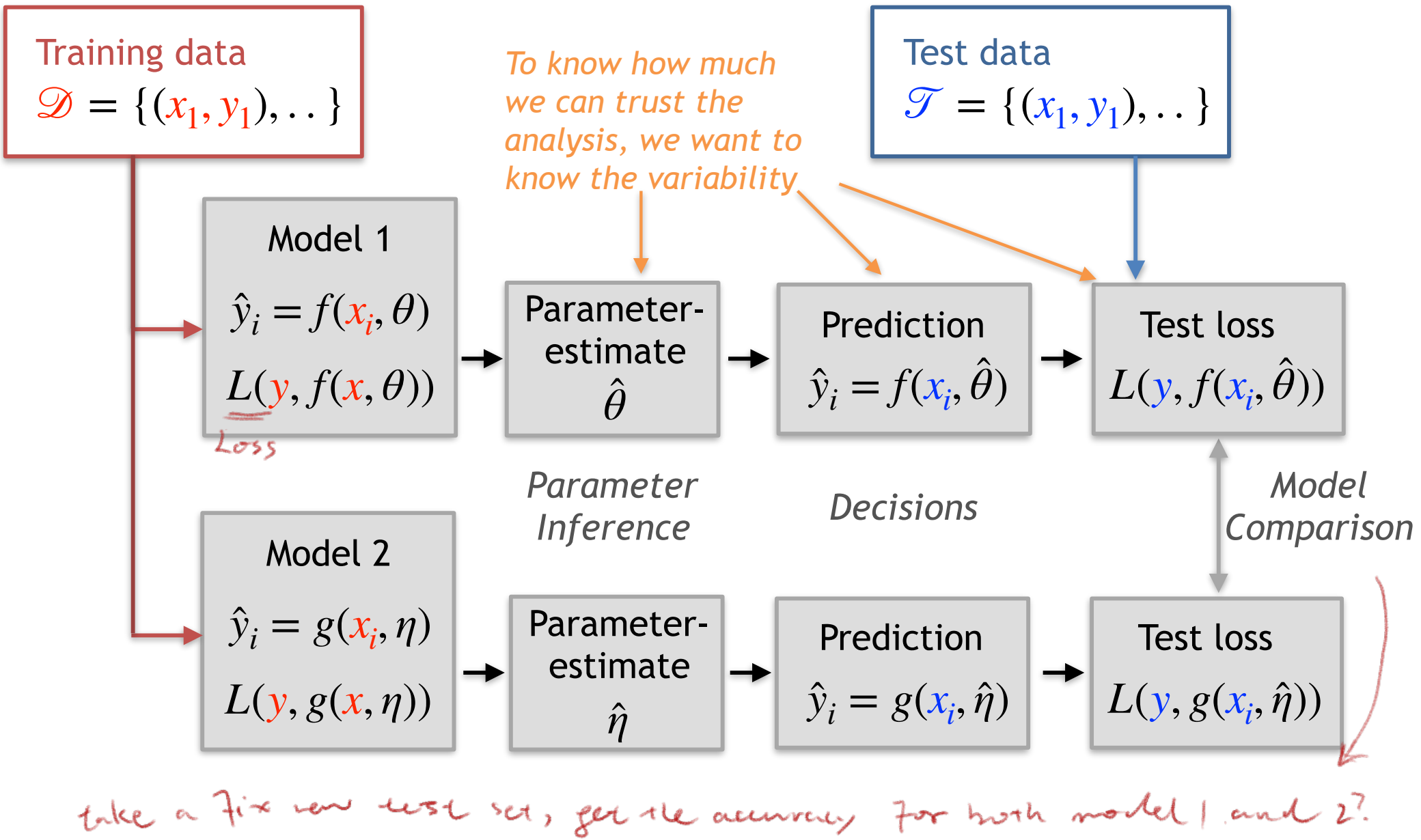
Quantifying Uncertainty

Week 4
Quantifying uncertainty

Week 4.1
Introduction

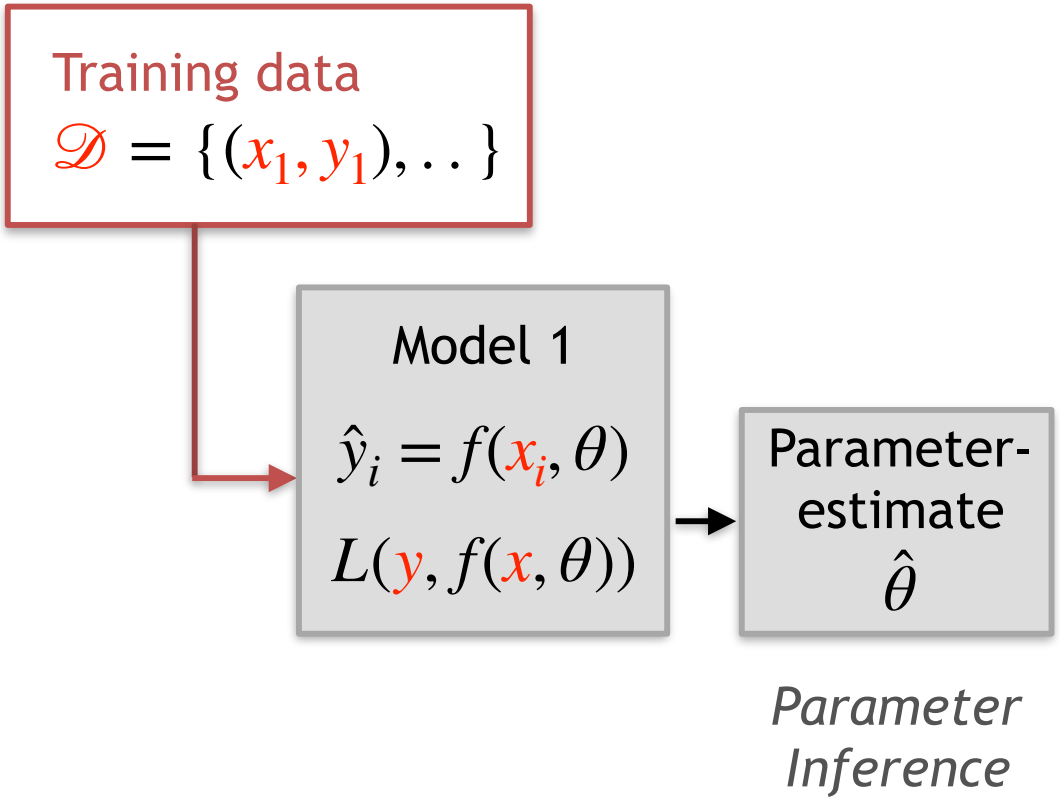
Supervised learning - overview



No! Test data is just a sample, so it could not take all circumstances into account.

Week 4.2
Parameter uncertainty, estimation, and sampling distribution

Parameter uncertainty



Statistics, Parameters, and Estimation

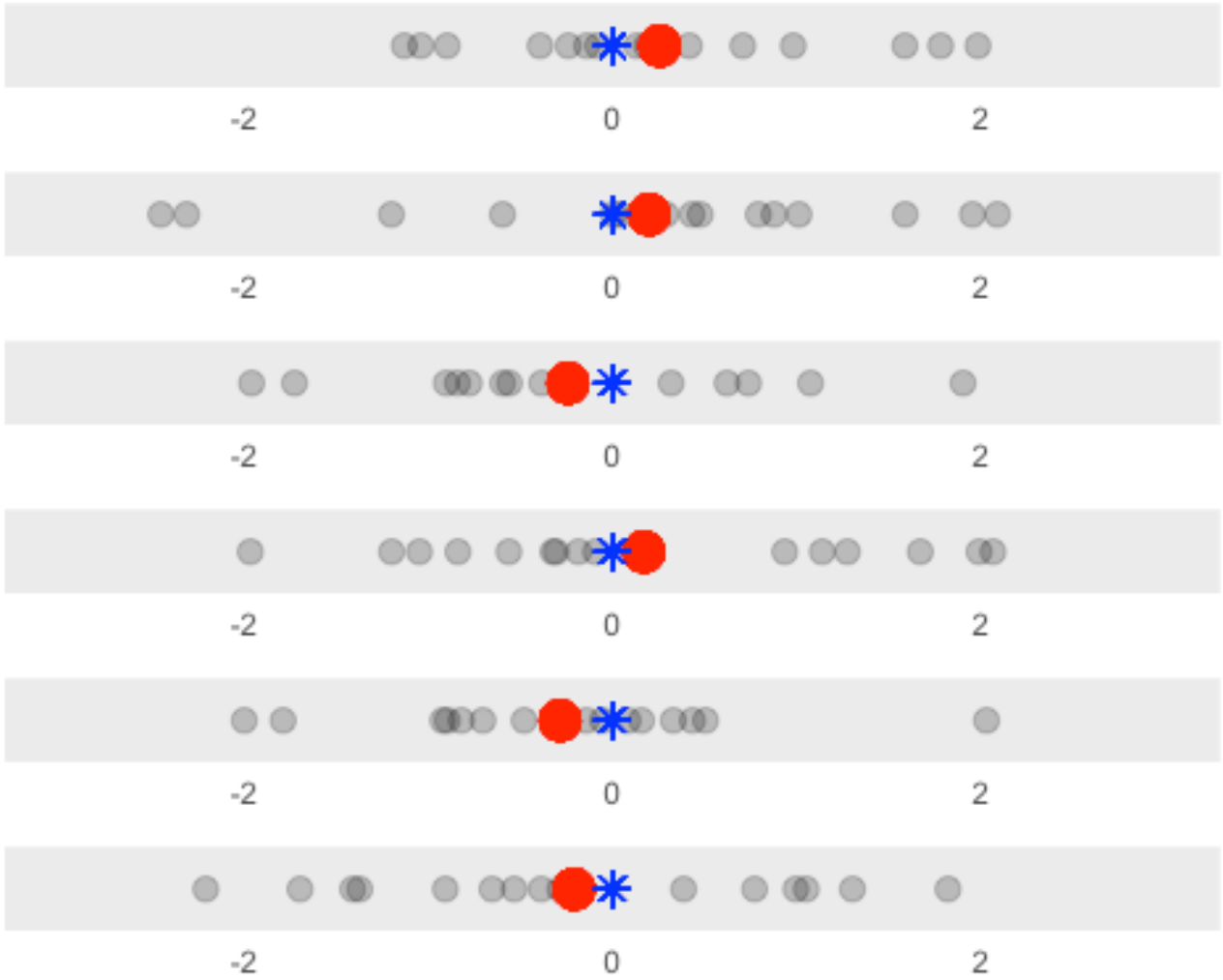
- A ***parameter*** is a value that characterizes the population that we want to study. Parameters are often expectations (like the mean) or values that describe the relationship between input and output (slope of a linear model).
- A ***statistic*** is any summary of a dataset. (E.g. $\overline{x}_n$)
  A statistic is the result of a ***function*** applied to a dataset.
- ***Estimation*** uses a ***statistic*** (e.g. $\overline{x}_n$) to estimate a ***parameter*** (e.g. μ_X) of the ***distribution*** of a ***random variable***.
  - ***Estimate***: value obtained from a specific dataset
  - ***Estimator***: function (e.g. “sum-and-divide by *n*” or “compute the maximum likelihood estimate”) used to compute the estimate
  - ***Estimand***: parameter of interest

A simple parameter: mean

- A simple model is one that predicts the mean $\hat{y} = \theta$
- Given a ***dataset*** (collection of realizations) $x_1, x_2, \ldots, x_n$ of X, the estimate for this parameter is the **sample mean**:
- $$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$$
- Given a dataset, $\hat{\theta}$ is a fixed number.
- Across possible randomly drawn dataset of size n, the mean is a ***random variable*** $\bar{X}_n$

Datasets and sample means

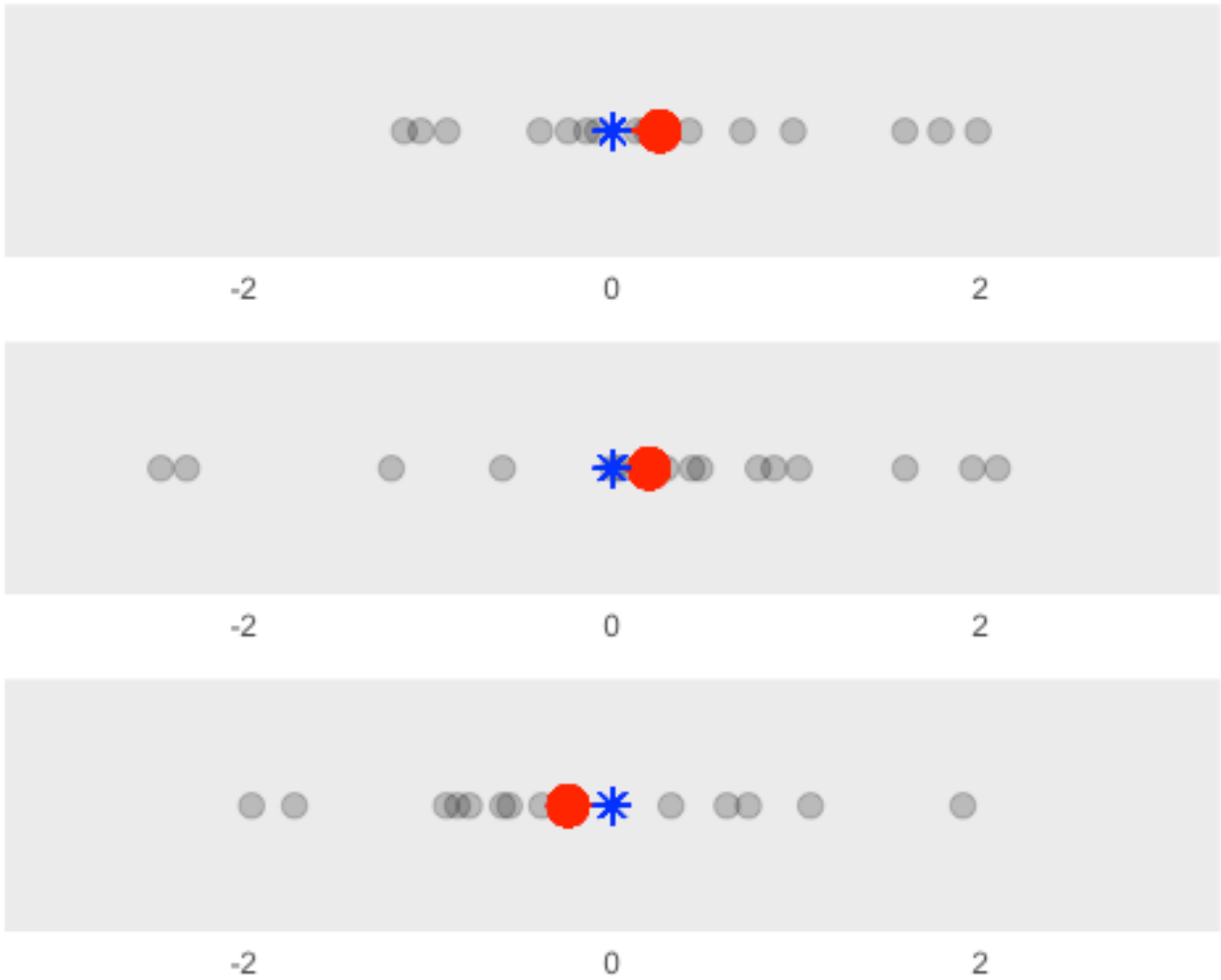
Datasets of size $n = 15$, sample means plotted in red. Truth is blue star.



Sampling Distributions

Given an estimate, how good is it?

The distribution of an estimator is called its ***sampling distribution***.



Bias

- The **expected difference** between estimator ($\hat{\theta}$) and estimand/parameter (θ). For example,

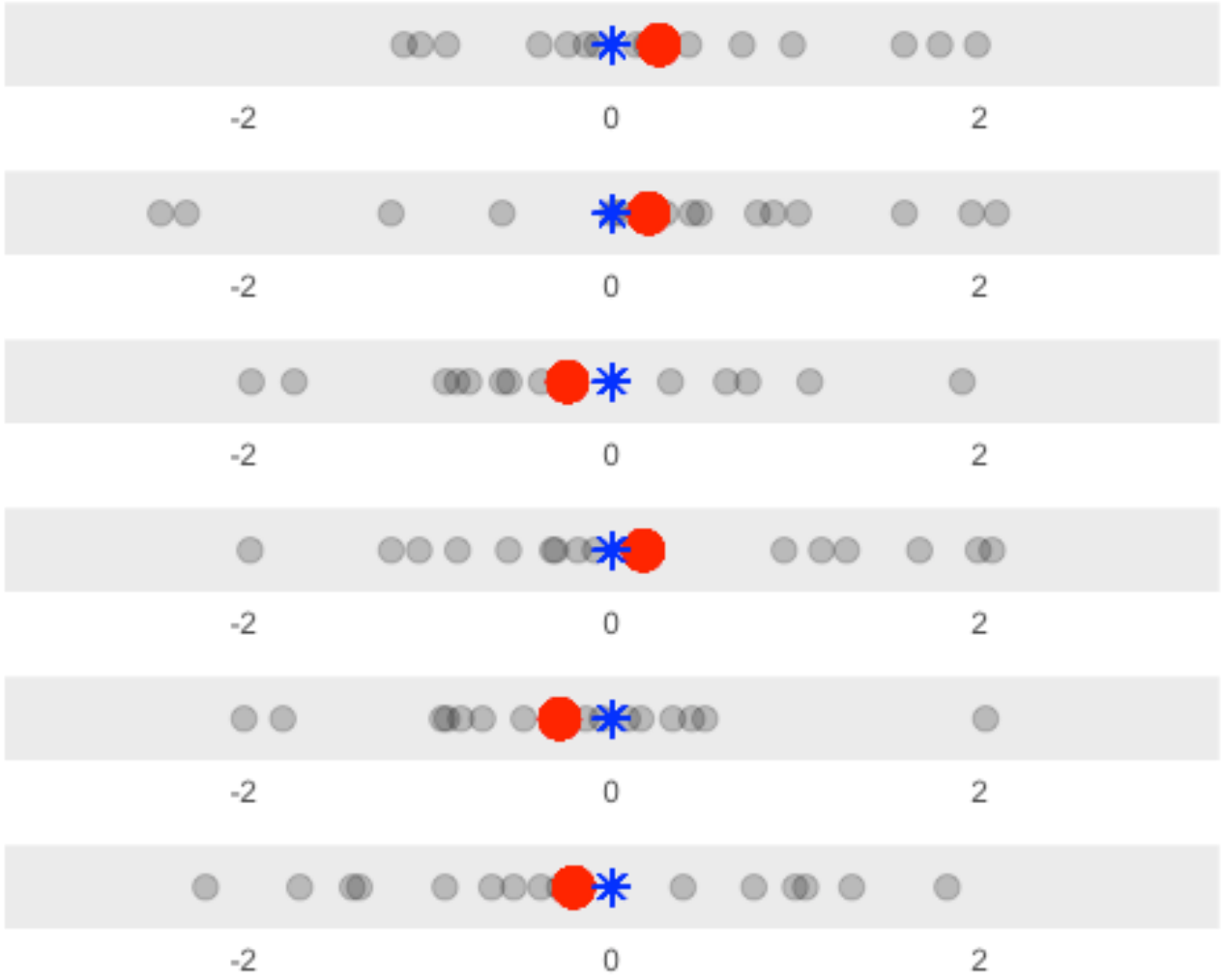
- $E[\overline{X}_n - \mu_X]$
  - Note: by convention $\mu_X = E[X]$, the mean of r.v. X.
  - If 0, estimator is **unbiased**.

- Sometimes, $\overline{x}_n > \mu_X$, sometimes $\overline{x}_n < \mu_X$, but the long run average of these differences will be zero.

Variance

- The **expected squared difference** between estimator and its mean. For example,
- $E[(\overline{X}_n - E[\overline{X}_n])^2]$
  - Positive for all non-trivial estimators. Higher variance means distribution of estimates is more "spread out."
- Because $\overline{X}_n$ is unbiased, we can write $E[(\overline{X}_n - \mu_X)^2]$
  - Sometimes, $\overline{x}_n > \mu_X$, sometimes $\overline{x}_n < \mu_X$, but the ***squared differences*** are all positive and do not cancel out.
- *Standard deviation* is the square root of the variance.

Assessing the precision of an estimate



How can we get an estimate of our sampling distribution?

Week 4.3
Confidence intervals via Central Limit theorem

Central Limit Theorem

- Informally: For *many different distributions* of X, the sampling distribution of $\overline{X}_n$ is approximately normal if n is big enough.
- More formally, for X with finite variance:

$$\bar{x}_n \sim N\left(\mu, \sigma^2_{\bar{X}_n}\right)$$

where

$$\sigma_{\overline{X}_n} = \frac{\sigma_X}{\sqrt{n}}$$

is called the ***standard error of*** $\overline{X}_n$ and σ^2_X is the variance of X.

CLT Consequence

95% of the mass of a normal with mean μ and standard deviation σ is between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$

So, 95% of the time, the sample mean $\overline{X}_n$ will be between

$$\mu - 1.96\frac{\sigma_X}{\sqrt{n}} \text{ and } \mu + 1.96\frac{\sigma_X}{\sqrt{n}}$$

So, 95% of the time, the true mean μ will be between

$$\bar{x}_n - 1.96\frac{\sigma_X}{\sqrt{n}} \text{ and } \bar{x}_n + 1.96\frac{\sigma_X}{\sqrt{n}}$$

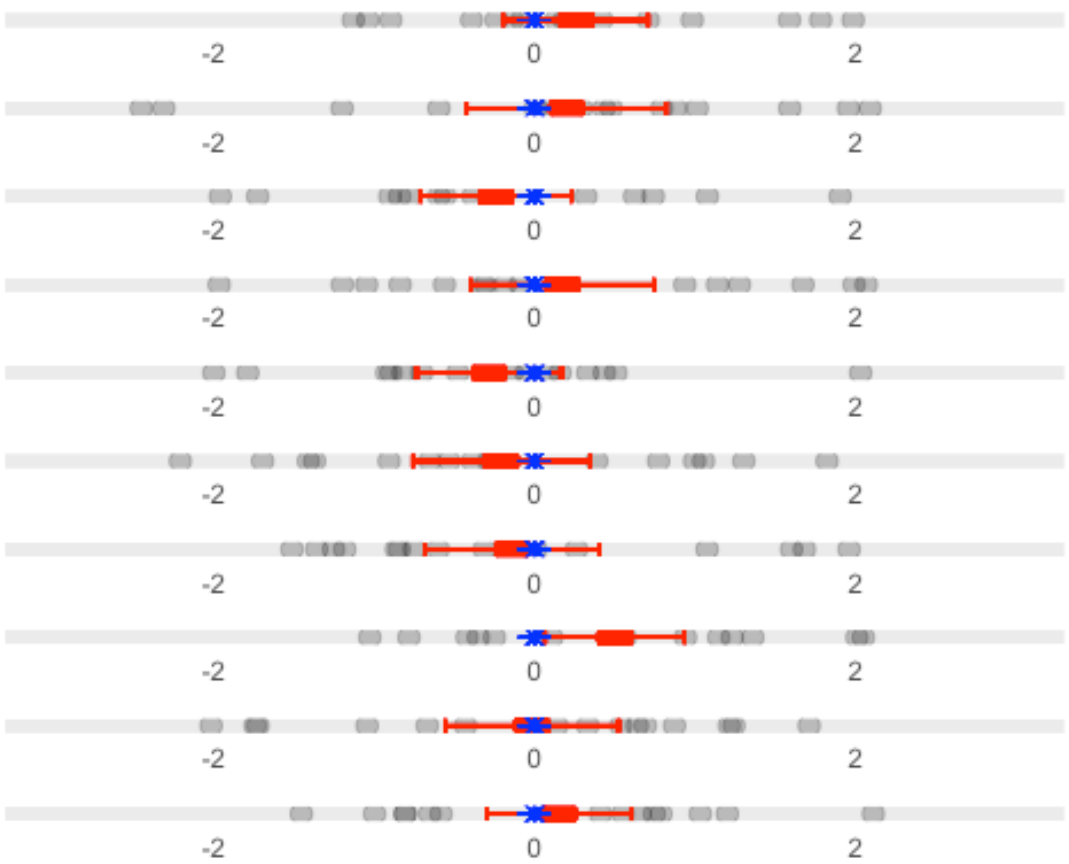
Confidence Intervals

Typically, we specify ***confidence*** given by $1 - \alpha$

Use the sampling distribution to get
an interval that traps the parameter (estimand) with probability $1 - \alpha$.

(We almost always use $1 - \alpha = 0.95$)

What a Confidence Interval Means



On average 95 out of 100 confidence intervals made from sample should include the real mean

Quantifying Precision

Eruptions dataset has $n = 272$ observations.

Our estimate of the mean of eruption times is $\overline{x}_{272}$ = 3.4877831.

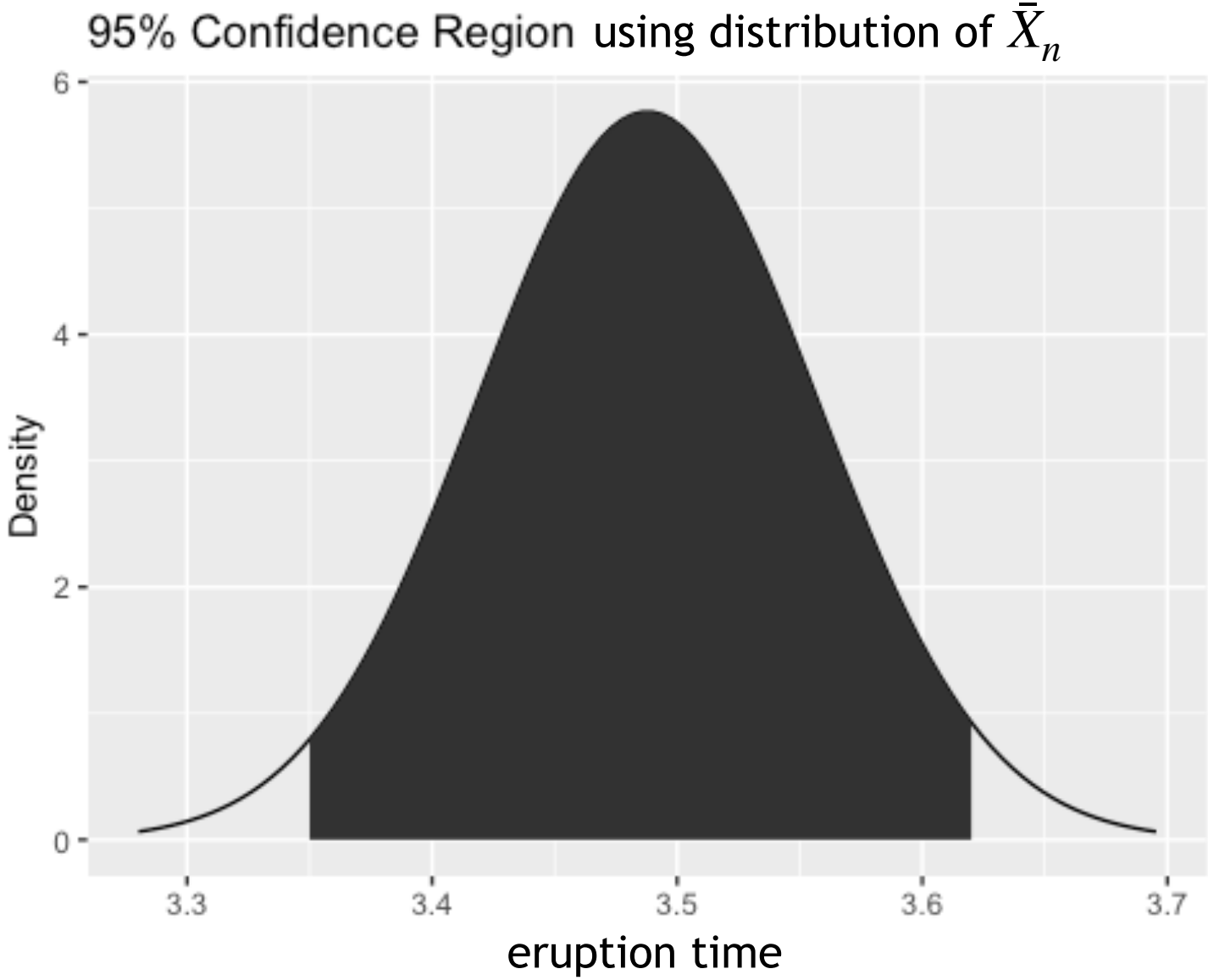
How good is this estimate?

Constructing a Confidence Interval

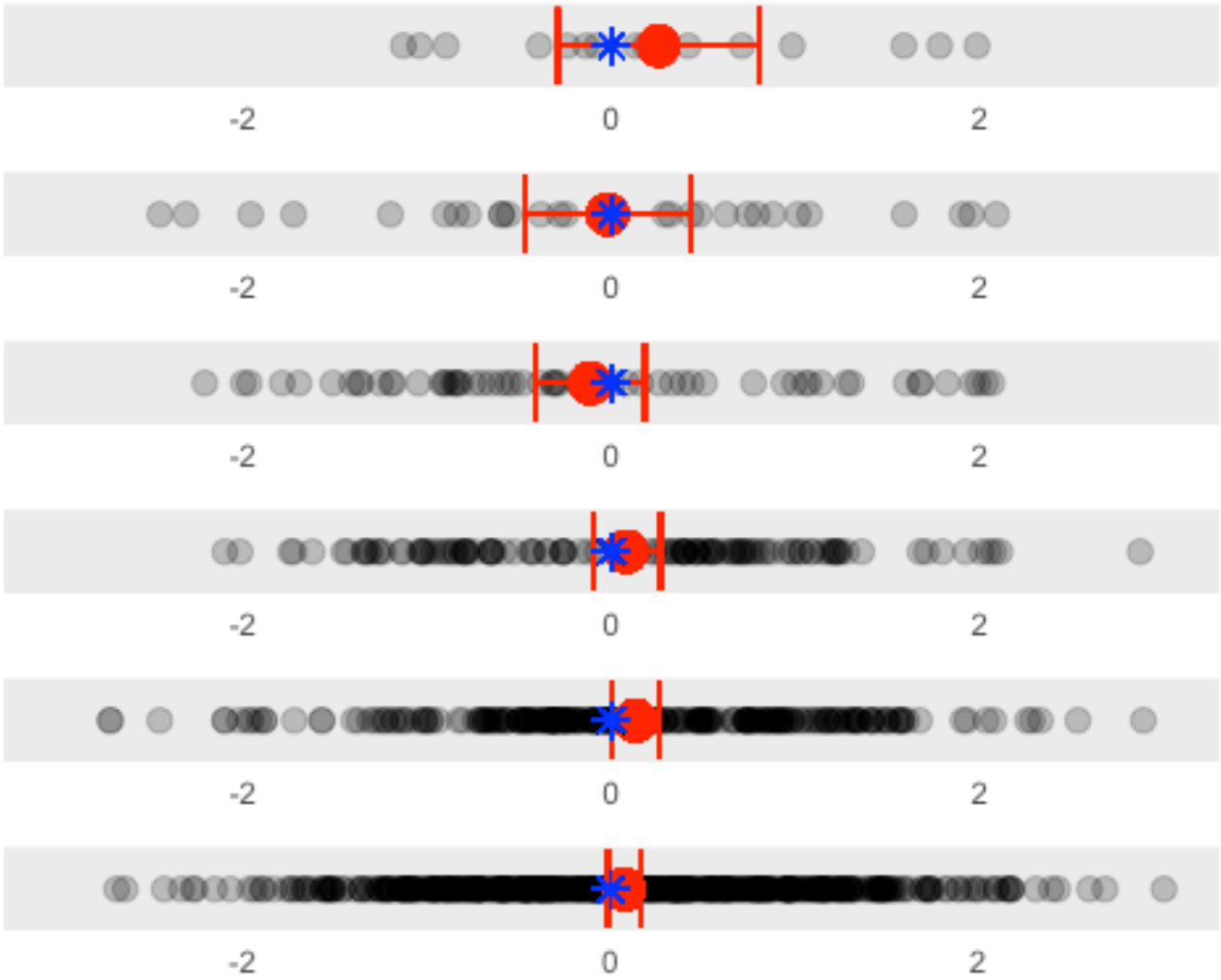
Mean = 3.49
StdDev = 1.14
n = 272

SE = $1.14/\sqrt{272}$
= 0.07

CI = 3.49 +- 1.96*0.07
CI = 3.49 +- 0.14



Effect of n on width

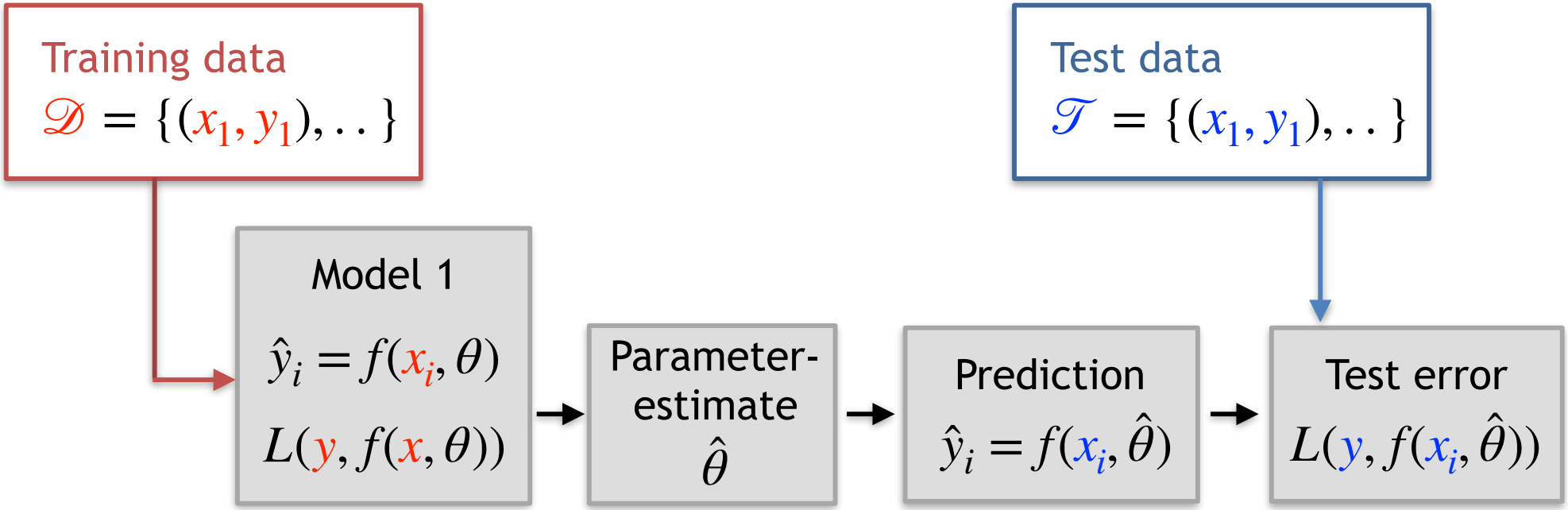


Aside: t versus z

- To make the CI, we had to estimate σ_X (or σ_l when talking about loss)
- This adds additional uncertainty
- Correct for this by using the *t-distribution with* $n-1$ *degrees of freedom* instead of normal.
- Doesn't really matter for n bigger than about 30

Week 4.4
Lab: Uncertainty of the test error

Prediction performance



Ultimately, we want our model to work well on new data

To test this, we can collect a new data or reserve a part of the data as test a test dataset

Test Error

Given a ***dataset*** (collection of realizations) $(x_1, y_1), \ldots, (x_n, y_n)$ of (X, Y), ***that were not used to find f*** the **test error** is:

$$l_{\mathcal{T}} = \frac{1}{n} \sum_i L(y_i, f(x_i, \hat{\theta}))$$

MENTAL HEALTH

A Blood Test Might One Day Mass Screen Military Personnel for PTSD

An assay that measures 28 variables could identify individuals who need further treatment

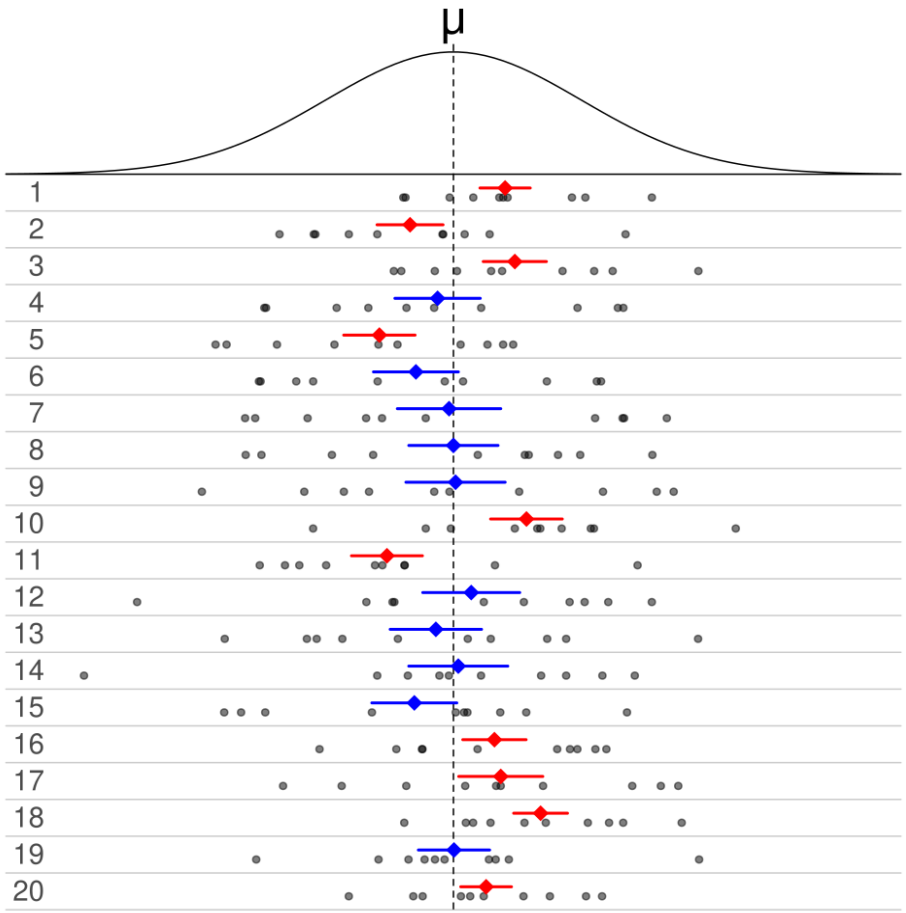
By Emily Willingham on September 10, 2019

https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=https://www.nature.com/articles/s41380-019-0496-z

To develop the 28-factor screen, Marmar and his colleagues began with almost a million genomic, protein, metabolic, and other molecular candidates that they assessed in blood samples of 77 veterans with diagnosed PTSD and 74 veterans without the condition. All participants were male, had been in a war zone in either Iraq or Afghanistan and had experienced at least one traumatic war zone event.

The research team then tested how well the markers they found worked in a separate pool of 52 veterans—26 with diagnosed PTSD and 26 without it. The markers achieved an accuracy of 81 percent in distinguishing the two groups.

- Confidence interval (CI) is a range of estimates for an unknown parameter.
- A CI is computed at a designated confidence level (CL) with 95% being the most common value.
- CL represents the long-run proportion of corresponding CIs that contain the true value of the parameter.



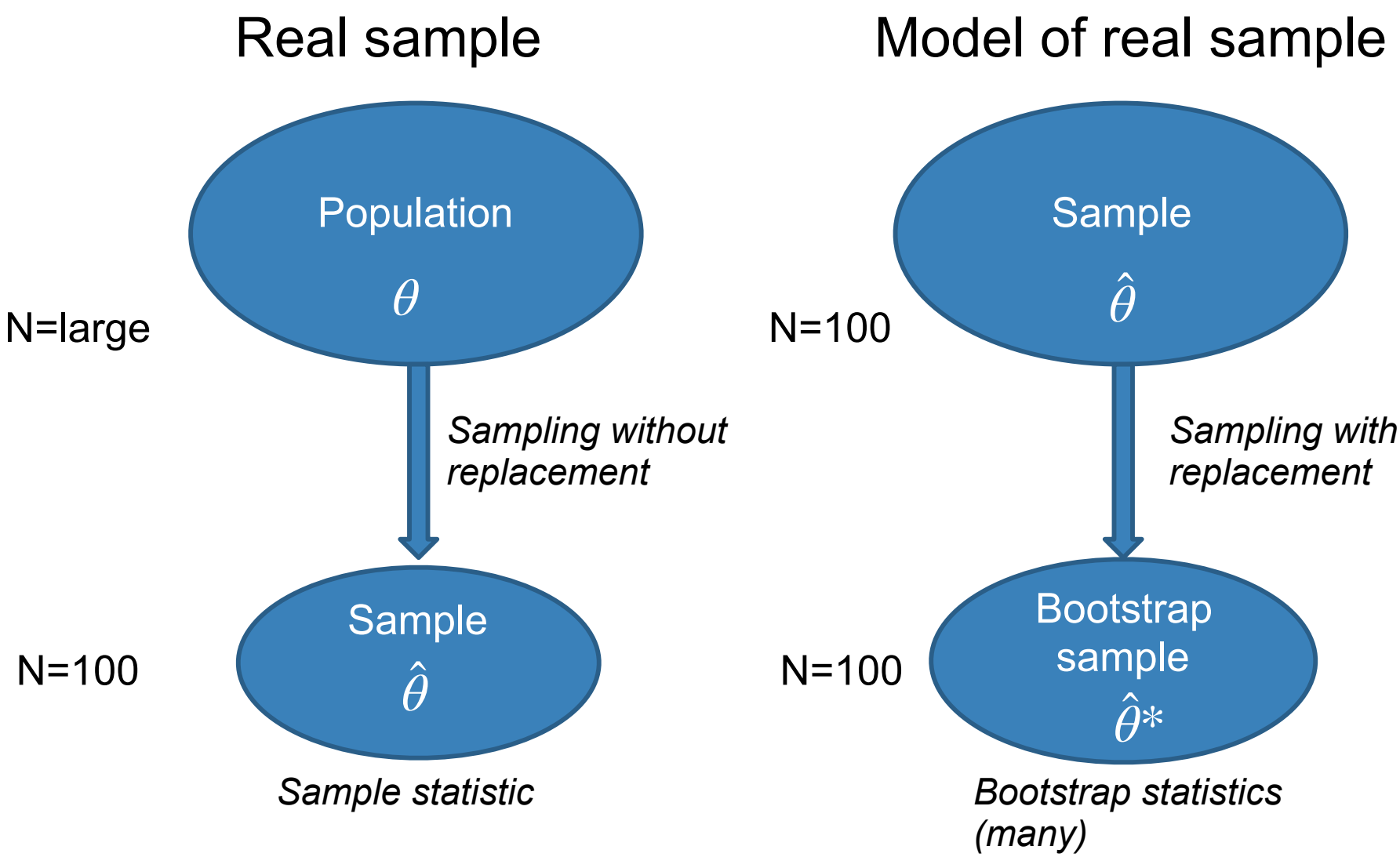
CL 50%

Week 4.5
The Bootstrap

The Bootstrap

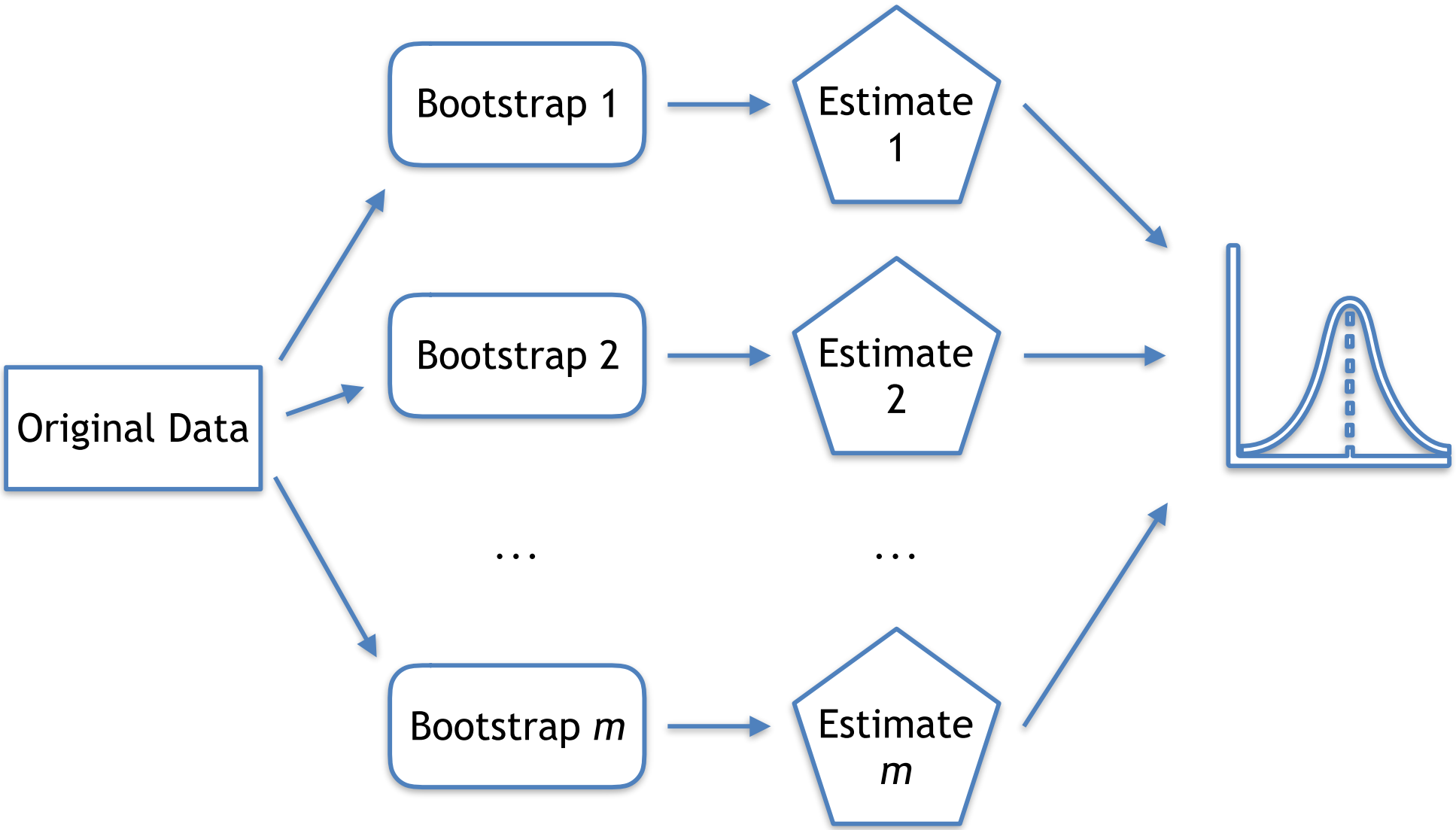
- Some estimators have complicated standard errors or are not normal
- If we could only draw many samples….
- Idea: Use the data we have to "pretend" we can sample more data

The Bootstrap



Efron (1979) showed that under specific conditions, the distribution of $\hat{\theta}* - \hat{\theta}$ *approximates the distribution of* $\hat{\theta} - \theta$.

The Bootstrap



The Bootstrap

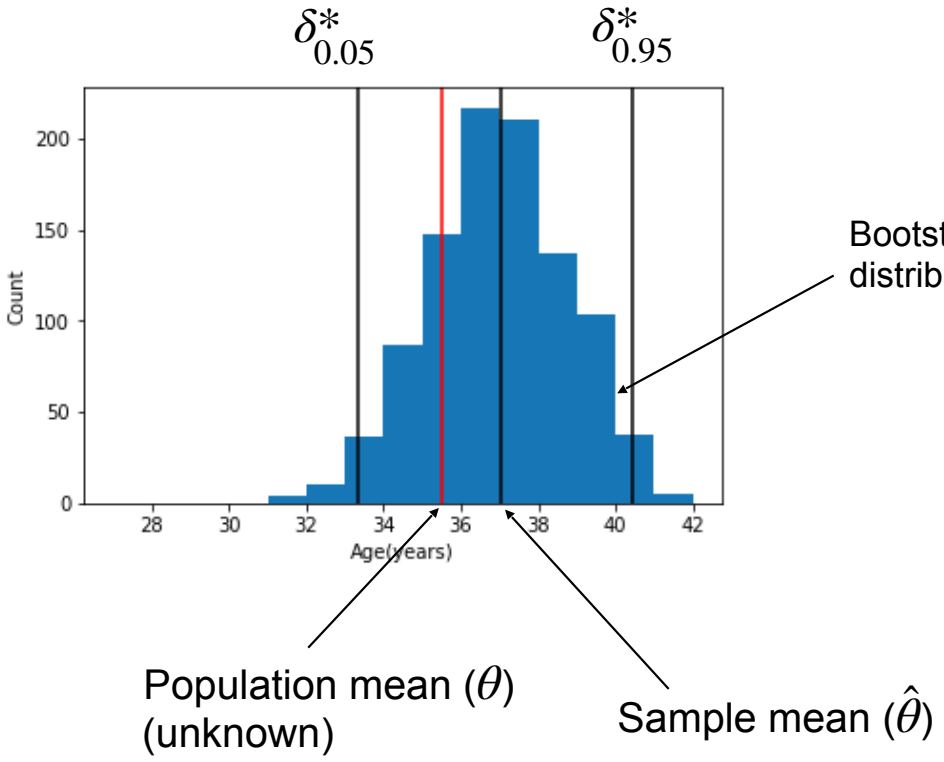
```
Your Sample S has N observations
For b in 1:numBootstrap:
      resample N from S with replacement -> S*
      Fit model to S* -> θ̂*(bootstrap statistics)
      Record your bootstrap statistics

Return the distribution of bootstrap statistics
```

The Bootstrap

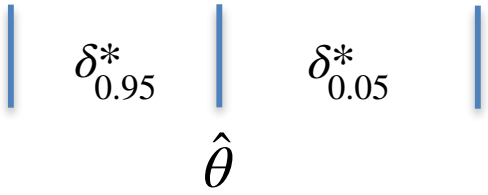
Once you have your bootstrap distribution, you can construct confidence interval:

$$\delta^* = \hat{\theta}^* - \hat{\theta}$$

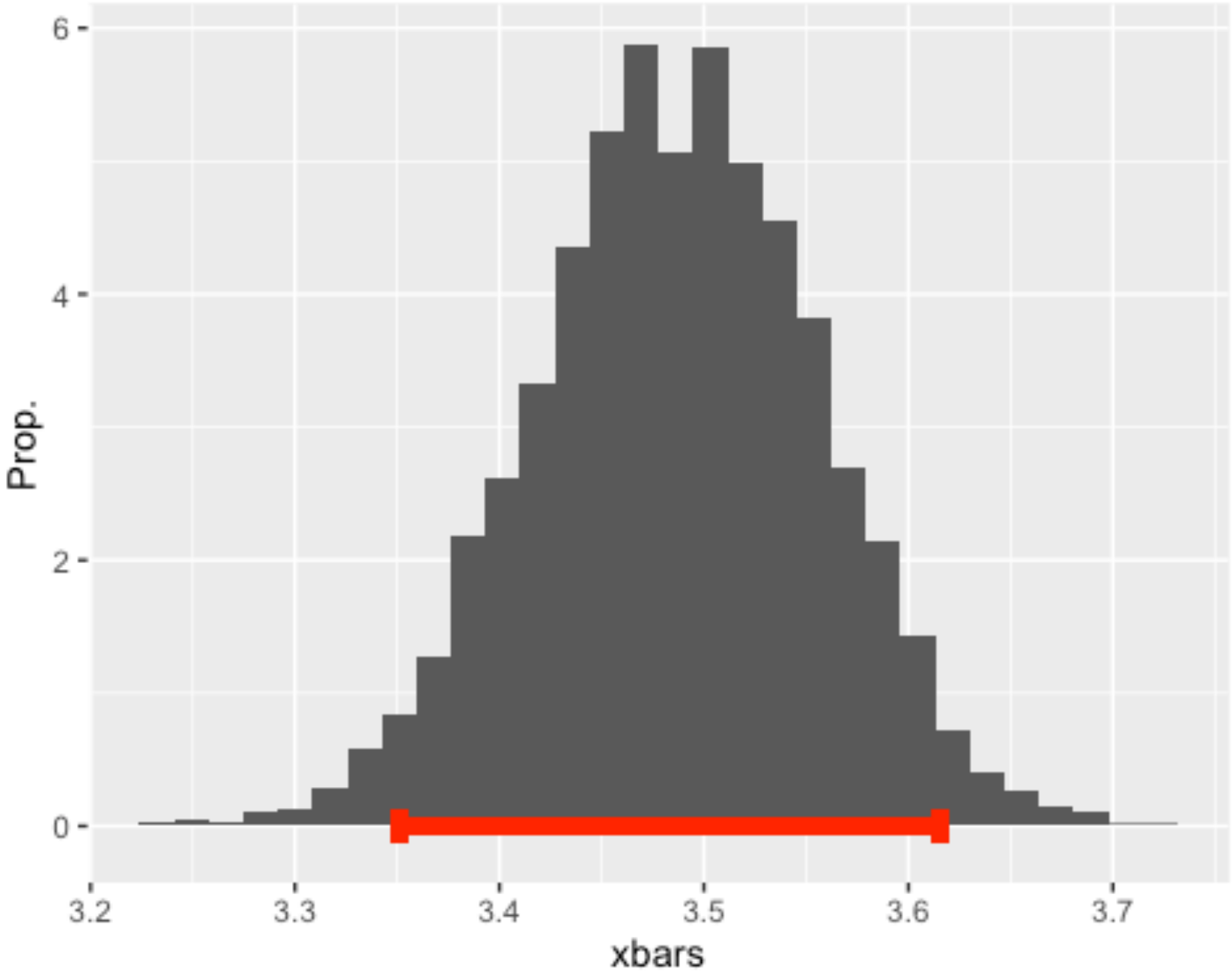
$\delta^*_{0.05}$: 5% percentile of bootstrap distribution



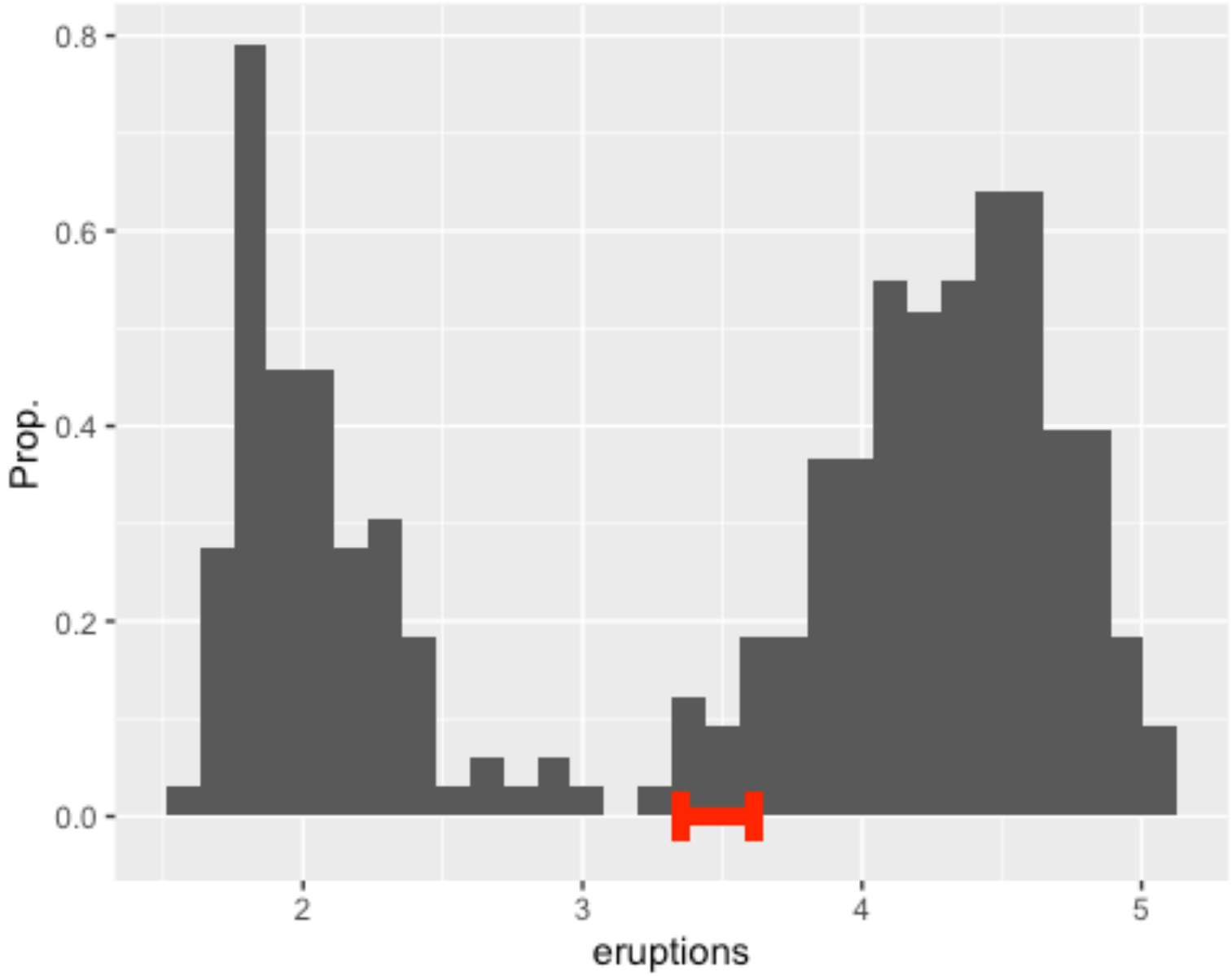
90% Confidence interval:



$$CI = \left[\hat{\theta} - \delta^*_{0.95}, \hat{\theta}^* - \delta^*_{0.05}\right]$$


6
4
2
0
Prop.
3.2
3.3
3.4
3.5
3.6
3.7
xbars

Reality Check: Geyser example



Summary: The Bootstrap

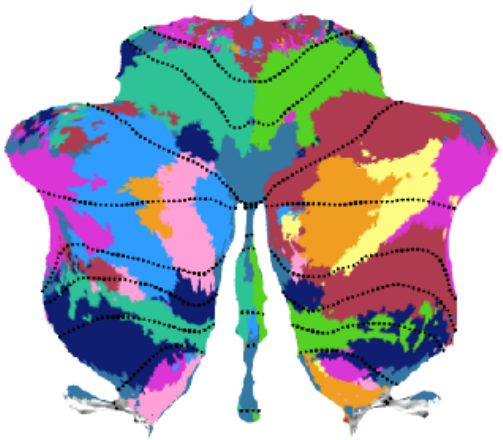
- Universal technique to obtain confidence intervals
- Can be applied to any statistics
- Confidence intervals are *asymptotically* correct
- Does not make assumptions about the underlying distribution

Disadvantages:

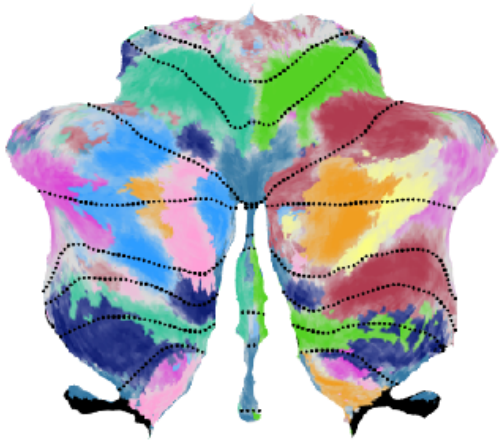
- Extra programming
- Can become very unstable with small *N*
- Takes computation time

Summary

- For Linear models and data with approximately gaussian noise, the CLT provides excellent closed-form expressions for the distribution of parameter estimates
- However, bootstrap is a universal technique applicable to all data distributions and models / sample statistics.



Clustering of the human cerebellum into functional regions on original sample of N=24 subjects



Bootstrap estimate of the assignment probability with grey areas showing high uncertainty

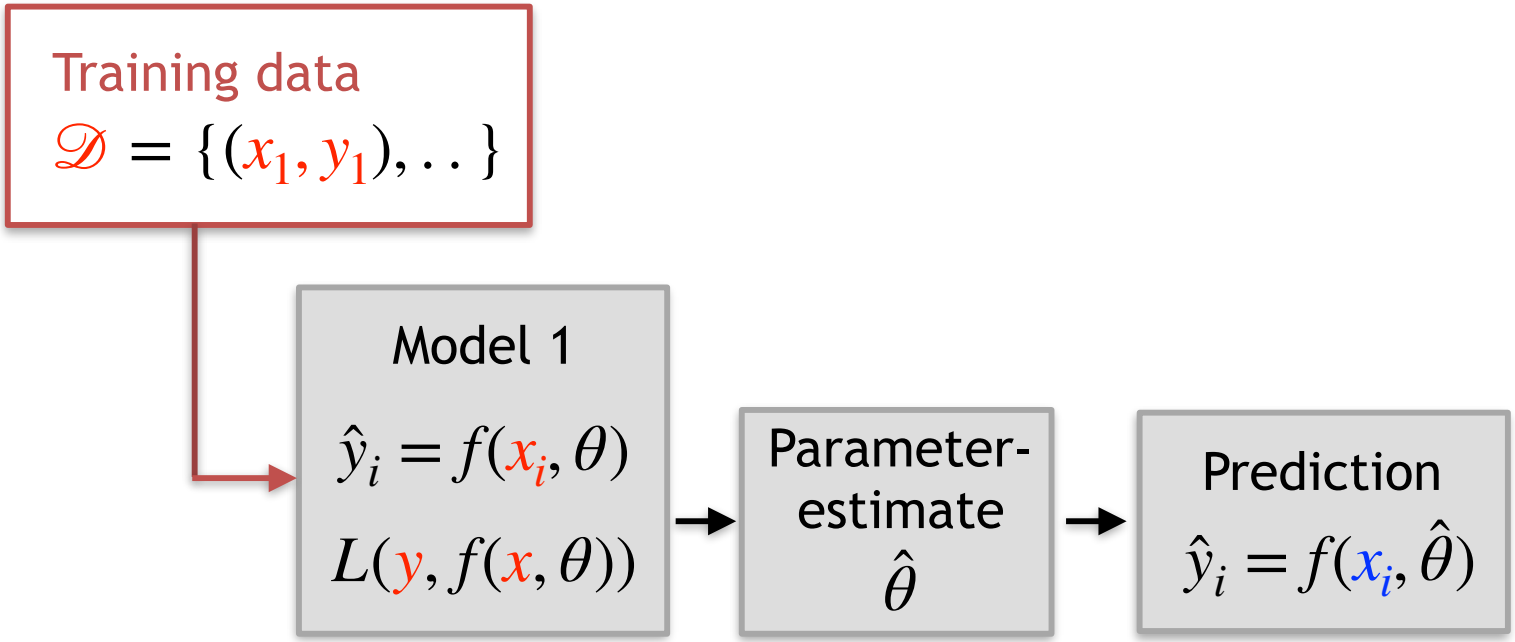
King et al. (2019). Nature Neuroscience

Week 4.6
Lab: Constructing Confidence intervals for test error via bootstrap

See Lab04_part1

Week 4.7
Prediction uncertainty

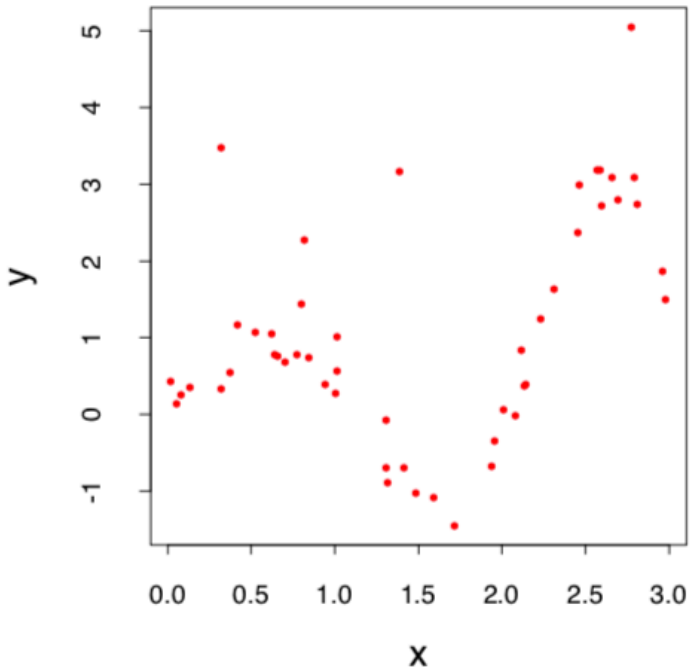
Prediction uncertainty



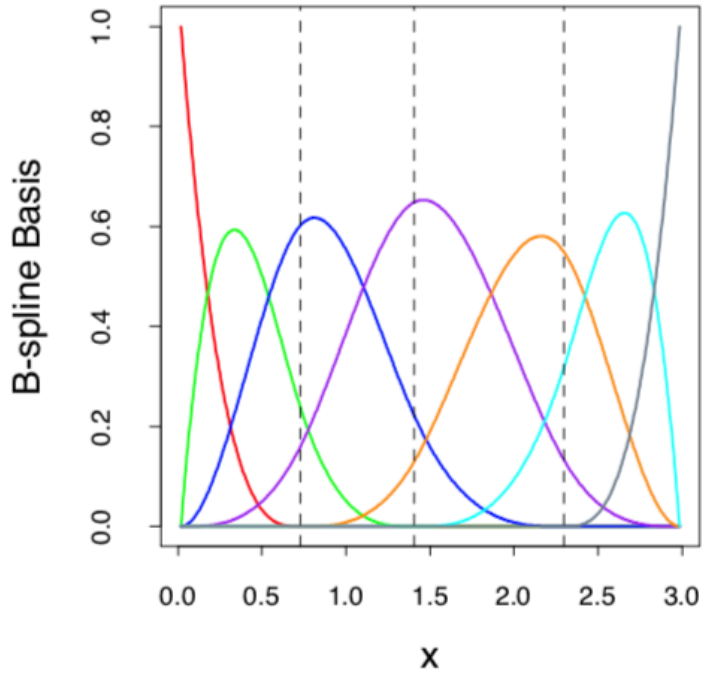
How does the uncertainty in the parameter estimates influence the uncertainty of the prediction?

How much would our prediction change, if we had drawn a different set of training data?

Prediction uncertainty



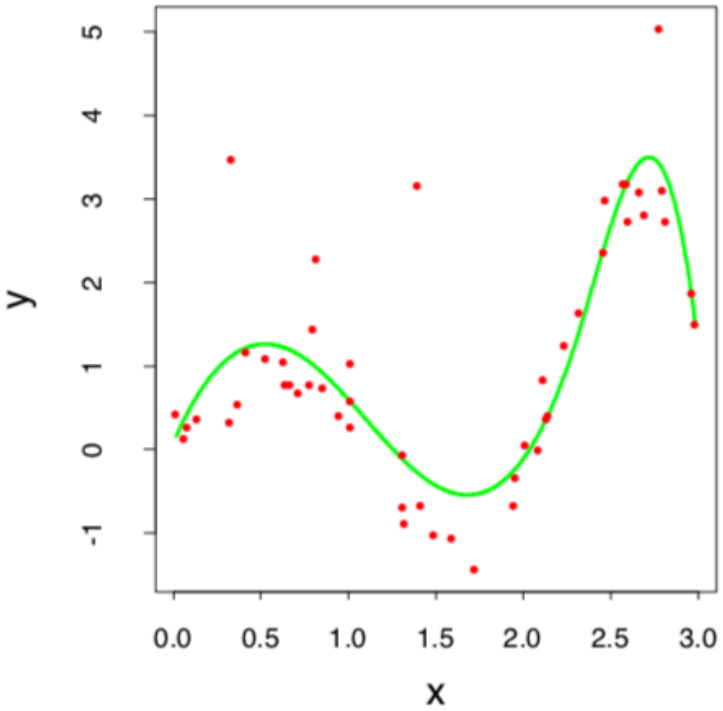
Say we have a data set with a clearly non-linear dependence between x and y



- We can define 7 nonlinear *basis functions* $b_i(x)$
- Here we use B-splines, but it could be any other (polynomials, gaussians, fourier-basis)
- We concatenate them into a design matrix X fit the model $\hat{y} = f(x) = X\theta$.

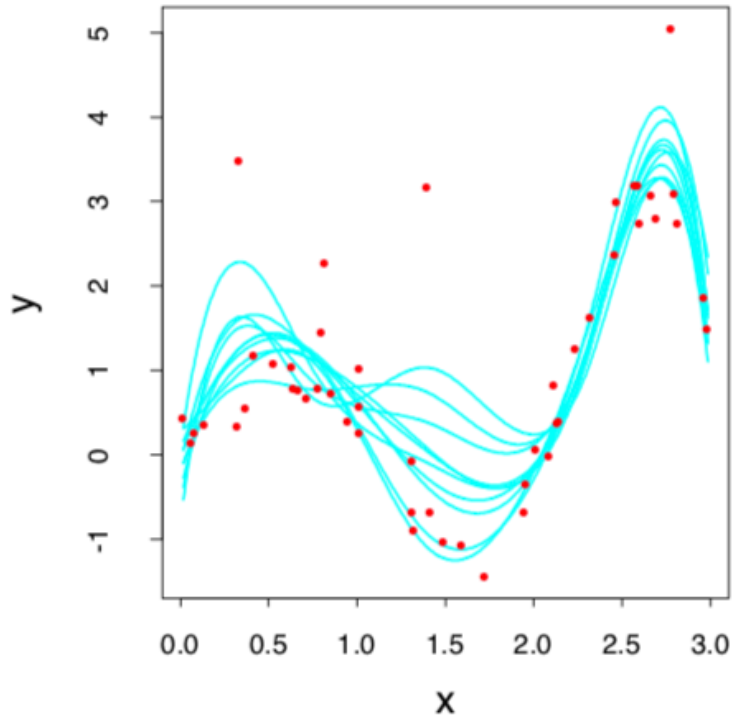
HTF, p261-264

Prediction uncertainty



Our best fit on the training data
$\hat{y} = \hat{f}(x) = X\hat{\theta}$
looks pretty good, but how certain can we be?

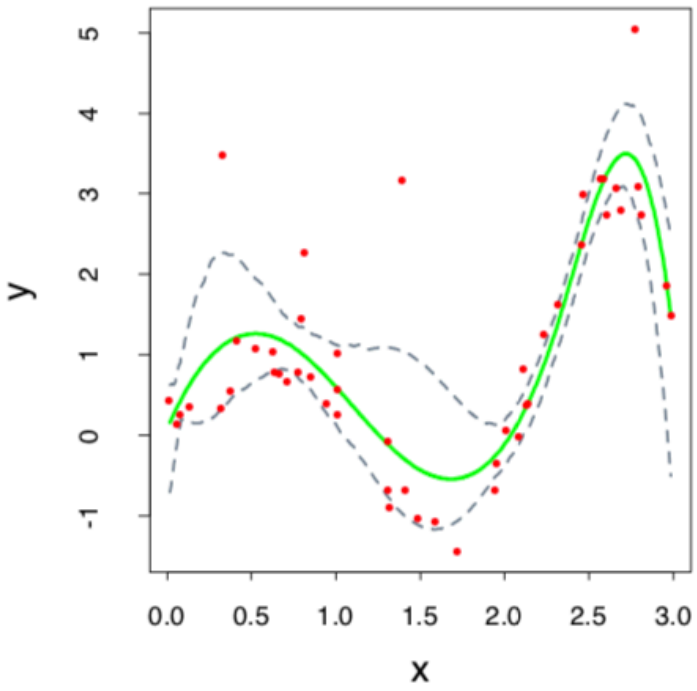
We do a bootstrap, each time getting a new sample, and each time getting a new parameter estimate $\hat{\theta}^*_b$



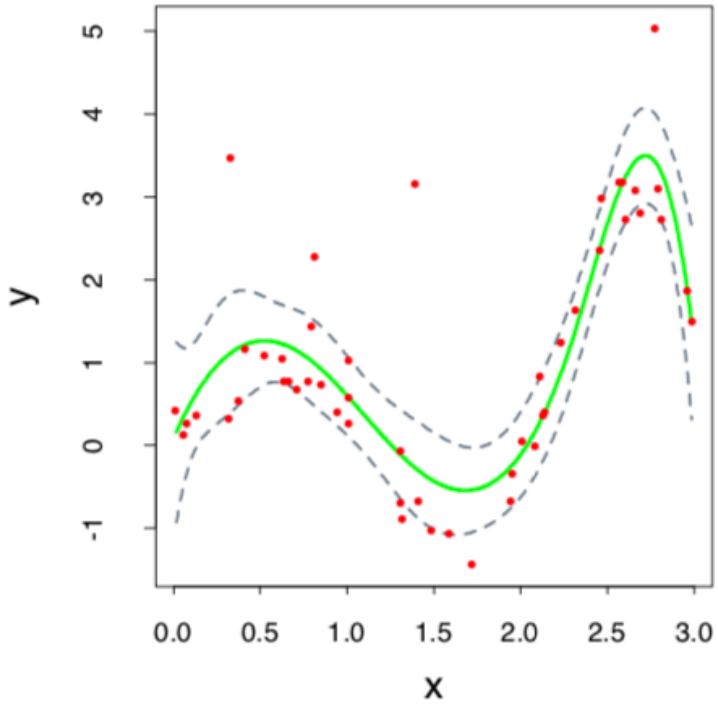
For each parameter vector, we can now plot a new prediction
$\hat{y}^*_b = \hat{f}^*_b(x) = X\hat{\theta}^*_b$

HTF, p261-264

Prediction uncertainty



For each x we can then compute a 95% confidence interval based on $\hat{y}_b^*$



In the case of linear regression, we can also use the central limit theorem to get the CIs - as you can see they look fairly similar.

HTF, p261-264

Prediction uncertainty

- Note that the confidence interval tells us that the true value ($f(x)$) is with 95% probability in this interval
- It does NOT mean that for a new new observation (x_n, y_n) is within the interval with 95% chance
- This is because y_n will differ from $f(x_n)$ by some random variability (σ^2)
- A confidence interval for new data (the *predictive density)* needs to take into account uncertainty of the predicted value, $var(\hat{f}(x_n))$, and the random variability (σ^2)