

LECTURE 17

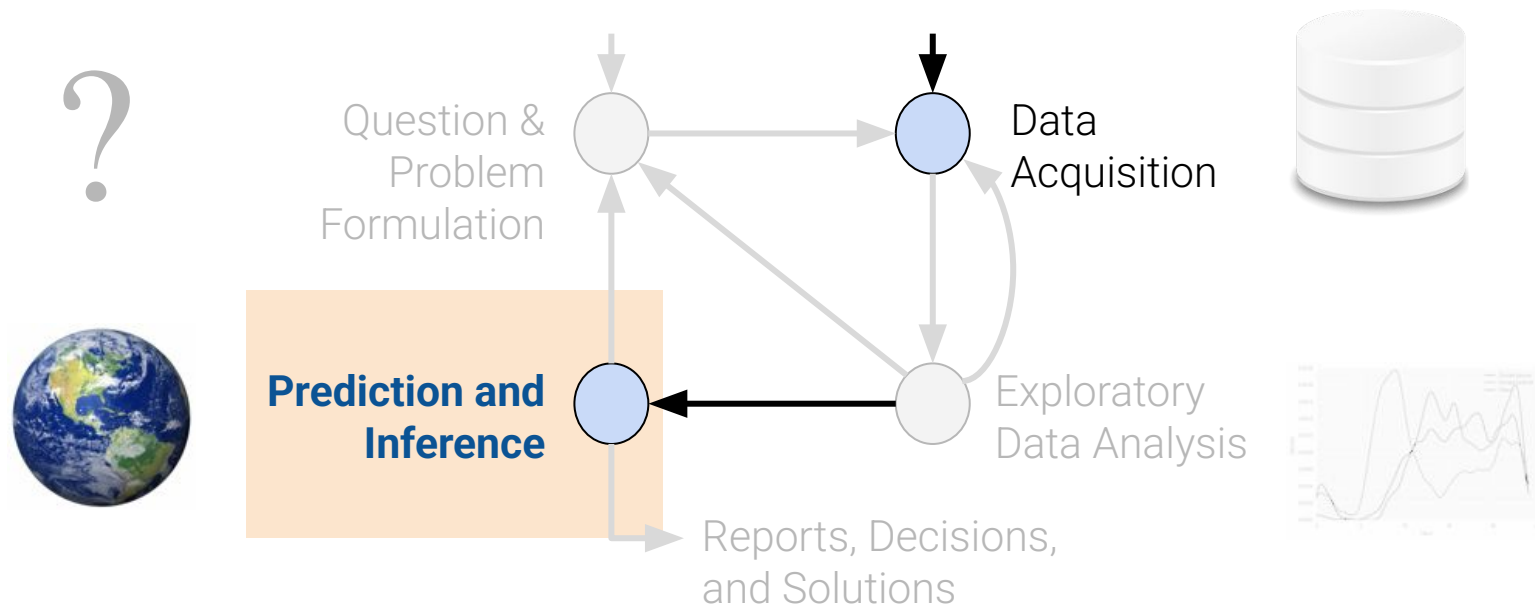
Estimators, Bias, and Variance

Exploring the different sources of error in the predictions that our models make.

Data 100/Data 200, Spring 2022 @ UC Berkeley

Josh Hug and Lisa Yan

Why Probability?



(today)

Model Selection Basics:

Cross Validation
Regularization

Probability I:

Random Variables
Estimators

Probability II:

Bias and Variance
Inference/Multicollinearity

Today's Roadmap

Lecture 17, Data 100 Spring 2022

Sample Statistics (from last time)

Prediction vs. Inference

- Modeling: Assumptions of Randomness

The Bias-Variance Tradeoff

Interpreting Slopes

[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance
Decomposition

From Populations to Samples

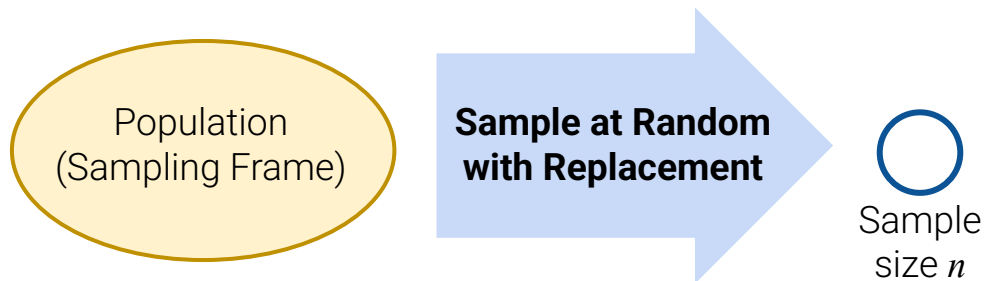
Today, we've talked extensively about **populations**:

- If we know the **distribution of a random variable**, we can reliably compute expectation, variance, functions of the random variable, etc.

However, in Data Science, we often collect **samples**.

- We don't know the distribution of our population.
- We'd like to use the distribution of your sample to estimate/infer properties of the population.

The **big assumption** we make in modeling/inference: Our random sample datapoints are **IID**.



The Sample is a Set of IID Random Variables

Population
(Sampling Frame)

Sample at Random
with Replacement

Sample
size n

x	P(X = x)
3	0.1
4	0.2
6	0.4
8	0.3

or

X(s)	
0	3
1	4
2	4
3	6
4	8
...	...
79995	6
79996	6
79997	4
79998	6
79999	6
...	...

`df.sample(n,
replace=True)
\[documentation\]`

x	
0	6
1	8
2	6
3	6
4	3
...	...
95	8
96	6
97	6
98	3
99	8

Each observation in our sample is a **Random Variable** drawn **IID** from our population distribution.

Sample
($n \ll N$)

$$X_1, X_2, \dots, X_n$$

Population
(really large N)

The Sample is a Set of IID Random Variables

Population
(Sampling Frame)

x	P(X = x)
---	----------

3	0.1
---	-----

4	0.2
---	-----

6	0.4
---	-----

8	0.3
---	-----

or

X(s)

0	3
---	---

1	4
---	---

2	4
---	---

3	6
---	---

4	8
---	---

...	...
-----	-----

79995	6
-------	---

79996	6
-------	---

79997	4
-------	---

79998	6
-------	---

79999	6
-------	---

...	...
-----	-----

Sample at Random
with Replacement

○
Sample
size n

`df.sample(n,
replace=True)
\[documentation\]`

x

0	6
---	---

1	8
---	---

2	6
---	---

3	6
---	---

4	3
---	---

...	...
-----	-----

95	8
----	---

96	6
----	---

97	6
----	---

98	3
----	---

99	8
----	---

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Sample Mean

A **random variable**!

Depends on our randomly
drawn sample!!

`np.mean(...)` = 5.71

Sample X_1, X_2, \dots, X_n

$$E[X] = 5.9$$

Population Mean

A **number**,
i.e., fixed value

μ

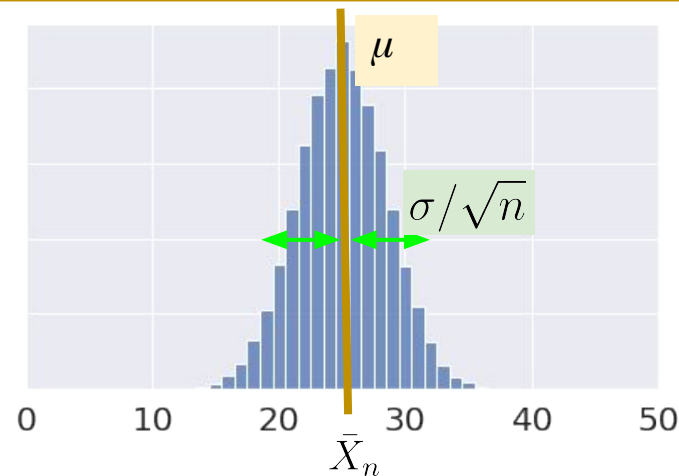
The Central Limit Theorem

No matter what population you are drawing from:

If an IID sample of size n is large,
the probability distribution of the **sample mean**
is **roughly normal** with mean μ and SD σ/\sqrt{n} .

(STAT 140/EECS 126)

(pop mean μ , pop SD σ
next slide)



Any theorem that provides the rough distribution of a statistic
and **doesn't need the distribution of the population** is valuable to data scientists.

- Because we rarely know a lot about the population!

For a more in-depth demo: https://onlinestatbook.com/stat_sim/sampling_dist/

Properties of the Sample Mean

Consider an IID sample X_1, X_2, \dots, X_n drawn from a numerical population with **mean μ and SD σ** .

Define the **sample mean**:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Expectation:

$$\begin{aligned} \mathbb{E}[\bar{X}_n] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] \\ &= \frac{1}{n} (n\mu) = \mu \end{aligned}$$

Variance/Standard Deviation:

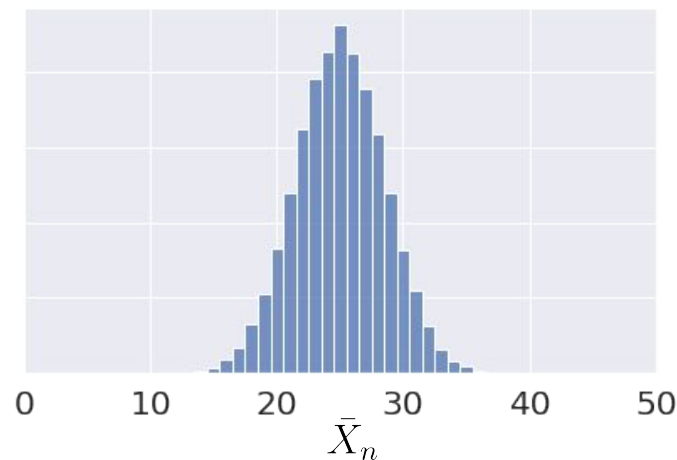
$$\begin{aligned} \text{Var}(\bar{X}_n) &= \frac{1}{n^2} \text{Var} \left(\sum_{i=1}^n X_i \right) = \frac{1}{n^2} \underbrace{\left(\sum_{i=1}^n \text{Var}(X_i) \right)}_{\text{IID} \rightarrow \text{Cov}(X_i, X_j) = 0} \\ &= \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n} \end{aligned}$$

$$\text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

How Large Is “Large”?

No matter what population you are drawing from:

If an IID **sample of size n is large**,
the probability distribution of the sample mean
is **roughly normal** with mean μ and SD σ/\sqrt{n} .



How large does n have to be for the normal approximation to be good?

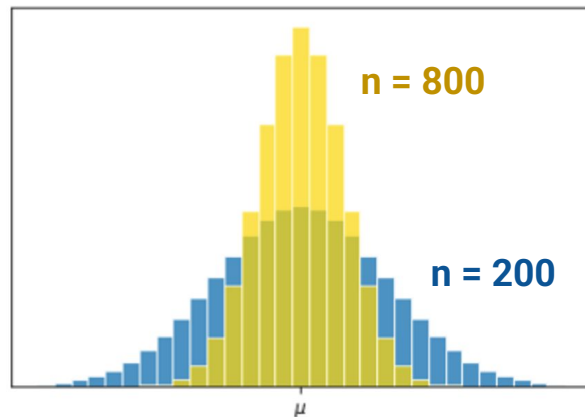
- ...It depends on the shape of the distribution of the population...
- If population is **roughly symmetric and unimodal**/uniform, could need as few as **$n = 20$** .
If population is very skewed, you will need bigger n .
- If in doubt, you can bootstrap the sample mean and see if the bootstrapped distribution is bell-shaped.

Using the Sample Mean to Estimate the Population Mean

Our goal is often to **estimate** some characteristic of a population.

- Example: average height of Cal undergraduates.
- We typically can collect a **single sample**. It has just one average.
- Since that sample was random, it *could have* come out differently.

We should consider the **average value and spread** of all possible sample means, and how it scales with the sample size n .



$$\mathbb{E}[\bar{X}_n] = \mu$$

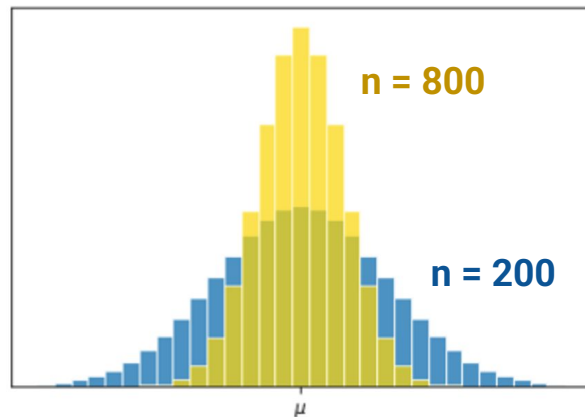
$$\text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

Using the Sample Mean to Estimate the Population Mean

Our goal is often to **estimate** some characteristic of a population.

- Example: average height of Cal undergraduates.
- We typically can collect a **single sample**. It has just one average.
- Since that sample was random, it *could have* come out differently.

We should consider the **average value and spread** of all possible sample means, and what this means for how big n should be.



$$\mathbb{E}[\bar{X}_n] = \mu$$

$$\text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

For every sample size, the expected value of the sample mean is the population mean.

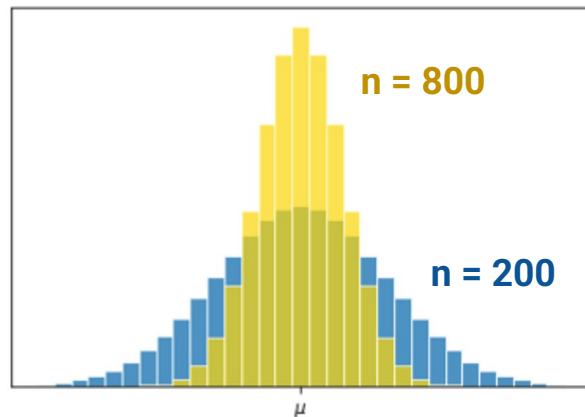
We call the sample mean an **unbiased estimator** of the population mean.
(more on this in a bit)

Using the Sample Mean to Estimate the Population Mean

Our goal is often to **estimate** some characteristic of a population.

- Example: average height of Cal undergraduates.
- We typically can collect a **single sample**. It has just one average.
- Since that sample was random, it *could have* come out differently.

We should consider the **average value and spread** of all possible sample means, and what this means for how big n should be.



$$\mathbb{E}[\bar{X}_n] = \mu$$

$$\text{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

For every sample size, the expected value of the sample mean is the population mean.

We call the sample mean an **unbiased estimator** of the population mean.
(more on this in a bit)

Square root law ([Data 8](#)): If you increase the sample size by a factor, the SD decreases by the square root of the factor.

The sample mean is more likely to be close to the population mean if we have a larger sample size.

Prediction vs. Inference

Lecture 17, Data 100 Spring 2022

Sample Statistics (from last time)

Prediction vs. Inference

- Modeling: Assumptions of Randomness

The Bias-Variance Tradeoff

Interpreting Slopes

[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance
Decomposition

Linear Algebra Resources: [Ed post](#)

Prediction vs. Inference

Why do we build models? (From Intro to Modeling lecture: [link](#))

To make **accurate predictions**
about unseen data.

Prediction is the task of using our model to make predictions for the response (output) of unseen data.

To understand **complex phenomena**
occurring in the world we live in.

Inference is the task of using our model to draw conclusions about the underlying true relationship(s) between our features and response.

Prediction vs. Inference

Why do we build models? (From Intro to Modeling lecture: [link](#))

To make **accurate predictions**
about unseen data.

Prediction is the task of using our model to make predictions for the response (output) of unseen data.

To understand **complex phenomena**
occurring in the world we live in.

Inference is the task of using our model to draw conclusions about the underlying true relationship(s) between our features and response.

Example: Suppose we are interested in studying the relationship between the value of a home and a view of a river, school districts, property size, income level of community, etc.

Prediction: Given the attributes of some house, how much is it worth?

We care more about making accurate predictions, don't care so much about how.

Inference: How much extra will a house be worth if it has a view of the river?

We care more about having model parameters that are interpretable and meaningful.

Inference and Statistical Inference

Inferences are steps in reasoning, **moving from premises to logical consequences**; etymologically, the word infer means to "carry forward".

[Wikipedia](#)

Statistical inference is the process of using data analysis to **deduce properties of an underlying distribution** of probability.

Oxford Dictionary of Statistics, Upton & Cook, 2008

Statistical inference, or "learning" as it is called in computer science, is the process of **using data to infer the distribution** that generated the data.

All of Statistics, L. Wasserman, 2004

The goal of empirical research is--or should be--to increase our understanding of the phenomena, **rather than displaying our mastery of the technique**.

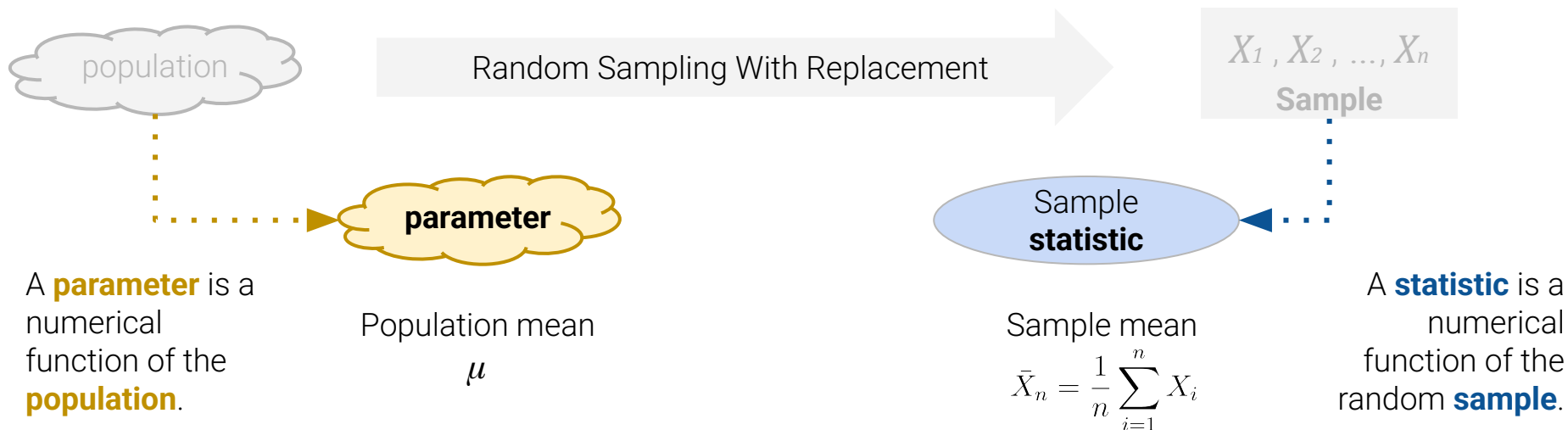
Statistical Models, D. Freedman, 2009

Inference is all about **drawing conclusions** about **population parameters**, given only a **random sample**.



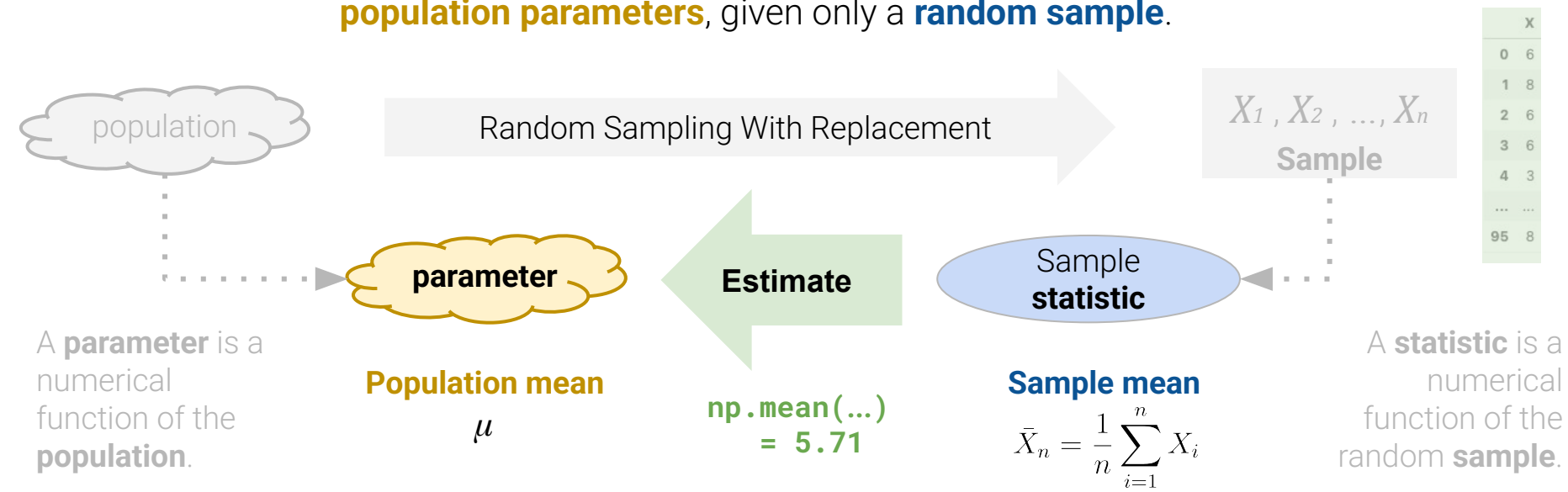
[Terminology] Parameters, Statistics, and Estimators

Inference is all about **drawing conclusions** about **population parameters**, given only a **random sample**.



[Terminology] Parameters, Statistics, and Estimators

Inference is all about **drawing conclusions** about **population parameters**, given only a **random sample**.



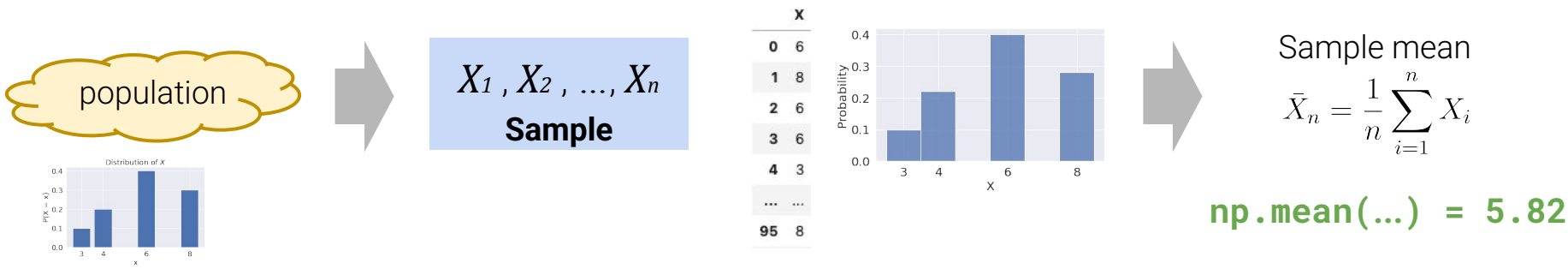
We can then use the sample statistic as an **estimator** of the true population parameter.

Since our **sample is random**, our statistic (which we use as our estimator) could have been different.

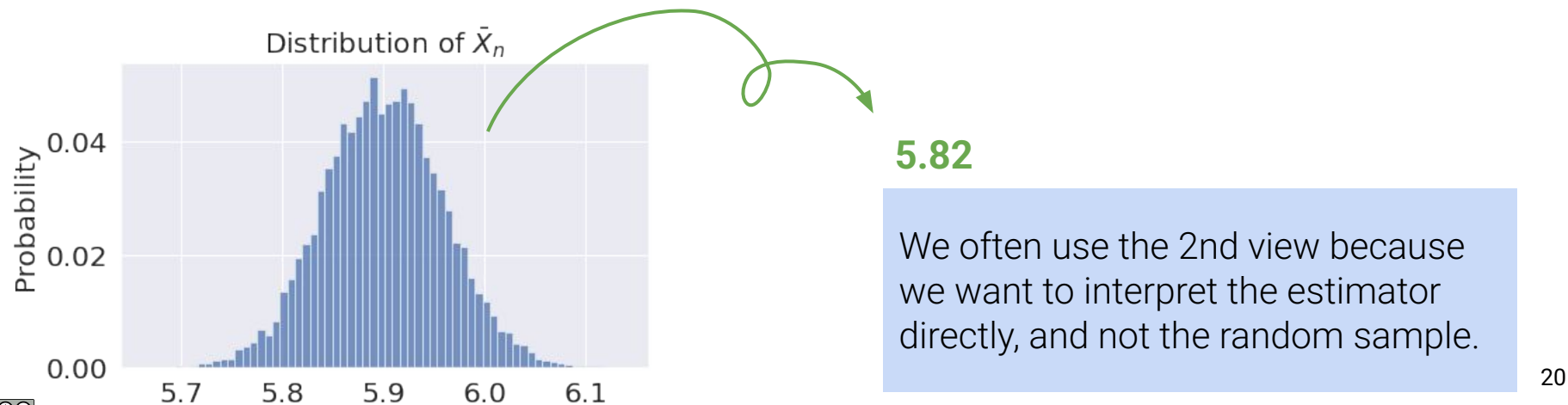
Example: When we use the sample mean to estimate the population mean, our estimator is almost always going to be somewhat off.

Data Generation Process: Estimating a Value

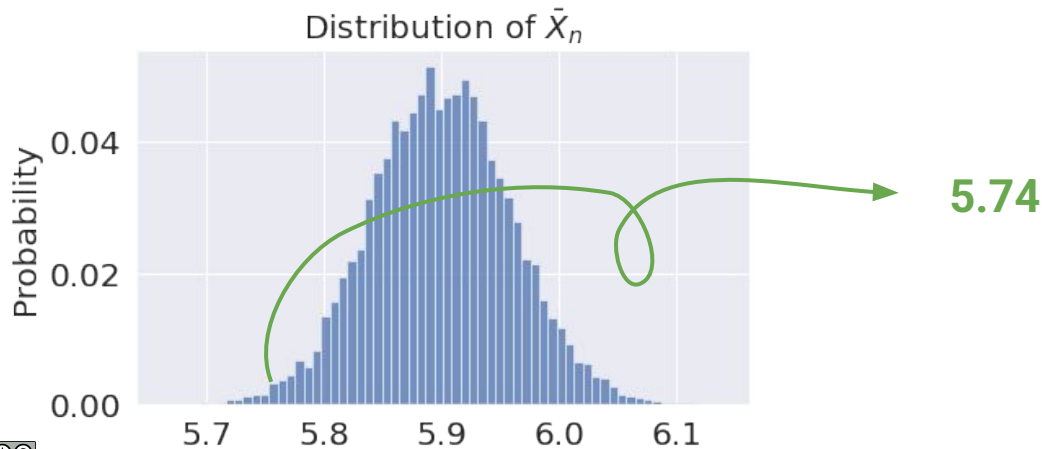
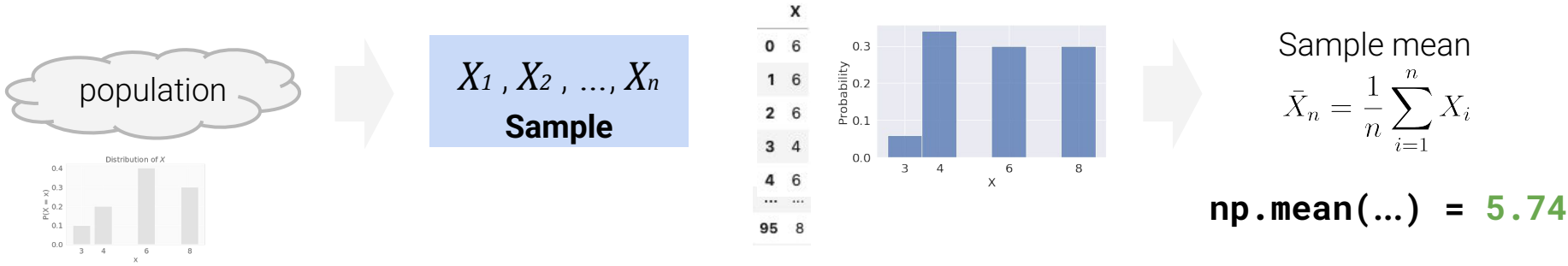
One View: Randomly draw a random sample, then compute the statistic for that sample.



Another View: Randomly draw from the distribution of the statistic (generated from all possible samples).



If We Drew a Different Sample, We'd Get A Different Estimator



The value of our estimator is a function of the random sample. The estimator is therefore also random.

Modeling: Estimating a Relationship

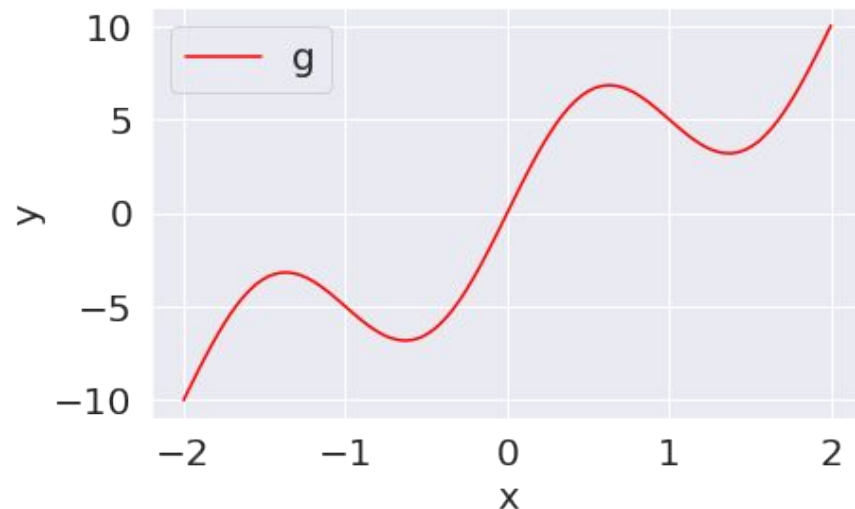
What if we wanted to estimate the relationship between input x and random response Y ?

$$Y = g(x) + \epsilon$$

We would like to find the true relationship g .

Each individual in the population has:

- **Fixed features** x , and hence fixed $g(x)$.



Modeling: Estimating a Relationship

What if we wanted to estimate the relationship between input x and random response Y ?

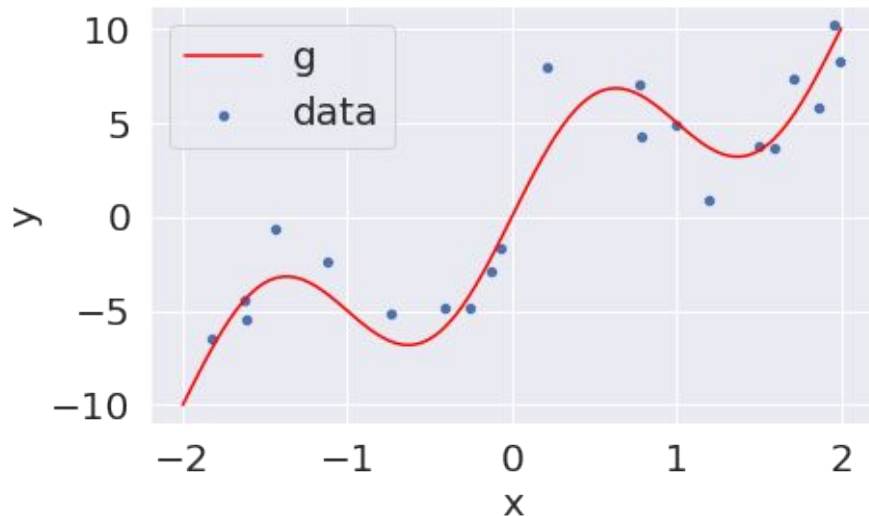
$$Y = g(x) + \epsilon$$

We would like to find the true relationship g .

Each individual in the population has:

- **Fixed features** x , and hence fixed $g(x)$.
- Random **error/noise** ϵ
- Random **observation/response** $Y = g(x) + \epsilon$

Errors ϵ are assumed expectation 0 (“zero mean”) and i.i.d. across individuals



Modeling: Estimating a Relationship

What if we wanted to estimate the relationship between input x and random response Y ?

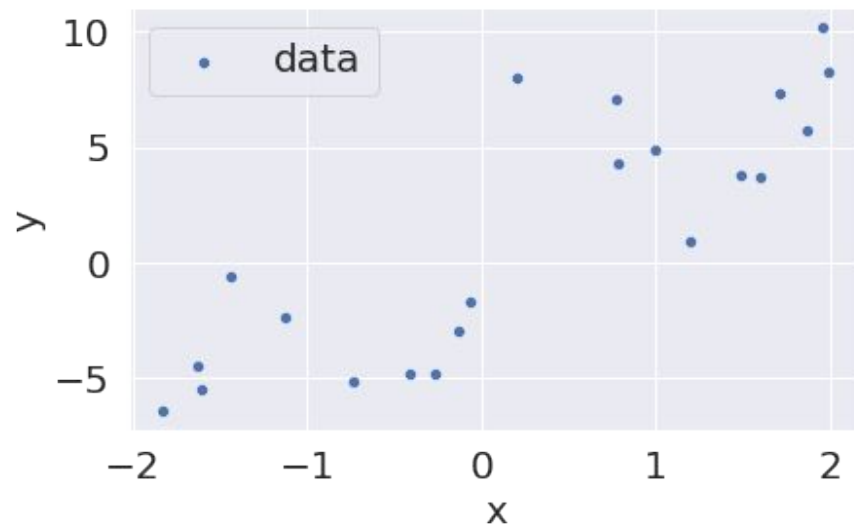
$$Y = g(x) + \epsilon$$

We would like to find the true relationship g .

Each individual in the population has:

- **Fixed features** x , and hence fixed $g(x)$.
- Random **error/noise** ϵ
- Random **observation/response** $Y = g(x) + \epsilon$

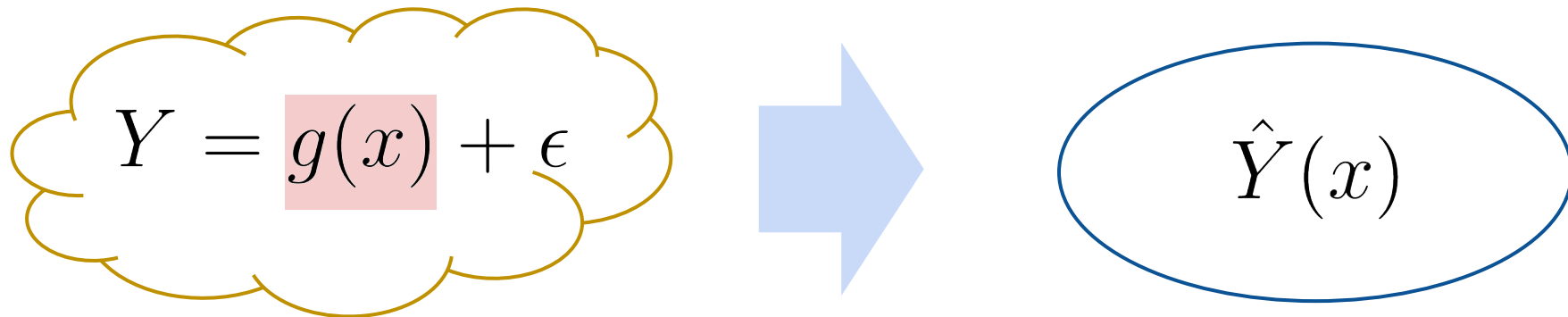
Errors ϵ are assumed expectation 0 (“zero mean”) and i.i.d. across individuals



We only can only observe our random sample. From this we'd like to estimate the true relationship g .

Modeling: Estimating a Relationship

What if we wanted to estimate the relationship between input x and random response Y ?



We would like to find the true relationship g .

Each individual in the population has:

- **Fixed features** x , and hence fixed $g(x)$.
- Random **error/noise** ϵ
- Random **observation/response** $Y = g(x) + \epsilon$

Errors ϵ are assumed expectation 0 (“zero mean”) and i.i.d. across individuals

We build a **model** for predictions based on our observed sample of (x, y) pairs. Our model **estimates** the true relationship g .

At every x , our **prediction** for Y is $\hat{Y}(x)$.

Modeling: Estimating a Relationship

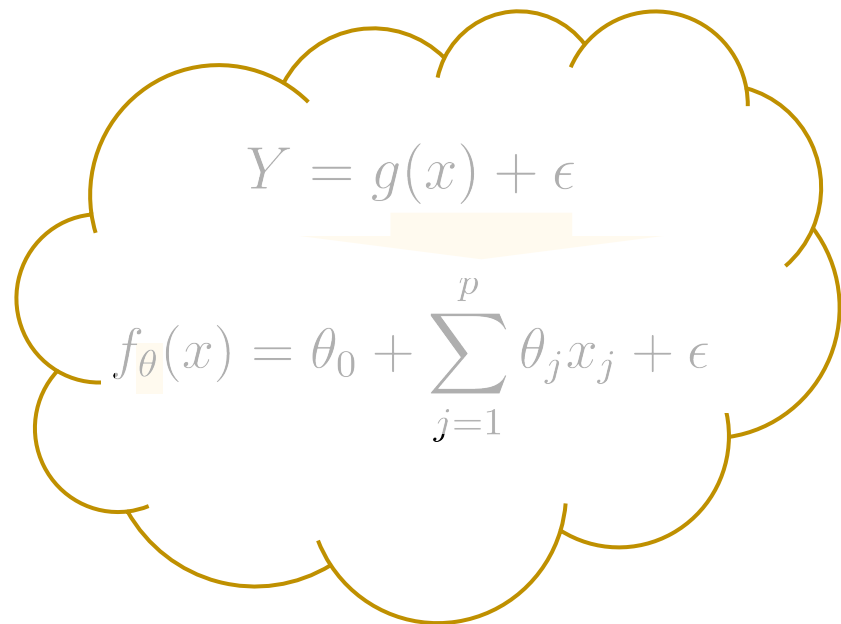
If we assume our true relationship g is **linear**, then we express the response as $Y = f_{\theta}(x)$.

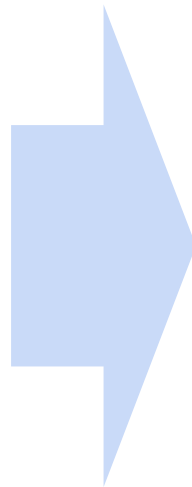


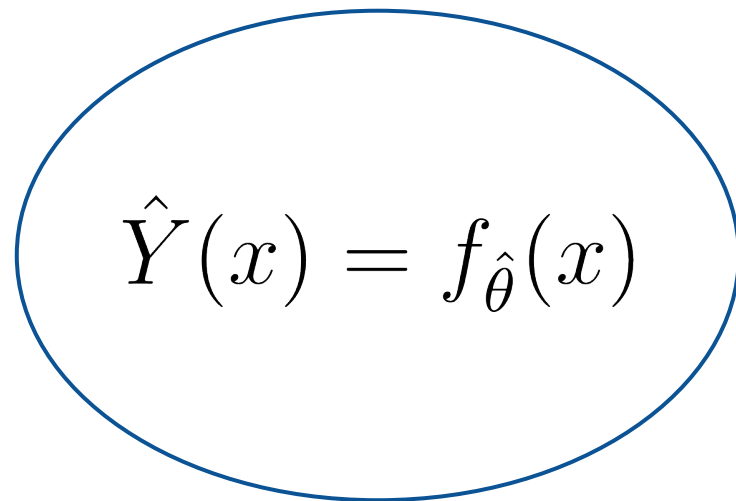
- The true relationship has true, unobservable parameters θ .
- There is still random noise ϵ , so we still can never observe the true relationship.

Estimating a linear relationship

If we assume our true relationship g is **linear**, then we express the response as $Y = f_{\theta}(x)$.


$$Y = g(x) + \epsilon$$
$$f_{\theta}(x) = \theta_0 + \sum_{j=1}^p \theta_j x_j + \epsilon$$




$$\hat{Y}(x) = f_{\hat{\theta}}(x)$$

- The true relationship has true, unobservable parameters θ .
- There is still random noise ϵ , so we still can never observe the true relationship.

- Obtain a sample \mathbb{X}, \mathbb{Y} of n observed relationships (x, Y) .
- Train a model and obtain estimates $\hat{\theta}$.



$$\hat{Y}(x) = f_{\hat{\theta}}(x)$$

Hats mean estimates.

Which expressions are random?

Suppose we have an individual with fixed input x . Assume the true relationship g is linear.



$$Y = g(x) + \epsilon$$

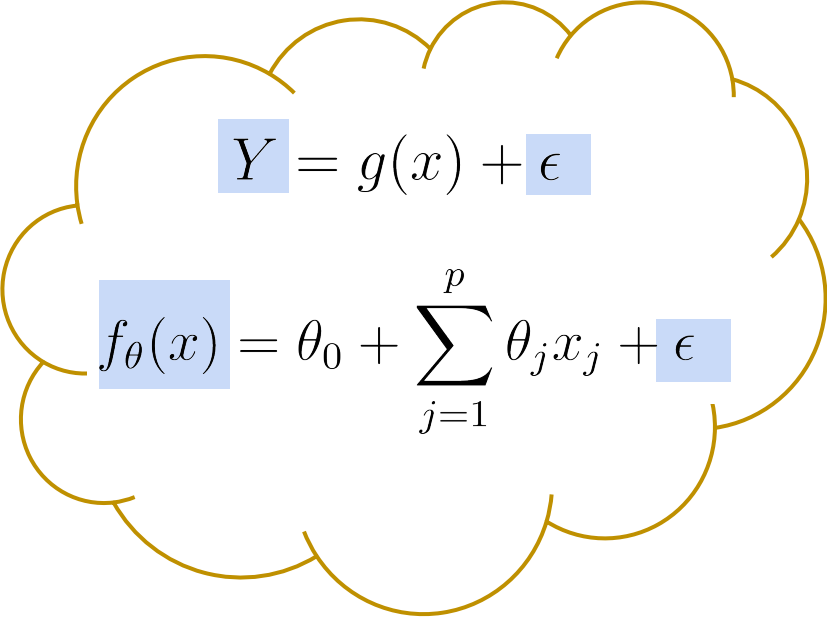
$$f_{\theta}(x) = \theta_0 + \sum_{j=1}^p \theta_j x_j + \epsilon$$

- The true relationship has true, unobservable parameters θ .
- There is still random noise ϵ , so we still can never observe the true relationship.

$$\begin{aligned}\hat{Y}(x) &= f_{\hat{\theta}}(x) \\ &= \hat{\theta}_0 + \sum_{j=1}^p \hat{\theta}_j x_j\end{aligned}$$

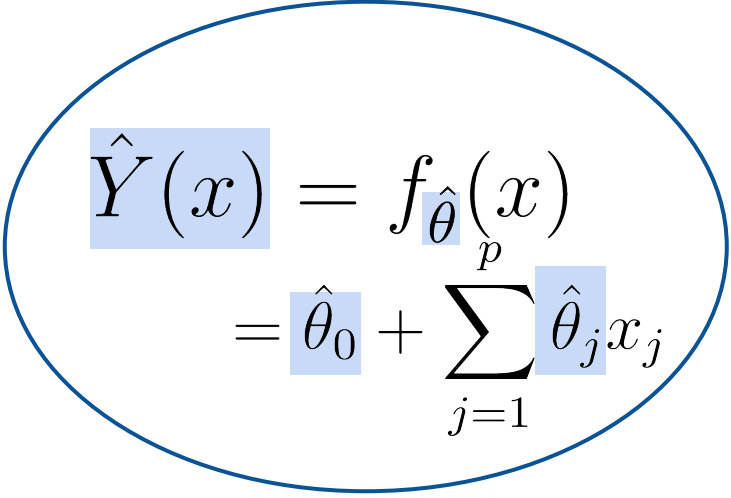
- Obtain a sample \mathbb{X}, \mathbb{Y} of n observed relationships (x, Y) .
- Train a model and obtain estimates $\hat{\theta}$.

Suppose we have an individual with fixed input x . Assume the true relationship g is linear.


$$Y = g(x) + \epsilon$$

$$f_{\theta}(x) = \theta_0 + \sum_{j=1}^p \theta_j x_j + \epsilon$$

- The true relationship has true, unobservable parameters θ .
- There is still random noise ϵ , so we still can never observe the true relationship.


$$\begin{aligned}\hat{Y}(x) &= f_{\hat{\theta}}(x) \\ &= \hat{\theta}_0 + \sum_{j=1}^p \hat{\theta}_j x_j\end{aligned}$$

- Obtain a sample \mathbb{X}, \mathbb{Y} of n observed relationships (x, Y) .
- Train a model and obtain estimates $\hat{\theta}$.

The Bias-Variance Tradeoff

Lecture 17, Data 100 Spring 2022

Sample Statistics (from last time)

Prediction vs. Inference

- Modeling: Assumptions of Randomness

The Bias-Variance Tradeoff

Interpreting Slopes

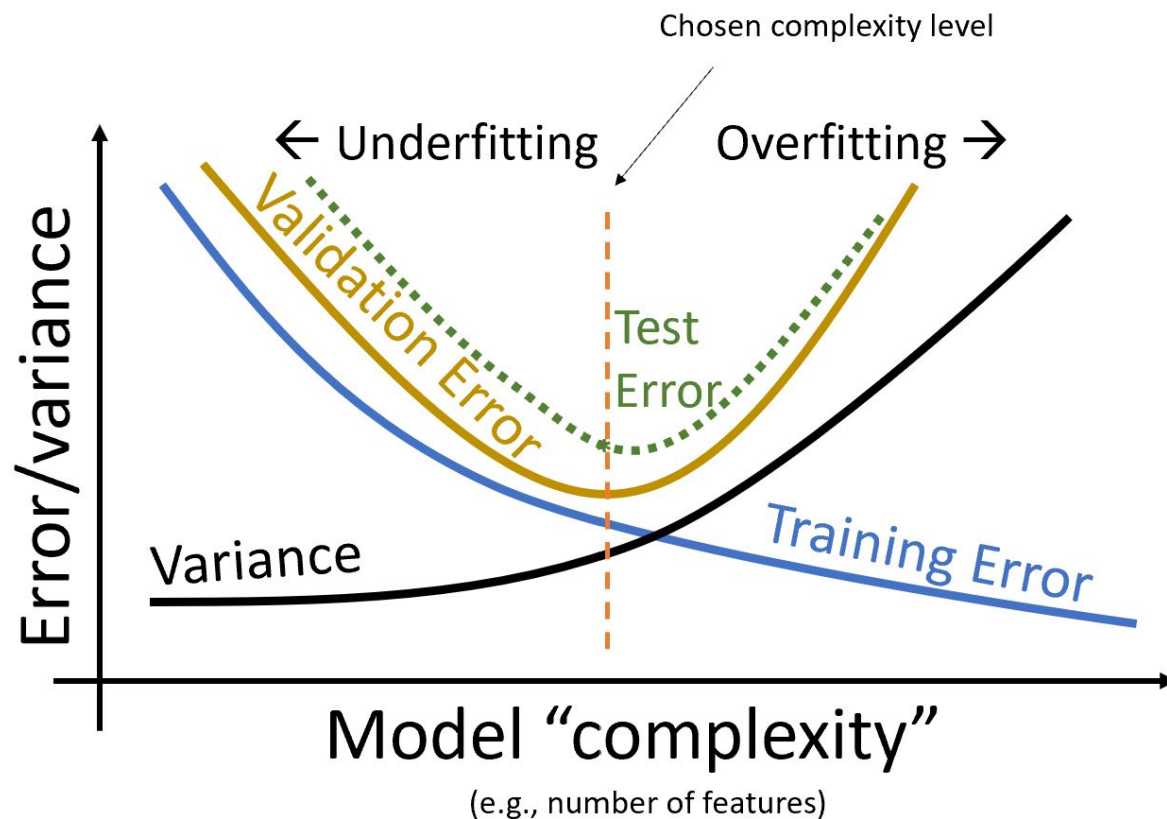
[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance
Decomposition

Linear Algebra Resources: [Ed post](#)

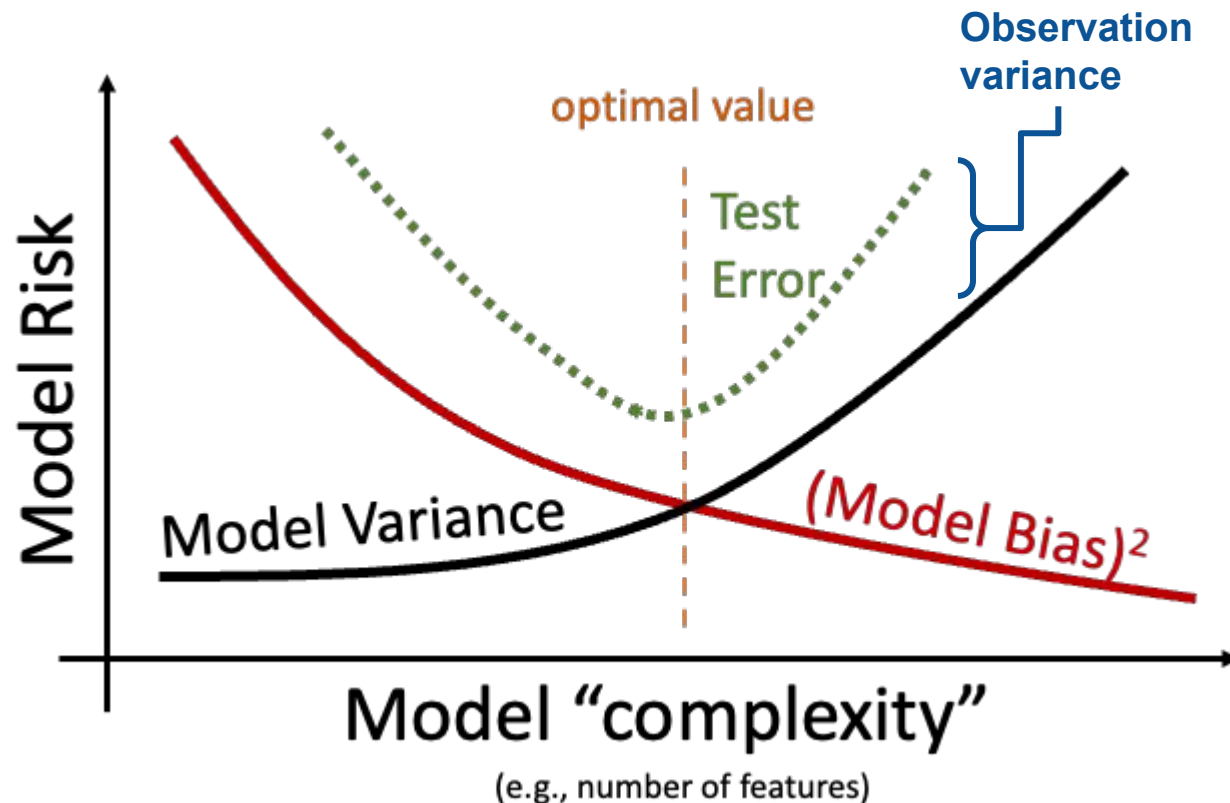
Prediction: The Bias-Variance Tradeoff

With this reformulated modeling goal we can now revisit the Bias-Variance Tradeoff.



The Bias-Variance Tradeoff

This is the abstract version of the plot on the previous slide:



Terms we will define:

- Model Risk
- Observation Variance
- Model Bias
- Model Variance

Model Risk

For a new individual at (x, Y) :

Model Risk is the mean squared prediction error.

$$\text{model risk} = \mathbb{E}[(Y - \hat{Y}(x))^2]$$

Expectation over **multiple** random variables X_1, X_2, \dots, X_n, Y :

- All possible samples we could have gotten when fitting our model
- All possible new observations at this fixed x

For a new individual at (x, Y) :

Model Risk is the mean squared prediction error.

$$\text{model risk} = \mathbb{E}[(Y - \hat{Y}(x))^2]$$

Expectation over **multiple** random variables X_1, X_2, \dots, X_n, Y :

- All possible samples we could have gotten when fitting our model
- All possible new observations at this fixed x

Contrast with numerical functions of the actual collected sample:

$$(\text{L2 loss})_i = (y_i - \hat{y}_i)^2$$

- The i -th collected response $Y = y_i$
- The prediction \hat{y}_i using the model you fit to the collected sample

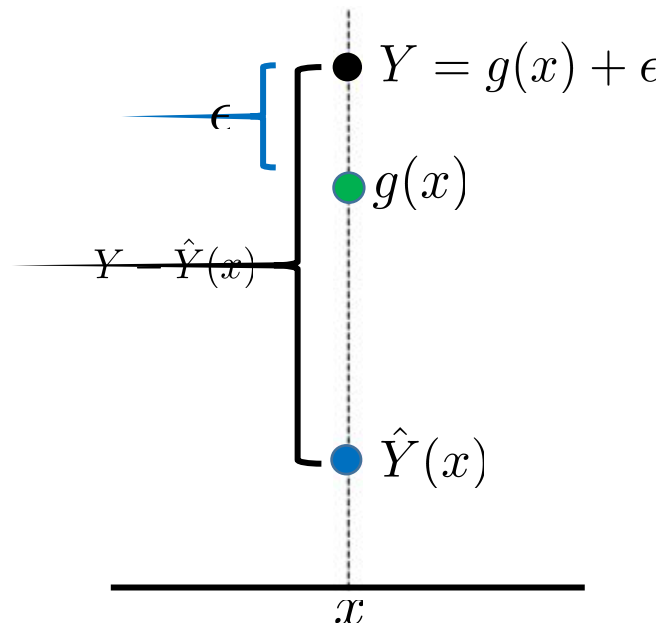
$$\text{empirical risk} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Mean Squared error over all individuals in the collected sample $\{(x_1, y_1), \dots, (x_n, y_n)\}$

The Three Components of Model Risk

There are three types of error that contribute to model risk:

1. **Observation variance**,
because Y has random noise ϵ ;
2. **Model variance**,
because sample X_1, X_2, \dots, X_n is random; and
3. **Model bias**,
because our model is different from
the true underlying function g .



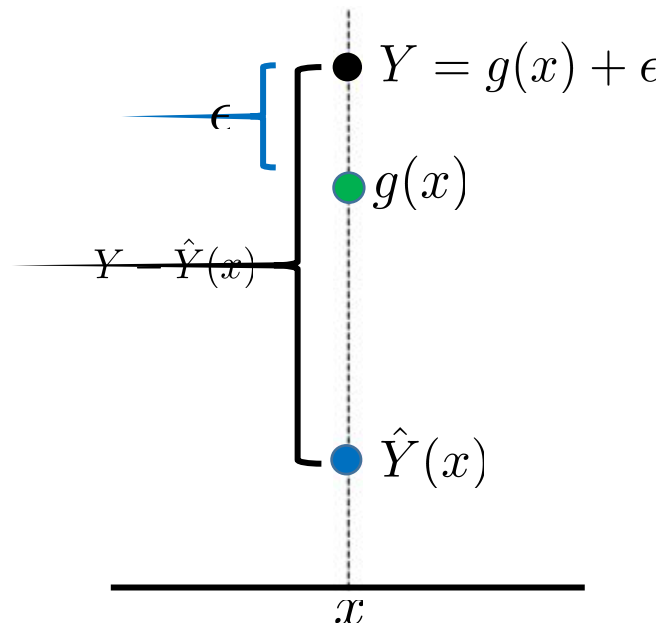
How do you think each of the types of error are encoded into the diagram?



The Three Components of Model Risk

There are three types of error that contribute to model risk:

1. **Observation variance**,
because Y has random noise ϵ ;
2. **Model variance**,
because sample X_1, X_2, \dots, X_n is random; and
3. **Model bias**,
because our model is different from
the true underlying function g .



We'll spend this section **defining** each component of the below equation. If you're interested in the derivation, check out the extra slides.


$$\text{model risk} = \text{observation variance} + (\text{model bias})^2 + \text{model variance}$$

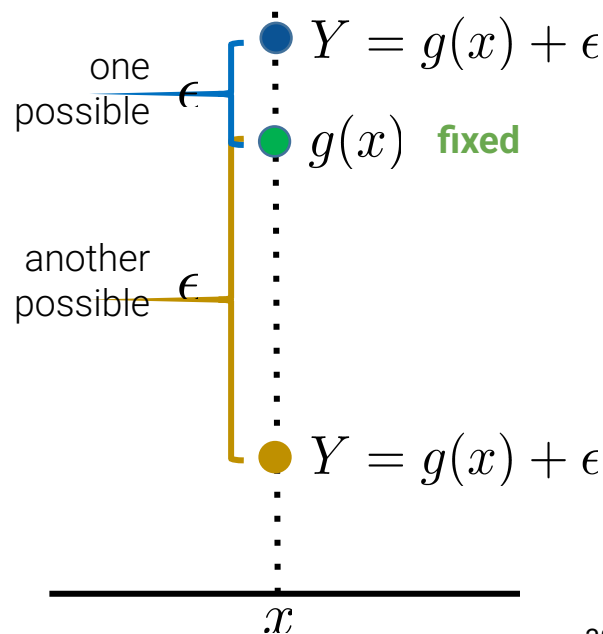
1. Observation Variance

The observation Y is random because by definition, our observation is noisy.

We assume **random error** ϵ to have zero mean and variance σ^2 .

$$Y = g(x) + \epsilon$$

 **random error**



1. Observation Variance

The observation Y is random because by definition, our observation is noisy.

We assume **random error** ϵ to have zero mean and variance σ^2 .

Define **observation variance** as the variance of random error:

$$\text{observation variance} = \sigma^2$$

Reasons:

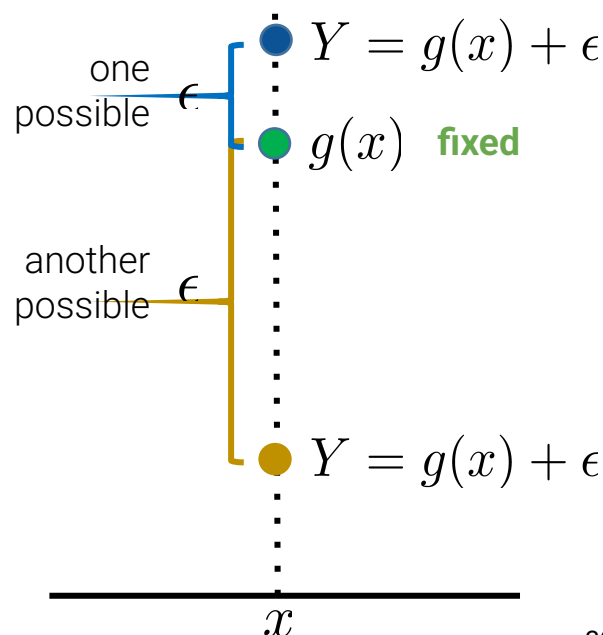
- Measurement error
- Missing information acting like noise

Remedies:

- Could try to get more precise measurements
- But often this is **beyond the control** of the data scientist.

$$Y = g(x) + \epsilon$$

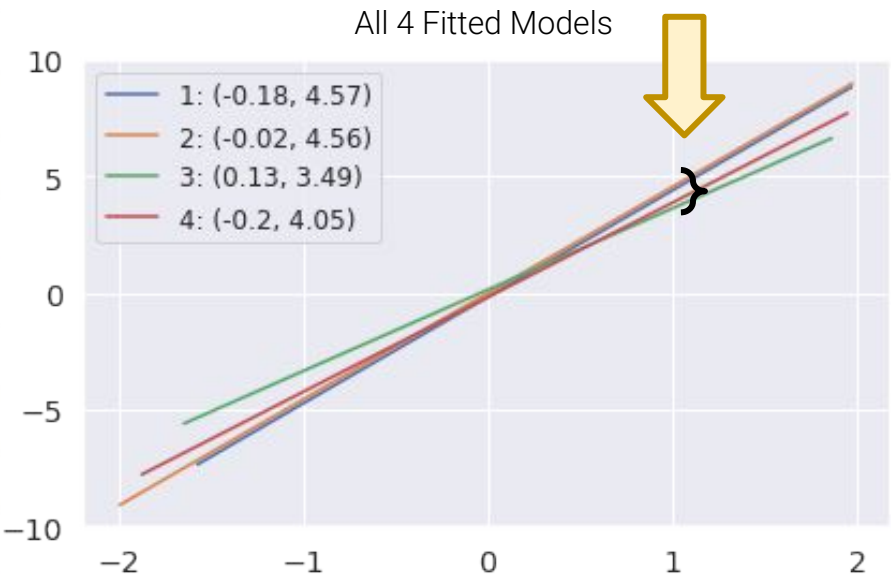
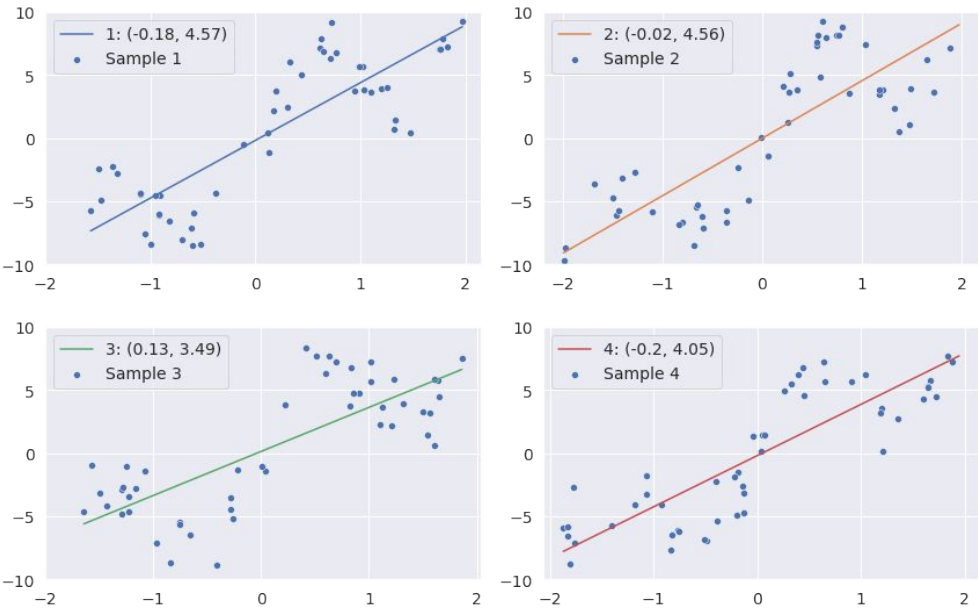
↑
random error



2. Model Variance

Our fitted model is based on a **random sample**.
If the sample came out differently, then the fitted model would have been different.

$\hat{Y}(x)$ Prediction for the individual at x
A random variable



Response vs 1-D x . Fitted SLR model legend: $(\hat{\theta}_0, \hat{\theta}_1)$

2. Model Variance

Our fitted model is based on a **random sample**.

If the sample came out differently, then the fitted model would have been different.

$\hat{Y}(x)$ Prediction for the individual at x .
A random variable

Define the **model variance** as the variance of our prediction at x :

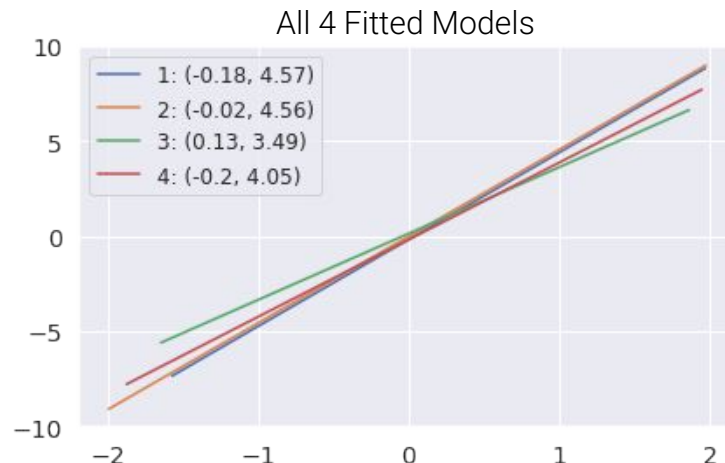
$$\text{model variance} = \text{Var}(\hat{Y}(x))$$

Main Reason:

- **Overfitting**. Small differences in random samples lead to large differences in the fitted model

Remedy:

- Reduce model complexity
- Don't fit the noise



3. Model Bias

Define the **model bias** as the difference between our predicted value and the true $g(x)$.

- The fit of our model (for a linear model, the estimate $\hat{\theta}$) is based on a random sample.
- So model bias is averaged over all possible samples.

$\hat{Y}(x)$ Prediction for the individual at x .
A random variable



3. Model Bias

Define the **model bias** as the difference between our predicted value and the true $g(x)$.

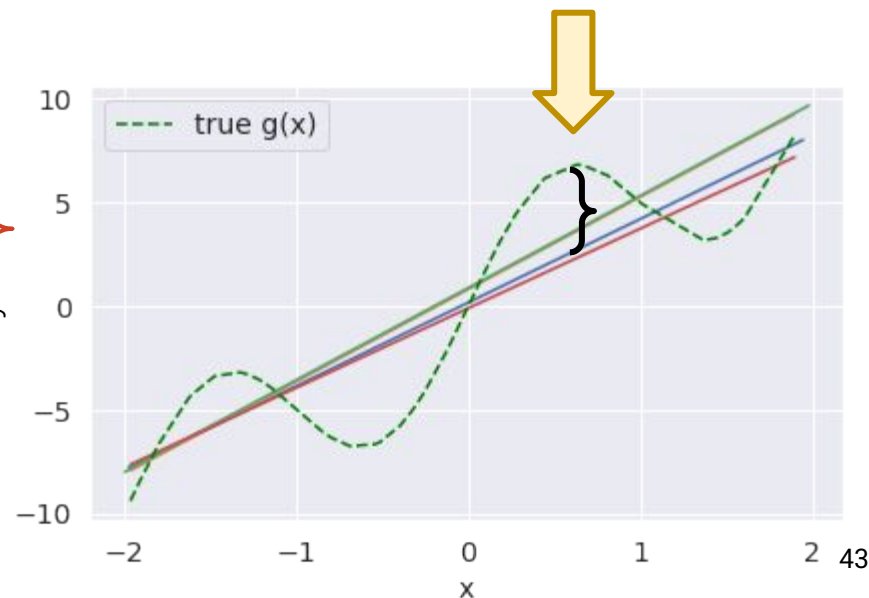
- The fit of our model (for a linear model, the estimate $\hat{\theta}$) is based on a random sample.
- So model bias is averaged over all possible samples.

$\hat{Y}(x)$ Prediction for the individual at x .
A random variable

$$\text{model bias} = \mathbb{E}[\hat{Y}(x)] - g(x)$$

Bias is an average measure for a specific individual x :

- If positive, the model tends to overestimate at this x .
- If negative, the model tends to underestimate at this x .
- If zero, the model is **unbiased**.



3. Model Bias

Define the **model bias** as the difference between our predicted value and the true $g(x)$.

- The fit of our model (for a linear model, the estimate $\hat{\theta}$) is based on a random sample.
- So model bias is averaged over all possible samples.

$\hat{Y}(x)$ Prediction for the individual at x .
A random variable

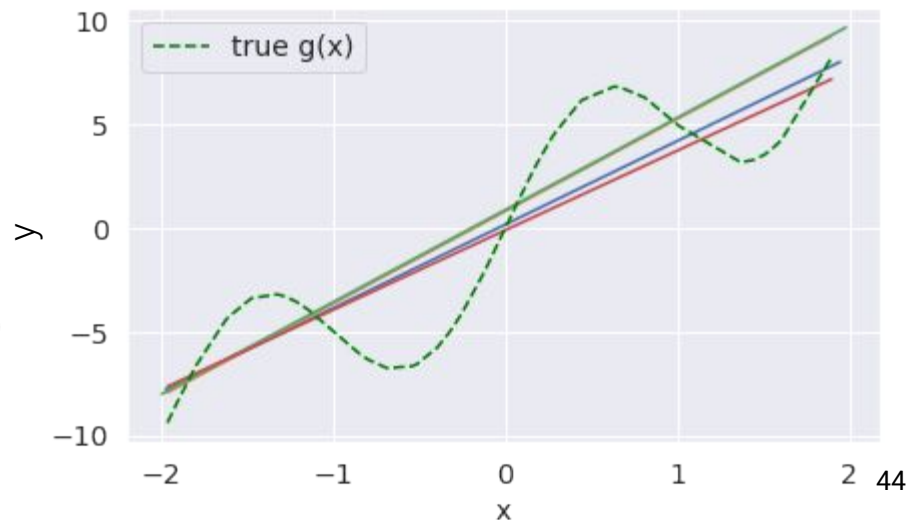
$$\text{model bias} = \mathbb{E}[\hat{Y}(x)] - g(x)$$

Reasons:

- **Underfitting.**
- Lack of domain knowledge.

Remedies:

- Increase model complexity (but don't overfit!)
- Consult domain experts to see which models make sense.



3. [Definition] Unbiased Estimators

$$\text{model bias} = \mathbb{E}[\hat{Y}(x)] - g(x)$$

An **unbiased model** is one where $\text{model bias} = 0$.

In other words, on average, the model predicts $g(x)$.

We can define bias for estimators, too.

For example, the sample mean is an **unbiased estimator** of the population mean.

- By the CLT, $\mathbb{E}[\bar{X}_n] = \mu$.
- Therefore estimator bias $= \mathbb{E}[\bar{X}_n] - \mu = 0$.

The Bias-Variance Decomposition

We've spent this section **defining** each component of the below equation.

$$\text{model risk} = \text{observation variance} + (\text{model bias})^2 + \text{model variance}$$



$$\mathbb{E}[(Y - \hat{Y}(x))^2] = \sigma^2 + \left(\mathbb{E}[\hat{Y}(x)] - g(x)\right)^2 + \text{Var}(\hat{Y}(x))$$

Interpret:

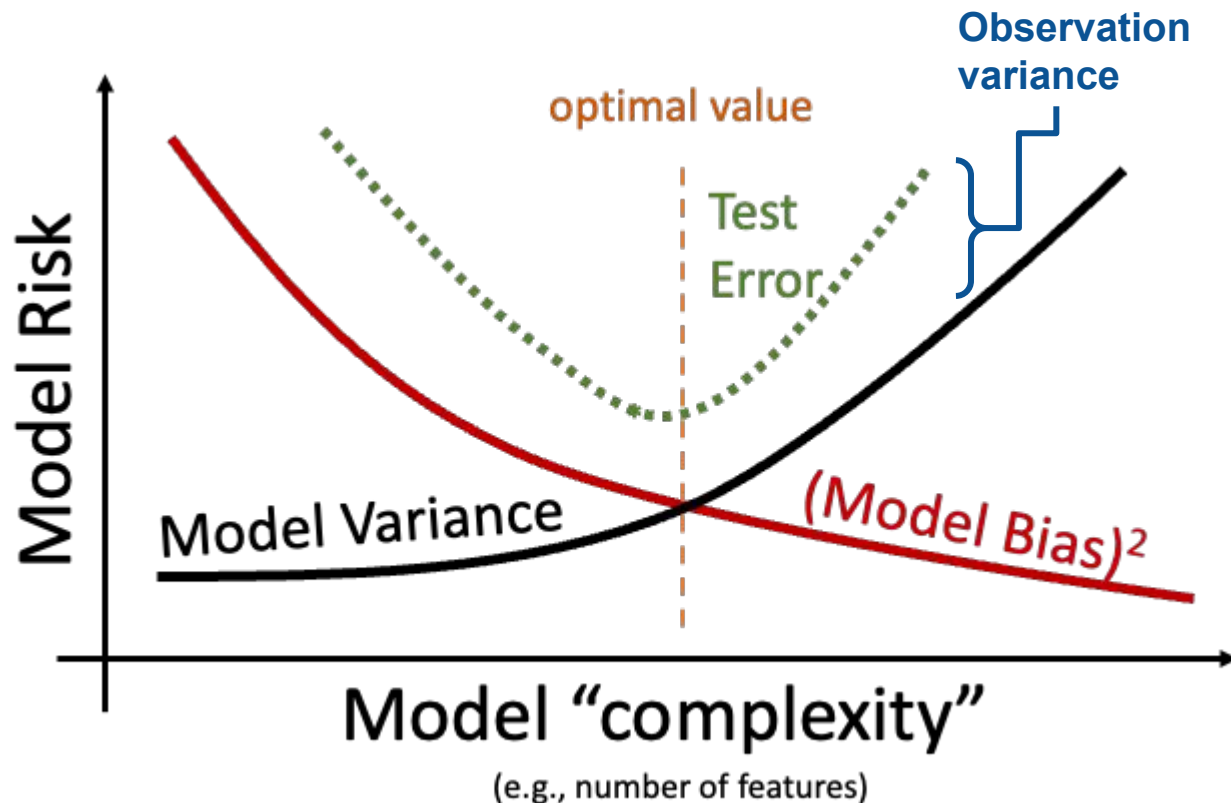
- Model risk is an expectation and is therefore a fixed number (for a given x and model $\hat{Y}(x)$).
- Observation variance is irreducible.
- As models **increase in complexity**, they **overfit** the sample data and will have **higher model variance**. This often corresponds to a decrease in bias.
- As models **decrease in complexity**, they **underfit** the sample data and have lower model variance. This corresponds to an **increase in bias**.

This is the **Bias-Variance Tradeoff**.

Interested in the derivation?
Check out the extra slides!

The Bias-Variance Tradeoff

model risk = observation variance + (model bias)² + model variance



Break (3 min)

Interlude

Interpreting Slopes

Lecture 17, Data 100 Spring 2022

Sample Statistics (from last time)

Prediction vs. Inference

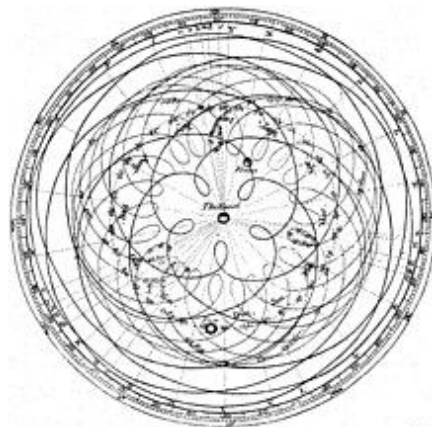
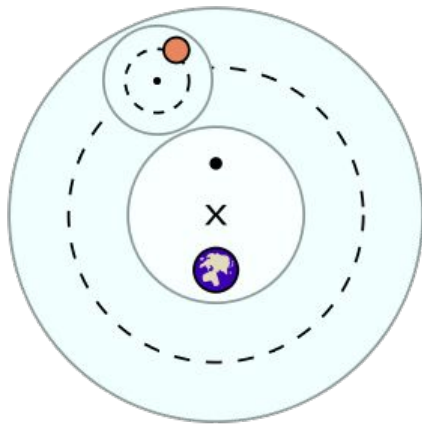
- Modeling: Assumptions of Randomness

The Bias-Variance Tradeoff

Interpreting Slopes

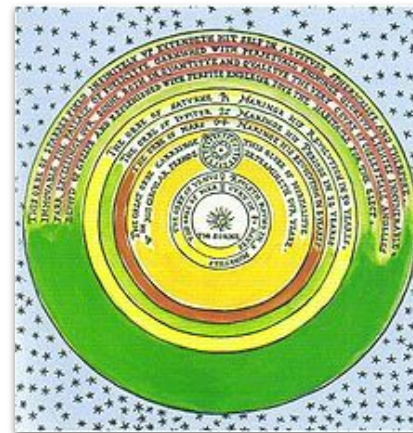
[Extra] Derivation of Bias-Variance
Decomposition

Inference: The right model structure matters!



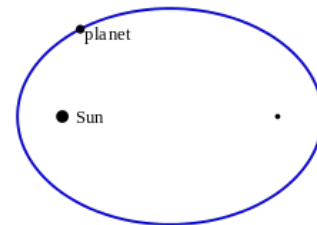
Ptolemaic Astronomy, a geocentric model based on circular orbits (*epicycles* and *deferents*).

High accuracy but very high model complexity.



Copernicus and Kepler: a heliocentric model with elliptical orbits.

Small model complexity yet high accuracy.



Inference for Linear Regression

Assume the true relationship is linear:

$$f_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots + \theta_p x_p + \epsilon$$

Unknown true parameters θ



Our estimation from our sample (design matrix \mathbb{X} , response vector \mathbb{Y}):

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 \cdots + \hat{\theta}_p x_p$$

Estimated parameters $\hat{\theta}$



The meaning of “slope”:

1. What if the true parameter θ_1 is 0?
2. What does the parameter θ_1 even mean?

What can we **infer** about our true parameter given our estimate $\hat{\theta}_1$?

Inferring Zero slope

Our **estimation** from our sample (design matrix \mathbb{X} , response vector \mathbb{Y}):

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 \cdots + \hat{\theta}_p x_p$$



This estimate $\hat{\theta}_1$ for the true parameter θ_1 depends on our sample.

What if the true parameter θ_1 was 0?

Then the feature x_1 has **no effect** on the response!

Approximate confidence interval for true slope

How do we test if the true parameter θ_1 was 0?

- We get one estimate $\hat{\theta}_1$ from our sample of size n .
- But we must imagine all the other random samples that could have happened, and draw our conclusion based on this distribution of estimates.

Enter **hypothesis testing**!

Null hypothesis: The true parameter θ_1 is 0.

Alternative hypothesis: The true parameter θ_1 is not 0.

If your p-value is small, reject the null hypothesis at the cutoff level (say, 5%).

Equivalently ([duality argument](#)):

- Compute an approximate 95% confidence interval with **bootstrapping**.
- If the interval does not contain 0, reject the null hypothesis at the 5% level.
- Otherwise, data are consistent with null hypothesis (the true parameter *could* be 0).

The Snowy Plover

Data on the tiny [Snowy Plover](#) bird was collected by a [former Berkeley student](#) at the Point Reyes National Seashore.

The bigger a newly hatched chick, the more likely it is to survive.



Demo

[Data 100 textbook](#)

Assumed true relationship for newborn weight $Y = f_{\theta}(x)$:

$$f_{\theta}(x) = \theta_0 + \theta_1 \text{egg_weight} + \theta_2 \text{egg_length} + \theta_3 \text{egg_breadth} + \epsilon$$

Estimating the Snowy Plover

Assumed true relationship for newborn weight $Y = f_{\theta}(x)$:

$$f_{\theta}(x) = \theta_0 + \theta_1 \text{egg_weight} + \theta_2 \text{egg_length} + \theta_3 \text{egg_breadth} + \epsilon$$

Estimated model for newborn weight $\hat{Y} = f_{\hat{\theta}}(x)$:

		theta_hat
$\hat{\theta}_0$	intercept	-4.605670
$\hat{\theta}_1$	egg_weight	0.431229
$\hat{\theta}_2$	egg_length	0.066570
$\hat{\theta}_3$	egg_breadth	0.215914

Demo

[Data 100 textbook](#)

Is this the right linear model for newborn weight?

Let's test the **null hypothesis**: The true parameter θ_1 is 0.

Bootstrapped Confidence Interval for θ_1

We can estimate the distribution of $\hat{\theta}_1$ by bootstrapping.

Bootstrap the sample to build an **approximate 95% confidence interval** for the parameter θ_1 :

```
sample_df = ... # call this the bootstrap population
n = len(sample_df)
estimates = []
repeat 10000 times:
    # resample ... ? times with replacement
    resample = ...
    ...
    estimate = ...
    estimates.append(estimate)
lower = np.percentile(estimates, ...)
upper = np.percentile(estimates, ...)
conf_interval = (lower, upper)
```

Demo

[Data 100 textbook](#)

1. (Bootstrap review) Why must we resample **with replacement**?
2. What goes in the blanks?



Bootstrapped Confidence Interval for θ_1

We can estimate the distribution of $\hat{\theta}_1$ by bootstrapping.

Bootstrap the sample to build an **approximate 95% confidence interval** for the parameter θ_1 :

```
sample_df = ... # call this the bootstrap population
n = len(sample_df)
estimates = []
repeat 10000 times:
    # resample n times with replacement
    resample = sample_df.sample(n, replace=True)
    ... # fit the new model to the new resampled X, y
    estimate = get_theta_hat1(model)
    estimates.append(estimate)
lower = np.percentile(estimates, 2.5)
upper = np.percentile(estimates, 97.5)
conf_interval = (lower, upper)
```

Demo

[Data 100 textbook](#)

Bootstrapped Confidence Interval for θ_1

We can estimate the distribution of $\hat{\theta}_1$ by bootstrapping.

Bootstrap the sample to build an **approximate 95% confidence interval** for the parameter θ_1 :

Our bootstrapped 95% confidence interval for the true θ_1 :

$$(-0.262, 1.115)$$

Demo

[Data 100 textbook](#)

We cannot reject the null hypothesis at cutoff 5% (our true parameter θ_1 could be 0).



Are all of our true parameters 0?

Let's bootstrap 95% confidence intervals for all our parameters:

True param		lower	upper
θ_0	intercept	-15.457398	5.518540
θ_1	theta_egg_weight	-0.271299	1.136913
θ_2	theta_egg_length	-0.102671	0.212089
θ_3	theta_egg_breadth	-0.271769	0.765737

Demo

[Data 100 textbook](#)

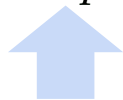


Wait....something's off here!



Our estimation from our sample (design matrix \mathbb{X} , response vector \mathbb{Y}):

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 \cdots + \hat{\theta}_p x_p$$

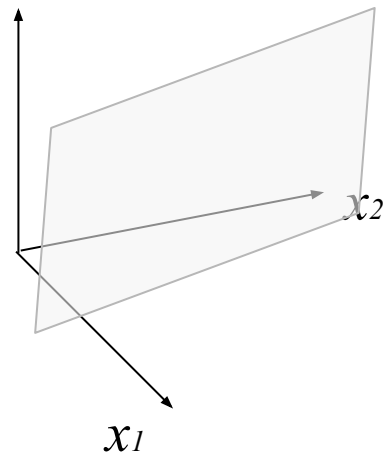


The **slope** $\hat{\theta}_p$ measures the change in y per unit change in x_p ,
provided all the other variables are held constant.

predicted weight = $a_0 + a_1 \cdot \text{length} + a_2 \cdot \text{sleep}$



If two chihuahuas have a 1 inch height difference **and the same hours of sleep**, their estimated weight difference is a_1 .



If variables are **related** to each other, then **interpretation fails!**
E.g., if a change in length always came with a change in sleep

If features are related to each other, it might not be possible to have a change in one of them **while holding the others constant**.

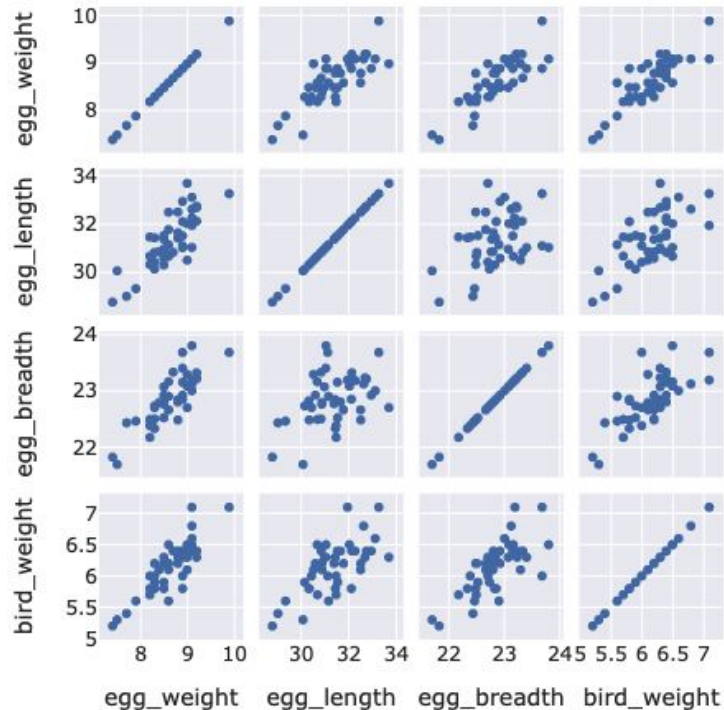
- Then the individual slopes will have no meaning!

Collinearity: When a feature can be predicted pretty accurately by a **linear** function of the others, i.e., the feature is highly correlated with the others.

- Slopes can't be interpreted
- $\mathbb{X}^T \mathbb{X}$ might not be invertible, i.e., solution might not be uniquely determined
- Small changes in the data sample can lead to big changes in the estimated slopes
- Also known as **multicollinearity**

Cross-wise comparison of egg features

Demo



```
px.scatter_  
matrix(eggs)
```

```
eggs.corr()
```

A more interpretable model

If we instead assume a true relationship using only egg weight:

$$f_{\theta}(x) = \theta_0 + \theta_1 \text{egg_weight} + \epsilon$$

	theta_hat	
$\hat{\theta}_0$	intercept	-0.058272
$\hat{\theta}_1$	egg_weight	0.718515

This model performs almost as well as our other model (RMSE 0.0464, old RMSE 0.0454), and the confidence interval for the true parameter θ_1 doesn't contain zero:

(0.604, 0.819)

In retrospect, it's no surprise that the weight of an egg best predicts the weight of a newly-hatched chick.

A model with **highly correlated variables** prevents us from interpreting how the variables are related to the prediction.

Demo

Reminder: Assumptions matter

Keep the following in mind:

- All inference assumes that the regression model holds.
- If the model doesn't hold, the inference might not be valid.
- If the [assumptions of the bootstrap](#) don't hold...
 - Sample size n is large
 - Sample is representative of population distribution (drawn IID, unbiased)...then the results of the bootstrap might not be valid.



This section should be review from Data 8.

The [Data 8 textbook](#) does a fantastic job of teaching bootstrapping if you've never seen it before.

[Extra] Review of the Bootstrap

Lecture 17, Data 100 Spring 2022

Sample Statistics (from last time)

Prediction vs. Inference

- Modeling: Assumptions of Randomness

The Bias-Variance Tradeoff

Interpreting Slopes

[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance Decomposition

Linear Algebra Resources: [Ed post](#)

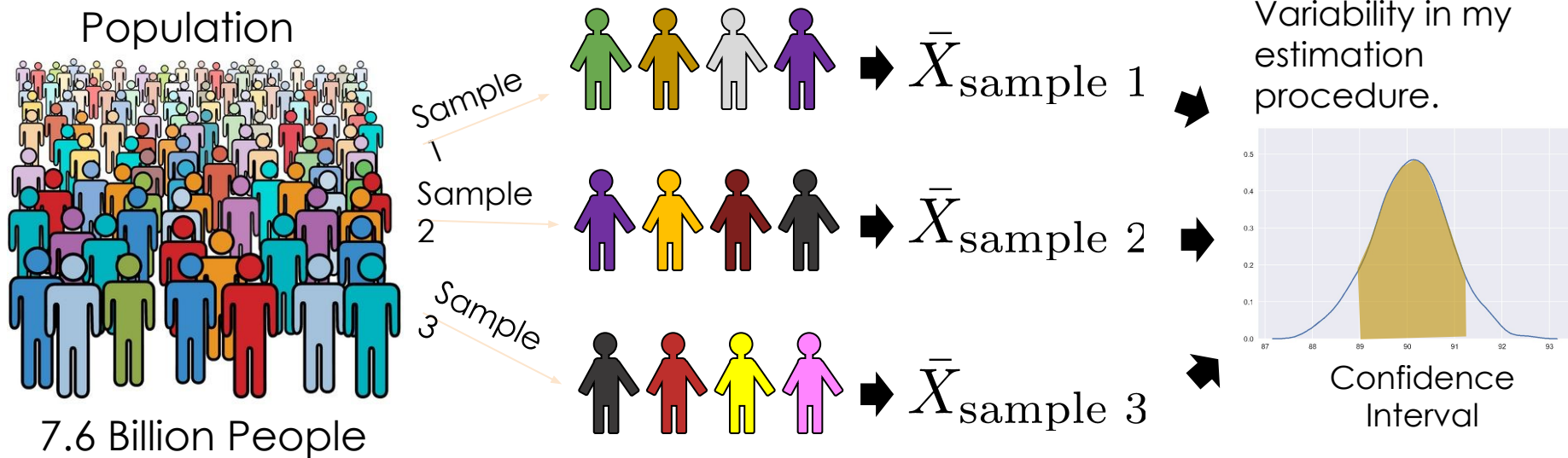
- To determine the properties (e.g. variance) of the sampling distribution of an estimator, we'd need to have access to the population.
 - We would have to consider all possible samples, and compute an estimate for each sample.
- But we don't, we only have one random sample from the population.

Idea: Treat our random sample as a “population”, and resample from it.

- Intuition: a random sample resembles the population, so a random resample resembles a random sample.

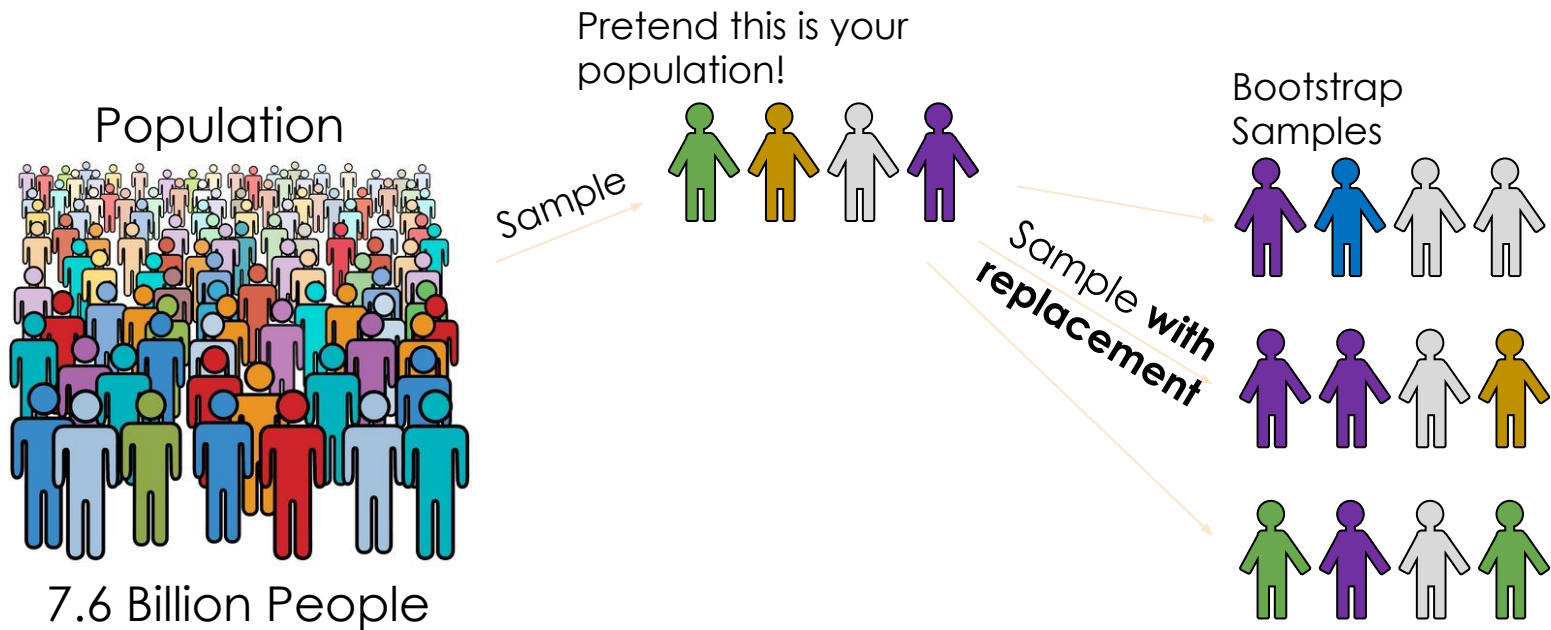
The Distribution of an Estimator

Resampling the population to estimate the sample distribution.



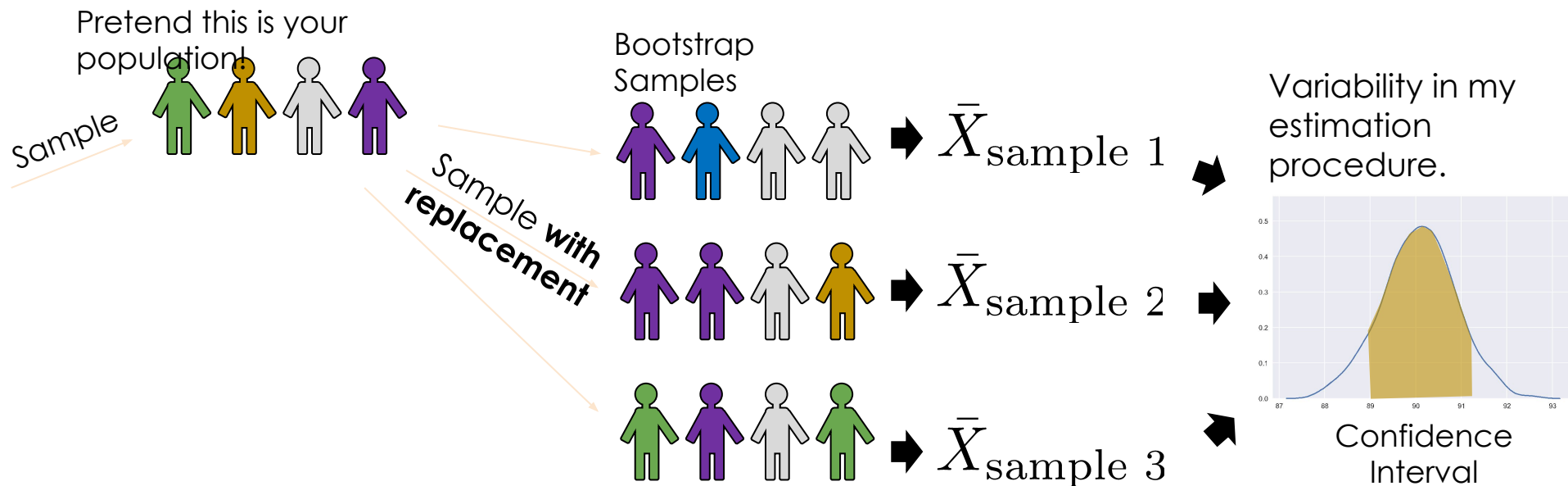
Bootstrap the Distribution of an Estimator

Simulation method to estimate the sample distribution.



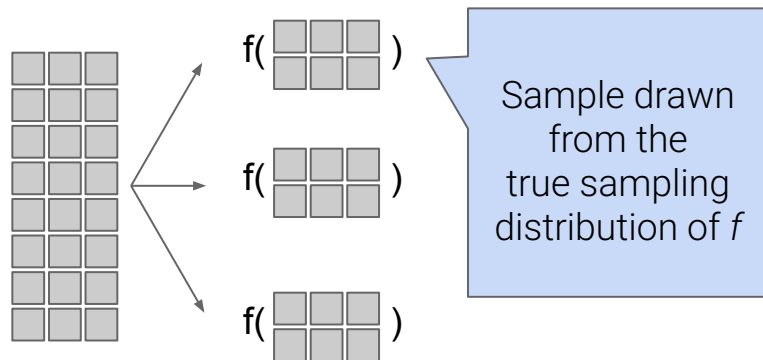
Bootstrap the Distribution of an Estimator

Simulation method to estimate the sample distribution.

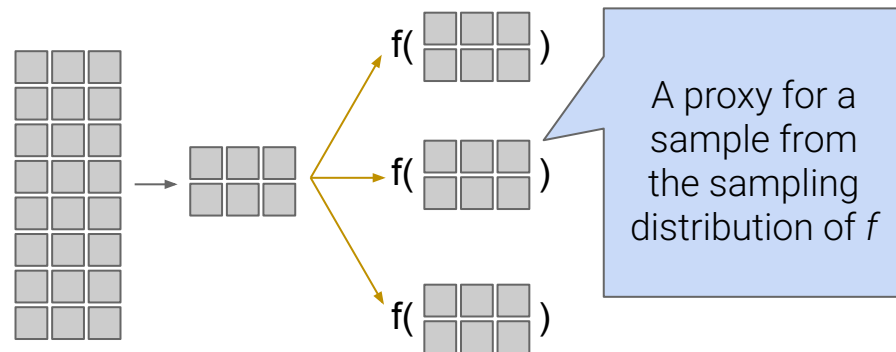


Bootstrap resampling is a technique for estimating the sampling distribution of an estimator.

Impractical:



Bootstrap:



(demo)

Bootstrapping pseudocode

```
collect random sample of size  $n$  (called the bootstrap population)
initiate list of estimates
repeat 10,000 times:
    resample with replacement  $n$  times from bootstrap population
    apply estimator  $f$  to resample
    store in list
list of estimates is the bootstrapped sampling distribution of  $f$ 
```

Why **must** we resample **with replacement**?

The **bootstrapped sampling distribution of an estimator** does not exactly match the **sampling distribution of that estimator**.

- The center and spread are both wrong (but often close).

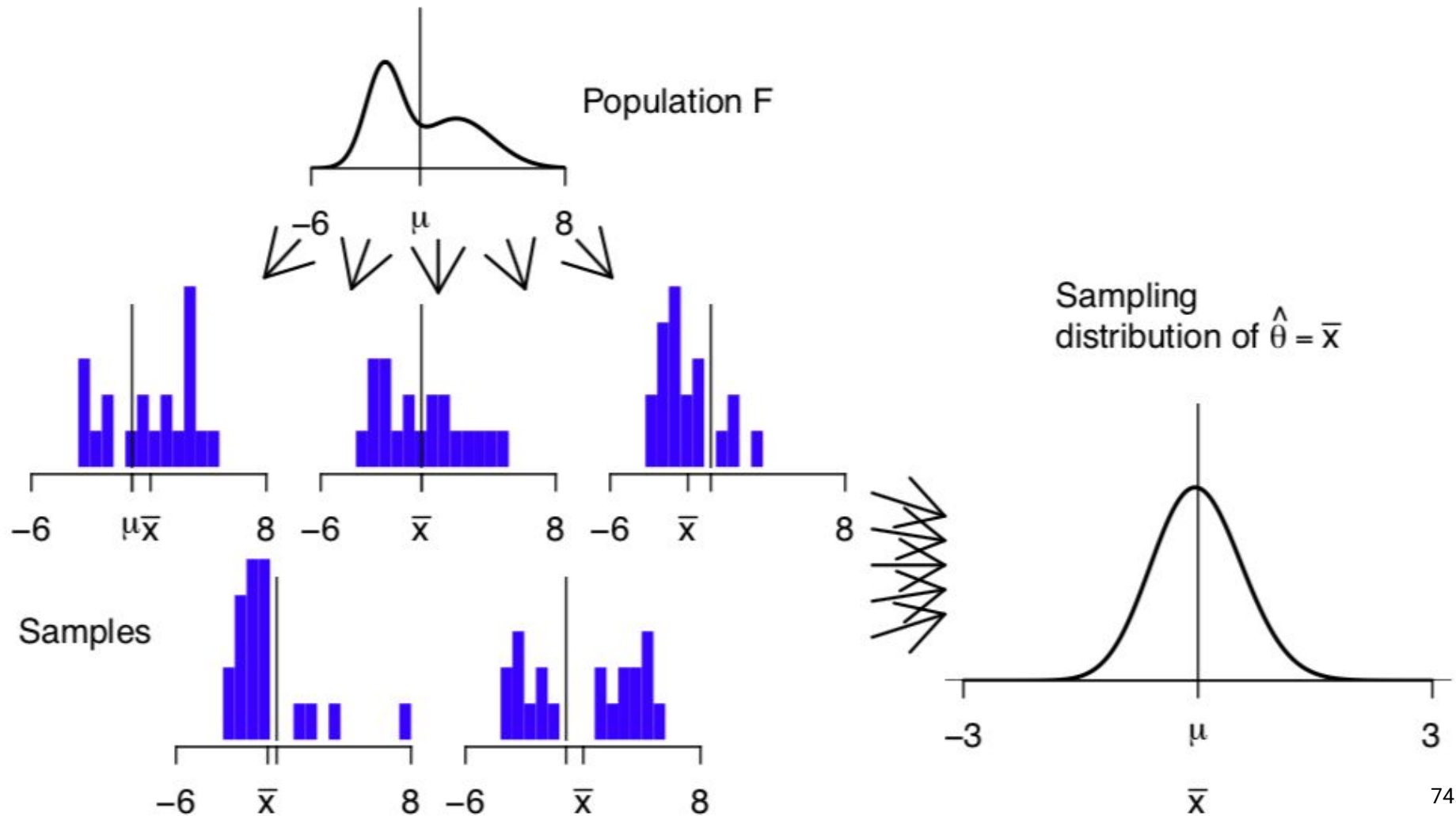
The center of the bootstrapped distribution is the estimator applied to our original sample.

- We have no way of recovering the estimator's true expected value.

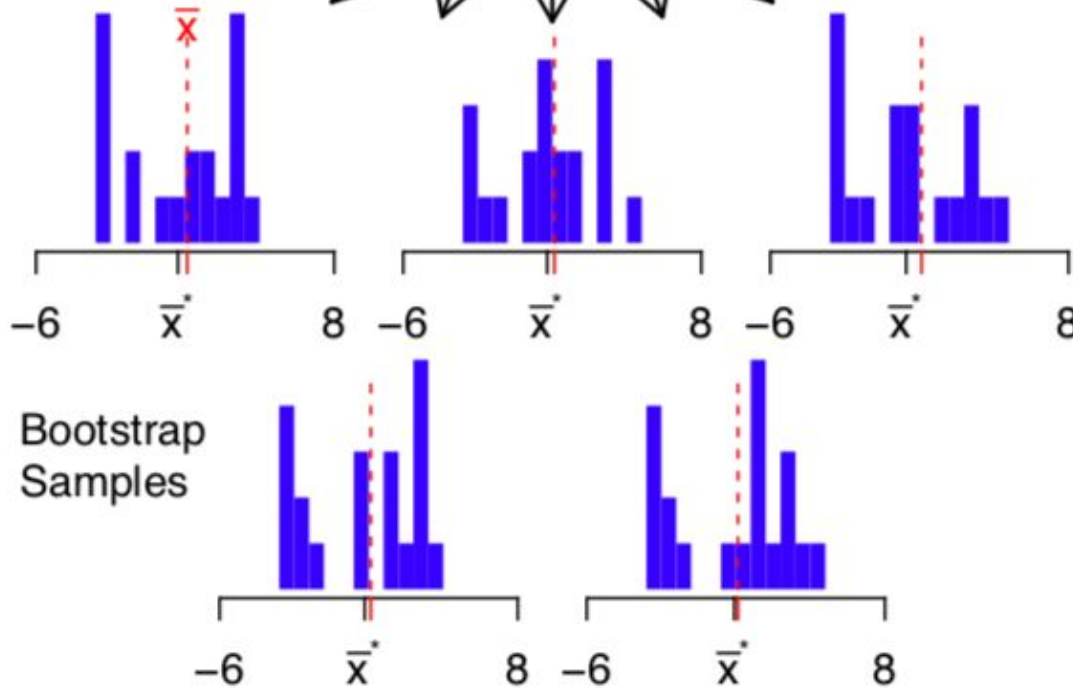
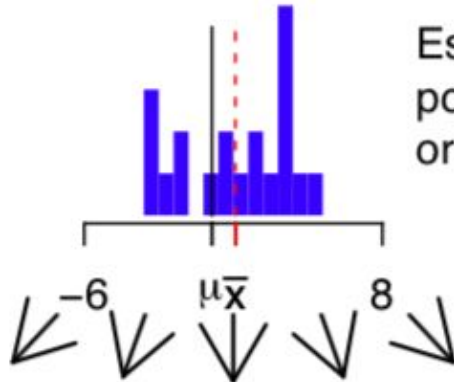
The variance of the bootstrapped distribution is often close to the true variance of the estimator.

The quality of our bootstrapped distribution depends on the quality of our original sample.

- If our original sample was not representative of the population, bootstrap is next to useless.



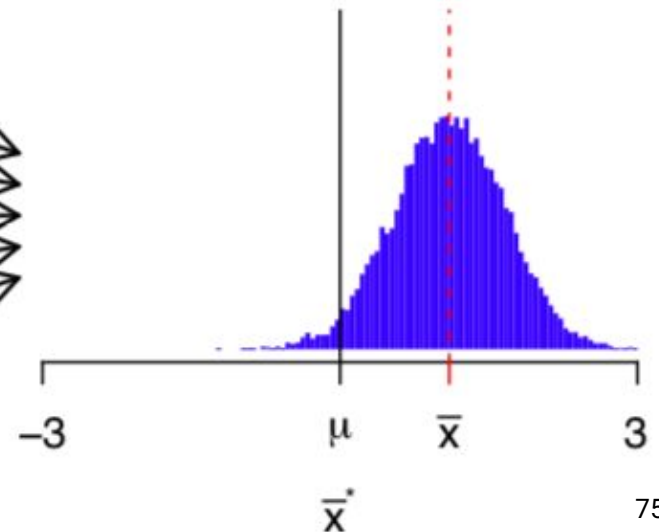
Estimate of
population=
original data \hat{F}



Bootstrap
Samples



Bootstrap
distribution of $\hat{\theta}^* = \bar{X}^*$



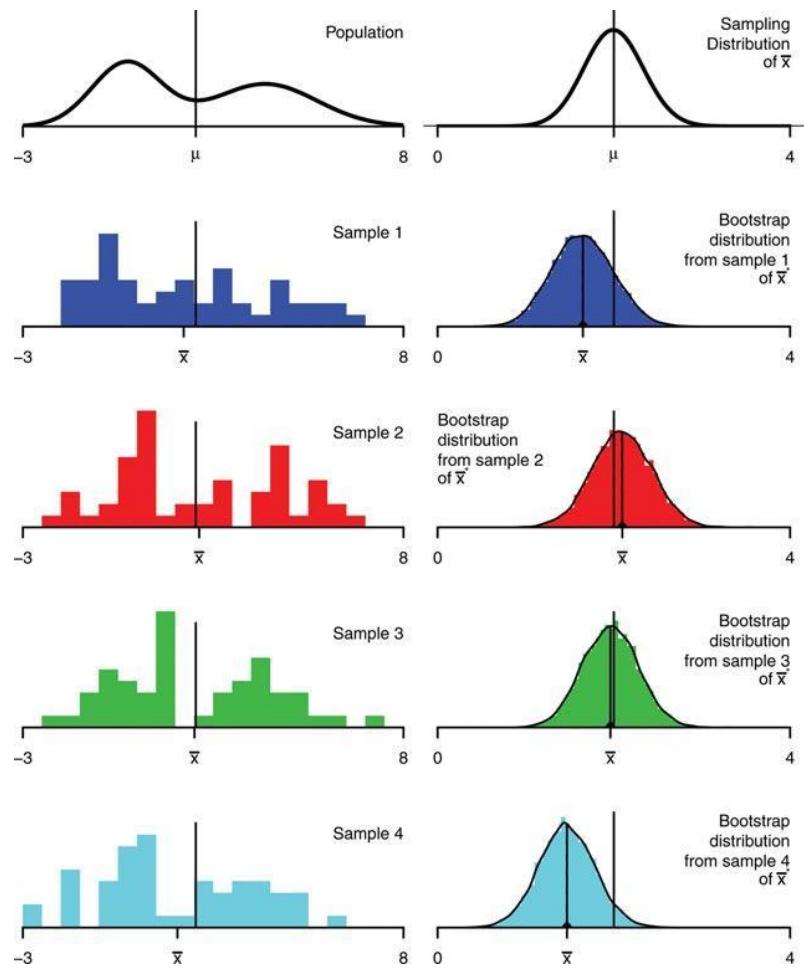
What Teachers Should Know About the Bootstrap

Resampling in the Undergraduate Statistics Curriculum

- The bootstrap is based on the *plug-in principle*—if something is unknown, we substitute an estimate for it.
- Instead of plugging in an estimate for a single parameter, we plug in an estimate for the whole population.
- *The bootstrap distribution is centered at the observed statistic, not the population parameter, for example, at \bar{x} not μ .*
- For example, we cannot use the bootstrap to improve on \bar{x} ; no matter how many bootstrap samples we take, they are centered at \bar{x} , not μ . Instead we use the bootstrap to tell how accurate the original estimate is.

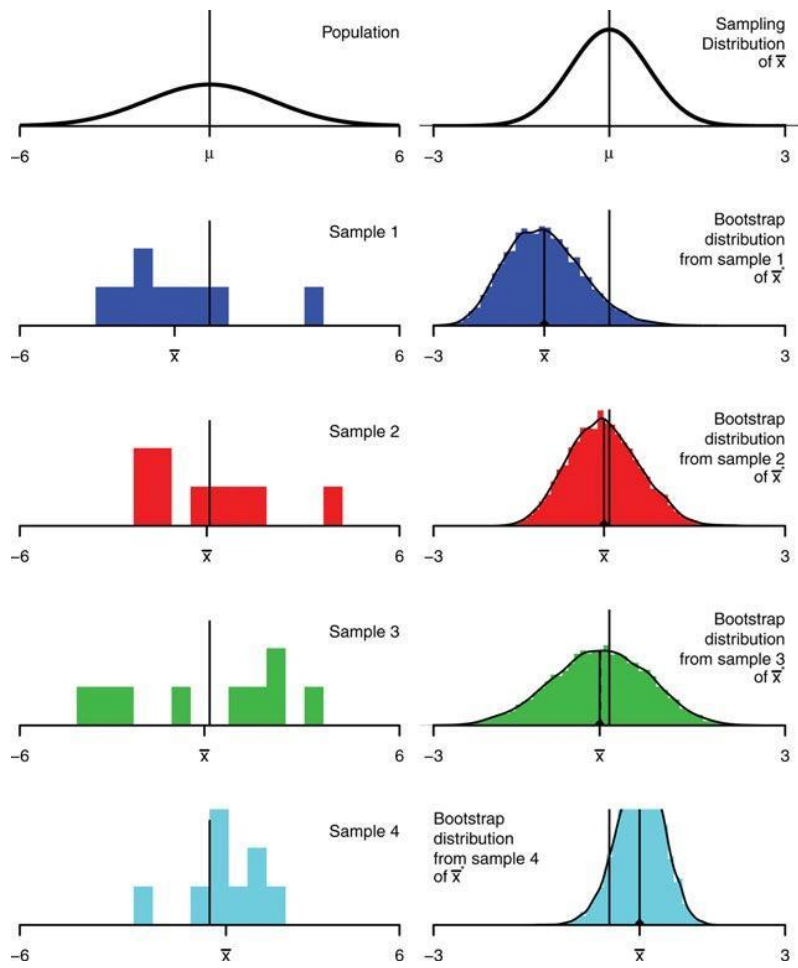
[Tim C. Hesterberg \(2015\)](#)

Bootstrap for the mean, n=50

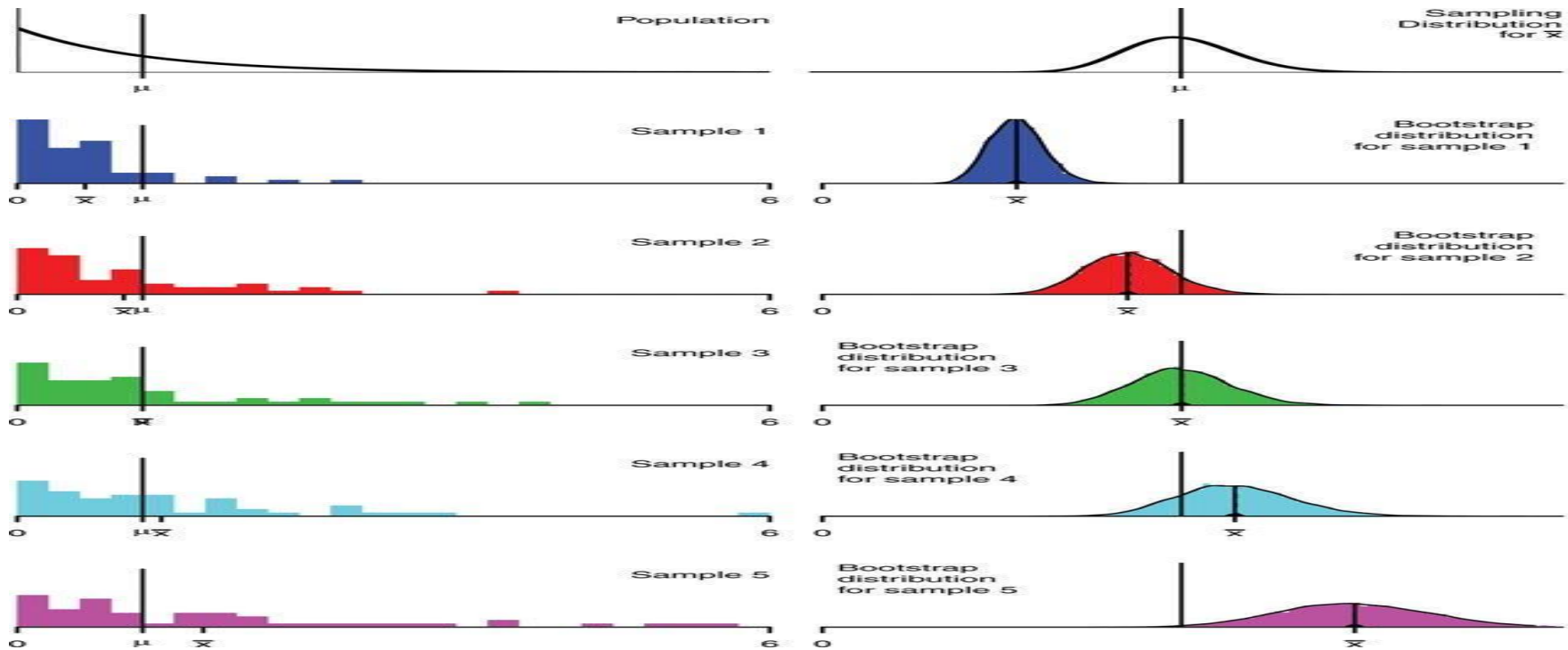


From Tim C. Hesterberg (2015)

Bootstrap distributions for the mean, $n = 9$



Bootstrap distributions for the mean, $n = 50$, exponential population.



The ordinary bootstrap tends not to work well for some statistics:

- Such as the median, or other quantiles in small samples that depend heavily on a small number of observations out of a larger sample.
- The bootstrap depends on the sample accurately reflecting what matters about the population, and those few observations cannot do that.

Bootstrapping does not overcome the weakness of small samples as a basis for inference.

Indeed, for the very smallest samples, it may be better to make additional assumptions such as a parametric family.

These screenshots are here
for your reference.

For more details, please
check the notebook.

[Extra] Derivation of Bias-Variance Decomposition

Lecture 17, Data 100 Spring 2022

Sample Statistics (from last time)

Prediction vs. Inference

- Modeling: Assumptions of Randomness

The Bias-Variance Tradeoff

Interpreting Slopes

[Extra] Review of the Bootstrap

**[Extra] Derivation of Bias-Variance
Decomposition**

Linear Algebra Resources: [Ed post](#)

For more details, please check the notebook. These screenshots are here for your reference.

1.0.2 Preliminary

Before proceeding with this derivation, you should be familiar with the Random Variables lecture (Lecture 16 in Spring 2022). In particular, you really need to understand expectation and variance.

This result will be used below. You don't have to know how to prove it.

If V and W are independent random variables then $\mathbb{E}(VW) = \mathbb{E}(V)\mathbb{E}(W)$.

Proof: We'll do this in the discrete finite case. Trust that it's true in greater generality.

The job is to calculate the weighted average of the values of VW , where the weights are the probabilities of those values. Here goes.

$$\begin{aligned}\mathbb{E}(VW) &= \sum_v \sum_w vwP(V = v \text{ and } W = w) \\ &= \sum_v \sum_w vwP(V = v)P(W = w) \quad \text{by independence} \\ &= \sum_v vP(V = v) \sum_w wP(W = w) \\ &= \mathbb{E}(V)\mathbb{E}(W)\end{aligned}$$

$$\begin{aligned}\text{model risk} &= \mathbb{E}((Y - \hat{Y}(x))^2) \\ &= \mathbb{E}((g(x) + \epsilon - \hat{Y}(x))^2) \\ &= \mathbb{E}((\epsilon + (g(x) - \hat{Y}(x)))^2) \\ &= \mathbb{E}(\epsilon^2) + 2\mathbb{E}(\epsilon(g(x) - \hat{Y}(x))) + \mathbb{E}((g(x) - \hat{Y}(x))^2)\end{aligned}$$

On the right hand side:

- The first term is the observation variance σ^2 .
- The cross product term is 0 because ϵ is independent of $g(x) - \hat{Y}(x)$ and $\mathbb{E}(\epsilon) = 0$
- The last term is the mean squared difference between our predicted value and the value of the true function at x

Derivation: Step 2

At this stage we have

$$\text{model risk} = \text{observation variance} + \mathbb{E}((g(x) - \hat{Y}(x))^2)$$

We don't yet have a good understanding of $g(x) - \hat{Y}(x)$. But we do understand the deviation $D_{\hat{Y}(x)} = \hat{Y}(x) - \mathbb{E}(\hat{Y}(x))$. We know that

- $\mathbb{E}(D_{\hat{Y}(x)}) = 0$
- $\mathbb{E}(D_{\hat{Y}(x)}^2) = \text{model variance}$

So let's add and subtract $\mathbb{E}(\hat{Y}(x))$ and see if that helps.

$$g(x) - \hat{Y}(x) = (g(x) - \mathbb{E}(\hat{Y}(x))) + (\mathbb{E}(\hat{Y}(x)) - \hat{Y}(x))$$

The first term on the right hand side is the model bias at x . The second term is $-D_{\hat{Y}(x)}$. So

$$g(x) - \hat{Y}(x) = \text{model bias} - D_{\hat{Y}(x)}$$

Derivation: Step 3

Remember that the model bias at x is a constant, not a random variable. Think of it as your favorite number, say 10. Then

$$\begin{aligned}\mathbb{E}((g(x) - \hat{Y}(x))^2) &= \text{model bias}^2 - 2(\text{model bias})\mathbb{E}(D_{\hat{Y}(x)}) + \mathbb{E}(D_{\hat{Y}(x)}^2) \\ &= \text{model bias}^2 - 0 + \text{model variance} \\ &= \text{model bias}^2 + \text{model variance}\end{aligned}$$

Derivation: Step 4 (Bias-Variance Decomposition)

In Step 2 we had

$$\text{model risk} = \text{observation variance} + \mathbb{E}((g(x) - \hat{Y}(x))^2)$$

Step 3 showed

$$\mathbb{E}((g(x) - \hat{Y}(x))^2) = \text{model bias}^2 + \text{model variance}$$

Thus we have shown the bias-variance decomposition

$$\text{model risk} = \text{observation variance} + \text{model bias}^2 + \text{model variance}$$

That is,

$$\mathbb{E}((Y - \hat{Y}(x))^2) = \sigma^2 + \mathbb{E}((g(x) - \mathbb{E}(\hat{Y}(x)))^2) + \mathbb{E}((\hat{Y}(x) - \mathbb{E}(\hat{Y}(x)))^2)$$

Derivation: Special case for parameterized models

In the case where we are making our predictions by fitting some function f that involves parameters θ , our estimate \hat{Y} is $f_{\hat{\theta}}$ where $\hat{\theta}$ has been estimated from the data and hence is random.

In the bias-variance decomposition

$$\mathbb{E}((Y - \hat{Y}(x))^2) = \sigma^2 + \mathbb{E}((g(x) - \mathbb{E}(\hat{Y}(x)))^2) + \mathbb{E}((\hat{Y}(x) - \mathbb{E}(\hat{Y}(x)))^2)$$

just plug in the particular prediction $f_{\hat{\theta}}$ in place of the general prediction \hat{Y} :

$$\mathbb{E}((Y - f_{\hat{\theta}}(x))^2) = \sigma^2 + \mathbb{E}((g(x) - \mathbb{E}(f_{\hat{\theta}}(x)))^2) + \mathbb{E}((f_{\hat{\theta}}(x) - \mathbb{E}(f_{\hat{\theta}}(x)))^2)$$

model risk = observation variance + model bias + model variance

That is,

$$\mathbb{E}((Y - \hat{Y}(x))^2) = \sigma^2 + \mathbb{E}((g(x) - \mathbb{E}(\hat{Y}(x)))^2) + \mathbb{E}((\hat{Y}(x) - \mathbb{E}(\hat{Y}(x)))^2)$$

LECTURE 17

Estimators, Bias, and Variance

Content credit: Lisa Yan, Suraj Rampure, Ani Adhikari, Deborah Nolan,
Joseph Gonzalez