LECTURE 20

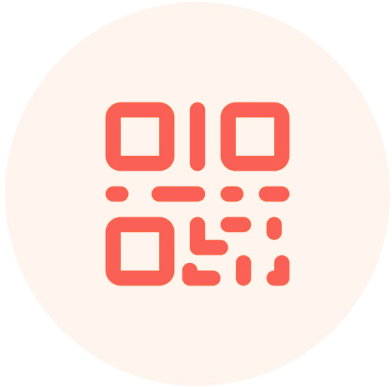
Bias, Variance, and Inference

Bias-Variance Tradeoff, regression coefficients, causal effects, and knowing the difference.

Data 100/Data 200, Spring 2023 @ UC Berkeley

Narges Norouzi and Lisa Yan

slido



Join at slido.com
#1352355

ⓘ Start presenting to display the joining instructions on this slide.

#1352355

Model Notation Review

Lecture 20, Data 100 Spring 2023

Model Notation Review

The Bias-Variance Tradeoff

Interpreting Regression Coefficients

Bootstrapping Test for a Regression Coefficient

Collinearity

Correlation vs. Causation

[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance Decomposition

Modeling: Estimating a Relationship

#1352355

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

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

$$\hat{Y}(x)$$

Ideally, we'd 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 based on our observed sample of (x, y) pairs. Our model **estimates** the true relationship g.

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

The model's prediction $\hat{Y}(x)$ is **random** because our sample is random.

#1352355

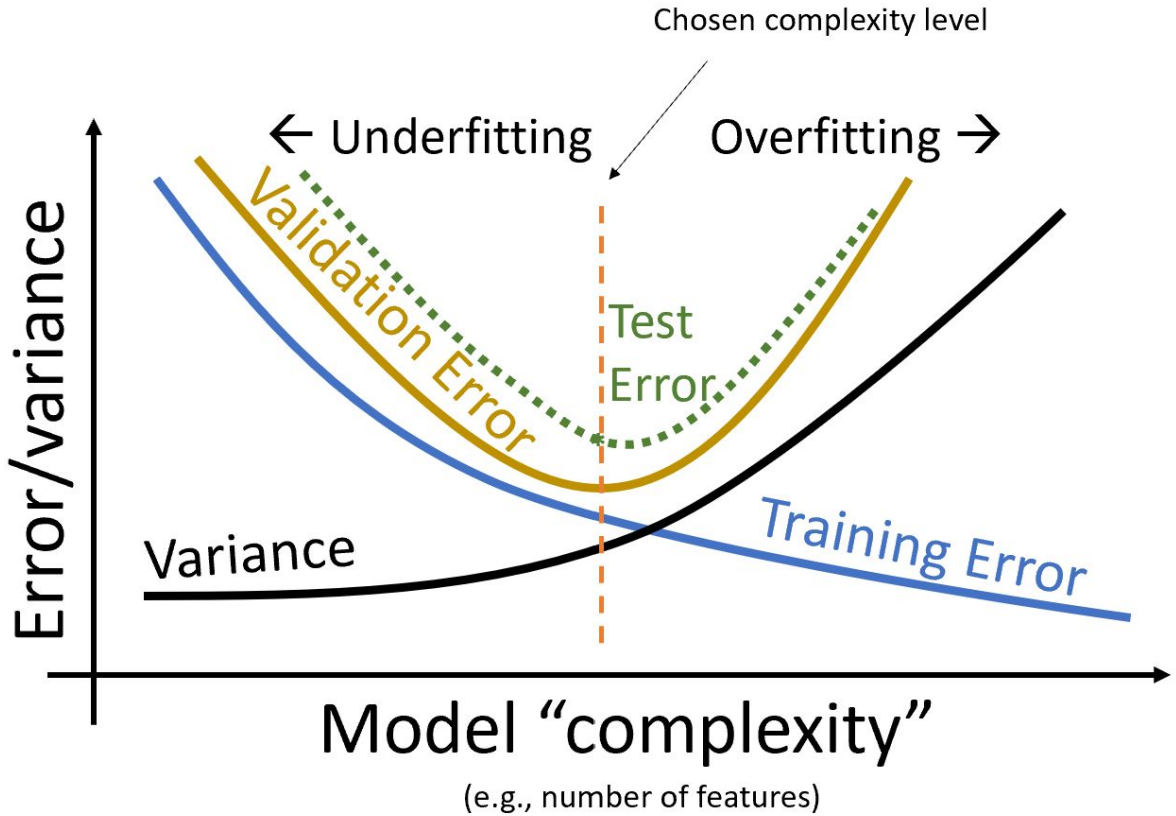
The Bias-Variance Tradeoff

Lecture 20, Data 100 Spring 2023

Prediction: The Bias-Variance Tradeoff

#1352355

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

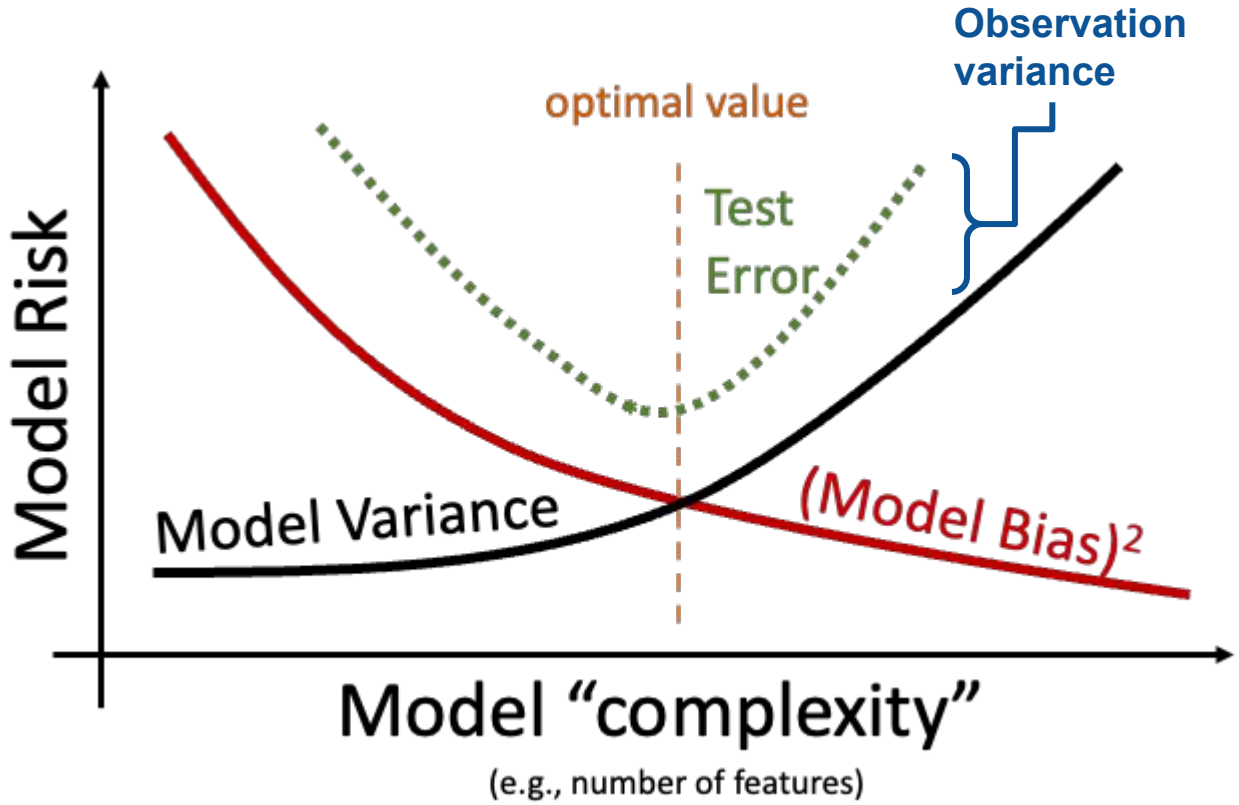


Covered up to here on 14/3.
Will cover the rest of the slides in lecture 20 on 23/3

The Bias-Variance Tradeoff

#1352355

This is the more mathematical 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 $X_1, X_2, \ldots, X_n, Y$:

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

slido

What is the difference between model risk and empirical risk?

ⓘ Start presenting to display the poll results on this slide.

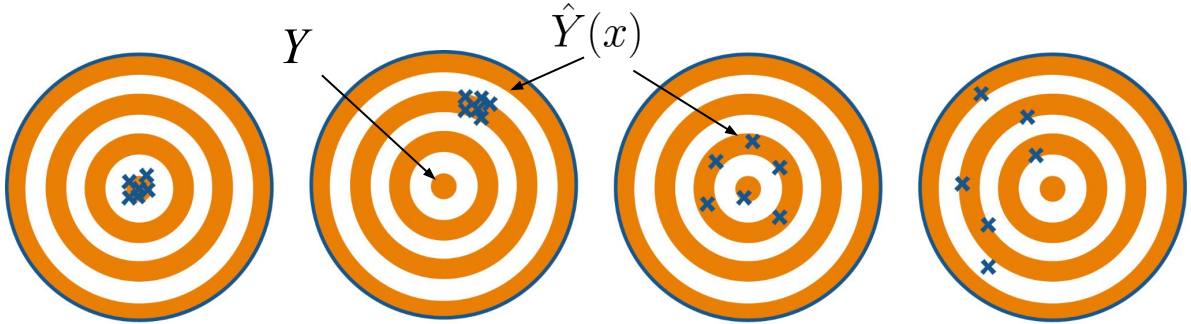
Performance of an Estimator

#1352355

Suppose we want to estimate a target Y using an estimator $\hat{Y}(x)$.

How good is the estimator? Questions we might ask:

- Do we get the right answer on average? **(Bias)**
- How variable is the answer? **(Variance)**
- How close do we get to Y ? **(Risk / MSE)**

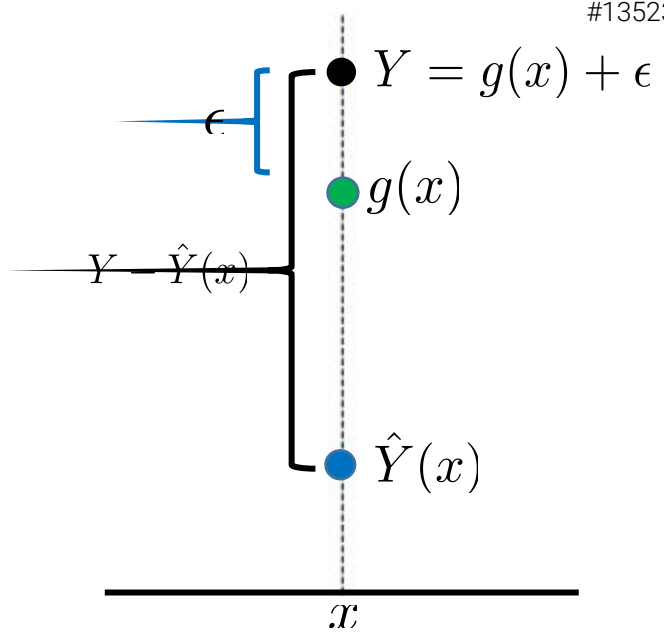


The Three Components of Model Risk

#1352355

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 $X_1, X_2, \ldots, X_n$ is random; and
3. **Model bias**,
   because our model is different from
   the true underlying function g.



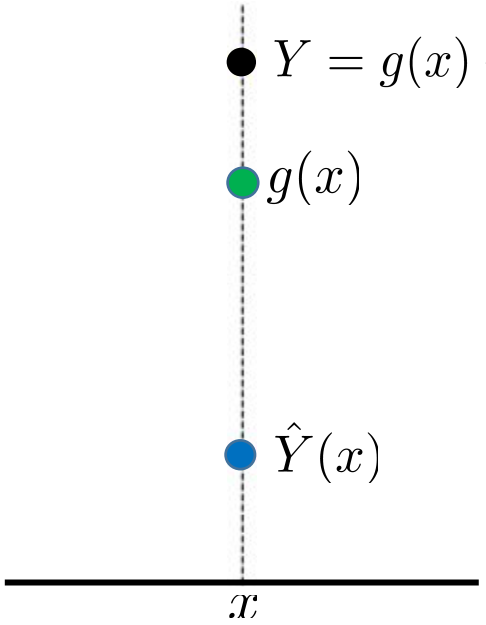
How do you think each of the types of error are encoded into the diagram?

(no slido)

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 $X_1, X_2, \ldots, X_n$ 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

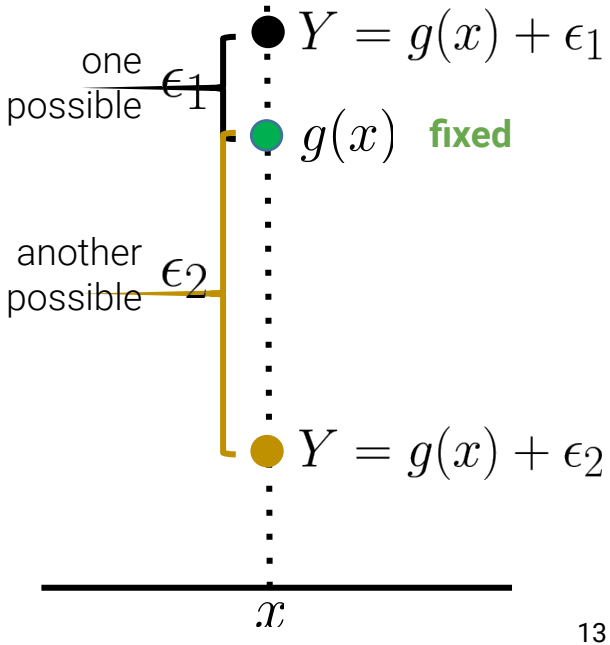
#1352355

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.

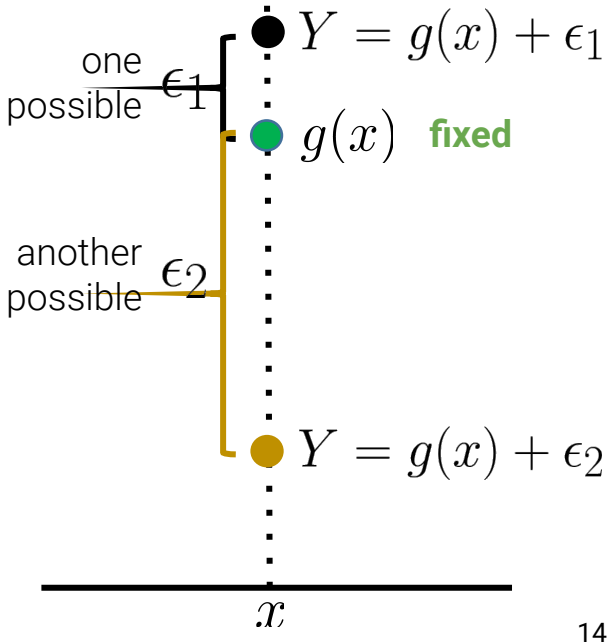
We assume random error ϵ to have zero mean and variance σ^2.

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

random error

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.

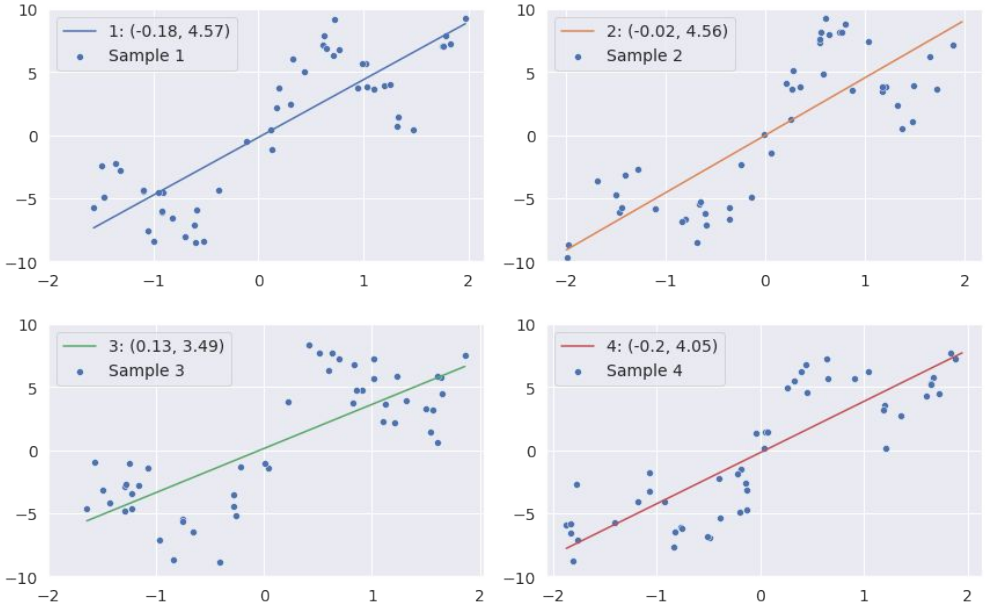
2. Model Variance

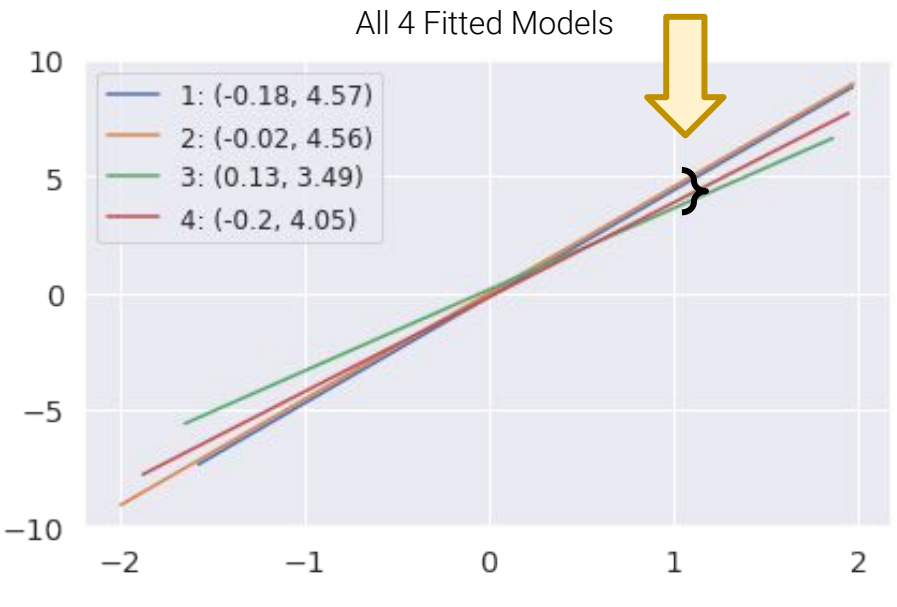
$\hat{Y}(x)$ Prediction for individual x
A random variable

#1352355

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

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





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

2. Model Variance

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

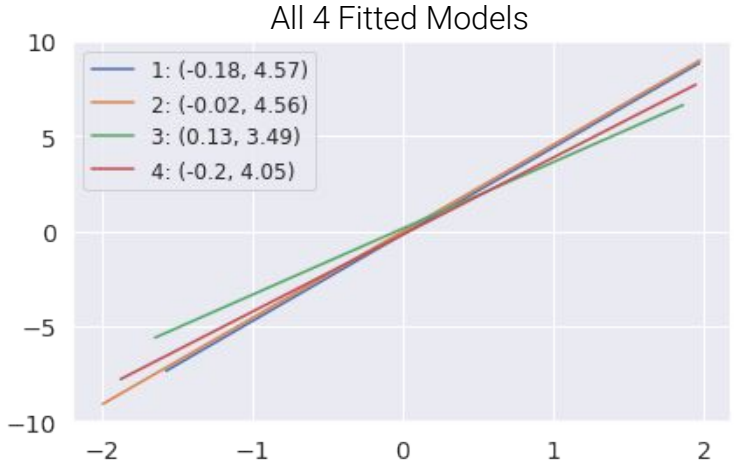
#1352355

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

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

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

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



2. Model Variance

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

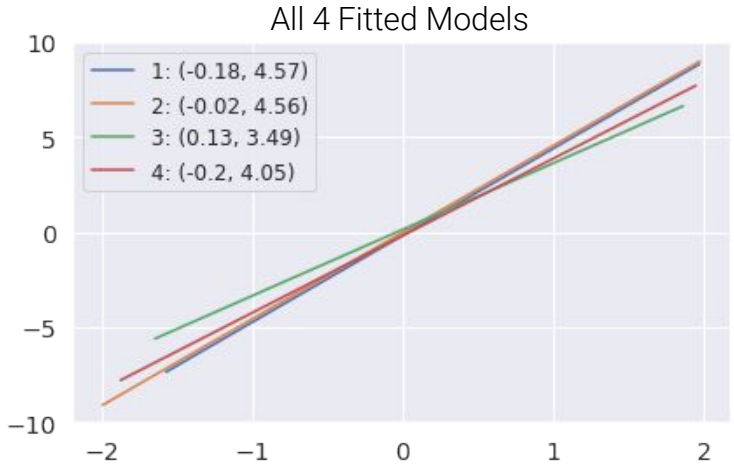
#1352355

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

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

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

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



Main Reason:

- Different samples → different model estimates
- **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

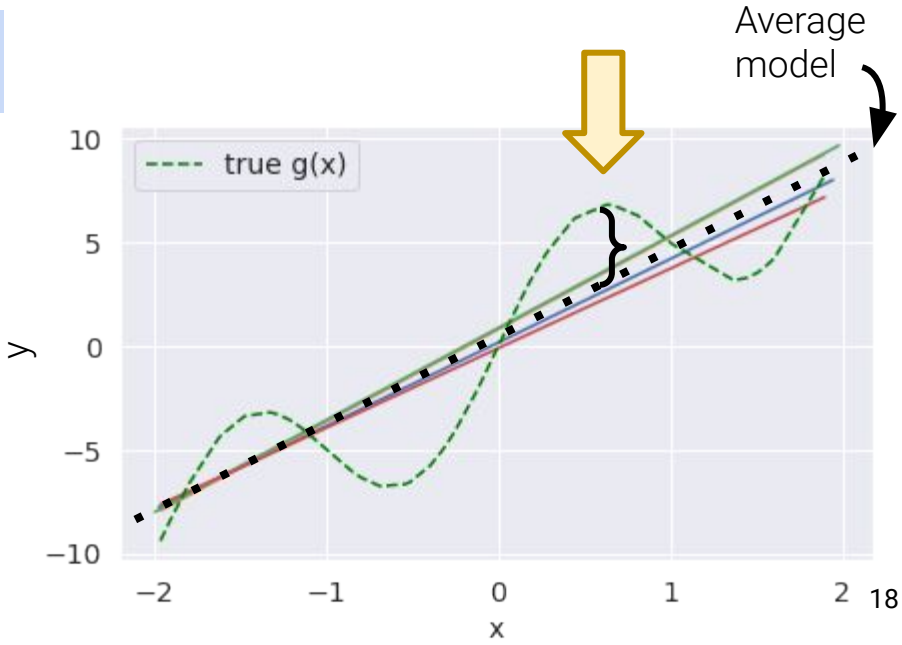
$\hat{Y}(x)$ Prediction for individual x
A random variable

#1352355

Define the **model bias** as the average 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.

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



3. Model Bias

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

#1352355

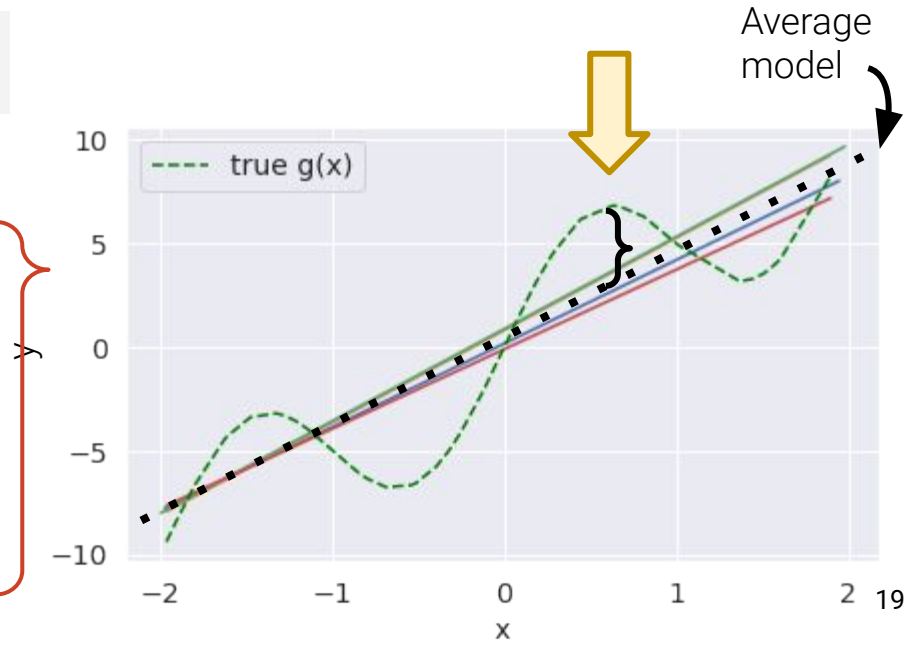
Define the **model bias** as the difference between our predicted value and the true $g(x)$.

- 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.

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

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

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

#1352355

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.

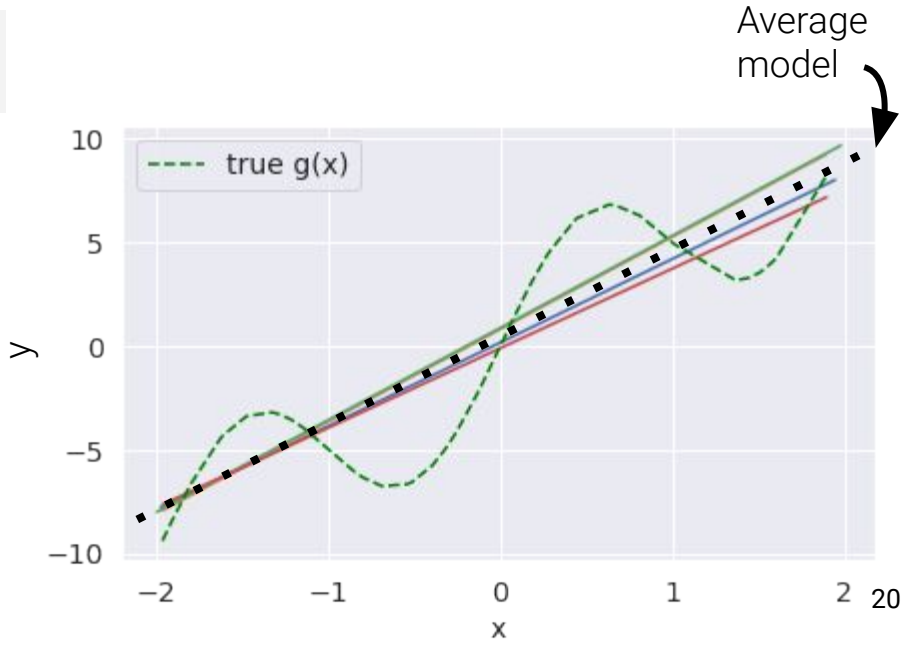
$$\text{model bias} = \mathbb{E}[\hat{Y}(x)] - g(x)$$

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

#1352355

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

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 $\text{estimator bias} = \mathbb{E}[\bar{X}_n] - \mu = 0$.

Matching Problem

1. $\mathbb{E}[(Y - \hat{Y}(x))^2]$
2. σ^2
3. $\mathbb{E}[\hat{Y}(x)] - g(x)$
4. $\mathrm{Var}(\hat{Y}(x))$

A. Observational variance

B. Model variance

C. Model bias

D. Model risk

E. Empirical risk

slido

Matching Question (Order: 1, 2, 3, 4)

ⓘ Start presenting to display the poll results on this slide.

The Bias-Variance Decomposition: Prediction

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

#1352355

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

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

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

Interested in the derivation?
Check out the extra slides!

The Bias-Variance Decomposition: Prediction

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

#1352355

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

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

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

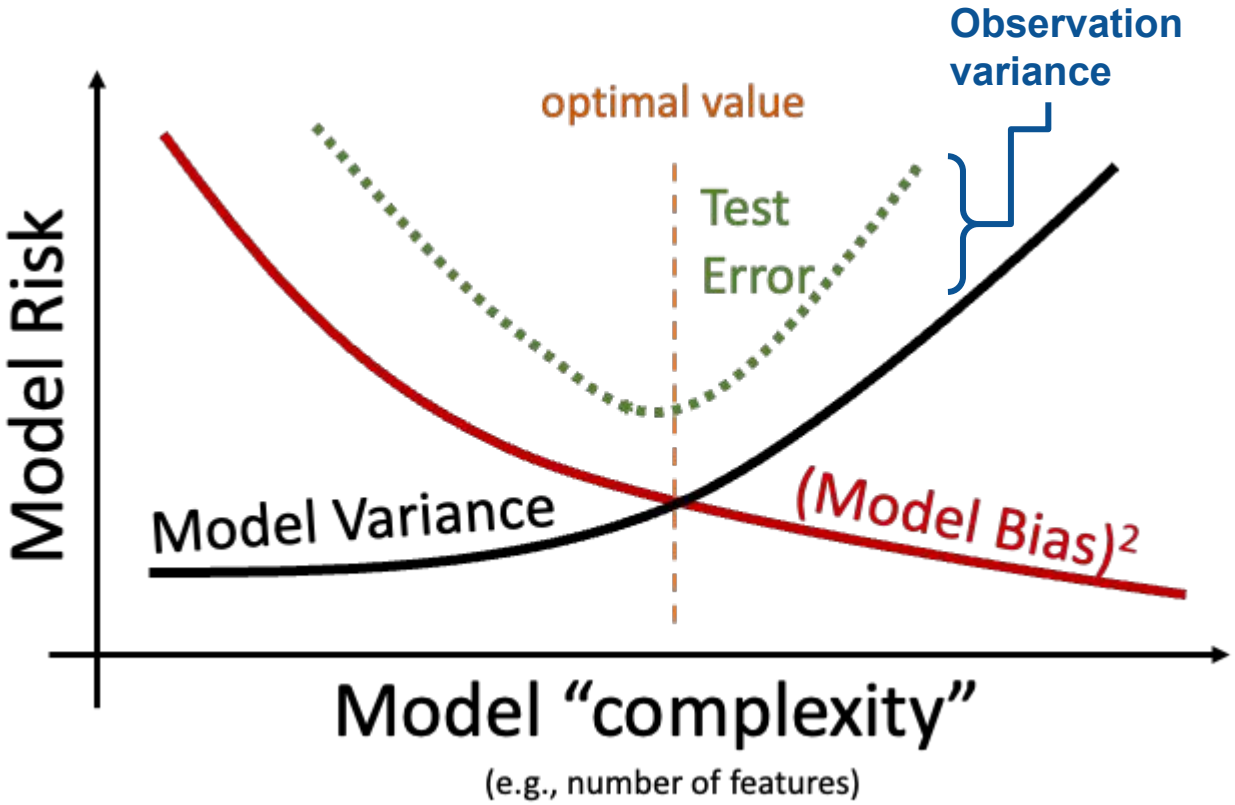
Notes:

- 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 often **overfit** the sample data and will have **higher model variance**. This often corresponds to a decrease in bias.
- As models **decrease in complexity**, they often **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!

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



Interlude

Instructor Office Hours

Questions about courses? research? data science? Come on by!

Gradebook Report Soon

We're working on an overall grade calculator. More information soon.

Lab 11, HW07

Due **after** spring break, when we return.

Interlude

Real World Data Scientist:
Jennifer Chayes (website)

Associate Provost of Associate Provost of the Division of Computing, Data Science, and Society

Dean of the School of Information

Research areas: Graph algorithms, network models, social choice/recommendation systems, phase transitions in CS, responsible AI, etc.

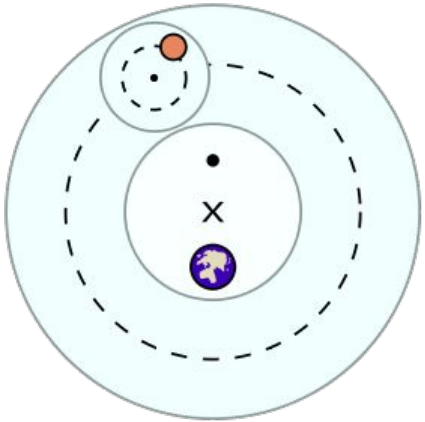
#1352355

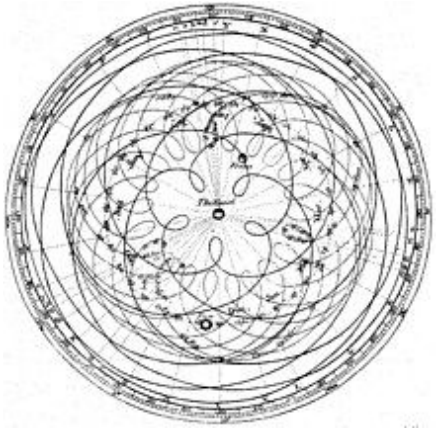
Interpreting Regression Coefficients

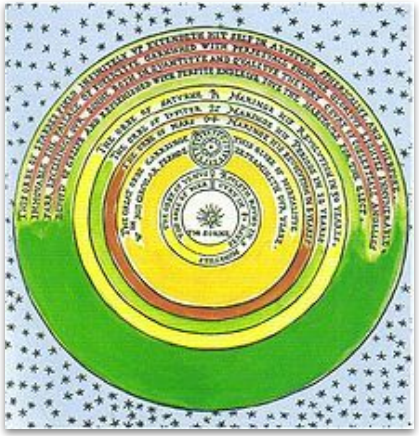
Lecture 20, Data 100 Spring 2023

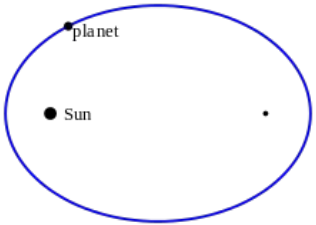
Inference: The right model structure matters!

#1352355









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

#1352355

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. Can we figure out whether it is positive or negative?
3. 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 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

#1352355

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%)

Ruling out 0 almost always means determining the sign of θ_1

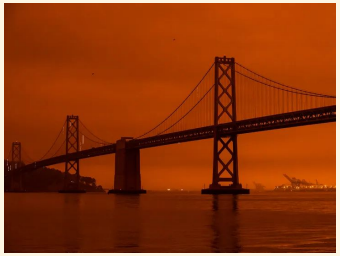
Equivalently (duality argument):

- 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).

#1352355

Bootstrapping Test for a Regression Coefficient

Lecture 20, Data 100 Spring 2023



SF Bay Bridge, 09/2020

Demo

Data 100 textbook (Ch12, Ch 17)

How accurate are air quality measurements?

#1352355

Two common sources of air quality information:
Air Quality System (AQS):

- (+) High-quality, well-calibrated, publicly available, government-run. Gold standard for accuracy
- (−) Expensive (~$15k-40k) and far apart.
- (−) Hourly/delayed reports because of extensive calibration

PurpleAir sensors (link)

- (+) Cheap (~$250), can be installed at home for personal use
- (+) Measurements every two minutes, denser coverage
- (−) Less accurate than AQS (see Josh Hug's post)

How do we use nearby AQS sensor measurements to improve PurpleAir measurements?
Focus on PM2.5 particles (particles < 2.5μm)

Calibration Model

#1352355

Goal: Create a model that predicts PM2.5 readings as accurately as possible.

- Build a model that adjusts PurpleAir (PA) measurements based on nearby **AQS measurements** (AQS, true air quality).

$$PA \approx \theta_0 + \theta_1 AQS$$

- Then, invert model to predict **true air quality** from PA measurements.

$$\text{True Air Quality} \approx -\frac{\theta_0}{\theta_1} + \frac{1}{\theta_1} PA$$

Demo

Data 100 textbook (Ch12, Ch 17)

Side note: Why perform this "inverse regression"?

- Intuitively, AQS measurements are "true" and have no error.
- **A linear model takes a "true" x value input and minimizes the error in the y direction.**
- Algebraically identical, but **statistically different**.

Calibration Model

#1352355

Focus on original linear model (instead of algebraic step 2):

1. Build a model that adjusts PurpleAir (PA) measurements based on nearby **AQS measurements** (AQS, true air quality).

$$PA \approx \theta_0 + \theta_1 AQS$$

2. Karoline Barkjohn, Brett Gannt, and Andrea Clements from the US Environmental Protection Agency developed a model to improve the PuprleAir measurements from the AQS sensor measurements by incorporating Relative Humidity:

$$\mathrm{PA} \approx \theta_0 + \theta_1 \mathrm{AQS} + \theta_2 \mathrm{RH}$$

Barkjohn and group's work is now used in the official US government maps, like the AirNow Fire and Smoke map, includes both AQS and PurpleAir sensors, and applies Barkjohn's correction to the PurpleAir data.

Demo

#1352355

Collinearity

Lecture 20, Data 100 Spring 2023

The Snowy Plover

#1352355

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

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

$$f_\theta(x) = \theta_0 + \theta_1 \text{egg\_weight} + \theta_2 \text{egg\_length} + \theta_3 \text{egg\_breadth} + \epsilon$$

Demo

Estimating the Snowy Plover

#1352355

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

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

#1352355

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

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

Demo

Bootstrapped Confidence Interval for θ_1

#1352355

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)
```

Bootstrapped Confidence Interval for θ_1

#1352355

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

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.

#1352355

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

⚠️ **Wait….something's off here!** ⚠️

Collinearity

#1352355

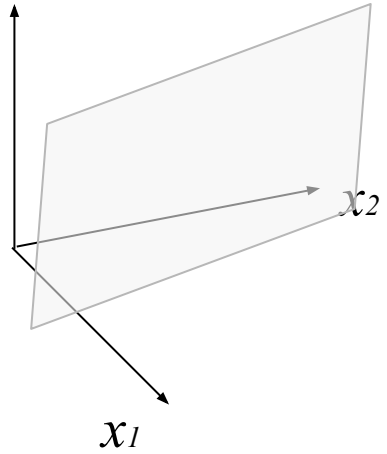
Our estimation from our sample (design matrix $\mathbb{X}$, response vector $\mathbb{Y}$):

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

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

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

If two cats 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

Collinearity

#1352355

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**.

- **Example:** we can't change only one column of a one-hot encoding
- Then the individual slopes are more difficult to interpret

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.

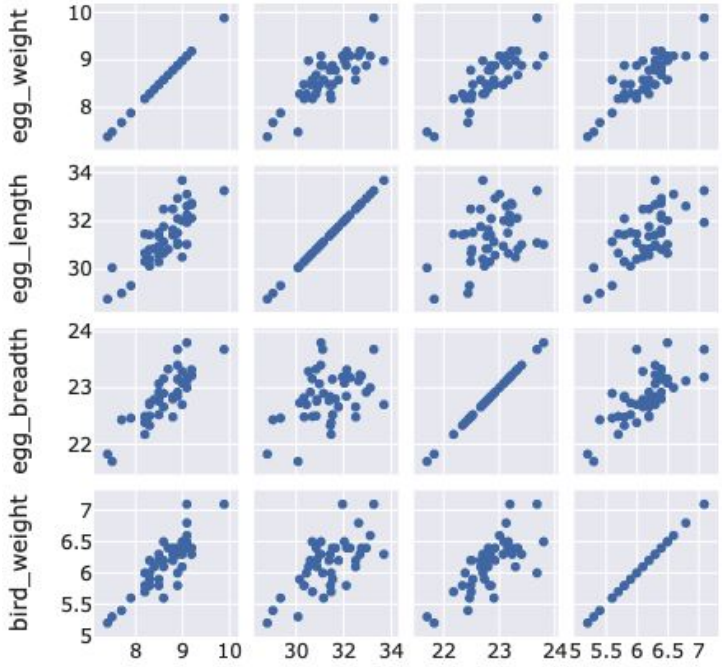
- Slopes are hard to interpret
- $\mathbb{X}^T\mathbb{X}$ might not be invertible, i.e., solution might not be uniquely determined
- Small changes in the data sample can lead to big changes in the estimated slopes
- Also known as **multicollinearity**

Why? Suppose p = 3, and $X_1 \approx X_2 + X_3$. What if we **increase** θ_1 by 10 and also **decrease** θ_2, θ_3 by 10?

- Predictions hardly change at all (but coefficients changed a lot!)
- Means there are very dissimilar models that are nearly indistinguishable from the data

Cross-wise comparison of egg features

#1352355




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

	egg_weight	egg_length	egg_breadth	bird_weight
egg_weight	1.000000	0.792449	0.839077	0.847228
egg_length	0.792449	1.000000	0.402764	0.676142
egg_breadth	0.839077	0.402764	1.000000	0.733687
bird_weight	0.847228	0.676142	0.733687	1.000000

```
eggs.corr()
```

Demo

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

#1352355

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, i.e.
  - Sample size n is large
  - Sample is representative of population distribution (drawn IID, unbiased)

  …then the results of the bootstrap might not be valid.

#1352355

Correlation vs. Causation

Lecture 20, Data 100 Spring 2023

Correlation and Causation

#1352355

What does θ_j mean in our regression?

- Holding other variables fixed, how much should our **prediction** change with X_j?
- For simple linear regression, this boils down to the **correlation coefficient**:
  - Does having more x predict more y (and by how much)?

Which of these questions can be answered using the data alone?

- Is college GPA higher for students who win a certain scholarship?
- Does getting the scholarship **improve** students' GPAs?

slido

Which of these questions can be answered using the data alone?

ⓘ Start presenting to display the poll results on this slide.

Prediction vs Causation

#1352355

Questions about **correlation / prediction**:

- Are homes with granite countertops worth more money?
- Is college GPA higher for students who win a certain scholarship?
- Are breastfed babies less likely to develop asthma?
- Do cancer patients given some aggressive treatment have a higher 5-year survival rate?
- Are people who smoke more likely to get cancer?

These sound like **causal questions**, but **they are not**!

Prediction vs Causation

#1352355

Questions about **correlation / prediction**:

- Are homes with granite countertops worth more money?
- Is college GPA higher for students who win a certain scholarship?
- Are breastfed babies less likely to develop asthma?
- Do cancer patients given some aggressive treatment have a higher 5-year survival rate?
- Are people who smoke more likely to get cancer?

Questions about **causality**:

- How much do granite countertops **raise** the value of a house?
- Does getting the scholarship **improve** students' GPAs?
- Does breastfeeding **protect** babies against asthma?
- Does the treatment **improve** cancer survival?
- Does smoking **cause** cancer?

Causal questions are about the **effects** of **interventions**, not just passive observation.

Note: Regression coefficients θ are sometimes called "effects." This can be deceptive!

Prediction vs Causation

#1352355

Only one of these questions can be answered using the data alone:

✅ **Predictive question:** Are breastfed babies healthier?

❌ **Causal question:** Does breastfeeding improve babies' health?

Possible explanations for **why** breastfed babies would be healthier, on average:

1. **Causal effect:** breastfeeding makes babies healthier
2. **Reverse causality**: healthier babies more likely to successfully breastfeed
3. **Common cause**: healthier / richer parents have healthier babies **and** are more likely to breastfeed

We cannot tell which explanations are true (or to what extent) just by observing (x,y) pairs!

Causal questions implicitly involve **counterfactuals** (an event that did not happen):

- **Would** the **same** breastfed babies have been less healthy **if** they **hadn't** been breastfed?
- Explanation 1 implies they would be, explanations 2 and 3 do not

Confounders

Let T: Treatment, e.g., alcohol use.

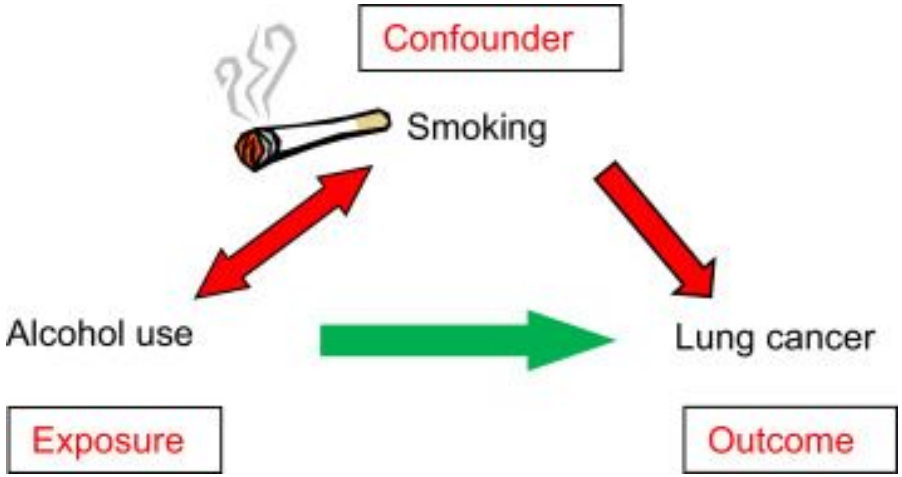
Y: Outcome, e.g., lung cancer.

Suppose we observe that people who drink are more likely to have lung cancer.

A **confounder** is a variable that affects both T and Y, distorting the correlation between them.

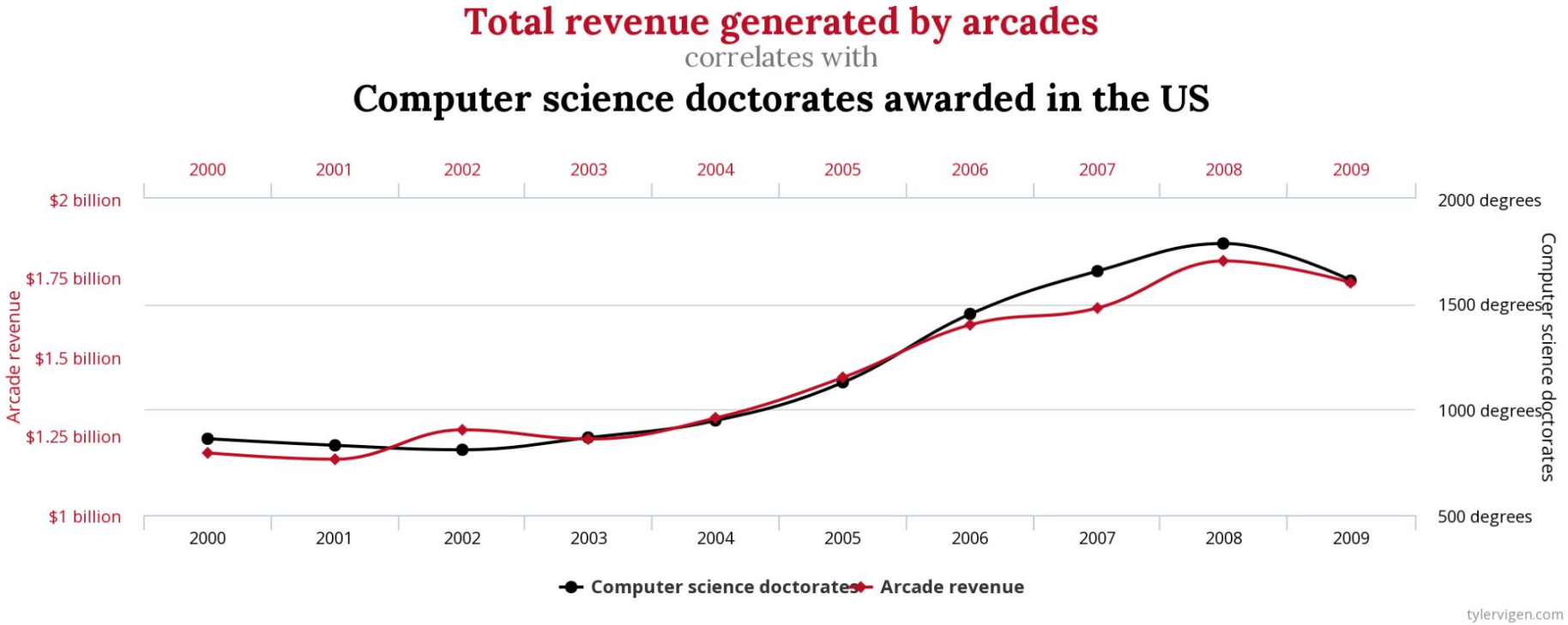
- (e.g. rich parents → breastfeeding, baby's health)
- Can be a measured covariate (feature), or unmeasured variable we don't know about!

Confounders generally cause problems.

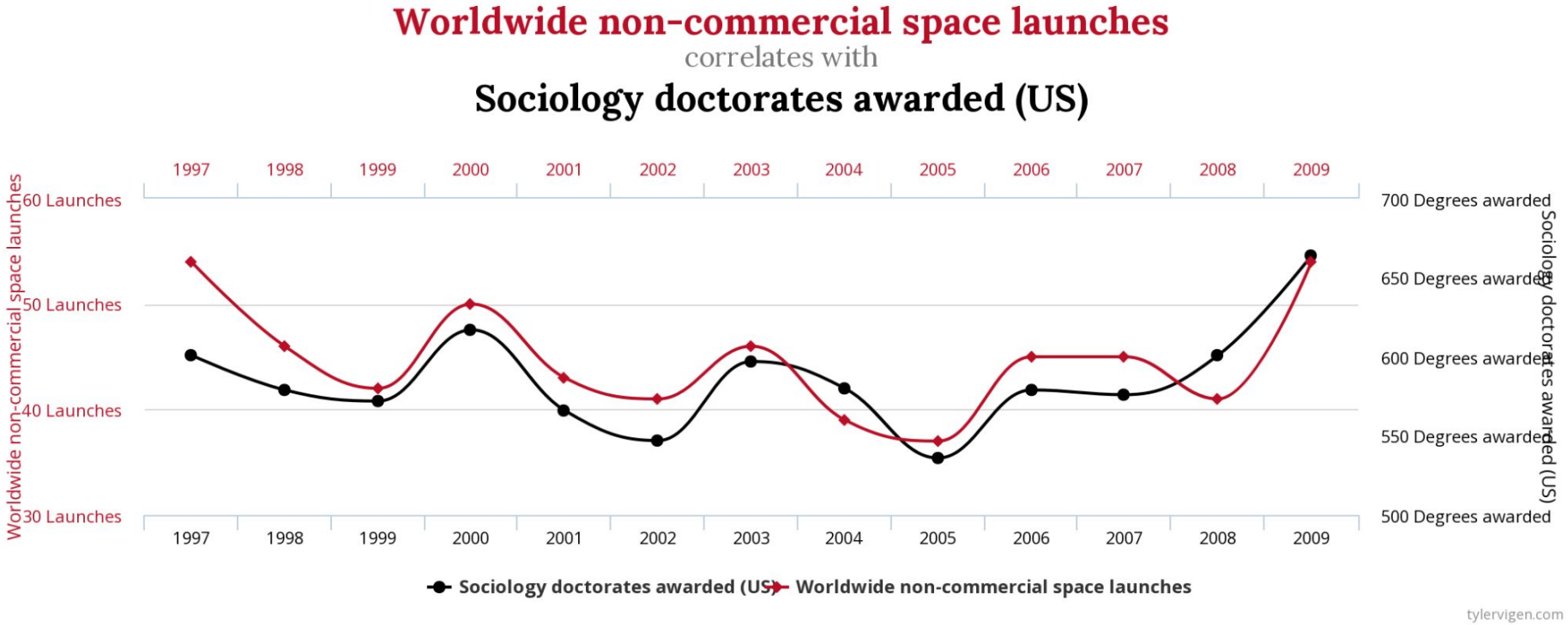


Common assumption: all confounders are observed (**ignorability**).

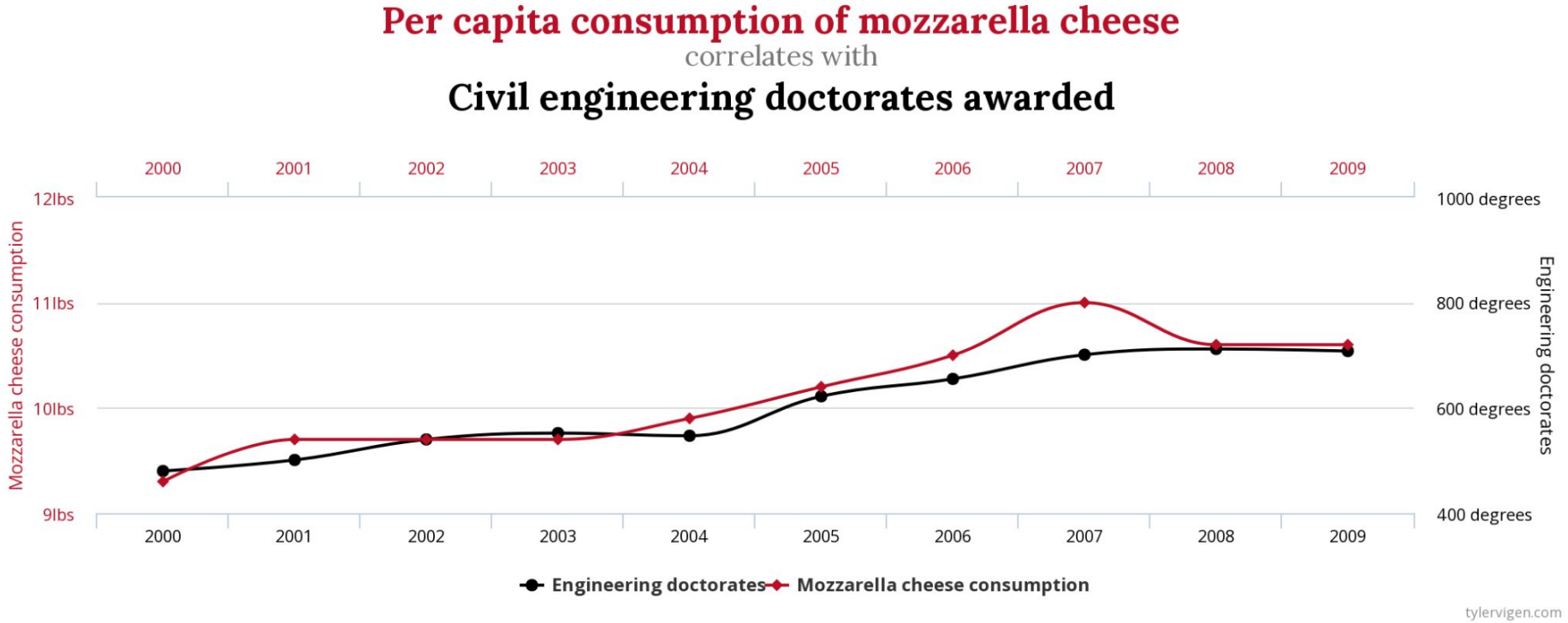
Spurious Correlations



Spurious Correlations



Spurious Correlations



How to perform causal inference?

#1352355

Ignorability: All confounders are observed, i.e., covariates (data features) contain confounders.

In a **randomized experiment**:

- Randomly assign participants into two groups (the **treatment** and the **control** group) and then apply the treatment to the treatment group only.
- We assume ignorability and gather as many measurements as possible.
- Often not practical: randomly assigning treatments to participants is impractical or unethical!

In an **observational study**:

- Obtain two participant groups separated based on some identified treatment variable.
- Cannot assume ignorability: the participants could have separated into the two groups based on other covariates!
- There could also be unmeasured confounders!

There is an entire field of statistics called **causal inference** which studies causal models (i.e., treatment/covariate/response variables) in the context of observational studies.

Take STAT 156! (Fall 2022 catalog)

#1352355

Have a lovely Spring Break!

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 20, Data 100 Spring 2023

#1352355

Bootstrap Resampling

#1352355

- 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.

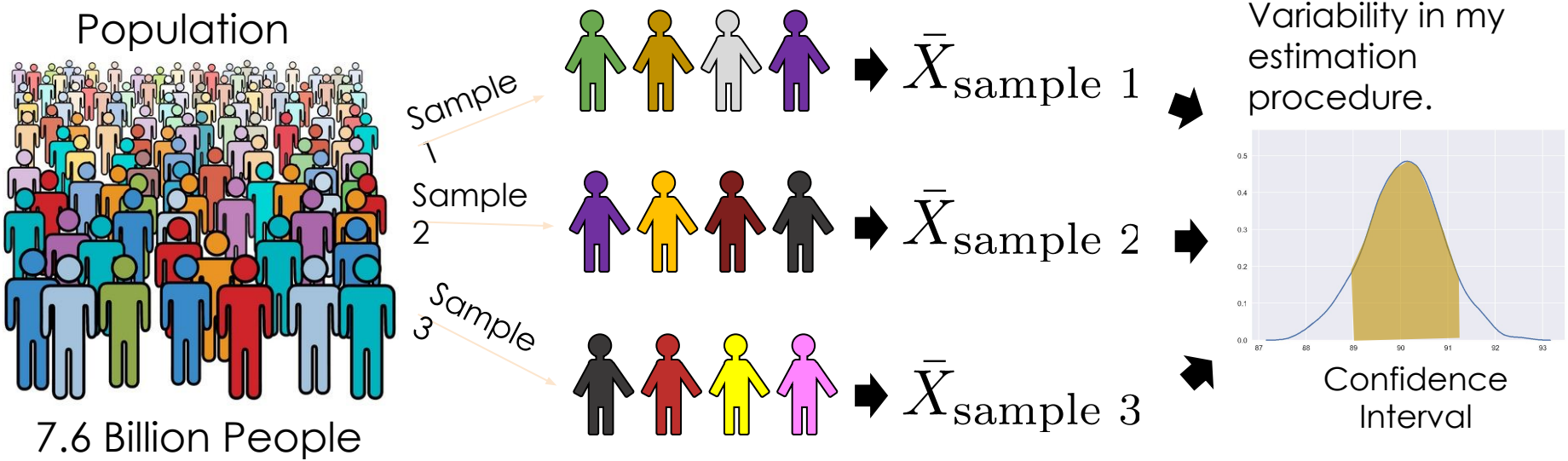
Idea: Treat our random sample as a "population", and resample from it.

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

The Distribution of an Estimator

#1352355

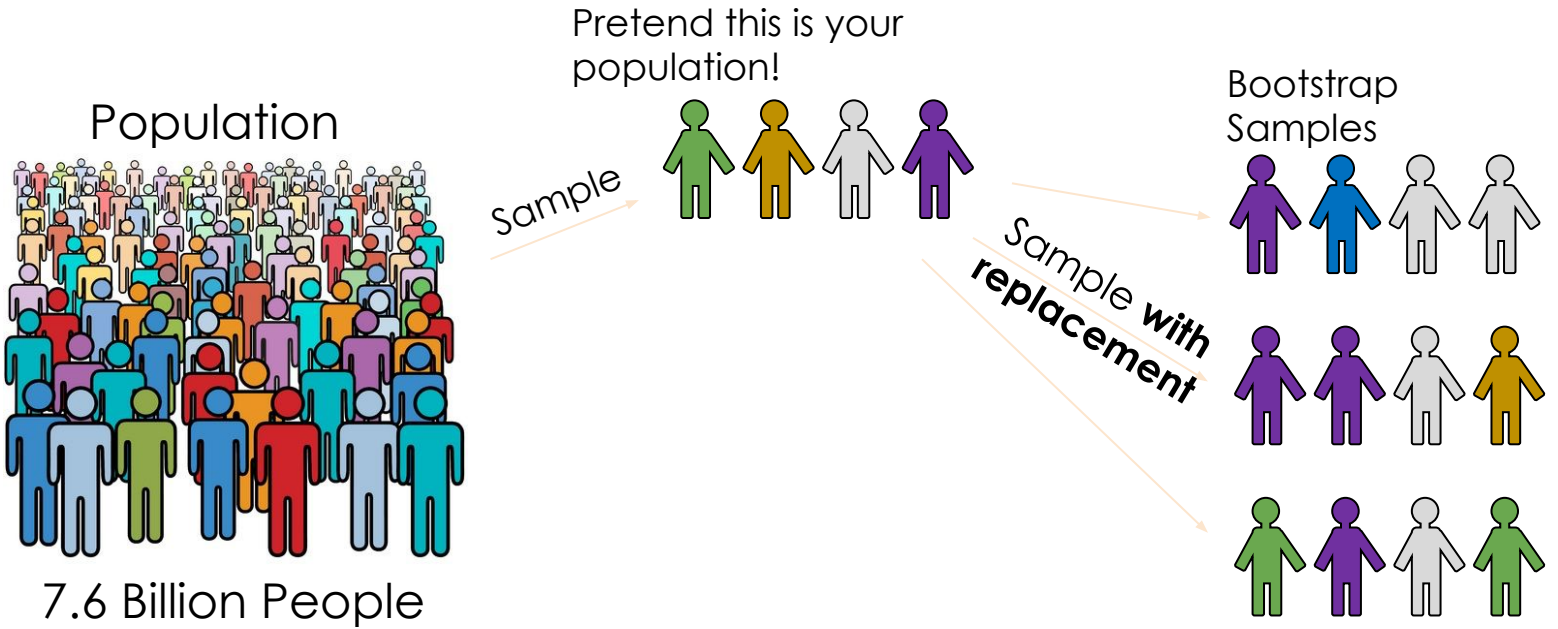
Resampling the population to estimate the sample distribution.



Bootstrap the Distribution of an Estimator

Simulation method to estimate the sample distribution.

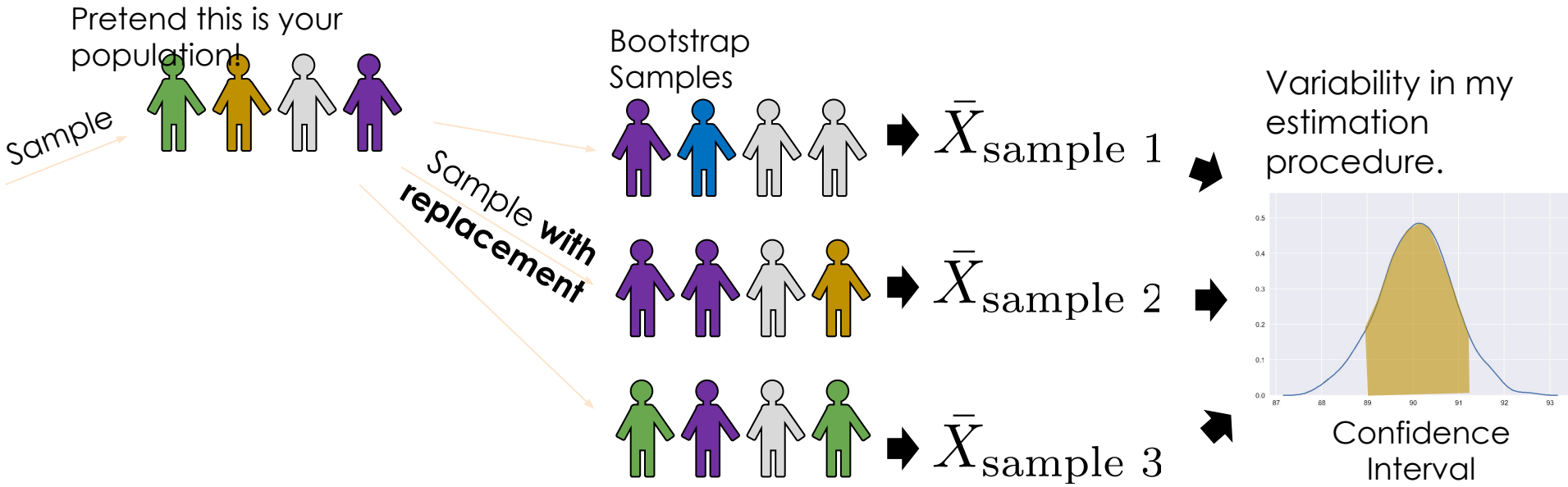
#1352355



Bootstrap the Distribution of an Estimator

#1352355

Simulation method to estimate the sample distribution.

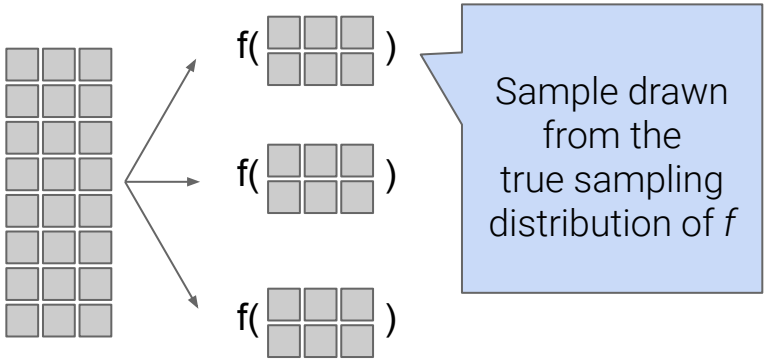


Bootstrap Resampling

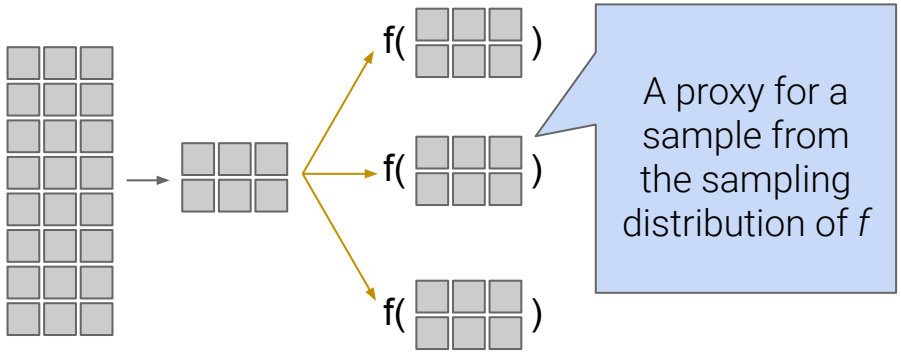
#1352355

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**?

Bootstrap Discussion

#1352355

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).

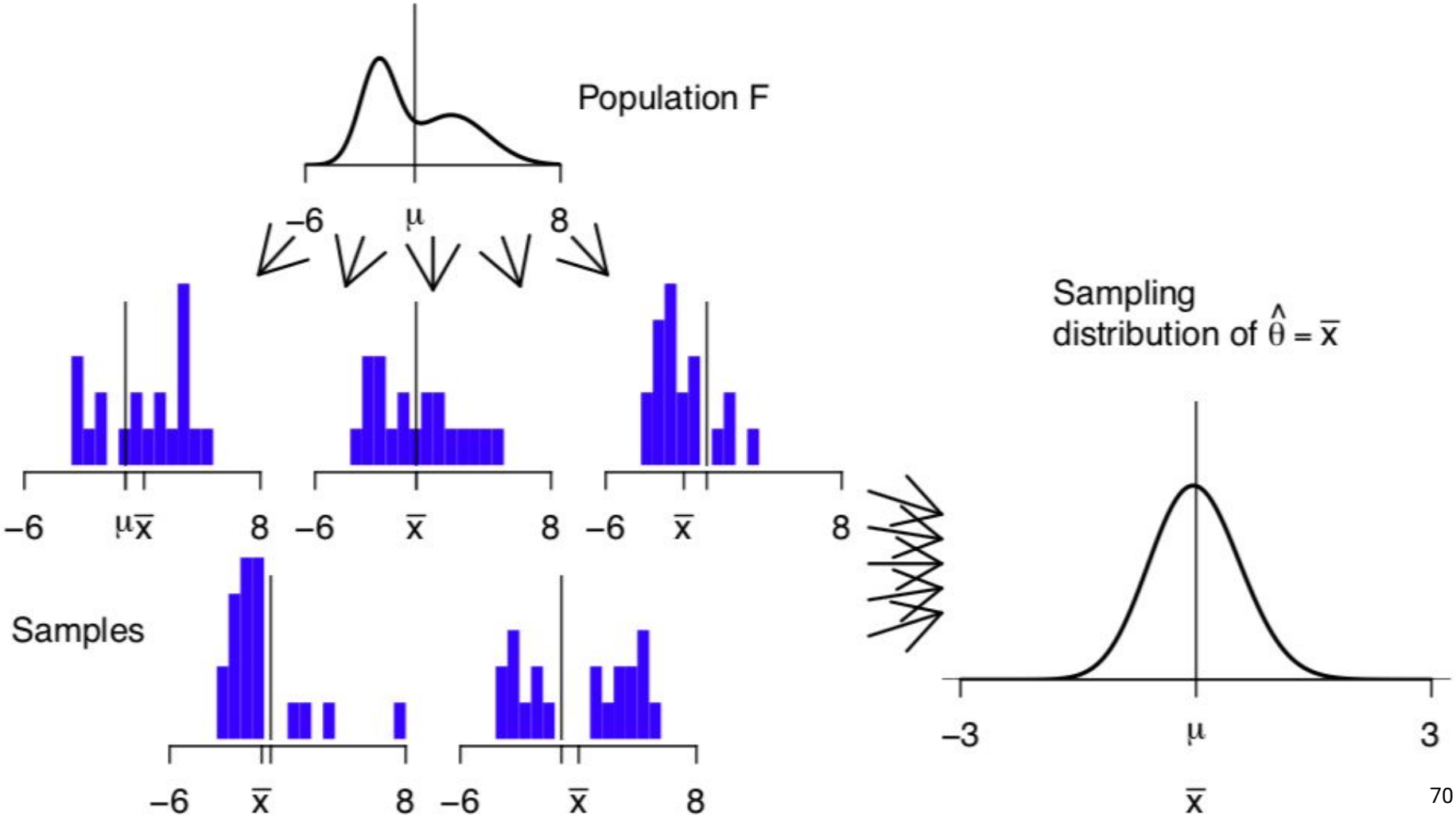
The center of the bootstrapped distribution is the estimator applied to our original sample.

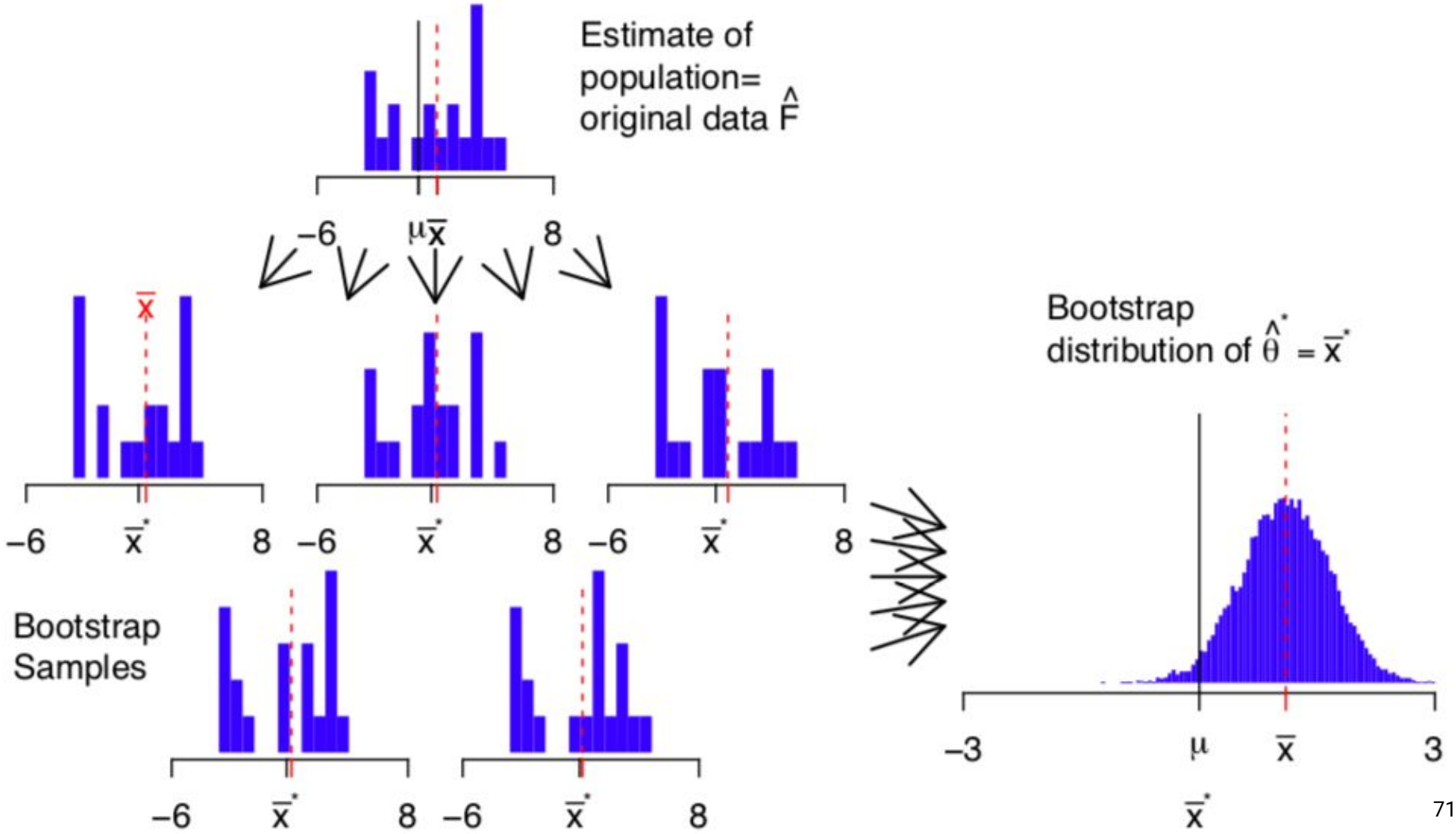
- We have no way of recovering the estimator's true expected value.

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.


Population F
−6
μ
8
Samples
−6
μx̄
8
−6
x̄
8
−6
x̄
8
−6
x̄
8
Sampling
distribution of θ̂ = x̄
−3
μ
3
x̄


Estimate of
population=
original data $\hat{F}$
−6
$\mu_{\bar{x}}$
8
$\bar{x}$
−6
$\bar{x}^*$
8
Bootstrap
Samples
Bootstrap
distribution of $\hat{\theta}^* = \bar{x}^*$
−3
μ
$\bar{x}$
3
$\bar{x}^*$

What Teachers Should Know About the Bootstrap

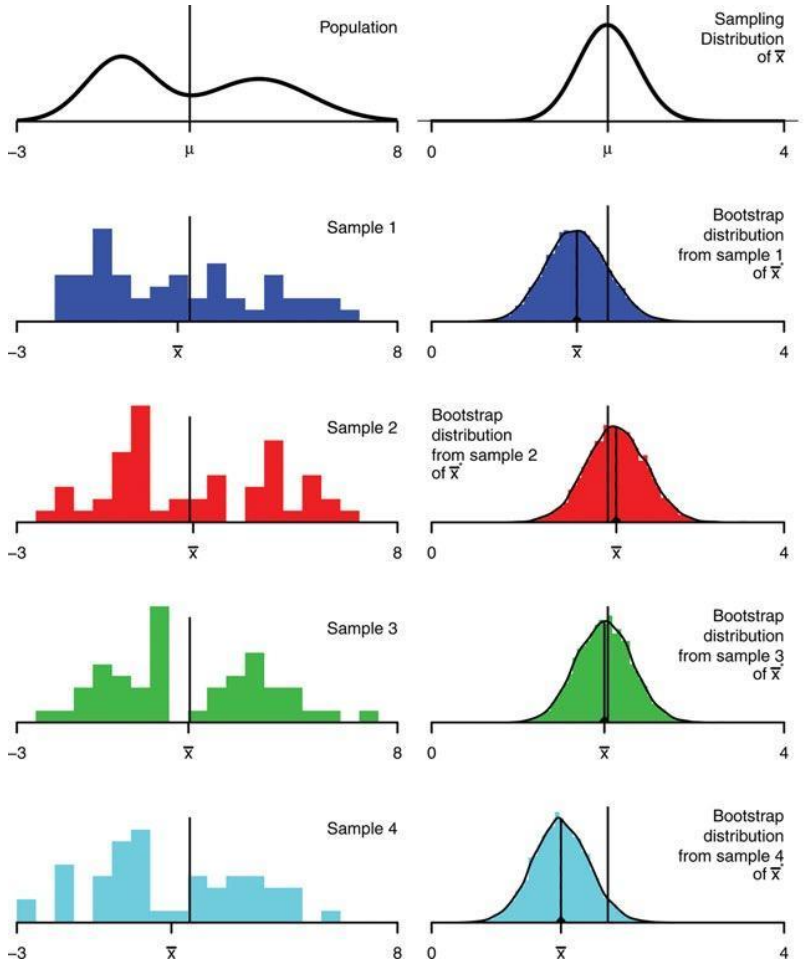
#1352355

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)

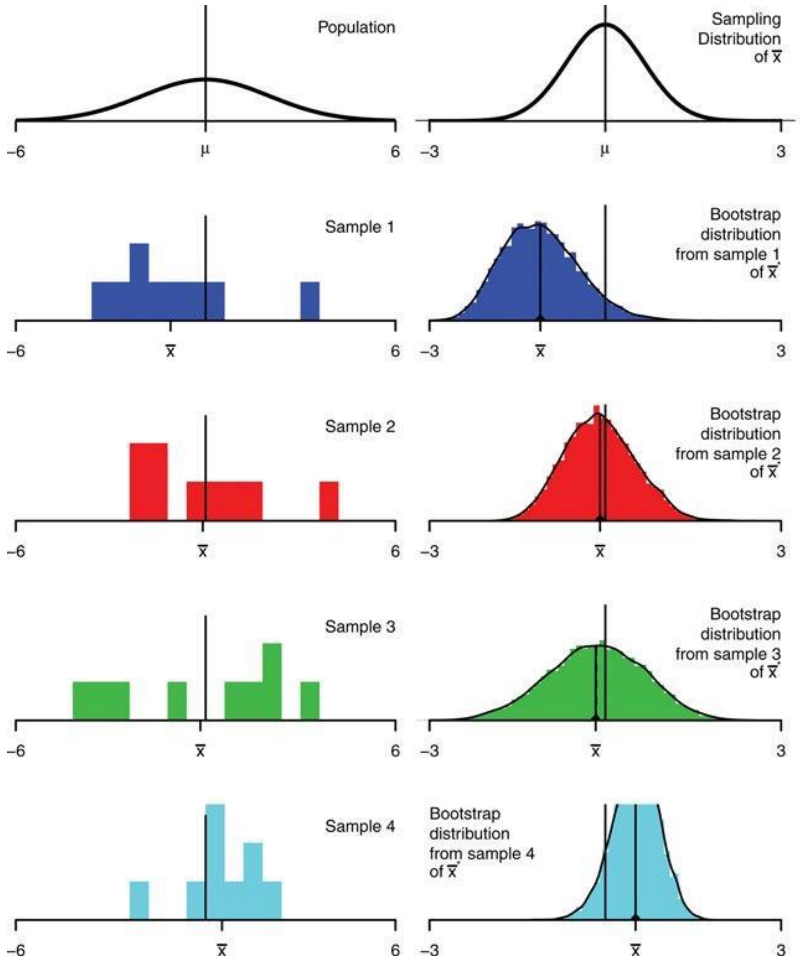
Bootstrap for the Mean, n=50



From Tim C. Hesterberg (2015)

#1352355

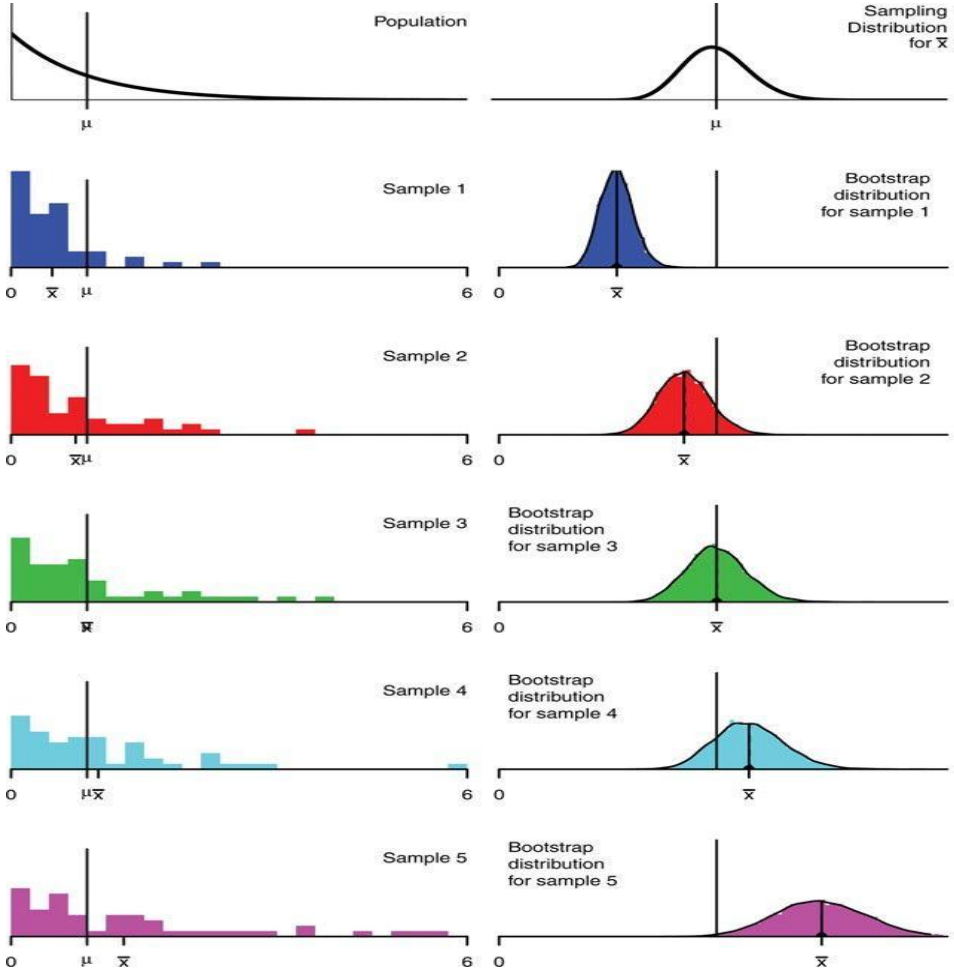
Bootstrap Distributions for the Mean, n = 9



#1352355

Bootstrap Distributions for the Mean, n = 50, Exponential Population.

#1352355



Some More Lessons from Hesterberg

#1352355

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 20, Data 100 Spring 2023

#1352355

Derivation: Preliminary

#1352355

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

Preliminary

Before proceeding with this derivation, you should be familiar with the Random Variables lecture (Lecture 16 in Spring 2023). 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.

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

#1352355

Step 1

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

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

#1352355

Step 2

At this stage we have

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

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

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

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

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

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

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

Step 4: Bias-Variance Decomposition

In Step 2 we had

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

Step 3 showed

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

Thus we have shown the bias-variance decomposition

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

That is,

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

Derivation: Special Case for Parameterized Models

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

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

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

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

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

LECTURE 20

Bias, Variance, and Inference

Content credit: Acknowledgments