slido

8422703

Join at slido.com #8422703

Click **Present with Slido** or install our Chrome extension to display joining instructions for participants while presenting.

LECTURE 18

Estimators, Bias, and Variance

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

Data 100/Data 200, Spring 2025 @ UC Berkeley

Narges Norouzi and Josh Grossman

Content credit: Acknowledgments

8422703

Announcements

8422703

You don't have to come to instructor office hours with an agenda. We're happy to talk about anything on your mind, and **we are excited to talk to you**!

Reminder about coffee chats with Josh. Slots are are beginning to fill up as we near the end of the term.

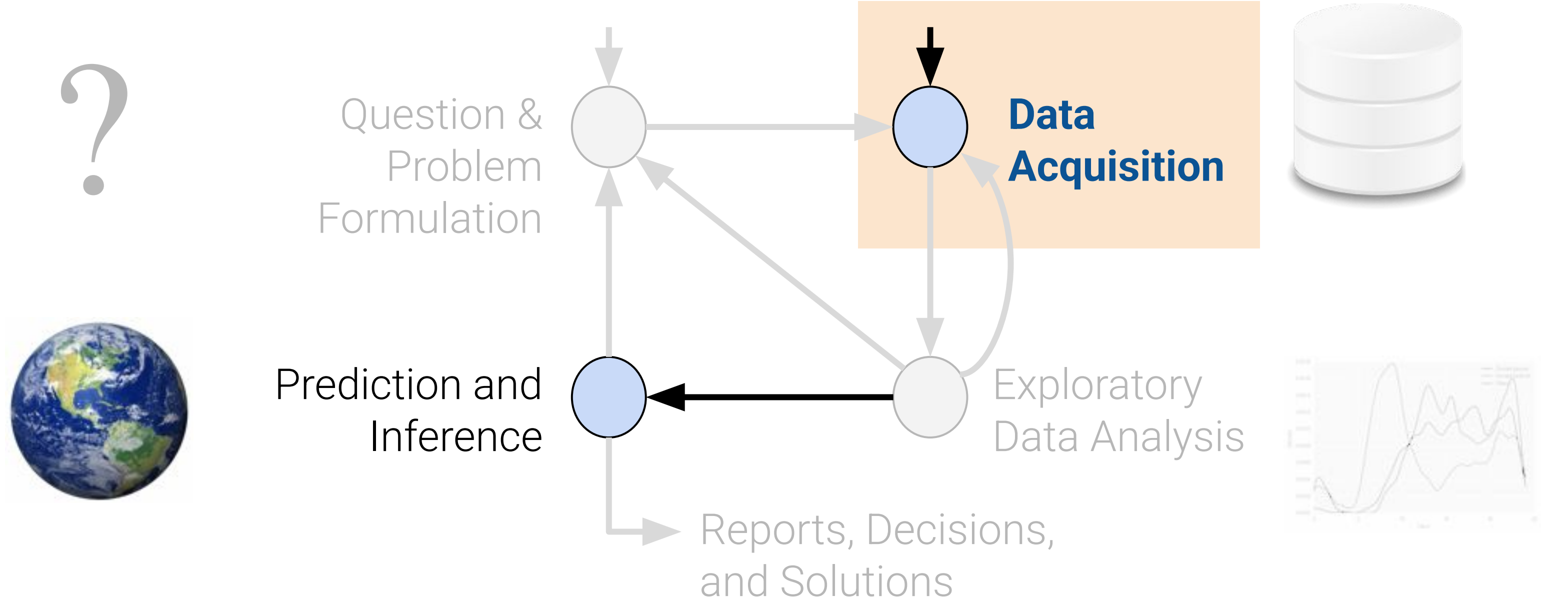
Folks joining from home: I'm switching from the laser pointer to the tablet pointer, and in the future we can separately post videos of physical demonstrations like OLS.

Probability is challenging to learn! Not a prereq of Data 100. If it feels tough, that's expected 🙂

Would you prefer that Josh's office hours:

Why Probability?

8422703



(today)

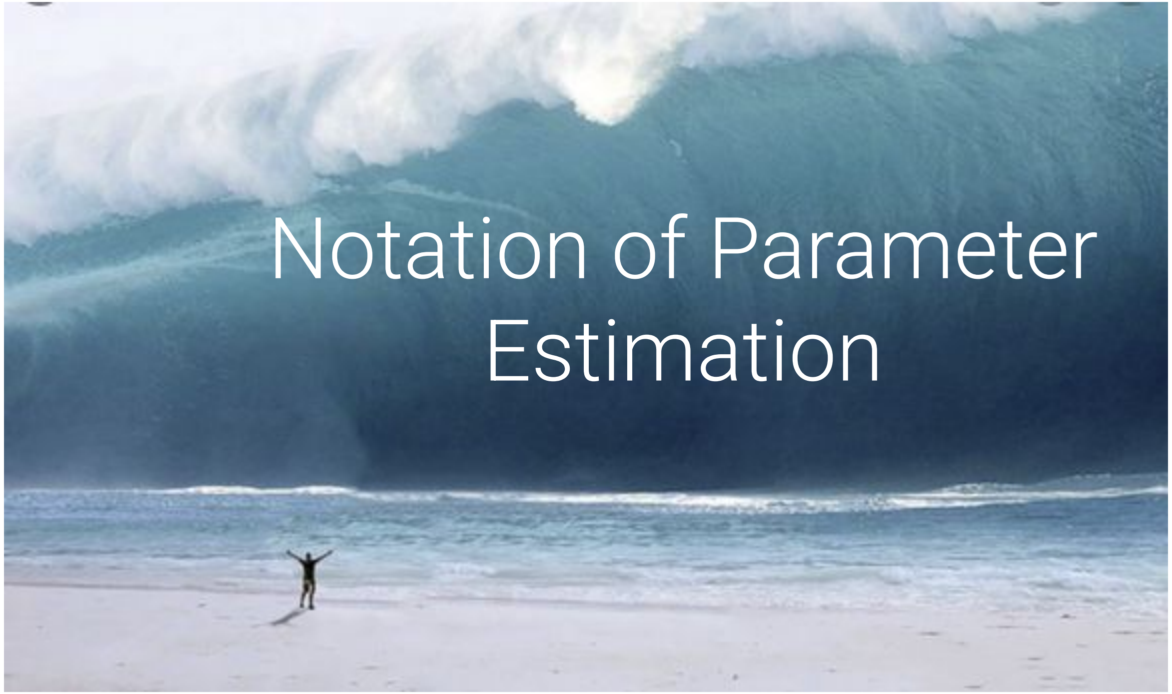
Model Selection Basics:
Cross Validation
Regularization

Probability I:
Random Variables
Estimators

Probability II:
Bias and Variance
Inference/Multicollinearity

Today's lecture is conceptually challenging!

8422703



Be kind to yourself, especially after spring break! 💙

Last time: Coins!

8422703

P(Heads) = 0.5
P(Tails) = 0.5

Let X_i be a **random variable (r.v.)** representing the i^{th} outcome of a series of coin flips.

If heads, $X_i = 1$. If tails, $X_i = 0$

$P(X_i=1) = P(X_i=0) = 0.5$

$\mathbf{X_i \sim}$ **Bernoulli(p=0.5)**, where the X_i's are independent and identically distributed (i.i.d.)

If an r.v. X follows a Bernoulli distribution, **P(X=1) = p** and **P(X=0) = 1-p**

New terminology: Data-generating Process (DGP)

8422703

Data-generating process (DGP):

$$X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5)$$

$X_1 = 1$

$X_2 = 0$

$X_3 = 0$

$X_\infty = 1$

. . .

E(X_i) [Expected value]:
What is the long run average of the X_i's ?

$E(X_i) = p = 0.5$

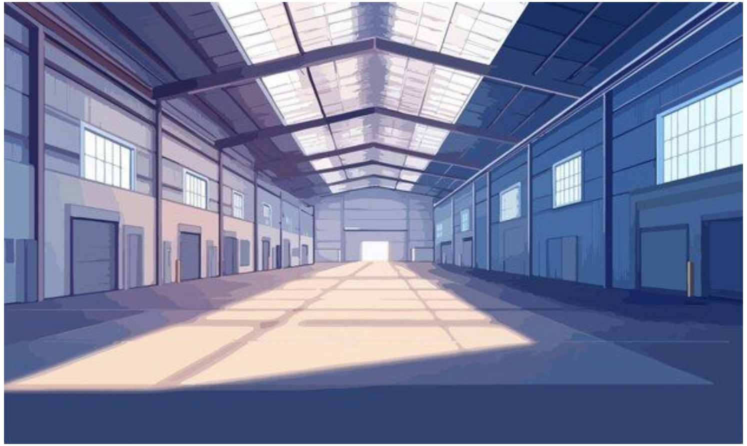
Var(X_i) [Variance]:
How spread out are the X_i's around their average?

$Var(X_i) = p(1-p) = 0.25$

An equivalent way to think about the coin flip DGP

8422703

Randomly sampling with replacement from a warehouse with an **infinite** number of **random** coin flip outcomes (i.e., a **population** of coin flips):



$X_1 = 1$

$X_2 = 0$

$X_3 = 0$

. . .

$X_\infty = 1$

E(X_i) [Expected value]:
What is the average value of the X_i's ?

$E(X_i) = p = 0.5$

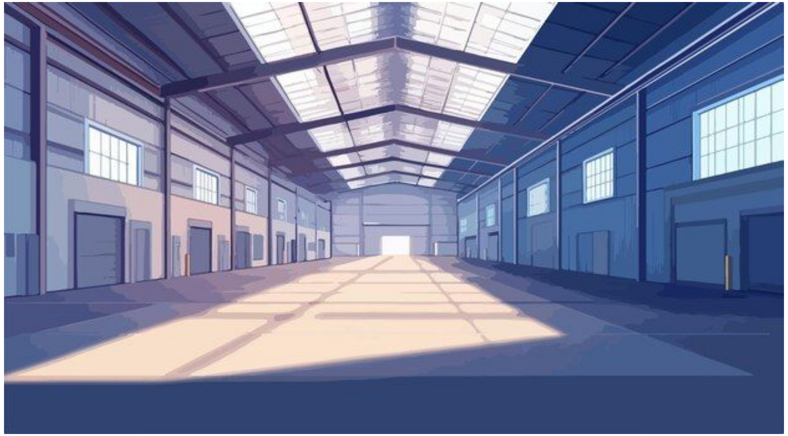
Var(X_i) [Variance]:
How spread out are the X_i's around their average?

$Var(X_i) = p(1-p) = 0.25$

The structure of the population is usually unknown

8422703

Randomly sampling with replacement from a warehouse with all **32,000 heights** of Berkeley undergrads on slips of paper (a **population** of heights)**:**



X_1 = 69 in

X_2 = 71 in

X_3 = 64 in

. . .

X_∞ = 60 in

E(X_i) [Expected value]:
What is the average value of the X_i's ?

We don't know!

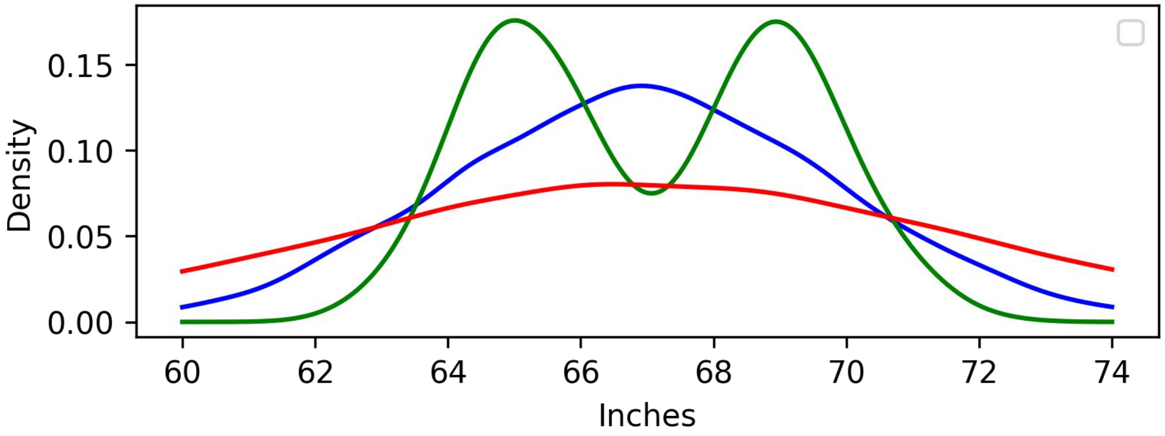
Var(X_i) [Variance]:
How spread out are the X_i's around their average?

We don't know!

Possible population distribution of heights

8422703

Some possible distributions of the 32,000 heights of Berkeley undergrads:



We do not know the true distribution of heights. But, we may want to estimate its properties.

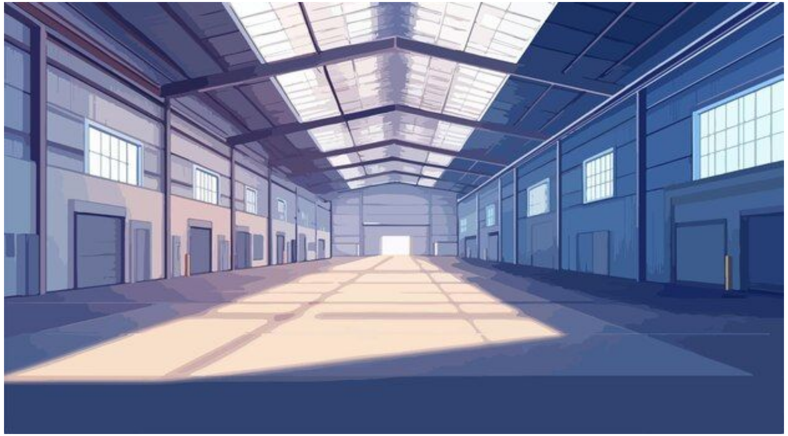
For example, we might want to estimate the **true average height** of Berkeley undergrads.
[Perhaps we are designing doors in a new building.]

A method we know: Randomly sample 100 undergrads and calculate the **sample mean**.

A familiar approach: Estimate the true mean with a sample mean

8422703

Warehouse with all **32,000 heights** of Berkeley undergrads on slips of paper (a **population**):



i.i.d. random sample of 100 heights: $X_1, X_2, \ldots X_{100}$

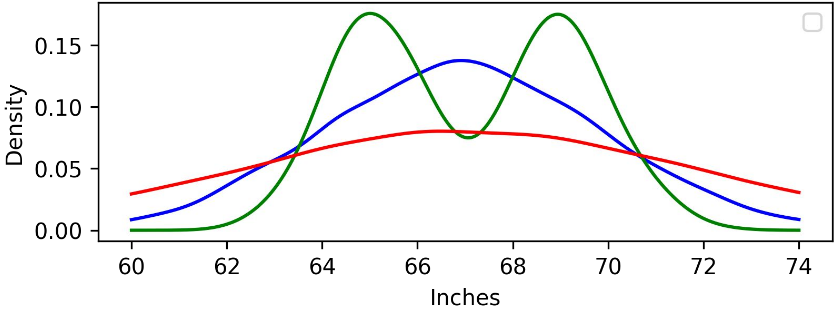
Sample mean is 68.1 inches.

$\bar{X}_{100}$ = 68.1 inches

Our "best guess" for the population mean is 68.1 inches.

Harder Q: **How do we know if 68.1 inches is a "good" estimate?**

Today, we address this question!

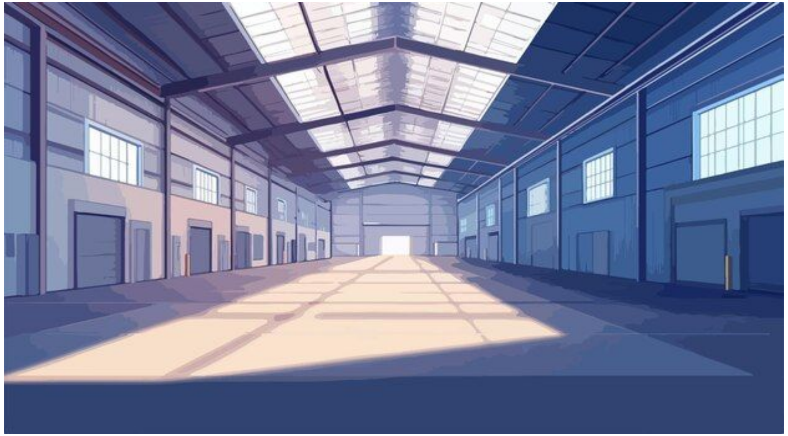


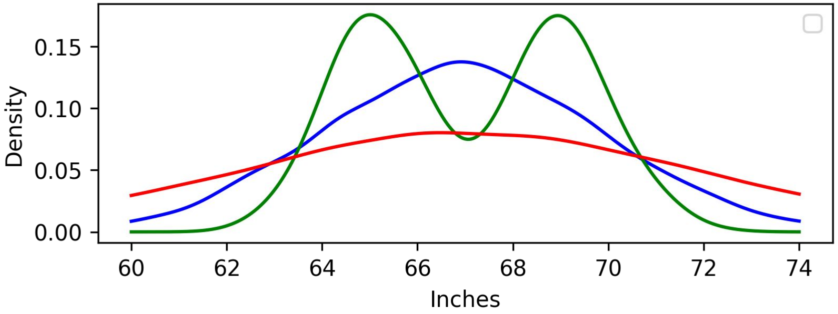
Possible distributions of the raw data

Thinking about a sample we could have observed

8422703

Warehouse with all **32,000 heights** of Berkeley undergrads on slips of paper (a **population**):





Possible distributions of the raw data

Our universe (Observed sample):

i.i.d. random sample of 100 heights:
$X_1, X_2, \ldots X_{100}$

Sample mean is 68.1 inches.

$\bar{X}_{100}$ = 68.1 inches

A parallel universe (An unobserved sample):

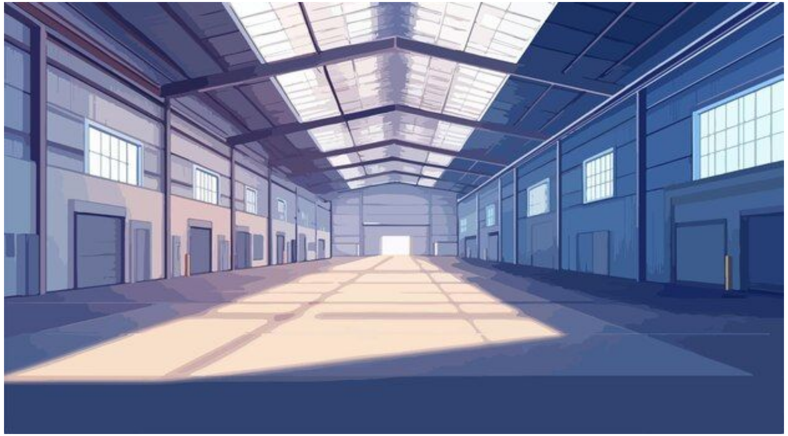
i.i.d. random sample of 100 heights:
$X_1, X_2, \ldots X_{100}$

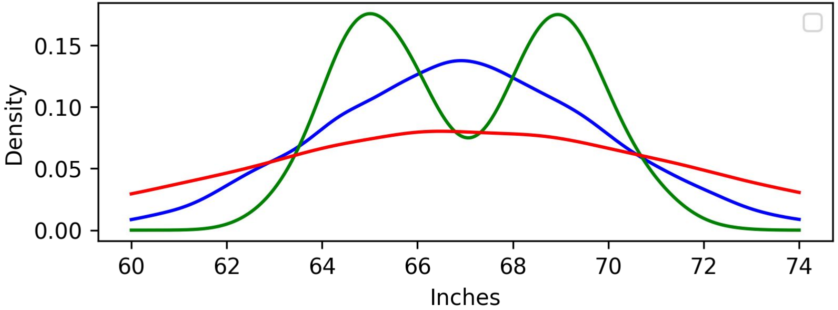
Sample mean is **69.2** inches.

$\bar{X}_{100}$ = **69.2** inches

There are many possible samples we could have observed!

Warehouse with all **32,000 heights** of Berkeley undergrads on slips of paper (a **population**):





Possible distributions of the raw data

There are (effectively) infinite possible samples of size 100 we could have drawn! But, we observe **just one sample**.

This is also a distribution!

$\bar{X}_{100,\text{Sample 1}}$ = 68.1 inches

$\bar{X}_{100,\text{Sample 2}}$ = 69.2 inches

$\bar{X}_{100,\text{Sample 3}}$ = 67.9 inches

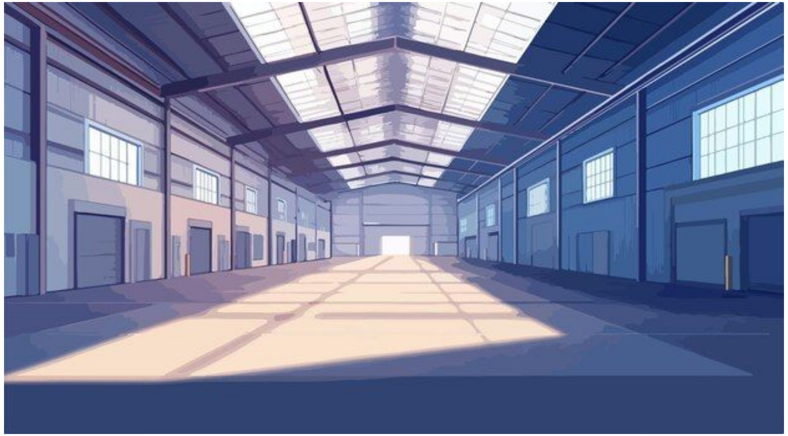
$\cdots$

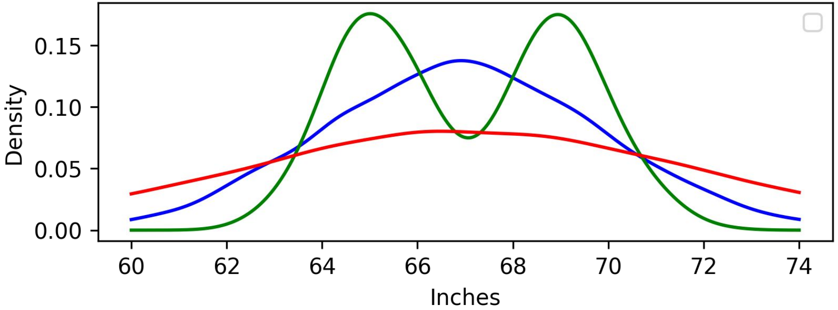
$\bar{X}_{100,\text{Sample }\infty}$ = 68.5 inches

8422703

The CLT is a story of repeated sampling (i.e., parallel universes)

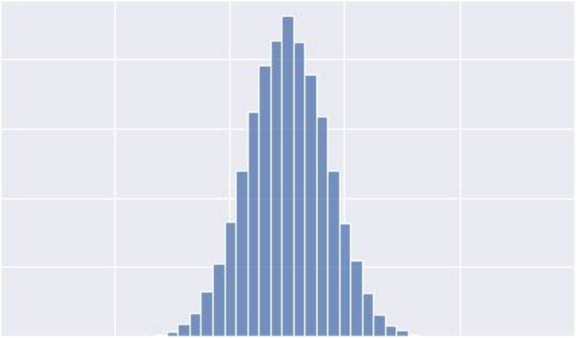
Warehouse with all **32,000 heights** of Berkeley undergrads on slips of paper (a **population**):





Possible distributions of the raw data

Central Limit Theorem (CLT) (Data 8)



$\bar{X}_{100}$ for multiple samples of size 100

For i.i.d. samples of X_i's of size n ($X_1, \ldots, X_n$),
Where n is "big enough",
the distribution of $\bar{X}_n$, the **sample mean** of X_i's ,
is **roughly normal.**

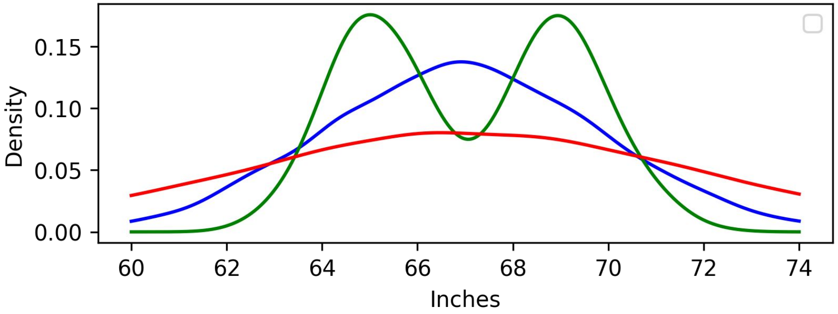
The same CLT story, but represented as a DGP

8422703

Data-generating process (DGP):

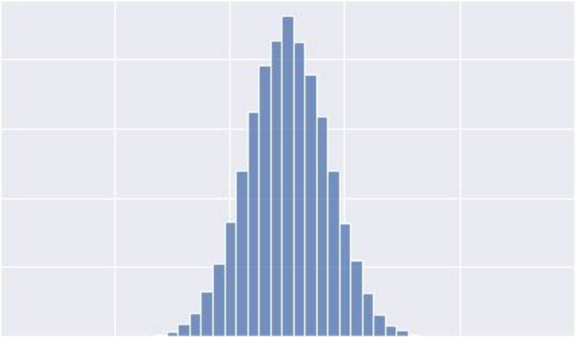
$$X_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(X_i) = \mu \qquad \text{Var}(X_i) = \sigma^2$$



Possible distributions of the raw data

Central Limit Theorem (CLT) (Data 8)



$\bar{X}_{100}$ for multiple samples of size 100

For i.i.d. samples of X_i's of size n ($X_1, \ldots, X_n$),
Where n is "big enough",
and X_i ~ Unknown, where $E(X_i)=\mu$ and $SD(X_i)=\sigma$,
the distribution of $\bar{X}_n$, the **sample mean** of X_i's ,
is **roughly normal.**

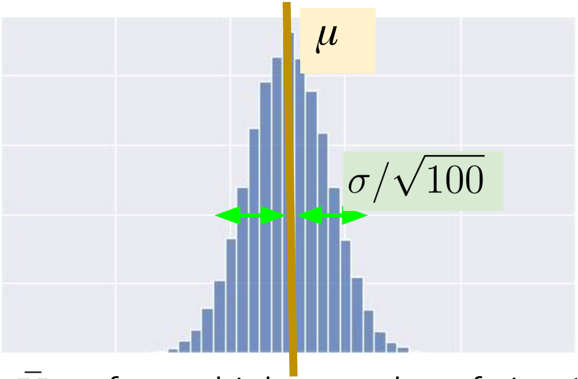
The same CLT story, but represented as a DGP

Data-generating process (DGP):

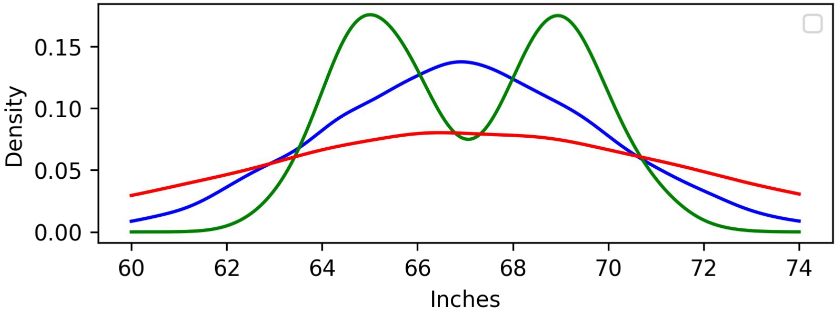
$$X_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(X_i) = \mu \qquad \text{Var}(X_i) = \sigma^2$$

Central Limit Theorem (CLT) (Data 8)



$\bar{X}_{100}$ for multiple samples of size 100



Possible distributions of the raw data

For i.i.d. samples of X_i's of size n ($X_1, \ldots, X_n$),
Where n is "big enough",
and X_i ~ Unknown, where $E(X_i)=\mu$ and $SD(X_i)=\sigma$,
the distribution of $\bar{X}_n$, the **sample mean** of X_i's ,
is **roughly normal with mean μ and SD $\sigma/\sqrt{n}$.**

8422703

Central Limit Theorem (Data 8 + today's terminology)

For an i.i.d. sample of X_i's of size n,

Where n is "big enough",

and X_i ~ Unknown, where $E(X_i)=\mu$ and $SD(X_i)=\sigma$,

the distribution of $\bar{X}_n$, the **sample mean** of X_i's ,

is $\underbrace{\textbf{roughly normal}}_{\text{Proof out of scope}}$ with $\underbrace{\text{mean } \mu \text{ and SD } \sigma/\sqrt{n}}_{\text{(Let's prove it!)}}$.

Sample mean of X

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

Central Limit Theorem (Data 8 + today's terminology)

8422703

For an i.i.d. sample of X_i's of size n,
Where n is "big enough",
and X_i ~ Unknown, where $E(X_i)=\mu$ and $SD(X_i)=\sigma$,
the distribution of $\bar{X}_n$, the **sample mean** of X_i's ,
is **roughly normal** with $\underbrace{\text{mean } \mu \text{ and SD } \sigma/\sqrt{n}}_{\text{(Let's prove it!)}}$

Sample mean of X

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

Expectation:

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

Variance/Standard Deviation:

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

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

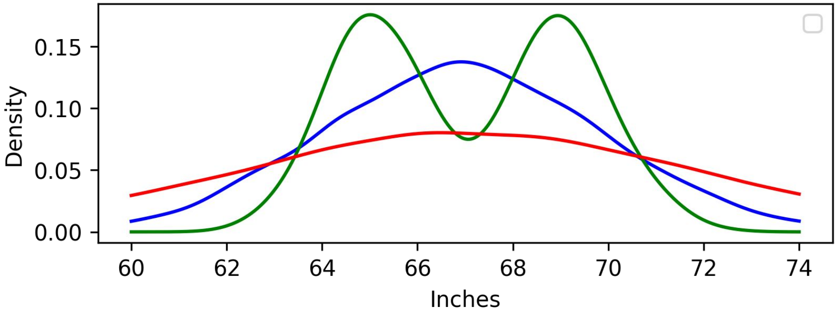
The Central Limit Theorem (CLT) in Data 100

8422703

Data-generating process (DGP):

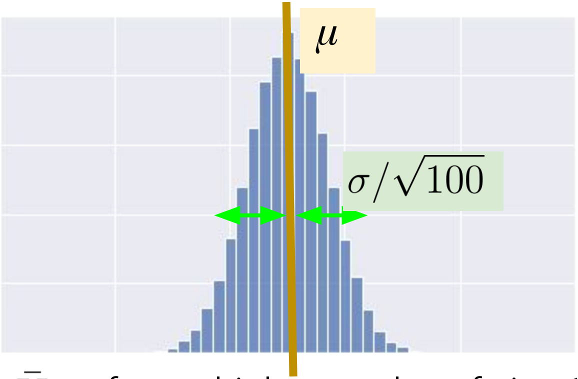
$$X_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(X_i) = \mu \qquad \text{Var}(X_i) = \sigma^2$$



Possible distributions of the raw data

Central Limit Theorem (CLT) (Data 8)



$\bar{X}_{100}$ for multiple samples of size 100

Understanding the "parallel universe" setup of the CLT is critical to the rest of this lecture.

Next lecture, we'll learn how to construct parallel universes. Today, take them for granted 🙂.

Which of the following is true about a data-generating process (DGP)? Select all that apply.

DGP Definitions

8422703

Which of the following is true about a data-generating process (DGP)? Select all that apply.

✅ A DGP is a model for how data are randomly drawn from a true distribution or population.

✅ We typically do not observe the true structure of a DGP.

✅ We typically use an observed sample of data to estimate properties of a DGP.

❌ After our analysis is complete, we often confirm whether estimated DGP properties are equal to the true DGP properties.

We rarely observe the DGP! Our analysis often assumes the data is generated with a certain structure, and we estimate components of that assumed structure.

Like before, "All models are wrong, but some are useful."

Properties of the estimator $\bar{X}_n$

8422703

Data-generating process (DGP):

$$X_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(X_i) = \mu \qquad \text{Var}(X_i) = \sigma^2$$

There are infinite possible samples of size **n** we could have drawn! But, we observe **just one sample**.

What is the behavior of $\bar{X}_n$ across parallel sampling universes?

$\bar{X}_{n,\text{Sample 1}}$

$\bar{X}_{n,\text{Sample 2}}$

$\bar{X}_{n,\text{Sample 3}}$

$\cdots$

$\bar{X}_{n,\text{Sample } \infty}$

Bias of $\bar{X}_n$: On average, how close are the $\bar{X}_n$'s to μ?

Variance of $\bar{X}_n$: How spread out are the $\bar{X}_n$'s from each other?

MSE of $\bar{X}_n$: What's the expected squared difference between $\bar{X}_n$ and μ?

Generalizing our setup to $\boldsymbol{\theta}$, an arbitrary property of the DGP

8422703

Data-generating process (DGP):

$$X_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$\boldsymbol{\theta}$ is a property of the unknown distribution.
[μ, $\boldsymbol{\sigma}^2$, median are some example $\boldsymbol{\theta}$'s]

$\hat{\theta}_n$ is an **estimator** of $\boldsymbol{\theta}$ calculated with a sample of X_i's of size **n**. For example, $\bar{X}_n$ is an estimator of μ.

What is the behavior of $\hat{\theta}_n$ across parallel sampling universes?

$\hat{\theta}_{n,\text{Sample } 1}$

$\hat{\theta}_{n,\text{Sample } 2}$

$\hat{\theta}_{n,\text{Sample } 3}$

. . .

$\hat{\theta}_{n,\text{Sample } \infty}$

Bias of $\hat{\theta}_n$: On average, how close are the $\hat{\theta}_n$'s to $\boldsymbol{\theta}$?

Variance of $\hat{\theta}_n$: How spread out are the $\hat{\theta}_n$'s from each other?

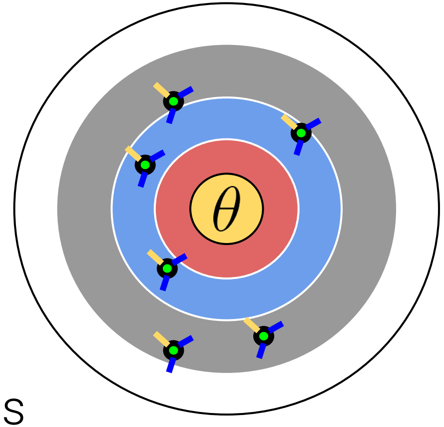
MSE of $\hat{\theta}_n$: What's the expected squared difference between $\hat{\theta}_n$ and $\boldsymbol{\theta}$?

What is a good estimator?

8422703

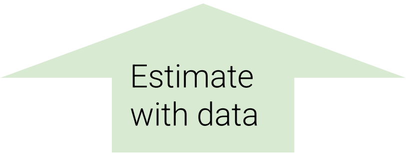
Archery Analogy:

- Center of the target is the **true** θ
- Each arrow corresponds to a separate **parameter estimate** $\hat{\theta}$ obtained from a different random sample.



Population parameter
True parameter
DGP property
Estimand

θ



Sample statistic
Estimator

$\hat{\theta}$

For UC Berkeley heights:

- Center of the target is μ, the **true average height** of Berkeley undergrads
- Each arrow corresponds to a **sample mean** $\bar{X}_n$ computed from a different random sample.

What is a good estimator?

8422703

To evaluate the quality of an **estimator** $\hat{\theta}$, we can think about its behavior across parallel sampling universes:

On average, how close is the estimator to $\boldsymbol{\theta}$?

$$\mathrm{Bias}\left(\hat{\theta}\right) = E\left[\hat{\theta} - \theta\right] = E\left[\hat{\theta}\right] - \theta$$

How variable is the estimator across different random samples?

$$\mathrm{Var}\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - E\left[\hat{\theta}\right]\right)^2\right]$$

What's the average squared difference between the estimator and $\boldsymbol{\theta}$?

$$\mathrm{MSE}\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - \theta\right)^2\right]$$

If the bias of an estimator $\hat{\theta}$ is **zero**, then it is said to be an **unbiased estimator**.

Population parameter
True parameter
DGP property
Estimand

θ

Estimate with data

Sample statistic
Estimator

$\hat{\theta}$

8422703

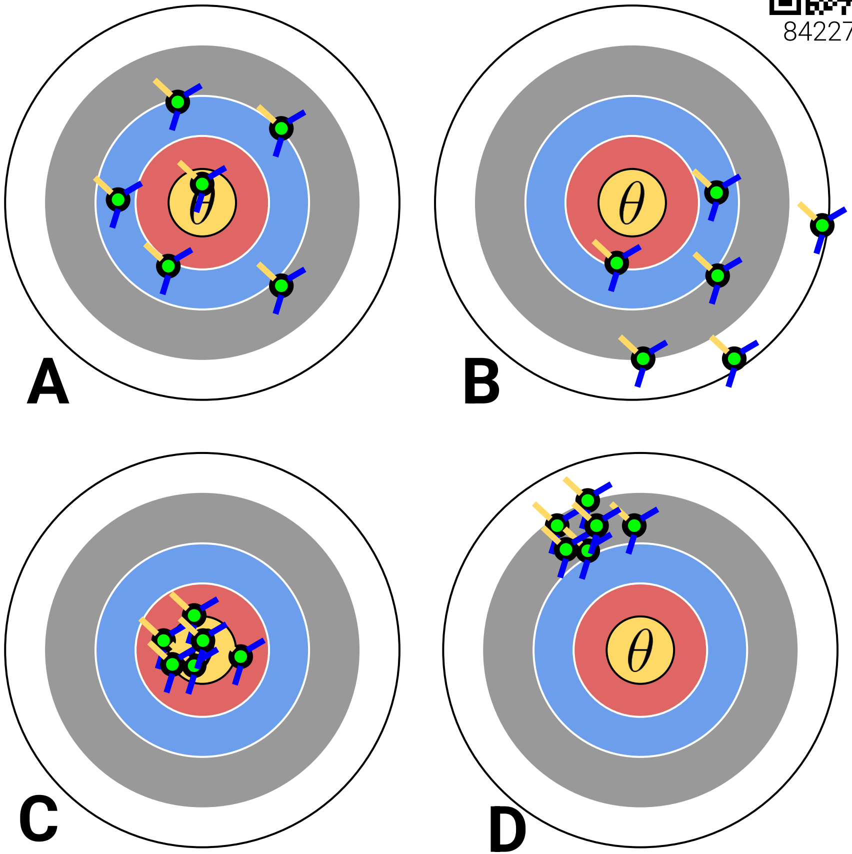
Archery Analogy:

- Center of the target is the **true** θ
- Each arrow corresponds to a separate **parameter estimate** $\hat{\theta}$ obtained from a different random sample.

$$\text{Bias}\left(\hat{\theta}\right) = E\left[\hat{\theta} - \theta\right] = E\left[\hat{\theta}\right] - \theta$$

$$\text{Var}\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - E\left[\hat{\theta}\right]\right)^2\right]$$

Slido: Which target demonstrates
high variance and low bias?



Which target demonstrates high variance and low bias?

What is a good estimator?

8422703

Archery Analogy:

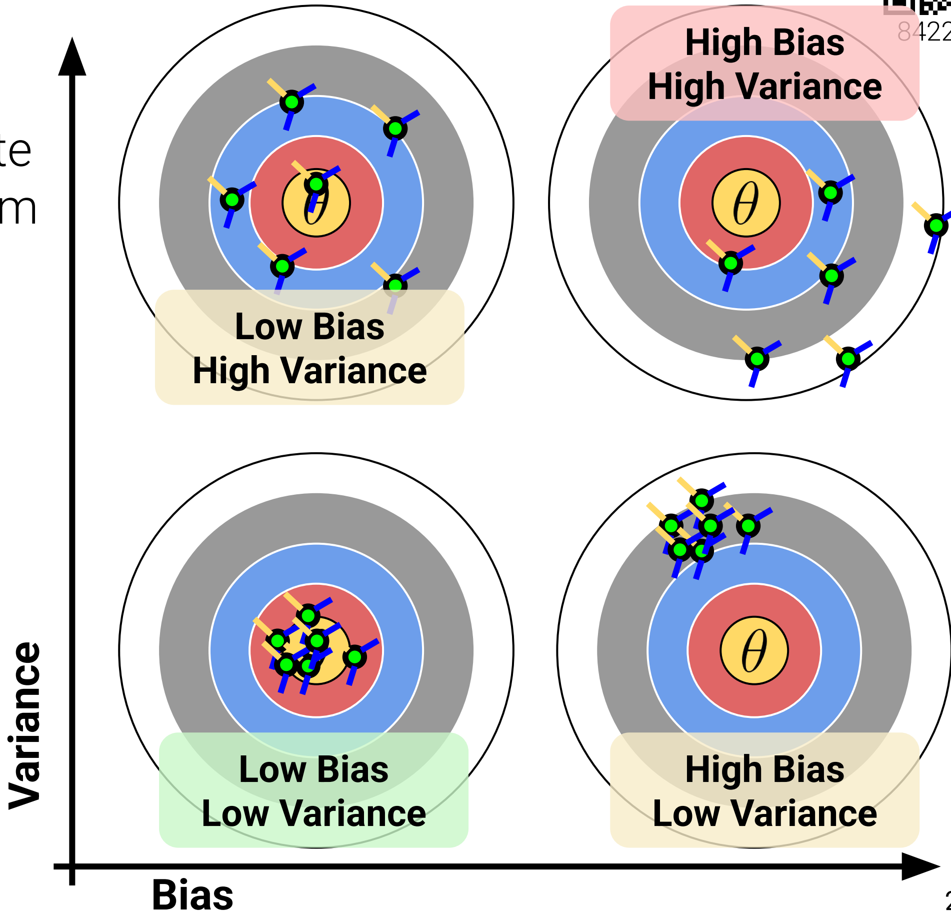
- Center of the target is the **true** θ
- Each arrow corresponds to a separate **parameter estimate** $\hat{\theta}$ obtained from a different random sample.

On average, how close is the estimator to $\boldsymbol{\theta}$?

$$\mathrm{Bias}\left(\hat{\theta}\right) = E\left[\hat{\theta} - \theta\right] = E\left[\hat{\theta}\right] - \theta$$

How variable is the estimator across different random samples?

$$\mathrm{Var}\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - E\left[\hat{\theta}\right]\right)^2\right]$$



A new data-generating process

8422703

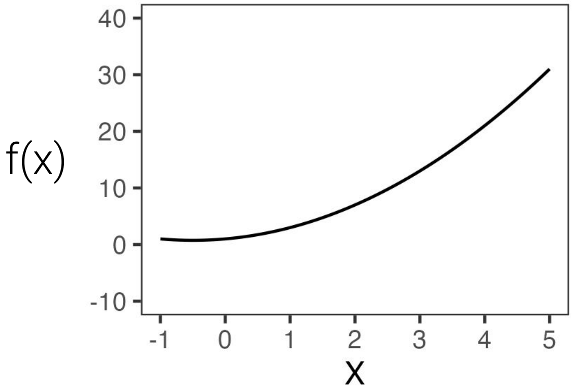
Data-generating process (DGP):

For a fixed set of features X_i ,

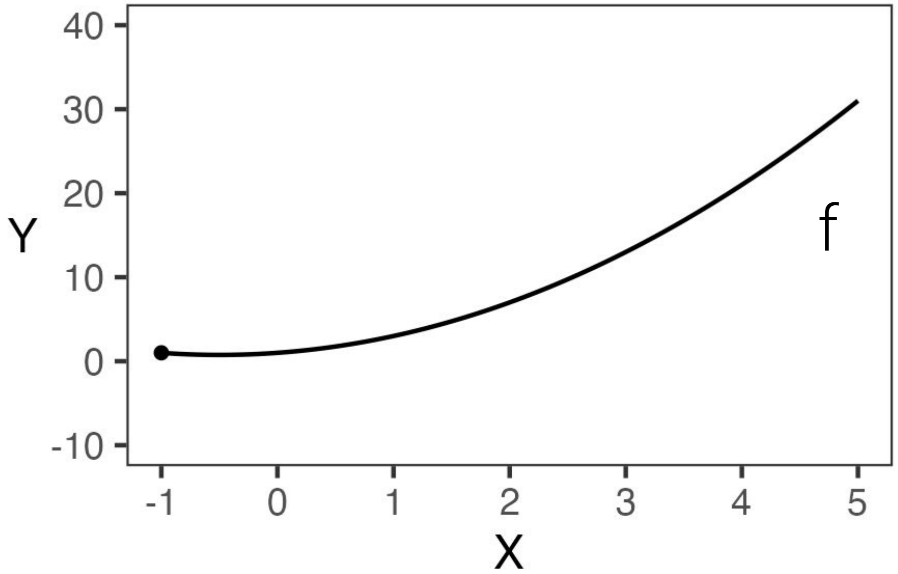
$$Y_i = f(X_i) + \epsilon_i$$

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(\epsilon) = 0 \qquad \text{Var}(\epsilon) = \sigma^2$$



Black points are the $f(X_i)$'s
Black lines are the random ϵ_i's
Blue points are what we observe.



Goal of modeling: How well can we reconstruct **f** with just the **blue** points?

The image above is a GIF. Be sure to view in slideshow mode!

A new data-generating process

8422703

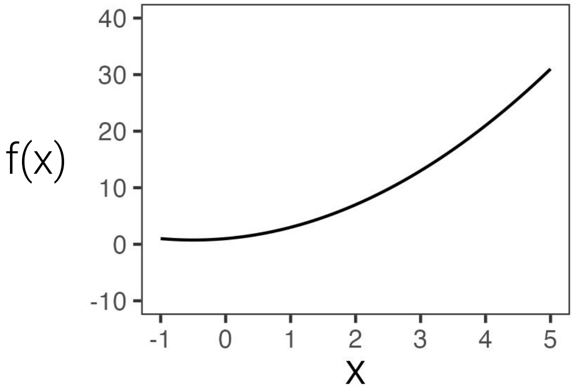
Data-generating process (DGP):

For a fixed set of features X_i ,

$$Y_i = f(X_i) + \epsilon_i$$

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(\epsilon) = 0 \qquad \text{Var}(\epsilon) = \sigma^2$$

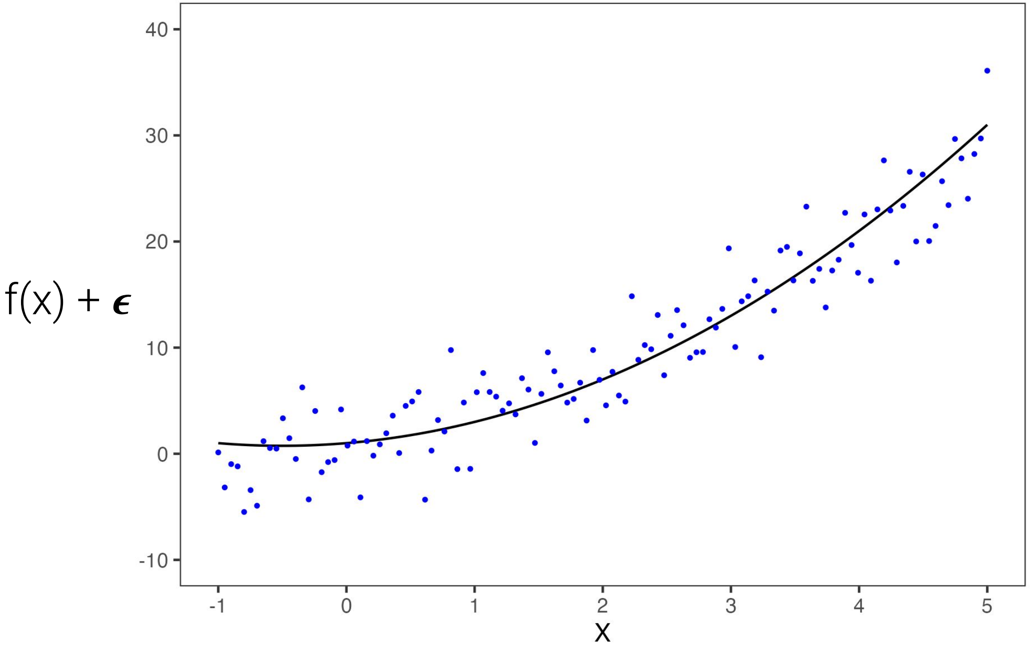


X is 120 evenly spaced points from -1 to 5.

X is fixed/given/constant!

$f(X) = 1 + X + X^2$ $\quad$ $\text{Var}(\boldsymbol{\epsilon}) = 9$

We assume f is **fixed but unknown** to us.



The image above is a GIF. Be sure to view in slideshow mode!

Our model tries to reconstruct the underlying function, but it cannot address noise

8422703

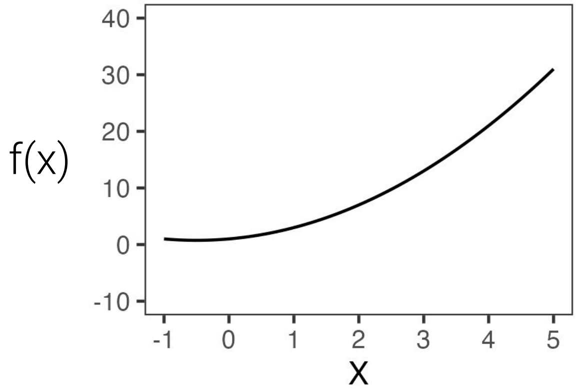
Data-generating process (DGP):

For a fixed set of features X_i ,

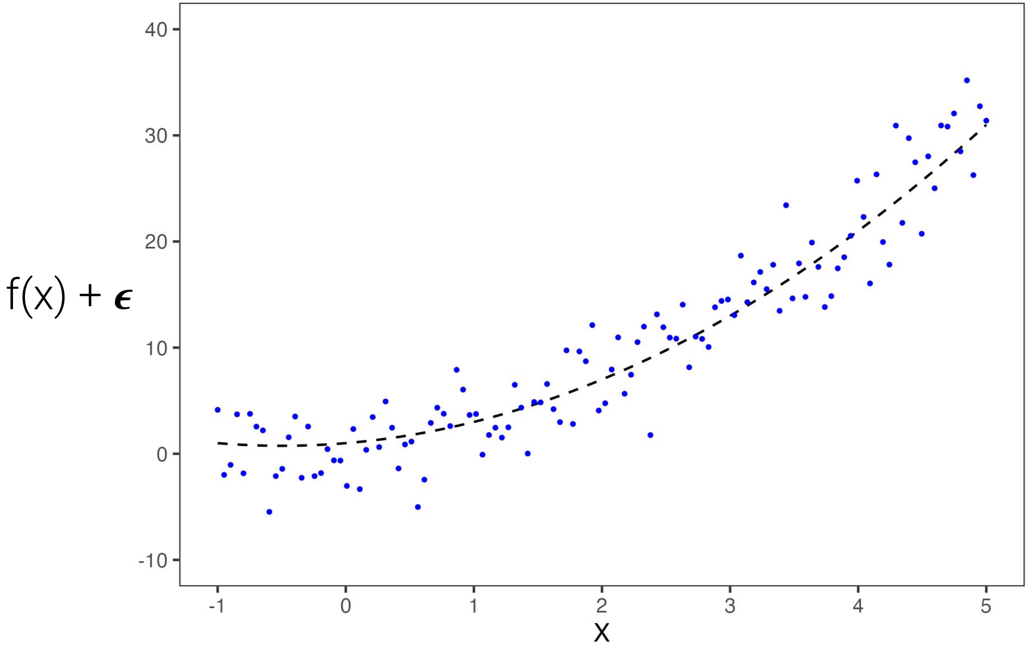
$$Y_i = f(X_i) + \boxed{\epsilon_i}$$

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(\epsilon) = 0 \qquad \text{Var}(\epsilon) = \sigma^2$$



Suppose we fit the model $\hat{f}(X) = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 X + \boldsymbol{\theta}_2 X^2$.
On average, our model predicts the same as f.
Model is **unbiased**, but not perfect. Random noise!



The image above is a GIF. Be sure to view in slideshow mode!

If our model has low complexity, it will likely be biased and low variance

8422703

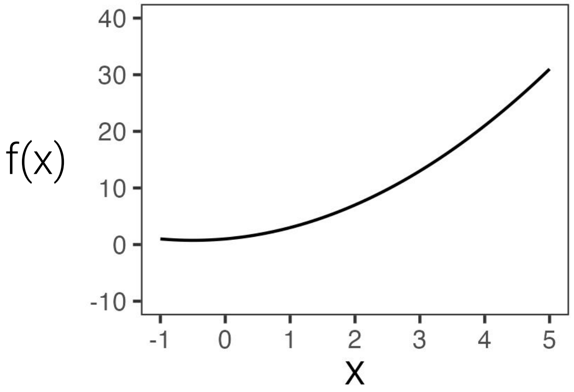
Data-generating process (DGP):

For a fixed set of features X_i ,

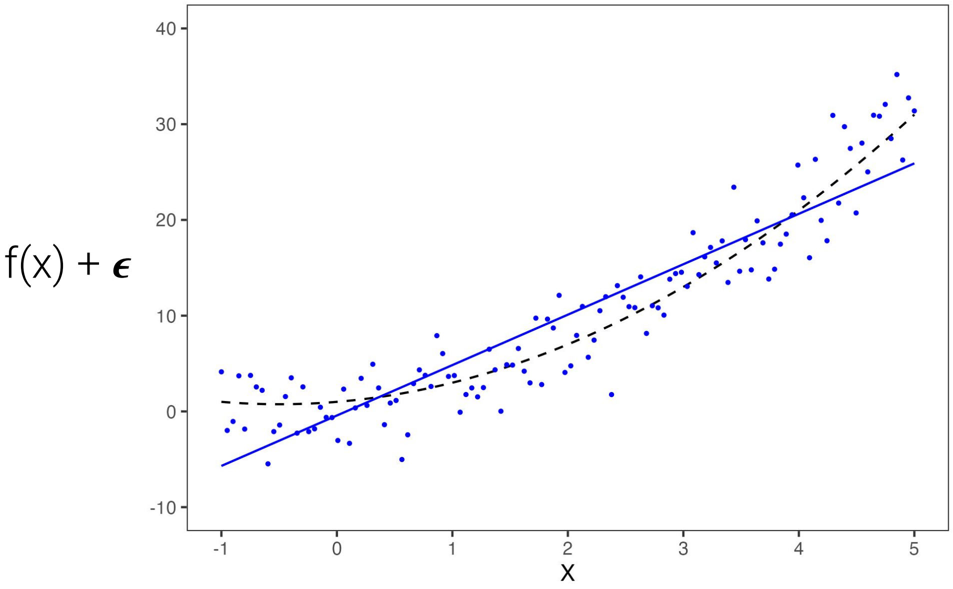
$$Y_i = f(X_i) + \epsilon_i$$

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(\epsilon) = 0 \qquad \text{Var}(\epsilon) = \sigma^2$$



This time, we fit the model $\hat{f}(X) = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 X$. Model is systematically incorrect, on average. Model is **biased**! But, it looks similar across datasets. So, model has **low variance**.



The image above is a GIF. Be sure to view in slideshow mode!

If our model has <u>high</u> complexity, it will likely have <u>low bias</u> and <u>high variance</u>

8422703

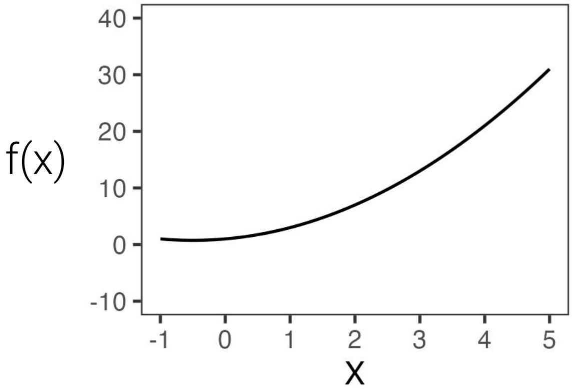
Data-generating process (DGP):

For a fixed set of features X_i ,

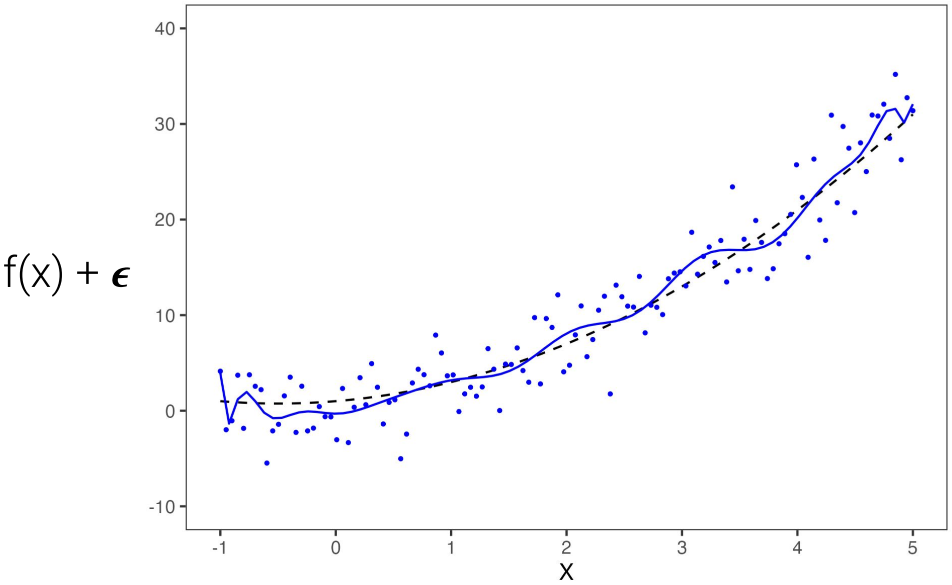
$$Y_i = f(X_i) + \epsilon_i$$

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(\epsilon) = 0 \qquad \text{Var}(\epsilon) = \sigma^2$$



We fit a 20th degree polynomial.
Model is correct on average. **Unbiased**!
But, big changes to $\hat{f}$ across datasets. **High variance**!



The image above is a GIF. Be sure to view in slideshow mode!

Suppose we fit an OLS model to data randomly generated by the given DGP. Which of the given OLS specifications will have the lowest bias?

Bias and variance with polynomials

8422703

Suppose $Y_i = f(X_i) + c_i$, where c_i is an i.i.d. r.v. with $E(c) = 0$ and $Var(c) = 1$.

As it turns out, $f(x) = 1 + x$

Suppose we fit an OLS model to data randomly generated by the DGP above. Which of the following OLS specifications will have the **lowest** bias?

A. $\boldsymbol{\theta}_0$ → Biased! Insufficient complexity to model $f(x) = 1 + x$.
B. $\boldsymbol{\theta}_0+\boldsymbol{\theta}_1 X$ → Unbiased.
C. $\boldsymbol{\theta}_0+\boldsymbol{\theta}_1 X+\boldsymbol{\theta}_2 X^2$ → Also unbiased! On average, $\boldsymbol{\theta}_2$ will be 0.

Suppose we fit an OLS model to data randomly generated by the given DGP. Which of the given OLS specifications will have the lowest variance?

Bias and variance with polynomials

8422703

Suppose $Y_i = f(X_i) + c_i$, where c_i is an i.i.d. r.v. with $E(c) = 0$ and $Var(c) = 1$.

As it turns out, $f(x) = 1 + x$

Suppose we fit an OLS model to data randomly generated by the DGP above. Which of the following OLS specifications will have the **lowest** bias?

A. $\boldsymbol{\theta}_0$ → Biased! Insufficient complexity to model $f(x) = 1 + x$.
B. $\boldsymbol{\theta}_0+\boldsymbol{\theta}_1X$ → Unbiased.
C. $\boldsymbol{\theta}_0+\boldsymbol{\theta}_1X+\boldsymbol{\theta}_2X^2$ → Also unbiased! On average, $\boldsymbol{\theta}_2$ will be 0.

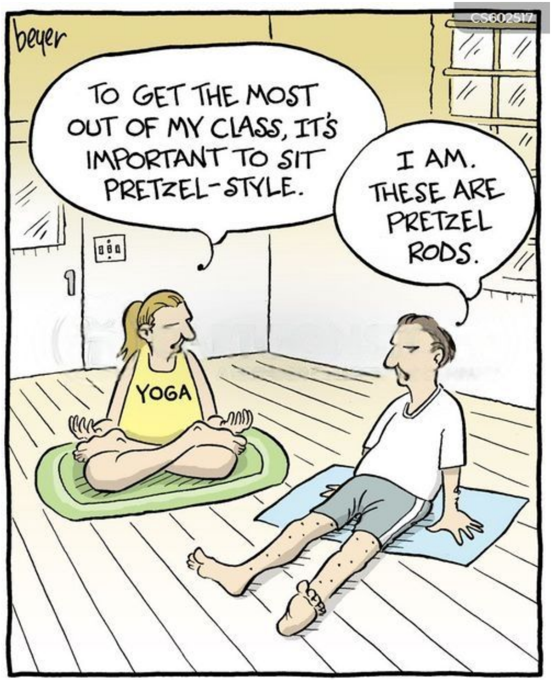
Which of the following OLS specifications will have the **lowest** variance?

A. $\boldsymbol{\theta}_0$ → Lowest variance. Least model complexity to vary from sample to sample.
B. $\boldsymbol{\theta}_0+\boldsymbol{\theta}_1X$ → Higher variance, but the ideal model since unbiased!
C. $\boldsymbol{\theta}_0+\boldsymbol{\theta}_1X+\boldsymbol{\theta}_2X^2$ → Even higher variance.

2-minute stretch break!

Lecture 18, Data 100 Spring 2025

8422703



Prediction model represented as a DGP

8422703

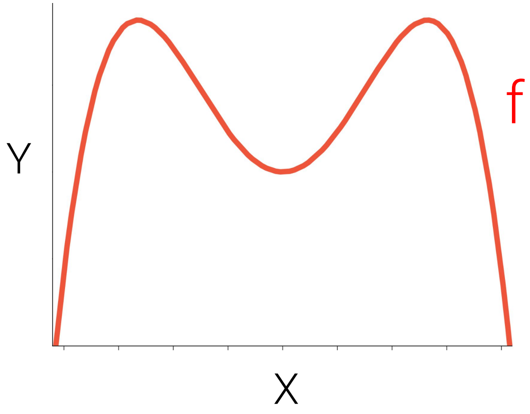
Data-generating process (DGP):

For a fixed set of features X_i ,

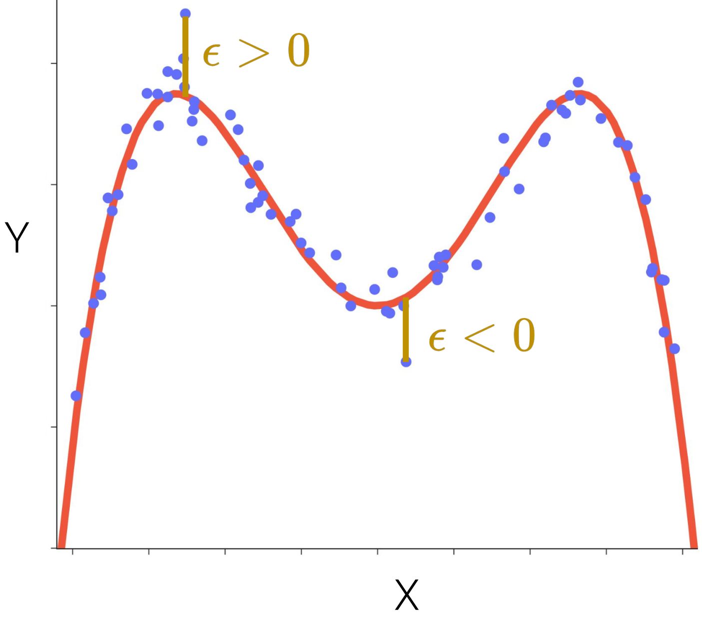
$$Y_i = f(X_i) + \epsilon_i$$

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(\epsilon) = 0 \qquad \text{Var}(\epsilon) = \sigma^2$$



Blue points: Random sample of (X_i,Y_i)'s



Prediction model: How well can we reconstruct f with a sample of (X_i,Y_i)'s?

Estimated model parameters are random

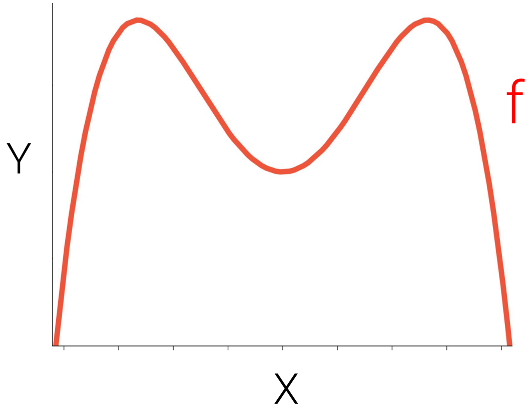
Data-generating process (DGP):

For a fixed set of features X_i ,

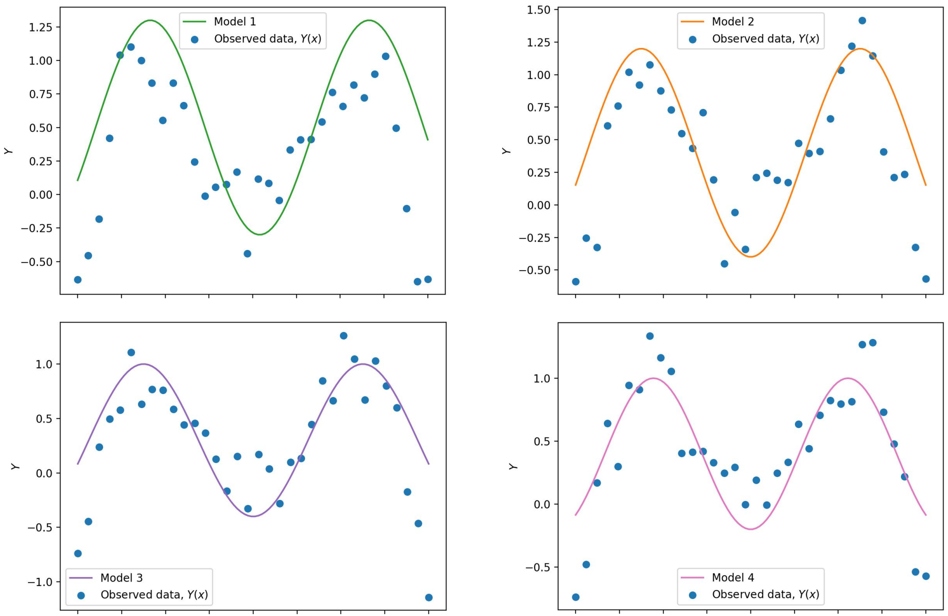
$$Y_i = f(X_i) + \epsilon_i$$

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(\epsilon) = 0 \qquad \text{Var}(\epsilon) = \sigma^2$$



The parameters of our fitted model depend on our training data. **If the data are random, the fitted model is random, too!**



Evaluating the quality of a model across parallel universes

8422703

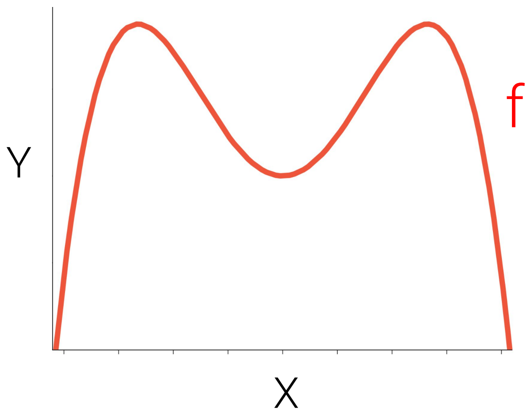
Data-generating process (DGP):

For a fixed set of features X_i ,

$$Y_i = f(X_i) + \epsilon_i$$

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(\epsilon) = 0 \qquad \text{Var}(\epsilon) = \sigma^2$$



Just like an estimator, we can evaluate a **model's quality** by considering its behavior across different training datasets (i.e., parallel sampling universes):

Model bias: How close is our fitted model to f, on average?

Model variance: How much does our fitted model vary across random samples?

Model risk (MSE): What's the typical squared error between our model's predictions and the actual outcomes?

8422703

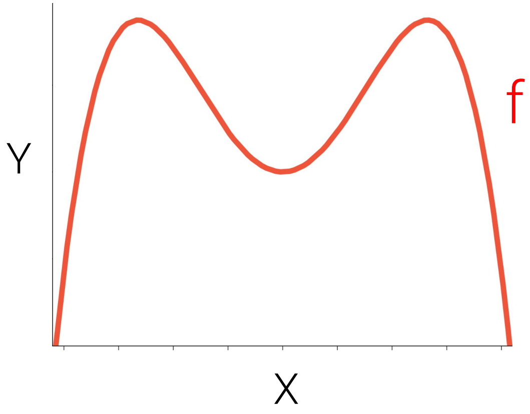
Data-generating process (DGP):

For a fixed set of features X_i ,

$$Y_i = f(X_i) + \epsilon_i$$

$$\epsilon_i \overset{\text{iid}}{\sim} \text{Unknown}$$

$$\mathbb{E}(\epsilon) = 0 \qquad \text{Var}(\epsilon) = \sigma^2$$



For a <u>fixed/given/constant</u> set of features X:

Model bias: How close is our fitted model to f, on average?

$$\text{Bias}\left(\hat{f}(X)\right) = \mathbb{E}\left[\hat{f}(X)\right] - f(X)$$

Model variance: How much does the fitted model's prediction vary across samples?

$$\text{Var}\left(\hat{f}(X)\right) = \mathbb{E}\left[\left(\hat{f}(X) - \mathbb{E}\left[\hat{f}(X)\right]\right)^2\right]$$

Model risk (MSE): What's the average squared error between our model's prediction and the actual outcome, across samples? $\mathbb{E}\left[\left(Y - \hat{Y}\right)^2\right]$

Prediction model setup and notation

8422703

Goal: What is the model risk for a single observation $\vec{X}$? $\vec{X}$ is given, so it is not random.

1a. The true DGP (i.e., population model) has the form $Y = f(\vec{X}) + \epsilon$

1b. We assume the function **f** is fixed but unknown. In other words, **f** is not random.

1c. **ϵ** is random noise generated i.i.d. from a distribution with mean 0 and variance $\boldsymbol{\sigma}^2$.

1d. **Y** is the observed outcome. **Y** depends on **ϵ**, so **Y** is random.

2a. We have a random sample of training data.

2b. We fit our own model $\hat{f}$ to this random training data. So, $\hat{f}$ is random, too.

2c. We get a prediction by plugging X into $\hat{f}$. In other words, $\hat{Y} = \hat{f}(\vec{X})$. So, $\hat{Y}$ is random.

3. To calculate model risk, we compute $\mathbb{E}[(Y - \hat{Y})^2]$.

Decomposition of model risk for a single observation

8422703

Goal: Compute $\mathbb{E}[(Y - \hat{Y})^2]$.

f and $\vec{X}$ are fixed.

$$Y = f(\vec{X}) + \epsilon$$
$$\mathbb{E}(\epsilon) = 0$$
$$\text{Var}(\epsilon) = \sigma^2$$
$$\hat{Y} = \hat{f}(\vec{X})$$

Probability rules:

$$\text{Var}(X) = E[X^2] - (E[X])^2$$
$$E[aX + b] = aE[X] + b$$
$$E[X + Y] = E[X] + E[Y]$$
$$\text{Var}(aX + b) = a^2\text{Var}(X)$$
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

if X and Y are independent!

Decomposition of model risk for a single observation

8422703

$$\text{Goal: Compute } \mathbb{E}[(Y - \hat{Y})^2].$$

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

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

f and $\vec{X}$ are fixed.

$$Y = f(\vec{X}) + \epsilon$$
$$\mathbb{E}(\epsilon) = 0$$
$$\text{Var}(\epsilon) = \sigma^2$$
$$\hat{Y} = \hat{f}(\vec{X})$$

Probability rules:

$$\boxed{\text{Var}(X) = E[X^2] - (E[X])^2}$$
$$E[aX + b] = aE[X] + b$$
$$E[X + Y] = E[X] + E[Y]$$
$$\text{Var}(aX + b) = a^2\text{Var}(X)$$
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

if X and Y are independent!

Decomposition of model risk for a single observation

8422703

Goal: Compute $\mathbb{E}[(Y-\hat{Y})^2]$.

$$\mathrm{Var}(Y-\hat{Y}) = \mathbb{E}[(Y-\hat{Y})^2] - \left(\mathbb{E}[Y-\hat{Y}]\right)^2$$

$$\mathbb{E}[(Y-\hat{Y})^2] = \boxed{\mathrm{Var}(Y-\hat{Y})} + \left(\mathbb{E}[Y-\hat{Y}]\right)^2$$

$$\boxed{\begin{aligned}\mathrm{Var}(Y-\hat{Y}) &= \mathrm{Var}\left(f(\vec{X}) + \epsilon - \hat{f}(\vec{X})\right)\\ &= \mathrm{Var}\left(\epsilon - \hat{f}(\vec{X})\right)\\ &= \mathrm{Var}(\epsilon) + \mathrm{Var}\left(\hat{f}(\vec{X})\right)\\ &= \sigma^2 + \mathrm{Var}\left(\hat{f}(\vec{X})\right)\end{aligned}}$$

f and $\vec{X}$ are fixed.

$$Y = f(\vec{X}) + \epsilon$$
$$\mathbb{E}(\epsilon) = 0$$
$$\mathrm{Var}(\epsilon) = \sigma^2$$
$$\hat{Y} = \hat{f}(\vec{X})$$

Probability rules:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$
$$E[aX+b] = aE[X] + b$$
$$E[X+Y] = E[X] + E[Y]$$
$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$$
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

if X and Y are independent!

Decomposition of model risk for a single observation

8422703

Goal: Compute $\mathbb{E}[(Y - \hat{Y})^2]$.

$$\mathrm{Var}(Y - \hat{Y}) = \mathbb{E}[(Y - \hat{Y})^2] - \left(\mathbb{E}[Y - \hat{Y}]\right)^2$$

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

f and $\vec{X}$ are fixed.

$$Y = f(\vec{X}) + \epsilon$$
$$\mathbb{E}(\epsilon) = 0$$
$$\mathrm{Var}(\epsilon) = \sigma^2$$
$$\hat{Y} = \hat{f}(\vec{X})$$

Probability rules:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$
$$E[aX + b] = aE[X] + b$$
$$E[X + Y] = E[X] + E[Y]$$
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$
$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

if X and Y are independent!

Decomposition of model risk for a single observation

8422703

Goal: Compute $\mathbb{E}[(Y-\hat{Y})^2]$.

$$\mathrm{Var}(Y-\hat{Y}) = \mathbb{E}[(Y-\hat{Y})^2] - \left(\mathbb{E}[Y-\hat{Y}]\right)^2$$

$$\mathbb{E}[(Y-\hat{Y})^2] = \mathrm{Var}(Y-\hat{Y}) + \boxed{\left(\mathbb{E}[Y-\hat{Y}]\right)^2}$$

$$\boxed{\begin{aligned}\left(\mathbb{E}[Y-\hat{Y}]\right)^2 &= \left(\mathbb{E}[Y]-\mathbb{E}[\hat{Y}]\right)^2 \\ &= \left(\mathbb{E}\left[f(\vec{X})+\epsilon\right]-\mathbb{E}\left[\hat{f}(\vec{X})\right]\right)^2 \\ &= \left(\mathbb{E}\left[f(\vec{X})\right]+\mathbb{E}[\epsilon]-\mathbb{E}\left[\hat{f}(\vec{X})\right]\right)^2 \\ &= \left(f(\vec{X})+0-\mathbb{E}\left[\hat{f}(\vec{X})\right]\right)^2 \\ &= \mathrm{Bias}\left(\hat{f}(\vec{X})\right)^2\end{aligned}}$$

f and $\vec{X}$ are fixed.

$$Y = f(\vec{X}) + \epsilon$$

$$\mathbb{E}(\epsilon) = 0$$

$$\mathrm{Var}(\epsilon) = \sigma^2$$

$$\hat{Y} = \hat{f}(\vec{X})$$

Probability rules:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

$$E[aX+b] = aE[X]+b$$

$$E[X+Y] = E[X]+E[Y]$$

$$\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$$

if X and Y are independent!

Decomposition of model risk for a single observation

8422703

Goal: Compute $\mathbb{E}[(Y - \hat{Y})^2]$.

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

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

$$\mathbb{E}[(Y - \hat{Y})^2] = \boxed{\sigma^2 + \text{Var}\left(\hat{f}(\vec{X})\right)} + \boxed{\text{Bias}\left(\hat{f}(\vec{X})\right)^2}$$

f and $\vec{X}$ are fixed.

$$Y = f(\vec{X}) + \epsilon$$
$$\mathbb{E}(\epsilon) = 0$$
$$\text{Var}(\epsilon) = \sigma^2$$
$$\hat{Y} = \hat{f}(\vec{X})$$

Probability rules:

$$\text{Var}(X) = E[X^2] - (E[X])^2$$
$$E[aX + b] = aE[X] + b$$
$$E[X + Y] = E[X] + E[Y]$$
$$\text{Var}(aX + b) = a^2\text{Var}(X)$$
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

if X and Y are independent!

8422703

Model Risk = **Irreducible error** + **Model Variance** + **(Model Bias)^2**

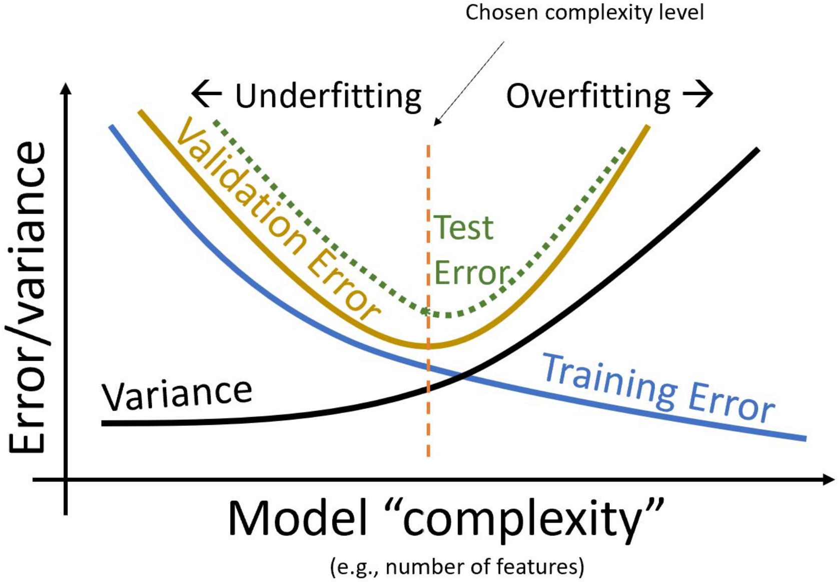
$$\mathbb{E}[(Y - \hat{Y})^2] = \sigma^2 + \text{Var}\left(\hat{f}(\vec{X})\right) + \text{Bias}\left(\hat{f}(\vec{X})\right)^2$$

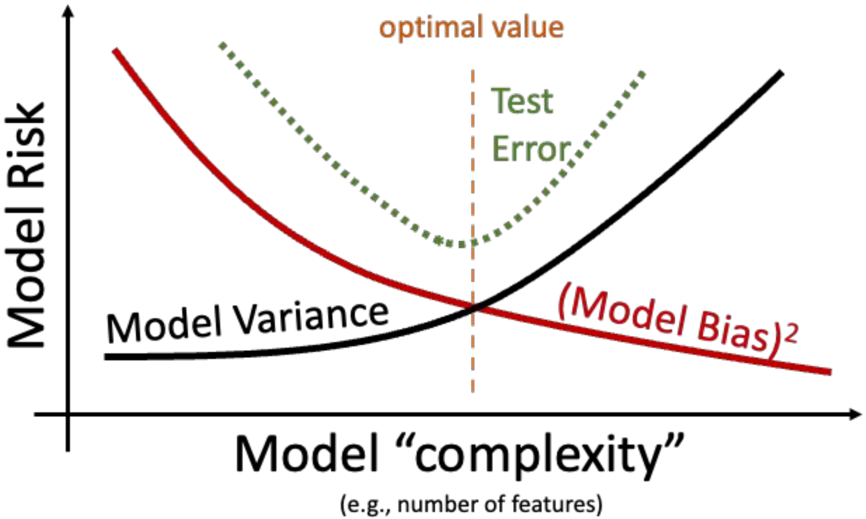
Interpretation:

- **Irreducible error / observational variance / noise** cannot be addressed by modeling.
- **Bias-Variance Tradeoff**:
  - To decrease model bias, we **increase model complexity.** As a result, the model will have **higher model variance**.
  - To decrease model variance, we **decrease model complexity**. The model may **underfit** the sample data and may have **higher model bias**.

The Bias-Variance Tradeoff has been with us all along!

8422703





Practical tips for addressing bias, variance, and irreducible error

8422703

High variance corresponds to **overfitting**.

- Your model may be too complex.
- You can reduce the # of parameters, or regularize.

High bias corresponds to **underfitting**.

- Your model may be too simple to capture complexities in the data.
- You may have overregularized → Regularization biases us towards a constant model in exchange for reduced variance!

Irreducible error

- For a fixed dataset, nothing you can do. That's why it's irreducible.

8422703

LECTURE 18

Estimators, Bias, and Variance

Data 100/Data 200, Spring 2025 @ UC Berkeley

Narges Norouzi and Josh Grossman

Content credit: [Acknowledgments]