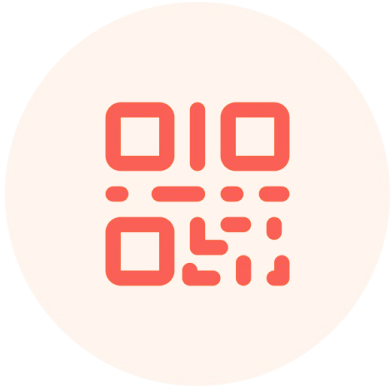
slido



Join at slido.com
#3678970

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

LECTURE 17

Estimators, Bias, and Variance

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

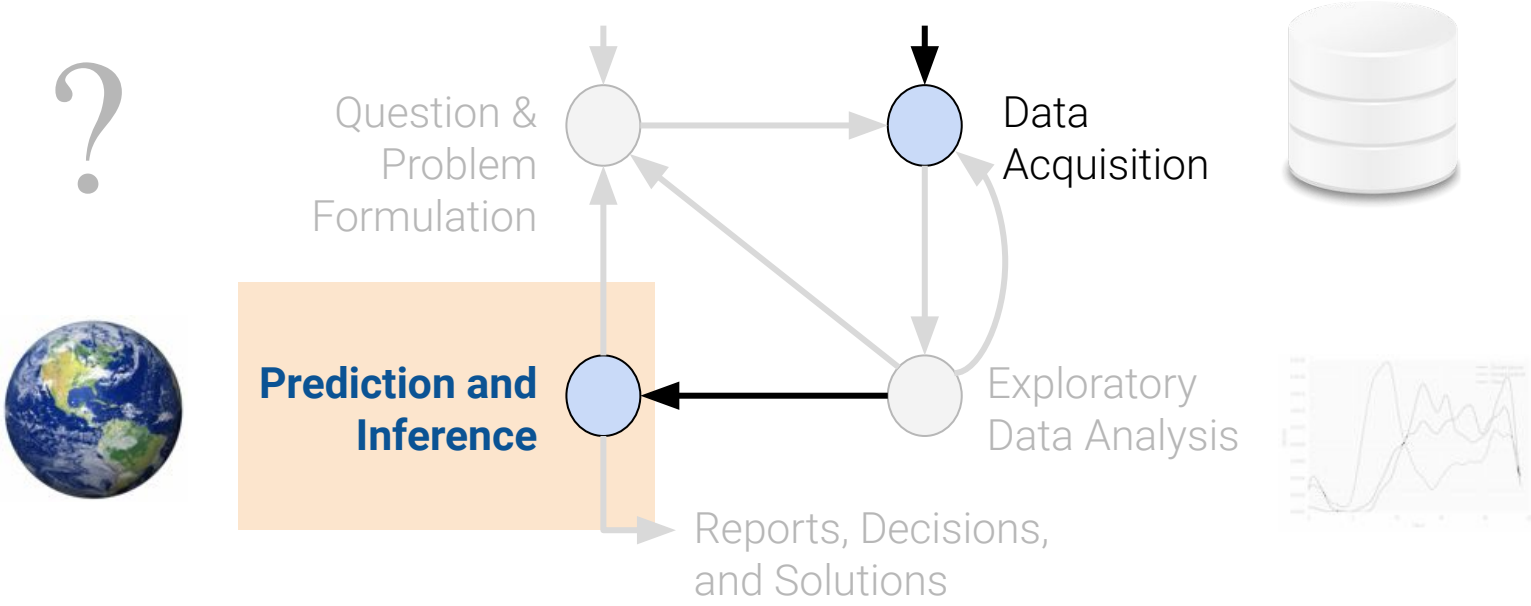
Data 100/Data 200, Spring 2023 @ UC Berkeley

Narges Norouzi and Lisa Yan

Content credit: Acknowledgments

Why Probability?

3678970



Model Selection Basics:
Cross Validation
Regularization

Probability I:
Random Variables

(today)

Probability II:
Estimators
Bias and Variance

Today's Roadmap

Lecture 17, Data 100 Spring 2023

Sums of Random Variables

- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- Sample Mean
- Central Limit Theorem

Prediction vs. Inference

- Modeling: Assumptions of Randomness

The Bias-Variance Tradeoff

[Extra] Derivations

[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance Decomposition

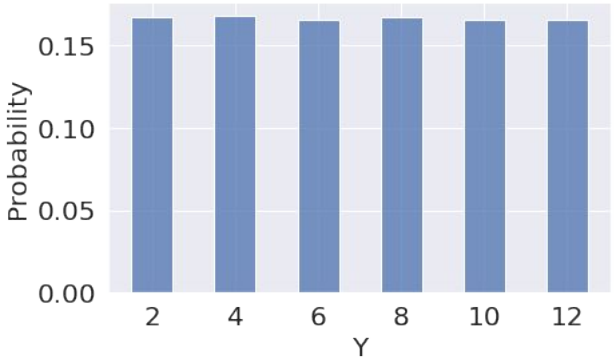
Distributions of Sums

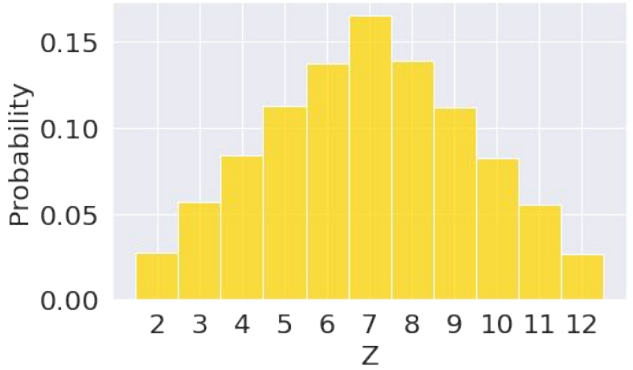
3678970

Let X_1 and X_2 be numbers on two rolls of a die.

- X_1, X_2 are **IID**, so X_1, X_2 have the same distribution.
- But the sums $Y = X_1 + X_1 = 2X_1$ and $Z = X_1 + X_2$ have different distributions!

Let's show this through simulation:





- Same expectation…
- But $Y = 2X_1$ has larger variance!

	Y	Z
E[•]	6.984400	6.984950
Var(•)	11.669203	5.817246
SD(•)	3.416021	2.411897

How can we directly compute $E[Y]$, $\mathrm{Var}(Y)$, **without** simulating distributions?

Demo

Properties of Expectation [1/3]

3678970

Instead of simulating full distributions, we often just compute expectation and variance directly.

Recall definitions of expectation:

$$\mathbb{E}[X] = \sum_{x} xP(X = x)$$

$$\mathbb{E}[X] = \sum_{\substack{\text{all samples} \\ s}} X(s)P(s)$$

Properties:

1. **Expectation is linear**.
Intuition: summations are linear. Proof

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Properties of Expectation [2/3]

3678970

Instead of simulating full distributions, we often just compute expectation and variance directly.

Recall definitions of expectation:

$$\mathbb{E}[X] = \sum_{x} xP(X = x)$$

$$\mathbb{E}[X] = \sum_{\substack{\text{all samples} \\ s}} X(s)P(s)$$

Properties:

1. **Expectation is linear**.
Intuition: summations are linear. Proof

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

2. Expectation is linear in sums of RVs, for any relationship between X and Y. Proof

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Properties of Expectation [3/3]

3678970

Instead of simulating full distributions, we often just compute expectation and variance directly.

Recall definitions of expectation:

$$\mathbb{E}[X] = \sum_x xP(X = x)$$

$$\mathbb{E}[X] = \sum_{\substack{\text{all samples} \\ s}} X(s)P(s)$$

Properties:

1. **Expectation is linear**.
Intuition: summations are linear. Proof

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

2. Expectation is linear in sums of RVs,
for any relationship between X and Y. Proof

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

3. If g is a non-linear function, then in general

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$$

.

- Example: if X is *-1* or *1* with equal probability, then $E[X] = 0$ but $E[X^2] = 1 \neq 0$.

Properties of Variance [1/2]

3678970

Recall definition of variance: $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$

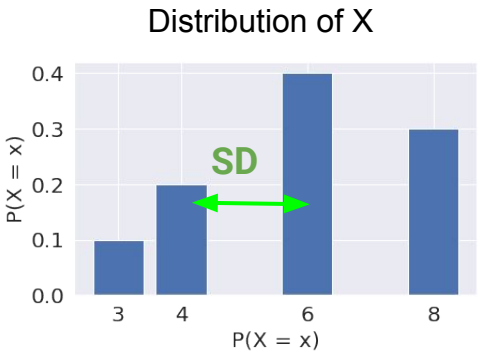
Properties:

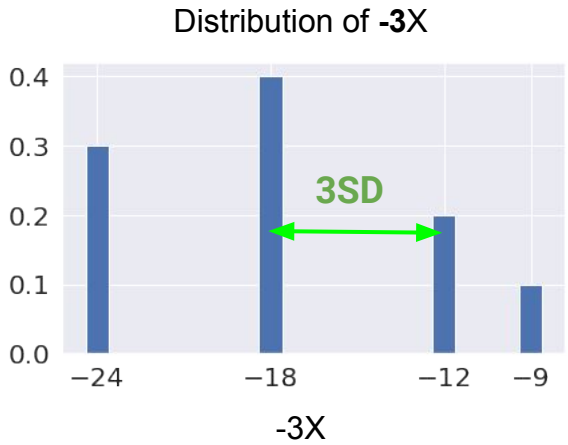
1. **Variance is non-linear**:

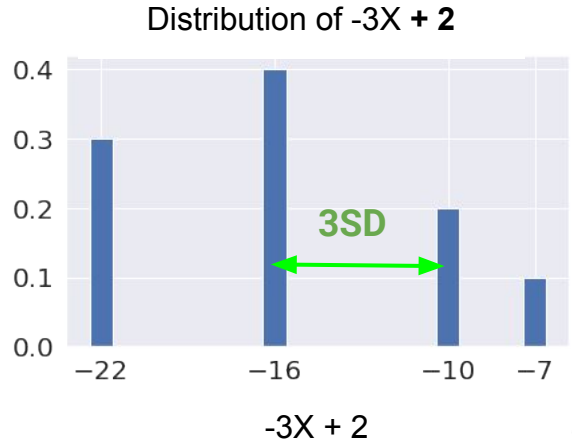
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

Intuition (full proof): Consider the Standard Deviation for $Y = -3X+2$:

$$\mathrm{SD}(aX + b) = |a|\mathrm{SD}(X)$$







Properties of Variance [2/2]

3678970

Recall definition of variance: $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$

Properties:

1. **Variance is non-linear**:

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

Intuition (full proof): Consider the Standard Deviation for $Y = -3X+2$:

$$\mathrm{SD}(aX + b) = |a|\mathrm{SD}(X)$$

2. Variance of sums of RVs is affected by the (in)dependence of the RVs (derivation):

$$Var(X + Y) = Var(X) + Var(Y) + 2\mathrm{Cov}(X, Y)$$

Covariance of X and Y (next slide).
If X, Y independent,
then Cov(X, Y) = 0.

Covariance and Correlation: The Basics

3678970

Covariance is the expected product of deviations from expectation.

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

- A generalization of variance. Note $\mathrm{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathrm{Var}(X)$.
- Interpret by defining **correlation** (yes, *that* correlation!):

$$r(X, Y) = \mathbb{E}\left[\left(\underbrace{\frac{X - \mathbb{E}[X]}{\mathrm{SD}(X)}}_{\text{standard units of } X \text{ (link)}}\right)\left(\frac{Y - \mathbb{E}[Y]}{\mathrm{SD}(Y)}\right)\right] = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\mathrm{SD}(Y)}$$

Correlation (and therefore covariance) measures a linear relationship between X and Y.

Covariance and Correlation: The Basics

3678970

Covariance is the expected product of deviations from expectation.

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

- A generalization of variance. Note $\text{Cov}(X, X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}(X)$.
- Interpret by defining **correlation** (yes, *that* correlation!):

$$r(X, Y) = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\text{SD}(X)}\right)\left(\frac{Y - \mathbb{E}[Y]}{\text{SD}(Y)}\right)\right] = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}$$

standard units of X (link)

Correlation (and therefore covariance) measures a linear relationship between X and Y.

- If X and Y are correlated, then knowing X tells you something about Y.
- "X and Y are uncorrelated" is the same as "Correlation and covariance equal to 0".
- **Independent X, Y are uncorrelated**, because knowing X tells you nothing about Y.
- The converse is not necessarily true: **X, Y could be uncorrelated but not independent**.
- For more info, see extra slides + take STAT 140/EECS 70.

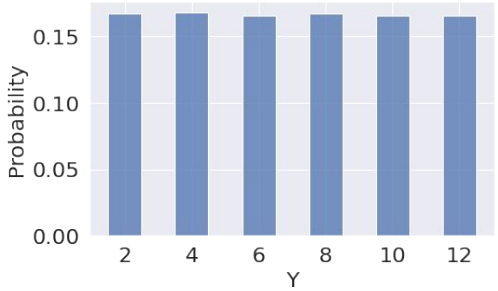
Dice, Our Old Friends: Expectation

3678970

Let X_1 and X_2 be numbers on two rolls of a die.

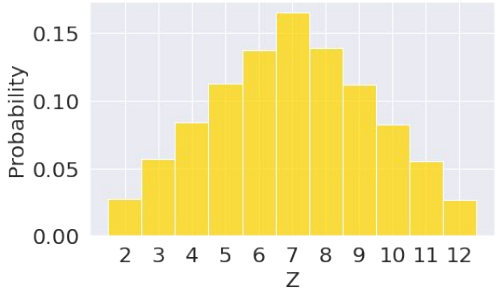
- X_1, X_2 are **IID**, so X_1, X_2 have the same distribution.
- Therefore $E[X_1] = E[X_2] = 7/2 \quad \text{Var}(X_1) = \text{Var}(X_2) = 35/12$

$Y = 2X_1$



$$E[Y] = E[2X_1] = 2E[X_1] = 7$$

$Z = X_1 + X_2$



$$E[Z] = E[X_1] + E[X_2] = (7/2) + (7/2) = 7$$

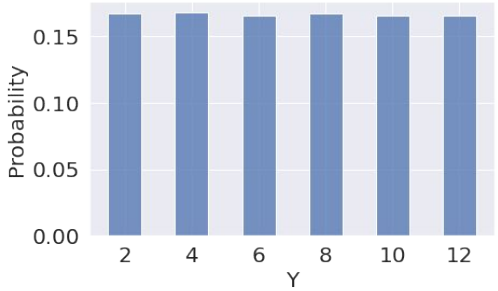
Dice, Our Old Friends: Variance

3678970

Let X_1 and X_2 be numbers on two rolls of a die.

- X_1, X_2 are **IID**, so X_1, X_2 have the same distribution.
- Therefore $E[X_1] = E[X_2] = 7/2$ $\quad \text{Var}(X_1) = \text{Var}(X_2) = 35/12$

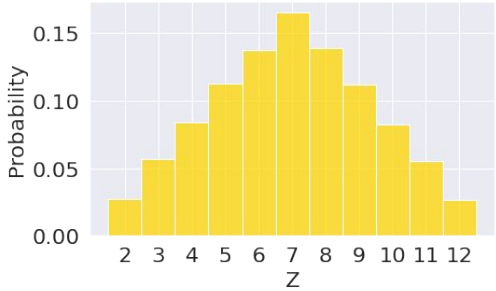
$Y = 2X_1$



$E[Y] = E[2X_1] = 2E[X_1] = 7$

$$\text{Var}(Y) = \text{Var}(2X_1) = 4\text{Var}(X_1)$$
$$= 4(35/12)$$
$$\approx \mathbf{11.67}$$

$Z = X_1 + X_2$



$E[Z] = E[X_1] + E[X_2] = (7/2) + (7/2) = 7$

$$\text{Var}(Z) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X1, X2)$$
$$= (35/12) + (35/12) + 0$$
$$\approx \mathbf{5.83}$$

0

X_1, X_2 independent

[Summary] Expectation and Variance for Linear Functions of Random Variables

3678970

Let X be a random variable with distribution $P(X = x)$.

$$\mathbb{E}[X] = \sum_x xP(X = x)$$

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] && \text{(definition)} \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 && \text{(easier computation)} \end{aligned}$$

Let a and b be scalar values.

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

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

Let Y be another random variable.

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$Var(X + Y) = Var(X) + Var(Y) + 2\mathrm{Cov}(X, Y)$$

Zero if X, Y independent.

Bernoulli and Binomial Random Variables

Lecture 17, Data 100 Spring 2023

3678970

Sums of Random Variables

- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- Sample Mean
- Central Limit Theorem

Prediction vs. Inference

- Modeling: Assumptions of Randomness

The Bias-Variance Tradeoff

[Extra] Derivations

[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance Decomposition

Common Random Variables

3678970

Bernoulli(p)

- Takes on value 1 with probability p, and 0 with probability 1 - p
- AKA the “indicator” random variable.

Binomial(n, p)

- Number of 1s in n independent Bernoulli(p) trials

Uniform on a finite set of values

- Probability of each value is 1 / (size of set)
- For example, a standard die

Uniform on the unit interval(0, 1)

- Density is flat on (0, 1) and 0 elsewhere

Normal(μ, σ^2)

We’ll now revisit these to solidify our understanding of expectation/variance.

Properties of Bernoulli Random Variables

$$\mathbb{E}[X] = \sum_x xP(X = x)$$
$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$
$$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Definitions

3678970

Let X be a **Bernoulli**(p) random variable.

- Takes on value 1 with probability p, and 0 with probability 1 - p.
- AKA the "indicator" random variable.

Expectation:

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

We will get an average value of p across many, many samples

Variance:

$$\mathbb{E}[X^2] = 1^2 \cdot p + 0 \cdot (1 - p) = p$$
$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$
$$= p - p^2 = p(1 - p)$$

Lower Var: p = 0.1 or 0.9
Higher Var: p close to 0.5

More info: google("plot x(1 - x)")

Properties of Binomial Random Variables

3678970

Let Y be a **Binomial**(n, p) random variable.

- Y is the number (i.e., count) of 1s in n independent Bernoulli(p) trials.

We can write: $Y = \sum_{i=1}^{n} X_i$

A count is a **sum** of 0's and 1's.

- X_i is the indicator of success on trial i. X_i = 1 if trial i is a success, else 0.
- All X_i s are **IID** (independent and identically distributed) and **Bernoulli**(p).

Properties of Binomial Random Variables

3678970

Let Y be a **Binomial**(n, p) random variable.

- Y is the number (i.e., count) of 1s in n independent Bernoulli(p) trials.
- Distribution of Y given by the binomial formula (Lecture 2).

We can write: $Y = \sum_{i=1}^{n} X_i$

A count is a sum of 0's and 1's.

- X_i is the indicator of success on trial *i*. X_i = 1 if trial i is a success, else 0.
- All X_i s are **IID** (independent and identically distributed) and **Bernoulli**(p).

Expectation: $\mathbb{E}[Y] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np$

Variance: Because all X_i s are independent, $\mathrm{Cov}(X_i, X_j) = 0$ for all i, j.

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p)$$

3678970

Example

Suppose you win cash based on the number of heads you get in a series of 20 coin flips.

Let $X_i = 1$ if the i-th coin is heads,
0 otherwise

Which payout strategy would you choose?

Hint: Compare expectations and variances.

A. $Y_A = 10 \cdot X_1 + 10 \cdot X_2$

B. $Y_B = \left(\sum_{i=1}^{20} X_i \right)$

C. $Y_C = 20 \cdot X_1$

slido

Suppose you win cash based on the number of heads you get in a series of 20 coin flips.Let Xi = 1 if the i-th coin is heads, and 0 otherwise. Which payout strategy would you choose?

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

Which Would You Pick?

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X, Y)$$
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

3678970

Suppose you win cash based on the number of heads you get in a series of 20 coin flips.

Let X_1 , X_2 , ..., X_{20} be 20 **IID** Bernoulli(0.5) random variables.

- Since X_i s are independent: $\mathrm{Cov}(X_i, X_j) = 0$ for all i, j.
- Since X_i is Bernoulli(p = 0.5): $E[X_i] = p = 0.5,\ \mathrm{Var}(X_i) = p(1\text{-}p) = 0.25.$

Which payout strategy would you choose?

	A. $Y_A = \$10 \cdot X_1 + \$10 \cdot X_2$	**B.** $Y_B = \$\left(\sum_{i=1}^{20} X_i\right)$	**C.** $Y_C = \$20 \cdot X_1$
Expectation			
Variance			
Std. Deviation			

Which Would You Pick?

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X, Y)$$
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

3678970

Suppose you win cash based on the number of heads you get in a series of 20 coin flips.

Let X_1 , X_2 , ..., X_{20} be 20 **IID** Bernoulli(0.5) random variables.

- Since X_i s are independent: $\mathrm{Cov}(X_i, X_j) = 0$ for all i, j.
- Since X_i is Bernoulli(p = 0.5): $E[X_i] = p = 0.5,\ \mathrm{Var}(X_i) = p(1\text{-}p) = 0.25.$

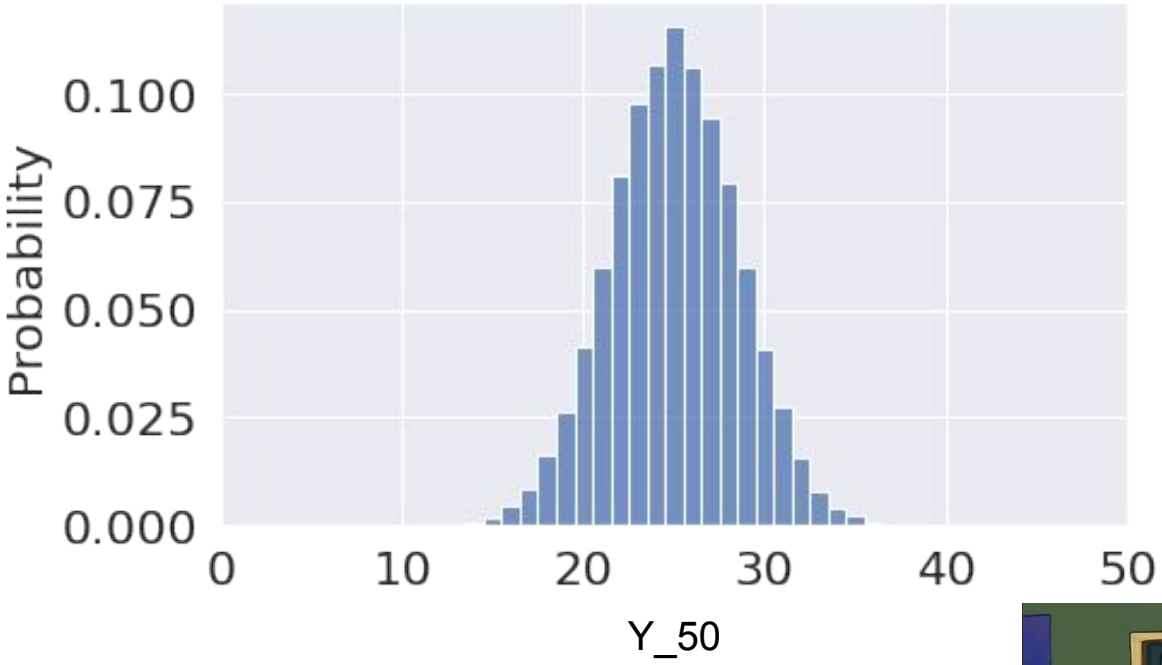
Which payout strategy would you choose?

	A. $Y_A = \$10 \cdot X_1 + \$10 \cdot X_2$	**B.** $Y_B = \$\left(\sum_{i=1}^{20} X_i\right)$	**C.** $Y_C = \$20 \cdot X_1$
Expectation	$E[Y_A] = 10(0.5) + 10(0.5) = 10$	$E[Y_B] = 0.5 + \ldots + 0.5 = 10$	$E[Y_C] = 20(0.5) = 10$
Variance	$\mathrm{Var}(Y_A) = 10^2(0.25) + 10^2(0.25) = 50$	$\mathrm{Var}(Y_B) = 0.25 + \ldots + 0.25 = 20(0.25) = 5$	$\mathrm{Var}(Y_C) = 20^2(0.25) = 100$
Std. Deviation	$\mathrm{SD}(Y_A) \approx \mathbf{7.07}$	$\mathrm{SD}(Y_B) \approx \mathbf{2.24}$	$\mathrm{SD}(Y_C) = \mathbf{10}$

Binomial(n, p) for Large n

3678970

For p = 0.5, n = 50 (i.e. number of heads in 50 fair coin flips):



Sample Statistics

Lecture 17, Data 100 Spring 2023

3678970

Sums of Random Variables

- Properties of Expectation and Variance
- Covariance, Correlation

Bernoulli and Binomial Random Variables

Sample Statistics

- **Sample Mean**
- **Central Limit Theorem**

Prediction vs. Inference

- Modeling: Assumptions of Randomness

The Bias-Variance Tradeoff

[Extra] Derivations

[Extra] Review of the Bootstrap

[Extra] Derivation of Bias-Variance Decomposition

From Populations to Samples

3678970

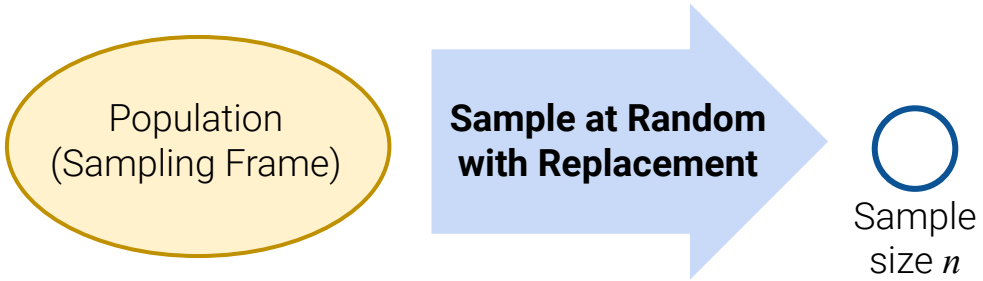
Today, we've talked extensively about **populations**:

- 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

3678970

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
...	...

Population
(really large 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

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

Sample
(n << N)

$X_1, X_2, \ldots, X_n$

The Sample is a Set of IID Random Variables

3678970

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
...	...

$E[X] = 5.9$

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

μ

Sample at Random with Replacement

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

Sample size n

	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 $X_1, X_2, \ldots, X_n$

The Central Limit Theorem

3678970

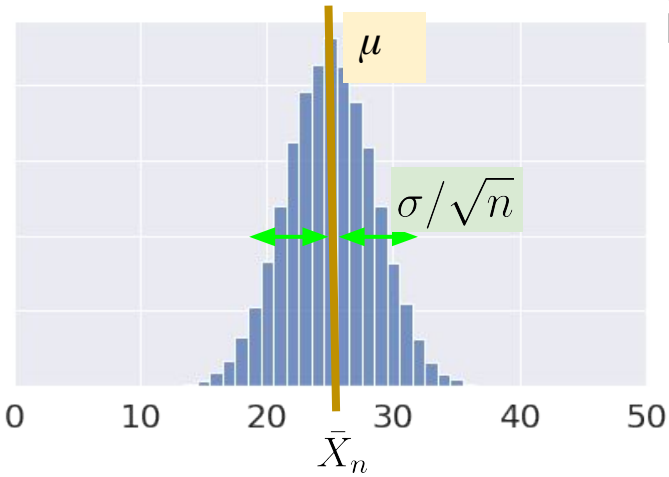
No matter what population you are drawing from:

If an IID sample of size n is large,

the probability distribution of the **sample mean**

is **roughly normal** with mean μ and SD $\sigma/\sqrt{n}$.

(STAT 140/EECS 126) (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/

[Terminology] Sample Mean

3678970

Consider an IID sample $X_1, X_2, \ldots, X_n$ drawn from a 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:

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

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

Distribution?

$\bar{X}_n$ is **normally distributed** by the **Central Limit Theorem**.

How Large Is “Large”?

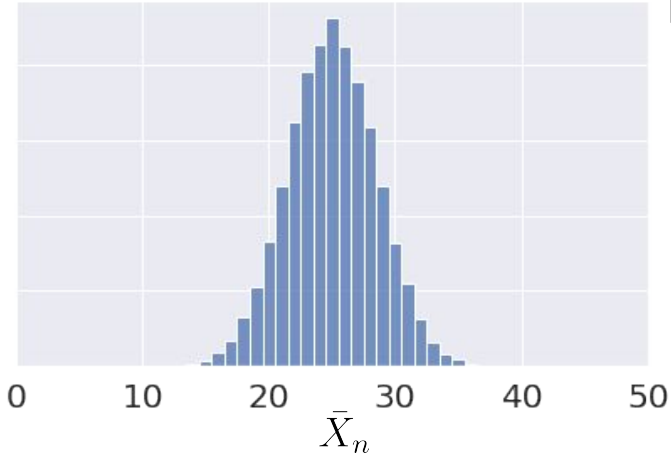
3678970

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 $\sigma/\sqrt{n}$.



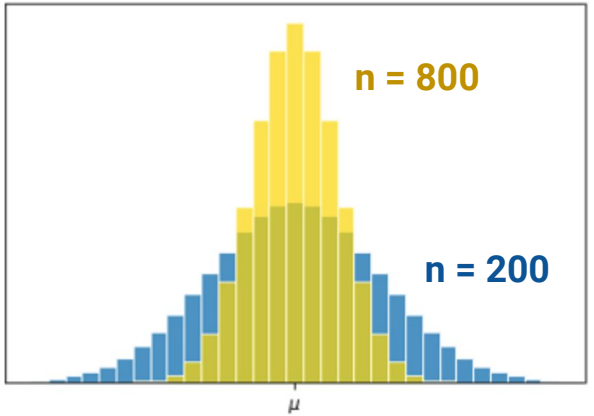
How large does n have to be for the normal approximation to be good?

- …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.

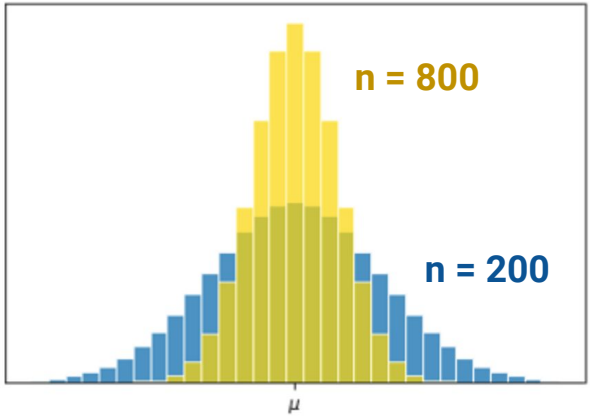
$$\mathbb{E}[\bar{X}_n] = \mu$$

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

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 how it scales with the sample size n.

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

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

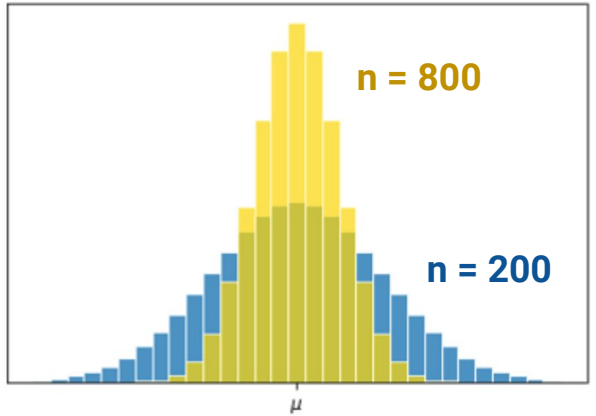
For every sample size, the expected value of the sample mean is the population mean.

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 how it scales with the sample size n.

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

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

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.

slido

Sample mean and sample standard deviation are unbiased estimators of population mean and population standard deviation, respectively.

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

Prediction vs. Inference

Lecture 17, Data 100 Spring 2023

3678970

Prediction vs. Inference

3678970

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

3678970

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 more are houses with river views worth (holding other variables fixed)?

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

[Terminology] Parameters, Statistics, and Estimators

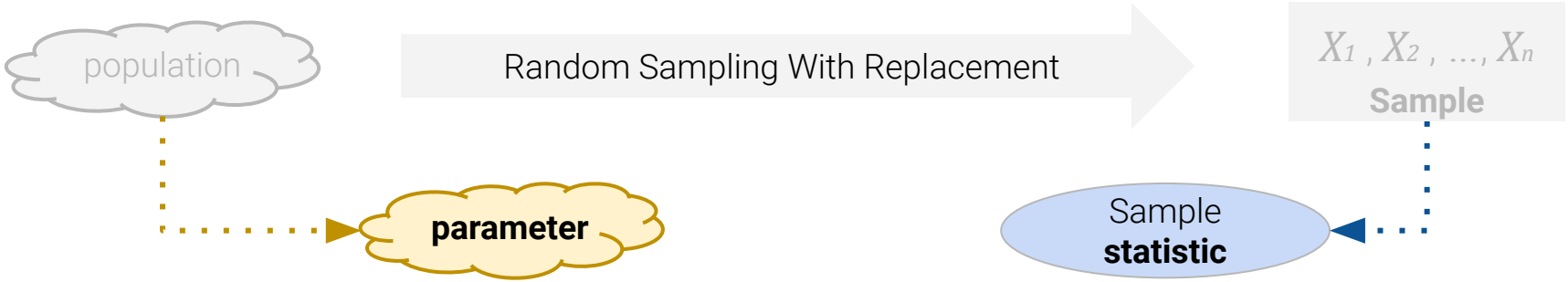
3678970

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

[Terminology] Parameters, Statistics, and Estimators

3678970

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



A **parameter** is a numerical function of the **population**.

Population mean

$$\mu$$

Sample mean

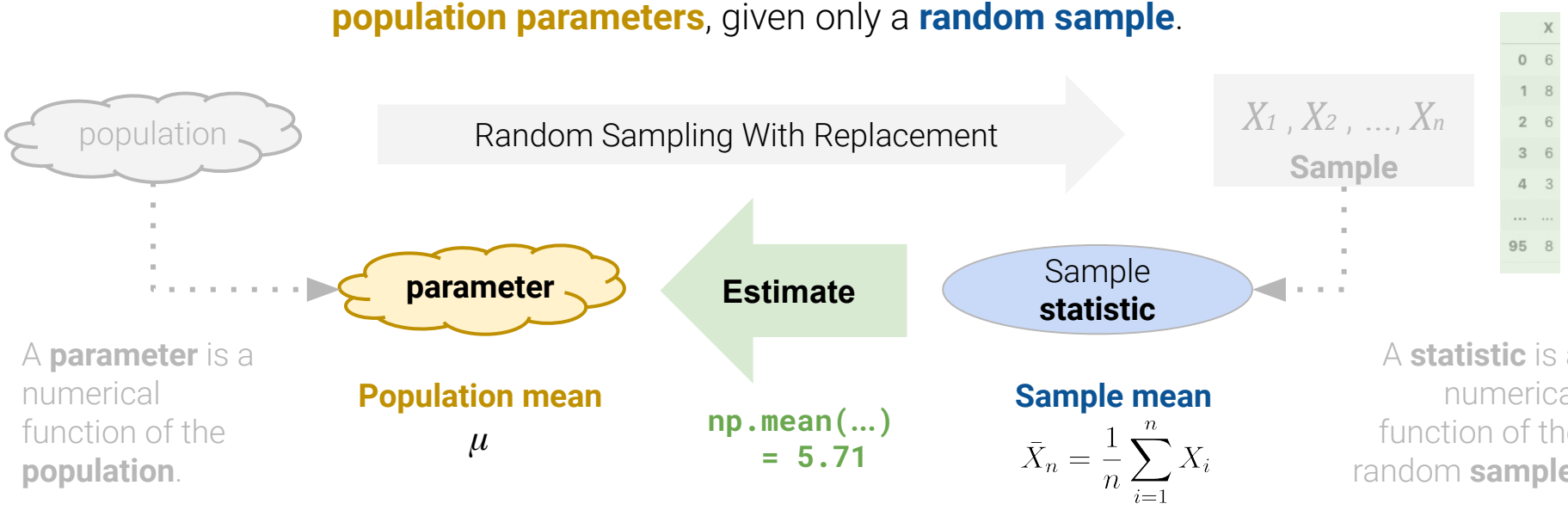
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

A **statistic** is a numerical function of the random **sample**.

[Terminology] Parameters, Statistics, and Estimators

3678970

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



	X
0	6
1	8
2	6
3	6
4	3
...	...
95	8

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

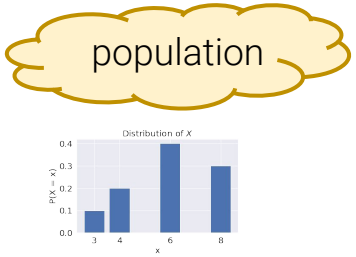
Since our **sample is random**, our statistic (which we use as our estimator) could have been different.

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

3678970

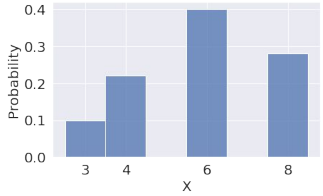
One View: Randomly draw a random sample, then compute the statistic for that sample.



$X_1, X_2, \ldots, X_n$

Sample

	X
0	6
1	8
2	6
3	6
4	3
...	...
95	8

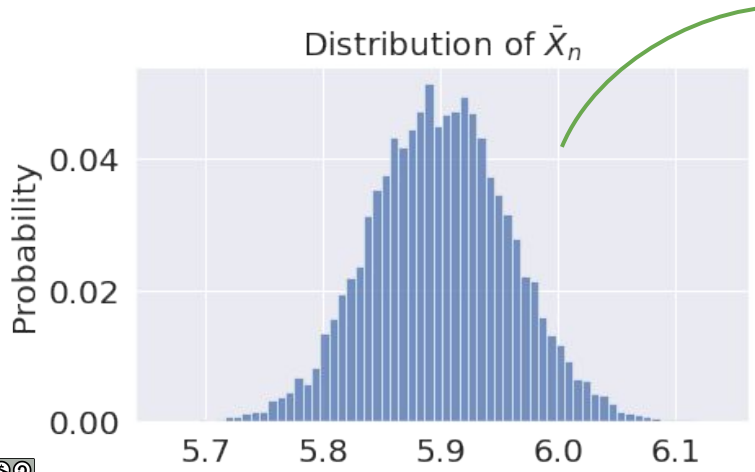


Sample mean

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

`np.mean(...) = 5.82`

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



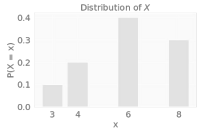
5.82

We often use the 2nd view because we want to interpret the estimator directly, and not the random sample.

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

3678970

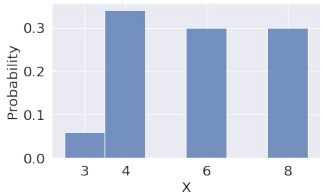
population



$X_1, X_2, \ldots, X_n$

Sample

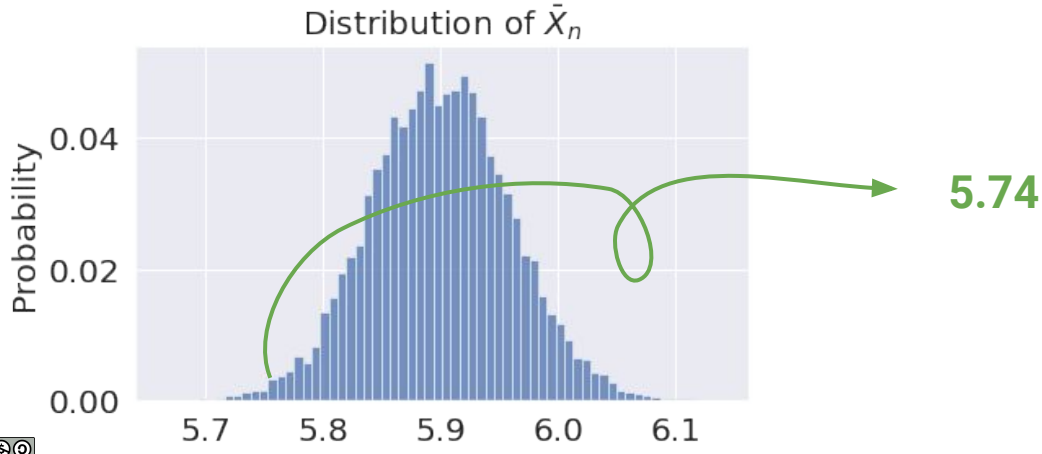
	X
0	6
1	6
2	6
3	4
4	6
...	...
95	8



Sample mean

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

`np.mean(...)` = **5.74**



5.74

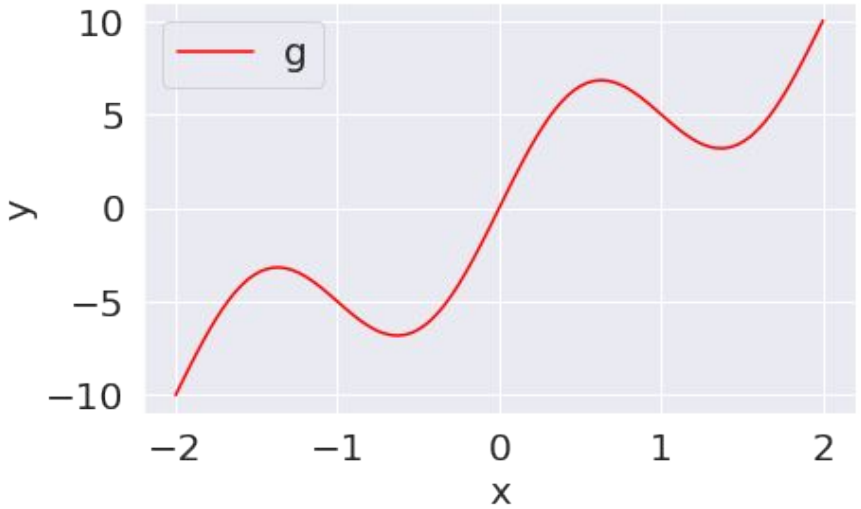
The value of our estimator is a function of the random sample. The estimator is therefore also random.

Modeling: Estimating a Relationship

3678970

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:

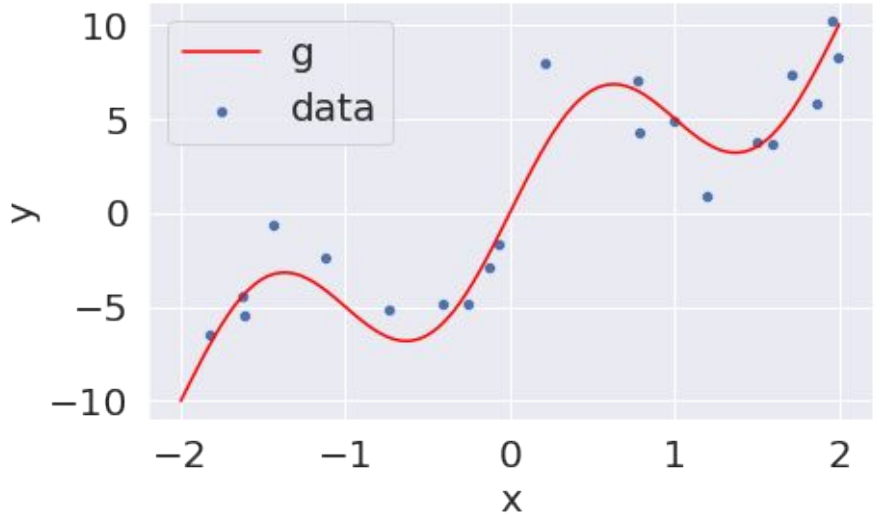
- **Fixed features** x, and hence fixed $g(x)$.

Modeling: Estimating a Relationship

3678970

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

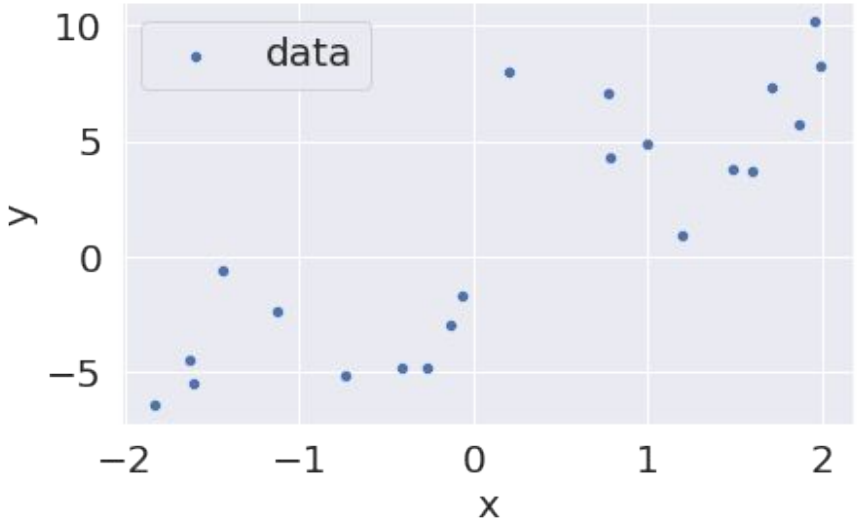
Errors ϵ are assumed expectation 0 ("zero mean") and i.i.d. across individuals.

Modeling: Estimating a Relationship

3678970

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

3678970

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

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

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

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 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 $Y = f_\theta(x)$.

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

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

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

- 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

3678970

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 $\hat{\theta}$.

We Have Finally Formalized the Distinction Between Estimators and Parameters

3678970

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

Hats mean estimates.

Which Expressions Are Random?

3678970

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

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

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

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

- 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 $\hat{\theta}$.

Which Expressions Are Random?

Random in blue

3678970

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

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

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

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

- 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 $\hat{\theta}$.

The Bias-Variance Tradeoff

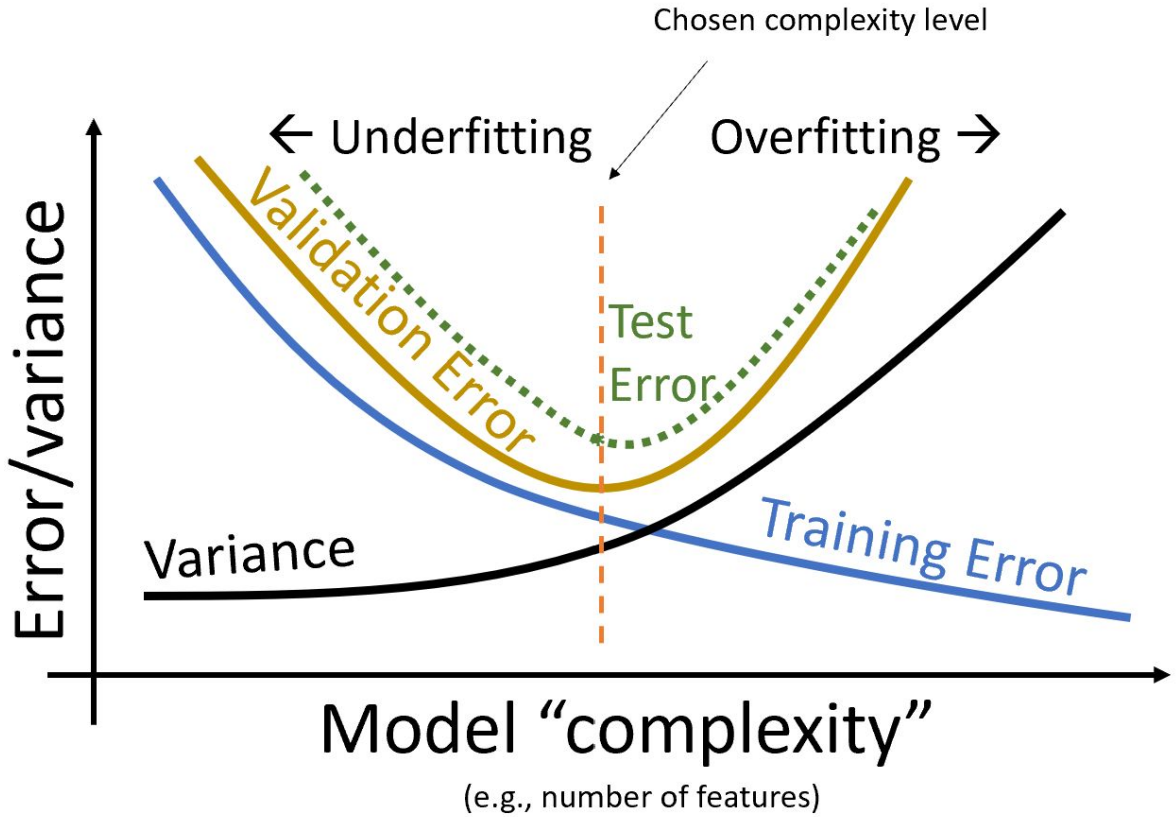
Lecture 17, Data 100 Spring 2023

3678970

Prediction: The Bias-Variance Tradeoff

3678970

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

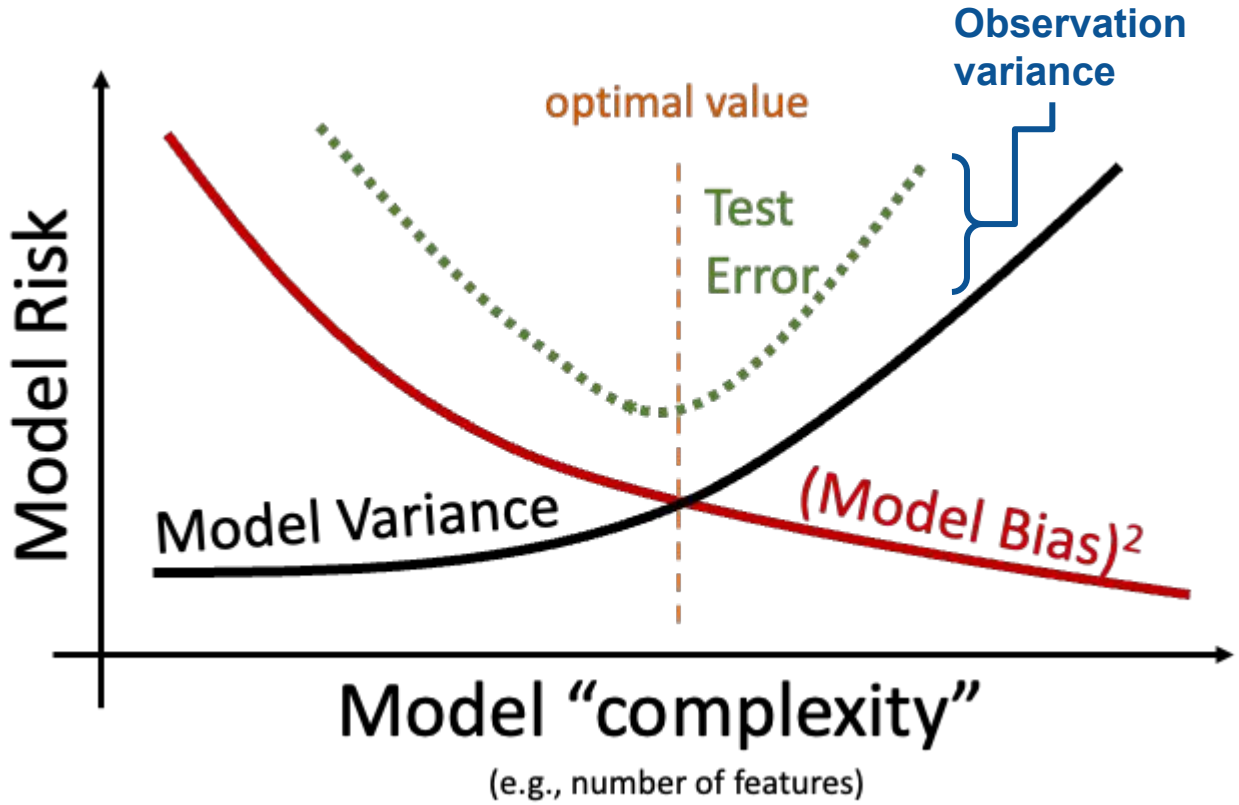


Covered up to here on 14/3.
Will cover the rest of the slides in lecture 20 on 23/3

The Bias-Variance Tradeoff

3678970

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

This Entire Section is Theoretical

3678970

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

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 $\{(x_1, y_1), \ldots, (y_n, y_n)\}$

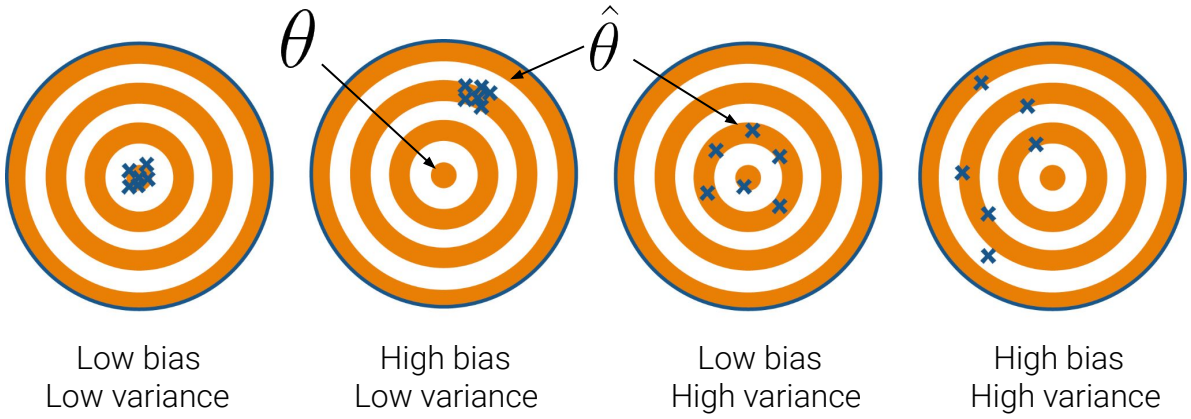
Performance of an Estimator

3678970

Suppose we want to estimate a population parameter θ using an estimator $\hat{\theta}$.

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 θ? **(Risk / MSE)**

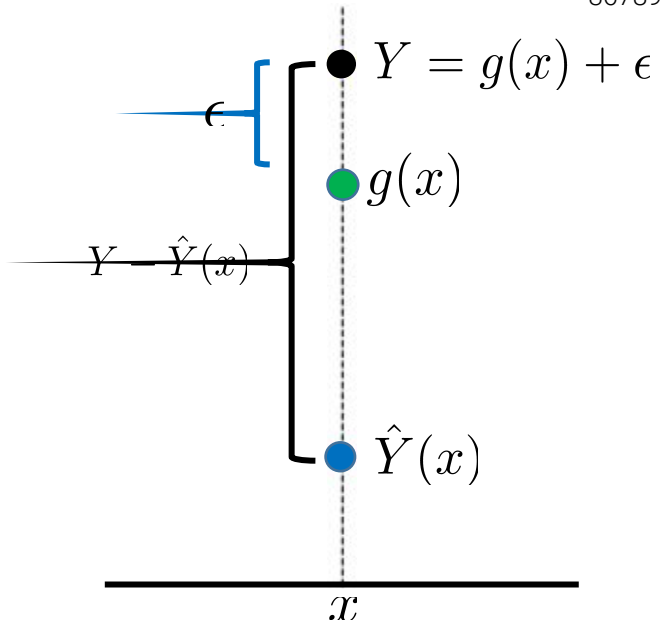


The Three Components of Model Risk

3678970

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.



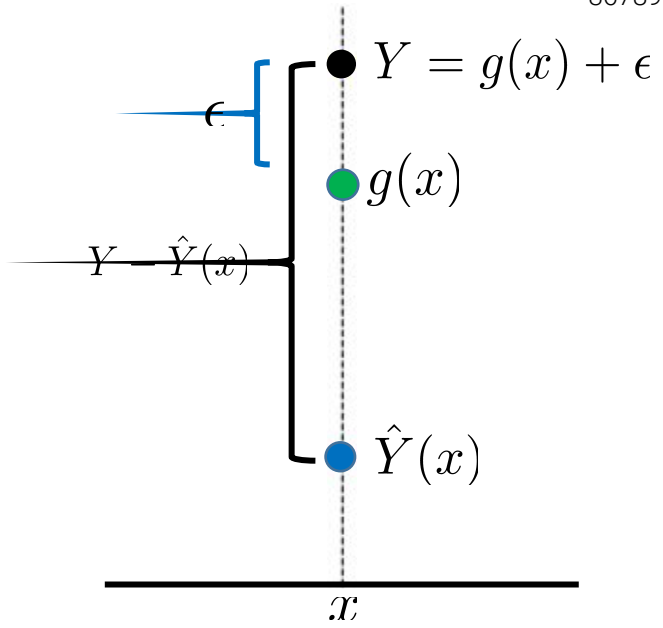
How do you think each of the types of error are encoded into the diagram?

The Three Components of Model Risk

3678970

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.

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

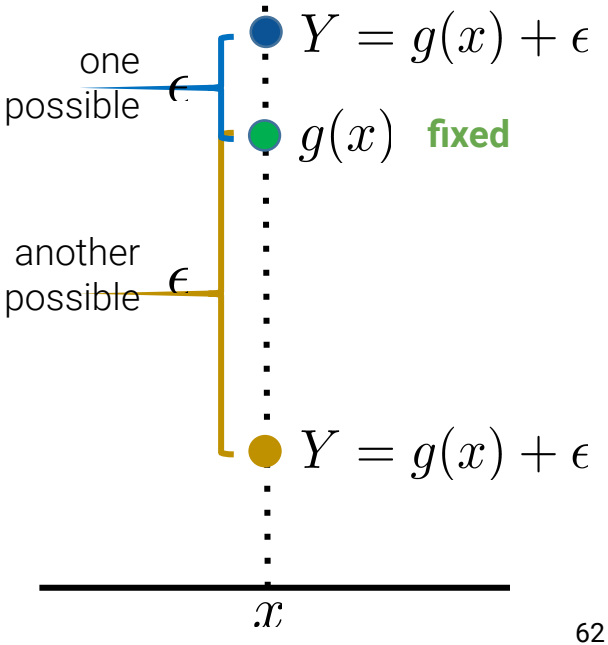
1. Observation Variance

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.

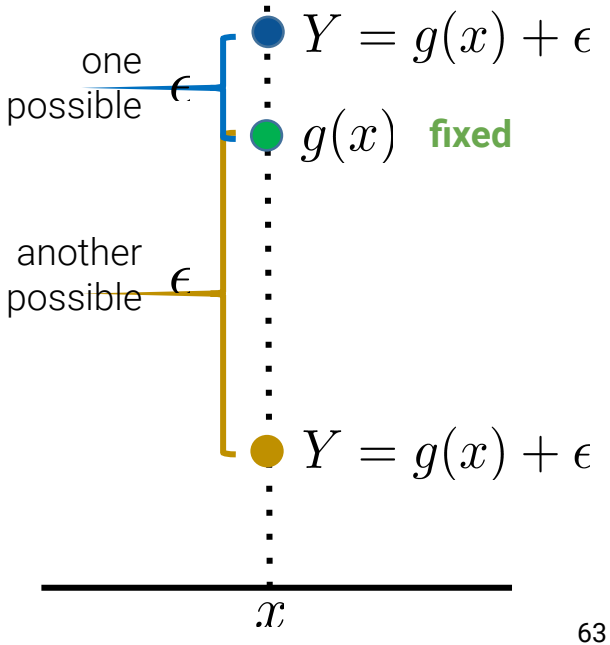
We assume **random error** ϵ to have zero mean and variance σ^2.

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

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.

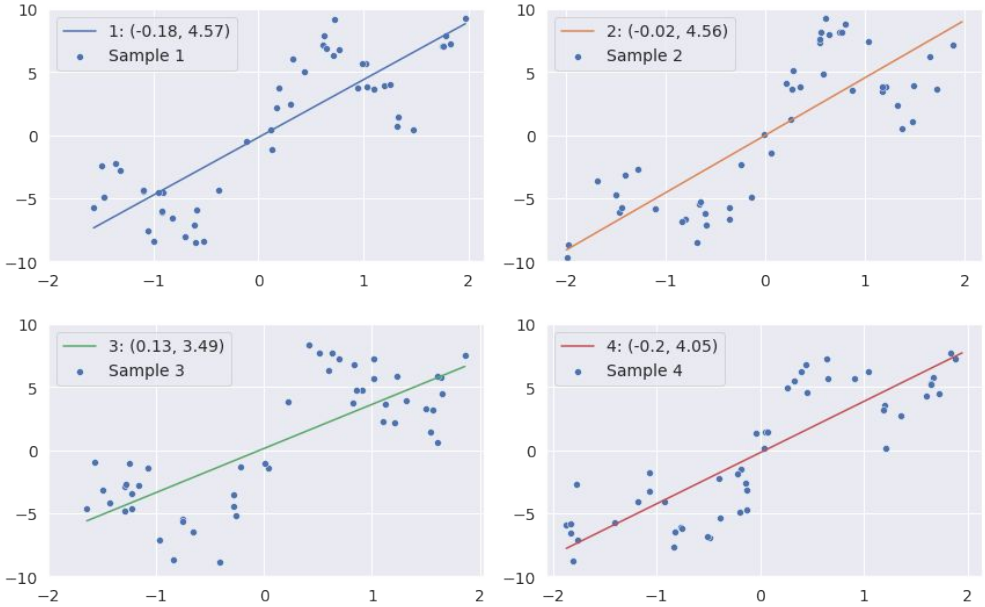
2. Model Variance

3678970

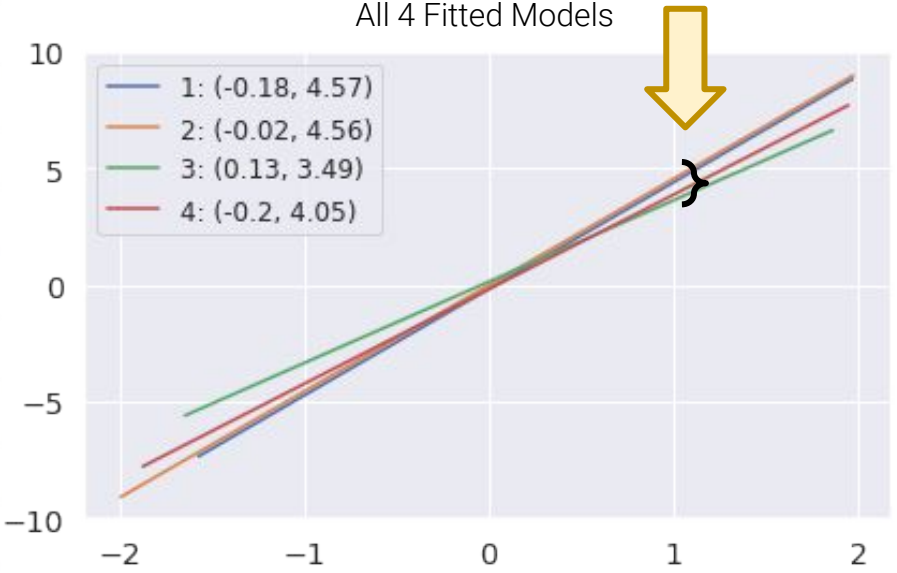
Our fitted model is based on a **random sample**.

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



All 4 Fitted Models



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

2. Model Variance

3678970

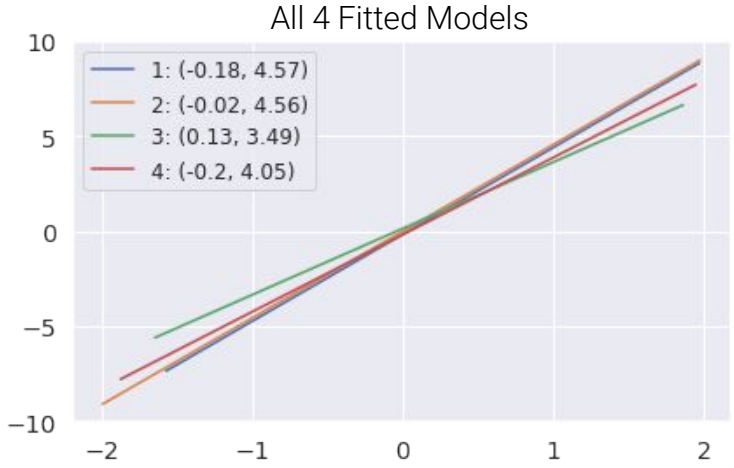
Our fitted model is based on a **random sample**.

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:

- Different samples → different model estimates
- **Overfitting**. Small differences in random samples lead to large differences in the fitted model

Remedy:

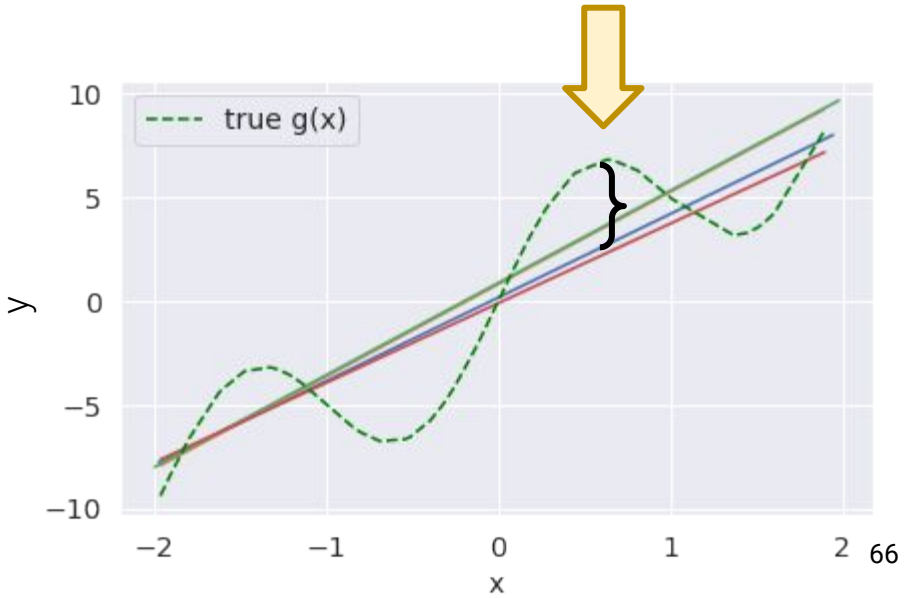
- Reduce model complexity
- Don't fit the noise

3. Model Bias

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 $g(x)$.

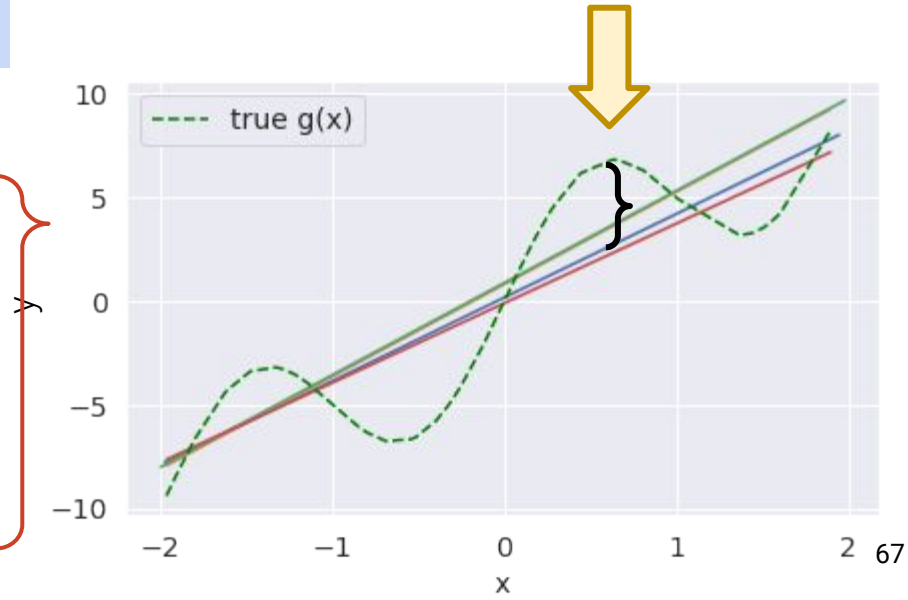
$\hat{Y}(x)$ Prediction for the individual at x
A random variable

- 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

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

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

- 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.

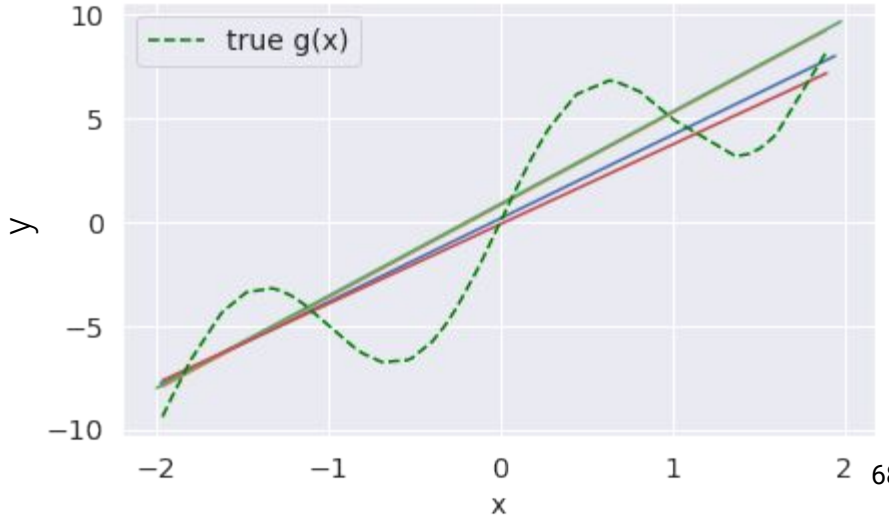
$$\text{model bias} = \mathbb{E}[\hat{Y}(x)] - g(x)$$

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

3678970

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

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

The Bias-Variance Decomposition: Prediction

3678970

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

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

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

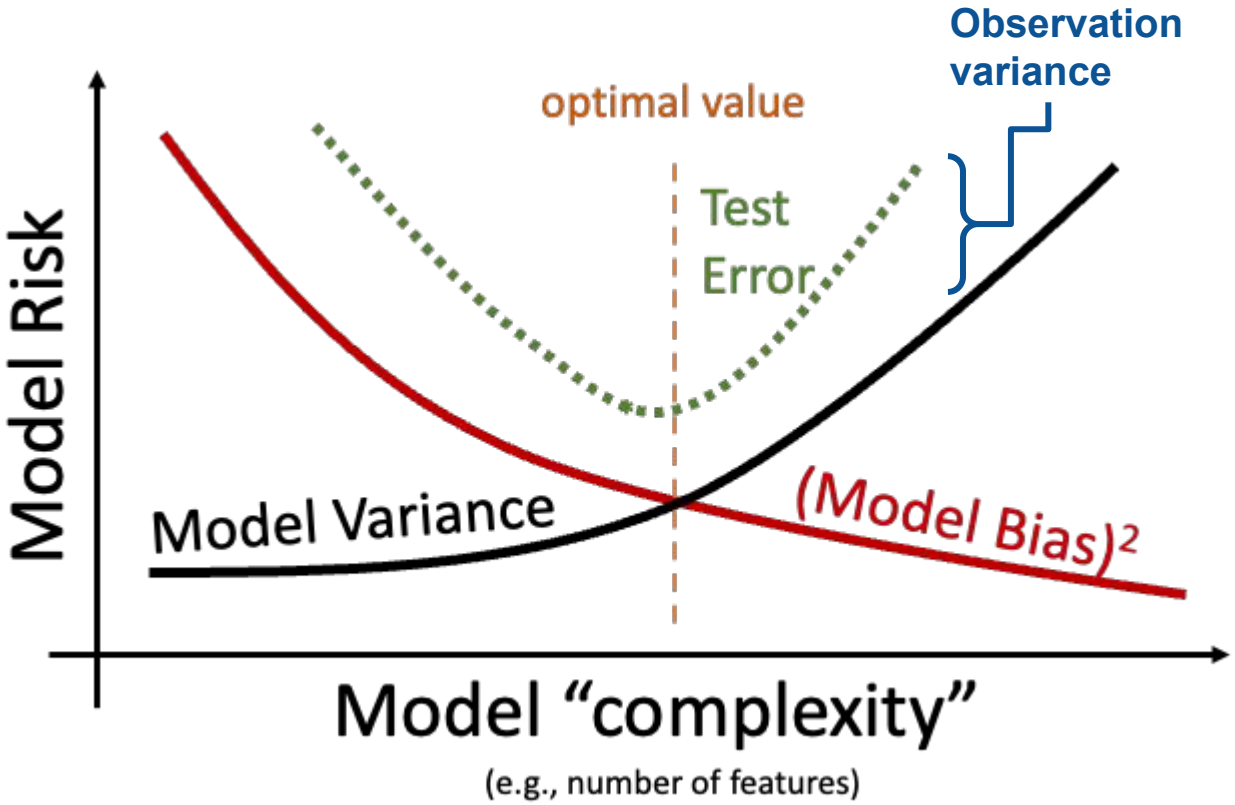
Interpret:

- 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!

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



Have a Normal Day!

3678970

[Extra] Derivations

Lecture 17, Data 100 Spring 2023

Standardization of random variables

Jump back: link

3678970

X in **standard units** is the random variable $X_{su} = \frac{X - \mathbb{E}(X)}{\mathbb{SD}(X)}$.

X_{su} measures X on the scale **"number of SDs from expectation**."

- It is a linear transformation of X. By the linear transformation rules for expectation and variance:

$$\mathbb{E}(X_{su}) = 0, \quad \mathbb{SD}(X_{su}) = 1$$

- Since X_{su} is centered (has expectation 0):

$$\mathbb{E}(X_{su}^2) = \mathbb{V}ar(X_{su}) = 1$$

You should prove these facts yourself.

Variance: An Alternate Derivation

Jump back: link

3678970

There's a more convenient form of variance for use in calculations.

$$\mathbb{V}ar(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

To derive this, we make repeated use of the linearity of expectation.

$$\begin{aligned} Var(X) &= E((X - E(X))^2) \\ &= E\big(X^2 - 2XE(X) + (E(X)^2)\big) \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

Properties of Expectation #1

Jump back: link

3678970

Recall definition of expectation: $\mathbb{E}[X] = \sum_x xP(X = x)$

$$\mathbb{E}[X] = \sum_{\substack{\text{all samples} \\ s}} X(s)P(s)$$

1. **Expectation is linear**:

(intuition: summations are linear)

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Proof:

$$\begin{aligned}\mathbb{E}[aX + b] &= \sum_x (ax + b)P(X = x) = \sum_x \big(axP(X = x) + bP(X = x)\big) \\ &= a\sum_x xP(X = x) + b\sum_x P(X = x) \\ &= a\mathbb{E}[X] + b \cdot 1\end{aligned}$$

Properties of Expectation #2

Jump back: link

Recall definitions of expectation: $\mathbb{E}[X] = \sum_x xP(X = x)$

$$\mathbb{E}[X] = \sum_{\substack{\text{all samples} \\ s}} X(s)P(s)$$

3. **Expectation is linear in sums of RVs**:

For any relationship between X and Y.

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Proof:

$$\mathbb{E}[X+Y] = \sum_s (X+Y)(s)P(s) = \sum_s \big(X(s)+Y(s)\big)P(s)$$

$$= \sum_s \big(X(s)P(s) + Y(s)P(s)\big)$$

$$= \sum_s X(s)\,P(s) + \sum_s Y(s)P(s)$$

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

Properties of Variance #1

Jump back: link

3678970

We know that $\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$

In order to compute Var(aX + b), consider:

- A shift by **b** units **does not** affect spread. Thus, Var(aX + b) = Var(aX).
- The multiplication by **a does** affect spread!

Then,

$$\begin{aligned} Var(aX+b) = Var(aX) &= E((aX)^2) - (E(aX))^2 \\ &= E(a^2X^2) - (aE(X))^2 \\ &= a^2\left(E(X^2) - (E(X))^2\right) \\ &= a^2 Var(X) \end{aligned}$$

In summary:

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$$

$$\mathbb{V}ar(aX+b) = a^2\mathbb{V}ar(X)$$

$$\mathbb{SD}(aX+b) = |a|\mathbb{SD}(X)$$

Don't forget the absolute values and squares!

Properties of Variance #2

Jump back: link

3678970

The variance of a sum is affected by the dependence between the two random variables that are being added. Let's expand out the definition of Var(X + Y) to see what's going on.

Let $\mu_x = E[X], \mu_y = E[Y]$

$$
\begin{aligned}
Var(X+Y) &= E\big[(X+Y-E(X+Y))^2\big] \\
&= E\big[((X-\mu_x)+(Y-\mu_y))^2\big] \\
&= E\big[(X-\mu_x)^2 + 2(X-\mu_x)(Y-\mu_y) + (Y-\mu_y)^2\big] \\
&= E\big[(X-\mu_x)^2\big] + E\big[(Y-\mu_y)^2\big] + 2E\big[(X-\mu_x)(Y-\mu_y)\big] \\
&= Var(X) + Var(Y) + 2E\big[(X-E(X))(Y-E(Y))\big]
\end{aligned}
$$

By the linearity of expectation, and the substitution.

We see

Addition rule for variance

3678970

If X and Y are **uncorrelated** (in particular, if they are **independent**), then

$$\mathbb{V}ar(X+Y) = \mathbb{V}ar(X) + \mathbb{V}ar(Y)$$

Therefore, under the same conditions,

$$\mathbb{SD}(X+Y) = \sqrt{\mathbb{V}ar(X) + \mathbb{V}ar(Y)} = \sqrt{(\mathbb{SD}(X))^2 + (\mathbb{SD}(Y))^2}$$

- Think of this as “Pythagorean theorem” for random variables.
- Uncorrelated random variables are like orthogonal vectors.

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 2023

Bootstrap Resampling

3678970

- 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.

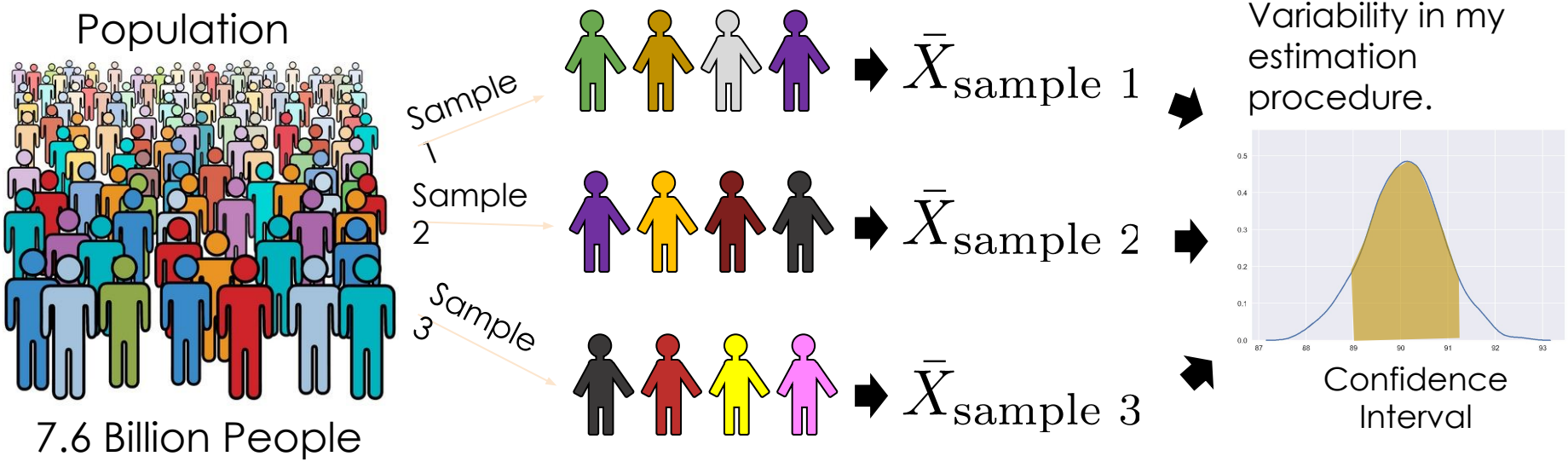
Idea: Treat our random sample as a "population", and resample from it.

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

The Distribution of an Estimator

3678970

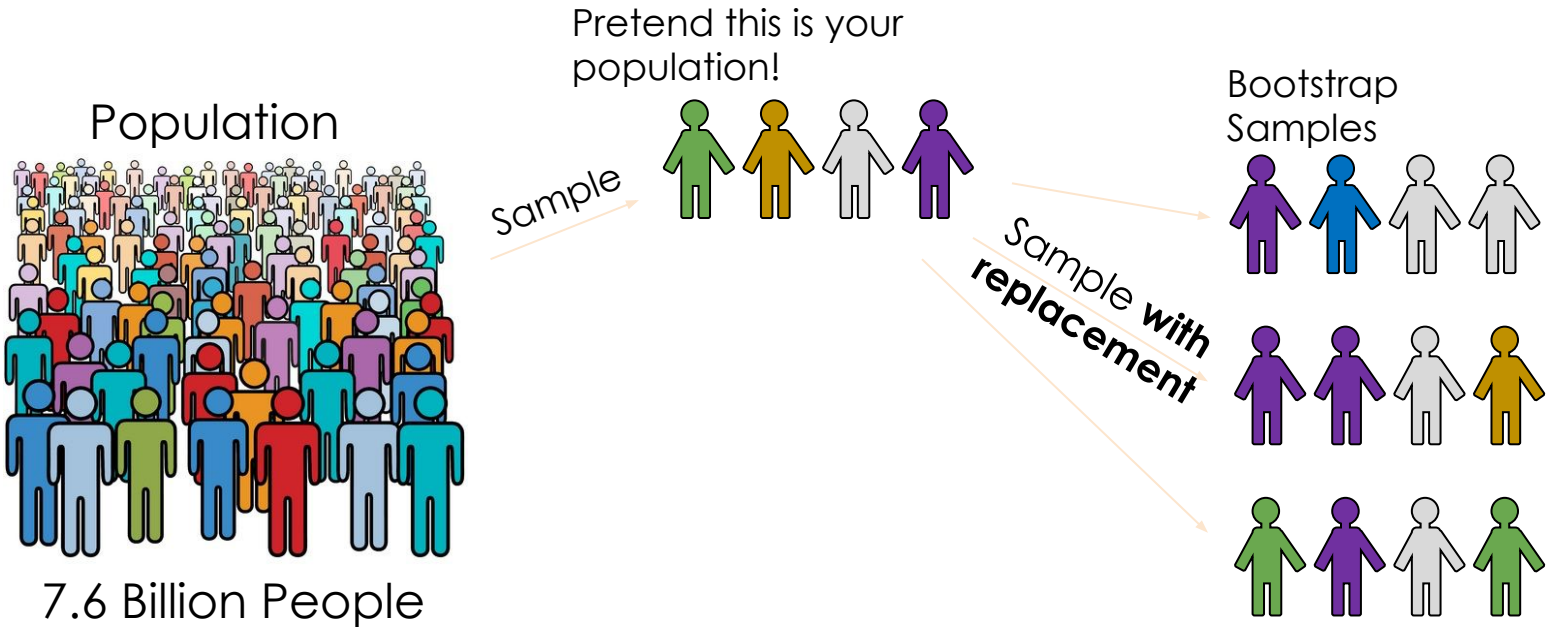
Resampling the population to estimate the sample distribution.



Bootstrap the Distribution of an Estimator

Simulation method to estimate the sample distribution.

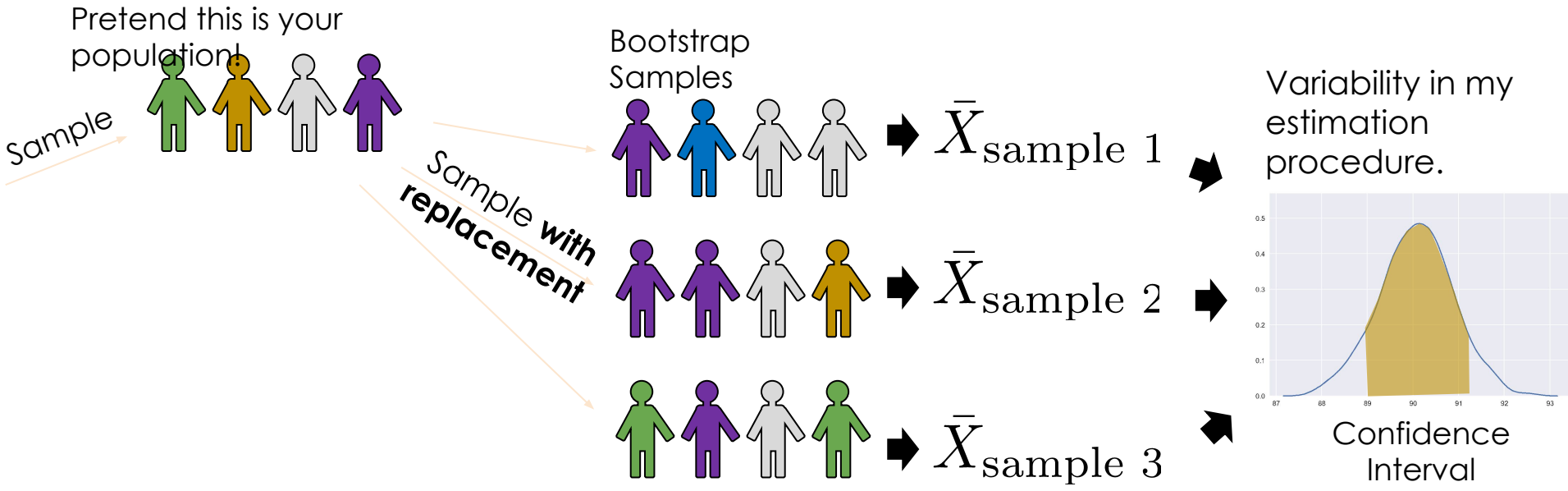
3678970



Bootstrap the Distribution of an Estimator

3678970

Simulation method to estimate the sample distribution.

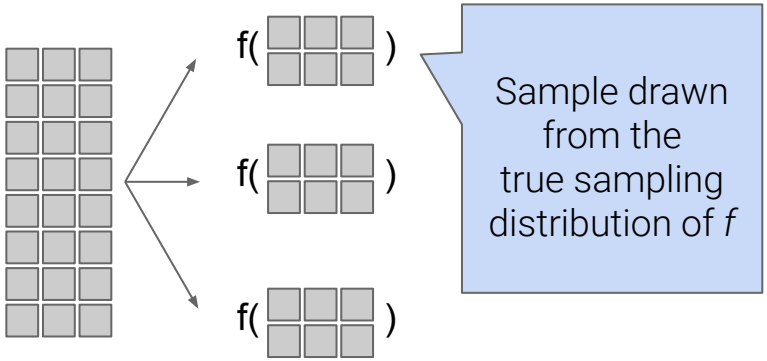


Bootstrap Resampling

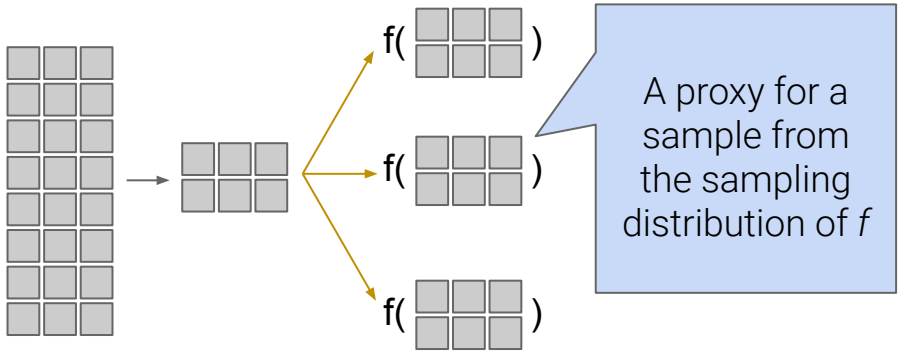
3678970

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

Impractical:



Bootstrap:



(demo)

Bootstrapping Pseudocode

3678970

```
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

3678970

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).

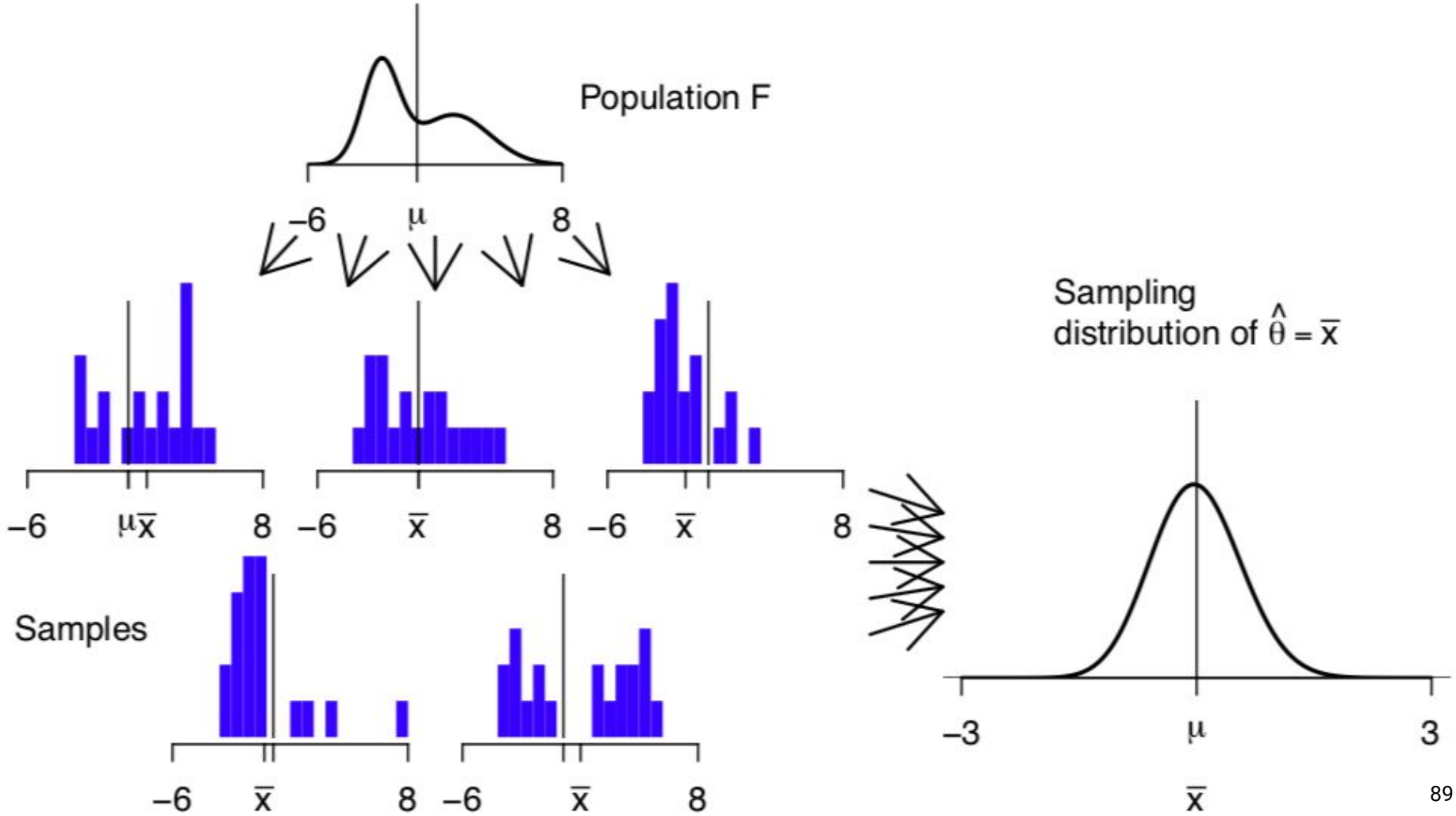
The center of the bootstrapped distribution is the estimator applied to our original sample.

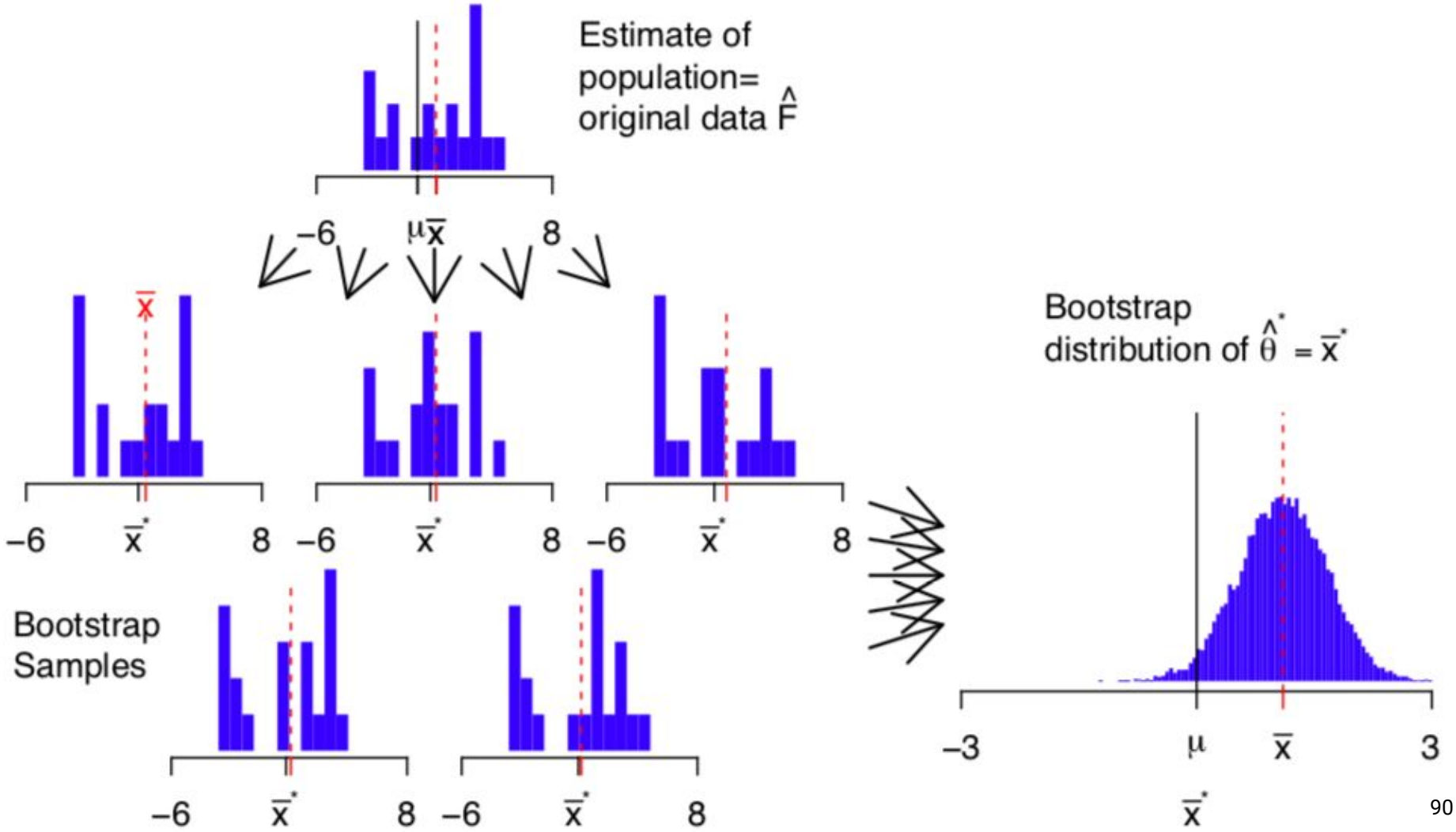
- We have no way of recovering the estimator's true expected value.

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
−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
−6
$\bar{x}^*$
8
−6
$\bar{x}^*$
8
Bootstrap
Samples
−6
$\bar{x}^*$
8
−6
$\bar{x}^*$
8
Bootstrap
distribution of $\hat{\theta}^* = \bar{x}^*$
−3
μ
$\bar{x}$
3
$\bar{x}^*$

What Teachers Should Know About the Bootstrap

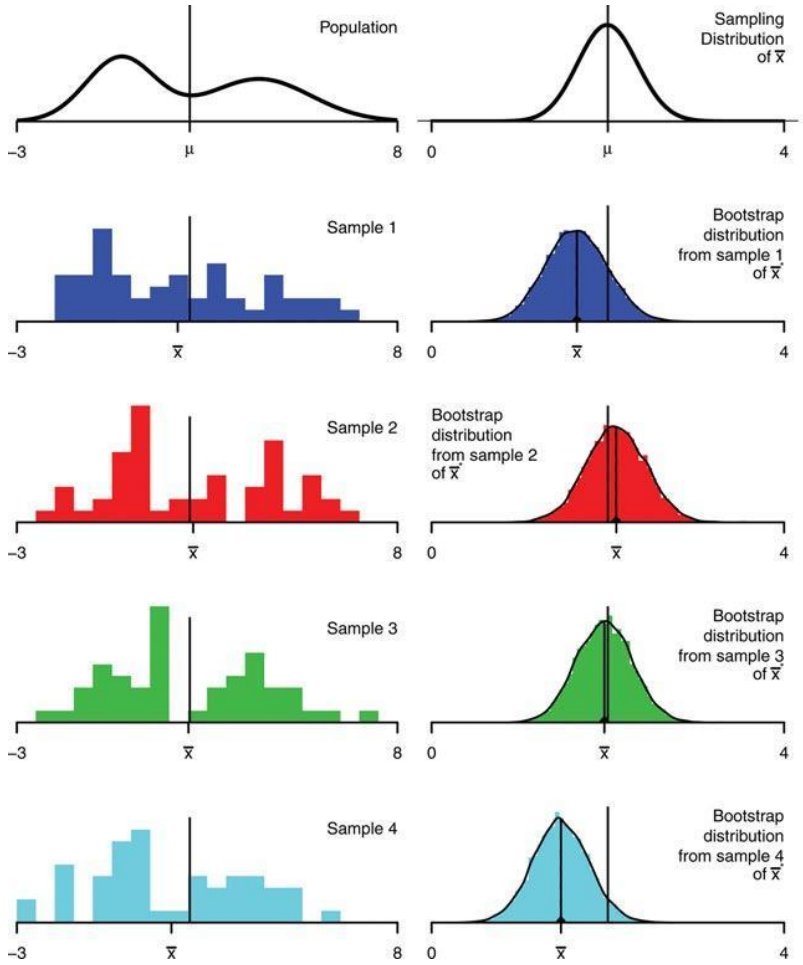
3678970

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.

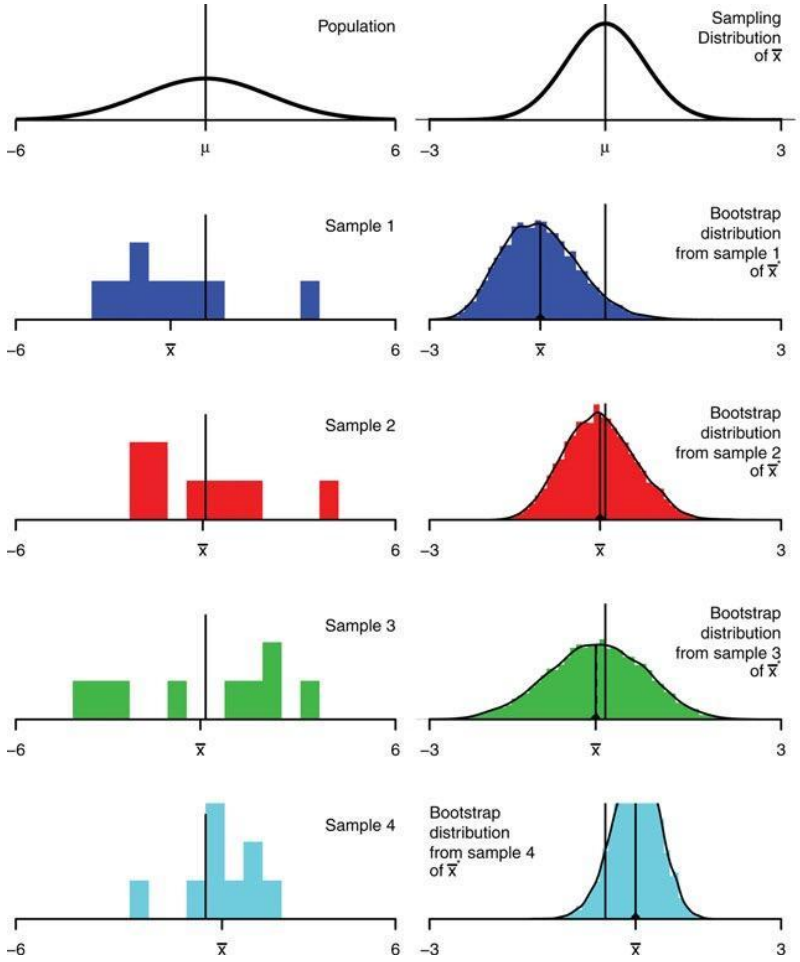
Tim C. Hesterberg (2015)

Bootstrap for the Mean, n=50

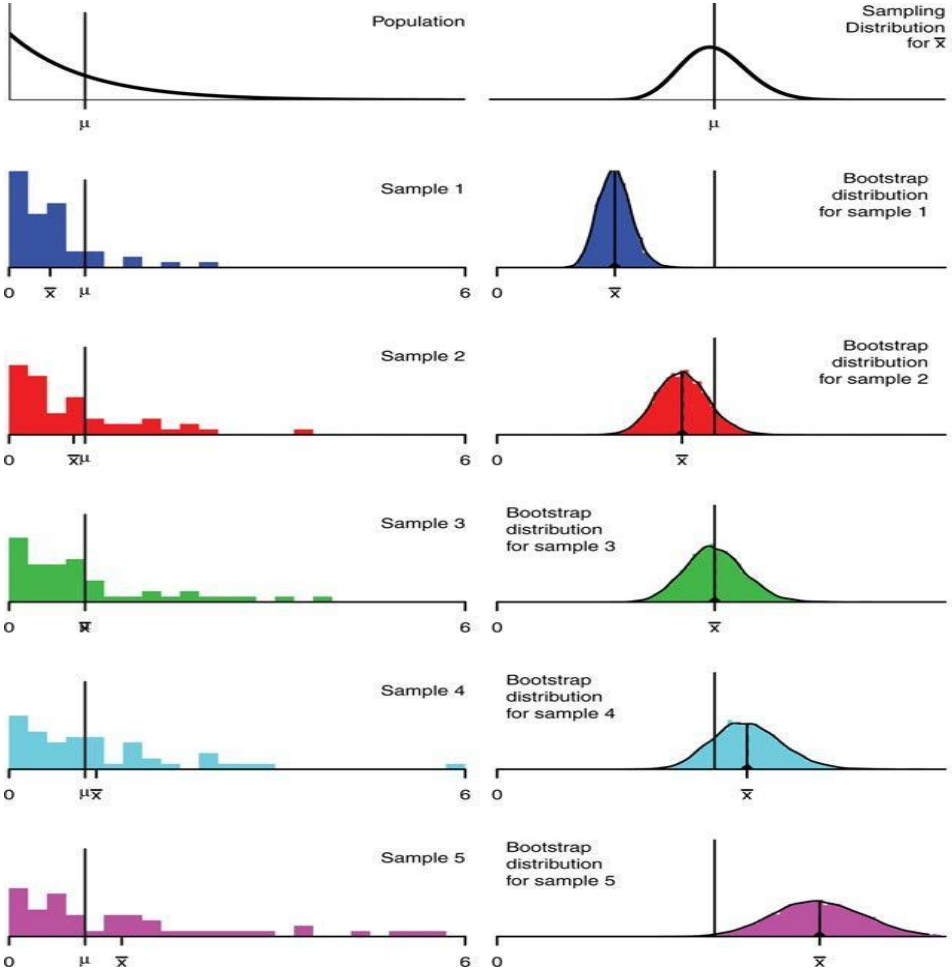


From Tim C. Hesterberg (2015)

Bootstrap Distributions for the Mean, n = 9



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



Some More Lessons from Hesterberg

3678970

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 2023

3678970

Derivation: Preliminary

3678970

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.

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

3678970

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

3678970

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 17

Estimators, Bias, and Variance

Narges Norouzi and Lisa Yan

Content credit: Acknowledgments