

# Copyright Notice

These slides are distributed under the Creative Commons License.

[DeepLearning.ai](#) makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite [DeepLearning.AI](#) as the source of the slides.

For the rest of the details of the license, see <https://creativecommons.org/licenses/by-sa/2.0/legalcode>



DeepLearning.AI

# Probability and Statistics for Machine Learning and Data Science

---

## Week 3: Sampling and Point Estimates

# W3 Lesson 1



DeepLearning.AI

## Sample and Population

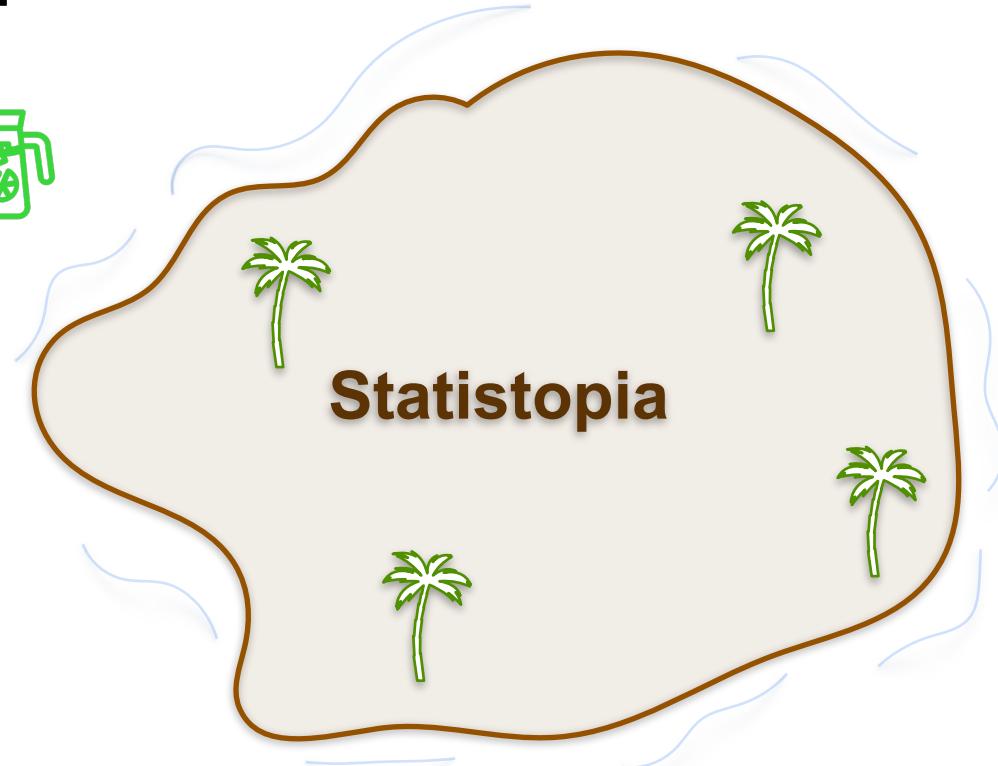
---

## Population and Sample

# Population and Sample



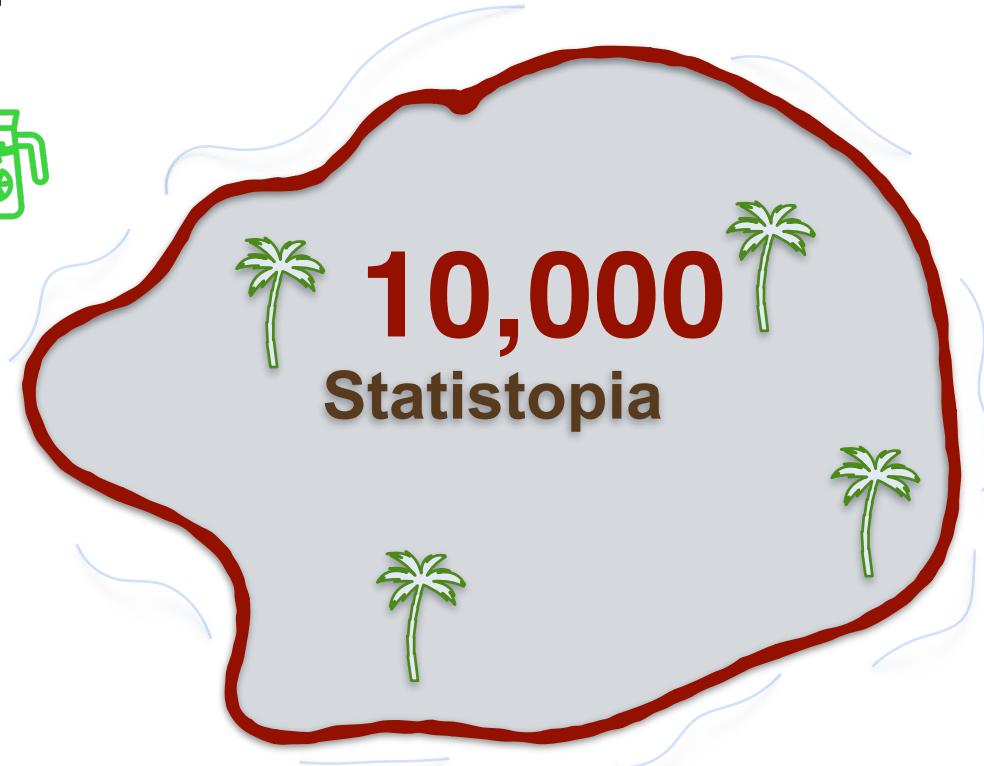
Find the **average height** of  
the people living on  
Statistopia



# Population and Sample



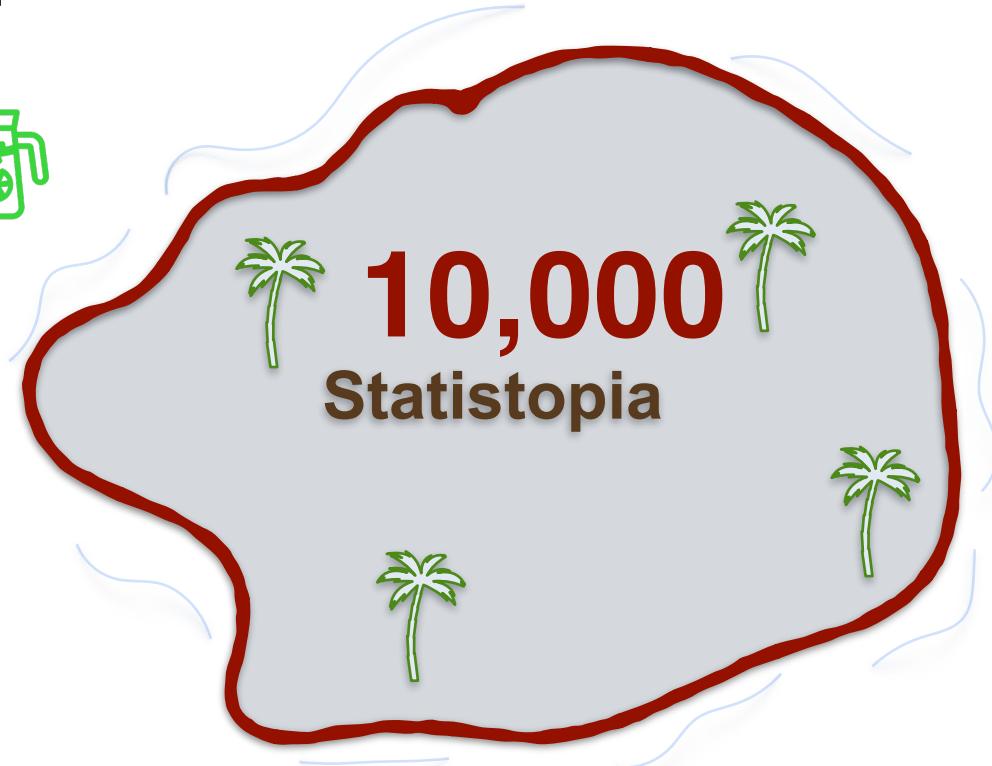
- Ask everyone on the island for their height.
- Divide by the total number



# Population and Sample



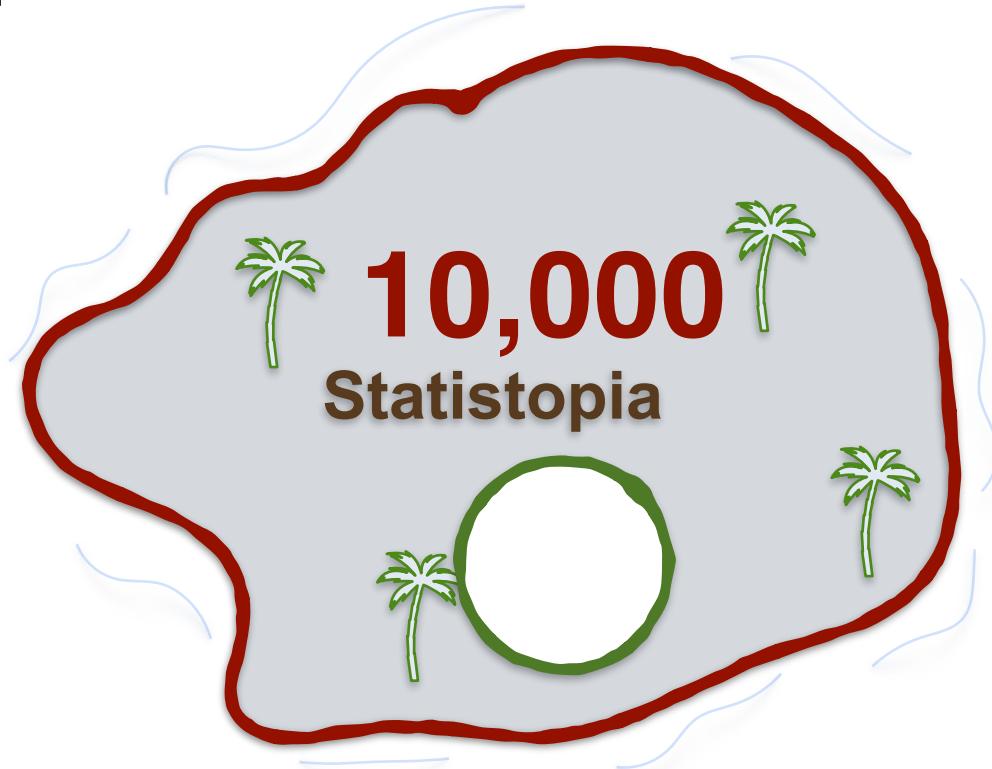
- Ask everyone on the island for their height.
- Divide by the total number



# Population and Sample



- Only ask a subset of the group to estimate the average height



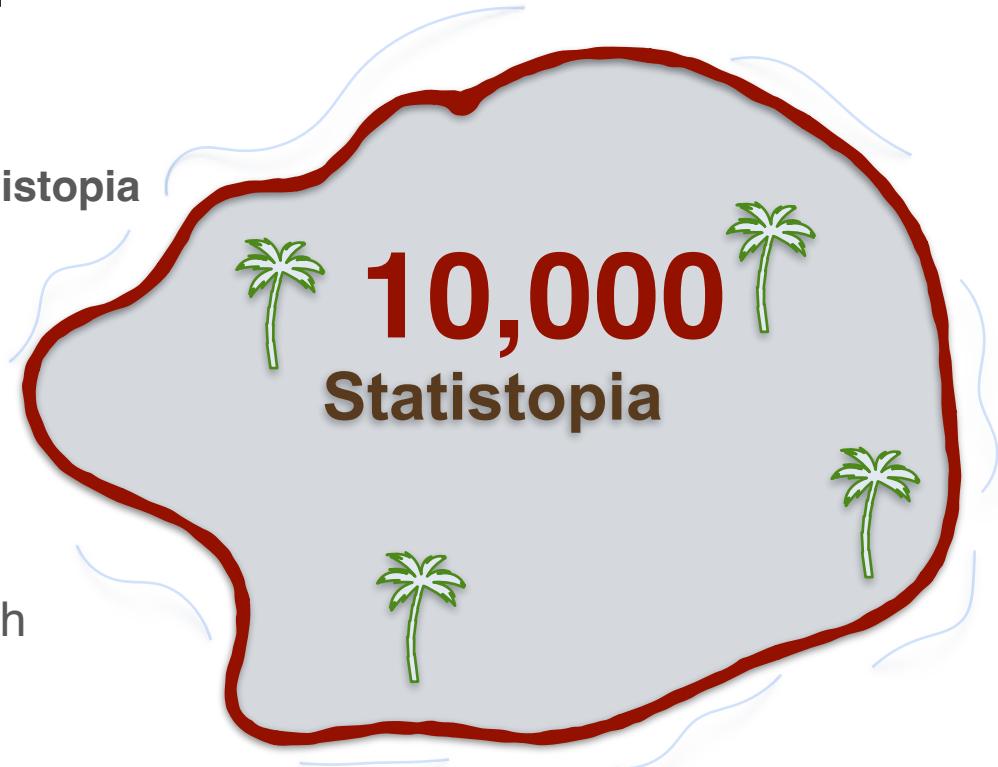
# Population and Sample



The people of statistopia

## Population:

the entire group of individuals or elements you want to study which share a common behaviour



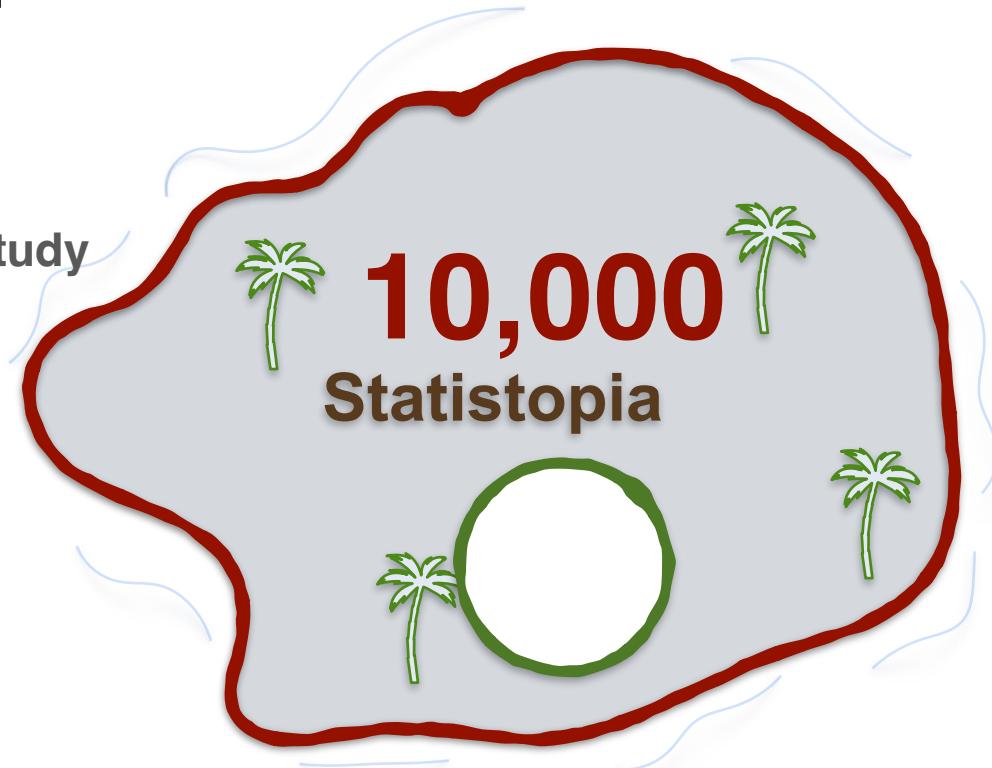
# Population and Sample



The people you  
select for your study

## Sample:

subset of the population you use  
to draw conclusions about the  
population as a whole



# Population and Sample

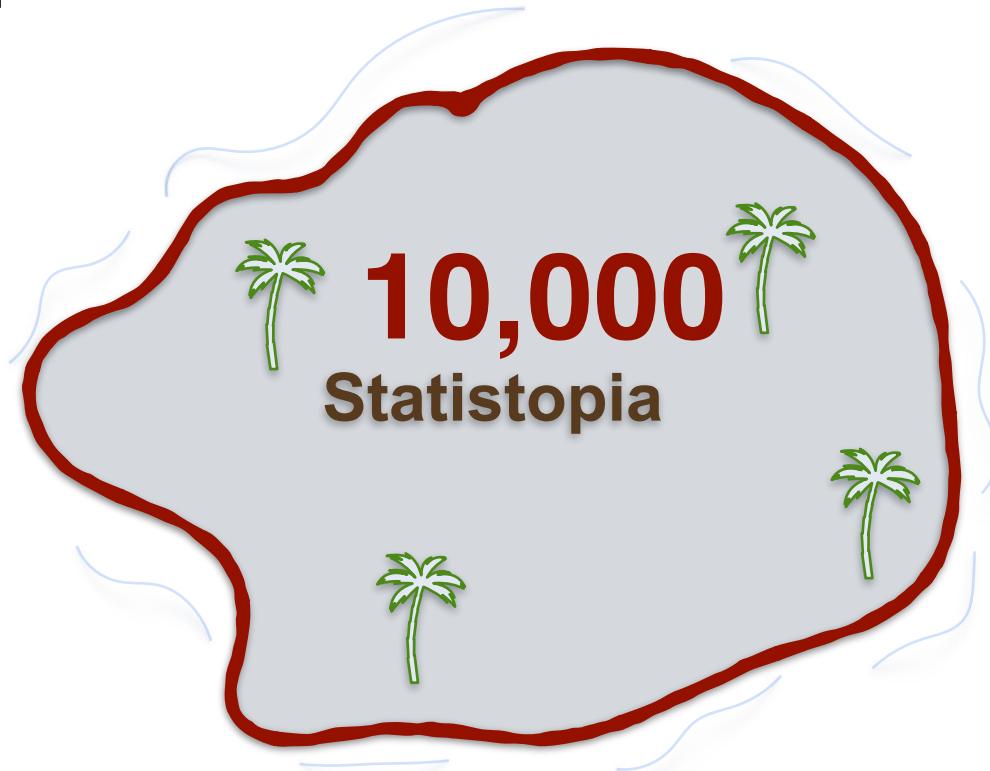


**Population Size (N)**

10,000

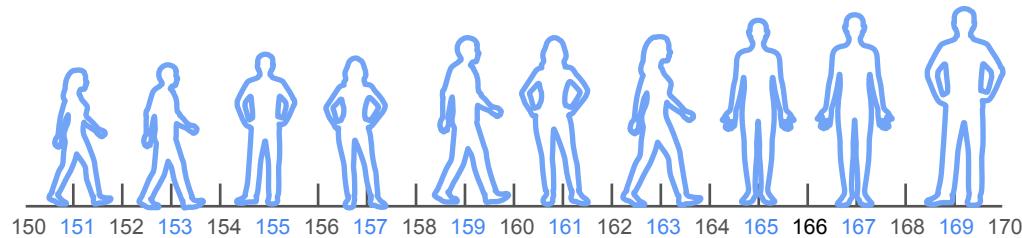
**Sample Size (n)**

1 - 9,999



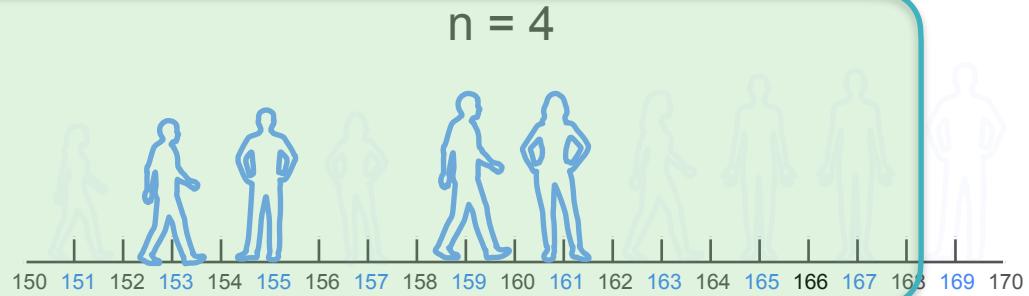
# Population and Sample

$N = 10$



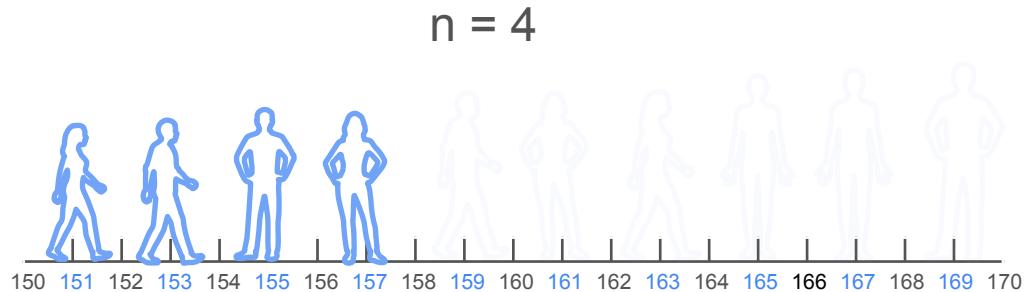
# Random Sampling

A



Which is the better sample  
to estimate the population  
mean height?

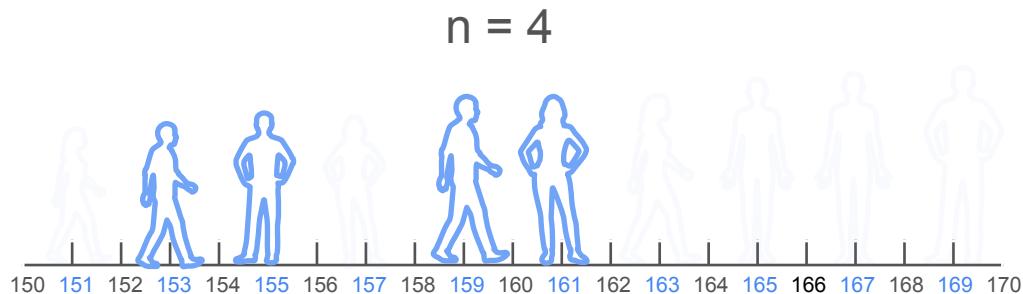
B



# Independent Sample

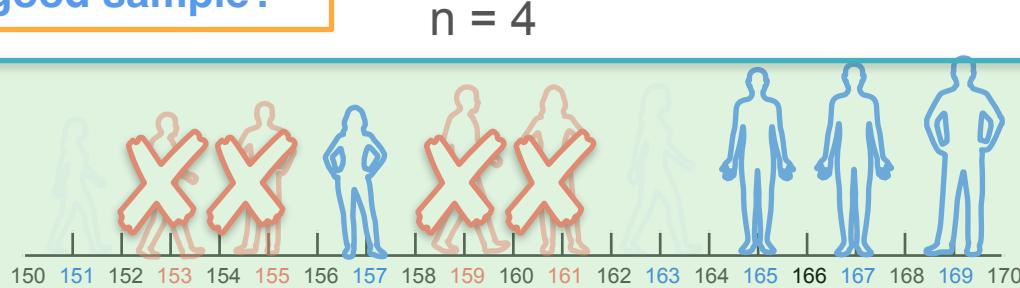
Example 1

1st sample set



Why is sample set two not a good sample?

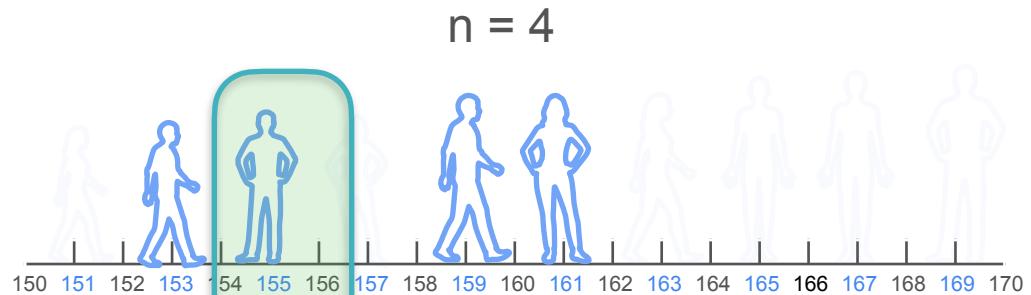
2nd sample set



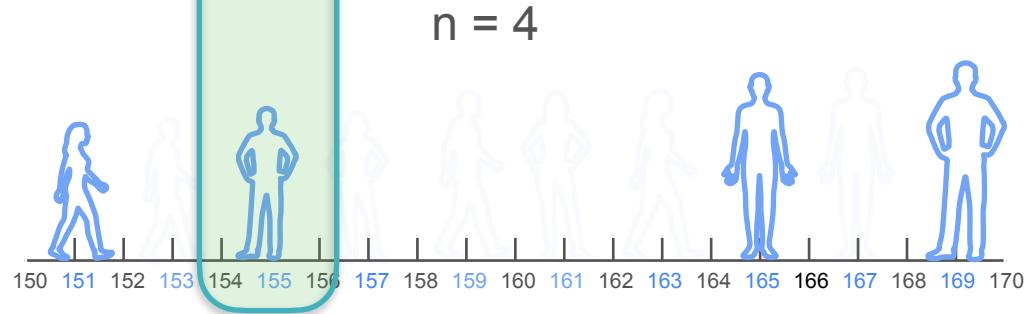
# Independent Sample

## Example 2

1st sample set



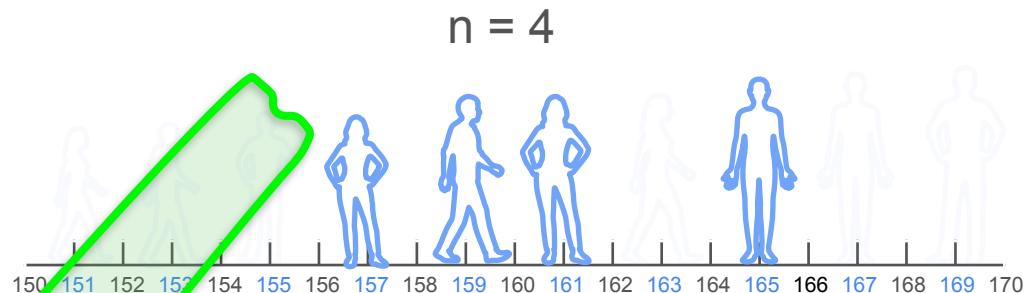
2nd sample set



# Identically Distributed Samples

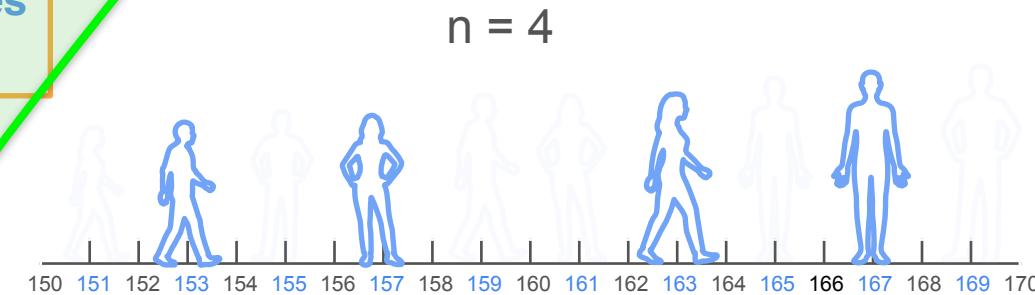
## Example 1

A



Which of the following samples  
are identically distributed?

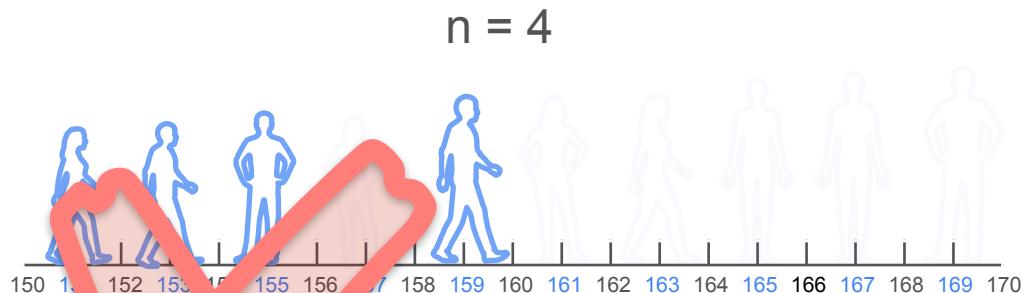
B



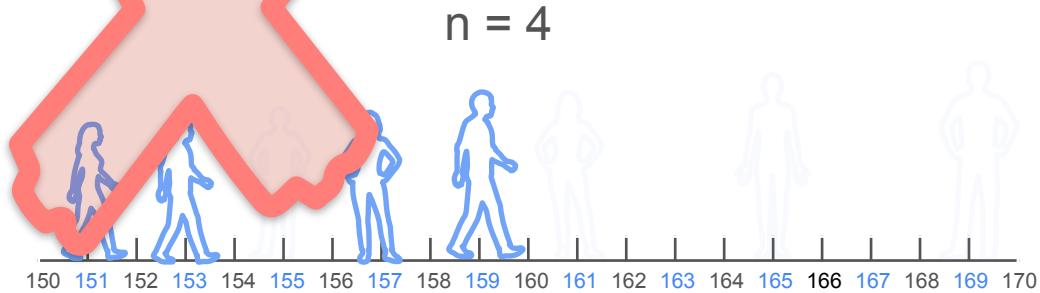
# Identically Distributed Samples

Example 2

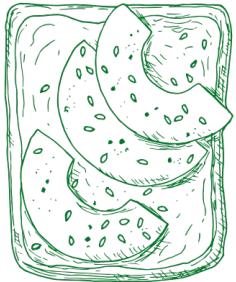
A



B



# The Avocado Toast Trend

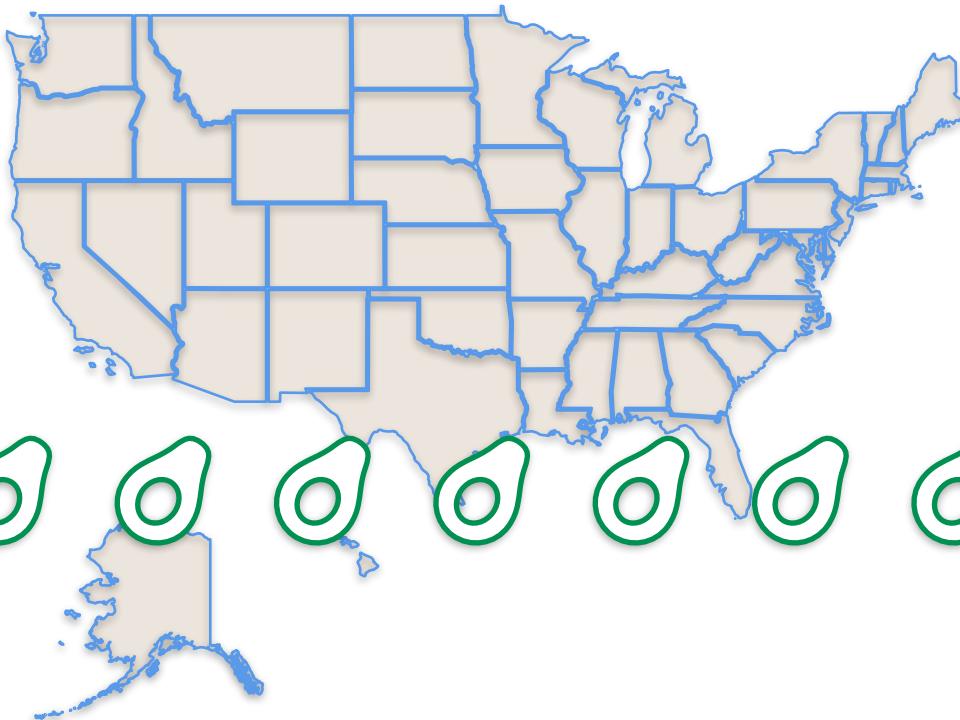


Study the price of avocados  
in the United States



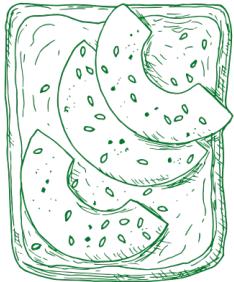
What is the population of your study?

# The Avocado Toast Trend



All avocados  
sold in the US

# The Avocado Toast Trend

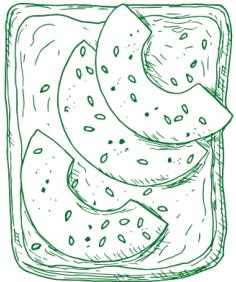


Study the price of avocados  
in the United States



What is the sample of your study?

# The Avocado Toast Trend



Study the price of avocados  
in the United States



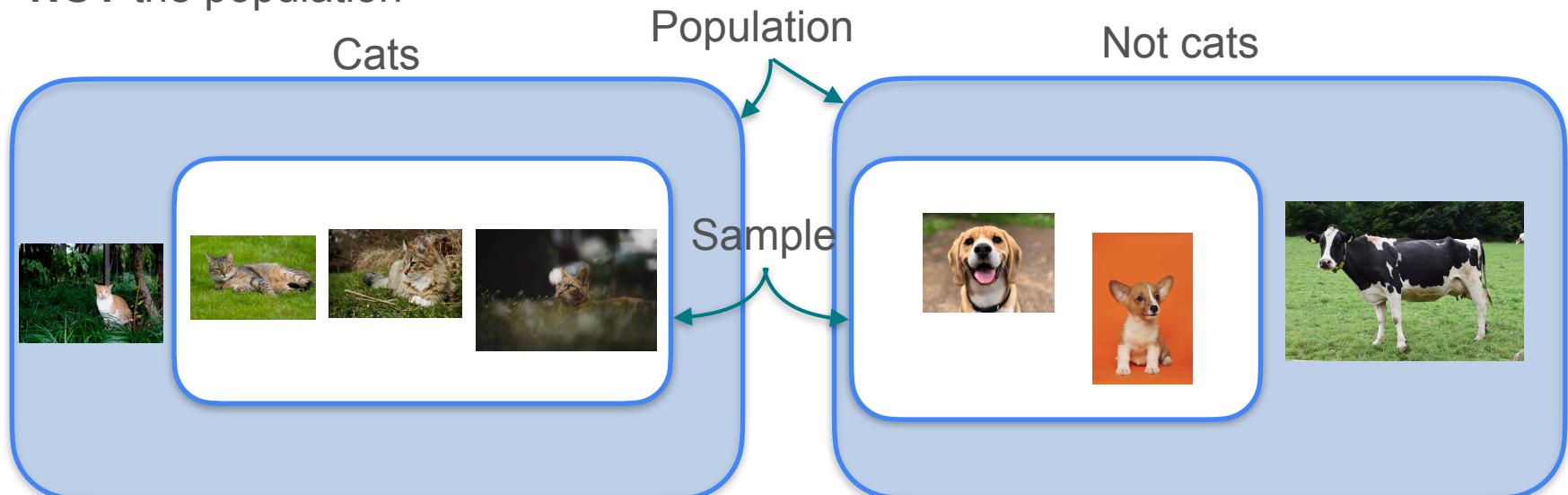
What is the sample of your study?

**Avocados sold  
in the 4 stores  
you selected**

# Population and Sample in Machine Learning

Every dataset you work with in machine learning is a sample

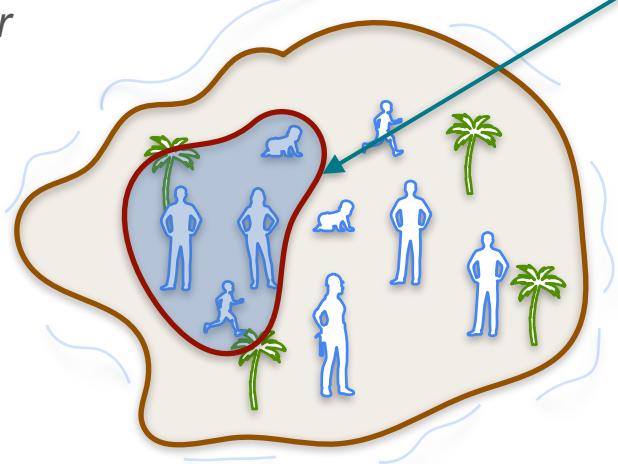
**NOT** the population



# Recap

## Population

*the entire group of individuals or elements you want to study which share a common behaviour*



## Sample

*subset of the population you use to draw conclusions about the population as a whole*

Population Size:

 $N$ 

Sample Size:

 $n$



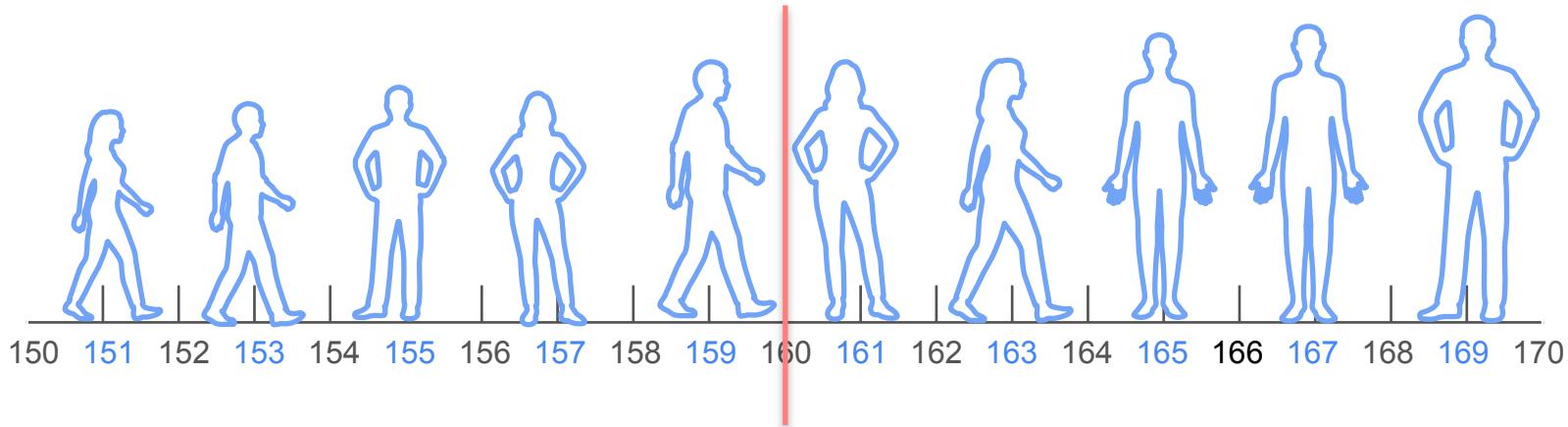
DeepLearning.AI

# Sample and Population

---

## Sample Mean

# Population and Sample Mean



What is the average height in statistopia?

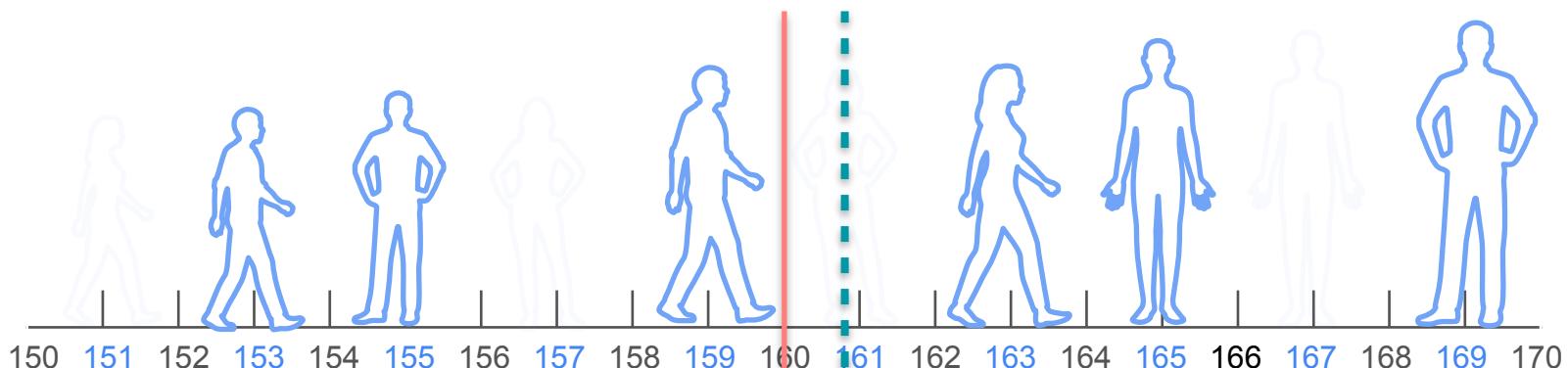
$$\frac{151 + 153 + 155 + 157 + 159 + 161 + 163 + 165 + 167 + 169}{10}$$

$$= \frac{1600}{10} = 160\text{cm}$$

Population mean

$\mu$

# Population and Sample Mean



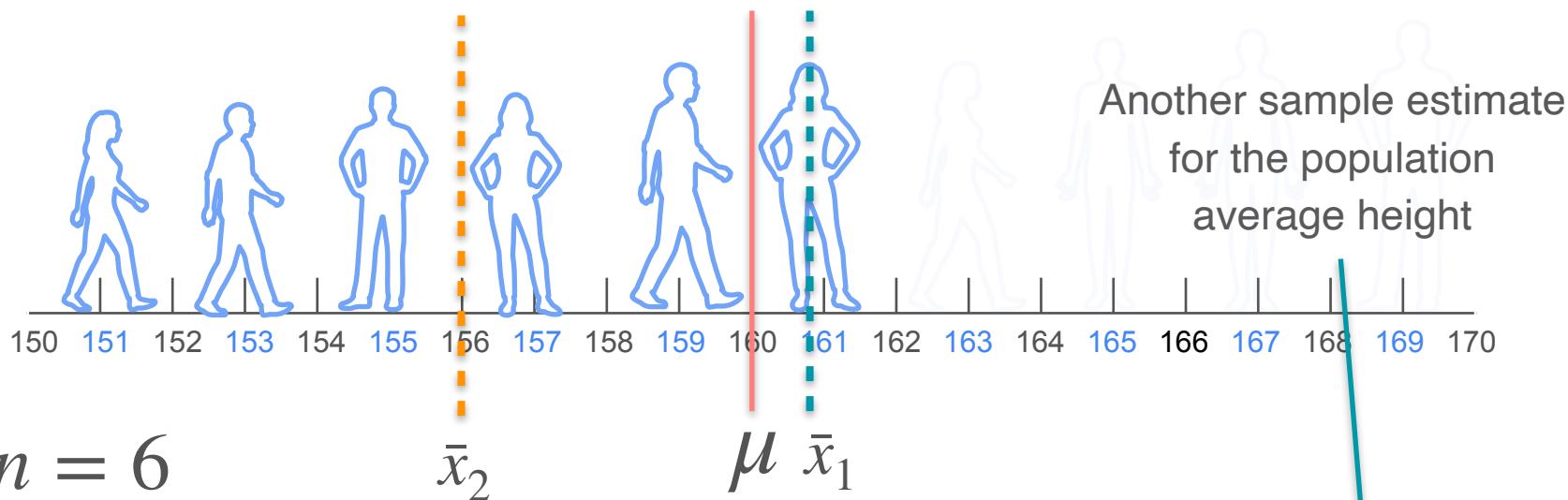
$$n = 6$$

What is the average height in statistopia?

$$\frac{153 + 155 + 159 + 163 + 165 + 169}{6} = \frac{964}{6} = 160.97$$

$\bar{x}$

# Population and Sample Mean



What is the average height in statistopia?

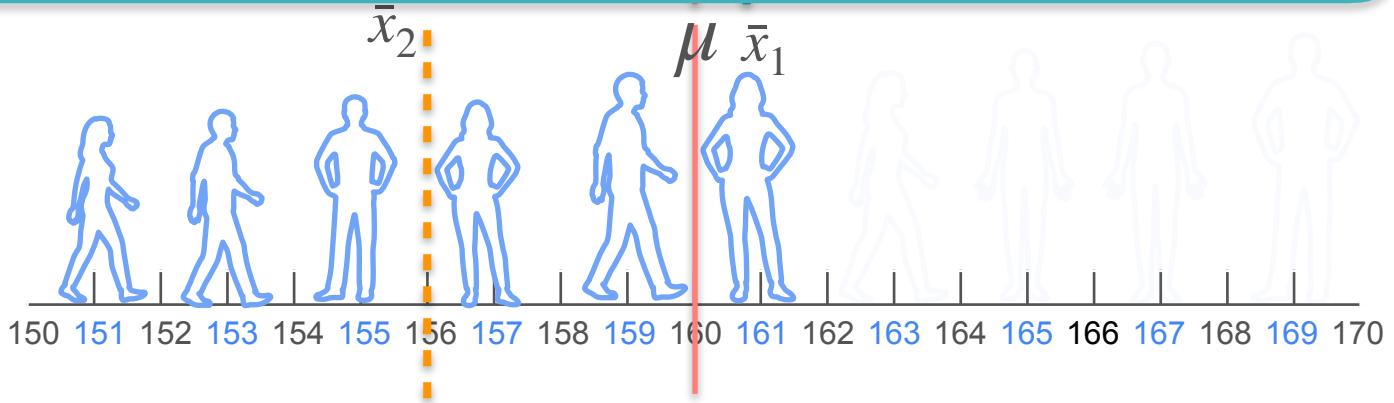
$$\frac{151 + 153 + 155 + 157 + 159 + 161}{6} = \frac{936}{6} = 156\text{cm}$$

# Population and Sample Mean

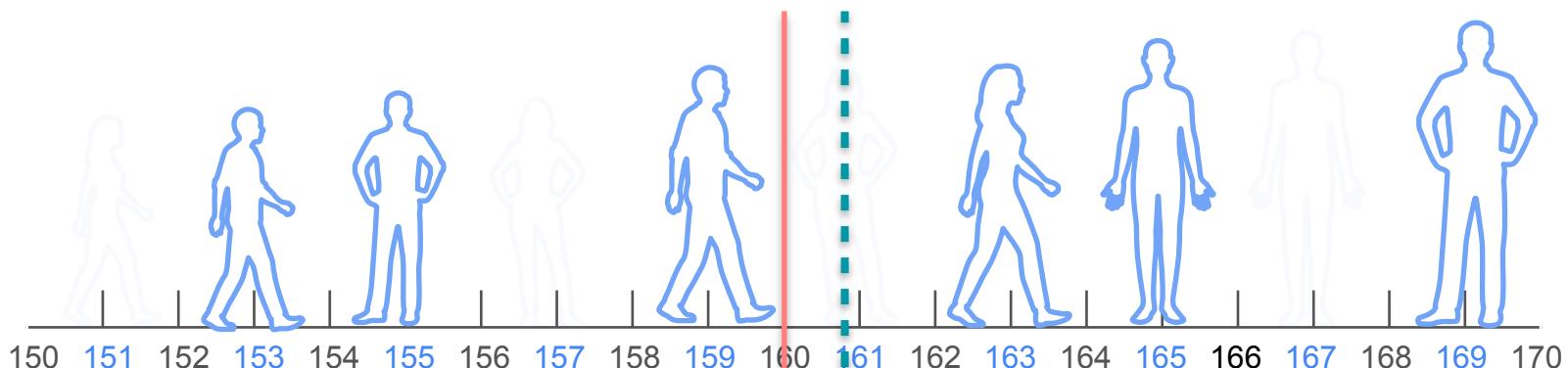
Better estimate of the population mean height

150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170

$$n = 6$$



# Population and Sample Mean

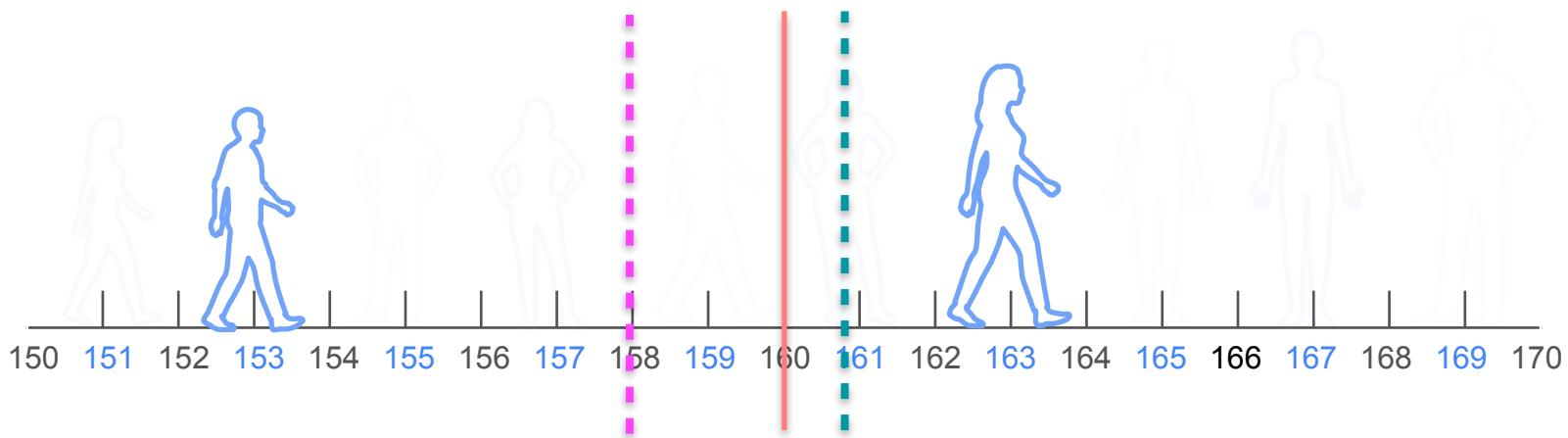


$$n = 6$$

$$n = 2$$

$$\mu \bar{x}_1$$

# Population and Sample Mean



$$n = 6$$

$$n = 2$$

What is the average height in statistopia?

$$\frac{153 + 163}{2} = \frac{316}{2} = 158\text{cm}$$



DeepLearning.AI

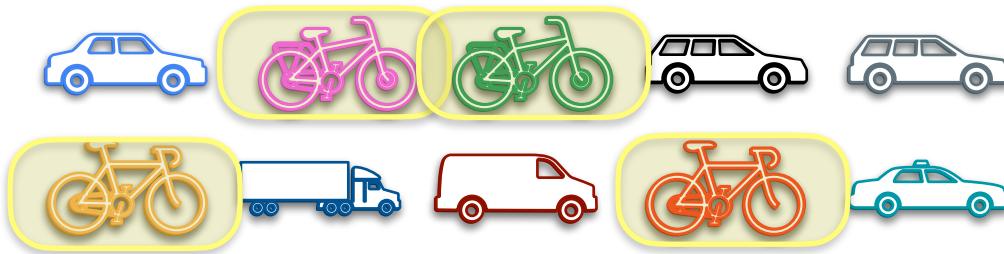
# Sample and Population

---

## Sample Proportion

# Proportion

Population size: 10



What proportion of people own a bicycle?

$$p \quad \text{population proportion} = \frac{4}{10} = 0.4 = 40\%$$

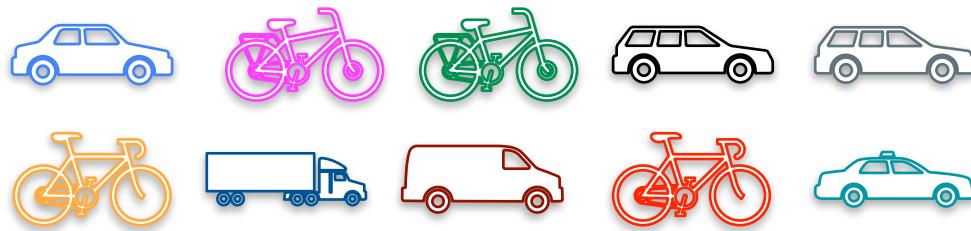
# Proportion

population proportion

$$P = \frac{\text{number of items with a given characteristic } (x)}{\text{population } (N)}$$

# Proportion

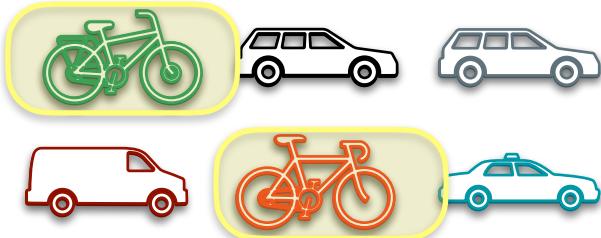
Population size: 10



# Sample Proportion

Sample size: 6

estimate of the population proportion



What proportion of people own a bicycle?

$$\hat{p} \text{ sample proportion } = \frac{2}{6} = 0.333 = 33.3\%$$



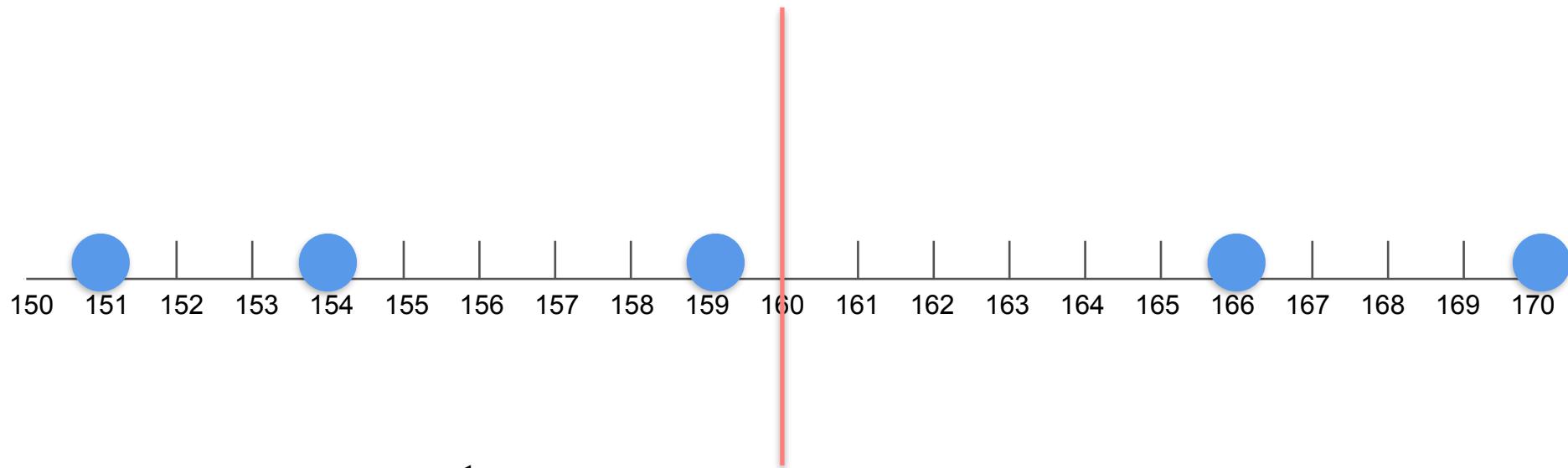
DeepLearning.AI

# Sample and Population

---

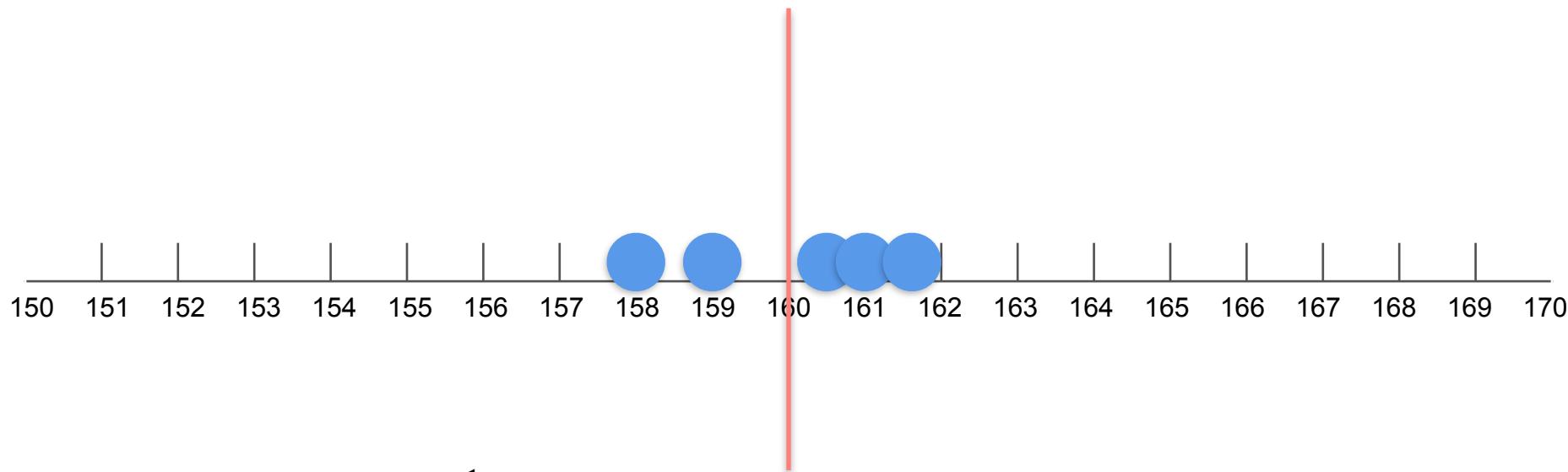
## Sample Variance

# Sample Variance



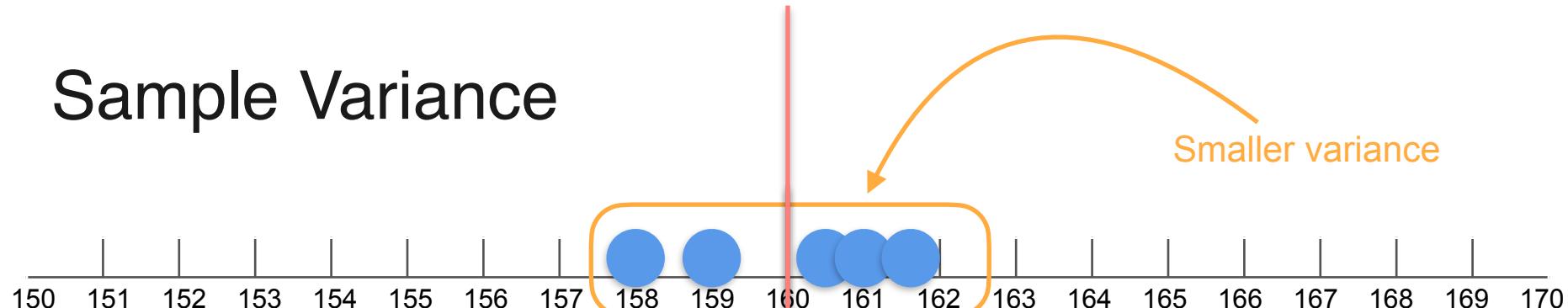
$$\mu = \frac{1}{5} (151 + 154 + 159 + 166 + 170) = 160$$

# Sample Variance



$$\mu = \frac{1}{5} (158 + 159 + 160.5 + 161 + 161.5) = 160$$

# Sample Variance



$$Var(X) = \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

population mean

population size

How to estimate population variance with a sample?

# Sample Variance

Let's cheat and use  
the sample mean

$$Var(X) = \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \longrightarrow \widehat{Var(X)} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

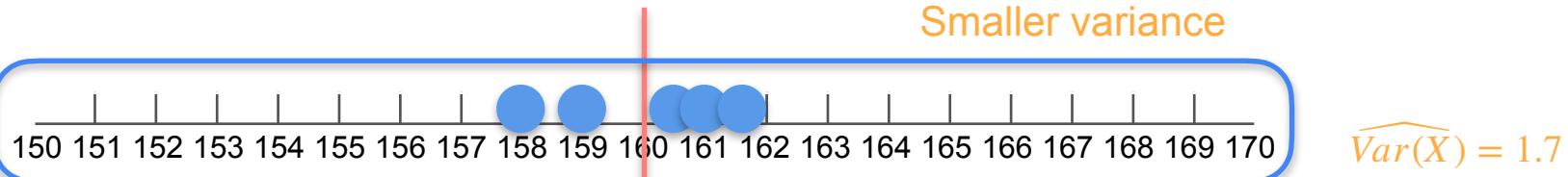
$$\downarrow \qquad \qquad Y = (X - \mu)^2 \qquad \qquad \uparrow$$
$$\mathbb{E}[Y] = \mu_Y = \frac{1}{N} \sum_{i=1}^N y_i \longrightarrow \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

The population mean of  $Y$

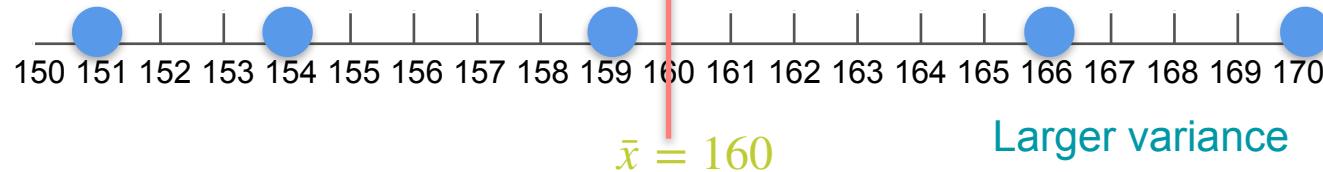
The sample mean of  $Y$

# Sample Variance

$$Var(X) = \sigma^2 = \frac{1}{N} \sum (x - \mu)^2 \longrightarrow \widehat{Var(X)} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$



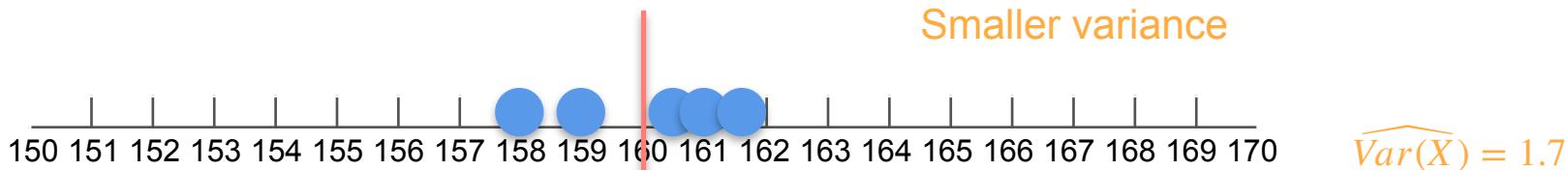
$$\widehat{Var(X)} = \frac{1}{5}((158-160)^2 + (159-160)^2 + (160.5-160)^2 + (161-160)^2 + (161.5-160)^2) = 1.7$$



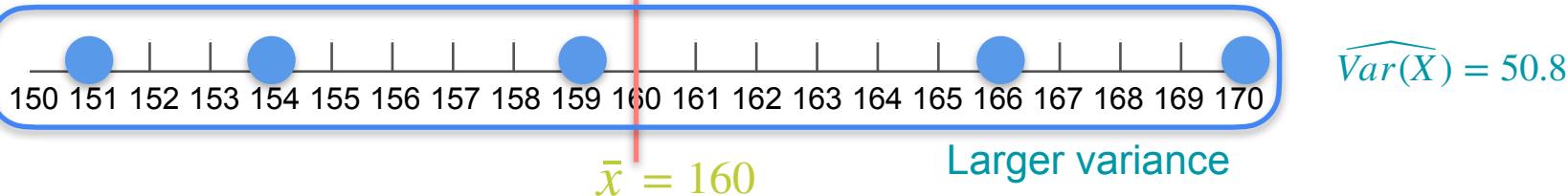
# Sample Variance

This equation is “biased”  
It underestimates the population variance

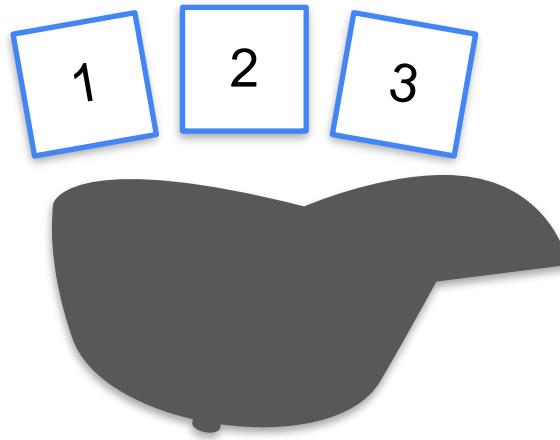
$$Var(X) = \sigma^2 = \frac{1}{N} \sum (x - \mu)^2 \longrightarrow \widehat{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$



$$\widehat{Var}(X) = \frac{1}{5} ((151-160)^2 + (154-160)^2 + (159-160)^2 + (166-160)^2 + (170-160)^2) = 50.8$$



# Variance Estimation



$$\mu = \frac{1 + 2 + 3}{3} = \frac{6}{3} = 2$$

# Variance Estimation

1	2	3
---	---	---

$$\mu = 2$$

$$\sigma^2 = \frac{1}{N} \sum (x - \mu)^2$$

$x$	$x - \mu$	$(x - \mu)^2$
1	-1	1
2	0	0
3	1	1

$$\frac{\sum (x - \mu)^2}{N} = \frac{2}{3}$$

$\sigma^2$   
Population variance

# Variance Estimation

1 2 3

$$\mu = 2$$

$$\sigma^2 = \frac{1}{N} \sum (x - \mu)^2$$

$$\sigma^2 = \frac{2}{3}$$

$n = 2$   
Samples

1	1
1	2
1	3
2	1
2	2
2	3
3	1
3	2
3	3

# Variance Estimation

1 2 3

$$\mu = 2$$

$$\sigma^2 = \frac{1}{N} \sum (x - \mu)^2$$

$$\sigma^2 = \frac{2}{3}$$

$$n = 2 \text{ Samples}$$
$$\bar{x}$$
$$\widehat{Var}(X) = \frac{\sum (x - \bar{x})^2}{n}$$

1	1	1
1	2	1.5
1	3	2
2	1	1.5
2	2	2
2	3	2.5
3	1	2
3	2	2.5
3	3	3

$$\widehat{Var}(X) = \frac{\sum (x - \bar{x})^2}{n}$$

0
0.25
1
0.25
0
0.25
1
0.25
0

estimated variance

$$= 0.333$$

$$= \frac{1}{3}$$

$$\begin{matrix} 1 \\ 2 \\ 3 \end{matrix}$$

$$\mu = 2$$

$$\sigma^2 = \frac{1}{N} \sum (x - \mu)^2$$

$$\sigma^2 = \frac{2}{3}$$

$$\begin{matrix} n = 2 \\ \text{Samples} \end{matrix}$$

1	1
1	2
1	3
2	1
2	2
2	3
3	1
3	2
3	3

$$\bar{x}$$

$$\widehat{Var}(X) = \frac{\sum (x - \bar{x})^2}{n - 1}$$

0
0.25
1
0.25
0
0.25
1
0.25
0

estimated variance

$$= 0.333$$

$$= \frac{1}{3}$$

# Variance Estimation

$$\begin{matrix} 1 & 2 & 3 \end{matrix}$$

$$\mu = 2$$

$$\sigma^2 = \frac{1}{N} \sum (x - \mu)^2$$

$$\sigma^2 = \frac{2}{3}$$

$$\begin{matrix} n = 2 \\ \text{Samples} \end{matrix}$$

1	1
1	2
1	3
2	1
2	2
2	3
3	1
3	2
3	3

$$\bar{x}$$

1
1.5
2
1.5
2
2.5
2
2.5
3

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

0
0.5
2
0.5
0
0.5
2
0.5
0

estimated variance

$$= 0.667$$

$$= \frac{2}{3}$$

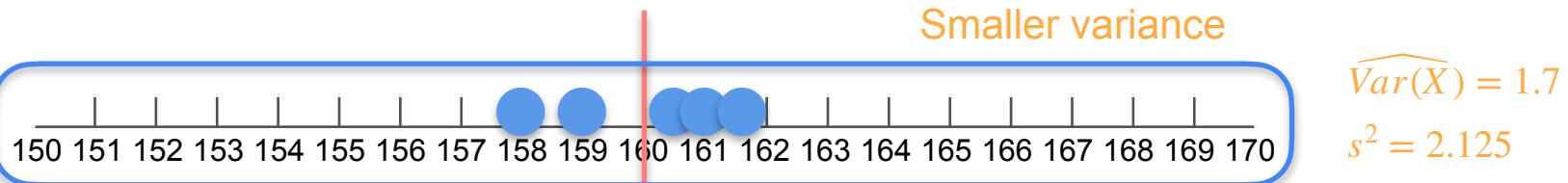
# Variance Estimation

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

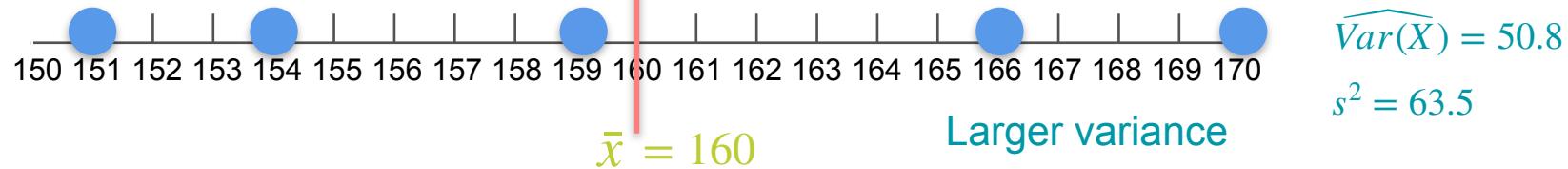
- $n - 1$  fixes bias when all you have is a sample
- As  $n$  gets big, the difference matters less
- If it matters, you may have too little data
- Some accepted statistical techniques use  $n$

# Sample Variance

$$Var(X) = \sigma^2 = \frac{1}{N} \sum (x - \mu)^2 \longrightarrow s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$



$$s^2 = \frac{1}{5-1} ((158 - 160)^2 + (159 - 160)^2 + (160.5 - 160)^2 + (161 - 160)^2 + (161.5 - 160)^2)$$



# Variance Estimation

## Population Variance Formula

$$Var(X) = \sigma^2 = \frac{1}{N} \sum (x - \mu)^2$$

## Sample Variance Formula

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

$$\widehat{\sigma}^2 = \widehat{Var(X)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$



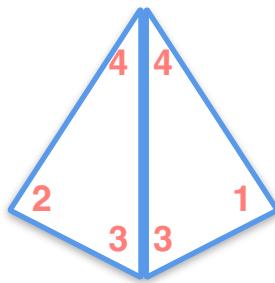
DeepLearning.AI

## Sample and Population

---

## Law of Large Numbers

# Law of Large Numbers



$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \mu = 2.5 \end{array}$$

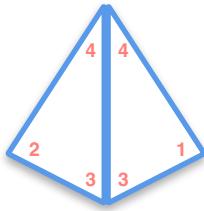
	1	2	3	4
1	1	1.5	2	2.5
2	1.5	2	2.5	3
3	2	2.5	3	3.5
4	2.5	3	3.5	4

	1	2	3	4
1	1			
2	2			
3	3			
4	4			

Experiment:

Toss the 4-sided dice twice and record the average of your outcomes

# Law of Large Numbers



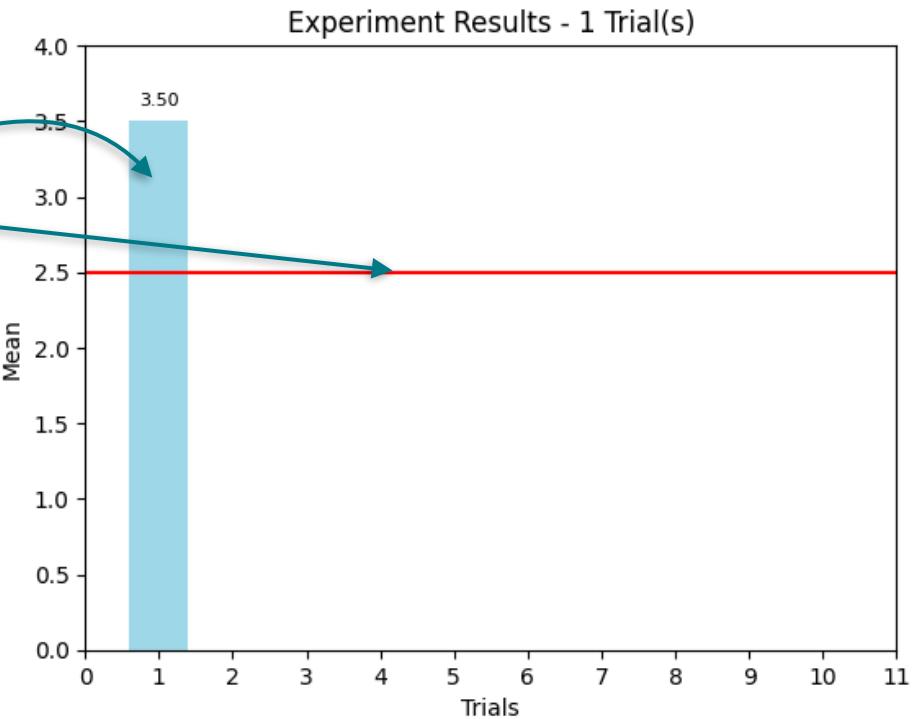
$$1 \ 2 \ 3 \ 4$$
$$\mu = 2.5$$

1 trial

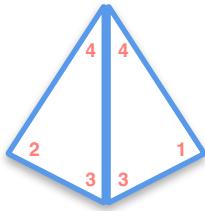
4,3

$$\bar{x}_1$$

	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4



# Law of Large Numbers



1 2 3 4  
 $\mu = 2.5$

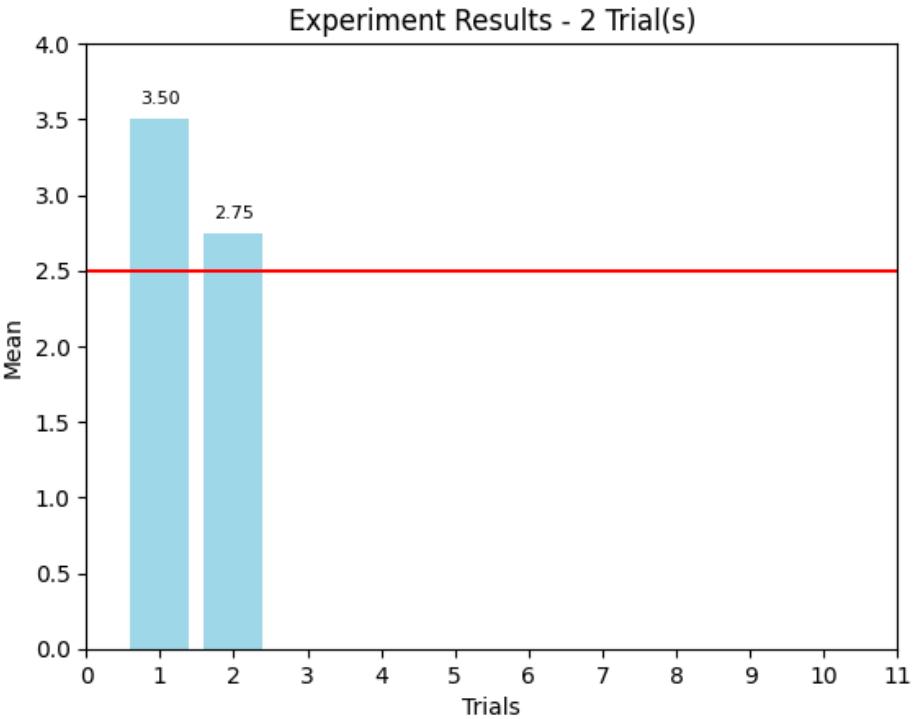
	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

1 trial

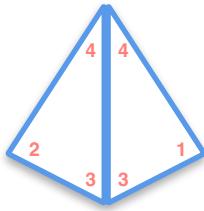
4,3

2 trials

3,4  
1,3



# Law of Large Numbers



1 2 3 4  
 $\mu = 2.5$

	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

1 trial

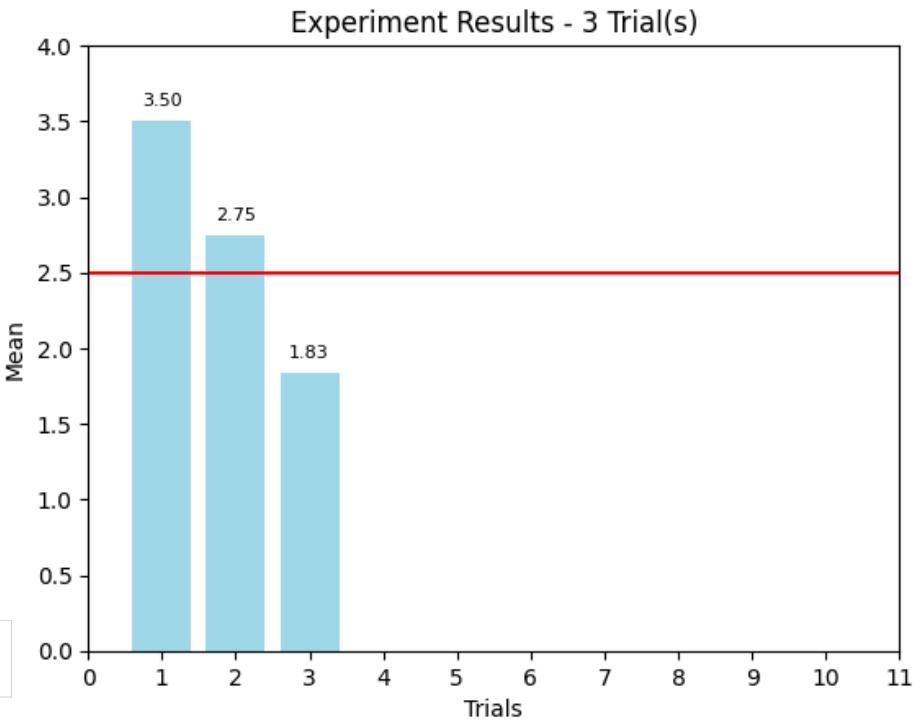
4,3

2 trials

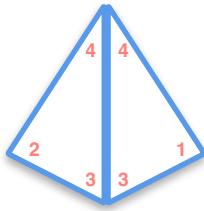
3,4  
1,3

3 trials

3,1    1,4    1,1



# Law of Large Numbers



1 2 3 4  
 $\mu = 2.5$

	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

2 trials

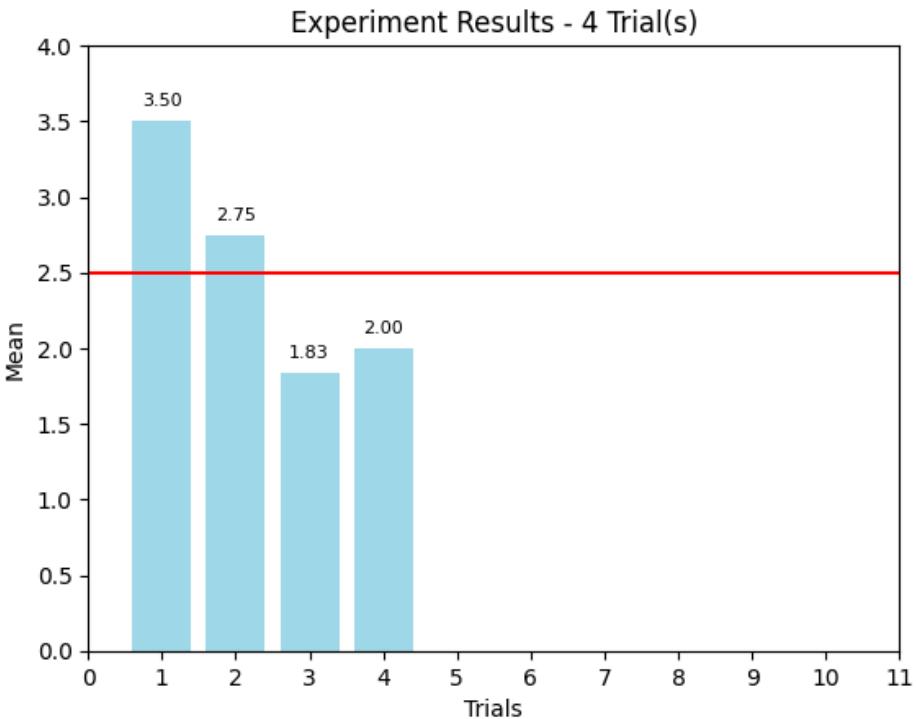
3,4
1,3

3 trials

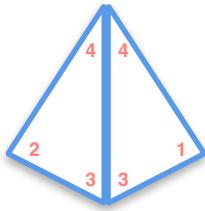
3,1	1,4	1,1
-----	-----	-----

4 trials

3,1	3,1
1,2	3,2

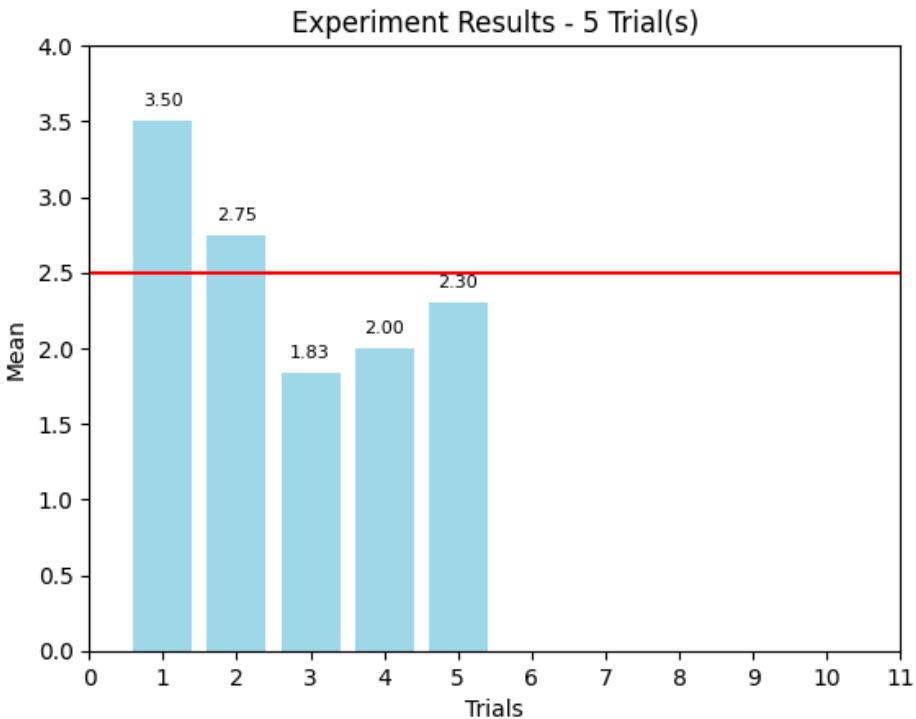


# Law of Large Numbers

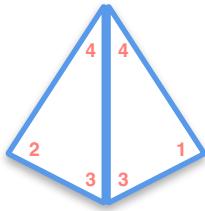


1 2 3 4  
 $\mu = 2.5$

	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

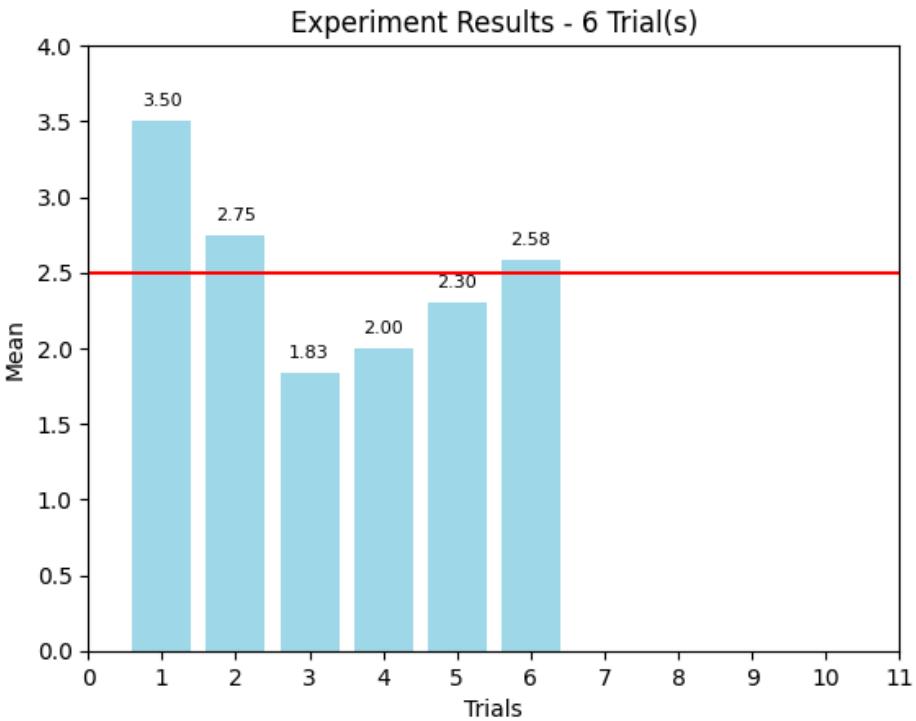


# Law of Large Numbers

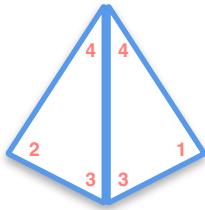


1 2 3 4  
 $\mu = 2.5$

	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

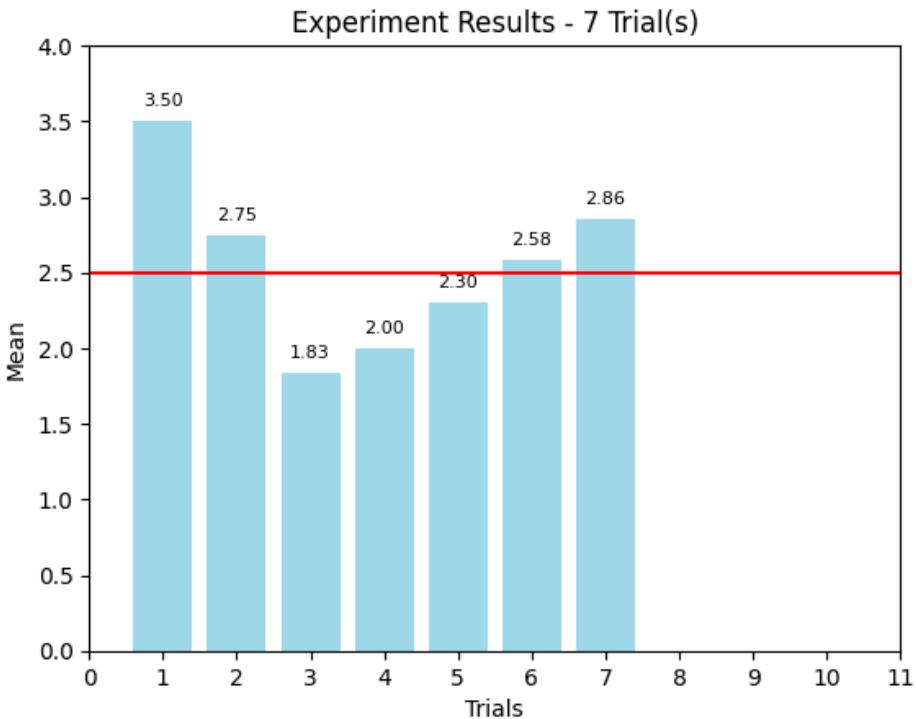


# Law of Large Numbers

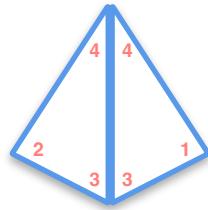


1 2 3 4  
 $\mu = 2.5$

	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

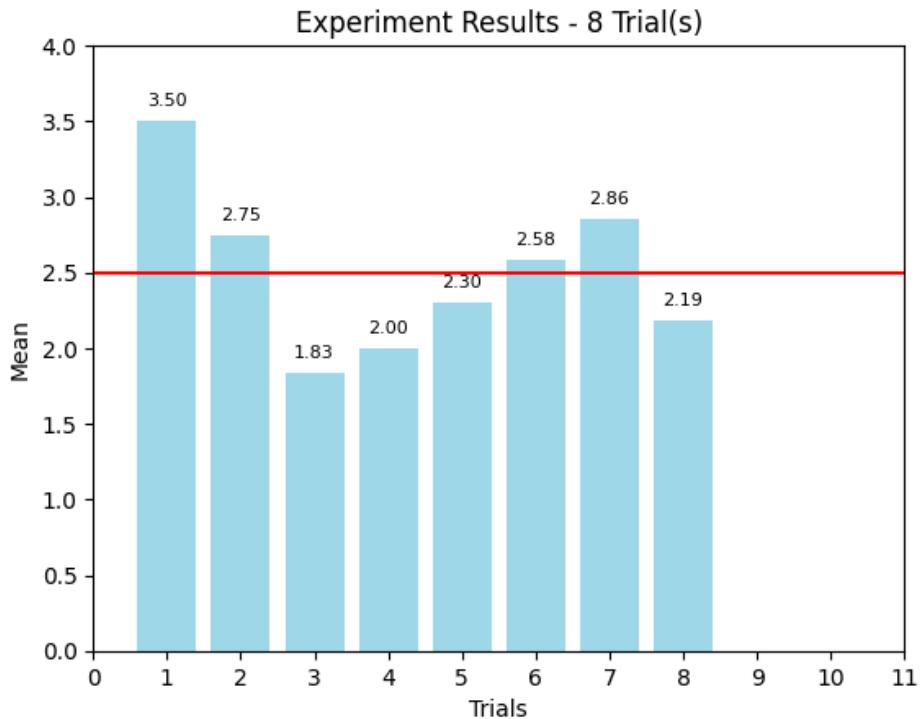


# Law of Large Numbers

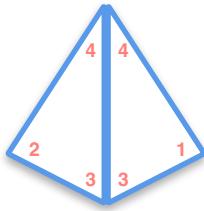


1 2 3 4  
 $\mu = 2.5$

	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

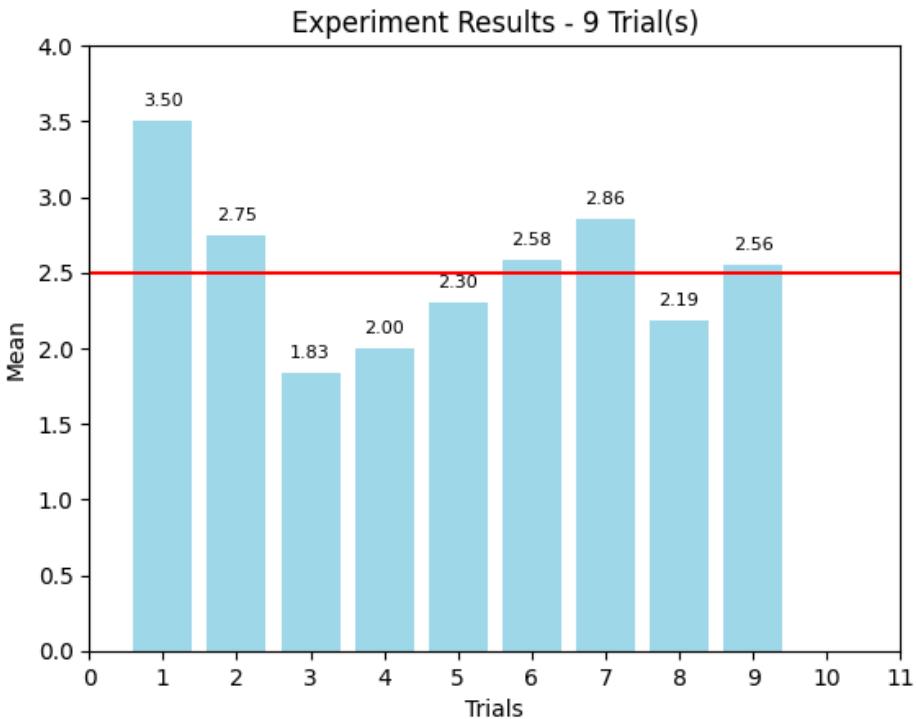


# Law of Large Numbers

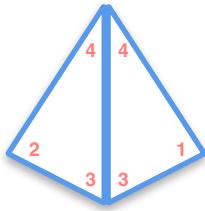


1 2 3 4  
 $\mu = 2.5$

	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

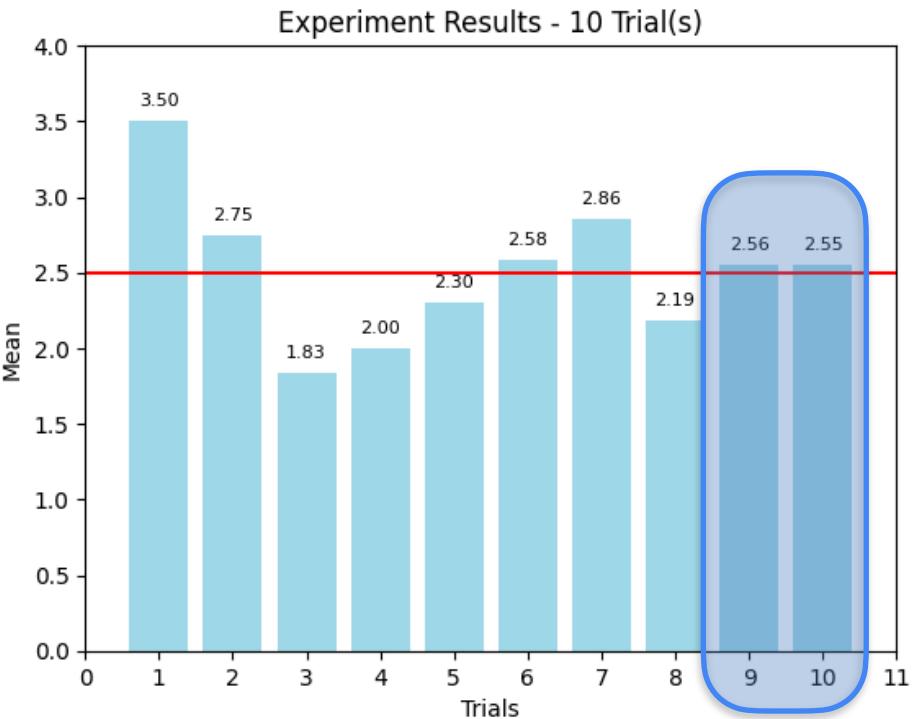


# Law of Large Numbers



1 2 3 4  
 $\mu = 2.5$

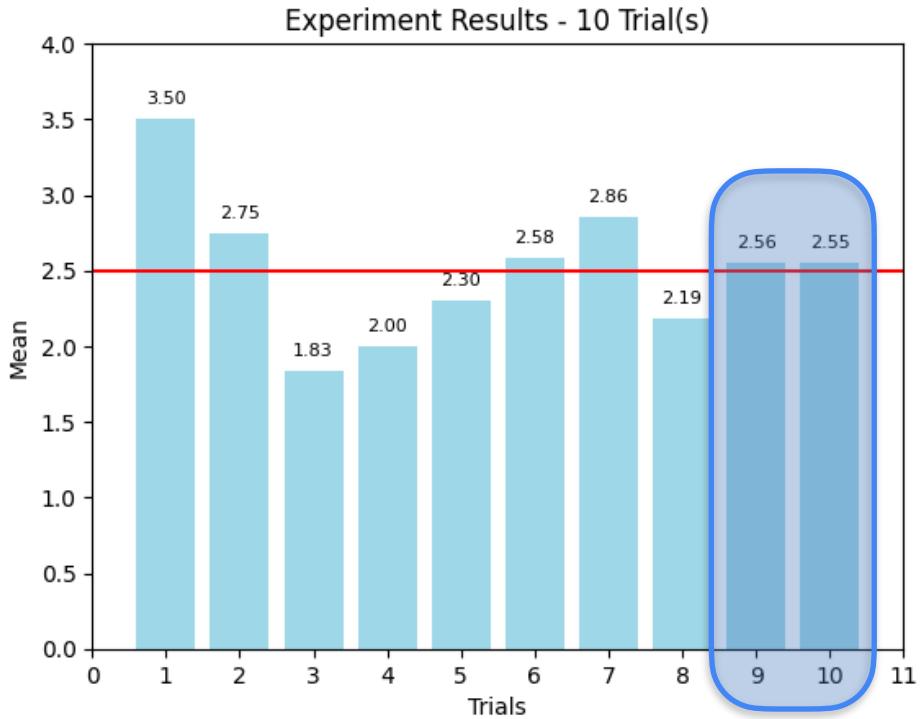
	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4



# Law of Large Numbers

As the sample size increases, the average of the sample will tend to get closer to the average of the entire population.

## Law of Large Numbers



# Law of Large Numbers

## Law of Large Numbers

$n$  : number of samples

$X_i$  : is the  $i$ -th random sample from the population.

Each  $X_i$  are independent and identically distributed (i.i.d.)

as  $n \rightarrow \infty$

$$\frac{X_1 + X_2 + X_3 + \dots + X_n}{n} \longrightarrow \mathbb{E}[X] = \mu_X$$

UNDER CERTAIN CONDITIONS

as  $n \rightarrow \infty$

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mathbb{E}[X] = \mu_X$$

# Law of Large Numbers

## UNDER CERTAIN CONDITIONS

- Sample is randomly drawn.
- Sample size must be sufficiently large.
- Independent observations.



DeepLearning.AI

## Sample and Population

---

**Central Limit Theorem  
Discrete Random Variable**

# Central Limit Theorem (CLT) - Example 1



$$P(H) = 0.5$$



$$P(T) = 0.5$$

Random variable  $\rightarrow X$  number of heads when a coin is flipped n times

$$\text{If } n = 1$$

$$X = 1$$

$$X = 0$$

Discrete Random Variable

# Central Limit Theorem (CLT) - Example 1



$$P(H) = 0.5$$

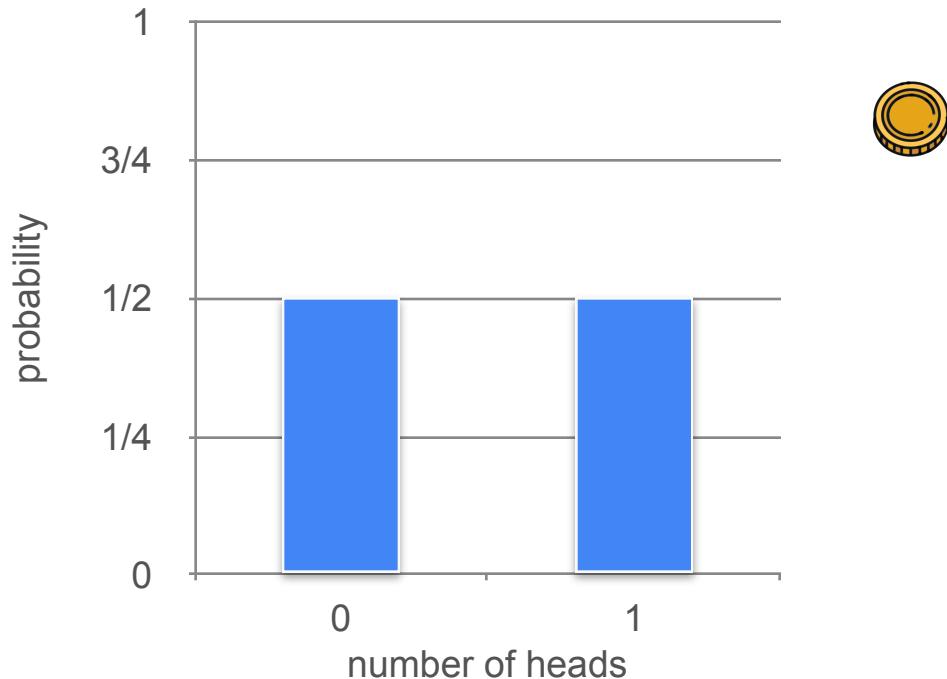


$$P(T) = 0.5$$

$$X = 1$$

$$X = 0$$

What can we say about the probability distribution when the number of coin flips increases?



# Central Limit Theorem (CLT) - Example 1



$$P(H) = 0.5$$

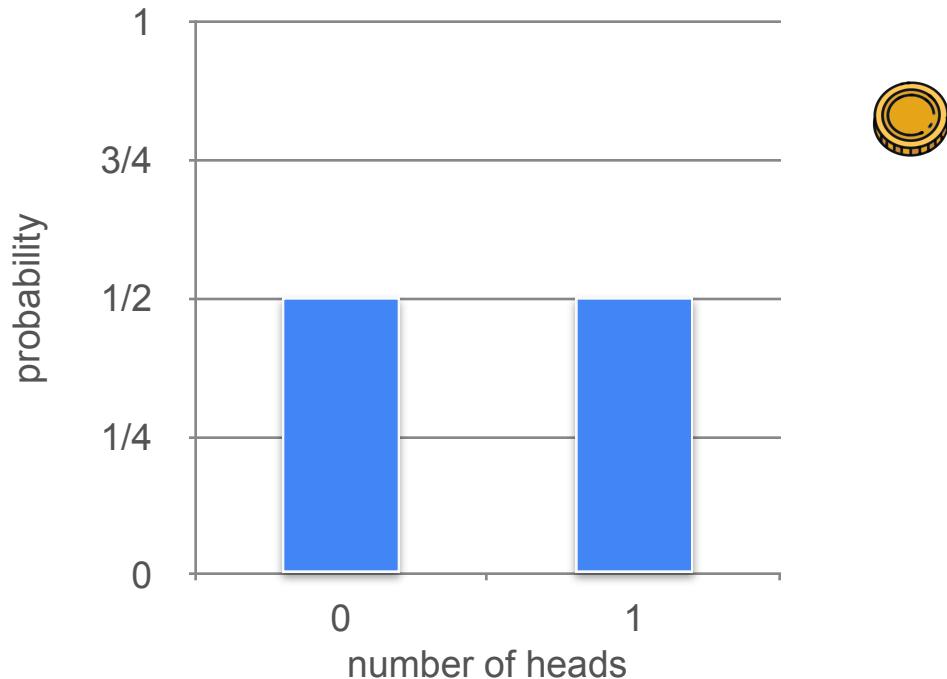


$$P(T) = 0.5$$

$$X = 1$$

$$X = 0$$

What can we say about the probability distribution when the number of coin flips increases?



# Central Limit Theorem (CLT) - Example 1



$$P(H) = 0.5$$

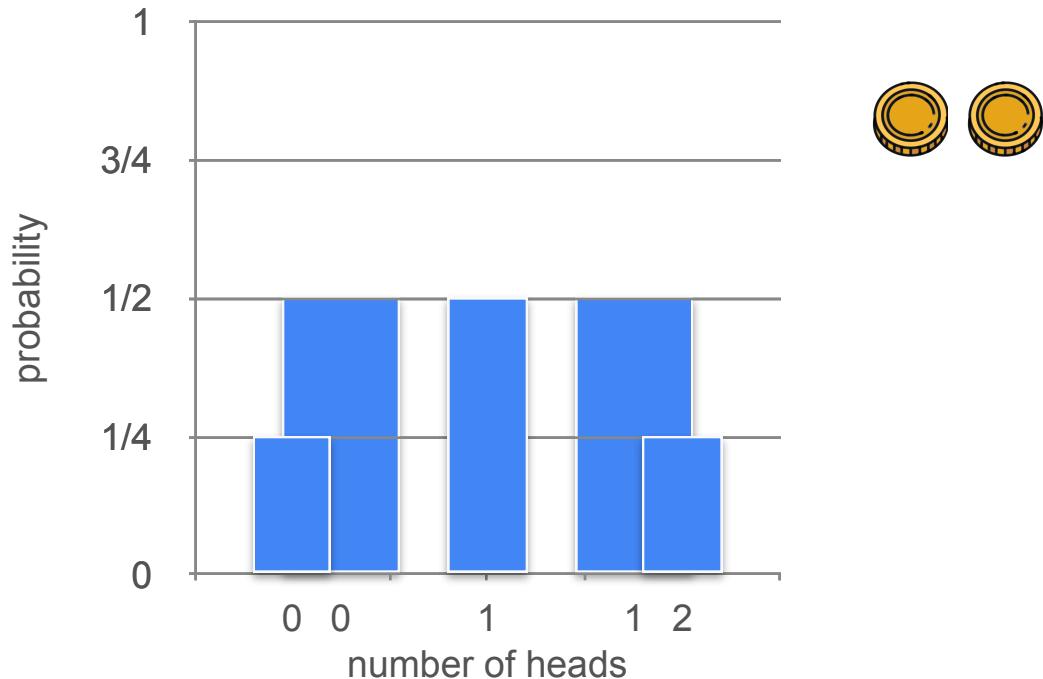


$$P(T) = 0.5$$

$$X = 1$$

$$X = 0$$

What can we say about the probability distribution when the number of coin flips increases?



# Central Limit Theorem (CLT) - Example 1



$$P(H) = 0.5$$

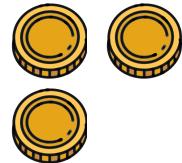
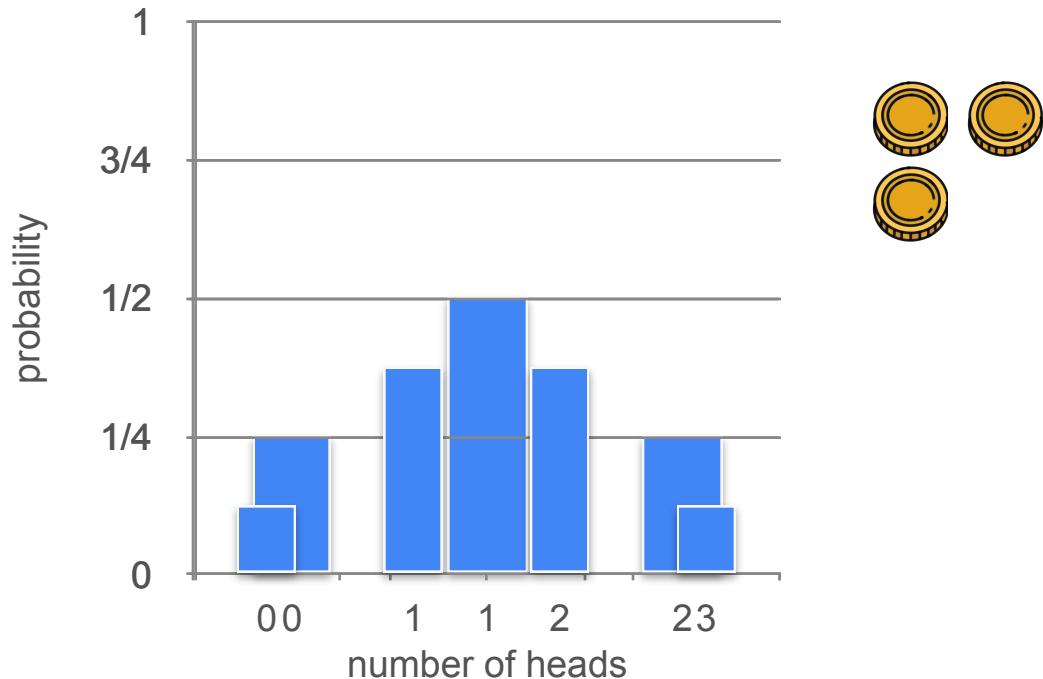


$$P(T) = 0.5$$

$$X = 1$$

$$X = 0$$

What can we say about the probability distribution when the number of coin flips increases?



# Central Limit Theorem (CLT) - Example 1



$$P(H) = 0.5$$

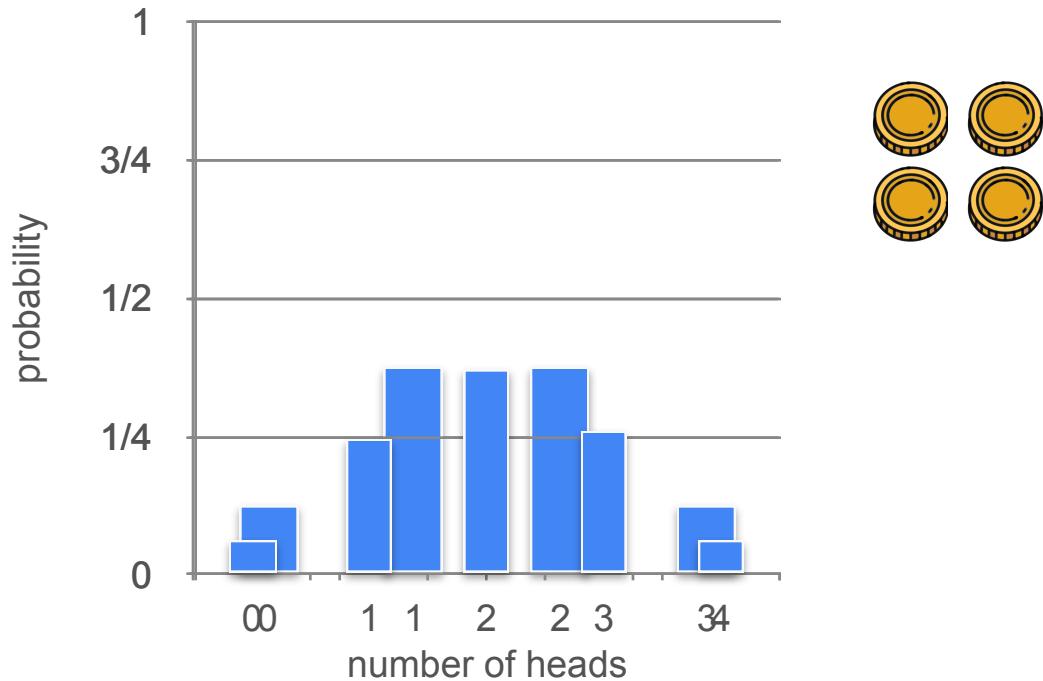


$$P(T) = 0.5$$

$$X = 1$$

$$X = 0$$

What can we say about the probability distribution when the number of coin flips increases?



# Central Limit Theorem (CLT) - Example 1



$$P(H) = 0.5$$

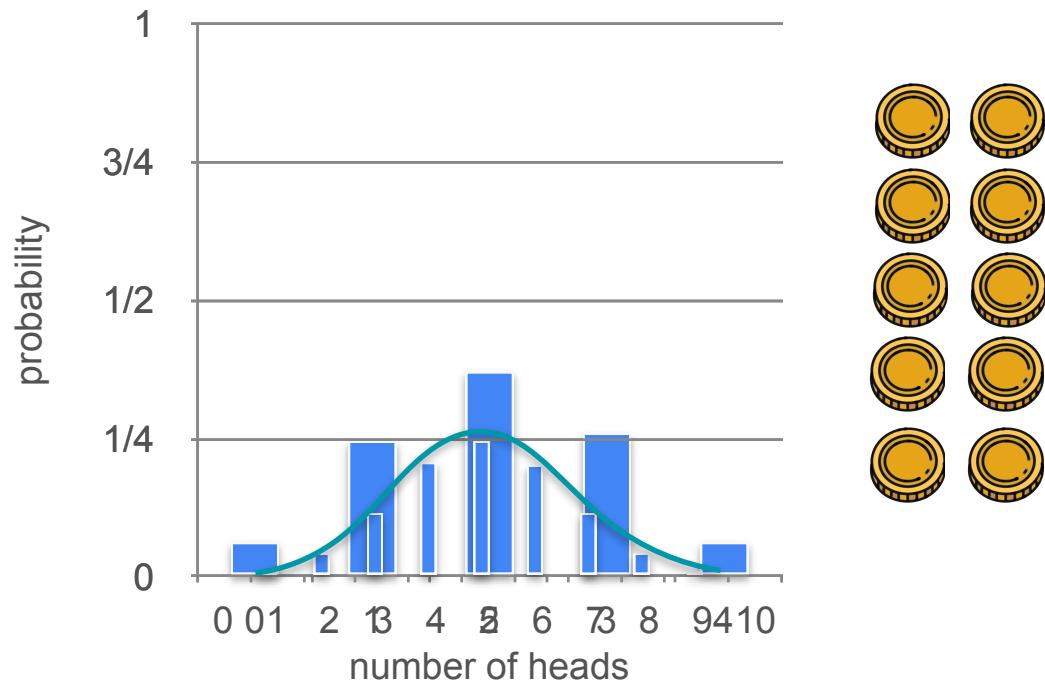


$$P(T) = 0.5$$

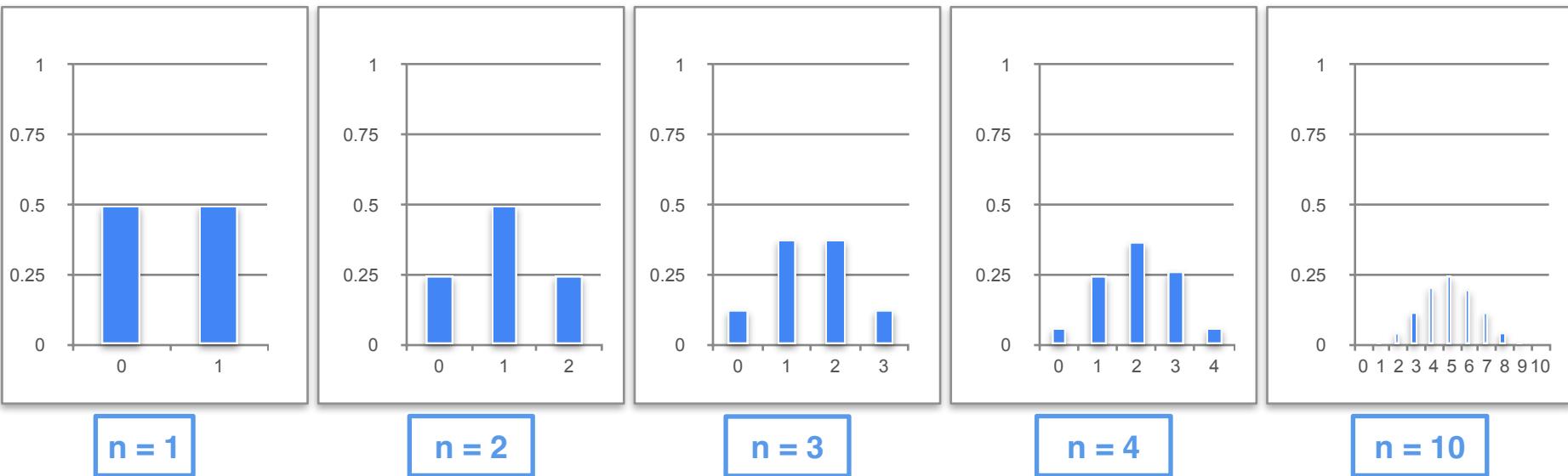
$$X = 1$$

$$X = 0$$

What can we say about the distribution when the number of coins we flip increases?



# Central Limit Theorem (CLT) - Example 1



As  $n$  increases, the probability distribution becomes closer to a Gaussian distribution

# Central Limit Theorem (CLT) - Example 1



$$\mathbf{P}(H) = 0.5$$



$$\mathbf{P}(T) = 0.5$$

Random variable



**$X$  number of heads when a coin is flipped n times**

$$\mu = np = n\mathbf{P}(H)$$



$$\sigma^2 = np(1 - p) = n\mathbf{P}(H)\mathbf{P}(T)$$



# Central Limit Theorem (CLT) - Example 1



$$\mathbf{P}(H) = 0.5 \quad \mathbf{P}(T) = 0.5$$

$$\mu = np = n\mathbf{P}(H)$$
Two gold-colored coins, one with 'H' and one with 'T'.

$$\sigma^2 = np(1 - p) = n\mathbf{P}(H)\mathbf{P}(T)$$
Two gold-colored coins, one with 'H' and one with 'T'.



$$n = 1$$

$$\mu = np$$



$$\mu = 1 \times 0.5 = 0.5$$

$$\sigma^2 = np(1 - p)$$

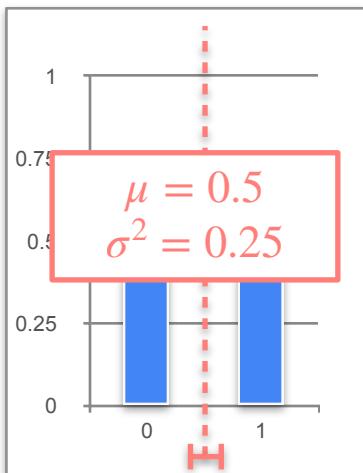


$$\sigma^2 = (1 \times 0.5)(0.5) = 0.25$$

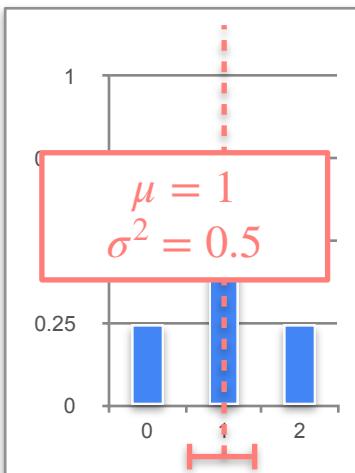
# Central Limit Theorem (CLT) - Example 1

$$\mu = np$$

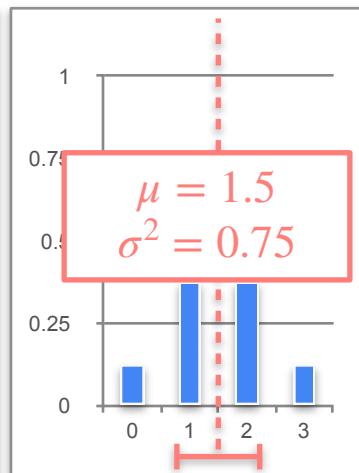
$$\sigma^2 = np(1 - p)$$



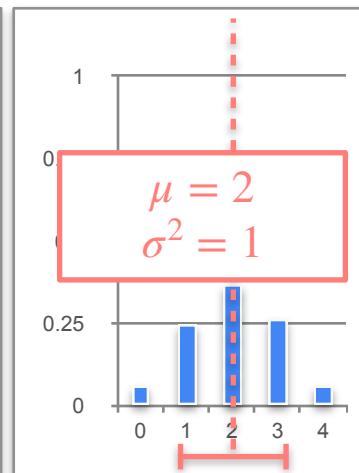
$n = 1$



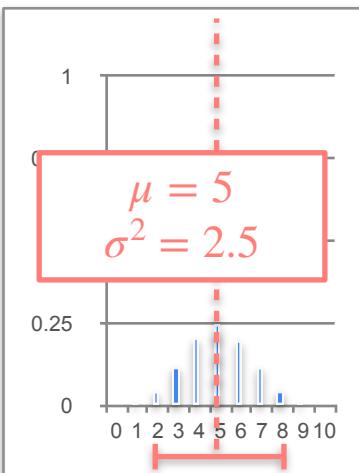
$n = 2$



$n = 3$



$n = 4$



$n = 10$

As  $n$  increases, the probability distribution becomes closer to a gaussian distribution

# Central Limit Theorem (CLT) - Example 1

$$\mu = np$$

$$\sigma^2 = np(1 - p)$$

$$\begin{aligned}\mu &= 0.5 \\ \sigma^2 &= 0.25\end{aligned}$$

$$\begin{aligned}\mu &= 1 \\ \sigma^2 &= 0.5\end{aligned}$$

$$\begin{aligned}\mu &= 1.5 \\ \sigma^2 &= 0.75\end{aligned}$$

$$\begin{aligned}\mu &= 2 \\ \sigma^2 &= 1\end{aligned}$$

$$\begin{aligned}\mu &= 5 \\ \sigma^2 &= 2.5\end{aligned}$$

as  $n$  become sufficiently large we will get a normal distribution with parameters

$$\mu = np$$

$$\sigma^2 = np(1 - p)$$



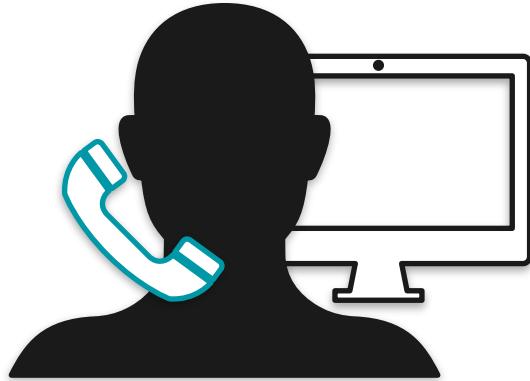
DeepLearning.AI

## Sample and Population

---

**Central Limit Theorem  
Continuous Random Variable**

# Uniform Distribution: Motivation



You're calling a tech support line. They can answer any time between zero and 15 minutes and if they don't answer in this time, the line is disconnected.

$X$  = "Wait time for a called to be answered "

$$X \sim \mathcal{U}(0,15)$$

# Central Limit Theorem (CLT) - Example 2

$$n = 1$$

$$Y_1 = \frac{X_1}{1}$$

Record the average of all  $n$  experiments

$$n = 2$$

$$Y_2 = \frac{X_1 + X_2}{2}$$

$$Y_n = \frac{1}{n} \sum_{i=1}^n X_i$$

$$n = 3$$

$$Y_3 = \frac{X_1 + X_2 + X_3}{3}$$

⋮

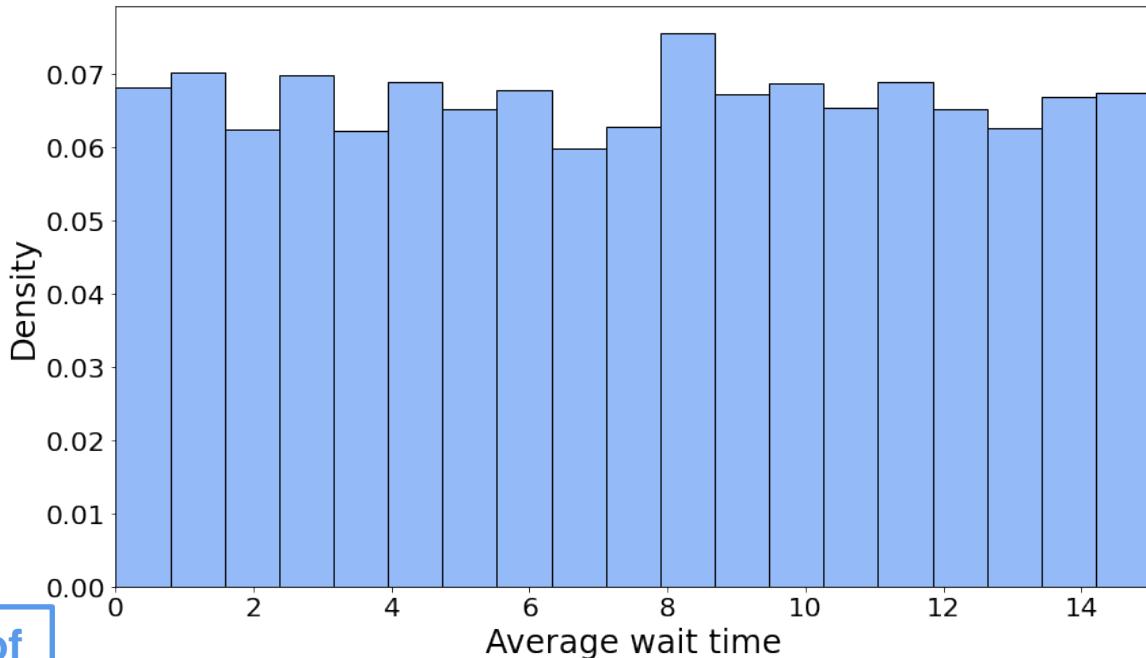
⋮

Can we say anything about the distribution of this average?

# Central Limit Theorem (CLT) - Example 2

$$n = 1 \quad Y_1 = \frac{X_1}{1}$$

Create many samples of  $Y_1$  so you can get a pretty histogram



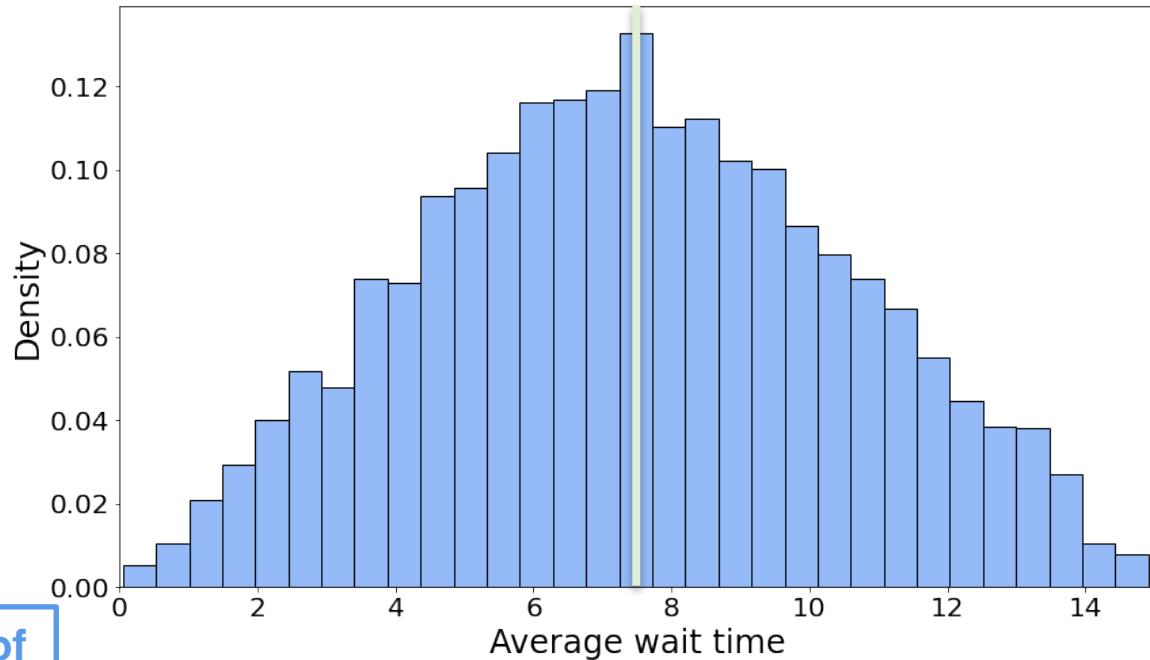
What happens to the distribution of these averages as  $n$  increases?

# Central Limit Theorem (CLT) - Example 2

$$n = 2 \quad Y_2 = \frac{X_1 + X_2}{2}$$

Create many samples of  $Y_2$  so you can get a pretty histogram

What happens to the distribution of these averages as  $n$  increases

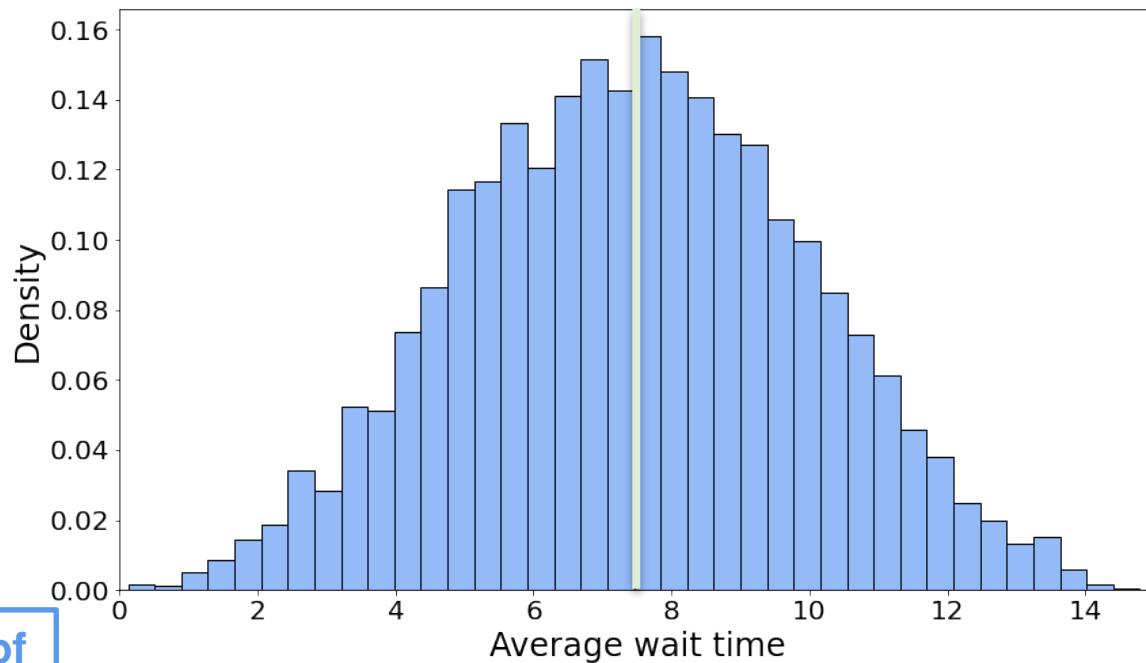


# Central Limit Theorem (CLT) - Example 2

$$n = 3 \quad Y_3 = \frac{X_1 + X_2 + X_3}{3}$$

Create many samples of  $Y_3$  so you can get a pretty histogram

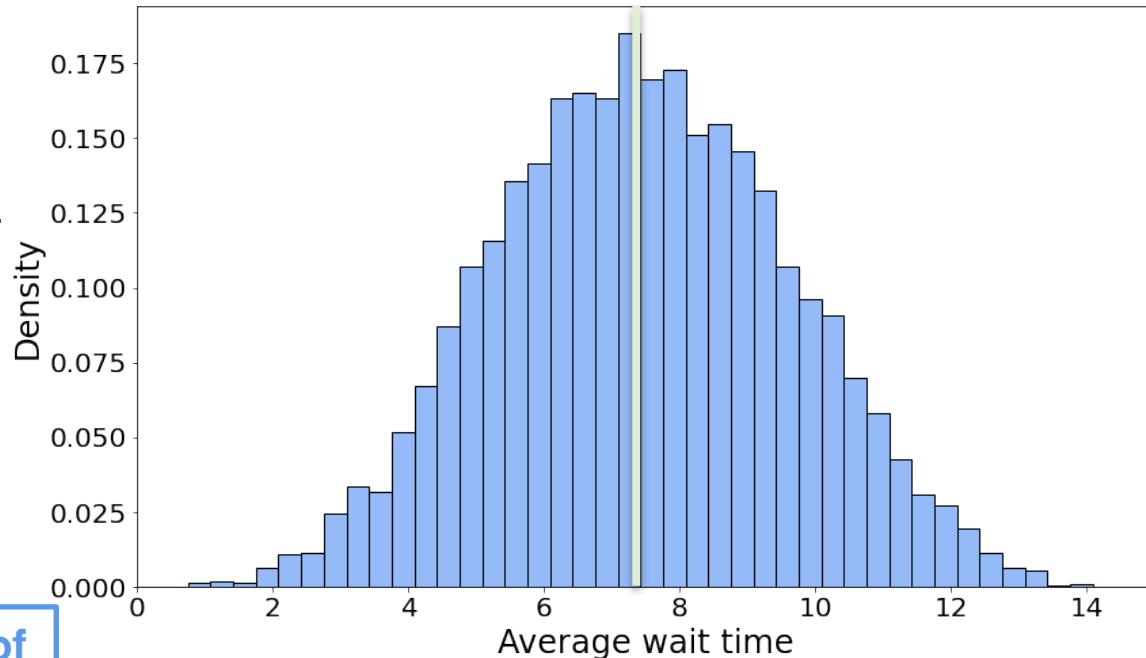
What happens to the distribution of these averages as  $n$  increases



# Central Limit Theorem (CLT) - Example 2

$$n = 4 \quad Y_4 = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

Create many samples of  $Y_4$  so you can get a pretty histogram

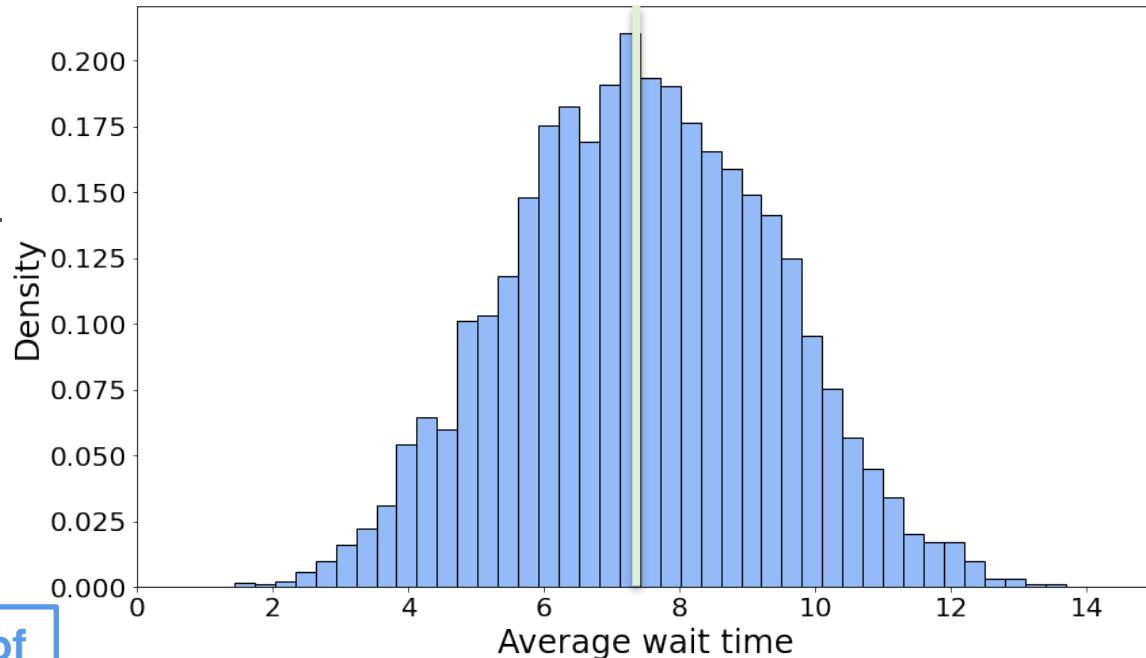


What happens to the distribution of these averages as  $n$  increases

# Central Limit Theorem (CLT) - Example 2

$$n = 5 \quad Y_5 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

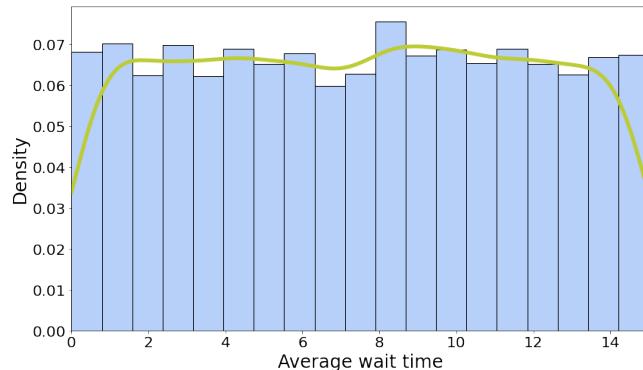
Create many samples of  $Y_5$  so you can get a pretty histogram



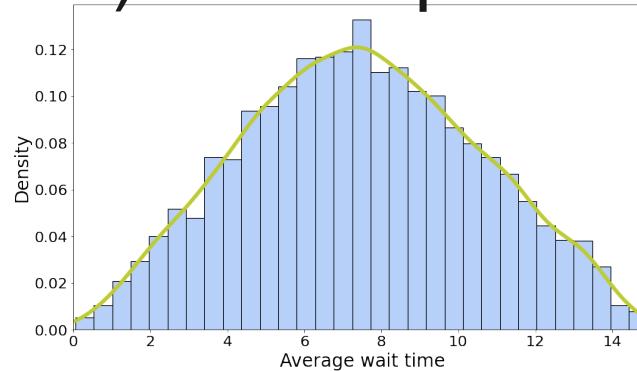
What happens to the distribution of these averages as  $n$  increases

# Central Limit Theorem (CLT) - Example 2

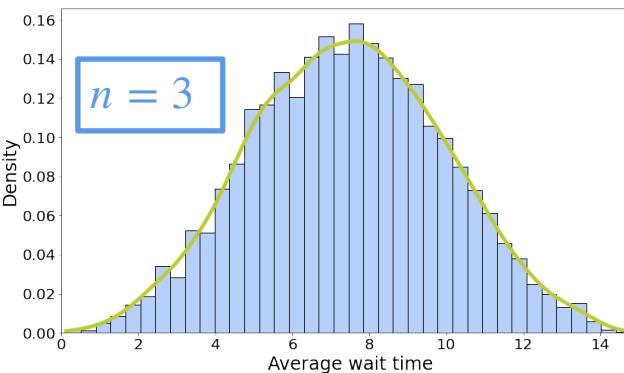
$n = 1$



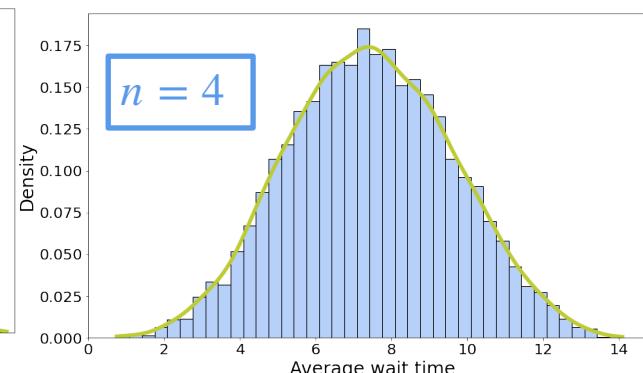
$n = 2$



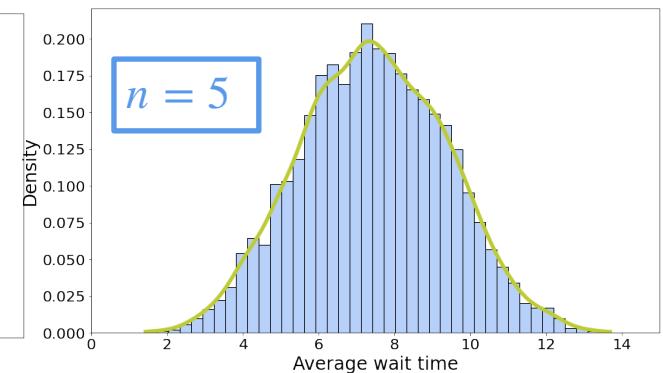
$n = 3$



$n = 4$



$n = 5$



# Central Limit Theorem (CLT) - Example 2

$$\mathbb{E}[Y_n] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} n \mathbb{E}[X] = \mathbb{E}[X] = 7.5$$

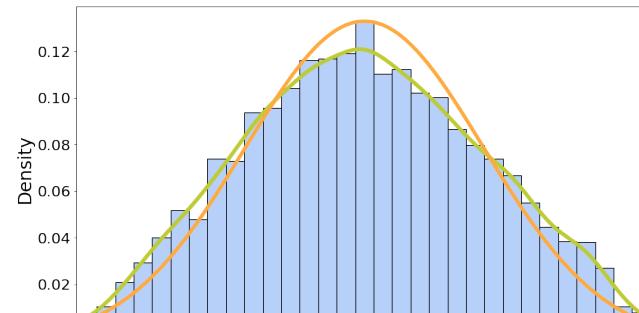
$$\begin{aligned} Var[Y_n] &= Var \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) \\ &= \frac{1}{n^2} n Var(X) = \frac{Var(X)}{n} = \frac{18.75}{n} \end{aligned}$$

# Central Limit Theorem (CLT) - Example 2

$n = 1$

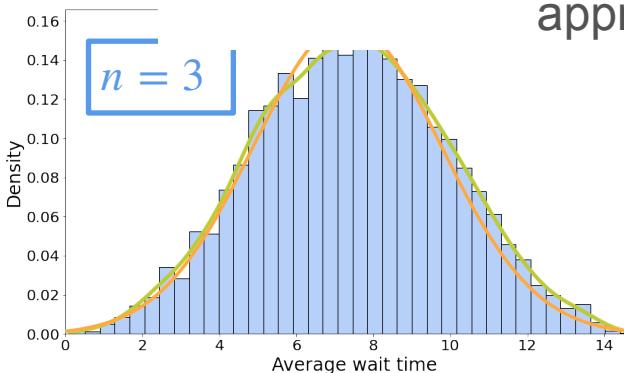


$n = 2$

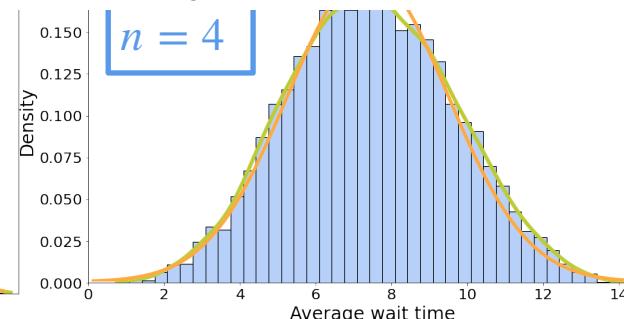


When you average a large enough number of variables, the distribution will approximately follow a normal distribution

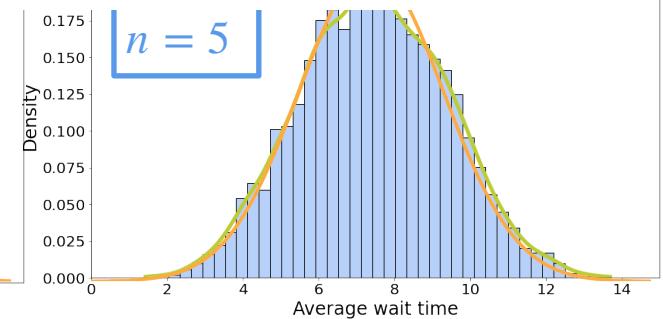
$n = 3$



$n = 4$

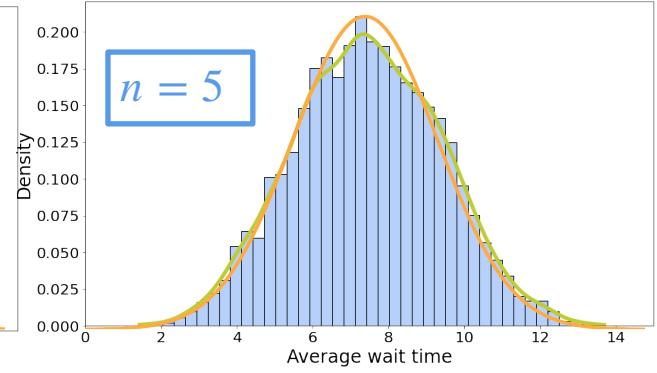
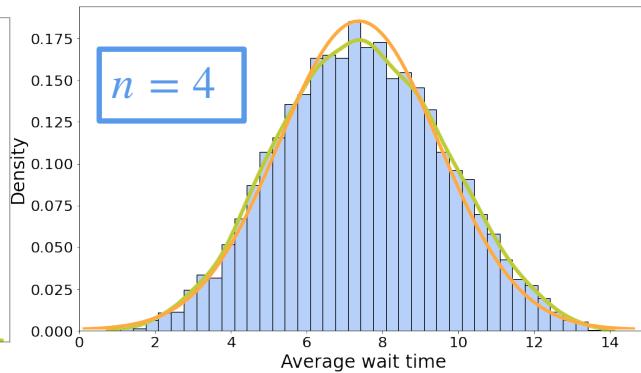
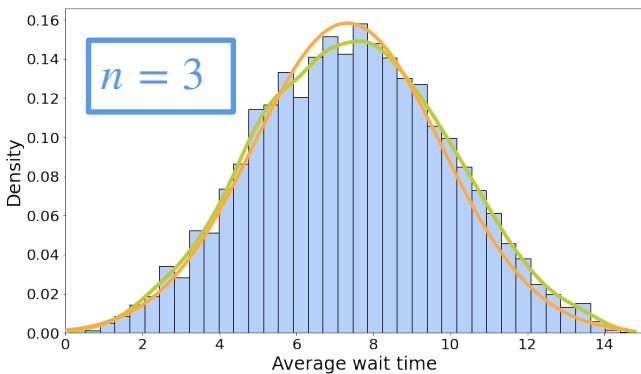


$n = 5$



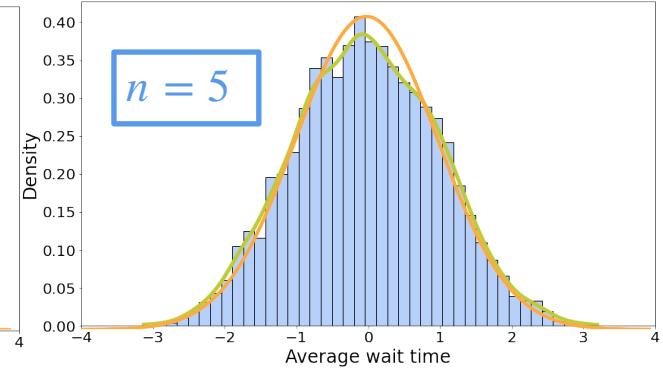
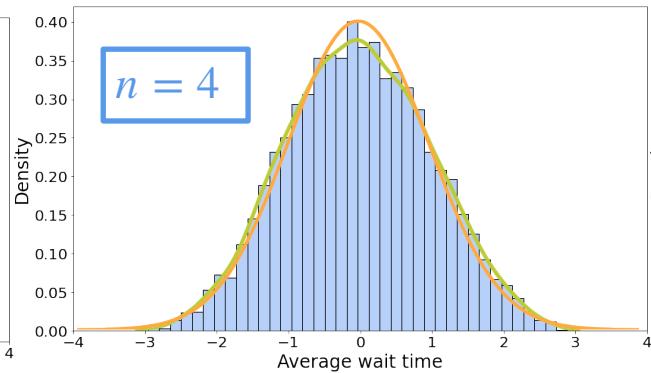
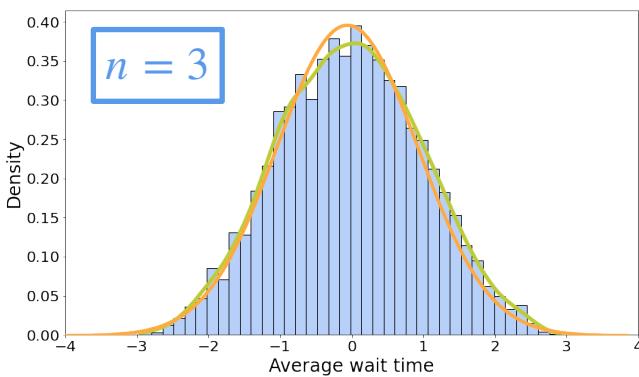
# Central Limit Theorem (CLT) - Example 2

$$\frac{Y_n - 7.5}{\sqrt{18.75/n}}$$



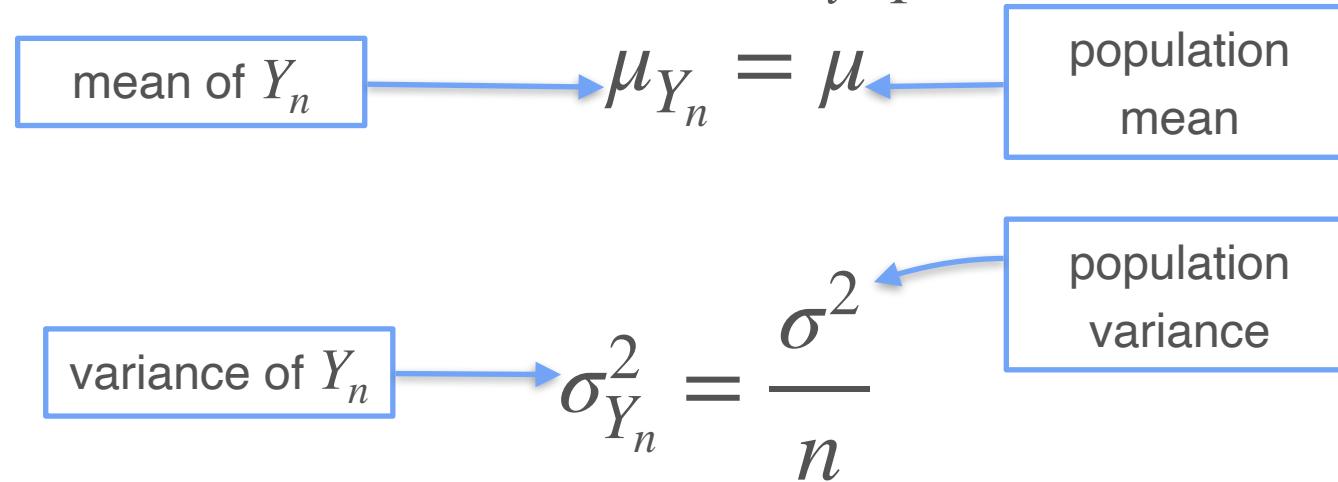
# Central Limit Theorem (CLT) - Example 2

$$\frac{Y_n - 7.5}{\sqrt{18.75/n}} \xrightarrow{n \uparrow} \mathcal{N}(0,1)$$



# Central Limit Theorem (CLT) - Example 2

$$Y_n = \frac{1}{n} \sum_{i=1}^n X_i$$



# Central Limit Theorem (CLT) - Formal Definition

$$\text{As } n \rightarrow \infty \quad \frac{\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X]}{\sigma_X} \sqrt{n} \sim \mathcal{N}(0, 1^2)$$

$$\text{As } n \rightarrow \infty \quad \frac{1}{\cancel{\sqrt{n}}} \left( \frac{\sum_{i=1}^n X_i - \cancel{\frac{1}{n} n \mathbb{E}[X]}}{\sigma_X} \right) \cancel{\sqrt{n}} \sim \mathcal{N}(0, 1^2)$$

# Central Limit Theorem (CLT) - Formal Definition

$$\text{As } n \rightarrow \infty \quad \frac{\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X]}{\sigma_X} \sqrt{n} \sim \mathcal{N}(0, 1^2)$$

$$\text{As } n \rightarrow \infty \quad \frac{\sum_{i=1}^n X_i - n\mathbb{E}[X]}{\sqrt{n}\sigma_X} \sim \mathcal{N}(0, 1^2)$$

# W3 Lesson 2



DeepLearning.AI

# Point Estimation

---

## Point Estimation



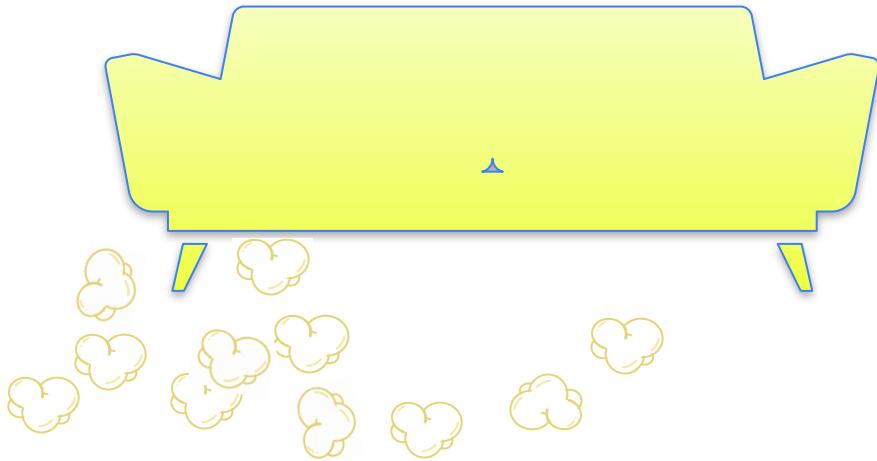
DeepLearning.AI

## Point Estimation

---

**Maximum Likelihood  
Estimation: Motivation**

# There's Popcorn on the Floor. What Happened?



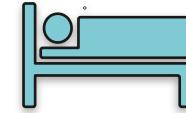
Movies



Board Games



Nap

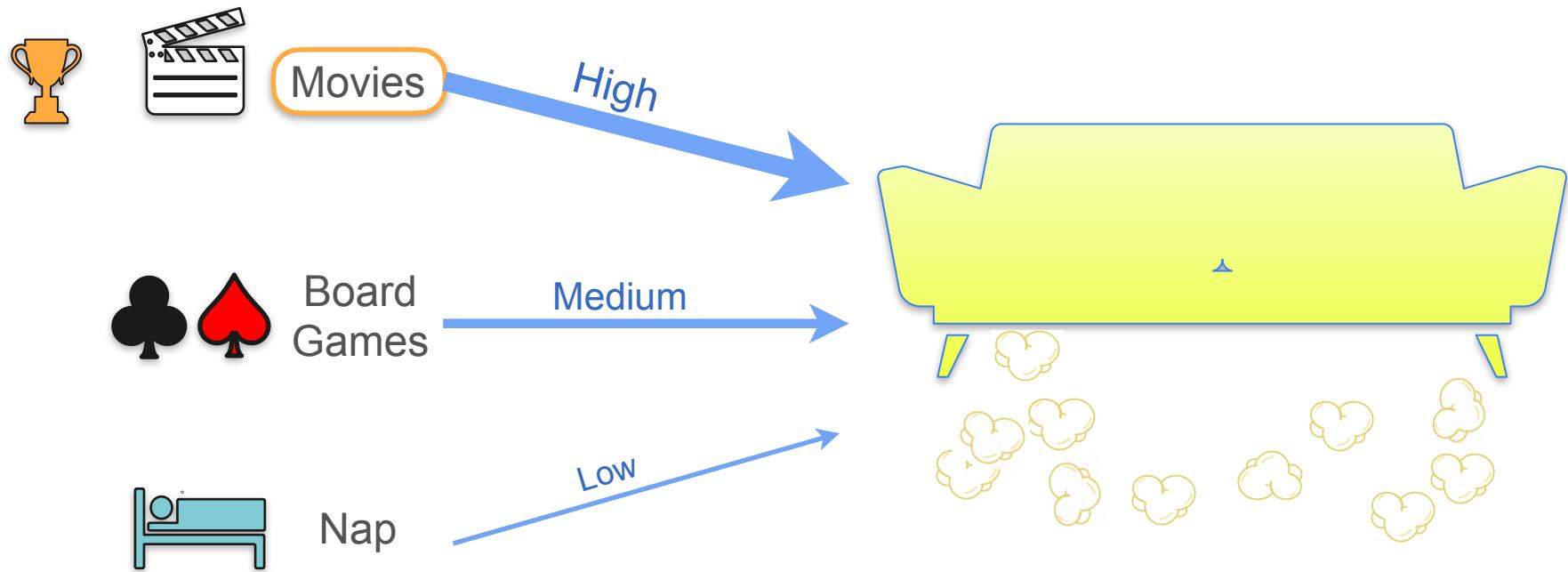


# Quiz

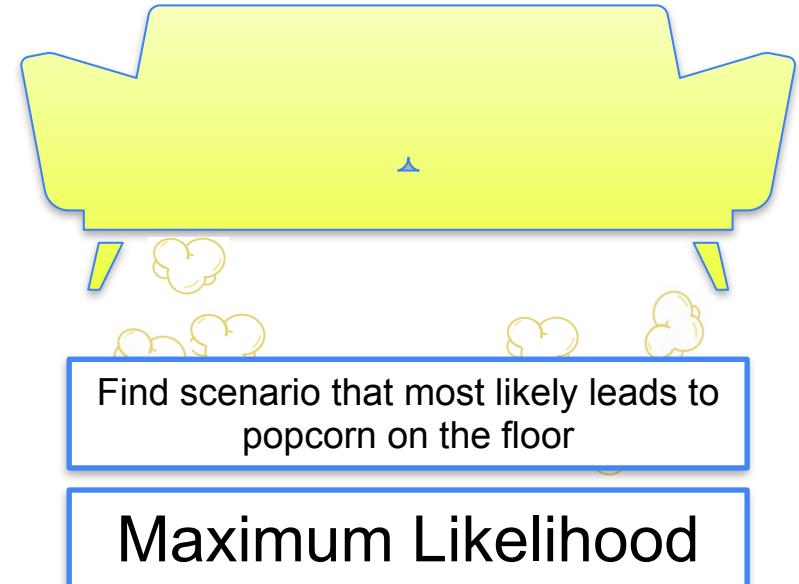
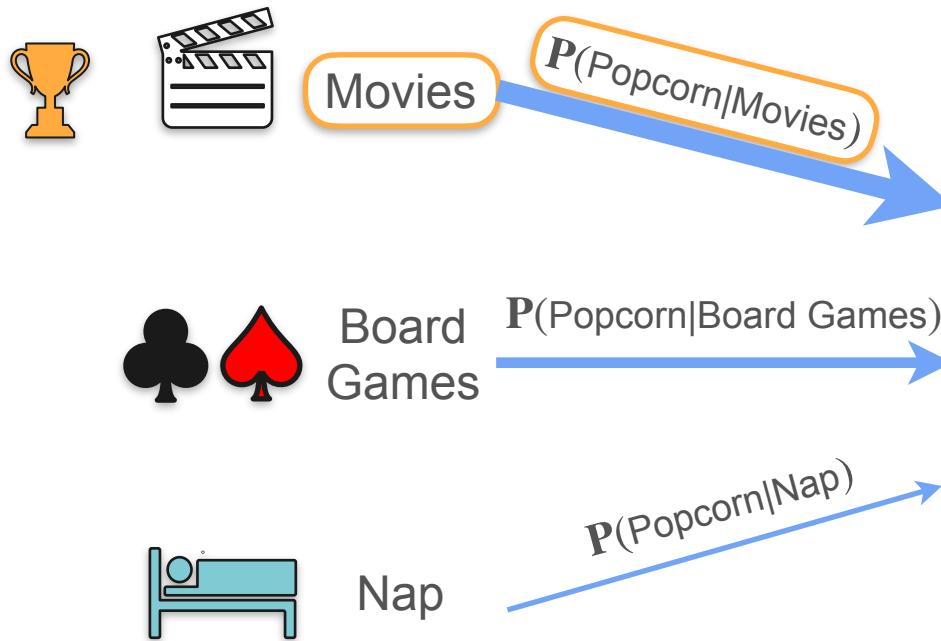
What do you think happened?

- A. People were watching a movie
- B. People were playing boardgames
- C. People were taking a nap

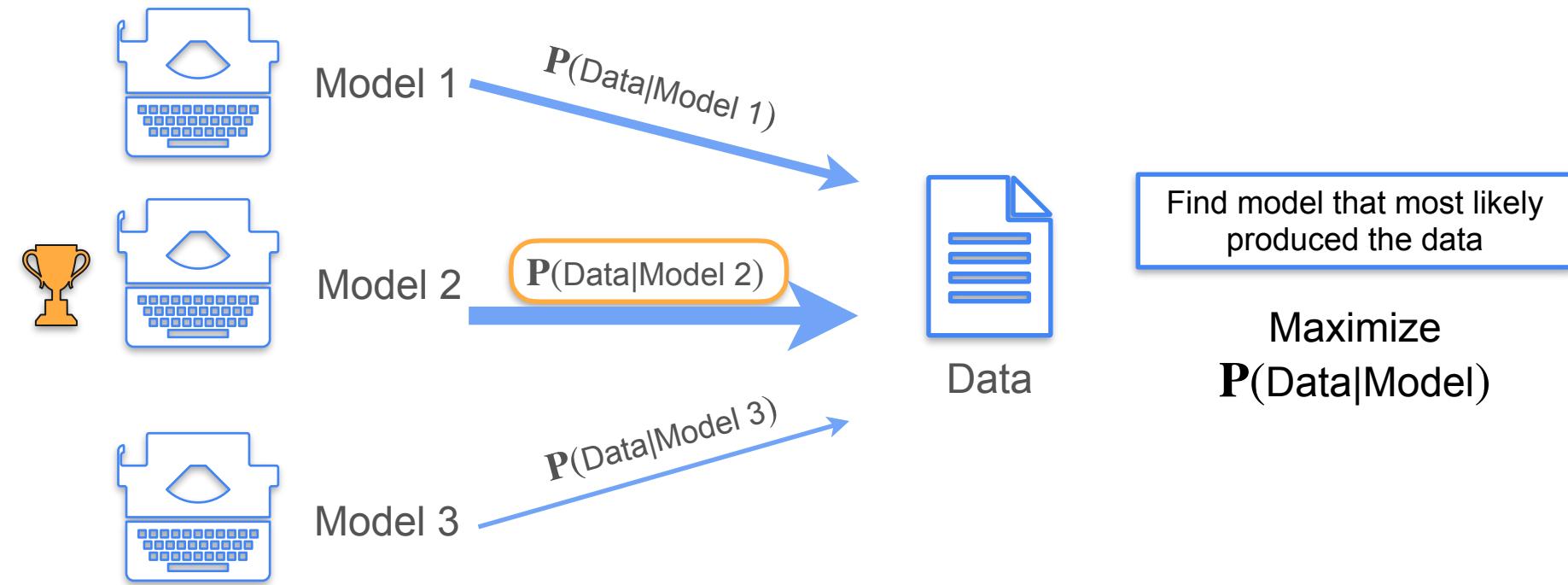
# There's Popcorn on the Floor. What Happened?



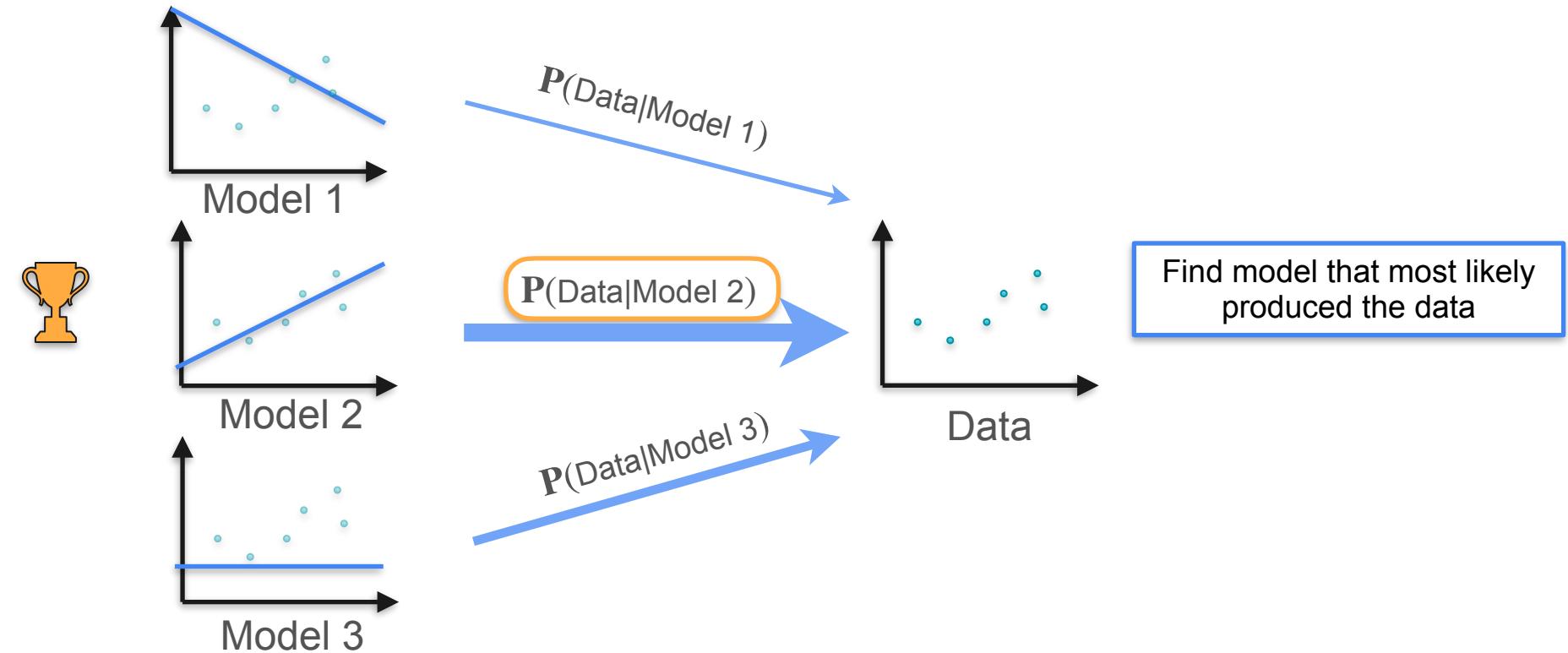
# There's Popcorn on the Floor. What Happened?



# Maximum Likelihood



# Example: Linear Regression





DeepLearning.AI

## Point Estimation

---

**MLE: Bernoulli Example**

# Maximum Likelihood: Bernoulli Example



Coin 1



$$P(H) = 0.7$$

$$P(T) = 0.3$$

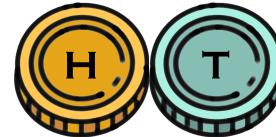
Coin 2



$$P(H) = 0.5$$

$$P(T) = 0.5$$

Coin 3



$$P(H) = 0.3$$

$$P(T) = 0.7$$

# Maximum Likelihood: Bernoulli Example



<b>Coin 1</b>	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.3	0.3 = 0.0051
---------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	--------------

<b>Coin 2</b>	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5 = 0.0010
---------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	--------------

<b>Coin 3</b>	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.30	0.7	0.7 = 0.00003
---------------	-----	-----	-----	-----	-----	-----	-----	------	-----	---------------

# Maximum Likelihood: Bernoulli Example



**Coin 1**

$$P(H) = 0.7$$

$$P(8H, 2T | C_1) = 0.0051$$

**Coin 2**

$$P(H) = 0.5$$

$$P(8H, 2T | C_2) = 0.0010$$

**Coin 3**

$$P(H) = 0.3$$

$$P(8H, 2T | C_3) = 0.00003$$

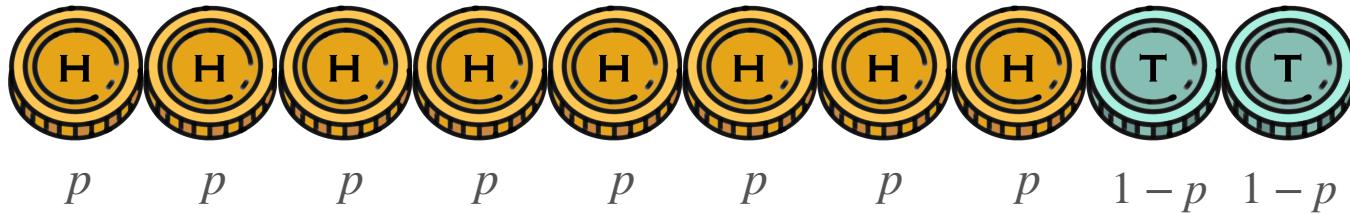
8H 2T



Find coin that most likely produced the 8 heads and 2 tails

Maximize  
 $P(8H, 2T | \text{Coin})$

# Maximum Likelihood: Bernoulli Example



Can you do any better?

$$p = \mathbf{P}(H)$$

$$p^8(1 - p)^2$$

You want  $p$  that maximizes the chances of seeing 8H

# Maximum Likelihood: Bernoulli Example



$$p = \mathbf{P}(H) \quad \text{Likelihood} \quad L(p; 8H) = p^8(1-p)^2 \quad \text{Function of } p$$

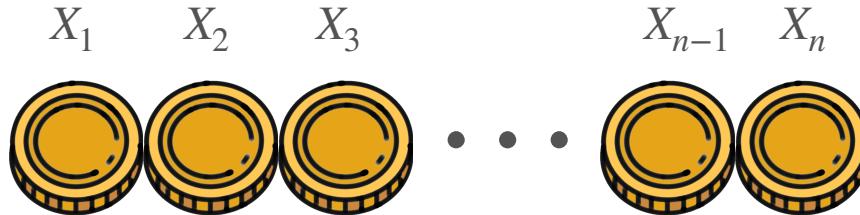
You want  $p$  that maximizes the chances of seeing 8H

$$\text{Log-likelihood} \quad \ell(p; 8H) = \log((p^8(1-p)^2)) = 8\log(p) + 2\log(1-p)$$

$$\frac{d}{dp} (8\log(p) + 2\log(1-p)) = \frac{8}{p} + \frac{2}{1-p}(-1) = 0 \rightarrow \hat{p} = \frac{8}{10}$$

# Maximum Likelihood: Bernoulli Example

$n$  coins  
 $k$  heads



$$\mathbf{X} = (X_1, \dots, X_n)$$

$$X_i \stackrel{i.i.d}{\sim} \text{Bernoulli}(p)$$

Likelihood

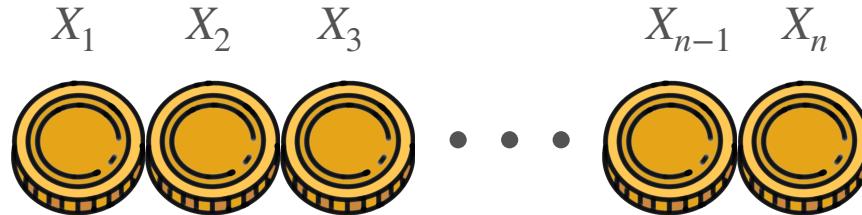
$$L(p; \mathbf{x}) = P_p(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n p_{X_i}(x_i) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$$

If  $x_i = 1$ ,  $p^{[x_i]}(1-p)^{[1-x_i]} = p$   
If  $x_i = 0$ ,  $p^{[x_i]}(1-p)^{[1-x_i]} = (1-p)$

$$\sum_{i=1}^n x_i = \# \text{ heads}$$

$$n - \sum_{i=1}^n x_i = \# \text{ tails}$$

# Maximum Likelihood: Bernoulli Example



$$\mathbf{X} = (X_1, \dots, X_n)$$

$$X_i \stackrel{i.i.d}{\sim} \text{Bernoulli}(p)$$

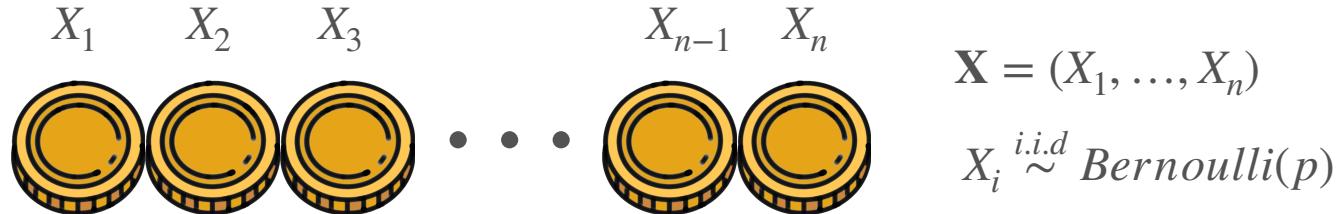
Likelihood

$$L(p; \mathbf{x}) = P_p(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n p_{X_i}(x_i) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\left(\sum_{i=1}^n x_i\right)}(1-p)^{\left(n - \sum_{i=1}^n x_i\right)}$$

Log-likelihood

$$\ell(p; \mathbf{x}) = \log \left( (p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}) \right) = \left( \sum_{i=1}^n x_i \right) \log(p) + \left( n - \sum_{i=1}^n x_i \right) \log(1-p)$$

# Maximum Likelihood: Bernoulli Example



$$\ell(p; \mathbf{x}) = \log \left( (p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}) \right) = \left( \sum_{i=1}^n x_i \right) \log(p) + \left( n - \sum_{i=1}^n x_i \right) \log(1-p)$$

Find the maximum!

$$\begin{aligned} \frac{d}{dp} \ell(p; \mathbf{x}) &= \frac{d}{dp} \left( \left( \sum_{i=1}^n x_i \right) \log(p) + \left( n - \sum_{i=1}^n x_i \right) \log(1-p) \right) \\ &= \frac{\sum_{i=1}^n x_i}{p} + \frac{n - \sum_{i=1}^n x_i}{1-p} (-1) = 0 \end{aligned} \quad \rightarrow \quad \hat{p} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$



DeepLearning.AI

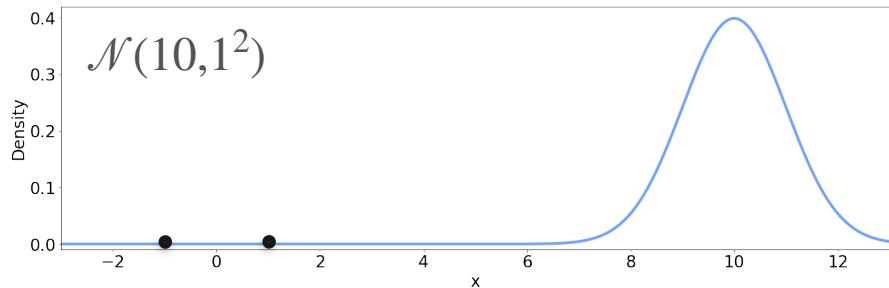
## Point Estimation

---

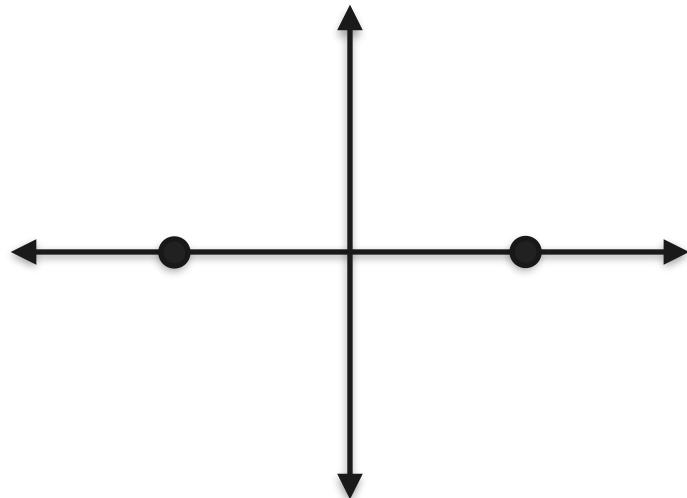
### MLE: Gaussian Example

# Maximum Likelihood: Gaussian Example

Candidates

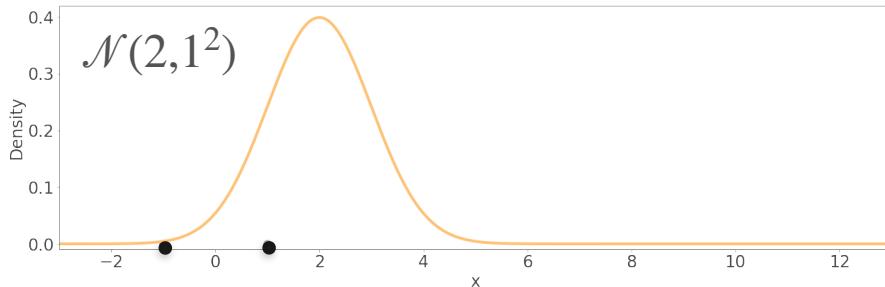
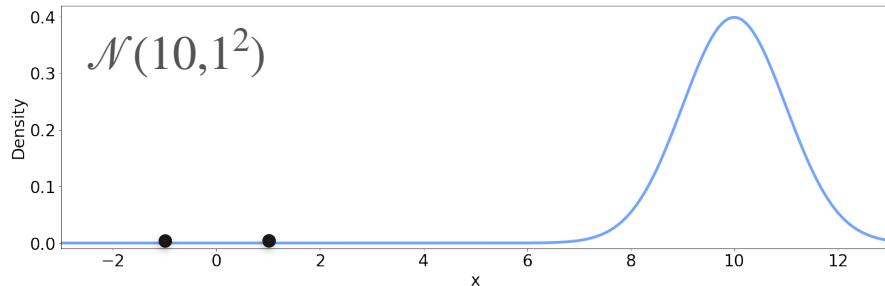


Observations

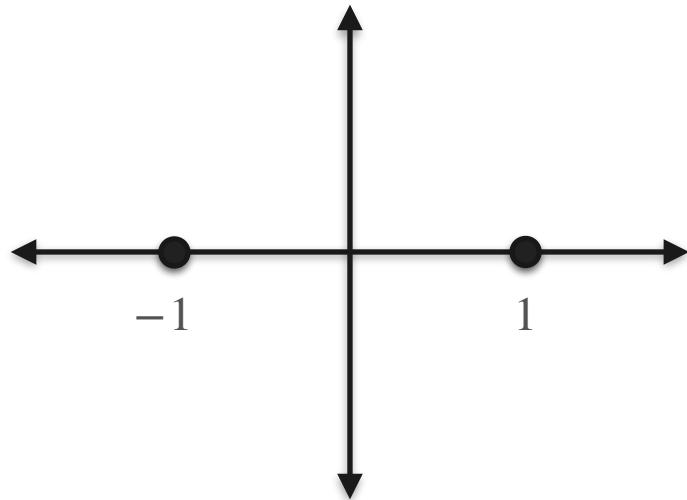


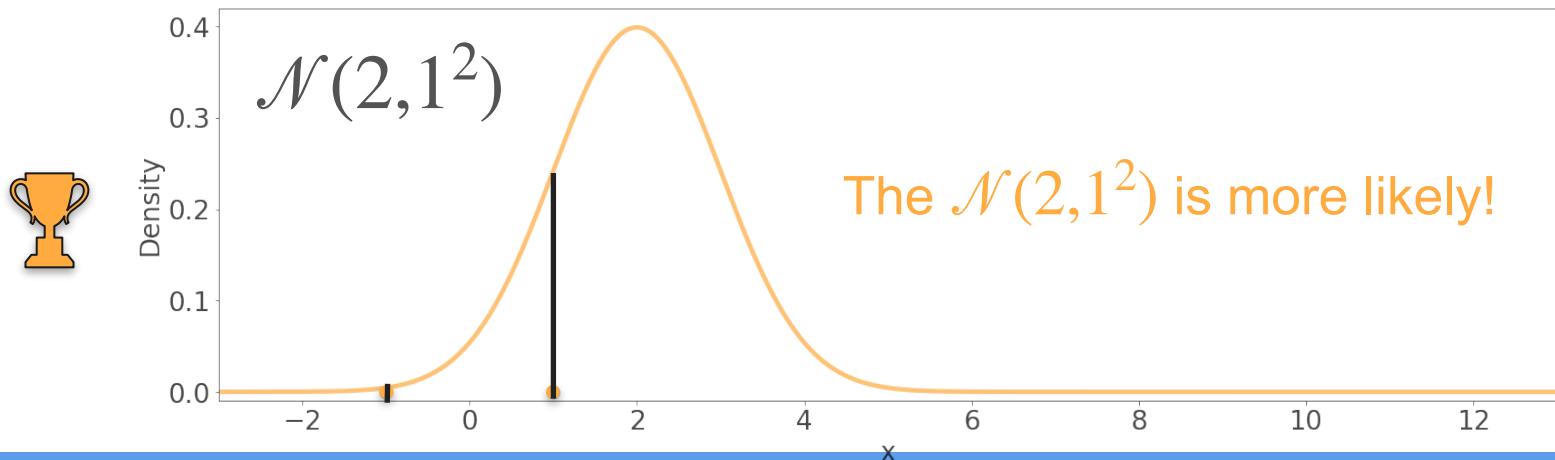
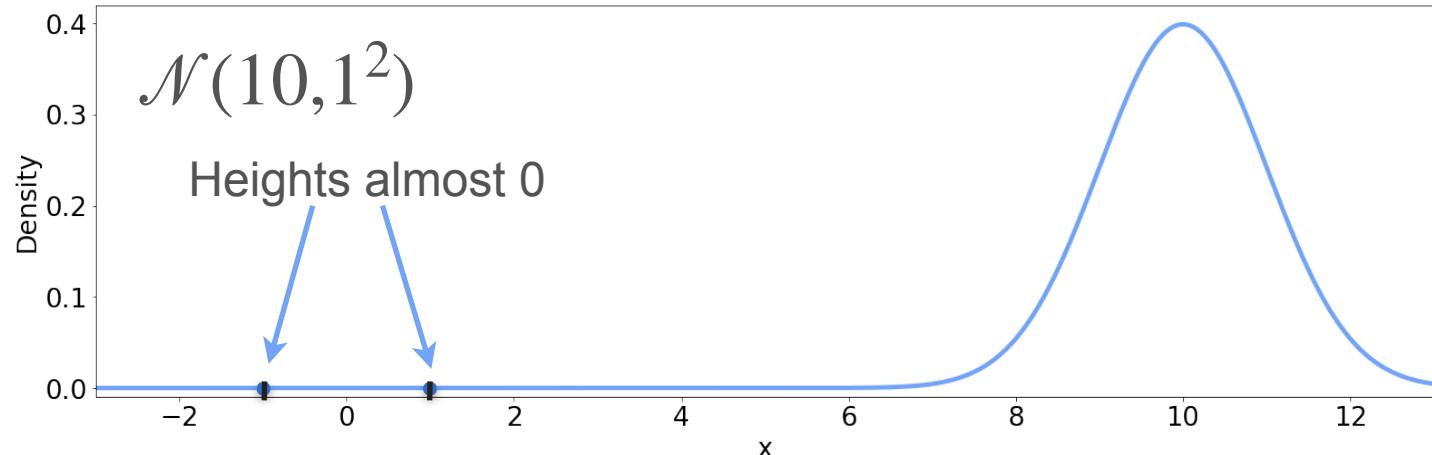
# Maximum Likelihood: Gaussian Example

Candidates

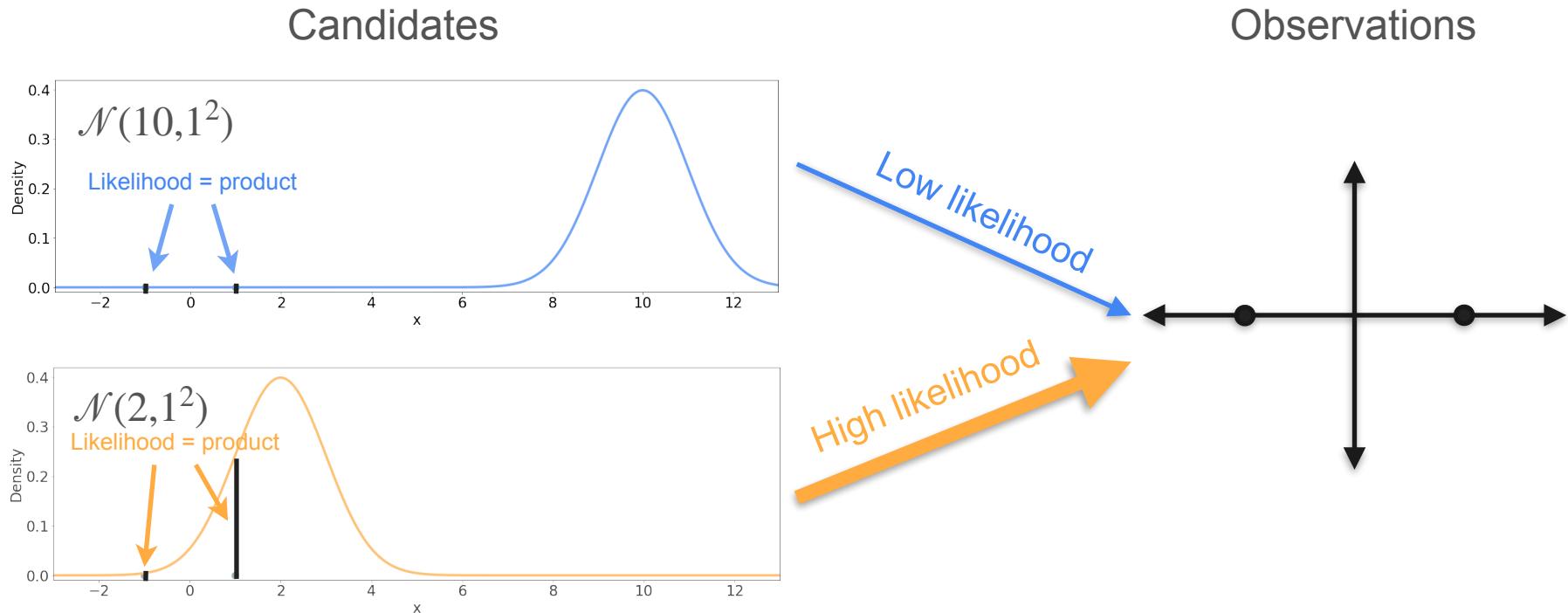


Observations





# Maximum Likelihood: Gaussian Example



# Gaussians With Three Different Means

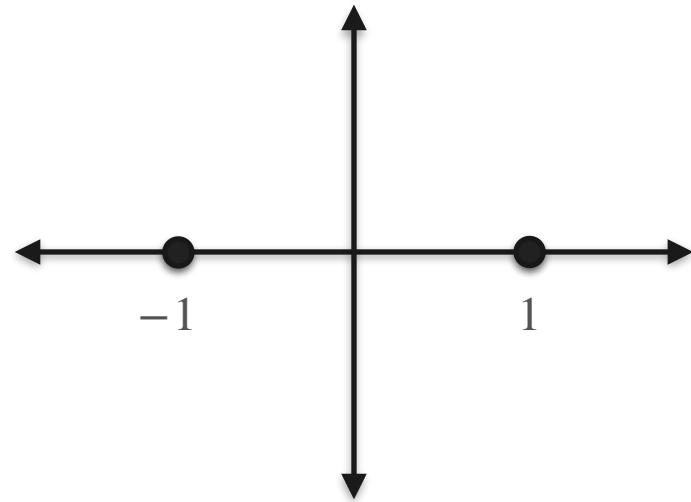
Candidates

$$\mathcal{N}(-1, 1^2)$$

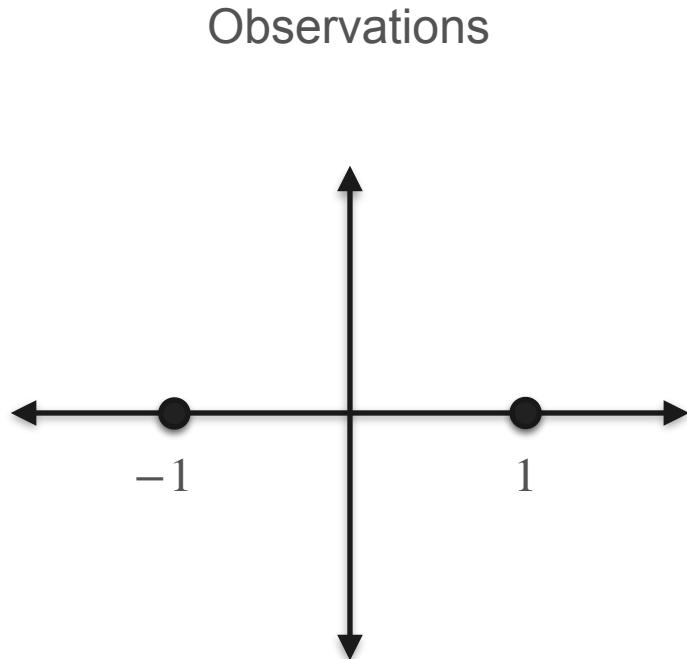
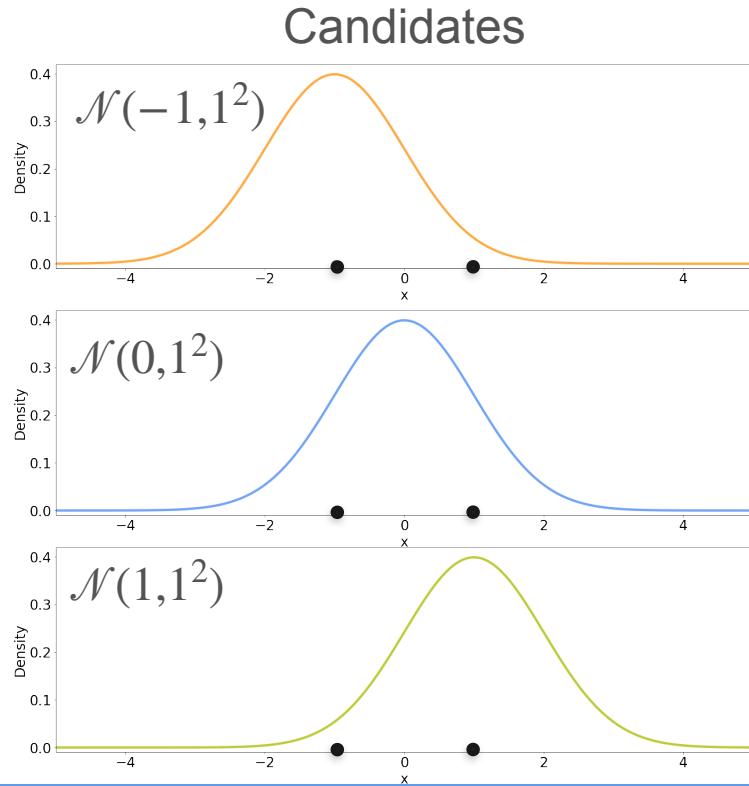
$$\mathcal{N}(0, 1^2)$$

$$\mathcal{N}(1, 1^2)$$

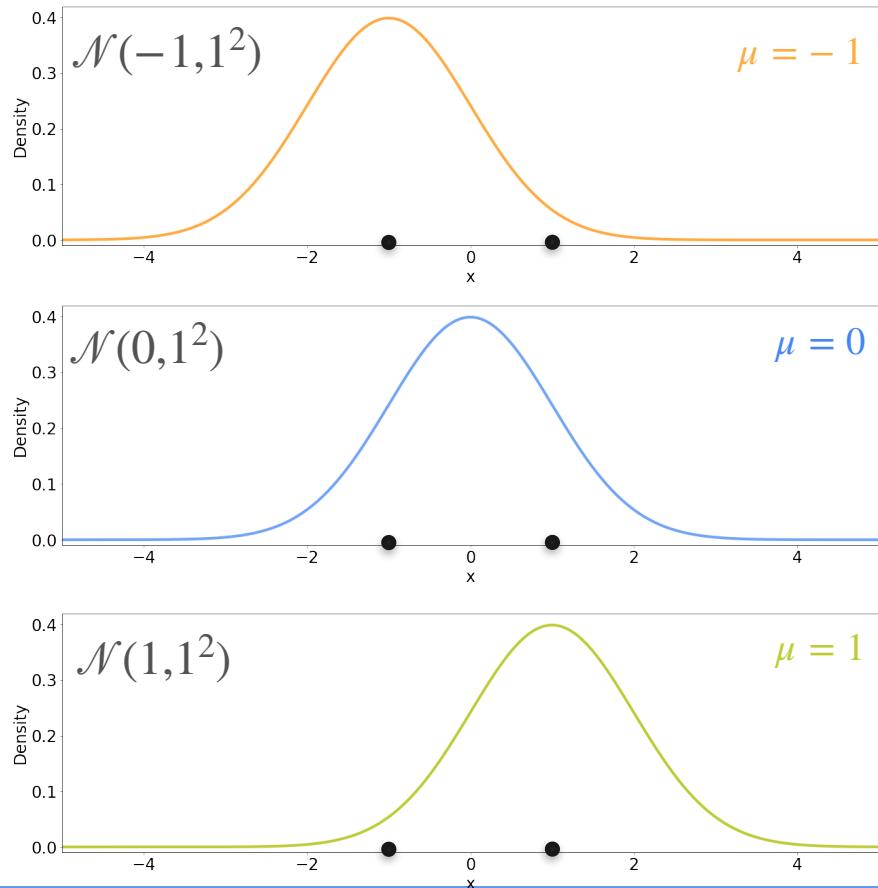
Observations



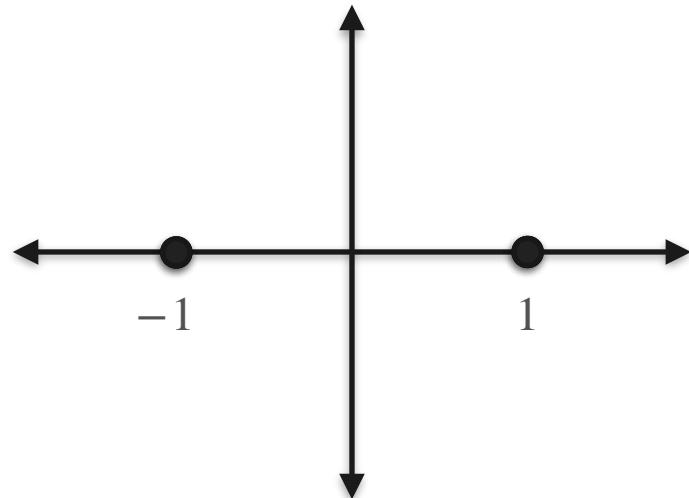
# Gaussians With Three Different Means



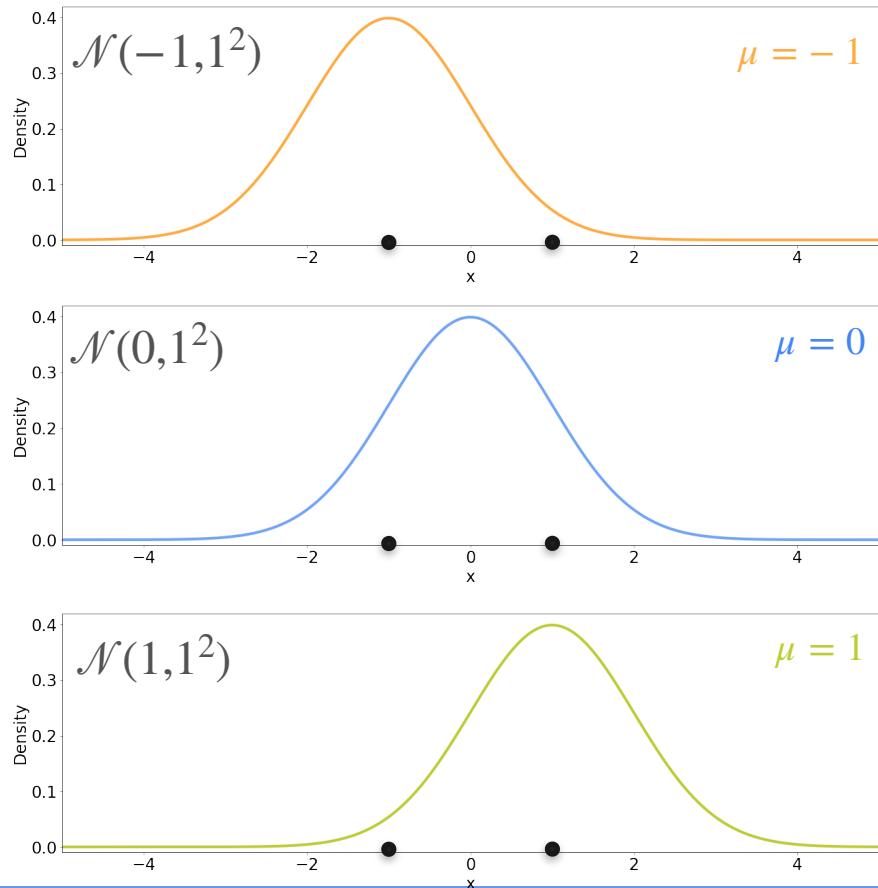
## Candidates



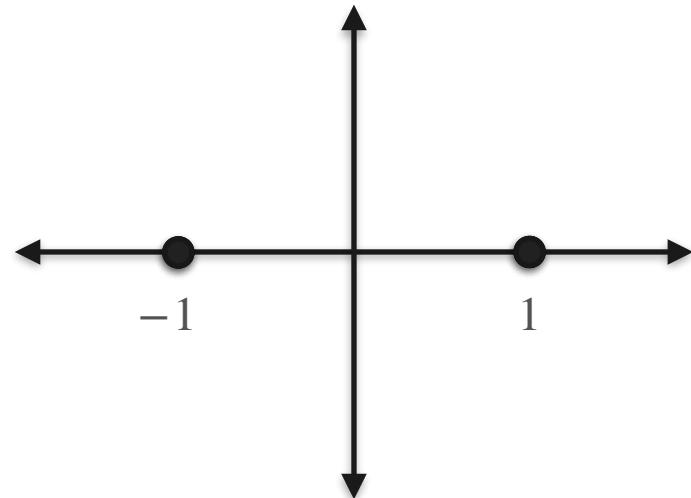
## Observations



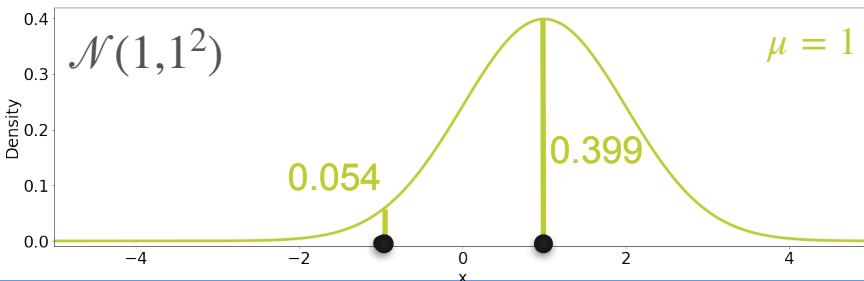
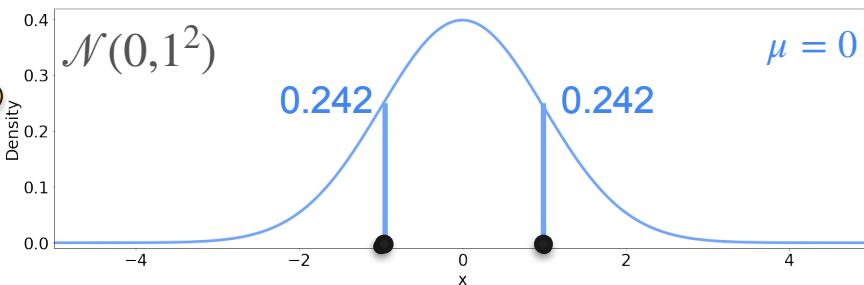
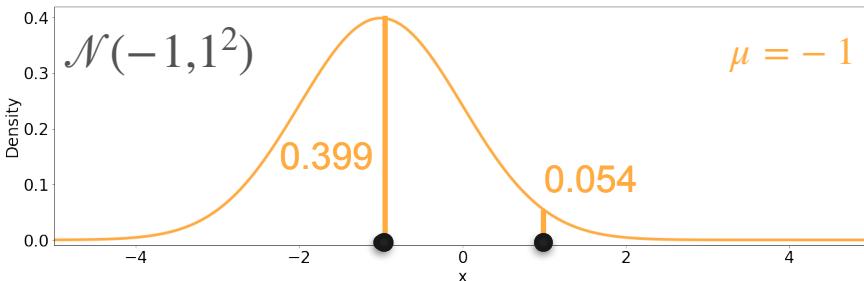
## Candidates



## Observations



## Candidates



## Observations

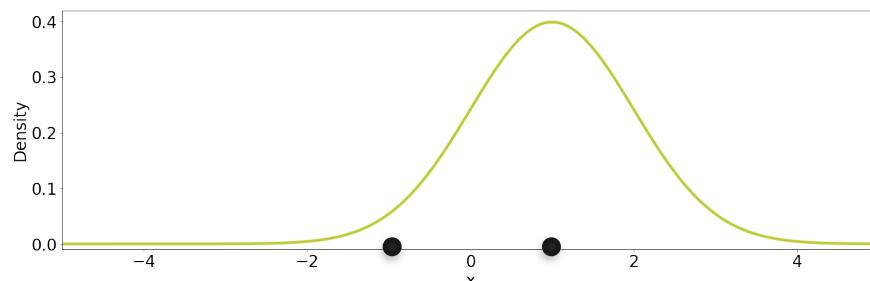
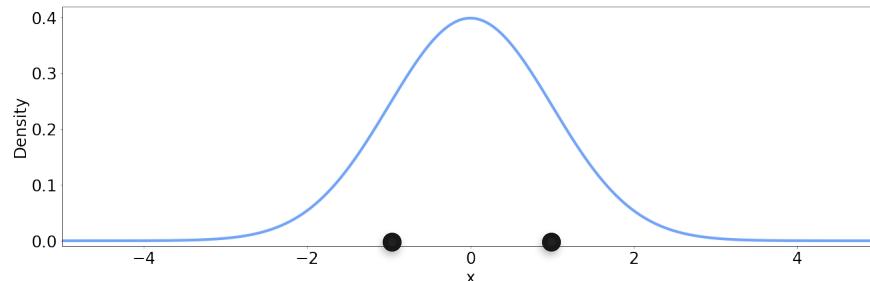
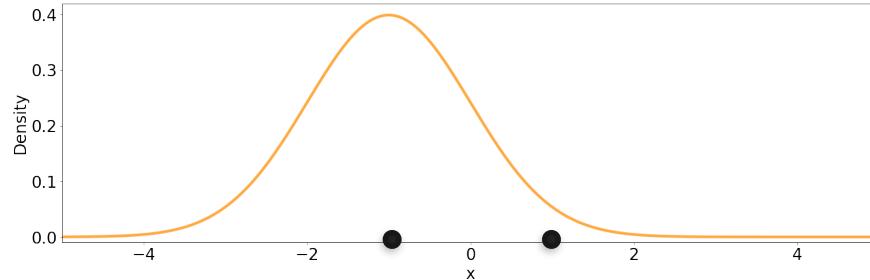
= 0.022

The  $\mathcal{N}(0, 1)$  is more likely!

= 0.059

= 0.022

# Candidates

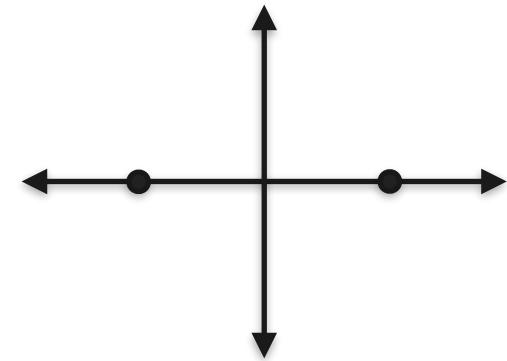


Likelihood = 0.022

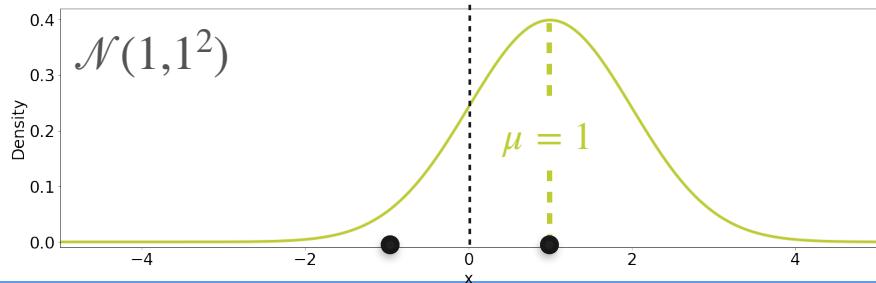
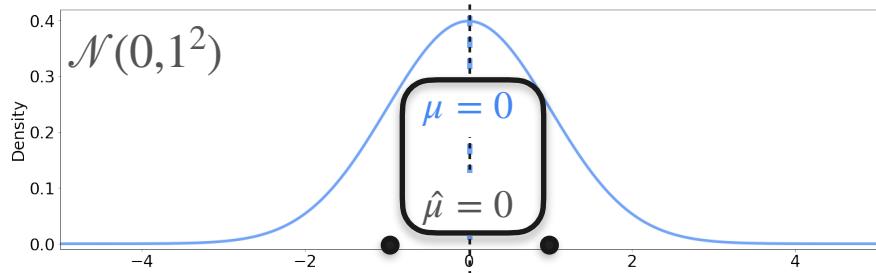
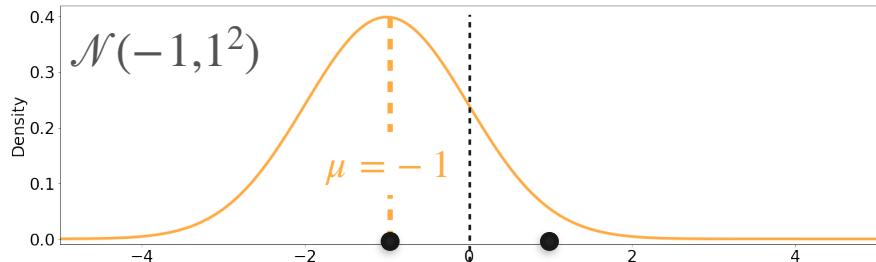
Likelihood = 0.059

Likelihood = 0.022

# Observations



## Candidates



The best distribution is the one where  
the **mean** of the distribution is the  
**mean** of the sample

# Gaussians With Three Different Variance

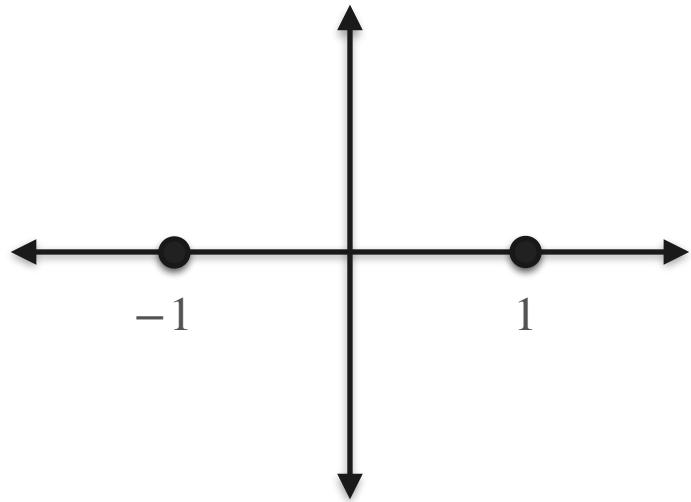
Candidates

$$\mathcal{N}(0,0.5^2)$$

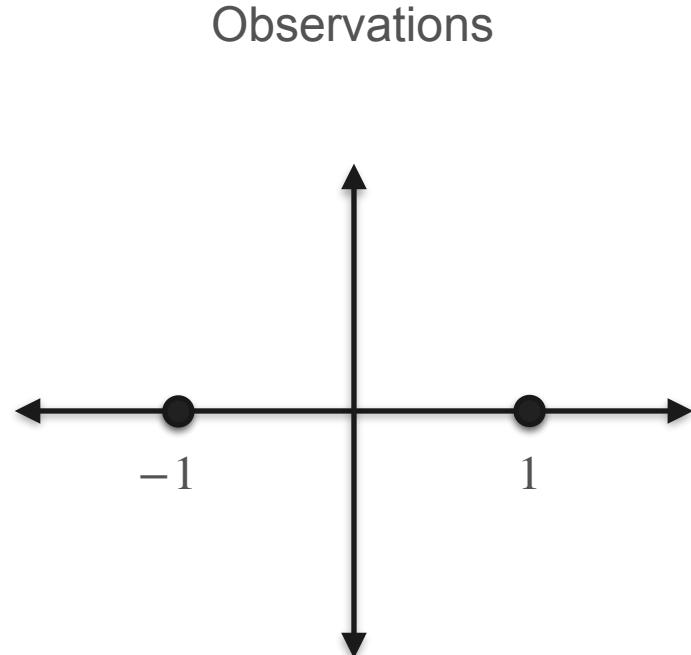
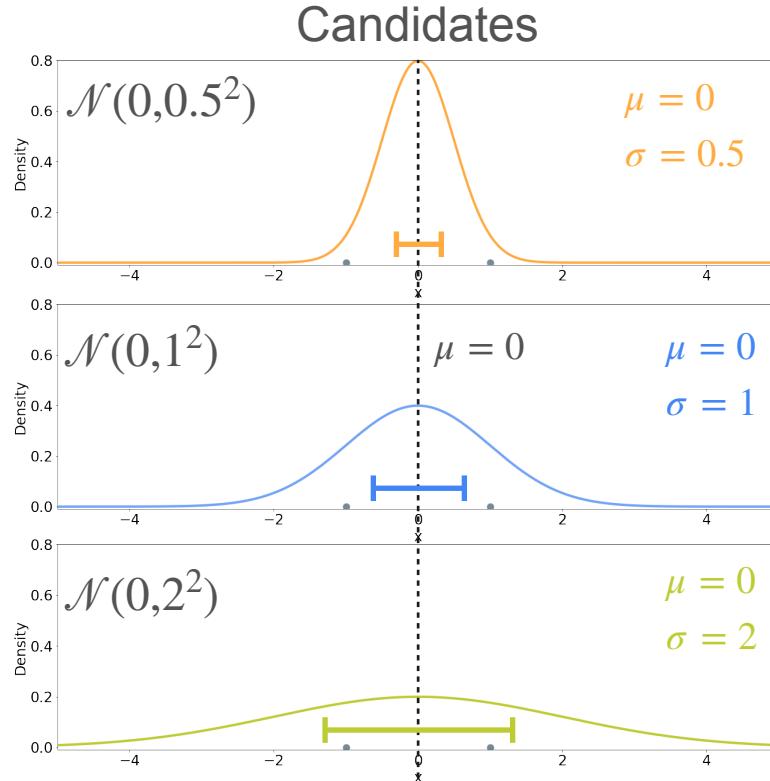
$$\mathcal{N}(0,1^2)$$

$$\mathcal{N}(0,2^2)$$

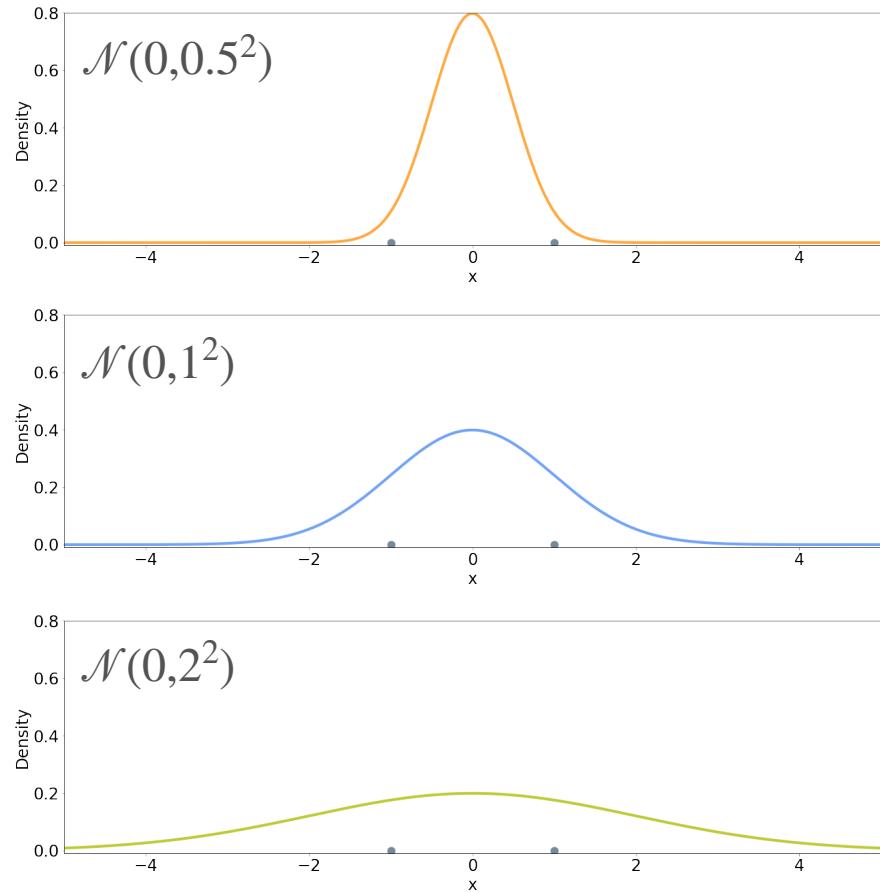
Observations



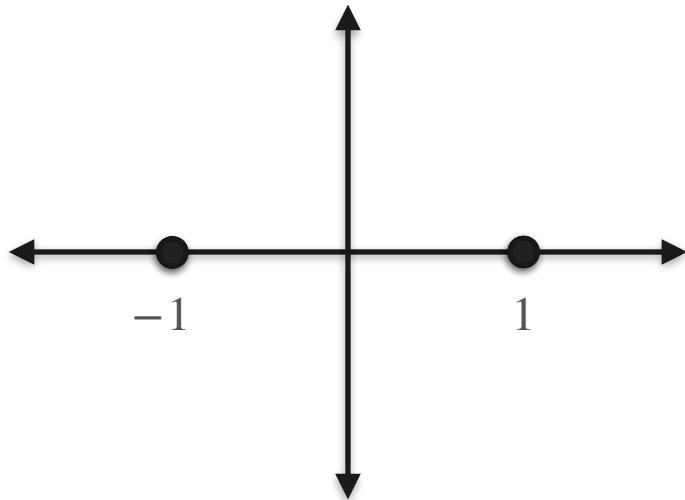
# Gaussians With Three Different Variance



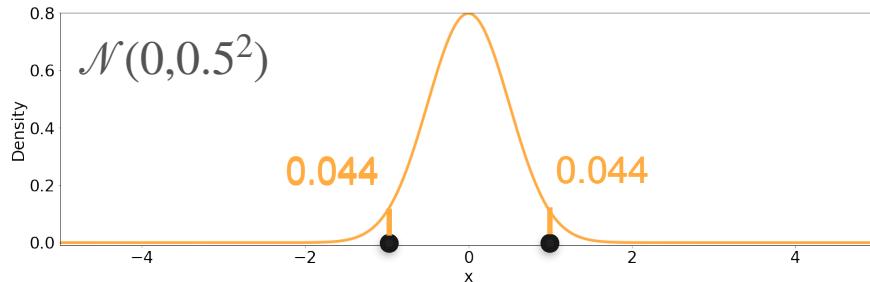
## Candidates



## Observations



## Candidates

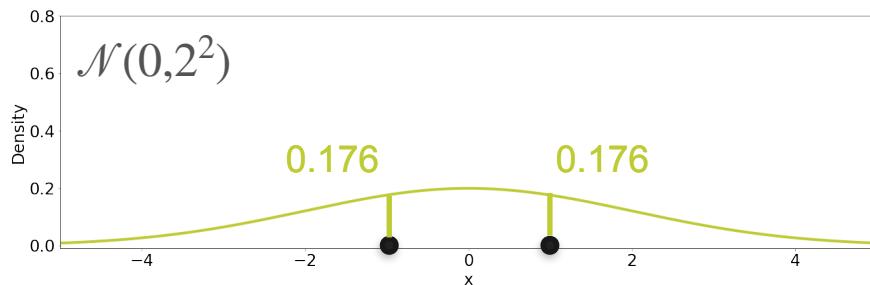
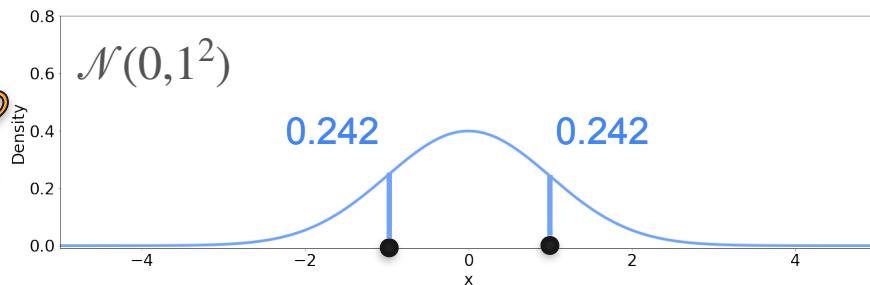


## Observations

= 0.002

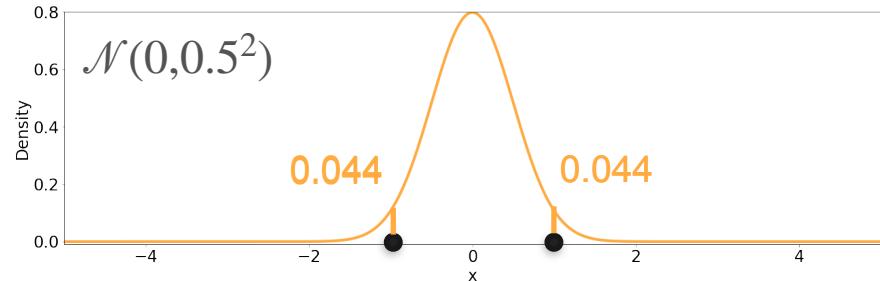
The  $\mathcal{N}(0, 1^2)$  is more likely!

= 0.059



= 0.031

## Candidates

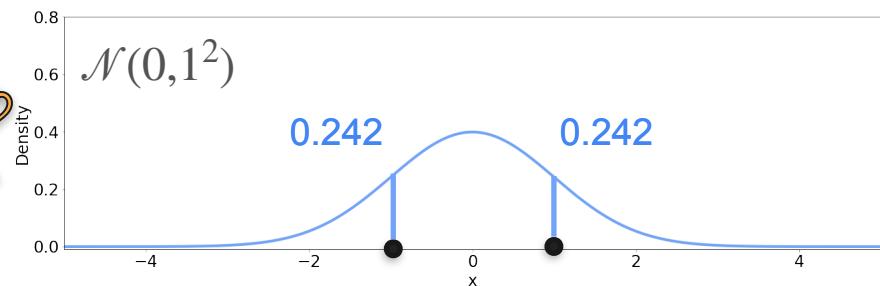


0.044

0.044



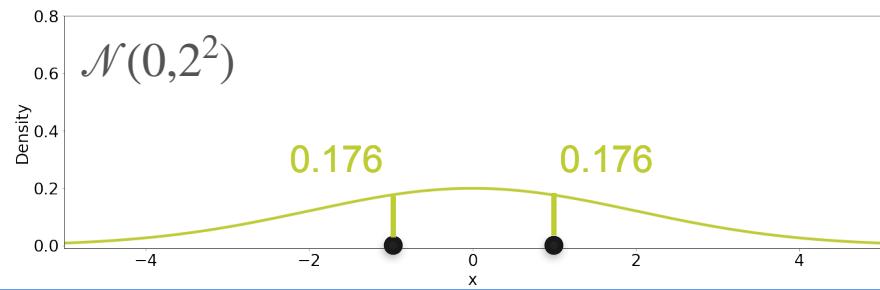
## $\mathcal{N}(0, 1^2)$



0.242

0.242

## $\mathcal{N}(0, 2^2)$



0.176

0.176

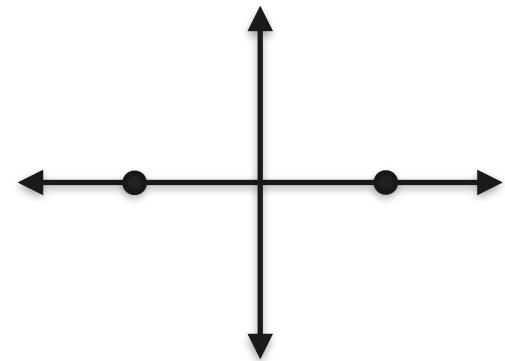
## Observations

Likelihood = 0.002

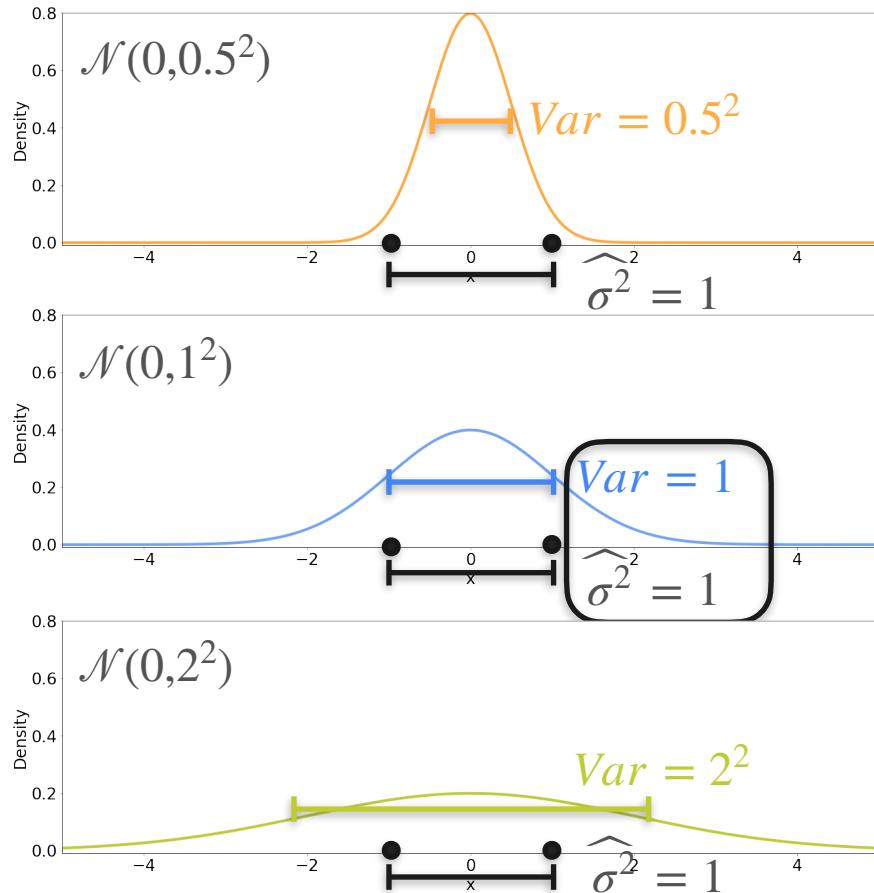
Likelihood = 0.059

Likelihood = 0.031

## Observations



## Candidates



## Observations

Variance of the observations

$$\widehat{\sigma}^2 = \frac{1}{2} ((0 - 1)^2 + (0 + 1)^2) = 1$$

The best distribution is the one where the **variance** of the distribution is the **variance** of the sample



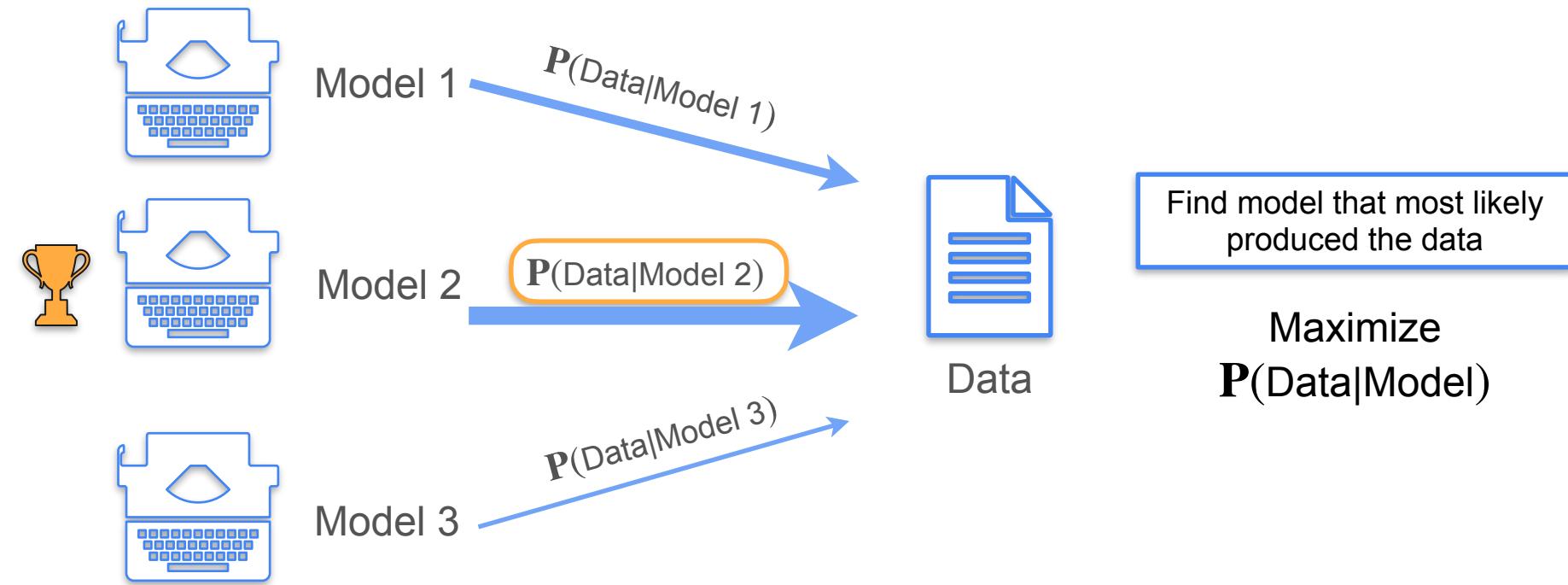
DeepLearning.AI

## Point Estimation

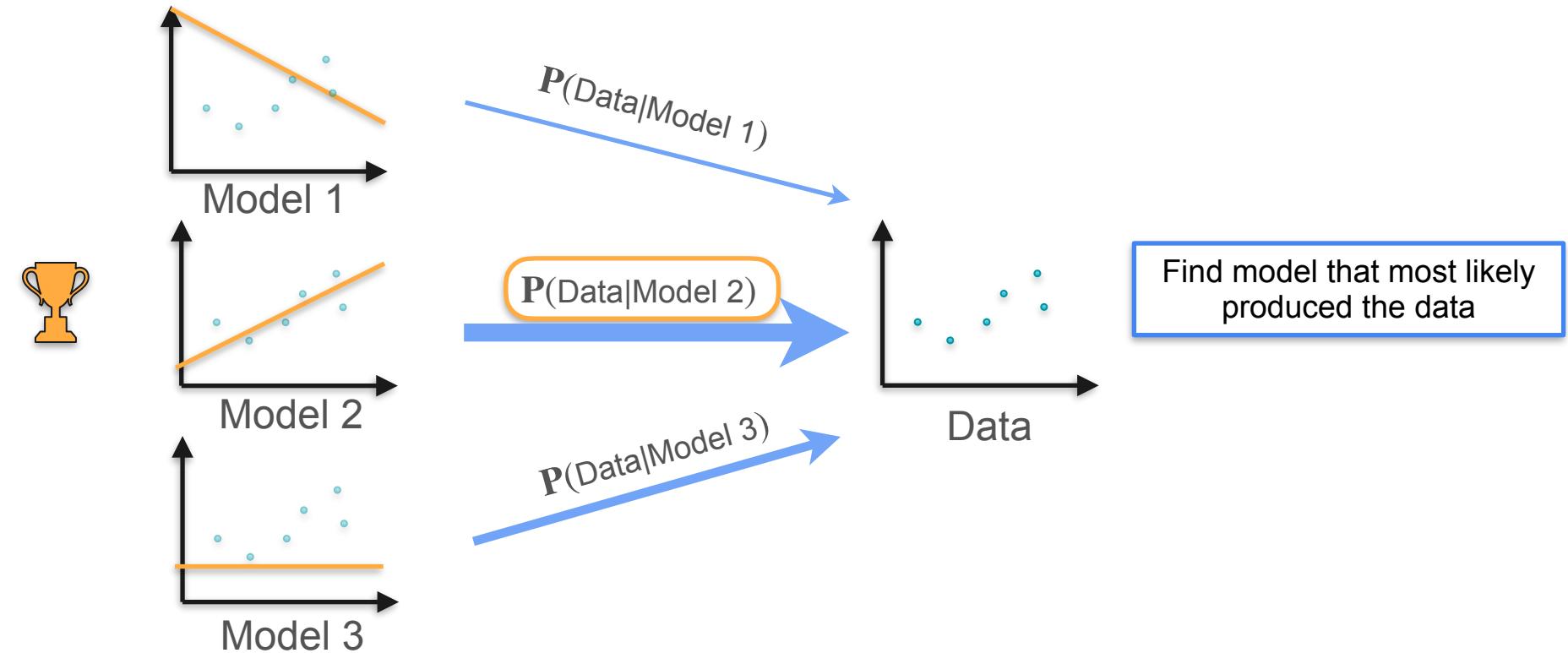
---

# MLE: Linear Regression

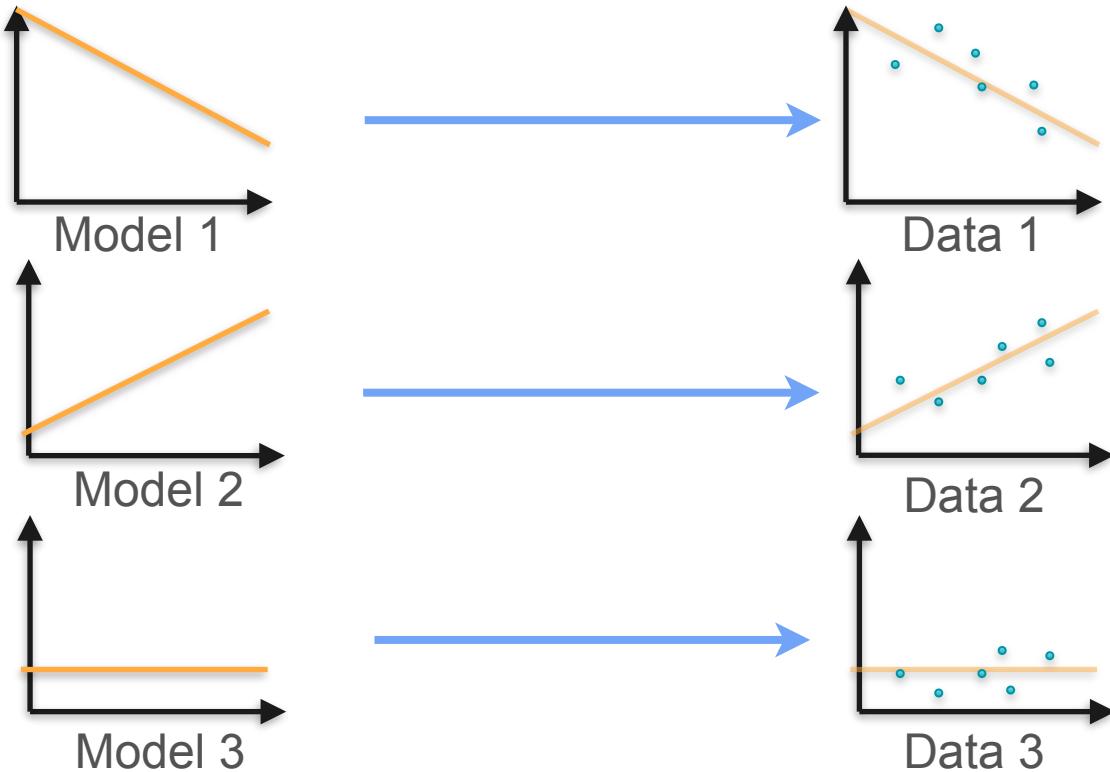
# Maximum Likelihood



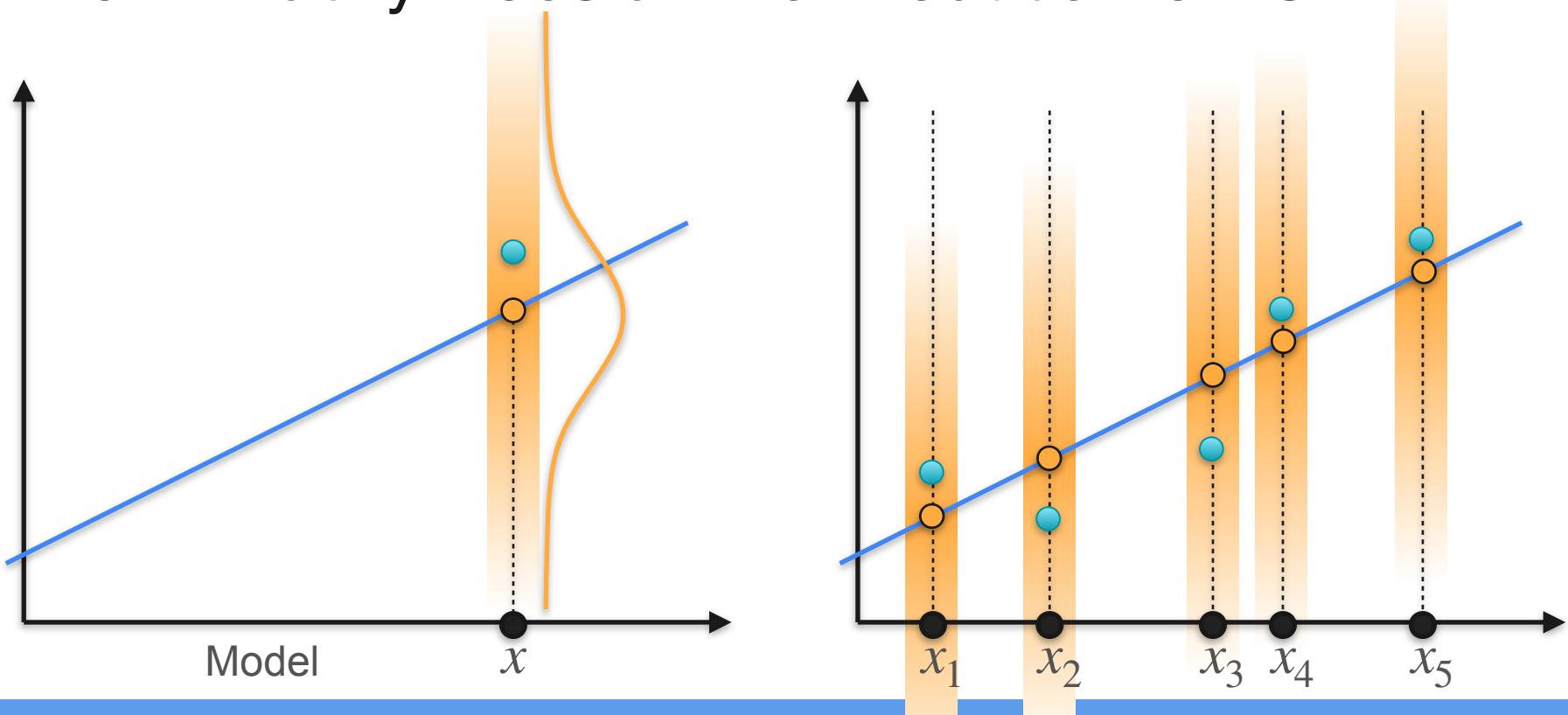
# Example: Linear Regression



# How Exactly Does a Line Produce Points?



# How Exactly Does a Line Produce Points?

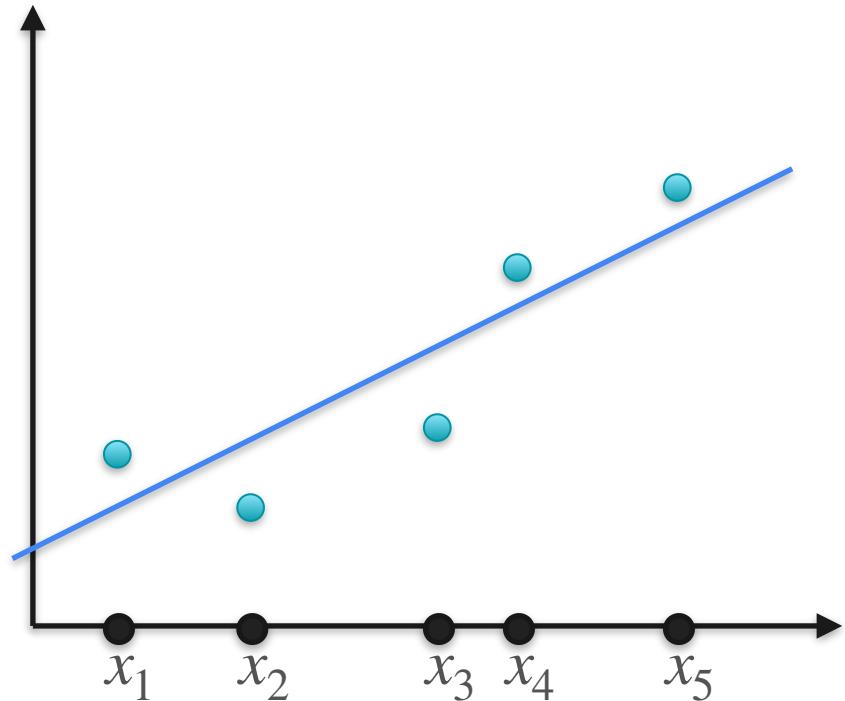


# Linear Regression

Line that best produced the points

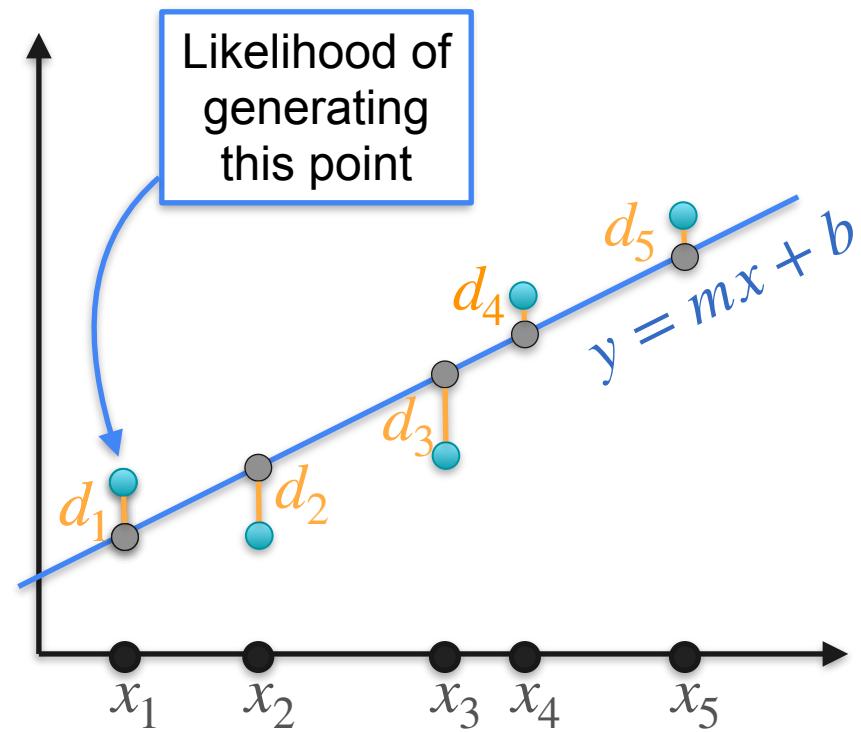
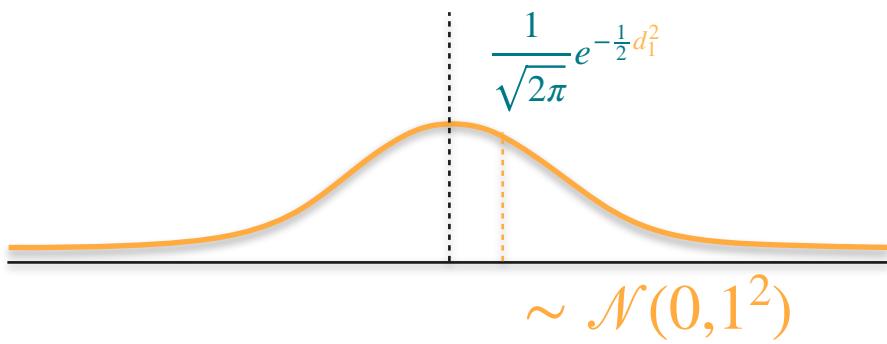
How?

Line that best fits the data  
(linear regression)



# Linear Regression and Likelihood

Likelihood:



# Linear Regression and Likelihood

Likelihood:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_2^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_3^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_4^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_5^2}$$

Maximize

$$e^{-\frac{1}{2}d_1^2} \cdot e^{-\frac{1}{2}d_2^2} \cdot e^{-\frac{1}{2}d_3^2} \cdot e^{-\frac{1}{2}d_4^2} \cdot e^{-\frac{1}{2}d_5^2}$$

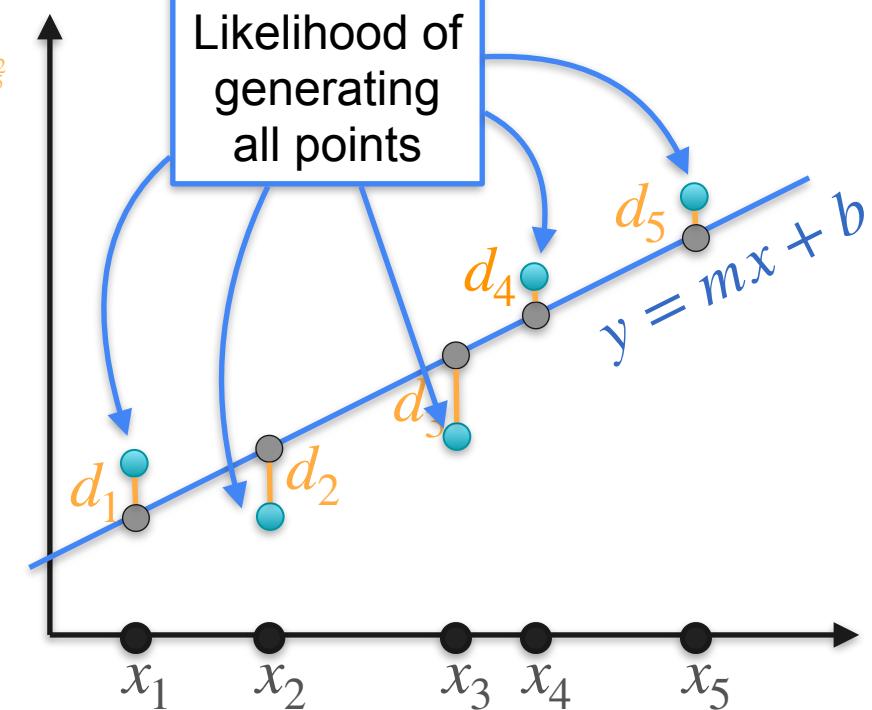
$$\cancel{e^{-\frac{1}{2}(d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2)}}$$

Minimize

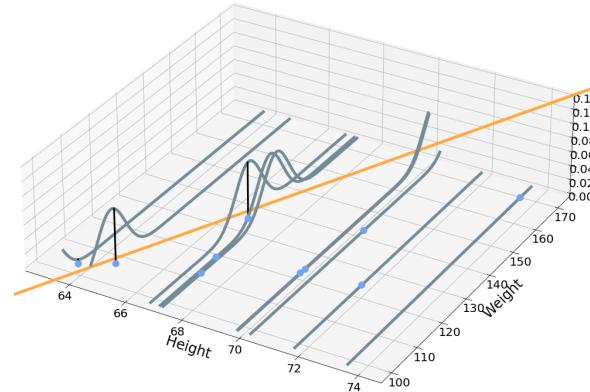
$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2$$

Linear regression!

Least squares error!

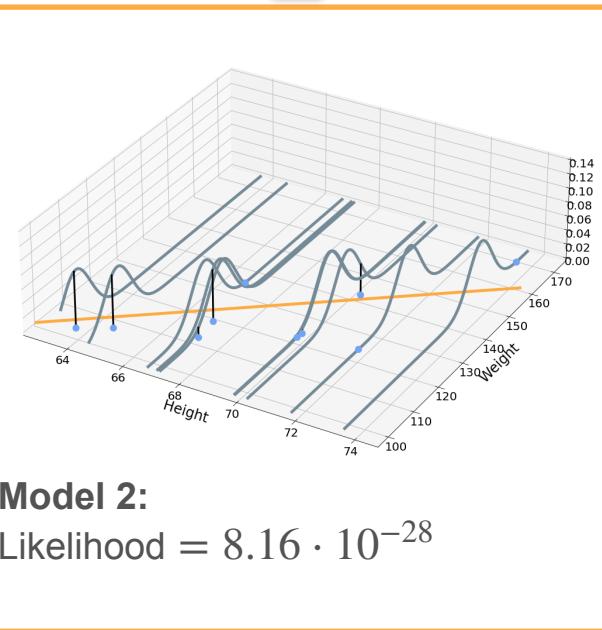


# Picking the Right Model



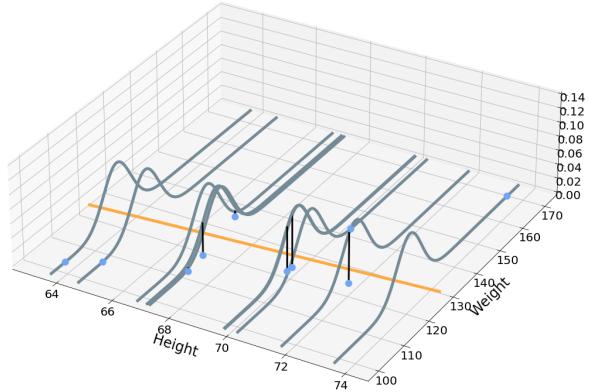
**Model 1:**

$$\text{Likelihood} = 4.91 \cdot 10^{-260}$$



**Model 2:**

$$\text{Likelihood} = 8.16 \cdot 10^{-28}$$



**Model 3:**

$$\text{Likelihood} = 3.48 \cdot 10^{-49}$$



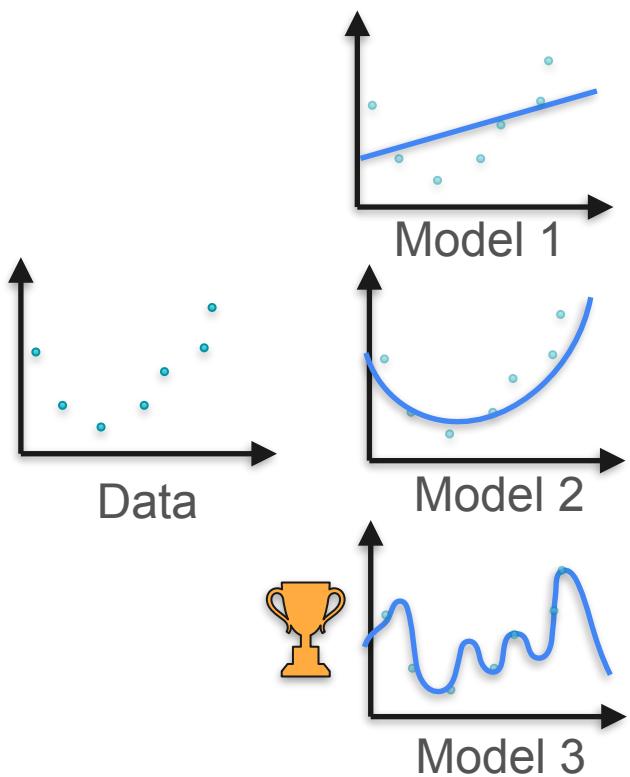
DeepLearning.AI

## Point Estimation

---

## Regularization

# Example: Polynomial Regression



Loss	Equation	Penalty	New loss
10	$y = 4x + 3$	$L_2 = 4^2 = 16$	26
2	$y = 2x^2 - 4x + 5$	$L_2 = 2^2 + (-4)^2 = 20$	22
0.1	$y = 4x^{10} - 9x^8 - 2x^6 + 3x^5 - 6x^4 - 10x + 4$	$L_2 = 4^2 + (-9)^2 + (-2)^2 + 3^2 + (-6)^2 + (-10)^2 = 246$	246.1

# Regularization Term

Model:  $y = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$

Log-loss:  $\ell\ell$

L2 Regularization Error:  $a_n^2 + a_{n-1}^2 + \dots + a_1^2$

Regularization parameter:  $\lambda$

Regularized error:  $+ (\text{L2 Regularization Error})$



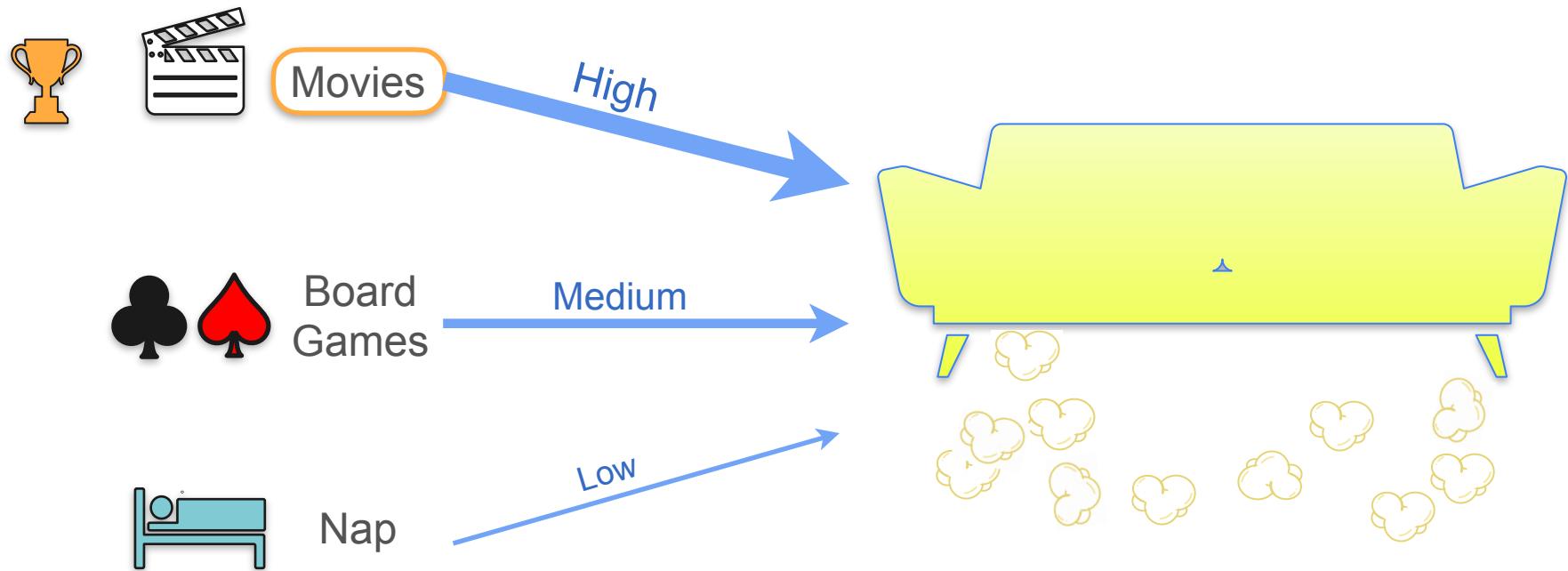
DeepLearning.AI

## Point Estimation

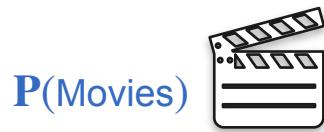
---

**Back to “Bayesics”**

# There's Popcorn on the Floor. What Happened?



# There's Popcorn on the Floor. What Happened?



$P(\text{Movies})$

Movies



$P(\text{Movies}) \gg P(\text{Contest})$



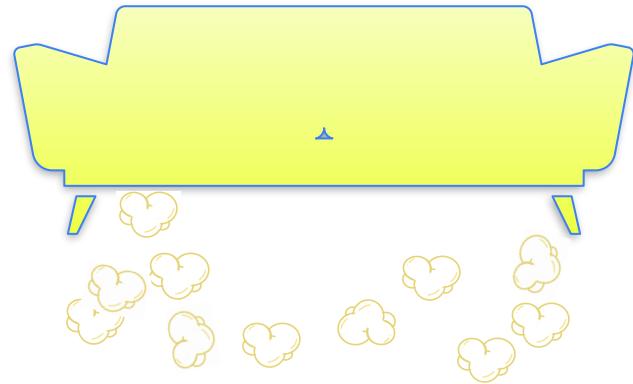
$P(\text{Contest})$

Popcorn  
throwing  
contest



$$P(\text{Popcorn}|\text{Movies}) < P(\text{Popcorn}|\text{Contest})$$

$$P(\text{Popcorn} \cap \text{Movie}) > P(\text{Popcorn} \cap \text{Contest})$$



$$P(A | B)P(B) = P(A \cap B)$$



DeepLearning.AI

## Point Estimation

---

**Frequentist vs Bayesian  
Statistics**

# Frequentists vs. Bayesians

Frequentist



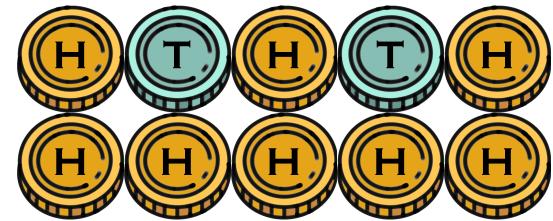
0.8

Bayesian



Belief (prior)

0.52



# Frequentist Vs. Bayesian Statistics

## Frequentists

- Probabilities represent long term frequency of events
- Concept of Likelihood
- Goal: Find the model that most likely generated the observed data

## Bayesians

- Probabilities represent the degree of belief (or certainty)
- Concept of Prior
- Goal: update prior belief based on observations



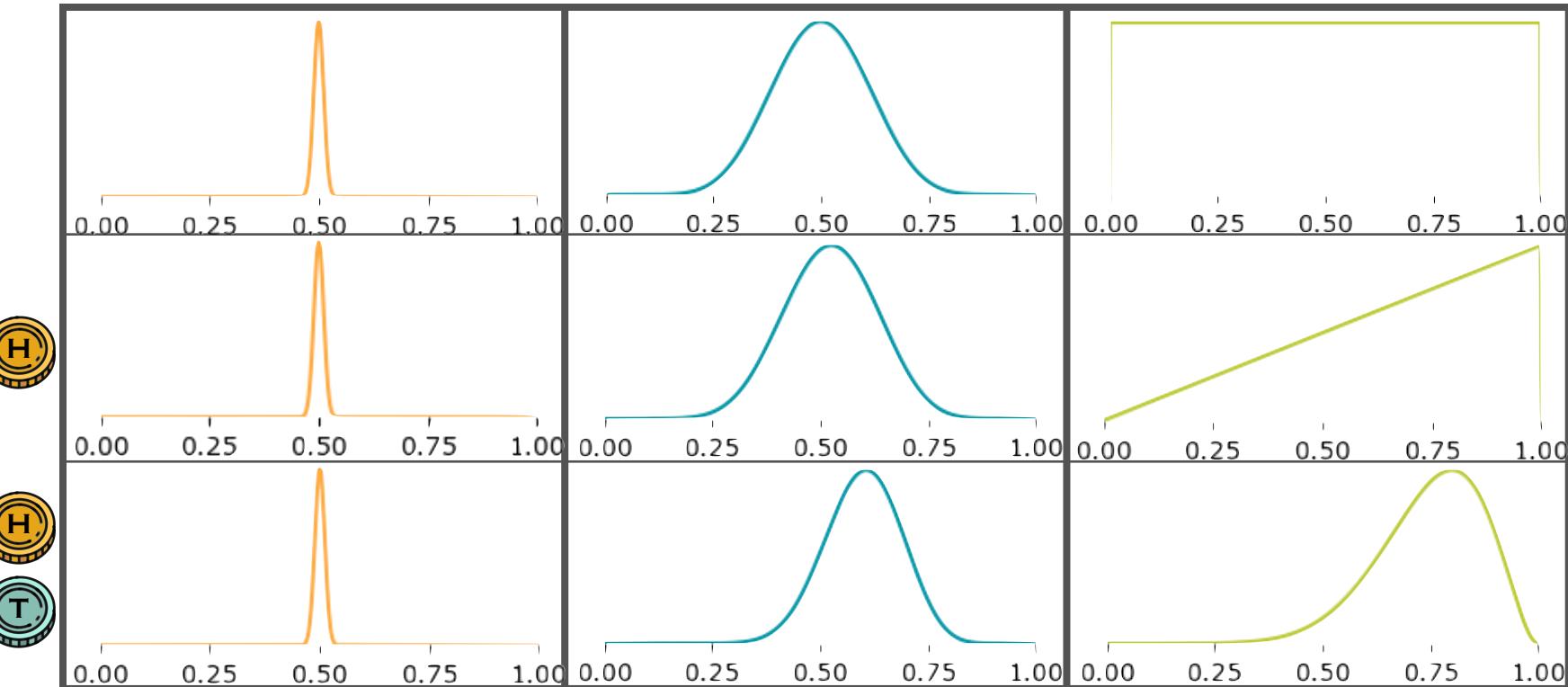
DeepLearning.AI

## Point Estimation

---

# Bayesian Statistics MAP

# Updating Your Beliefs



# Maximum A Posteriori (MAP)

1 value for the parameter?



Choose the one with highest probability

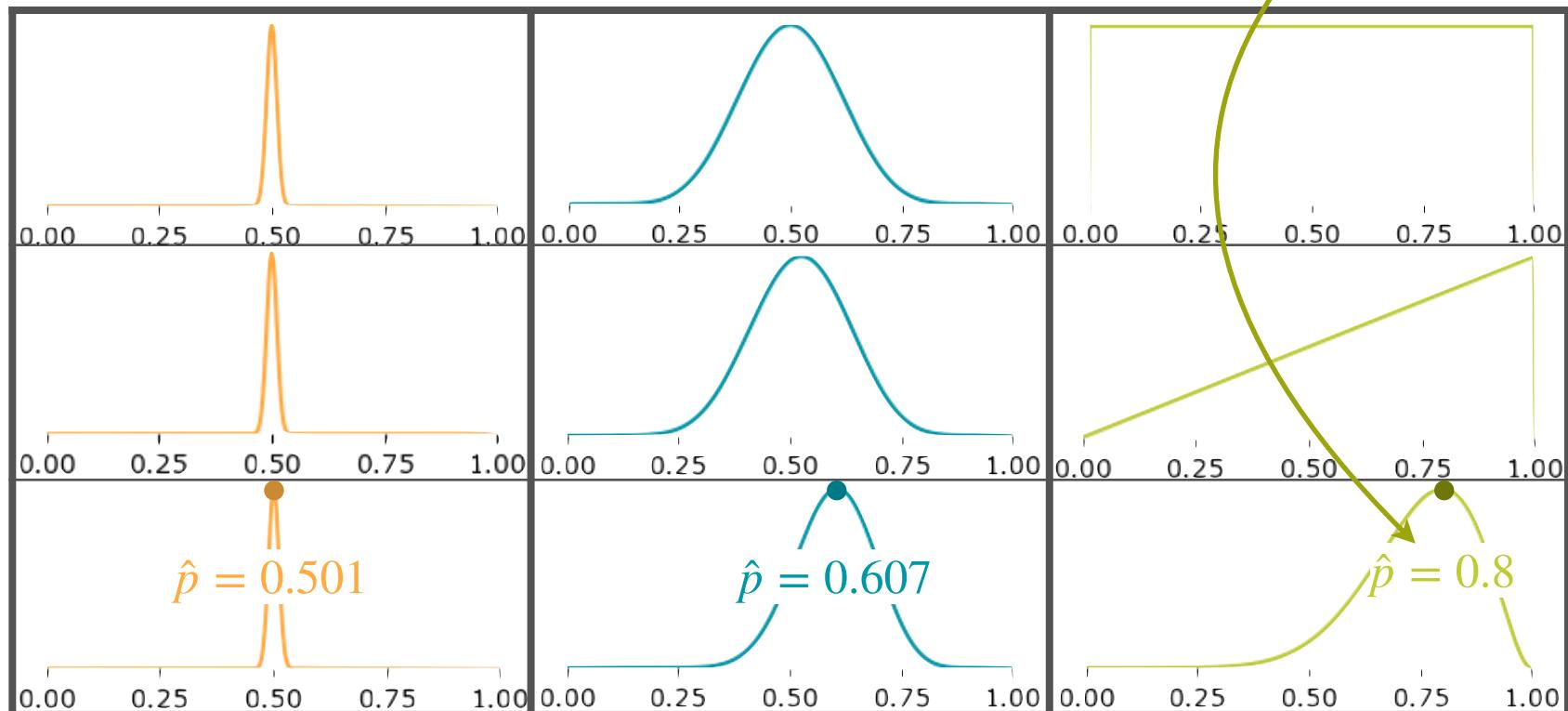


Mode of the updated belief

Posterior

# Maximum A Posteriori (MAP)

Same result as frequentist!





DeepLearning.AI

## Point Estimation

---

**Bayesian Statistics  
Updating Priors**

# Bayesian Statistics

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

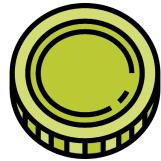
The diagram illustrates the Bayesian formula with annotations:

- "Posterior" Belief that A will happen after considering evidence B:  $P(A | B)$  (orange arrow pointing to the formula)
- Likelihood of evidence B appearing, given A happened:  $P(B | A)$  (purple arrow pointing to the numerator)
- "Prior" Belief that A will happen, before considering evidence B:  $P(A)$  (blue arrow pointing to the numerator)
- Probability of evidence B in any circumstances:  $P(B)$  (green arrow pointing to the denominator)
- $P(B | A)P(A) + P(B | A')P(A')$ : The denominator of the formula, shown as a curved arrow pointing from the bottom right towards the denominator.

$A$ : an event you are trying to predict (you are offered a job)

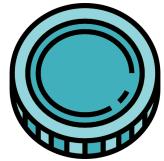
$B$ : another event, or evidence, that helps refine your prediction (you were asked to a follow-up phone call)

# Bayesian Statistics



“Fair”

$$P(H) = 0.5$$



“Biased”

$$P(H) = 0.8$$



“Mystery”

Either Fair or Biased

$$p_{Y|X=1}(0.5)$$

$$= \frac{p_{X|Y=0.5}(1) \quad p_Y(0.5)}{p_X(1)} =$$

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Event to Predict

$A \rightarrow Y$  takes some value

$Y$ : odds of H for your coin

$$Y = \begin{cases} 0.5 & \text{if coin is fair} \\ 0.8 & \text{if coin is biased} \end{cases}$$

Evidence

$B \rightarrow X$  take some value

$X$ : result of coin flip

$$X = \begin{cases} 0 & \text{if } T \\ 1 & \text{if } H \end{cases}$$

Priors

$$\begin{aligned} P(Y = 0.5) &= 0.75 \\ P(Y = 0.80) &= 0.25 \end{aligned}$$

$x = 1$

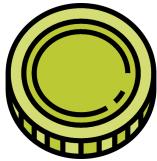
0.5 0.75

0.5 0.75 0.8 0.25

$$\begin{aligned} P(Y = 0.5 | X = 1) &= 0.652 \\ P(Y = 0.8 | X = 1) &= 0.348 \end{aligned}$$

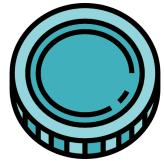
Posterior

# Bayesian Statistics



“Fair”

$$P(H) = 0.5$$



“Biased”

$$P(H) = 0.8$$



“Mystery”

Either Fair or Biased

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Event to Predict

$A \rightarrow Y$  takes some value

$Y$ : odds of H for your coin

$$Y = \begin{cases} 0.5 & \text{if coin is fair} \\ 0.8 & \text{if coin is biased} \end{cases}$$

$$P(Y = 0.5) = 0.75$$

$$P(Y = 0.80) = 0.25$$

Evidence

$B \rightarrow X$  take some value

$X$ : result of coin flip

$$X = \begin{cases} 0 & \text{if } T \\ 1 & \text{if } H \end{cases}$$

$$x = 1$$

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x) p_Y(y)}{p_X(x)}$$

# Bayesian Statistics

Y is discrete

X is discrete

X is continuous

Posterior

Prior

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)p_Y(y)}{p_X(x)}$$

Posterior

Prior

$$p_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)p_Y(y)}{f_X(x)}$$

Y is continuous

Prior

Posterior

$$f_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)f_Y(y)}{p_X(x)}$$

Posterior

Prior

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x)f_Y(y)}{f_X(x)}$$

# Bayesian Statistics

$X$  is discrete

$\Theta$  is discrete

$$p_{\Theta|X=x}(\theta) = \frac{p_{X|\Theta=\theta}(x)p_{\Theta}(\theta)}{p_X(x)}$$

$X$  is continuous

$\Theta$  is continuous

$$f_{\Theta|X=x}(\theta) = \frac{p_{X|\Theta=\theta}(x)f_{\Theta}(\theta)}{f_X(x)}$$

$$p_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x)p_{\Theta}(\theta)}{f_X(x)}$$

$$f_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x)f_{\Theta}(\theta)}{f_X(x)}$$



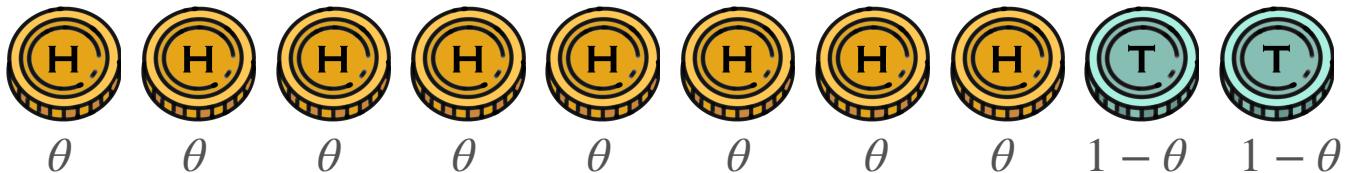
DeepLearning.AI

## Point Estimation

---

**Bayesian Statistics  
Full Worked Example**

# Bayesian Statistics: Bernoulli Example



$$\Theta = P(H)$$

$$\mathbf{X} = (X_1, X_2, \dots, X_{10})$$

$X_i = 1$  if  $H$ ,  $0$  if  $T$

$$X_i | \Theta = \theta \sim \text{Bernoulli}(\theta)$$

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) = \frac{p_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) f_{\Theta}(\theta)}{p_{\mathbf{X}}(\mathbf{x})}$$

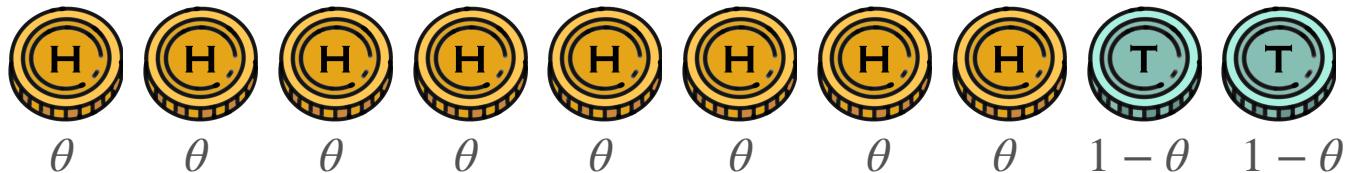
$\Theta$  is a continuous random variable

$\mathbf{X}$  is a discrete random variable

Use this version of Bayes' Theorem

$$p_{\mathbf{X}|\Theta=\theta}(1,1,\dots,1,0,0) = p(X_1 = 1, X_2 = 1, \dots, X_8 = 1, X_9 = 0, X_{10} = 0 | \Theta = \theta) = \theta^8(1 - \theta)^2$$

# Bayesian Statistics: Bernoulli Example



$$X_i | \Theta = \theta \sim \text{Bernoulli}(\theta)$$

$$\Theta \sim \text{Uniform}(0,1)$$

$$f_{\Theta|X=x}(\theta) = \frac{p_{\mathbf{X}|\Theta=\theta}(\mathbf{x})f_{\Theta}(\theta)}{p_{\mathbf{X}}(\mathbf{x})}$$

$$p_{\mathbf{X}|\Theta=\theta}(1,1,\dots,1,0,0) = \theta^8(1-\theta)^2 \quad f_{\Theta}(\theta) = 1, \quad 0 \leq \theta \leq 1$$

$$f_{\Theta|X=x}(\theta) = \frac{\theta^8(1-\theta)^2}{\text{constant}} = \frac{1}{\text{constant}}\theta^8(1-\theta)^2$$

This is a Beta Distribution, it's possible to calculate this constant, but it's unnecessary

# Bayesian Statistics: Bernoulli Example

$$f_{\Theta|X=x}(\theta) = \frac{\theta^8(1-\theta)^2 1}{\text{constant}}$$



$$f_{\Theta}(\theta) = 1$$

$$f_{\Theta|X=x}(\theta) = \frac{1}{\text{constant}} \theta^8(1-\theta)^2$$

# Bayesian Statistics: MAP Estimator

$\theta$ : a model of the coin  
where  $P(H) = \theta$

$$f_{\Theta|X=x}(\theta) \propto \theta^8(1-\theta)^2 [1] \quad \text{Likelihood of the model}$$

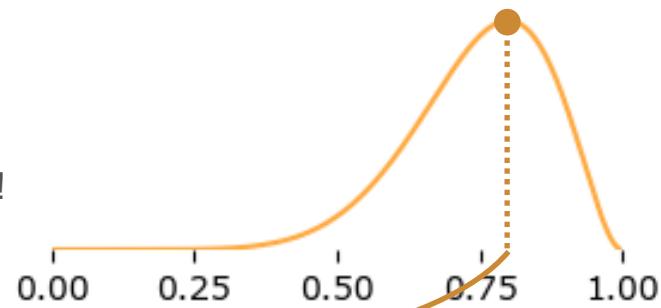
$P(\text{data} | \text{model})$

Maximizing this is all MLE does!

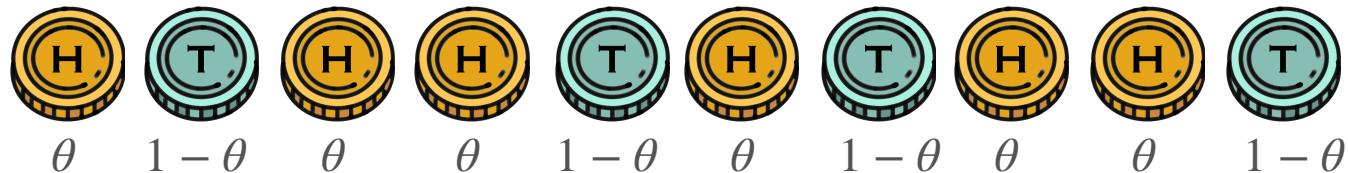
Maximum a Posteriori (MAP)

$$\hat{\theta} = \arg \max_{\theta} f_{\Theta|X=x}(\theta) = \frac{8}{10}$$

Non informative prior: MAP = MLE



# Bayesian Statistics: Bernoulli Example Continued



$$f_{\Theta|X=x}(\theta) = \theta^8(1 - \theta)^2$$

$$p_{X|\Theta=\theta}(1,1,0,\dots,1,1,0) = \theta^6(1 - \theta)^4$$

New prior!

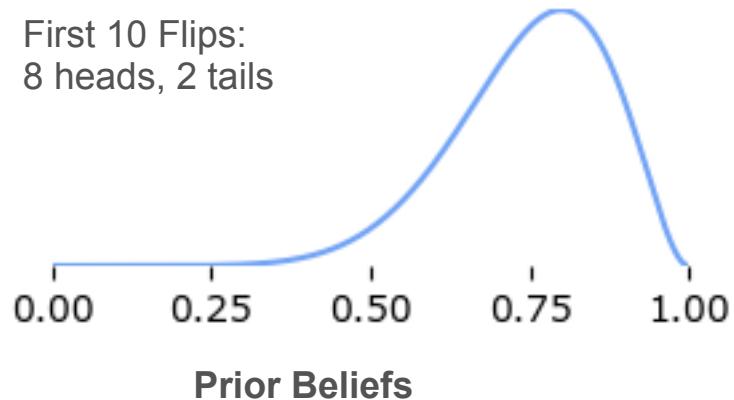
Repeat process

$$f_{\Theta|X=x}(\theta) = \frac{\text{constant}}{\theta^6(1 - \theta)^4}$$

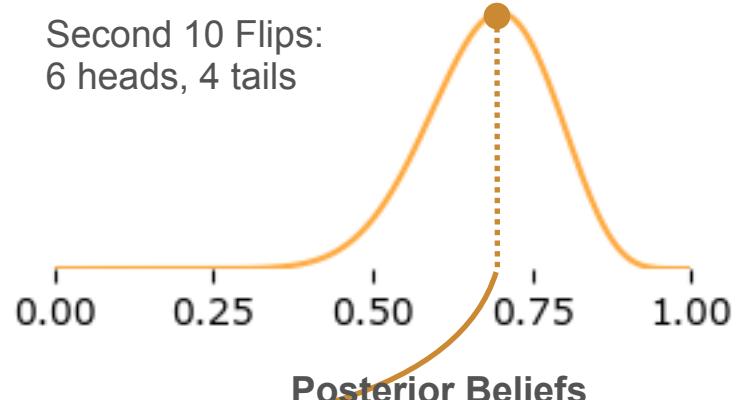
# Bayesian Statistics: Bernoulli Example Continued

$$f_{\Theta|X=x}(\theta) = \frac{\theta^6(1-\theta)^4 \theta^8(1-\theta)^2}{\text{constant}}$$

First 10 Flips:  
8 heads, 2 tails



Second 10 Flips:  
6 heads, 4 tails



MAP

$$\hat{\theta} = 0.7$$

14 out of 20 heads  
Same result as Frequentist

$$f_{\Theta}(\theta) = \frac{1}{\text{constant}} \theta^8(1-\theta)^2$$

# Bayesian Statistics: Final Summary



- Bayesians update prior beliefs
- MAP with uninformative priors is just MLE
- With enough data, MLE and MAP estimates usually converge
- Good for instances when you have limited data or strong prior beliefs
- Wrong priors, wrong conclusions



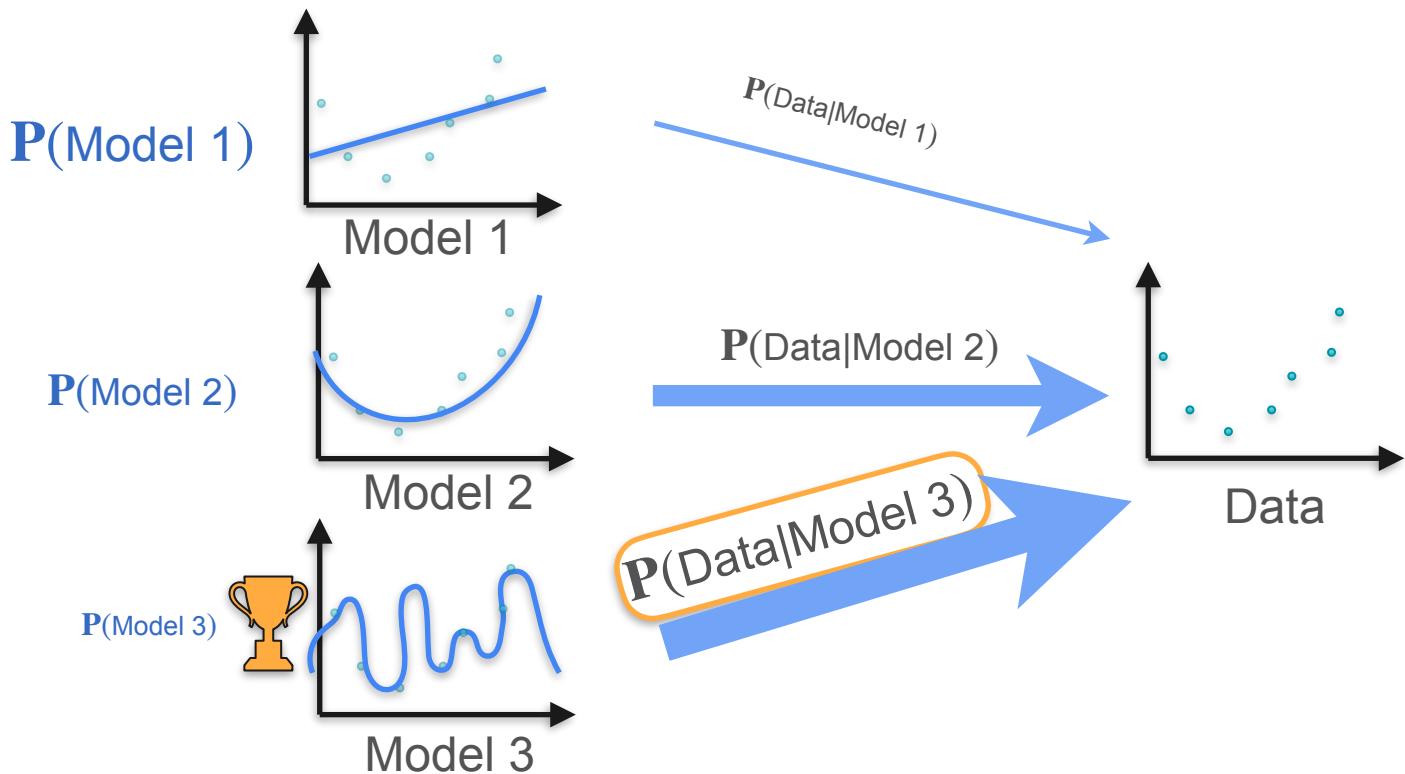
DeepLearning.AI

## Point Estimation

---

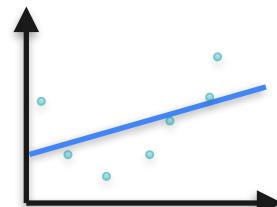
**Relationship between MAP,  
MLE, and Regularization**

# Example: Polynomial Regression

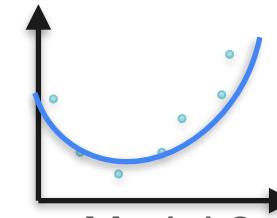


# Example: Polynomial Regression

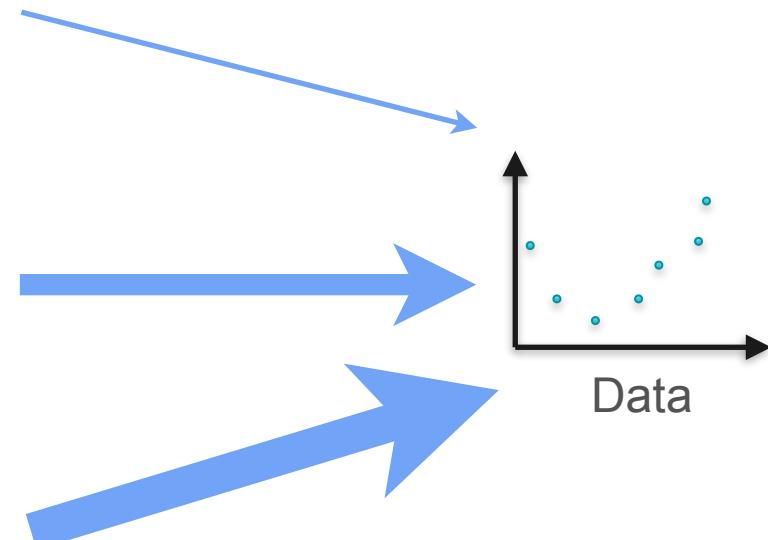
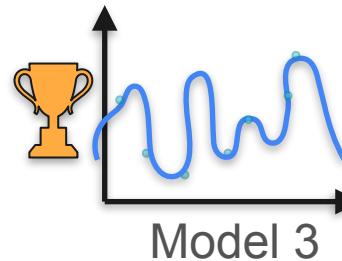
$P(\text{Model 1}) P(\text{Data}|\text{Model 1})$



$P(\text{Model 2}) P(\text{Data}|\text{Model 2})$



$P(\text{Model 3}) P(\text{Data}|\text{Model 3})$



Maximum likelihood  
with Bayes

Polynomial regression  
with regularization

$P(\text{Data}|\text{Model})$

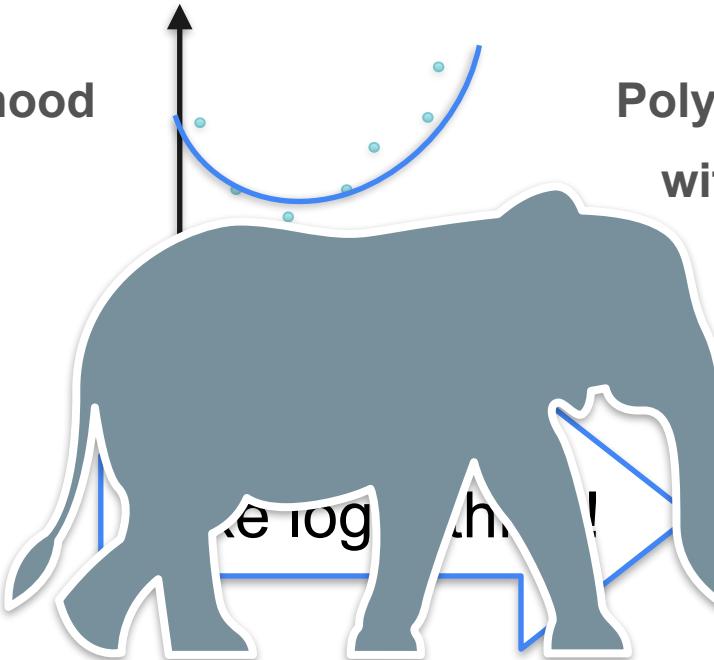
$P(\text{Model})$

?

Log-loss

+

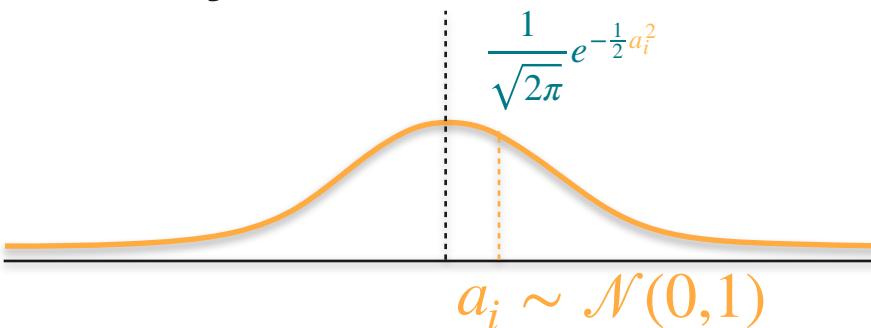
Regularization term



# What Is the Probability of a Model?

$$P(\text{Model 1}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_1^2}$$

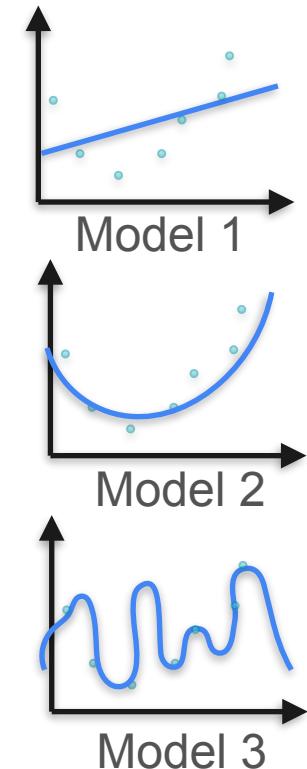
$a_1x + b$



$$P(\text{Model 2}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_2^2}$$

$a_1x^2 + a_2x + b$

$$P(\text{Model 3}) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_i^2}$$



# Bayes and Regularization

$P(\text{Data}|\text{Model})$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_2^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_3^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_4^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}d_5^2}$$

$P(\text{Model})$

Maximize

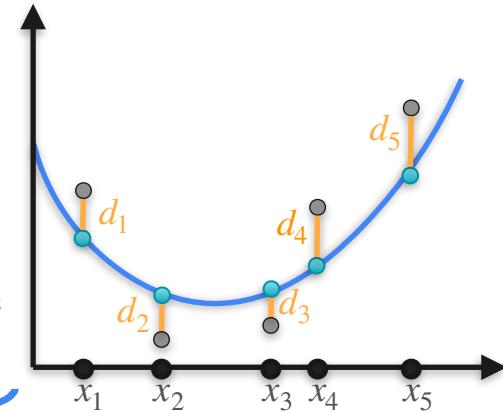
$$-\frac{1}{2}(d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2)$$

$\log$

$\log$

$\log$

$$+ \quad -\frac{1}{2}(a_1^2 + a_2^2)$$



$$a_1x^2 + a_2x + b$$

Minimize

$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + a_1^2 + a_2^2$$

# Regularization

P(Model 1)

$$\text{Minimize } x_1^2 + d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2$$

P(Model 2)

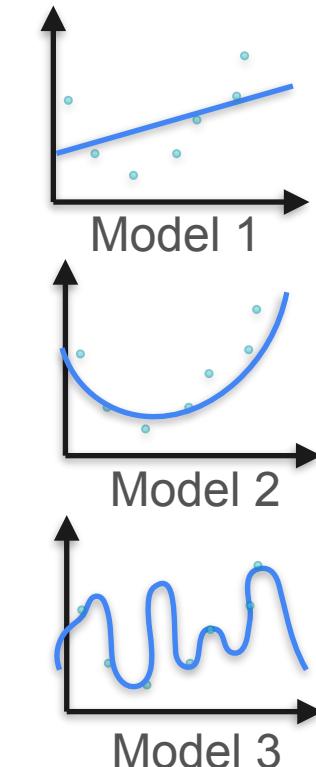
$$\text{Minimize } x_1^2 + x_2^2 + d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2$$

P(Model 3)

$$\text{Minimize } x_1^2 + \dots + x_{10}^2 + d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2$$

New Loss

P(Data|Model 1)





DeepLearning.AI

## Point Estimation

---

## Conclusion