Copyright Notice

DeepLearning.AI

Probability and Statistics for Machine Learning and Data Science

-

Week 3: Sampling and Point Estimates

W3 Lesson 1

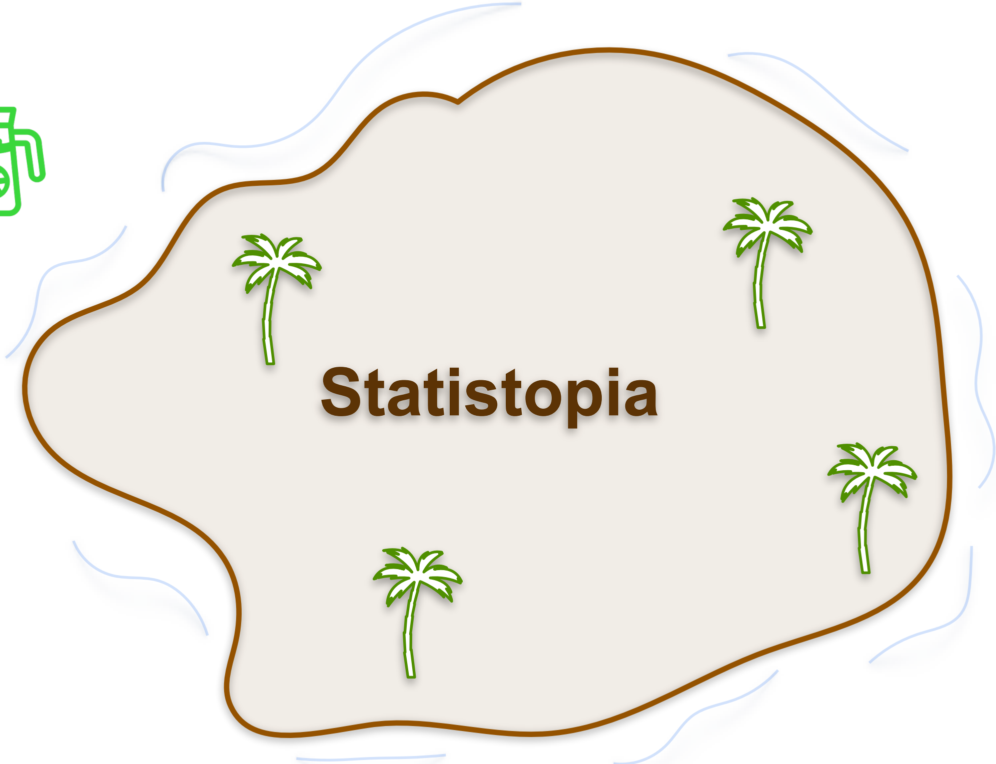
DeepLearning.AI

Sample and Population

Population and Sample

Population and Sample

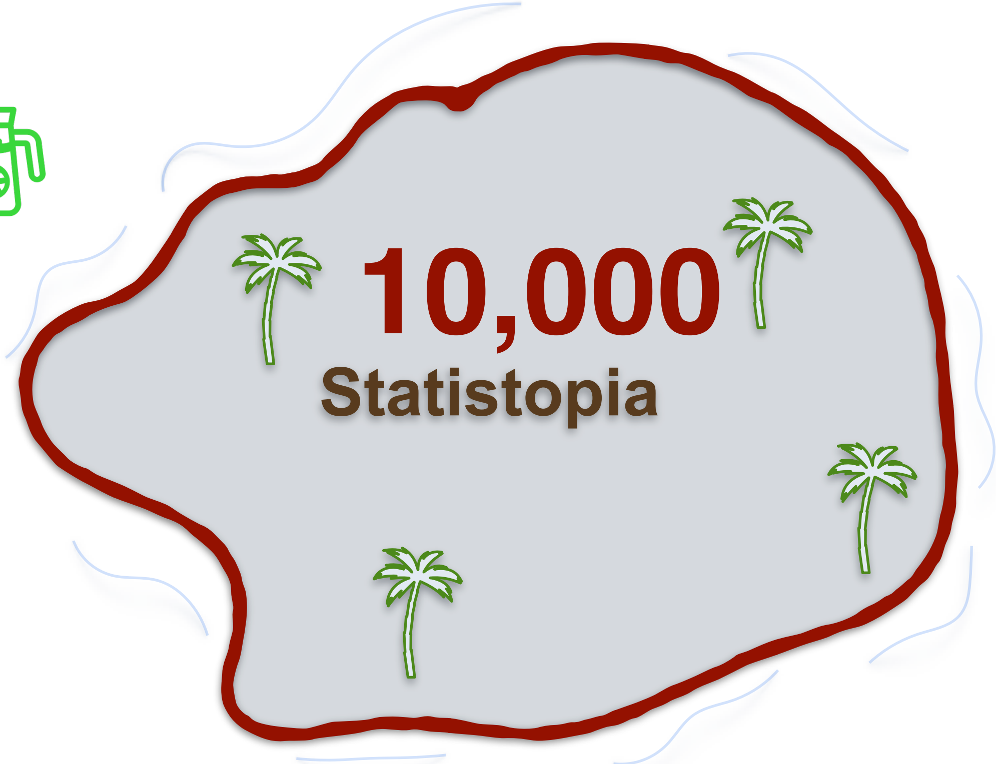
Find the **average height** of the people living on Statistopia



Population and Sample

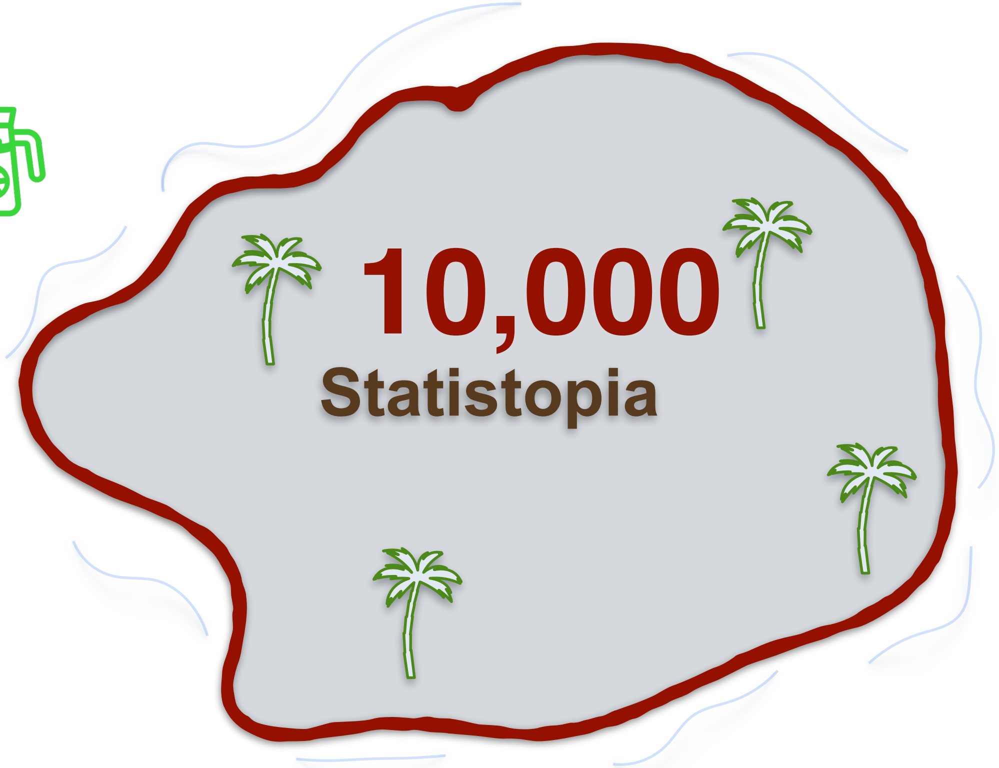
- Ask everyone on the island for their height.

- Divide by the total number



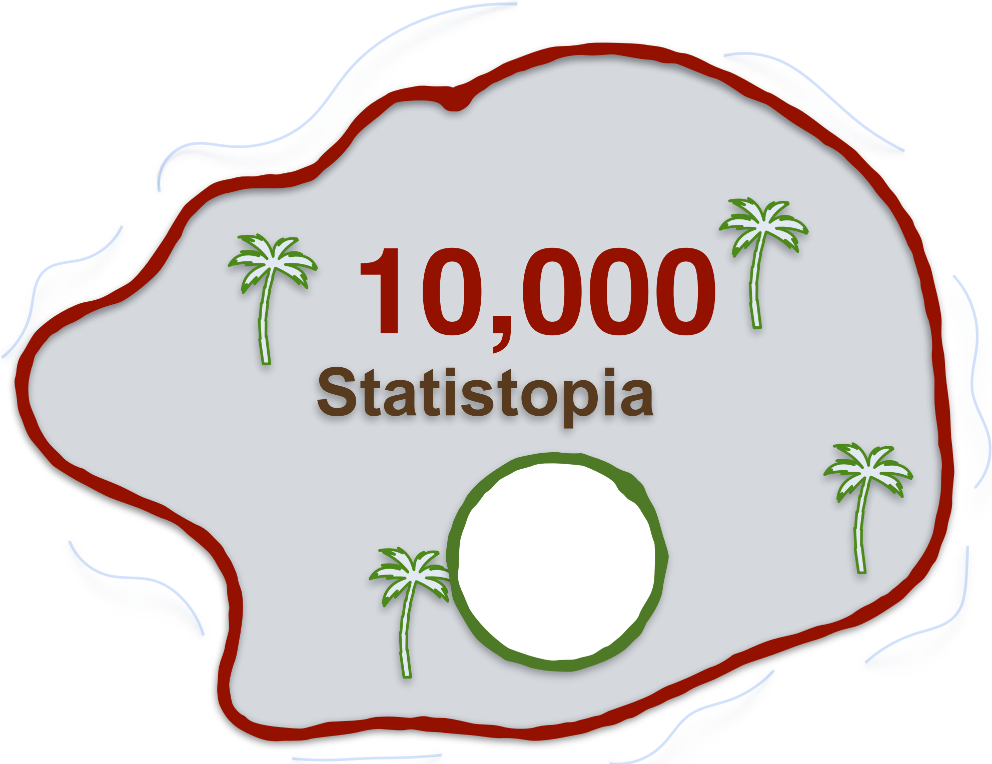
Population and Sample

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



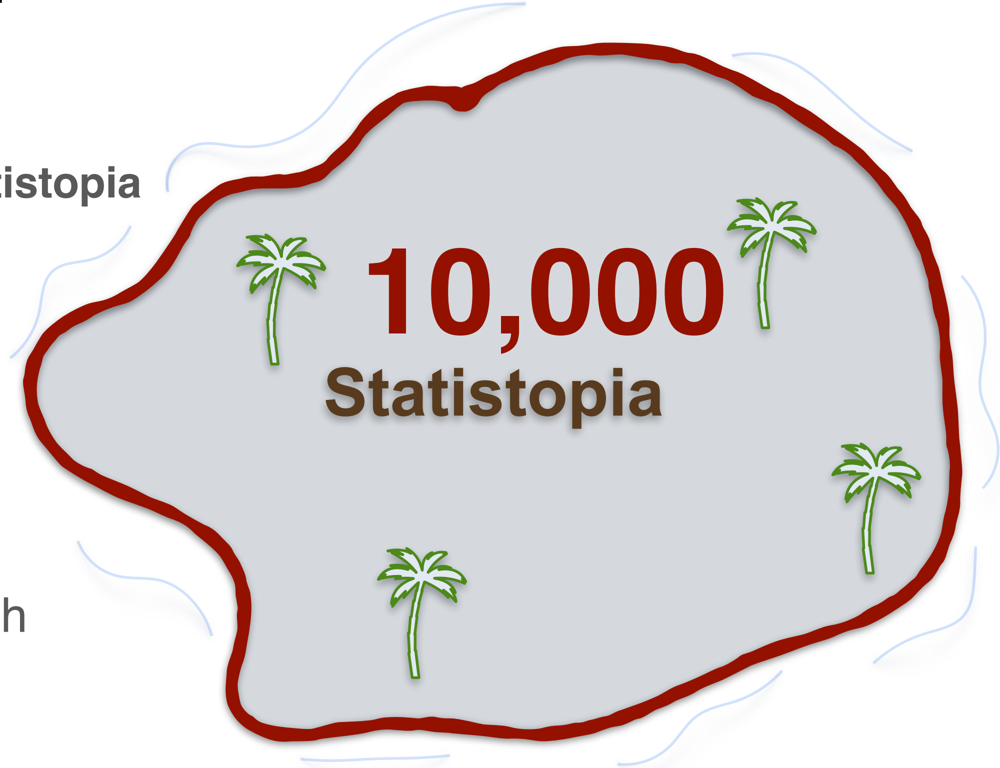
Population and Sample

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



Population and Sample

The people of statistopia

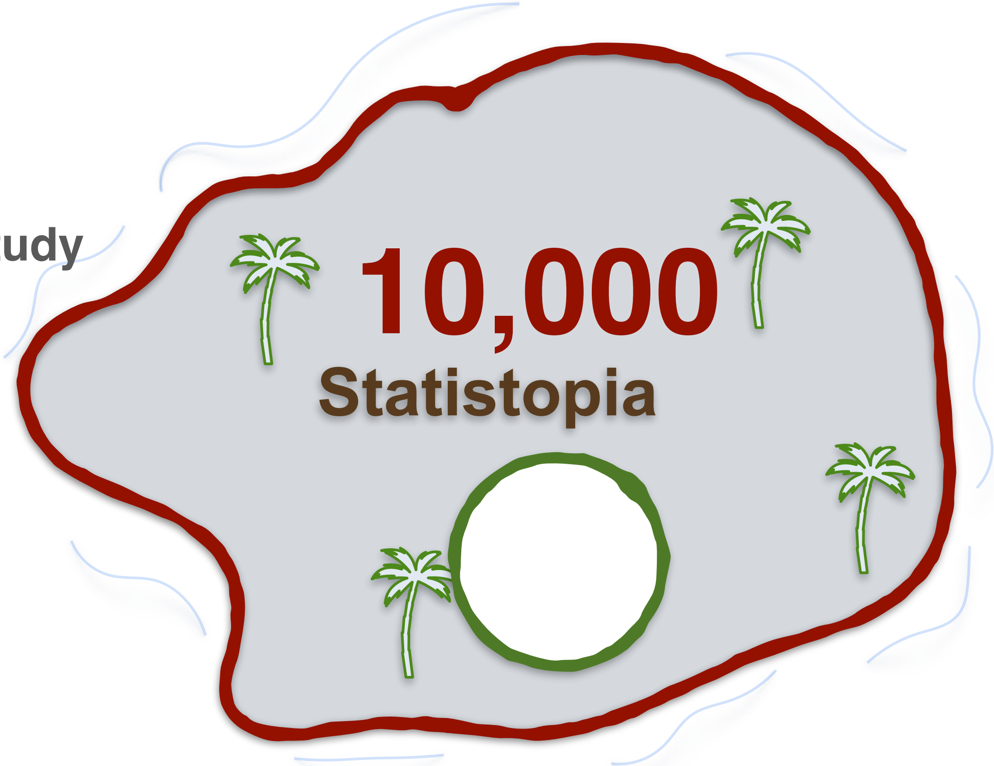


Population:
the entire group of individuals or elements you want to study which share a common behaviour

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



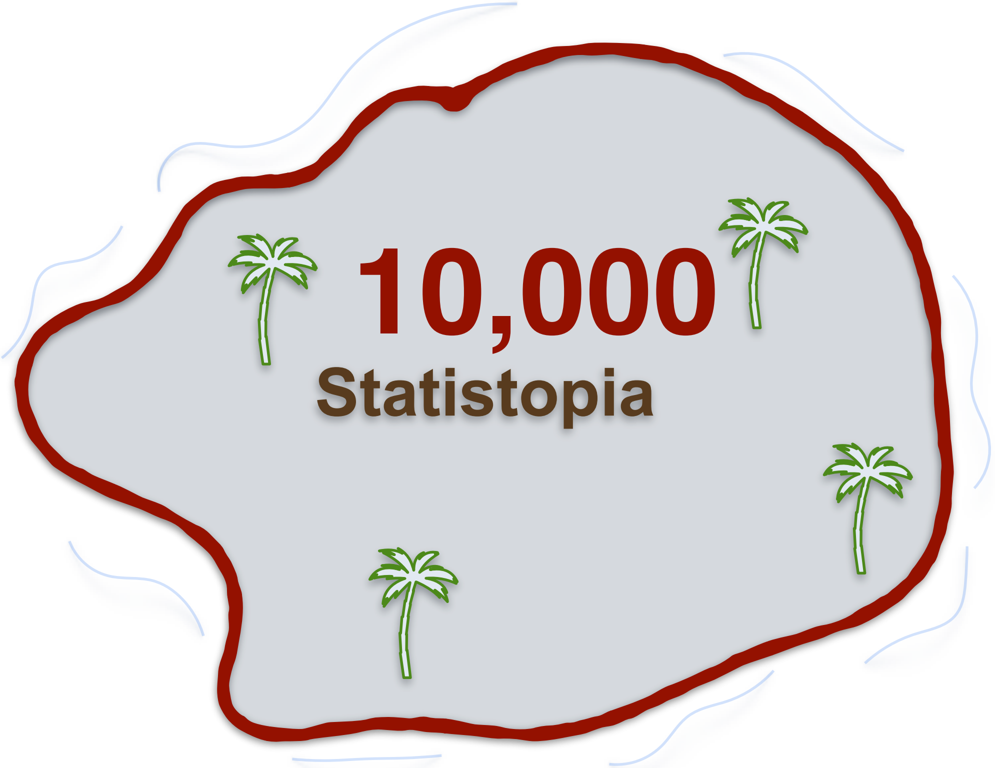
Population and Sample

Population Size (N)

10,000

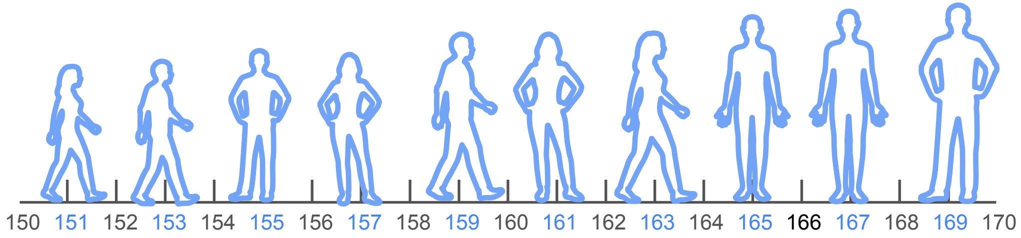
Sample Size (n)

1 - 9,999



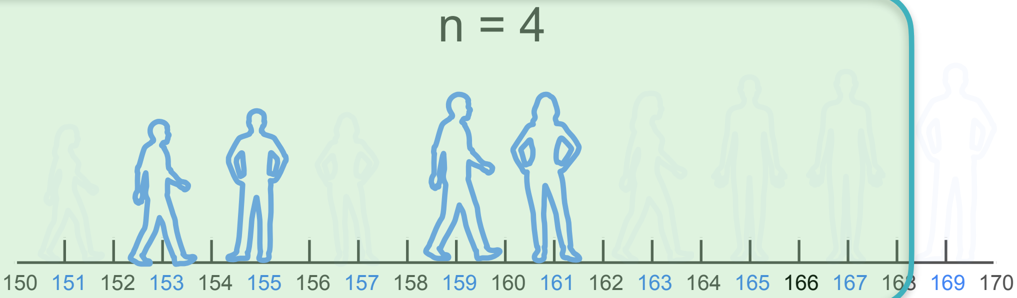
Population and Sample

N = 10



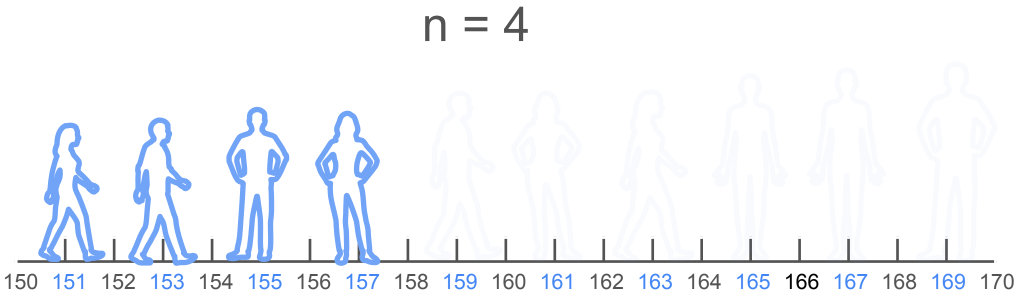
Random Sampling

A



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

B

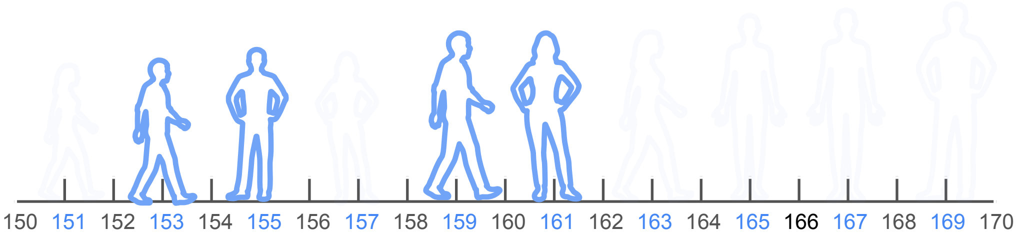


Independent Sample

Example 1

n = 4

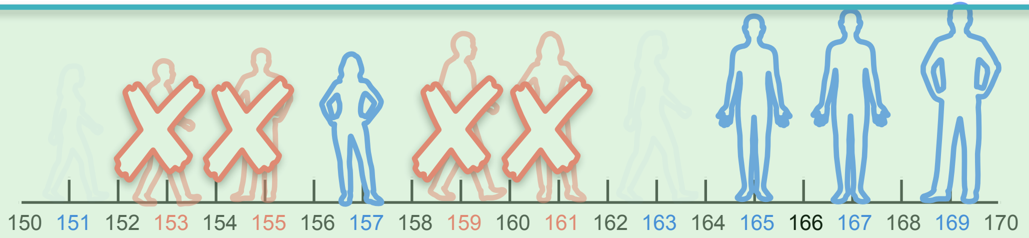
1st sample set



Why is sample set two not a good sample?

n = 4

2nd sample set

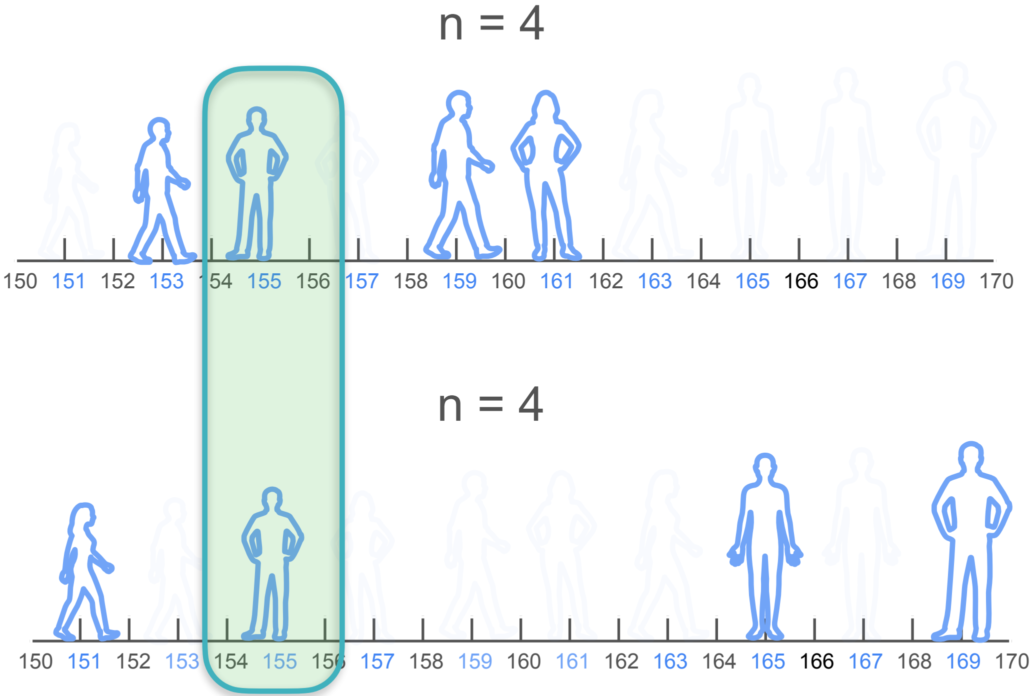


Independent Sample

Example 2

1st sample set

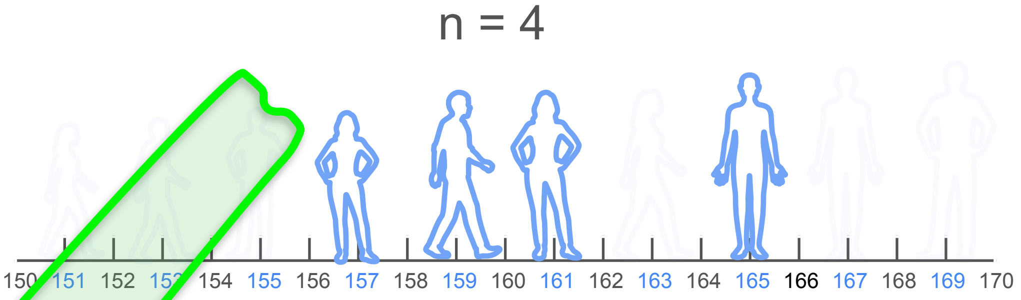
2nd sample set



Identically Distributed Samples

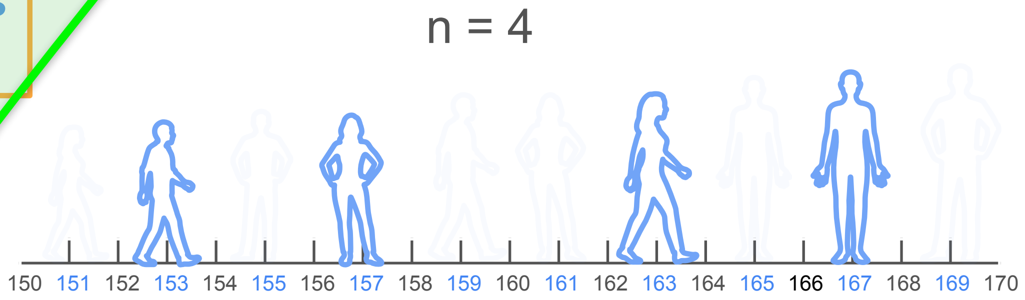
Example 1

A



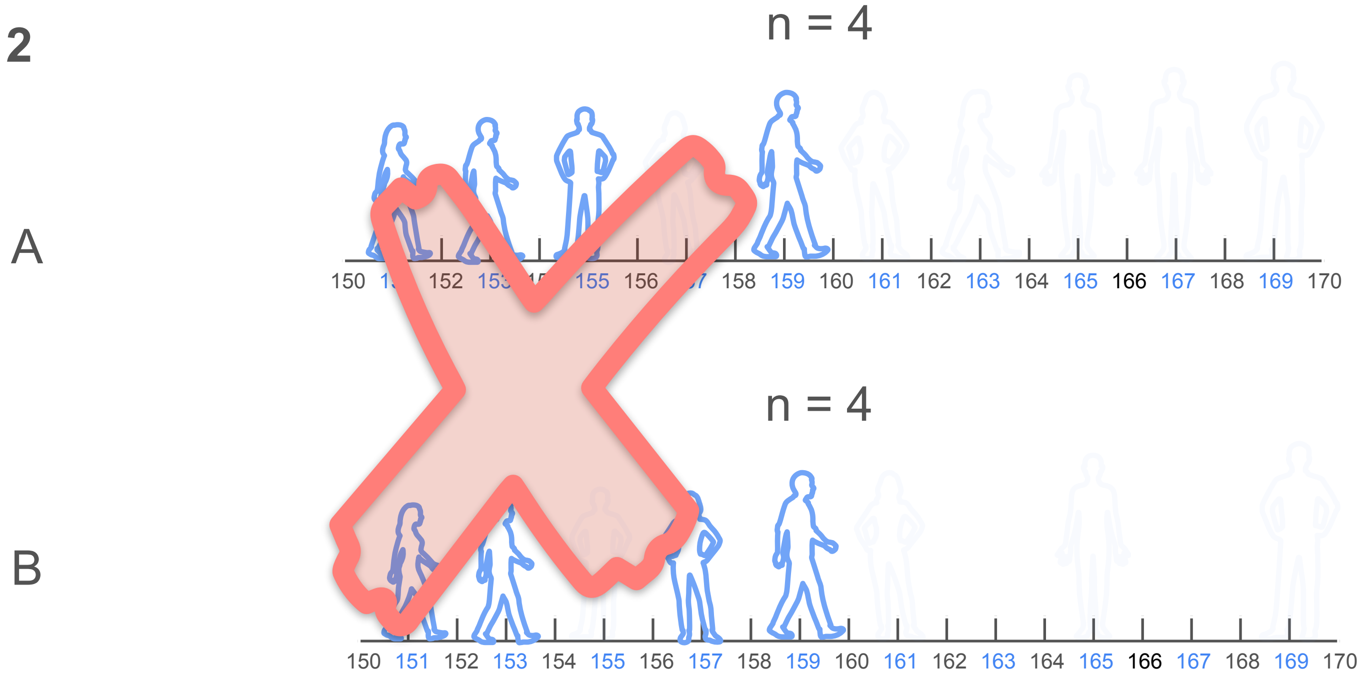
Which of the following samples are identically distributed?

B



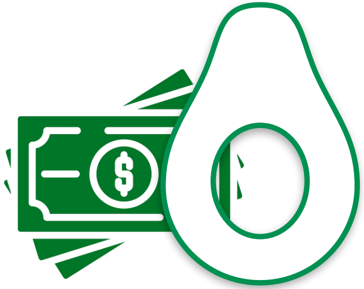
Identically Distributed Samples

Example 2



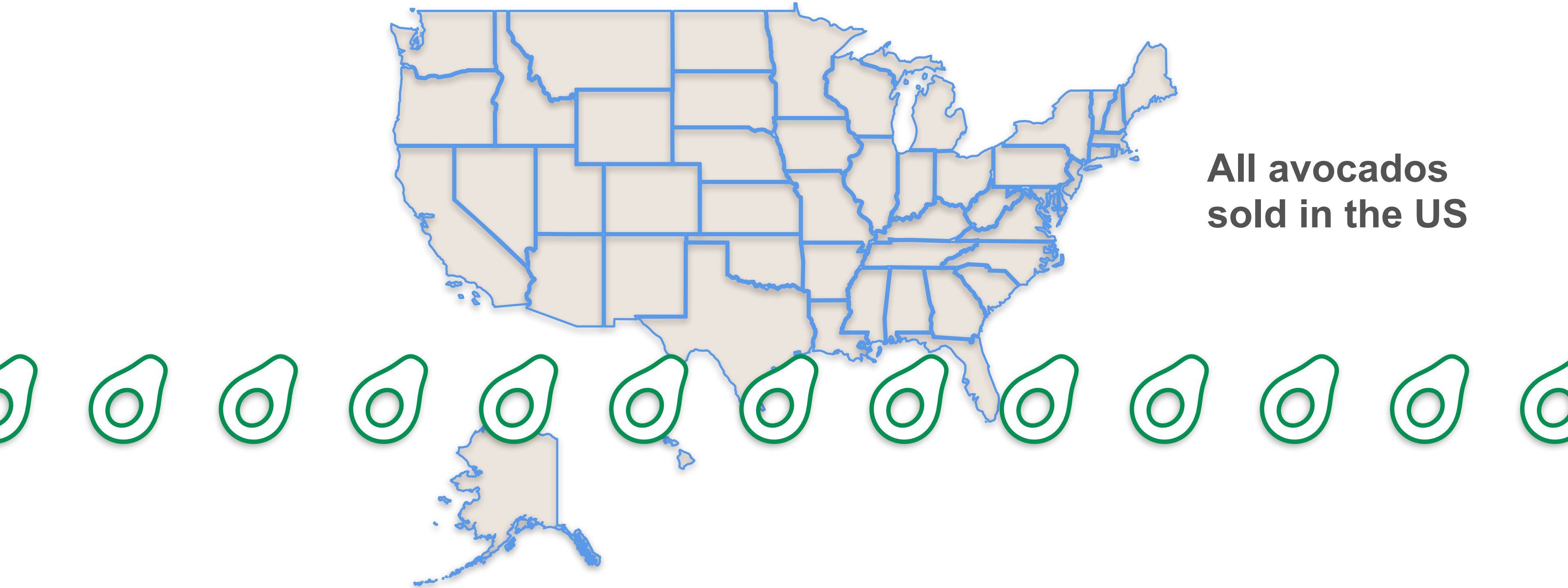
The Avocado Toast Trend

Study the price of avocados
in the United States



What is the population of your study?

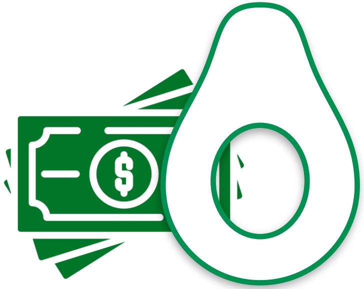
The Avocado Toast Trend



All avocados
sold in the US

The Avocado Toast Trend

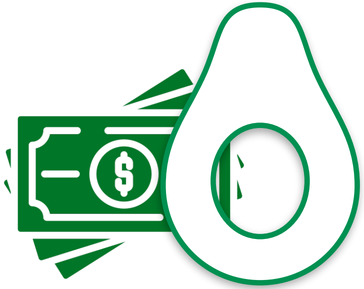
Study the price of avocados
in the United States



What is the sample of your study?

The Avocado Toast Trend

Study the price of avocados
in the United States



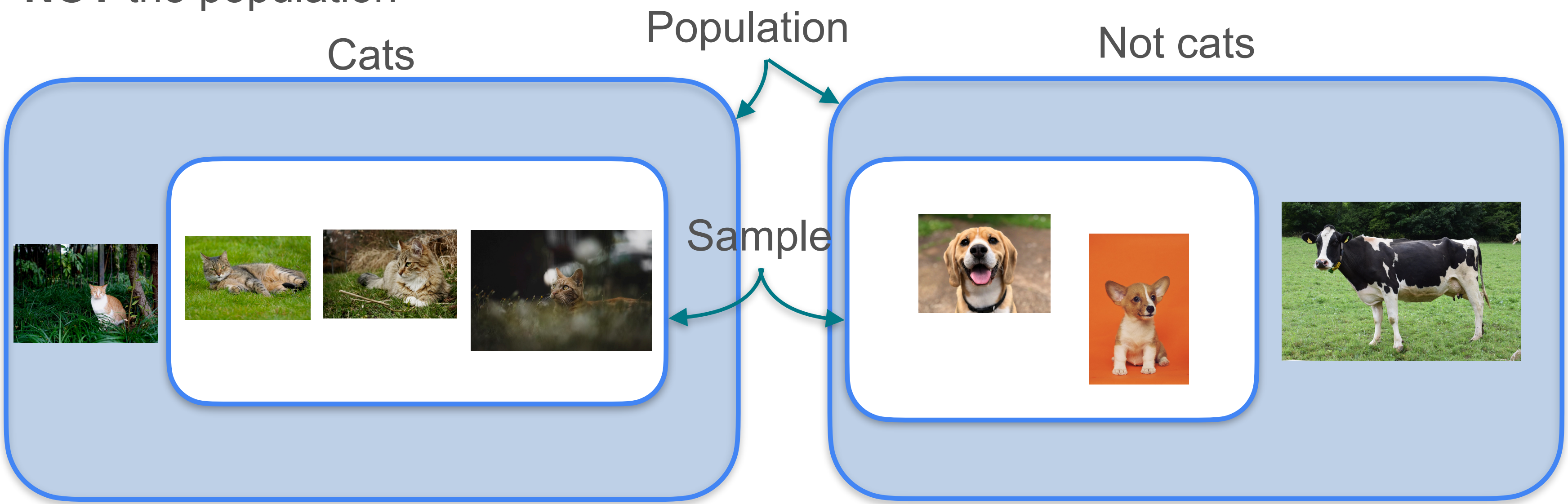
What is the sample of your study?

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

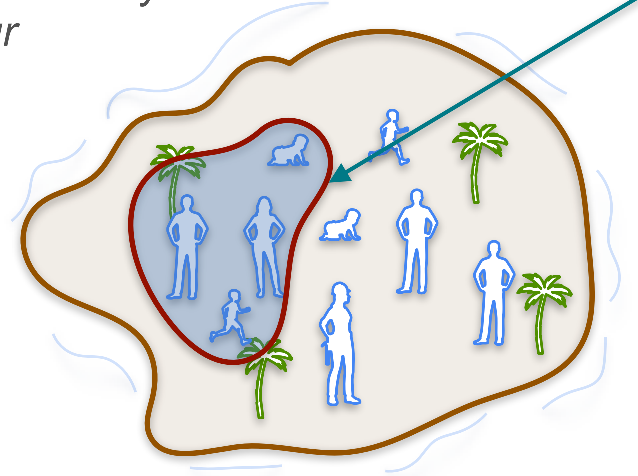
NOT the population



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

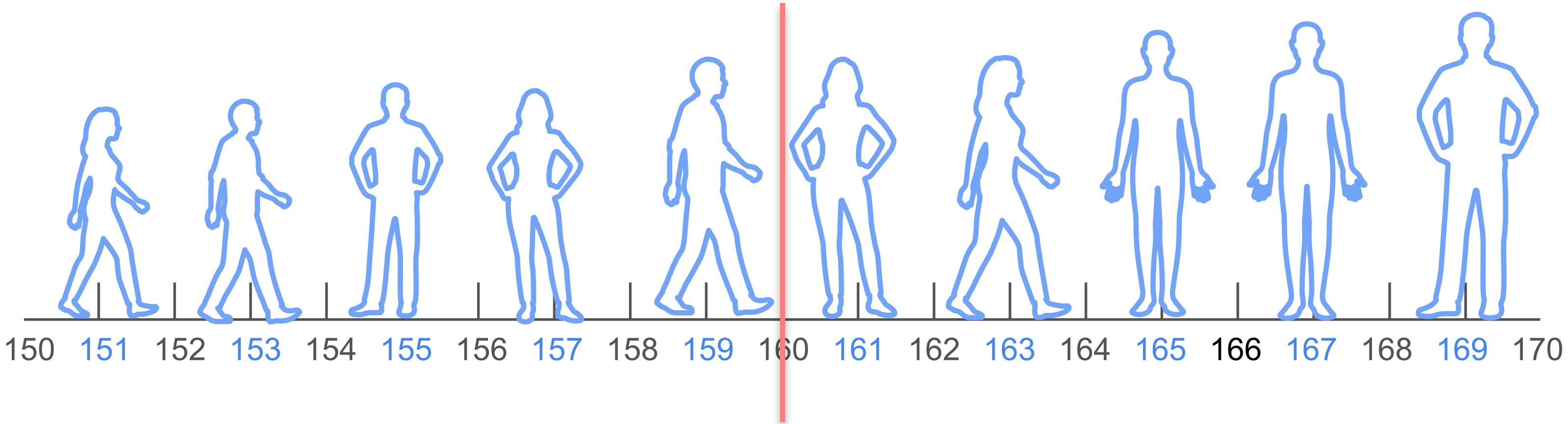
Sample Size: n

DeepLearning.AI

Sample and Population

Sample Mean

Population and Sample Mean



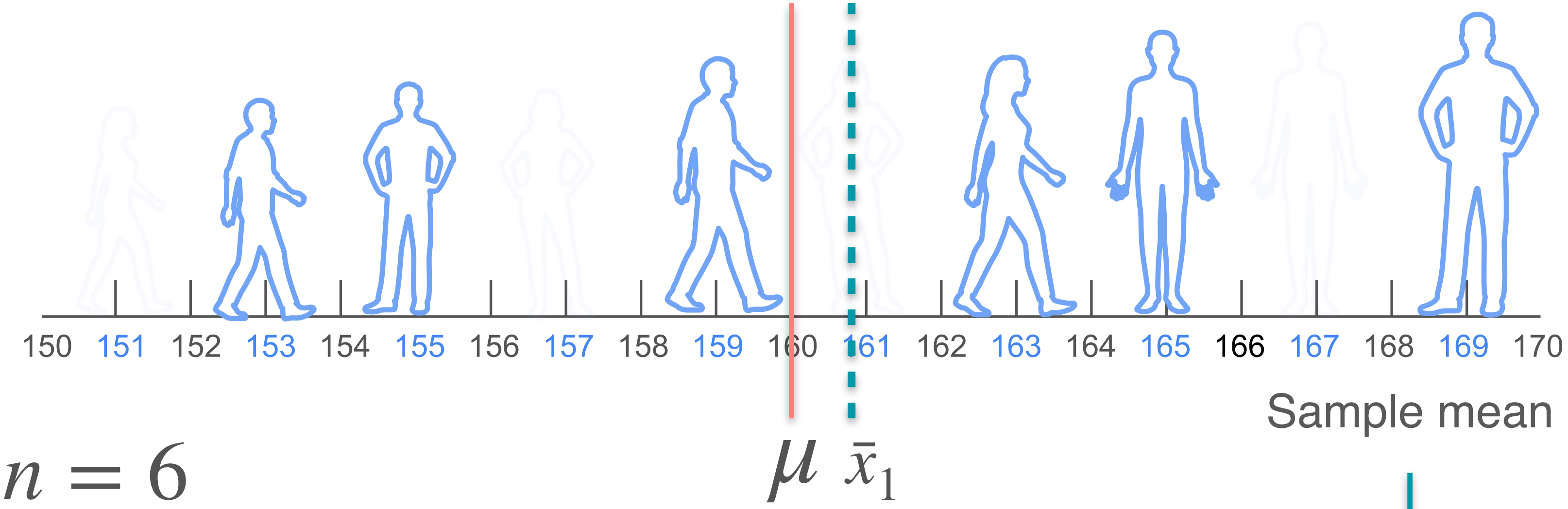
What is the average height in statistopia?

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

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

Population mean μ

Population and Sample Mean



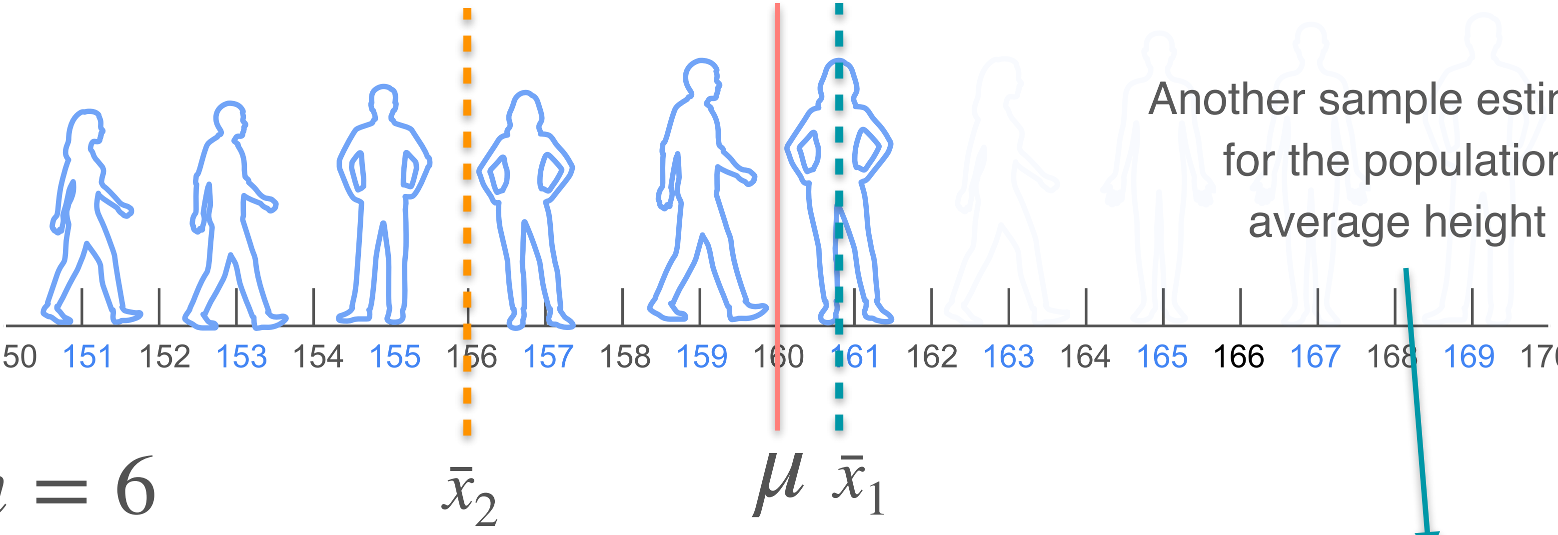
Sample mean

$$n = 6$$

What is the average height in statistopia?

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

Population and Sample Mean



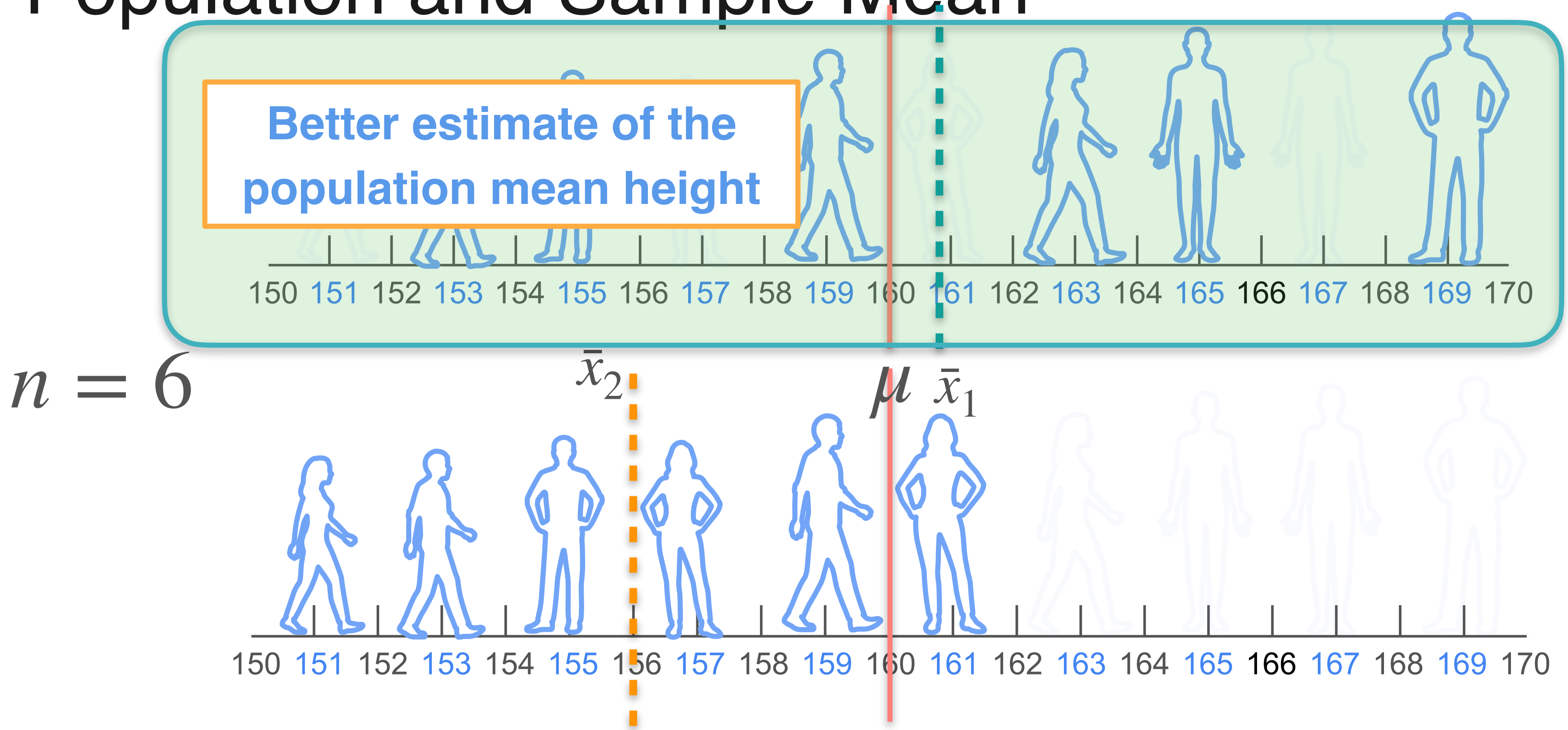
Another sample estimate for the population average height

$n = 6$

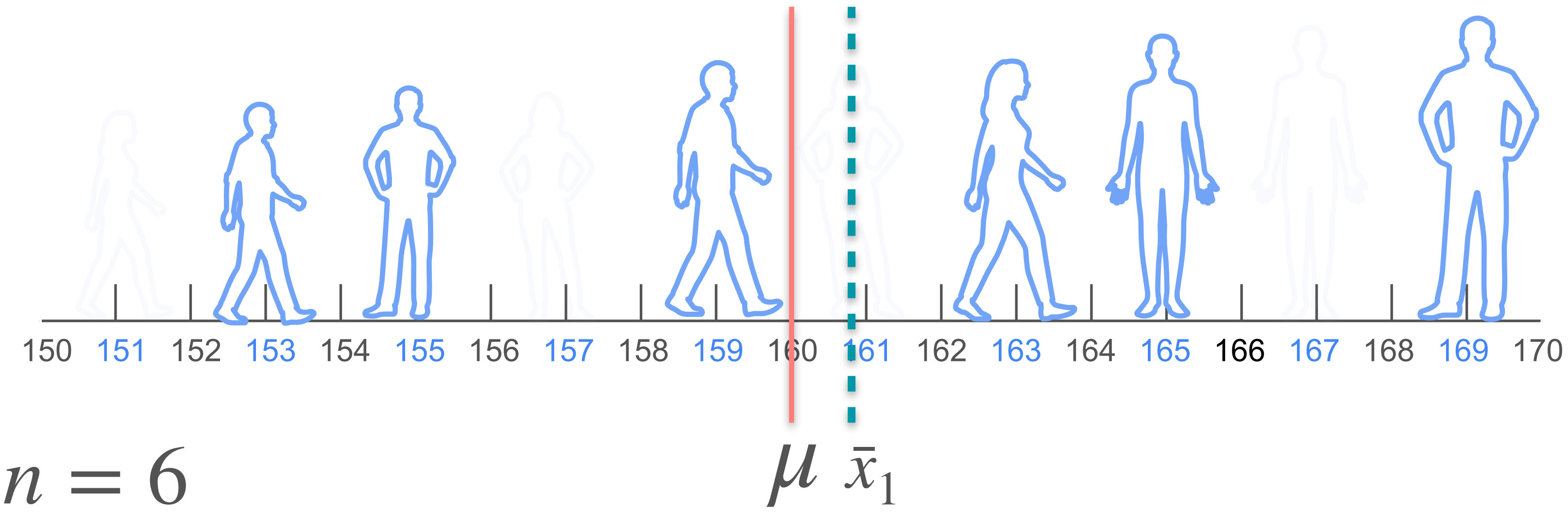
What is the average height in statistopia?

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

Population and Sample Mean



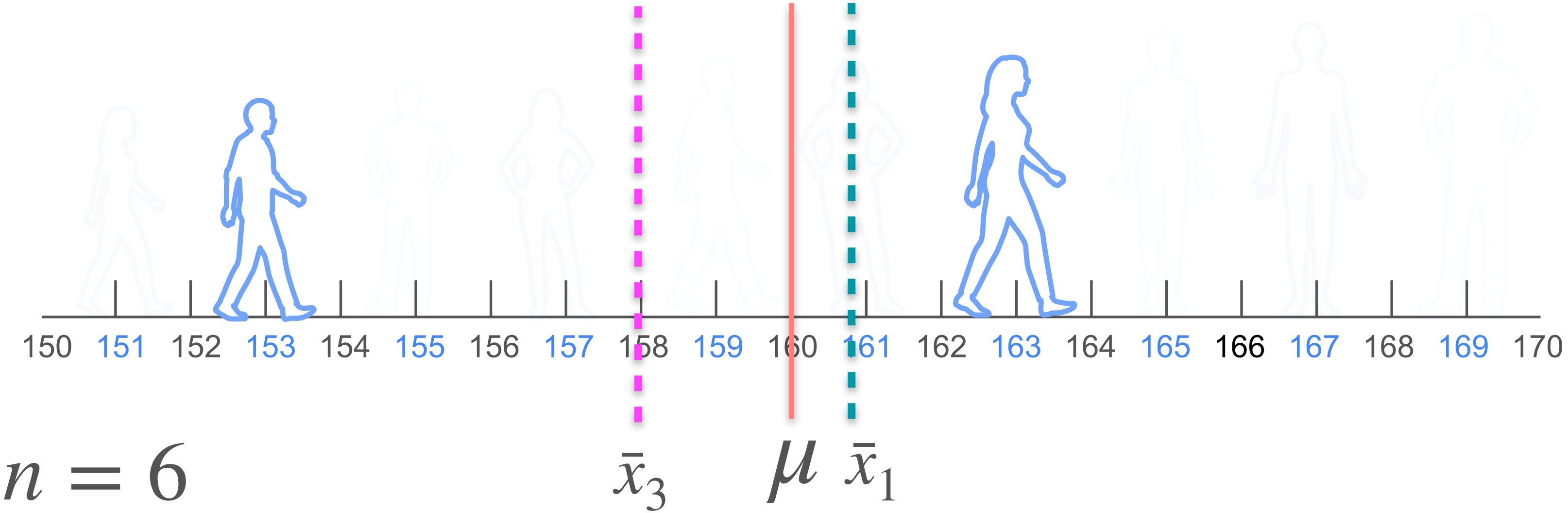
Population and Sample Mean



$n = 6$

$n = 2$

Population and Sample Mean



$n = 6$

$n = 2$

What is the average height in statistopia?

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

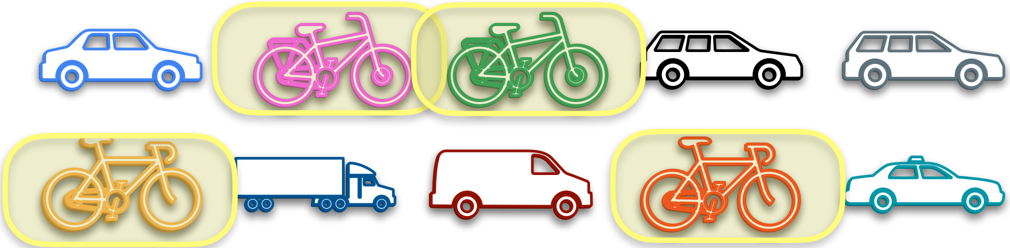
DeepLearning.AI

Sample and Population

Sample Proportion

Proportion

Population size: 10



What proportion of people own a bicycle?

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

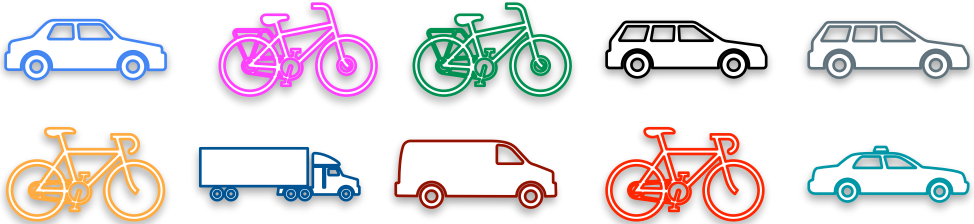
Proportion

population proportion

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

Proportion

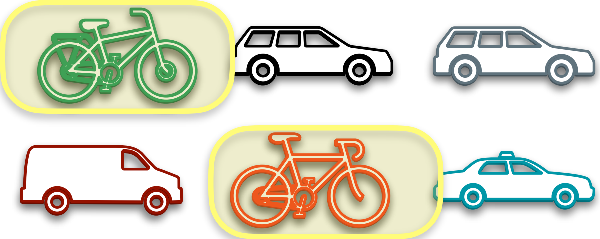
Population size: 10



Sample Proportion

Sample size: 6

estimate of the population proportion



What proportion of people own a bicycle?

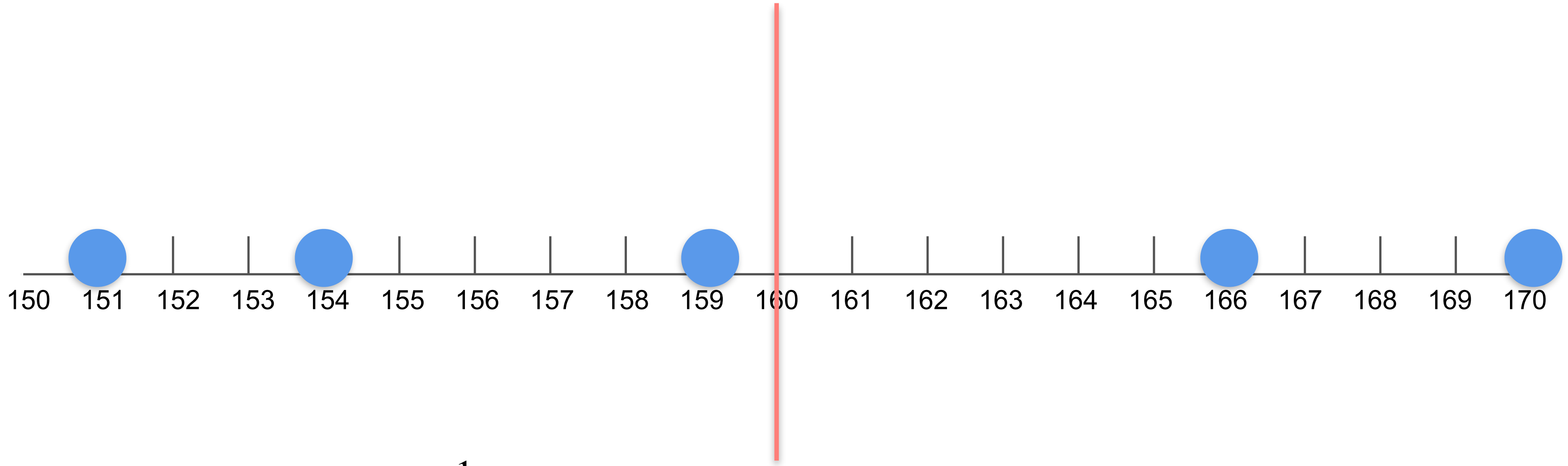
$\hat{p}$ sample proportion $= \frac{2}{6} = 0.333 = 33.3\,\%$

DeepLearning.AI

Sample and Population

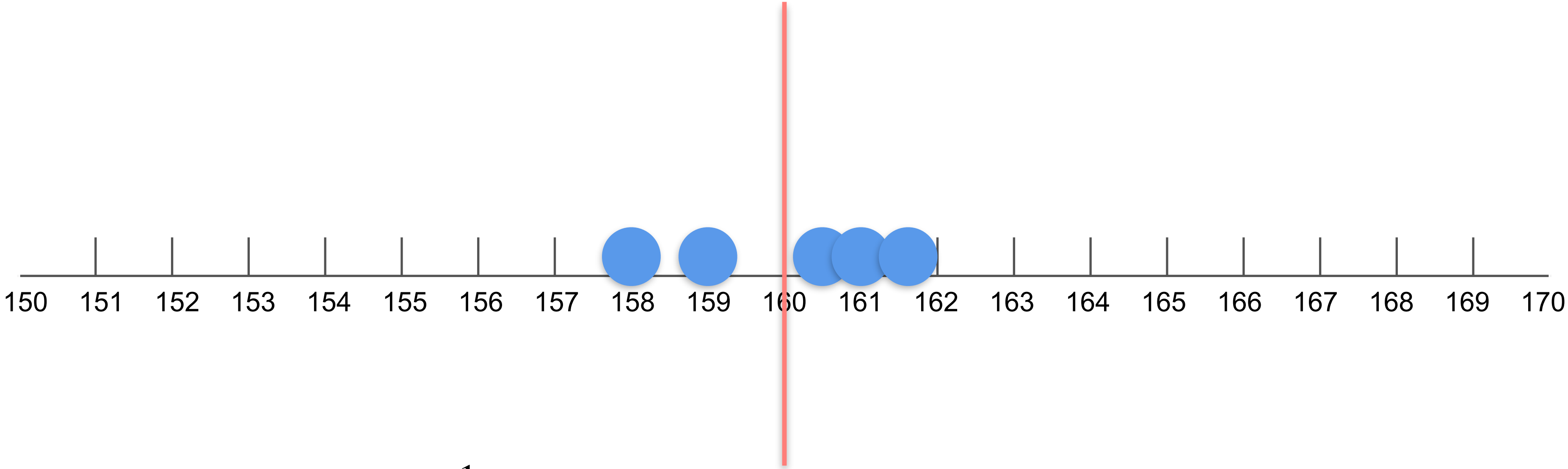
Sample Variance

Sample Variance



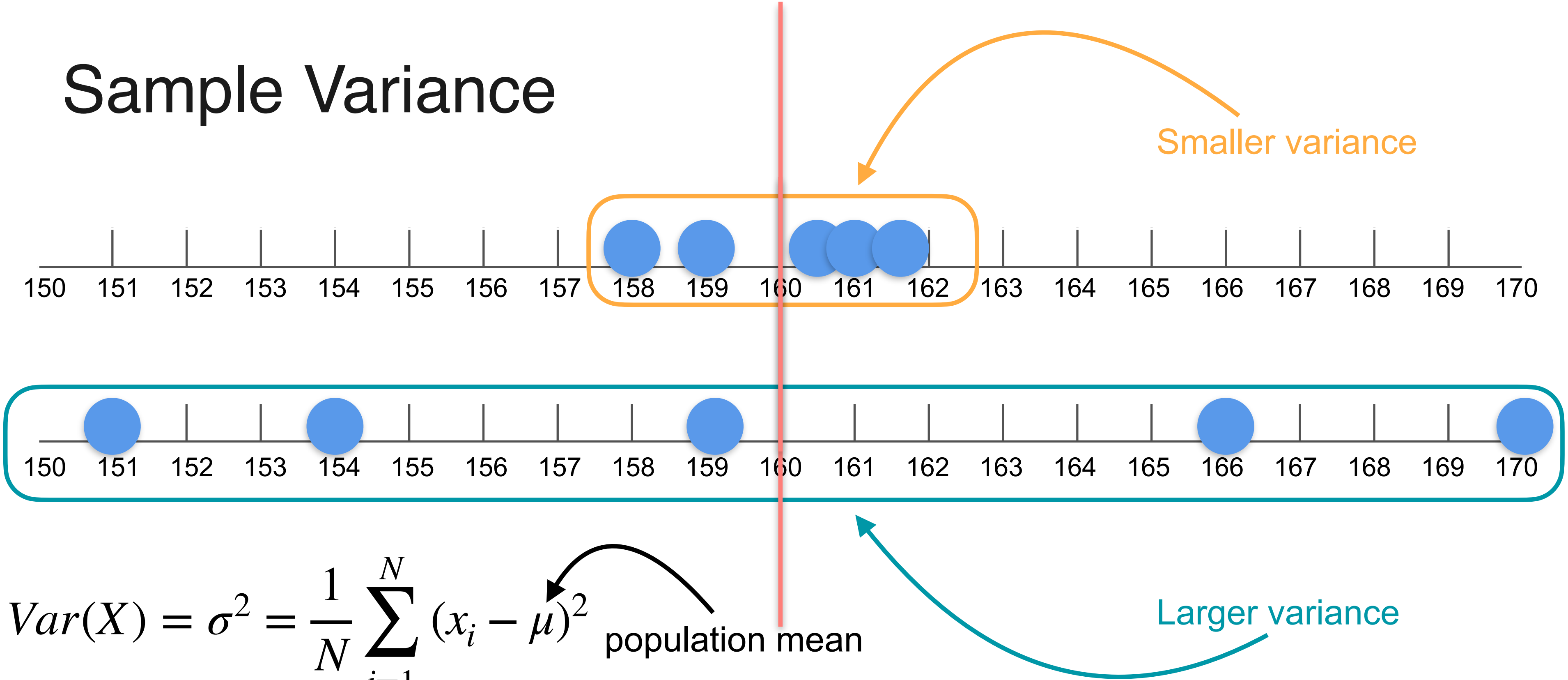
$$\mu = \frac{1}{5}\left(151 + 154 + 159 + 166 + 170\right) = 160$$

Sample Variance



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

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?

DeepLearning.AI

Sample Variance

Let's cheat and use the sample mean

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

$$Y = (X - \mu)^2$$

$$\mathbb{E}[Y] = \mu_Y = \frac{1}{N}\sum_{i=1}^{N} y_i \longrightarrow \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

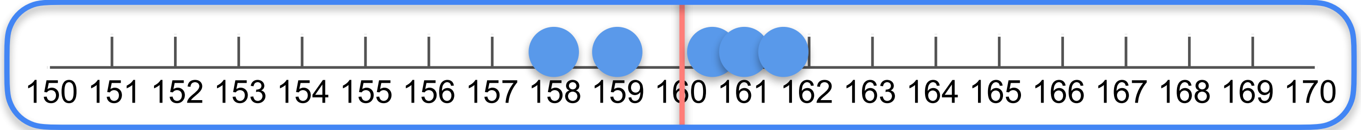
The population mean of Y

The sample mean of Y

Sample Variance

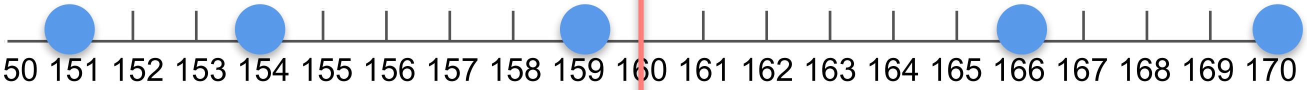
$$Var(X) = \sigma^2 = \frac{1}{N}\sum (x-\mu)^2 \longrightarrow \widehat{Var(X)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Smaller variance



$\widehat{Var(X)} = 1.7$

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



$\bar{x} = 160$

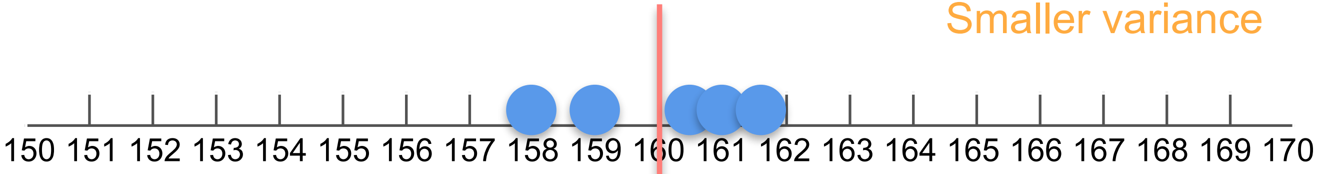
Larger variance

Sample Variance

This equation is "biased"
It underestimates the population variance

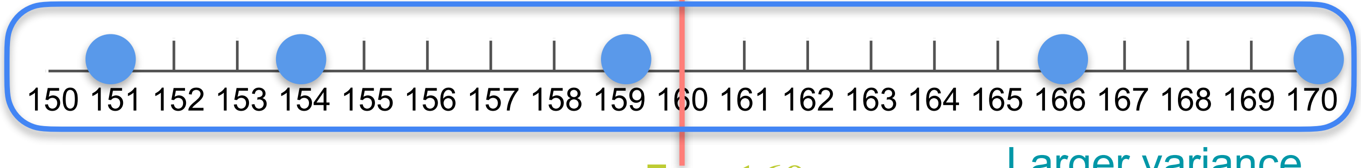
$$Var(X) = \sigma^2 = \frac{1}{N}\sum (x-\mu)^2 \longrightarrow \widehat{Var(X)} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Smaller variance



$\widehat{Var(X)} = 1.7$

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

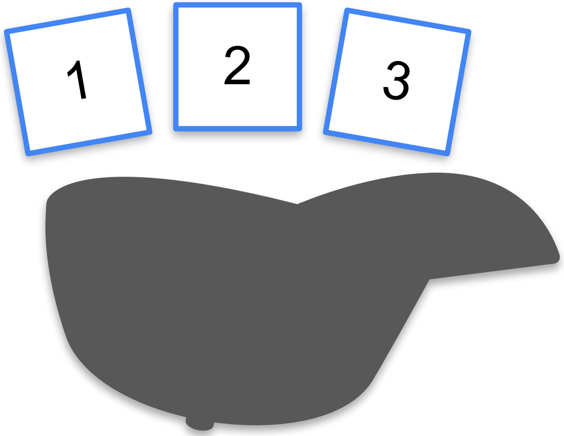


$\widehat{Var(X)} = 50.8$

Larger variance

$\bar{x} = 160$

Variance Estimation



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

Variance Estimation

1	2	3

$\mu = 2$

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

$x - 2$

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

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

σ^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

$\mu = 2$

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

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

$n = 2$

Samples		$\bar{x}$	$\widehat{Var}(X) = \dfrac{\sum (x-\bar{x})^2}{n}$
1	1	1	0
1	2	1.5	0.25
1	3	2	1
2	1	1.5	0.25
2	2	2	0
2	3	2.5	0.25
3	1	2	1
3	2	2.5	0.25
3	3	3	0

estimated variance

$$= 0.333$$

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

1	2	3

$\mu = 2$

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

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

$n = 2$

Samples		$\bar{x}$	$\widehat{Var(X)} = \dfrac{\sum (x-\bar{x})^2}{n\ \cancel{n}\ 1}$
1	1	1	0
1	2	1.5	0.25
1	3	2	1
2	1	1.5	0.25
2	2	2	0
2	3	2.5	0.25
3	1	2	1
3	2	2.5	0.25
3	3	3	0

estimated variance

$= 0.333$

$= \dfrac{1}{3}$

Variance Estimation

1	2	3

$\mu = 2$

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

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

$n = 2$ Samples		$\bar{x}$	$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$
1	1	1	0
1	2	1.5	0.5
1	3	2	2
2	1	1.5	0.5
2	2	2	0
2	3	2.5	0.5
3	1	2	2
3	2	2.5	0.5
3	3	3	0

estimated variance

$= 0.667$

$= \frac{2}{3}$

Variance Estimation

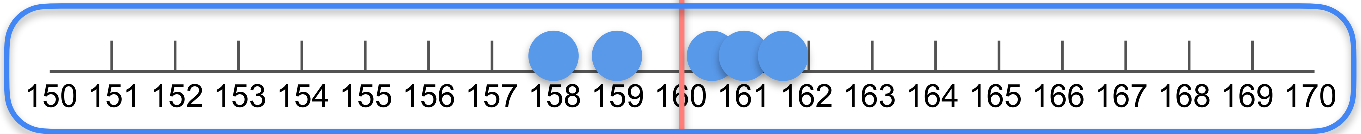
$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

- 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

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

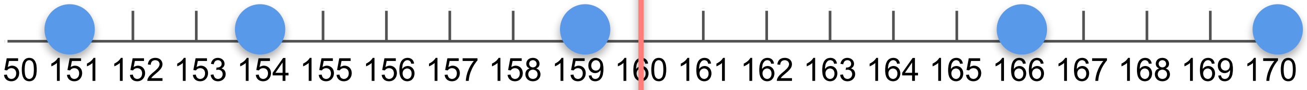
Smaller variance



$\widehat{Var(X)} = 1.7$

$s^2 = 2.125$

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



$\widehat{Var(X)} = 50.8$

$s^2 = 63.5$

$\bar{x} = 160$

Larger variance

Variance Estimation

Population Variance Formula

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

Sample Variance Formula

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

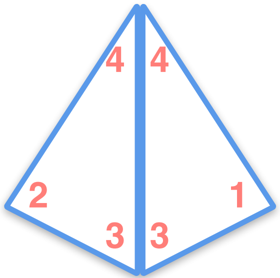
$$\widehat{\sigma^2} = \widehat{Var(X)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

DeepLearning.AI

Sample and Population

Law of Large Numbers

Law of Large Numbers



1 2 3 4

$\mu = 2.5$

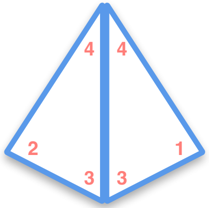
Experiment:

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

	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

<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th></tr>
<tr><td>1</td><td rowspan="4" colspan="4">μ
2.5</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td></tr>
</table>

Law of Large Numbers



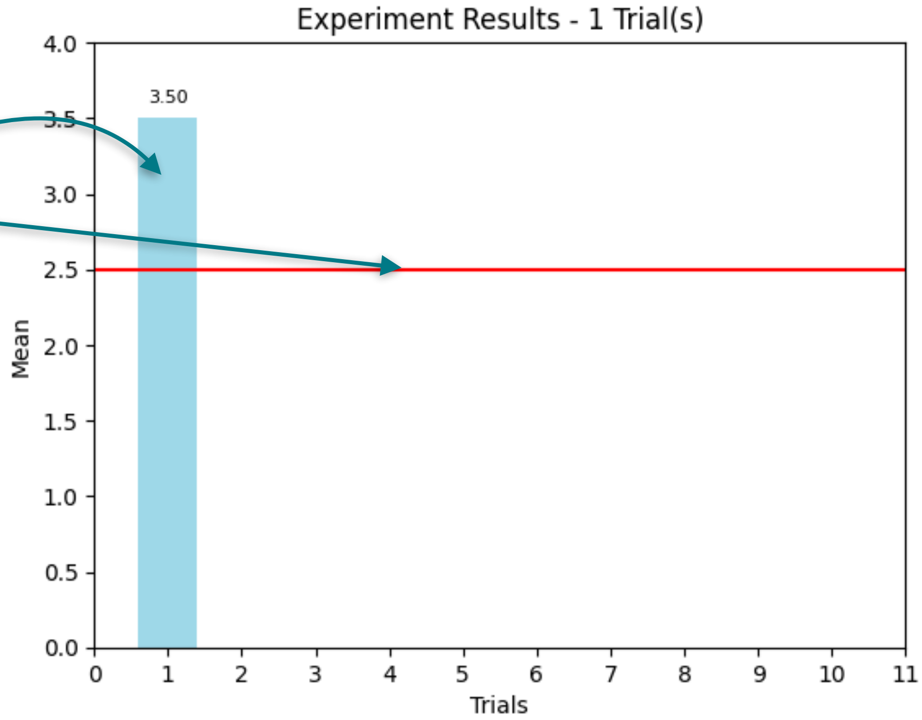
1 2 3 4

$\mu = 2.5$

1 trial

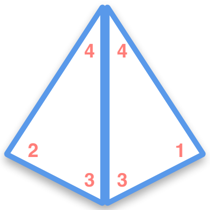
4,3

$\bar{x}_1$



	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

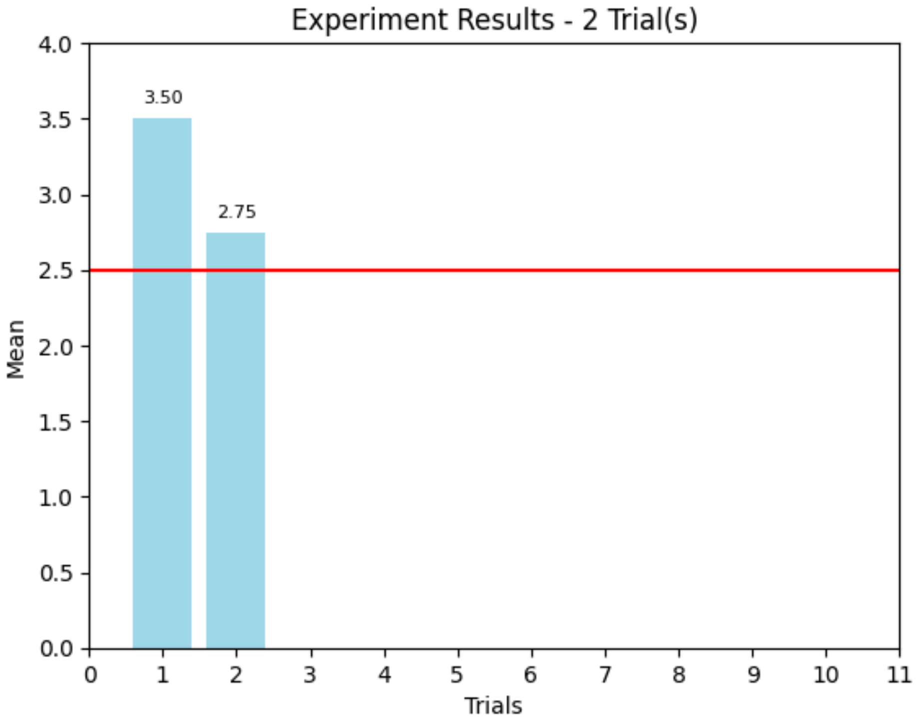
1 trial

4,3

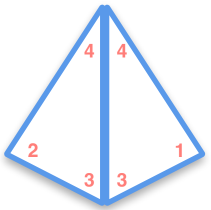
2 trials

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

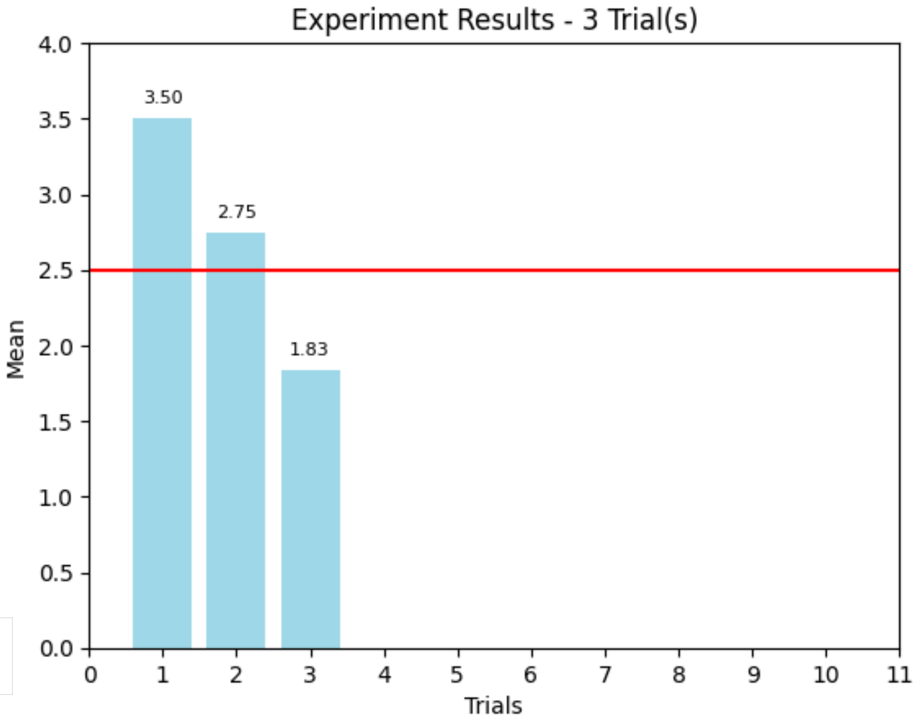
4,3

2 trials

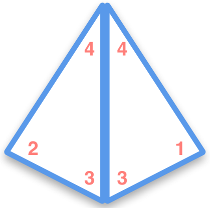
3,4
1,3

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

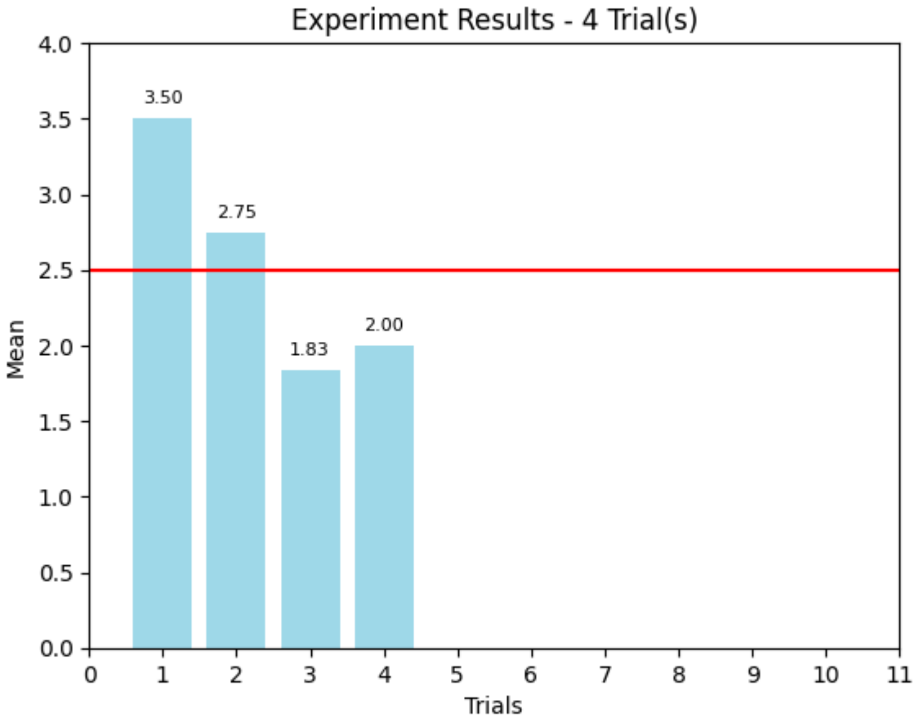
3,4
1,3

3 trials

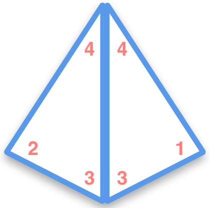
3,1	1,4	1,1

4 trials

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



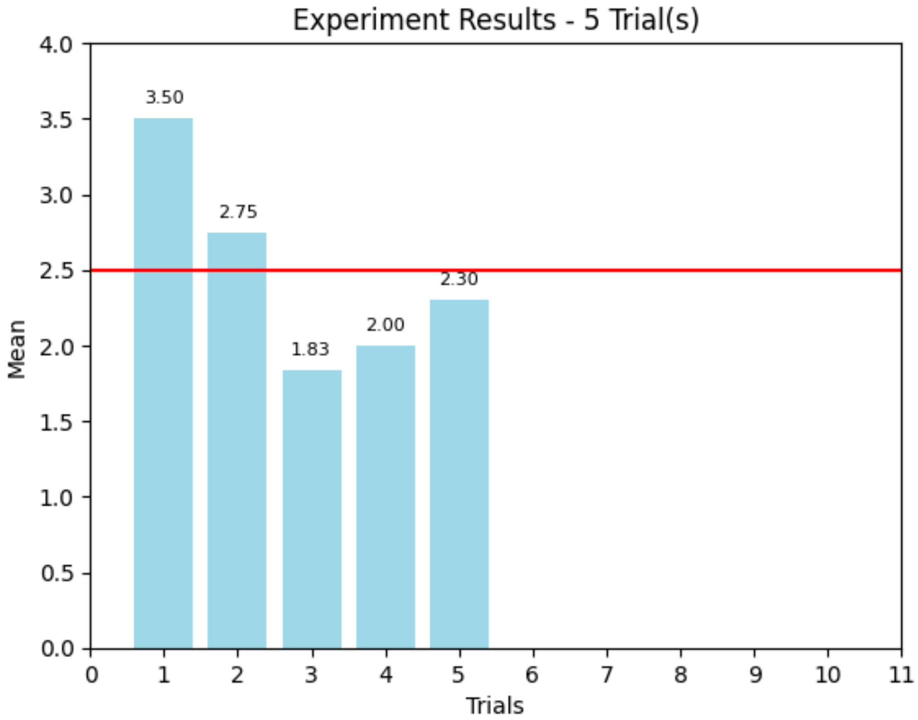
Law of Large Numbers



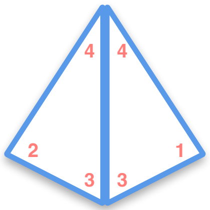
1 2 3 4

$\mu = 2.5$

	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



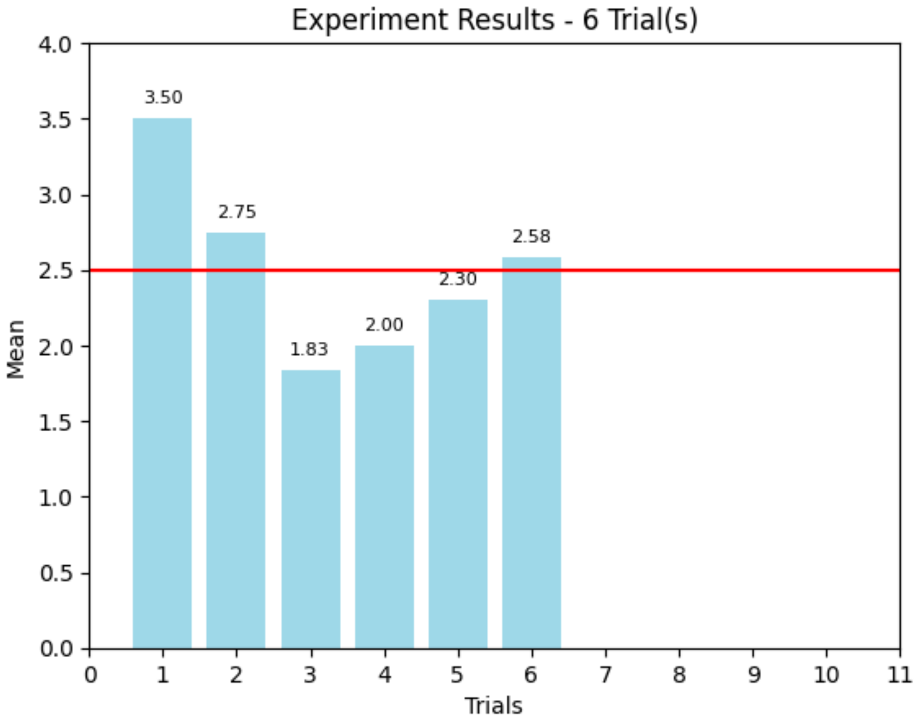
Law of Large Numbers



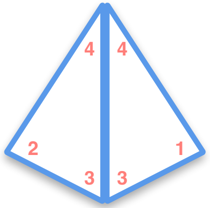
1 2 3 4

$\mu = 2.5$

	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



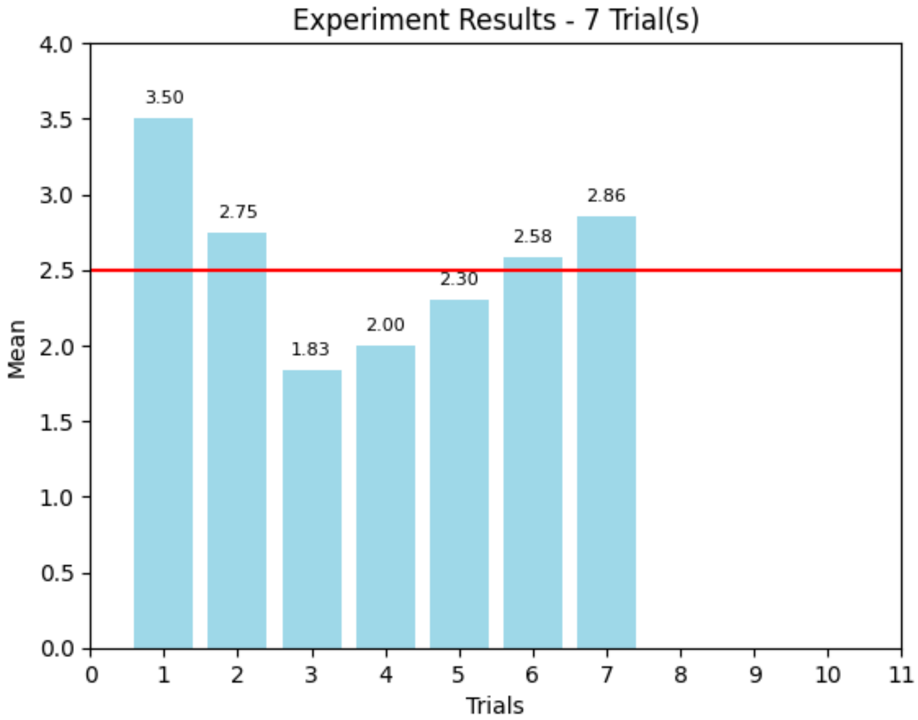
Law of Large Numbers



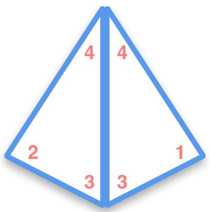
1 2 3 4

$$\mu = 2.5$$

	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



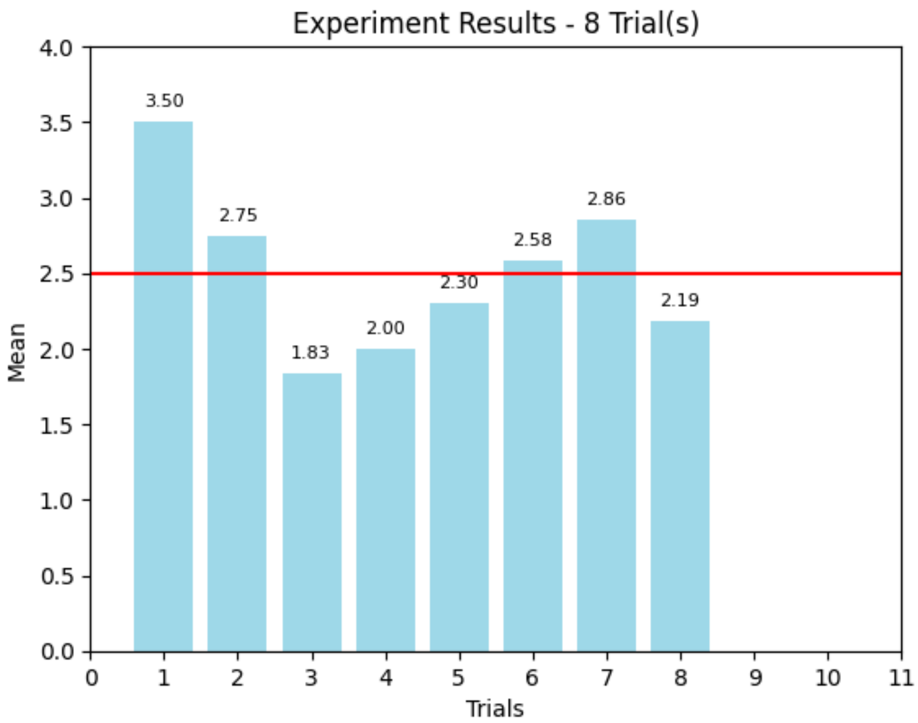
Law of Large Numbers



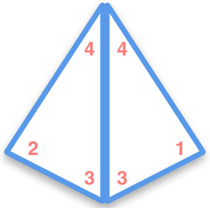
1 2 3 4

$$\mu = 2.5$$

	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



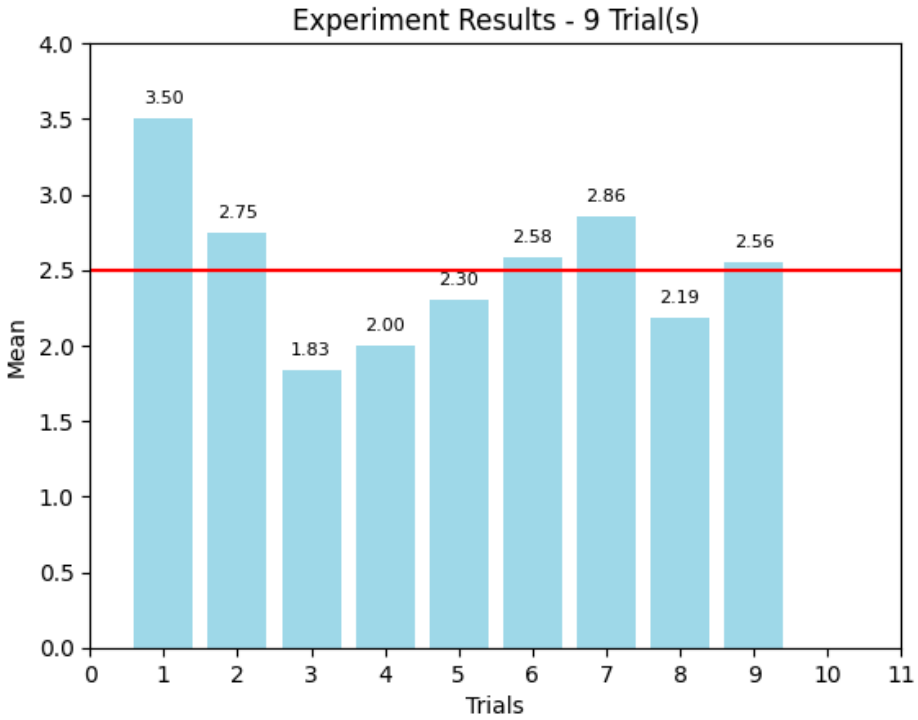
Law of Large Numbers



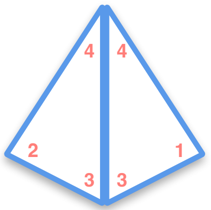
1 2 3 4

$$\mu = 2.5$$

	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



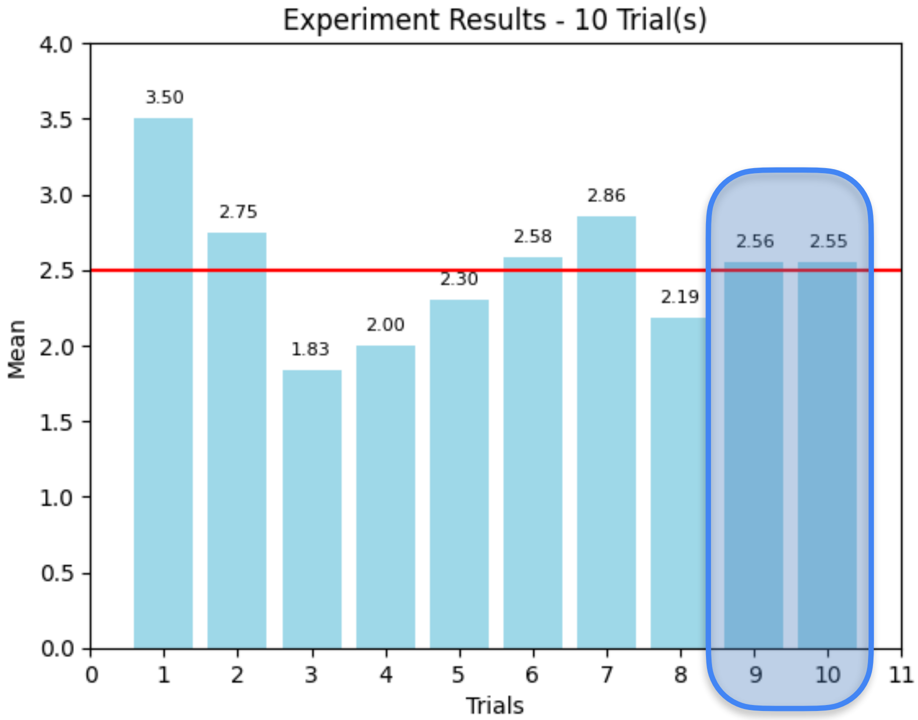
Law of Large Numbers



1 2 3 4

$$\mu = 2.5$$

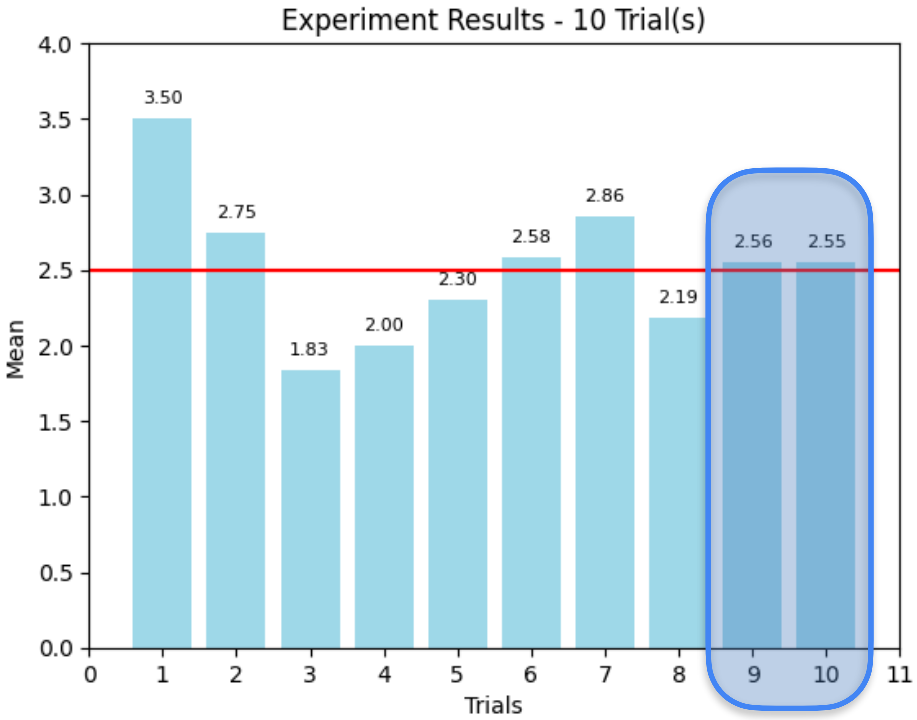
	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4



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 $n \to \infty$

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

UNDER CERTAIN CONDITIONS

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

Law of Large Numbers

UNDER CERTAIN CONDITIONS

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

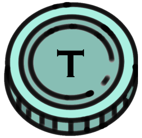
DeepLearning.AI

Sample and Population

Central Limit Theorem
Discrete Random Variable

Central Limit Theorem (CLT) - Example 1

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



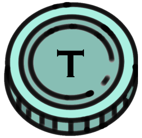
$\mathbf{P}(T) = 0.5$

Random variable → X **number of heads when a coin is flipped n times**

If $n = 1$ $X = 1$ $X = 0$

Discrete Random Variable

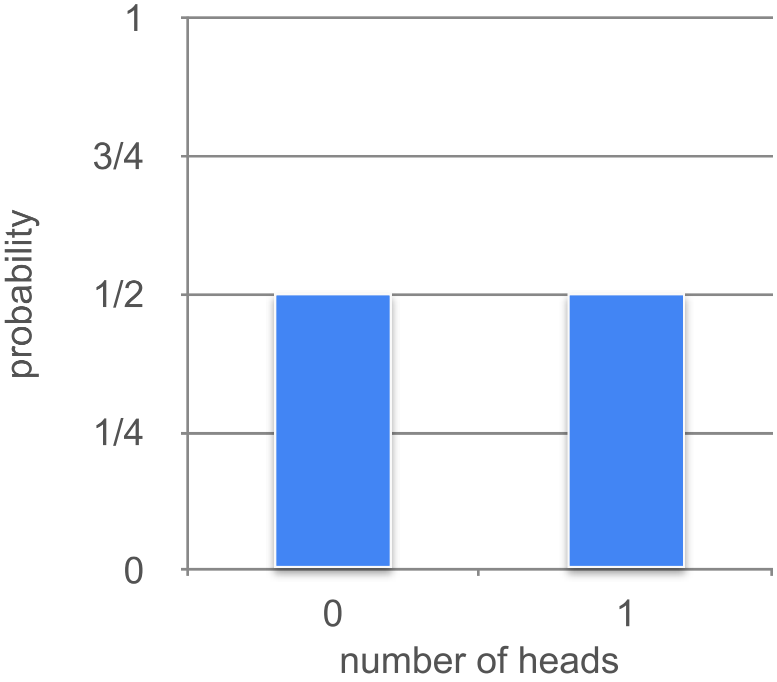
Central Limit Theorem (CLT) - Example 1



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

$X = 1$ $\quad$ $X = 0$

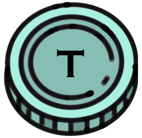
What can we say about the probability distribution when the number of coin flips increases?



Central Limit Theorem (CLT) - Example 1

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

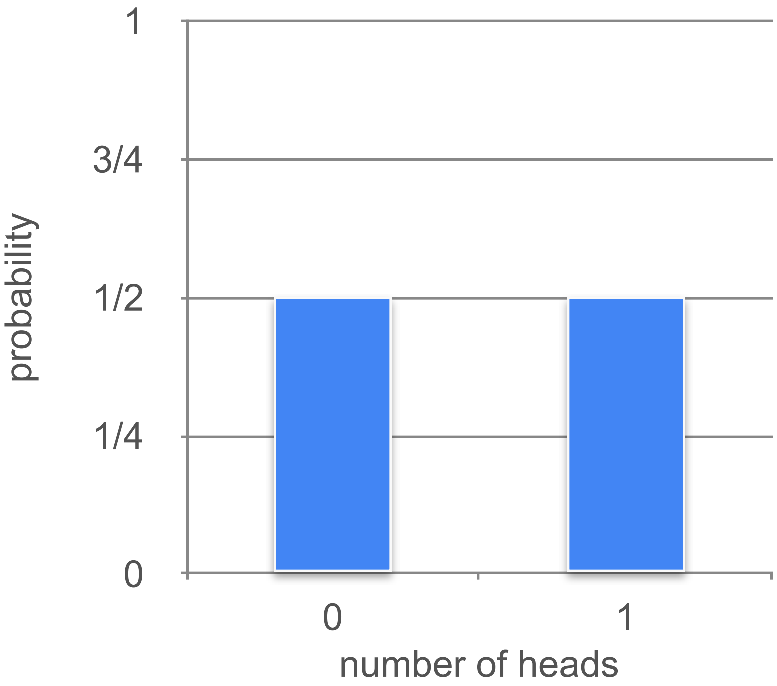
$X = 1$



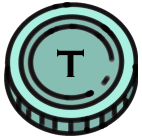
$\mathbf{P}(T) = 0.5$

$X = 0$

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



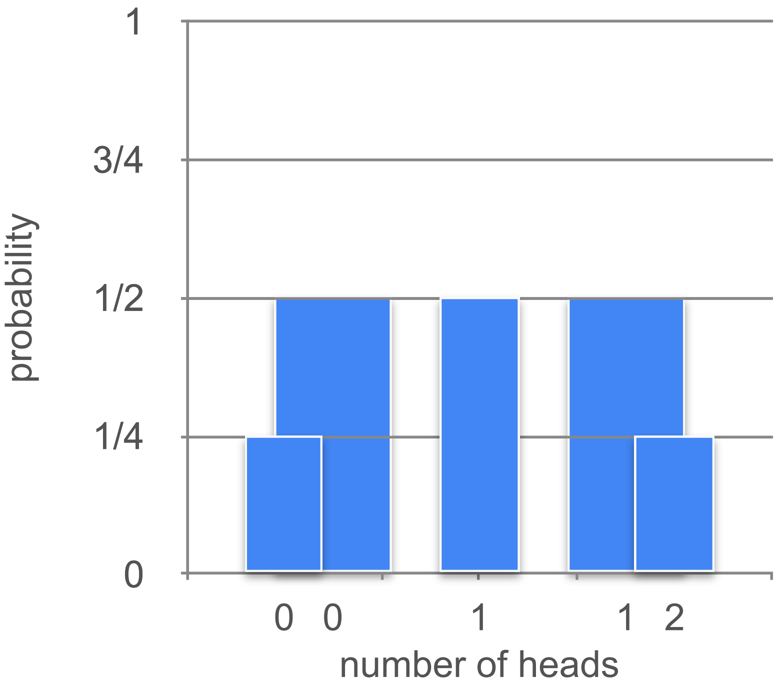
Central Limit Theorem (CLT) - Example 1



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

$X = 1$ $X = 0$

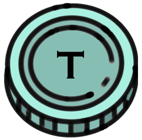
What can we say about the probability distribution when the number of coin flips increases?



Central Limit Theorem (CLT) - Example 1

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

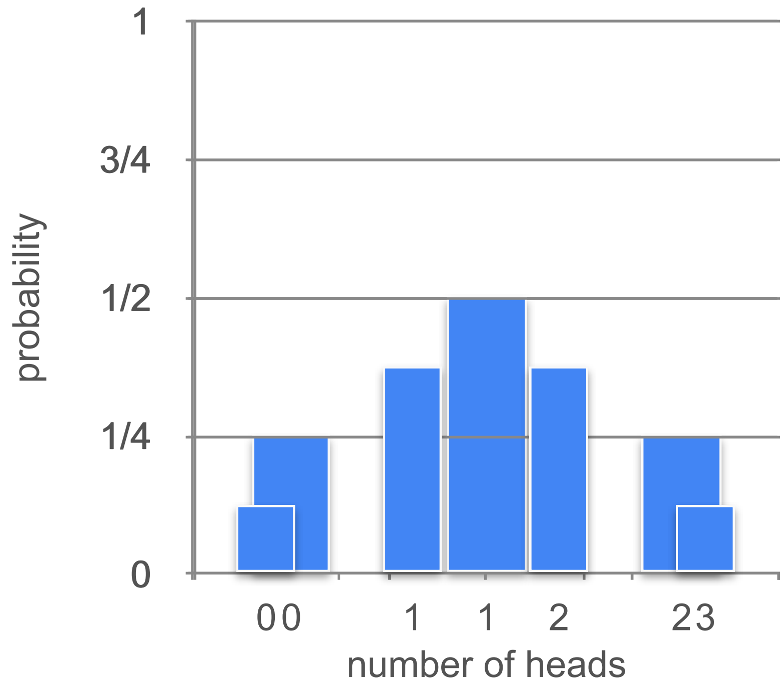
$X = 1$

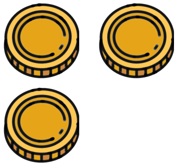


$\mathbf{P}(T) = 0.5$

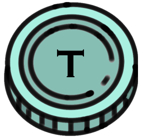
$X = 0$

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





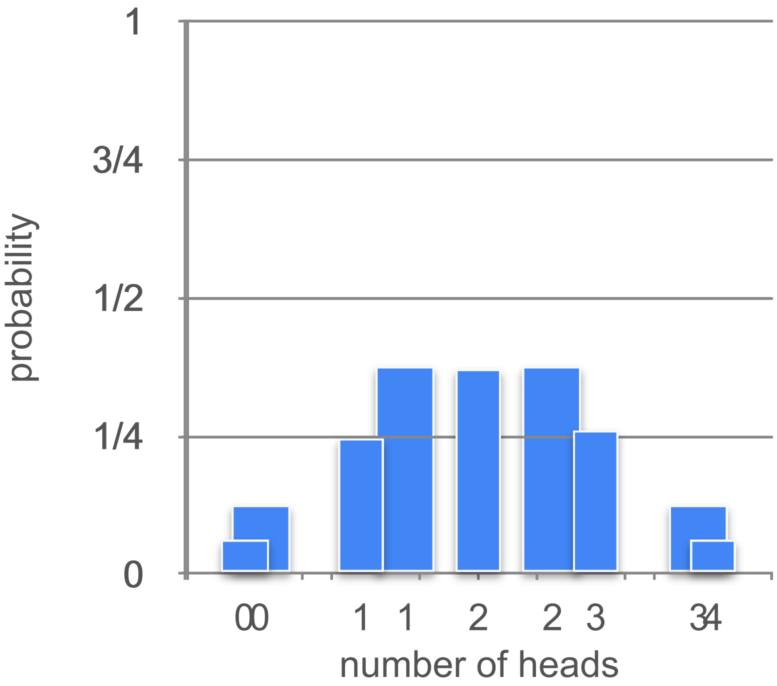
Central Limit Theorem (CLT) - Example 1

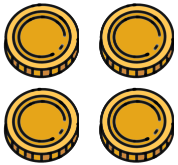


$\mathbf{P}(H) = 0.5$ $\mathbf{P}(T) = 0.5$

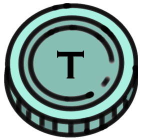
$X = 1$ $X = 0$

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





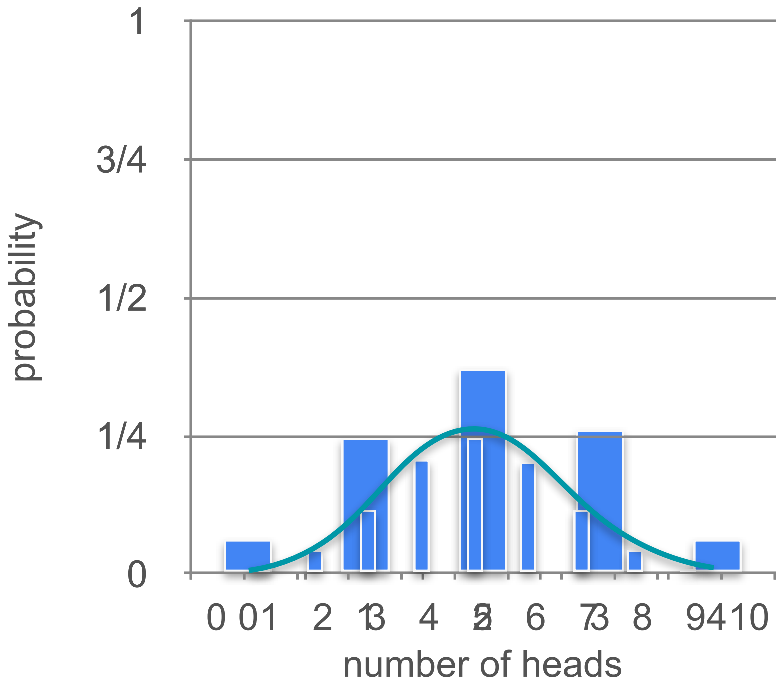
Central Limit Theorem (CLT) - Example 1

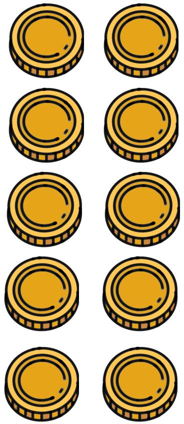


$\mathbf{P}(H) = 0.5$ $\mathbf{P}(T) = 0.5$

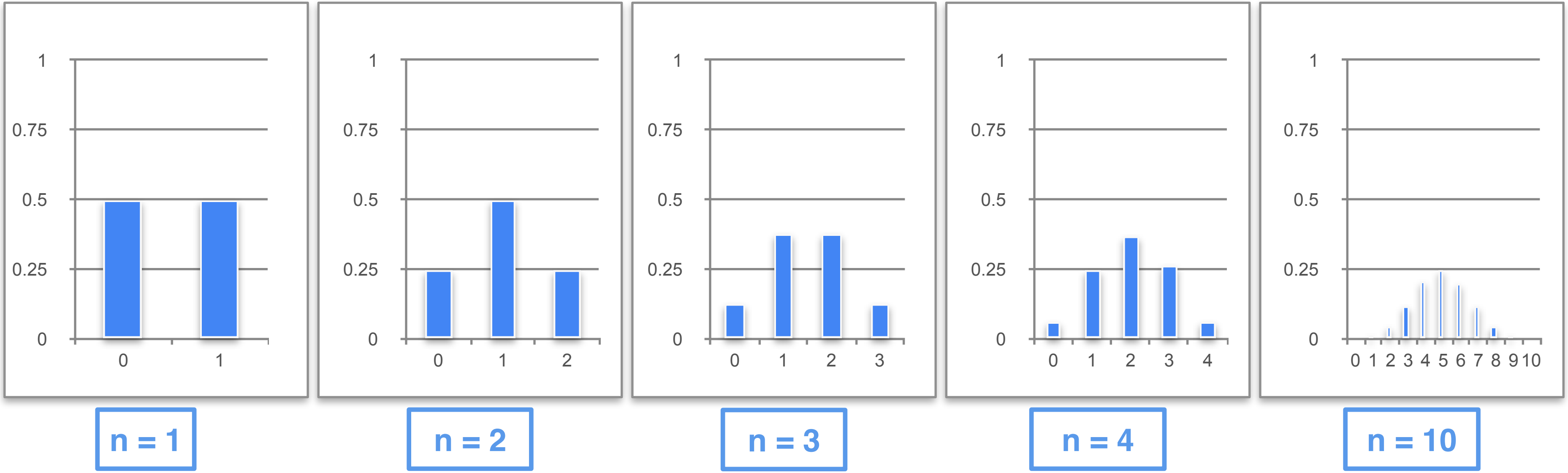
$X = 1$ $X = 0$

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





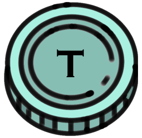
Central Limit Theorem (CLT) - Example 1



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

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

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

Central Limit Theorem (CLT) - Example 1

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

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

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

$n = 1$

$$\mu = np$$

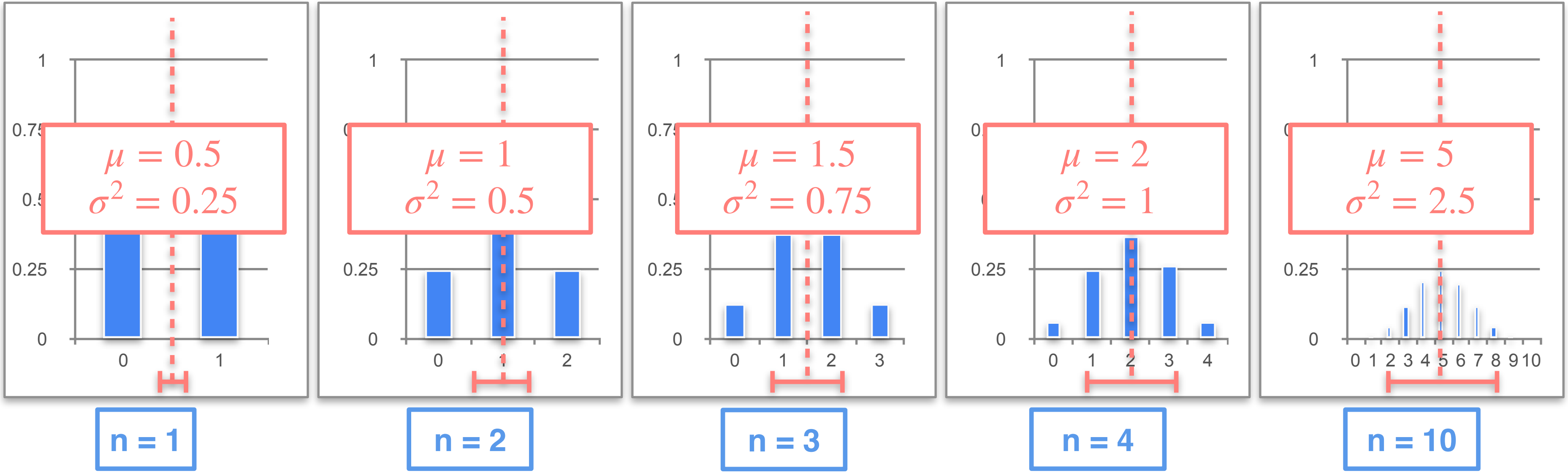
$$\mu = 1 \times 0.5 = 0.5$$

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

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

Central Limit Theorem (CLT) - Example 1

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



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

Central Limit Theorem (CLT) - Example 1

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

$\mu = 0.5$ $\sigma^2 = 0.25$	$\mu = 1$ $\sigma^2 = 0.5$	$\mu = 1.5$ $\sigma^2 = 0.75$	$\mu = 2$ $\sigma^2 = 1$	$\mu = 5$ $\sigma^2 = 2.5$

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

$$\mu = np$$

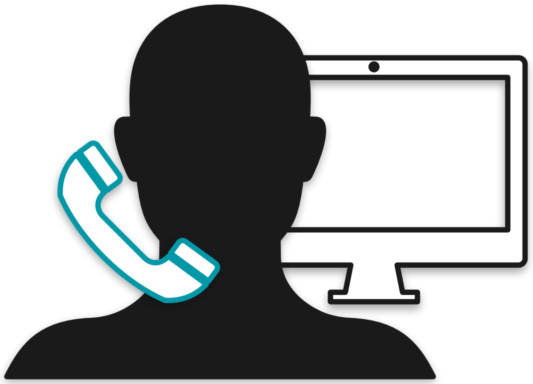
$$\sigma^2 = np(1-p)$$

DeepLearning.AI

Sample and Population

Central Limit Theorem
Contiuous 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

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

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

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

$\vdots \qquad \vdots$

Record the average of all n experiments

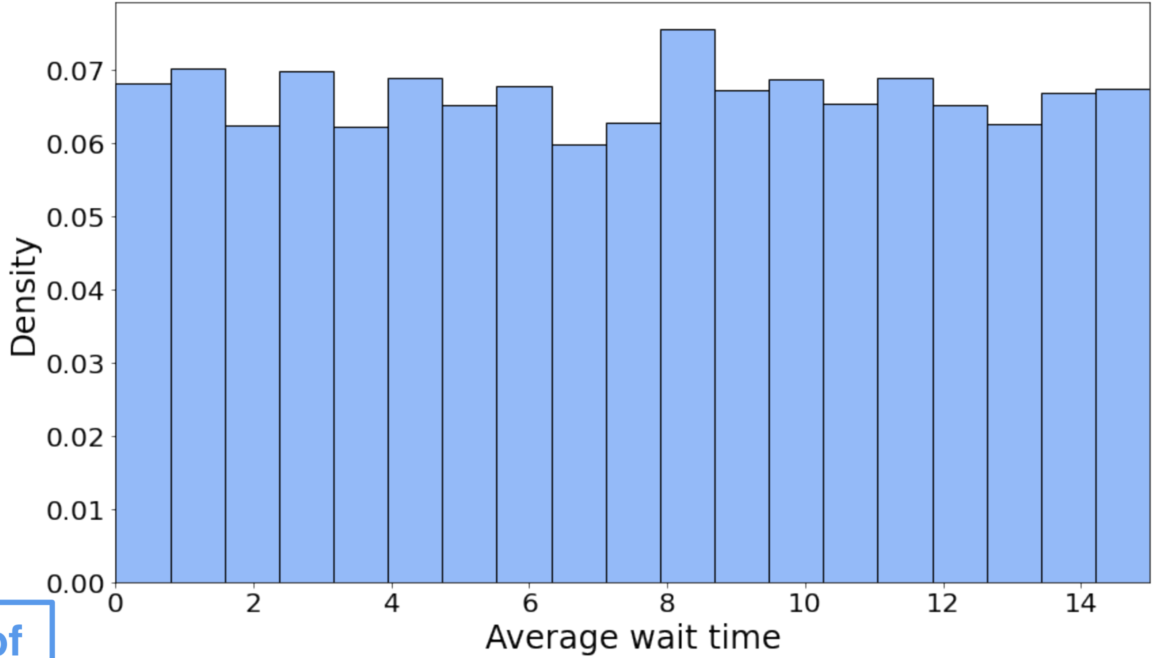
$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Can we say anything about the distribution of this average?

Central Limit Theorem (CLT) - Example 2

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

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

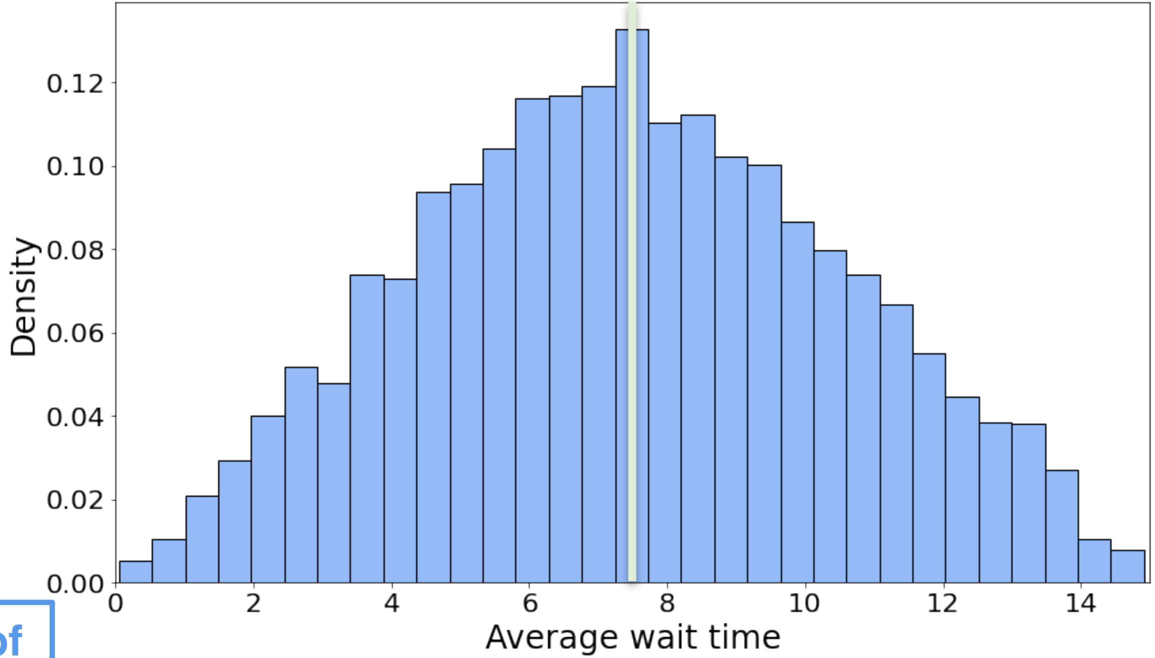


What happens to the distribution of these averages as n increases?

Central Limit Theorem (CLT) - Example 2

$n = 2$ $\qquad$ $Y_2 = \dfrac{X_1 + X_2}{2}$

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

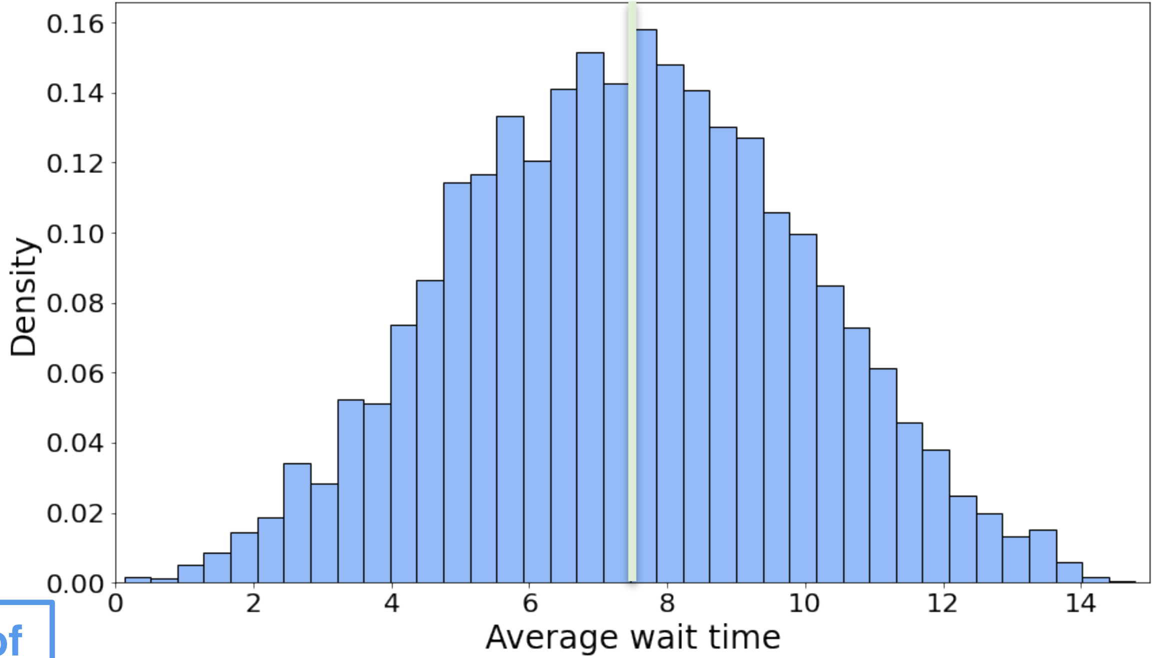


What happens to the distribution of these averages as n increases

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

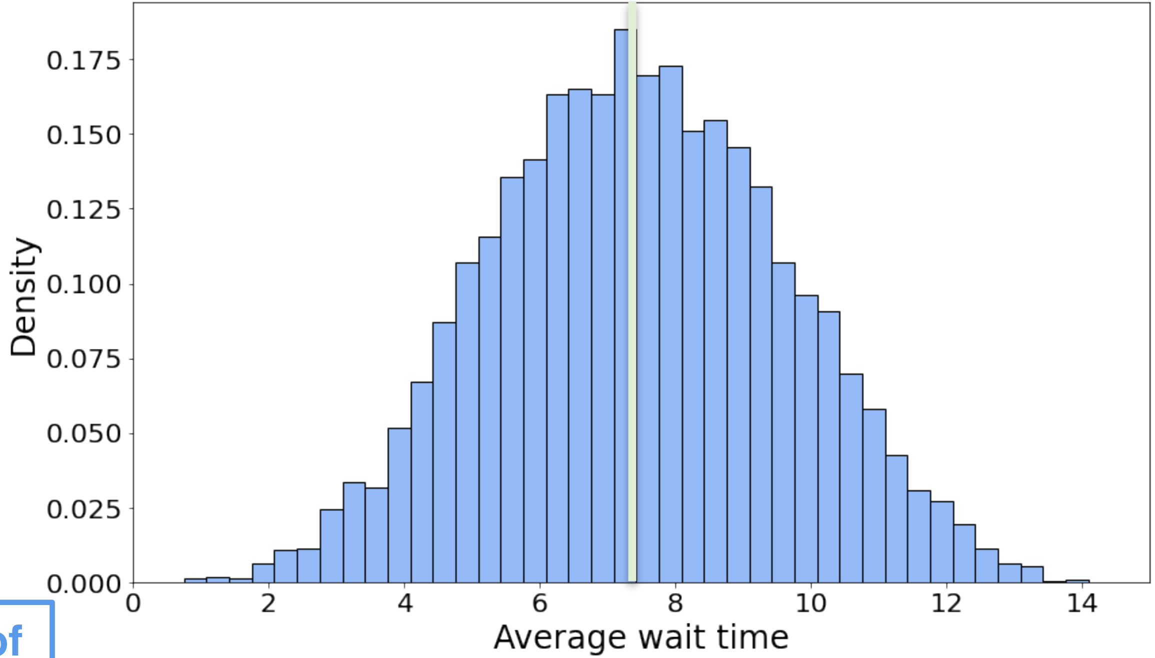


What happens to the distribution of these averages as n increases

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

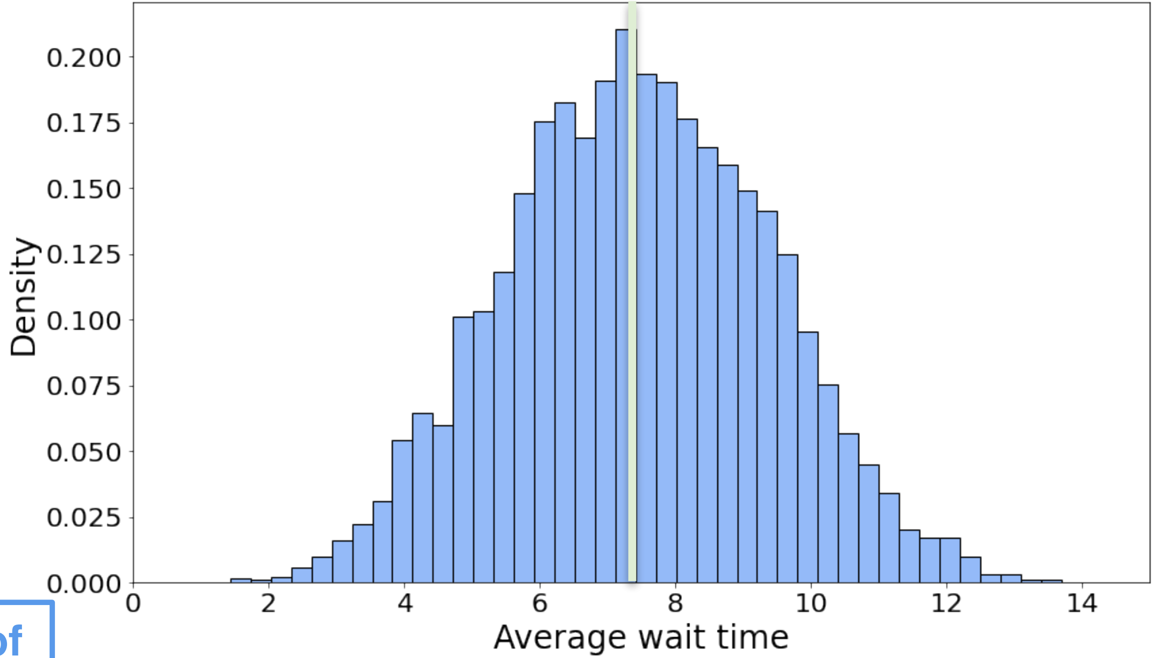


What happens to the distribution of these averages as n increases

Central Limit Theorem (CLT) - Example 2

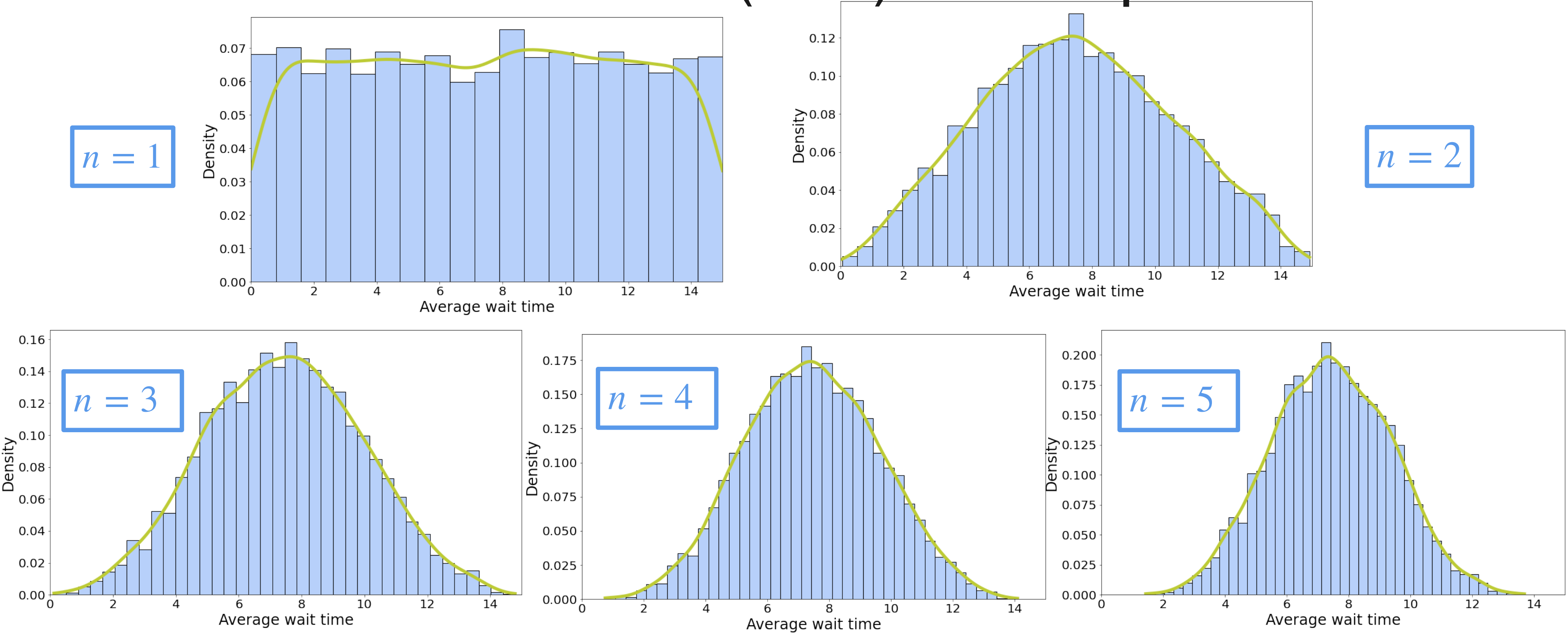
$$n = 5 \quad Y_5 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

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



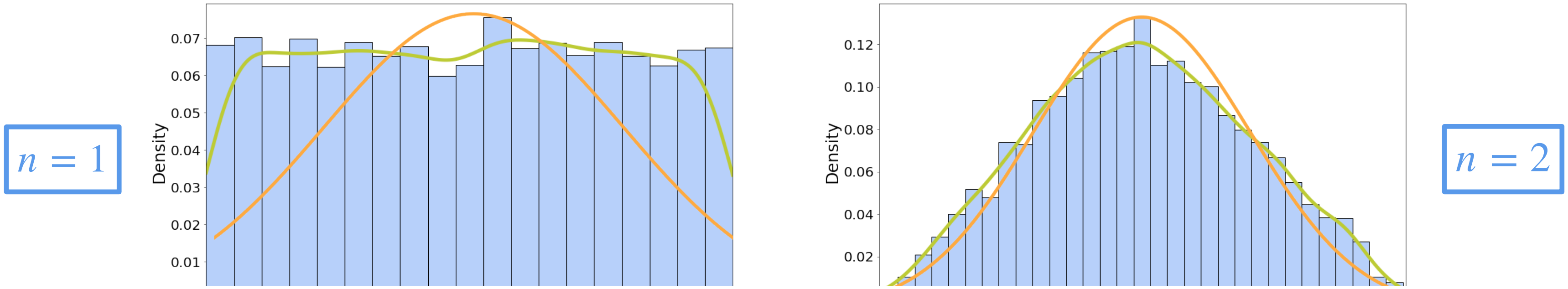
Central Limit Theorem (CLT) - Example 2

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

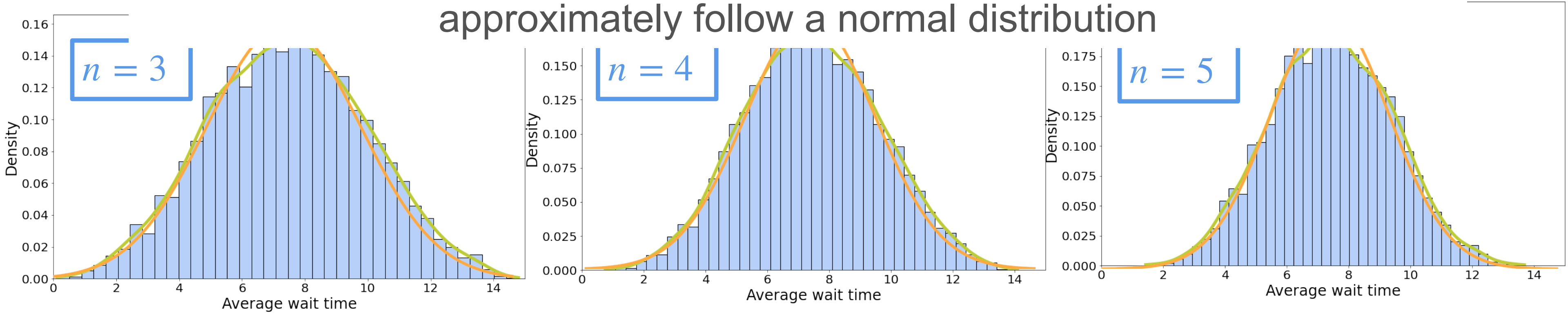
$$Var[Y_n] = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{1}{n^2} n Var(X) = \frac{Var(X)}{n} = \frac{18.75}{n}$$

Central Limit Theorem (CLT) - Example 2



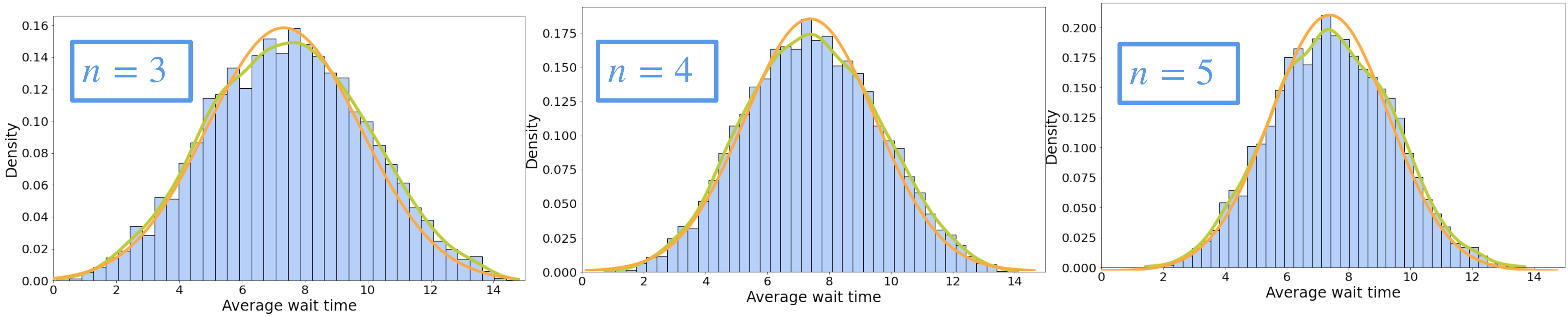
When you average a large enough number of variables, the distribution will approximately follow a normal distribution



DeepLearning.AI

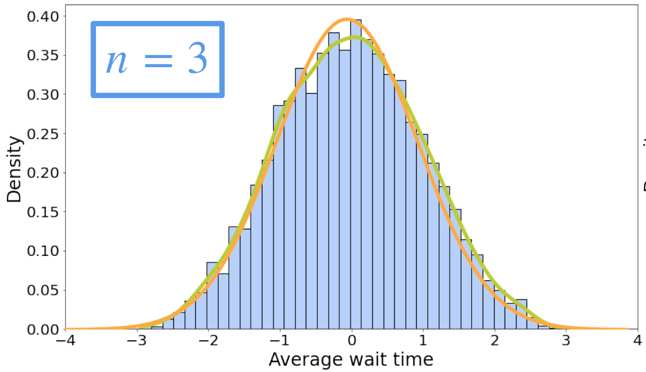
Central Limit Theorem (CLT) - Example 2

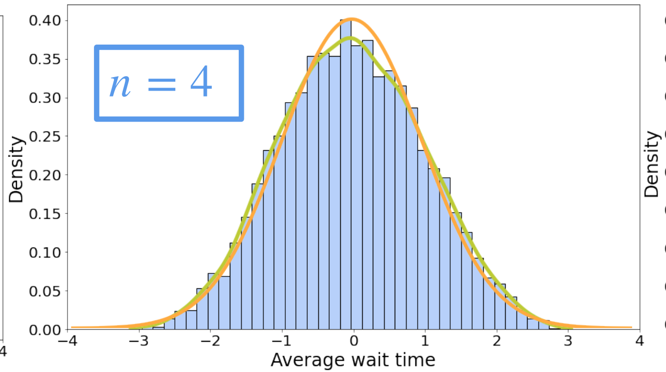
$$\frac{Y_n - 7.5}{\sqrt{18.75/n}}$$

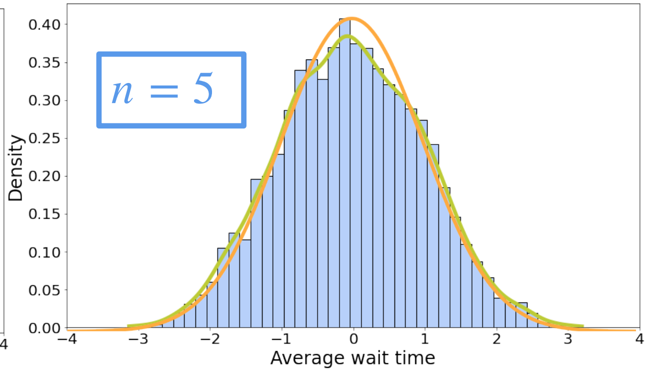


Central Limit Theorem (CLT) - Example 2

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







Central Limit Theorem (CLT) - Example 2

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

mean of Y_n $\rightarrow$ $\mu_{Y_n} = \mu$ $\leftarrow$ population mean

variance of Y_n $\rightarrow$ $\sigma_{Y_n}^2 = \frac{\sigma^2}{n}$

population variance

Central Limit Theorem (CLT) - Formal Definition

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

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

Central Limit Theorem (CLT) - Formal Definition

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

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

W3 Lesson 2

DeepLearning.AI

Point Estimation

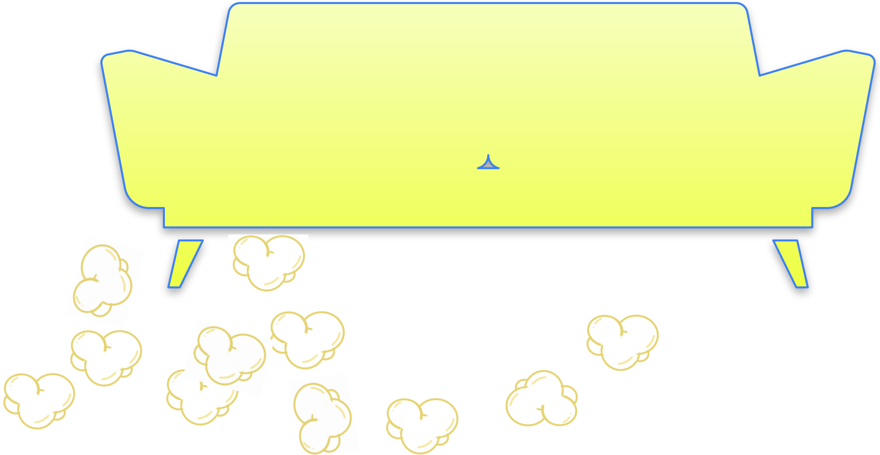
Point Estimation

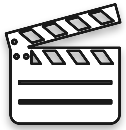
DeepLearning.AI

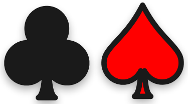
Point Estimation

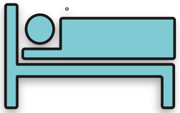
Maximum Likelihood Estimation: Motivation

There’s Popcorn on the Floor. What Happened?



Movies 

Board Games 

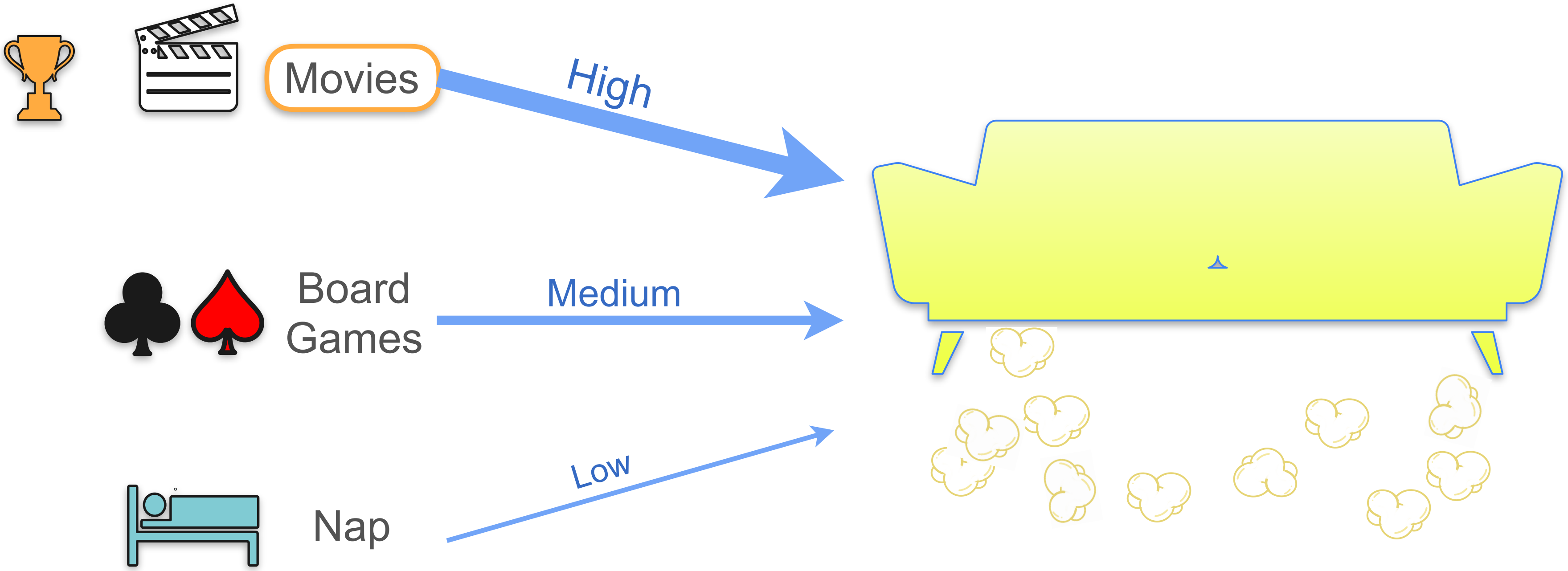
Nap 

Quiz

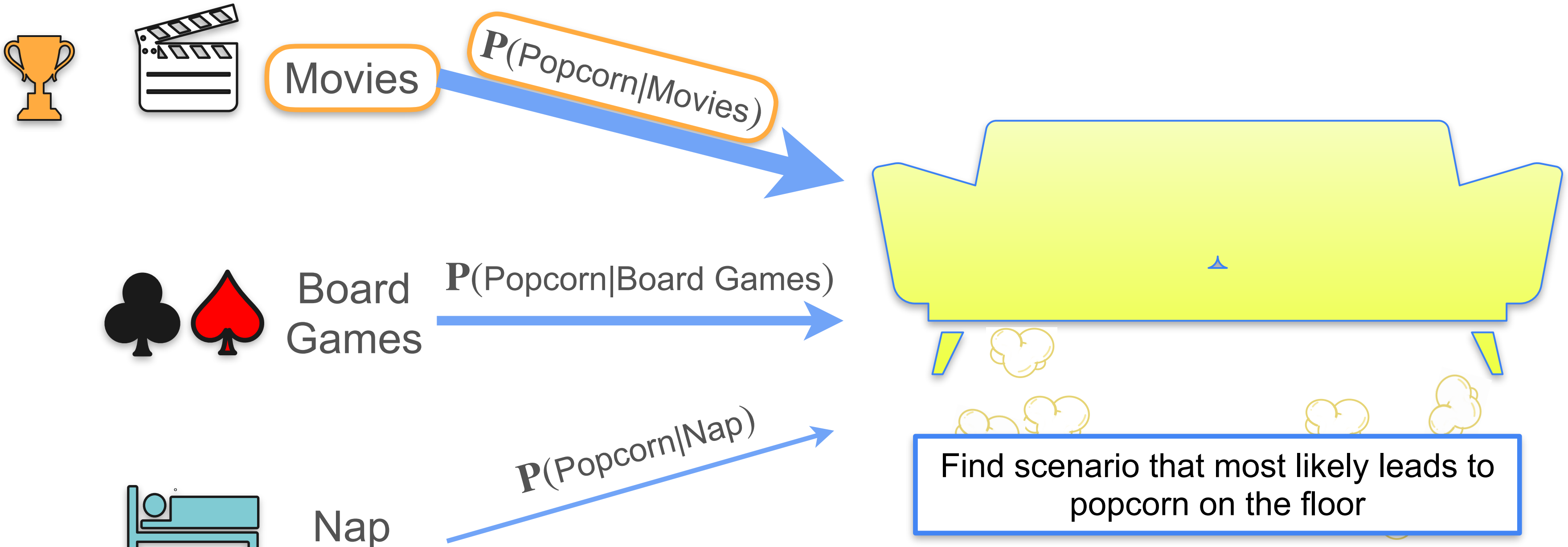
What do you think happened?

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?



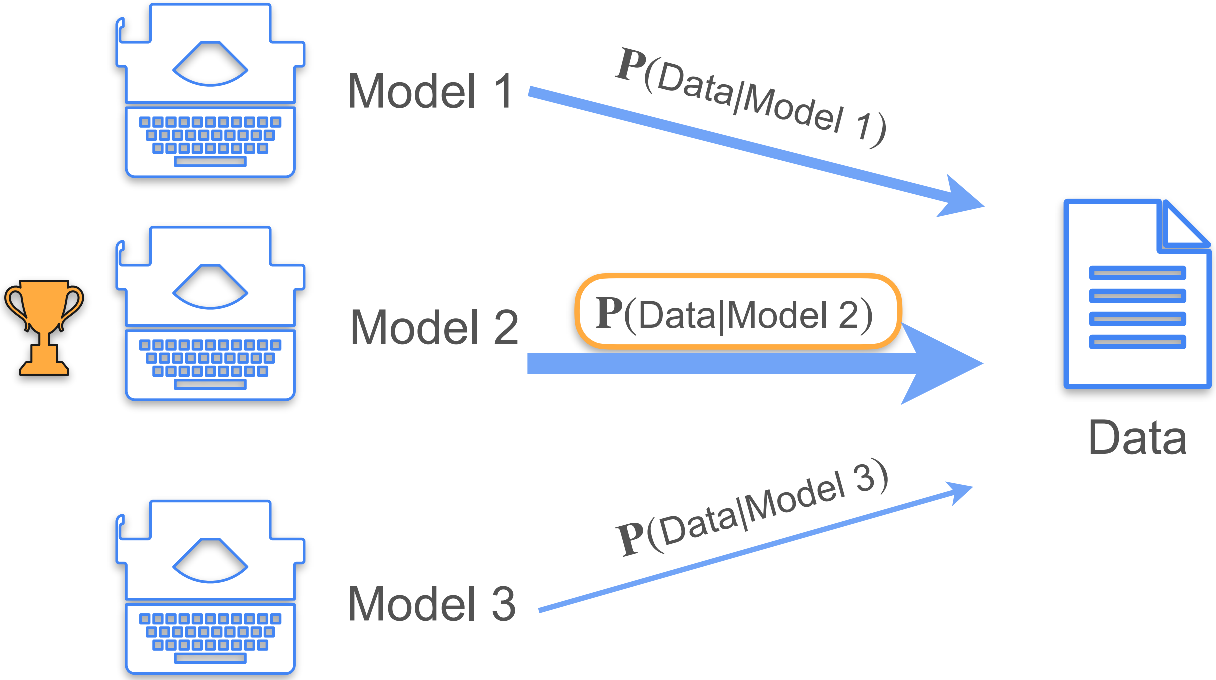
There's Popcorn on the Floor. What Happened?



Find scenario that most likely leads to popcorn on the floor

Maximum Likelihood

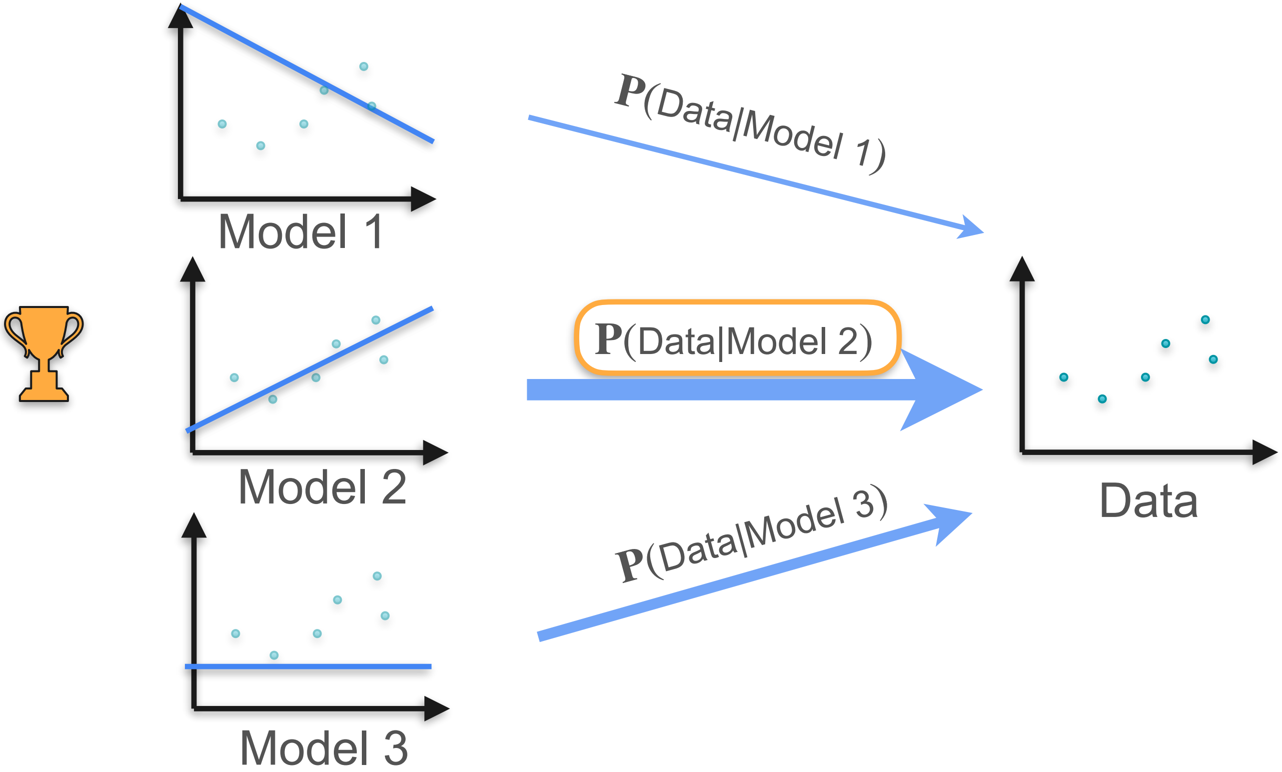
Maximum Likelihood



Find model that most likely produced the data

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

Example: Linear Regression

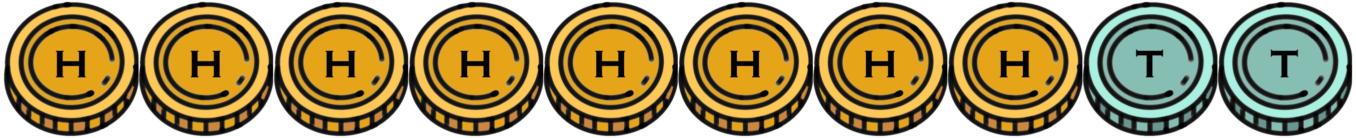


Find model that most likely produced the data

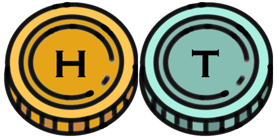
Point Estimation

MLE: Bernoulli Example

Maximum Likelihood: Bernoulli Example

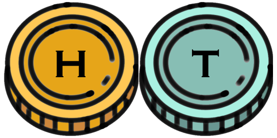


Coin 1



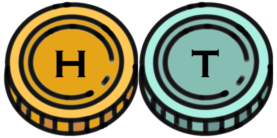
$\mathbf{P}(H) = 0.7$
$\mathbf{P}(T) = 0.3$

Coin 2



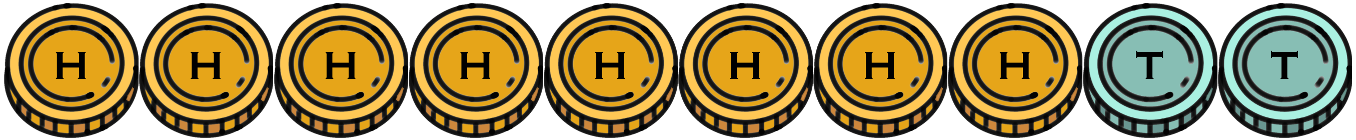
$\mathbf{P}(H) = 0.5$
$\mathbf{P}(T) = 0.5$

Coin 3



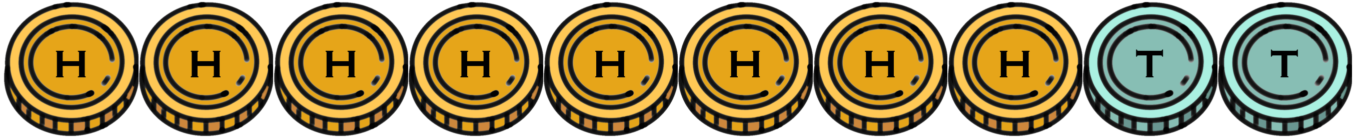
$\mathbf{P}(H) = 0.3$
$\mathbf{P}(T) = 0.7$

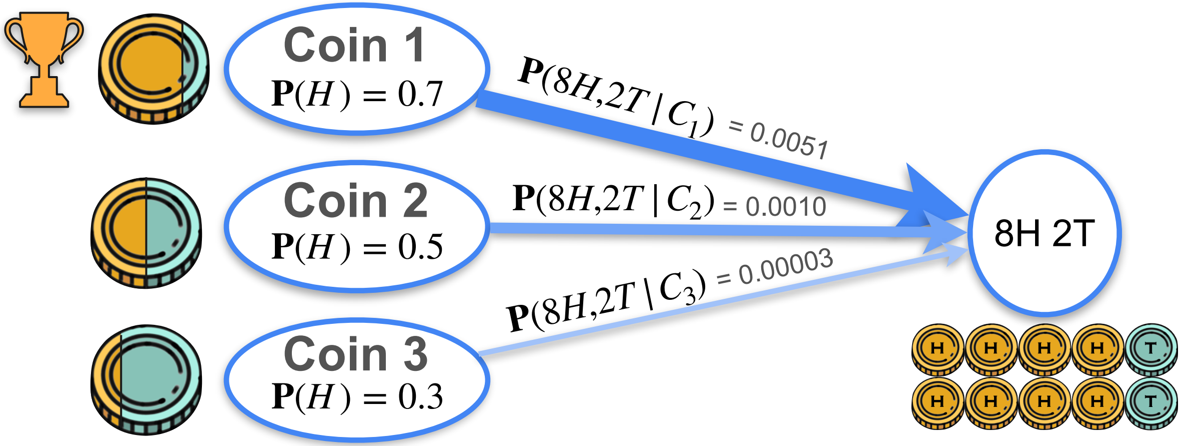
Maximum Likelihood: Bernoulli Example



Coin 1	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.3	0.3 = 0.0051
Coin 2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5 = 0.0010
Coin 3	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



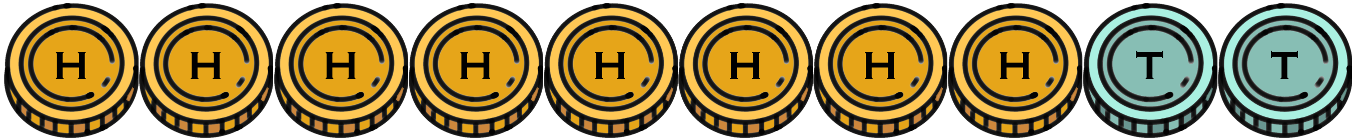


Find coin that most likely produced the 8 heads and 2 tails

Maximize

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

Maximum Likelihood: Bernoulli Example



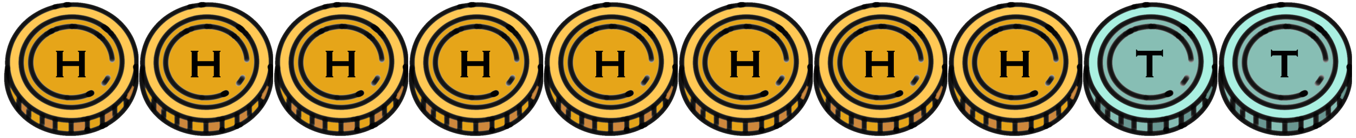
$p \quad p \quad p \quad p \quad p \quad p \quad p \quad p \quad 1-p \quad 1-p$

Can you do any better?

$p = \mathbf{P}(H)$ $\qquad$ $p^8(1-p)^2$

You want p that maximizes the chances of seeing 8H

Maximum Likelihood: Bernoulli Example



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

Function of p

You want p that maximizes the chances of seeing 8H

Log-likelihood $\ell(p; 8H) = \log\left((p^8(1-p)^2\right) = 8\log(p) + 2\log(1-p)$

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

Maximum Likelihood: Bernoulli Example

n coins

k heads

$X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_{n-1} \quad X_n$

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

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

Likelihood

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

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

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

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

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

Maximum Likelihood: Bernoulli Example

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

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

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

$X_1 \quad X_2 \quad X_3 \quad \cdots \quad X_{n-1} \quad X_n$

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

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

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

Find the maximum!

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

$$= \frac{\sum_{i=1}^n x_i}{p} + \frac{n - \sum_{i=1}^n x_i}{1-p}(-1) = 0 \quad \rightarrow \quad \hat{p} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

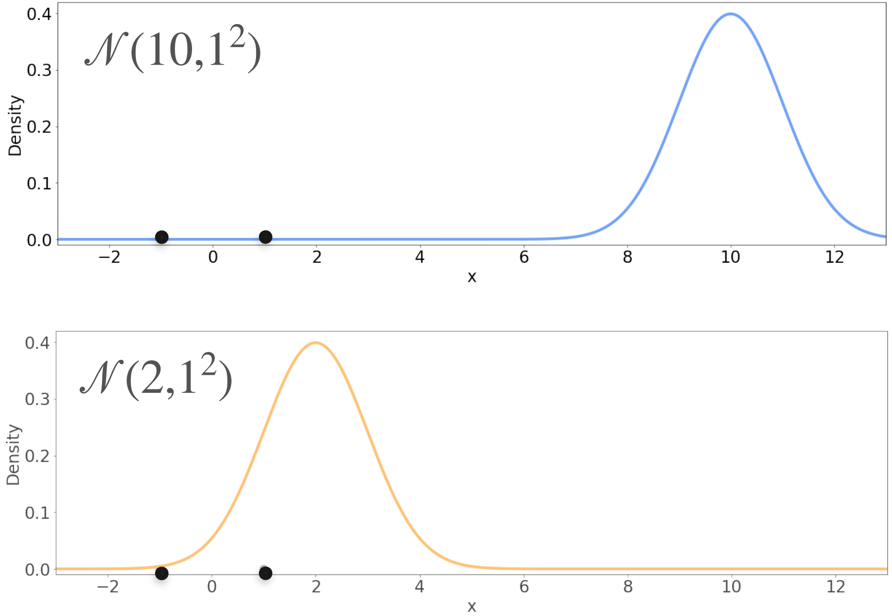
Point Estimation

MLE: Gaussian Example

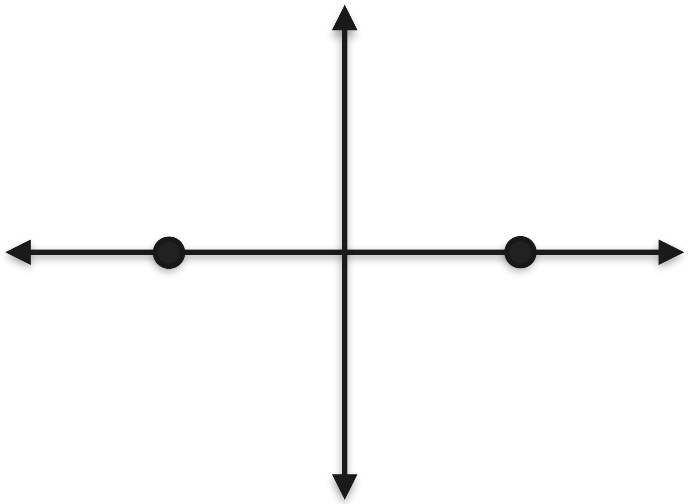
DeepLearning.AI

Maximum Likelihood: Gaussian Example

Candidates

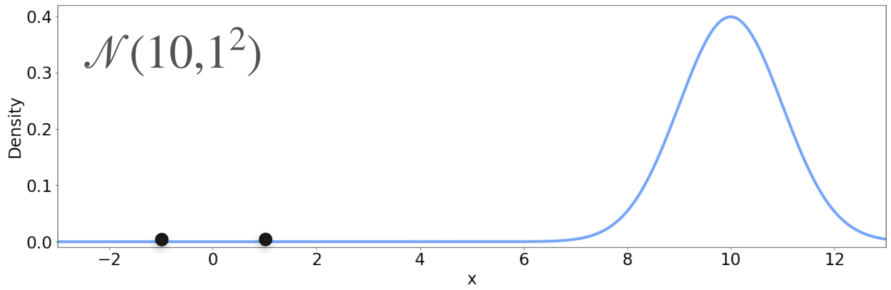


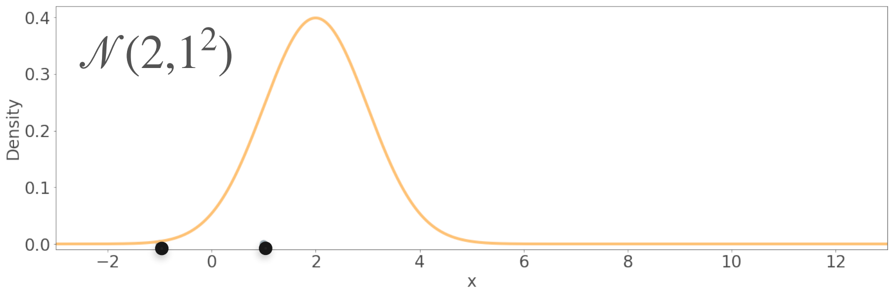
Observations



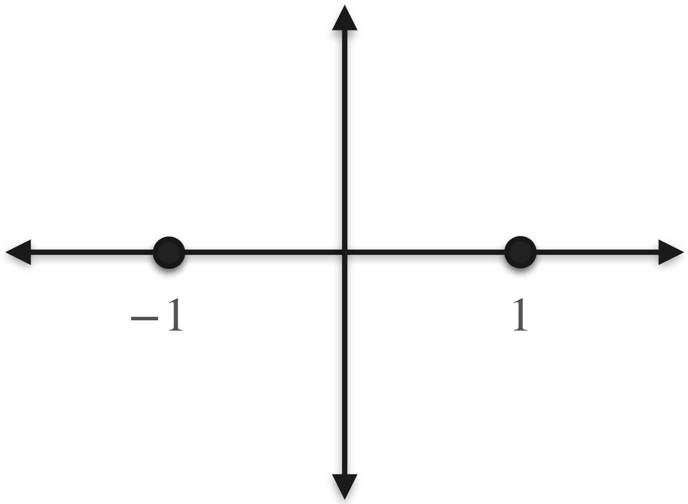
Maximum Likelihood: Gaussian Example

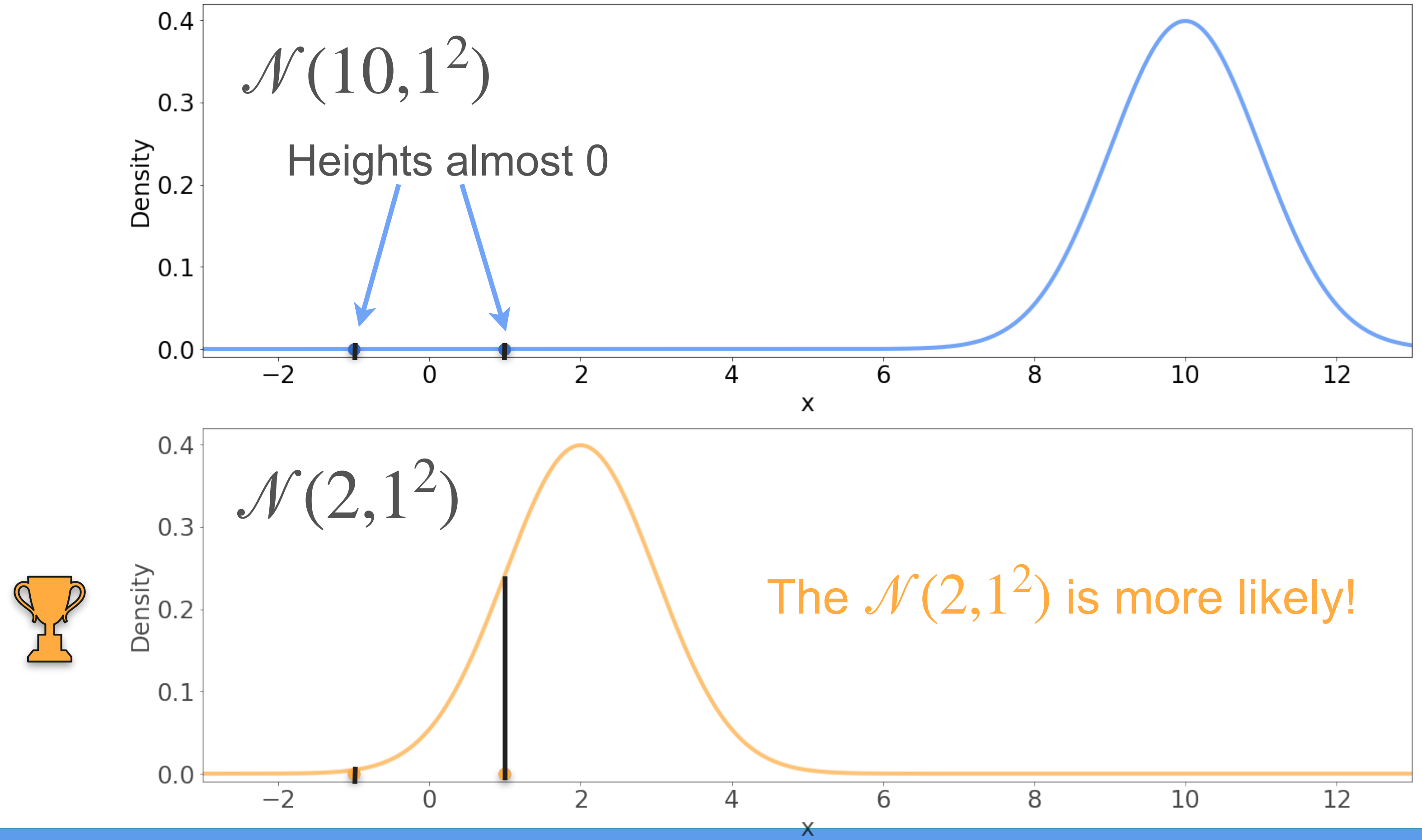
Candidates





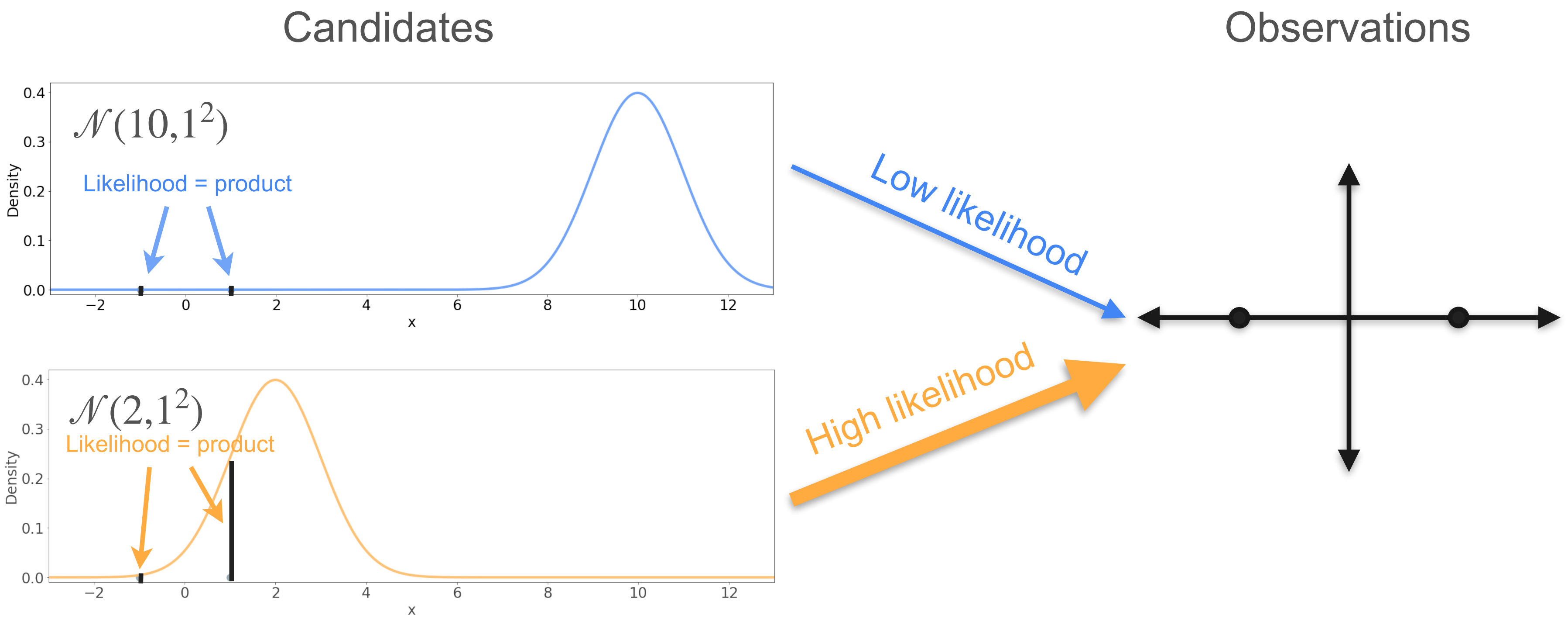
Observations





$\mathcal{N}(10,1^2)$
Heights almost 0
Density
x
$\mathcal{N}(2,1^2)$
The $\mathcal{N}(2,1^2)$ is more likely!
Density
x

Maximum Likelihood: Gaussian Example



Gaussians With Three Different Means

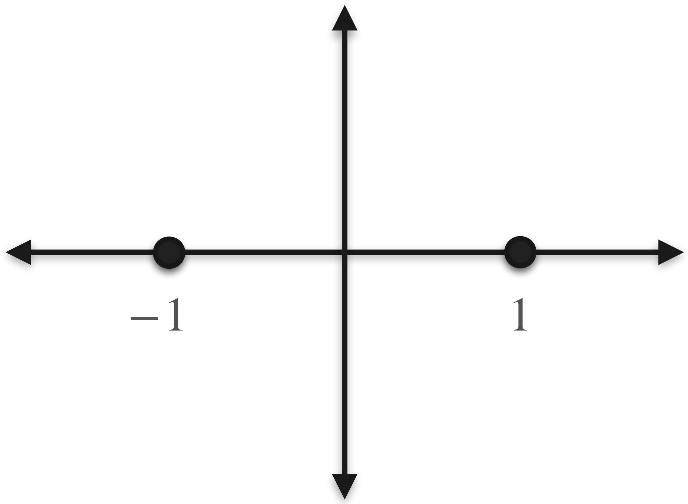
Candidates

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

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

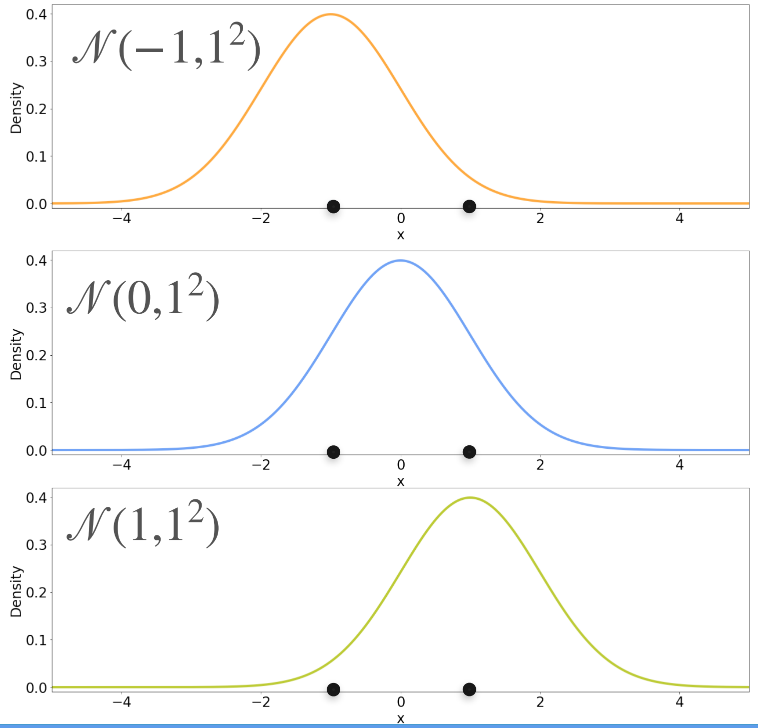
$\mathcal{N}(1,1^2)$

Observations

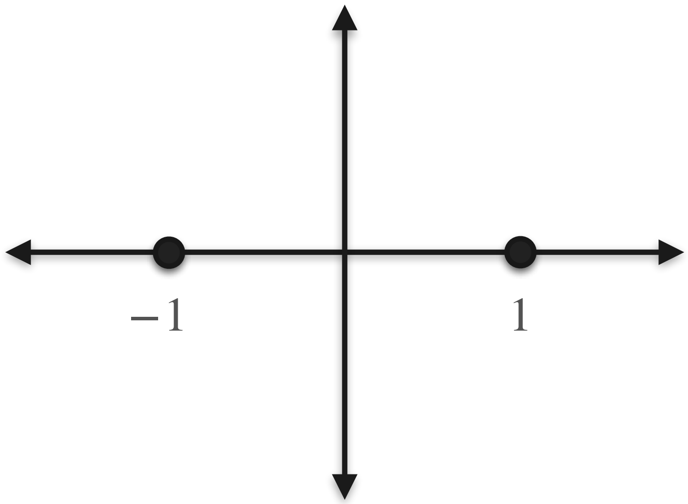


Gaussians With Three Different Means

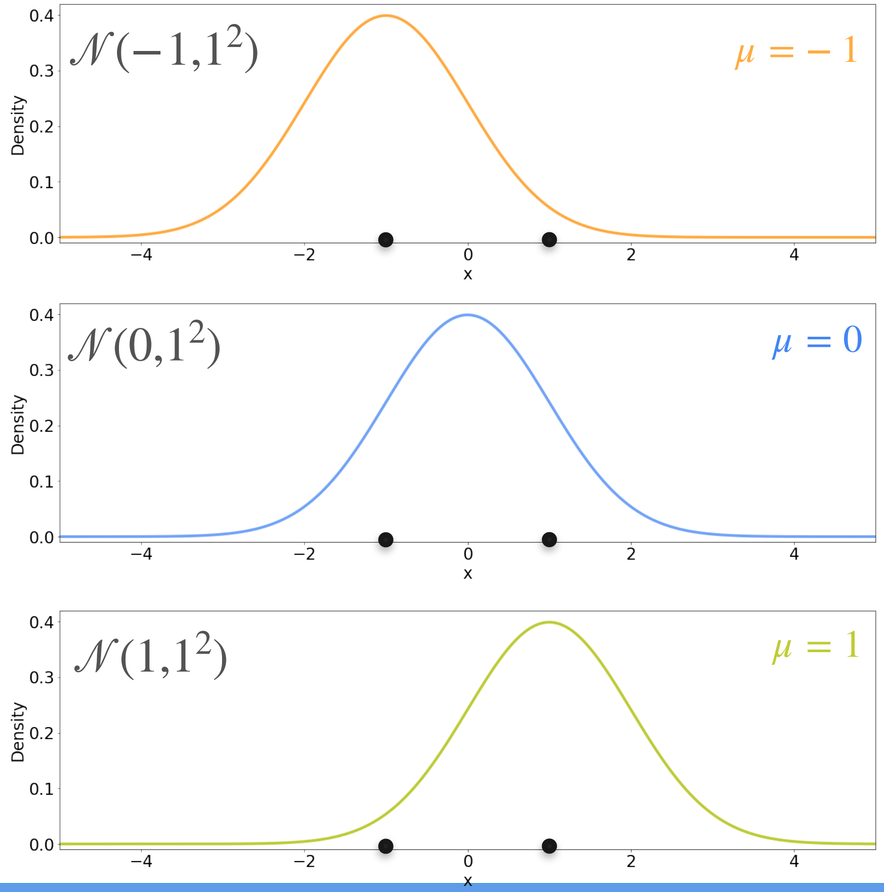
Candidates



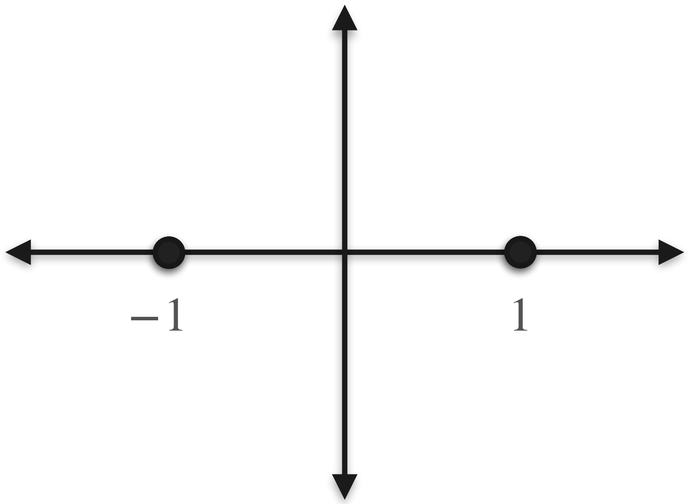
Observations



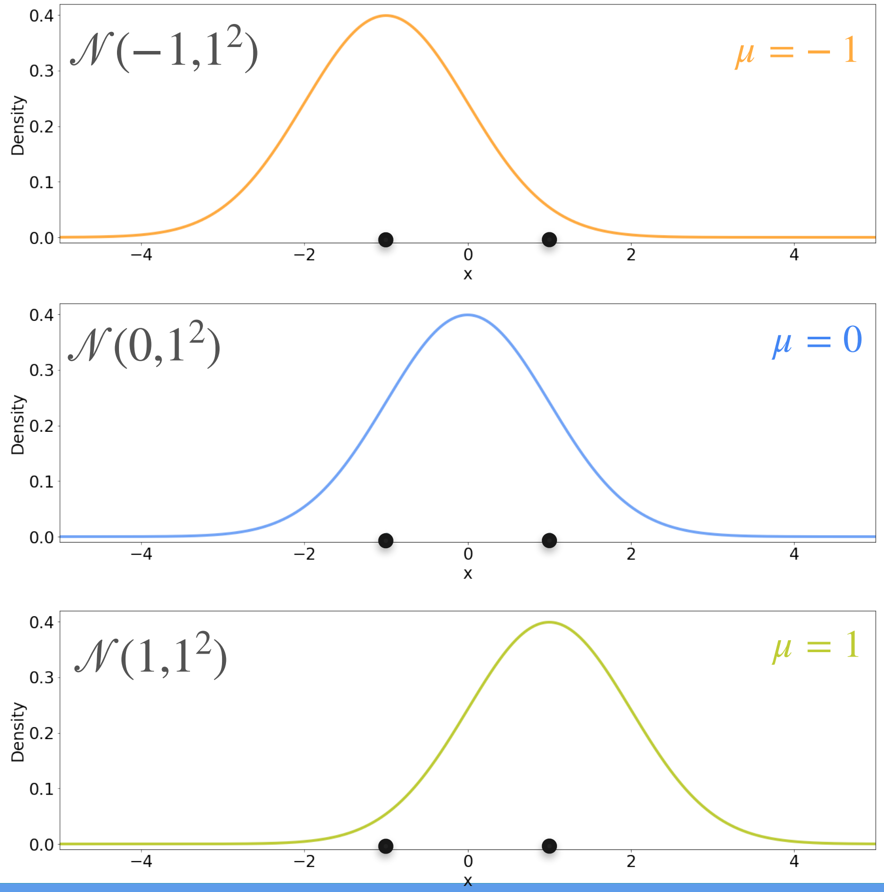
Candidates



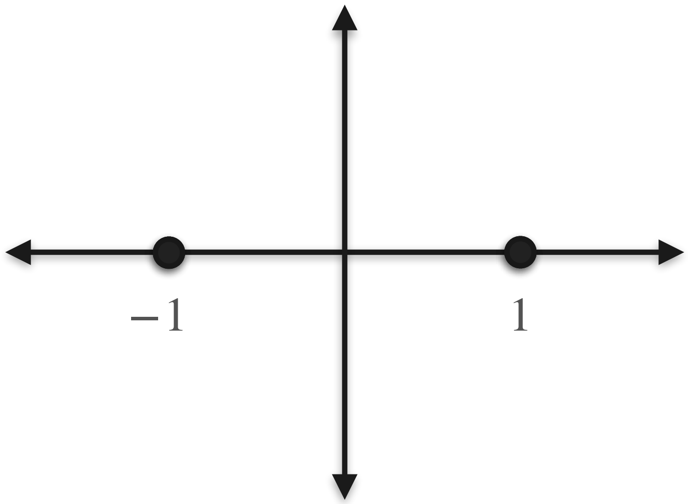
Observations



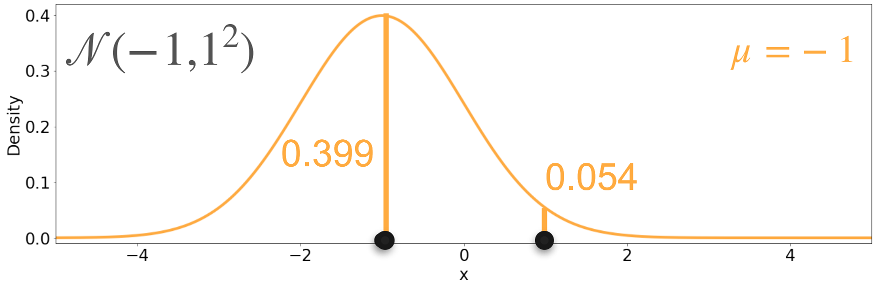
Candidates



Observations

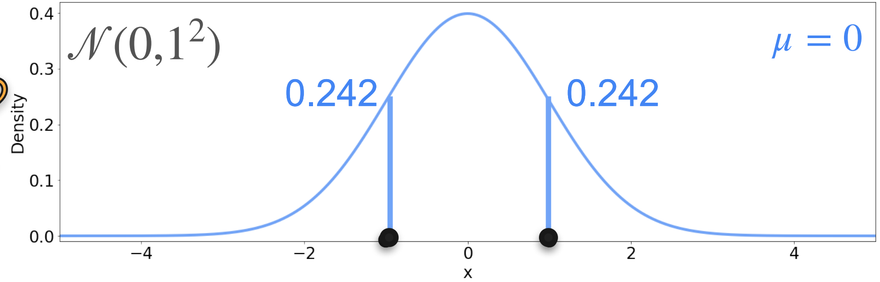


Candidates | Observations

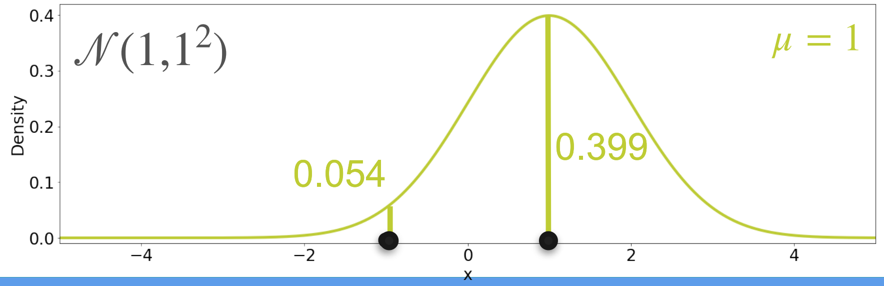


$\cdot$ = 0.022

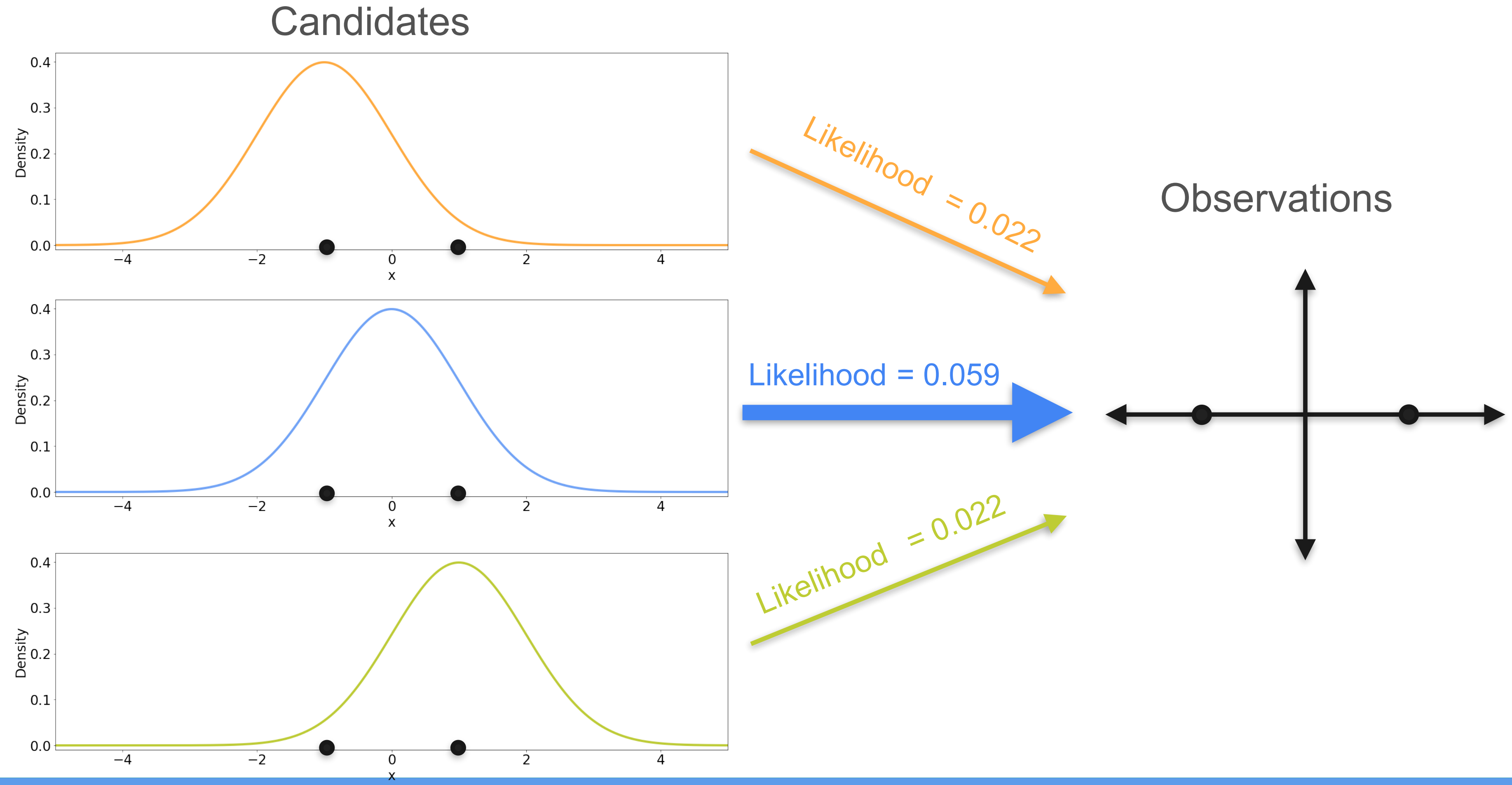
The $\mathcal{N}(0,1)$ is more likely!

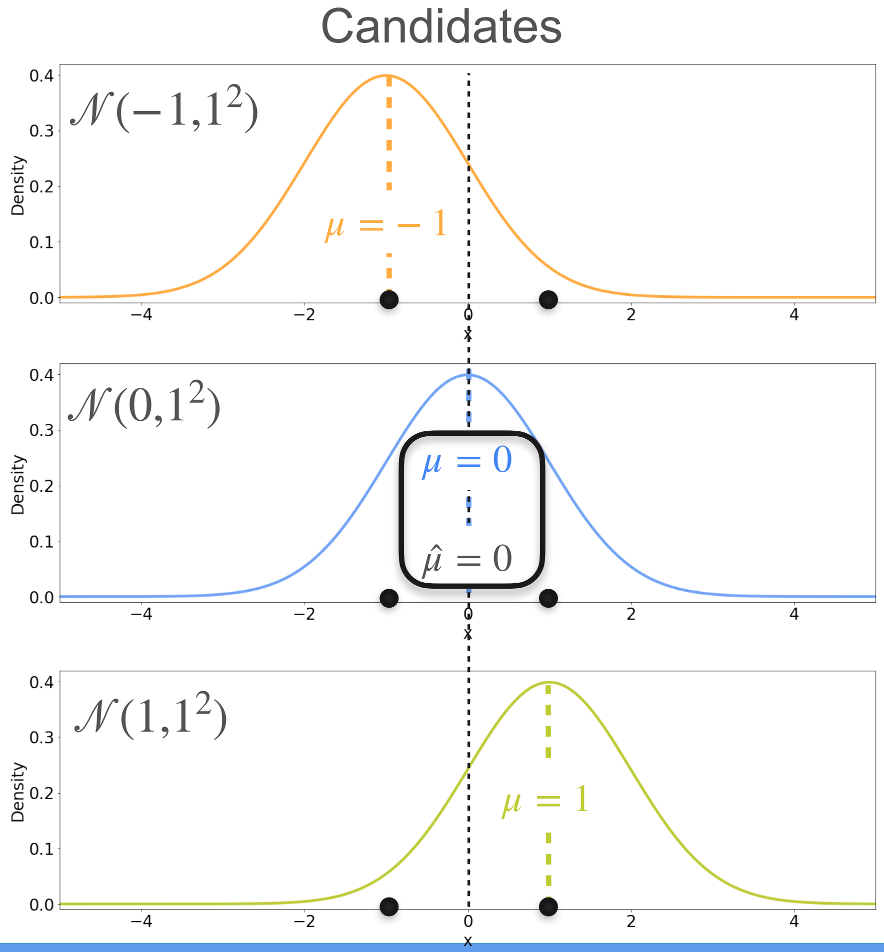


$\cdot$ = 0.059



$\cdot$ = 0.022


Candidates
Density
x
Likelihood = 0.022
Likelihood = 0.059
Likelihood = 0.022
Observations



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

Gaussians With Three Different Variance

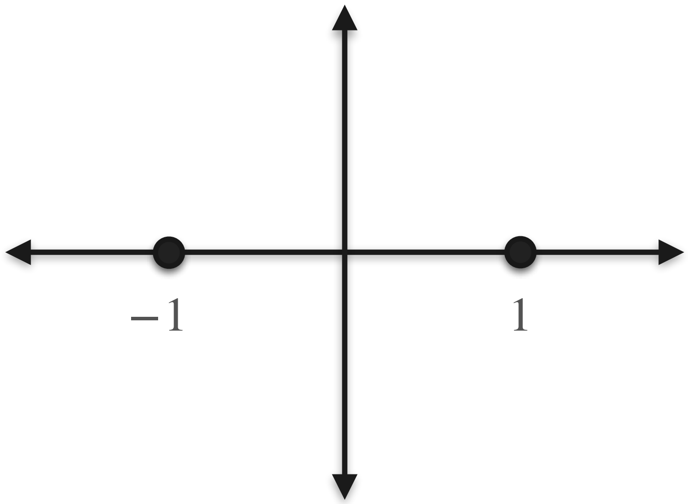
Candidates

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

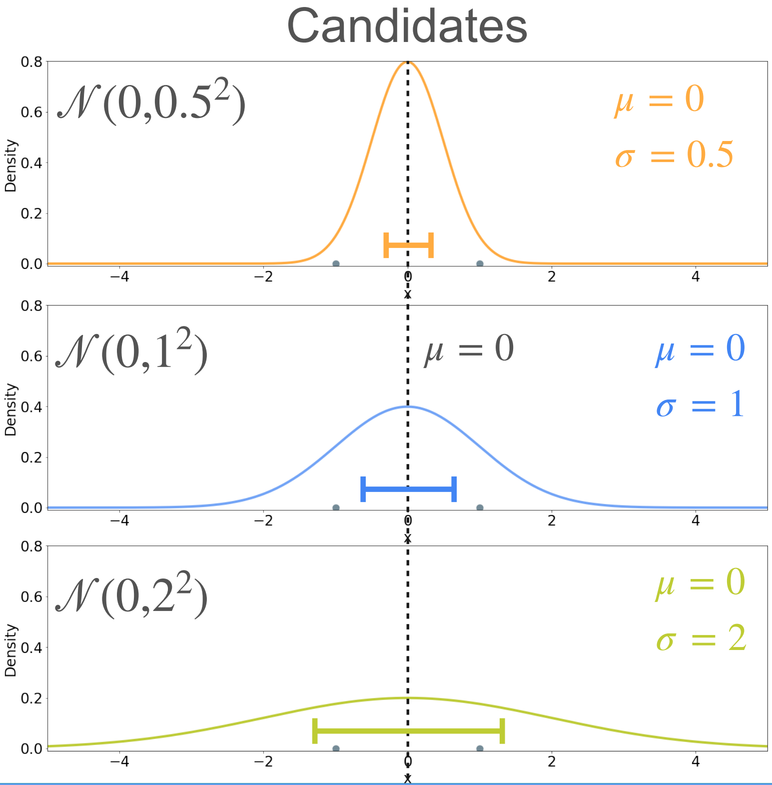
$\mathcal{N}(0, 1^2)$

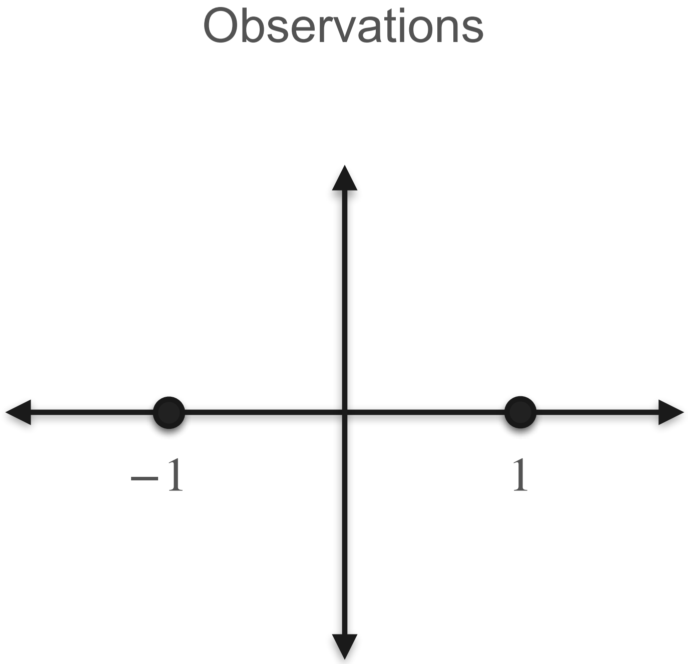
$\mathcal{N}(0, 2^2)$

Observations

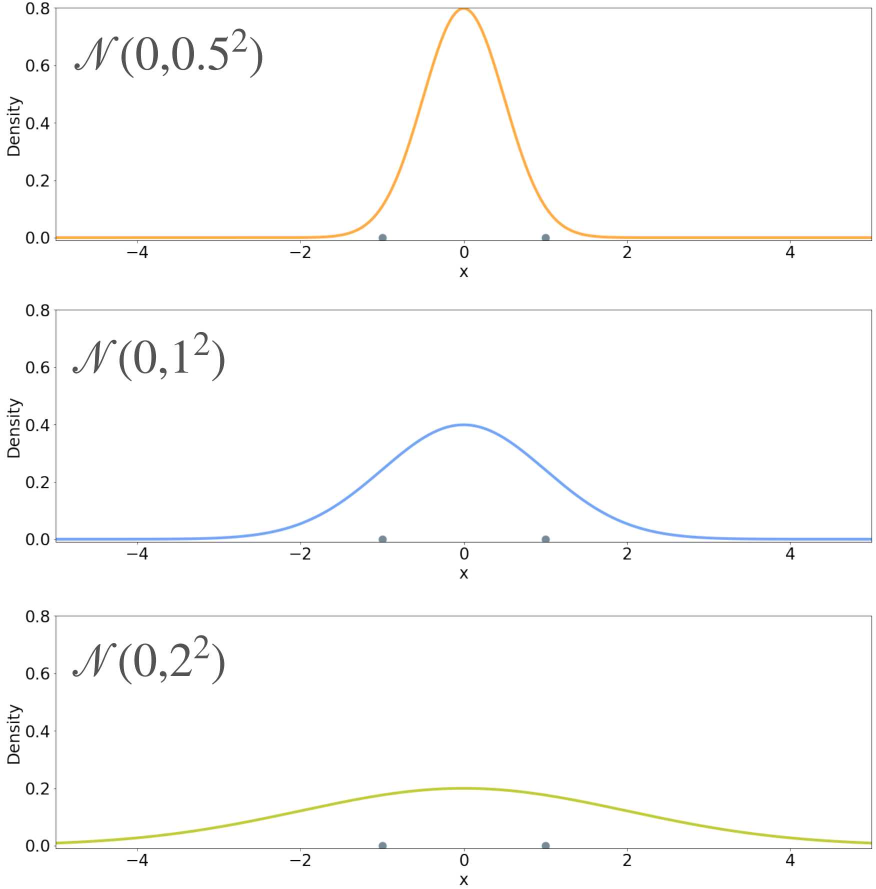


Gaussians With Three Different Variance

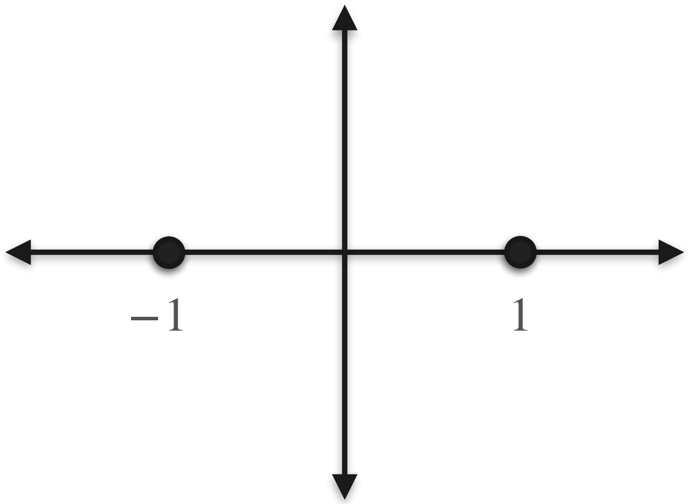




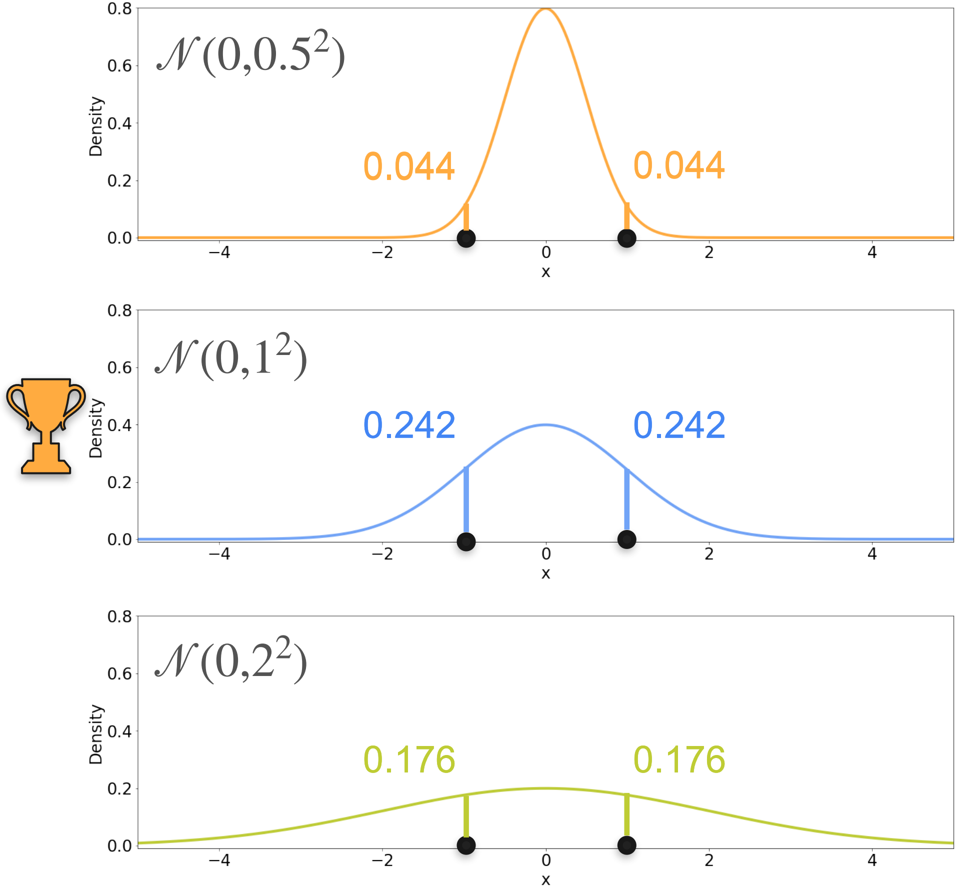
Candidates



Observations



Candidates



Observations

$\cdot \quad = 0.002$

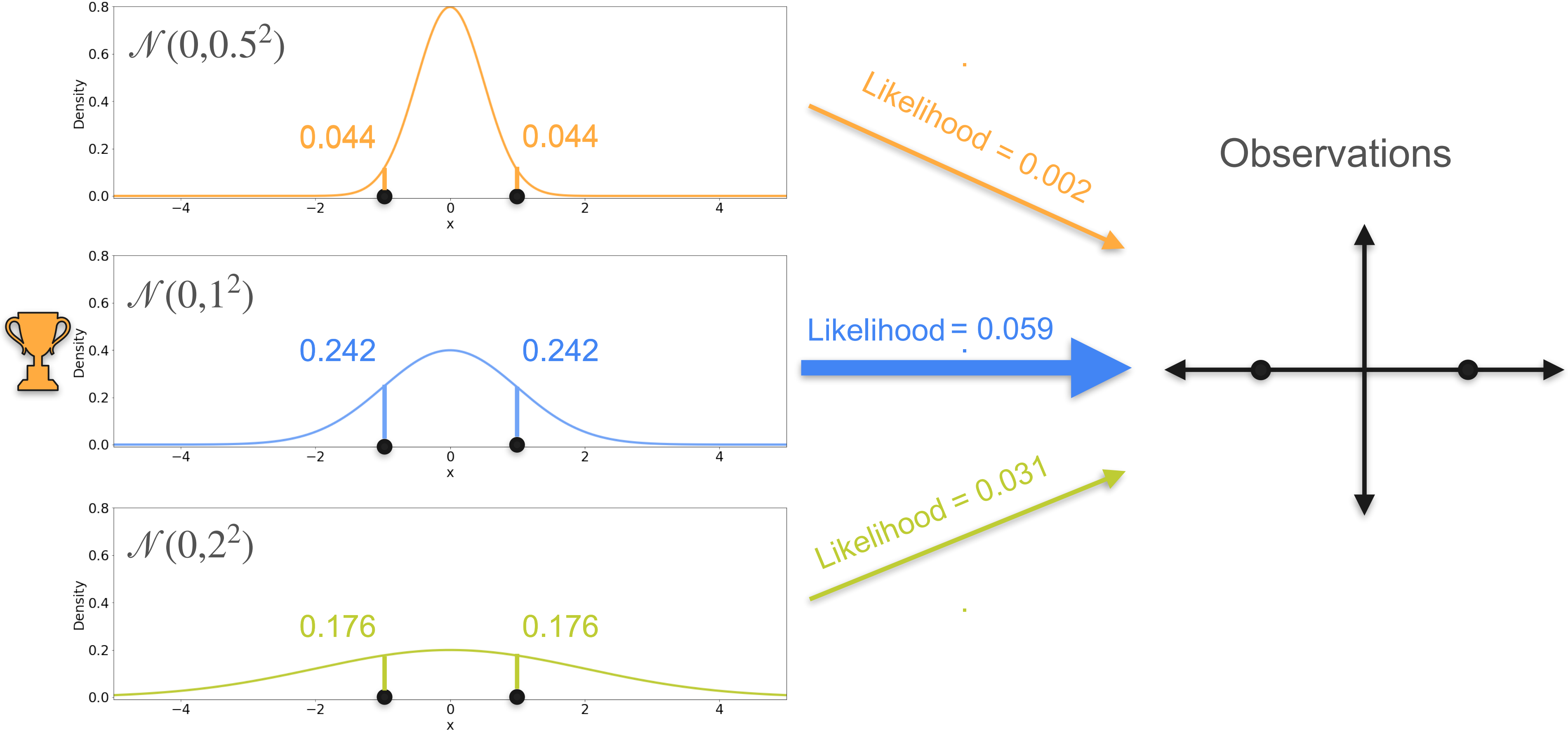
The $\mathcal{N}(0,1^2)$ is more likely!

$\cdot \quad = 0.059$

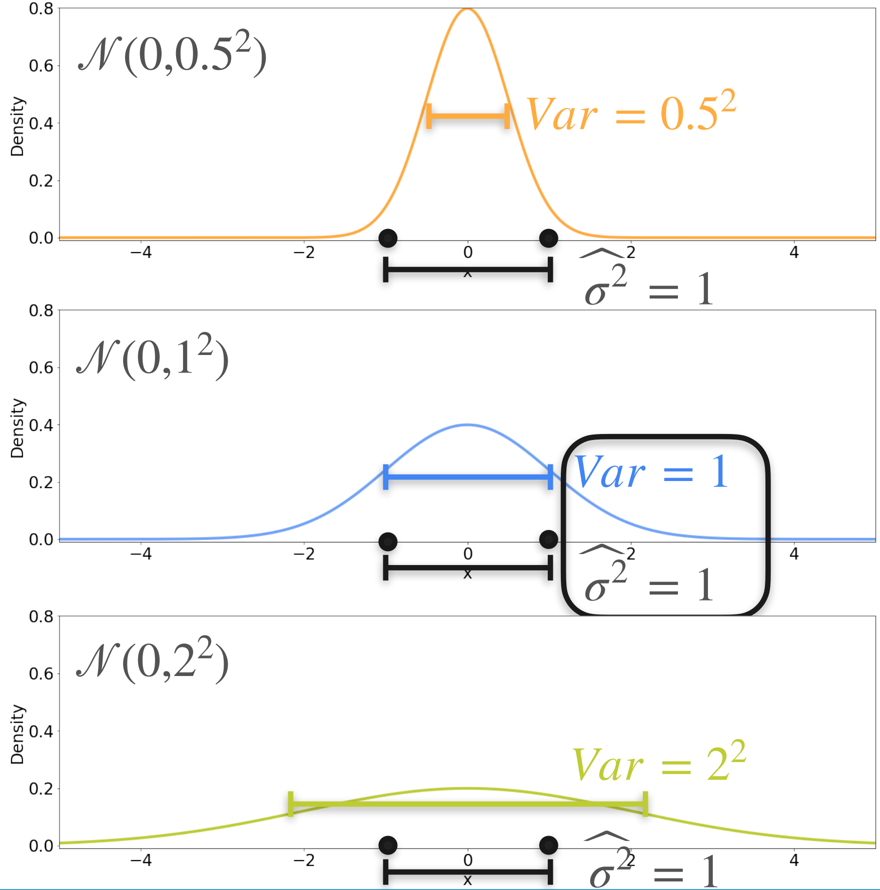
$\cdot \quad = 0.031$

Candidates

Observations



Candidates



Observations

Variance of the observations

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

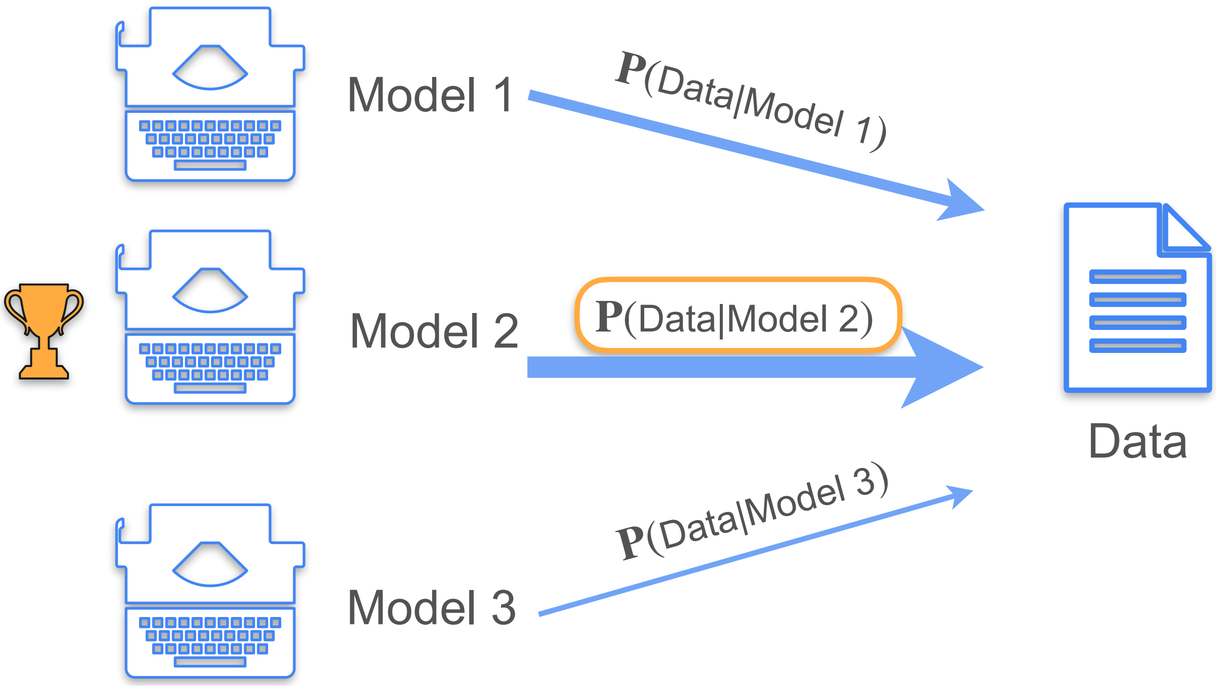
The best distribution is the one where the **variance** of the distribution is the **variance** of the sample

DeepLearning.AI

Point Estimation

MLE: Linear Regression

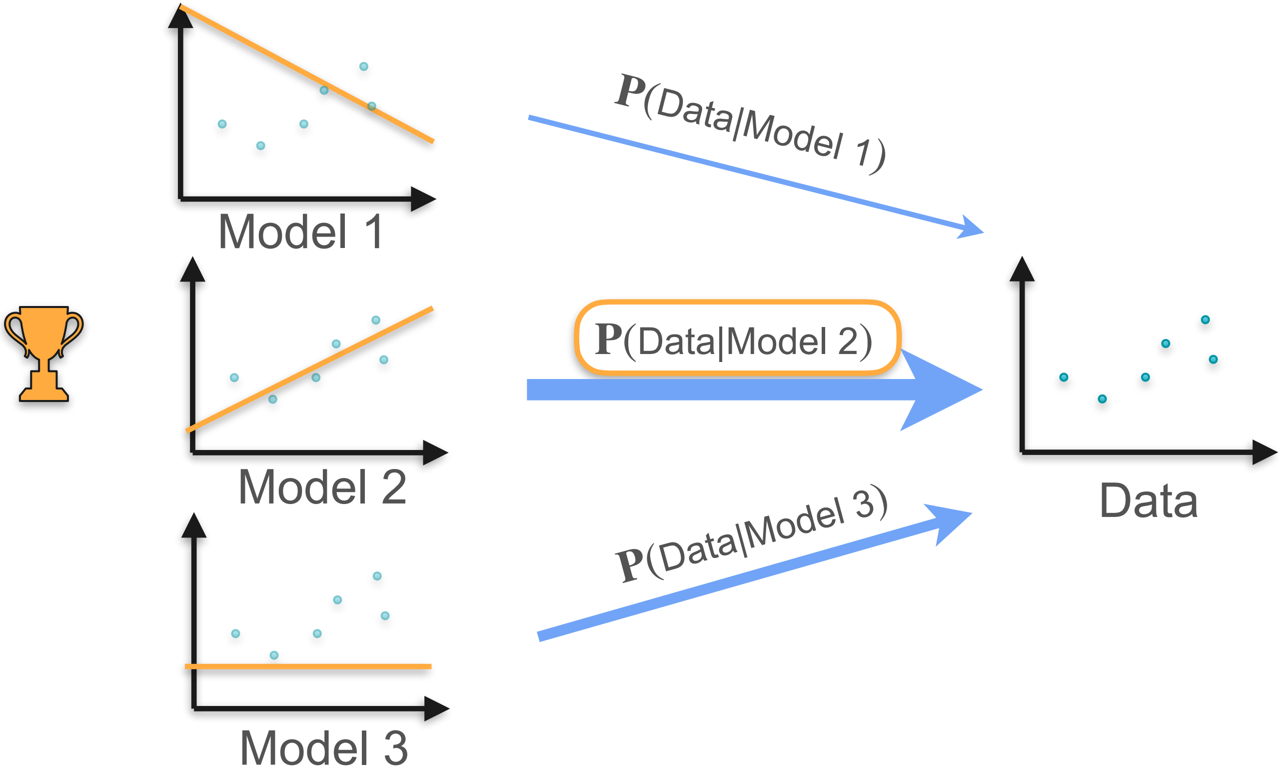
Maximum Likelihood



Find model that most likely produced the data

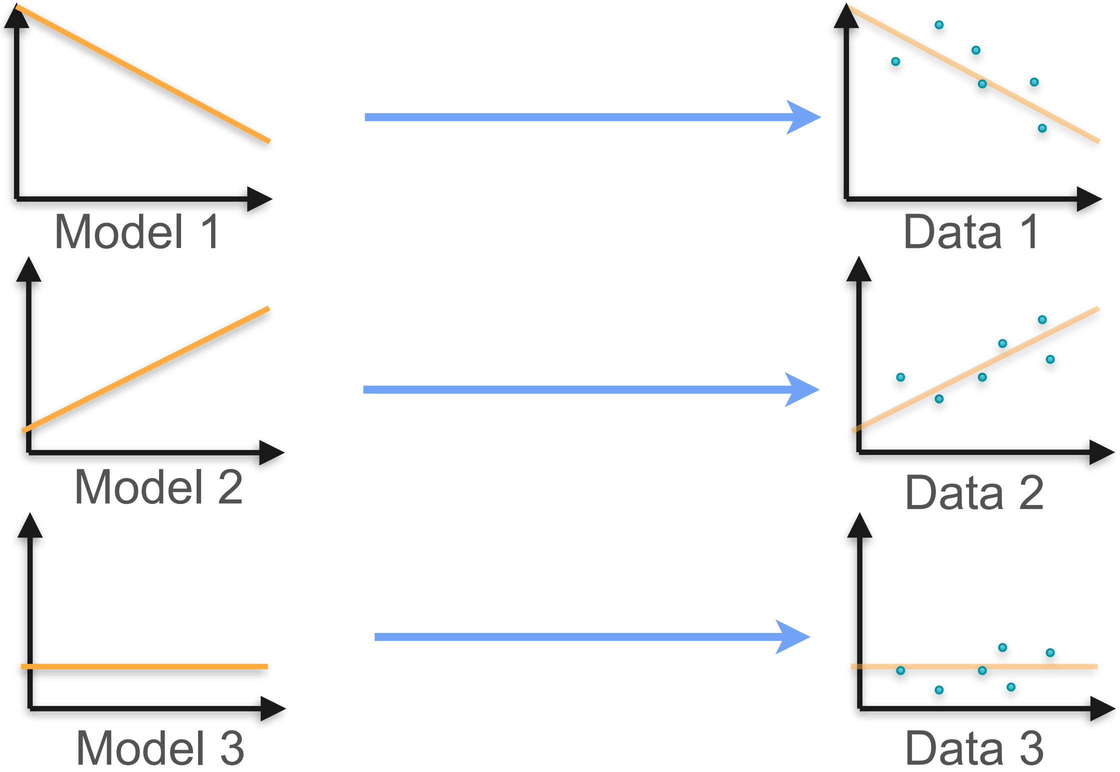
Maximize
$\mathbf{P}$(Data|Model)

Example: Linear Regression

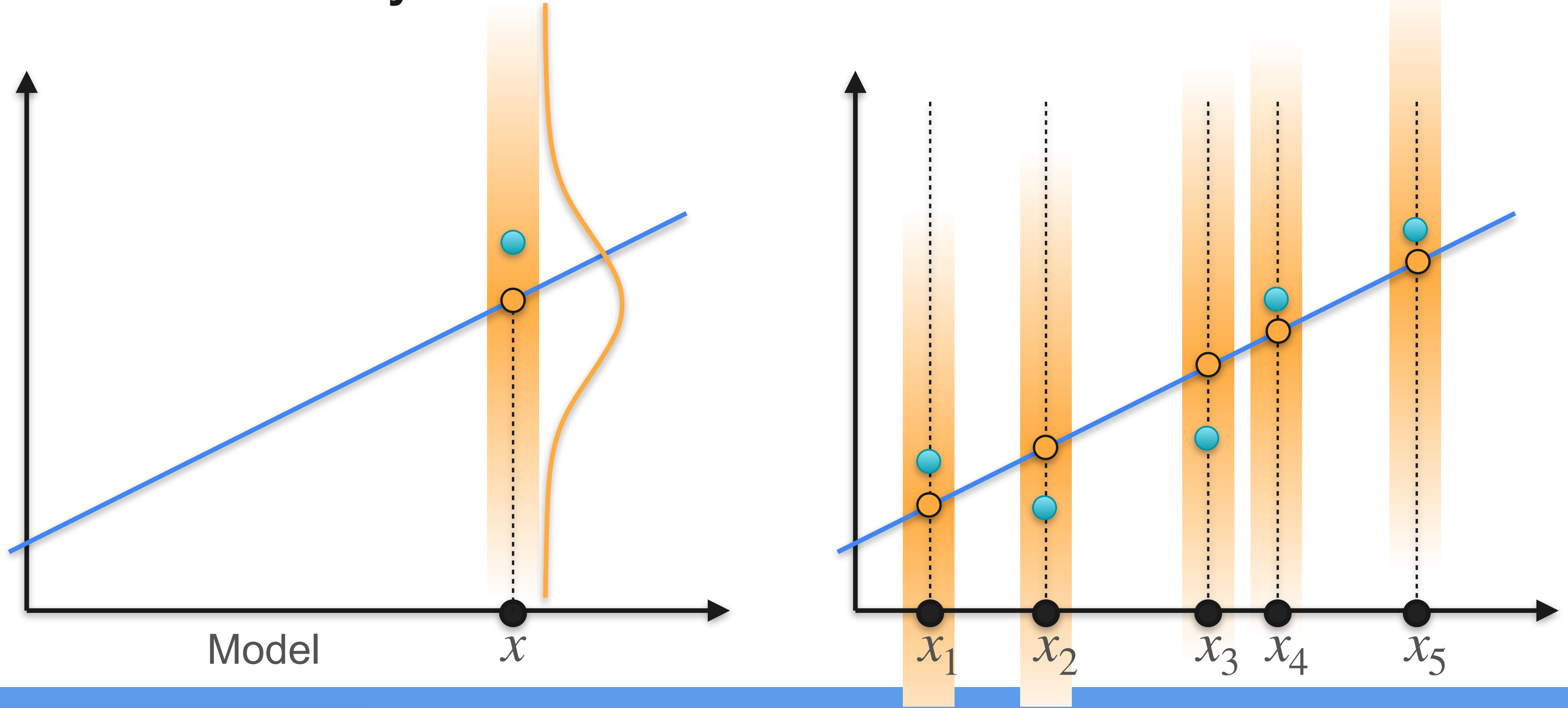


Find model that most likely produced the data

How Exactly Does a Line Produce Points?



How Exactly Does a Line Produce Points?

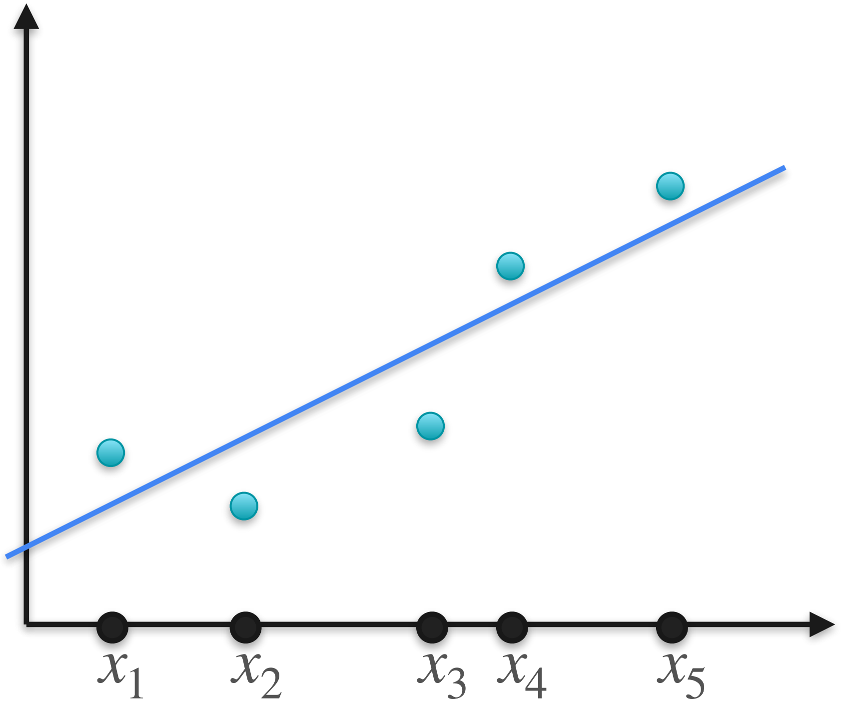


Linear Regression

Line that best produced the points

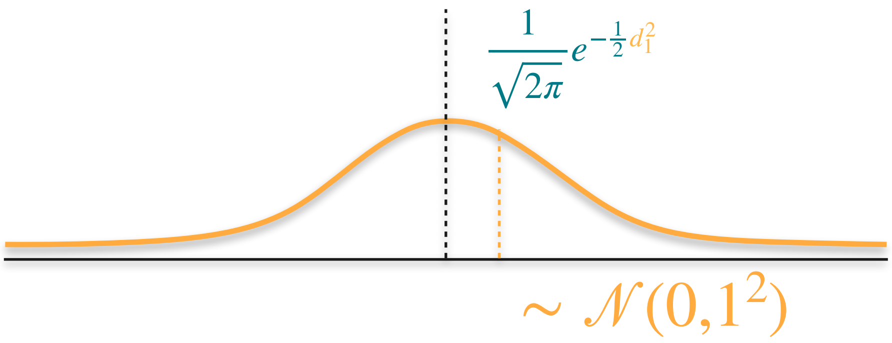
How?

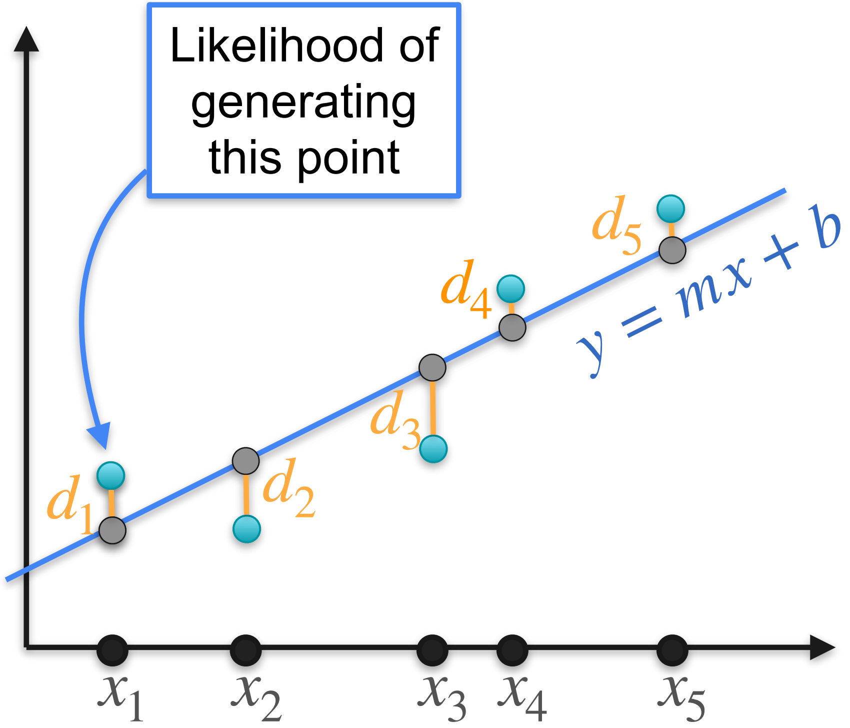
Line that best fits the data (linear regression)



Linear Regression and Likelihood

Likelihood:





Linear Regression and Likelihood

Likelihood:

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

Maximize

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

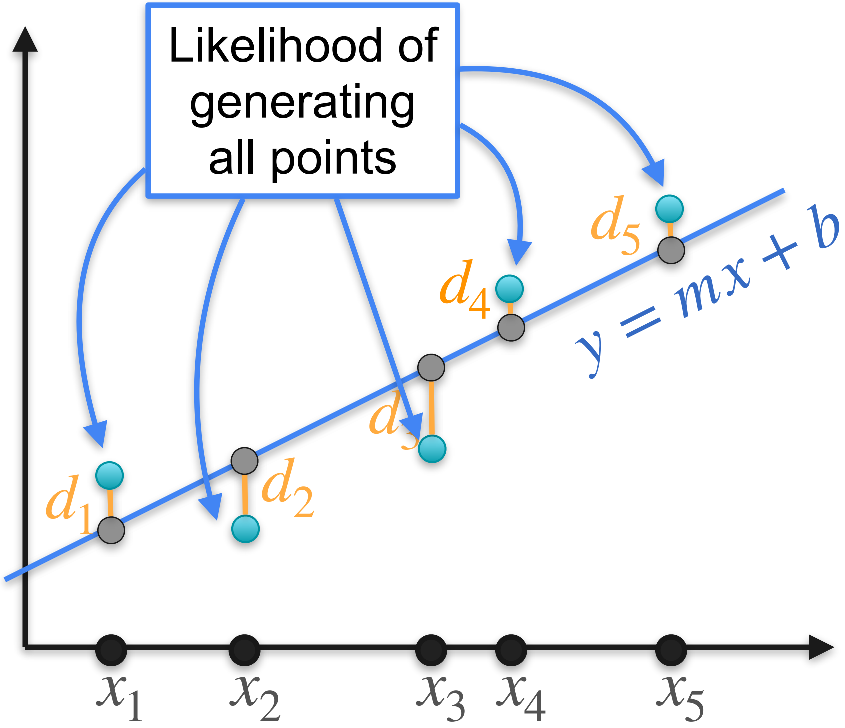
$$\cancel{e^{\boxed{-}\frac{1}{2}}}(d_1^2+d_2^2+d_3^2+d_4^2+d_5^2)$$

Minimize

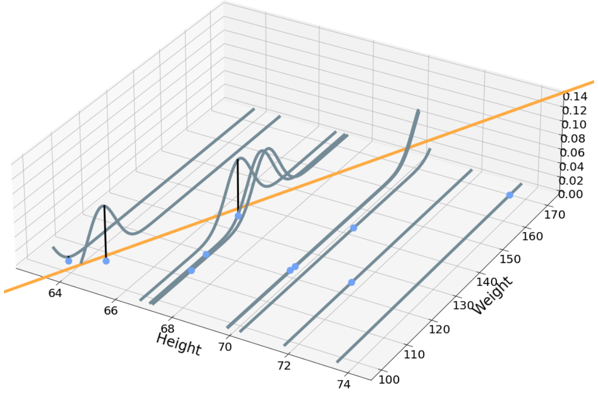
$$d_1^2+d_2^2+d_3^2+d_4^2+d_5^2$$

Least squares error!

Linear regression!

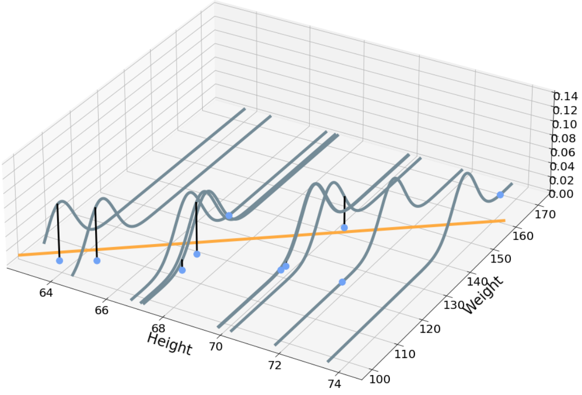


Picking the Right Model



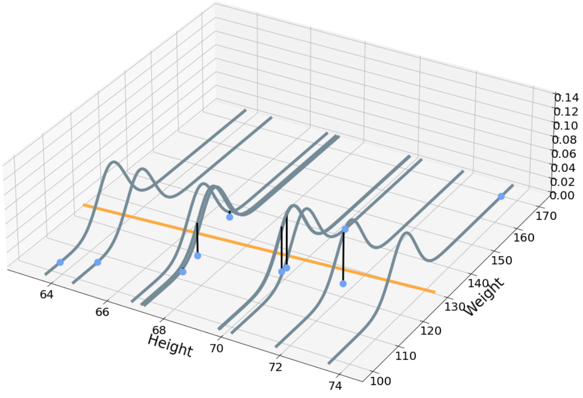
Model 1:

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



Model 2:

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



Model 3:

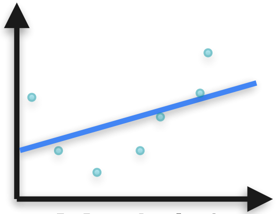
Likelihood $= 3.48 \cdot 10^{-49}$

DeepLearning.AI

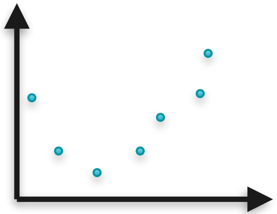
Point Estimation

Regularization

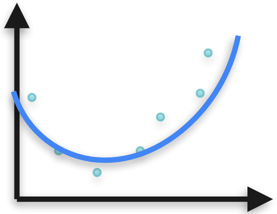
Example: Polynomial Regression



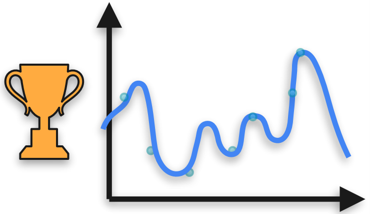
Model 1



Data



Model 2



Model 3

	Loss	Equation	Penalty	New loss
Model 1	$\boxed{10}$	$y = \boxed{4}x + 3$	$L_2 = 4^2 = \boxed{16}$	26
Model 2	$\boxed{2}$	$y = \boxed{2}x^2 \boxed{-4}x + 5$	$L_2 = 2^2 + (-4)^2 = \boxed{20}$	22
Model 3	$\boxed{0.1}$	$y = \boxed{4}x^{10} \boxed{-9}x^8 \boxed{-2}x^6 \boxed{+3}x^5 \boxed{-6}x^4 \boxed{-10}x + 4$	$L_2 = 4^2 + (-9)^2 + (-2)^2 + 3^2 + (-6)^2 + (-10)^2 = \boxed{246}$	246.1

Regularization Term

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

Log-loss: $\ell\ell$

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

Regularization parameter: λ

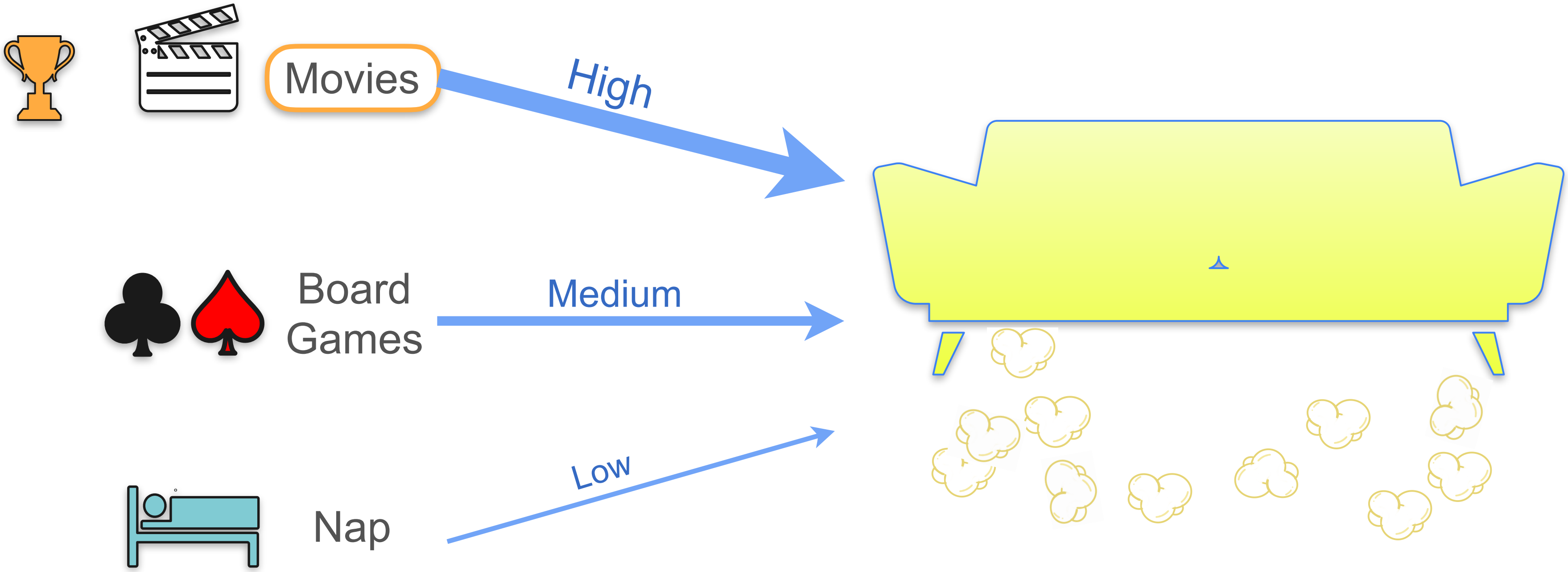
Regularized error: $+ \quad (\qquad)$

DeepLearning.AI

Point Estimation

Back to “Bayesics”

There's Popcorn on the Floor. What Happened?



There's Popcorn on the Floor. What Happened?

$\mathbf{P}(\text{Movies})$ Movies

High

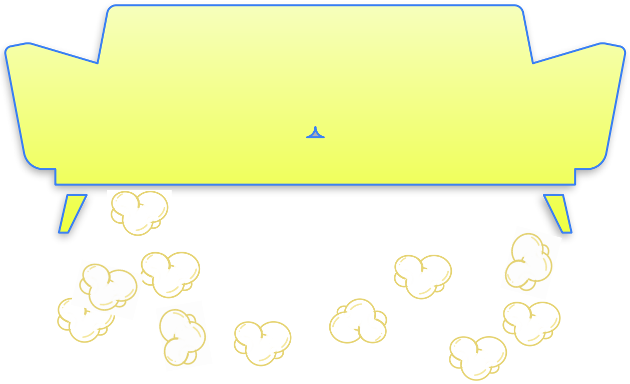
$\mathbf{P}(\text{Popcorn}|\text{Movies})$

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

Very high

$\mathbf{P}(\text{Popcorn}|\text{Contest})$

$\mathbf{P}(\text{Contest})$ Popcorn throwing contest



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

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

$\boxed{\mathbf{P}(A\,|\,B)\mathbf{P}(B) = \mathbf{P}(A \cap B)}$

DeepLearning.AI

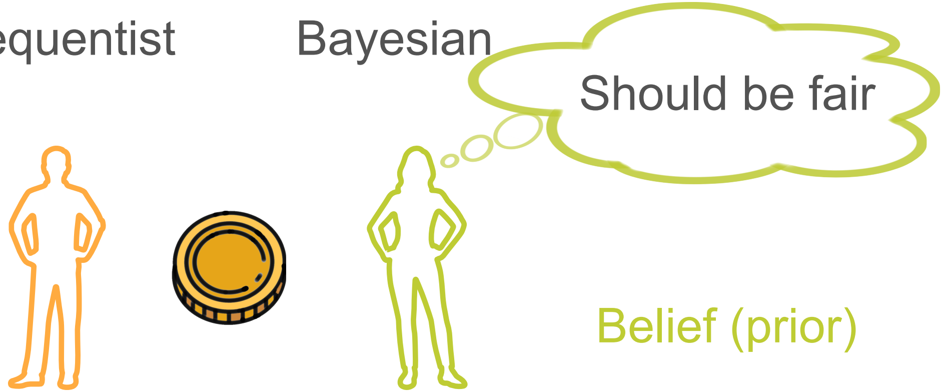
Point Estimation

Frequentist vs Bayesian Statistics

Frequentists vs. Bayesians

Frequentist

Bayesian



Belief (prior)

0.8

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

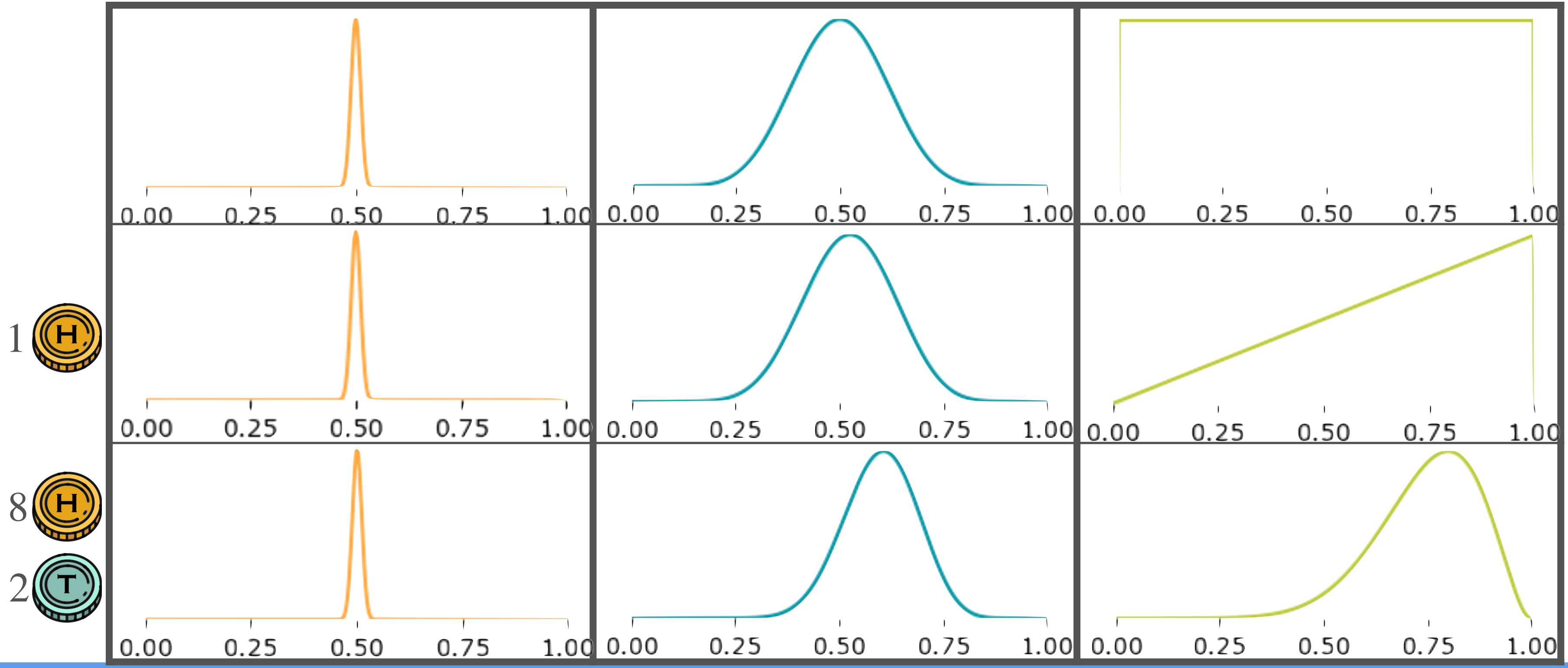
- Probabilities represent the degree of belief (or certainty)
- Concept of Prior
- Goal: update prior belief based on observations

DeepLearning.AI

Point Estimation

Bayesian Statistics
MAP

Updating Your Beliefs



DeepLearning.AI

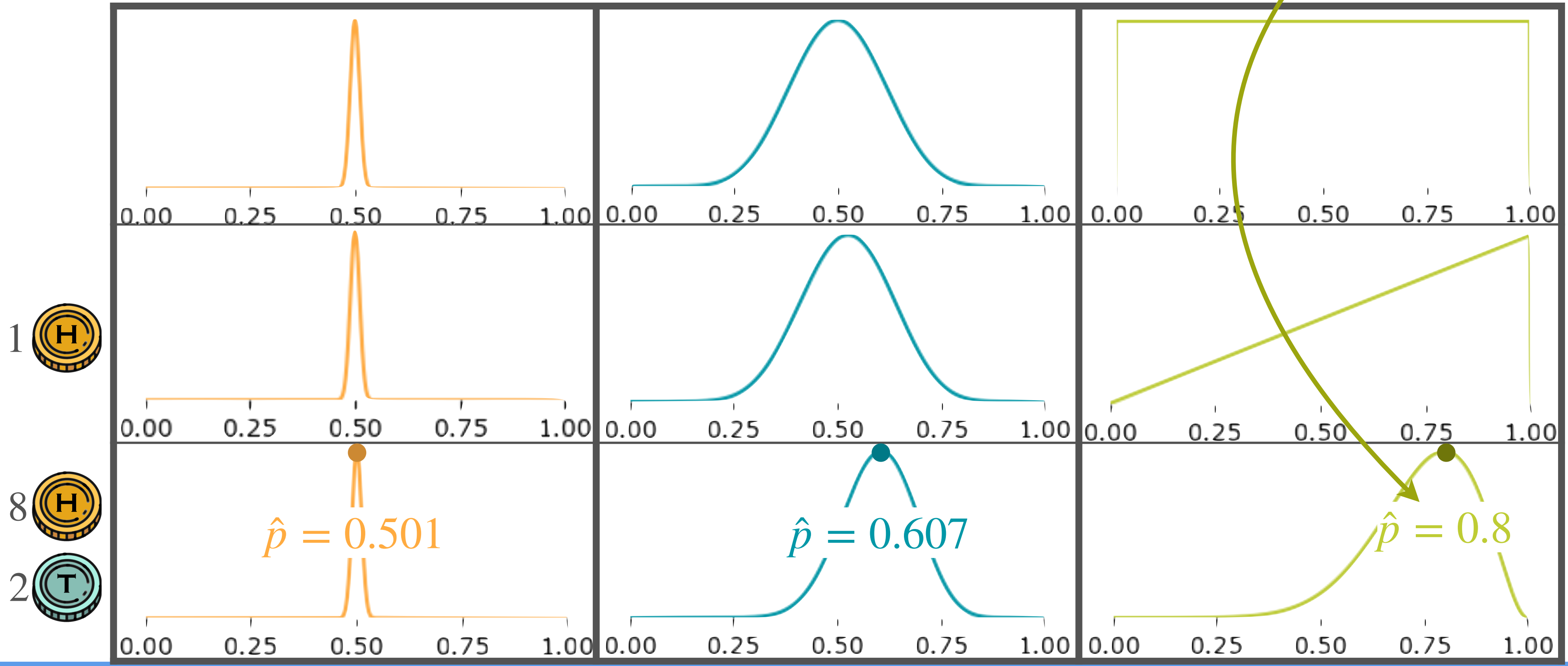
Maximum A Posteriori (MAP)

1 value for the parameter? → Choose the one with highest probability → Mode of the $\boxed{\text{updated belief}}$

Posterior

Maximum A Posteriori (MAP)

Same result as frequentist!



DeepLearning.AI

Point Estimation

Bayesian Statistics
Updating Priors

Bayesian Statistics

Likelihood of evidence B appearing, given A happened

"Posterior"
Belief that A will happen after considering evidence B

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

"Prior"
Belief that A will happen, before considering evidence B

Probability of evidence B in any circumstances

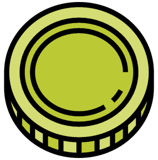
$$P(B\,|\,A)P(A) + P(B\,|\,A')P(A')$$

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)

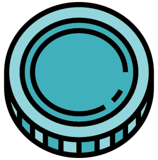
Bayesian Statistics

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A)\mathbf{P}(A)}{\mathbf{P}(B)}$$



“Fair”

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



“Biased”

$\mathbf{P}(H) = 0.8$

“Mystery”

Either Fair or Biased

Event to Predict

$A \rightarrow Y$ takes some value

Y: odds of H for your coin

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

Evidence

$B \rightarrow X$ take some value

X: result of coin flip

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

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

Priors

$x = 1$

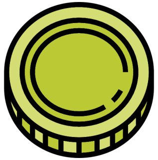
$$p_{Y|X=1}(0.5) = \frac{p_{X|Y=0.5}(1) \quad p_Y(0.5)}{p_X(1)} = \frac{0.5 \; 0.75}{0.5 \; 0.75 + 0.8 \; 0.25}$$

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

Posterior

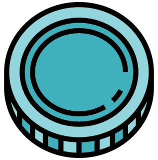
Bayesian Statistics

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A)\mathbf{P}(A)}{\mathbf{P}(B)}$$



"Fair"

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



"Biased"

$\mathbf{P}(H) = 0.8$

"Mystery"

Either Fair or Biased

Event to Predict

$A \rightarrow Y$ takes some value

Y: odds of H for your coin

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

$\mathbf{P}(Y = 0.5) = 0.75$

$\mathbf{P}(Y = 0.80) = 0.25$

Evidence

$B \rightarrow X$ take some value

X: result of coin flip

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

$x = 1$

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

Bayesian Statistics

	Y is discrete	Y is continuous
X is discrete	$\underbrace{\boxed{p_{Y\mid X=x}(y)}}_{\text{Posterior}} = \dfrac{p_{X\mid Y=y}(x)\overbrace{\boxed{p_Y(y)}}^{\text{Prior}}}{p_X(x)}$	$\underbrace{\boxed{f_{Y\mid X=x}(y)}}_{\text{Posterior}} = \dfrac{p_{X\mid Y=y}(x)\overbrace{\boxed{f_Y(y)}}^{\text{Prior}}}{p_X(x)}$
X is continuous	$\underbrace{\boxed{p_{Y\mid X=x}(y)}}_{\text{Posterior}} = \dfrac{f_{X\mid Y=y}(x)\overbrace{\boxed{p_Y(y)}}^{\text{Prior}}}{f_X(x)}$	$\underbrace{\boxed{f_{Y\mid X=x}(y)}}_{\text{Posterior}} = \dfrac{f_{X\mid Y=y}(x)\overbrace{\boxed{f_Y(y)}}^{\text{Prior}}}{f_X(x)}$

Bayesian Statistics

	Θ is discrete	Θ is continuous
X is discrete	$p_{\Theta\|X=x}(\theta) = \dfrac{p_{X\|\Theta=\theta}(x)p_{\Theta}(\theta)}{p_X(x)}$	$f_{\Theta\|X=x}(\theta) = \dfrac{p_{X\|\Theta=\theta}(x)f_{\Theta}(\theta)}{p_X(x)}$
X is continuous	$p_{\Theta\|X=x}(\theta) = \dfrac{f_{X\|\Theta=\theta}(x)p_{\Theta}(\theta)}{f_X(x)}$	$f_{\Theta\|X=x}(\theta) = \dfrac{f_{X\|\Theta=\theta}(x)f_{\Theta}(\theta)}{f_X(x)}$

DeepLearning.AI

Point Estimation

Bayesian Statistics Full Worked Example

Bayesian Statistics: Bernoulli Example

$\theta \quad \theta \quad \theta \quad \theta \quad \theta \quad \theta \quad \theta \quad \theta \quad 1-\theta \quad 1-\theta$

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

$\mathbf{X} = \left(X_1, X_2, \ldots, X_{10}\right)$

$X_i = 1$ if H, 0 if T

$X_i \,|\, \Theta = \theta \sim Bernoulli(\theta)$

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

Θ is a continuous random variable

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

Use this version of Bayes' Theorem

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

Bayesian Statistics: Bernoulli Example

$\theta \quad \theta \quad \theta \quad \theta \quad \theta \quad \theta \quad \theta \quad \theta \quad 1-\theta \quad 1-\theta$

$X_i \mid \Theta = \theta \sim Bernoulli(\theta)$

$\Theta \sim Uniform(0,1)$

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

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

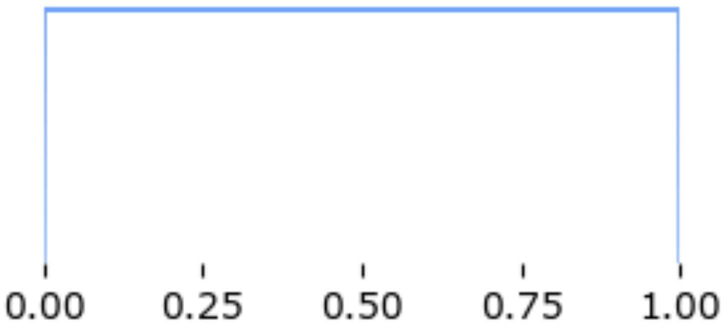
$f_\Theta(\theta) = 1, \quad 0 \le \theta \le 1$

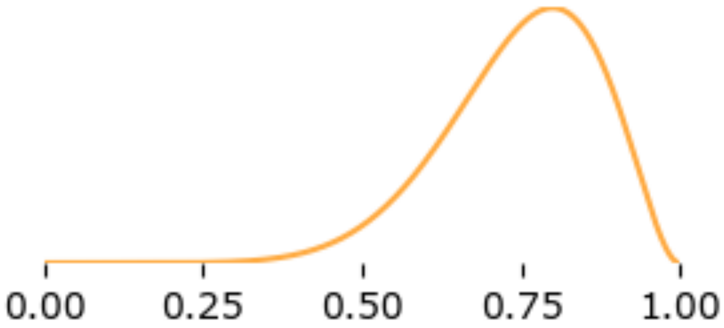
$$f_{\Theta|X=x}(\theta) = \frac{}{\text{constant}} = \frac{1}{\text{constant}}\theta^8(1-\theta)^2$$

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

Bayesian Statistics: Bernoulli Example

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





Prior Beliefs

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

Posterior Beliefs

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

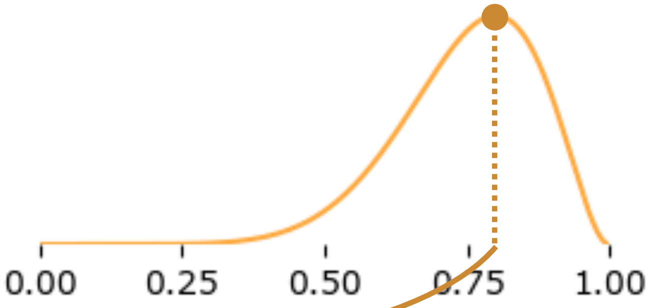
Bayesian Statistics: MAP Estimator

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

Likelihood of the model

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

$\mathbf{P}$(data | model)
Maximizing this is all MLE does!

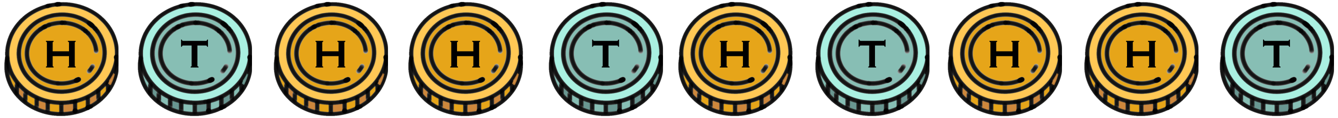


Maximum a Posteriori (MAP)

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

Non informative prior: MAP = MLE

Bayesian Statistics: Bernoulli Example Continued



$\theta \quad 1-\theta \quad \theta \quad \theta \quad 1-\theta \quad \theta \quad 1-\theta \quad \theta \quad \theta \quad 1-\theta$

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

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

New prior!

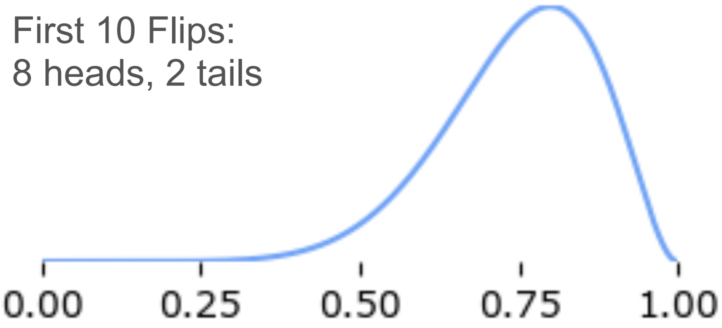
Repeat process

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) = \frac{\qquad}{\text{constant}} = \frac{1}{\text{constant}}\theta^{14}(1-\theta)^6$$

Bayesian Statistics: Bernoulli Example Continued

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

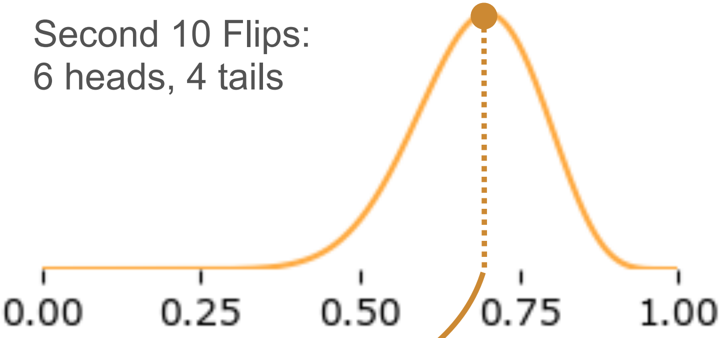
First 10 Flips:
8 heads, 2 tails



Prior Beliefs

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

Second 10 Flips:
6 heads, 4 tails



Posterior Beliefs

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) = \frac{1}{\text{constant}}\theta^{14}(1-\theta)^6$$

MAP

$\hat{\theta} = 0.7$

14 out of 20 heads

Same result as Frequentist

Bayesian Statistics: Final Summary

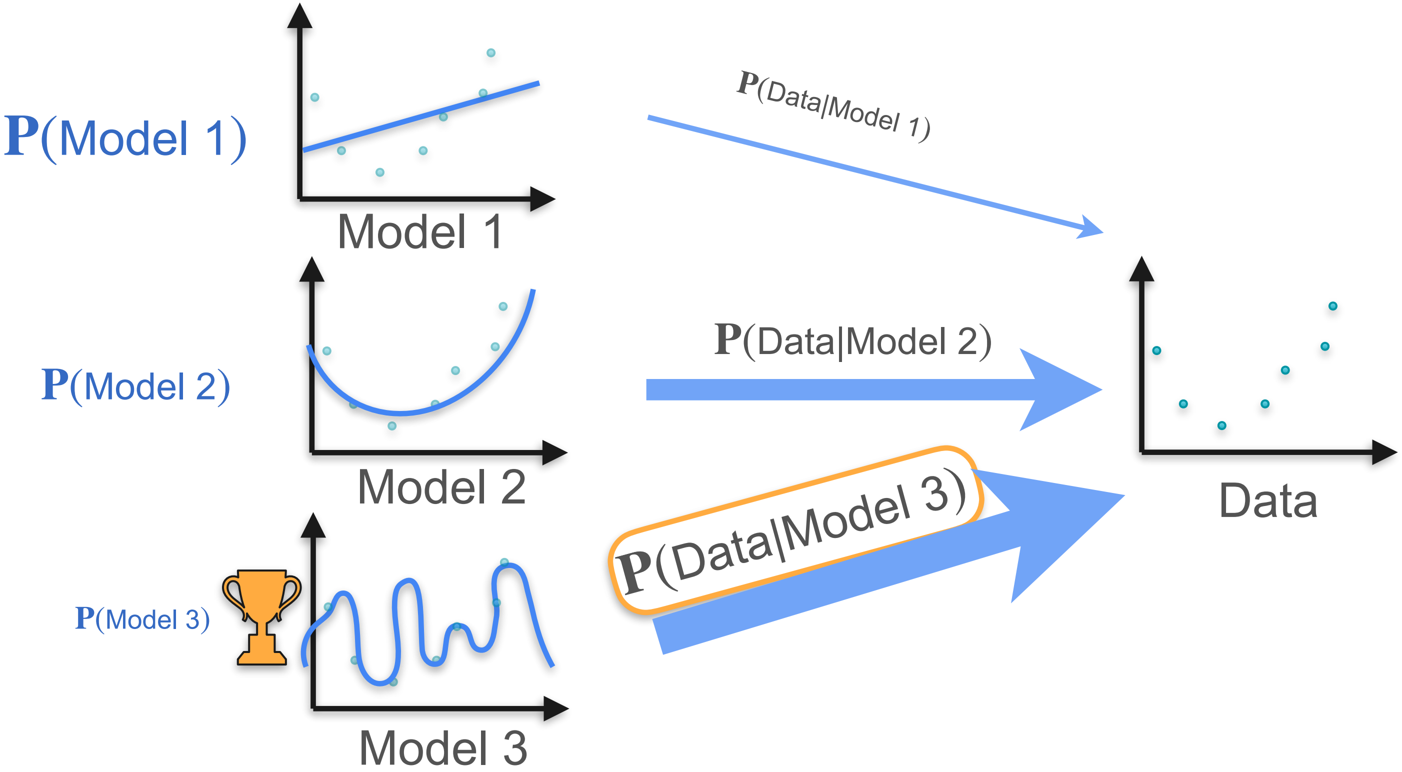
- Bayesians update prior beliefs
- MAP with uninformative priors is just MLE
- With enough data, MLE and MAP estimates usually converge
- Good for instances when you have limited data or strong prior beliefs
- Wrong priors, wrong conclusions

DeepLearning.AI

Point Estimation

Relationship between MAP, MLE, and Regularlization

Example: Polynomial Regression

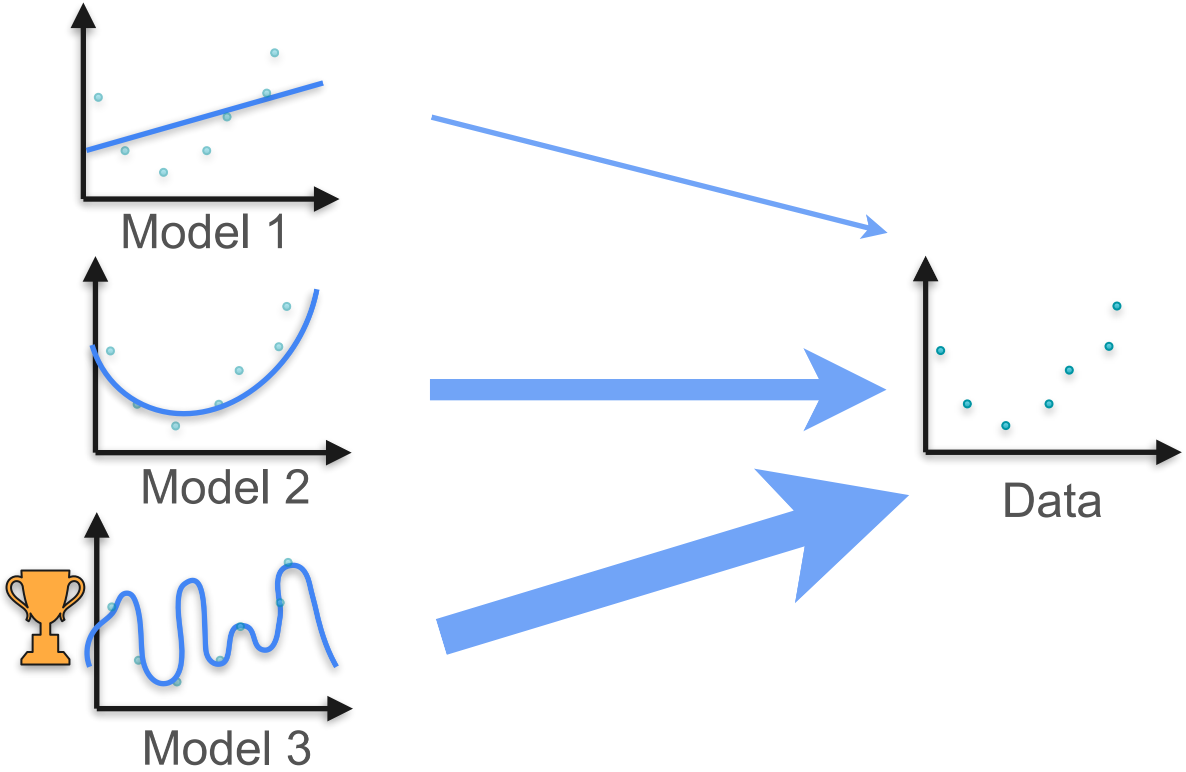


Example: Polynomial Regression

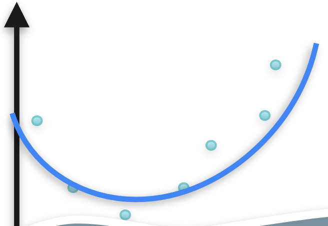
$\mathbf{P}$(Model 1) $\mathbf{P}$(Data|Model 1)

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

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



Maximum likelihood with Bayes



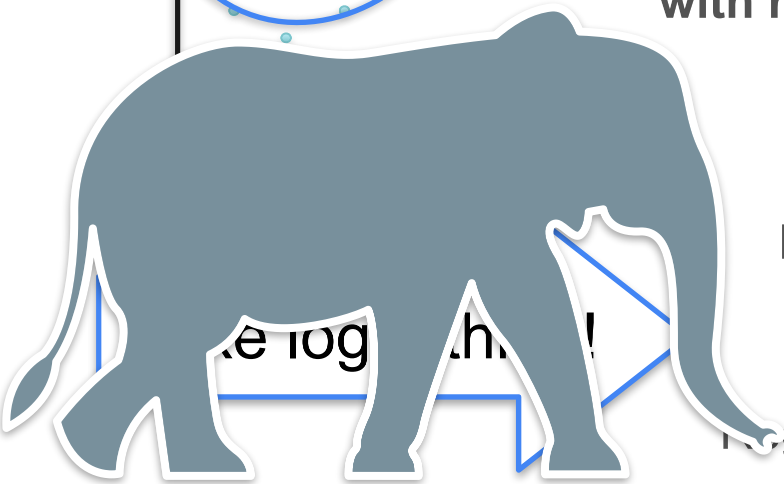
Polynomial regression with regularization

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

$\cdot$

$\boxed{\mathbf{P}(\text{Model})}$

?



e log thi !

Log-loss

+

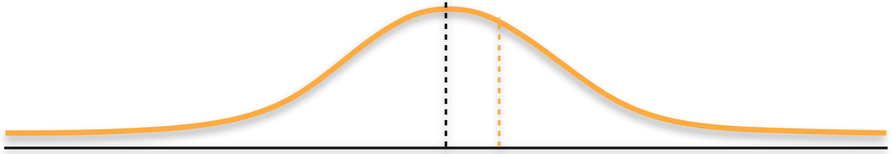
gularization term

What Is the Probability of a Model?

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

$$\boxed{a_1}x + b$$

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_i^2}$$



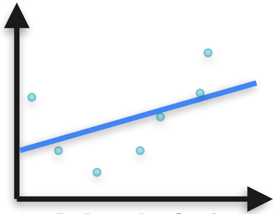
$$a_i \sim \mathcal{N}(0,1)$$

$$\mathbf{P}(\text{Model 2})$$

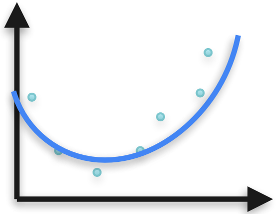
$$\boxed{a_1}x^2 + \boxed{a_2}x + b = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_1^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_2^2}$$

$$\mathbf{P}(\text{Model 3})$$

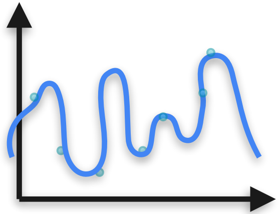
$$\boxed{a_1}x^{10} + \boxed{a_2}x^9 + \cdots + \boxed{a_{10}}x + b = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a_i^2}$$



Model 1



Model 2

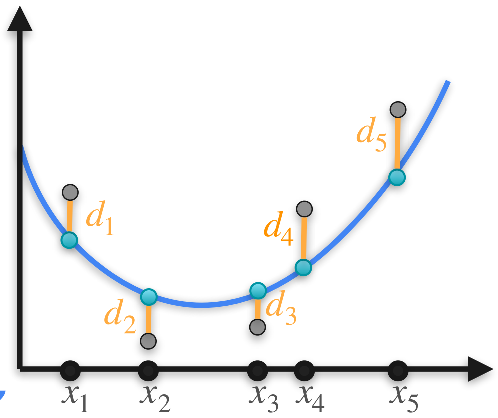


Model 3

Bayes and Regularization

$\mathbf{P}$(Data|Model) $\qquad\qquad$ $\mathbf{P}$(Model)

$$\underbrace{\cancel{\frac{1}{\sqrt{2\pi}}}e^{-\frac{1}{2}d_1^2}\cdot\cancel{\frac{1}{\sqrt{2\pi}}}e^{-\frac{1}{2}d_2^2}\cdot\cancel{\frac{1}{\sqrt{2\pi}}}e^{-\frac{1}{2}d_3^2}\cdot\cancel{\frac{1}{\sqrt{2\pi}}}e^{-\frac{1}{2}d_4^2}\cdot\cancel{\frac{1}{\sqrt{2\pi}}}e^{-\frac{1}{2}d_5^2}}_{\boxed{\log}}\;\underset{\boxed{\log}}{\cdot}\;\underbrace{\cancel{\frac{1}{\sqrt{2\pi}}}e^{-\frac{1}{2}a_1^2}\cdot\cancel{\frac{1}{\sqrt{2\pi}}}e^{-\frac{1}{2}a_2^2}}_{\boxed{\log}}$$



$$\boxed{a_1}x^2 + \boxed{a_2}x + b$$

Maximize

$$-\frac{1}{2}(d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2) \quad + \quad -\frac{1}{2}(a_1^2 + a_2^2)$$

Minimize

$$\boxed{d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 \quad + \quad a_1^2 + a_2^2}$$

Regularization

New Loss

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

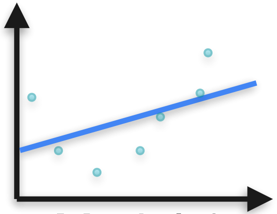
Minimize $\boxed{x_1^2 \quad + \quad d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2}$

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

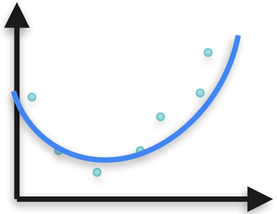
Minimize $\boxed{x_1^2 + x_2^2 \quad + \quad d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2}$

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

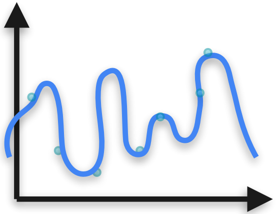
Minimize $\boxed{x_1^2 + \cdots + x_{10}^2 \quad + \quad d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2}$



Model 1



Model 2



Model 3

DeepLearning.AI

Point Estimation

Conclusion