Copyright Notice

DeepLearning.AI

Probability and Statistics for Machine Learning and Data Science

-

Week 3: Sampling and Point Estimates

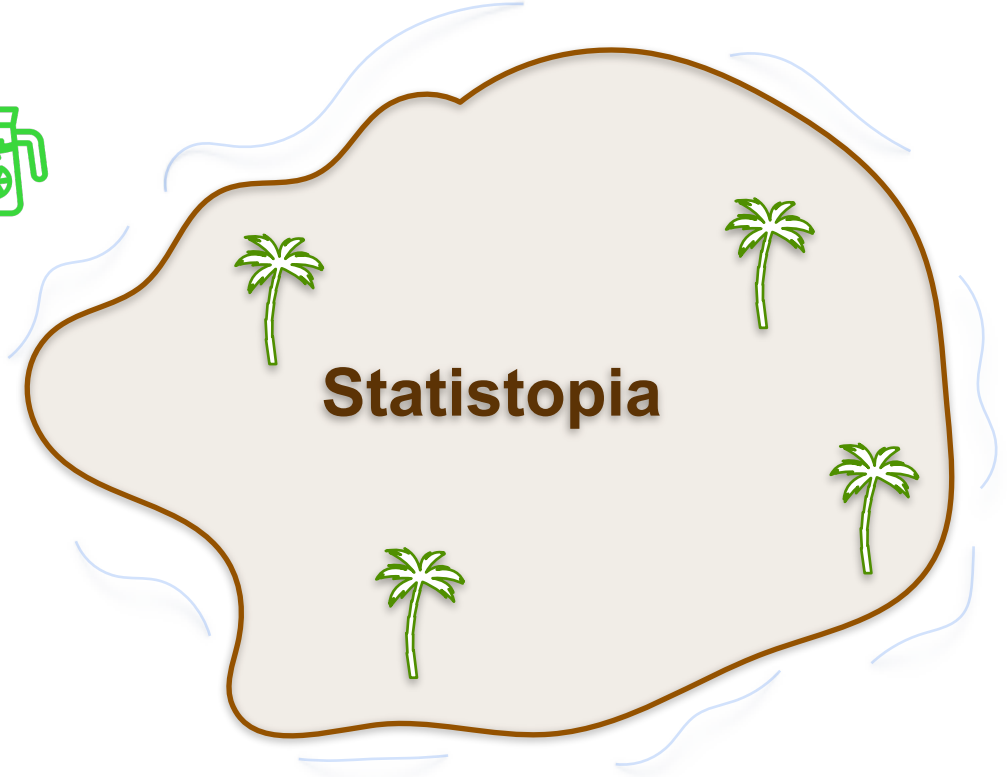
W3 Lesson 1

DeepLearning.AI

Sample and Population

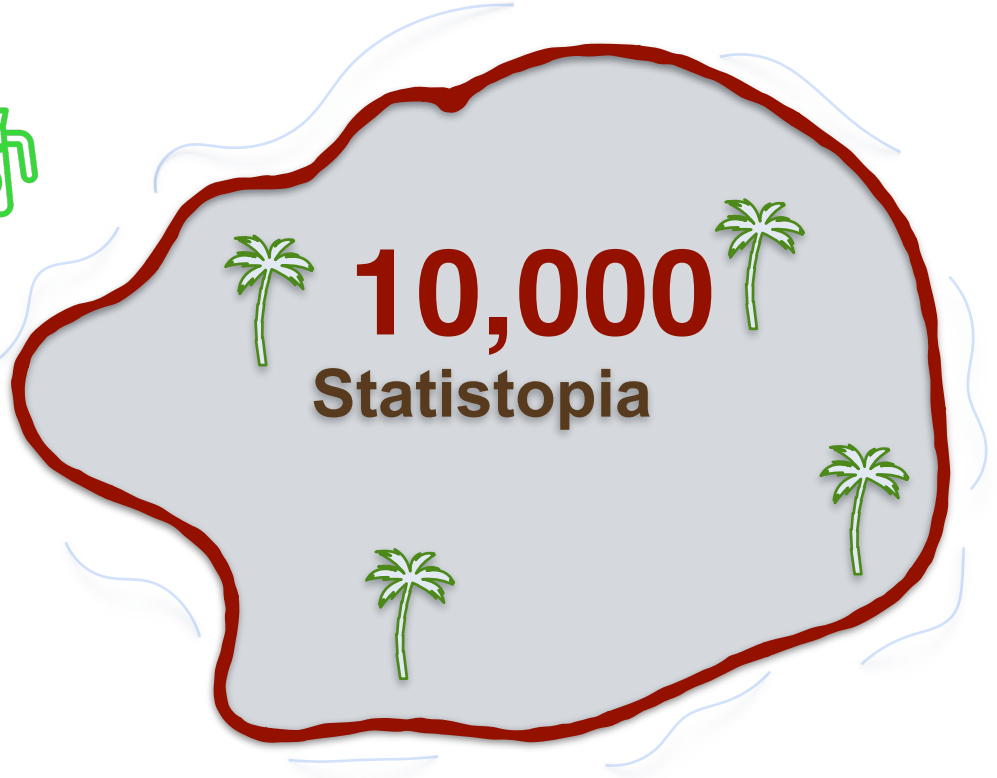
Population and Sample

Population and Sample



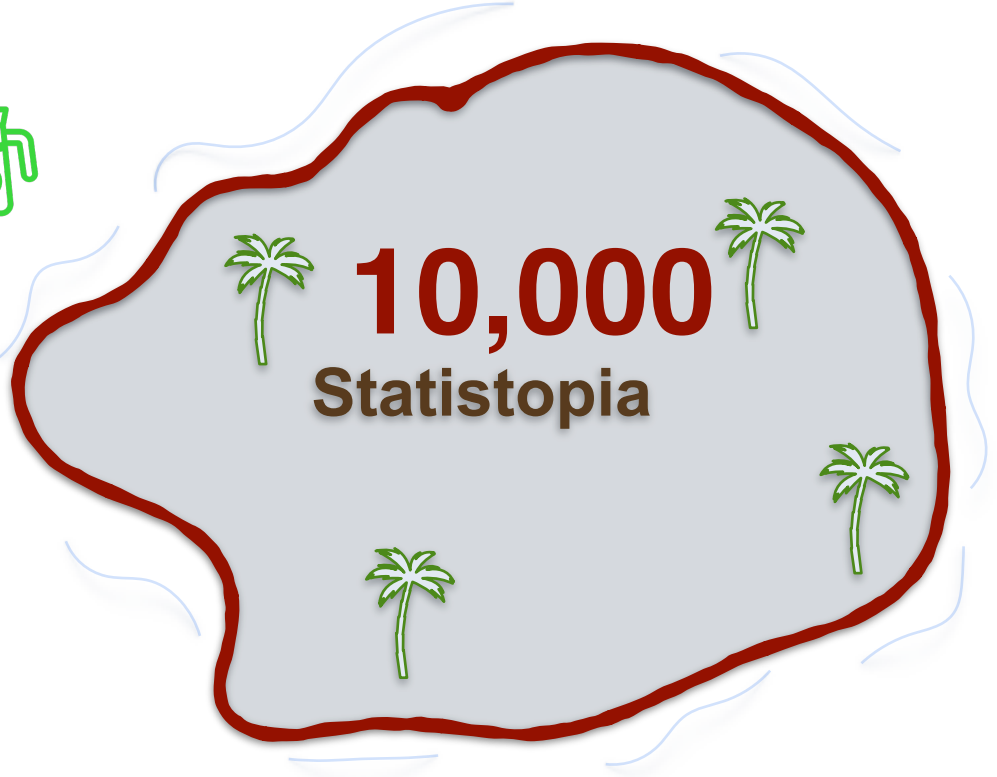
Find the **average height** of the people living on Statistopia

Population and Sample



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

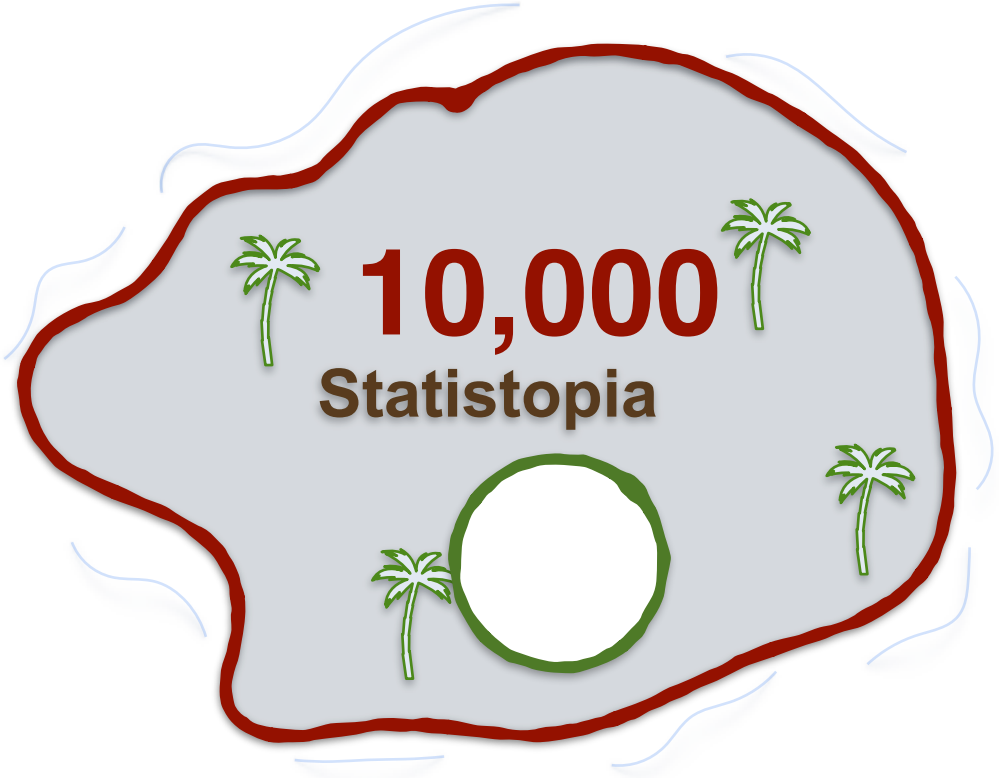
Population and Sample



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

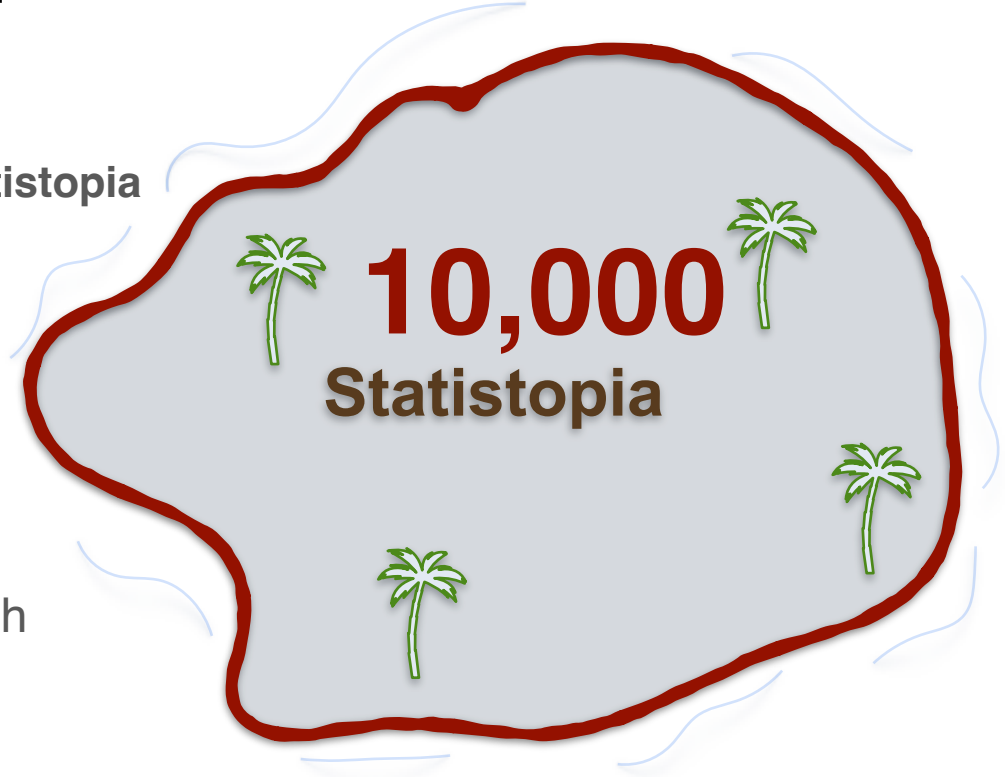
Population and Sample

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



Population and Sample

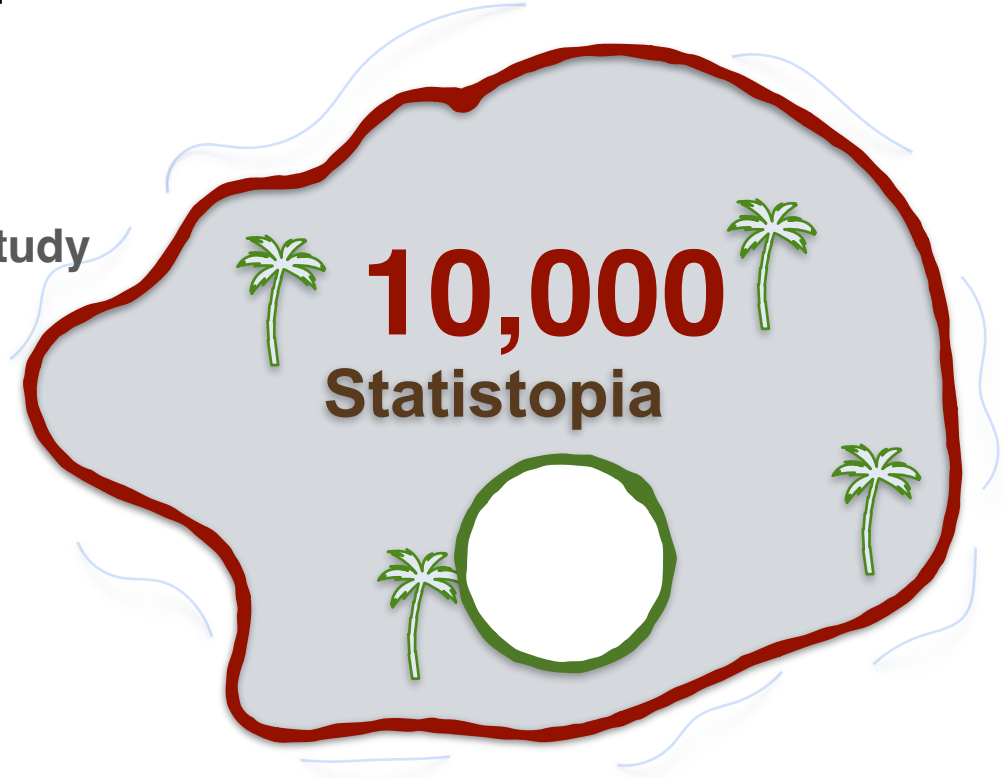
The people of statistopia



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

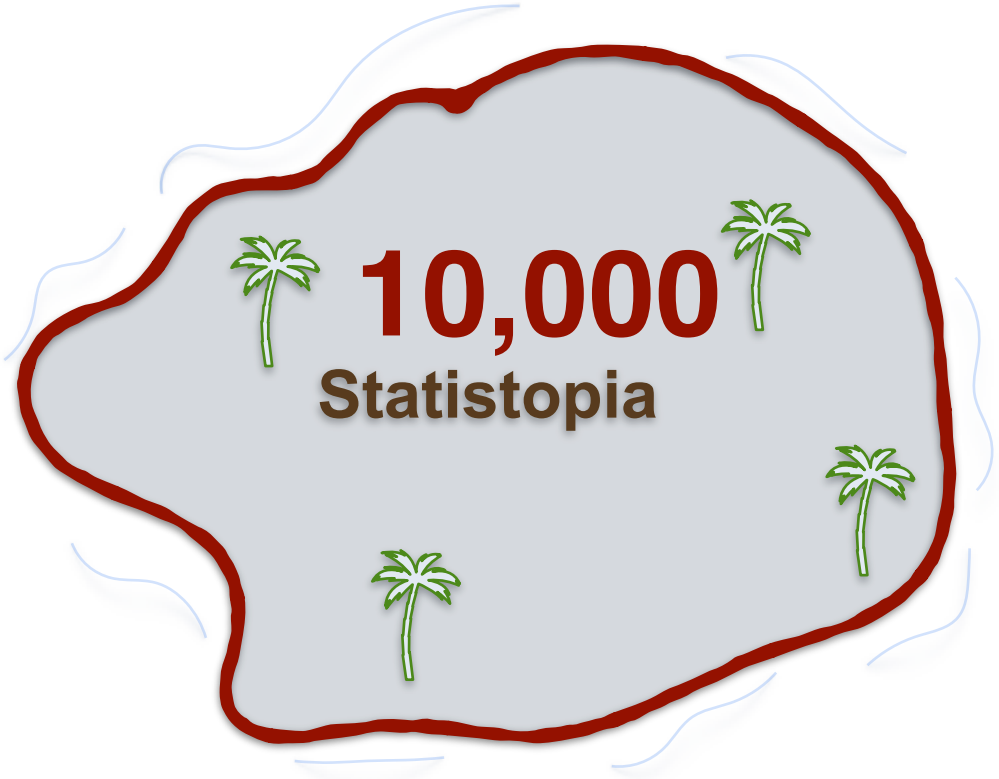
Population and Sample

Population Size (N)

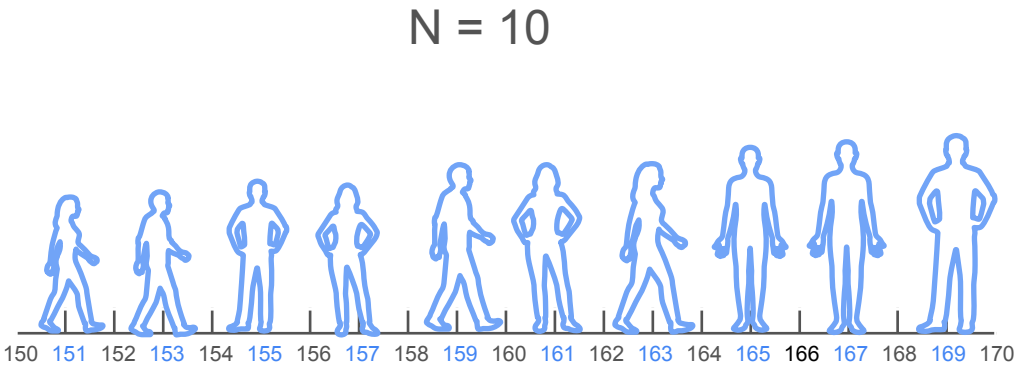
10,000

Sample Size (n)

1 - 9,999

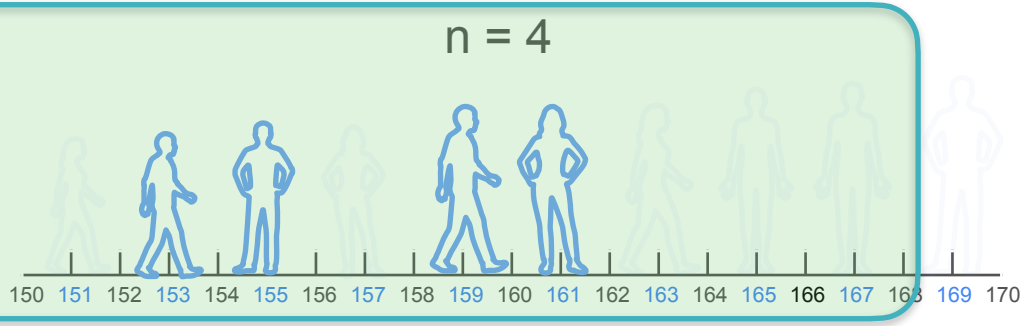


Population and Sample



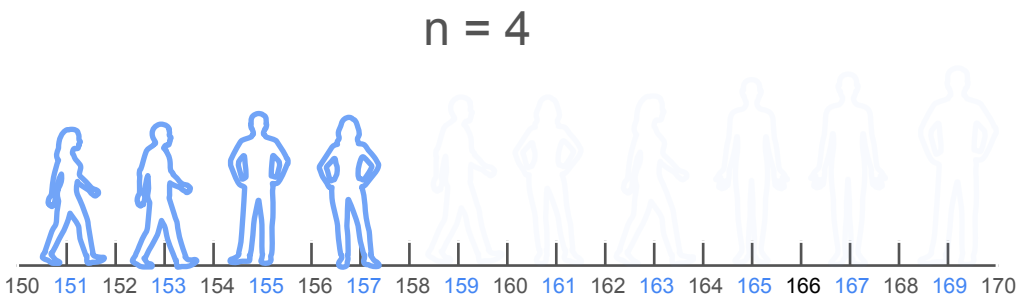
Random Sampling

A



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

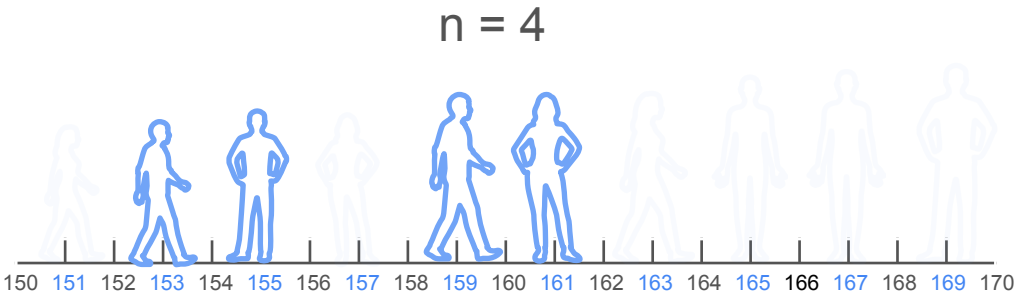
B



Independent Sample

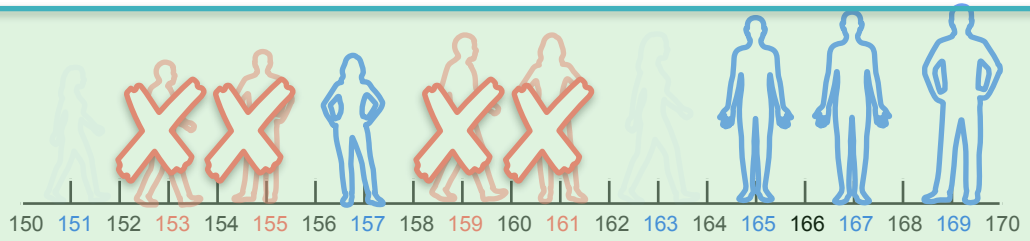
Example 1

1st sample set



Why is sample set two not a good sample?

2nd sample set

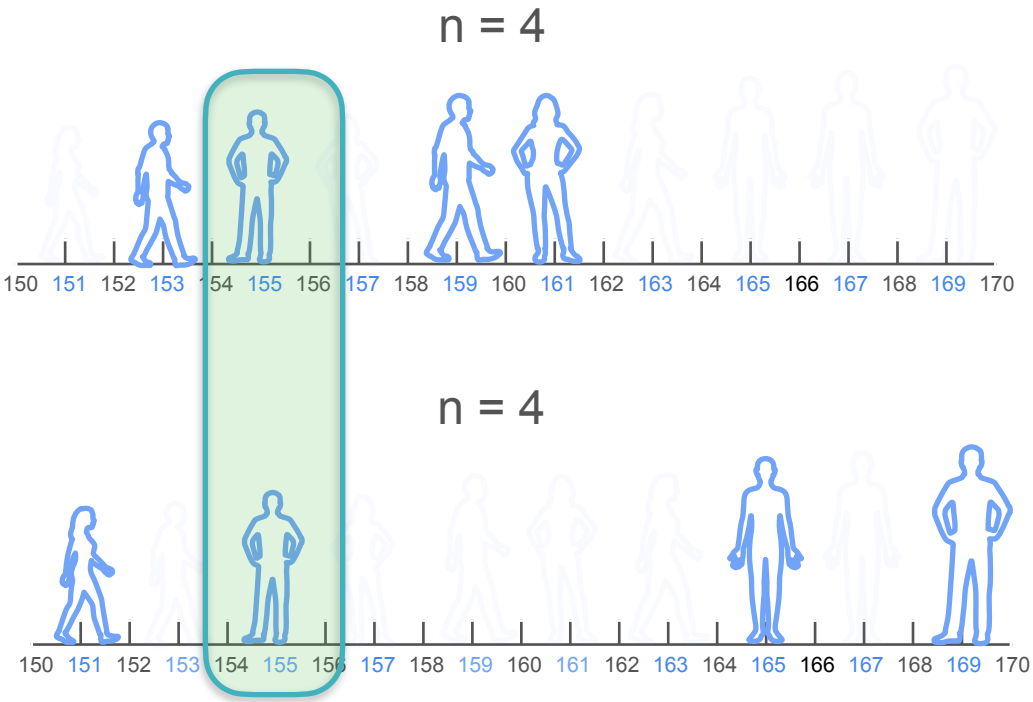


Independent Sample

Example 2

1st sample set

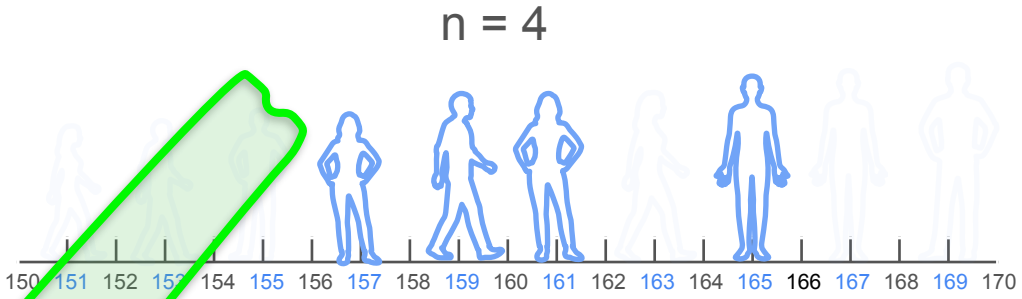
2nd sample set



Identically Distributed Samples

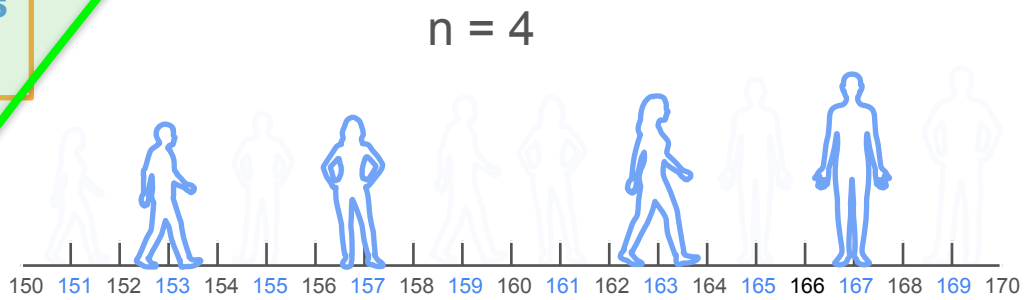
Example 1

A



Which of the following samples are identically distributed?

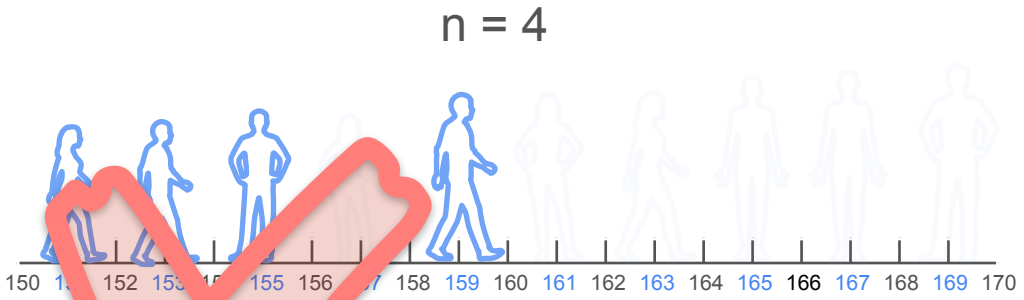
B



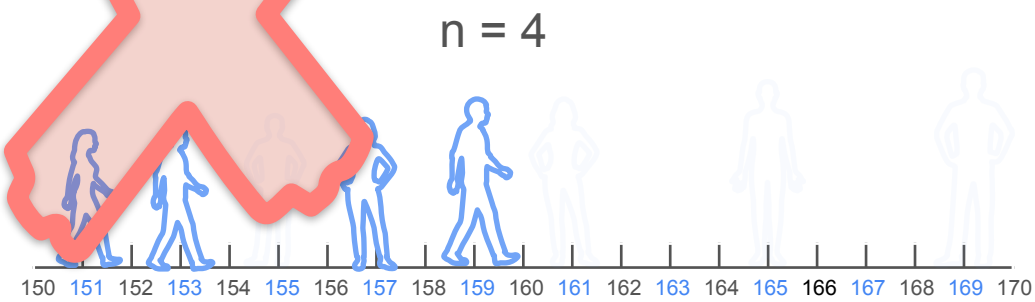
Identically Distributed Samples

Example 2

A

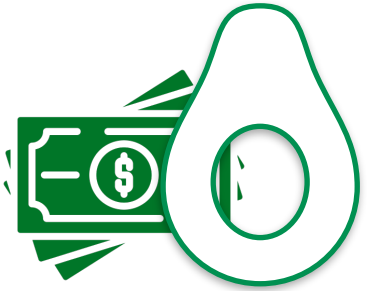


B



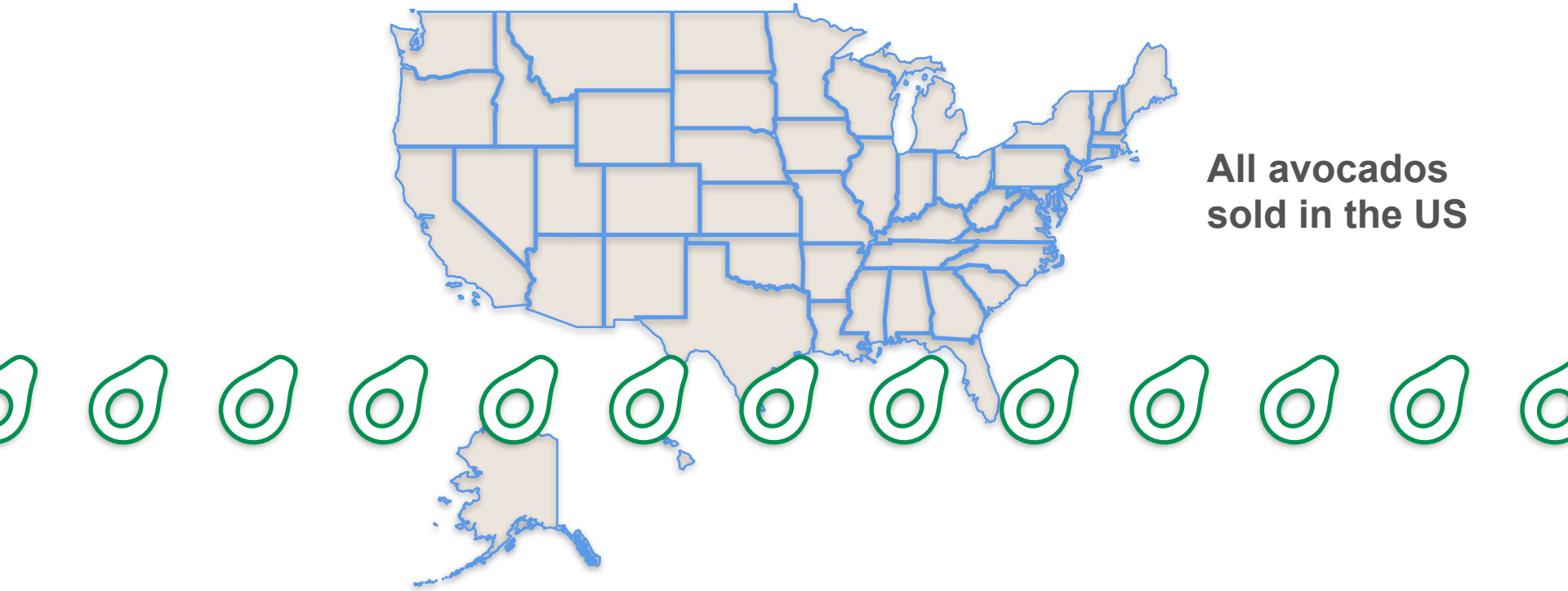
The Avocado Toast Trend

Study the price of avocados in the United States



What is the population of your study?

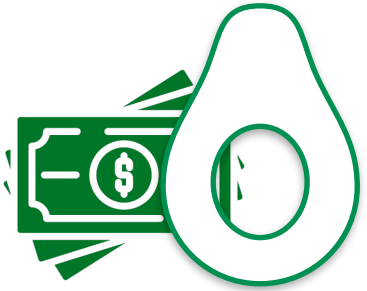
The Avocado Toast Trend



All avocados sold in the US

The Avocado Toast Trend

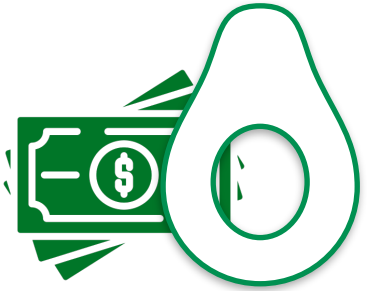
Study the price of avocados in the United States



What is the sample of your study?

The Avocado Toast Trend

Study the price of avocados in the United States



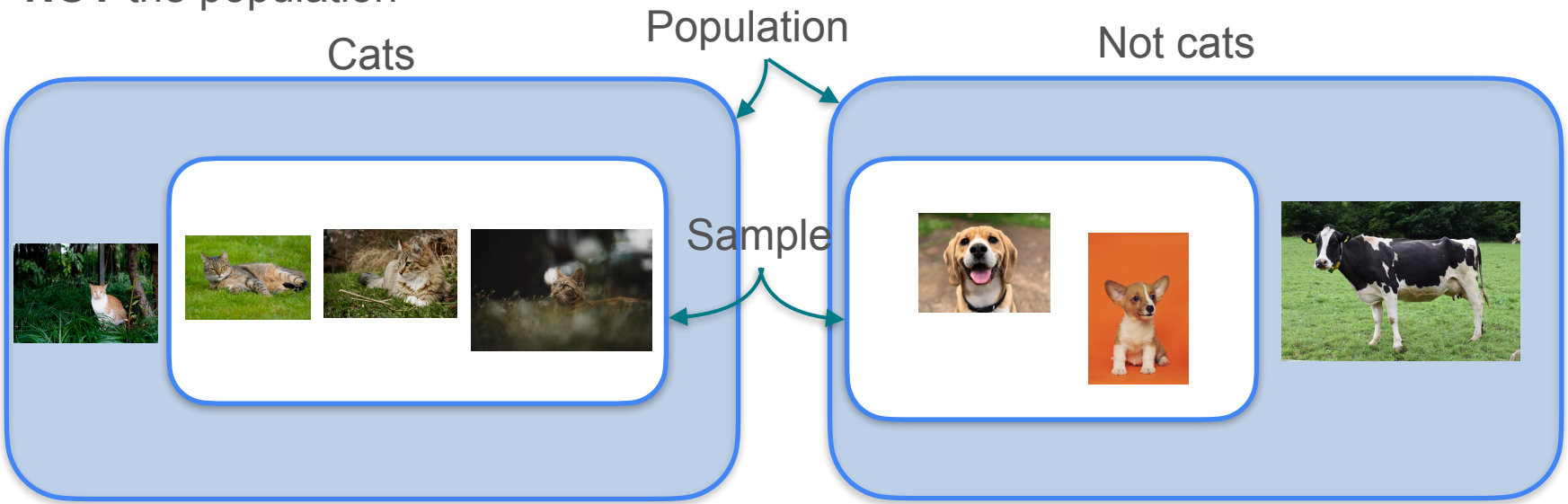
What is the sample of your study?

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

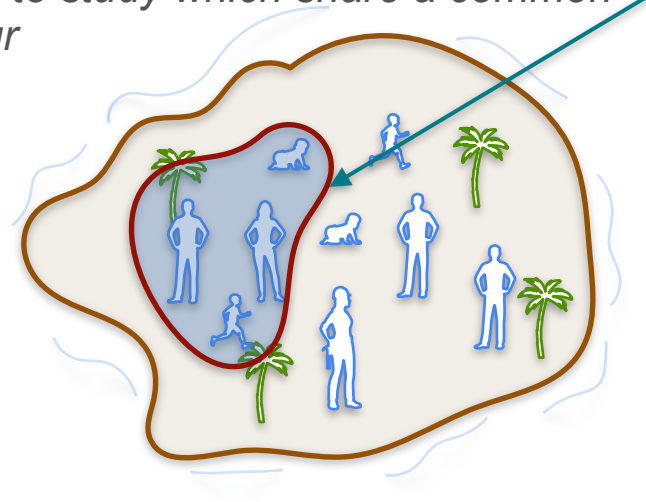
NOT the population



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?

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

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

Population mean μ

Population and Sample Mean

$n = 6$

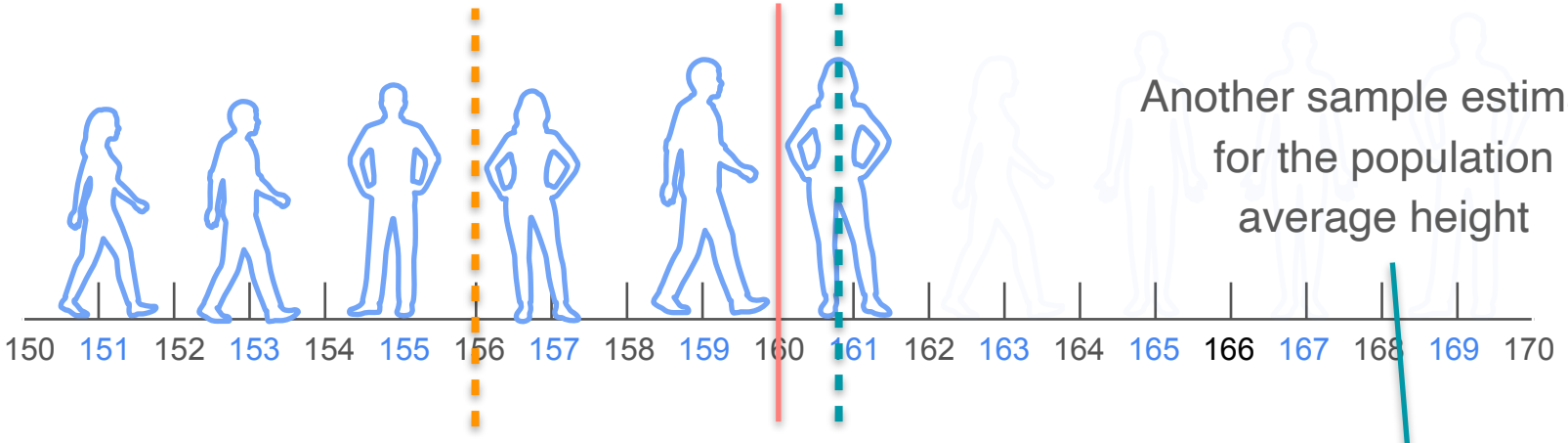
μ $\bar{x}_1$

What is the average height in statistopia?

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

$\bar{x}$

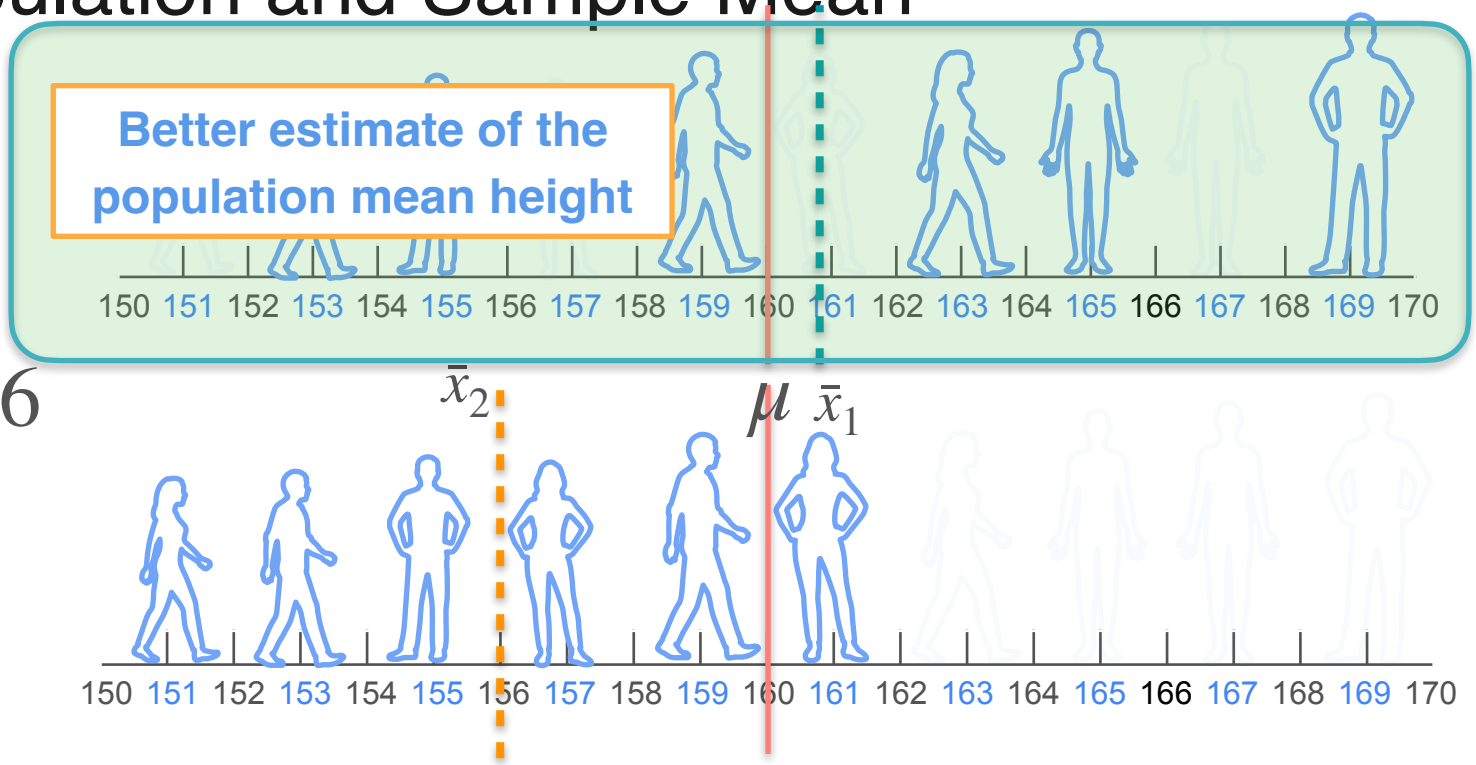
Population and Sample Mean



What is the average height in statistopia?

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

Population and Sample Mean

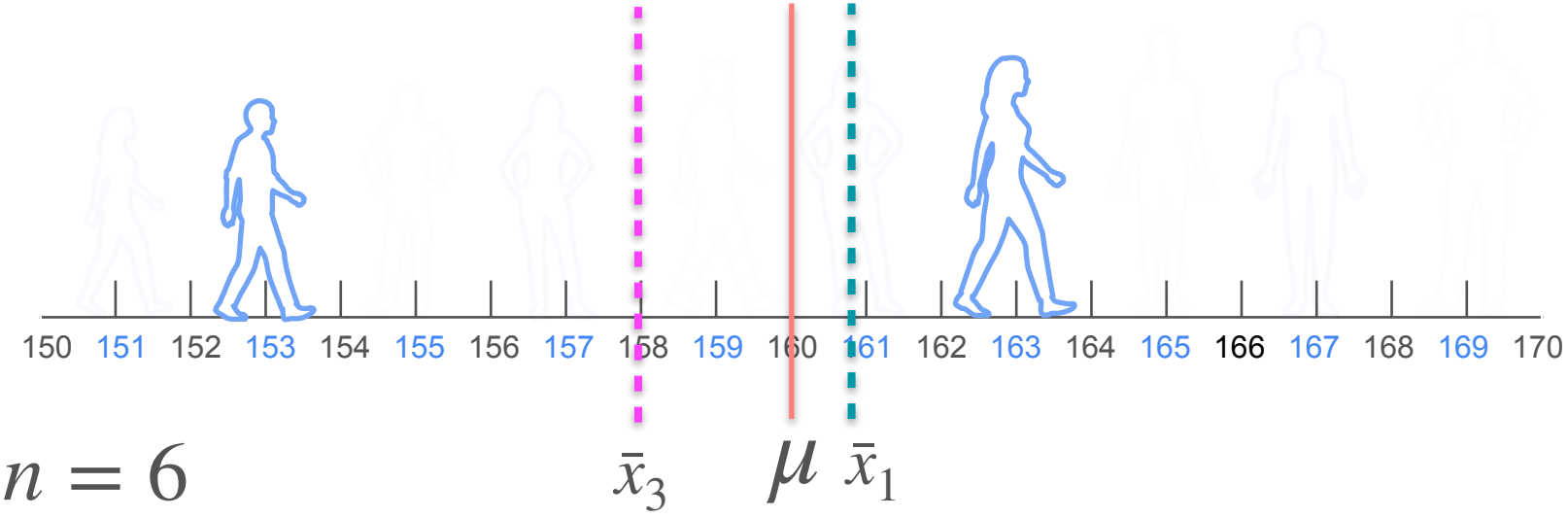


Population and Sample Mean

$n = 6$

$n = 2$

Population and Sample Mean



$n = 6$

$n = 2$

What is the average height in statistopia?

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

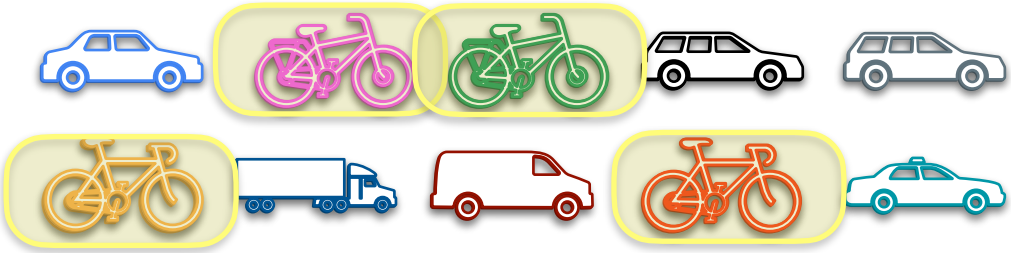
DeepLearning.AI

Sample and Population

Sample Proportion

Proportion

Population size: 10



What proportion of people own a bicycle?

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

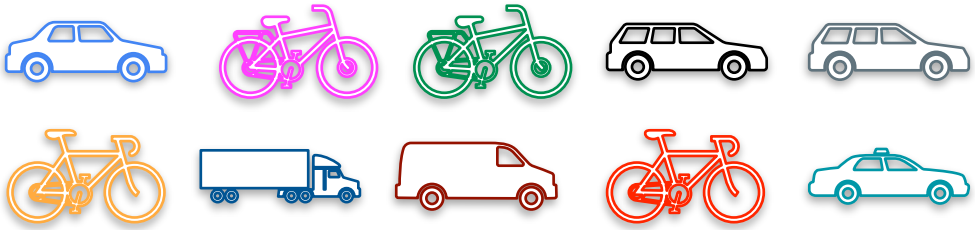
Proportion

population proportion

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

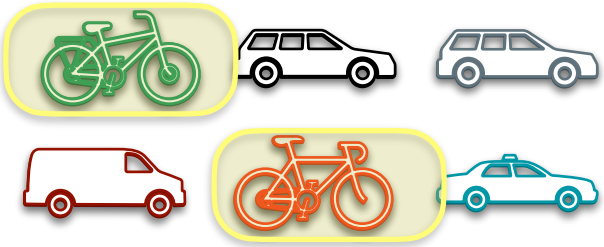
Proportion

Population size: 10



Sample Proportion

Sample size: 6



What proportion of people own a bicycle?

estimate of the population proportion

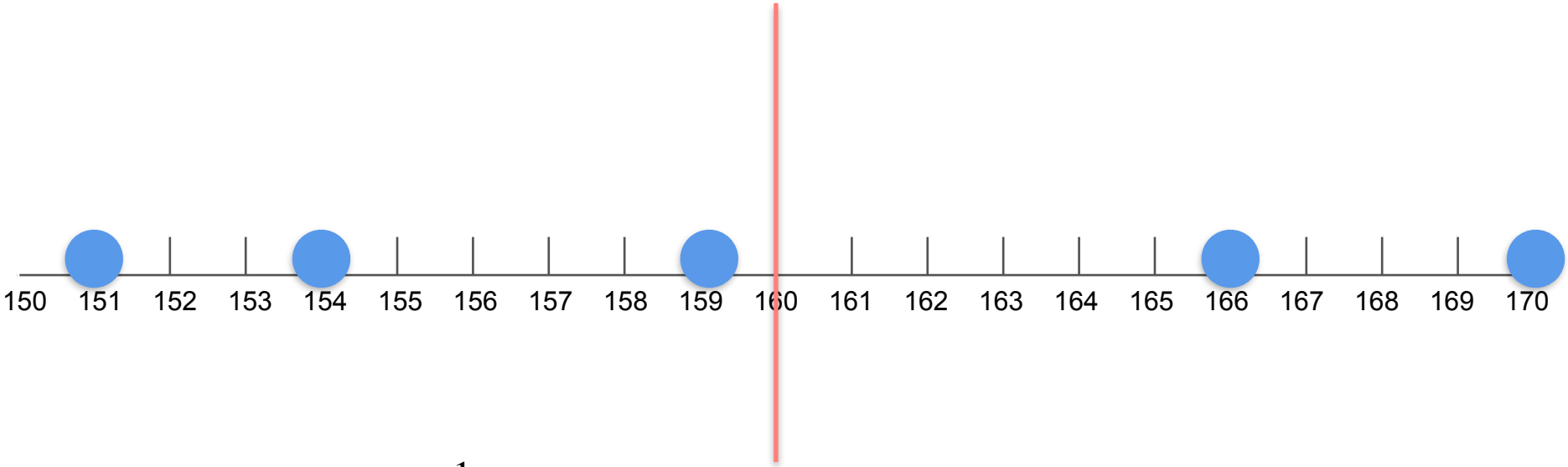
$$\hat{p} \text{ sample proportion } = \frac{2}{6} = 0.333 = 33.3\,\%$$

DeepLearning.AI

Sample and Population

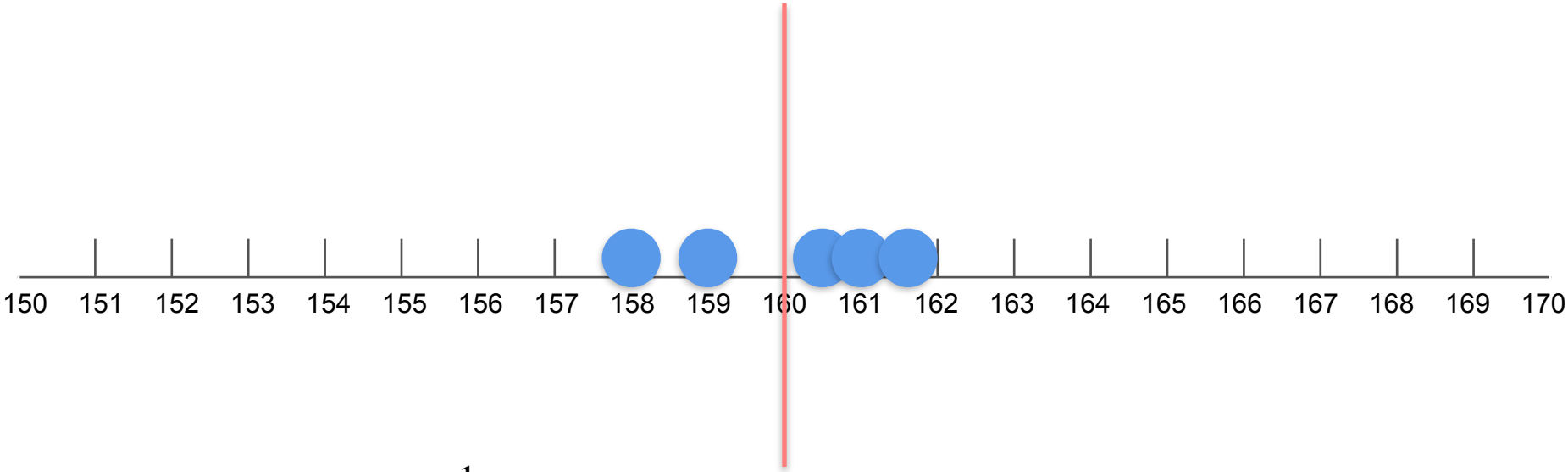
Sample Variance

Sample Variance



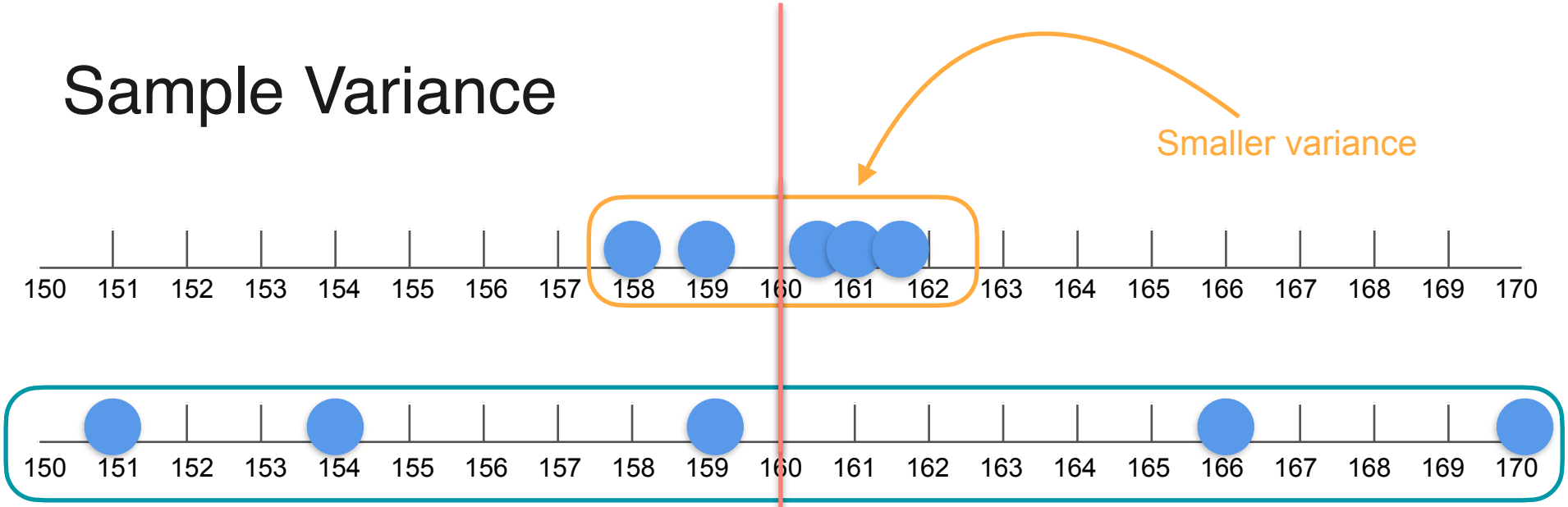
$$\mu = \frac{1}{5}\left(151 + 154 + 159 + 166 + 170\right) = 160$$

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?

Sample Variance

Let's cheat and use the sample mean

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

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

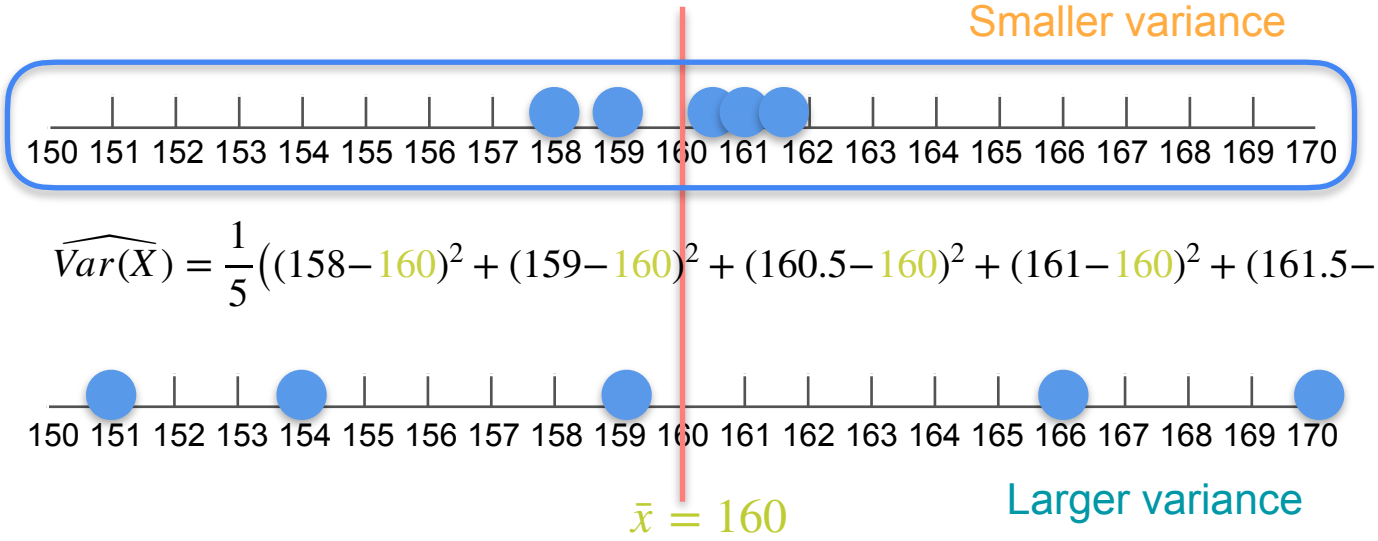
$$\mathbb{E}[Y] = \mu_Y = \frac{1}{N}\sum_{i=1}^{N} y_i \longrightarrow \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The population mean of Y

The sample mean of Y

Sample Variance

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

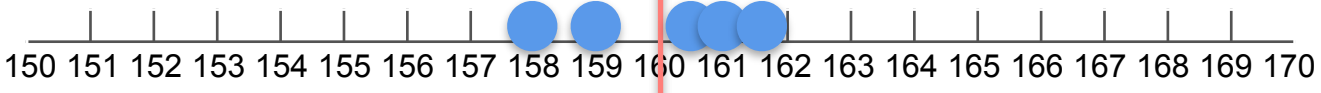


Sample Variance

This equation is “biased”
It underestimates the population variance

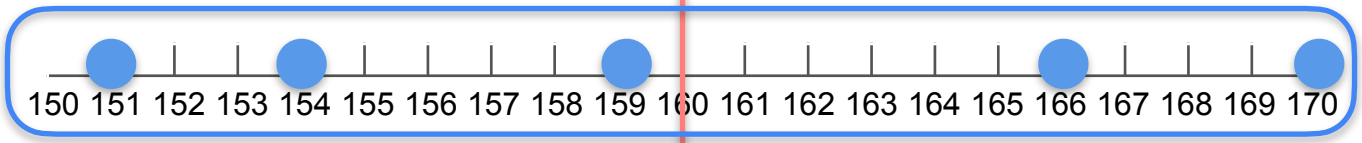
$$Var(X) = \sigma^2 = \frac{1}{N}\sum (x-\mu)^2 \longrightarrow \widehat{Var(X)} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Smaller variance



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

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

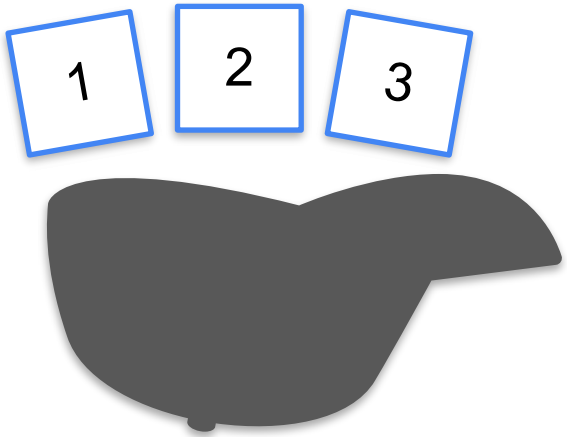


$\widehat{Var(X)} = 50.8$

$\bar{x} = 160$

Larger variance

Variance Estimation



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

Variance Estimation

$\mu = 2$

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

$x - 2$

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

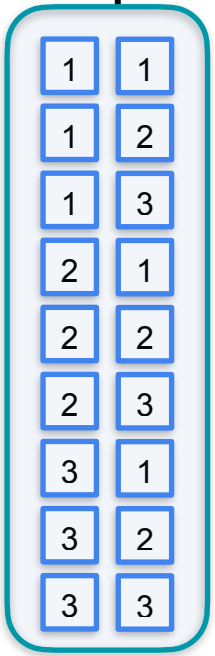
$\mu = 2$

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

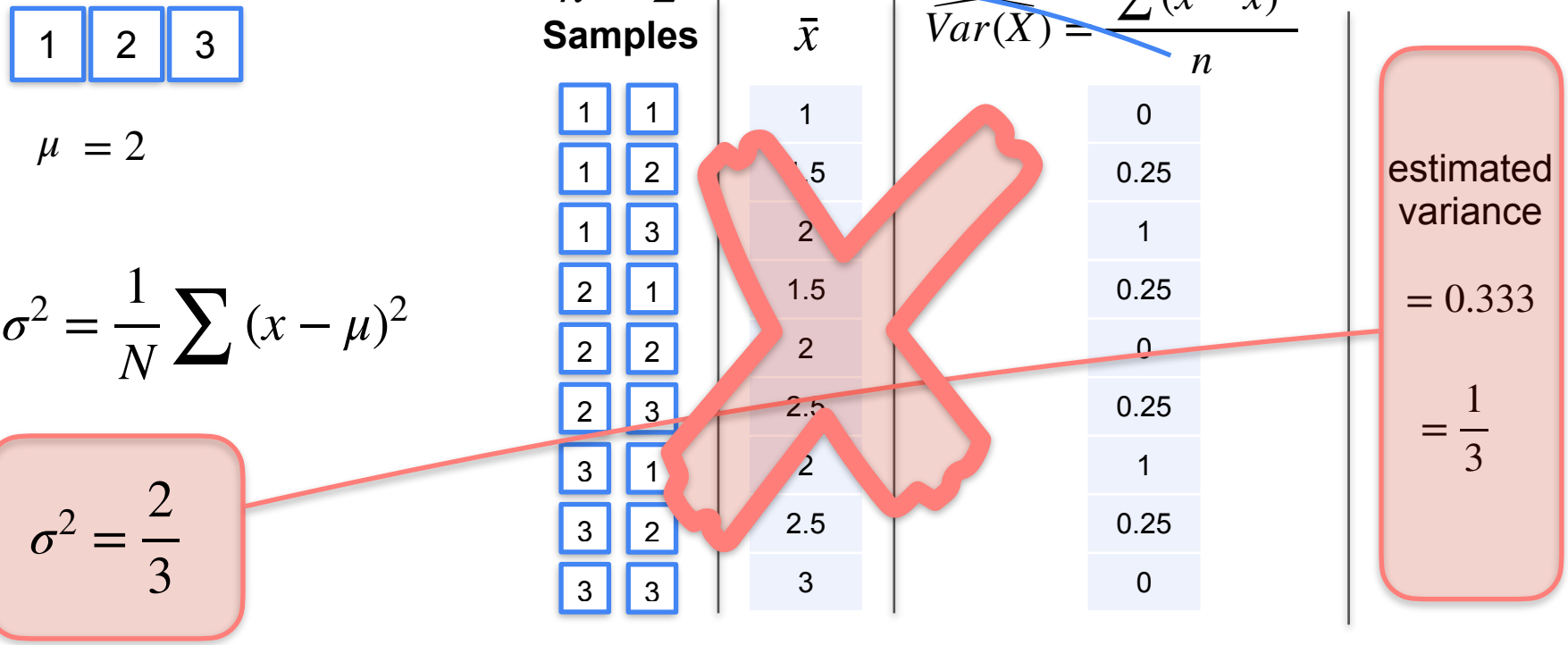
$$\sigma^2 = \frac{2}{3}$$

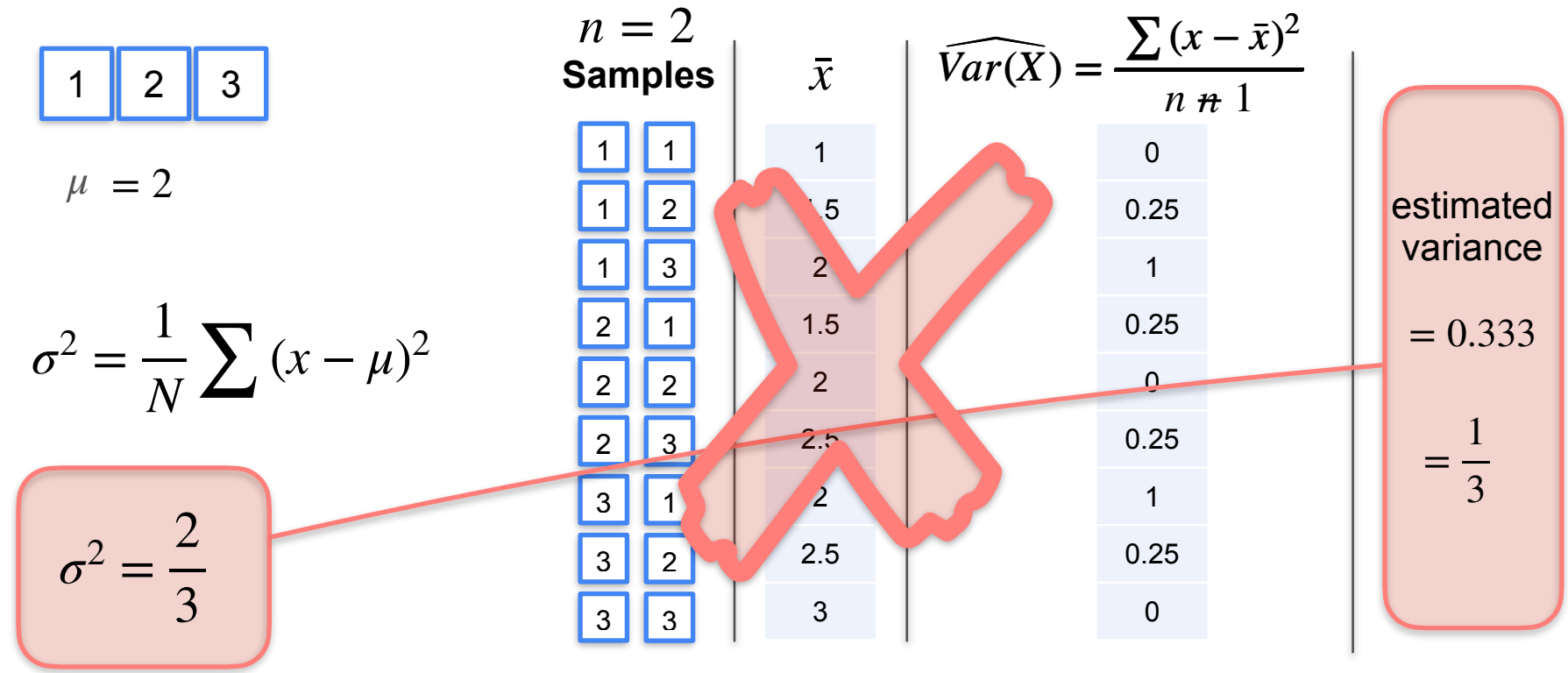
$n = 2$

Samples



Variance Estimation




1 2 3
$\mu = 2$
$\sigma^2 = \frac{1}{N}\sum(x-\mu)^2$
$\sigma^2 = \frac{2}{3}$
$n = 2$
Samples
$\bar{x}$
$\widehat{Var(X)} = \frac{\sum(x-\bar{x})^2}{n-1}$
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
$= 0.333$
$= \frac{1}{3}$

Variance Estimation

$\mu = 2$

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

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

$n = 2$

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

$= 0.667$

$= \frac{2}{3}$

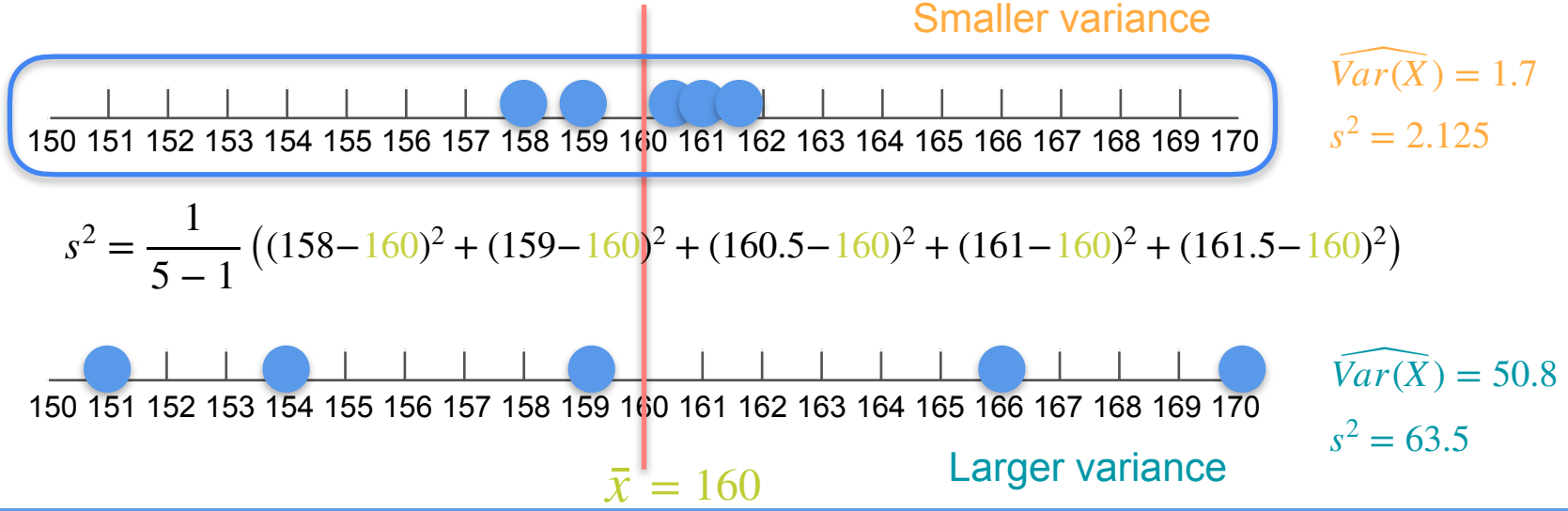
Variance Estimation

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

- 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

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



Variance Estimation

Population Variance Formula

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

Sample Variance Formula

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

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