STAT 135
4. An introduction to inference
Spring 2022

Lecturer: Dr Rebecca Barter (*she/her*)
Office hours: Tu 9:30-10:30, Th 1:00-2:00
Office: Evans 339

Email: rebeccabarter@berkeley.edu
Twitter: @rlbarter
GitHub: rlbarter

Probability prerequisites (e.g. STAT 134)

Be familiar with common distributions:

$N(\mu, \sigma)$, $Binom(p)$, $Unif(min, max)$, $\Gamma(\alpha, \beta)$ (gamma), etc

And know how to work with their densities and compete their expectations and variances.

E.g. if $X \sim N(\mu, \sigma)$, then

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

And $E(X) = \mu$, $Var(X) = \sigma^2$

Probability prerequisites (e.g. STAT 134)

Other topics you should be familiar with are:

Limiting results such as the **central limit theorem** (CLT), e.g., $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$

Conditional probabilities, e.g. $P(X = x \mid Y)$

Independence of random variables, e.g. X and Y and independent if:
$F_{X,Y}(x, y) = F_X(x)F_Y(y)$

Probabilistic properties of random variable distributions, e.g., if X has a symmetric dist then $P(\mid X - a \mid > b) = P(X - a < -b) + P(X - a > b)$

Calculus prerequisites

Know how to find the value of x that minimizes (or maximizes) a function f(x)

i.e., know how to compete derivatives. E.g.: if

$$f(x) = e^{(x-5)^2}$$

Then to find the value of x that minimizes or maximizes this function, differentiate with respect to x

$$\frac{df(x)}{dx} = 2(x-5)e^{(x-5)^2}$$

Setting to zero and solving for x gives $x = 5$

Linear algebra prerequesits

Know how to:

- Invert a matrix
- Multiply matrices together
- (Compute the expected value and variance of a random matrix)

Parameters and estimates

Parameters and populations

A **population** is the complete set of individuals or entities that we are interested in. We usually only have data on a subset of them.

A **parameter** is any quantifiable feature of a **population**.

Bushfire example

What is the **population**? All animals that were living in the burnt area

What is the **parameter** of interest? The proportion of the animals that were killed by the fires

Photo credit: NYTimes

What is the <u>population</u> and <u>parameter</u> of interest for the following questions:

An insurance company hoping to update its rates has conducted a review of its members' data to determine what the average annual claim amounts are for Male customers aged 18-25.

What is the population? *All male customers aged 18-25*

What is the parameter of interest? The average annual claim amounts

What is the population and parameter of interest for the following questions:

Kaiser is using data collected from their patients to provide an estimate of the proportion of American women who will develop breast cancer over their lifetimes.

What is the population? *All American women*

What is the parameter of interest? *The proportion who will develop breast cancer in their lifetimes*

Common parameters of interest in statistics

In statistics, the most *common* population parameters we are interested in are:

- **Mean**
- **Proportions** (Note: Proportions are essentially averages of binary data)

And since we will often assume a distribution, explicit examples of parameters we're interested in include:

- μ and σ from a $N(\mu, \sigma)$ distribution
- p from a $Binomial(n, p)$ distribution
- Etc

Inference

Inference involves using **data** to compute an **estimate** of a **population parameter** of interest

The “population” should always be defined in the context of where the results will be applied

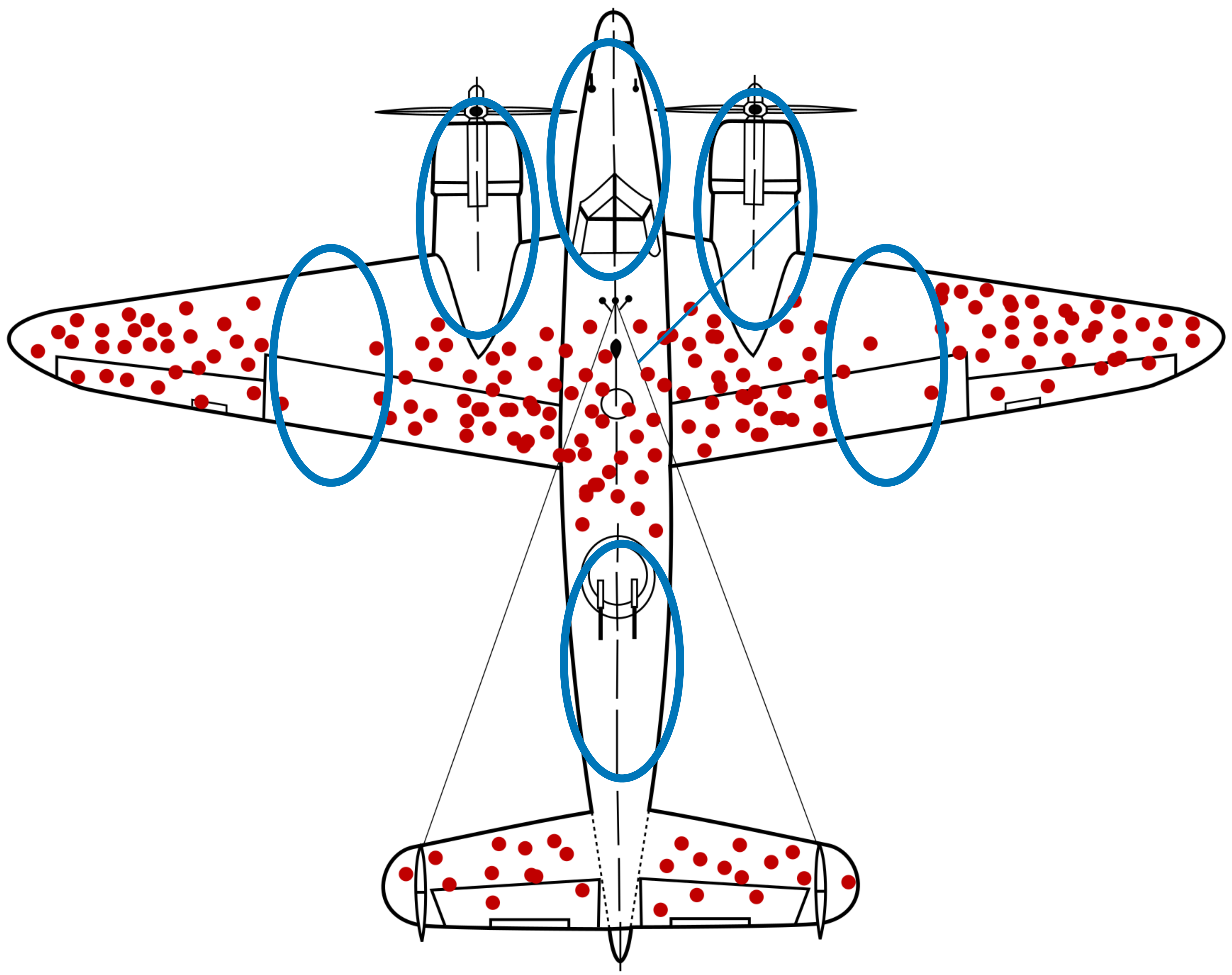
Accurate inference is only possible when the data is representative of the population (i.e., the data is **unbiased**)

Bias in data

Bias in data

Each dot corresponds to a place that a returning plane has been hit

Where should you reinforce the plane's armor?



If the bullets hit here, the plane goes down and does not return

The data is a **biased** representation of where the planes are getting hit

This is known as "survivorship bias"

Is this a random sample from the "population" of all places that the planes get hit?

Bias in data

Data is biased if it does not reflect the population it was designed to represent

Biased data leads to biased results

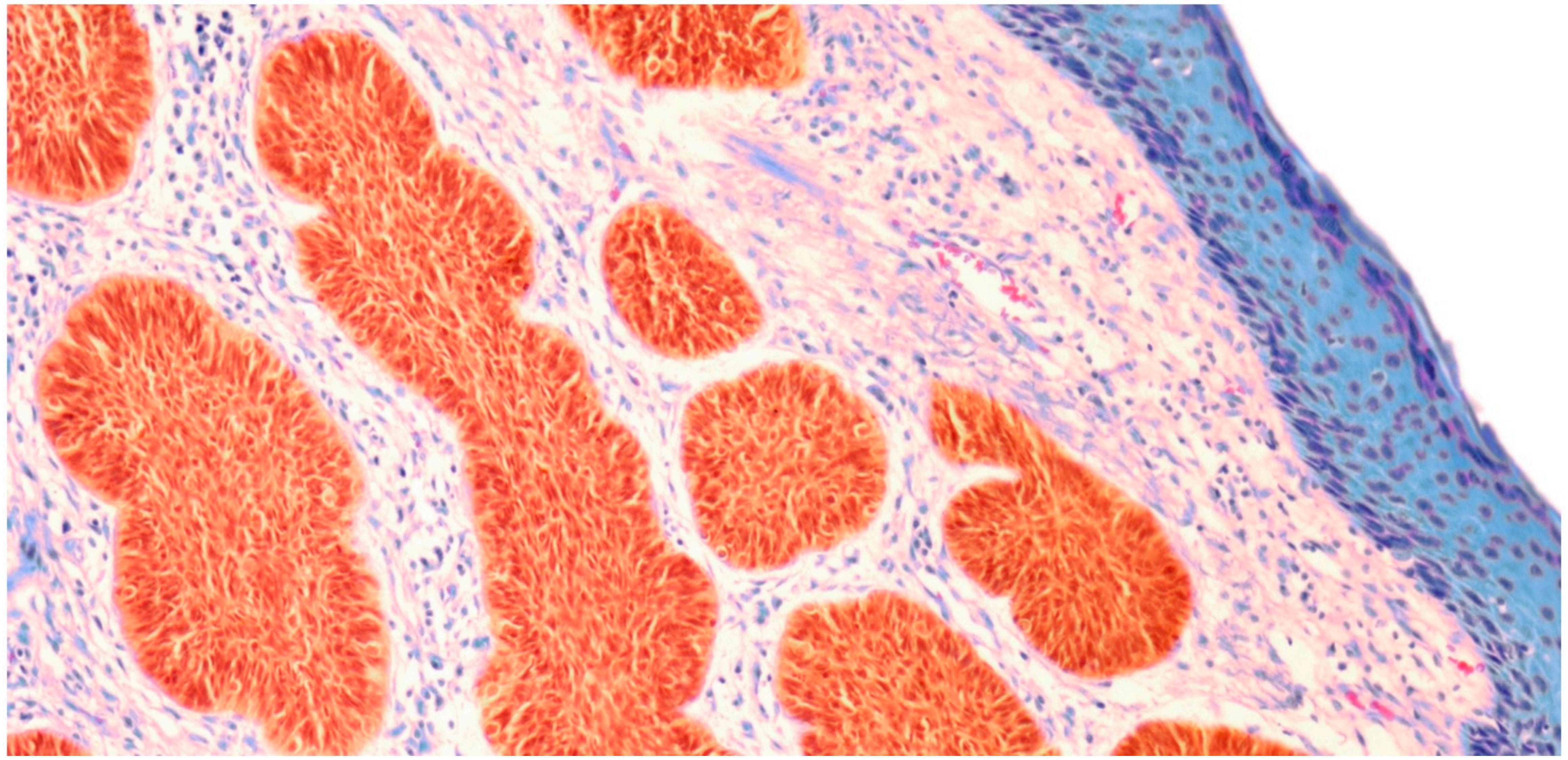
Example: If AI-driven skin cancer detection is built only using patients with light skin tones but is used to detect skin cancer in racially diverse patients, the algorithm might be biased

Popular Latest *The Atlantic* Sign In

AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind

Machine learning has the potential to save thousands of people from skin cancer each year—while putting others at greater risk.

By Angela Lashbrook



https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/

Bias in data

Does this example involve biased data?

Population: All American adults
Parameter: Proportion who are worried about climate change
Data: A street-corner survey asking people whether they are worried about climate change

BIASED!

Bias in data

(This is called cluster sampling)

Does this example involve biased data?

Population: Children in the Alameda school district
Parameter: The average commute time from home to school
Data: All children from ***10 randomly selected schools*** in Alameda county

BIASED!

E.g., If there are:

- a *large number of small private schools* and
- a *small number of large public schools*

This sampling mechanism will over-sample the private school students, and under-sample the public school students.

But this will be unbiased if the schools are approximately the same size

Bias in data

Does this example involve biased data?

Population: Everyone enrolled in STAT 135 this semester
Parameter: The proportion of you who prefer zoom class versus in person
Data: Everyone who fills out Sam Tan's survey

https://tinyurl.com/2p879xn7

Bias in data

Does this example involve biased data?

Population: Everyone enrolled in STAT 135 this semester
Parameter: The proportion of you who prefer zoom class versus in person
Data: Everyone who fills out Sam Tan's survey

https://berkeley.qualtrics.com/jfe/form/SV_5pVUeyVW2MvKtUi

BIASED?

Because there are undoubtedly students who are enrolled in the class that didn't fill out the survey… This is called **response bias.**

Random sampling

Random samples

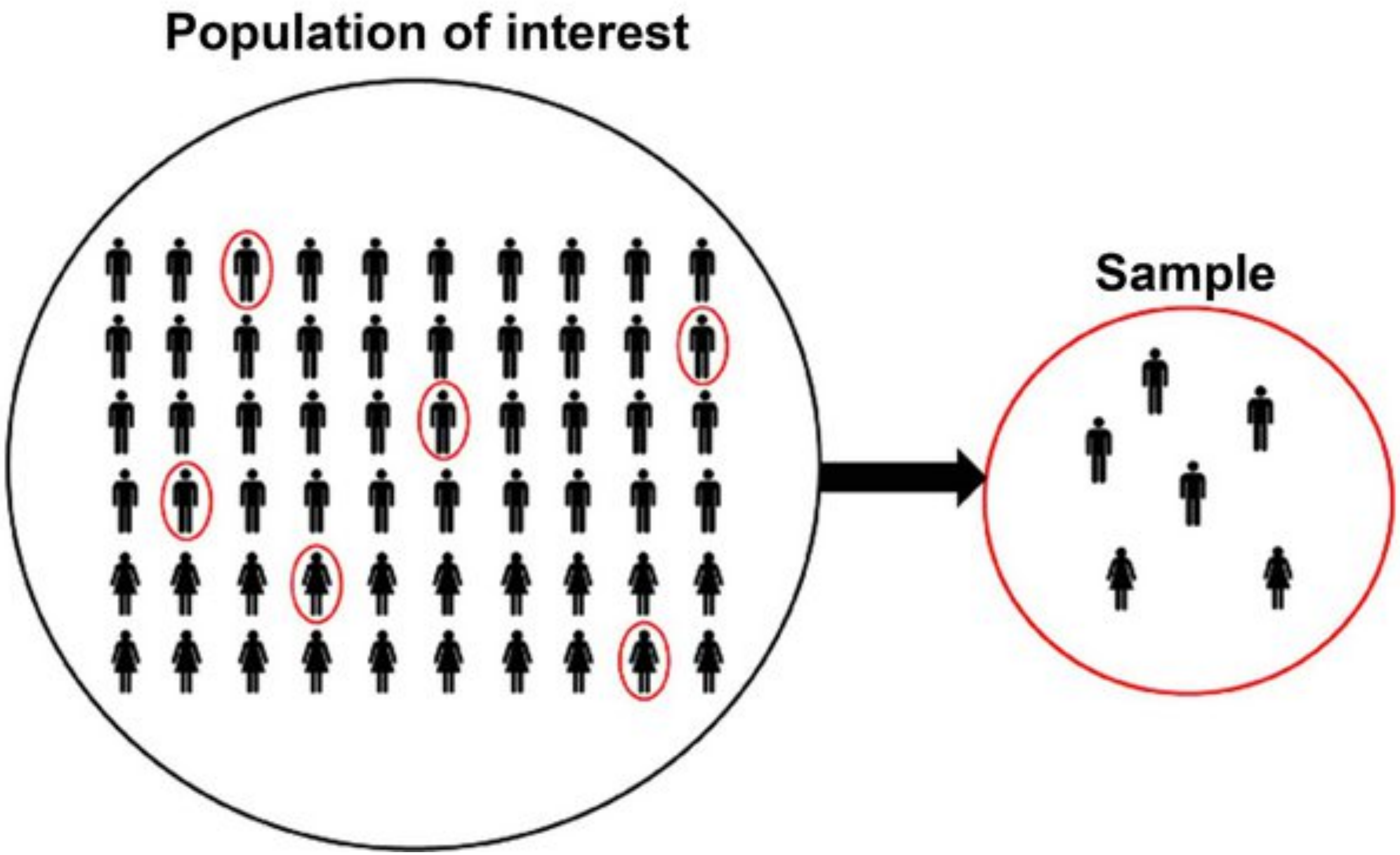


Image stolen from this quora post, which was probably stolen from somewhere else

When a dataset consists of **randomly sampled** observations from the population of interest, it is considered *representative* of that population and is thus **unbiased**.

Under random sampling, everyone in the population has an *equal* chance of being included

Random variables

Random variables

We use **random variables** to represent all possible values that an unknown quantity could take when we observe it

Population:

X denotes a randomly selected observation from the population

The randomness in X comes from the random sampling process

Observed sample:

6 1 4 9 2 0 1 3

x denotes an observed data point in our sample

Random variables

There are many different possible samples that we *could* have taken.

Population:

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

$\mu = 3.9$

Observed sample:

7 2 3 8 3 5 4 2

Sample mean: $\bar{x} = 4.6$

Observed sample:

6 1 4 9 2 0 1 3

Sample mean: $\bar{x} = 3.3$

Observed sample:

3 4 4 8 1 7 4 5

Sample mean: $\bar{x} = 4.4$

Random variables

But we usually only ever observe one:

Population:

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

$\mu = 3.9$

↓

Observed sample:

6 1 4 9 2 0 1 3

Sample mean: $\bar{x} = 3.3$

Where does the randomness come from?

Every random variable has a real-world source of "uncertainty"/ "randomness" that it is supposed to reflect. This could be:

1. The random sampling of the dataset (this is the source of randomness that we will consider in this course)
2. Measurement error
3. …

Whenever you're working with a random variable, pause to consider where the randomness or "source of uncertainty" came from.

In statistical inference, this randomness is almost always the random sampling.

Random variables

Is the following quantity a random variable?

Population:

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

↓

Observed sample:

6 1 4 9 2 0 1 3

Sample mean: $\bar{x} = 3.3$

$\boxed{\mu = ?}$

No, the population mean is a **fixed** property of the population

(…Unless we're Bayesians)

Random variables

Is the following quantity a random variable?

No, the underlying population value is not a random variable.

The randomness in the random variable comes from the random sampling

Population:

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
$\boxed{?}$? ? ? ? ? ? ? ? ? ? ?

$\mu = ?$

Observed sample:

6 1 4 9 2 0 1 3

Sample mean: $\bar{x} =$ 3.3

Random variables

Is the following quantity a random variable?

Population:

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

$\mu = ?$

Observed sample:

$\boxed{6}$ 1 4 9 2 0 1 3

Sample mean: $\bar{x} = 3.3$

X

No, an observed data point in our sample is a fixed number

But… the **hypothetical version** of this sampled data point (i.e. before we look at it) ***is* a random variable**

Random variables

Is the following quantity a random variable?

Population:

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

$\mu = ?$

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

Observed sample:

6 1 4 9 2 0 1 3

Sample mean: $\boxed{\bar{x}} = 3.3$

No, the sample mean of our observed sample is a **fixed** number

But… the **sample mean *based on the random variables* is a random variable**

Independent and identically distributed data

Independent and Identically Distributed (IID) is essentially equivalent to random sampling

IID means that every unit in the population is essentially equivalent and that drawing one into the sample says nothing about how likely the others are to be drawn into the sample

Population:

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

$$\mu = ?$$

$$X$$

The expected value of IID random variables is equal to the population mean:

$$E(X) = \mu$$

Estimating parameters

The sample mean as an estimate of the population mean

Population:

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

$\mu = ?$

Observed sample:

6 1 4 9 2 0 1 3

Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 3.3$

The sample mean is a common estimate of the population mean

But how **good** is the sample mean as an estimate of the population mean?

Evaluating parameter estimates

Two common metrics of parameter estimate performance:

Bias

How close is the sample mean to the population mean?

Variance

How much would the sample mean differ if we had collected a slightly different sample?

Theoretical evaluations of parameter estimates:

1. **Bias**

Parameter bias

A parameter estimate is biased if it is inherently not capturing the unknown population parameter that it is supposed to represent

The **bias** of an estimate, $\hat{\theta}$, of population parameter, θ, is:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

A parameter estimate is **unbiased** if the bias is equal to 0.

Notation:

- θ is the population parameter
- $\hat{\theta}$ is an estimate of the population parameter from the data

The sample mean is an unbiased estimate of the population mean

Theorem: when the observations in a sample *are IID from a population with mean* μ, the sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, is an unbiased estimate of μ.

Proof:

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{n} n\mu$$

$$= \mu$$

Parameter bias

True or False: a parameter estimate from a sample is biased if it is *not* equal to the underlying population quantity it is supposed to represent?

False. Even if the parameter estimate is unbiased, there is no guarantee that the parameter computed from a specific sample of data points will be **exactly** equal to the underlying population parameter

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

$\mu = 3.9$

Random sample

9 6 2 9 3 5 1 2

$\bar{x} = 4.6$

Unbiasedness is referring to the Expected value of the estimate, not the sample estimate itself

The sample standard deviation is a *biased* for the population standard deviation

Theorem: when the observations in a sample are IID from a distribution with mean μ and standard deviation, σ:
The sample standard deviation, $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, is a *biased* estimate of σ.

Proof: Homework 1

You will show that the "sample-size adjusted" sample variance estimate is unbiased for the population variance

$$E\left[\frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \sigma^2$$

What is an *intuitive, real-world* explanation for this?

Theoretical evaluations of parameter estimates:

1. Bias
2. **Variance**

Parameter variance

The **variance** of a parameter estimate tells you how much it generally changes across alternative equivalent versions of the data

Population

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

$\mu = ?$

Observed sample:

9 6 2 9 3 5 1 2

Sample mean: $\hat{\mu} =$ **4.6**

Observed sample:

6 1 4 9 2 0 1 3

Sample mean: $\hat{\mu} =$ **3.3**

Observed sample:

3 4 5 8 1 1 4 9

Sample mean: $\hat{\mu} =$ **4.4**

The **variance** of an estimate, $\hat{\theta}$, of population parameter, θ, is:

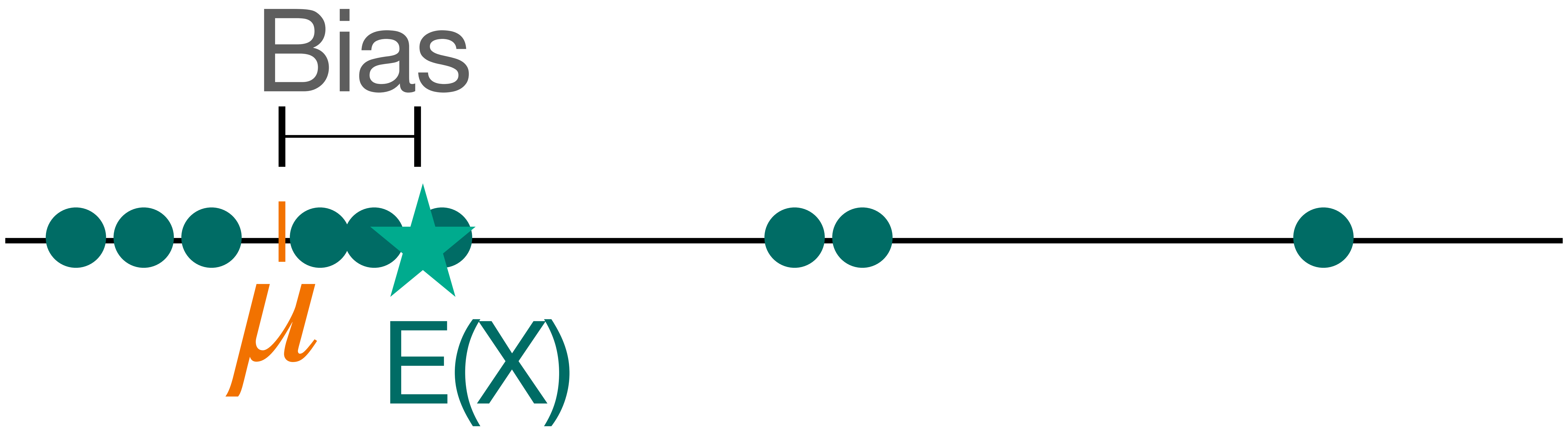
$$Var(\hat{\theta}) = E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right] = E\left[\hat{\theta}^2\right] - E[\hat{\theta}]^2$$

Bias vs variance

● Represents the sample mean of a *hypothetical* sample drawn from the population

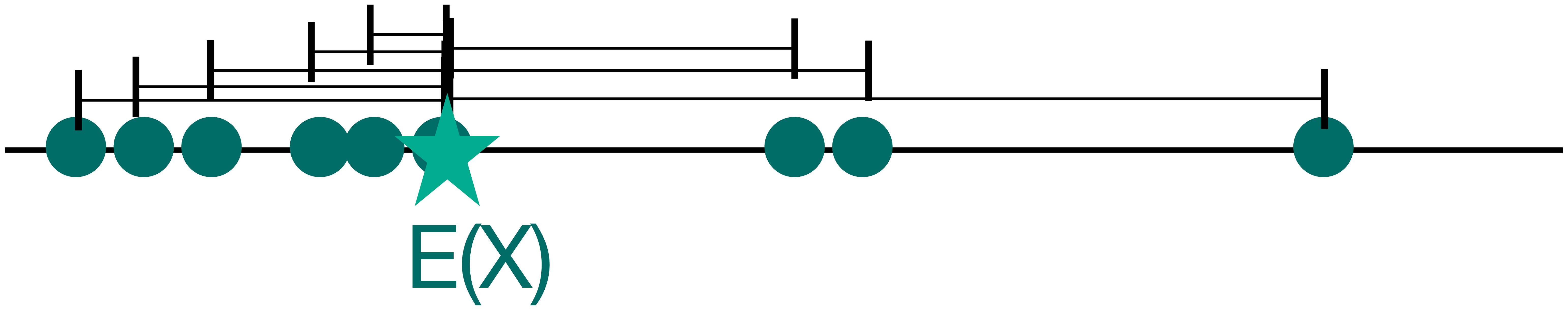
Bias

The distance from the expected value to the true mean



Variance

The average of the (squared) distances from the individual sample means to the expected value



The variance of the sample mean

Theorem: when X1, …, Xn are an IID sample from a population with mean μ and standard deviation σ, the sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, has variance $\frac{\sigma^2}{n}$

Proof:

$$Var(\hat{\mu}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

(Since the X_is are IID)

$$= \frac{1}{n^2} n\sigma^2$$

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

This means that the variance *decreases* as the sample size increases

Theoretical evaluations of parameter estimates:

1. Bias
2. Variance
3. **MSE**

Mean Squared Error

The Mean Squared Error (MSE) is a measure of how "good" an estimate $\hat{\theta}$ is. The MSE is equal to:

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

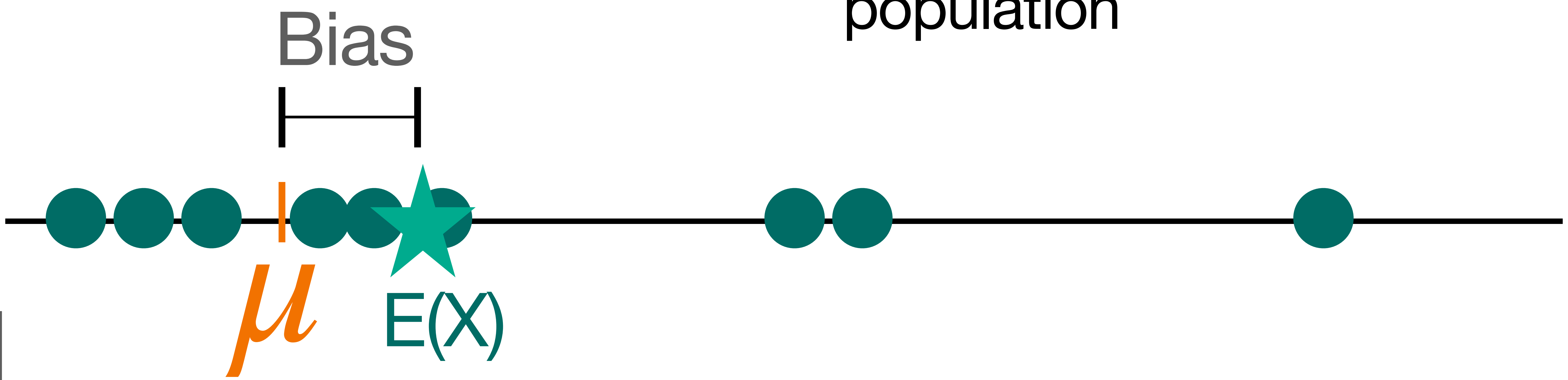
Bias vs variance vs MSE

● Represents the sample mean of a *hypothetical* sample drawn from the population

Bias

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

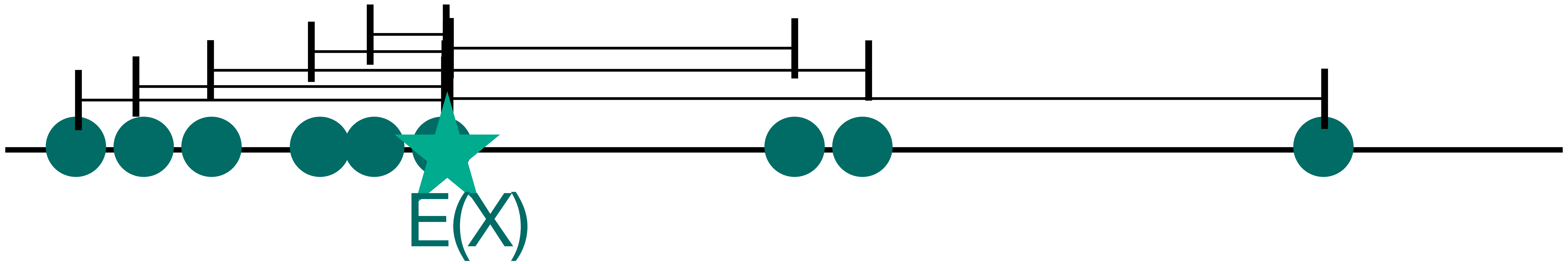
The distance from the **expected value** to the **true mean**



Variance

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

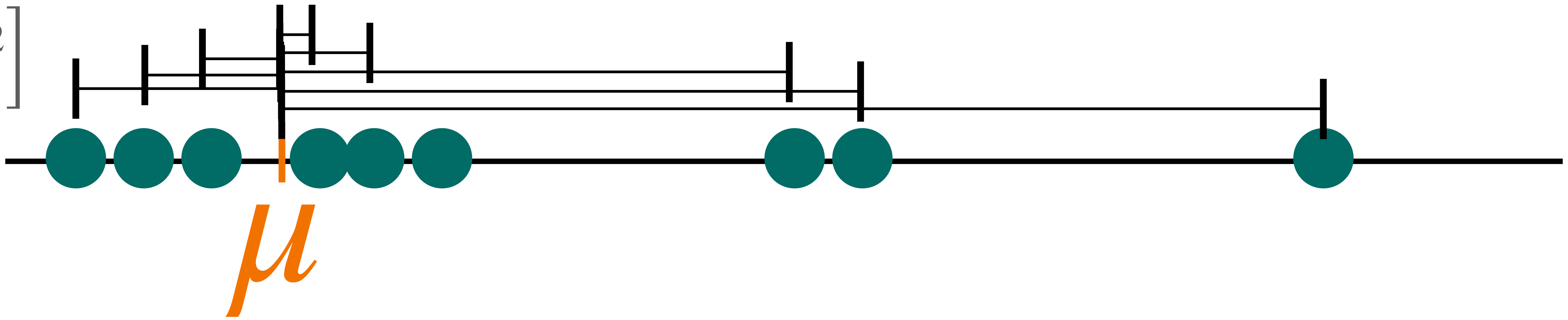
The average of the (squared) distances from the **individual sample means** to the **expected value**



MSE

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

The average of the (squared) distances from the **individual sample means** to the **true mean**



Mean Squared Error

Theorem: The Mean Squared Error (MSE) can be decomposed into the sum of the squared bias and the variance of $\hat{\theta}$:

$$MSE(\hat{\theta}) = Bias(\hat{\theta})^2 + Var(\hat{\theta})$$

Proof:

$$\begin{aligned} MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] &= E\left[(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2)\right] \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= Var(\hat{\theta}) + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\ &= Var(\hat{\theta}) + \left(E[\hat{\theta}] - \theta\right)^2 \\ &= Var(\hat{\theta}) + Bias(\hat{\theta})^2 \end{aligned}$$

(Since θ is not a random variable, $E[\theta] = \theta$)

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

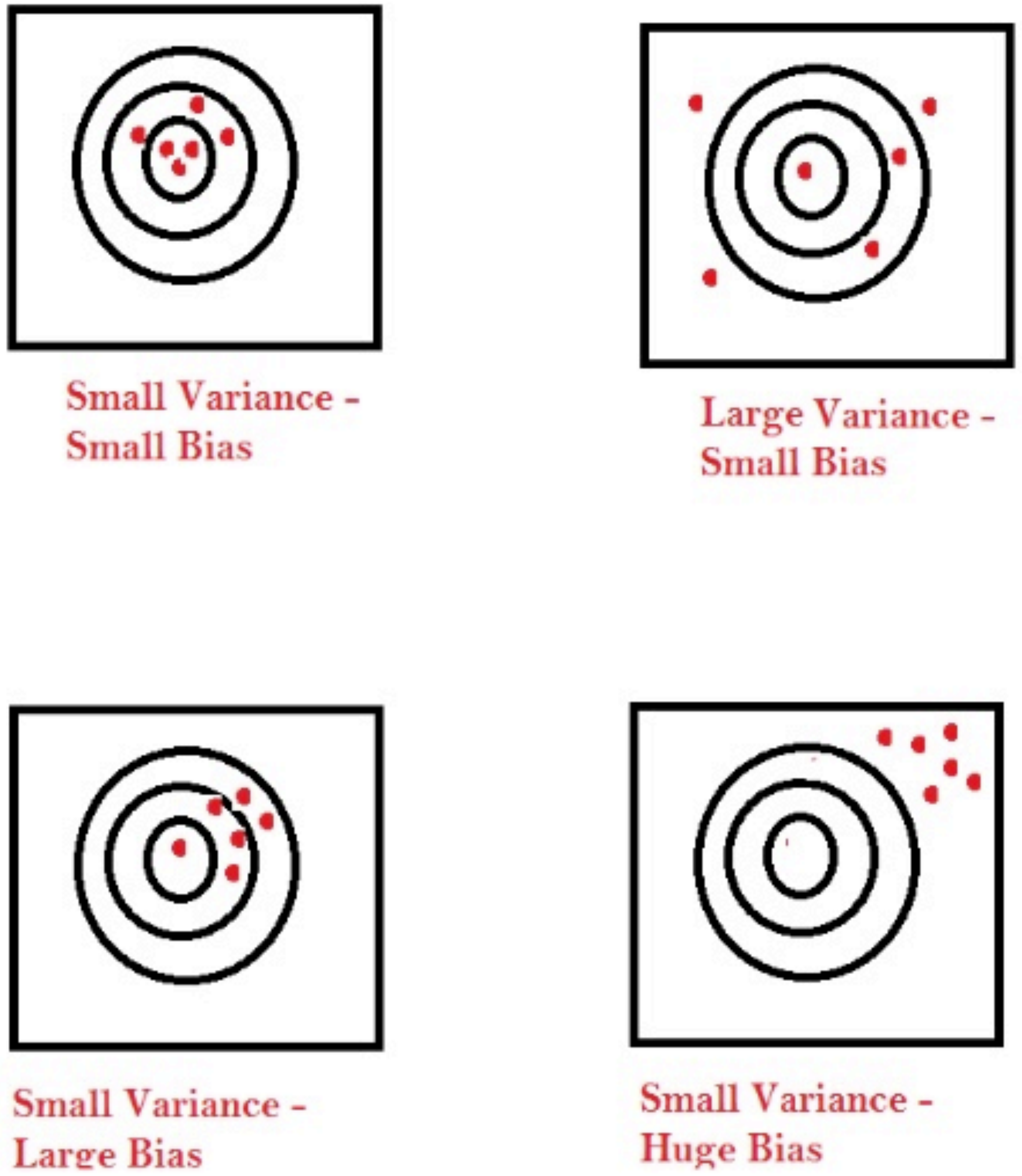
Note that this is true for all general parameter estimates $\hat{\theta}$

The bias-variance tradeoff

The bias-variance tradeoff

In general:

- Low variance (good) corresponds to high bias (bad)
- Low bias (good) corresponds to high variance (bad)



More Bias, Variance, and MSE examples

Bernoulli example

If $X_1, \ldots, X_n$ are an IID sample from a $Bernoulli(p)$ distribution, and we decide estimate p using $\hat{p} = \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$ $\quad (E(X_1) = p)$

What is the **bias** of $\hat{p}$?

$$\begin{aligned} Bias(\hat{p}) &= E(\hat{p}) - p \\ &= \frac{1}{n} E\left[\sum_i X_i\right] - p \\ &= \frac{n}{n} E\left[X_1\right] - p \\ &= p - p = 0 \end{aligned}$$

Bernoulli example

If $X_1, \ldots, X_n$ are an IID sample from a $Bernoulli(p)$ distribution, and we decide estimate p using $\hat{p} = \bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$

$(Var(X_1) = p(1-p))$

What is the **variance** of $\hat{p}$?

$$Var(\hat{p}) = Var\left(\frac{\sum_i X_i}{n}\right) = \frac{1}{n^2} Var\left(\sum_i X_i\right)$$

$$= \frac{n}{n^2} Var\left(X_1\right)$$

$$= \frac{p(1-p)}{n} \quad \rightarrow 0 \text{ as } n \rightarrow \infty$$

The problem…

There's a catch:

How can we compute the **bias of the estimator** when we don't know the value of the true parameter estimate?

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

We don't know what this is equal to

How can we compute the **Variance of the (mean) estimator** when we don't know the value of the population standard deviation?

$$Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

We don't know what this is equal to

Techniques for estimating bias, variance, and MSE from a single data sample:

1. Non-parametric bootstrap

Non-parametric bootstrap

Note that we are focusing on the mean parameter, but this could be any parameter of interest

While we usually *can't* draw another sample from the **original population**…

Population:

? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?

$\mu = ?$

We *can* draw samples from our ***sample***

Observed sample: (Sample mean: $\hat{\mu} = 3.3$)

6 1 4 9 2 0 1 3

Since we want our "new" samples to be the same size as our original sample, we must sample ***with replacement***

6 6 4 2 2 0 1 3
Bootstrap sample mean:
$\hat{\mu}^*_1 = 3$

1 1 4 4 9 0 0 1
Bootstrap sample mean:
$\hat{\mu}^*_2 = 2.5$

6 1 9 9 2 1 1 3
Bootstrap sample mean:
$\hat{\mu}^*_3 = 4$

Estimating bias of an estimator using non-parametric bootstrap

Observed sample:

6 1 4 9 2 0 1 3

6 6 4 2 2 0 1 3

Bootstrap sample mean:

$\hat{\mu}_1^* = 3$

1 1 4 4 9 0 0 1

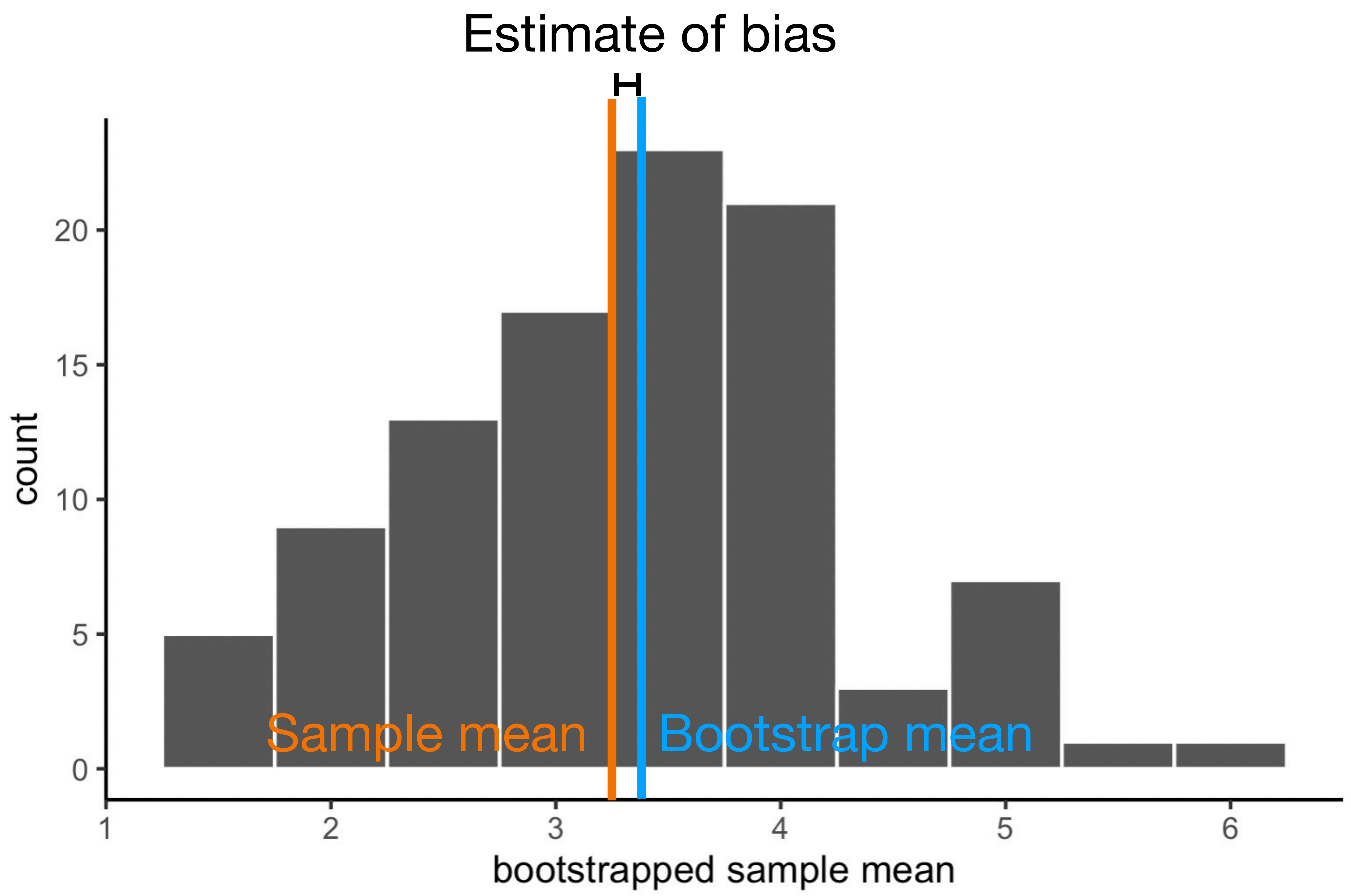
Bootstrap sample mean:

$\hat{\mu}_2^* = 2.5$

6 1 9 9 2 1 1 3

Bootstrap sample mean:

$\hat{\mu}_3^* = 4$



Idea:

- treat the original sample as the population
- treat the bootstrapped sample as the sample

Use these to estimate the bias

NP Bootstrap bias estimate:

$$\widehat{Bias}(\hat{\mu}) \approx \left(\frac{1}{N}\sum_{k=1}^{N} \hat{\mu}_k^*\right) - \hat{\mu}$$

Where N is the number of bootstrap samples

(N = 3) in this example

Estimating <u>bias</u> of an estimator using non-parametric bootstrap

Observed sample:

6 1 4 9 2 0 1 3

(Sample mean: $\hat{\mu} = 3.3$)

6 6 4 2 2 0 1 3

Bootstrap sample mean:

$\hat{\mu}_1^* = 3$

1 1 4 4 9 0 0 1

Bootstrap sample mean:

$\hat{\mu}_2^* = 2.5$

6 1 9 9 2 1 1 3

Bootstrap sample mean:

$\hat{\mu}_3^* = 4$

NP Bootstrap bias estimate:

$$\widehat{Bias}(\hat{\mu}) \approx \left(\frac{1}{N}\sum_{k=1}^{N} \hat{\mu}_k^*\right) - \hat{\mu}$$

For this example,

$$\widehat{Bias}(\hat{\mu}) = \frac{3 + 2.5 + 4}{3} - 3.3$$
$$= -0.133$$

See bootstrap_mean.R for code for this example

Estimating <u>variance</u> of an estimator using non-parametric bootstrap

Observed sample:

6 1 4 9 2 0 1 3

6 6 4 2 2 0 1 3

Bootstrap sample mean:

$\hat{\mu}_1^* = 3$

1 1 4 4 9 0 0 1

Bootstrap sample mean:

$\hat{\mu}_2^* = 2.5$

6 1 9 9 2 1 1 3

Bootstrap sample mean:

$\hat{\mu}_3^* = 4$

NP Bootstrap variance estimate:

$$\widehat{Var}\,(\hat{\mu}) \approx = \frac{1}{N}\sum_{k=1}^{N}(\hat{\mu}_k^* - \overline{\hat{\mu}^*})^2$$

Where $\overline{\hat{\mu}^*} = \frac{1}{N}\sum_{k=1}^{N}\hat{\mu}_k^*$

Idea:

- treat the sample as the population
- treat the bootstrapped sample as the sample

Use these to estimate the variance

Estimating <u>bias</u> of an estimator using non-parametric bootstrap

Observed sample:

6 1 4 9 2 0 1 3

(Sample mean: $\hat{\mu} = 3.3$)

6 6 4 2 2 0 1 3

Bootstrap sample mean:

$\hat{\mu}_1^* = 3$

1 1 4 4 9 0 0 1

Bootstrap sample mean:

$\hat{\mu}_2^* = 2.5$

6 1 9 9 2 1 1 3

Bootstrap sample mean:

$\hat{\mu}_3^* = 4$

NP Bootstrap variance estimate:

$$\widehat{Var}(\hat{\mu}) \approx = \frac{1}{N} \sum_{k=1}^{N} (\hat{\mu}_k^* - \overline{\hat{\mu}^*})^2$$

For this example,

$$\overline{\hat{\mu}^*} = \frac{3 + 2.5 + 4}{3} = 3.167$$

$$\widehat{Variance}(\hat{\mu}) = \frac{(3 - 3.167)^2 + (2.5 - 3.167)^2 + (4 - 3.167)^2}{3}$$

$$= 0.389$$

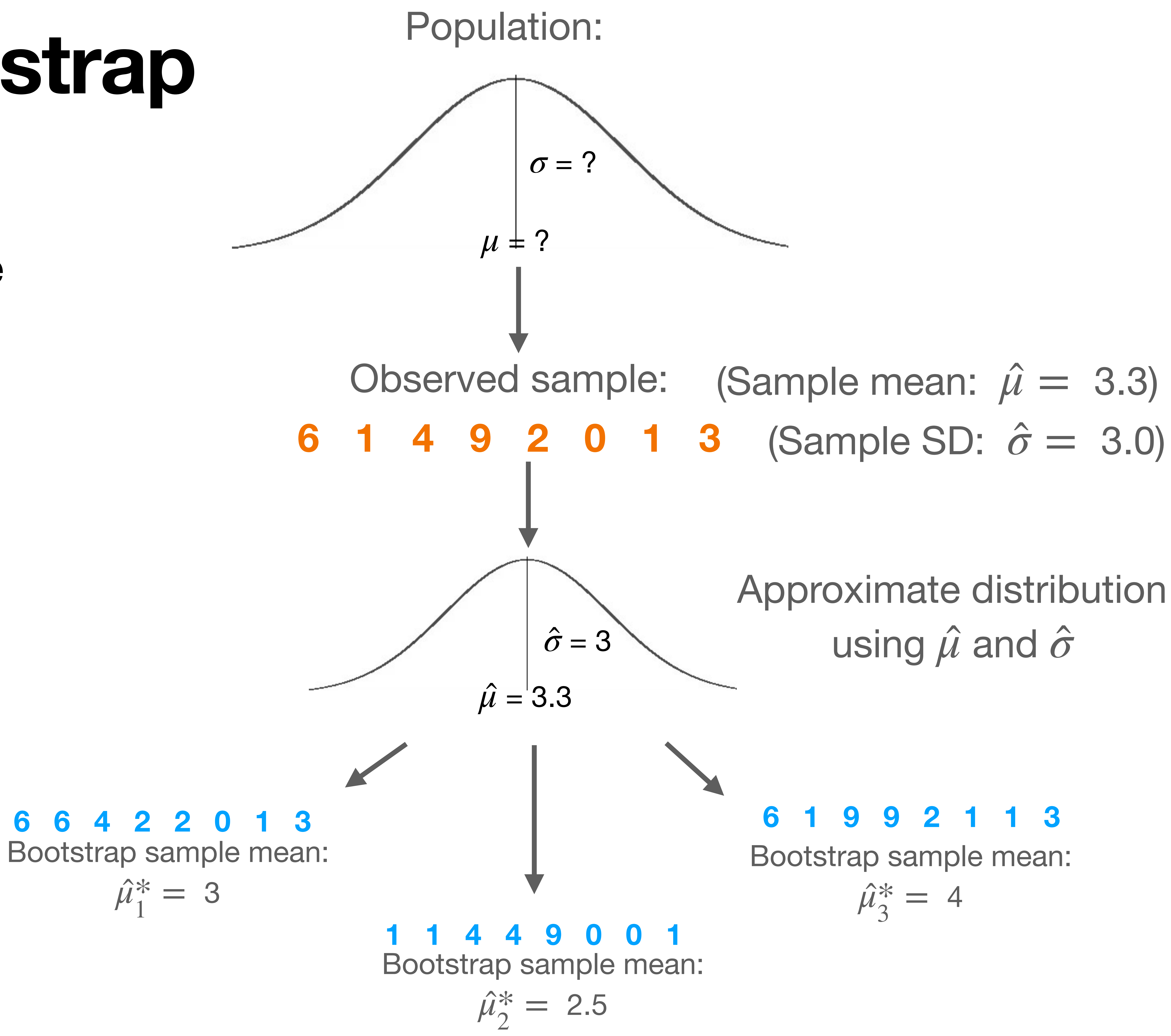
Techniques for estimating bias, variance, and MSE from a single data sample:

1. Non-parametric bootstrap
2. Parametric bootstrap

Parametric bootstrap

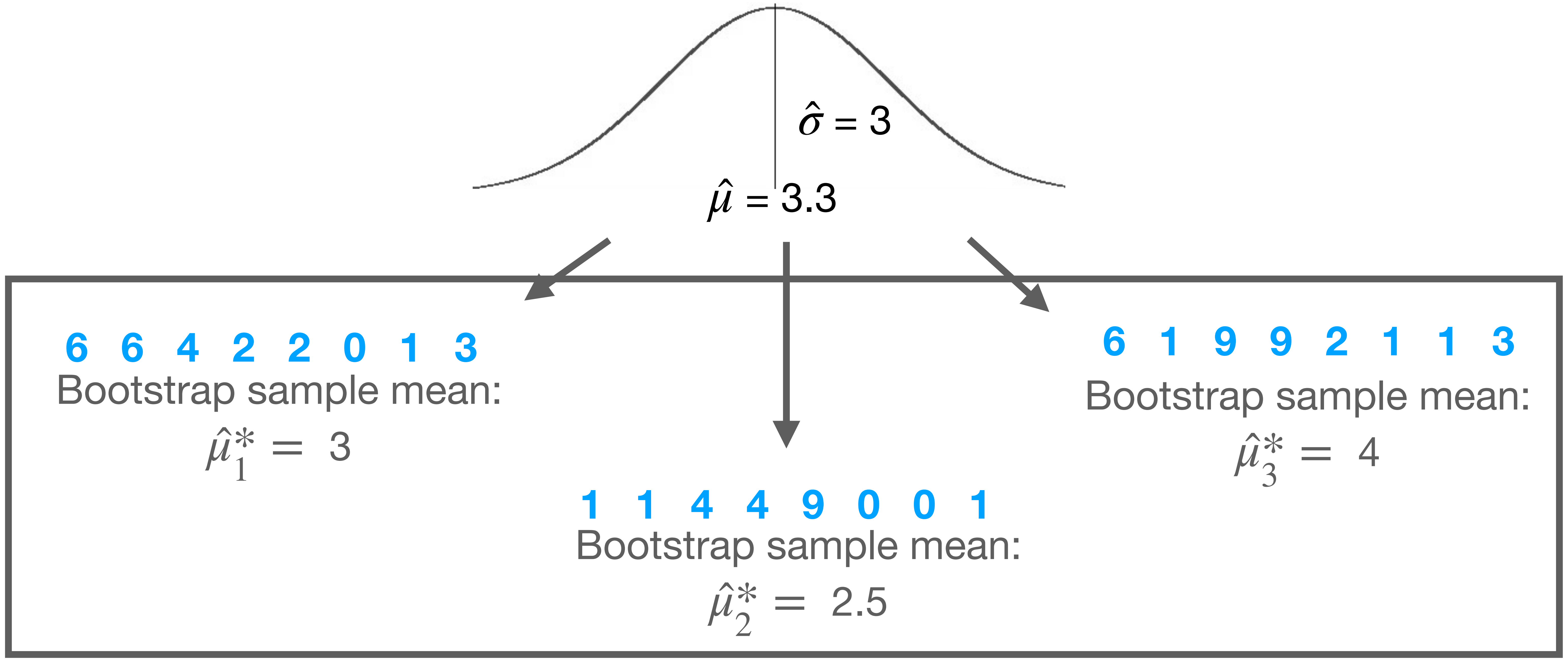
If you know that your data came from a particular distribution, e.g., a normal distribution

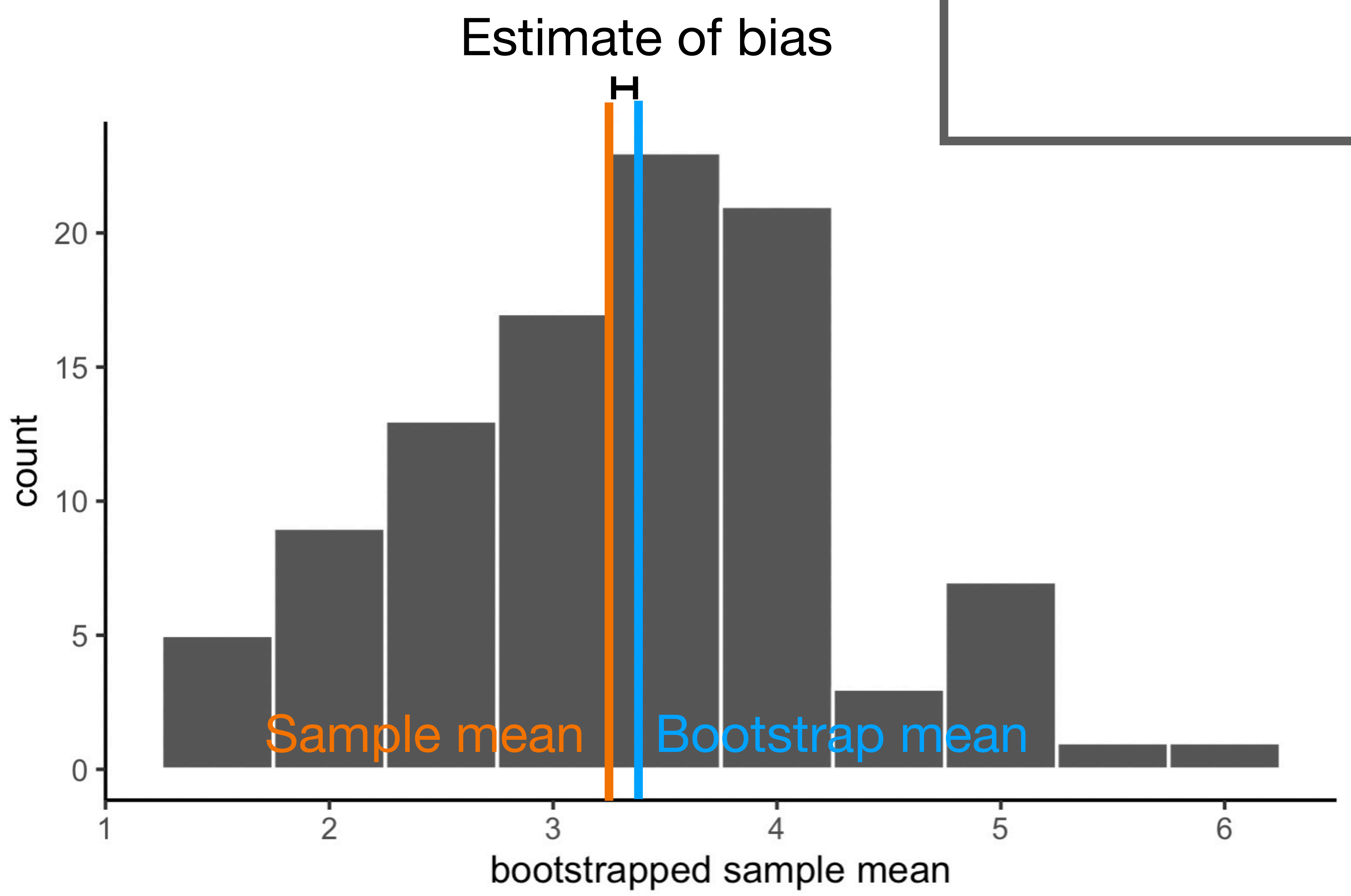
Then use the distribution with the estimated parameters to *draw* many "parametric bootstrap" samples



Estimating <u>bias</u> of an estimator using parametric bootstrap

Once you have the bootstrap parameter estimates, estimating the bias and variance is then the same as in the non-parametric bootstrap case





Parametric bootstrap bias estimate:

$$\widehat{Bias}(\hat{\mu}) \approx \frac{1}{N}\sum_{k=1}^{N} \hat{\mu}_k^* - \hat{\mu}$$

See bootstrap_mean.R for code for this example

Asymptotic properties of the sample mean

The Law of Large Numbers (LLN)

Theorem: if $X_1, X_2, \ldots, X_n$ is an IID sample, then:

$$\overline{X} \xrightarrow{P} E(X_1), \quad \text{as } n \to \infty$$

The Law of Large Numbers (LLN)

Theorem: if $X_1, X_2, \ldots, X_n$ is an IID sample, from a population with mean μ then:

$$\overline{X} \xrightarrow{P} \mu, \quad \text{as } n \to \infty$$

Random **Not random!**

Note: We had already shown that $E(\bar{X}) = E[X_1] = \mu$. This statement is stronger.

This essentially means that when n is really, really large, the sample mean is no longer random.

E.g., if we sample enough people, our results will become more and more accurate (not just in expectation!)

The Central Limit Theorem (CLT)

Theorem: if $X_1, X_2, \ldots, X_n$ is an IID sample from a population with mean μ and standard deviation σ, then:

$$\overline{X} \xrightarrow{D} N\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{as } n \to \infty$$

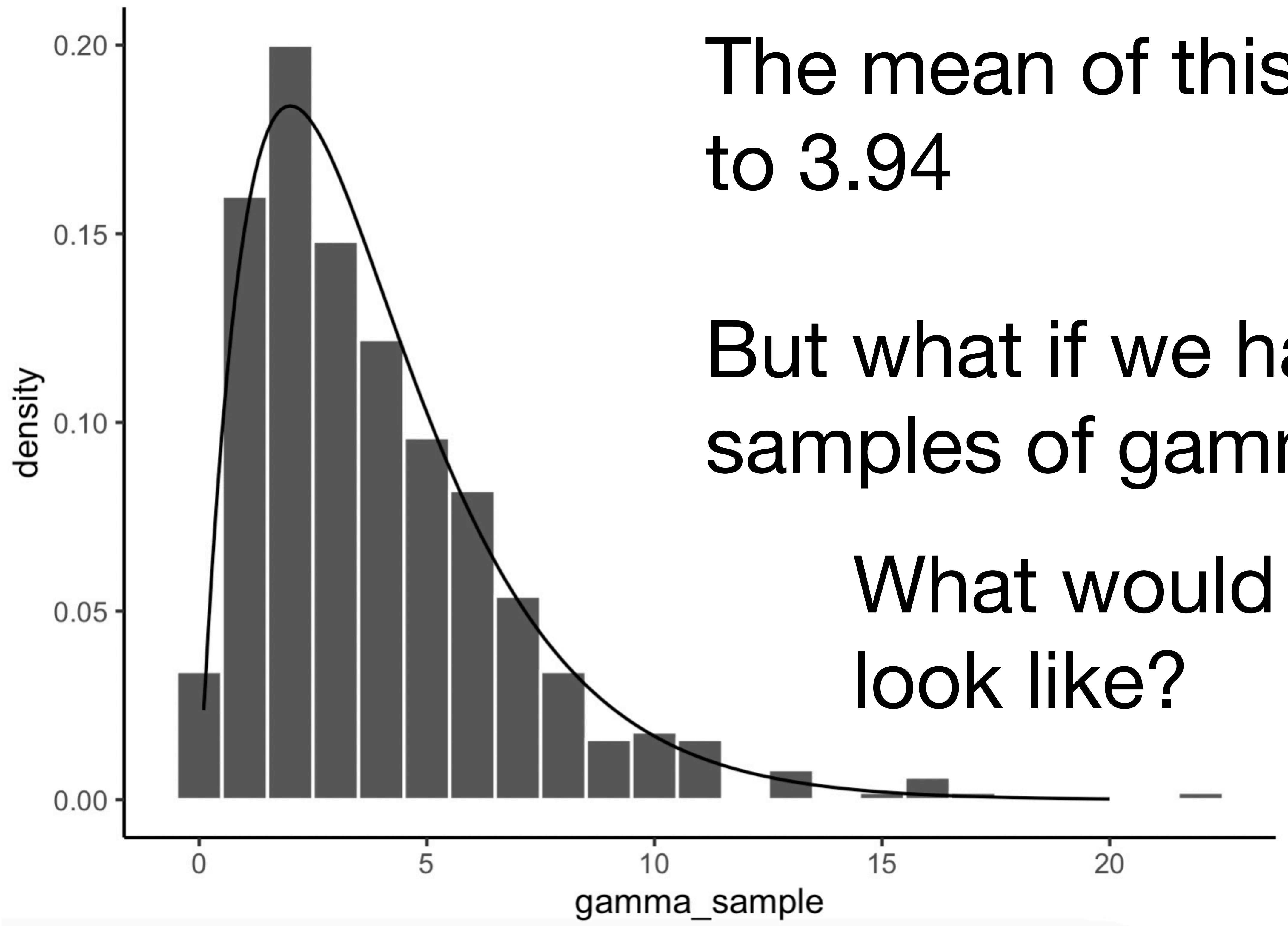
Regardless of the original distribution of the X_i.
("as the sample size increases, the sample mean behaves as if it is from a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$)

Note: We had already shown that $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$, but the CLT also says that the *distribution* of $\overline{X}$ is *Normal* when the sample size is large enough

Empirical CLT example

If $X_1, X_2, \ldots, X_n$ correspond to an IID sample from a *Gamma*($\alpha =$ *2*, $\beta =$ *0.5*) distribution.

The distribution of a *single sample* of size n = 500 looks like:



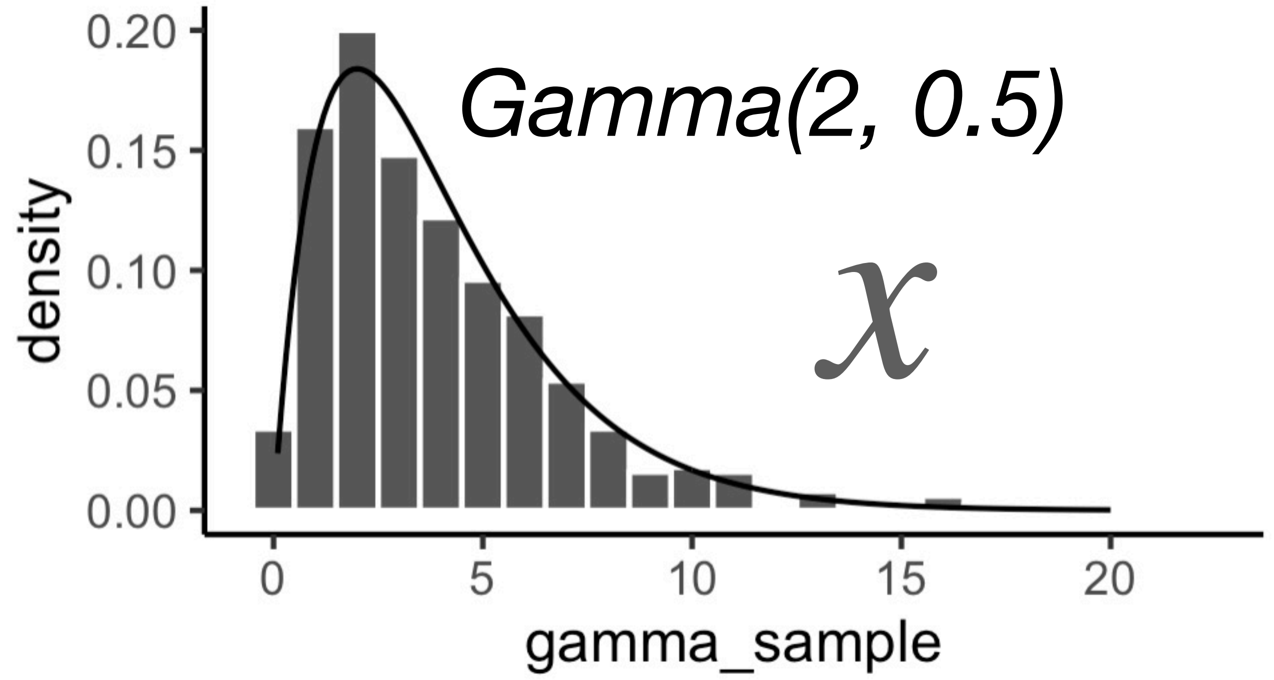
The mean of this sample specific sample is equal to 3.94

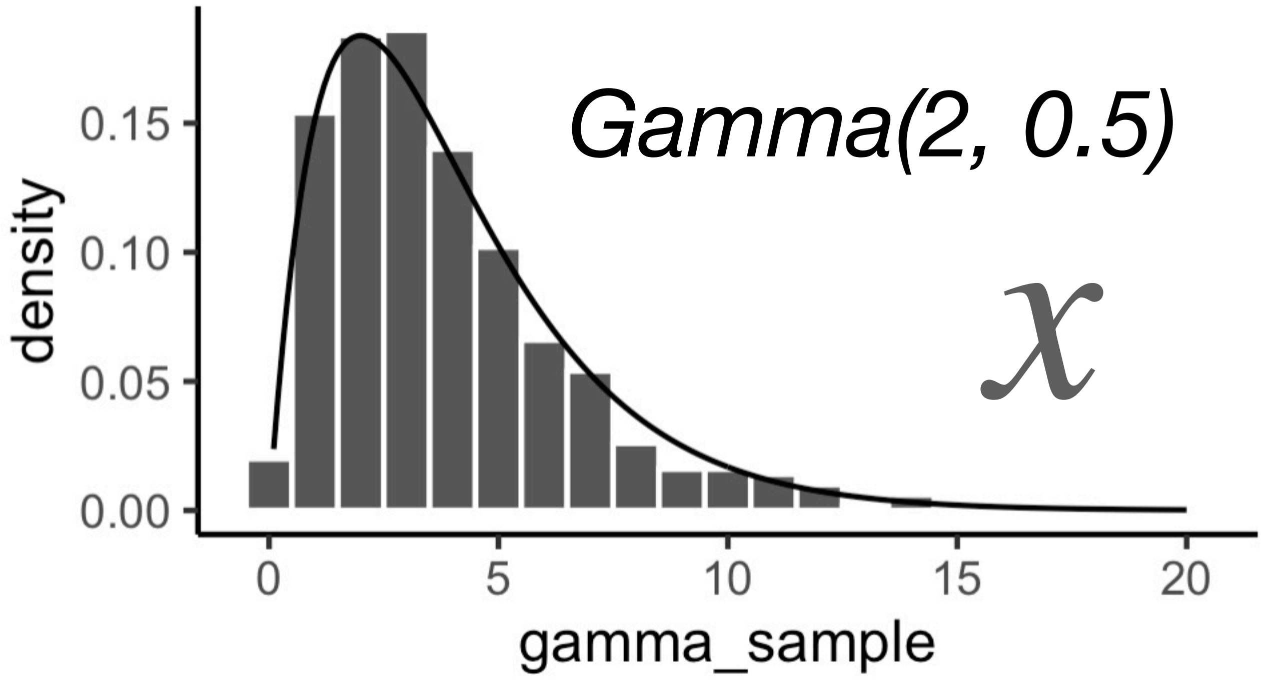
But what if we had produced many different samples of gamma-distributed numbers?

What would the distribution of the sample *means* look like?

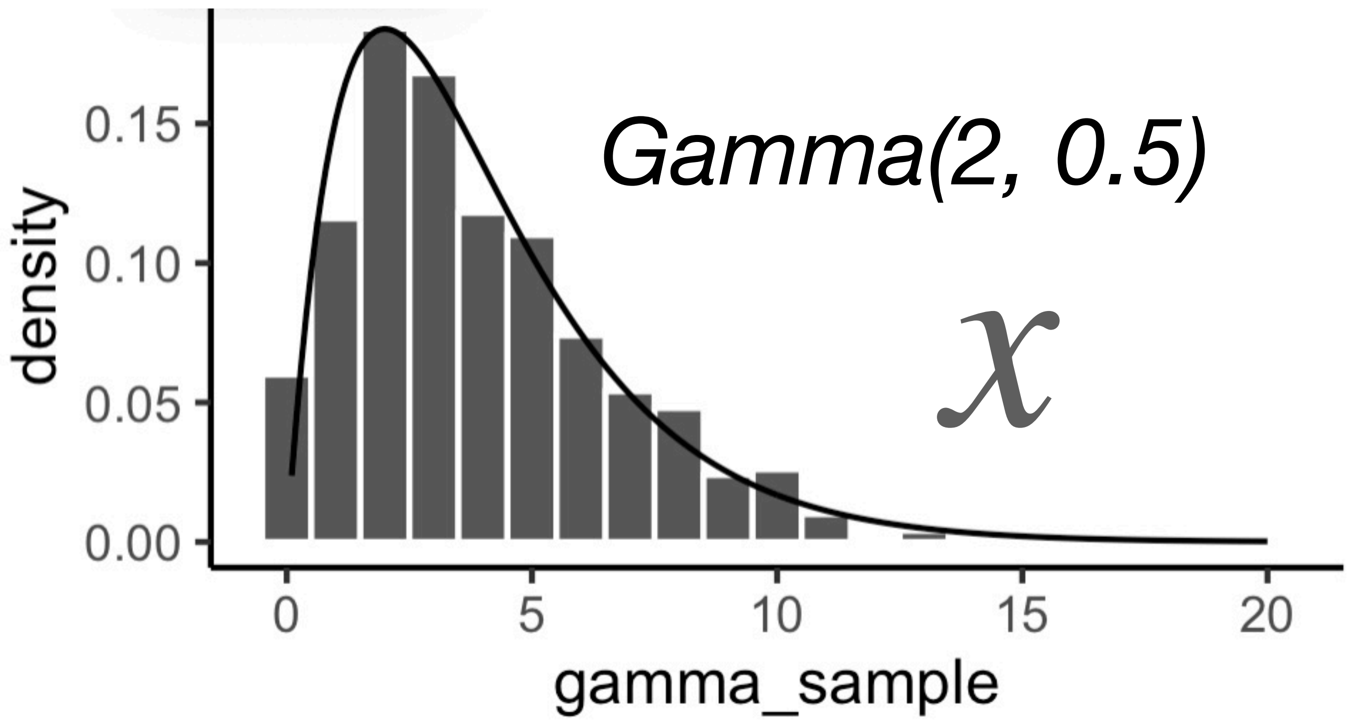
Empirical CLT example

Sample 1 Sample 2 … Sample 1000





…



Sample mean: 3.94 Sample mean: 3.89 … Sample mean: 3.92

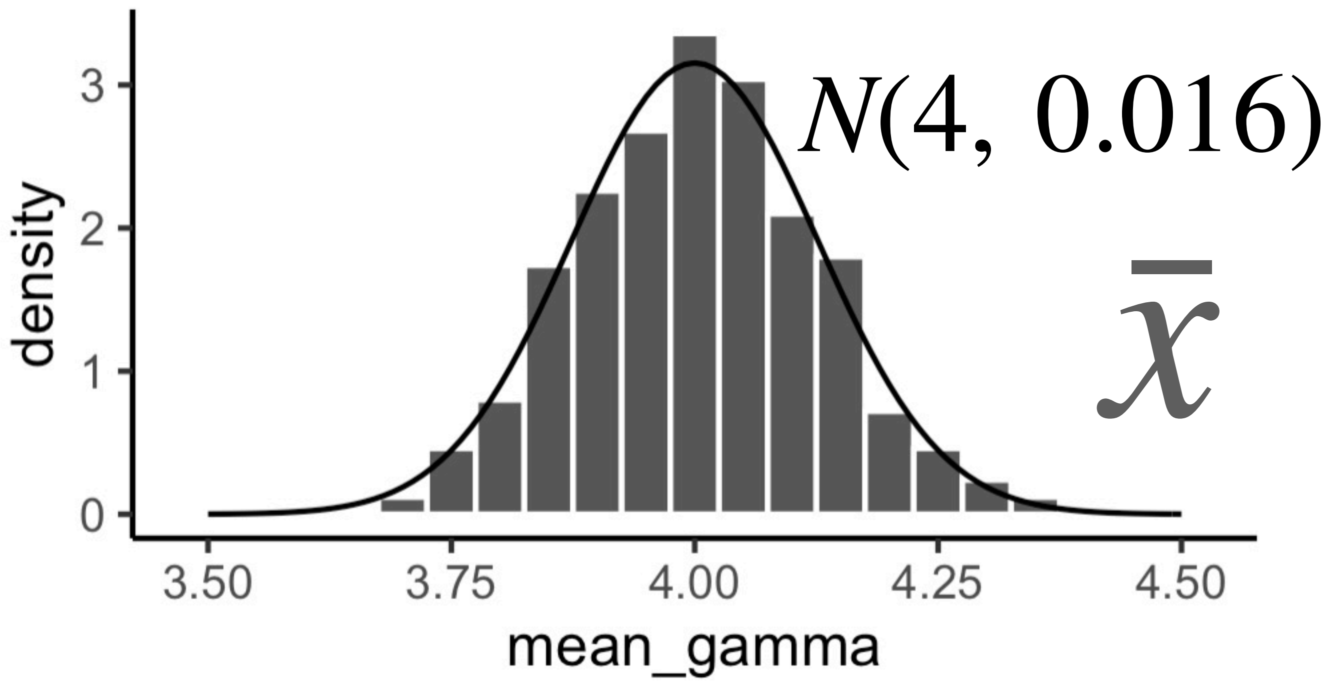
Since:

$\mu = E(X) = \frac{\alpha}{\beta}$, and

$\sigma^2 = Var(X_i) = \frac{\alpha}{\beta^2}$

The CLT implies that:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to N\left(\frac{\alpha}{\beta}, \frac{\alpha}{n\beta^2}\right), \quad \text{as } n \to \infty$$



Even though the original data itself isn't Normal, the distribution of the sample means is!

See clt.R for code for this example

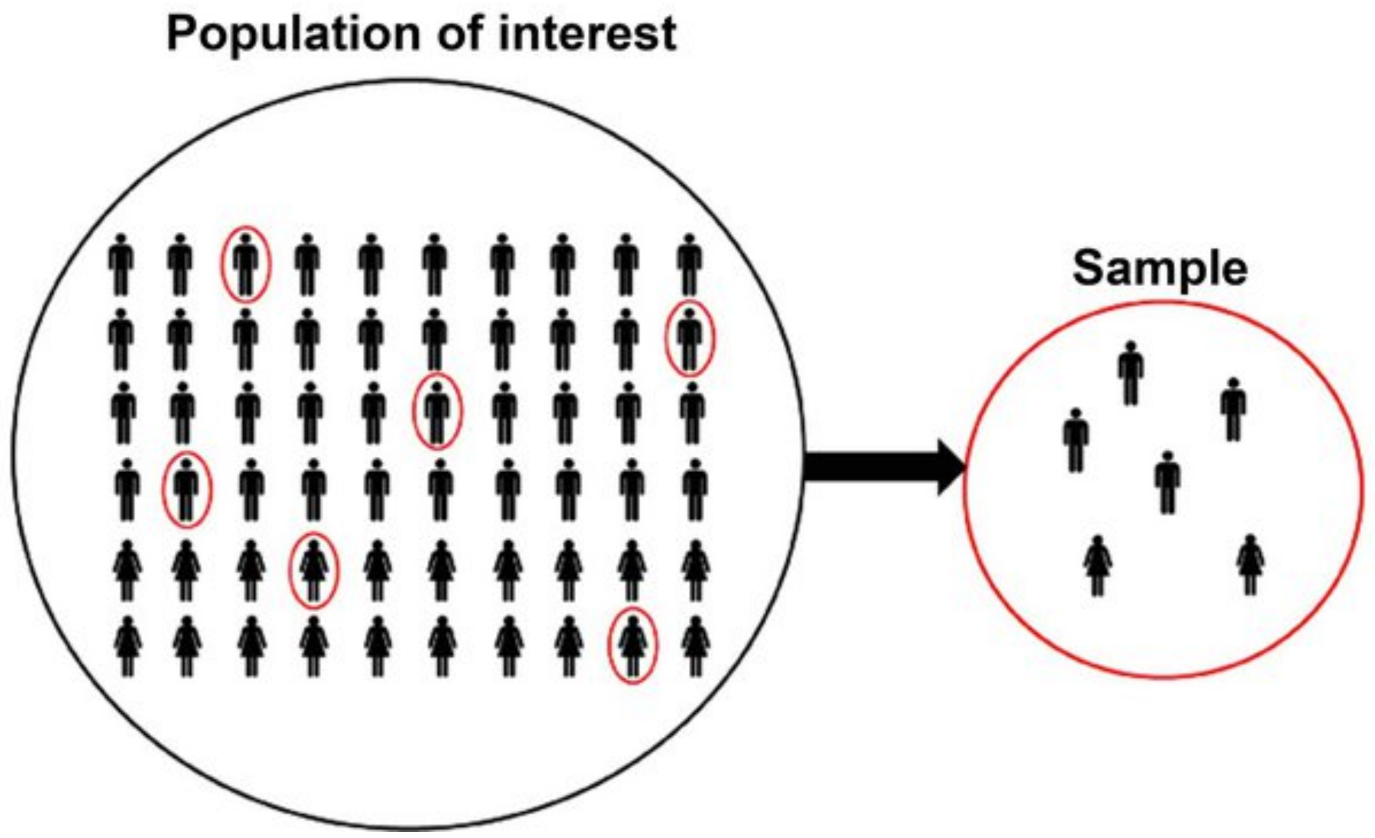
Recap

What we’ve covered so far

A **parameter** is an unknown feature of a population (such as a mean or proportion).

Data that comes from the population can be used to generate **estimates** of the parameters if the data is ***representative of the population***.

Random samples tend to be representative of the population



Random variables represent the hypothetical versions of the random samples that can be drawn from the population

What we've covered so far

An estimate of a parameter can be assessed in terms of its **bias, variance and MSE**

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

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

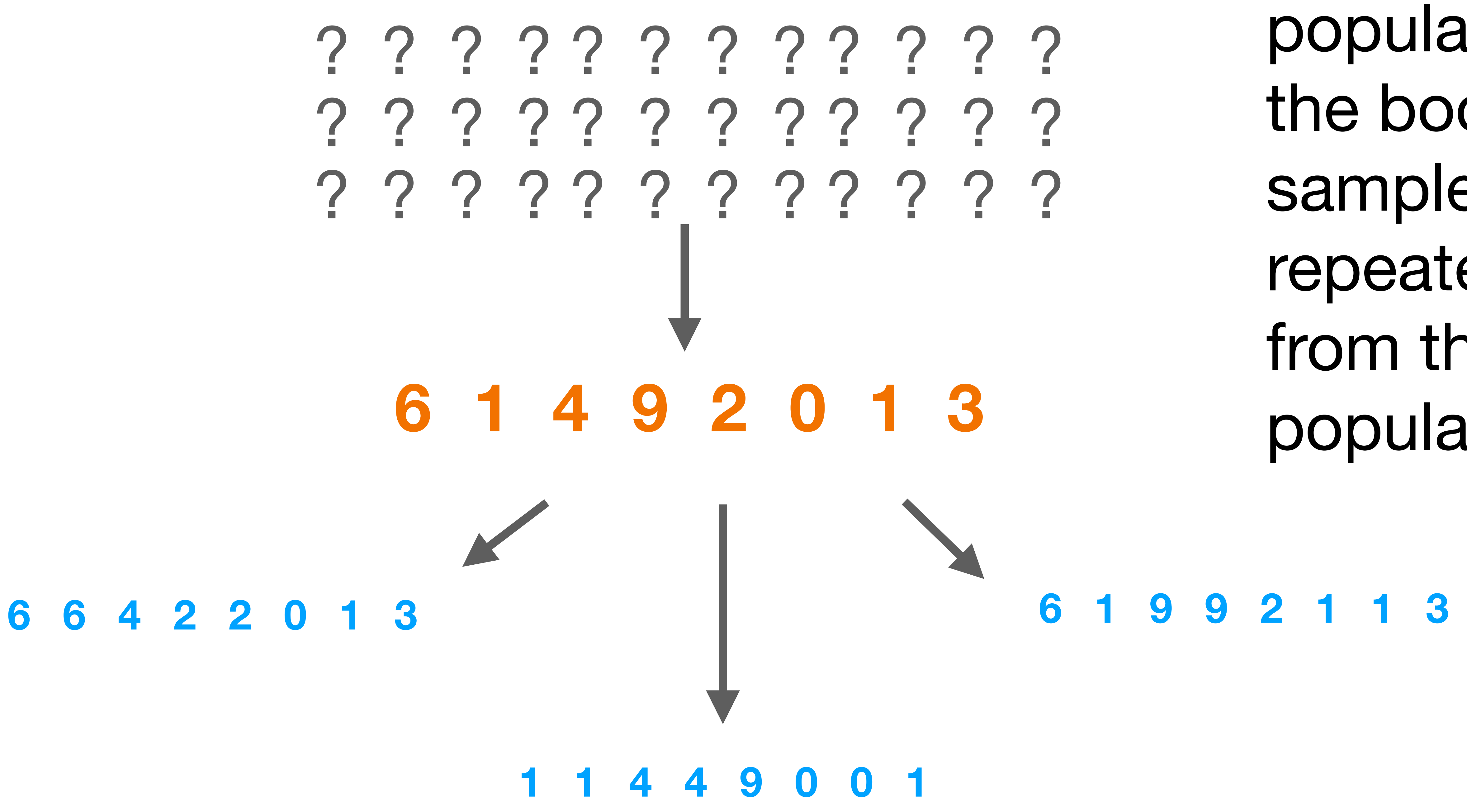
$$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = Bias(\hat{\theta})^2 + Var(\hat{\theta})$$

But these quantities cannot be estimated from the data alone because they require knowledge of the original population parameters

What we've covered so far

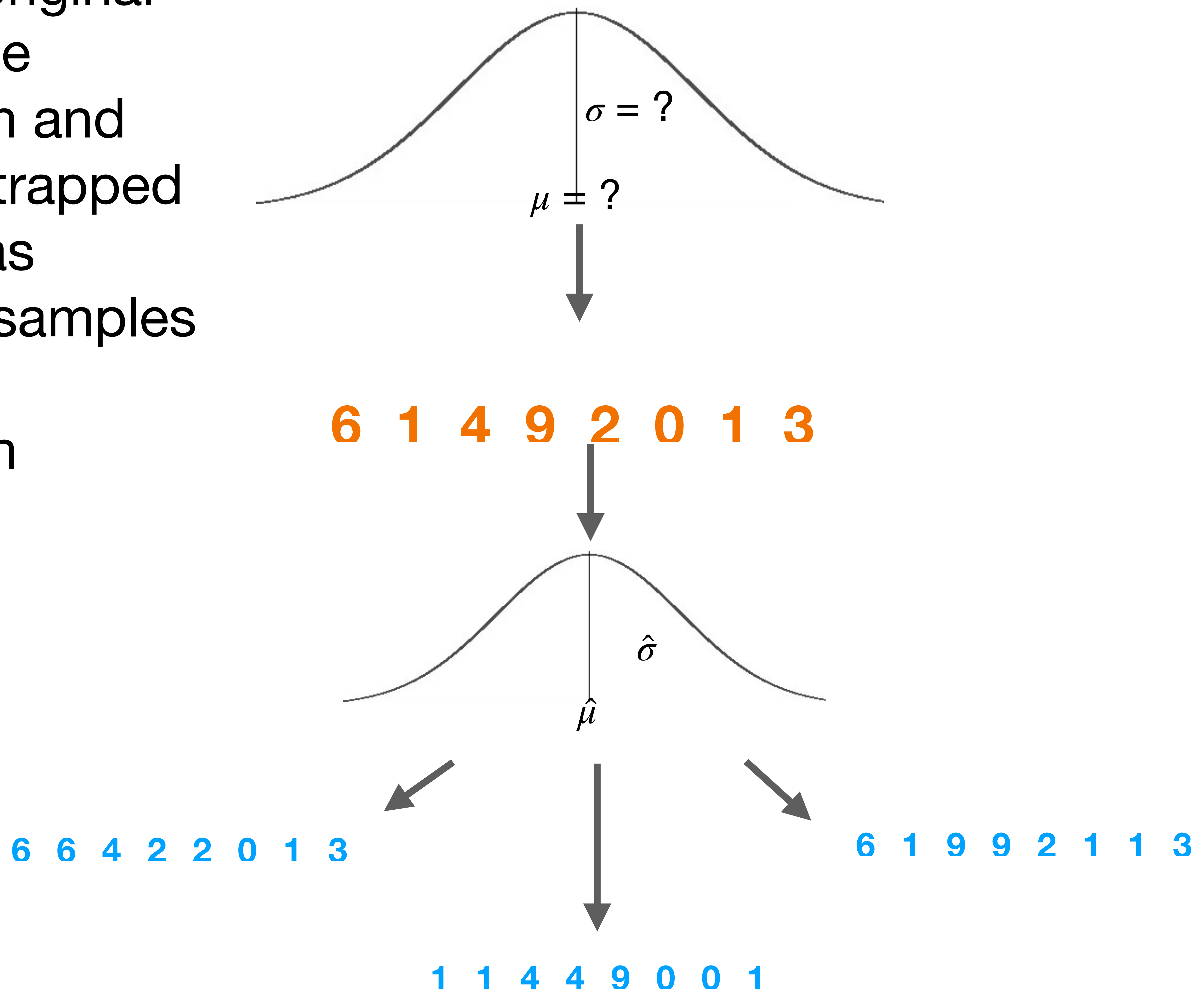
We can estimate the **bias, variance and MSE** using a bootstrapping technique, which can be:

Non-parametric



The idea is to treat the original data as the population and the bootstrapped samples as repeated samples from the population

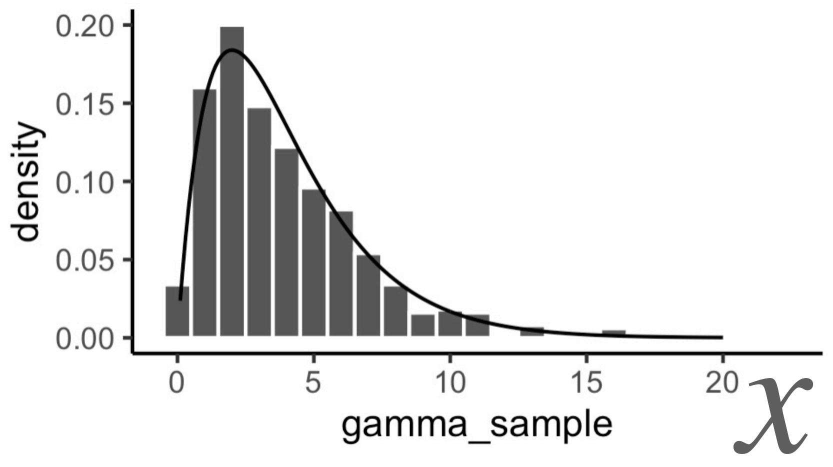
Parametric

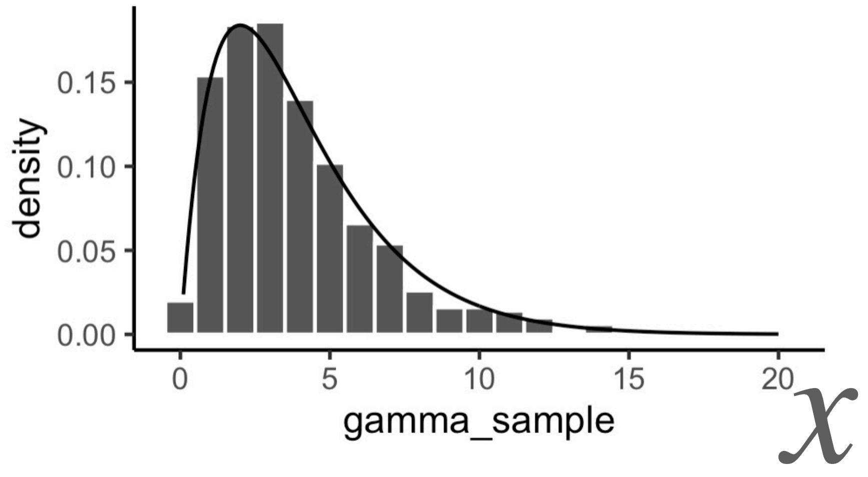


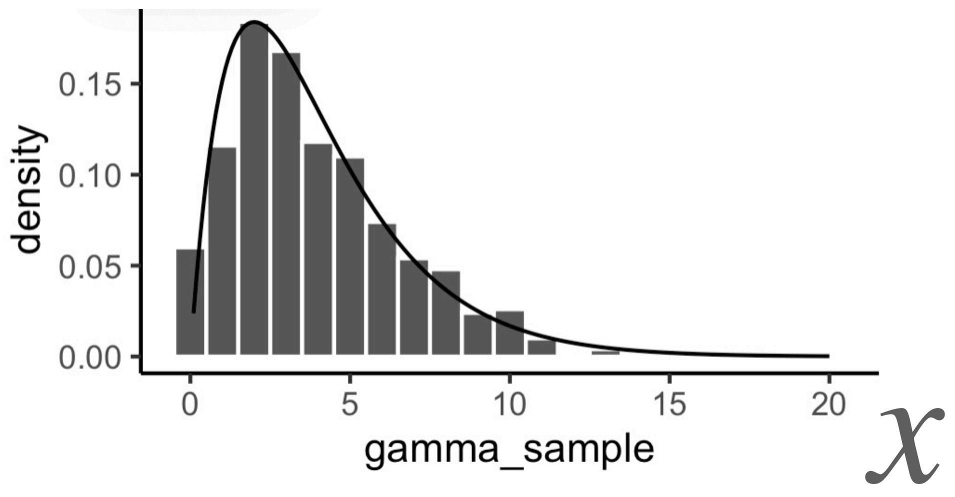
What we’ve covered so far

The **Central Limit Theorem** tells us that if our data are IID, the sample mean tends to wards a normal distribution whose mean is equal to the population mean and whose variance is equal to the population variance divided by n (sample size)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to N\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{as } n \to \infty$$







Sample mean: 3.94

Sample mean: 3.89

Sample mean: 3.92