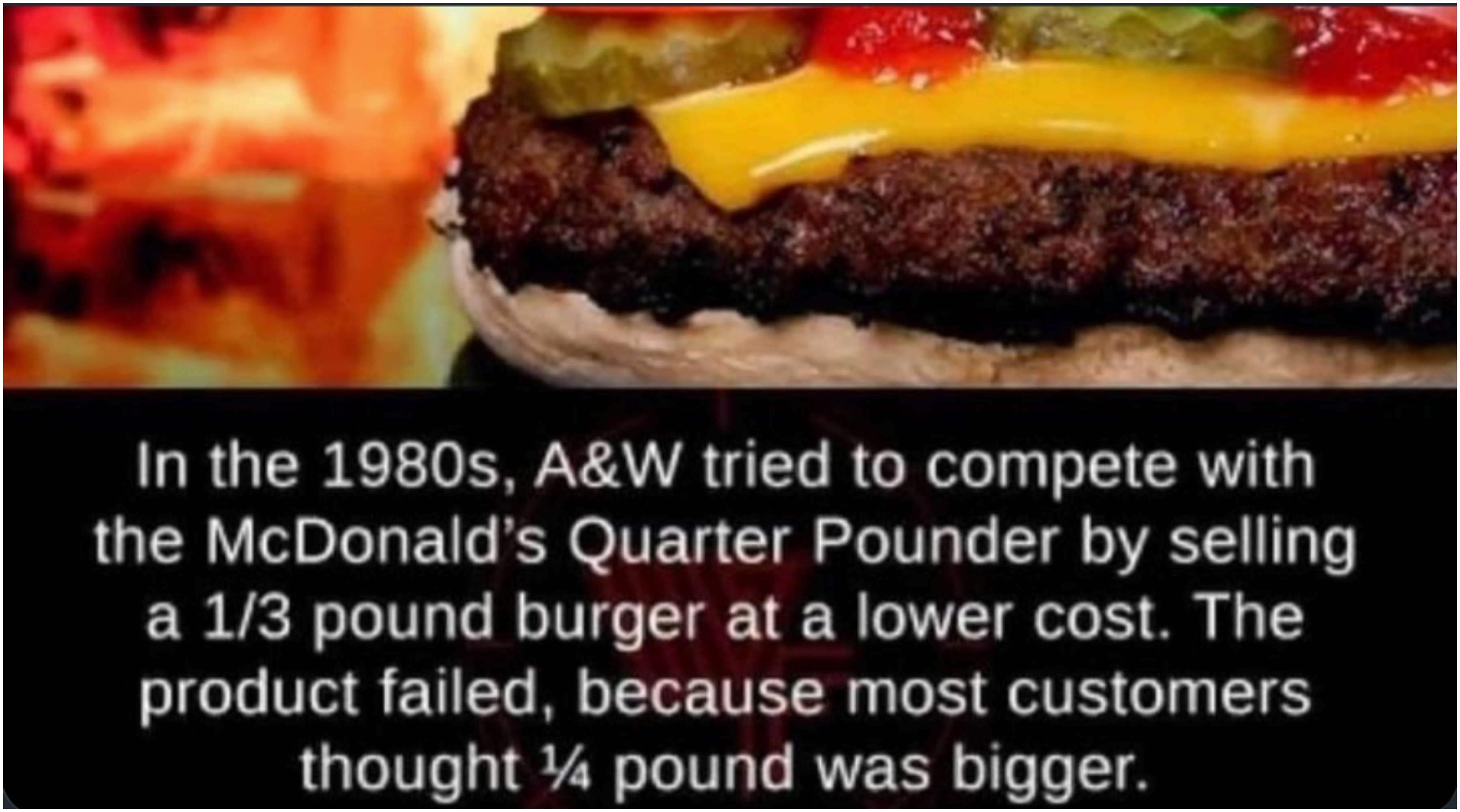
A snap demonstration of why numeracy is key to society



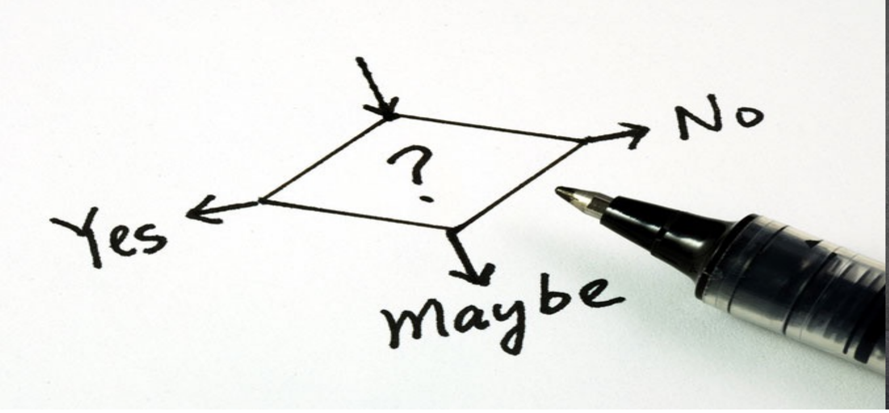
In the 1980s, A&W tried to compete with the McDonald's Quarter Pounder by selling a 1/3 pound burger at a lower cost. The product failed, because most customers thought ¼ pound was bigger.

Lecture 8: Estimating with uncertainty, but with a degree of certainty (i.e., with some confidence).

Statistics is the science of aiding decision-making with incomplete information

"While nothing is more uncertain than a single life, nothing is more certain than the average duration of a thousand lives"

Elizur Wright (mathematician & “the father of life insurance”)



Statistics is the study of uncertainty

Statistics - like life itself - is all about making big conclusions from (small) samples.

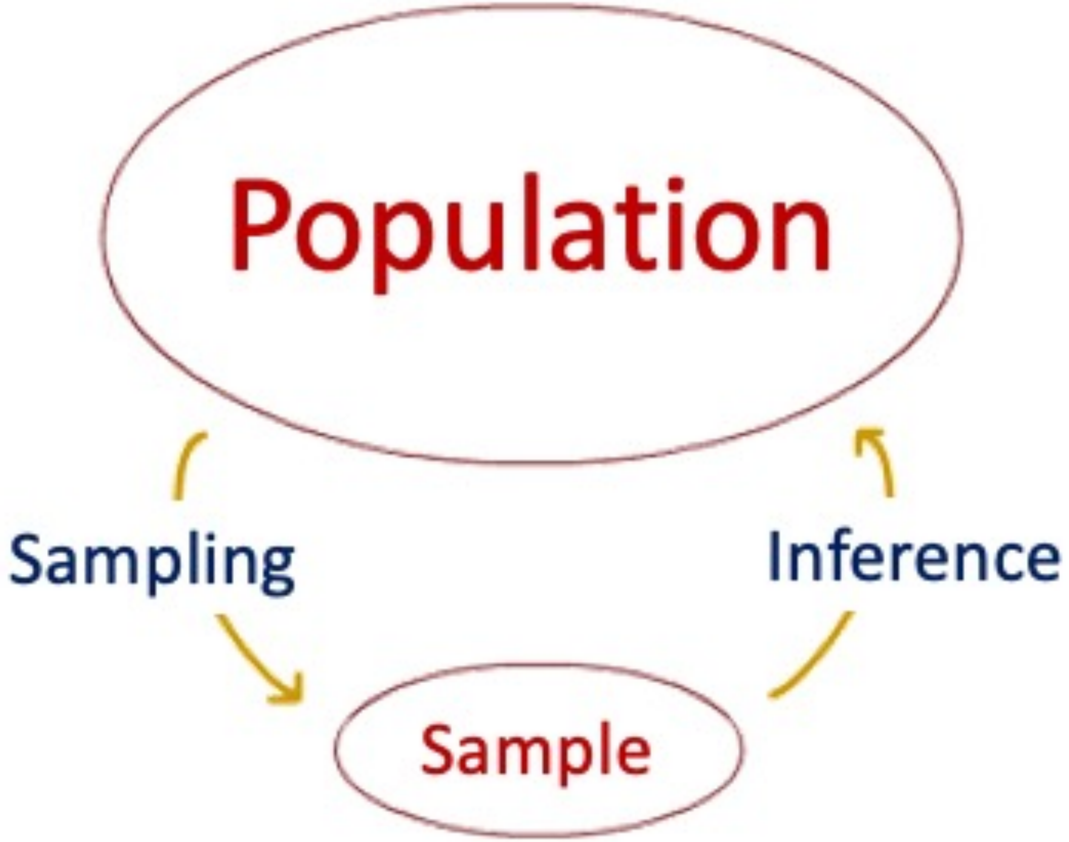
One primary goal of statistics is to estimate (infer) an unknown quantity (parameter) of a population based on sample data.

Estimation involves inferring a population parameter (e.g., mean, standard deviation, median) from a sample.

We use estimates to make decisions. Statistics is fundamentally the science of making decisions with incomplete knowledge, often using samples from populations of unknown sizes.

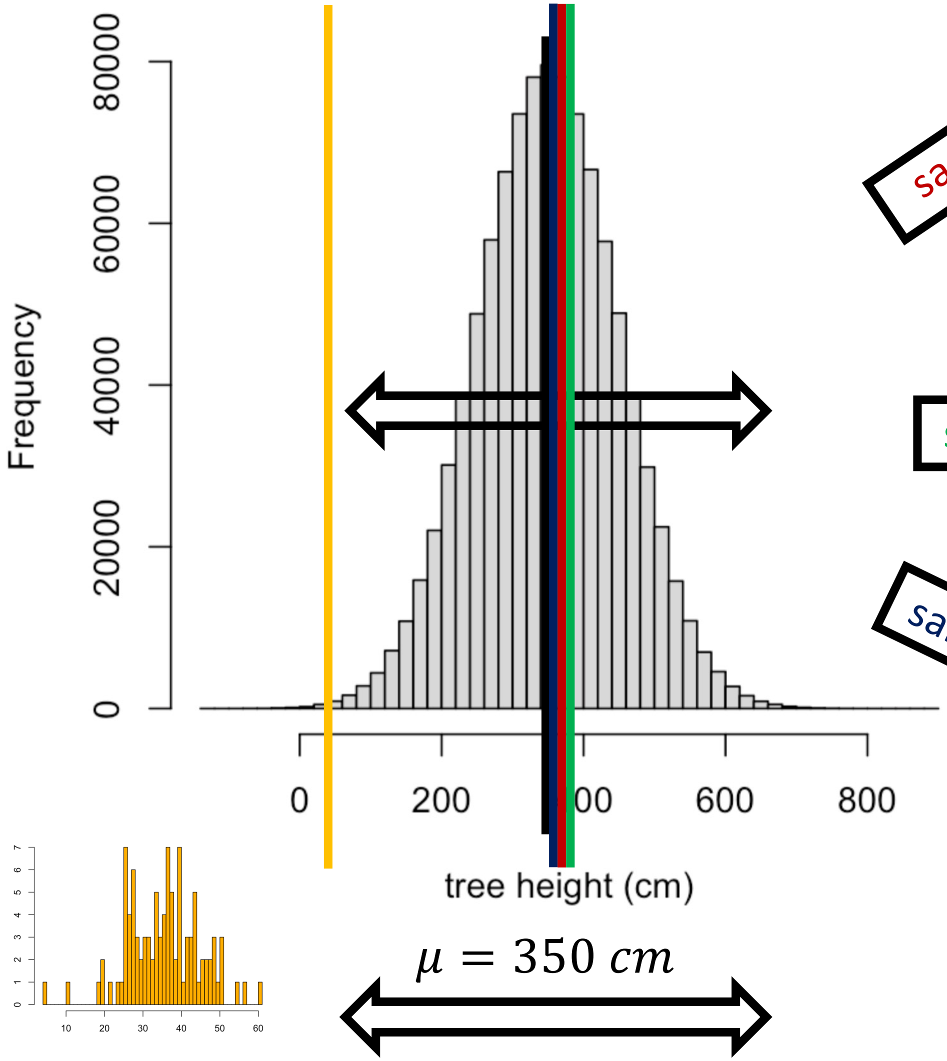
However, sample-based statistics (e.g., mean, median, standard deviation) vary from one sample to another. This variation introduces uncertainty, known as sampling variation.

How to estimate with uncertainty, but with some degree of certainty (i.e., with some confidence)?



Sampling variation generates uncertainty

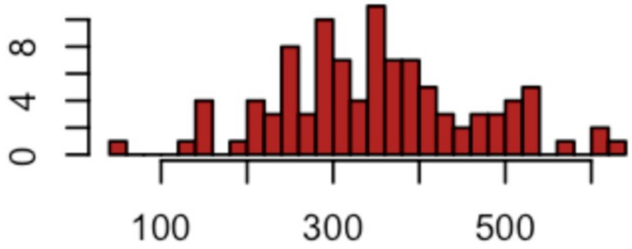
$\mu = 350\ cm;\ \sigma = 100\ cm$



$\mu = 350\ cm$

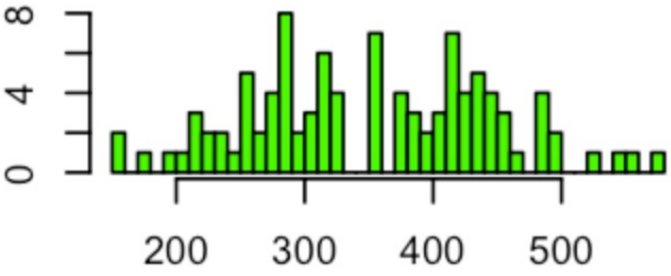
sampling

$\bar{\mathbf{X}} = \mathbf{351.5}\ \boldsymbol{cm};\ \boldsymbol{s} = \mathbf{114.2}\ \boldsymbol{cm}$



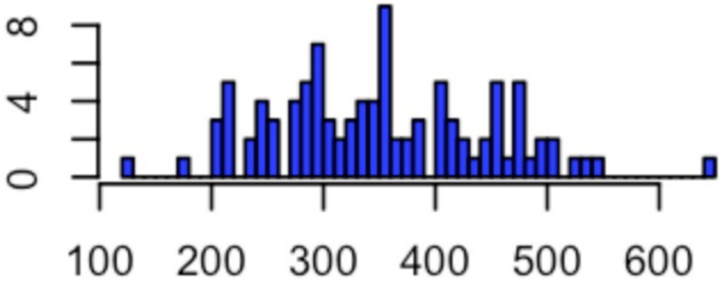
sampling

$\bar{\mathbf{X}} = \mathbf{352.3}\ \boldsymbol{cm};\ \boldsymbol{s} = \mathbf{94.0}\ \boldsymbol{cm}$



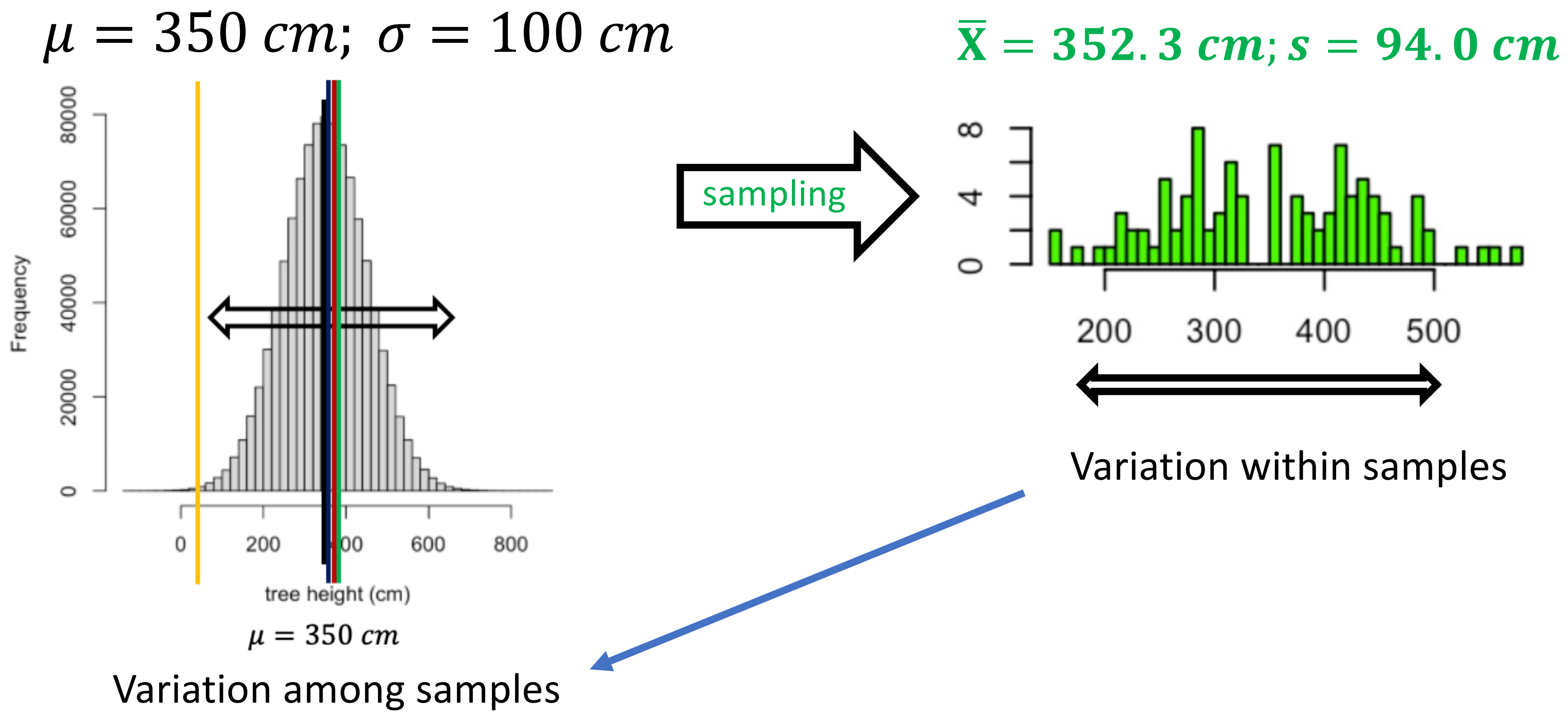
sampling

$\bar{\mathbf{X}} = \mathbf{351.4}\ \boldsymbol{cm};\ \boldsymbol{s} = \mathbf{96.6}\ \boldsymbol{cm}$



Uncertainty (samples means vary around the true population mean)

The variation within a sample (measured by the standard deviation) gives us insight into how much sample means (averages) might differ from the true population mean (average)—essentially estimating how far off we might be



$\mu = 350\ cm;\ \sigma = 100\ cm$

sampling

$\bar{\mathrm{X}} = 352.3\ cm; s = 94.0\ cm$

$\mu = 350\ cm$

Variation within samples

Variation among samples

Statistical 'superpower': Variation within samples (among observations) allows us to estimate just how confident we are in our uncertainty (the variation among sample means).

Population parameters versus sample estimates

A parameter describes a quantity in a statistical population, while an estimate (or statistic) is a similar quantity derived from a sample.

For example, the mean of a population is a parameter, whereas the mean of a sample is an estimate (or statistic) of the population mean.

Similarly, the standard deviation of a population is a parameter, and the standard deviation of a sample is an estimate (or statistic) of the population’s standard deviation.

Estimating with uncertainty
(i.e., error around the true parameter)

An estimate (derived from a sample) is rarely, if ever, exactly the same as the population parameter being estimated—especially in large populations—because sampling is influenced by chance.

For example, two people could sample 100 trees from the same forest and get different mean values. Neither of these sample means will be exactly equal to the population mean.

The critical question in statistics is: **In the face of uncertainty (due to random chance), how much can we trust an estimate and the decisions based on it?** In other words, how accurate is the estimate (i.e., how close is the sample value to the true population value)?

The goal is to deal with uncertainty with a degree of certainty!

How to estimate with uncertainty, but with some degree of certainty (i.e., with some confidence)?

We need to understand the properties of estimators (such as the mean, variance, and standard deviation).

These **properties** are examined through the sampling distribution of the statistic or estimate of interest (e.g., sample mean, standard deviation).

A sampling distribution represents the probability distribution of an estimate based on random sampling from the population. It shows what we might observe if we were to repeatedly sample from the population.

While sampling distributions resemble frequency distributions, sampling distributions are made of probabilities instead of frequencies.

Statistical symbols

μ = population mean (we say “mu”, Greek alphabet).

σ = population standard deviation (we say “sigma”).

σ^2 = population variance (we say “sigma squared”).

Important statistical symbols regarding inference

μ = population mean (we say "mu", Greek alphabet).

σ = population standard deviation (we say "sigma").

σ^2 = population variance (we say "sigma squared").

$\overline{X}$ = sample mean (we say "X bar", Latin or Roman alphabet).

s = sample standard deviation.

s^2 = sample variance.

While μ always represent the mean of the population for any variable you're measuring (e.g., X), the symbol for the sample mean (as discussed before) can vary depending on the variable. For example, it might be written as $\bar{X}$ for the mean of X, or $\bar{Y}$ for the sample mean of Y. However, the key is that it always includes a bar on top of the variable, regardless of which variable you're referring to.

Properties of sampling distributions - the case of a tiny statistical population of 5 numbers

1,2,3,4,5; population mean (parameter) = **3.0**

All possible 15 samples (with replacement) and their means for *n* = 2:

(1,1) = 1.0
(2,2) = 2.0
(3,3) = 3.0
(4,4) = 4.0
(5,5) = 5.0

(1,2) = 1.5
(1,3) = 2.0
(1,4) = 2.5
(1,5) = 3.0

(2,3) = 2.5
(2,4) = 3.0
(2,5) = 3.5

(3,4) = 3.5
(3,5) = 4.0

(4,5) = 4.5

Notice that permutations, i.e., (1,2) = (2,1) are not shown but should be considered

Property 1: The mean of all sample means is always equal to the population mean:

$$(1.0 + 2.0 + 3.0 + 4.0 + 5.0 + 1.5 + 2.0 + 2.5 + 3.0 + 2.5 + 3.0 + 3.5 + 3.5 + 4.0 + 4.5) / 15 = \mathbf{3.0}$$

Sample means of the sample population varied from 1.0 to 5.0

sample size (i.e., number of observational units) is represented by the letter "n". Here, n = 2 observational units.

Properties of estimators are based on the sampling distribution under random sampling of the estimate of interest (here, sample mean).

Property 1: The average of the sample means will always be equal to the true population mean; as such, the

When the mean of all possible sample means—i.e., the mean of the sampling distribution of an estimate (such as the sample mean or standard deviation)—equals the population parameter, the estimate is said to be **unbiased**. This holds true when sampling is done randomly, meaning that each observation in the population has an equal chance of being selected.

In this case, the sample mean is unbiased because, under random sampling, the sample means do not systematically tend to be either larger or smaller than the true population mean

$$(\mathbf{1.0} + \mathbf{2.0} + 3.0 + \mathbf{4.0} + \mathbf{5.0} + \mathbf{1.5} + \mathbf{2.0} + \mathbf{2.5} + 3.0$$
$$+ \mathbf{2.5} + 3.0 + \mathbf{3.5} + \mathbf{3.5} + \mathbf{4.0} + \mathbf{4.5}) / 15 = \mathbf{3.0}$$

6 sample means smaller than the true population value [**in red**]

6 sample means greater than the true population value [**in green**]

3 sample means equal to the true population value [**in black**]

Random sampling reduces both sampling error and inferential bias, which refers to how close or far the sample values are from the true population value for the statistic of interest.

Remember: a random sample is one that fulfills two criteria:

1) Every observational unit in the population (e.g., individual tree) have an **equal chance** of being included in the sample.

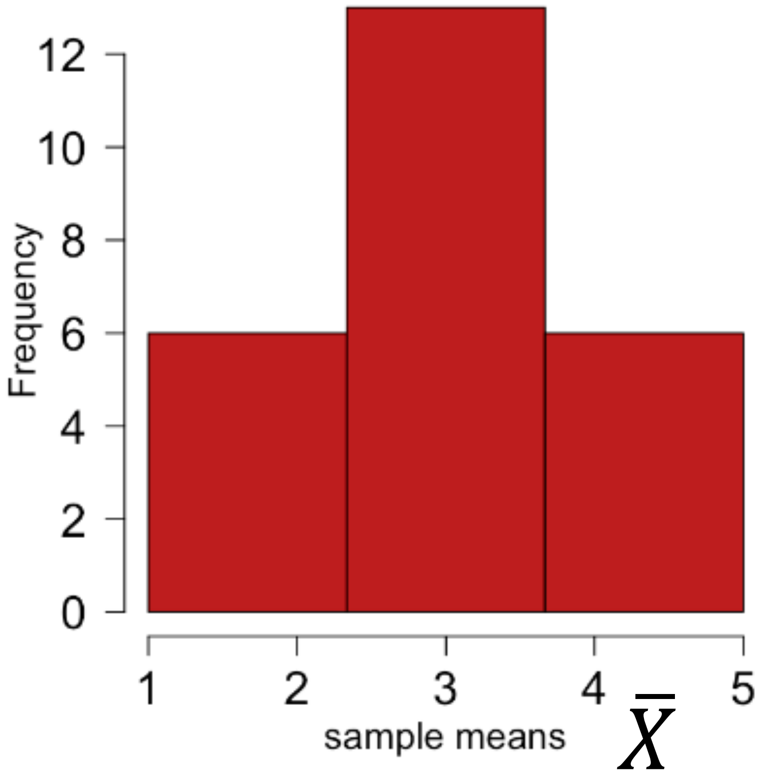
2) The selection of observational units in the population (e.g., individual tree) must be **independent**, i.e., the selection of any unit (e.g., individual tree) of the population must not influence the selection of any other unit.

Sampling bias occurs when some observational units are more or less likely to be selected, leading to an unrepresentative sample.

As a result, not all possible combinations of observational units are equally likely to be sampled, which can indeed lead to a sampling distribution that has a different mean from the true population mean. This difference is a form of bias and can skew statistical estimates, making them less reliable.

Estimating with uncertainty: the sampling distribution of the mean based a tiny statistical population of 5 numbers

25 possible different combinations of 2 numbers (i.e.,25 different potential samples; with repetition of observational units, i.e., (1,2),(2,1), etc) from 1,2,3,4,5 (population)



$\mu = 3$

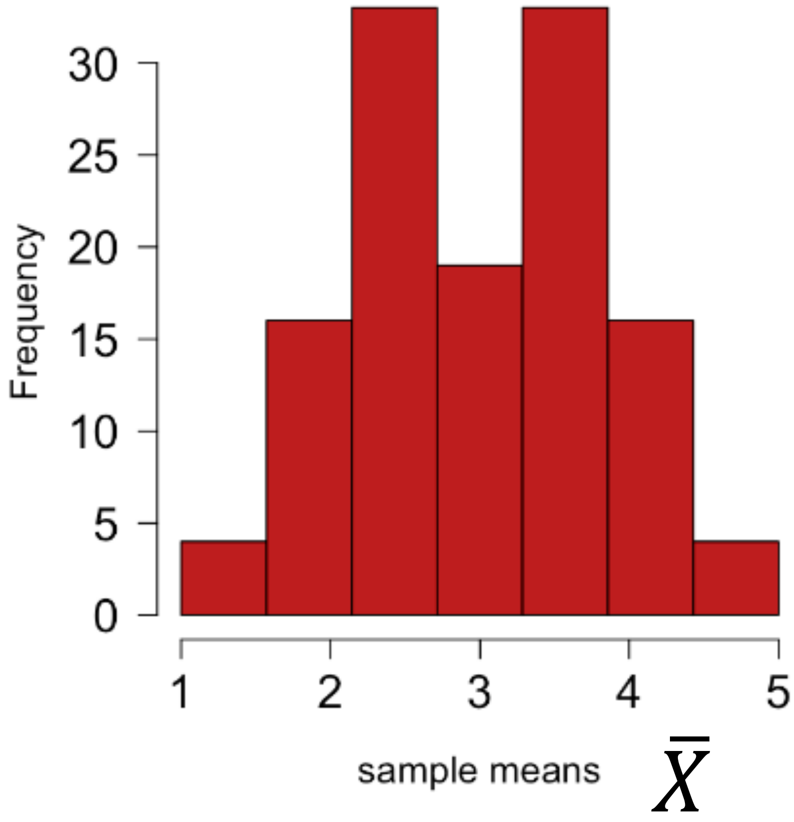
Mean of all samples means = 3.0

n = 2

μ (*symbol for the population mean*)

Estimating with uncertainty: the sampling distribution of the mean based a tiny statistical population of 5 numbers

125 possible different combinations of 3 numbers (i.e.,125 different potential samples; with repetition of observational units, i.e., (1,2,1),(2,1,1), etc) from 1,2,3,4,5 (population)



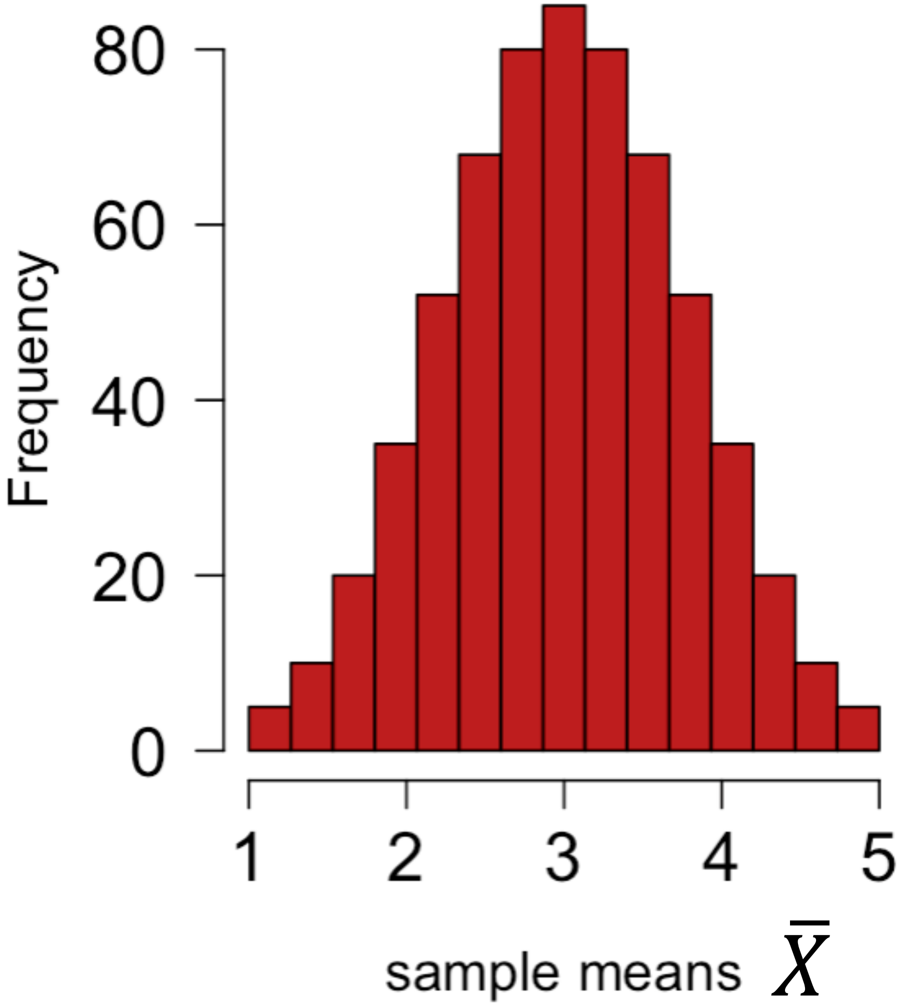
$\mu = 3$

Mean of all samples means = 3.0

n = 3

Estimating with uncertainty: the sampling distribution of the mean based a tiny statistical population of 5 numbers

625 possible different combinations of 4 numbers (i.e.,625 different potential samples; with repetition of observational units, i.e., (1,2,1,3),(2,1,1,4), etc) from 1,2,3,4,5 (population)



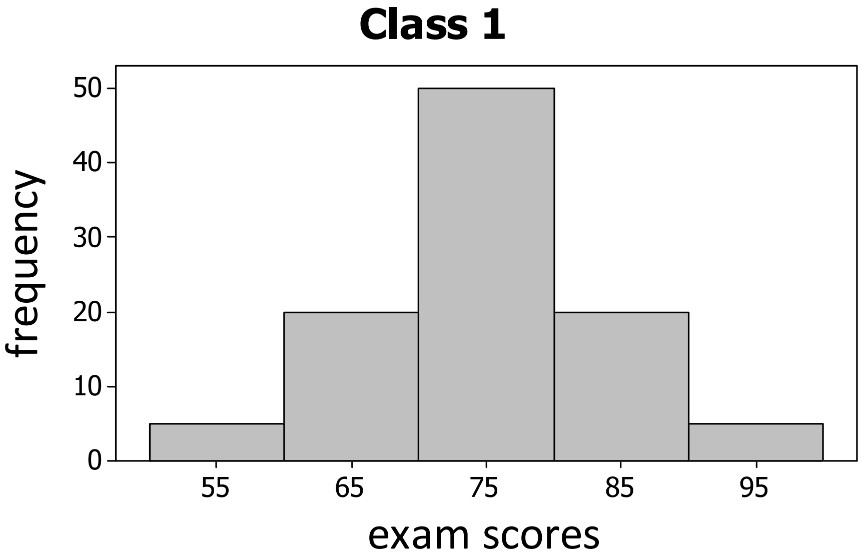
$\mu = 3$

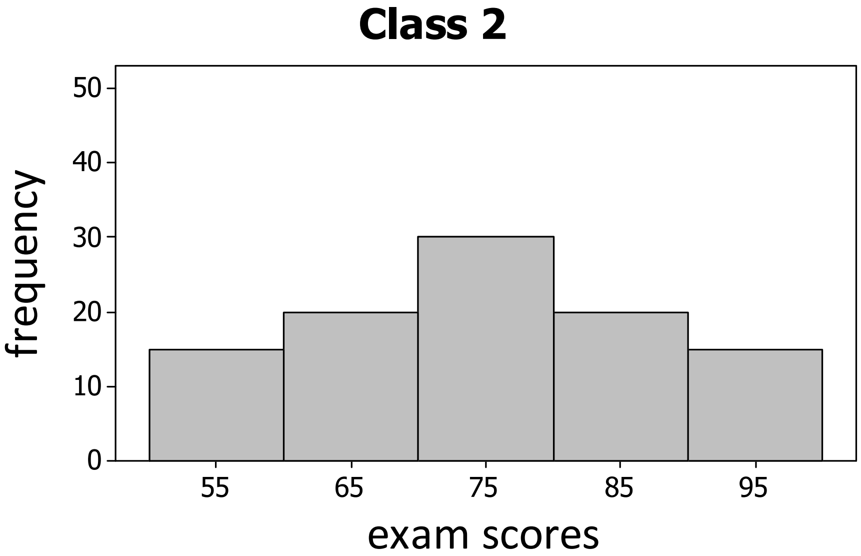
Mean of all samples means = 3.0

n = 4

Remember Variability in frequency distributions?!!

In which class exam scores vary the most?

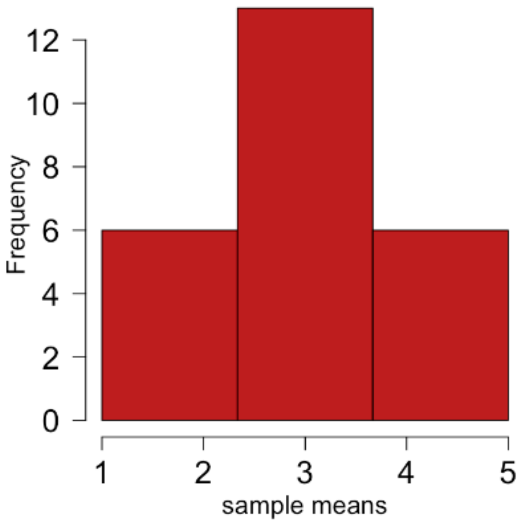




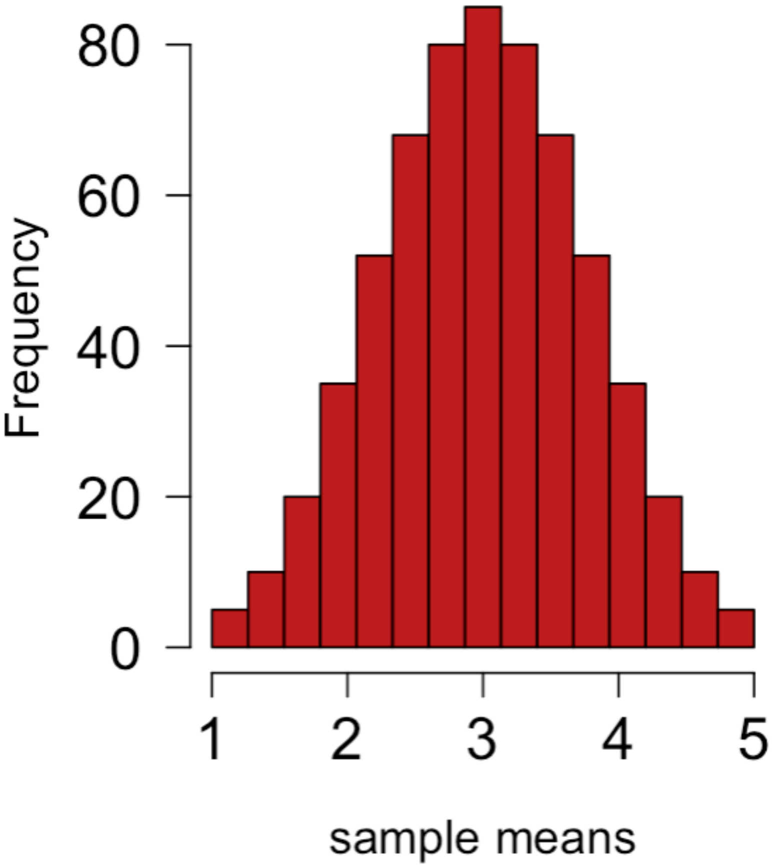
- Source: Cooper & Shore; Journal of Statistics Education (vol. 18, #2)

What sample size leads to more precise sample estimates (i.e., less variation around the true population value) when using random sampling?

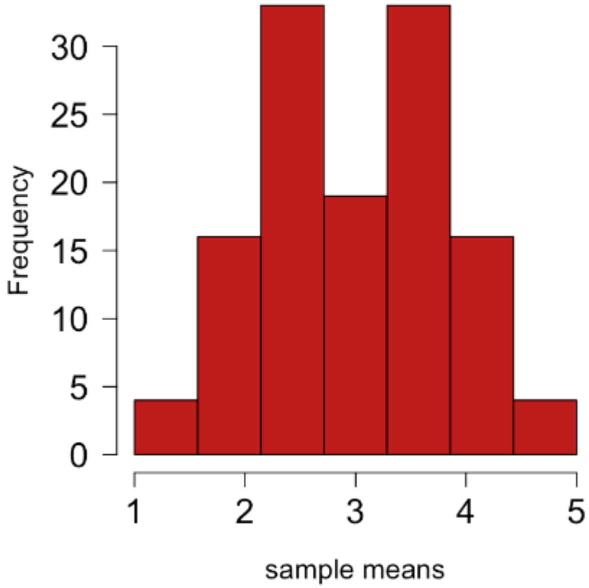
n = 2



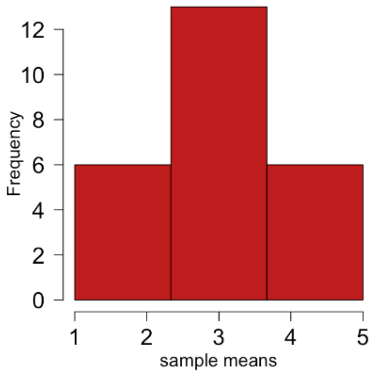
n = 4

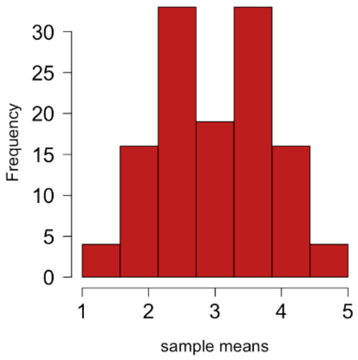


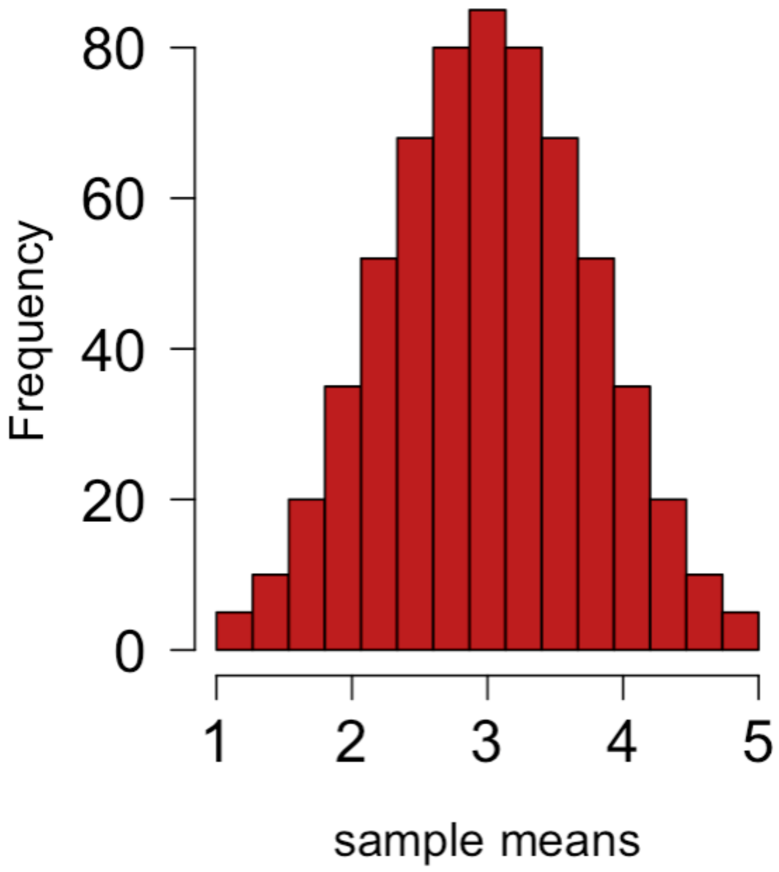
n = 3



What sample size leads to more precise sample estimates (i.e., less variation around the true population value) when using random sampling?



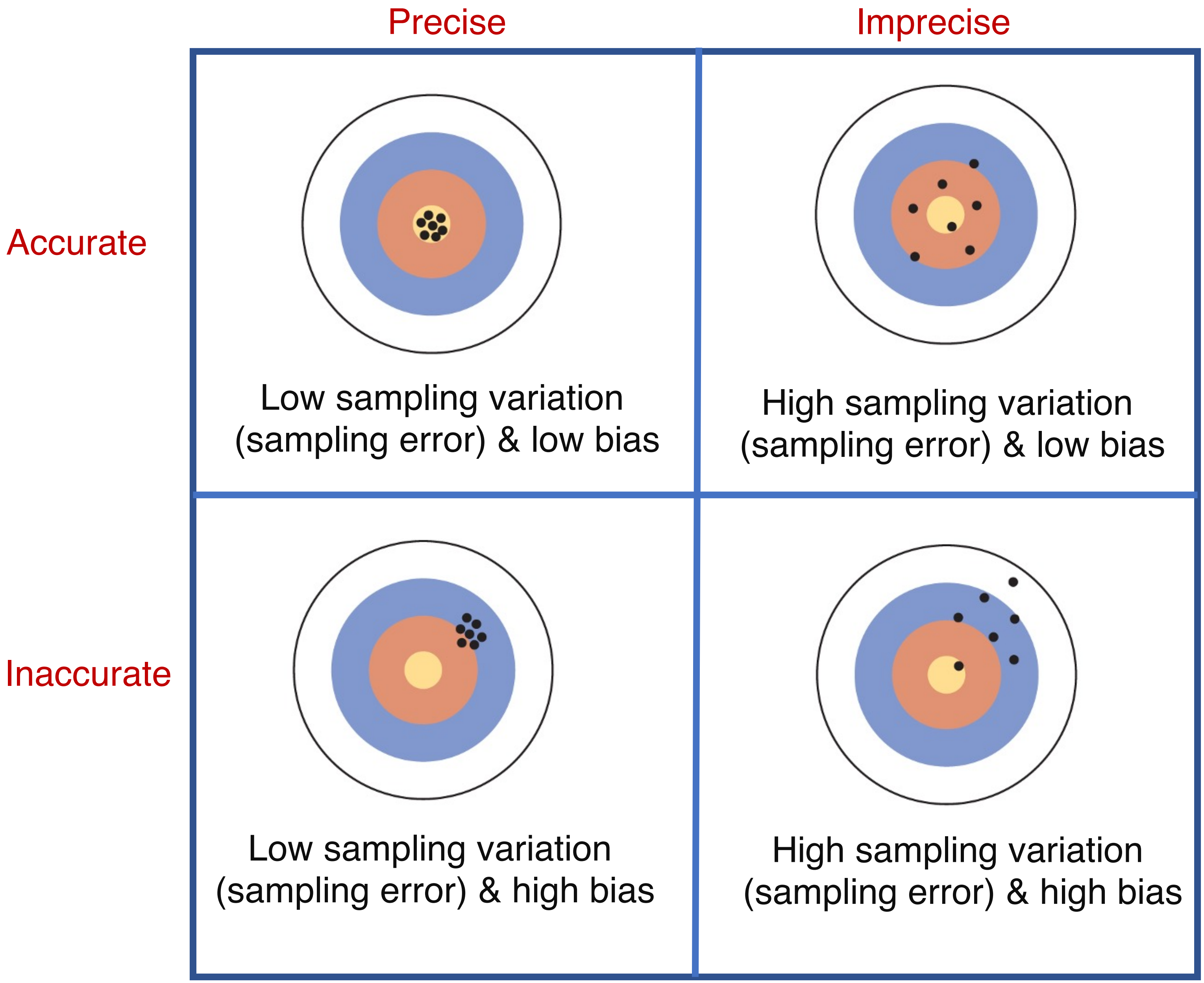




As the sample size increases, the likelihood that a random sample will be closer to the true population mean also increases, leading to more **precise** estimates.

Since the sampling was random, the sample mean is **accurate**, meaning it is unbiased. The mean of all sample means equals the population mean (the true parameter).

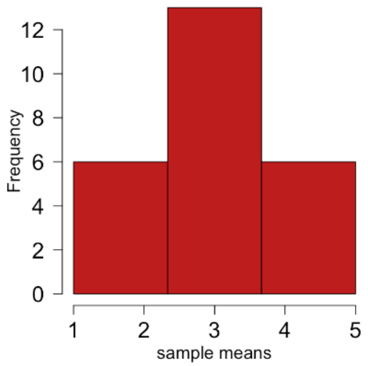
Random sampling minimizes sampling error and allows for the measurement of sampling error (next lectures)



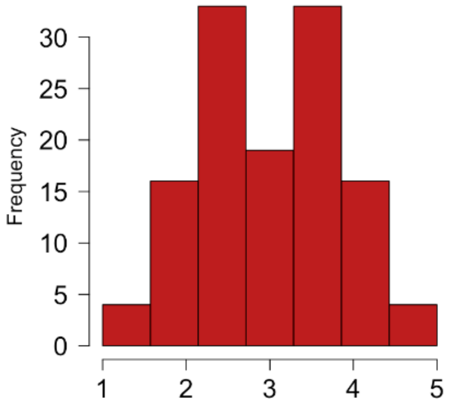
	Precise	Imprecise
Accurate	Low sampling variation (sampling error) & low bias	High sampling variation (sampling error) & low bias
Inaccurate	Low sampling variation (sampling error) & high bias	High sampling variation (sampling error) & high bias

All sample sizes here lead to accurate estimates, but which one results in greater precision?

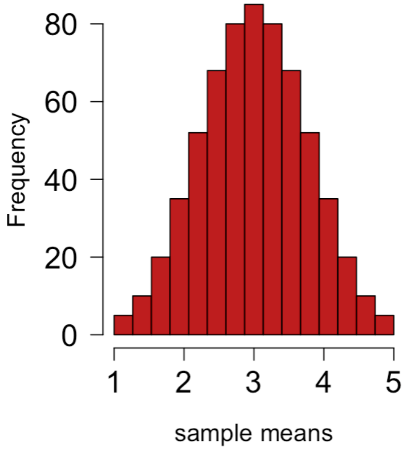
n = 2

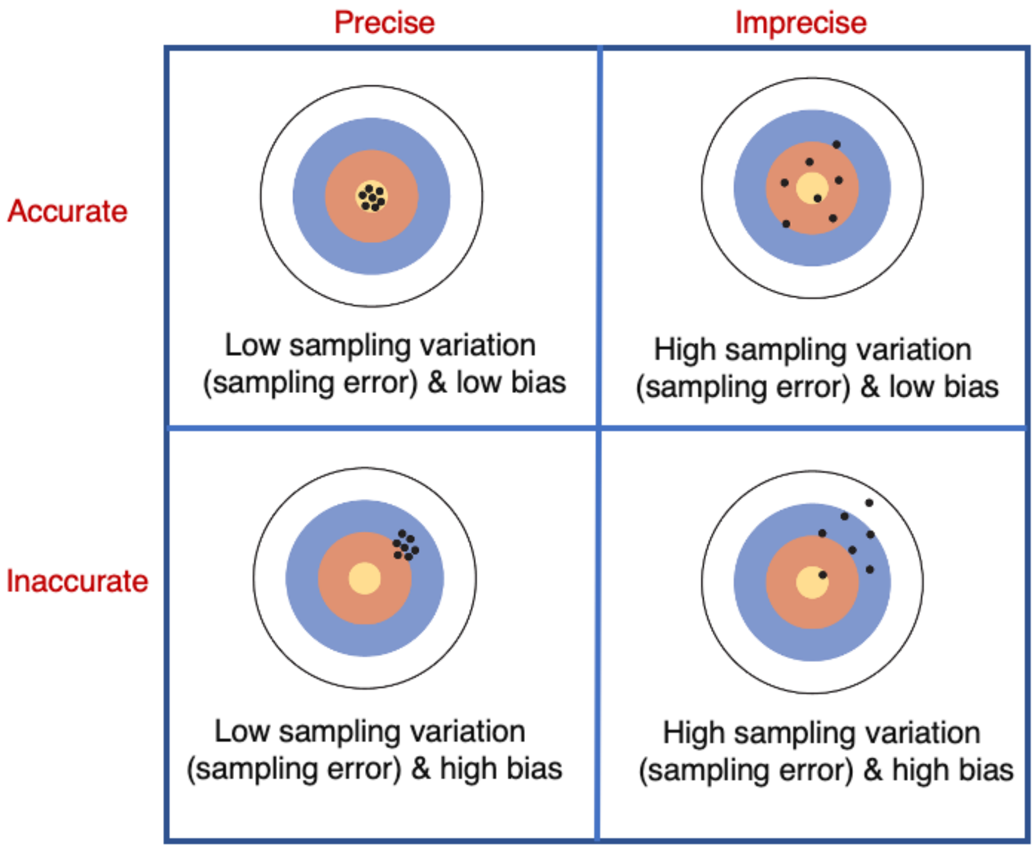


n = 3



n = 4





Precise

Imprecise

Accurate

Low sampling variation (sampling error) & low bias

High sampling variation (sampling error) & low bias

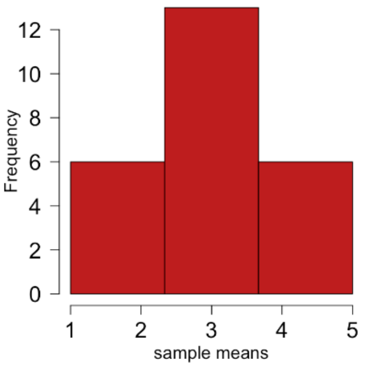
Inaccurate

Low sampling variation (sampling error) & high bias

High sampling variation (sampling error) & high bias

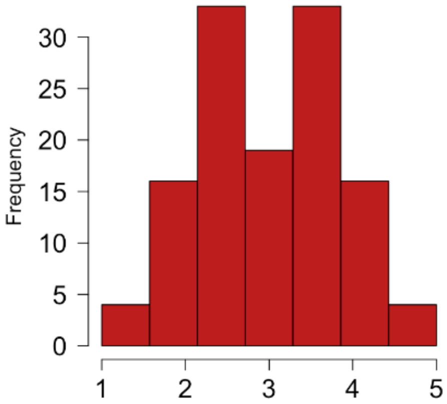
Properties of sampling distributions

n = 2



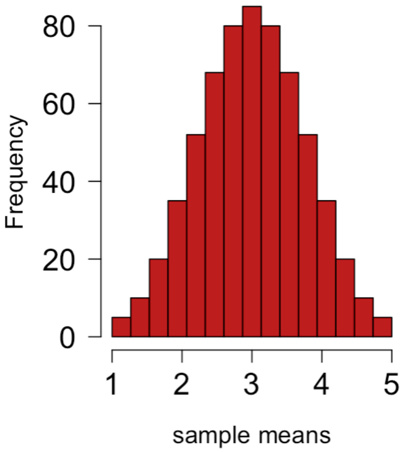
All sample sizes here lead to accurate estimates, but which one results in greater precision?

n = 3



Property 2: With random sampling, larger sample sizes increase the likelihood that the sample mean will be closer to the true population mean; this increasing estimator precision

n = 4

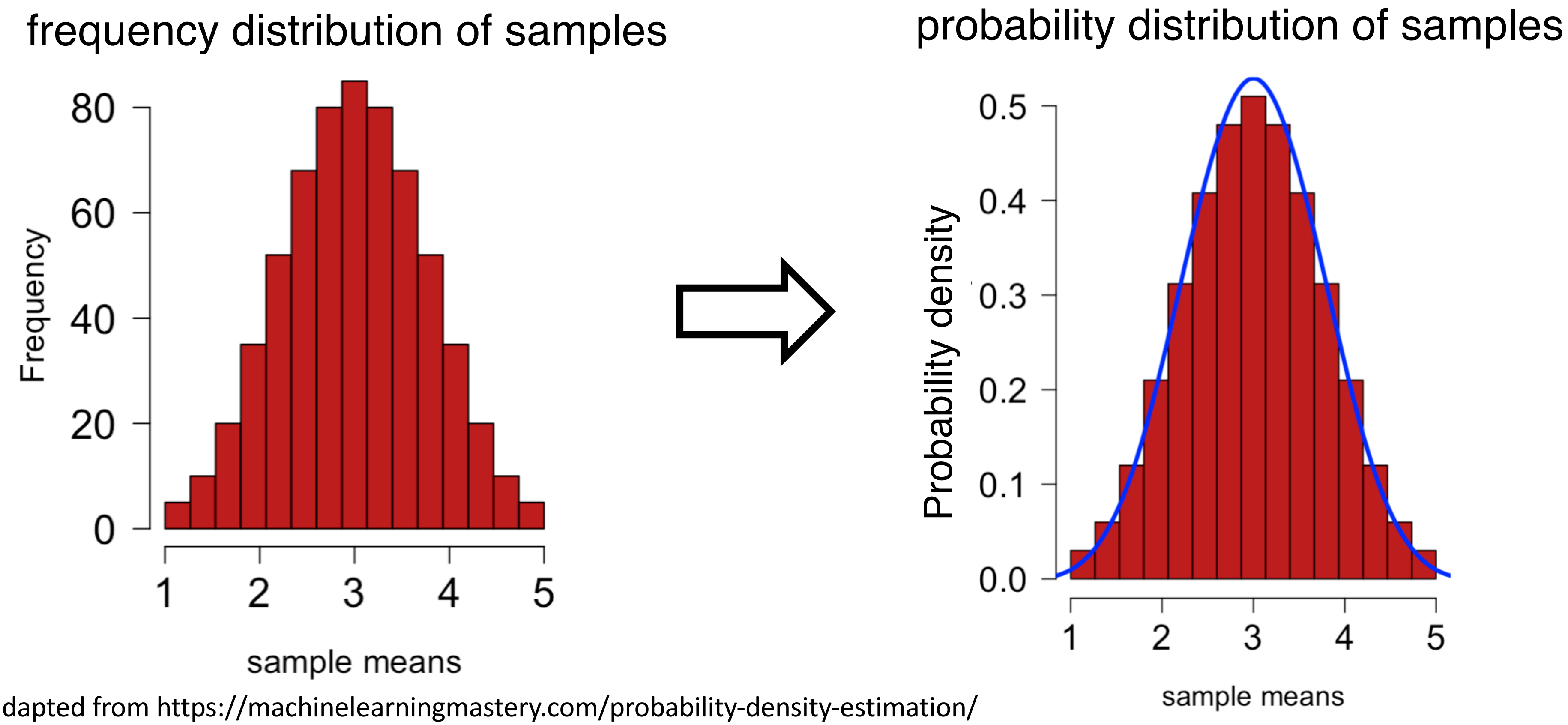


Sampling distributions are best represented by probability distributions

Probability density describes the relationship between observations (in this case, sample means) and their likelihood of occurring.

Some outcomes (sample means) will have a low probability density, while others will have a high probability density.

The overall shape formed by these densities is called the probability distribution, and specific probabilities for outcomes of a random variable are calculated using a probability density function (PDF).

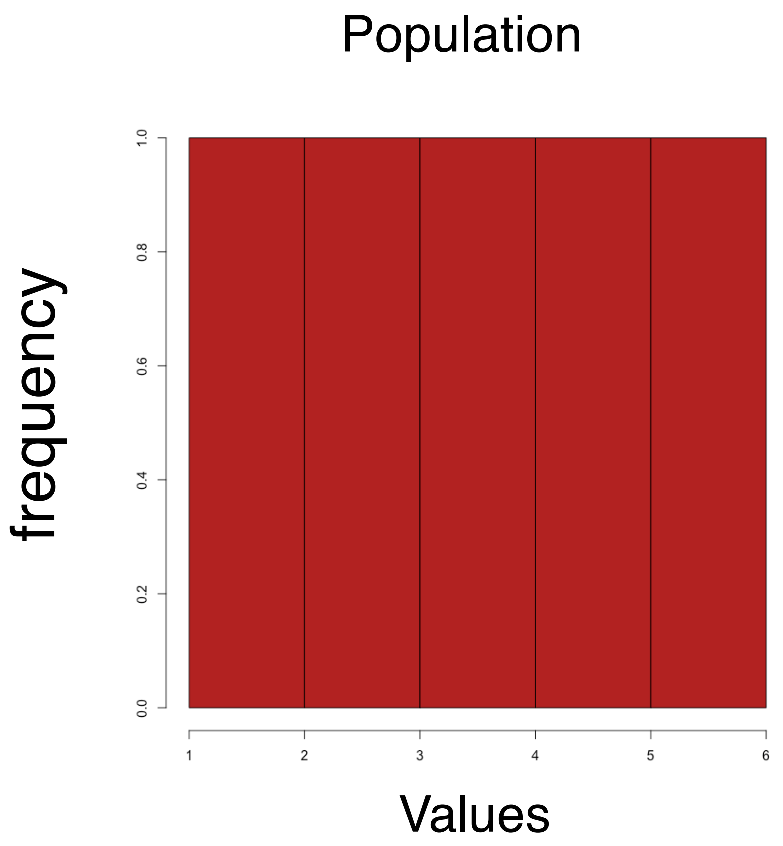


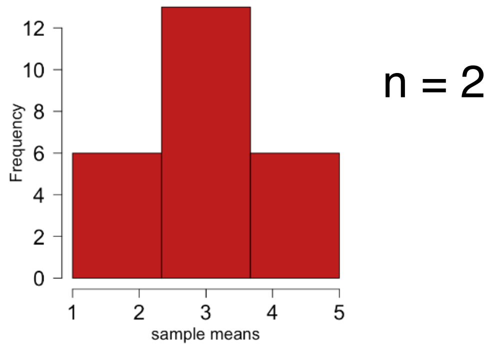
frequency distribution of samples

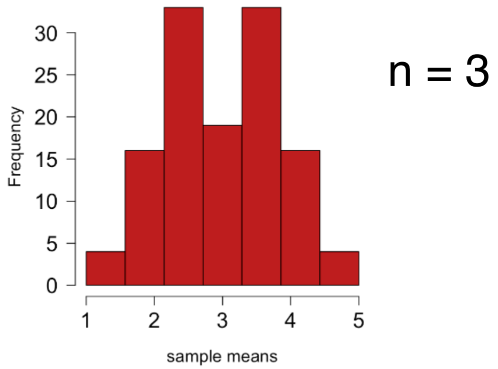
probability distribution of samples

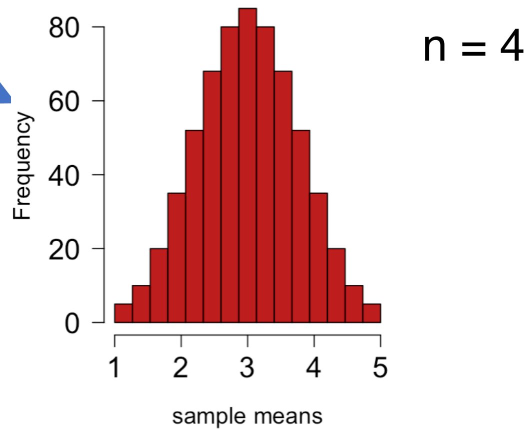
Text adapted from https://machinelearningmastery.com/probability-density-estimation/

Critical: The shape of the population's frequency distribution (called "marginal distribution") is not necessarily the same as the frequency distribution of sample estimates (e.g., The sampling distribution of mean values) drawn from that population.







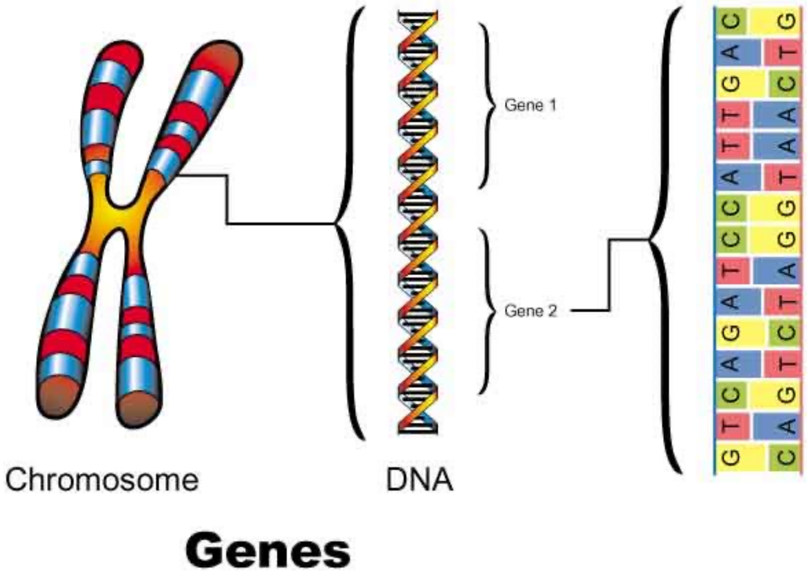


population: 1,2,3,4,5;
population mean=**3.0**

Let’s take a break – 1 minute

The length of protein-coding genes in humans is a rare example of an almost complete statistical population in biology

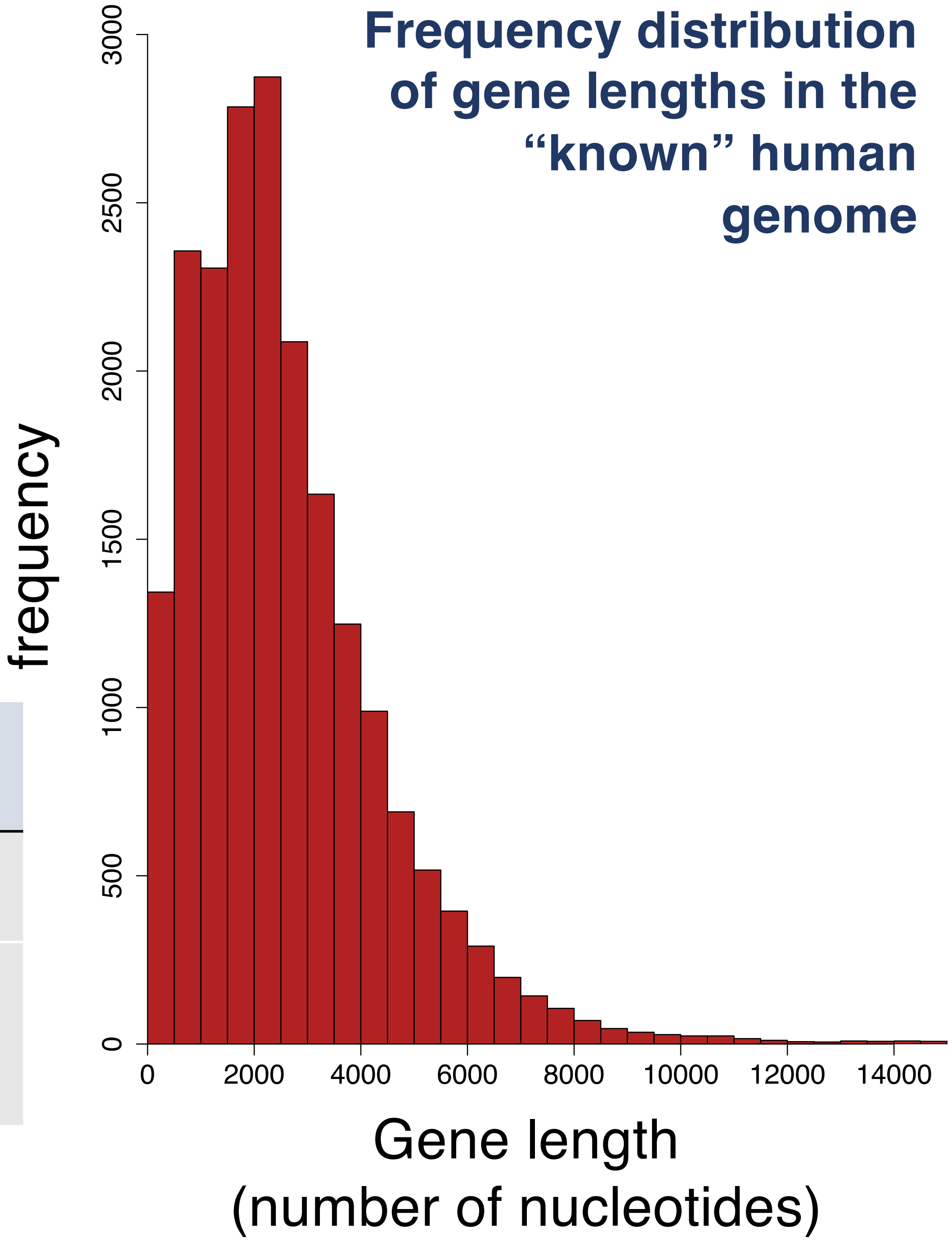
The International Human Genome Project produced the complete DNA sequence for all 23 human chromosomes, each containing millions of nucleotides and more than 23,000 protein-coding genes. The project began in 1990 and was completed in 2006 with the sequencing of the last chromosome. For BIOL 322 tutorials, the available data includes 20,290 genes.



The length of human genes

It involves the length of almost all human genes, i.e., these is very close to the true *population* of genes!

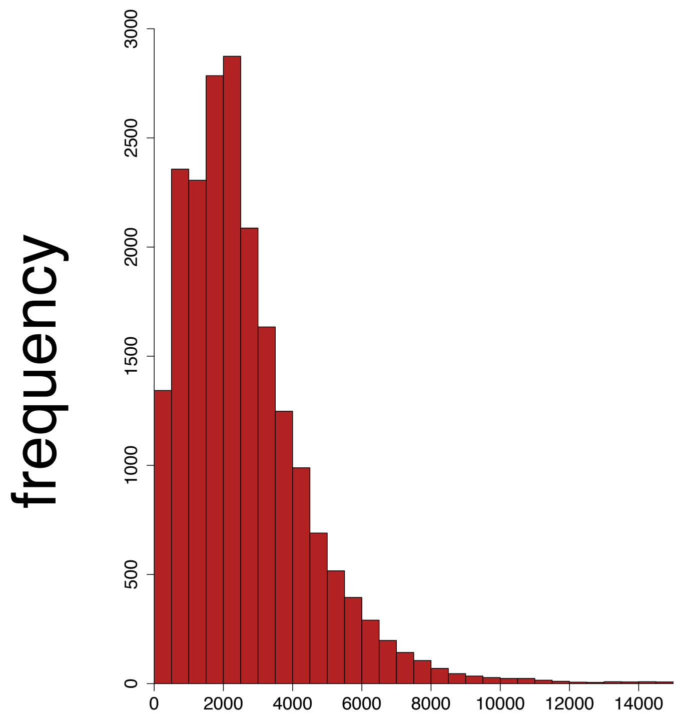
Names	Parameter	Value (nucleotides)
Mean (mu)	μ	2622.0
Standard deviation (sigma)	σ	2036.9



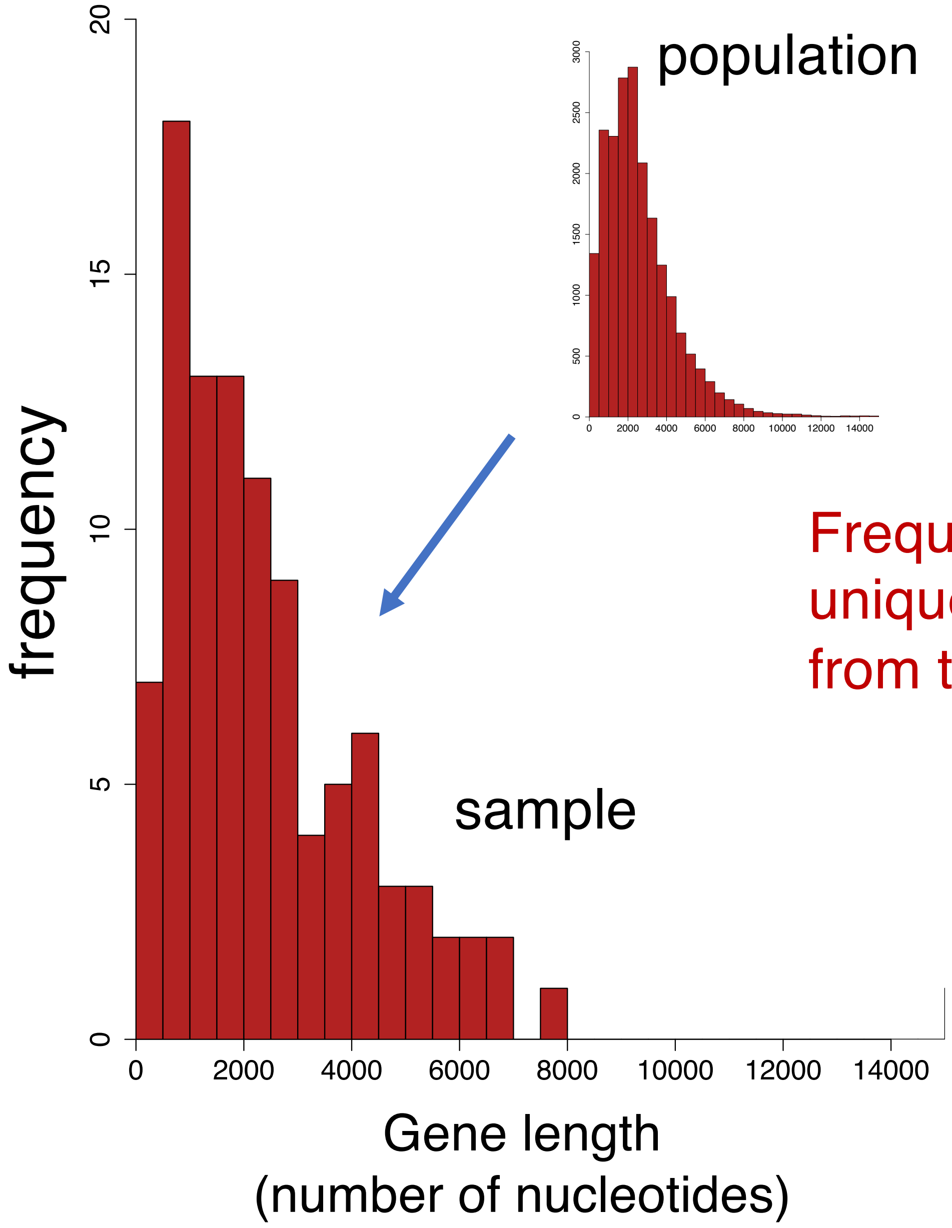
In real situations, we typically don't know the parameter values of the study population, but in this case, we (almost) do!

So, we'll take advantage of this gene population to illustrate the processes of sampling, uncertainty, accuracy, precision, and how to estimate with uncertainty—yet with some level of confidence!

Names	Parameter	Value (nucleotides)
Mean	μ	2622.0
Standard deviation	σ	2036.9



Estimating mean gene length with a random sample of 100 genes (random sampling out of 20,290 genes)



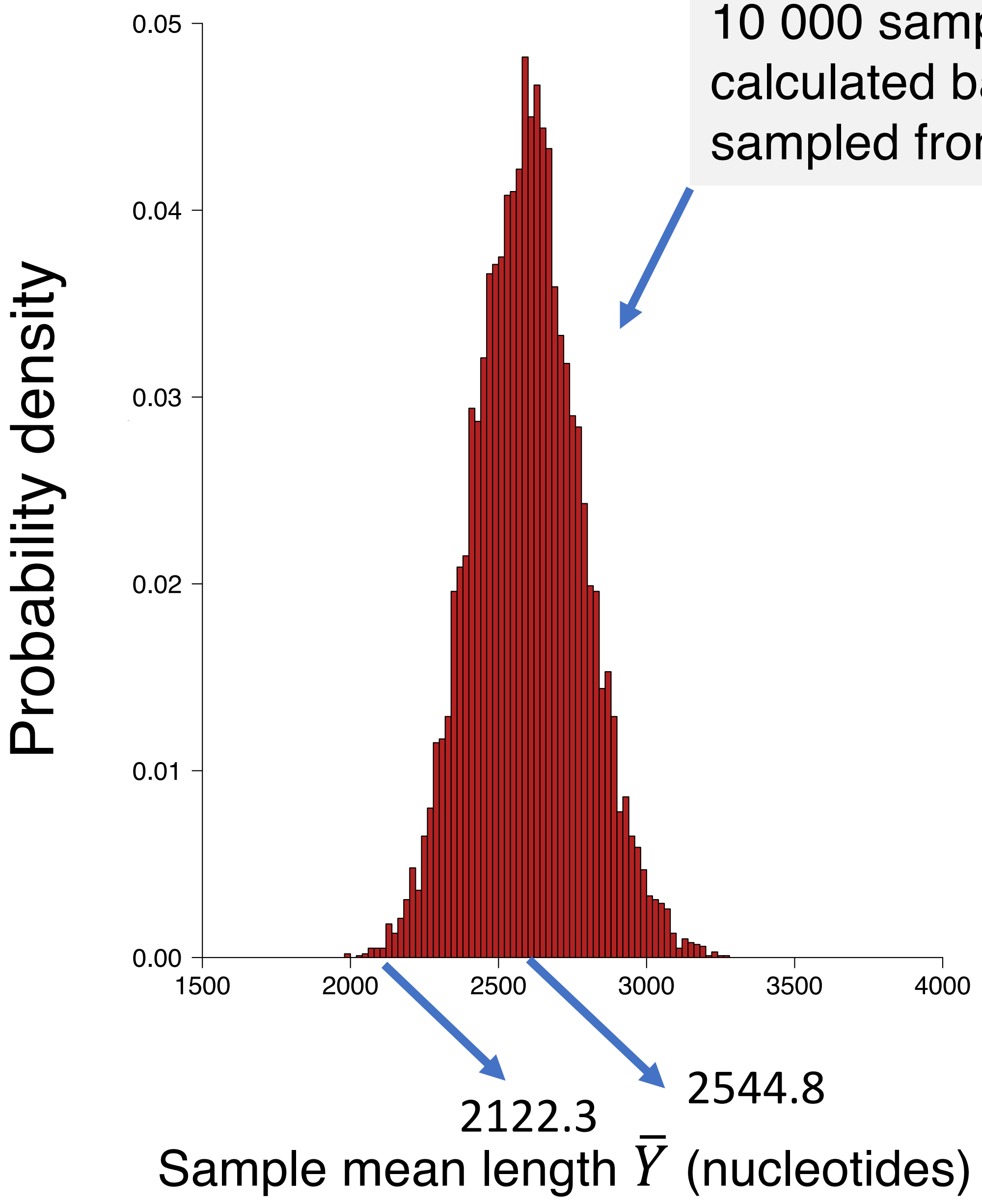
Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2544.8
Standard deviation	s	2125.3

Frequency distribution of gene lengths in a unique random sample of n = 100 genes from the human genome.

Imagine a group in Canada and another in France in 1985 working on the same problem, i.e., estimating the average gene length in the human genome; they would have different sample means

The sampling distribution of sample means ($\bar{Y}$)

Sampling distribution of means based on 10 000 sample mean values. Each sample mean is calculated based on the lengths of 100 genes randomly sampled from the population of 20,290 genes.



Mean and standard deviation of two possible samples from the same population (out of the 10,000 samples):

Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2544.8
Standard deviation	s	2125.3

Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2122.3
Standard deviation	s	2423.1

Estimating mean gene length with a random sample of 100 genes (random sampling out of 20,290 genes) – variation due to pure chance (i.e., random sampling)

Population

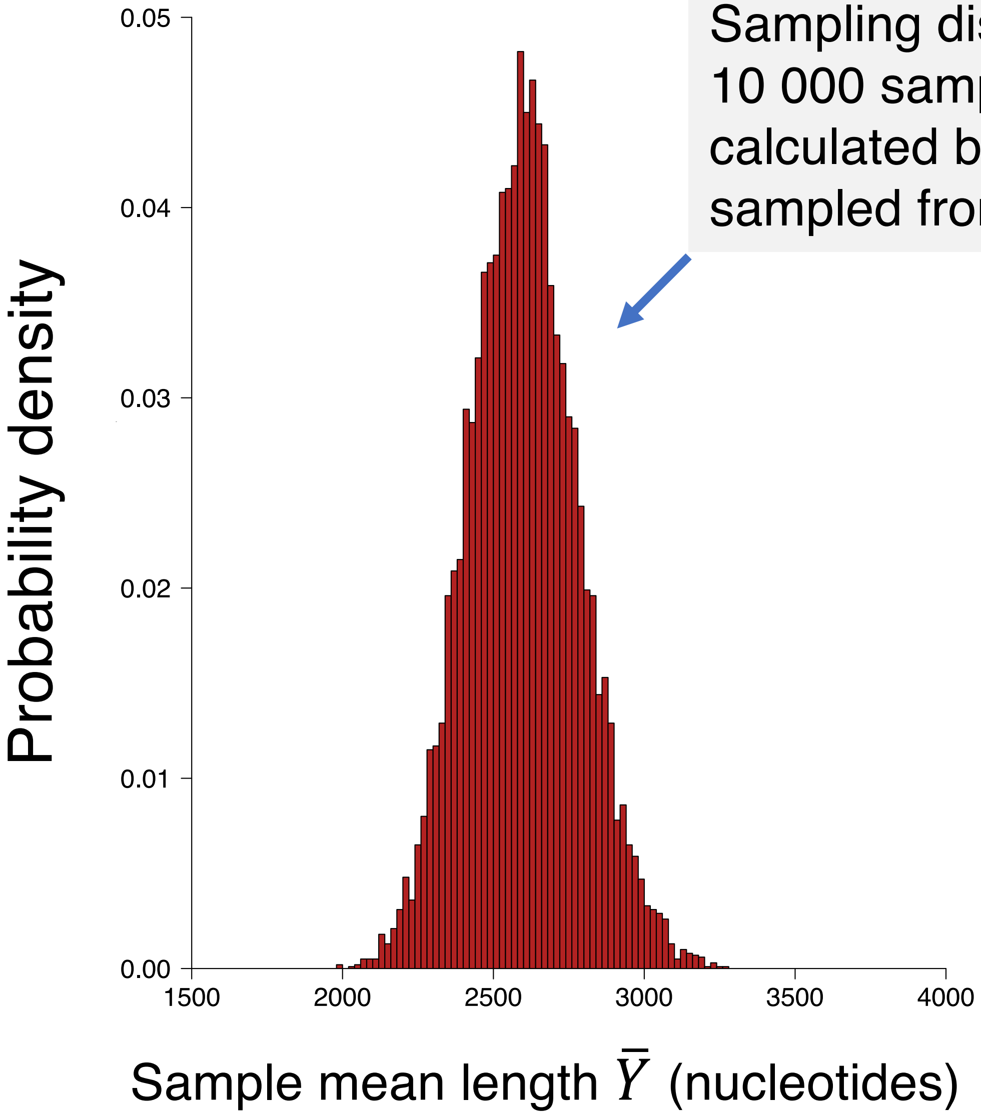
Names	Parameter	Value (nucleotides)
Mean	μ	2622.0
Standard deviation	σ	2036.9

Sample

Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2544.8
Standard deviation	s	2125.3

The sample mean is approximately 77 nucleotides shorter than the true population value. We shouldn't be surprised that the sample estimates differ from the population parameter; such differences are virtually inevitable due to random sampling variation.

The sampling distribution of sample means ($\bar{Y}$)

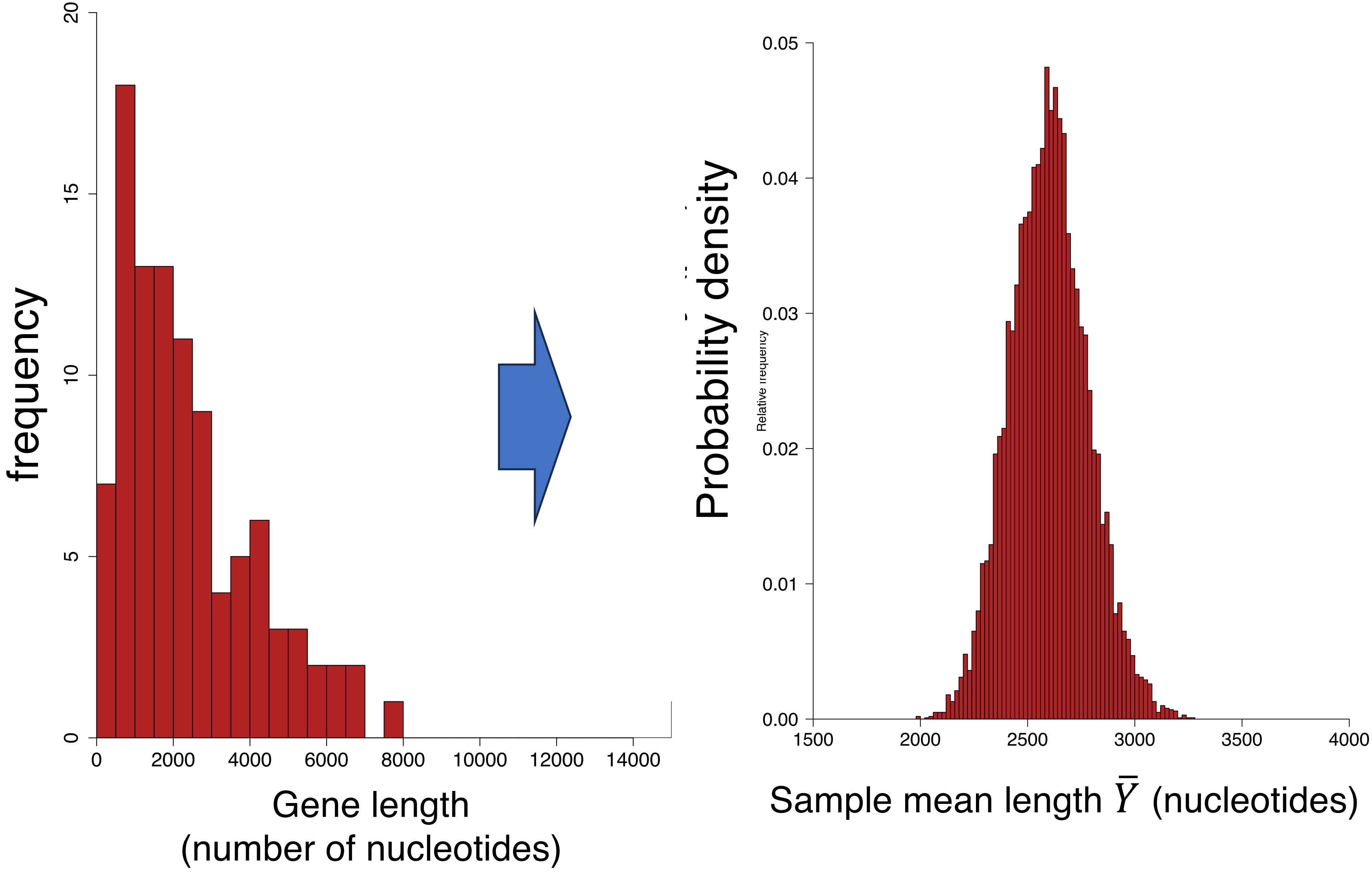


Sampling distribution of means based on 10 000 sample mean values. Each sample mean is calculated based on the lengths of 100 genes randomly sampled from the population of 20,290 genes.

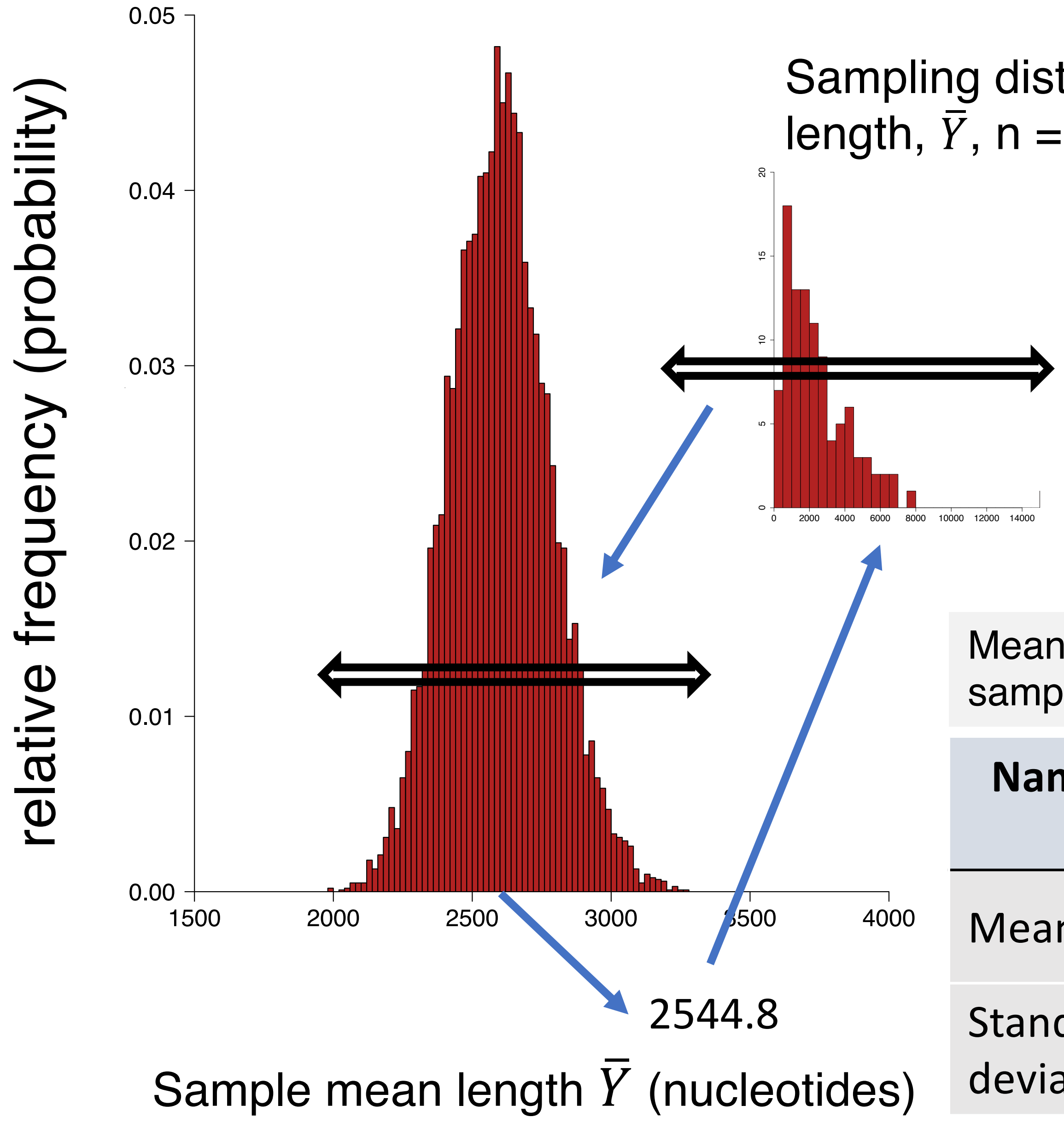
Here, 10,000 sample means were drawn from the population using a computational approach.

However, in statistics, we use an analytical, calculus-based approach to estimate the sampling distribution (or probability distribution) of all possible sample means, whether based on 100 genes or any other sample size. This was critical as most important probability distributions were developed without computers and many more than 100 years ago.

Critical: The shape of the population's frequency distribution (called “marginal distribution”) is not necessarily the same as the frequency distribution of sample estimates (e.g., The sampling distribution of mean values) drawn from that population.



The sampling distribution (probability distribution) of sample means ($\bar{Y}$)



Sampling distribution of mean gene length, $\bar{Y}$, n = 100.

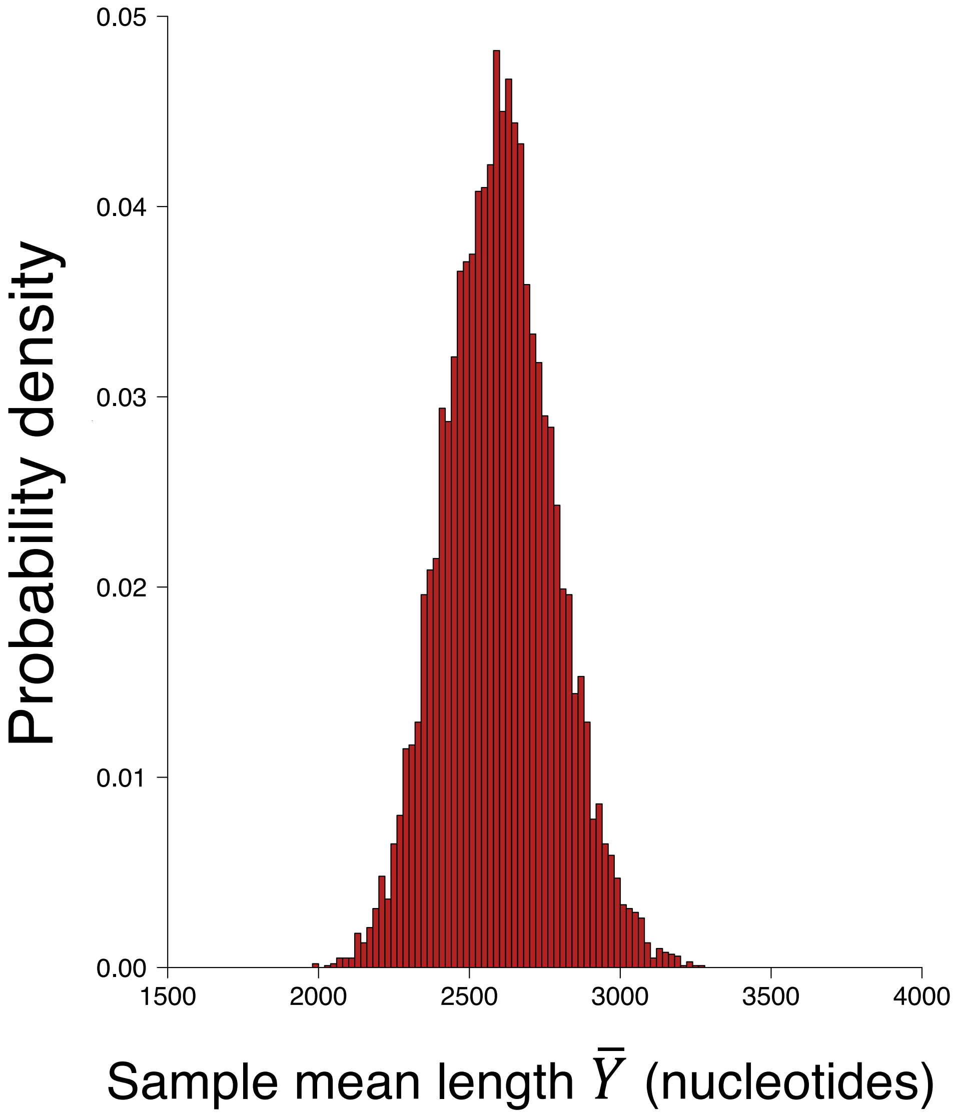
Statistical Insight: We will learn that the variation within a single sample can be used to estimate the uncertainty across all possible sample values from a population.

Mean and standard deviation of one single sample of 100 genes out of 20,290

Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2544.8
Standard deviation	s	2125.3

Let’s take a break – 1 minute

The sampling distribution (probability distribution) of sample means ($\bar{Y}$)



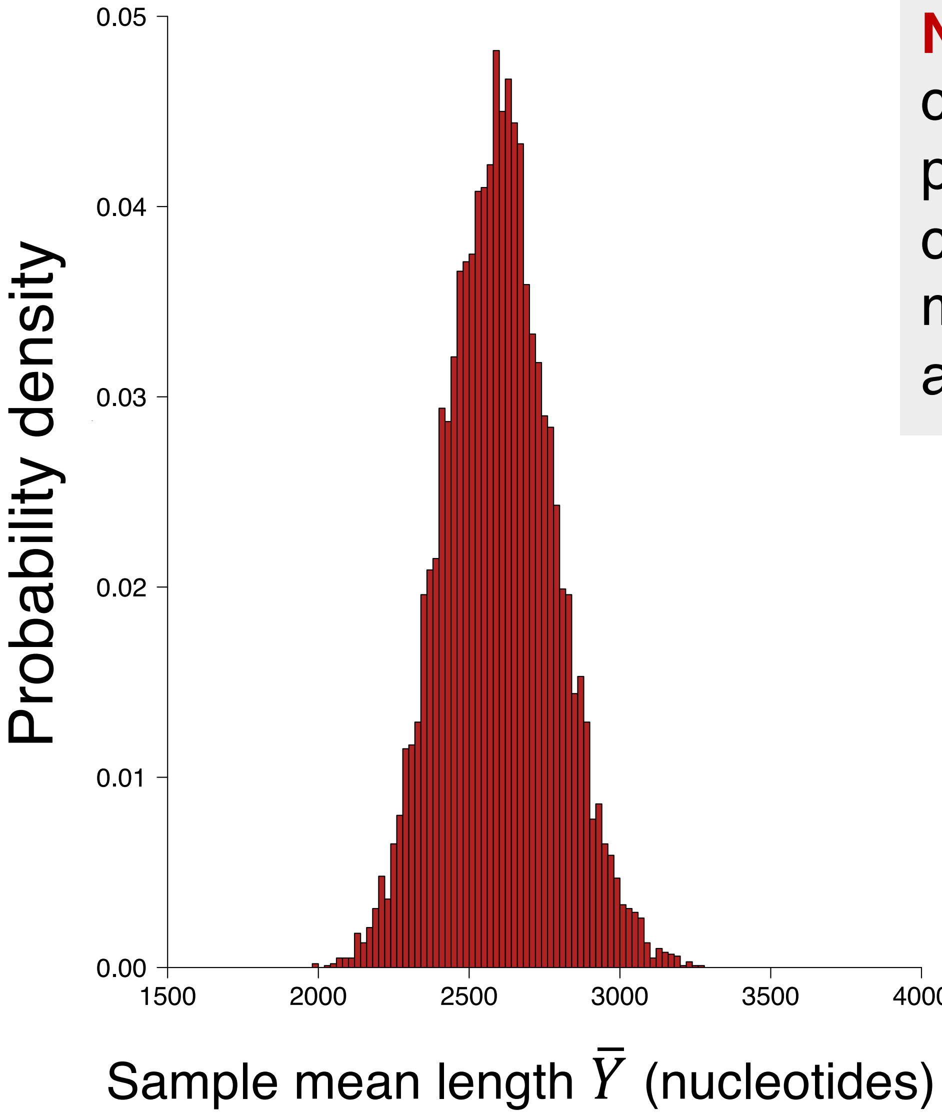
Note 1: We typically work with a single sample, which gives us just one sample mean value ($\bar{Y}$).

However, understanding how sampling distributions are constructed is essential for estimating uncertainty (i.e., how sample mean values vary from one sample to another) and determining the confidence in inferences based on those samples.

If samples show high variability, we will have less confidence in our estimates, whereas lower variability among samples leads to greater confidence.

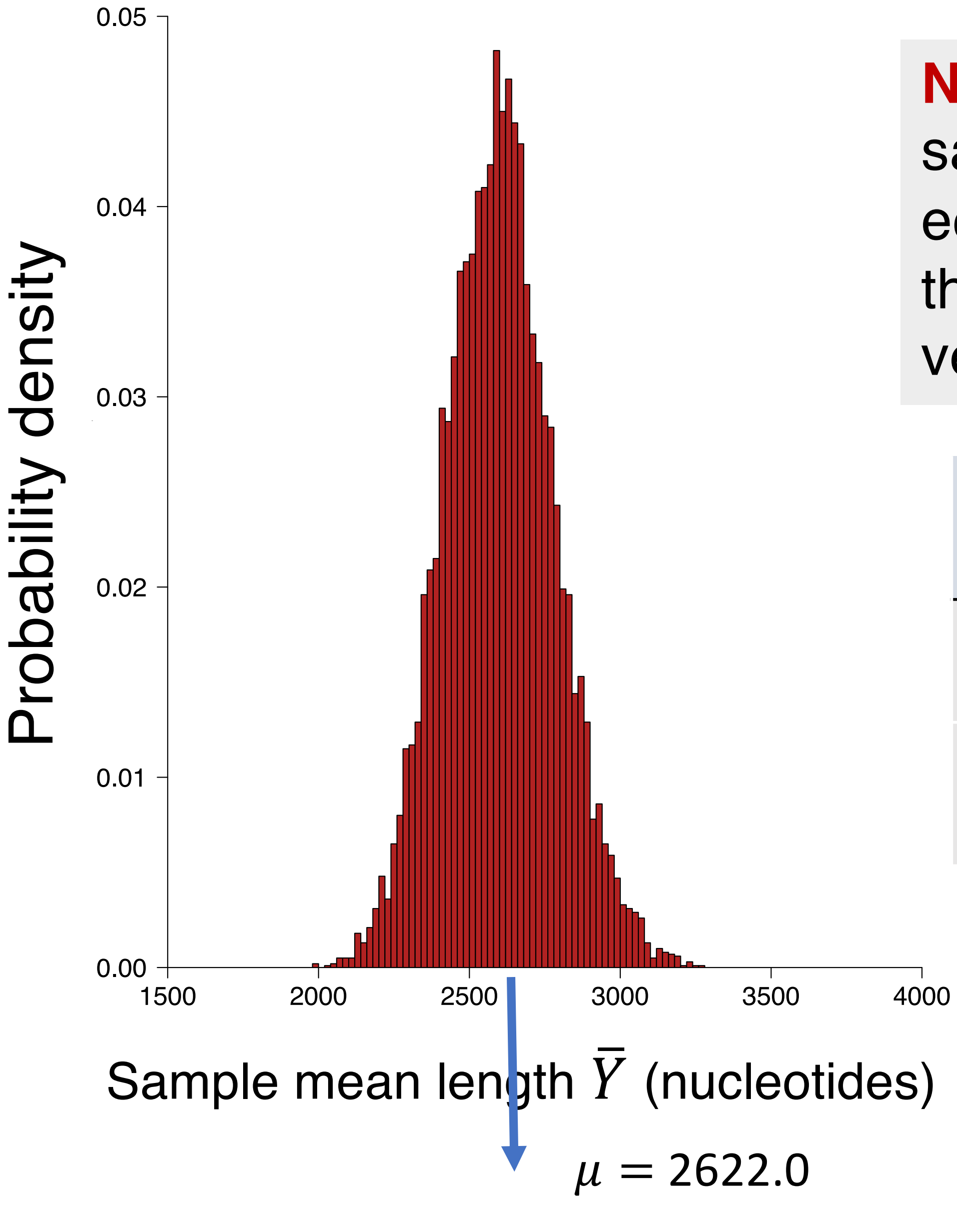
Interestingly, variation within a single sample can provide insights into variation among samples (we will explore this in upcoming lectures).

The sampling distribution (probability distribution) of sample means ($\bar{Y}$)



Note 2: The sampling distribution clearly shows that while the population mean (μ = 2622.0) is considered a constant, the sample mean ($\bar{Y}$) is a variable that fluctuates across different samples.

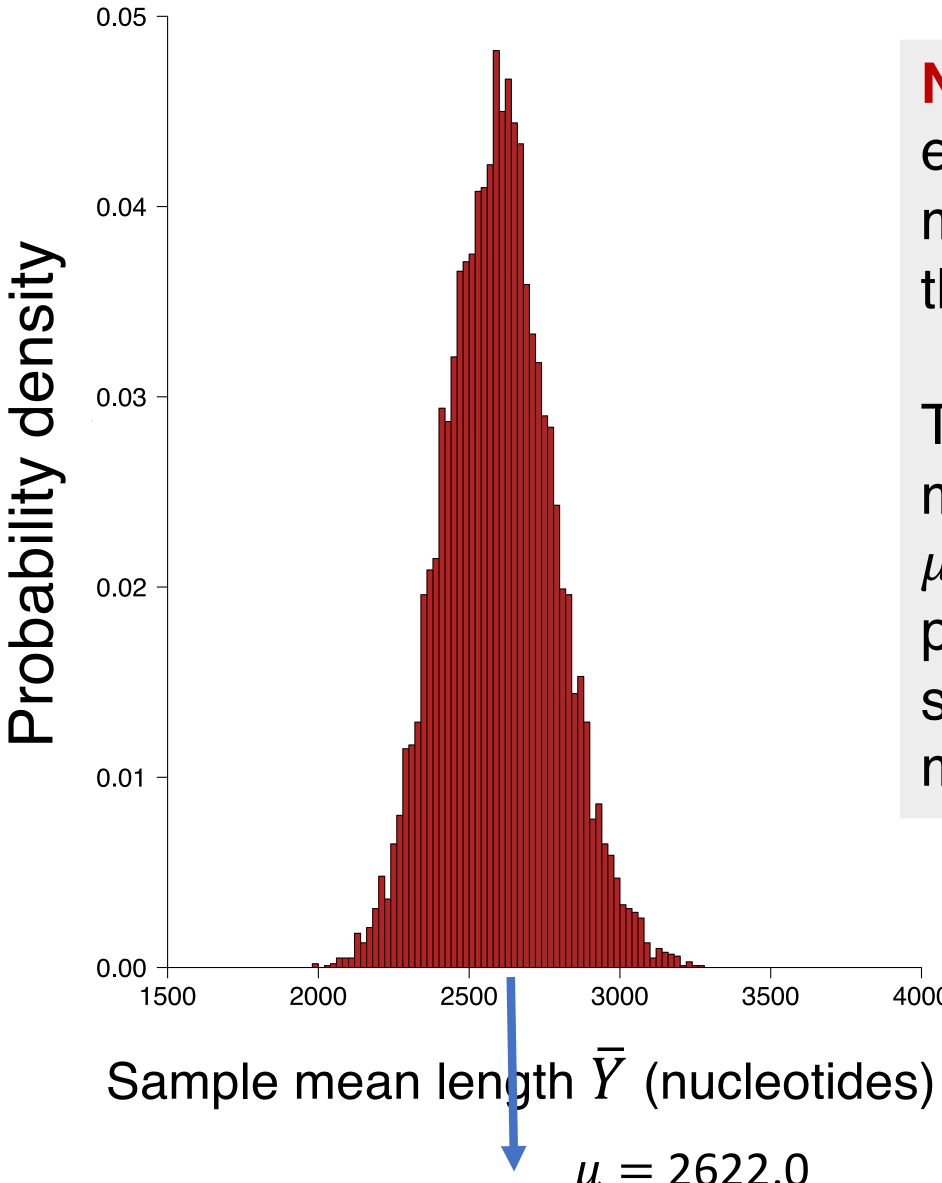
The sampling distribution (probability distribution) of sample means ($\bar{Y}$)



Note 3 (again): The mean of all sample estimates of the mean is equal to the population mean. Even the mean of 10,000 sample means is very close to it.

Names	Parameter	Value (nucleotides)
Mean	μ	2622.0
Standard deviation	σ	2036.9

The sampling distribution (probability distribution) of sample means ($\bar{Y}$)

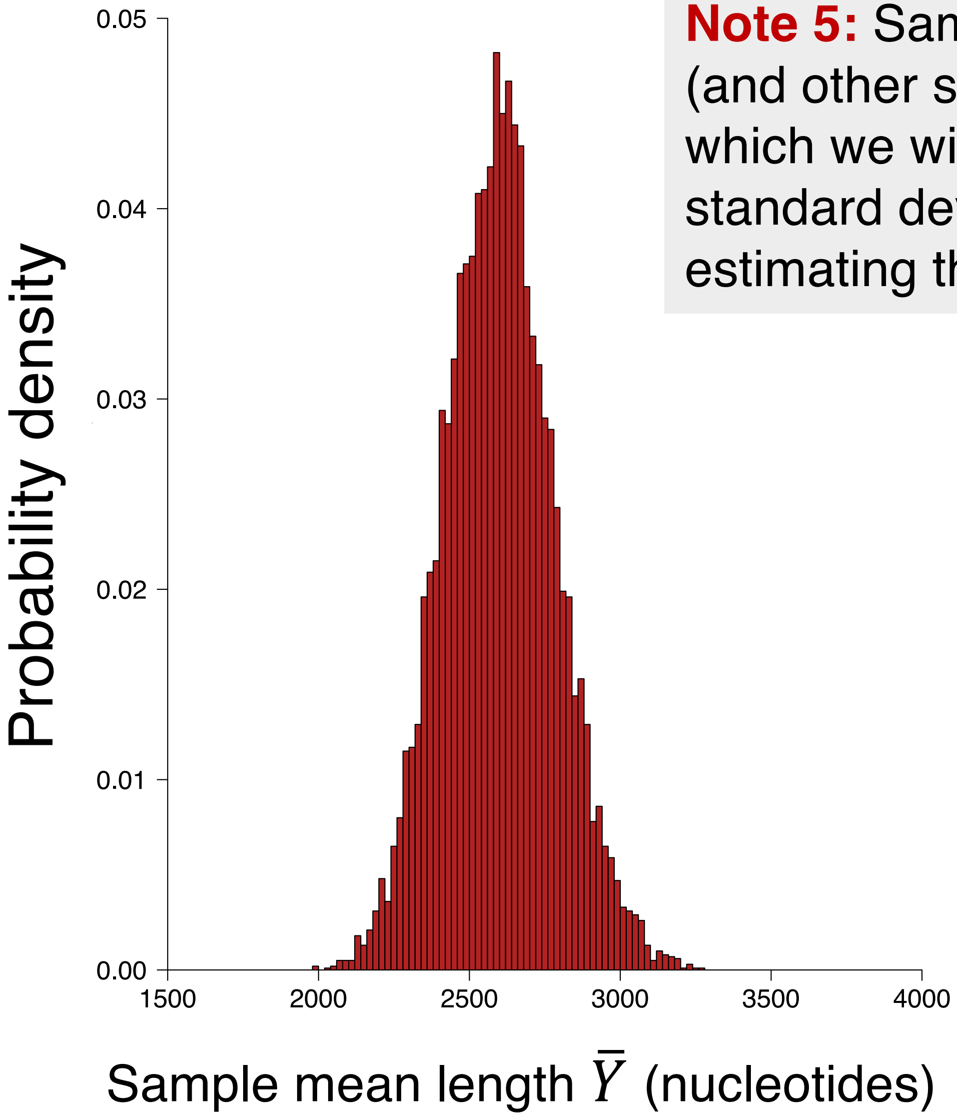


$\mu = 2622.0$

Note 4: The mean of all sample estimates equals the population mean (μ) and is perfectly centered on the true population mean.

This demonstrates that the sample mean ($\bar{Y}$) is an unbiased estimate of μ, assuming random sampling was performed, because on average, the sample mean equals the population mean.

The sampling distribution (probability distribution) of sample means ($\bar{Y}$)



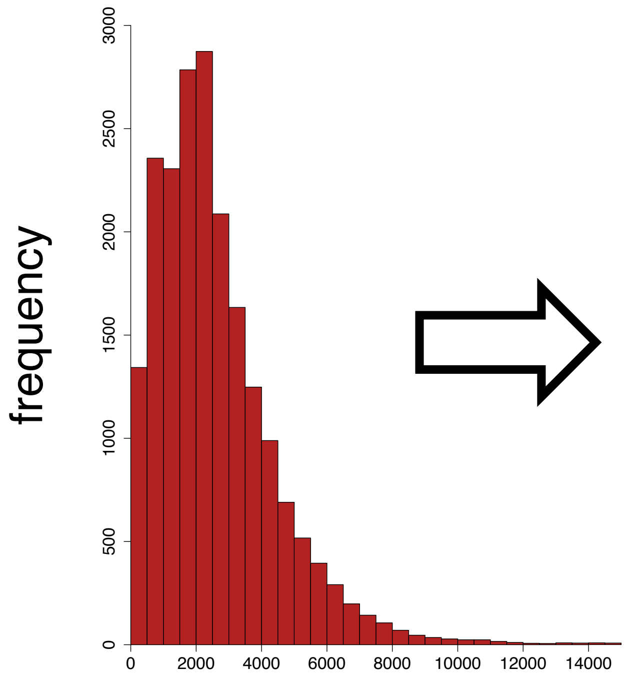
Note 5: Sample values for the standard deviation (and other statistics) also vary among samples, which we will discuss in our next lecture. The standard deviation of samples is crucial for estimating the uncertainty of a sample mean.

Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2544.8
Standard deviation	s	2125.3

Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2122.3
Standard deviation	s	2423.1

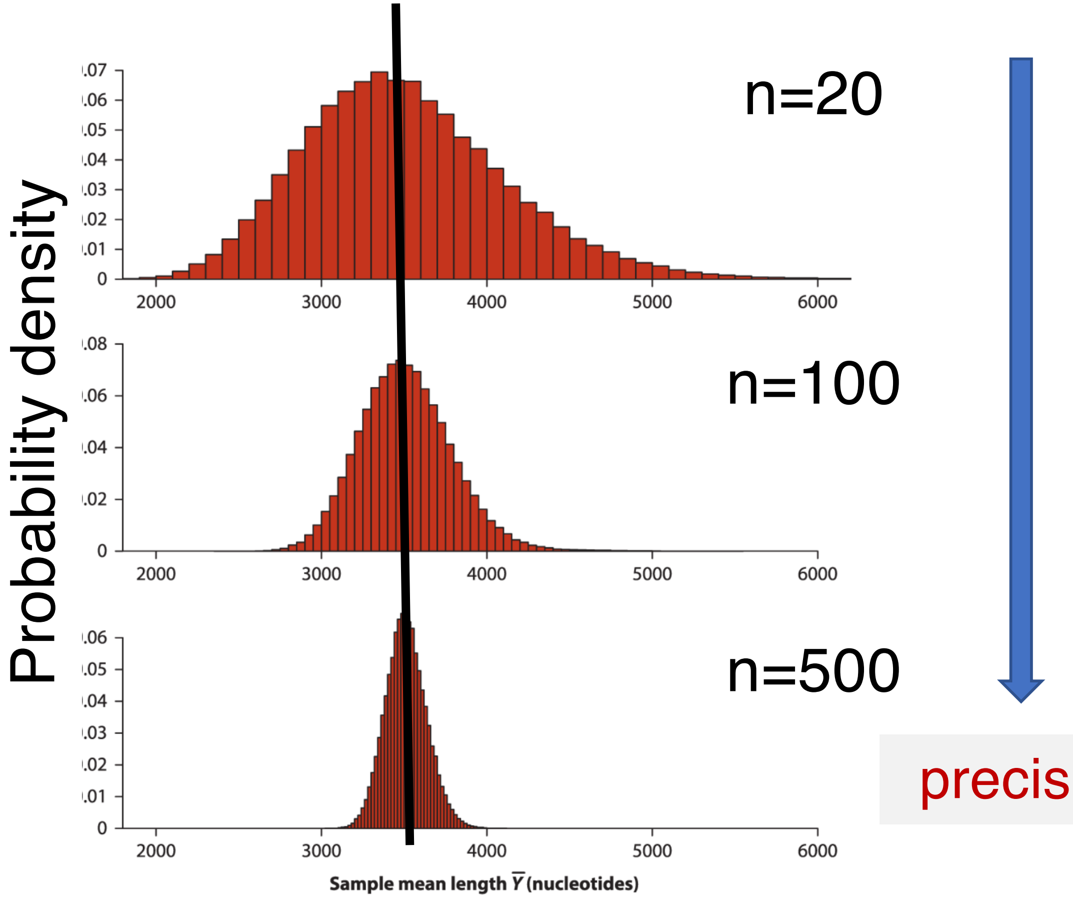
The effects of sample size (n) on the sampling distribution of sample means ($\bar{Y}$)

Frequency distribution of the gene length Population



Gene length
(number of nucleotides)

Sampling distributions for the sample means of the gene population (varying n)



precision

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Sample mean length $\bar{Y}$ (nucleotides)