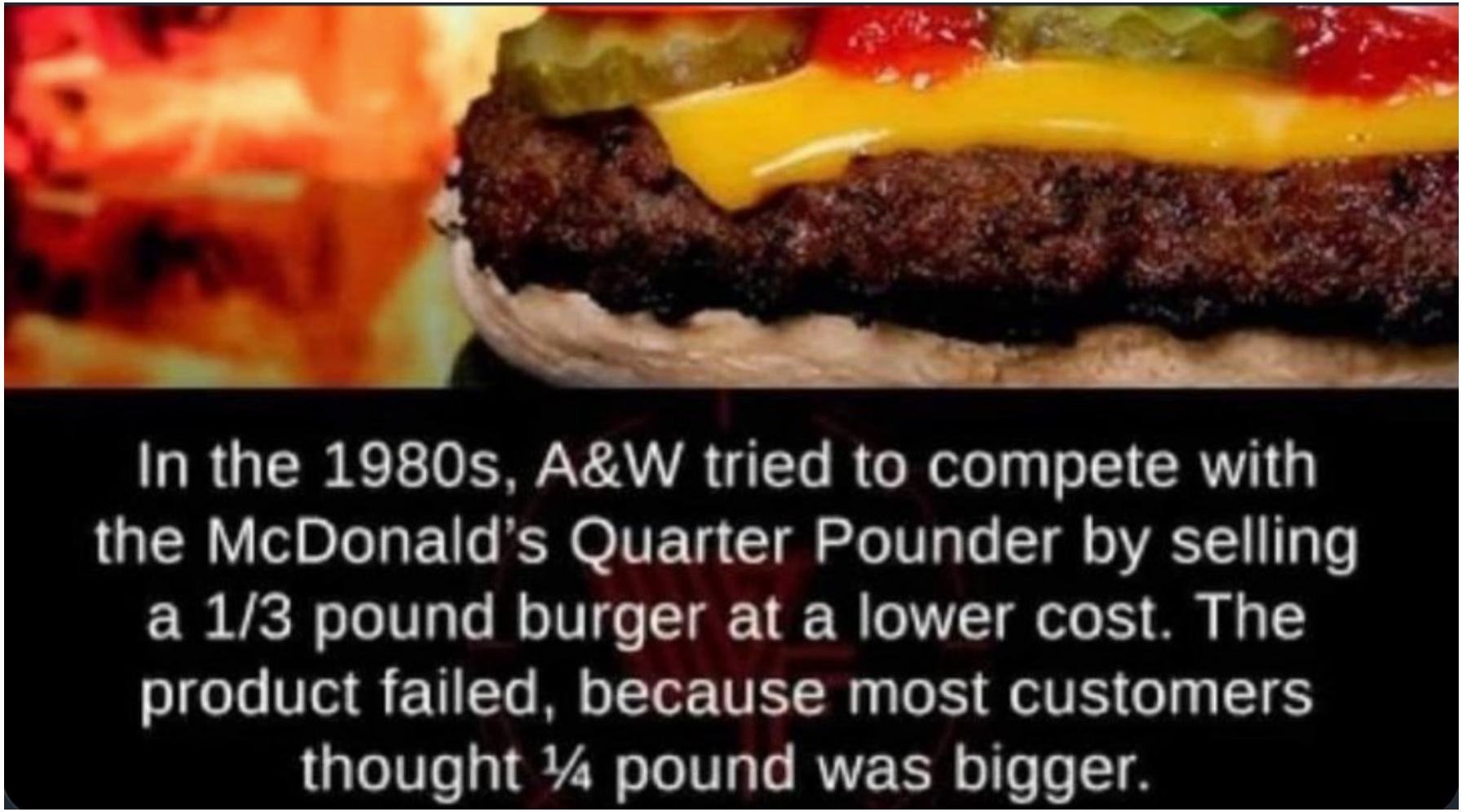
A snap demonstration of why numeracy is key to society

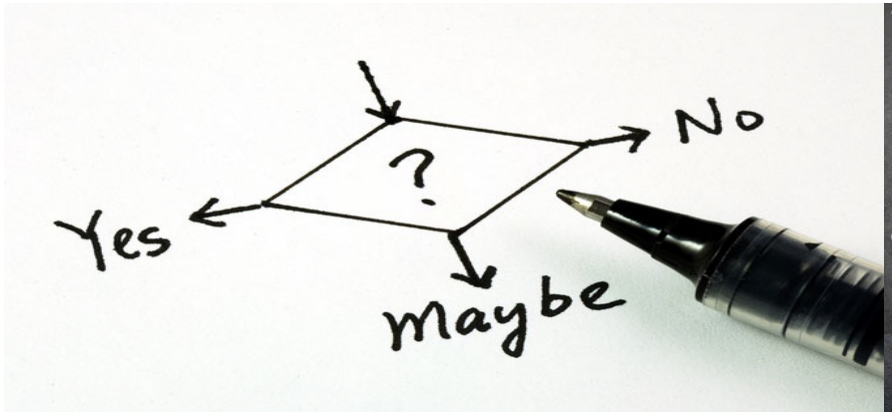


Lecture 8: Estimating with uncertainty, but with a degree of certainty (i.e., with some confidence).

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.

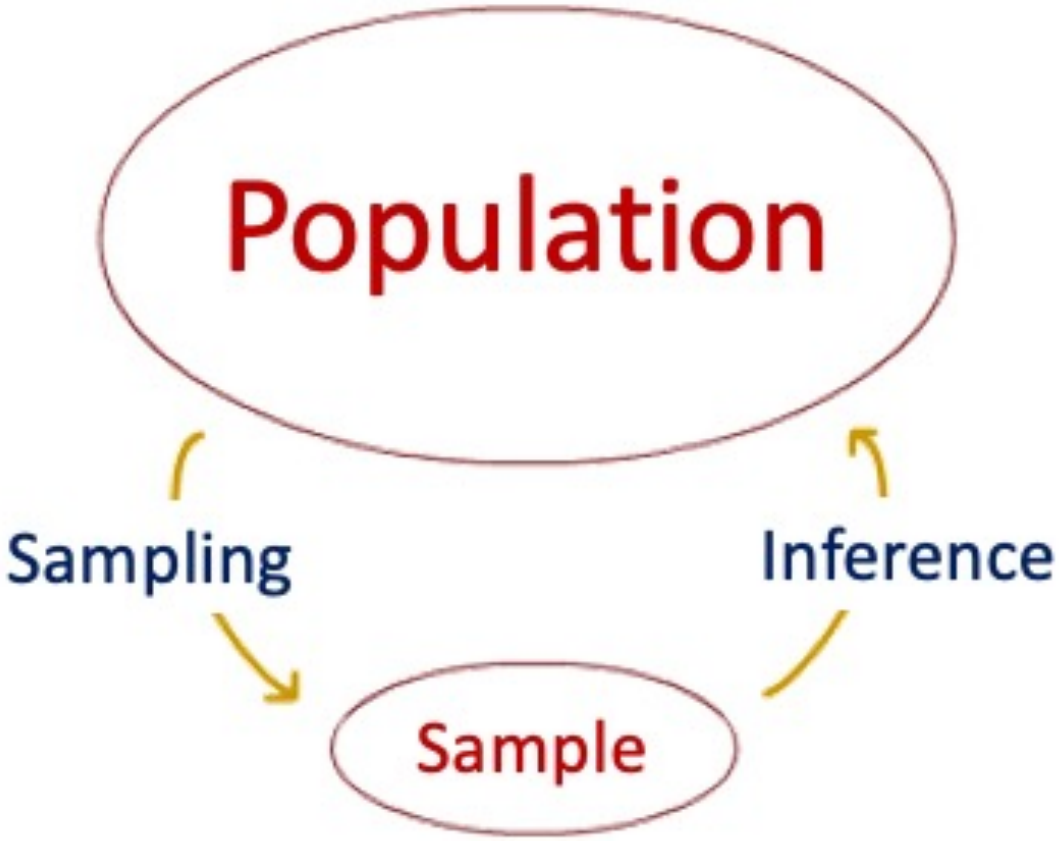
One primary goal of statistics is to estimate (infer) an unknown quantity (parameter) of a population based on sample data.

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

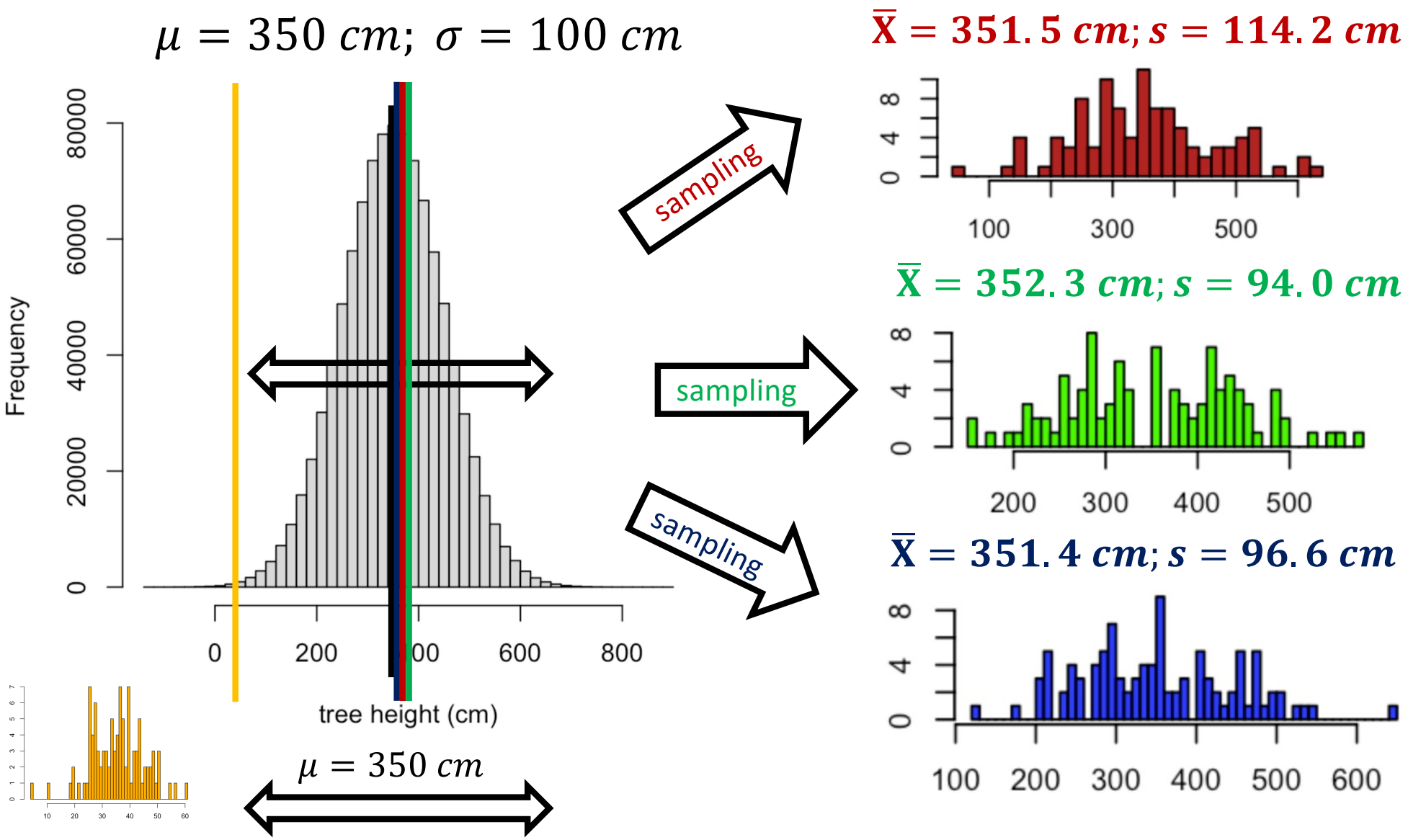
We use estimates to make decisions. Statistics is fundamentally the science of making decisions with incomplete knowledge, often using samples from populations of unknown sizes.

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)?



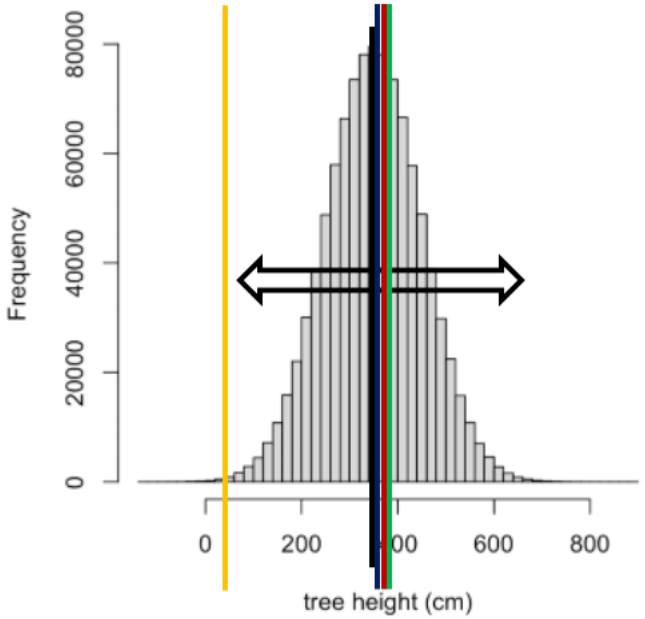
Sampling variation generates uncertainty



Uncertainty (samples means vary around the true population mean)

The variation within a sample, summarized by the standard deviation, informs us about the expected variability of sample means around the true population mean—that is, how much our sample average might deviate simply because of random sampling.

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

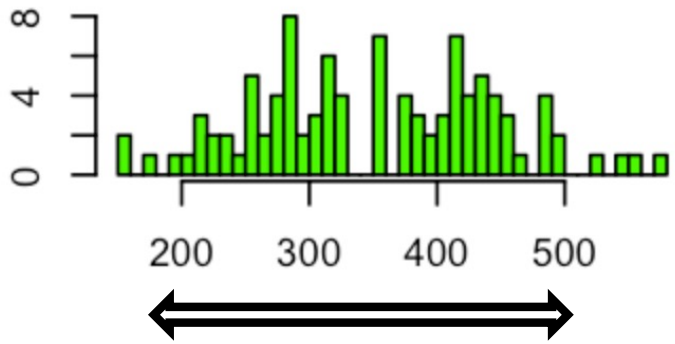


$\mu = 350\ cm$

Variation among samples

sampling

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



Variation within samples

Statistical “superpower”: By measuring variation within a sample, we can estimate how uncertain our estimate is—that is, how much sample means would fluctuate if we repeated the study.

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.

The Cognitive Discomfort of Statistical Thinking

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.

At its core, statistics asks: given uncertainty arising from random sampling, how reliable is an estimate, and how much confidence should we place in decisions based on it? Put differently, how close is a sample-based estimate expected to be to the true population value?

The goal is not to eliminate uncertainty (one can't), but to quantify it, so that we can make decisions with a known degree of certainty.

How can we estimate quantities for an entire population in the presence of uncertainty, while still expressing results with a meaningful degree of 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., the sample mean or sample standard deviation).

A sampling distribution is the probability distribution of an estimator generated by random sampling from a population. It describes what values the estimate would take if we were to repeatedly sample from the same population under identical conditions.

Although sampling distributions may resemble frequency distributions, they differ in a crucial way: sampling distributions describe probabilities of possible estimates, not frequencies of observed data values.

Important Statistical Symbols for Statistical Inference

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 for Statistical 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 represents the **population mean** of a variable (e.g., X), the symbol for the **sample mean** depends on the variable being measured. For example, the sample mean of X is written as $\bar{X}$, and the sample mean of Y is written as $\bar{Y}$. The key point is that the sample mean is always indicated by a bar over the variable, regardless of which variable is being considered.

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 samples with replacement for n = 2. Mirror duplicates (e.g., (1,2) and (2,1)) are shown only once for clarity, but their full contribution is accounted for in the sampling distribution.

(1,1) = 1.0	(1,2) = 1.5	(2,3) = 2.5	(3,4) = 3.5	(4,5) = 4.5
(2,2) = 2.0	(1,3) = 2.0	(2,4) = 3.0	(3,5) = 4.0	
(3,3) = 3.0	(1,4) = 2.5	(2,5) = 3.5		
(4,4) = 4.0	(1,5) = 3.0			
(5,5) = 5.0				

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.

[FIXED]

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 inferential bias by ensuring that estimates are not systematically shifted away from the true population value, and it allows sampling error—the random deviation of a sample estimate from the population value—to be quantified.

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.

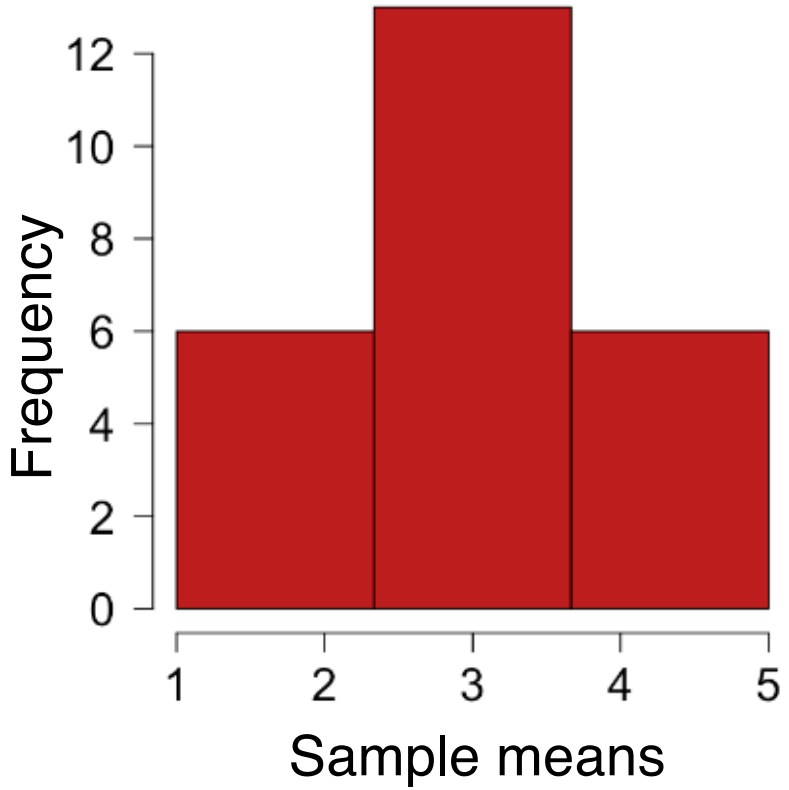
2) The selection of observational units in the population (e.g., individual tree) must be **independent**, i.e., the selection of any unit (e.g., individual tree) of the population must not influence the selection of any other unit.

When sampling is biased, differences between the sample estimate and the population value are no longer due solely to random variation—they reflect a systematic distortion of the sampling process.

Let’s take a break – 1 minute

Estimating with Uncertainty: The Sampling Distribution of the Mean

25 (5^2) possible different combinations of 2 numbers (i.e., $5^2 = 255$ different potential samples; with repetition (i.e., mirror duplicates) of observational units, i.e., (1,2),(2,1), etc) from 1,2,3,4,5 (population)



$\mu = 3$

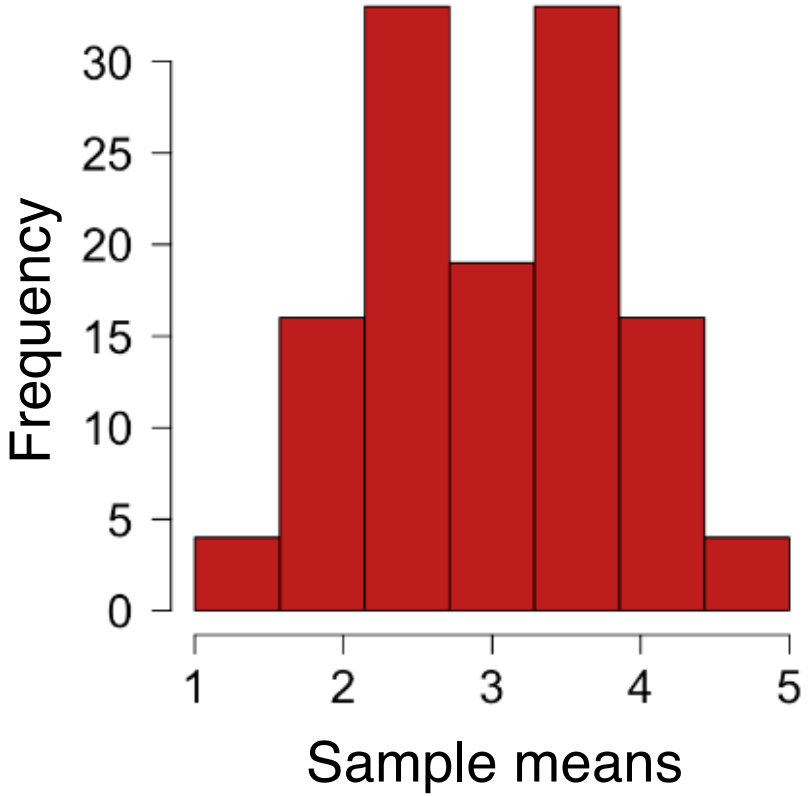
Mean of all samples means = 3.0

n = 2

$\mu\ (symbol\ for\ the\ population\ mean)$

Estimating with Uncertainty: The Sampling Distribution of the Mean

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



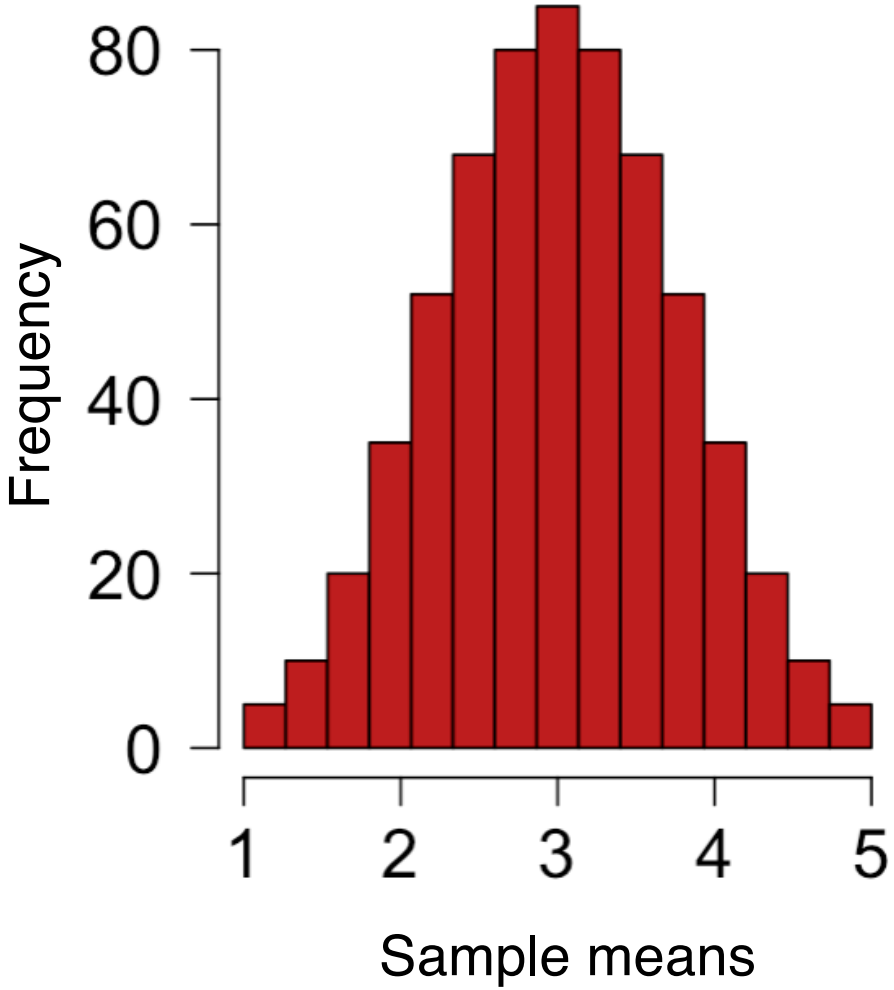
$\mu = 3$

Mean of all samples means = 3.0

n = 3

Estimating with Uncertainty: The Sampling Distribution of the Mean

625 possible different combinations of 4 numbers (i.e., $5^{4=}$ 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)



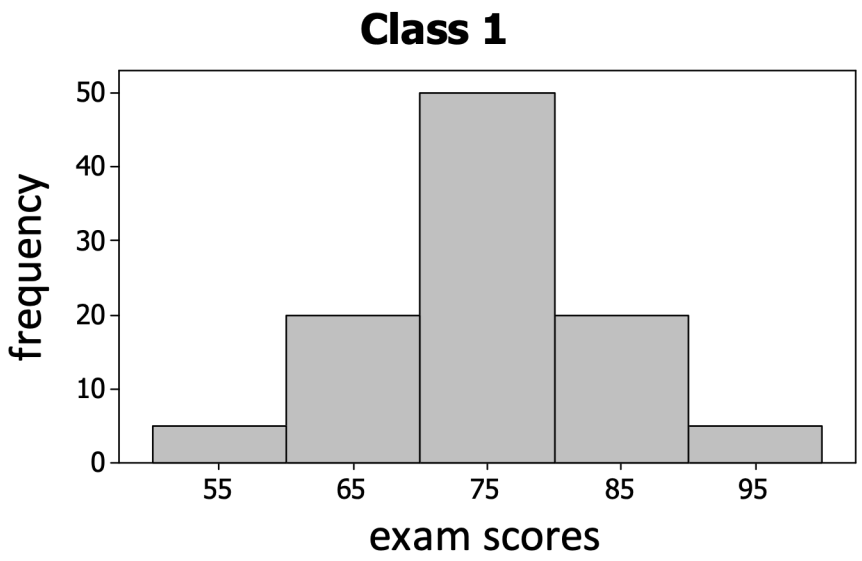
$\mu = 3$

Mean of all samples means = 3.0

n = 4

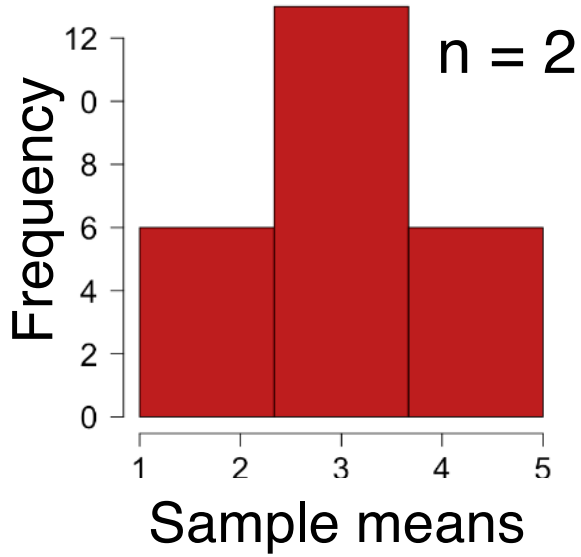
Remember Variability in frequency distributions?!!

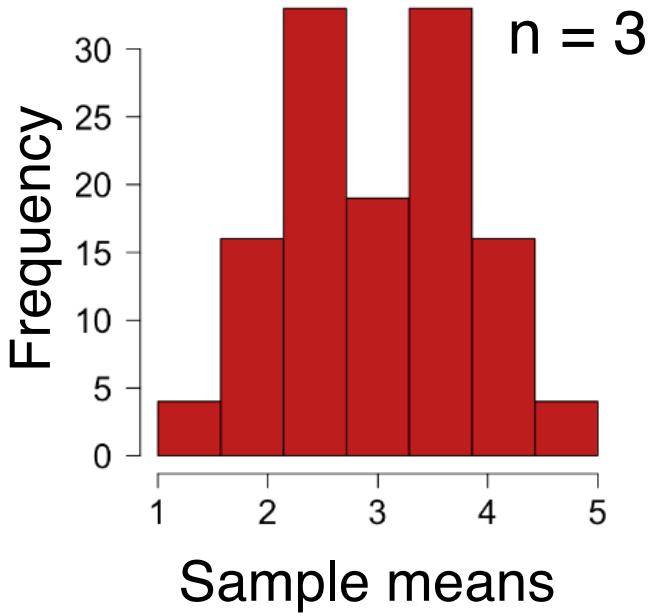
In which class do exam scores vary the most?

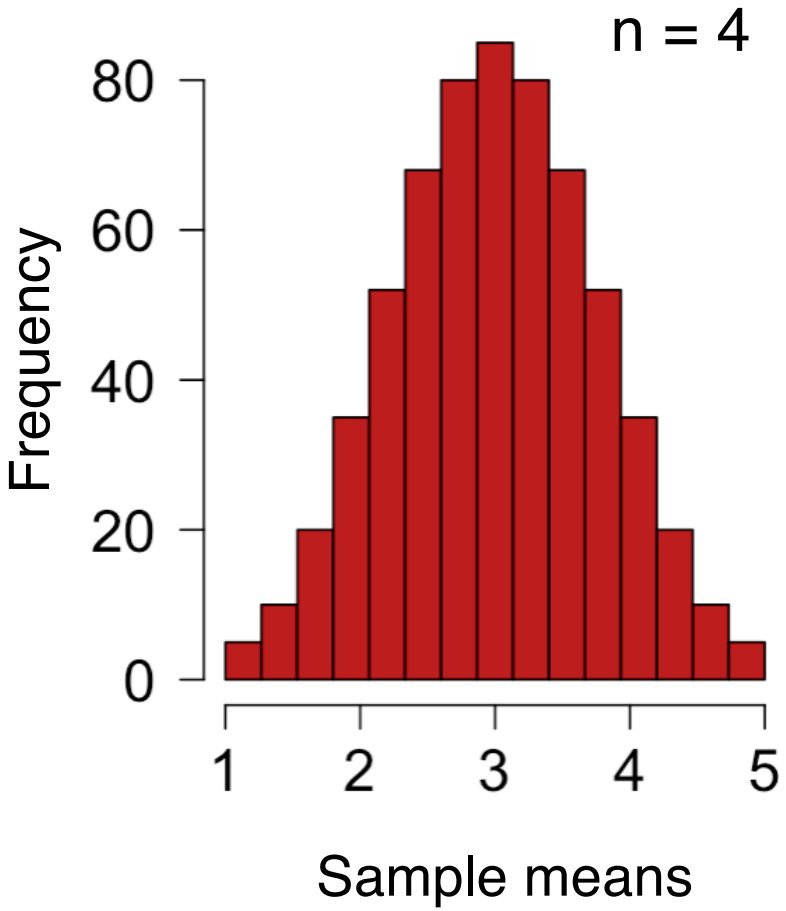


- Source: Cooper & Shore; Journal of Statistics Education (vol. 18, #2)

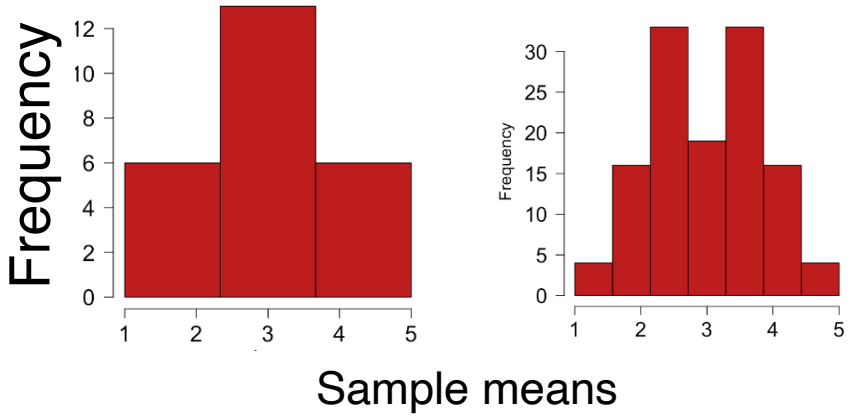
How does sample size affect the precision of sample estimates under random sampling?



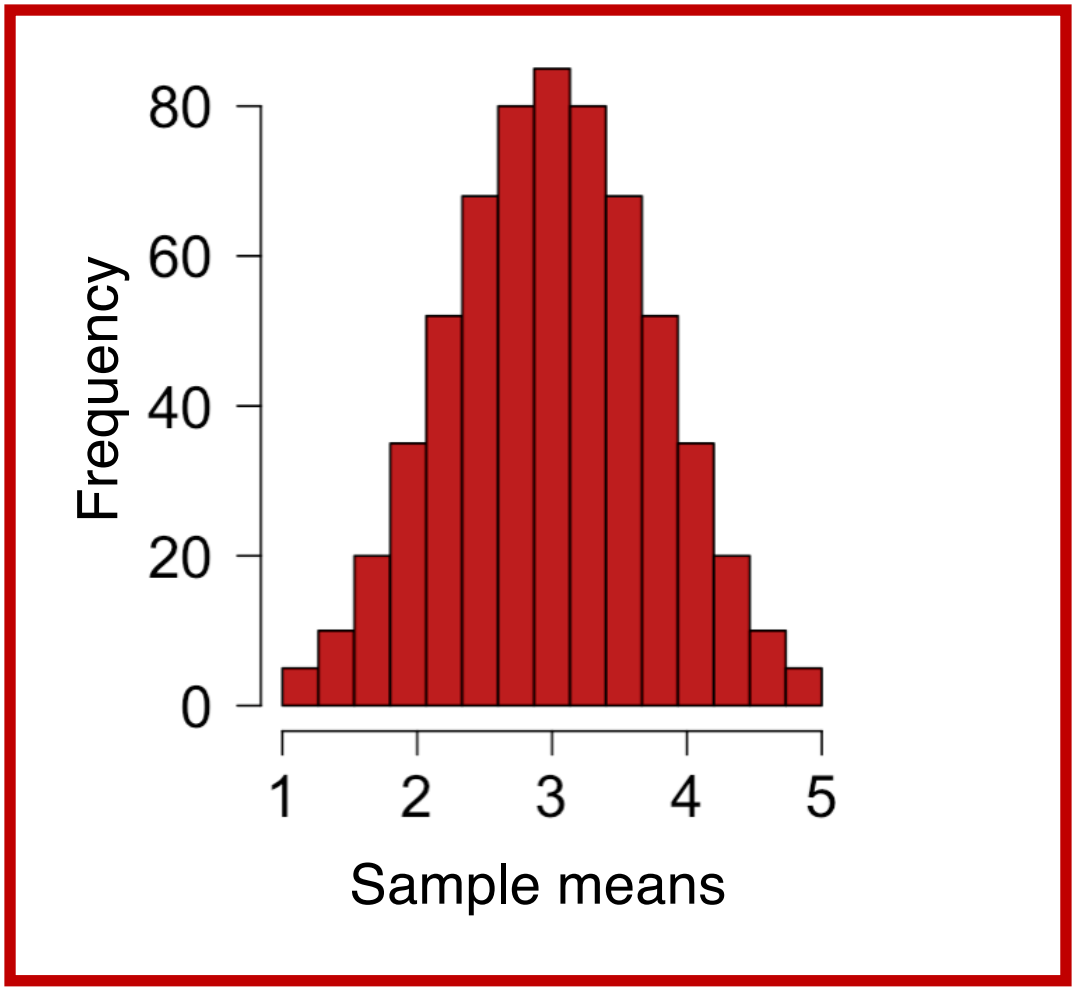




Larger samples produce narrower sampling distributions (smaller standard errors), even though the variability of the population itself does not change.

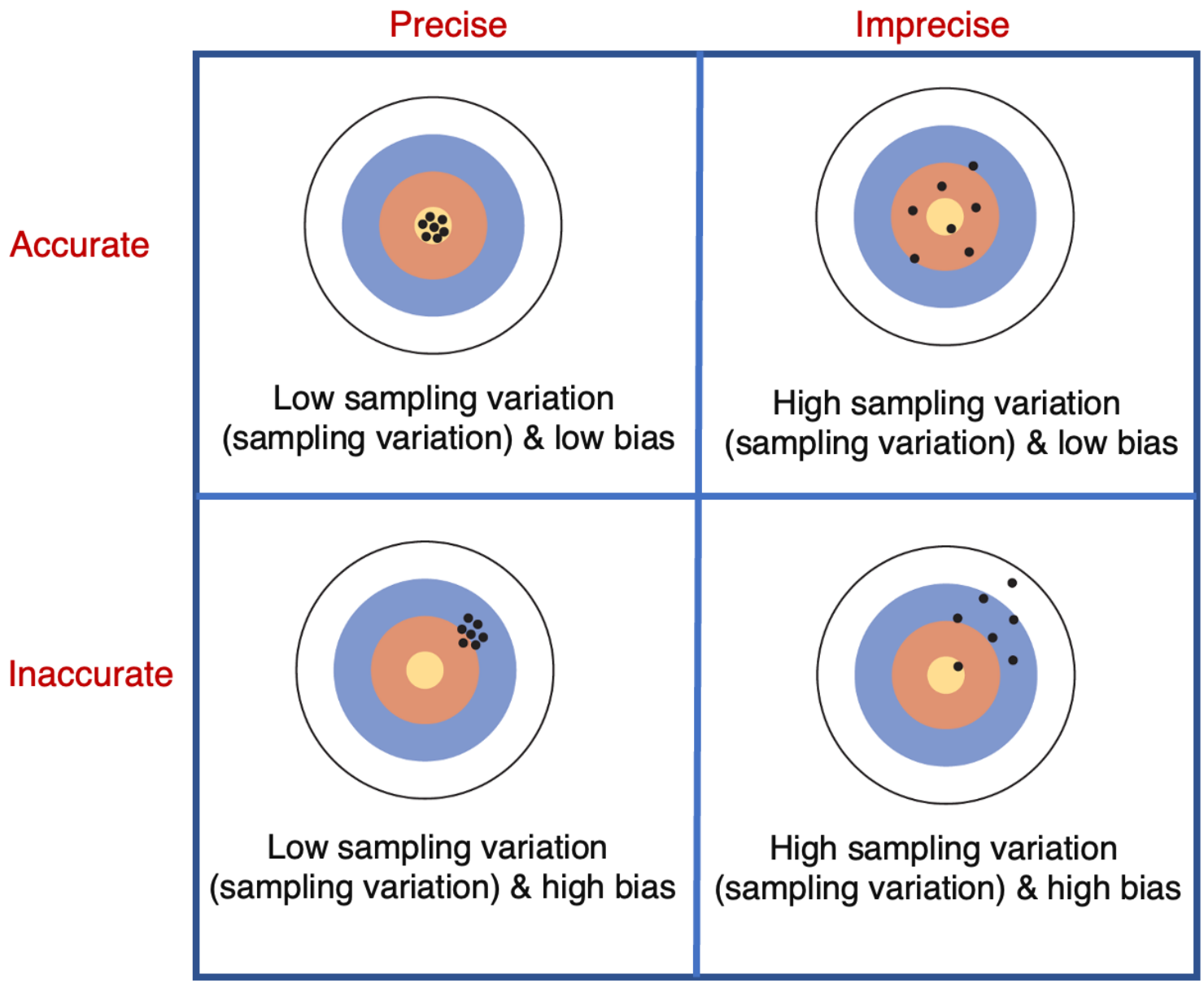


As sample size increases, sampling variability decreases, so sample means cluster more tightly around the true population mean, yielding more **precise** estimates.



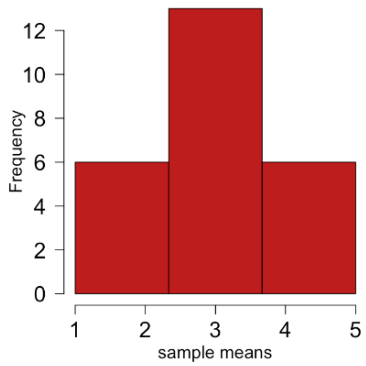
Under random sampling, the sample mean is unbiased—meaning that, although any single sample mean may differ from the population mean, the mean of all possible sample means equals the true parameter.

Random sampling does not eliminate sampling error, but it prevents systematic bias and allows sampling error to be measured (covered in the next lectures).

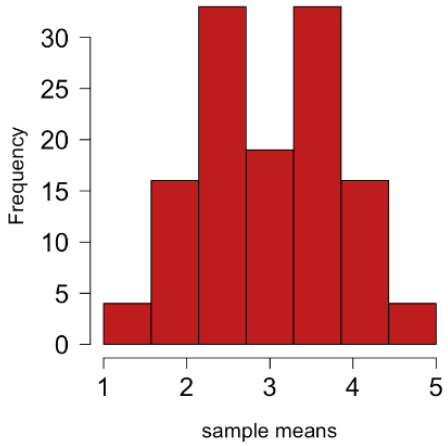


All sample sizes here produce unbiased estimates, but which one yields greater precision?

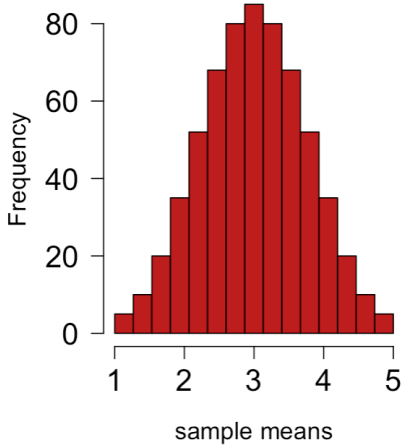
n = 2

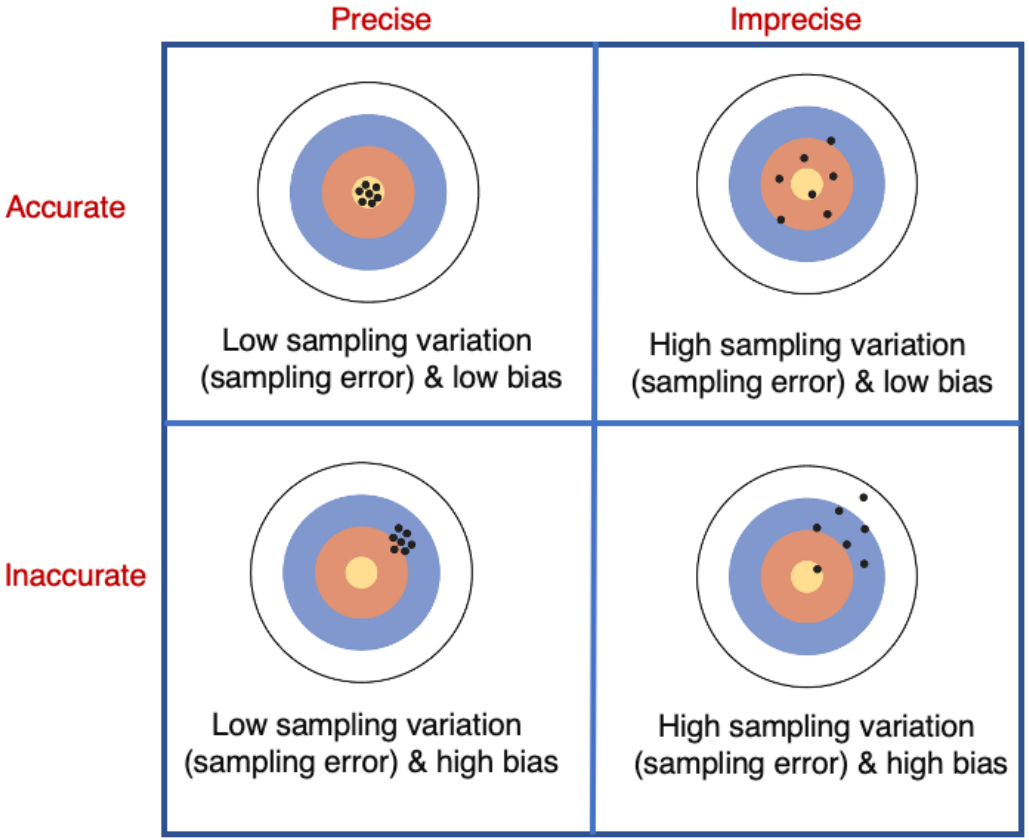


n = 3



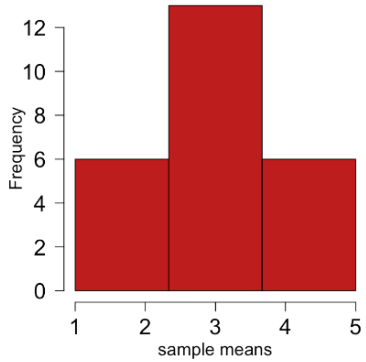
n = 4





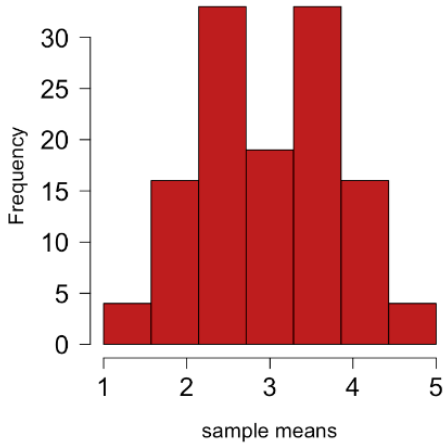
Properties of sampling distributions

n = 2



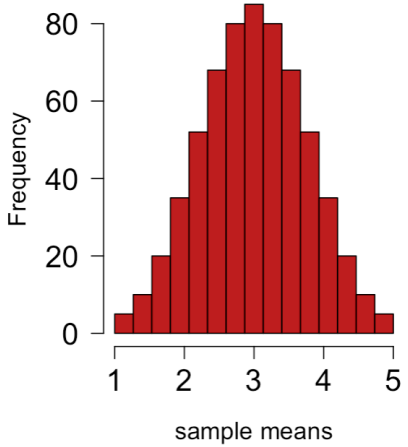
All sample sizes here produce unbiased estimates, but which one yields greater precision?

n = 3



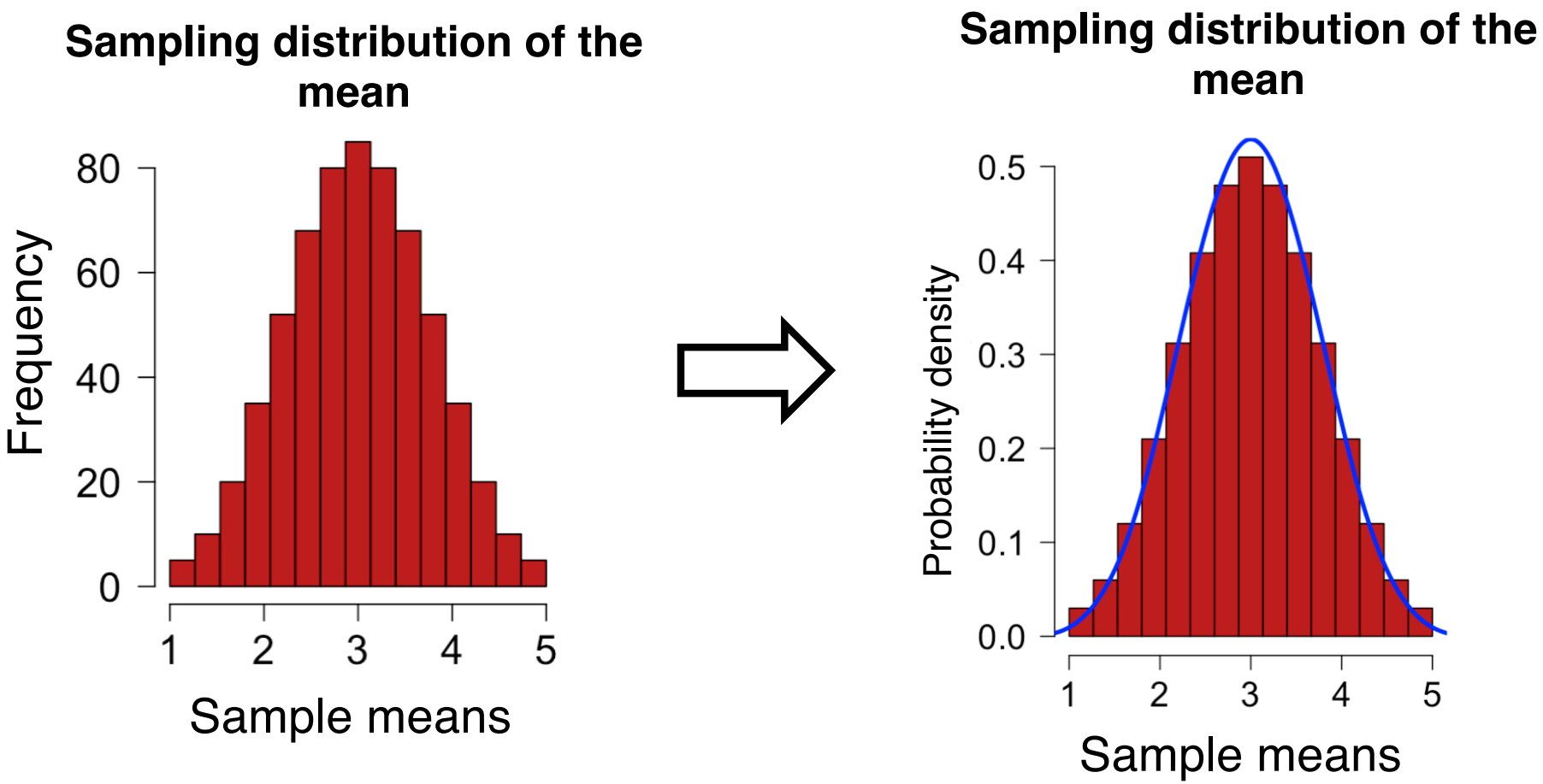
Property 2: Under random sampling, increasing sample size reduces sampling variability, so sample means cluster more tightly around the true population mean, resulting in greater precision.

n = 4



Sampling distributions are best represented by probability density functions (PDFs) when the estimator is continuous, as is the case for sample means except in trivial or extremely small populations.

Probability density does not give the probability of a single exact value, but rather how densely probability is concentrated around that value. Probabilities are obtained by integrating the PDF over an interval.



Sampling distributions are best represented by probability density functions (PDFs) when the estimator is continuous, as is the case for sample means except in trivial or extremely small populations.

Example:

Imagine repeatedly taking random samples from a population and calculating the sample mean each time.

Sample means near the true population mean occur **often** → they have **high probability density**.

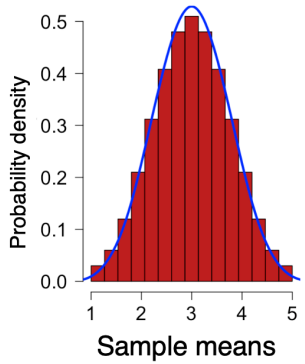
Sample means far from the population mean occur **rarely** → they have **low probability density**.

If you plot all possible sample means and how frequently they occur, you get a smooth curve. That curve is the **probability distribution** of the sample mean.

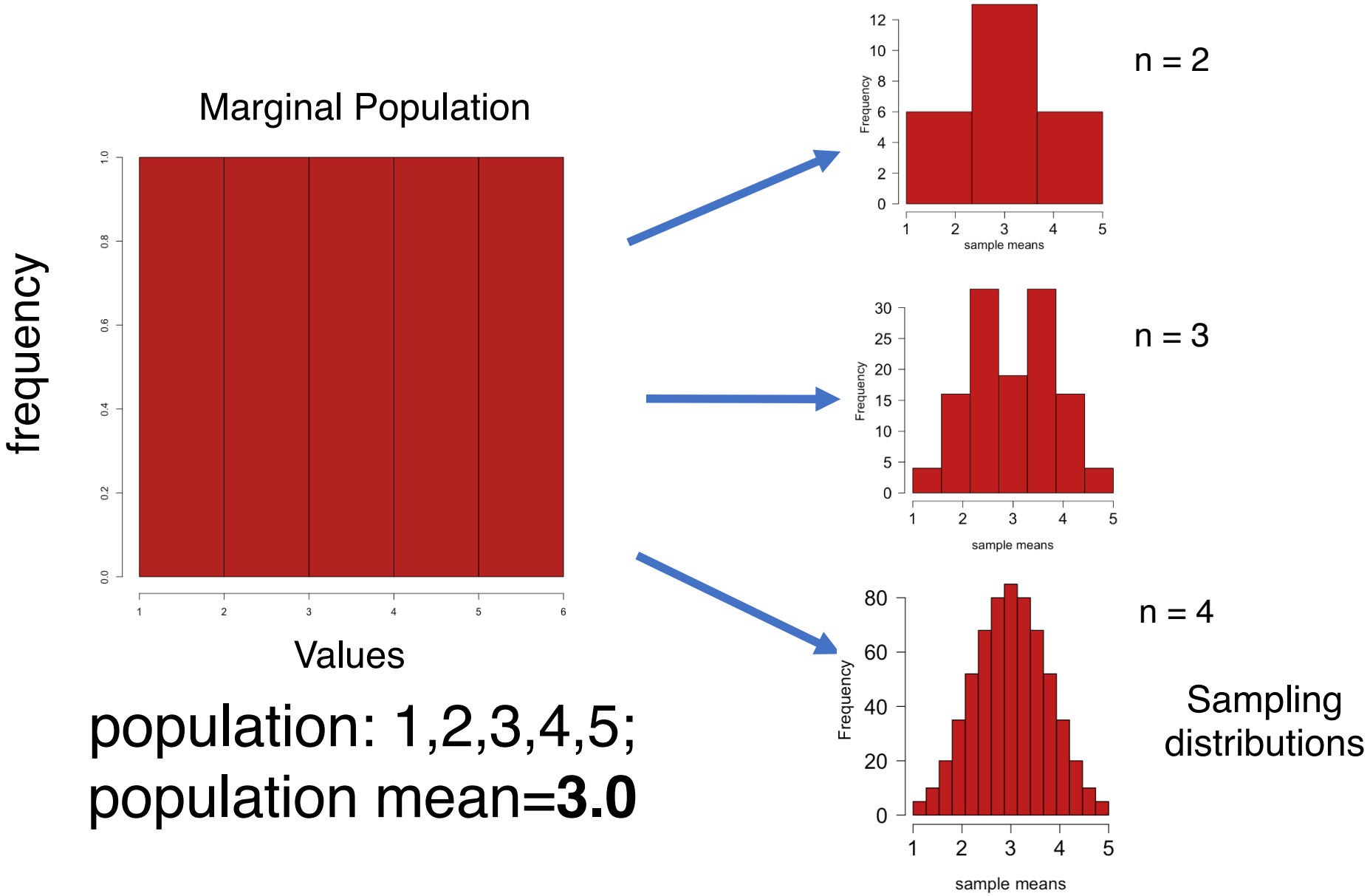
Now the key idea:

The **height of the curve** at a value tells you the probability *density* (how common nearby values are).

The **area under the curve over an interval** gives you an actual probability.



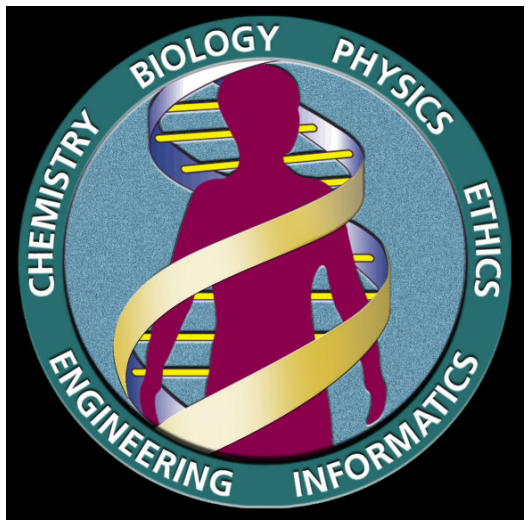
Critical: The shape of a population's frequency distribution (often called its marginal distribution) is not necessarily the same as the distribution of sample-based estimates drawn from that population (e.g., the sampling distribution of the mean).

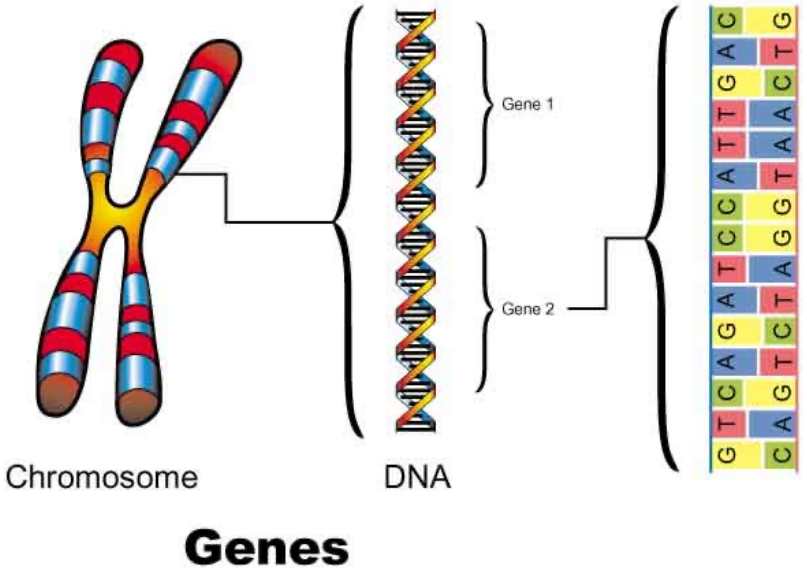


Let’s take a break – 1 minute

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.

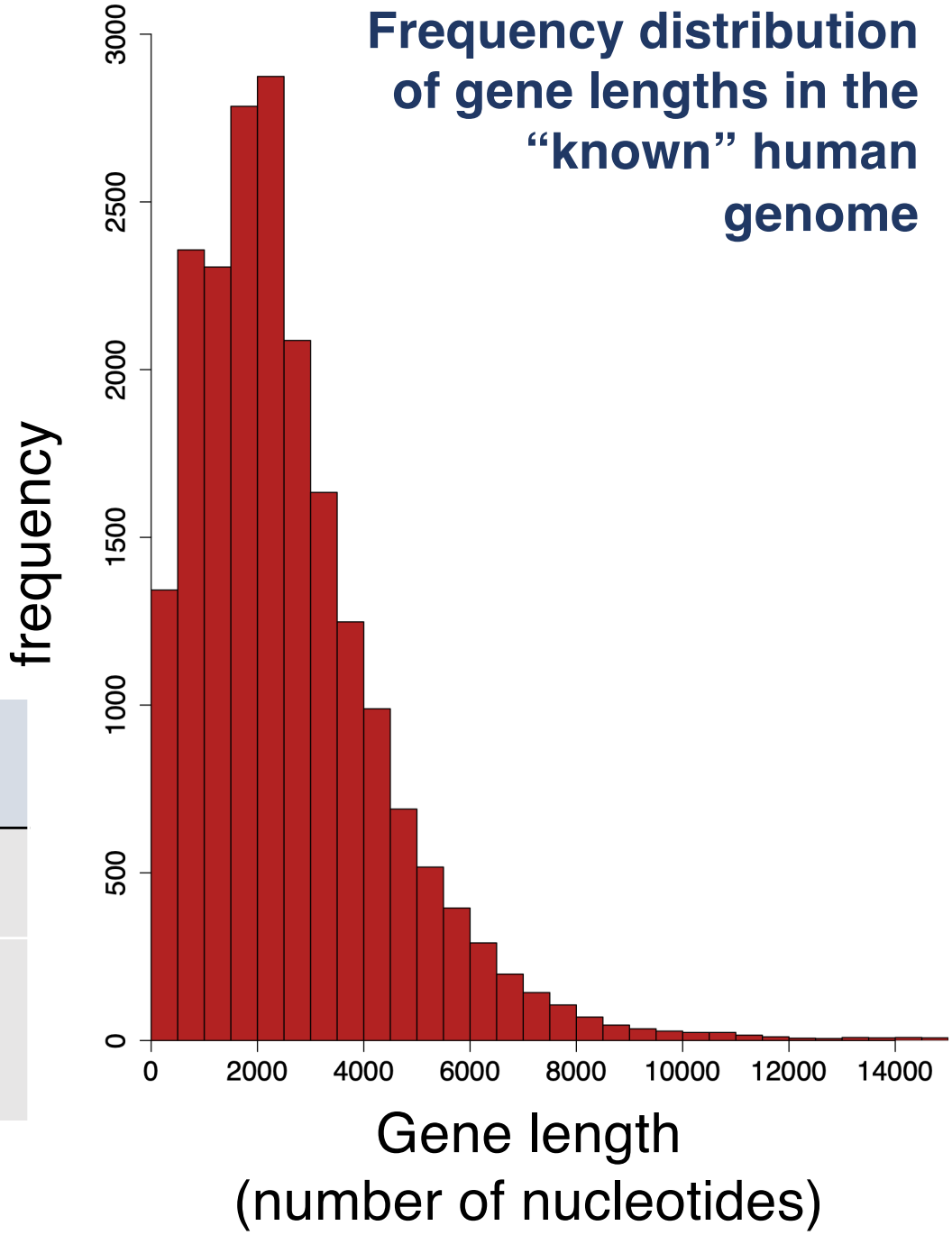




The length of human genes

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

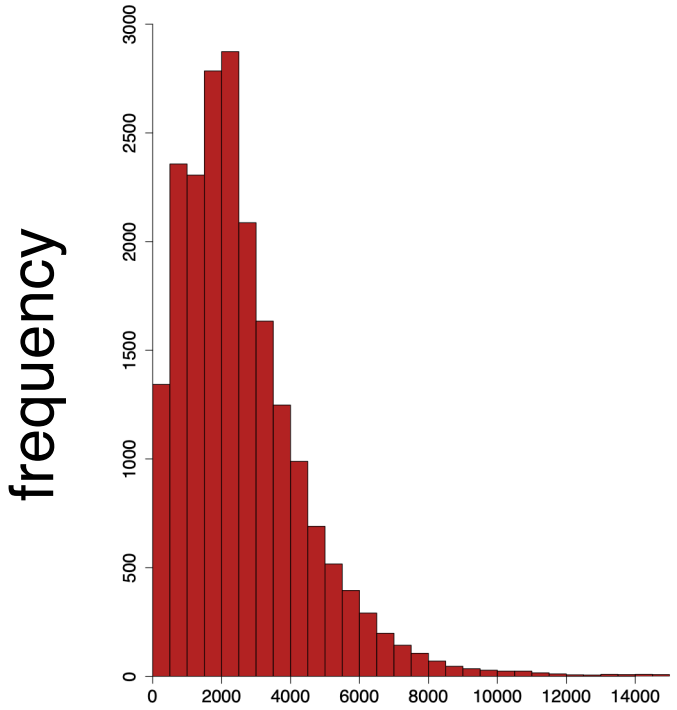
Names	Parameter	Value (nucleotides)
Mean (mu)	μ	2622.0
Standard deviation (sigma)	σ	2036.9



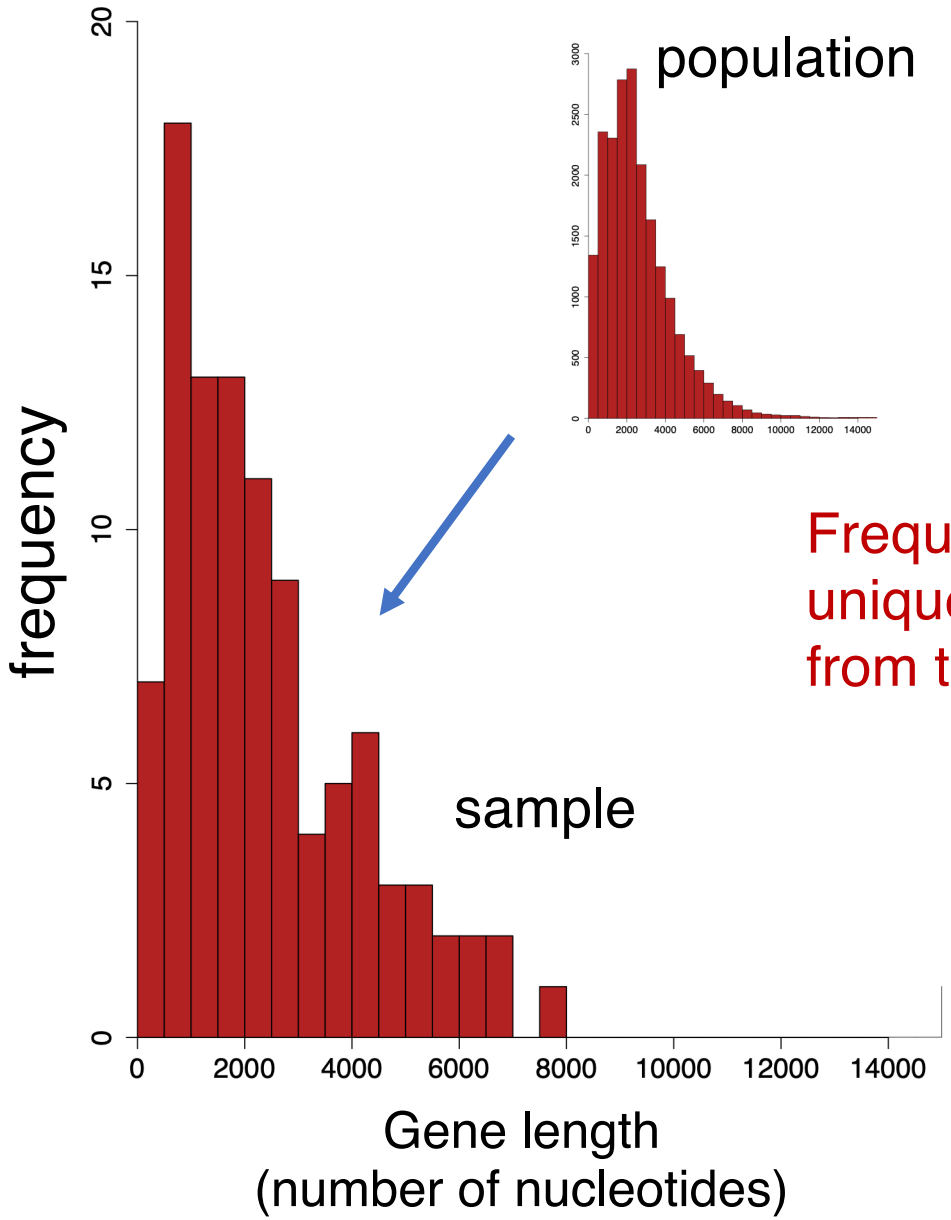
In real-world studies, we rarely know the true parameter values of the population. In this case, however, we (almost) do!

We can therefore use this gene population to illustrate the key ideas behind sampling, uncertainty, accuracy, and precision—and, ultimately, how we estimate under uncertainty while still expressing results with a quantified 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)

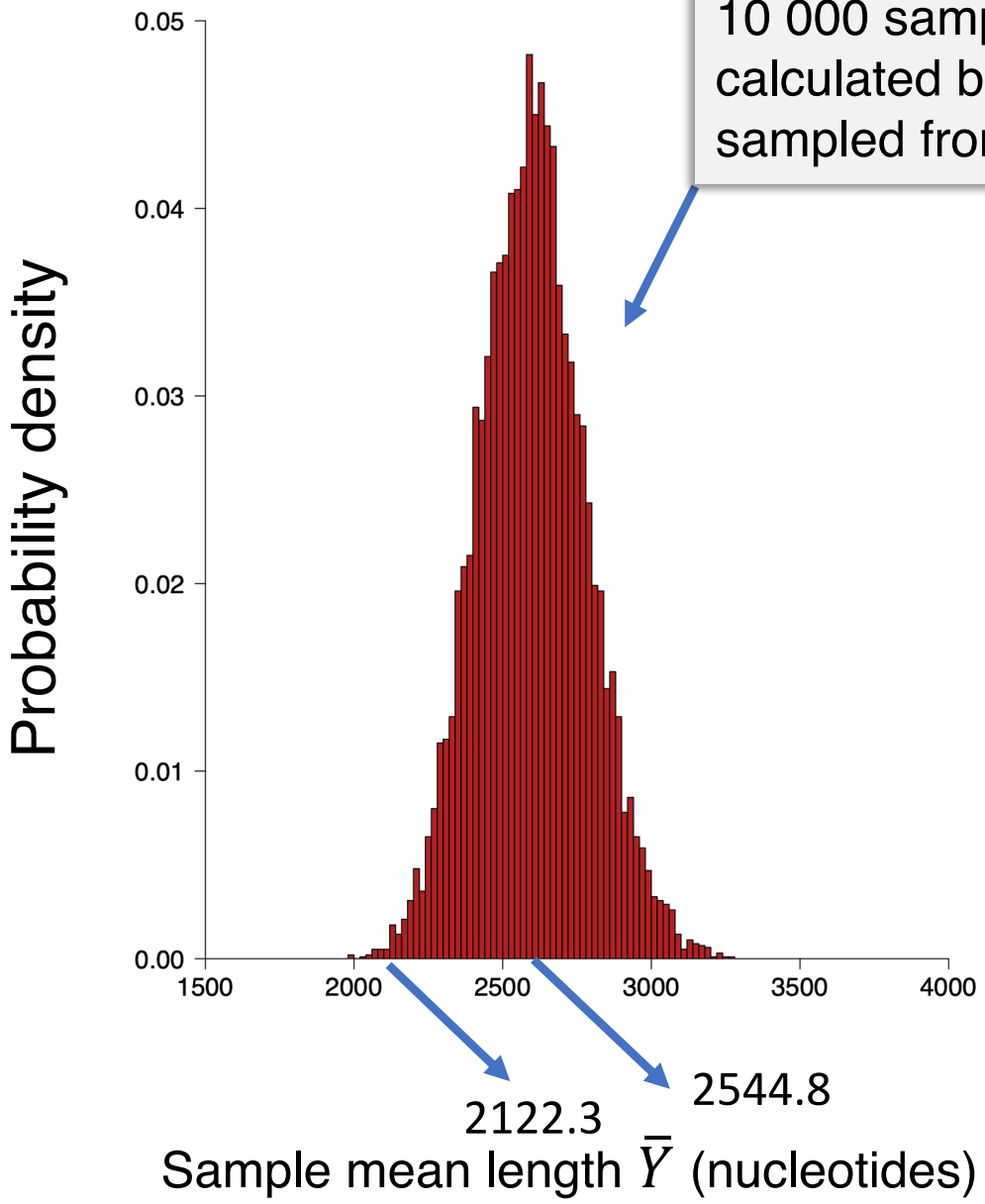


Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2544.8
Standard deviation	s	2125.3

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

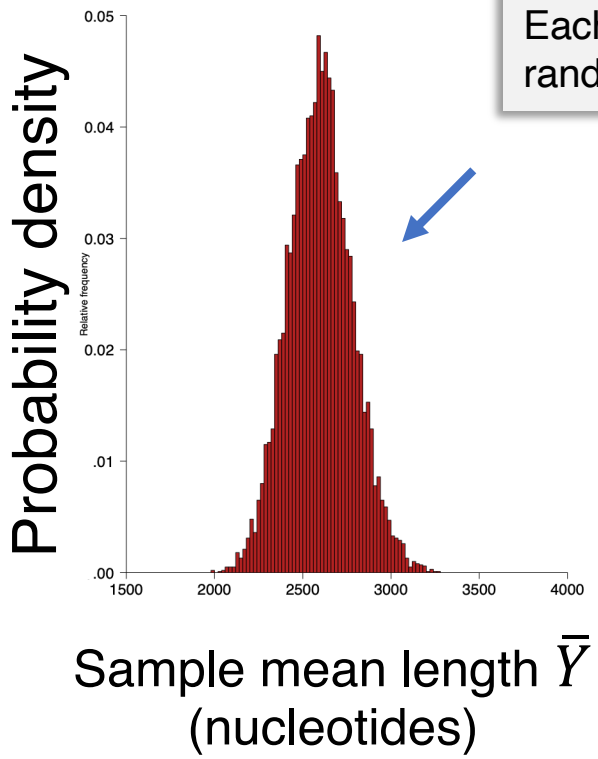
Names	Parameter	Value (nucleotides)
Mean	μ	2622.0
Standard deviation	σ	2036.9

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. This difference is not surprising: sample estimates rarely equal the population parameter exactly, as deviations due to random sampling variation are virtually inevitable.

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



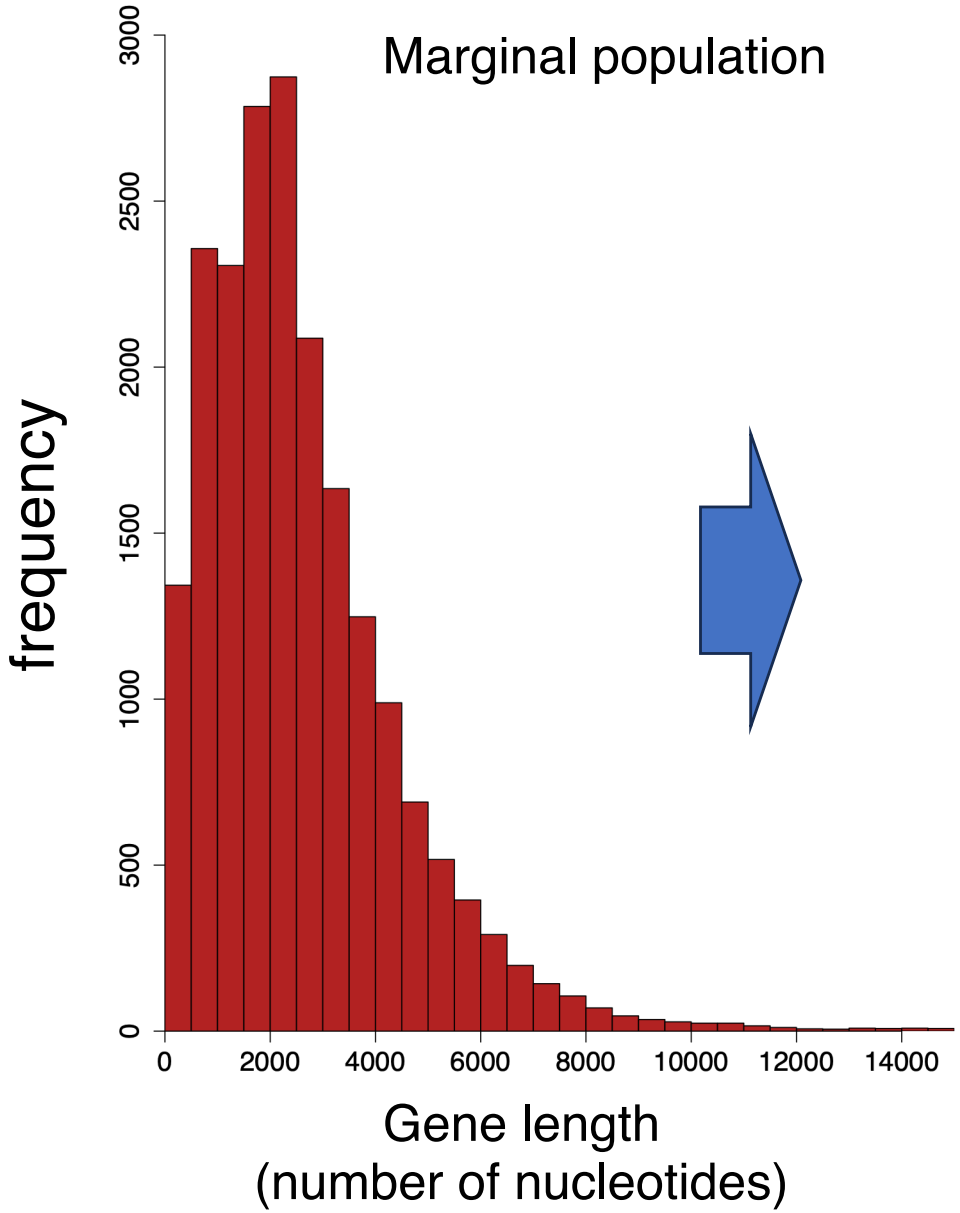
Sampling distribution of means based on 10 000 sample mean values. Each sample mean is calculated based on the lengths of 100 genes randomly sampled from the population of 20,290 genes.

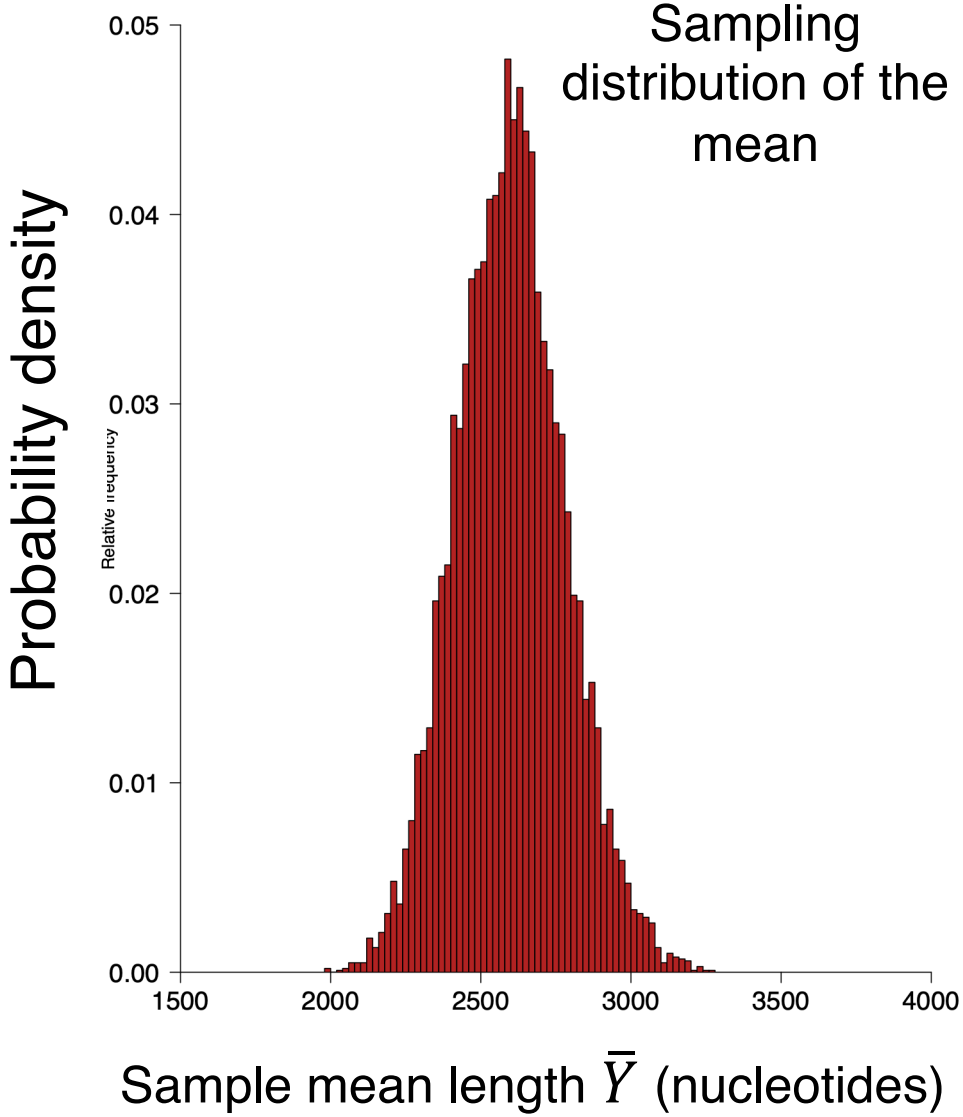
Here, 10,000 sample means were generated from the population using a computational (simulation-based) approach.

In statistics, however, sampling distributions are typically derived analytically using calculus-based methods. These approaches allow us to characterize the probability distribution of all possible sample means for a given sample size (whether based on 100 genes or any other number) without explicitly simulating them.

This analytical framework is essential historically and conceptually: most foundational probability distributions were developed long before computers existed, often more than a century ago.

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

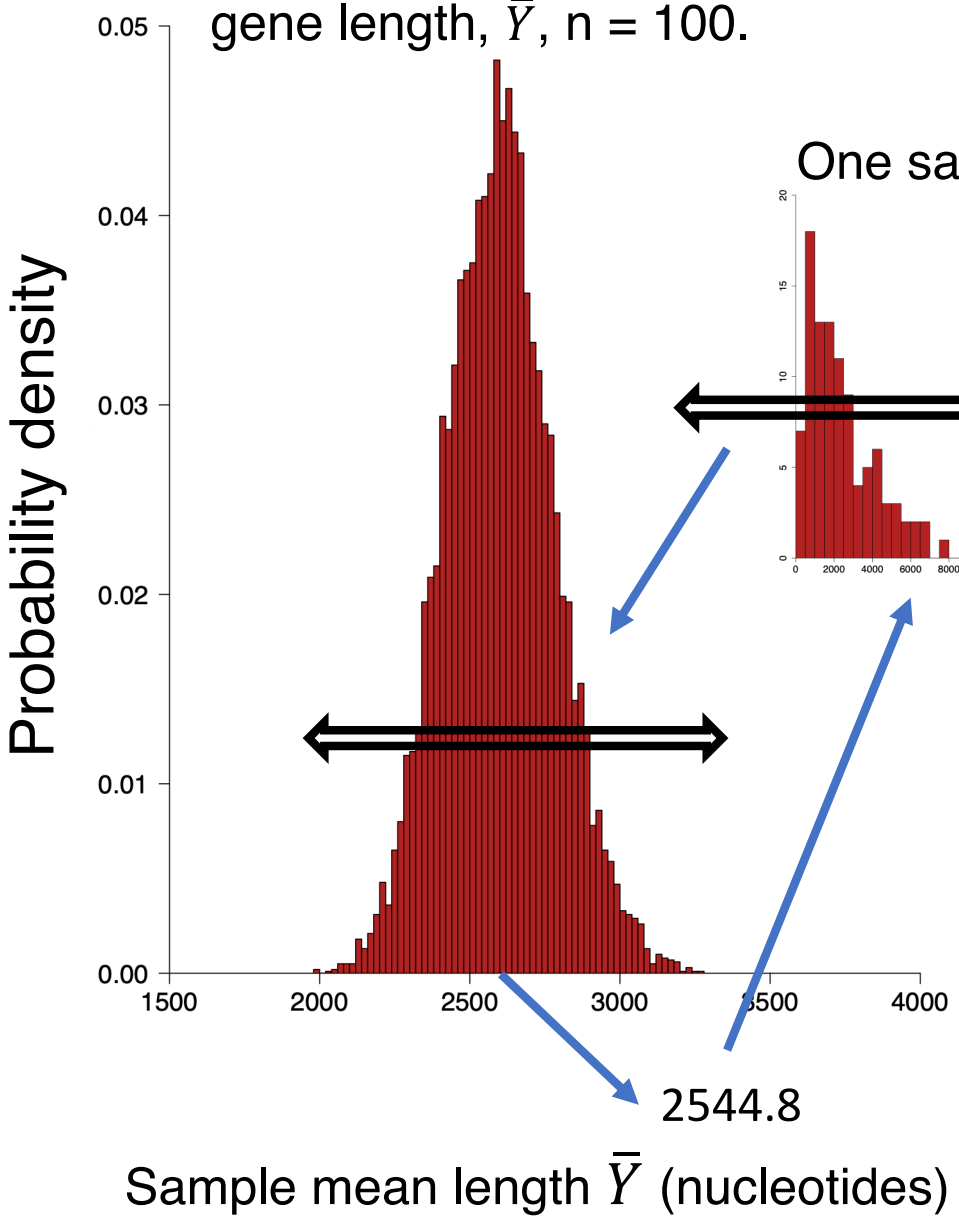




[FIXED]

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

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



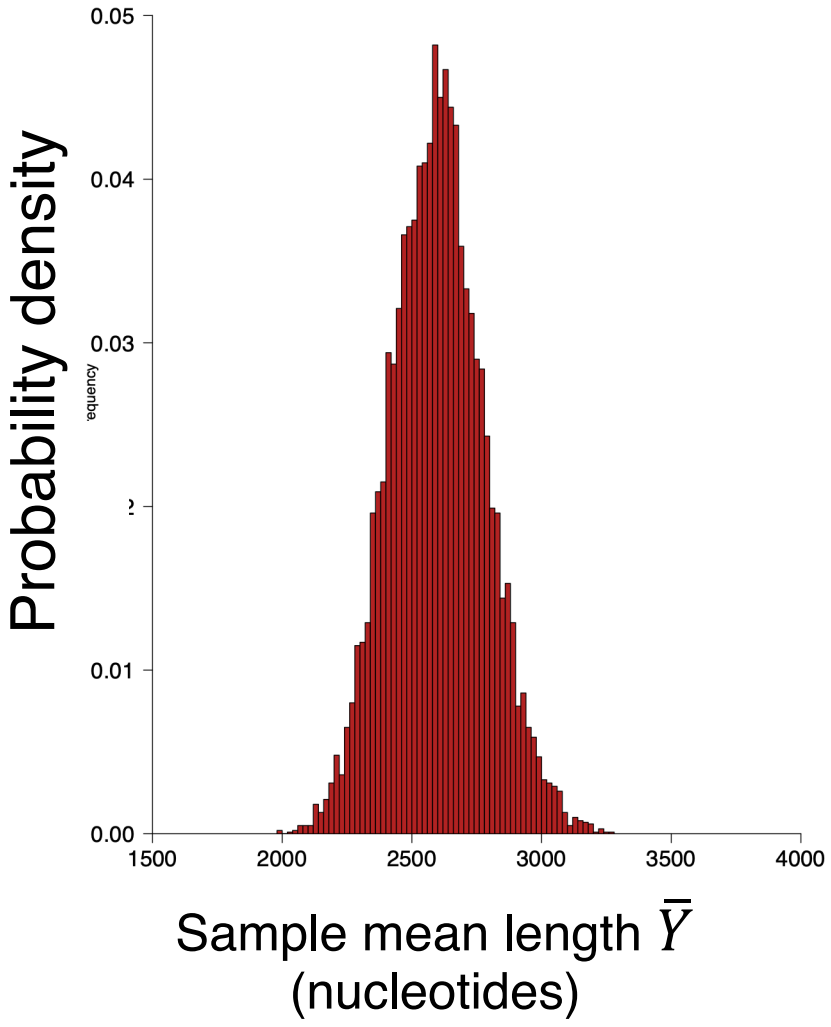
One sample (n = 100).

Statistical insight: The variation within a single sample can be used to estimate the uncertainty of sample-based estimates across all possible samples drawn 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

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



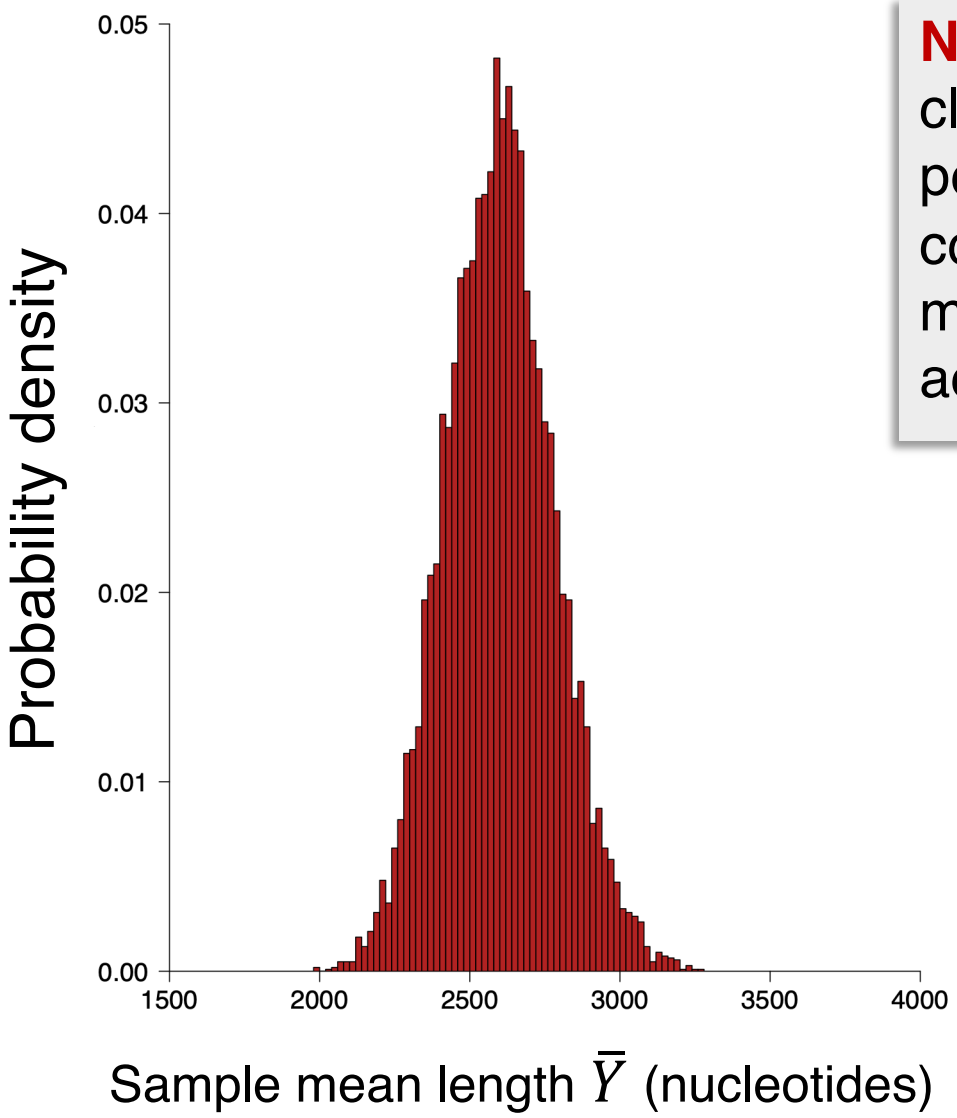
Note 1: We typically work with a single sample, which gives us just one sample mean value ($\bar{Y}$).

Understanding how sampling distributions are constructed is essential for quantifying uncertainty—that is, for understanding how sample means vary from one sample to another—and for assessing how much confidence we should place in inferences based on those samples.

When sample estimates show high variability across samples, our confidence in any single estimate is lower; when variability is smaller, our estimates are more precise and our confidence increases.

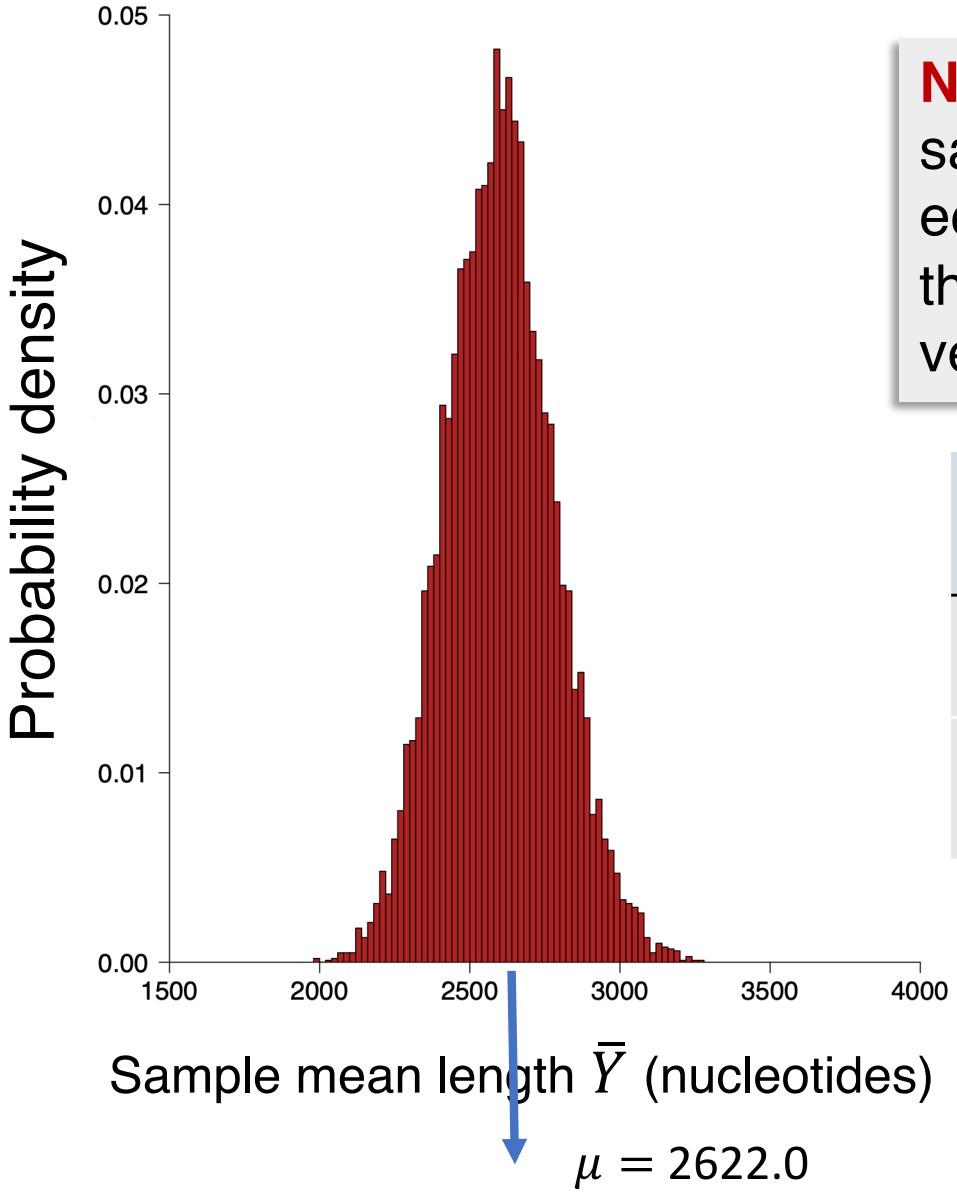
Crucially, the variation observed within a single sample contains information about the variation expected among samples, a result we will explore in upcoming lectures.

The sampling 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.

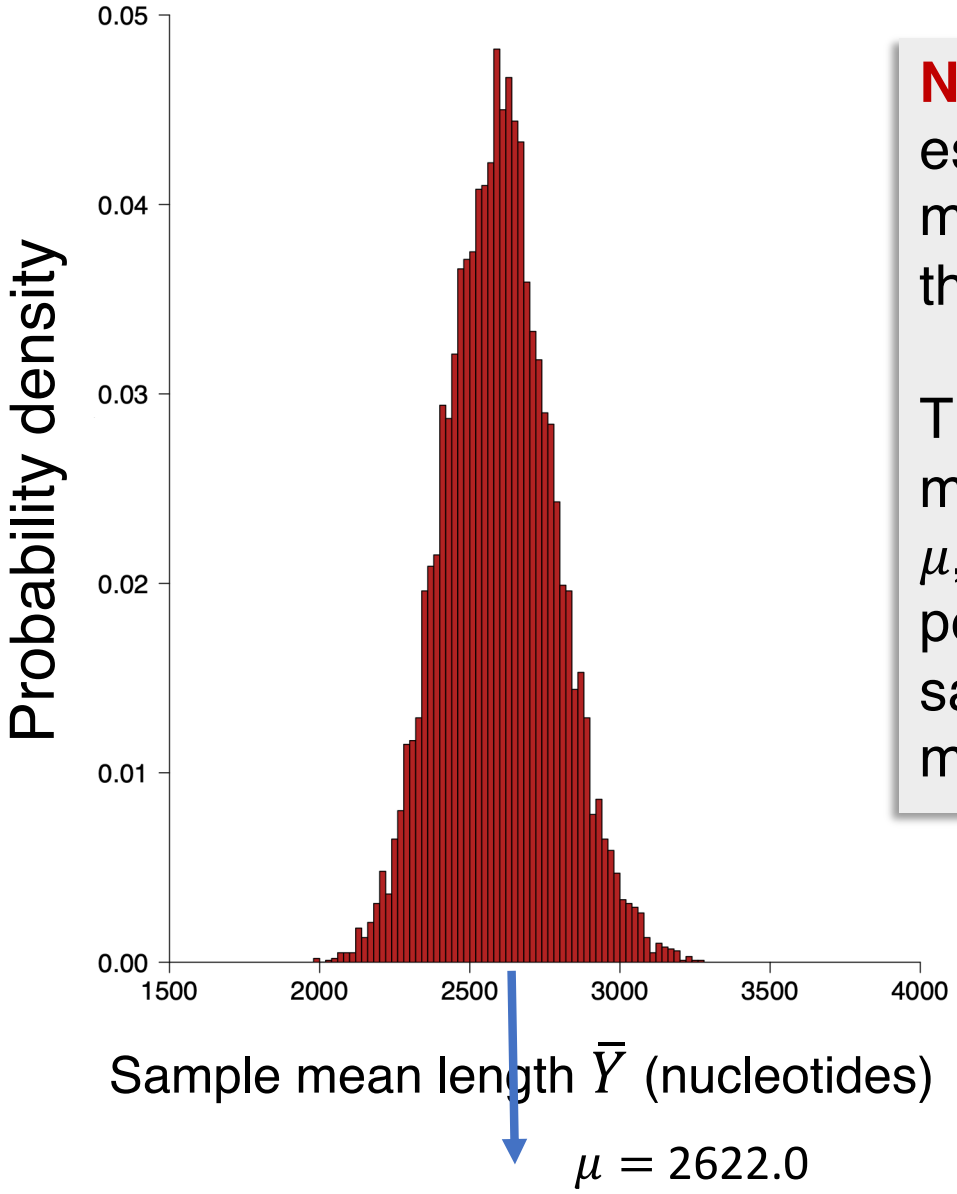
The sampling distribution of sample means ($\bar{Y}$)



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

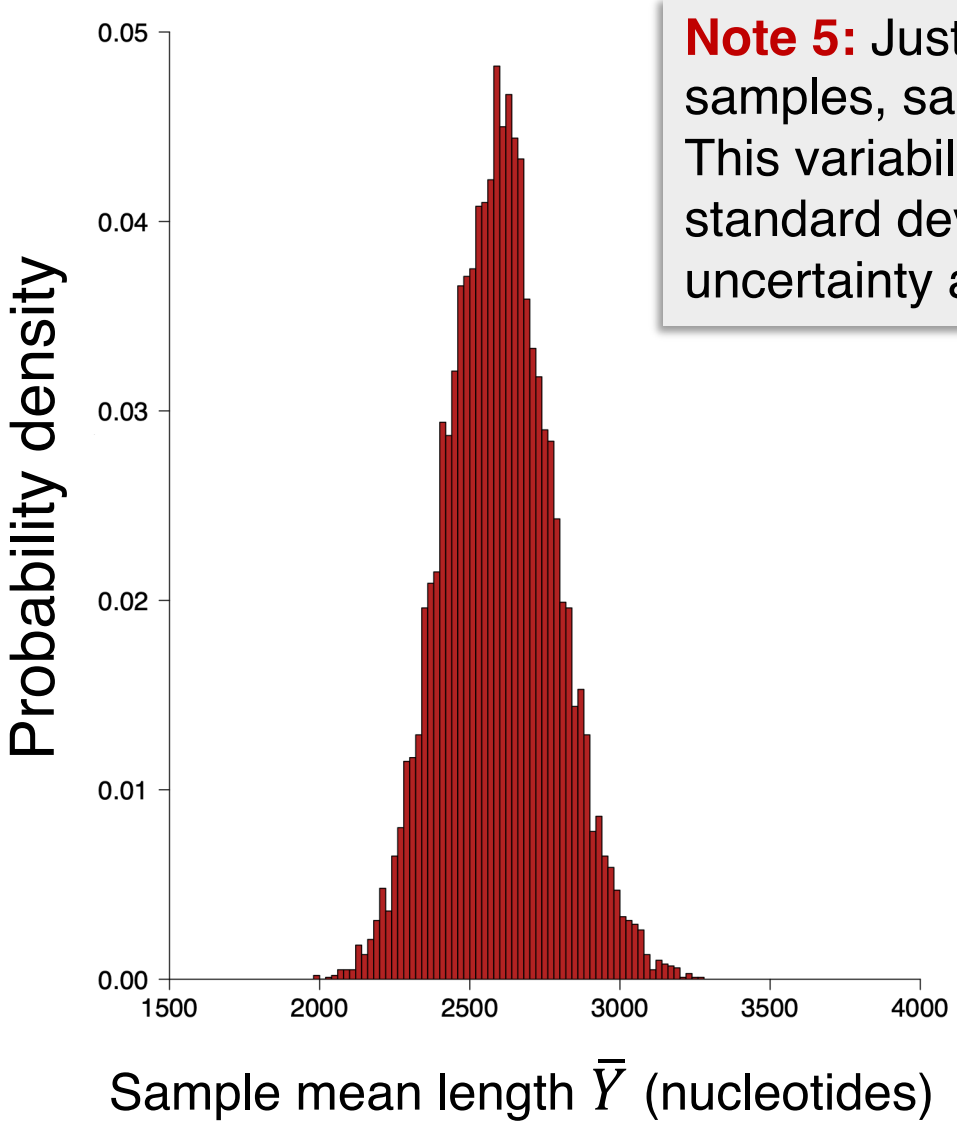
The sampling distribution of sample means ($\bar{Y}$)



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 of sample means ($\bar{Y}$)



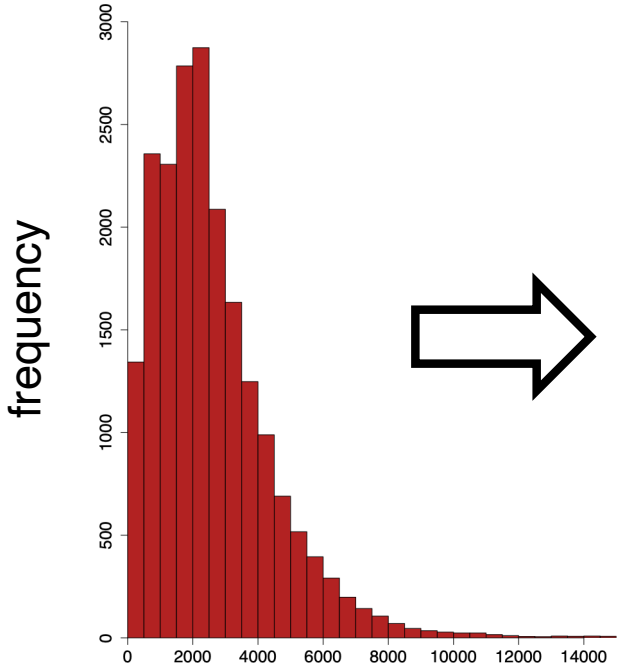
Note 5: Just as sample means vary across samples, sample standard deviations do as well. This variability matters because the sample standard deviation is what allows us to quantify the uncertainty associated with a sample mean.

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

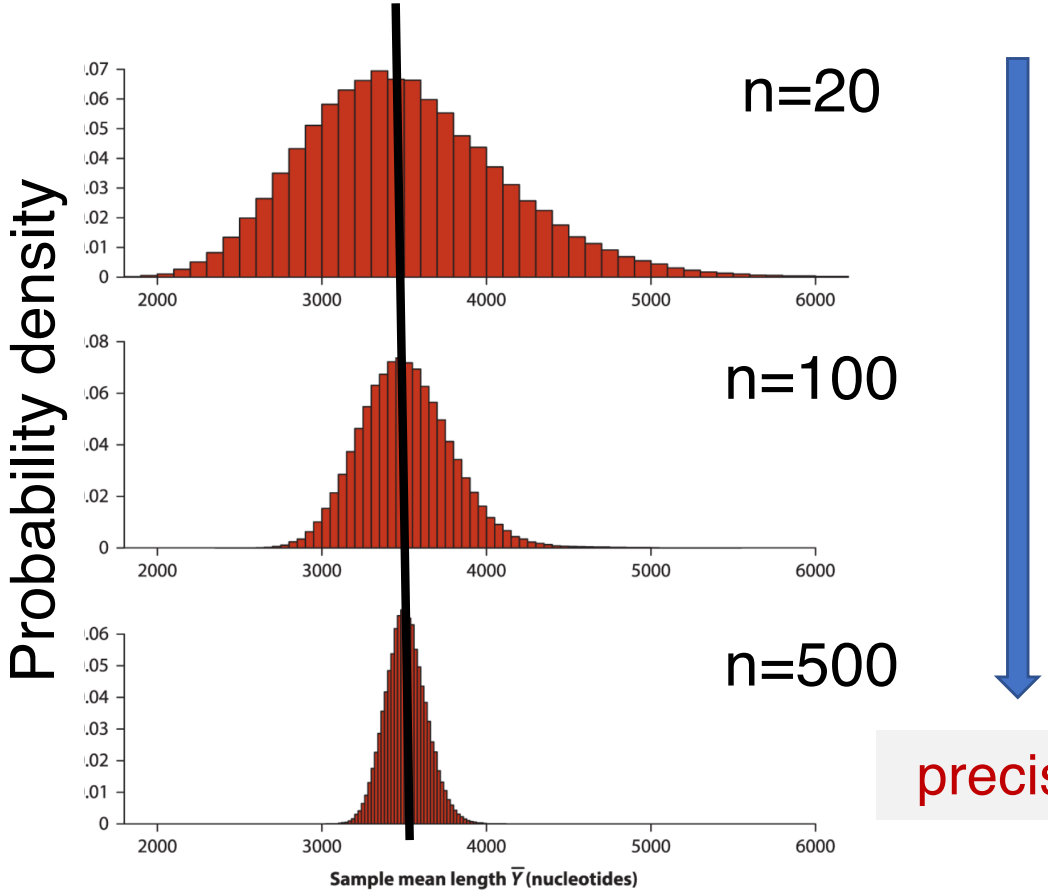
Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2122.3
Standard deviation	s	2423.1

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

Frequency distribution of the gene length Population



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)