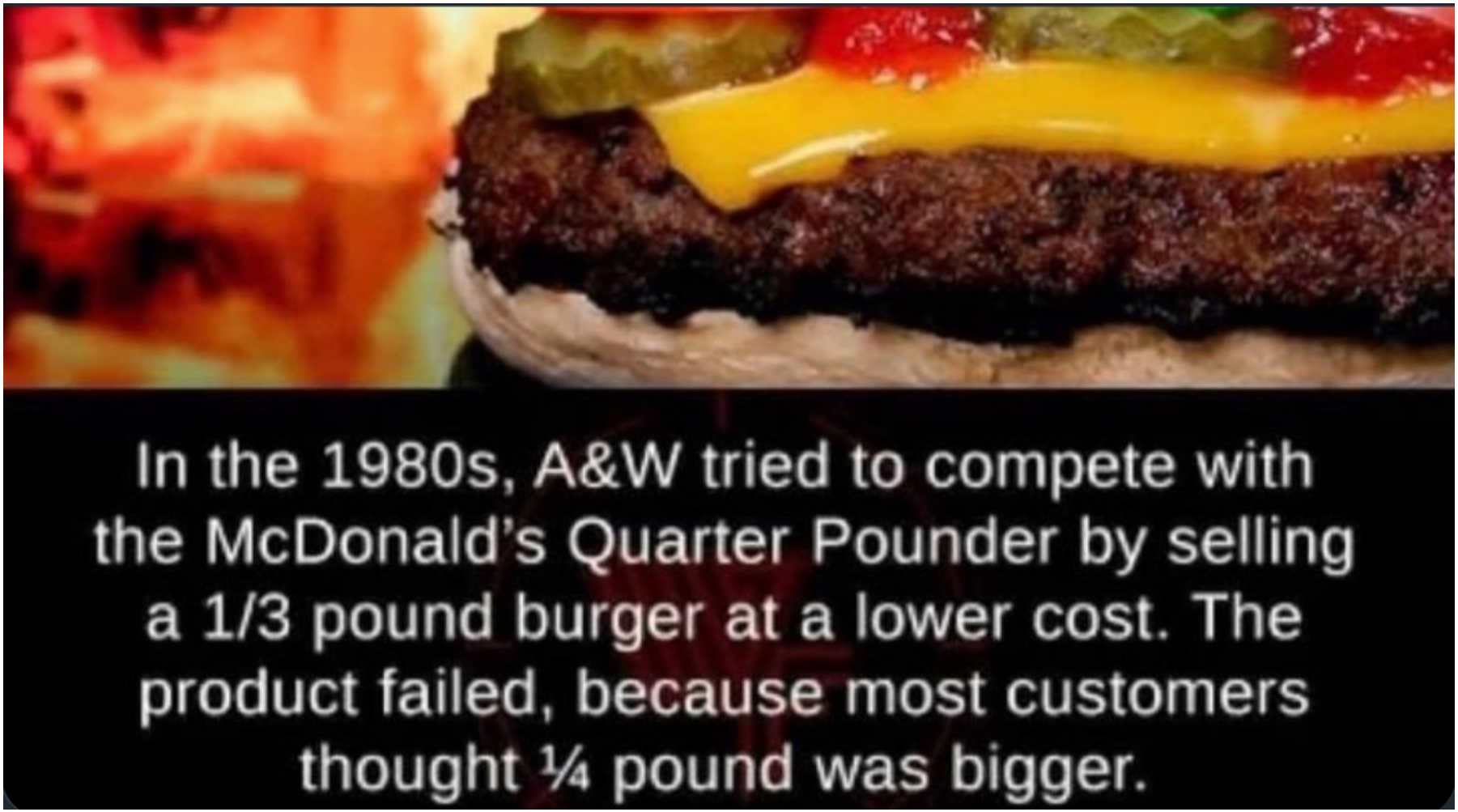
A snap demonstration of why numeracy is key to society

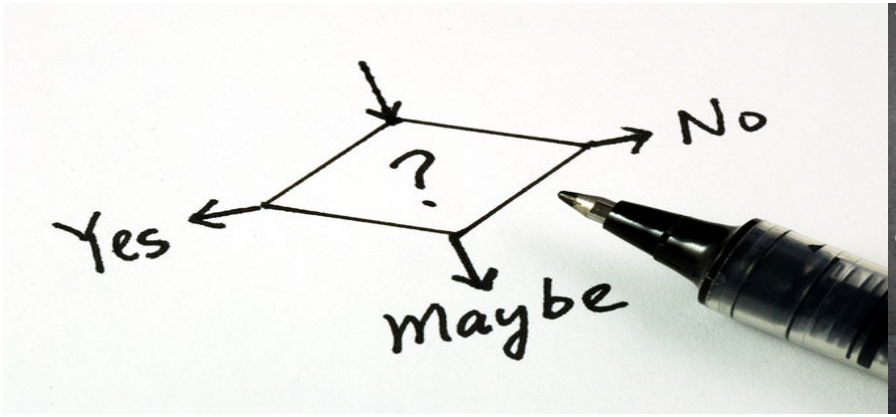


Lecture 8: estimating with uncertainty with certainty (i.e., with some confidence)

Statistics is the science of assisting in decision making with incomplete knowledge

"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 and pretty much everything else is based on samples!

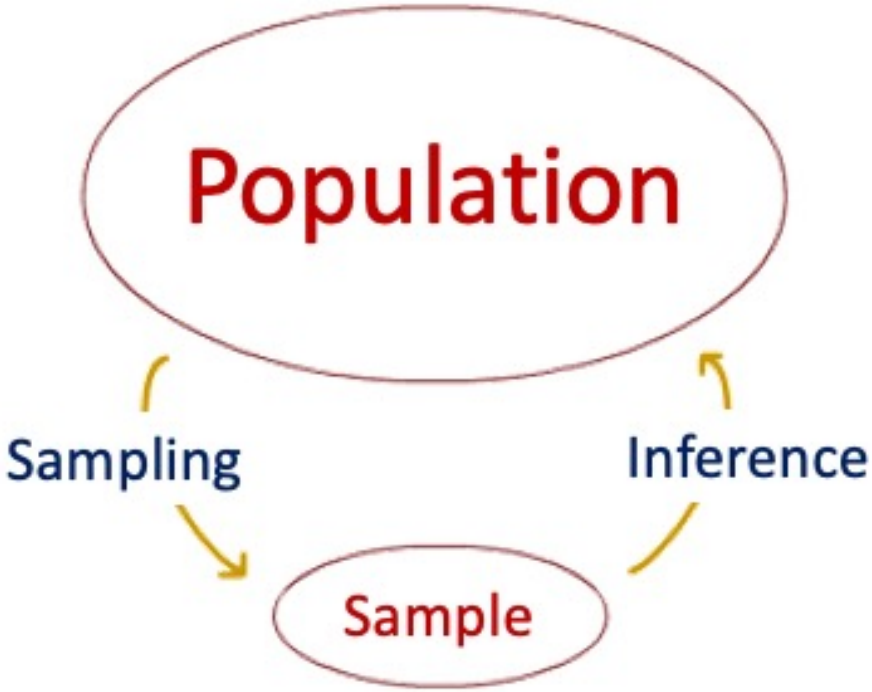
- The most important goal of statistics is to estimate (infer) an unknown quantity (parameter) of an entire population based on sample data (often one single sample from the population).

- Estimation is the process of inferring a population parameter (mean, standard deviation, median, etc) from sample data!

We use estimates to make decisions - Statistics is the science of making decisions with incomplete knowledge (i.e., based on samples) based on populations that too often have unknown sizes.

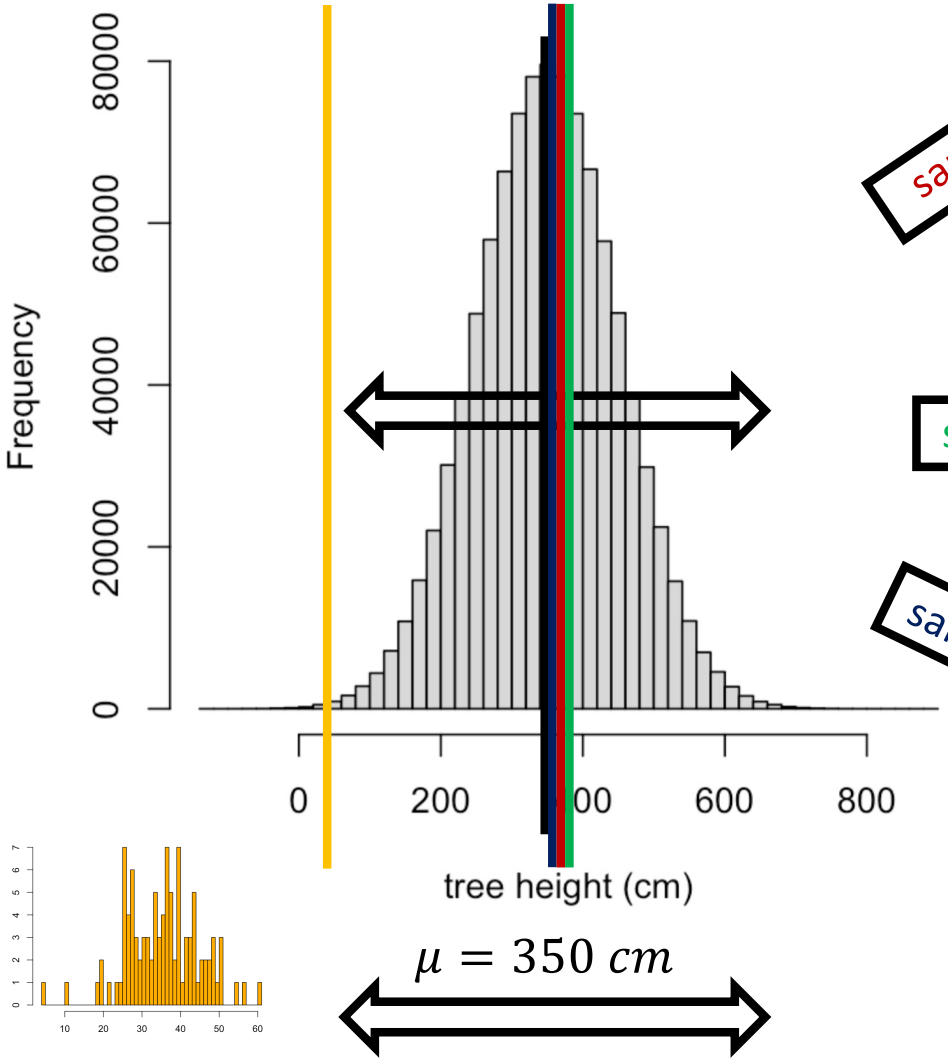
But sample-based statistics (mean, median, standard deviation, etc) vary from sample to sample (i.e., they have some level of uncertainty) - we call this variation as “sampling variation”.

How to estimate under uncertainty (sample variation) with certainty (i.e., with some confidence)?



Sampling variation generates uncertainty

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

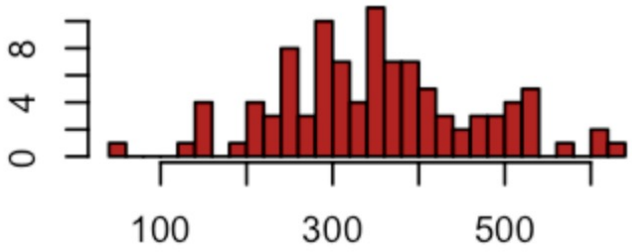


sampling

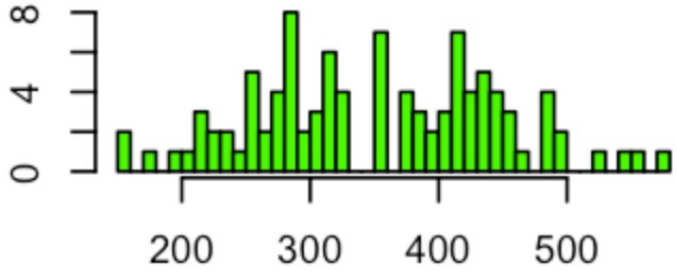
sampling

sampling

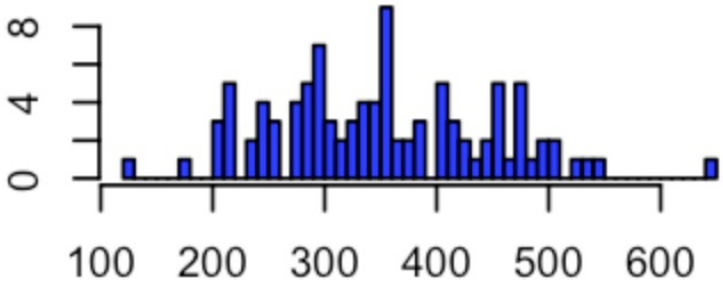
$\bar{X} = 351.5\ cm; s = 114.2\ cm$



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



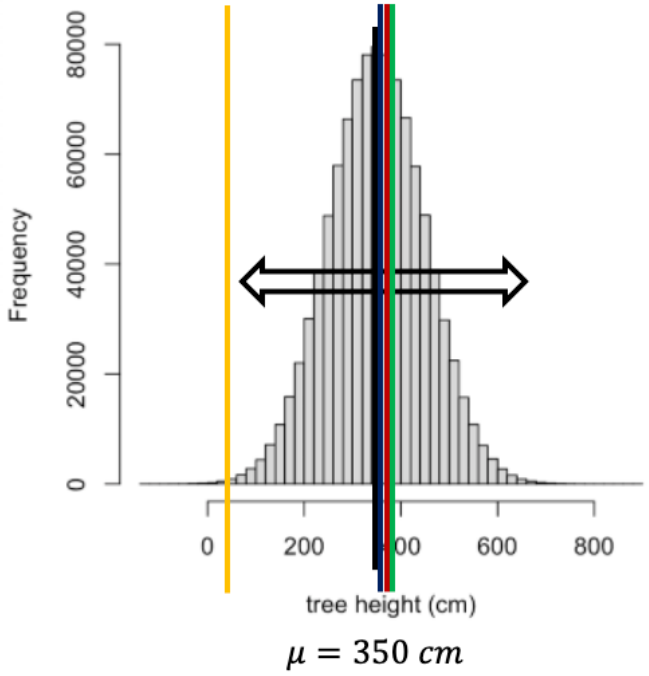
$\bar{X} = 351.4\ cm; s = 96.6\ cm$



Uncertainty (samples means varying around the true population mean)

The variation among observations within samples (standard deviation) can inform us about how far sample means in general might be from the true population mean (estimate how wrong one could be).

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

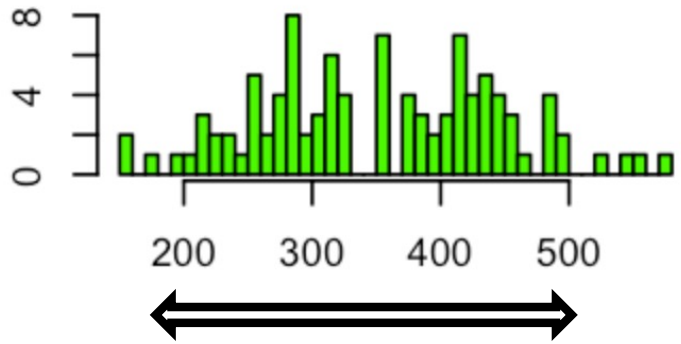


$\mu = 350\ cm$

Variation among samples

sampling

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



Variation within samples

Variation within samples (among observations) can generate estimates of certainty (confidence) about uncertainty (variation among sample means)

Parameters (populations) versus Estimates (samples)

A **parameter** is a quantity describing a statistical population, whereas an **estimate** or **statistic** is a related quantity from a sample.

The mean of of a statistical population is a parameter; and the mean of a sample is an estimate (or statistic) of the mean of the population.

The standard deviation of a statistical population is a parameter; and the standard deviation of a sample is an estimate (or statistic) of the standard deviation of the population.

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

A value from an estimate (i.e., from a sample) is never (especially in large populations) the same as the value of the population parameter being estimated, because sampling is influenced by chance.

Two people could sample 100 trees from the same forest and get different mean values for the two samples; and the two samples would almost definitely not be equal to the population mean.

The critical question is statistics is: In the face of uncertainty (due to random chance), how much can we trust an estimate and the decisions we make based on that estimate, i.e., what is its accuracy? (i.e., how close the sample value is to the true population value?).

Deal with uncertainty with some certainty!!!

How to estimate with uncertainty with certainty (with some confidence)?

We need to understand properties of estimators (mean, standard deviation, etc).

Properties of estimators are understood via the sampling distribution of the estimate or statistic of interest (e.g., sample mean, standard deviation, etc).

Sampling distributions are the probability distributions of an estimate (i.e., sample-based) what we might have obtained when we sample the population (randomly). They look like frequency distributions but transformed into probabilities.

Statistical symbols

μ = population mean (we say “mu”, Greek alphabet). σ = population standard deviation (we say “sigma”).

Important statistical symbols regarding inference

μ = population mean (we say "mu", Greek alphabet). σ = population standard deviation (we say "sigma").

$\bar{X}$ = sample mean (mean of the sample) - we say "X bar", Latin or Roman alphabet).
s = sample standard deviation.

Note - Although μ is always the mean of the population for whatever variable you are measuring (e.g., X), symbols for the sample mean can take other values (e.g., $\bar{X}$, $\bar{Y}$) depending how you call the variable of interest (X or Y or something else), but it always has the bar at the top.

Estimating with uncertainty: the sampling distribution of an estimate the case of a tiny statistical population of 5 numbers

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

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

(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

Properties of estimators are understood via the sampling distribution of the estimate (e.g., sample mean).

Property 1:

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

If the mean of all possible sample means, i.e., the mean of the sampling distribution of the estimate (e.g., sample mean, standard deviation based on samples), **the sample estimate is said to be unbiased when sampling is performed randomly (i.e., all observations in the population have equal chance to be sample).**

In this case, the mean is unbiased because sample means (under random sampling) don't have the tendency to be more often bigger or more often smaller than the true population mean.

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

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 minimizes sampling error & inferential bias (i.e., how close or far the sample values from the statistic of interest are from the true population value for that statistic)

A random sample is one that fulfills two criteria:

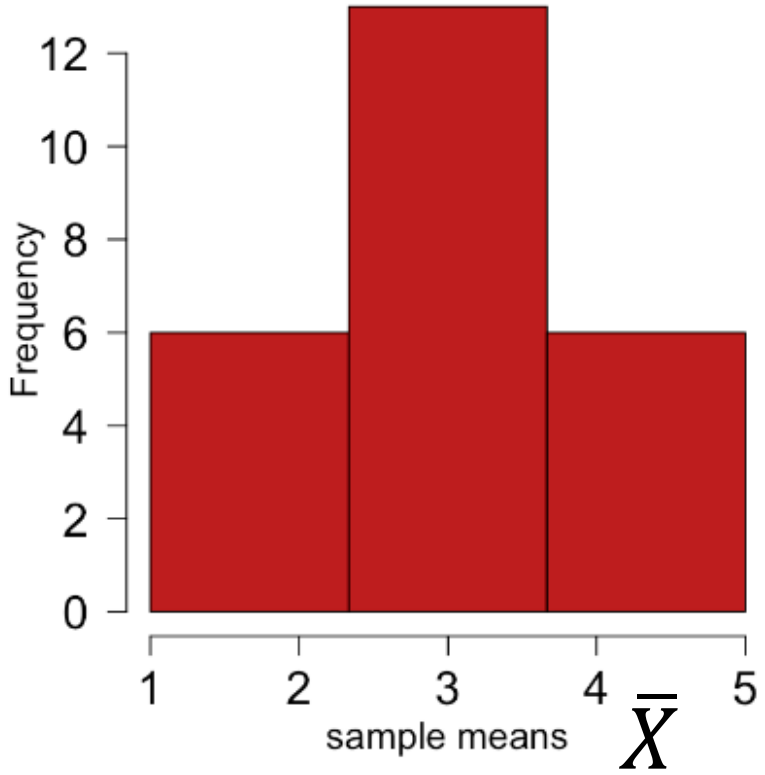
1) Every observational unit in the population (e.g., individual tree) have an **equal chance** of being included in the sample.

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.

Samples are biased when some observational units of the intended population have lower or higher probabilities to be sampled.

Estimating with uncertainty: the sampling distribution of an estimate the case of a tiny statistical population of 5 numbers

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



$\mu = 3$

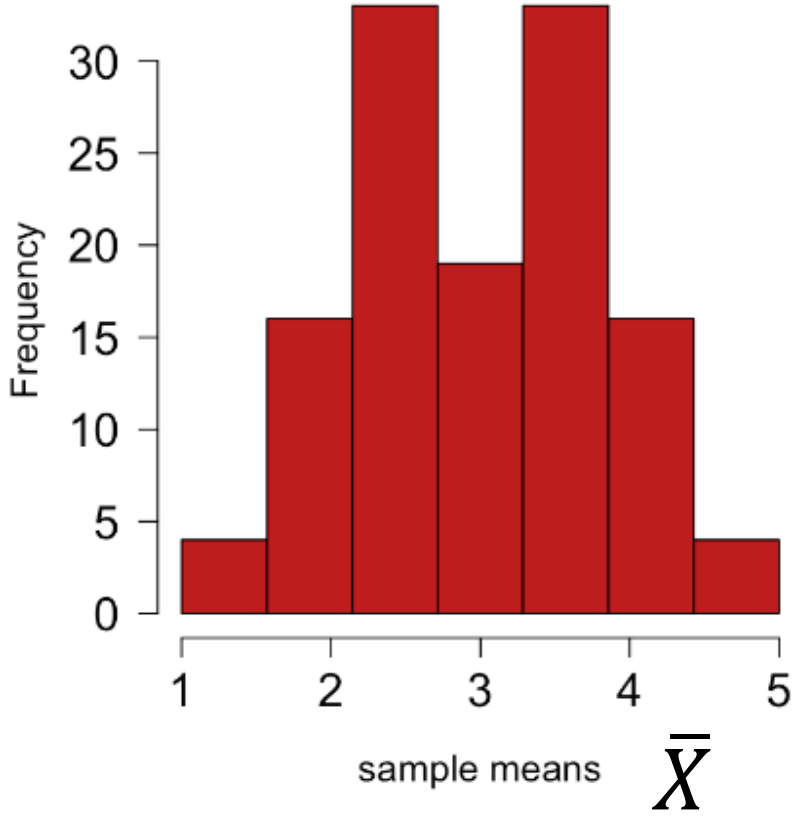
Mean of all samples means = 3.0

n = 2

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

Estimating with uncertainty: the sampling distribution of an estimate the case of a tiny statistical population of 5 numbers

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



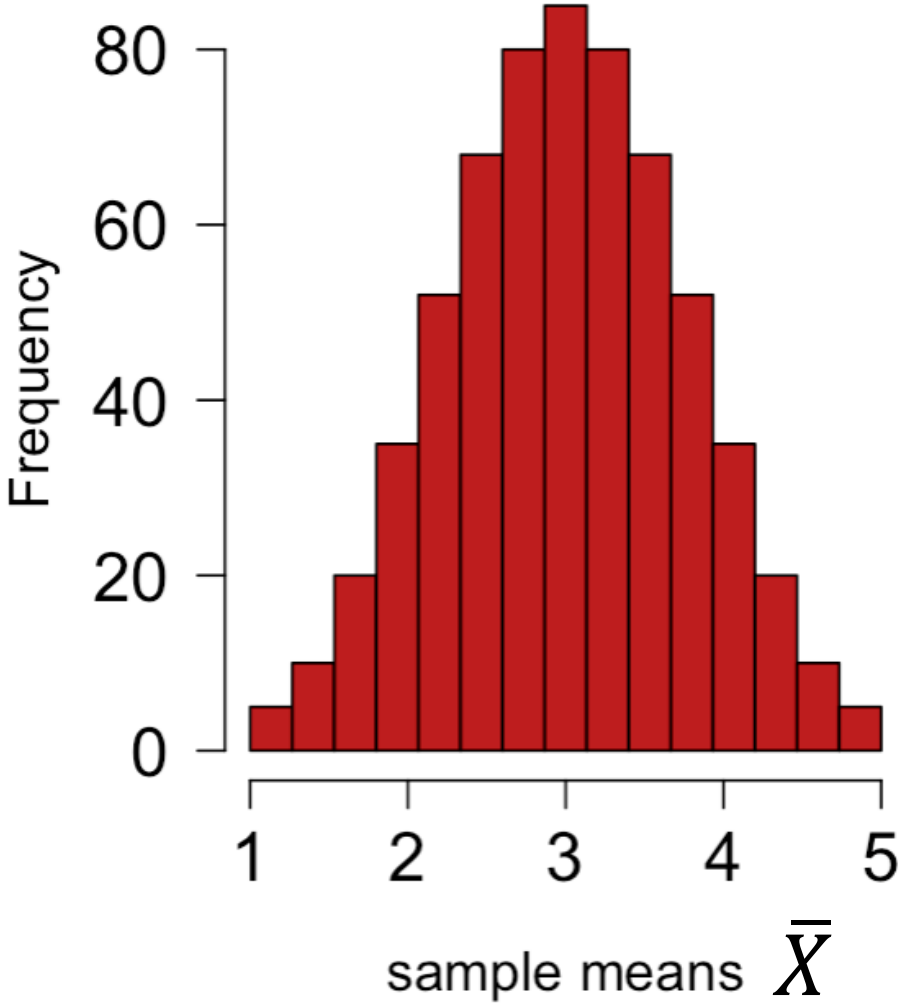
$\mu = 3$

Mean of all samples means = 3.0

n = 3

Estimating with uncertainty: the sampling distribution of an estimate the case of a tiny statistical population of 5 numbers

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



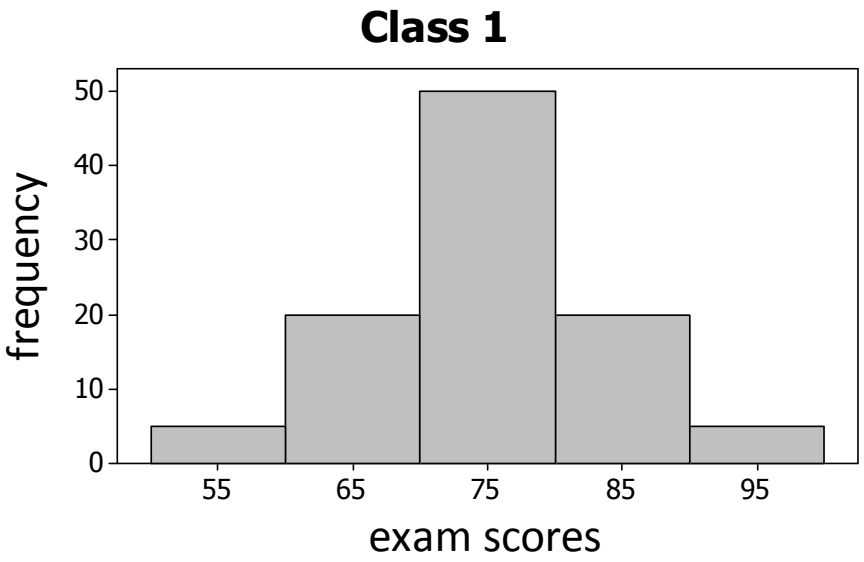
$\mu = 3$

Mean of all samples means = 3.0

n = 4

Remember Variability in frequency distributions?!!

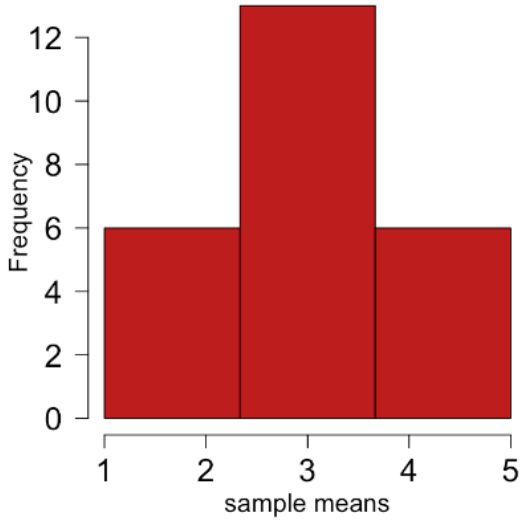
In which class exam scores vary the most?



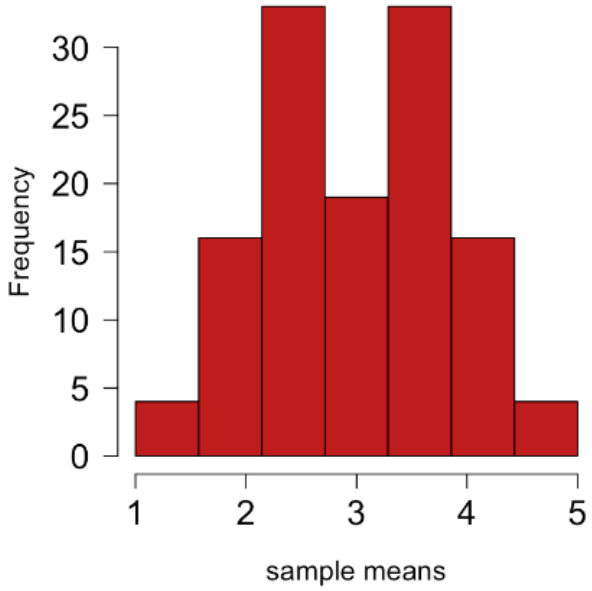
- Source: Cooper & Shore; Journal of Statistics Education (vol. 18, #2)

Which sample size leads to more precise (less variation around the true population value) sample estimates based on random sampling?

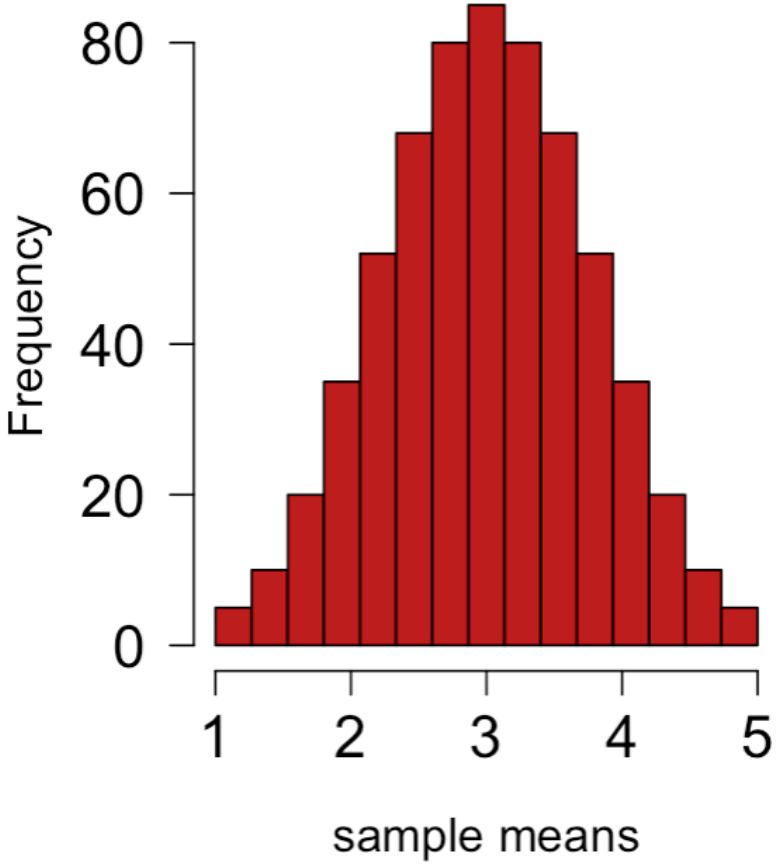
n = 2



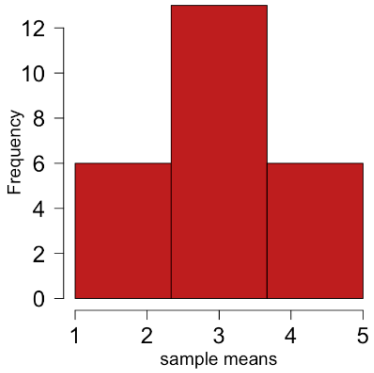
n = 3

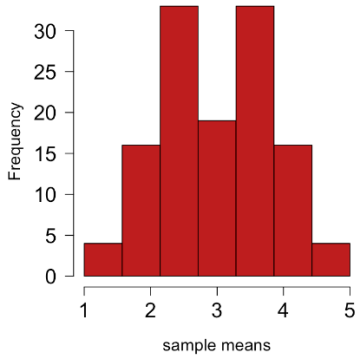


n = 4

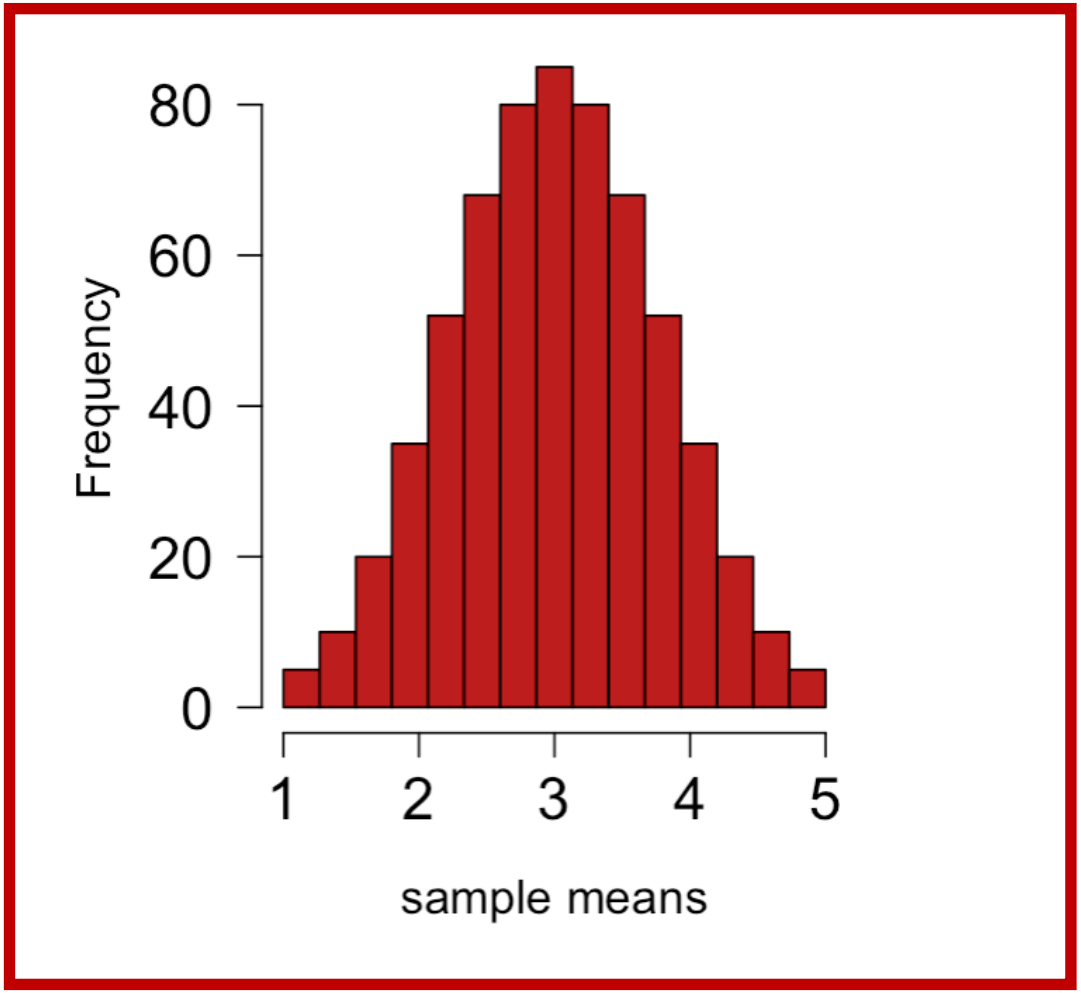


Which sample size leads to more precise (closer to the true population value) sample estimates based on random sampling?



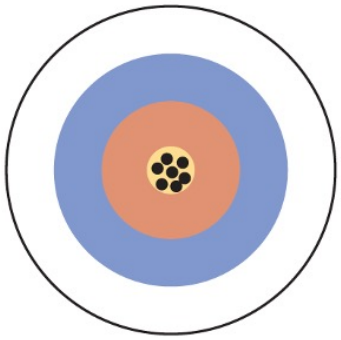
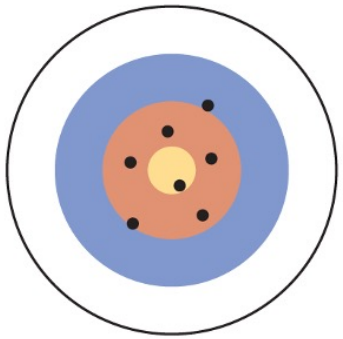
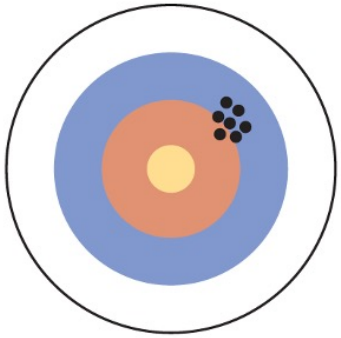
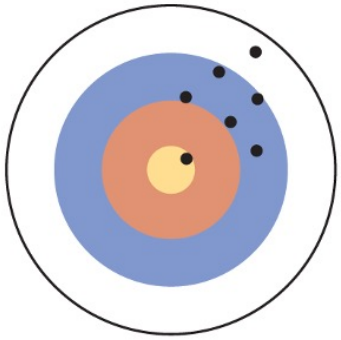


As sample size increases, there is a greater probability (under random samples) that a given sample will be closer to the true population mean; i.e., they become more precise.



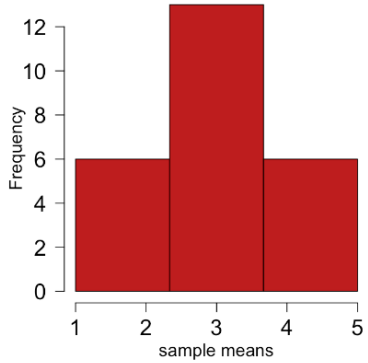
Because sampling was random, then the sample mean is accurate (i.e., unbiased); the mean of all sample means equal the population mean (parameter)

Random sampling minimizes bias and makes it possible to measure the amount of sampling error (next lectures)

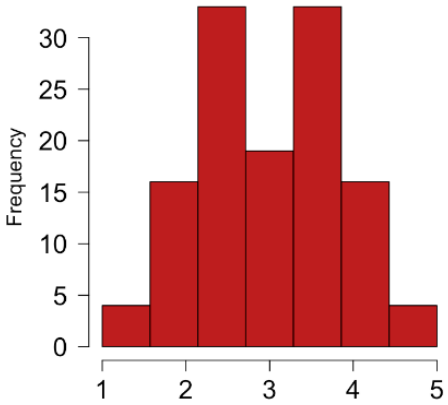
	Precise	Imprecise
Accurate	 Low sampling variation (sampling error) & low bias	 High sampling variation (sampling error) & low bias
Inaccurate	 Low sampling variation (sampling error) & high bias	 High sampling variation (sampling error) & high bias

Which sample size leads to greater accuracy and precision?

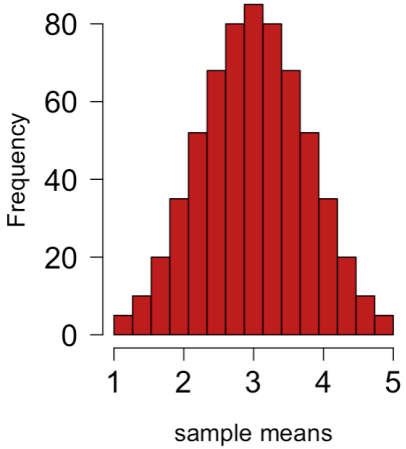
n = 2

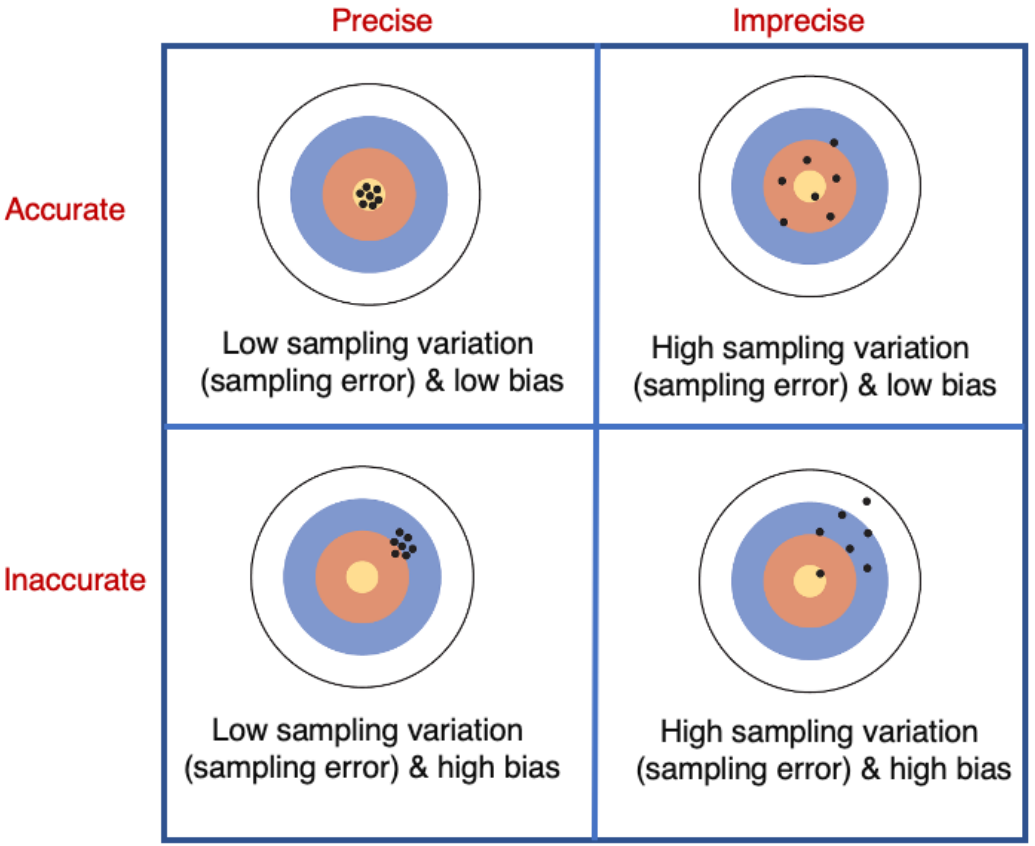


n = 3



n = 4





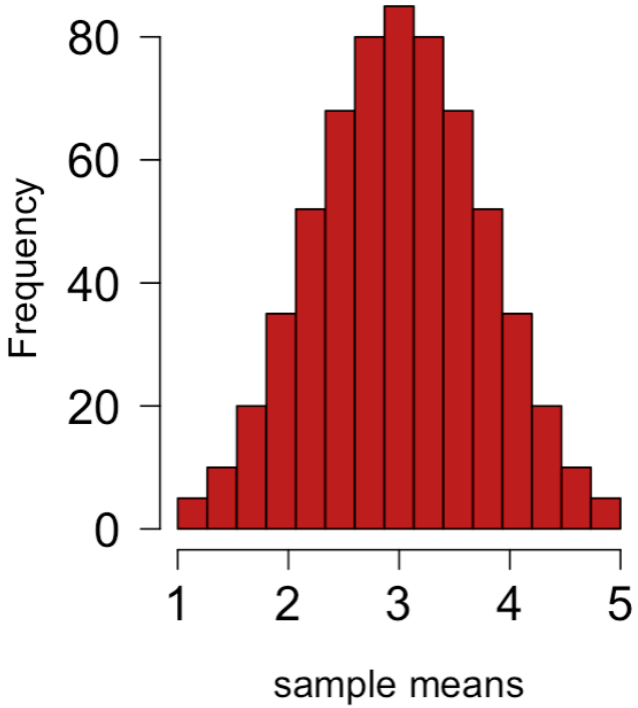
Sampling distributions are best represented by probability distributions

Probability density is the relationship between observations (here sample means) and their probability.

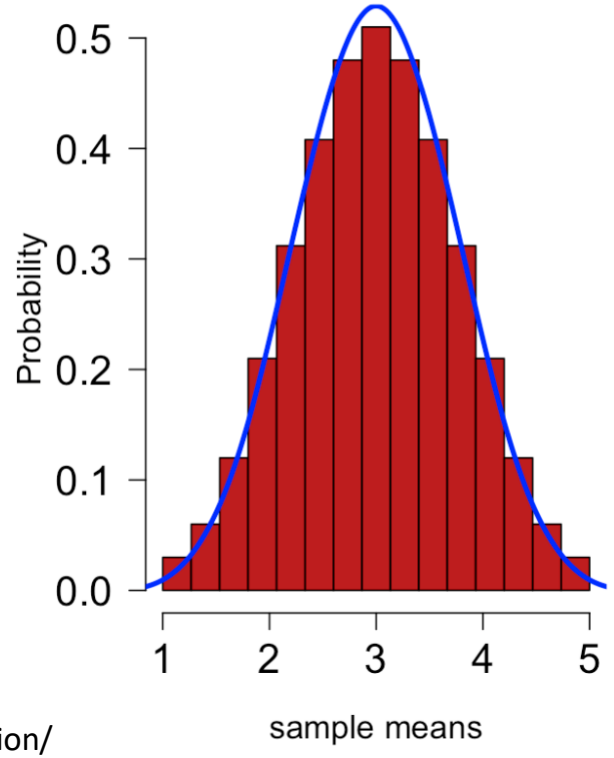
Some outcomes (samples) of a random variable (sample means) will have low probability density and other outcomes will have a high probability density.

The overall shape of the probability density is referred to as a probability distribution, and the calculation of probabilities for specific outcomes of a random variable is performed by a probability density function, or PDF for short.

frequency distribution of samples

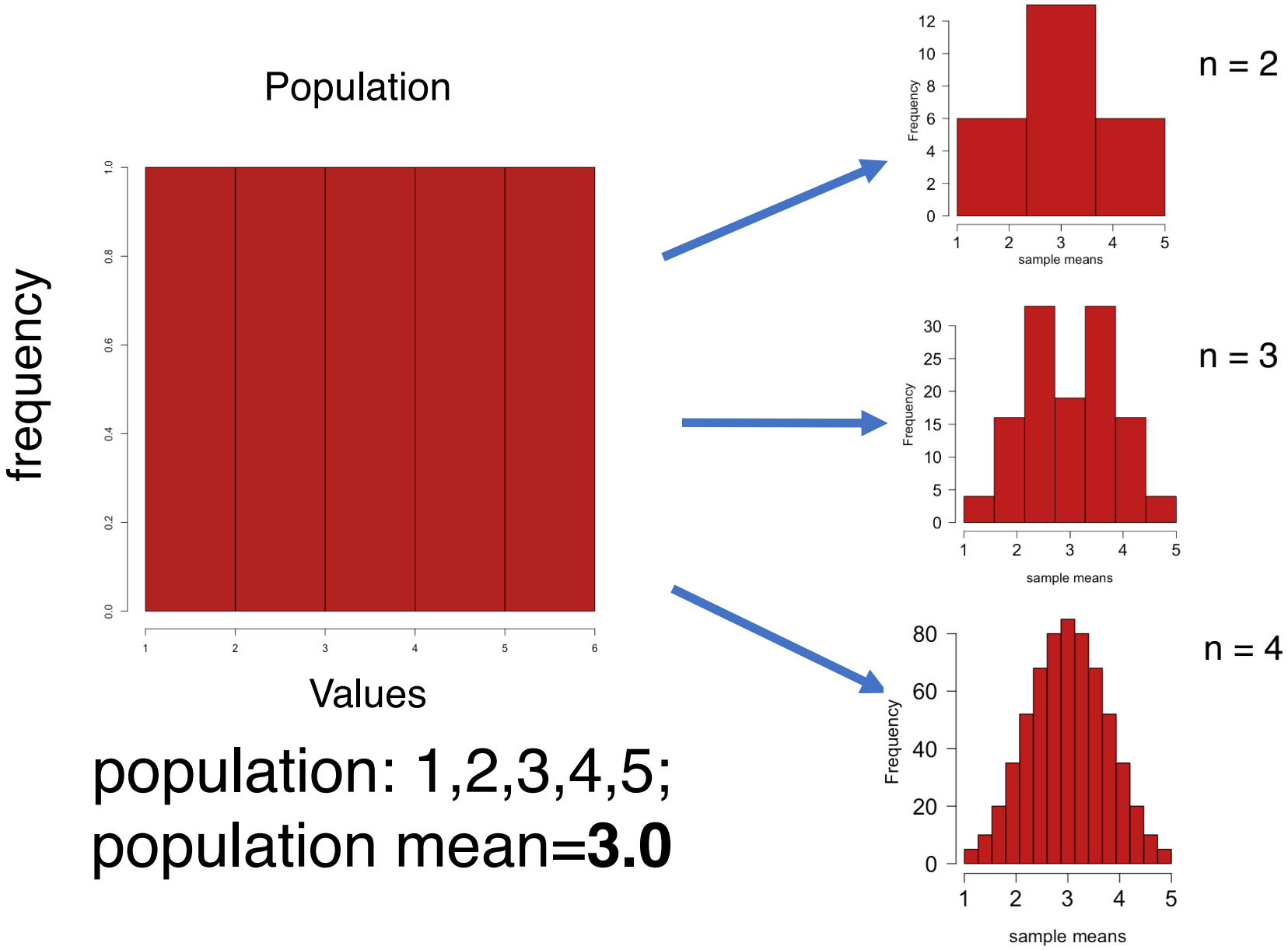


probability distribution of samples



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

Critical: The shape of the frequency distribution of the population is not necessarily similar to the frequency distribution of the sample estimates (e.g., here the distribution of sample mean values) from the population

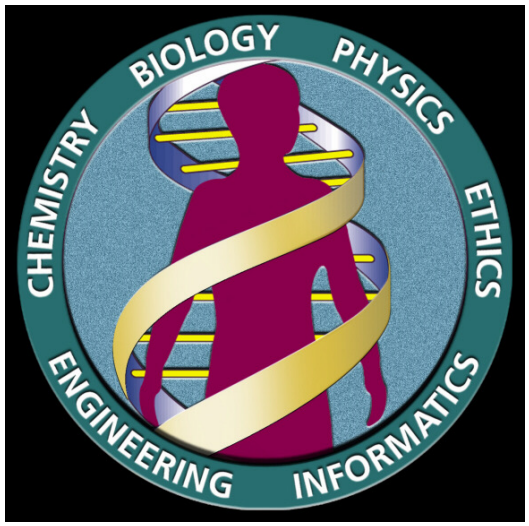


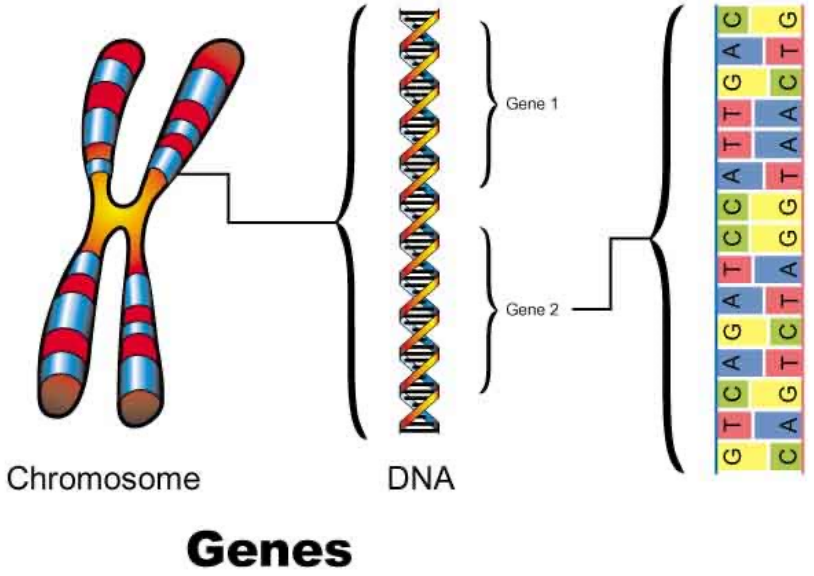
population: 1,2,3,4,5;
population mean=**3.0**

Let’s take a break – 2 minutes

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

The International Human Genome Project generated the DNA sequence of all 23 human chromosomes, each containing millions of nucleotides (more than 23,000 protein-coding genes)! Started in 1990 and finished in 2006 (sequence of the last chromosome). The available data that we have for BIOL 322 (tutorials) is 20,290 genes.

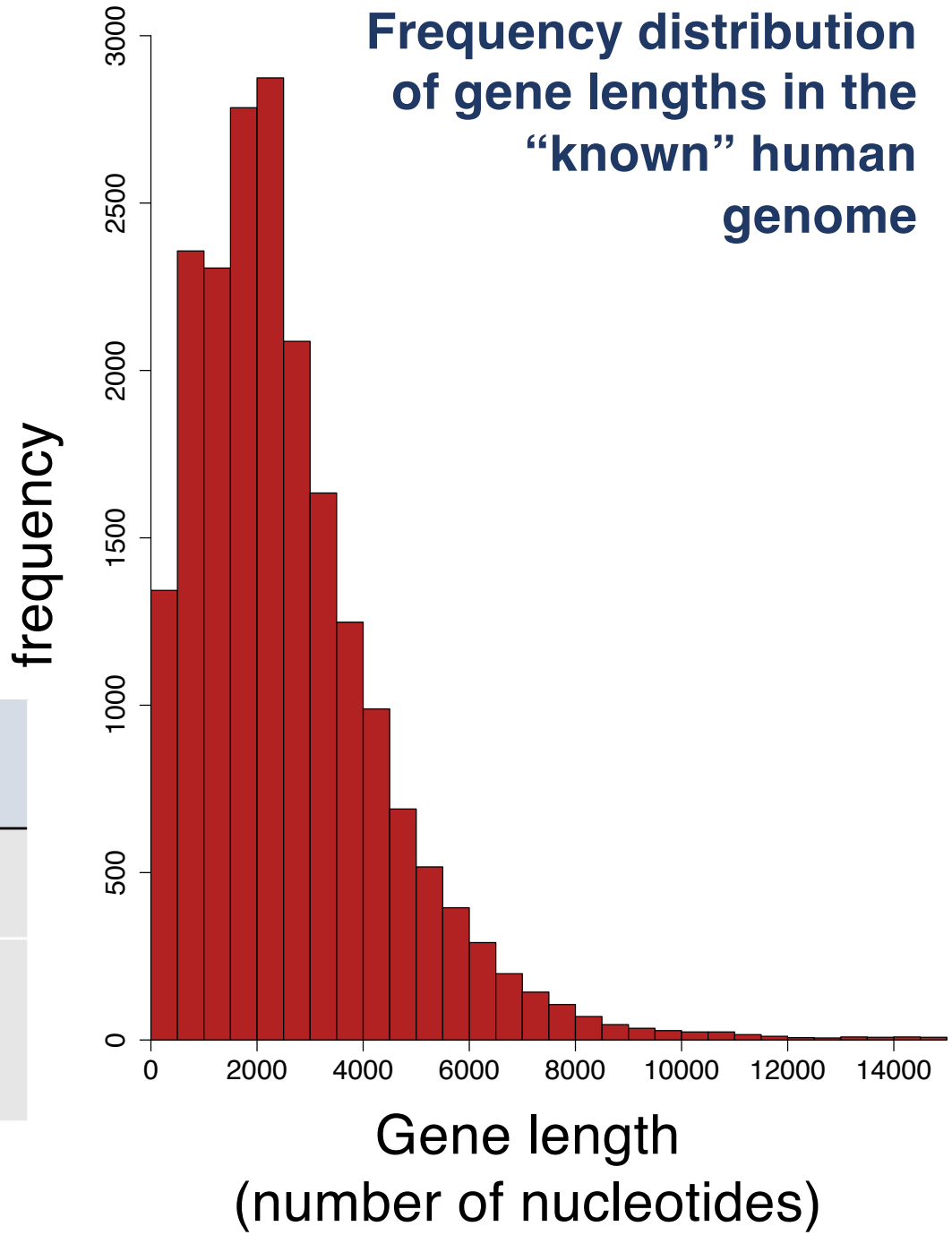




The length of human genes

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

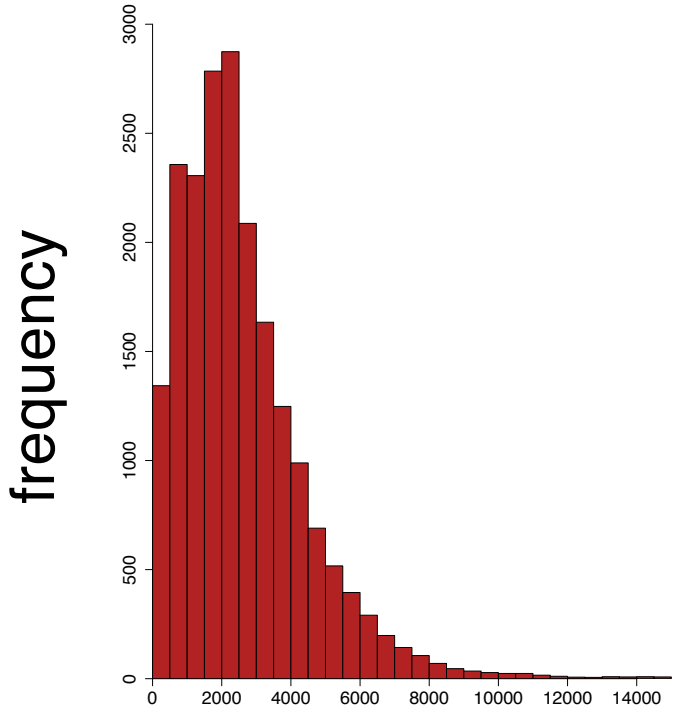
Names	Parameter	Value (nucleotides)
Mean (mu)	μ	2622.0
Standard deviation (sigma)	σ	2036.9



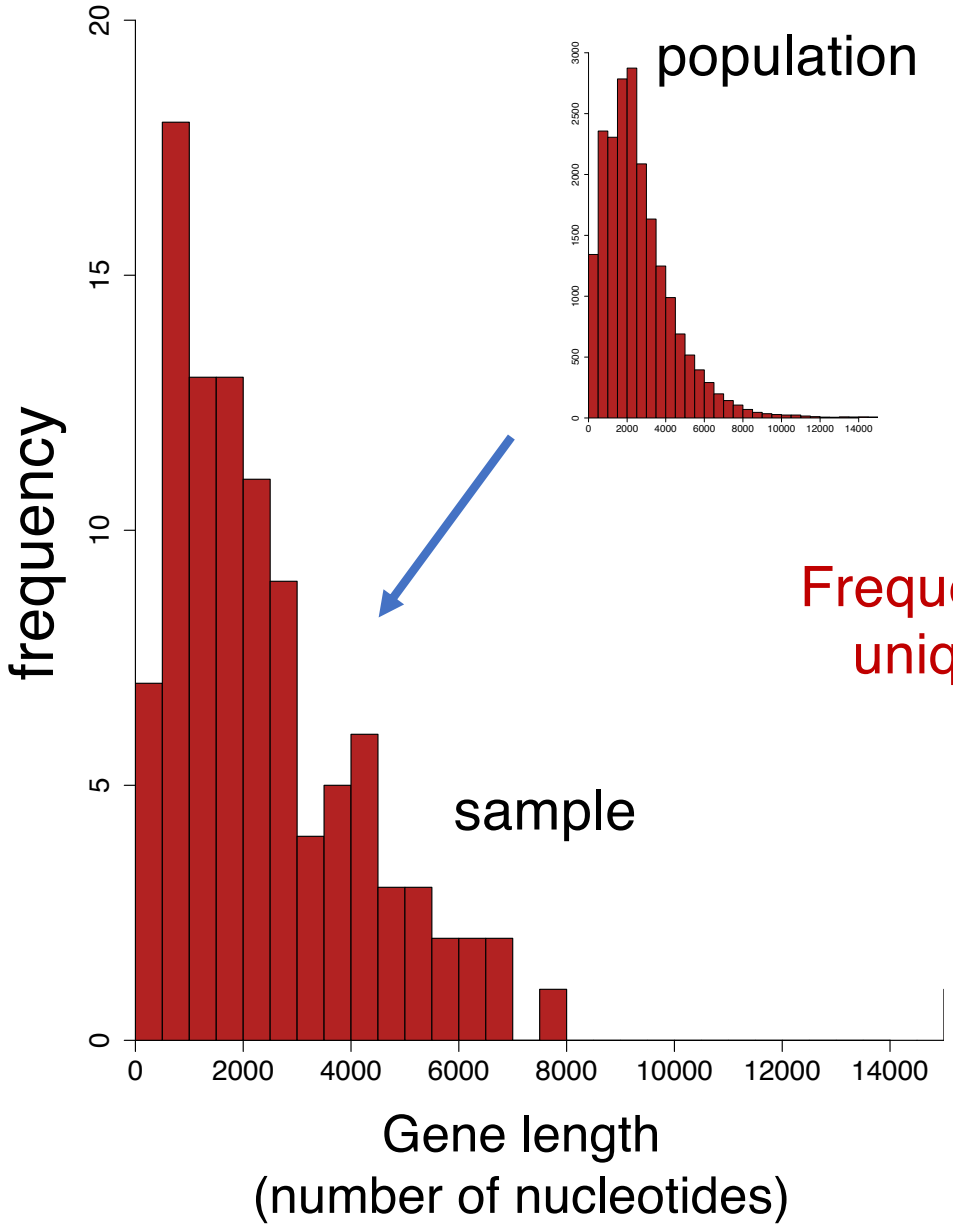
In real life we would not usually know the parameter values of the study population, but in this case we (almost) do!

So, we'll take advantage of this gene population to illustrate the process of sampling, uncertainty, accuracy, precision and how estimate with uncertainty with certainty (with some 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)

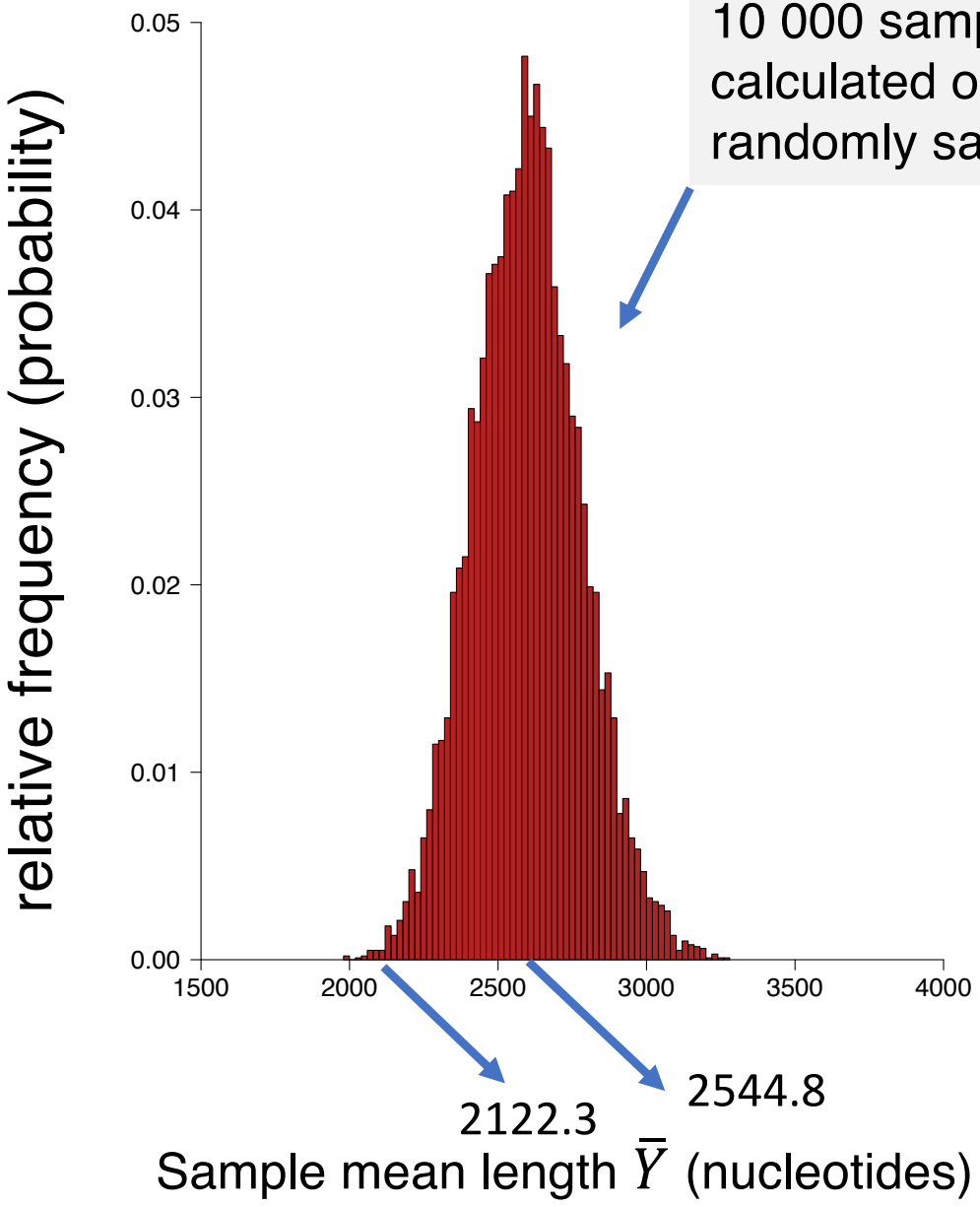


Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2544.8
Standard deviation	s	2125.3

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.

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



Sampling distribution of means based on 10 000 sample mean values. Each sample mean is the calculated on the basis of 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)

Population

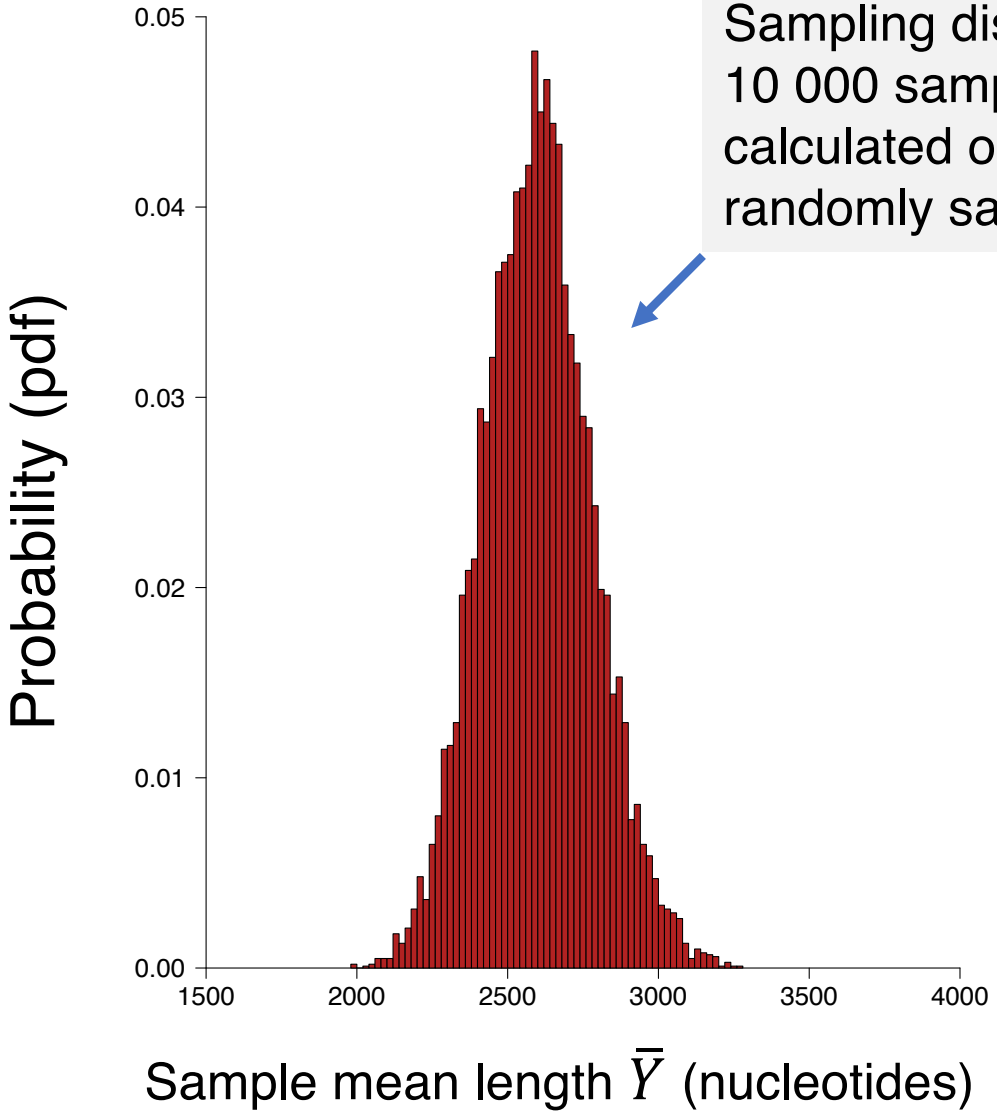
Names	Parameter	Value (nucleotides)
Mean	μ	2622.0
Standard deviation	σ	2036.9

Sample

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

The sample mean is about 77 nucleotides shorter than the true population value. We shouldn't be surprised that the sample estimates differ from the parameter (population) values. Such differences are virtually inevitable because of chance in the random sampling process (i.e., sampling variation).

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

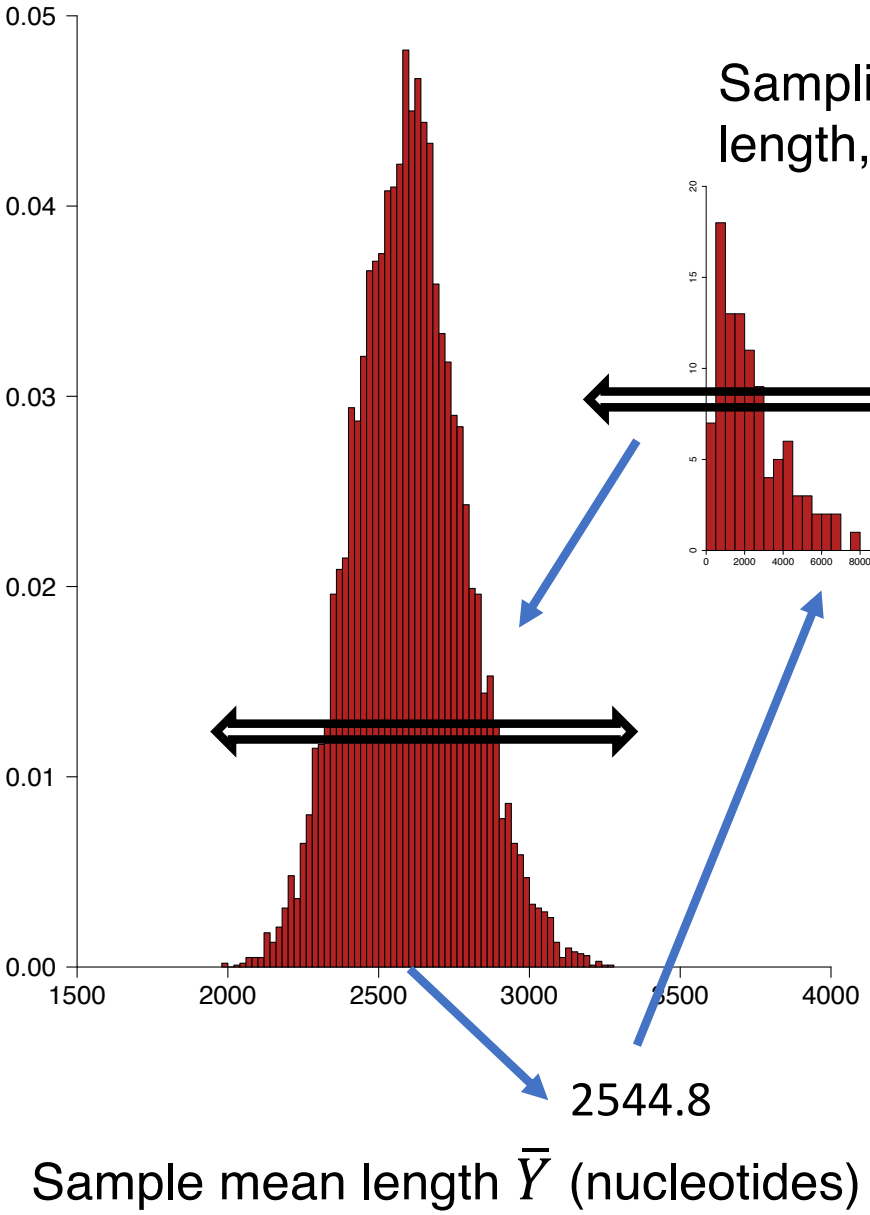


Sampling distribution of means based on 10 000 sample mean values. Each sample mean is calculated on the basis of the lengths of 100 genes randomly sampled from the population of 20,290 genes

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

But we can use an analytical approach (calculus based) to estimate the sampling distribution (probability distribution) of all infinite sample means based on 100 genes or any other sample size).

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



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

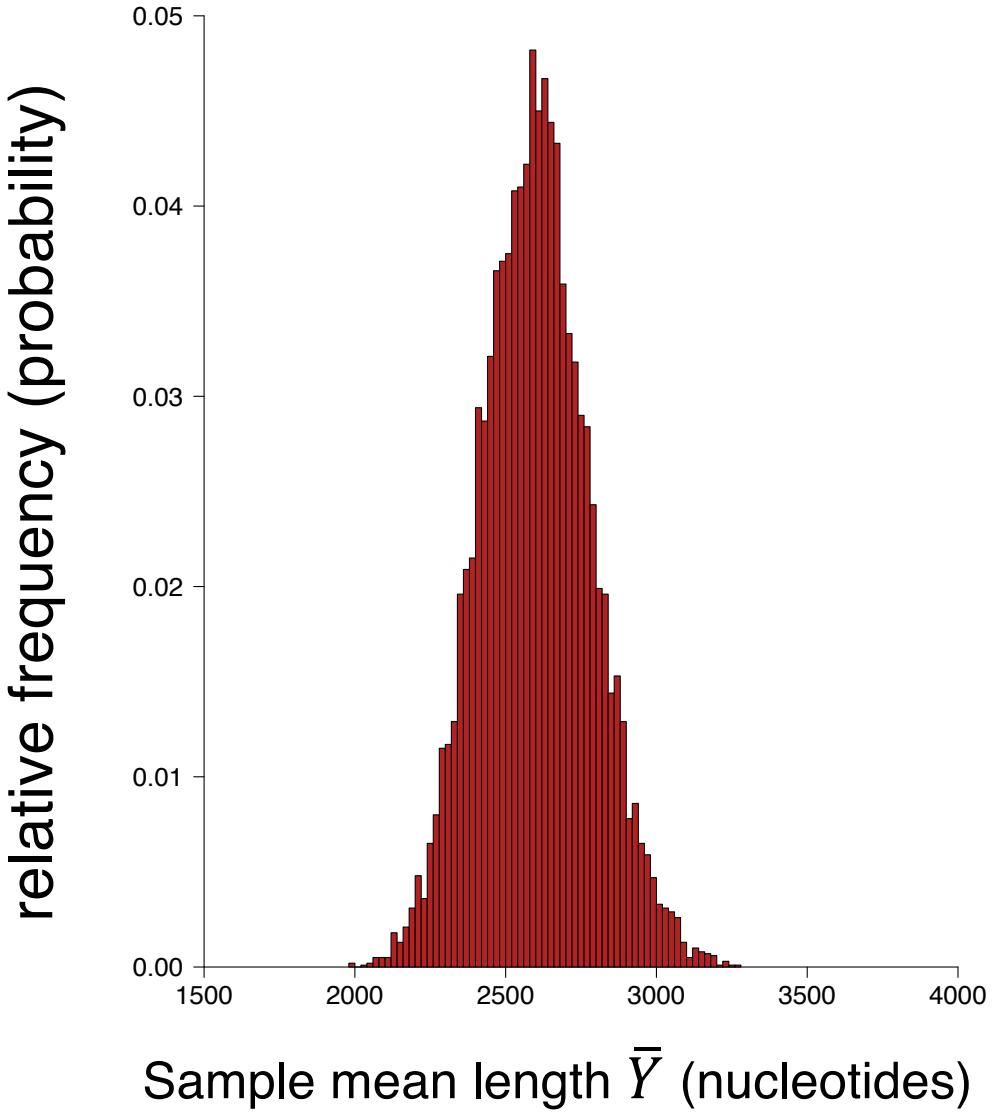
Statistical Wow: We will learn that variation within a single sample can estimate uncertainty among all possible sample values from a population.

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 – 2 minutes

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



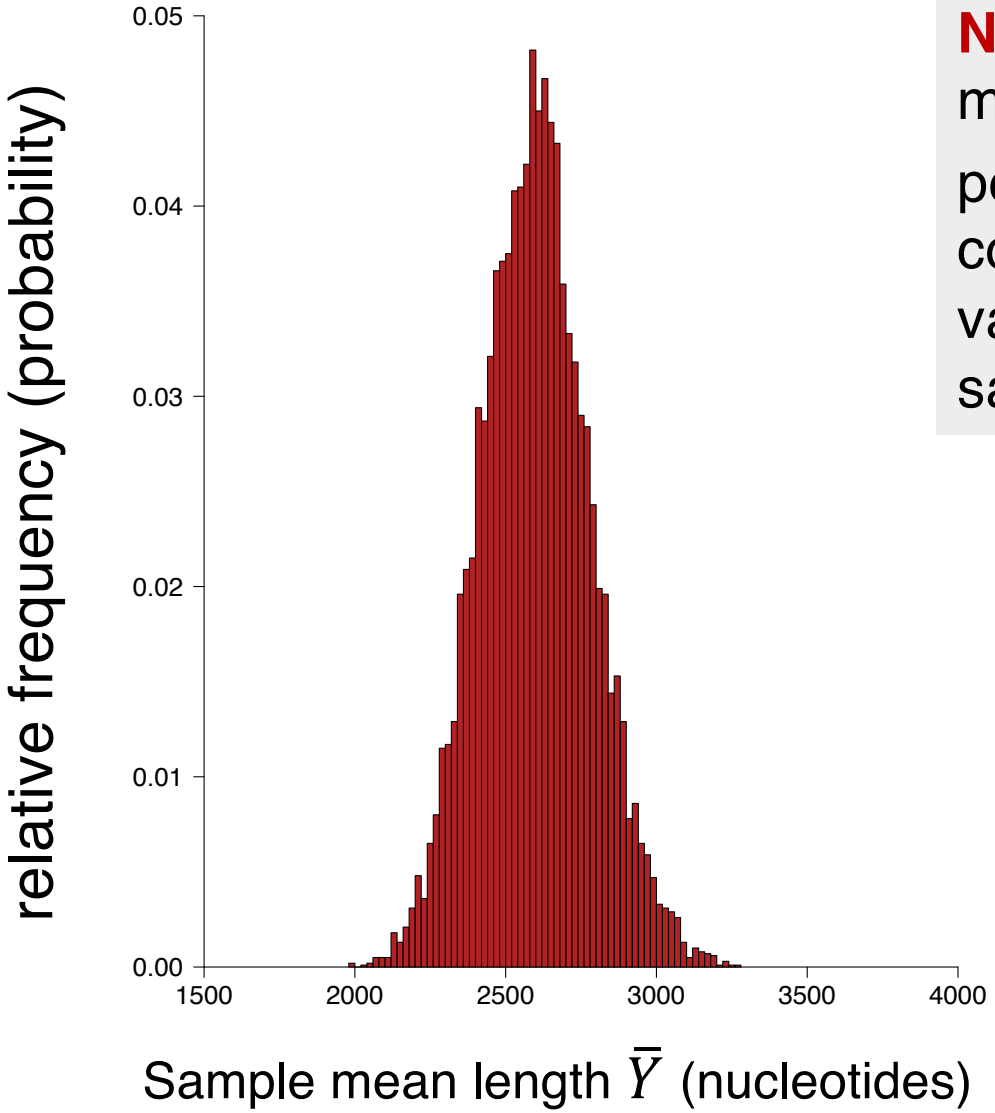
Note 1: We usually work with one single sample, and therefore only one sample mean value $\bar{Y}$.

But, understanding how sampling distributions are built is necessary to understand the process of estimating uncertainty (i.e., sample mean values vary from one sample to another) to determine confidence on inferences based on samples.

If samples vary too much among them, we would then have less confidence than if they vary little among them.

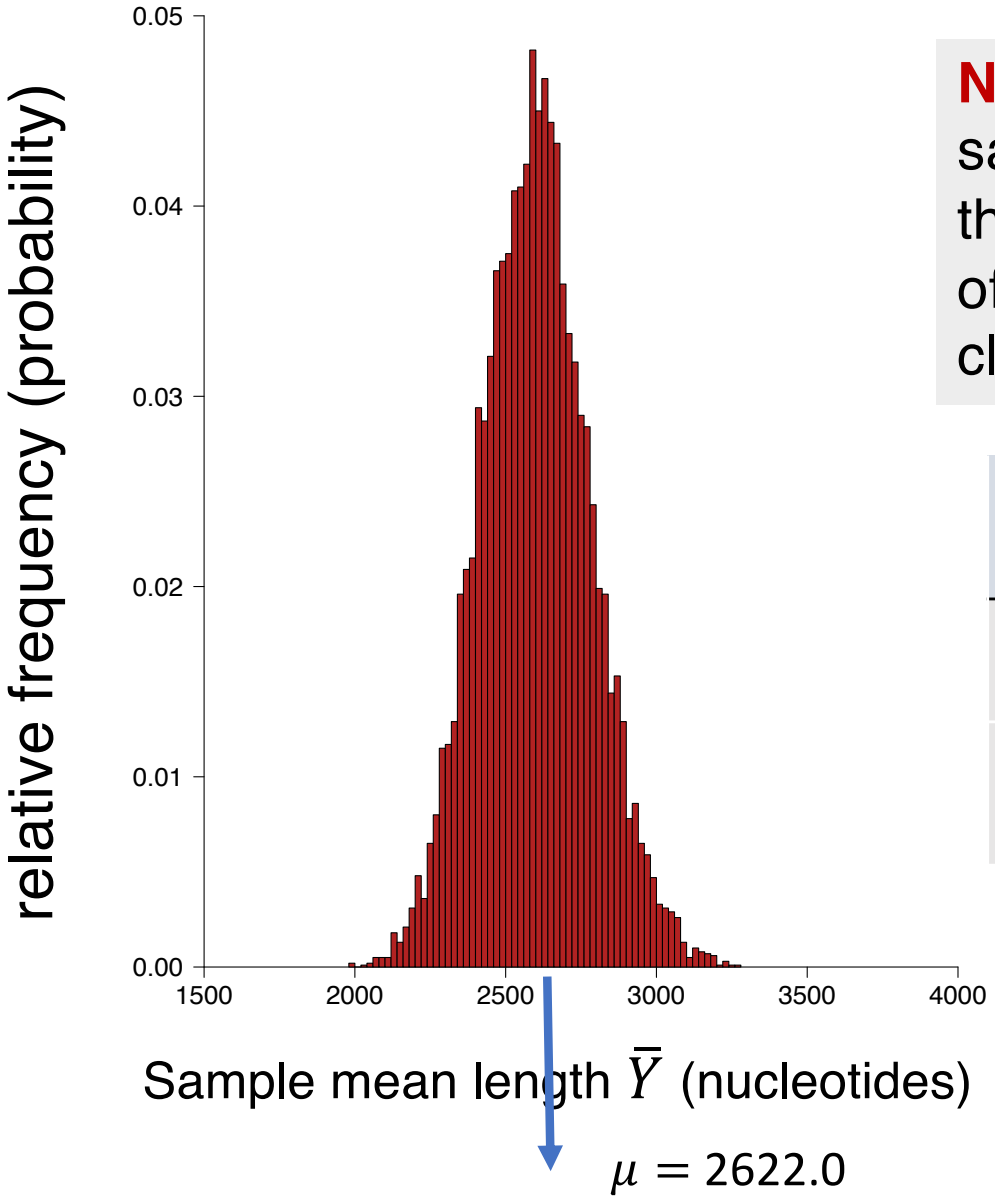
And variation within a single sample can tell use about variation among samples (we will see that in the next lectures).

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



Note 2: The sampling distribution makes it obvious that although the population mean μ is assumed as a constant (2622.0), its estimate $\bar{Y}$ is a variable (i.e., they vary among samples).

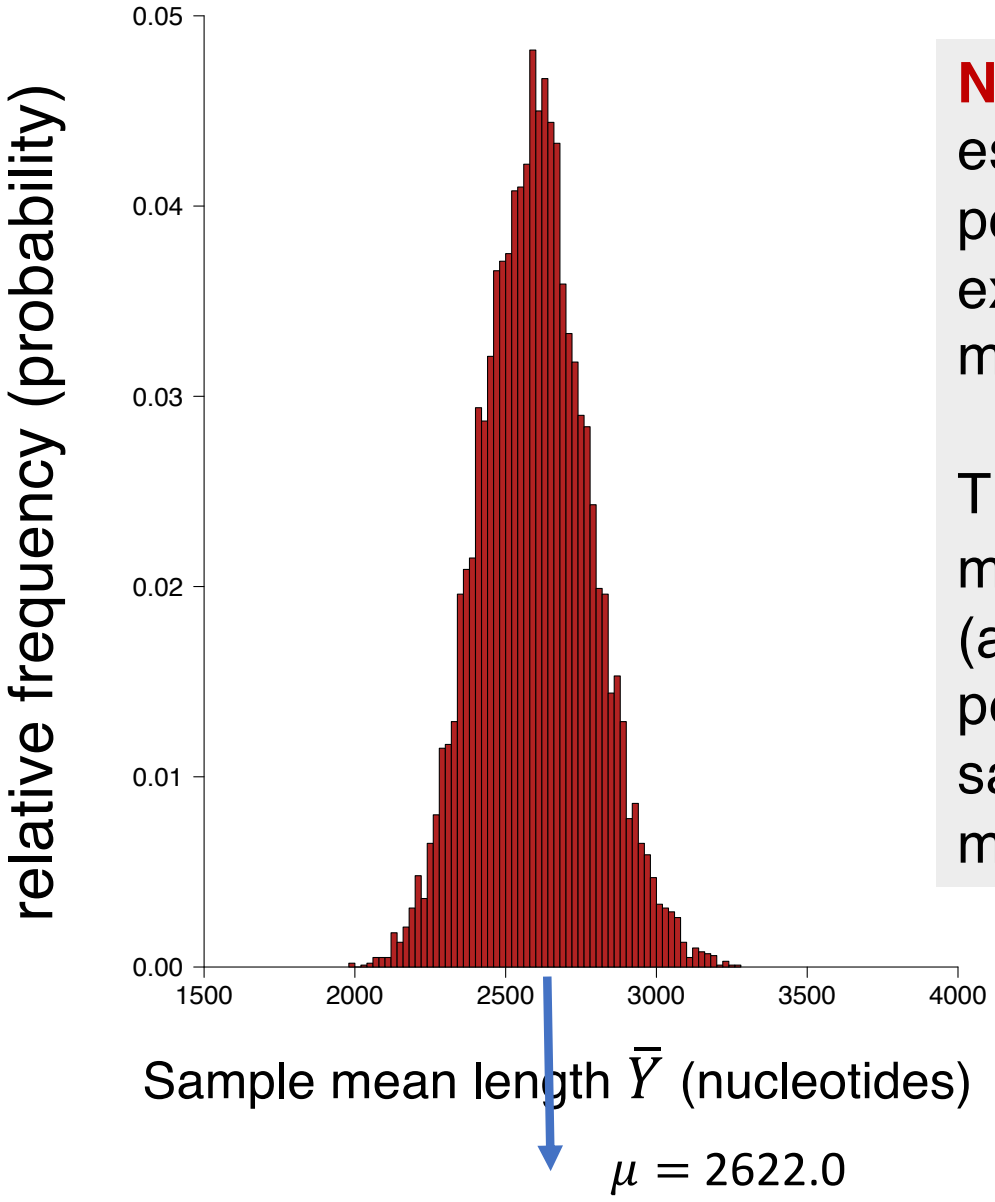
The sampling distribution (probability distribution) of sample means ($\bar{Y}$)



Note 3 (again): The mean of all sample estimates of the mean equals the population mean. Even the mean of 10000 sample means is pretty close.

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

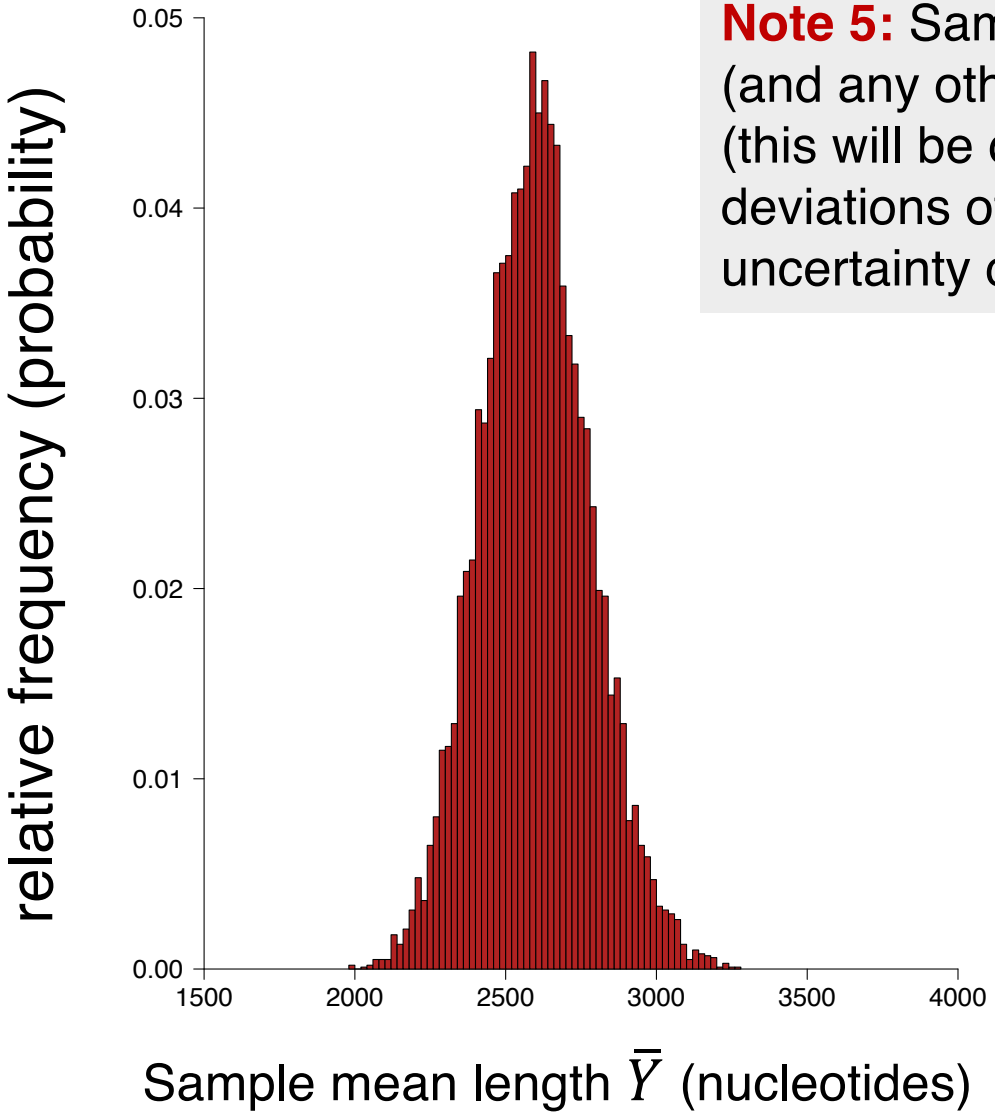
The sampling distribution (probability distribution) of sample means ($\bar{Y}$)



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

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

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



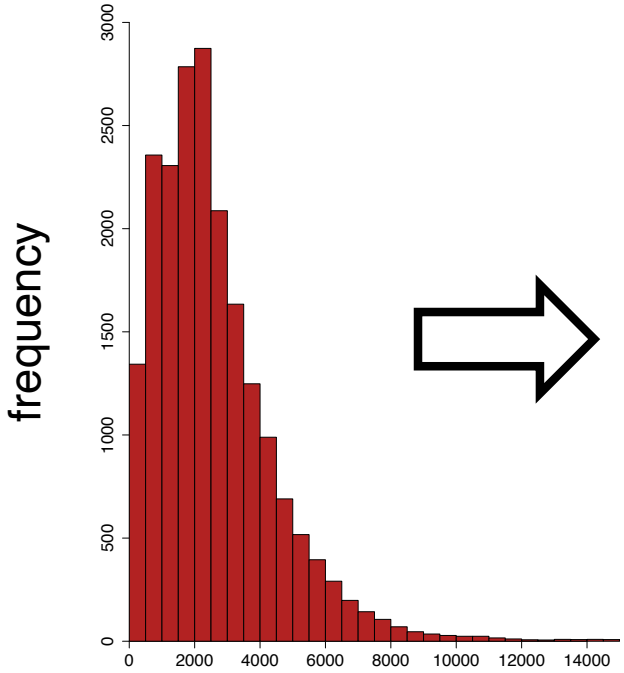
Note 5: Sample values for the standard deviation (and any other statistic) also vary among samples (this will be discussed in our next lecture). Standard deviations of samples are key to estimate uncertainty of a sample mean.

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

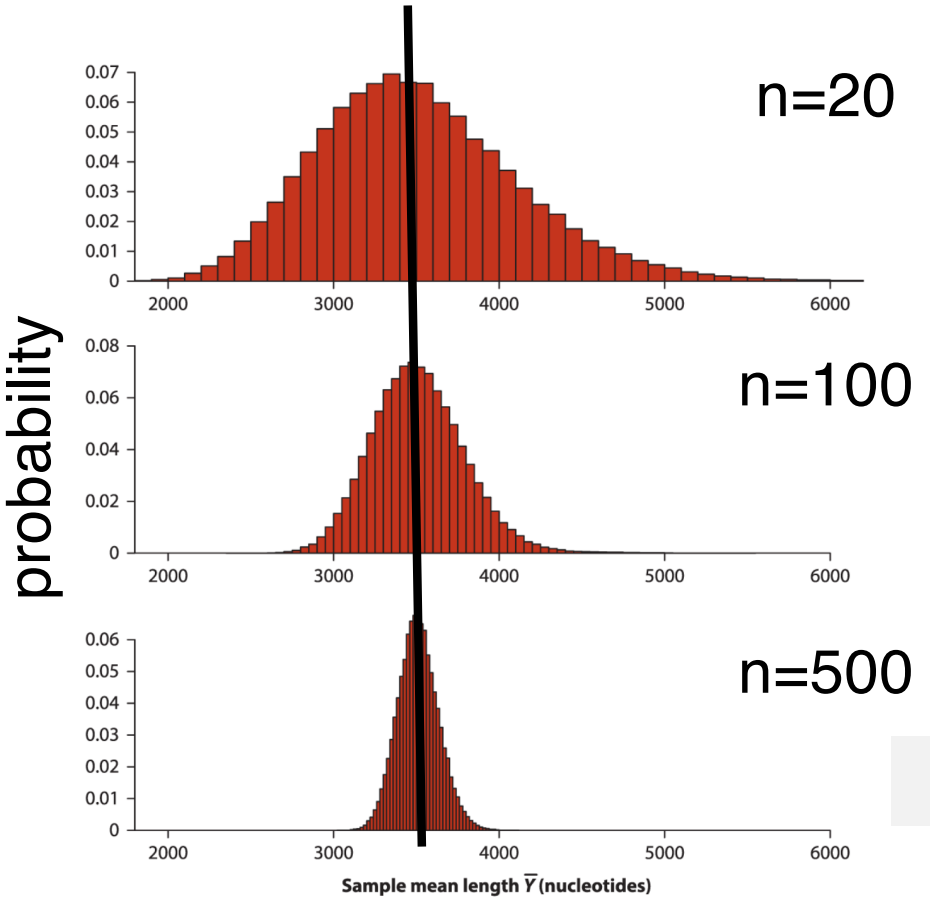
Names	Statistic	Value (nucleotides)
Mean	$\bar{Y}$	2122.3
Standard deviation	s	2423.1

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

Frequency distribution of the gene 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)