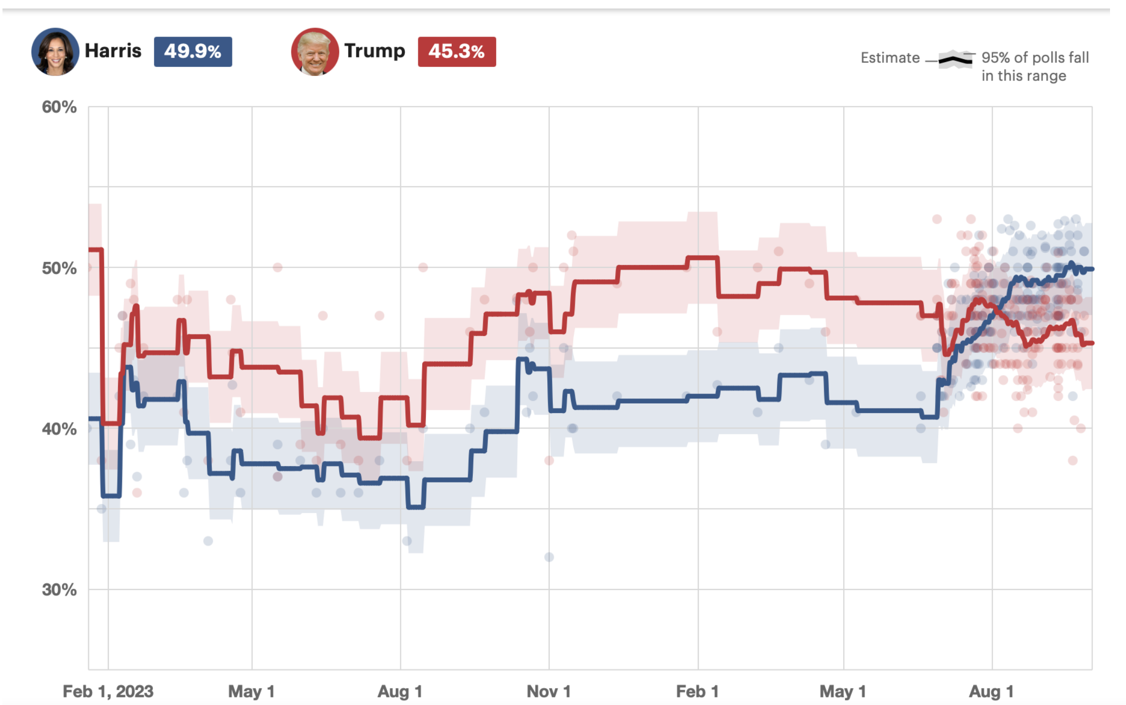
In polling, a 95% confidence interval means that if the same poll were repeated many times (i.e., sampling variation), the true population parameter (e.g., the percentage of people supporting a candidate) would fall within the calculated range in 95% of those polls.

The confidence interval accounts for the margin of error, giving us a range of values that we are fairly confident includes the true result. However, it's important to remember that this does not mean there is a 95% probability that the true value is in that specific range from one poll—it means that, over many polls, 95% of those calculated intervals will capture the true population parameter.



Lecture 10: Videos integrating lecture 10 (confidence intervals part 2) and tutorial 5.

Tutorial 1: Introducing R

Lecture 2: Key Jargon

Lecture 3: Displaying data

Tutorial 2: The R environment

Lecture 4: Frequency distributions

Lecture 5: Describing data

Tutorial 3: Graphs

Lecture 6: Describing data (part 2)

Lecture 7: Sampling variation

Tutorial 4: Describing data

Lecture 8: Sampling distributions

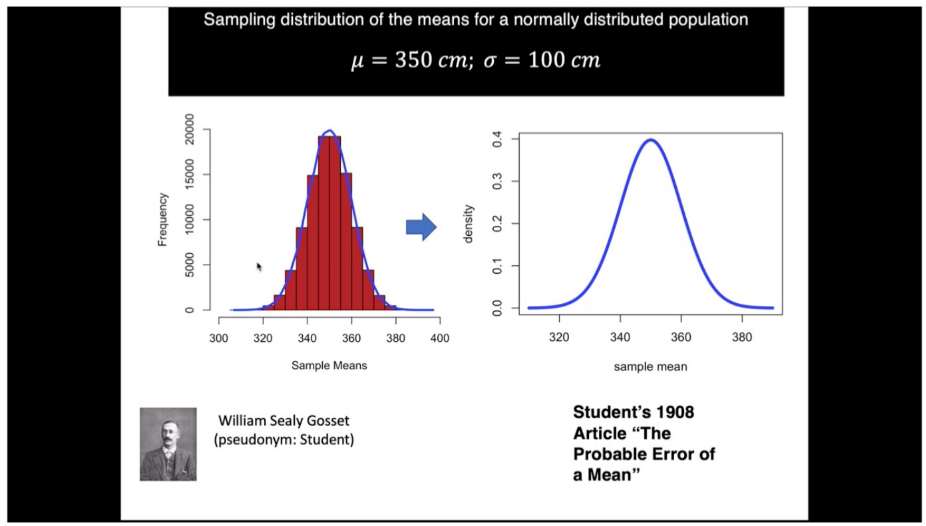
Lecture 9: Confidence Intervals

Tutorial 5: Sampling variation

Lecture 10: Confidence Intervals par…

part 1

General theory: from a computational approach to develop sampling distributions ba
number of sample means to a general statistical theory that can be used for “infinite
number) of sample means.



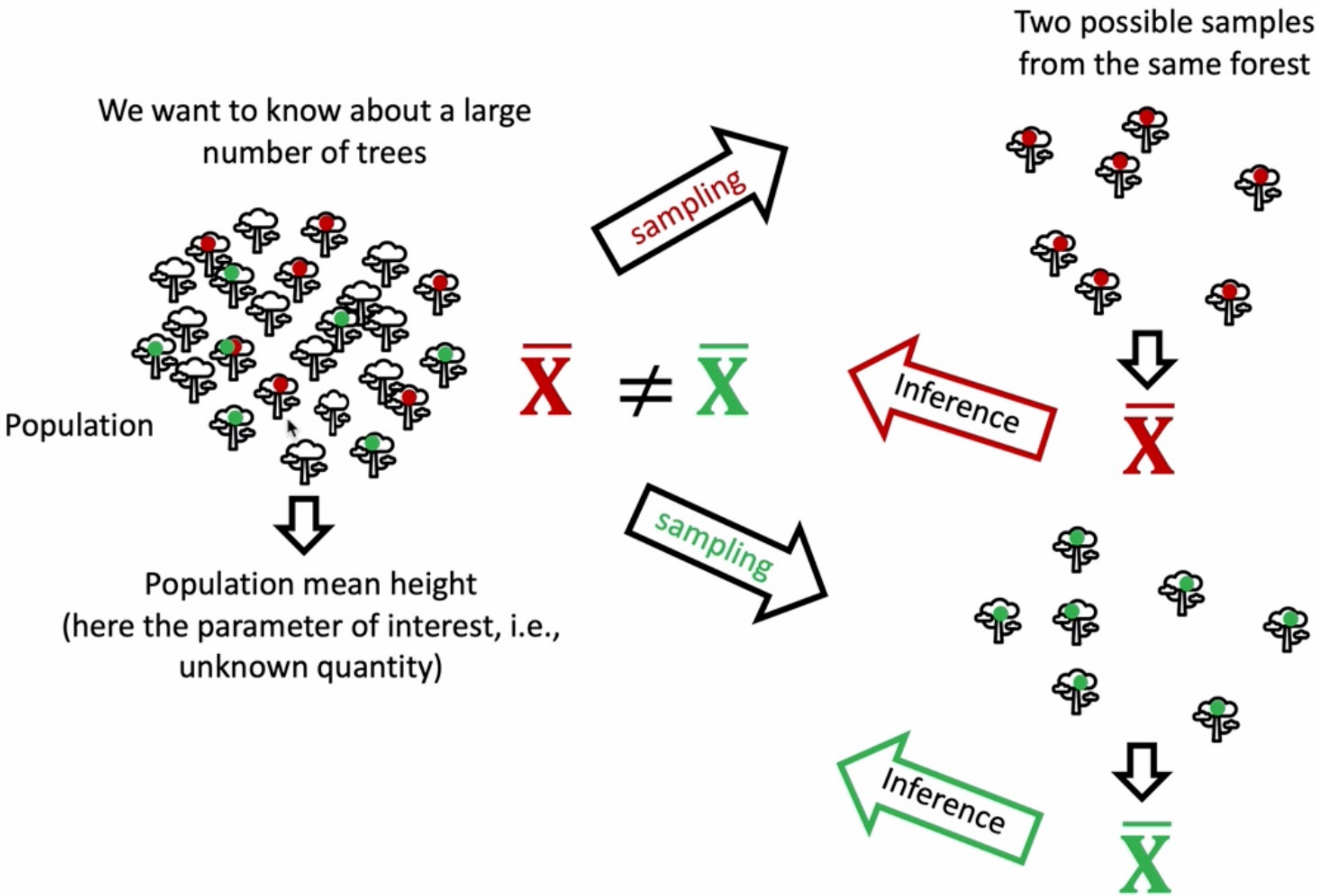
part 2

Estimating with uncertainty
with some certainty

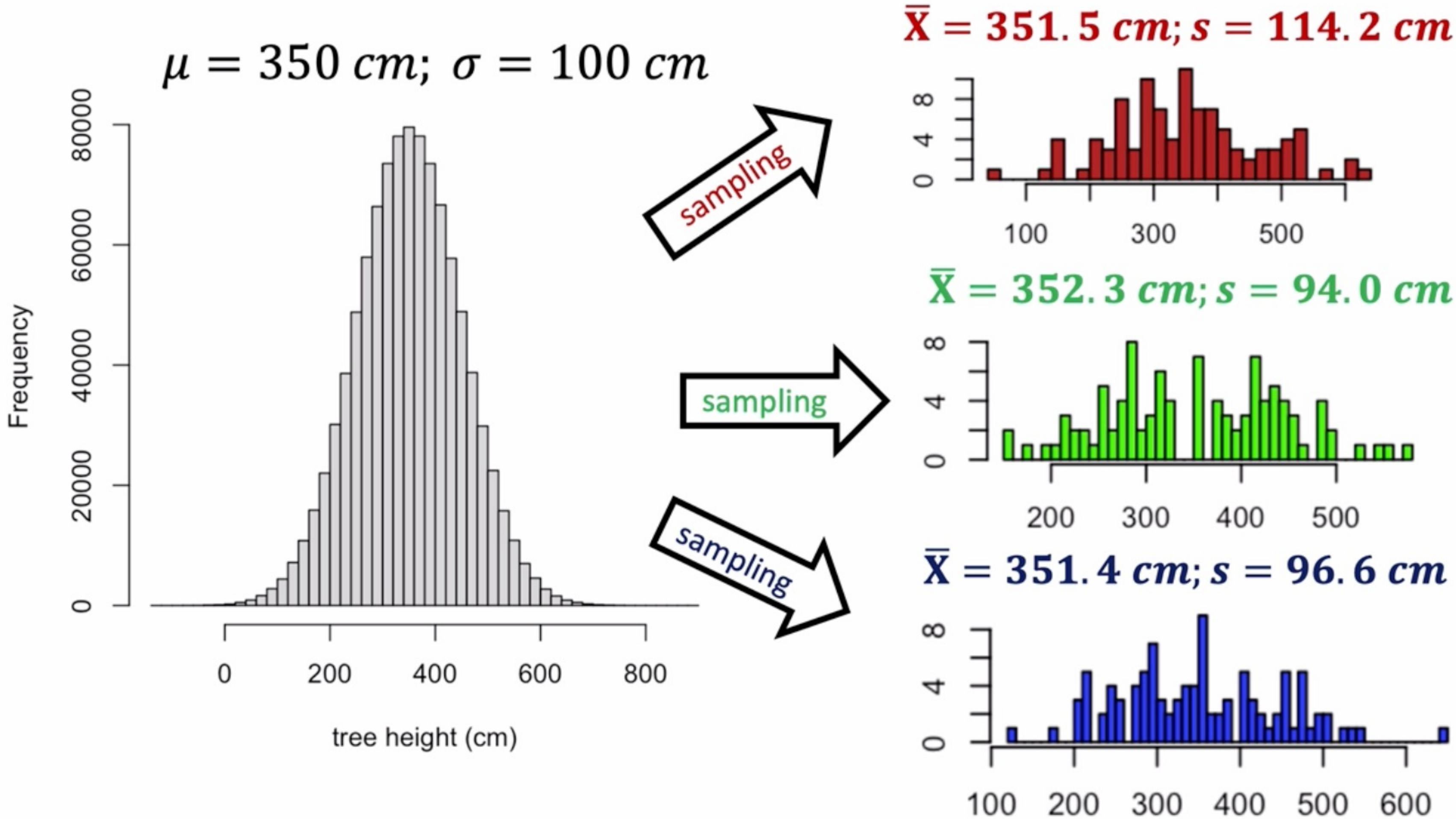
The statistical road: estimate with uncertainty but know how confident you can be!

Sampling variation and Sampling distributions

Sampling variation: two or more sample means of tree height from the same population will always differ from the true population value (parameter)! So we estimate and make inferences with uncertainty (without certainty)

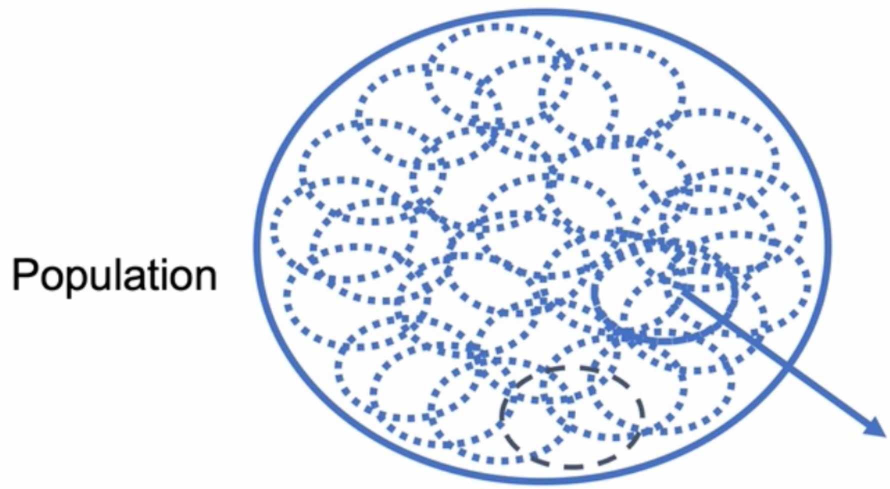


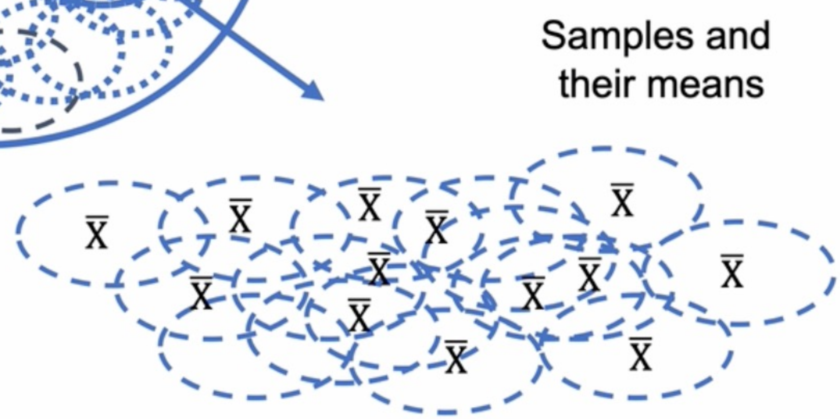
Sampling variation: linking frequency distributions of populations & frequency distributions of samples

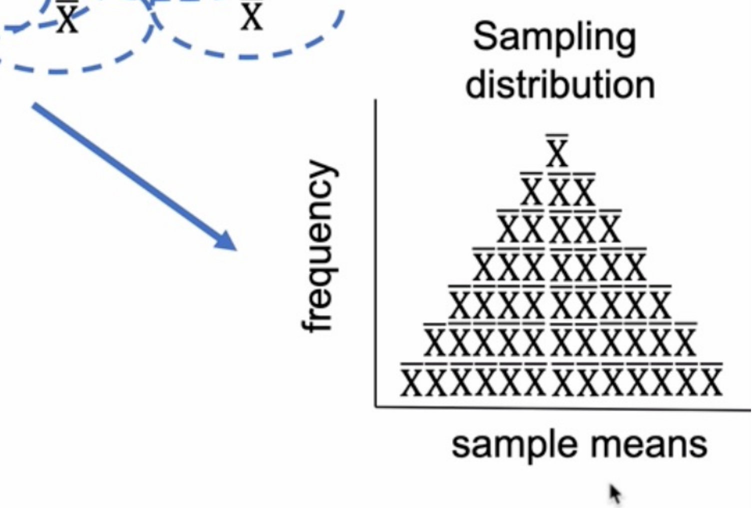


Assume (hypothetically, i.e., for demonstration purposes only) a statistical population of tree heights in cm (1000000 trees) & 3 possible samples of 100 trees each.

Sampling distribution of means.



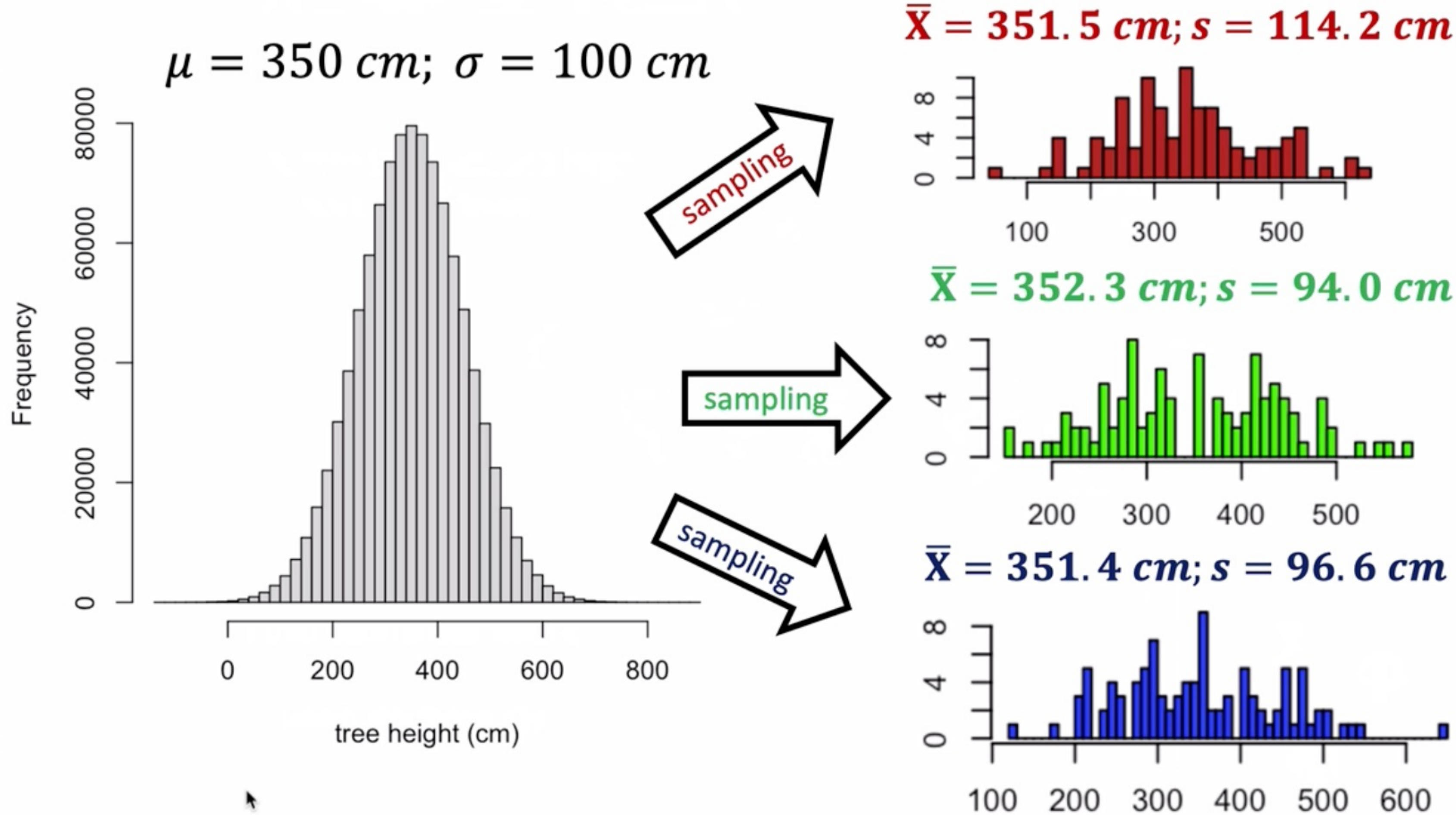




Samples have the same sample size n.

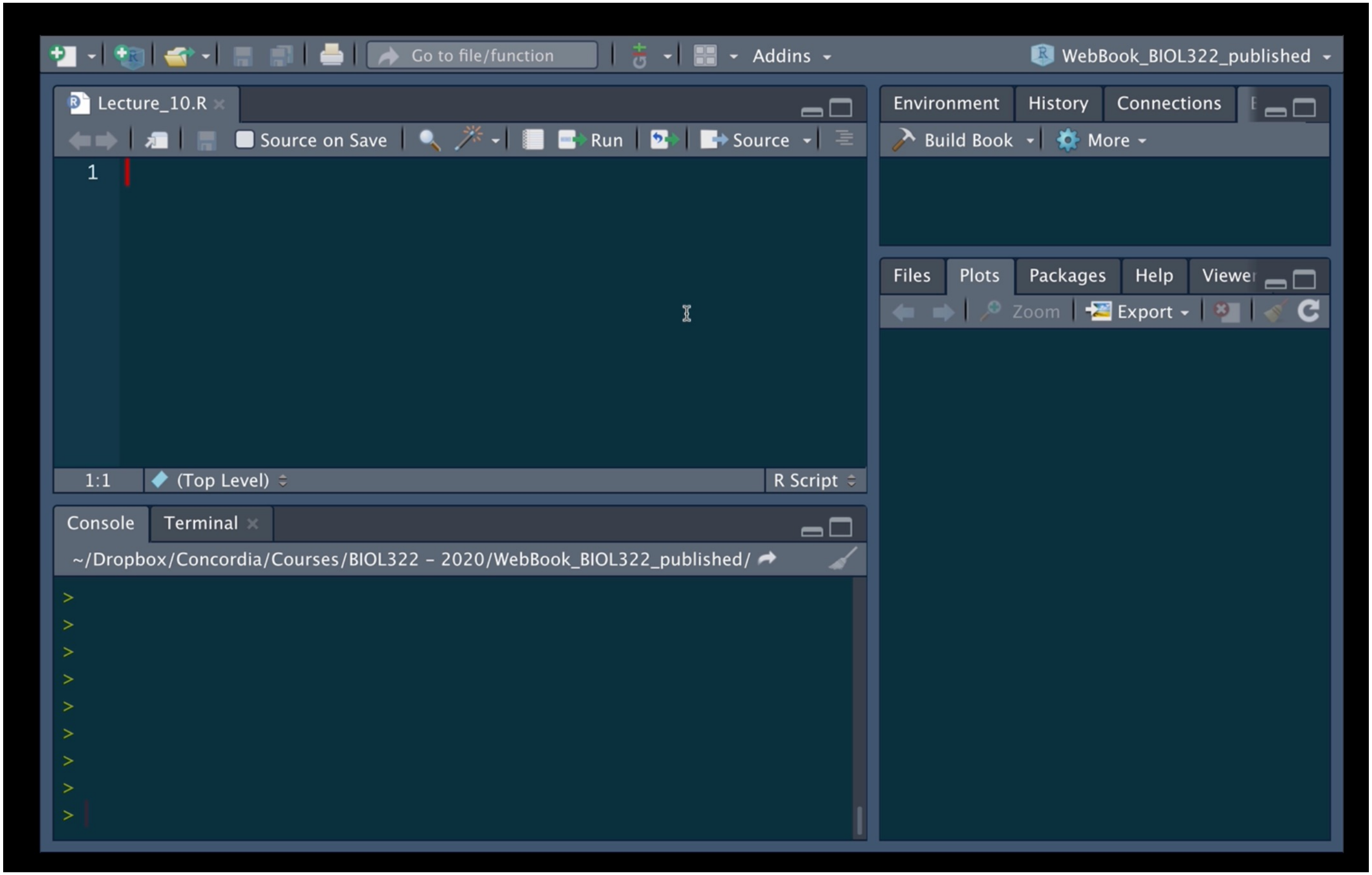
The variation among sample means is due to **sampling error**, i.e., error between the true population mean value and the sample mean value.

Sampling variation: linking frequency distributions of populations & frequency distributions of samples



Assume (hypothetically, i.e., for demonstration purposes only) a statistical population of tree heights in cm (1000000 trees) & 3 possible samples of 100 trees each.

Understanding sampling distributions (of them mean) based on simulations



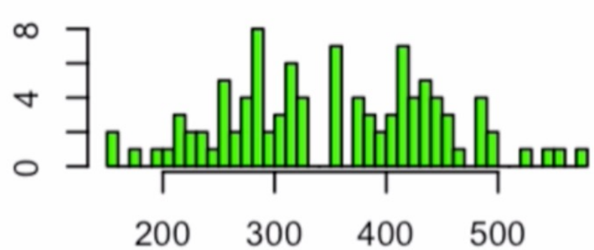
Estimating with uncertainty with certainty (i.e., with some confidence)

A confidence interval is a range of values surrounding the sample estimate that is likely to contain the population parameter.

A large confidence interval (e.g., 95% or 99%) provides a most plausible range for a parameter. Values lying within the interval are most plausible, whereas those outside are less plausible, based **ON A SINGLE sample data alone.**

$$\bar{Y} \pm 2SE_{\bar{Y}} \therefore SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

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



$$352.3 \pm 2 \times \frac{94.0}{\sqrt{100}}$$

333.5 cm 371.1 cm

The properties of sampling distributions: location and spread

The mean of all sample means equal
the population mean:

$$\mu = \sum_{i=i}^{\infty} \frac{\bar{Y}}{\infty}$$

The number of samples is so large that
for mathematical purposes it can be considered infinite (∞)

The properties of sampling distributions: location and spread

Sampling error - the difference between sample means and the population mean. The estimate of this error is the standard deviation of the sampling distribution, i.e., the average difference between all sample means and the true mean:

The standard deviation of the sampling distribution $\sigma_{\bar{Y}}$ is called standard error (SE) and is exactly:

$$\sigma_{\bar{Y}} = \sqrt{\sum_{i=i}^{\infty} \frac{(\bar{Y}_i - \mu)^2}{\infty}} = \mathrm{SE}_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

The number of samples is so large that can be considered infinite (∞)

$\sigma = \textit{the standard deviation of the population}$

The properties of sampling distributions: location and spread

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

Given that we almost never know the population standard deviation, we estimate it with the sample value:

$$\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

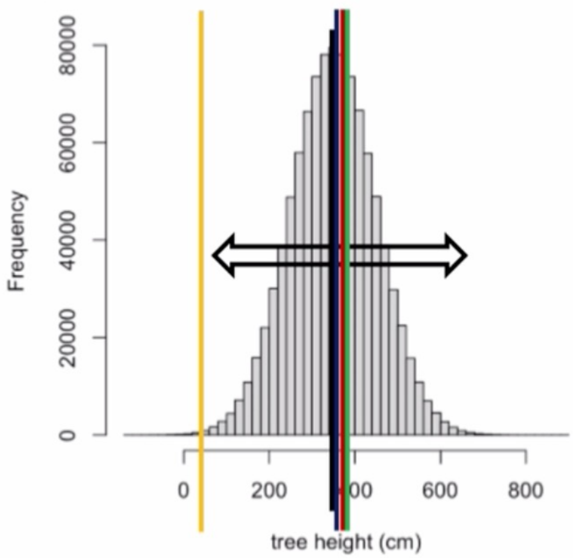
$\sigma_{\bar{Y}}$ = the standard deviation of the sampling distribution of means

σ = the standard deviation of the population

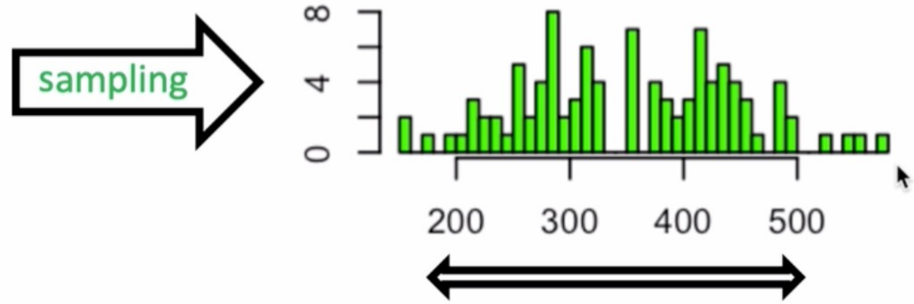
The variation among observations within samples (standard deviation) can inform us about how far sample means in general might be from the true population mean (estimate how wrong one could be).

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

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



$\mu = 350\ cm$

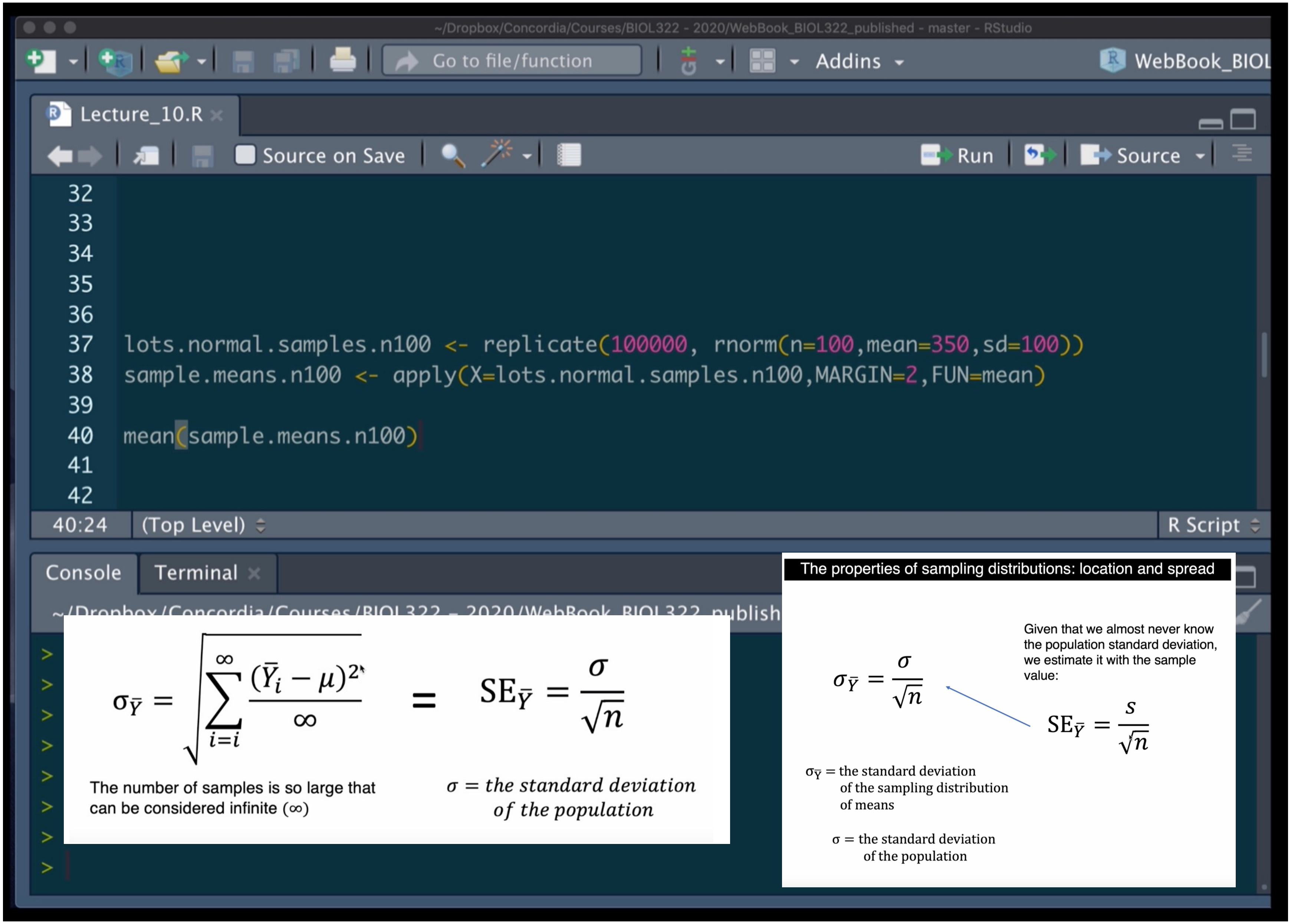


Variation within samples

Variation among samples

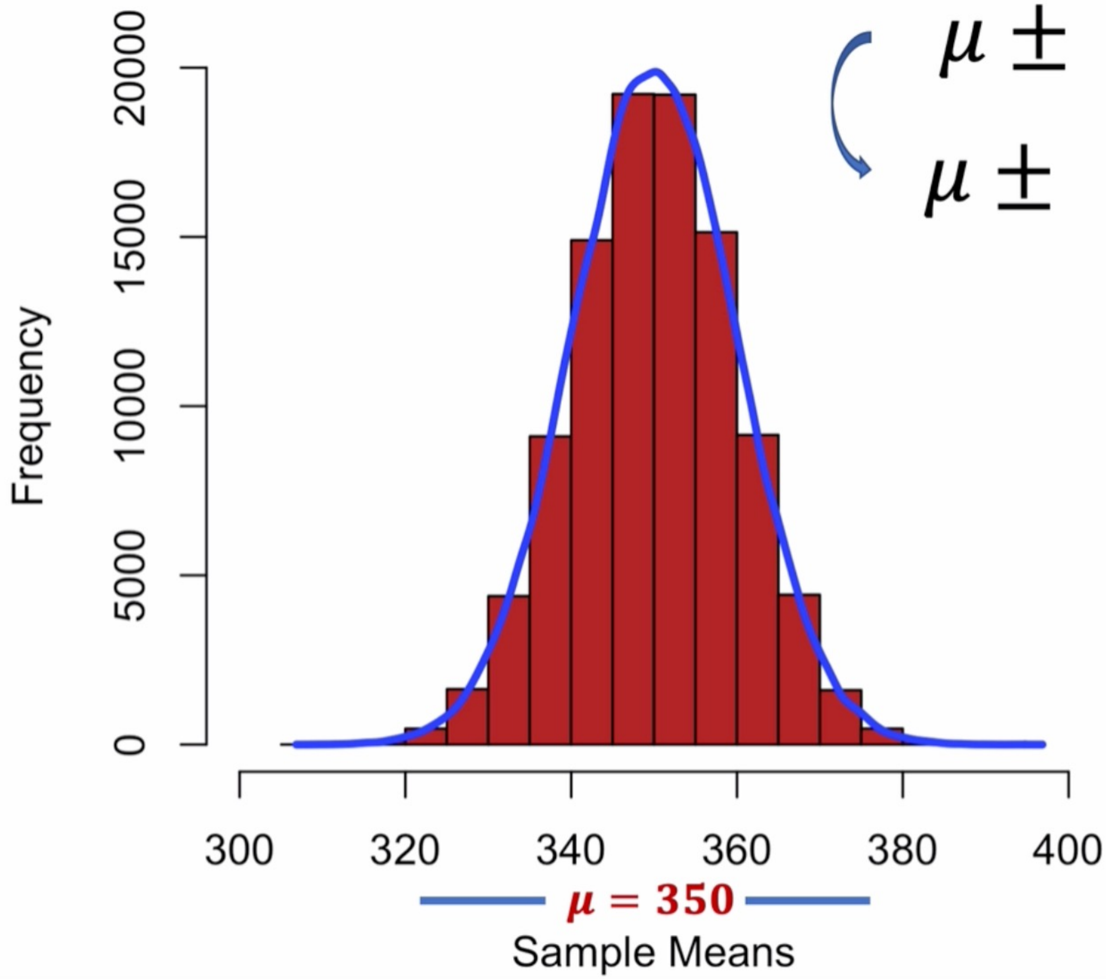
Variation within samples (among observations) can estimate some certainty (confidence) about uncertainty (variation among sample means)

Understanding standard errors via simulations



Sampling distribution of the means for a normally distributed population

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



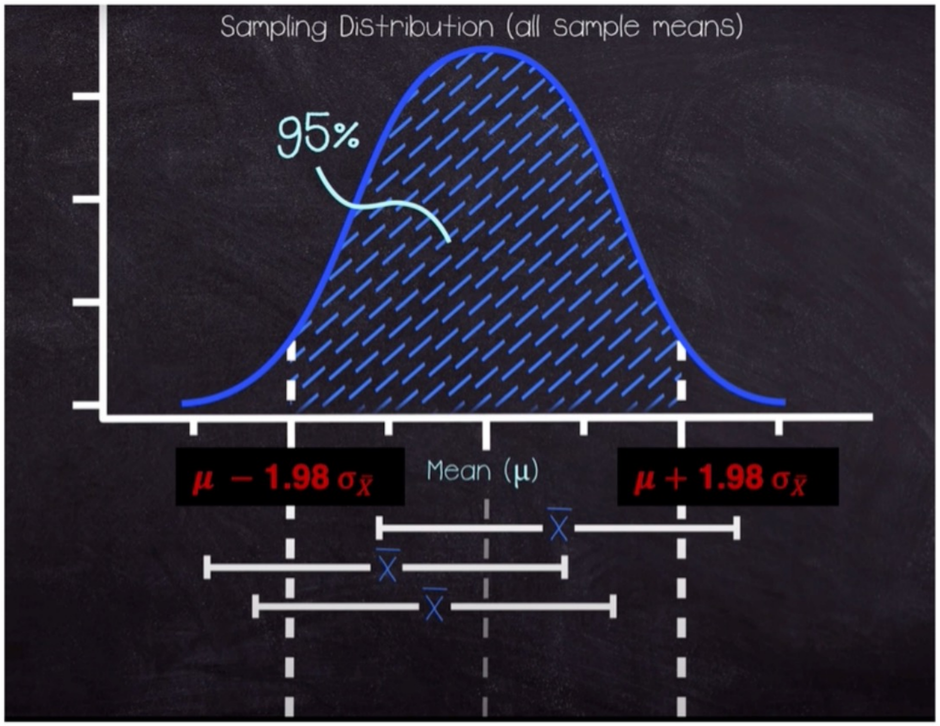
$$\mu \pm\ 2\times\sigma_{\bar{Y}}$$

$$\mu \pm\ 1.98\times\sigma_{\bar{Y}}$$

Student's 1908 Article "The Probable Error of a Mean"

William Sealy Gosset
(pseudonym: Student)

We use the sampling distribution of all sample means to calculate confidence intervals



Because the distributions is symmetric, if the the $\mu \pm$ **1.98** x $\sigma_{\bar{X}}$ encompasses 95% of the sample means $\bar{X}$, then 95% of $\bar{X} \pm$ **1.98** x $\sigma_{\bar{X}}$ will encompass the population mean!

How many possible samples of 100 trees out of 100000 trees?

1e+15 (zeros)

The **human body** consists of some 37.2 trillion **cells**

(3.72e+13 zeros**)**

Sampling variation: linking frequency distributions of populations & frequency distributions of samples

How many possible samples of 100 trees out of 1000000 trees?

1e+15 (zeros)

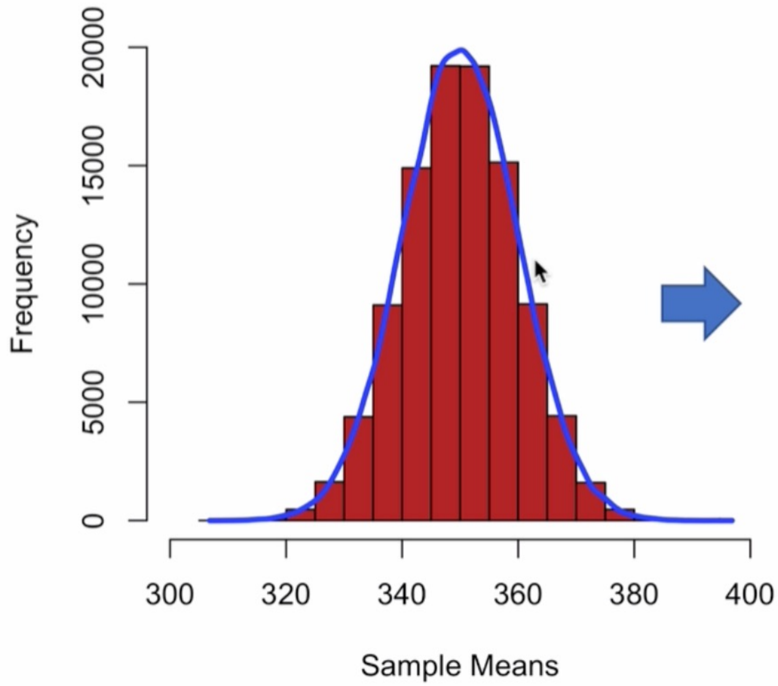
How many possible samples of 100 trees out of 1000000?

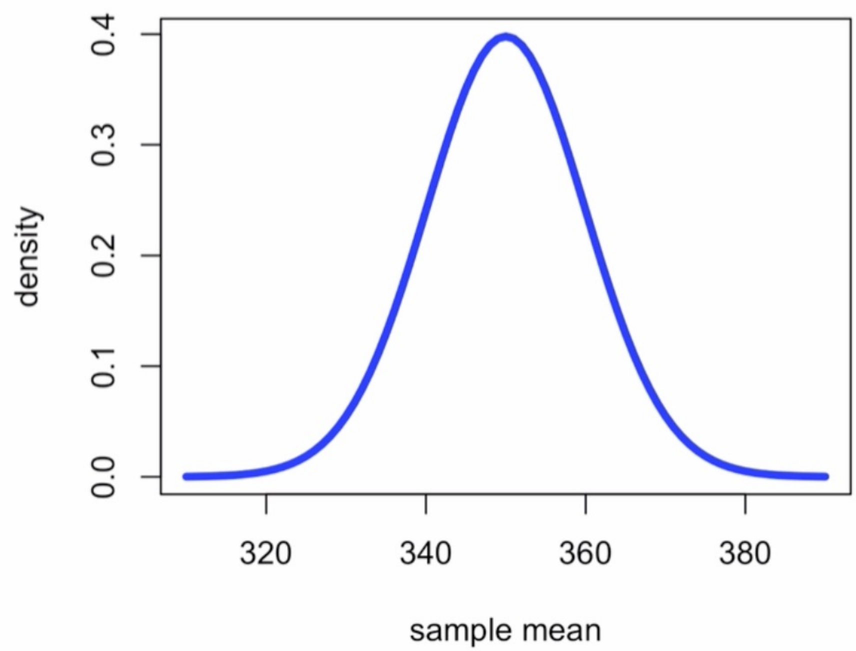
10768272362e+432 (zeros)

The **human body** consists of some 37.2 trillion **cells**
(3.72e+13 zeros)

Sampling distribution of the means for a normally distributed population

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





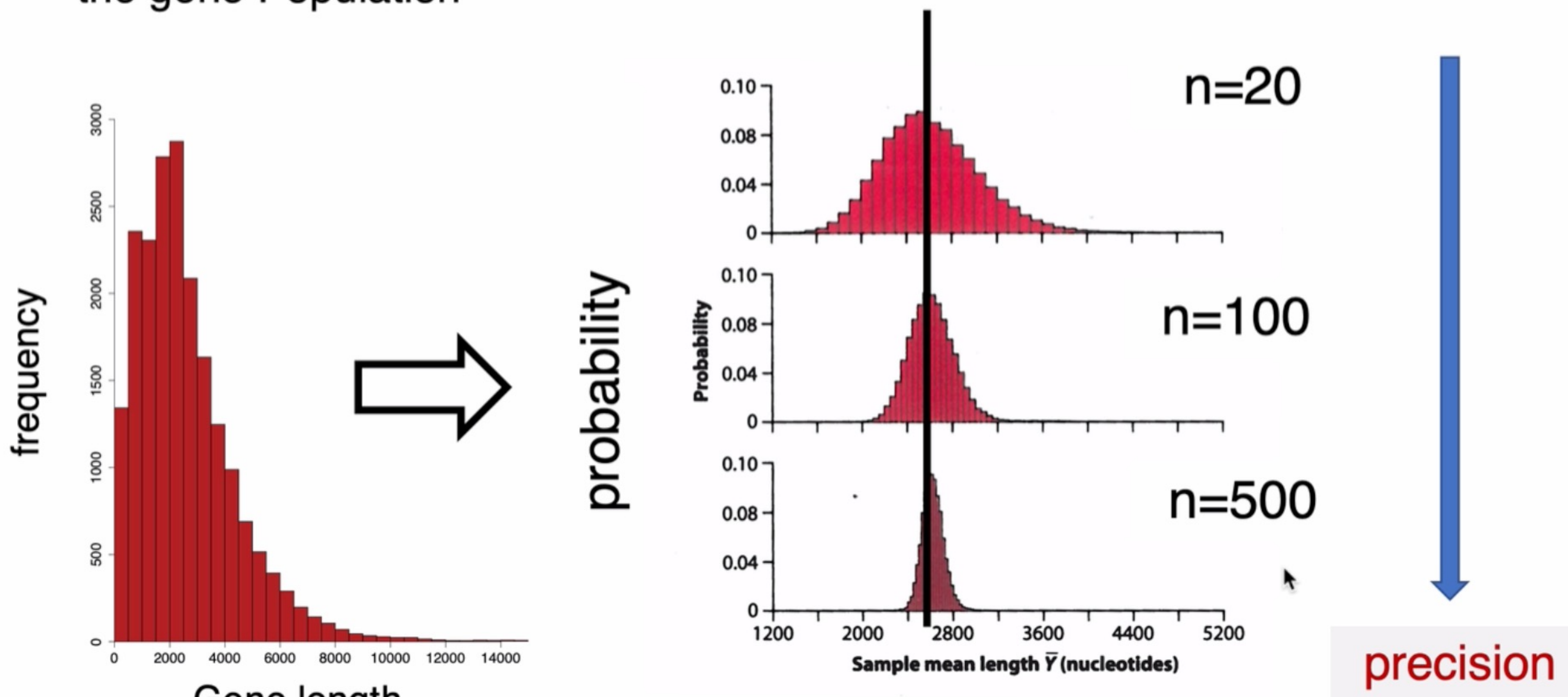
William Sealy Gosset
(pseudonym: Student)

Student's 1908 Article "The Probable Error of a Mean"

Sample size increases precision

Frequency distribution of the gene Population

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



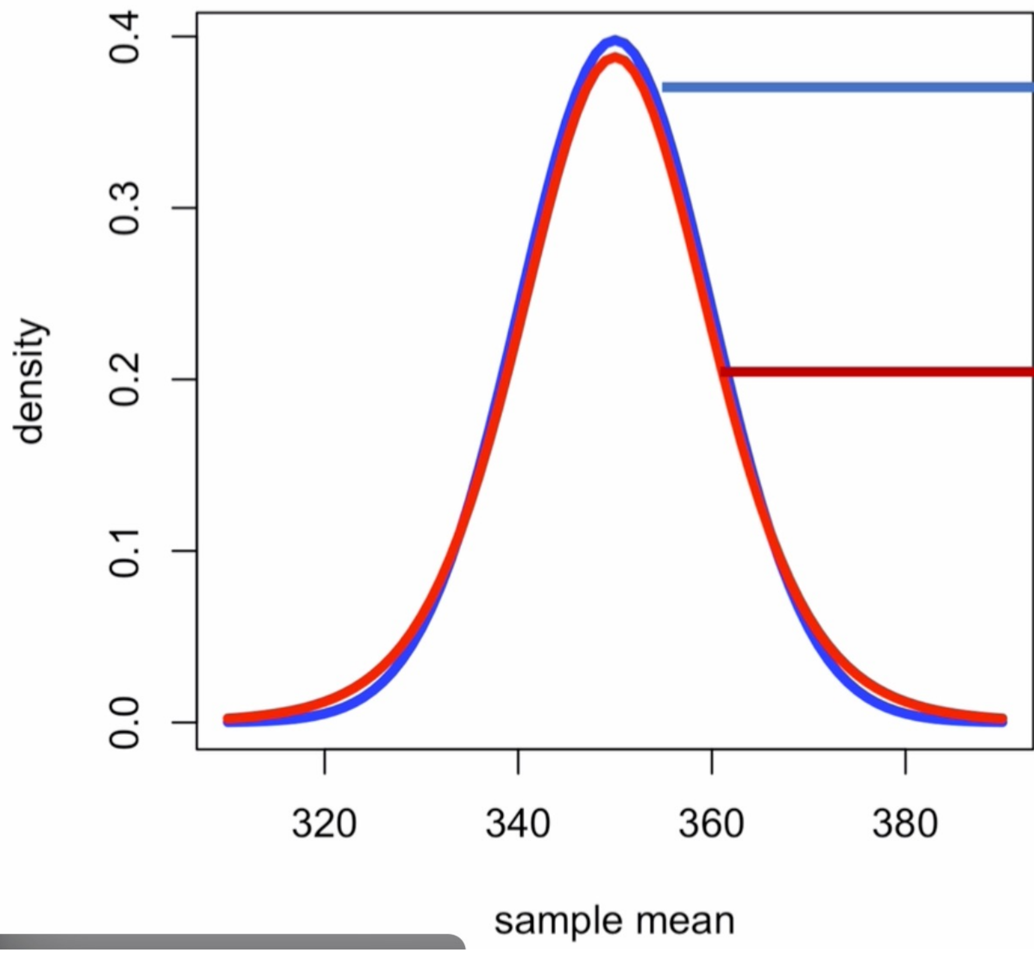
Gene length
(number of nucleotides)

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

Whitlock & Schluter, 2nd edition; 3rd edition has a different set of genes.

Sampling distribution of the means for a normally distributed population

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



$n = 100$

$\mu \pm\ 1.98 \times \sigma_{\bar{Y}}$

$n = 10$

$\mu \pm\ 2.26 \times \sigma_{\bar{Y}}$

The properties of sampling distributions: location and spread

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

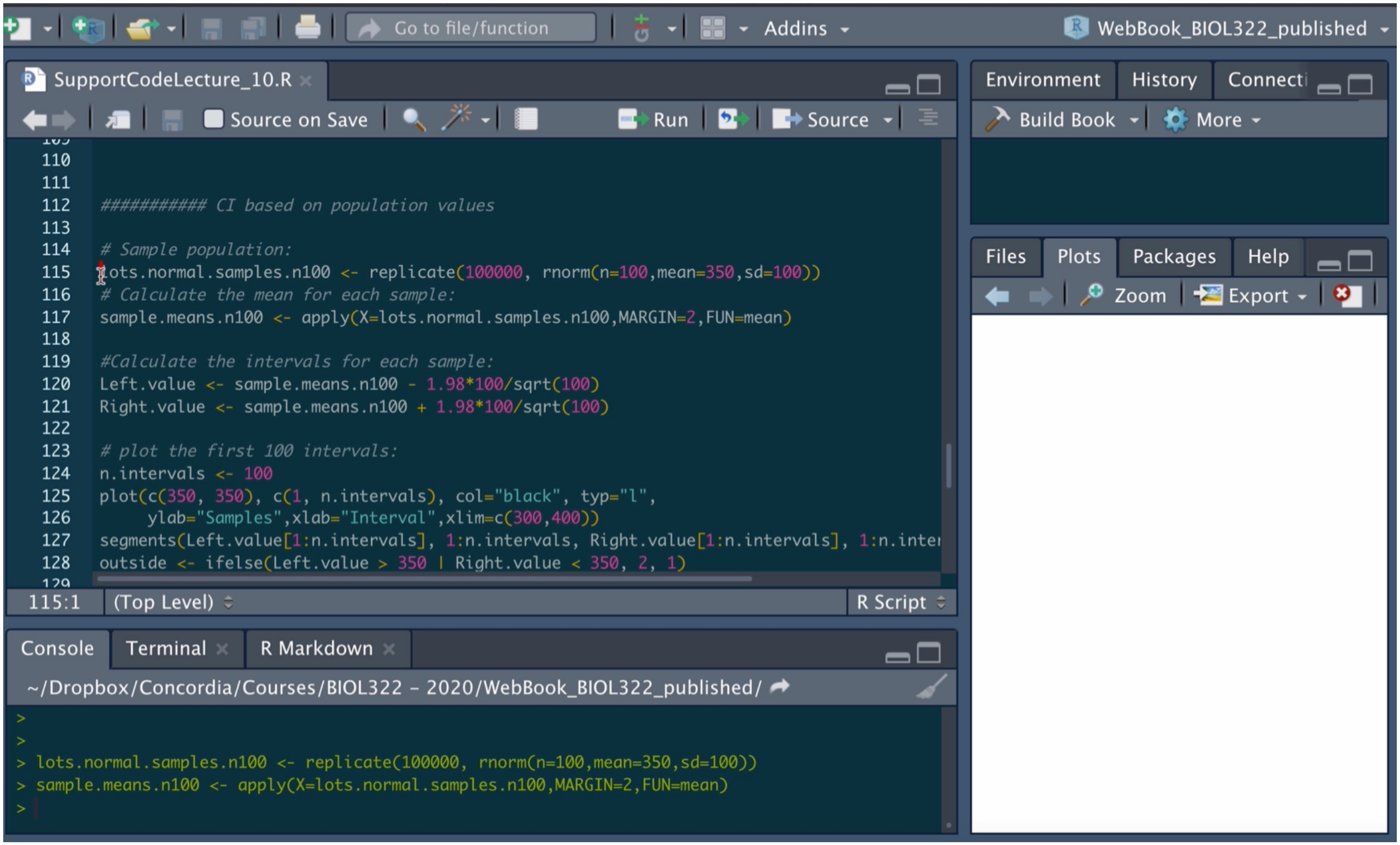
Given that we almost never know the population standard deviation, we estimate it with the sample value:

$$\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

$\sigma_{\bar{Y}}$ = the standard deviation of the sampling distribution of means

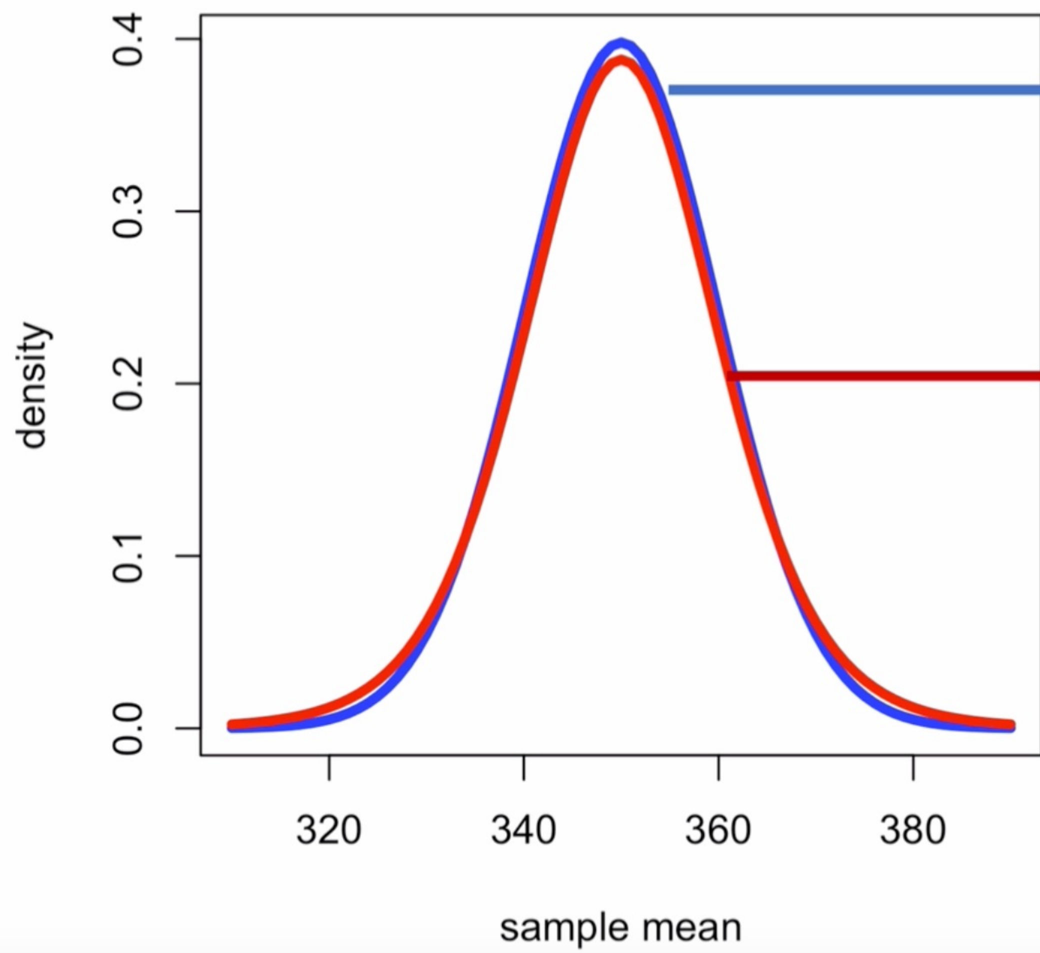
σ = the standard deviation of the population

Connecting standard errors to confidence intervals
Understanding confidence intervals via simulations



Sampling distribution of the means for a normally distributed population
IS NOT NORMALLY DISTRIBUTED; it is t-distributed

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



$n = 100$

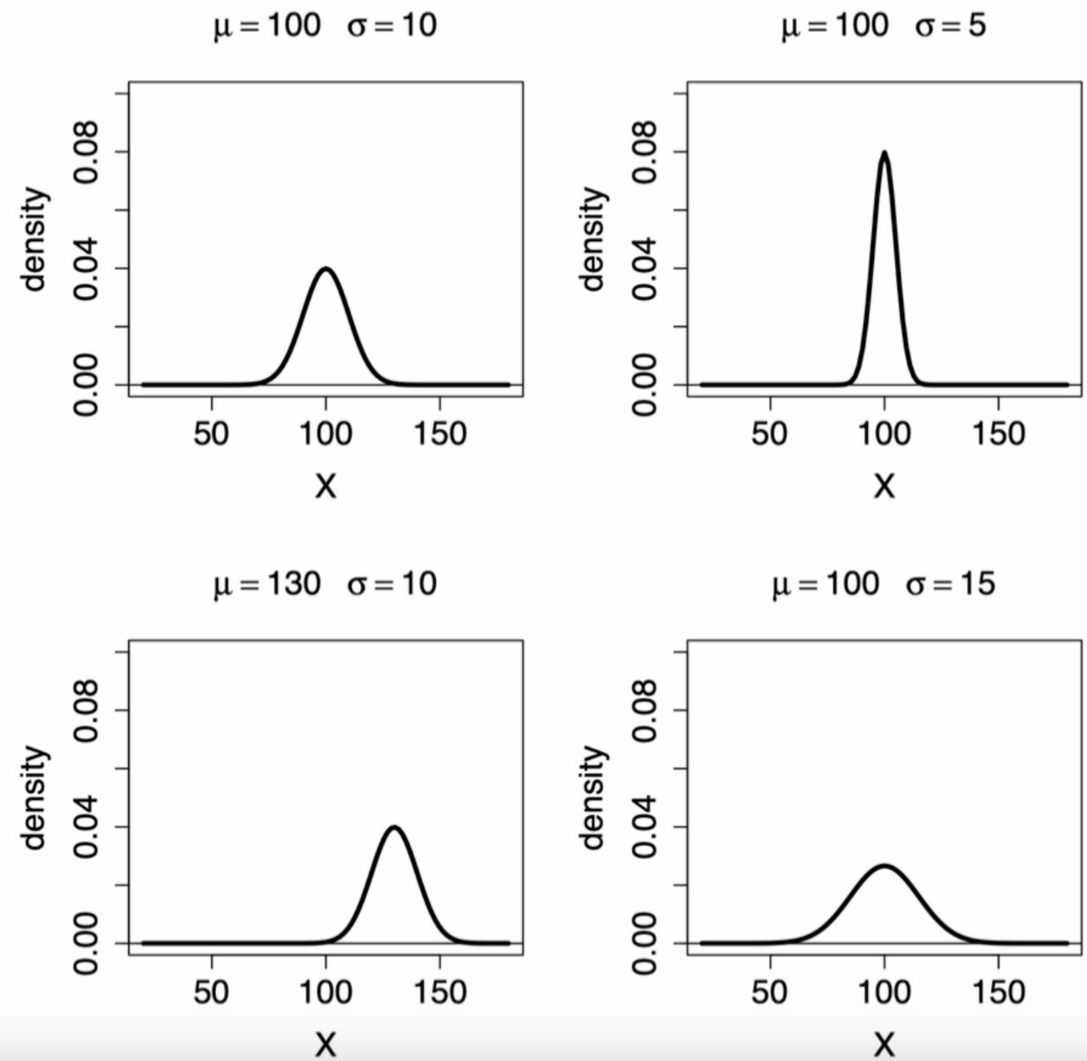
$\mu \pm\ 1.98 \times \sigma_{\bar{Y}}$

$n = 10$

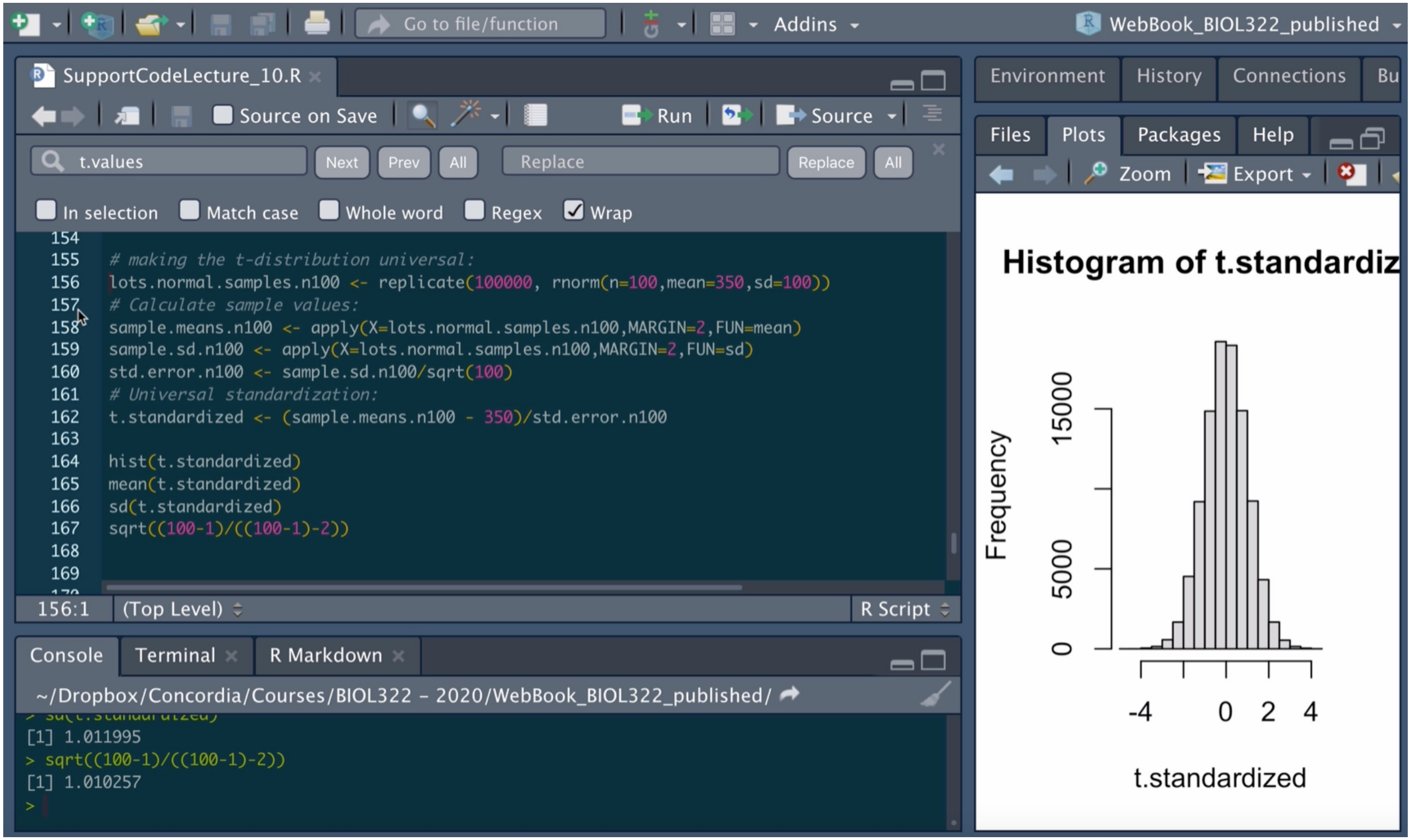
$\mu \pm\ 2.26 \times \sigma_{\bar{Y}}$

There are infinite normal distributions that can lead to infinite sampling distributions of means and associated t-distributions based on the combination of simply two parameters: μ and σ and a constant n (sample size).

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

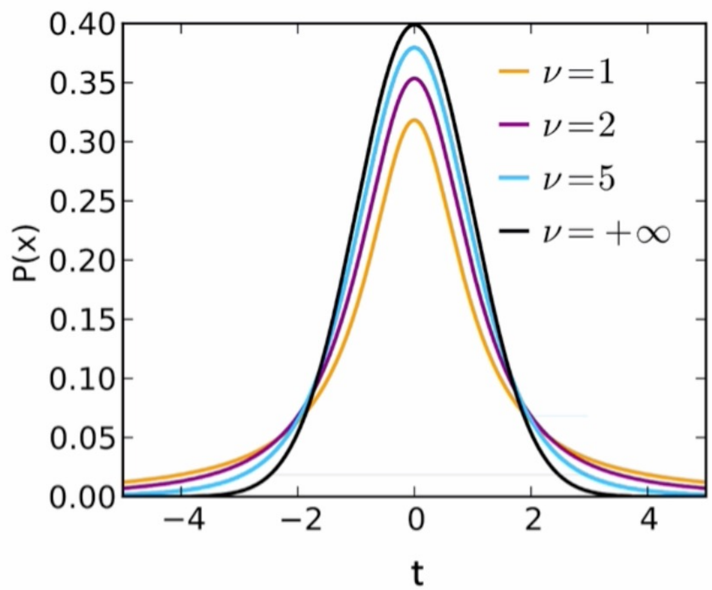


- Source: Jonathan Marchini


Go to file/function
Addins
WebBook_BIOL322_published
SupportCodeLecture_10.R
Source on Save
Run
Source
t.values
Next
Prev
All
Replace
Replace
All
In selection
Match case
Whole word
Regex
Wrap
# making the t-distribution universal:
lots.normal.samples.n100 <- replicate(100000, rnorm(n=100,mean=350,sd=100))
# Calculate sample values:
sample.means.n100 <- apply(X=lots.normal.samples.n100,MARGIN=2,FUN=mean)
sample.sd.n100 <- apply(X=lots.normal.samples.n100,MARGIN=2,FUN=sd)
std.error.n100 <- sample.sd.n100/sqrt(100)
# Universal standardization:
t.standardized <- (sample.means.n100 - 350)/std.error.n100
hist(t.standardized)
mean(t.standardized)
sd(t.standardized)
sqrt((100-1)/((100-1)-2))
156:1
(Top Level)
R Script
Console
Terminal
R Markdown
~/Dropbox/Concordia/Courses/BIOL322 - 2020/WebBook_BIOL322_published/
[1] 1.011995
> sqrt((100-1)/((100-1)-2))
[1] 1.010257
>
Environment
History
Connections
Files
Plots
Packages
Help
Zoom
Export
Histogram of t.standardiz
Frequency
0
5000
15000
-4
0
2
4
t.standardized

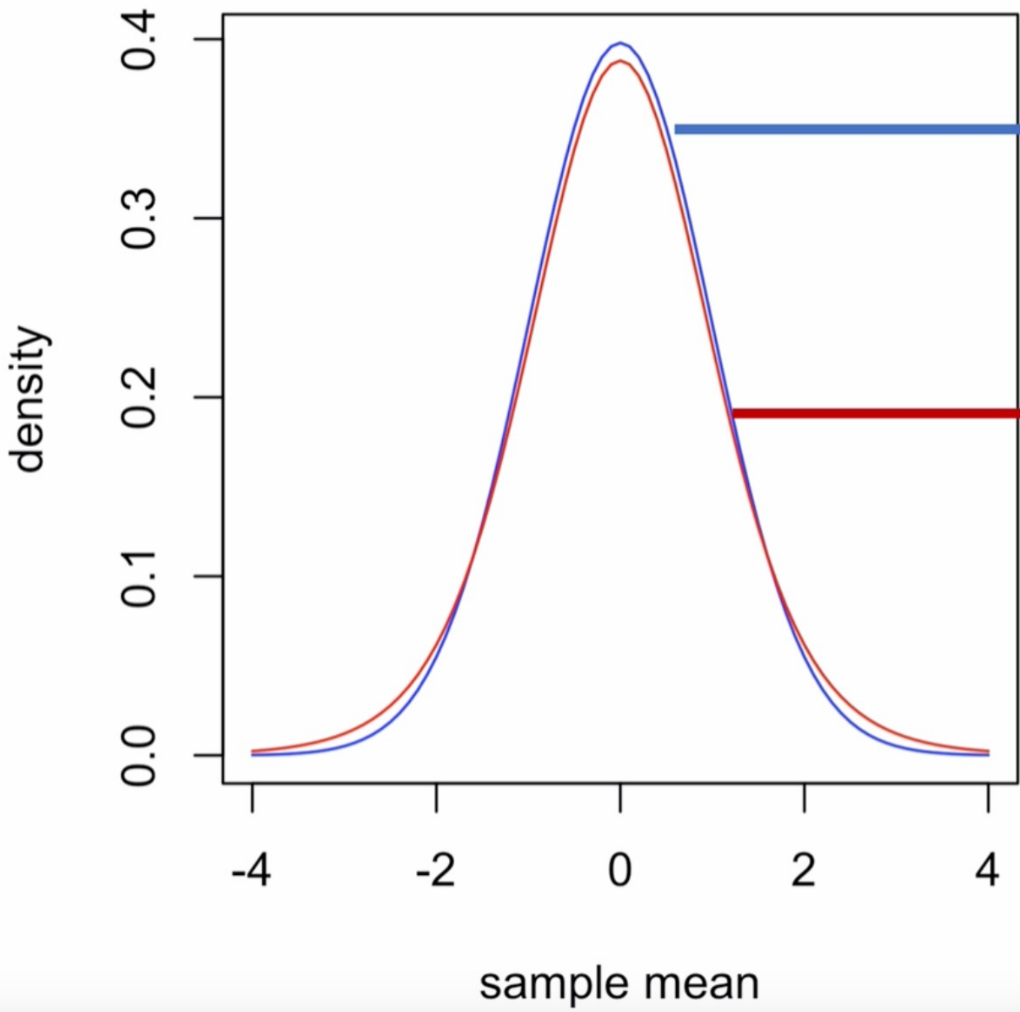
By now, you should suspect that one of the inconveniences is that the exact value needed to be multiplied by SE to create 95% confidence intervals changes as a function of sample size.

The sampling distribution of means that varies as a function of the sample size (here v = degrees of freedom; v = n - 1).



This *t* distribution is a sampling distribution of the the number of sample standard errors away from the mean (now always 0 after the standardization) necessary to produce a confidence interval of the desired coverage (e.g., 95%).

$$t = \frac{\bar{X}_i - \mu}{SE_{\bar{X}_i}} \longrightarrow \bar{X}_i \pm \; t \times SE_{\bar{X}_i}$$



$n = 100$

$\overline{X}_i \pm 1.98 \times SE_{\overline{X}_i}$

$n = 10$

$\overline{X}_i \pm 2.26 \times SE_{\overline{X}_i}$

How to find the appropriate values of t?

the old days of tables allow to understand the principle – in practice (today) we use software (e.g., R).

Degrees of freedom

Two-sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.080	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318

Assume a sample size of n = 9, then the degrees of freedom would be 8 for the t value to calculate the confidence interval for the sample mean.

$$\bar{\mathrm{X}}_{\mathrm{i}} \pm t \times \mathrm{SE}_{\bar{\mathrm{X}}_{\mathrm{i}}} \therefore \bar{\mathrm{X}}_{i} \pm 2.306 \frac{s_i}{\sqrt{9}}$$