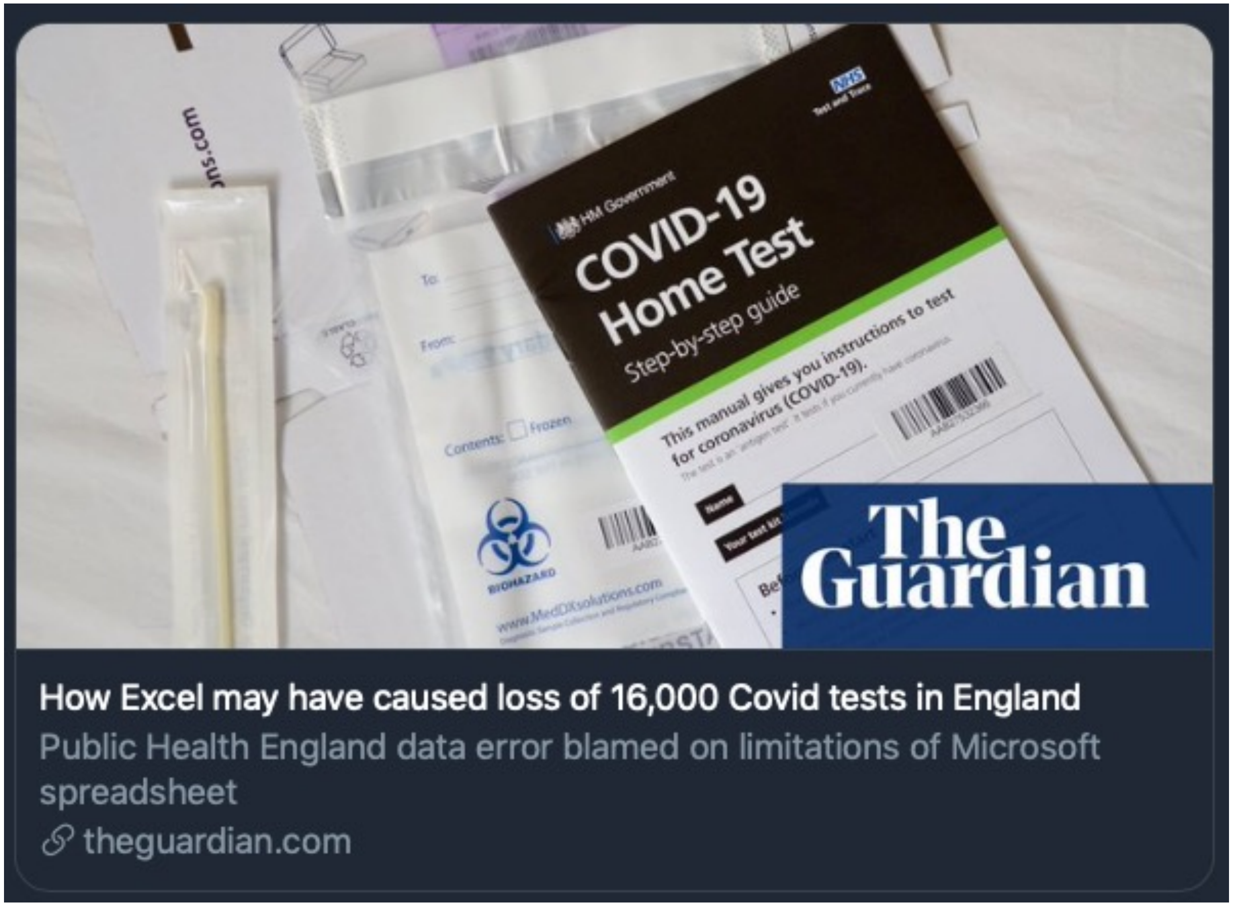
A snap demonstration why data entry and file formats are critical – .csv would never allow that to happen!

16,000 Covid cases in the UK were missed because an 'Excel spreadsheet maxed out and wouldn't update' - meaning thousands of potentially infected contacts were not performed. Details were not passed to contact tracers, meaning people exposed to the virus were not tracked down.



How Excel may have caused loss of 16,000 Covid tests in England

Public Health England data error blamed on limitations of Microsoft spreadsheet

theguardian.com

Let’s go to our WebBook and watch What is Confidence Interval? By Mike Marin

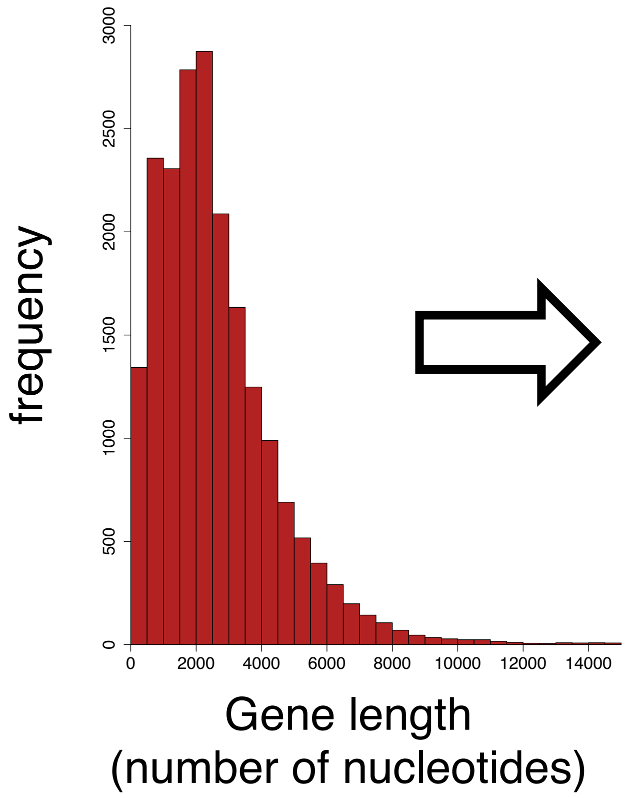
https://www.youtube.com/watch?v=9jTJD5SLweY

Lecture 9: Estimating with uncertainty, but with a degree of certainty (i.e., with some confidence), part 2

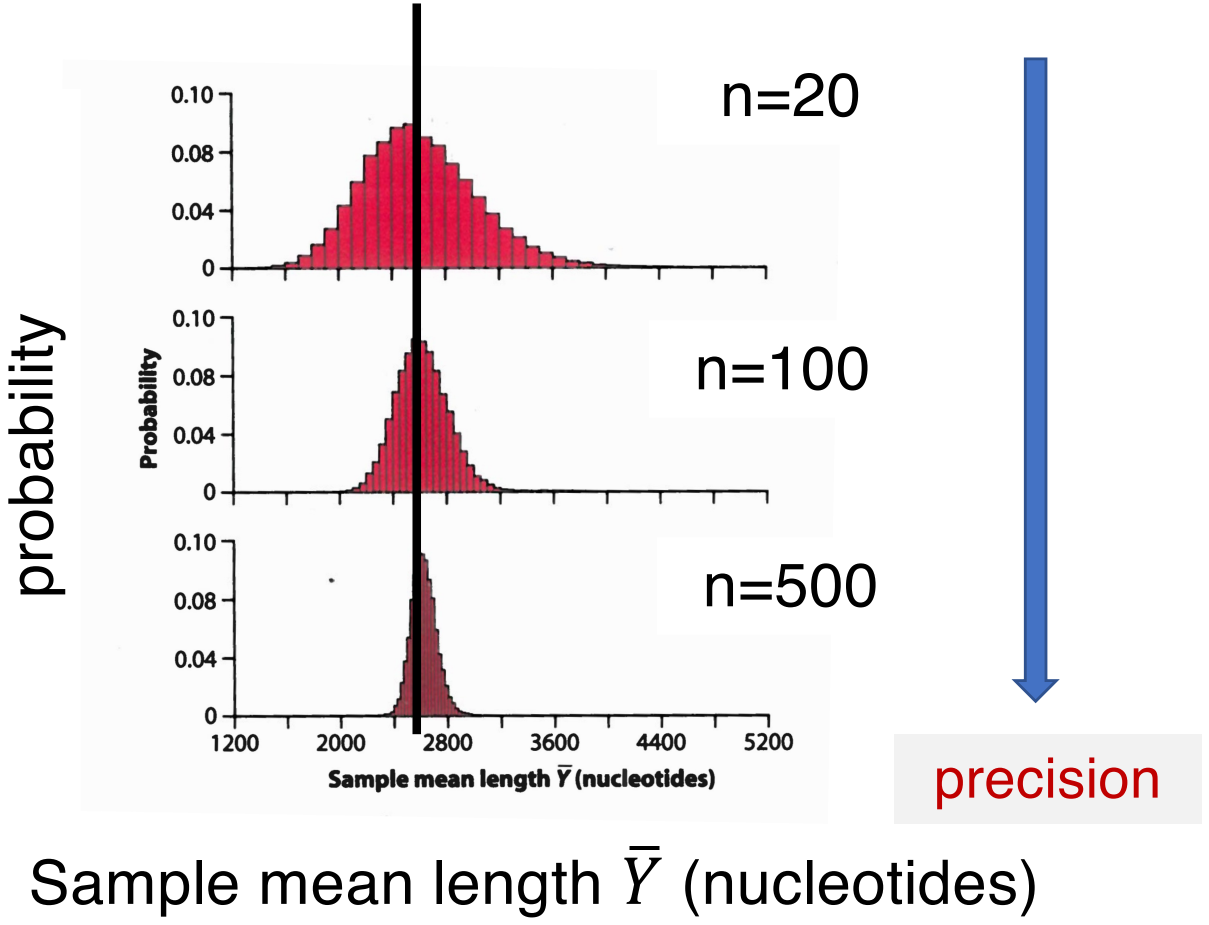
The statistical road: estimate with uncertainty but measure your confidence.

Sample size increases precision

Frequency distribution of the gene Population



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



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

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)

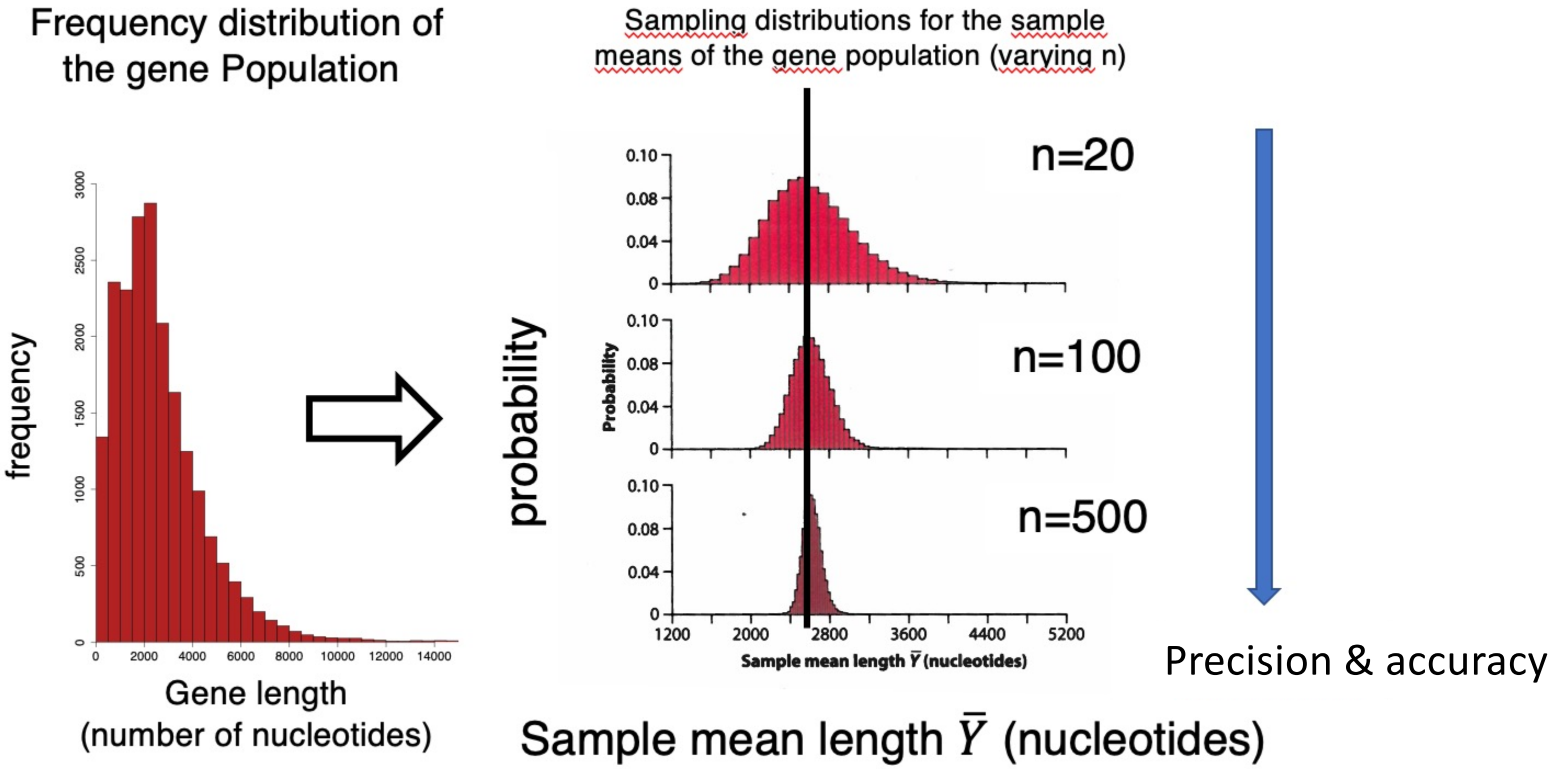
The common requirement of the methods presented in this course (and in statistics in general) is that data come from a **random sample**. A random sample is one that fulfills two criteria:

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.

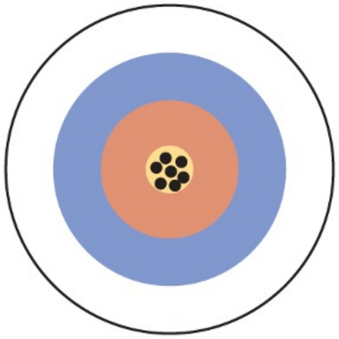
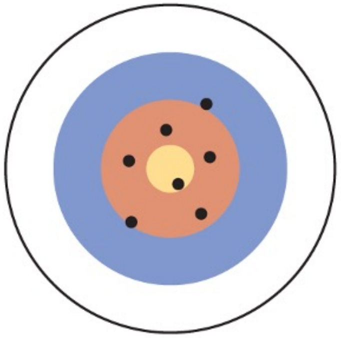
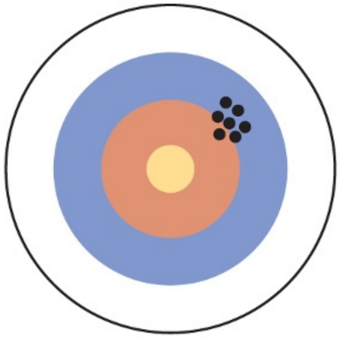
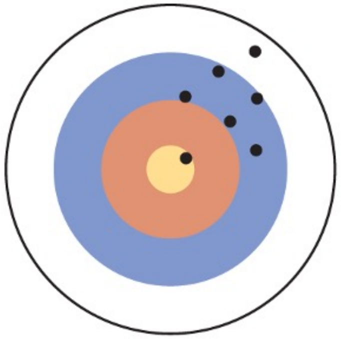
Samples are biased when some observational units of the intended population have lower or higher probabilities to be sampled.

While increasing sample size reduces random error (which improves precision), it doesn't necessarily improve accuracy if there is systematic bias in the study design, sampling method, or data collection process. So, under unbiased sampling, then increasing sample sizes improves precision and accuracy



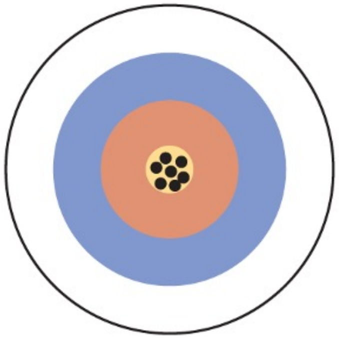
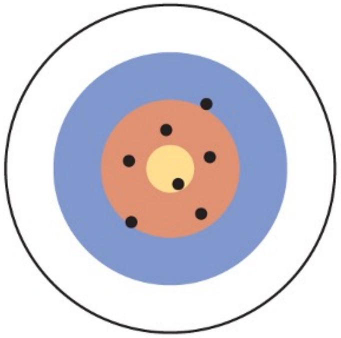
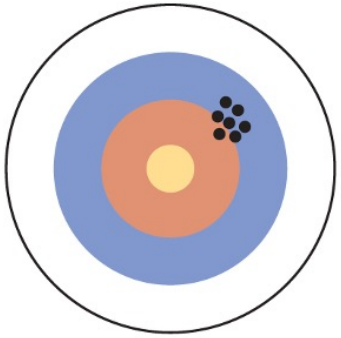
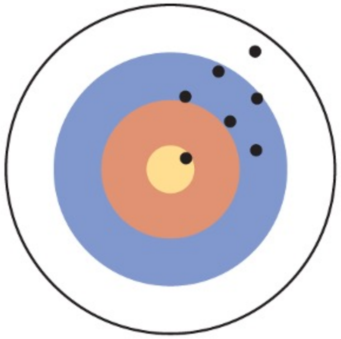
Whitlock & Schluter, 2nd edition; 3rd edition has a different set of genes.

Regarding the estimation of population means, what does random sampling assure? Accuracy!

	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

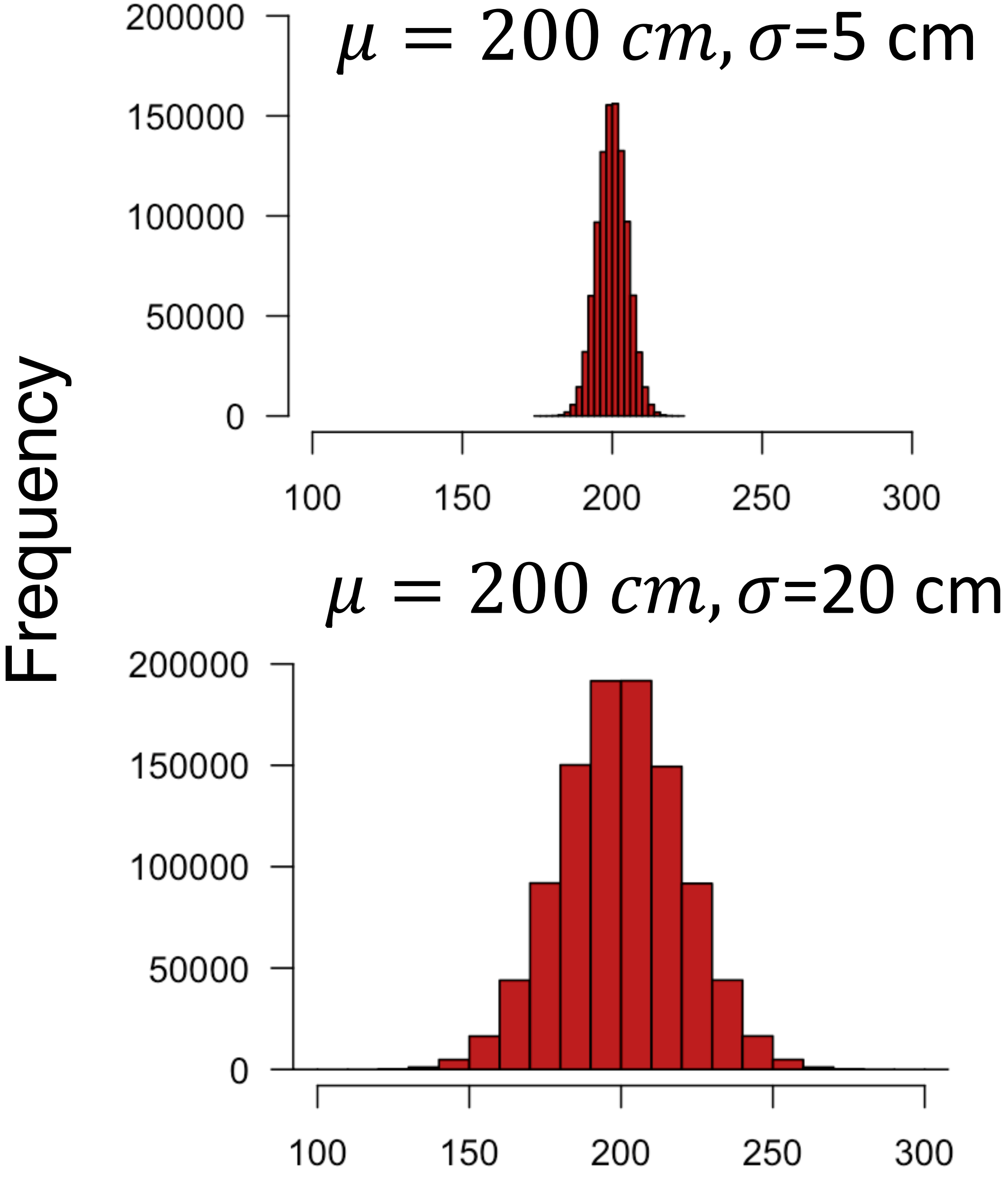
A single sample mean is said to be unbiased under random sampling because the mean of all sample means equal the population mean.

Regarding the estimation of population means, what does random sampling assure as sample size increases? **Precision!**

	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

As sample size increases, there is less variation of sample means around the true population mean.

What else affects precision assuming that accuracy is correct? The standard deviation (or variance) of the statistical population

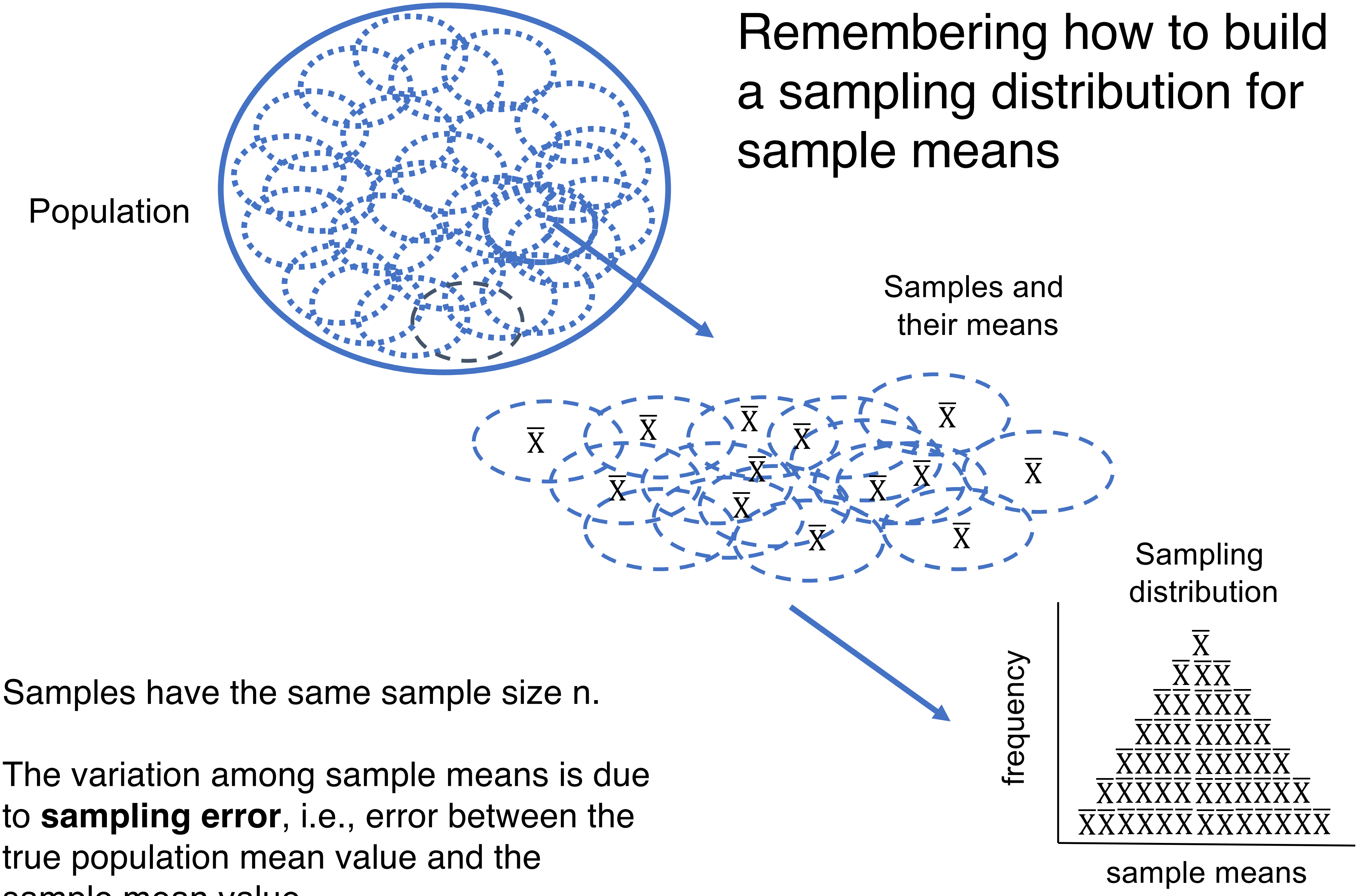


$\mu = 200\ cm, \sigma$=5 cm

$\mu = 200\ cm, \sigma$=20 cm

Two statistical populations with the same population mean μ = 200 but differing in their population standard deviations σ.

Remembering how to build a sampling distribution for sample means



Population

Samples and their means

Sampling distribution

frequency

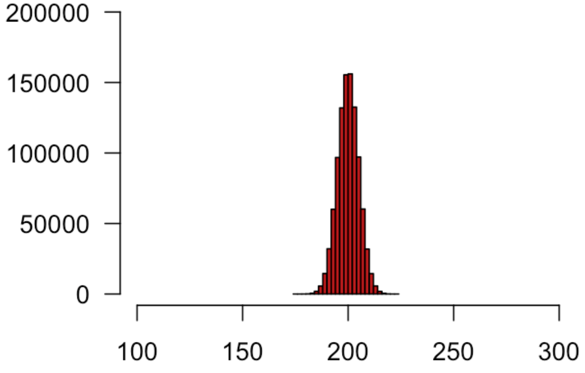
sample means

Samples have the same sample size n.

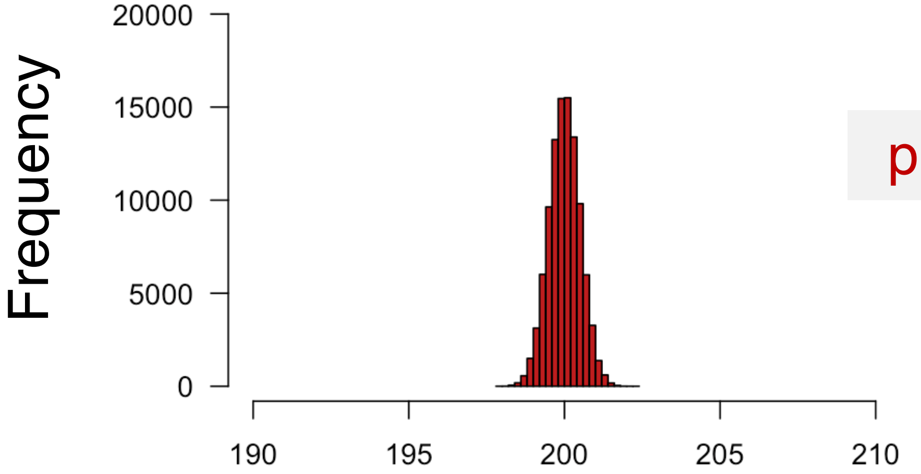
The variation among sample means is due to **sampling error**, i.e., error between the true population mean value and the sample mean value.

What else affects precision assuming that accuracy is correct? The standard deviation (or variance) of the statistical population!

$\mu = 200\ cm, \sigma$=5 cm

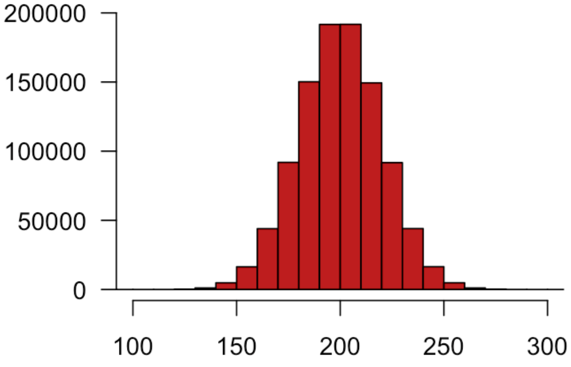


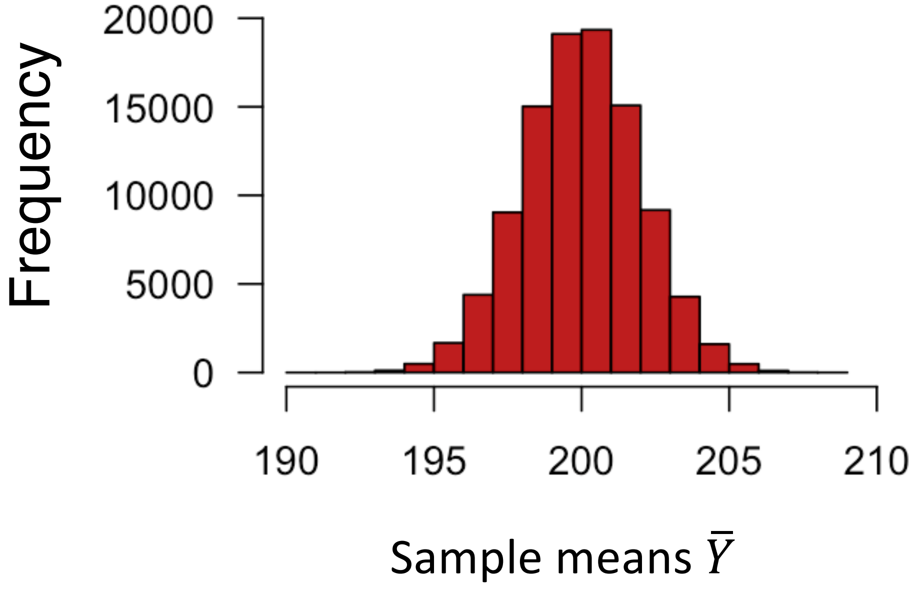
Sampling distributions of sample means.
n=100 for each sample based on 100,000 samples.



precision

$\mu = 200\ cm, \sigma$=20 cm





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

Example: Voting polls in the news claims about **accuracy & precision** (under unbiased sampling):

"43% of the voting intention goes to the XXX party. The sample size was 1020; for a sample of this size the maximum margin of error is about 3%."

Do you know what that means? (assuming that the sample is random, we're pretty confidence that the true value in the voting population is between 43 ±3%, i.e., somewhere between 40% and 46%.")

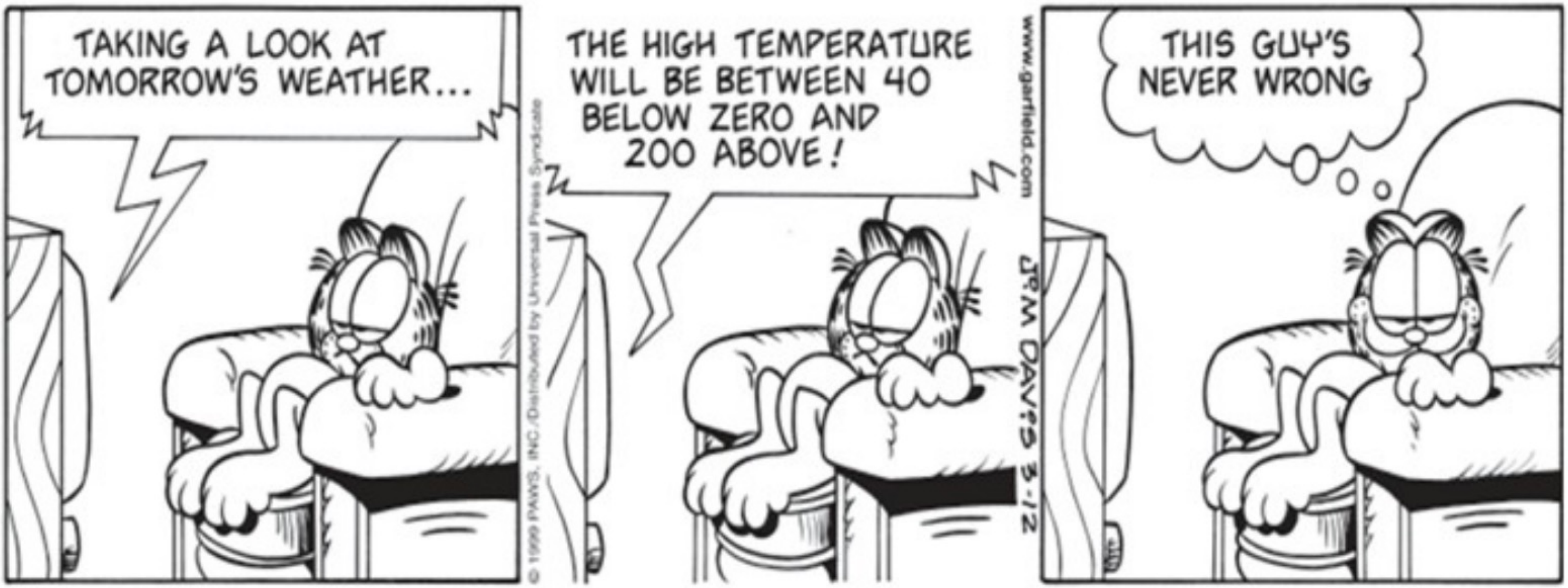
- Source - http://www.scc.ms.unimelb.edu.au/whatisstatistics/ssize.html

How to trust an estimate?
(i.e., a value based on a single sample)

Estimating with uncertainty, but with a degree of certainty (i.e., with some confidence), part 2

We are confident (assuming unbiased sampling) that the true proportion of the voting population supporting party XXX is between 40% and 46%, with a point estimate of 43% ± 3%.

How can we trust (or be confident in) a sample estimate?


TAKING A LOOK AT TOMORROW'S WEATHER...
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
www.garfield.com
JIM DAVIS 3-12
THIS GUY'S NEVER WRONG

source - https://slideplayer.com/slide/7575691/

Estimating with uncertainty, but with a degree of certainty

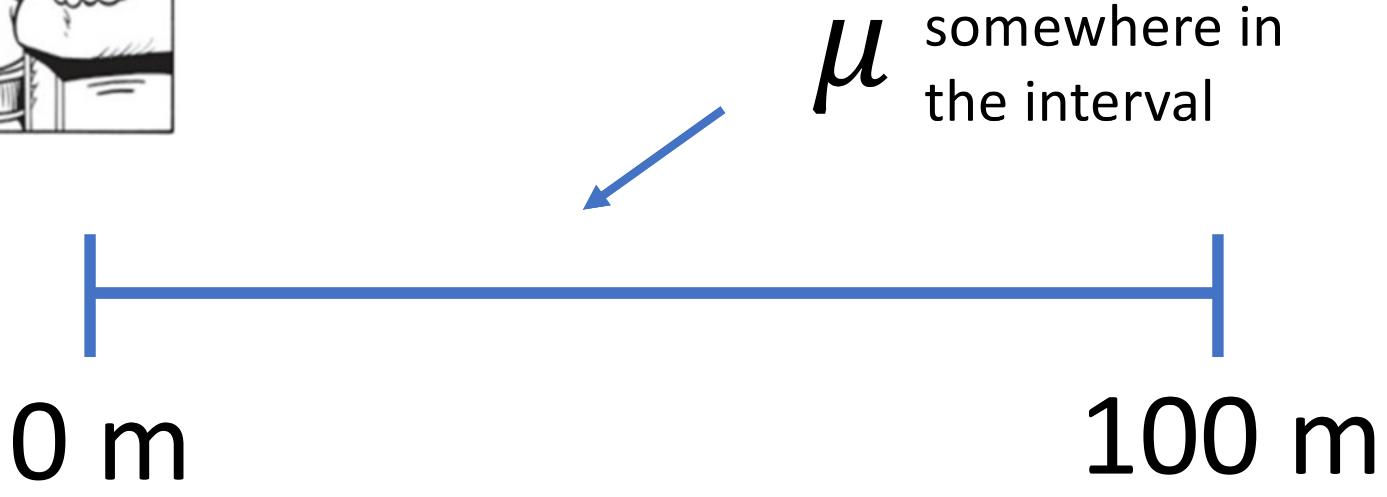
- Most conclusions are drawn from samples, meaning we always have incomplete knowledge about the population of interest.

- Now, imagine a method that allows us to say, "We are confident" that the true parameter of interest (e.g., the mean height of humans or trees) lies within a specific range of values.

- Example 1: The average height of all humans (i.e., the entire population) is between 0 m and 100 m. While this statement is technically true, it's useless because it provides no meaningful precision—obviously, all humans are taller than 0 m and shorter than 100 m, but this range doesn't help us estimate the true average accurately.

Estimating with uncertainty, but with a degree of certainty

Example 1: The average height of all humans (i.e., the entire population) is between 0 m and 100 m. While this statement is technically true, it's useless because it provides no meaningful precision—obviously, all humans are taller than 0 m and shorter than 100 m, but this range doesn’t help us estimate the true average precision (it’s accurate though but “useless”).

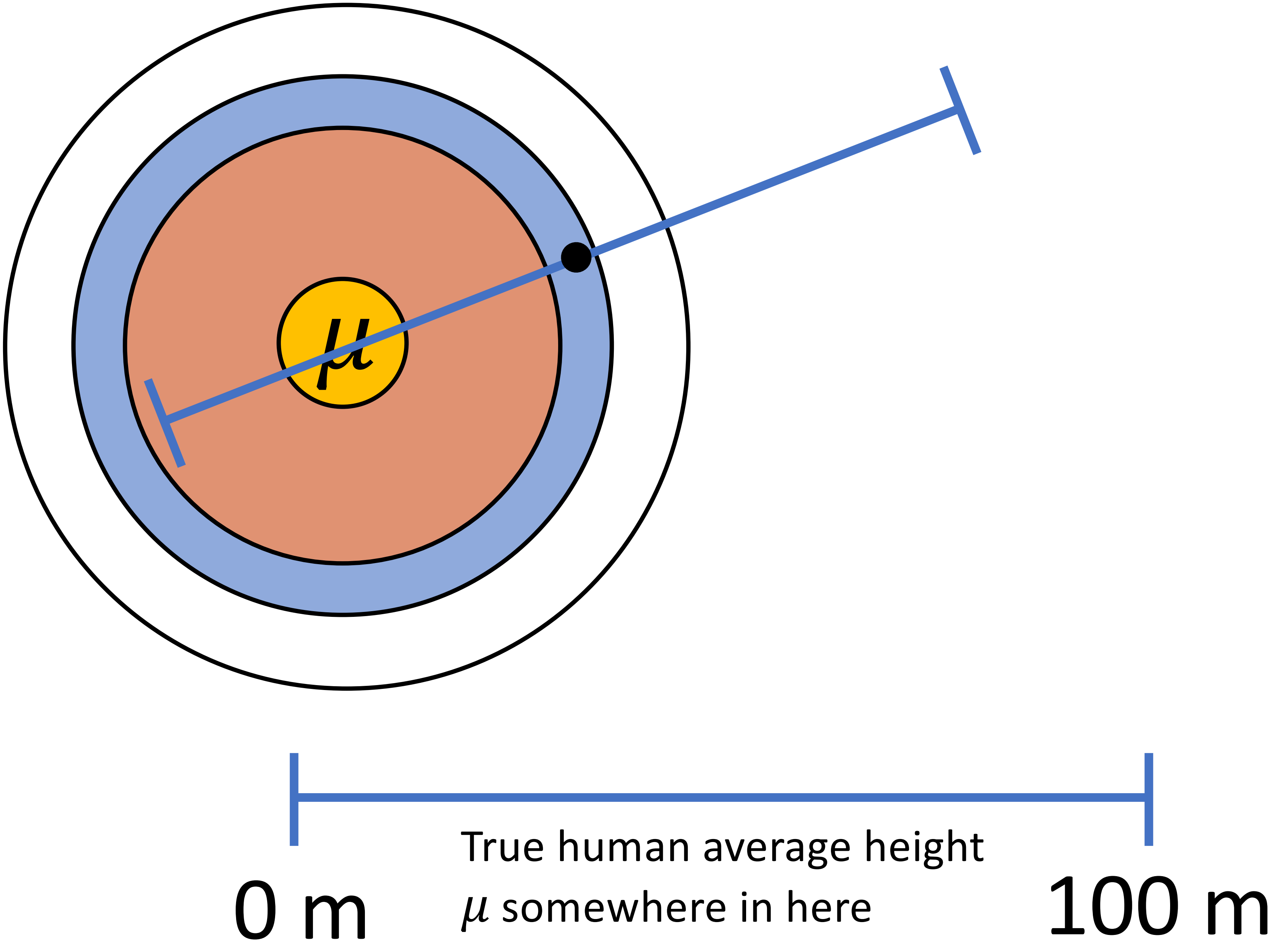


μ somewhere in the interval

0 m

100 m

Estimating with uncertainty, but with a degree of certainty

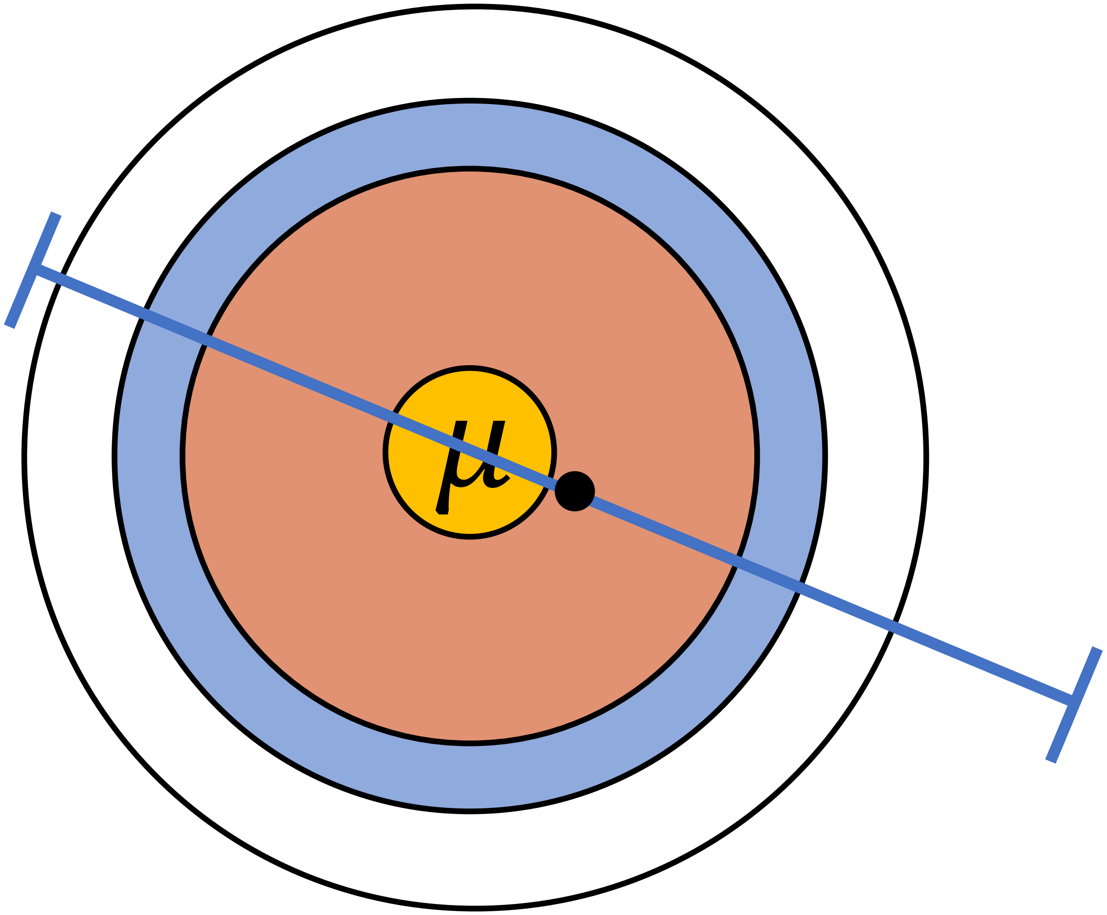


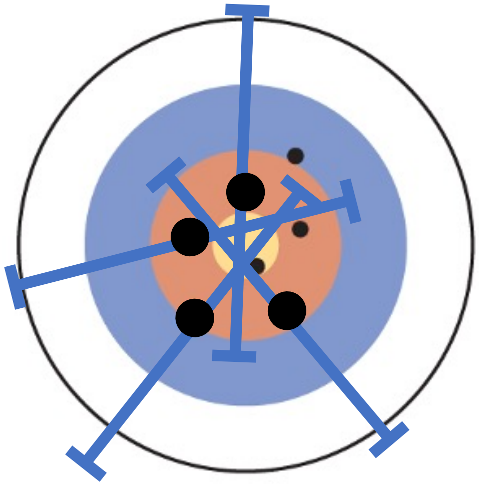
0 m

True human average height
μ somewhere in here

100 m

Estimating with uncertainty, but with a degree of certainty





For any given sample mean, we can construct an interval that likely contains the true population parameter, with a certain level of confidence.

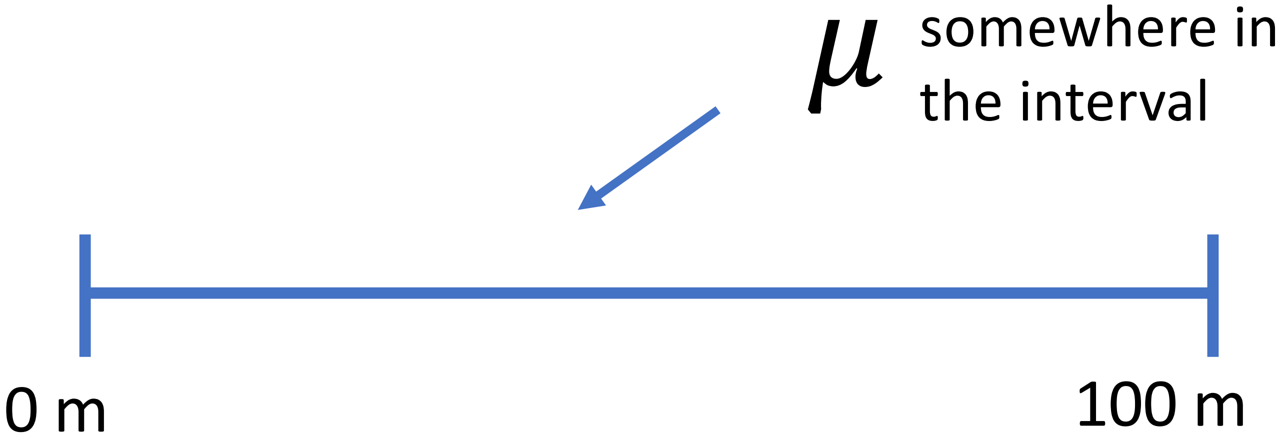
True human average height
μ somewhere in here

0 m

100 m

Let’s take a break – 1 minute!

Estimating with uncertainty, but with a degree of certainty

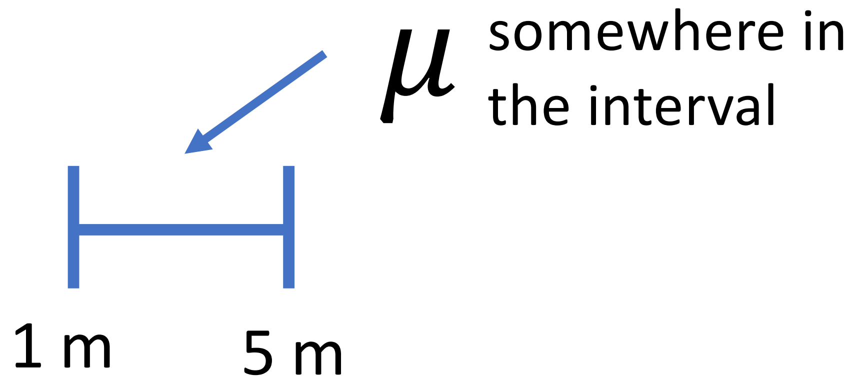


μ somewhere in the interval

0 m

100 m

- **Example 2:** The average height of all adult humans is between 1 m and 5 m. While this interval is more useful than the first, it’s still not very helpful because we know that most adults are taller than 1 m and shorter than 5 m. Therefore, the true average will fall within this range, but it doesn't provide much precision.



μ somewhere in the interval

1 m

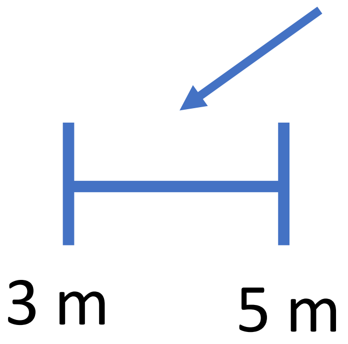
5 m

Estimating with uncertainty, but with a degree of certainty

- **Example 2:** The average height of all adult humans is between 1 m and 5 m. While this interval is more useful than the first, it’s still not very helpful because we know that most adults are taller than 1 m and shorter than 5 m. Therefore, the true average will fall within this range, but it doesn't provide much precision.

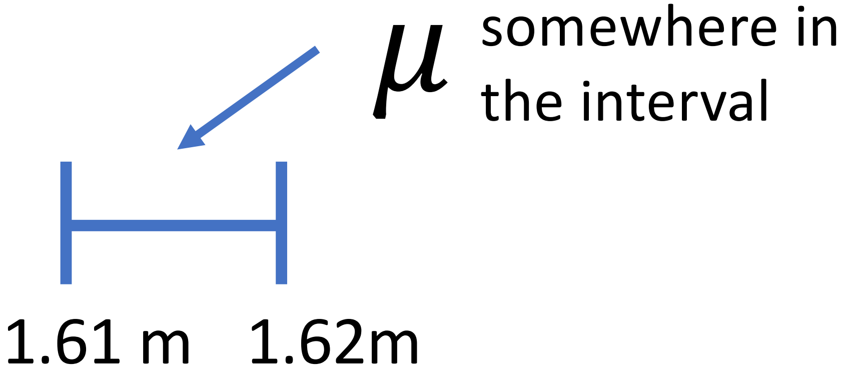
- **Example 3:** The average height of all adult humans is between 3 m and 5 m. This interval is incorrect because it’s impossible for the average height of all adults to be greater than 3 m and less than 5 m. Therefore, the true average cannot be within this range.

μ somewhere in the interval



Estimating with uncertainty, but with a degree of certainty

Example 4: The average height of all adult humans is between 1.61 m and 1.62 m. While this interval might be accurate, it is likely too narrow, and therefore could be very misleading.



μ somewhere in the interval

1.61 m 1.62m

Estimating with uncertainty, but with a degree of certainty

- **Example 5:** The average height of all adult humans is between 1.51 m and 1.80 m. This interval is the most reliable.

The method for constructing confidence intervals for the population mean relies on the sampling distribution of means, the sample mean, and the sample standard deviation.

So, although we don’t know the true value (parameter) with certainty, we can estimate an interval that gives us a certain degree of confidence about where that true value lies.

By the way, we can also construct confidence intervals for other statistics such as the standard deviation, variance, median, and more.

Estimating with uncertainty, but with a degree of certainty

Making the claims we just did, i.e., building confidence (intervals) for the true population statistic of interest (e.g., mean) requires that we trust our sample estimates & increase precision when possible.

Estimating with uncertainty, but with a degree of certainty

To estimate confidence intervals for the true population statistic (e.g., mean) requires trusting our sample estimates and increasing precision by boosting sample size when possible.

We can trust our samples by using random sampling (to ensure accuracy) and increasing the sample size (to improve precision).

Estimating with uncertainty, but with a degree of certainty

To estimate confidence intervals for the true population statistic (e.g., mean) requires trusting our sample estimates and increasing precision by boosting sample size when possible.

We can trust our samples by using random sampling (to ensure accuracy) and increasing the sample size (to improve precision).

Statistical populations with smaller variances increase precision, but this is a luxury researchers can’t always control. However, it may be achievable by defining more specific problems—for example, comparing the average height of all humans versus the average height of adult humans.

Estimating with uncertainty, but with a degree of certainty

The discussion on confidence intervals for human height was intended to give students some intuition about what "confidence" means in statistical terms.

That said, in real-world scenarios, these kinds of contrasts can be challenging because we often don’t have a clear idea of the potential range of values.

As such, many confidence intervals are likely built without a clear understanding of the true range of values, which can lead to imprecision or misinterpretation in the results.

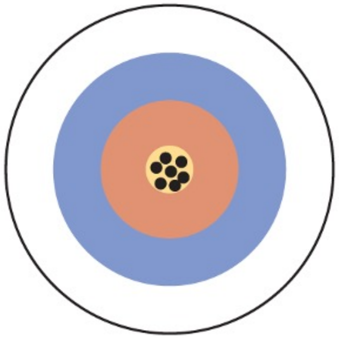
Biologists need to invest more effort in determining which confidence intervals are plausible for a given problem.

An important goal is to increase precision while ensuring accuracy in sampling (assuming the estimator is accurate)

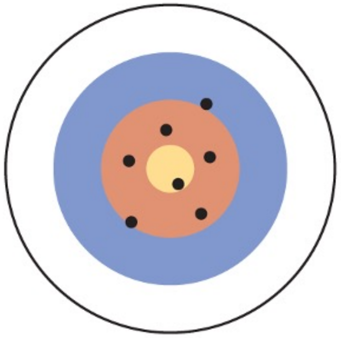
Precise

Imprecise

Accurate



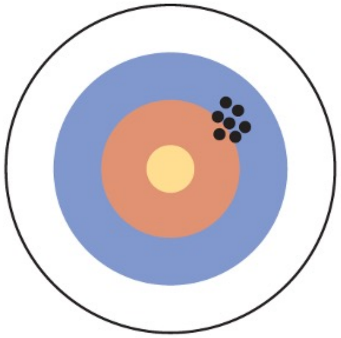
Accurate & precise



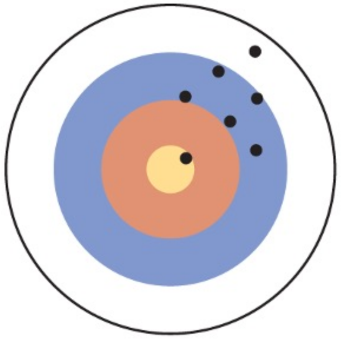
Accurate but not precise

As the sample size increases, the variation in sample means around the true population mean decreases.

Inaccurate



Not accurate but precise

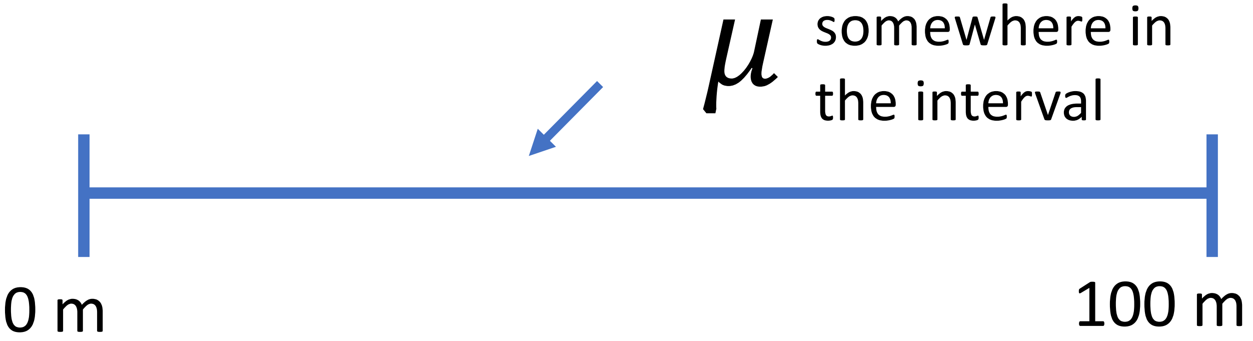


Not accurate & not precise

The sample mean is an accurate estimator of the population mean because the average of all possible sample means equals the population mean (parameter); if sampling is unbiased.

Estimating with uncertainty, but with a degree of certainty

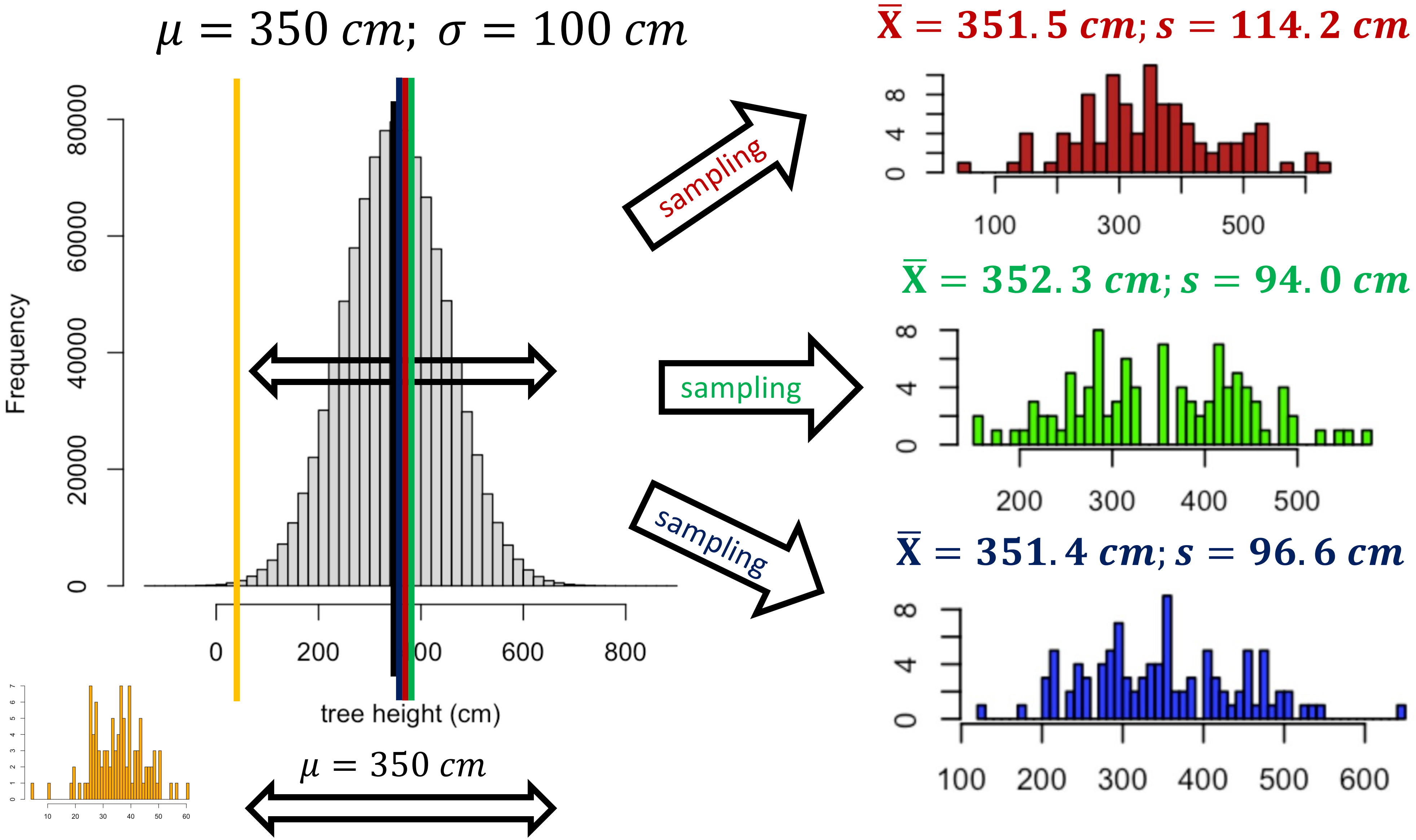
How to build certainty?



This interval is built on the basis of a single sample!

But our confidence on it is based on random sampling.

Sampling variation generates uncertainty, i.e., sampling error



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

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

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

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

sampling

sampling

sampling

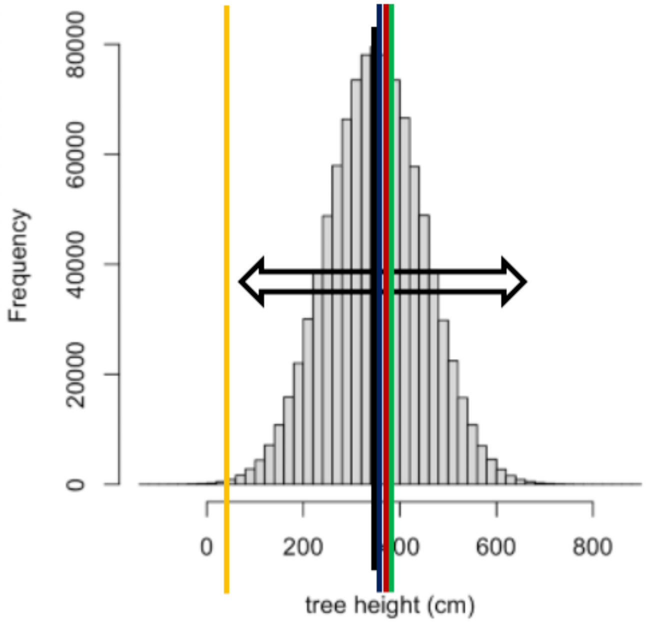
$\mu = 350\ cm$

Uncertainty (samples means varying around the true population mean)

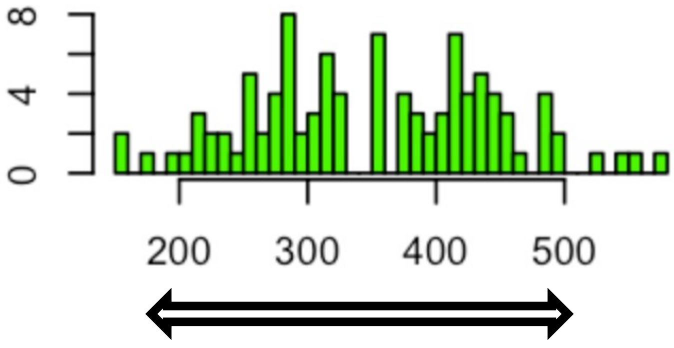
The variation within a sample (measured by the standard deviation) gives us insight into how much sample means (averages) might differ from the true population mean (average)—essentially estimating how far off we might be

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

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



sampling



Variation within samples

$\mu = 350\ cm$

Variation among samples

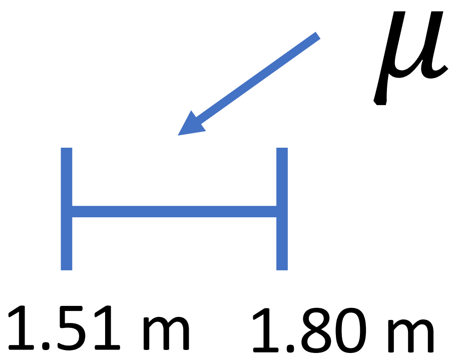
Variation within samples (among observations) can estimate some certainty (confidence) about uncertainty (variation among sample means)

The statistical road: estimate with uncertainty but measure your confidence.

Imagine an interval referred as to “95% confidence interval”:

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

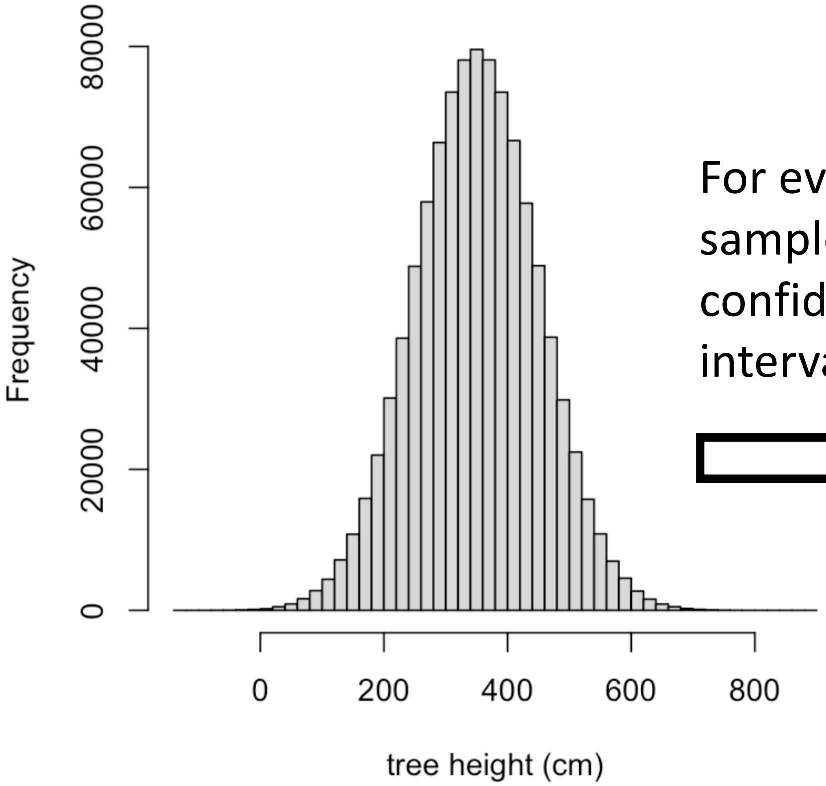
A larger confidence level (e.g., 95% or 99%) provides a more plausible range for the parameter. Values within the interval are considered more plausible, while those outside are less plausible, based solely on the sample data.



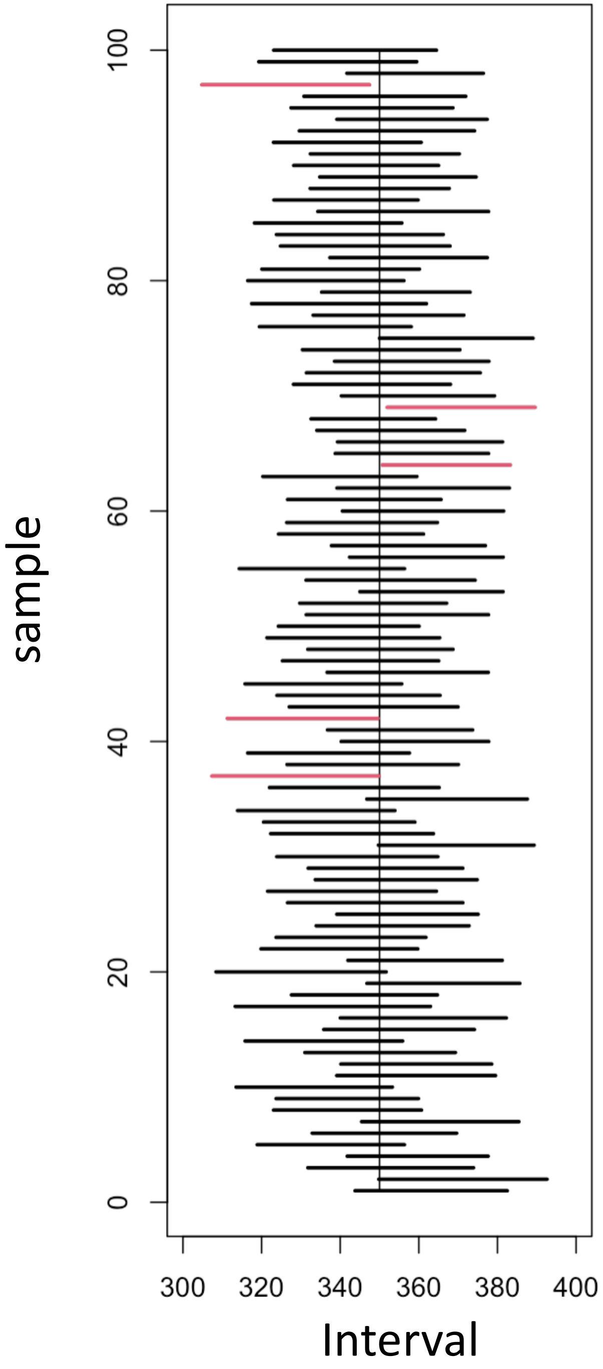
Very plausible (high confidence) that the population parameter is somewhere within The 95% confidence interval.

Imagine an interval referred as to “95% confidence interval”:

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



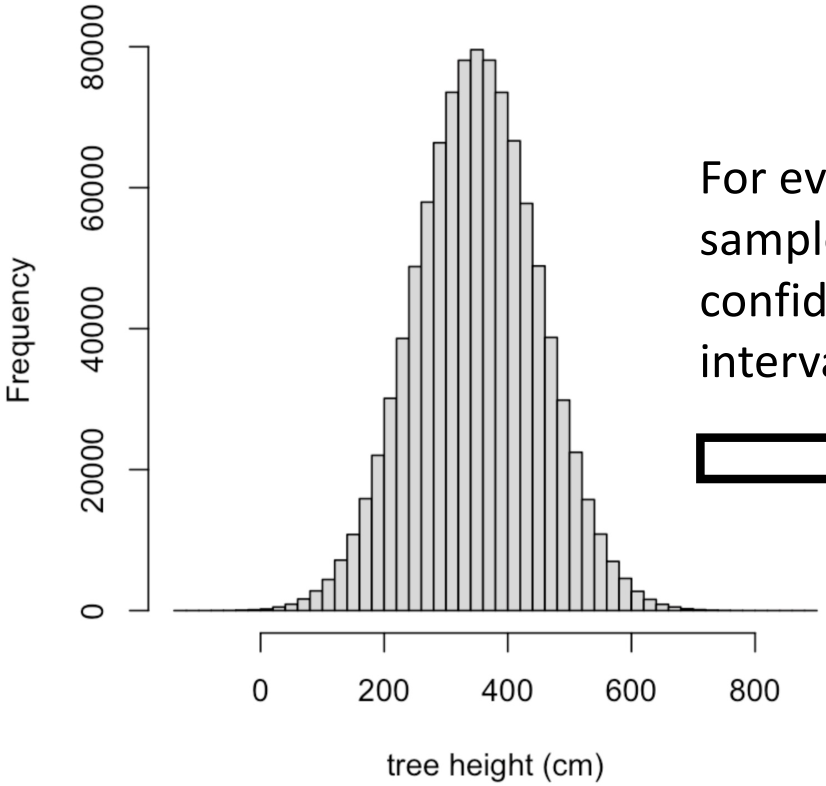
For every possible sample build a confidence interval



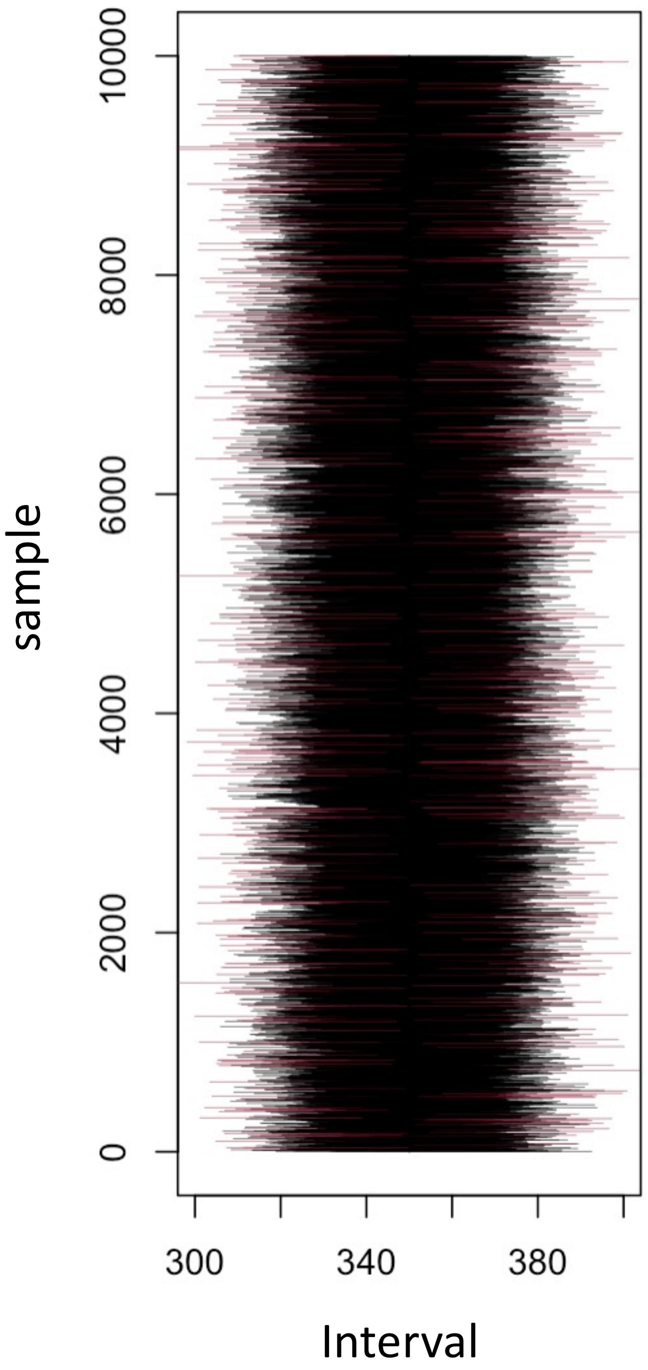
If sampling is random (unbiased) and the population distribution has certain properties (e.g., approximately normal), 95 out of 100 (95%) confidence intervals will contain the true population parameter. *The intervals that do not contain the true parameter are shown in red (5%).*

Imagine an interval referred as to “95% confidence interval”:

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



For every possible sample build a confidence interval

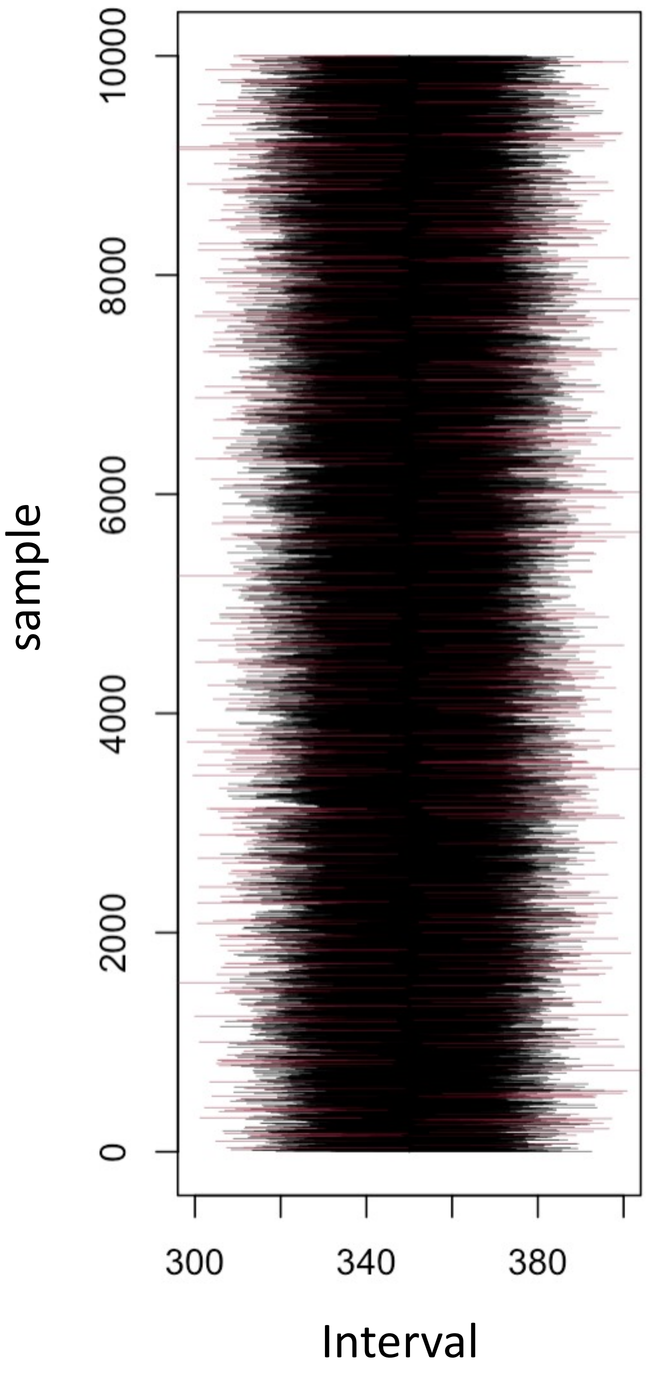


If sampling is random (unbiased) and the population distribution has certain properties (e.g., approximately normal), 9500 out of 10000 (95%) confidence intervals will contain the true population parameter. *The intervals that do not contain the true parameter are shown in red (5%).*

Very important!

For any given confidence interval, *we can say*, **'We are 95% confident that the true population mean lies between the lower and upper limits of the interval.**'

However, *we cannot say*, **'There is a 95% probability that the true population mean lies within the confidence interval.'** The true parameter either lies within the interval or it doesn't—there's no probability attached to this specific condition.

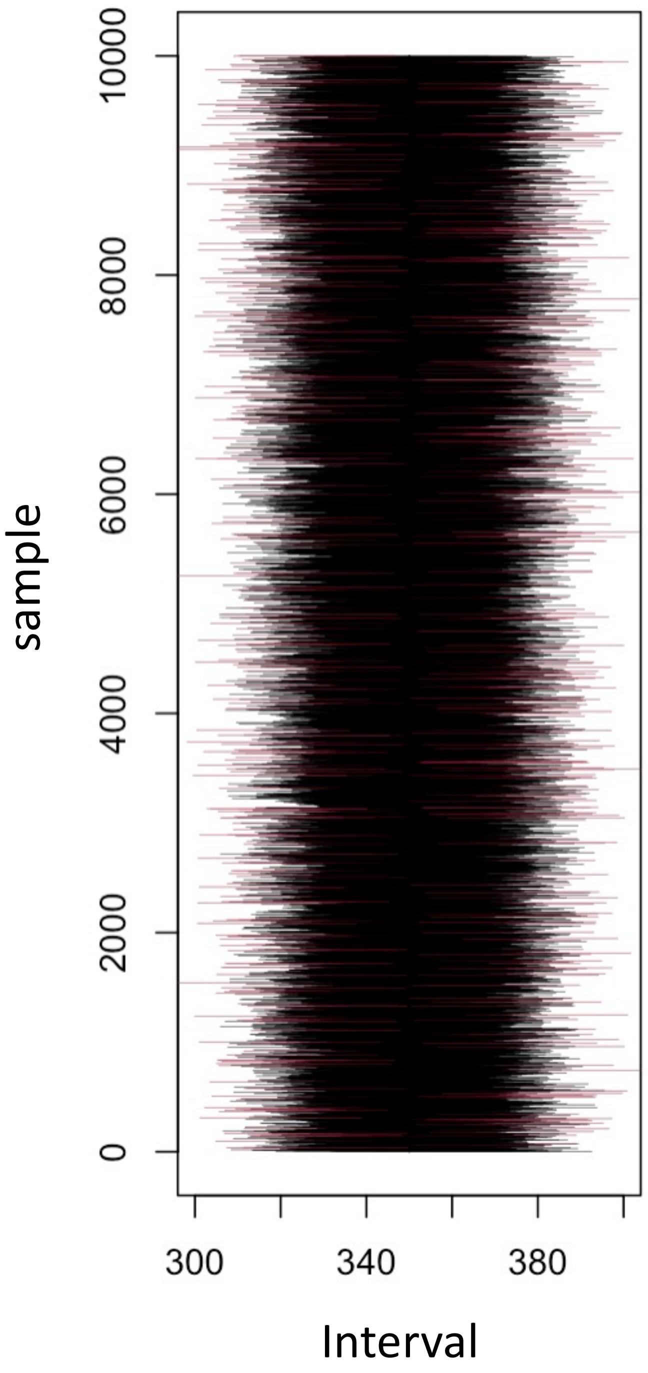


Very important!

Confidence intervals are based on the principle of '**repeated sampling**' and not on the idea of assigning a probability to whether the interval contains the true population value (parameter).

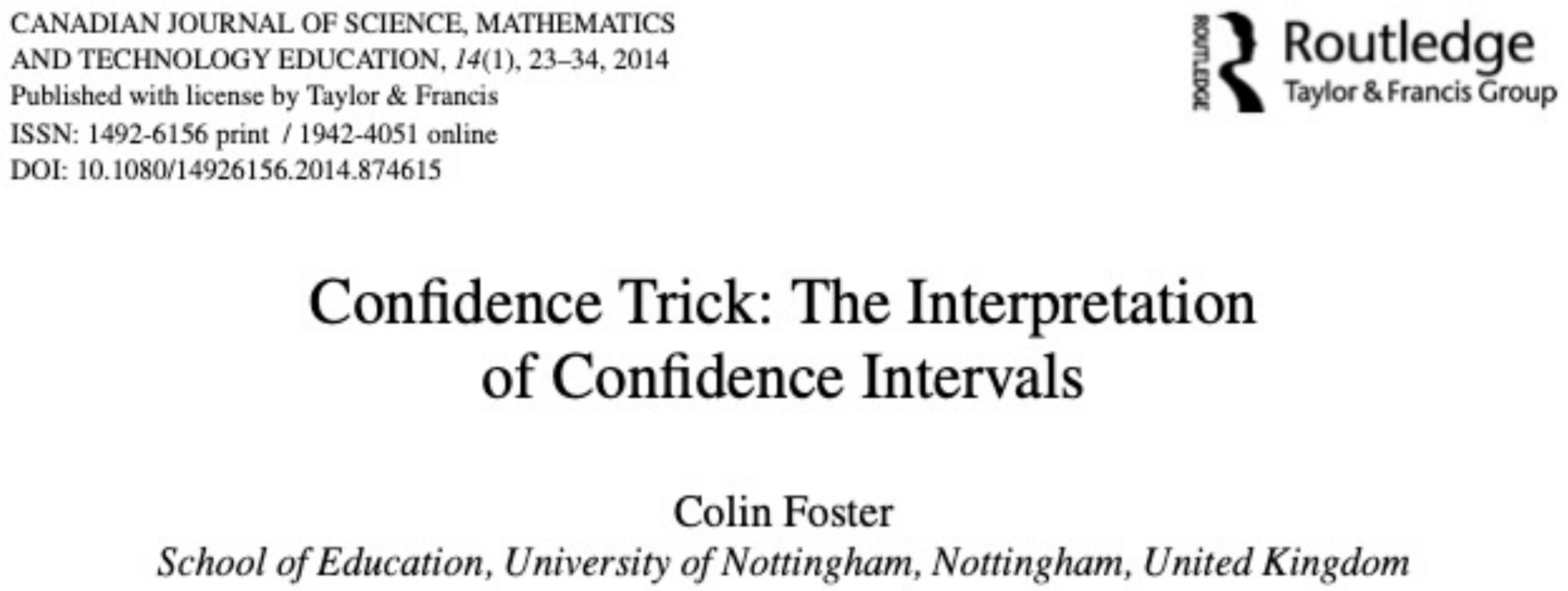
This means that if we repeatedly take samples and construct confidence intervals from them, a certain percentage of those intervals (e.g., 95%) will contain the true population parameter. The interval itself varies from sample to sample, but the parameter is fixed.

Thus, confidence intervals describe the likelihood that, through repeated sampling, the interval will capture the true, fixed parameter. They do not represent the probability that any single interval contains the true parameter.



Let’s take a break – 1 minute!

Many users of statistics struggle to fully understand confidence intervals because of the concept of “repeated sampling”



CANADIAN JOURNAL OF SCIENCE, MATHEMATICS
AND TECHNOLOGY EDUCATION, *14*(1), 23–34, 2014
Published with license by Taylor & Francis
ISSN: 1492-6156 print / 1942-4051 online
DOI: 10.1080/14926156.2014.874615

ROUTLEDGE
Routledge
Taylor & Francis Group

Confidence Trick: The Interpretation of Confidence Intervals

Colin Foster
School of Education, University of Nottingham, Nottingham, United Kingdom

TRUE: confidence intervals describe the likelihood that, through repeated sampling, the interval will capture the true, fixed parameter.

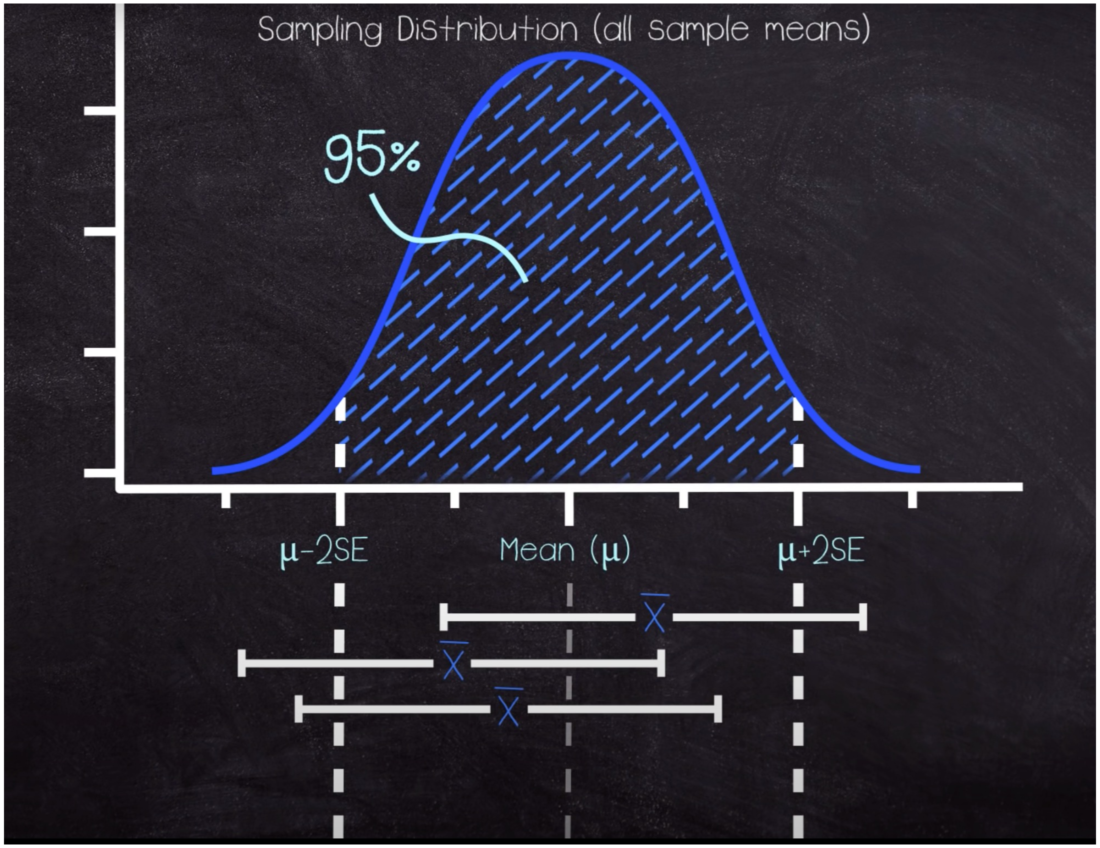
But Many think (wrongly) that confidence intervals represent the probability that the interval contains the true parameter.

Assumptions underlying confidence intervals:

If sampling is random and the population's distribution (marginal distribution) is approximately normal, then exactly 95% of all possible confidence intervals will contain the true population parameter (i.e., “a sampling distribution of confidence intervals”).

If the population distribution is not normal, a number close to 95% (e.g., 92%, 97%, etc) of the intervals will contain the true parameter. This number, as we’ll discuss later, depends on the population's distributional properties (e.g., asymmetry) - see tutorial 5.

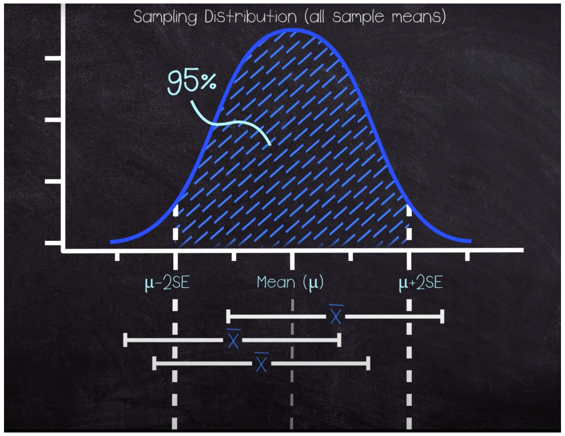
We use the sampling distribution of all sample means to calculate confidence intervals



The interval based on **population mean $\pm$ 2 $\times$ SE** contains 95% of all possible sample means.

Because the distribution is symmetric, then 95% of the intervals based on **sample mean $\pm$ 2 $\times$ SE** will contain the population mean.

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



The interval based on **population mean ± 2 × SE** contains 95% of all possible sample means.

Because the distribution is symmetric, then 95% of the intervals based on **sample mean ± 2 × SE** will contain the population mean.

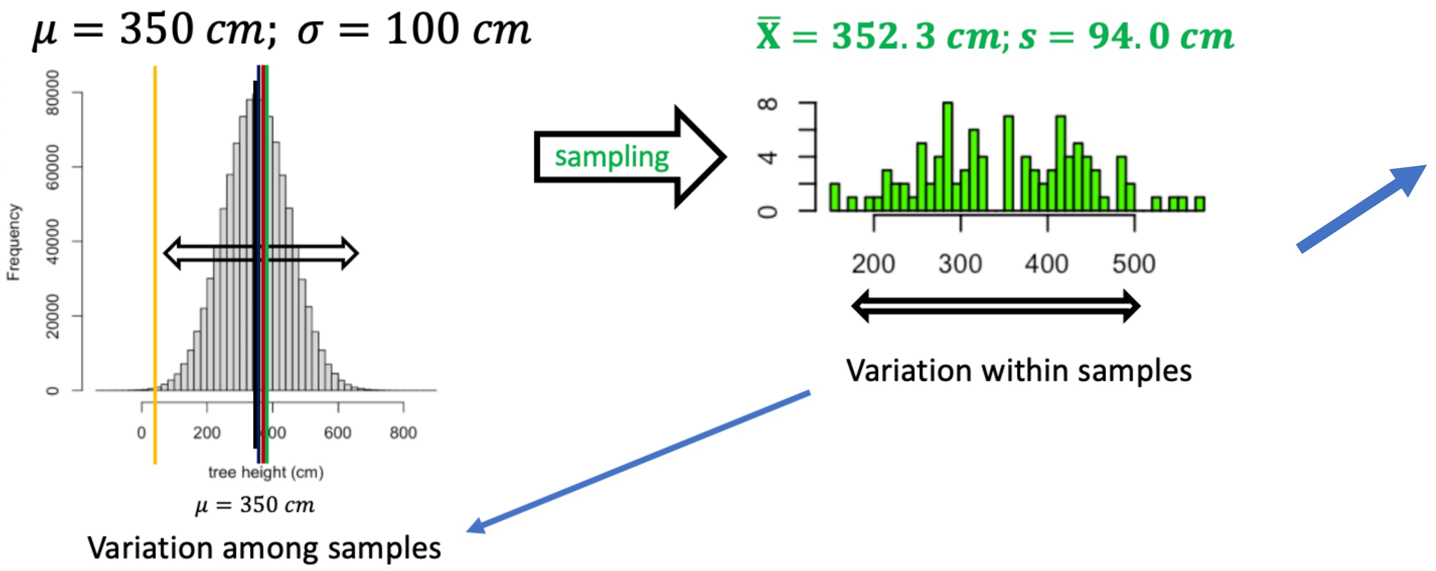
What is $\mathrm{SE}_{\bar{Y}}$? It is called the standard error, which is the standard deviation of all sample means; it is the average difference between each sample mean and the true population mean. Importantly, we don't need to know the true population parameter (mean) to estimate the sampling distribution (more on this in the next lecture).

If the standard error of the mean ($\mathrm{SE}_{\bar{Y}}$) is small, then we expect a greater probability of our sample be closer to the true population mean (parameter).

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

If the standard error of the mean ($\mathrm{SE}_{\bar{Y}}$) is small, then we expect a greater probability of our sample be closer to the true population mean (parameter).

$\mathrm{SE}_{\bar{Y}}$ can be estimated from the sample standard deviation *s* as follows:



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

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

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

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:

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

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

$$\sigma = the\ standard\ deviation\ of\ the\ population$$

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

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

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

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

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

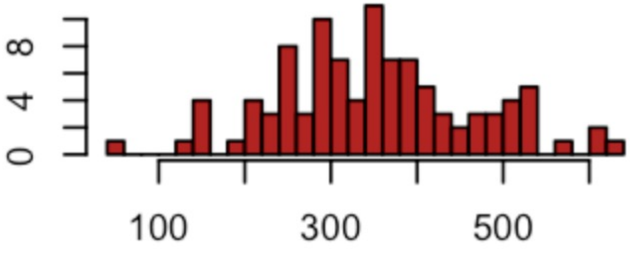
$\sigma =$ the standard deviation of the population

$\mathrm{SE}_{\bar{Y}}$ estimates the average value in which the sample means differ from the true population mean. And this estimate is produced from the sample alone (in tutorial 5 you will learn this principle in detail).

How to calculate a "95% confidence interval" in practice:

$$\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{114.2}{\sqrt{100}} = 11.42$$

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



$\mathrm{SE}_{\bar{Y}}$ estimates the average value in which the sample means differ from the true population mean.

Based on this sample, in average, samples are estimated to differ from the true population value by 11.42 *cm.*

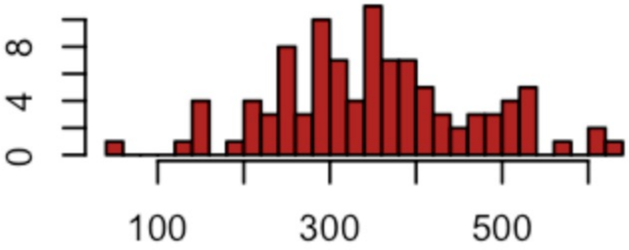
Obviously, a different sample may give a different estimate of this error.

How to calculate a “95% confidence interval” in practice:

If sampling is random, if the frequency distribution of the population is roughly normal, and if sample size is relatively large, then this interval can be calculated based on the sample standard error $\mathrm{SE}_{\bar{Y}}$ (why? Next lecture).

$$\bar{Y} \pm 2\mathrm{SE}_{\bar{Y}} \therefore \mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

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



$$351.5 \pm \overbrace{2\times\frac{114.2}{\sqrt{100}}}^{\text{Margin of error}}$$

328.66 cm 374.34 cm

How to calculate a “95% confidence interval” in practice:

The margin of error is introduced here being calculated based on **2** to facilitate understanding what confidence intervals are (“pedagogical approach”).

If [a] sampling is random, [b] the frequency distribution of the population is normal or roughly normal & [c] **sample size is relatively large** (30 or more observational units), then **2** as the multiplier is a good approximation (the exact value will be smaller than **2** though). We will see these details in our next lecture.

When sample sizes are less than 30 observations, then the multiplier of the $\mathrm{SE}_{\bar{Y}}$ will be bigger than **2**; and when the sample size is huge (“infinite”), the multiplier is exactly 1.96 instead of **2**. Basically, the multiplier changes as a function of sample size by tend to be around **2** when n > 30.

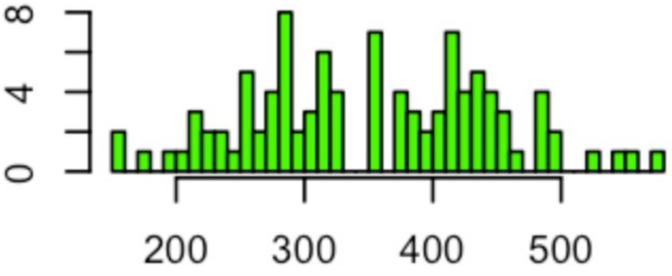
$$351.5 \pm \mathbf{2} \times \frac{114.2}{\sqrt{100}}$$

How to calculate a “95% confidence interval” in practice:

If sampling is random, if the frequency distribution of the population is roughly normal, and if sample size is relatively large, then this interval can be calculated as:

$$\bar{\mathrm{Y}} \pm \ 2\mathrm{SE}_{\bar{Y}} \therefore \mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

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



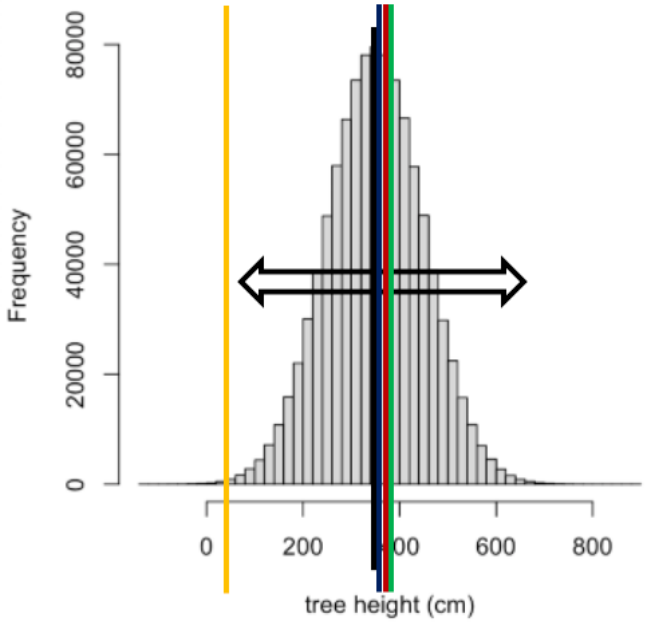
$$352.3 \pm 2 \times \frac{114.2}{\sqrt{100}}$$

333.5 cm

371.1 cm

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

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

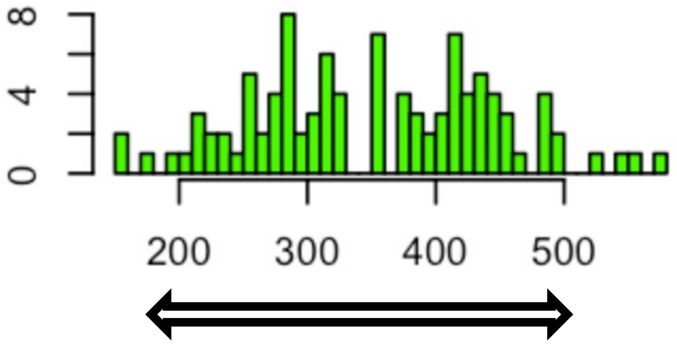


$\mu = 350\ cm$

Variation among samples

sampling

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



Variation within samples

$$352.3 \pm 2\times\frac{114.2}{\sqrt{100}}$$

333.5 cm 371.1 cm

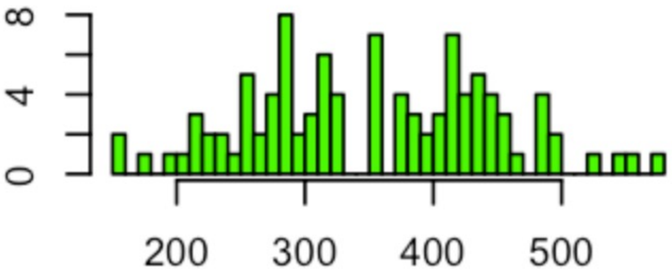
Estimating with uncertainty, but with a degree of certainty

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.**

$$\bar{\mathrm{Y}} \pm \ 2\mathrm{SE}_{\bar{Y}} \therefore \mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

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



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

333.5 cm
371.1 cm

NOTE: Confidence intervals are calculated and not estimated!

Calculated: Confidence intervals are derived using a mathematical formula based on sample data, the sampling distribution, and assumptions about the population (e.g., normality). Since a confidence interval involves precise computation, "calculated" is the more appropriate term.

Estimated: While the confidence interval gives us a range to estimate where the true population parameter lies, the interval itself is not estimated but rather **calculated** based on the sample.

How do you know if the interval is useful?! How wide is too wide?

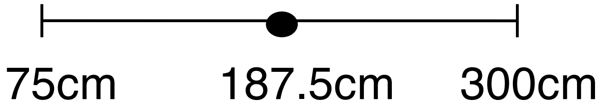
In general, the 95% confidence interval is a good measure of our uncertainty about the true value of the parameter (population value).

If the confidence interval is broad, then uncertainty is high and the data are not very informative about the value of the population parameter (i.e., location in the sampling distribution; more on that in the next lecture).

Is the interval useful? This is not a statistical question per se. The answer is often based on the problem at hands and/or your expertise able to defend that uncertainty given by the interval. Does this interval allow you to say something that is important with scientific confidence?

I can guarantee that it's between one and one million!
That's not exactly useful...

e.g., 100% sure:
Average adult height
of people living in
Montreal


75cm
187.5cm
300cm