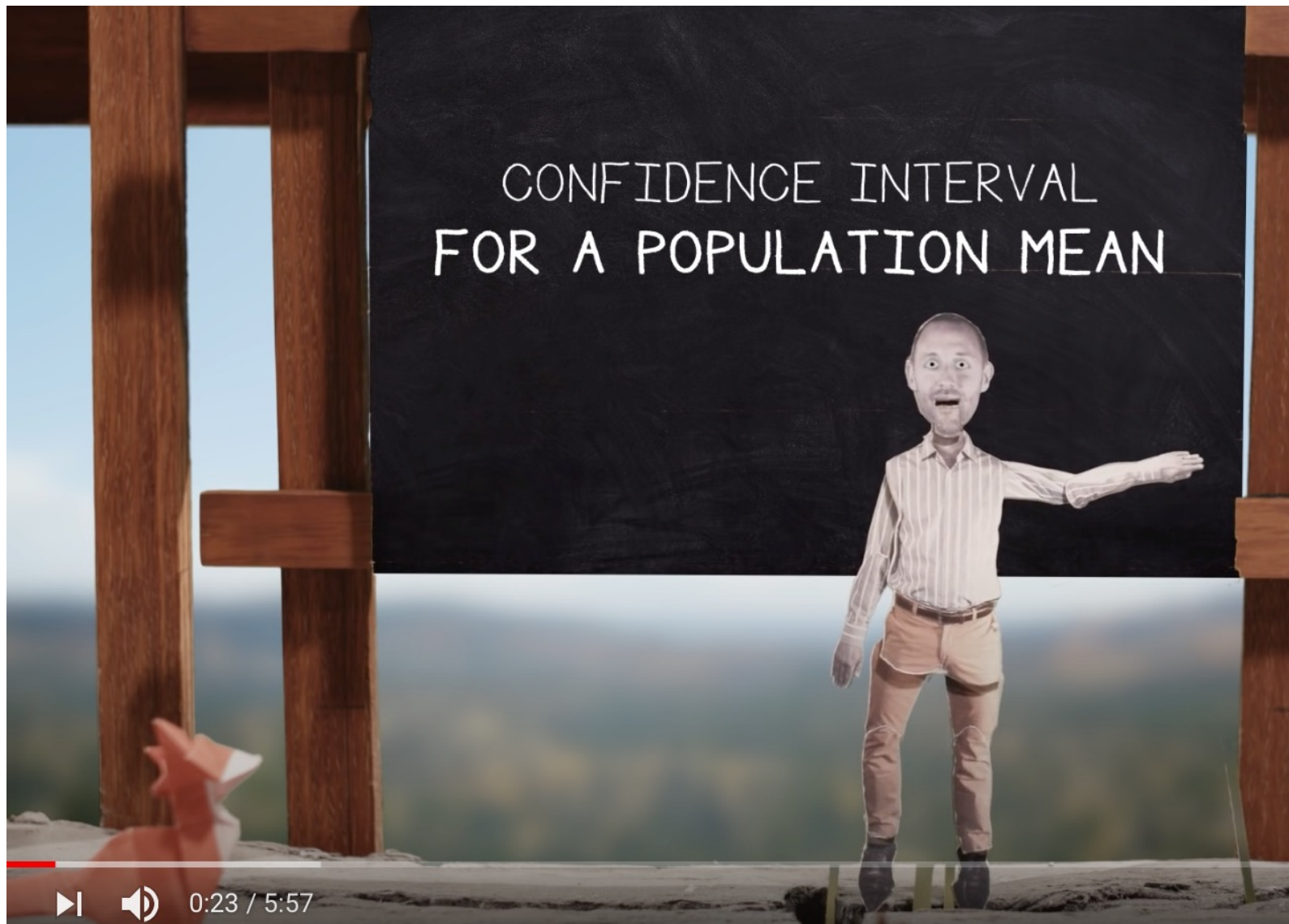


# A snap demonstration why data entry and file formats are critical – .csv would never allow that to happen!

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.



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 with certainty (i.e., with some confidence), part 2

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



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:

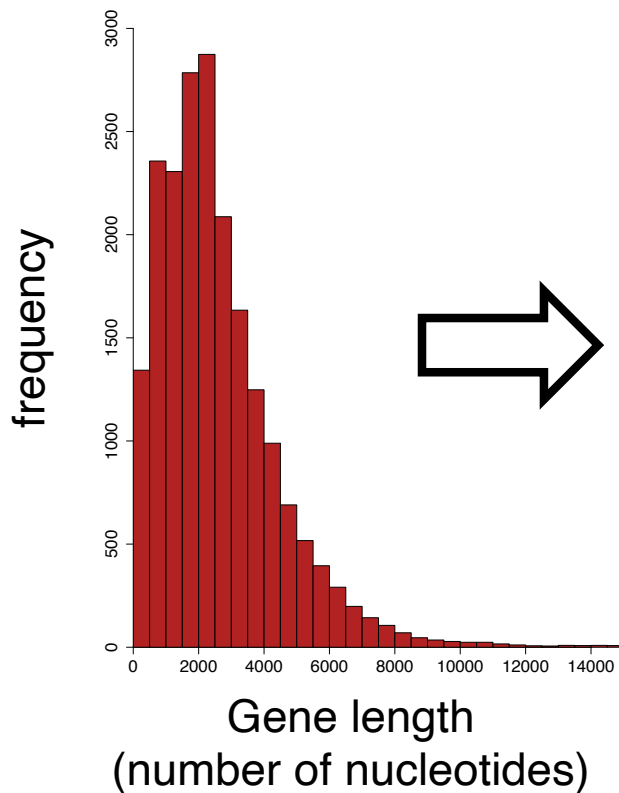
1) Every observational unit in the population (e.g., individual tree) have an **equal chance** of being included in the sample.

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.

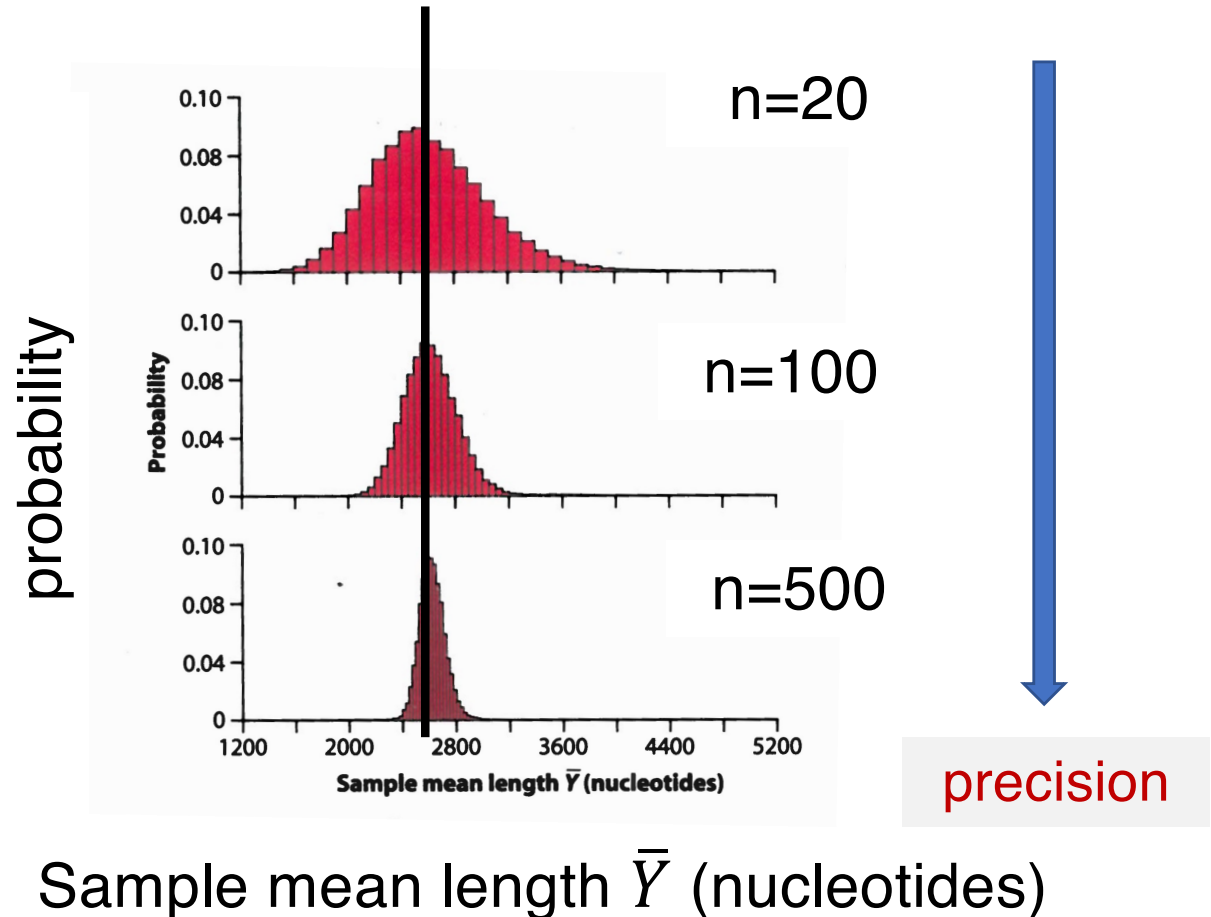
Samples are biased when some observational units of the intended population have lower or higher probabilities to be sampled.

# Sample size increases precision

Frequency distribution of the gene Population

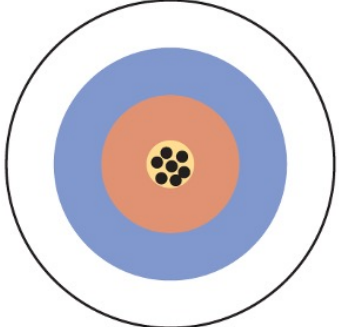
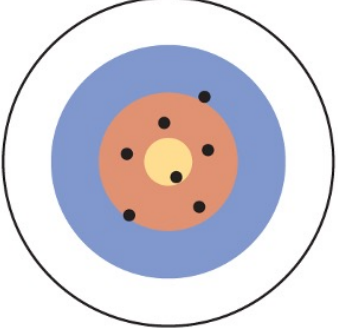
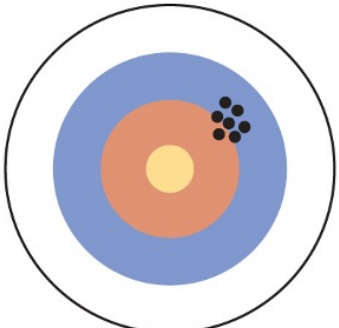
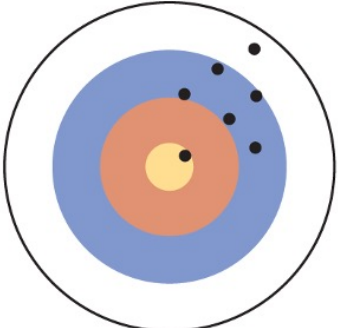


Sampling distributions for the sample means of the gene population (varying n)



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

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

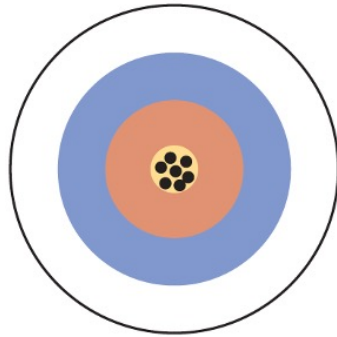
	Precise	Imprecise	
Accurate	 <p>Low sampling variation (sampling error) &amp; low bias</p>	 <p>High sampling variation (sampling error) &amp; low bias</p>	A single sample mean is said to be unbiased under random sampling because the mean of all sample means equal the population mean.
Inaccurate	 <p>Low sampling variation (sampling error) &amp; high bias</p>	 <p>High sampling variation (sampling error) &amp; high bias</p>	

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

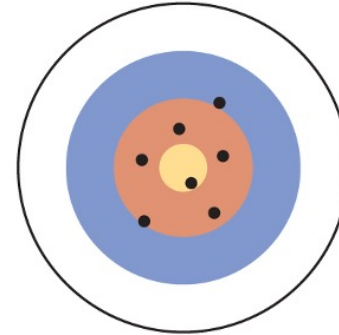
Precise

Imprecise

Accurate



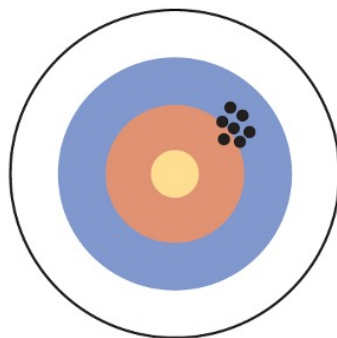
Low sampling variation  
(sampling error) & low bias



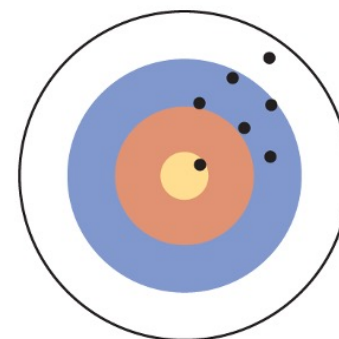
High sampling variation  
(sampling error) & low bias

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

Inaccurate

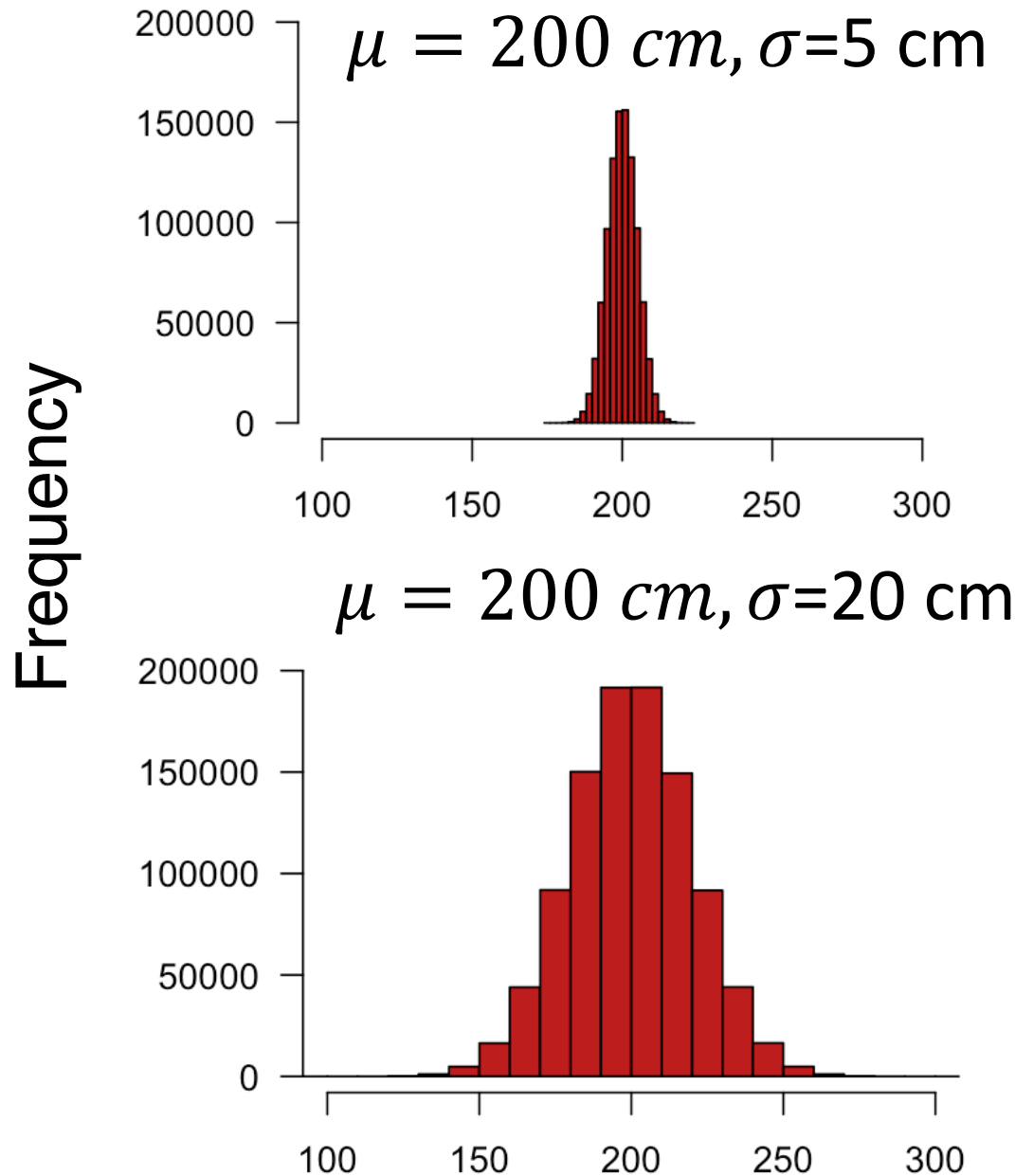


Low sampling variation  
(sampling error) & high bias



High sampling variation  
(sampling error) & high bias

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

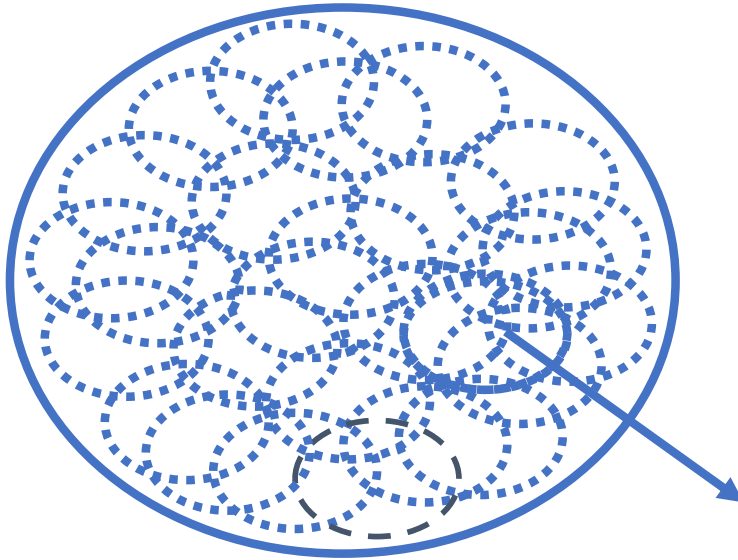


Two statistical populations with the same population mean  $\mu = 200$  but differing in their population standard deviations  $\sigma$ .

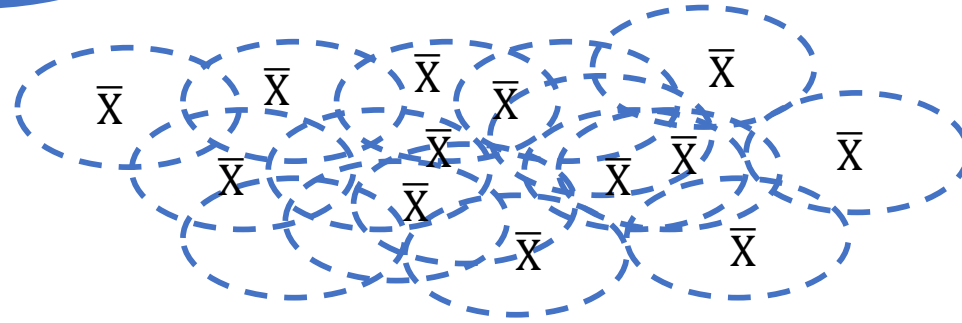


# Remembering how to build a sampling distribution for sample means

Population



Samples and their means

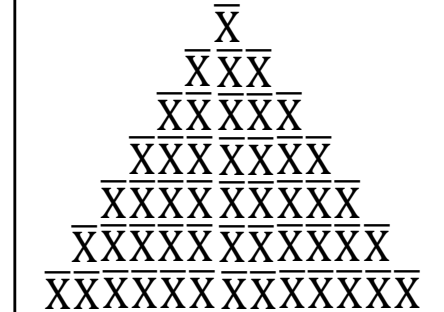


Samples have the same sample size  $n$ .

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 distribution

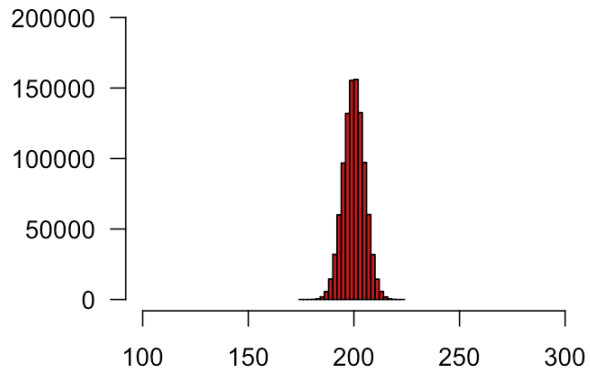
frequency



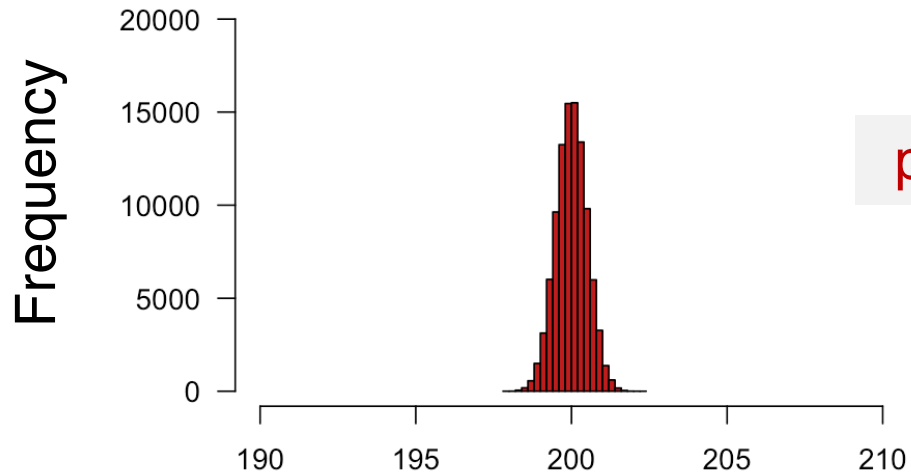
sample means

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

$$\mu = 200 \text{ cm}, \sigma = 5 \text{ cm}$$



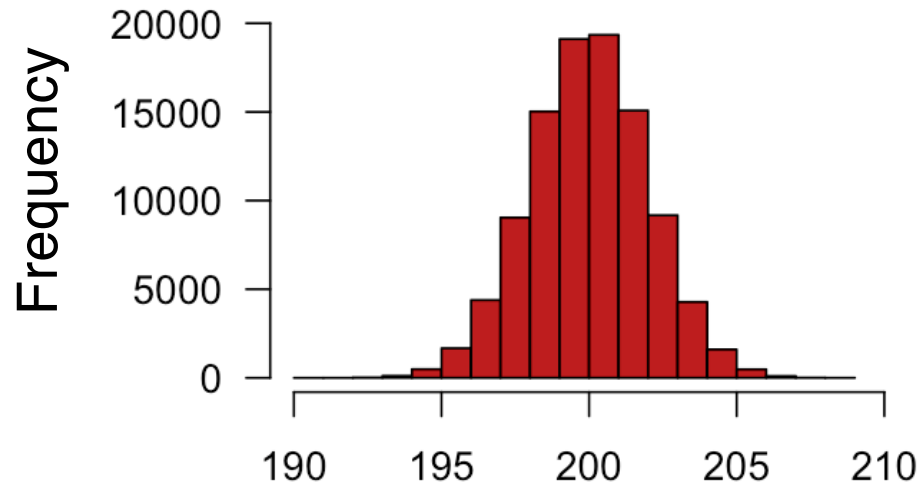
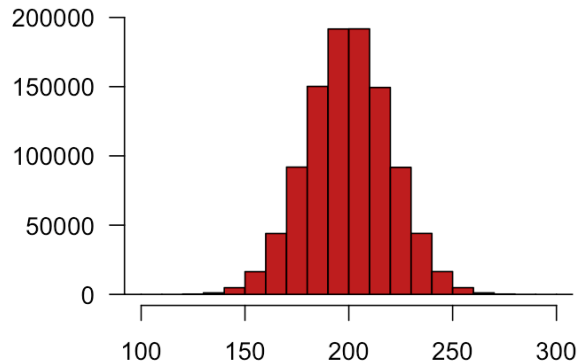
Sampling distributions of sample means.  
 $n=100$  for each sample. 100,000 samples.



precision



$$\mu = 200 \text{ cm}, \sigma = 20 \text{ cm}$$



Sample means  $\bar{Y}$

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

Example: Voting polls in the news which make a claim about **precision**; example:

"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%."

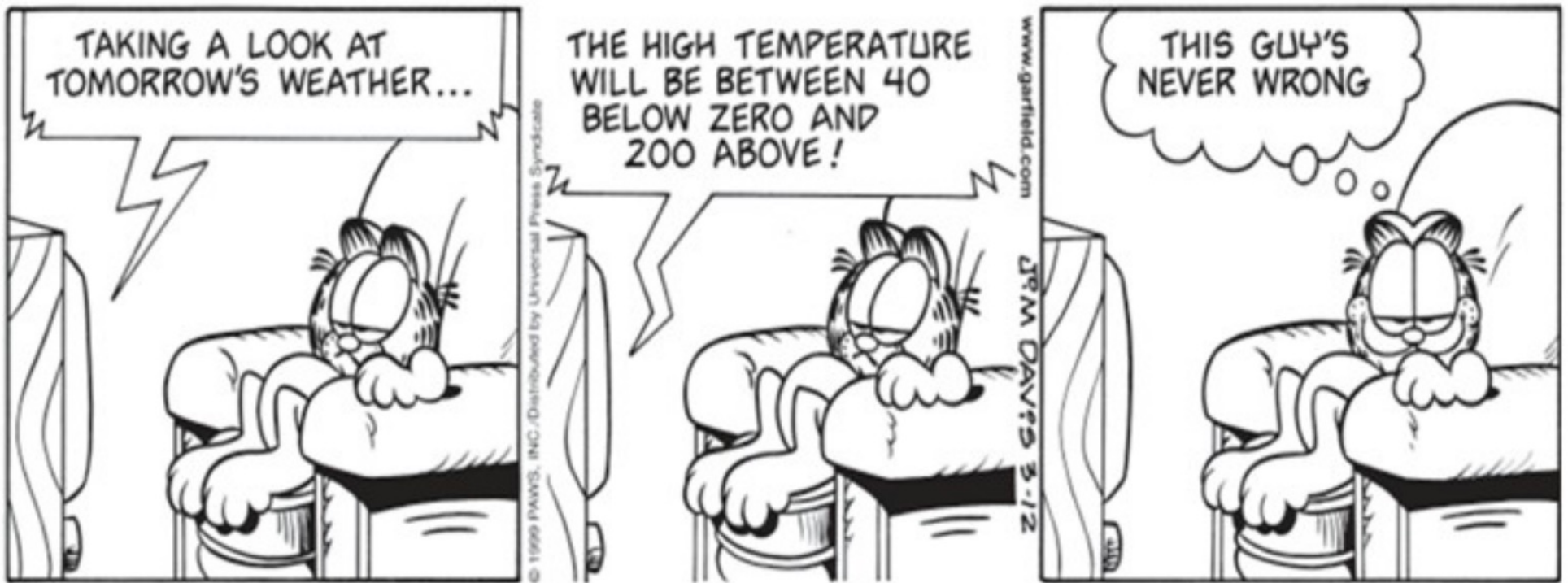
*Do you know what that means?* ("we're pretty sure the true value in the voting population is between  $43 \pm 3\%$ , i.e., somewhere between 40% and 46%.")

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

## Estimating with uncertainty with certainty (i.e., with some confidence), part 2

We're pretty sure the true value in the voting population is between  $43 \pm 3\%$ , i.e., somewhere between 40% and 46%. ”)

# How to trust (be confident) a sample estimate?



# Estimating uncertainty, with certainty!

- Most findings are based on samples, i.e., we always have incomplete knowledge about the population of interest.
- Imagine now a method in which we can state that “we have some confidence” that the true parameter of interest (say mean height of humans, or trees, etc) is between two values:
- Example 1: The average height of all humans (i.e., statistical population) is between 0 m and 100 m. This is absolutely true but useless, i.e., all humans are taller than 0 m and shorter than 100 m; so...the average has to be in this interval but this doesn't help me to make any precision.

# Estimating uncertainty, with certainty!

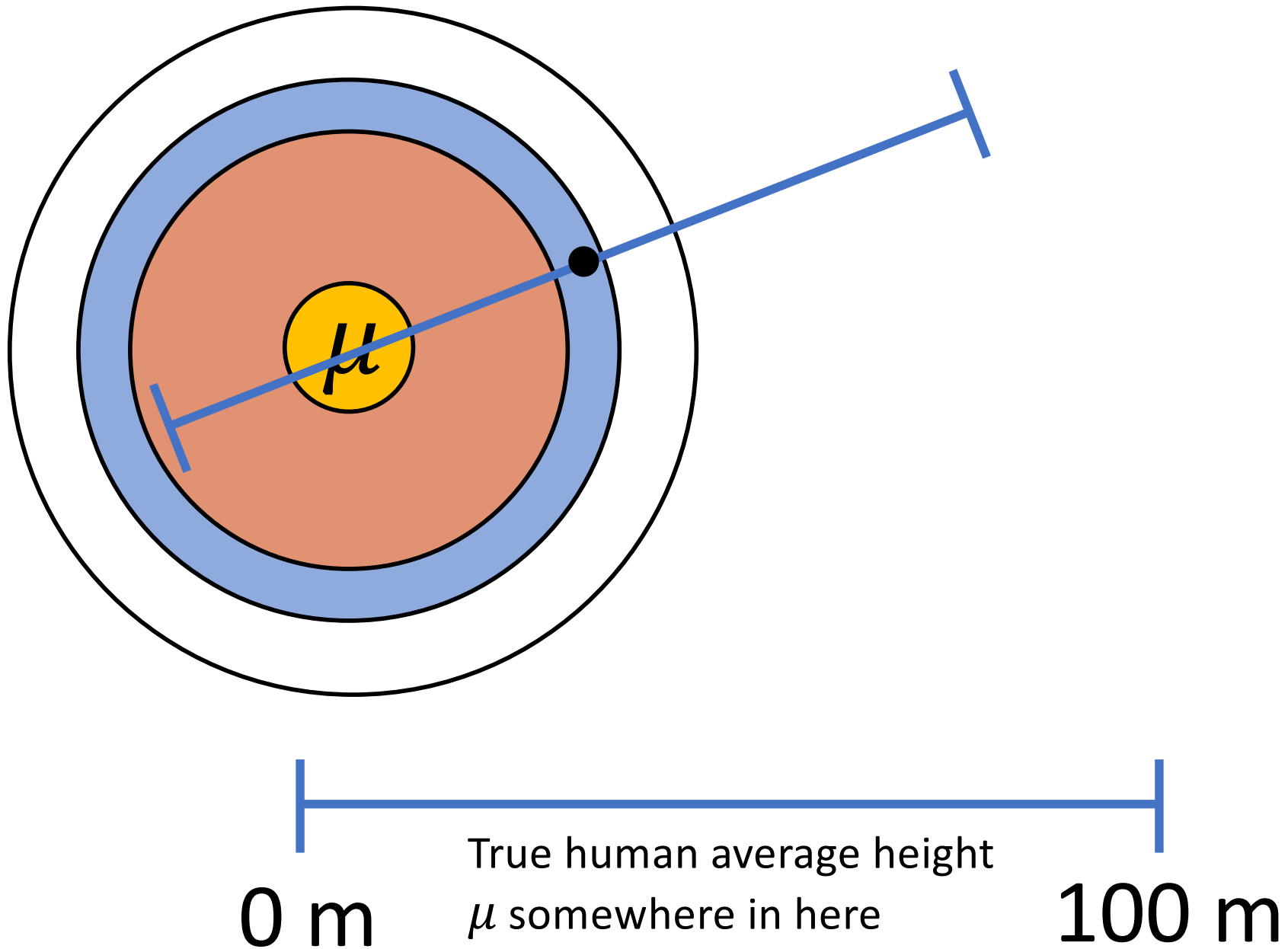
Example 1: The average height of all humans (i.e., statistical population) is between 0 m and 100 m. This is absolutely true but useless, i.e., all humans are taller than 0 m and shorter than 100 m; so...the average has to be in this interval but this interval doesn't help me to make any precise claims about the population parameter of interest (i.e. gain some precise knowledge).



$\mu$  somewhere in the interval

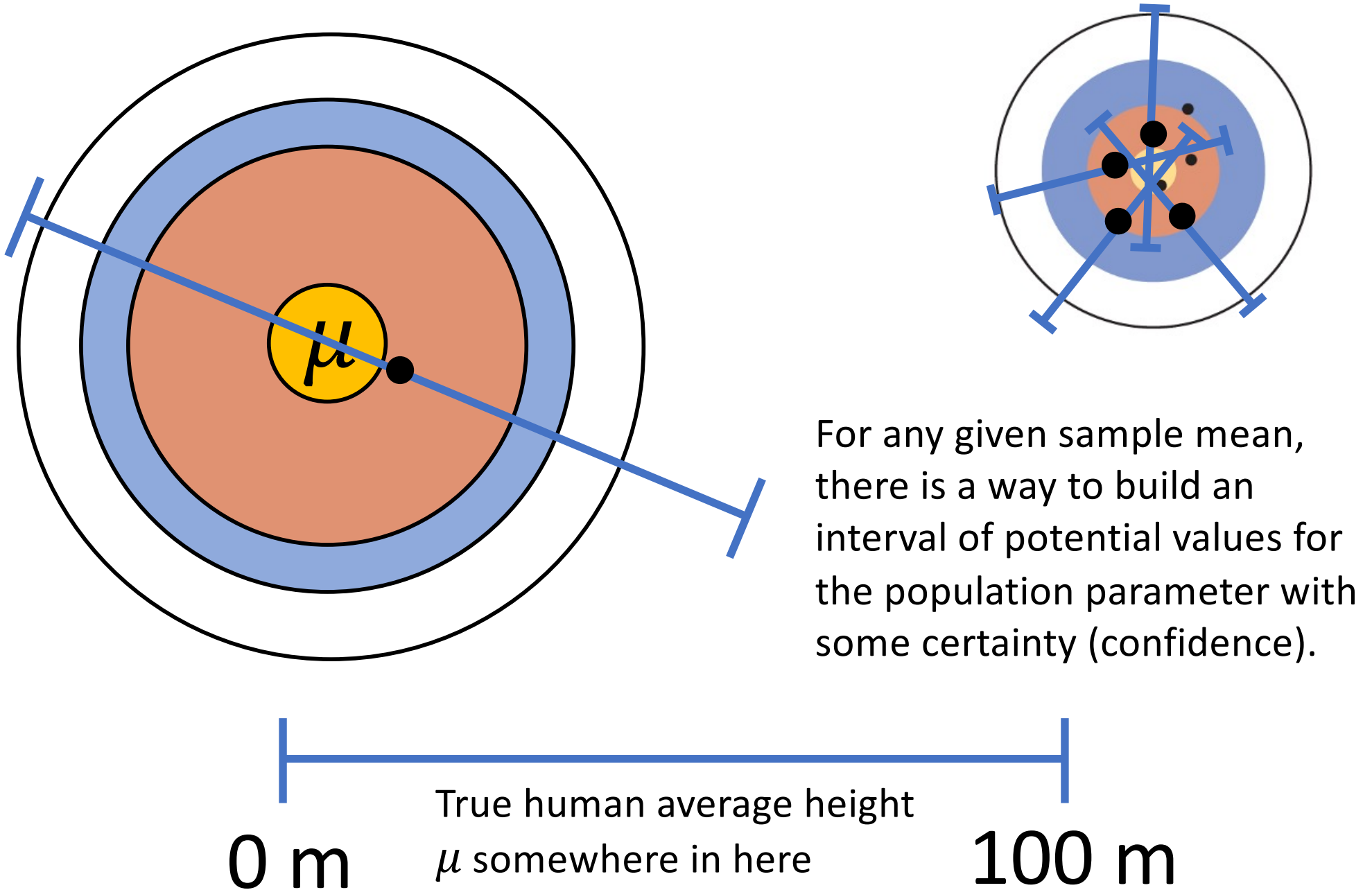


# Estimating uncertainty, with certainty!





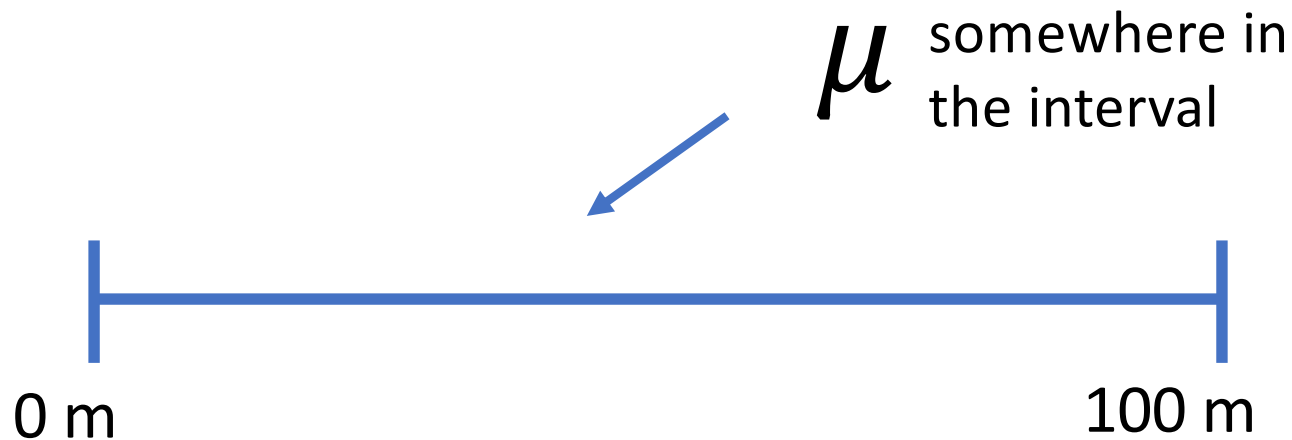
# Estimating uncertainty, with certainty!



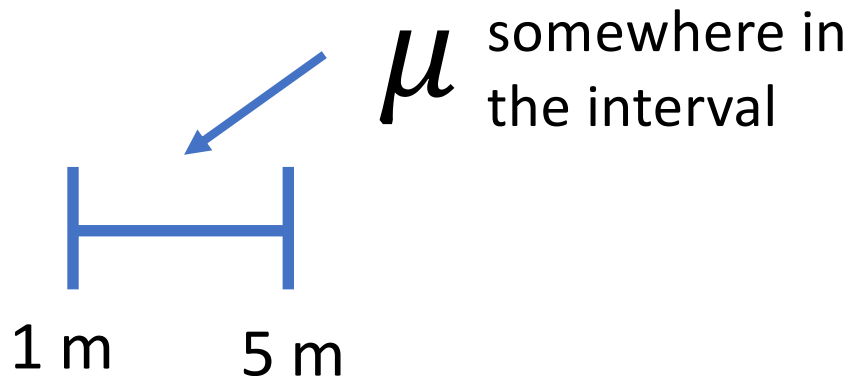
Let's take a break – 2 minutes!



# Estimating uncertainty, with certainty!



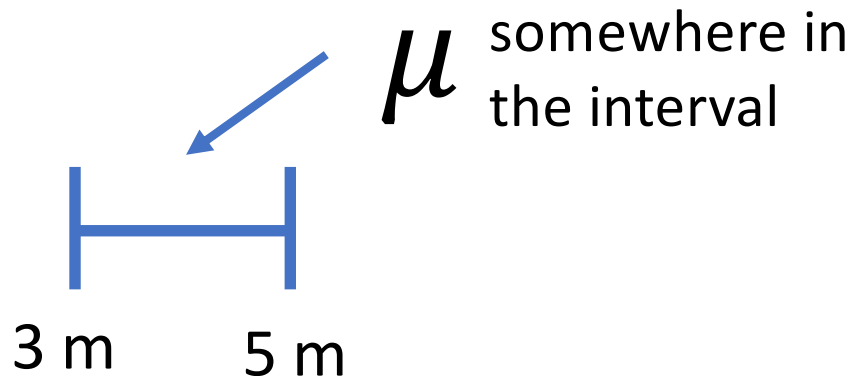
- Example 2: The average height of all adult humans is between 1 m and 5 m. This interval is more useful than the first but still not very useful because it is likely that the average of all humans are taller than 1 m and shorter than 5 m; so...the true average will be within this interval.



# Estimating uncertainty, with certainty!

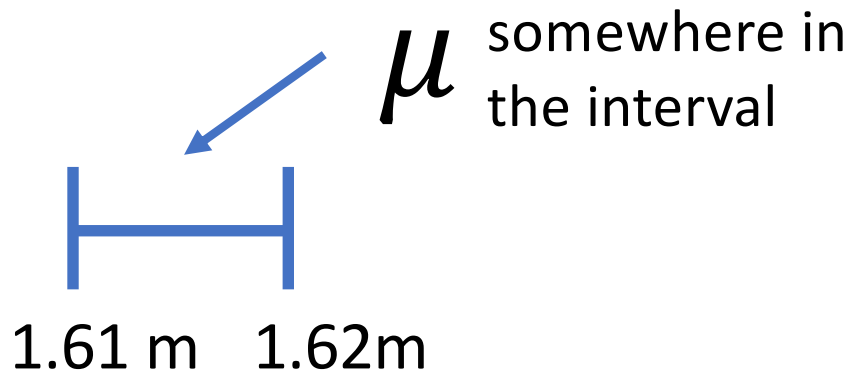
- Example 2: The average height of all adult humans is between 1 m and 5 m. This interval is more useful than the first but still not very useful because it is likely that the average of all humans are taller than 1 m and shorter than 5 m; so...the true average will be within this interval.

- Example 3: The average height of all adult humans is between 3 m and 5 m. This interval is actually wrong because it is impossible that the average of all humans are taller than 3 m and shorter than 5 m; so...the true average is not within this interval.



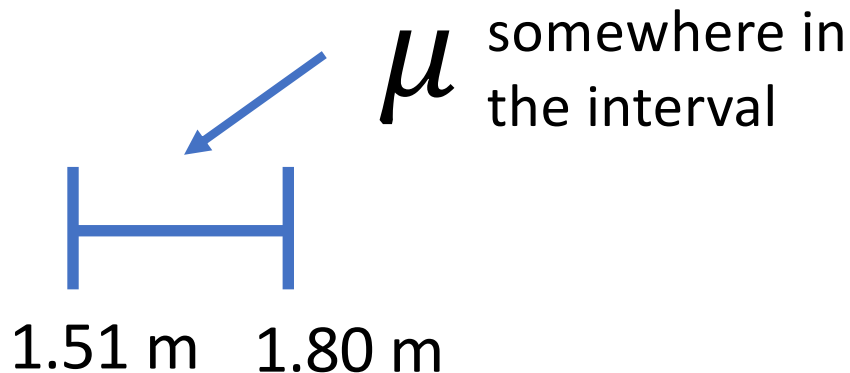
# Estimating uncertainty, with certainty!

- Example 3: The average height of all adult humans is between 3 m and 5 m. This interval is actually wrong because it is impossible that the average of all humans are taller than 3 m and shorter than 5 m; so...the true average is not within this interval.
- Example 4: The average height of all adult humans is between 1.61 m and 1.62 m. This interval may be true but it may be very wrong (perhaps too narrow).



# Estimating uncertainty, with certainty!

- Example 3: The average height of all adult humans is between 3 m and 5 m. This interval is actually wrong because it is impossible that the average of all humans are taller than 3 m and shorter than 5 m; so...the true average is not within this interval.
- Example 4: The average height of all adult humans is between 1.61 m and 1.62 m. This interval may be true but it may be very wrong.
- Example 5: The average height of all adult humans is between 1.51 m and 1.80 m. This interval is likely the most trustworthy; but is it useful?



# Estimating uncertainty, with certainty!

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

The method to build confidence intervals for the population mean is based on the sampling distribution of means, the sample mean and the sample standard deviation!

So....although we don't know what the true value (parameter) is for sure, we can estimate an interval that can tell us with a certain degree of confidence what that true value is!

BTW, we can build confidence intervals for other statistics such as the standard deviation, variance, medians, etc.

## **Estimating uncertainty, with 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 uncertainty, with 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.

The way we can trust samples is through random sampling (accuracy is assured) and increasing sample size (increase precision).

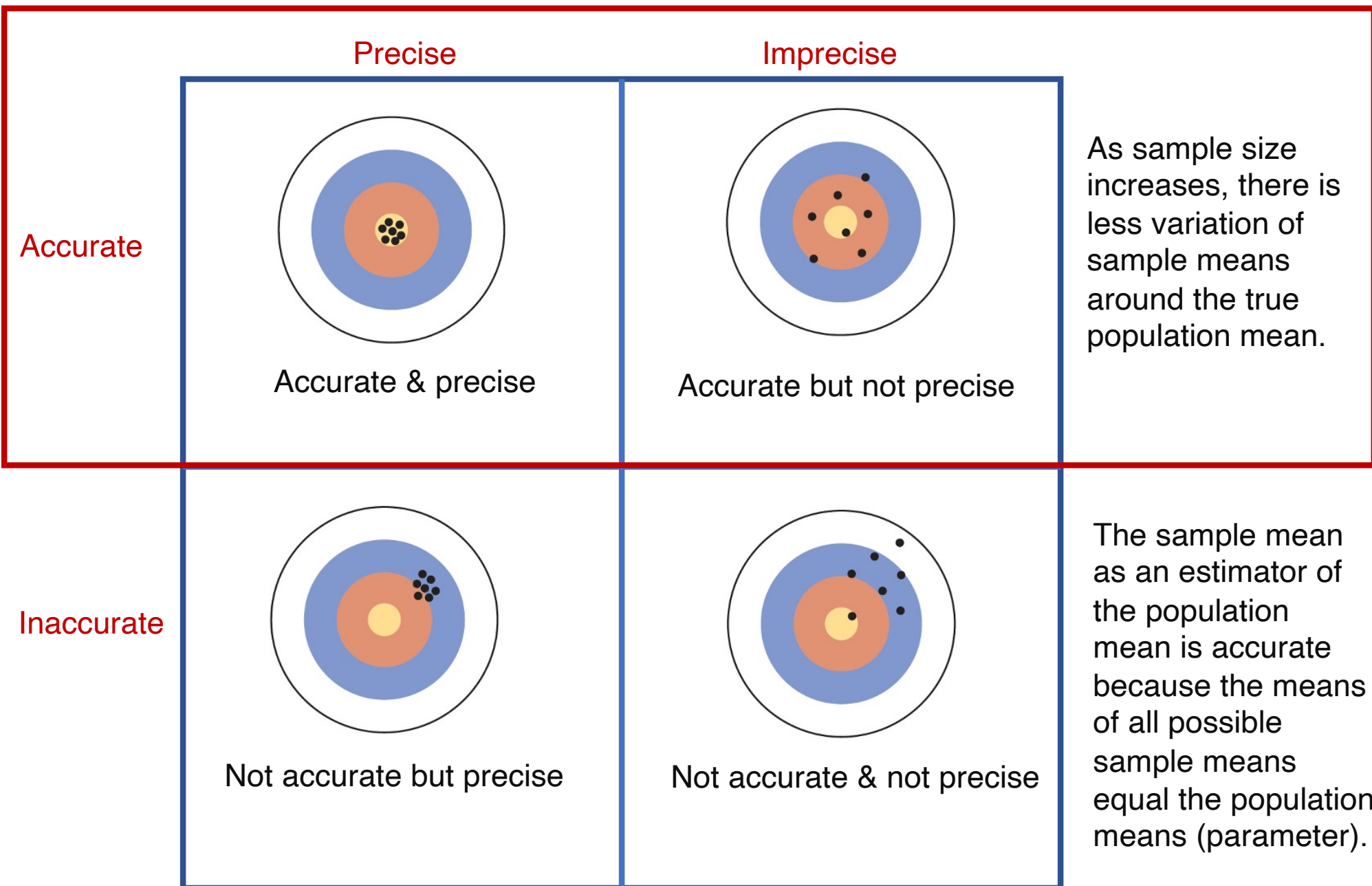
# Estimating uncertainty, with certainty!

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.

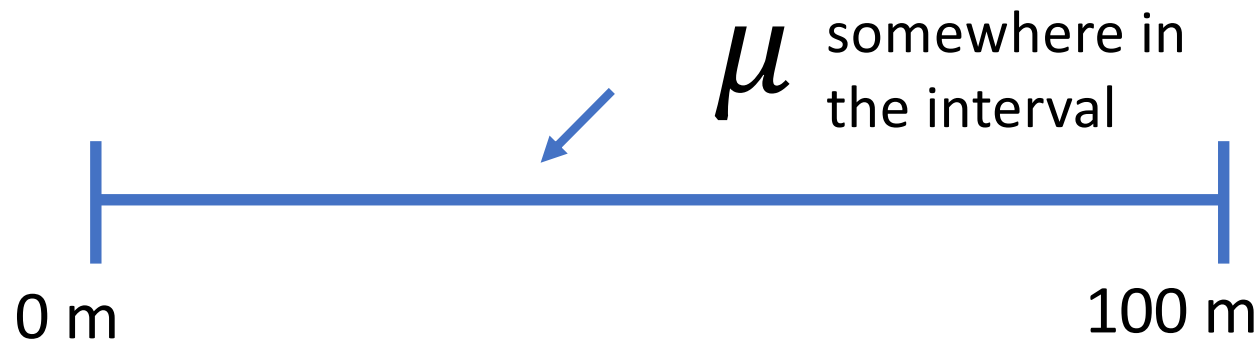
The way we can trust samples (i.e., they are accurate) is through random sampling (accuracy is assured) and increasing sample size (increase precision).

Statistical populations with relatively smaller variances increase precision, but this is a luxury that researchers don't always control. Or perhaps possible by defining more specific problems: e.g., average height of humans *versus* average height of adult humans.

# What we want is to increase precision while being accurate while sampling (assuming that estimator is accurate)



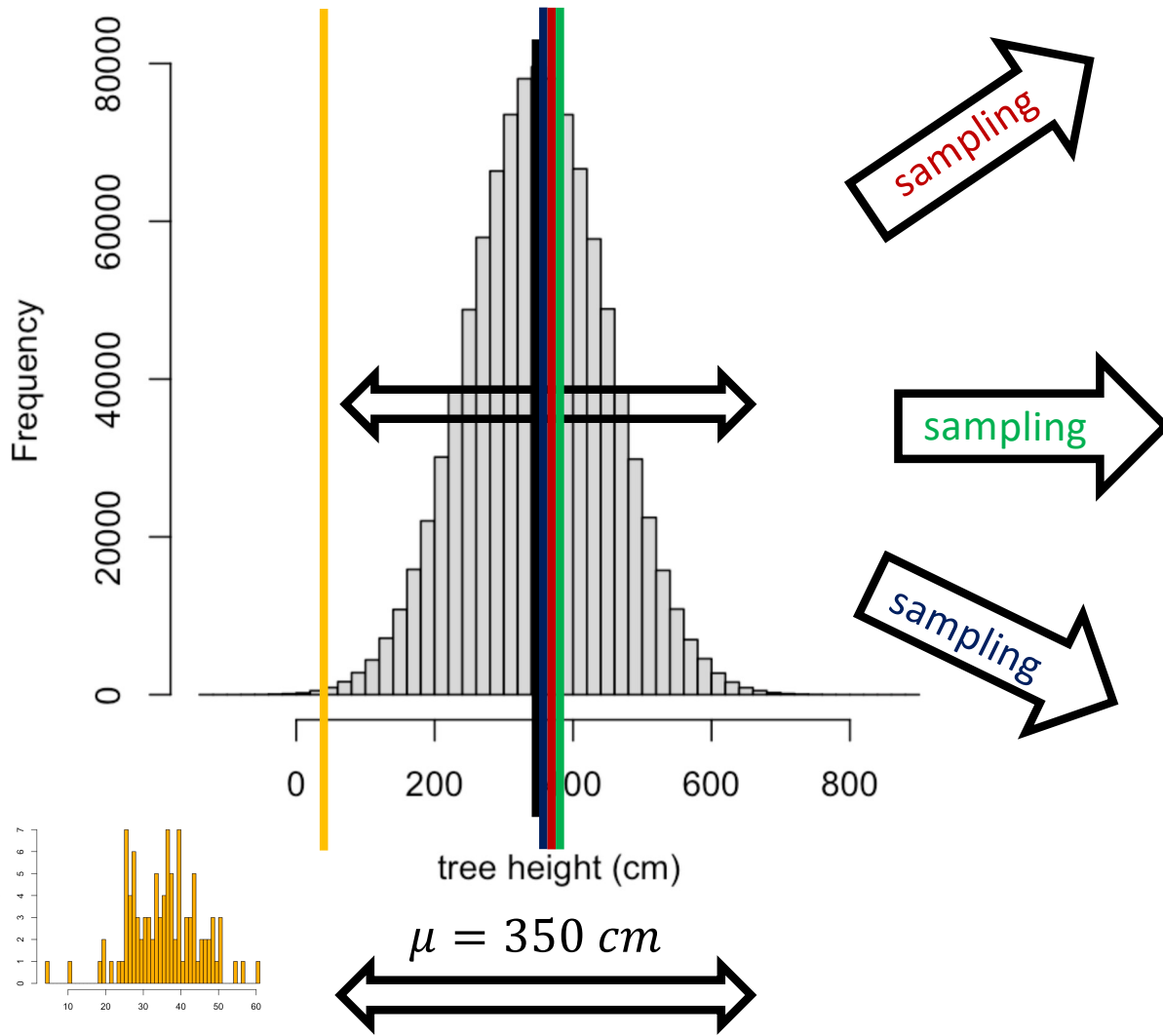
# Estimating uncertainty, with certainty! How to build certainty?



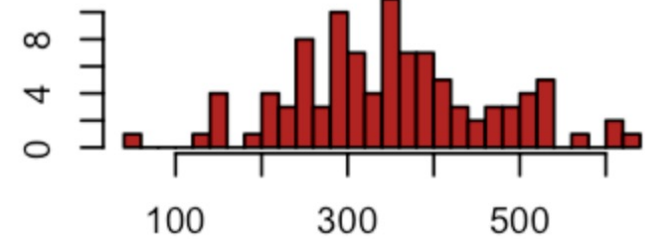
**Again, this interval is estimated on  
the basis of a single sample!**

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

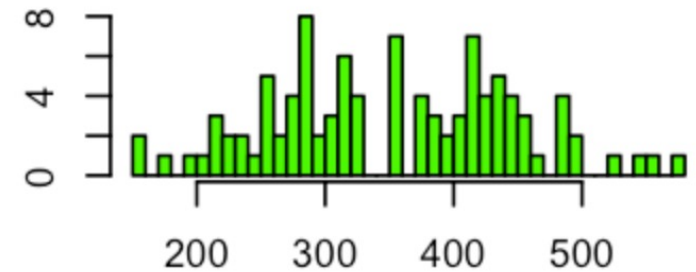
$\mu = 350 \text{ cm}; \sigma = 100 \text{ cm}$



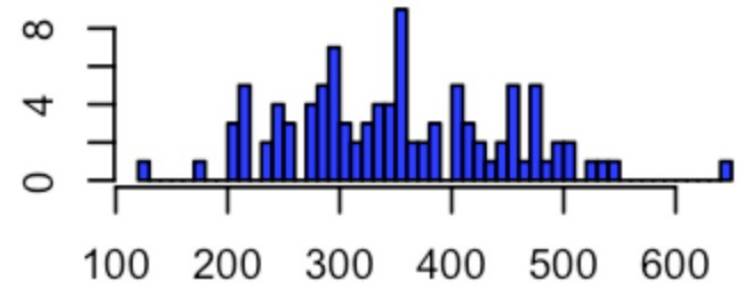
$\bar{X} = 351.5 \text{ cm}; s = 114.2 \text{ cm}$



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



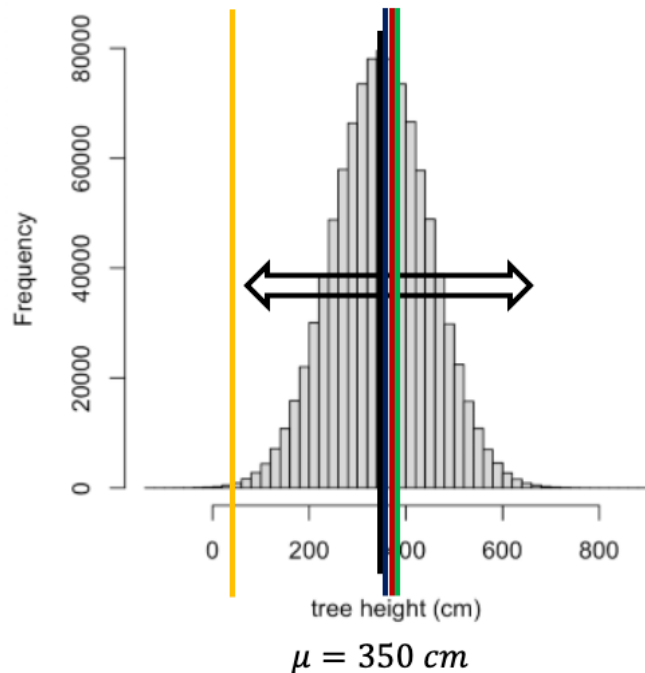
$\bar{X} = 351.4 \text{ cm}; s = 96.6 \text{ cm}$



Uncertainty (samples means varying around the true population mean)

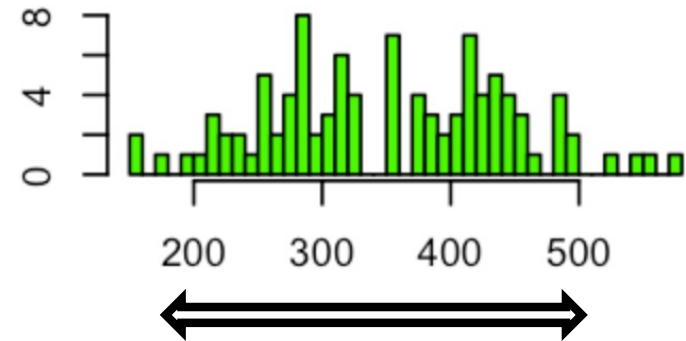
The variation among observations within samples (standard deviation) can inform us about how far sample means in general might be from the true population mean (estimate how wrong one could be).

$$\mu = 350 \text{ cm}; \sigma = 100 \text{ cm}$$



Variation among samples

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



Variation within samples

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

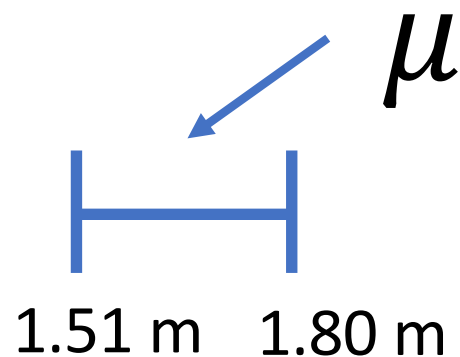
**The statistical road:  
estimate with  
uncertainty but with  
confidence**



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

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 values outside are less plausible, based on the sample data alone.

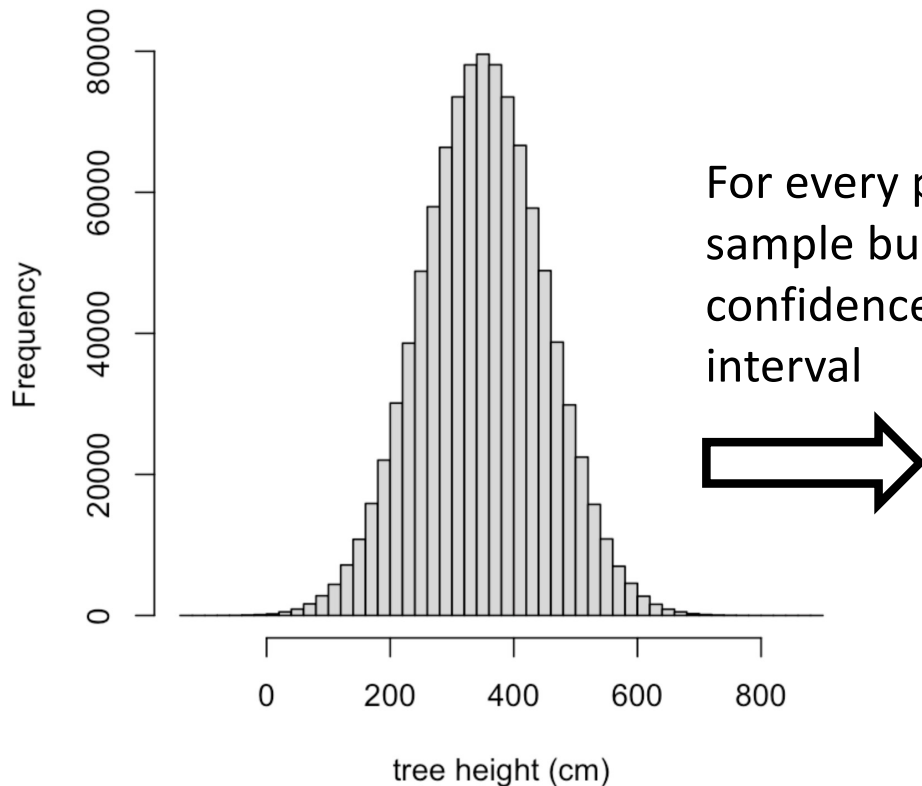


Very plausible (high confidence) that the population parameter is somewhere within The 95% confidence interval.



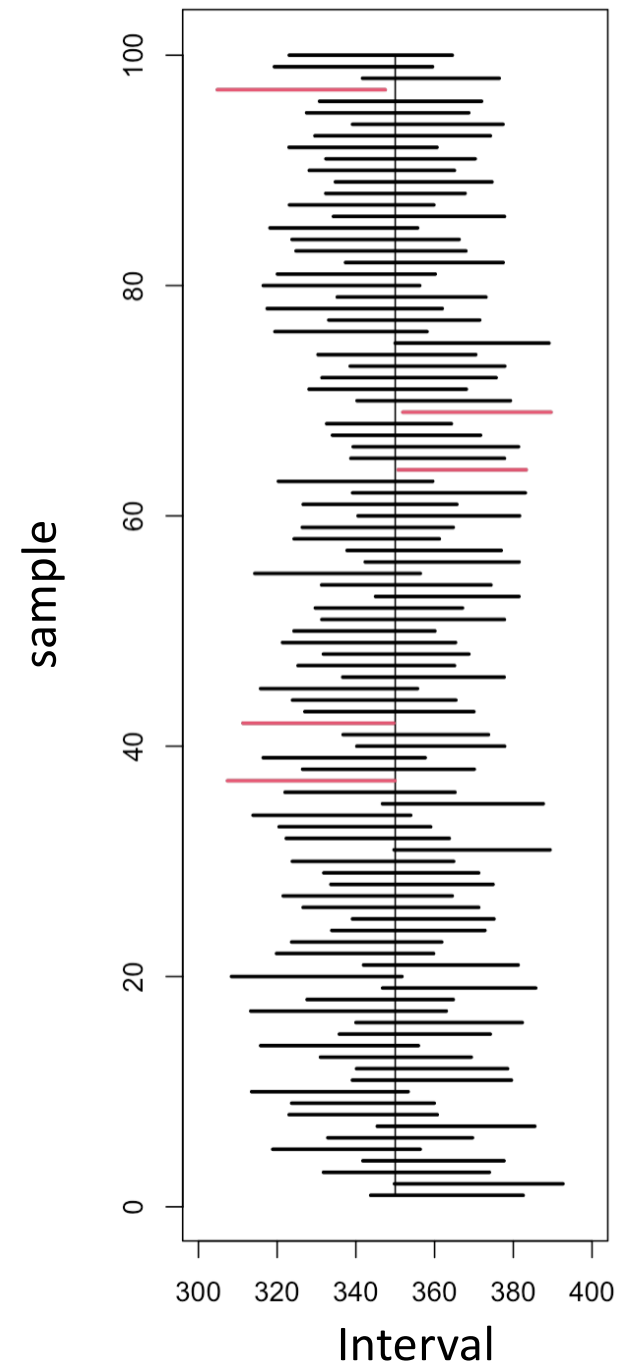
Imagine an interval referred as to “95% confidence interval”:

$$\mu = 350 \text{ cm}; \sigma = 100 \text{ cm}$$



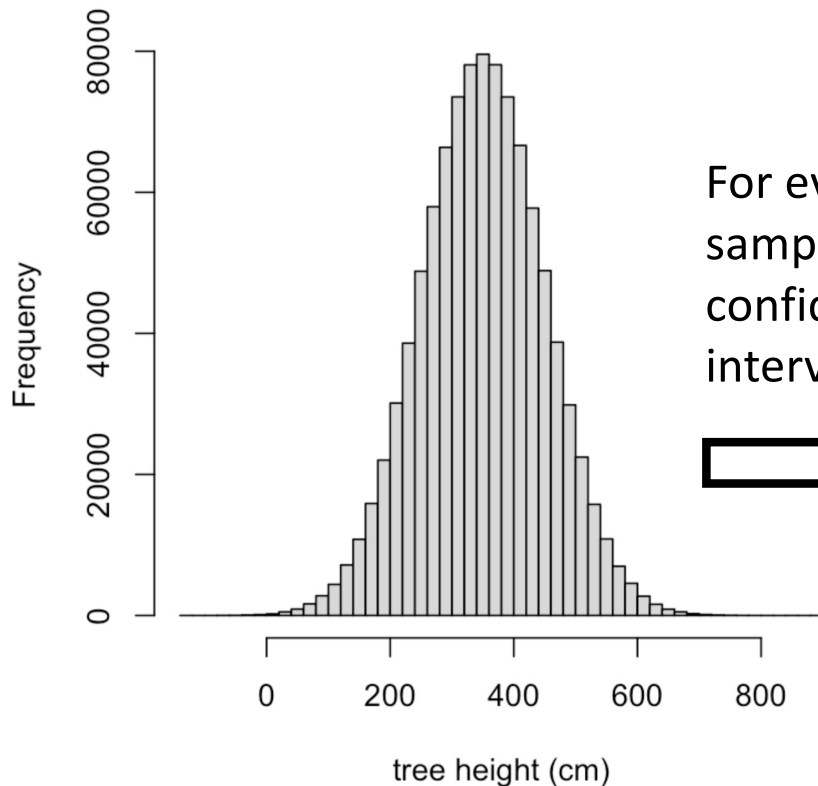
For every possible  
sample build a  
confidence  
interval

If sampling is random and distributional properties of the population (e.g., close to normality), 95 out of **100** (95%) sample intervals will contain the true population parameter; intervals not containing the true parameter are plotted in red (5%).



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

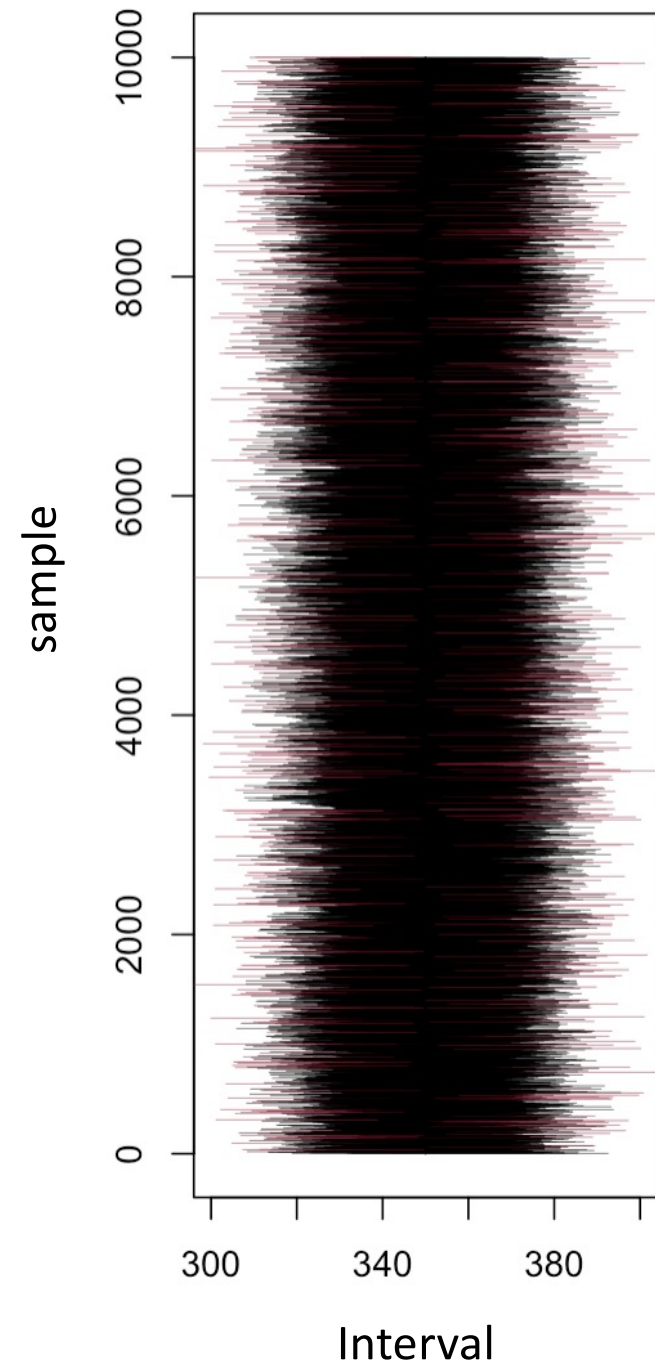
$$\mu = 350 \text{ cm}; \sigma = 100 \text{ cm}$$



For every possible  
sample build a  
confidence  
interval



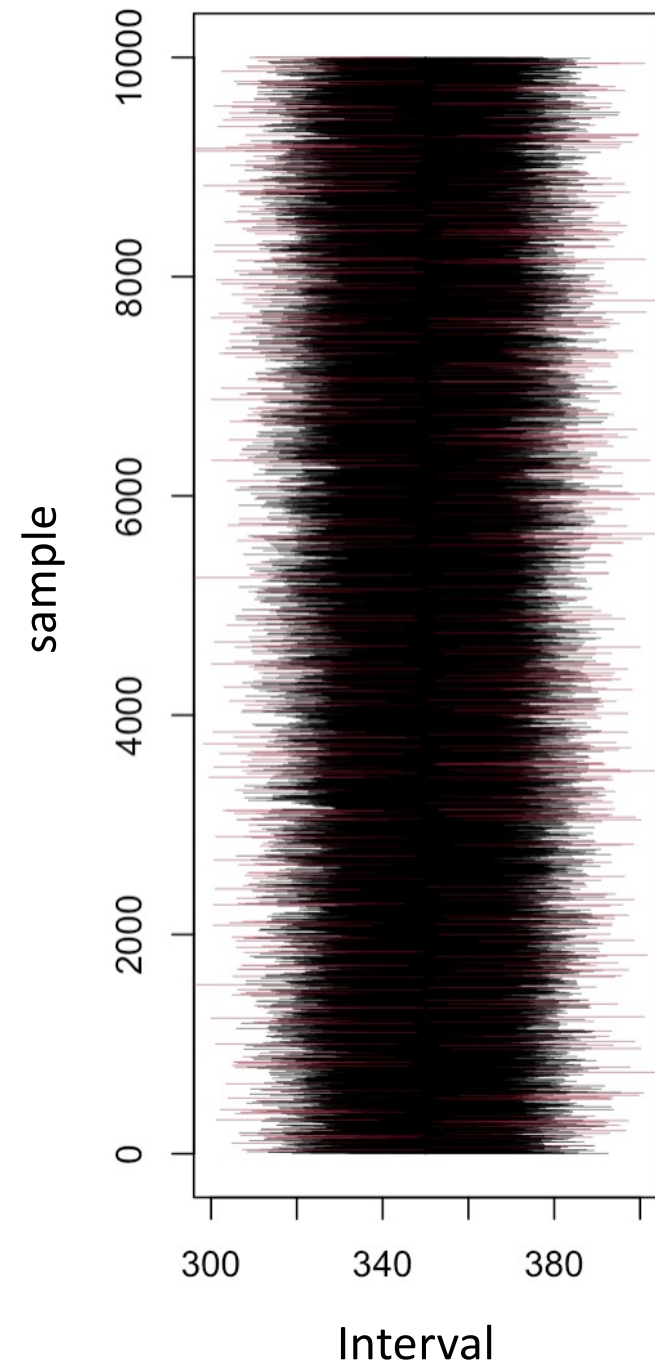
If sampling is random and distributional properties of the population (e.g., close to normality), 9500 out of 10,000 (95%) sample intervals will contain the true population parameter; intervals not containing the true parameter are plotted in red.



## Very important!

For any given sample confidence interval, we can state that “we are 95% confident that the true population mean lies between the lower and upper limits of the interval”.

We cannot say that “there is a 95% probability that the true population mean lies within the confidence interval”. Either the parameter is within the interval or not! So, no probability attached to this condition.



Let's take a break – 2 minutes!



Confidence intervals are not well grasped by a large number of users of statistics!

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



## Confidence Trick: The Interpretation of Confidence Intervals

Colin Foster

*School of Education, University of Nottingham, Nottingham, United Kingdom*

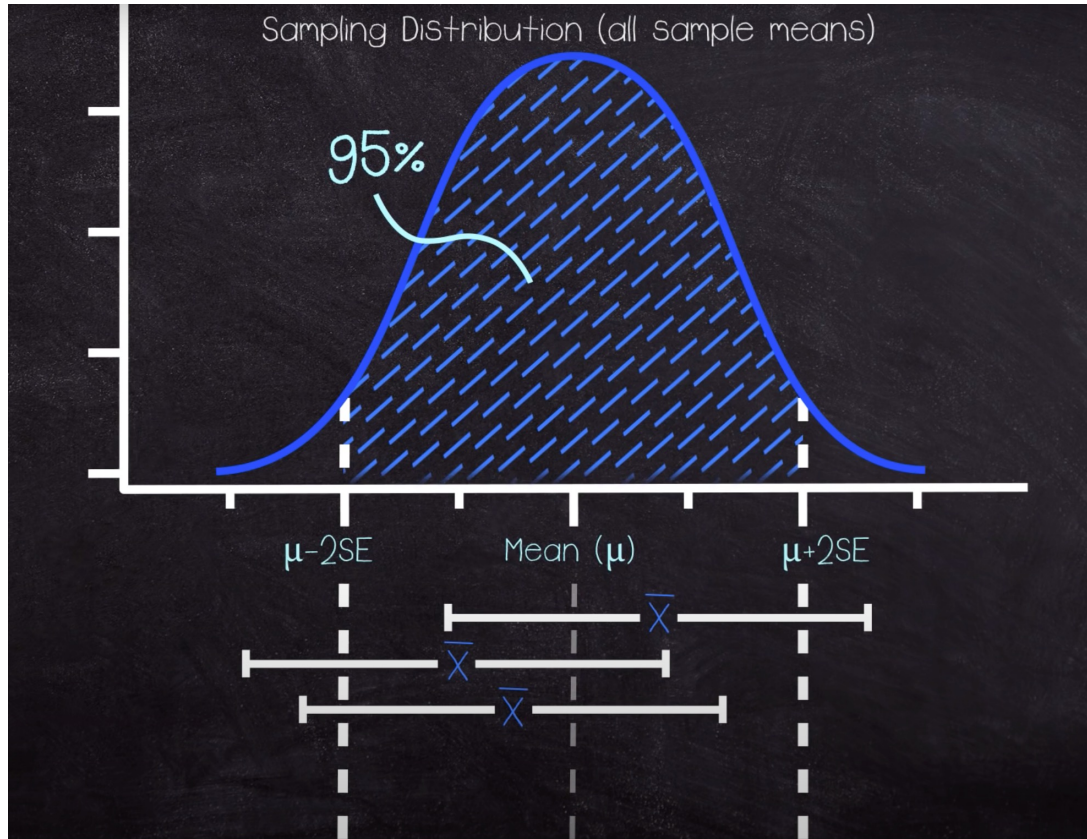
**Often stated by students and practitioners of statistics:** We cannot say that “there is a 95% probability that the true population mean lies within the confidence interval”. Either the parameter is within the interval or not! So, no probability attached to this condition.

## Assumptions underlying confidence intervals

If sampling is random and if the frequency distribution of the population is roughly normal, then exactly 95% out of the infinite possible sample intervals will contain the true population parameter in the way we calculate confidence intervals!!

If the frequency distribution of the population is not normal, a number close to 95% (e.g., 92% or 97%) out of the infinite possible sample intervals will contain the true population parameter!! That number (as we will see later) will depend on the distributional properties of the population (asymmetry, etc).

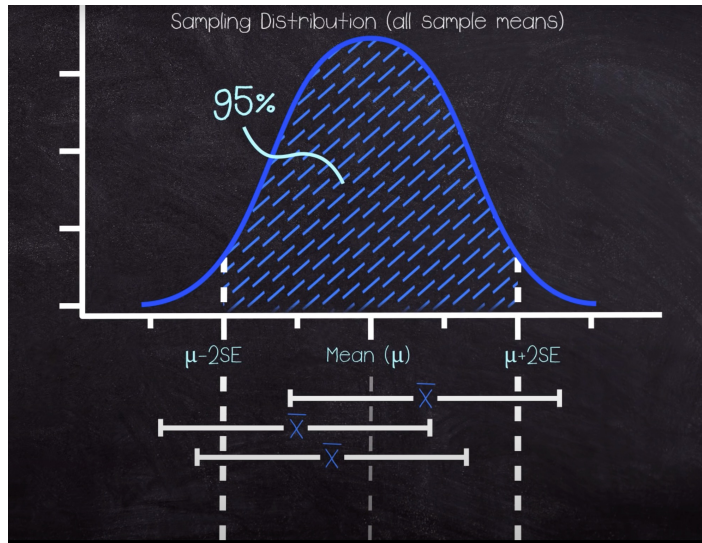
We use the sampling distribution of all sample means to calculate confidence intervals



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  $\pm 2 \times SE$**  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.

What does  $SE_{\bar{y}}$  mean? It is the standard deviation of all sample means, i.e., the average difference between each sample mean and the true population mean. We don't need to know the population parameter (true mean value) to estimate the sampling distribution (next lecture).

$SE_{\bar{y}}$  can be estimated from the sample standard deviation  $s$  as follows:

$$SE_{\bar{y}} = \frac{s}{\sqrt{n}}$$



# We use the sampling distribution of all sample means to calculate 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:

$$\sigma_{\bar{Y}} = \sqrt{\sum_{i=1}^{\infty} \frac{(\bar{Y}_i - \mu)^2}{\infty}} = \text{SE}_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

The number of samples is so large that can be considered infinite ( $\infty$ )

$\sigma =$  *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:

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

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

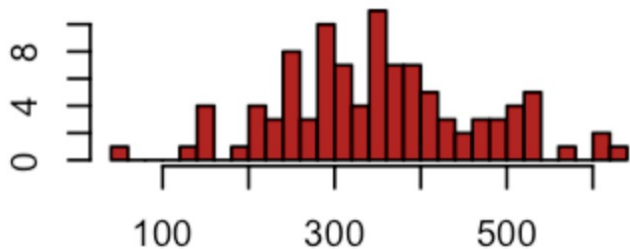
$\sigma$  = the standard deviation of the population

$SE_{\bar{Y}}$  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:

$$SE_{\bar{y}} = \frac{s}{\sqrt{n}} = \frac{114.2}{\sqrt{100}} = 11.42$$

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



$SE_{\bar{y}}$  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 sample size is relatively large, then this interval can be calculated based on the sample standard error  $SE_{\bar{Y}}$  (why? Next lecture).

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

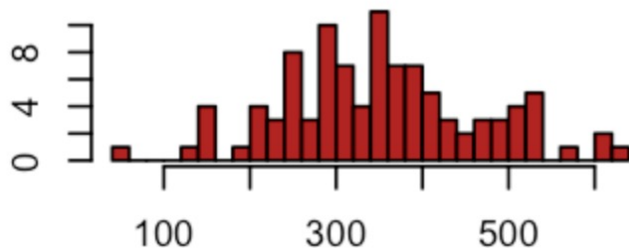
$$\bar{X} = 351.5 \text{ cm}; s = 114.2 \text{ cm}$$

Margin of error

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

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  $SE_{\bar{y}}$  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 somewhat with sample size by tend to be around **2** when  $n > 30$ .

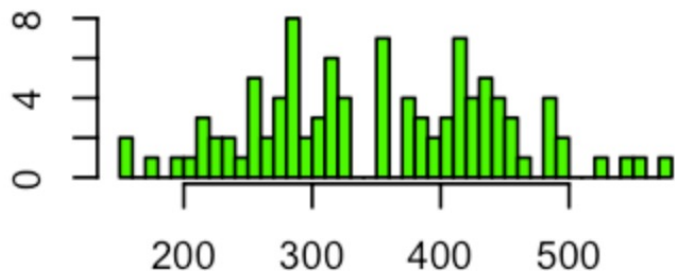
$$351.5 \pm \mathbf{2} \times \frac{114.2}{\sqrt{100}}$$

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

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

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

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



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

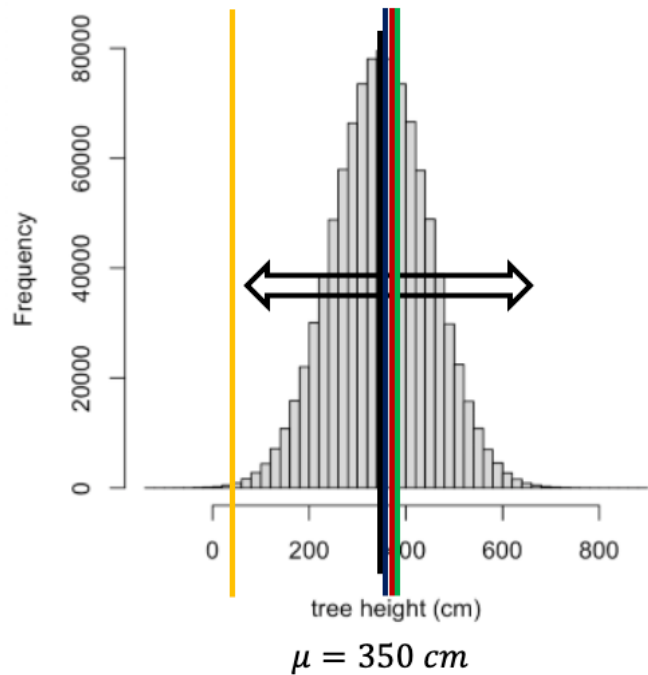
333.5 cm

371.1 cm



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).

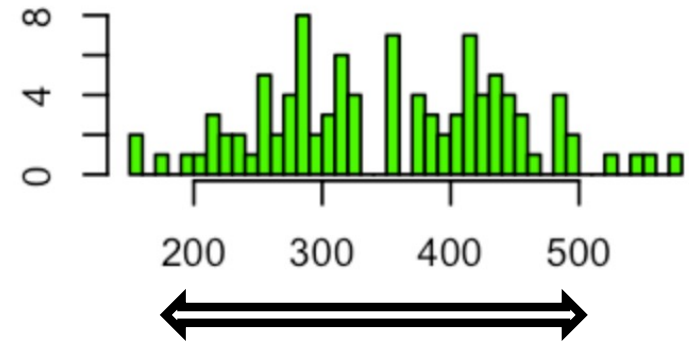
$$\mu = 350 \text{ cm}; \sigma = 100 \text{ cm}$$



Variation among samples



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



Variation within samples

333.5 cm

371.1 cm



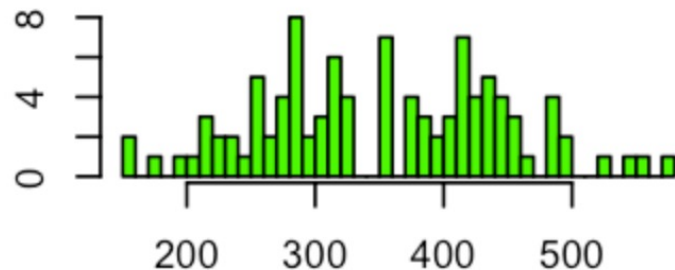
$$352.3 \pm 2 \times \frac{114.2}{\sqrt{100}}$$

## 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**.

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



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

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

333.5 cm

371.1 cm





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

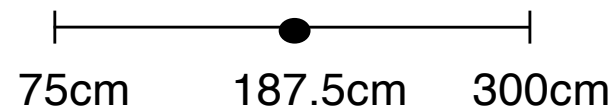
In general, the 95% confidence interval is a good measure of our uncertainty about the true value of the parameter (population value).

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?



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



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

51% of the voting intention is not among the most plausible values, therefore more work needs to be done!

“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?* (“we're pretty sure the true value in the voting population is between  $43 \pm 3\%$ , i.e., somewhere between 40% and 46%.”)