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.

Lecture 9: Estimating with uncertainty, but with a degree of certainty (i.e., with some confidence), part 1

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

Random sampling minimizes sampling error and inferential bias by preventing systematic over- or under-representation of observational units, so that sample estimates vary randomly around the true population value rather than being consistently displaced from it.

Sampling error refers to how much a statistic calculated from a sample differs from the true population value simply due to random variation in which units happen to be included in the sample.

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.

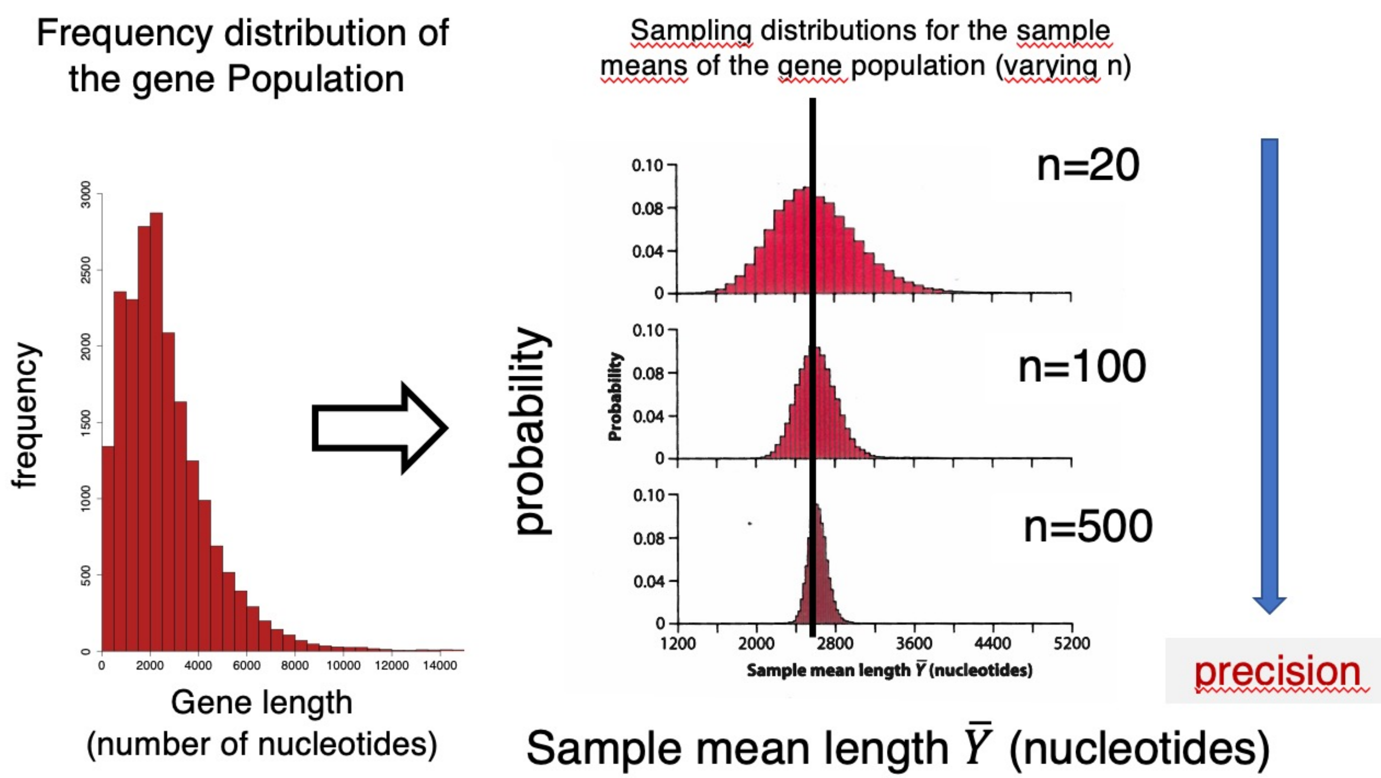
2) The selection of observational units in the population (e.g., individual tree) must be **independent**, i.e., the selection of any unit (e.g., individual tree) of the population must not influence the selection of any other unit.

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

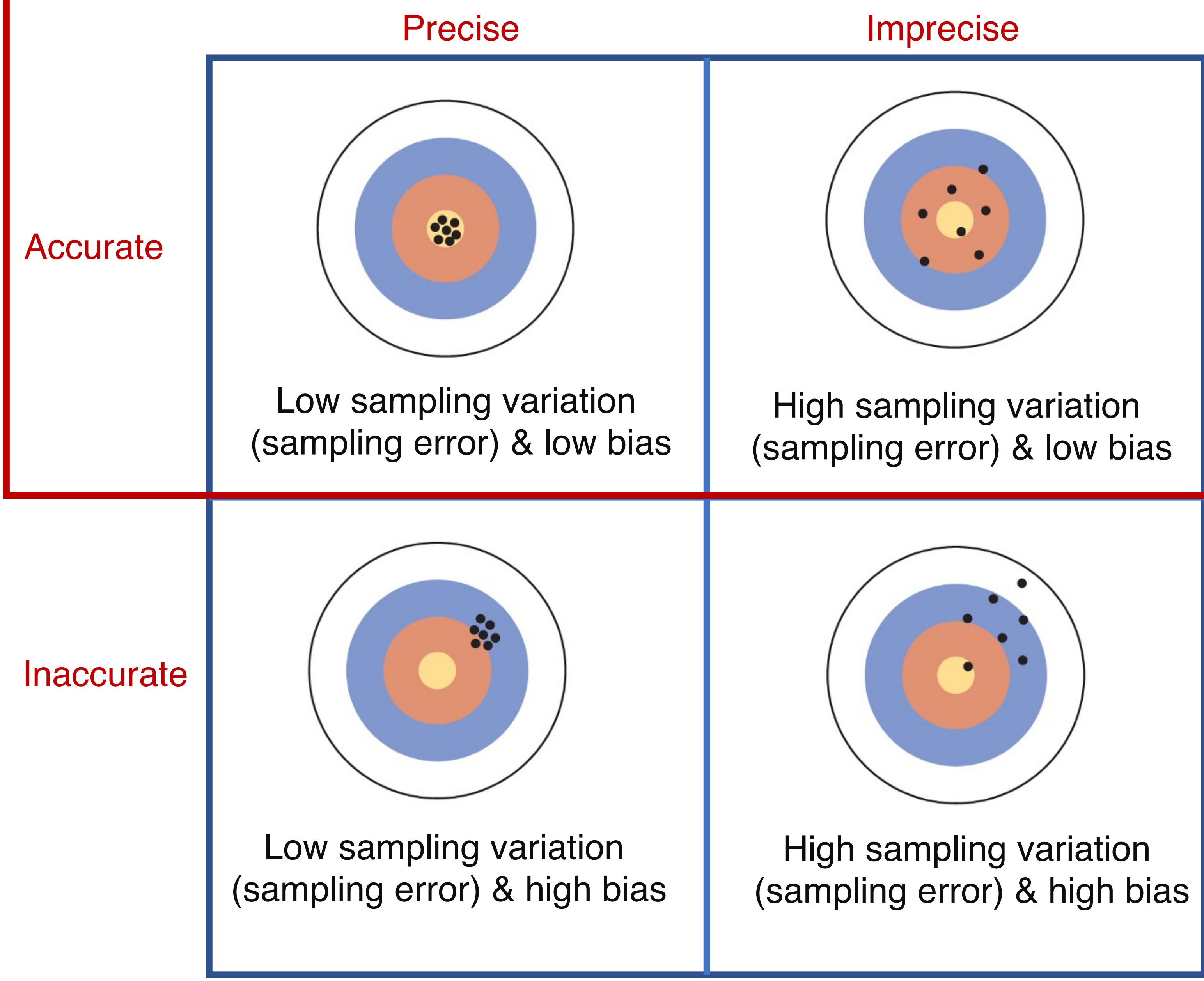
Increasing sample size reduces **sampling error**, which improves precision (estimates vary less across repeated samples).

Increasing sample size **does not** correct systematic bias; biased studies become more precisely biased.

Under unbiased sampling, larger samples increase the probability that an estimate is close to the true population value, but sample size itself controls precision—not accuracy.

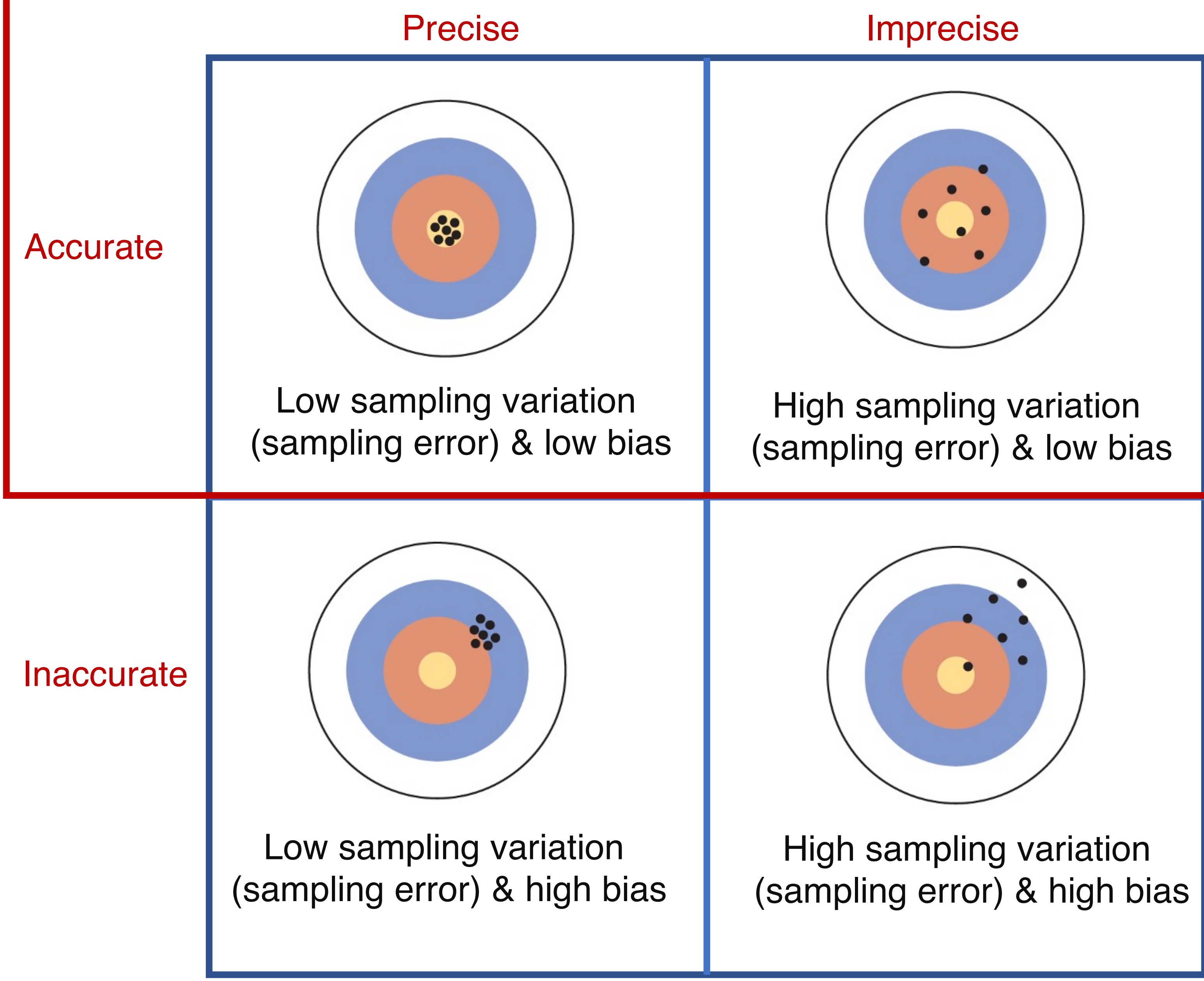


Regarding the estimation of population means, what does random sampling assure? **Accuracy!**



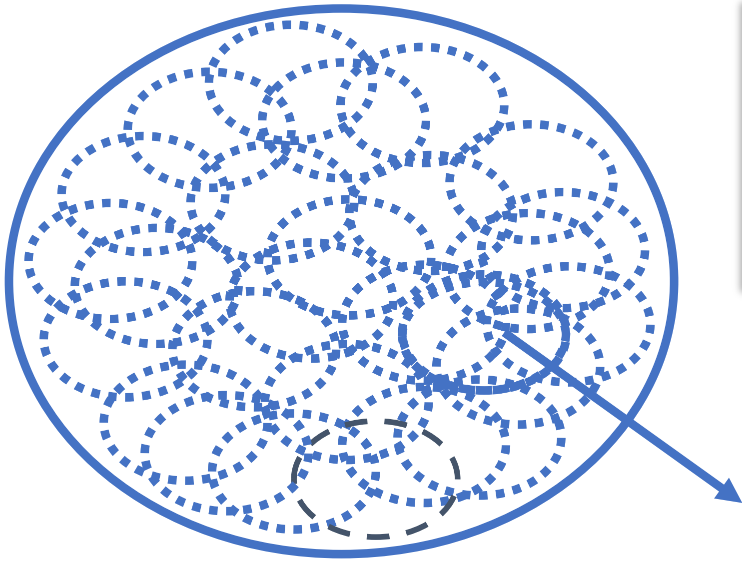
Under random sampling, the sample mean is an unbiased estimator of the population mean: although any single sample mean may differ from the true value, the average of sample means over repeated samples equals the population mean.

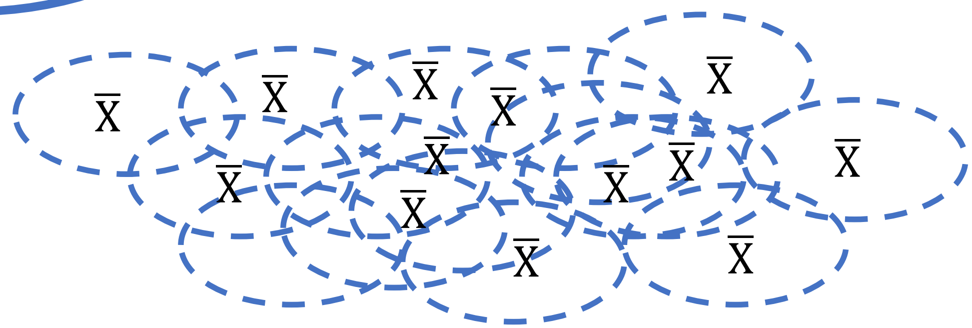
Regarding the estimation of population means, what does random sampling assure as sample size increases? **Precision!**

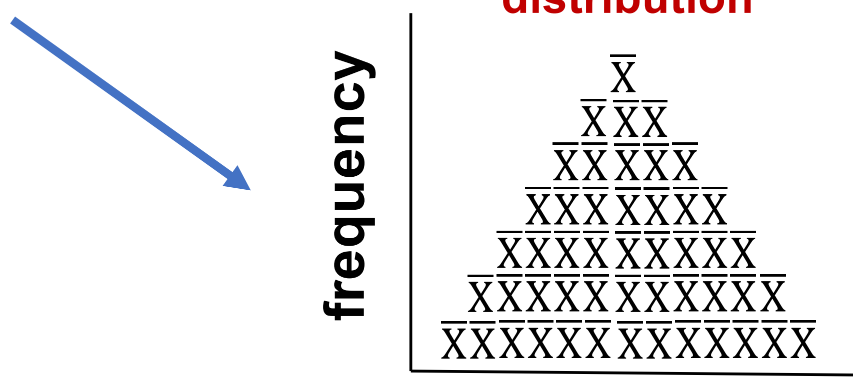


As sample size increases, the variation of sample means around the true population mean decreases.

Recap: Building the sampling distribution of the sample mean



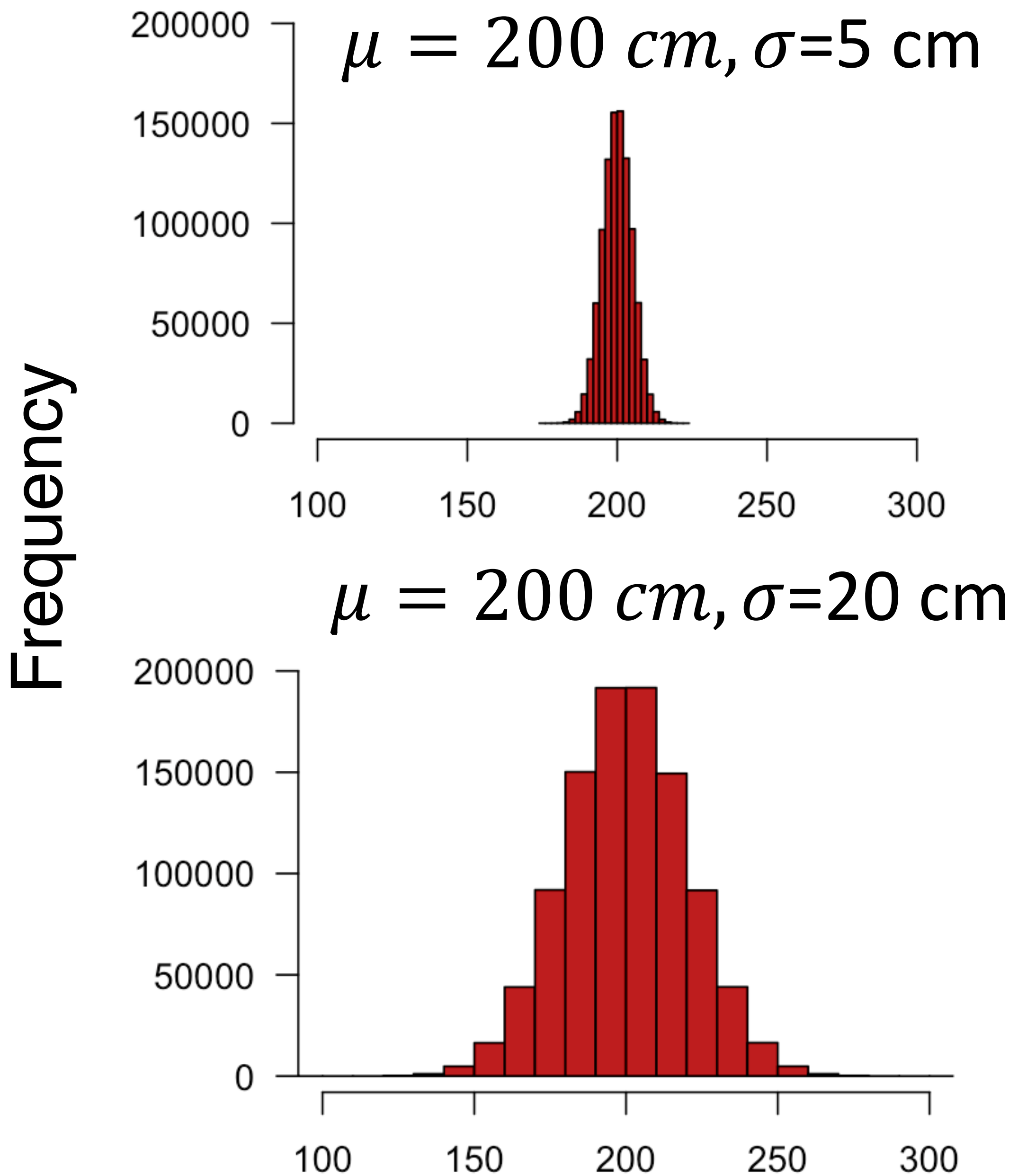




Samples have the same sample size n.

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

Assuming unbiased sampling, precision depends not only on sample size but also on the population standard deviation: populations with higher variability produce less precise estimates for the same sample size.

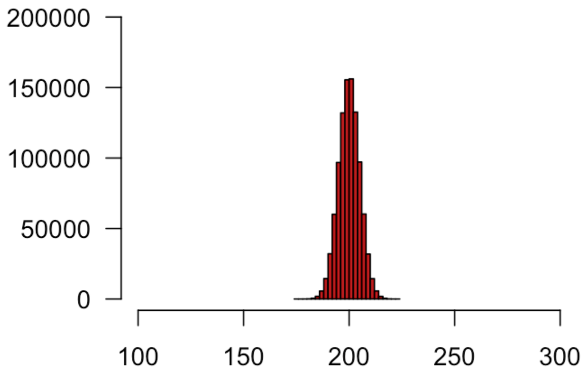


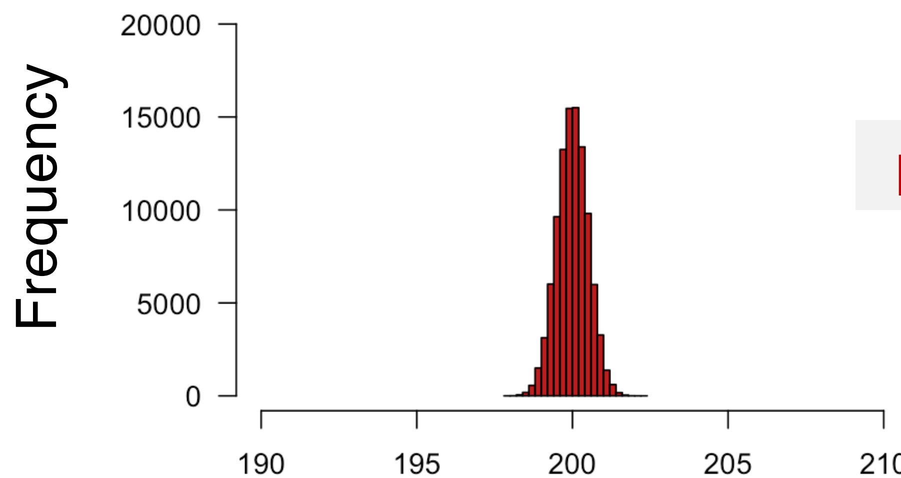
Two populations with identical means (μ = 200) but differing population standard deviations σ.

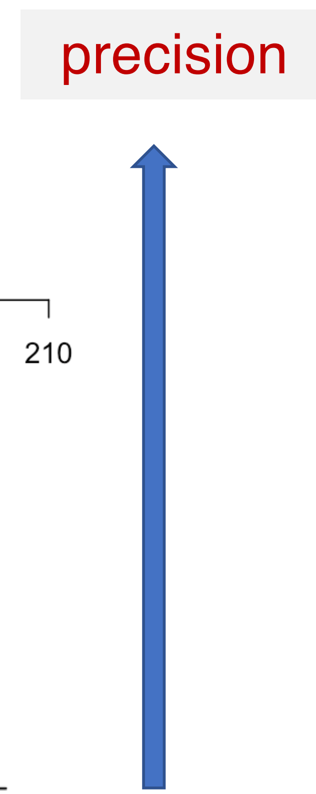
What else affects precision assuming that accuracy is correct?
The standard deviation (or variance) of the statistical population!

Sampling distributions of the sample mean for two populations with identical means (μ = 200) but differing population standard deviations σ. n=100; 100,000 samples.

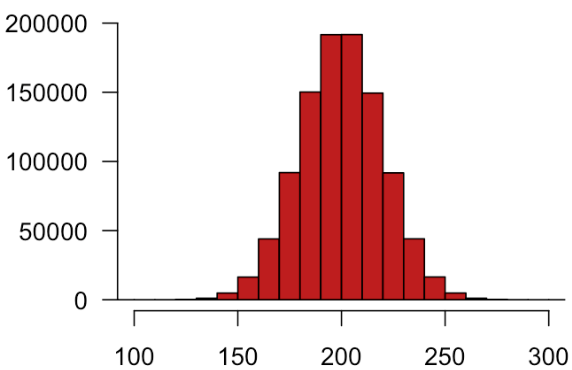
$\mu = 200\ cm, \sigma$=5 cm

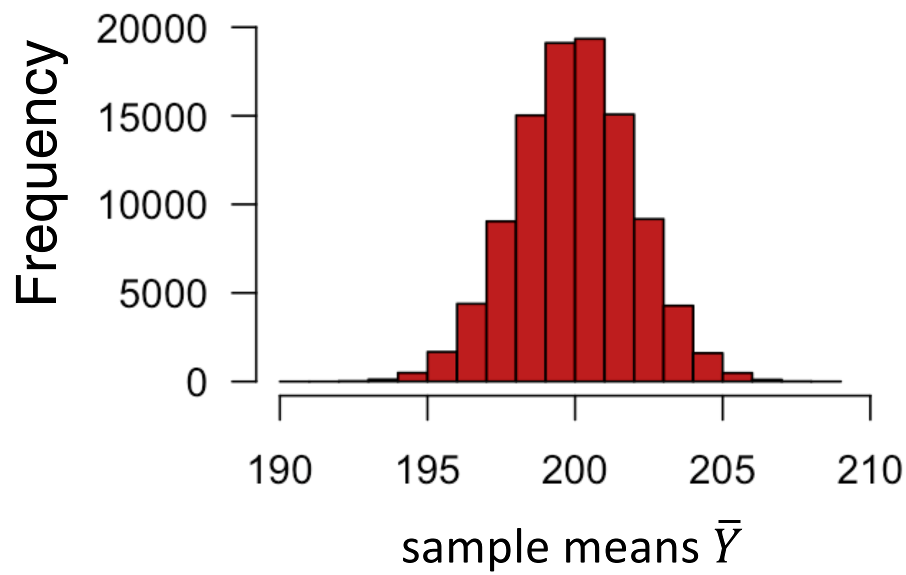






$\mu = 200\ cm, \sigma$=20 cm





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

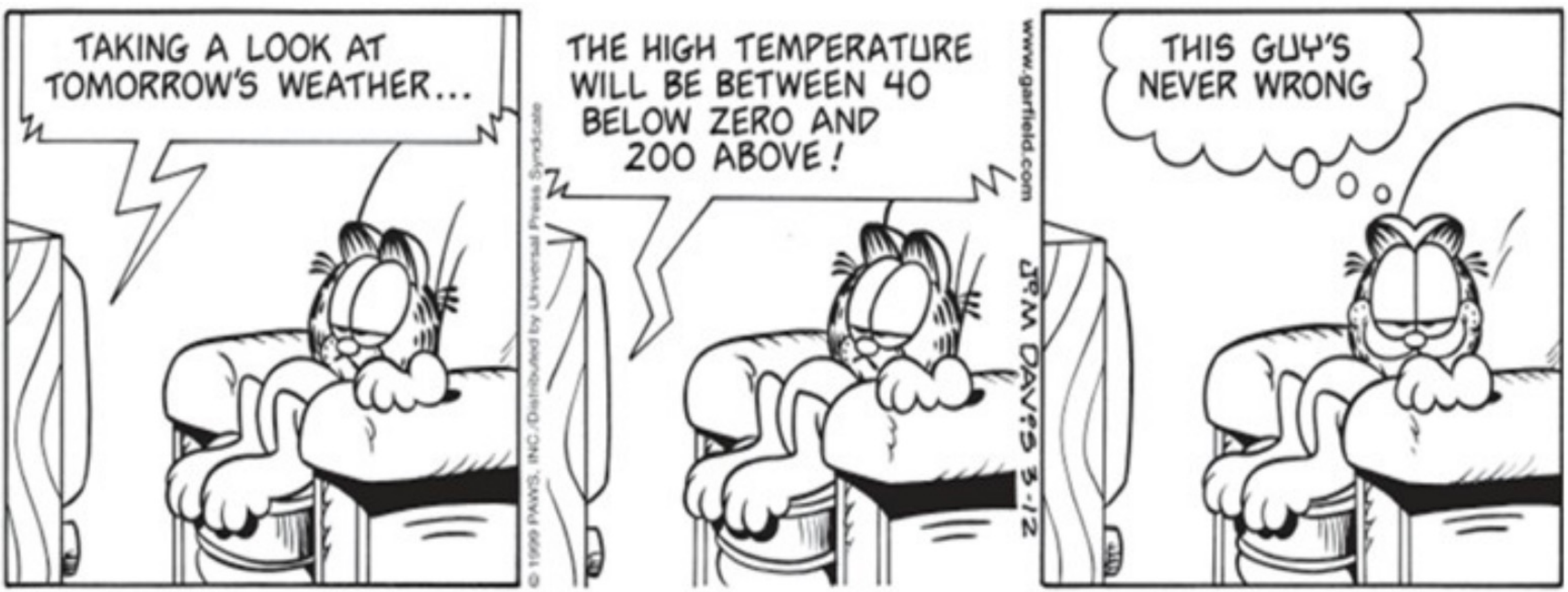
Uncertainty reflects how estimates vary across repeated samples. How can we quantify that uncertainty when we observe only one sample?

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 quantify our confidence in a sample estimate, given sampling variation?

Said differently: How can we evaluate the reliability of a sample estimate when it is only one realization from a sampling process?



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

Estimating with uncertainty, but with a degree of certainty

- Most scientific conclusions are drawn from samples, not entire populations. As a result, we always have incomplete knowledge about the true population parameters of interest.

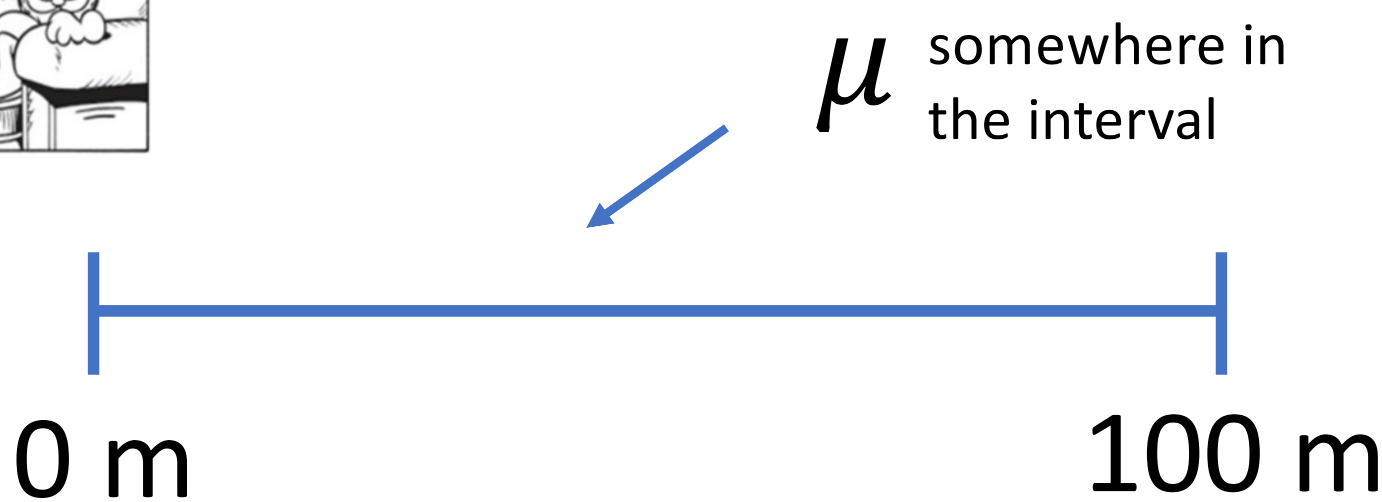
- Now imagine a method that allows us to say, "We are confident that the true value of a population parameter (e.g., the mean height of humans or trees) lies within a specific range." The usefulness of such a method depends entirely on how informative that range is. Let's call this range "confidence interval".

- Example 1. "The average height of all humans lies between 0 m and 100 m." In this example, 50 m would be the average estimated and the margin of error would be also 50 m; i.e., 50 ± 50

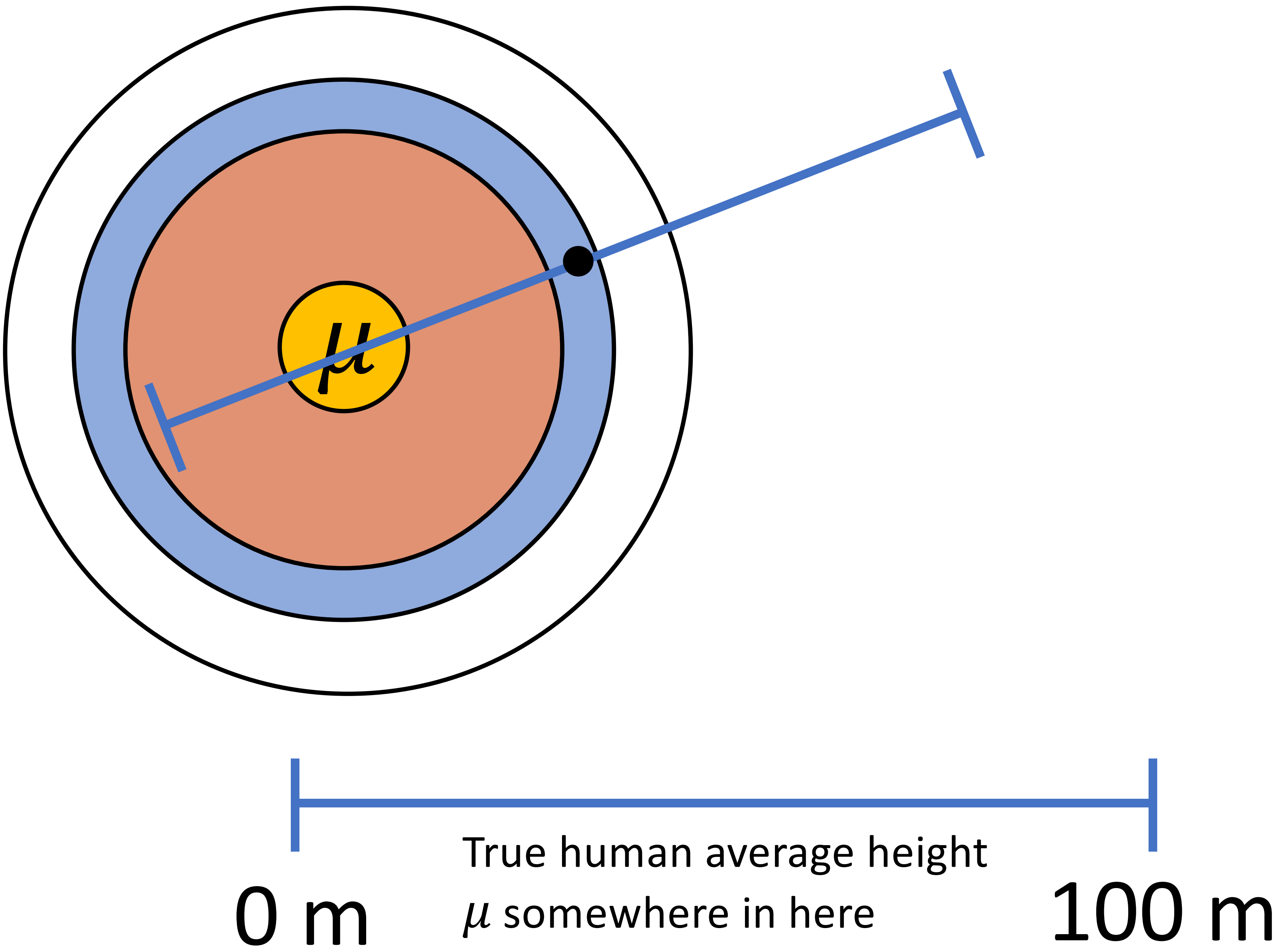
- Although this statement is technically true, it is scientifically useless. The range is so wide that it provides no meaningful information about the true average height. In other words, the statement offers certainty without precision, which does not help us make reliable inferences about the population.

Estimating with uncertainty, but with a degree of certainty

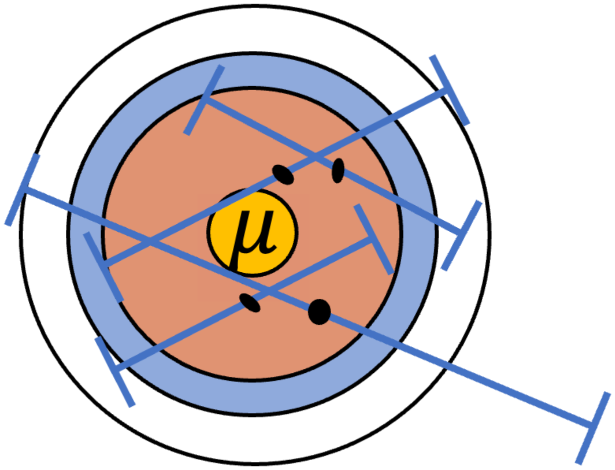
Example 1: The average height of all humans (i.e., the entire population) is between 0 m and 100 m. While this statement is technically true, it's useless because it provides no meaningful precision; obviously, all humans are taller than 0 m and shorter than 100 m, but this range doesn't help us estimate the true average precision (it's accurate though but "useless"). In other words, the statement provides confidence, but at the cost of extreme imprecision.



Estimating with uncertainty, but with a degree of certainty



Estimating with uncertainty, but with a degree of certainty

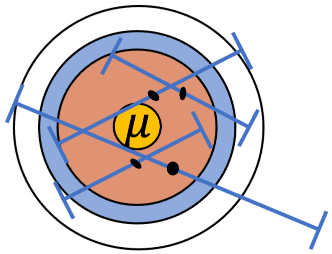


For any given sample mean, we can construct an interval that is likely to contain the true population parameter, with a specified level of confidence (say 95%).

This means that the procedure used to construct the interval has a 95% success rate in capturing the true parameter over repeated samples, not that there is a 95% probability that the true value lies within this particular interval (we will discuss this later).

So, if we were to take all possible sample means from a population, and build the confidence interval, 95% of these intervals would contain the true population value.

Estimating with uncertainty, but with a degree of certainty



For any given sample mean, we can construct an interval that is likely to contain the true population parameter, with a specified level of confidence (e.g., 95%).

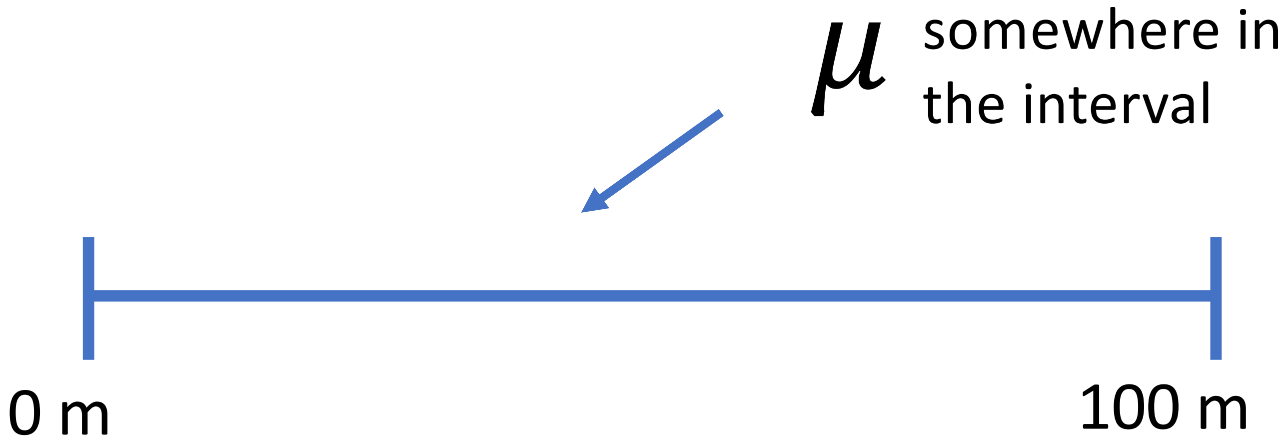
A **confidence interval** is centered on the observed sample mean, and its half-width (the margin of error) reflects the **uncertainty of the mean estimate**.

This **uncertainty** is determined by the **standard error**, which quantifies **sampling variability**, and by the chosen confidence level (in biology usually 95% or 99%). As such, the full width of the confidence interval represents the sampling uncertainty range.

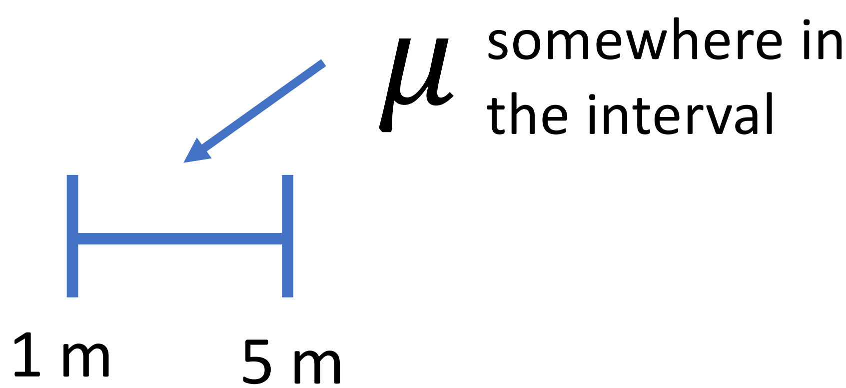
The **margin of error** is the maximum typical deviation we expect between a sample estimate and the true population value due to sampling variability alone, at a chosen confidence level.

Let’s take a break – 1 minute!

Estimating with uncertainty, but with a degree of certainty



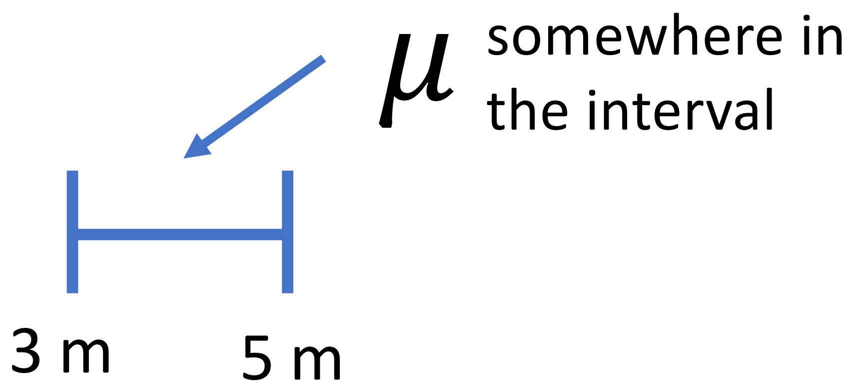
- **Example 2:** The average height of all adult humans is between 1 m and 5 m. While this interval is more useful than the first, it’s still not very helpful because we know that most adults are taller than 1 m and shorter than 5 m. Therefore, the true average will fall within this range, but it doesn't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



Estimating with uncertainty, but with a degree of certainty

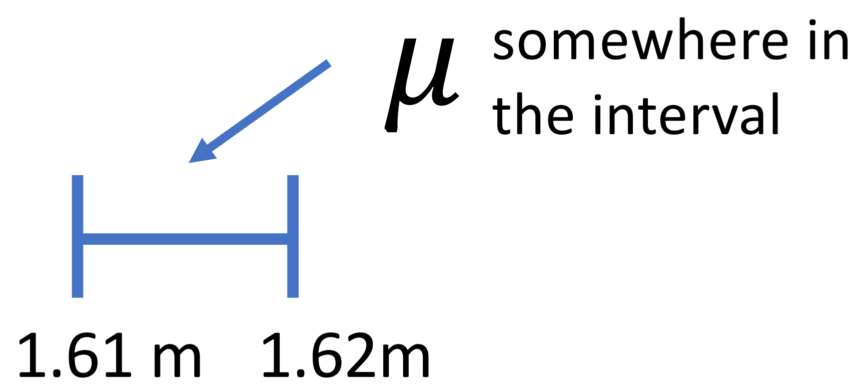
- **Example 2:** The average height of all adult humans is between 1 m and 5 m. While this interval is more useful than the first, it's still not very helpful because we know that most adults are taller than 1 m and shorter than 5 m. Therefore, the true average will fall within this range, but it doesn't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



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.



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 construction of confidence intervals for a population mean is based on the sampling distribution of the mean, the observed sample mean, and an estimate of variability from the data. This variability is summarized by the standard error, which quantifies how much sample means are expected to vary due to sampling alone.

Although we do not know the true population parameter with certainty, this framework allows us to construct an interval that provides a specified level of confidence (e.g., 95%) about ***where the true value is likely to lie***, because the interval is built from the known long-run behavior of sample means under repeated random sampling.

Importantly, confidence intervals are not limited to means: similar principles can be used to construct confidence intervals for other statistics, such as standard deviations, variances, medians, and many other parameters.

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

Confidence intervals are themselves estimates because they vary from sample to sample.

Estimating confidence intervals for a true population parameter (e.g., the mean) requires that we trust our sample estimates and, when possible, increase precision by using larger samples.

We can trust our estimates by using random sampling, which ensures accuracy by avoiding systematic bias. We can improve precision by increasing sample size, which reduces sampling error and narrows confidence intervals.

Estimating with uncertainty, but with a degree of certainty

Statistical populations with smaller variances lead to more precise estimates, but population variability is often not something researchers can directly control.

However, precision can sometimes be improved by defining more specific research questions; for example, estimating the average height of adult humans rather than the average height of all humans, which reduces variability by focusing on a more homogeneous population.

Estimating with uncertainty, but with a degree of certainty

The discussion of confidence intervals using human height was intended to build intuition about what “confidence” means in statistical terms.

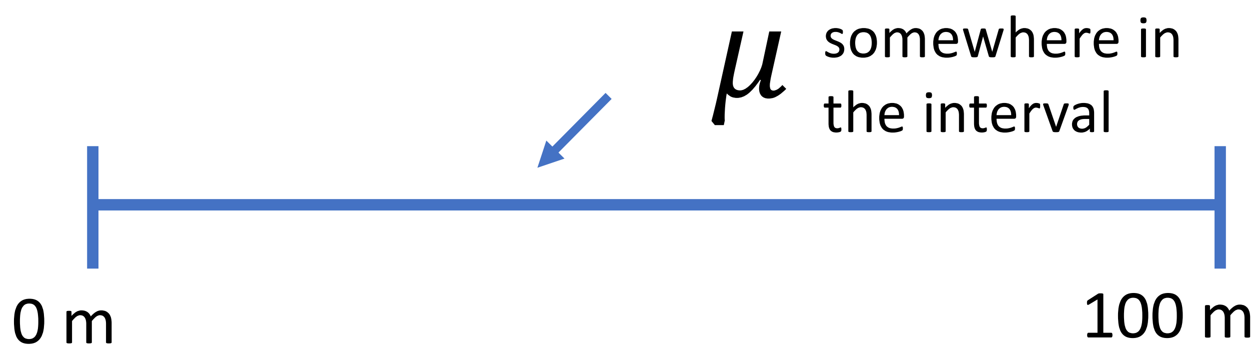
In real-world applications, however, such contrasts are often less straightforward because we rarely know the plausible range of values for the parameter of interest in advance.

As a result, confidence intervals are often constructed without a clear sense of what ranges are biologically meaningful, which can lead to imprecise or poorly interpreted results.

For this reason, biologists should invest effort in defining plausible parameter ranges and biological context when interpreting confidence intervals, rather than relying on them as purely technical outputs.

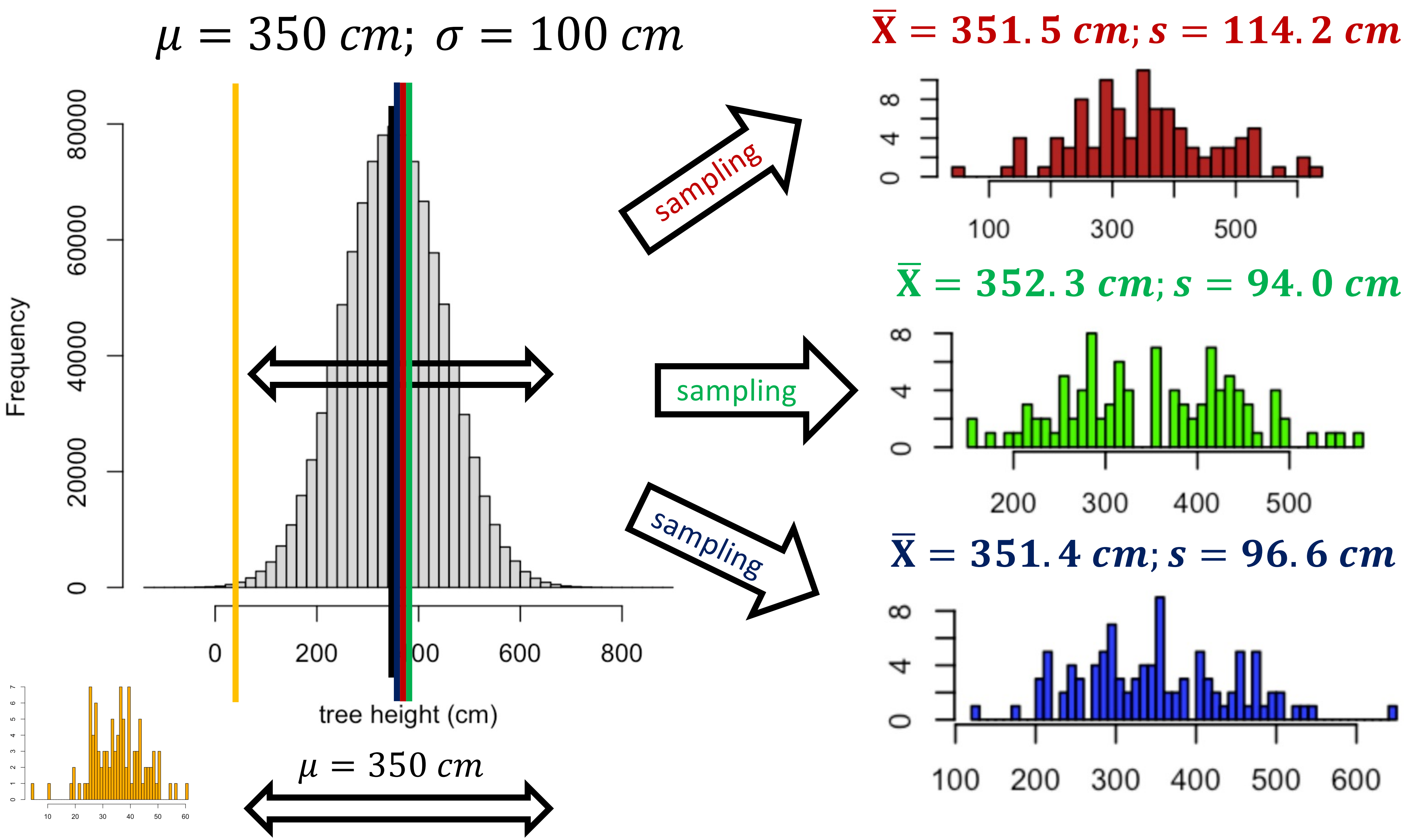
Estimating with uncertainty, but with a degree of certainty

How to build a confidence interval?



This confidence interval is built from a single observed sample, but its interpretation depends on the assumption that the sample was obtained randomly, ensuring that the sampling process is unbiased.

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

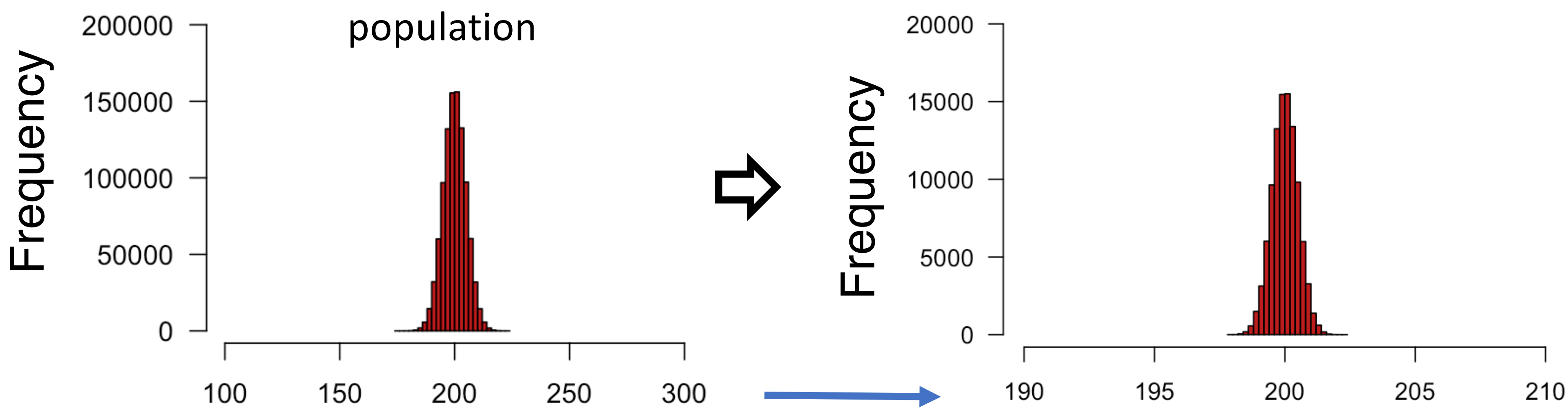


Uncertainty (samples means varying around the true population mean)

The variation within a sample, summarized by the sample standard deviation, provides information about how much sample means are expected to fluctuate around the true population mean—allowing us to estimate the typical uncertainty in our estimate.

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

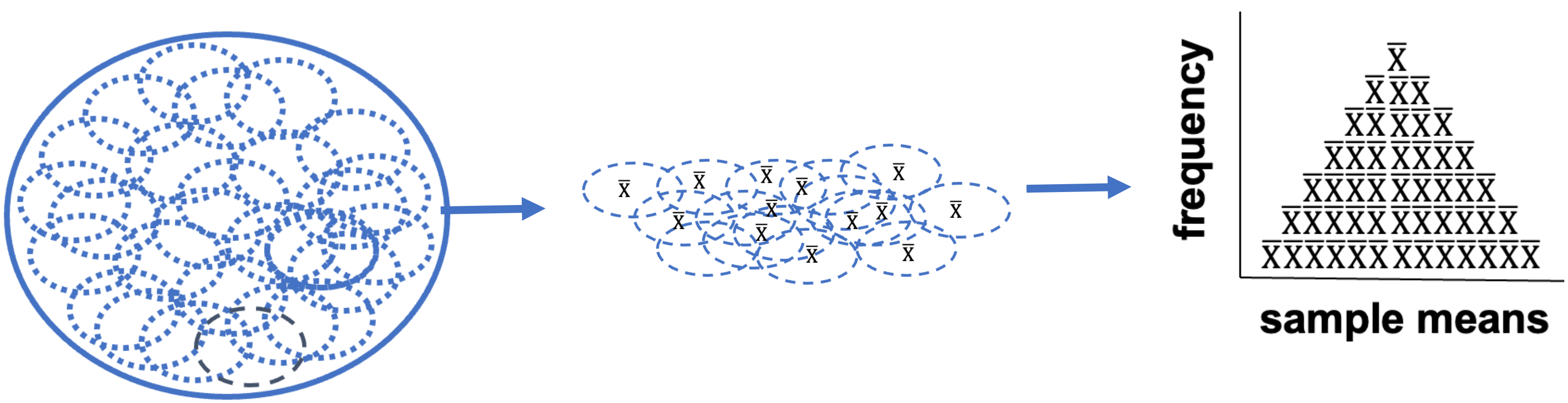
Sampling distribution of means



Variation among trees (small trees)

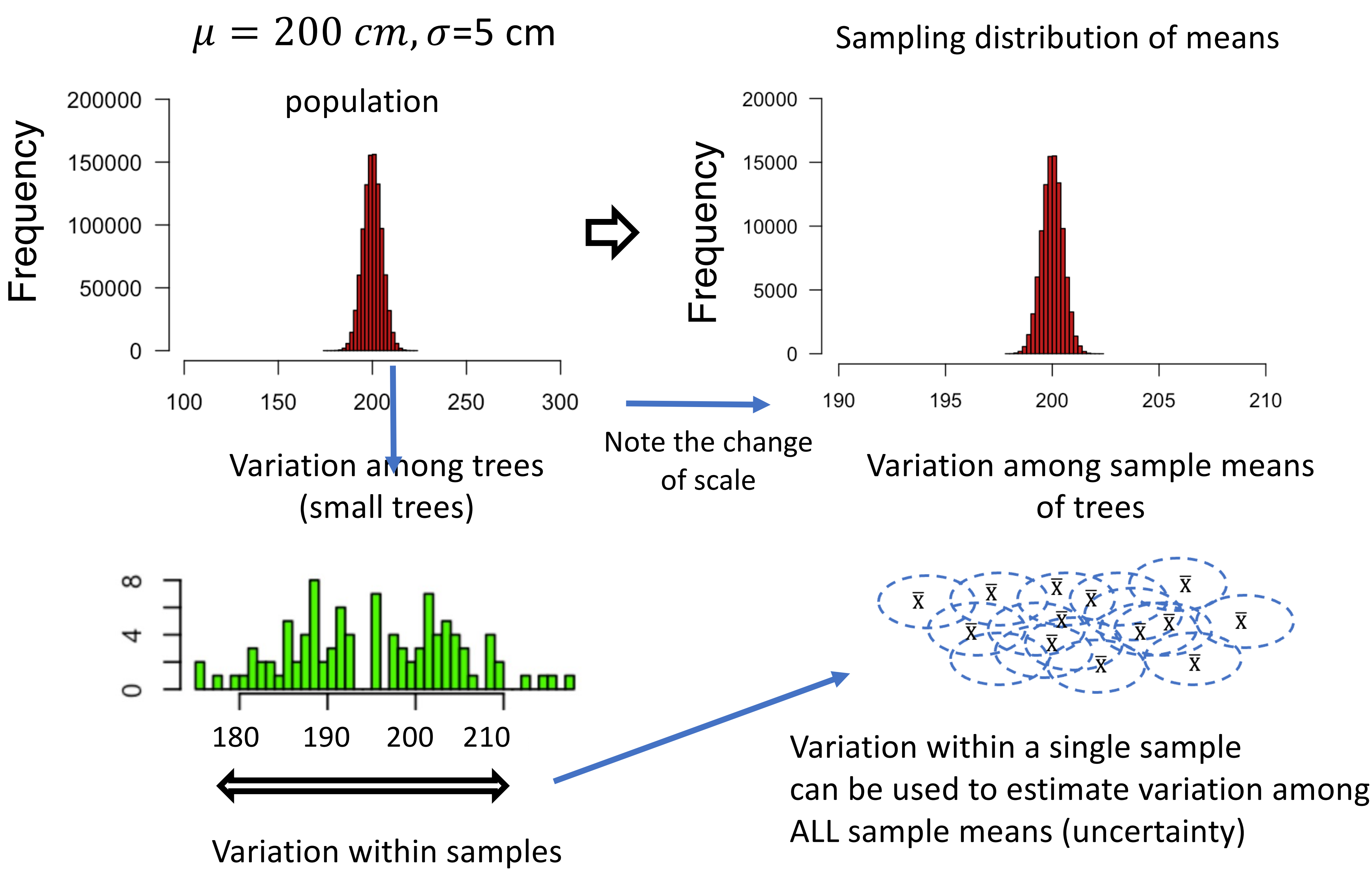
Note the change of scale

Variation among sample means of trees

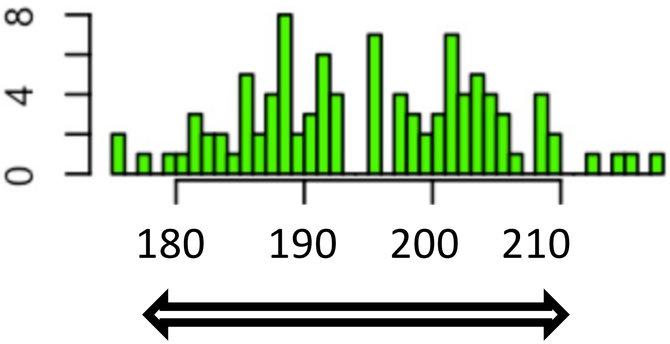


Population

The variation within a sample, summarized by the sample standard deviation, provides information about how much sample means are expected to fluctuate around the true population mean—allowing us to estimate the typical uncertainty in our estimate.

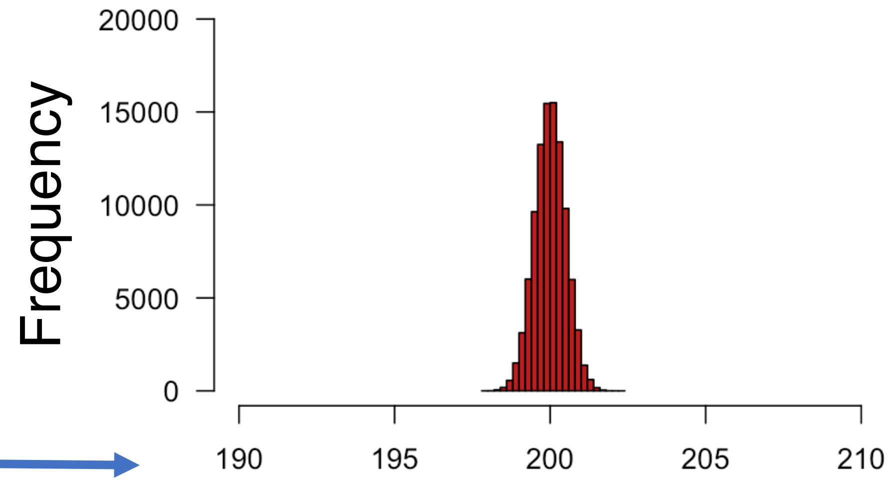


The variation within a sample, summarized by the sample standard deviation, provides information about how much sample means are expected to fluctuate around the true population mean—allowing us to estimate the typical uncertainty in our estimate.



Variation within samples

Sampling distribution of means



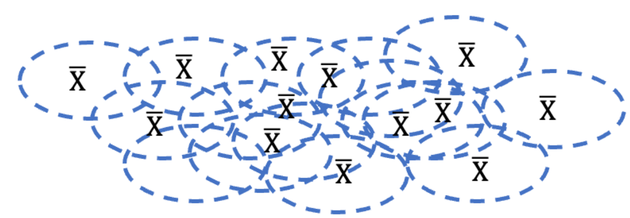
Note the change of scale

Variation among sample means of trees

s Standard deviation of the sample

n Sample size

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$



Variation within a single sample can be used to estimate variation among ALL sample means (uncertainty)

$s_{\bar{X}}$ Standard error of the mean = the standard deviation of the sampling distribution of the mean (uncertainty).

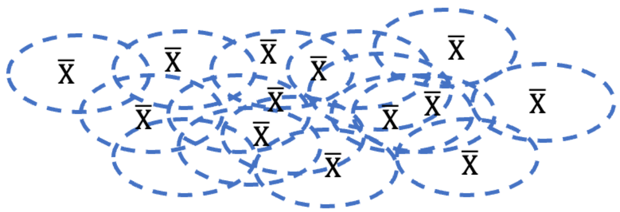
$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

THE STANDARD ERROR OF THE MEAN

Standard deviation (*s*): If you measure 25 students and the SD is 10 cm, it means a typical student's height differs from the sample average by about 10 cm.

If you repeatedly (randomly) sampled different groups of 25 students, their averages would vary slightly around the true population mean. And the same for any number of students that build a sample.

$$s_{\bar{X}} = \frac{10}{\sqrt{25}} = 2\text{cm}$$

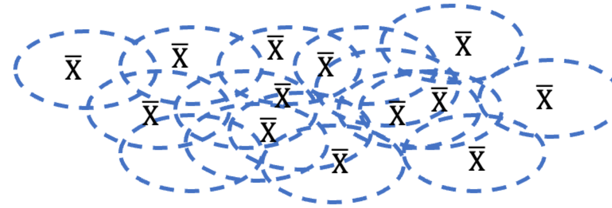


This 2 cm is the standard error and is a measure of uncertainty. It tells you: if you repeatedly took random samples of 25 students and calculated their average height, those averages would typically differ from the true population mean by about 2 cm.

UNCERTAINTY: $s_{\bar{X}}$ (the standard error) estimates how wrong our sample mean is expected to be, on average, from the true population mean—purely because we relied on a sample.

CONFIDENCE INTERVAL FOR THE MEAN

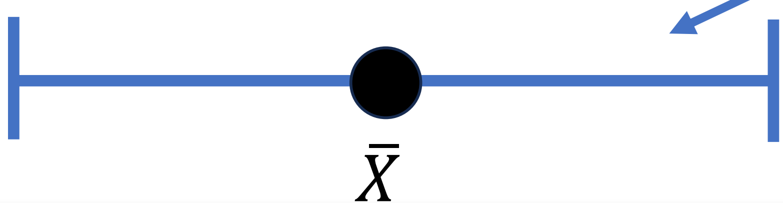
$$s_{\bar{X}} = \frac{10}{\sqrt{25}} = 2\text{cm}$$



UNCERTAINTY: $s_{\bar{X}}$ (the standard error) estimates how wrong our sample mean is expected to be, on average, from the true population mean—purely because we relied on a sample.

$$Confidence\ interval = \bar{X} \pm quantity * s_{\bar{X}}$$

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

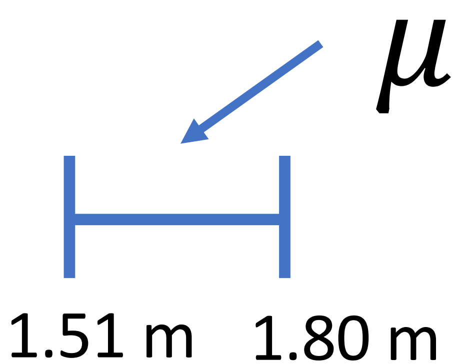


The quantity (i.e., which establishes the critical value or margin of error) varies with the confidence level we choose (e.g., 95% or 99%).

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.

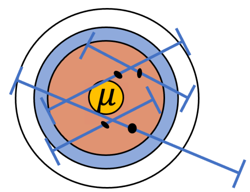
A larger confidence level (e.g., 95% or 99%) provides a more plausible range for the parameter. Values within the interval are considered more plausible, while those outside are less plausible, based solely on the sample data.

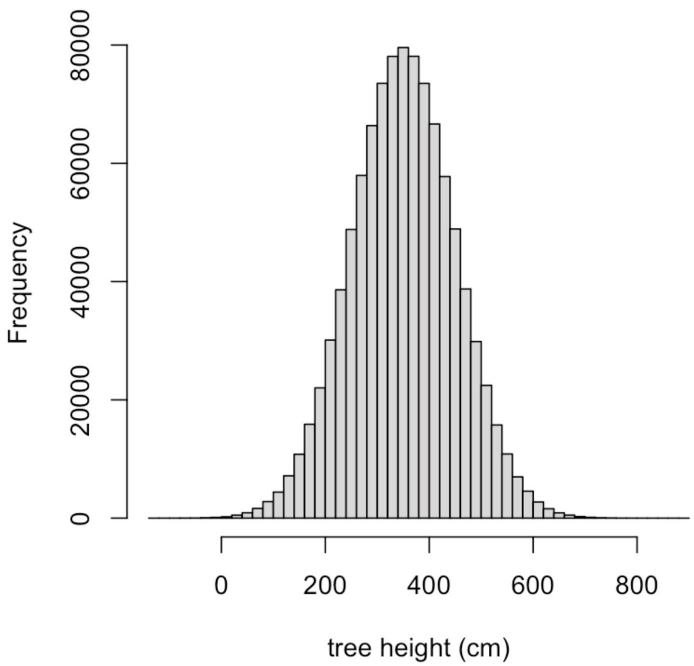


Very plausible (high confidence) that the population parameter is somewhere within The 95% confidence interval.

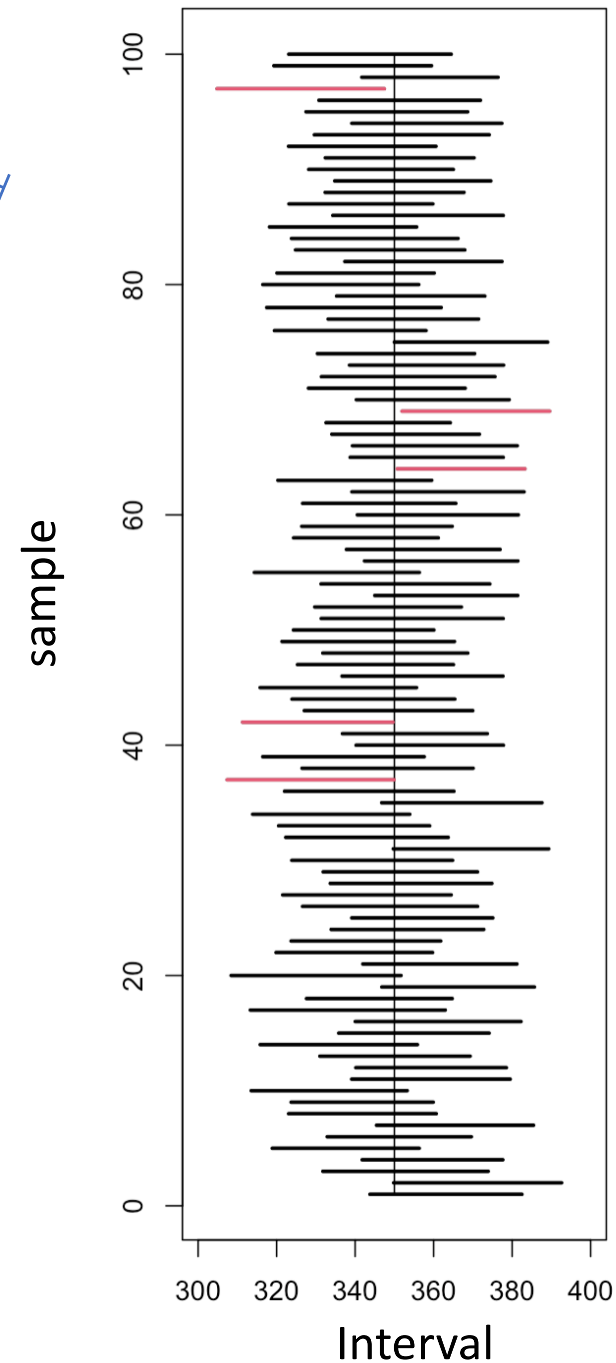
Imagine an interval referred to as "95% confidence interval":

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





For every possible sample build a confidence interval

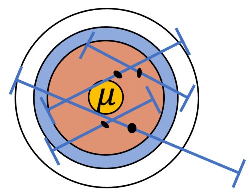


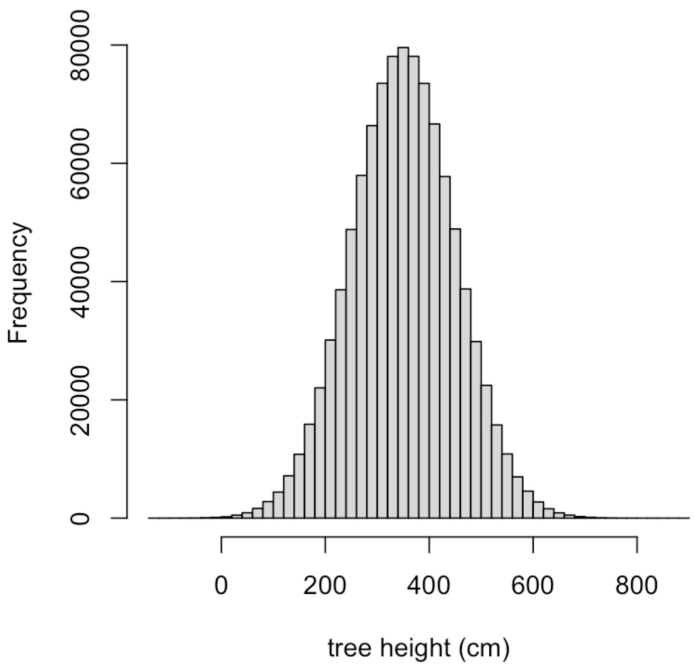
If sampling is random (unbiased) and the population distribution satisfies the required assumptions (e.g., is approximately normal), then 95 out of 100 confidence intervals constructed using this method will contain the true population parameter.

The remaining intervals—about 5%—will not contain the true parameter; these are shown in red.

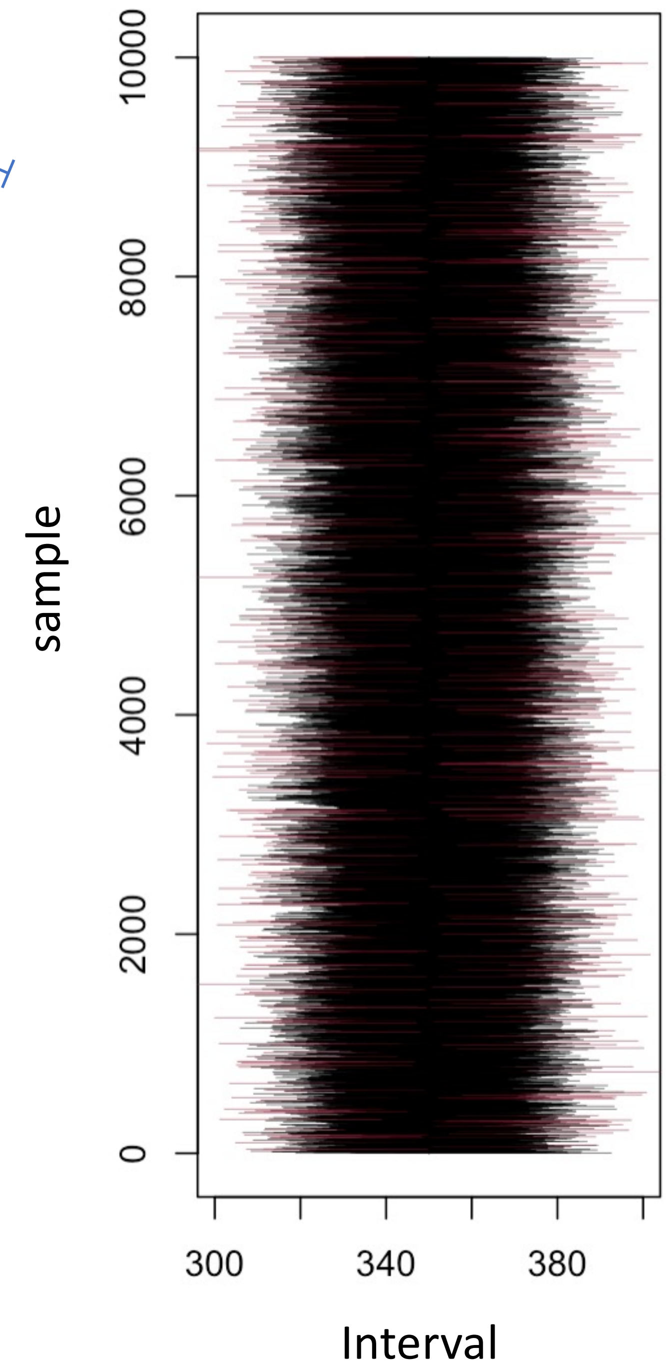
Imagine an interval referred as to “95% confidence interval”:

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





For every possible sample build a confidence interval



If sampling is random (unbiased) and the population distribution satisfies the required assumptions (e.g., is approximately normal), then 95 out of 100 confidence intervals constructed using this method will contain the true population parameter.

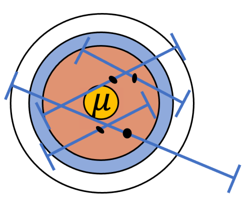
The remaining intervals—about 5%—**will not** contain the true parameter; these are shown in red.

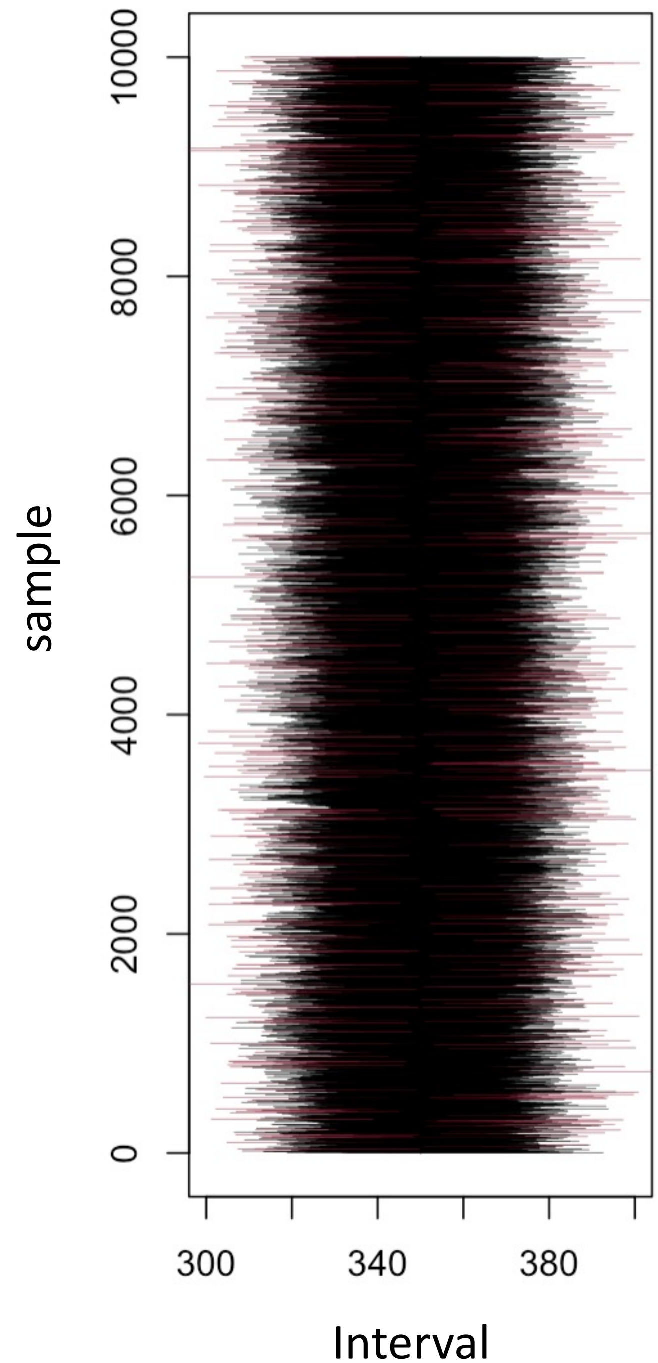
Very important!

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 this interval."

However, we cannot say, "There is a 95% probability that the true population mean lies within this interval." Once the interval is constructed, the true parameter either lies within it or it does not—there is no probability associated with this specific event (*this is a challenging concept - more or this later in the course*).

The 95% confidence level refers instead to the long-run performance of the interval-construction method across repeated samples.



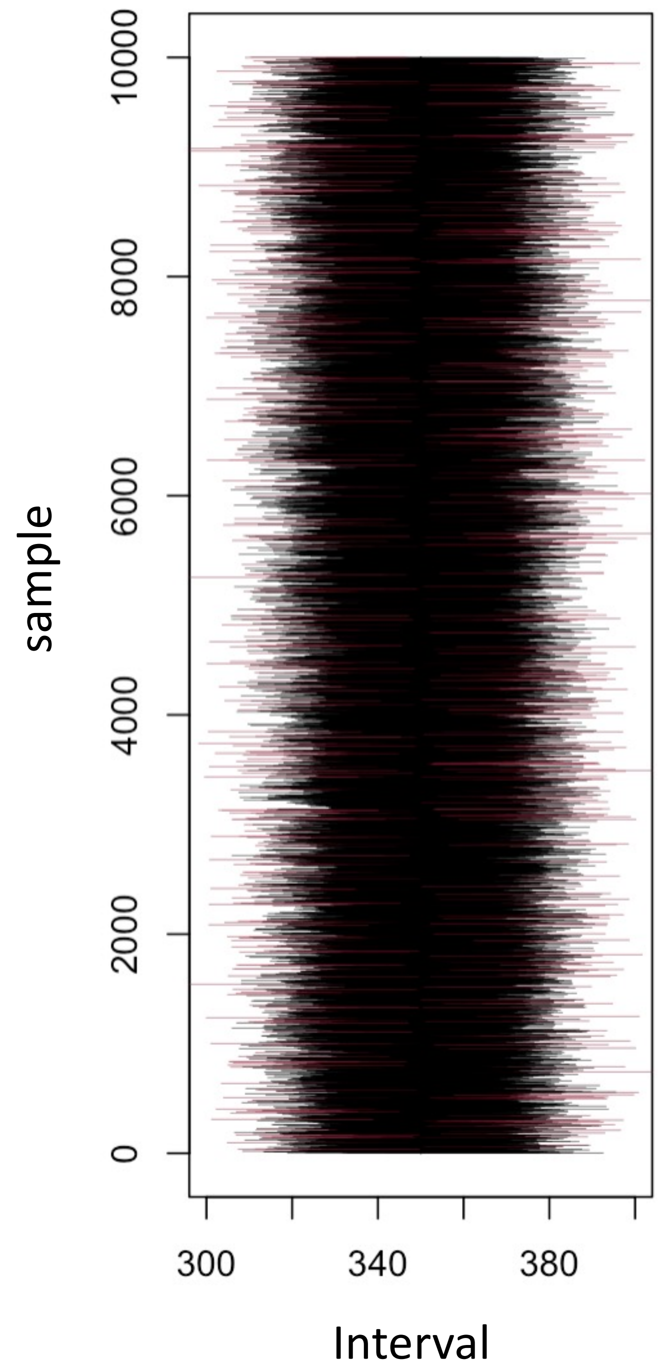


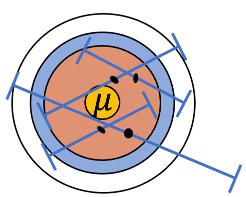
Very important!

Confidence intervals are based on the principle of repeated sampling, not on assigning a probability to whether a specific interval contains the true population parameter.

Under this framework, if we were to repeatedly draw samples from the same population and construct a confidence interval from each sample, a fixed proportion of those intervals (e.g., 95%) would contain the true population parameter.

The confidence interval itself varies from sample to sample, while the population parameter remains fixed.





LEVEL OF CONFIDENCE

Statistical inference is about balancing **confidence and precision:**

- Higher confidence → wider interval → less precision
- Lower confidence → narrower interval → more precision

As the confidence level approaches 100%, the critical value becomes extremely large. At exactly 100%, the interval width must go to infinity (or to the full logical range of the parameter).

A 100% confidence interval sacrifices all precision for certainty. It tells us nothing useful about where the parameter actually lies.

A CI of 20% looks very precise (narrow) but it is rarely correct.

Why 95% or 99%?

1. **Historical convention** — Fisher popularized the 5% significance level, which corresponds to 95% confidence.
2. **Practical compromise** — it balances reliability and precision reasonably well.
3. **Compare across studies.**

Let's take a break – 1 minute!

Many users of statistics struggle to fully understand confidence intervals because of the concept of “repeated sampling” (i.e., sampling variation) – cognitive discomfort

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
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 long-run likelihood that, through repeated sampling, the interval-construction method captures the true, fixed population parameter.

Common misconception: Confidence intervals represent the probability that a specific interval contains the true parameter.

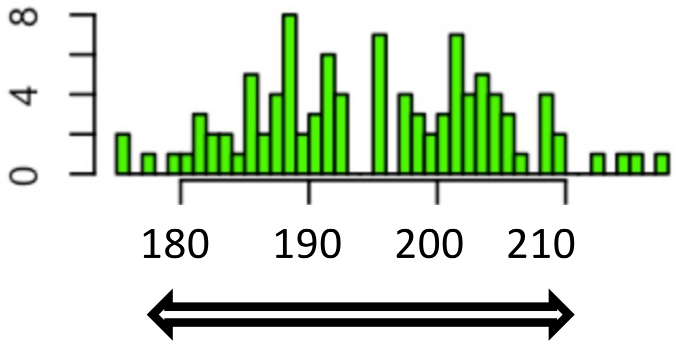
In reality, once an interval is constructed, the parameter either lies within it or it does not; the confidence level refers to the reliability of the method, not to a probability statement about that particular interval.

Assumptions underlying confidence intervals:

If sampling is random and the population's marginal distribution is approximately normal, then about 95% of all confidence intervals constructed using this method will contain the true population parameter (that is, across repeated samples, the interval-construction procedure has a 95% **coverage rate**).

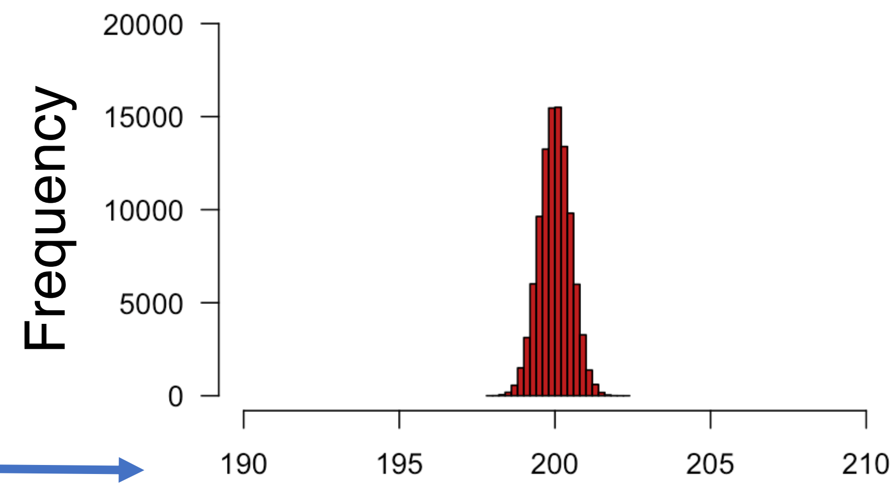
If the population distribution is not normal, the coverage rate may differ slightly from 95% (e.g., 92%, 97%, etc.). As we will discuss later (see Tutorial 6), this deviation depends on the distributional properties of the population, such as skewness or heavy tails.

Although we have not yet shown how to estimate a confidence interval, we already know that it depends on the standard error of the mean, a measure of how much sample means vary due to sampling.



Variation within samples

Sampling distribution of means



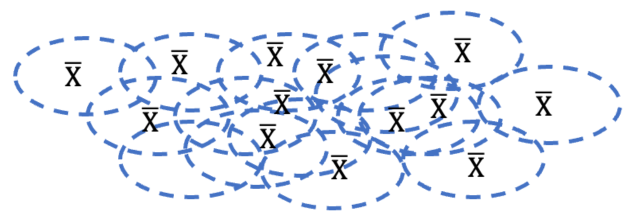
Note the change of scale

Variation among sample means of trees

s Standard deviation of the sample

n Sample size

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$



Variation within a single sample can be used to estimate variation among ALL sample means (uncertainty)

$s_{\bar{X}}$ Standard error of the mean = the standard deviation of the sampling distribution of the mean (uncertainty).

Although we have not yet shown how to estimate a confidence interval, we already know that it depends on the standard error of the mean, a measure of how much sample means vary due to sampling.

Sampling error refers to the random difference between a sample mean and the true population mean that arises because we observe only a finite sample.

The magnitude of this error is quantified by the **standard deviation of the sampling distribution of the mean, known as the standard error**, which measures how much sample means are expected to vary around the true population mean across repeated samples.

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

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

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

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

Although we have not yet shown how to estimate a confidence interval, we already know that it depends on the standard error of the mean, a measure of how much sample means vary due to sampling.

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

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

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

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

$\sqrt{\sum_{i=i}^{\infty} \frac{(\bar{Y}_i - \mu)^2}{\infty}}$ → Not observable: infinitely many samples

$\frac{\sigma}{\sqrt{n}}$ → closed-form result from probability theory.
variability of sample means depends only on:
the population standard deviation σ, and the sample size n.

Although we have not yet shown how to estimate a confidence interval, we already know that it depends on the standard error of the mean, a measure of how much sample means vary due to sampling.

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

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

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

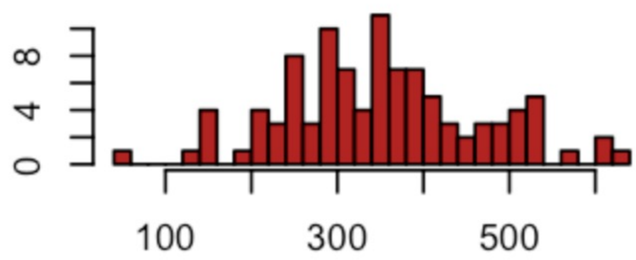
$\sigma_{\bar{Y}}$ = the standard deviation of the sampling distribution of means (standard error) ; σ = the standard deviation of the population.

The standard error of the mean, $\mathrm{SE}_{\bar{Y}}$, estimates the standard deviation of the sampling distribution of the mean; that is, how much sample means are expected to vary around the true population mean across repeated samples. Remarkably, this quantity can be estimated from a single sample (a principle we will examine in detail in Tutorial 5).

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

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

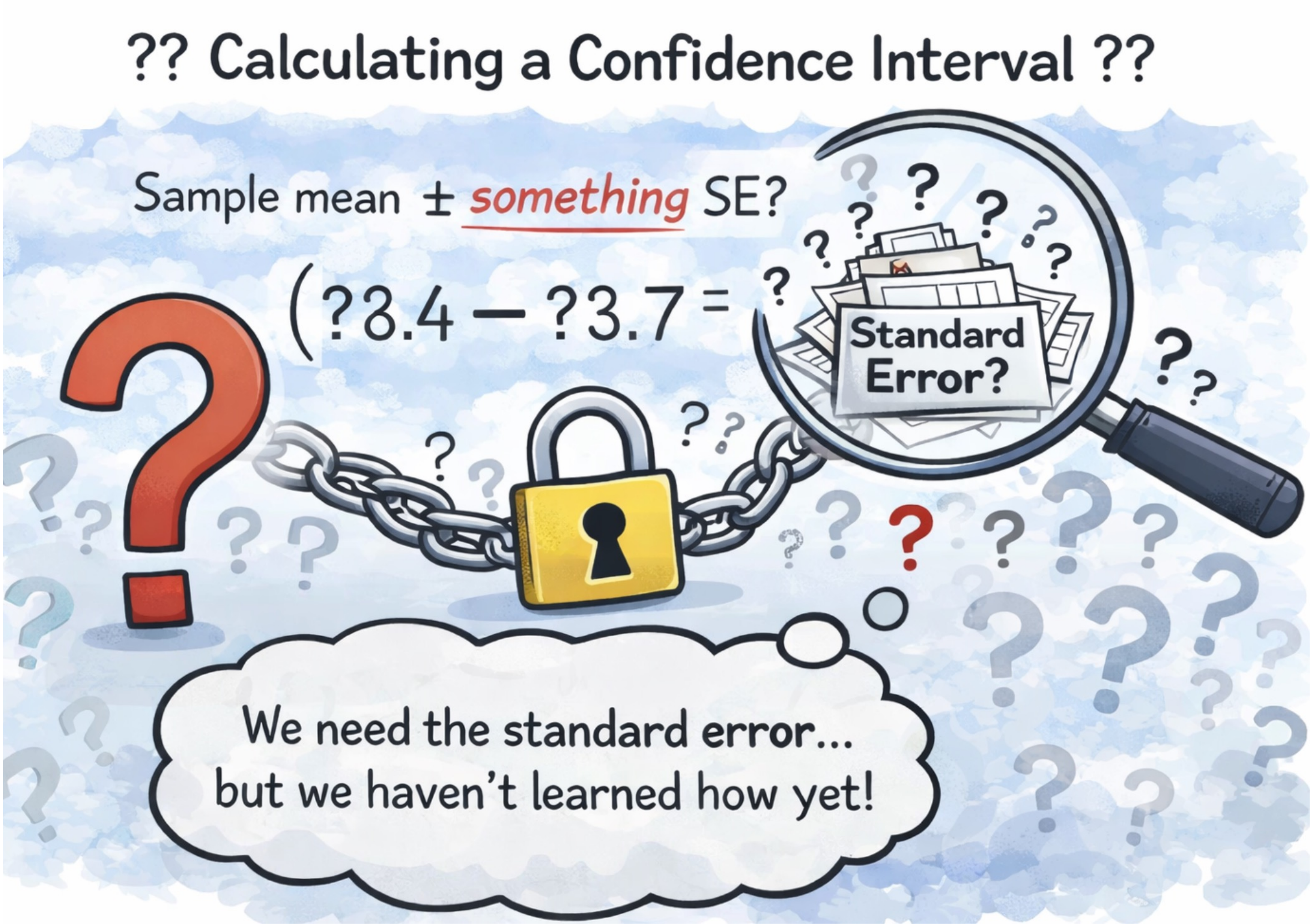
$\bar{\mathbf{X}} = \mathbf{351.5}\ \boldsymbol{cm};\ \boldsymbol{s} = \mathbf{114.2}\ \boldsymbol{cm}$



$\mathrm{SE}_{\bar{Y}}$ estimates the standard deviation of the sampling distribution of the mean (uncertainty).

Based on this sample, we estimate that sample means typically differ from the true population value by about 11.42 cm due to sampling variability.

Obviously, a different sample will generate a different estimate of this error.


?? Calculating a Confidence Interval ??
Sample mean ± something SE?
(?3.4 − ?3.7 =
Standard Error?
We need the standard error...
but we haven't learned how yet!

But for now

In general, a 95% confidence interval provides a useful summary of our uncertainty about the true population parameter.

If the interval is wide, uncertainty is high and the data are relatively uninformative about the parameter's value. If the interval is narrow, uncertainty is lower and the estimate is more precise.

But an important question remains: Is the interval useful? This is not purely a statistical question. It is a scientific one.

Whether a confidence interval is useful depends on the context of the problem and your substantive expertise. Does the range of plausible values allow you to make a meaningful scientific conclusion? Does it rule out values that would change your interpretation? Is the remaining uncertainty acceptable given the biological question?