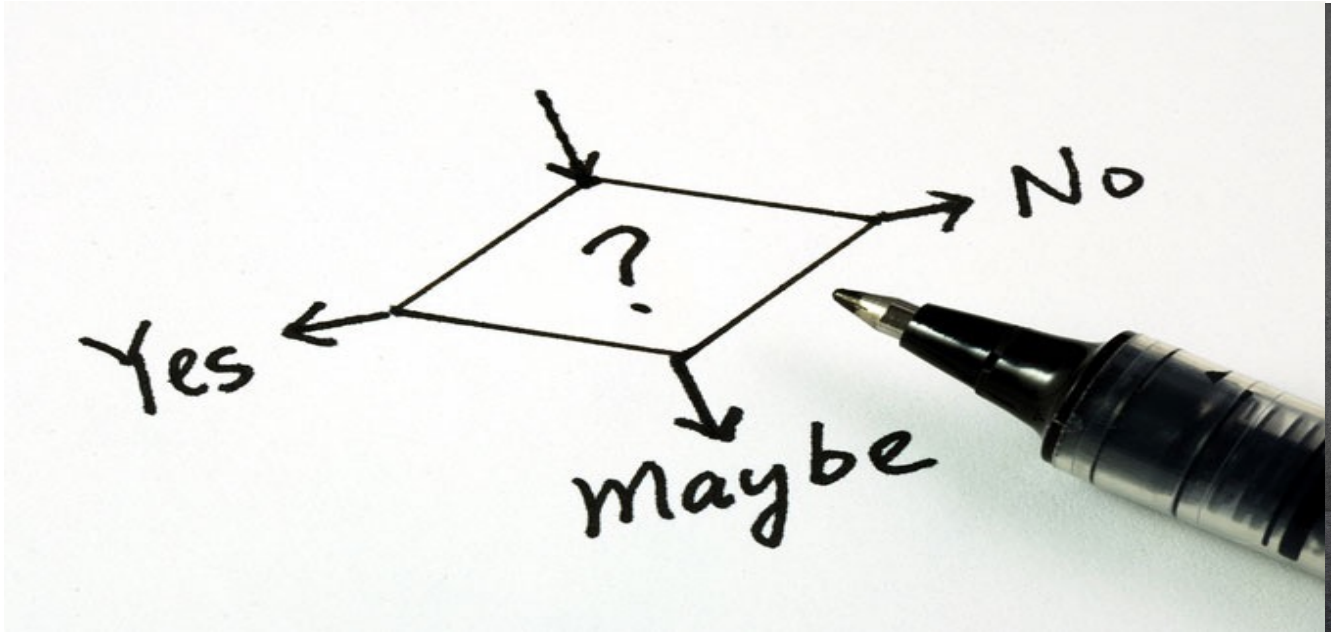
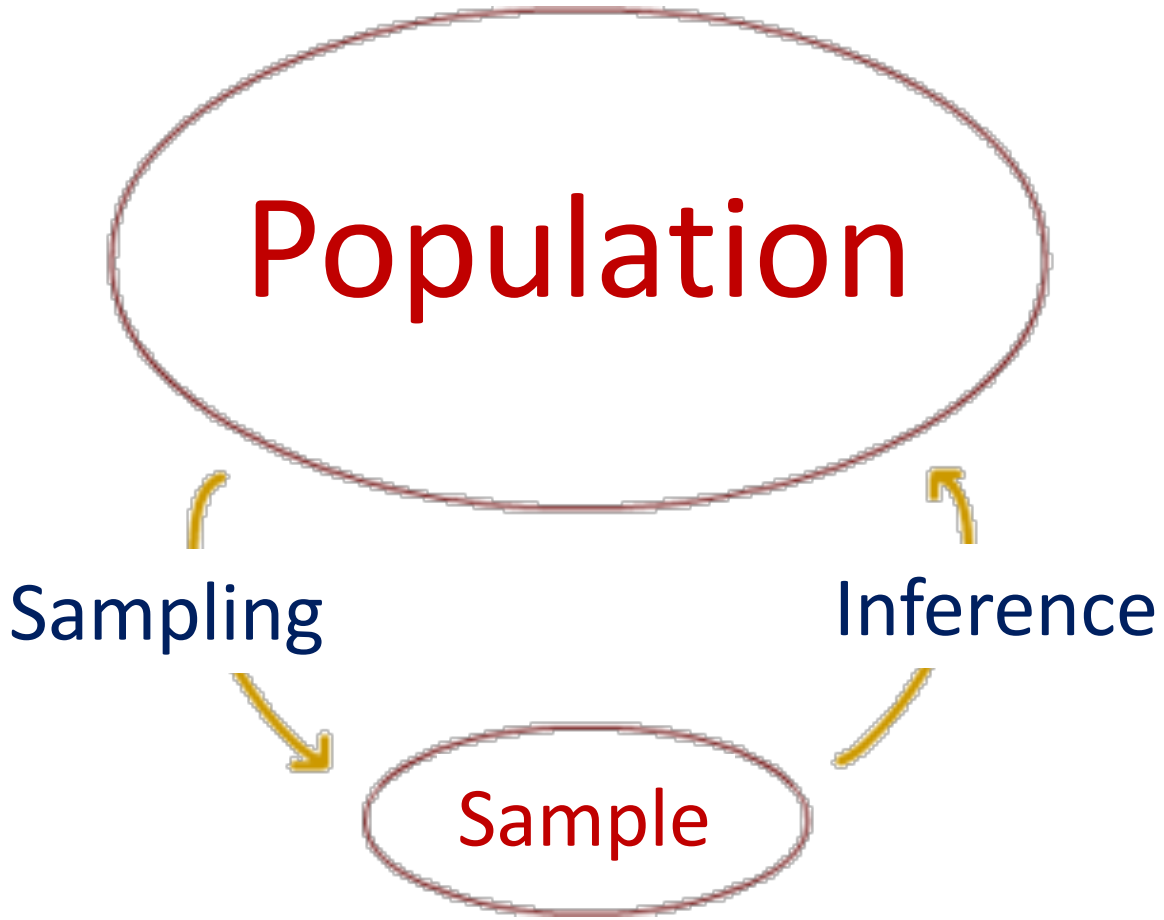
Lecture 7: estimating & making inferences with uncertainty – samples and biases

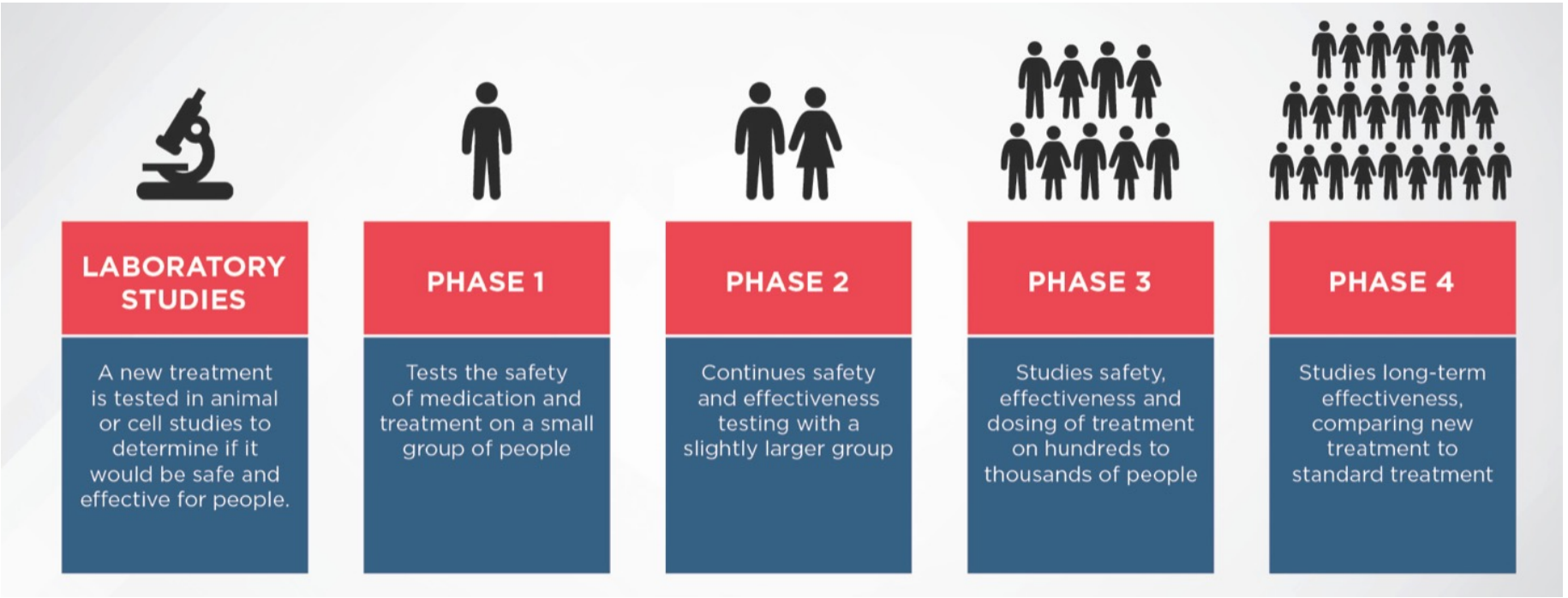
The science of aiding decision-making with incomplete information
(i.e., without certainty)



Inferential process



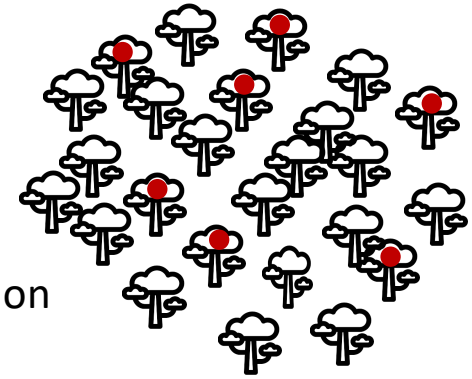
A good example of sampling:
Stages of Clinical Trials



https://news.sanfordhealth.org/coronavirus-disease-2019-covid-19/coronavirus-expert-q-a/sanford-research-vaccine-phases/

One of the most important goal of statistics is to infer an unknown quantity (e.g., height) of a population based on sample data!

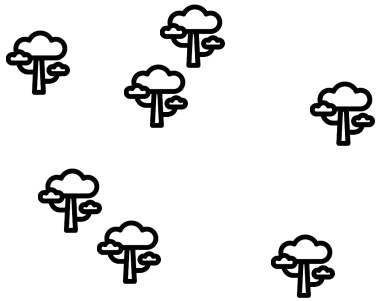
We want to know about a large number of trees



Population

Selected trees to measure height

We have these trees to work with



Sample

Inference

Population mean height
(here the parameter of interest, i.e., unknown quantity)

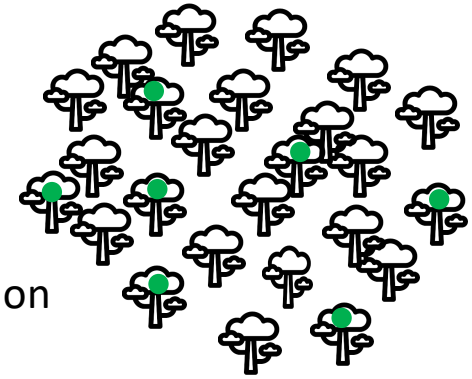
Sample mean height

A ***population*** is all the obervational units of interest, whereas a ***sample*** is a subset of observational units taken from the population.

Inspired by https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics

One of the most important goal of statistics is to infer an unknown quantity (e.g., height) of a population based on sample data!

We want to know about a large number of trees



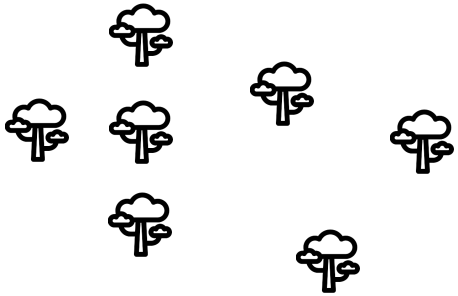
Population

Population mean height
(here the parameter of interest, i.e., unknown quantity)

Selected trees to measure height

Inference

We have these trees to work with

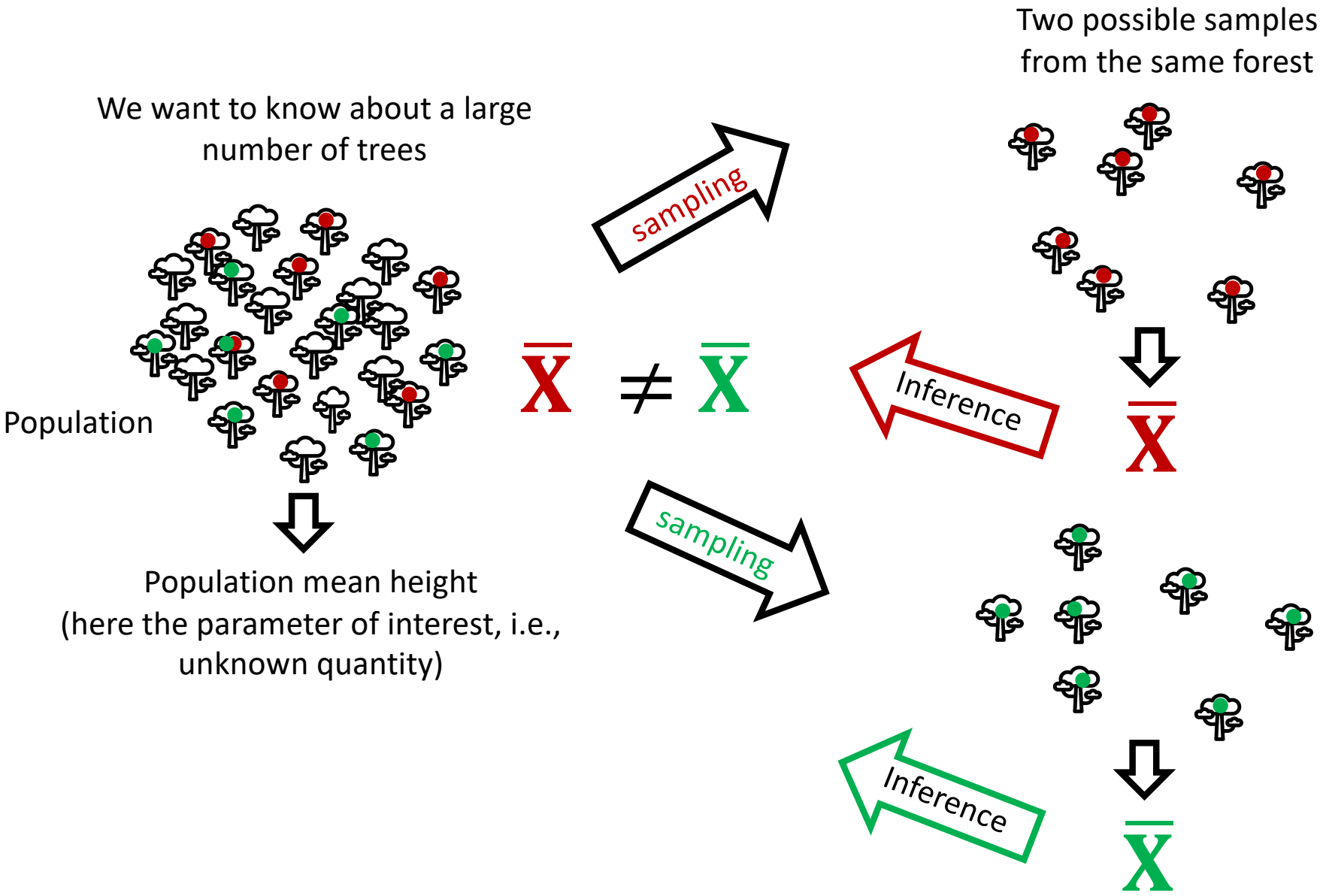


Sample

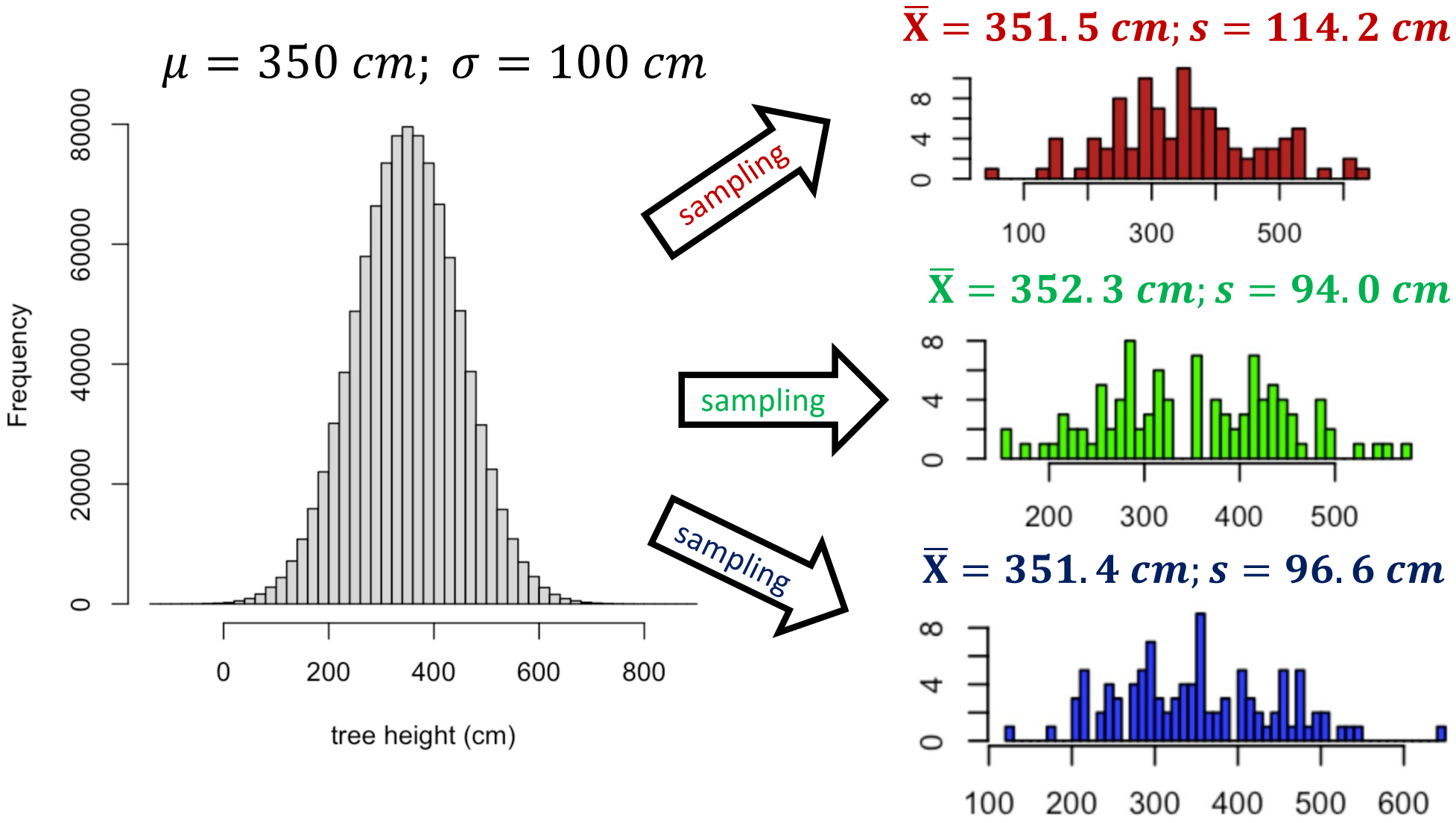
Sample mean height

Inspired by https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics

Sampling variation: The means of tree height from two or more samples of the same population will always differ from the true population parameter. Therefore, we estimate and make inferences while accounting for uncertainty.



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



Assume (hypothetically, i.e., for demonstration purposes only) a statistical population of tree heights in cm (1,000,000 trees) & 3 possible samples of 100 trees each.

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

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

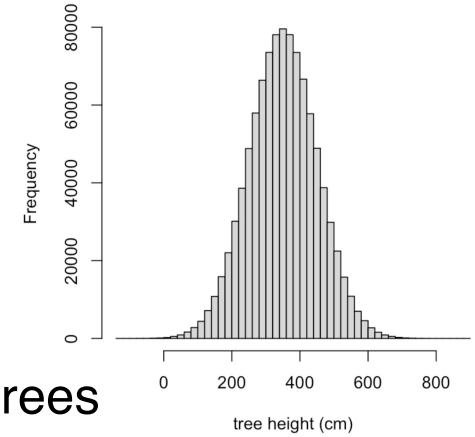
How many possible samples of 100 trees out of a population with 1000000 trees?

1076827236 2e+432 (zeros)

For comparison: the **human body** consists of about 37.2 trillion **cells**
(3.72e+13 zeros)

In most real-world studies, however, we typically take only one sample from the target statistical population.

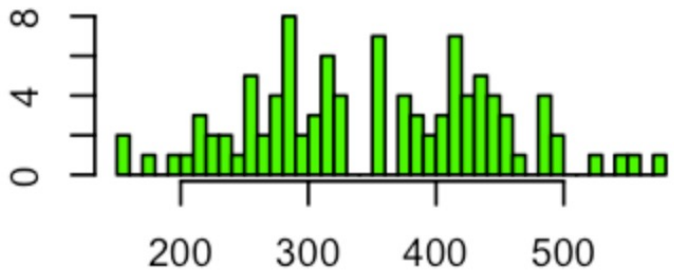
$\mu = 350\ cm;\ \sigma = 100\ cm$



1,000,000 trees

sampling

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



100 trees

Critical understanding:

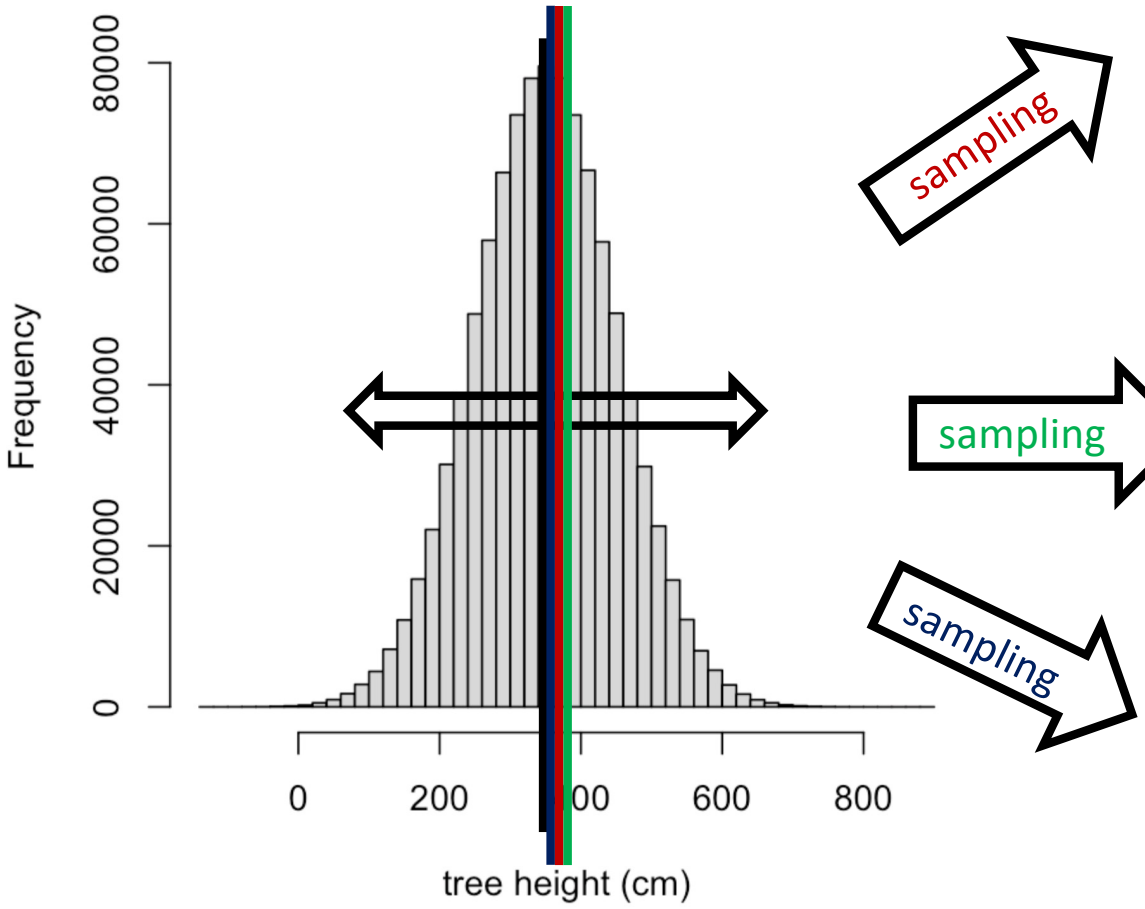
[1] Values of descriptive statistics from samples are never exactly the same as the corresponding population values due to sampling error.

[2] However, this does not mean that inferences drawn from samples are incorrect (we'll explore this in more detail later). Sample values serve as good approximations of the true population values.

[3] These approximations can range from good (when the sample statistic is close to the true population value) to poor (when it is far from the true value).

Sampling variation: Some samples are closer to the population mean, while others are further away, indicating that samples vary from one another.

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



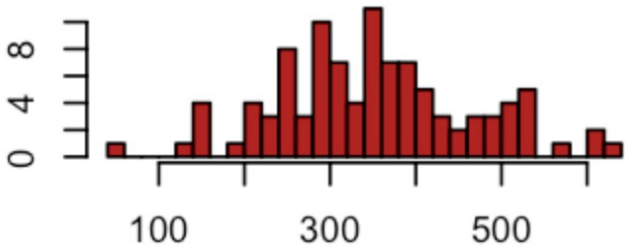
$\mu = 350\ cm$

sampling

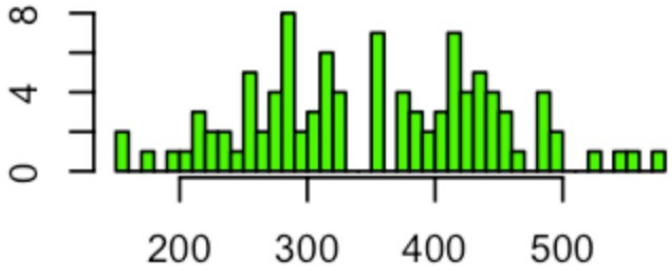
sampling

sampling

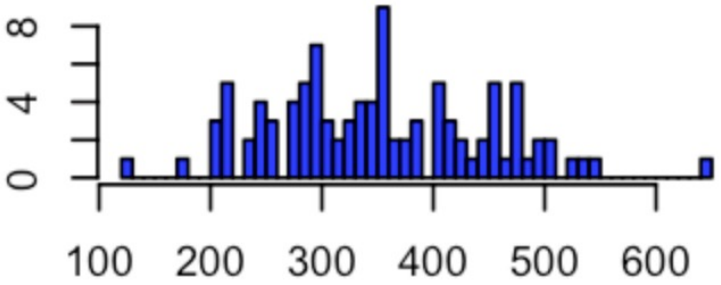
$\bar{X} = 351.5\ cm; s = 114.2\ cm$



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

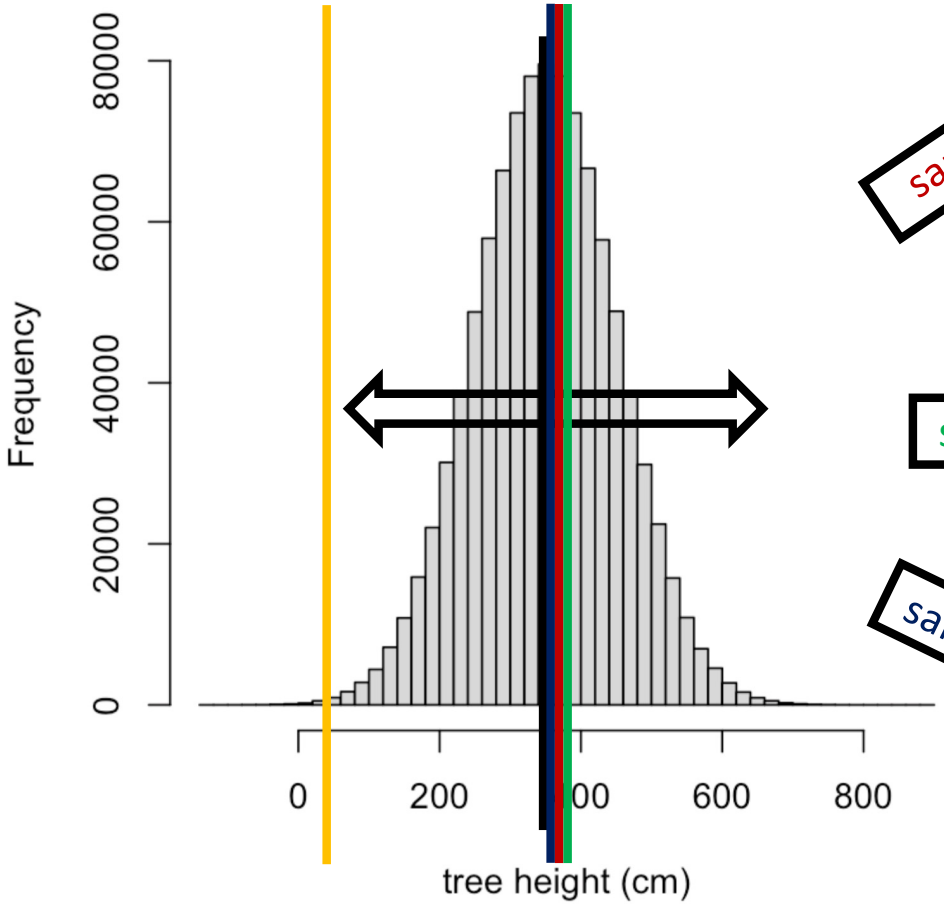


$\bar{X} = 351.4\ cm; s = 96.6\ cm$



Sampling variation: Some samples are closer to the population mean, while others are further away, indicating that samples vary from one another.

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

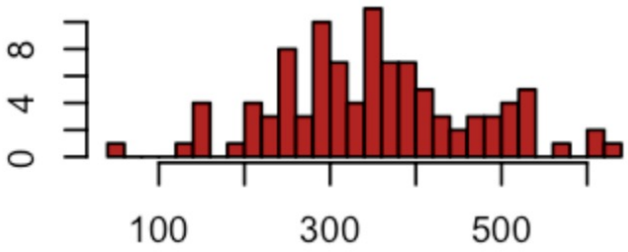


sampling

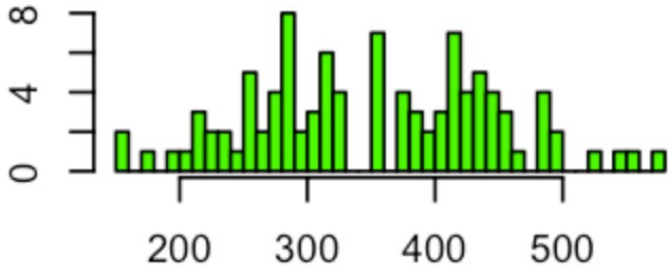
sampling

sampling

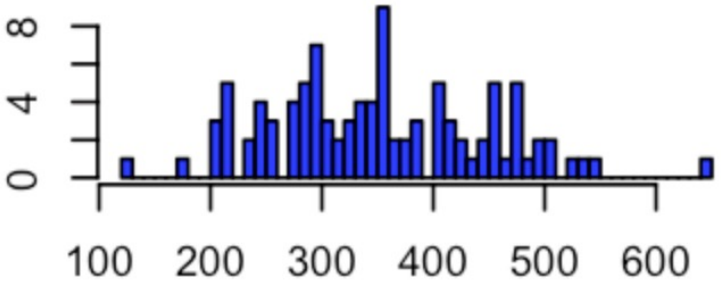
$\bar{X} = 351.5\ cm; s = 114.2\ cm$



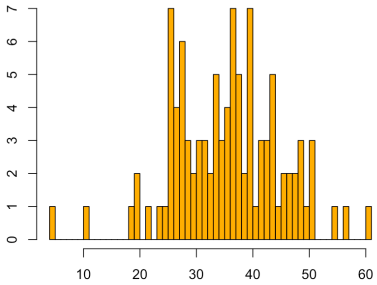
$\bar{X} = 352.3\ cm; s = 94.0\ cm$



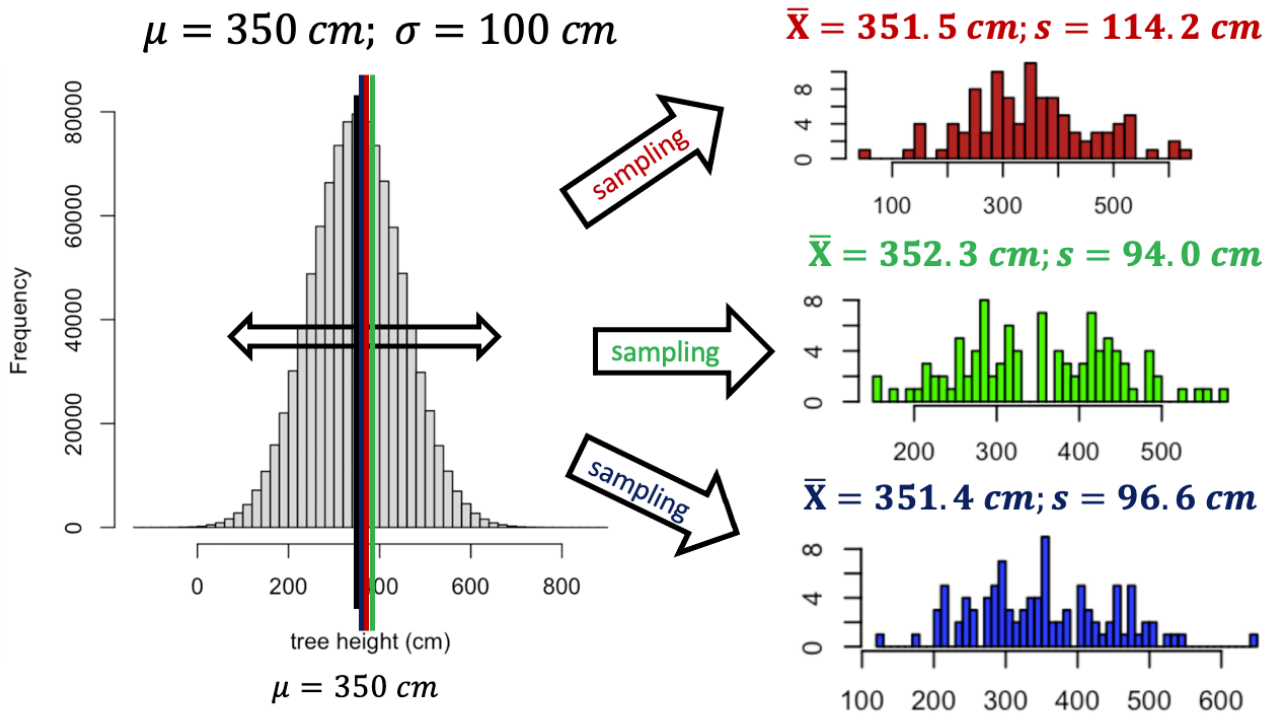
$\bar{X} = 351.4\ cm; s = 96.6\ cm$



$\mu = 350\ cm$



Sampling variation: Some samples are closer to the population mean, while others are further away, indicating that samples vary from one another.

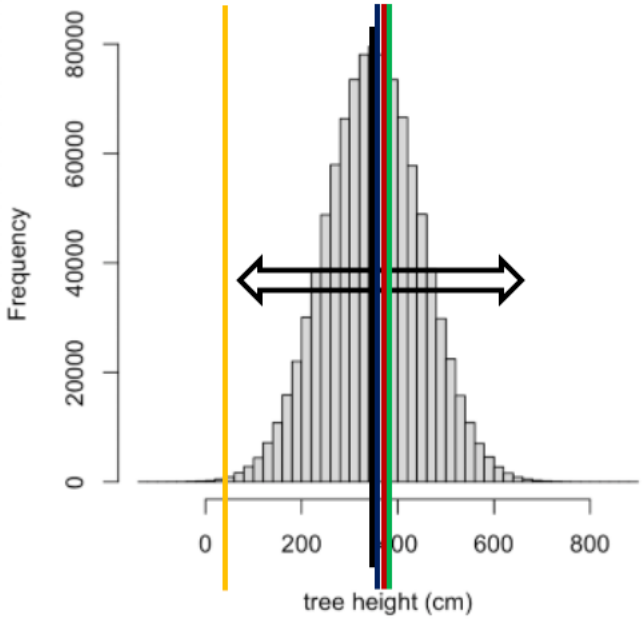


As we will explore later in the semester, an unbiased estimator (e.g., mean, standard deviation, median) is one that, on average across multiple repeated samples, equals the population parameter (i.e., the sampling error around the true mean is random).

Since we typically work with only one sample, we need to demonstrate mathematically that the estimator we are using is unbiased. We will discuss this in more detail later in the semester.

In most real-world studies, however, we typically take only one sample from the target statistical population

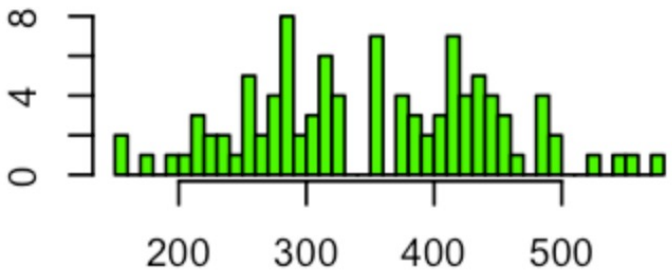
$\mu = 350\ cm;\ \sigma = 100\ cm$



$\mu = 350\ cm$

sampling

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



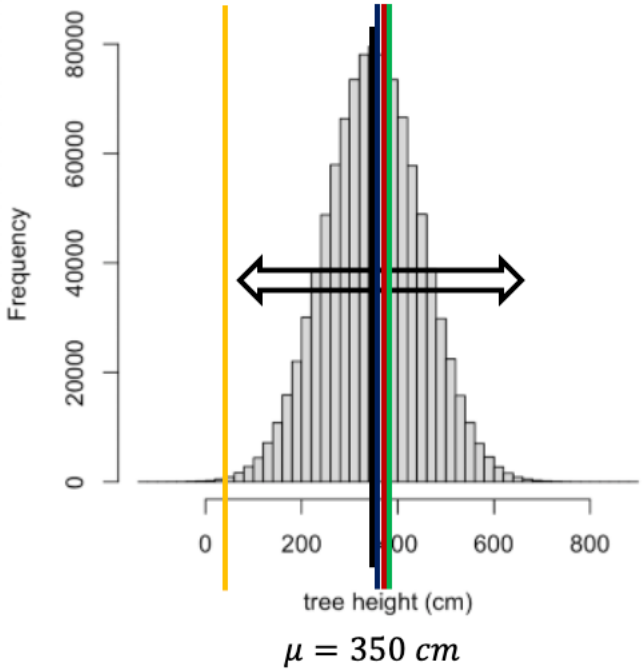
Critical understanding:

[3] Again, approximations can be either good (when the sample value is close to the true population value) or bad (when the sample value is far from the true value). Later, you will understand why we use terms like "close" and "far" to describe how samples relate to their populations.

[4] To build confidence, it would be helpful to estimate the expected margin of error.

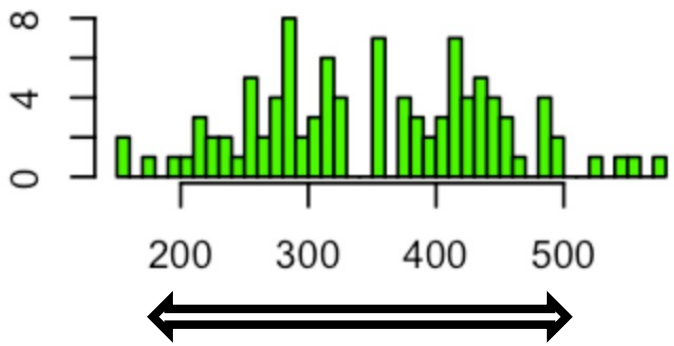
How wrong one could be in trusting their sample values to estimate the population value (i.e., parameter)?

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



sampling

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



Critical understanding:

[5] As we will see later, the variation among observations within a single sample (standard deviation) can help estimate how far sample means might be from the true population mean, providing an idea of potential error.

[3] Again, approximations can be either good (when the sample value is close to the true population value) or bad (when the sample value is far from the true value). Later, you will understand why we use terms like "close" and "far" to describe how samples relate to their populations.

[4] To build confidence, it would be helpful to estimate the expected margin of error.

Key concepts underlying statistics and statistical thinking

- Uncertainty: never being able to determine the true population parameter.
- The risk of error arises when decisions are based on estimates. If the estimate is close to the true value, the impact is minimal. However, when the estimate is far from the true value, decisions may lead to significant errors.
- Assessing uncertainty (error/risk) then becomes crucial.
- Sample variability - Answer may change with different sample data.
- Accuracy (how close to the true population value).

Key concepts:
Statistics is based on samples!

Sample quantities (e.g., mean, median, standard deviation, interquartile range) almost always vary from sample to sample, introducing some level of uncertainty.

Therefore, any estimates or inferences we make about population values inherently involve some degree of uncertainty, both in terms of variation between samples and the difference between sample estimates and the true population value.

As we will explore later, the variation among observations within a sample (standard deviation) can help estimate how far sample means may deviate from the true population mean, allowing us to assess uncertainty and associated risks.

Don’t forget to watch all the material in our WebBook.

Understanding sampling variation with dance.

https://www.youtube.com/watch?v=5fGu8hvdZ6s

Let’s take a break - 1 minute

Random sampling helps minimize sampling error (i.e., how close or far sample values are from the true population value for the statistic of interest) and reduces inferential bias.

The key requirement for the methods presented in this course (and in statistics in general) is that the data come from a random sample. A random sample must meet two essential 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 certain observational units in the target population have a lower or higher probability of being selected.

Before I forget!!!!

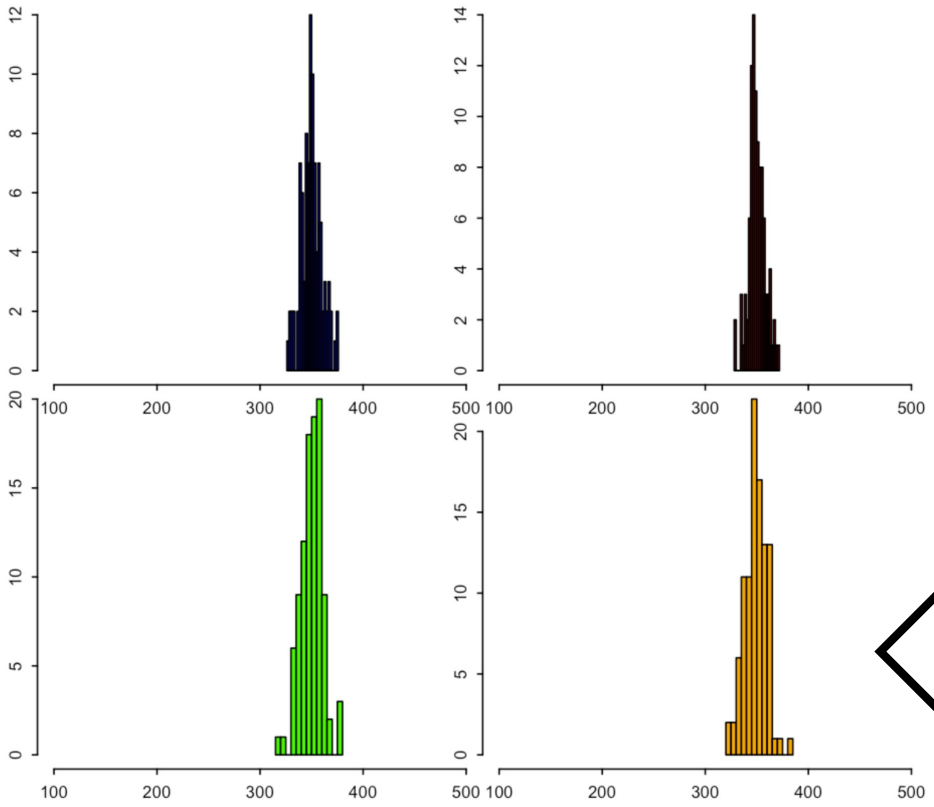
2) The selection of observational units in the population (e.g., individual tree) must be **independent**, **i.e.**, the selection of any unit (e.g., individual tree) of the population must not influence the selection of any other unit.

***i.e. = id est* (it is)**

Properties of samples - Precision and accuracy: the major goal of sampling is to increase accuracy & precision

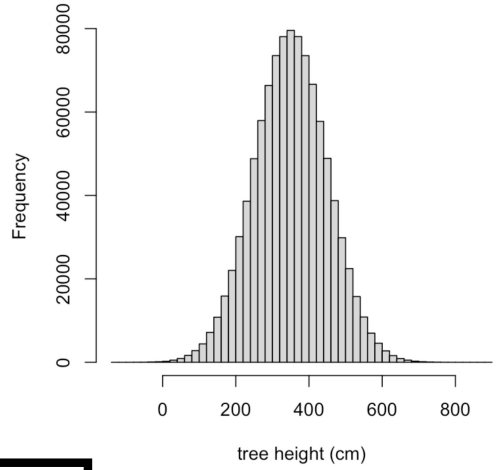
Precise

Accurate



Tree population

$\mu = 350\ cm$



sampling

1 000 000 trees (heights in cm)

4 random samples of 100 trees each

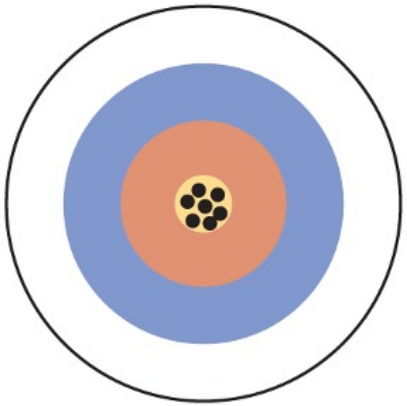
The average sample mean shows little variation from sample to sample (precision) and, on average, is similar to the population mean (accuracy).

These 4 samples are precise and accurate

Properties of samples - Precision and accuracy: the major goal of sampling is to increase accuracy & precision

Precise

Accurate



Imagine the bull's eye as the population parameter (in this case, the mean tree height), and the points as possible sample mean values for tree height (i.e., estimates).

Accurate = sample values (e.g., sample means) tend to be close to the true population value.

Precise = Sample values (e.g., sample means) tend to be similar to each other, regardless of whether they are close to or far from the true population value.

Properties of samples - Precision and accuracy: the major goal of sampling is to increase accuracy & precision

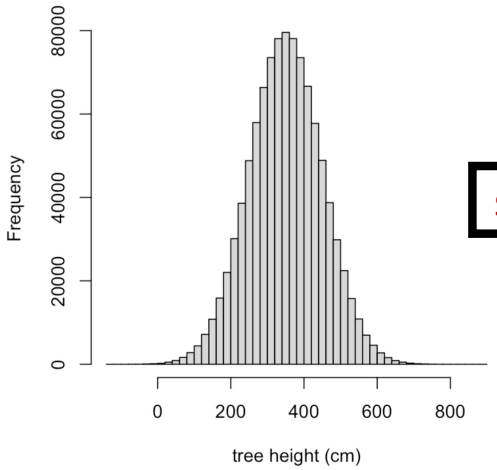
Accurate

Imprecise

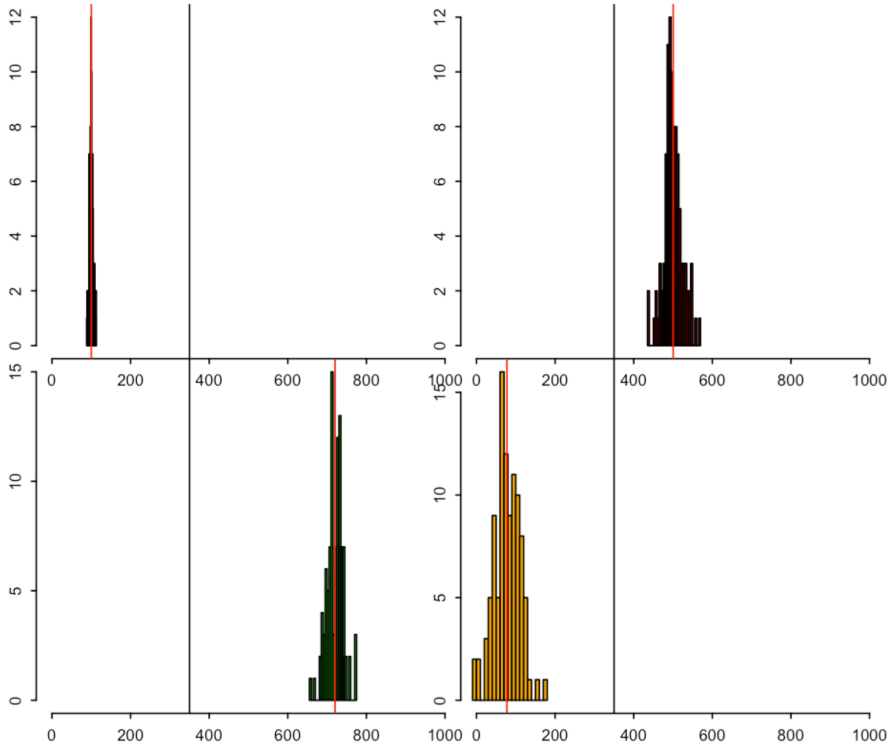
Black line = Population mean
Red line = sample mean

Tree population

$\mu = 350\ cm$



sampling



1 000 000 trees
(heights in cm)

4 random samples
of 100 trees each

The average sample mean varies considerably from sample to sample (imprecise), but on average, it is close to the population mean (accurate). In other words, the average of these four samples is very close to the true population value.

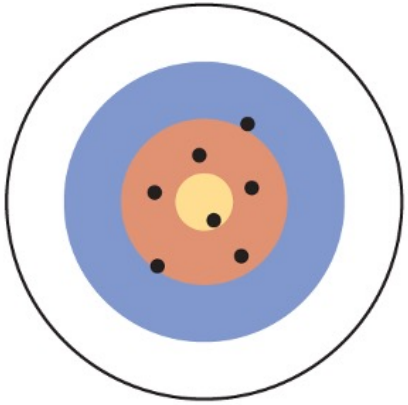
These 4 samples are imprecise but accurate

Properties of samples - Precision and accuracy: the major goal of sampling is to increase accuracy & precision

Imprecise

Accurate

Imagine the bull's eye as the population parameter (in this case, the mean tree height), and the points as possible sample mean values for tree height (i.e., estimates).



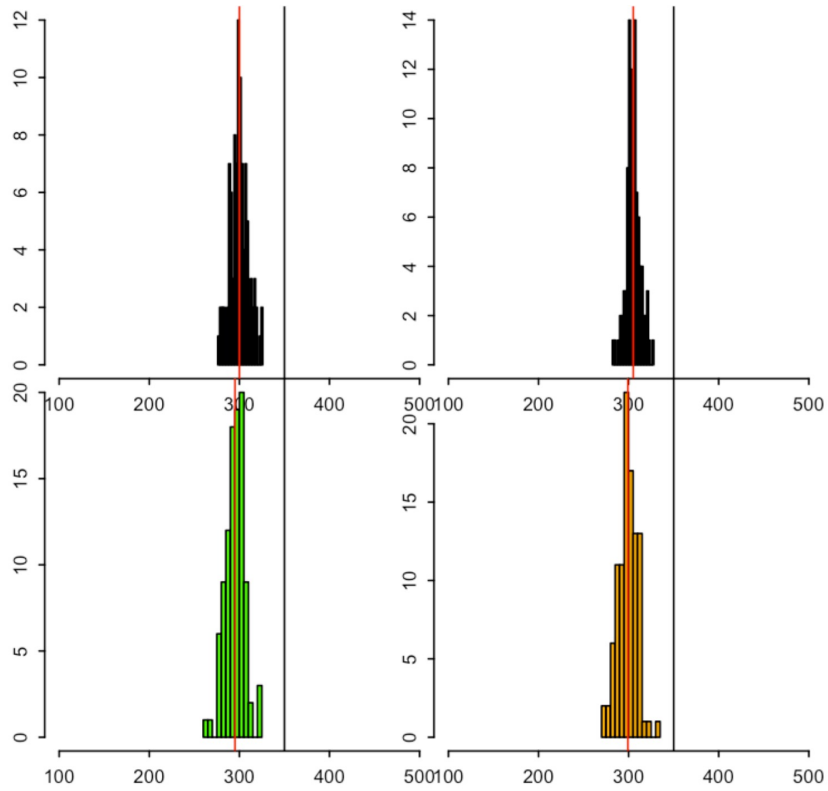
Accurate = sample values (e.g., sample means) tend to be close to the true population value.

Imprecise: Sample values (e.g., sample means) tend to vary widely from each other, regardless of whether they are close to or far from the population value.

Properties of samples - Precision and accuracy: the major goal of sampling is to increase accuracy & precision

Precise

Inaccurate

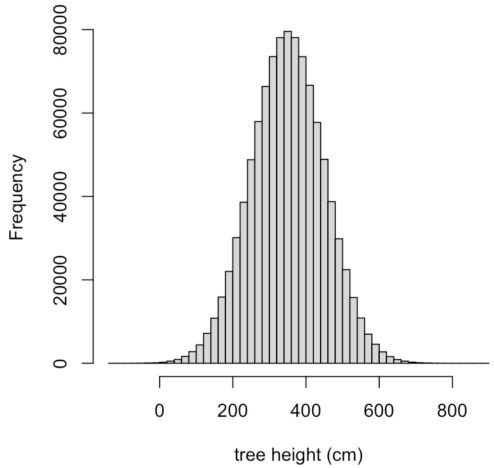


4 random samples of 100 trees each

Black line = Population mean
Red line = sample mean

sampling

$\mu = 350\ cm$



1 000 000 trees (heights in cm)

The average sample mean varies very little from sample to sample (precise), but on average, it differs from the population mean (inaccurate).

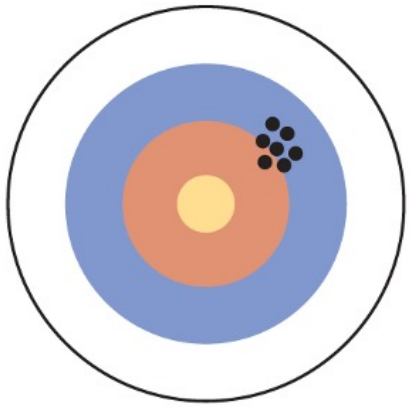
These 4 samples are precise but inaccurate

Properties of samples - Precision and accuracy: the major goal of sampling is to increase accuracy & precision

Precise

Imagine the bull's eye as the population parameter (in this case, the mean tree height), and the points as possible sample mean values for tree height (i.e., estimates).

Inaccurate



Inaccurate: Sample values (e.g., sample means) tend to be far from the true population value.

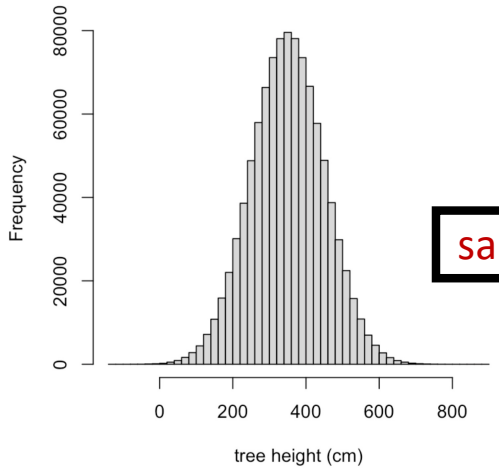
Precise = Sample values (e.g., sample means) tend to be similar to each other, regardless of whether they are close to or far from the true population value.

Properties of samples - Precision and accuracy: the major goal of sampling is to increase accuracy & precision

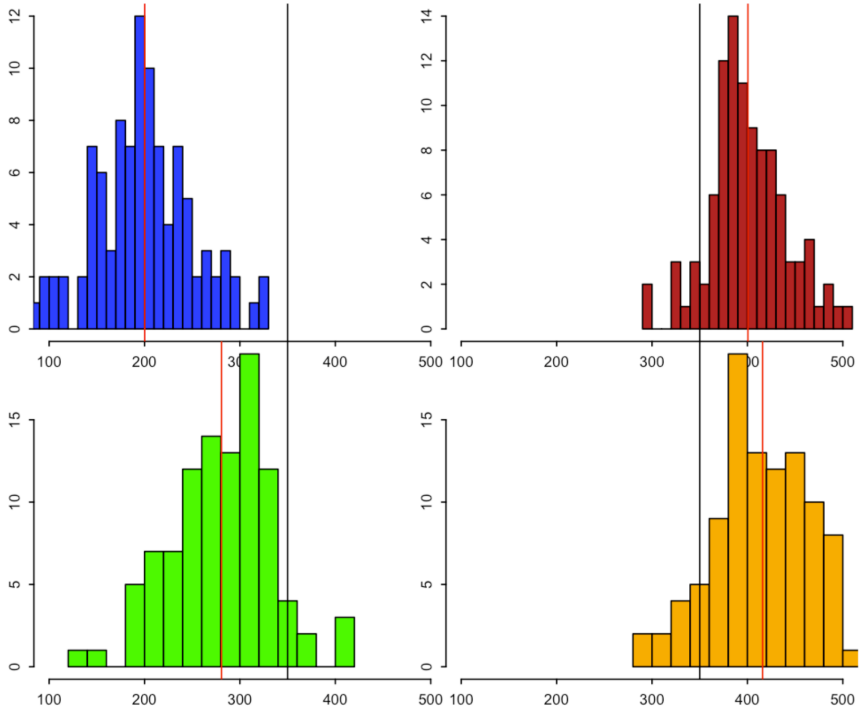
Imprecise

Black line = Population mean
Red line = sample mean

$\mu = 350\ cm$



sampling



1 000 000 trees
(heights in cm)

Inaccurate

4 random samples
of 100 trees each

The average sample mean varies significantly from sample to sample (imprecise) and, on average, differs from the population mean (inaccurate).

These 4 samples are imprecise and inaccurate

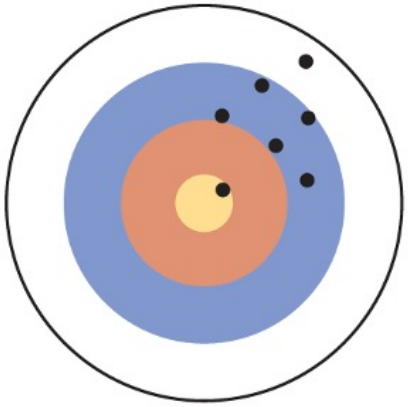
Properties of samples - Precision and accuracy: the major goal of sampling is to increase accuracy & precision

Imprecise

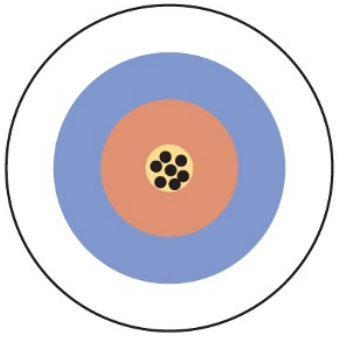
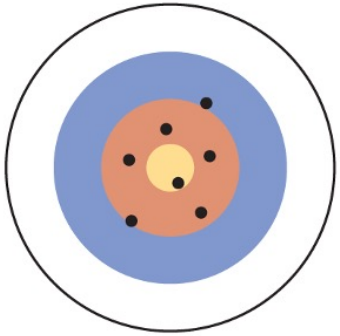
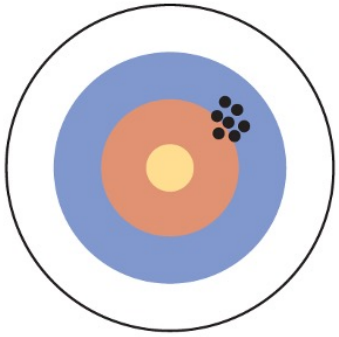
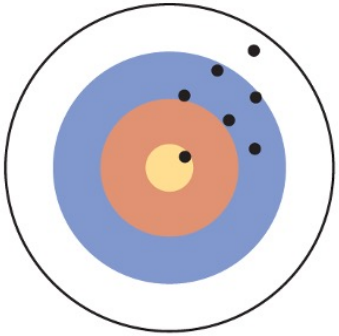
Inaccurate

Inaccurate: Sample values (e.g., sample means) tend to be far from the true population value.

Imprecise: Sample values (e.g., sample means) tend to vary widely from each other, regardless of whether they are close to or far from the population value.



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

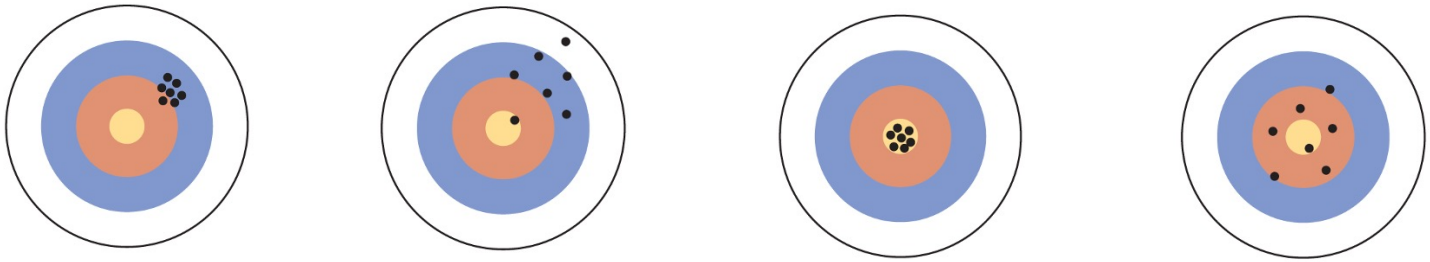
	Precise	Imprecise
Accurate	 Low sampling variation (sampling error) & low bias	 High sampling variation (sampling error) & low bias
Inaccurate	 Low sampling variation (sampling error) & high bias	 High sampling variation (sampling error) & high bias

Random sampling minimizes bias and allows for the measurement of sampling error (to be covered in upcoming lectures).

Sample bias: Occurs when certain observational units in the target population have a lower or higher probability of being sampled.

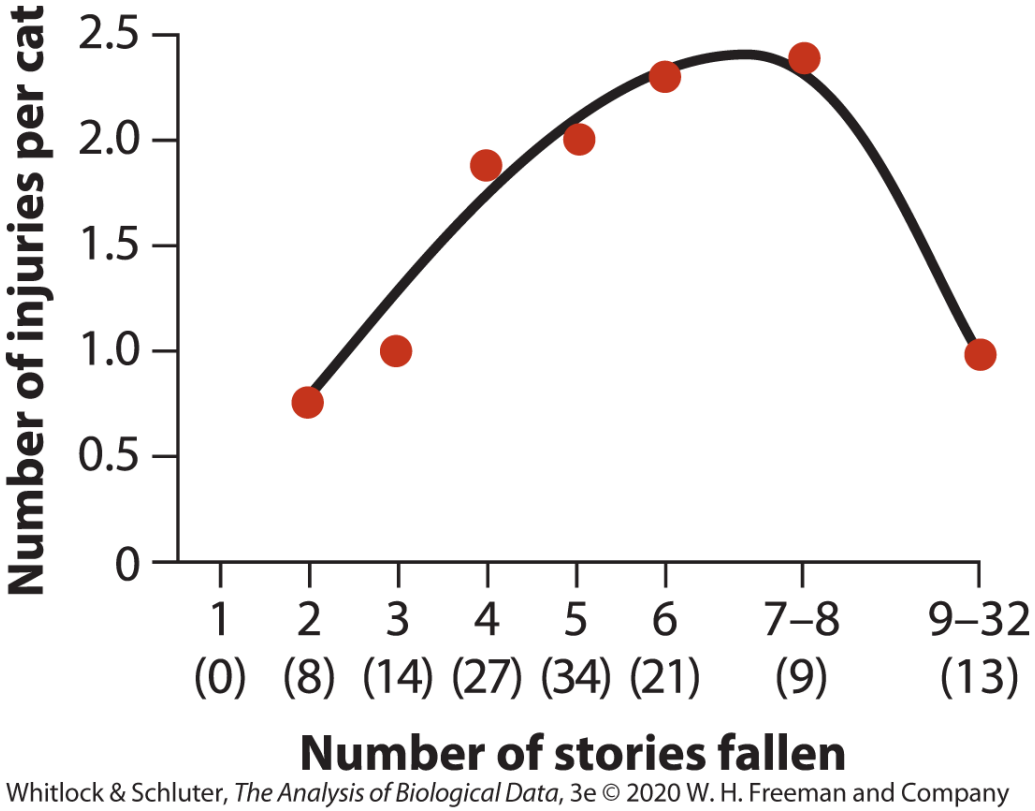
Inferential bias: Arises when the average of all sample values for the statistic of interest (e.g., mean tree height) differs from the true population value. There are many sources of inferential bias, including a lack of random sampling (other sources will be discussed later).

Sampling variation: Refers to the variation that occurs from sample to sample in terms of the statistic of interest. High sampling variation leads to imprecision.



Sampling populations - what can go wrong?
Issues with biased samples based on **sampling of convenience**

Sampling bias occurs when certain members of a population are systematically more likely to be selected in a sample than others

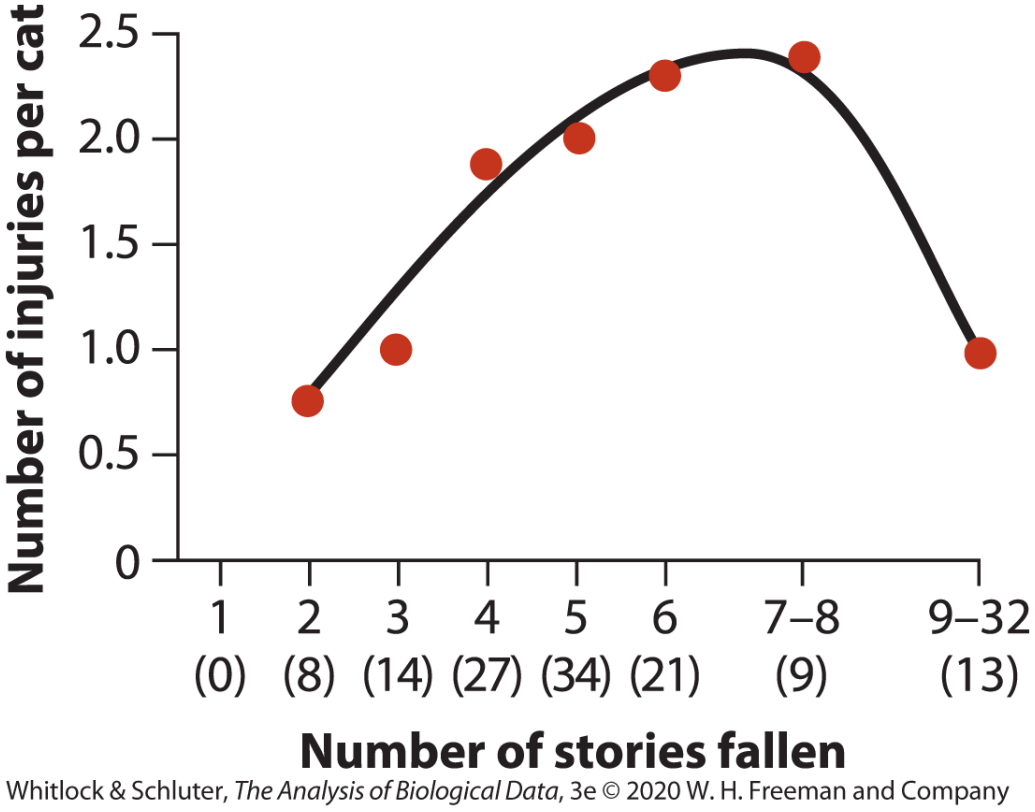


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

The author proposed that dropping more than 8 floors allows cat to relax and change muscles to cushion their impact

Mehlaff (1987) – Journal of the American Veterinary Medical Association

Sampling populations - what can go wrong?
Issues with biased samples based on **sampling of convenience**



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

Critics of the study pointed out that instantly fatal falls were not included.

Issue with samples of convenience = existent data not collected for the purposes of the study.

Sampling populations - what can go wrong?

Observational units may vary in other aspects that may lead to sampling biases

Oecologia (2013) 171:339–345
DOI 10.1007/s00442-012-2426-5

METHODS

Are most samples of animals systematically biased? Consistent individual trait differences bias samples despite random sampling

Peter A. Biro

Sampling bias occurs when certain members of a population are systematically more likely to be selected in a sample than others.

Sampling populations - what can go wrong?

Observational units may vary in other aspects that may lead to sampling biases

Volunteer bias

In a large experiment to test the effectiveness of a polio vaccine, schoolchildren were randomly selected to receive either the vaccine or a saline solution as a control.

The vaccine was shown to be effective, but it was later discovered that the rate of polio infection among children in the saline group was higher than that of the general population.

One possible explanation (as suggested by Bland, 2000) is that parents of children who had not been exposed to polio—and therefore had no immunity—were more likely to volunteer their children for the study than parents of children who had already been exposed.

Bland, M. 2000. An Introduction to Medical Statistics, 3rd ed. Oxford, UK: Oxford University Press.

Sampling populations - what can go wrong?

Observational units may vary in other aspects that may lead to sampling biases

Volunteer bias

In a large experiment to test the effectiveness of a polio vaccine, schoolchildren were randomly selected to receive either the vaccine or a saline solution as a control.

The vaccine was shown to be effective, but it was later discovered that the rate of polio infection among children in the saline group was higher than that of the general population.

One possible explanation (as suggested by Bland, 2000) is that parents of children who had not been exposed to polio—and therefore had no immunity—were more likely to volunteer their children for the study than parents of children who had already been exposed.

Compared to the general population, volunteers may:

- Be more health-conscious and proactive;
- Have lower incomes (especially if volunteers are compensated);
- Be more ill, particularly if the therapy carries risk, as individuals with severe illnesses may be willing to try anything;
- Have more free time (e.g., retirees or unemployed individuals are more likely to participate in telephone surveys);
- Be more upset or angry, as people who are dissatisfied may be more inclined to express their views (e.g., surveys);
- And so on.

Bland, M. 2000. An Introduction to Medical Statistics, 3rd ed. Oxford, UK: Oxford University Press.

Look into notes & additional material in the WebBook

Survivorship bias: great video explaining sample bias (also covered in Whitlock & Schluter). This is a great video where wrong understanding of sampling can lead to wrong decisions.