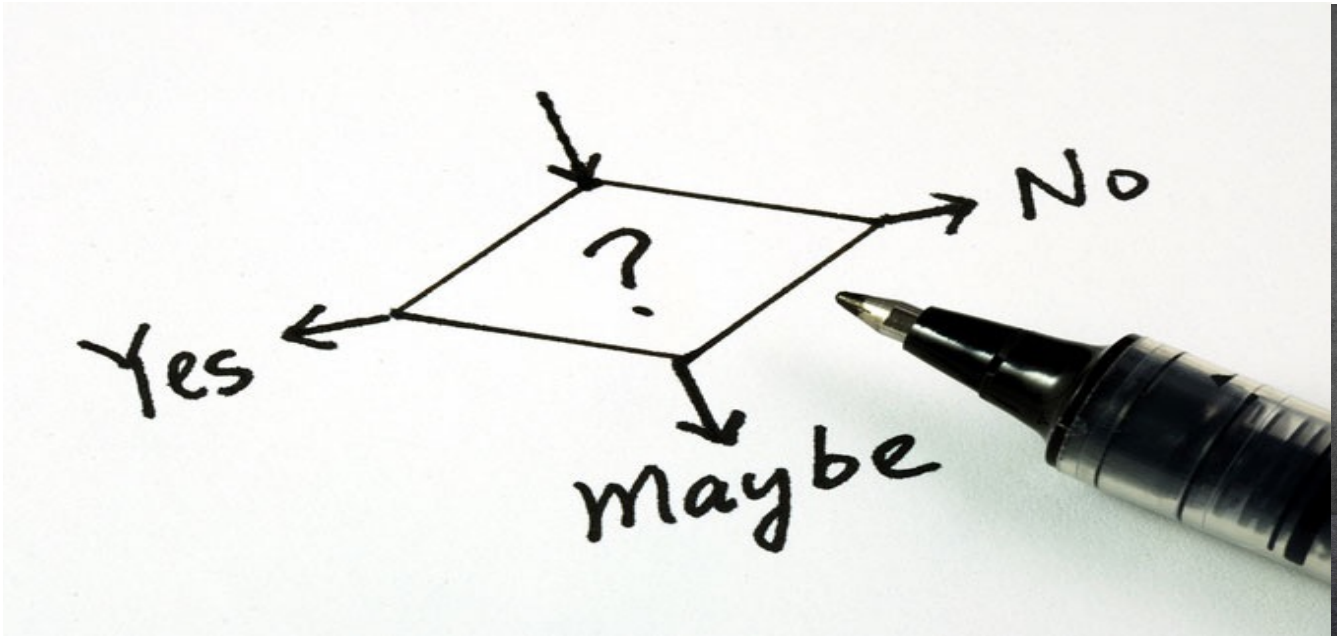
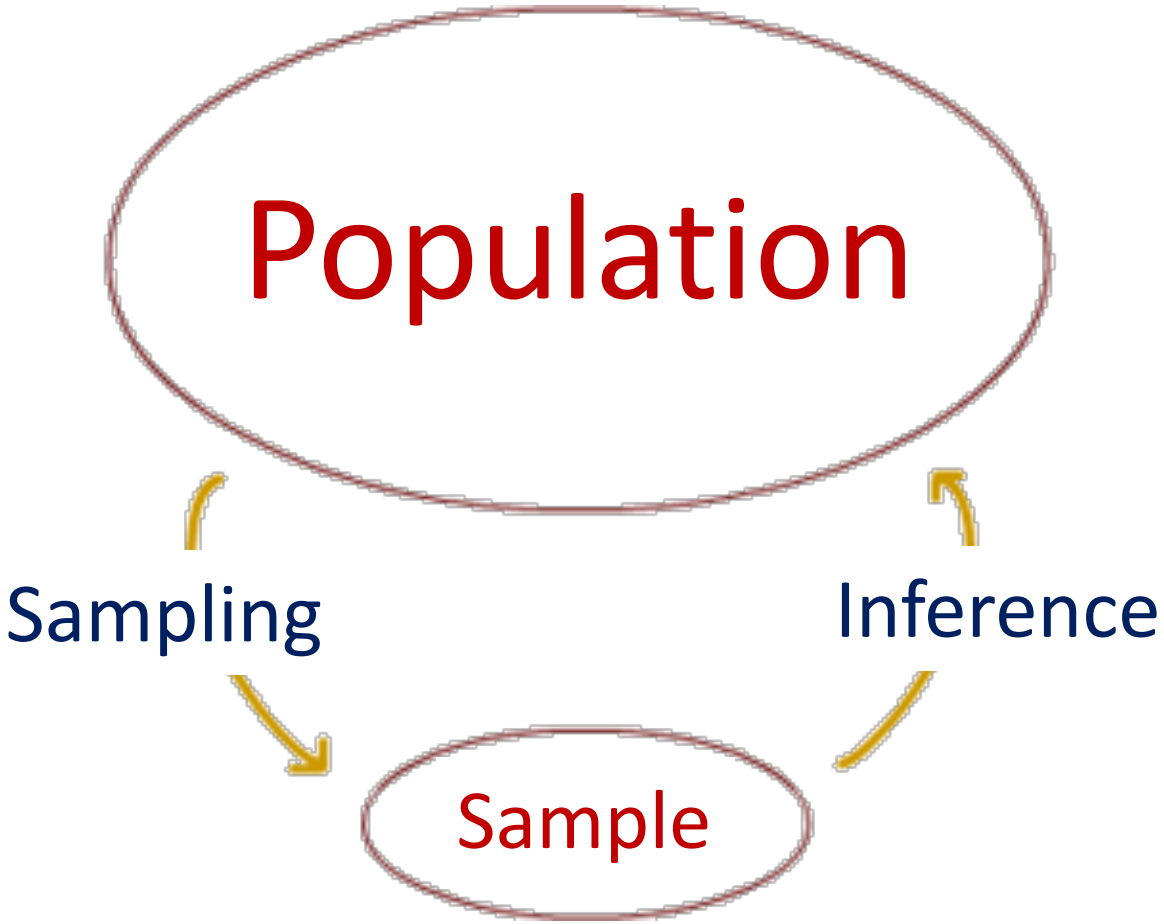
Lecture 7: estimating & making inferences with uncertainty – samples and biases

Statistics is the science of assisting in decision making with incomplete knowledge (i.e., without certainty)

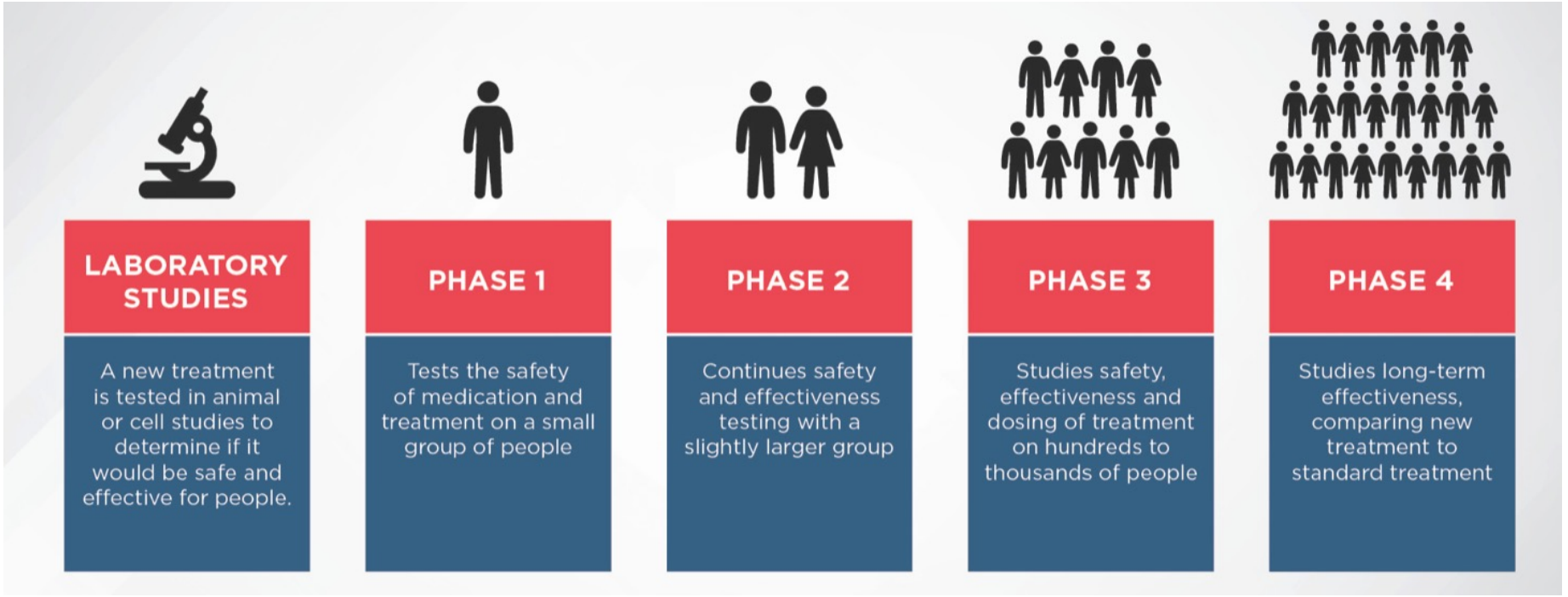


Inferential process



A good example of sampling:

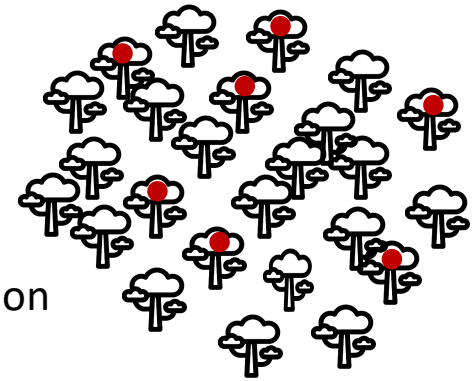
Stages of Clinical Trials



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

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

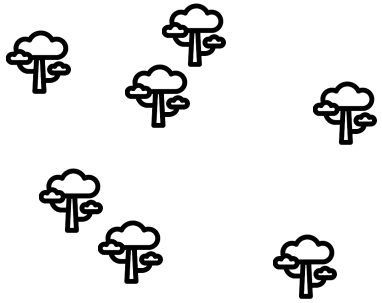
We want to know about a large number of trees



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)

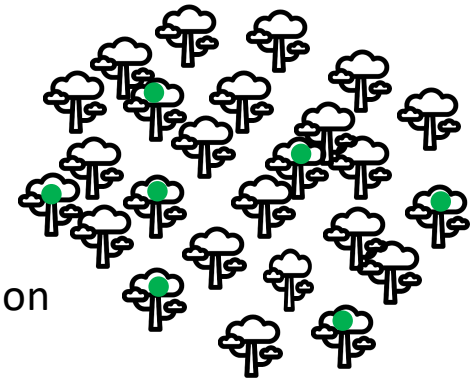
Sample mean height

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

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

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



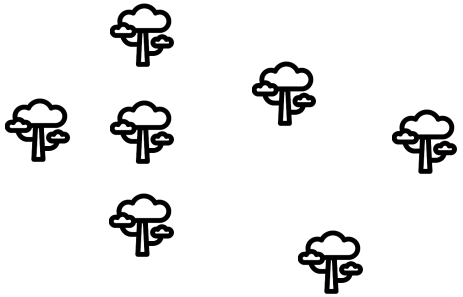
Population

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

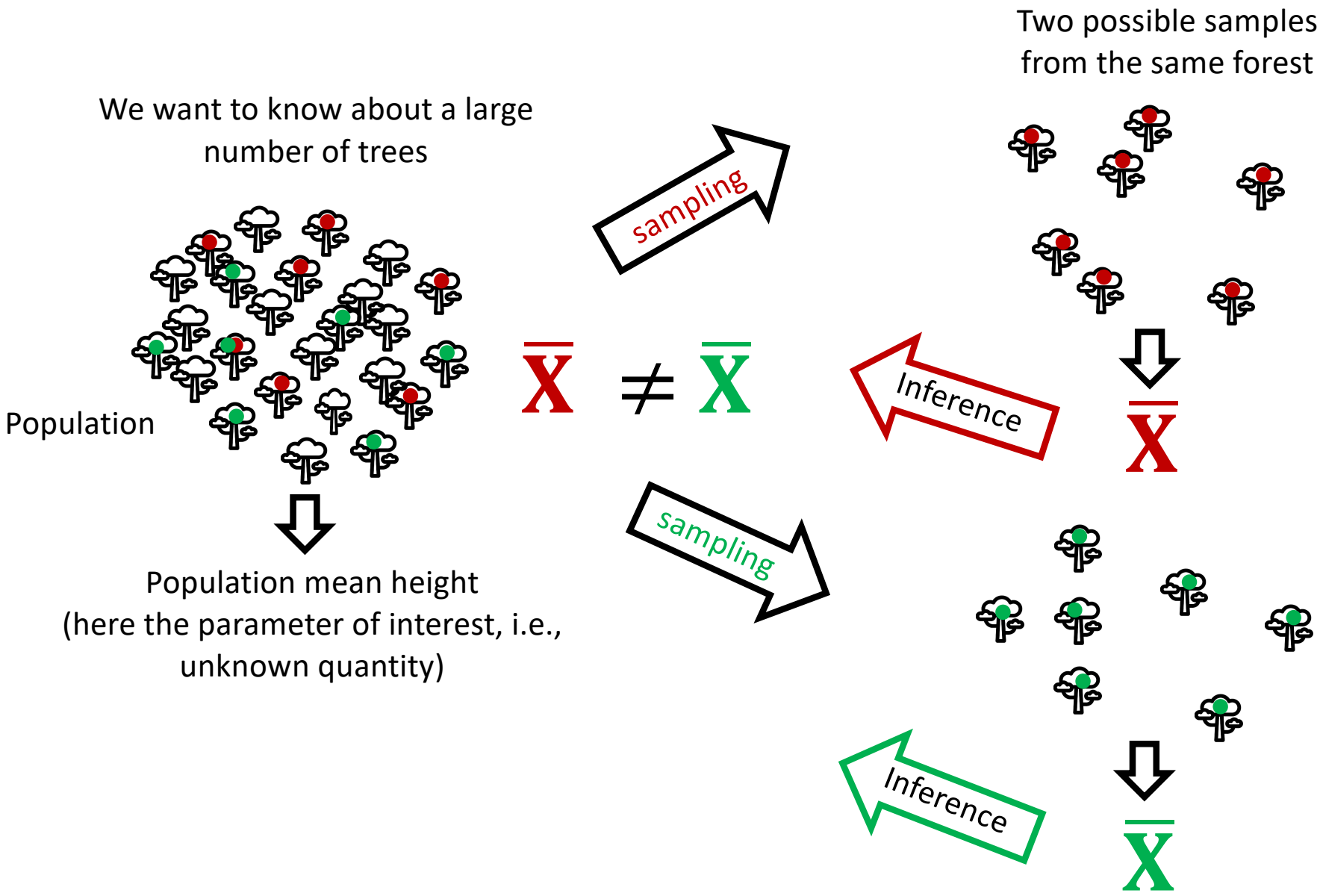


Sample

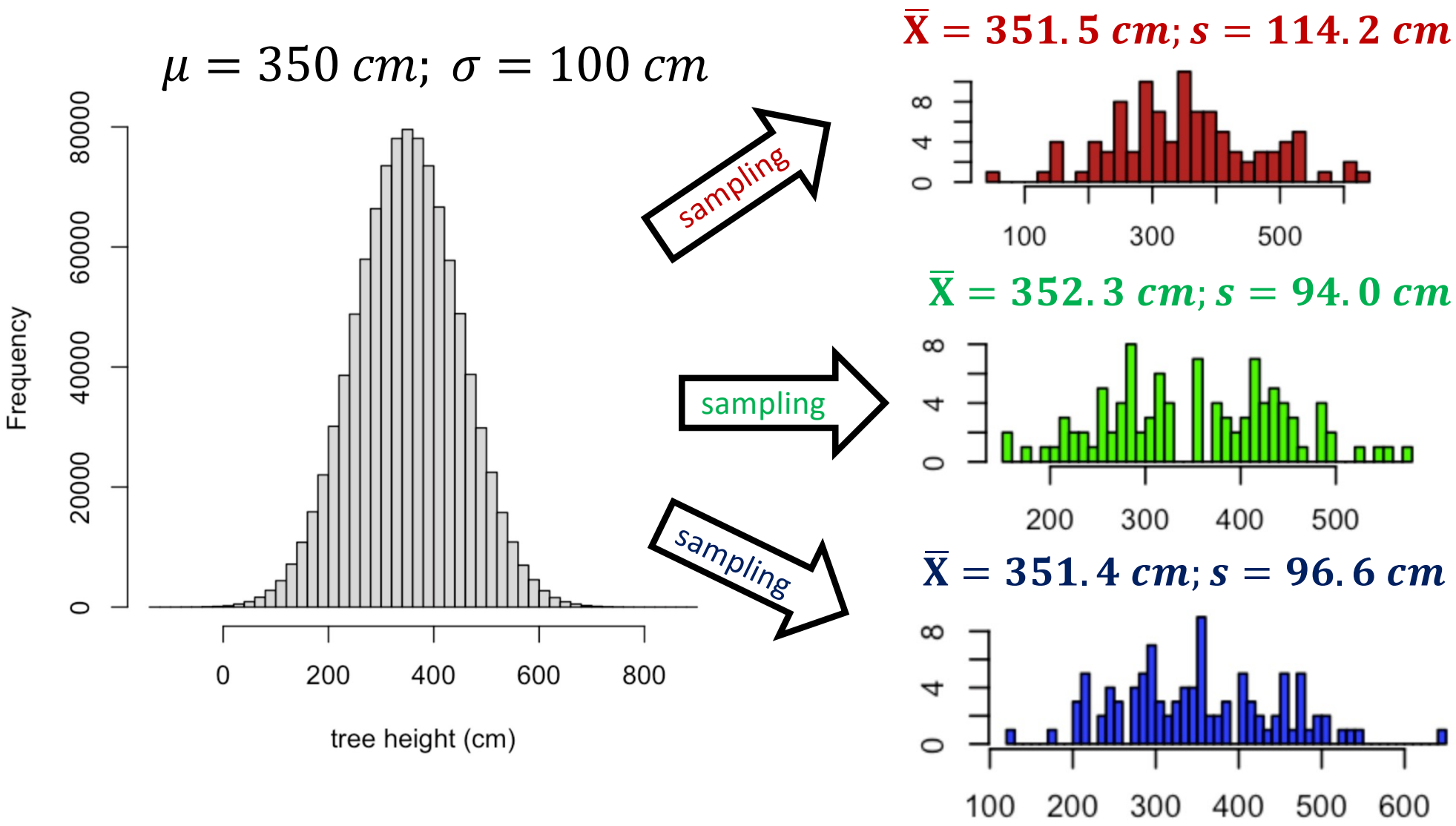
Sample mean height

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

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



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



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

Sampling 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.

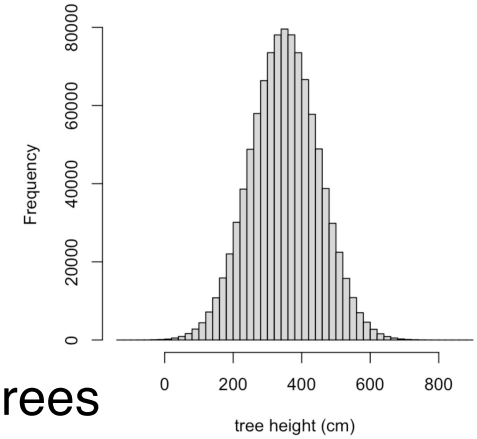
How many possible samples of 100 trees out of a population with 1000000 trees?

10768272362e+432 (zeros)

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

In real studies, most of the time though, we only take one sample from the intended statistical population

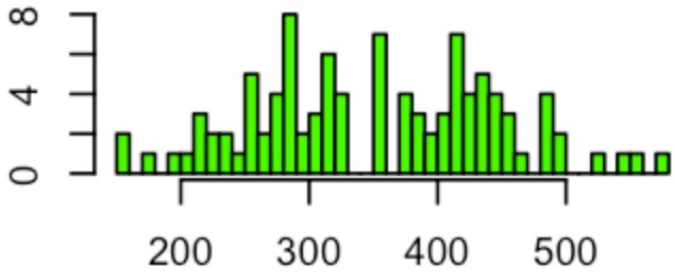
$\mu = 350\ cm;\ \sigma = 100\ cm$



1 000 000 trees

sampling

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



100 trees

Critical understanding:

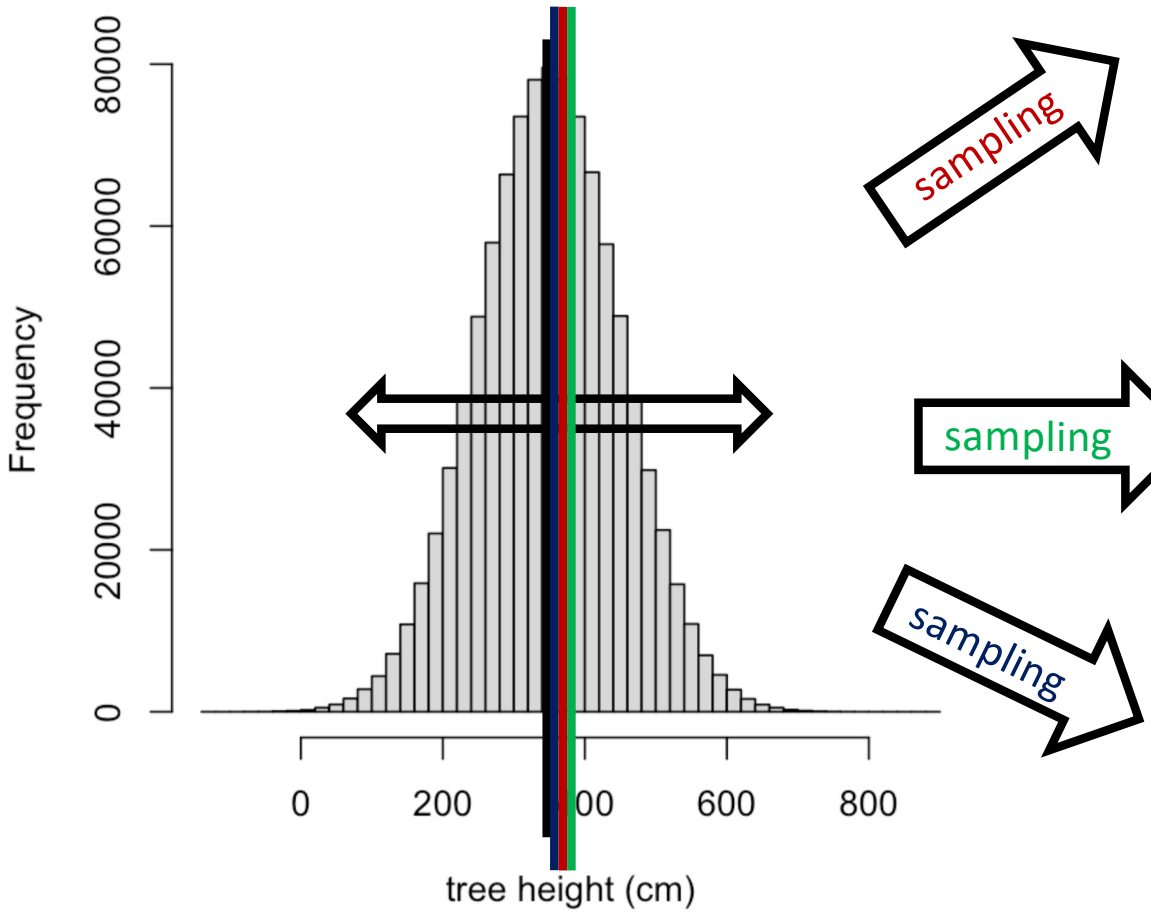
[1] Values for descriptive statistics based on samples are never (exactly) the same as their values for the populations (there is always “sampling error”).

[2] That does not mean that inference based on samples are wrong (more on that later). Sample values can be a very good approximation of the true value.

[3] Approximations can be good (sample value for the statistic of interest close to the true population value) or bad (sample value far from the true value).

Sampling variation: some samples are **closer** to the mean whereas others are **far** from the mean (i.e., samples vary among each other)

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



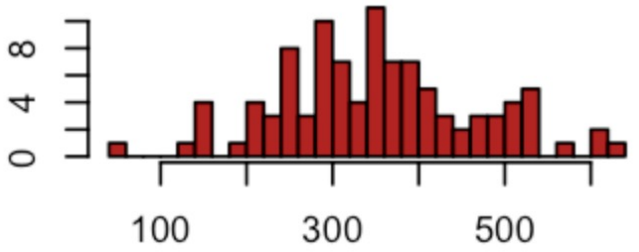
$\mu = 350\ cm$

sampling

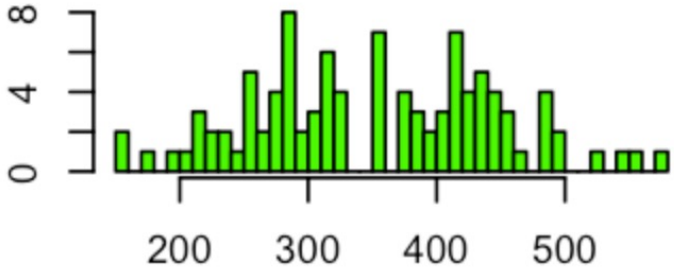
sampling

sampling

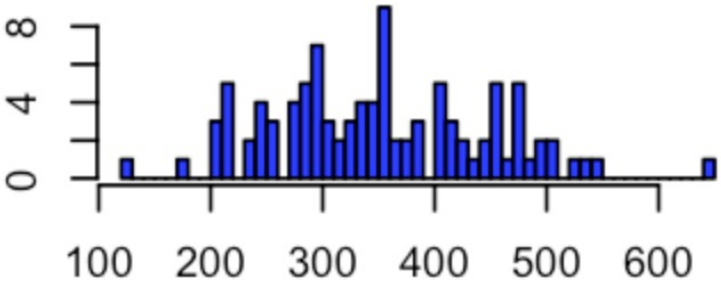
$\bar{X} = 351.5\ cm; s = 114.2\ cm$



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

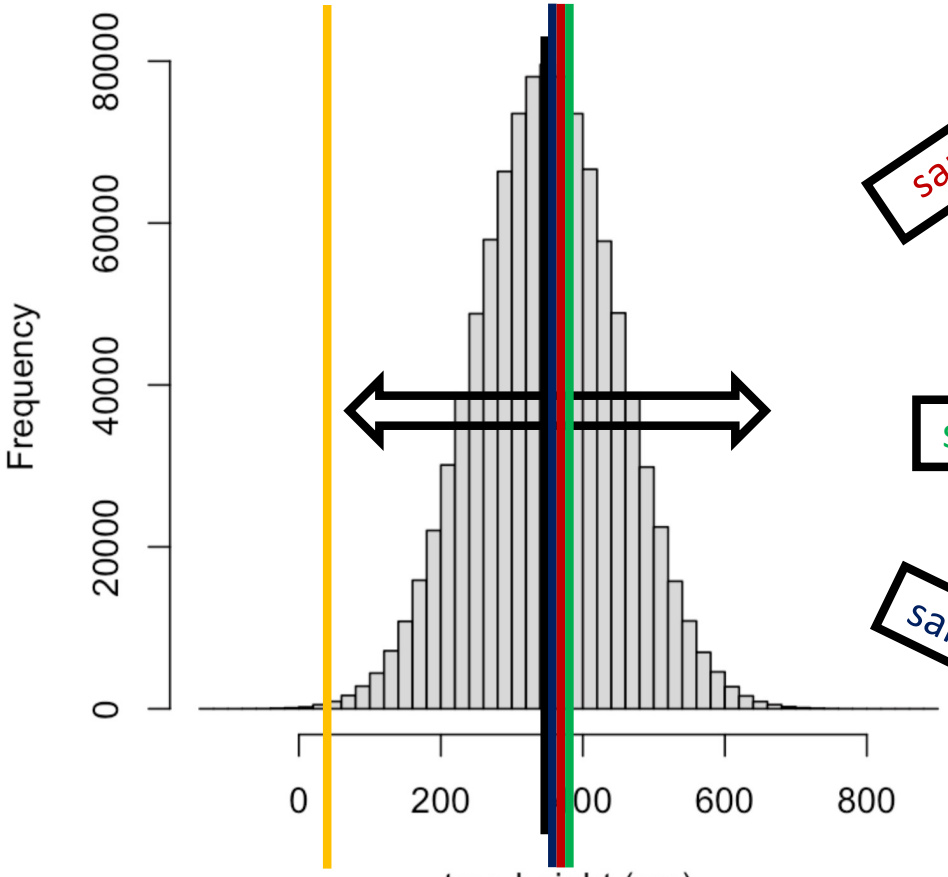


$\bar{X} = 351.4\ cm; s = 96.6\ cm$



Sampling variation: some samples are **closer** to the mean whereas others are **far** from the mean (i.e., samples vary among each other)

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

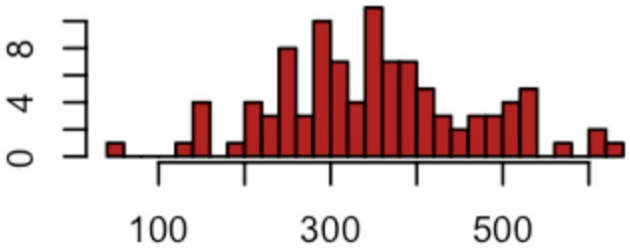


sampling

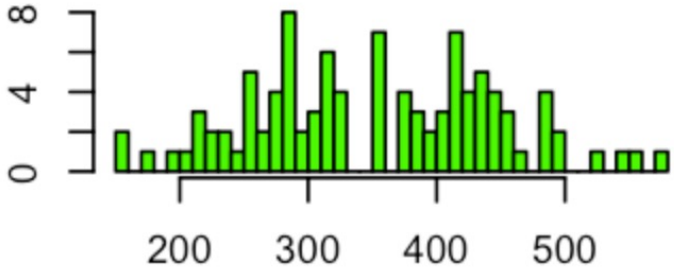
sampling

sampling

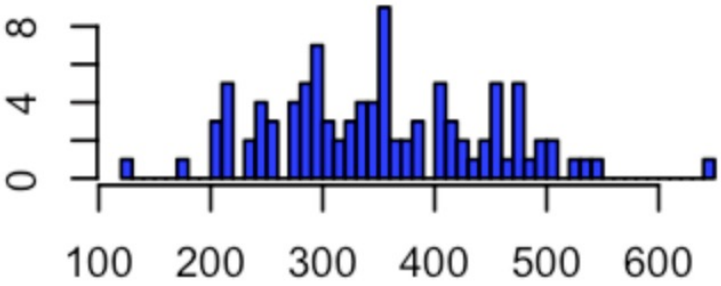
$\bar{X} = 351.5\ cm; s = 114.2\ cm$



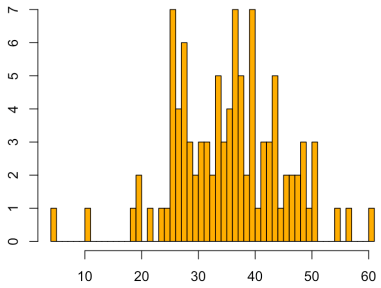
$\bar{X} = 352.3\ cm; s = 94.0\ cm$



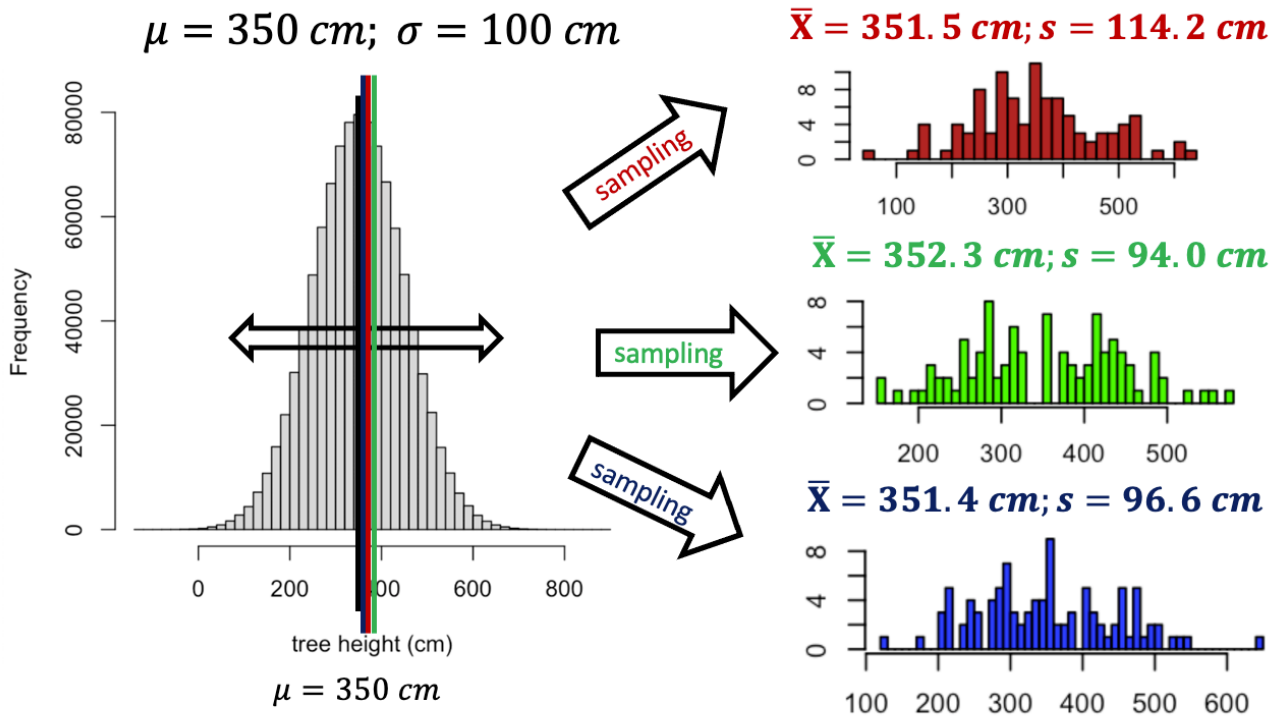
$\bar{X} = 351.4\ cm; s = 96.6\ cm$



$\mu = 350\ cm$



Sampling variation: some samples are **closer** to the mean whereas others are **far** from the mean (i.e., samples vary among each other)

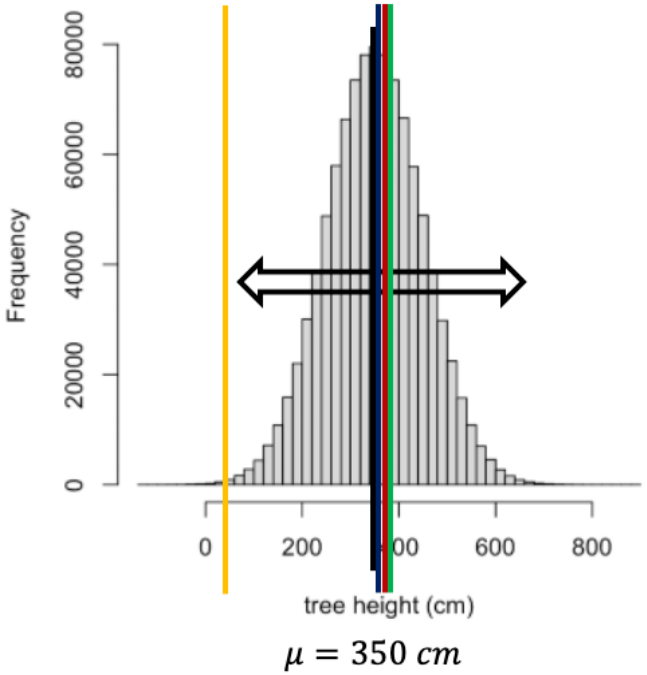


As we will see later in the semester, an unbiased estimator (e.g., mean, standard deviation, median) is one that in average across multiple (repeated) samples equals the population parameter.

Because we usually only have one sample, we need to demonstrate mathematically that the estimator we are using is unbiased. More on that later in the semester.

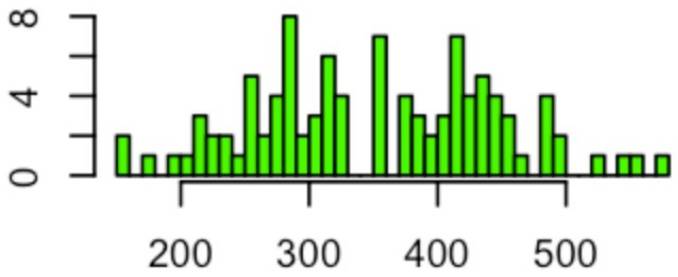
In real studies, though, most of the time, we only take one sample from the intended statistical population

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



sampling

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



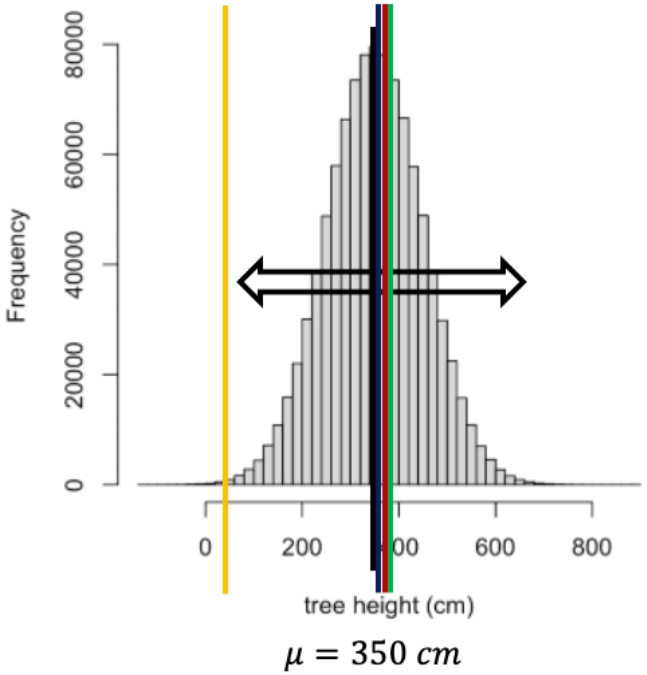
Critical understanding:

[3] Again, approximations can be good (sample value close to the true population value) or bad (sample value far from the true value). You will understand later why we use terms “close” & ”far” to describe samples in relation to their populations next.

[4] But to feel “safe”, it would be great to have a measure that estimates how wrong one could be.

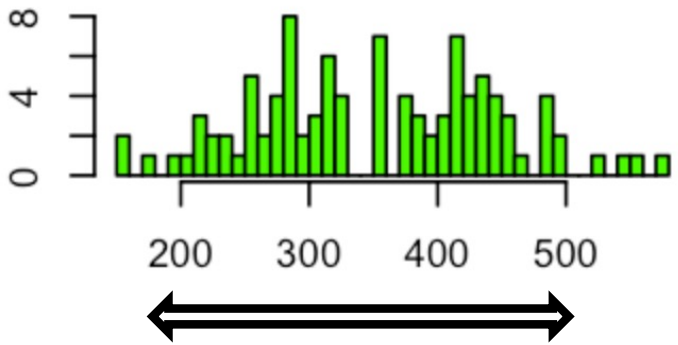
How wrong one could be in trusting their sample values to estimate the population value (i.e., parameter)?

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



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 inform us about how far sample means in general can be from the true population mean (estimate how wrong one could be).

[3] Approximations can be good (sample value close to the true population value) or bad (sample value far from the true value). You will understand later why we use terms “close” & ”far” to describe samples in relation to their populations next.

[4] But to feel “safe”, it would be great to have a measure that estimates how wrong one could be.

Key concepts underlying statistics and statistical thinking

- Uncertainty (never being able to know the true population parameter).
- Risk of being wrong (error); decisions based on estimates closer to the true value may be not problematic; but when far from the true value…then decisions may be “wrong”.
- Evaluating risk then becomes key!
- Sample variability - Answer may change with different sample data.
- Accuracy (close to reality).

Key concepts:
Statistics is based on samples!

Sample quantities (mean, median, standard deviation, interquartile range, etc) almost always vary from sample to sample (i.e., they have some level of uncertainty).

As such, we always estimate and make inferences with some level of uncertainty about the population true values.

As we will see later, the variation among observations within samples (standard deviation) can inform us about how far sample means in general can be from the true population mean (estimate uncertainty and the risks involved).

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

Understanding sampling variation with dance.

Let’s take a break - 2 minutes

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.

Before I forget!!!!

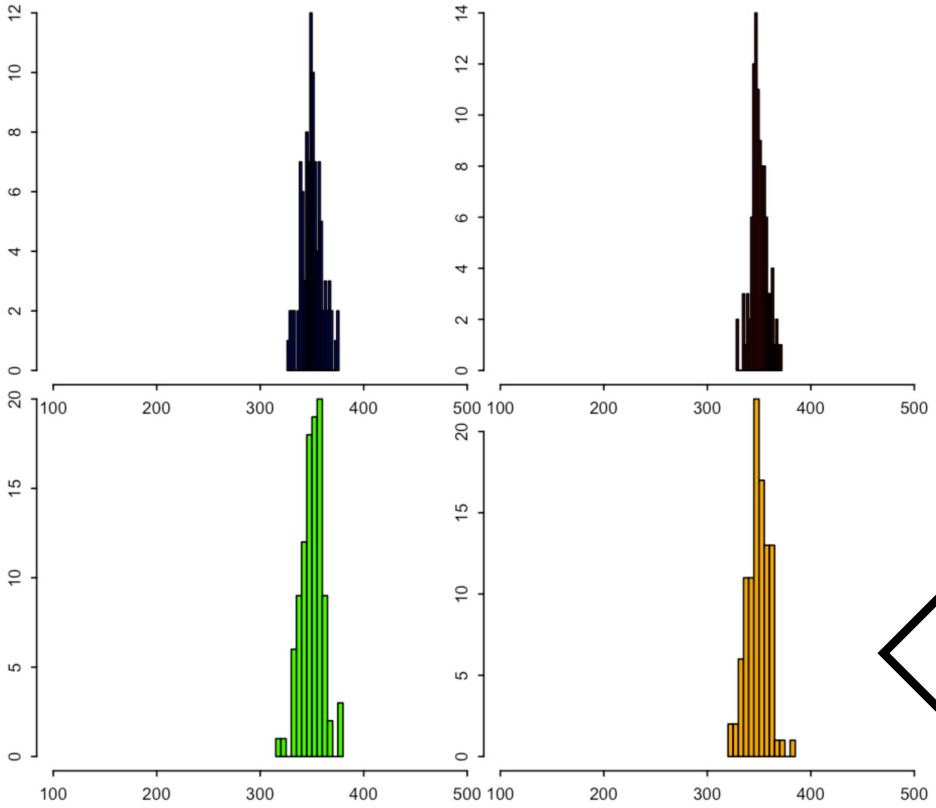
2) The selection of observational units in the population (e.g., individual tree) must be **independent**, **i.e.**, the selection of any unit (e.g., individual tree) of the population must not influence the selection of any other unit.

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

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

Precise

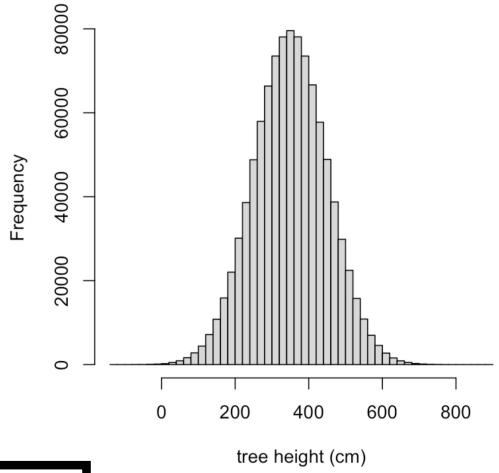
Accurate



4 random samples of 100 trees each

Tree population

$\mu = 350\ cm$



sampling

1 000 000 trees (heights in cm)

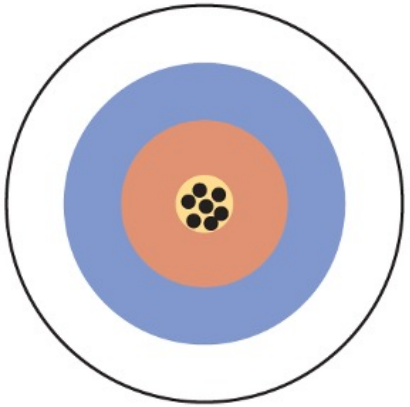
The average sample mean vary very little from sample to sample (precision) and in average they are similar to the population mean (accurate).

These 4 samples are precise and accurate

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

Precise

Accurate



Imagine the bull's eye as the population parameter (here mean tree height) and the points are possible different sample mean values of tree heights (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 whether they are close or far from the population value.

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

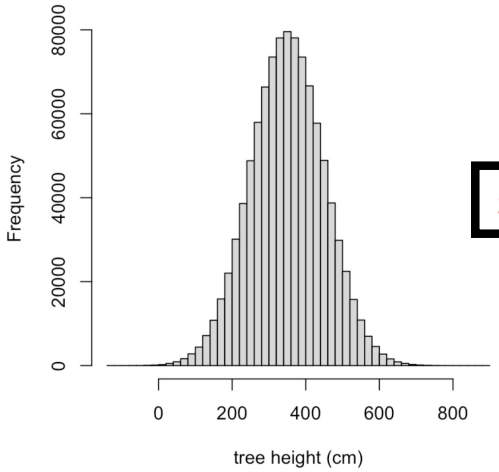
Imprecise

Black line = Population mean
Red line = sample mean

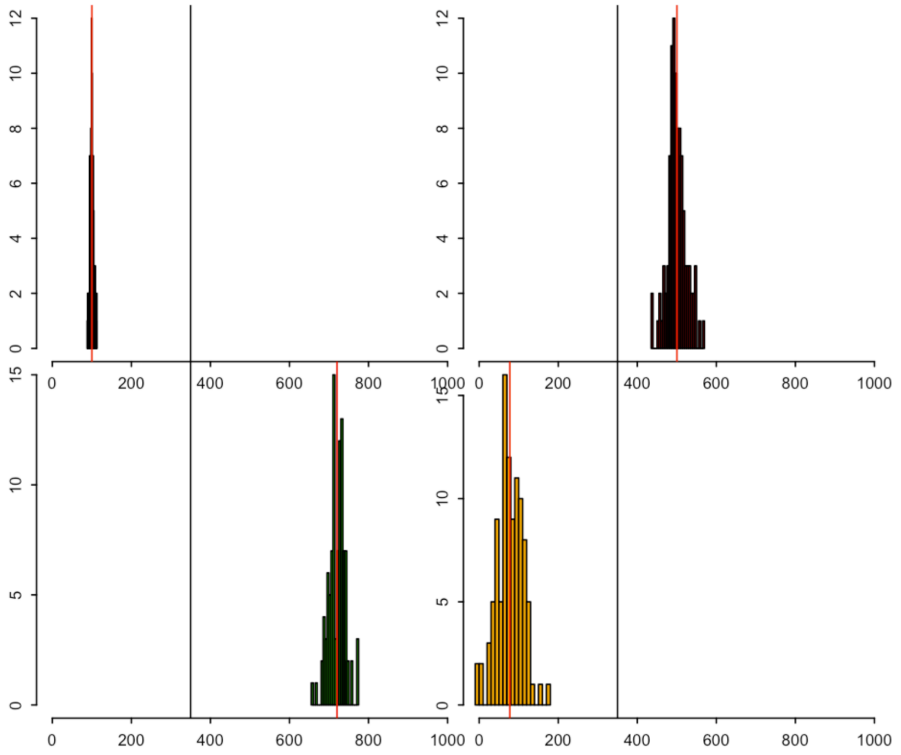
Accurate

Tree population

$\mu = 350\ cm$



sampling



1 000 000 trees
(heights in cm)

4 random samples
of 100 trees each

The average sample mean vary a lot from sample to sample (imprecise) but in average they are similar to the population mean (accurate), i.e., the average of these 4 samples are very close to the true population value

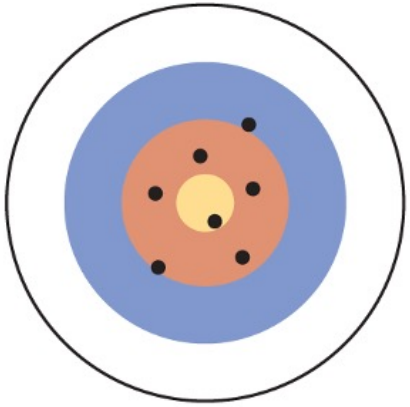
These 4 samples are imprecise but accurate

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

Imprecise

Accurate

Imagine the bull's eye as the population parameter (here mean tree height) and the points are possible different sample mean values of tree heights (i.e., estimates).



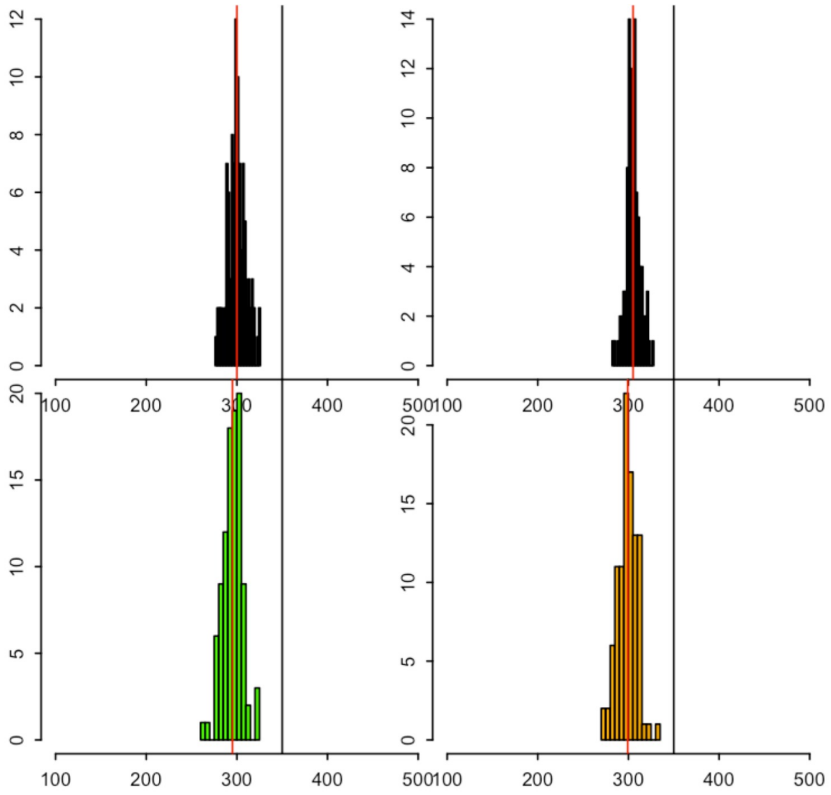
Accurate = sample values (e.g., sample means) tend to be close to the true population value.

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

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

Precise

Inaccurate

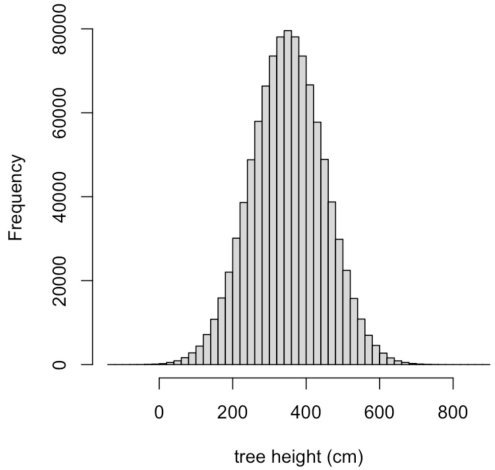


4 random samples of 100 trees each

Black line = Population mean
Red line = sample mean

sampling

$\mu = 350\ cm$



1 000 000 trees
(heights in cm)

The average sample mean vary very little from sample to sample (precise) but in average they are different from the population mean (inaccurate).

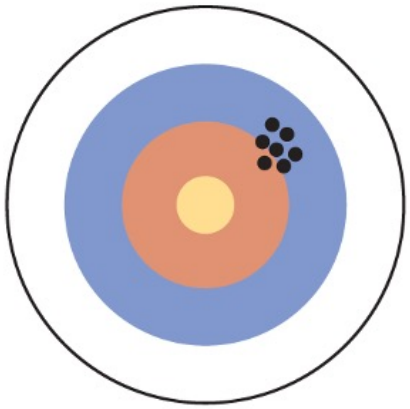
These 4 samples are precise but inaccurate

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

Precise

Imagine the bull's eye as the population parameter (here mean tree height) and the points are possible different sample mean values of tree heights (i.e., estimates).

Inaccurate



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

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

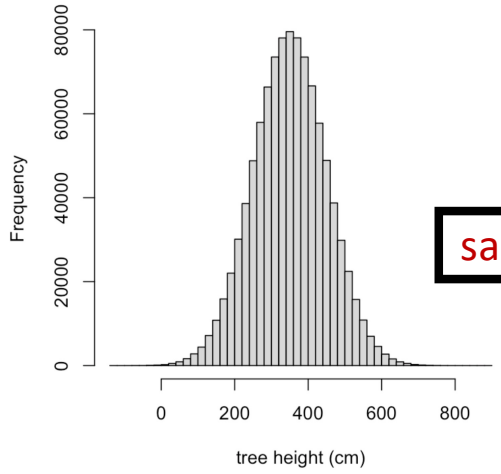
Precision and accuracy: Properties of samples
the major goal of sampling is to increase accuracy and precision

Black line = Population mean
Red line = sample mean

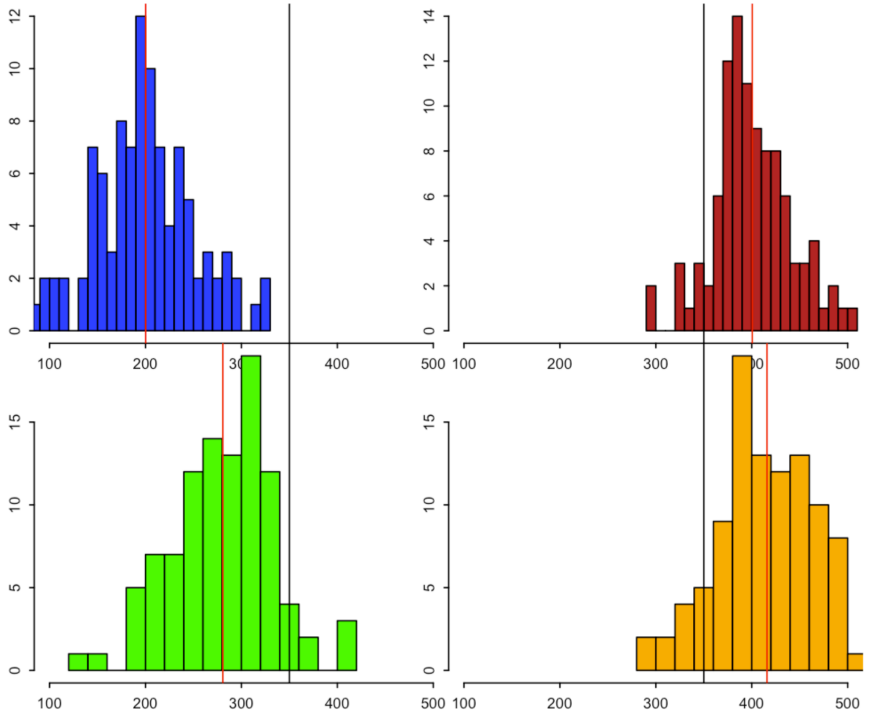
Imprecise

Inaccurate

$\mu = 350\ cm$



sampling



1 000 000 trees
(heights in cm)

4 random samples
of 100 trees each

The average sample mean vary a lot from sample to sample (imprecise) and in average they are different from the population mean (inaccurate).

These 4 samples are imprecise and inaccurate

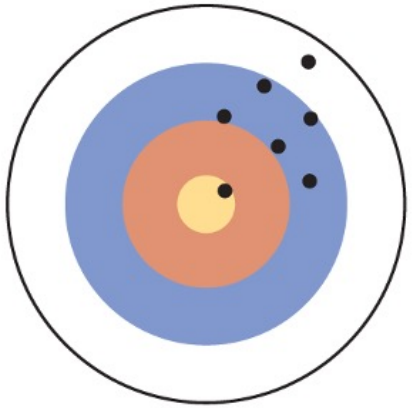
Precision and accuracy: Properties of samples
the major goal of sampling is to increase accuracy and precision

Imprecise

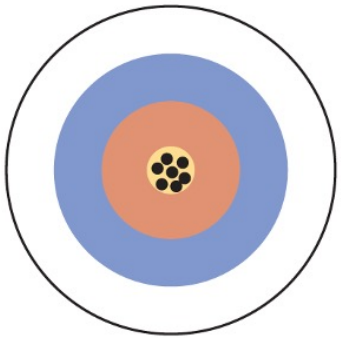
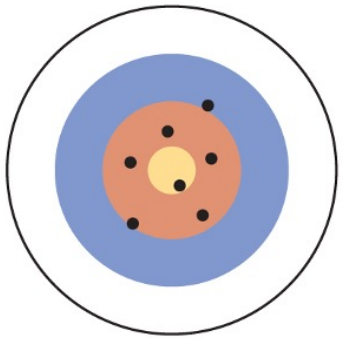
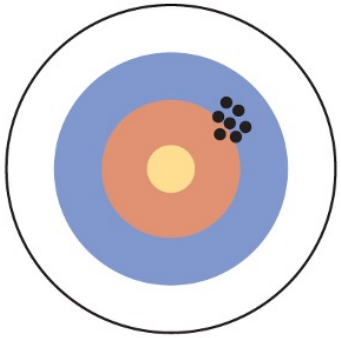
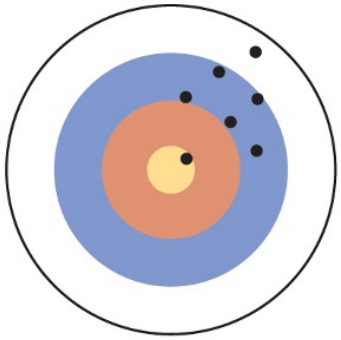
Inaccurate

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

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



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

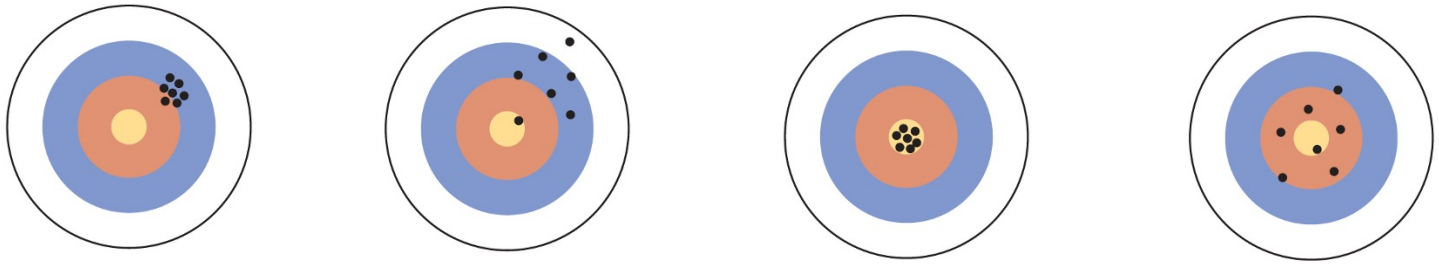
	Precise	Imprecise
Accurate	 Low sampling variation (sampling error) & low bias	 High sampling variation (sampling error) & low bias
Inaccurate	 Low sampling variation (sampling error) & high bias	 High sampling variation (sampling error) & high bias

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

Sample bias: when some observational units of the intended population have lower or higher probabilities to be sampled.

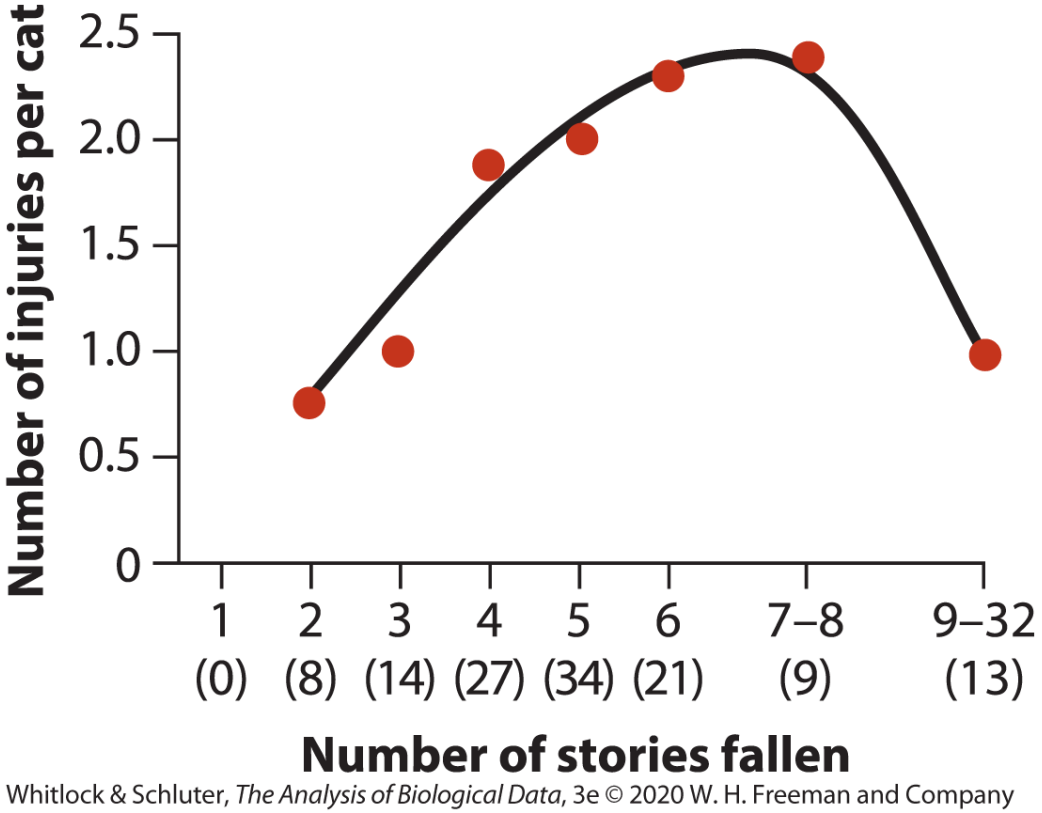
Inferential bias: when the average of all sample values of the statistic of interest (mean tree height) is different from the true population value. There are many sources of inferential biases, including lack of random sampling (other sources will be covered soon).

Sampling variation: the variation that exists 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 some members of a population are systematically more likely to be selected in a **sample** than others.

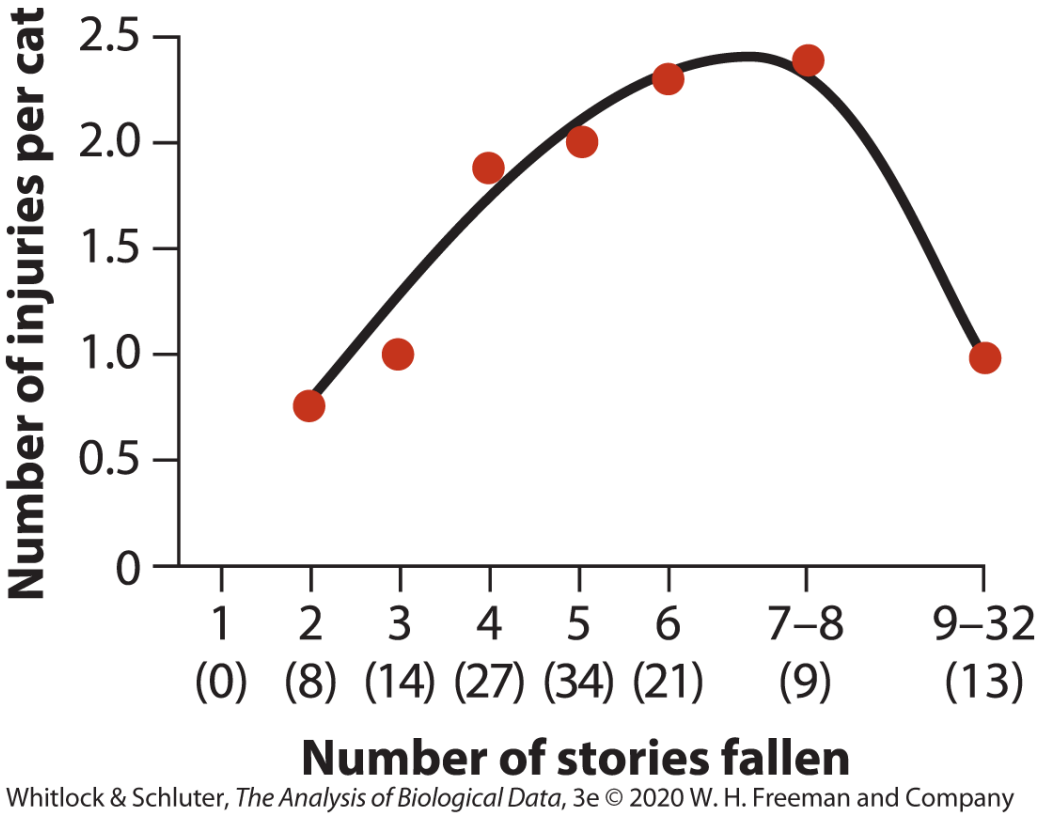


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

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.

Issues 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 some 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 benefits of a polio vaccine, for example, participating school children were randomly chosen to receive either the vaccine or a saline solution (serving as the control).

The vaccine proved effective, but the rate at which children in the saline group contracted polio was found (later on after the study was over) to be higher than in the general population.

Perhaps parents of children who had not been exposed to polio prior to the study, and therefore had no immunity, were more likely to volunteer their children for the study than parents of kids who had been exposed (Bland 2000).

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 benefits of a polio vaccine, for example, participating school children were randomly chosen to receive either the vaccine or a saline solution (serving as the control).

The vaccine proved effective, but the rate at which children in the saline group contracted polio was found (later on after the study was over) to be higher than in the general population.

Perhaps parents of children who had not been exposed to polio prior to the study, and therefore had no immunity, were more likely to volunteer their children for the study than parents of kids who had been exposed (Bland 2000).

Compared with the rest of the population, volunteers might be:

- more health conscious and more proactive;
- low-income (if volunteers are paid);
- more ill, particularly if the therapy involves risk, because individuals who are dying anyway might try anything;
- more likely to have time on their hands (e.g., retirees and the unemployed are more likely to answer telephone surveys);
- more angry, because people who are upset are sometimes more likely to speak up.
- etc

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

Look into notes and 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.