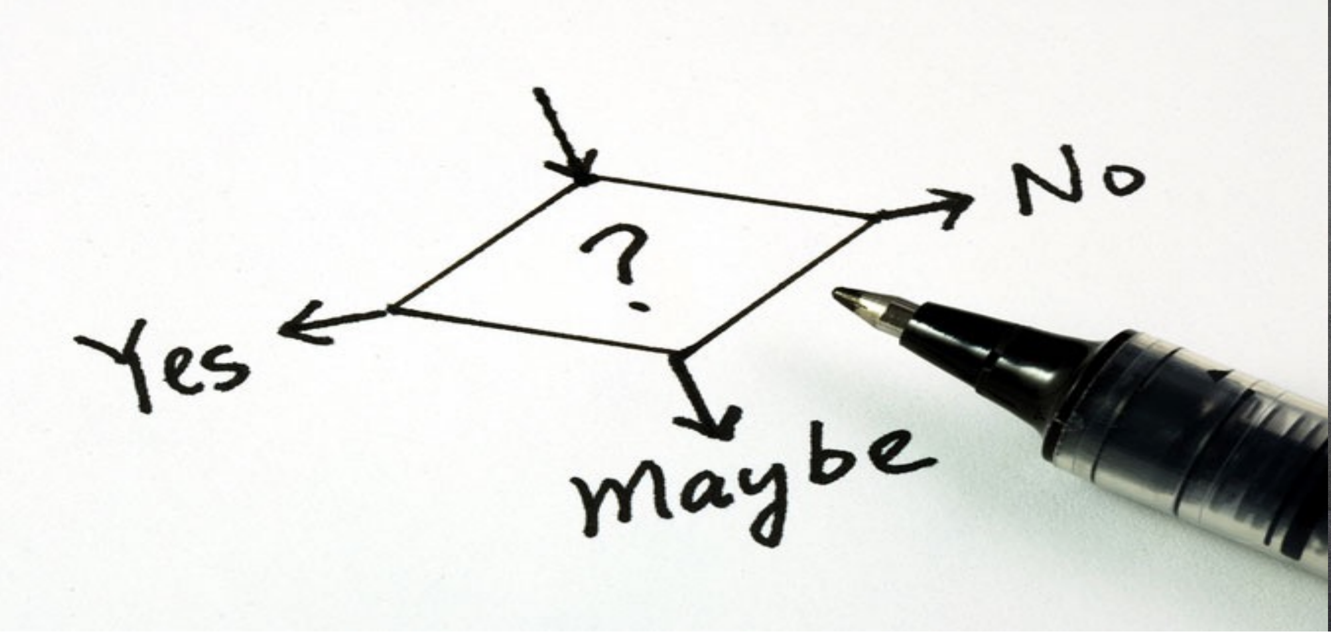
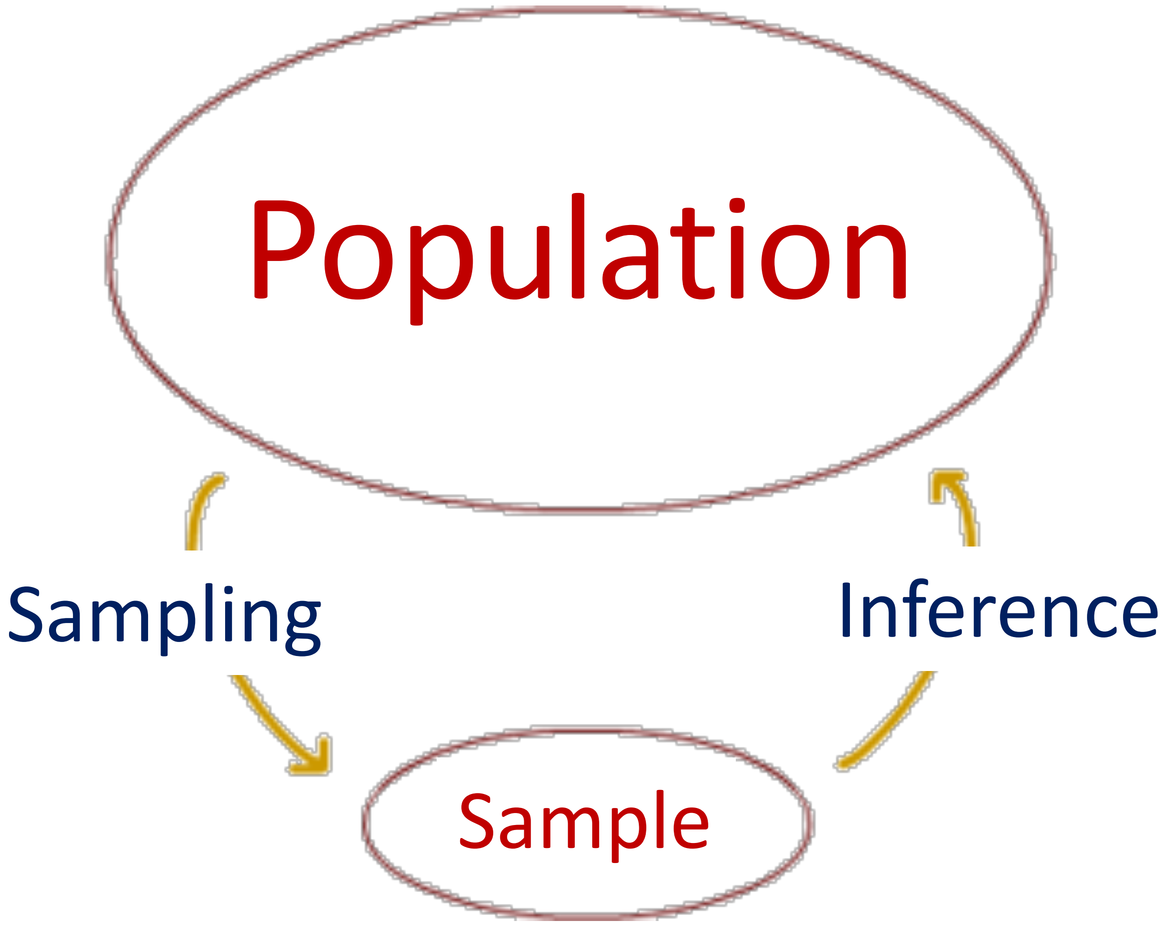
Lecture 7: estimating & making inferences with uncertainty – samples and biases

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



Inferential process



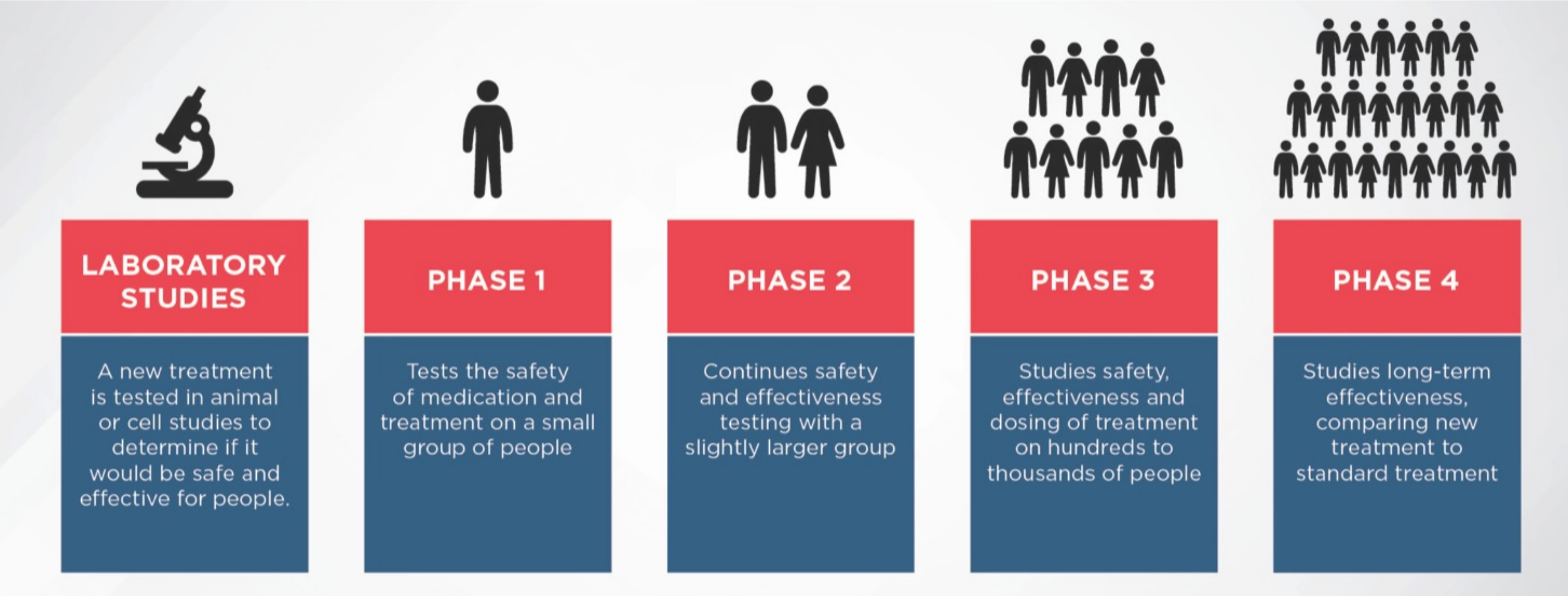
Population

Sampling

Inference

Sample

A good example of sampling:
Stages of Clinical Trials and related samples



LABORATORY STUDIES

A new treatment is tested in animal or cell studies to determine if it would be safe and effective for people.

PHASE 1

Tests the safety of medication and treatment on a small group of people

PHASE 2

Continues safety and effectiveness testing with a slightly larger group

PHASE 3

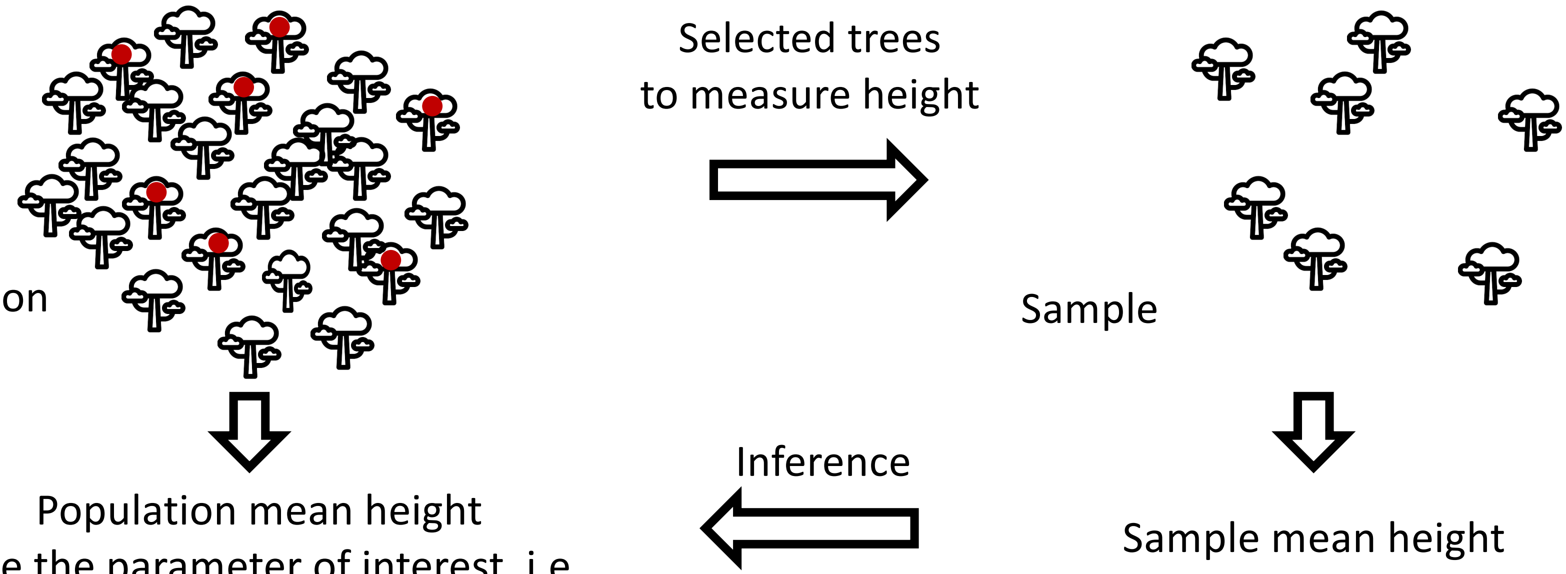
Studies safety, effectiveness and dosing of treatment on hundreds to thousands of people

PHASE 4

Studies long-term effectiveness, comparing new treatment to standard treatment

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

We have these trees to work with

Selected trees to measure height

Population

Sample

Inference

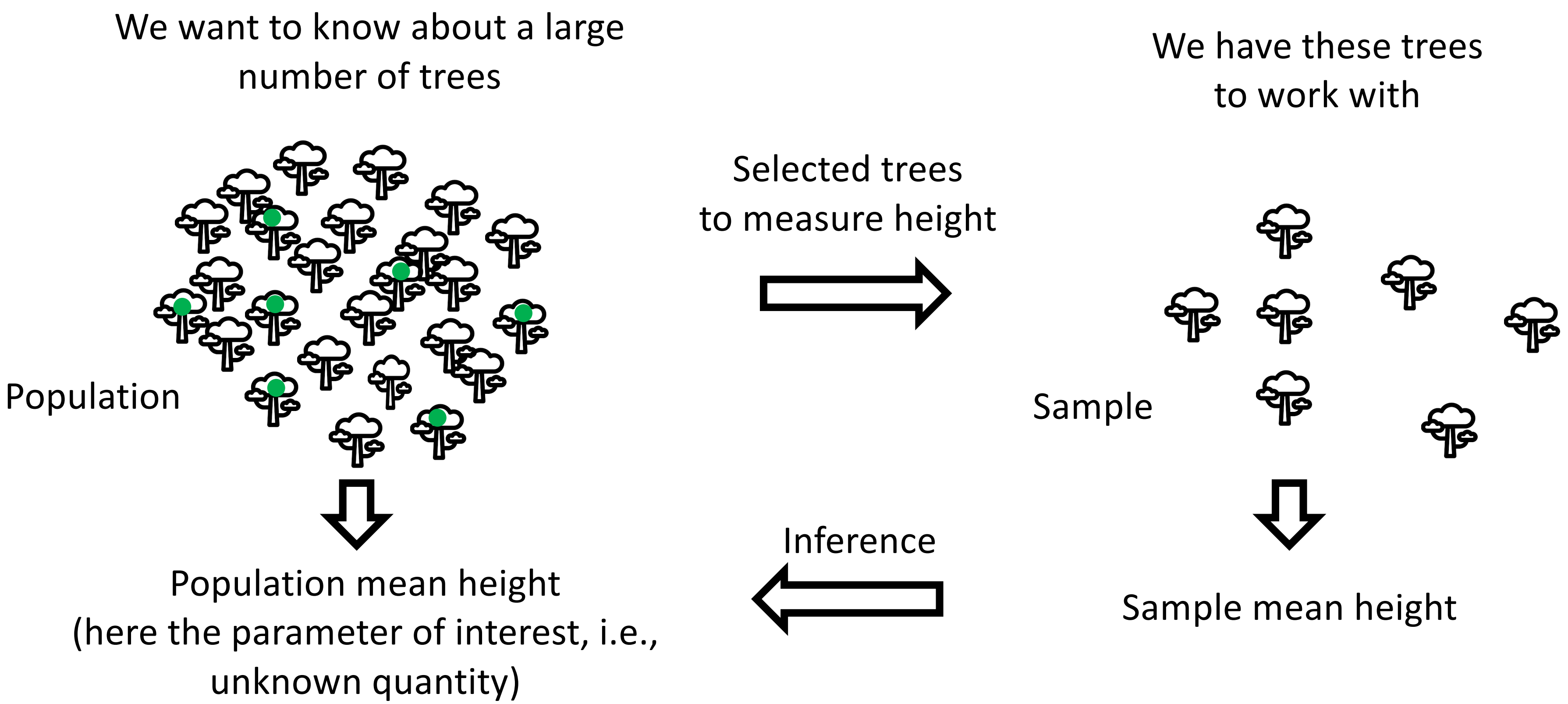
Population mean height (here the parameter of interest, i.e., unknown quantity)

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

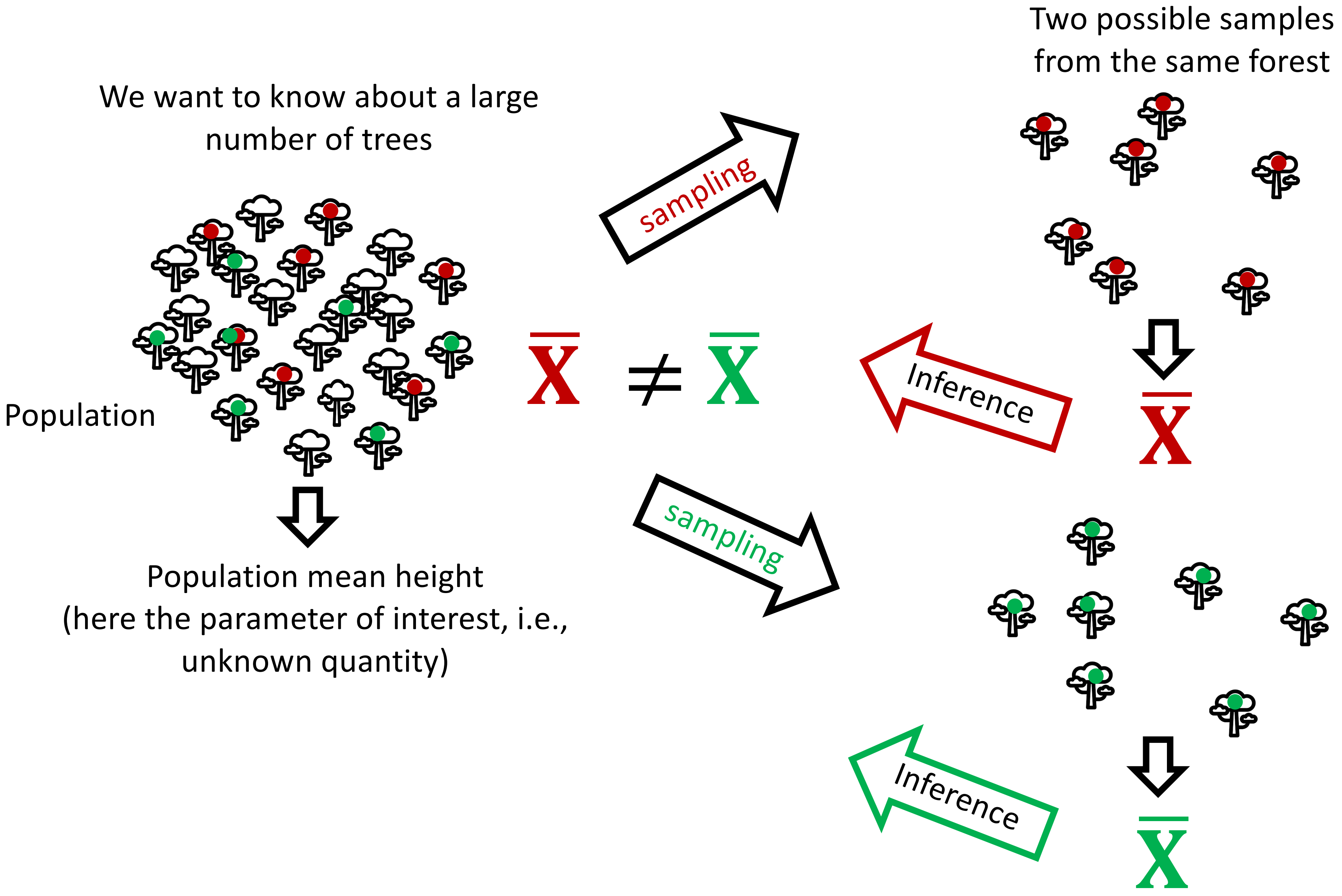
Sample

Sample mean height

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

Sampling variation – connecting samples and populations

The mean tree height estimated from different samples drawn from the same population will generally differ from each other and from the true population mean. Because of this unavoidable variability, we estimate population parameters and make inferences while explicitly accounting for uncertainty due to sampling variation.



Two possible samples from the same forest

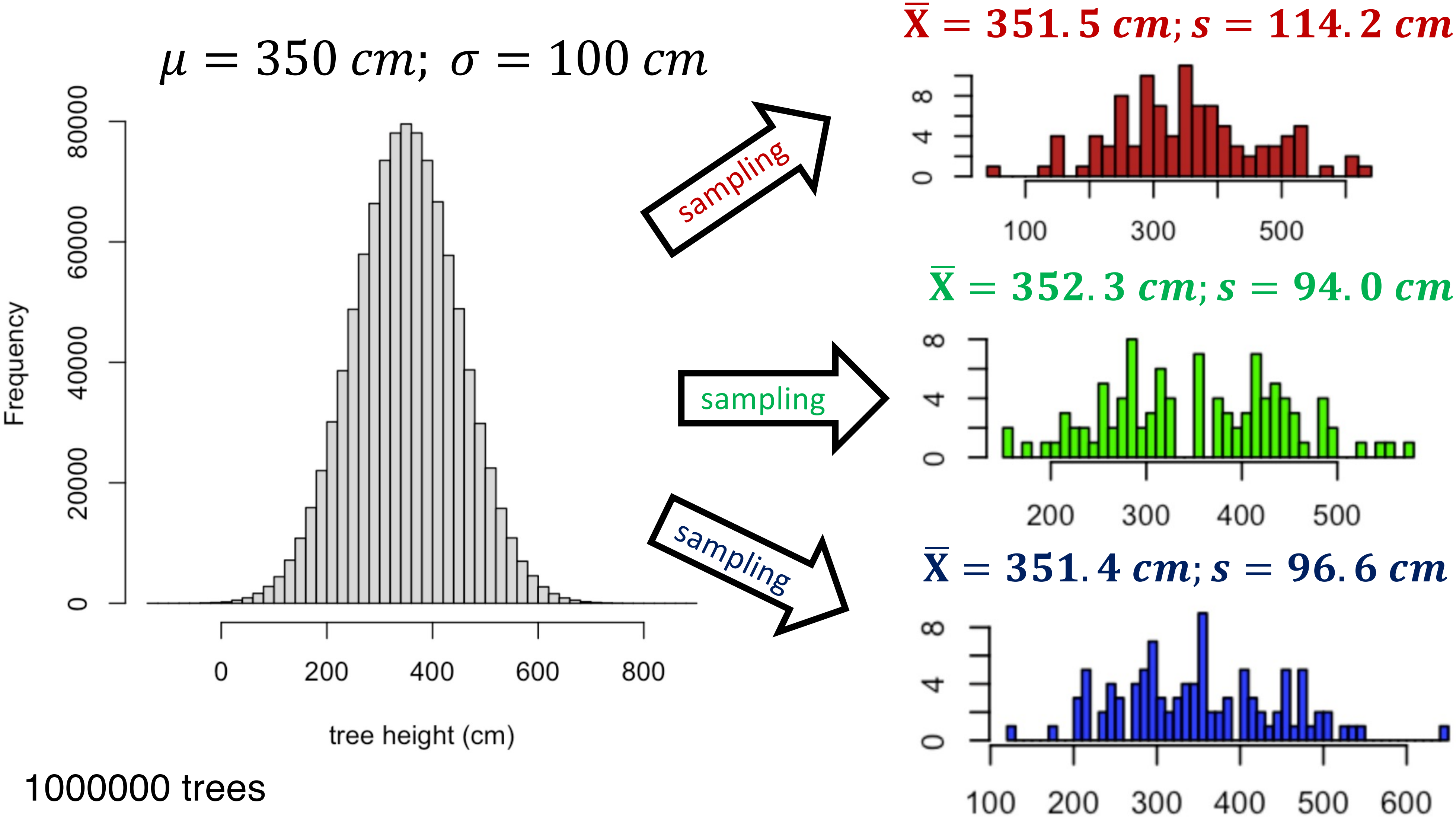
We want to know about a large number of trees

Population

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

$\bar{X} \neq \bar{X}$

Sampling variation: connecting populations and samples



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

1000000 trees

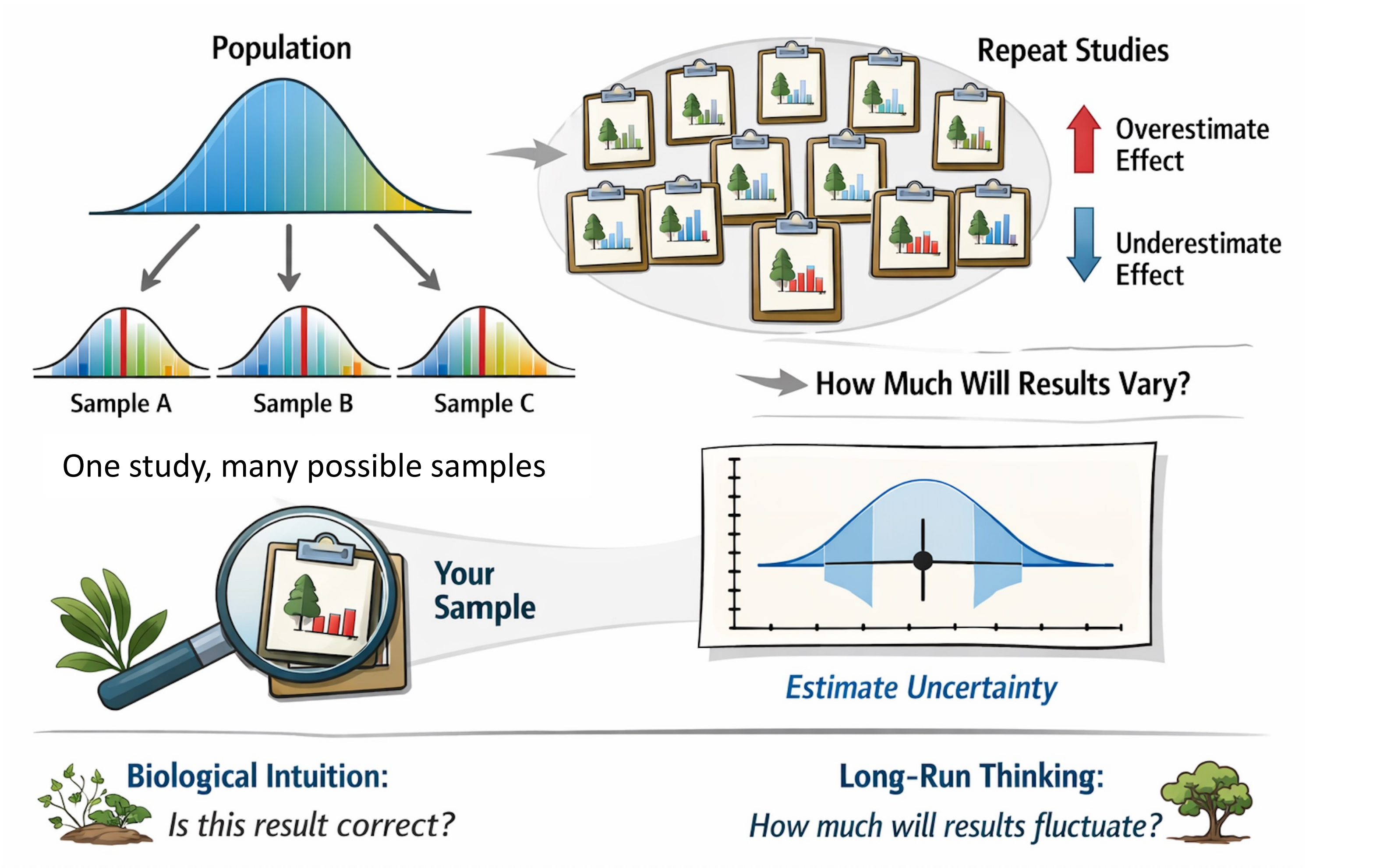
$\bar{\mathbf{X}} = \mathbf{351.5}\ \boldsymbol{cm}; \boldsymbol{s} = \mathbf{114.2}\ \boldsymbol{cm}$

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

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

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.

From one sample to the long run



Although we usually observe only one sample, statistics uses its structure to estimate how much results would vary across these hypothetical repetitions; that is, the uncertainty of our inference.

Why sampling variation is not about a single dataset

Sampling variation is a population-level concept because it describes how estimates and test results vary across many repeated samples, not what happens in a single dataset. Any sample-based statistic (e.g., mean) is just one realization from a larger sampling process.

Long-run thinking means asking what would happen if the same study were repeated many times under identical conditions. Some samples would overestimate the mean, others would underestimate it, purely by chance. Statistics is designed to quantify this variability.

Although we usually observe only one sample, statistics uses its structure to estimate how much results would vary across these hypothetical repetitions; that is, quantify sampling variability to estimate uncertainty of our inference.

This perspective often clashes with biological intuition ("cognitive discomfort"), which focuses on single systems and specific mechanisms. Sampling variation does not tell us whether a sample mean is correct; it tells us how much sample means from the same population are expected to fluctuate and how reliable our inference is in the long run.

Sampling variation: connecting populations and 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)

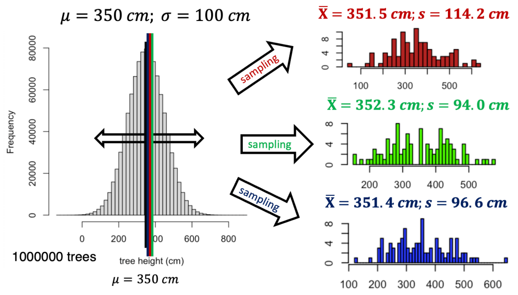
One population, one observed sample - Critical understanding

[1] Sample statistics (e.g., means, variances) almost never equal the corresponding population parameters exactly. This difference arises from sampling variation, often called sampling error.

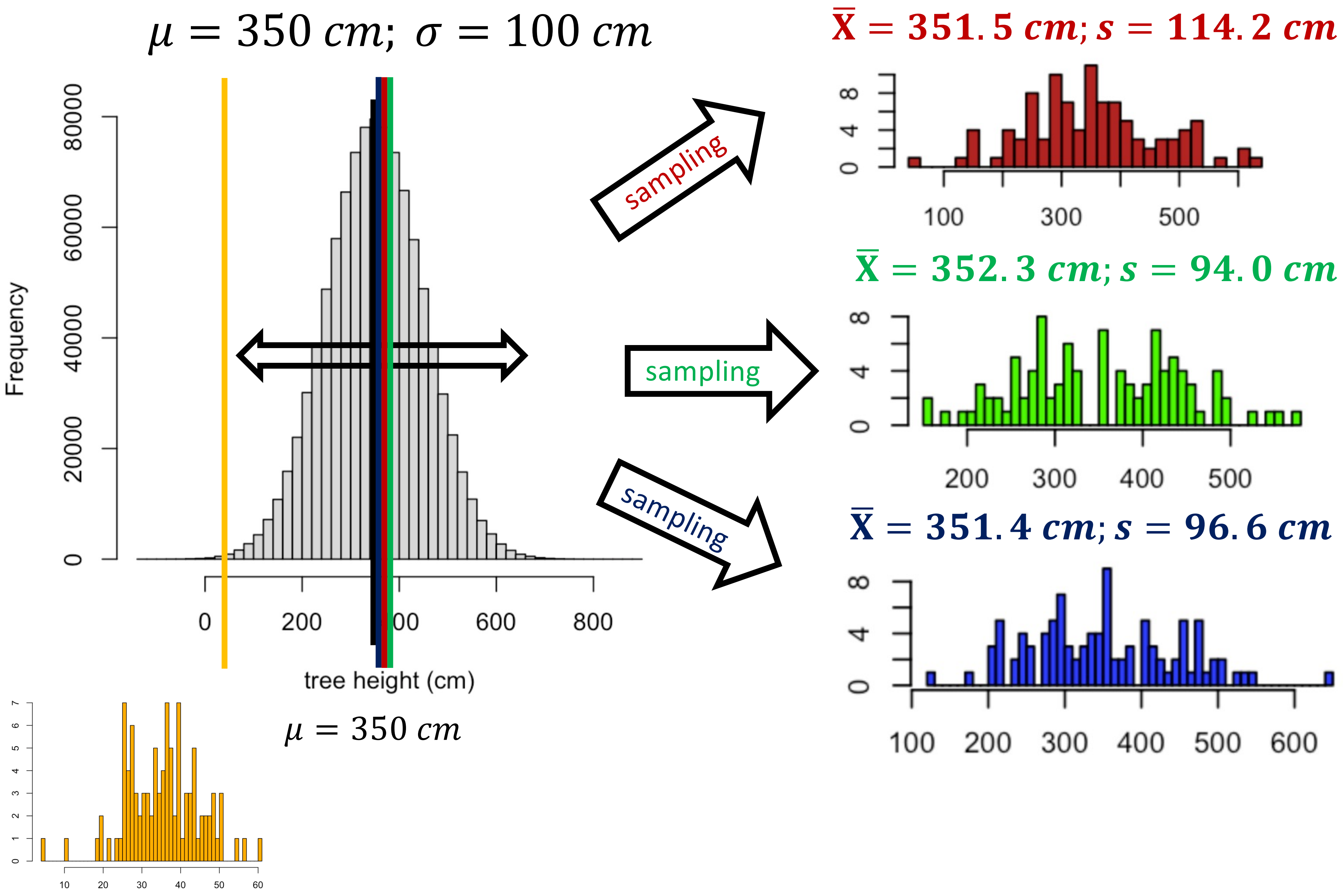
[2] This does not imply that inferences based on samples are incorrect. Sample statistics are estimates of population values, and inference is about assessing how close those estimates are likely to be to the truth.

[3] The quality of an estimate varies: some samples yield statistics close to the true population value, while others are farther away—purely by chance.

[4] Statistics provides tools (e.g., sampling distributions, confidence intervals, hypothesis tests) to quantify this uncertainty, allowing us to judge how reliable our estimates and conclusions are in the long run.



Sampling variation: Some samples produce means close to the population mean, while others fall farther away, reflecting natural variability across samples.



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

$\mu = 350\ cm$

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

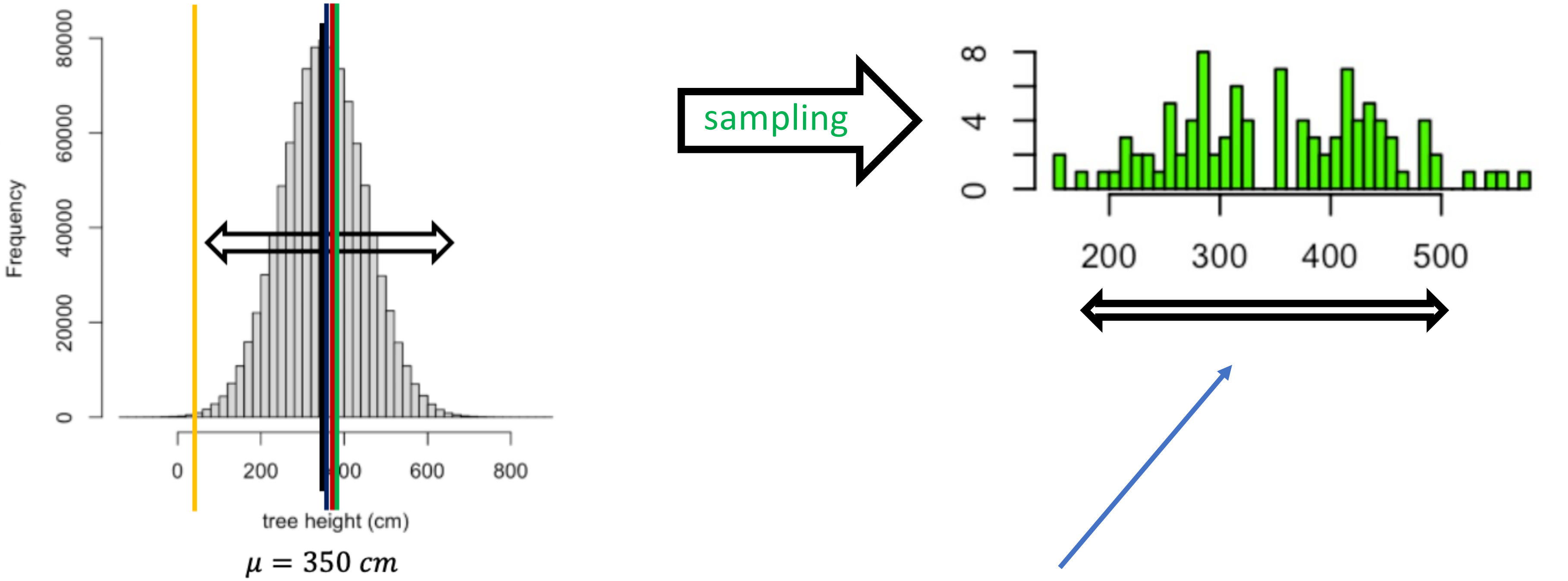
$\bar{\mathbf{X}} = \mathbf{352.3}\ \boldsymbol{cm}; \boldsymbol{s} = \mathbf{94.0}\ \boldsymbol{cm}$

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

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

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

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



As we will see later, the variation among observations within a sample (measured by the standard deviation) helps us estimate how far the sample mean is likely to be from the true population mean, providing a measure of uncertainty.

Key concepts underlying statistics and statistical thinking

- We can never know the true population parameter with certainty; we can only estimate it from sample data.

- Decisions are therefore made using estimates, and the risk of error depends on how close those estimates are to the true value.

- When an estimate is close to the population parameter, the consequences of error are small; when it is far, decisions can be seriously misleading.

- Assessing uncertainty is thus essential—it allows us to evaluate the risk associated with our conclusions.

- Uncertainty arises from sample variability: different samples would produce different estimates.

- Accuracy refers to how close an estimate is to the true population value, while uncertainty tells us how reliable that estimate is.

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

Understanding sampling variation with dance.

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

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.

A sample is biased when certain units in the target population are systematically more or less likely to be selected, leading to unrepresentative data.

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.

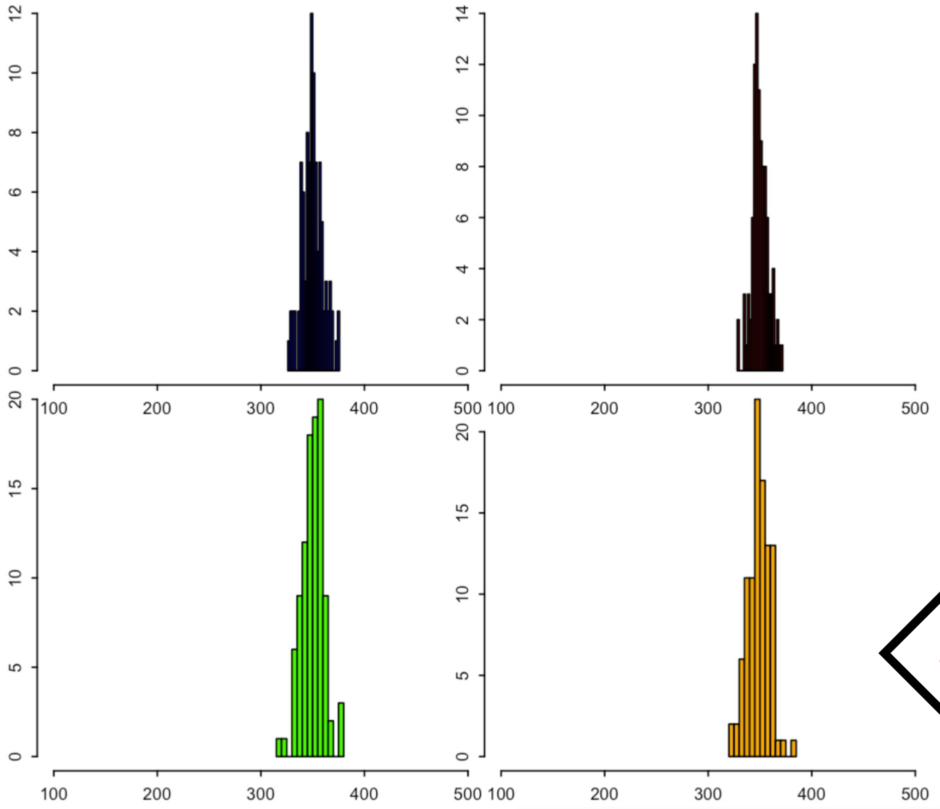
***i.e. = id est* (it is)**

Properties of samples: accuracy and precision

Effective sampling aims to improve both **accuracy**: how close an estimate is to the population parameter; and **precision**: how much estimates vary across samples.

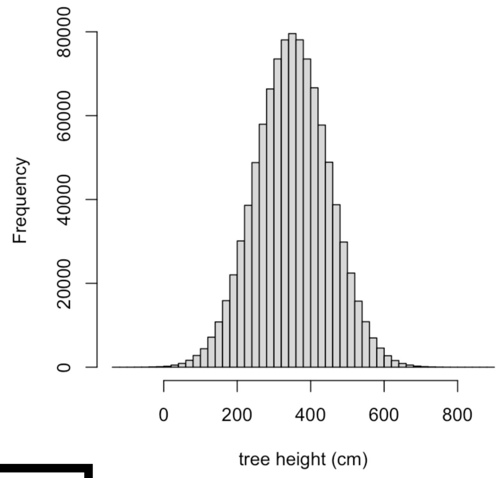
Precise

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).

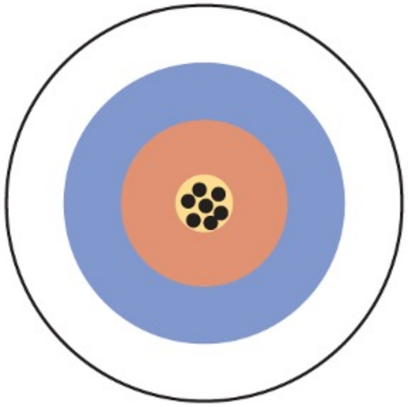
These 4 samples are precise and accurate

accuracy: how close an estimate is to the population parameter

precision: how much estimates vary across samples.

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.

accuracy: how close an estimate is to the population parameter

precision: how much estimates vary across samples.

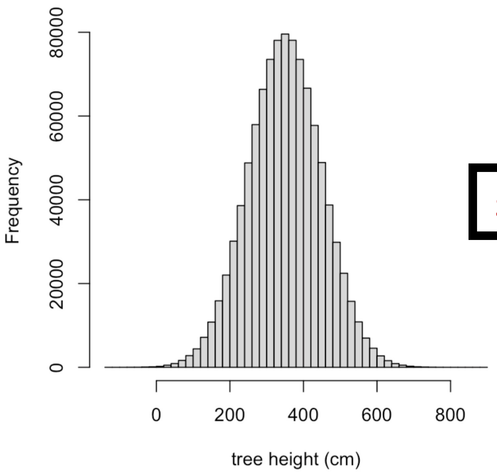
Imprecise

Black line = Population mean
Red line = sample mean

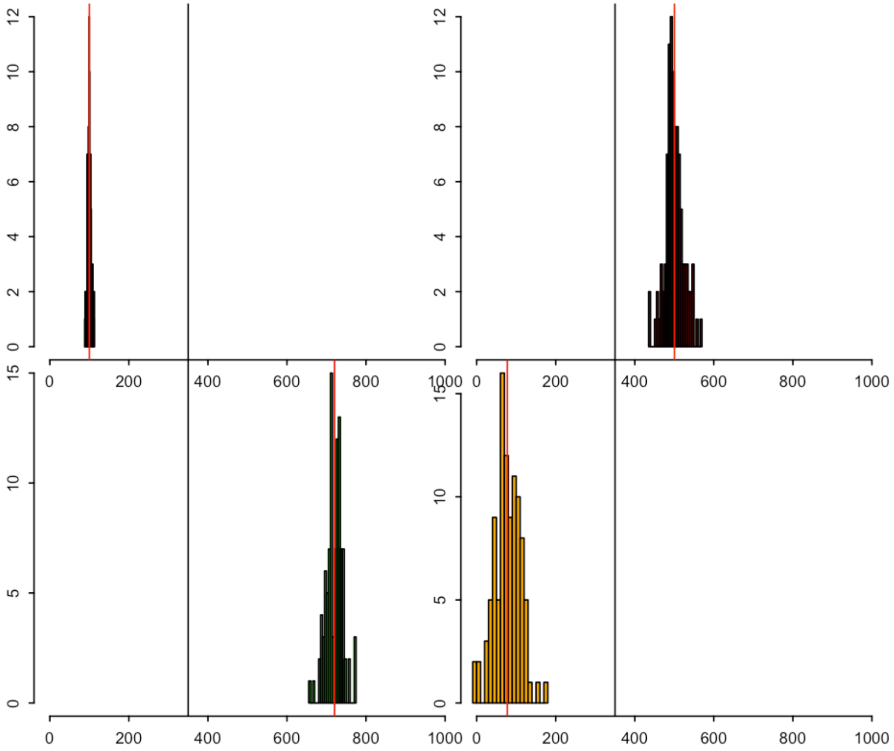
Accurate

Tree population

$$\mu = 350\ cm$$



sampling



1 000 000 trees
(heights in cm)

4 random samples
of 100 trees each

The average sample mean 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

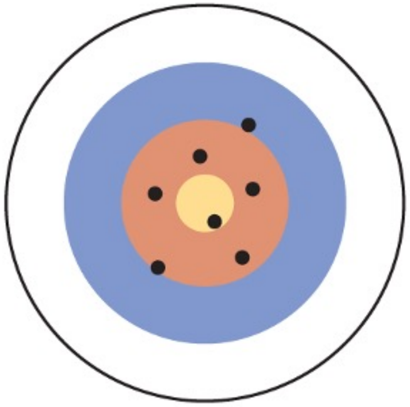
accuracy: how close an estimate is to the population parameter

precision: how much estimates vary across samples.

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.

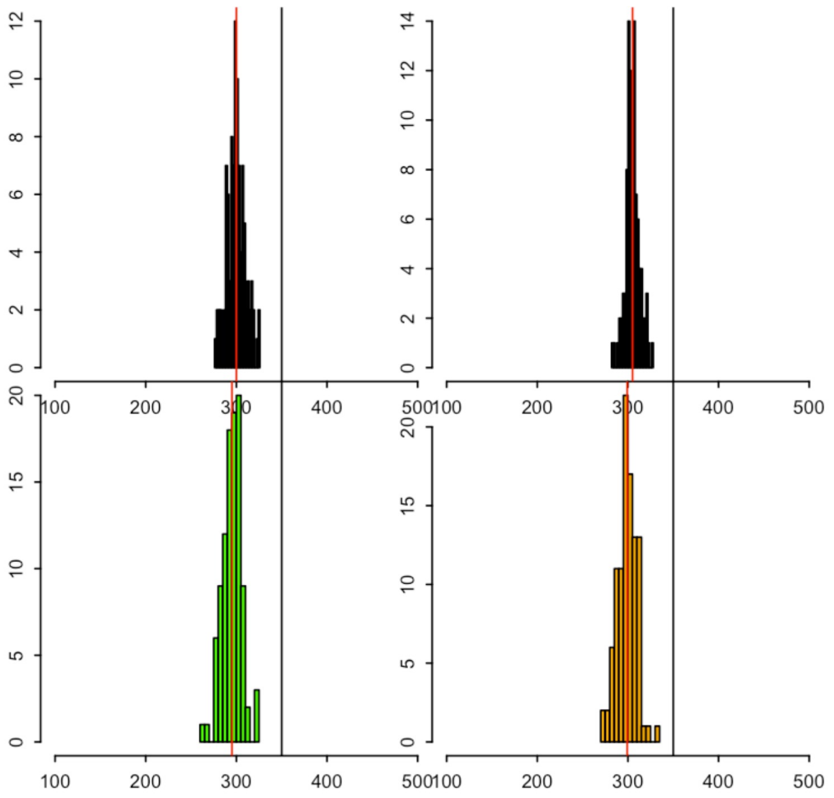
Imprecise = Sample estimates (e.g., sample means) vary widely across repeated samples, regardless of whether they are close to or far from the true population value.

accuracy: how close an estimate is to the population parameter

precision: how much estimates vary across samples.

Precise

Inaccurate

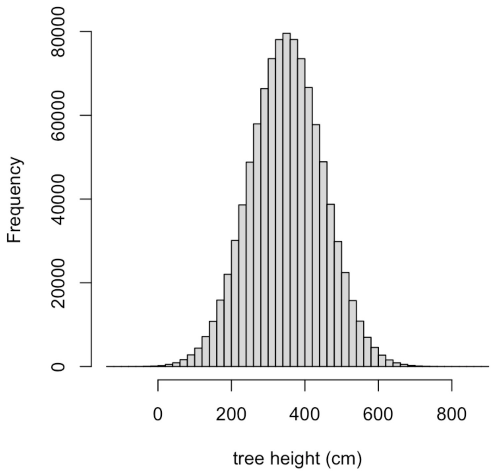


4 random samples of 100 trees each

Black line = Population mean

Red line = sample mean

$\mu = 350\ cm$



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).

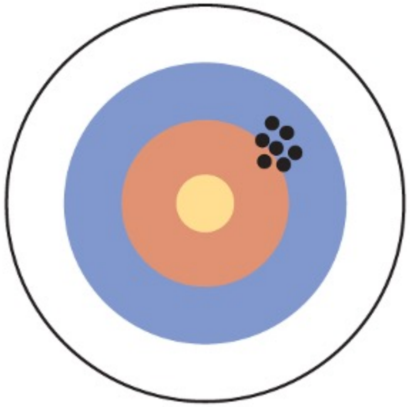
These 4 samples are precise but inaccurate accuracy: how close an estimate is to the population parameter

precision: how much estimates vary across samples.

Precise

Inaccurate

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 = Sample estimates systematically deviate 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.

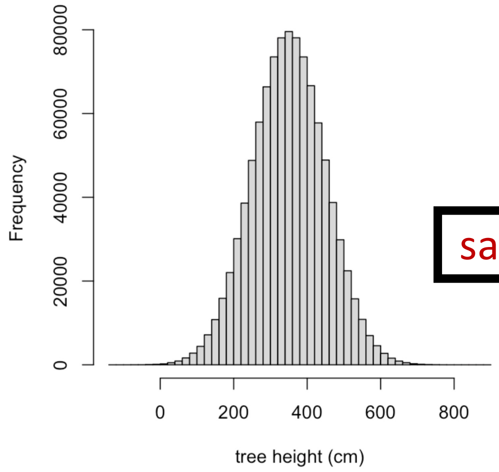
accuracy: how close an estimate is to the population parameter

precision: how much estimates vary across samples.

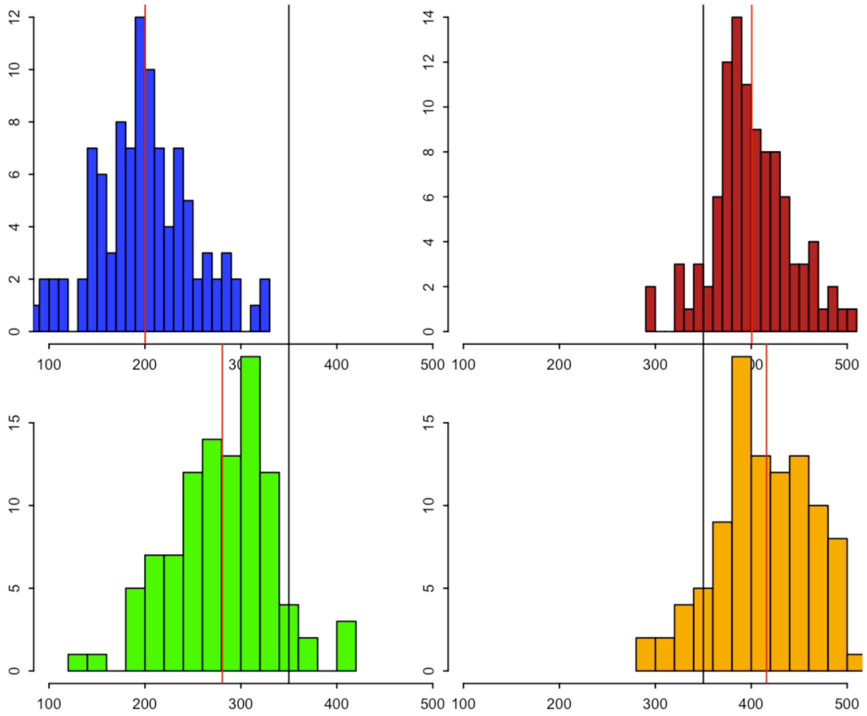
Imprecise

Black line = Population mean
Red line = sample mean

$$\mu = 350\ cm$$



sampling



1 000 000 trees
(heights in cm)

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).

Inaccurate

These 4 samples are imprecise and inaccurate

accuracy: how close an estimate is to the population parameter

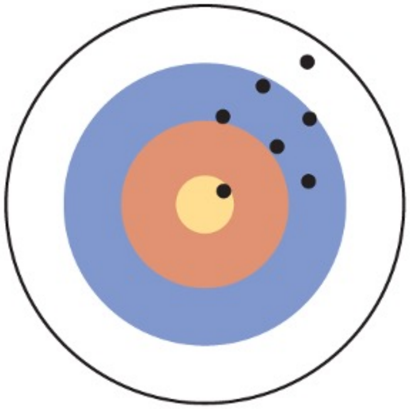
precision: how much estimates vary across samples.

Imprecise

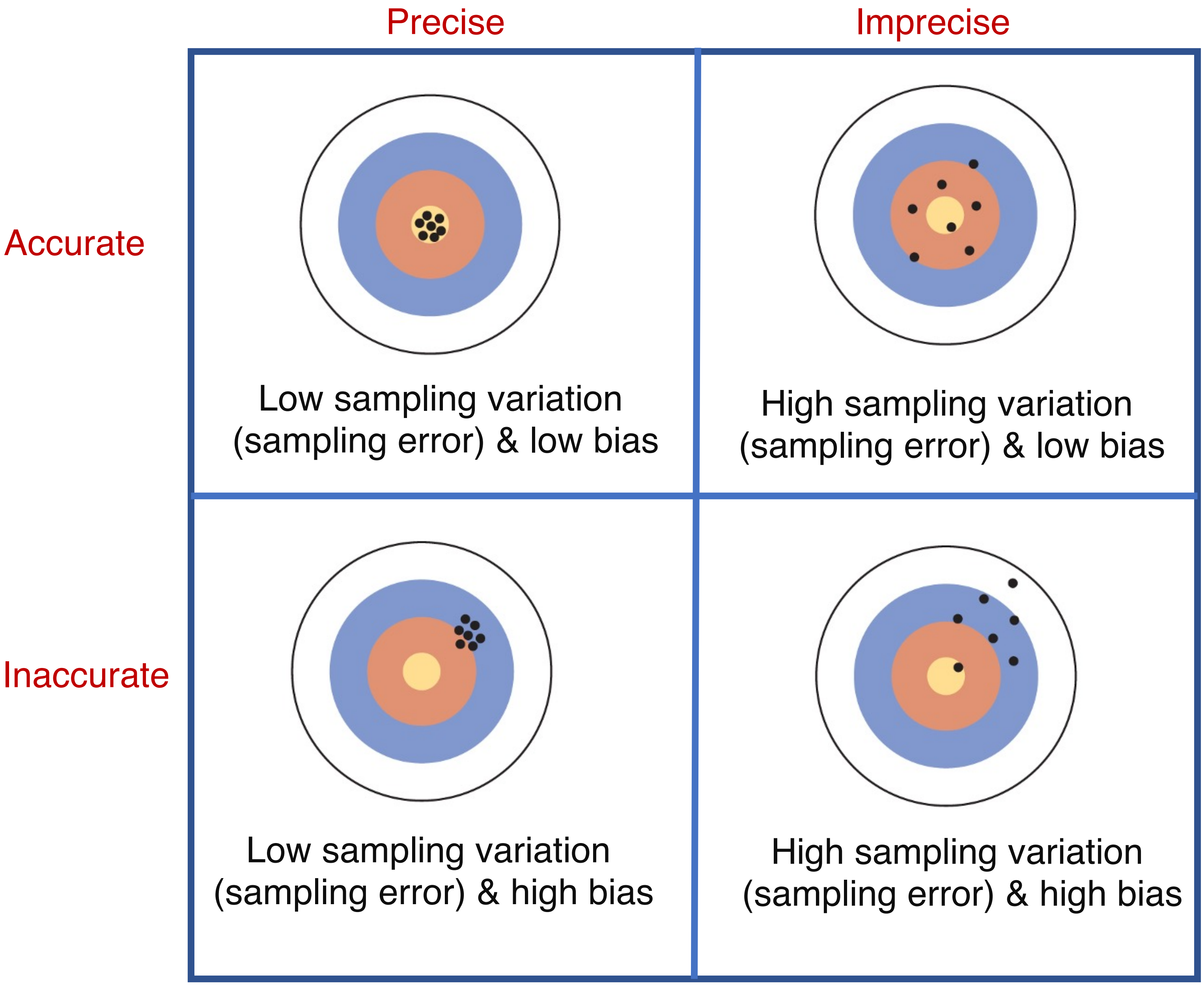
Inaccurate

Inaccurate: Sample estimates systematically deviate from the true population value.

Imprecise: Sample estimates (e.g., sample means) vary widely across repeated samples, regardless of whether they are close to or far from the true population value.



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

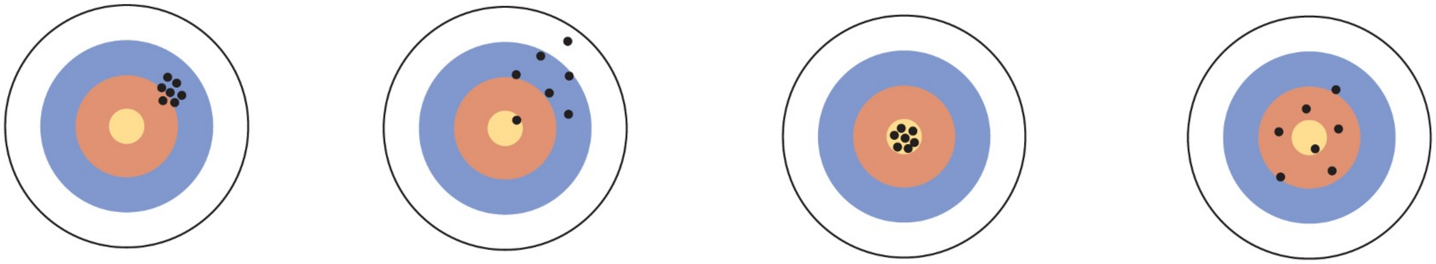


	Precise	Imprecise
Accurate	Low sampling variation (sampling error) & low bias	High sampling variation (sampling error) & low bias
Inaccurate	Low sampling variation (sampling error) & high bias	High sampling variation (sampling error) & high bias

Random sampling: Minimizes bias and allows us to quantify sampling variation (sampling error), which we will cover in upcoming lectures.

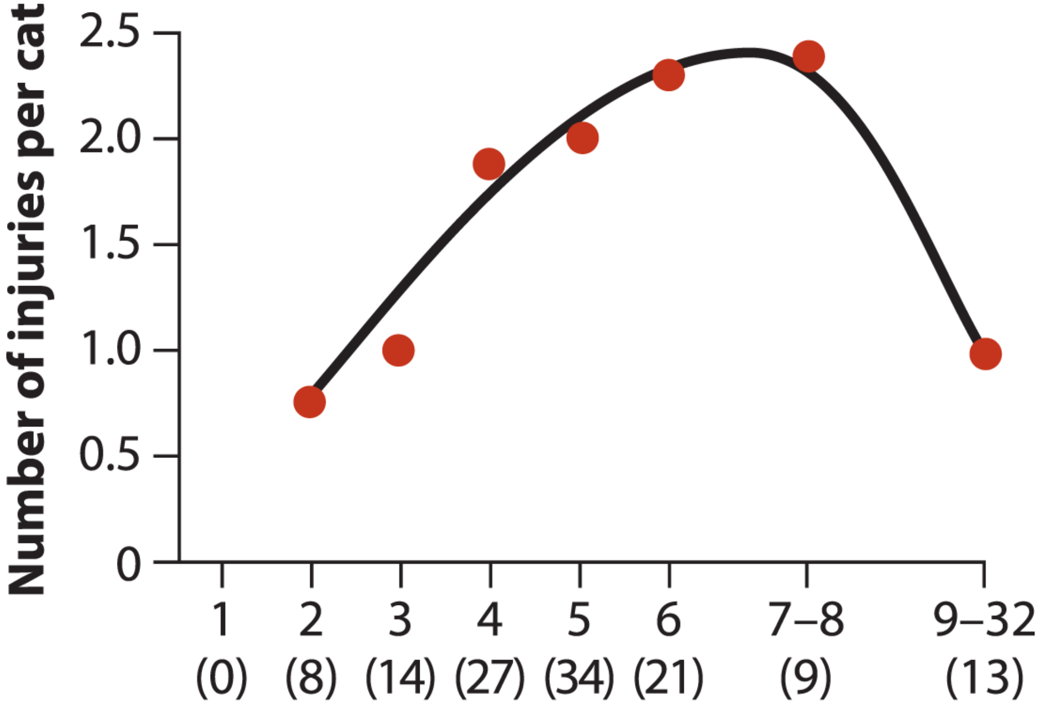
Sample bias: Occurs when some observational units in the target population have a higher or lower probability of being sampled than others, leading to systematically inaccurate estimates of population parameters (true values).

Sampling variation: Refers to the natural variation in a statistic (e.g., the mean) across different samples drawn from the same population. High sampling variation leads to low precision, but when sampling is random, this low precision does not imply inaccuracy on average.



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



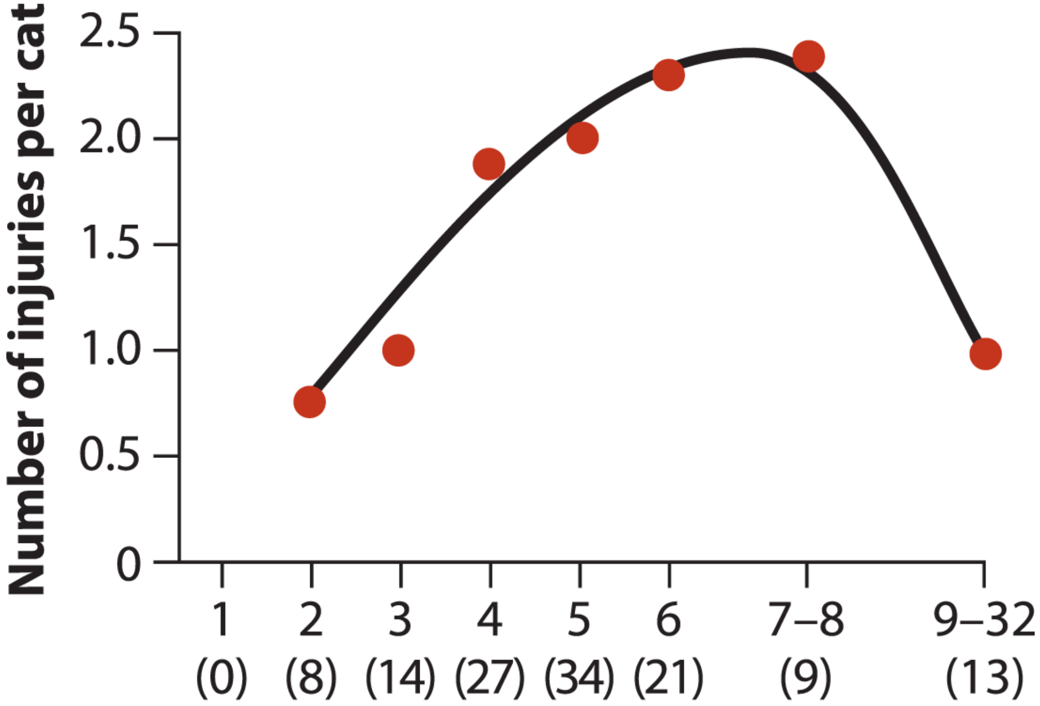
Number of stories fallen

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



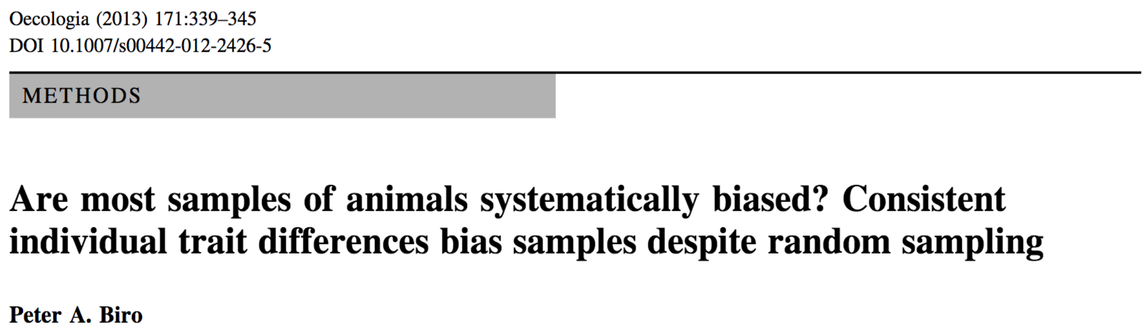
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

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

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

Sample biases due to true biological variation – an example

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



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

Four small, fishless lakes were stocked with equal densities of slow-, intermediate-, and fast-growing fish (known from their genetic background), creating initial populations with known and identical growth-rate distributions.

Fish were sampled using non-size-selective methods (well-established in fisheries), ensuring that body size did not influence capture probability.

Despite this, fast-growing fish were up to twice as likely to be captured, because growth rate was correlated with behavioural traits such as higher moving activity that increased catchability, revealing systematic sampling bias rather than true population differences among lakes.

Sampling populations - what can go wrong?

The case of 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.

Children entered the study through voluntary participation, and this volunteer pool contained more high-risk (non-immune) children than the general population. Parents who believed their child was at higher risk of polio—because the child had not been previously exposed and therefore lacked immunity, had more incentive to participate, hoping for protection from the vaccine.

Randomization protects comparisons within the study by ensuring that the vaccine and control groups have the same baseline risk composition. *Because both groups started from the same volunteer pool, differences in polio infection rates between them can be attributed to the vaccine.*

However, the control group showed a higher polio rate than the general population not because the saline caused harm, but because it revealed the higher baseline risk of the volunteer pool.

Randomization ensures internal validity, but it does not make the study population representative of the population at large.

Sampling populations - what can go wrong?

The case of Volunteer bias

> *The control group showed a higher polio rate than the general population not because the saline caused harm, but because it revealed the higher baseline risk of the volunteer pool.*

The key **consequence** of this finding is that **two different conclusions must be kept separate**:

The causal conclusion is valid:

Because randomization made the vaccine and control groups comparable, the difference in polio rates *between those two groups* correctly shows that the vaccine is effective. This is **internal validity**, and it is protected by randomization.

The estimated effectiveness percentage is numerically valid for the studied individuals, but it is not automatically generalizable to the whole population.

The descriptive comparison to the general population is misleading:

The higher polio rate in the control group compared to the general population **does not** imply that the saline caused harm or that the experiment increased risk. It simply reflects that the study population started out at higher baseline risk due to volunteer bias. This limits **external validity**, i.e., how well results generalize to the population.

Sampling populations - what can go wrong?

The case of Volunteer bias

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);
- More politically engaged, especially in opinion polls or civic surveys, leading to overrepresentation of strong viewpoints.
- More environmentally concerned, in studies on conservation, climate change, or sustainability.
- More physically active, in fitness or exercise studies, because inactive individuals are less likely to volunteer.
- More technologically savvy, in online surveys or app-based studies, excluding people with limited internet access.
- More socially isolated or lonely, in studies offering interaction or attention.

Sampling populations - what can go wrong?

Survivorship biases

Survivorship bias occurs when we draw conclusions based only on the individuals or cases that **remain observable (**i.e., sampled observations), while ignoring those that **did not survive, failed, or disappeared (i.e., unobserved)**.

Because the missing cases are invisible to us, the data we see are systematically unrepresentative. This leads us to overestimate success, resilience, or effectiveness by learning only from what *made it through*.

The famous airplane example (World War II)

During World War II, analysts examined returning military airplanes and mapped where bullet holes appeared. ***Most damage was observed on the wings and tail, leading to the initial suggestion that these areas should be reinforced with armor.***

Planes that returned from missions were not hit in critical areas because hits to those areas—such as the engine, cockpit, fuel system, or control systems—were fatal. Any plane struck there was unlikely to stay airborne and therefore never made it back to be observed. As a result, surviving planes disproportionately show damage in non-critical areas (like wings or tail), where a hit could be tolerated without immediate loss of the aircraft.

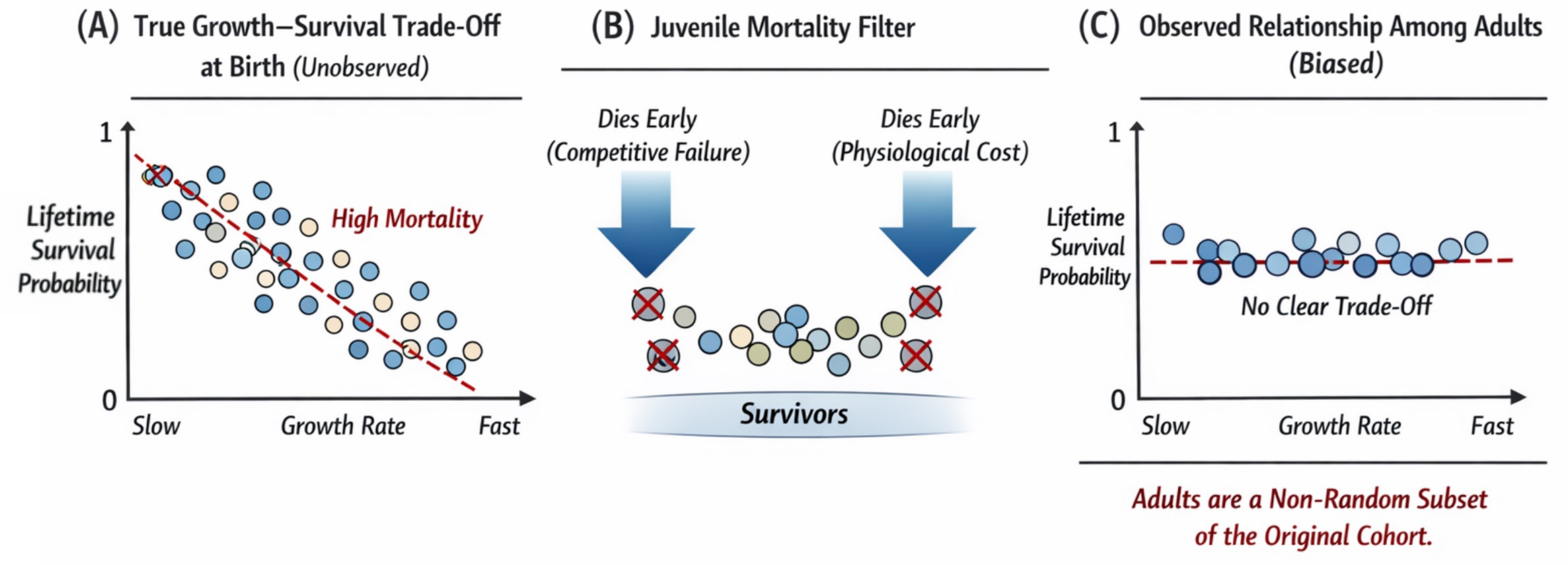
In other words, the pattern of bullet holes on returning planes reflects where planes can be hit and still survive, not where they were most often hit overall. The absence of bullet holes in critical areas is itself evidence that hits there prevented survival and removed those planes from the sample entirely.

Sampling populations - what can go wrong?

Survivorship biases

Estimating growth–survival trade-offs in natural populations

Imagine a study aiming to understand the relationship between **growth rate and longevity** in a wild animal population. Researchers measure growth rates and physiological traits in **adult individuals captured during a breeding season** and conclude that fast growth is not associated with reduced survival (common in many species).


(A) True Growth–Survival Trade-Off at Birth (Unobserved)
1
0
Lifetime Survival Probability
High Mortality
Slow
Growth Rate
Fast
(B) Juvenile Mortality Filter
Dies Early (Competitive Failure)
Dies Early (Physiological Cost)
Survivors
(C) Observed Relationship Among Adults (Biased)
1
0
Lifetime Survival Probability
No Clear Trade-Off
Slow
Growth Rate
Fast
Adults are a Non-Random Subset of the Original Cohort.

Sampling populations - what can go wrong?

Survivorship biases

Estimating growth–survival trade-offs in natural populations

Imagine a study aiming to understand the relationship between **growth rate and longevity** in a wild animal population. Researchers measure growth rates and physiological traits in **adult individuals captured during a breeding season** and conclude that fast growth is not associated with reduced survival (common in many species).

The hidden problem is that individuals with **very fast growth and poor survival prospects may have died earlier in life** and therefore never reached adulthood to be sampled. Likewise, individuals with **very slow growth and low competitive ability may also have died young**. The adults that remain represent a **non-random subset** of the original cohort—those that survived juvenile mortality filters.

As a result, the observed relationship between growth and survival among adults can be very difference compared to the true relationship operating at birth.

There is here a misidentification of trade-offs, concluding that rapid growth has little or no survival cost when, in reality, the cost was paid earlier in life.

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.