Module 1C

Random Variables & Population versus Samples

Module 1 : Descriptive Statistics

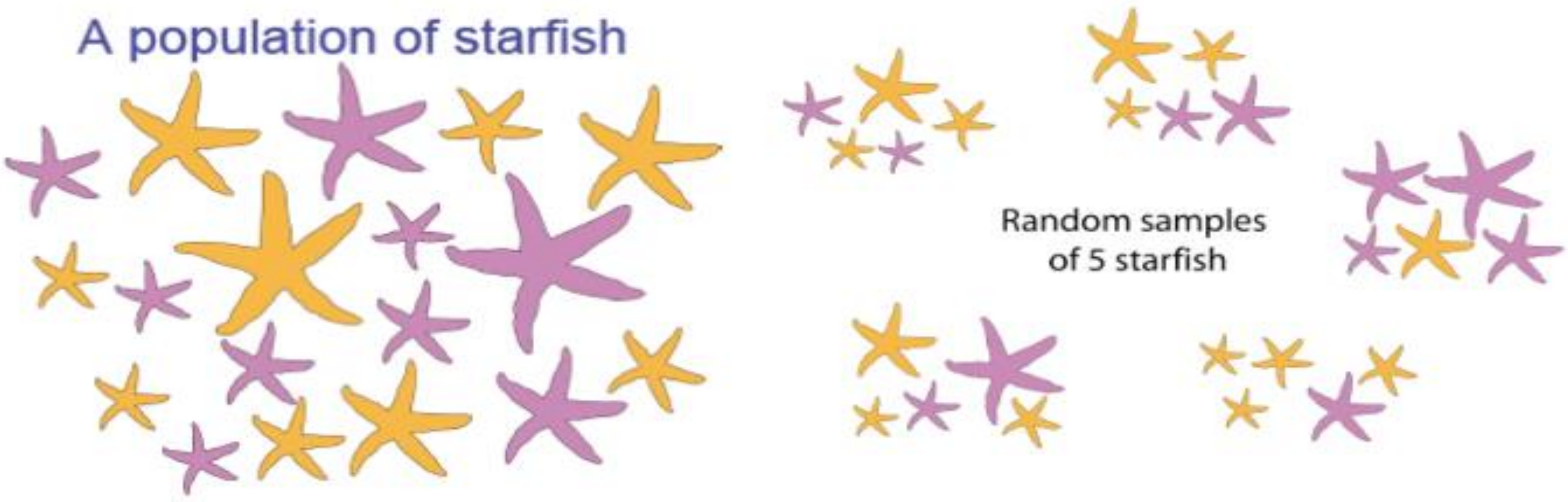
Measurements of ***location*** and ***spread*** of data

<u>Agenda:</u>

- ~~Mean, mode, median~~
- ~~Variability, variation, range~~
- ~~Simpson’s paradox~~
- ~~Intuitions about uncertainty: Fermi Estimation~~
- Accuracy/Bias and Precision/Spread

Random Variables:

- Characteristics measured on individuals drawn from the population
- Value is not constant; it is subject to **VARIATION**
- **Categorical (Nominal, Ordinal)** or **Numeric (Discrete, Continuous)**



Types of data:

Categorical Variable

- AKA Class variables or Nominal variables
- They do not have magnitude on a numerical scale
- **Nominal**
    - Lack inherent order
- **Ordinal**
    - Inherent order **i.e., age (0-18, 19-30, 30-45, etc.)**
- Ex: blood type, genotype, sex, state, survival (live or die), drug treatment (aspirin vs ibuprofen)

Quantitative Variables

- AKA Numerical variables
- Random Variable is a Quantitative variable
- **Continuous**
    - Ability to take any value ex.. Human weight, **age**
    - **They can be measured**
- **Discrete**
    - Spaces between possible values ex. Number of offspring, **age**
    - **They can be counted**

A research team is studying the health and fitness habits of a group of individuals. They collect the following data for each participant:

1. **Resting heart rate (beats per minute)**
2. **Favorite type of exercise (running, swimming, cycling, pilates, etc.)**
3. **Number of hours exercised per week**
4. **Body Mass Index (BMI)**
5. **Member status at a gym (yes or no)**

Which of the following (A, B, C, or D) correct classifies these variables:

A. Resting heart rate: **Nominal**
Favorite exercise: **Ordinal**
Number of hours of exercise per week: **Discrete**
BMI: **Continuous**
Membership status: **Nominal**

B. Resting heart rate: **Continuous**
Favorite exercise: **Nominal**
Number of hours of exercise per week: **Continuous**
BMI: **Continuous**
Membership status: **Categorical**

C. Resting heart rate: **Ordinal**
Favorite exercise: **Nominal**
Number of hours of exercise per week: **Continuous**
BMI: **Ordinal**
Membership status: **Nominal**

D. Resting heart rate: **Discrete**
Favorite exercise: **Continuous**
Number of hours of exercise per week: **Discrete**
BMI: **Continuous**
Membership status: **Ordinal**

Populations

have

PARAMETERS

- Represented by Greek Letters
- **μ; σ**

Samples

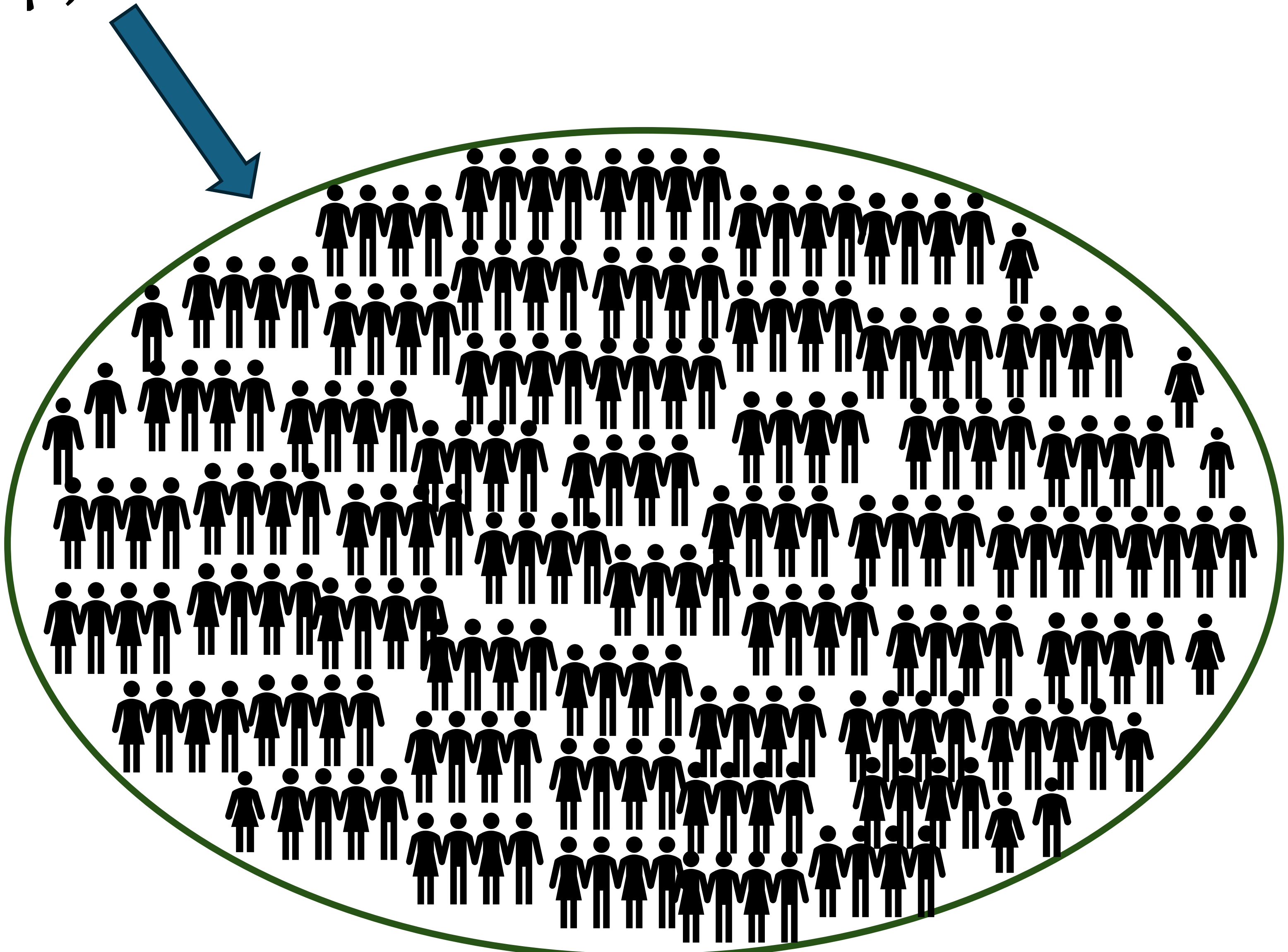
have

E***S***TIMATES

- Represented by Roman Letters
- $\bar{x}$ **; s**

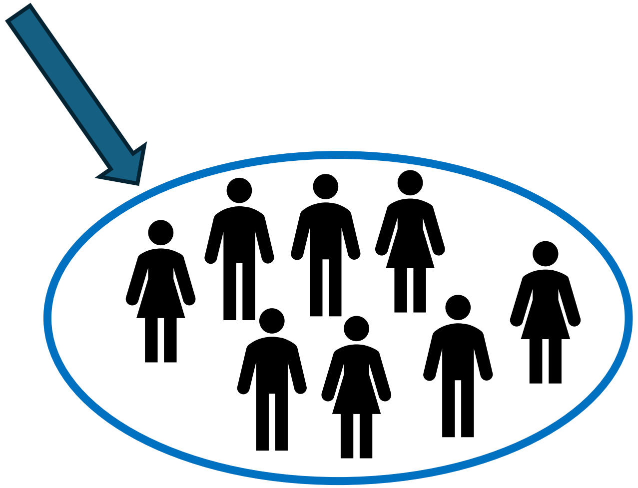
Populations have ***P***ARAMETERS

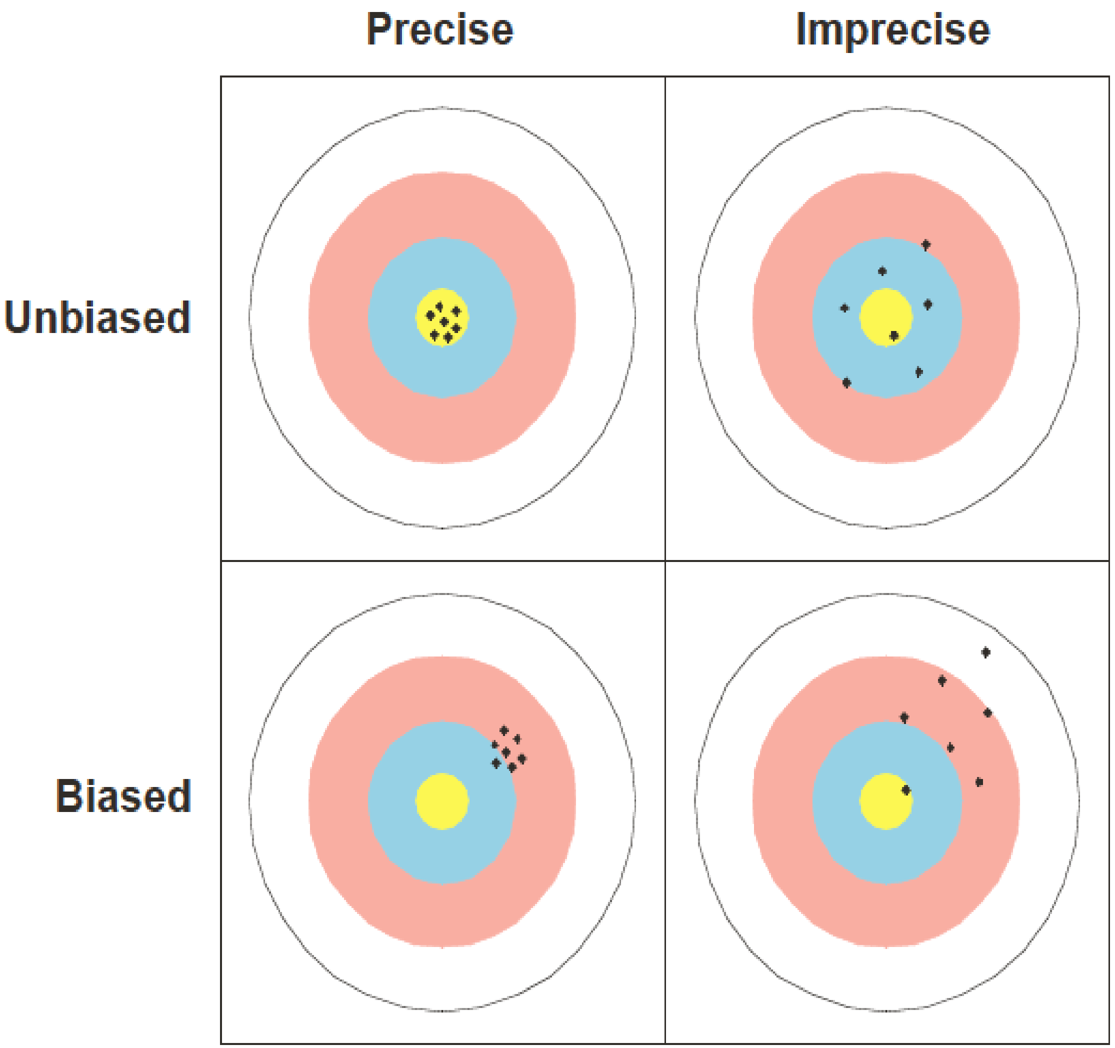
μ; σ



Samples have E***S***TIMATES

$\bar{x}$; **s**





Two major considerations:

1. Accuracy/biased

Bias:

a systematic discrepancy between estimates and the true population characteristic

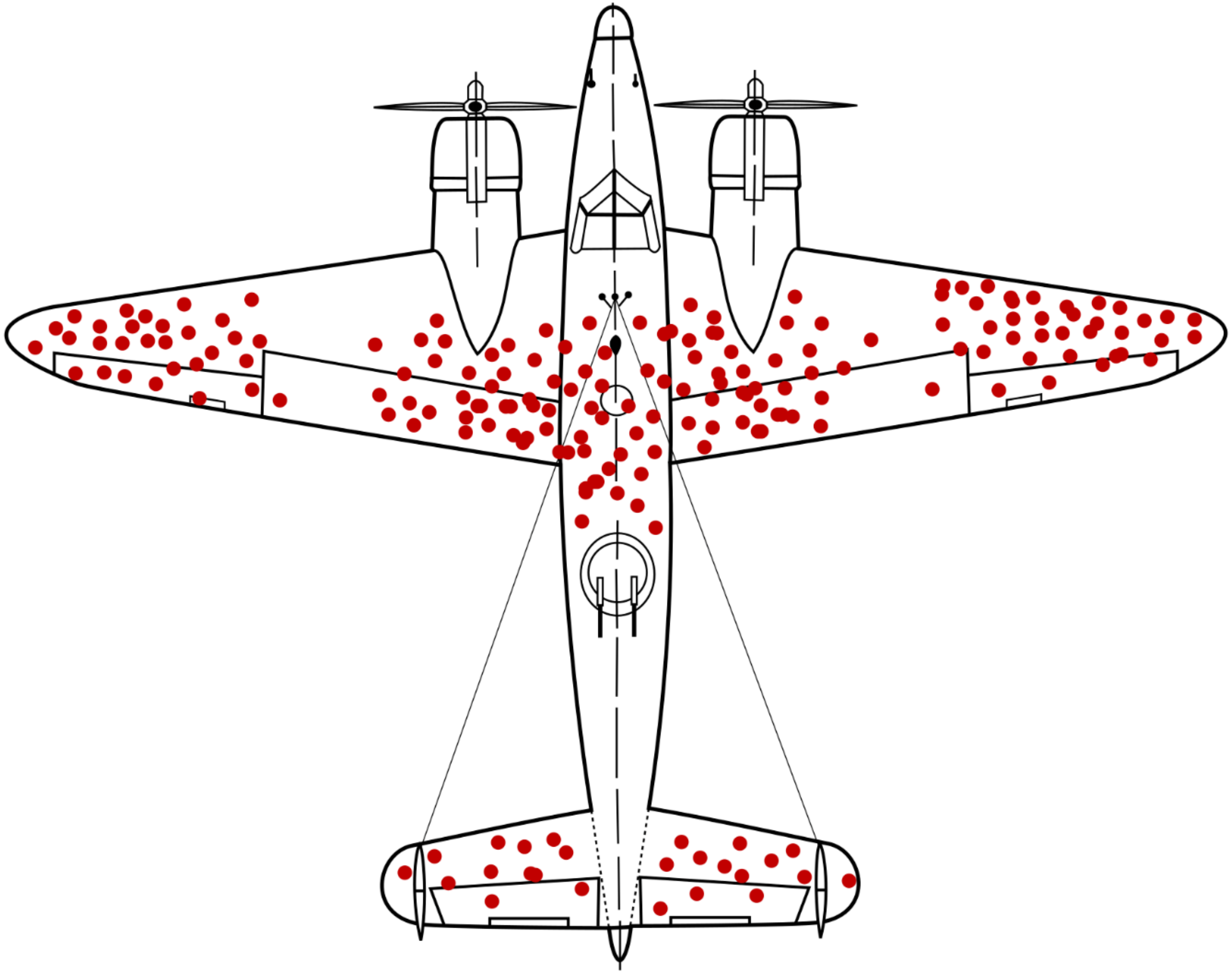
2. Precision/Spread

- Low Sampling Error, high precision

$$SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

To address these, you typically need:

1. A sufficiently large sample
2. Randomly Sampled data points that are independent of each other



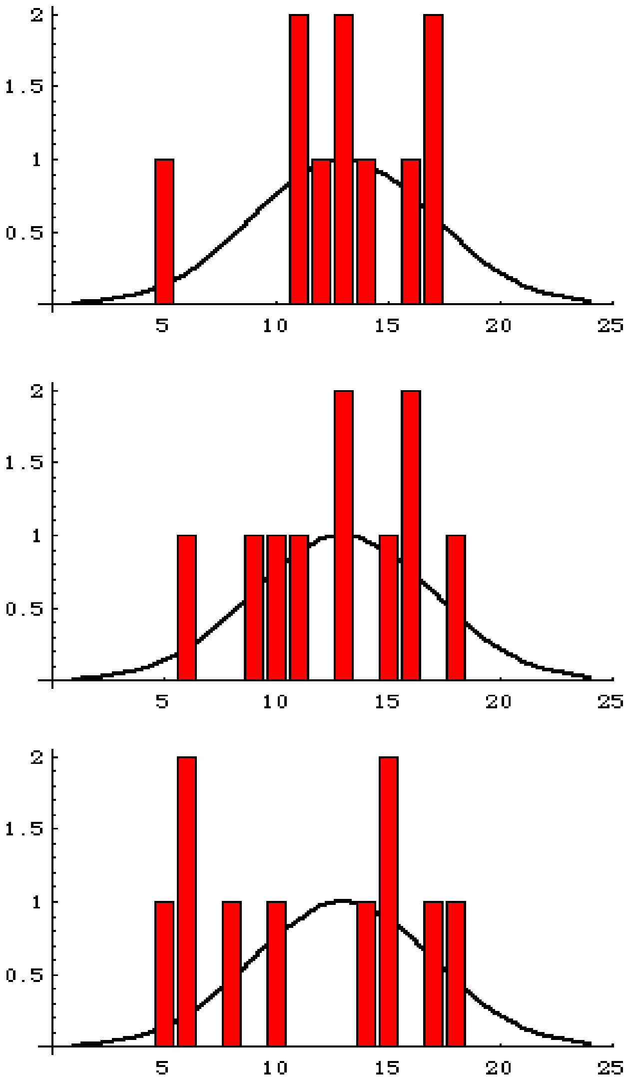
Question: Which of the following statements best describes the difference between accuracy and precision?

A. Accuracy refers to how close measurements are to each other, while precision refers to how close measurements are to the true value.
B. Accuracy refers to how close measurements are to the true value, while precision refers to how consistent measurements are with each other.
C. Accuracy and precision are the same and both refer to how close measurements are to the true value.
D. Accuracy and precision are unrelated to measurements and focus only on data variability.

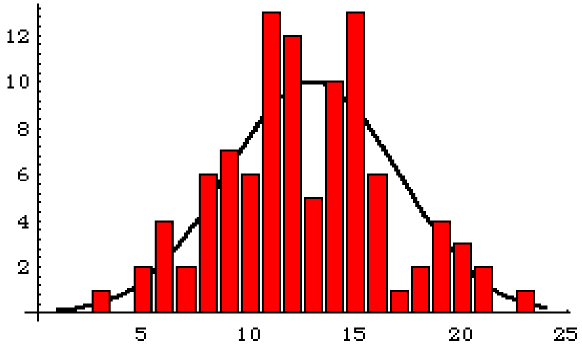
Is it Bias/Accuracy, Variation/Precision?

	Scenario
1	A scale always reads 0.5 grams too high, no matter who uses it.
2	Five repeated pipettings of the same solution yield 1.00, 1.02, 0.98, 1.01, and 0.99 mL.
3	Blood pressure readings vary by ±10 mmHg when measured multiple times on the same subject.
4	A survey systematically oversamples urban participants compared to rural ones.
5	A thermal sensor gives nearly identical readings every time—but all are 2°C too high.
6	A small RNA-seq experiment shows inconsistent fold changes because of low read depth.
7	In a pilot study, different technicians get very similar results from replicate samples.
8	A researcher adjusts their data until it matches an expected pattern.

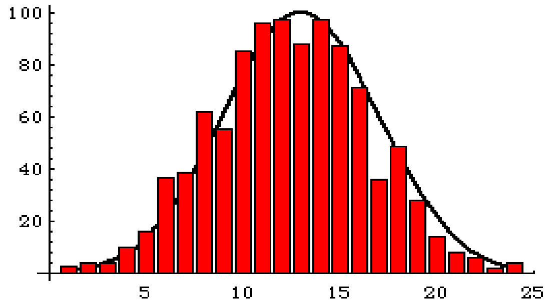
#	Scenario		
1	A scale always reads 0.5 grams too high, no matter who uses it.	**Bias**	Systematic error; accurate shape but wrong center.
2	Five repeated pipettings of the same solution yield 1.00, 1.02, 0.98, 1.01, and 0.99 mL.	**Precision**	High precision (tight grouping) even if mean might be off.
3	Blood pressure readings vary by ±10 mmHg when measured multiple times on the same subject.	**Variability**	Random fluctuation = low precision.
4	A survey systematically oversamples urban participants compared to rural ones.	**Bias**	Sampling bias.
5	A thermal sensor gives nearly identical readings every time—but all are 2°C too high.	**Precision + Bias**	Discuss that precision ≠ accuracy.
6	A small RNA-seq experiment shows inconsistent fold changes because of low read depth.	**Variability**	Low precision due to sampling noise.
7	In a pilot study, different technicians get very similar results from replicate samples.	**Precision**	High precision; good repeatability.
8	A researcher adjusts their data until it matches an expected pattern.	**Bias**	Analytical bias (confirmation bias).



N=10



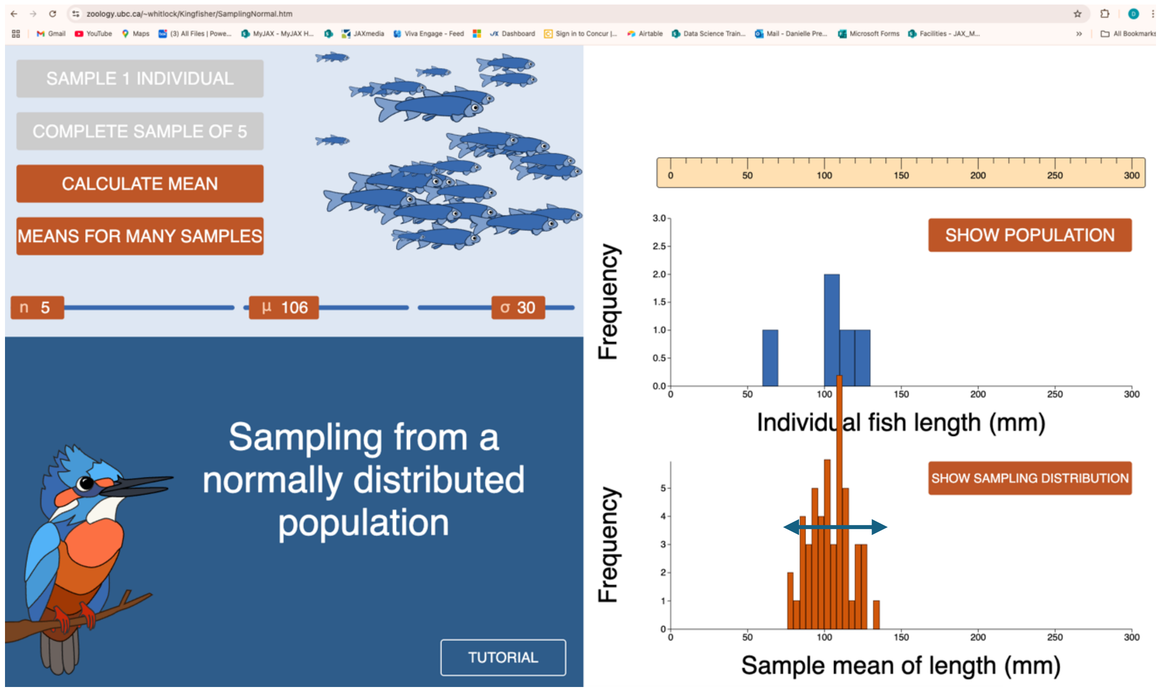
N=100

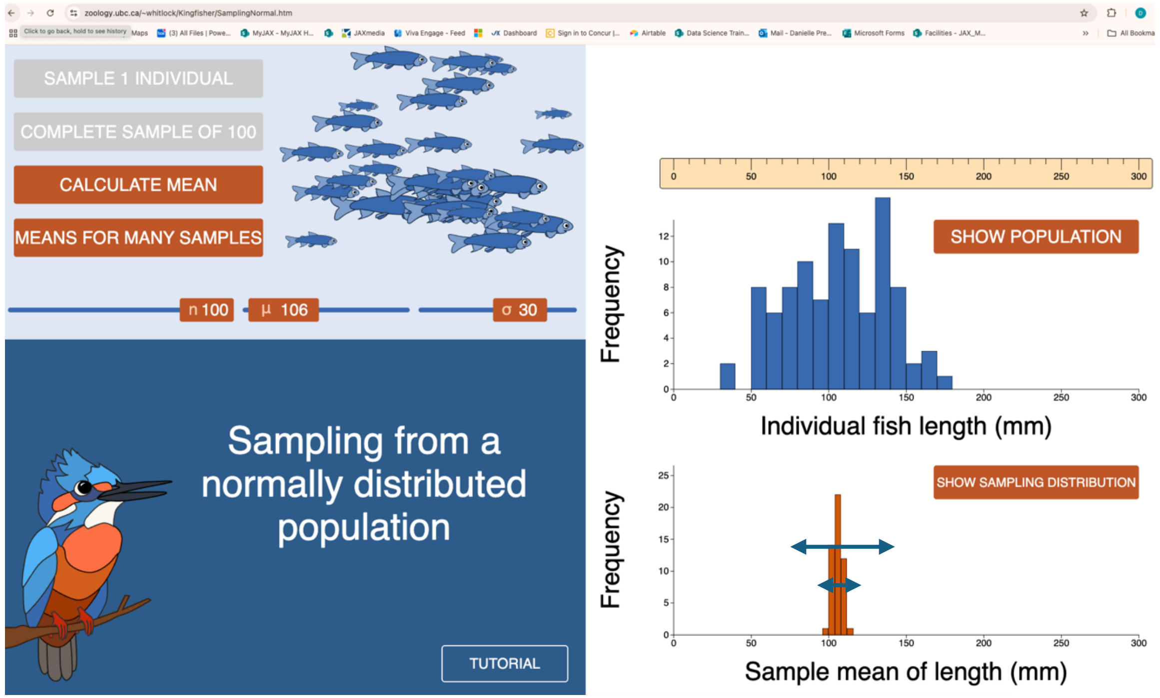


N=1000

n(individual sample sizes) = 10
N is the number of repeats of sample. THIS value ranges from 10 samples to 1000 samples (each one of size 10).

https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm





Many samples, each sample is size 5 individuals

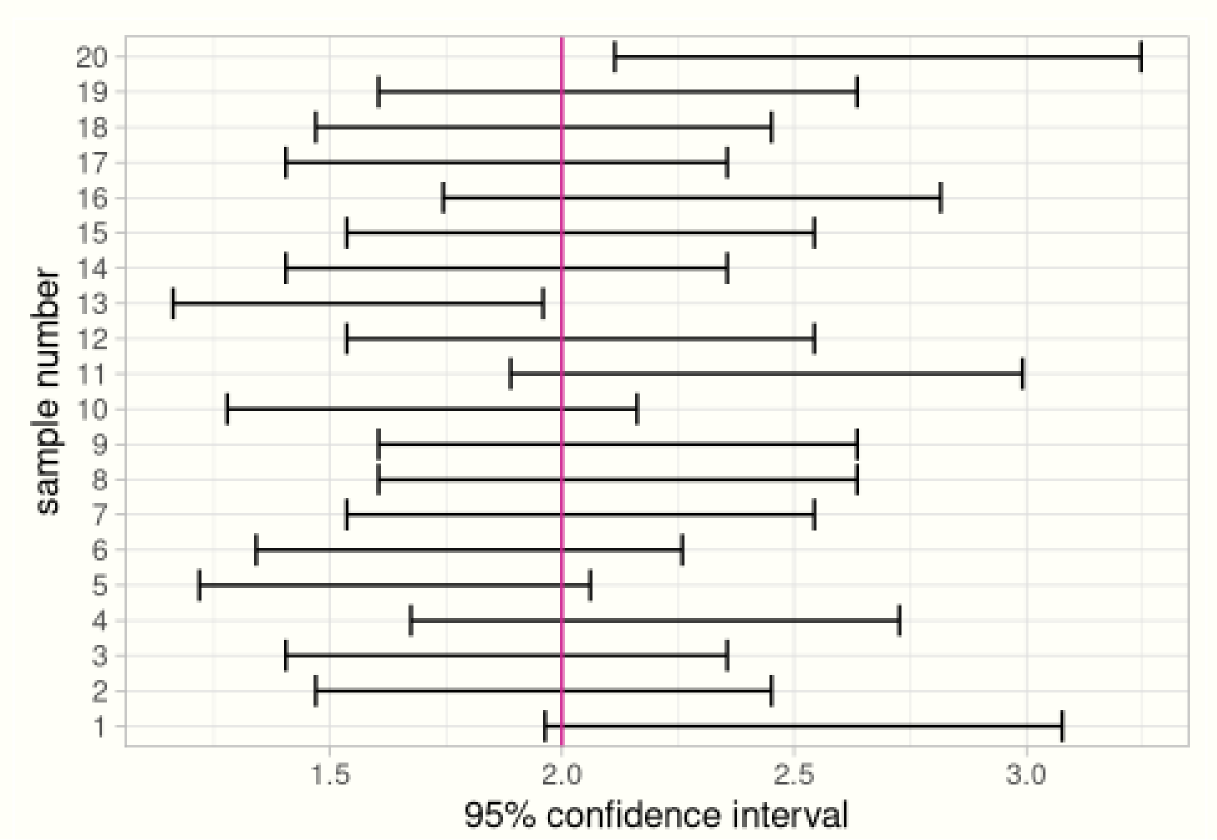
Many samples, each sample is size 100 individuals

Produces a sampling distribution that is
Much narrower than the sampling distribution
produced from n=5, on the left.
(two arrows compare widths)

95% Confidence Intervals

95% Confidence Interval is calculated:
$\bar{x} - 1.96 * SE_{\bar{x}} < \mu < \bar{x} + 1.96 * SE_{\bar{x}}$

We care a lot about precision and sample sizes because (along with alpha and some other assumptions) that is going to create our confidence intervals!



https://www.zoology.ubc.ca/~whitlock/Kingfisher/CIMean.htm

https://stats103.com/confidence-intervals/

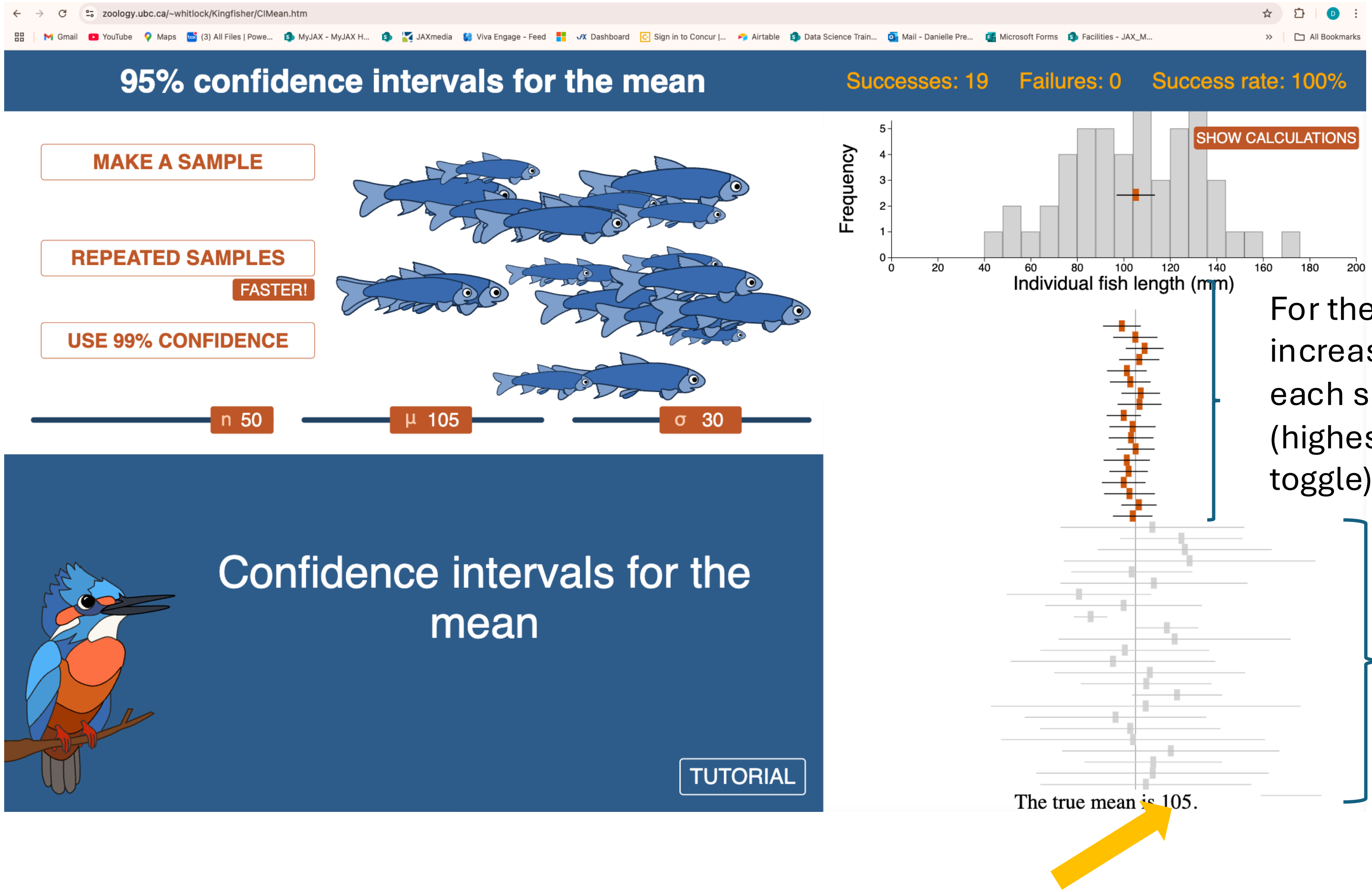
https://onlinestatbook.com/2/estimation/ci_sim.html

95% Confidence Intervals

95% Confidence Interval is calculated:

$$\bar{x} - 1.96 * SE_{\bar{x}} < \mu < \bar{x} + 1.96 * SE_{\bar{x}}$$

https://www.zoology.ubc.ca/~whitlock/Kingfisher/CIMean.htm



For there intervals, I increased n (number in each sample) to be 50 (highest setting on the toggle).

For there intervals, n (number of individuals in each sample) was 5 (the lowest setting on the toggle)

Summary

1. Average:

- mean, median, mode all are legitimate ways of summarizing the average
- They are impacted differently by features of the data set
- Summary statistics, like average, hide a lot of heterogeneity, but are often useful

2. Philosophical core of frequentist statistics (mostly what we use):

We use **samples** to infer information about **populations**

- **Samples** are **noisy.** You estimate a value that jumps around from sample to sample and isn't constant.
- **Populations** have a **TRUE AND CONSTANT PARAMETER VALUE** that you usually don't know (and are thus using samples to estimate the parameter value)

3. Accuracy ("Signal") versus Precision ("Noise")

- **Bias is bad** and almost impossible to fix (try to avoid with good experimental design and sampling protocol)
- **Precision** can be fixed by increasing sample size: