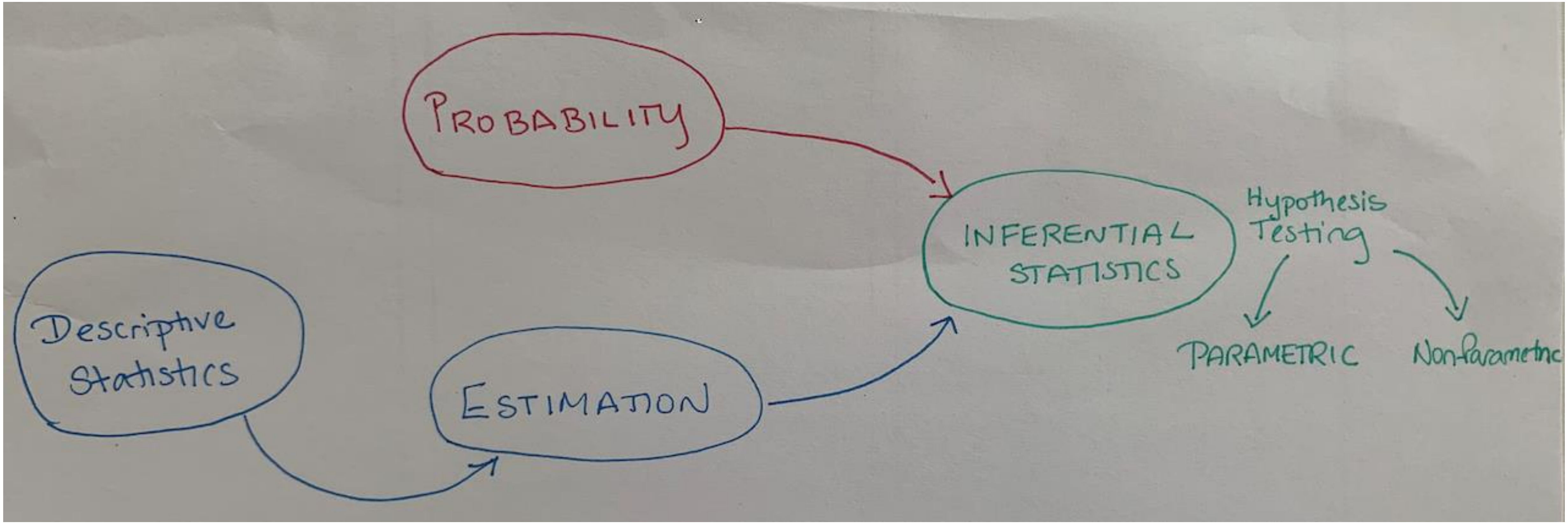

PROBABILITY
Descriptive Statistics
ESTIMATION
INFERENTIAL STATISTICS
Hypothesis Testing
PARAMETRIC
NonParametric

Module 1A : Descriptive Statistics

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

<u>Agenda:</u>

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

You are considering buying a house in a certain neighbourhood. You find a potential house and, to appeal to perceived snobbiness as you are making your decision, your realtor mentions that the **average income in this neighbourhood is $100,000 per year.**

You buy the house.

A year later, the same realtor knocks on your door, this time acting as a representative of the neighbourhood taxpayers' association. He would like you to sign a petition to decrease property taxes because, he says, the residents can't afford an increase in property taxes since the **average family income in the neighbourhood is only $25,000 per year.**

How is this possible, if the realtor is telling the truth, and no one in the neighbourhood has moved or changed jobs in the last year?

The two common descriptions of data:

1. Location:

- Central Tendency
- Where is the weight of the data?

Average

2. Spread:

- How far apart are the data points? Especially: how far apart are the largest and smallest data points?

Range

You will also see:

1. **Skew –** The third standardized moment; positive or negative skew. The shape of the distribution is not symmetric.
2. **Kurtosis –** The fourth standardized moment; sort of ‘peakness’ of the distribution (fatness of the tails)

A story about central location of the data

Waiter	$35,000
Cook	$30,000
Dishwasher	$25,000
Customer 1	$80,000
Customer 2	$50,000
Customer 3	$30,000
Customer 4	$45,000

“Average” is approx. **$42,143**

“Average” is $**125,000,037**

Waiter	$35,000
Cook	$30,000
Dishwasher	$25,000
Customer 1	$80,000
Customer 2	$50,000
Customer 3	$30,000
Customer 4	$45,000
Software or Social Engineer	$1,000,000,000

$35,000
$30,000
$25,000
$80,000
$50,000
$30,000
$45,000

Reorder data

$25,000
$30,000
$30,000
$35,000
$45,000
$50,000
$80,000

(Arithmetic) **Mean** = $\frac{\sum_1^n x_i}{n}$

Median = middle value (odd), mean of middle value (even)

Mode = most frequent value

$35,000
$30,000
$25,000
$80,000
$50,000
$30,000
$45,000
$1,000,000,000

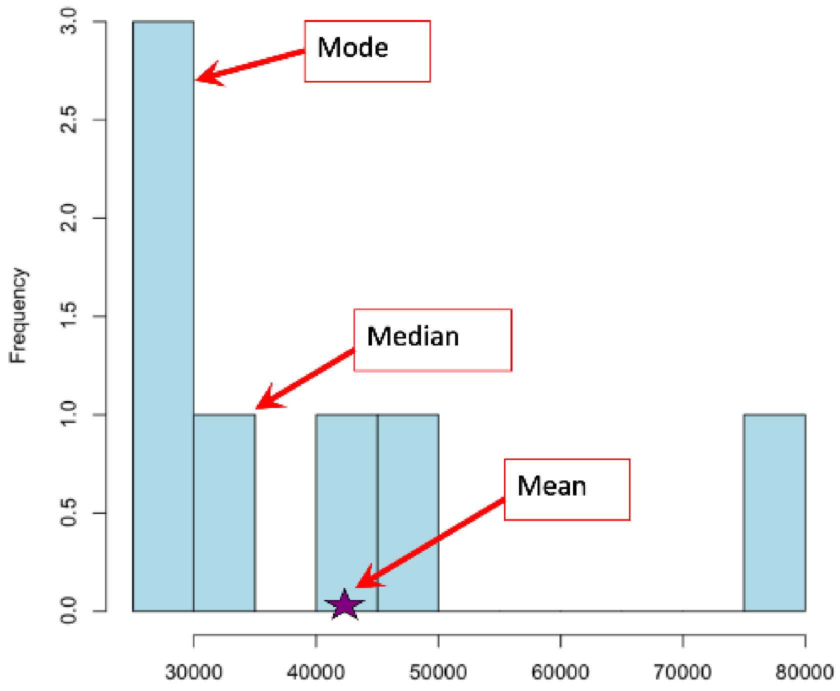
Reorder data

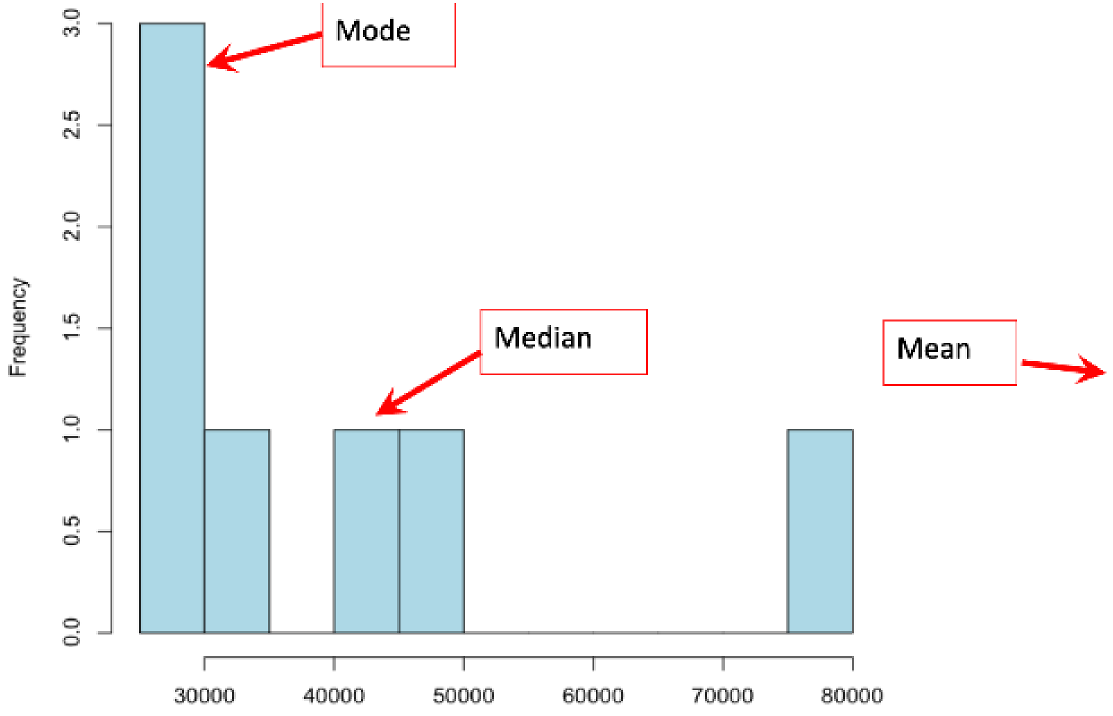
$25,000
$30,000
$30,000
$35,000
$45,000
$50,000
$80,000
$1,000,000,000

	Scenario 1	Scenario 2
mean	$42 143	$125,000,037
median	$35,000	$40,000
mode	$30,000	$30,000

Mean, Mode, and Median can give you different information and they have different benefits

- If the data are skewed or have an outlier, median is often a fairer reflection of the data
- Median can give quick information abut the data without having to calculate anything
- (arithmetic) mean can be a theoretical abstract (2.2 children per woman doesn’t actually exist), but it allows you to use normal distribution to answer questions about the whole population





Will Rogers Phenomenon

"When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states."

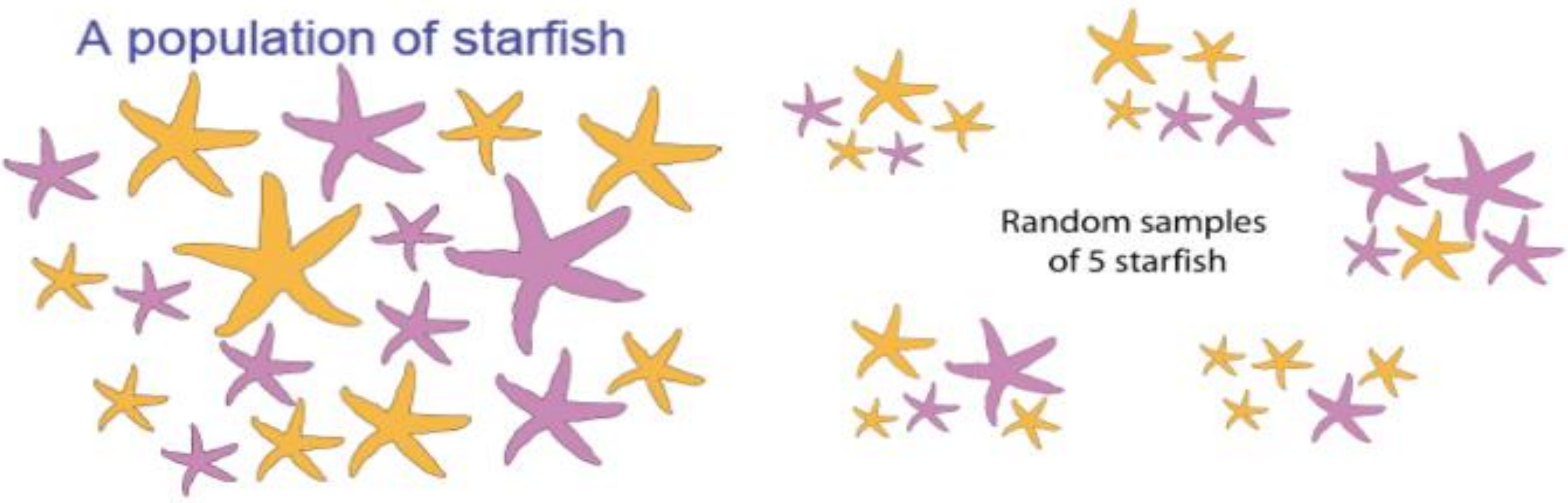
https://en.wikipedia.org/wiki/Will_Rogers_phenomenon

Actual medical phenomenon: "medical stage migration"

<u>Example:</u> There were more COVID deaths among the vaccinated than the unvaccinated. In September 2022, 12,593 COVID deaths occurred in the United States. Of those, 39% were unvaccinated, while 61% were vaccinated. WHY?

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**

A story about spread (and shift of location) of the data

Spread of Data:

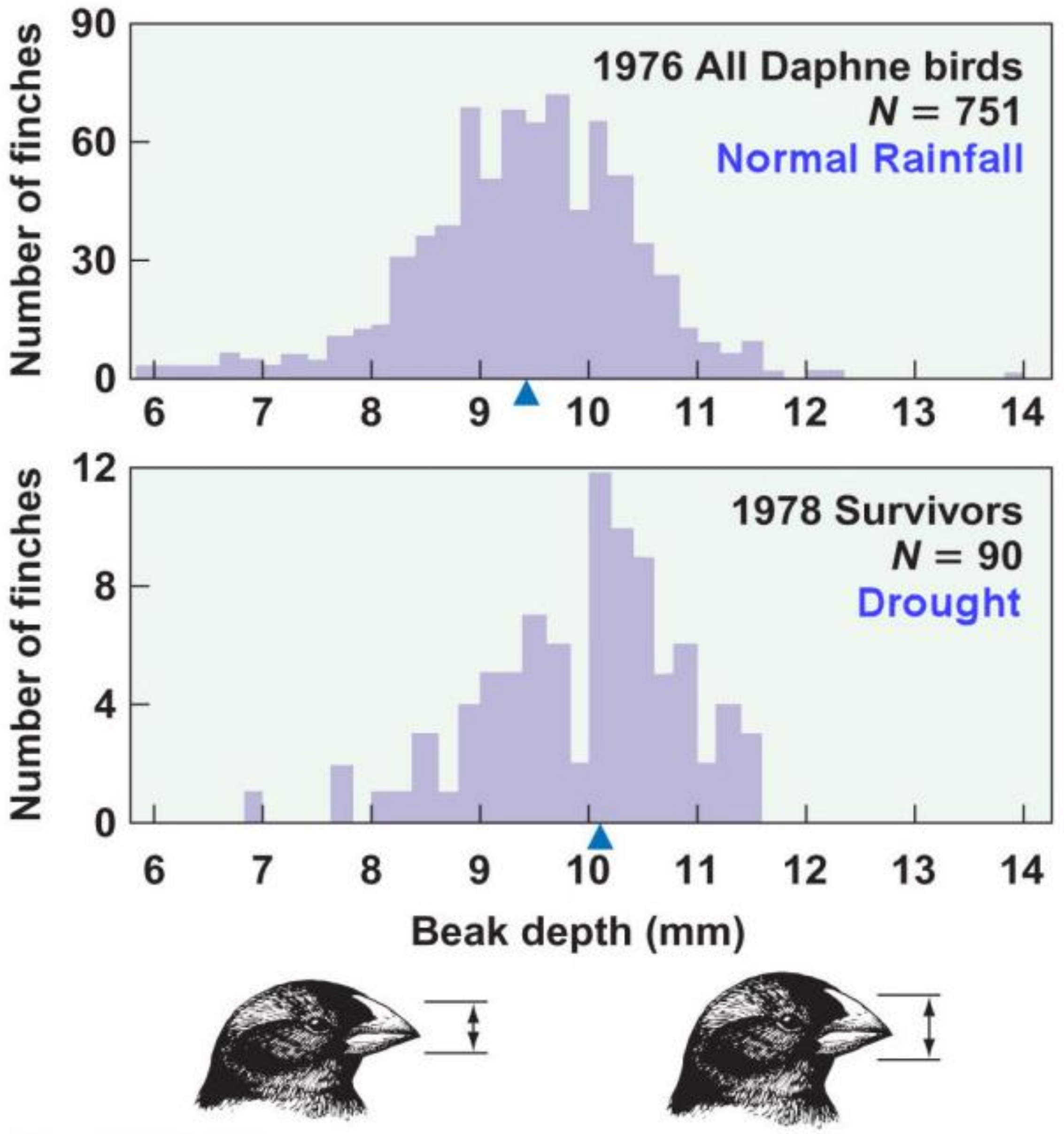
1. Variance
$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

2. Standard Deviation
   - Same units as data
   - σ

3. Range
   - largest – smallest value

4. Interquartile Range
   - 25^{th} to 75^{th} percentile

Peter and Rosemary Grant and the Ongoing Evolution of Galapagos Finches



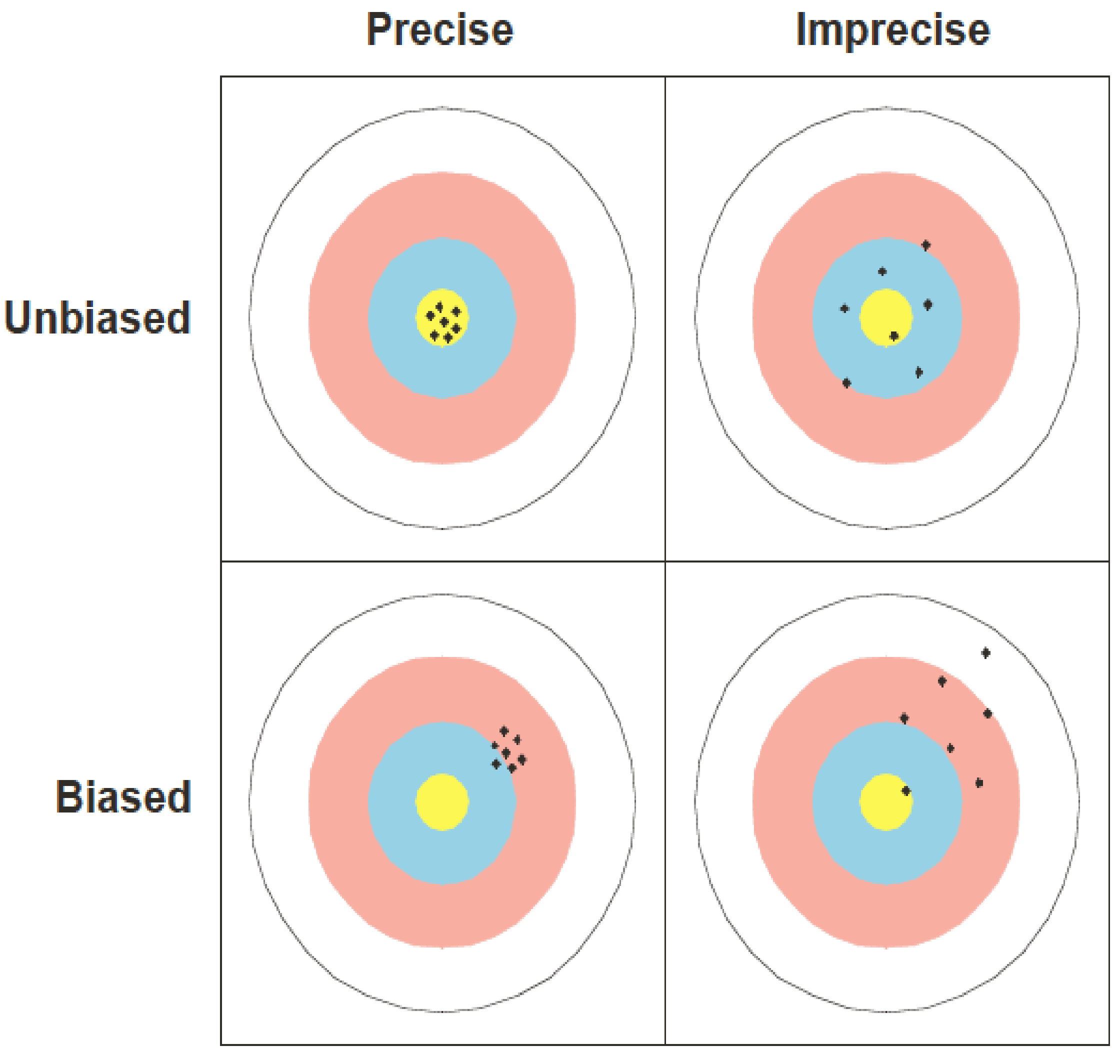
Fermi Estimation*

- A good way to practice so you aren't bamboozled so easily
- https://www.njaapt.org/resources/Documents/Physics Olympics -- All/Fermi Questions - Worksheet and Answers.pdf

- **Question:** Genes are composed of exons (and introns) and all the exons in the genome comprise the exome. Using a Fermi estimation, what is a reasonable estimate for the size of the human genome?

* won't usually be time to discuss this, but the link to how to improve/justify your estimation is here in case you want to use it as a party trick

What Makes a ’good’ sample?



Two major considerations:

1. **Accuracy/biased**

Bias:
a systematic discrepancy between estimates and the true population characteristic

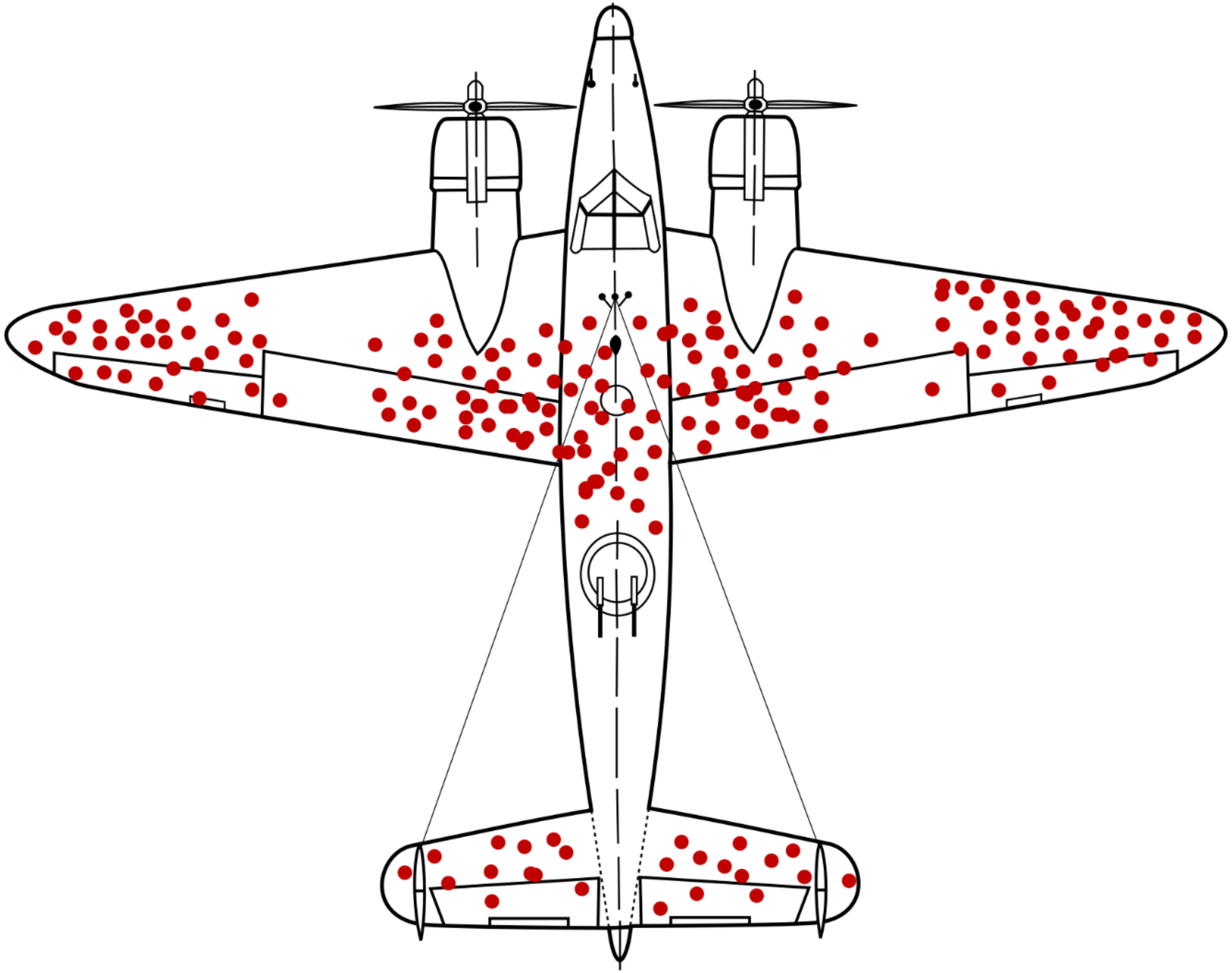
2. **Precision/Spread**

- 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.

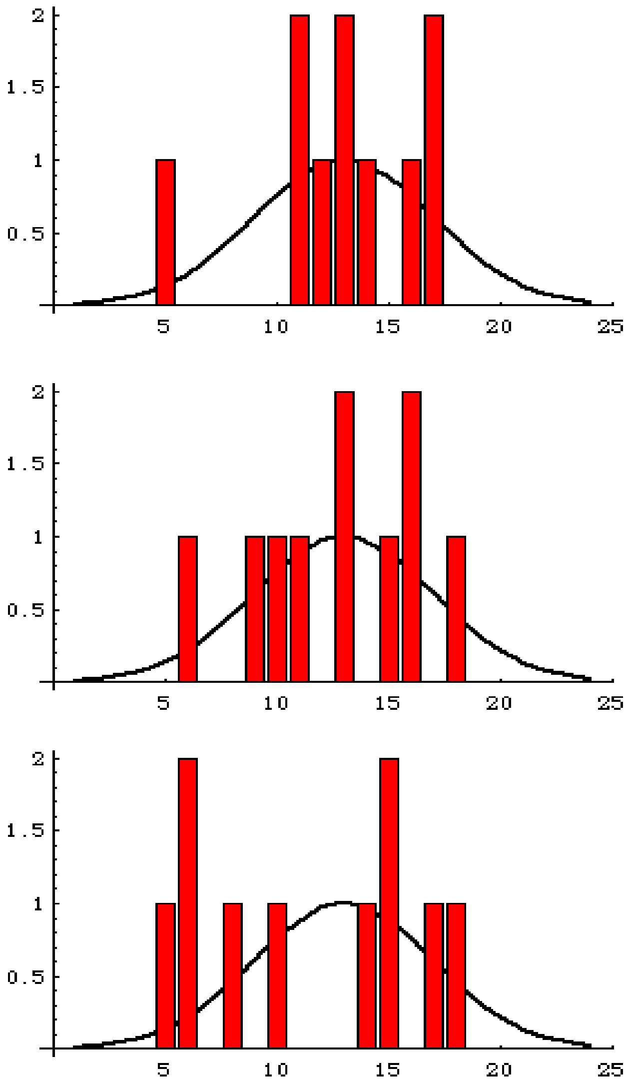
C. Accuracy and precision are the same and both refer to how close measurements are to the true value.

D. Accuracy and precision are unrelated to measurements and focus only on data variability.

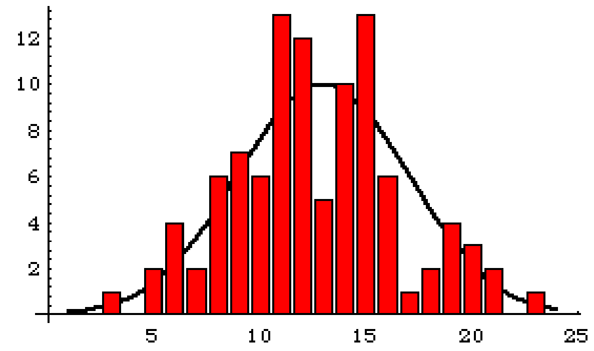
Is it Bias/Accuracy, Variation/Precision?

	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.

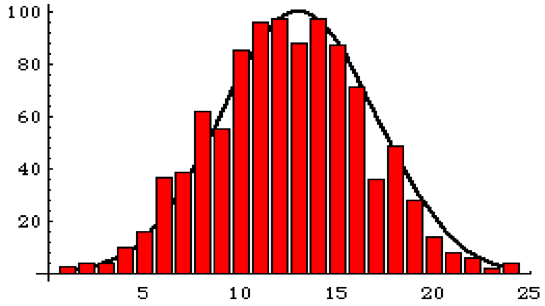
#	Scenario		
1	A scale always reads 0.5 grams too high, no matter who uses it.	**Bias**	Systematic error; accurate shape but wrong center.
2	Five repeated pipettings of the same solution yield 1.00, 1.02, 0.98, 1.01, and 0.99 mL.	**Precision**	High precision (tight grouping) even if mean might be off.
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



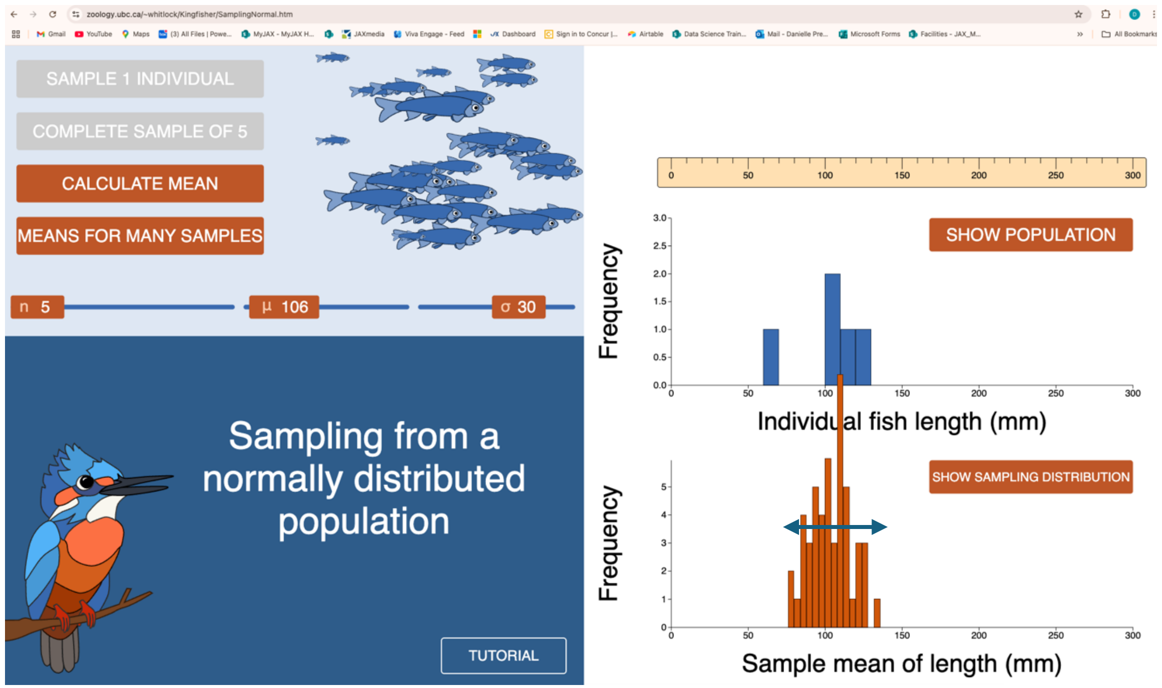
N=100



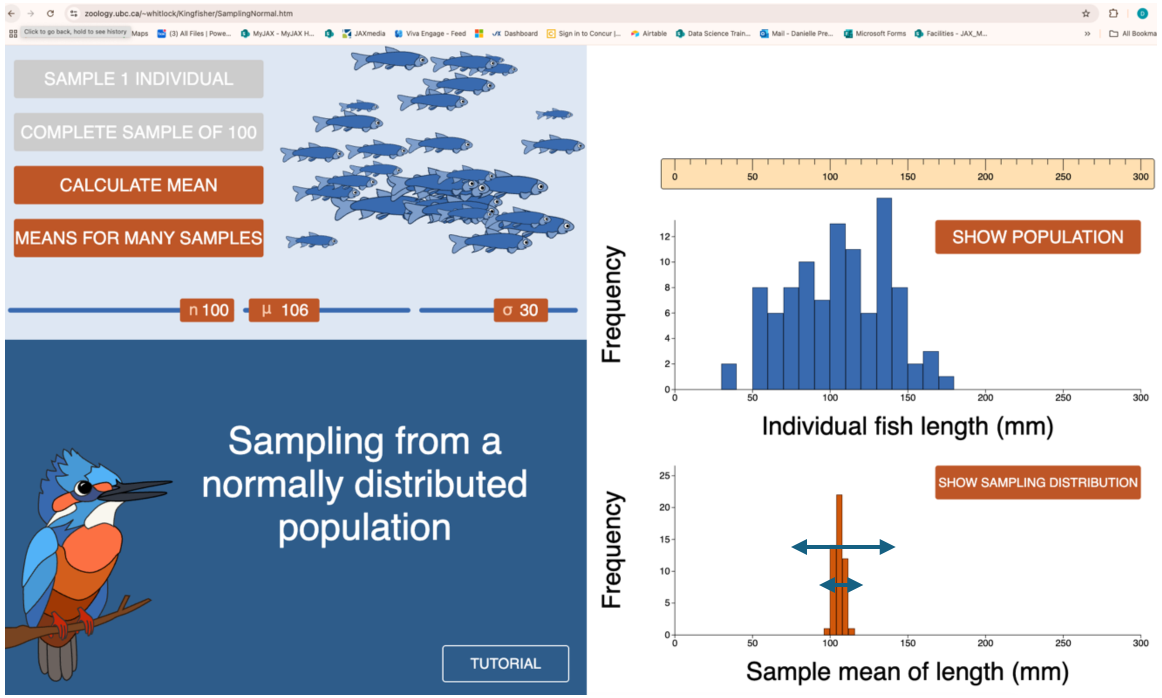
N=1000

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



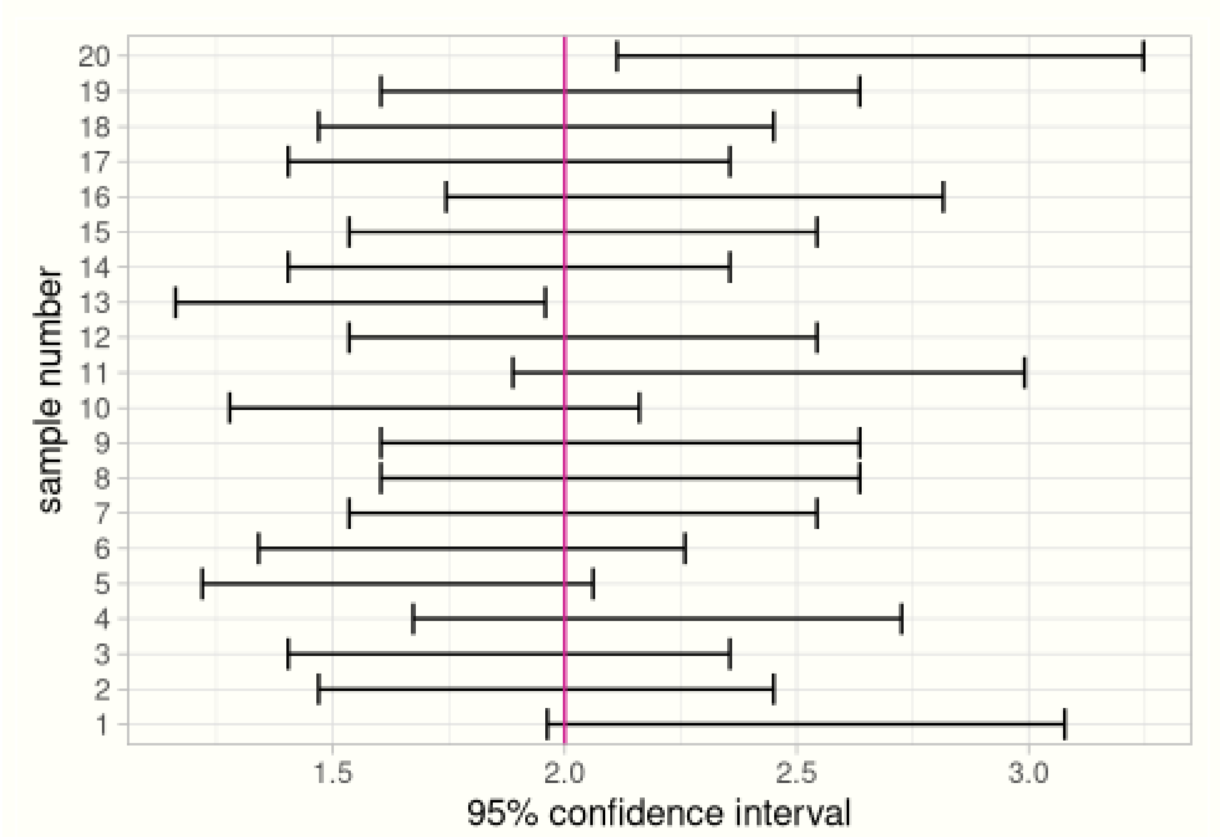
Many samples, each sample is size 100 individuals

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/

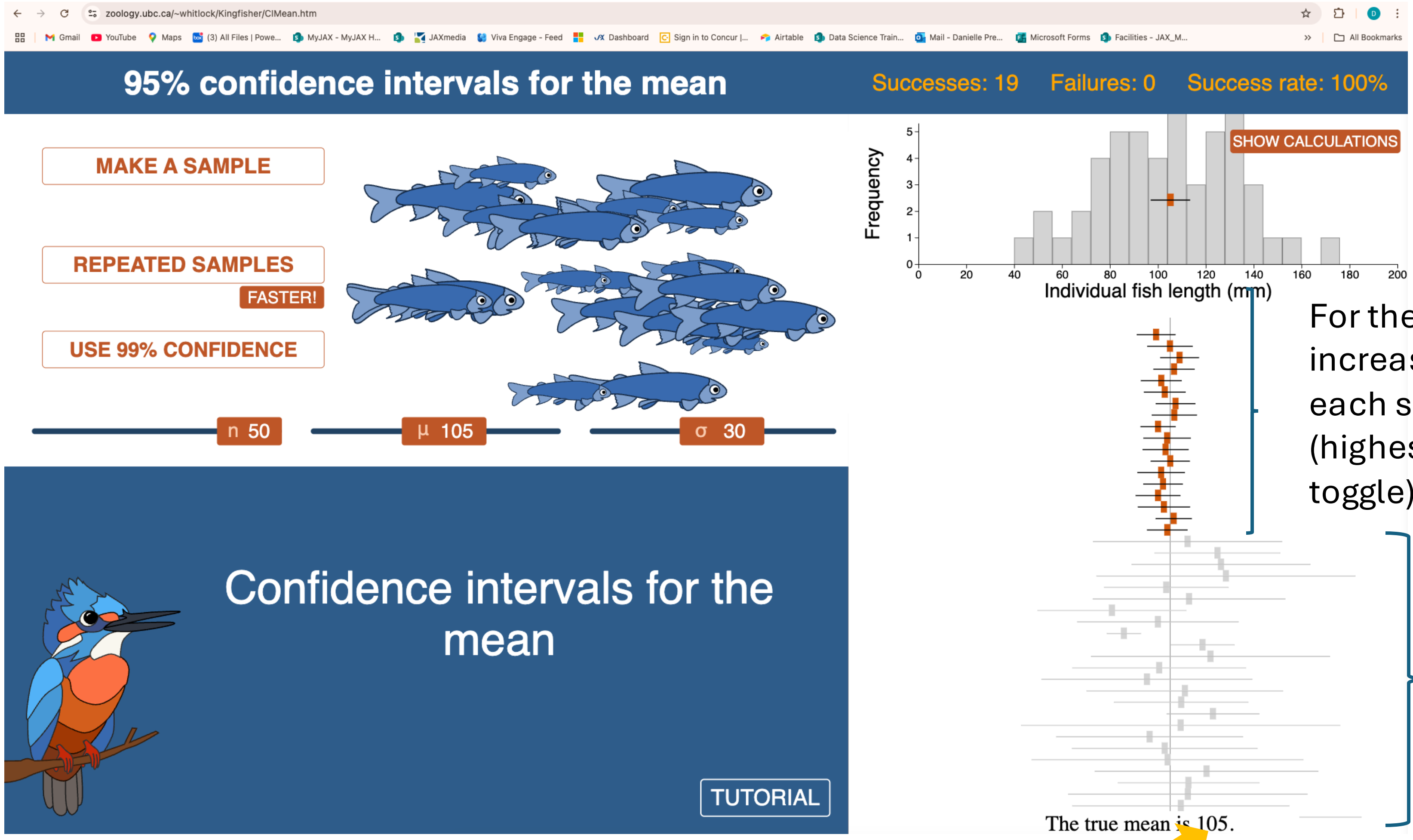
https://onlinestatbook.com/2/estimation/ci_sim.html

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:

Upload to Canvas as part of your Certification of Completion

Module 1A Questions:

A. Can the Standard deviation ever be zero? If so, when would that situation occur?

B. Imagine that we have a population that is skewed to the left. This population has a mean of 112 and a standard deviation of 16. Using a simulation program, Tyler simulated drawing 1000 samples of size 2 from the population. He then plotted the means for each of the samples that he drew. Alex simulated drawing 1000 samples of size 30, and he also plotted the means for each of the samples that he drew.

i. Would you expect the shape of Tyler’ s distribution of sample means to differ from the shape of Alex’s distribution of sample means? Please explain your answer (i.e., If you do expect the shapes to differ, how will they differ? If you do not expect the shapes to differ, why not?)

ii. Is the mean of Tyler’s distribution of sample means <,>, or = to the mean of Alex’s distribution of sample means and to the mean of the sample?

iii. How would you rank, from largest to smallest, the following: the standard deviation of Tyler’s distribution of sample means, the standard deviation of Alex’s distribution of sample means, and the standard deviation of the sample itself?