

Module 1 : Descriptive Statistics

Measurements of *location* and *spread* of data

- Mean, mode, median
- Variability, variation, range
- Simpson's paradox
- Intuitions about uncertainty: Fermi Estimation
- Accuracy/Bias and Precision/Spread
- Data types and their common visualizations:
 - Which type of plot is governed by data type

\$35,000	Reorder data→	\$25,000
\$30,000		\$30,000
\$25,000		\$30,000
\$80,000		\$35,000
\$50,000		\$45,000
\$30,000		\$50,000
\$45,000		\$80,000

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

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

Mode = most frequent value(s)

\$35,000	Reorder data→	\$25,000
\$30,000		\$30,000
\$25,000		\$30,000
\$80,000		\$35,000
\$50,000		\$45,000
\$30,000		\$50,000
\$45,000		\$80,000
\$1,000,000,000		\$1,000,000,000

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

\$35,000	Reorder data→	\$25,000
\$30,000		\$30,000
\$25,000		\$30,000
\$80,000		\$35,000
\$50,000		\$45,000
\$30,000		\$50,000
\$45,000		\$80,000

\$35,000	Reorder data→	\$25,000
\$30,000		\$30,000
\$25,000		\$30,000
\$80,000		\$35,000
\$50,000		\$45,000
\$30,000		\$50,000
\$45,000		\$80,000
\$1,000,000,000		\$1,000,000,000

	Scenario 1	Scenario 2
mean	\$42 143	\$125,000,037
median	\$35,000	\$40,000
mode	\$30,000	\$30,000
Variance (Population/Sample)	306122450/ 357142860	1.0936578E+17/ 1.2498946E+17
Standard deviation (Population/Sample)	17496.4/ 18898.2	330704974.26/ 353538484.5

Variance (Population): $\sigma^2=\frac{1}{N}\sum_{i=1}^N(X_i-\mu)^2$

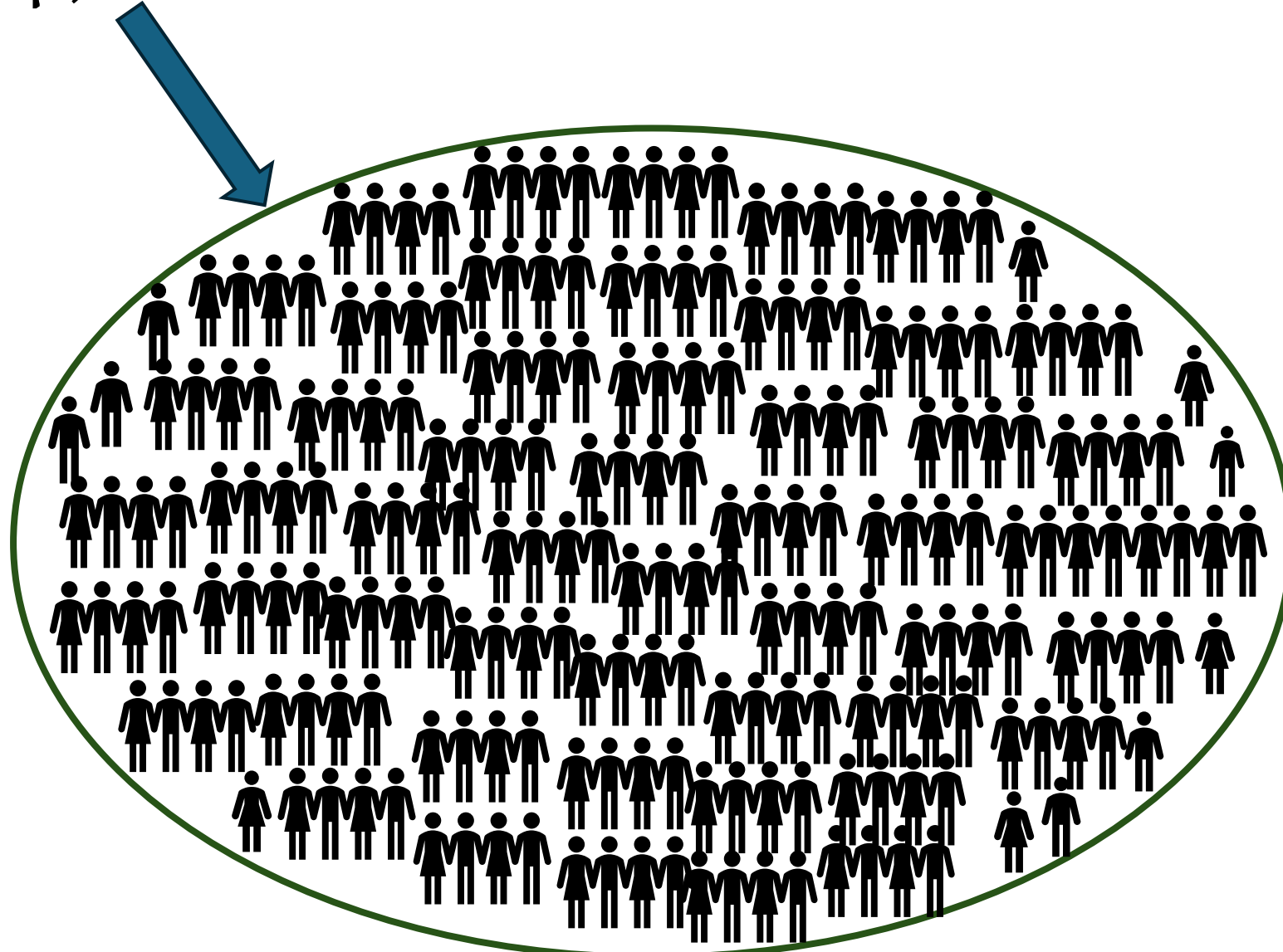
Variance (Sample): $s^2=\frac{1}{n-1}\sum_{i=1}(X_i-\bar{x})^2$

Range: $X_{\max}-X_{\min}$

Inter-Quartile Range (IQR): $R=X_{75}-X_{25}$

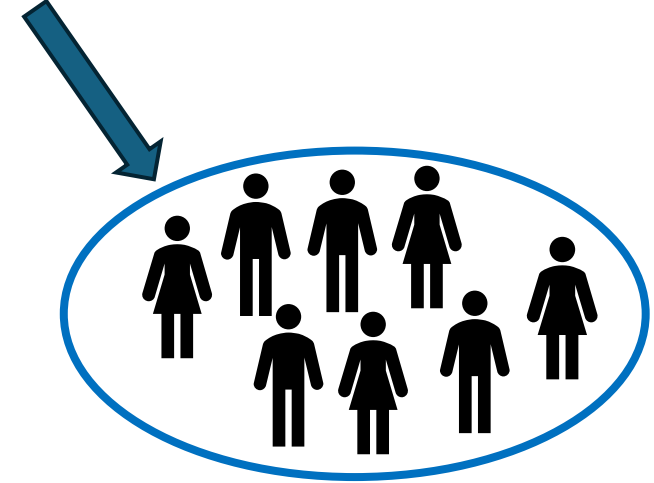
Populations have **P**ARAMETERS

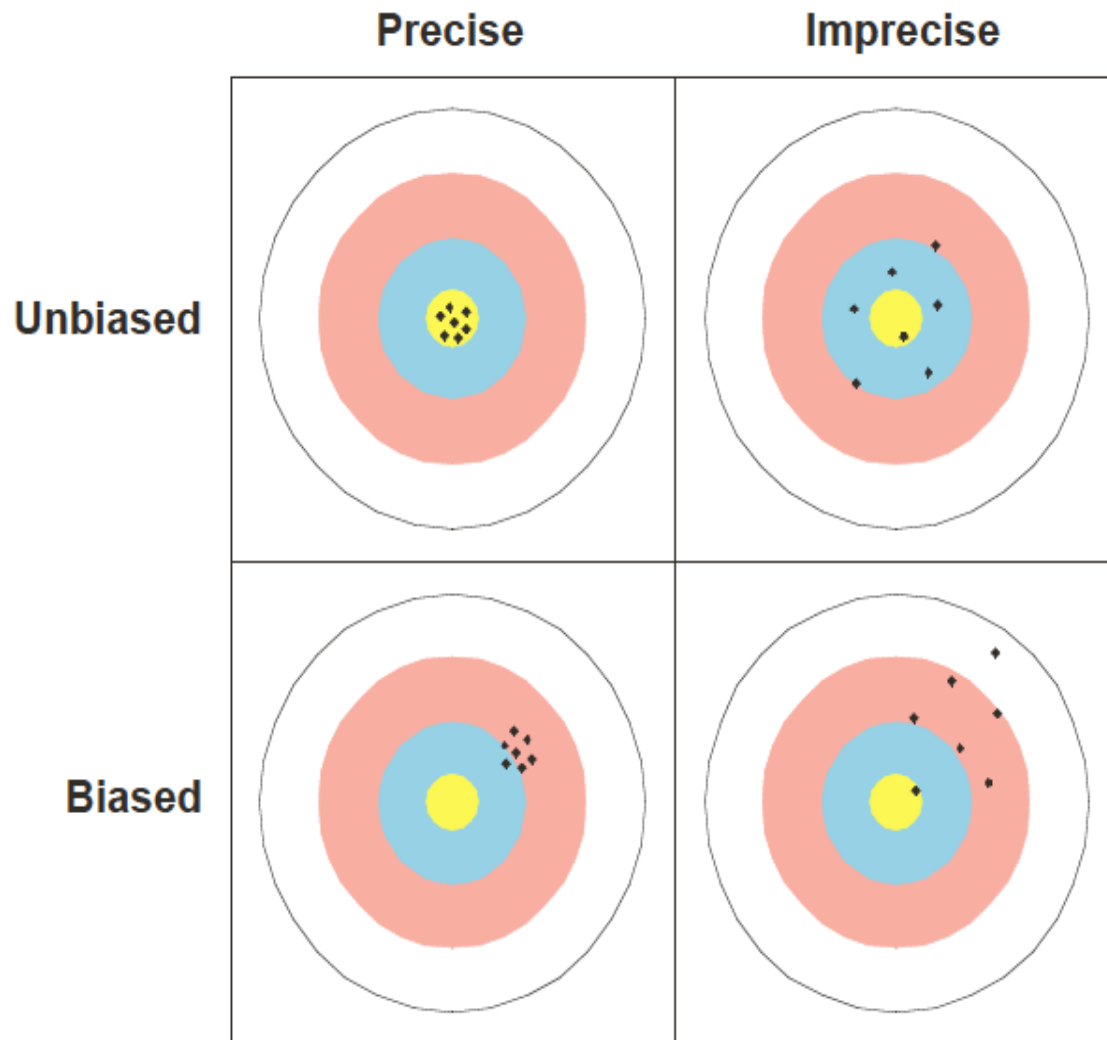
$\mu; \sigma$



Samples have **E**STIMATES

$\bar{x}; s$





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

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:

Summary

1. The appropriate visualization will depend on the type of variable(s) you are graphing

# variables	Variable Type	Recommended Plots	Use Case
1 (univariate)	Categorical	Bar Chart, Pie Chart	Comparing category frequencies
	Numerical	Histogram, Boxplot, Density Plot	Understanding distributions
2 (Bivariate)	Categorical & Categorical	Grouped Bar Chart, Mosaic Plot	Comparing proportions of two groups
	Numerical & Categorical	Boxplot, Violin Plot, Strip Plot	Comparing distributions across categories
	Numerical & Numerical	Scatter Plot, Line Plot, Hexbin Plot	Examining relationships or trends
3+ (Multivariate)	Multiple Categorical	Stacked Bar Chart	Analyzing categorical interactions
	Multiple Numerical	Scatterplot Matrix	Comparing multiple numeric relationships
	Mixed	Faceted Plots, Heatmap, Bubble Chart	Visualizing mixed data relationships

<https://www.data-to-viz.com/>

2. Everything else is (mostly) artistry and **being clear** in what you are revealing to your audience (See: Edward Tufte for “rules”)