Basic Statistical Concepts

Rethinking statistics

- Statistics are more than just facts and figures
- Statistics is a way to make sense of large data
  - Involves analyzing, interpreting, displaying, and making decisions based on data

Basic terminology

Variables

- Most research begins with a general question about the relationship between two variables for a specific group of individuals

- A **variable** is a characteristic or condition that can change or take on different values
  - Height and weight
  - Willingness to get vaccinated

Statistics

Numerical representations of our data can be:

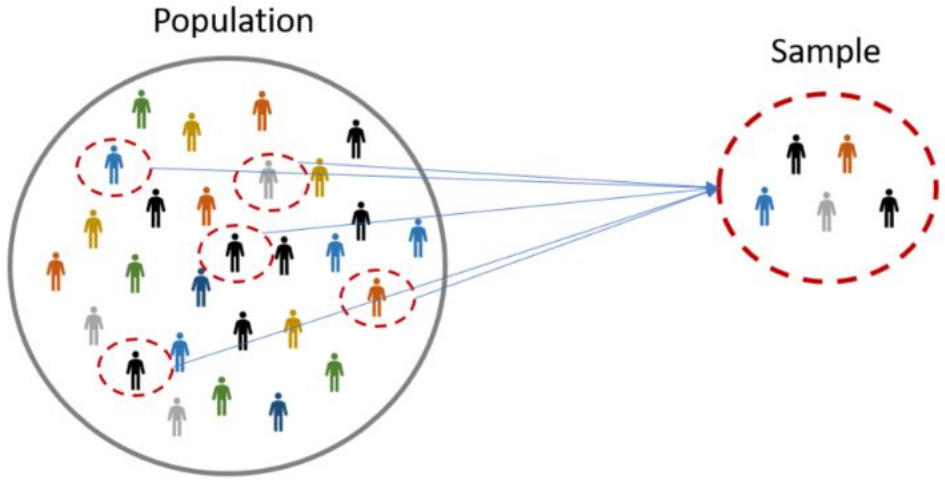
- Descriptive
  - Organize and summarize data
- Inferential
  - Indicate how much confidence we can have when we generalize from a sample to a populatio

Population

- The entire group of individuals is called the **population**
- For example, a researcher may be interested in the relation between class size (variable 1) and academic performance (variable 2) for college students in the U.S.
- **Parameter** - any summary number, like an average or percentage, that describes the entire population

Why sampling?

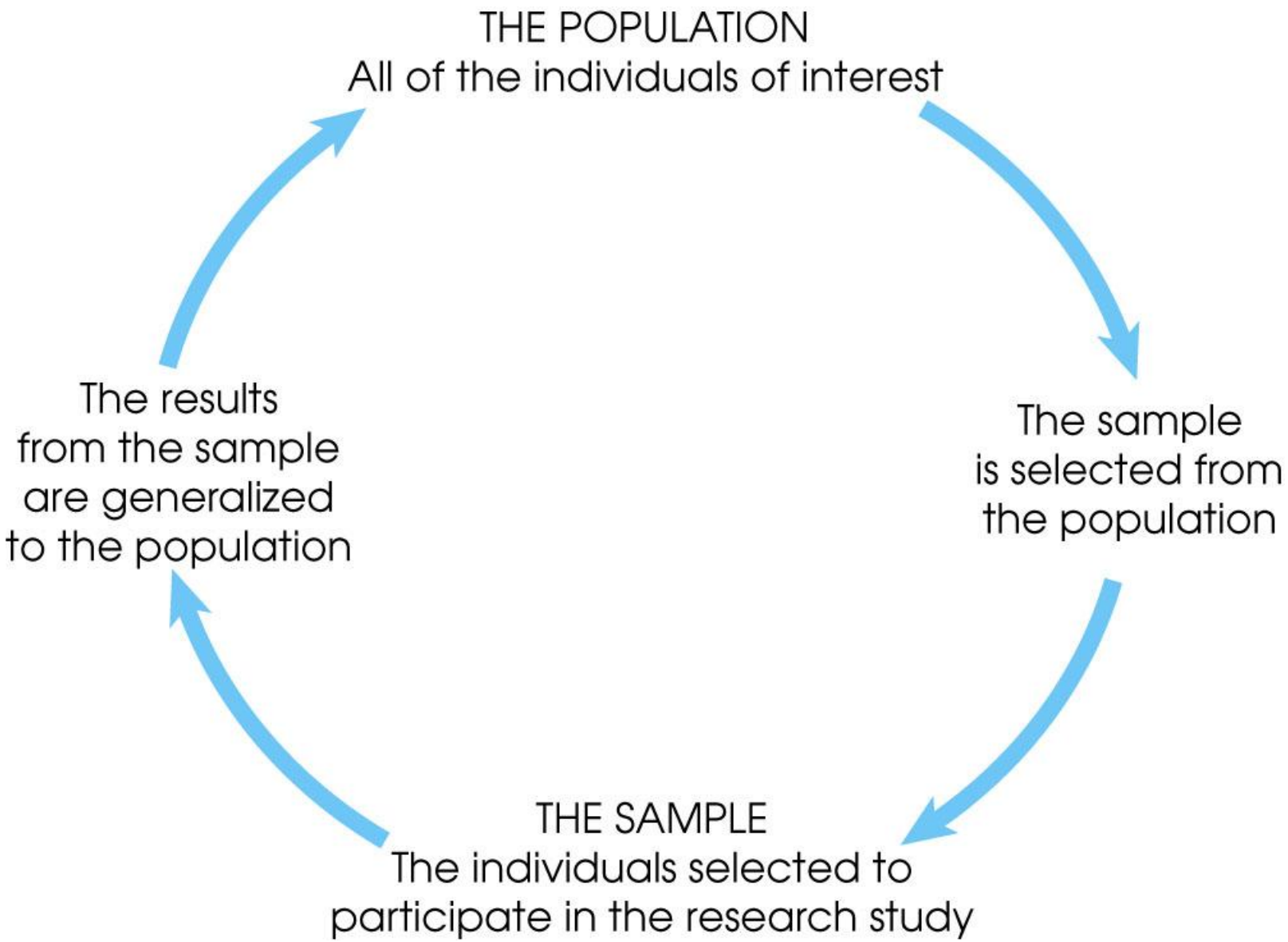
- Usually, populations are so large that a researcher cannot examine the entire group
- A **sample** is selected to represent the population in a research study
- Many reasons to choose sampling:
  - Less costs
  - Less field time
  - More accuracy
  - When it's impossible to study the population



Population vs. Sample notation

- A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**

Parameter name	Population parameter symbol	Sample statistic
Number of cases	N	n
Mean	μ (mu)	$\bar{x}$ (Sample mean)
Proportion	π (Pi)	P (Sample proportion)
Variance	σ^2 (Sigma-square)	s^2 (Sample variance)
Standard deviation	σ (Sigma)	s (sample standard deviation)
Correlation	ρ (rho)	r (Sample correlation)
Regression Coefficient	β (beta)	b (sample regression coefficient)


THE POPULATION
All of the individuals of interest
The sample
is selected from
the population
THE SAMPLE
The individuals selected to
participate in the research study
The results
from the sample
are generalized
to the population

Learning check 1 ☑

Use the following scenarios to identify populations and samples

1. A gaming website wanted to find out which console its visitors owned. Which choice **best** represents a population?

   A) Visitors to the XboxOne section

   B) All of the website visitors

   C) Visitors to the PS5 section

   D) Visitors who are on the website for more than 5 minutes

Learning check 1 ☑

Use the following scenarios to identify populations and samples

1. A gaming website wanted to find out which console its visitors owned. Which choice **best** represents a population?

   A) Visitors to the XboxOne section

   B) All of the website visitors

   C) Visitors to the PS5 section

   D) Visitors who are on the website for more than 5 minutes

Learning Check 2 ☑

2. Before the 2020 U.S. Presidential election, a poll was trying to estimate who would win the election. Which choice represents the **best** sample for the poll?

A) A selection of voters over the age of 50

B) All registered voters in the U.S.

C) Democratic voters

D) A selection of voters from all ages and political backgrounds

Learning Check 2 ☑

2. Before the 2020 U.S. Presidential election, a poll was trying to estimate who would win the election. Which choice represents the **best** sample for the poll?

A) A selection of voters over the age of 50

B) All registered voters in the U.S.

C) Democratic voters

D) A selection of voters from all ages and political backgrounds

Mathematical notation

Mathematical notation (bringing algebra back in)

$\sum$	Summation
s	The standard deviation of sample data
σ	The standard deviation of population data
s^2	The variance of sample data
σ^2	The variance of population data
R	The range of data
$\overline{R}$	The average range of data
k	Multi-purpose notation, i.e. # of subgroups, # of classes
$\lvert y \rvert$	The absolute value of some term
$>, <$	Greater than, less than
$\geq, \leq$	Greater than or equal to, less than or equal to
x	An individual value, an observation
x_1	A particular (1st) individual value
x_i	For each, all, individual values
$\overline{x}$	The mean, average of sample data
$\overline{\overline{x}}$	The grand mean, grand average
μ	The mean of population data
p	A proportion of sample data
P	A proportion of population data
n	Sample size
N	Population size

Mathematical notation

- Individual measurements or scores can be identified by the letter *X* (or *X* and *Y* if there are multiple scores for each individual)

- The number of scores in a dataset are identified by *N* for a population, n for a sample

Mathematical notation

- Summing a set of values in statistics has its own notation: the Greek letter sigma, Σ. This will be used to stand for "the sum of."
  - ΣX identifies the sum of the X scores
  - ΣY identifies the sum of the Y scores
  - ΣXY identifies the sum of $X*Y$
  - ΣX^2 identifies sum of (X^2)

Notation Examples – Try this on your own!

X	X^2
3	9
1	1
7	49
4	16

1. $\Sigma X =$
2. ΣX^2
3. $(\Sigma X)^2$

Notation Examples - Solutions

X	X^2
3	9
1	1
7	49
4	16

1. $\Sigma X =$

   $\Sigma X = 3 + 1 + 7 + 4$

   $\boldsymbol{\Sigma X = 15}$

2. ΣX^2

   $\Sigma X^2 = 9 + 1 + 49 + 16$

   $\boldsymbol{\Sigma X^2 = 75}$

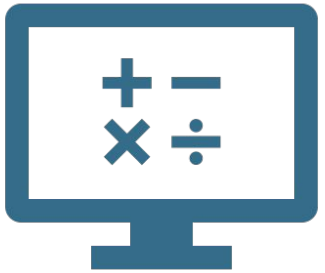
3. $(\Sigma X)^2$

   $(\Sigma X)^2 = (15)^2$

   $\boldsymbol{(\Sigma X)^2 = 225}$

Order of operations - Review

Remember the order of operations? They're useful here too!



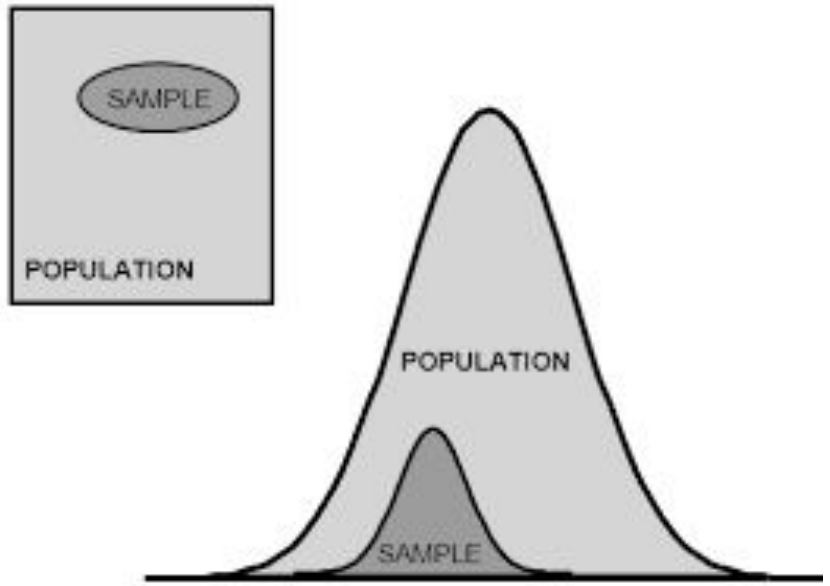
Please **E**xcuse **M**y **D**ear **A**unt **S**ally

1. **P**arentheses - All calculations within parentheses are done first
2. **E**xponents - Squaring or raising to other exponents is done second
3. **M**ultiplication and **D**ivision - Multiplying, and dividing are done third, and should be completed in order from left to right
4. **A**ddition and **S**ubtraction - Summation with the Σ notation is done next. Any additional adding and subtracting is done last and should be completed in order from left to right

Central Tendency

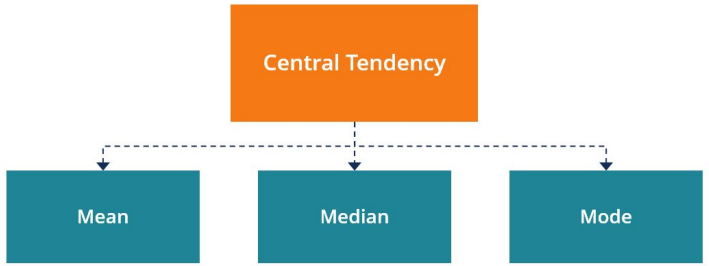
Populations & samples

- Population
  - Parameter
  - Exact value
  - Population mean = μ
- Sample
  - Statistic
  - Estimate of parameter
  - Introduces error
  - Sample mean = $\overline{X}$



Central tendency

- A single score to define the center of a distribution
- Purpose: find the single score that is most typical or best represents the entire group



Mean as calculation

- The mean is the sum of all the scores divided by the number of scores in the data.

- Population:

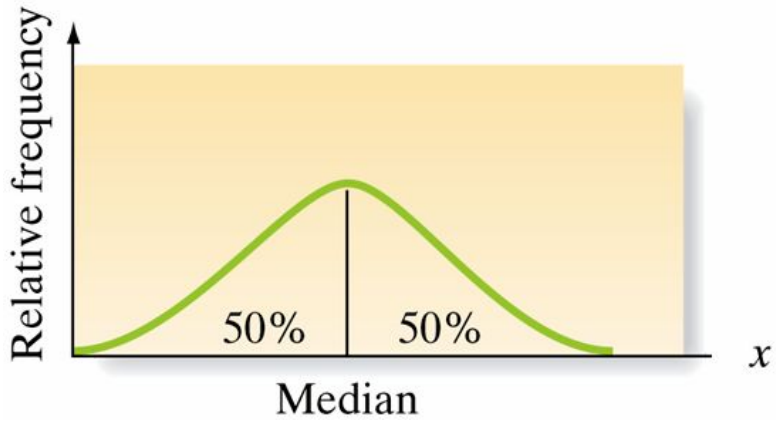
$$\mu = \frac{\sum X}{N}$$

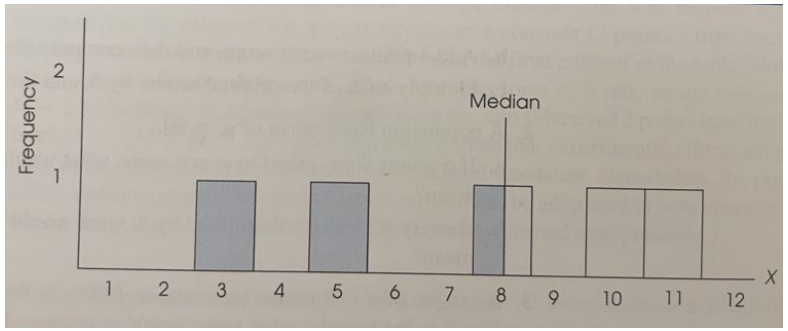
- Sample:

$$M = \frac{\sum X}{n}$$

THE MEDIAN

- Midpoint of the scores in a distribution when they are listed in order from smallest to largest
  - Divides the scores into two groups of equal size
- Finding the median
  - Arrange the *n* measurements from the smallest to the largest
  - If n is odd, M is the middle number
  - If n is even, M is the mean of the middle two numbers





Locating the median (odd *N*)

Assume you had the following data: 10, 5, 2, 11, 8

Step 1: Put scores in order

Step 2: Identify the "middle" score to find median

2 5 8 10 11

Answer: "Middle" score is 8 so median = 8

Locating the median (even *N*)

Assume you had the following data: 9, 7, 1, 1, 5, 4

Step 1: Put scores in order

Step 2: Average middle pair to find median

1 1 4 5 7 9

(4 + 5) / 2 = 4.5

The mode

- The mode is the score or category that has the greatest frequency of any score in the frequency distribution
  - Can be used with any scale of measurement
  - Corresponds to an actual score in the data

- It is possible to have more than one mode

Learning objectives

By the end of this lecture, you should be able to:

- Differentiate populations from samples
- Read scientific notation
- Identify different measures of central tendency