

# LA's BEST @ USC

Introduction to Biostatistics &  
A Motivation for Statistical Learning

Trevor A. Pickering, PhD

Assistant Professor

USC Division of Biostatistics

# Overview

- “About me” - AKA - “Why the heck do I like stats?”
- What is biostatistics?
- Types of data
- Thinking statistically
- Describing a sample: distributions, percentiles, central tendency, variation

# An FAQ About... Me?

## **Who am I?**

- Assistant Professor of Biostatistics at the Keck School of Medicine of USC

## **What do I do?**

- Short answer: Teaching (40%), Research (40%), "Service" (20%)
- Co-Director of the CTSI BERD Biostatistics Consulting core
- Director of the MS programs in Biostatistics and Public Health Data Science

# An FAQ About... Me?

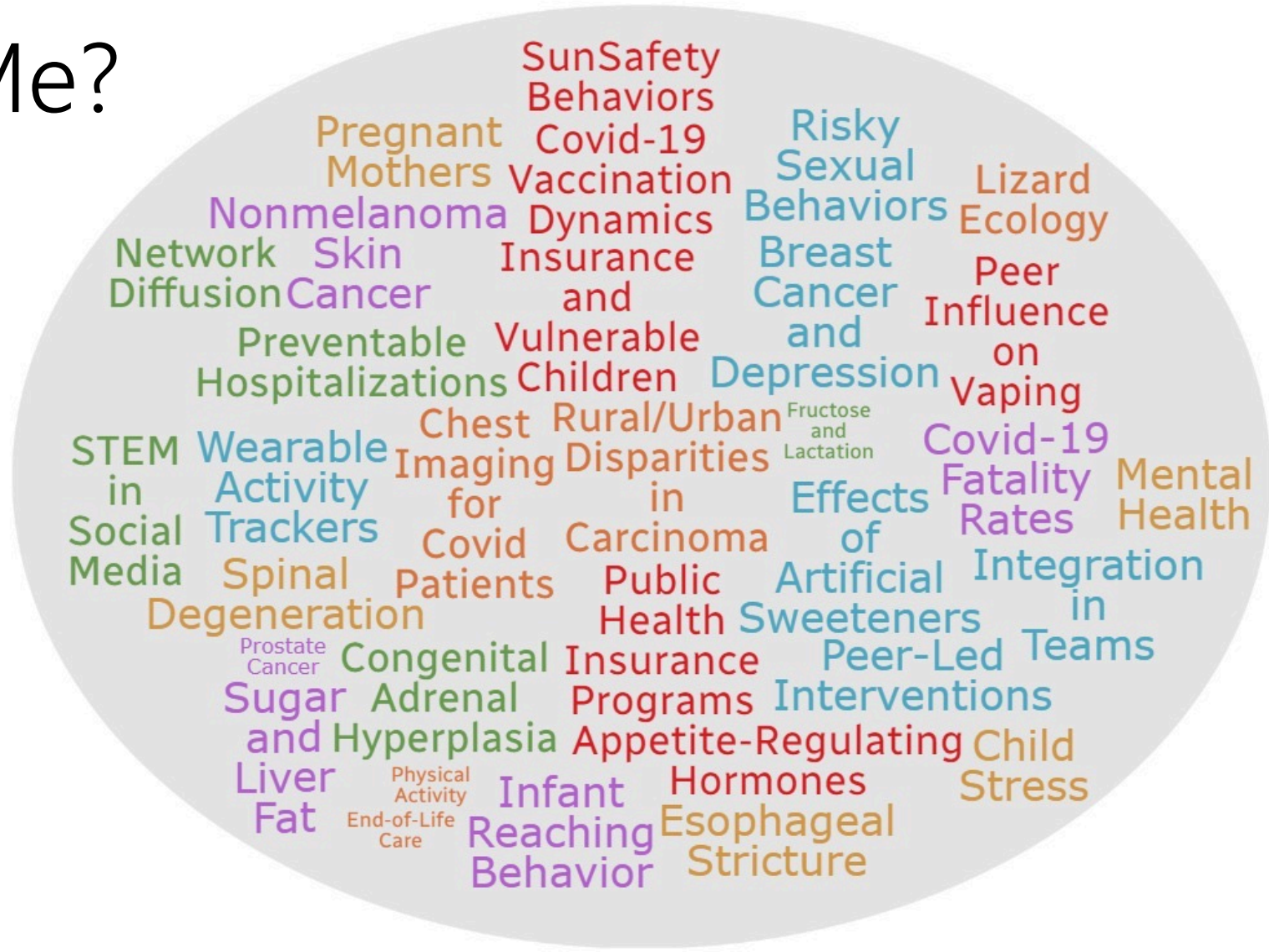
## **How did I end up here?**

- BS in Mathematical Biology studying measurement of lizards (lizard ecology)
- MS in Biostatistics studying air pollution and lung function in adolescents
- PhD in Preventive Medicine - Health Behavior studying peer-led interventions in schools



# An FAQ About... Me?

**And I enjoy analyzing data!**

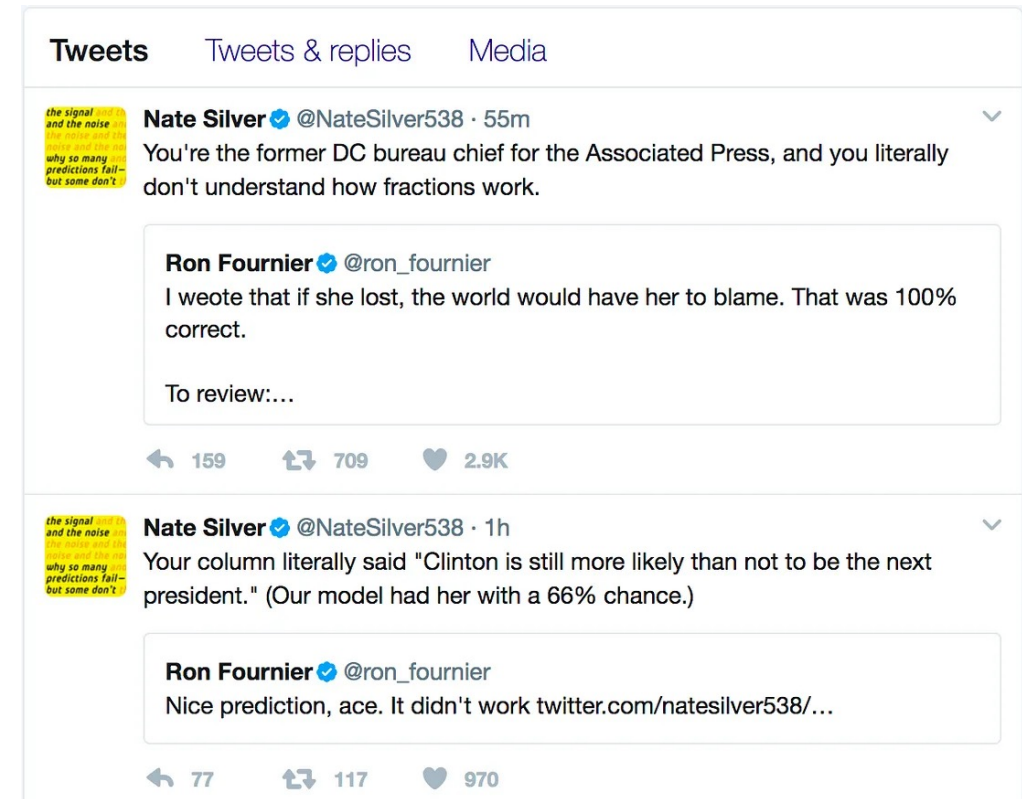


# Why do I enjoy statistics?

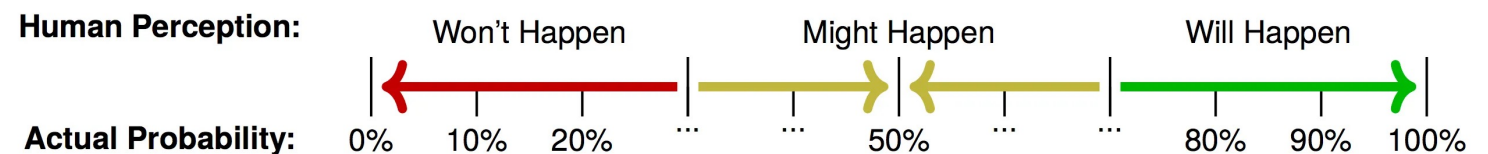
## Humans are **BAD** at assessing things!

Example: People frequently misinterpret probability

<https://towardsdatascience.com/humans-are-bad-at-probability-834980e719a3>.



### Why Humans are Bad at Thinking About Chance



# Why do I enjoy statistics?

**Statistics makes us more informed about our world!**

Example: Group Fitness



# GROUP FITNESS: THE 3 ELEMENTS THAT Make It Work

DID YOU KNOW THERE ARE 3 KEY FACTORS THAT AFFECT  
A NEW PERSON'S GROUP FITNESS EXPERIENCE?

## SMARTSTART

The level of satisfaction that our members experience in class will determine whether or not they come back the following week. Consider this: 50% of people who start a new exercise regime will drop out within the first 6 months. If we understand how we can influence our participants' satisfaction levels, this will help us to get new exercisers addicted to fitness.

### THE THREE KEY FACTORS

Past research has highlighted three variables that are key to improving satisfaction during exercise classes. These are: intensity, social connection and competence.

#### Intensity

Previous research has indicated that exercising at high intensity can result in a level of discomfort that can be associated with reduced satisfaction, particularly for beginners.

#### Social Connection

Feeling connected to the group and the instructor has been found to be important to

### THE TEST

To see how these variables affect class satisfaction in LES MILLS™ programs, we completed a 30-week intervention with 25, non-active, but otherwise healthy, adults between the ages of 25 and 40. The participants completed a 6-week familiarization protocol followed by two 12-week blocks of 6 or 7 Les Mills classes per week. After each class, they completed surveys that included questions related to their levels of satisfaction with the class, what they thought of the instructor, the levels of intensity, their connection to the group and the instructor, and their levels of competence.

### THE RESULTS

There was a difference between the levels of satisfaction and **Intensity** depending on the type of class. For example, when the participants completed a cardio class such as BODYATTACK™, BODYCOMBAT™, BODYSTEP™, or RPM™ and rated the class as being intense and challenging, they also rated the class as being highly satisfying. On the other hand, if they found a BODYPUMP™ or BODYBALANCE™/BODYFLOW® class to be more intense than anticipated, then they were not as satisfied with the overall experience.

Feeling **Connected** to the group and the instructor was extremely important to the levels of satisfaction in the class, and this was particularly evident during the first 12-week block. This connection allowed participants to cope with the discomfort of exercise, and therefore increased their enjoyment of the workout.

Finally, when we looked at **Competence**, the beginner participants were more satisfied when there was technical and clear instruction, enabling them to execute the movements well. Then, once they felt they had developed the skills required to complete

### WHAT DOES YOUR CLASS NEED?


1. Remember more structure than your pushed to but we need slowly in E and BODY
2. Using clear people to in the early levels, giving energy of
3. It's vital to you the For tips or connection the Educators Directors

Don't forget to **template** to provide new after the first

[www.lesmills.com](http://www.lesmills.com)



# Perceptions of the activity, the social climate, and the self during group exercise classes regulate intrinsic satisfaction

 **Jaclyn P. Maher**<sup>1,2\*</sup>,  **Jinger S. Gottschall**<sup>1</sup> and  **David E. Conroy**<sup>1,3</sup>

<sup>1</sup>Department of Kinesiology, Pennsylvania State University, University Park, PA, USA

<sup>2</sup>Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

<sup>3</sup>Department of Preventive Medicine, Northwestern University, Chicago, IL, USA

Engaging in regular physical activity is a challenging task for many adults. Intrinsic satisfaction with exercise classes is thought to promote adherence to physical activity. This study examined the characteristics of exercise classes that impact within-person changes in intrinsic satisfaction over the course of an extended group exercise program. A 30-week physical activity trial was conducted with assessments at the end of each class. Community-living adults ( $n = 29$ ) were instructed to complete at least six group exercise classes each week and, following each exercise class, complete a questionnaire asking about the characteristics of the class and the participant's evaluation of the class. Intrinsic satisfaction was high, on average, but varied as much within-person from class-to-class as it did between exercisers.

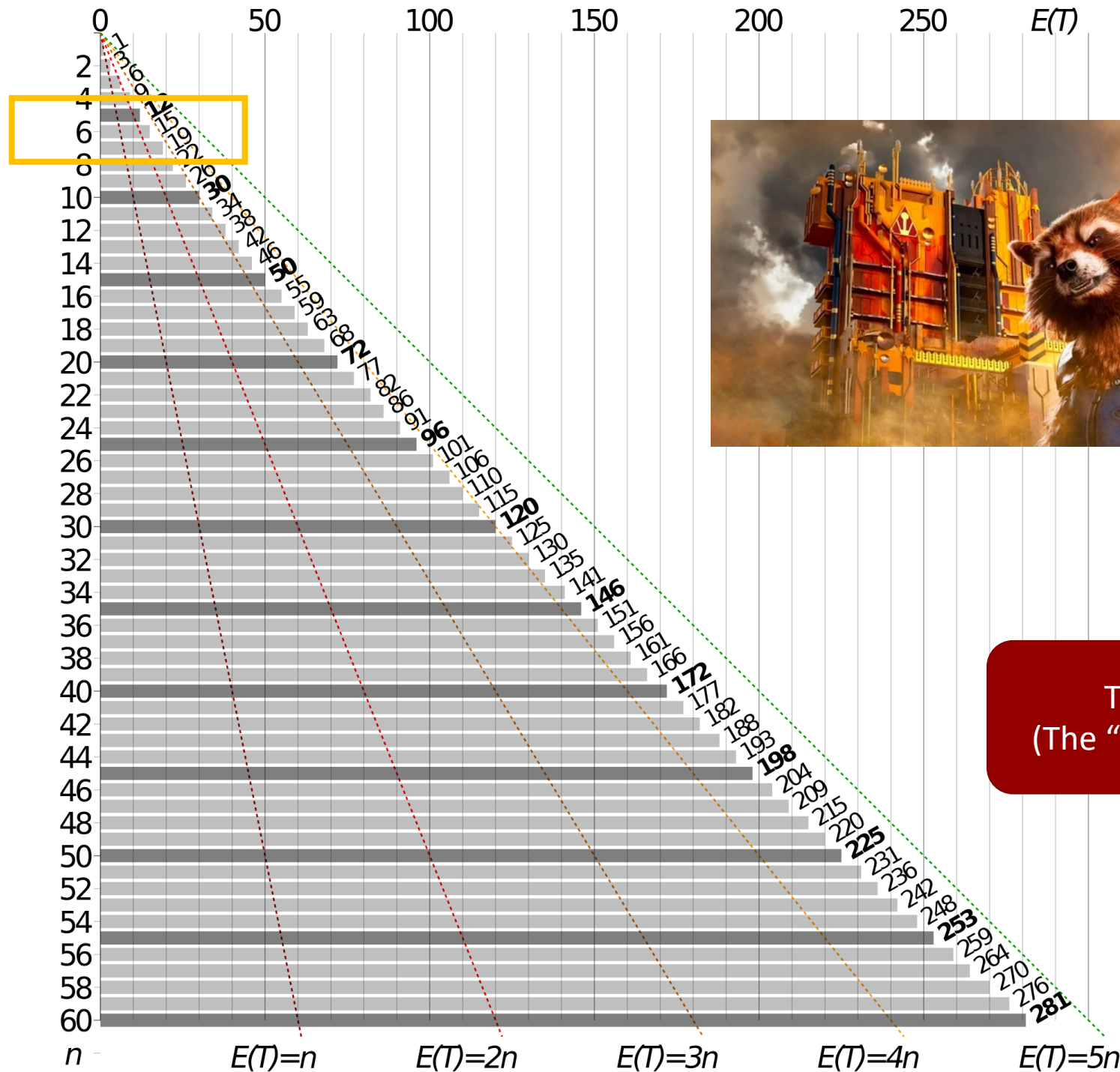
Participants reported the greatest intrinsic satisfaction when classes placed greater emphasis on exercisers' involvement with the group task, feelings of competence, and encouragement from the instructor. For the most part, exercise classes that were more intense than usual were perceived by exercisers as less intrinsically satisfying. Some overall characteristics of the exercise classes were also associated

with intrinsic satisfaction. The social and motivational characteristics of group

# Why do I enjoy statistics?

**Statistics makes us more informed about our world!**

Example: Probabilities of Events



The “coupon collector” problem  
(The “Happy Meal toy collector” problem)

# Why do I enjoy statistics?

**Statistics takes the “I’ve noticed...” or “I think that...” and puts a number to it!**

Picture it – London – 1710.

It seems like more males tend to be born than females.

John Arbuthnot examines birth records from 1629 to 1710.

Every year there were more males born than females.

*Could this happen due to chance?*



Arbuthnot thought this was *divine providence*.  
“It is art, not chance, that governs.”

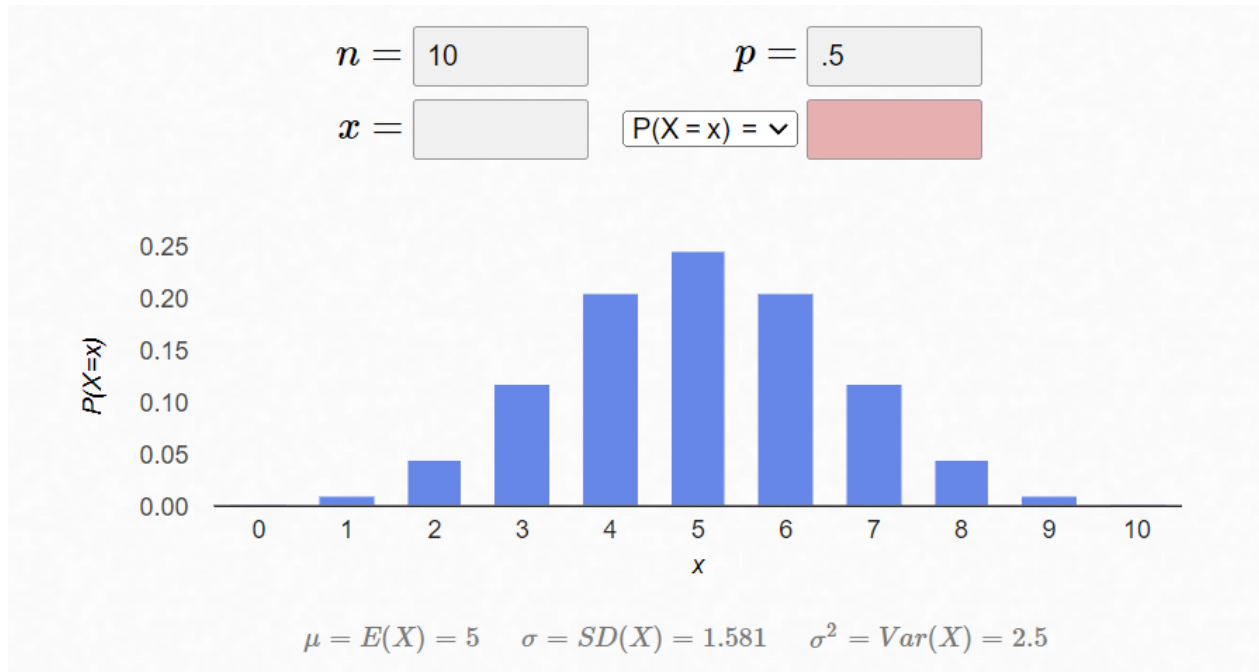
# Why do I enjoy statistics?

**I will now share psychic predictive powers with the class.**

# Why do I enjoy statistics?

## Some questions to consider:

- How uncommon/improbable would it be for all 10 cards to be guessed correctly?
- At what number of correct guesses out of 10 do we stop considering it improbable?
- Would it be just as surprising if all 10 guesses were wrong?



X	P
10	0.00098
9	0.00977
8	0.04395
7	0.11719



# What is biostatistics?

**Misconception: it's just calculating things like mean, median, mode**

Descriptive Statistics					
Variable	<u>Obs</u>	Mean	<u>Std.Dev.</u>	Min	Max
price	74	6165.257	2949.496	3291	15906
mpg	74	21.297	5.786	12	41
rep78	69	3.406	.99	1	5
headroom	74	2.993	.846	.846	5
trunk	74	13.757	4.277	5	23
weight	74	3019.459	777.194	1760	4840
length	74	187.932	22.266	142	233
turn	74	39.649	4.399	31	51
displacement	74	197.297	91.837	79	425
<u>gear_ratio</u>	74	3.015	.456	2.19	3.89
foreign	74	.297	.46	0	1

# What is Biostatistics?

**A mathematical body of science dealing with the collection, analysis, interpretation, and presentation of (biological) data.**

## **Many specialized branches**

- Statistical genetics
- Environmental statistics
- Clinical trials
- Psychometric statistics

# Modern Biostatistics

## Really took off in the early 1900s

- Francis Galton – 1907 – examined 787 guesses of the weight of an ox at a county fair. The median was 1208; the actual weight was 1,198. *The Wisdom of the Crowd*.
- Karl Pearson – Early 1900s – *The “standard deviation.” The correlation coefficient.*
- Ronald Fisher – 1920s-30s – *Design of Experiments, Rothamsted Experimental Station*

# Modern Biostatistics

## **Biostatistics in 2023**

- The need for biostatisticians is expected to increase 31% from 2021 to 2031.
- Some top jobs for individuals in biostatistics are:
  - Data Analyst/Scientist
  - Biostatistician
  - Software Engineer
  - Research Analyst
  - Bioinformaticist

# Modern Biostatistics

## PERSONAL AND PROFESSIONAL SKILLS OF BIOSTATISTICIANS

Biostatisticians work in offices, in laboratories and in the field conducting a range of tasks, from designing research studies to analyzing and reporting on their results. In addition to technical and statistical skills, biostatisticians require several personal and professional skills.



Written and oral  
communication



Problem-solving



Critical  
thinking



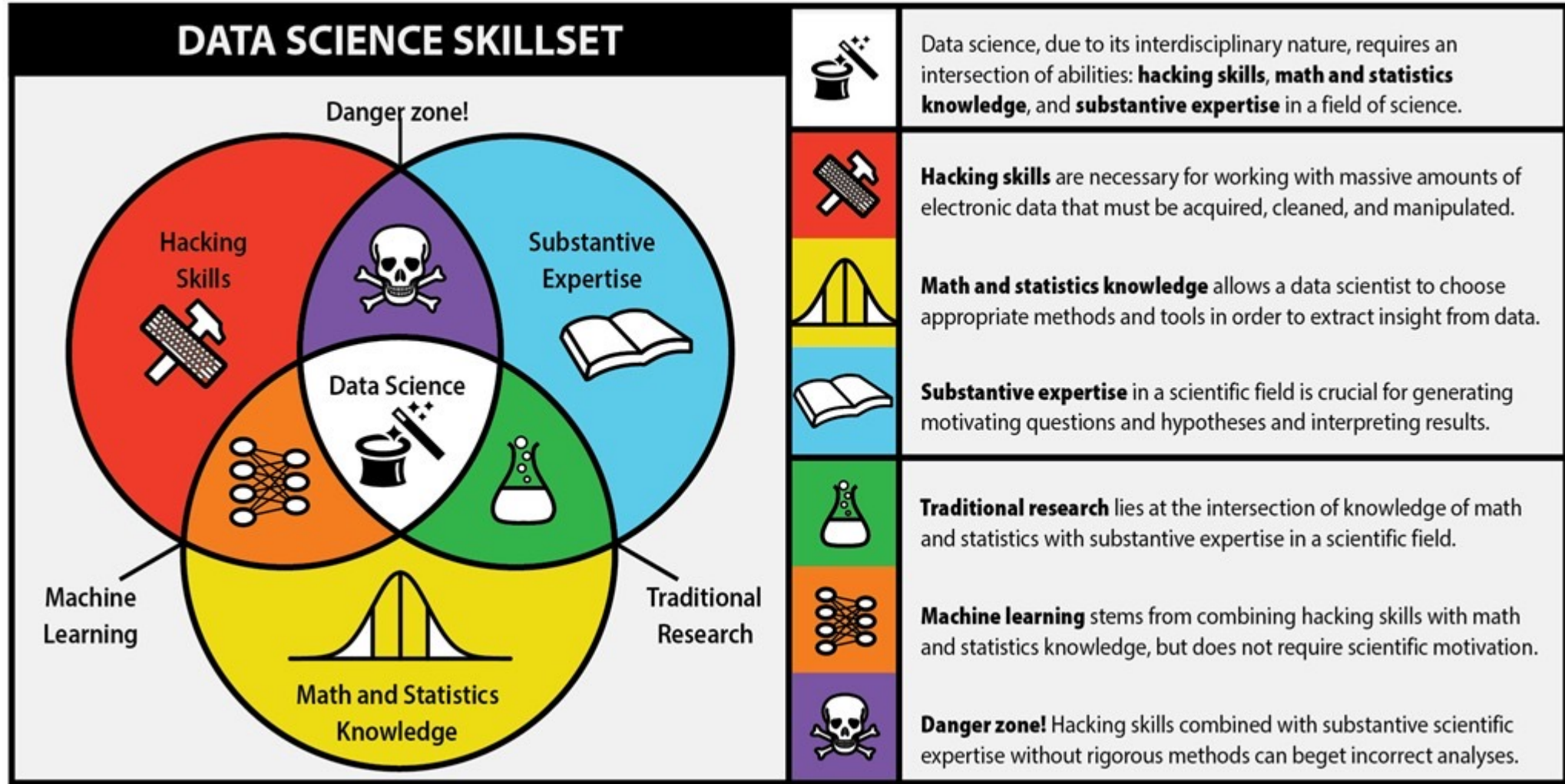
Ability to work  
autonomously



Adaptability

Source: Indeed

# Modern Biostatistics





# Types of Data

In order to analyze the data, it is helpful to know the types of variables we are using.

## Categorical

Binary/Dichotomous: only two options (*card color red vs. black*)

Nominal: no natural order (*card suit*)

Ordinal: has a natural order (*A, 2, 3, ..., 10, J, Q, K*)



# Types of Data

## **Count/Discrete**

Countable, ordered whole numbers (*# of students, # of strokes*)

## **Continuous**

Ordered numerical data that can, in theory, take on any value (*height, weight, age, cholesterol level*)

Note:

- Continuous data can sometimes be treated as categorical.
- Discrete data can sometimes be approximated as continuous.

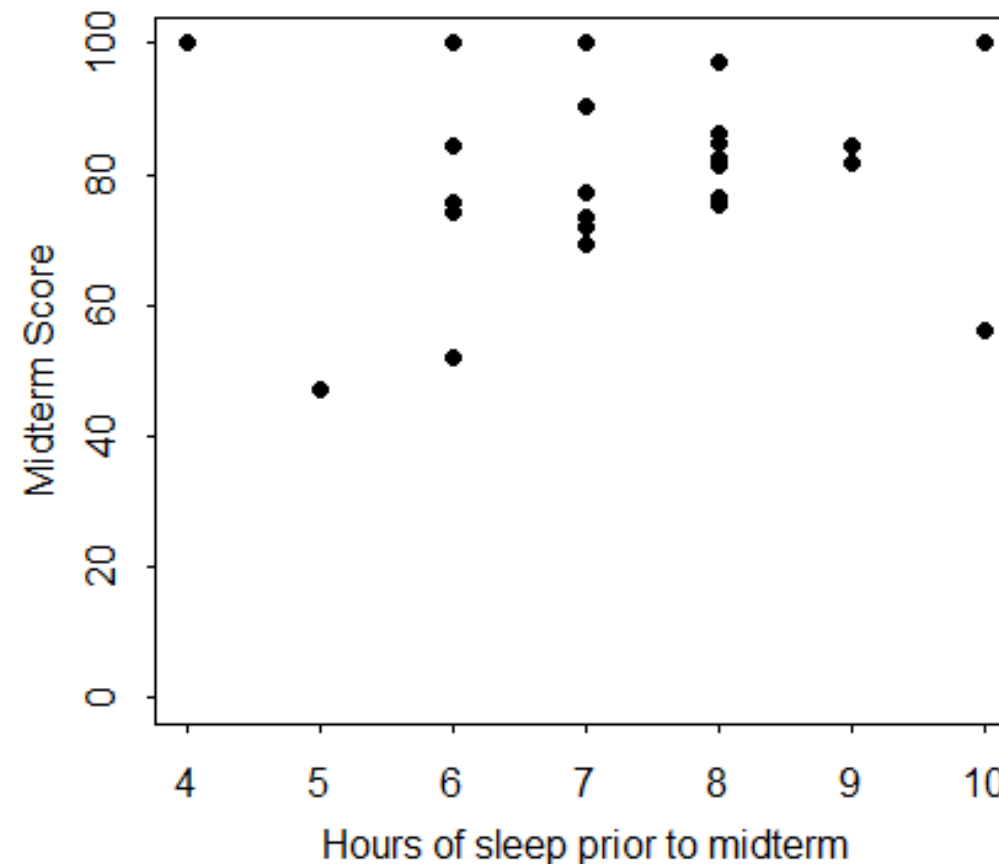
# How can we begin to think statistically?

Biggest advice: don't take things that seem obvious or like "common sense" at face value!

Recognize the role that randomness and chance plays in the processes you are studying.

# Quantifying Statistical Evidence

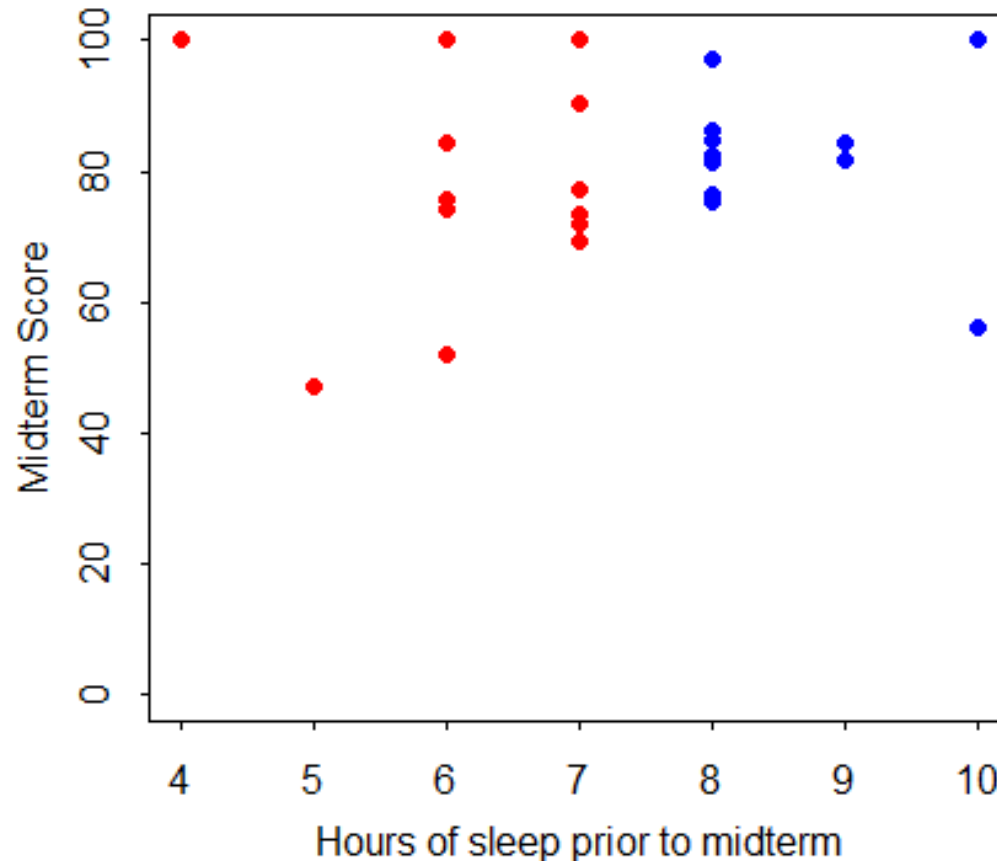
Example: Is midterm grade associated with sleep duration prior to midterm?



Is there an association?  
Does it seem obvious?  
Could you quantify this association?

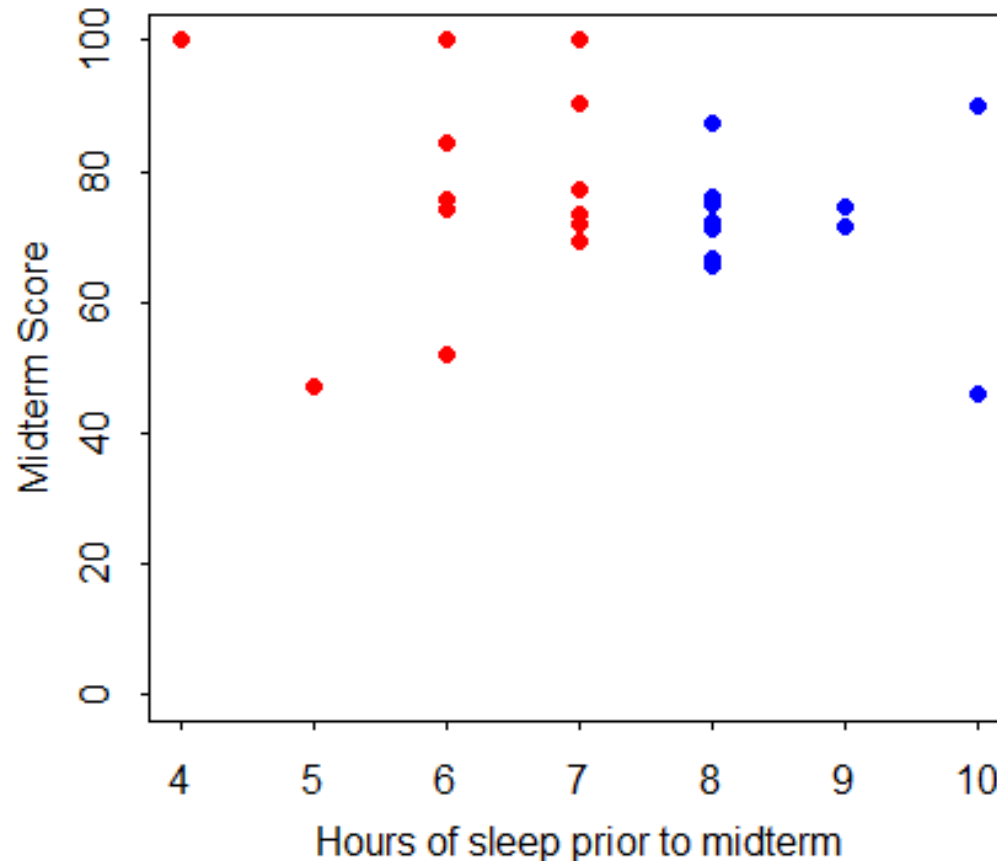
# Quantifying Statistical Evidence

What if we just want to know if there's a difference in midterm score based on whether you slept a short ( $<8$  hours) vs. long ( $\geq 8$  hours) amount of time?



# Quantifying Statistical Evidence

What if we just want to know if there's a difference in midterm score based on whether you slept a short ( $<8$  hours) vs. long ( $\geq 8$  hours) amount of time?

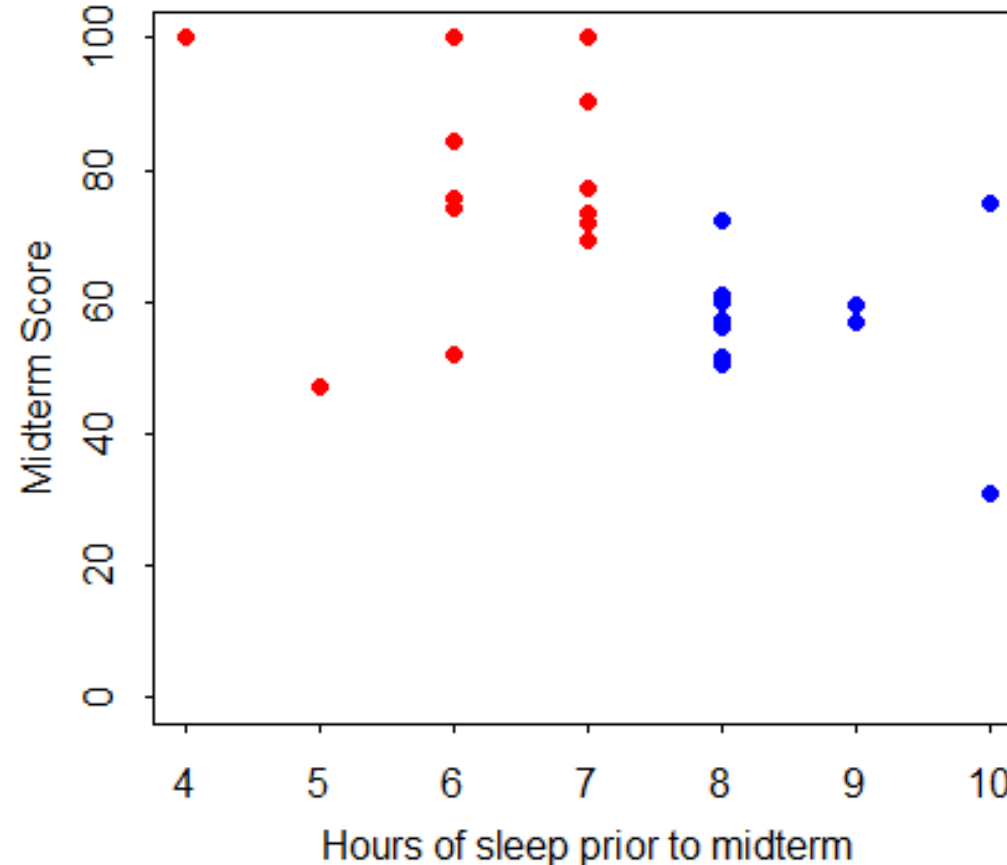


How about now?



# Quantifying Statistical Evidence

What if we just want to know if there's a difference in midterm score based on whether you slept a short ( $<8$  hours) vs. long ( $\geq 8$  hours) amount of time?



How about now?

# The Sample

One of the reasons we need statistics is that we aren't omniscient.

- We don't know who citizens are going to vote for in the next election.
- We don't know which team is going to win the super bowl.
- We don't know what everybody thought about that new restaurant.

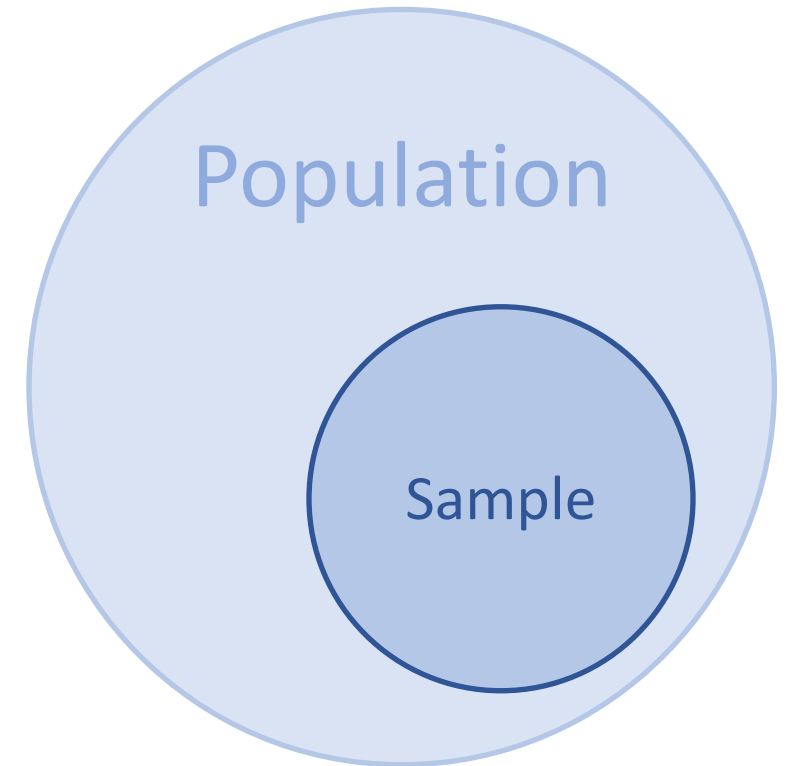
So what do we do?



# Take A Sample!

We sample people/things that belong to the population of interest.

- We don't know who citizens are going to vote for in the next election.
- We don't know what everybody thought about that new restaurant.
- We don't know the vaping habits of everybody in Los Angeles county.

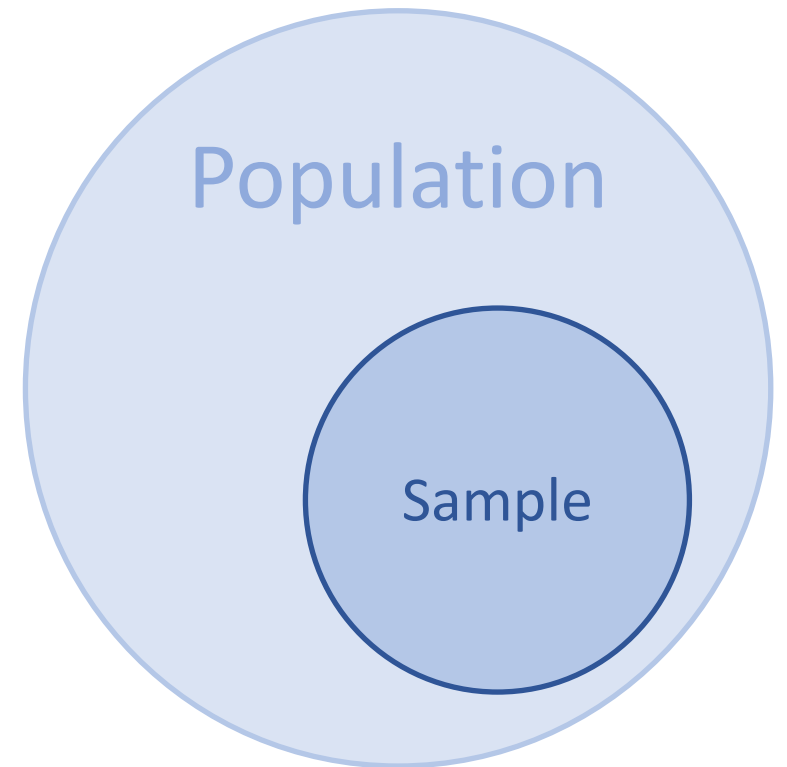


# Take A Sample!

We sample people/things that belong to the population of interest.

- We don't know who citizens are going to vote for in the next election.  
*Ask the people in this room.*
- We don't know what everybody thought about that new restaurant.  
*Check Yelp.*
- We don't know the vaping habits of everybody in Los Angeles county.  
*Take a survey of your friends/family.*

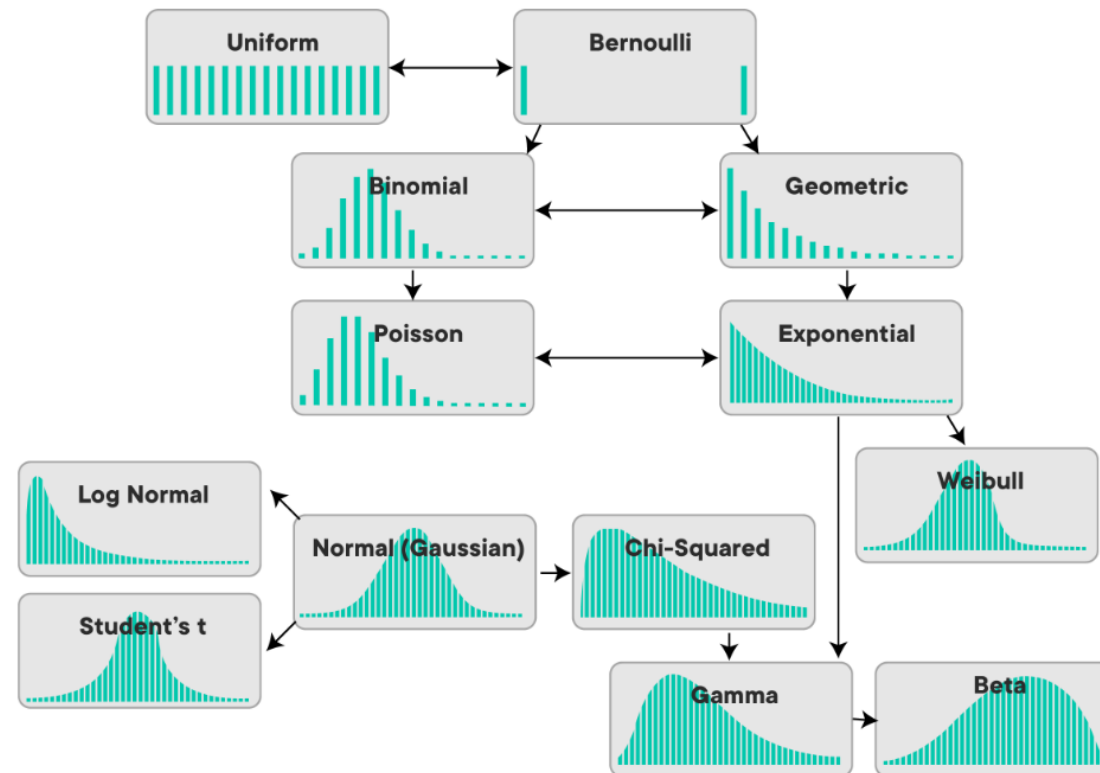
Do you think these are good ways to sample?

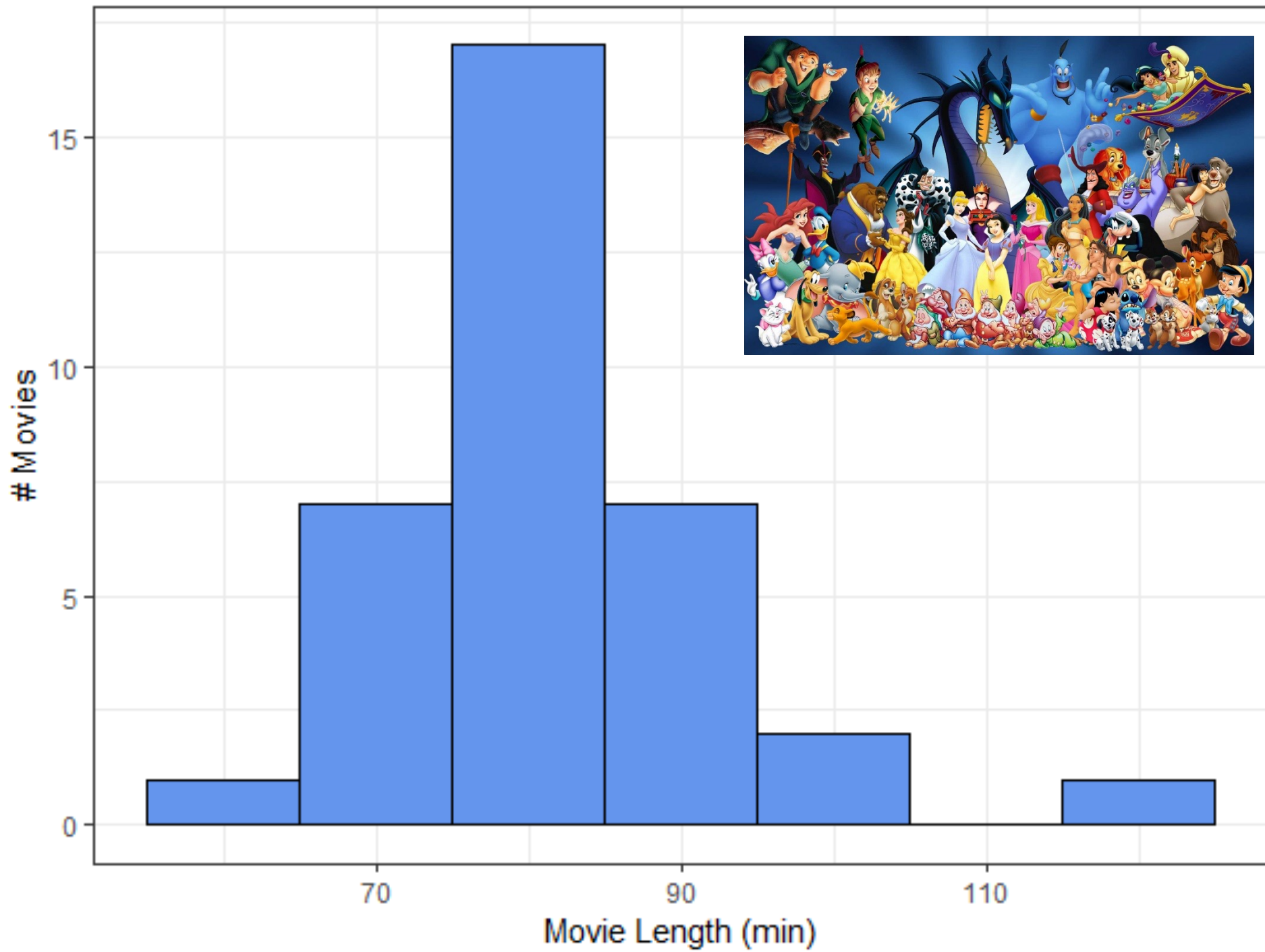


As you can see, this approach is subject to bias.

# Distributions

When we sample, we get data. And when we get data, see that different variables are distributed in different ways.







# Some Notation

These three quantities will be useful

Sum:

$$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

Add up all the values.

Sum of squares:

$$\sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2$$

Square each value and then add them all up.

Square of the sum:

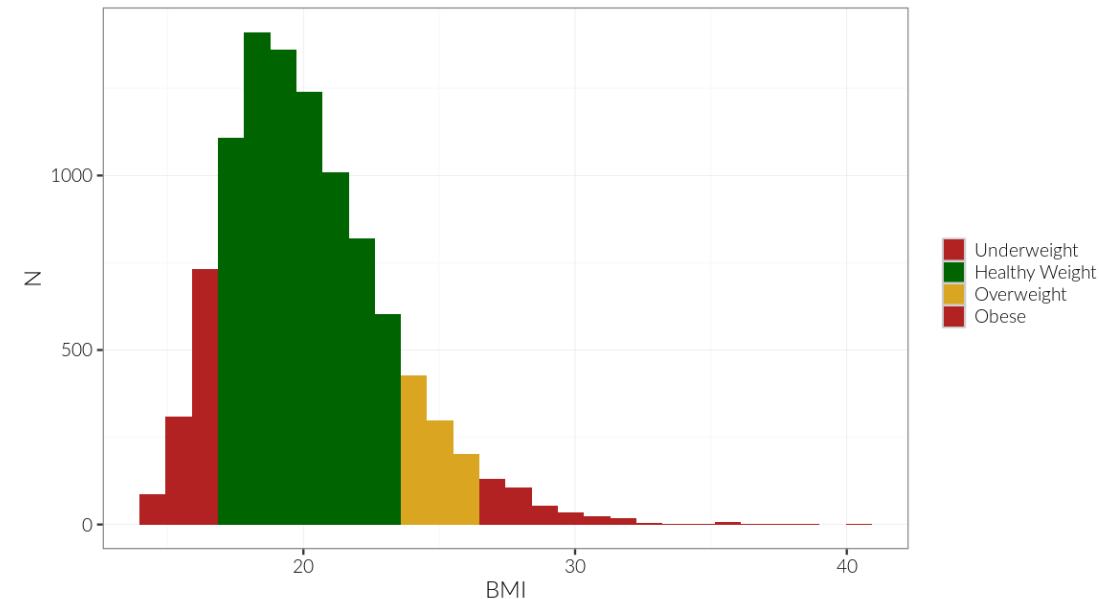
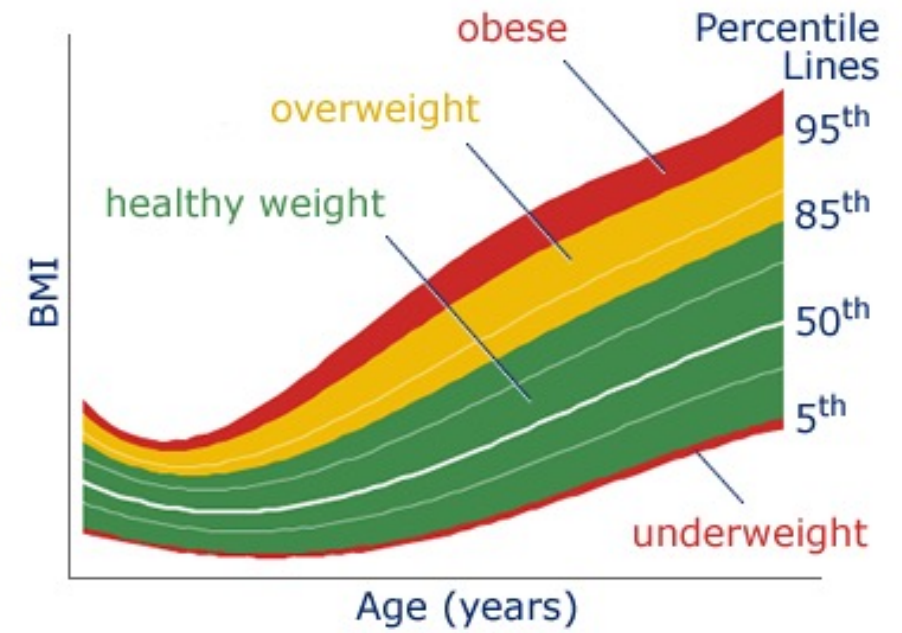
$$\left( \sum_{i=1}^n x_i \right)^2 = (x_1 + x_2 + x_3 + \cdots + x_n)^2$$

Add the values up and then square that result.

# Percentiles

An  $X^{\text{th}}$  percentile is the value of which  $X\%$  of the data is to the left of on the distribution.

Example: BMI in children



# Central Tendency

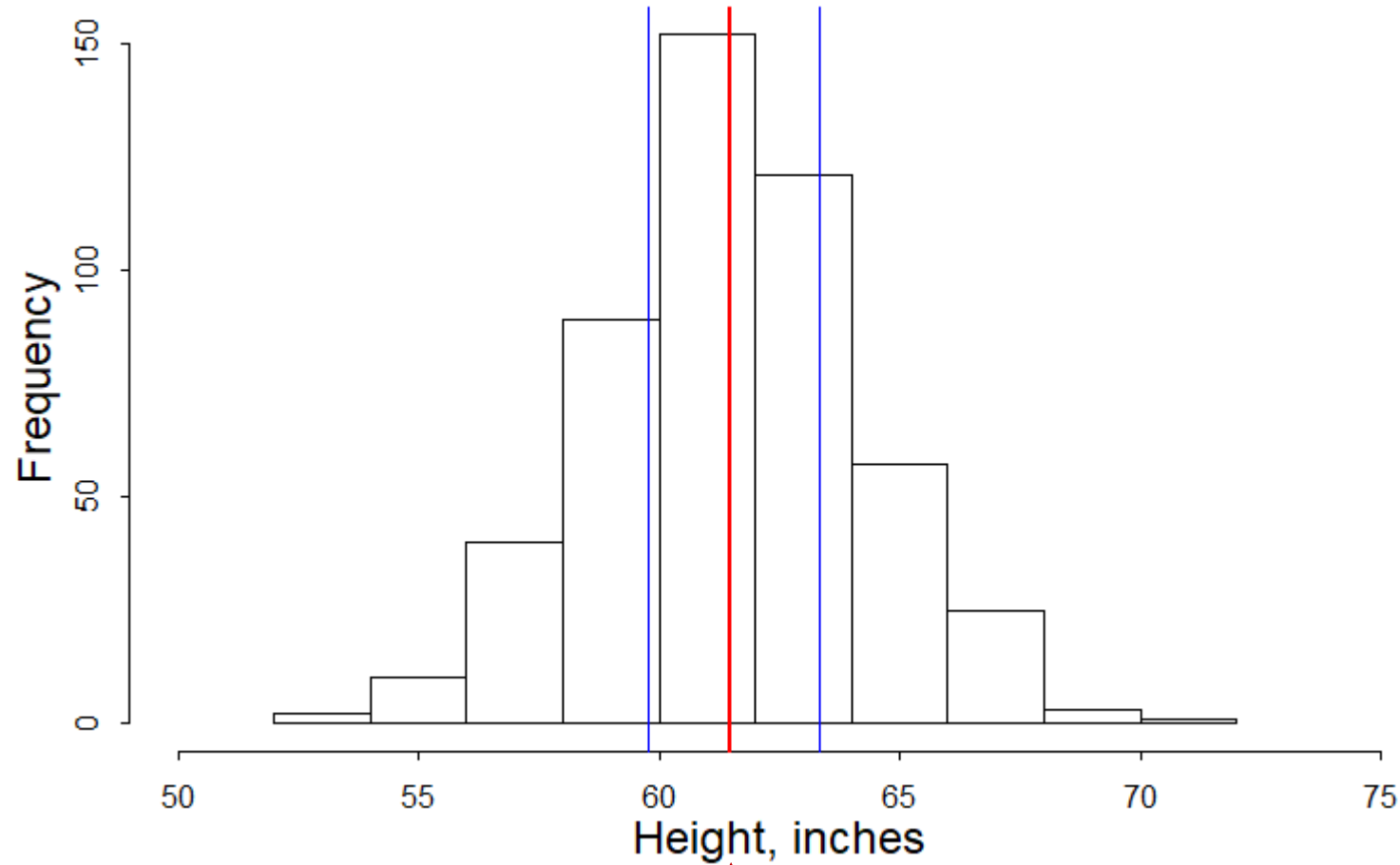
Central Tendency is a way of describing a “typical” value in the data set.

Measures of “central tendency”; “middle” of the data

- **Median:** the 50<sup>th</sup> percentile; the middle value
- **Mean (average):**
- **Mode:** value occurring most often (discrete/categorical data)
  - For continuous data, the value at the “peak”

*Most people have an above-average number of legs.*

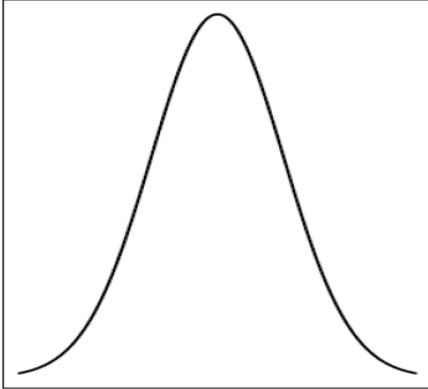
# An Example with Height



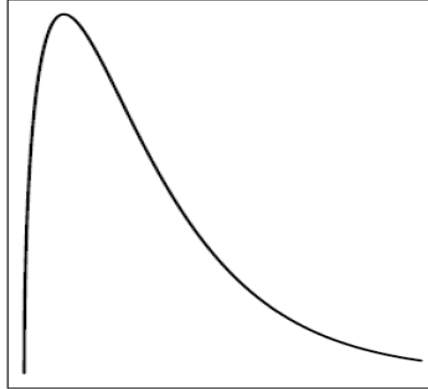
Mean  $\approx$  Median  $\approx$  Mode  
61.5 in

# A Note on Common Distribution Shapes

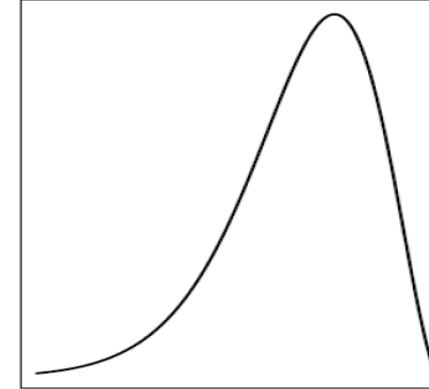
**Symmetrical  
and bell shaped**



**Positively skewed  
or skewed to the right**



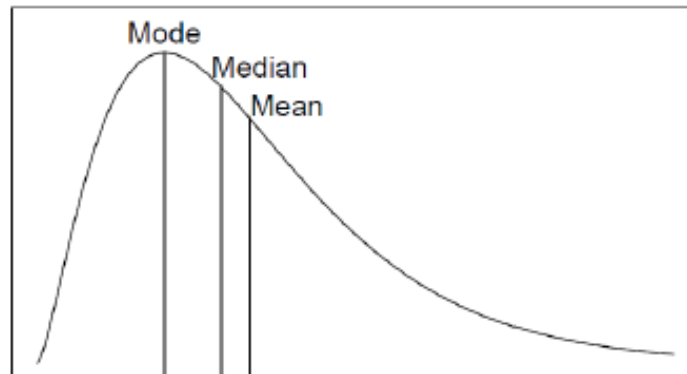
**Negatively skewed  
or skewed to the left**



Positive skew/skewed to the right

- Longer tail in high values
- Mean > median > mode

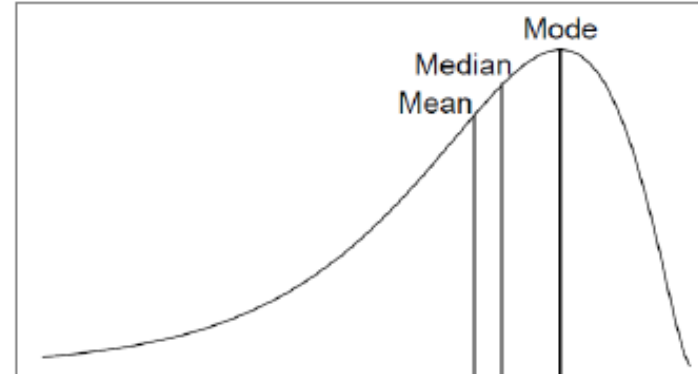
**Positively skewed or skewed to the right**



Negative skew/skewed to the left

- Longer tail in low values
- Mode > Median > Mean

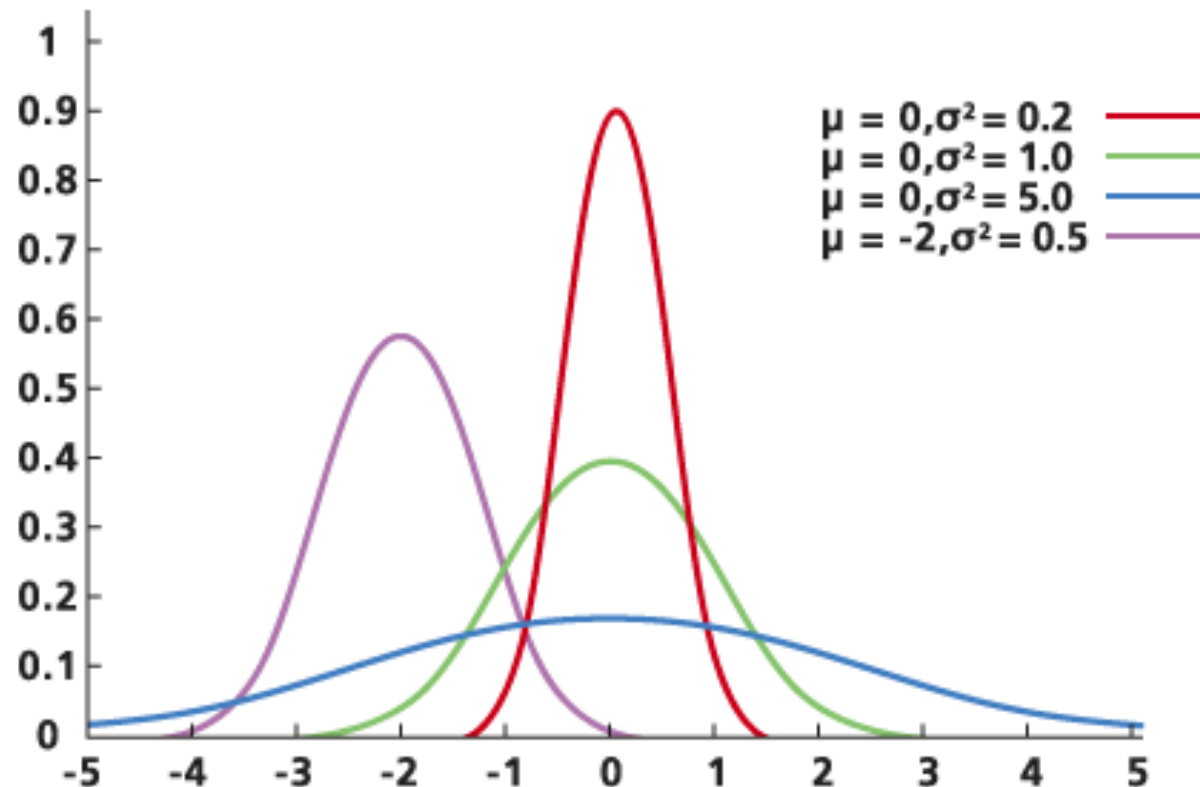
**Negatively skewed or skewed to the left**



What types of variables  
would follow these  
distributions?

# Central Tendency vs. Spread

Measures of central tendency alone do not adequately describe the distribution of the data.

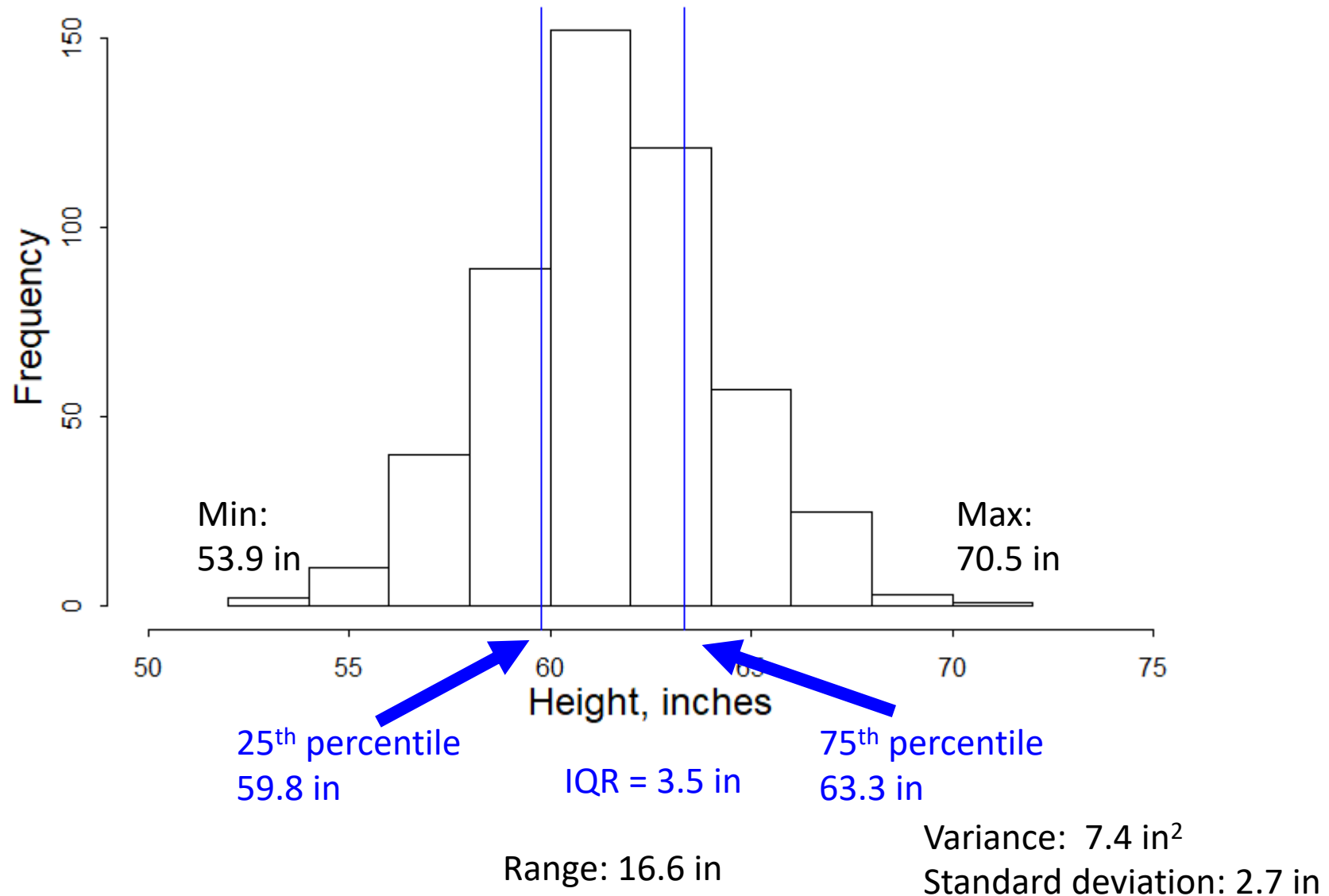


We also need to know  
the spread!

# Measures of Spread

- Range: difference between largest and smallest value
- Interquartile range (IQR):  
difference between 75<sup>th</sup> and 25<sup>th</sup> percentile values
- Variance: “average” squared deviation from the mean
  - units: inches<sup>2</sup>
- Standard deviation (SD): square root of variance
  - unit: inches

# Height: Measures of Spread



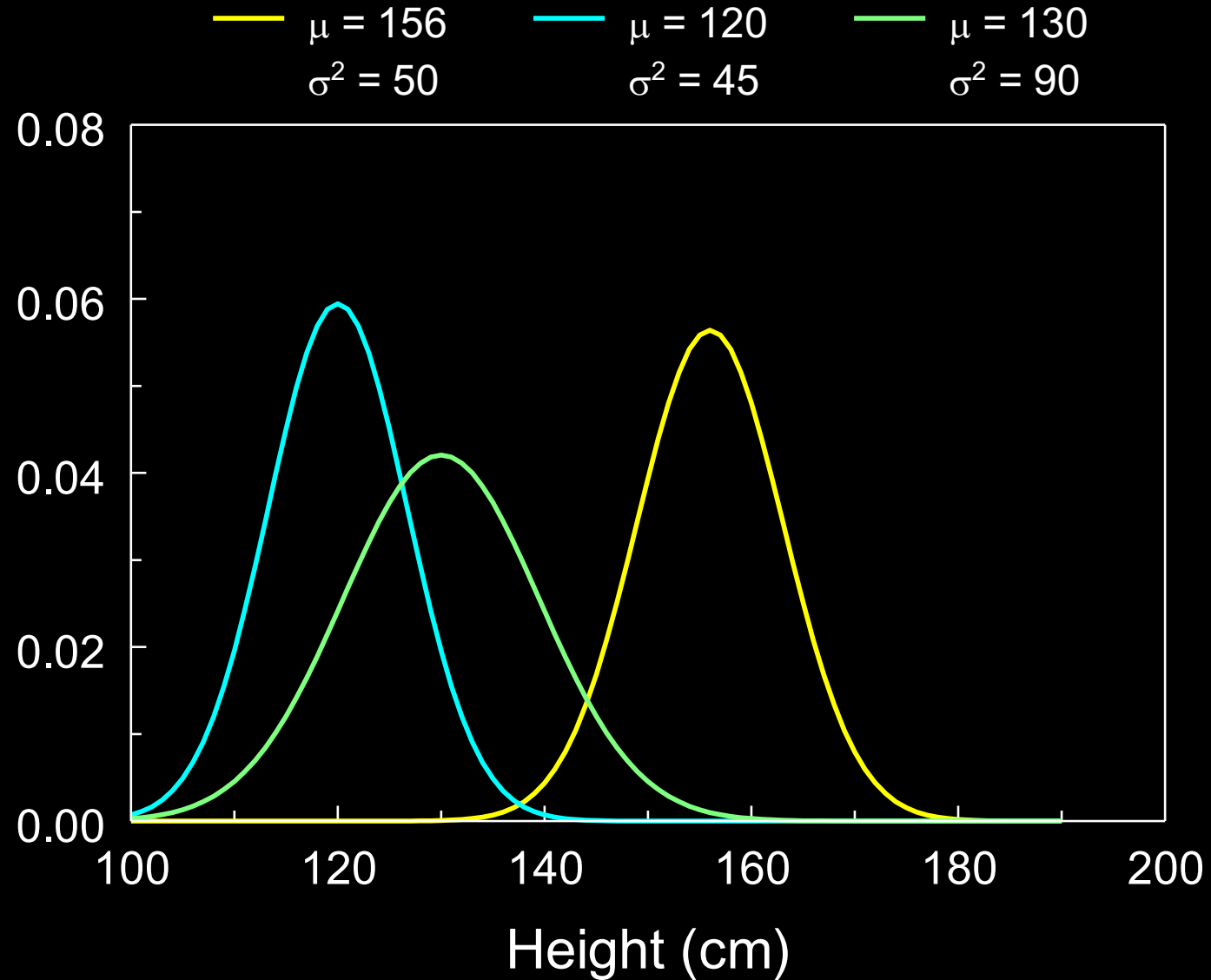


# Standardization: Scaling and Centering

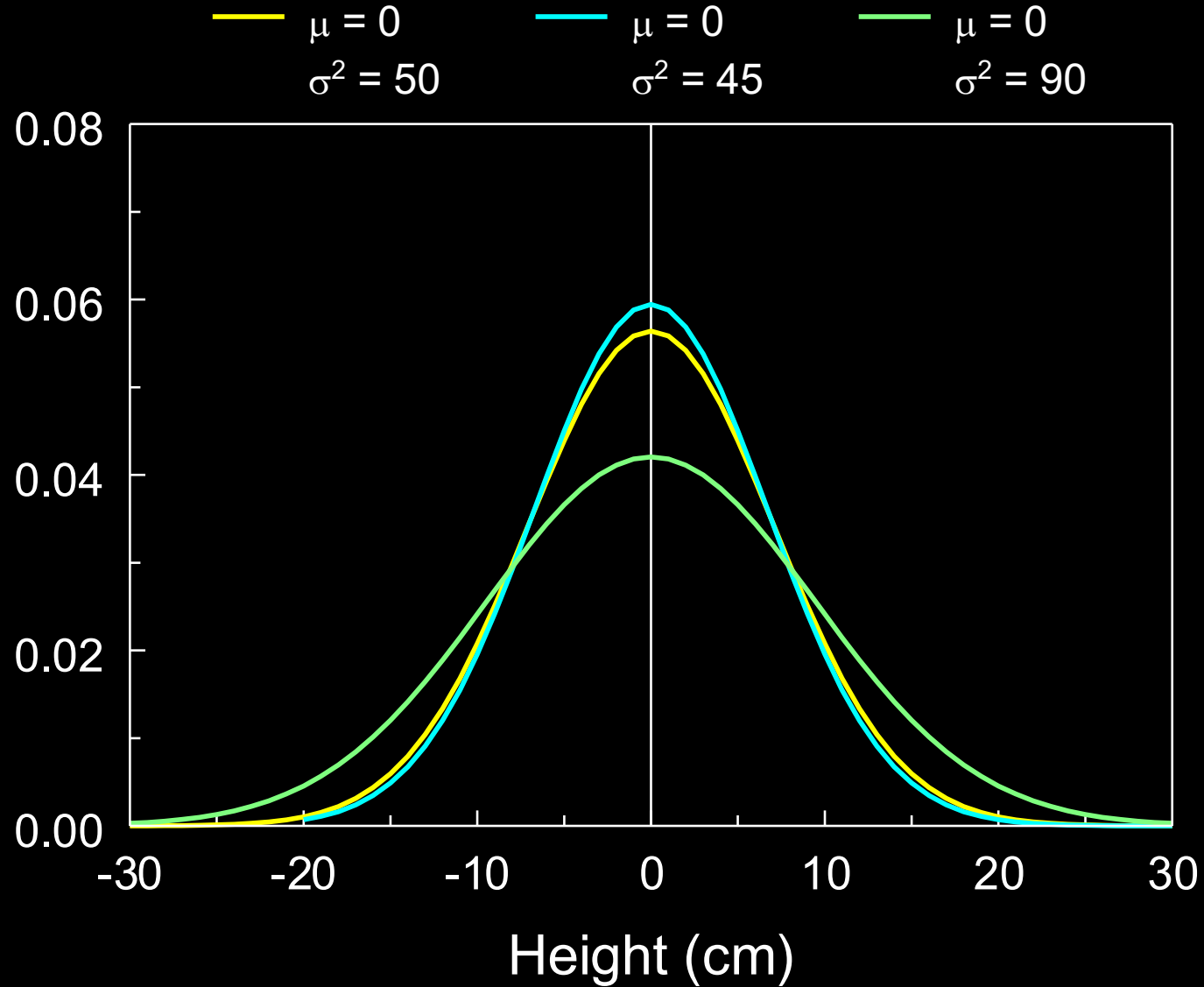
- You might want to compare values coming from distributions in different populations
- Standardization can help:
- “Center” by subtracting the mean  $X'_i = X_i - \bar{X}$ 
  - New mean = 0
- “Scaled” by dividing by the standard deviation
  - New standard deviation = 1
  - Now a 1 unit change in  $X''$  is equivalent to a SD change in  $X$

$$X''_i = \frac{X'_i}{s} = \frac{X_i - \bar{X}}{s}$$

# 3 Distributions - Original

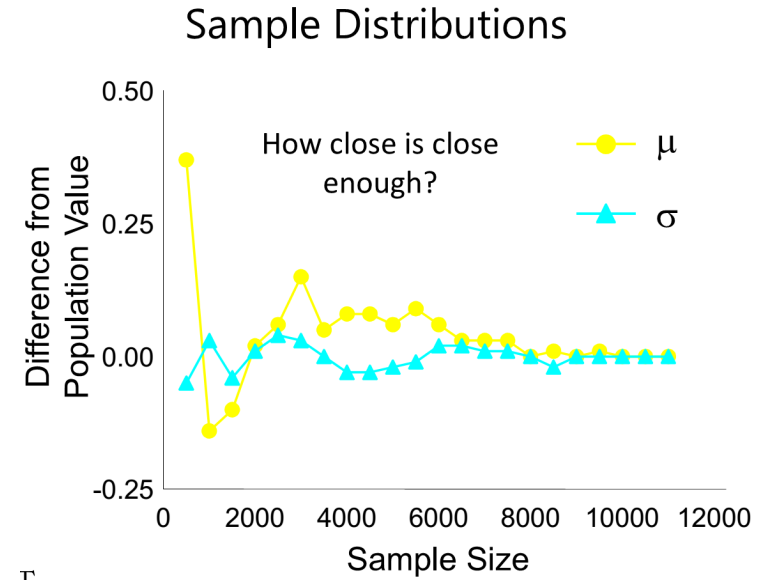


# 3 Distributions - Centered



# Sample vs. Population Distribution

- Typically we never get to observe the actual population distribution
- Estimate population parameters by sampling from the population
- Ability to accurately characterize population depends on representativeness of sample and sample size
- In practice, not all data are normal
- Statistical methods fall into two broad categories
  - Parametric: assumes data follow a given distribution
  - Non-parametric: does not assume a distribution
- Some statistics are more robust than others to deviation from the assumed distribution



# Recap

- “About me” - AKA - “Why the heck do I like stats?”
- What is biostatistics?
- Types of data
- Thinking statistically
- Describing a sample: distributions, percentiles, central tendency, variation