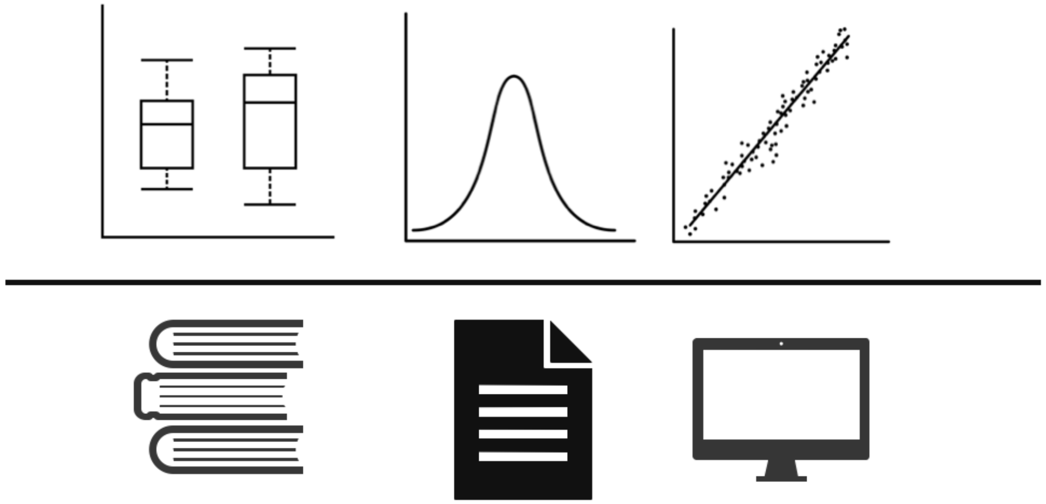
PSYC 2300

Introduction to Statistics



Lecture 02: Central Tendency and Variability

Reminder: Install JASP

Download link here

Outline for Today

Measures of central tendency

- Mean, Median, Mode

Scales of Measurement

- Nominal, Ordinal, Interval, and Ratio scales

Measures of variability

- Range, Standard Deviation, Variance

Measures of Central Tendency

Measures of Central Tendency

Measures of central tendency: Numbers that represent the *center* or *middle* of a distribution of data

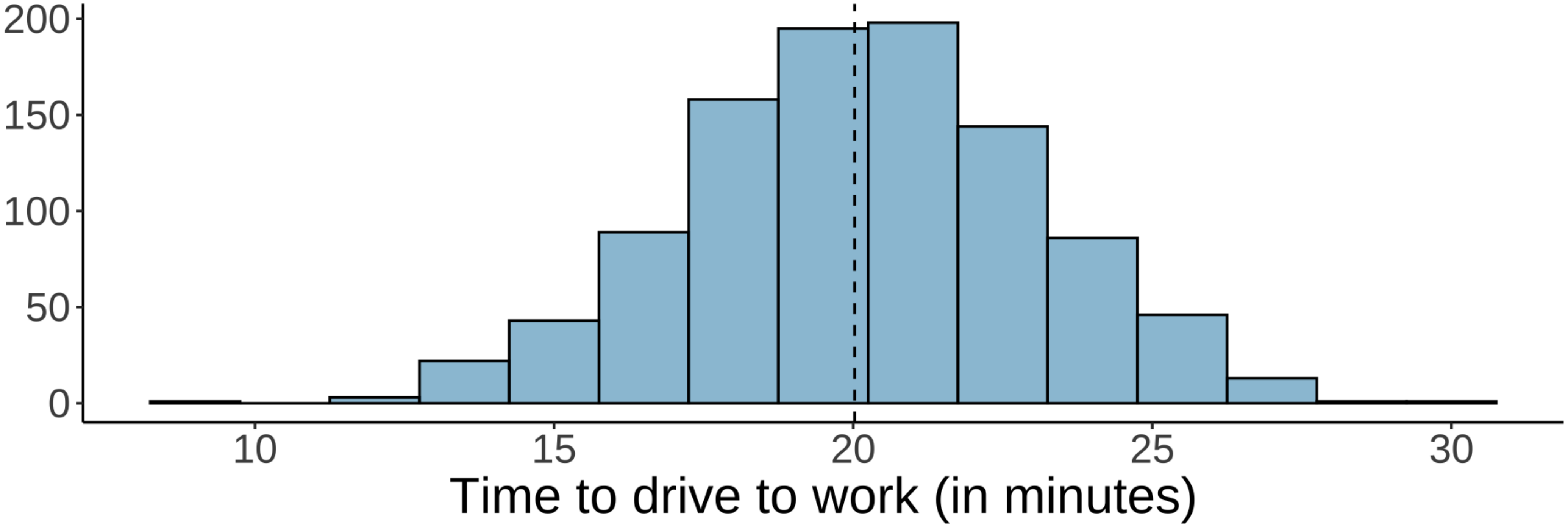
Mean

Median

Mode

The Mean

Mean: The sum of a set of scores divided by the total number of scores in the set



The Mean

Populations

μ

"mew"

Samples

$\bar{x}$

"x-bar"

M

"M"

The Mean: Populations

Calculating the mean for populations

$$\mu = \frac{Sum\ of\ scores}{Number\ of\ scores}$$

Mean for Populations

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

The Mean: Populations

"Take the sum of"

"Each score in the dataset"

Mean for Populations

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

Population mean

Number of scores

The Mean: Populations

i	x
1	6
2	3
3	7
4	12

$\sum X$ means *"Take the sum of all x values"*

$$\sum X = 6 + 3 + 7 + 12 = 28$$

The Mean: Populations

i	x
1	6
2	3
3	7
4	12

$\sum X_i$ uses in *index notation*

$$\sum X_i = X_1 + X_2 + X_3 + X_4$$

$$\sum X_i = 6 + 3 + 7 + 12$$

$$\frac{\sum X_i}{N} = \frac{6+3+7+12}{4} = 7$$

The Mean: Populations

Mean for Populations

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

The Mean: Samples

Mean for Samples

$$\overline{x} = \frac{\sum X_i}{n}$$

Same calculations, different notation

N = population size
n = sample size

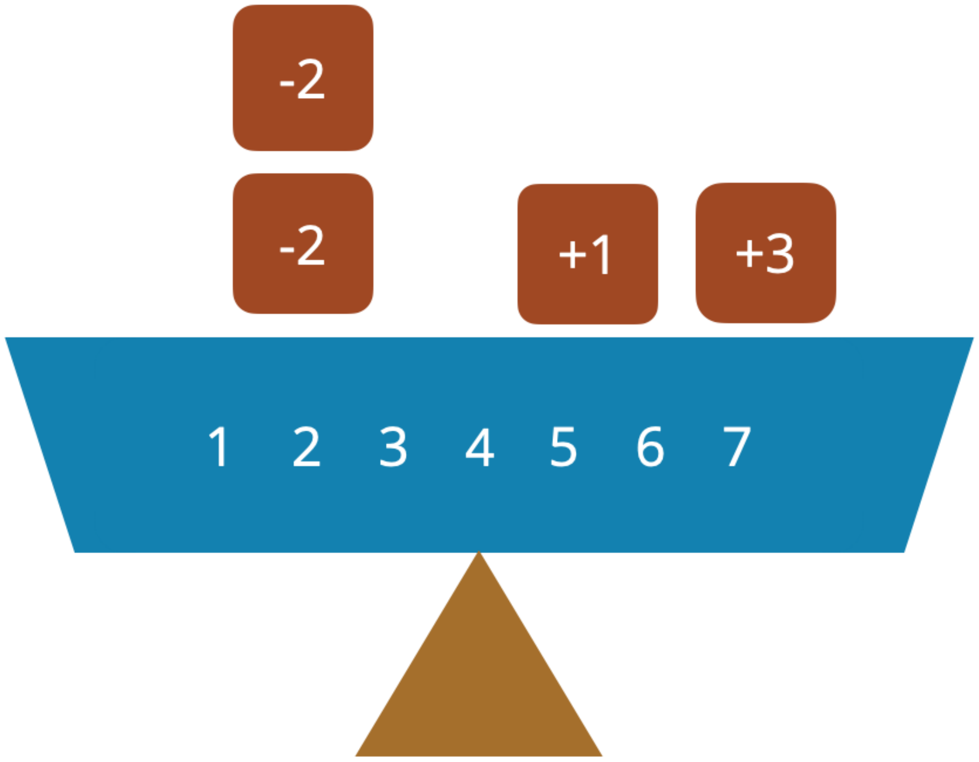
What *is* the mean?

The arithmetic center or "balancing point" of a distribution, where the sum of the signed deviations from the mean always equals zero

Data: $2, 2, 5, 7$

$\overline{x} = 4$

Signed deviations from the mean:
$(-2) + (-2) + (1) + (3) = 0$



Measures of Central Tendency

Measures of central tendency are just that: Numbers that represent the *center* or *middle* of a distribution of data

Mean

Median

Mode

The Median

The *median* is also an “average,” but of a very different kind

Median: the point at which half (50%) of the values are above and half (50%) of the values are below

The Median

Calculating the median

1. List all values in the set in ascending order
2. The **middle-most score** is the median of the set

The Median

Example: Student GPA's

Raw Scores	Ordered Scores
2.85	1.90
2.55	2.55
3.59	2.85
4.00	3.59
1.90	4.00

The **median** is 2.85

The Median

What if you have an even number of values in the set?

Take the **mean** of the two middle-most values

Ordered Scores
1.90
2.55
2.59
3.00
3.15
3.95

$$\frac{2.59+3.00}{2} = 2.795$$

Measures of Central Tendency

Measures of central tendency are just that: Numbers that represent the *center* or *middle* of a distribution of data

Mean

Median

Mode

The Mode

The mode is our third and final measure of central tendency

Mode: the value in a distribution of data that occurs most frequently

2, 4, 5, 5, 6, 7, 7, 7, 8, 9

The beauty of the mode is that you can calculate it even if your dataset doesn’t contain numbers

The Mode

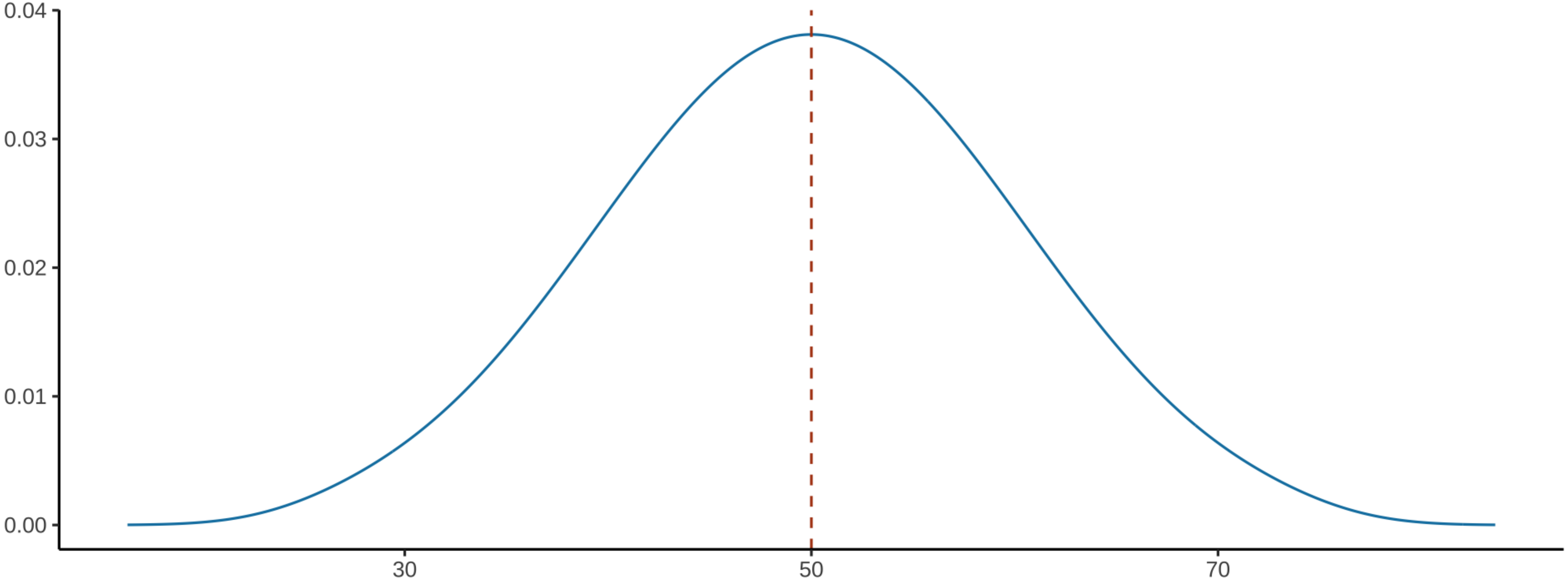
Coffee Shop Customer Data

Date	Customer	Beverage
12/1/2021	Charlie	Espresso
12/5/2021	Tiffany	Coffee
12/17/2021	Hayley	Latte
12/17/2021	Chris	Chai Tea
12/18/2021	Becca	Peppermint Mocha
12/18/2021	Laura	Peppermint Mocha
12/18/2021	Jill	Peppermint Mocha

Beverage	Frequency
Espresso	1
Coffee	1
Latte	1
Chai Tea	1
Peppermint Mocha	3

The mean ain't perfect

In a perfectly “normal” (bell-curve) distribution, Mean = Median = Mode = 50

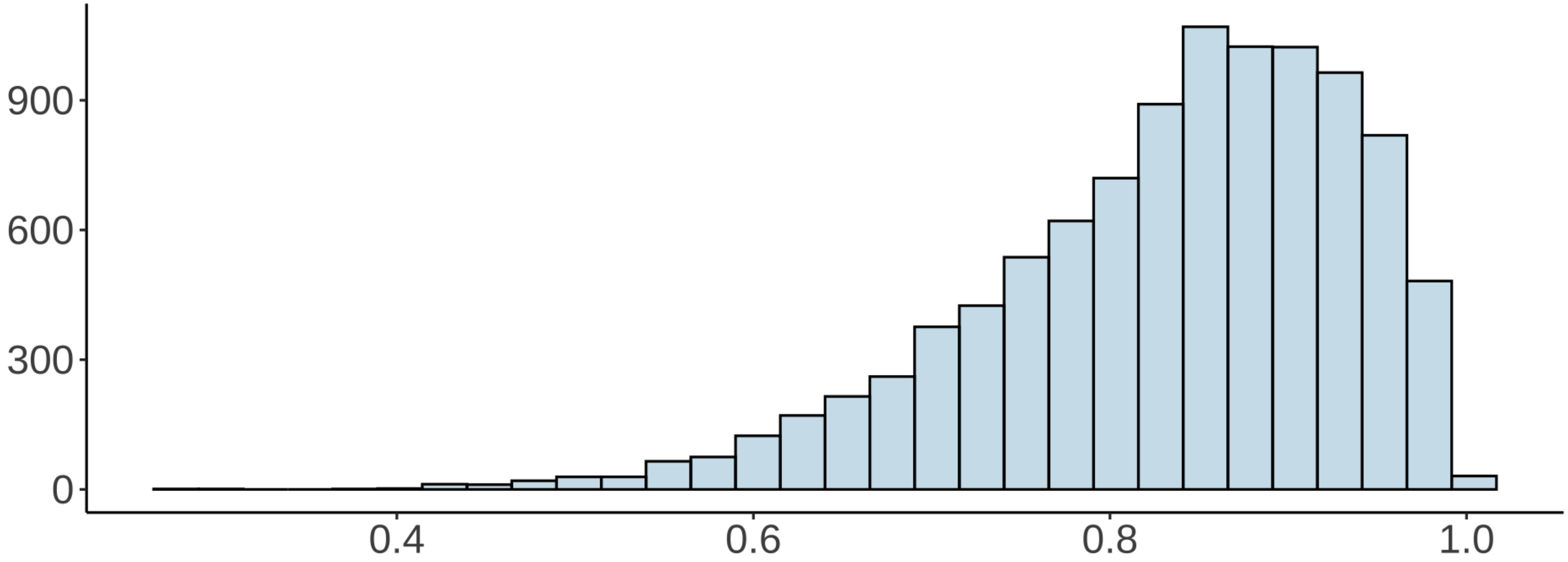


The mean ain't perfect

But in a non-normal (or “skewed”) distribution, the three are differentially influenced:

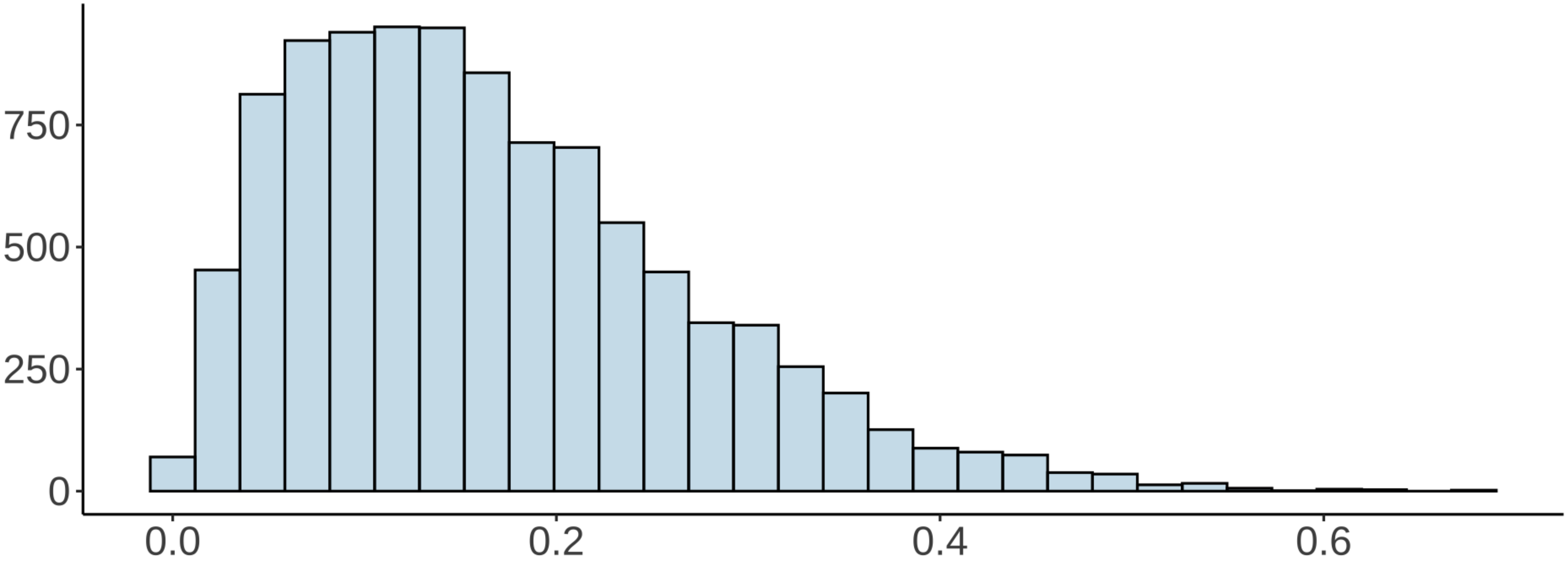
Mean	Median	Mode
Heavily influenced by skew	Moderately influenced by skew	Relatively immune to skew

The mean ain't perfect



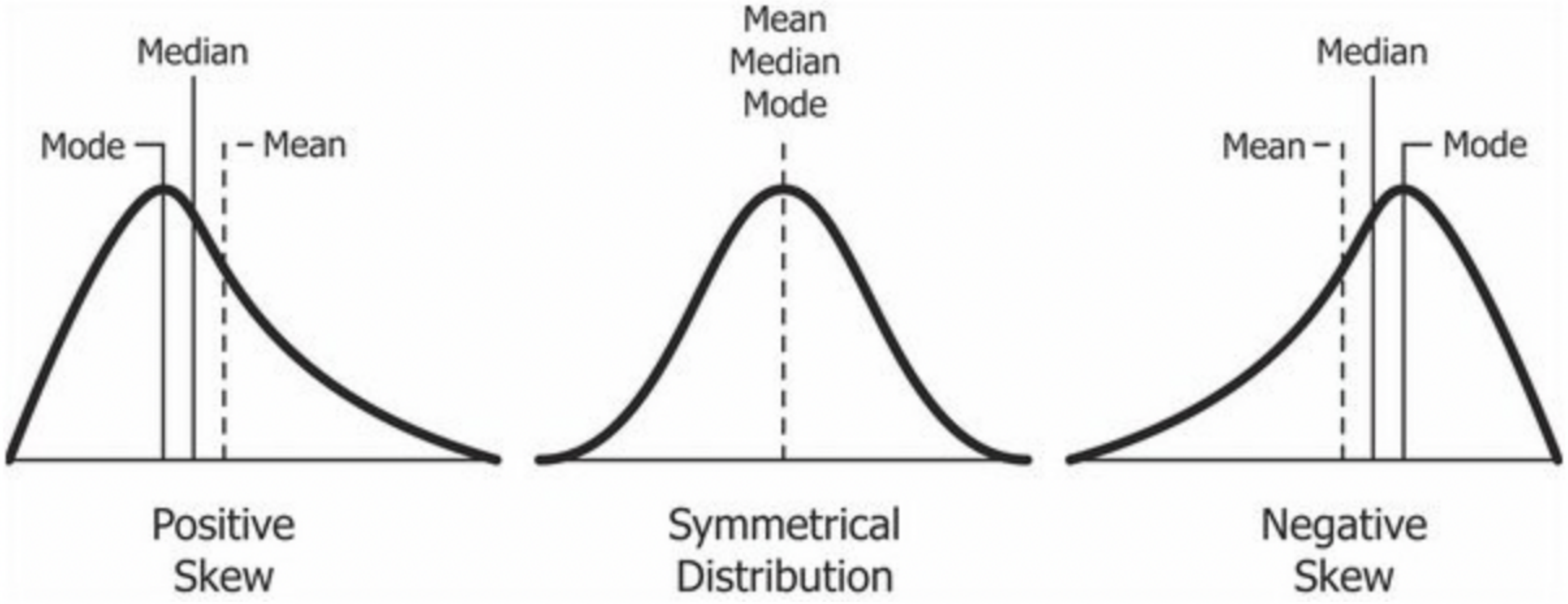
This distribution is “negatively” or **left-skewed**

The mean ain't perfect



This distribution is “positively” or **right-skewed**

The mean ain't perfect



The mean ain't perfect

The mean is especially susceptible to *extreme values* in a dataset

$$1, 1, 3, 3, 3, 4, 100$$

$\overline{x} = 16.43$

Median $= 3$

Mode $= 3$

Scales of Measurement

Scales of Measurement

We have three options for measures of central tendency – how do we know when to use which?

Scales of measurement: describes the nature of the information contained in a given set of data

Scales of Measurement

Nominal

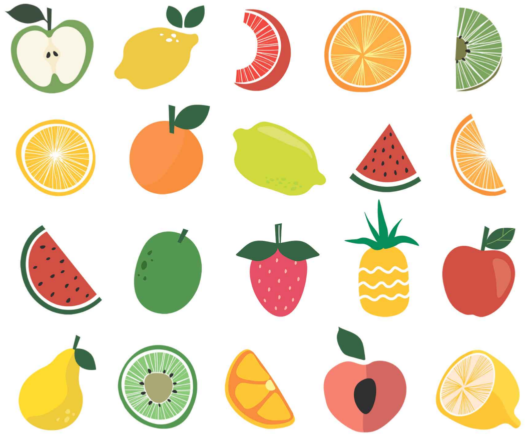
Ordinal

Interval

Ratio

Nominal Scale

- Non-numerical (can only be *qualitative*)
- Each item in the set belongs to a class or category



What are some other examples of nominally-scaled data?

-

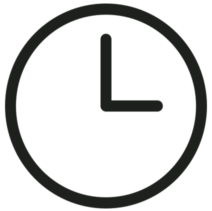
Ordinal Scale

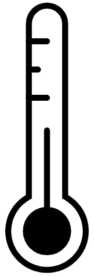
- Items are **ordered** in a meaningful direction
    - Can be quantitative or qualitative
    - The “ord” stands for “order”
    - Distance between items is *not necessarily* equal

What are some other examples of ordinally-scaled data?

Interval Scale

- Numerical (can only be quantitative)
- Distance between points is **equal** and **meaningful**
- But relationship between points is *not* meaningful
- Can have values below 0



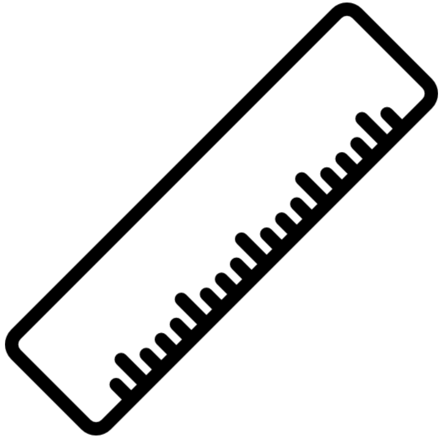


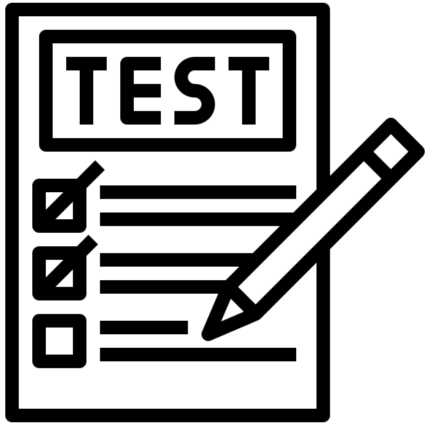
Example: degrees in Fahrenheit

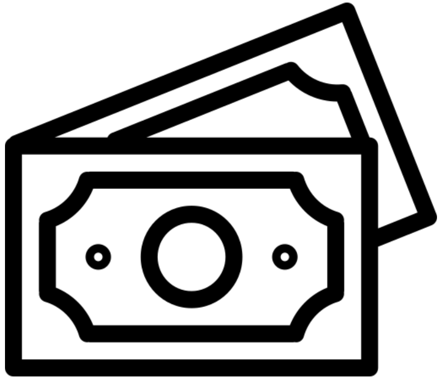
- 5 $^\circ F$ vs. 6 $^\circ F$ is a difference of 1 $^\circ F$
- A 1 $^\circ F$ degree difference is meaningful because it will always be 1 degree hotter
- **But**, 10 $^\circ F$ is *not* twice as hot as 5 $^\circ F$

Ratio Scale

- Numerical (can only be quantitative)
- All the qualities of the interval scale plus a **true zero point**
    - The *absence* of whatever is being measured is possible
- Relationship between points is **meaningful**







Scales of Measurement: Test Yourself

Variable	Level of Measurement
Eye color	
Rating of well-being on a 5-point scale	
Reaction time at a computer task	
Order of finishers in a 5K race	
Parents’ marital status	
Blood alcohol content	

Nominal

Ordinal

Interval

Ratio

00:30

Scales of Measurement: Test Yourself

Variable	Level of Measurement
Eye color	Nominal
Rating of well-being on a 5-point scale	Ordinal
Reaction time at a computer task	Ratio
Order of finishers in a 5K race	Ordinal
Parents’ marital status	Nominal
Blood alcohol content	Ratio

Scales of Measurement

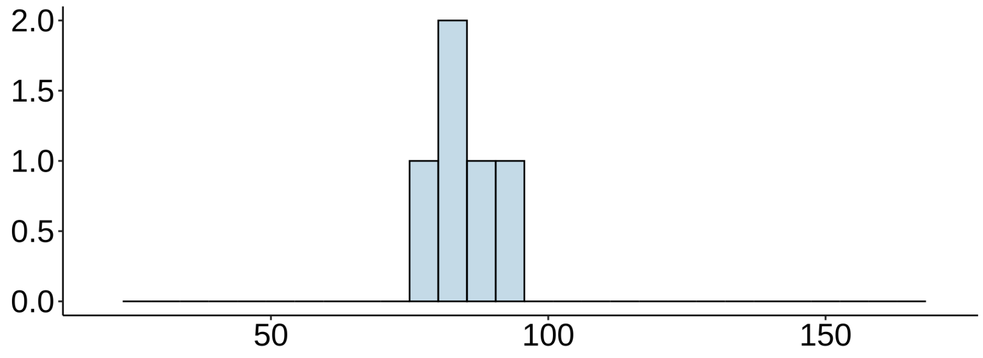
The relationship between measures of central tendency and scales of measurement

	Nominal	Ordinal	Interval	Ratio
Mean	-	-	✓	✓
Median	-	✓	✓	✓
Mode	✓	✓	✓	✓

Measures of Variability

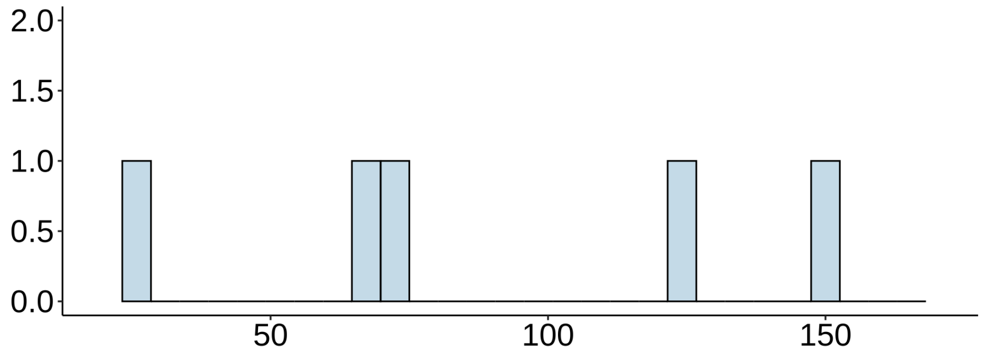
Why Variability Matters

$$80, 85, 85, 90, 95$$



$\overline{x}$ = 87

$$25, 65, 70, 125, 150$$



$\overline{x}$ = 87

Measures of Variability

How can we describe these differences statistically?

In statistics, **measures of variability** describe how scores in a given dataset differ from one another (e.g., the spread or clustering of points)

Range

Standard Deviation

Variance

The Range

The range is the simplest measure of variability (or dispersion), and is defined as follows:

Range

$$r = h - l$$

h = highest score in the set

l = lowest score in the set

The Range

Sample A

$$80, 85, 85, 90, 95$$

$$\bar{x} = 87$$

$$r = 95 - 80 = 15$$

Sample B

$$25, 65, 70, 125, 150$$

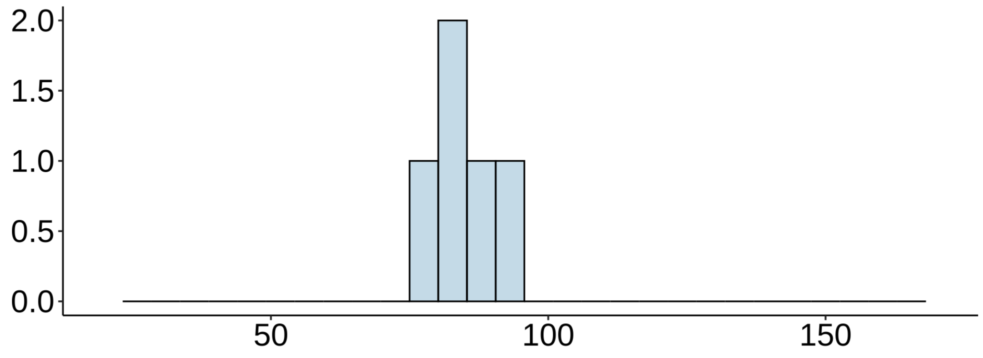
$$\bar{x} = 87$$

$$r = 150 - 25 = 125$$

- Although the range is a simple and useful calculation, it can often miss important information of a dataset's variability

The Range

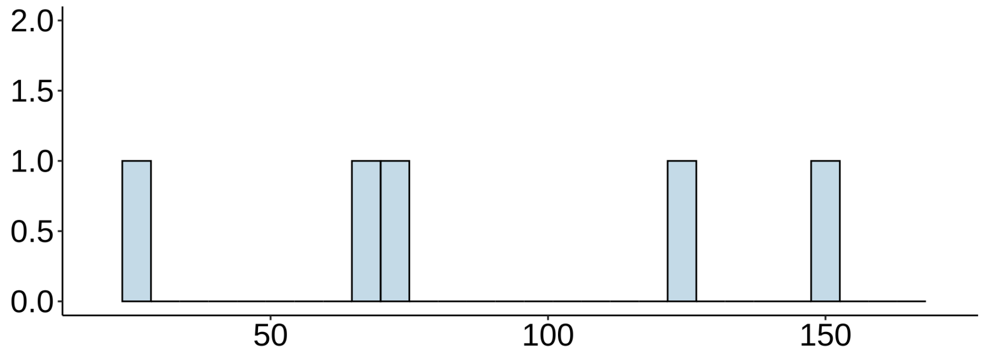
$$80, 85, 85, 90, 95$$



$\overline{x}$ = 87

$$r = 15$$

$$25, 65, 70, 125, 150$$



$\overline{x}$ = 87

$$r = 125$$

Measures of Variability

How can we describe these differences statistically?

In statistics, **measures of variability** describe how scores in a given dataset differ from one another (e.g., the spread or clustering of points)

Range

Standard Deviation

Variance

Standard Deviation

The *standard deviation* takes into account how far each point in a set is from the mean of the set

Standard Deviation: The standard (or typical) amount that scores deviate from the mean

Standard Deviation

Populations

σ

"sigma"

Samples

S

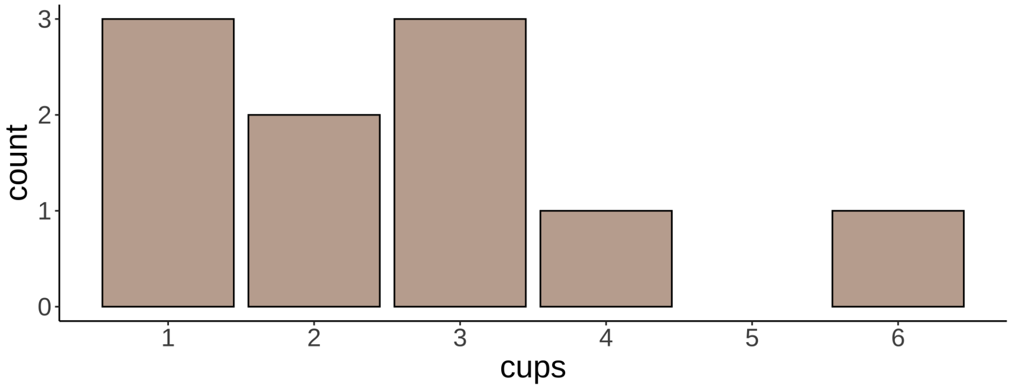
"s"

Standard Deviation

Population: University of Denver PhD students, N = 10

How many cups of coffee do you drink each day?

$$1, 2, 1, 4, 3, 3, 6, 1, 2, 3$$



Standard Deviation: Calulation

Standard Deviation: The standard (or typical) amount that scores deviate from the mean

$$deviation_i = X_i - \mu$$

Standard Deviation: Calulation

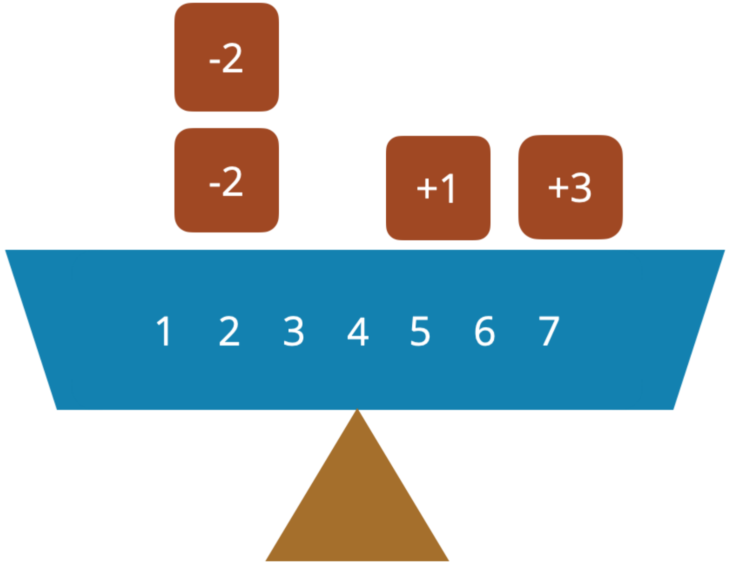
i	x	$x_i - \mu$
1	1	-1.6
2	2	-0.6
3	1	-1.6
4	4	1.4
5	3	0.4
6	3	0.4
7	6	3.4
8	1	-1.6
9	2	-0.6
10	3	0.4

μ = 2.6

$$\sum(X_i - \mu) = 0$$

Remember this?

The arithmetic center or "balancing point" of a distribution, where the sum of the signed deviations from the mean always equals zero



$$(-2) + (-2) + (1) + (3) = 0$$

Standard Deviation: Calulation

i	X	$X_i - \mu$
1	1	-1.6
2	2	-0.6
3	1	-1.6
4	4	1.4
5	3	0.4
6	3	0.4
7	6	3.4
8	1	-1.6
9	2	-0.6
10	3	0.4

μ = 2.6

$$\sum(X_i - \mu) = 0$$

Since we can't do much with 0, we'll need to do something with the negative signs. Only then will we be able to work with these numbers to find the standard deviation.

What's one possible way to handle these negative signs?

Standard Deviation: Calculation

Absolute Deviation

$$|X_i - \mu|$$

Squared Deviation

$$(X_i - \mu)^2$$

Squared deviations (or “squares”) have been shown to better approximate the population, so we use it when calculating standard deviations

Standard Deviation: Calculation

i	X	$X_i - \mu$	$(X_i - \mu)^2$
1	1	-1.6	2.56
2	2	-0.6	0.36
3	1	-1.6	2.56
4	4	1.4	1.96
5	3	0.4	0.16
6	3	0.4	0.16
7	6	3.4	11.56
8	1	-1.6	2.56
9	2	-0.6	0.36
10	3	0.4	0.16

$$\sum (X_i - \mu)^2 = 22.4$$

Sum of Squares

What we just calculated is important for statistics and is called the **sum of squared deviations** or simply the "sum of squares" (SS)

Sum of Squares

$$\sum(X_i - \mu)^2$$

Standard Deviation: Calculation

We can use the sum of squares to calculate the standard deviation of scores from the mean:

$$\sigma = \sqrt{\frac{SS}{N}} = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

Standard Deviation: Calculation

Back to the coffee example.

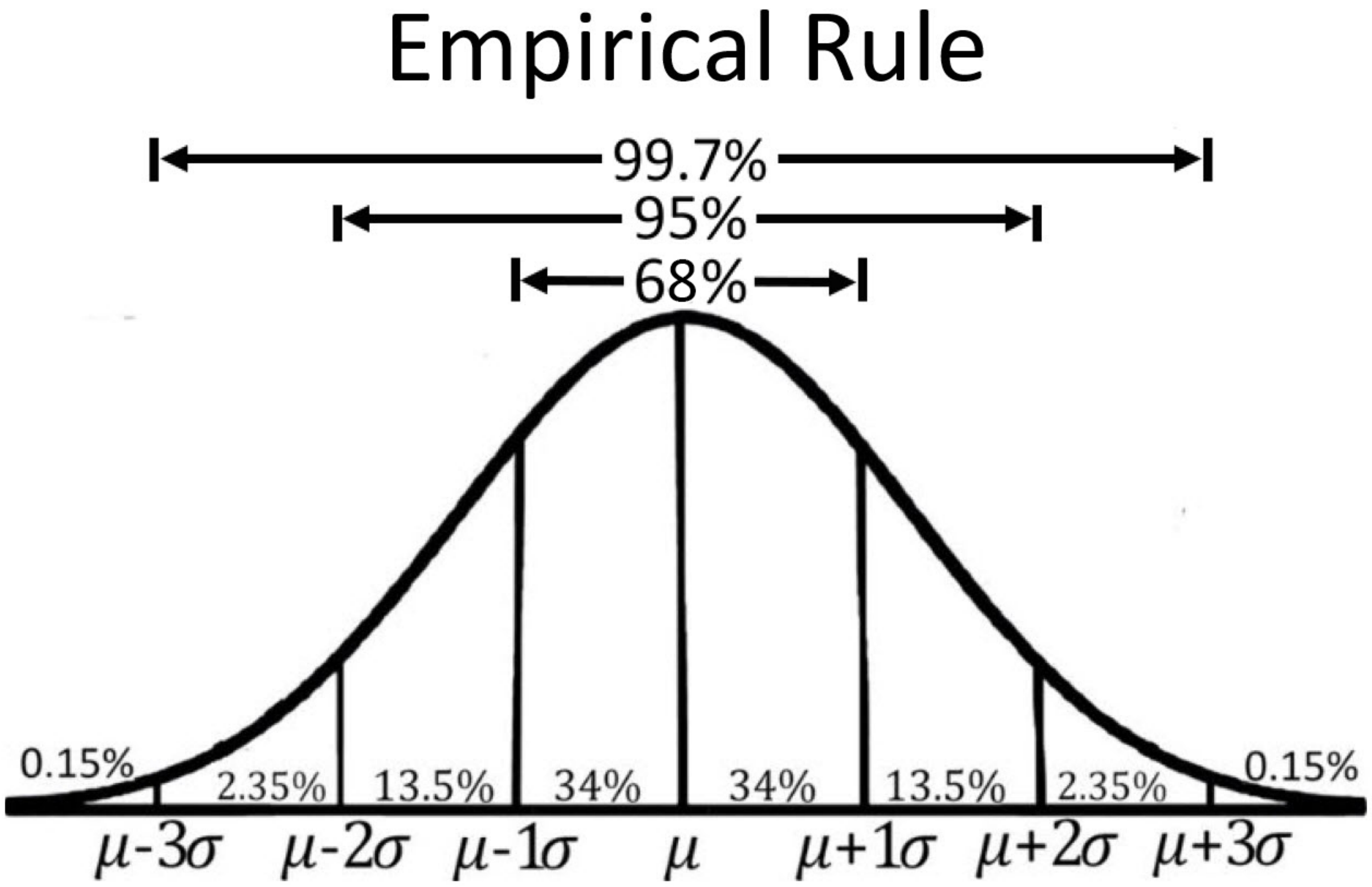
$$N = 10$$

$$\mu = 2.6$$

$$\sum (X_i - \mu)^2 = 22.4$$

$$\sigma = \sqrt{\frac{SS}{N}} = \sqrt{\frac{22.4}{10}} = 1.5$$

Why is this useful?



Measures of Variability

How can we describe these differences statistically?

In statistics, **measures of variability** describe how scores in a given dataset differ from one another (e.g., the spread or clustering of points)

Range

Standard Deviation

Variance

Variance

Variance is the averaged **squared** deviation from the mean

Population Standard Deviation
"Sigma"

$$\sigma = \sqrt{\frac{SS}{N}}$$

Population Variance
“Sigma squared”

$$\sigma^2 = \frac{SS}{N}$$

Variance

This leads us to our final formula for variance:

Population Variance

$$\sigma^2 = \frac{SS}{N} = \frac{\sum(X_i - \mu)^2}{N}$$

Standard Deviation vs. Variance

Populations

$$\sigma = \sqrt{\sigma^2}$$

Samples

$$s = \sqrt{s^2}$$

Whether for populations or samples, the standard deviation is equal to the square root of variance.

Calculating the Variance

Back to the coffee example.

$$N = 10$$

$$\mu = 2.6$$

$$\sum (X_i - \mu)^2 = 22.4$$

$$\sigma^2 = \frac{SS}{N} = \frac{22.4}{10} = 2.24$$

$$\sqrt{2.24} = 1.5 = \sigma$$

What About *Samples*?

- The formula thus far have been for **populations**, but usually you calculate these descriptive statistics for **samples.** What changes?

What About *Samples*?

Population Parameter

Standard Deviation

$$\sigma = \sqrt{\frac{\sum(X_i - \mu)^2}{N}}$$

Variance

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$$

Sample Statistic

Standard Deviation

$$s = \sqrt{\frac{\sum(x_i - \overline{x})^2}{(n-1)}}$$

Variance

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{(n-1)}$$

Why *n - 1* instead of *N*?

Why not just keep using N (instead of $n - 1$) in the sample statistics?

- **Answer**: Turns out that $n - 1$ is better because it estimates the population parameters better (i.e., it is an **unbiased** estimator of the population)

Next time

Lecture

- Visualizing data with graphs

Reading

- Chapter Two
- Chapter Three