CH. 3 – Measures of Central Tendency and Spread

PY 221 Statistics for Research

Dr. Valenti

Reminders

- Make sure to set a reminder for yourself to complete a LSR T, W, Th

- Lab #1 due next Tuesday by 12:30 pm – turn in to Moodle.
- Student office hours are:
  - Tues 8:30 – 9:30 / Wed 2:15 – 4:15 / Thurs 9:30 – 10:30

- The peer-tutors have availability M-F
  - Alanna Gaines (who you met yesterday), Ada Weems, Mary Blake Zeron
  - Hanna McNamara (our class' "embedded peer-tutor")

Outline for Ch. 3

1. Frequency distributions
   - Including skew and outliers
2. Measures of central tendency
3. Measures of spread
4. Combining central tendency & spread

Frequency distribution

*How are people **distributed** across the various answer choices?*

*What is the **frequency** of people who gave each answer?*

Please indicate your gender identity by choosing one option.

1. Male
2. Female
3. Transgender male
4. Transgender female
5. Not listed
6. Non-binary

Frequency distribution in TABLE form

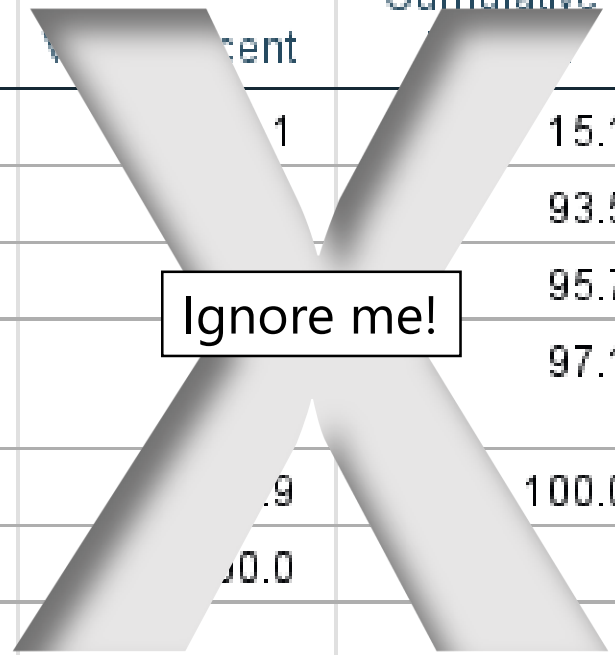
*How are people **distributed** across the answer choices?*

Please indicate your gender identity.

1. Male
2. Female
3. Transgender male
4. Transgender female
5. Not listed
6. Non-binary

gender

		# of Ps Frequency	% of Ps Percent	[illegible]cent	Cumulative
Valid	1 Male	21	14.7	1	15.1
	2 Female	109	76.2		93.5
	3 Transgender male	3	2.1		95.7
	5 Not listed (please specify if you wish)	2	1.4		97.1
	6 Non-binary	4	2.8	.9	100.0
	Total	139	97.2	0.0	
Missing	System	4	2.8		
Total		143	100.0		



Frequency distribution in TABLE form

*How are people **distributed** across the answer choices?*

How much do you like or dislike cats?

extremely dislike
moderately dislike
somewhat dislike
neither dislike nor like
somewhat like
moderately like
extremely like

How much do you like or dislike cats?

		# of Ps Frequency	% of Ps Percent	Percer
Valid	Somewhat dislike them	1	5.3	
	Somewhat like them	4	21.1	
	Moderately like them	8	42.1	
	Extremely like them	6	31.6	31
	Total	19	100.0	100.0

Ignore me!

Frequency distribution in GRAPH form (*aka* a Histogram)

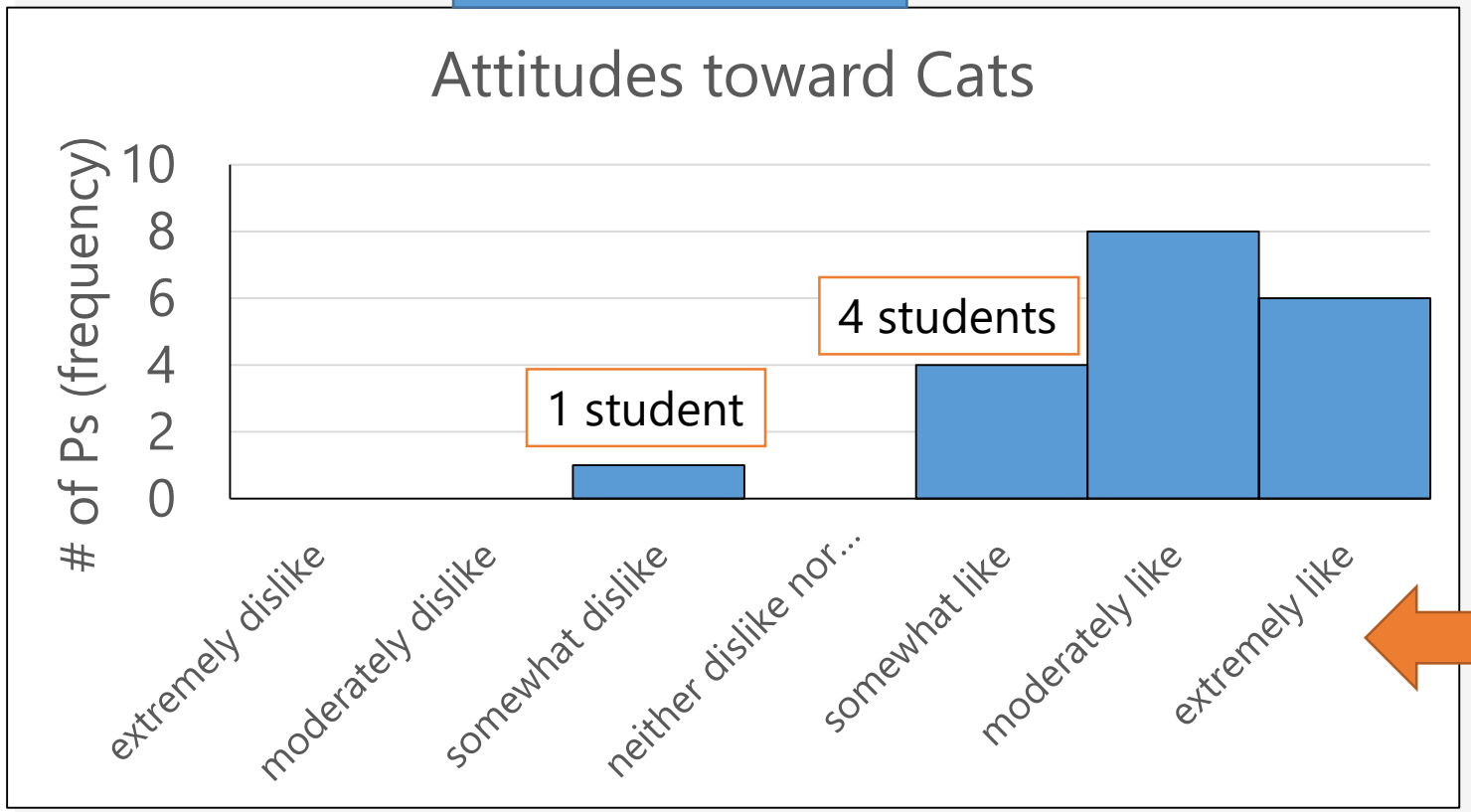
How much do you like or dislik

FREQUENCY **TABLE**

# of Ps

		Frequency	Percent
Valid	Somewhat dislike them	1	5.3
	Somewhat like them	4	21.1
	Moderately like them	8	42.1
	Extremely like them	6	31.6
	Total	19	100.0

HISTOGRAM



How much do you like or dislike cats?

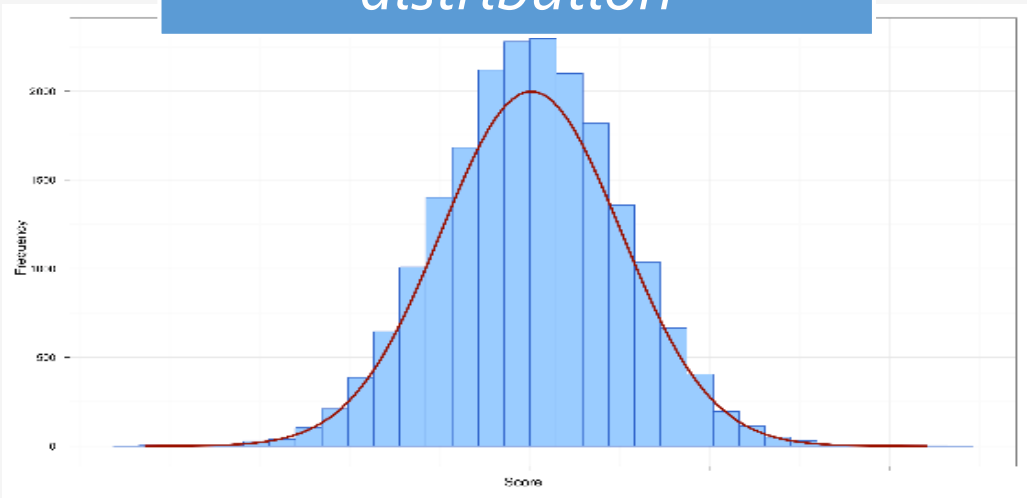
extremely dislike
moderately dislike
somewhat dislike
neither dislike nor like
somewhat like
moderately like
extremely like

Skewed distributions

- **skew:** a measure of the symmetry of a frequency distribution

- Non-skewed distributions are nearly symmetrical

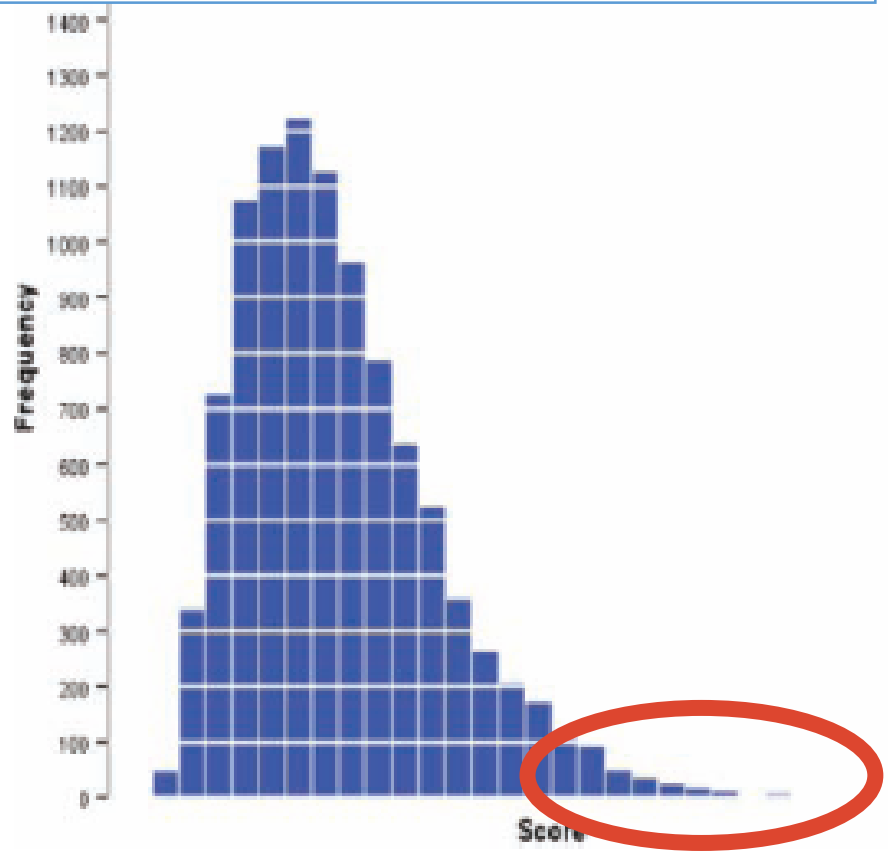
Here's a histogram with a non-skewed distribution



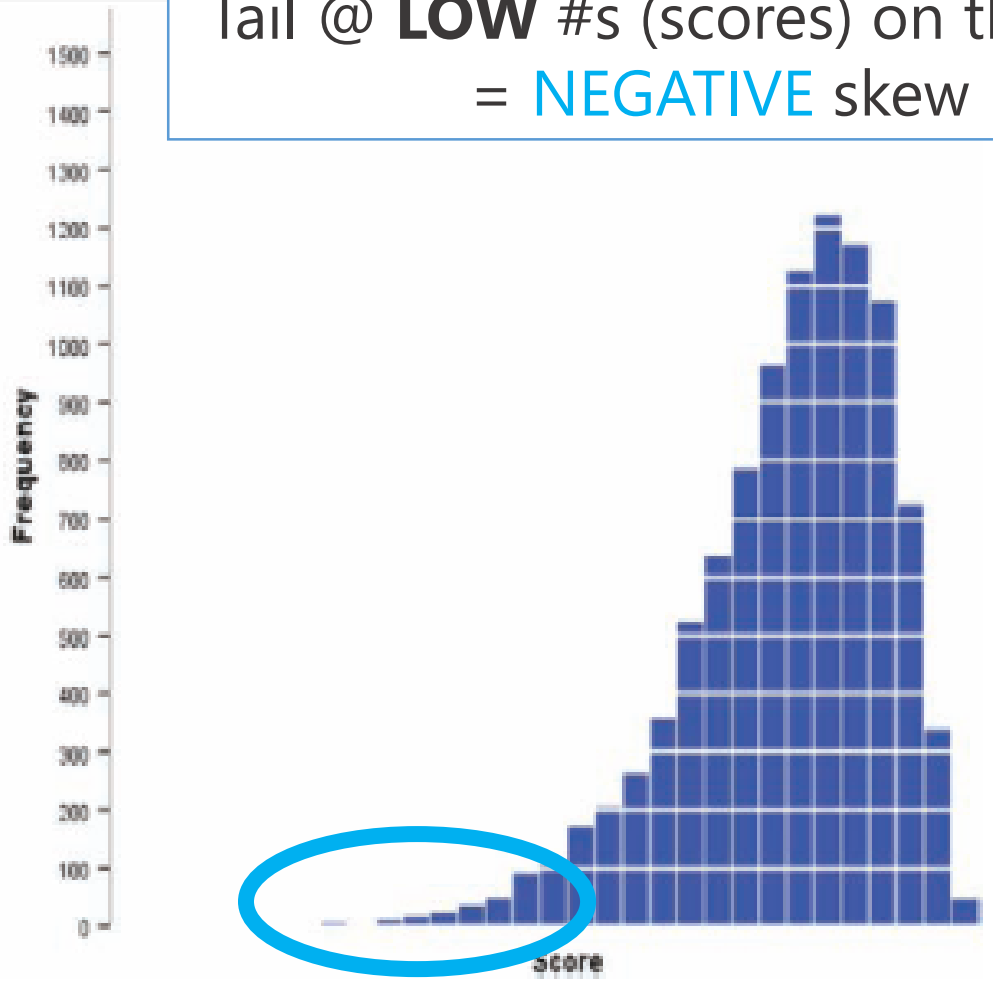
- Skewed distributions have a tail . . .

Two Types of Skew

Tail @ **HIGH** #s (scores) on the x-axis = POSITIVE skew

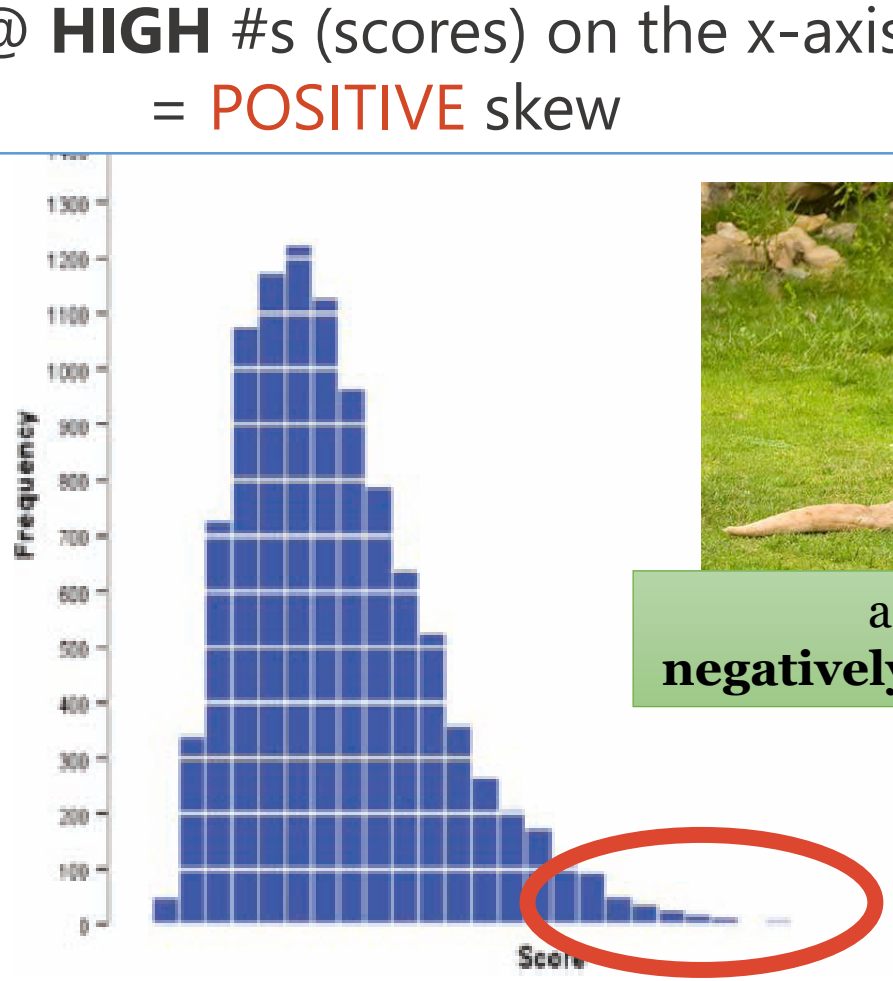


Tail @ **LOW** #s (scores) on the x-axis = NEGATIVE skew



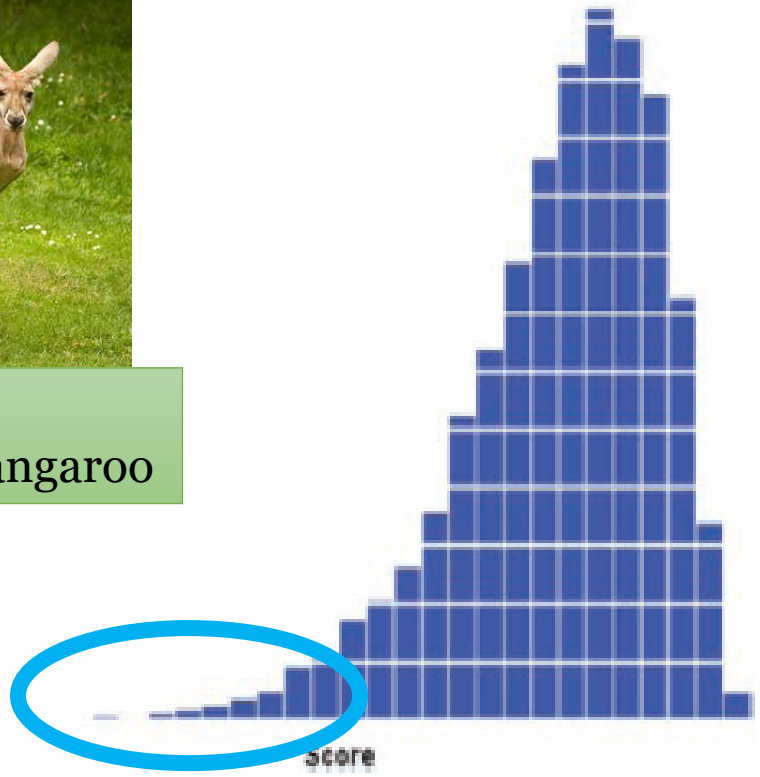
Two Types of Skew

Tail @ **HIGH** #s (scores) on the x-axis
= POSITIVE skew

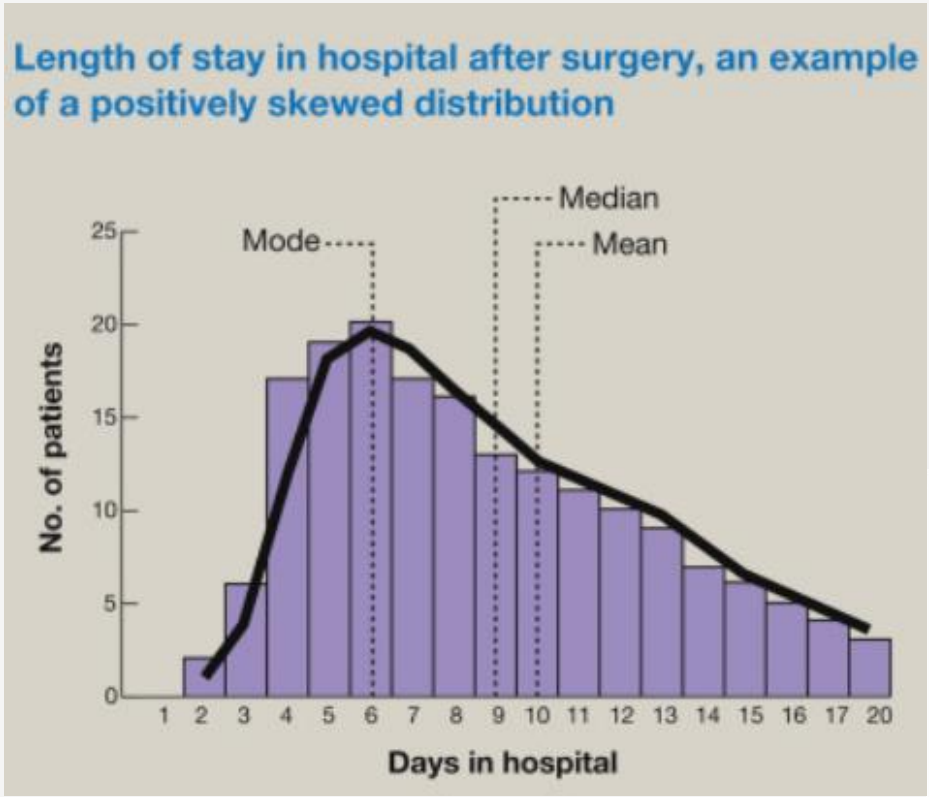


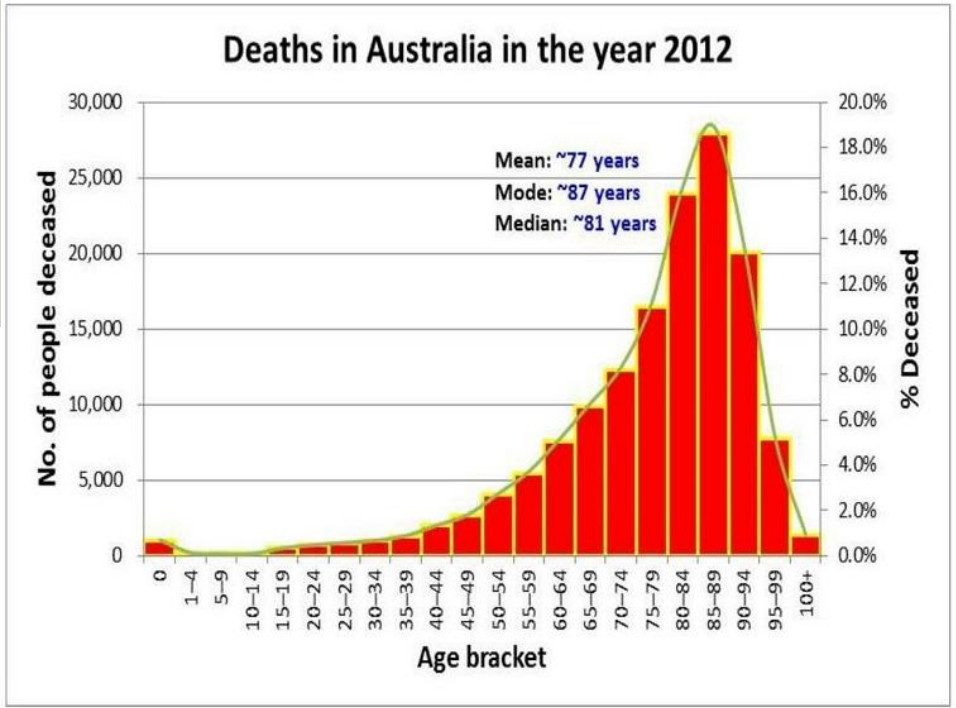
an adorable
negatively-skewed kangaroo

Tail @ **LOW** #s (scores) on x-axis
= NEGATIVE skew



More images of positive and negative skew



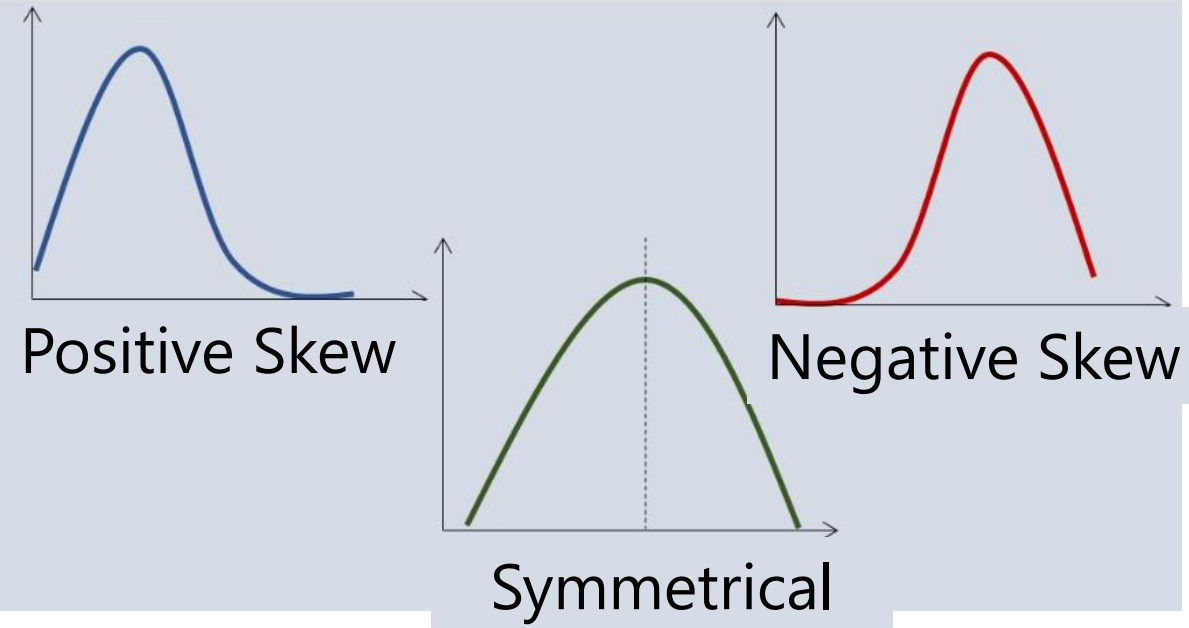


PRACTICE with Skewed vs. Normal distributions

Identify three quantitative variables that, when measured among BSC students, would be characterized each of the following distributions.

Try to think of variables you could measure quantitatively that relate to BSC students' behaviors, habits, skills, or opinions. Be creative!

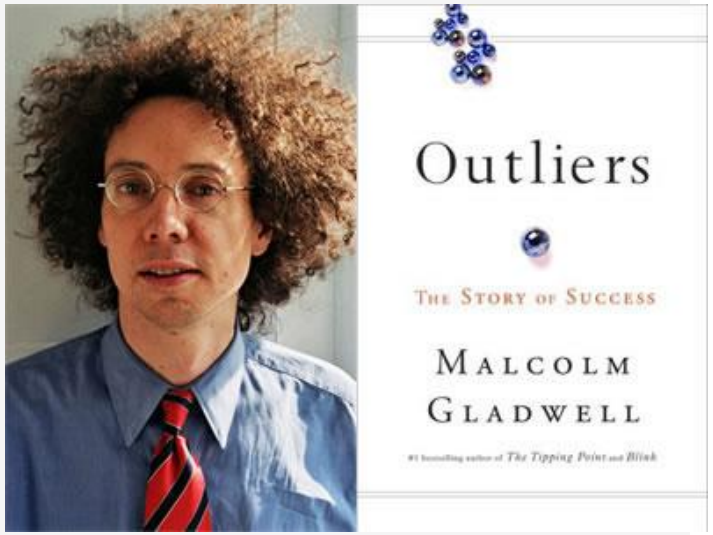
1. Symmetrical distribution
2. Positively-skewed distribution
3. Negatively-skewed distribution



What is an outlier?

- a score very different from the rest of the data; an extreme score

Also a really interesting book by Malcolm Gladwell

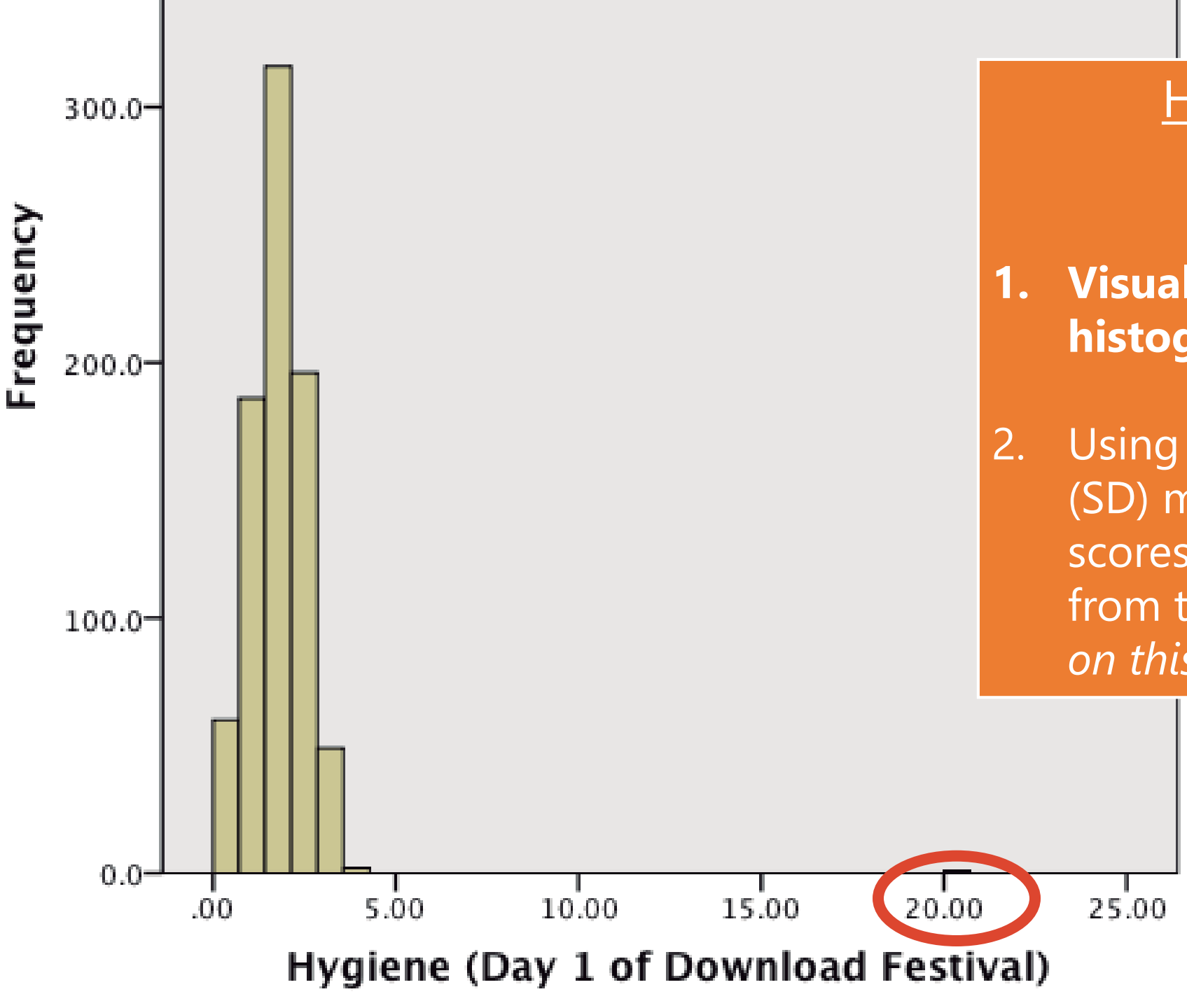


How do you spot an outlier? Weird example.

- A biologist was worried about the potential health effects of music festivals.
- Measured the hygiene of 810 concert-goers over the three days of the festival.
- Hygiene was measured on a six-point scale from . . .

    0 = you smell like a corpse rotting up a skunk's arse *to*

    5 = you smell of sweet roses on a fresh spring day



How do you spot an outlier?

1. **Visually (e.g., by looking at a histogram of your data)**
2. Using the "standard deviation (SD) method," and labeling scores that are a certain # of SDs from the mean as outliers (*more on this later in semester!*)

Why might you have outliers in your data?

- Outlier: a score very different from the rest of the data; an extreme score

- Where might outliers come from?
  - Researcher's data entry mistake
  - Participant mistake or misunderstanding
  - Participant intentionally misleads researcher
  - It's a real, meaningful score that's just very different

Outline for Ch. 3

1. Frequency distributions
2. **Measures of central tendency**
3. Measures of spread
4. Combining central tendency & spread

Mode
Median
Mean

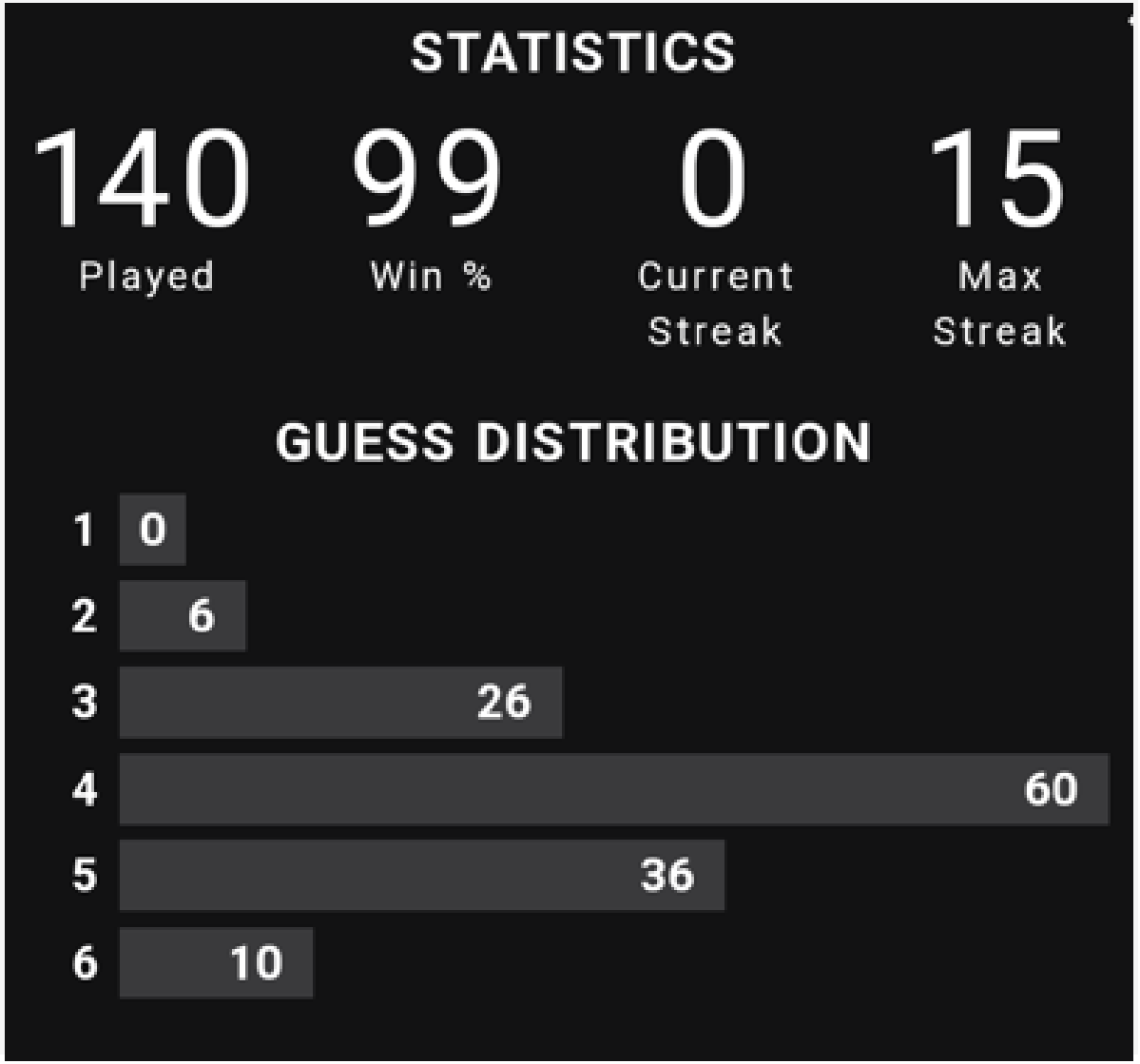
Histogram of Dr. V's Wordle Scores

Wordle scores can vary from 1 to 6.

Which score do I get **most** days?

4!

"4" is the **mode** of my Wordle scores.



Which score (response) occurs most frequently for a variable in our data set?

By "score" or "response" I mean, *the value or answer provided by or chosen by, or calculated for*, the participant.

Mode (an example of a measure of central tendency):

- Defined as: the score/response that occurs most frequently in the data set
  - In some cases, mode is defined as: *the answer that most people gave*

Here are the possible "scores" (aka "responses" aka answers)

Here are the # of people who gave each response

H ike or dislike cats?

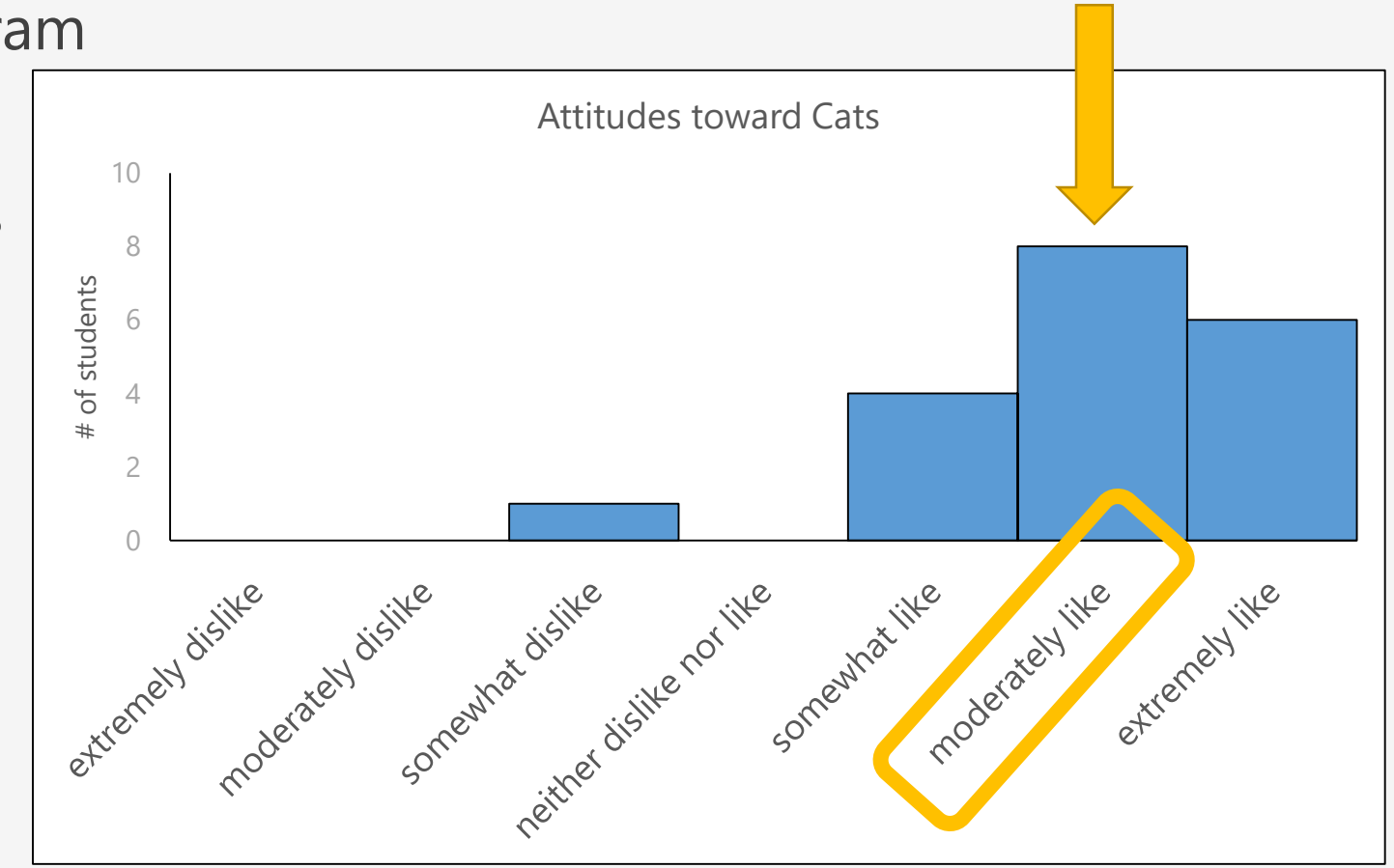
		Frequen	Percent	Valid Percent	Cumulative Percent
Valid	Somewhat dislike them	1	5.3	5.3	5.3
	Somewhat like them	4	21.1	21.1	26.3
	Moderately like them	8	42.1	42.1	68.4
	Extremely like them	6	31.6	31.6	100.0
	Total	19	100.0	100.0	

the mode!

Which score (response) occurs most frequently in our data set?

Mode (a measure of central tendency)

- Tallest/longest bar in histogram
  - **"Moderately like" is the mode.**
  - What if 2 bars are equally tall?



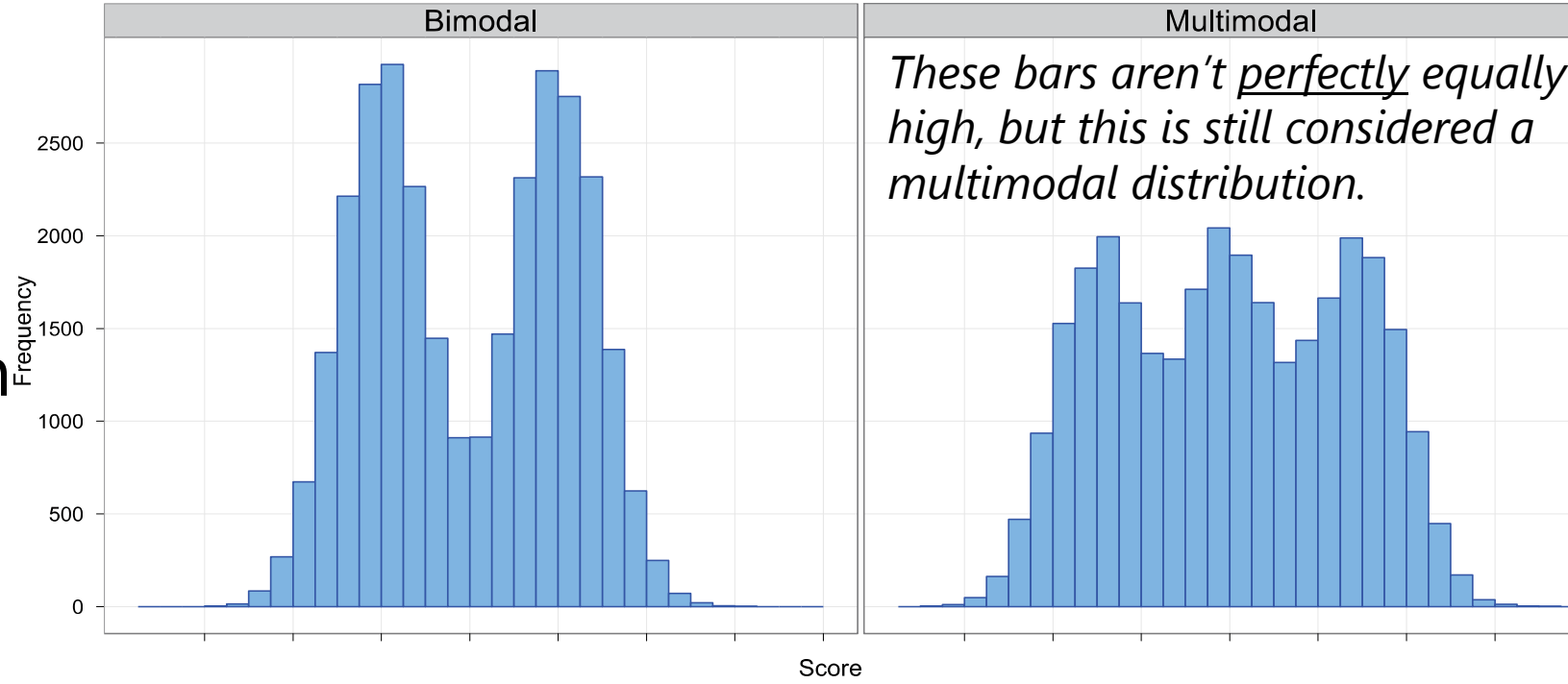
Bimodal and Multimodal Distributions

a BIMODAL distribution

has 2 modes

a MULTIMODAL distribution

has more than 2 modes



1 **How many Harry Potter movies have you seen?**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	2	11.1	11.8	11.8
	2.00	1	5.6	5.9	17.6
	3.00	1	5.6	5.9	23.5
	4.00	1	5.6	5.9	29.4
	6.00	2	11.1	11.8	41.2
	8.00	10	55.6	58.8	100.0
	Total	17	94.4	100.0	
Missing	System	1	5.6		
Total		18	100.0		

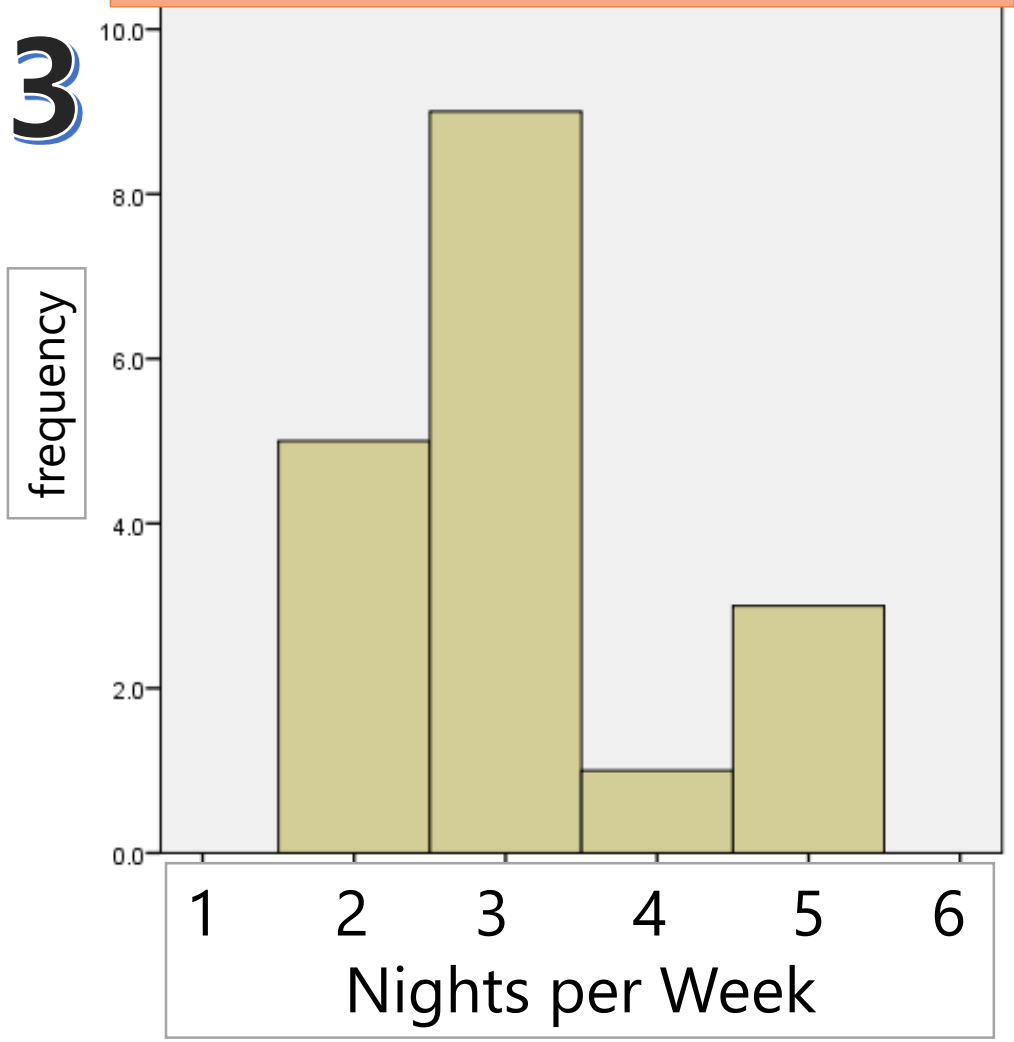
2

# of X (formerly known as Twitter) followers each of five people have:

Frank: 100 followers
Mary: 572 followers
Abby: 78 followers
Kevin: 154 followers
Jennifer: 238 followers

3

On average, how many nights per week does the typical college student drink?



What is the mode for each of these 3 examples?
(Make sure to include the *units* in your answer.)

1 **How many Harry Potter movies have you seen?**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	2	11.1	11.8	11.8
	2.00	1	5.6	5.9	17.6
	3.00	1	5.6	5.9	23.5
	4.00	1	5.6	5.9	29.4
	6.00	2	11.1	11.8	41.2
	8.00	10	55.6	58.8	100.0
	Total	17	94.4		
Missing	System	1	5.6		
Total		18	100.0		

Mode =
8 Harry Potter movies

2

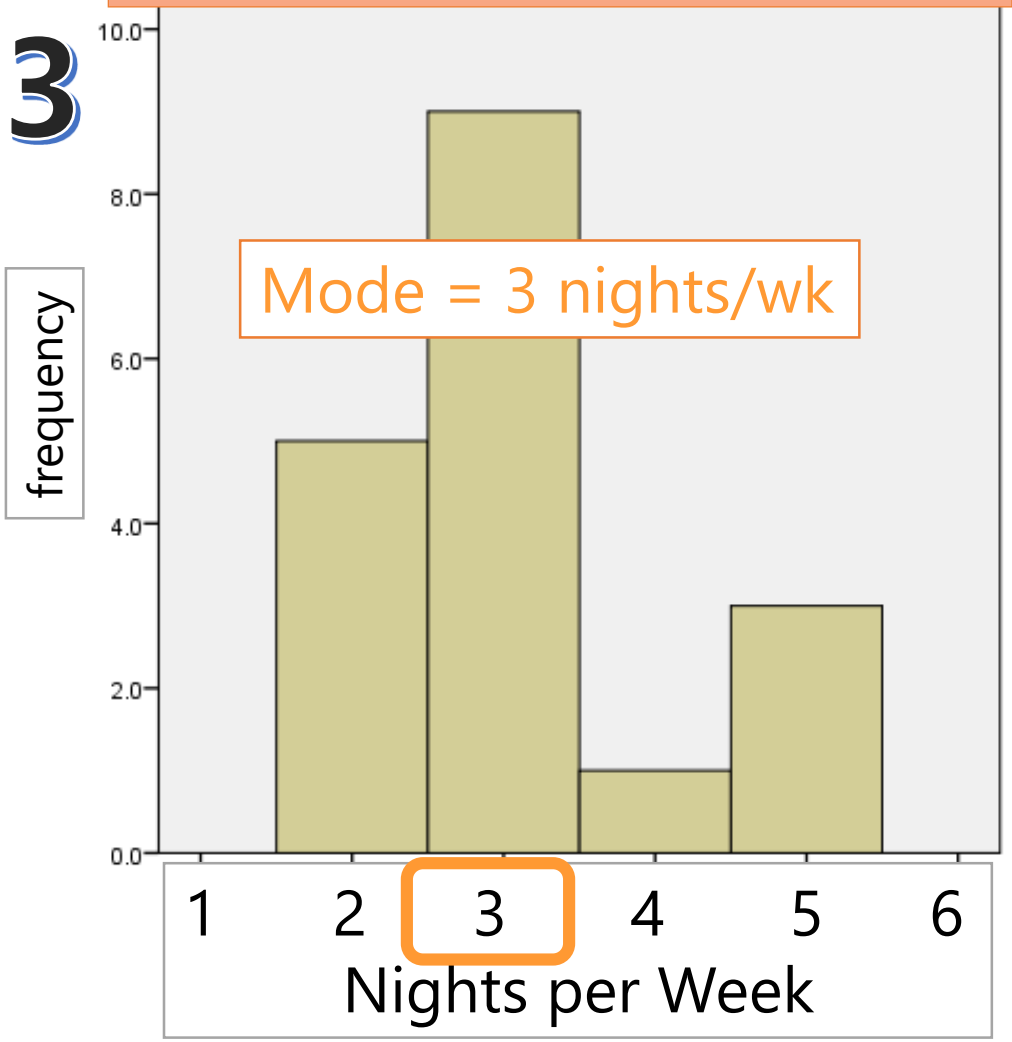
# of X followers each of five people have:

Frank: 100 followers
Mary: 572 followers
Abby: 78 followers
Kevin: 154 followers
Jennifer: 238 followers

There is no mode.

3

On average, how many nights per week does the typical college student drink?



What is the mode for each of these 3 examples?

The Median (another measure of central tendency)

- Defined as: the midpoint of the distribution of scores
  - This means that the same # of scores is *above* the median as is *below* it
- Also defined as: the value associated w/the middle data point, when the data points are put in order

***EXAMPLE:* # of X (Twitter) followers each of five people have**

Frank: 100 followers
Abby: 78 followers
Kevin: 154 followers
Mary: 572 followers
Jennifer: 238 followers

PUT IN ORDER

Abby: 78 followers
Frank: 100 followers
Kevin: 154 followers
Jennifer: 238 followers
Mary: 572 followers

The median number of X followers in our sample of 5 participants is: ***154 followers***.

The Median (another measure of central tendency)

- Defined as: the value associated w/the middle data point, when the data points are ordered
  - by middle data point, I mean (N + 1) / 2
  - *This formula is not a formula for the median itself; it is a way to figure out which data point is the middle one. The median is the* ***value associated with*** *that middle data point.*

Remember that N = the number of data points in the data set (aka, *N* = the sample size)

The Median (another measure of central tendency)

- Defined as: the value associated w/the middle data point, when the data points are ordered
  - by middle, I mean (N + 1) / 2

EXAMPLE: # of FB friends each of 11 users has

N = 11 data points

middle means the (11+1)/2 = **6th data pt**

Then, put data pts in order from low to high. Count until you get to the 6th one.

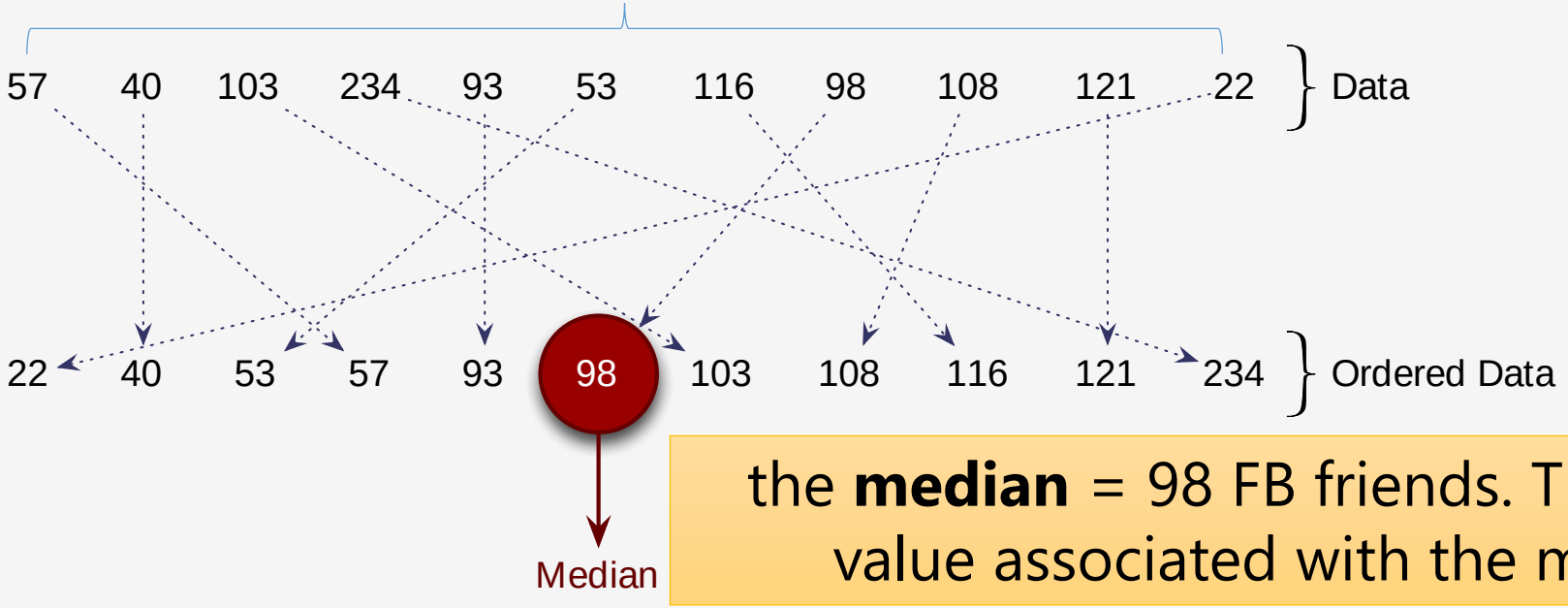
The Median (another measure of central tendency)

- Defined as: the value associated w/the middle data point, when the data points are ordered
  - by middle, I mean (N + 1) / 2

N = 11 data points

middle means the (11+1)/2 = **6th data pt**

EXAMPLE: # of FB friends each of 11 users has



Then, put data pts in order from low to high. Count until you get to the 6th one.

the **median** = 98 FB friends. This is because "98" is the value associated with the middle/6th data point.

The Median (another measure of central tendency)

- by middle, I mean (N + 1) / 2, where N = the # of scores

- If your *middle* is a decimal number (e.g., 3.5), take an average (e.g., average the 3rd and 4th scores (i.e., data points)

N = 8 data pts (i.e., 8 scores)

middle = (8+1)/2 = 4.5th data pt

Example – try it on your own.
Find the median of these scores:
7, 9, 9, 1, 6, 8, 2, 4

Put data points in order. Count til you get to the 4th & 5th ones, and average those two data pts.

1 2 4 6 7 8 9 9

Median = (6+7)/2 = **6.5**

The Mean (another measure of central tendency)

- The **sum** of scores divided by the **number** of scores (N).

$$= \frac{(x_1 + x_2 + x_3 + x_4 + x_5 + x_6)}{6}$$

We say this "x sub 5" → x_5

We say this "x sub 6" → x_6

and it just means the score of the 6th participant.

("sub" is for "subscript")

N = 6

scores	scores
x_1	1
x_2	5
x_3	20
x_4	23
x_5	79
x_6	83

The Mean (another measure of central tendency)

- The **sum** of scores divided by the **number** of scores (N).

$$\frac{(x_1 + x_2 + x_3 + x_4 + x_5 + x_6)}{6}$$

$\bar{X}$ = 211 / 6 = 35.17

N = 6

scores	scores
x_1	1
x_2	5
x_3	20
x_4	23
x_5	79
x_6	83

Do you remember what symbol we use to represent the *mean in a sample?*

The Mean (another measure of central tendency)

- The **sum** of scores divided by the **number** of scores (N).

Sum the scores (i.e., the x's), from the 1st score to the Nth (last) score

Summation sign (sigma)

Represents the mean in your sample ("x-bar").

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{(x_1 + x_2 + x_3 + x_4 + x_5 + x_6)}{6}$$

Represents the total # of scores in your sample (i.e., sample size)

N = 6

scores	scores
x_1	1
x_2	5
x_3	20
x_4	23
x_5	79
x_6	83

72, 90, 90, 91, 69 412/5 = 82.4

Practice: Calculate the mean.