CH. 3 – Measures of Central Tendency and Spread

PY 221 Statistics for Research

Dr. Valenti

Reminders

- Complete Quiz #4 by start of class tomorrow. Study all the slides thus far.

- Lab #1 due Tuesday by 12:30 pm – turn in to Moodle.
  - Remember that the key for practice lab #1 is on Moodle for you to carefully check your work from yesterday.

- Dr. V has office hours today, tomorrow morning, & Monday afternoon https://gvalenti.youcanbook.me/
  - No office hours Tuesdays and Fridays, so plan ahead if you'd like to meet!

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 in TABLE form

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

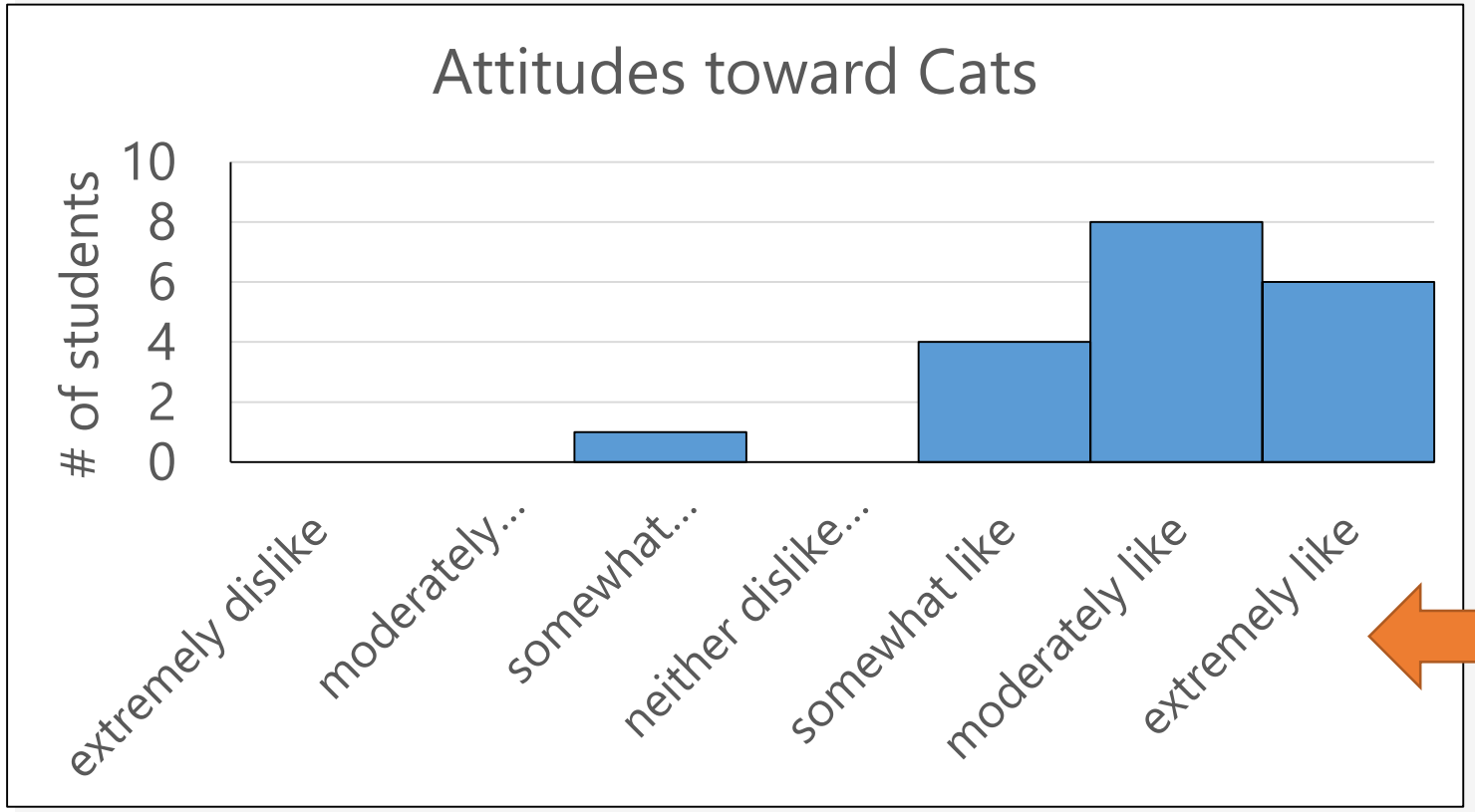
Ignore me!

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

How much do you like

FREQUENCY **TABLE**

		Frequency
Valid	Somewhat dislike them	1
	Somewhat like them	4
	Moderately like them	8
	Extremely like them	6
	Total	19



How much do you like or dislike cats?

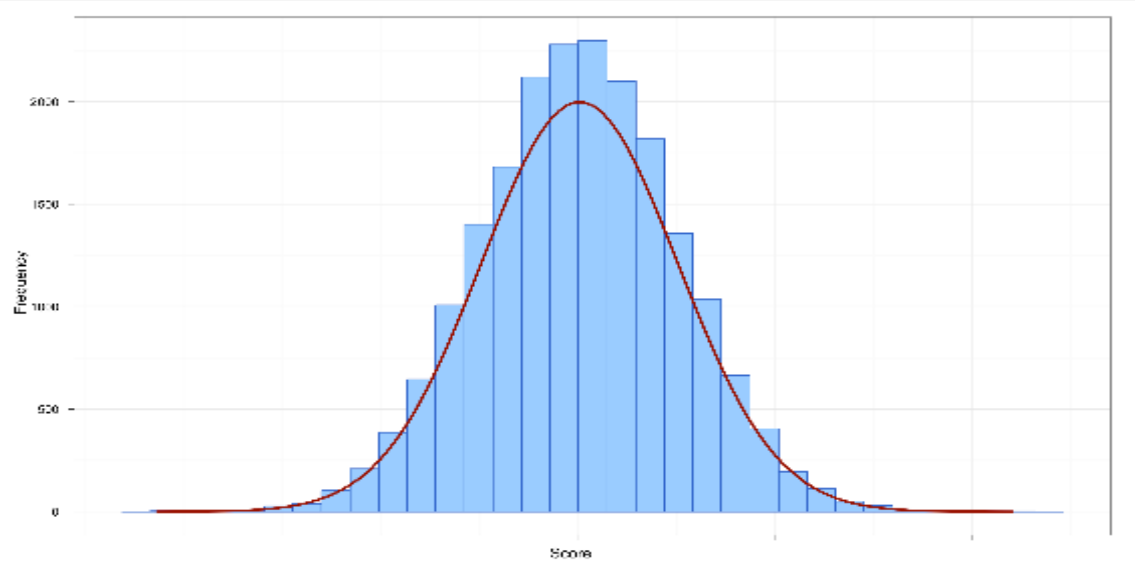
extremely dislike
moderately dislike
somewhat dislike
neither dislike nor like
somewhat like
moderately like
extremely like

Skewed distributions

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

- Non-skewed distributions are nearly symmetrical

A non-skewed distribution



- Skewed distributions have a tail . . .

Two Types of Skew

Tail @ **HIGH** #s (scores) → POSITIVE skew

Tail @ **LOW** #s (scores) → NEGATIVE skew

an adorable
negatively-skewed kangaroo

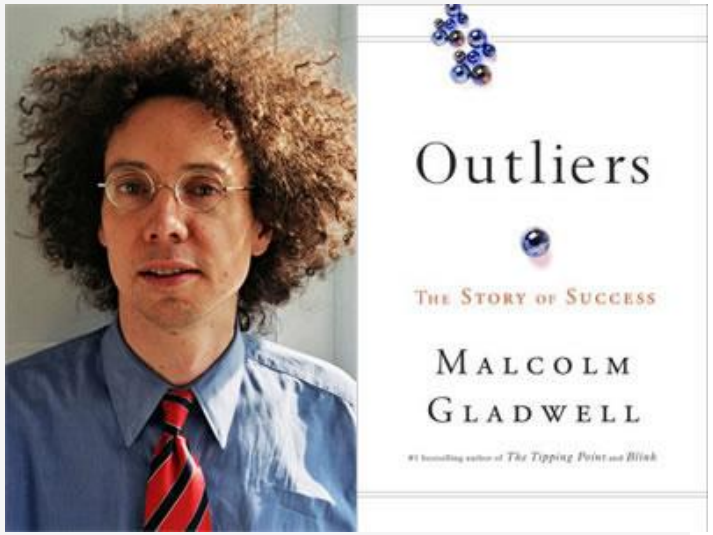
What is an outlier?

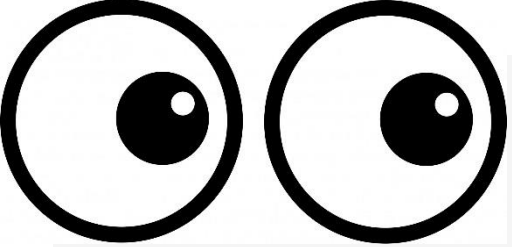
- a score very different from the rest of the data; an extreme score

Reasons for outliers

- Researcher's data entry mistake
- Participant mistake or misunderstanding
- Participant intentionally misleads researcher
- It's a real, meaningful score that's just very different

Also a really interesting book by Malcolm Gladwell



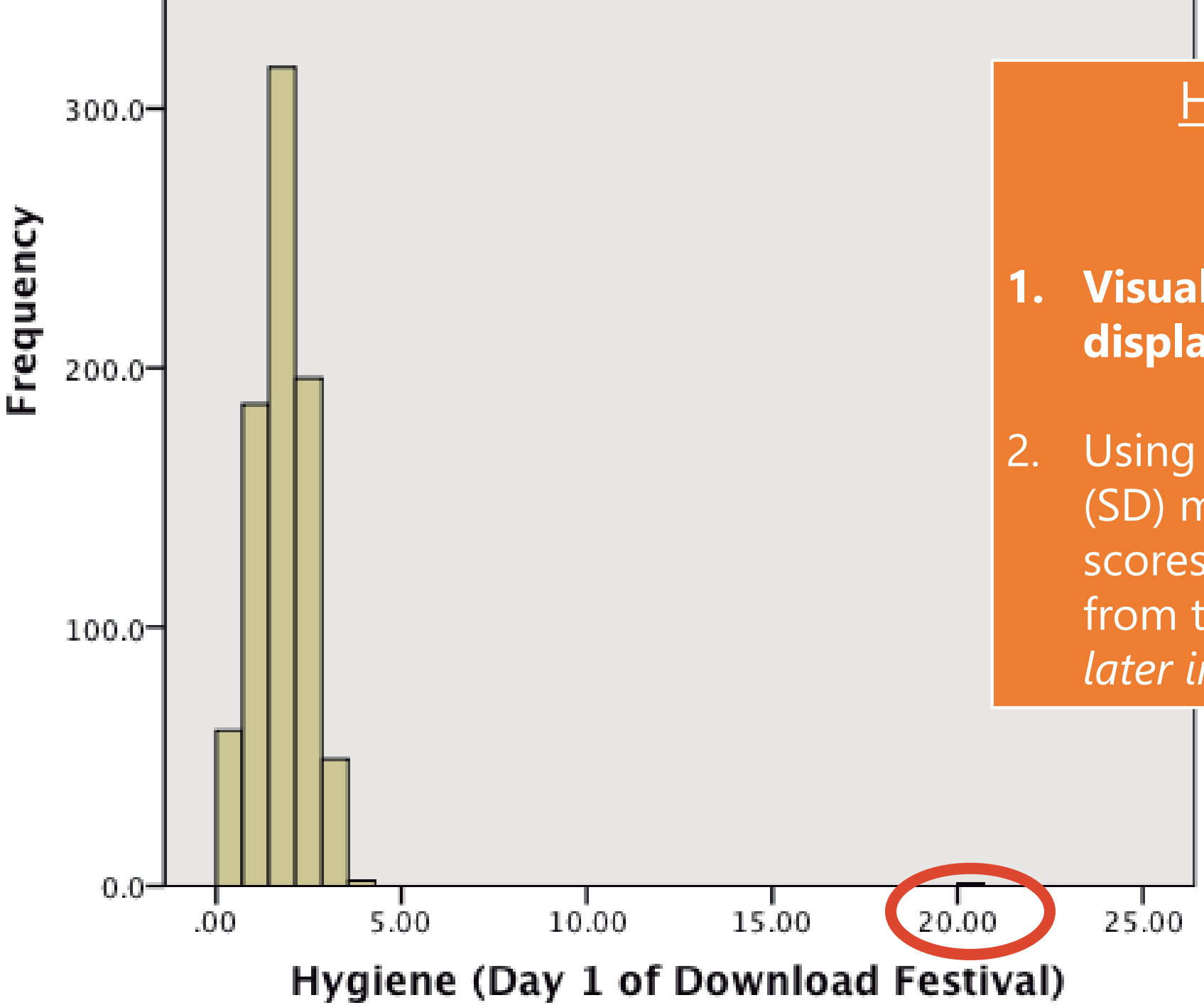


How do you spot an outlier? Weird ex.

- 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 (using some graphical display of your data)**
2. Using the "standard deviation (SD) method," and excluding scores that are a certain # of SDs from the mean (*more on this later in semester!*)

Outline for Ch. 3

1. Frequency distributions
2. **Measures of central tendency**
   - Mode
   - Median
   - Mean
3. Measures of spread
4. Combining central tendency & spread

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

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

Mode (a measure of central tendency):

- 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" or "responses"

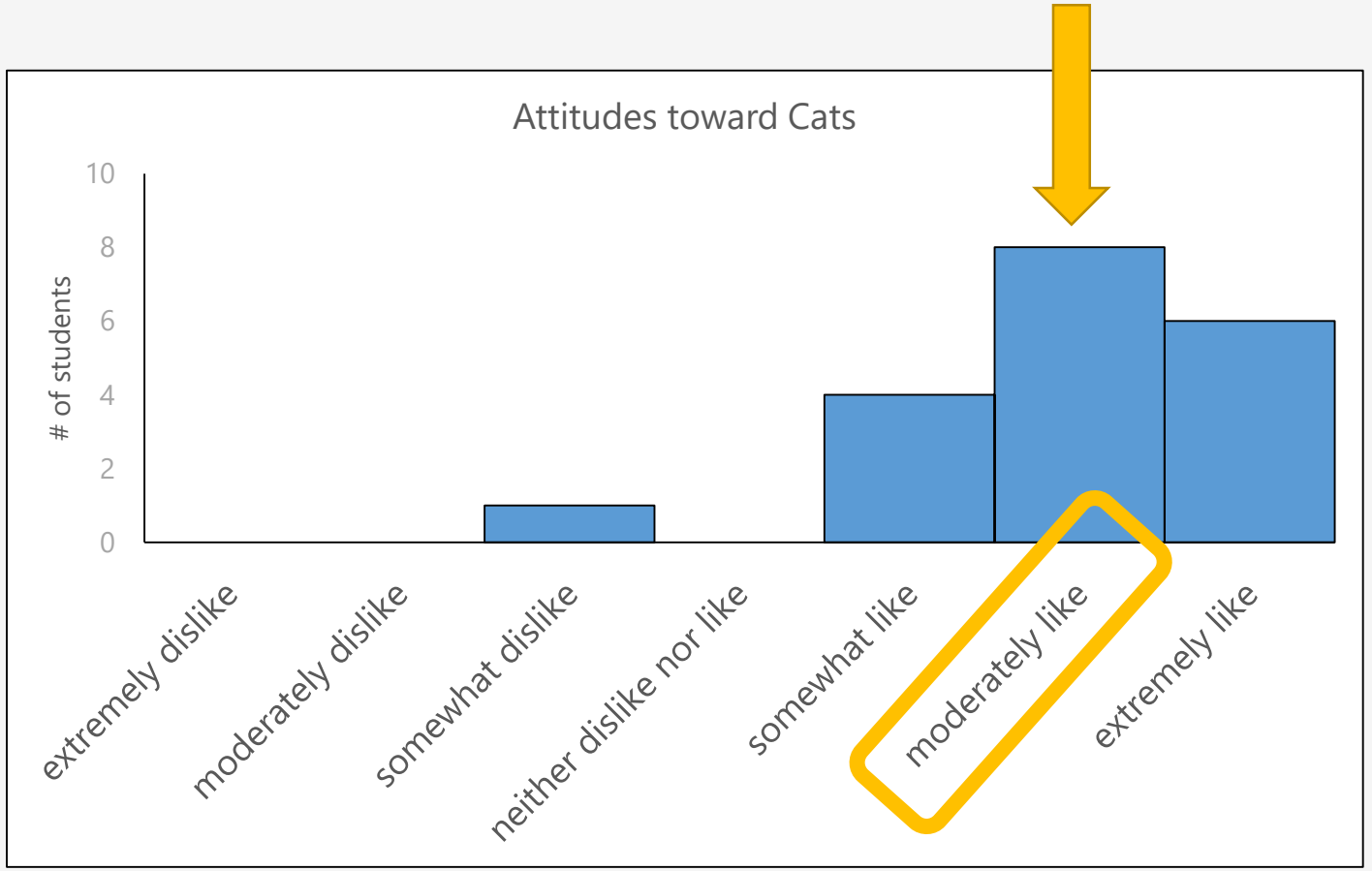
How much do you like or dislike cats?

		Frequency	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	

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

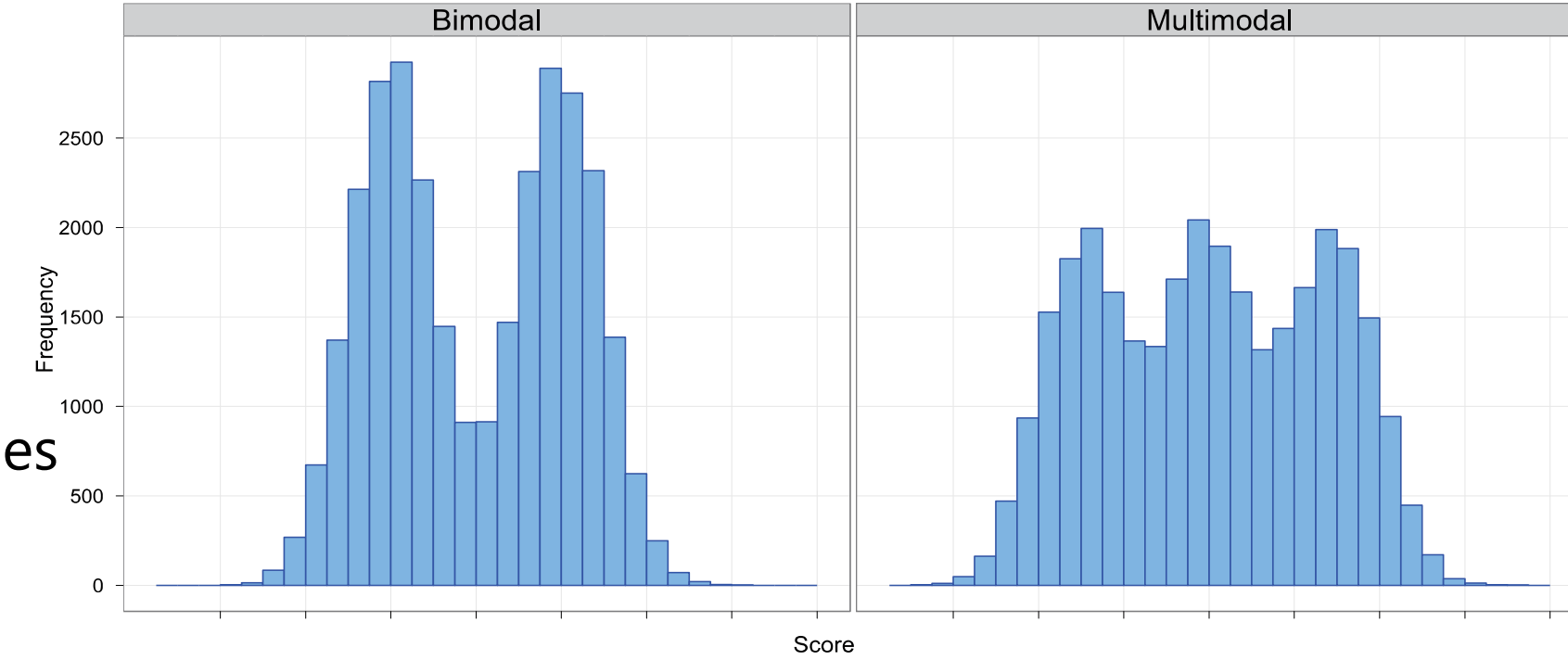
Mode (a measure of central tendency)

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



Bimodal and Multimodal Distributions

- Bimodal distribution
  - Having 2 modes
- Multimodal distribution
  - Having more than 2 modes

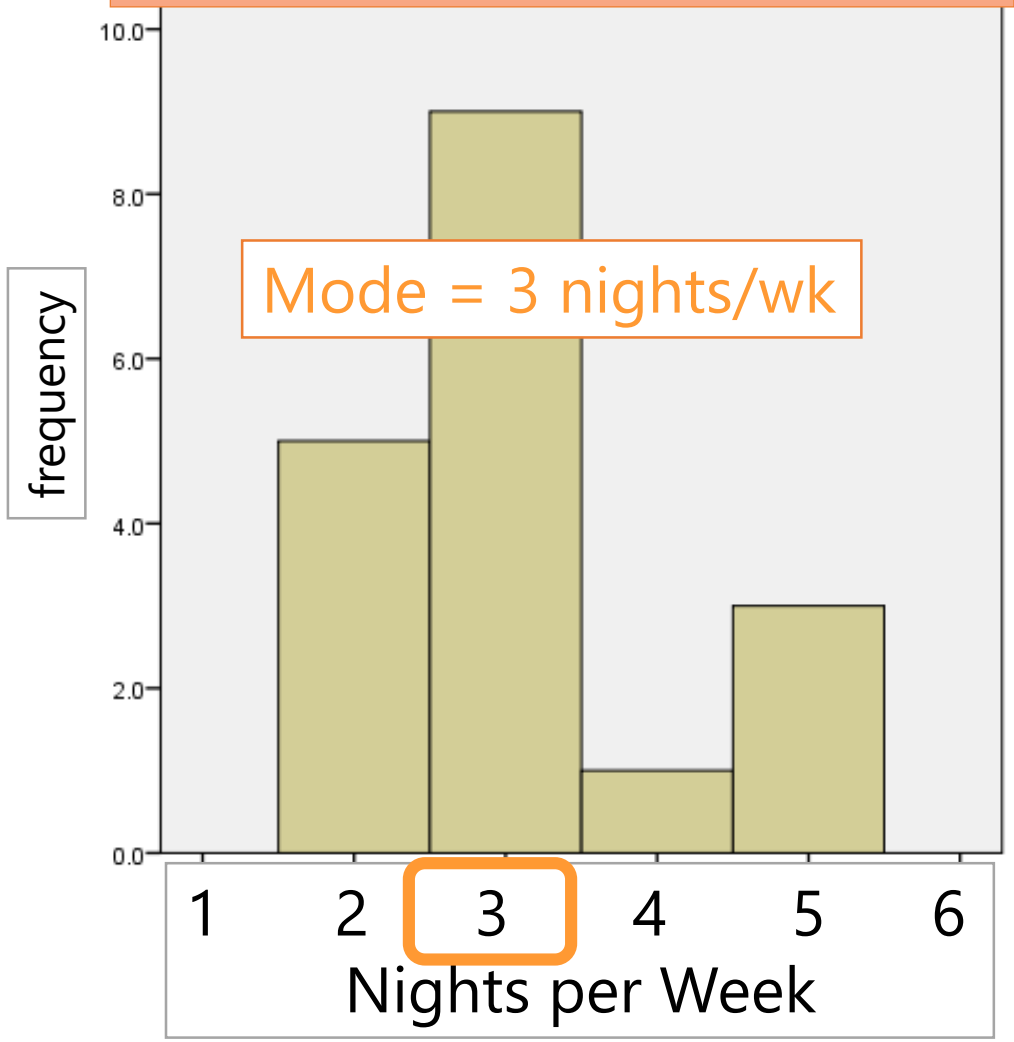


How many Harry Potter movies have you seen?

		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		

Mode = 8 HP movies

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



# of Twitter 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.

What is the mode for each of these 3 examples?

The Median (another measure of central tendency)

- the midpoint of the distribution of scores
  - The same # of scores is *above* the median as is *below* it

- the value associated w/the middle data point, when the data points are ordered

***EXAMPLE:* # of 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 is: 154 followers.

The Median (another measure of central tendency)

- the midpoint of the distribution of scores
  - The same # of scores is *above* the median as is *below* it
- the value associated w/the middle data point, when the data points are ordered
  - by middle data point, I mean (n + 1) / 2, where n = the number of data points in the data set (aka, where n = the sample size)
  - *note that 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.*

The Median (another measure of central tendency)

- the value associated w/the middle data point, when the data points are ordered
  - by middle, I mean (n + 1) / 2, where n = the number

EXAMPLE: # of FB friends each user has

n = 11 data points

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

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

Median = 98

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)

n = 8 data pts

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

Example – try it on your own.
Find the median of 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).

Summation sign (sigma)

Sum the scores, from the 1st score to the nth (last) score

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

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

$$= 211 / 6 = 35.17$$

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

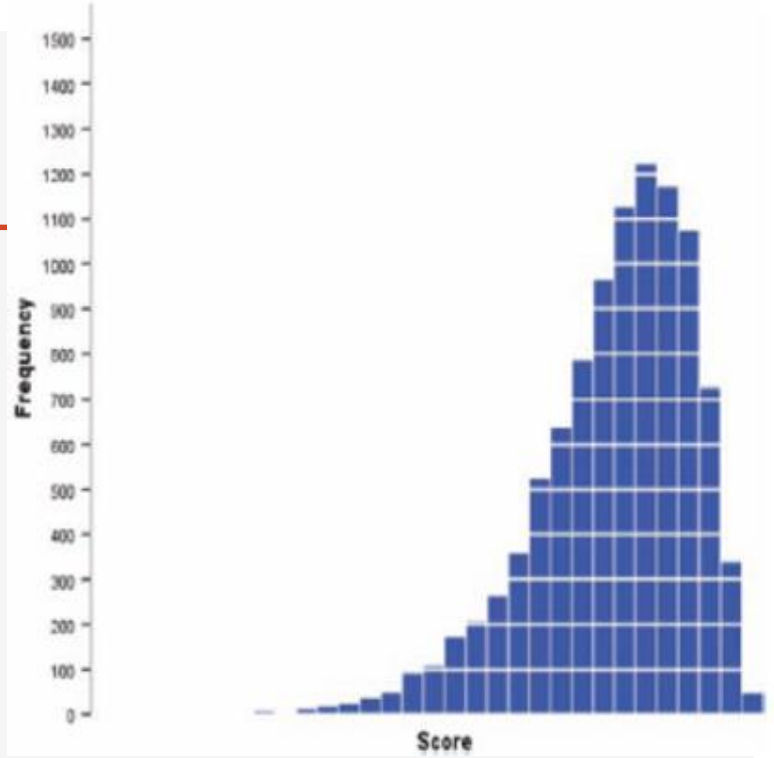
n = 6

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
Calculate the mean.

412/5 = 82.4

Mode, Median, & Mean Critical Thinking Qs



1. If the skew is negative (see diagram), which will be *higher*, the median or the mean?
2. Is the median or the mean more sensitive to extreme scores (outliers)?
3. Which is the only measure of central tendency (mode, median, or mean) that may not exist for a quantitative variable?
4. Which measure of central tendency is the only one that incorporates every single score into its calculation?
5. Which is the best measure of central tendency for *qualitative* data?

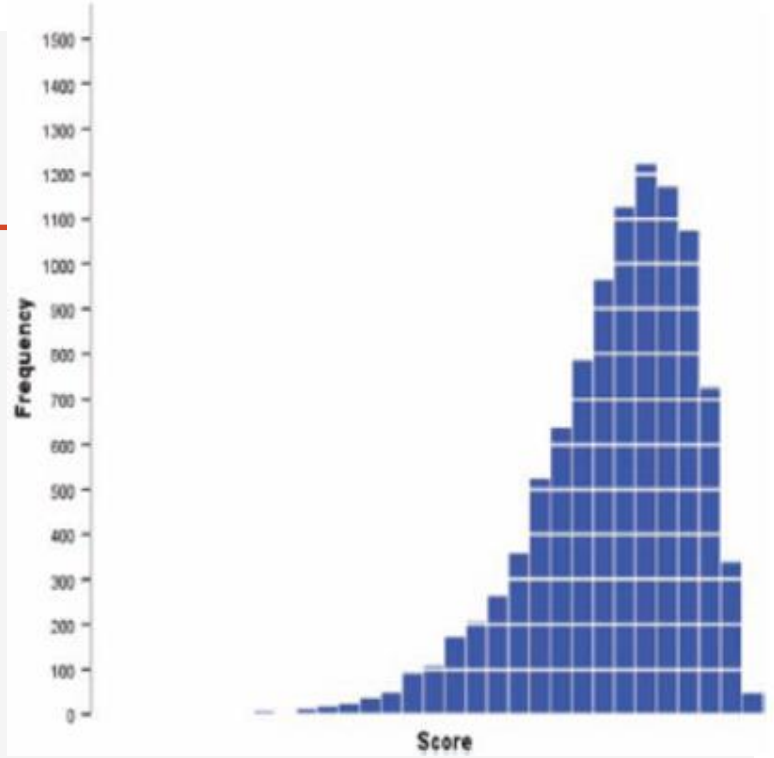
Mode, Median, and Mean Critical Thinking Qs & Answers

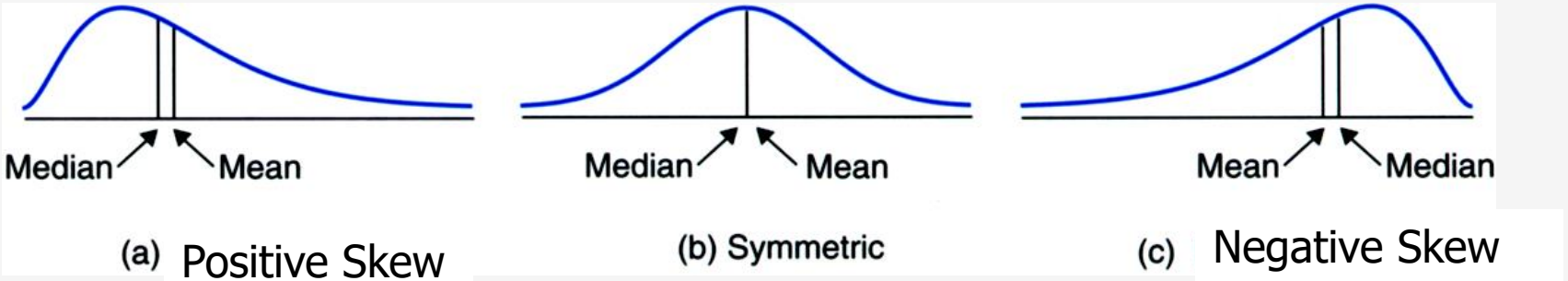
1. If the skew is negative (see diagram), which will be *higher*, the median or the mean? median

FACTS:

- ***In perfectly symmetrical distributions, the median and mean are identical***
- ***Means get pulled in direction of tail***

2. Is the median or the mean more sensitive to extreme scores (outliers)? mean





Mode, Median, and Mean Trivia Qs

1. If the skew is negative (see diagram), which will be *higher*, the median or the mean?
2. Is the median or the mean more sensitive to extreme scores (outliers)?
3. Which is the only measure of central tendency (mode, median, or mean) that may not exist for a quantitative variable?

mode

Each score may occur an equal number of times → there is no mode.

4. Which measure is the only one that incorporates every single score into its calculation?

mean

5. Which is the best measure of central tendency for *qualitative* data? mode

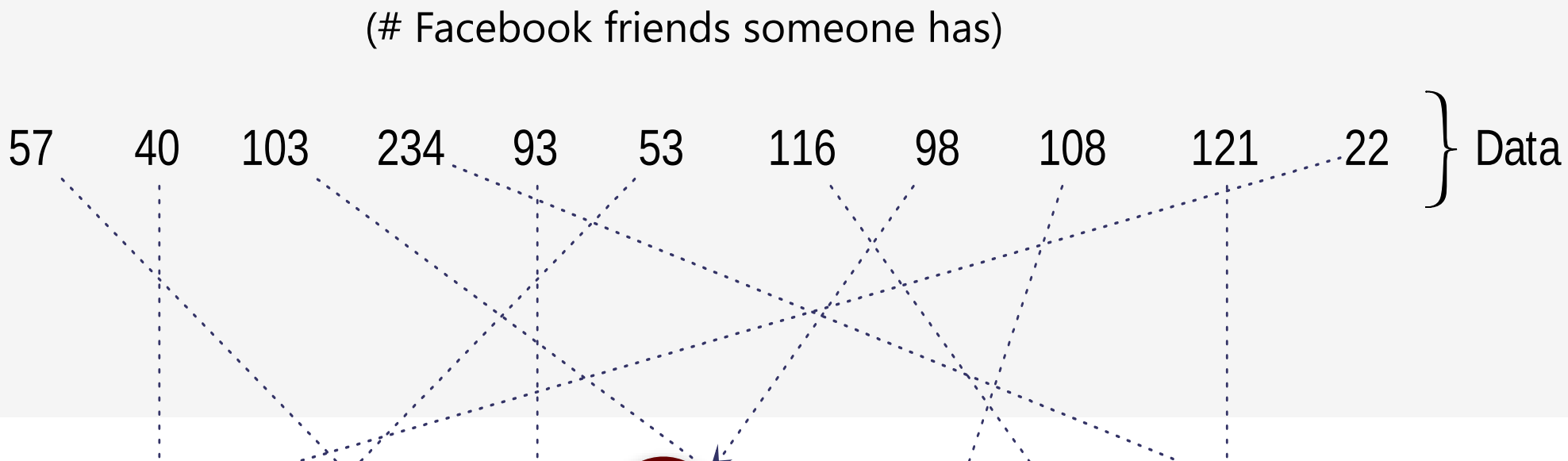
Ex: if measure favorite color, can't calculate a median because you cannot put the responses in order, and can't calculate mean because responses are not values

Outline for Ch. 3

1. Review of frequency distributions
2. Measures of central tendency
3. **Measures of spread (aka, measures of dispersion, aka measures of variability)**
4. Combining central tendency & spread

Range (a measure of spread (*aka dispersion, variability*))

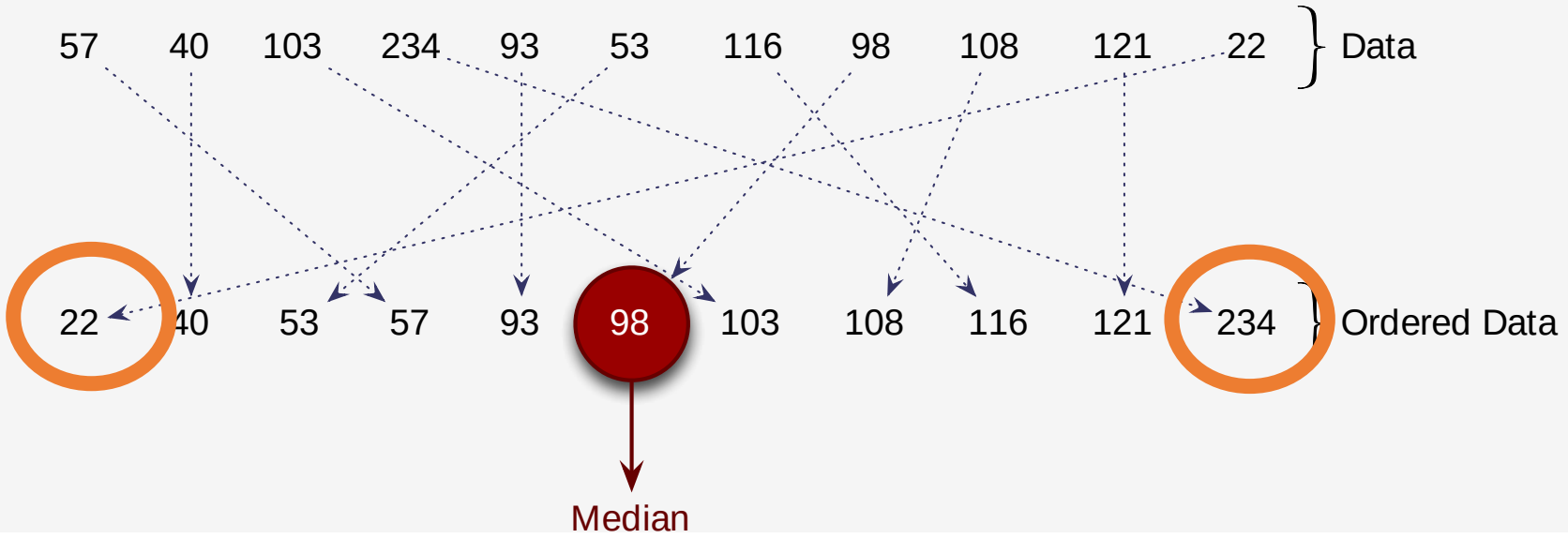
- The distance covered by the scores in a distribution
- To calculate, subtract the smallest score from the largest
- *Practice*: What is the **range** for this data set?



Range (a measure of spread (*aka dispersion, variability*))

- The distance covered by the scores in a distribution
- To calculate, subtract the smallest score from the largest
- ***Weakness of range*: only uses 2 scores**
- Practice: What is the **range** for this data set

Range = 234 - 22 = **212**
Range = 212 Facebook friends



Deviation (a measure of spread)

- how different a given score is from the center of a distribution (i.e., from *the mean*)
  - You can calculate a deviation for each individual score in the data set.

$$\text{deviation} = x_i - \bar{x}$$

(# Facebook friends someone has)

57 40 103 234 93 53 116 98 108 121 22 } Data

Score (x_i)
22
40
53
57
Etc.
27

Deviation (a measure of spread)

$$\text{deviation} = x_i - \bar{x}$$

- how different a given score is from the center of a distribution (i.e., from *the mean*)
  - You can calculate a deviation for each individual score in the data set

- **In practice, researchers rarely use individual deviations.**

Score (x_i)	Mean (x-bar)	deviation $(x_i - \bar{x})$
22	95	-73
40	95	-55
53	95	-42
57	95	-38
Etc.	Etc.	Etc.

(# Facebook friends someone has)

57 40 103 234 93 53 116 98 108 121 22 } Data