

Organizing Data

Statistics is the science and art of *collecting, analyzing, and drawing conclusions* from data.

Student Name	Favorite Movie	Genre	Streaming Platform	Rating (out of 10)
Alice	Inception	Sci-Fi	Netflix	9
Bob	The Dark Knight	Action	Hulu	10
Charlie	The Shawshank Redemption	Drama	Amazon Prime	8
Diana	Parasite	Thriller	Netflix	9
Edward	Toy Story	Animation	Disney+	7
Fiona	Titanic	Romance	Hulu	8
George	Avengers: Endgame	Superhero	Disney+	9
Hannah	Get Out	Horror	HBO Max	8
Isaac	Forrest Gump	Drama	Amazon Prime	9
Julia	The Matrix	Sci-Fi	Netflix	10

Statistics help us make sense of data. This year, we will explore data related to sports, medicine, weather, tech and more!

Investigative Questions

A statistical study starts with an investigative question. A **valid statistical investigative question** is one based on data that *vary*.

Valid ✓

"Can Joy Milne really smell Parkinson's?" Some shirts were identified correctly, others not — the data vary.

Not valid ✗

"What is the world record for the 100-meter dash?" A single data value answers it — nothing varies.



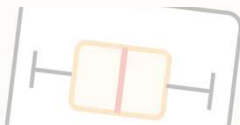
Population, Census, and Sample

The **population** in a statistical study is the entire group of individuals we want information about.

A **census** collects data from every individual in the population.

Why not use a census to collect data about a population?

- Very expensive
- Perhaps impossible
- If using destructive methods, you'd destroy the whole population
 - Lifetime of flashlight batteries
 - Safety ratings of cars



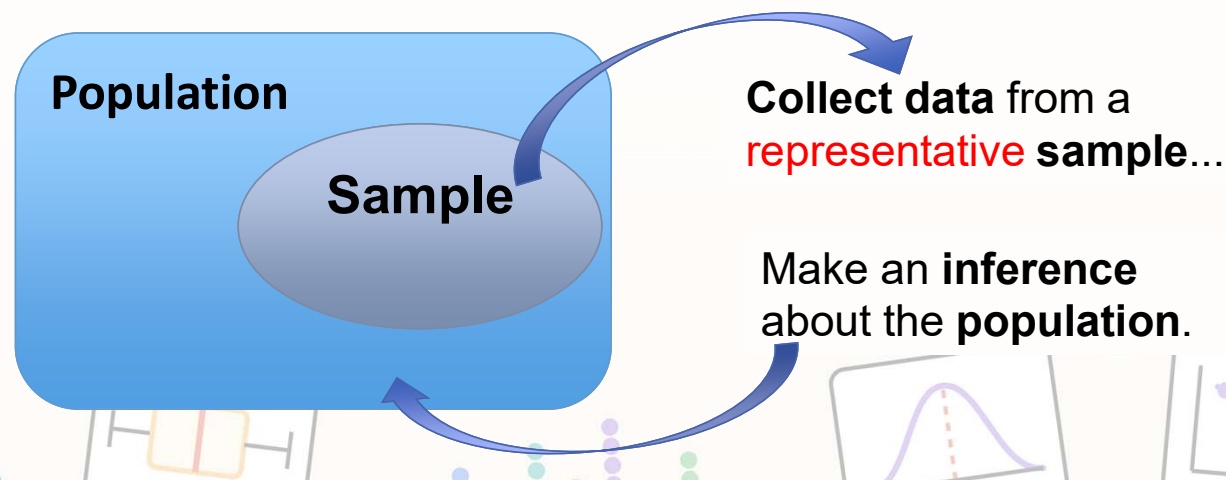


Population, Census, and Sample

The **population** in a statistical study is the entire group of individuals we want information about.

A **census** collects data from every individual in the population.

A **sample** is a subset of individuals in the population from which we actually collect data.



Understanding Check

Identify the population and the sample:

- (a) To estimate this week's average ER wait time, hospital staff time 50 patients from arrival to treatment. Identify the population and sample.
- (b) A company selects 20 lifeguard shifts from the next 2 months' schedule for unannounced skills tests. Identify the population and sample.

Statistic vs. Parameter

Imagine we are interested in the mean length of fish in Lake Malawi (Tanzania).



What would it be reasonable to say to estimate it?

Statistic vs. Parameter

Any number that describes or summarizes a set of quantitative data from a sample is called a **Statistic**

Any number that describes or summarizes a set of quantitative data from a population is called a **Parameter**

Examples

Minimum
Maximum
Median
Mean
Range
Standard Deviation
A % above some value
The 3rd value

They are the same thing! (*Sort of*)



Parameters and Statistics

A **parameter** is a number that describes some characteristic of a population. A **statistic** is a number that describes some characteristic of a sample.

- We use the letter **s** for a sample statistic to denote the sample mean, $\bar{x}$; N = population

AP® Exam Tip

Many students lose points on the statistics exam when defining parameters because their description refers to a sample instead of the population or because the description isn't clear about which group the parameter is describing. When defining a parameter, we suggest including the word *all* or the word *true* in your description to make it clear that you aren't referring to a sample statistic.

Remember **s** and **p**
statistics come from **s**amples and
parameters come from **p**opulations

Organizing Data

An **observational unit** is an object described in a set of data. Observational units can be people, animals, or things.

Variable - an attribute that can take different values for different observational units

Categorical Variable

assigns labels that place each observational unit into a particular group, called a category.

Quantitative Variable

takes number values that are quantities — counts or measurements.



Why is a Zip Code not considered a Quantitative Variable?

Non-Numeric Nature: Zip codes are composed of numerical digits, but they are not used to measure a quantity. Each digit or group of digits represents a specific geographic area or region rather than a numerical value.

No Meaningful Arithmetic Operations: You cannot perform meaningful arithmetic operations on zip codes. Adding or subtracting zip codes doesn't convey any meaningful information.

Categories with No Inherent Order: Zip codes do not have an inherent order or ranking. The numerical values themselves do not imply any quantitative relationship or order.



Categorical vs. Quantitative

Please separate the following list of variables into a **categorical** bucket and a **quantitative** bucket.

Annual income of households in dollars

Average daily temperature in degrees Celsius

Blood type

Car brand owned by students

Distance traveled by students to reach Soffer in miles

Grade level of students (Freshman, Sophomore, Junior, Senior)

Height of students in centimeters

Number of pages in a textbook

Number of siblings each student has

Preferred genre of music

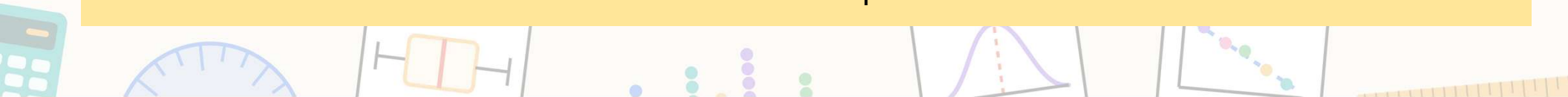
Shirt sizes

Time taken to complete a math quiz in minutes

Type of transportation used by students to get to school

Weight of backpacks carried by students

Zip Codes of Soffer Staff Households



Let's Practice:

Airport name	Elevation (ft)	Municipality	Country	Radio frequency (mHz)	Runway length (ft)	Runway surface
Los Angeles Int'l	125	Los Angeles	United States	36.07	8925	Concrete
Kempegowda Int'l	3000	Bangalore	India	121.25	13,123	Asphalt
Sendai	15	Natori	Japan	120.4	9842	Asphalt
Warsaw Chopin	362	Warsaw	Poland	128.8	9186	Concrete
Brisbane Int'l	13	Brisbane	Australia	124.35	10,827	Asphalt
Jorge Newbery	18	Buenos Aires	Argentina	120.6	7710	Concrete
Seychelles Int'l	10	Mahe Island	Seychelles	119.7	9800	Concrete
Sofia	1742	Sofia	Bulgaria	123.7	11,811	Asphalt
Hewanorra Int'l	14	Vieux Fort	Saint Lucia	119.8	9003	Asphalt
La Guardia	21	New York	United States	118	7000	Asphalt

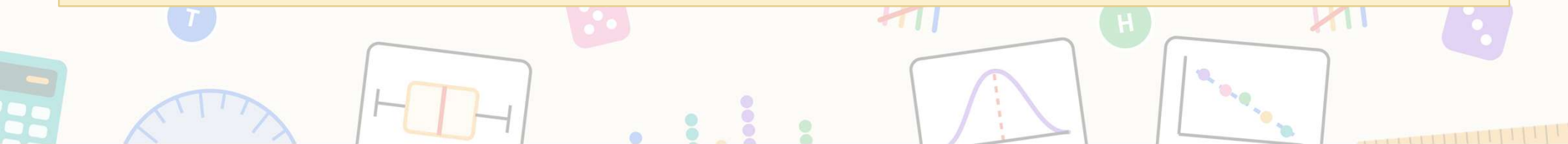
- (a) Identify the observational units and variables in this data set.
- (b) Classify each variable as categorical or quantitative.

The Statistical Problem-Solving Process

Every statistical study follows the same four-step rhythm — we will use it all year.

The four steps:

- 1. Ask questions:** pose a valid investigative question about a population.
- 2. Collect data:** design and carry out a plan for gathering data (a sample, an experiment).
- 3. Analyze data:** organize, display, and summarize what you collected.
- 4. Interpret results:** answer the investigative question in context.

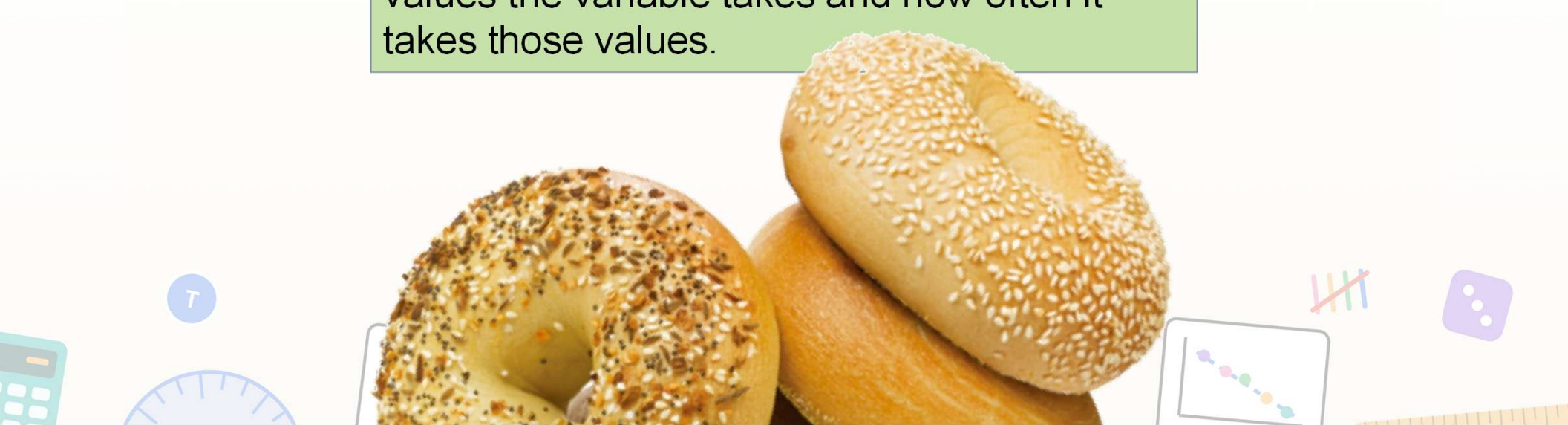


Analyzing Data

A **variable** generally takes values that vary. We are interested in the pattern of that variation.

Distribution

The distribution of a variable tells us what values the variable takes and how often it takes those values.



Organizing Categorical Data

Province	Gender	Number of languages spoken	Handedness	Height (cm)	Wrist circumference (mm)	Preferred communication
Saskatchewan	Male	1	Right	175.0	180	In person
Ontario	Female	1	Right	162.5	160	In person
Alberta	Male	1	Right	178.0	174	Facebook
Ontario	Male	2	Right	169.0	160	Cell phone
Ontario	Female	2	Right	166.0	65	In person
Nunavut	Male	1	Right	168.5	160	Text messaging
Ontario	Female	1	Right	166.0	165	Cell phone
Ontario	Male	4	Left	157.5	147	Text messaging
Ontario	Female	2	Right	150.5	187	Text messaging
Ontario	Female	1	Right	171.0	180	Text messaging

Categorical variable (arrow pointing to Preferred communication)

Values (These are the data) (arrow pointing to the values in the Preferred communication column)

Organizing Categorical Data

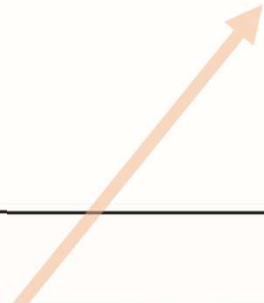
Frequency table	
Preferred method	Frequency
Cell phone	2
Facebook	1
In person	3
Text messaging	4

Count



Relative frequency table	
Preferred method	Relative frequency
Cell phone	$2/10 = 0.20$ or 20%

Proportion



Percent



Organizing Categorical Data

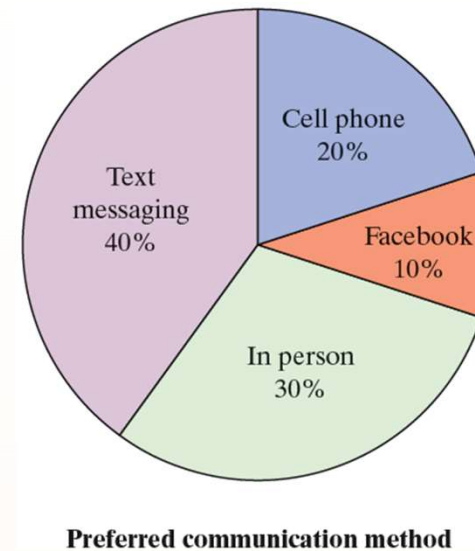
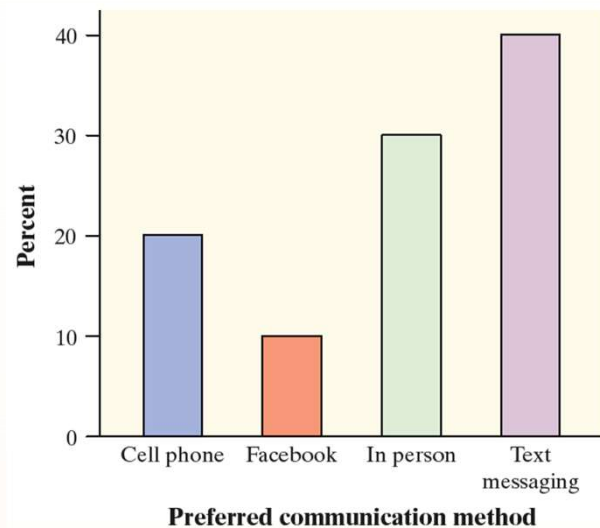
Province	Gender	Number of languages spoken	Handedness	Height (cm)	Wrist circum- ference (mm)	Preferred communication
Saskatchewan	Male	1	Right	175.0	180	In person
Ontario	Female	1	Right	162.5	160	In person
Alberta	Male	1	Right	178.0	174	Facebook
Ontario	Male	2	Right	169.0	160	Cell phone
Ontario	Female	2	Right	166.0	65	In person
Nunavut	Male	1	Right	168.5	160	Text messaging
Ontario	Female	1	Right	166.0	165	Cell phone
Ontario	Male	4	Left	157.5	147	Text messaging
Ontario	Female	2	Right	150.5	187	Text messaging
Ontario	Female	1	Right	171.0	180	Text messaging

On your guided notes, create a frequency table and relative frequency table for the variable, “Province”.



Displaying Categorical Data

To display the distribution of categorical data, make a **bar graph** or a **pie chart**.



Bar Graph

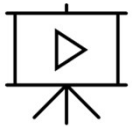
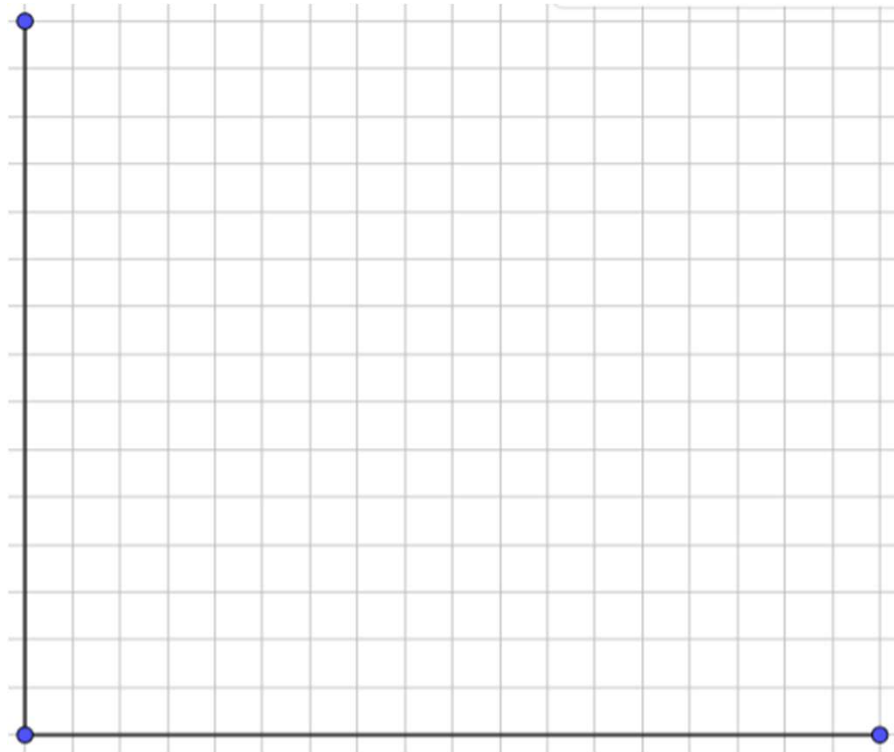
A **bar graph** shows each category as a bar. The heights of the bars show the category frequencies or relative frequencies.



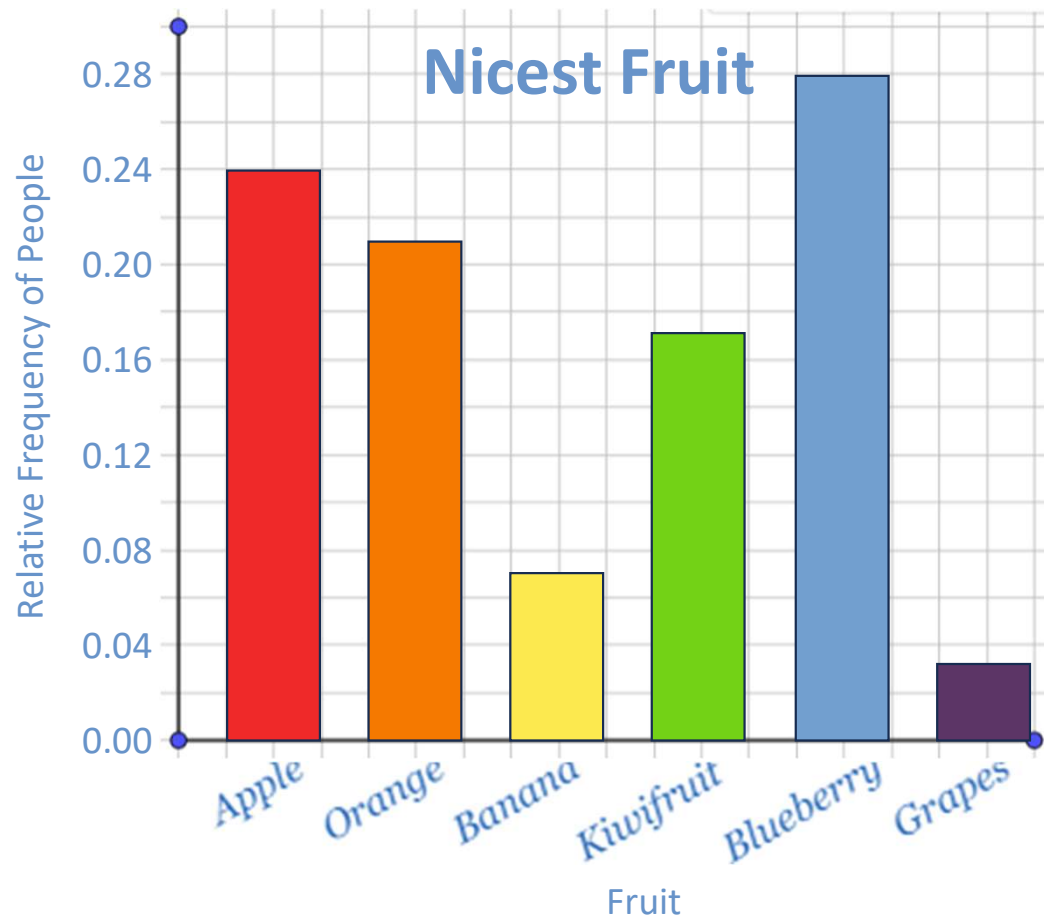
How to make a bar graph:

- 1) Draw and label the axes. Make sure to indicate whether the vertical axis is measuring frequency or relative frequency.
- 2) Write the names of the categories at equally spaced intervals under the horizontal axis.
- 3) Use equal spacing for same sized intervals on the vertical axis.
- 4) Draw bars (heights corresponding to frequencies or relative frequencies) above the category names. Make the bars equal in width and leave gaps between them.

Practice: Bar Graph



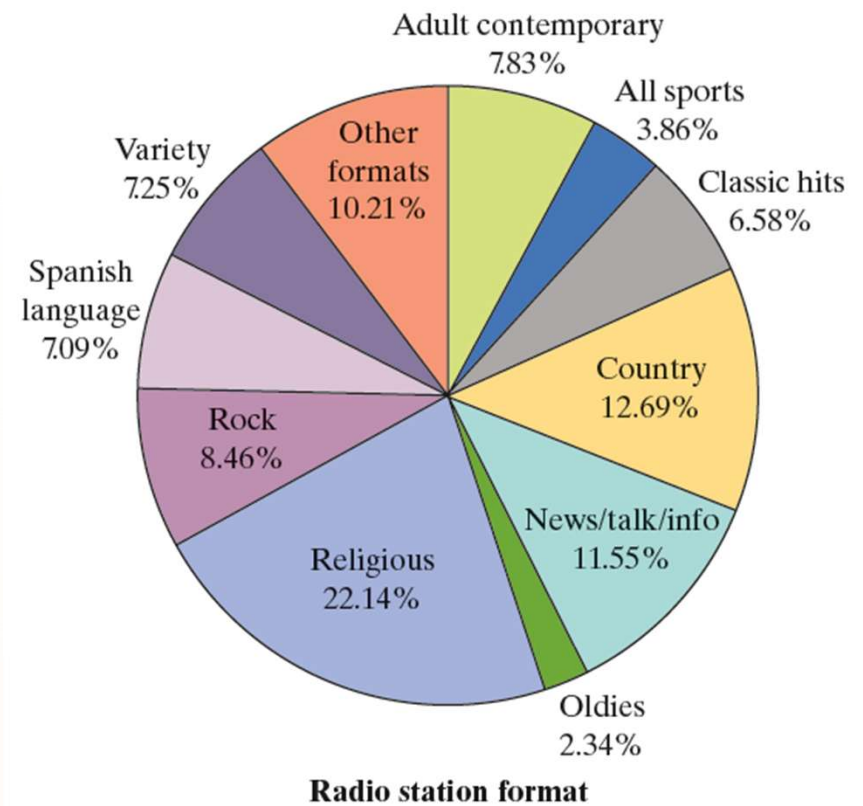
Practice: Bar Graph



Pie Charts

A **pie chart** is a graph of data for one categorical variable that displays each category as a slice of the "pie." The area of each slice is proportional to the category frequency or relative frequency.

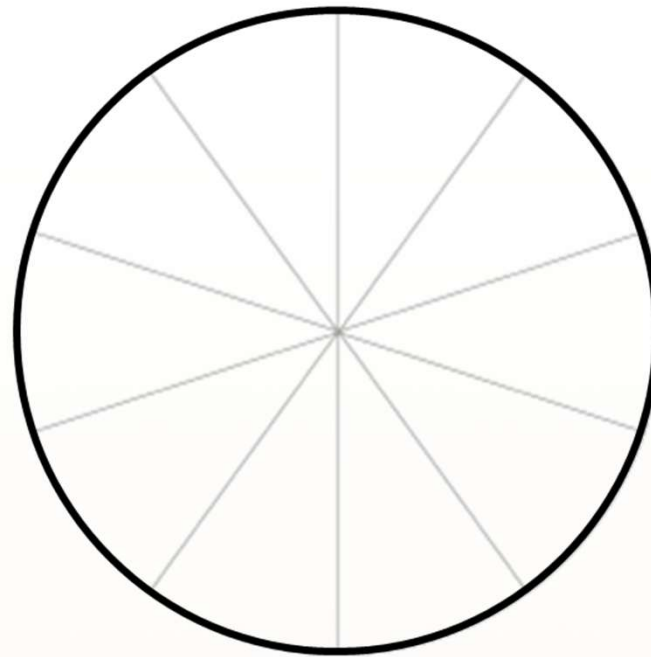
- Use a pie chart to emphasize each category's relation to the whole.
- Must include ALL observational units — add an "other" category if needed.
- Bar and pie charts also help justify a claim about the variable in context.



Practice: Pie Chart

Frequency table

Preferred method	Frequency
Cell phone	2
Facebook	1
In person	3
Text messaging	4



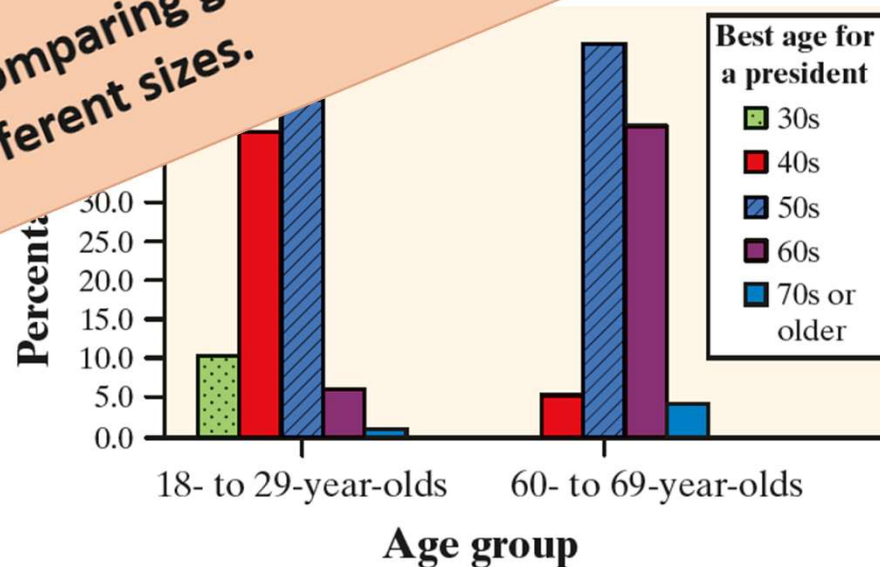
Comparing Distributions

A **side-by-side bar chart** displays the distribution of a categorical variable in two or more groups: a separate bar for each group in every category. Bar heights show the frequency or relative frequency within that group.

Pew Research asked U.S. adults the best age for a president. The chart compares 18- to 29-year-olds with 60- to 69-year-olds.

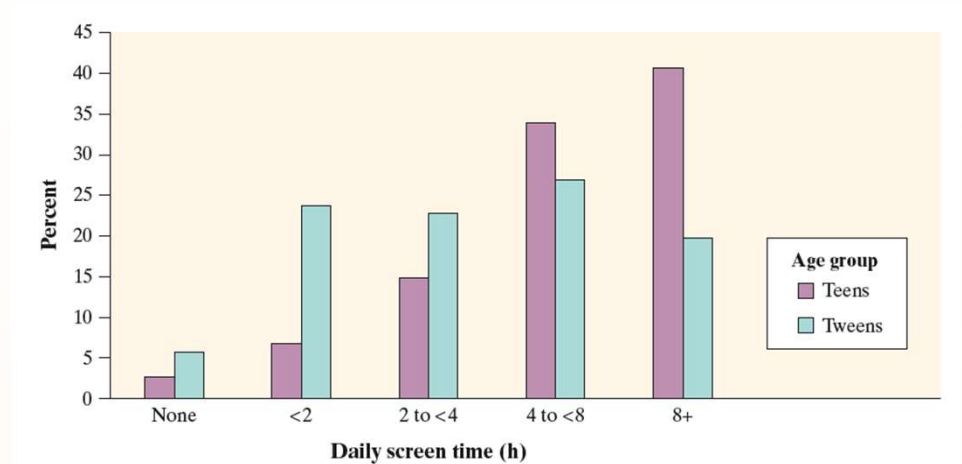
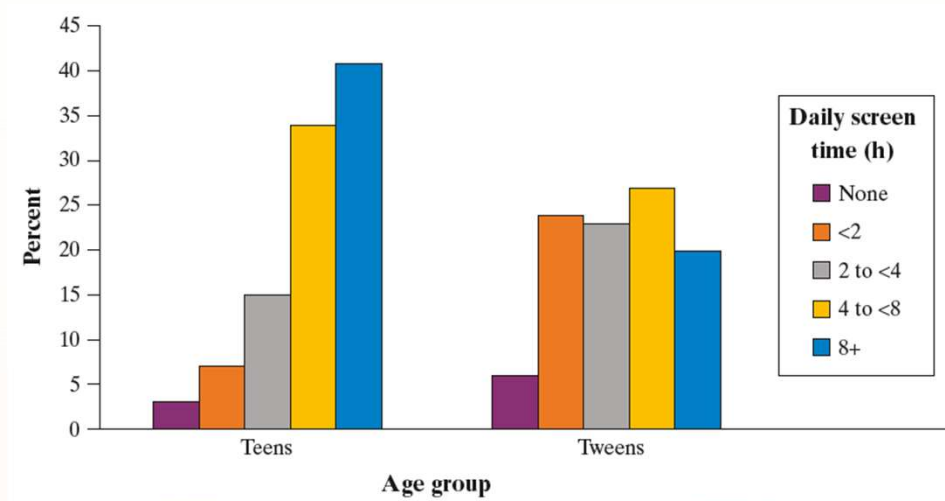
Compare the best age for a president for 18- to 29-year-olds with 60- to 69-year-olds.

AP® Exam Tip
Use relative frequencies when comparing groups — especially if the groups have different sizes.



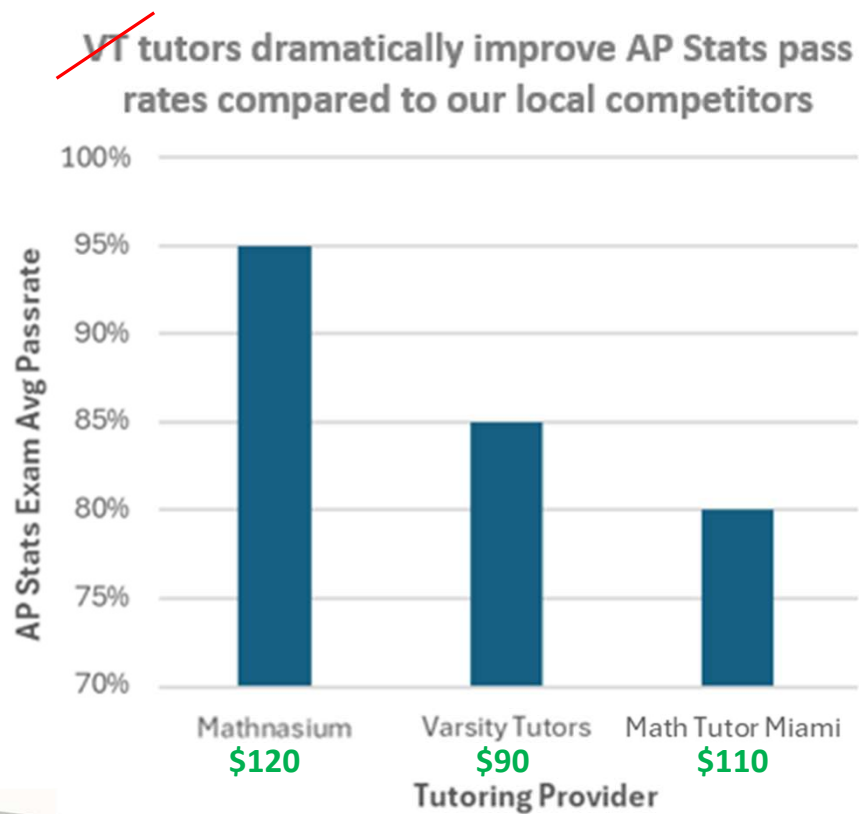
Two Ways to Group the Bars

- Same screen-time data, grouped two ways: by **age group** or by **daily screen time**.
- Grouping by screen time makes category-by-category comparisons easier.
- Always include a key!

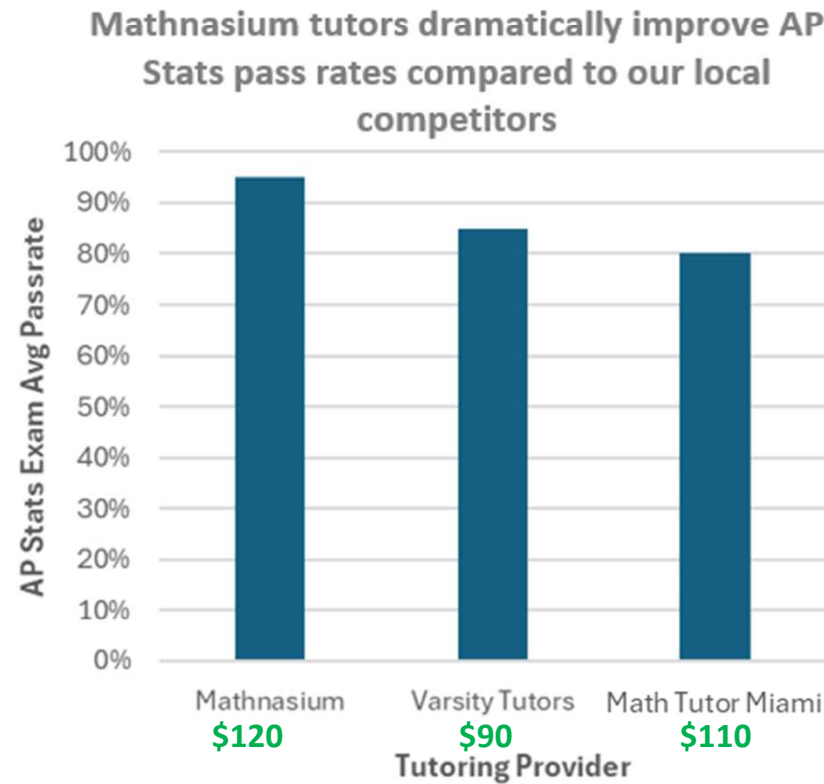


Interpreting Graphical Representations of Data

Mathnasium



What attributes makes this bar graph complete?



Reaching Conclusions from Graphs

- What is misleading about the graph?
- How could the graph be changed to reflect the data more accurately?
- Why would someone make a graph that is misleading in this way?

Spotting a misleading Graph:

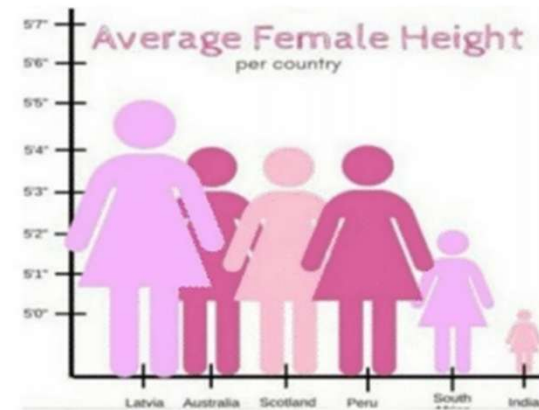
- X or Y-axis scale is too big or too small
- X or Y-axis skips numbers, or does not start at zero
- Axes are not labeled
- Data is left out
- Uses pictures for bar graphs (pictograph)



Pictograph

Bar graphs are a bit dull to look at. It is tempting to replace the bars with pictures or to use special 3-D effects to make the graphs seem more interesting.

Don't do it!



CAUTION:

- 1) beware the pictograph
- 2) watch those scales

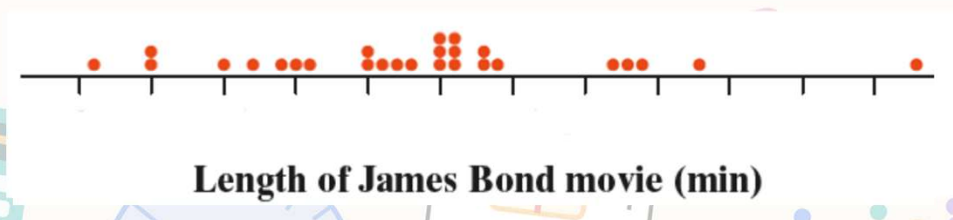
Dotplots

A **dotplot** shows each data value as a dot above its location on a number line.

- Simplest graph for a quantitative variable
- Easy to make by hand for small data sets
- Can also be oriented vertically

110 115 110 130 117 142 120 121 125 125 126 127
131 131 130 133 130 119 128 133 144 106 143 148
163 131 134

(a) Make a dotplot of these data.



(b) A movie over two hours is considered long. Can the graph justify the claim that more than 75% of these movies are long? Explain.

7 movies are 120 minutes or shorter, so $27 - 7 = 20$, and $20/27 = 74.1\%$ last longer than 2 hours. Because 74.1% is less than 75%, the dotplot cannot be used to justify the claim.

Stemplots

A **stem-and-leaf plot** displays each data value split into a **stem** (the leftmost digits) and a **leaf** (the final digit). Stems are ordered least to greatest in a vertical column; leaves are arranged in increasing order out from the stems.

These data represent the responses of 20 female AP Statistics students to the question, “How many pairs of shoes do you have?” Construct a stemplot.

50	26	26	31	57	19	24	22	23	38
13	50	13	34	23	30	49	13	15	51

1		93335
2		664233
3		1840
4		9
5		0701

Stems

Add leaves

1		33359
2		233466
3		0148
4		9
5		0017

Order leaves

Key: 4|9
represents a
female student
who reported
having 49
pairs of shoes.

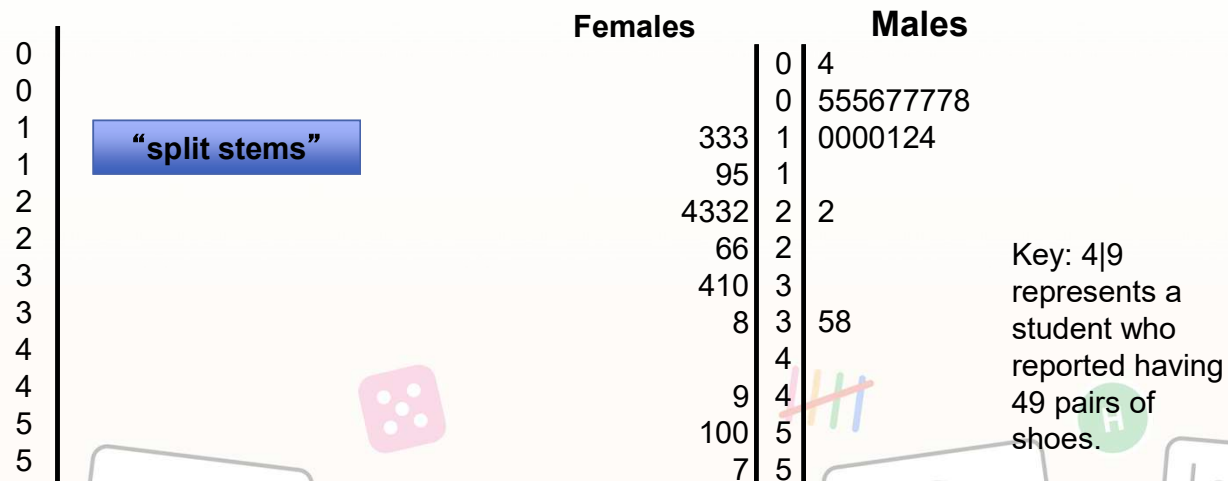
Add a key

Stemplots

When data values are “bunched up”, we can get a better picture of the distribution by **splitting stems**.

Two distributions of the same quantitative variable can be compared using a **back-to-back stemplot** with common stems.

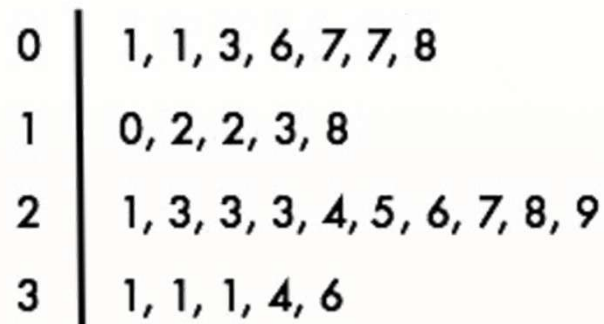
Females										Males									
50	26	26	31	57	19	24	22	23	38	14	7	6	5	12	38	8	7	10	10
13	50	13	34	23	30	49	13	15	51	10	11	4	5	22	7	5	10	35	7



Practice: Stemplots

- Create a split-stem stemplot using the data in the stemplot below:

Stem and Leaf Plot



Key: 1 | 2 = 12 inches

What is the range of the data?

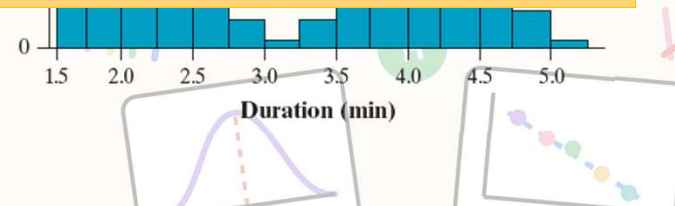
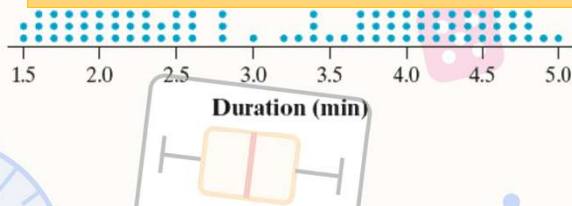
Histograms

A **histogram** shows each interval of values as a bar. The heights of the bars show the frequency of values in each interval.

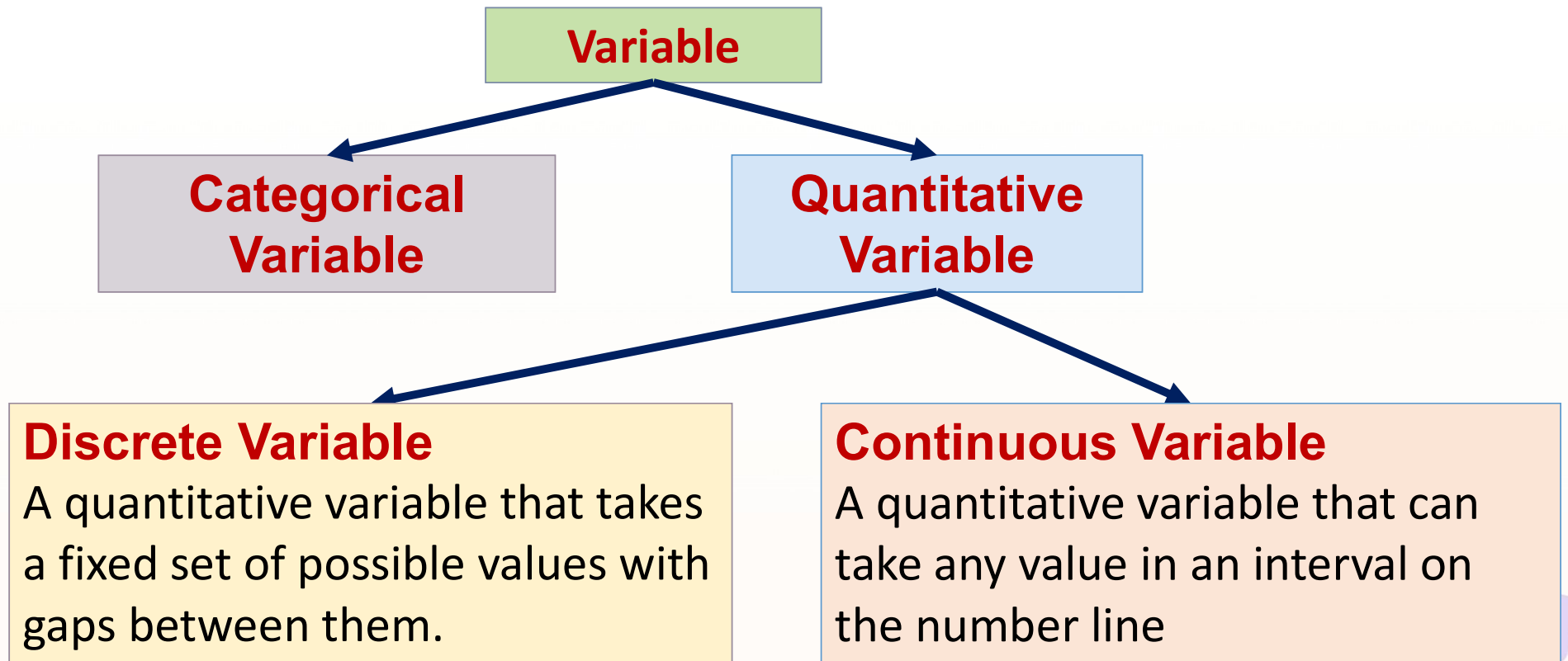


CAUTION:

- 1) Don't confuse histograms and bar graphs.
- 2) Use percents or proportions instead of counts on the vertical axis when comparing distributions with different numbers of observations.
- 3) Just because a graph looks nice doesn't make it a meaningful display of data.



Organizing Data



Discrete vs. Continuous Quantitative Variables

A **discrete variable** can take on a countable number of numeric values within a given range.

- These values are specific, and separate
- The number of values may be finite or countably infinite, as with counting numbers.
- Ex) number of visits to the doctor, number of skittles in a bag

A **continuous variable** represents a measurable quantity. They can take on infinitely many values, but those values cannot be counted.

- No matter how small the interval between two values of a continuous variable, it is always possible to determine another value between them.
- Ex) height of a student, pH of a solution

Let's Practice

Identify the following variables as Categorical, Discrete Quantitative, or Continuous Quantitative:

1. Number of Marvel Movies watched
2. Favorite Fast Food Restaurant
3. The Weight of an Elephant
4. Tickets sold to different Music Festivals
5. Time spent commuting to school
6. Number of steps taken between 1st and 2nd period
7. Favorite social media platform
8. Amount of lemonade in a jug at Chick-Fil-A

There are many ways to visualize quantitative variables. Some are better suited for discrete variables while others are more appropriate for continuous variables.

Describing Distributions

HOW TO DESCRIBE THE DISTRIBUTION OF A QUANTITATIVE VARIABLE

In any graph, look for the *overall pattern* and for clear *departures* from that pattern.

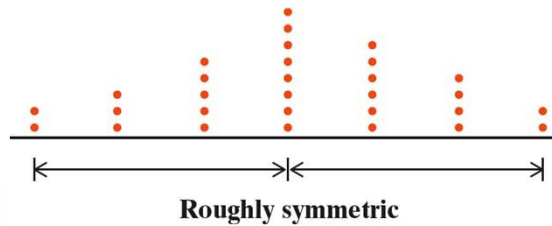
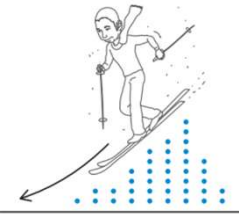
- You can describe the overall pattern of a distribution by its *shape*, *center*, and *variability*.
- An important kind of departure is an **outlier**, an observation that falls outside the overall pattern.

Don't forget your

SOCS!



Describing Shape

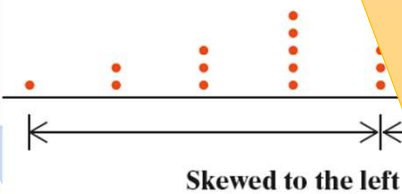


A distribution is **skewed to the right** if the right side of the graph is much longer than the left side.

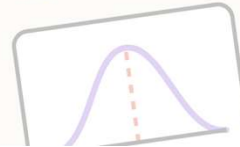
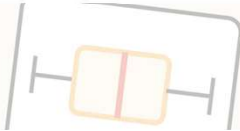


CAUTION:
The direction of skewness is toward the long tail, not the direction where most observations are clustered.

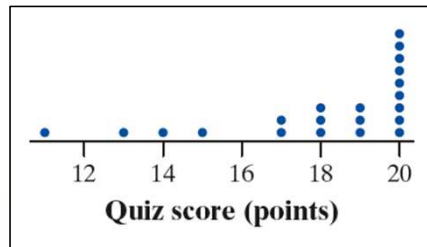
Skewed to the right



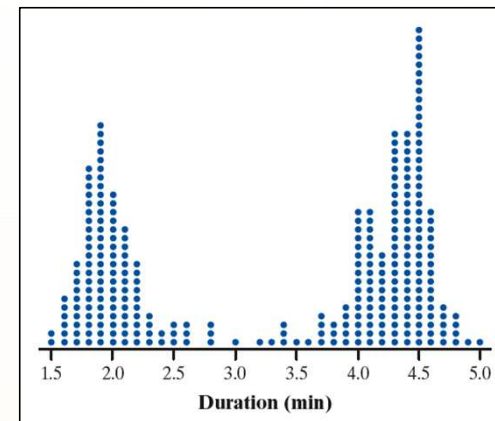
A distribution is **skewed to the left** if the left side of the graph is much longer than the right side.



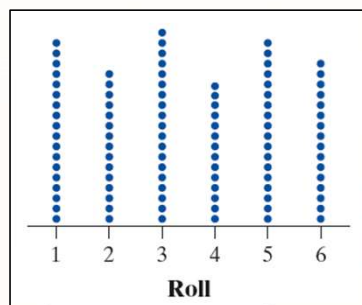
Describing Shape



The distribution of a quantitative variable is **unimodal** if it has a single peak.

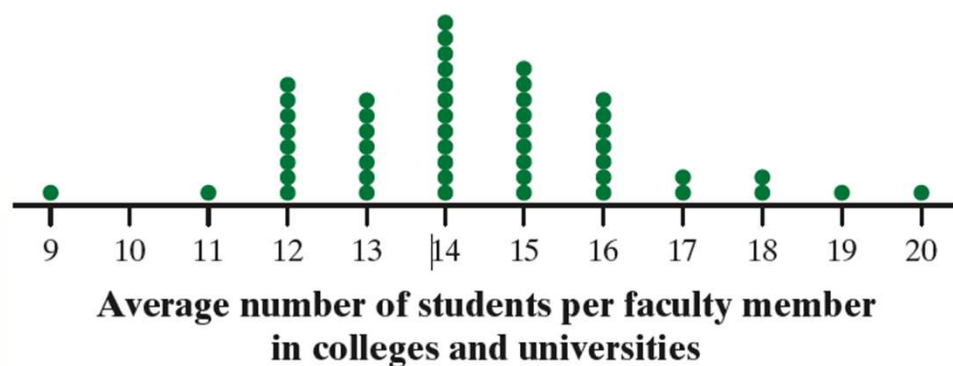


The distribution of a quantitative variable is **bimodal** if it has two distinct clusters and peaks.

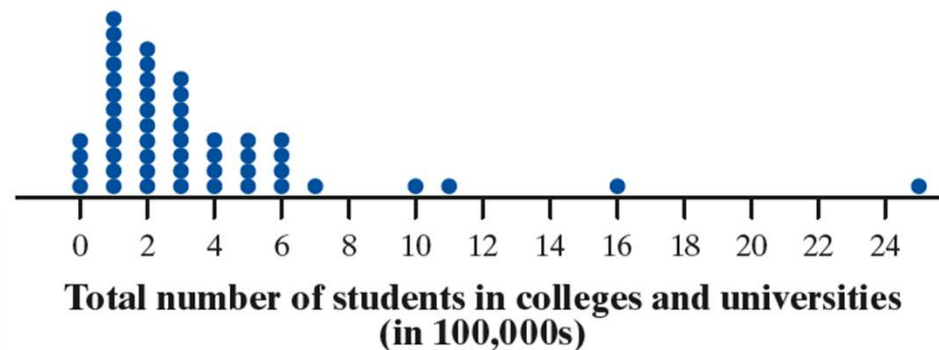


The distribution of a quantitative variable is **approximately uniform** if the frequencies are about the same for all values.

Identifying Shape: Describe the shape of each graph:



(a) Roughly symmetric, with a single peak at about 14 students per faculty member.



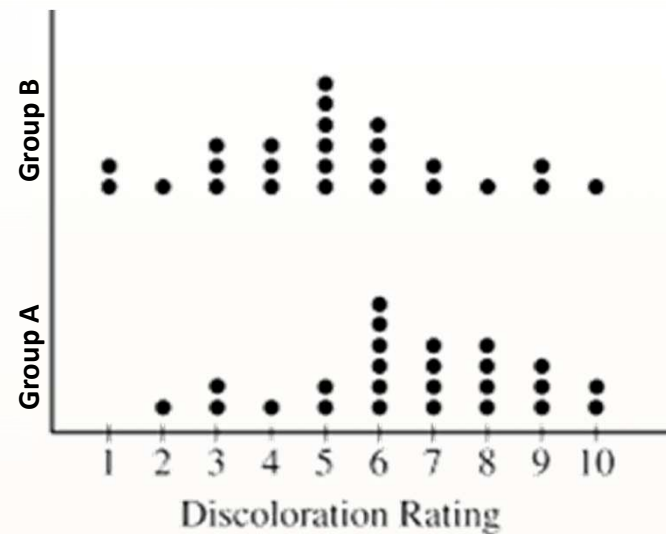
(b) Strongly skewed to the right, with a single peak at 100,000 students. Gaps between 0.7-1.0, 1.1-1.6, and 1.6-2.5 million students.

Describing Center

Here we are looking for one value that you feel best describes all of the data.

Females		Males
	0	4
	0	555677778
333	1	0000124
95	1	
4332	2	2
66	2	
410	3	
8	3	58
	4	
9	4	
100	5	
7	5	

Key: 2|2 represents
a male student with
22 pairs of shoes.



Describing Unusual Features: Outliers

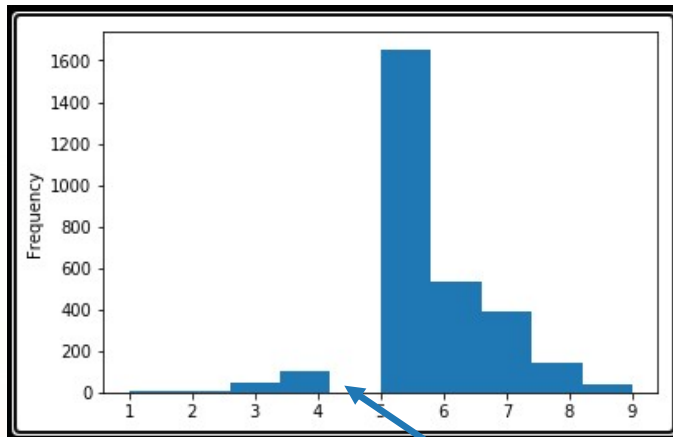
An **outlier** is a value that lies away from the rest of the data. Always think for possible reasons for outliers.

2	5									
3										
4	0									
5	8									
6	8	9	9							
7	0	0	1	3	3	3	4	6	7	8
8	0	3	3	6	7					
9	0	0	0	2	3	5	7			
10	0	0								

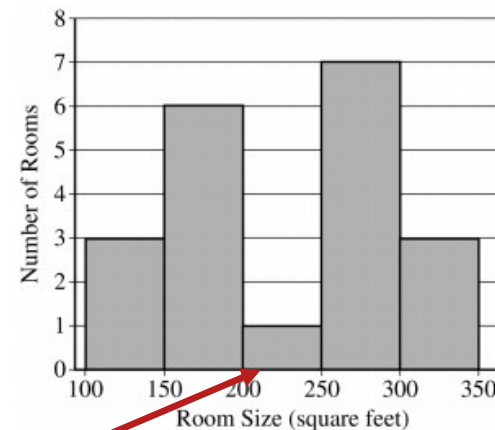


Describing Unusual Features: Gaps

A **gap** is a region of a distribution between two data values where there are no observed data. Think what gaps could mean.



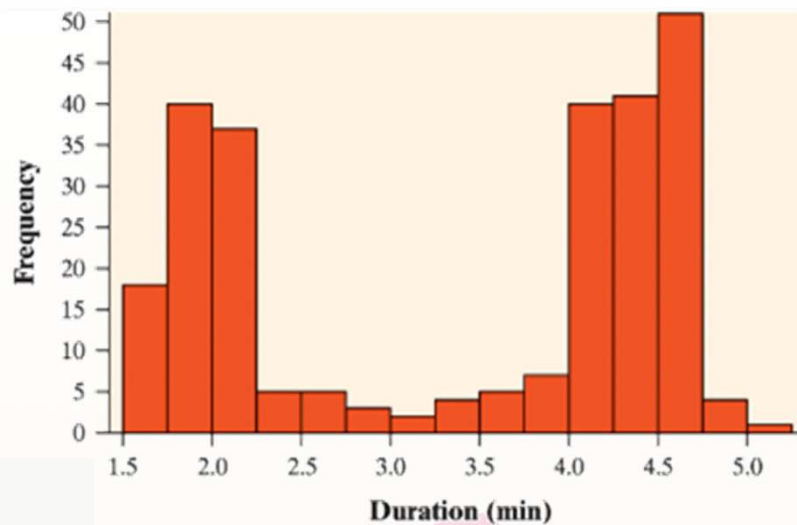
Gap



Not a Gap

Describing Unusual Features: Clusters

Clusters are concentrations of data usually separated by gaps. What could cause this?



There seem to be two distinct clusters with one group taking roughly 2 minutes and the other taking closer to 4.5 minutes to solve the puzzle.

Describing Variability (Spread)

Here you need to talk in simple terms about the range the data values fall in and where maybe the majority of the data lies.

6	8
7	
8	8
9	79
10	08
11	15566
12	012223444457888999
13	01233333444899
14	02666
15	23
16	8

The variability of the patient heart rates were from 68 to 168, but the large majority of patients had heart rates from about 110 to 140 beats per minute while exercising.



Describing Distributions

- **Shape** - symmetric or skewed?
peaks, clusters, gaps
- **Outliers** - any values that stand apart?
- **Center** - a typical or middle value
- **Spread** - the variability of the data values

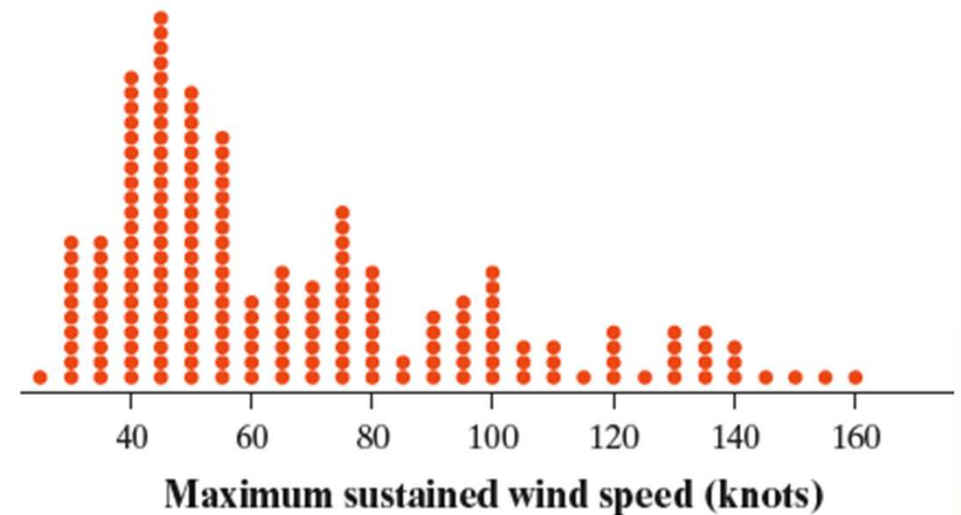
Don't forget your

SOCS!



Cyclone speeds

- Tropical cyclones are classified by maximum sustained wind speed: tropical depression, tropical storm, hurricane, major hurricane
- Dotplot: maximum sustained wind speed (knots) for the 193 Atlantic-basin cyclones over an 11-year period
- One knot is a little faster than 1 mph



Describe the distribution.



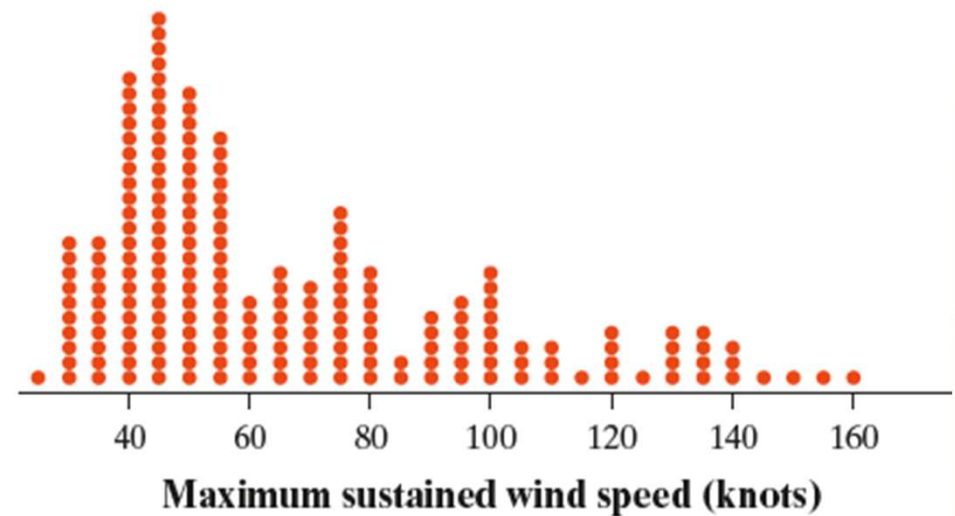
Solution: Cyclone speeds

Shape: strongly skewed to the right, with a single peak at about 45 knots; no major gaps.

Outliers: there do not appear to be any outliers.

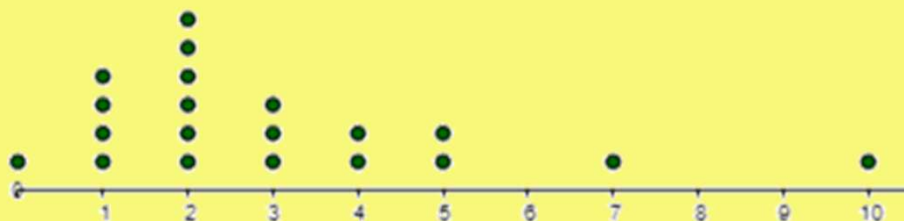
Center: the middle value is about 55 knots (median = 55 knots).

Spread: wind speeds vary from about 25 to about 160 knots.



Practice FRQ!!!

The dotplot below shows the number of goals scored by the 20 teams playing in a city's high school soccer games on a particular day.

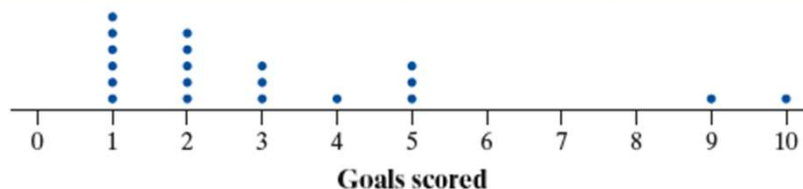


- (a) Describe the distribution.
- (b) One superstar scored six goals, but his team still lost. What are all possible final scores for that game? Explain.
- (c) Is it possible that each of the teams that scored exactly two goals were the winners of that game? Explain.

Describing Distributions

The dotplot below shows the number of goals scored by the 20 teams playing in a city's high school soccer games on a particular day.

Did we include context? **YES!**



Shape: The distribution of goals scored is skewed to the right, with a single peak at 1 goal. There is a gap between 5 and 9 goals.

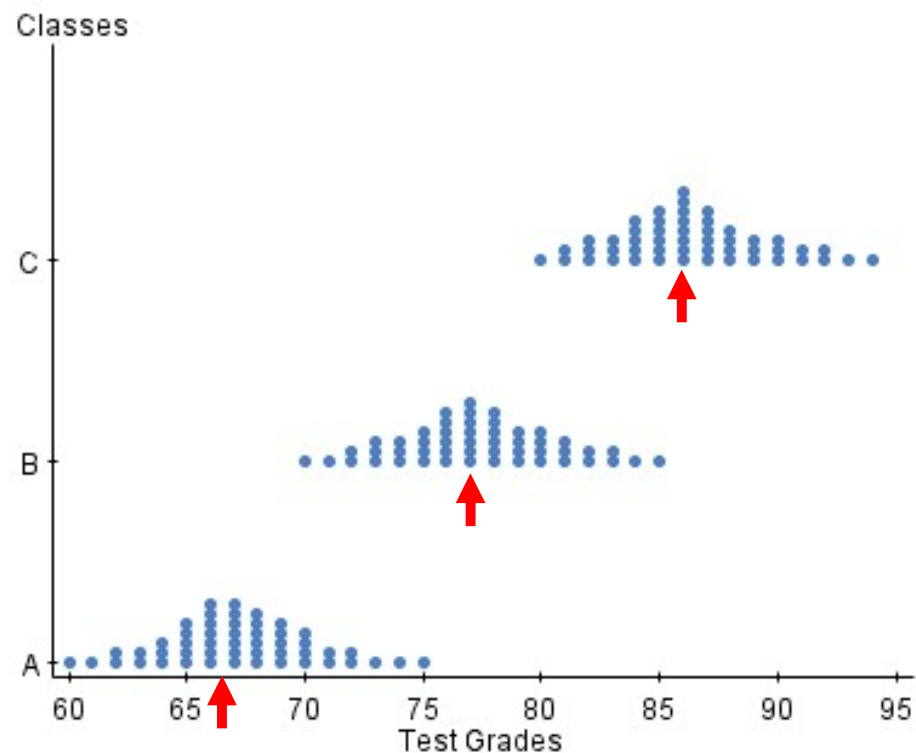
Outliers: The games when the team scored 9 and 10 goals appear to be outliers.

Center: The center seems to be around 2 goals scored.

Variability: The number of goals varies from 1 to 10 goals scored.

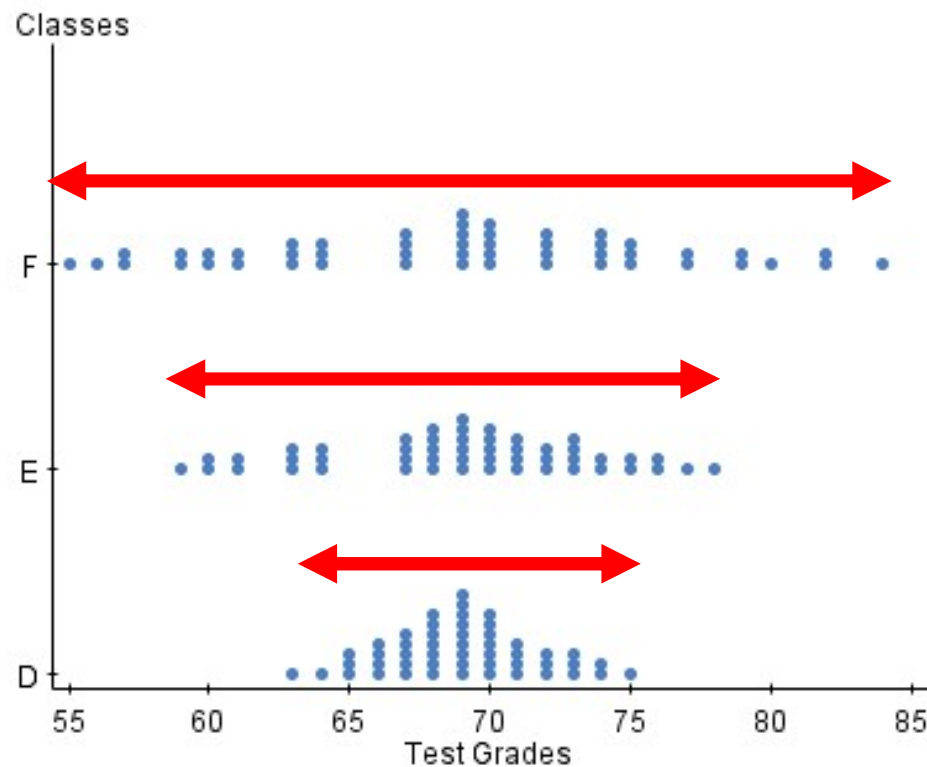
Comparing Distributions

What strikes you as the most distinctive **difference** among the distributions of exam scores in classes A, B, & C ?



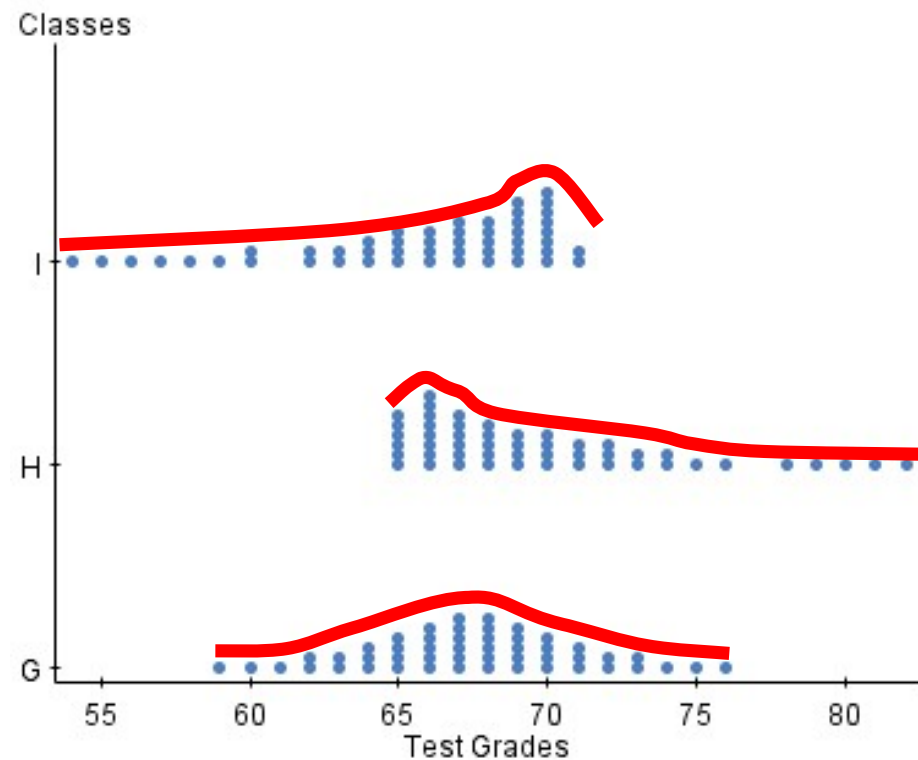
Comparing Distributions

What strikes you as the most distinctive **difference** among the distributions of scores in classes D, E, & F?



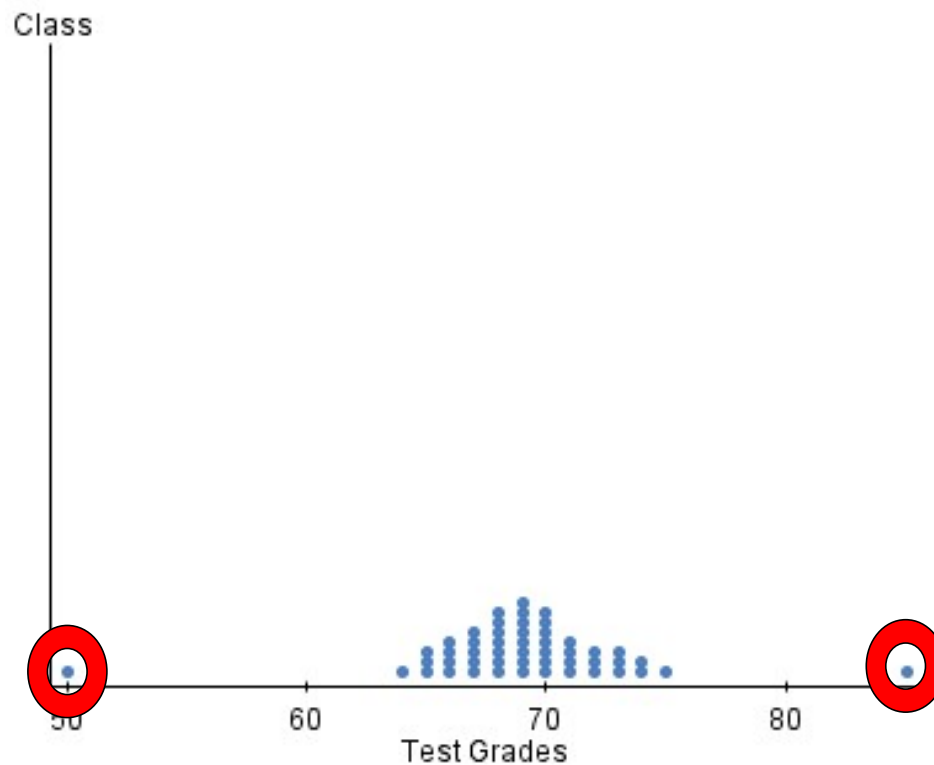
Comparing Distributions

What strikes you as the most distinctive **difference** among the distributions of exam scores in classes G, H, & I?



Comparing Distributions

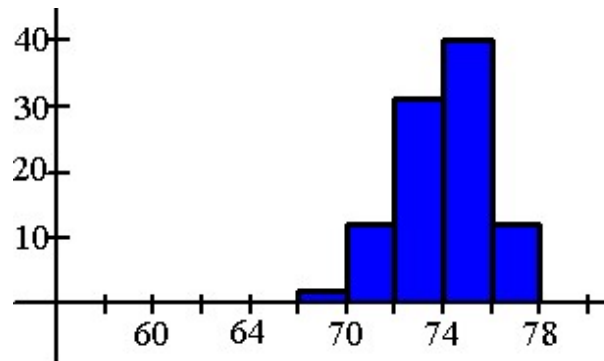
What strikes you as the most distinctive **characteristic** among the distribution of exam scores in class J ?



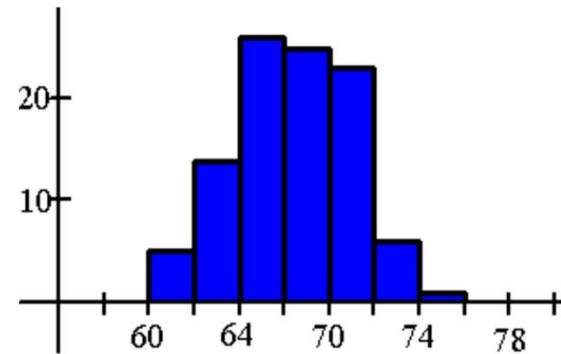
If the Shoe Fits...

Suppose a shoe store (that has only 16 shoes) is
distributed for the heights of men, with
distribution of heights of female basketball
players? Why?

Which is which? Why?



Heights – Figure A



Heights – Figure B

Measuring Center: The Median

The **median** is the midpoint of a distribution. It divides an ordered dataset in half.

To find the median, arrange the data in numerical order.

- If the number n of data values is odd, the median is the middle value in the ordered list.
- If the number n of data values is even, the median is the average of the two middle values in the ordered list.

$$\text{Position of the median} = \frac{N + 1}{2}$$

Measuring Center: The Median

Suppose we catch a sample of 5 fish from the lake. The lengths of the fish (in inches) are listed below.

Find the median length of the fish.

3 4 5 8 10

This is an ordered list with odd length.
Look for the middle #. The median is **5**.

Measuring Center: The Median

Suppose we caught a sample of 6 fish from the lake. The lengths of the fish (in inches) are listed below.

Find the median length of the fish.

3 4 5 6 8 10

This is an ordered list with even length.

Look for the middle 2 #s and take their average.

The median is $\frac{5+6}{2} = 5.5$.

Measuring Center: The Median

Let's Practice: Find the median length of a sample of 12 fish.

8 4 2 1 9 2
3 7 5 4 8 6

This is an unordered list with even length. First order the list.

Look for the middle 2 #s and take their average.

The median is $\frac{4+5}{2} = 4.5$.

Measuring Center: The Mean

The **mean** of a distribution of quantitative data is its arithmetic average.

If the n observations are $x_1, x_2, \dots, x_n$, the sample mean $\bar{x}$ (pronounced “x-bar”) is given by the following formula:

$$\bar{x} = \frac{\text{sum of data values}}{\text{number of data values}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

Measuring Center: The Mean

The symbol $\bar{x}$ refers to the mean of a *sample*.

← Statistic

The notation μ refers to the mean of a *population*.

← Parameter

REMEMBER

A **statistic** is a number that describes some characteristic of a *sample*.

A **parameter** is a number that describes some characteristic of a *population*.

Measuring Center: The Mean

Suppose we caught a sample of 6 fish from the lake. The lengths of the fish (in inches) are listed below.

Find the mean length of the fish.

3 4 5 6 8 10

$$\bar{x} = \frac{3 + 4 + 5 + 6 + 8 + 10}{6} = 6$$

Measuring Center: The Mean

Suppose we caught a sample of 6 fish from the lake. The lengths of the fish (in inches) are listed below.

Find how each observation deviates from the mean.

x	$x - \bar{x}$
3	$3 - 6 = -3$
4	$4 - 6 = -2$
5	$5 - 6 = -1$
6	$6 - 6 = 0$
8	$8 - 6 = 2$
10	$10 - 6 = 4$
Sum	0

The mean is considered the **balance point** of the distribution because it “balances” the positive and negative deviations.



Will this sum always equal zero?

YES!

Comparing Measures of Center

3 4 5 6 8 15

What happens if we found out that due to a measuring error, the 10in fish was actually 15in?

What happens to the median?

Still 5.5 in

What happens to the mean?

It increases from 6 in to 6.833 in

Comparing Measures of Center

3 4 5 6 7 8 9 10

What happens
in?

What happens

What happens to

fish was actually 20

CAUTION:

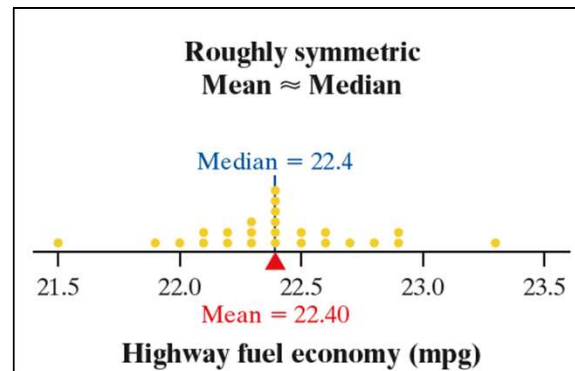
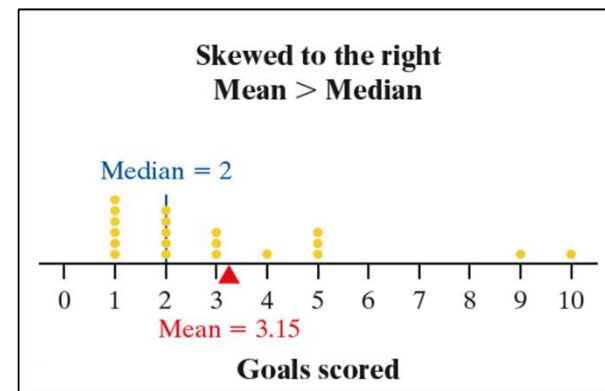
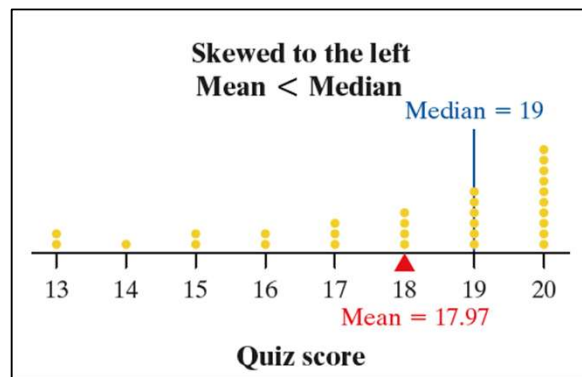


The mean is sensitive to extreme values in a distribution. The mean is **not** a resistant measure of center.

It increases from 6.83 in to 7.67 in

A statistical measure is **resistant** if it isn't sensitive to extreme values.

Comparing the Mean and the Median



Comparing the Mean and the Median

Effect of Skewness and Outliers on Measures of Center

- If a distribution of quantitative data is roughly symmetric and has no outliers, the mean and median will be similar.
- If the distribution is strongly skewed, the mean will be pulled in the direction of the skewness but the median won't.
 - For a right-skewed distribution, we expect the mean to be greater than the median.
 - For a left-skewed distribution, we expect the mean to be less than the median.
- The median is resistant to outliers but the mean isn't.
 - **Make sure to use an appropriate measure of center based on the shape of the distribution.**

Challenge: Extensions of Mean (#1)

- Winter Break is approaching and as promised, a class is gearing up for their Unit 5 Exam.
- So far, a students' exams scores have been:
 - 95, 83, 91, and 87
- What does she need to score on the Unit 5 Exam to have an exam average of 90 for the first semester?

$$\frac{95 + 83 + 91 + 87 + x}{5} = 90$$

$$x = 94$$

Challenge: Extensions of Mean (#2)

- On the first knowledge check, Mr. Levy's 7th Period (31 students) averaged 3.2 points. 8th Period (18 students) averaged 4.1. What is the average for Mr. Levy's 7th and 8th Period classes combined?
- We need to account for the fact that the two classes are not the same size.

$$\frac{3.2(31) + 4.1(18)}{31 + 18} = 3.531$$

Motivating Example for Position

- Imagine that on the Unit 1 Exam, you score a 74%. How do you feel about this score?
- What if I tell you the average score was a 66%?
- What if I tell you that you had the 2nd highest score in the class?
- **Numbers without context are meaningless!**



Measuring Location: Percentiles

One way to describe the location of a value in a distribution is to tell what percent of observations are less than or equal to it.

An individual's **percentile** is the percent of values in a distribution that are less than or equal to the individual's data value.

Jenny earned a score of 86 on her test. How did she perform relative to the rest of the class?

6 | 7

7 | 2334

7 | 5777899

8 | 00123334

8 | 569

9 | 03



Measuring Location: Percentiles

One way to describe the location of a value in a distribution is to tell what percent of observations are less than or equal to it.

An individual's **percentile** is the percent of values in a distribution that are less than or equal to the individual's data value.

Jenny earned a score of 86 on her test. How did she perform relative to the rest of the class?

6 | 7

7 | 2334

7 | 5777899

8 | 00123334

8 | 569

9 | 03

Jenny's score was greater than 21 of the 25 observations. Since 22 of the 25, or 88%, of the scores are at or below hers, Jenny's score is at the 88th percentile in the distribution of test scores for the class.

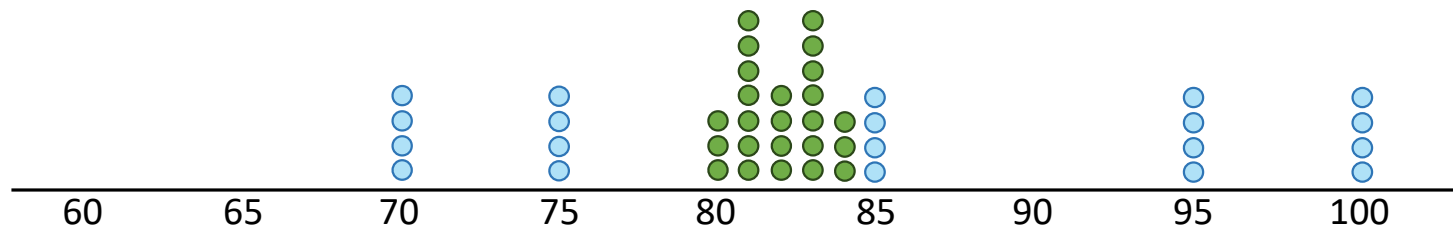


Measuring Variability

Imagine you learn that there are two professors who teach a course you want to take as a first-year college student. You have heard that both professors are brilliant in their field, have high RateMyProfessor scores and glowing reviews!

They teach using the same curriculum (slides, assignments, tests, etc.) however create their own Final Exams. Professor A's Final Exam always tends to have a higher average score than Professor B's (85 vs. 82).

Which Professor do you choose and why?



Measuring Variability: The Range

The **range** of a distribution is the distance between the minimum value and the maximum value. That is,
$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

$$\text{Range} = 10 - 1 = 9 \text{ goals}$$



CAUTION:

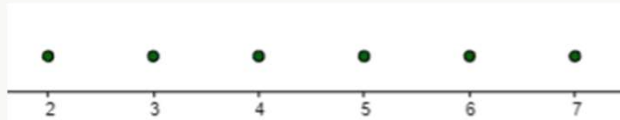
- The range of a data set is a single number.
- The range is **not** a resistant measure of variability.



Measuring Variability: The Range

Please note, that all three of these samples have the same range of $7 - 2 = 5$.

A



B



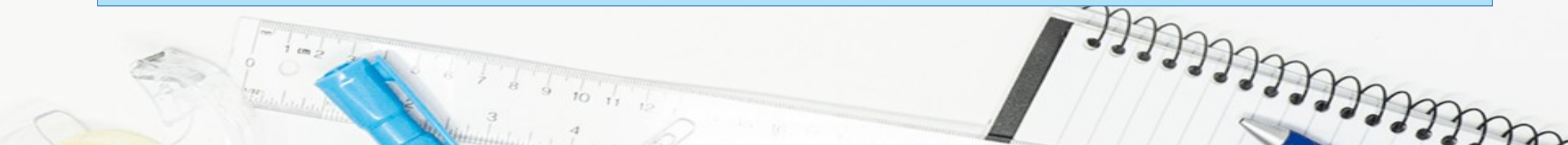
C



CAUTION:

- The range of a data set is a single number.
- The range is **not** a resistant measure of variability.

Another idea to quantify variability (spread) is to consider the deviations from the mean of the individual observations in the distribution. . .



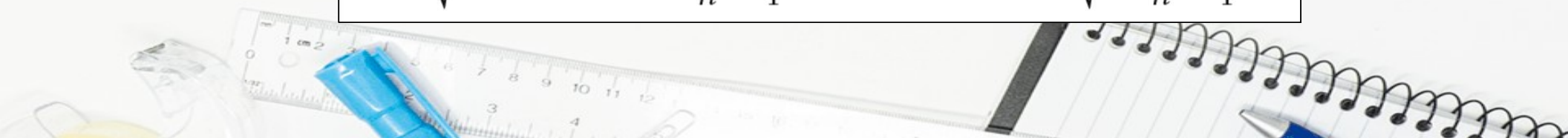
Measuring Variability: The Standard Deviation

How to calculate standard deviation and variance:

- 1) Find the mean of the distribution.
- 2) Calculate the *deviation* (value – mean) of each value from the mean.
- 3) Square each of the deviations.
- 4) Add all the squared deviations, divide by $n - 1$. This is the **sample variance**.
- 5) Take the square root. This is the **sample standard deviation**.

The **standard deviation** measures the typical distance of the values in a distribution from the mean.

$$s_x = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$



Measuring Variability: The Standard Deviation

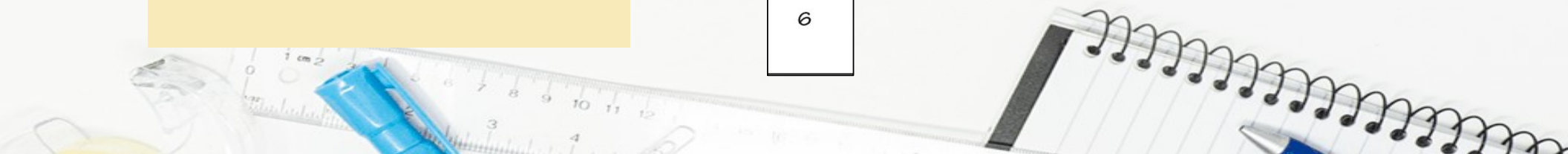
Eleven high school students were asked how many “close” friends they have. Here are their responses: 1 2 2 2 3 3 3 3 3 4 4 6

How to calculate standard deviation, s_x :

- 1) Find the mean of the distribution.

$$\bar{x} = \frac{1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 4 + 4 + 6}{11} = 3$$

x_i
1
2
2
2
3
3
3
3
3
4
4
6



Measuring Variability: The Standard Deviation

Eleven high school students were asked how many “close” friends they have. Here are their responses: 1 2 2 2 3 3 3 3 3 4 4 6

How to calculate standard deviation, s_x :

- 1) Find the mean of the distribution.
- 2) Calculate the *deviation* of each value from the mean: deviation = value – mean.

$$\bar{x} = \frac{1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 4 + 4 + 6}{11} = 3$$

x_i	$x_i - \bar{x}$
1	$1 - 3 = -2$
2	$2 - 3 = -1$
2	$2 - 3 = -1$
2	$2 - 3 = -1$
3	$3 - 3 = 0$
3	$3 - 3 = 0$
3	$3 - 3 = 0$
3	$3 - 3 = 0$
4	$4 - 3 = 1$
4	$4 - 3 = 1$
6	$6 - 3 = 3$

Measuring Variability: The Standard Deviation

Eleven high school students were asked how many “close” friends they have. Here are their responses: 1 2 2 2 3 3 3 3 3 4 4 6

How to calculate standard deviation, s_x :

- 1) Find the mean of the distribution.
- 2) Calculate the *deviation* of each value from the mean: deviation = value – mean.
- 3) Square each of the deviations.

$$\bar{x} = \frac{1 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 4 + 4 + 6}{11} = 3$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	$1 - 3 = -2$	$(-2)^2 = 4$
2	$2 - 3 = -1$	$(-1)^2 = 1$
2	$2 - 3 = -1$	$(-1)^2 = 1$
2	$2 - 3 = -1$	$(-1)^2 = 1$
3	$3 - 3 = 0$	$0^2 = 0$
3	$3 - 3 = 0$	$0^2 = 0$
3	$3 - 3 = 0$	$0^2 = 0$
3	$3 - 3 = 0$	$0^2 = 0$
4	$4 - 3 = 1$	$1^2 = 1$
4	$4 - 3 = 1$	$1^2 = 1$
6	$6 - 3 = 3$	$3^2 = 9$
		Sum = 18

Measuring Variability: The Standard Deviation

Eleven high school students were asked how many “close” friends they have. Here are their responses: 1 2 2 2 3 3 3 3 4 4 6

How to calculate standard deviation, s_x :

- 1) Find the mean of the distribution.
- 2) Calculate the *deviation* of each value from the mean.
- 3) Square each of the deviations.
- 4) Add all the squared deviations, divide by $n - 1$.

$$s_x^2 = \frac{18}{11-1} = 1.80$$

This value is known as the **sample variance**. (In this case, the units would be “squared close friends.”)



Measuring Variability: The Standard Deviation

Eleven high school students were asked how many “close” friends they have. Here are their responses: 1 2 2 2 3 3 3 3 4 4 6

How to calculate standard deviation, s_x :

- 1) Find the mean of the distribution.
- 2) Calculate the *deviation* of each value from the mean.
- 3) Square each of the deviations.
- 4) Add all the squared deviations, divide by $n - 1$.
- 5) Take the square root.

$$s_x = \sqrt{\frac{18}{11 - 1}} = 1.34 \text{ close friends}$$

This value is known as the **sample standard deviation**.



Measuring Variability: The Standard Deviation

Properties of Standard Deviation

- s_x is always greater than or equal to 0.
- Larger values of s_x indicate greater variation.
- s_x is not a resistant measure of variability.
- s_x measures variation about the mean.



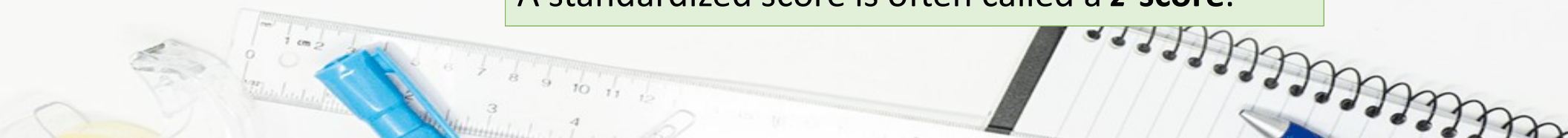
Measuring Position: z-Scores

A z-score tells us how many standard deviations from the mean an observation falls, and in what direction.

If x is an observation from a distribution that has known mean and standard deviation, the **standardized score** of x is:

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

A standardized score is often called a **z-score**.



Measuring Position: z-Scores

Jenny earned a score of 86 on her test. The class mean is 80 and the standard deviation is 6.07. Find and interpret her standardized score.



CAUTION:

Do not interpret the z-score as a distance "away" from the mean. Always indicate if the value is greater than (above) or less than (below) the mean.

Motivating Question

When comparing two applicants in an admissions office, would you prefer to admit the student who scored a 1100 on the SAT or the student who scored a 1500 on the SAT (all else equal)?

That should be an easy decision. . .

When comparing two applicants in an admissions office, would you prefer to admit the student who scored a 1400 on the SAT or the student who scored a 32 on the ACT (all else equal)?

That becomes a bit more tricky. . .

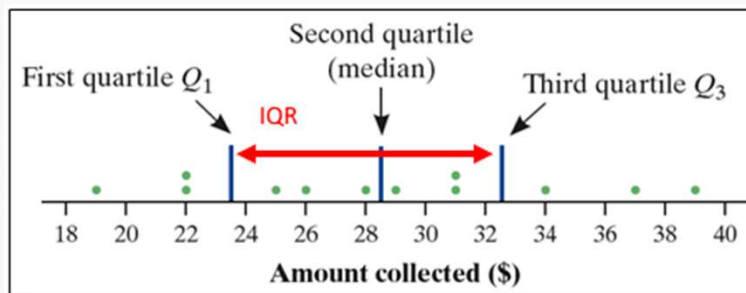


Measuring Variability: The Interquartile Range (IQR)

The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

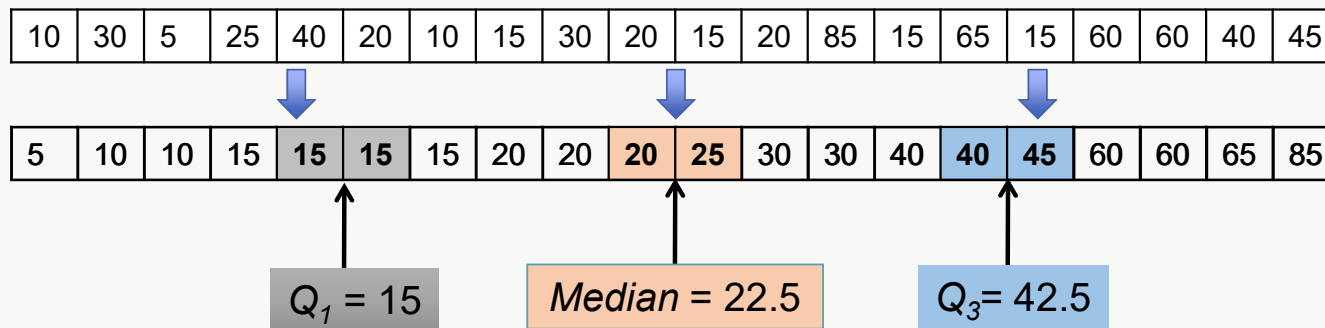
The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.



The **interquartile range (IQR)** is the distance between the first and third quartiles of a distribution. In symbols:
$$IQR = Q_3 - Q_1$$

Measuring Variability: The Interquartile Range (IQR)

Travel times for 20 New Yorkers:



$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 42.5 - 15 \\ &= 27.5 \text{ minutes} \end{aligned}$$

Interpretation: The range of the middle half of travel times for the New Yorkers in the sample is 27.5 minutes.

Measuring Variability: The Interquartile Range (IQR)

Examples

1	6	12	16	17	19	31	33	36
---	---	----	----	----	----	----	----	----

$$\text{IQR} = 32 - 9 = 23$$

3	5	7	8	9	11	15	16	20	21
---	---	---	---	---	----	----	----	----	----

$$\text{IQR} = 16 - 7 = 9$$

4	6	7	9	12	14	16	22	26	38	42	55
---	---	---	---	----	----	----	----	----	----	----	----

$$\text{IQR} = 32 - 8 = 24$$



Choosing Measures of Center/Spread

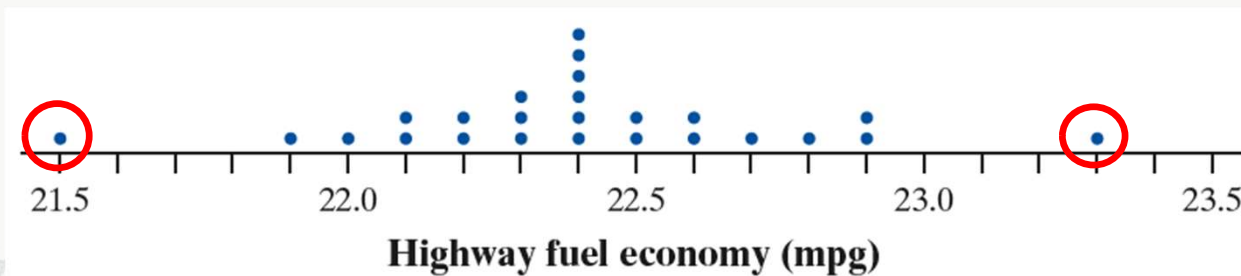
- The mean and standard deviation are always paired together to describe center and spread.
- **Non-Resistant**
- Use them when your data has a roughly symmetric shape with no major outliers.
- The median and IQR are always paired together to describe center and spread.
- **Resistant**
- Use them when your data is skewed to either side or has major outliers.



Identifying Outliers

Why look for outliers?

1. They might be inaccurate data values.
2. They can indicate a remarkable occurrence.
3. They can heavily influence the values of some summary statistics, like the mean, range, and standard deviation.



Identifying Outliers (Method 1)

Although there are several rules for outliers, one of the most common rules is the $1.5 \times IQR$ rule.

HOW TO IDENTIFY OUTLIERS: THE $1.5 \times IQR$ RULE

Call an observation an outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile. That is,

Low outliers $< Q_1 - 1.5 \times IQR$ High outliers $> Q_3 + 1.5 \times IQR$

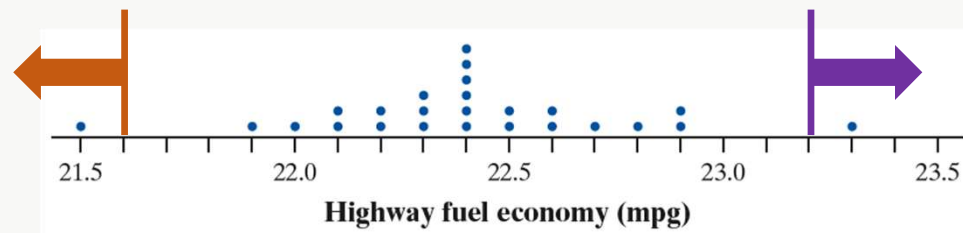


Identifying Outliers (Method 1)

Highway fuel economy ratings for twenty five 2018 Toyota 4Runners tested by the EPA:

21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

$Q_1 = 22.2$ mpg
 $Q_3 = 22.6$ mpg
 $IQR = 0.4$ mpg



Low outliers $< Q_1 - 1.5 \times IQR = 22.2 - 1.5 \times 0.4 = 21.6$

High outliers $> Q_3 + 1.5 \times IQR = 22.6 + 1.5 \times 0.4 = 23.2$

The cars with fuel economy ratings of 21.5 mpg and 23.3 mpg would be considered outliers by the $1.5 \times IQR$ rule.

Identifying Outliers (Method 2)

- An outlier is a value located more than 2 standard deviations above, or below, the mean.

0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	2	2	2
2	2	2	2	2	2	2	2	3	3
3	3	3	3	3	3	3	3	3	4
4	5	5	5	5	5	5	5	6	6
6	6	7	7	7	8	8	9	10	10
11	13	18	20	21	22	29	42	42	104

n	mean	SD	min	Q1	med	Q3	max
70	7.34	14.45	0	2	3	7	104

Low outlier $< \text{mean} - 2 \times \text{SD}$
 $< 7.34 - 2 \times 14.45$
 < -21.56

No low outliers!

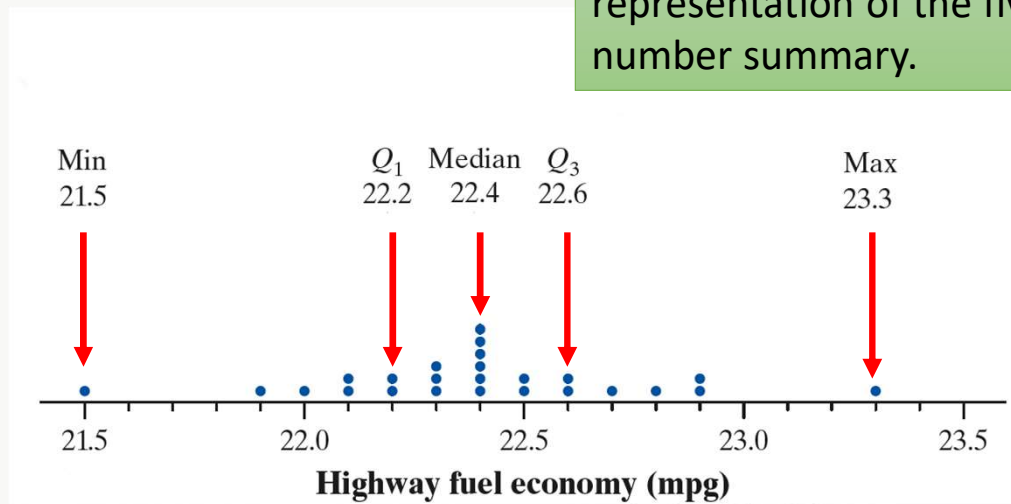
High outlier $> \text{mean} + 2 \times \text{SD}$
 $> 7.34 + 2 \times 14.45$
 > 36.24

42, 42, and 104 are outliers!

Making and Interpreting Boxplots

The **five-number summary** of a distribution of quantitative data consists of the minimum, the first quartile Q_1 , the median, the third quartile Q_3 , and the maximum.

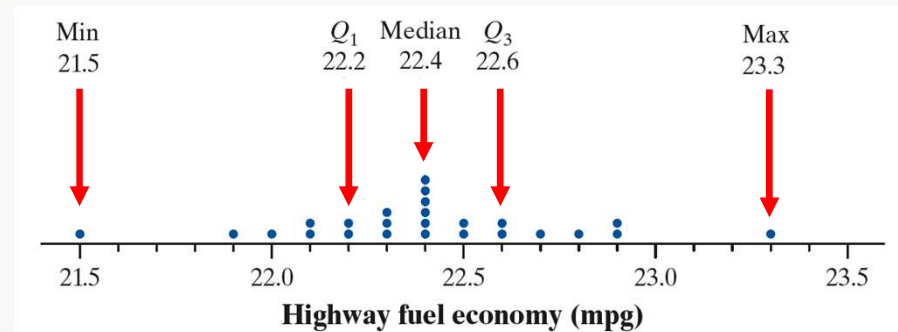
A **boxplot** is a visual representation of the five-number summary.



Making and Interpreting Boxplots

How to Make a Boxplot

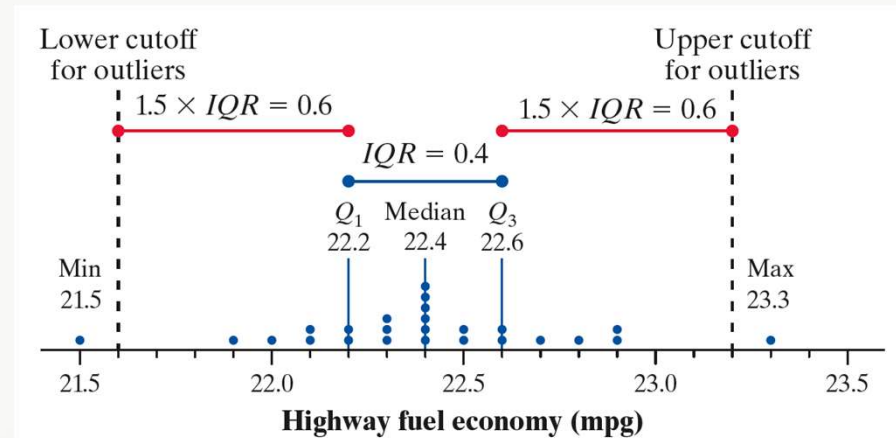
- Find the five-number summary.
- Identify outliers using the $1.5 \times \text{IQR}$ rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.



Making and Interpreting Boxplots

How to Make a Boxplot

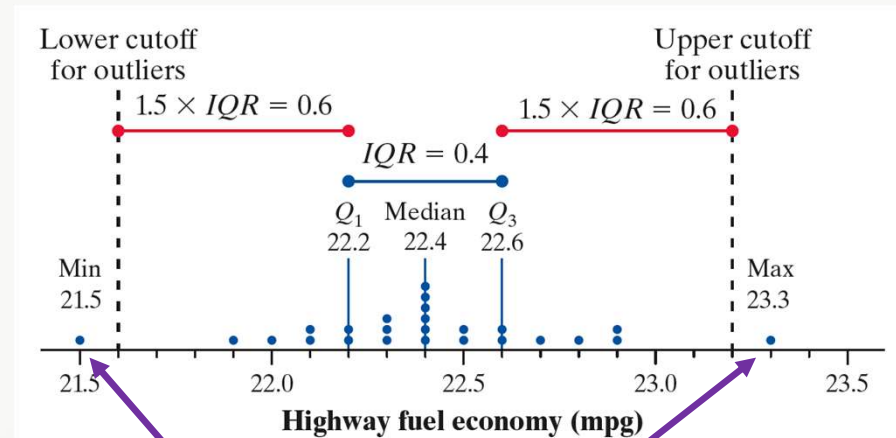
- Find the five-number summary.
- Identify outliers using the $1.5 \times IQR$ rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.



Making and Interpreting Boxplots

How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the $1.5 \times IQR$ rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.



Outliers

Making and Interpreting Boxplots

How to Make a Boxplot

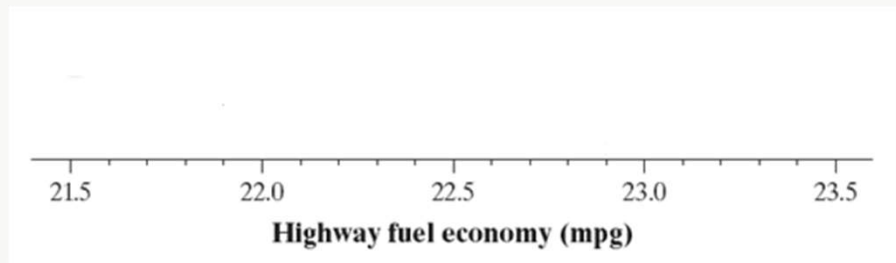
- Find the five-number summary.
- Identify outliers using the $1.5 \times \text{IQR}$ rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.



Making and Interpreting Boxplots

How to Make a Boxplot

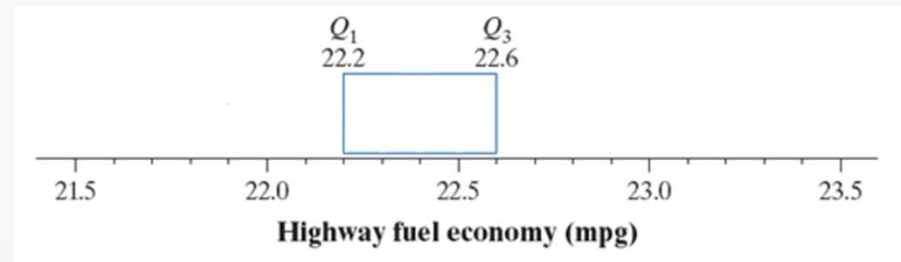
- Find the five-number summary.
- Identify outliers using the $1.5 \times \text{IQR}$ rule.
- Draw and label the horizontal axis.
- **Scale the axis.**
- Draw a box.
- Mark the median.
- Draw whiskers.



Making and Interpreting Boxplots

How to Make a Boxplot

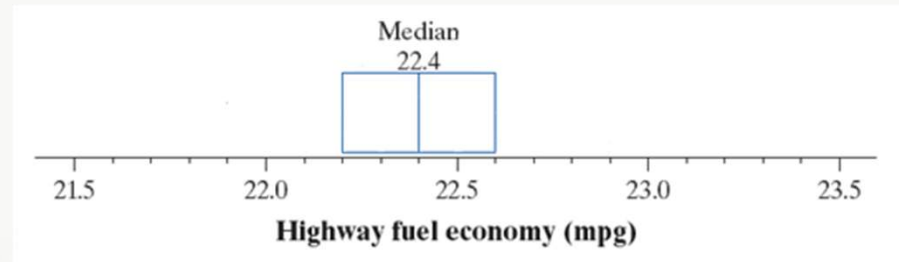
- Find the five-number summary.
- Identify outliers using the $1.5 \times \text{IQR}$ rule.
- Draw and label the horizontal axis.
- Scale the axis.
- **Draw a box.**
- Mark the median.
- Draw whiskers.



Making and Interpreting Boxplots

How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the $1.5 \times \text{IQR}$ rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- **Mark the median.**
- Draw whiskers.

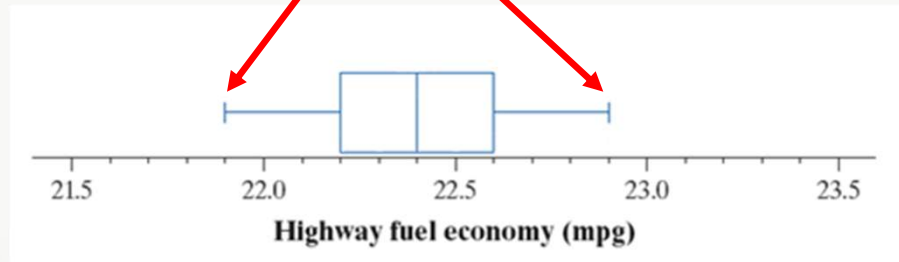


Making and Interpreting Boxplots

How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the $1.5 \times \text{IQR}$ rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

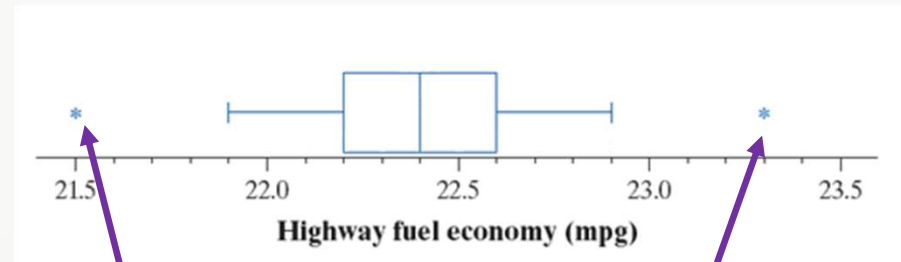
Whiskers extend to last data value that isn't an outlier



Making and Interpreting Boxplots

How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the $1.5 \times \text{IQR}$ rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.



Mark outliers as separate points

Making and Interpreting Boxplots

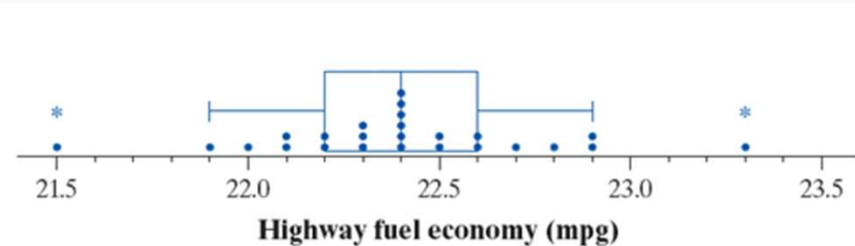
How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the $1.5 \times \text{IQR}$ rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.



CAUTION:

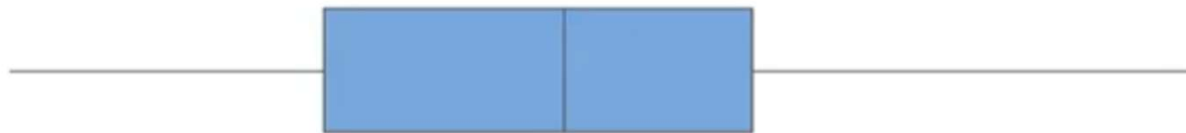
- Boxplots do not display each individual value in a distribution.
- Boxplots don't show gaps, clusters, or peaks.



Left-skewed

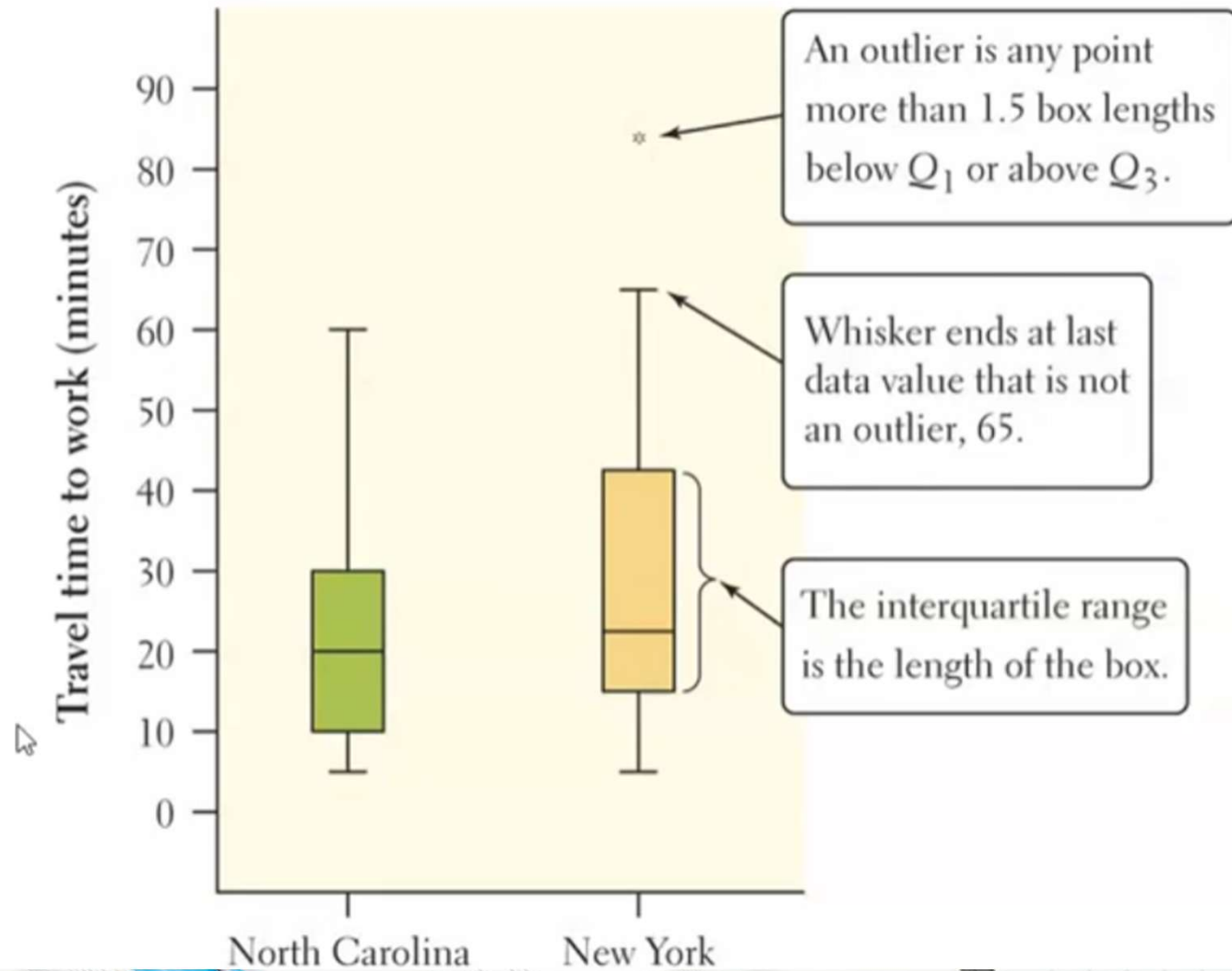


Symmetric



Right-skewed





Transforming Data

Transforming converts the original observations from the original units of measurements to another scale. Transformations can affect the shape, center, and spread of a distribution.

The Effect of Adding or Subtracting a Constant

Adding the same positive number a to (subtracting a from) each observation:

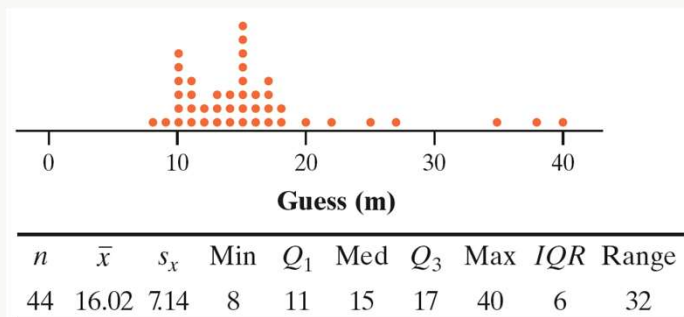
- Adds a to (subtracts a from) measures of center and location (mean, five-number summary, percentiles)
- Does not change measures of variability (range, IQR, standard deviation)
- Does not change the shape of the distribution



Transforming Data

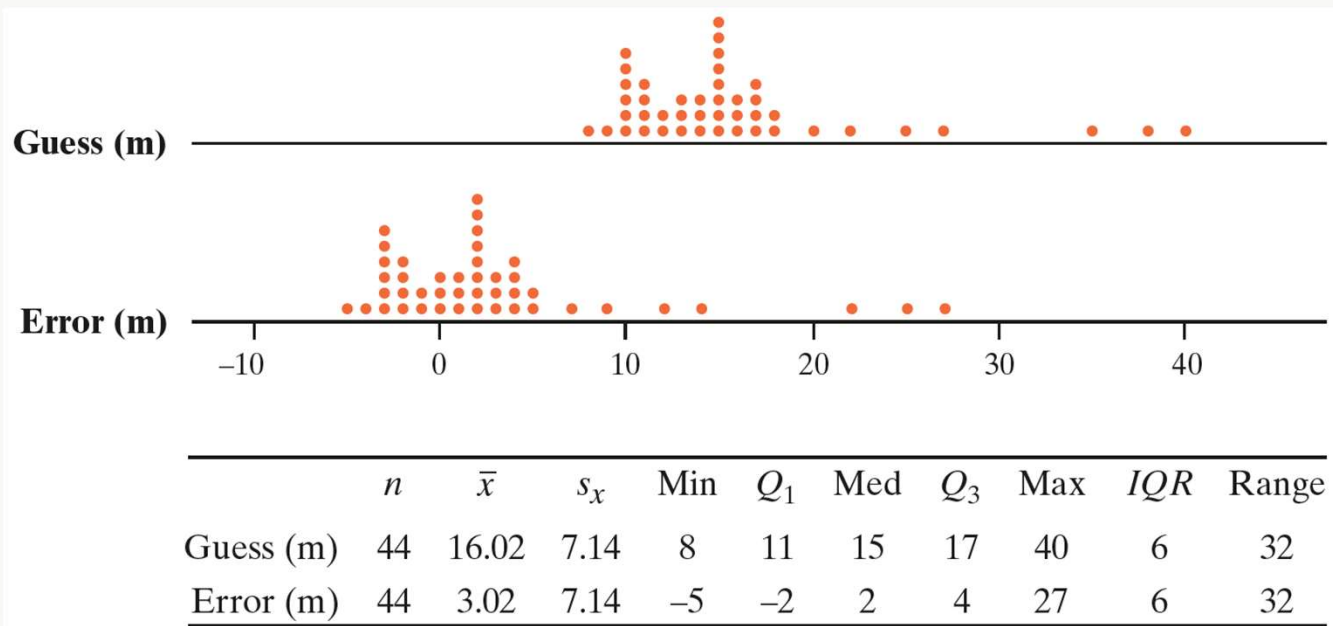
Soon after the metric system was introduced in Australia, a group of students was asked to guess the width of their classroom to the nearest meter. The actual width of the room was 13 meters. We can examine the distribution of students' errors by defining a new variable as follows: $error = guess - 13$.

- What shape does the distribution of $error$ have?
- Find the mean and the median of the distribution of $error$.
- Find the standard deviation and interquartile range (IQR) of the distribution of $error$.



- The same shape as the original distribution of guesses: skewed to the right with two distinct peaks.*
- Mean: $16.02 - 13 = 3.02$ meters;
Median: $15 - 13 = 2$ meters.*
- Standard deviation = 7.14 meters;
 $IQR = 6$ meters*

Transforming Data



Transforming Data

Transforming converts the original observations from the original units of measurements to another scale. Transformations can affect the shape, center, and spread of a distribution.

The Effect of Multiplying or Dividing by a Constant

Multiplying (or dividing) each observation by the same positive number b :

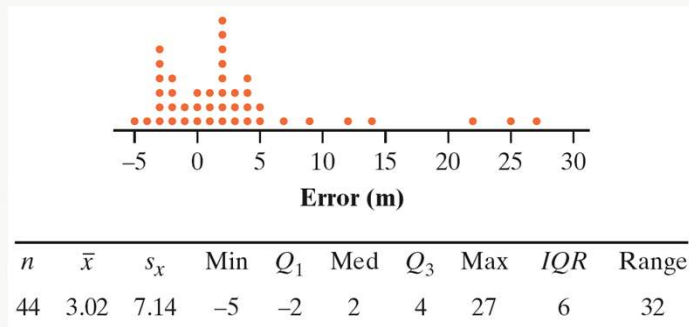
- Multiplies (divides) measures of center and location by b
(mean, five-number summary, percentiles) by b
- Multiplies (divides) measures of variability by b
(range, *IQR*, standard deviation)
- Does not change the shape of the distribution



Transforming Data

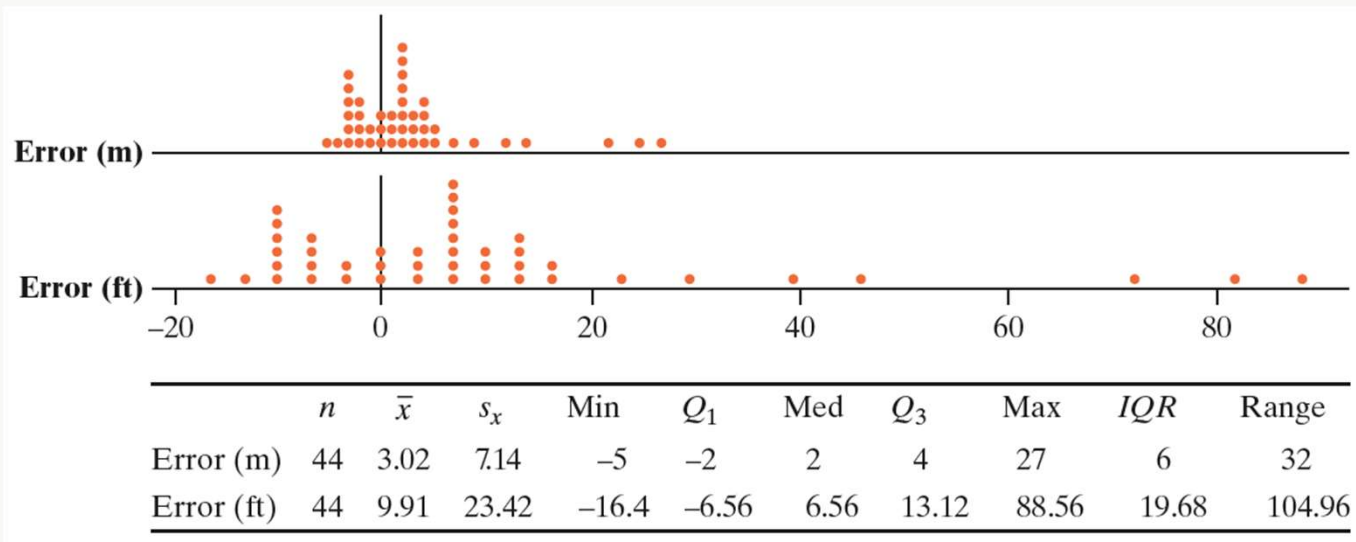
Because the students are having some difficulty with the metric system, it may not be helpful to tell them that their guesses tended to be about 2 meters too large. Let's convert the *error* data to feet before we report back to them. To convert from meters to feet, we multiply each of the error values by 3.28.

- (a) What shape does the resulting distribution of error have?
- (b) Find the median of the distribution of error in feet.
- (c) Find the interquartile range (IQR) of the distribution of error in feet.



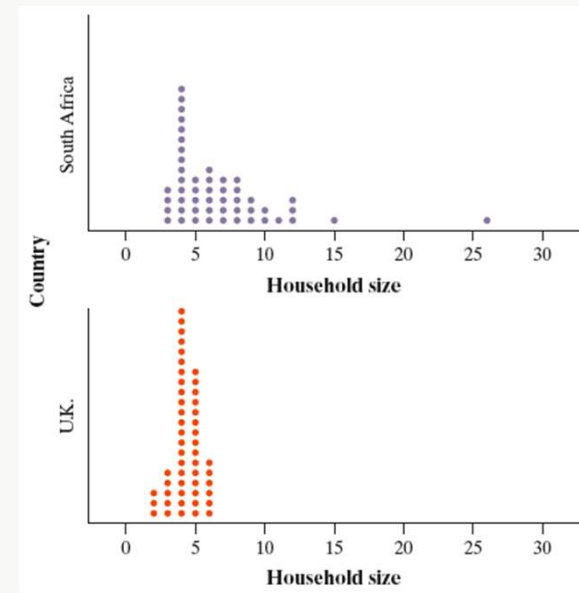
- a) *The same shape as the original distribution of guesses: skewed to the right with two distinct peaks.*
- b) *Median = $2 \times 3.28 = 6.56$ feet*
- c) *IQR = $6 \times 3.28 = 19.68$ feet*

Transforming Data



Comparing Distributions

We used Census At School's "Random Data Selector" to choose 50 students from each country. Here are dotplots of the household sizes reported by the survey respondents. Compare the distributions of household size for these two countries.



AP® Exam Tip

When comparing distributions of quantitative data, it's not enough just to list values for the center and variability of each distribution. You must explicitly compare these values, using words like "greater than," "less than," or "about the same as."

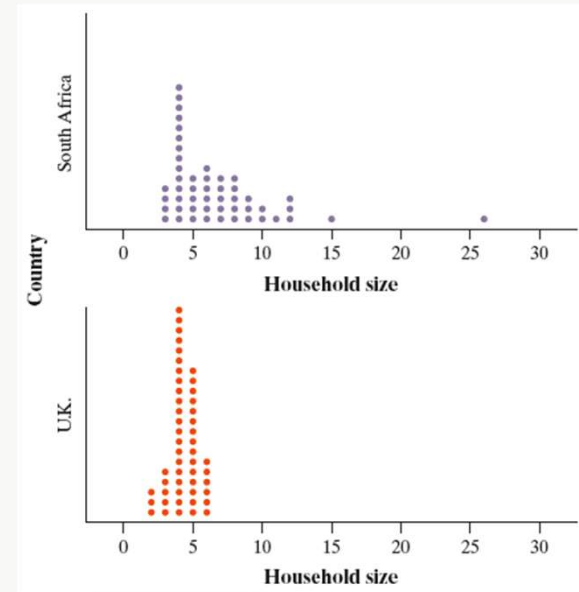
Comparing Distributions

Shape: The distribution of household size for the U.K. sample is roughly symmetric, with a single peak at 4 people. The distribution of household size for the South Africa sample is skewed to the right, with a single peak at 4 people and a clear gap between 15 and 26.

Outliers: There don't appear to be any outliers in the U.K. distribution. The South African distribution seems to have two outliers: the households with 15 and 26 people.

Center: Household sizes for the South African students tend to be larger (median 56 people) than for the U.K. students (median 54 people).

Spread: The household sizes for the South African students vary more (from 3 to 26 people) than for the U.K. students (from 2 to 6 people).



- ✓ Context
- ✓ Comparative language