

Examining Data

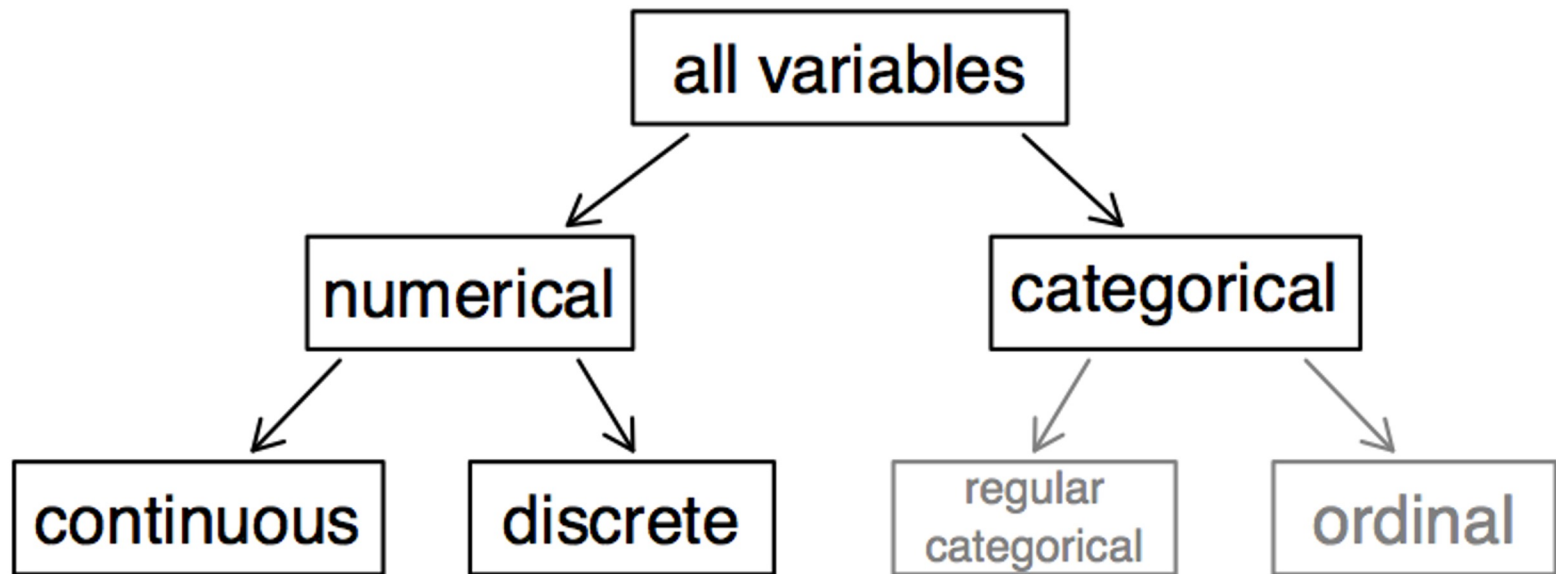
Recap

- Statistics is the science of making inferences based on data
- Data collection tries to take a representative sample of the population of interest
- Sampling is hard
 - Sampling bias results in a sample that is non-representative of our population and leads to wrong inferences
- Causal claims require experiments, observational studies only allow inferences about correlations

Data

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

Types of variables



Types of variables

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- **gender**: *categorical*
- **sleep**: *numerical, continuous*
- **bedtime**: *categorical, ordinal*
- **countries**: *numerical, discrete*
- **dread**: *categorical, ordinal - could also be used as numerical*

Let's collect our own data

- Who is more popular: Actors or Athletes?
- What type of data can we use?

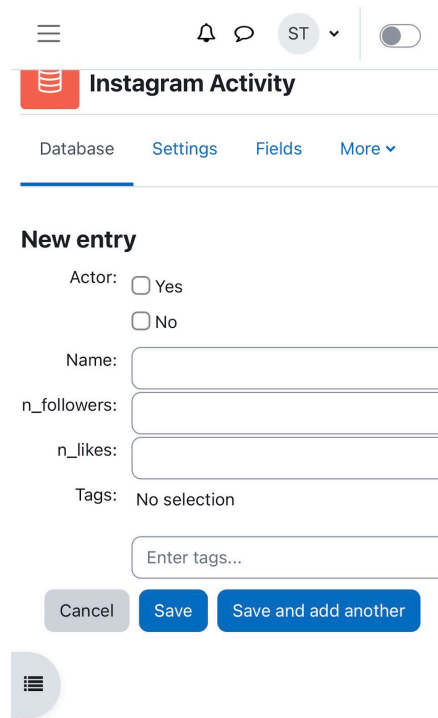
To Instagram!

- You will be split in two groups
- Left-side: Pick an Actor
- Right-Side: Pick an Athlete

- Go to their instagram page
- Write down:
 - Number of followers (in million)
 - Number of likes in their second post (in thousands)
 - Why?

Report your entry in the class website

- This week -> Instagram Activity-> Add new entry.



The screenshot shows a mobile application interface for 'Instagram Activity'. At the top, there is a navigation bar with a hamburger menu icon, a notification bell, a speech bubble, a user profile icon labeled 'ST', and a toggle switch. Below the navigation bar, the title 'Instagram Activity' is displayed with a red icon. Underneath the title, there are four tabs: 'Database', 'Settings', 'Fields', and 'More'. The 'New entry' section contains the following fields and options:

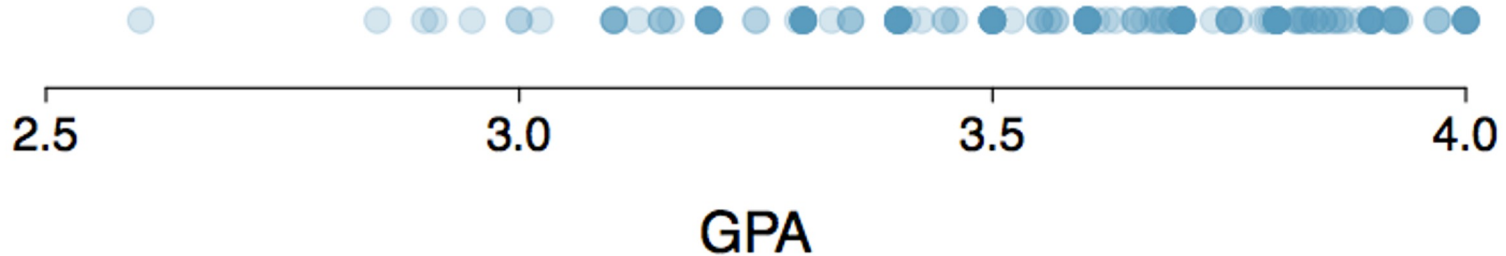
- Actor: Yes No
- Name:
- n_followers:
- n_likes:
- Tags: No selection
- Enter tags...

At the bottom of the form, there are three buttons: 'Cancel', 'Save', and 'Save and add another'. A hamburger menu icon is also visible at the bottom left of the screen.

Examining Numerical Data

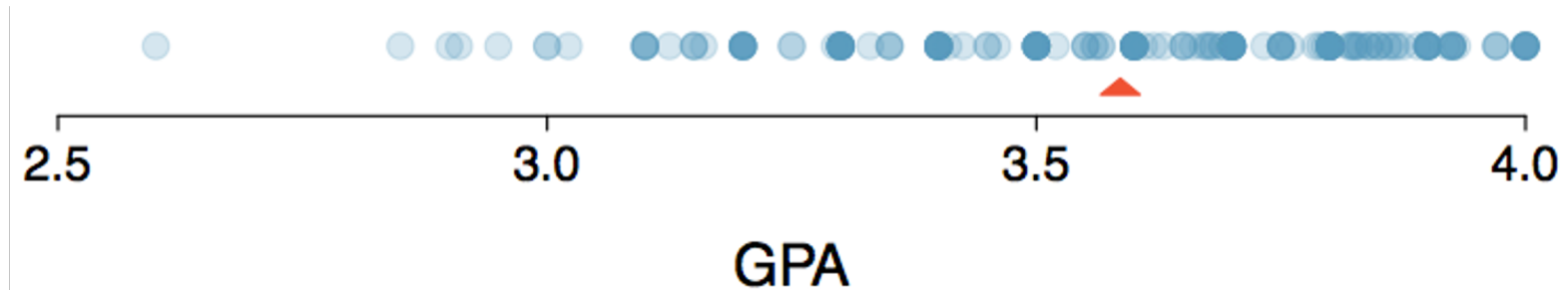
Dot Plots

Useful for visualizing one numerical variable. Darker colors represent areas where there are more observations.



How would you describe the distribution of GPAs in this data set? Make sure to say something about the center, shape, and spread of the distribution.

Dot Plots & Mean



The *mean*, also called the *average* (marked with a triangle in the above plot), is one way to measure the center of a *distribution* of data.

The mean GPA is 3.59.

Mean

The *sample mean*, denoted as \bar{x} , can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

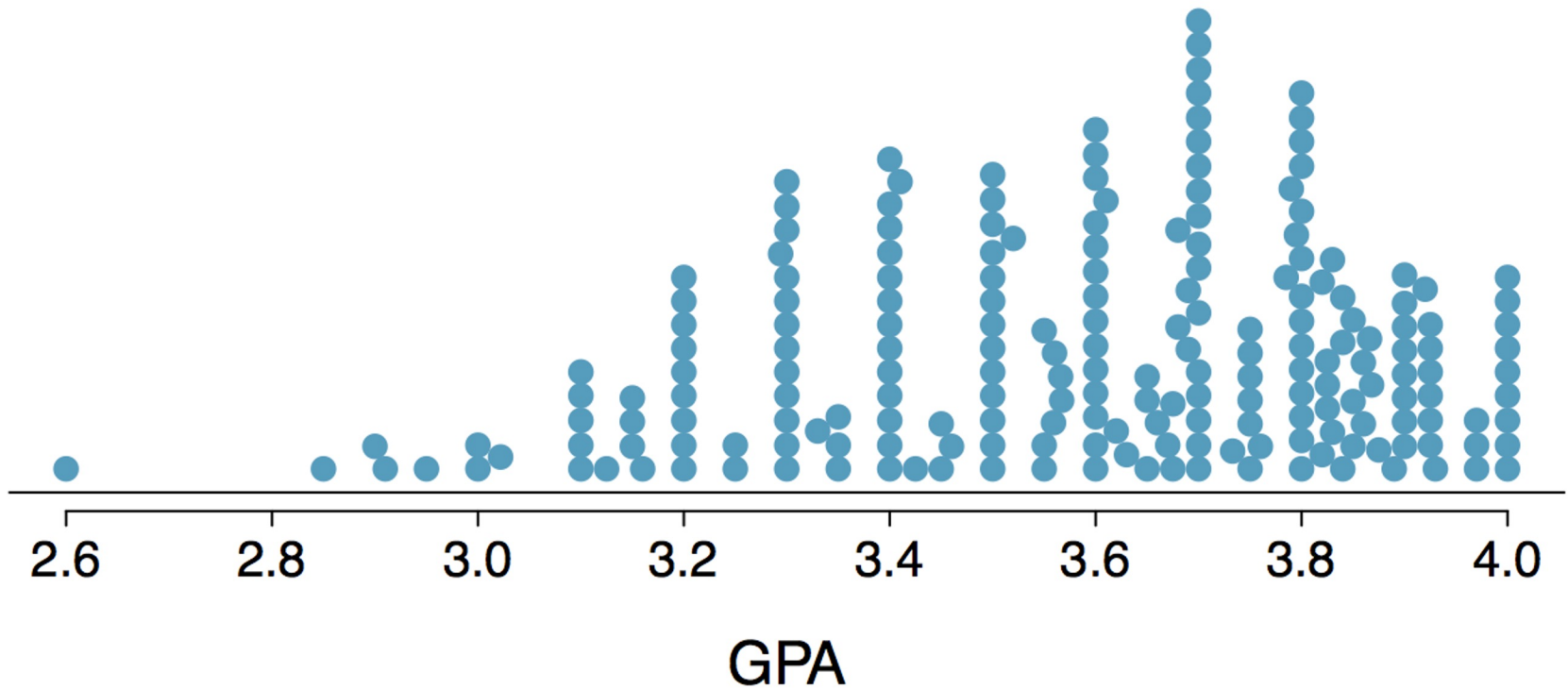
where x_1, x_2, \dots, x_n represent the n observed values.

The *population mean* is also computed the same way but is denoted as μ . It is often not possible to calculate μ since population data are rarely available.

The sample mean is a *sample statistic*, and serves as a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

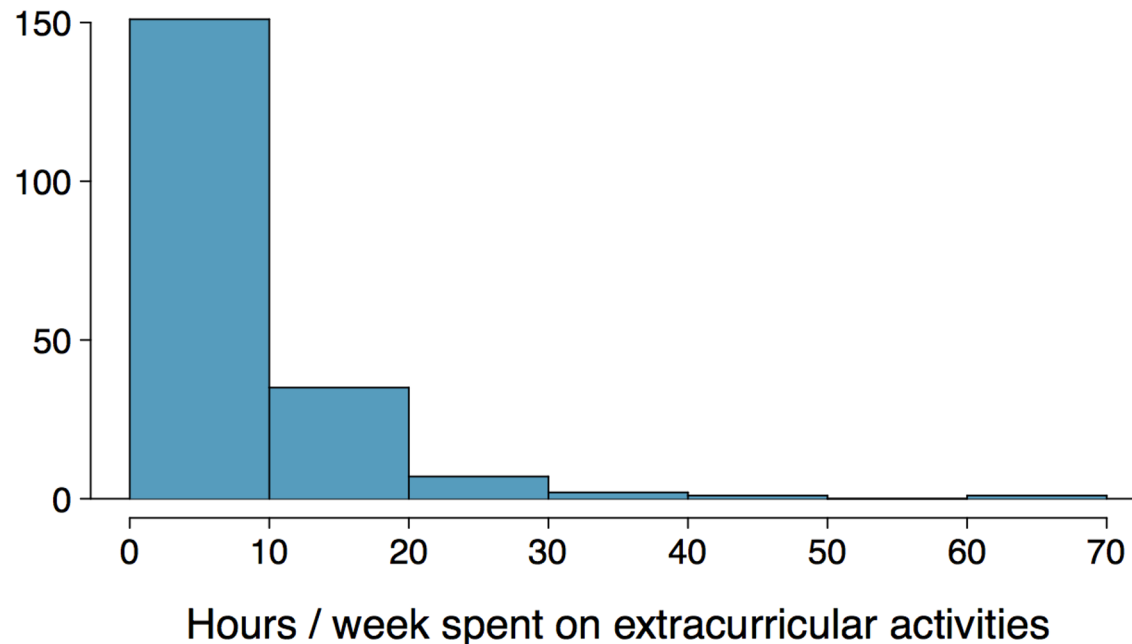
Stacked Dot Plot

Higher bars represent areas where there are more observations, makes it a little easier to judge the center and the shape of the distribution.



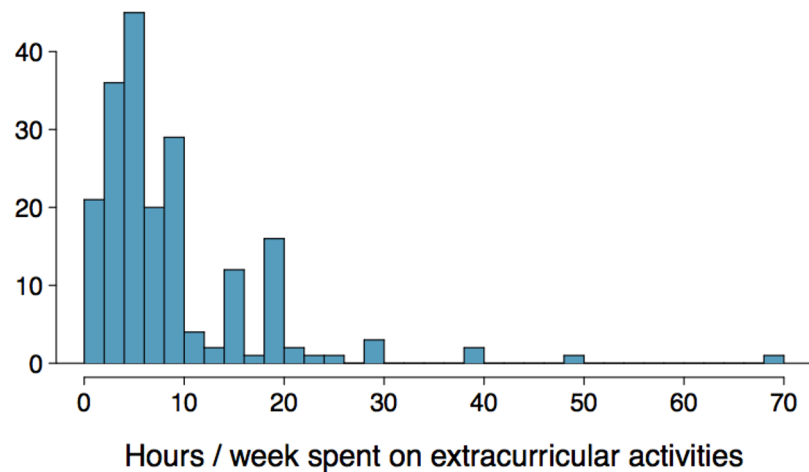
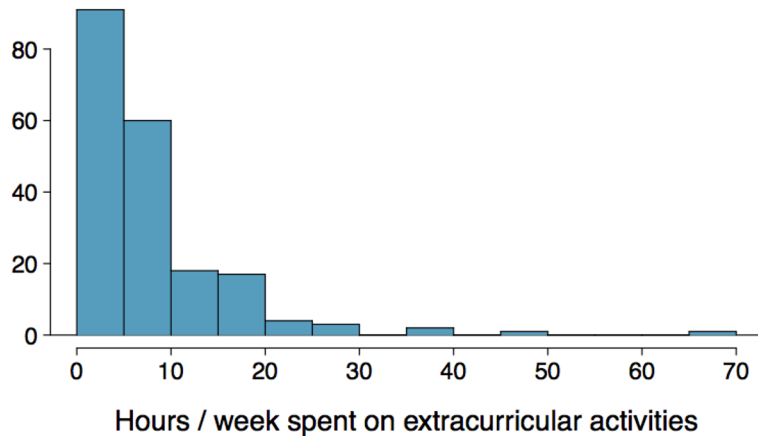
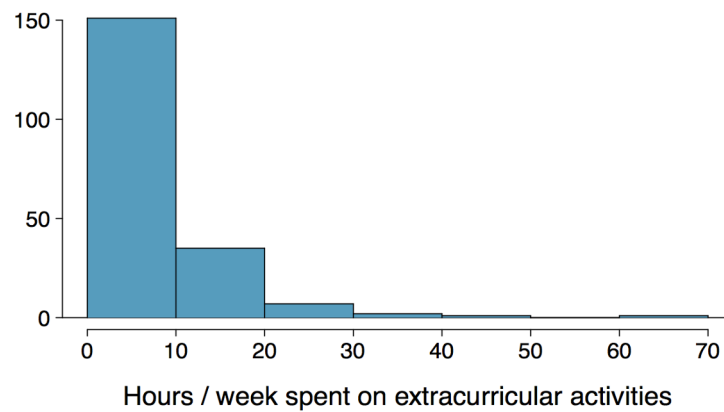
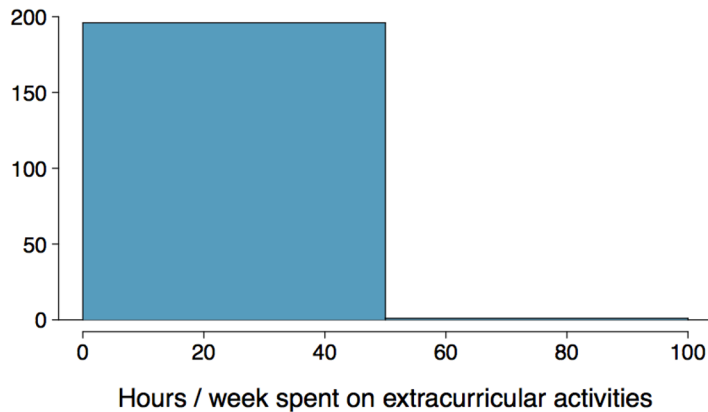
Histograms - Extracurricular Hours

- Histograms provide a view of the *data density*. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the *shape* of the data distribution.
- The chosen *bin width* can alter the story the histogram is telling.



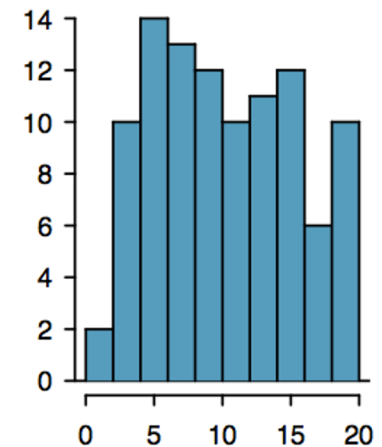
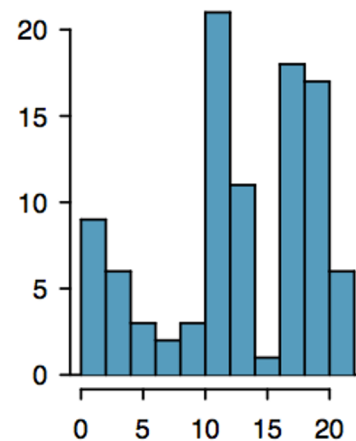
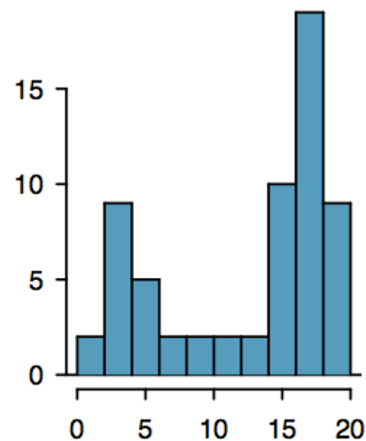
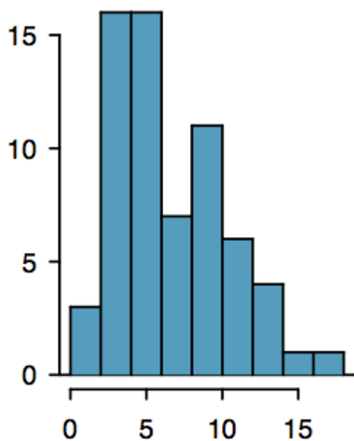
Bin Width

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



Shape of a Distribution: Modality

Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?

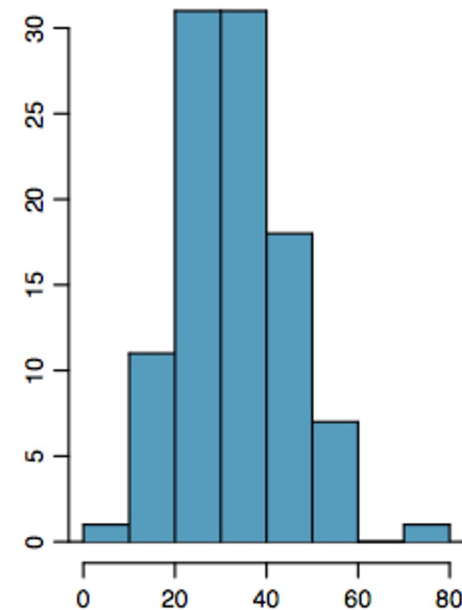
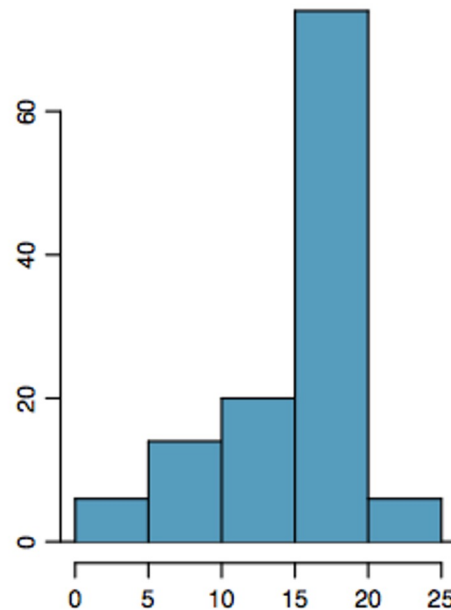
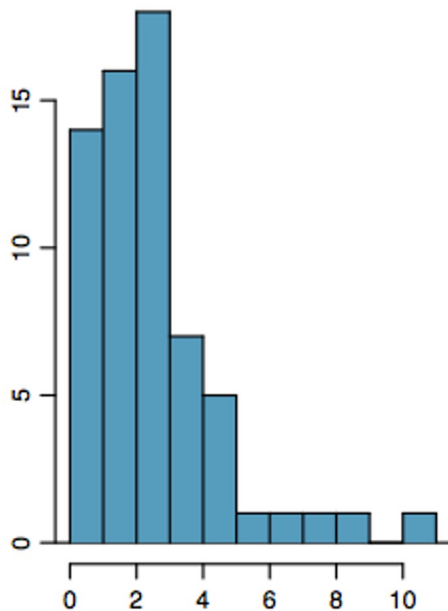


Note: In order to determine modality, step back and imagine a smooth curve over the histogram.

Mode: The most frequent value of a distribution.

Shape of a Distribution: Skewness

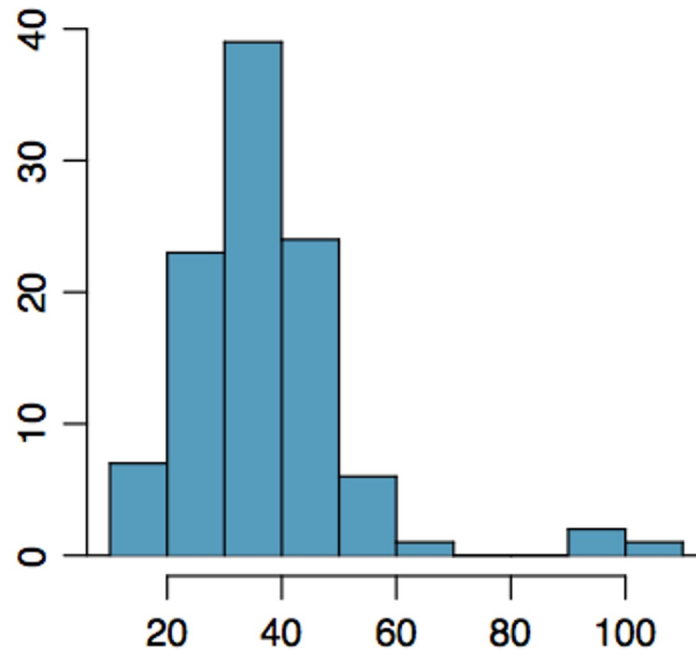
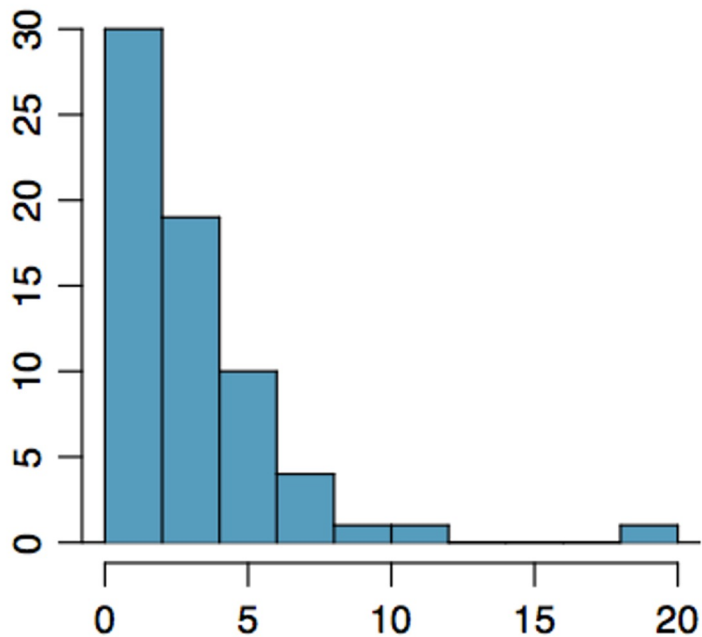
Is the histogram *right skewed*, *left skewed*, or *symmetric*?



Note: Histograms are said to be skewed to the side of the long tail.

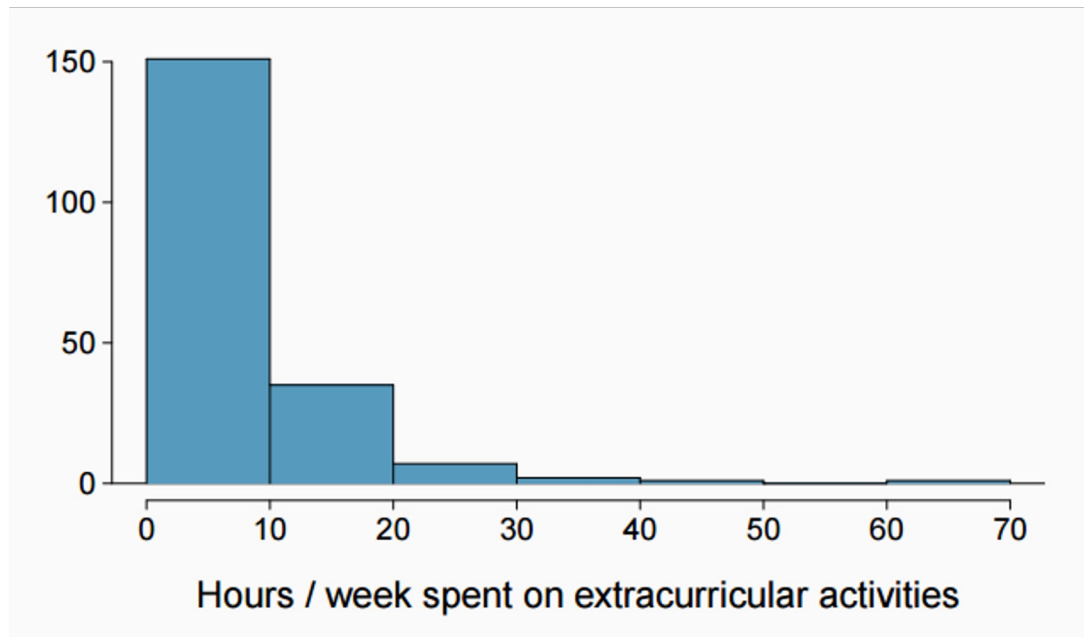
Shape of a Distribution: Unusual Observations

Are there any unusual observations or potential *outliers*?



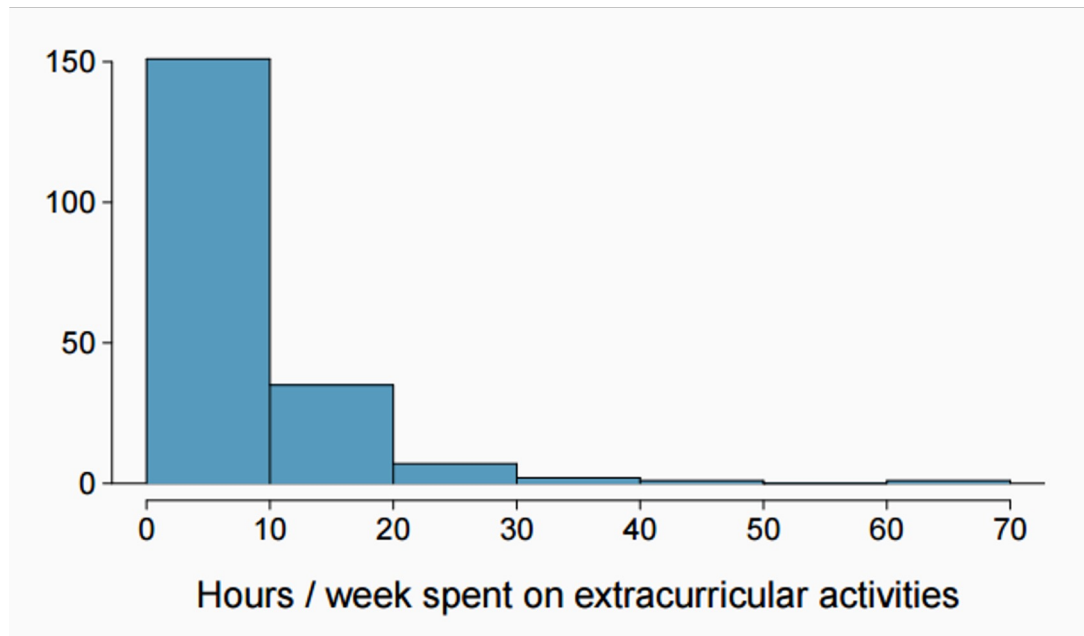
Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?

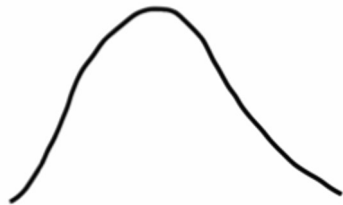


Unimodal and right skewed, with a potentially unusual observation at 60 hours/week.

Commonly observed shapes of distributions

Modality

unimodal



bimodal



multimodal



uniform



Skewness

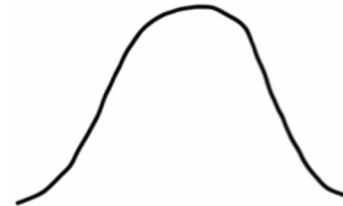
right skew



left skew



symmetric



Practice

Which of these variables do you expect to be uniformly distributed?

- (a) weights of adult females
- (b) salaries of a random sample of people from Crete
- (c) house prices
- (d) birthdays of classmates (day of the month)

Practice

Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

(b) salaries of a random sample of people from Crete

(c) house prices

(d) birthdays of classmates (day of the month)

Application Activity: Shapes of Distributions

Sketch the expected distributions of the following variables:

- number of piercings
- scores on an exam
- IQ scores

Come up with a concise way (1-2 sentences) to teach someone how to determine the expected distribution of any variable.

Summaries of Distributions

- Sample Mean: $\frac{x_1 + \dots + x_n}{n}$
- Sample Mode: The element that appears most frequently in x_1, \dots, x_n

Variance

Variance is roughly the average squared deviation from the mean.

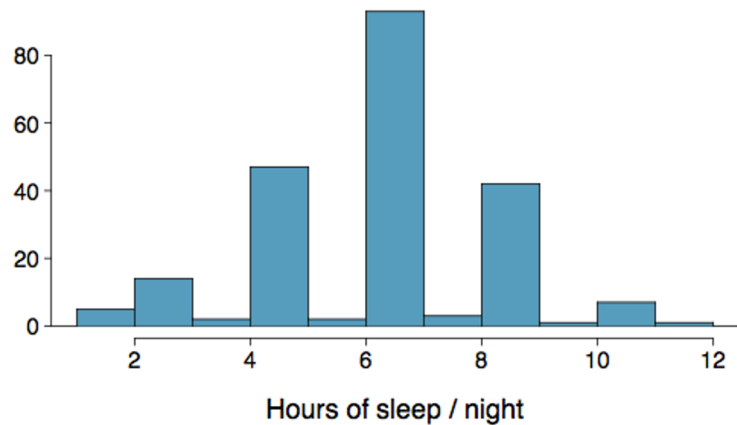
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Variance

Variance is roughly the average squared deviation from the mean.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.

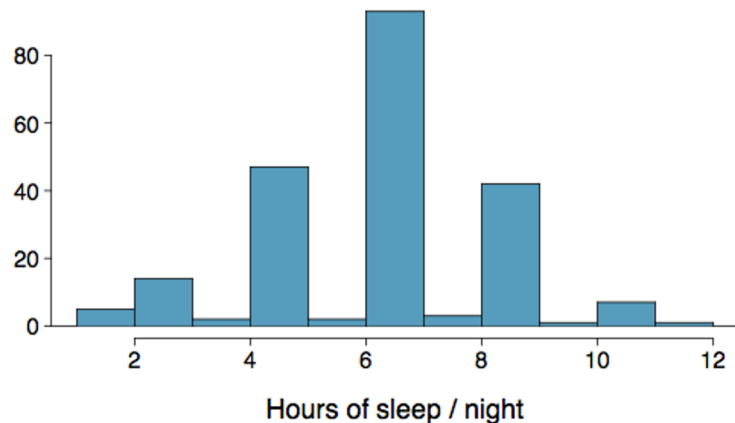


Variance

Variance is roughly the average squared deviation from the mean.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.
- The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5 - 6.71)^2 + (9 - 6.71)^2 + \dots + (7 - 6.71)^2}{217 - 1} = 4.11 \text{ hours}^2$$

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

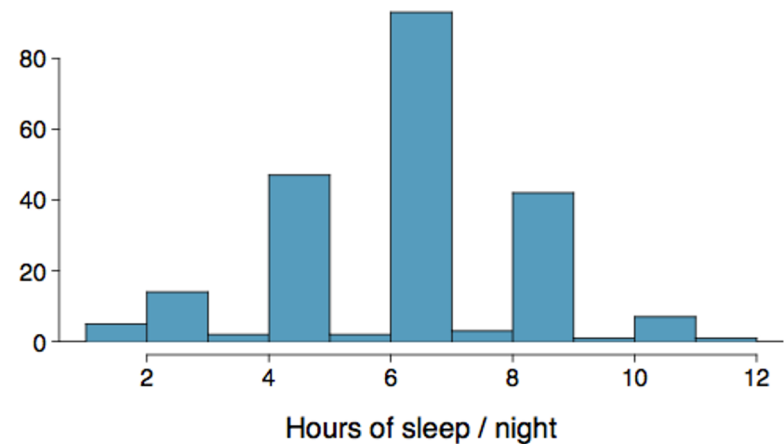
Standard Deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.

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

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \text{ hours}$$



- We can see that all of the data are within 3 standard deviations of the mean.

Summaries of Distributions

- Sample Mean: $\frac{x_1 + \dots + x_n}{n}$
- Sample Mode: The element that appears most frequently in x_1, \dots, x_n
- Sample Variance
- Sample Standard Deviation

Median

The *median* is the value that splits the data in half when ordered in ascending order.

0, 1, **2**, 3, 4

If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2}, \underline{3}, 4, 5 \rightarrow \frac{2 + 3}{2} = 2.5$$

Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the **50th percentile**.

Q1, Q3, and IQR

- The 25th percentile is also called the first quartile, *Q1*.
- The 50th percentile is also called the median.
- The 75th percentile is also called the third quartile, *Q3*.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the *interquartile range*, or the *IQR*.

$$IQR = Q3 - Q1$$

Q1, Q3, and IQR

- The 25th percentile is also called the first quartile, *Q1*.
- The 50th percentile is also called the median.
- The 75th percentile is also called the third quartile, *Q3*.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the *interquartile range*, or the *IQR*.

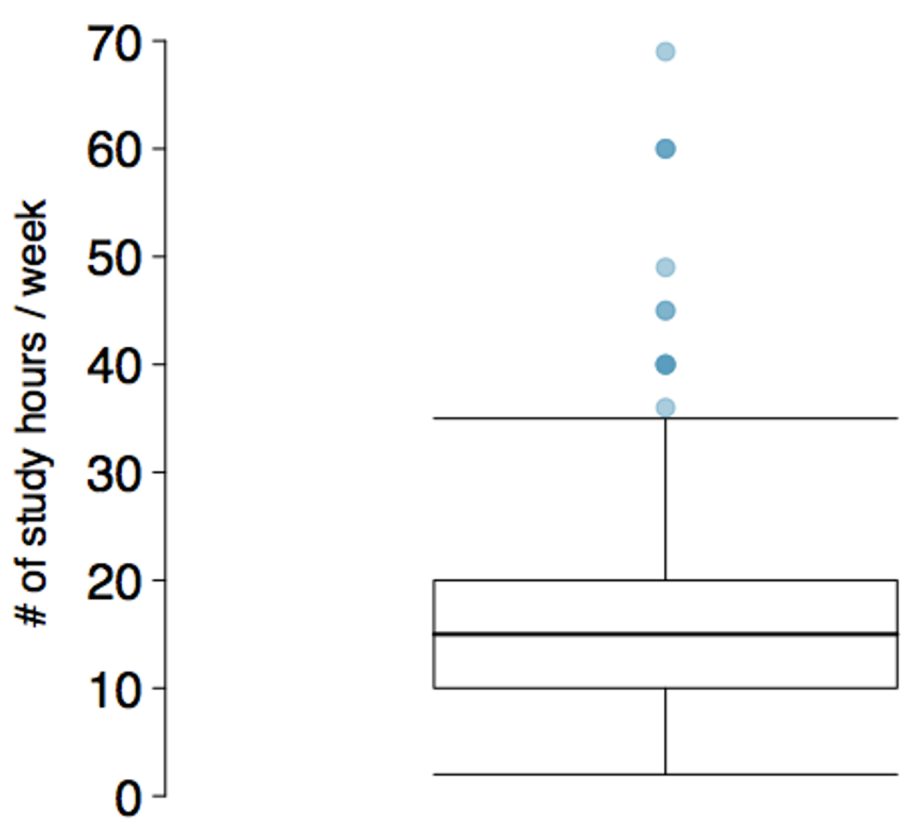
$$IQR = Q3 - Q1$$

Example: Age of retirement, 11 samples:

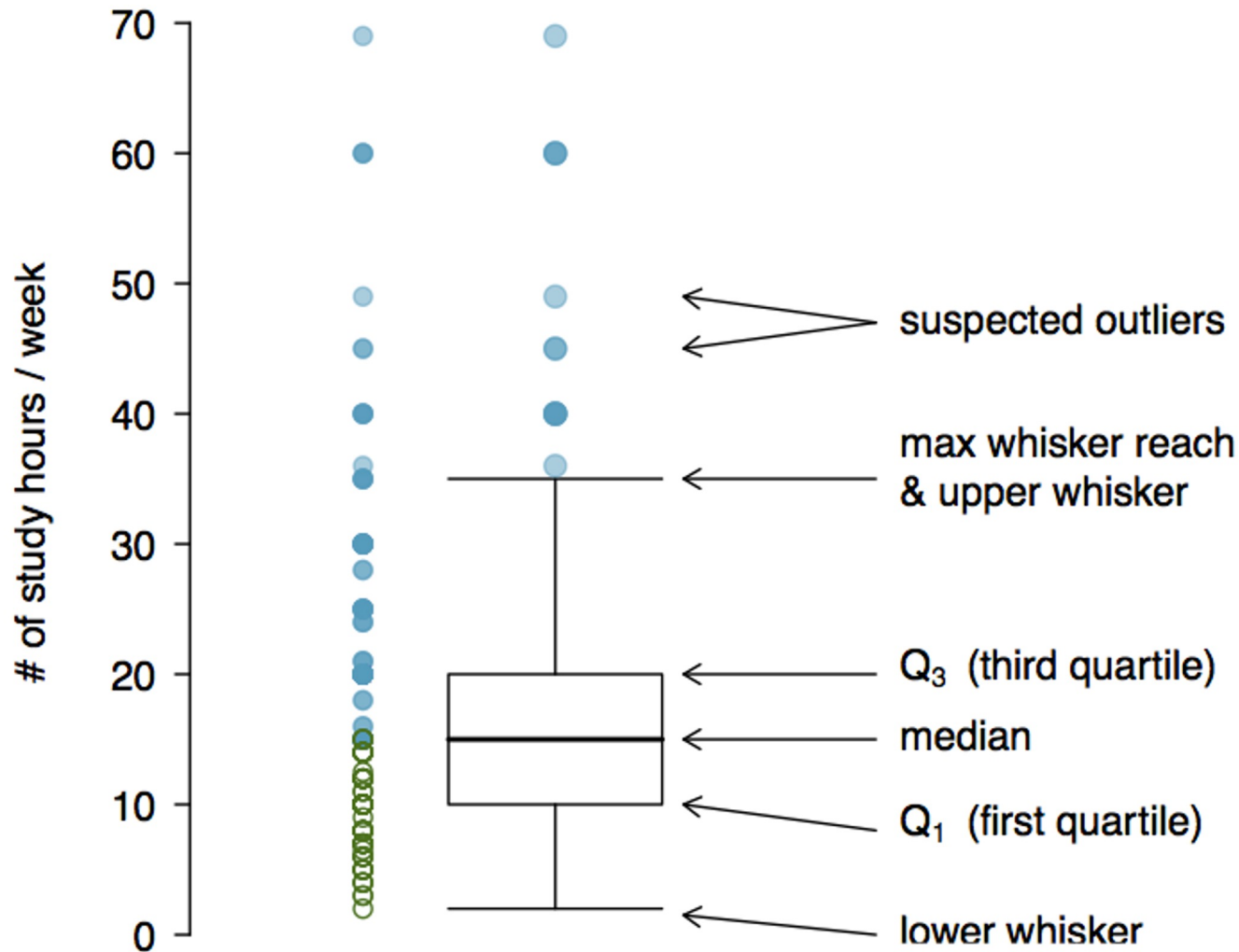
60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63

Box Plot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.



Anatomy of a Box Plot



Whiskers and Outliers

Whiskers of a box plot can extend up to $1.5 \times \text{IQR}$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times \text{IQR}$$

$$\text{max lower whisker reach} = Q1 - 1.5 \times \text{IQR}$$

$$\text{IQR: } 20 - 10 = 10$$

$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$

$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

Why is it important to look for outliers?

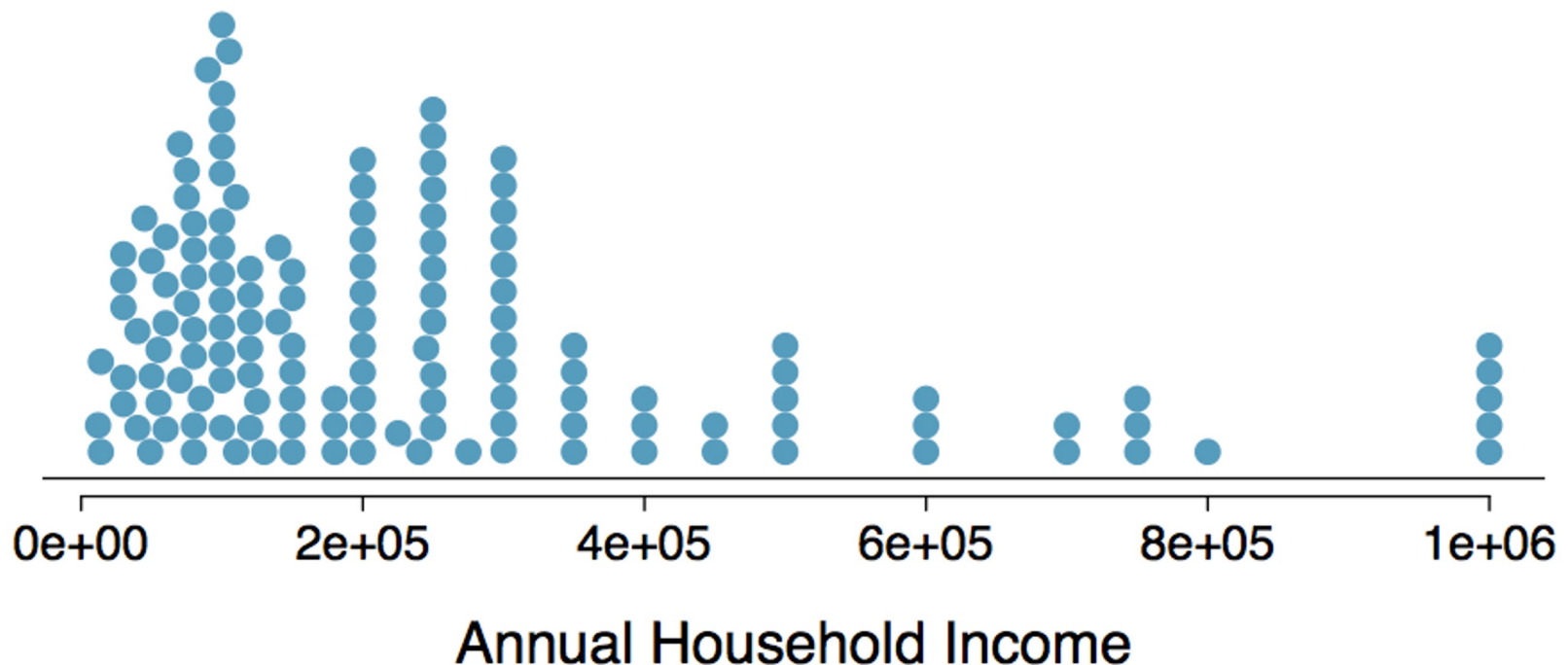
Outliers (cont.)

Why is it important to look for outliers?

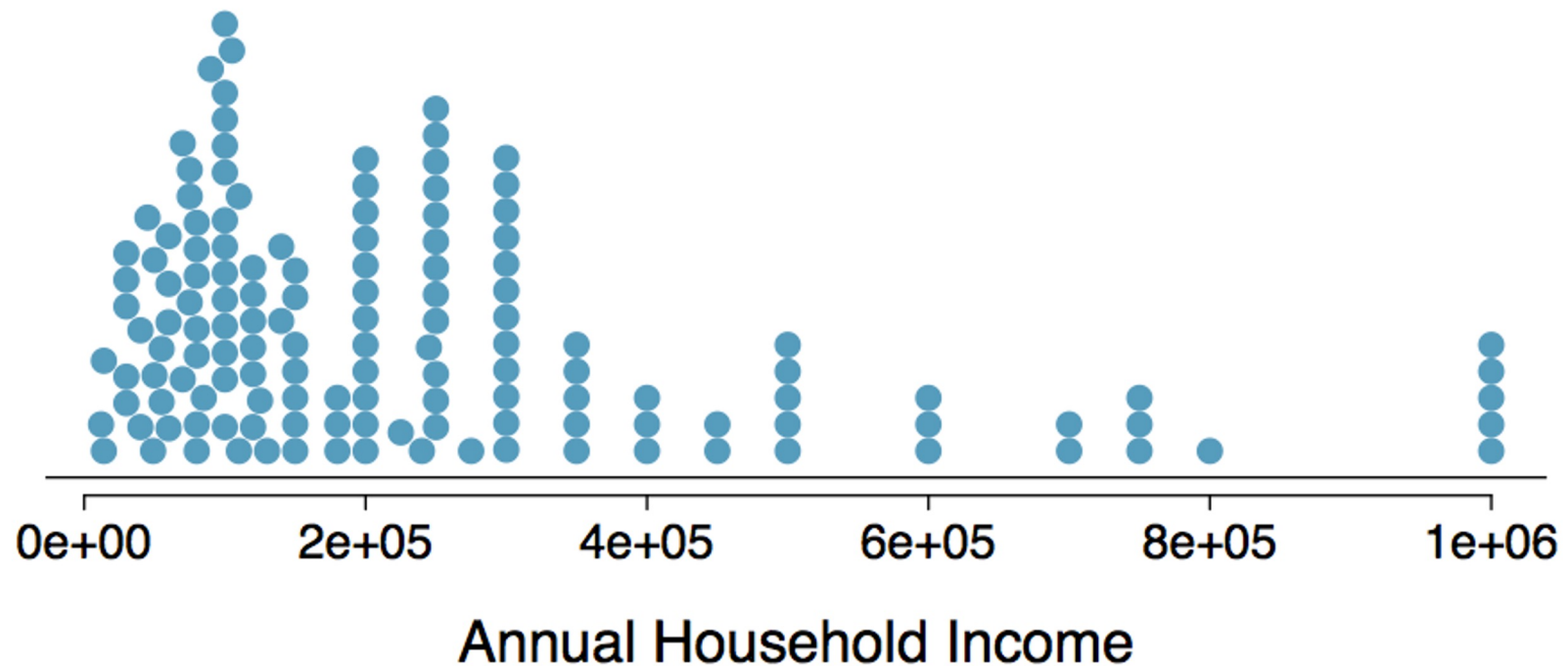
- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

Extreme Observations

How would sample statistics such as mean, median, SD, and IQR of household income be affected if the largest value was replaced with \$10 million? What if the smallest value was replaced with \$10 million?



Robust Statistics



scenario	robust		not robust	
	median	IQR	\bar{x}	s
original data	190K	200K	245K	226K
move largest to \$10 million	190K	200K	309K	853K
move smallest to \$10 million	200K	200K	316K	854K

Robust Statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

Robust Statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

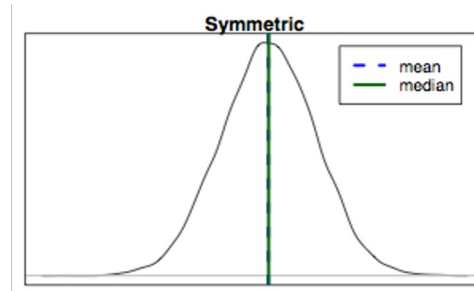
If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

Median

Mean vs. Median

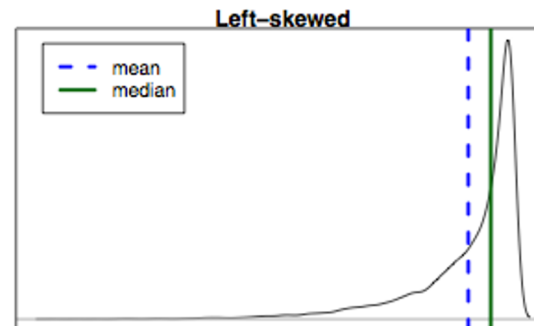
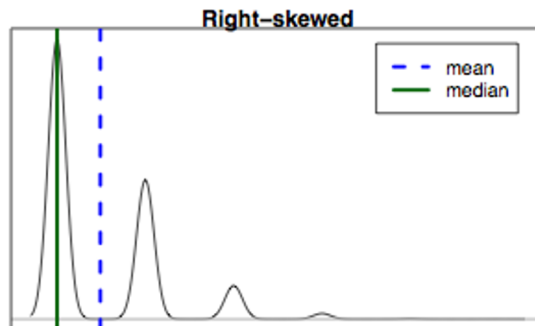
If the distribution is symmetric, center is often defined as the mean:

mean \sim median



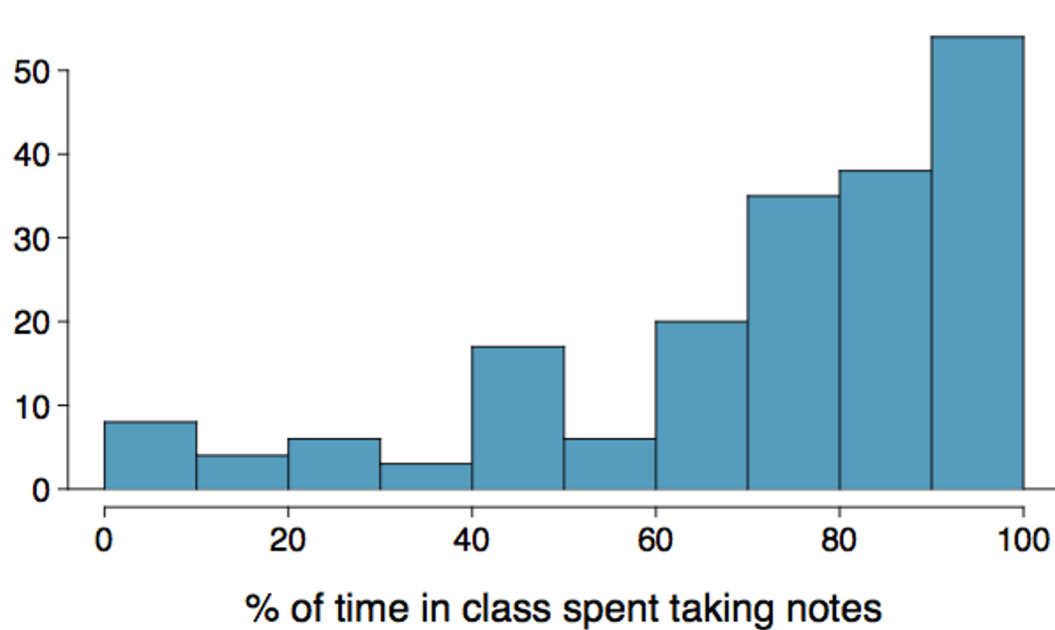
If the distribution is skewed or has extreme outliers, center is often defined as the median

- Right-skewed: mean $>$ median
- Left-skewed: mean $<$ median



Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



(a) mean > median

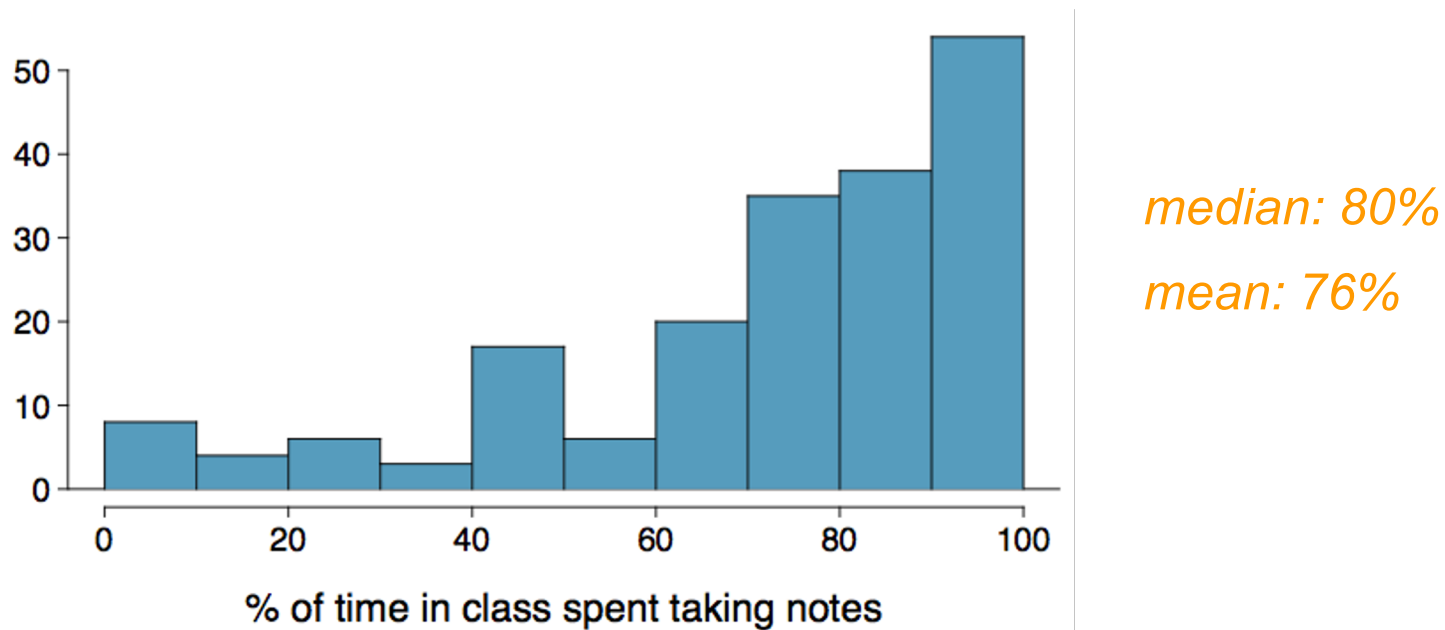
(c) mean < median

(b) mean ~ median

(d) impossible to tell

Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



(a) mean > median

(c) *mean < median*

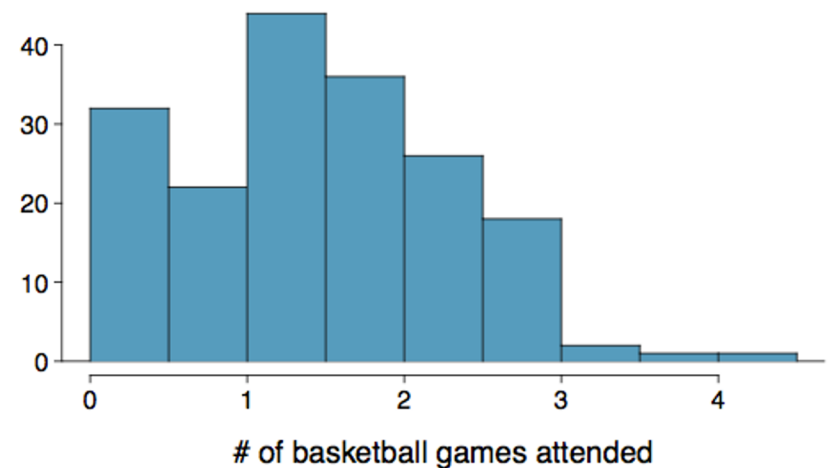
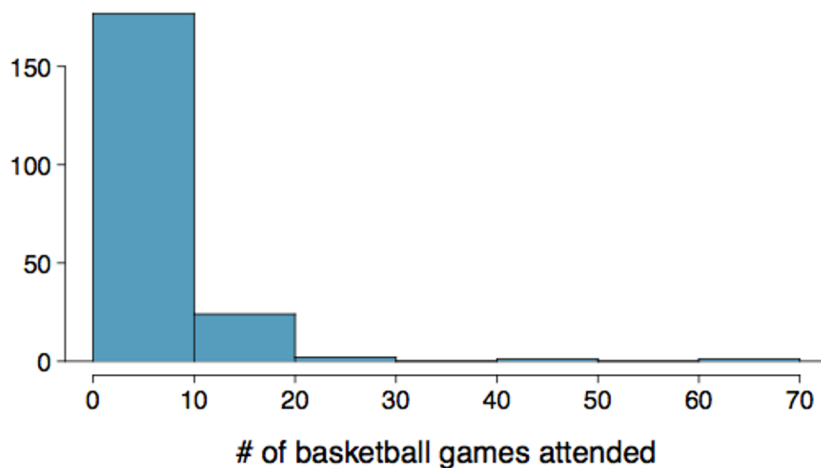
(b) mean ~ median

(d) impossible to tell

Extremely Skewed Data

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the [log transformation](#).

The histograms on the left shows the distribution of number of basketball games attended by students. The histogram on the right shows the distribution of log of number of games attended.



Pros and Cons of Transformations

- Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

# of games	70		50		25	...
# of games	4.25	3.91	3.22	...		

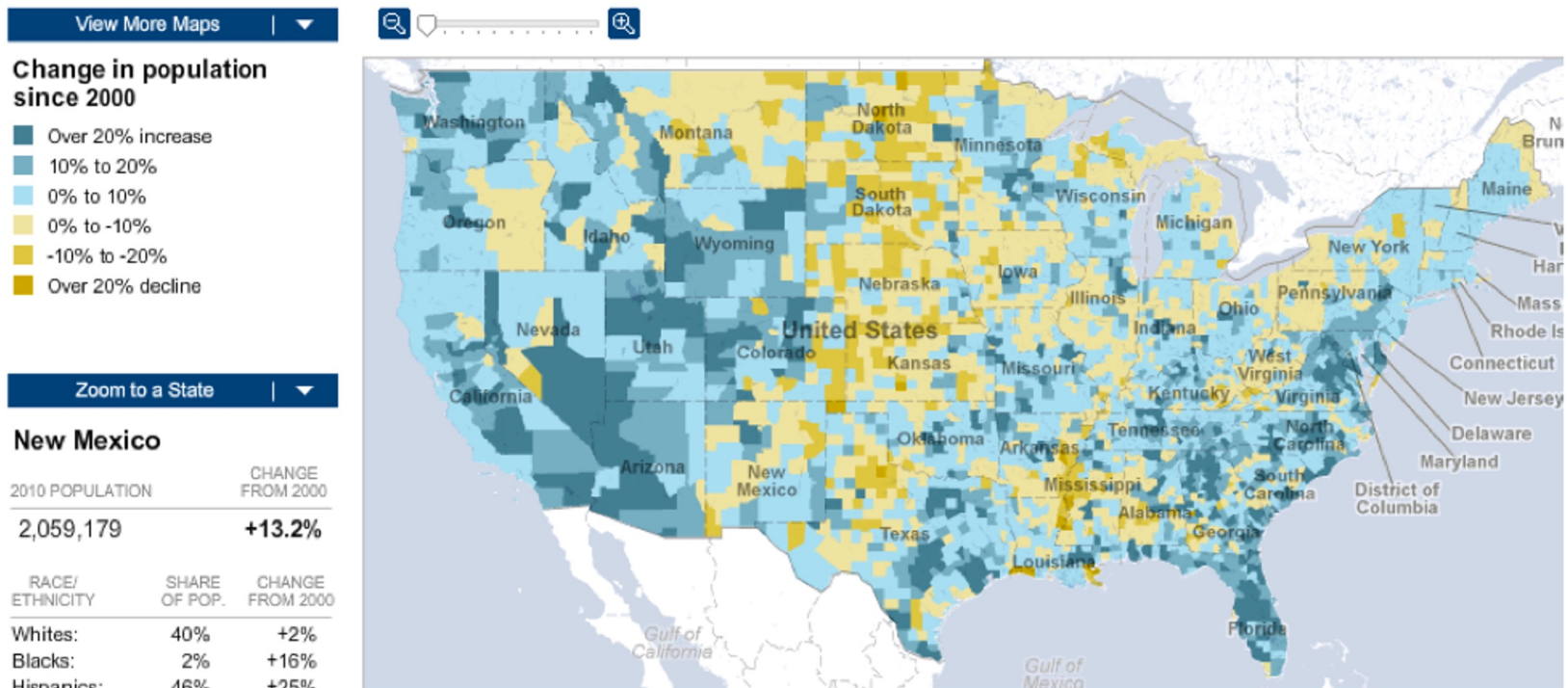
- However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

What other variables would you expect to be extremely skewed?

Salary, housing prices, etc.

Intensity Maps

What patterns are apparent in the change in population between 2000 and 2010?



<http://projects.nytimes.com/census/2010/map>

Scatterplot

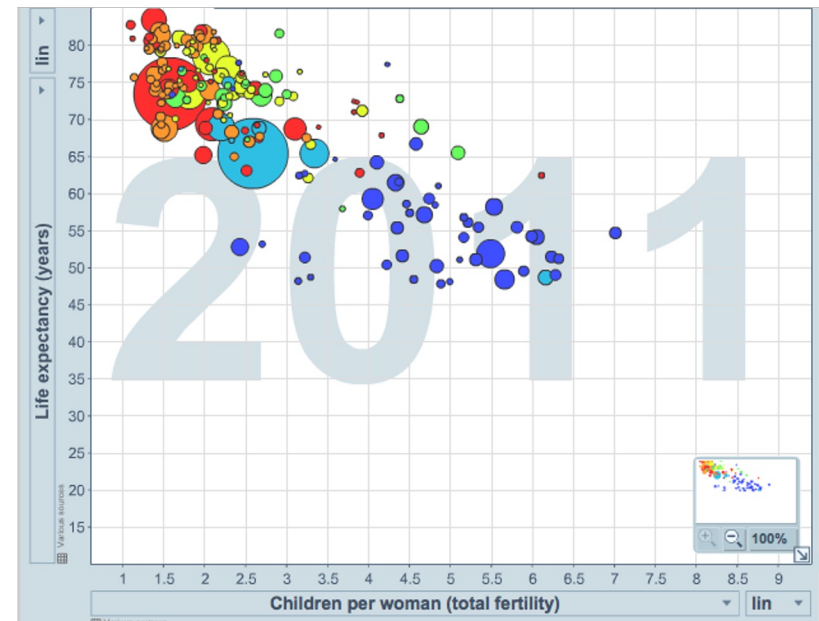
Scatterplots are useful for visualizing the relationship between two numerical variables.

Do life expectancy and total fertility appear to be *associated* or *independent*?

They appear to be linearly and negatively associated: as fertility increases, life expectancy decreases.

Was the relationship the same throughout the years, or did it change?

The relationship changed over the years.



<http://www.gapminder.org/world>

Considering Categorical Data

Contingency Tables

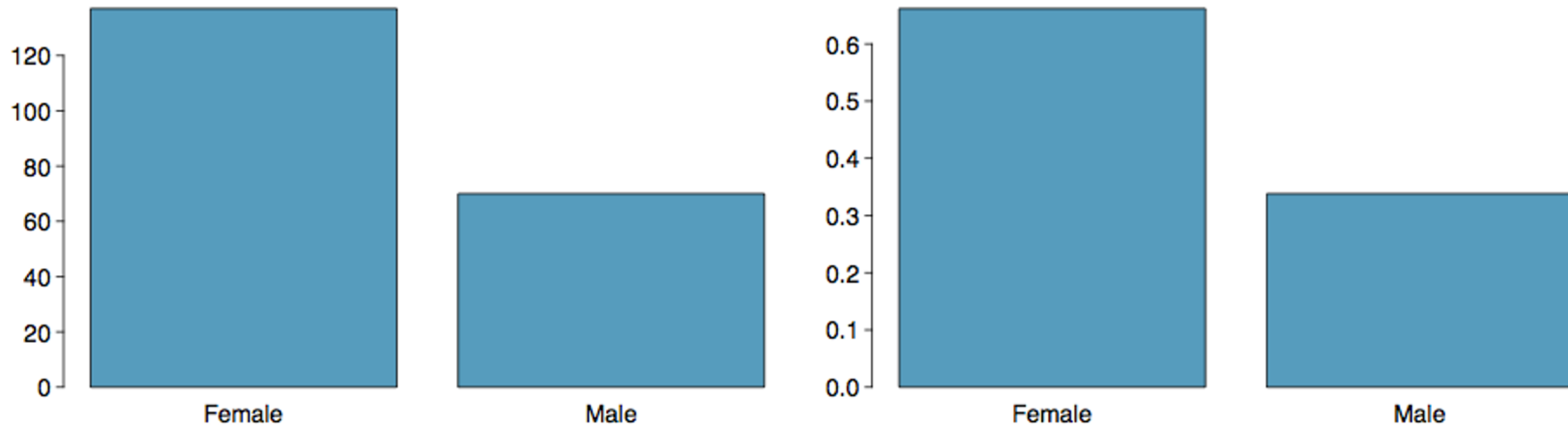
A table that summarizes data for two categorical variables is called a *contingency table*.

The contingency table below shows the distribution of students' genders and whether or not they are looking for a spouse while in college.

		looking for spouse		Total
		No	Yes	
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

Bar Plots

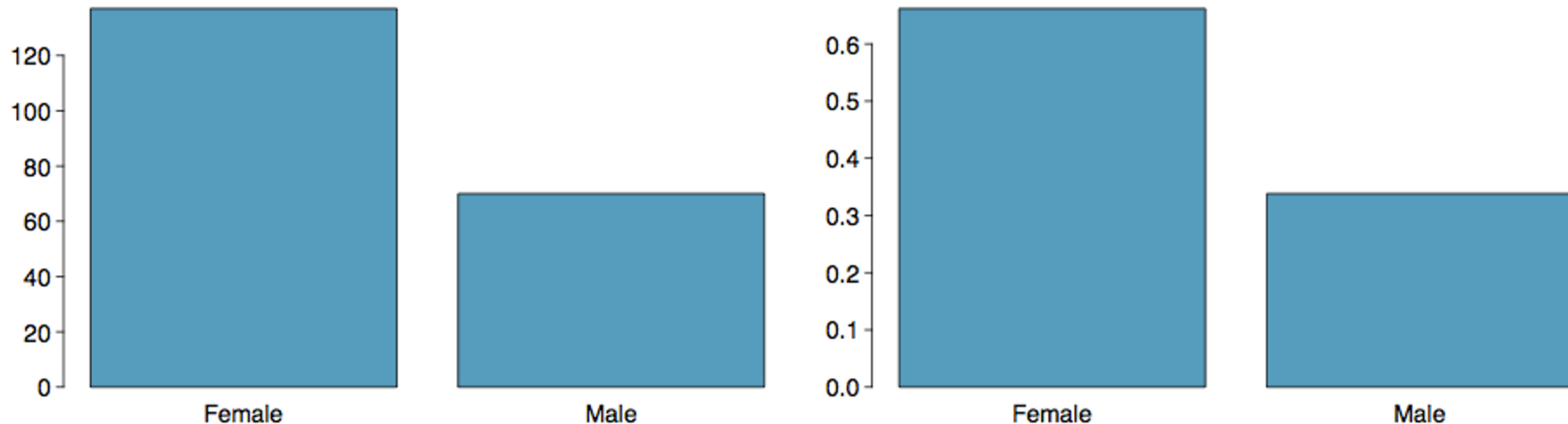
A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.



How are bar plots different than histograms?

Bar Plots

A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.



How are bar plots different than histograms?

Bar plots are used for displaying distributions of categorical variables, while histograms are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

Choosing the Appropriate Proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

		looking for spouse		Total
		No	Yes	
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

To answer this question we examine the row proportions:

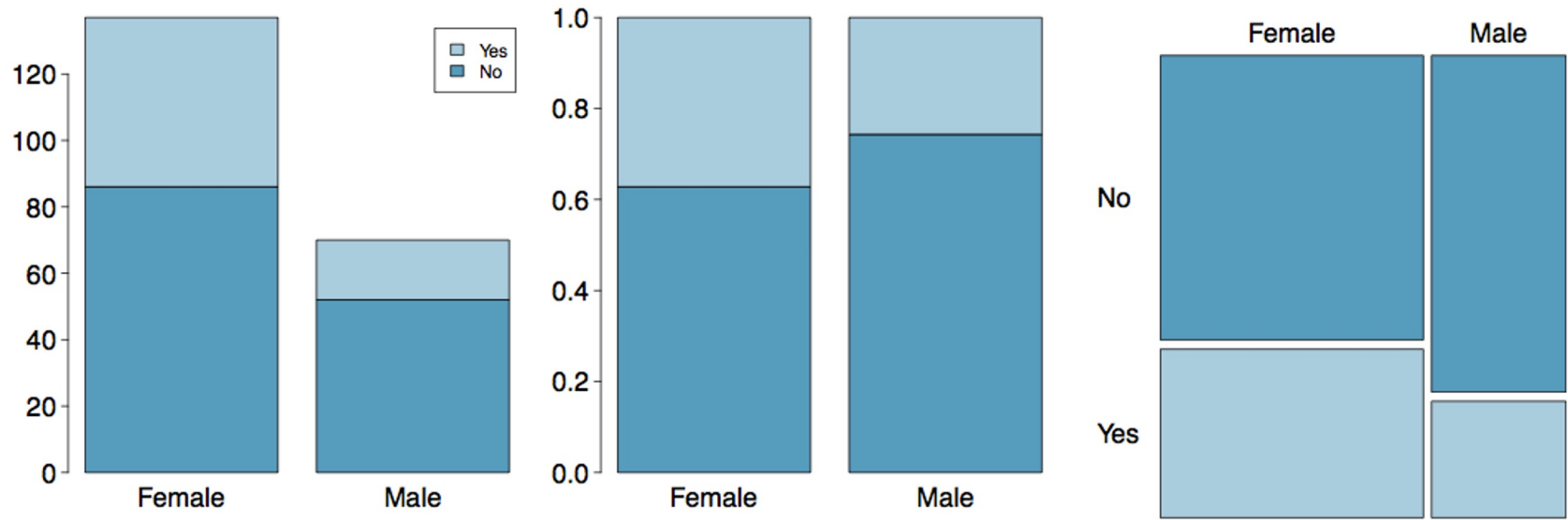
- % Females looking for a spouse: $51 / 137 \sim 0.37$
- % Males looking for a spouse: $18 / 70 \sim 0.26$

Bar plots with two variables

- *Stacked bar plot*: Graphical display of contingency table information, for counts.
- *Side-by-side bar plot*: Displays the same information by placing bars next to, instead of on top of, each other.
- *Standardized stacked bar plot*: Graphical display of contingency table information, for proportions.

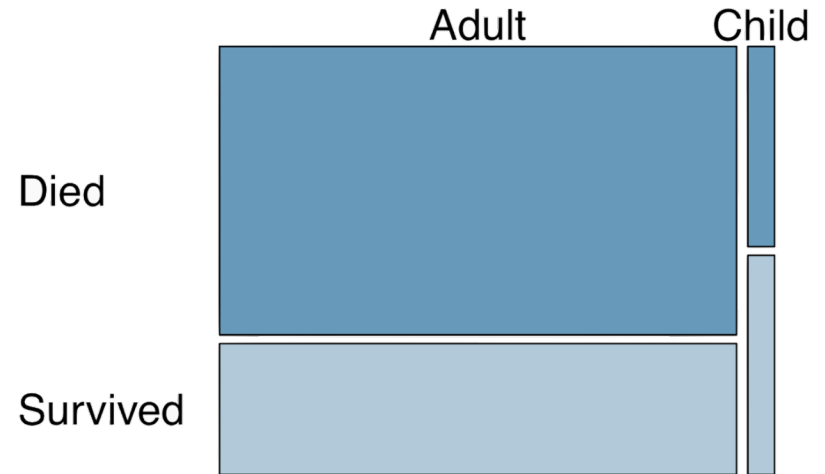
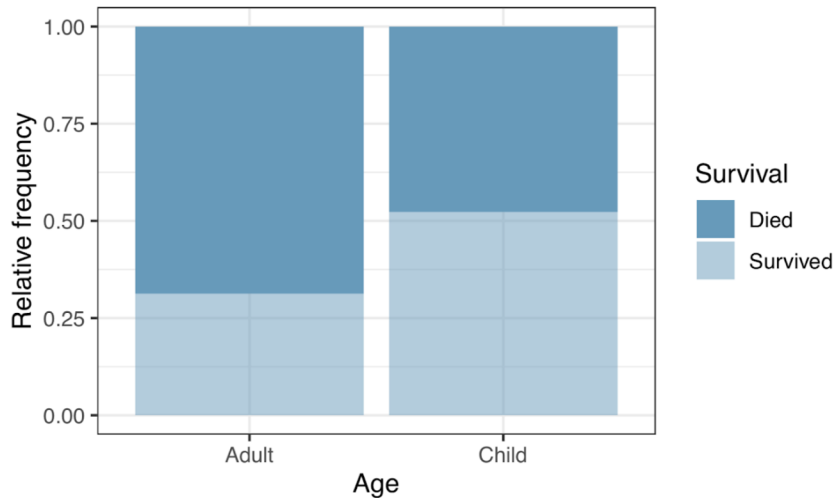
Segmented Bar and Mosaic Plots

What are the differences between the three visualizations shown below?



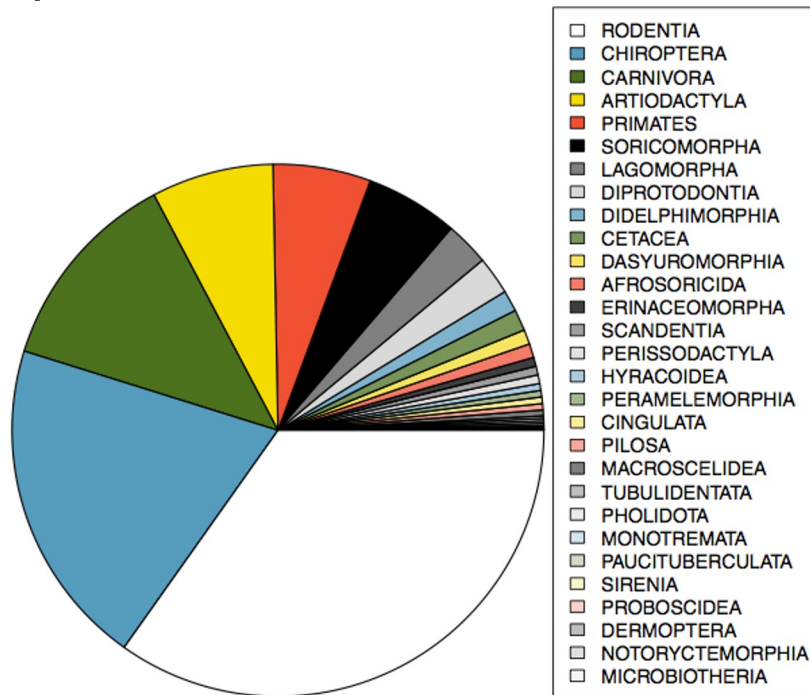
Mosaic plots

What are the differences between the two visualizations shown below?



Pie Charts

Can you tell which order encompasses the lowest percentage of mammal species?



Comparing Numerical Data Across Groups

Does there appear to be a relationship between class year and number of clubs students are in?

