

## Chapter 2

### Summarizing Data<sup>1</sup>

Department of Mathematics & Statistics  
North Carolina A&T State University

---

<sup>1</sup>These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

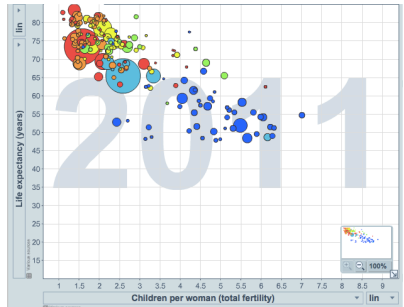
## Examining numerical data

# Scatterplot

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

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

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



# Scatterplot

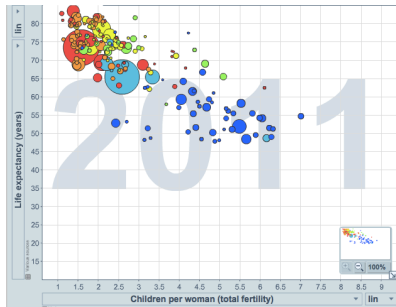
**Scatterplots** are useful for visualizing the relationship between two numerical variables.

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



## 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

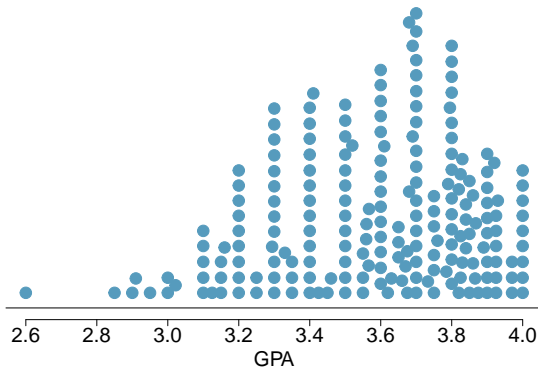
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

where  $x_1 + x_2 + \cdots + x_n$  represent the **n** observed values.

- ▶ The **population mean** is also computed the same way but is denoted as  $\mu$ . It is often not possible to calculate  $\mu$  since population data are rarely available.
- ▶ The sample mean is a **sample statistic**, and served 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

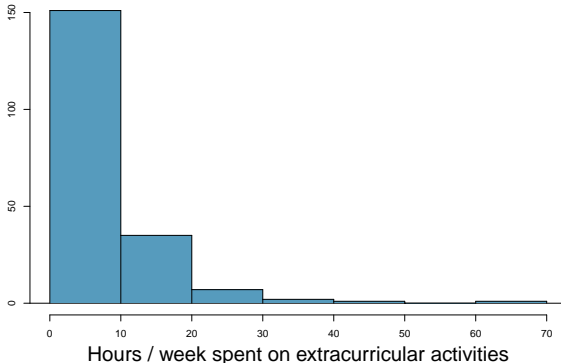
Higher bars represent areas where there are more observations, makes it a little easier to judge the center and the shape of the distribution.





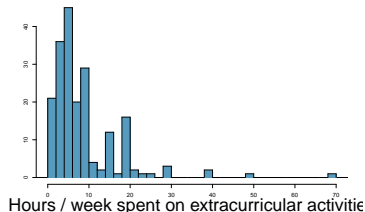
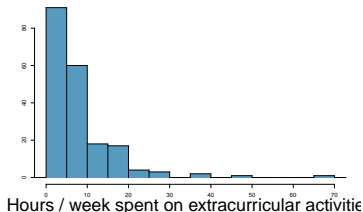
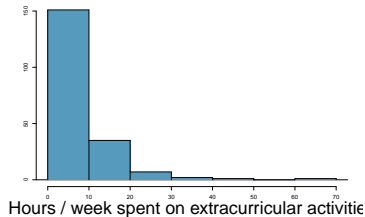
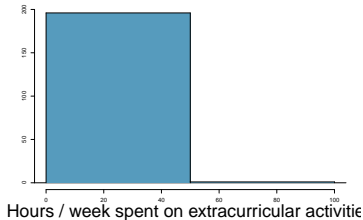
# Histograms - Extracurricular hours

- ▶ Histogram 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.



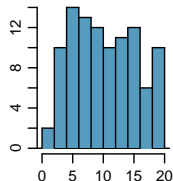
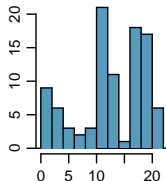
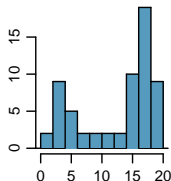
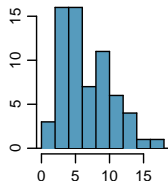
## Bin width

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



## Shape of the distribution: modality

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

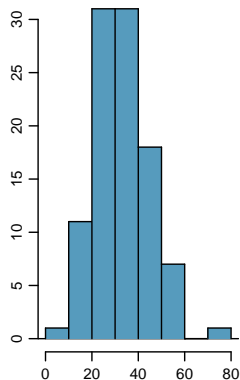
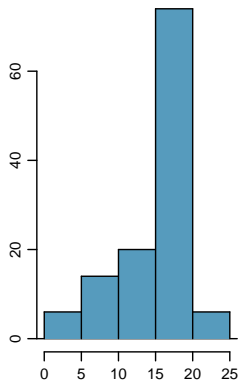
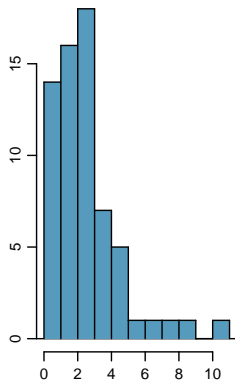


---

**Note:** In order to determine modality, step back and imagine a smooth curve over the histogram - imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

## Shape of the distribution: skewness

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

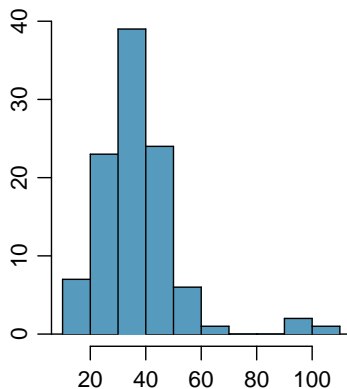
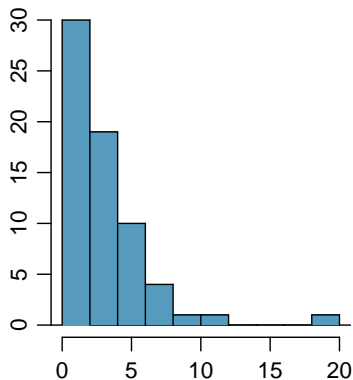


---

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

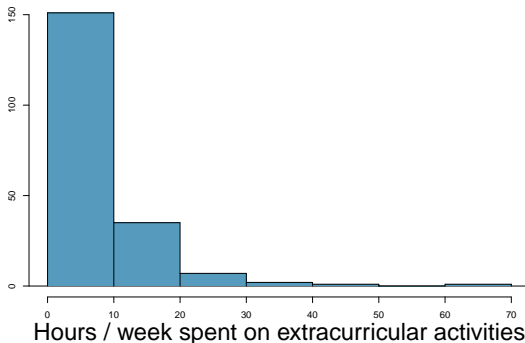
## Shape of the distribution: unusual observations

Are there any unusual observations or potential **outliers**?



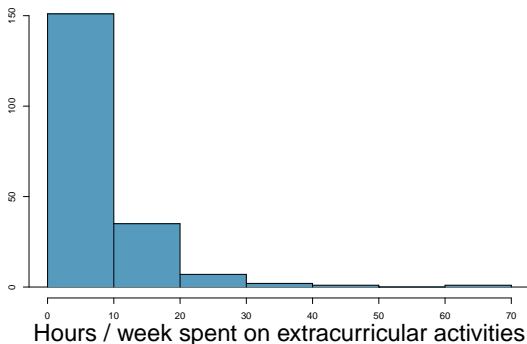
## Extracurricular activities

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



## Extracurricular activities

How would you describe the shape of the distribution of hours of 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





# Commonly observed shapes of distributions

## ► Modality

Unimodal



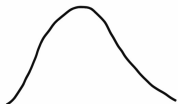
Bimodal



# Commonly observed shapes of distributions

## ► Modality

Unimodal



Bimodal



Multimodal



# Commonly observed shapes of distributions

## ► Modality

Unimodal



Bimodal



Multimodal



uniform



# Commonly observed shapes of distributions

## ► Modality

Unimodal



Bimodal



Multimodal



uniform



## ► Skewness

Right Skew



# Commonly observed shapes of distributions

## ► Modality

Unimodal



Bimodal



Multimodal



uniform



## ► Skewness

Right Skew



Left Skew



# Commonly observed shapes of distributions

## ► Modality

Unimodal



Bimodal



Multimodal



uniform



## ► Skewness

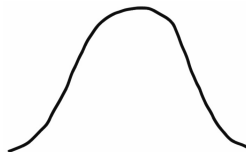
Right Skew



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 North Carolina
- C) House prices
- D) Birthdays of classmates (days 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 North Carolina
- C) House prices
- D) Birthdays of classmates (days of the month)



## Application activity: Shapes of distributions

Sketch the expected distribution 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.

Are you typical



<https://youtu.be/4B2xOvKFFz4>

Are you typical



<https://youtu.be/4B2xOvKFFz4>

How useful are centers alone for conveying the true characteristics of a distribution?

# Variance

**Variance** is roughly the average squared deviation from the mean.

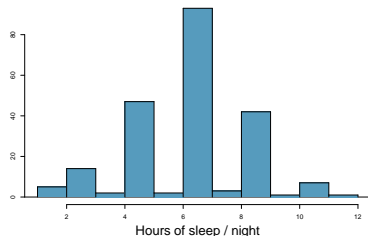
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

# 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  $n = 217$ .
- ▶ The variance of amount of sleep students get per night can be calculated as:

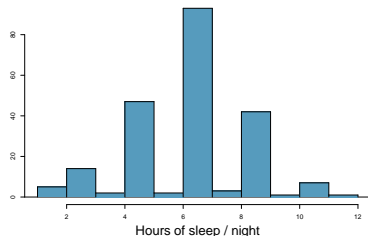


# Variance

**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

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

# Variance

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.

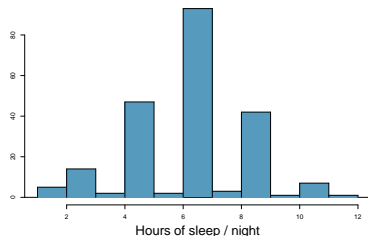
$$s = \sqrt{s^2}$$

# 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.



## Standard Deviation

The **standard deviation** is the square root of the variance  $s^2$ , and has the same units as the data.

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

Example: The following are samples of women's and men's ideal number of children. Find the standard deviation for each group.

Men: 0, 0, 0, 2, 4, 4, 4      Women: 0, 2, 2, 2, 2, 2, 4

**Answer:** The mean for men is  $\bar{x} = 14/7 = 2$  and the standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{(0 - 2)^2 + \dots + (4 - 2)^2}{7 - 1}} = \sqrt{\frac{24}{6}} = 2.0$$

► Similarly, the mean for women is  $\bar{x} = 14/7 = 2$  and the standard deviation is  $s = 1.2$  children.

# Standard Deviation

- ▶ The standard deviation is the typical deviation of an observation from the mean.
- ▶ The *larger* the value of standard deviation,  $s$ , the *greater* the variability of the data. As the spread of the data increases,  $s$  gets larger.
- ▶ Unlike the variance which has squared units, the standard deviation has the same units of measurement as the original observations.
- ▶ The standard deviation is zero ( $s = 0$ ) only when all observations have the same value, otherwise  $s > 0$ .
- ▶ The standard deviation  $s$  is less than the variance  $s^2$  unless  $s^2$  is smaller than 1.
- ▶ The standard deviation  $s$  is not resistant. That is, strong skewness or a few outliers can greatly increase  $s$ .

# Median

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

$$0, 1, \mathbf{2}, 3, 4$$

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

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

- ▶ Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the **50<sup>th</sup> percentile**.

# Median

- ▶ The **median** is the value that splits the data in half when ordered in ascending order.
- ▶ Example: the data below gives the per capita CO<sub>2</sub> emissions in 9 largest nations measured in metric tons per person. Find the value of the median.

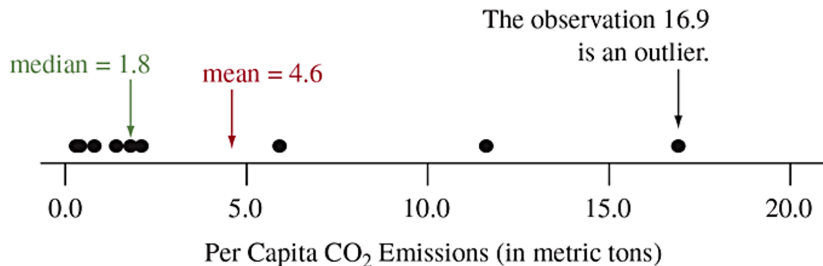
China 5.9; India 1.4; U.S. 16.9; Indonesia 1.8; Brazil 2.1; Pakistan 0.8; Nigeria 0.3; Bangladesh 0.4; Russia 11.6

## Solution:

- ▶ First, put the  $n = 9$  observations in order of their size.  
0.3, 0.4, 0.8, 1.4, 1.8, 2.1, 5.9, 11.6, 16.9
- ▶ Since  $n = 9$  is odd, the median is the middle observation:  
median = 1.8 metric tons.

## Median

- ▶ The **median** is the value that splits the data in half when ordered in ascending order.
- ▶ Unlike the mean, the *median* is a **resistant measure** as its value is not sensitive to outliers.



- ▶ If we drop out the U.S. value, what is the new median?
  - ▶ Now  $n = 8$  is even and the ordered values are:  
0.3, 0.4, 0.8, 1.4, 1.8, 2.1, 5.9, 11.6
  - ▶ The median is the average of the two middle observations:  
 $\text{median} = (1.4 + 1.8)/2 = 1.6$

## Q1, Q3, and IQR

- ▶ The 25<sup>th</sup> percentile is also called the first quartile, **Q1**.
- ▶ The 50<sup>th</sup> percentile is also called the median.
- ▶ The 75<sup>th</sup> 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

## Finding Quartiles

- ▶ Arrange the data in order.
- ▶ Consider the median. This is the second quartile,  $Q_2$ .
- ▶ Consider the lower half of the observations (excluding the median itself if  $n$  is odd). The median of these observations is the first quartile,  $Q_1$ .
- ▶ Consider the upper half of the observations (excluding the median itself if  $n$  is odd). Their median is the third quartile,  $Q_3$ .

## Q1, Q3, and IQR

Example: Consider the following sodium values in 20 brands of breakfast cereals. The sodium values, in ascending order, are:

					<b>Q1 = 135</b>				
									
0	50	70	100	<b>130</b>	<b>140</b>	140	150	160	<b>180</b>
<b>180</b>	180	190	200	<b>200</b>	<b>210</b>	210	220	290	340
									
					<b>Q3 = 205</b>				

What are the quartiles and IQR for the 20 cereal sodium values?

- ▶ The median of the 20 values is the average of the 10th and 11th observations, 180 and 180, which is  $Q_2 = 180$  mgs.
- ▶ The first quartile  $Q_1$  is the median of the 10 smallest values, which is the average of 130 & 140,  $Q_1 = 135$  mgs.
- ▶ The third quartile  $Q_3$  is the median of the 10 largest values, which is the average of 200 & 210,  $Q_3 = 205$  mgs.

## Q1, Q3, and IQR

Example: Consider the following sodium values in 20 brands of breakfast cereals. The sodium values, in ascending order, are:

					<b>Q1 = 135</b>								
					<b>130</b>	<b>140</b>							
0	50	70	100				140	150	160		<b>180</b>		
<b>180</b>	180	190	200		<b>200</b>	<b>210</b>	210	220	290	340			
					<b>Q3 = 205</b>								

- The interquartile range (IQR) is the distance between the third quartile and first quartile:

$$IQR = Q_3 - Q_1 = 205 - 135 = 70$$

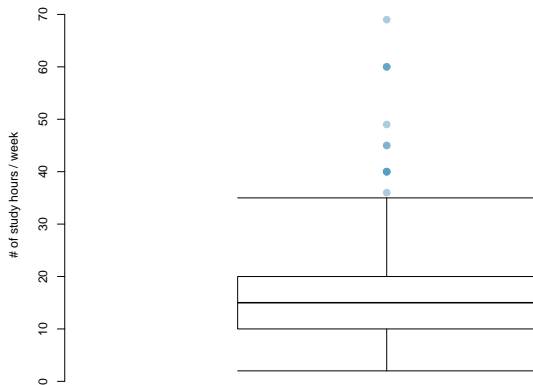
This means that the middle 50% of the distribution of sodium amount stretches over a distance of 70.

# The Five-Number Summary

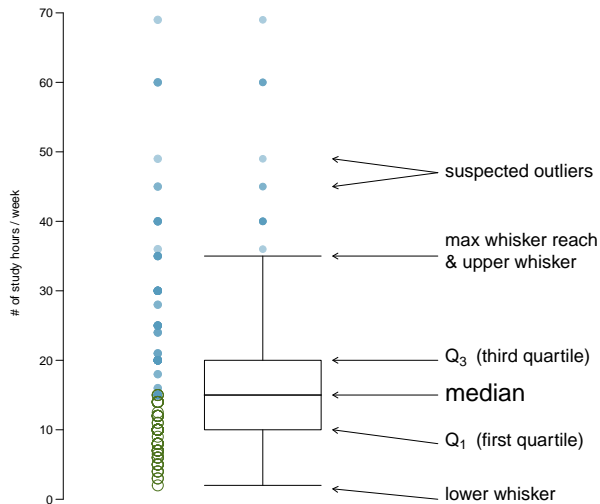
- ▶ The **five-number summary** is a numerical descriptive summary of the distribution of the data and it consists of the following:
  - ▶ Minimum value
  - ▶ First Quartile ( $Q_1$ )
  - ▶ Median ( $Q_2$ )
  - ▶ Third Quartile ( $Q_3$ )
  - ▶ Maximum value
- ▶ The **five-number summary** is the basis of a graphical display called the *box plot*.

## 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  $1.5 \times IQR$  away from the quartiles.

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

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

## Whiskers and Outliers

- **Whiskers** of a box plot can extend up to  $1.5 \times IQR$  away from the quartiles.

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

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

$$IQR : 20 - 10 = 10$$

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

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



## Whiskers and Outliers

- ▶ **Whiskers** of a box plot can extend up to  $1.5 \times IQR$  away from the quartiles.

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

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

$$IQR : 20 - 10 = 10$$

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

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

- ▶ 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

Why is it important to look for outliers?

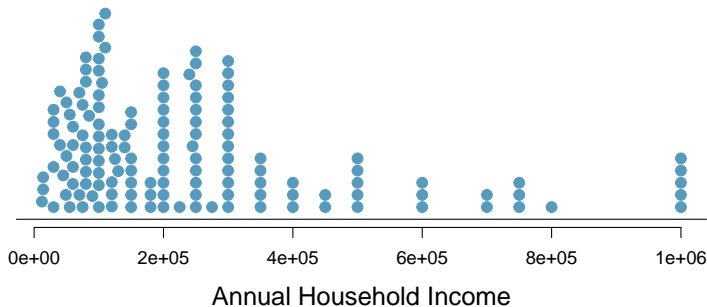
# Outliers

Why is it important to look for outliers?

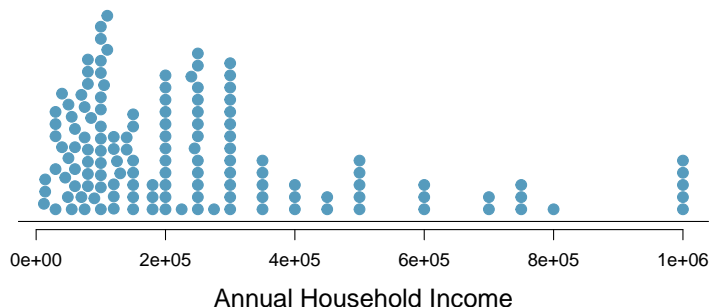
- ▶ Identify extreme skew in the distribution.
- ▶ Identify data collection and entry errors.
- ▶ Provide insight into interesting features of the data.

## 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.

# 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.

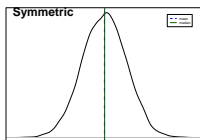
If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

Median

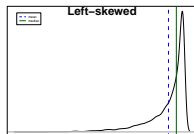
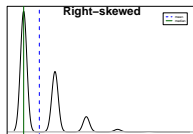


# Mean vs. Median

- ▶ If the distribution is symmetric, center is often defined as the mean:  $\text{mean} \approx \text{median}$ .

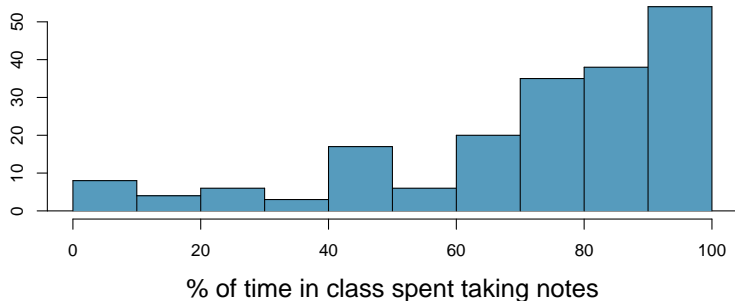


- ▶ If the distribution is skewed or has extreme outliers, center is often defined as the median.
  - ▶ Right-skewed:  $\text{mean} > \text{median}$
  - ▶ Left-skewed:  $\text{mean} < \text{median}$



## Practice

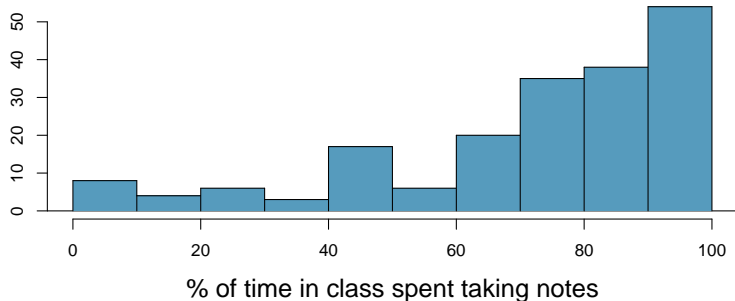
Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter etc.?



- A) Mean  $>$  Median C) Mean  $\approx$  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.?

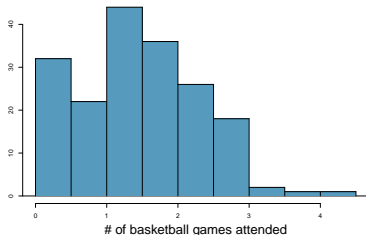
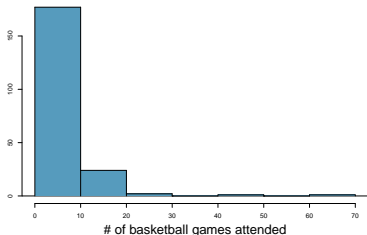


- A) Mean  $>$  Median C) Mean  $\approx$  Median  
B) Mean  $<$  Median D) Impossible to tell

## 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	...
$\log(\# \text{ of games})$	4.25	3.91	3.22	...

- ▶ However, results of an analysis in log units of the measured variable might be difficult to interpret.

# 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	...
$\log(\# \text{ of games})$	4.25	3.91	3.22	...

- ▶ However, results of an analysis in log units of the measured variable might be difficult to interpret.

What other variables would you expect to be extremely skewed?

# 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	...
$\log(\# \text{ of games})$	4.25	3.91	3.22	...

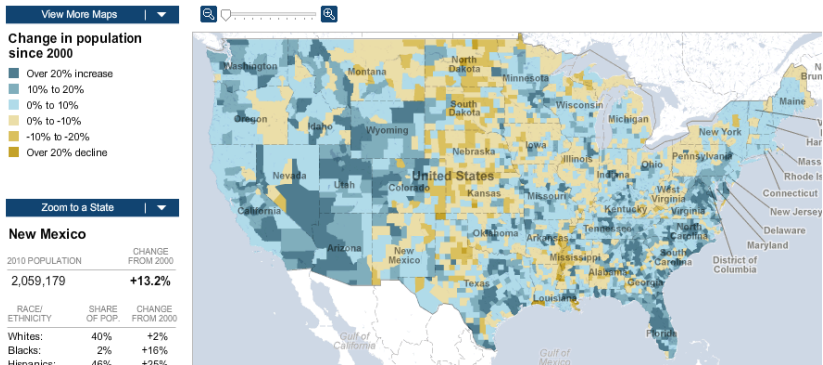
- ▶ However, results of an analysis in log units of the measured variable might be difficult to interpret.

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?



<https://www.nytimes.com/projects/census/2010/map.html>