

SEIS 631

Class 3



Descriptive Statistics

- **Center** - “What is a typical/common/representative value?”
- **Spread** - “How far are data values from each other, typically?”
- **Shape/Groupings** - “How are the data values distributed? Are there clusters of data?”
- **Extreme or Surprising Values**

Describing Populations vs. Samples

Population

Parameter

Example: Median income determined from the U.S. Census

Sample

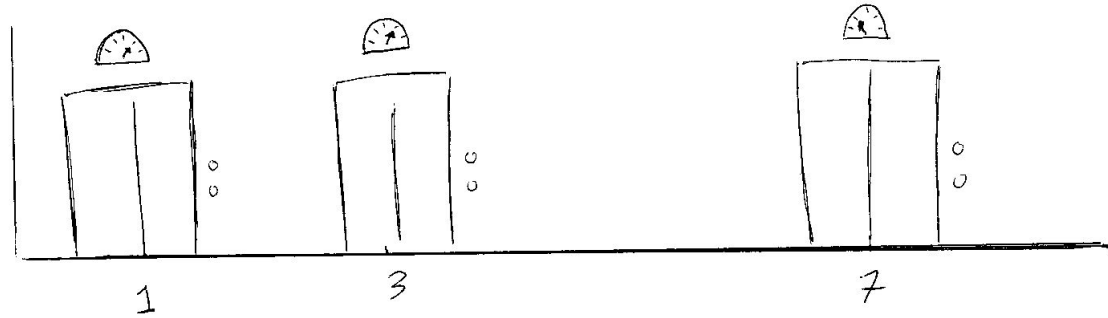
Statistic

Example: Proportion of Survey Respondents who prefer Coke to Pepsi

- If a measurement is calculated from a sample, it's called a sample statistic. **Sample statistics are point estimates for the unknown population parameters.**

Descriptive Statistics

Elevator Problem



All the floor indicators are broken.
Where should you stand while you
wait?

Elevator Problem Common Responses

“Average” 3.67, hedges toward the extreme

“Mid-Range” 4, minimizes the worst case scenario

“Median” 3, minimizes the total distance from all elevators

“Trimmed Mean” 2, like the average but ignores extreme values

“Min/Max” farthest away from the crowd

Measures of Center

What's a typical value? What's the middle of the data set? Where is the largest cluster of data values?

- Mean
- Median
- Mode
- **Notation:** Latin (regular) alphabet for sample statistics, Greek alphabet for population parameter.

How Far are You from Campus?

If you don't know, use Google Maps (or an equivalent service) to determine your distance from campus, rounded to the nearest mile.

Enter your information into this form.

<https://forms.gle/9dEzNUdUUyzYUwVx5>

Example

- 7 data points
 - 34 23 1 34 2 6 89
- **mean**: Arithmetic Average of all the numbers (add up numbers and divide by # of sample)
$$\bar{x} = 27$$
- **mode**: Most common value. In this case, 34
- **median**: 1 2 6 **23** 34 34 89
 - In case of even number of data points, take the mean of the middle two in sorted order.
 - Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50th percentile.
 - Data must be ordered before determining the median.

The Arithmetic Mean

- The balance point of the data.
 - The “center of mass”
- Equidistribution
 - If total assets/attributes were distributed equally
- Deviations above the mean are balanced by deviations below the mean

For n data values:

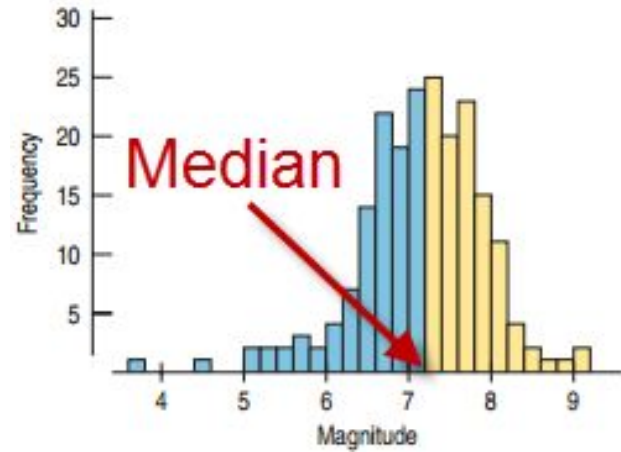
$$x_1, x_2, x_3, \dots, x_i, \dots, x_n$$

The mean is:


$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

The Median

- **Median:** The center of the data values
- Half of the data values are to the left of the median and half are to the right of the median.
- For symmetric distributions, the median is directly in the middle.



Calculating the Median: Odd Sample Size

- First order the numbers.
- If there are an odd number of numbers, n , the median is at position $\frac{n+1}{2}$
- Find the median of the numbers: 2, 4, 5, 6, 7, 9, 9.
- $\frac{n+1}{2} = \frac{7+1}{2} = 4$ 
- The median is the fourth number: 6
- Note that there are 3 numbers to the left of 6 and 3 to the right.

Calculating the Median: Even Sample Size

- First order the numbers.
- If there are an even number of numbers, n , the median is the average of the two middle numbers: $\frac{n}{2}, \frac{n}{2} + 1$
- Find the median of the numbers: 2, 2, 4, 6, 7, 8.
- $\frac{n}{2} = \frac{6}{2} = 3$
- The median is the average of the third and the fourth numbers:

Median

$$\text{Median} = \frac{4 + 6}{2} = 5$$

Measures of center in R

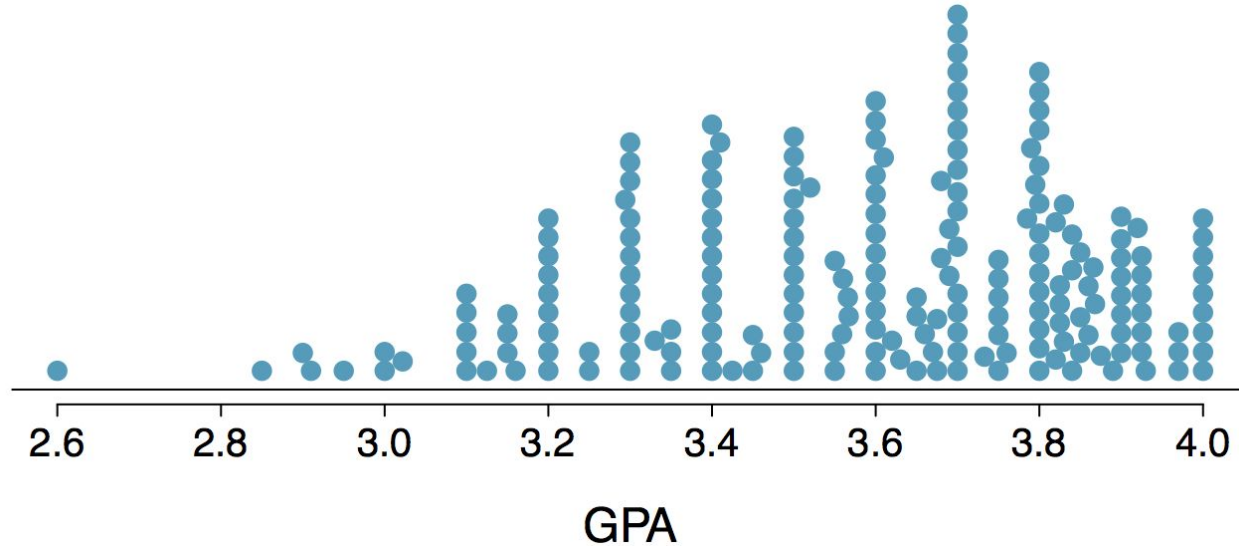
- `x <- c(34, 23, 1, 2, 34, 6, 89)` `# Create a list of numbers and store in x`
- `mean(x)`
- `median(x)`
- No base R function for mode.

Visual Summaries

Histograms and Stacked Dot Plots

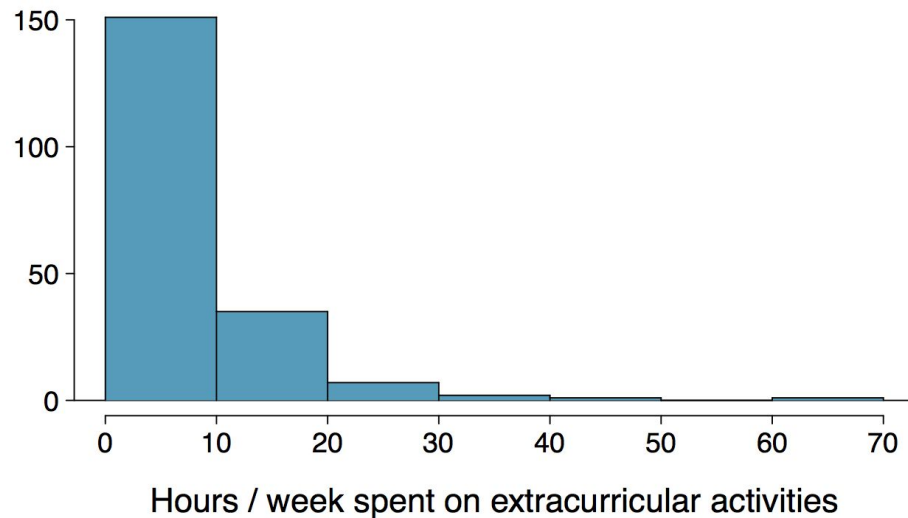
Stacked Dot Plot

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



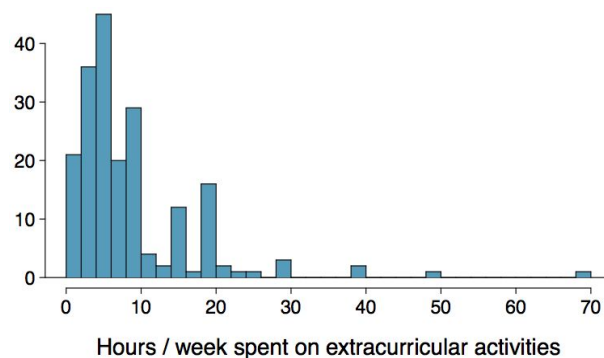
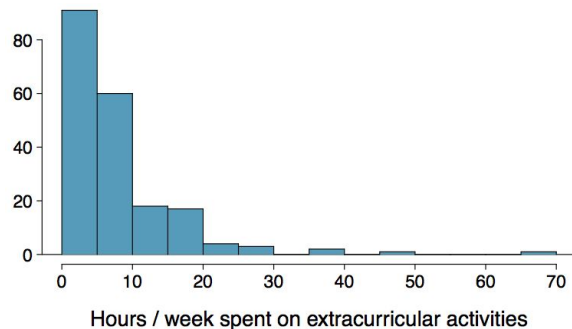
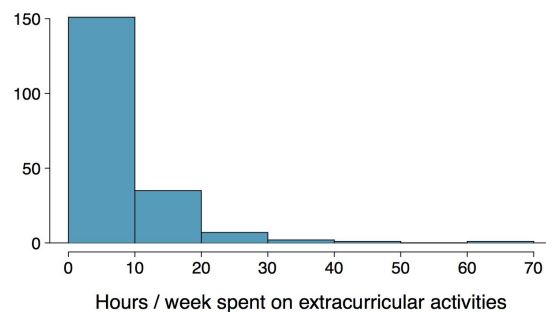
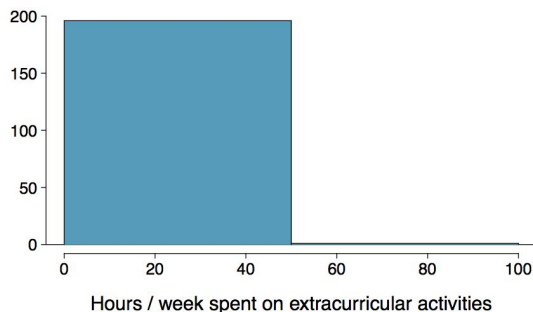
Histograms - Extracurricular Hours

- 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.

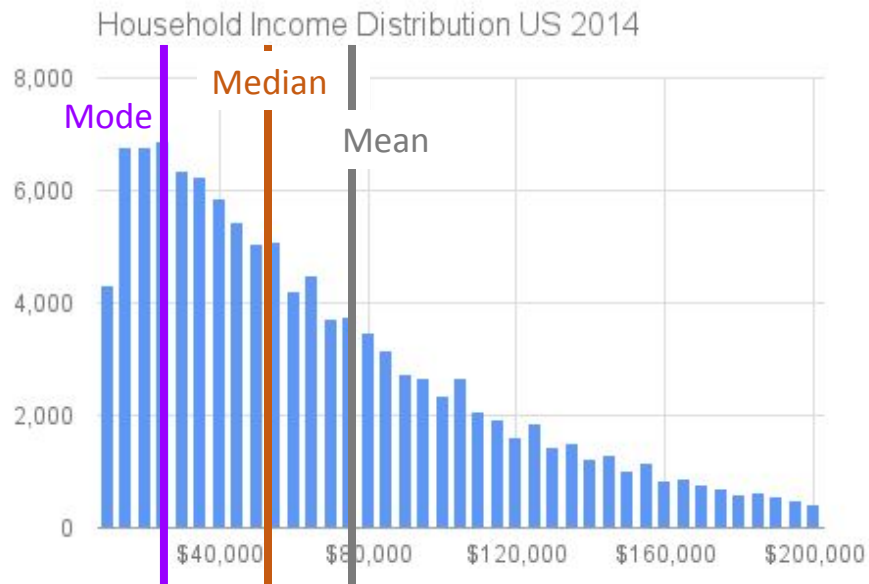


Bin Width

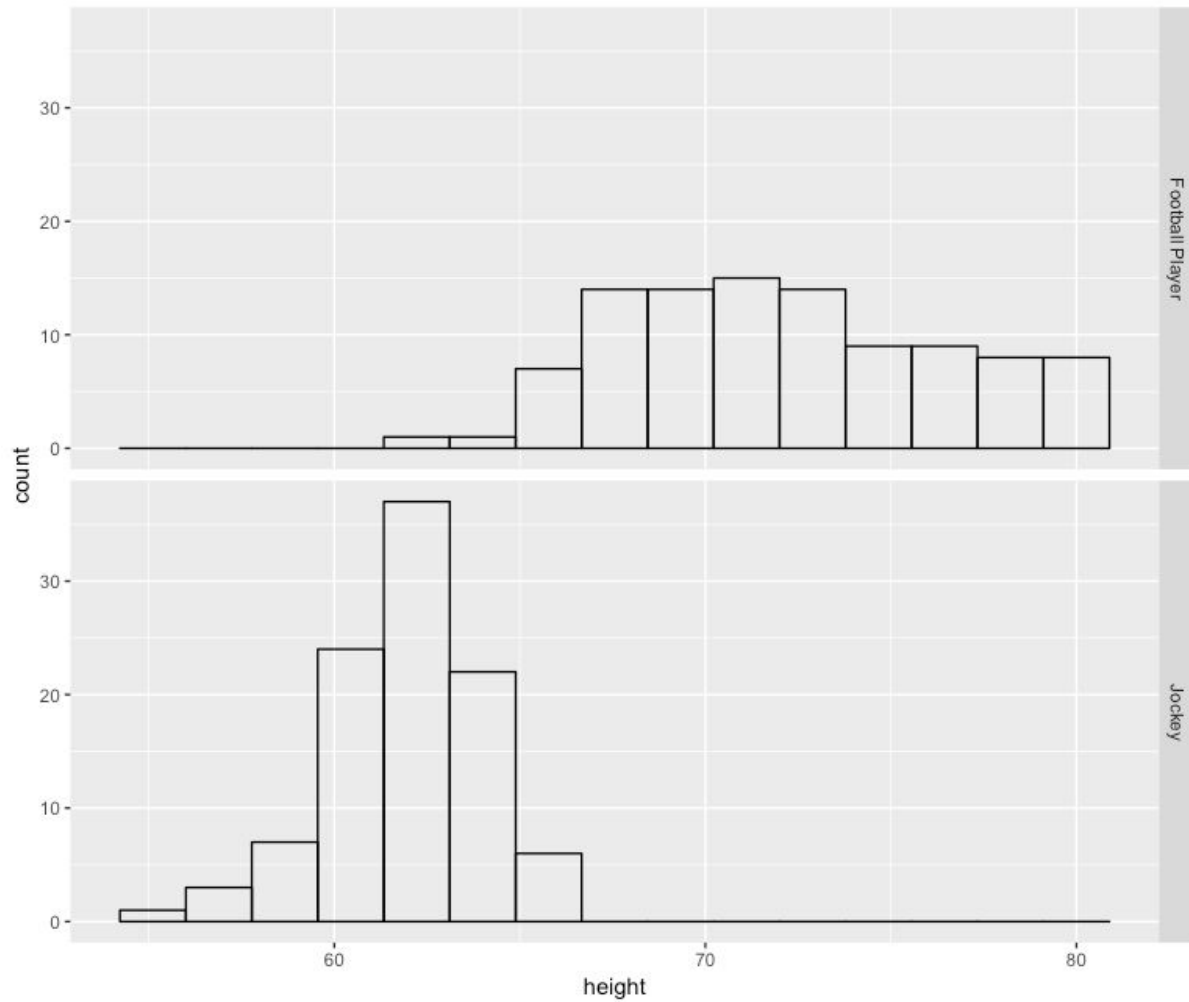
Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



Measures of Center in Histograms



Compare



measures of spread

- range: (max - min)
- Variance
- standard deviation
- inter-quartile range

Range

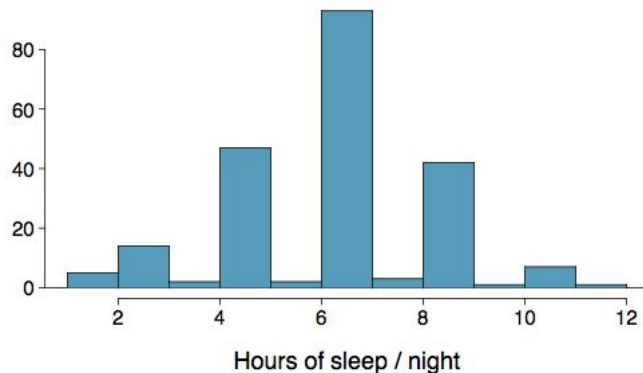
Range = max - min

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?

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

Why do we divide by $n-1$ and not n ?

- We only do this when calculating SAMPLE variance.
- If we don't, we don't get a reliable value. More on this later.

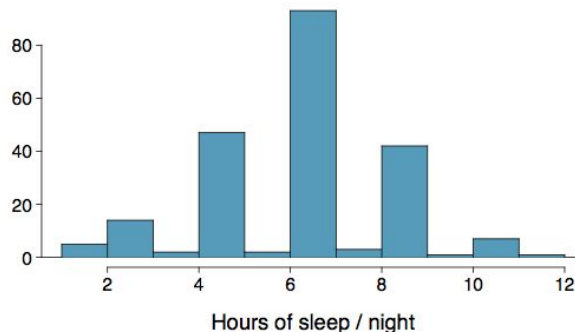
Standard Deviation

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:

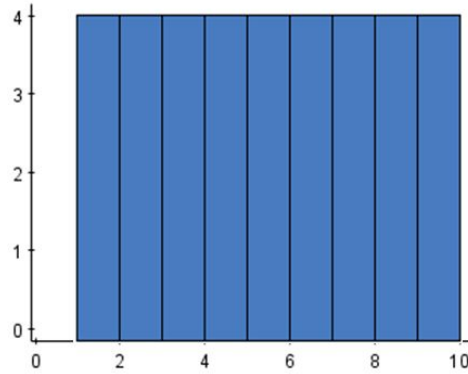
$$s = \sqrt{4.11} = 2.03 \text{ hours}$$



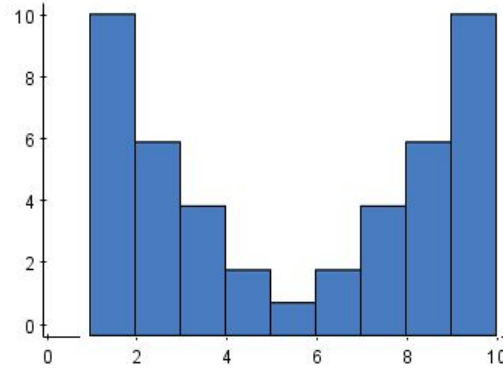
We can see that all of the data are within 3 standard deviations of the mean.

The Standard Deviation and Histograms

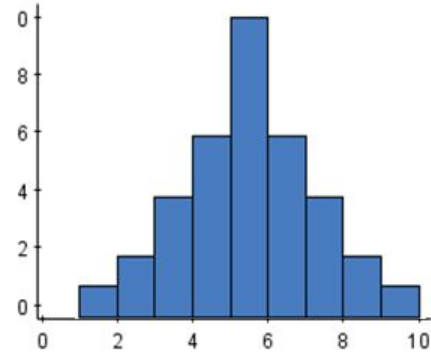
Order the histograms below from smallest standard deviation to largest standard deviation.



A



B



C

Answer: C, A, B

Percentiles and Quartiles

- Percentiles divide the data in one hundred groups.
- The n^{th} percentile is the data value such that n percent of the data lies below that value.
- For large data sets, the median is the 50^{th} percentile.
- The median of the lower half of the data is the 25^{th} percentile and is called the first quartile (Q1).
- The median of the upper half of the data is the 75^{th} percentile and is called the third quartile (Q3).

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.

$$IQR = Q3 - Q1$$

The Interquartile Range

- The **Interquartile Range (IQR)** is the difference between the upper quartile and the lower quartile

$$\text{IQR} = Q3 - Q1$$

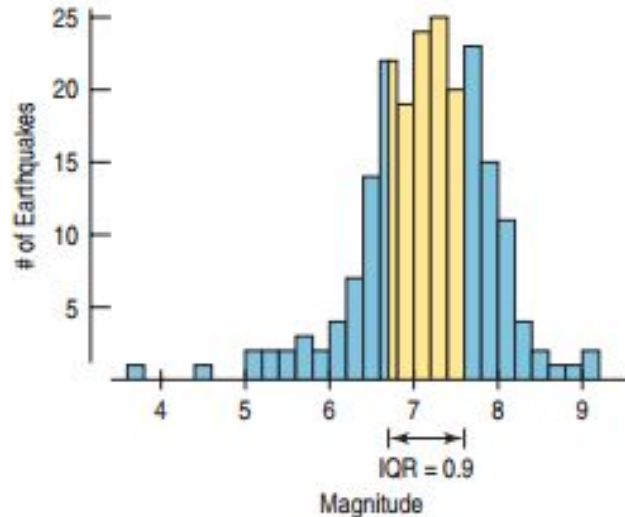
- The IQR measures the range of the middle half of the data.

- Example: If $Q1 = 23$ and $Q3 = 44$ then

$$\text{IQR} = 44 - 23 = 21$$

The Interquartile Range

- The Interquartile Range for earthquake causing tsunamis is 0.9.
- The picture below shows the meaning of the IQR.



Measures of Spread in R

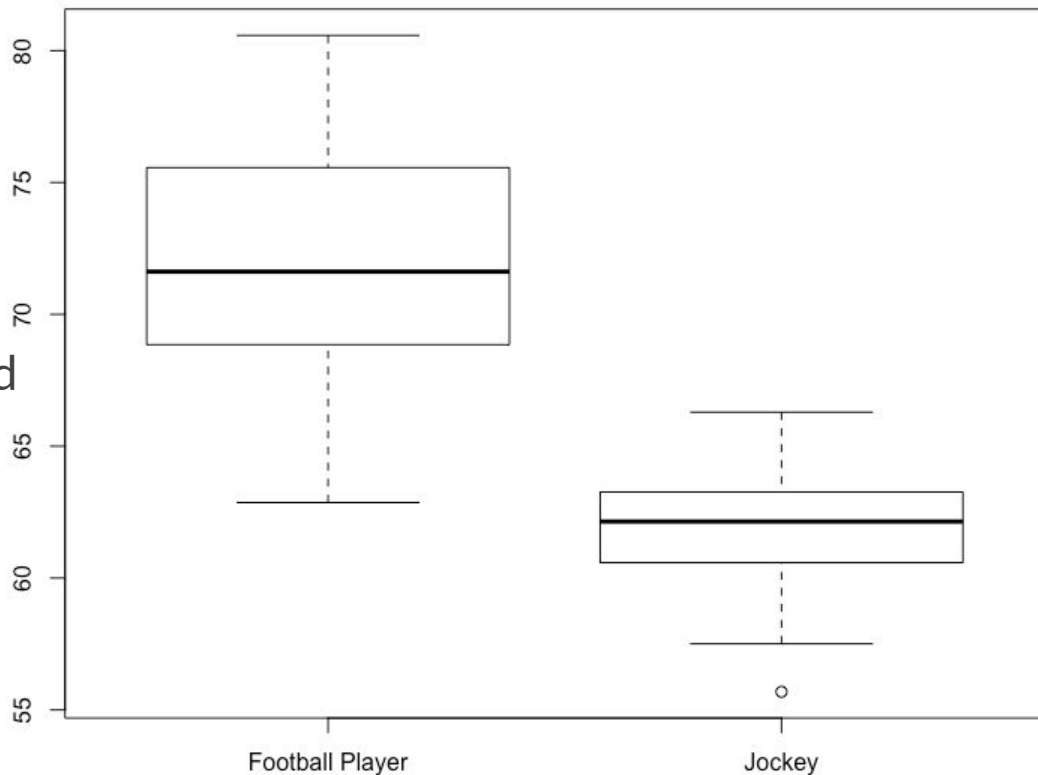
- `var(x)` # Variance
- `sd(x)` # standard deviation
- `IQR(x)` #IQR

Visual Summaries

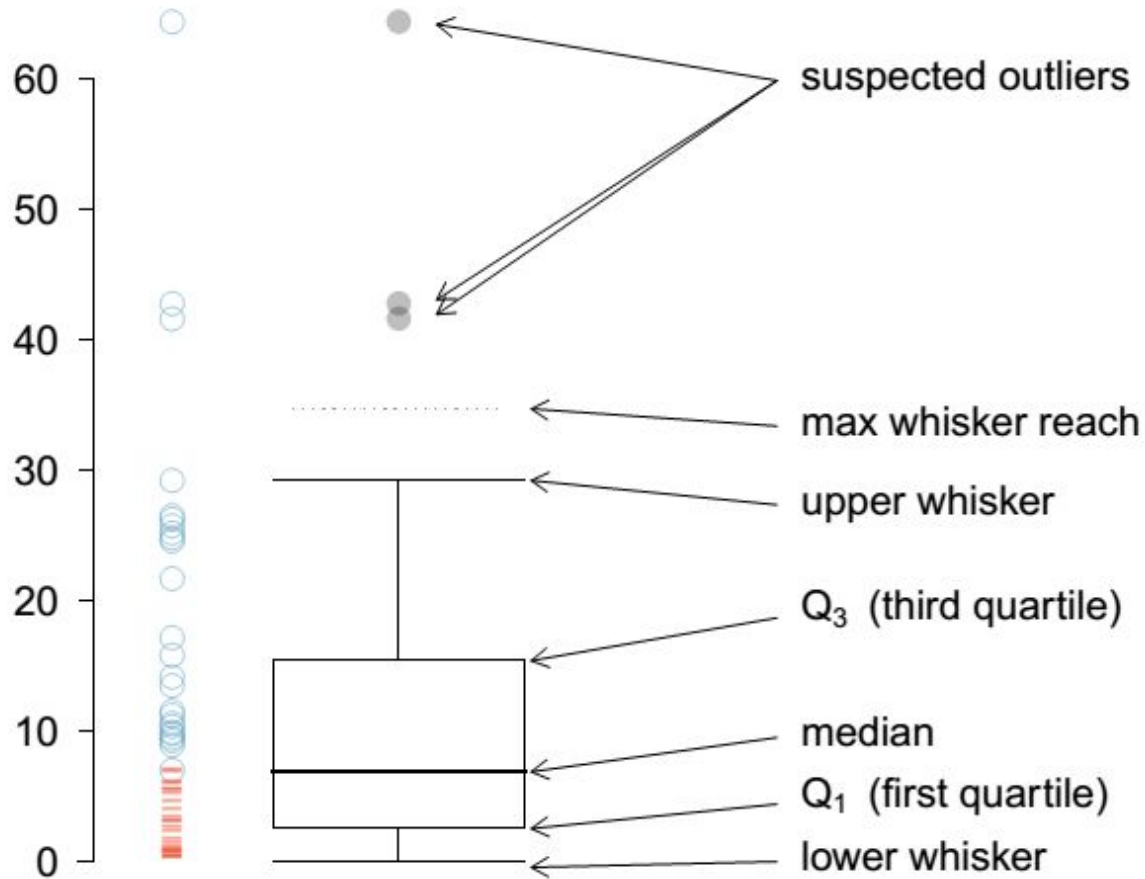
Box Plots (a.k.a. Box and Whisker Diagrams)

Boxplots

- Show center and spread of data
 - Median
 - IQR
- Potential Outliers are Flagged as circles or dots



Box plots, quartiles, and the median

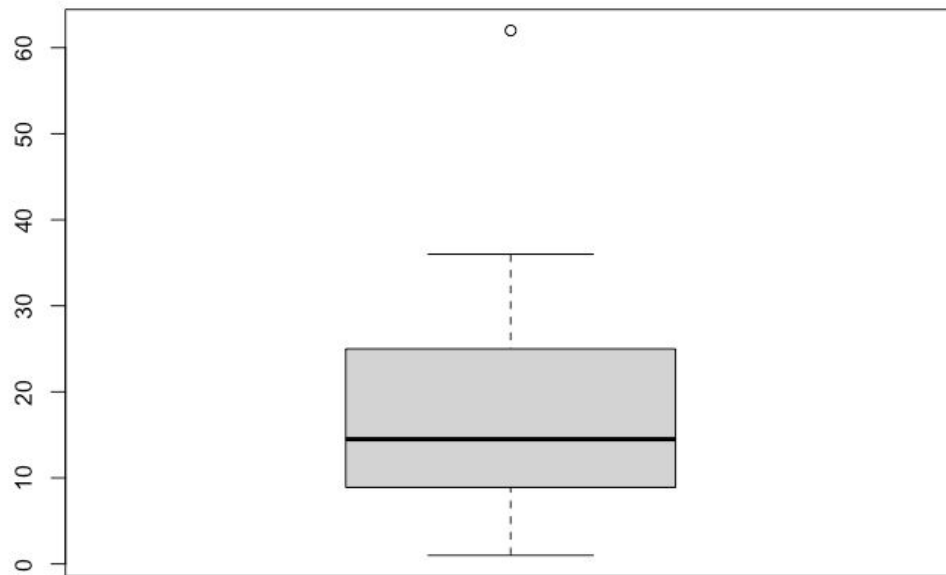


Box plots, quartiles, and the median

- The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half.
- The second step in building a box plot is drawing a rectangle to represent the middle 50% of the data.
- The two boundaries of the box are called the **first quartile** (the 25th percentile, i.e. 25% of the data fall below this value) and the **third quartile** (the 75th percentile).
- Length of the box is the **Interquartile range (IQR)** = $Q_3 - Q_1$
- Extending out from the box, the **whiskers** attempt to capture the data outside of the box, however, their reach is never allowed to be more than $1.5 \times IQR$.
- In the previous figure, the upper whisker does not extend to the last three points, which is beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit.

Boxplot of the Commute Data

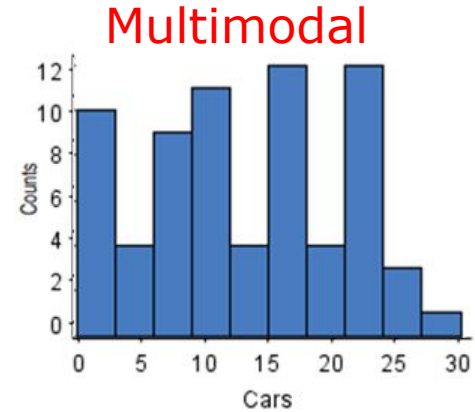
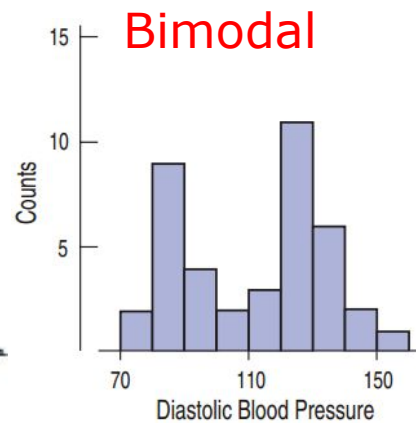
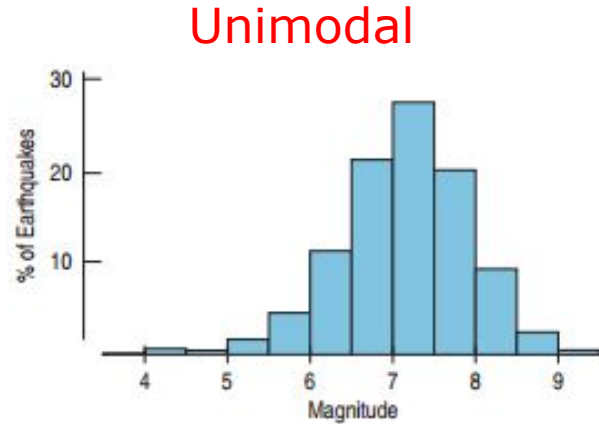
Boxplot of the Commute Data



Shapes of Distributions

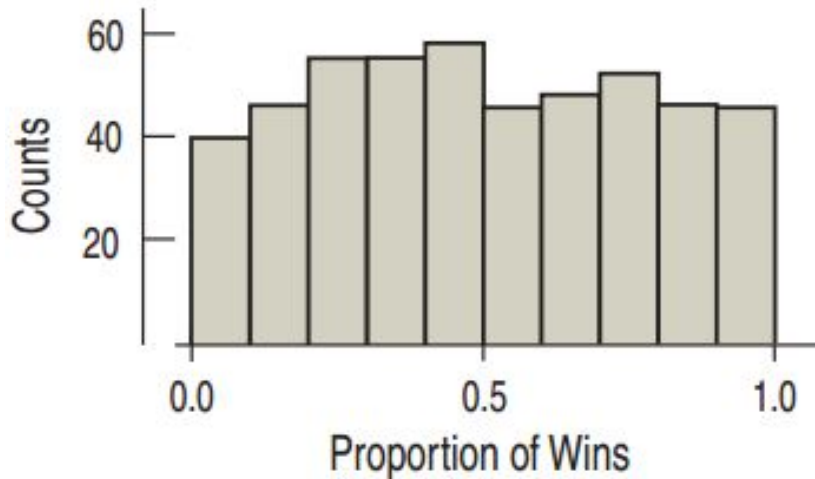
Modes

- A Mode of a histogram is a hump or high-frequency bin.
 - One mode → Unimodal
 - Two modes → Bimodal
 - 3 or more → Multimodal



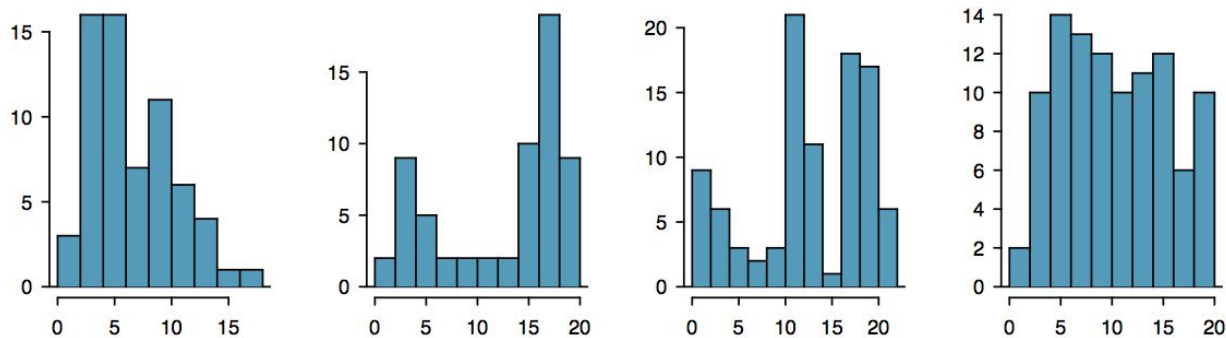
Uniform Distributions

- **Uniform Distribution:** All the bins have the same frequency, or at least close to the same frequency.
- The histogram for a uniform distribution will be **flat**.



Shape of a Distribution: Modality

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

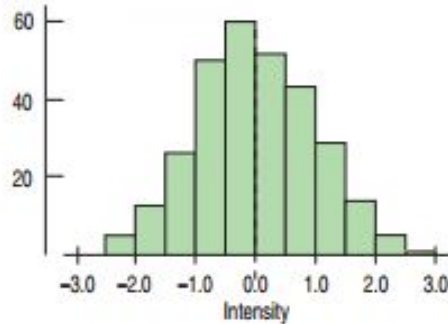


Note: In order to determine modality, step back and imagine a smooth curve over the histogram

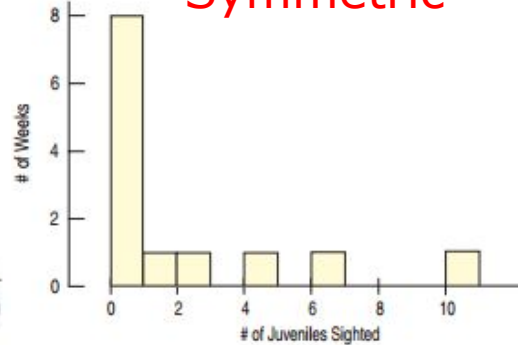
Symmetry

- The histogram for a **symmetric** distribution will look the same on the left and the right of its center.

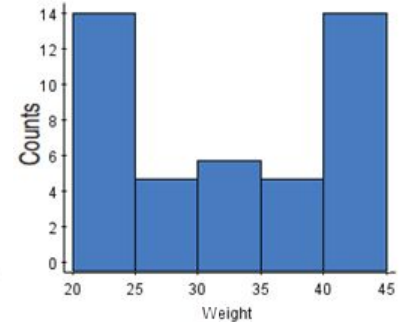
Symmetric



Not
Symmetric

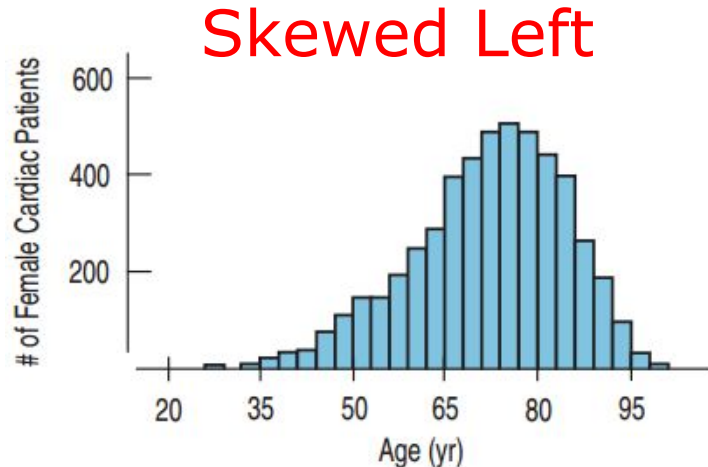
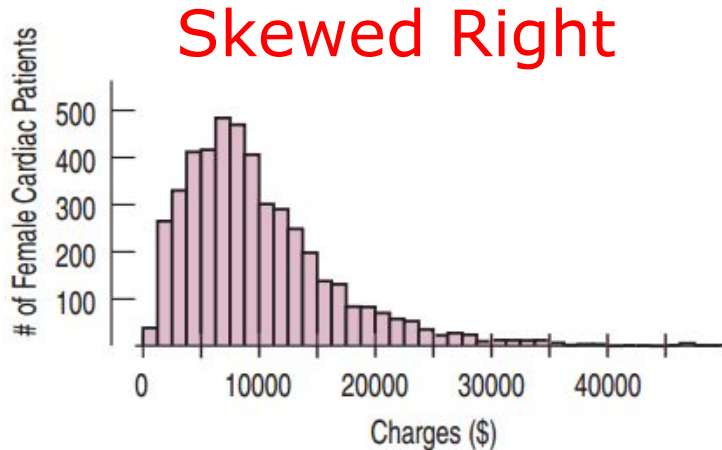


Symmetric



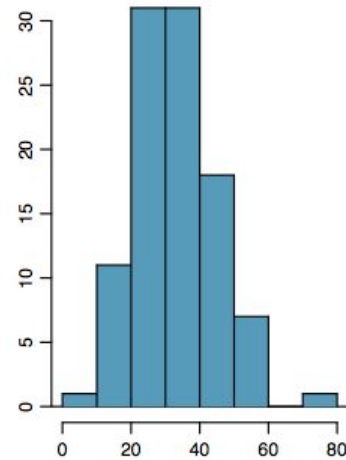
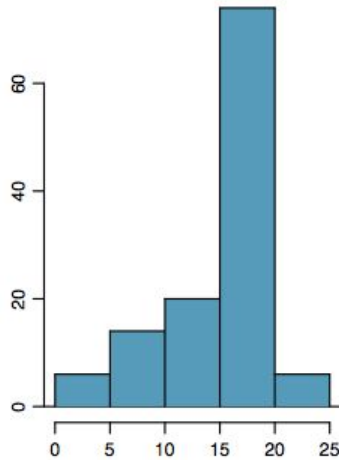
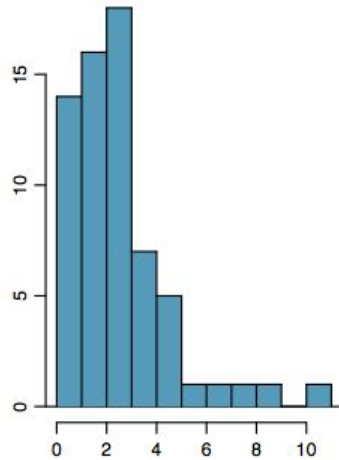
Skew

- A histogram is **skewed right** if the longer tail is on the right side of the mode.
- A histogram is **skewed left** if the longer tail is on the left side of the mode.



Shape of a Distribution: Skewness

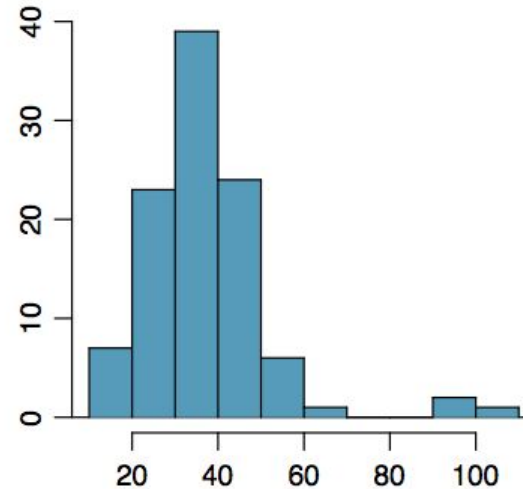
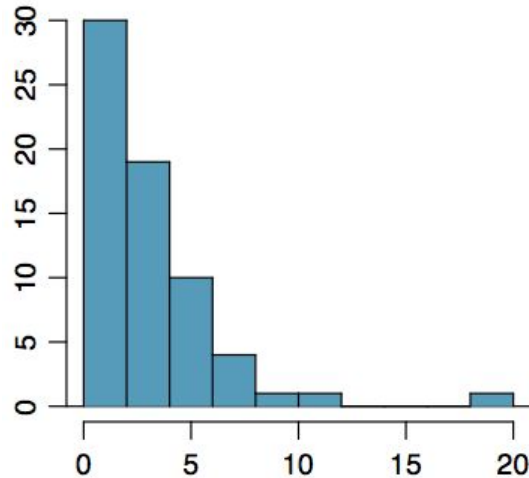
Is the histogram right skewed, left skewed, or symmetric?



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

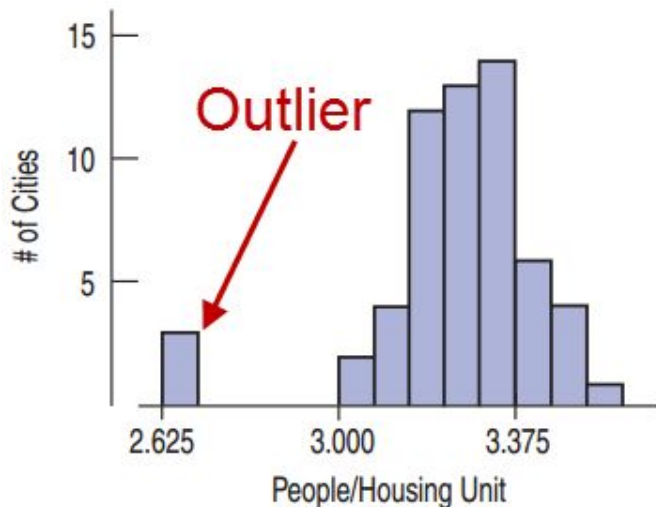
Shape of a Distribution: Unusual Observations

Are there any unusual observations or potential outliers?



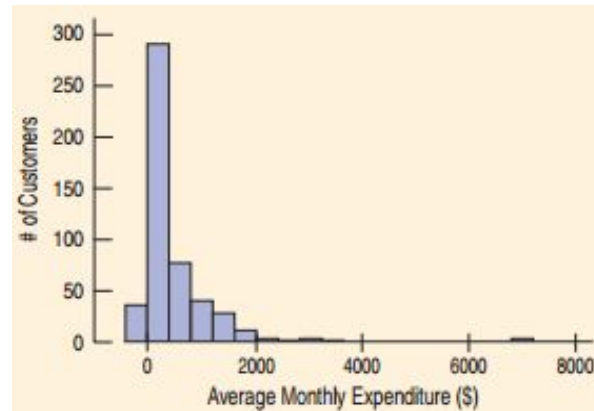
Outliers

- An **Outlier** is a data value that is far above or far below the rest of the data values.
- An outlier is sometimes just an error in the data collection.
- An outlier can also be the most important data value.
 - Income of a CEO
 - Pinocchio's nose length after lying
 - Elevation at Death Valley



Example

- The histogram shows the amount of money spent by a credit card company's customers. Describe and interpret the distribution.
 - The distribution is **unimodal**. Customers most commonly spent a small amount of money.
 - The distribution is **skewed right**. Many customers spent only a small amount and a few were spread out at the high end.
 - There is an **outlier** at around **\$7000**. One customer spent much more than the rest of the customers.



Commonly observed shapes of distributions

Modality

unimodal



bimodal



multimodal



uniform



Skewness

right skew



left skew



symmetric



A man with light-colored hair, wearing a dark suit, white shirt, and green tie, is seated in a chair. He is looking down. The background is a dark studio set with a large screen showing a crowd of people. A desk with the word "DAYS" is visible to the right.

Real Time with Bill Maher
July, 2014 Episode 324
Bill Maher & Reihan Salam

DAYS

Checking the Numbers

Bill Maher

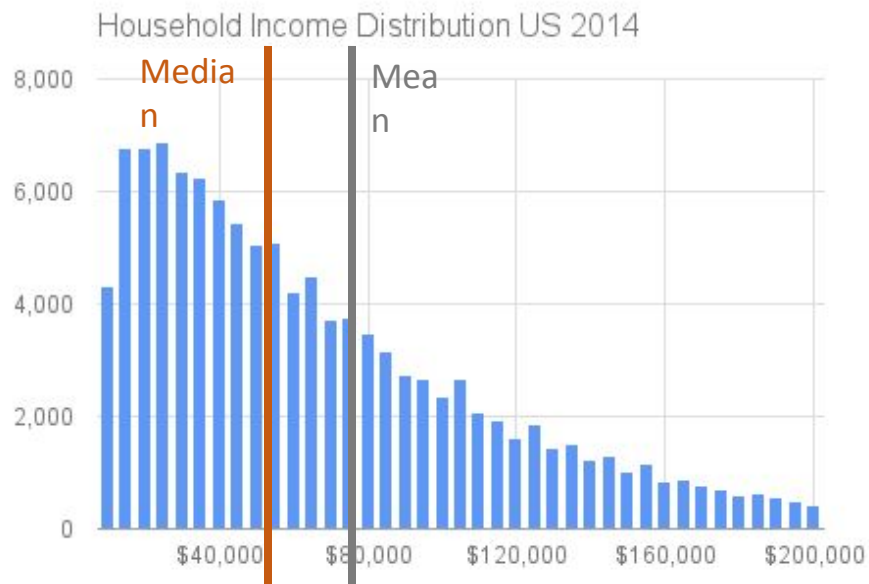
- Median Household Income
 - \$51,000
 - Actual: \$53,000

Reihan Salam

- Median Income: Family of 4
 - \$80,000
 - Actual: \$83,000
- Mean Income: Family of 4
 - \$100,000
 - Actual: \$108,000

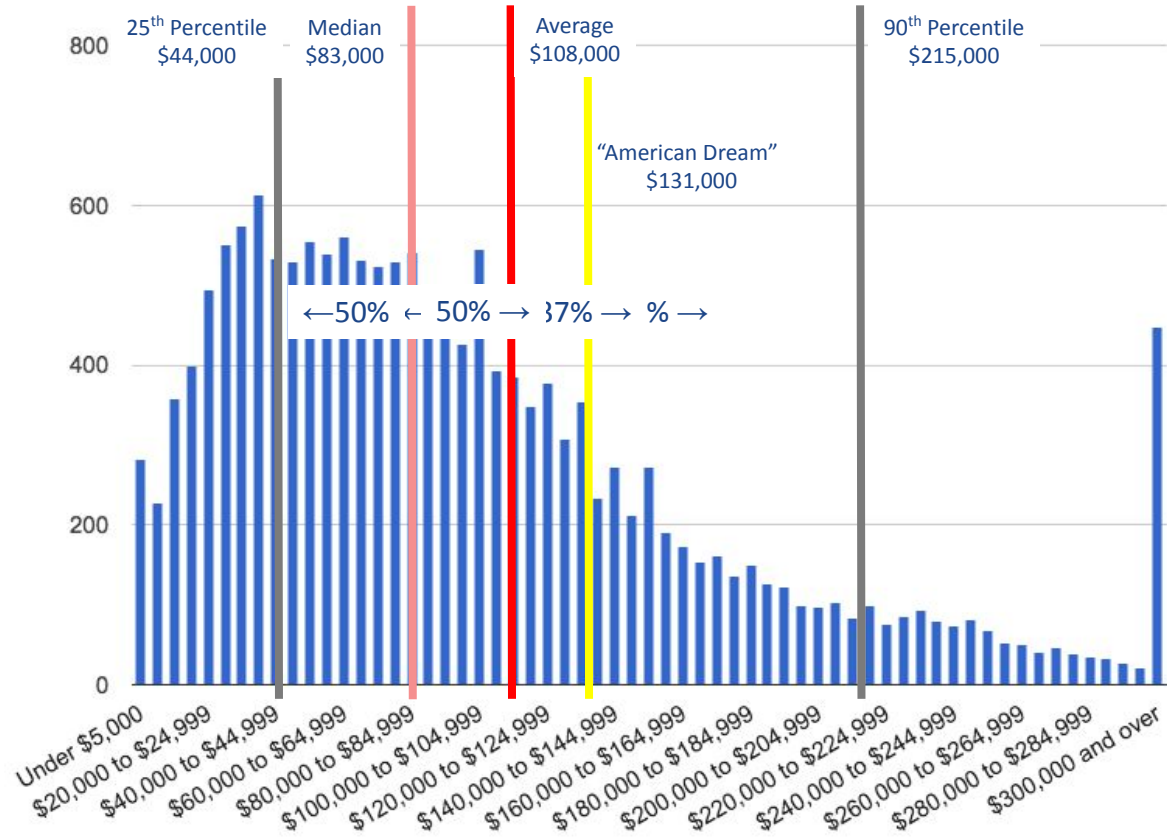
Actual Values taken from 2014 US Census Bureau Data

Looking at the Distribution



Households of 4 Total Income

2014



Another Example

- Two people have a total height of 11ft. What's the most likely guess for the heights of the two individuals?
 - 5 ft 6 in.
- Two people have a total income of \$36 million dollars. What's the most likely guess for the incomes of the two individuals?
 - \$0 and \$36,000,000

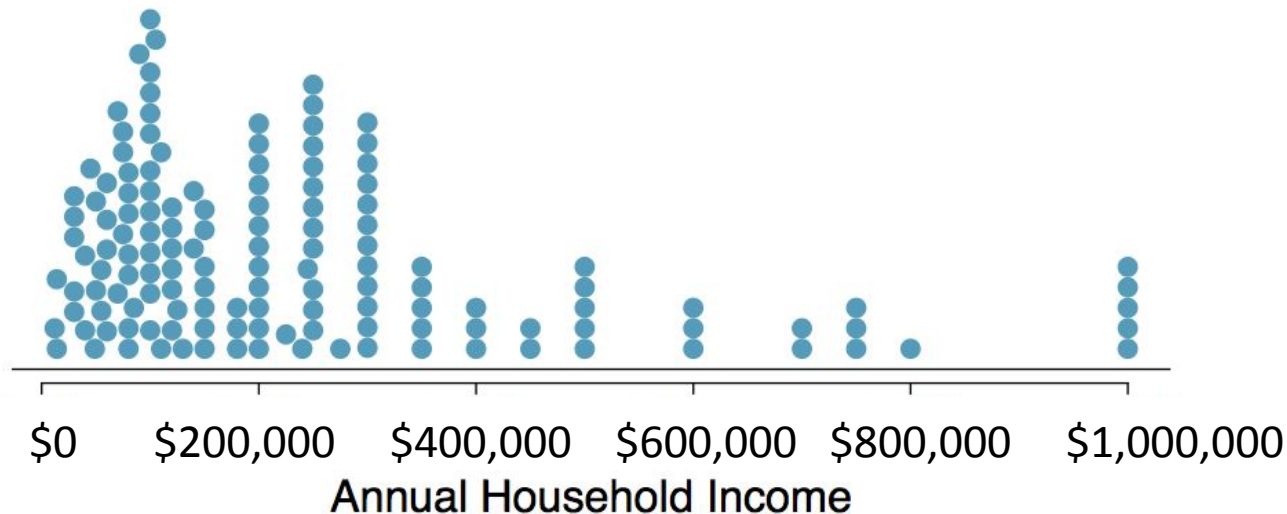
Outliers

Why is it important to look for outliers?

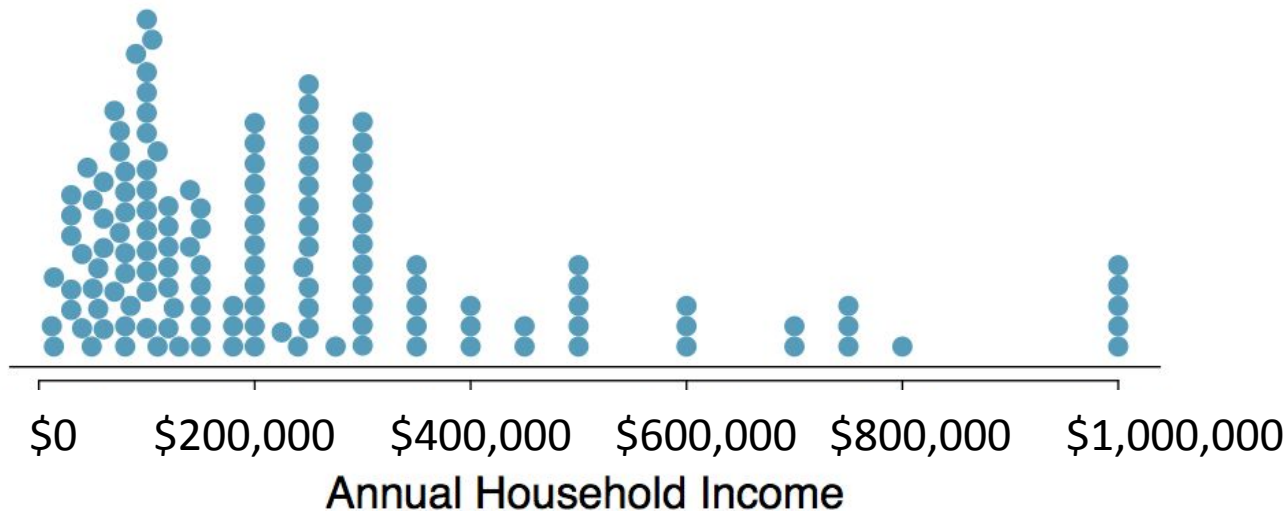
- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

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

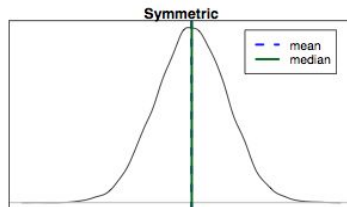
Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- 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

Mean vs. Median

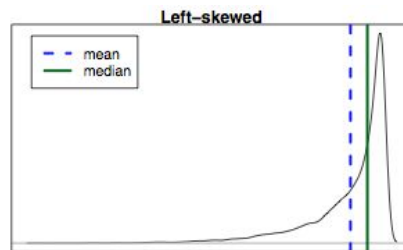
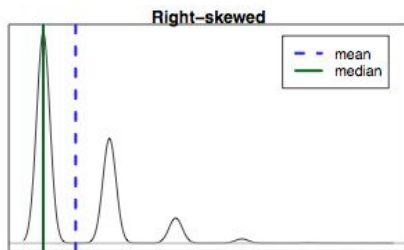
If the distribution is symmetric, center is often defined as the mean:

mean \sim median



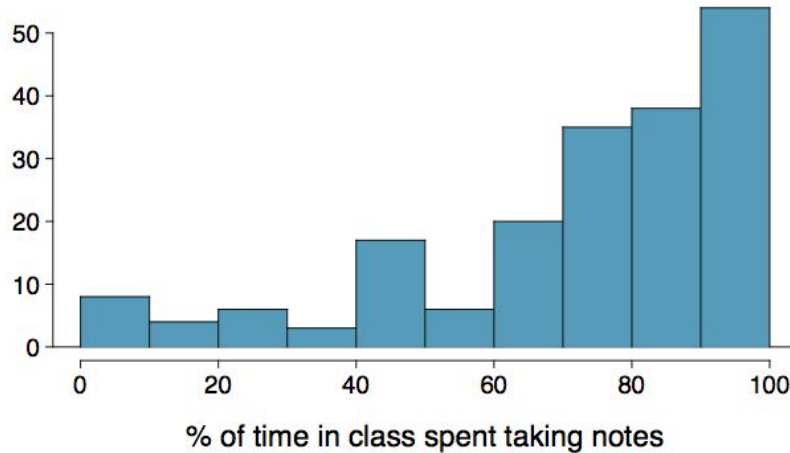
If the distribution is skewed or has extreme outliers, center is often defined as the median

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



Question

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

(b) mean ~ median

(c) mean < median

(d) impossible to tell

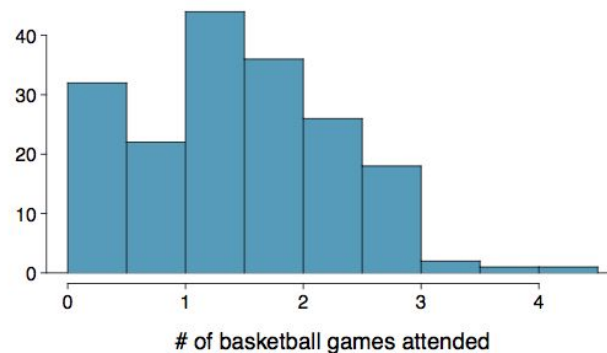
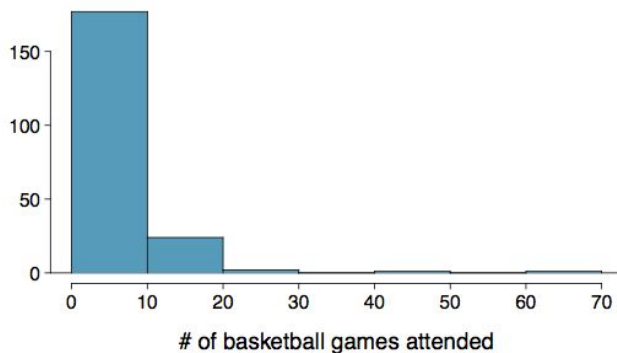
Data Transformations

- Transformation is a rescaling of the data using a function.
- When data are very strongly skewed, we sometimes transform them so they are easier to model.

Extremely Skewed Data

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.

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

Summary: Goals of Transformations

- To see the structure of the data differently.
- To reduce the skew to help modeling.
- To straighten a nonlinear relationship in a scatterplot.

Summarizing or Exploring a Numerical Dataset

- Focus on 5 aspects (at least) for Numerical Data
 - Center
 - Spread
 - Skew
 - Clusters/Modality
 - Extreme Values
- Always look at multiple data representations

Try It

- Find a numeric variable in one of the Fivethirtyeight or Openintro Data sets and explore and summarize it.
 - Make sure to try to describe each of the 5 aspects on the previous slide.

Anticipating Shape, Symmetry & Skew

- What modality and symmetry would you expect from the following data sets.
 - Heights of women in the U.S.
 - Income of individuals in California
 - Outcomes from rolling a 20 sided die 10,000 times
 - Magnitudes of earthquakes that occurred between 2010 and 2018
 - Weights of elephants worldwide

Anticipating Shape, Symmetry & Skew

- If data values have a hard limit on one side but not the other, they are more likely to be skewed (especially if the center of the distribution is near the limit)
 - For example:
 - You can't earn less than \$0
 - You can't score more than 100% on a test
- If the center is far from any limits, the distribution is likely to be symmetric.
 - For example:
 - You can't be less than 0 inches tall, but average heights are not at all close to 0. Heights are usually distributed symmetrically

Exploring Nominal/Categorical Variables

General Social Survey

Year	Marital Status	Age	Race	Income	Party ID	Religion	Denomination
2004	Married	61	White	Not applicable	Strong democrat	Catholic	Not applicable
2008	Married	84	White	Not applicable	Strong democrat	Protestant	United methodist
2010	Married	61	White	Refused	Not str republican	Catholic	Not applicable
2008	Never married	26	Other	\$25000 or more	Strong democrat	Catholic	Not applicable
2014	Never married	26	White	\$25000 or more	Ind,near rep	Other	Not applicable
2006	Married	51	White	\$25000 or more	Independent	None	Not applicable
2014	Married	65	White	\$25000 or more	Not str democrat	Buddhism	Not applicable
2002	Widowed	84	White	Not applicable	Not str republican	Protestant	No denomination
2002	Divorced	37	White	Lt \$1000	Independent	Catholic	Not applicable

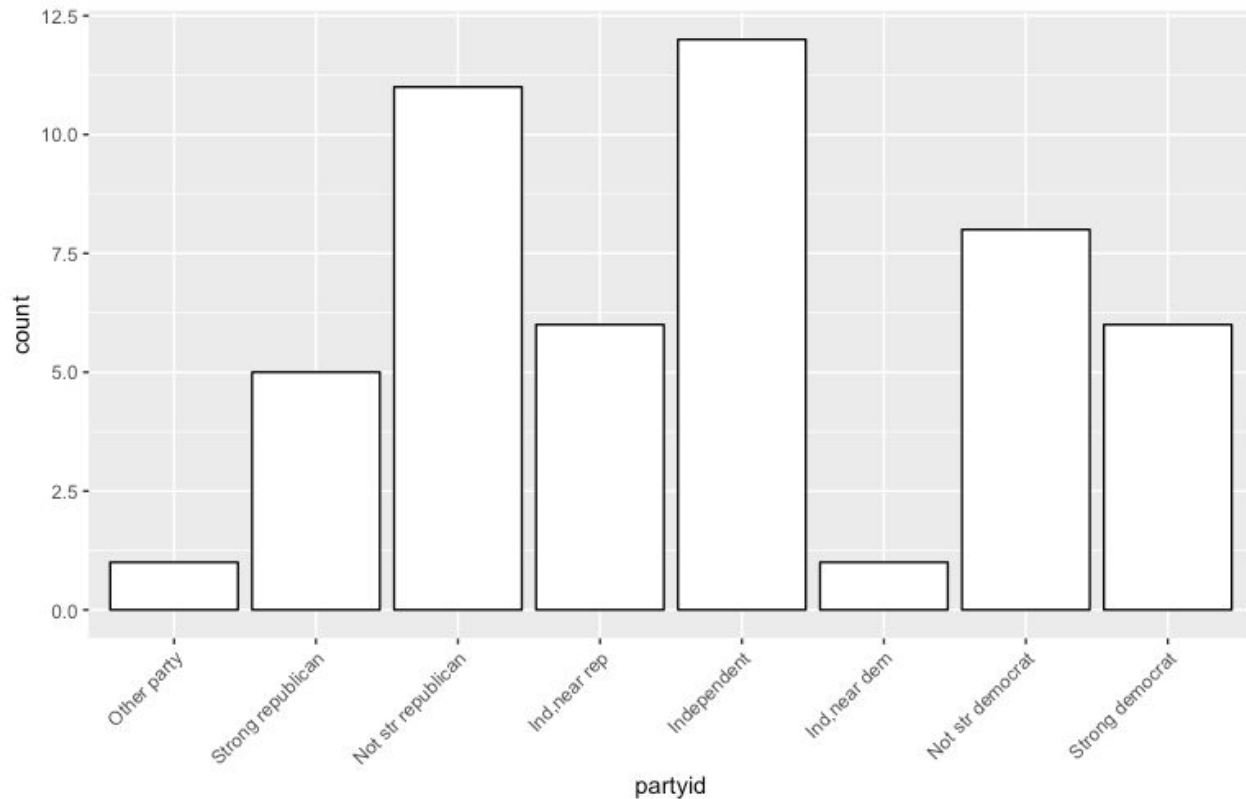
Categorical Data

Means, Medians, sd, etc. are not applicable

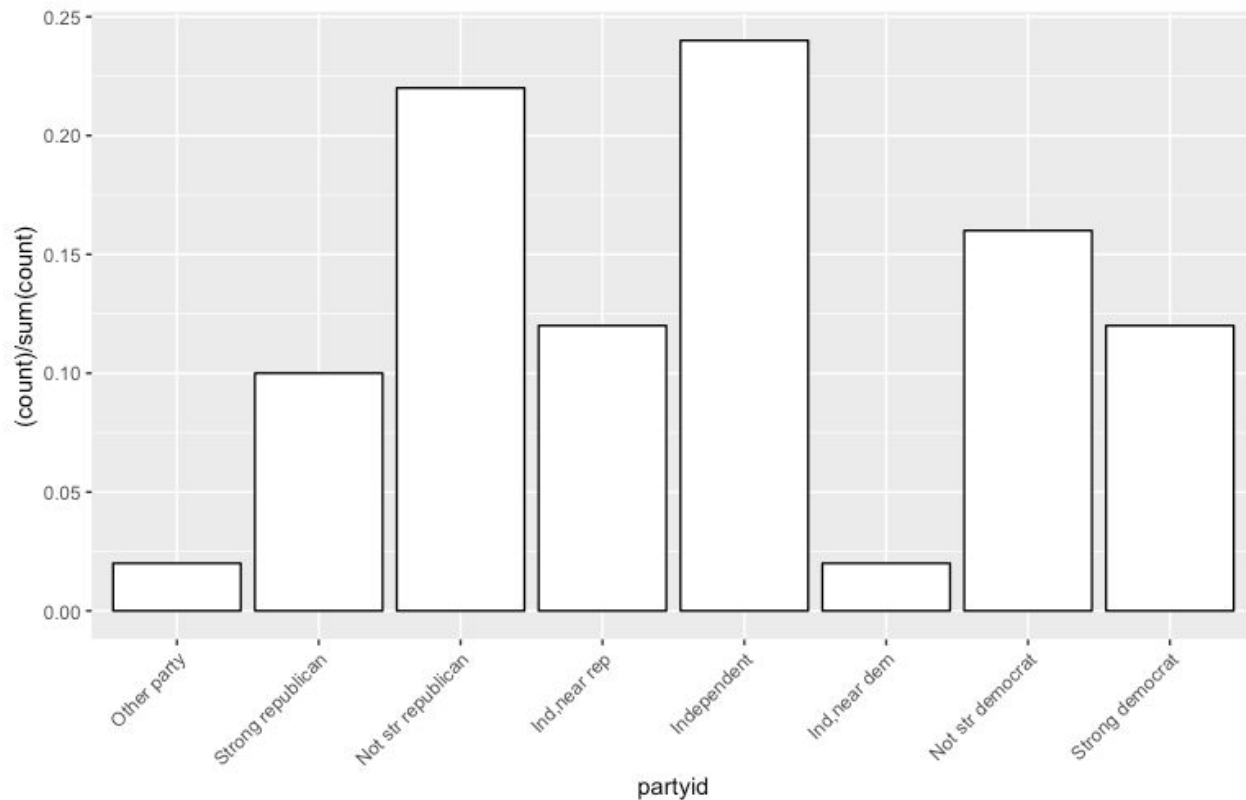
Instead:

- **Counts:** How many are in each group?
- **Proportions:** What % of the total are in each group?

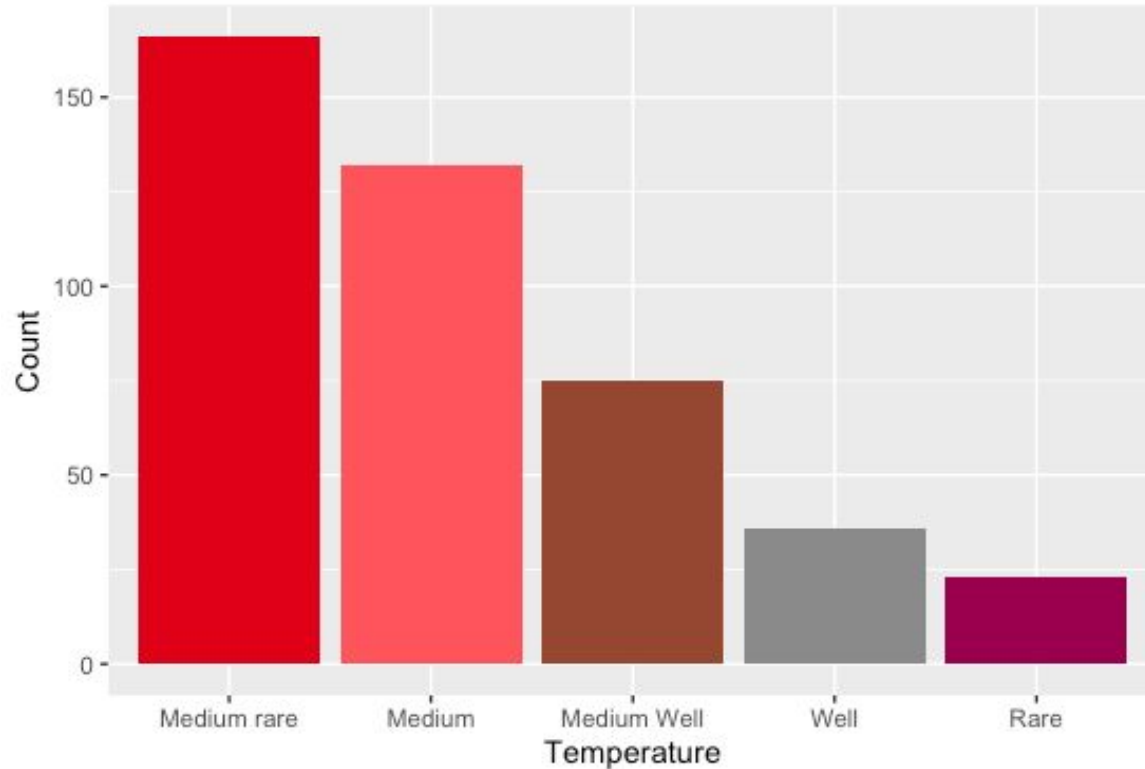
Visualizing Counts: Bar Charts



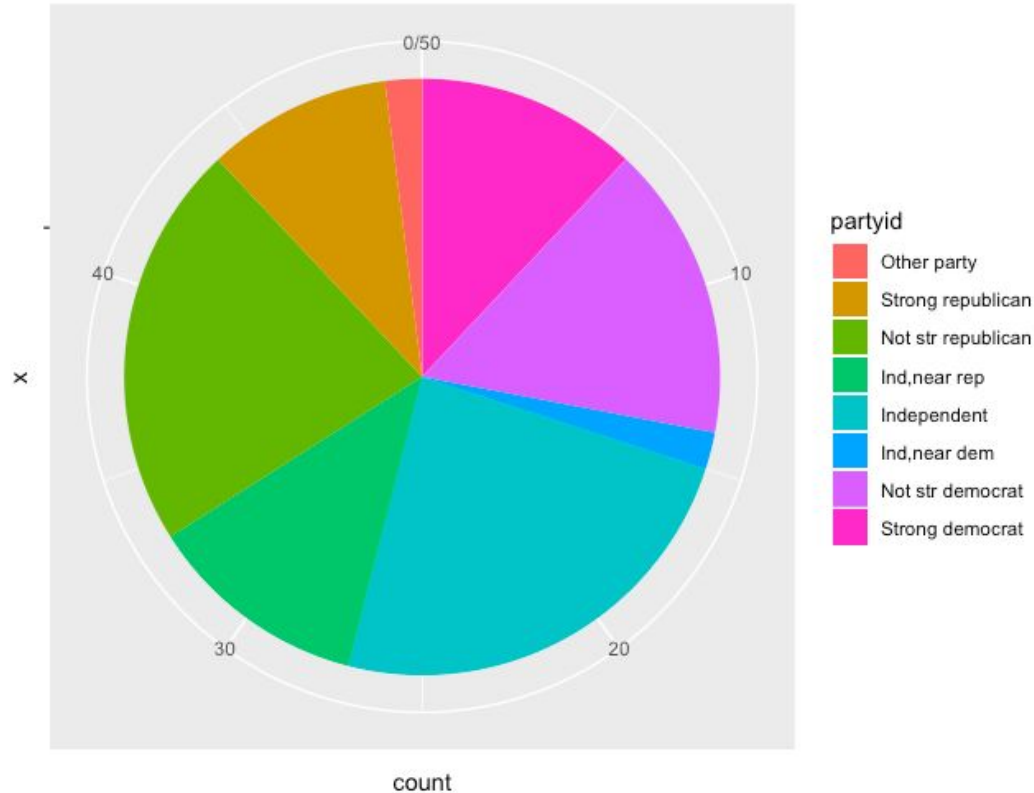
Visualizing Proportions: Bar Charts



Bar Charts: Another Example



Visualizing Proportions: Pie Charts



Relationships Between Variables:

Contingency Tables

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

	Rare	Medium Rare	Medium	Medium Well	Well	Total
No Smoke	16	136	111	63	30	356
Smoke	6	30	20	11	5	72
Total	22	166	131	74	35	428

Relationships Between Variables:

Contingency Tables

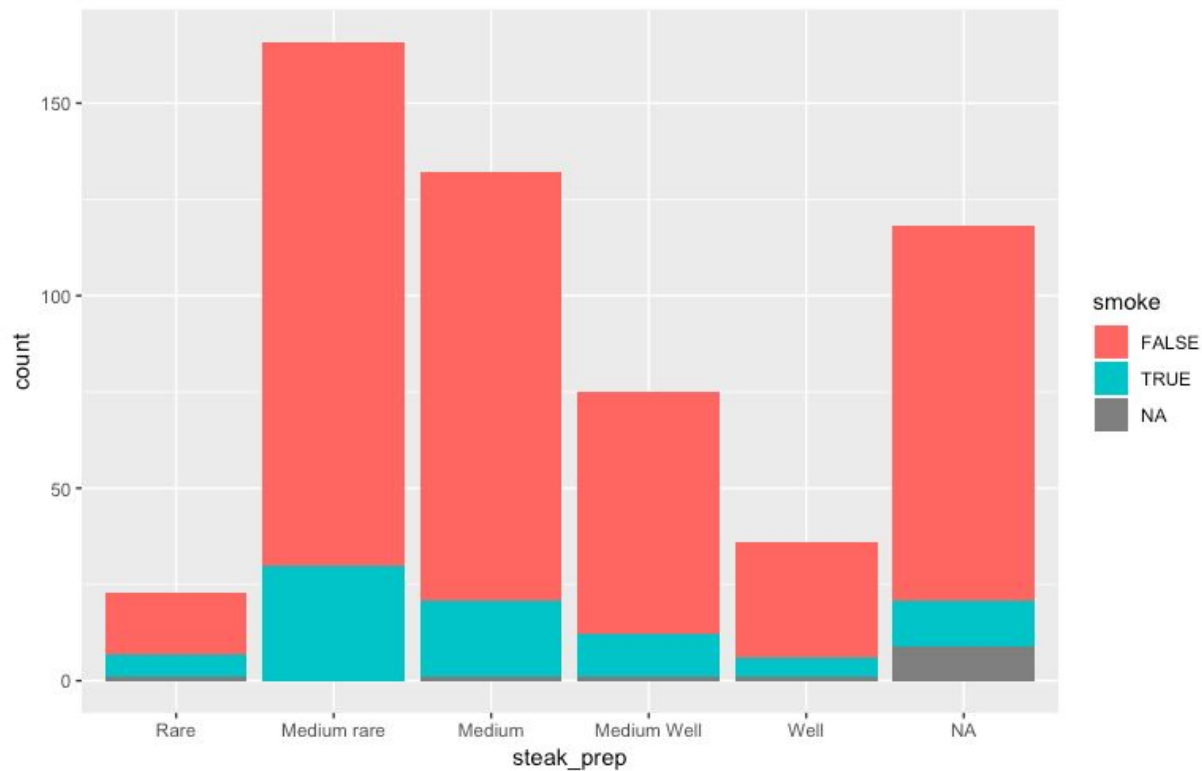
Does there appear to be a relationship between smoking and preference in steak preparation?

	Rare	Medium Rare	Medium	Medium Well	Well	Total
No Smoke	16	136	111	63	30	356
Smoke	6	30	20	11	5	72
Total	22	166	131	74	35	428

Contingency Table Proportions

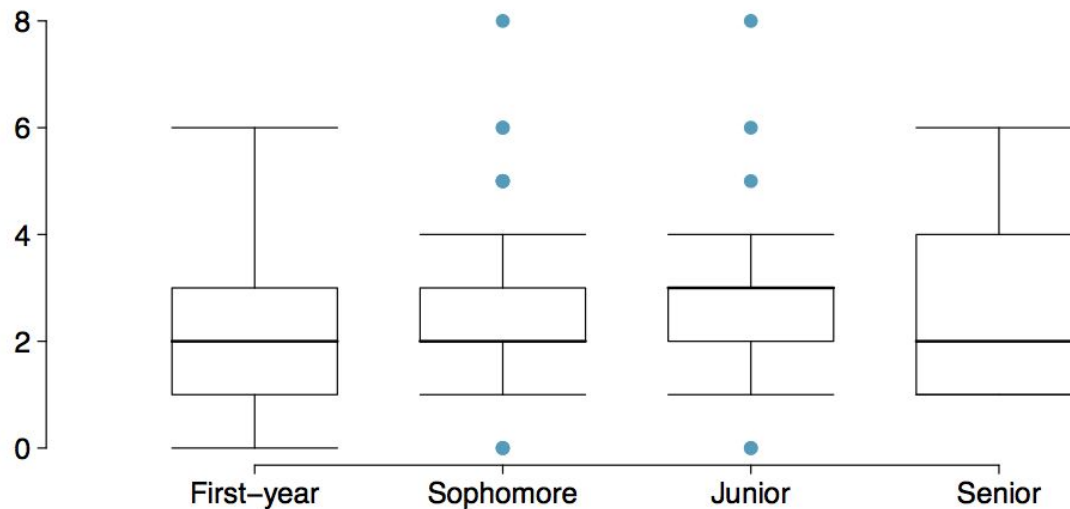
- Proportions can be calculated in three ways:
 - Based on overall totals (What % of the whole?)
 - Based on row totals (What % of smokers?)
 - Based on column totals (What % of Rare Steak Lovers?)
- Be very careful of the denominator!!

Segmented Bar Charts



Side by Side Box Plot

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



box plot of number of clubs college students are involved with and their class year

Data Analysis Step 1: Explore & Describe A Single Variable

Quantitative

- Center, Spread, Shape, Symmetry, Extreme Values
- Histogram
- Boxplot
- other (e.g. dotplot, lineplot...)

Qualitative

- Mode
- Proportions for each level
- Frequency Table
- Bar Chart
- Pie chart (last resort)

Data Analysis Step 1: Explore & Describe

Two Variables: Both Categorical

- Contingency Table
- Double Bar Chart
- Segmented Bar Chart
- Mosaic Plot

Are they independent?

Independence

Two variables are

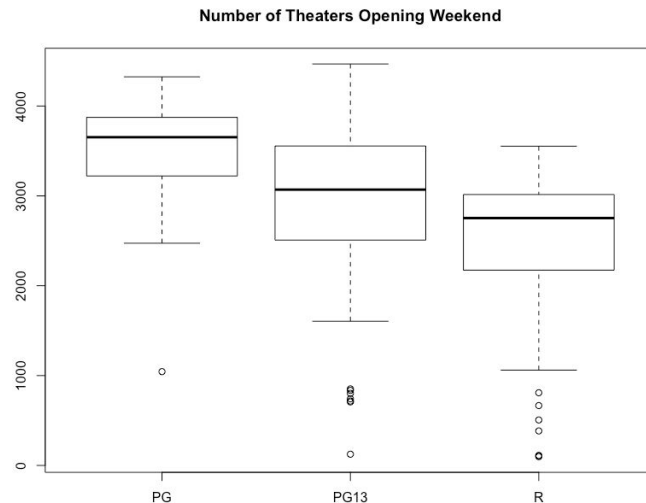
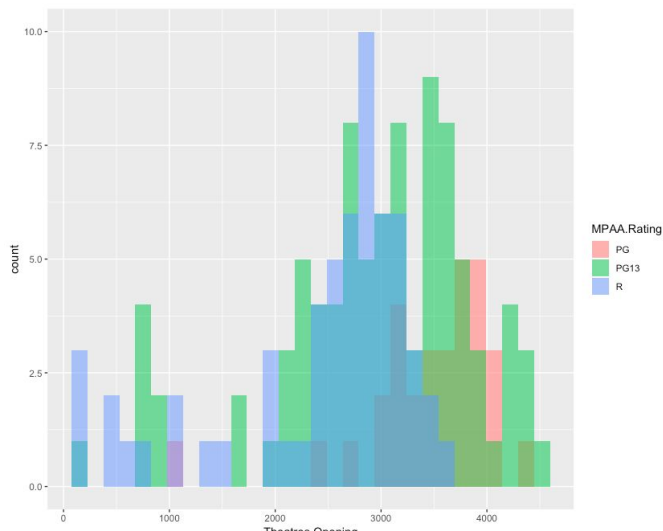
- **Independent** if they do not affect each other.
 - A change in one variable doesn't imply a change in another.
 - Patterns across levels of a variable are all similar
- **Dependent** if they do affect each other.
 - The variables change together in some way
 - Patterns for one level are different from patterns in another level.

Data Analysis Step 1: Explore & Describe

Two Variables: One Categorical, One Numeric

- Side by side boxplots
- Transparent Histograms

Are they independent?



Data Analysis Step 1: Explore & Describe

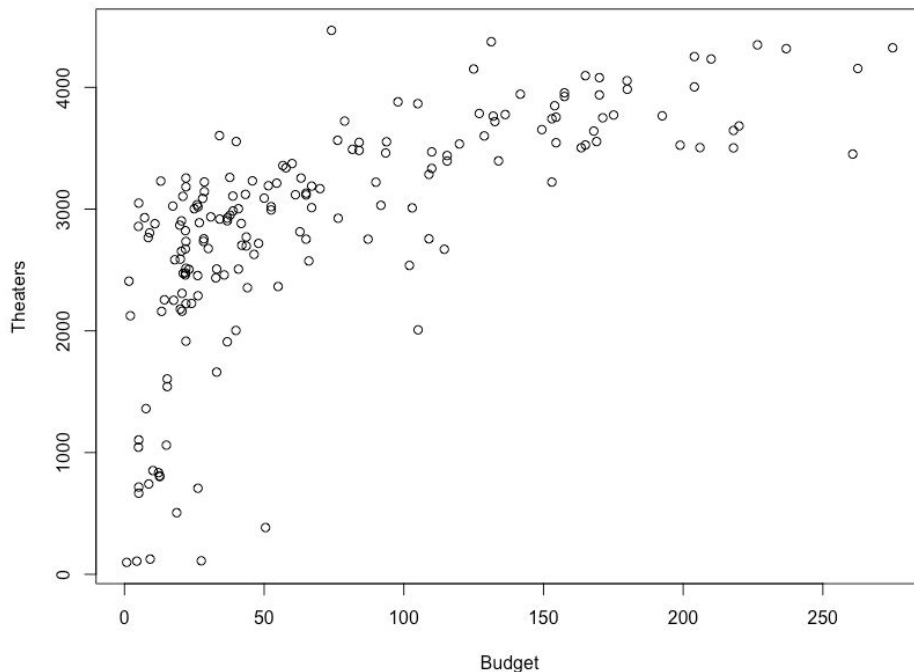
Two Variables: Both Numeric

- Scatterplot

Are they independent?

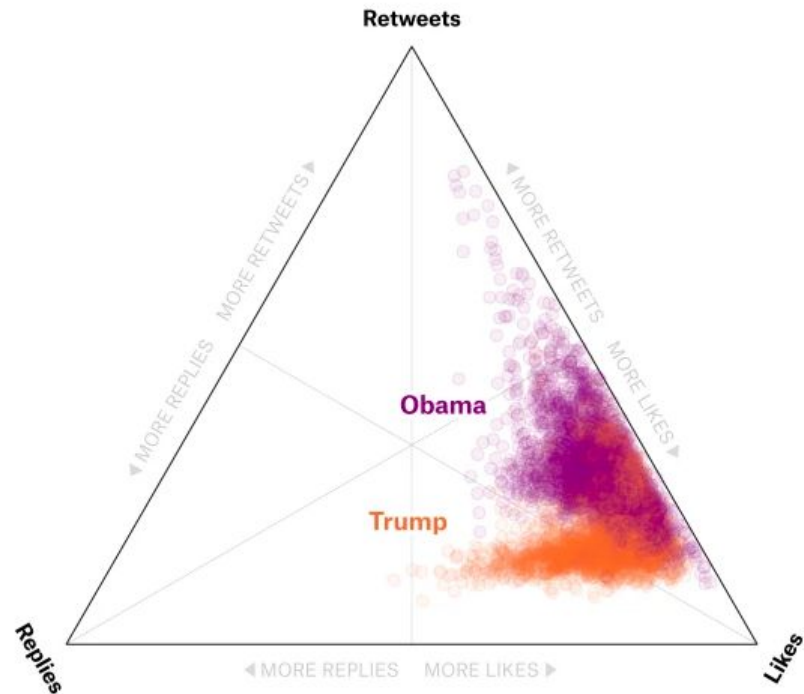
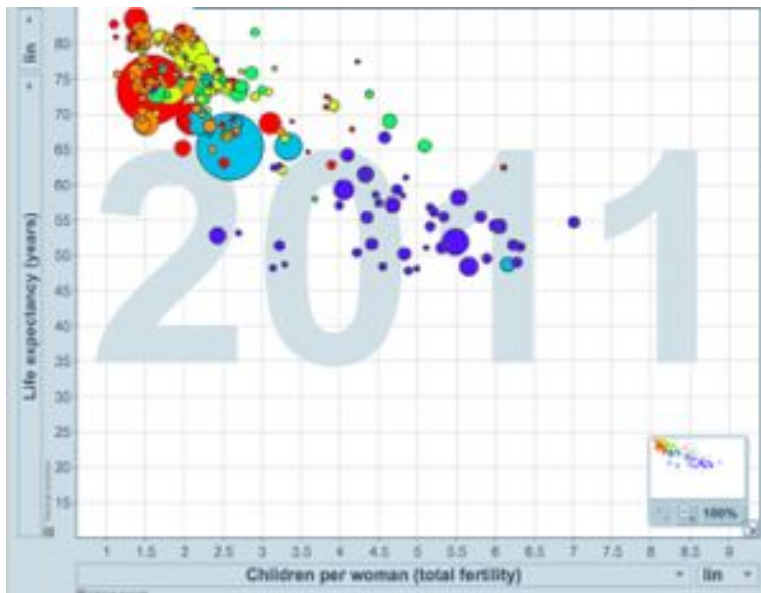
How do we describe the relationship?

- Fairly strong
- Positive
- Non-linear



Data Analysis Step 1: Explore & Describe More than Two Variables

Be creative!



When two variables are related

- Two variables are related when they change together.
 - Different values of variable A correspond with different values of variable B.
- When variables are related, we can do two things:
 - Describe the relationship
 - How strong is the relationship?
 - What pattern does the relationship show?
 - Make inferences
 - Is this relationship predictive of the larger population?
 - Does this relationship indicate the possibility of a cause-effect relationship.

When two variables are related

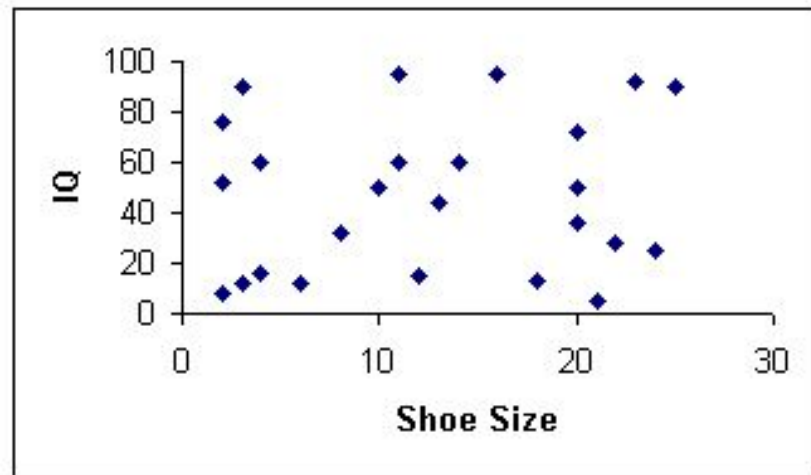
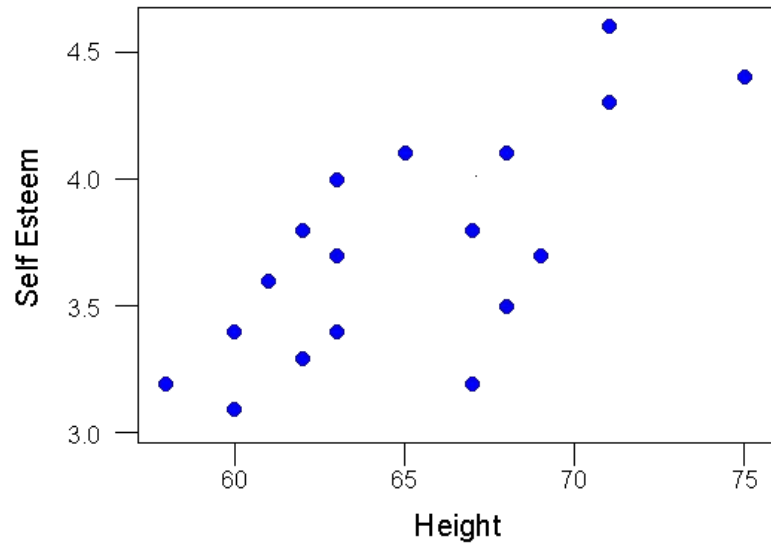
- Two variables are related when they change together.
 - Different values of variable A correspond with different values of variable B.
- When variables are related, we can do two things:
 - Describe the relationship
 - How strong is the relationship?
 - What pattern does the relationship show?
 - Make inferences
 - Is this relationship predictive of the larger population?
 - Does this relationship indicate the possibility of a cause-effect relationship.

LATER...

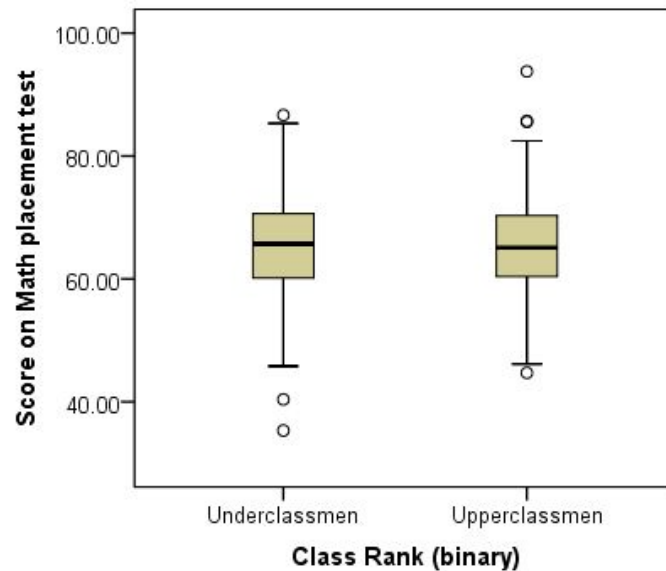
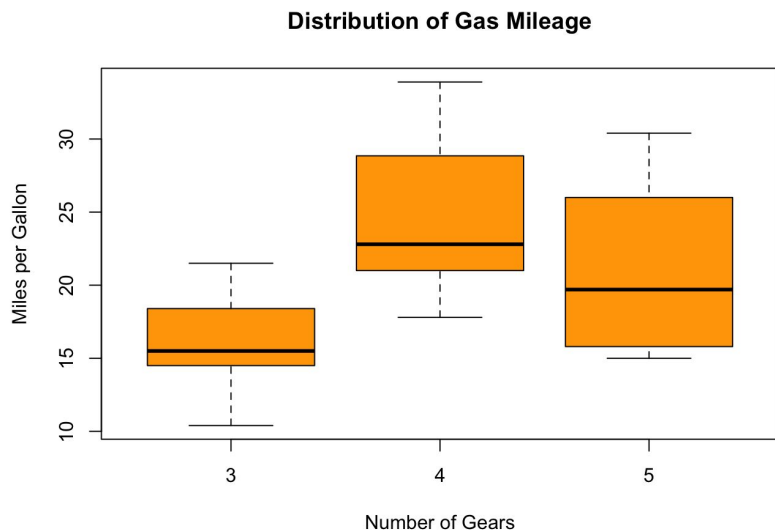
Relationships between Two Quantitative Variables

- Strength: Strong or Weak?
 - A strong relationship is indicated by how close to the trend the data fall
- Direction: Positive, Negative
 - Do the variables change in the same direction or opposite directions?
- Shape: Linear, Curved
- Outliers

Relationships in Variables



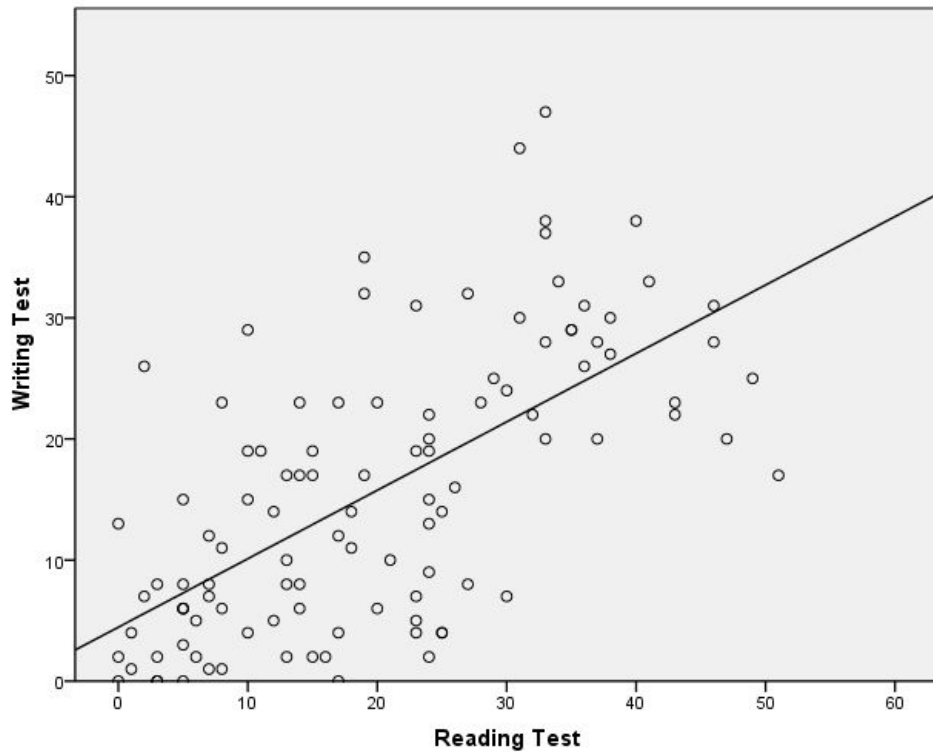
Relationships between Variables



Associated vs. independent

- When two variables show some connection with one another, they are called associated variables.
 - Associated variables can also be called dependent variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be independent.

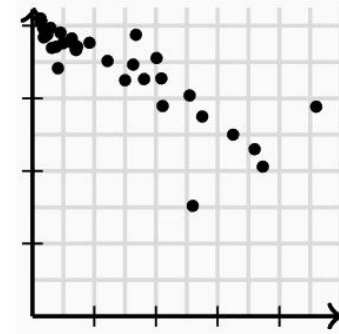
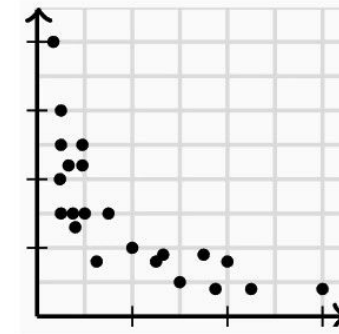
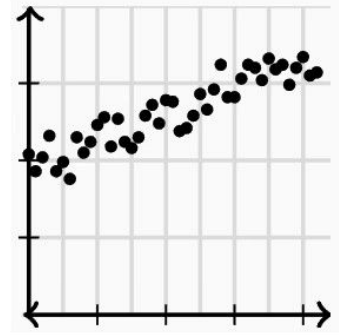
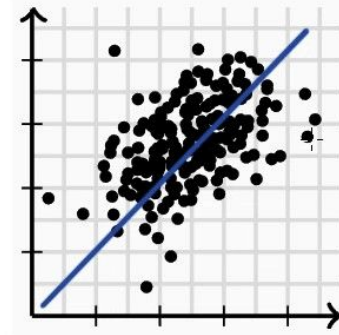
Example



Linear
Weak
Positive
Few extreme
values

More Examples

- Which ones are strong? Weak?
- Which ones are positive? negative?
- Which ones are linear? Curved?
- Which ones have outliers?



Identifying Relationships

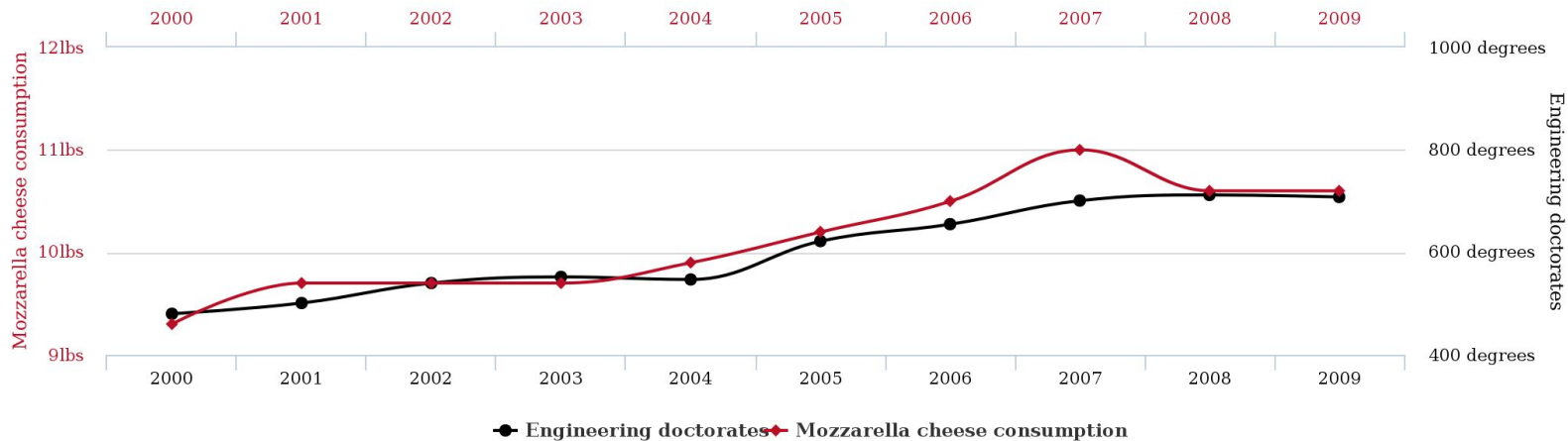
- To identify relationships we create studies
 - Correlations are relatively straightforward to find
 - Causation is extremely difficult

Correlation \neq Causation

Per capita consumption of mozzarella cheese

correlates with

Civil engineering doctorates awarded



Descriptive Statistics in the Wild

The Washington Post

Democracy Dies in Darkness

North America has lost 29% of its birds in 50 years

A sweeping new study says a steep decline in bird abundance, including among common species, amounts to "an overlooked biodiversity crisis."

By Karin Brulliard

The New York Times

3 Billion North American Birds Have Vanished: 'It's Just Staggering'

The number of birds in the United States and Canada has declined by 3 billion, or 29 percent, over the past half-century, scientists find.

7m ago [466 comments](#)