Introduction to Statistics I

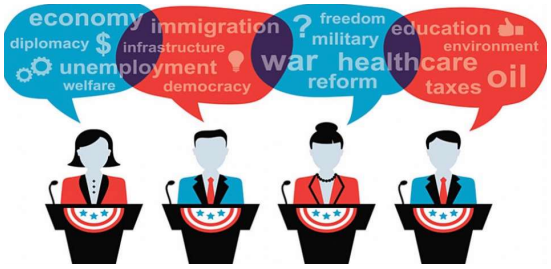
Textbook: Elementary Statistics (4th Edition, by Navidi and Monk), and Workshop Statistics (4th Edition by Rossman and Chance).

Previous Lecture

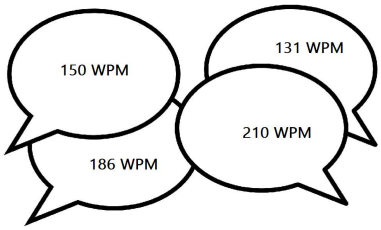
- Dot Plots & Distr's
- Distributions: shape/center/spread, skew, symmetry, outliers
- Stem-and-Leaf Plots
- Histogram: bins/frequencies
- Time Series

§3.1: Measures of Center

Debating Speeds of Presidential Candidates 2007



GOP Candidate	Rate
Sam Brownback	196
Rudy Giuliani	206
Mike Huckabee	207
Duncan Hunter	189
John McCain	175
Ron Paul	187
Mitt Romney	221
Tom Tancredo	196
Fred Thompson	182



How should we graph?

Dem. Candidate	Rate
Joe Biden	190
Hilary Clinton	183
Chris Dodd	222
John Edwards	201
Mike Gravel	174
Dennis Kucinich	189
Barack Obama	187
Bill Richardson	161

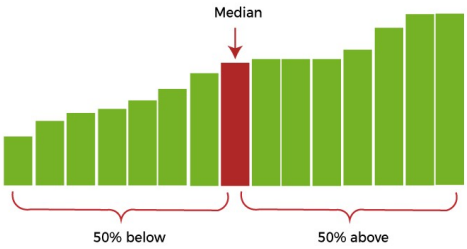
Republican		Democratic
	16	1
5	17	4
972	18	379
66	19	0
76	20	1
	21	
1	22	2

How should we compare their "centers"?

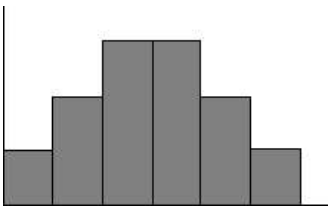
Measures of Center

- **Mean** is the "ordinary" average. Add up all values, divide by sample size.

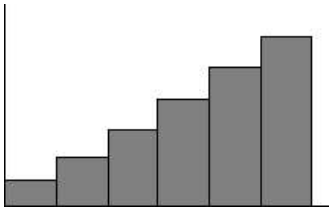
- **Median** is "middle-most" value, when arranged in numerical order. If two values are "middle-most" (when sample size is even), average those two values. Half the values are greater than the **median**, and half are less.



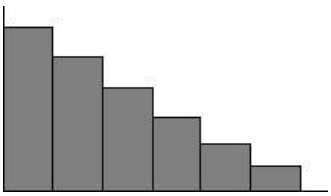
Skewness Detector (Usually works. Some exceptions exist)!



Symmetric: *mean* = *median*.



Skew left: *mean* < *median*.



Skew right: *median* < *mean*.

Activity 3.1a

Calculating Mean Speeds

Recall Dems: 190, 183, 222, 201, 174, 189, 187, 161. GOP: 196, 206, 207, 189, 175, 187, 221, 196, 182

Dems: Mean is ??

$$\frac{190+183+222+201+174+189+187+161}{8} = 188.375.$$

GOP: Mean is $\frac{196+206+207+189+175+187+221+196+182}{9} = 195.44.$

Before calculating the median, we need to ... ??

... put the data in order.

Dems: 161, 174, 183, 187, 189, 190, 201, 222. GOP: 175, 182, 187, 189, 196, 196, 206, 207, 221

Dem median: $\frac{187+189}{2} = 188.$

(when sample size is even, average the two center values to find median)

GOP median is 196.

(when sample size is odd, select middle-most value)

Who's faster: In both cases, GOP had faster debate speeds on average.

Mean vs Median

	GOP	Dems
Mean	195.44	188.38
Median	196	188

This data was chosen to point out the possibility of an exception.

Here, the GOP appears right skewed, but the mean is to the left of the median.

Mean & median here look similar - why have both, what do they represent?

Example: Suppose you are taking donations.

Nine people each bring a dollar, and one person brings 91 dollars.

So data looks like: 1 1 1 1 1 1 1 1 1 91. Mean/median?

Mean amount brought: $\frac{100}{10}$ or \$10.

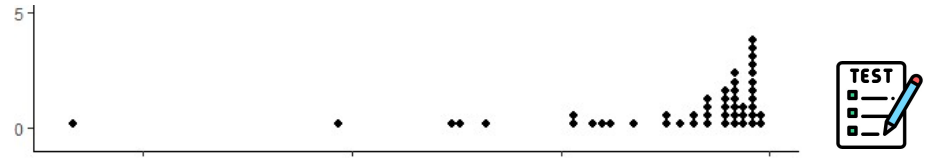
But median is \$1, so "typical" guest brought only \$1.

Mean is amount each person would get if you collected everyone's money and gave each person back an equal amount.

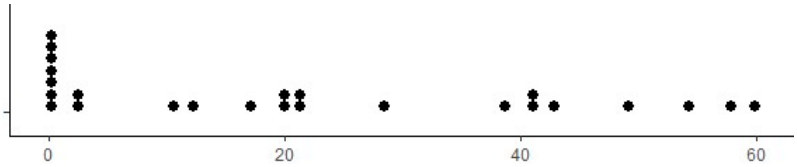
Median is amount exhibited by the "typical" (middlemost) person.

Mean & **Median** don't have to be same thing (and usually aren't).

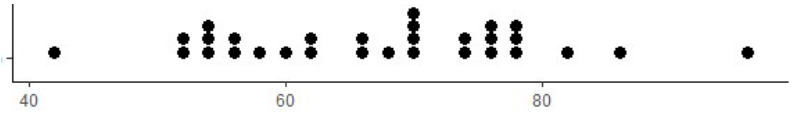
Data Examples: consider the following graphs. How does mean compare to median?



Quiz Percents



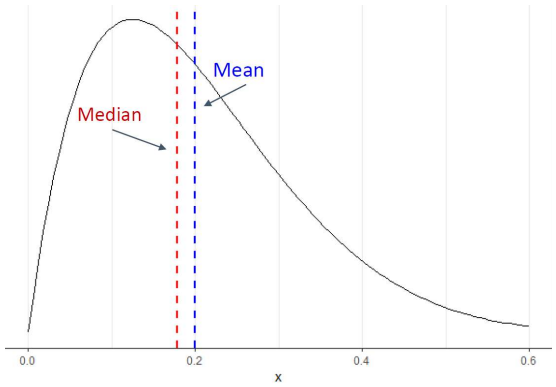
Snowfall



Blood Pressure

Shape and Measures of Center

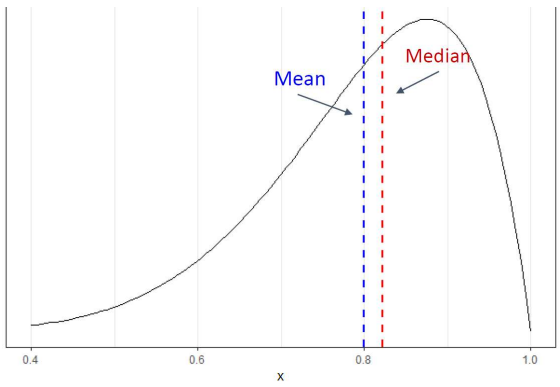
	Shape	Mean	Median	
Quiz Percents	L. Skewed	88.9	95	Left skewed: Mean < Median
Snowfall	R. Skewed	21.65	19.9	Right skewed: Mean > Median
Blood Pressure	Symmetric	68.8	69	Symmetric: Mean = Median (approximately)



Right Skewed Data

Mean is pulled from max in direction of the tail,

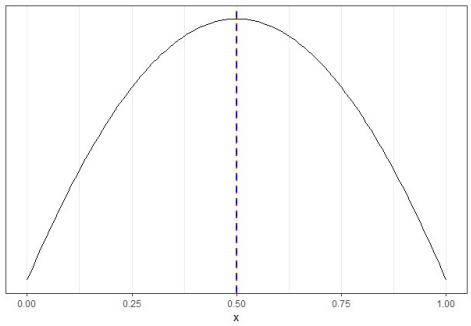
so mean is larger than median.



Left Skewed Data

Mean is pulled from max in direction of the tail,

so mean is smaller than median.



Symmetric Data

Mean is not pulled in either direction,

so mean and median (and max) coincide.

Rowers Activity

(Activity 3.1b: On Canvas, insufficient class time to do here. Good practice, though.)

Use applet and rower data to fill in the table below.

bit.ly/introstatsdata

Applets: DotPlots

Data: Rower

Mean vs. Median

Mean is heavily influenced by outliers because it sums all values when averaging.

Median is **resistant to outliers**. Same # of values are both greater and less than the median.
The median **doesn't care about magnitude** of those values.

Mean is computationally quicker to compute. It also has some nice mathematical properties.

So this course focuses more on mean than median.

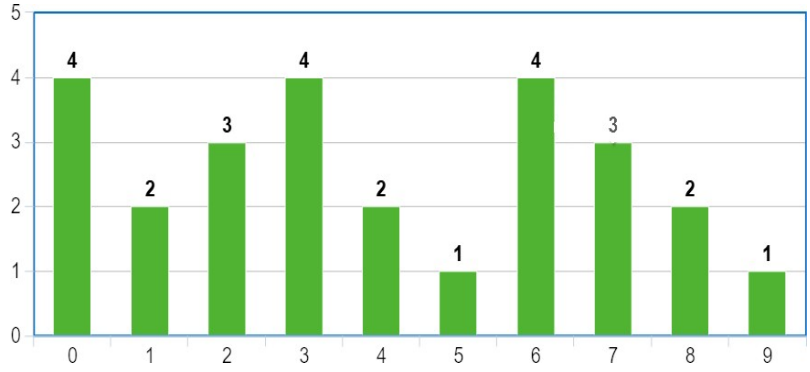
Mode

The modes of the data set are the values that appear most frequently.

Keep in mind, if several values are tied for how frequently they occur, you can have several modes.

Example: Examine the modes of the following data: 0 0 0 0 1 1 2 2 2 3 3 3 3 4 4 5 6 6 6 6 7 7 7 8 9 9.

You will notice that $0, 3, 6$ occur most frequently, they each occur 4 times. The other values occur less frequently (3 or less). Therefore, the modes are $0, 3,$ and 6.



What did we learn?

- Defined: Mean, Median
- Distr Shape: Symmetric/Skewed Data
- Mean vs. Median
- Mode