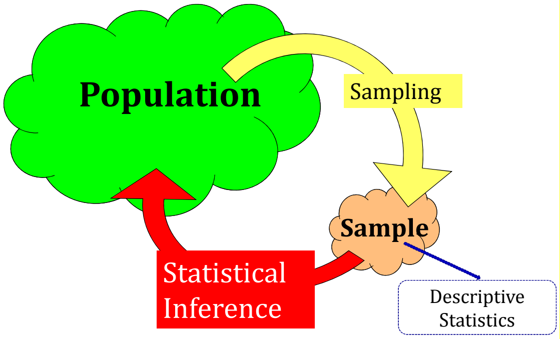
STAT 250
Dr. Kari Lock Morgan

Describing Data: One Quantitative Variable

SECTIONS 2.2, 2.3

- One quantitative variable (2.2, 2.3)

The Big Picture



Population

Sampling

Sample

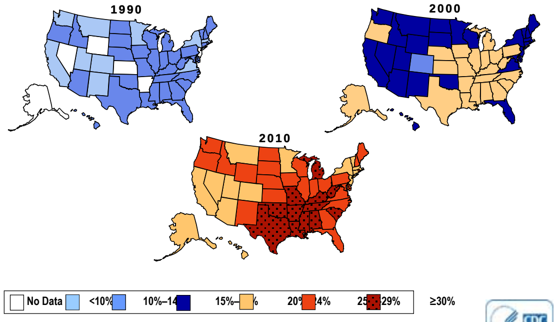
Statistical Inference

Descriptive Statistics

Descriptive Statistics

- In order to make sense of data, we need ways to *summarize* and *visualize* it
- Summarizing and visualizing variables and relationships between two variables is often known as ***descriptive statistics*** (also known as *exploratory data analysis*)
- Type of summary statistics and visualization methods depend on the type of variable(s) being analyzed (categorical or quantitative)
- Today: One quantitative variable

Obesity Trends* Among U.S. Adults
BRFSS, 1990, 2000, 2010
(*BMI ≥30, or about 30 lbs. overweight for 5' 4" person)

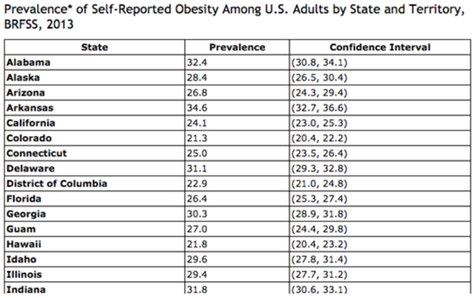


Source: Behavioral Risk Factor Surveillance System, CDC.

Obesity in America

- Obesity is a HUGE problem in America
- We'll explore this with two different types of data, both collected by the CDC:
  - ○ Proportion of adults who are obese in each state
  - ○ BMI for a random sample of Americans

Behavioral Risk Factor Surveillance System

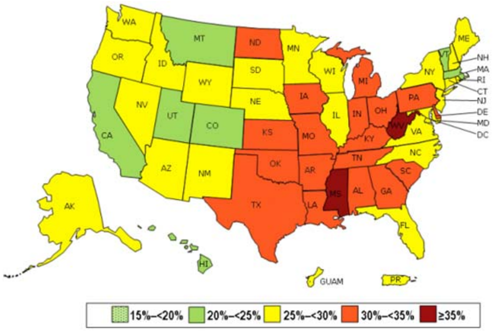


Prevalence* of Self-Reported Obesity Among U.S. Adults by State and Territory, BRFSS, 2013

State	Prevalence	Confidence Interval
Alabama	32.4	(30.8, 34.1)
Alaska	28.4	(26.5, 30.4)
Arizona	26.8	(24.3, 29.4)
Arkansas	34.6	(32.7, 36.6)
California	24.1	(23.0, 25.3)
Colorado	21.3	(20.4, 22.2)
Connecticut	25.0	(23.5, 26.4)
Delaware	31.1	(29.3, 32.8)
District of Columbia	22.9	(21.0, 24.8)
Florida	26.4	(25.3, 27.4)
Georgia	30.3	(28.9, 31.8)
Guam	27.0	(24.4, 29.8)
Hawaii	21.8	(20.4, 23.2)
Idaho	29.6	(27.8, 31.4)
Illinois	29.4	(27.7, 31.2)
Indiana	31.8	(30.6, 33.1)

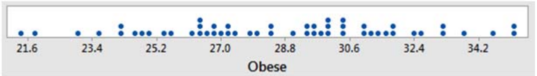
http://www.cdc.gov/obesity/data/table-adults.html

Obesity by State



Dotplot

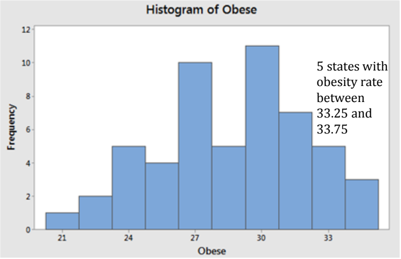
- In a ***dotplot***, each case is represented by a dot and dots are stacked.
- Easy way to see each case



Minitab: Graph -> Dotplot -> One Y -> Simple

Histogram

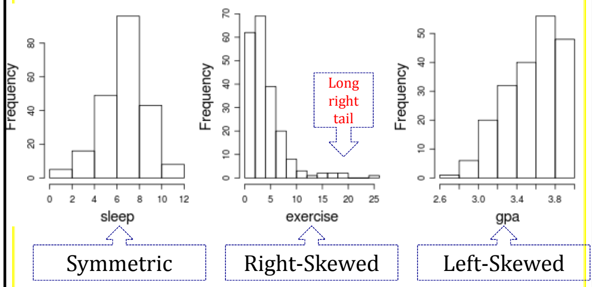
- The height of the each bar corresponds to the number of cases within that range of the variable



5 states with obesity rate between 33.25 and 33.75

Minitab: Graph -> Histogram -> Simple

Shape



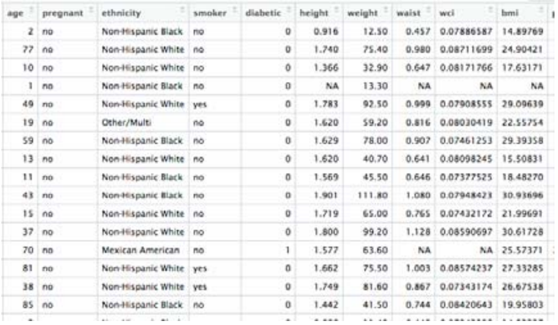
Long right tail

Symmetric

Right-Skewed

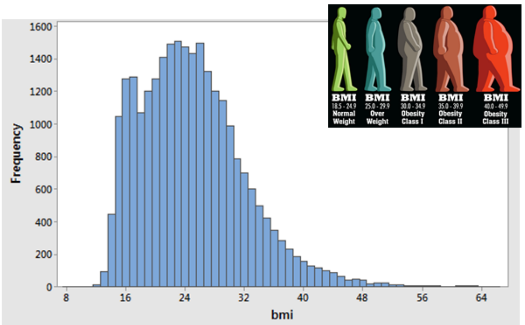
Left-Skewed

National Health and Nutrition Examination Survey



age	pregnant	ethnicity	smoker	diabetic	height	weight	waist	wci	bmi
2	no	Non-Hispanic Black	no	0	0.916	12.50	0.457	0.07886587	14.89769
77	no	Non-Hispanic White	no	0	1.740	75.40	0.980	0.08711699	24.90421
10	no	Non-Hispanic White	no	0	1.366	32.90	0.647	0.08171766	17.63171
1	no	Non-Hispanic Black	no	0	NA	13.30	NA	NA	NA
49	no	Non-Hispanic White	yes	0	1.783	92.50	0.999	0.07908555	29.09639
19	no	Other/Multi	no	0	1.620	59.20	0.816	0.08030419	22.55754
59	no	Non-Hispanic Black	no	0	1.629	78.00	0.907	0.07461253	29.39358
13	no	Non-Hispanic White	no	0	1.620	40.70	0.641	0.08098245	15.50831
11	no	Non-Hispanic Black	no	0	1.569	45.50	0.646	0.07377525	18.48270
43	no	Non-Hispanic Black	no	0	1.901	111.80	1.080	0.07948423	30.93696
15	no	Non-Hispanic White	no	0	1.719	65.00	0.765	0.07432172	21.99691
37	no	Non-Hispanic White	no	0	1.800	99.20	1.128	0.08590697	30.61728
70	no	Mexican American	no	1	1.577	63.60	NA	NA	25.57371
81	no	Non-Hispanic White	yes	0	1.662	75.50	1.003	0.08574237	27.33285
38	no	Non-Hispanic White	yes	0	1.749	81.60	0.867	0.07343174	26.67538
85	no	Non-Hispanic Black	no	0	1.442	41.50	0.744	0.08420643	19.95803

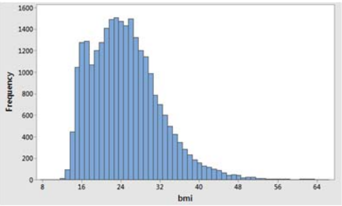
BMI of Americans



BMI of Americans

The distribution of BMI for American adults is

a) Symmetric
b) Left-skewed
c) Right-skewed



Notation

- The sample size, the number of cases in the sample, is denoted by *n*
- We often let *x* or *y* stand for any variable, and x_1, x_2, ..., x_n represent the *n* values of the variable *x*
- $x_1 = 32.4$, $x_2 = 28.4$, $x_3 = 26.8$, ...

State	Prevalence	Confidence Interval
Alabama	32.4	(30.8, 34.1)
Alaska	28.4	(26.5, 30.4)
Arizona	26.8	(24.3, 29.4)
Arkansas	34.6	(32.7, 36.6)
California	24.1	(23.0, 25.3)
Colorado	21.3	(20.4, 22.2)

Mean

The ***mean*** or average of the data values is

$$mean = \frac{sum\ of\ all\ data\ values}{number\ of\ data\ values}$$

$$mean = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

- Sample mean: $\bar{x}$
- Population mean: μ ("mu")

Minitab: Stat -> Basic Statistics -> Display Descriptive Statistics

Mean

```
Variable   N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3  Maximum
Obese     53   0  28.606    0.464  3.377   21.300  26.350  28.900  31.050   35.100
```

The average obesity rate across the 50 states is μ = 28.606.

```
Variable       N    N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3  Maximum
bmi        26159  4967  24.887   0.0437  7.064    7.987  19.573  24.163  28.995   66.437
```

The average BMI for Americans in this sample is $\bar{x} = 24.887$.

Median

The ***median***, *m*, is the middle value when the data are ordered.

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

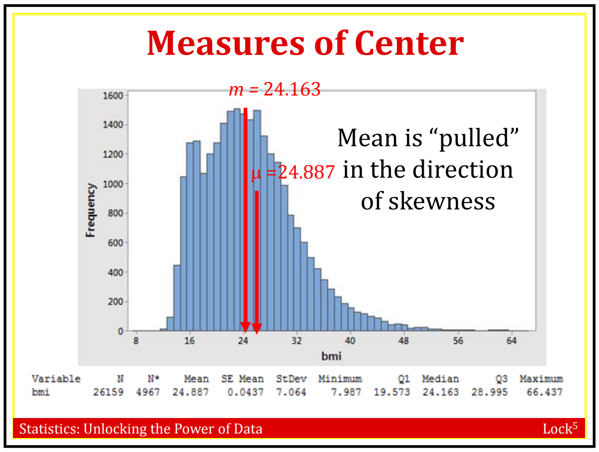
- The median splits the data in half.

```
Variable   N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3  Maximum
Obese     53   0  28.606    0.464  3.377   21.300  26.350  28.900  31.050   35.100
```

Minitab: Stat -> Basic Statistics -> Display Descriptive Statistics

Measures of Center

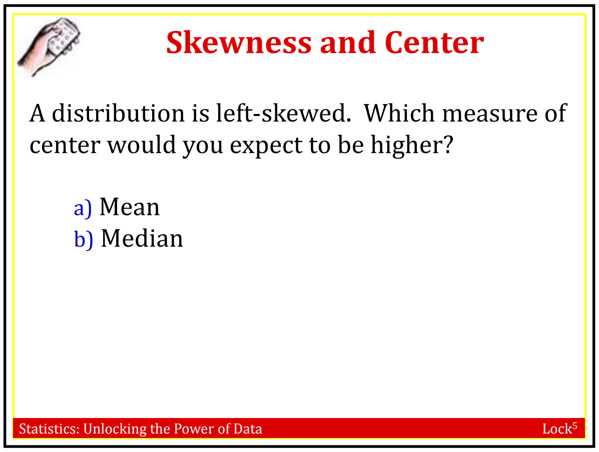
- For symmetric distributions, the mean and the median will be about the same
- For skewed distributions, the mean will be more pulled towards the direction of skewness



Measures of Center

Mean is “pulled” in the direction of skewness

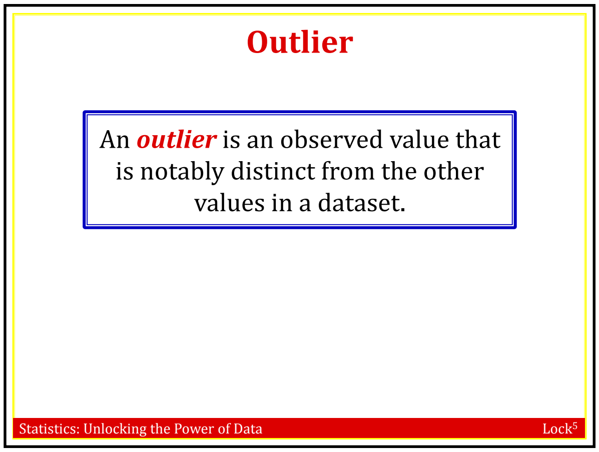
```
Variable      N     N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3  Maximum
bmi       26159   4967  24.887   0.0437  7.064    7.987  19.573  24.163  28.995   66.437
```



Skewness and Center

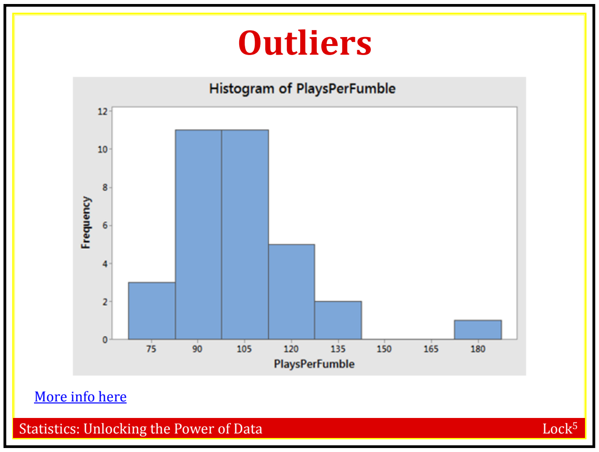
A distribution is left-skewed. Which measure of center would you expect to be higher?

a) Mean
b) Median



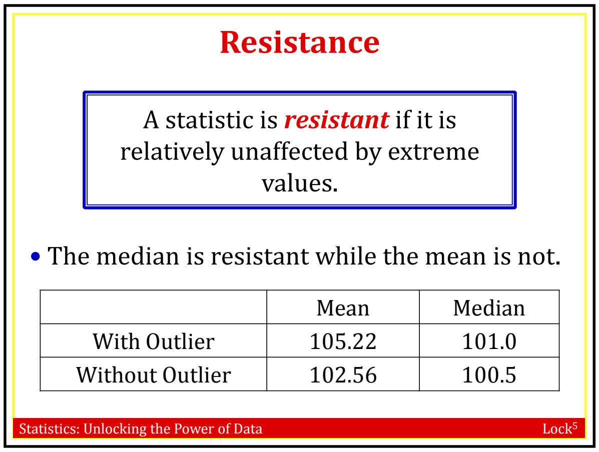
Outlier

An ***outlier*** is an observed value that is notably distinct from the other values in a dataset.



Outliers

More info here

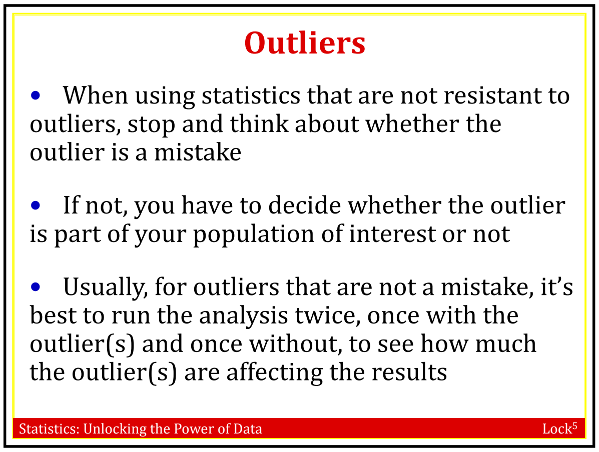


Resistance

A statistic is ***resistant*** if it is relatively unaffected by extreme values.

- The median is resistant while the mean is not.

	Mean	Median
With Outlier	105.22	101.0
Without Outlier	102.56	100.5



Outliers

- When using statistics that are not resistant to outliers, stop and think about whether the outlier is a mistake
- If not, you have to decide whether the outlier is part of your population of interest or not
- Usually, for outliers that are not a mistake, it’s best to run the analysis twice, once with the outlier(s) and once without, to see how much the outlier(s) are affecting the results

Standard Deviation

The ***standard deviation*** for a quantitative variable measures the spread of the data

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

- Sample standard deviation: *s*
- Population standard deviation: σ (“sigma”)

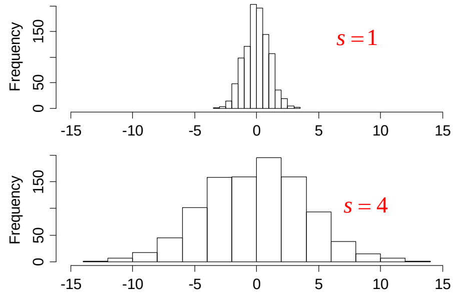
```
Variable   N  N*    Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
Obese     53   0  28.606    0.464  3.377   21.300 26.350  28.900 31.050   35.100
```

Minitab: Stat -> Basic Statistics -> Display Descriptive Statistics

Standard Deviation

- The standard deviation gives a rough estimate of the typical distance of a data values from the mean
- The larger the standard deviation, the more variability there is in the data and the more spread out the data are

Standard Deviation



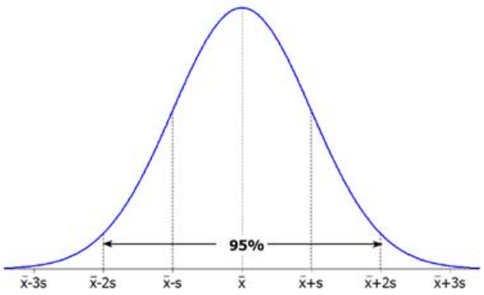
Both of these distributions are ***bell-shaped***

95% Rule

If a distribution of data is approximately symmetric and bell-shaped, about 95% of the data should fall within two standard deviations of the mean.

- For a population, 95% of the data will be between μ – 2σ and μ + 2σ
- For a sample, 95% of the data will be between $\bar{x} - 2s$ and $\bar{x} - 2s$

The 95% Rule



95% Rule

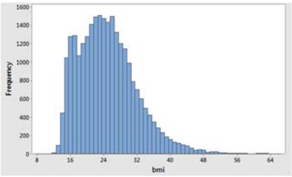
Give an interval that will likely contain 95% of obesity rates of states.

```
Variable   N  N*    Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
Obese     53   0  28.606    0.464  3.377   21.300 26.350  28.900 31.050   35.100
```

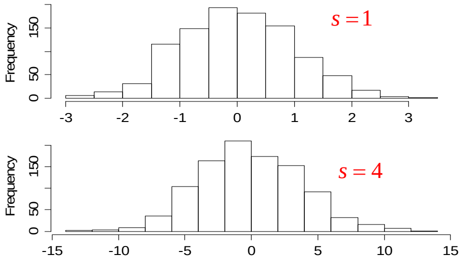
95% Rule

Could we use the same method to get an interval that will contain 95% of BMIs of American adults?

a) Yes
b) No

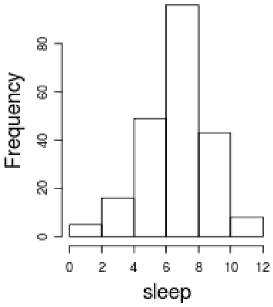


The 95% Rule



- StatKey

The 95% Rule



The standard deviation for hours of sleep per night is closest to

a) ½
b) 1
c) 2
d) 4
e) I have no idea

To Do

- Read Sections 2.2 and 2.3
- Do Homework 2.2 (due Friday, 2/6)