# Descriptive Statistics

Anastasiia Kim

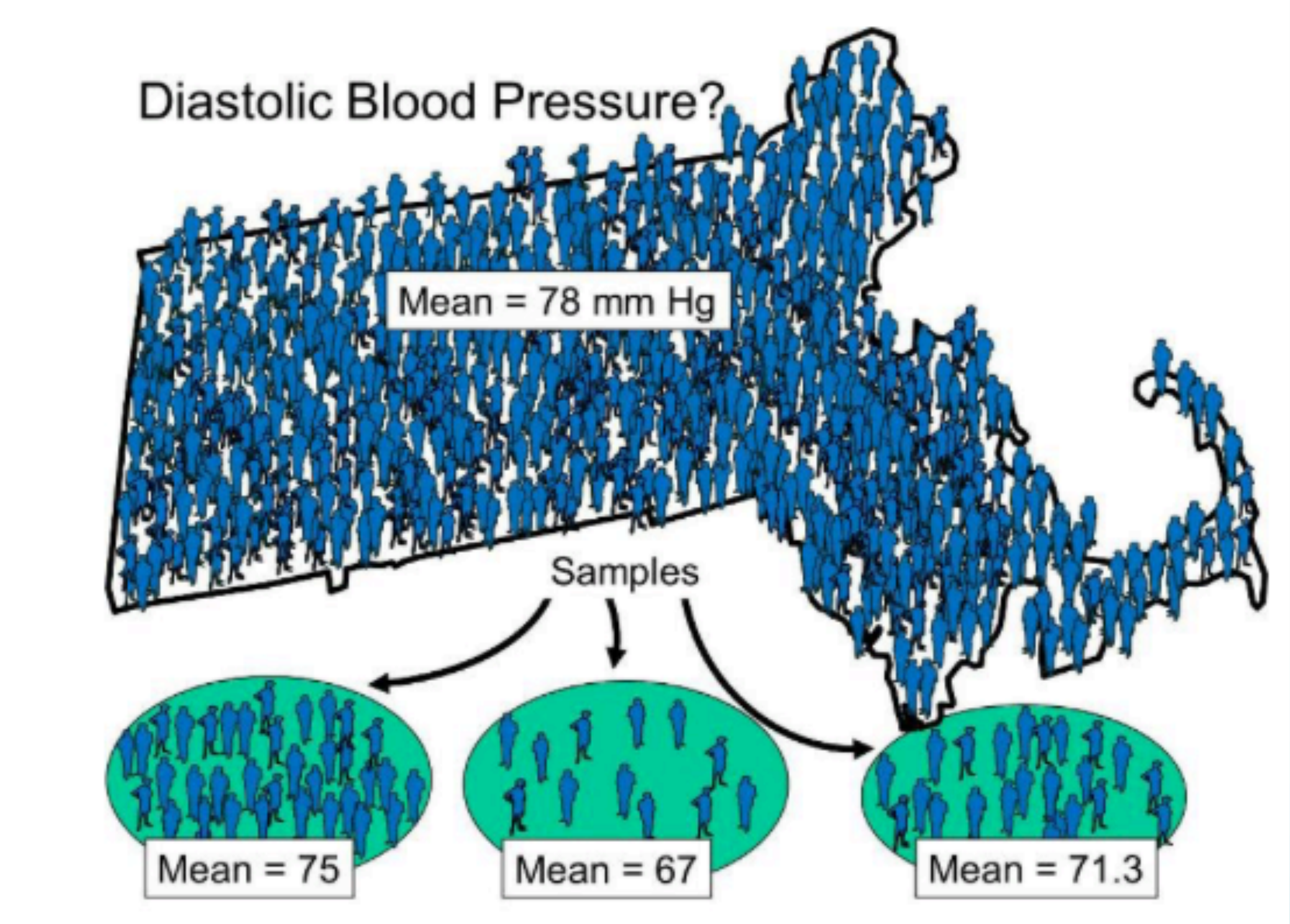
April 13, 2020

some examples are adapted from *The Basic Practice of Statistics (7th Edition),* by Moore, Notz and Fligner

# Population and Sample

- The **population** in a statistical study is the entire group of individuals about which we want information.
- A **sample** is the part of the population from which we actually collect information. We use information from a sample to draw conclusions about the entire population.
- In a study of work stress, 100 female restaurant workers were asked about the impact of work stress on their personal lives. Population: female restaurant workers. Sample: 100 workers.

# Population and Sample



http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_BiostatisticsBasics/BS704_BiostatisticsBasics2.html

# Probability Sampling: equal chances of being selected

- Simple Random Sampling:

use software to generate rand. numbers

- Systematic Sampling:

select every nth subject in the population

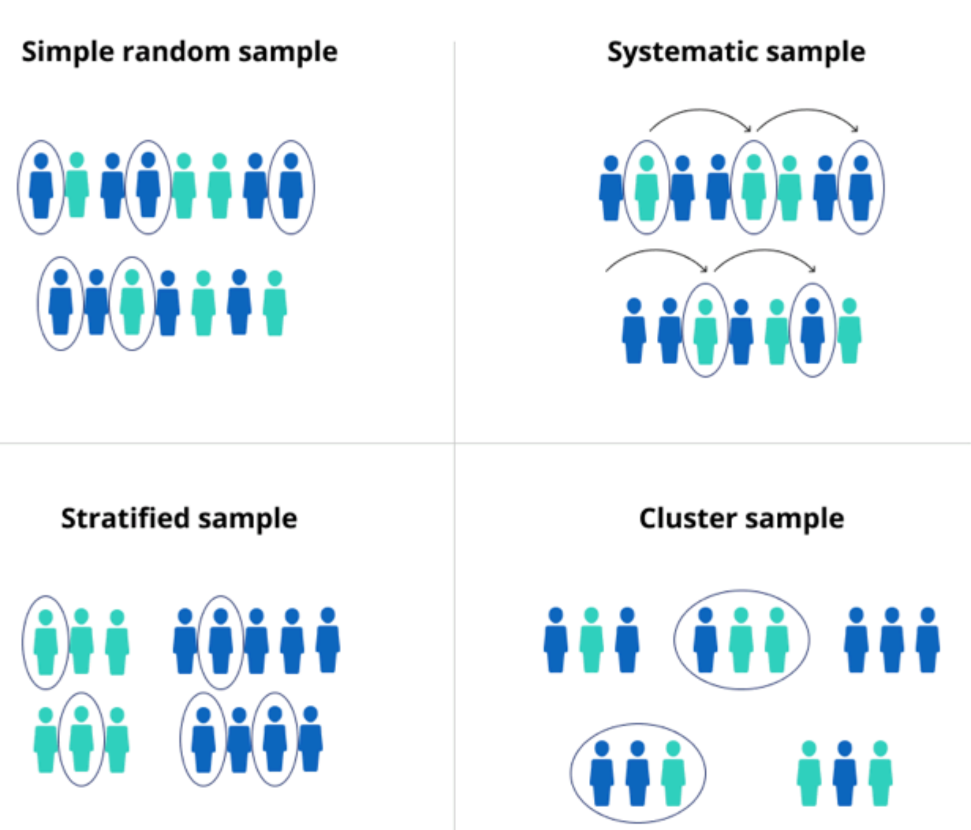
- Stratified Random Sampling:

separate the population into subgroups

and then take a random sample from

each subgroup. This ensures that certain

groups are represented in the sample

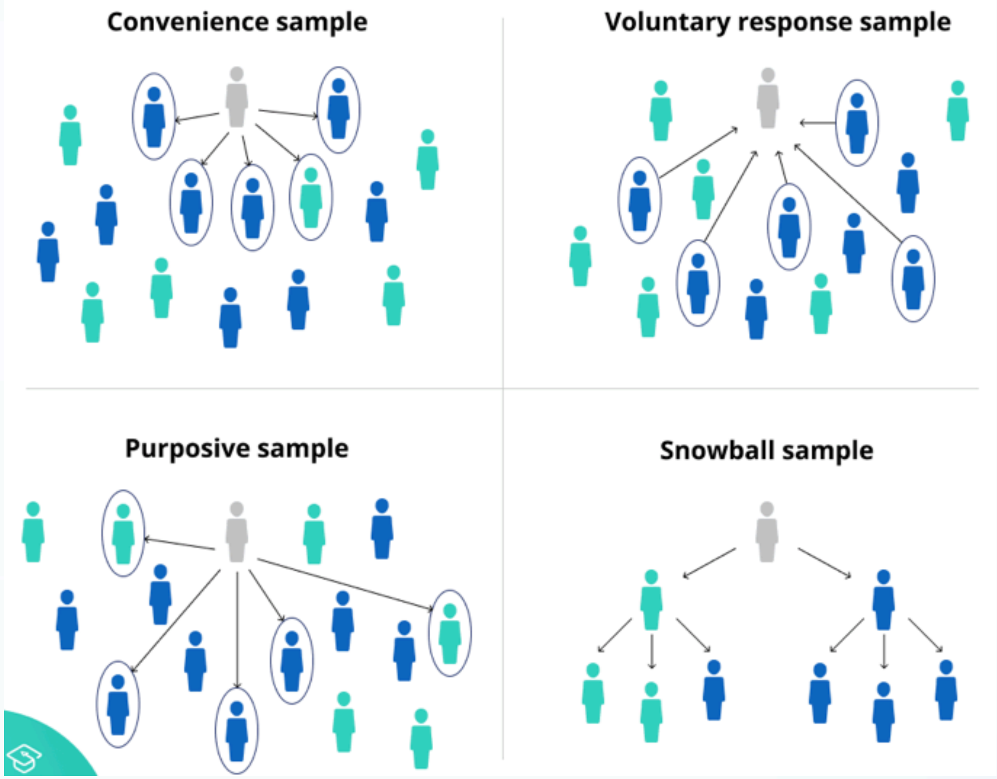


- Cluster Sampling: divide a population into smaller clusters, then select a random sample of the clusters. All of the subjects are measured within each selected cluster.

picture: https://www.scribbr.com/methodology/sampling-methods/

# Non-Probability Sampling

- Convenience Sampling
- Voluntary Response Sampling
- Purposive Sampling
- Snowball Sampling



picture: https://www.scribbr.com/methodology/sampling-methods/

# Categorical and Quantitative variables

| Categorical | Quantitative |
|---|---|
| Type of pet owned (cat, fish, dog) | Numbers of pets owned (2 pets) |
| Favorite book, song | Numbers of books in the library (100 books) |
| Car color | Weight in pounds |
| Model of car | Bank account balance |

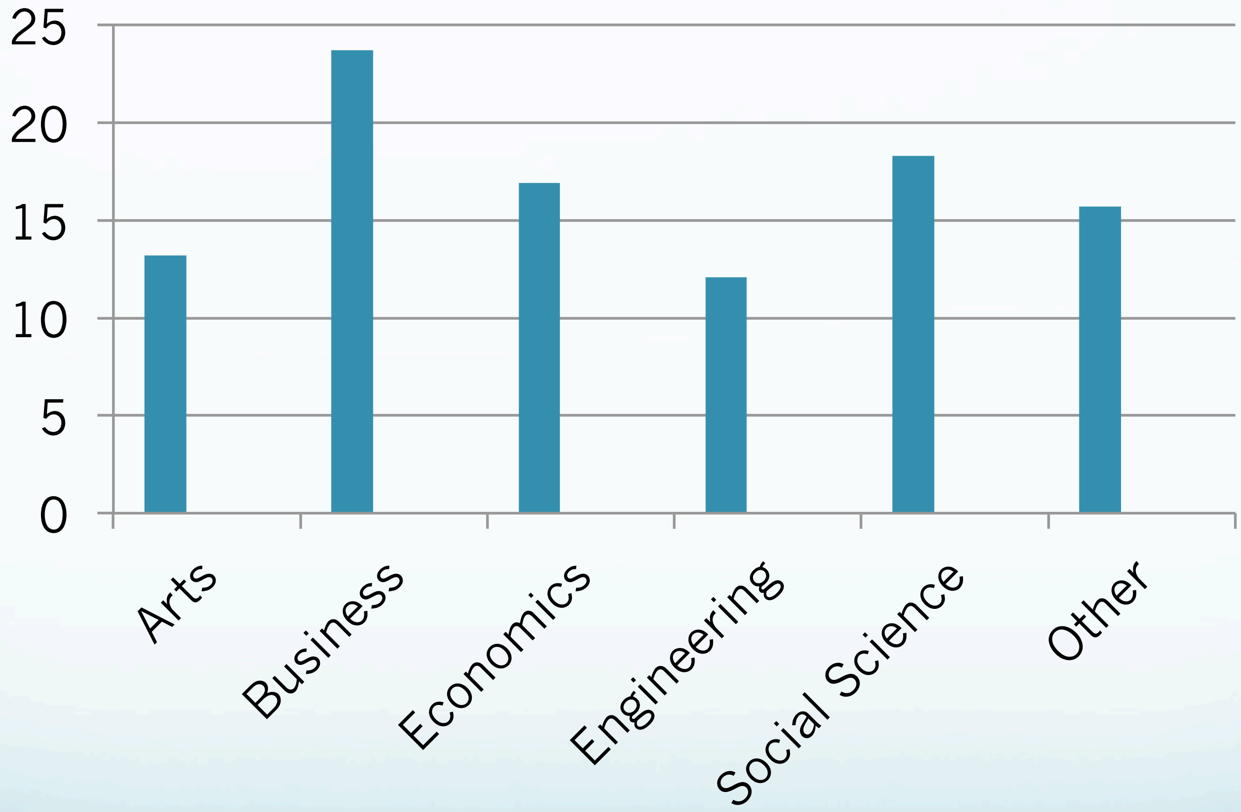
Gender is a categorical variable but looks like quantitative. Because arithmetic operations doesn't make sense for it.

Here are data on the percents of first-year students who plan to major in several areas:

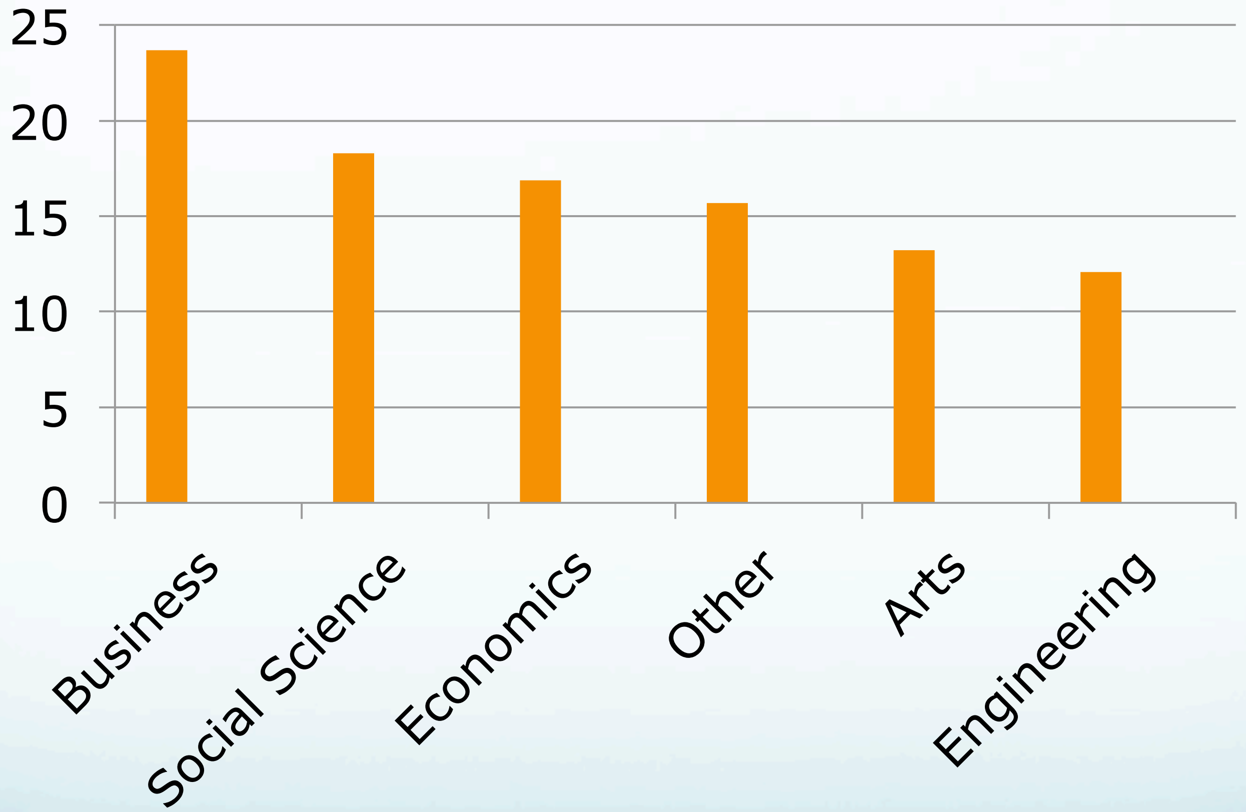
| Field of study | Percent of students |
|---|---|
| Arts | 13.2 |
| Social science | 18.3 |
| Economics | 16.9 |
| Engineering | 12.1 |
| Business | 23.7 |
| Other majors | 15,7 |
| Total | 99.9 |

Why not 100%? The exact percents would add to 100, but each percent is rounded to the nearest tenth. This is roundoff error.

## The bar heights show the category counts or percents (the bar in alphabetical order).



# In order of height



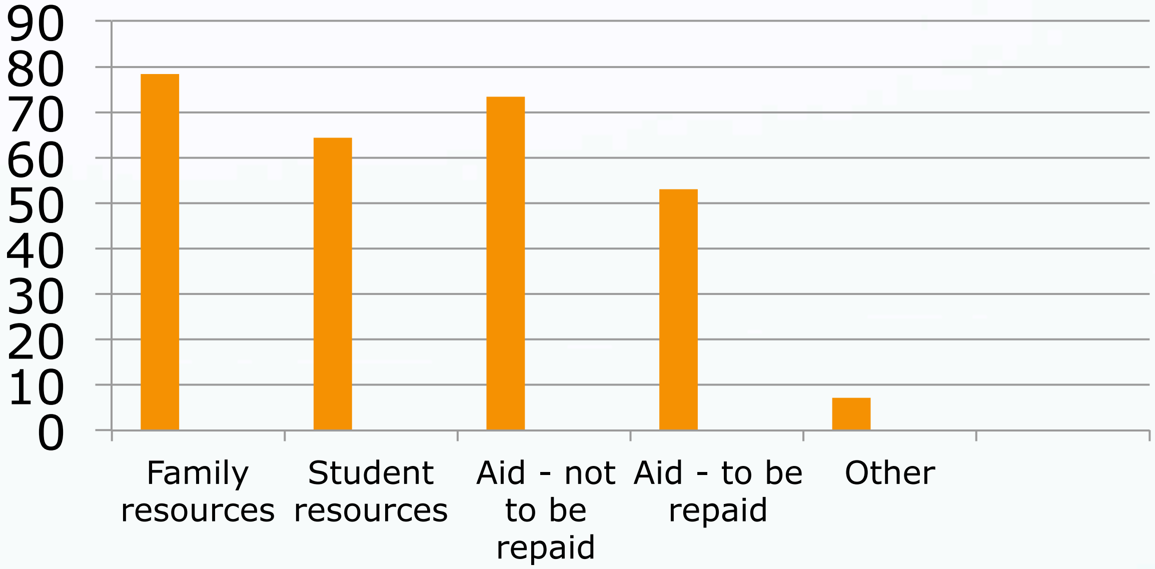
## Example

The Higher Education Research Institute's Freshman Survey reports the following data on the sources students use to pay for college expenses.

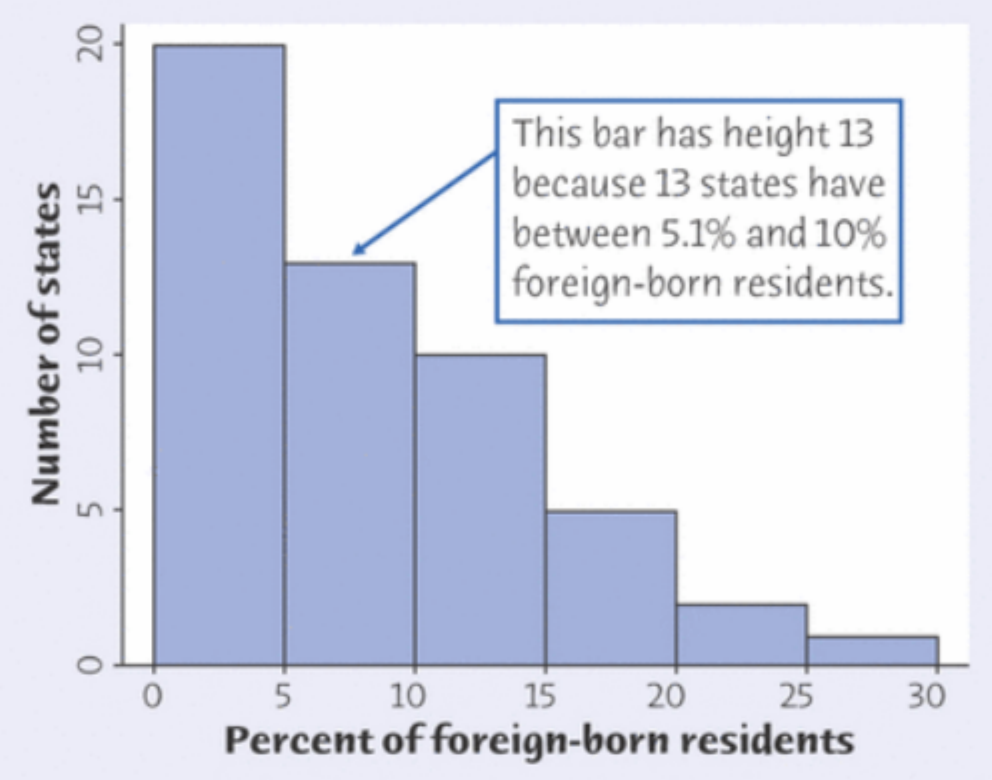
| Source for college expenses | Students |
|---|---|
| Family resources | 78,4% |
| Student resources | 64,3% |
| Aid – not to be repaid | 73,4% |
| Aid – to be repaid | 53,1% |
| Other | 7,1% |

Why it is not correct to use a pie chart? Because each percent in table refers to a different source of student payment.

# But we can build a bar for these data

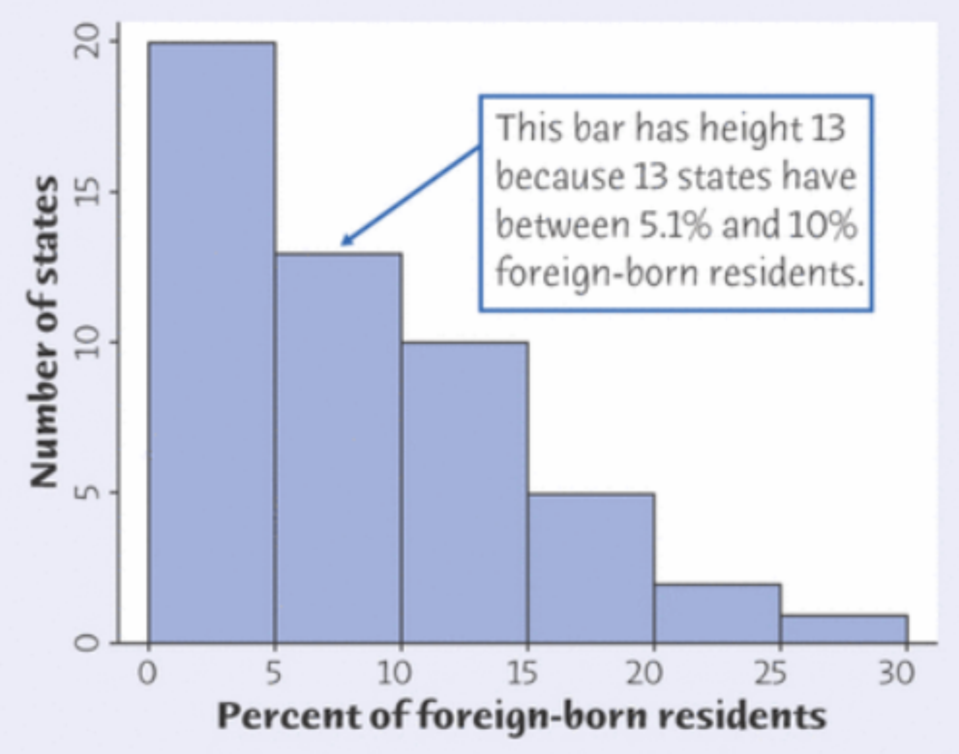


| Class | Count |
|---|---|
| 0.1 to 5.0 | 20 |
| 5.1 to 10.0 | 13 |
| 10.1 to 15.0 | 10 |
| 15.1 to 20.0 | 5 |
| 20.1 to 25.0 | 2 |
| 25.1 to 30.0 | 1 |



This bar has height 13 because 13 states have between 5.1% and 10% foreign-born residents.

- Histogram:
- Single peak (between 0% and 5%).
- A majority of states(33) have no more than 10% foreign-born residents. The distribution is skewed to the right.



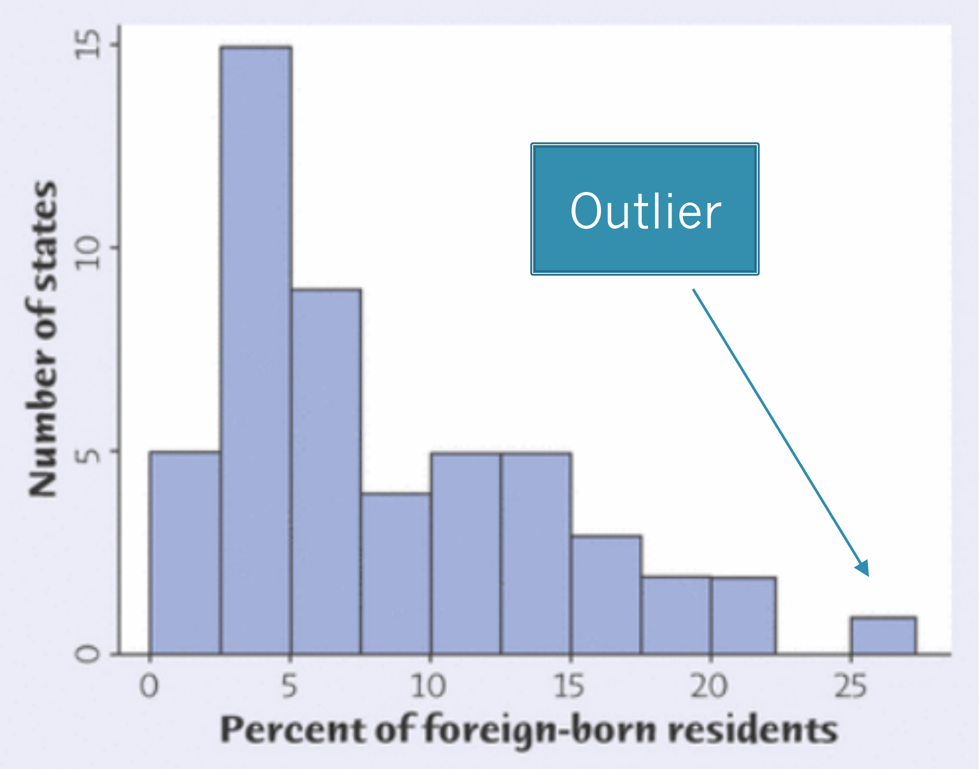
This bar has height 13 because 13 states have between 5.1% and 10% foreign-born residents.

- Spread: from 1.2% to 27.2%

The choice of classes can influence the appearance of a distribution.

Concentrate on the main features of a distribution:

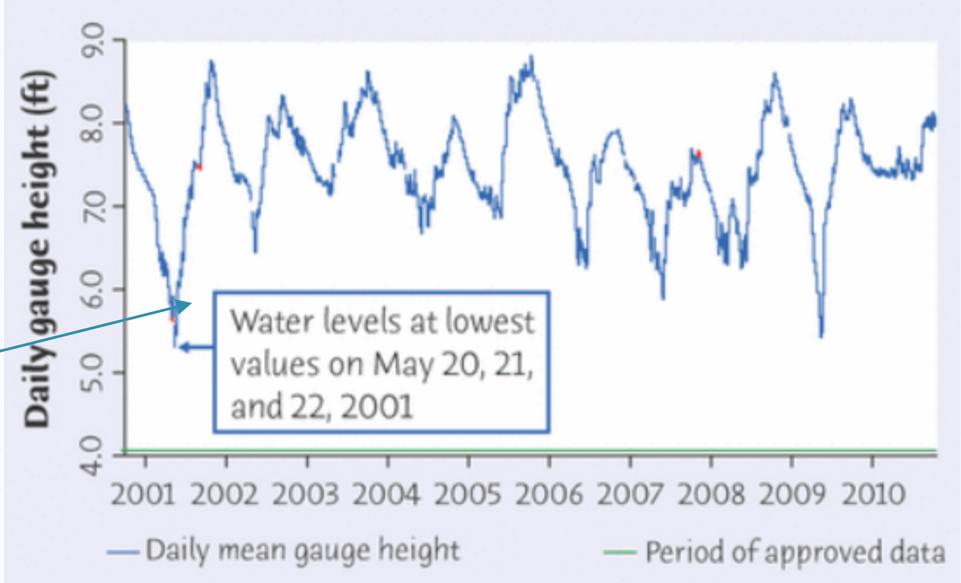
- Major peaks
- Clear outliers
- Approximate symmetry
- Clear skewness

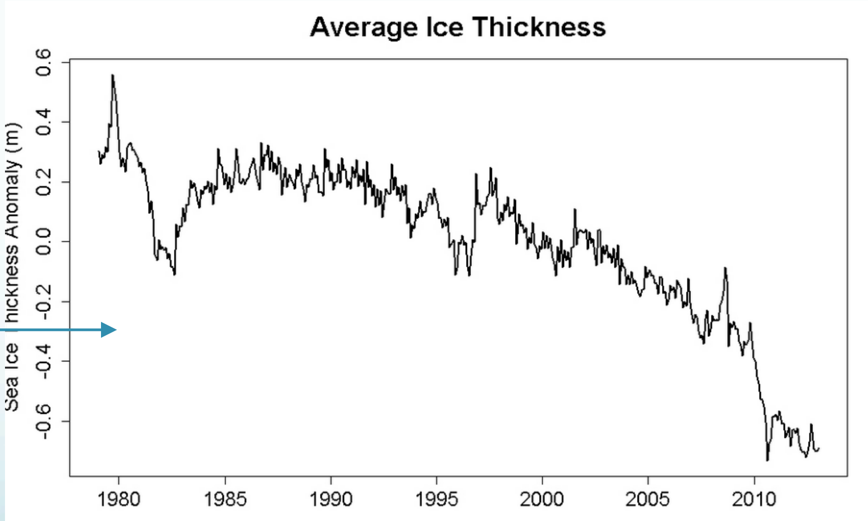


# Time plot

Common overall patterns:

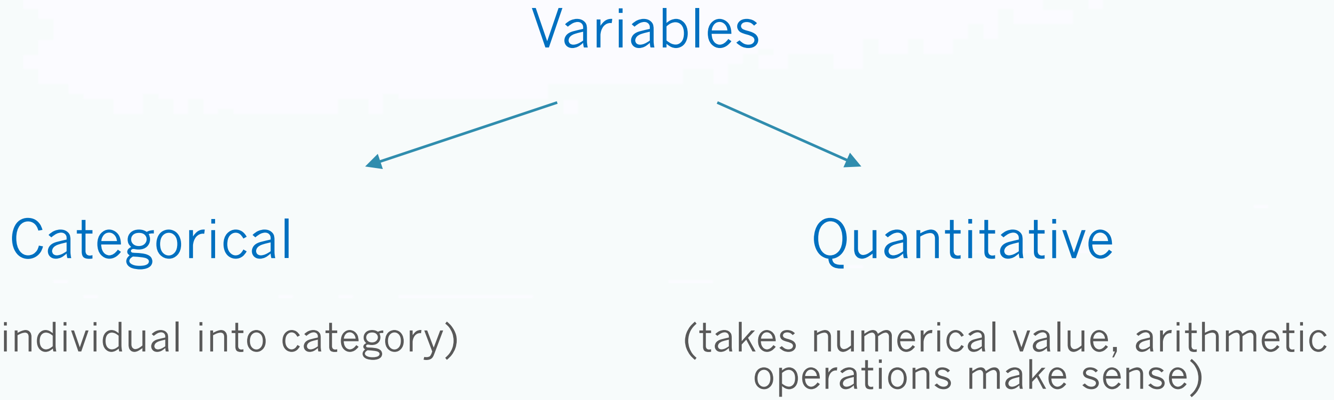
- Cycles (regular up-and-down movements)
- Trend (a long-term upward or downward movement over time)


Water levels at lowest values on May 20, 21, and 22, 2001


Average Ice Thickness

Individuals – the objects described by a set of data (people, animals, things).

Variables – characteristics of an individual (it can take different values for different individuals).



Variables

Categorical

(places an individual into category)

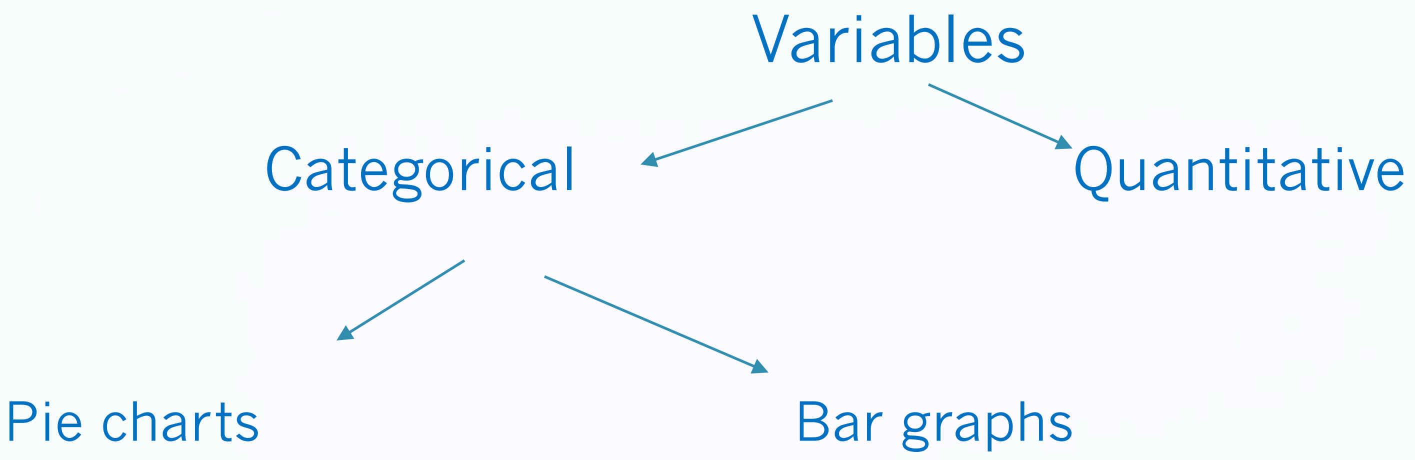
Quantitative

(takes numerical value, arithmetic operations make sense)

Two principles to organize our exploration of a set of data:

1. Exploring data (examining each variable by itself, study the relationships among the variables).

2. Distribution of a variable (what values it takes and how often it takes these values).



# Variables

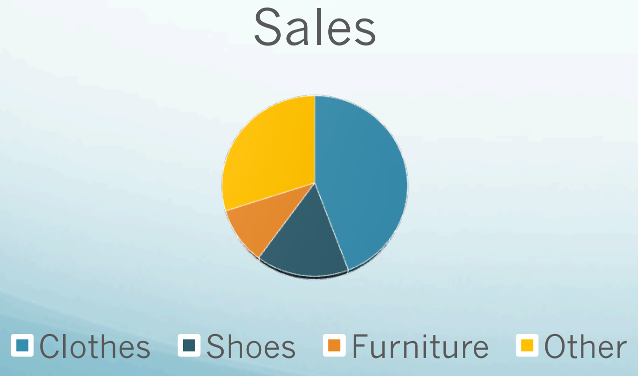
## Categorical

## Quantitative

### Pie charts

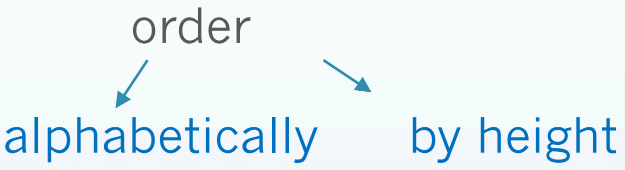
(slices are sized by the counts or percents for the categories)

Use to emphasize each category's relation to the whole



### Bar graphs

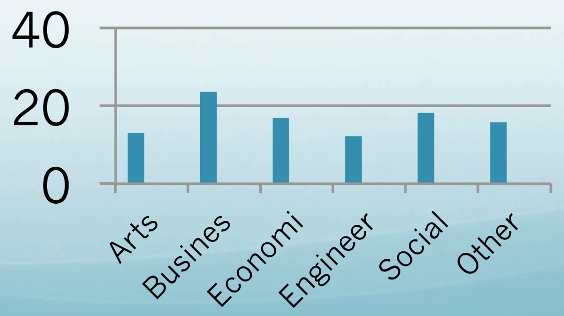
(represent each category as a bar. The bar heights show the category counts or percents)

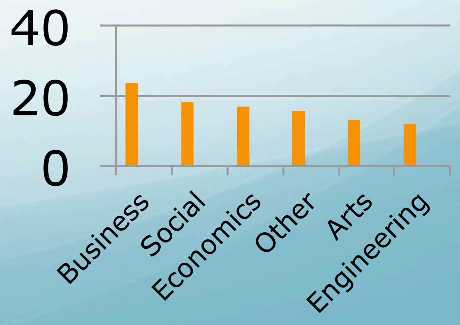


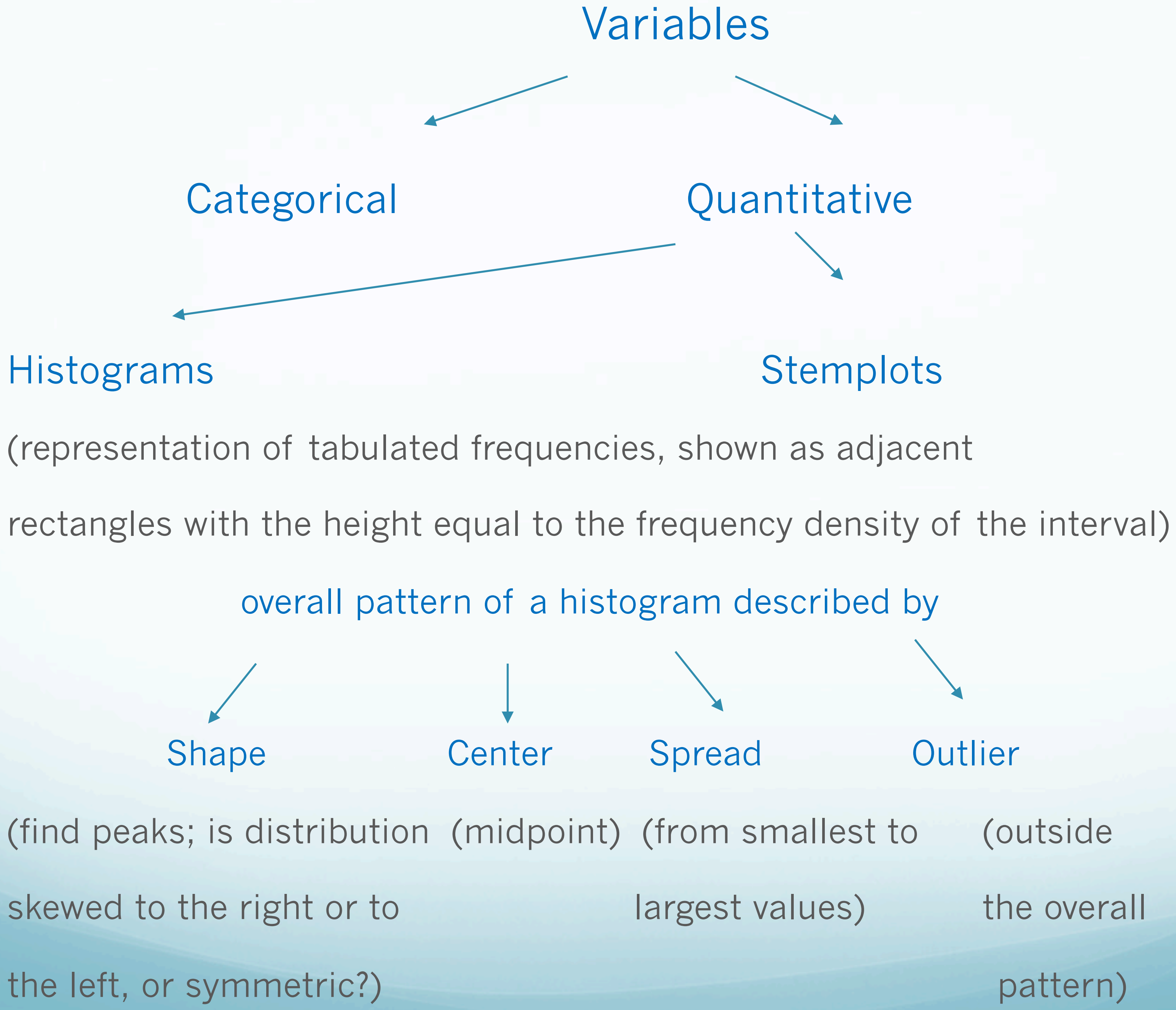
order

alphabetically

by height







Variables

Categorical

Quantitative

Histograms

Stemplots

(representation of tabulated frequencies, shown as adjacent rectangles with the height equal to the frequency density of the interval)

overall pattern of a histogram described by

Shape

(find peaks; is distribution skewed to the right or to the left, or symmetric?)

Center

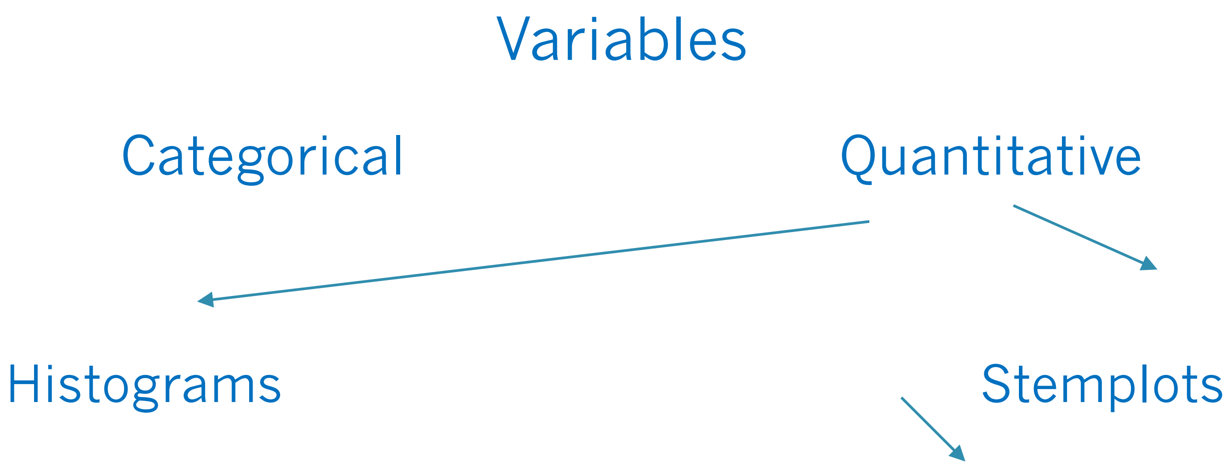
(midpoint)

Spread

(from smallest to largest values)

Outlier

(outside the overall pattern)



Variables

Categorical

Quantitative

Histograms

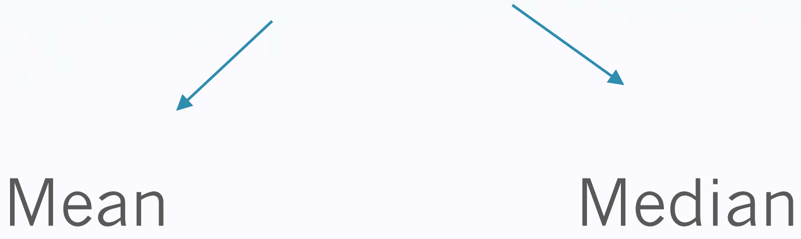
Stemplots

Looks like a histogram turned on end but preserves the actual value of each observation. Arrange the leaves in ascending order. Don’t use stemplots for large data sets)

| | |
|---|---|
| 12 | 2 |
| 13 | 1 3 |
| 14 | 2 5 8 |
| 15 | 0 2 5 7 9 |
| 16 | 1 3 5 7 |
| 17 | 1 2 |

# Describing distributions with numbers

# Measures of center



Mean

Median

- Mean: sum of all the observations divided by the number of observations
- Median: the middle number when observations are put in increasing order

## Example 1

Here are the data: 5 4 3 2 6 2 3 4 8

1) The mean:

2) The median:

n = 9

a) arrange in increasing order: 2 2 3 3 4 4 5 6 8

b) n is odd, location of median = element in the sorted list so M = 4.

## Example 2

Find the median for next data set: 4 6 7 2 8 3

1) 2 3 4 6 7 8

2) n = 6, n is even, location of median = so median is a mean of 3rd and 4th values

Let's consider 1st example again:

Change the last value from 8 to 34 and calculate the mean and the median again:

Data: 2 2 3 3 4 4 5 6 34

1) The mean:

2) The median:

n = 9

a) arrange in increasing order: 2 2 3 3 4 4 5 6 34

b) n is odd, location of median = element in the sorted list so M = 4.

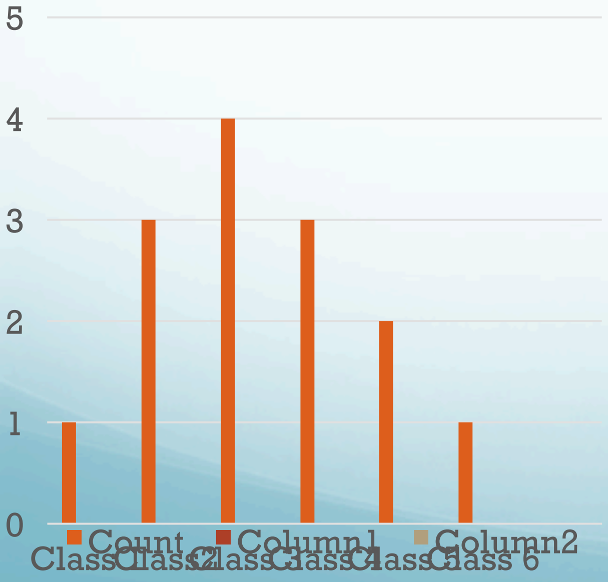
We has increased the observation and the mean has increased by 3 but the median hasn't change. The median is a resistant measure of center while the mean is not. Resistant measures are not influenced by outliers or skewness. In our case 34 is outlier.

# Facts: if a distribution is:

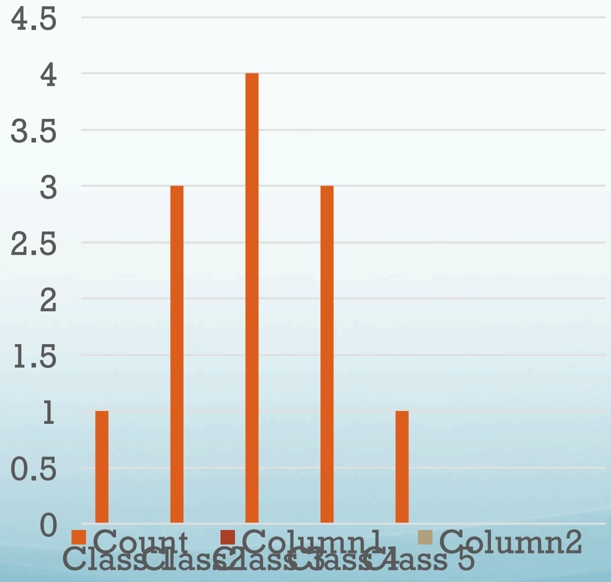
1) Roughly symmetric – the mean and median are close together;

2) Exactly symmetric – the mean and median are exactly the same;

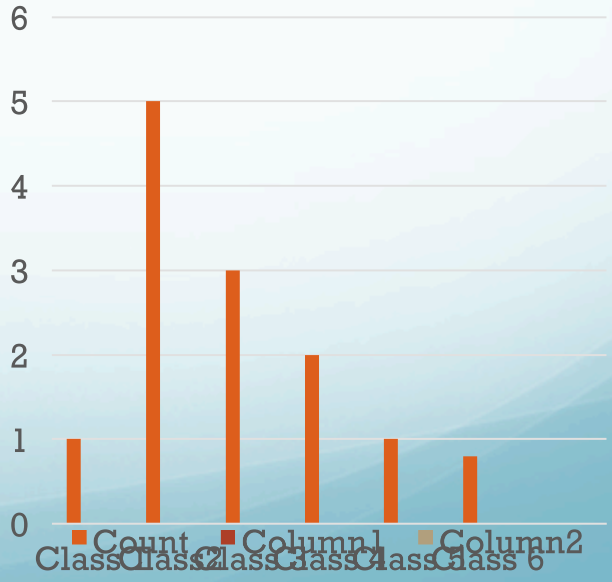
3) Skewed to the right or to the left – the mean is usually farther out in the long tail than is the median



Roughly symmetric



Exactly symmetric



Skewed to the right

# MEASURING SPREAD: The standard deviation

The sample variance $s^2$ of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of $n$ observations $x_1$, $x_2$,..., $x_n$ is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

or

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

The standard deviation $s$ is the square root of the sample variance $s^2$:

# Measuring spread: the quartiles

The quartiles are the 3 points that divide the data set into four equal groups.

To calculate the quartiles:

1. Sort the data in increasing order.
2. Find the median of the data set. It will be the second quartile, so
3. The first quartile (lower quartile) is the middle number (median) between the smallest number and the median of the data set.
4. The third quartile (upper quartile) is the middle number (median) between the highest number and the median of the data set.

Example 3

Find quartiles for the data set: 7 12 5 2 9 10 1

1. 1 2 5 7 9 10 12, n is odd
2. Q1 median of the data set: 1 2 5
3. Q3 median of the data set: 9 10 12

Example 4

Find quartiles for the data set: 1 3 7 1 10 10 10 13 8 1

1. 1 1 1 3 7 8 10 10 10 13, n is even
2. Q1 median of the data set: 1 1 1 3 7
3. Q3 median of the data set: 8 10 10 10 13

Let’s change the last number in the third example from 12 to 20, the quartiles will not change. The quartiles are resistant because they are not affected by a few extreme observations.

# The Five-number summary

The five–number summary of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five–number summary is

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

Example:

The number of sleeping hours for adults: 1 2 5 7 9 10 12

The number of sleeping hours for teens: 1 1 1 3 7 10 10 10 10 13

We can give a five-number summary:

| Sleeping hours | Min | | M | | Max |
|---|---|---|---|---|---|
| For adults | 1 | 2 | 7 | 10 | 12 |
| For teens | 1 | 1 | 8.5 | 10 | 13 |

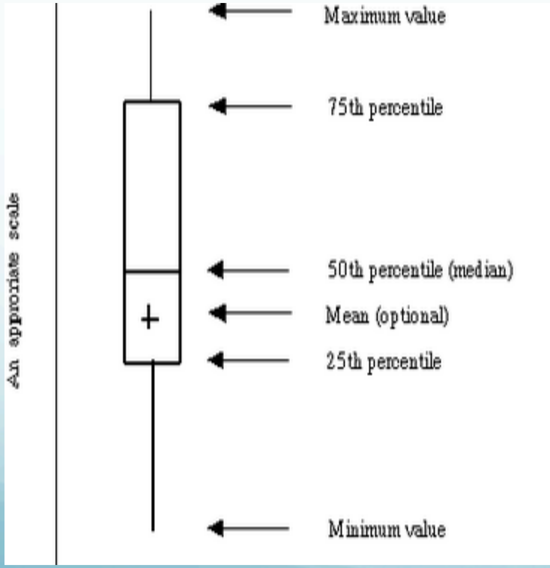
Boxplot is a graph of the five-number summary. How to make a boxplot?

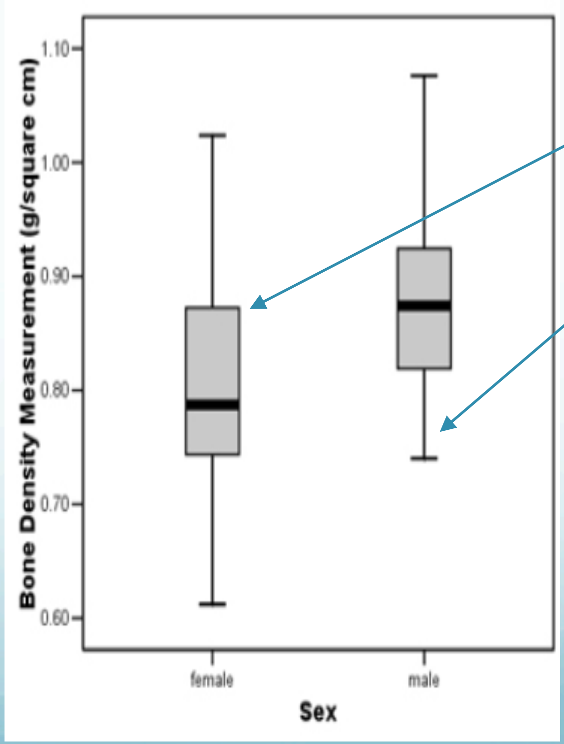
1. A central box spans the lower and upper quartiles.

2. A line in the box marks the median.

3. Lines extend from the box out to the smallest and largest observations.

Boxplots show less detail than histograms or stemplots, so they are best used for side-by-side comparison of more than one distribution.





Skewed to the right

# Spotting suspected outliers

The interquartile range *IQR* is the distance between the first and third quartiles,

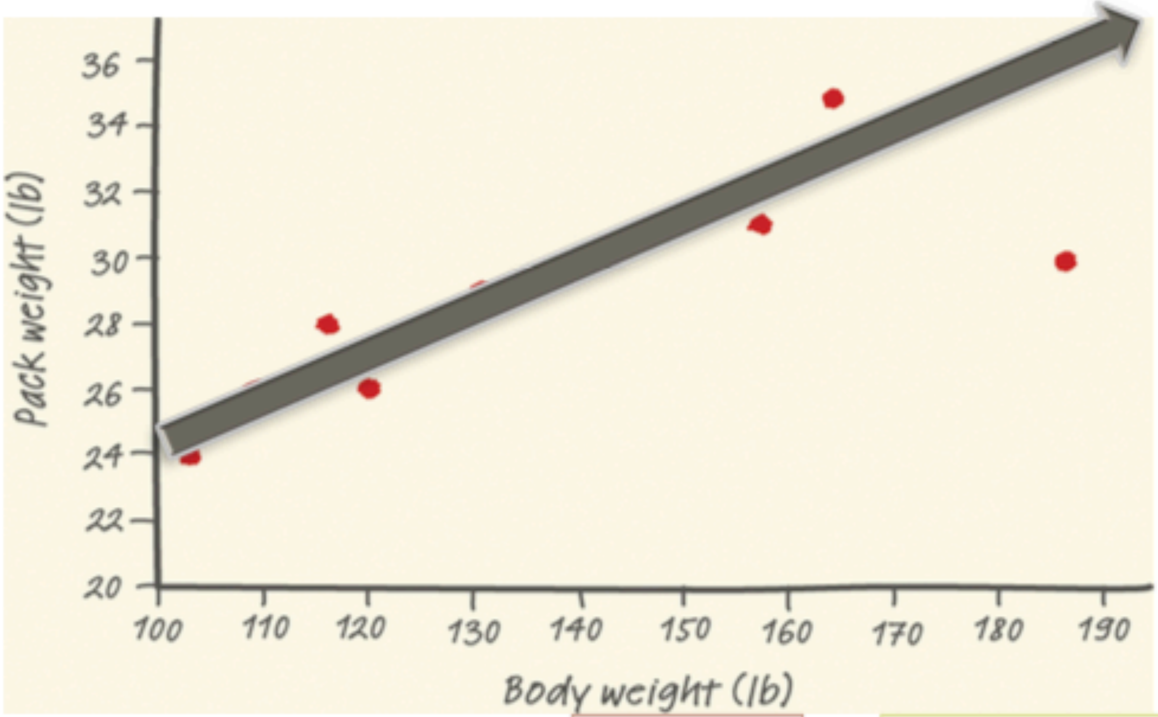
$IQR = Q_3 - Q_1$

## THE 1.5 × *IQR* RULE FOR OUTLIERS

Call an observation a suspected outlier if it falls more than 1.5 × *IQR* above the third quartile or below the first quartile.

Any values not falling between are flagged as suspected outliers. The 1.5 × *IQR* rule is not a replacement for looking at the data. It is most useful when large volumes of data are scanned automatically.

# Scatterplot



**Outlier**

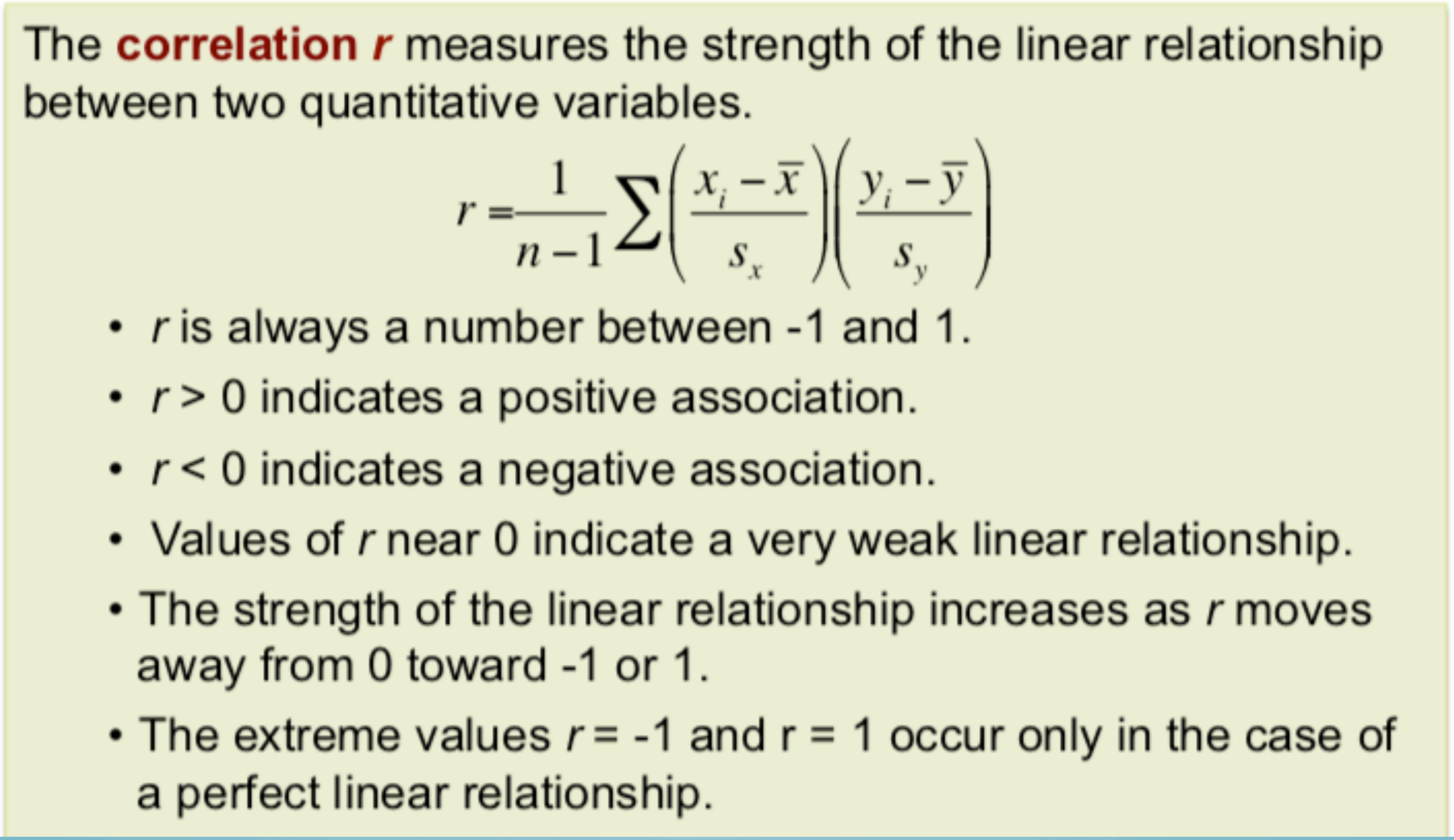
There is one possible outlier: the hiker with the body weight of 187 pounds seems to be carrying relatively less weight than are the other group members.

**Strength** **Direction** **Form**

- ✓ There is a moderately strong, positive, linear relationship between body weight and pack weight.
- ✓ It appears that lighter hikers are carrying lighter backpacks.

# Sample correlation

□ A scatterplot displays the strength, direction, and form of the relationship between two quantitative variables.



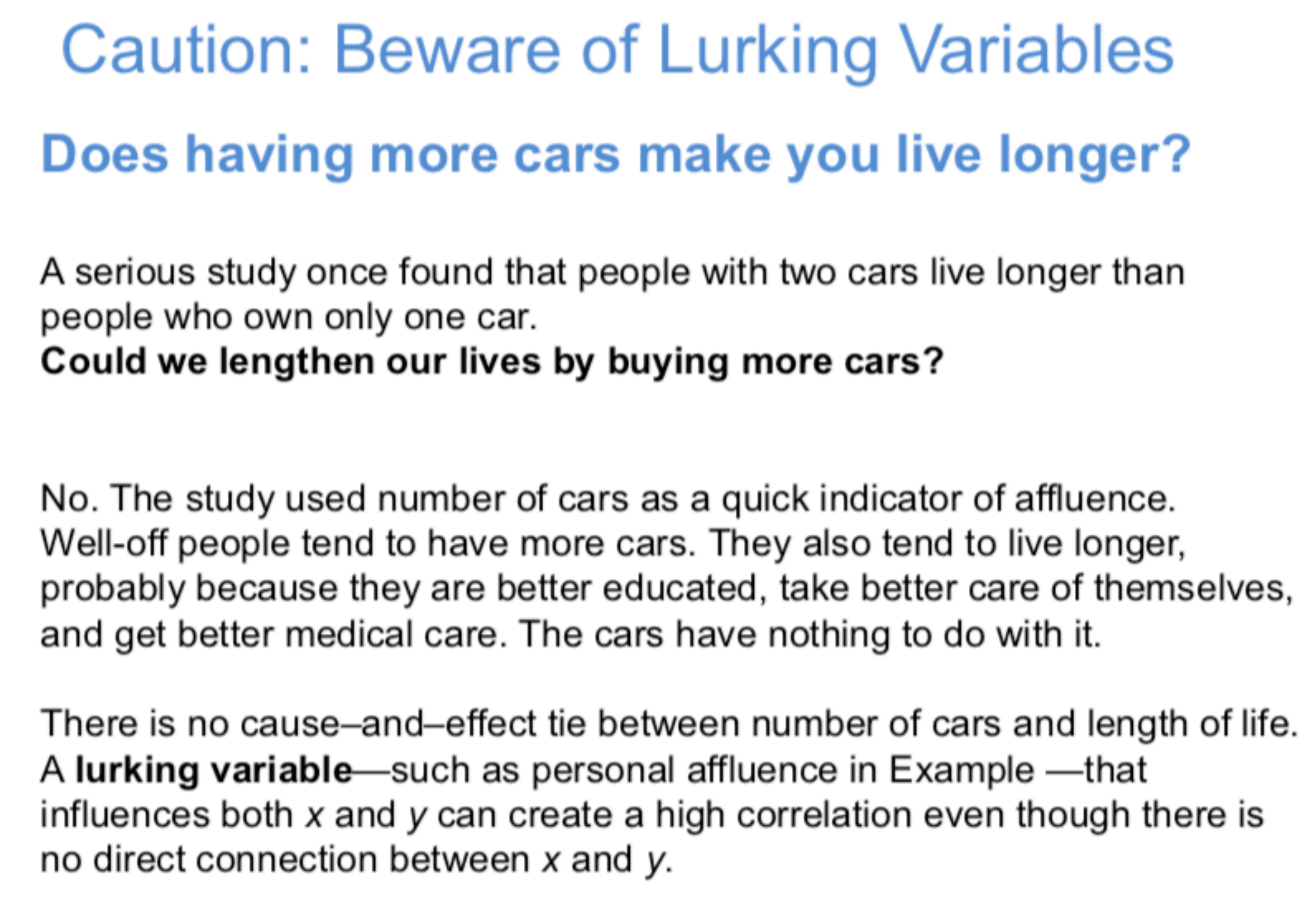
The **correlation *r*** measures the strength of the linear relationship between two quantitative variables.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- *r* is always a number between -1 and 1.
- *r* > 0 indicates a positive association.
- *r* < 0 indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 toward -1 or 1.
- The extreme values *r* = -1 and r = 1 occur only in the case of a perfect linear relationship.

# Association Does Not Imply Causation

- Even very strong correlations may *not* correspond to a real causal relationship (changes in *x* actually *causing* changes in *y*).
- Correlation may be explained by a lurking variable.



## Caution: Beware of Lurking Variables

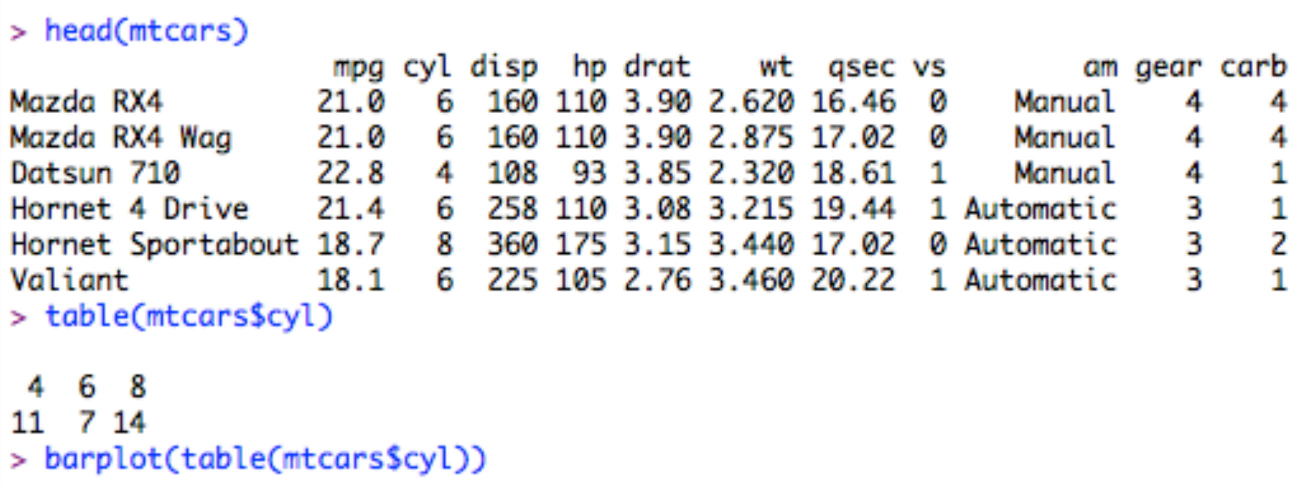
### Does having more cars make you live longer?

A serious study once found that people with two cars live longer than people who own only one car.
**Could we lengthen our lives by buying more cars?**

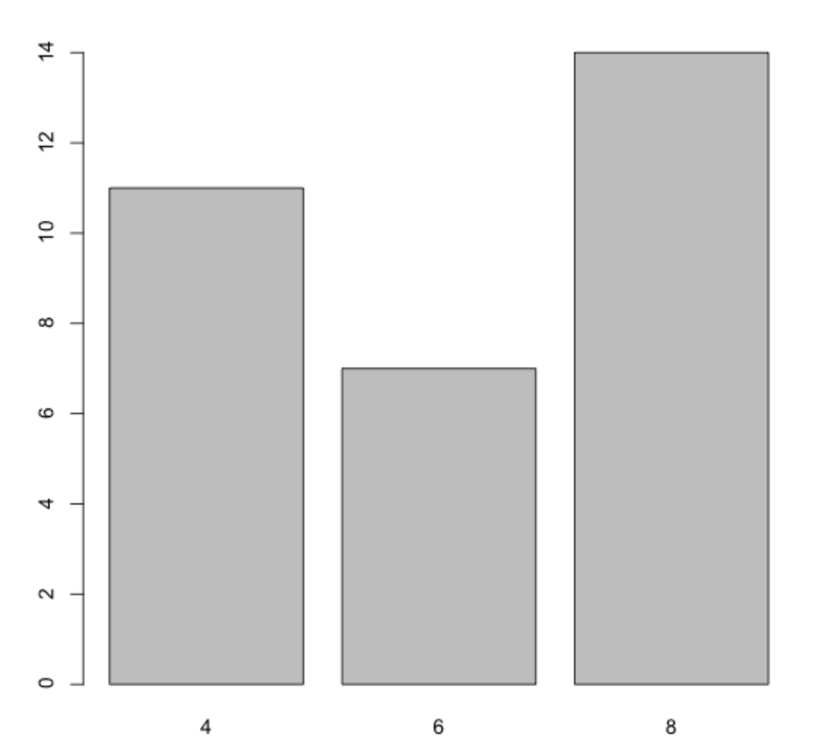
No. The study used number of cars as a quick indicator of affluence. Well-off people tend to have more cars. They also tend to live longer, probably because they are better educated, take better care of themselves, and get better medical care. The cars have nothing to do with it.

There is no cause–and–effect tie between number of cars and length of life. A **lurking variable**—such as personal affluence in Example —that influences both *x* and *y* can create a high correlation even though there is no direct connection between *x* and *y*.

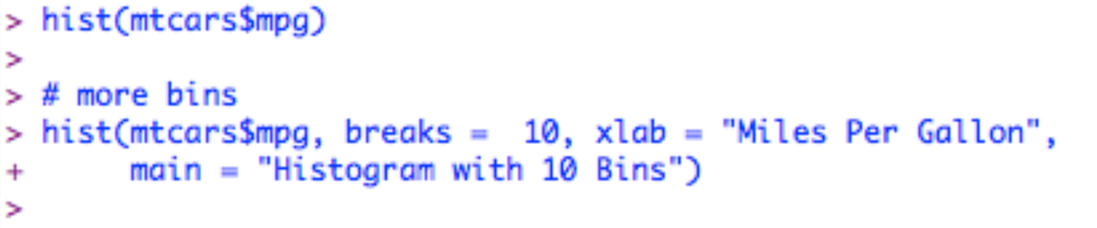
# R: bar plot



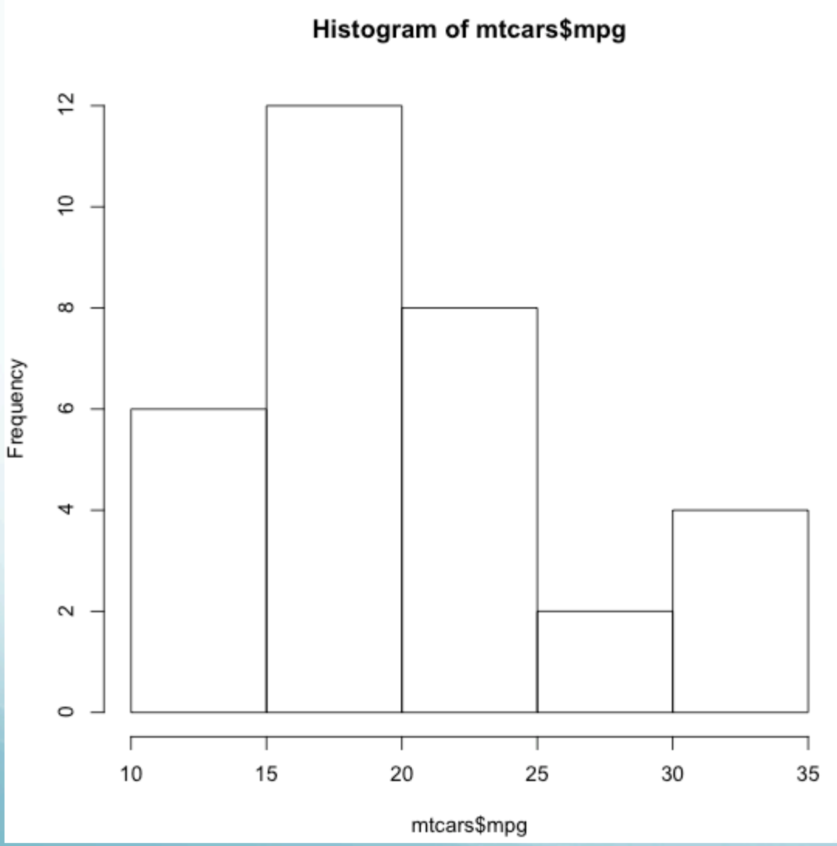
```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs        am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0    Manual    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0    Manual    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1    Manual    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1 Automatic    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0 Automatic    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1 Automatic    3    1
> table(mtcars$cyl)

 4  6  8
11  7 14
> barplot(table(mtcars$cyl))
```

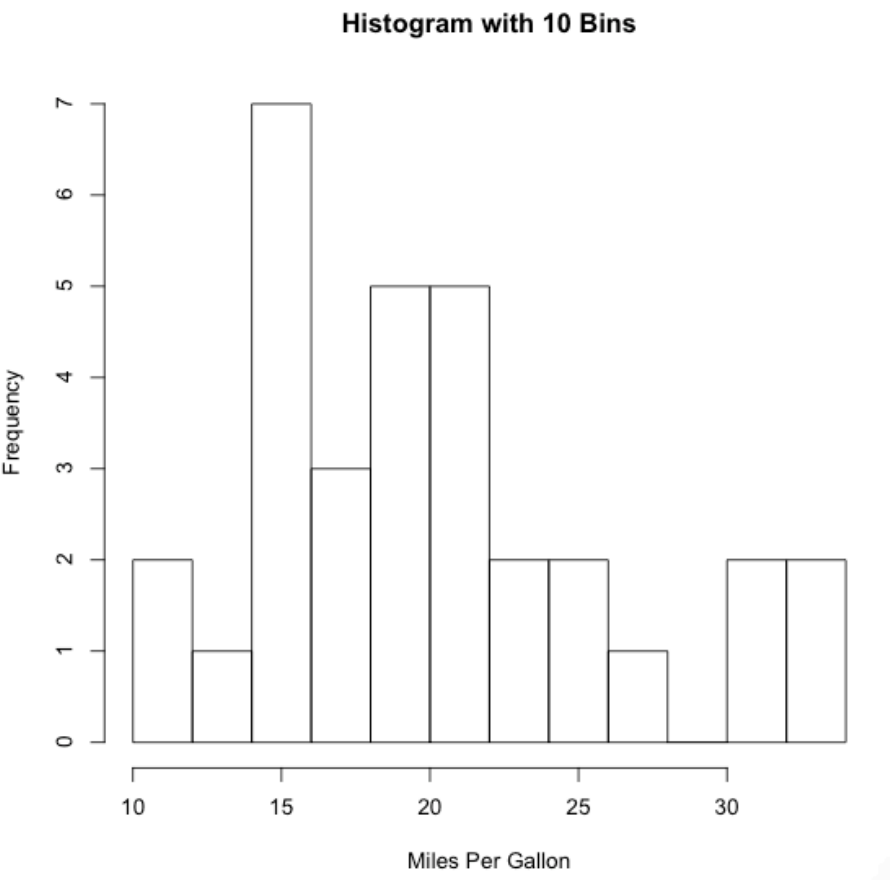


# R: histogram

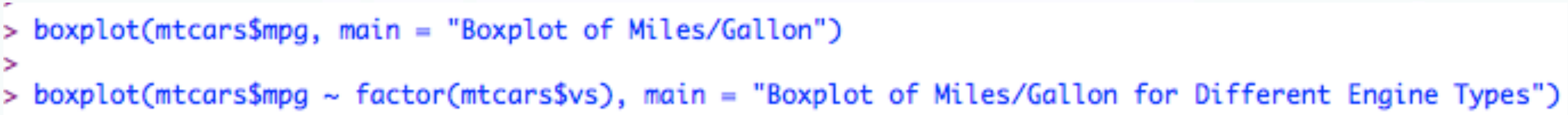


```
> hist(mtcars$mpg)
>
> # more bins
> hist(mtcars$mpg, breaks =  10, xlab = "Miles Per Gallon",
+       main = "Histogram with 10 Bins")
>
```

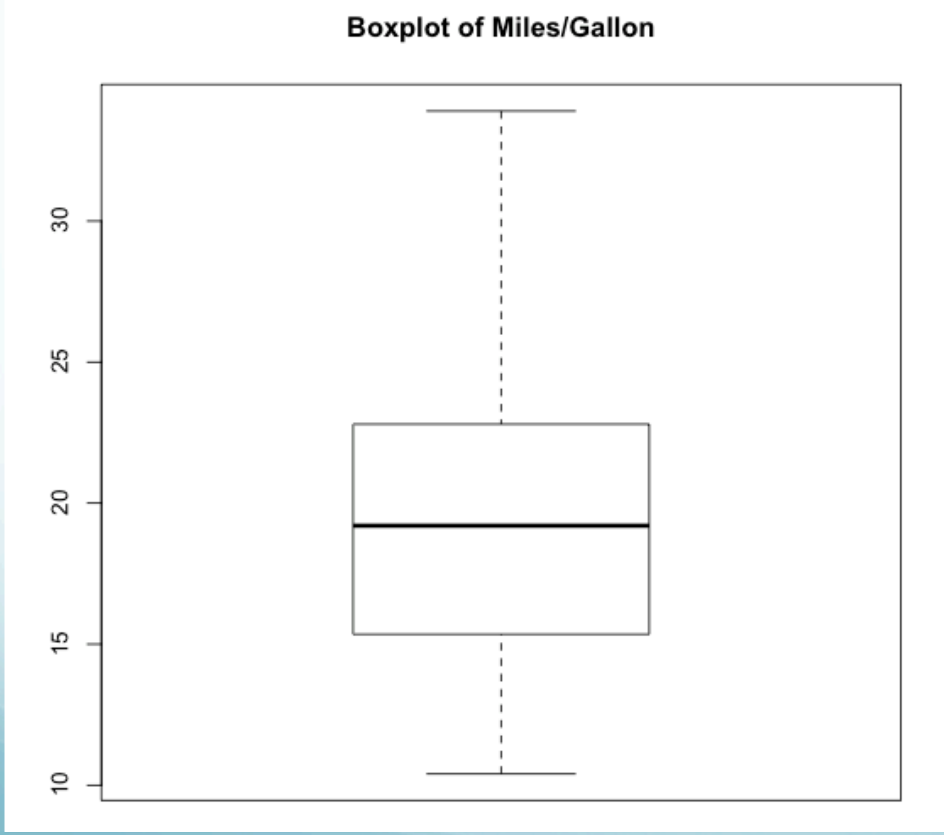


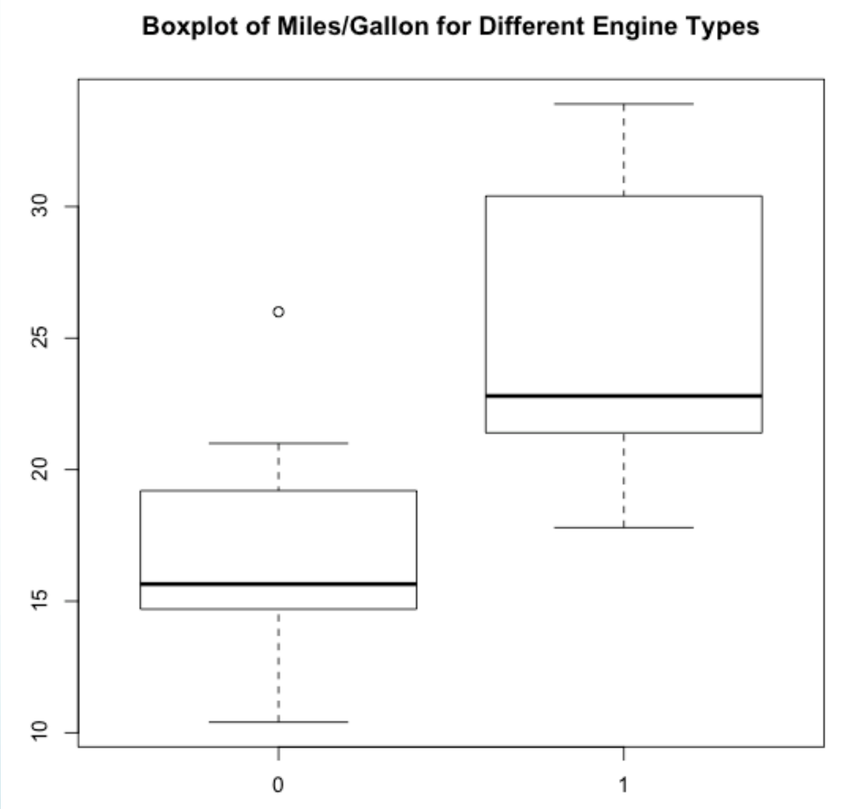


# R: boxplots

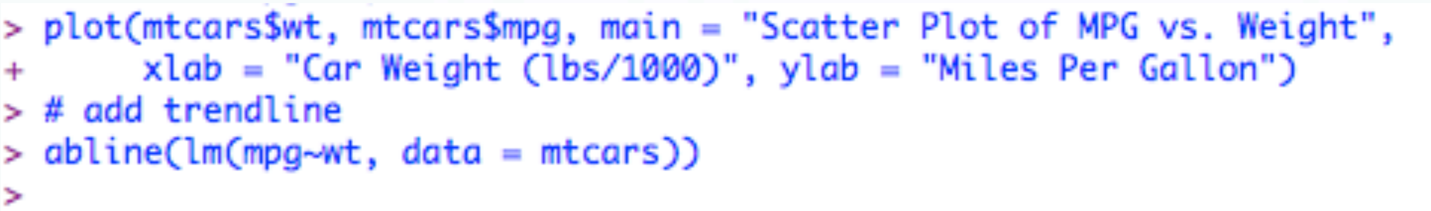


```
> boxplot(mtcars$mpg, main = "Boxplot of Miles/Gallon")
>
> boxplot(mtcars$mpg ~ factor(mtcars$vs), main = "Boxplot of Miles/Gallon for Different Engine Types")
```

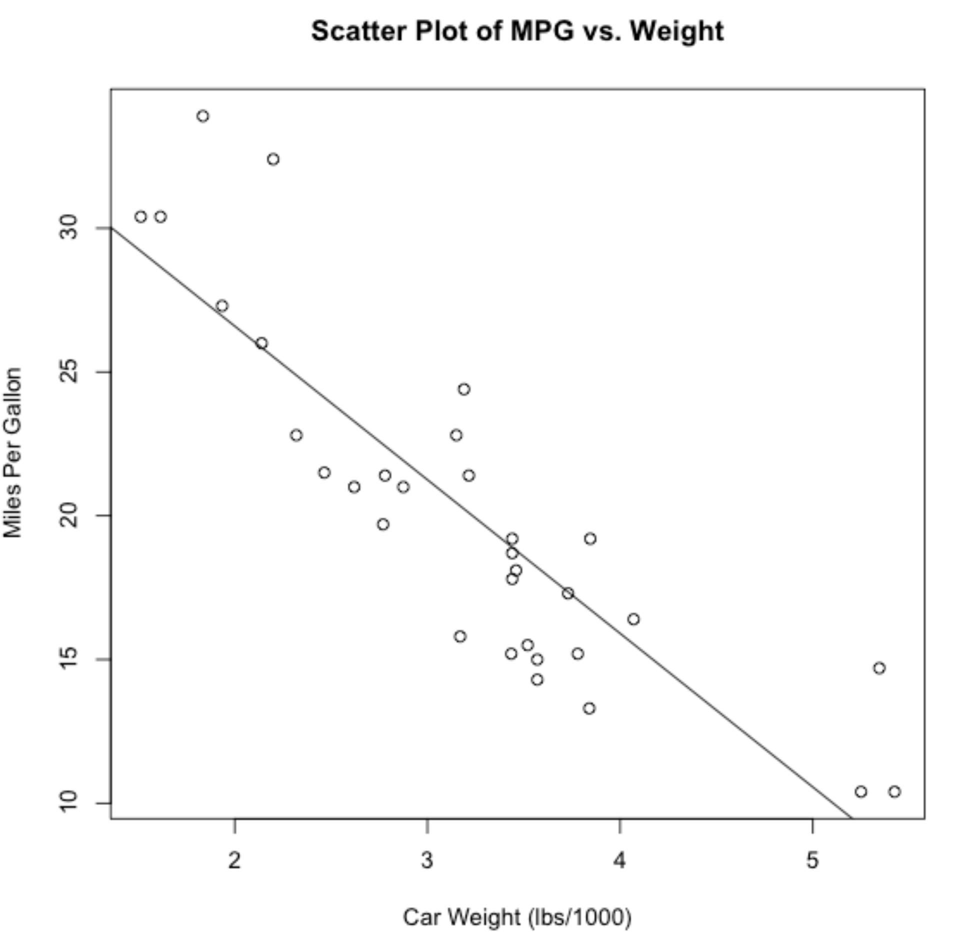




# R: scatterplot



```
> plot(mtcars$wt, mtcars$mpg, main = "Scatter Plot of MPG vs. Weight",
+      xlab = "Car Weight (lbs/1000)", ylab = "Miles Per Gallon")
> # add trendline
> abline(lm(mpg~wt, data = mtcars))
>
```


Scatter Plot of MPG vs. Weight
Miles Per Gallon
Car Weight (lbs/1000)