

Correlation continued and a review of descriptive statistics

Overview

Review and continuation of scatter plots and correlation

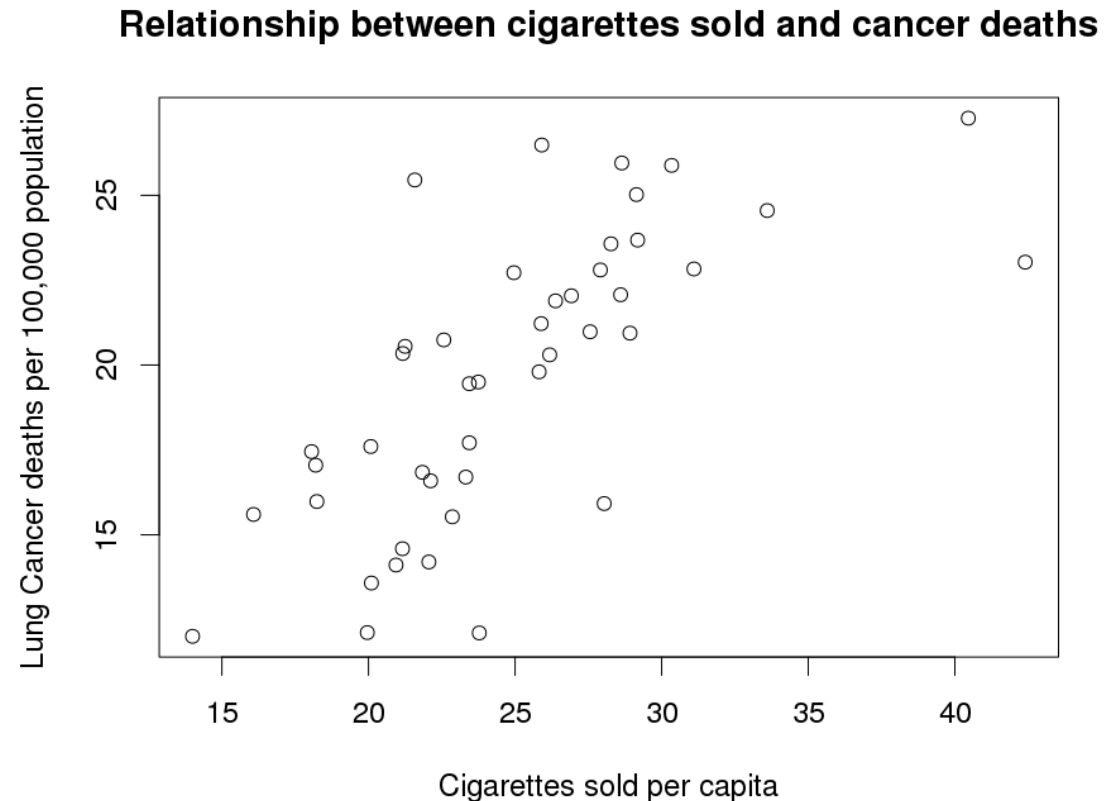
Worksheet 4

Review of what we have covered in descriptive statistics

Review of relationships between measures

Q: What is this type of plot called?

A: A scatter plot



R: `plot(x, y)`

Positive, negative, no correlation?

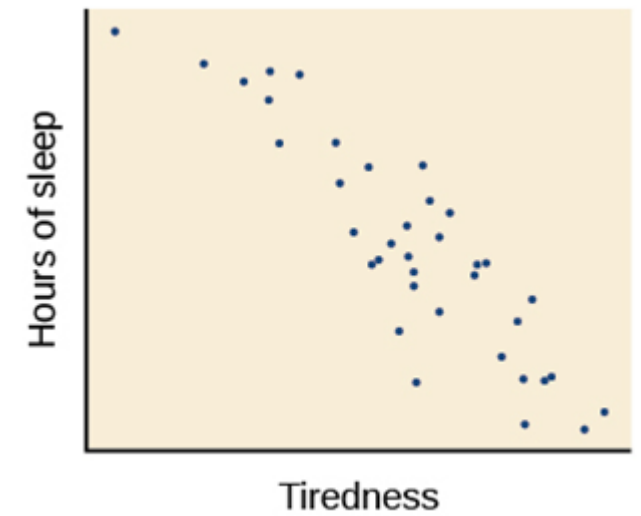
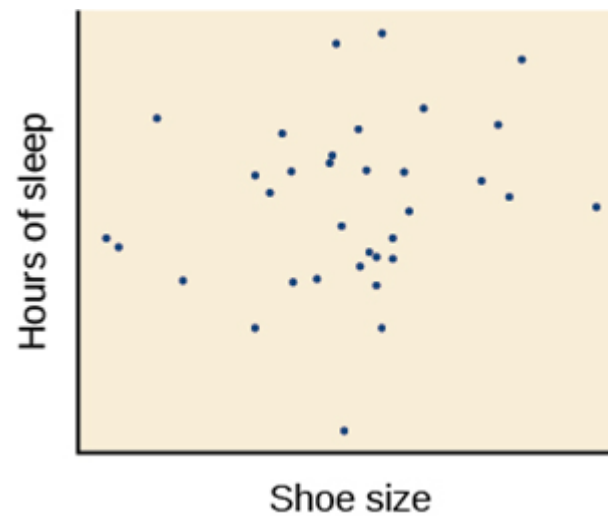
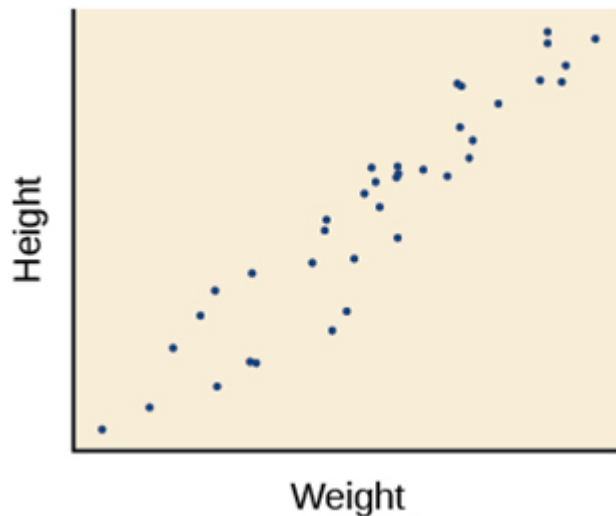
Do the points show a clear trend?

Does it go upward or downward?

How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

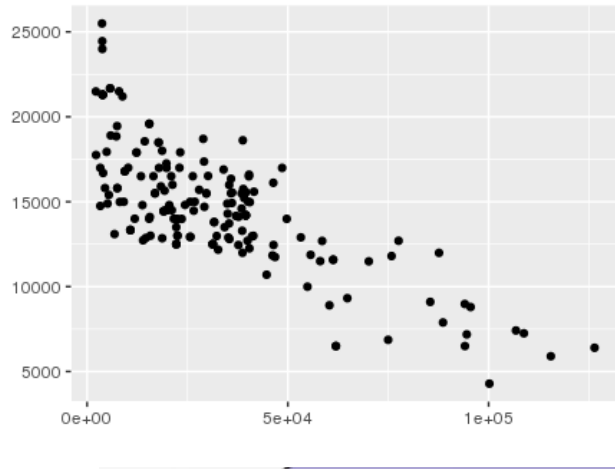


The correlation coefficient

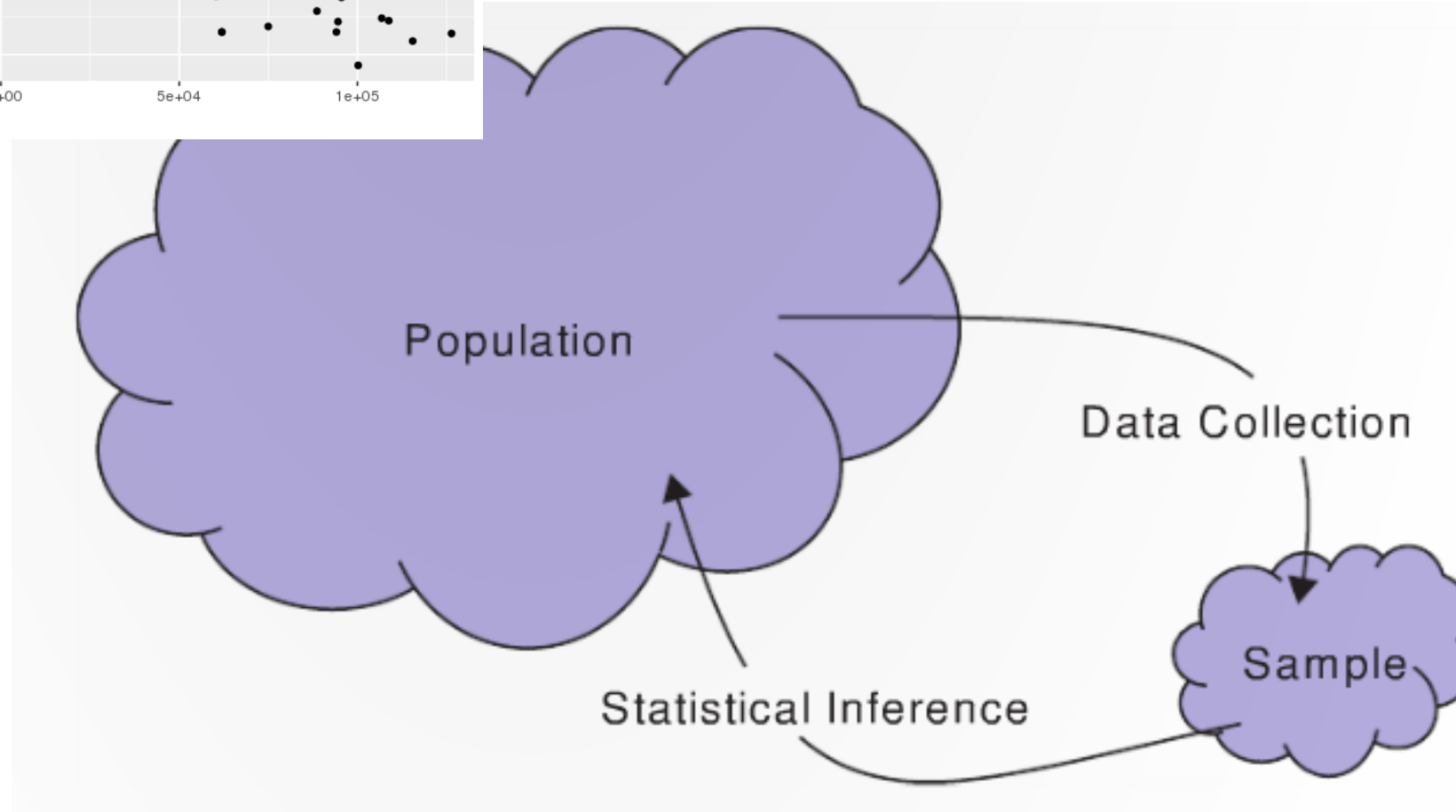
The **correlation** is measure of the strength and direction of a linear association between two variables

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

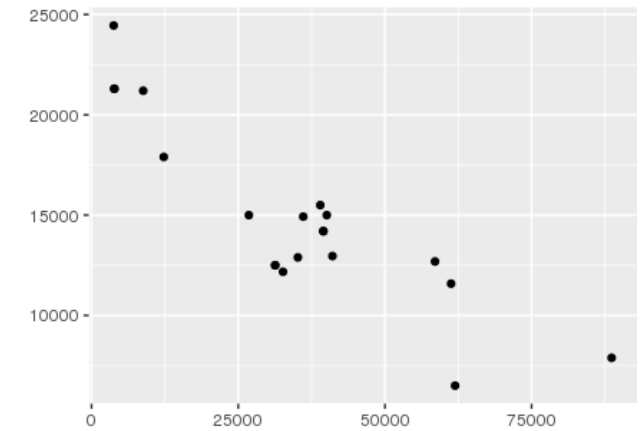
R: `cor(x, y)`



ρ parameter



r statistic



Properties of the correlation

Correlation is always between -1 and 1: $-1 \leq r \leq 1$

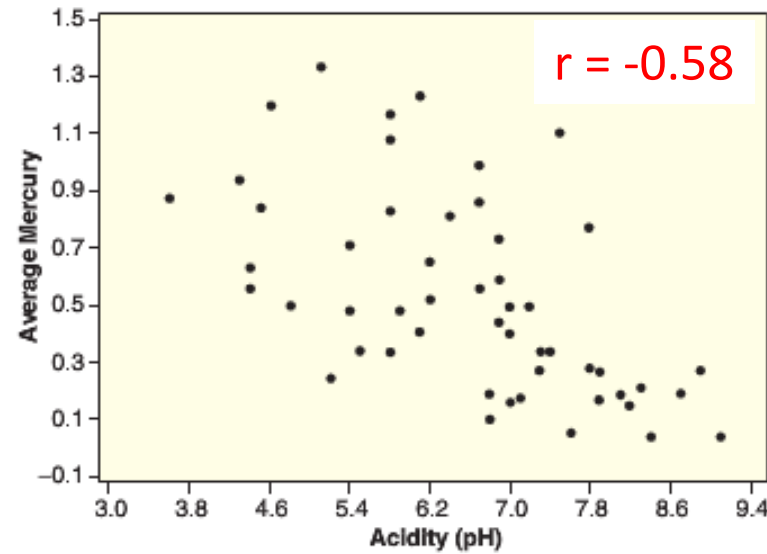
The sign of r indicates the direction of the association

Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

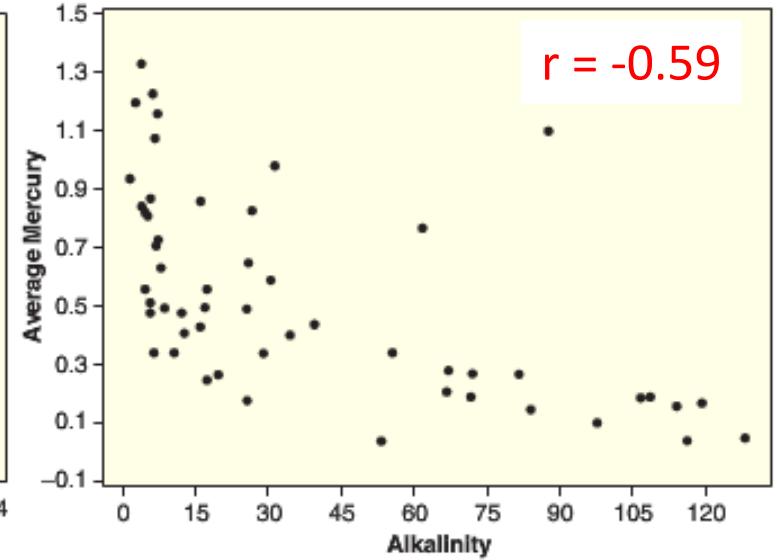
Correlation is symmetric: $r = \text{cor}(x, y) = \text{cor}(y, x)$

Florida lakes

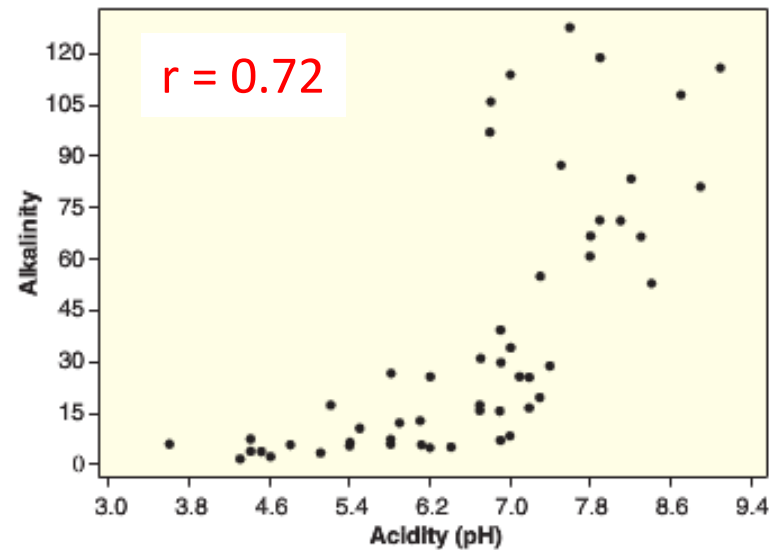
Correlation game



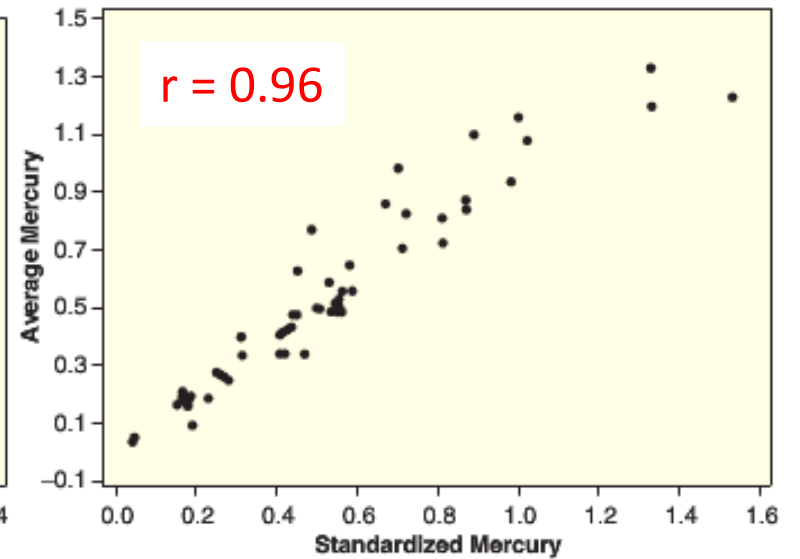
(a) Average mercury level vs acidity



(b) Average mercury level vs alkalinity



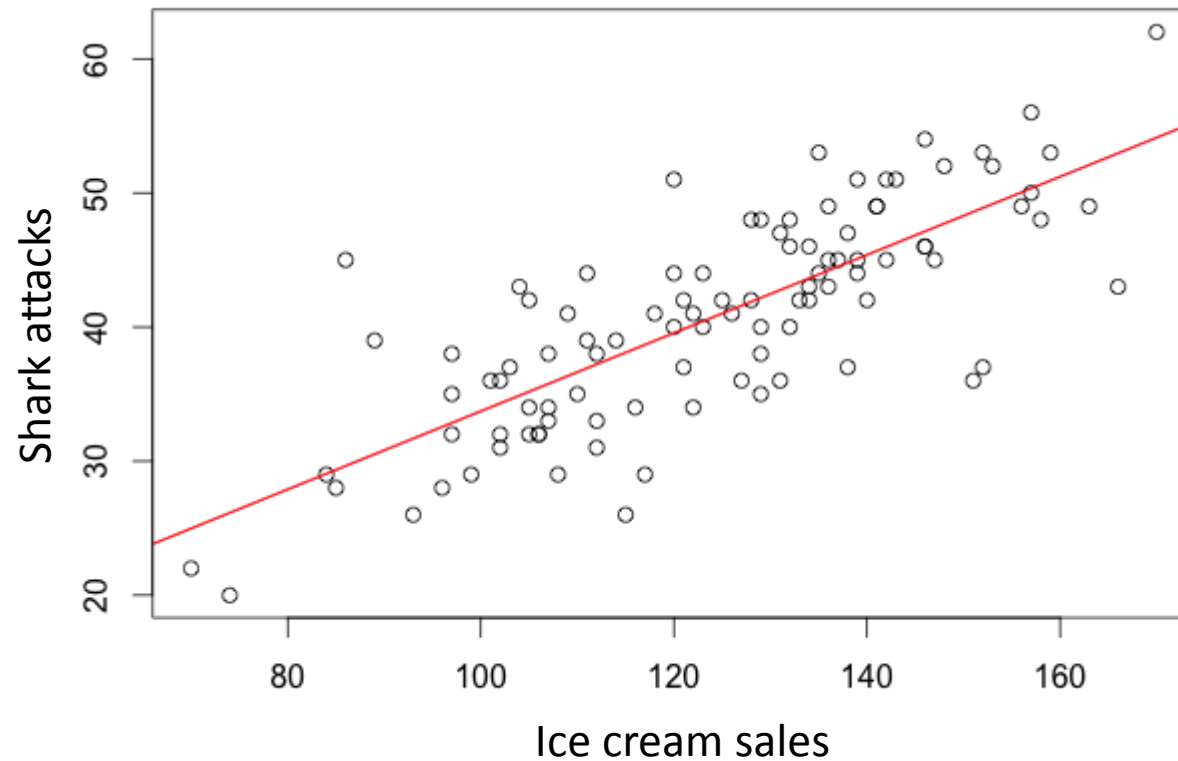
(c) Alkalinity vs acidity



(d) Average vs standardized mercury levels

Correlation caution #1

A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables

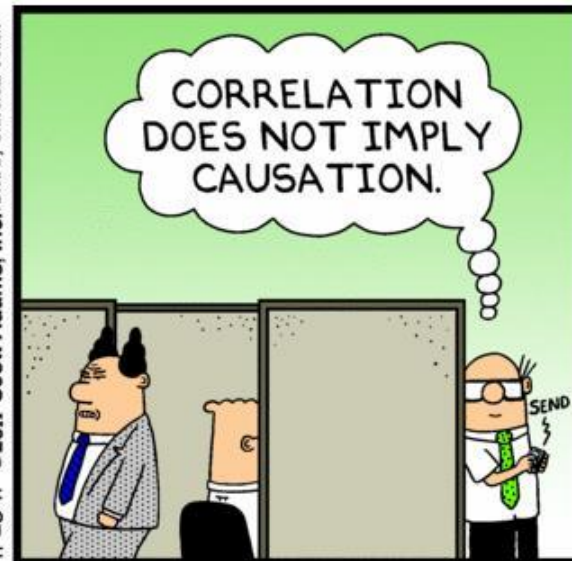




Dilbert.com DilbertCartoonist@gmail.com



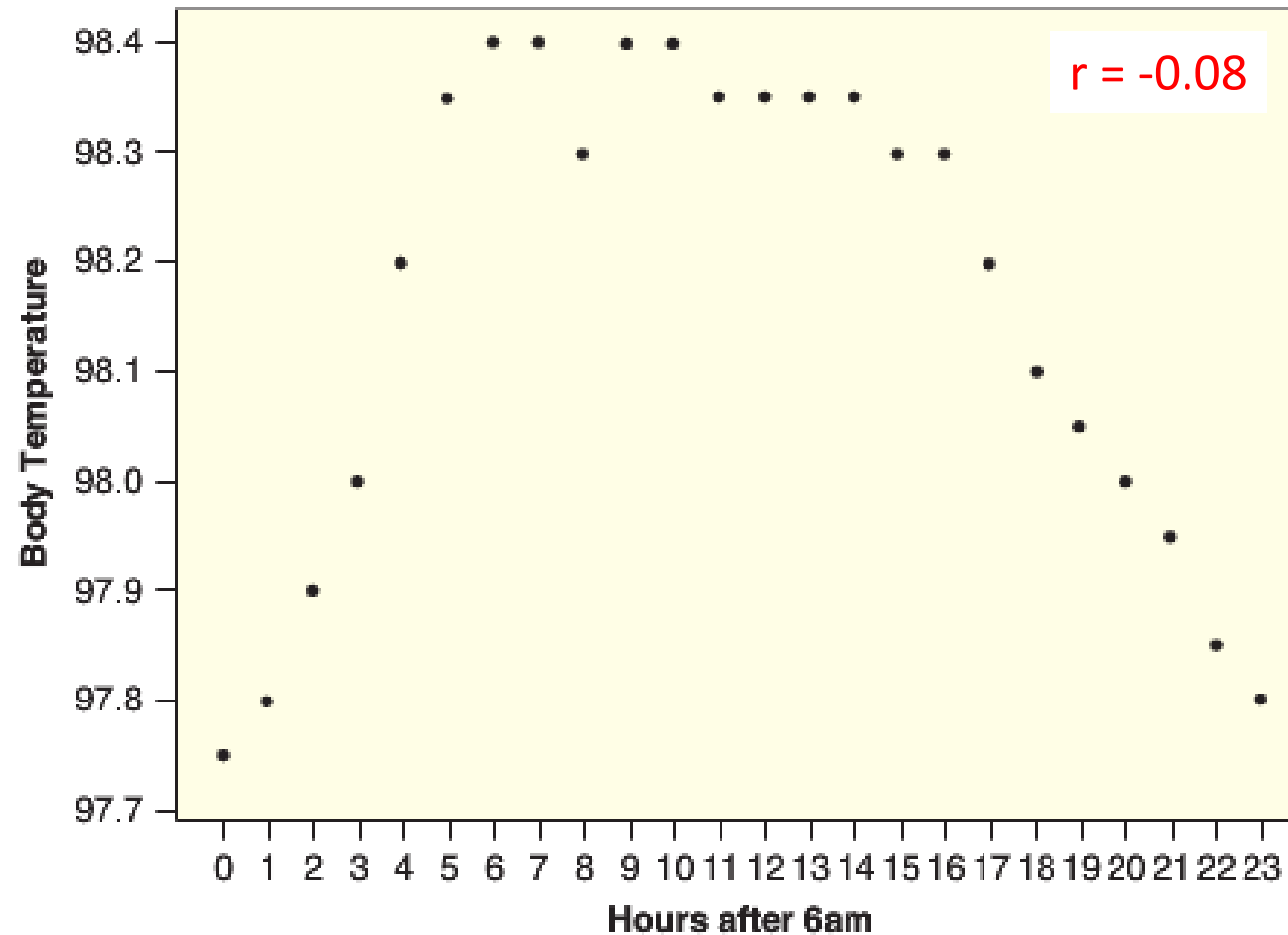
11-28-11 © 2011 Scott Adams, Inc. /Dist. by Universal Uclick



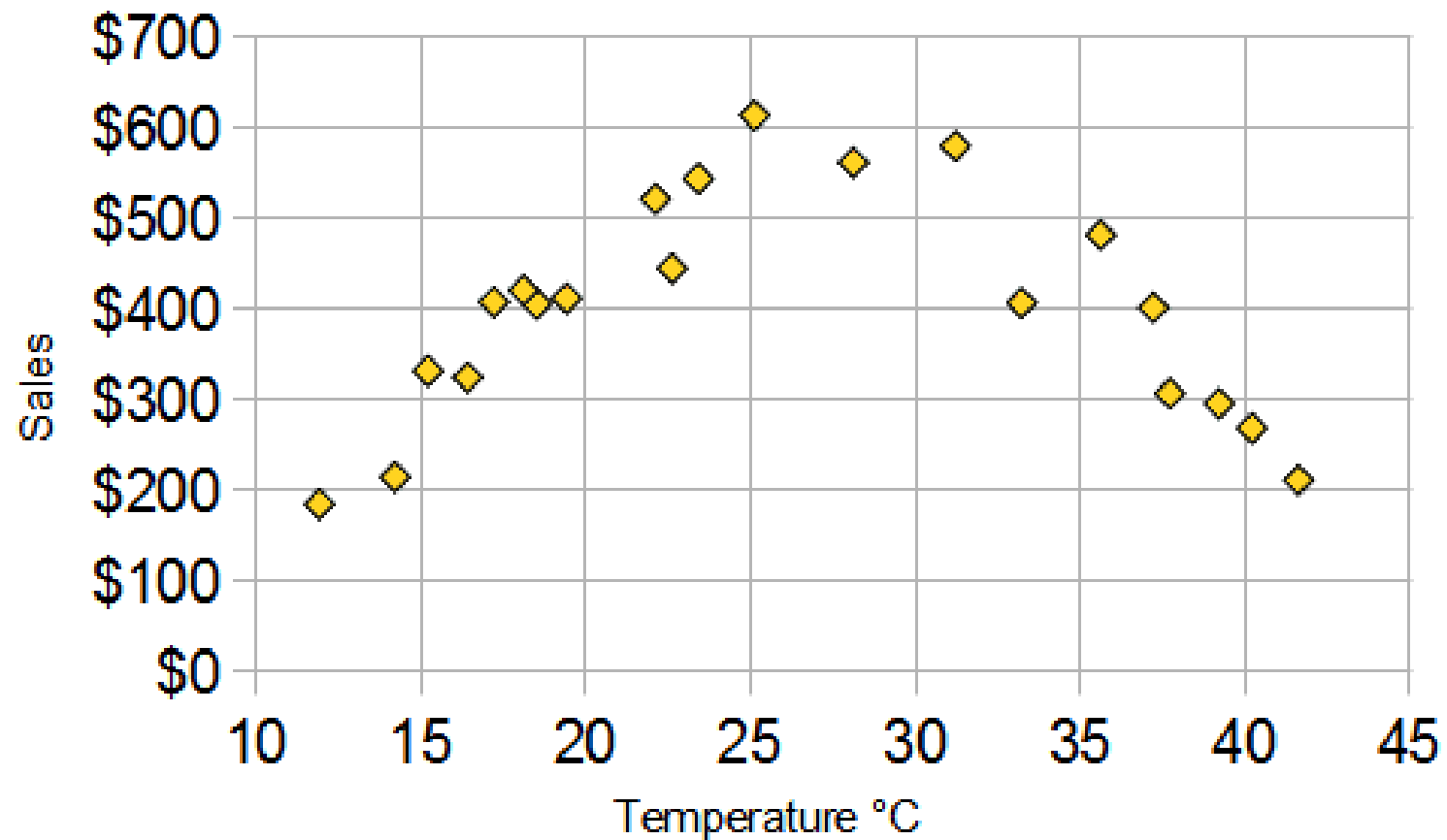
Correlation caution #2

A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.

Body temperature as a function of time of the day

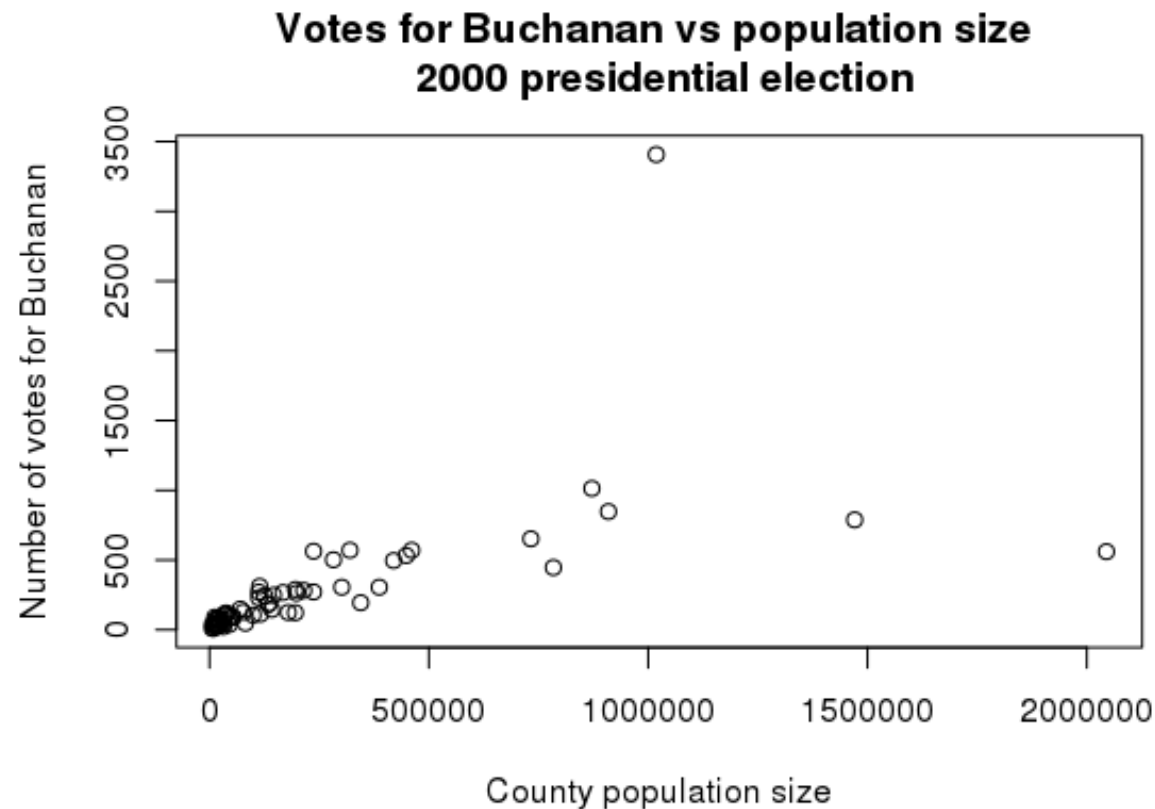


Ice cream sales and temperature



Correlation caution #3

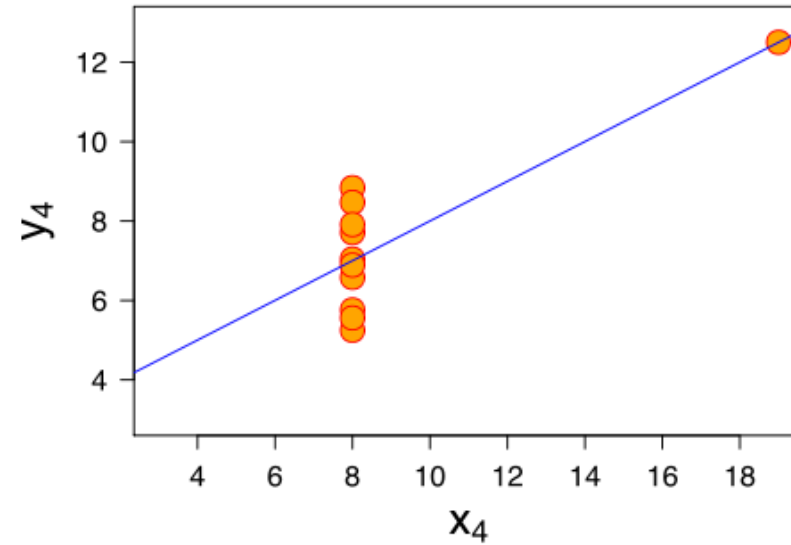
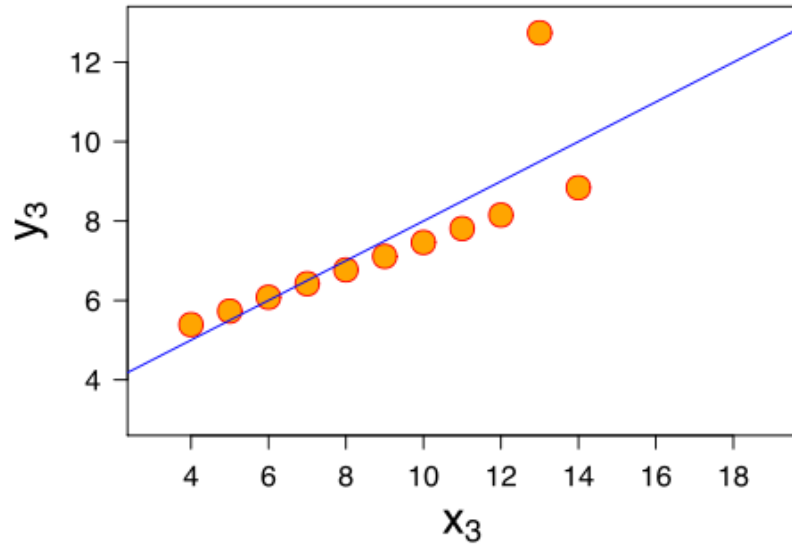
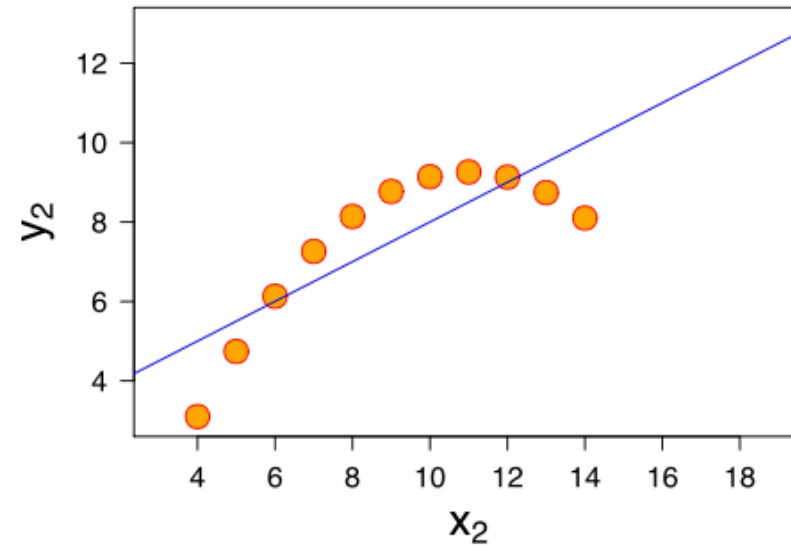
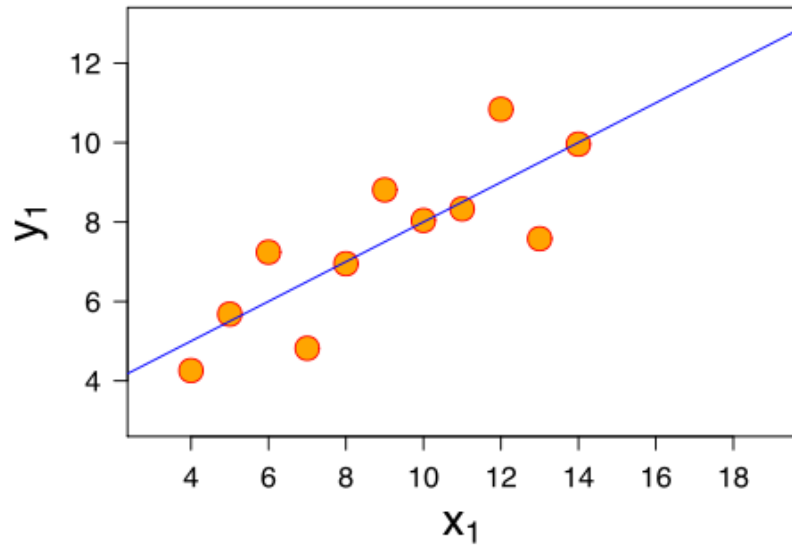
Correlation can be heavily influenced by outliers. Always plot your data!



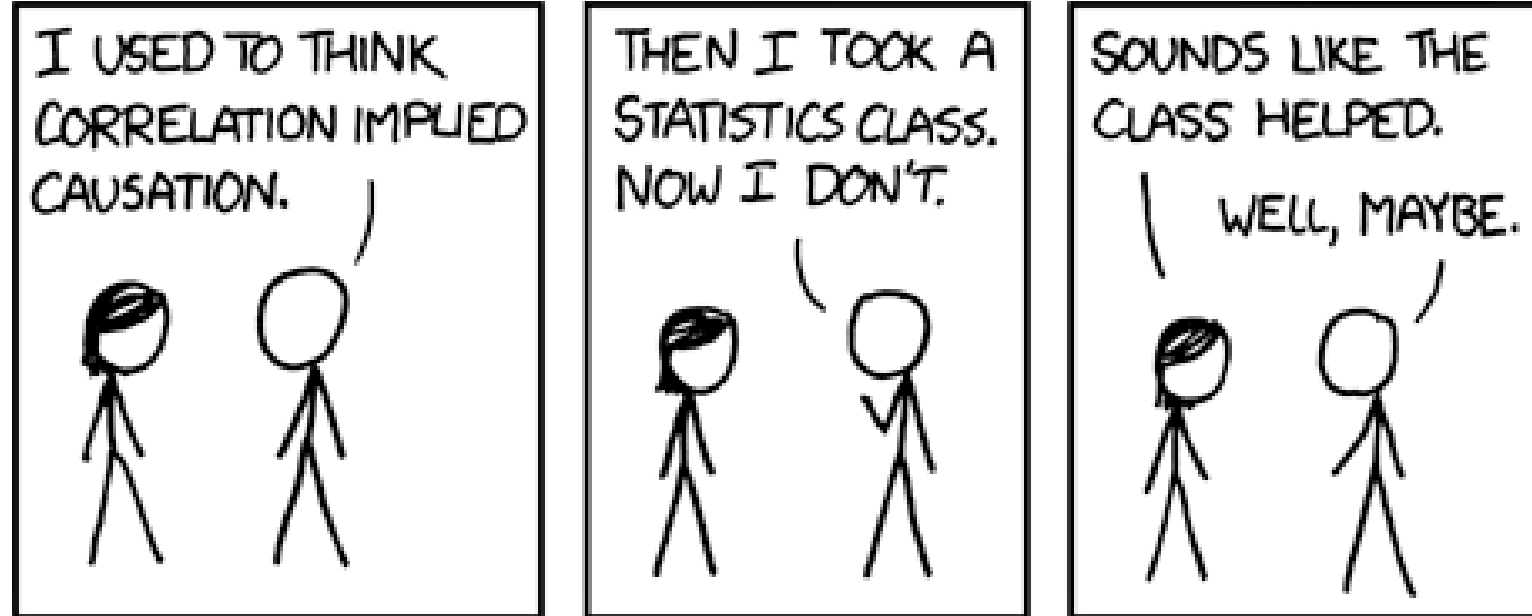
With Palm Beach
 $r = 0.61$

Without Palm Beach
 $r = .78$

Anscombe's quartet ($r = 0.81$)



Questions?



Worksheet 4

Worksheet is due at 11:59pm on Sunday October 7th

1. Load some class specific functions using the code

```
> source("/home/shared/intro_stats/cs206_functions.R")
```

2. Go to the console and copy the worksheet using the following commands:

```
> get_worksheet(4)
```

A list of function we used in class can be found at: <https://goo.gl/LxrU7J>

Worksheet 4

Worksheet is due at 11:59pm on Sunday October 7th

Part 2 involves exploring data:

<https://vincentarelbundock.github.io/Rdatasets/datasets.html>

Example code:

```
> library(boot)
> data(beaver)
> ? beaver
> beaver <- data.frame(beaver) # careful, some of the data sets are not data frames
> View(beaver)
> beaver_temp_vector <- beaver$temp
```

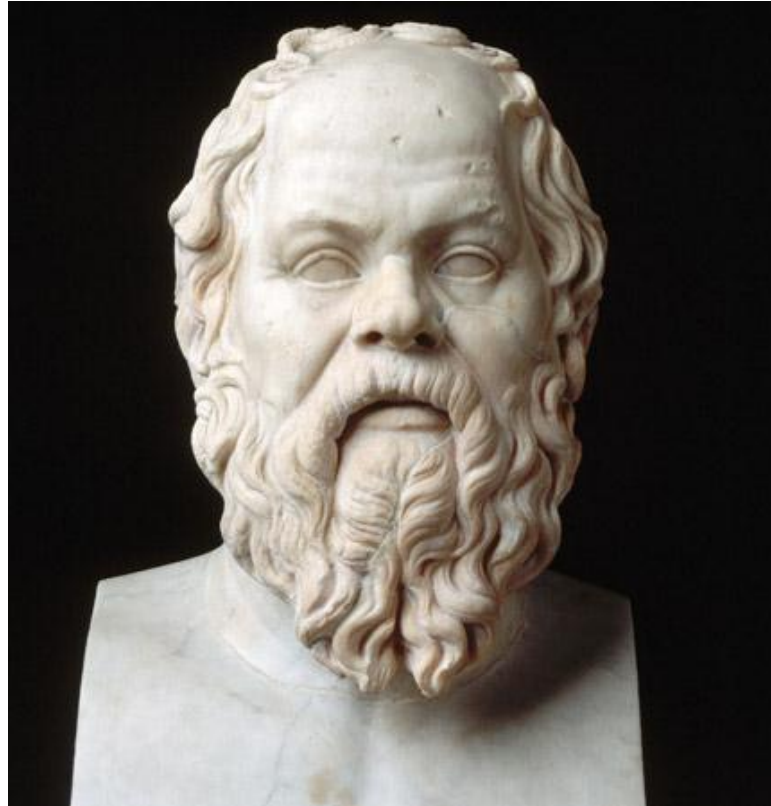
Review of what we have
covered so far...

Topics we have covered in class

1. Intro to data
2. Parameters and statistics
3. Categorical data statistics and visualization
4. Quantitative data
 - Visualization
 - Measures of central tendency
 - Measures of spread
 - Statistics and visualization for two quantitative variables

The TRUTH!

Who is this?



What can he tell us about Statistics?

1. Intro to data

What is Statistics?

What are...

Observational units?

Variables?

Categorical variables?

Quantitative variables?

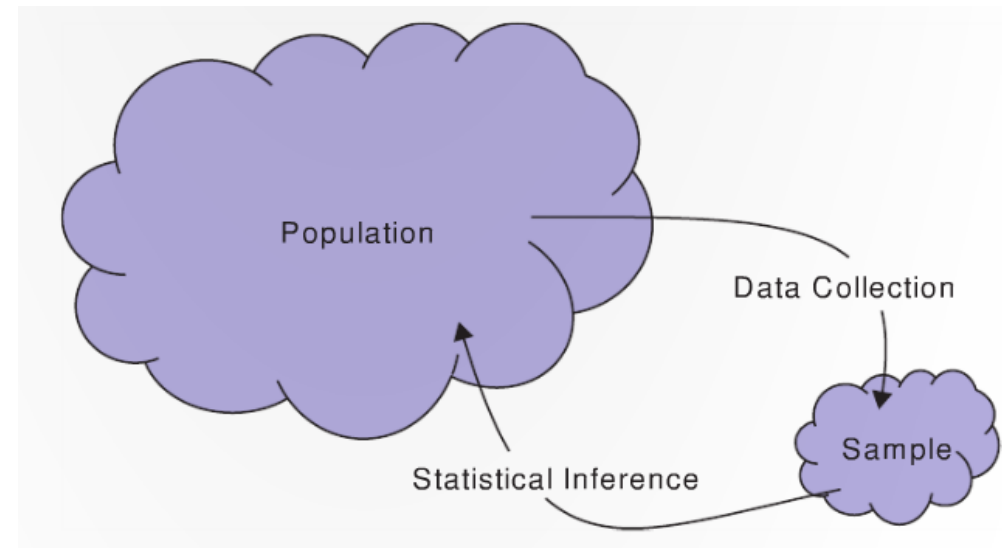
	flight 	date 	carrier 	origin 	dest 	air_time 	arr_delay 
1	1545	1-1-2013	UA	EWR	IAH	227	11
2	1714	1-1-2013	UA	LGA	IAH	227	20
3	1141	1-1-2013	AA	JFK	MIA	160	33
4	725	1-1-2013	B6	JFK	BQN	183	-18
5	461	1-1-2013	DL	LGA	ATL	116	-25
6	1696	1-1-2013	UA	EWR	ORD	150	12
7	507	1-1-2013	B6	EWR	FLL	158	19

2. Sampling

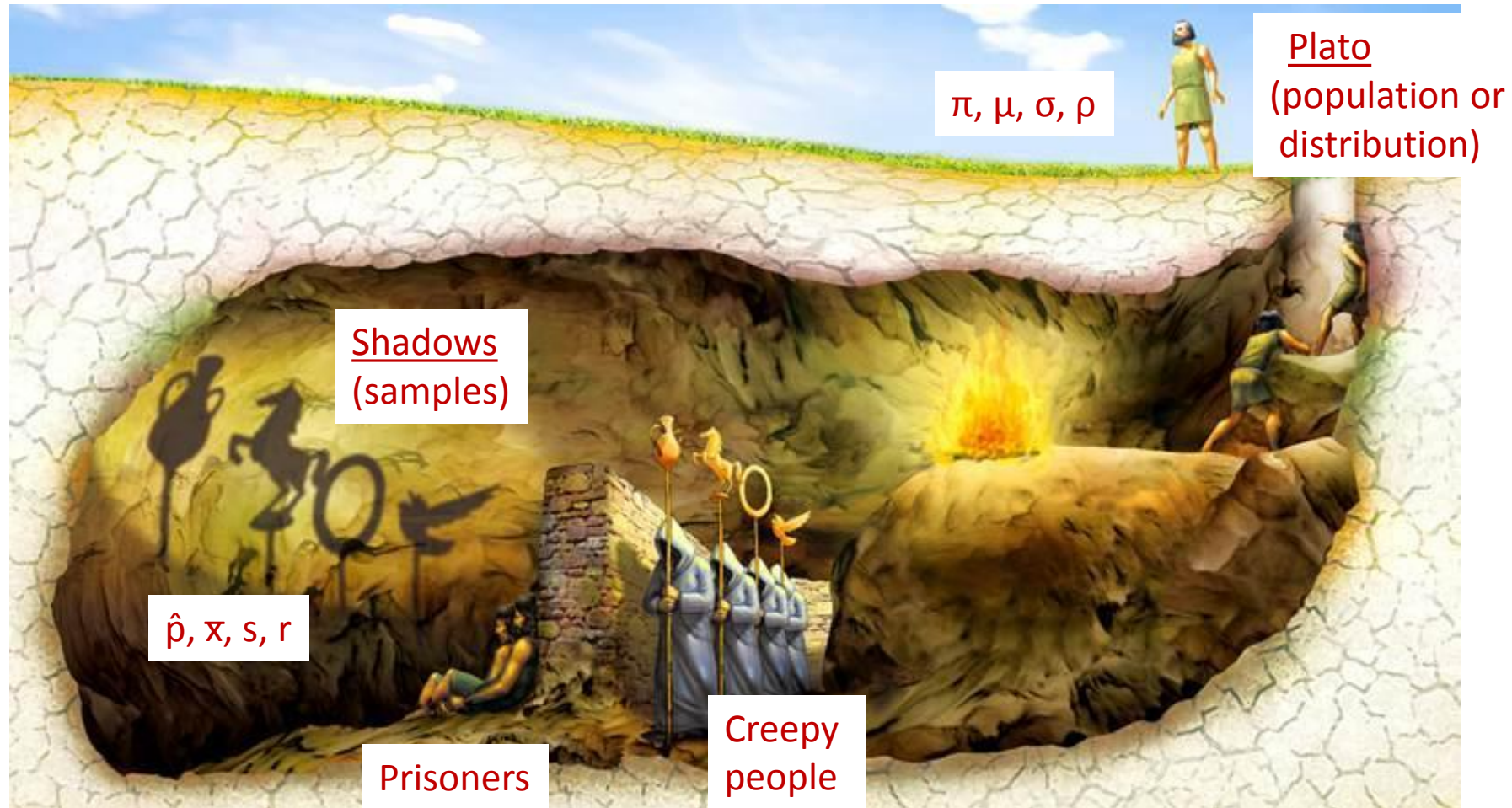
What is a ...?

- sample
- population
- statistic
- parameter

What is statistical inference?



Plato's cave

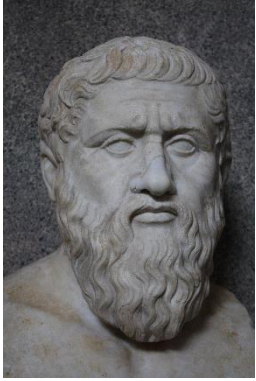


From The Republic (~ 380 BCE)

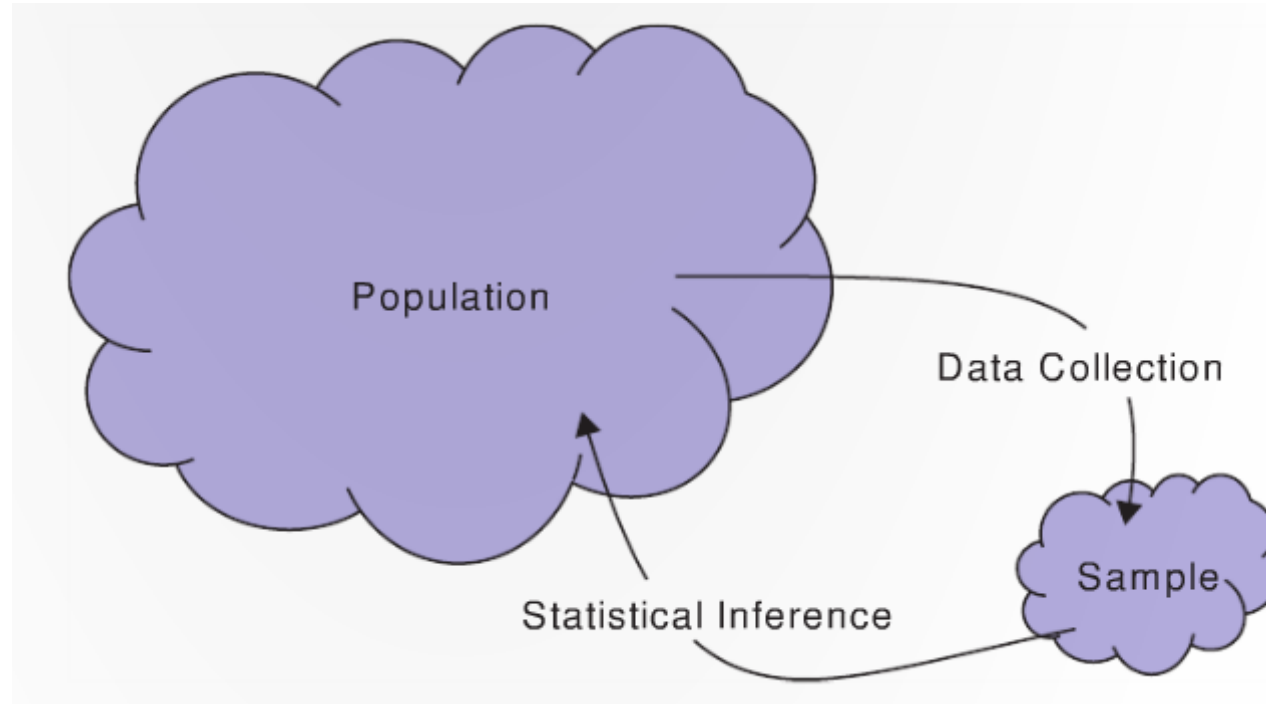
Quiz: parameters and statistics

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Proportion	$\hat{p}$	π
Correlation	r	ρ
Median	m	?

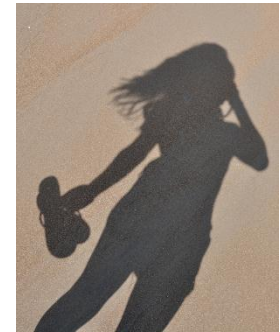
Population parameters vs. sample statistics



π, μ, σ, ρ



$\hat{p}, \bar{x}, s, r$

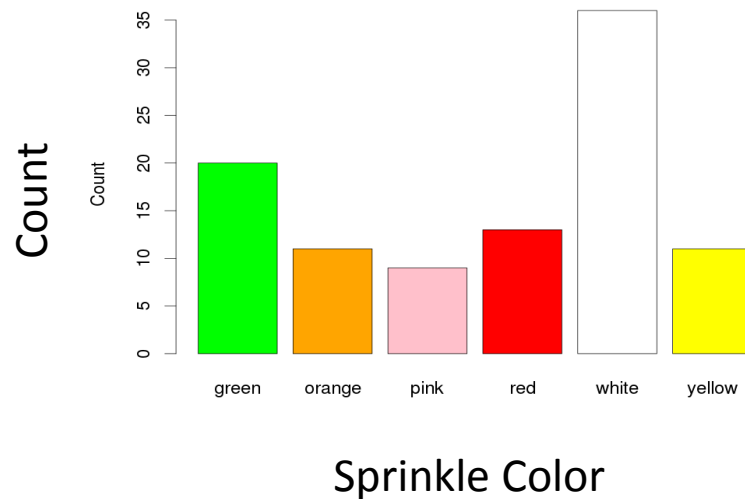


Categorical data

What is the main statistic we discussed for categorical data?

- π or $\hat{p}$
- proportion = number in category/total

How can we plot categorical data?



World's Most Accurate Pie Chart

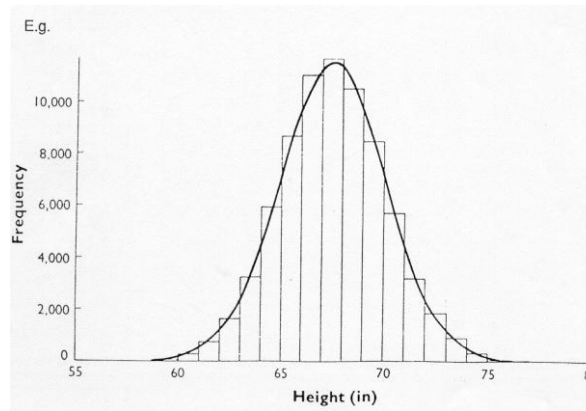
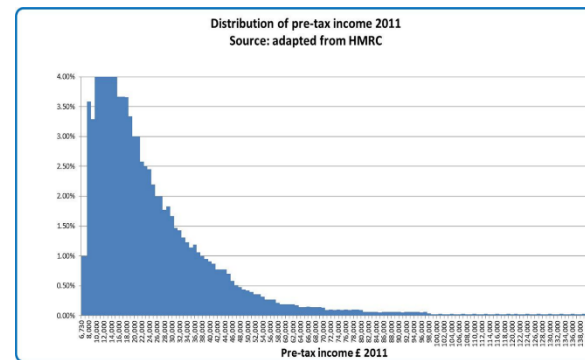


Quantitative data?

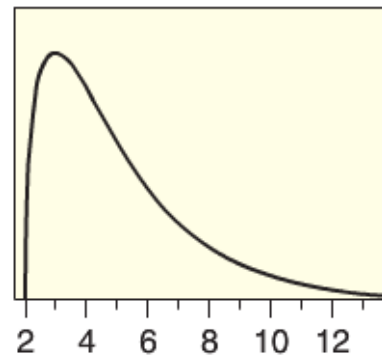
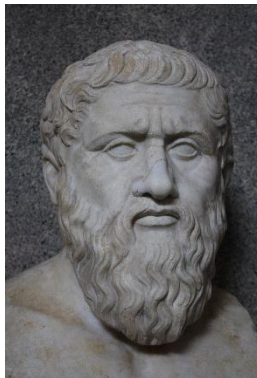
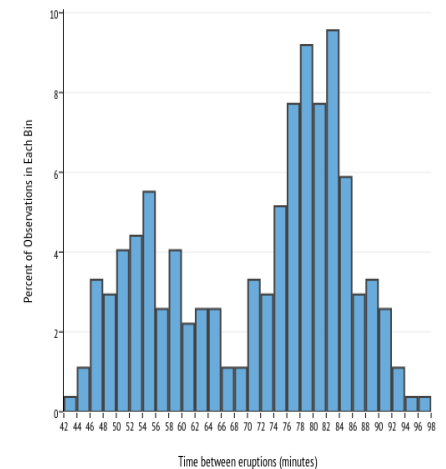
What is a good way to visualize the shape of quantitative data?



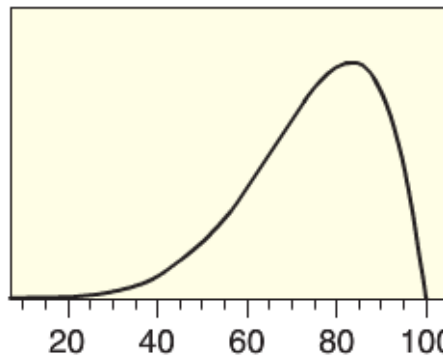
Income distribution



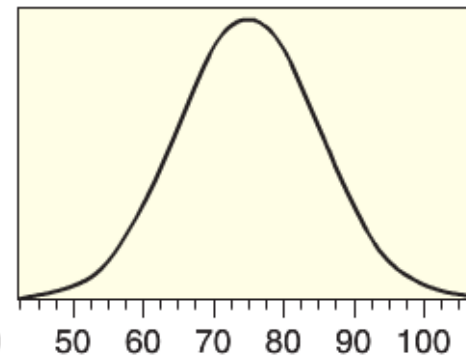
Old (Not So?) Faithful: A Bimodal Distribution



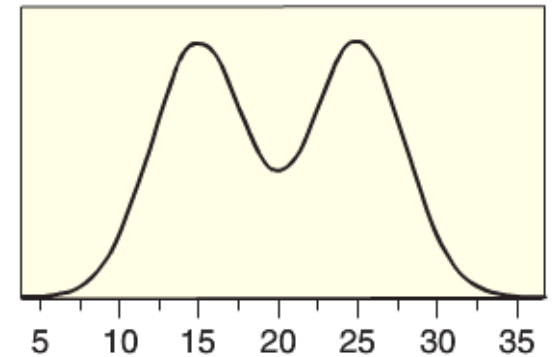
(a) Skewed to the right



(b) Skewed to the left

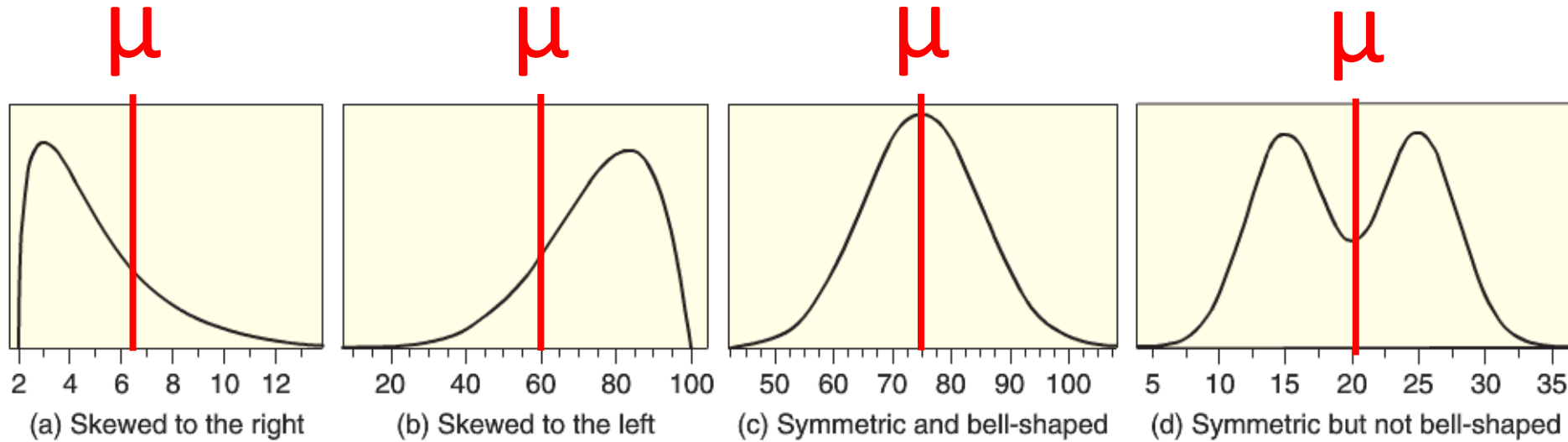


(c) Symmetric and bell-shaped



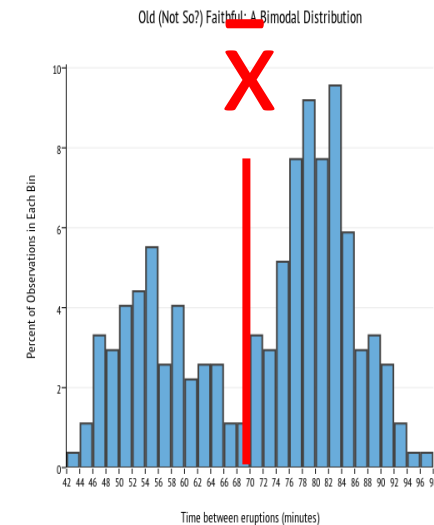
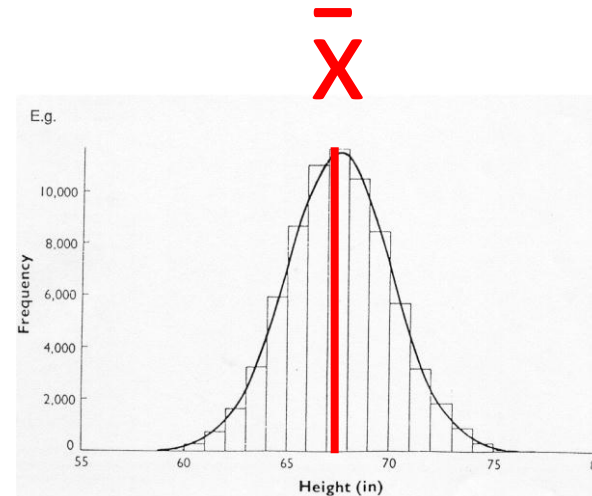
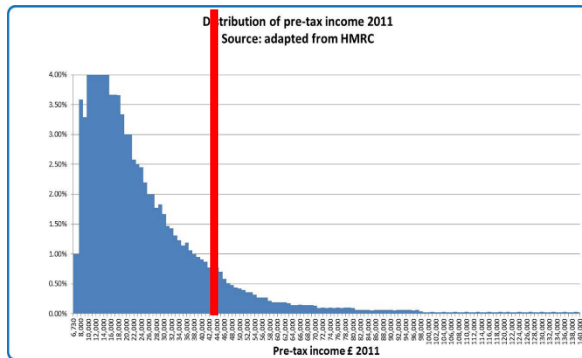
(d) Symmetric but not bell-shaped

Measure of central tendency: the mean

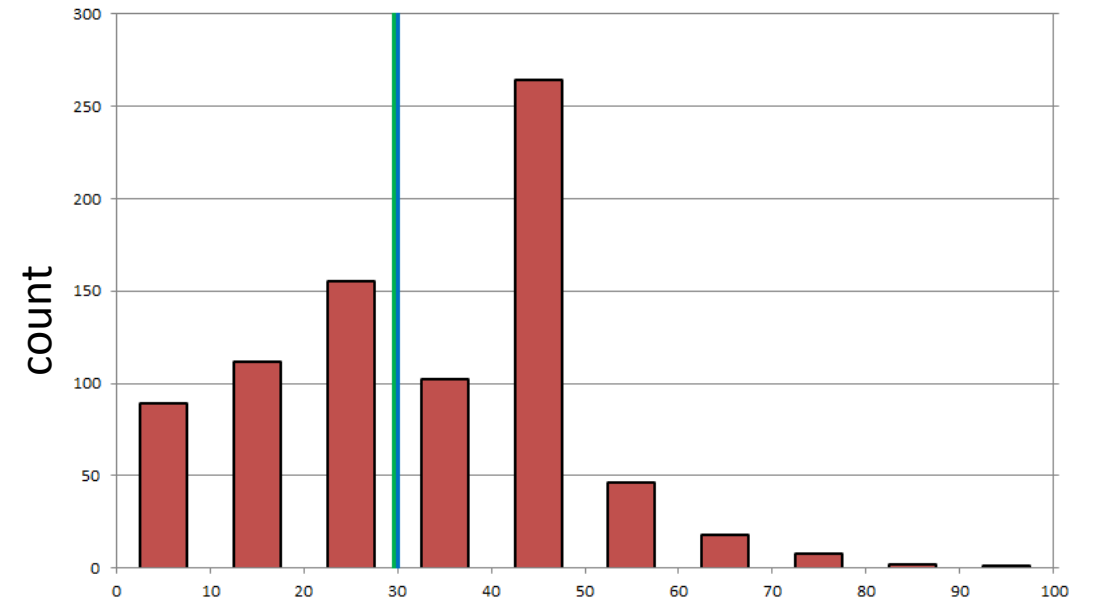
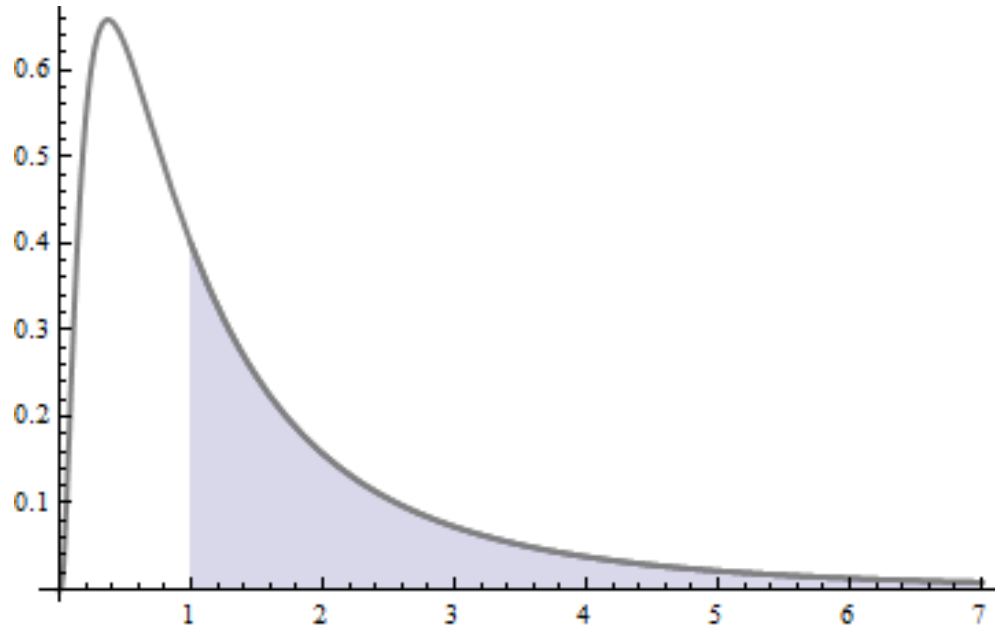


$$\frac{\sum_{i=1}^n x_i}{n}$$

$\bar{x}$



Measure of central tendency: the median

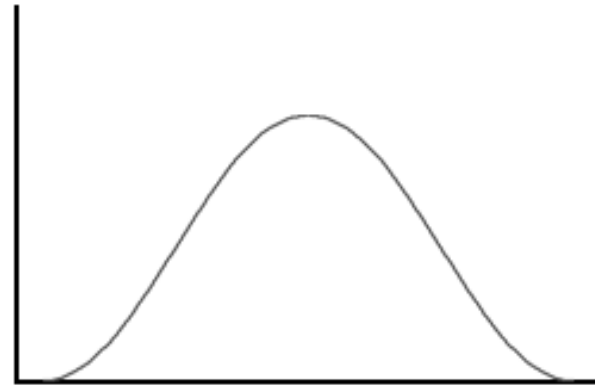
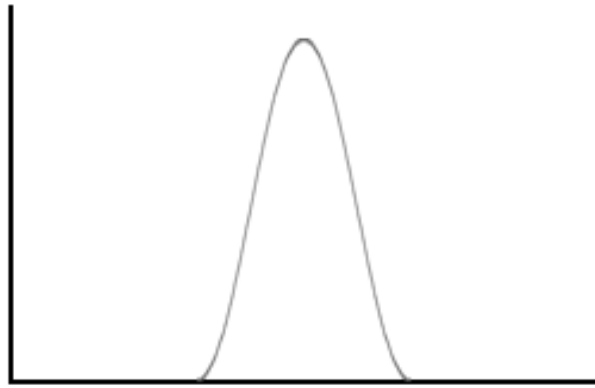


Which is resistant, the mean or the median?

The standard deviation

Which distribution has a larger standard deviation?

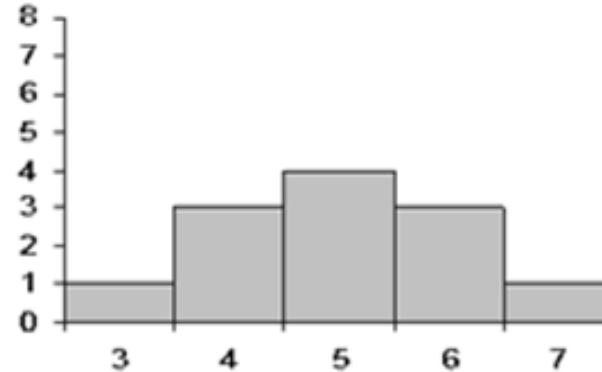
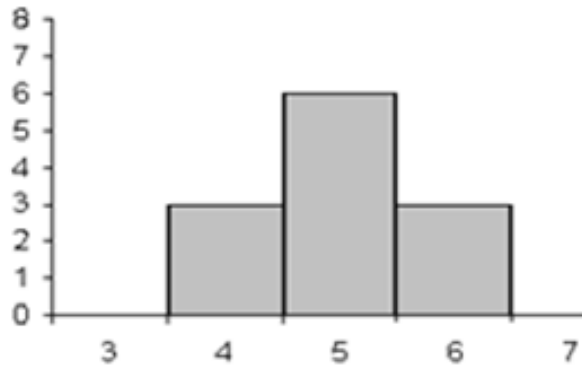
parameter σ



The standard deviation

Which distribution has a larger standard deviation?

statistic: s



What is the formula for the standard deviation?

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

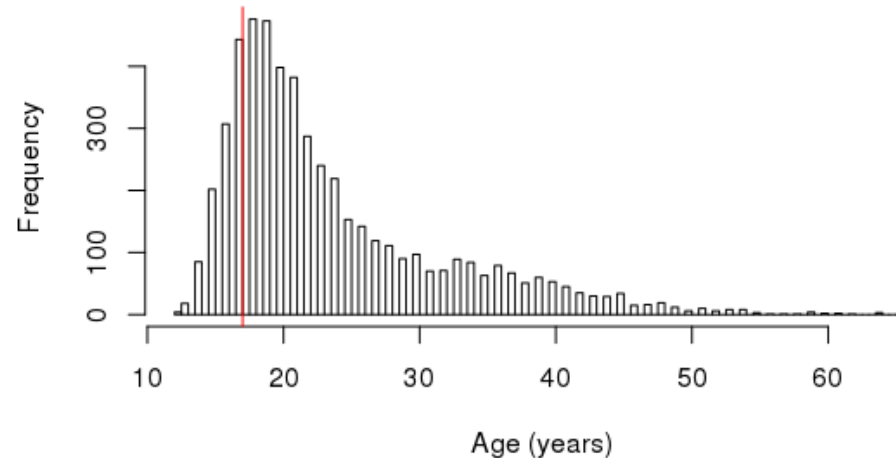
z-scores and percentiles

What is a z-score and why is it useful?

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{s}$$

What is the p^{th} percentile?

Histogram of Ages of people arrested for marijuana use

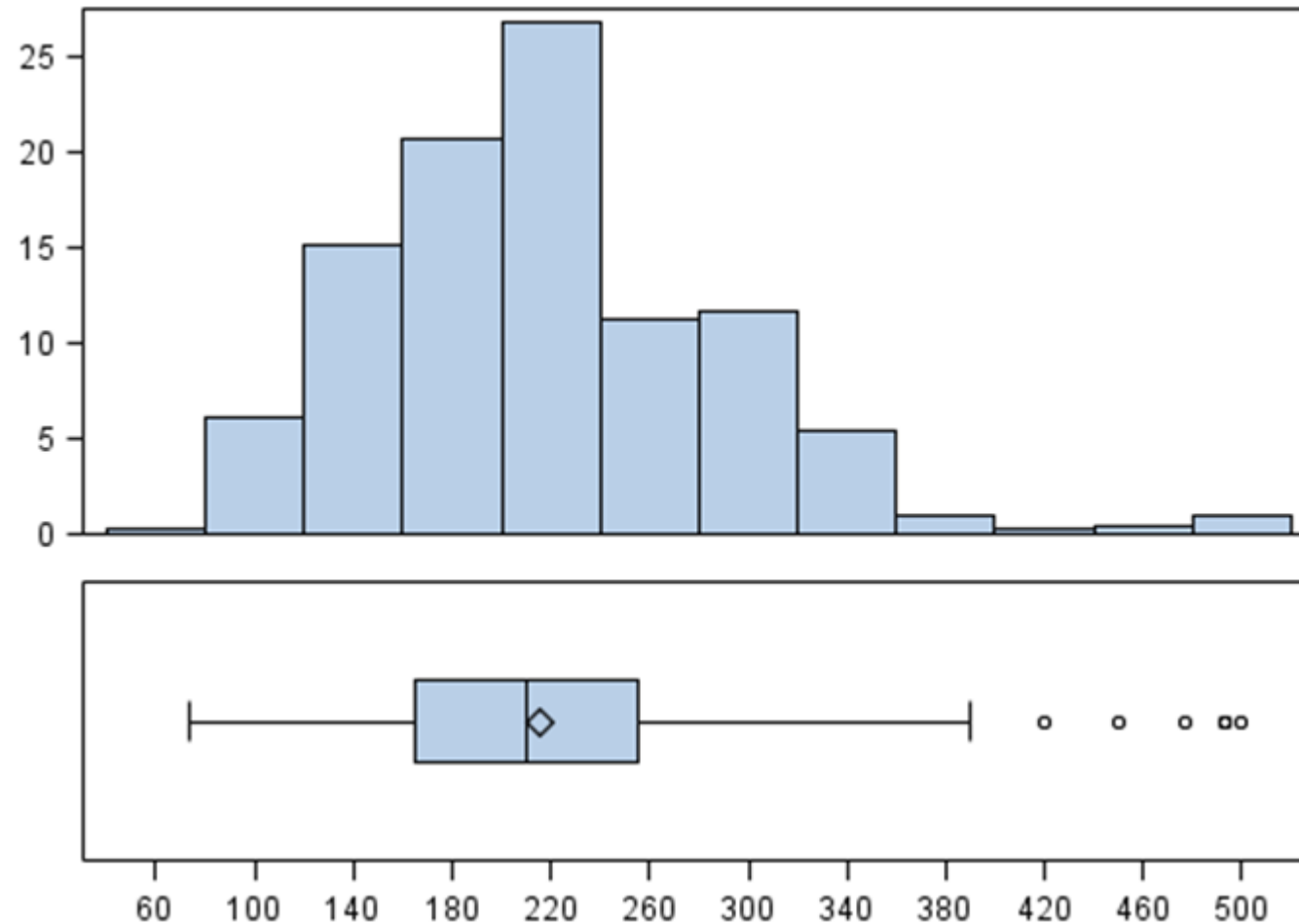


Normal pillow

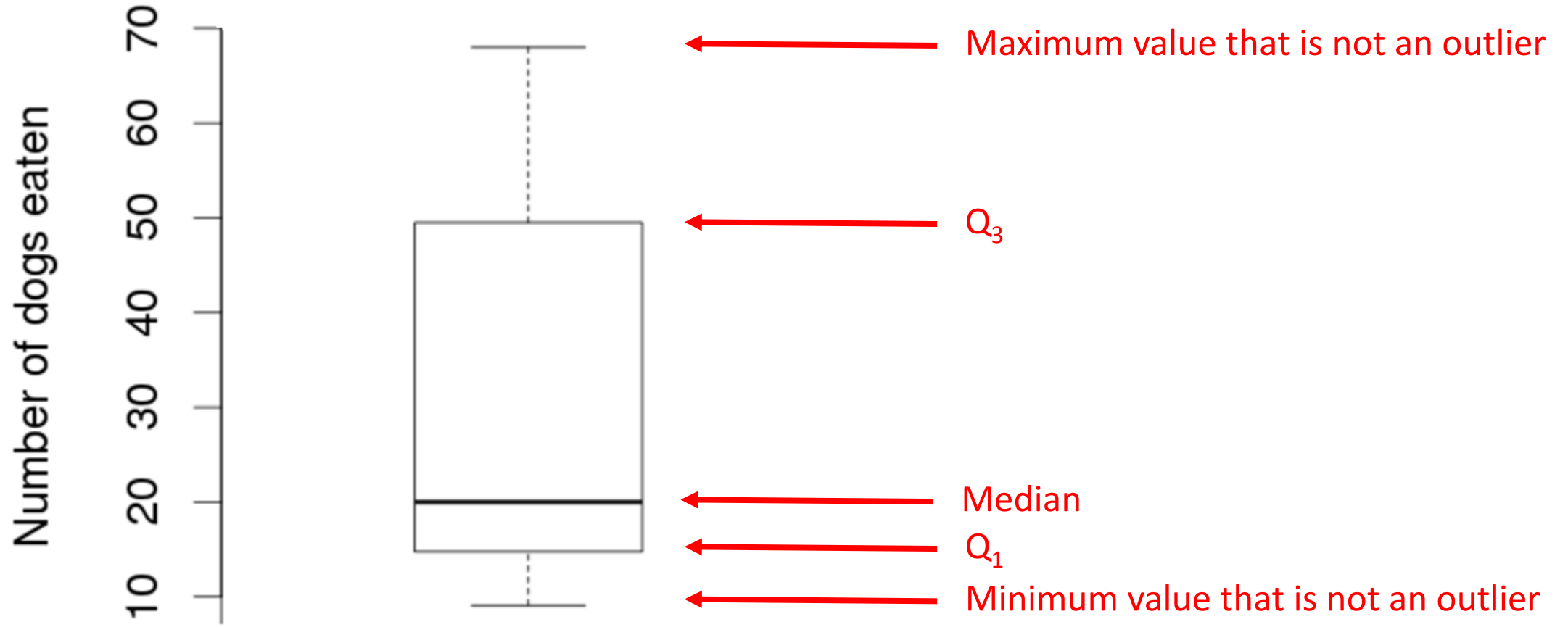


What percent of the pillow's mass is ± 2 standard deviations from the mean?

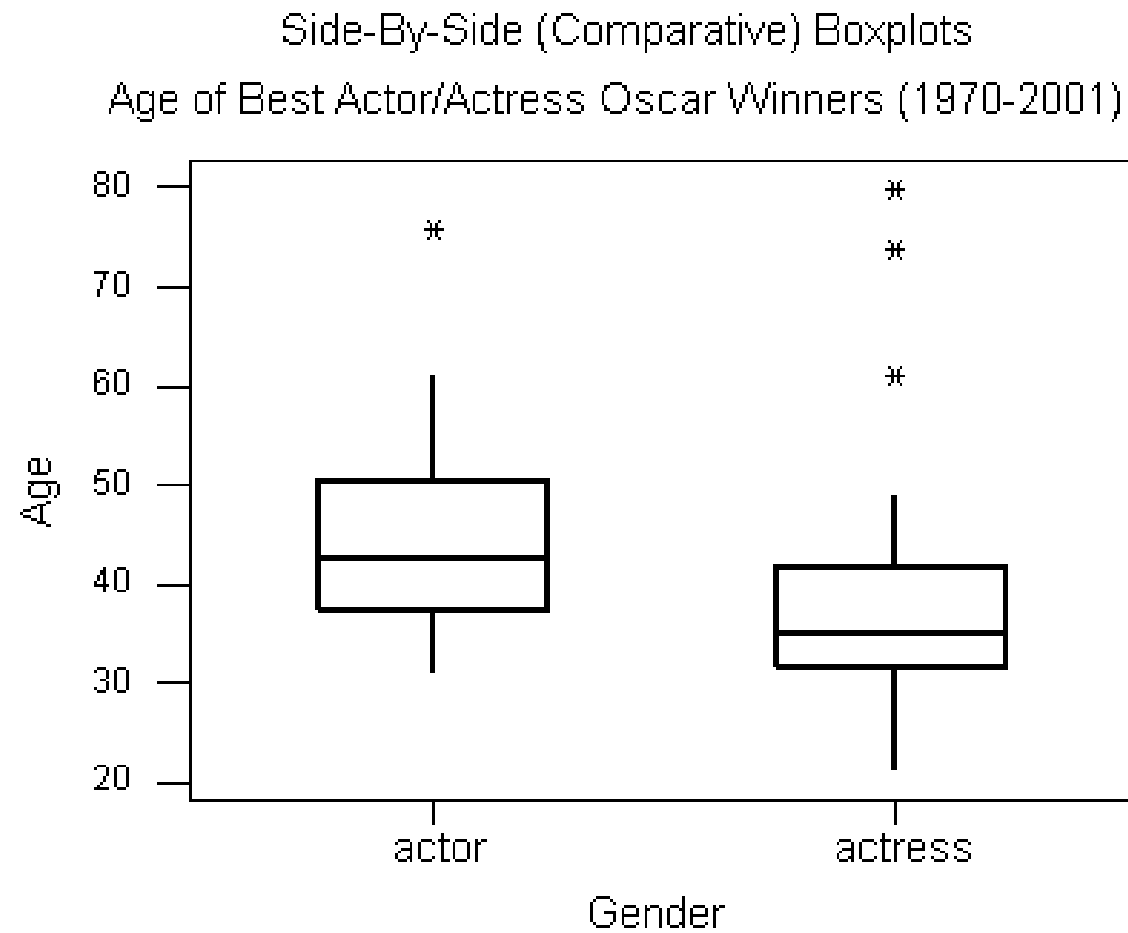
What is a 5 number summary and a boxplot?



What is a 5 number summary and a boxplot?

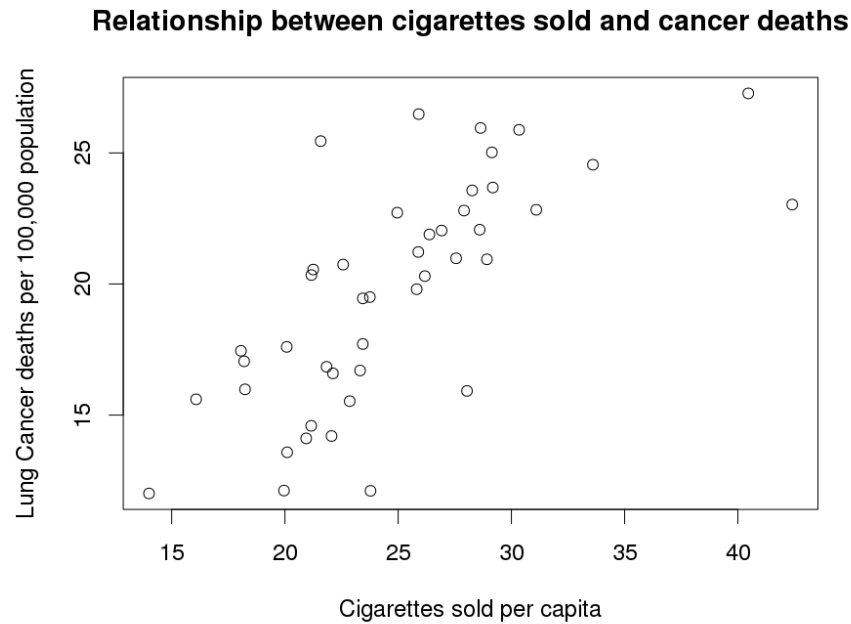


Side-by-side boxplots



Relationships between measures

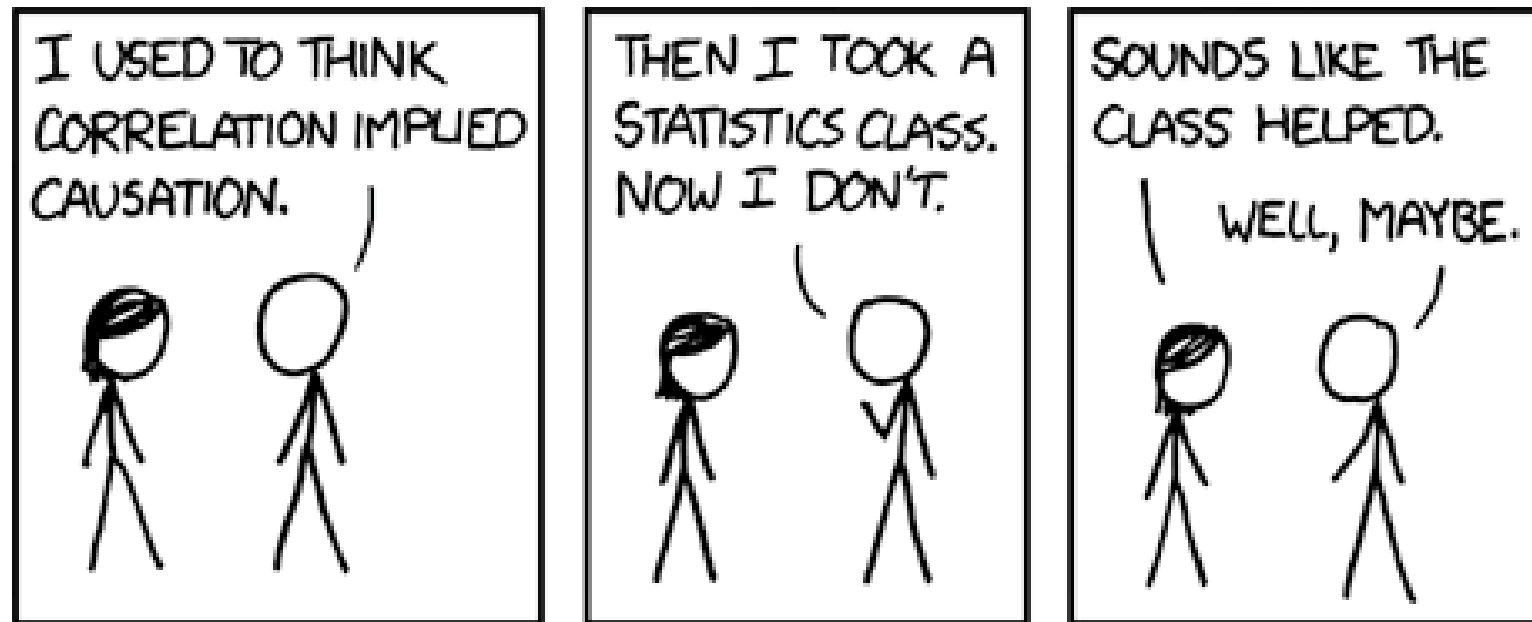
Q: What is this type of plot called?



Q: What statistic have we used to describe the linear relationship between quantitative variables?

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

Does correlation imply causation?



What is our primary focus in Statistics?

THE TRUTH IS OUT THERE



Have a good fall break!

Worksheet 4 due on 11:59pm on Sunday October 7th

After the break – statistical inference!

