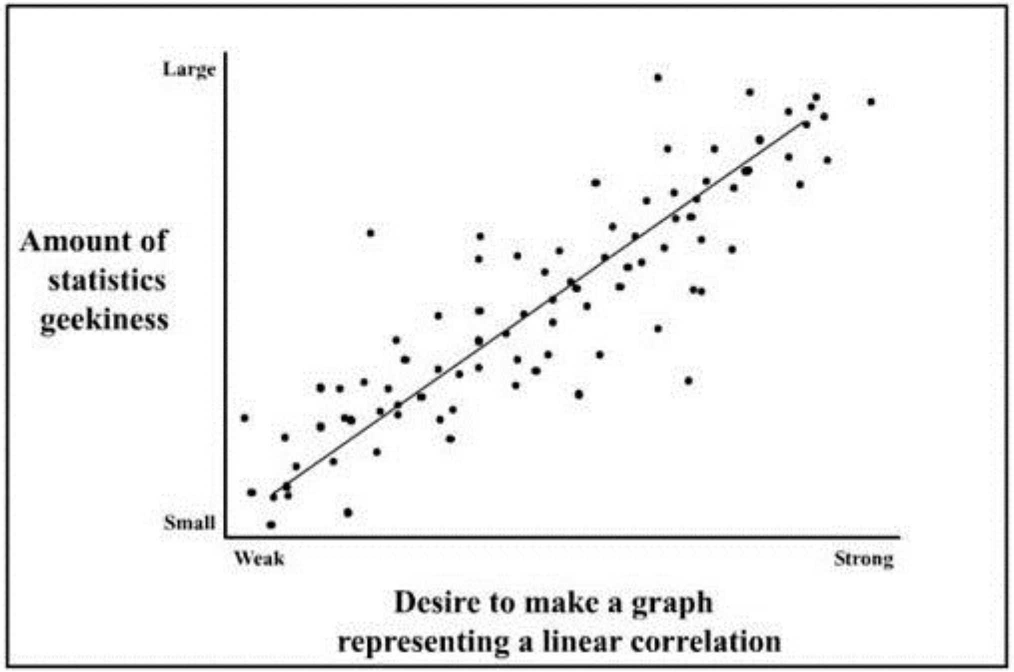
Simple linear regression

Overview

Quick review of correlation

Review of simple linear regression

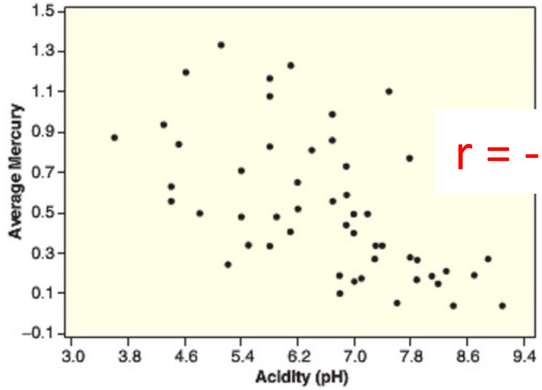
Practice problems



Review: The correlation coefficient

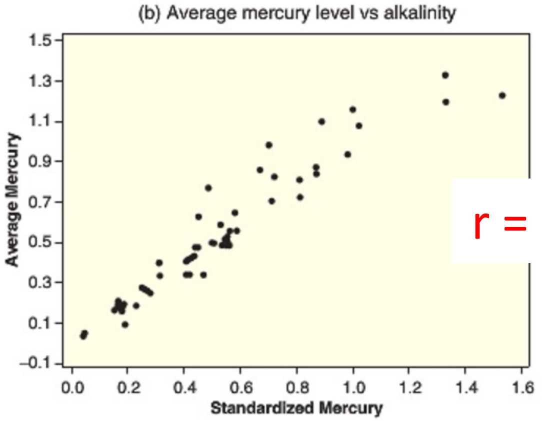
The **correlation** is measure of the strength and direction of a linear association between two variables

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



r = -0.58

Correlation as always between -1 and 1: -1 ≤ r ≤ 1

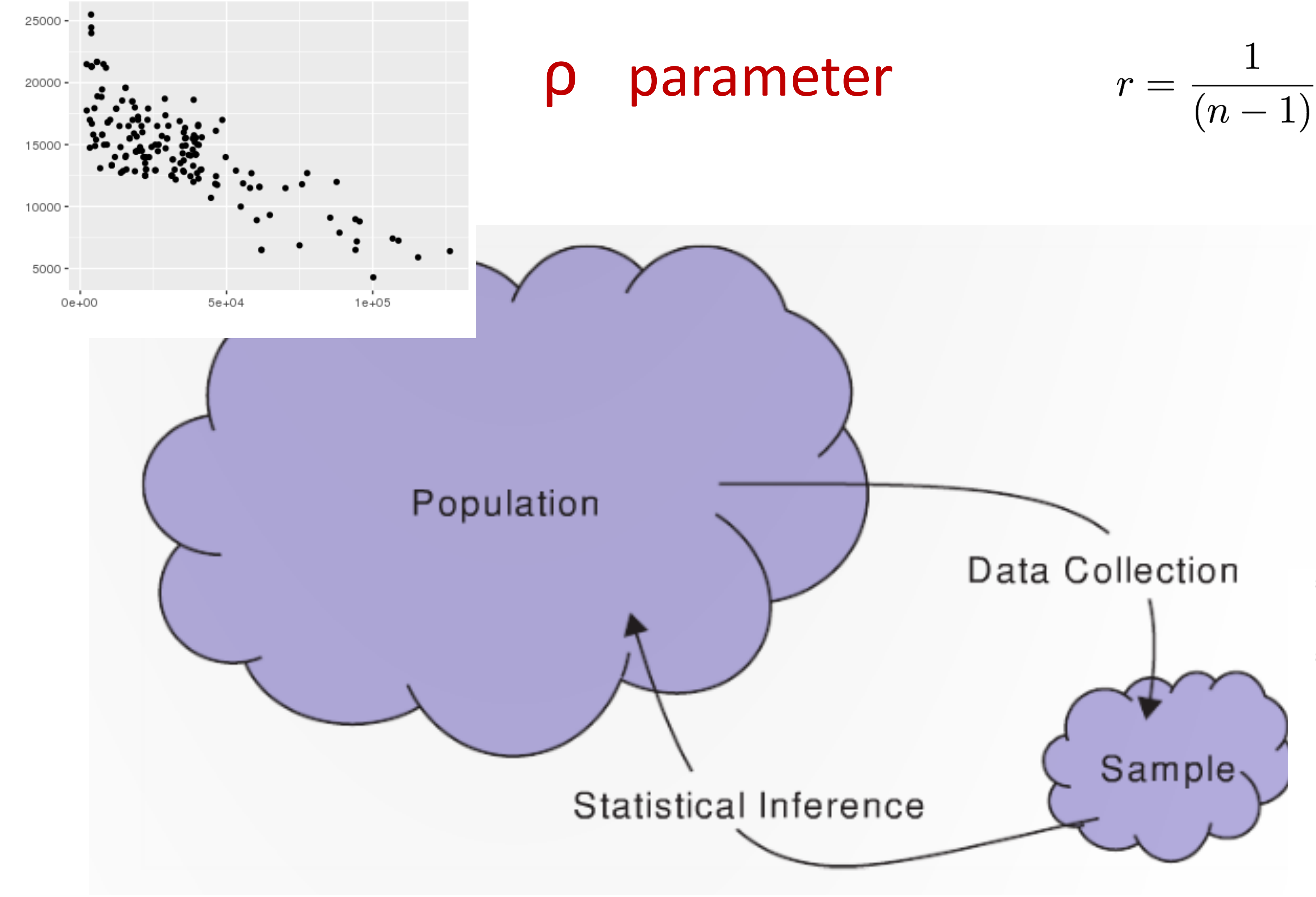


r = 0.96

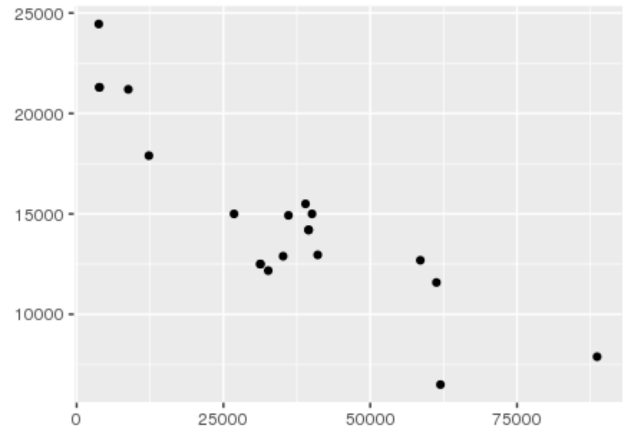
Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship ρ parameter

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

```
R: cor(x, y)
```

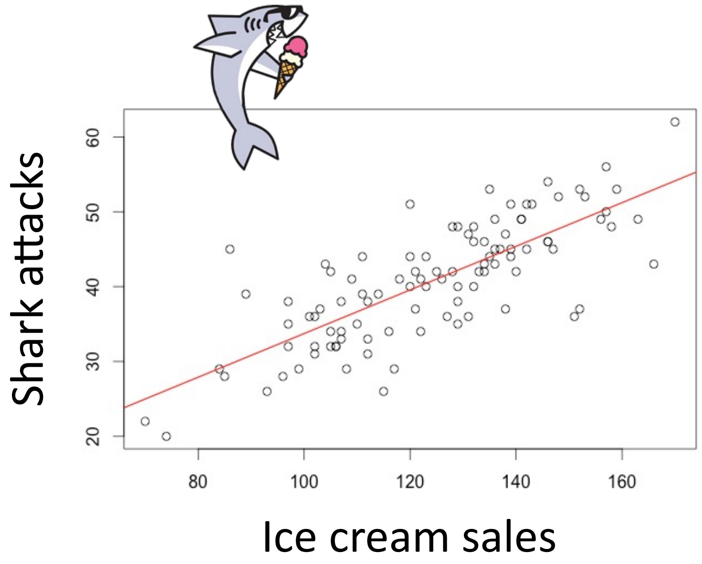


r statistic

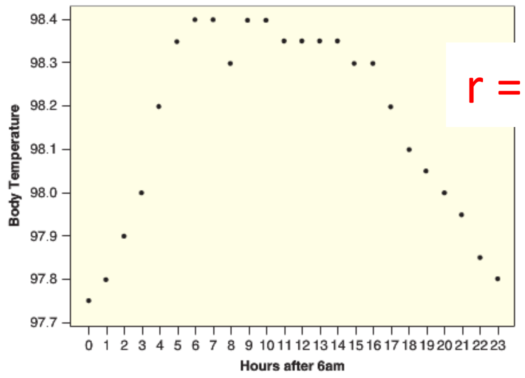


Review: correlation cautions

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

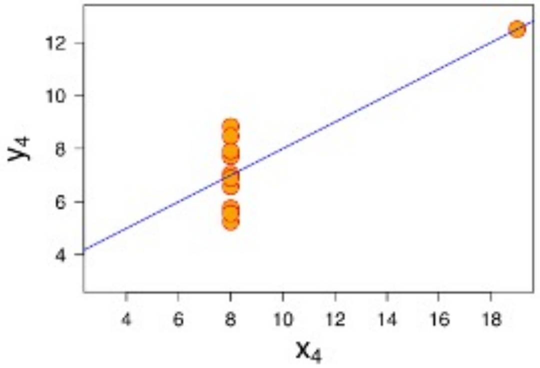


2. A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship



r = -0.08

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



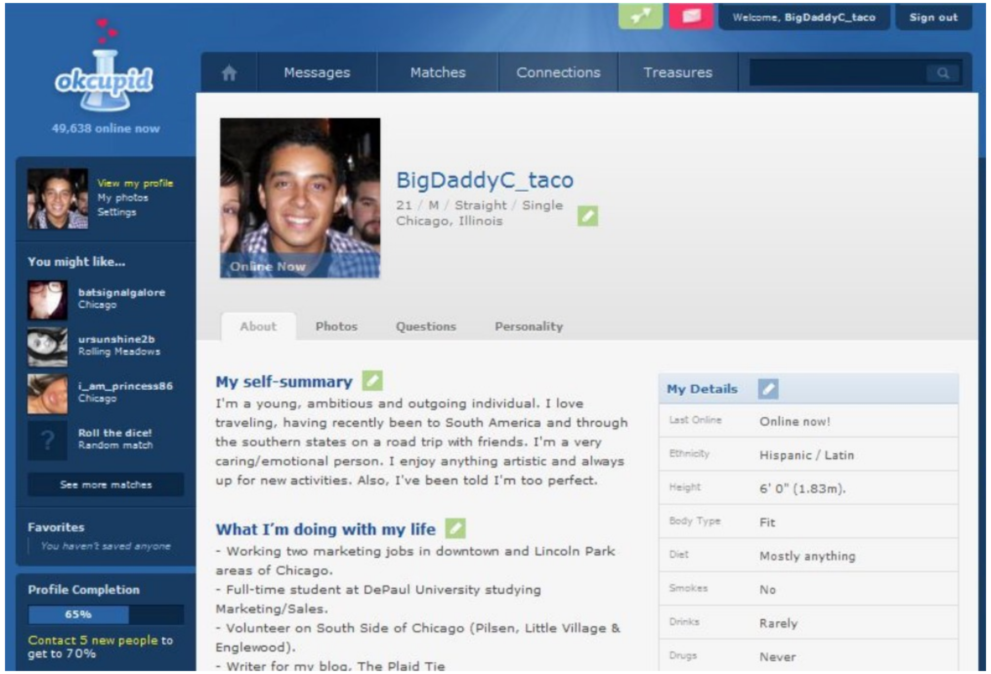
Regression

Regression is method of using one variable ***x*** *to predict* the value of a second variable ***y***

- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

OkCupid text and images



You might like...

batsignalgalore
Chicago

ursunshine2b
Rolling Meadows

i_am_princess86
Chicago

Roll the dice!
Random match

See more matches

Favorites

You haven't saved anyone

Profile Completion

65%

Contact 5 new people to get to 70%

BigDaddyC_taco

21 / M / Straight / Single
Chicago, Illinois

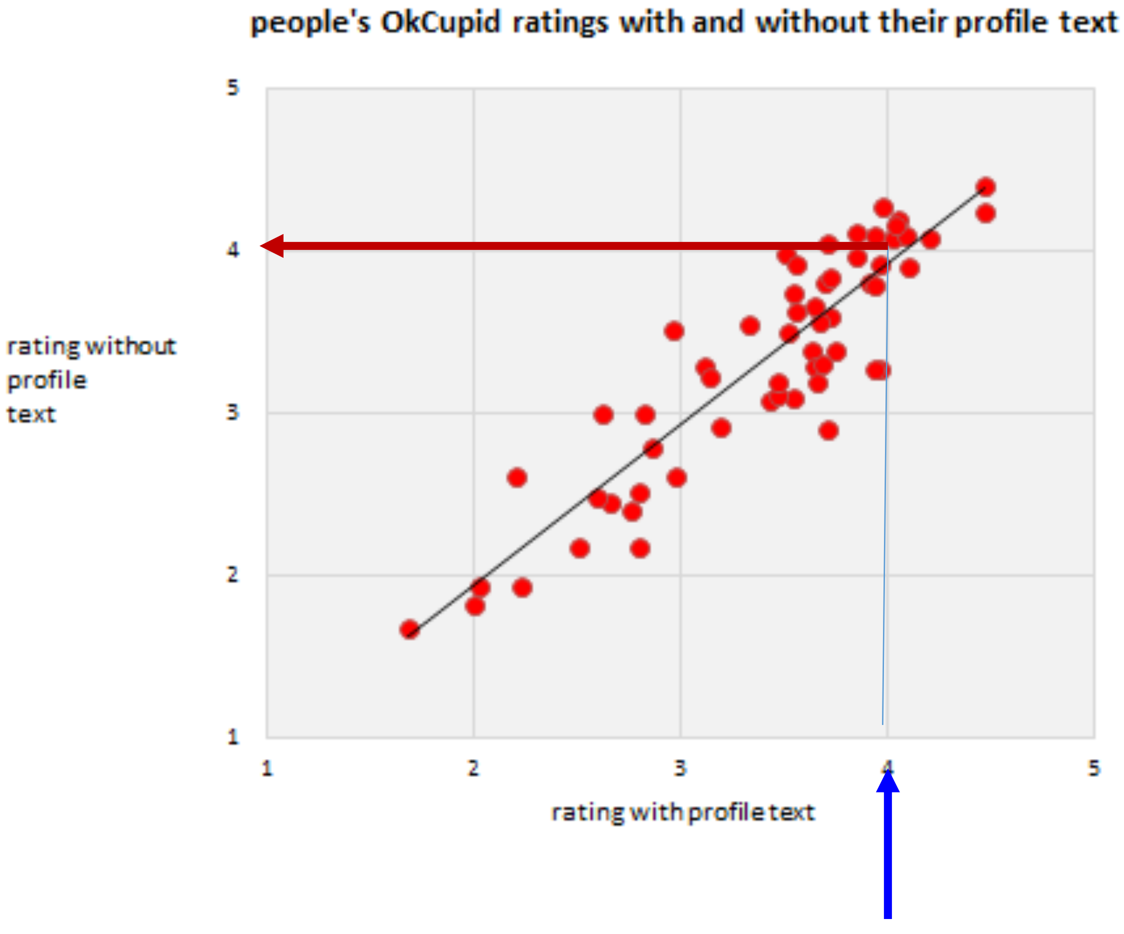
My self-summary

I'm a young, ambitious and outgoing individual. I love traveling, having recently been to South America and through the southern states on a road trip with friends. I'm a very caring/emotional person. I enjoy anything artistic and always up for new activities. Also, I've been told I'm too perfect.

What I'm doing with my life

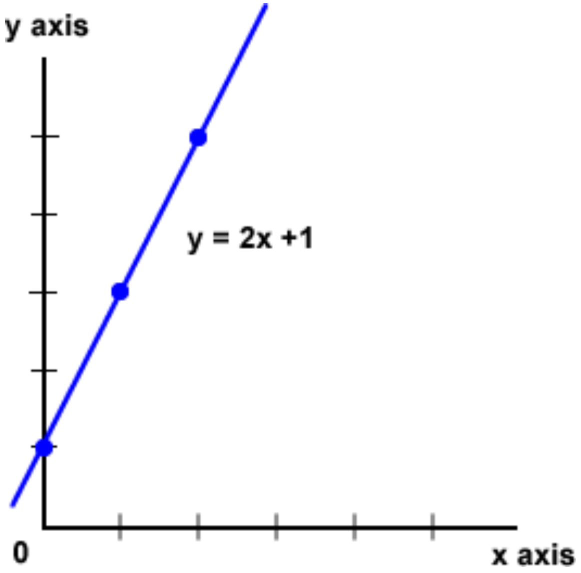
- Working two marketing jobs in downtown and Lincoln Park areas of Chicago.
- Full-time student at DePaul University studying Marketing/Sales.
- Volunteer on South Side of Chicago (Pilsen, Little Village & Englewood).
- Writer for my blog, The Plaid Tie

My Details	
Last Online	Online now!
Ethnicity	Hispanic / Latin
Height	6' 0" (1.83m).
Body Type	Fit
Diet	Mostly anything
Smokes	No
Drinks	Rarely
Drugs	Never



$x_i = 4$

Regression lines



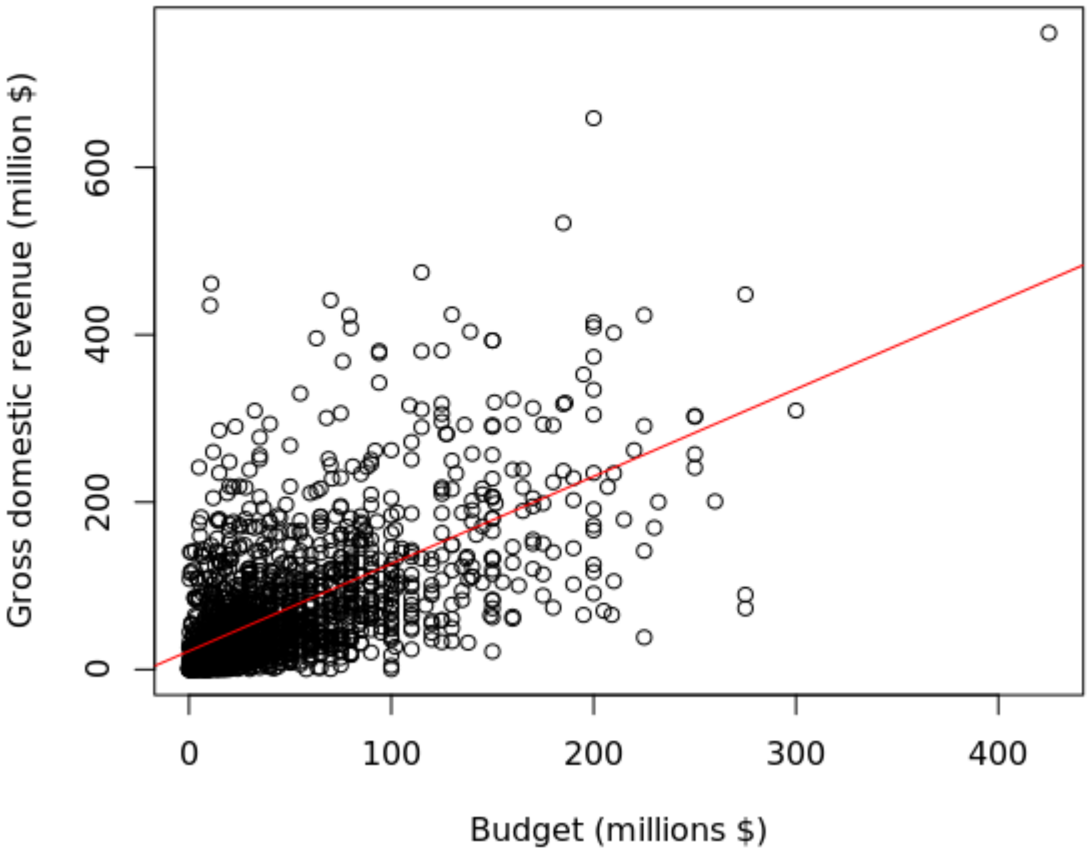
$\hat{y} \;=\; a \;+\; b \cdot x$

$Response \;=\; a \;+\; b \cdot Explanatory$

The slope ***b*** represents the predicted change in the response variable *y* given a one unit change in the explanatory variable *x*

The intercept ***a*** is the predicted value of the response variable y if the explanatory variable x were 0

Bechdel budget revenue regression line



$\hat{y} = a + b \cdot x$

a = 16.636

b = 1.088

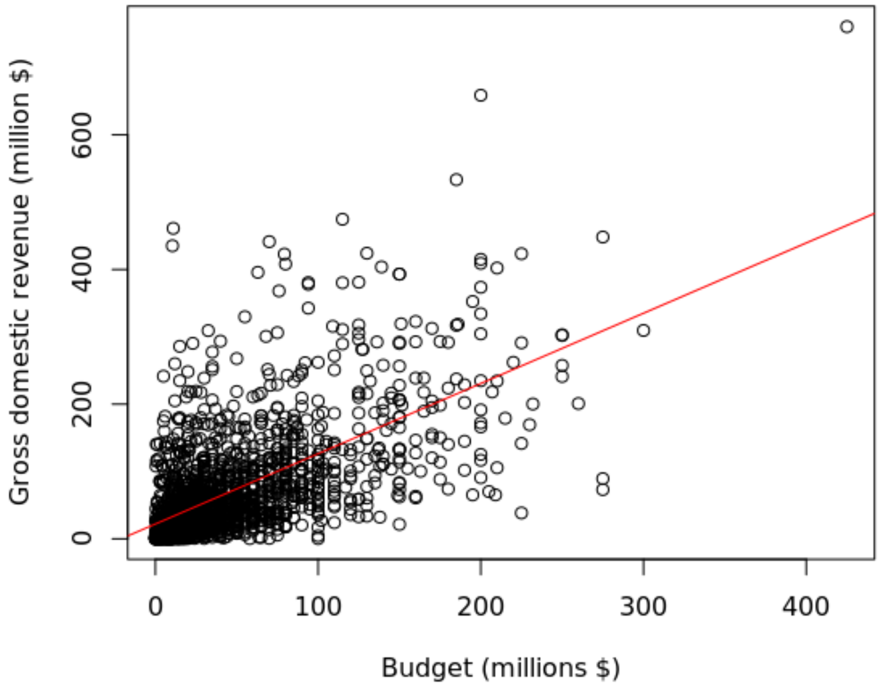
```
R: lm(y ~ x)
```

Using the regression line to make predictions

If a movie had a budget of $0, how much what would their gross domestic revenue be?

$a = 16.636, \quad b = 1.088$

$\hat{y} = 16.636 + 1.088 \cdot x$

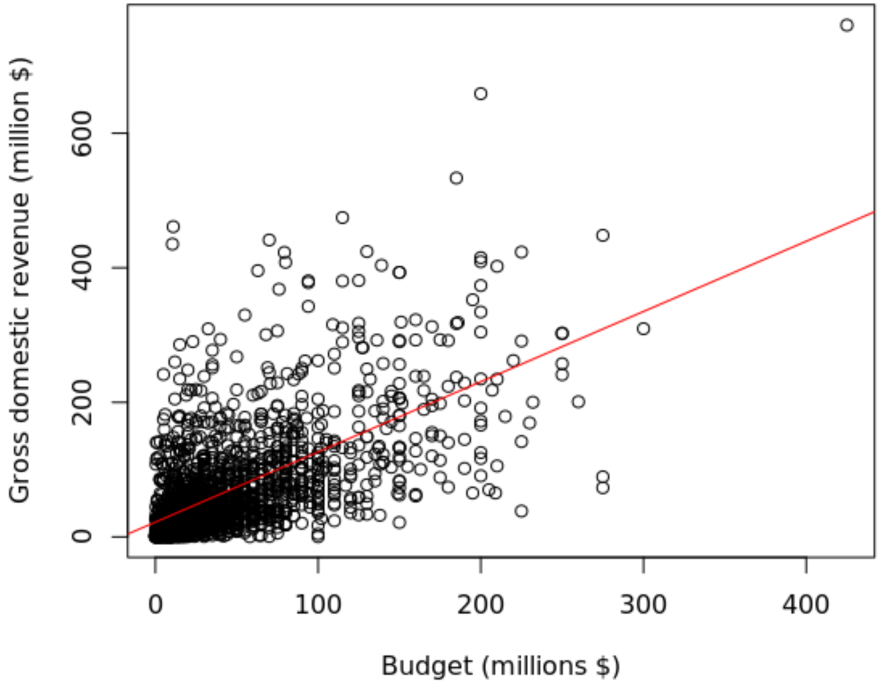


Using the regression line to make predictions

For every extra $1 spent, how much more would we predict their gross domestic revenue to be?

$a = 16.636, \quad b = 1.088$

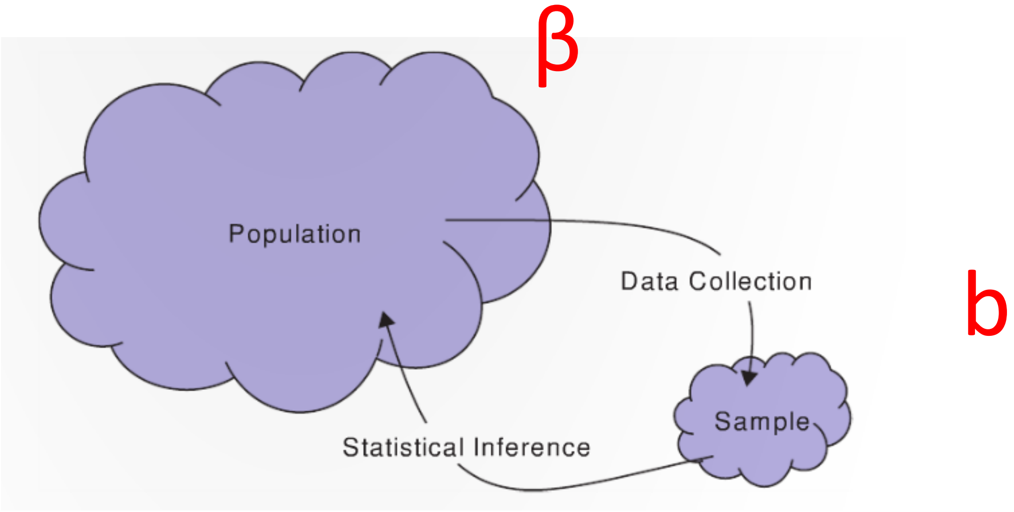
$\hat{y} = 16.636 + 1.088 \cdot x$



Notation

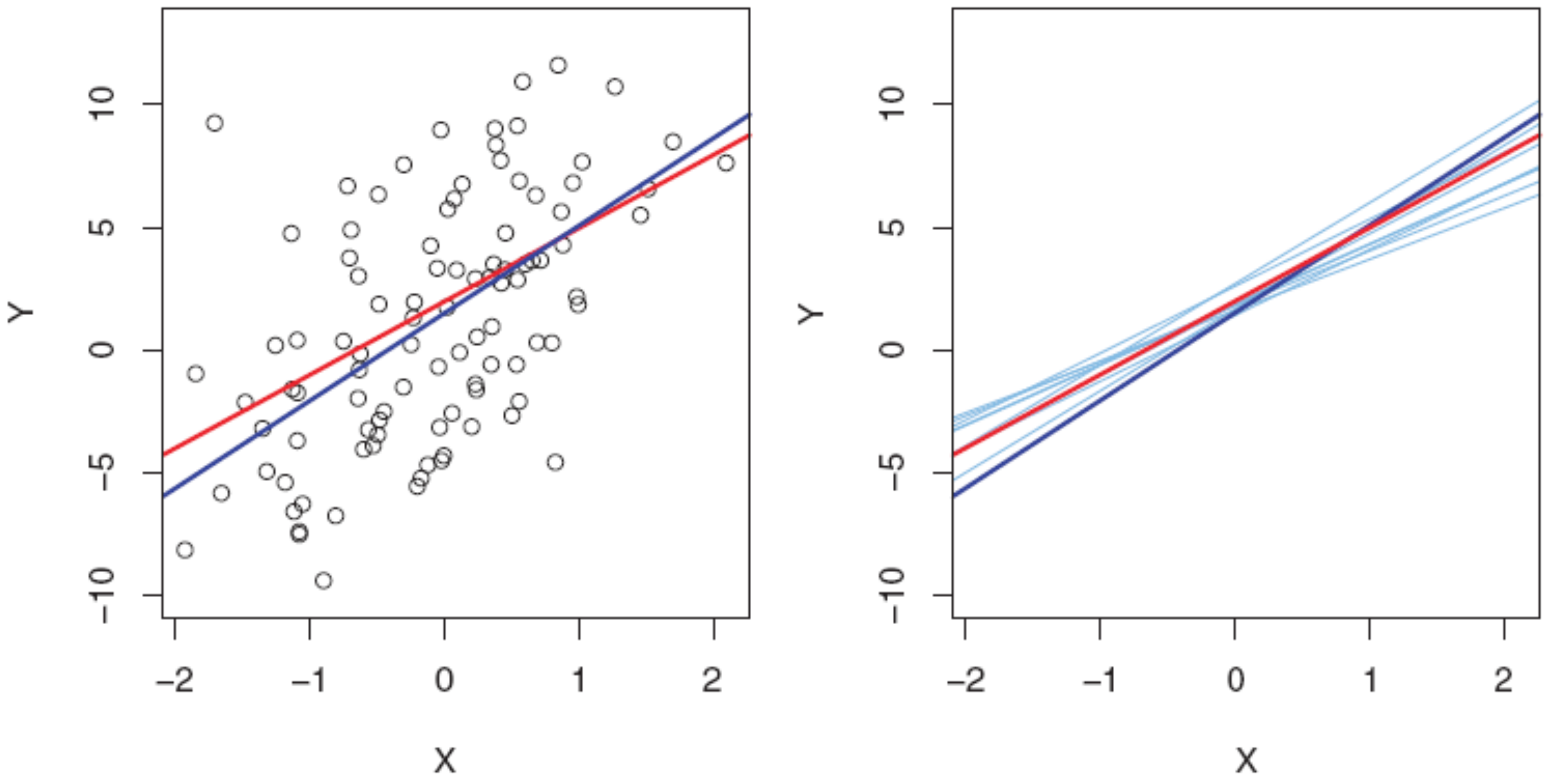
The letter **b** is typically used to denote the slope of the sample

The Greek letter **β** is used to denote the slope of the population



Population: β

Sample estimates: b

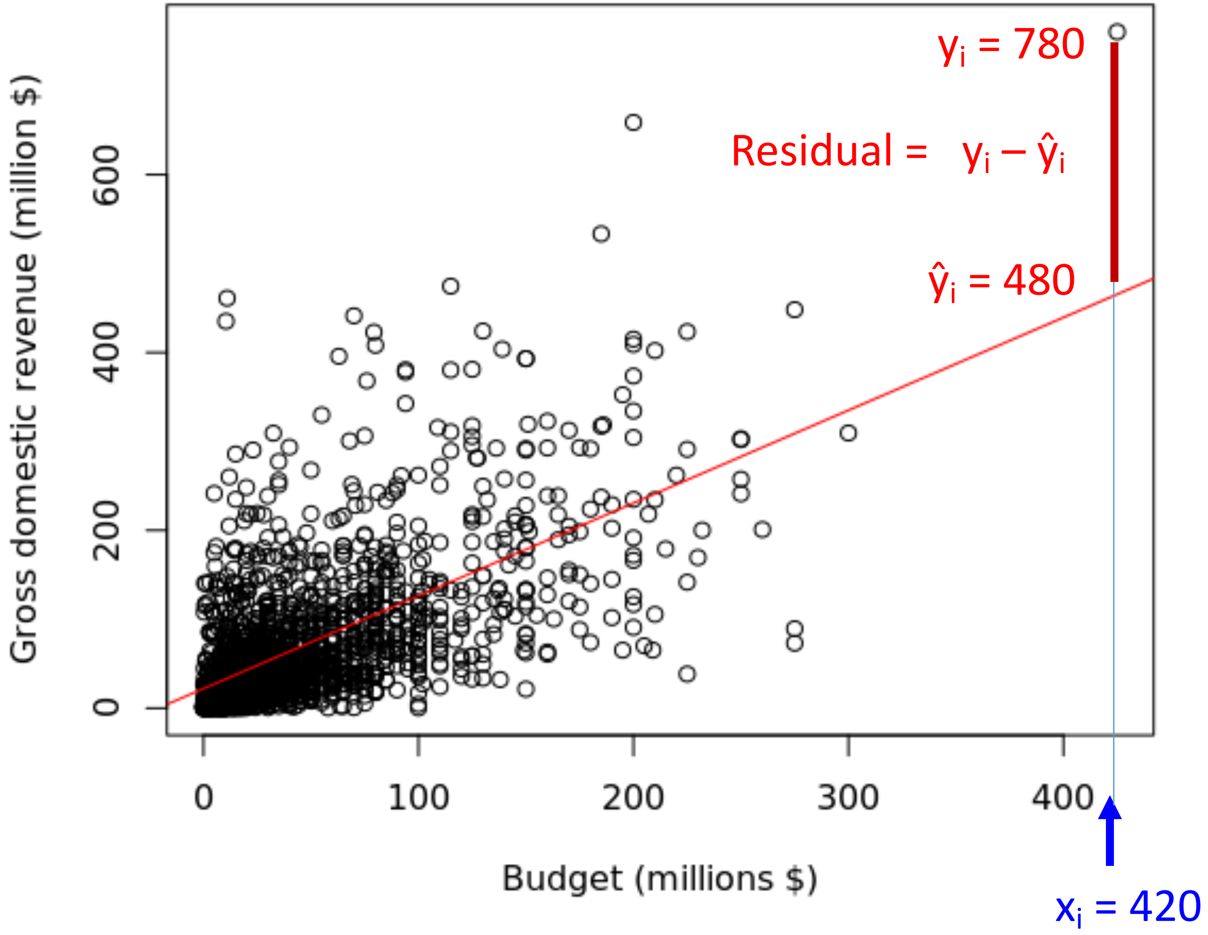


Residuals

The **residual** is the difference between an observed (y_i) and a predicted value ($\hat{y}_i$) of the response variable

$$\textit{Residual}_i \;=\; \textit{Observed}_i - \textit{Predicted}_i \;=\; y_i - \hat{y}_i$$

Budget revenue regression line



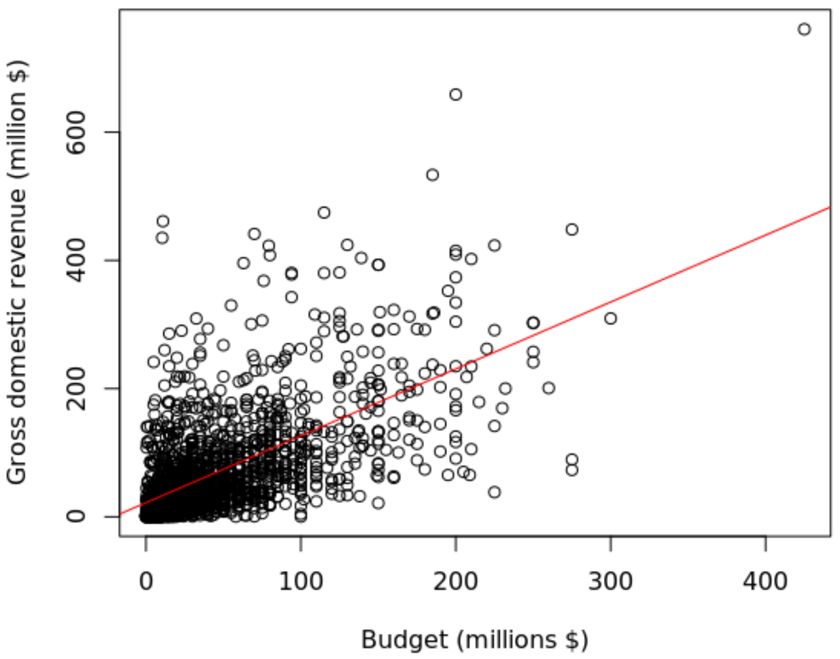
Domestic gross revenue residuals

$$\hat{y} = 16.636 + 1.088 \cdot budget$$

Budget (x)	domgross obs (y)	domgross pred (ŷ)
13	25.7	30.8
45	13.4	65.6
20	53.1	38.4
61	75.6	83.0
40	95.0	60.2
225	38.4	261.5
92	67.3	116.7
12	15.3	29.7

Line of ‘best fit’

The **least squares line**, also called **‘the line of best fit’**, is the line which minimizes the sum of squared residuals



Try to find the line of best fit

Domestic gross revenue residuals

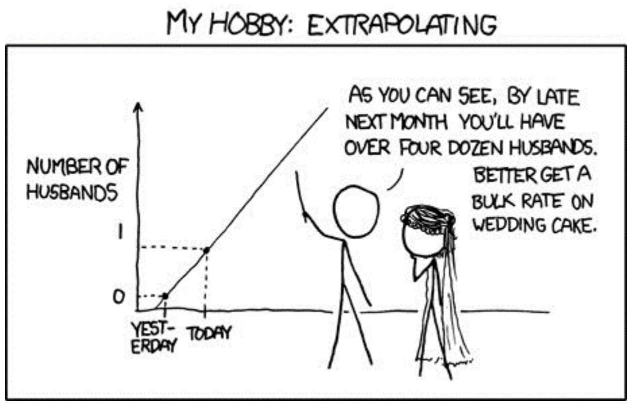
domgross obs **(y)**	**domgross pred** **($\hat{y}$)**	**Residuals** **($y - \hat{y}$)**	**Residuals^2** **$(y - \hat{y})^2$**
25.7	30.8	-5.1	26.0
13.4	65.6	-52.2	2723.2
53.1	38.4	14.7	216.4
75.6	83.0	-7.4	54.7
95.0	60.2	34.9	1215.3
38.4	261.5	-223.1	49769.2
67.3	116.7	-49.4	2439.3
15.3	29.7	-14.4	206.5

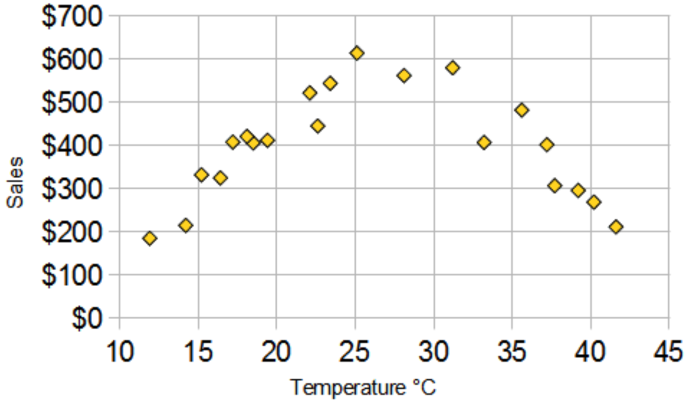
Regression cautions

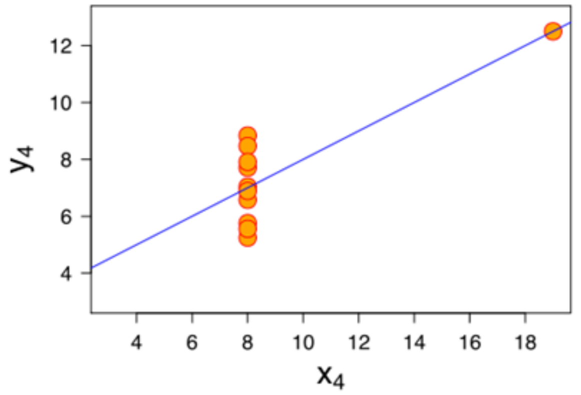
1. Avoid trying to apply the regression line to predict values far from those that were used to create the line.

2. Plot the data! Regression lines are only appropriate when there is a linear trend in the data.

3. Be aware of outliers – they can have an huge effect on the regression line.


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.





Linear regression in R

Regression lines in R – extracting the data

```
#  get the markdown document for today’s class
> SDS100::download_class_code(6)

# load the library with the data
> library(fivethirtyeight)

# remove missing values
> bechdel <- na.omit(bechdel)

# extract variables of interest
> budget <- bechdel$budget/10^6
> bechdel$domgross/10^6
```

Regression lines in R

```
# create a scatter plot
> plot(budget, domgross)

# fit a regression model
> lm_fit <- lm(domgross ~ budget)

# examine the a and b coefficients
> coef(lm_fit)

# add the regression line to the plot
> abline(lm_fit, col = "red")
```

Concepts for the relationship between two quantitative variables

A **scatterplot** graphs the relationship between two variables

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

- Value between -1 and 1

In **linear regression** we fit a <u>line</u> to the data, called the **regression line**

- We get coefficients for the slope (b) and the y-intercept (a)

The **residual** is the difference between <u>an observed</u> (y_i) and a <u>predicted value</u> ($\hat{y}_i$) of the response variable

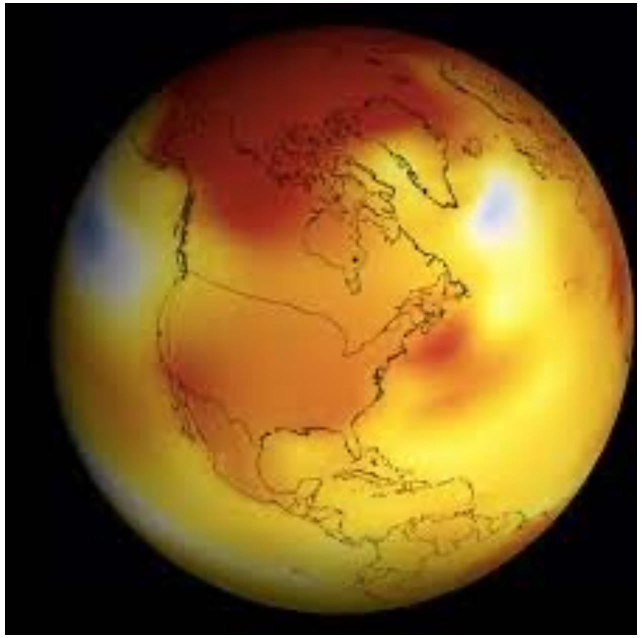
- The regression line minimizes the sum of squared residuals

Practice problems

Regression practice problem 1

Levels of carbon dioxide (CO_2) in the atmosphere are rising rapidly, far above any levels ever before recorded.

Levels were around 278 parts per million in 1800, before the Industrial Age, and had never, in the hundreds of thousands of years before that, gone above 300 ppm.



Levels are now over 400 ppm.

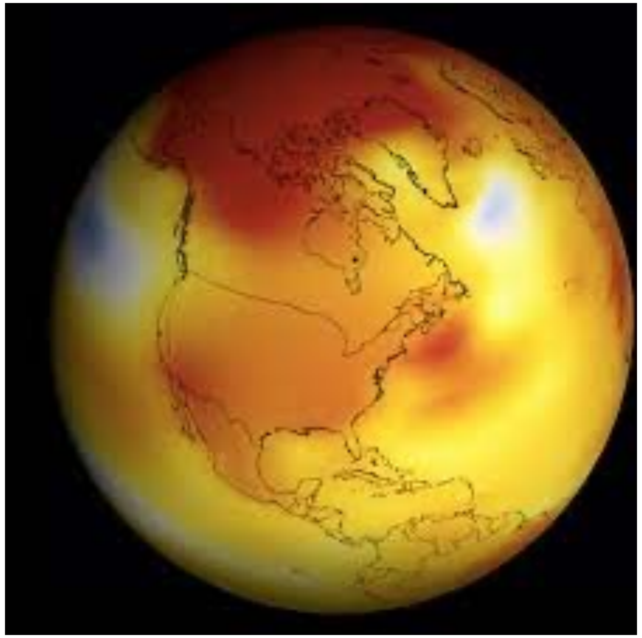
We can use this information to predict CO_2 levels in different years.

Regression practice problem 1

```
# Download the data
download_data("CarbonDioxide.csv")

# Load the data
carbon <- read.csv("CarbonDioxide.csv")

# Extract vectors of interest
year <- carbon$Year
co2 <- carbon$C02
```

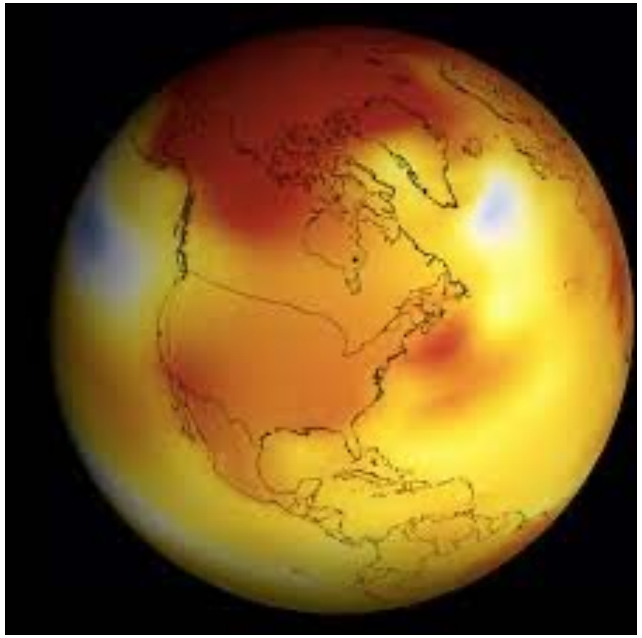


Regression practice problem 1

Please do the following:

1. Create a scatter plot of the data
2. Calculate the correlation coefficient
3. Fit a linear regression model
   - Write down the linear regression equation
4. Present what CO_2 levels will be in:
   - 2003, 2025, 2050, 2100

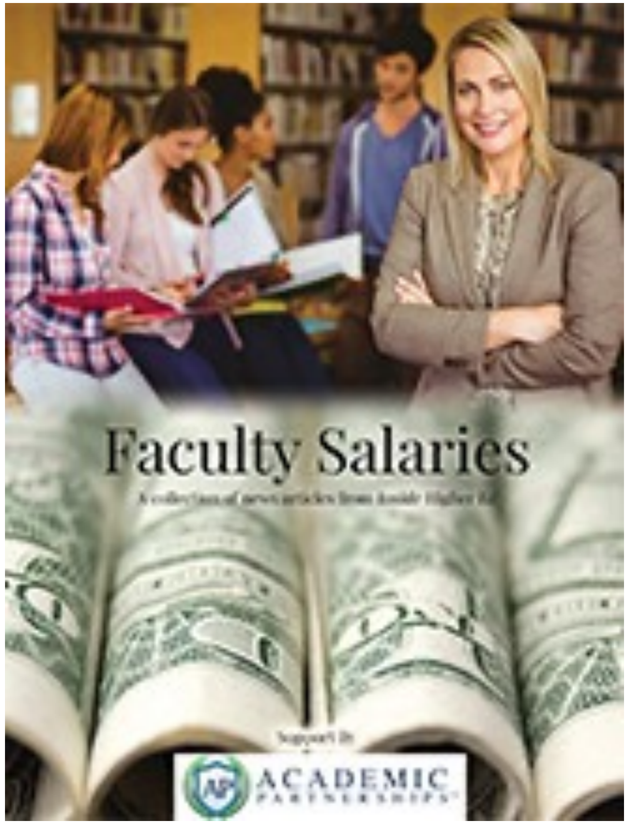
5. Report which predictions are reasonable



Should we work in groups?

Regression practice problem 2

Does paying faculty more lead to higher college graduation rates?



The CollegeScores4yr contains two variables of interest to help us answer this question:

1. **CompRate** records the percentage of students at each four-year school who graduate within six years (known as the completion or graduation rate).

2. **FacSalary** gives the average monthly salary for faculty (in dollars) at each school.

```
download_data("CollegeScores4yr.csv")
salary_data <- read.csv("CollegeScores4yr.csv")
```

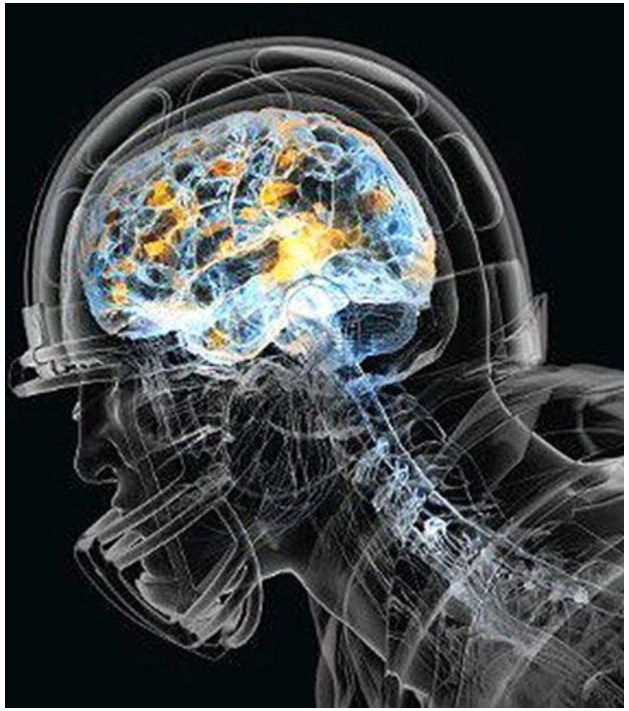
Explore other relationships in the data as well!

Does playing football affect brain size?

A study Singh et al (2014) published in the Journal of the American Medical Association (JAMA) examined the relationship between football and concussions on the brain.

The study included three groups with n = 25 participants in each group

- Healthy controls who had never played football.
- Football players with no history of concussions.
- Football players with a history of concussions.



Let's examine the following through visualizations and/or statistics:

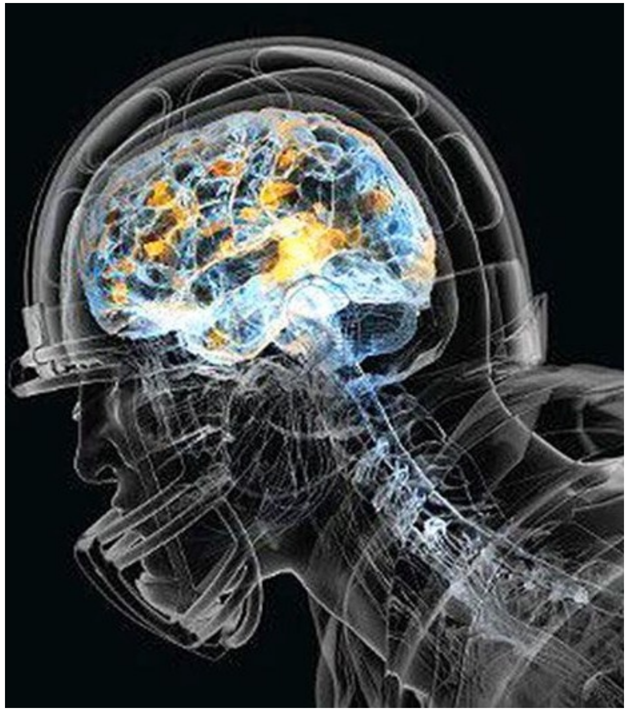
1. The relationship between number of years playing football and hippocampus volume
2. The relationship between hippocampus size and the three groups

Does playing football affect brain size?

```
# install.packages("Lock5Data")

library(Lock5Data)

data(FootballBrain)
```



Let's examine the following through visualizations and/or statistics:

1. The relationship between number of years playing football and hippocampus volume
2. The relationship between hippocampus size and the three groups

Life expectancies in different countries

Data about countries in the world can accesses in the Lock5Data package

- install.packages("Lock5Data")
- library(Lock5Data)
- View(AllCountries)

Create a histogram of life expectancies for all countries and…

- Describe the shape of the histogram
- From looking at the histogram, estimate the mean and median
  - Which will be larger?
- Check your answers using the mean() and median() functions

Student exercise

Let's look at the Lock5Data StudentSurvey data

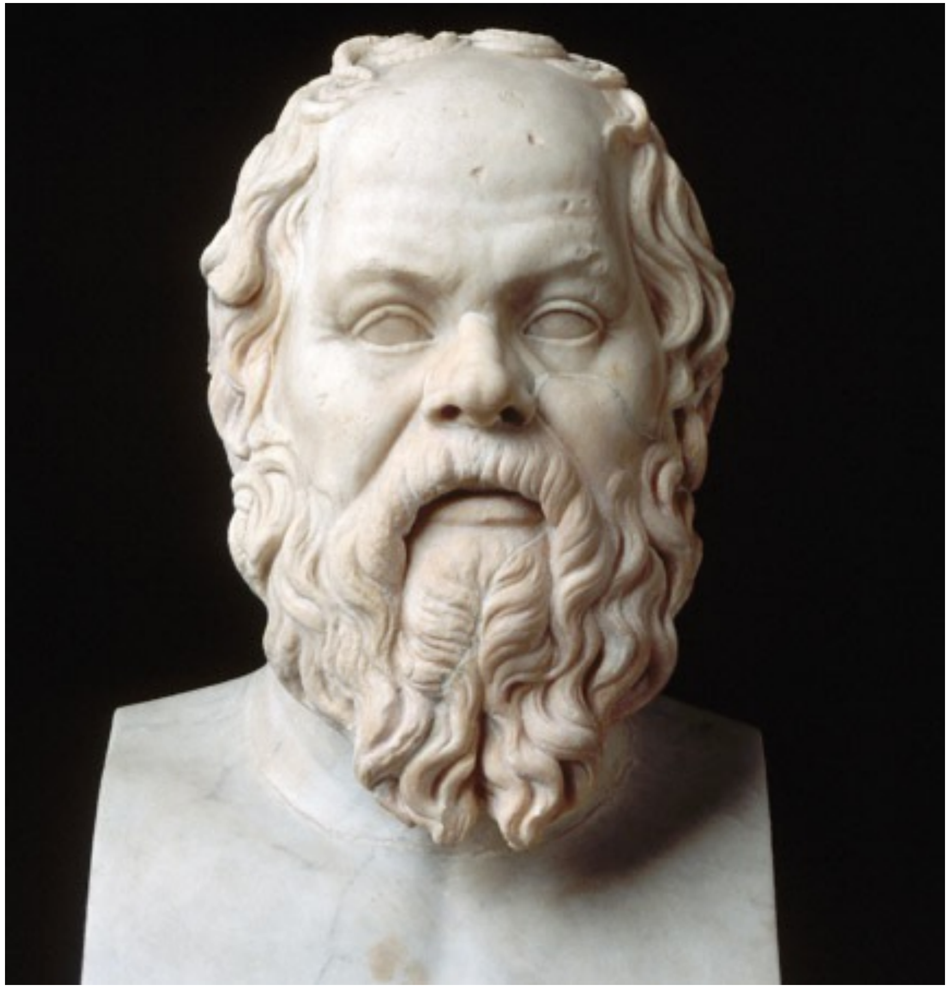
- View(StudentSurvey)
- male_data <- subset(StudentSurvey, Sex == "M")$Exercise
- female_data <- subset(StudentSurvey, Sex == "F")$Exercise

From this data calculate:

- $\overline{x}_f$, the mean number of hours spent exercises by the females
- $\overline{x}_m$, the mean number of hours spent exercises by the males
- Compute the difference $\overline{x}_m - \overline{x}_f$, and interpret it in context.

Review of descriptive statistics

Who is this?

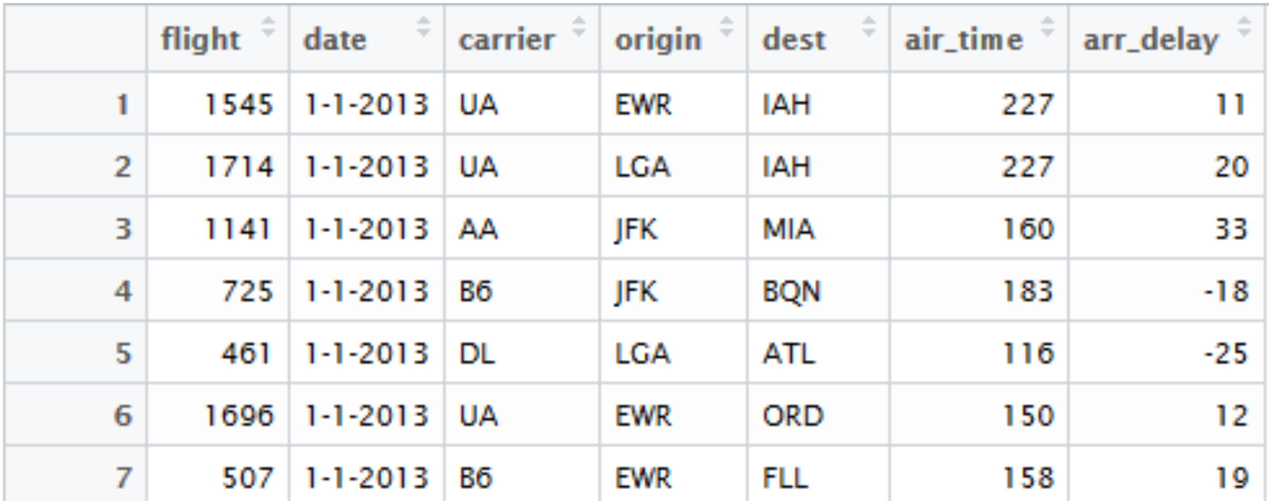


Intro to data

What is Statistics?

What are…

- Observational units?
- Variables?
- Categorical variables?
- Quantitative variables?



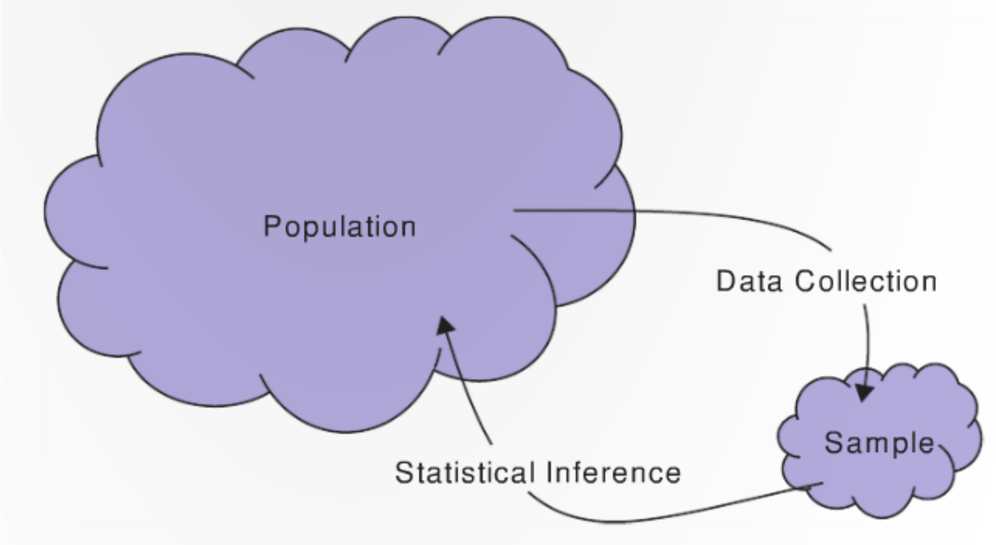
	flight	date	carrier	origin	dest	air_time	arr_delay
1	1545	1-1-2013	UA	EWR	IAH	227	11
2	1714	1-1-2013	UA	LGA	IAH	227	20
3	1141	1-1-2013	AA	JFK	MIA	160	33
4	725	1-1-2013	B6	JFK	BQN	183	-18
5	461	1-1-2013	DL	LGA	ATL	116	-25
6	1696	1-1-2013	UA	EWR	ORD	150	12
7	507	1-1-2013	B6	EWR	FLL	158	19

Sampling

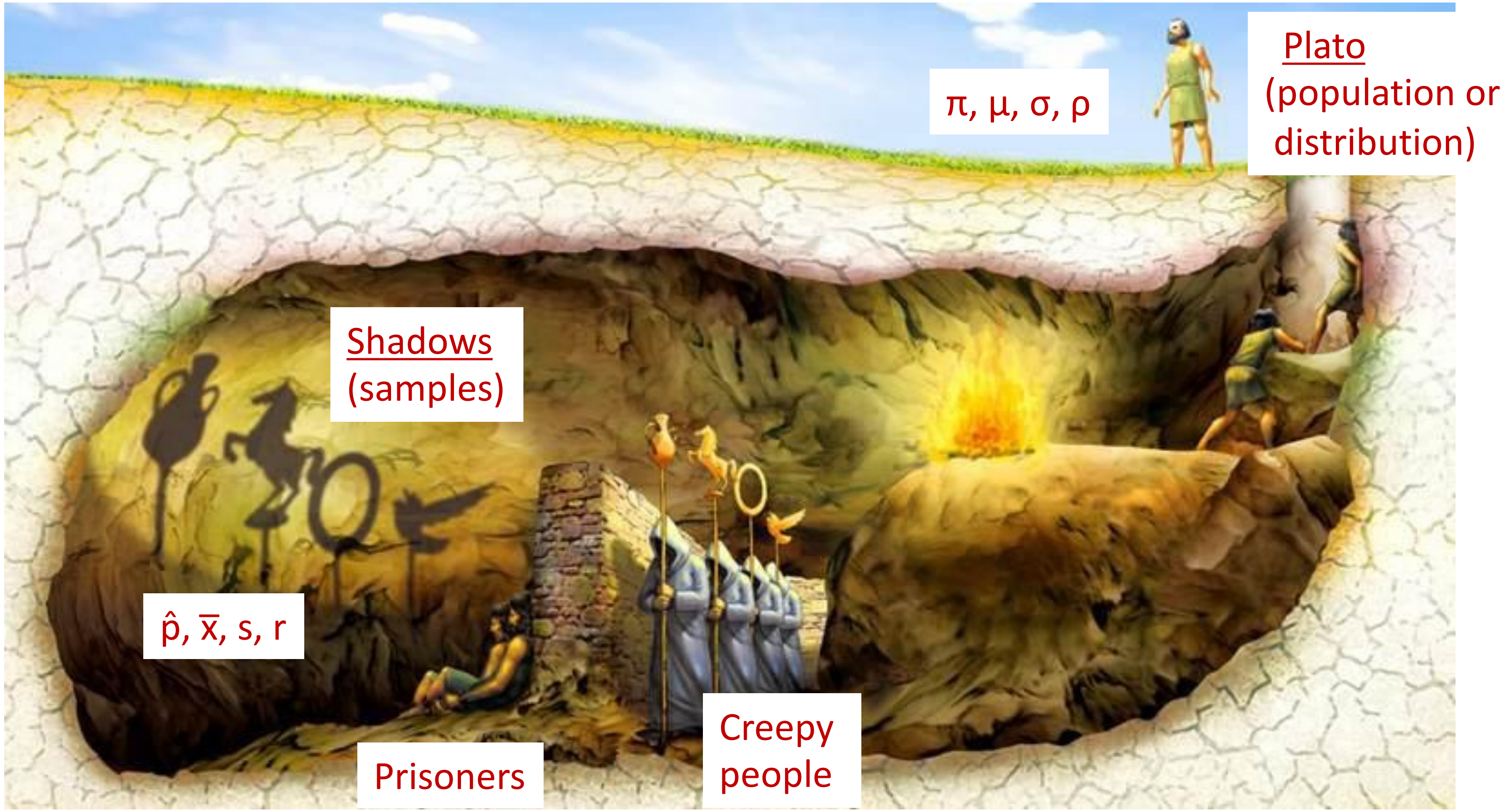
What is a …?

- sample
- population
- statistic
- parameter

What is statistical inference?



Plato’s cave



From The Republic (~ 380 BCE)

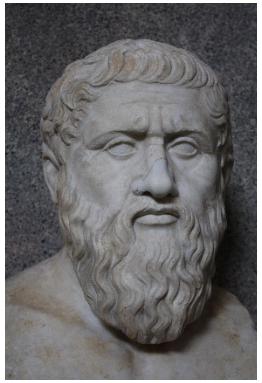
Quiz: parameters and statistics

	Sample Statistic	**Population Parameter**
Mean	$\bar{x}$	μ
Standard deviation		
Proportion		
Correlation		
Regression slope		

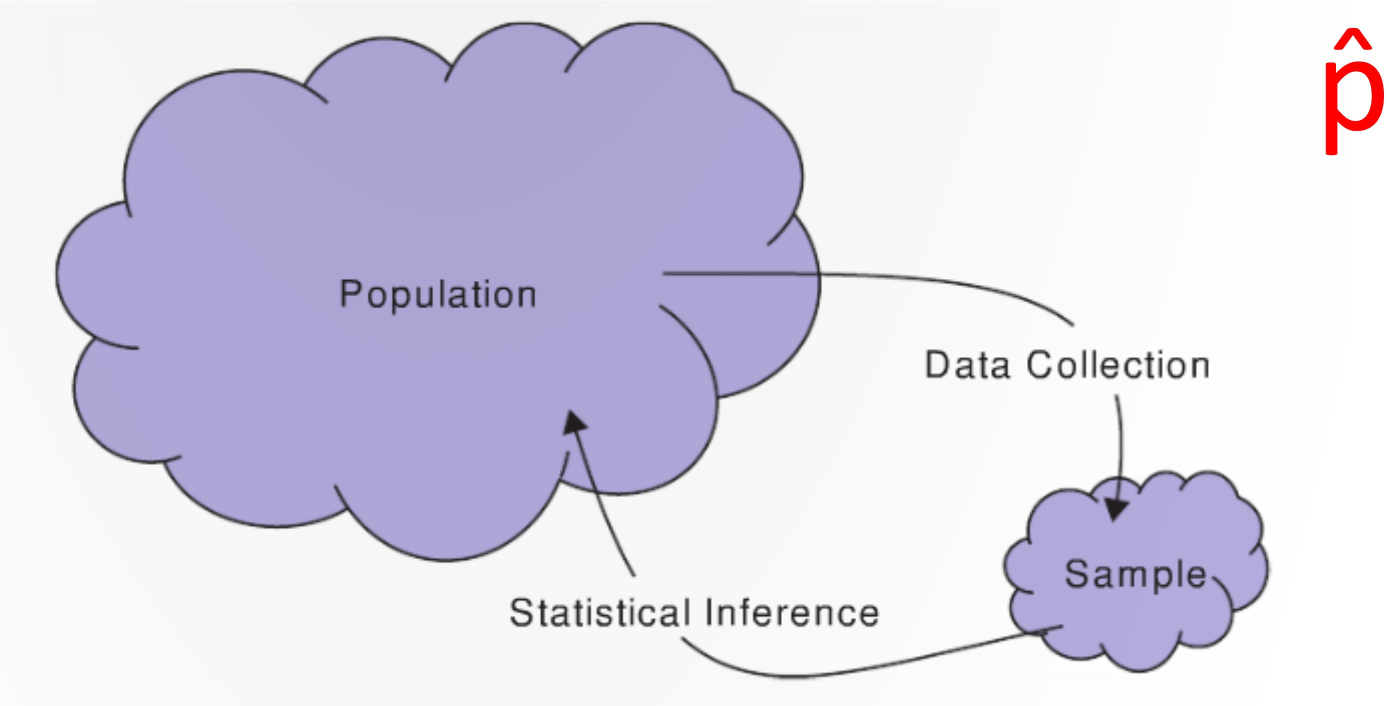
Quiz: parameters and statistics

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Proportion	$\hat{p}$	π
Correlation	r	ρ
regression slope	b	β

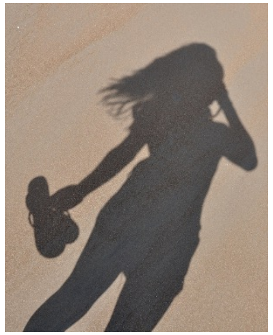
Population parameters vs. sample statistics



$\pi, \mu, \sigma, \rho, \beta$



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

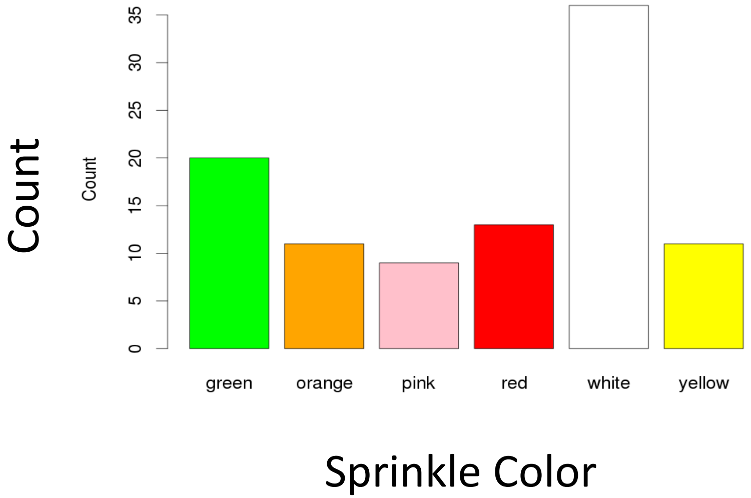


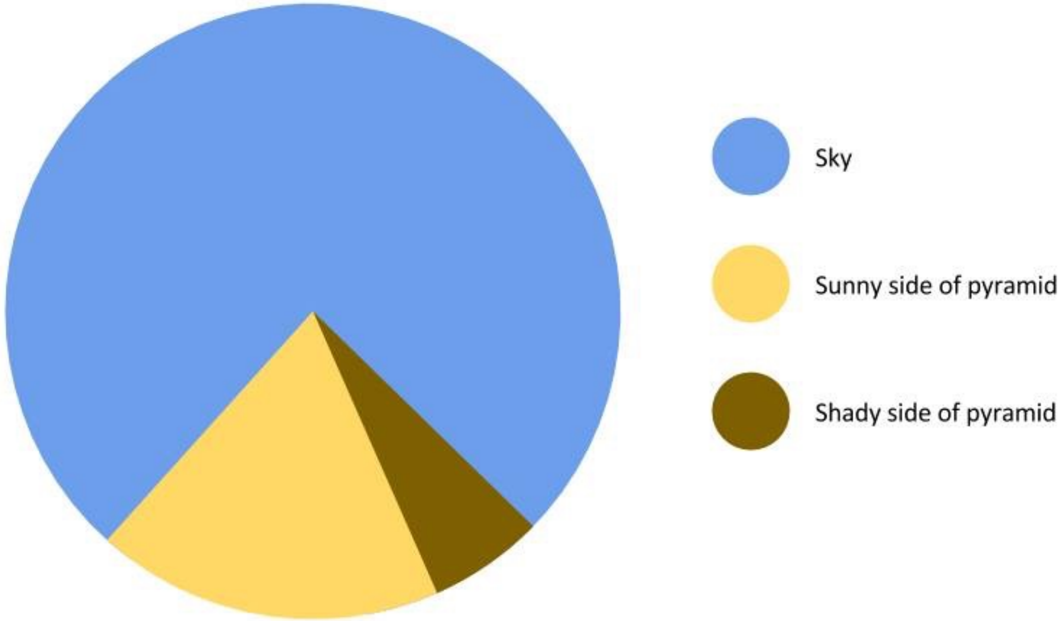
Categorical data

What is the main statistic we discussed for categorical data?

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

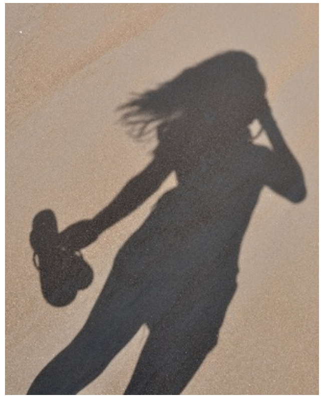
How can we plot categorical data?



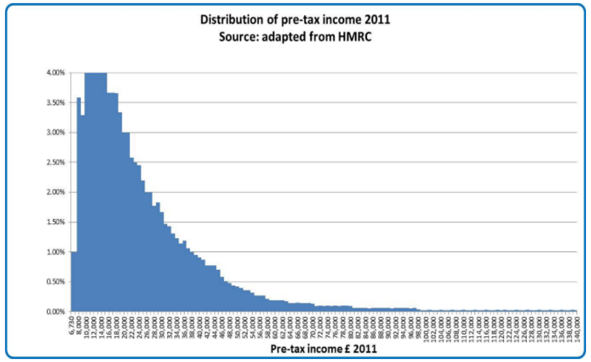


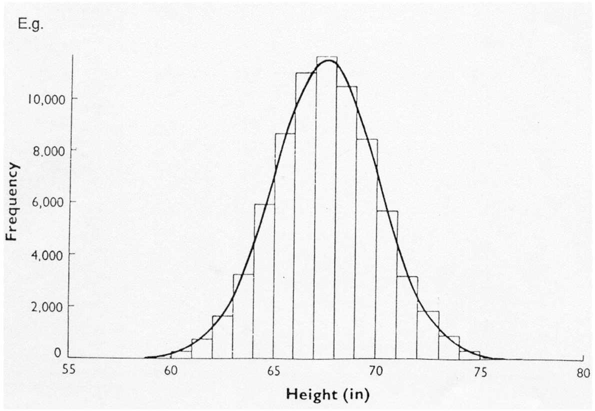
Quantitative data?

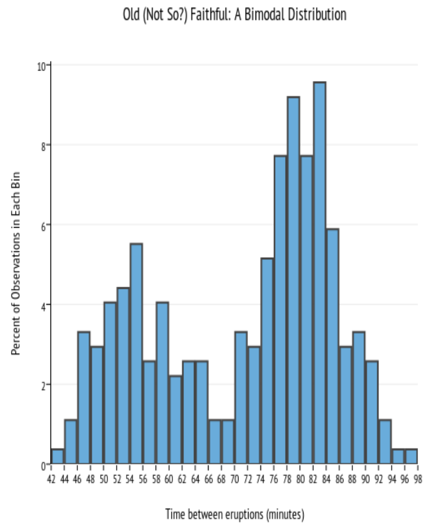
What is a good way to visualize the shape of quantitative data?

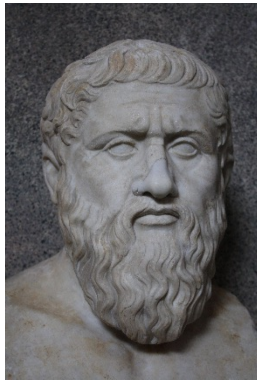


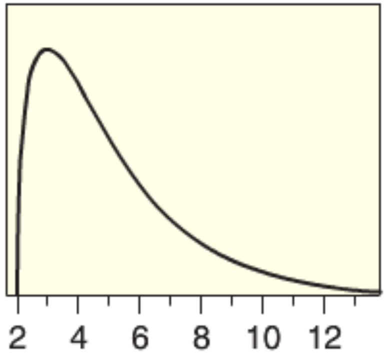
Income distribution

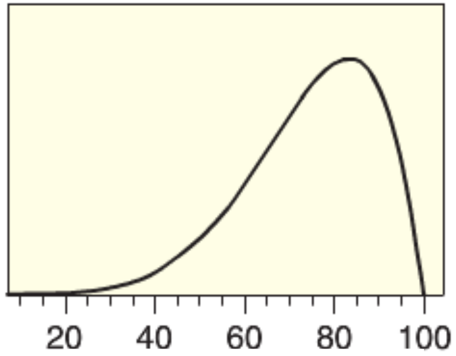


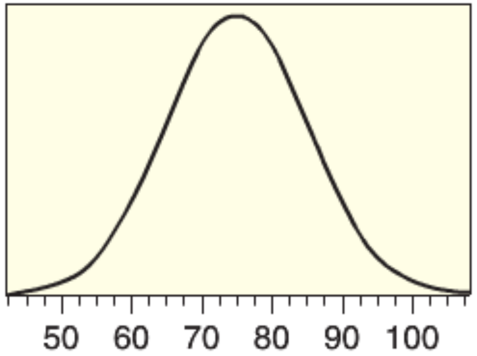


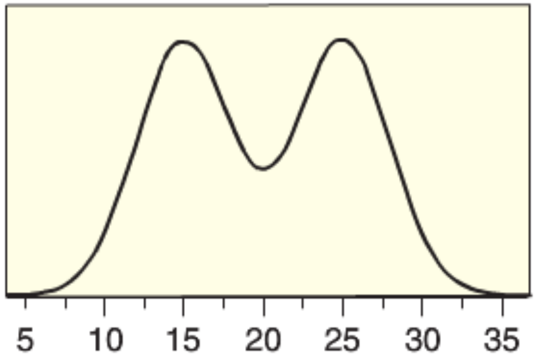




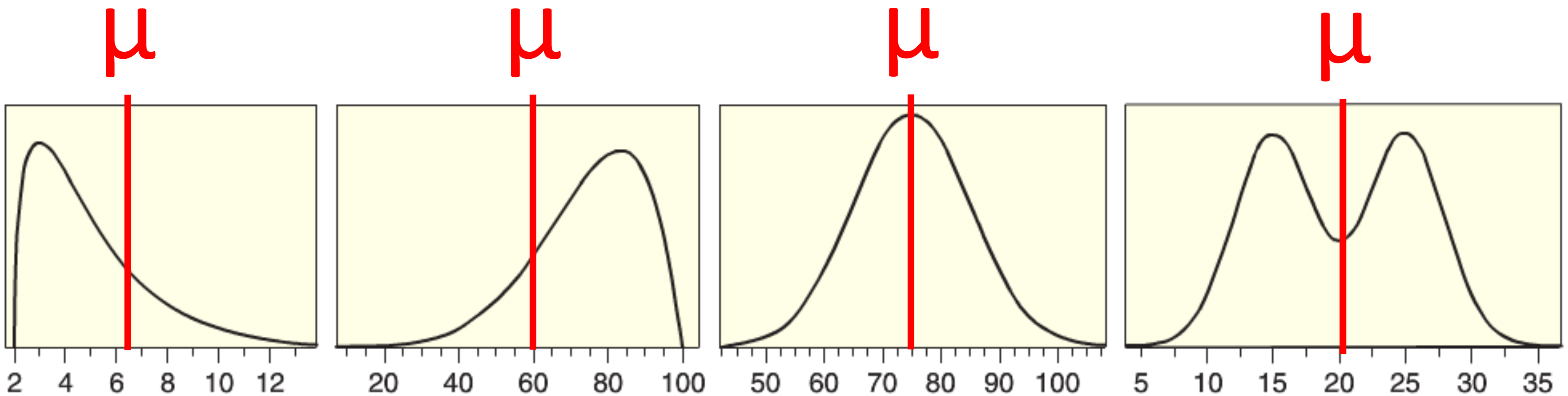






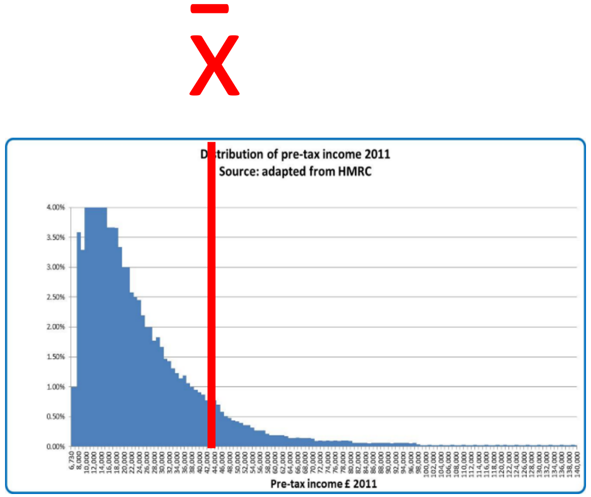


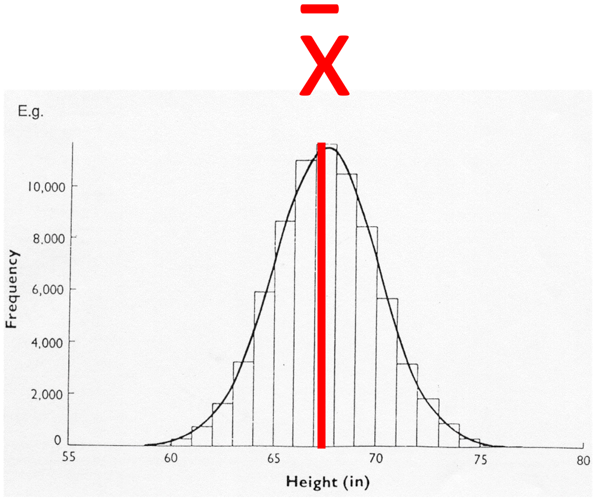
Measure of central tendency: the mean

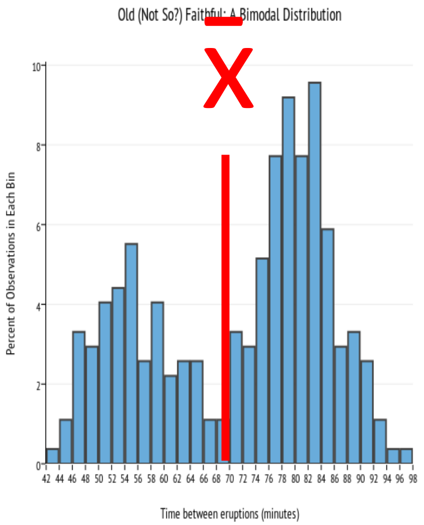


(a) Skewed to the right (b) Skewed to the left (c) Symmetric and bell-shaped (d) Symmetric but not bell-shaped

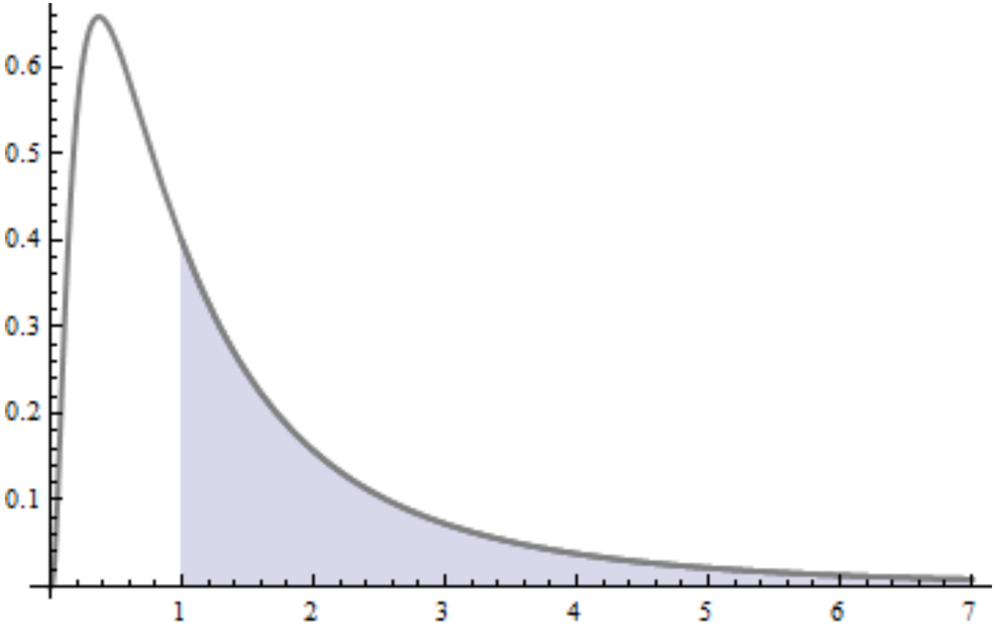
$$\frac{\Sigma_i^n x_i}{n}$$

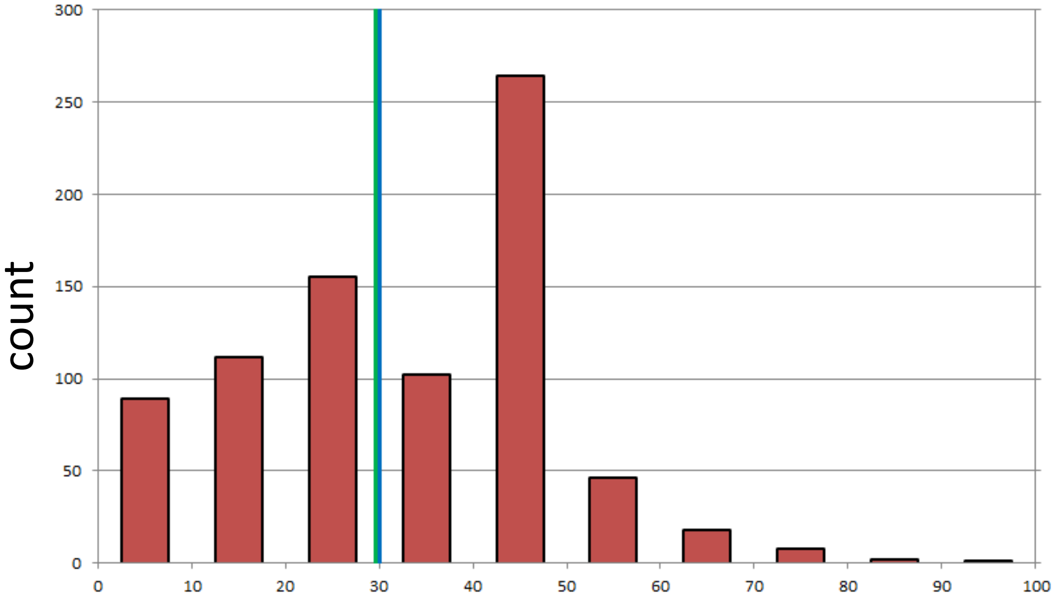






Measure of central tendency: the median



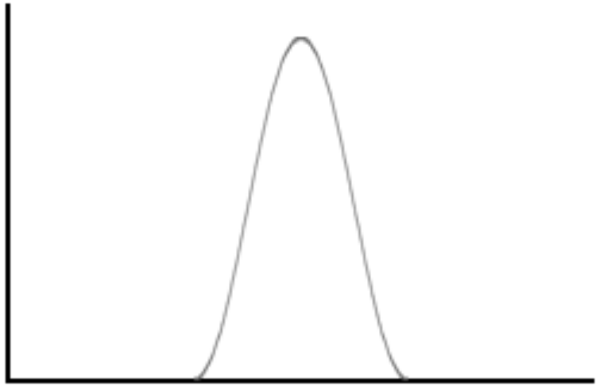


Which is resistant to outliers, the mean or the median?

The standard deviation

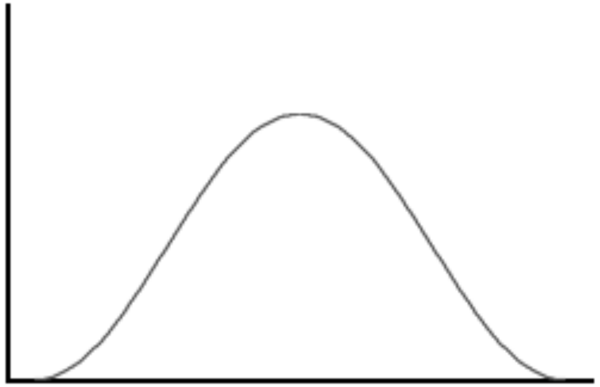
Which distribution has a larger standard deviation?

A



parameter σ

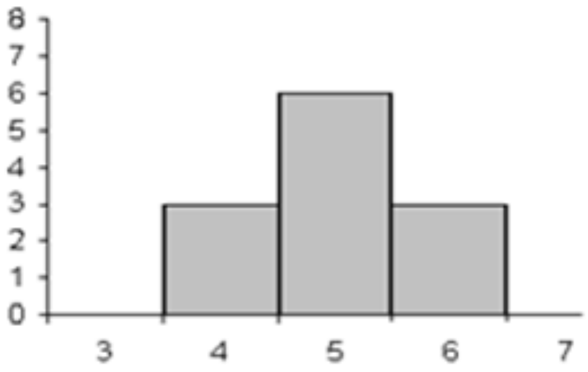
B

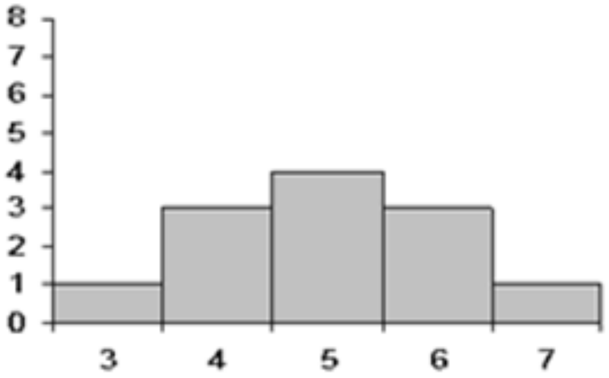


The standard deviation

Which distribution has a larger standard deviation?

statistic: s





What is the formula for the standard deviation?

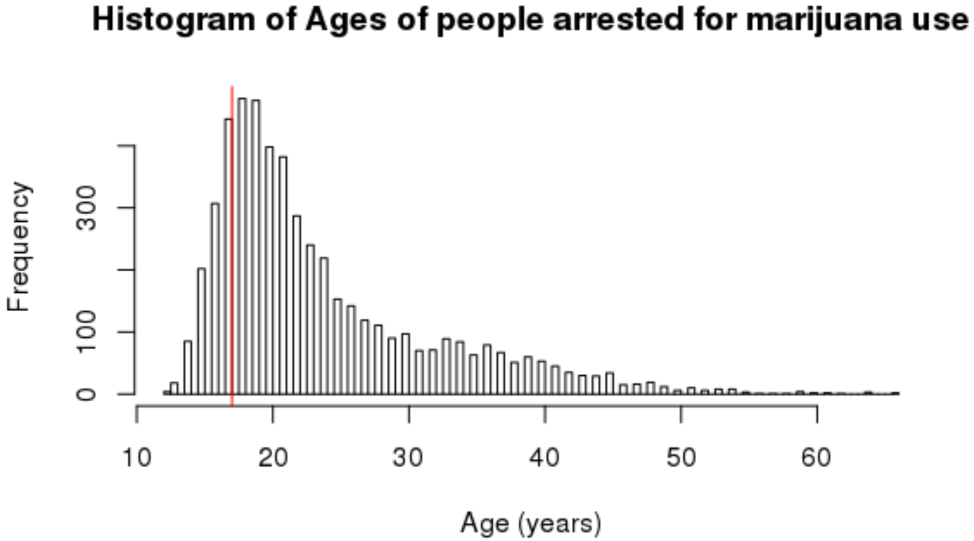
$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

z-scores and percentiles

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

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

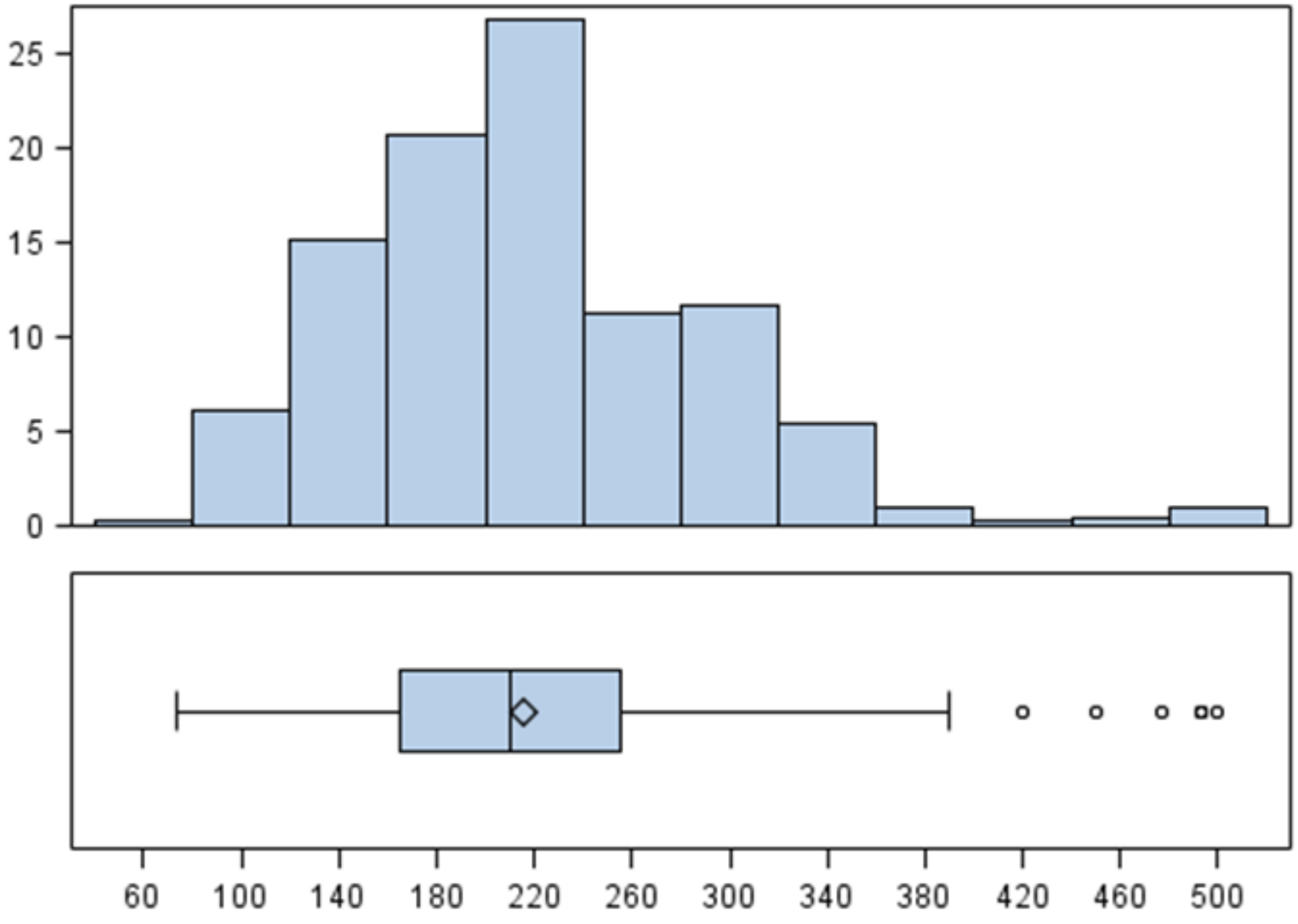
What is the p^{th} percentile?



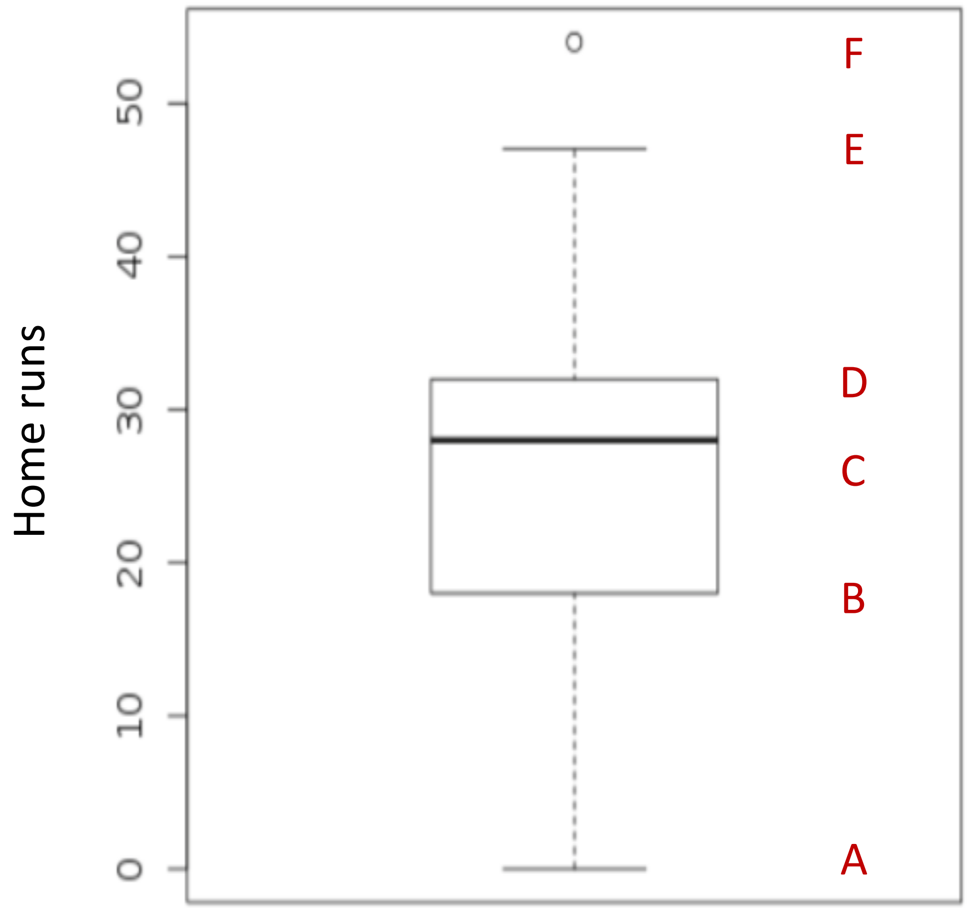
Normal pillow

What percent of the pillow’s mass is ± 1 standard deviations from the mean?

What is a five-number summary and a box plot?



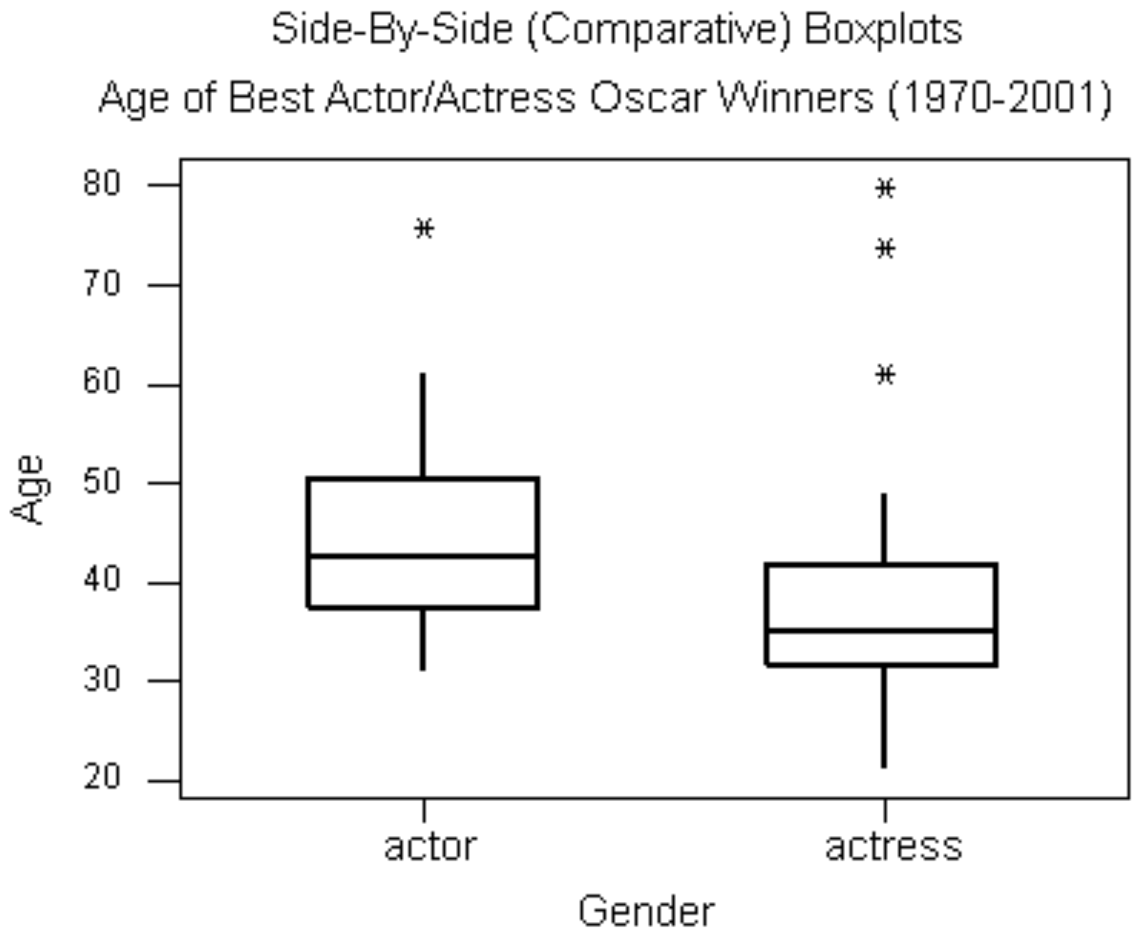
Box plot quiz



What is:

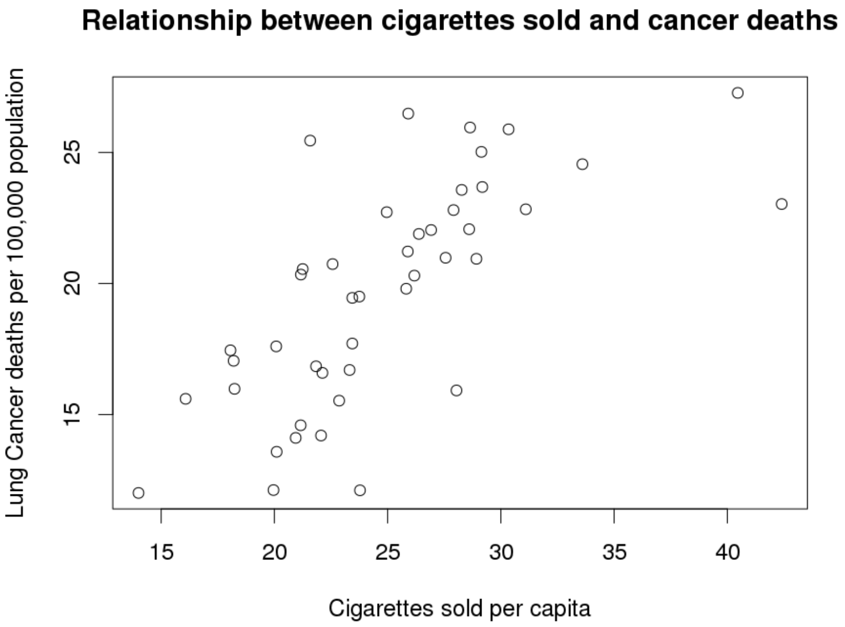
- Q1?
- Q3?
- The median?
- Most extreme values that are not outliers
- Outliers

Side-by-side boxplots



Relationships between measures

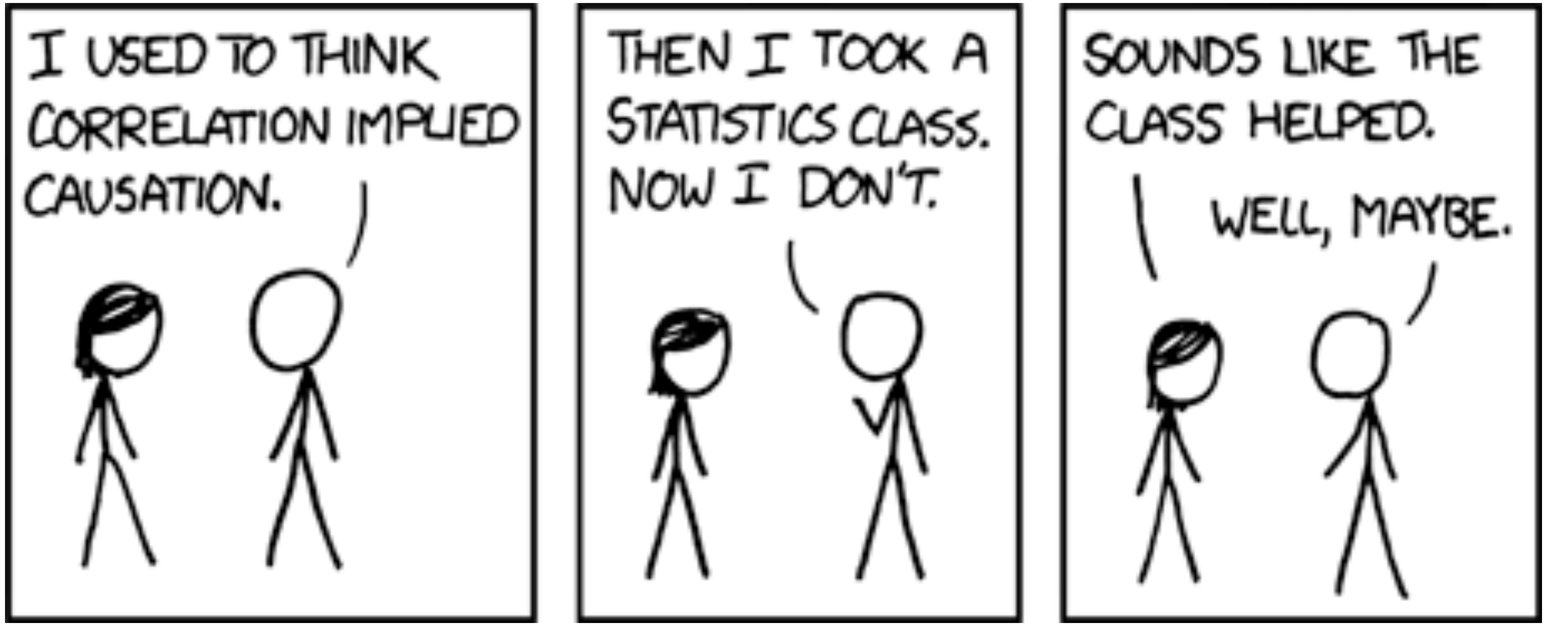
Q: What is this type of plot called?



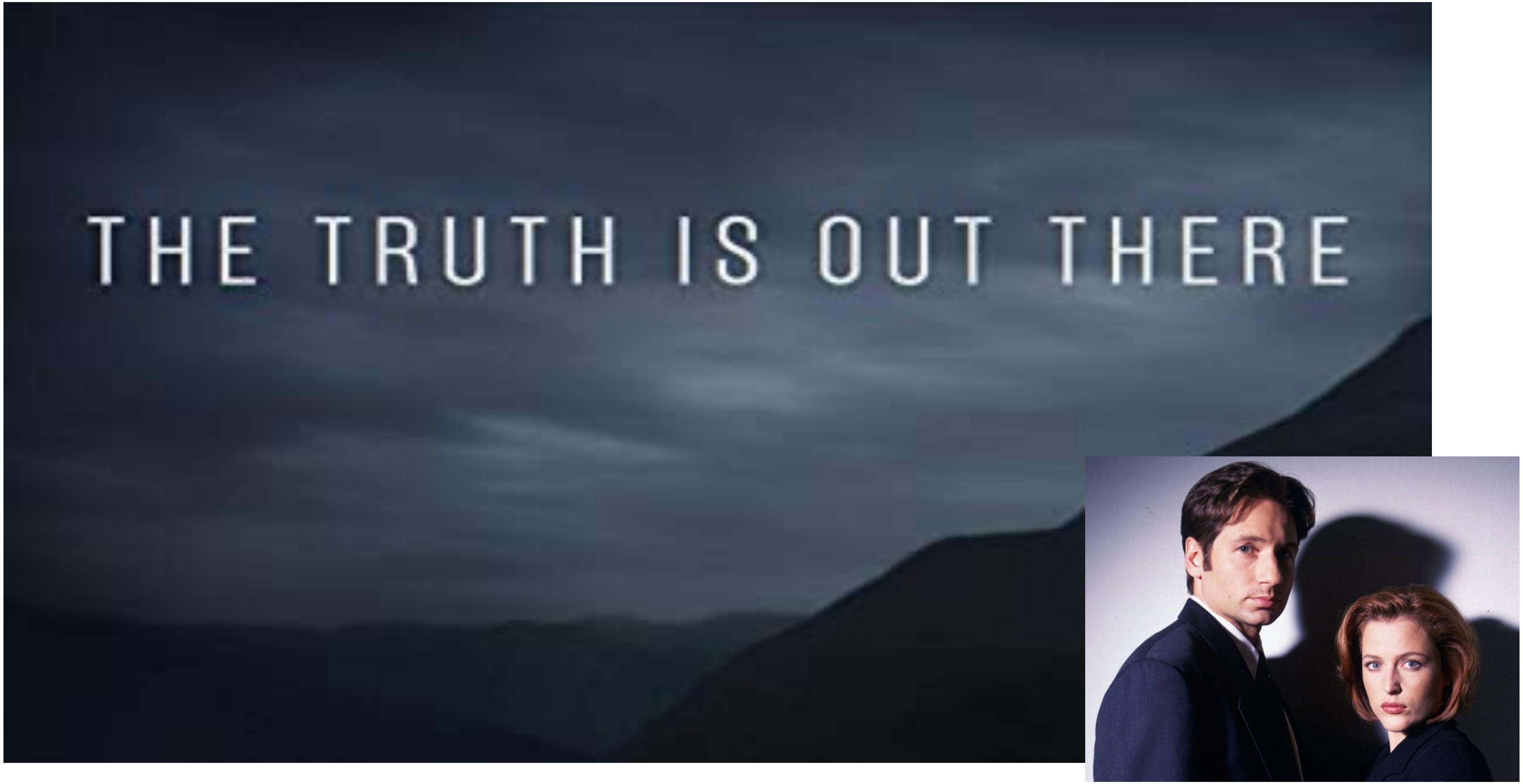
Q: What statistic have we used to describe the linear relationship between quantitative variables?

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

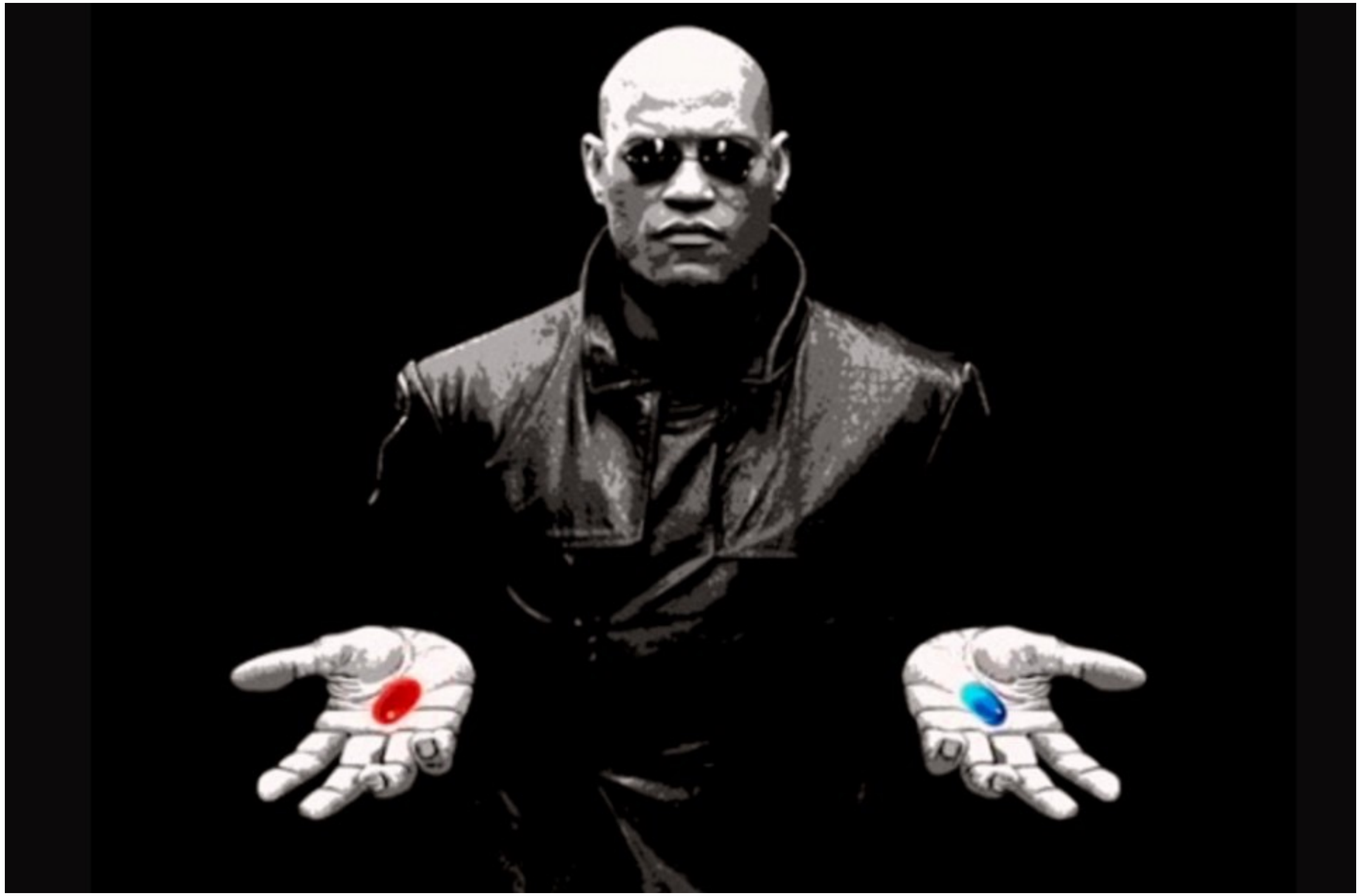
Does correlation imply causation?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

What is our primary focus in Statistics?



Can you handle The TRUTH®?



Ok, let’s ease our way into it…