Measures of spread continued and relationships between two quantitative variables

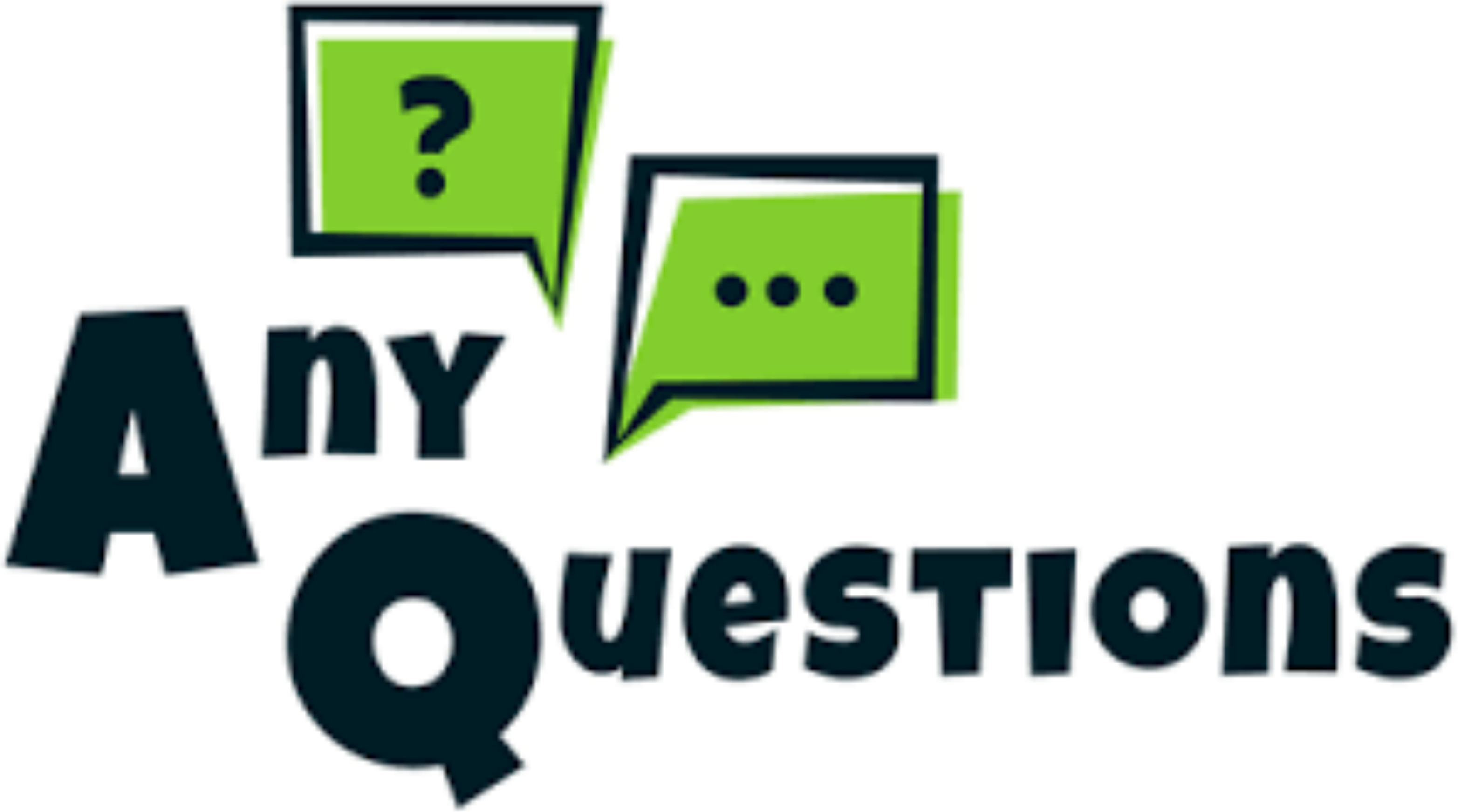
Overview

Quick review:

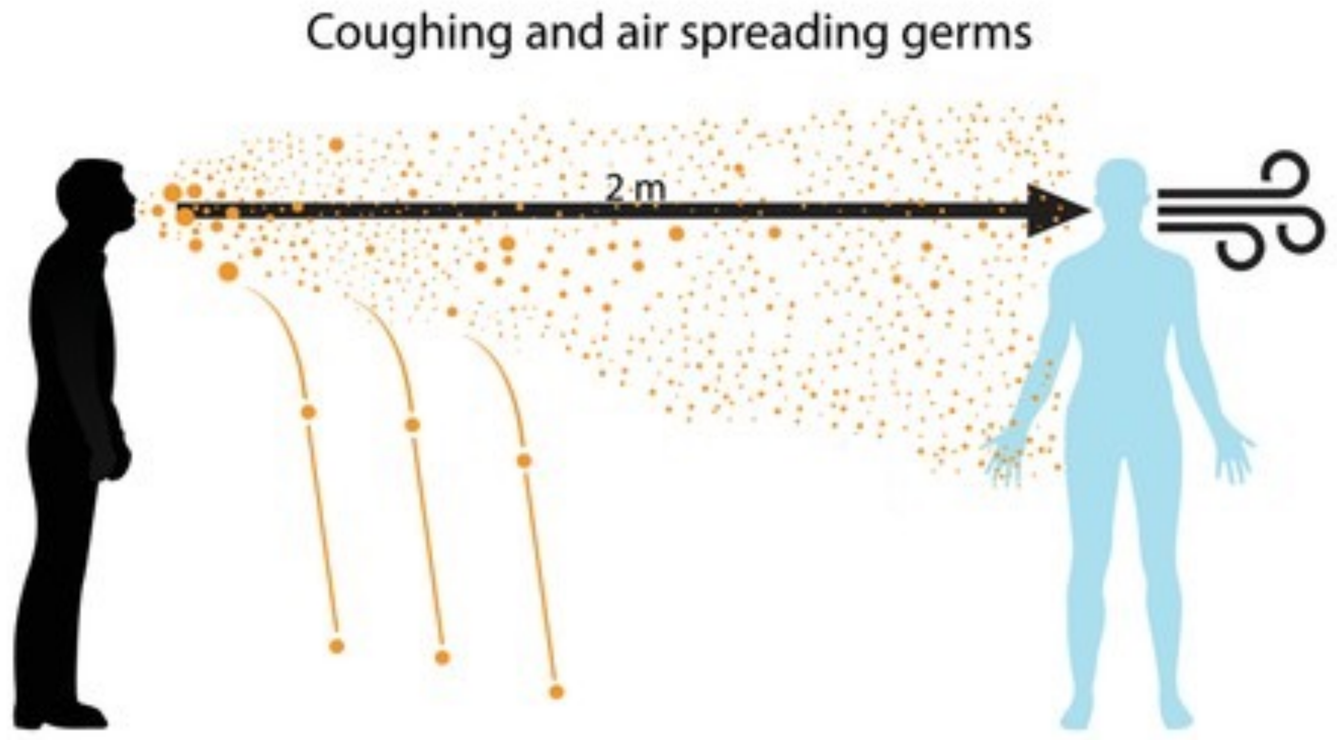
- Standard deviations, z-scores, percentiles

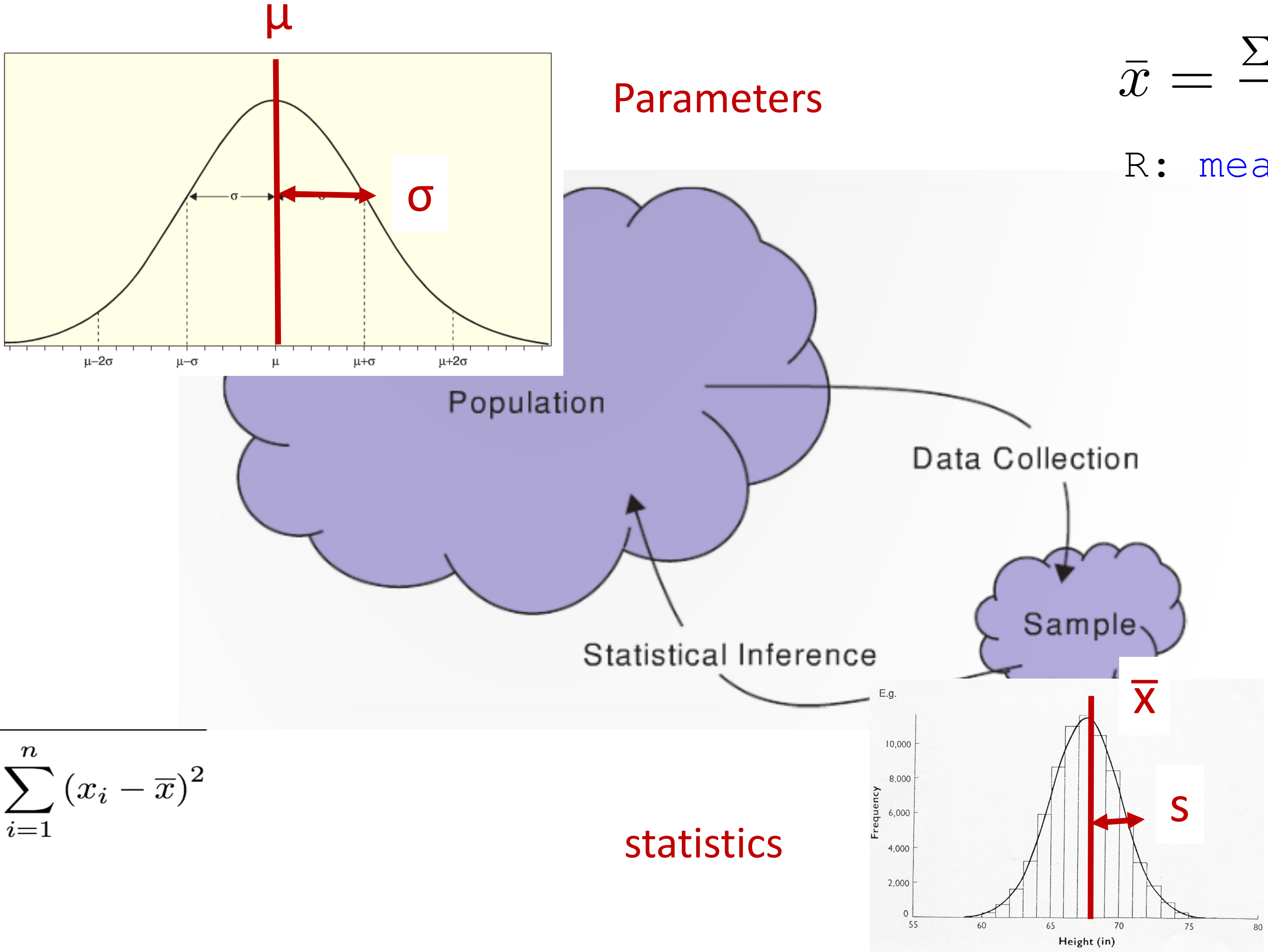
Boxplots

Correlation


?
...
Any Questions

Review and continuation of measures of spread…





Parameters

$$\bar{x} = \frac{\Sigma_i^n x_i}{n}$$

R: `mean(x)`

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

R: `sd(x)`

statistics

The variance

The **variance** is the standard deviation squared

- Population variance = σ^2
- Sample variance = s^2

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

R: `sd(x)`

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

R: `var(x)`

Review: z-scores

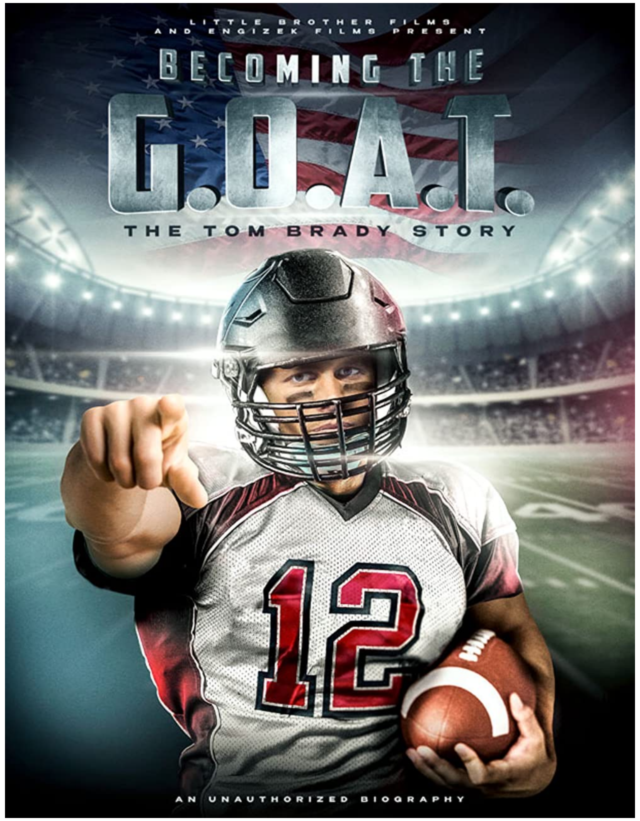
The z-scores tells how many standard deviations a value is from the mean

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

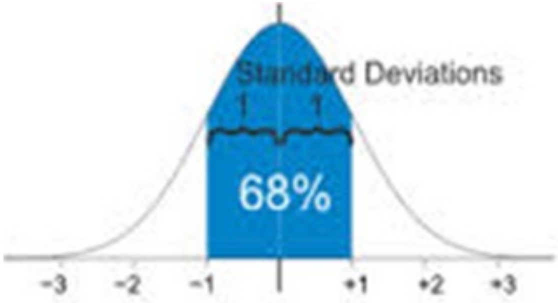
Which 2021-2022 statistic of Tom Brady's is most impressive?

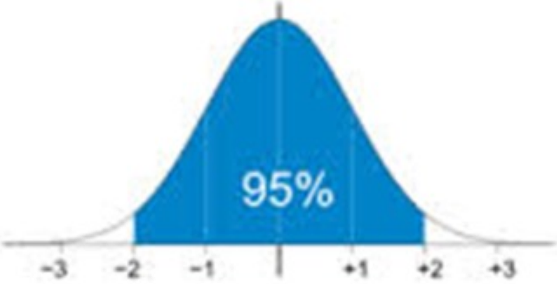
- Compared to all 2021 QBs who played at least 10 games, and attempted at least 100 passes (n = 33)

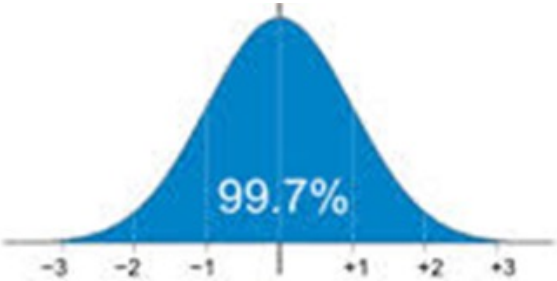
	Raw statistic value	Z-score
Yards	5316	1.73
Touchdowns	43	1.89
Completions	485	1.81
Age	44	2.86



Review: The normal pillow



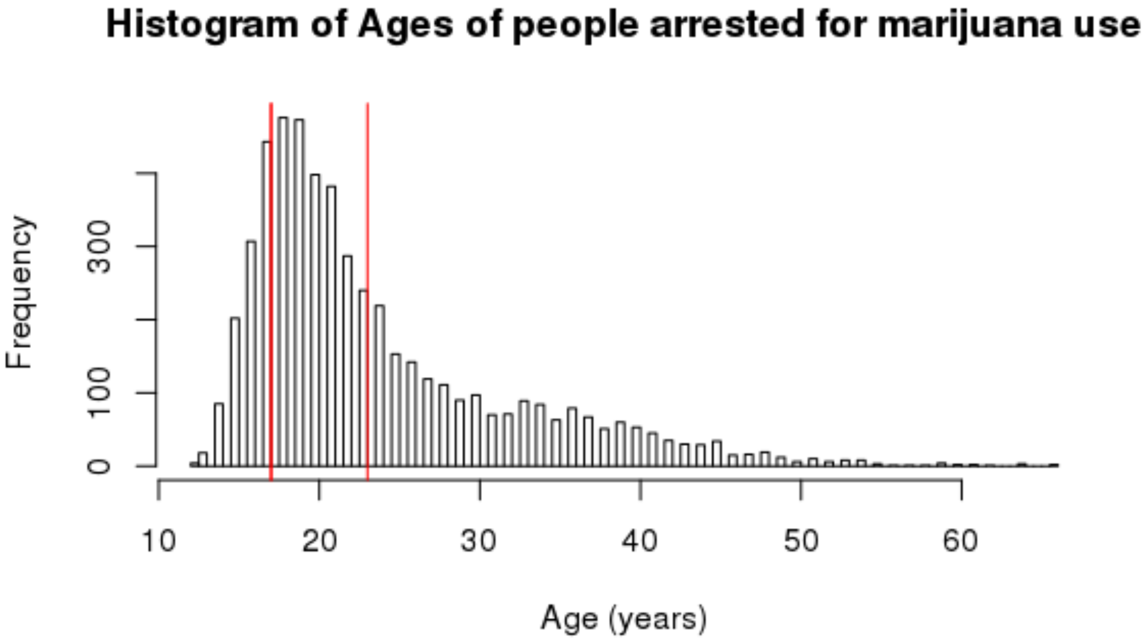




Question: What percent of the pillow's mass is ± 2 standard deviations from the mean?

Review: quantiles (percentiles)

The **pth percentile** is a quantitative value ***x*** which is greater than p percent of the data



60th percentile value is 23

i.e., 60% of the arrests were of ages 23 or less

In R: `quantile(Arrests$age, .6)`

The quantile universe

Five-Number Summary = (minimum, Q_1, median, Q_3, maximum)

Q_1 = 25th percentile, Q_3 = 75th percentile

Range = maximum – minimum

Interquartile range (IQR) = $Q_3 - Q_1$

As a rule of thumb, we call a data value an **outlier** if it is:

Smaller than: $Q_1 - 1.5 * \text{IQR}$

Larger than: $Q_3 + 1.5 * \text{IQR}$

In R: `fivenum(v)`

Box plots

A **box plot** is a graphical display of the five-number summary and consists of:

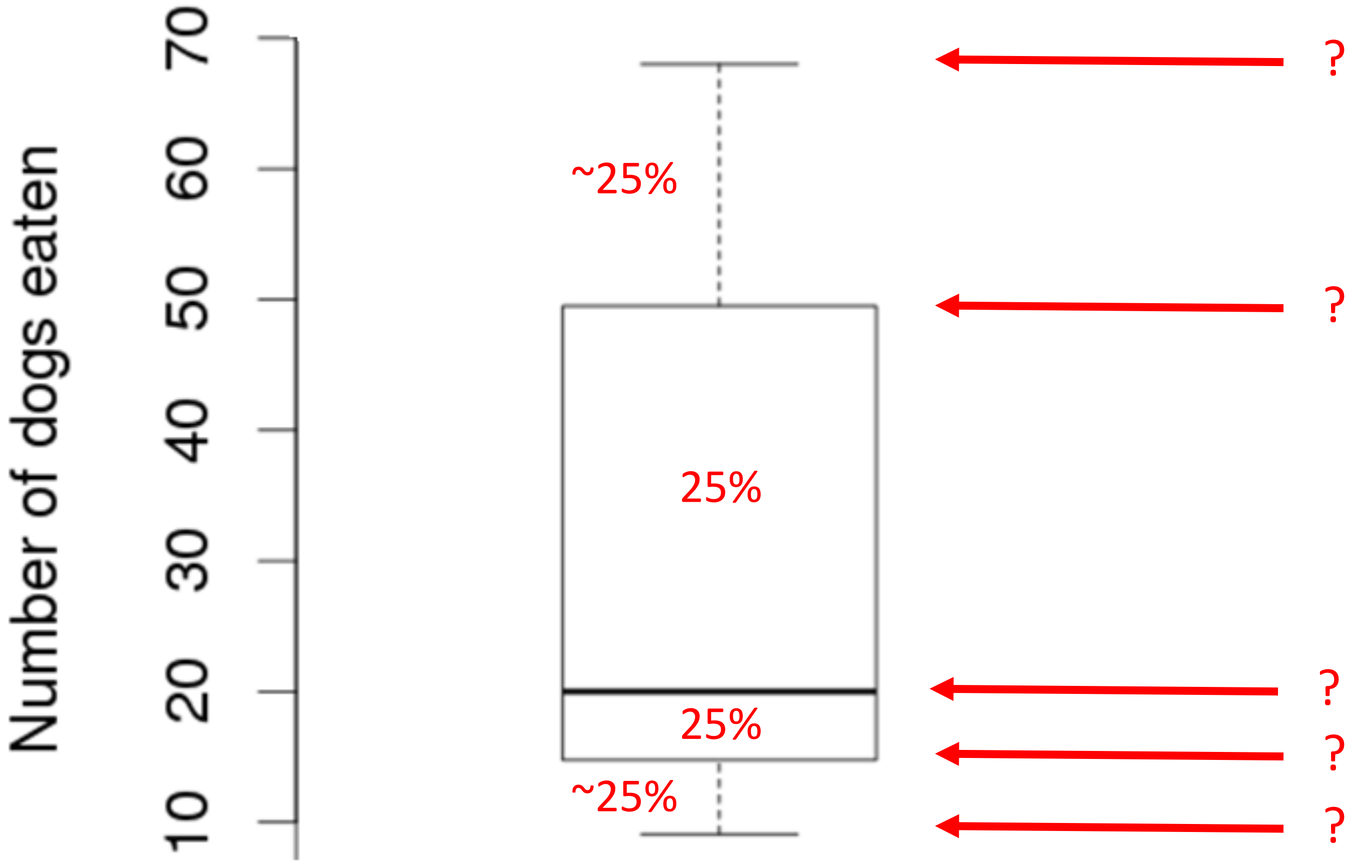
1. Drawing a box from Q_1 to Q_3

2. Dividing the box with a line (or dot) drawn at the median

3. Draw a line from each quartile to the most extreme data value that is not and outlier

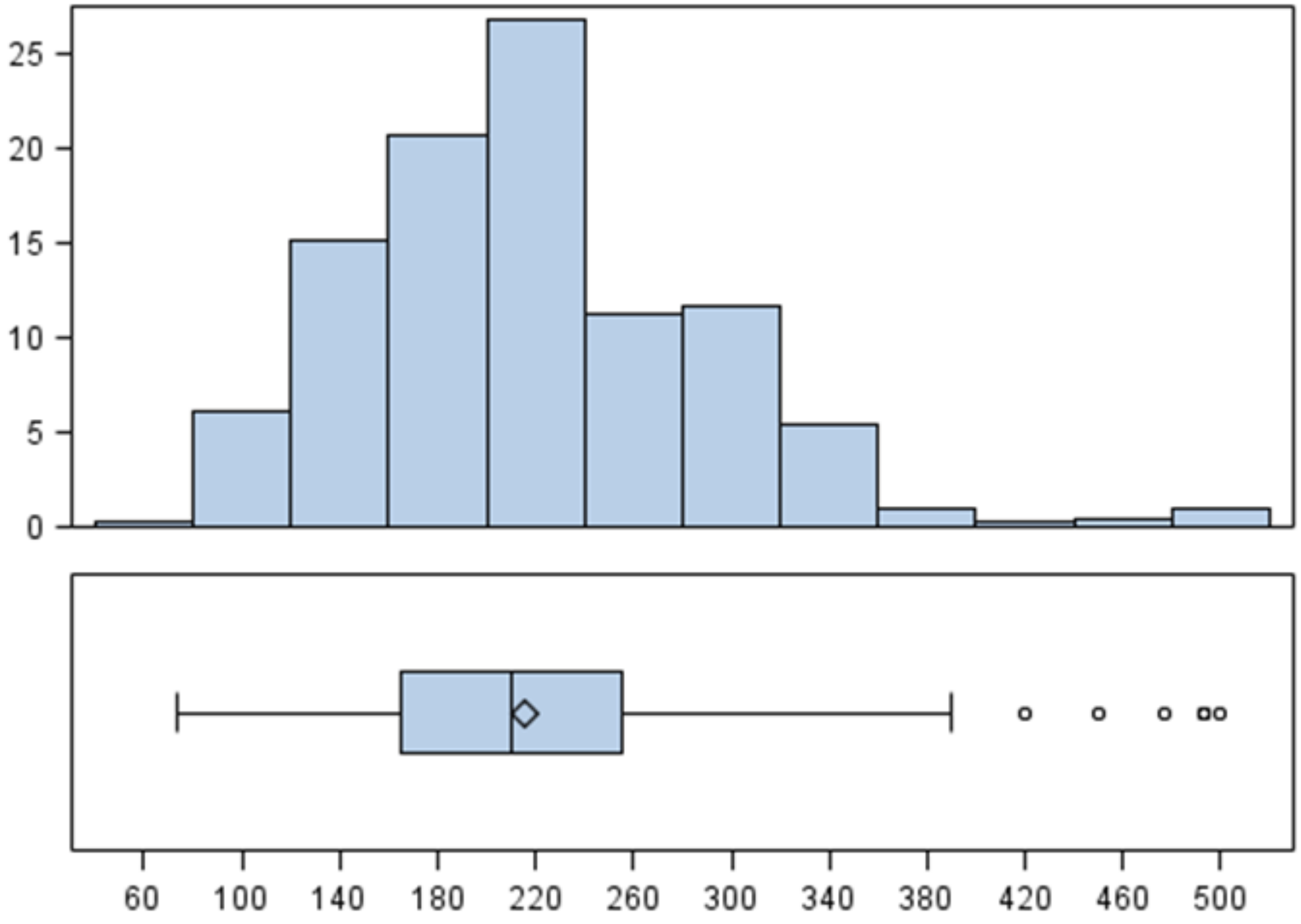
4. Draw a dot/asterisk for each outlier data point.

Box plot of the number of hot dogs eaten by the men's contest winners 1980 to 2010



R: `boxplot(v)`

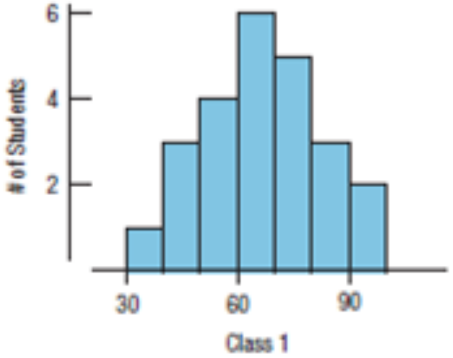
Box plots extract key statistics from histograms



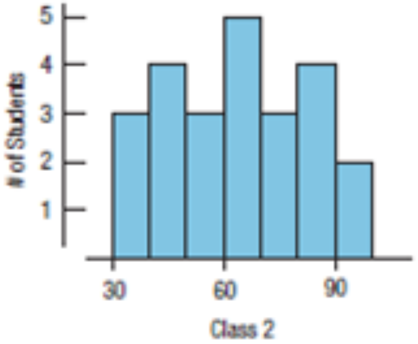
Box plots extract key statistics from histograms

Question: which Box plot goes with which histogram?

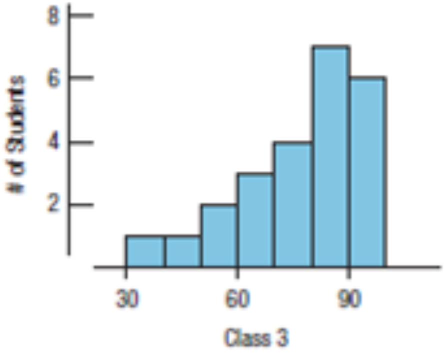
Histogram 1

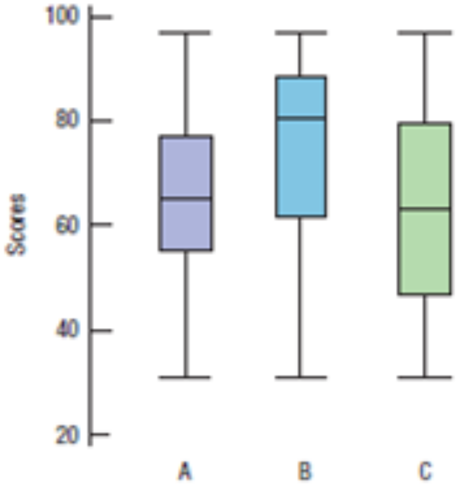


Histogram 2



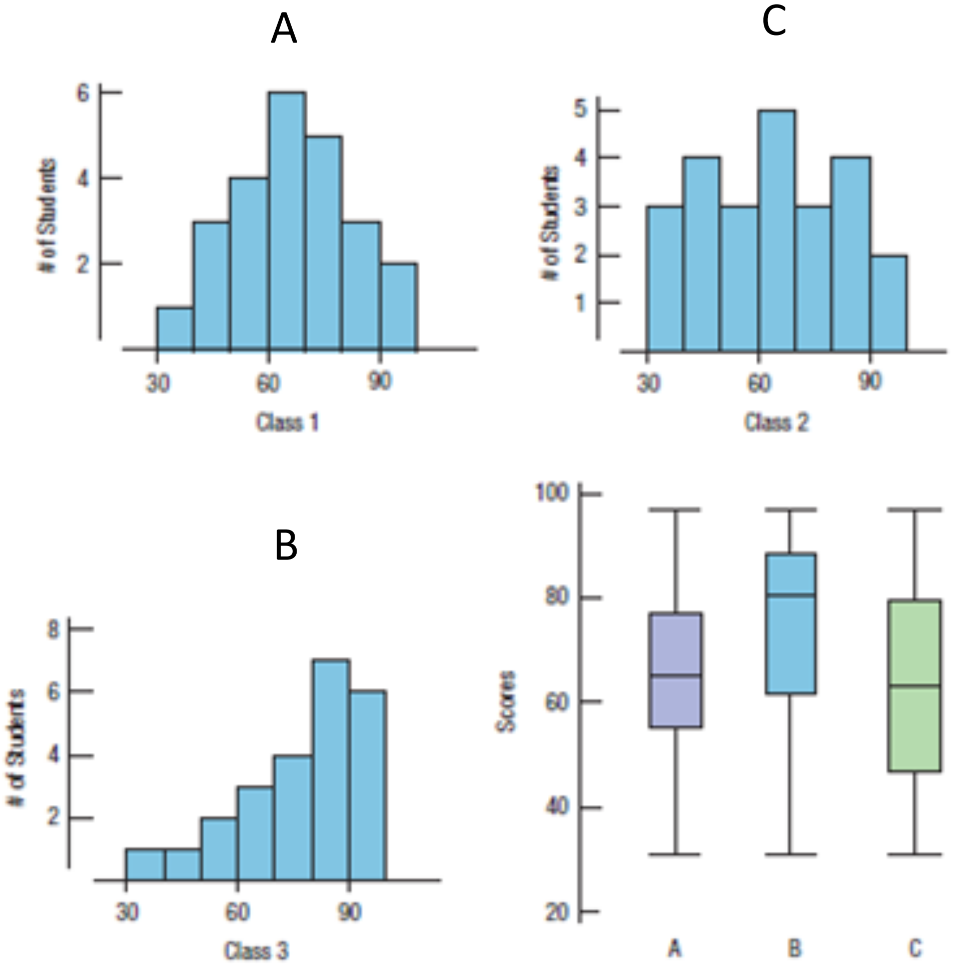
Histogram 3





Box plots extract key statistics from histograms

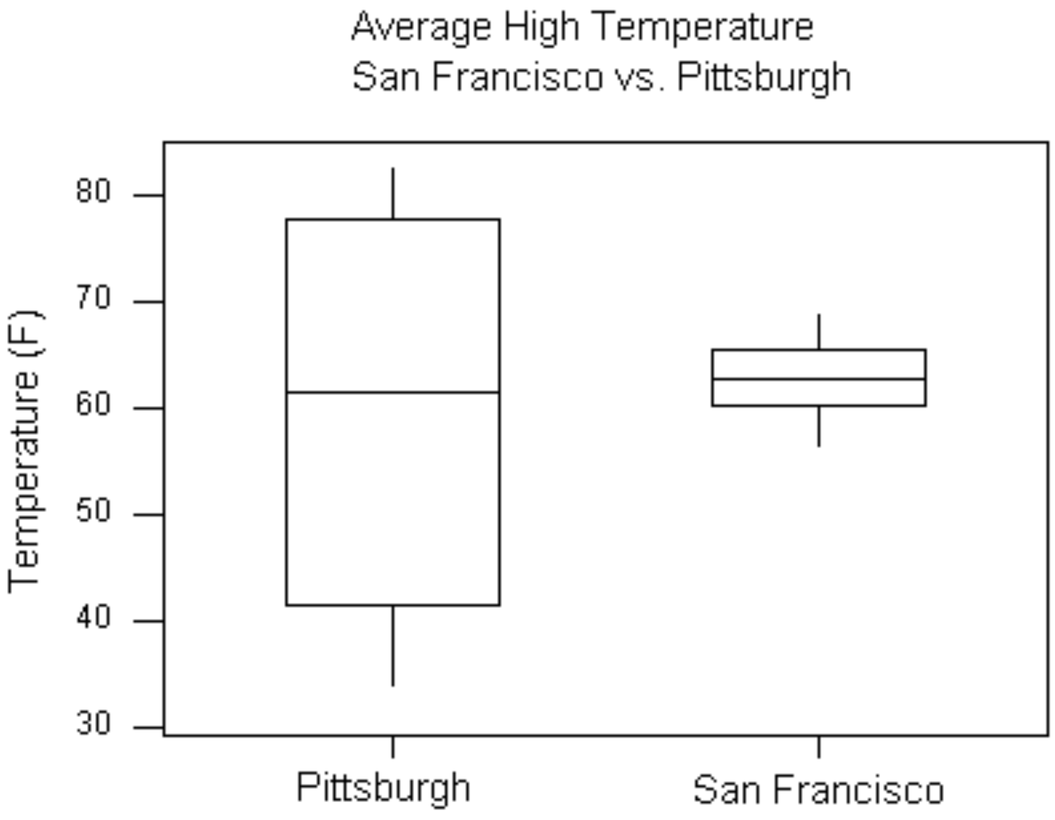
Question: which Box plot goes with which histogram?



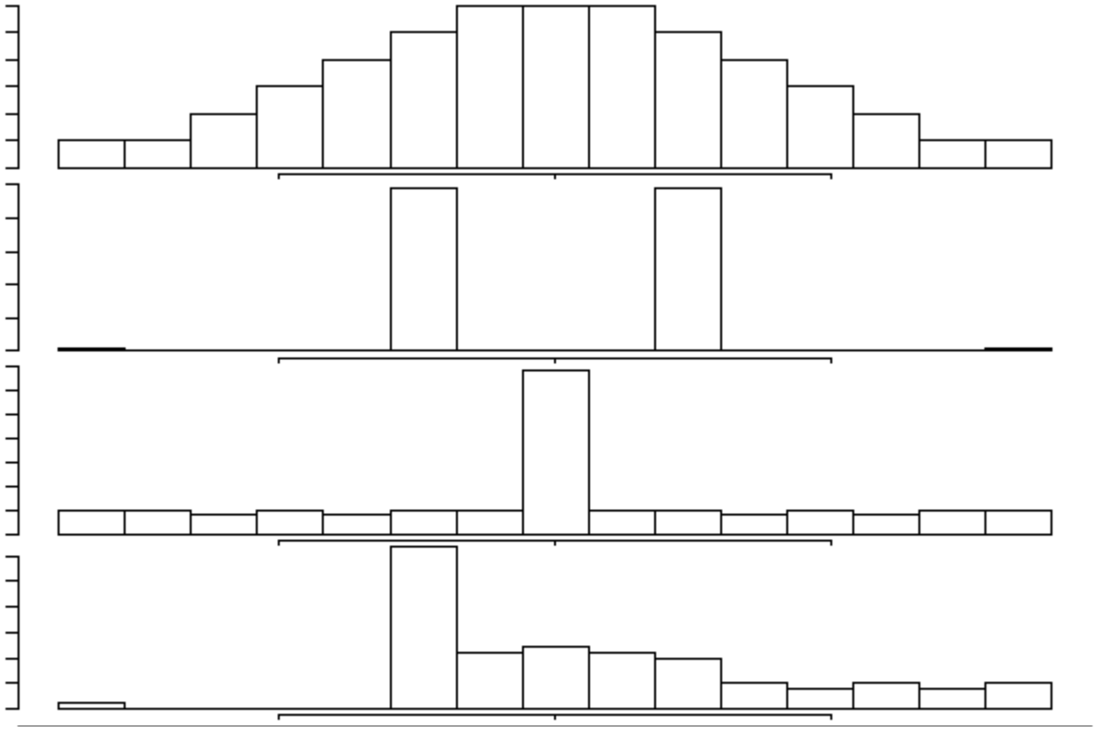
Comparing quantitative variables across categories

Often one wants to compare quantitative variables across categories

Side-by-Side graphs are a way to visually compare quantitative variables across different categories



Box plots don’t capture everything



Do you think the box plots for these distributions look similar?

Side-by-size boxplots in R

```
> boxplot(v1, v2,                                  # compare two vectors v1 and v2
     names = c("name 1", "name 2"),                # labels below each box plot
      ylab =  "y-axis name"                        # y-axis label name
  )
```

Let's explore side-by side boxplots on the Bechdel data to try to see if movies that pass the Bechdel test make a larger profit!

Relationships between two quantitative variables

Do movies with larger budgets make more money?

Q: How could we visualize the data to see if this is true?

A: Create a scatter plot!

Scatterplot

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

- Each axis represents the value of one variables
- Each point the plot shows the value for the two variables for a single data case

If there is an explanatory and response variable, then the explanatory variable is put on the x-axis and the response variable is put on the y-axis.

Do movies with larger budgets make more money?

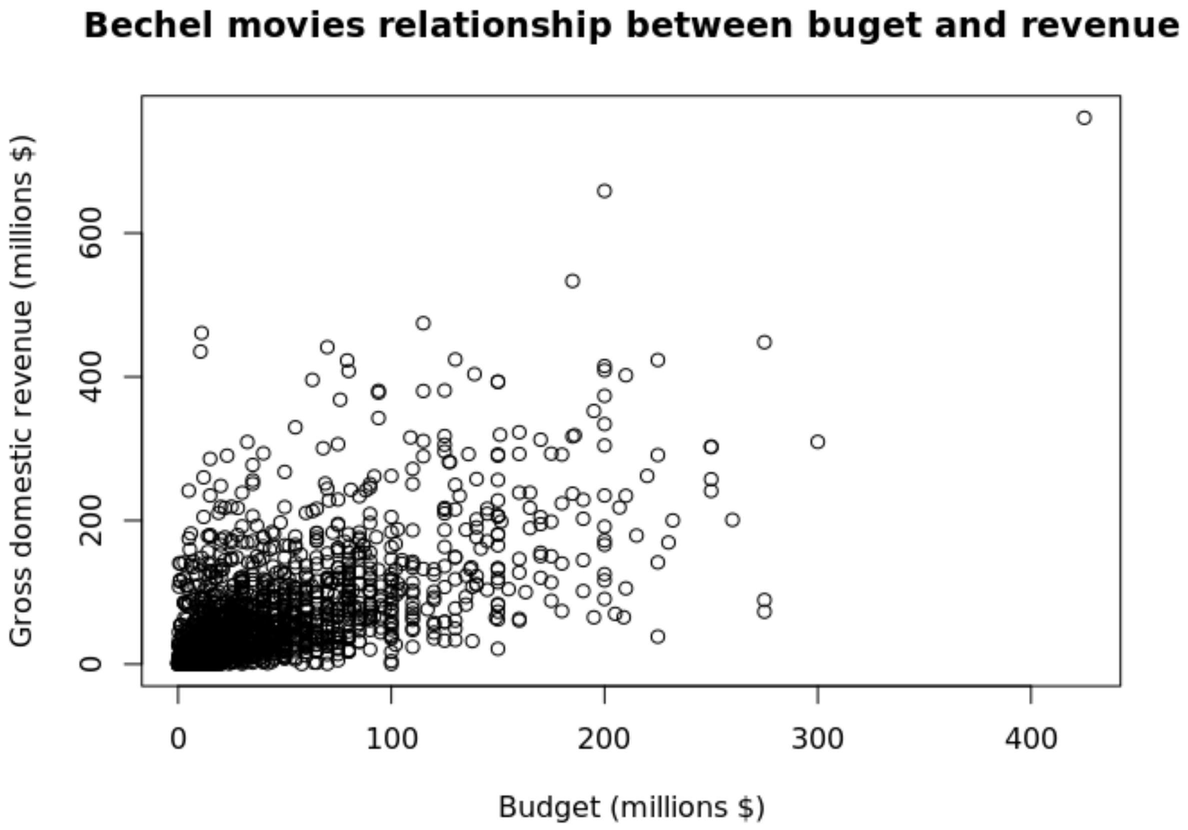
Q: If we want to create a scatter plot to address our question, what variables should we use in our plot?

A:

- budget_2013
- domgross_2013

Let's try it in R!

Relationship movie money spent and made



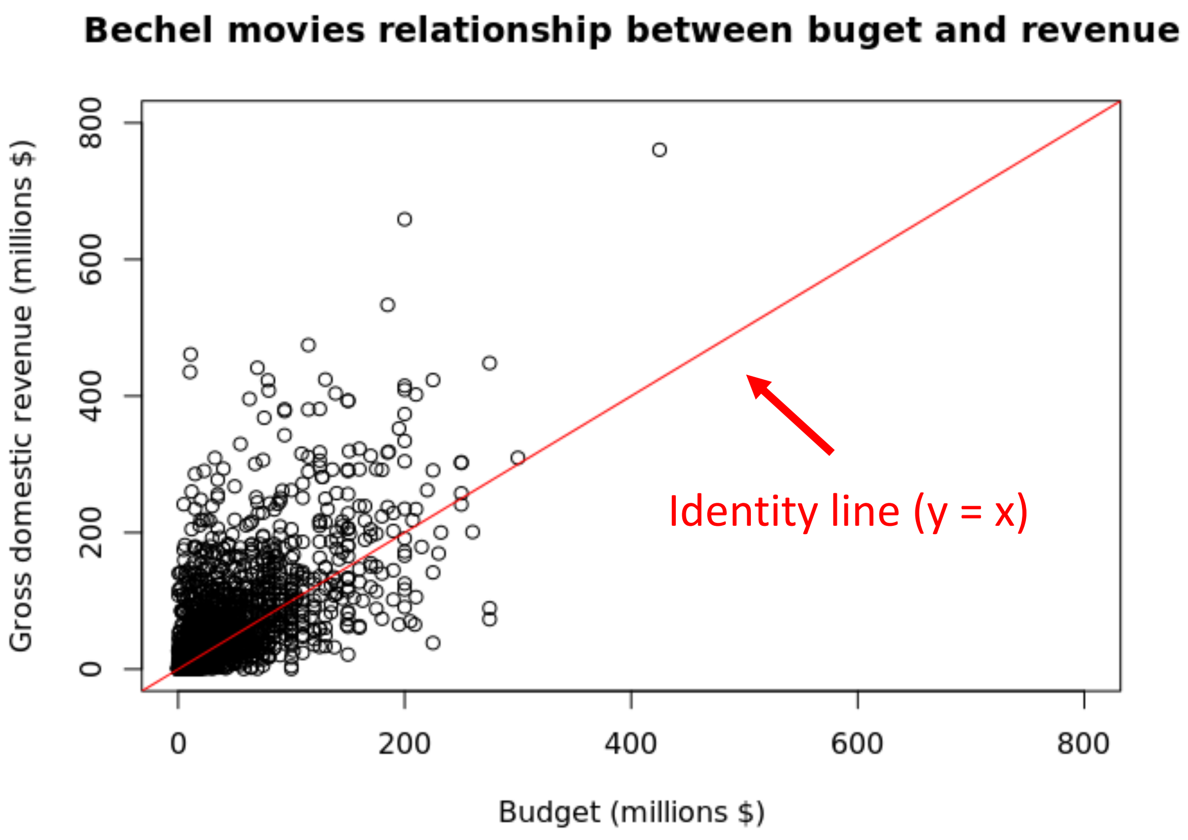
Do movies with larger budgets make more money?

Do most movies make money?

- How could we create a more informative scatter plot of this data?

R: `plot(x, y)`

Relationship movie money spent and made



Do movies with larger budgets make more money?

Do most movies make money?

- How could we create a more informative scatter plot of this data?

R: `plot(x, y)`

Questions when looking at scatterplots

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?

Questions when looking at scatterplots

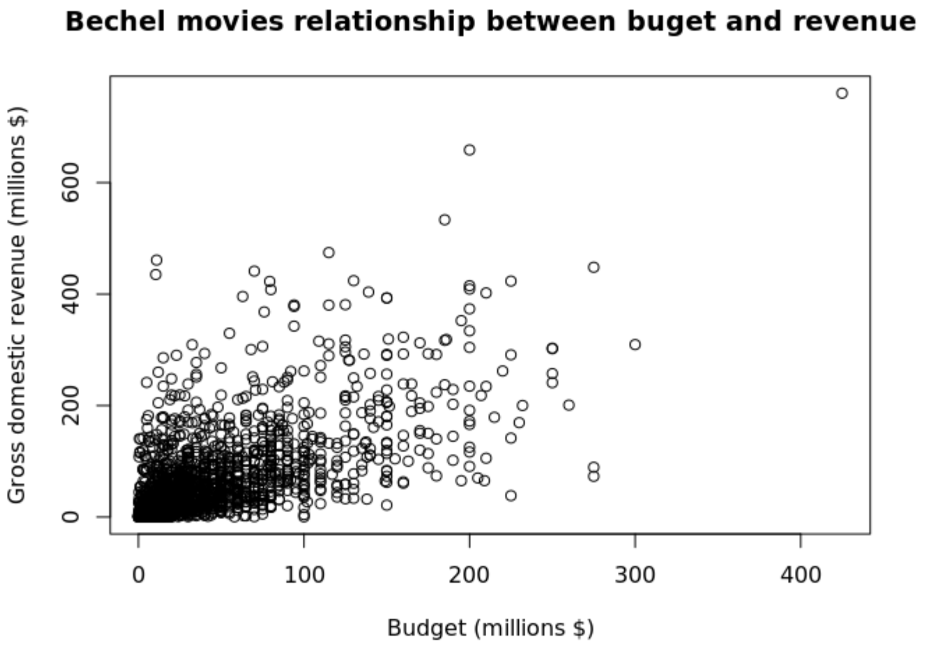
Do the points show a clear trend?

- 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?

Budget and revenue



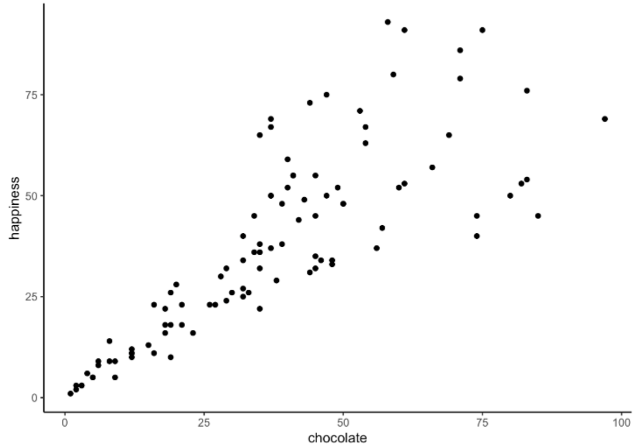
Positive, negative, no correlation

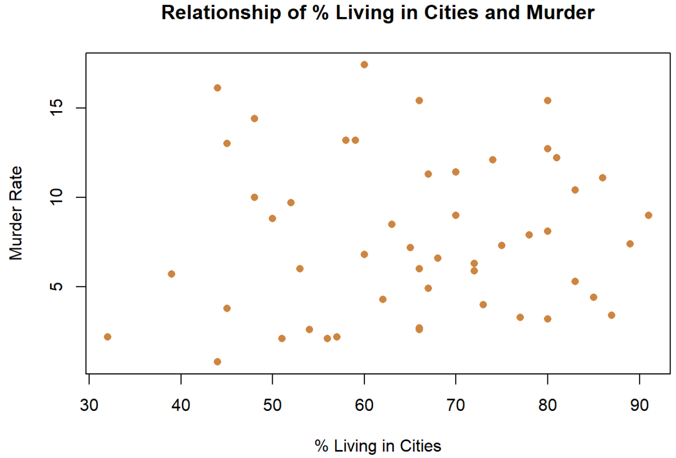
Do the points show a clear trend?

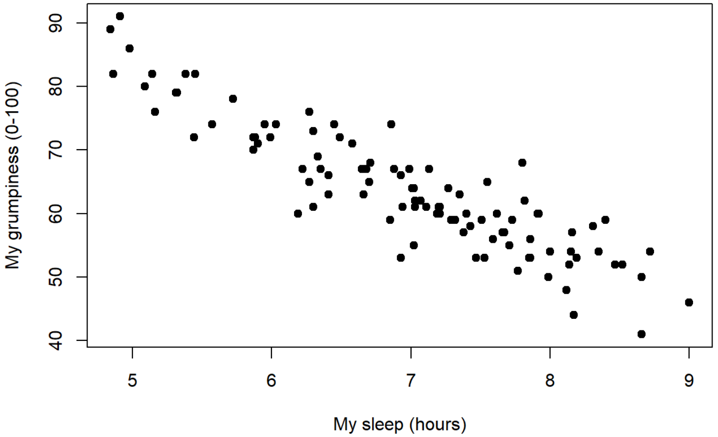
- Does it go upward or downward?
- How much scatter around the trend?

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

Are there any outlier points?




Relationship of % Living in Cities and Murder



The correlation coefficient

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

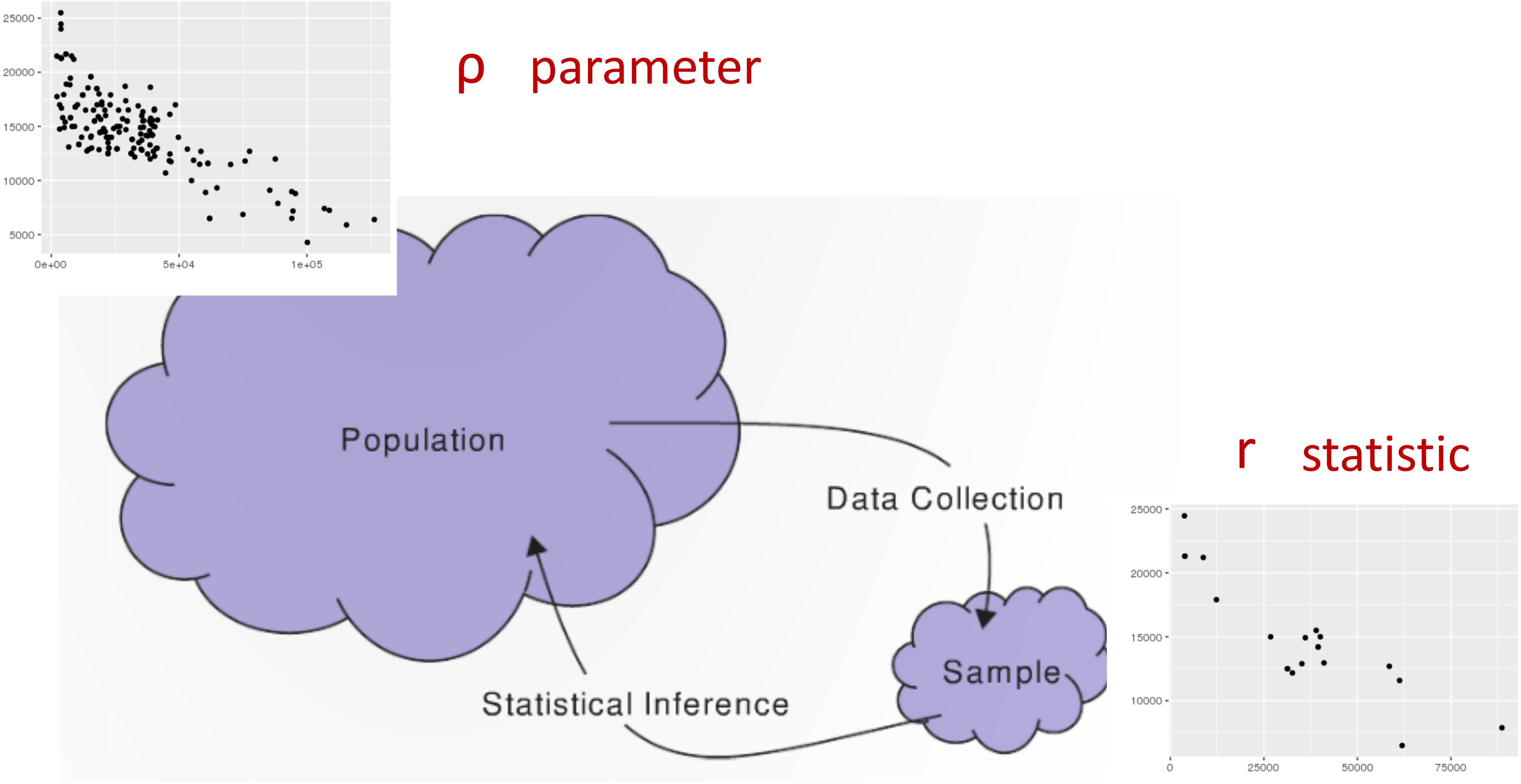
$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

- The correlation for a sample is denoted with **r**
- The correlation in the population is denoted with **ρ**

  (the Greek letter rho)

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

ρ parameter

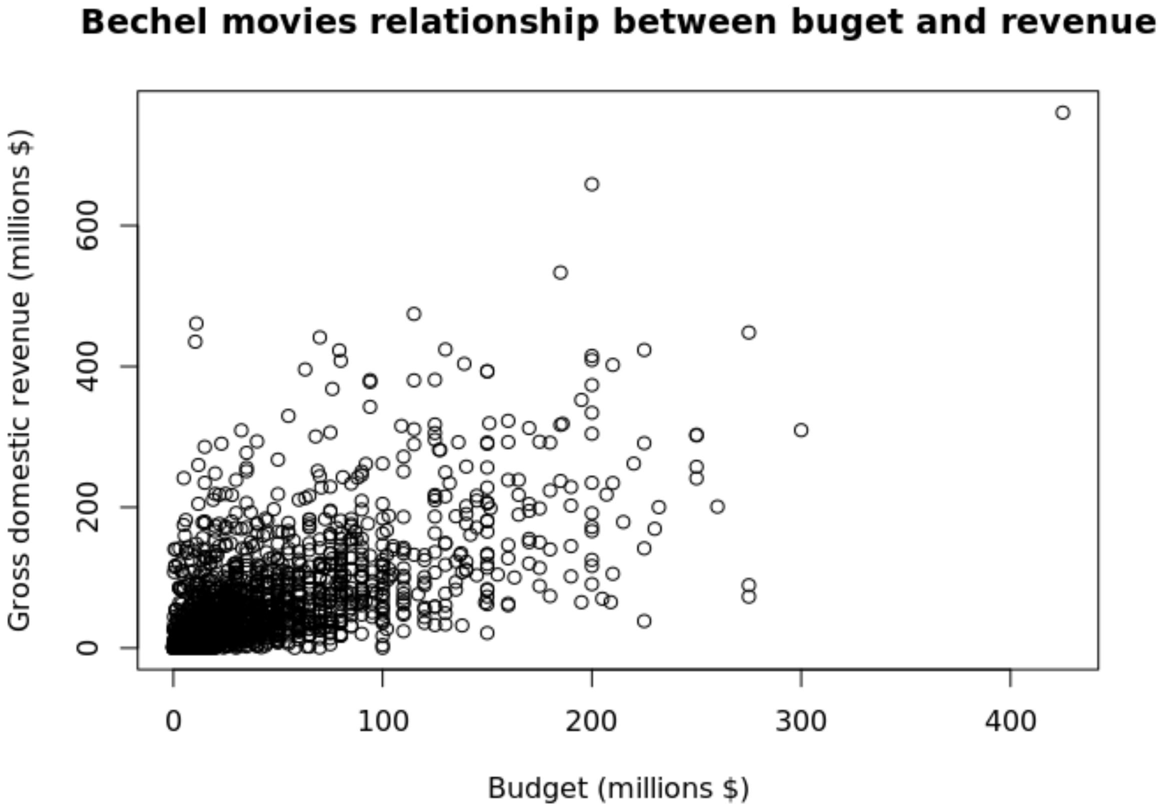


r statistic

Movie budget and revenue correlation?

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

r = ?



Properties of the correlation

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

The sign of r indicates the direction of the association

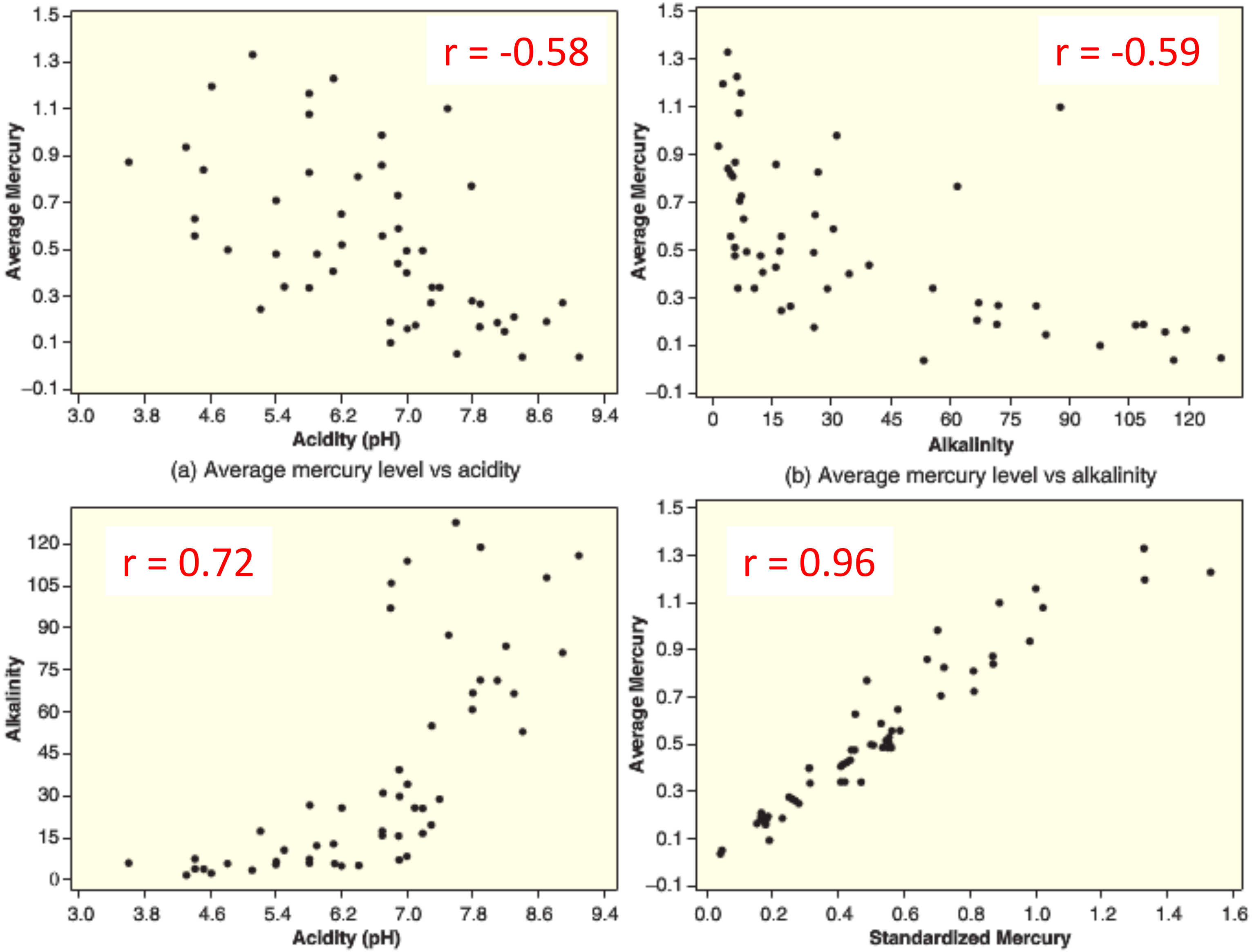
Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

Correlation is symmetric: r = cor(x, y) = cor(y, x)

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

Florida lakes

Correlation game



r = -0.58

(a) Average mercury level vs acidity

r = -0.59

(b) Average mercury level vs alkalinity

r = 0.72

(c) Alkalinity vs acidity

r = 0.96

(d) Average vs standardized mercury levels

Let’s calculate some correlations in R!

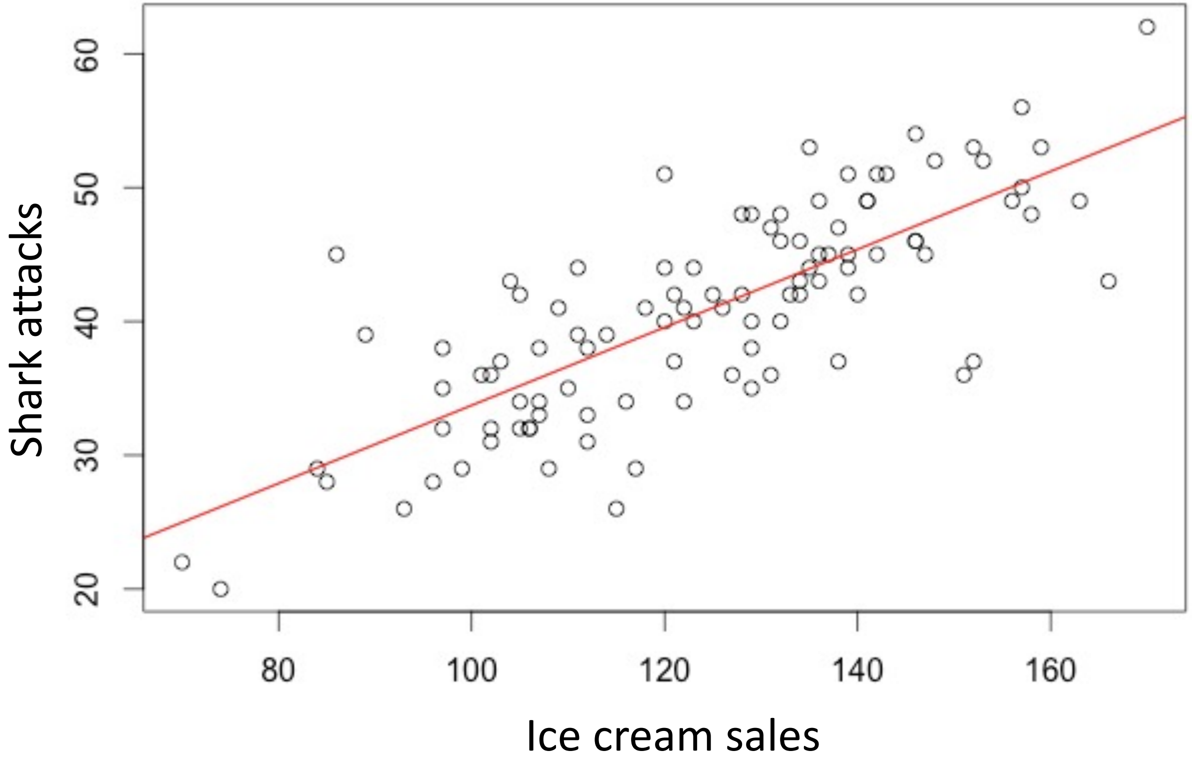
Let's examine the correlation between budget of a movie and the amount of revenue the movie made

```
# load the data
> library(fivethirtyeight)
> bechdel <- na.omit(bechdel)      # remove data points with missing values

# create a scatter plot and calculate the correlation
> plot(bechdel$budget_2013, bechdel$domgross_2013)
> cor(bechdel$budget_2013, bechdel$domgross_2013)
```

Correlation caution #1

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



© MARK ANDERSON

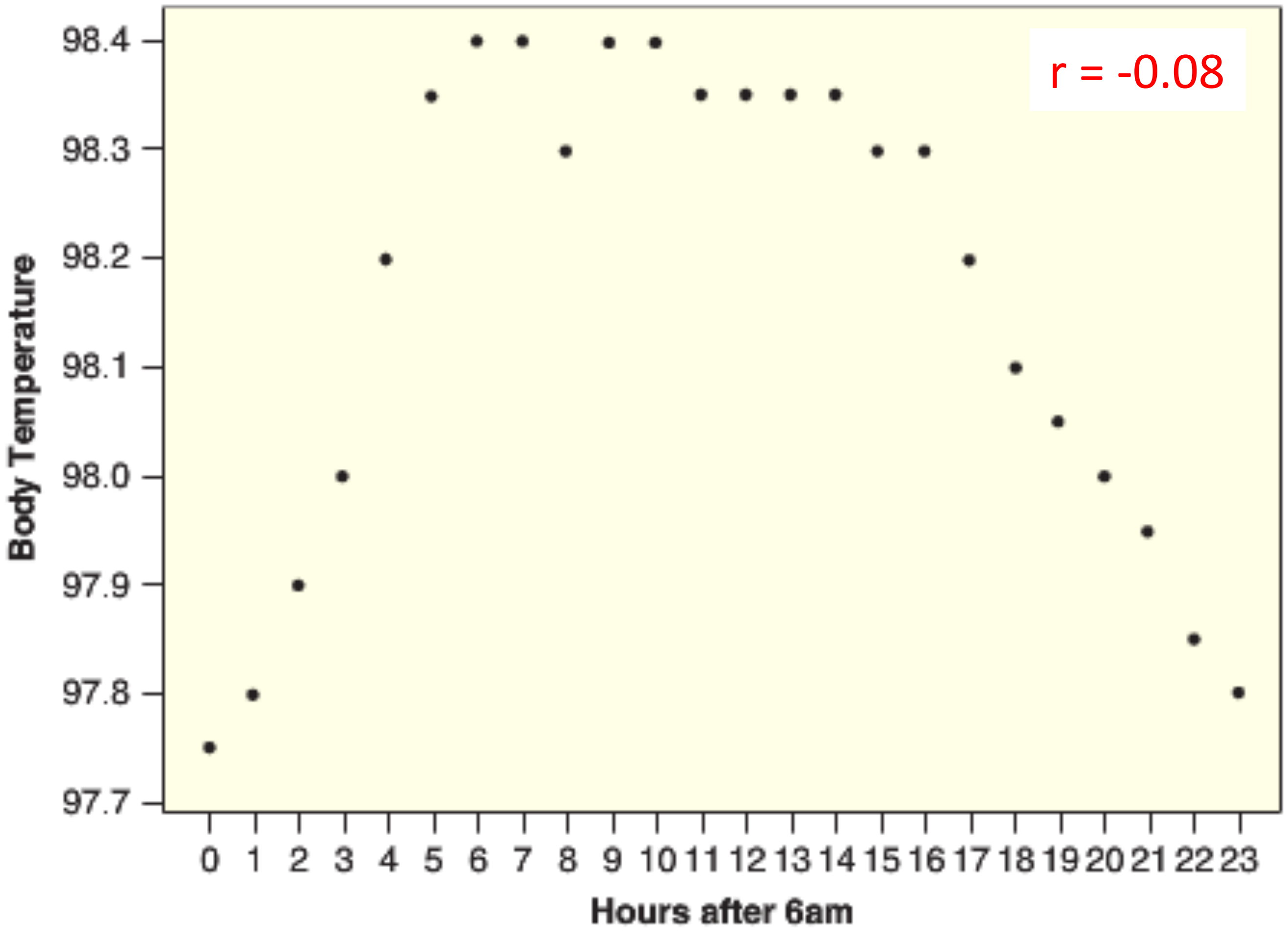
WWW.ANDERTOONS.COM

“It’s important to remember that correlation does not imply causation. Besides, we all know it was Brian.”

Correlation caution #2

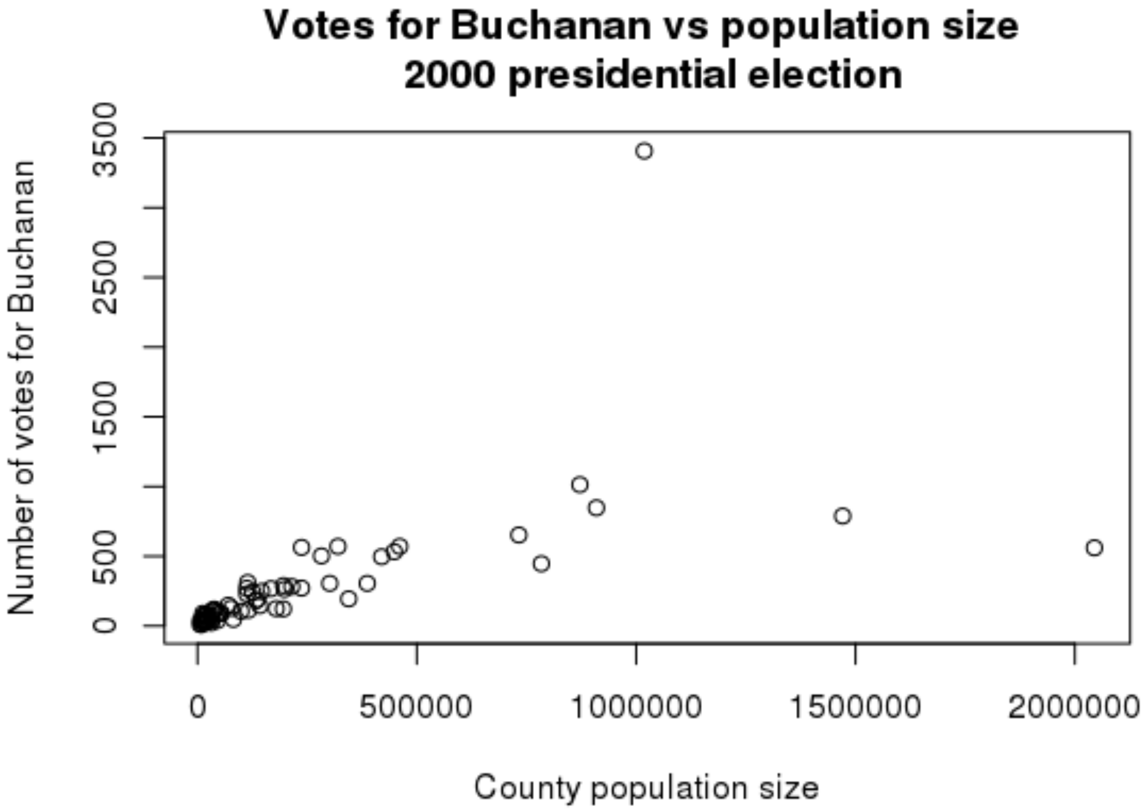
A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.

Body temperature as a function of time of the day



Correlation caution #3

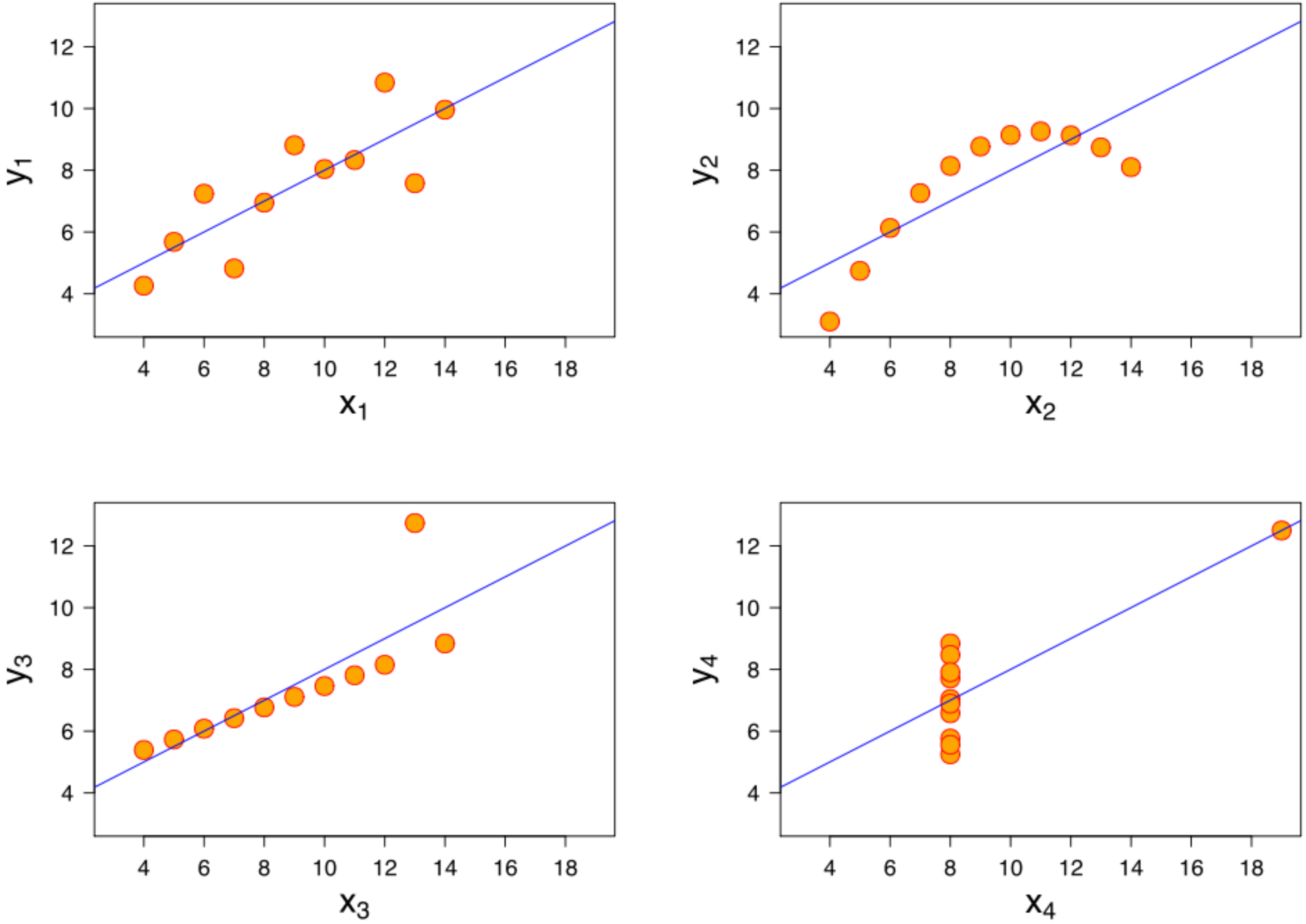
Correlation can be heavily influenced by outliers. Always plot your data!

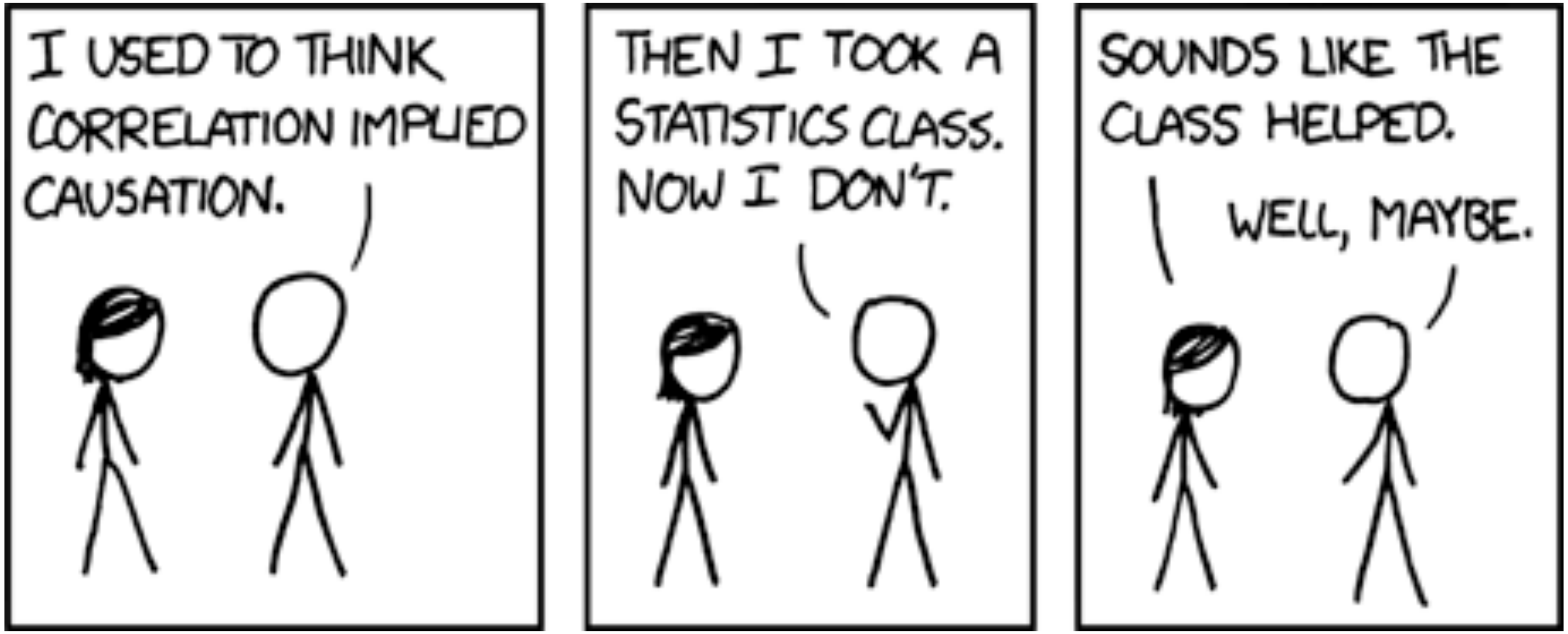


With Palm Beach
r = 0.61

Without Palm Beach
r = .78

Anscombe’s quartet (r = 0.81)




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

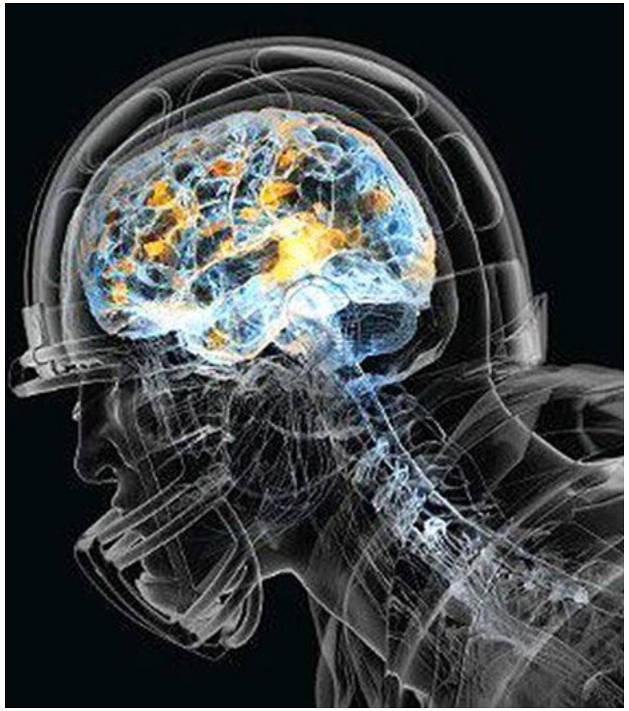
Practice questions

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:

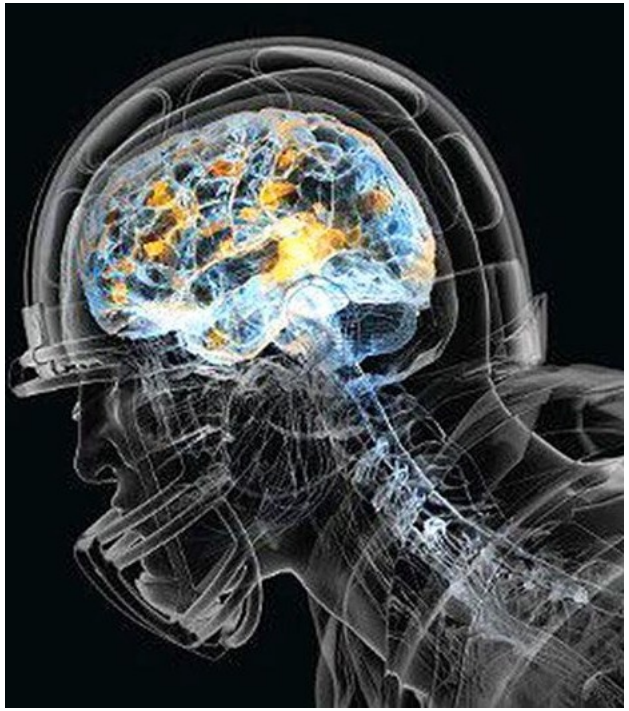
1. The relationship between number of years playing football and hippocampus volume
2. The relationship between hippocampus size and the three groups

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