Relationships between variables

Outline for today

Better know some players double header

Quick review of box plots and z-scores

Relationships between variables

- Scatter plots
- Correlation Coefficient

Better know some players

Michael Jordan and Buster Posey

Review of:

5 number summaries,

box plots,

and z-scores

Five Number Summary

Five Number Summary = (min, Q_1, median, Q_3, max)

Q_1 = 25th percentile (also called 1st quartile)

Q_3 = 75th percentile (also called 3rd quartile)

Range = maximum – minimum

Interquartile range (IQR) = $Q_3 - Q_1$

```
R: fivenum(x)
```

Boxplots

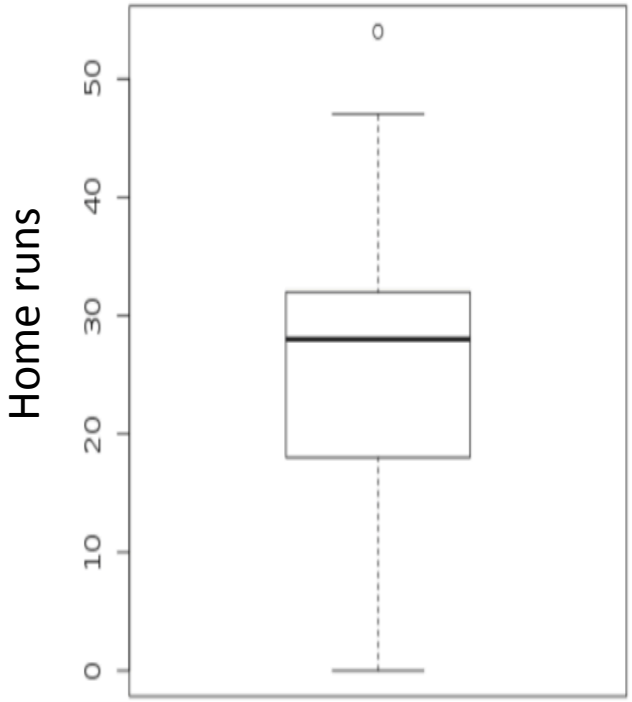
A **boxplot** is a graphical display of the 5 number summary and consists of:

1. Drawing a box from Q_1 to Q_3

2. Dividing the box with a line 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.




```
boxplot(Ted.data$HR, Wade.data$HR, names = c('Ted', 'Wade'))
```

z-scores

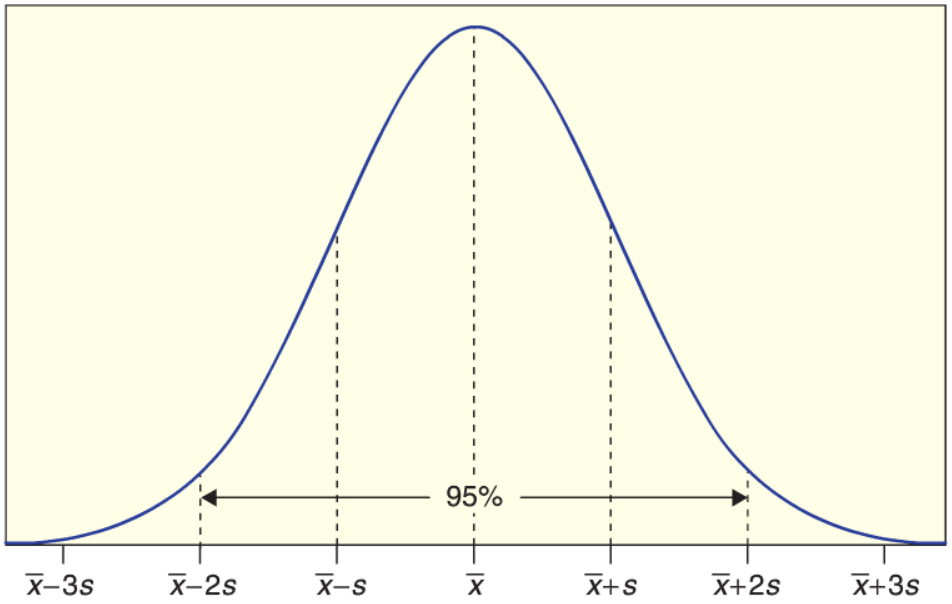
$$z_i = \frac{x_i - \bar{x}}{s}$$

The z-scores tells how many standard deviations a value x_i is from the mean ($\bar{x}$), in a way that is independent of the units of measurement.

The 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.

i.e., 95% of the data is in the interval: $\bar{x}$ -2s to $\bar{x}$ +2s



Relationships between variables

Do power hitters strike out more often?

Chris Davis in 2013:

53 home runs

199 strike outs

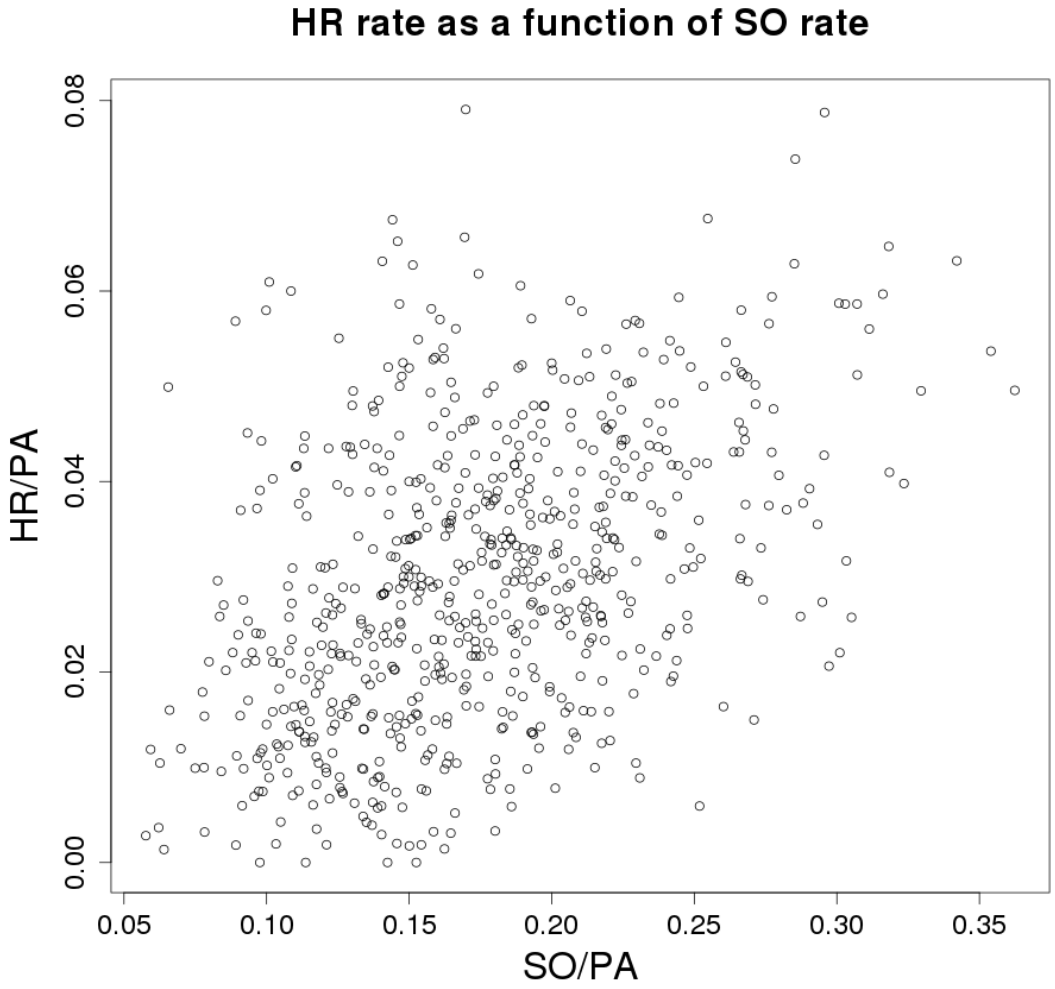
Scatter plots

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.

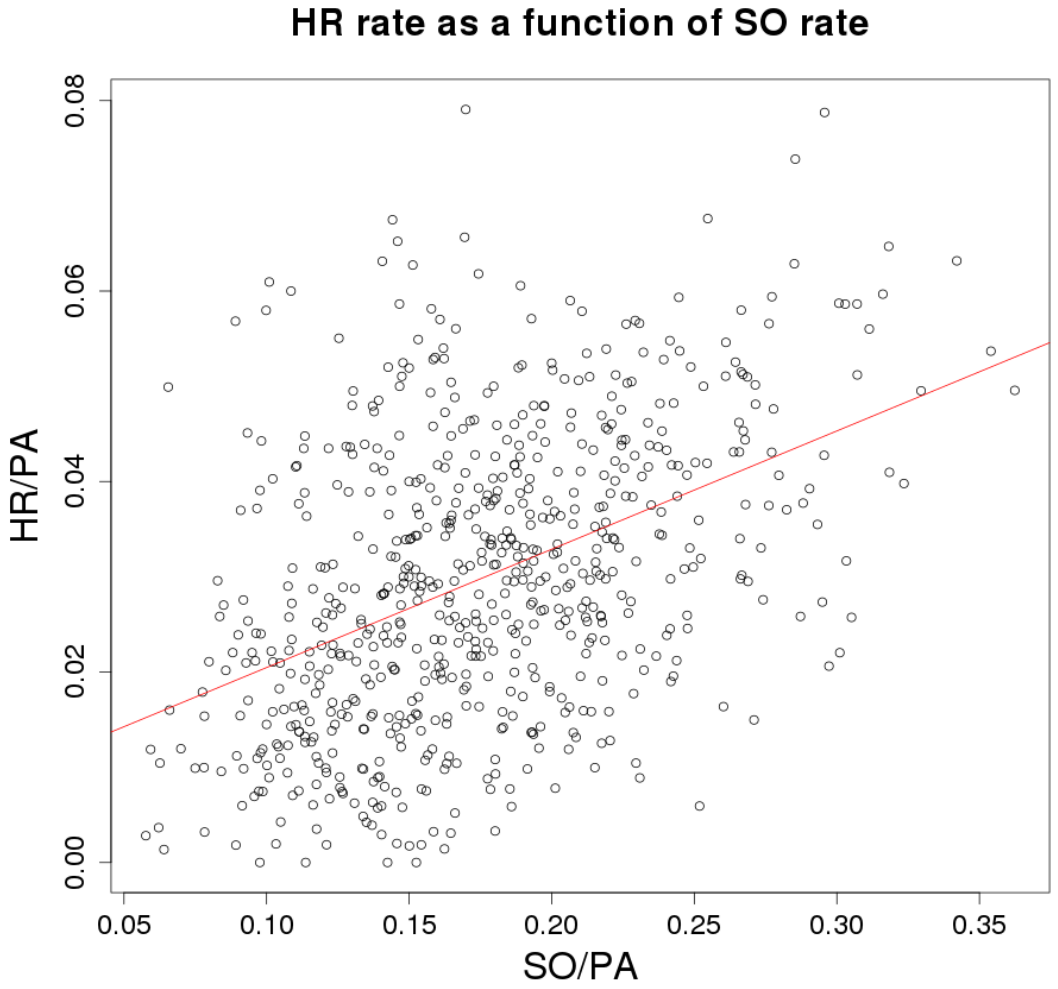
If there is an explanatory and response variable, then the explanatory variable is put on the x-axis and the response variable is put on the y-axis

Scatter plots: 2010-2014 home run rate as a function of strike out rate




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

Scatter plots: 2010-2014 home run rate as a function of strike out rate




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

Questions when looking at scatter plots

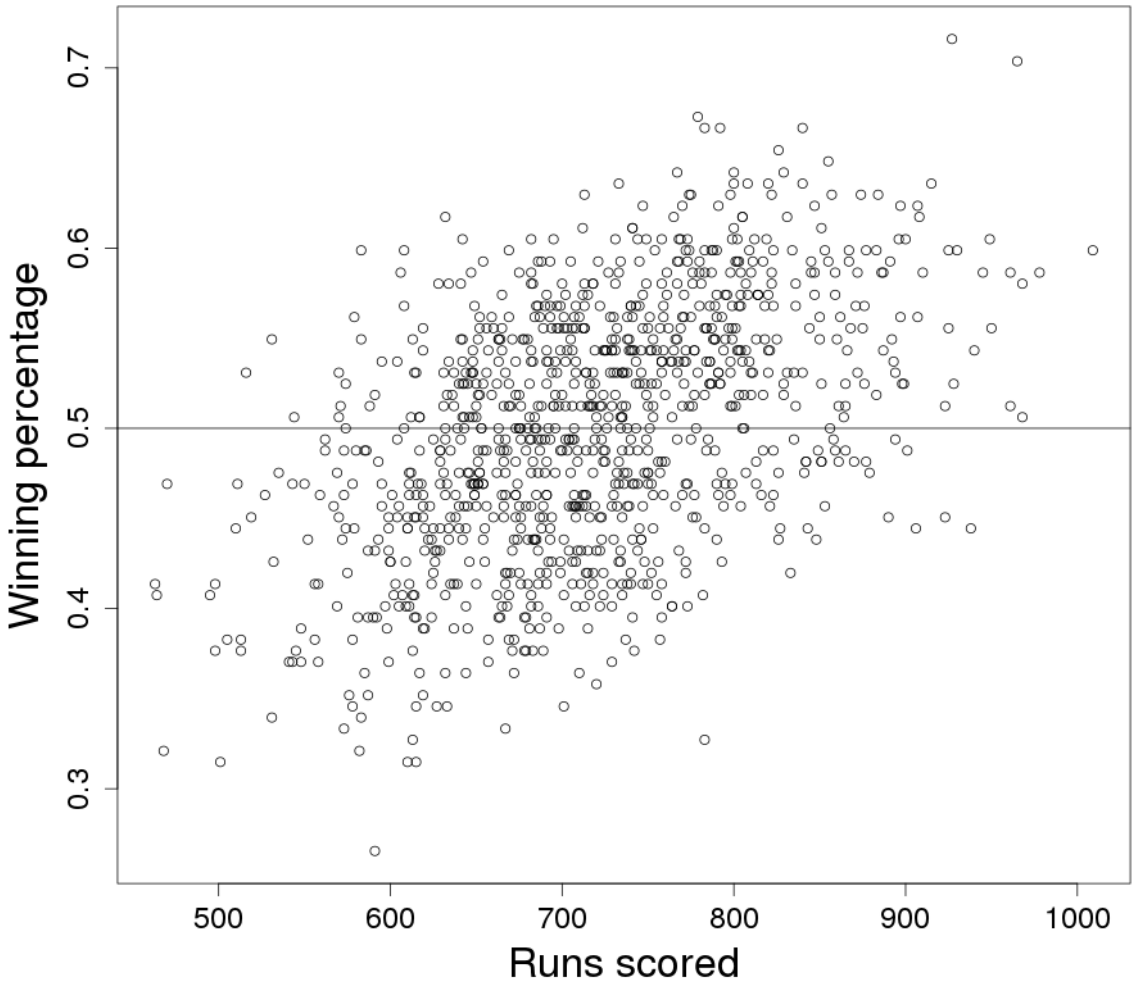
Do the points show a clear trend

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

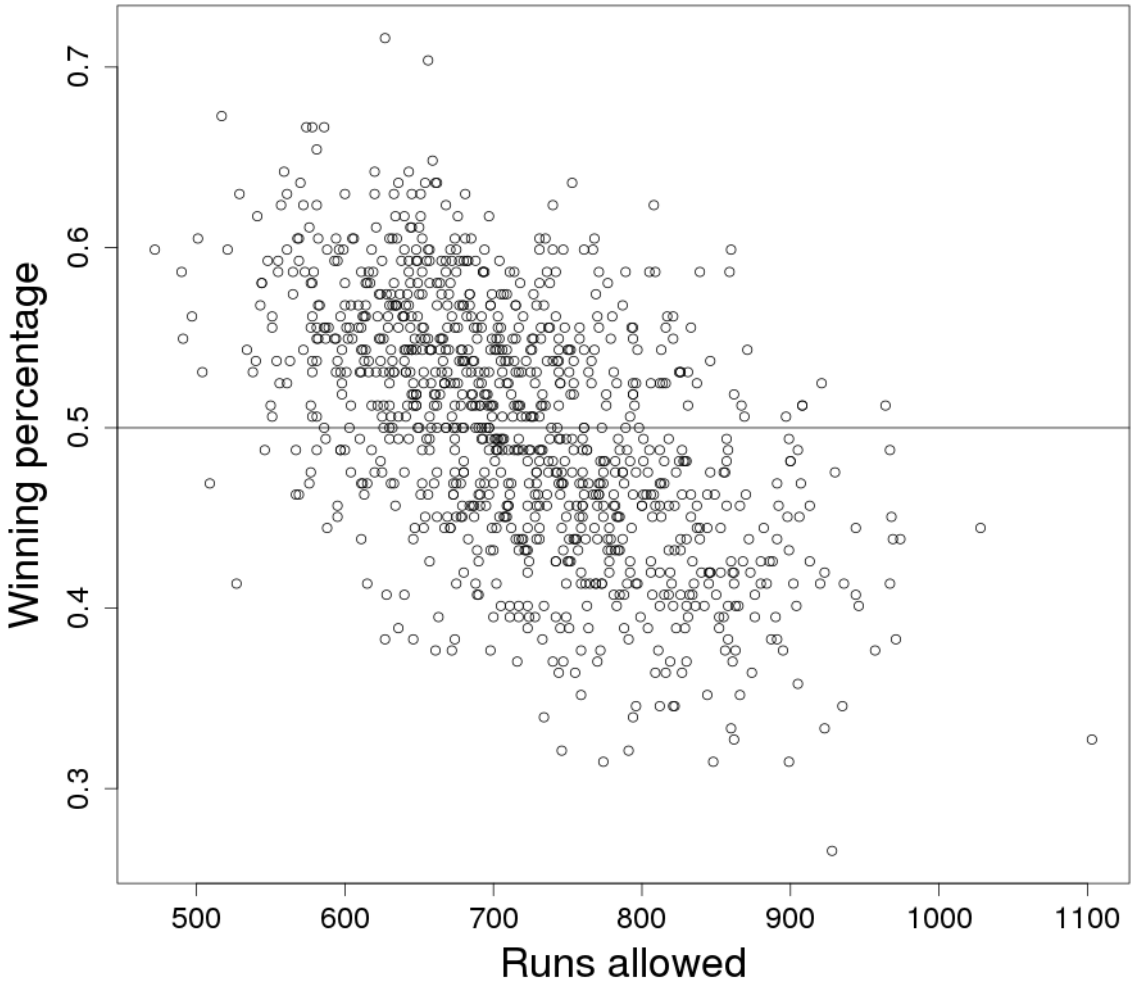
Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

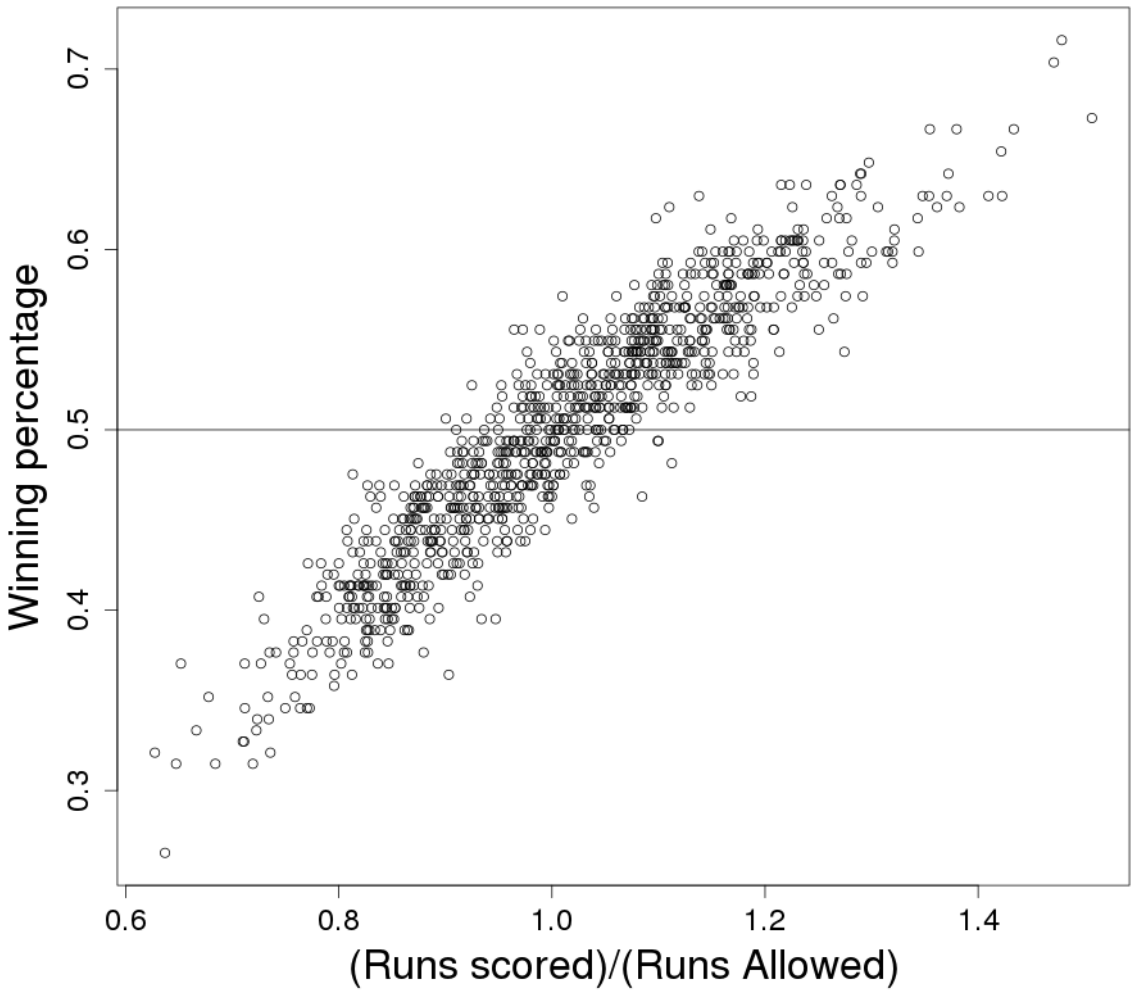
Relationship between runs scored and wins (each point is one team from one season)



Relationship between runs allowed and wins



Relationship between (runs scored)/(runs allowed) and wins



Correlation

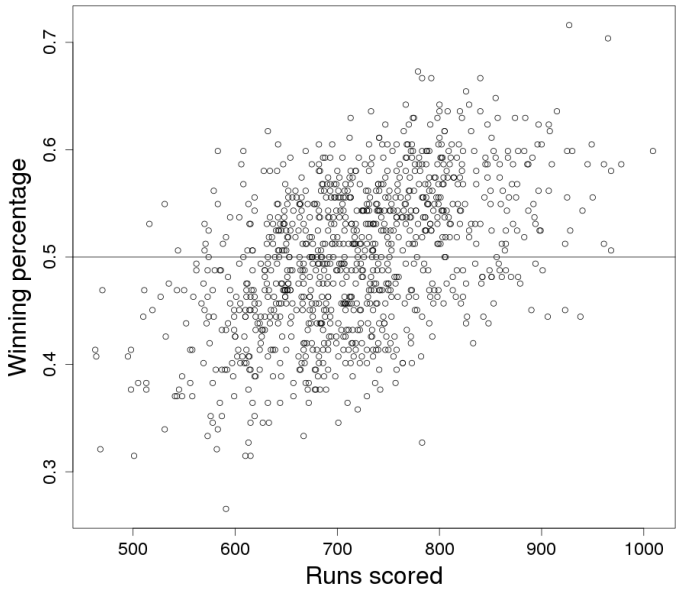
The **correlation** is measure of the strength and direction of a linear association between two variables.

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

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

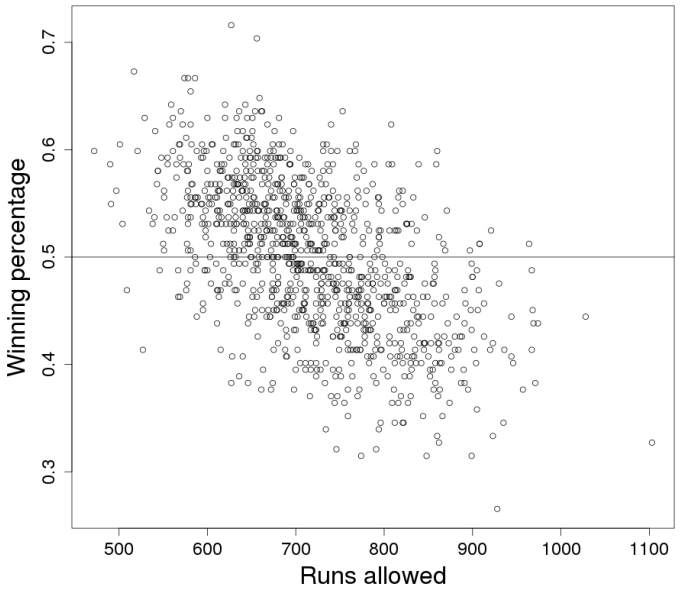
Correlation between runs and wins

Runs scored and wins



r = .50

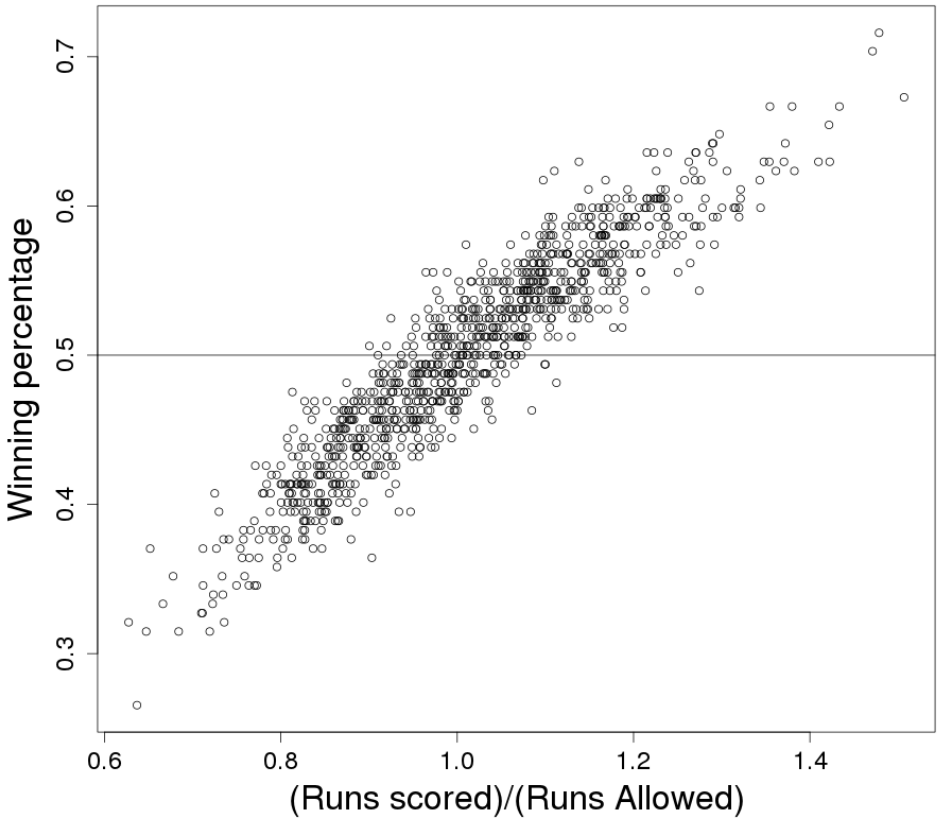
Runs allowed and wins



r = -.55

Correlation between runs and wins

(runs scored)/(runs allowed) and wins



r = .93

Properties of the correlation

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

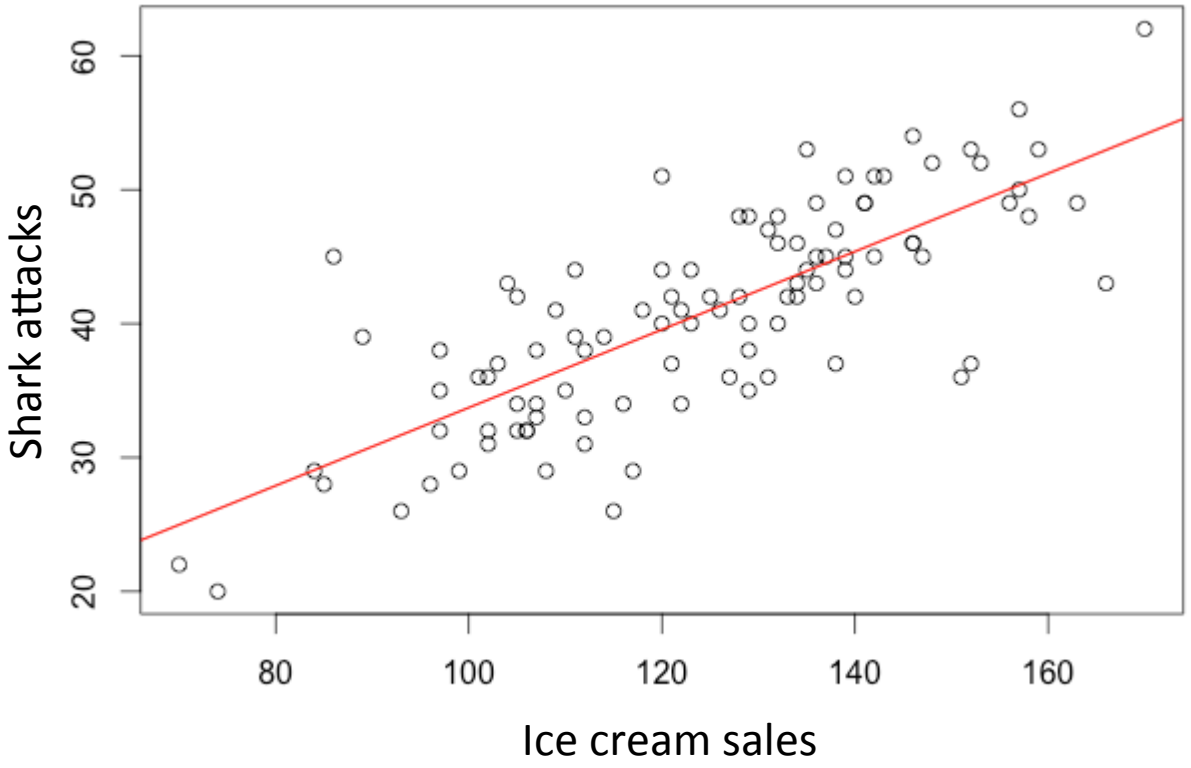
The sign of r indicates the direction of the association

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

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

Correlation caution #1

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



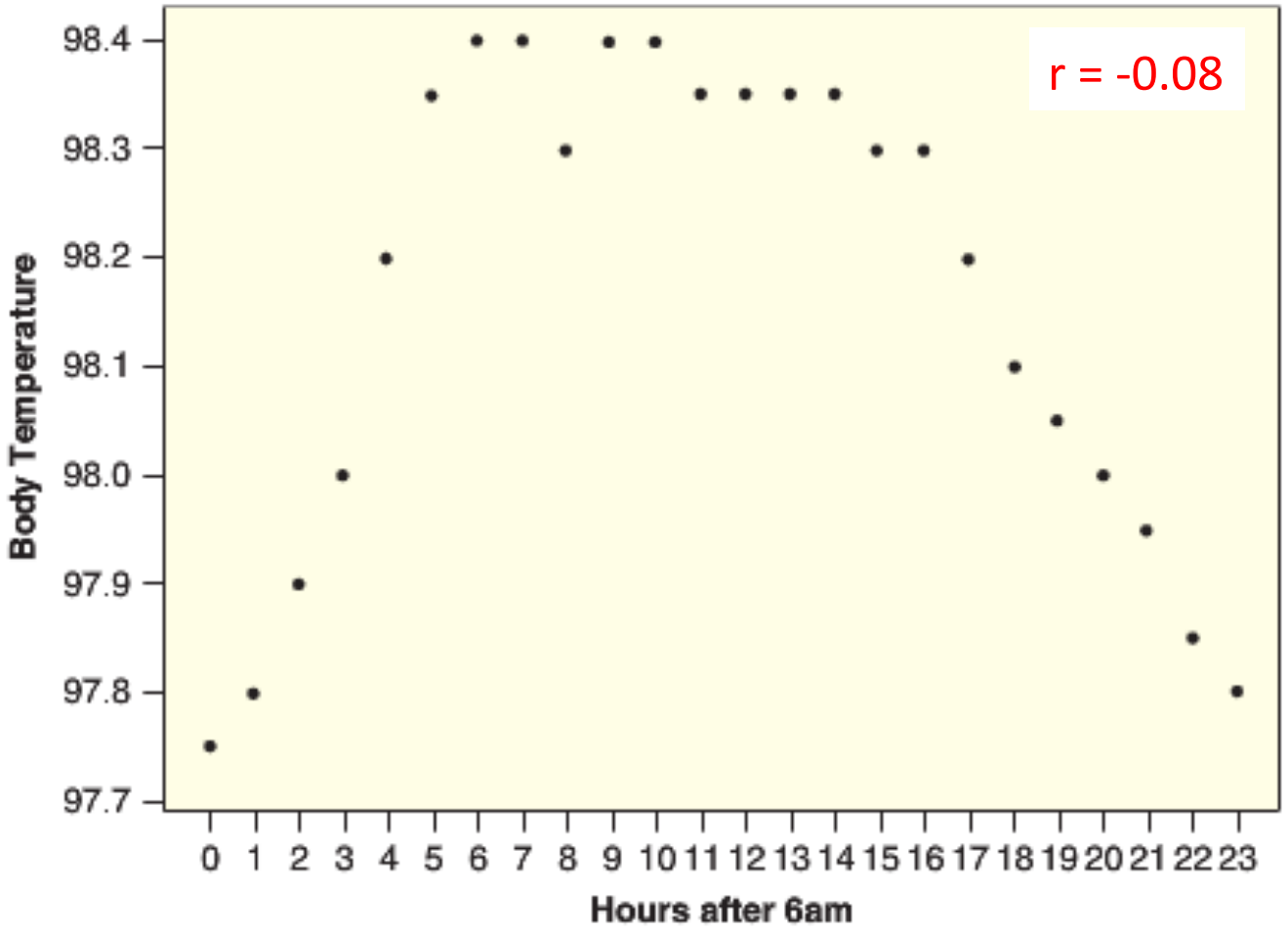
Correlation caution #1

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

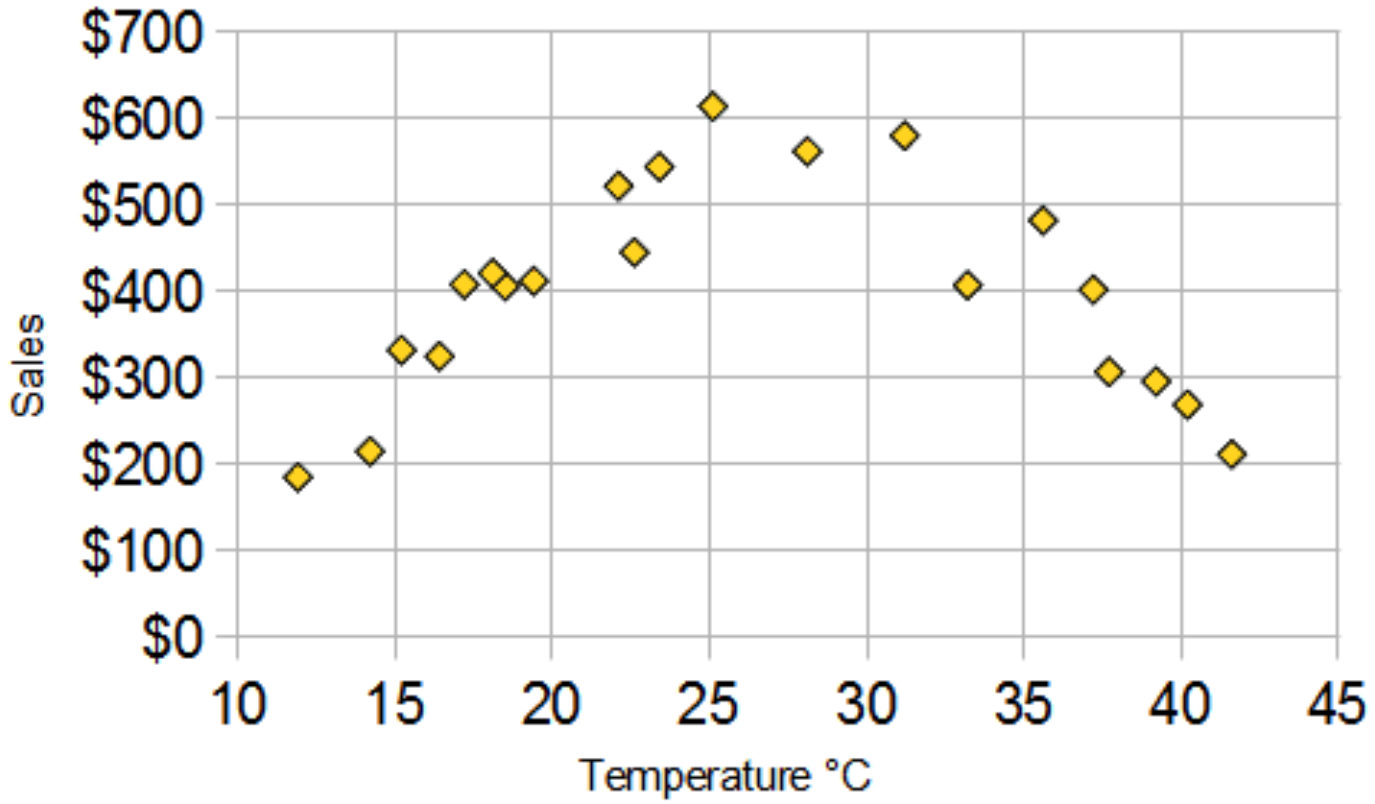
Correlation caution #2

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

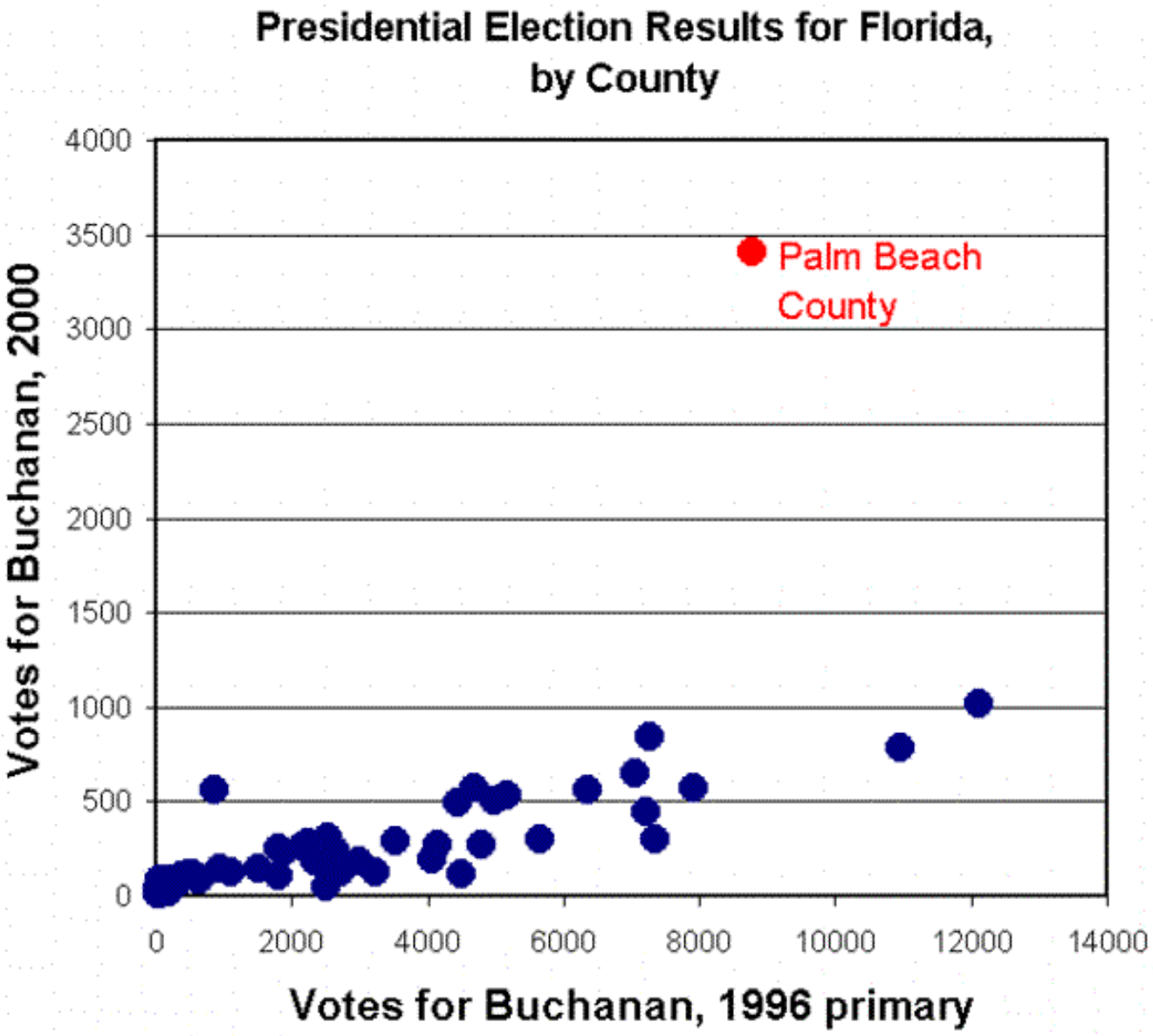


Ice cream sales and temperature



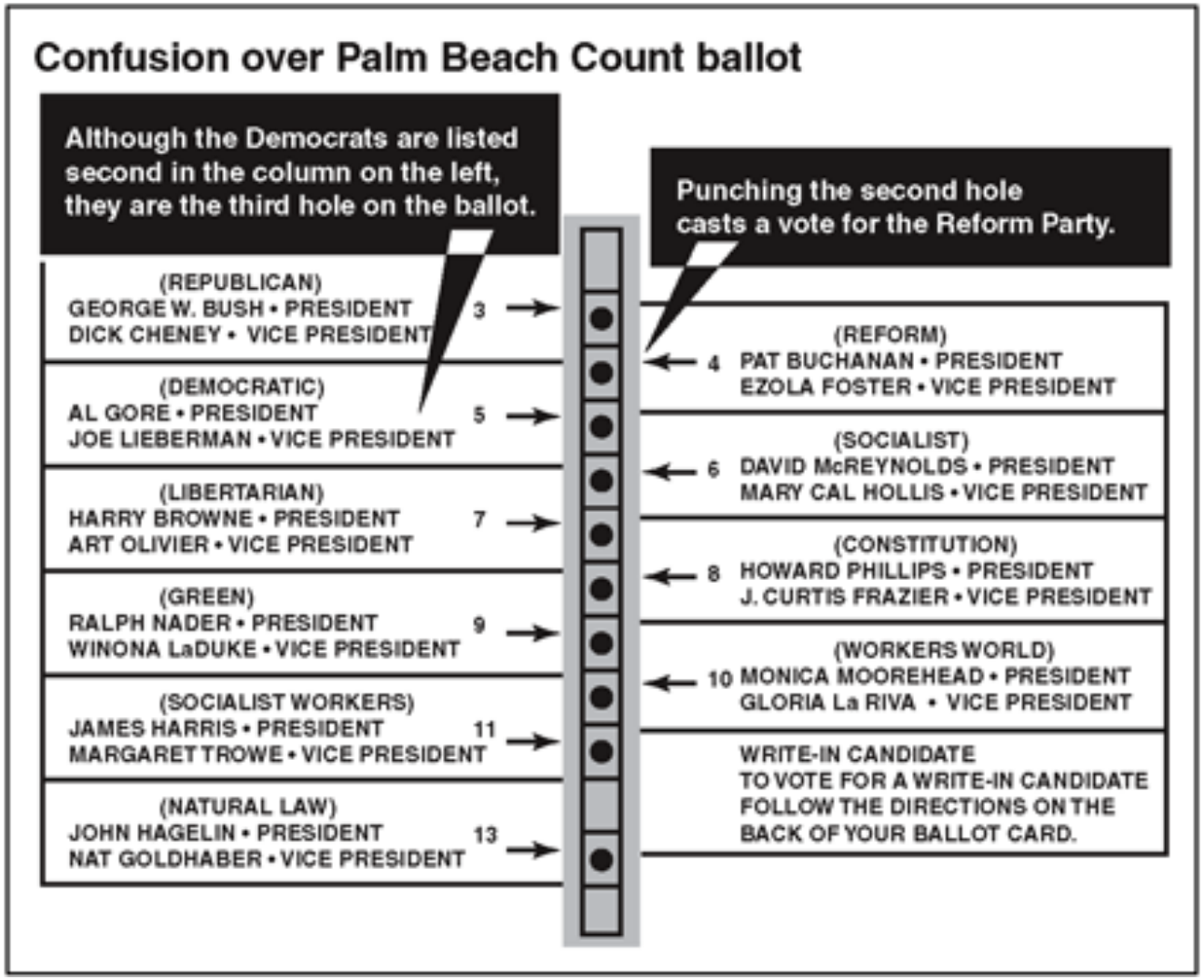
Correlation caution #3

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

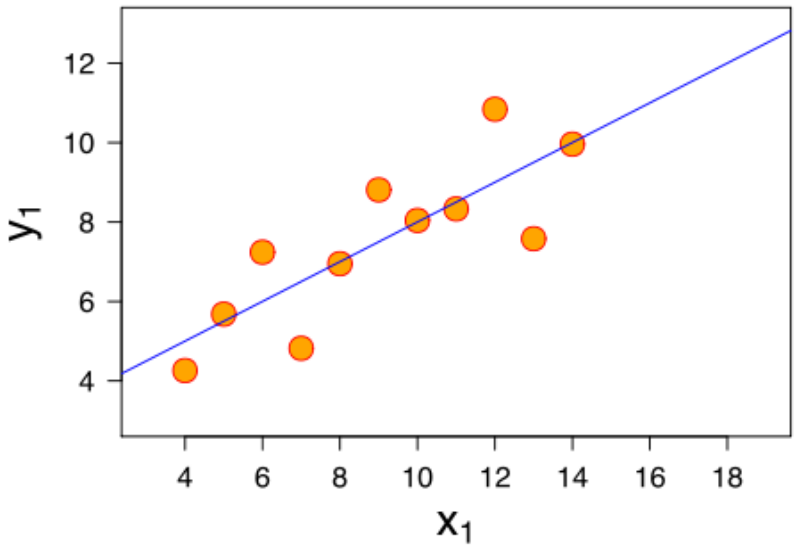


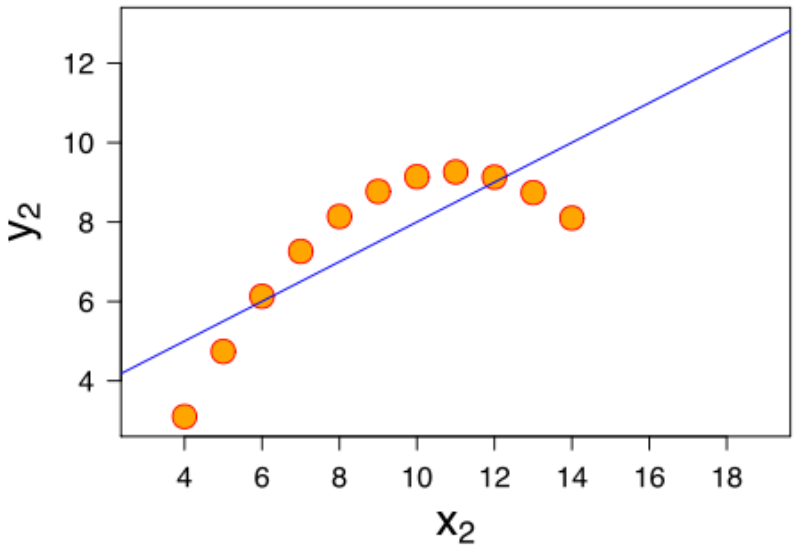
Prof. Greg Adams (Carnegie Mellon Univ.) *Prof. Chris Fastnow (Chatham College)*

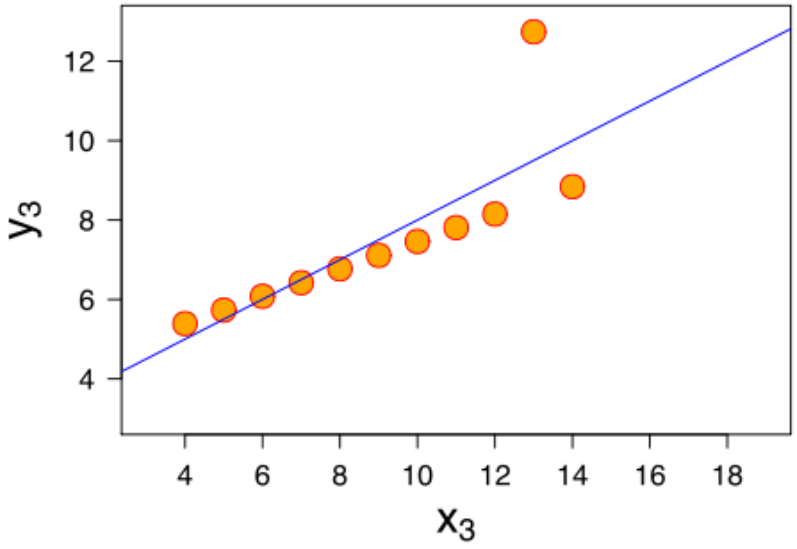
The butterfly ballot

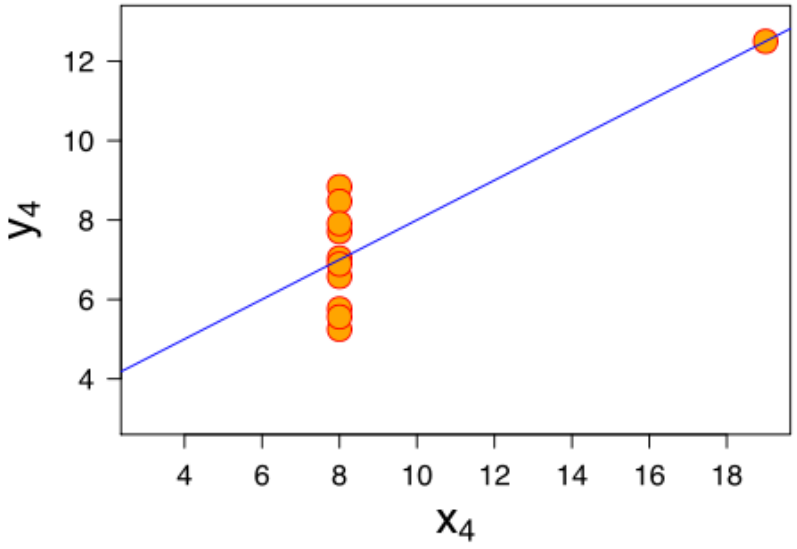


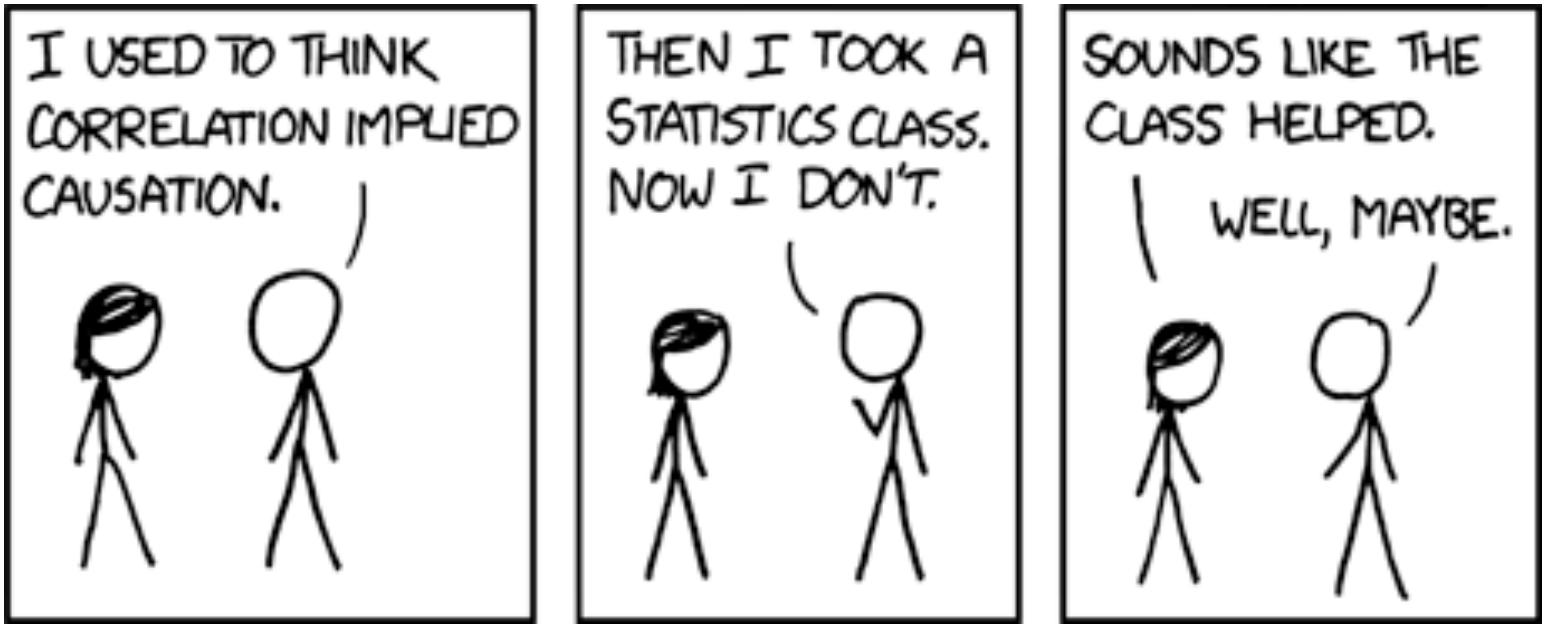
Anscombe's quartet (r = 0.81)










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

Let's start on worksheet 3!

A. Accessing R with the Hampshire server:
https://asterius.hampshire.edu

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

```
> get.worksheet(3)
```

You might also have to type:

```
> source('/home/shared/baseball_stats/baseball_class_functions.R')
```

Let's start on worksheet 3!

```
# Getting David Ortiz data
Ortiz.data <- get.Lahman.batting.data("David", "Ortiz")

# David Ortiz data from his 7th season
Ortiz.data[7, ]

# The actual year of David Ortiz's 7th season
Ortiz.year7  <-  Ortiz.data[7, ]$yearID

# all players from David Ortiz's 7th season who had over 500 PA
PA500.Ortiz7th <- get.Lahman.batting.data(year = Ortiz.year7, min.PA = 500)
```