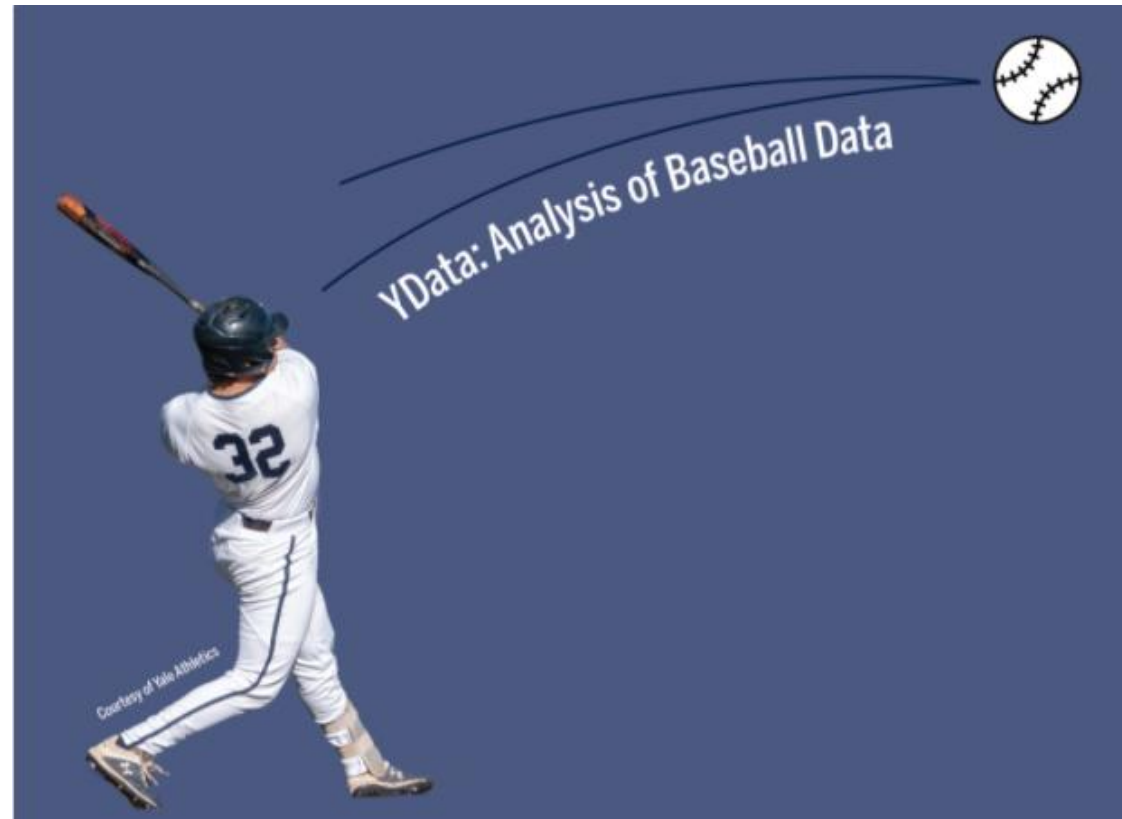


Confidence intervals continued and relationships between measures



Overview

Discussion of chapter 7 of Astroball

Calculating confidence intervals with the bootstrap

- Intuitions on why the bootstrap works
- Computing bootstrap confidence intervals in Python

Relationship between two quantitative variables

Logistics

Has everyone started thinking about the final project?

- Lab 9 will have a question checking in to see how it is going

I'm trying to decide on a class time for next year. Which time do you think would be best for students?

- Same as this year
- Monday 1:30-3:20pm
- Tuesday or Thursday 9:25-11:15
- Later in the evening, say Tuesday from 7-8:50pm

Events that occur in baseball at a rate of 1 out of 600,000?

Lisa Falkenberg

Editor of Opinion, Houston Chronicle








About this Author

Lisa Falkenberg is the Chronicle's vice president/editor of opinion. A Pulitzer Prize-winning journalist with more than 20 years' experience, Falkenberg leads the editorial board and the paper's opinion and outlook sections, including letters, op-eds and Gray Matters.

Falkenberg wrote a metro column at the Chronicle for more than a decade that explored a range of topics, including education, criminal justice and state, local and national politics. In 2015, Falkenberg was awarded the Pulitzer for commentary, as well as the American Society of News Editors' Mike Royko Award for Commentary/Column Writing for a series that exposed a wrongful conviction in a death case and led Texas lawmakers to reform the grand jury system. She was a Pulitzer finalist in 2014.

The MLB season is in week 2...

Zack Greinke eephus pitch

American League				National League			
AL East							
Team	W	L	Pct	GB	Home	Away	L10
 Red Sox	7	3	.700	-	3-3	4-0	7-3
 Orioles	5	6	.455	2.5	1-4	4-2	4-6
 Blue Jays	5	6	.455	2.5	2-3	3-3	4-6
 Yankees	5	6	.455	2.5	3-3	2-3	5-5
 Rays	5	6	.455	2.5	3-2	2-4	4-6

Astroball discussion

Let's discuss the chapter for 7 minutes in breakout rooms and then have a larger conversation as a group

- Discuss your quote and reaction to chapter 7

Thoughts on the chapter 7 of Astrobball?

I couldn't find the paper "In Search of David Ross" presented at the Sloan Conference

I did find another paper by the same authors:

Journal of Sports Analytics 5 (2019) 247–279
DOI 10.3233/JSA-190248
IOS Press

Uncovering the sources of team synergy: Player complementarities in the production of wins

Scott A. Brave^{a,*}, R. Andrew Butters^b and Kevin A. Roberts^c

^aFederal Reserve Bank of Chicago, Economic Research, IL, Chicago

^bBusiness Economics and Public Policy, Kelley School of Business, Indiana University, IN, Bloomington

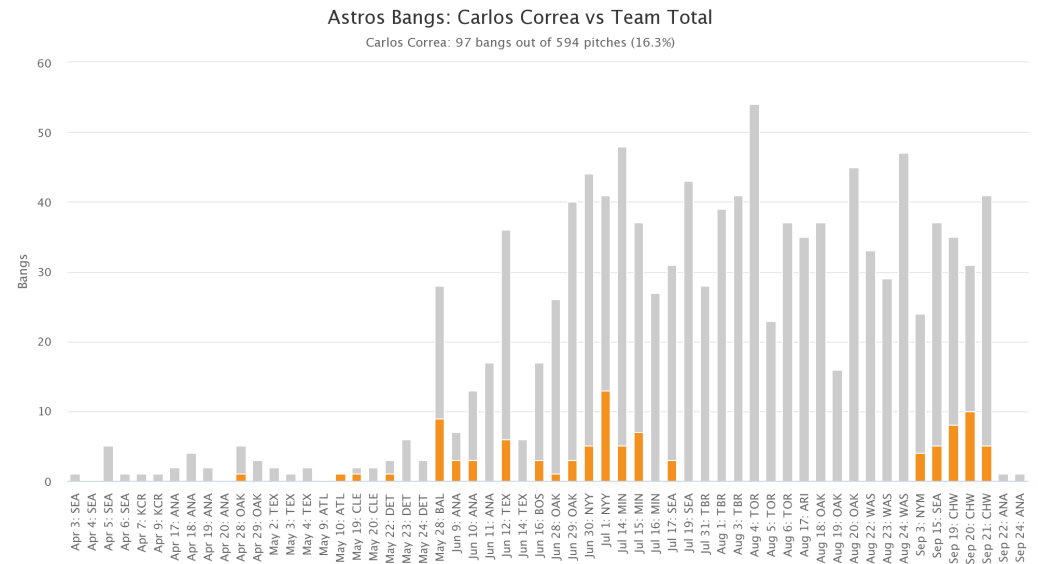
^cDepartment of Economics, Duke University, NC, Durham

$$\hat{\varepsilon}_{int} = \underbrace{w_{ii} f_{it} \lambda_n}_{\text{"Measurement Error"}} + \underbrace{\sum_{i:j \neq i} w_{ij} f_{jt} \lambda_n}_{\text{"Team Synergy"}}$$

$$\begin{aligned} \log(\text{salary}_{it}) &= \sum_c \gamma_c (FA_{cit} \sum_{n=1}^{t-1} WAR_{in}) + \sum_c \beta_c (FA_{cit} \sum_{n=1}^{t-1} pcWAR_{in}) + \\ &\sum_c \theta_c FA_{cit} + \sum_c \theta_c (FA_{cit} * teamExp_{it-1}) + \sum_p \rho_p pos_{pit} + \\ &\sum_p \phi_p (pos_{pit} * age_{it}) + \sum_p \lambda_p (pos_{pit} * mlbExp_{it-1}) + \\ &\sum_p \tau_p (pos_{pit} * mlbExp_{it-1}^2) + \alpha_i + \varepsilon_{it}, \end{aligned} \quad (17)$$

Thoughts on the chapter 7 of Astrobball?

“Beltran’s impact was impossible for the Nerd Cave to quantify, but Carlos Correa attempted to attach a number to it: seven. Of the 24 home runs he hit in 2017, by the end of the regular season Correa attributed precisely seven to Beltran’s influence, and to Beltran’s showing him how to use video to break down opposing pitchers to a depth Correa had never before imagined, to his identifying their tells.” pg 160



Only Astros 2017 home games with video available were logged (58 games).
The commissioner stated the Astros used other methods besides banging a trash can to indicate pitches. This analysis only looked at the banging.
SignStealingScandal.com

Are these home runs related to trashcan bangs and can we find these home runs?

Thoughts on the chapter 7 of Astroball?

Burial of Beltran's glove



ESPN: Is any team having more fun than the Houston Astros?

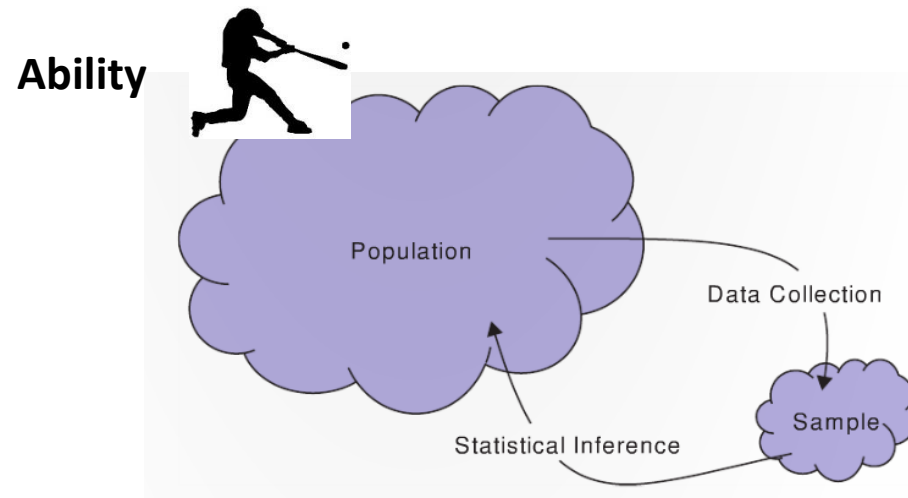
Review: statistical inference

Statistical inference: use sample of data to deduce properties of an underlying population or stochastic process

In the context of baseball this usually means: looking at a player's **performance** to tell something about the player's **ability**

- **Ability:** innate talent
- **Performance:** outcomes from playing a number of games

parameters



statistics

Performance: Hit, Out, Hit, Out, ...

Confidence intervals

Hypothesis tests: we test whether a parameter is equal to a particular value

- E.g., $\pi \neq 300$

Confidence intervals: We create a range of plausible values for a parameter

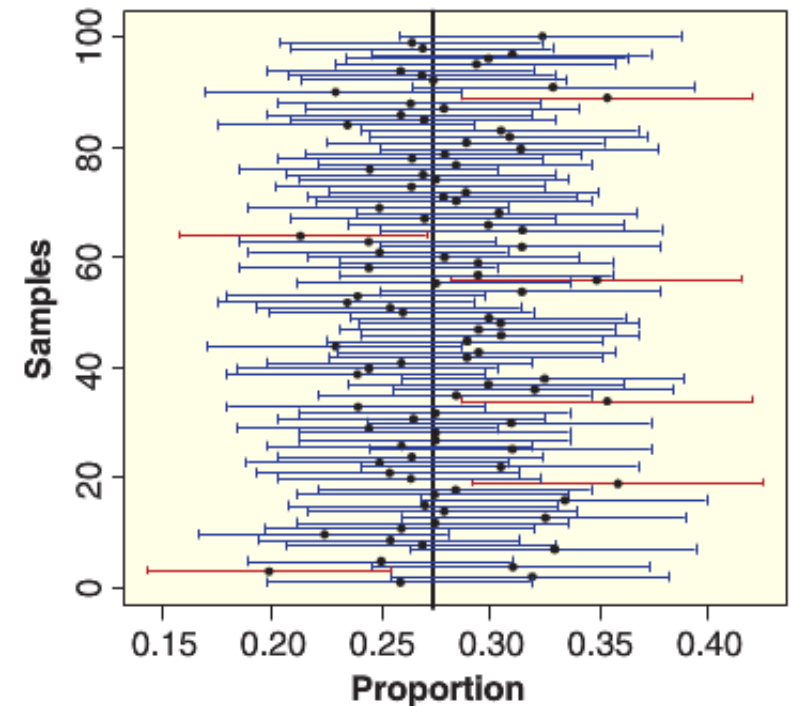
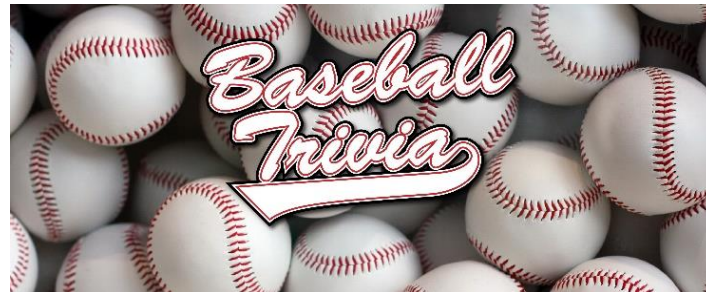
- E.g., π is in $[\text{.320 } \text{.387}]$

Confidence Intervals

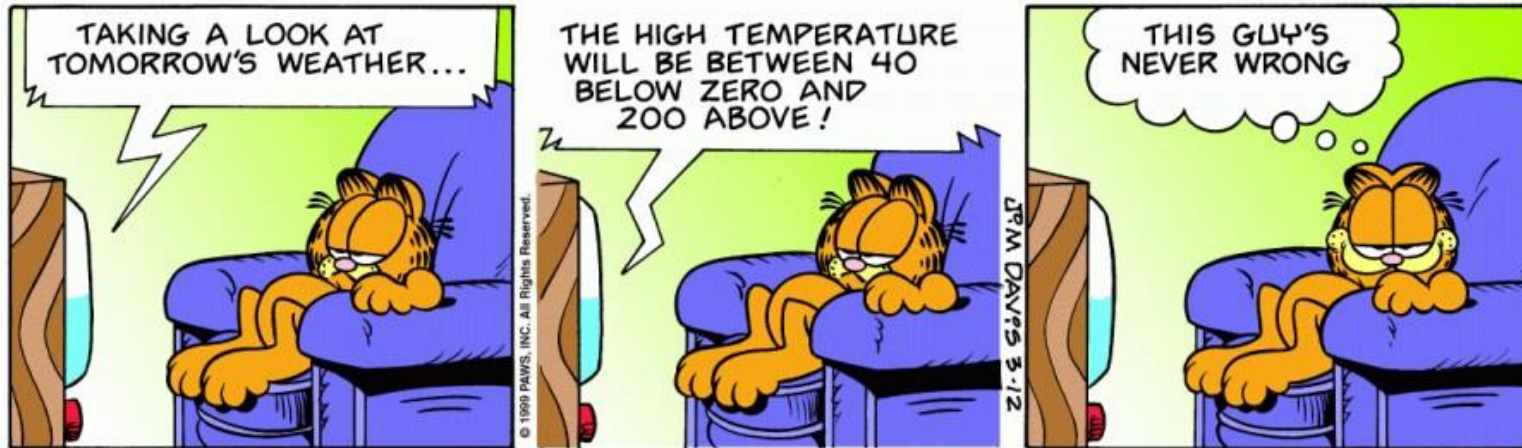
A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

The **confidence level** is the percent of all intervals that contain the parameter

- E.g., for a 90% confidence level, 90% of our CI will contain the parameter



100% confidence intervals



There is a tradeoff between:

- The **confidence level** (percent of times we capture the parameter)
- The **confidence interval size**

Note

For any given confidence interval we compute, we don't know whether it has really captured the parameter

But we do know that if we do this 100 times, 95 of these intervals will have the parameter in it

(for a 95% confidence interval)

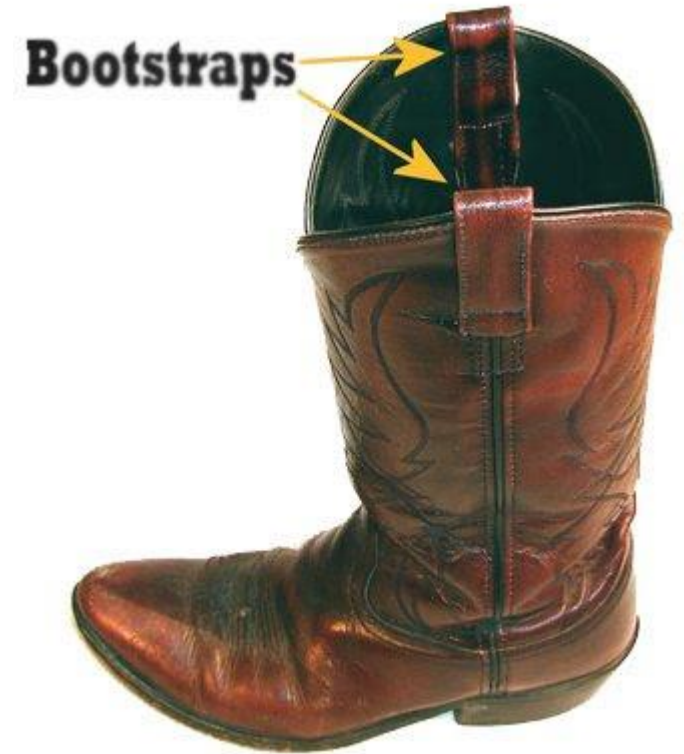
Creating confidence intervals

There are both parametric and computational methods we can use to create confidence intervals.

In this class we are going to focus on a computational method to create confidence intervals called **the bootstrap**

- Take Intro Stats to learn parametric methods!

In order to understand how the bootstrap works, we first need to understand the concept of a **sampling distribution**

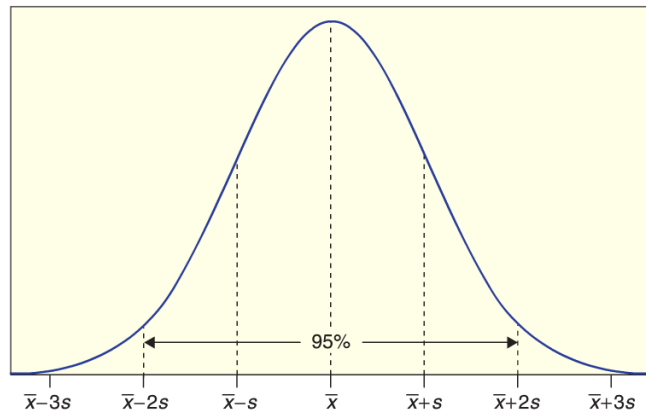
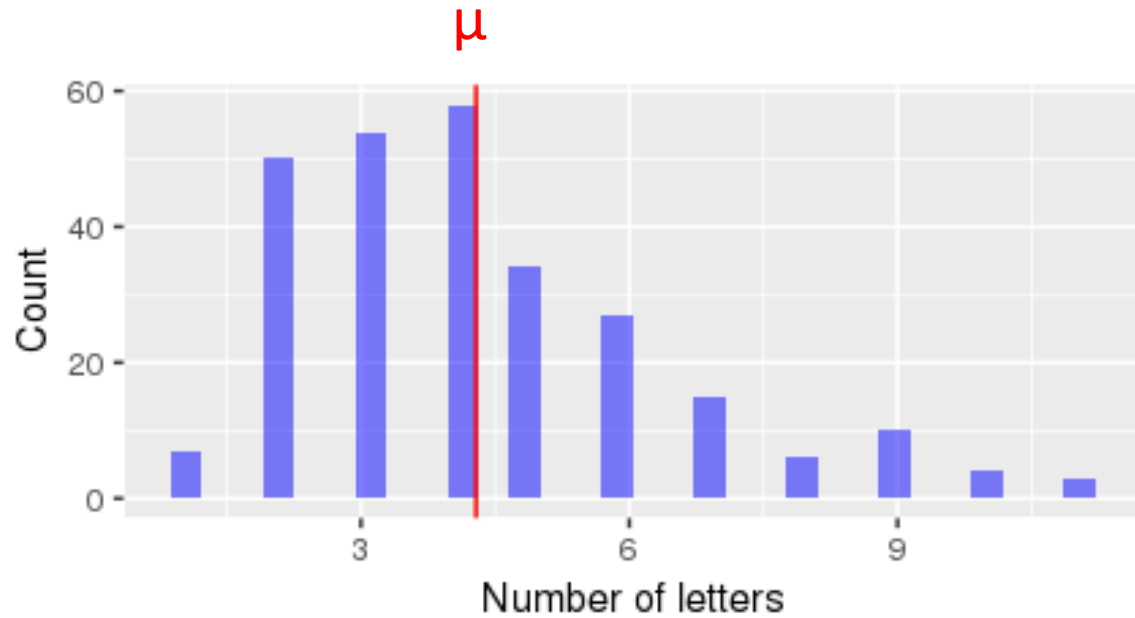


Sampling distributions

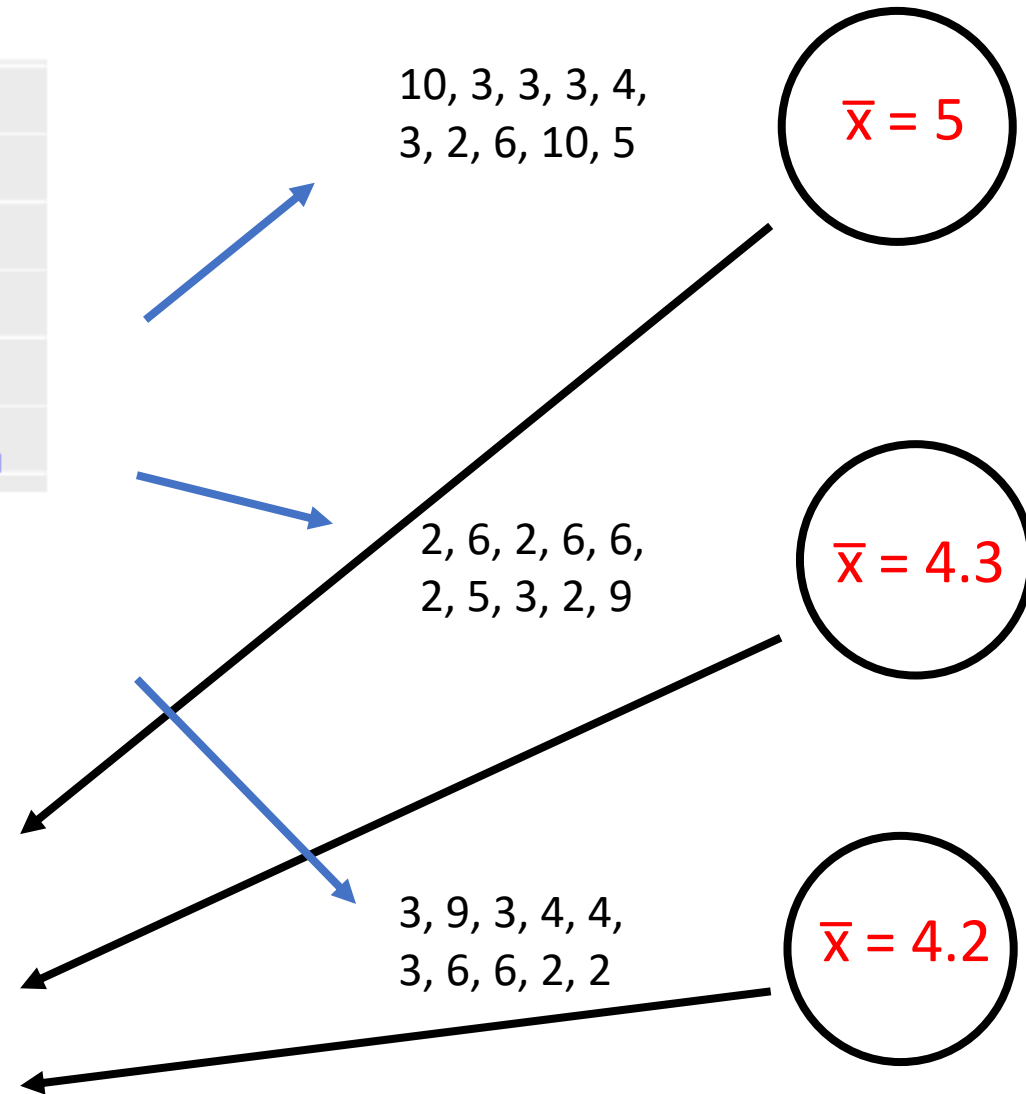
A **sampling distribution** is the distribution of sample statistics computed for different samples of the same size (n) from the same population

A sampling distribution shows us how the sample statistic varies from sample to sample

Sampling distribution



Sampling distribution!

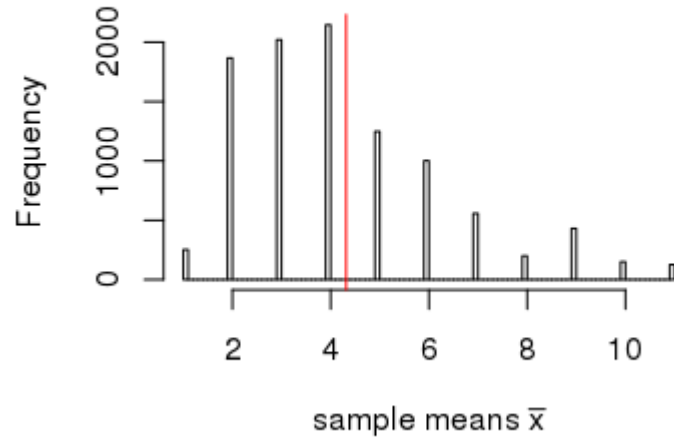


[Sampling distribution app](#)

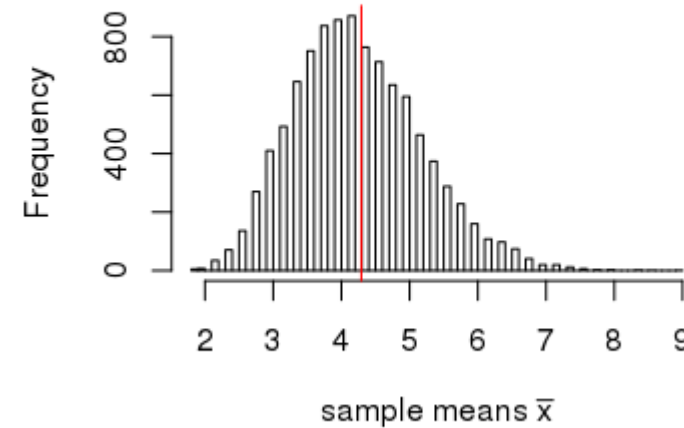
As the sample size n increases

1. The sampling distribution becomes more like a normal distribution
2. The sampling distribution points (\bar{x} 's) become more concentrated around the mean $E[\bar{x}] = \mu$

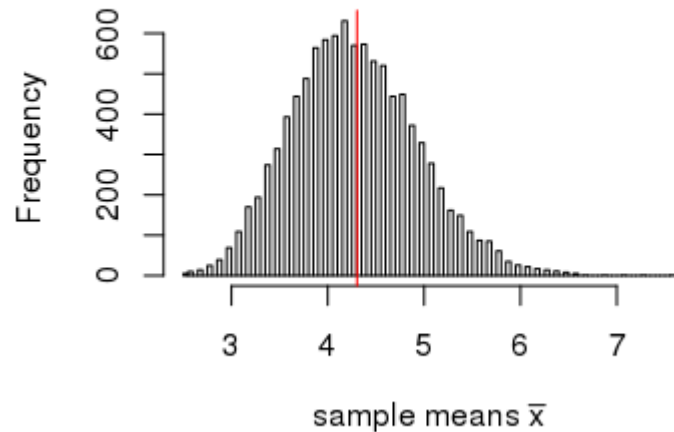
Sampling distribution ($n = 1$)



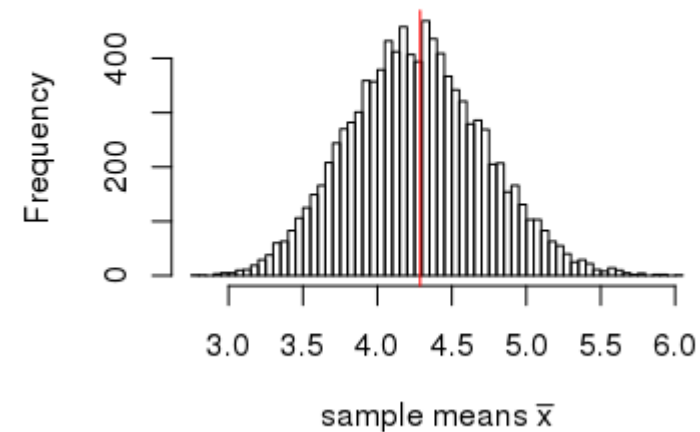
Sampling distribution ($n = 5$)



Sampling distribution ($n = 10$)



Sampling distribution ($n = 20$)

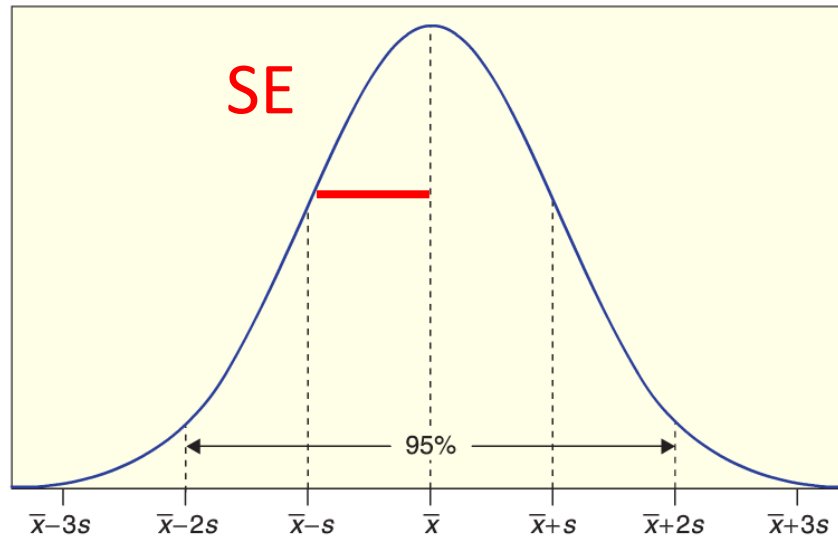


x-axis range 9 vs. 6

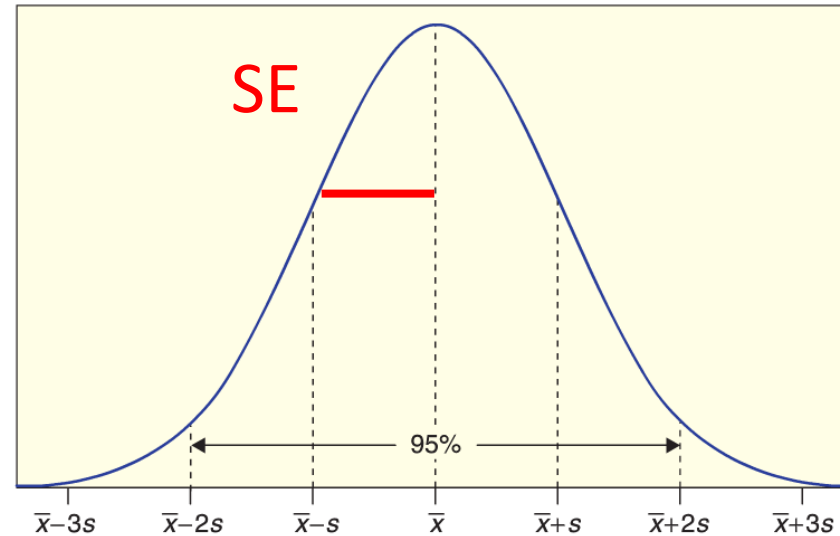
The standard error

The **standard error** of a statistic, denoted SE, is the standard deviation of the sample statistic

- i.e., SE is the standard deviation of the *sampling distribution*



What does the size of a standard error tell us?



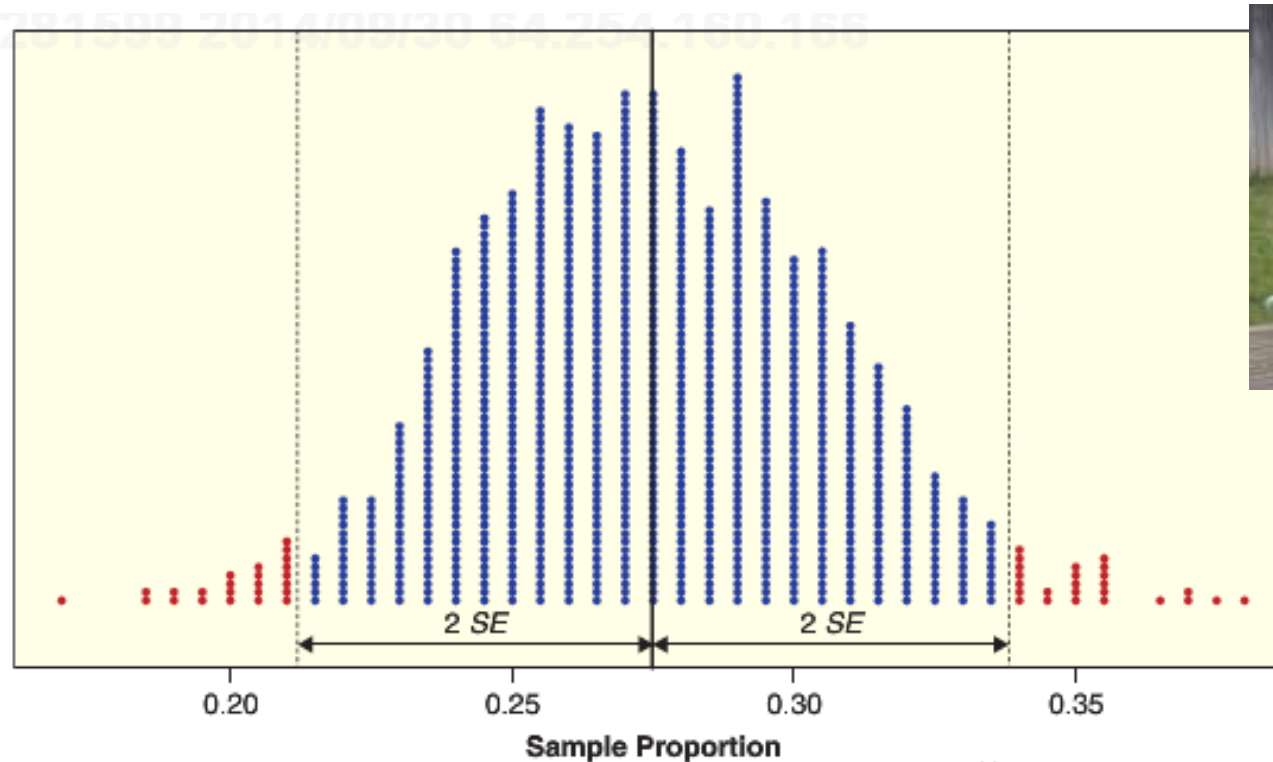
Q: If we have a large SE, would we believe a given statistic is a good estimate for the parameter?

- E.g., would we believe a particular \bar{x} is a good estimate for μ ?

A: A large SE means our statistic (point estimate) could be far from the parameter

- E.g., \bar{x} could be far from μ

By the Central Limit Theorem: sampling distributions are often normal!



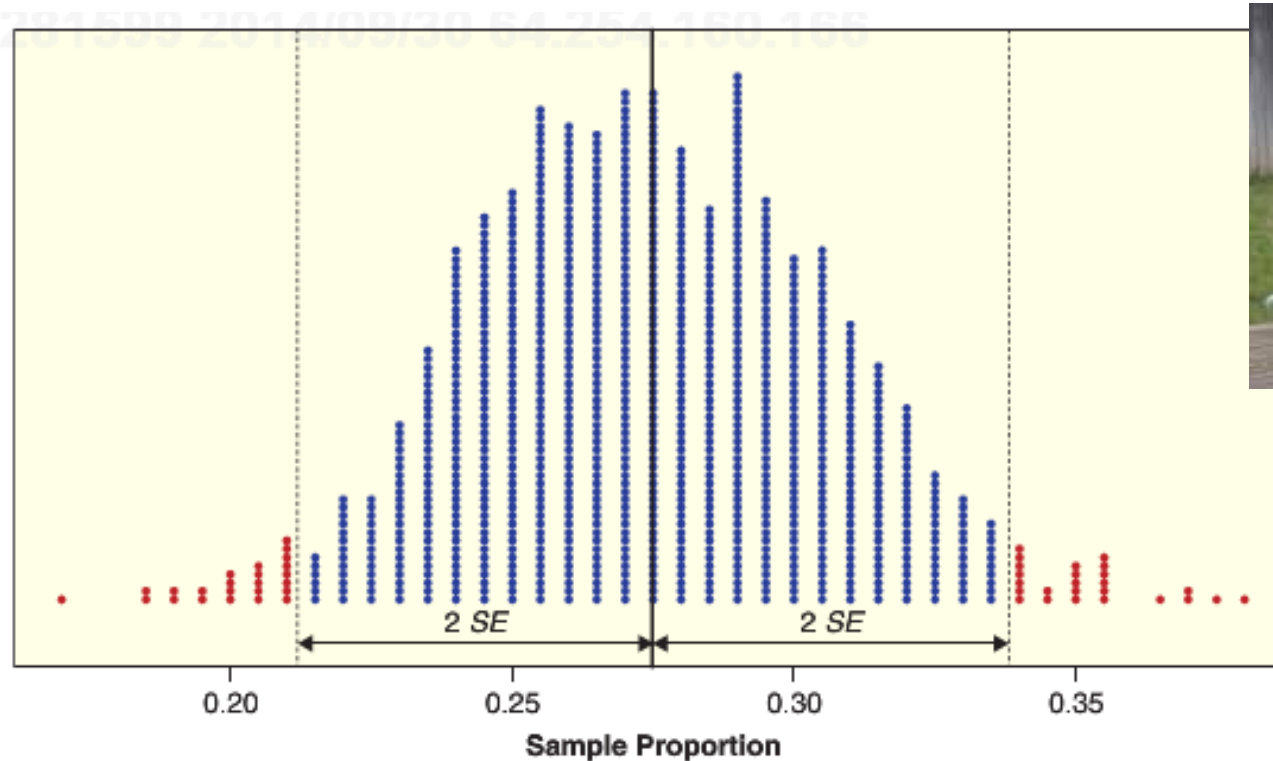
Actually, using 1.96 is more accurate than using 2

So 95% of the sample statistics will fall within ± 2 Standard Errors from the mean

Thus, if we had a statistic value and knew the SE, we could calculate a 95% confidence interval!

- Why is this true?

By the Central Limit Theorem: sampling distributions are often normal!



So 95% of the sample statistics will fall within ± 1.96 Standard Errors from the mean

95% CI is: statistic $\pm 1.96 \cdot SE$

The problem

Unfortunately we don't know the Standard Error ☹️

Q: Could we repeat the sampling process many times to estimate it?

A: No, we can't do an experiment that many times

We're just going to have to pick ourselves up from the bootstraps!

Estimate SE with \hat{SE}

Then use $\bar{x} \pm 1.96 \cdot \hat{SE}$ to get the 95% CI

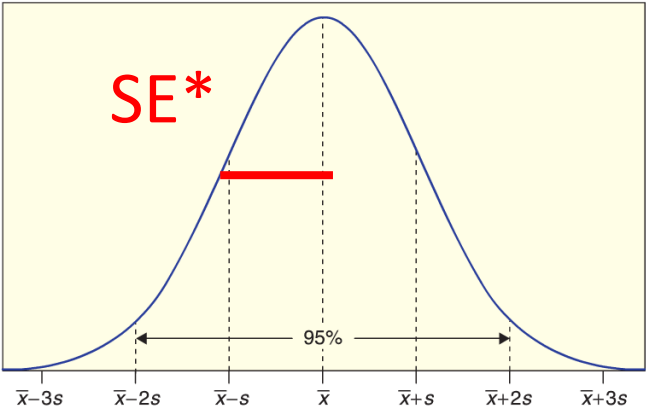
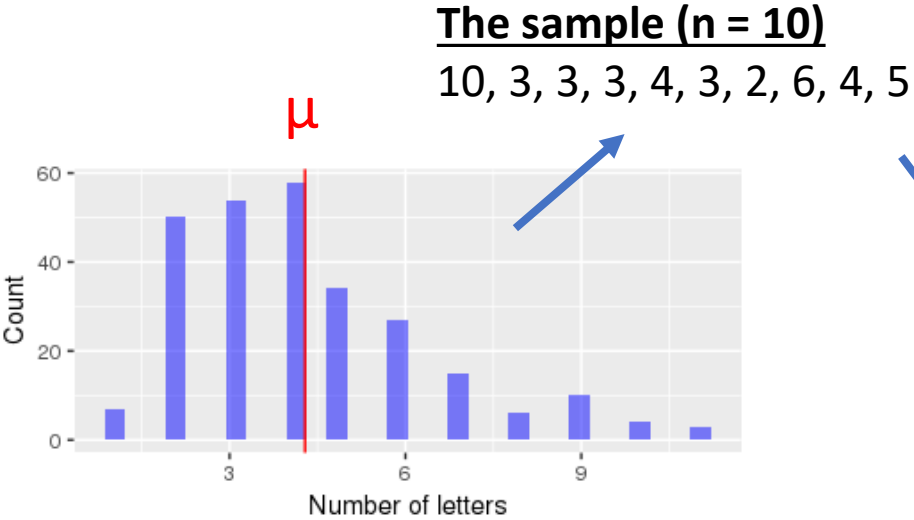
The bootstrap “plug-in principle”

Suppose we get a sample from a population of size n

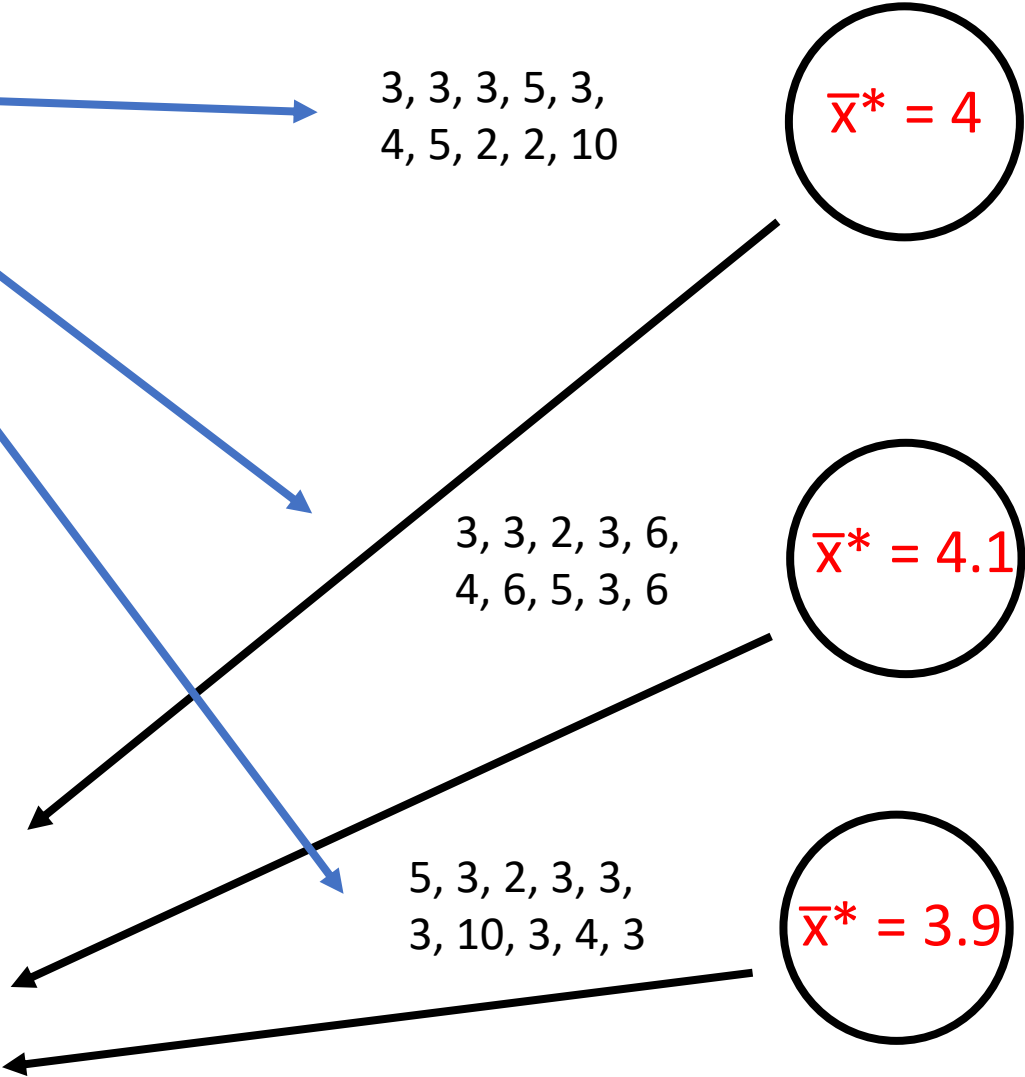
We pretend that this sample is the population (plug-in principle)

1. We then sample n points with replacement from our sample, and compute our statistic of interest
2. We repeat this process 1000's of times and get a *bootstrap* sample distribution
3. The standard deviation of this bootstrap distribution (SE* bootstrap) is a good approximate for standard error SE from the real sampling distribution

Bootstrap distribution



Bootstrap distribution!



Notice there is no 9's in the bootstrap samples

Bootstrap process

“Fake sampling distribution”

Original
Sample

95% Confidence Intervals

When a bootstrap distribution for a sample statistic is approximately normal, we can estimate a 95% confidence interval using:

$$\text{Statistic} \pm 1.96 \cdot SE^*$$

Where SE^* is the standard error estimated using the bootstrap

Empirically examining the bootstrap SE^*

Let's empirically examine that the bootstrap SE^* is a reasonable approximation for the true SE using an app

Lab 8 exercises 1 and 2!

Creating confidence intervals in Python

To create confidence intervals in Python, we can:

1. Sampling with replace from our original sample to create a bootstrap sample
2. Calculating a bootstrap statistic from this bootstrap sample
3. Repeating this process many times to get a bootstrap distribution

We can then get confidence intervals by:

1. Estimating the bootstrap standard error SE^* and then using the formula for a 95% confidence interval: $\text{stat} \pm 1.96 \cdot SE^*$
2. Taking percentiles from the bootstrap distribution (i.e., the 2.5 and 97.5 percentile)

Calculating a plausible range of A-Rod's OBP ability

For exercise 3, you will calculate a plausible range of values (i.e., a confidence interval) for A-Rod's ability π

Because using the bootstrap requires resampling with replacement from a **data sample**, the first step in doing this will be to create a data table that has the following properties:

1. A single column called ONBASE_EVENT
2. It should have the value **True** for how many times A-Rod got on base in the 2012 season and **False** for how many times he got out

We can then sample with replacement from this table to create bootstrap statistics

Calculating a plausible range of A-Rod's OBP ability

To create this confidence interval, please read chapter 13 of the class textbook

- You will write the `bootstrap_proportion()` function described in this chapter

If there is time at the end of class, you can work on this exercise while I'm on Zoom to get help

Relationships between variables

Do power hitters strike out more often?

Chris Davis in 2013:



53 home runs



199 strike outs

Scatter plots

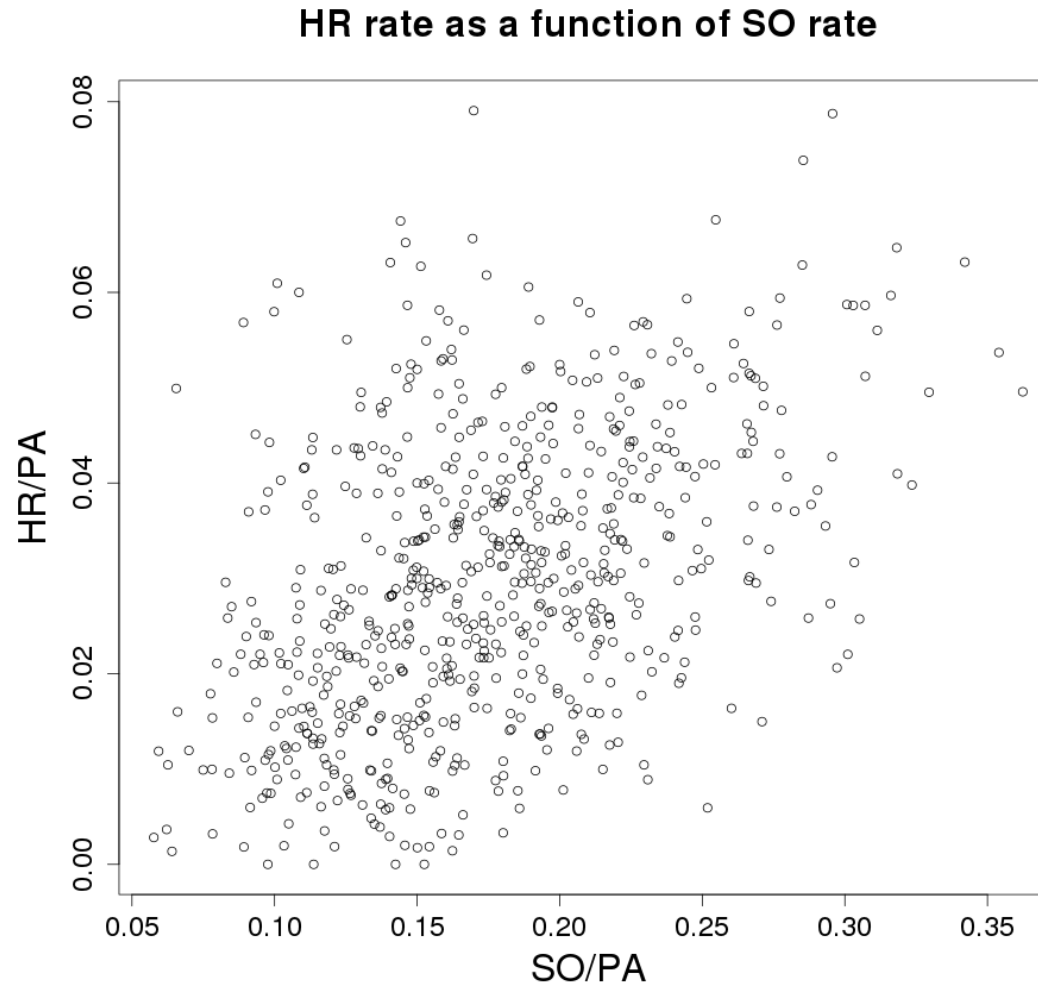
A **scatterplot** graphs the relationship between two variables

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



```
tb.scatter('x_col', 'y_col')
```

Questions when looking at scatter plots

Do the points show a clear trend

Does it go upward or downward?

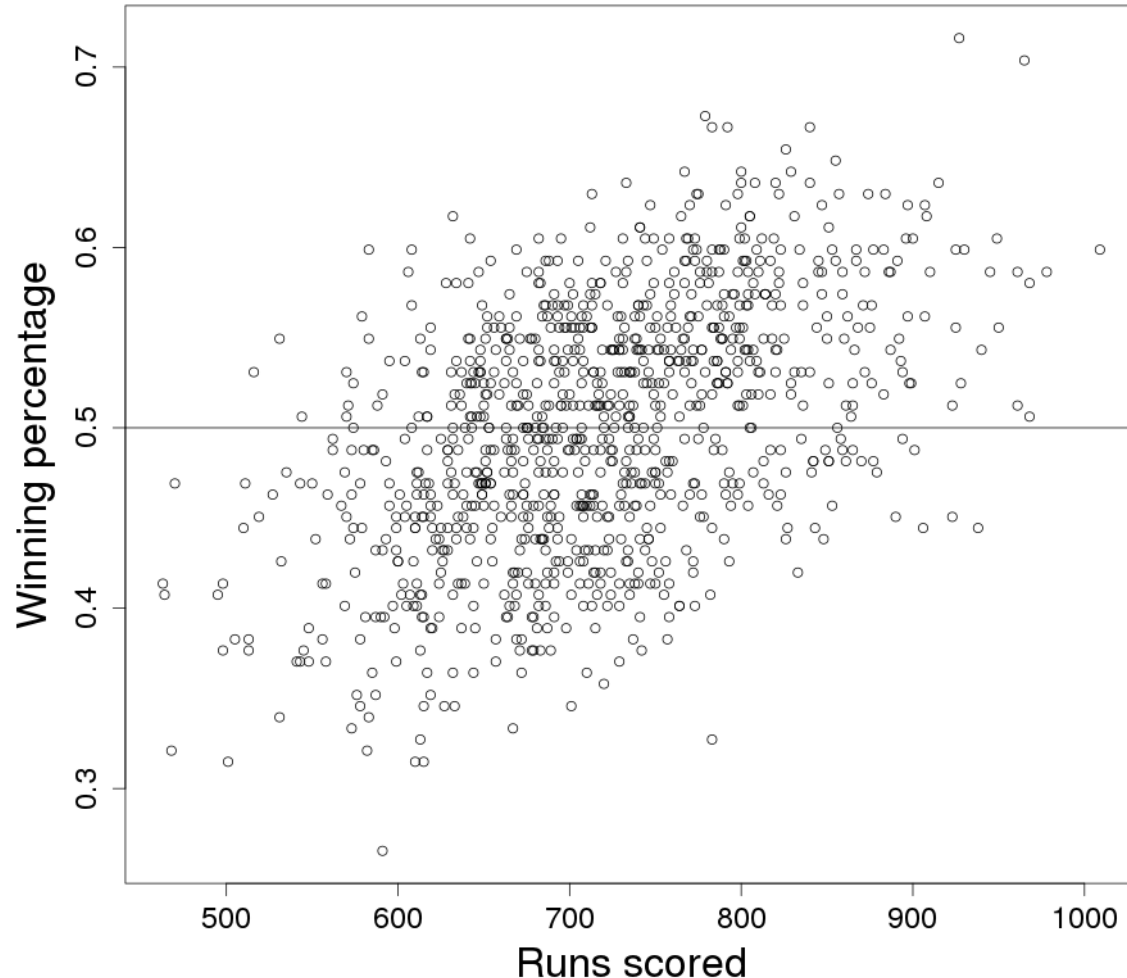
How much scatter around the trend?

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)

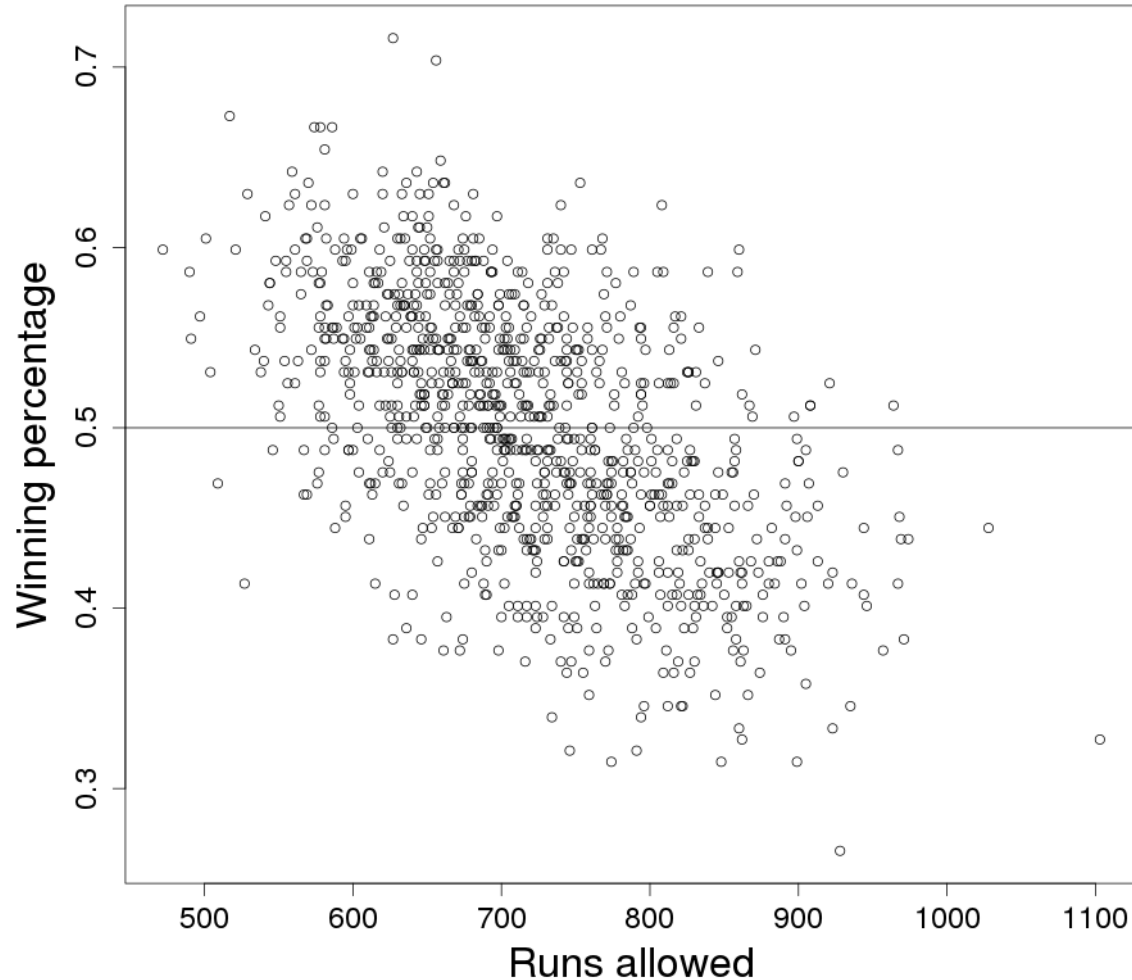


Do the points show a clear trend
Does it go upward or downward?
How much scatter around the trend?

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

Are there any outlier points?

Relationship between runs allowed and wins

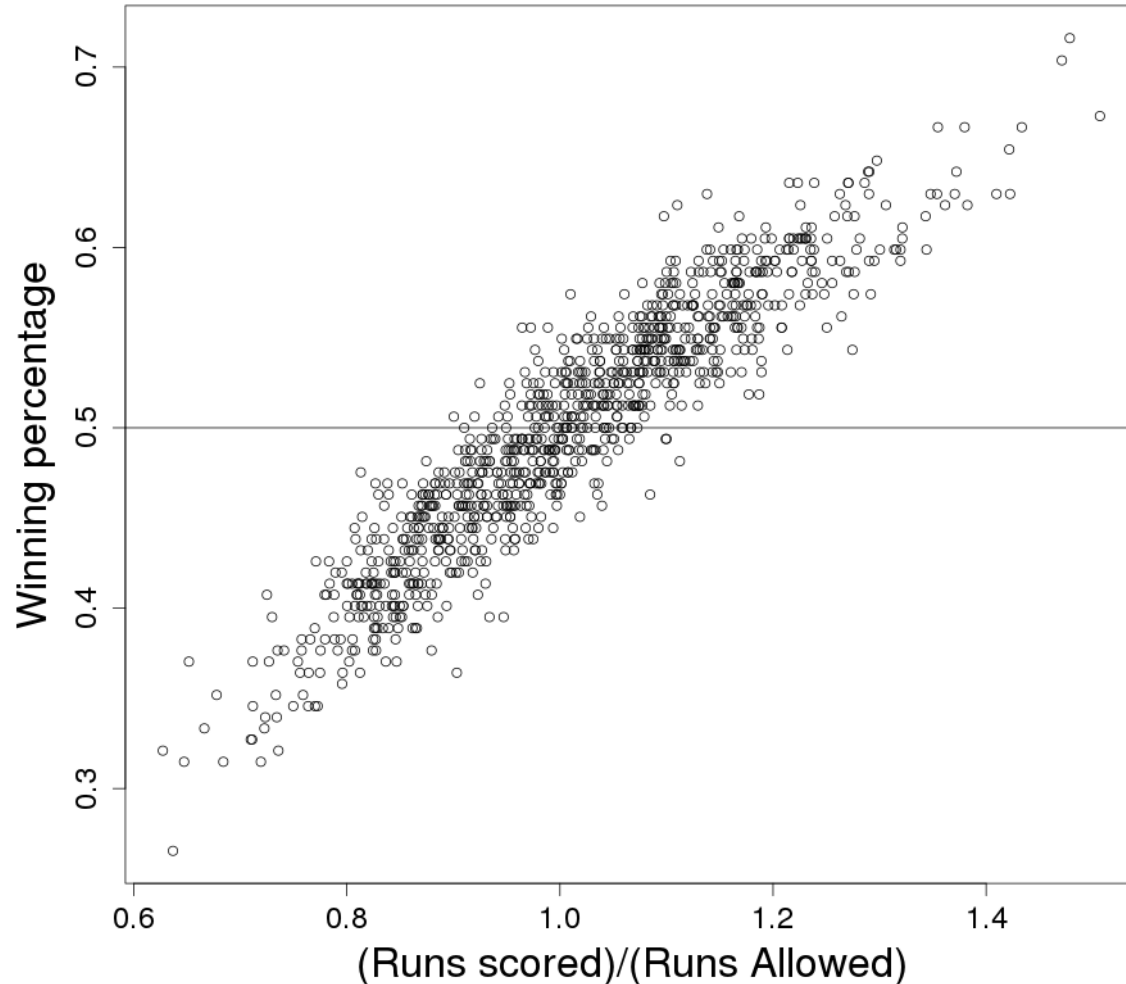


Do the points show a clear trend
Does it go upward or downward?
How much scatter around the trend?

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

Are there any outlier points?

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



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?

Correlation

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

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

Properties of the correlation

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

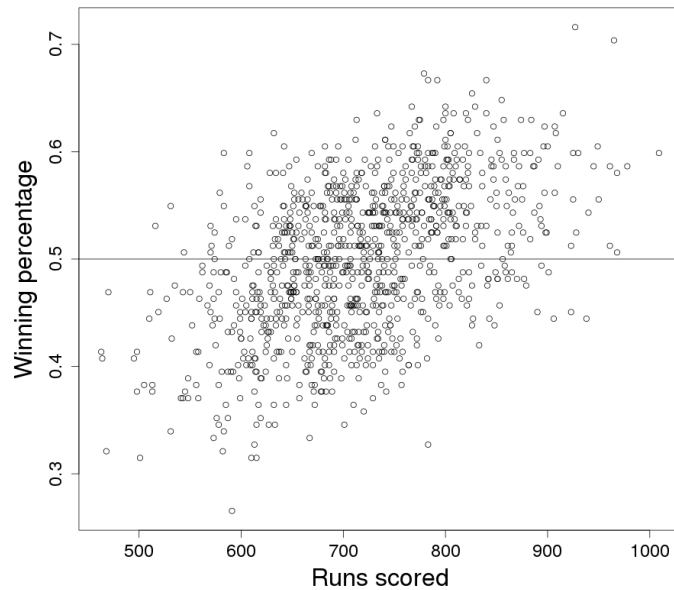
The sign of r indicates the direction of the association

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

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

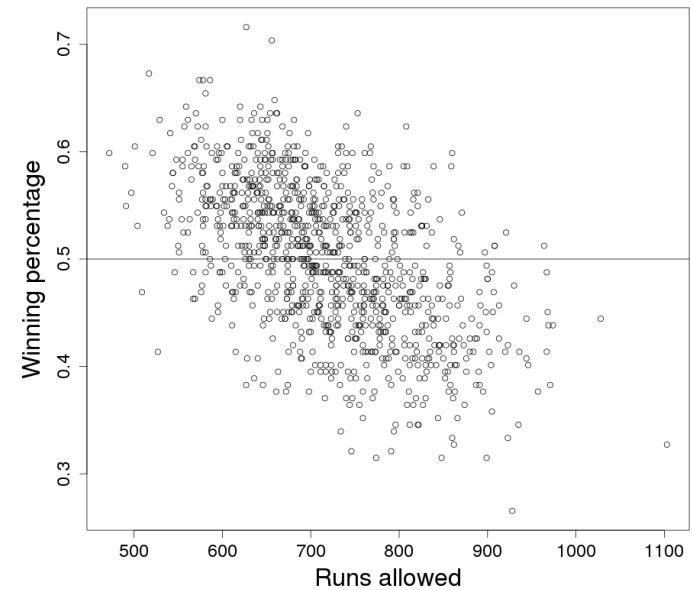
Correlation between runs and wins

Runs scored and wins



$$r = .50$$

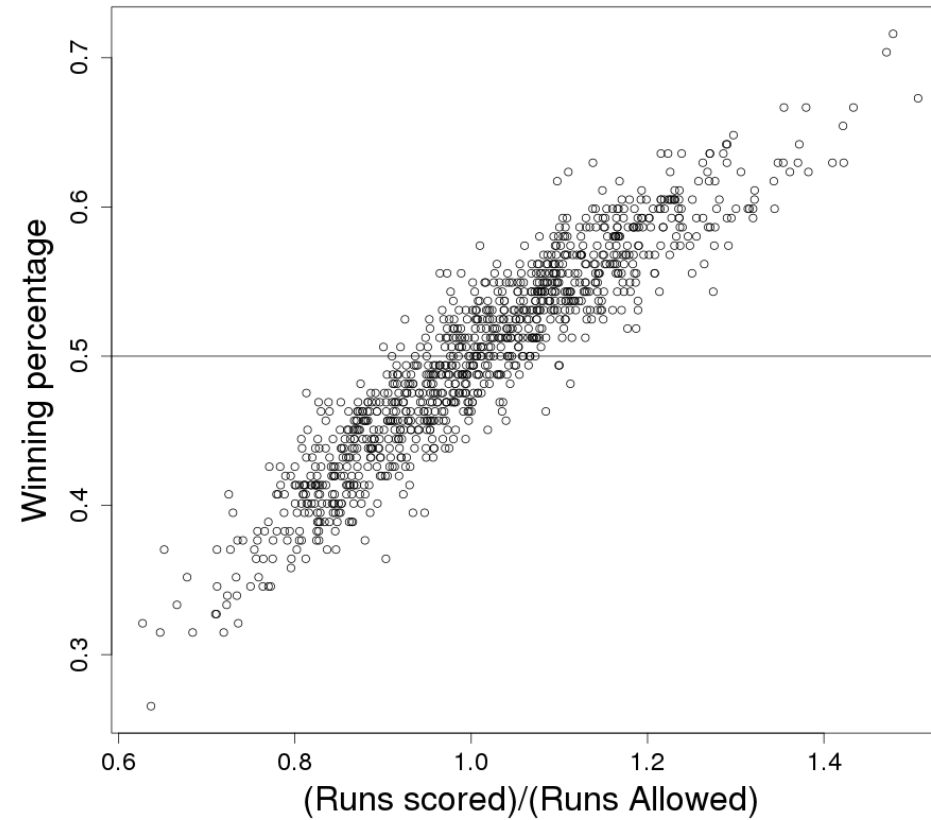
Runs allowed and wins



$$r = -.55$$

Correlation between runs and wins

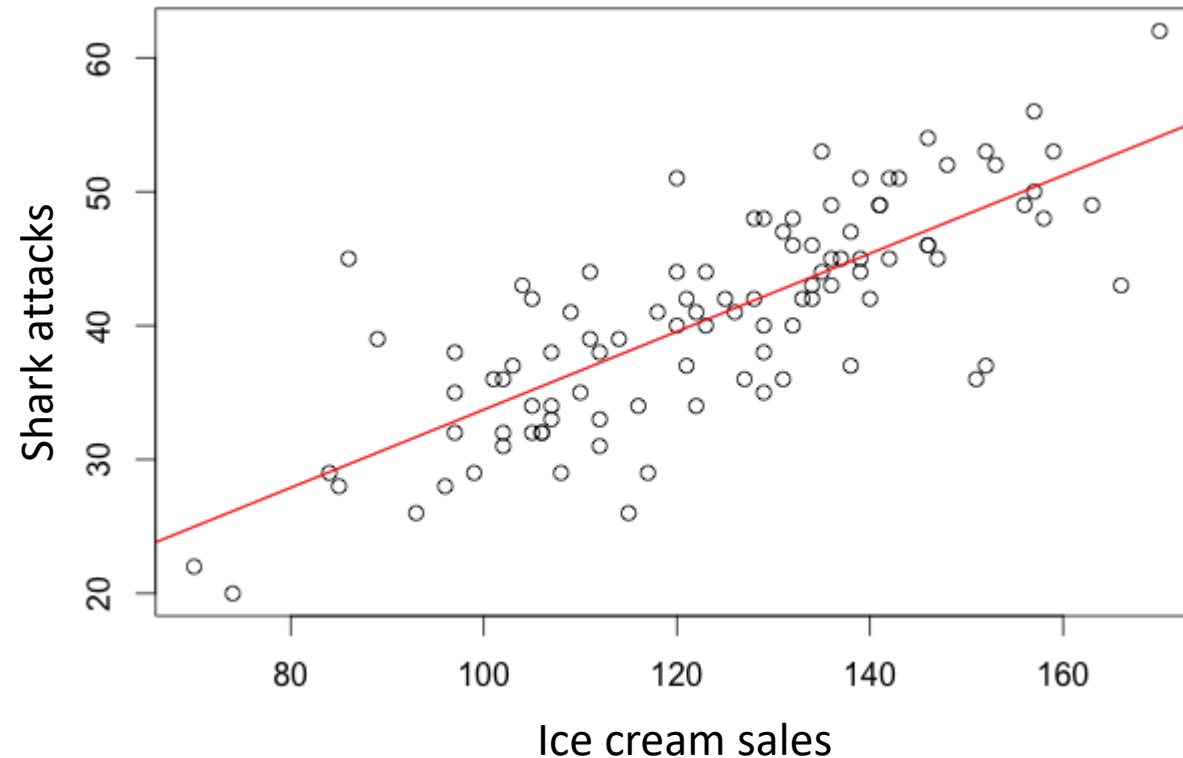
(runs scored)/(runs allowed) and wins



$r = .93$

Correlation caution #1

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



Correlation caution #1

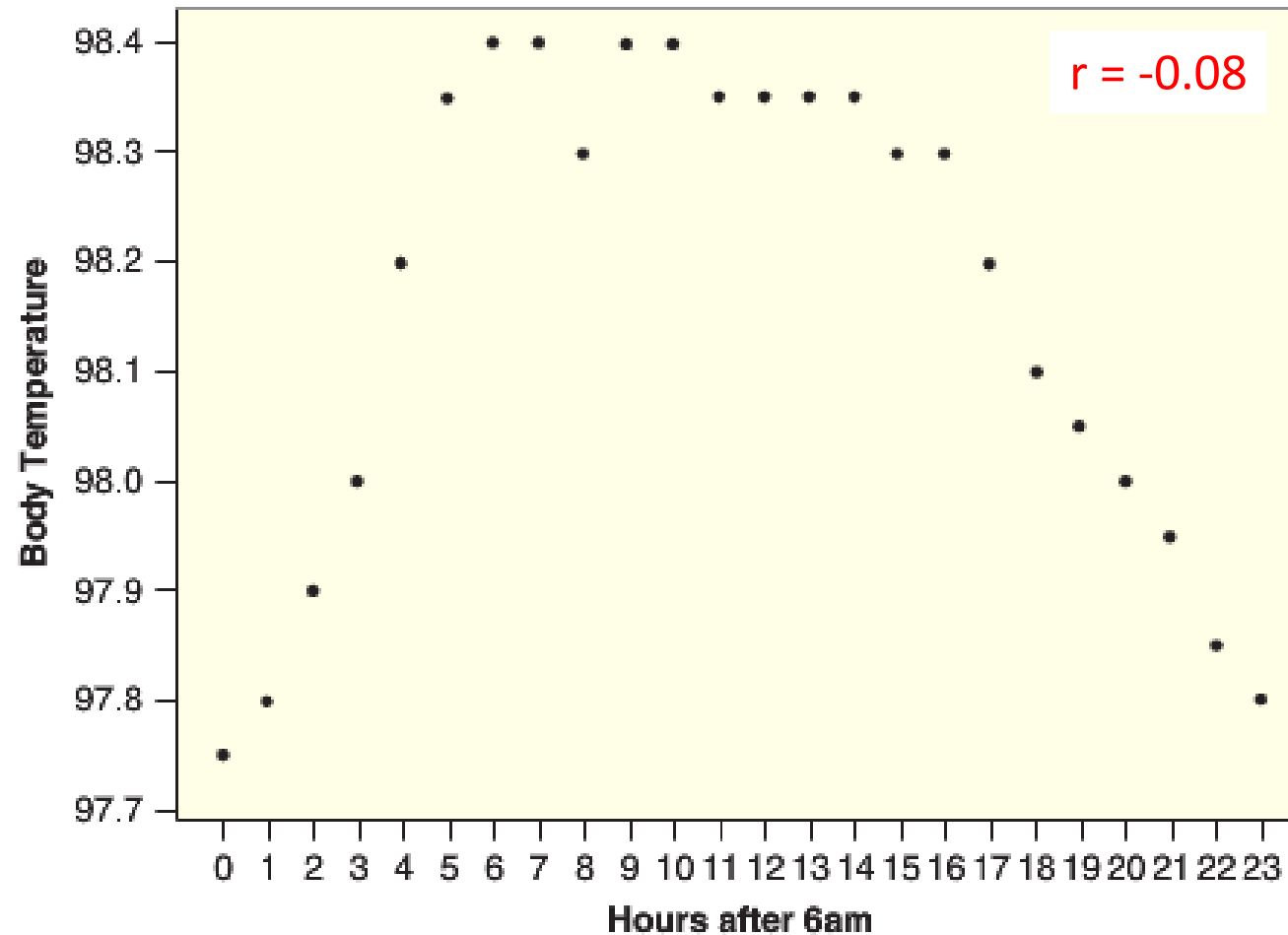
A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables



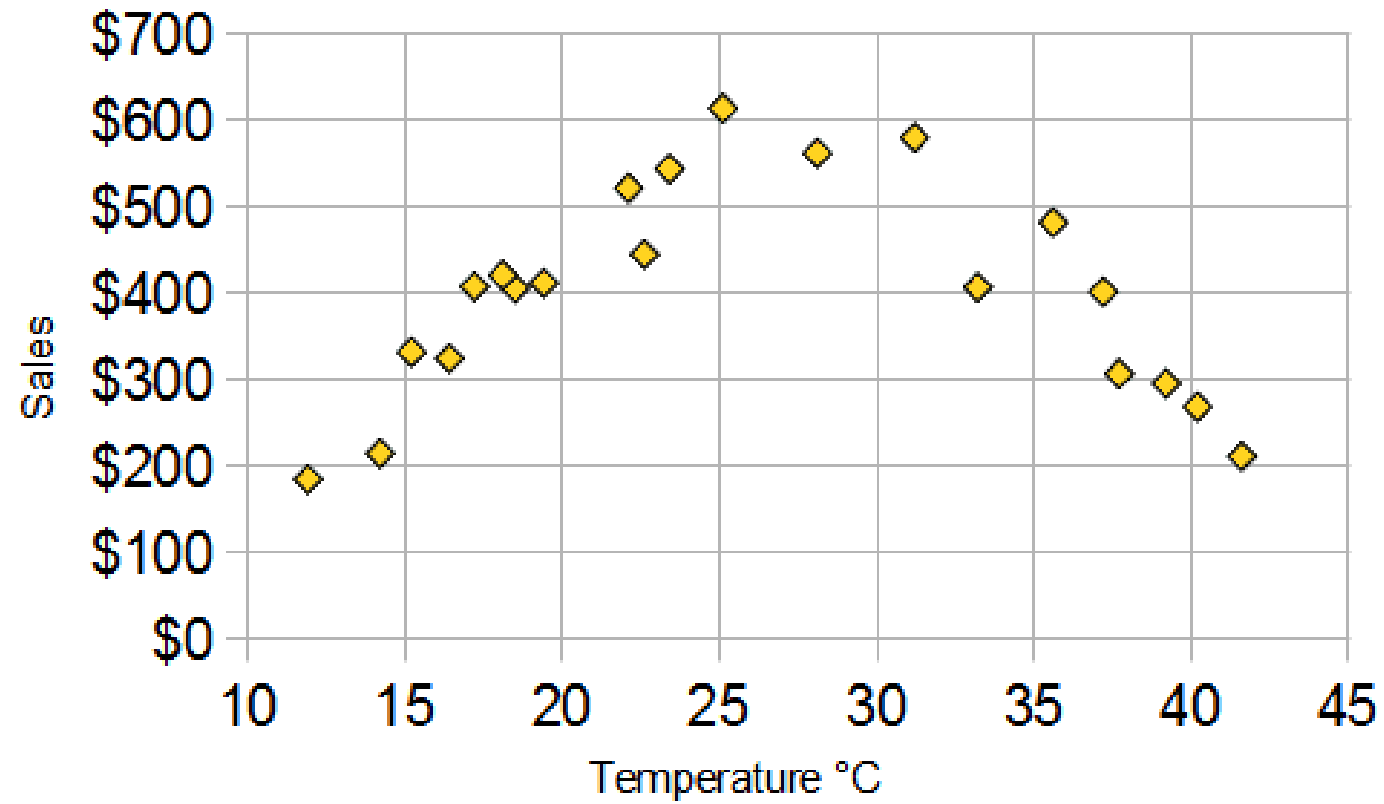
Correlation caution #2

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



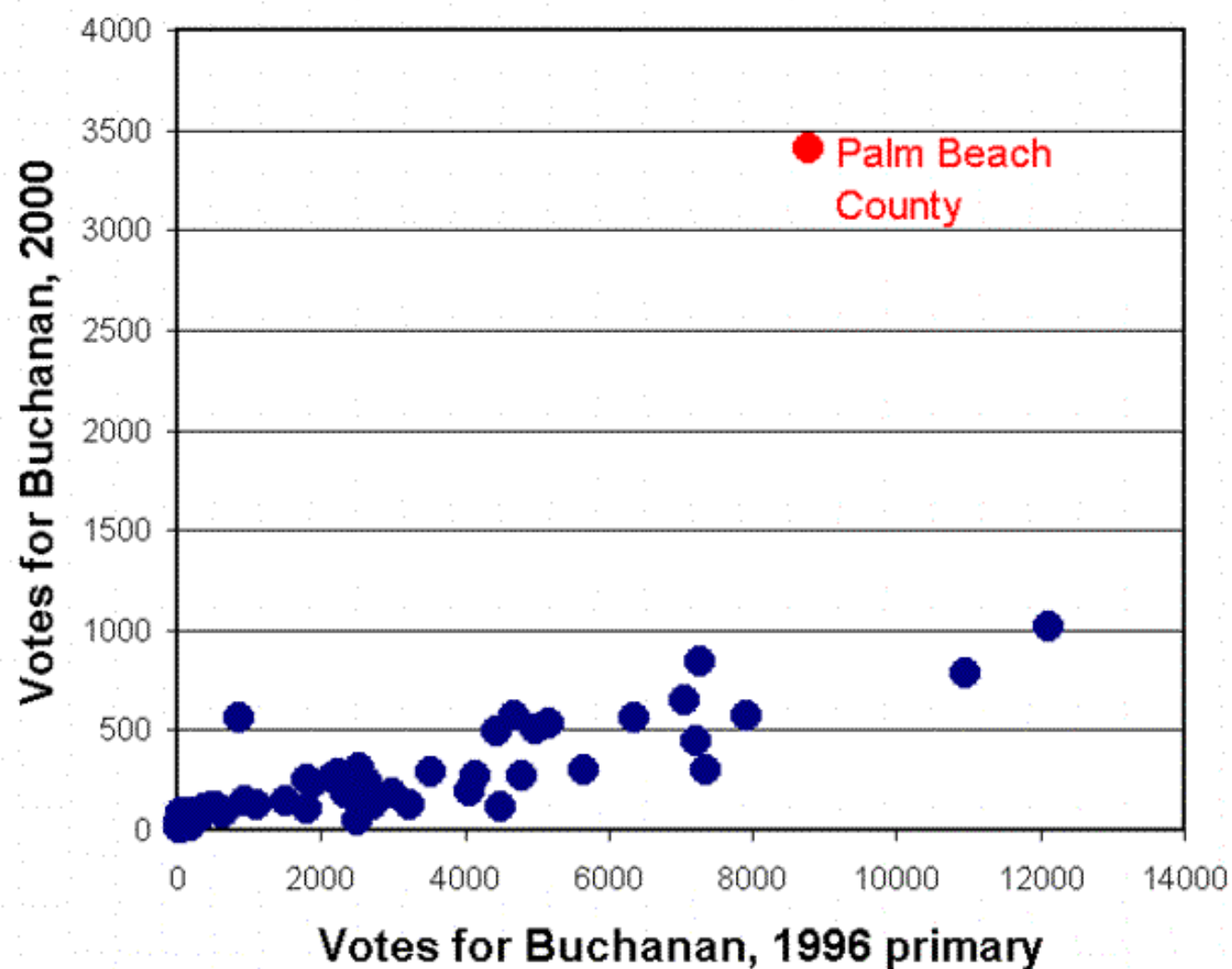
Ice cream sales and temperature



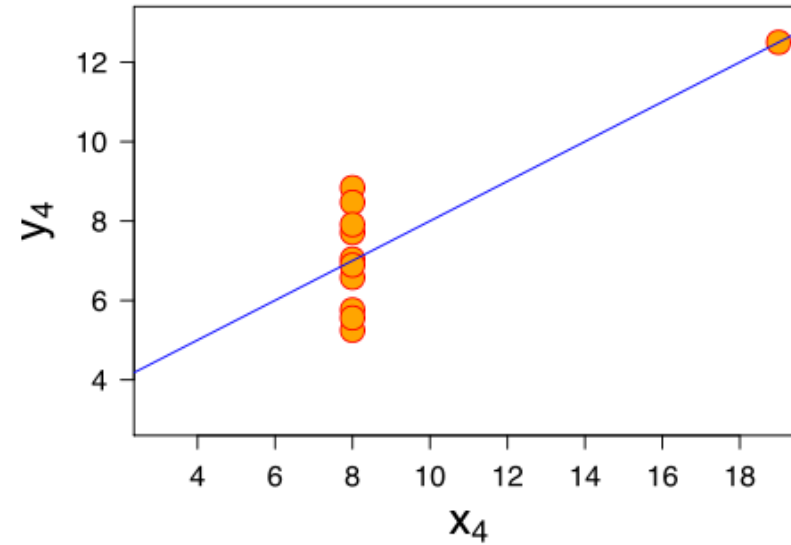
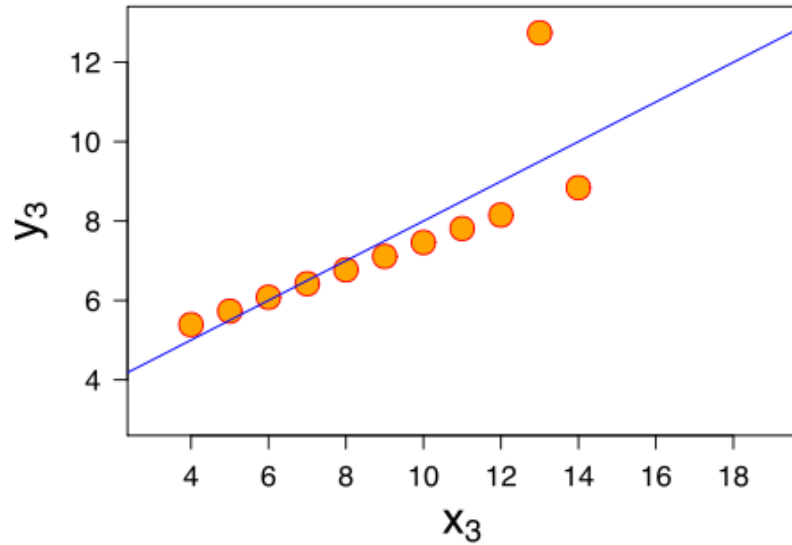
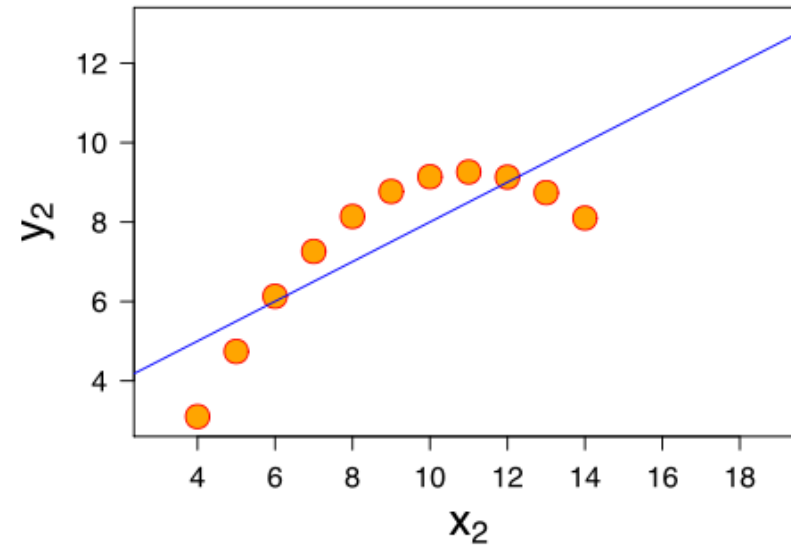
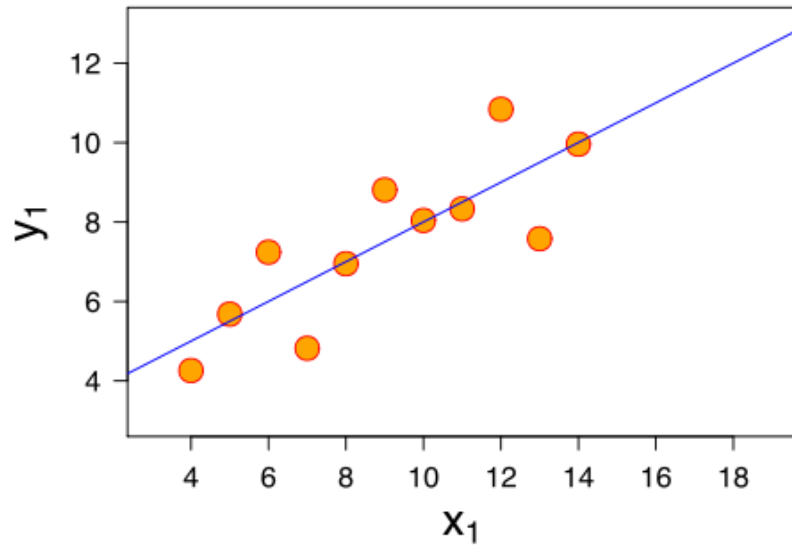
Correlation caution #3

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

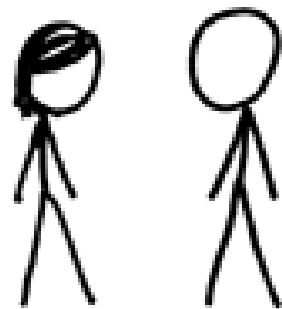
Presidential Election Results for Florida, by County



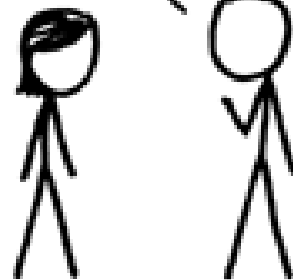
Anscombe's quartet ($r = 0.81$)



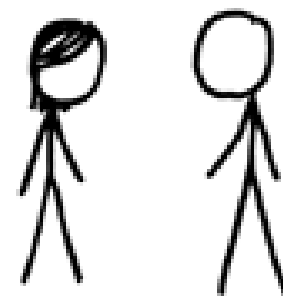
I USED TO THINK
CORRELATION IMPLIED
CAUSATION.



THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.



SOUNDS LIKE THE
CLASS HELPED.
WELL, MAYBE.



Who is a better hitter: Derek Jeter or David Ortiz?



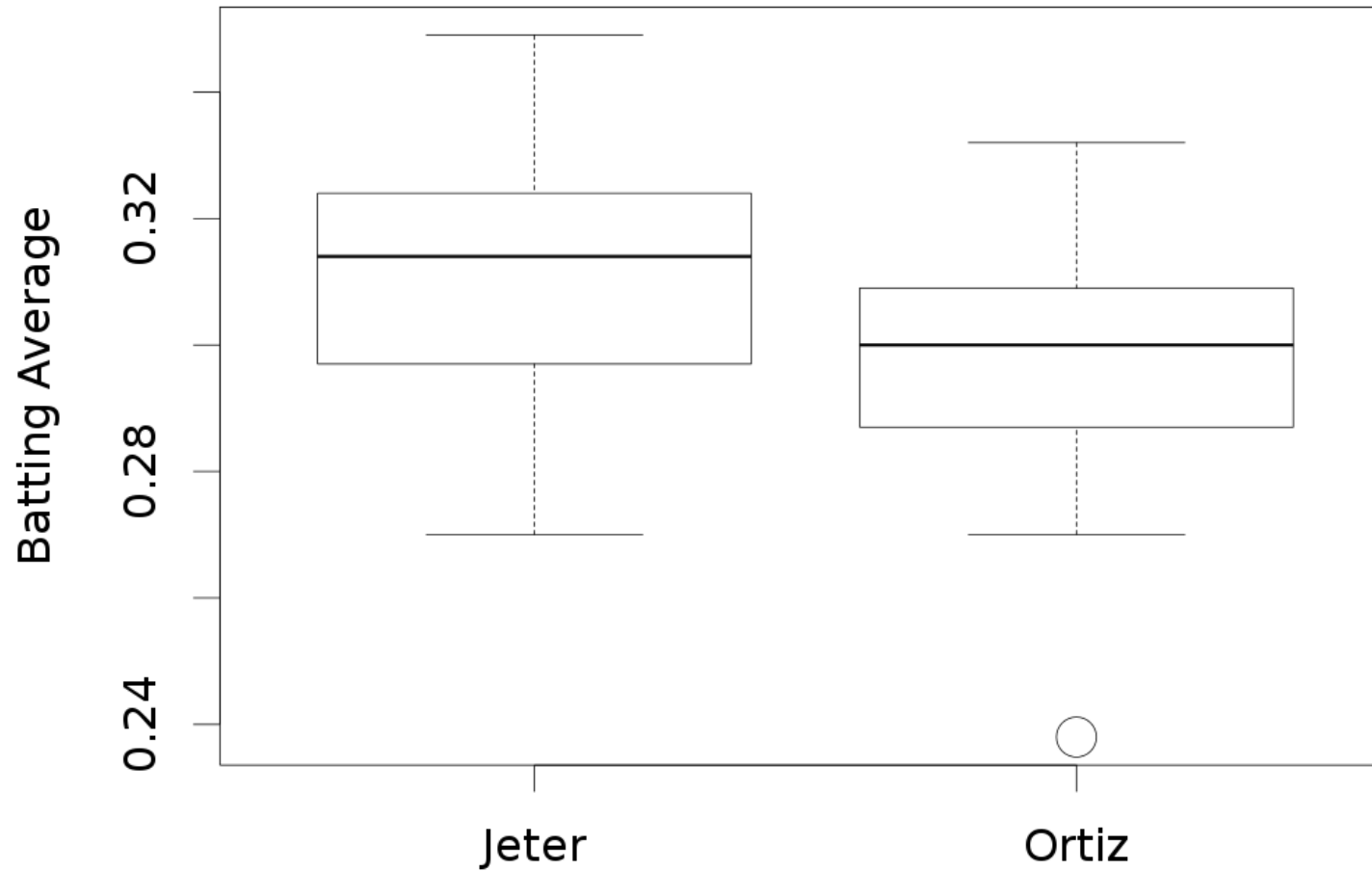
Derek Jeter



David Ortiz

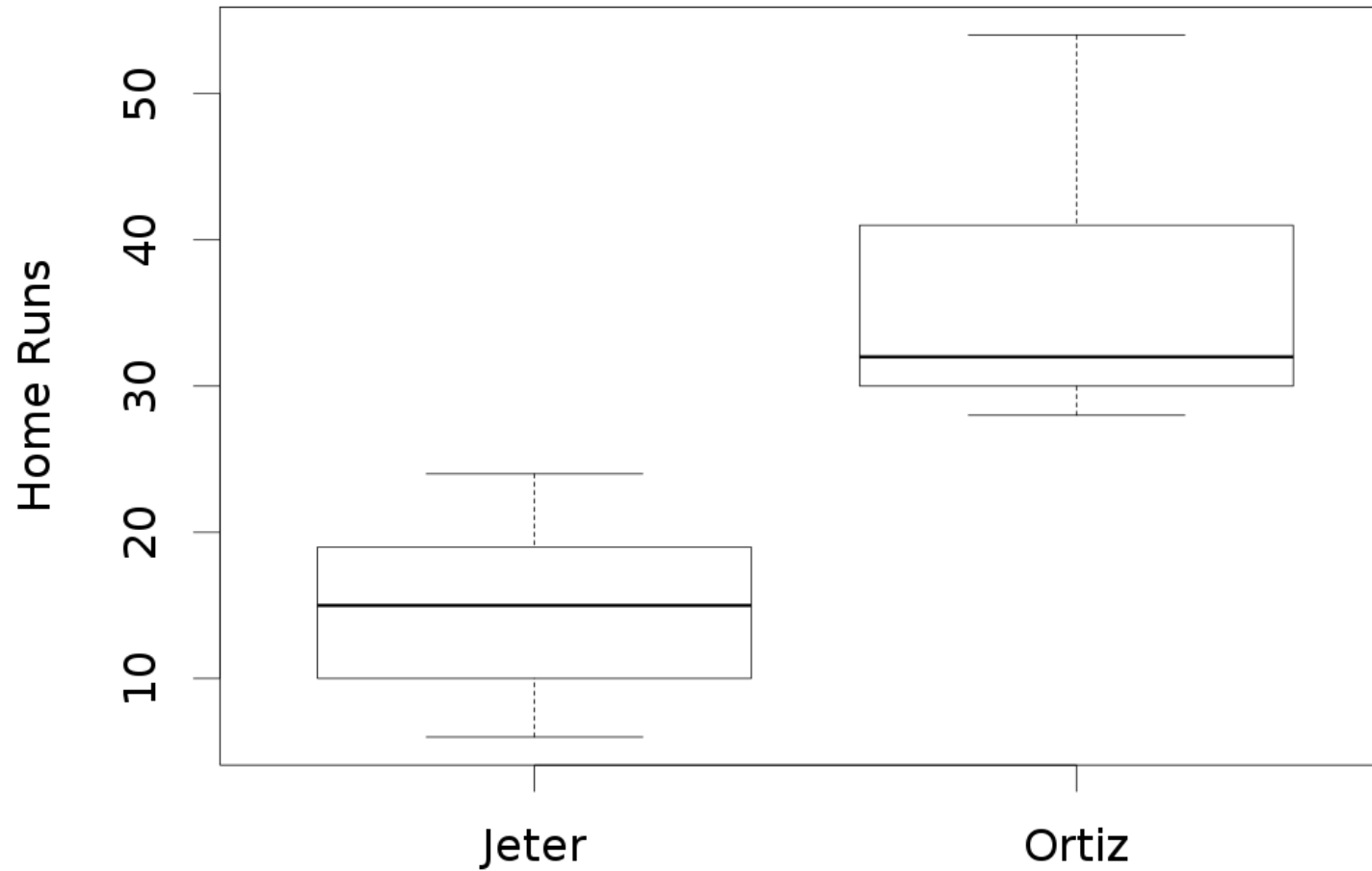
If we are going to pay these players millions of dollars, how can we assess who is best?

Who is a better hitter: Derek Jeter or David Ortiz?



Jeter has a better batting average

Who is a better hitter: Derek Jeter or David Ortiz?

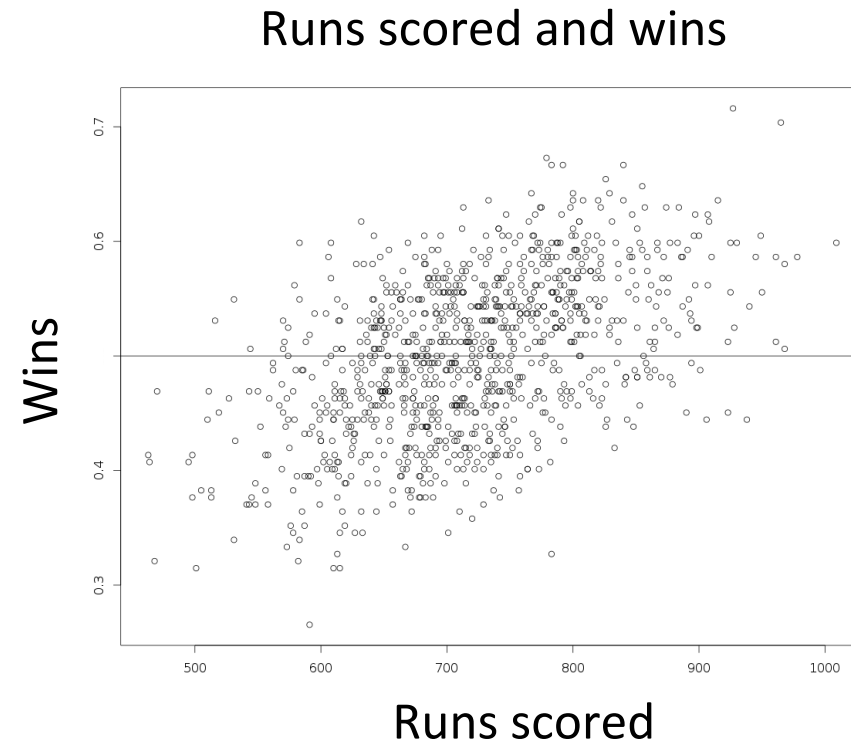


Ortiz hits more home runs

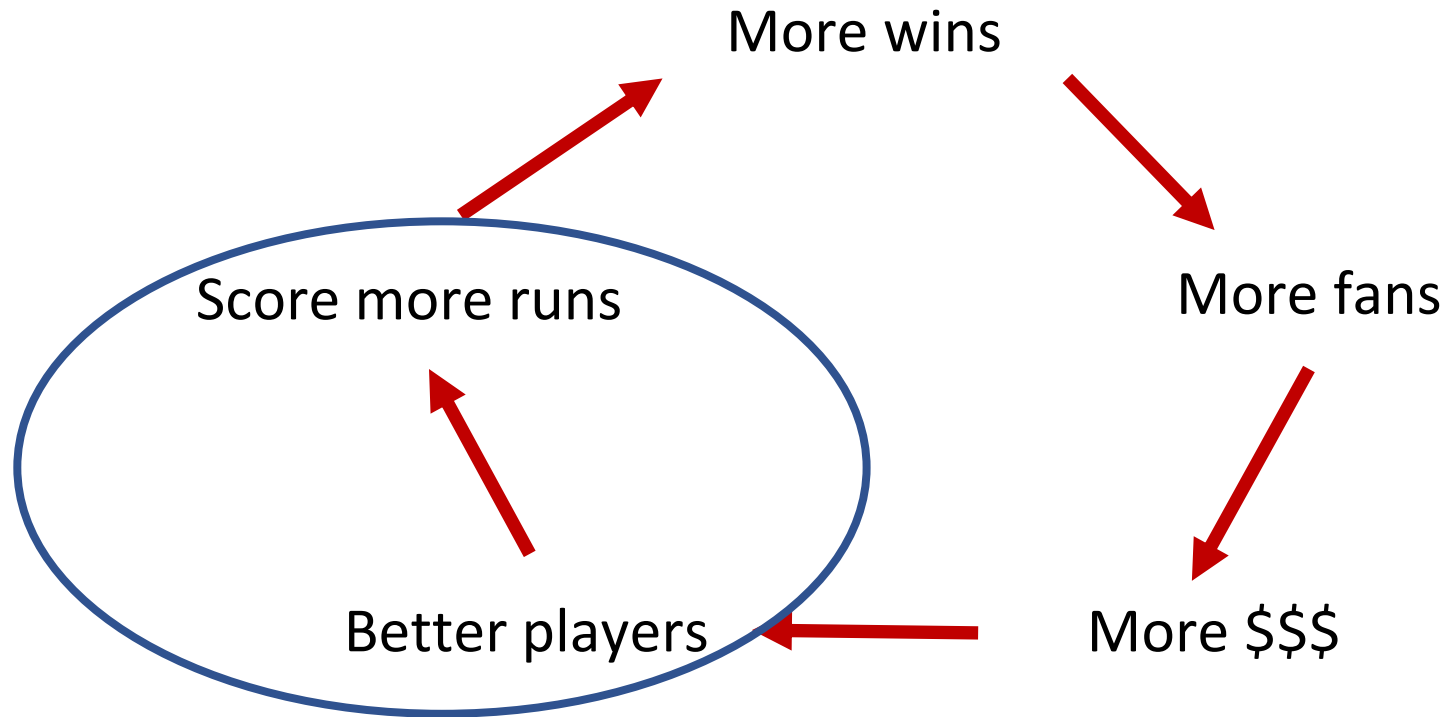
Is power or batting average more important?

It would be good to compare Jeter and Ortiz based on the “best” statistic

How do we determine which statistic is best?



The great cycle of baseball



We can evaluate how 'good' a statistic is based on how well it correlates with the number of runs a team scores

What is the best statistic to use?

One idea: the 'best' statistic to judge a player is the statistic that is most correlated with runs

- We can then use this to examine how good a hitter is

Descriptive statistics find the correlation with runs:

HR: Home runs

OBP: On-Base Percentage: $(H + BB)/PA$

BA: Batting Average: H/AB

SLG: Slugging percentage: $(1 \cdot 1B + 2 \cdot 2B + 3 \cdot 3B + 4 \cdot HR)/AB$

On-base plus Slugging (OPS): $OBP + SLG$

Calculating correlations in Python

To calculate the correlation with runs you can use numpy's `np.corrcoef()` function:

- (also see the class textbook for writing a correlation n function yourself):

```
corr_matrix = np.corrcoef(ndarray1, ndarray2)
```

`corr_matrix` will be a 2 x 2 matrix with:

- Diagonal elements are the correlation of an array and itself (i.e., 1)
- Off diagonal elements are the correlation between `ndarray1` and `ndarray2`

We can then get the (scalar) value of the correlation of `ndarray1` and `ndarray2` using:

- `corr_matrix[0, 1]`

Lab 8, problem 4!

In Lab 8, problem 4 you will find which statistic has the highest correlation with runs, and then compare Jeter and Ortiz on this statistic

Get started early, this homework is on the long side!

Next class: Linear regression for creating better performance metrics and for prediction