S&DS 173
Ydata: Analysis of Baseball Data

Ethan Meyers

Overview

Lab 0 discussion

Discussion of preface and prologue to Astroball

Watch an inning of the 2014 All-star game

Review of structured data and classic baseball statistics

Python!

Lab 0: questions?

How did it go?

Was everyone able to complete it?

Astroball discussion of Preface and prologue

Interesting quotes from the preface and prologue?

Preface and prologue of Astroball?

Season	League	Division	Finish[2]	Wins[2]	Losses[2]	Win%[2]	GB[2]
2010	NL	Central	4th	76	86	0.469	15
2011	NL	Central	6th	56	106	0.346	37½
2012	NL	Central	6th	55	107	0.34	42
2013	AL	West	5th	51	111	0.315	45
2014	AL	West	4th	70	92	0.432	28
2015	AL	West	2nd ¤	86	76	0.531	2

Jeopardy 11/18/2013

Preface and prologue of Astroball?

[Astrodome]

The lineup

National				American		
Order	Player	Position		Order	Player	Position
1	Andrew McCutchen	CF		1	Derek Jeter	SS
2	Yasiel Puig	RF		2	Mike Trout	LF
3	Troy Tulowitzki	SS		3	Robinson Canó	2B
4	Paul Goldschmidt	1B		4	Miguel Cabrera	1B
5	Giancarlo Stanton	DH		5	José Bautista	RF
6	Aramis Ramírez	3B		6	Nelson Cruz	DH
7	Chase Utley	2B		7	Adam Jones	CF
8	Jonathan Lucroy	C		8	Josh Donaldson	3B
9	Carlos Gómez	LF		9	Salvador Pérez	C
	Adam Wainwright	P			Félix Hernández	P

To learn the basics of baseball let's watch and keep score for the 2014 all-star game

Score card

#	Player	Pos	1	2
	Henry		1B	
	Clarke		2B	
	Navi		OUT	
	Terra		OUT	
	Gabe		K	
	Jake			BB

[More official score keeping](#)

2014 All-star game

National				American		
Order	Player	Position		Order	Player	Position
1	Andrew McCutchen	CF		1	Derek Jeter	SS
2	Yasiel Puig	RF		2	Mike Trout	LF
3	Troy Tulowitzki	SS		3	Robinson Canó	2B
4	Paul Goldschmidt	1B		4	Miguel Cabrera	1B
5	Giancarlo Stanton	DH		5	José Bautista	RF
6	Aramis Ramírez	3B		6	Nelson Cruz	DH
7	Chase Utley	2B		7	Adam Jones	CF
8	Jonathan Lucroy	C		8	Josh Donaldson	3B
9	Carlos Gómez	LF		9	Salvador Pérez	C
	Adam Wainwright	P			Félix Hernández	P

Retrosheet play-by-play data

Let’s take a quick dive into the retrosheet play-by-play data in Python

Please download Lab 1

We will take a quick look at the retrosheet data and you will do some exercises on it for homework

Retrosheet play-by-play data

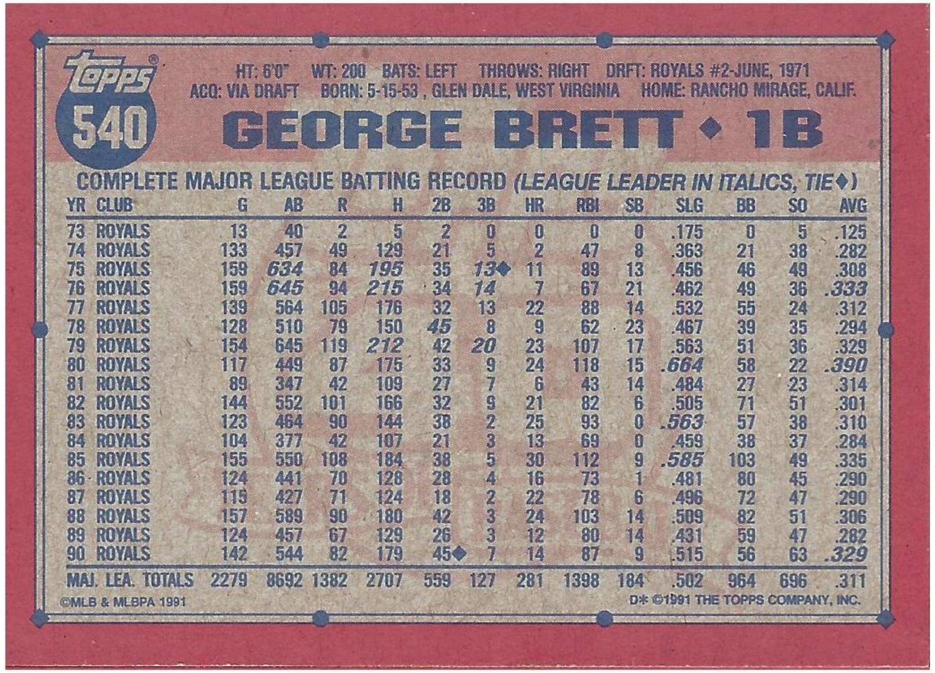
INN_CT	BAT_HOME_ID	OUTS_CT	RESP_BAT_ID	PITCH_SEQ_TX	EVENT_TX
1	0	0	mccua001	BX	S6/G+
1	0	0	puigy001	BB	WP.1-2
1	0	0	puigy001	BB.SFS	K
1	0	1	tulot001	C*BS>S	K+SB3
1	0	2	goldp001	BCX	53/G
1	1	0	jeted001	BX	D9/L+
1	1	0	troum001	FBBS*BX	T8/L+.2-H
1	1	0	canor001	SFBS	K
1	1	1	cabrm001	FX	HR/7/L.3-H

Interpreting each pitch event: **PITCH_SEQ_TX and EVENT_TX**

https://www.retrosheet.org/eventfile.htm#5

Common baseball statistics

Let’s look at some baseball cards!



topps 540

HT: 6'0" WT: 200 BATS: LEFT THROWS: RIGHT DRFT: ROYALS #2-JUNE, 1971
ACQ: VIA DRAFT BORN: 5-15-53 , GLEN DALE, WEST VIRGINIA HOME: RANCHO MIRAGE, CALIF.

GEORGE BRETT ◆ 1B

COMPLETE MAJOR LEAGUE BATTING RECORD *(LEAGUE LEADER IN ITALICS, TIE◆)*

YR	CLUB	G	AB	R	H	2B	3B	HR	RBI	SB	SLG	BB	SO	AVG
73	ROYALS	13	40	2	5	2	0	0	0	0	.175	0	5	.125
74	ROYALS	133	457	49	129	21	5	2	47	8	.363	21	38	.282
75	ROYALS	159	*634*	84	*195*	35	*13*◆	11	89	13	.456	46	49	.308
76	ROYALS	159	*645*	94	*215*	34	*14*	7	67	21	.462	49	36	*.333*
77	ROYALS	139	564	105	176	32	13	22	88	14	.532	55	24	.312
78	ROYALS	128	510	79	150	*45*	8	9	62	23	.467	39	35	.294
79	ROYALS	154	645	119	*212*	42	*20*	23	107	17	.563	51	36	.329
80	ROYALS	117	449	87	175	33	9	24	118	15	*.664*	58	22	*.390*
81	ROYALS	89	347	42	109	27	7	6	43	14	.484	27	23	.314
82	ROYALS	144	552	101	166	32	9	21	82	6	.505	71	51	.301
83	ROYALS	123	464	90	144	38	2	25	93	0	*.563*	57	38	.310
84	ROYALS	104	377	42	107	21	3	13	69	0	.459	38	37	.284
85	ROYALS	155	550	108	184	38	5	30	112	9	*.585*	103	49	.335
86	ROYALS	124	441	70	128	28	4	16	73	1	.481	80	45	.290
87	ROYALS	115	427	71	124	18	2	22	78	6	.496	72	47	.290
88	ROYALS	157	589	90	180	42	3	24	103	14	.509	82	51	.306
89	ROYALS	124	457	67	129	26	3	12	80	14	.431	59	47	.282
90	ROYALS	142	544	82	179	45◆	7	14	87	9	.515	56	63	*.329*
MAJ. LEA. TOTALS		2279	8692	1382	2707	559	127	281	1398	184	.502	964	696	.311

Let’s look at some baseball cards!

First Name	Last Name		First Name	Last Name
Austin	O'Toole		Al	Newman
Ben	Scher		Curt	Wilkerson
Gaby	Branin		Dion	James
Harry	Hegeman		Gary	Gaetti
Hassan	Siddiq		Jim	Presley
Jack	Klinger		John	Russell
Jonathan	Boulaphinh		Kirt	Manwaring
Krish	Maypole		Oddibe	McDowell
Matt	Leone		Rick	Cerone
Max	Krupnick		Sid	Bream
Raphael	Berz		Steve	Lyons
Rohan	Handa		Terry	Kennedy
Sorenie	Gudissa		Tim	Jones
Teddy	Hague		Tim	Teufel

http://bit.ly/baseball_cards https://github.com/emeyers/SDS173/tree/main/images/baseball_cards

statistics and structured data

statistics: a numerical summary of data

(technically a summary of a data sample)

Statistics: is the mathematics of collecting, organizing and interpreting data

Describing and summarizing data

statistics that are used to summarize a data set (sample of data) are called **descriptive statistics**

Examples:

- Maximum value in the data set
- Minimum value in the data set
- <u>Mean</u> value of the data set

Common baseball descriptive statistics

G = games

- Number of games a player participated in (out of 162 games in a season)

AB = at bats

- Number of times a batter was hitting and either got a hit or got out (does not include walks or reaching base on an error)

R = runs

- Number of runs the player scored

H = hit

- Number of times a player hit the ball on got on base or hit a home run (sum of 1B, 2B, 3B, HR)

Common baseball statistics

BB = base on balls (walks)

- Number of times a player got on base do to the pitcher throwing 4 balls

RBI = Runs batted in

- How many runs scored as a result of a player getting a hit

SB = stolen bases

- Number of times a runner advanced by ‘stealing a base’

Common derived baseball statistics

AVG= batting average

- Hits/(At bats) = H/AB = (1B + 2B + 3B + HR)/AB

SLG = slugging percentage

- (1 * 1B + 2 * 2B + 3 * 3B + 4 * 4B) /AB

Lahman Database – Individual player yearly batting statistics

As we saw in Lab 0, the Batting.csv file in the Lahman database contains batting information about all baseball players for each season from 1871 to 2018.

You will extract information the particular baseball player on your card from this dataset

But first let’s talk about some general terms for structured data

Structured data

Variables

Cases

	playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H
1	aardsda01	2004	1	SFN	NL	11	11	0	0	0
2	aardsda01	2006	1	CHN	NL	45	43	2	0	0
3	aardsda01	2007	1	CHA	AL	25	2	0	0	0
4	aardsda01	2008	1	BOS	AL	47	5	1	0	0
5	aardsda01	2009	1	SEA	AL	73	3	0	0	0
6	aardsda01	2010	1	SEA	AL	53	4	0	0	0
7	aardsda01	2012	1	NYA	AL	1	NA	NA	NA	NA
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131

Data taken from the Lahman Batting dataset

Structured data

Variables

Cases

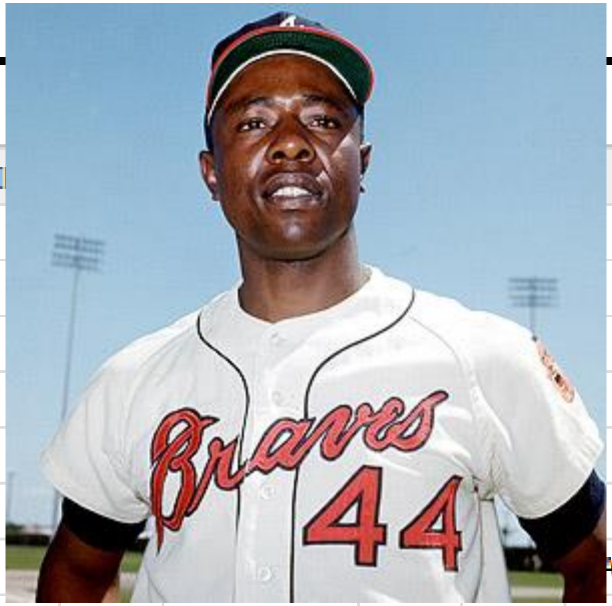
	playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H
1	aardsda01	2004	1					0	0	0
2	aardsda01	2006	1					2	0	0
3	aardsda01	2007	1					0	0	0
4	aardsda01	2008	1					1	0	0
5	aardsda01	2009	1					0	0	0
6	aardsda01	2010	1					0	0	0
7	aardsda01	2012	1					NA	NA	NA
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131

Structured data

Variables

Cases

	playerID	yearID	stint	teamID	lgI						
1	aardsda01	2004	1	SFN	NL						
2	aardsda01	2006	1	CHN	NL						
3	aardsda01	2007	1	CHA	AL						
4	aardsda01	2008	1	BOS	AL						
5	aardsda01	2009	1	SEA	AL						
6	aardsda01	2010	1	SEA	AL						
7	aardsda01	2012	1	NYA	AL						
8	aaronha01	1954	1	ML1	NL	122	122		468	58	131



Categorical and Quantitative Variables

Categorical Variable

Quantitative Variable

Cases

	playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H
1	aardsda01	2004	1	SFN	NL	11	11	0	0	0
2	aardsda01	2006	1	CHN	NL	45	43	2	0	0
3	aardsda01	2007	1	CHA	AL	25	2	0	0	0
4	aardsda01	2008	1	BOS	AL	47	5	1	0	0
5	aardsda01	2009	1	SEA	AL	73	3	0	0	0
6	aardsda01	2010	1	SEA	AL	53	4	0	0	0
7	aardsda01	2012	1	NYA	AL	1	NA	NA	NA	NA
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131

Explanatory and Response Variables

Sometimes we use one variable (the explanatory variable) to understand/predict another variable (the response variable)

	playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H
1	aardsda01	2004	1	SFN	NL	11	11	0	0	0
2	aardsda01	2006	1	CHN	NL	45	43	2	0	0
3	aardsda01	2007	1	CHA	AL	25	2	0	0	0
4	aardsda01	2008	1	BOS	AL	47	5	1	0	0
5	aardsda01	2009	1	SEA	AL	73	3	0	0	0
6	aardsda01	2010	1	SEA	AL	53	4	0	0	0
7	aardsda01	2012	1	NYA	AL	1	NA	NA	NA	NA
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131

Another Dataset – 2014 Team statistics

Variables

Cases

	Tm	X.Bat	BatAge	R.G	G	PA	AB	R	H
1	ARI	52	27.6	3.80	162	6089	5552	615	1379
2	ATL	39	26.8	3.54	162	6064	5468	573	1316
3	BAL	39	28.3	4.35	162	6130	5596	705	1434
4	BOS	50	29.2	3.91	162	6226	5551	634	1355
5	CHC	48	26.8	3.79	162	6102	5508	614	1315
6	CHW	36	27.7	4.07	162	6077	5543	660	1400
7	CIN	45	28.9	3.67	162	5978	5395	595	1282
8	CLE	43	28.5	4.13	162	6222	5575	669	1411

Finding data on the player on our card

To find information on the player on our card we need to:

1. Find our player’s playerID using the Player.csv dataset

2. Use our player’s playerID to filter out only the rows in the Batting.csv file that contain information about our player

You will do these steps in the second problem of lab 1

What are “good” statistics?

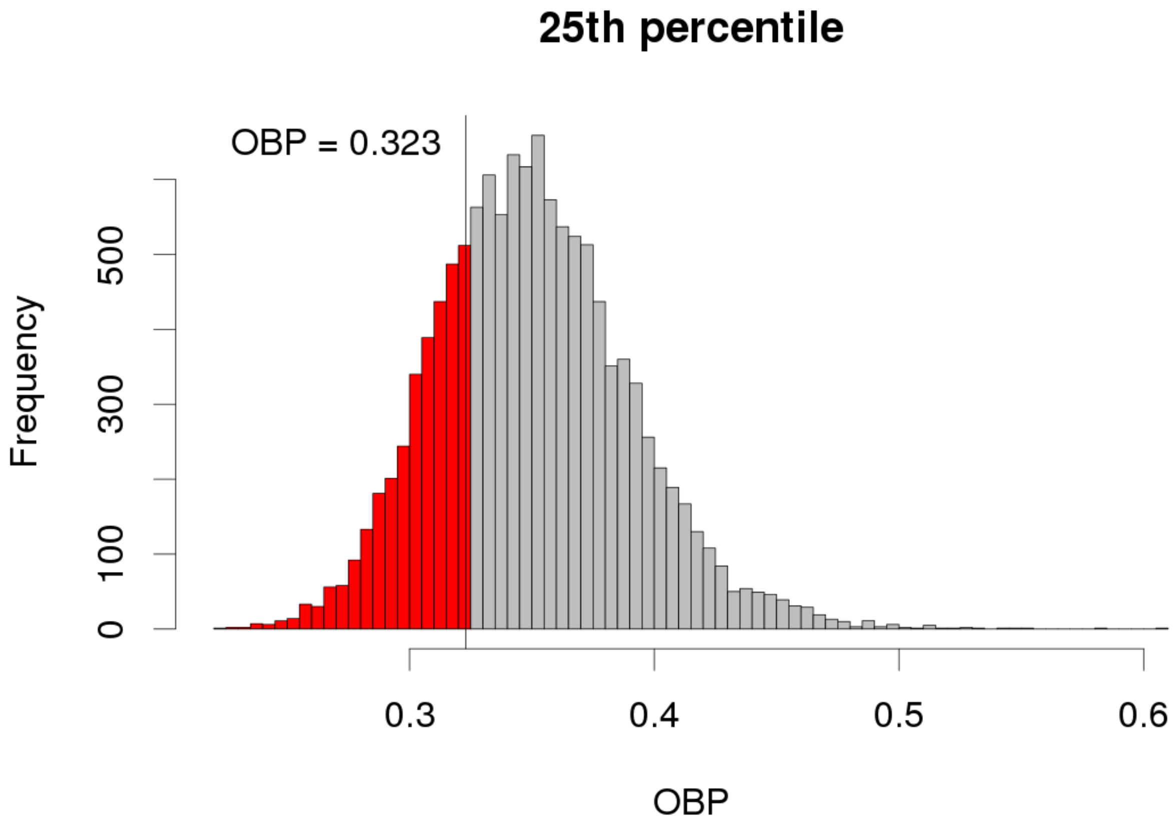
How could we determine what a “good” value for a statistic is?

- i.e., how many home runs would need to be hit to determine if a player is “good at hitting home runs”?

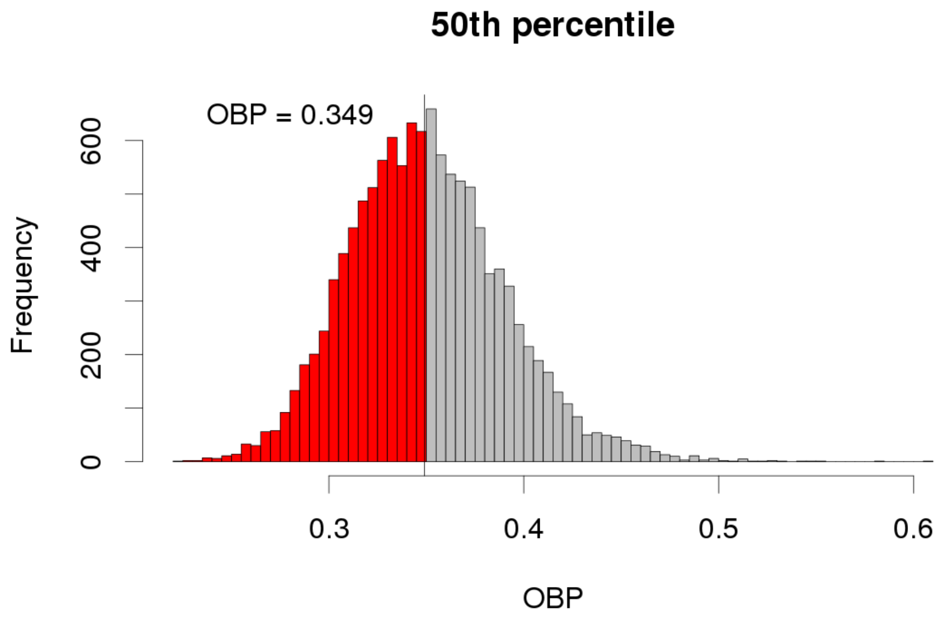
One method to determine what a “good” statistic is, would be a value that is say greater than 90% of baseball players

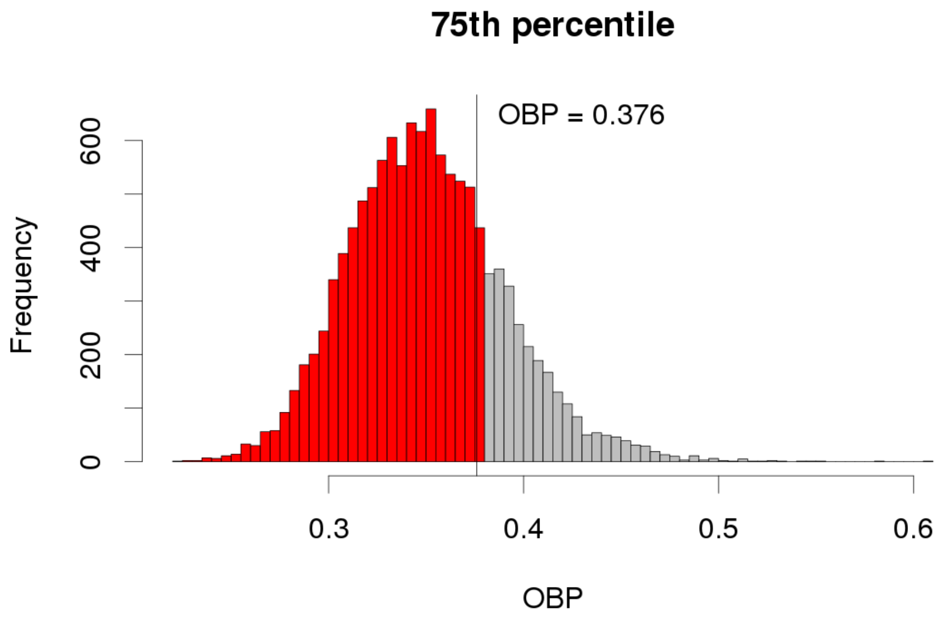
Percentiles

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



Percentiles/quantiles





https://emeyers.shinyapps.io/baseball_stat_percentiles/

What is a good statistic for…?

What are “good” values are for the following statistics:

- (I used the years from 1971 to 2014, min PA = 500)

<u>Home runs (HR)?</u>

- [see lab 1]

<u>On base percentage (OBP)?</u>

- .394

<u>Batting average (BA)</u>

- .313

<u>Strikeouts (SO)</u>

- 47
- Bad is 129 (90^{th} percentile)

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)

Roughly divides the data into fourths

Range and Interquartile Range

Range = maximum – minimum

Interquartile range (IQR) = $Q_3 - Q_1$

Detecting of outliers

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

Smaller than: Q_1 - 1.5 * IQR

Larger than: Q_3 + 1.5 * IQR

Are there any outlier years in David Ortiz home run numbers?

1. Five Number Summary: (23, 28.5, 32, 36, 54)
2. Range: 31
3. Interquartile range (IQR) = 7.5

Boxplots

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

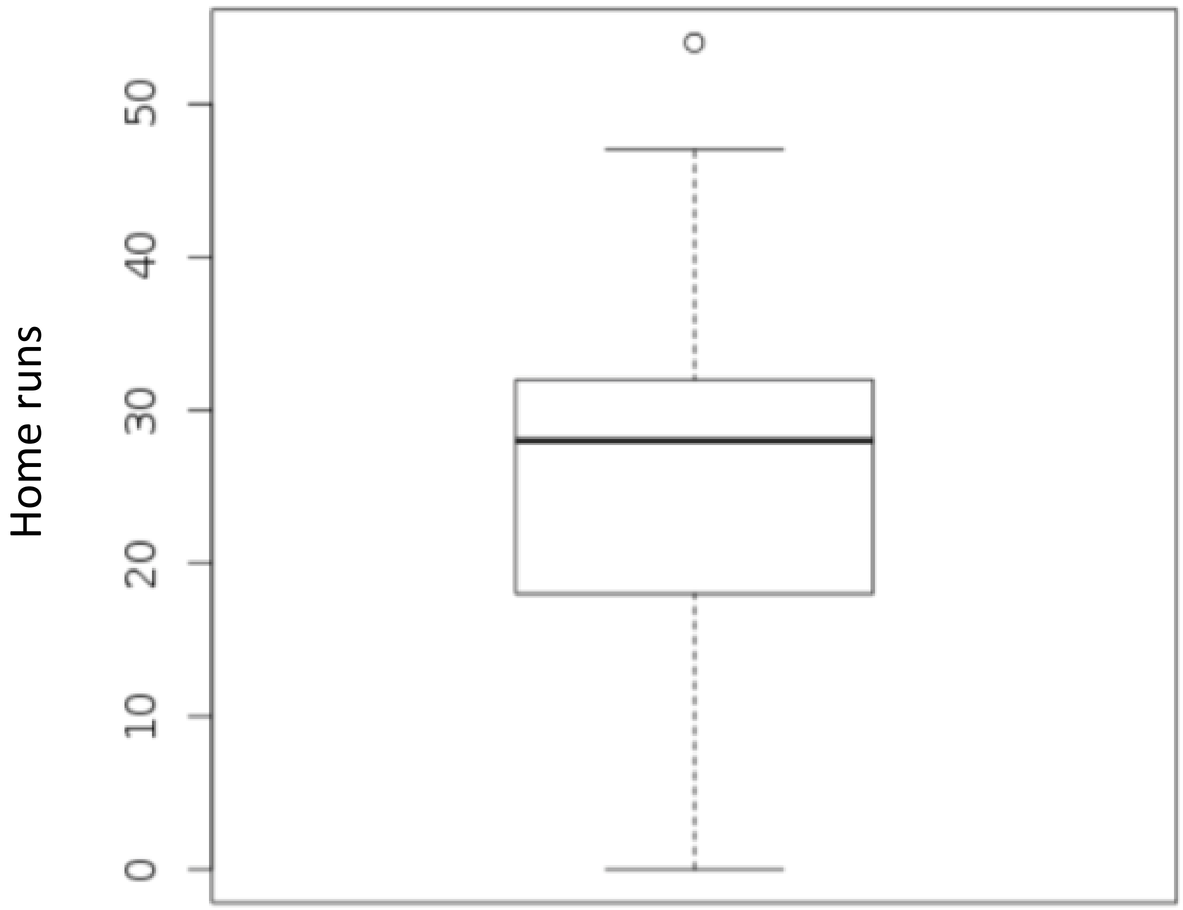
1. Drawing a box from Q_1 to Q_3

2. Dividing the box with a line drawn at the median

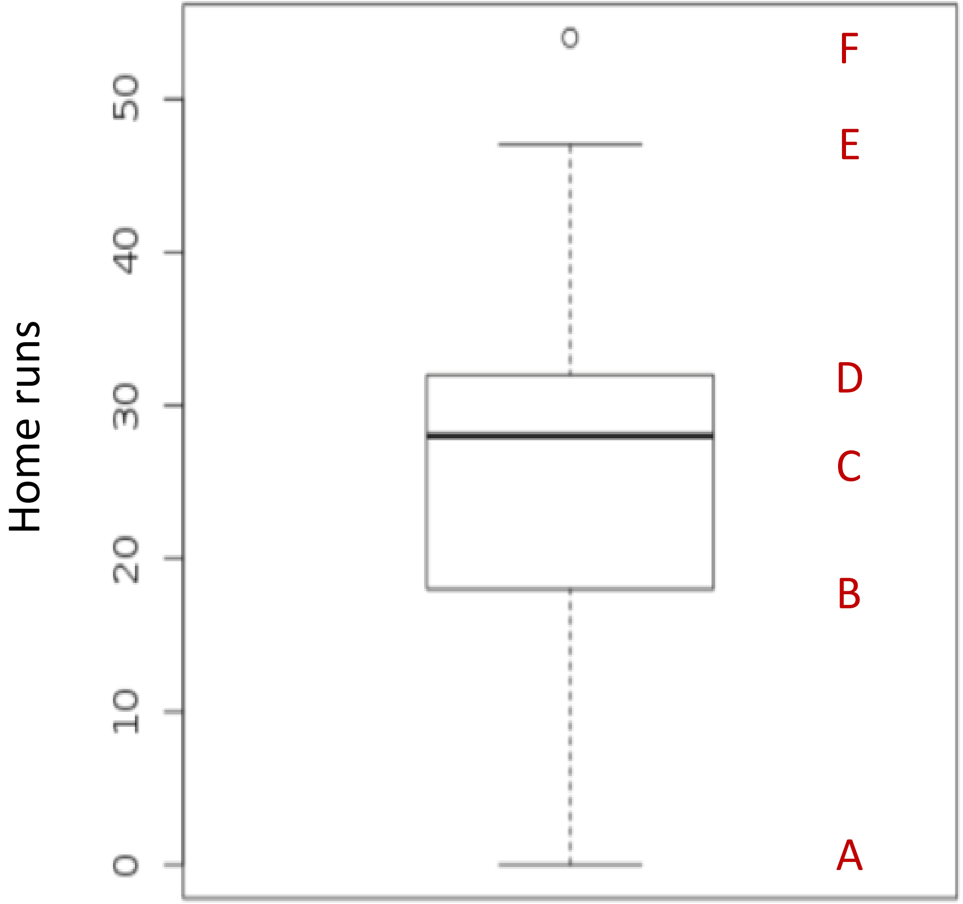
3. Draw a line from each quartile to the most extreme data value that is not and outlier

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

Box plot of David Ortiz home runs



Box plot quiz



What is:

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

Who is better?

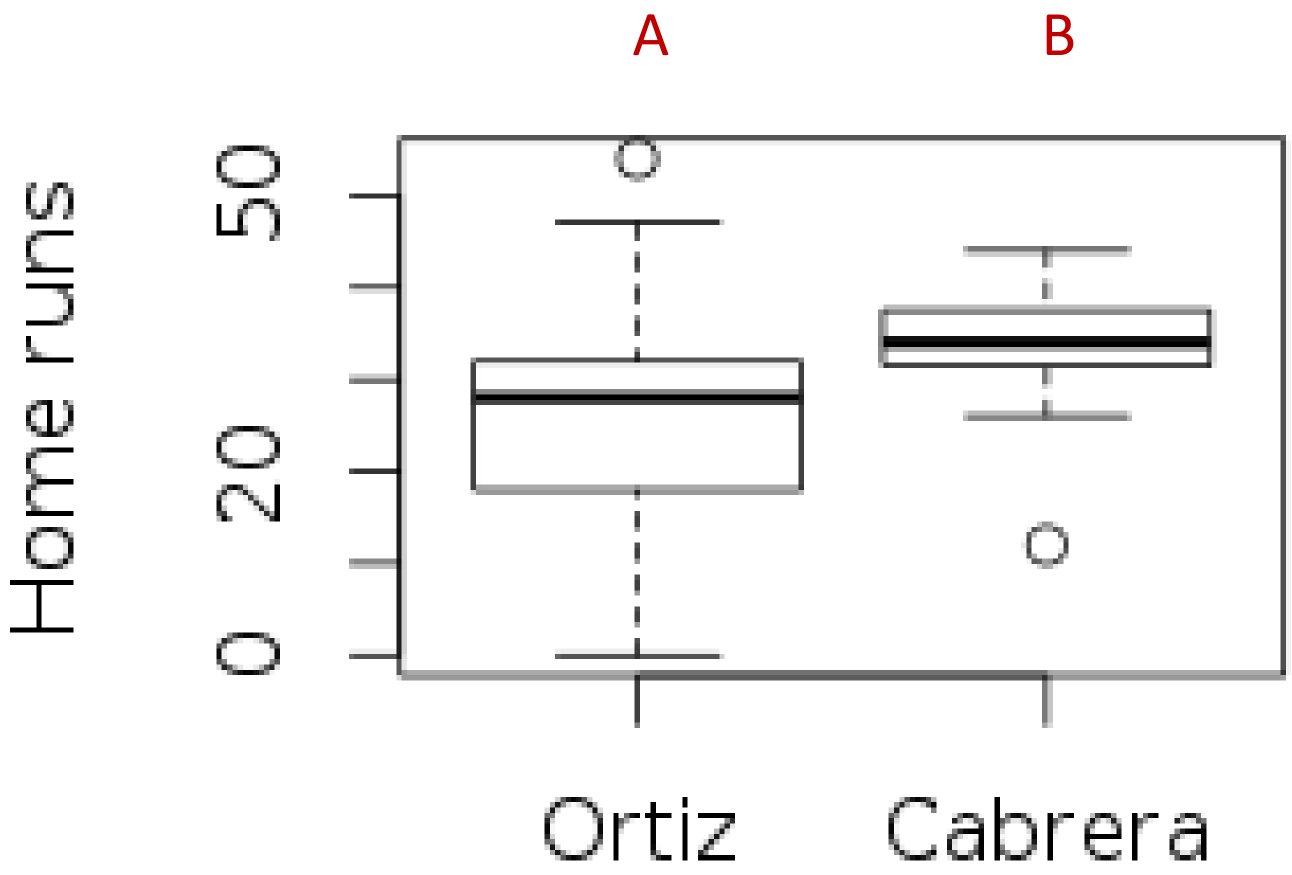
Miguel Cabrera:

HR in 2014 = 25

David Ortiz:

HR in 2014 = 35

Comparing players with side-by-side box plots



How would you describe the differences between these two players in terms of HRs?
Who is better?

Visualizing the ‘shape’ of how data is distributed

Boxplots can give us a sense of some key statistics about our data

There are other methods that can give us a better picture of the shape of how all the data is distributed

Stemplot for team HR in 2014

Let's look at the 2014 team data

	Tm	X.Bat	BatAge	R.G	G	PA	AB	R	H	X2B	X3B	HR
1	ARI	52	27.6	3.80	162	6089	5552	615	1379	259	47	118
2	ATL	39	26.8	3.54	162	6064	5468	573	1316	240	22	123
3	BAL	39	28.3	4.35	162	6130	5596	705	1434	264	16	211
4	BOS	50	29.2	3.91	162	6226	5551	634	1355	282	20	123
5	CHC	48	26.8	3.79	162	6102	5508	614	1315	270	31	157
6	CHW	36	27.7	4.07	162	6077	5543	660	1400	279	32	155

Sorted number of home runs hit by a team:

95 105 109 111 117 118 122 123 123 125 125 128 131 132 134 136 142 146 147 150 152 155 155 155 156 157 163 177 186 211

Stemplot for team HR in 2014

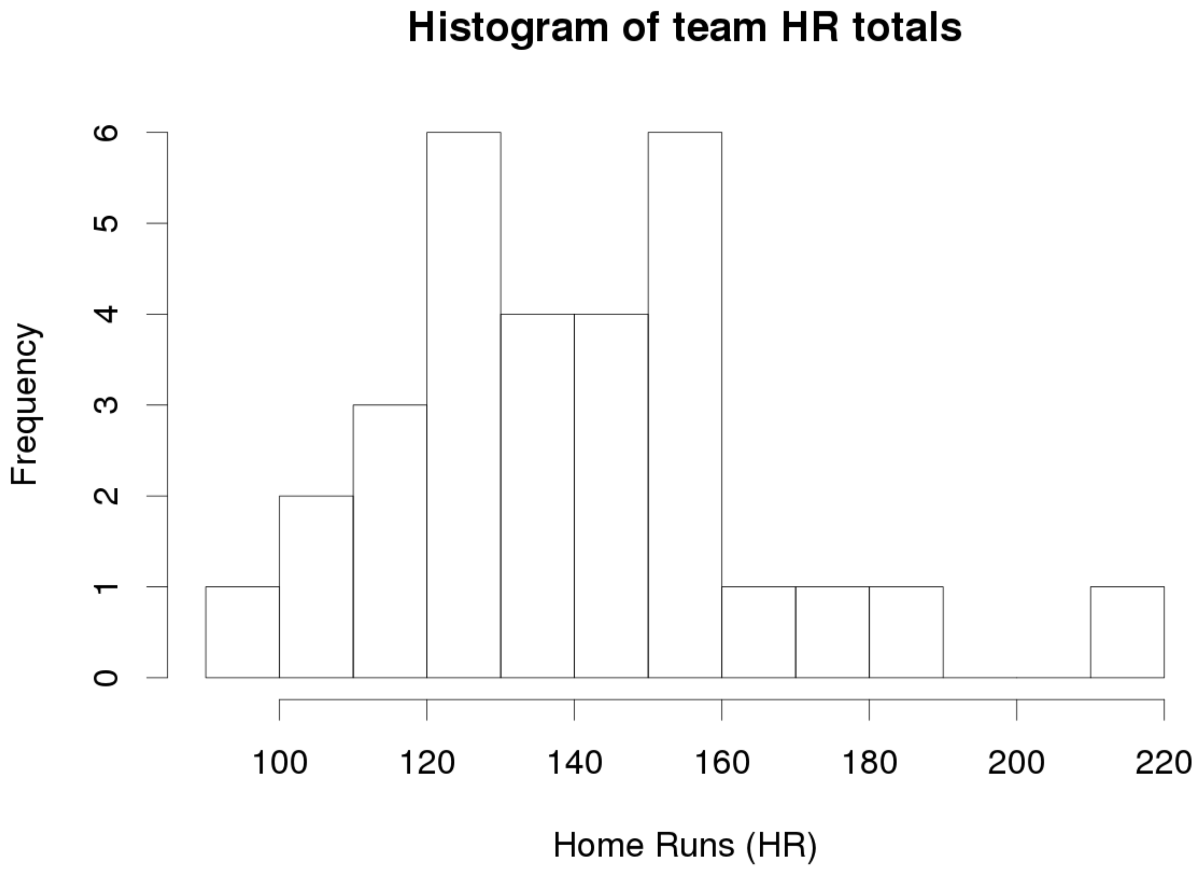
One way to get a sense of the shape of a distribution is to use a **stem plot**

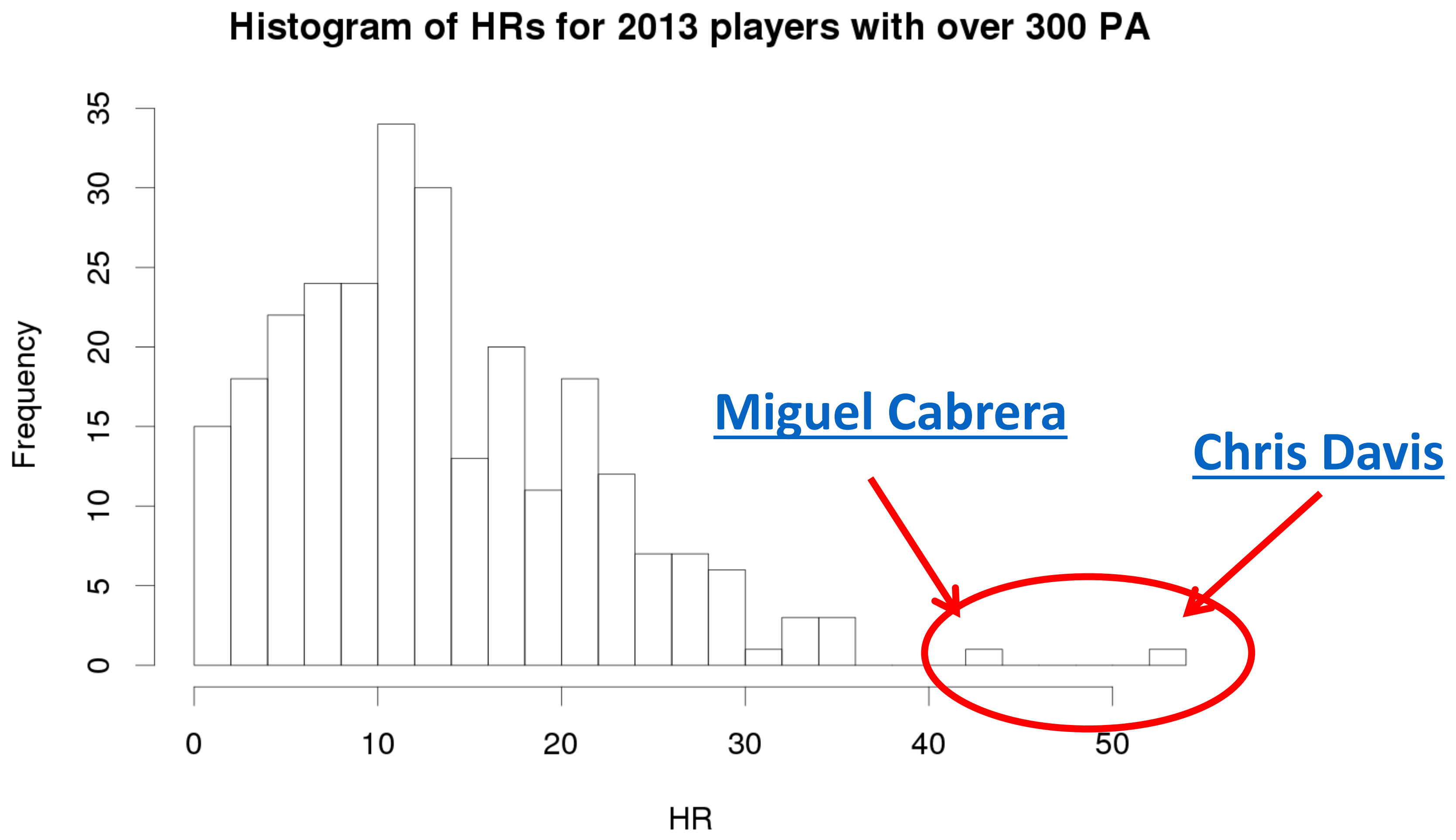
```
The decimal point is 1 digit(s) to the right of the |

 9 | 5
10 | 59
11 | 178
12 | 233558
13 | 1246
14 | 267
15 | 0255567
16 | 3
17 | 7
18 | 6
19 |
20 |
21 | 1
```

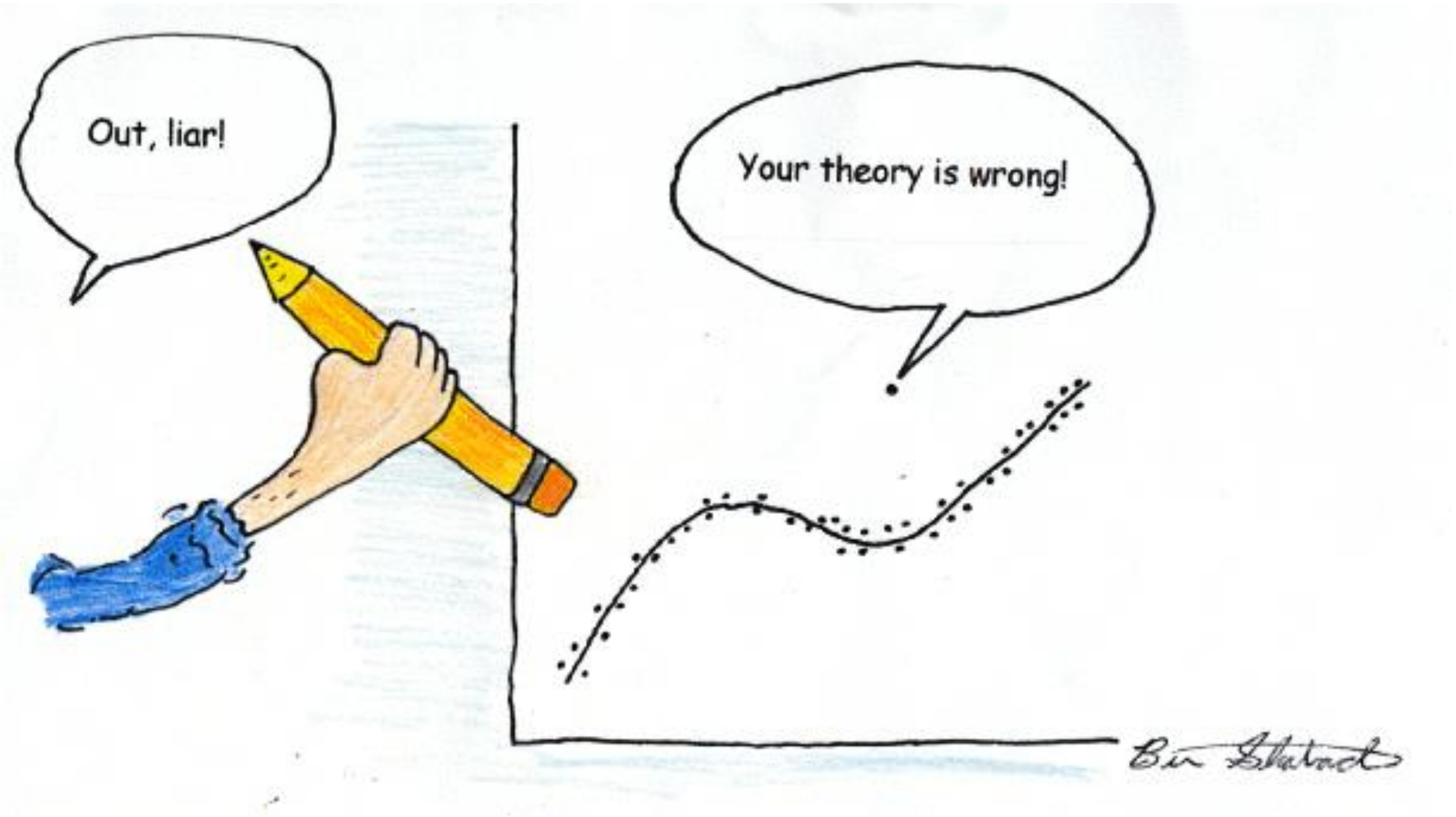
Histograms

Another related way to get a sense of the shape of a distribution is to use a **histogram**

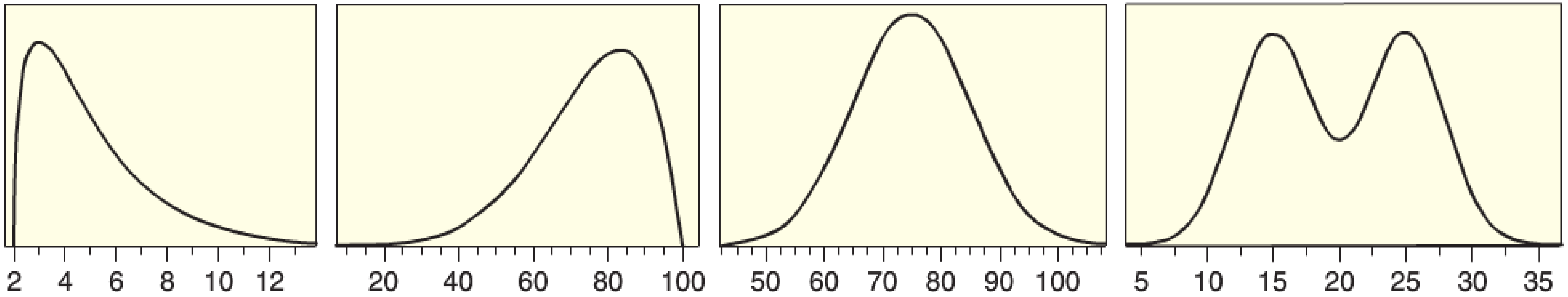




Observations about the distribution?


Out, liar!
Your theory is wrong!

Common shapes for Distributions



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

Try it in Python

Lab 1: section 4!

Reminder: Lab 1 is due on Monday February 15th at 11:30pm

Please turn a pdf in to Gradescope

Data manipulation code

From last class: datascience package

For a full list of Table functions see: http://data8.org/datascience/tables.html

Read in data into a table

- tb = Table.read_table('data.csv')

Table methods from last class:

- tb.show(5) # shows the first 5 rows of a Table
- tb.select() # select a subset of columns from a Table
- tb.take() # get a subset of rows from a Table
- tb.sum() # sums the values in a column
- tb.sort() # arrange the rows in a table based on the values in a column

Table object in the datascience package

Additional Table properties

- tb.num_rows # number of rows in a Table
- tb.num_columns # number of columns in a Table

Additional Table method to ‘filter’ data

gets a subset of rows that meet a particular criteria

- tb.where('col_name', value)

Arrays

The datascience package has wrapper for numpy

- make_array(4, 3, 5)   # creates a NumPy array

This is the same as:

- import numpy as np
- my_array2 = np.array([1, 2, 3])

These NumPy arrays have the same values but are not referring to the same object

- np.array_equal(my_array, my_array2)   # arrays have the same values
- my_array2 is my_array   # they do not refer to the same piece of memory

Comparing the datascience package with Pandas

Creating and viewing Tables/DataFrames

Description	datascience package	Pandas
Read in a csv file	tb = Table.read_table("data.csv")	df = pd.read_csv("data.csv")
Create Table/DataFrame	tb = Table().with_column("name", vals)	pd.DataFame(dict)
Get number of rows	tb.num_rows	df.shape[0]
Get number of columns	tb.num_columns	df.shape[1]
Show first 5 rows	tb.show(5)	df.head(5)

Selecting and filtering

Description	datascience package	Pandas
Select a single colum	tb.select("col1")	df.col1 also df["col1"]
Select multiple columns	tb.select("col1", "col2")	df[["col1", "col2"]]
Select 20th row	tb.take(20)	df.iloc[[20]]
Filtering rows equal to cond	tb.where("col", cond)	df[df.col == cond]
Sort descending by column	tb.sort("col", descending = True)	df.sort_values("col", ascending = False)

Statistics and visualization

Description	datascience package	Pandas
Get 90th percentile	tb.select("col").percentile(90)	df.col.quantile(.9)
Get min value	np.min(tb.select("col")).values[0][0]	np.min(df.col)
Create boxplot	tb.select("col").boxplot()	plt.boxplot(df.col)
Create histogram	tb.select("col").histogram	plt.hist(df.col, density = False)