

Summary statistics and plots
for a single batch of data

Outline for today

Review of Big Data Baseball chapter 2

Better know a player: David Ortiz

Central concepts in statistics: Populations, samples and statistical inference

Review and more R for data analysis:

- Plotting time series
- R Markdown
- Worksheet 1

Thoughts on BDB chapter 2?

Stadium effects – why are more home runs hit in some stadiums? (Julia)

Blaming the manager even though they have little control (Chris, Helen, James)

Statistical analyses vs. instincts (Aodhan)

Baseball players not living up to expectations (Henne, Matt)

Commitment to the game?

- Saying 'Baseball is just a game' does not show enough commitment (Ian)
- Hurdle changing lifestyle for the game shows a lot of commitment (Phillip)

Life

- A partner's role in changing one's life (Sheyla)
- Colors of emotions, and emotions toward Statistical analyses (Maddie)

Sports Illustrated Cover jinx and regression towards the mean (Ethan)

Better know a player: David Ortiz



David Ortiz (“Big Papi”) was born on November 18th, 1975 in the Dominican Republic

He has primarily played as a designated hitter (DH)

Slow career start



Major league debut on September 2nd 1997, for several years split his time between the major and minor leagues, and battled injuries

Twins released him in 2002 when they couldn't find a team willing to trade

Total of 58 home runs for the twins in 455 games over 6 seasons

Career with the Red Sox



Signed with the Red Sox as a free agent in 2003, played sparingly the first two months but took over for as Red Sox DH in June 1st 2003.

- In 2003: Hit 31 HR, batting average of .288

Career took off from there, big contributions to the regular and post season for the Red Sox

Career highlights

Three-time World Series Champion

Ten-time All-Star

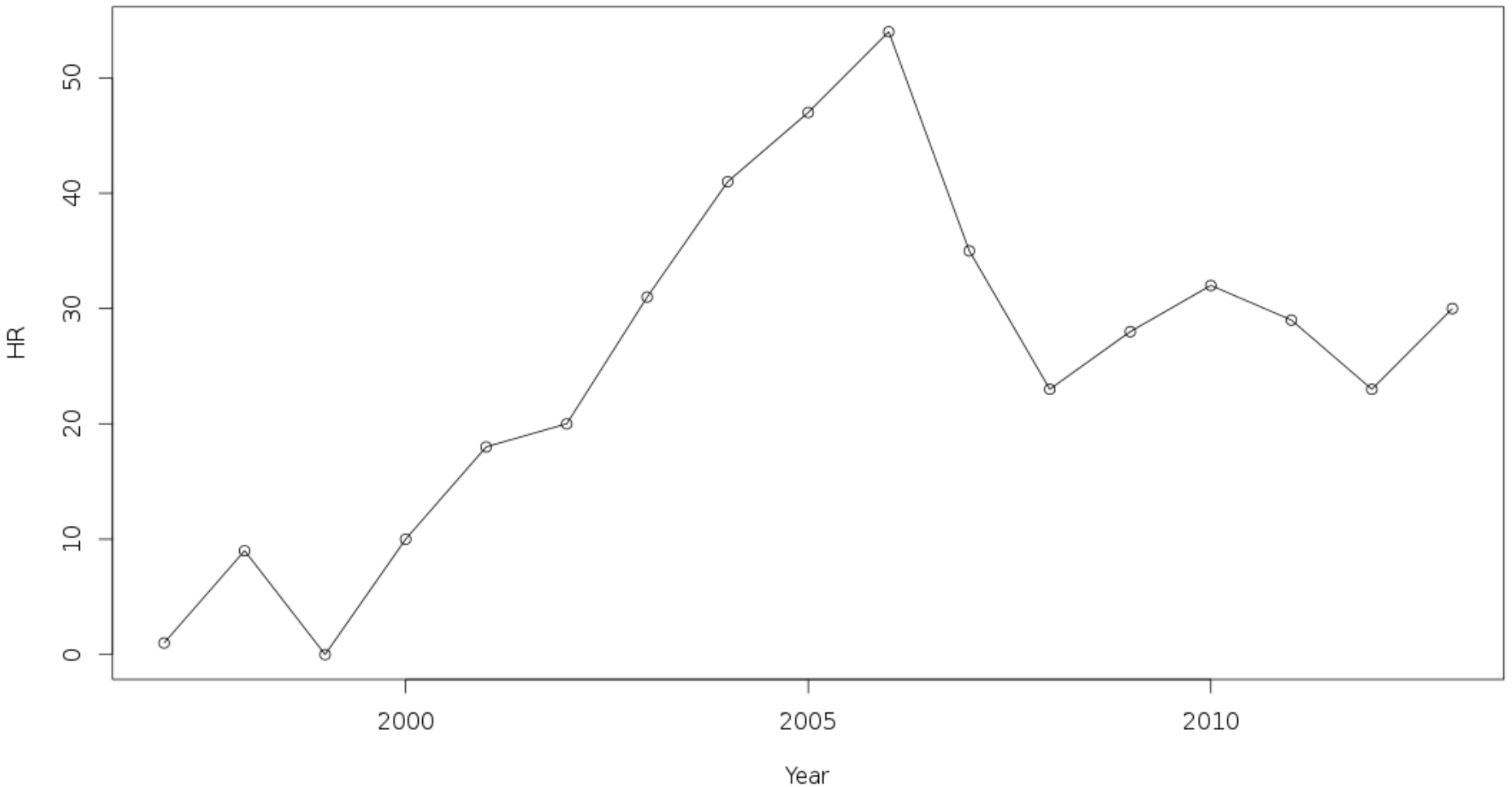
Red Sox single season HR record with 54 in 2006

- 17 on the MLB all time home run list (above Mickey Mantle)

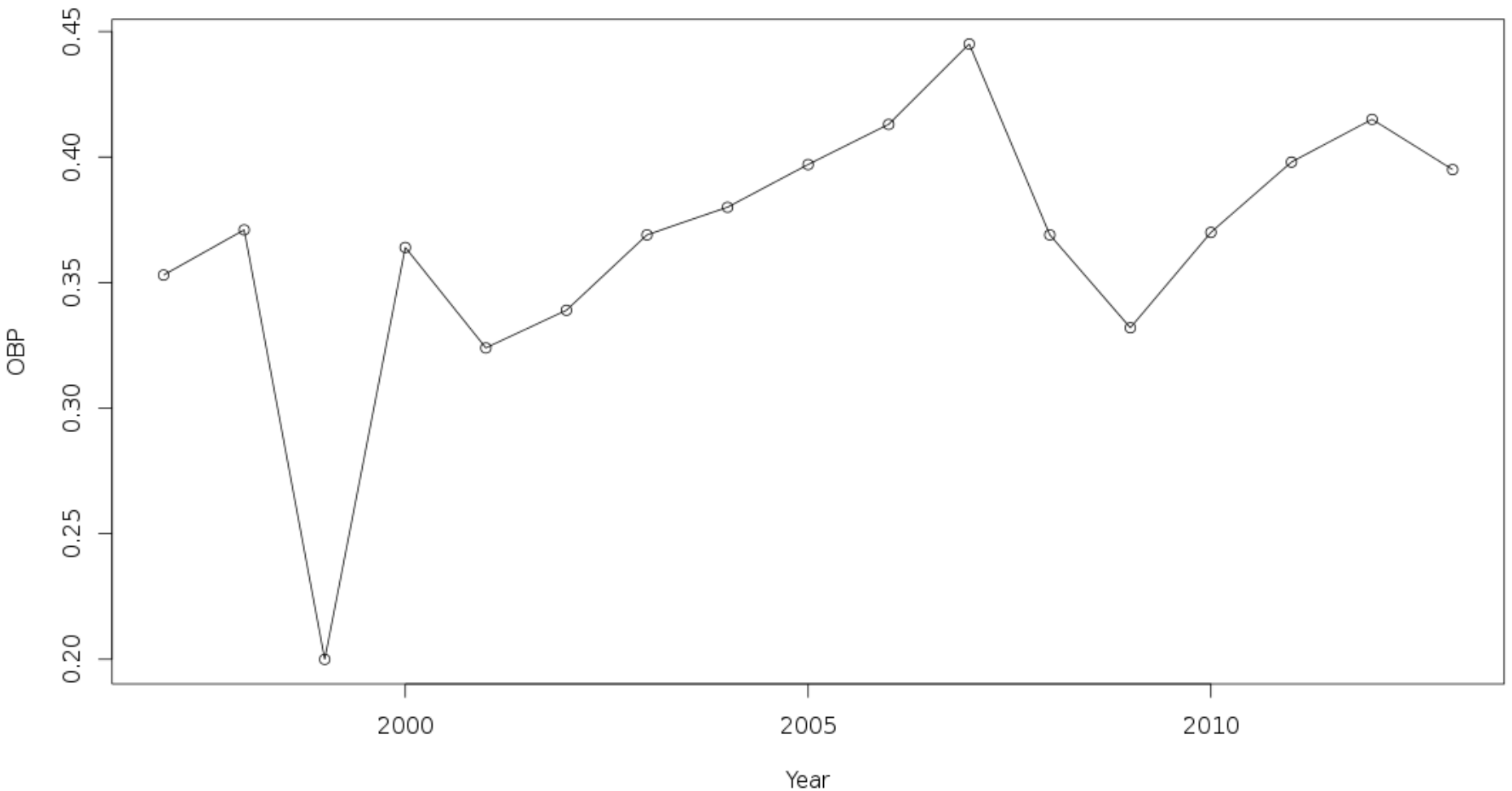
Clutch hitter?

- [Big hits against the Yankees](#)
- [2013 ALCS game 2](#)

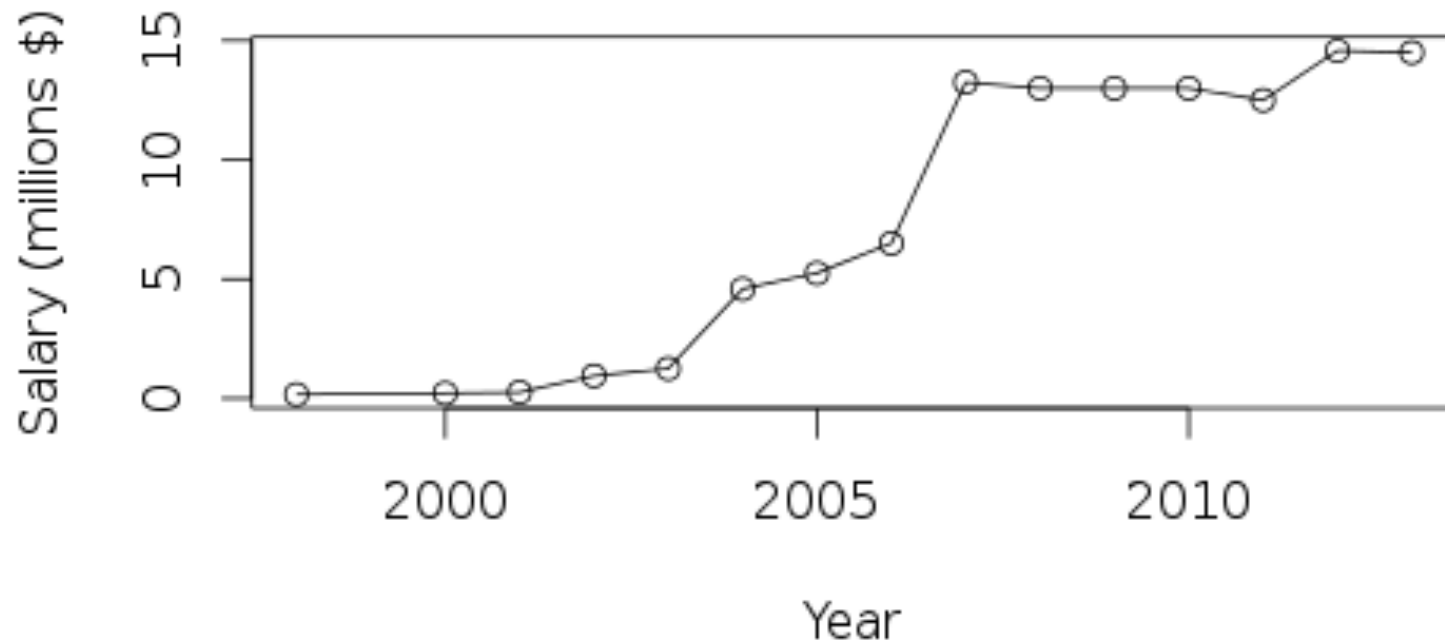
David Ortiz HR as a function of year



David Ortiz OBP as a function of year



Salary



Also received ~\$4.5 million in endorsements over the years

Samsung endorsement controversy



[Trip to the White House](#) April 1st 2014

Salsa



Retired this season



[Now drives for Lyft](#)

Presentation guidelines

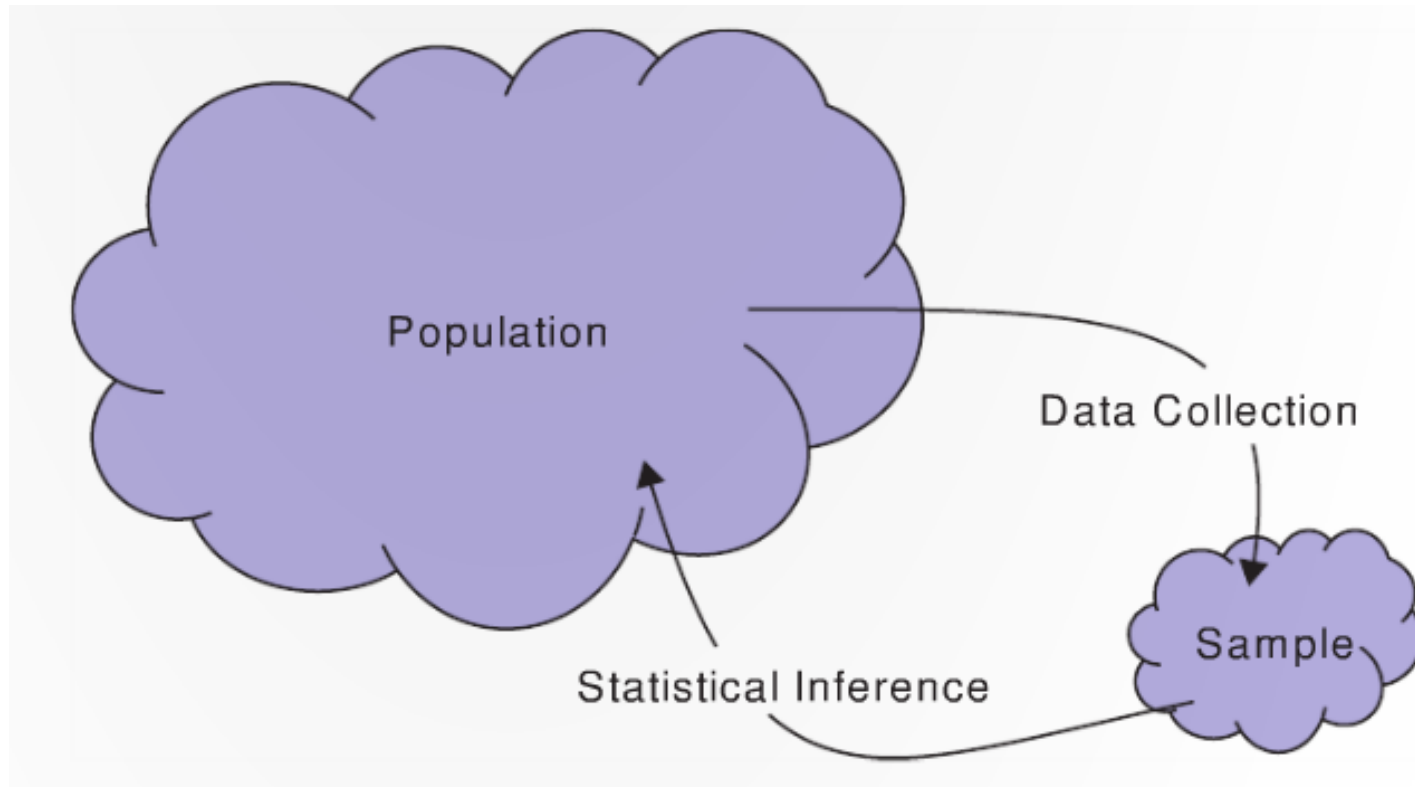
Pictures are better than words!

Keep it short: 5-10 slides, ~5 minute presentation

Keep it entertaining: focus on the personal and statistics!

Practice your talk!

Population and sample



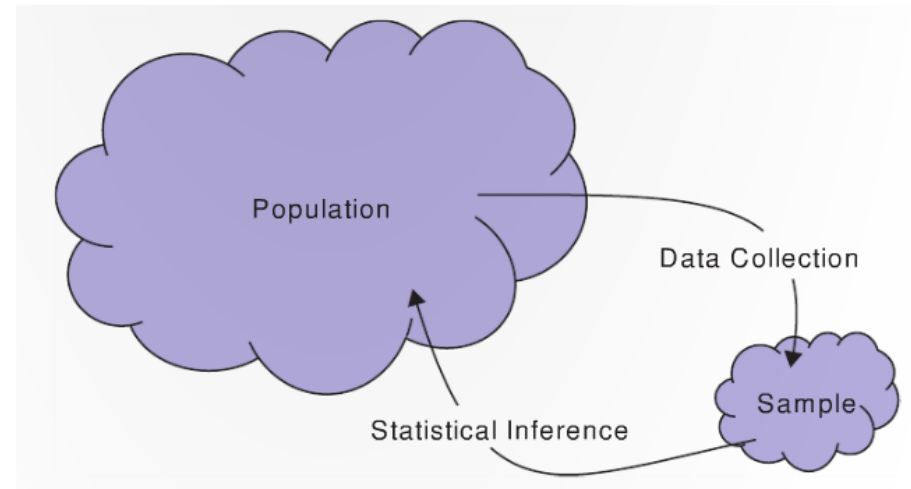
Sample from a Population

Population: all objects of interest (or an infinite processes)

Sample: A subset of the population

statistic: A number that is computed from data in a sample

Statistical inference: Making judgments about the population using data from the sample.



Examples of populations and samples

Population: all players who played in the major leagues from 1970 to 2015

Sample: the subset of players on baseball cards we have in this class

statistics (and statistical inference):

- Average (mean) number of games a player plays in a season
- Height of the tallest baseball player
- Number of hits (on average) player gets in a season

Flipping a coin



If I flip a coin 10 times and get 6 heads, do you think that the coin is fair?

i.e., observing a statistic from a sample of data (6/10) to infer something property of the coin (is it fair?)

Ability vs. performance

Ability: innate capacity of a baseball player

- Analogous to the bias of a coin
- e.g., if the player had played in infinitely many games, how well would he/she do?

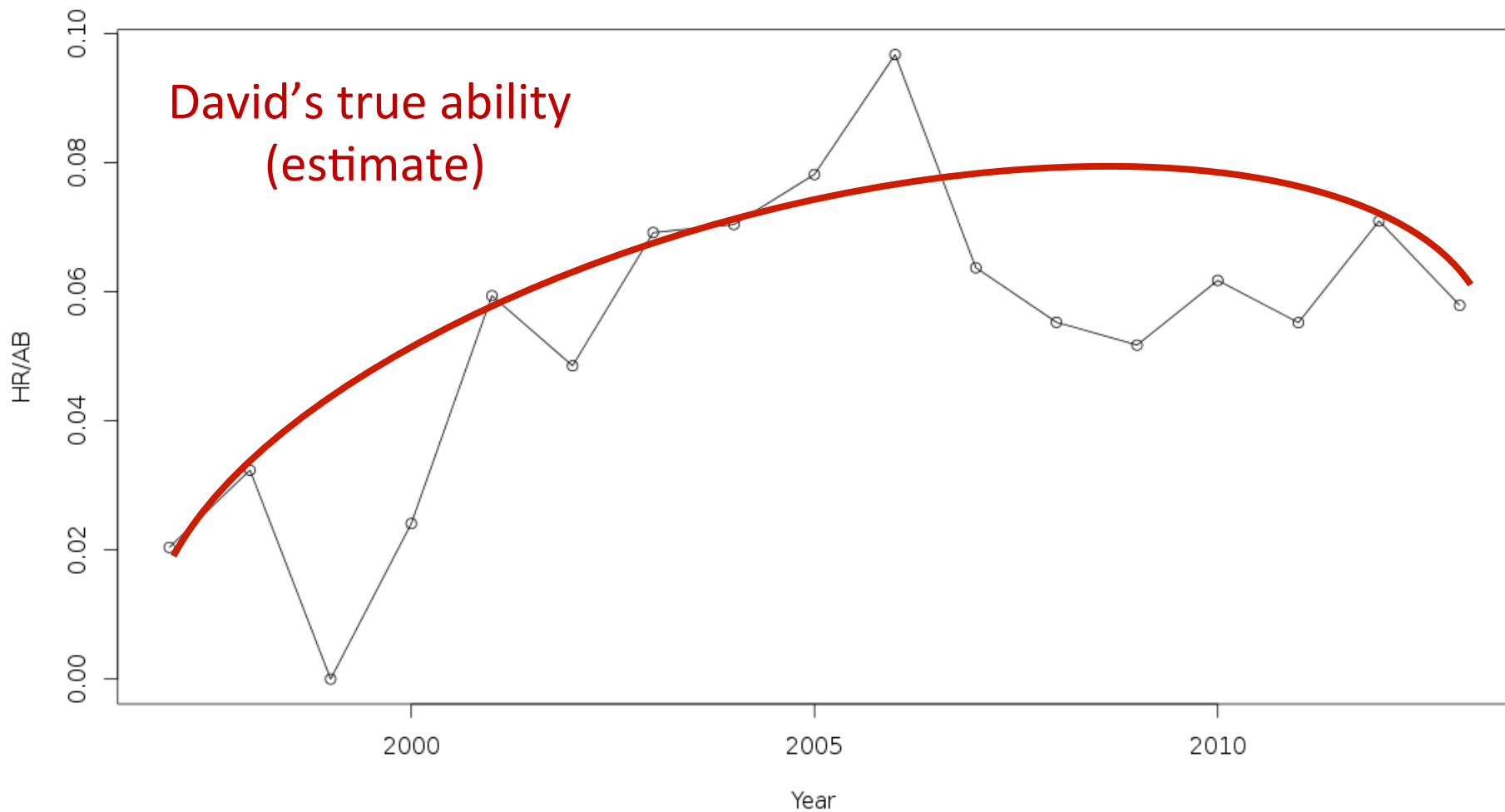
Performance: a statistic describing how the player performed

- Analogous to the observed number of heads when a coin is flipped

Statistical inference: inferring ability from performance

- i.e., looking at collected statistics (performance) and inferring ability
 - Should give insight into how the player will perform in the future

David Ortiz HR per AB



Describing and summarizing data

Descriptive statistics describe the sample of data you have

- E.g., description of performance

One can summarize the data with numbers or graphically

Descriptive statistics

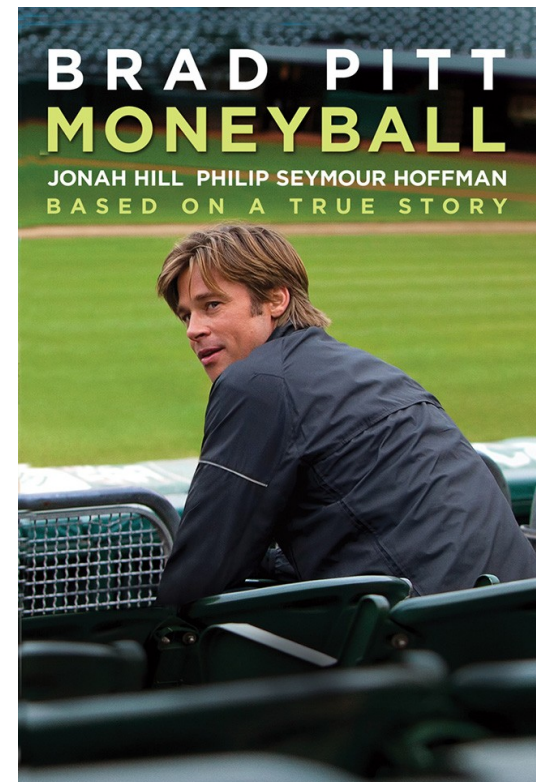
We will explore how to summarize and visual categorical and quantitative data next class...

Question: Does anyone know what the movie moneyball was rated?

A: PG-13

Q: Worst joke of the semester?

A: Unlikely



R review

Arithmetic: > 7 * 5

Assignment: > a <- 4

Functions: > sqrt(49) or 7^2

Vectors: > v <- c(5, 232, 5, 543)

Number journey:

- Create a vector v that contains the values 3, 4 and 5
- Square the values in this vector
- Sum the results

Answers:

```
> v <- c(3, 4, 5)
```

```
> v <- v^2
```

```
> sum(v)
```

```
[1] 50
```

Vector division

What would happen if you ran...

```
> v <- c(9, 18, 27)  
> v/3
```

What would happen if you ran...

```
> v1 <- c(9, 18, 36)  
> v2 <- c(3, 9, 12)  
> v1/v2
```

Let's look at a data frame

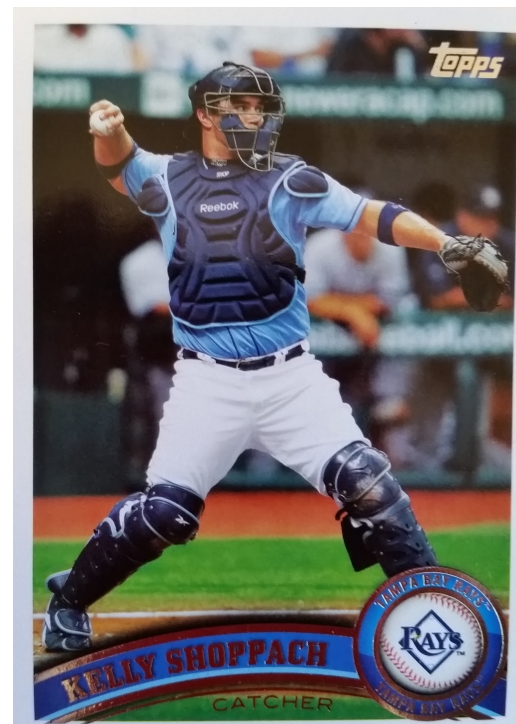
Load a function I wrote into R by typing:

```
source('/home/shared/baseball_stats_2017/  
baseball_class_functions.R')
```

If you load this correctly you should have a function in your Global Environment called `get.Lahman.batting.data()`

Let's look at a data frame

	playerID	nameFirst	nameLast	yearID	stint	teamID	lgID	G	AB
1	shoppke01	Kelly	Shoppach	2005	1	BOS	AL	9	15
2	shoppke01	Kelly	Shoppach	2006	1	CLE	AL	41	110
3	shoppke01	Kelly	Shoppach	2007	1	CLE	AL	59	161
4	shoppke01	Kelly	Shoppach	2008	1	CLE	AL	112	352
5	shoppke01	Kelly	Shoppach	2009	1	CLE	AL	89	271
6	shoppke01	Kelly	Shoppach	2010	1	TBA	AL	63	158
7	shoppke01	Kelly	Shoppach	2011	1	TBA	AL	87	221
8	shoppke01	Kelly	Shoppach	2012	2	NYN	NL	28	79
9	shoppke01	Kelly	Shoppach	2012	1	BOS	AL	48	140
10	shoppke01	Kelly	Shoppach	2013	2	CLE	AL	1	2
11	shoppke01	Kelly	Shoppach	2013	1	SEA	AL	35	107



Use this function to get batting data on a specific player:

```
> card.data <- get.Lahman.batting.data("Kelly", "Shoppach")  
> View(card.data) # The View() function only works in the console!
```

Let's look at a data frame

	playerID	nameFirst	nameLast	yearID	stint	teamID	lgID	G	AB
1	shoppke01	Kelly	Shoppach	2005	1	BOS	AL	9	15
2	shoppke01	Kelly	Shoppach	2006	1	CLE	AL	41	110
3	shoppke01	Kelly	Shoppach	2007	1	CLE	AL	59	161
4	shoppke01	Kelly	Shoppach	2008	1	CLE	AL	112	352
5	shoppke01	Kelly	Shoppach	2009	1	CLE	AL	89	271
6	shoppke01	Kelly	Shoppach	2010	1	TBA	AL	63	158
7	shoppke01	Kelly	Shoppach	2011	1	TBA	AL	87	221
8	shoppke01	Kelly	Shoppach	2012	2	NYN	NL	28	79
9	shoppke01	Kelly	Shoppach	2012	1	BOS	AL	48	140
10	shoppke01	Kelly	Shoppach	2013	2	CLE	AL	1	2
11	shoppke01	Kelly	Shoppach	2013	1	SEA	AL	35	107



Getting number of games (G) Kelly played each season:

```
> card.data$G
```

```
[1] 9 41 59 112 89 63 87 28 48 35 1
```

Computing statistics

One compute statistics on vectors (columns of a data frame)

```
> sum(card.data$G)  
[1] 572
```

Or we can assign vectors in a data frame to an object

```
> games <- card.data$G  
> games
```

Comments

We can use the # symbol to leave comments that describe what our code does

For example:

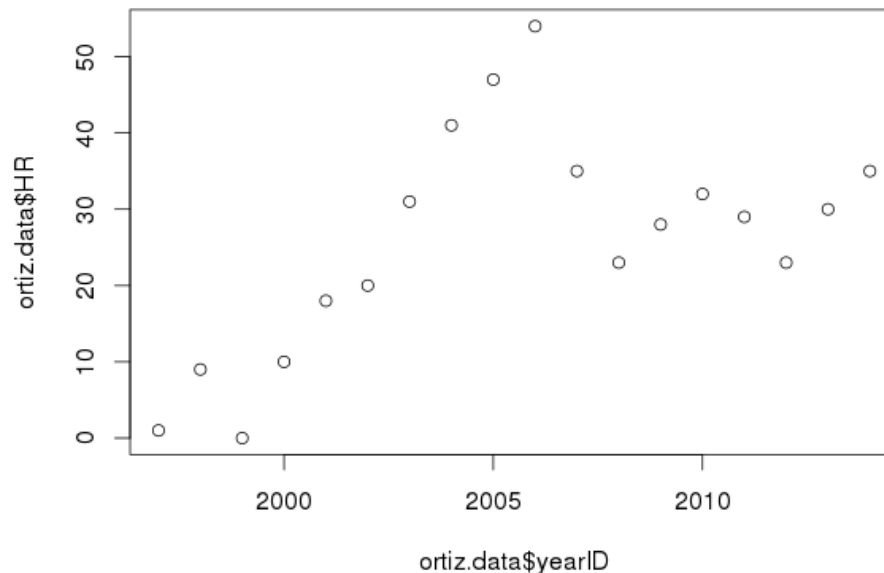
```
# this code adds two and two  
> 2 + 2
```

Please leave comments so that readers can understand your code better!

Plotting time series

We can use the `plot(x, y)` function to plot a vector of values `y`, as function of the vector of values in `x`

- > `ortiz.data <- get.Lahman.batting.data("David", "Ortiz")`
- > `plot(ortiz.data$yearID, ortiz.data$HR)`



Plotting time series

Usually we want to connect the lines in such a plot

```
> plot(ortiz.data$yearID, ortiz.data$HR, type = "o")
```



Even better

```
> plot(ortiz.data$yearID, ortiz.data$HR, type = "o", xlab = "Year", ylab = "HR")
```

Worksheet 1!

A. Accessing R with the Hampshire server:

<https://asterius.hampshire.edu>

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

```
> source('/home/shared/baseball_stats_2017/baseball_class_functions.R')  
> get.worksheet(1)
```

RMarkdown

RMarkdown (.Rmd files) allow you to embed written descriptions, R code and the output of that code into a nice looking document

Everything in R chunks is executed as code:

```
```${r}
 # this is a comment
 # the following code will be executed
 2 + 3
```
```

Everything outside R chunks appears as text

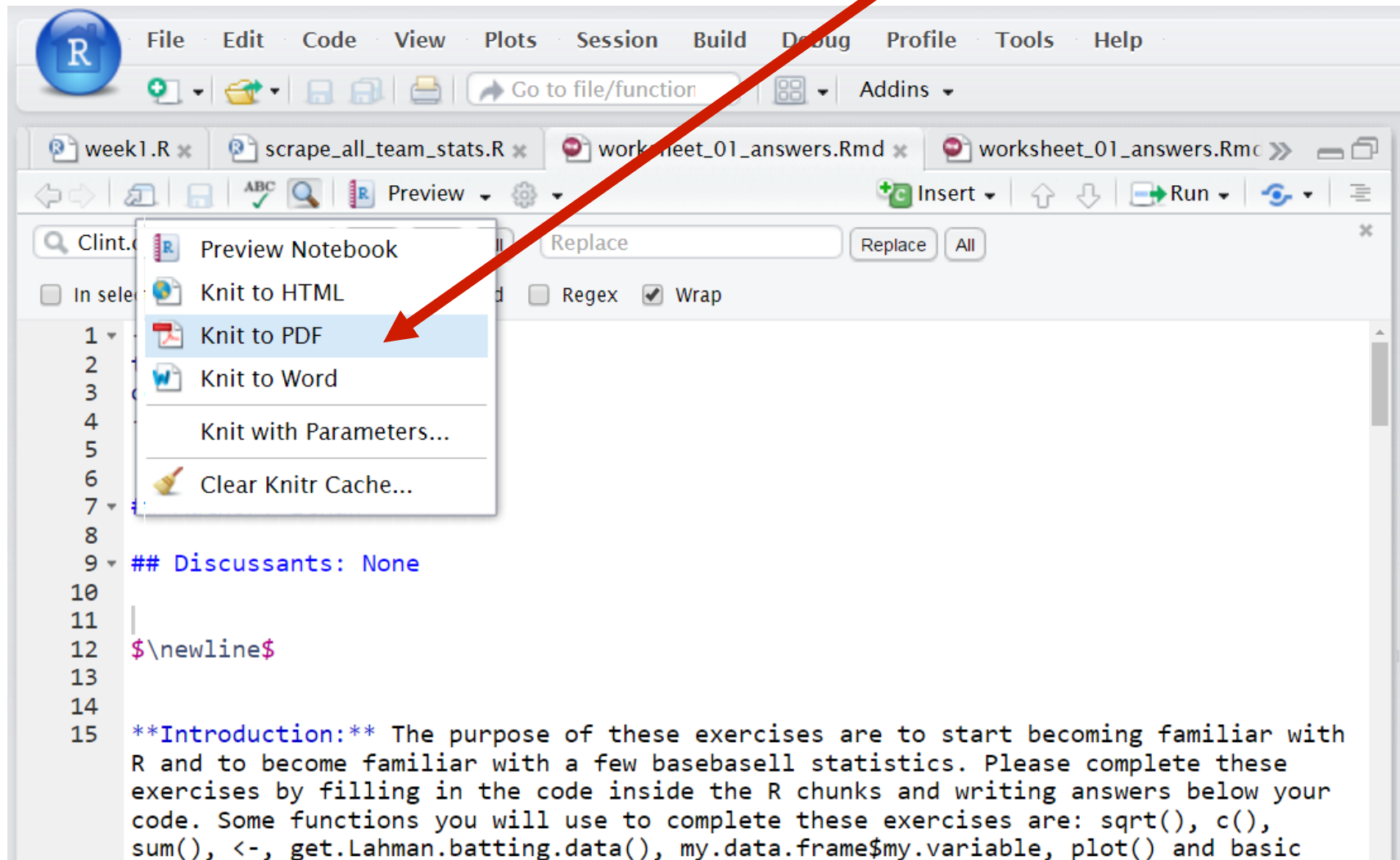
RMarkdown

Note: Rmarkdown files do not have access to variables in the global environment, but instead have their own environment.

Why is this a good thing???

Knitting to a pdf

Turn in a pdf of your solutions to Moodle



Avoid hard to debug code!

Only change a few lines at a time and then knit your document to make sure everything is working!

Comment out parts of the code that isn't working (using the # symbol) until you can find the line of code that is giving the error message

Questions? Use Piazza!

We will use Piazza for questions that arise on the homework:

<https://piazza.com/class/iy3nflk2izi6np>

Please post any questions you have to this site and help out by answering other student's questions

Worksheet 1 is due on midnight on Sunday Feb 5th

- All worksheets need to be turned in on time!

TA office hours: 1-3pm on the bridge