Quantifying central tendency and variability

Outline for today

Big Data Baseball – chapter 3

Better know a player: Bo Jackson

Review:

- Plotting categorical and quantitative data

More descriptive statistics:

- Measures of central tendency for quantitative data
- Measures of variability

Worksheet 2: due midnight Wednesday Feb 15th

DataFest 2017

- March 31st to April 2nd

Amherst is having pre-DataFest workshops

- First one is at 7:30 on February 15th

If you are interested in participating in DataFest let me know

Interested in data?
Want to learn some skills with R?
Prepping for DataFest?

Workshops in Statistics and R

Brought to you by the Statistics Fellows

**** Data Wrangling in R ****
Wednesday, February 15th - 7:30-8:30 pm in SM 204
(preceded by *DataFest Information Session 7-7:30*)
Hosted by Paul and Jonathan

**** Regression in R ****
Wednesday, February 22nd - 7-8 pm in SM 204
Hosted by Silvia, Leonard, and Connor

**** Intro to ggplot2 in R ****
Wednesday, March 1st - 7-8 pm in SM 204
Hosted by Pei and Sarah

**** Intro to Machine Learning in R ****
Monday, March 20th - 7-8 pm in SM 204
Hosted by Azka, Caleb, and Muling

Feel free to bring your laptops to follow along with the Workshops.
Open to all interested students, staff, and faculty.

Big Data Baseball Chapter 3

- Thoughts?

Better know a player

Bo Jackson

Review

Categorical and quantitative data

Descriptive statistics describes the sample of data you have

Categorical variables: fall into distinct categories

E.g., team (Red Sox, Yankees, Mets, etc.)

Quantitative variables: numerical data

E.g., Number of home runs

Categorical variables: proportion

The **proportion** of a category is found by:

$$\text{Proportion of category} = \frac{\text{Number in that category}}{\text{total number}}$$

Example: proportion of hits that are home runs

160 hits total

30/160 = .1875

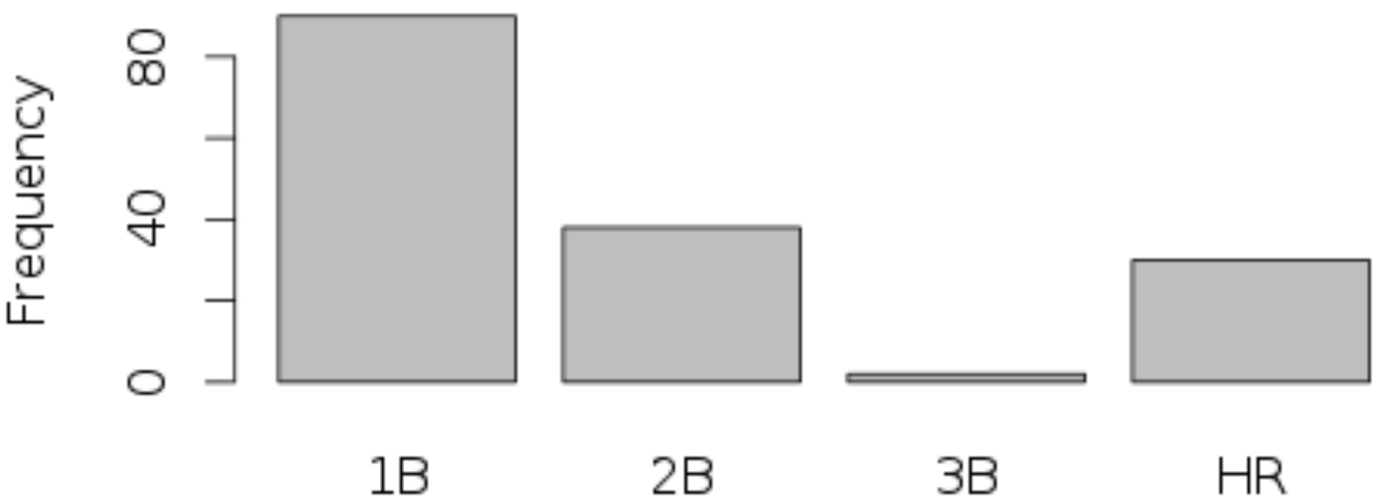
	1B	2B	3B	HR
Count	90	38	2	30
Proportion	0.56	0.24	0.01	0.19

```
> hit.types <- c(90, 38, 2, 30)
> hit.types/sum(hit.types)
```

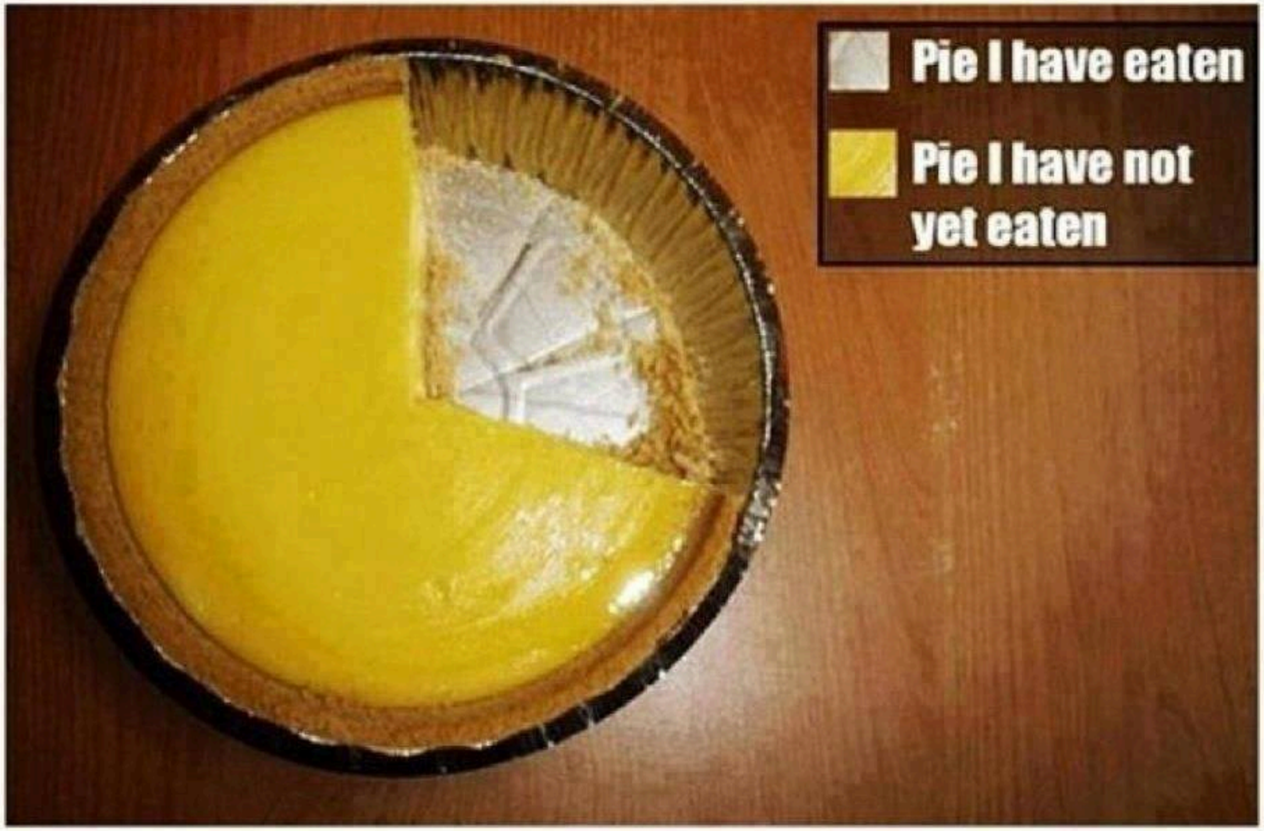
Plotting categorical data

R: barplot(x)

R: pie(x)



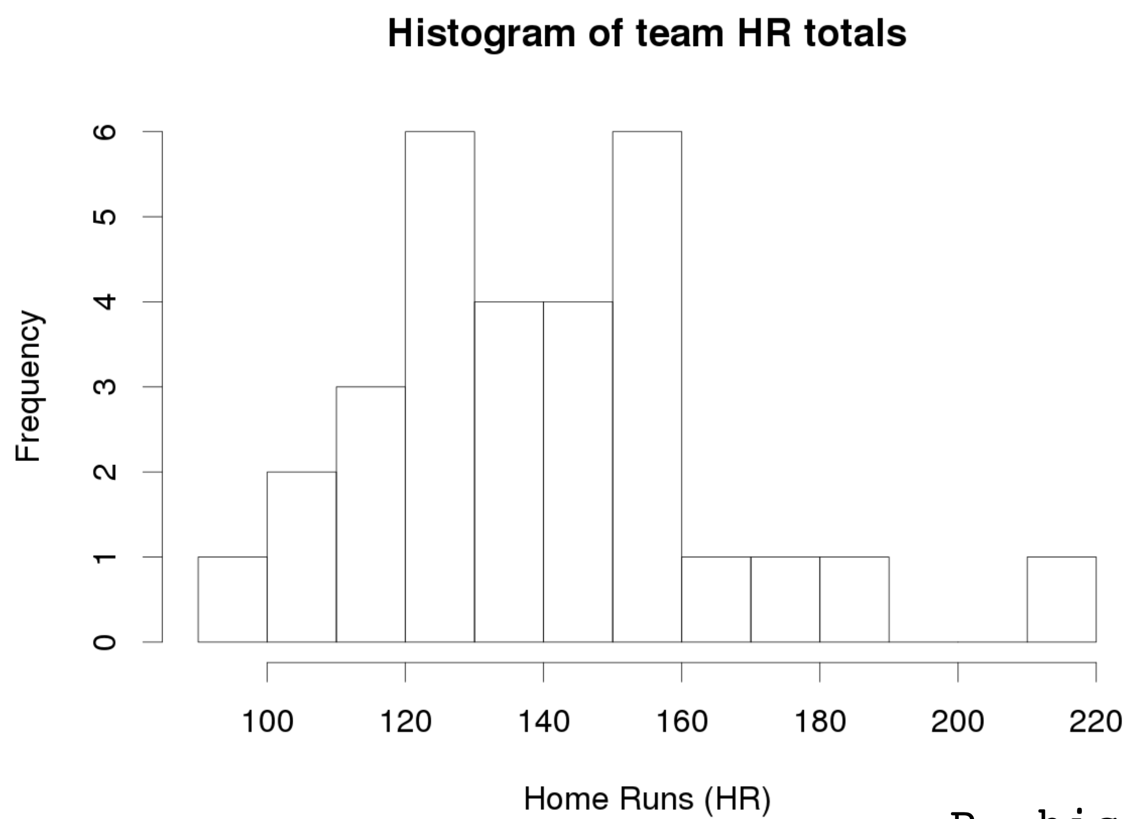
World's Most Accurate Pie Chart



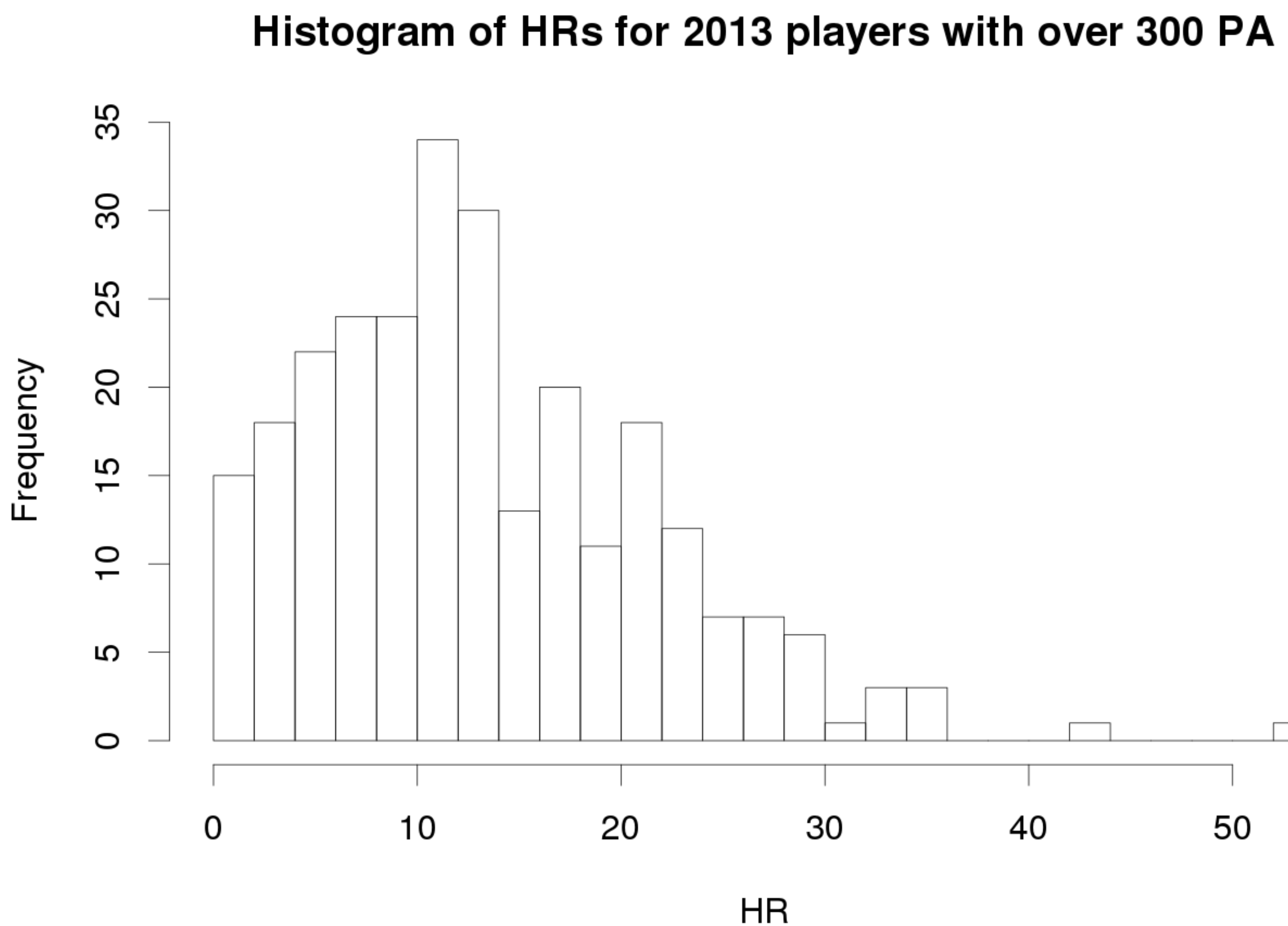
Describing quantitative variables

```
> stem(team.data.2014$HR, scale = 2)

  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
```

R: stem(x)



R: hist(x)

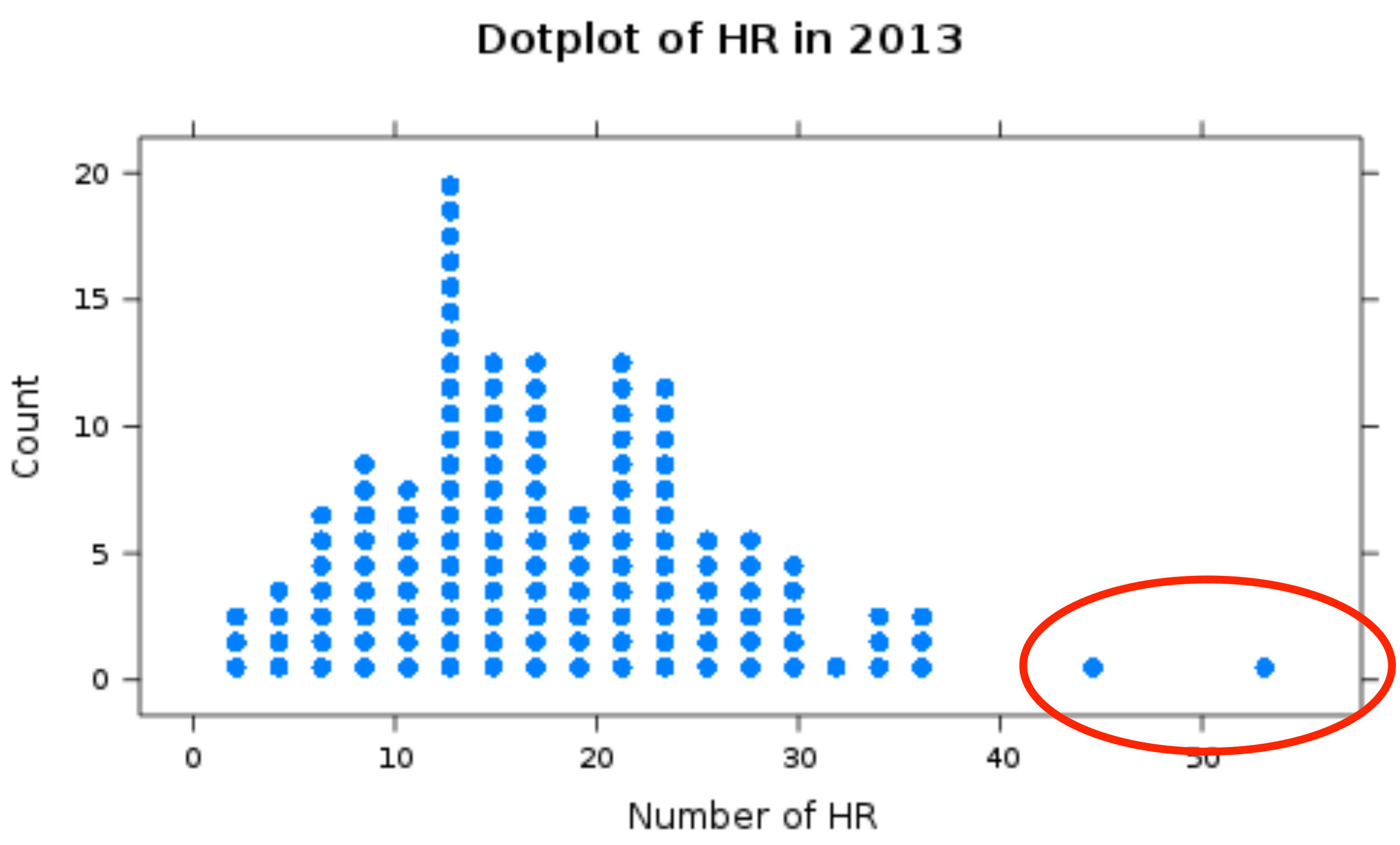



```
hist(player.data.2013$HR, n = 30, xlab = "HR",
     main = "Histogram of HRs for 2013 players with over 300 PA")
```

Observations about the distribution?

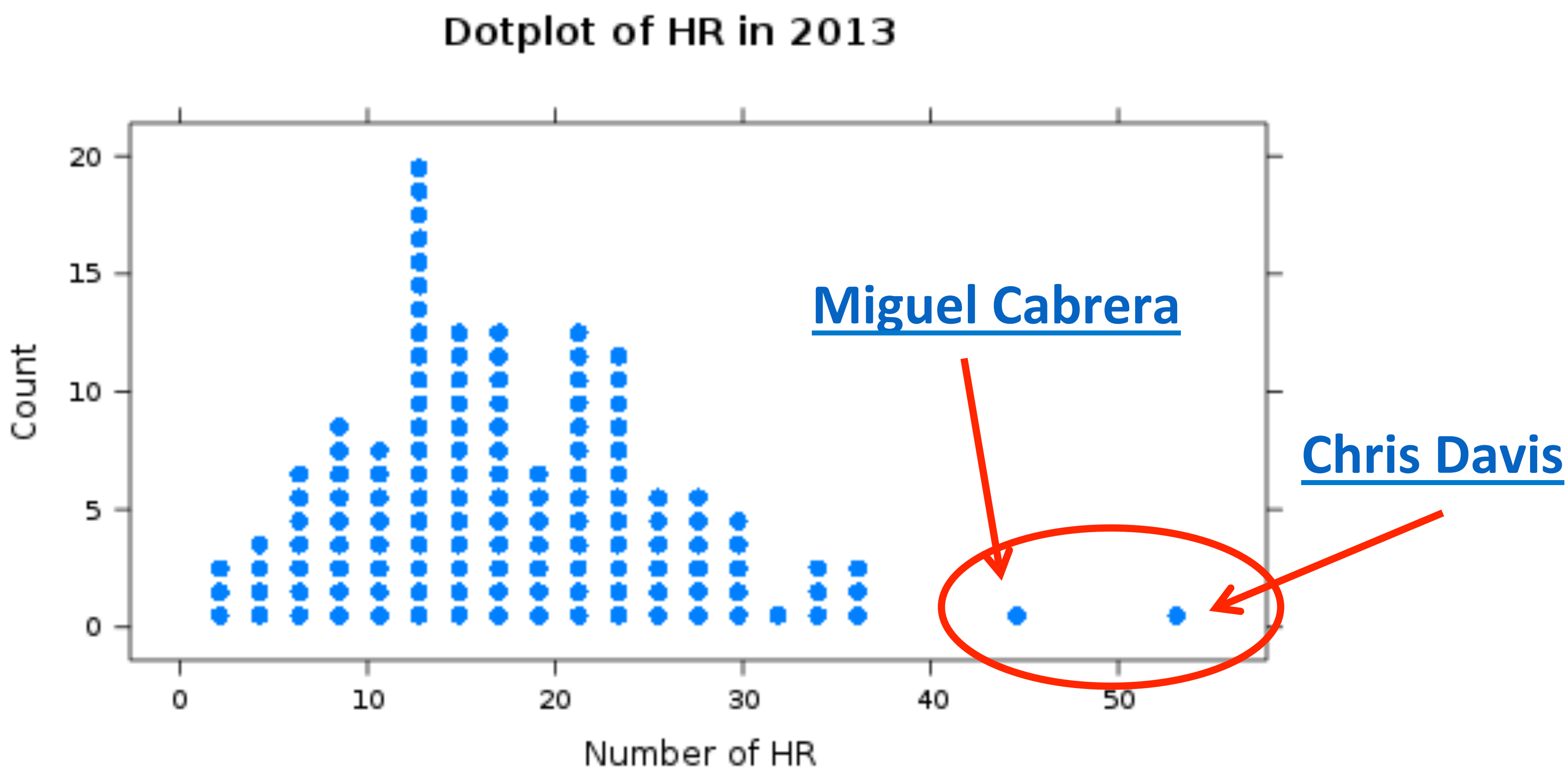
Dotplot for individuals' HR in 2013

When we have ***discrete data***, we can also use **dotplots** to get a sense of how the data is distributed

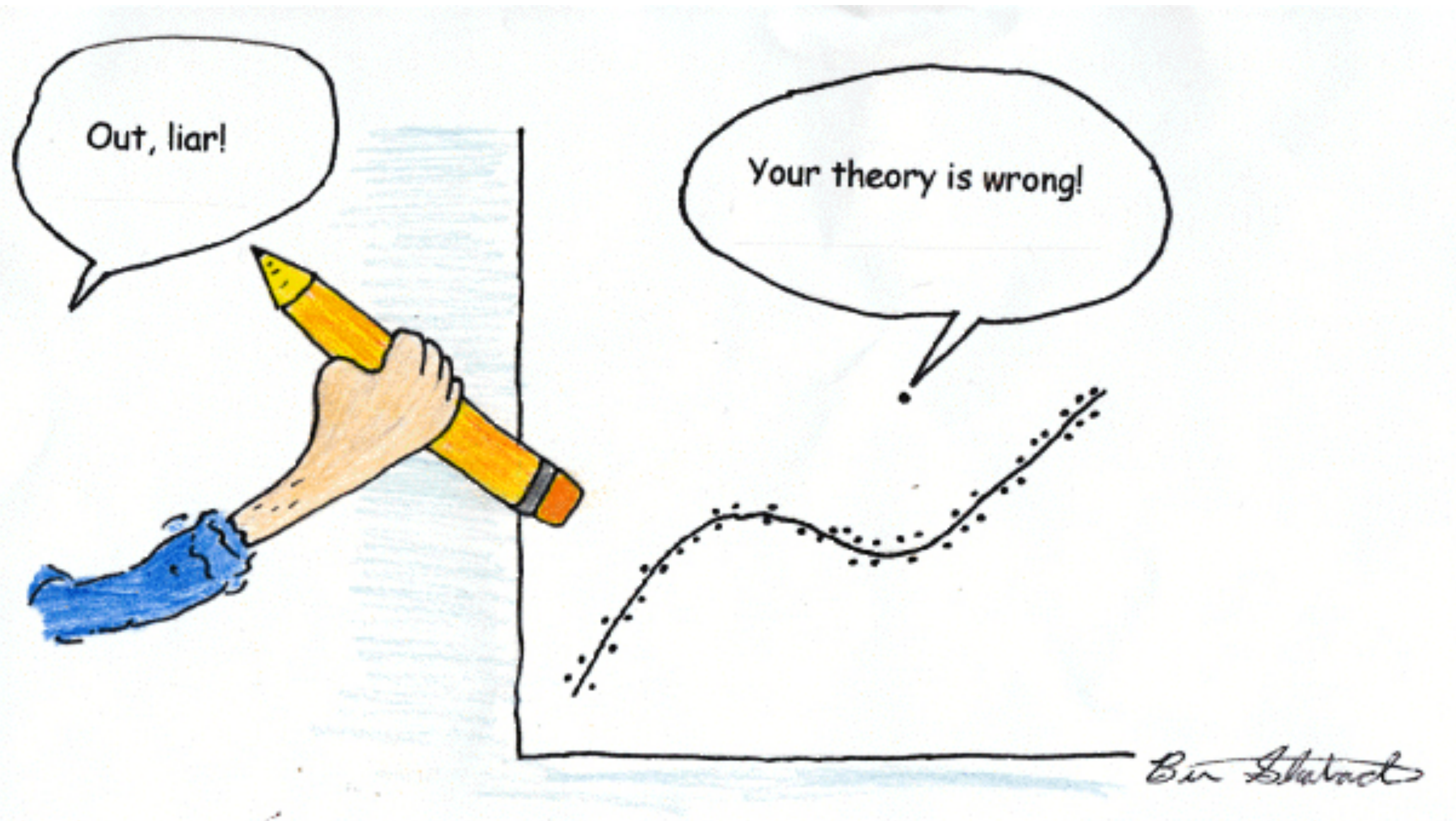


Dotplot for individuals HR in 2013

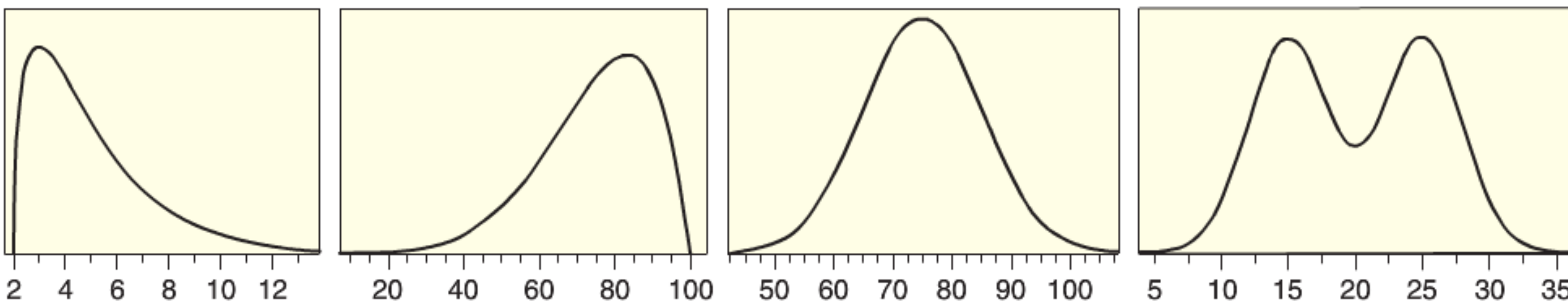
One way to get a sense of the shape of a distribution is to use a **dotplot**



R: mosaic::dotPlot(x)


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

statistics measuring the center of distribution

Graphs are useful for visualizing data to get a sense of what of what the data look like

We can also summarize data numerically

A numerical summary (function) of sample is called **statistic**

Two important statistics that can be used to describe the center of the data are the **mean** and the **median**.

The mean

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\Sigma x_i}{n}$$

R: `mean(x)`

If you data has missing values use:

```
mean(x, na.rm = TRUE)
```

Mean number of games played (G)

Can you calculate the mean number of games played for players who had 300 plate appearances in 2013

```
> players.2013 <- get.Lahman.batting.data(year = 2013, min.PA = 300)
```

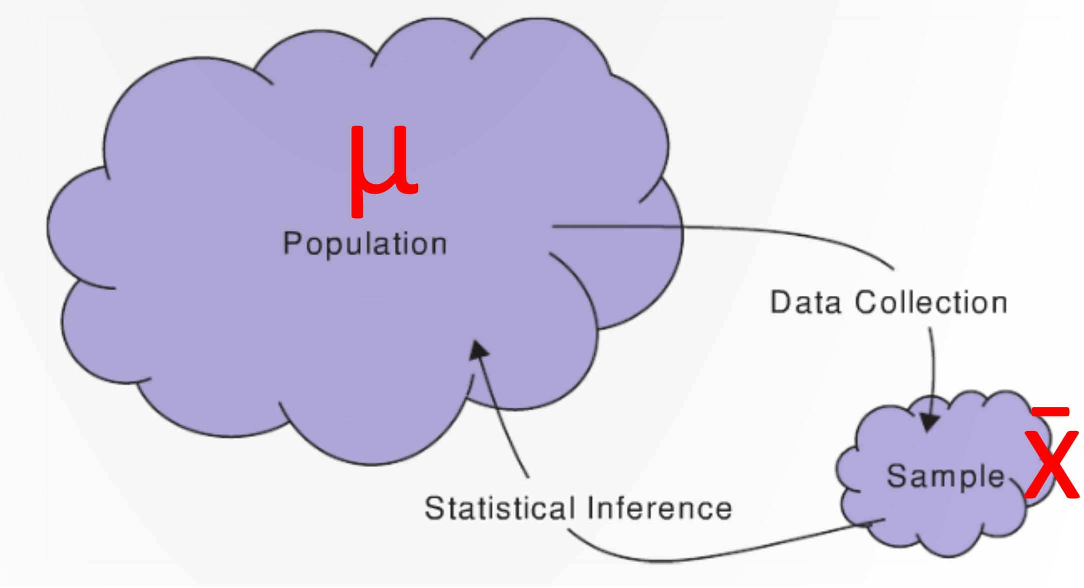
```
> mean(players.2013$G)
```

Do you think the mean number of games played would be higher if we calculated it from only players who had 500 plate appearances?

Sample vs. Population mean

The mean for a ***sample*** is denoted $\bar{x}$ (pronounced “x-bar”)

The mean for a ***population*** is denoted μ, which is the Greek letter “mu”



Give the proper notation: μ vs. $\bar{x}$?

We randomly select 50 baseball players and take their mean height?

We look at all professional baseball players and take their mean height?

The median

The **median** of a data set of size n is

- If n is odd: The middle value of the sorted data
- If n is even: The average of the middle two values of the sorted data

The median splits the data in half

`R: median(x)`

Resistance

We say that a statistics is **resistant** if it is relatively unaffected by extreme values (outliers).

The median is resistant when the mean is not

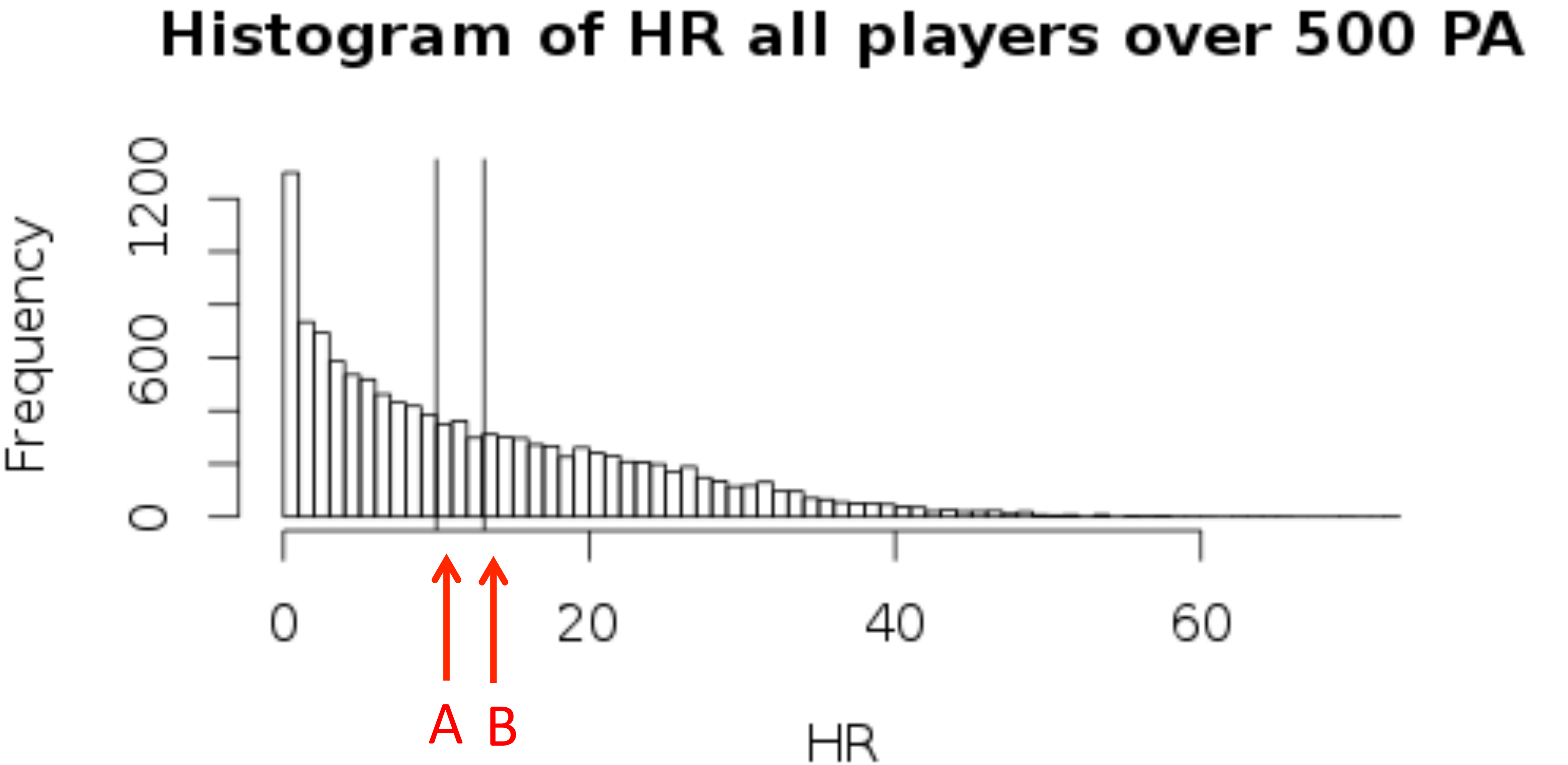
Football player salary examples

Some NFL football players are paid a lot more than others
(star quarterbacks can be paid more than $20 million)

Mean NFL salary = $1.87 million
Median NFL salary = $838,000

Mean and median salary for all US workers?
Distribution of salaries for US workers?

Which is the mean A or B?



Summary statistics quantifying the spread of quantitative variables

Standard deviation

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

```
R: sd(x)
```

In class worksheet: computing the standard deviation

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

Values	Deviations	Squared Deviations
54	20.91	437.19
35	1.91	3.64

Number of home runs David Ortiz had in the last 11 seasons:

54	35	23	28	32
29	23	30	35	37
38				