More descriptive statistics:

Percentiles, boxplots, and z-scores

Outline for today

Better know a player Wade Boggs

Review:

- Central tendency and measures of variation

More descriptive statistics:

- Percentiles, 5 number summaries, boxplots
- Z-scores

Better a player

Wade Boggs

Any questions about worksheet 2?

Descriptive statistics

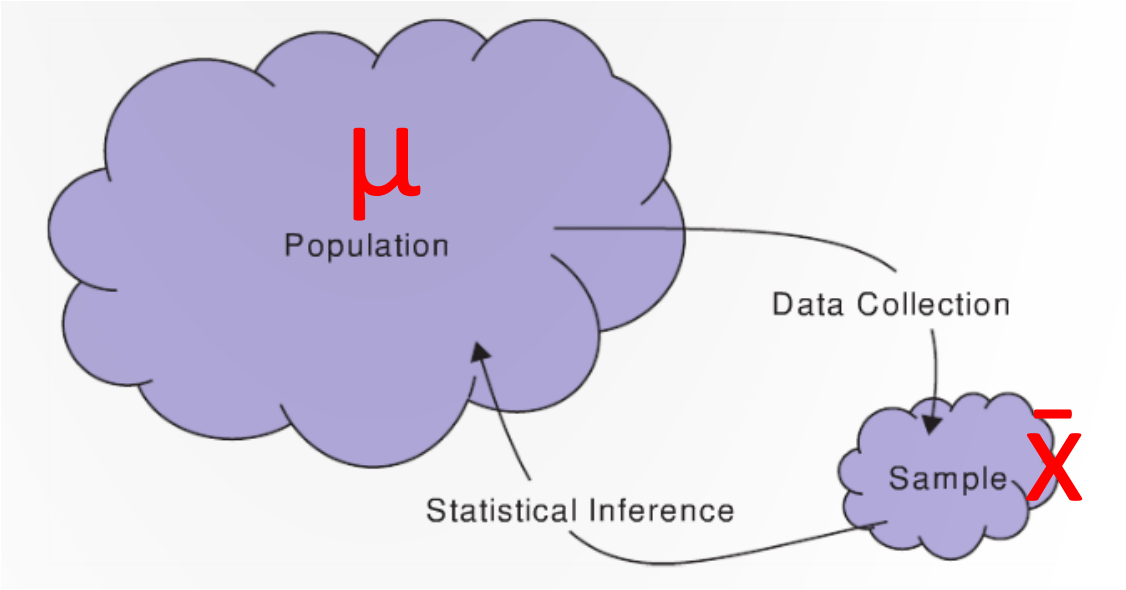
What is a **statistic**?

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

The mean

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

Sample mean ($\bar{x}$) vs. population mean (μ)



The median

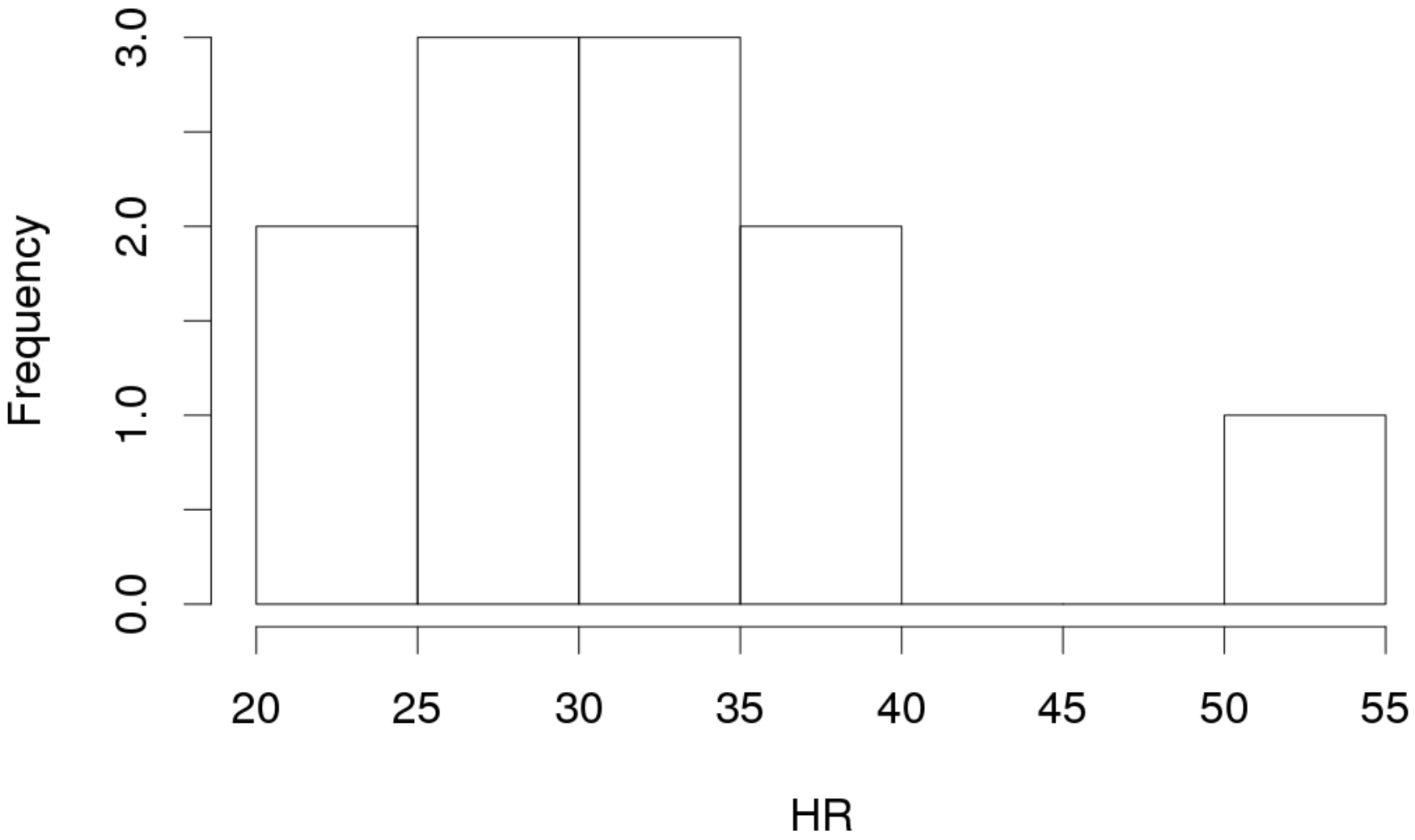
The **median** is the value in the middle of your data

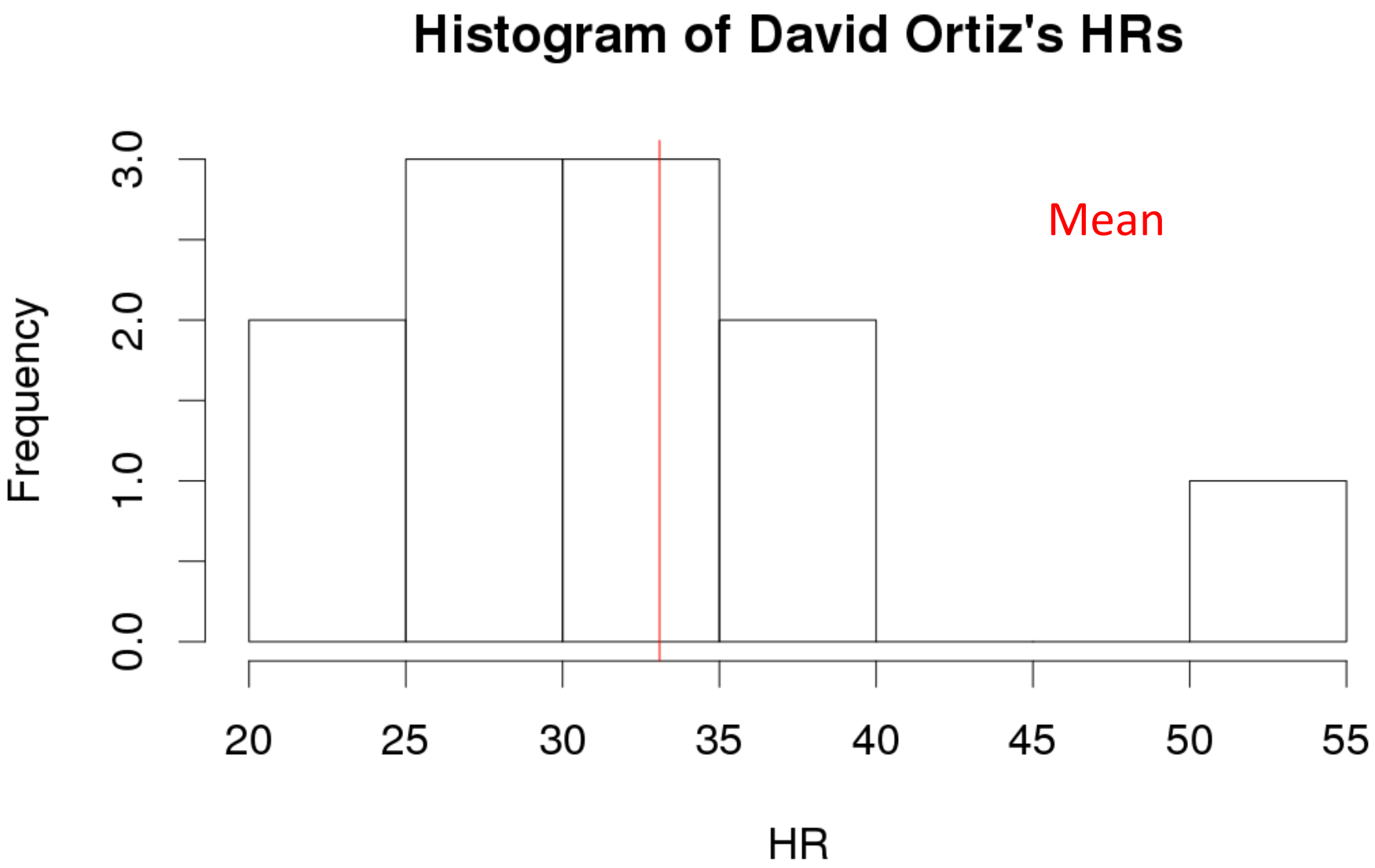
- ½ of the values are greater than the median and ½ are less

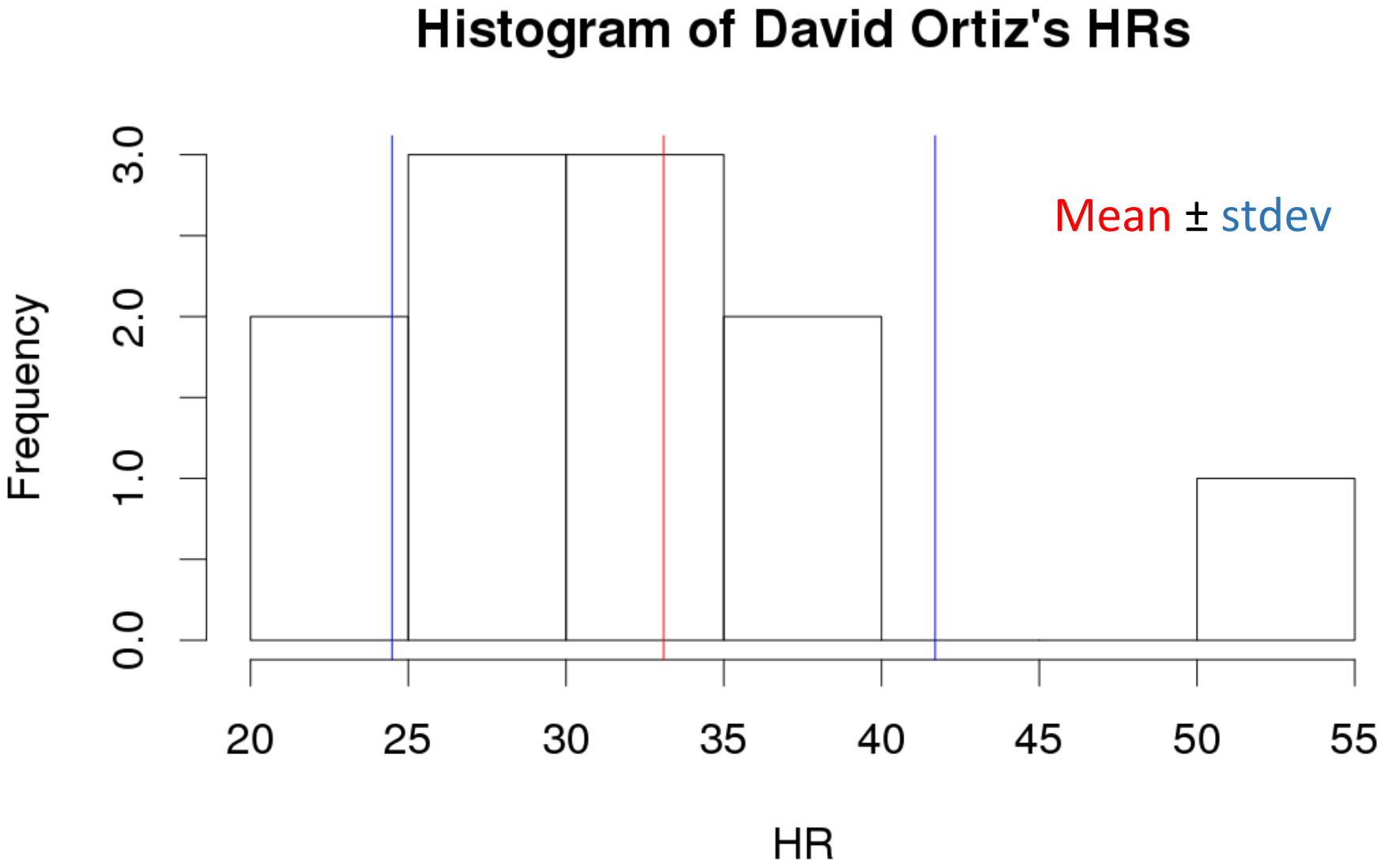
The median is resistant to outliers when the mean is not

The standard deviation

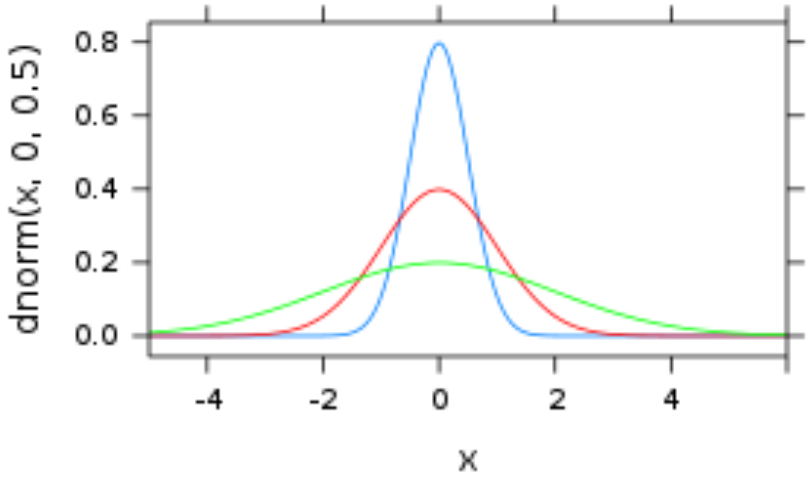
$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$


Histogram of David Ortiz's HRs
Frequency
0.0
1.0
2.0
3.0
20
25
30
35
40
45
50
55
HR

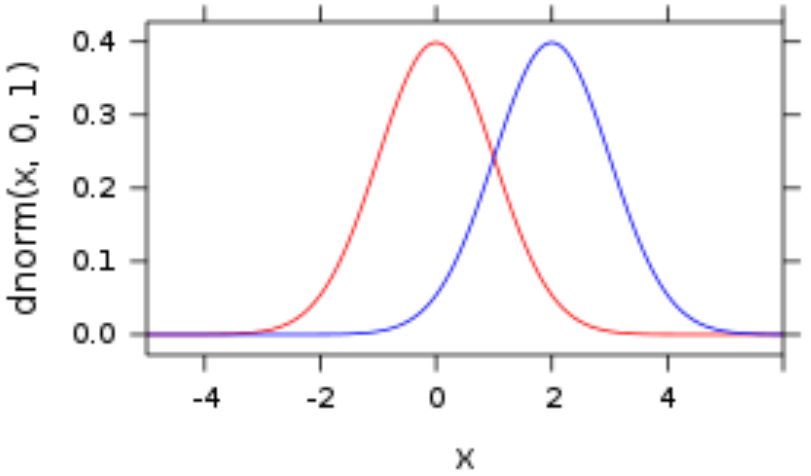

Histogram of David Ortiz's HRs
Mean
Frequency
0.0
1.0
2.0
3.0
20
25
30
35
40
45
50
55
HR


Histogram of David Ortiz's HRs
Mean ± stdev
Frequency
0.0
1.0
2.0
3.0
20
25
30
35
40
45
50
55
HR

Large vs small standard deviations



Same mean, different standard deviation



Same standard deviation, different mean

Which has a largest standard deviation?

A) Green
B) Red
C) Blue

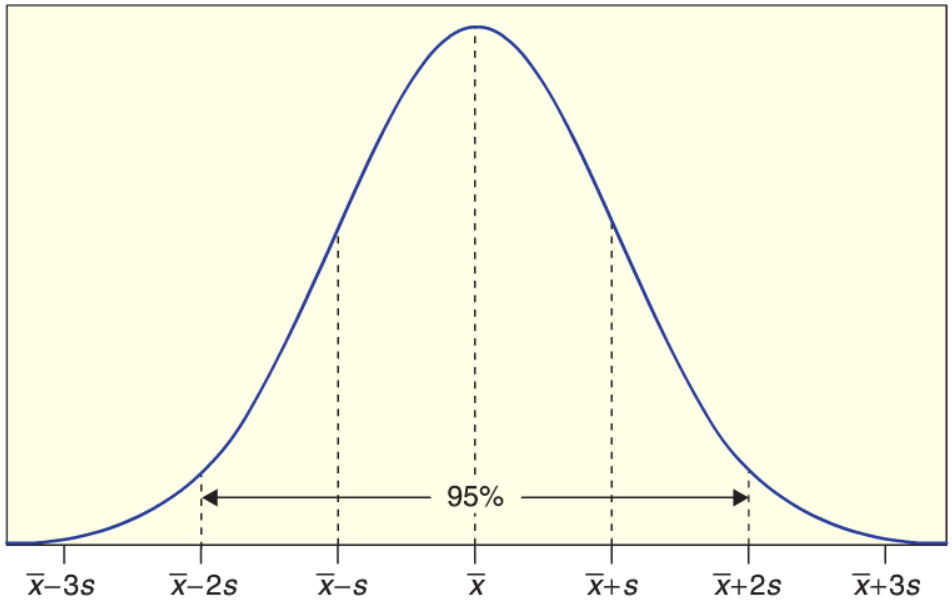
Which has a largest mean?

A) Red
B) Blue

The 95% rule (of thumb)

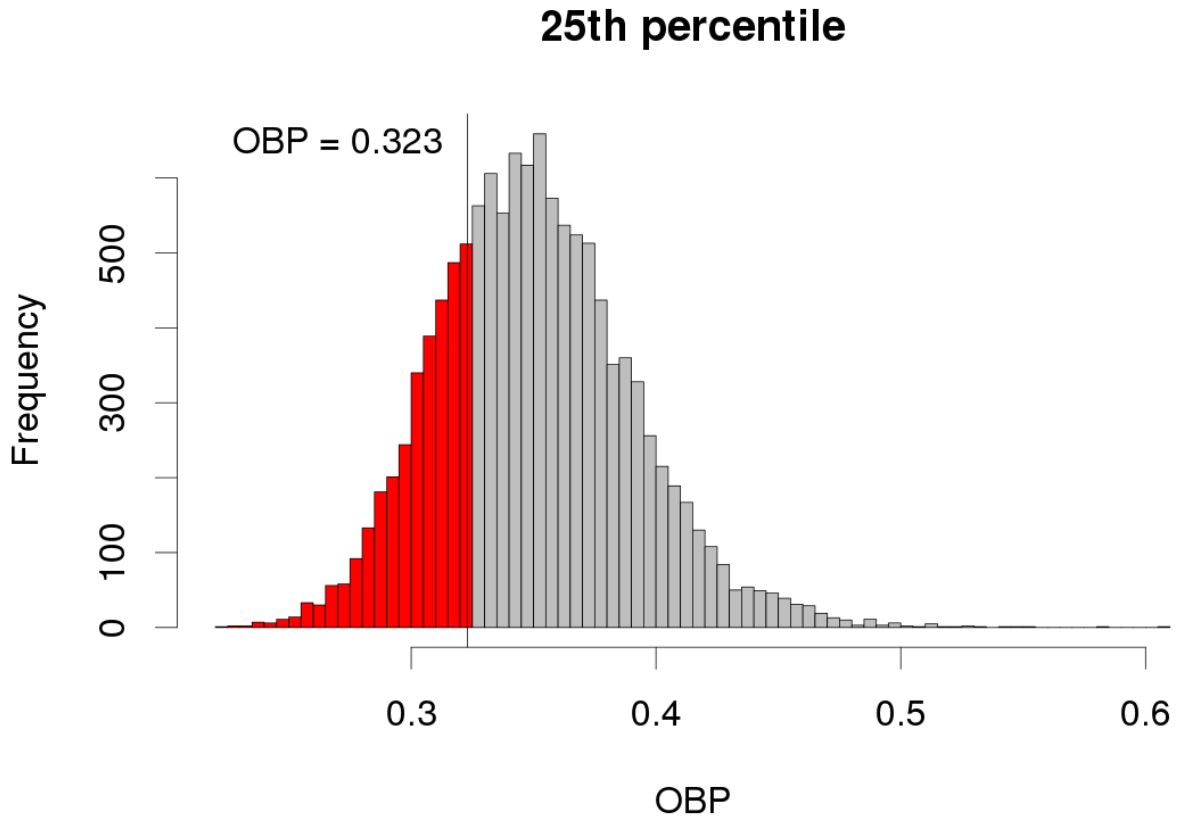
If a distribution of data is approximately symmetric and bell shaped, about 95% of the data should fall within two standard deviations of the mean.

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

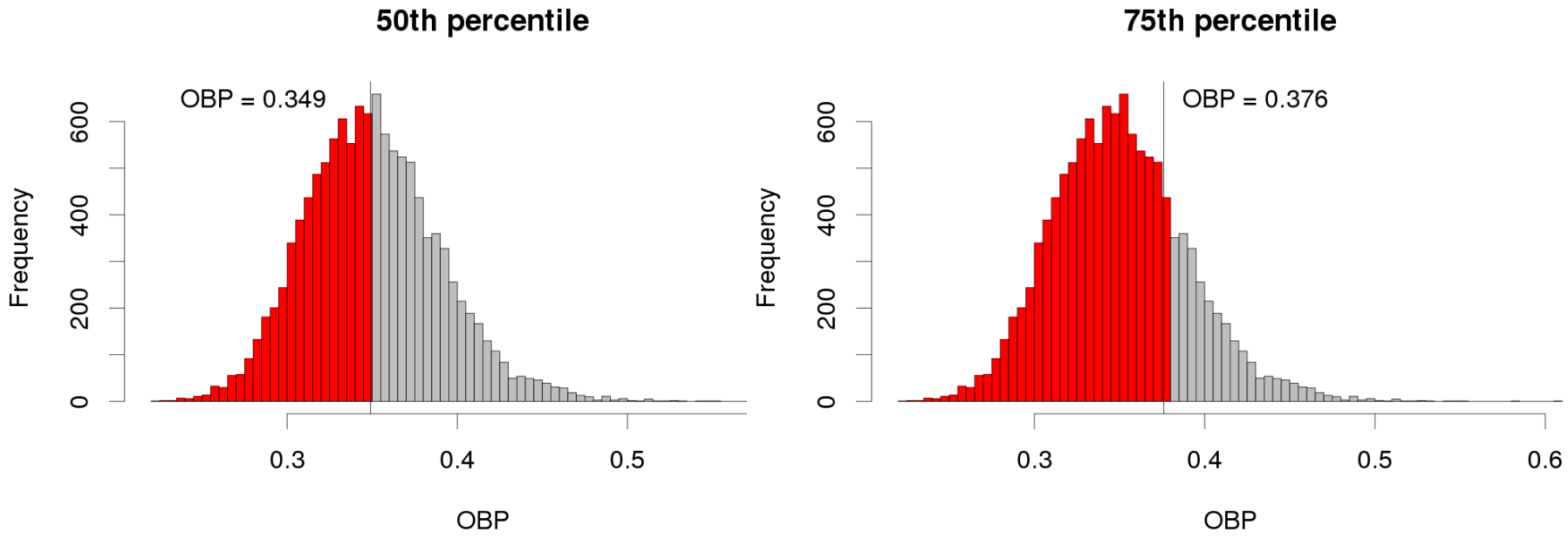


Percentiles

The **p^{th} 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/

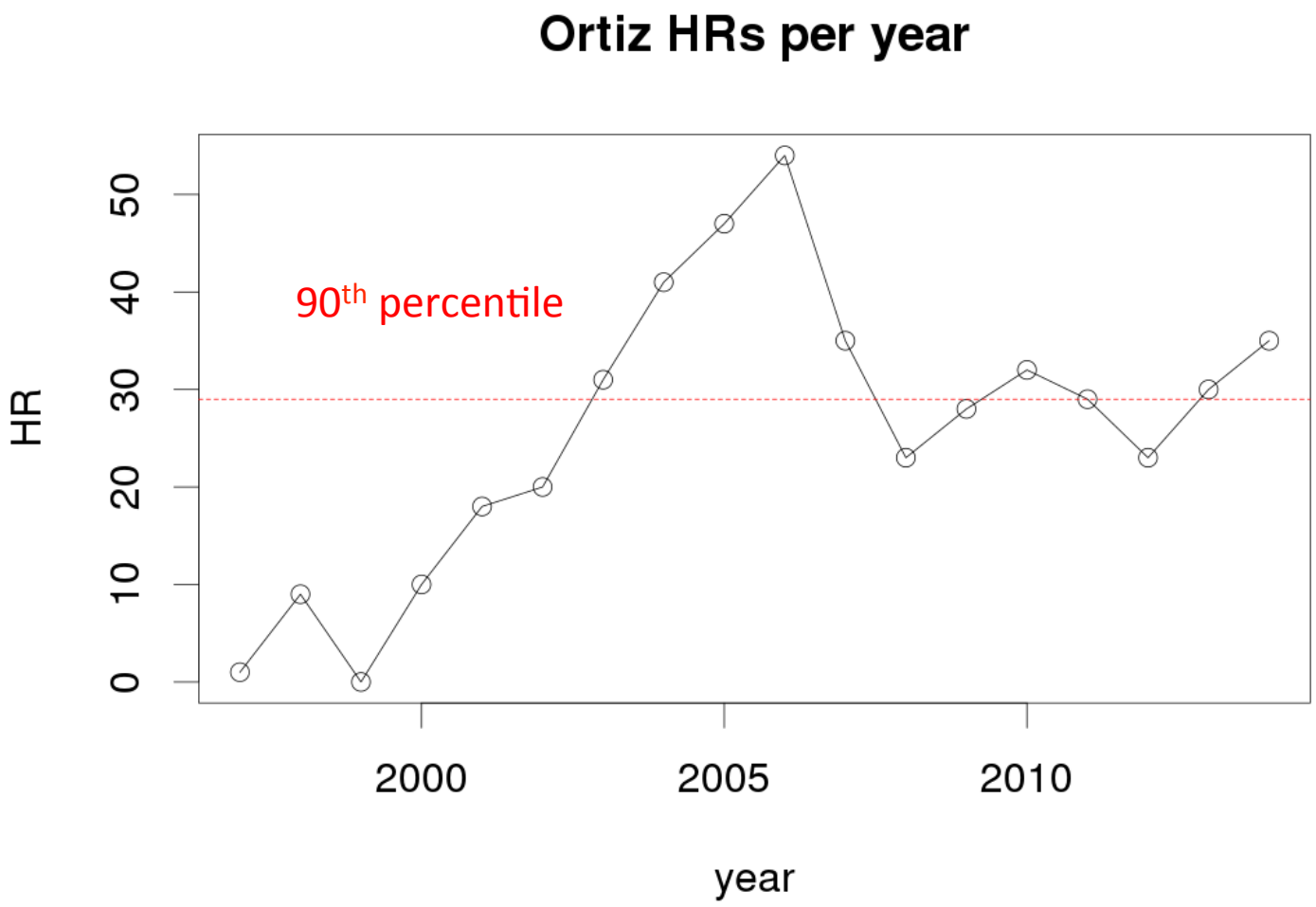
What is a good statistic for…?

Use the website to determine what “good” values are for the following statistics:

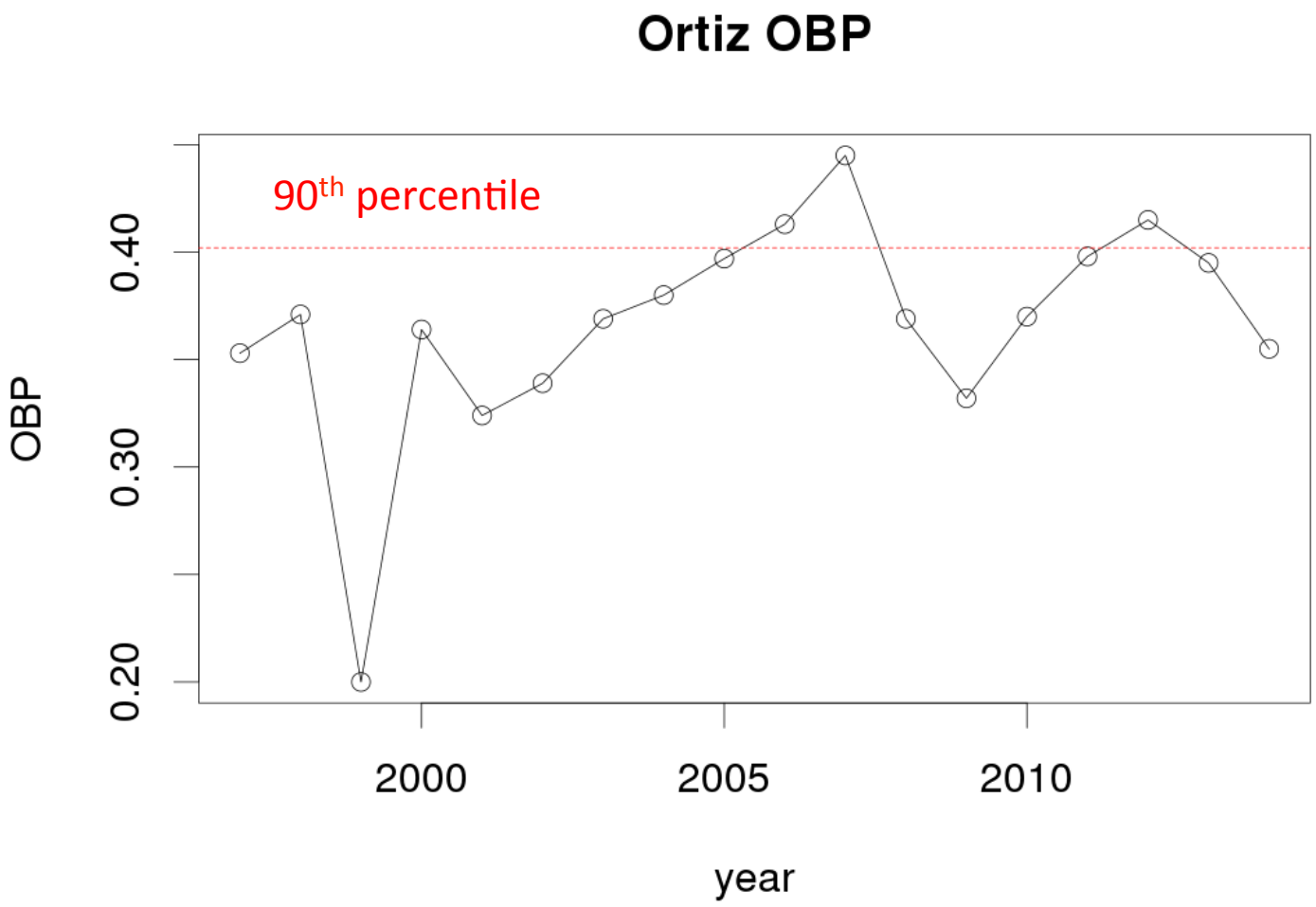
- Home runs (HR)
- On base percentage (OBP)
- Batting average (BA)
- Strikeouts (SO)

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

Putting statistics in context



Putting statistics in context



Calculating percentiles

The **p^{th} percentile** is the value of a quantitative variable which is greater than *p* percent of the data.

David Ortiz's Home run data:

Order	1	2	3	4	5	6	7	8	9	10	11
Sorted data	23	23	28	29	30	32	35	35	37	38	54
Percentile											

`R: quantile(x, probs)`

Calculating percentiles

The **p^{th} percentile** is the value of a quantitative variable which is greater than *p* percent of the data.

David Ortiz's Home run data:

Order	1	2	3	4	5	6	7	8	9	10	11
Sorted data	23	23	28	29	30	32	35	35	37	38	54
Percentile	0					50					100

`R: quantile(x, probs)`

Calculating percentiles

The **p^{th} percentile** is the value of a quantitative variable which is greater than *p* percent of the data.

David Ortiz's Home run data:

Order	1	2	3	4	5	6	7	8	9	10	11
Sorted data	23	23	28	29	30	32	35	35	37	38	54
Percentile	0	10	20	30	40	50	60	70	80	90	100

```
R: quantile(x, probs)
```

Calculating percentiles

Typically we ask for a value that is at the p^{th} percentile rather than calculate the percentiles for our data

- e.g., the 25% percentile value is 28.5
  - (weighted mean of the 20th and 30^{th} percentiles)

David Ortiz's Home run data:

Order	1	2	3	4	5	6	7	8	9	10	11
Sorted data	23	23	28	29	30	32	35	35	37	38	54
Percentile	0	10	20	30	40	50	60	70	80	90	100

```
R: quantile(x, probs)
```

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

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

Range and Interquartile Range

Range = maximum – minimum

Interquartile range (IQR) = $Q_3 - Q_1$

```
R: IQR(x)
```

Compute: 5 number summary, range, and IQR fro David Ortiz home runs

1. **Five Number Summary** = (min, Q_1, median, Q_3, max)
2. **Range** = maximum – minimum
3. **Interquartile range (IQR)** = $Q_3 - Q_1$

54	35	23	28	32
29	23	30	35	37
38				

Also use the percentile app to find the 5 number summary for HRs for all player-season with over 500 PA: https://emeyers.shinyapps.io/baseball_stat_percentiles/

5 number summary, range, and IQR fro David Ortiz home runs

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

54	35	23	28	32
29	23	30	35	37
38				

The 5 number summary for HRs for all player-season with over 500 PA is: (0, 4, 10, 20, 73)

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 outliers 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:

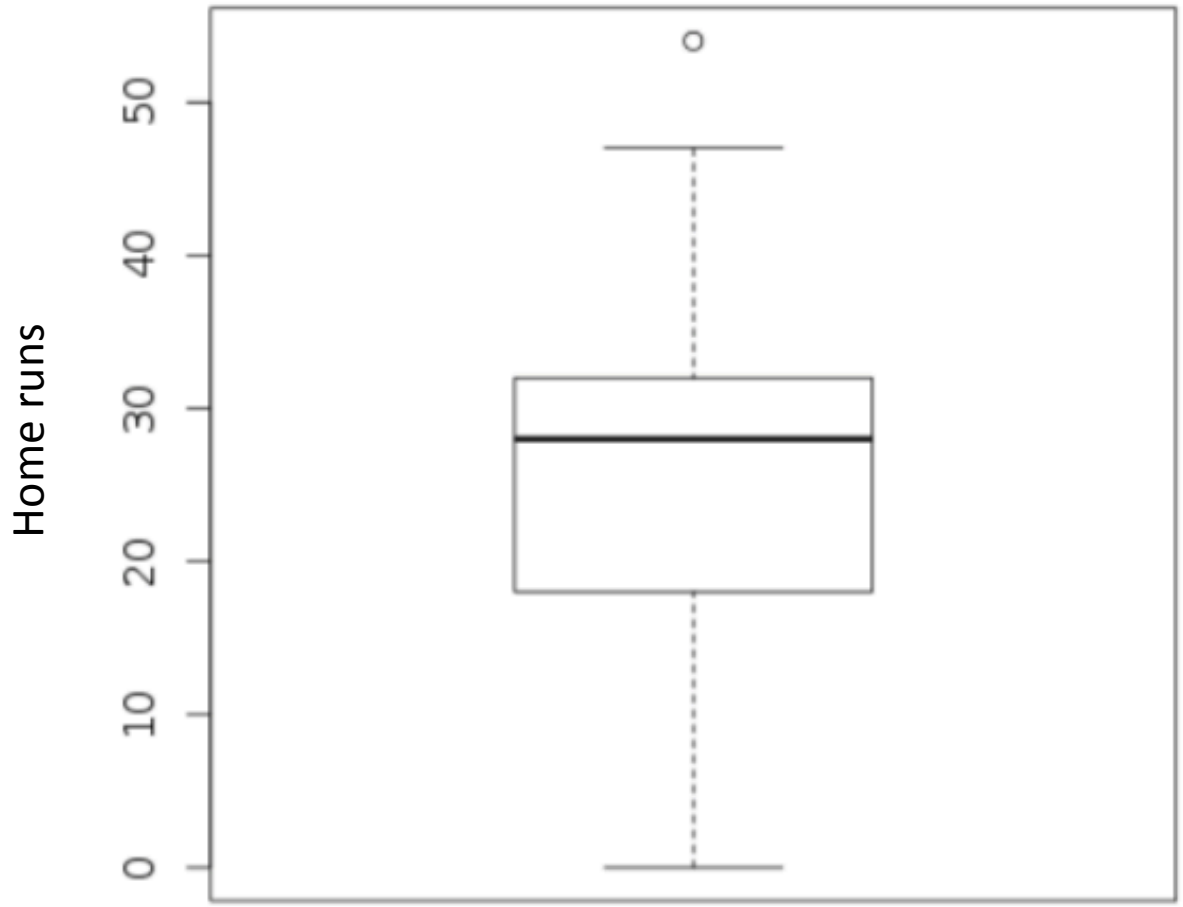
1. Drawing a box from Q_1 to Q_3

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

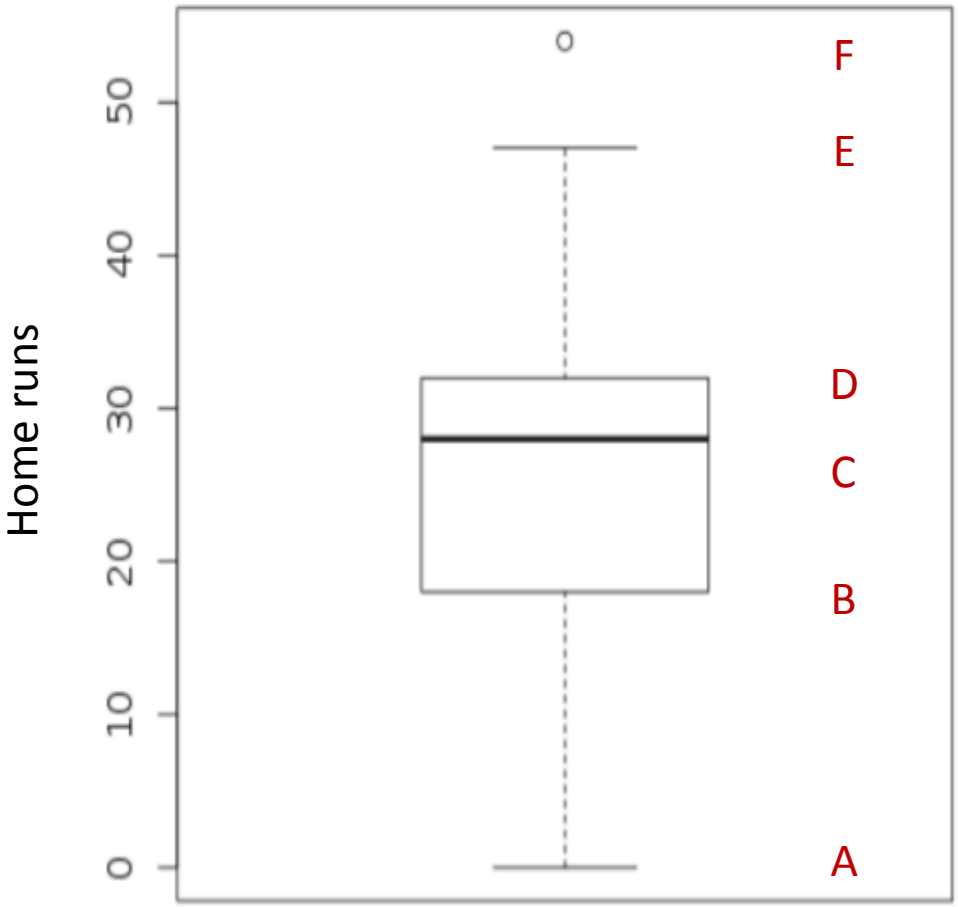
4. Draw a dot/asterisk for each outlier data point.

Box plot of David Ortiz home runs




```
R: boxplot(x)
```

Box plot quiz



What is:

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

Two current players: who is best?

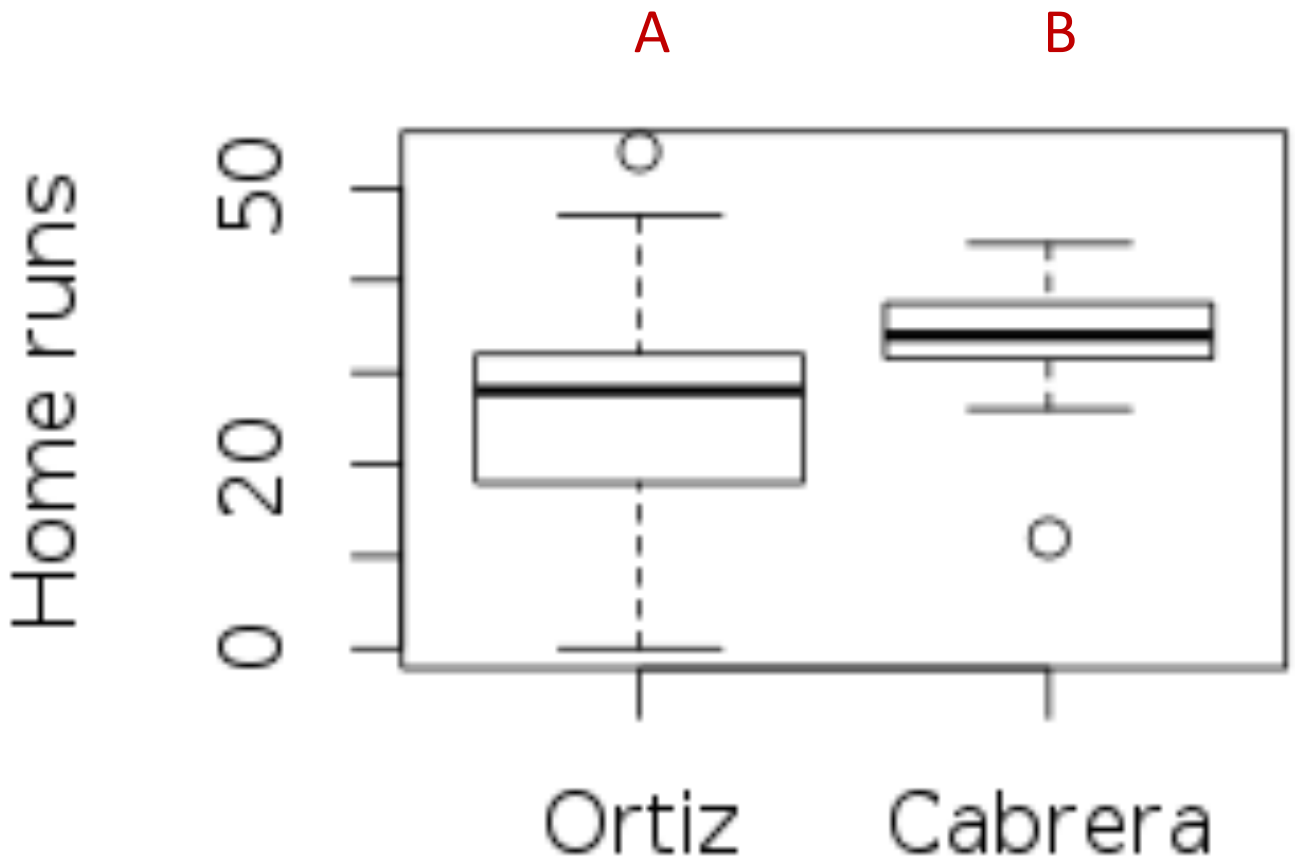
Miguel Cabrera:

HR in 2014 = 25

David Ortiz:

HR in 2014 = 35

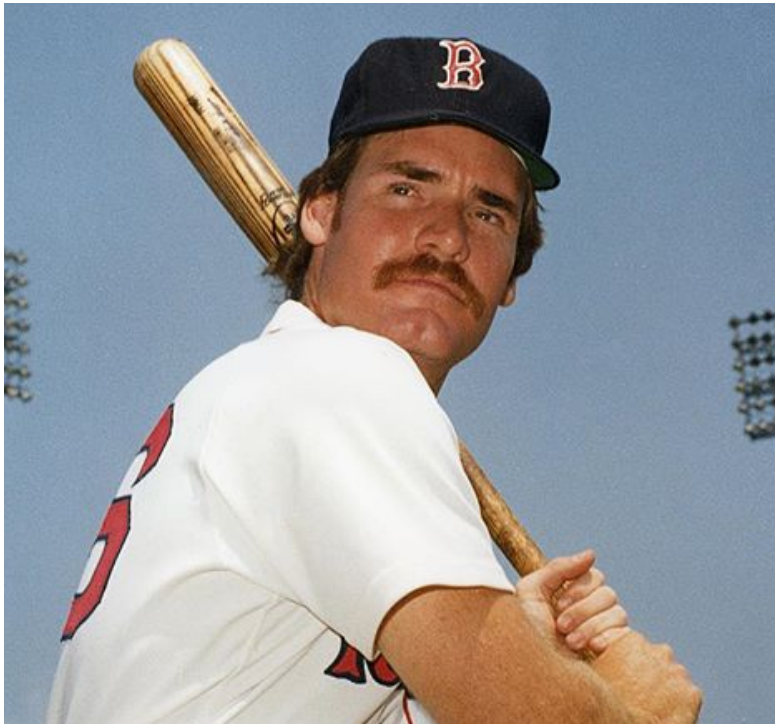
Comparing players with side-by-side box plots



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

Let's compare two more players

1985



Wade Boggs: BA = .368

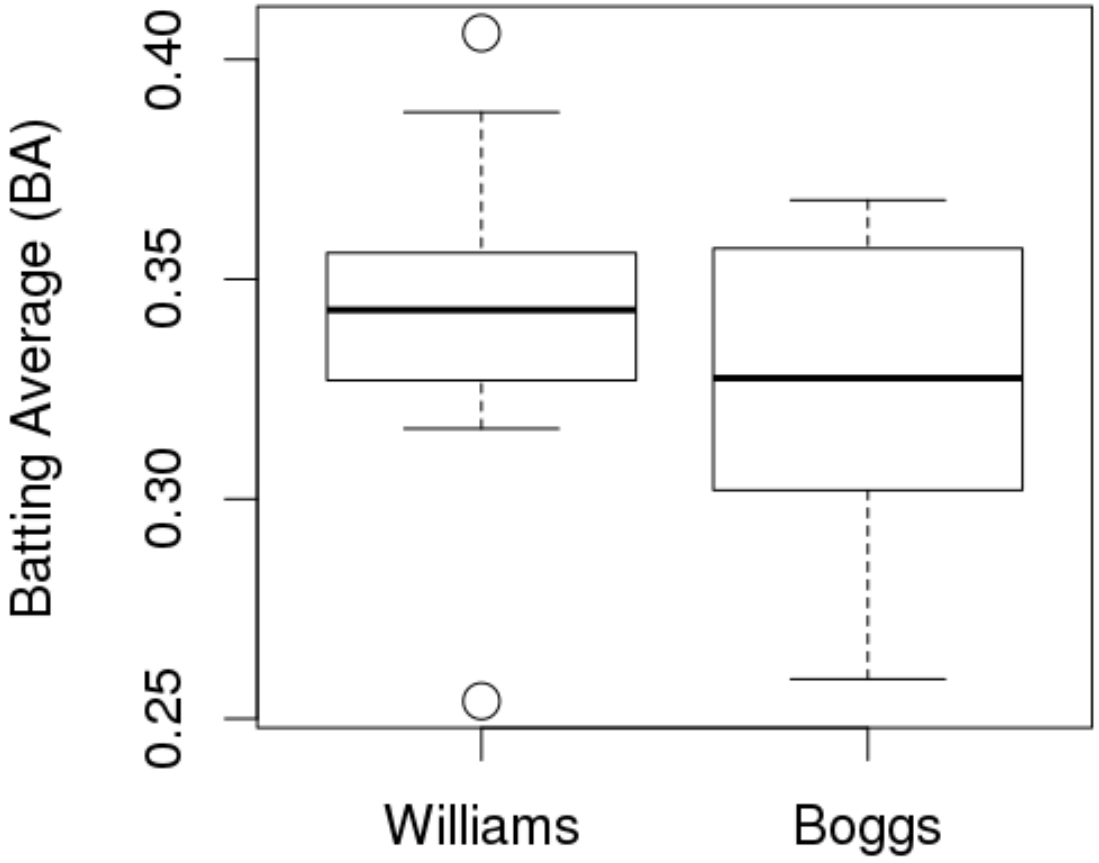
1941

Ted Williams: BA = .406

Career best seasons

Who is better?

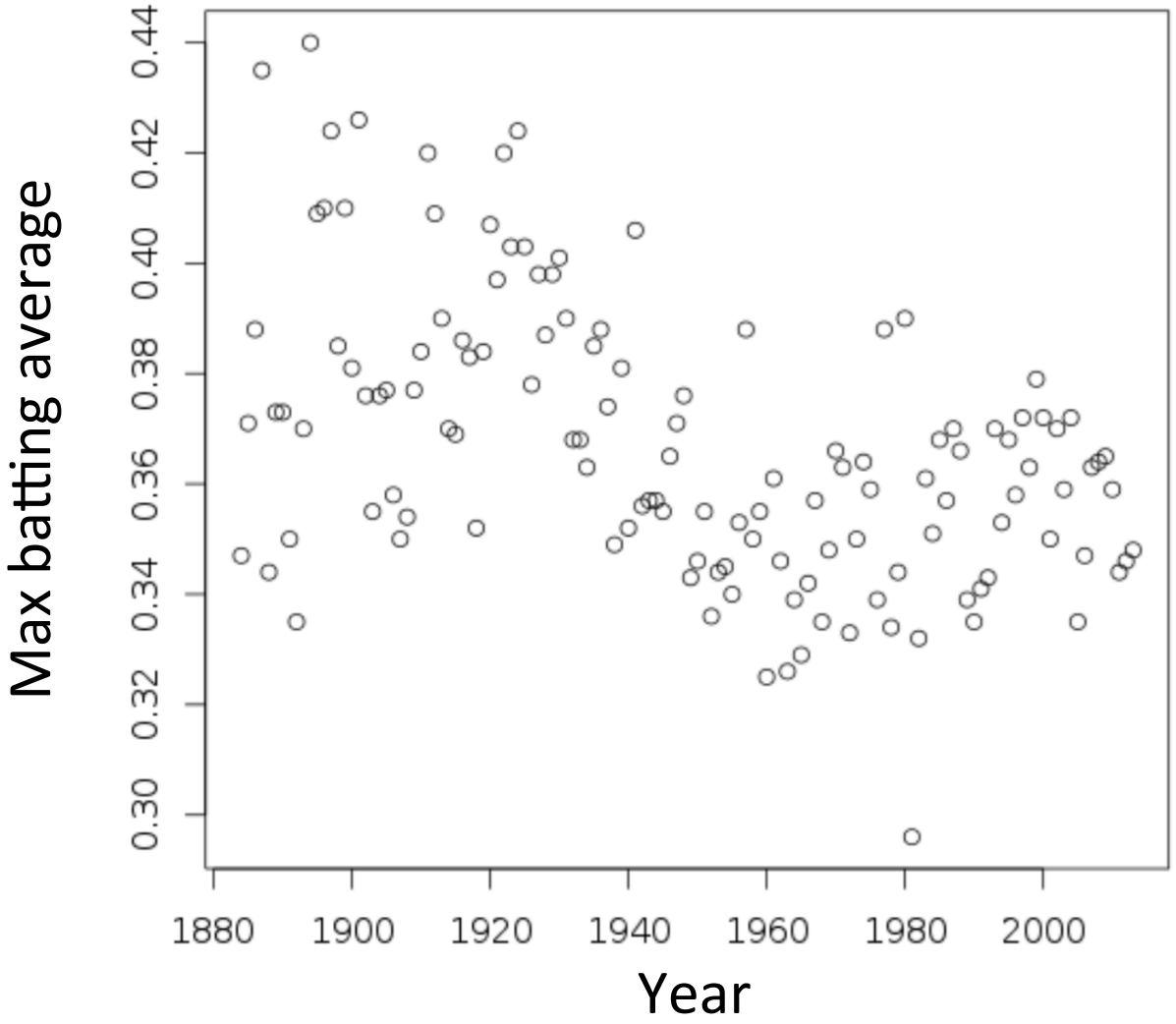
Who is best here?



Is Ted Williams better than Wade Boggs?

Ted Williams hit .406 in 1941
23 plenty of people hit over .400 before him
but no one has since…

Have the best players gotten worse at hitting over the past 140 years?



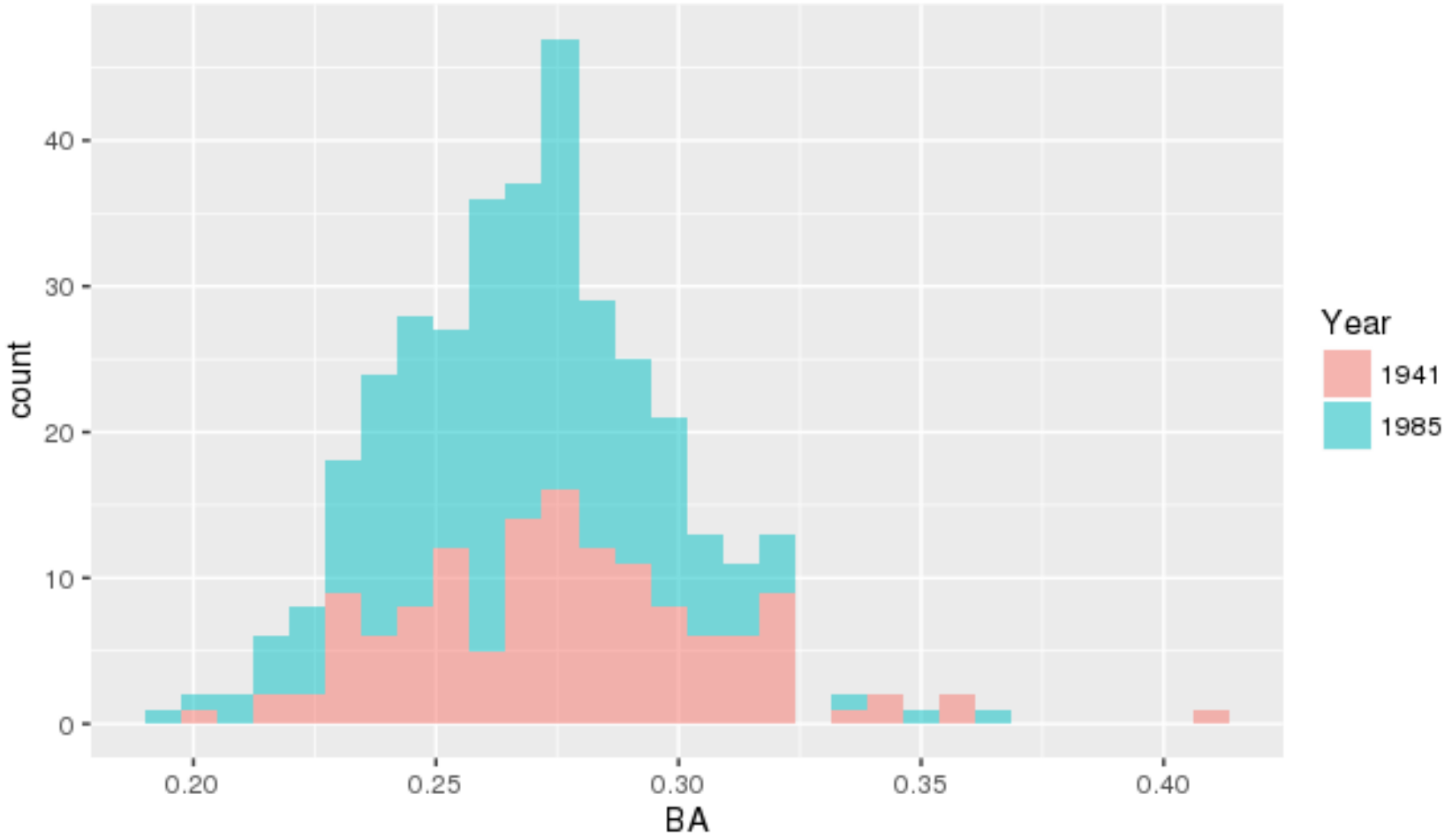
Comparing players across time periods

Problem: baseball has changed from 1871 to now

We can't simply compare statistics to judge how great a baseball player is when comparing across decades

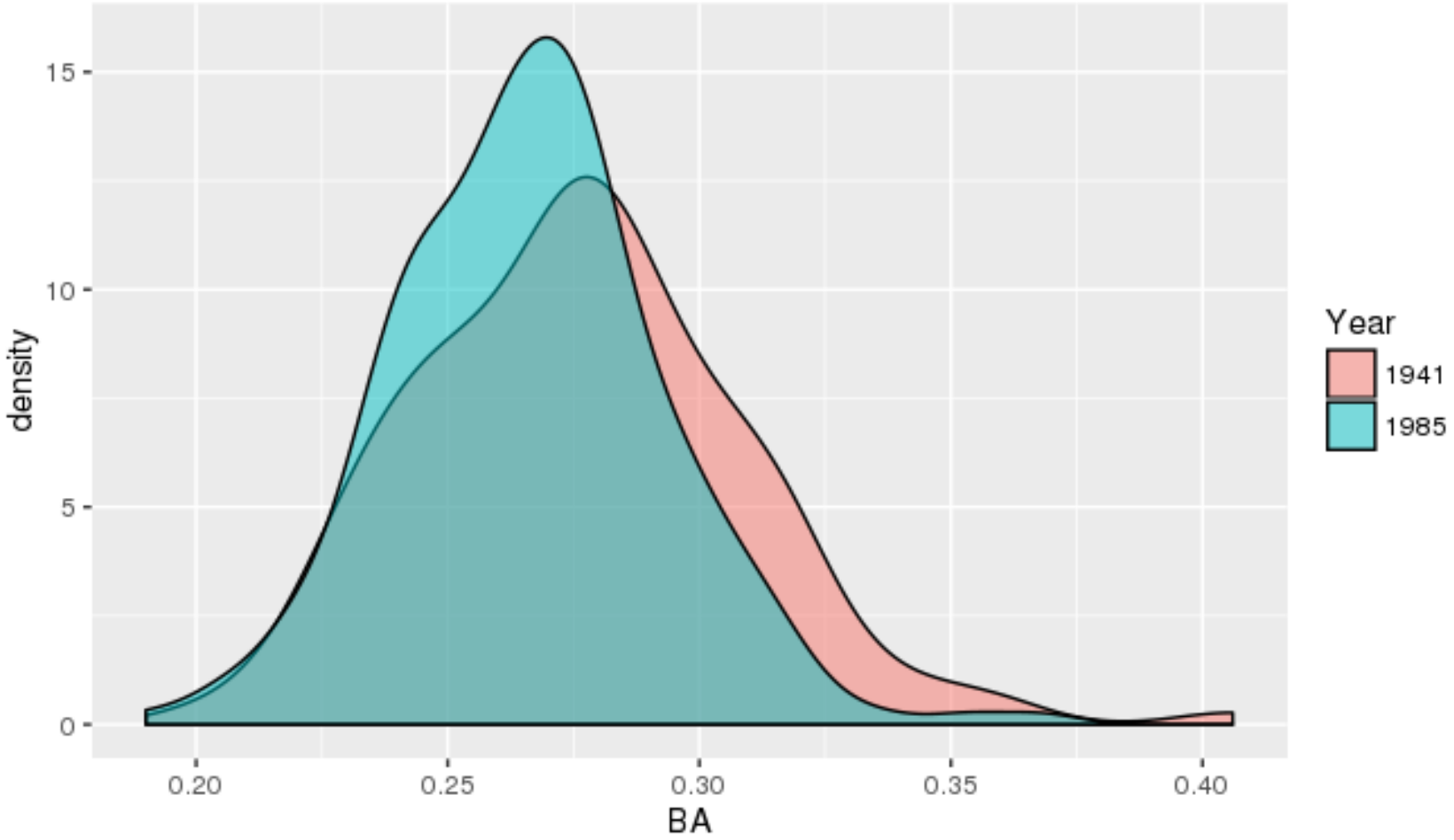
Useful to judge the 'greatness' of players relative to their peers

Histograms of batting average 1941 vs. 1985



Do the batting averages look similar in these years?

Density of batting average 1941 vs. 1985



Do the batting averages look similar in these years?

z-scores

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

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

z-scores for comparing players across eras

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

When comparing players across eras, we will use the mean ($\bar{x}$), and standard deviation (s) from each era separately.

This will give a measure of player performance relative to their peers in the same era

Comparing Ted and Wade to their peers

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

In 1941:

- Mean batting average was: .276
- Standard deviation in batting average was: .033
- Ted William's batting average was: .406

In 1985:

- Mean batting average was: .266
- Standard deviation in batting average was: .027
- Wade Bogg's batting average was: .368

Calculate z-scores for Ted and Wade's batting averages.

Who was better relative to their peers?

Comparing Ted and Wade to their peers

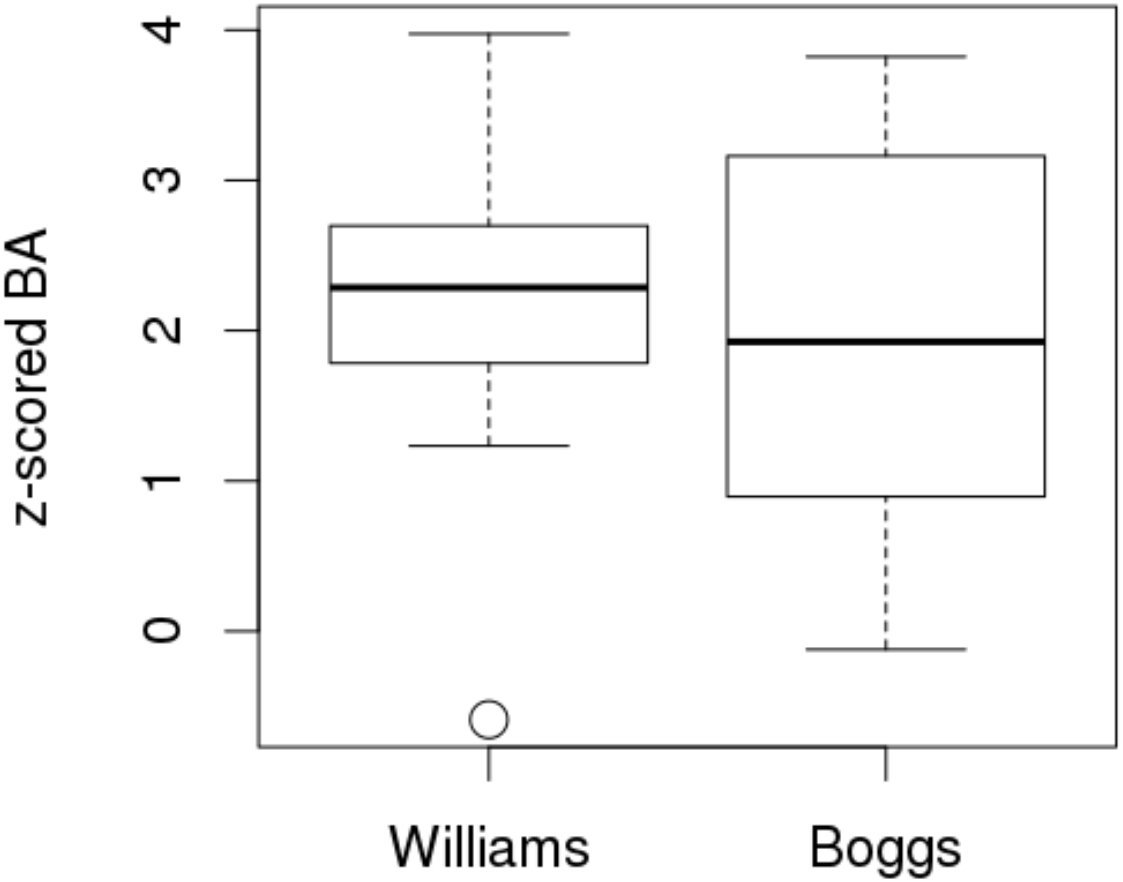
$$z_i = \frac{x_i - \bar{x}}{s}$$

Wade's batting average z-score: 3.82

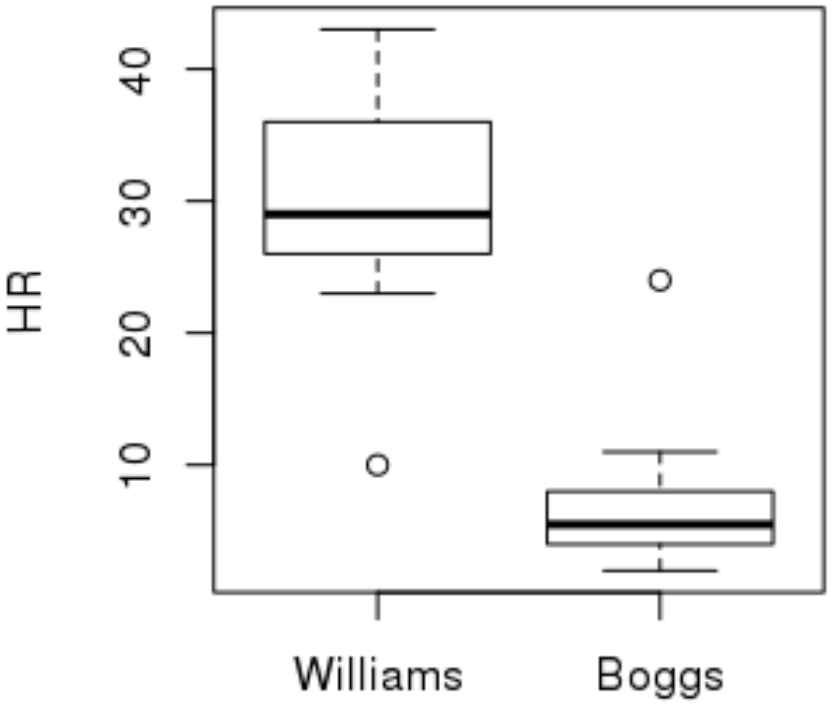
Ted's batting average z-score: 3.97

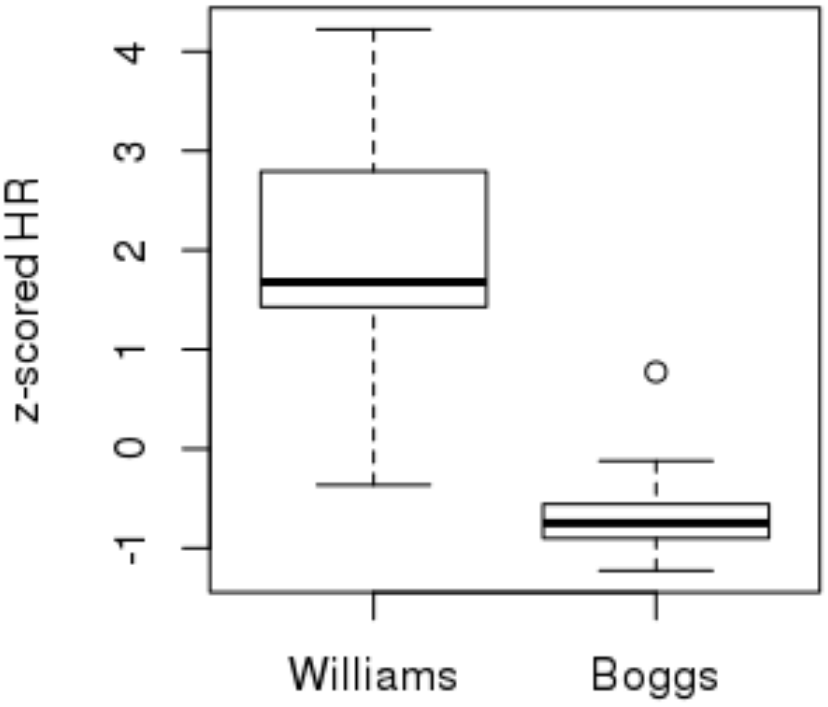
Who is the better hitter?

Career z-scored batting averages



What about Home Runs…





Next class: correlation!

Q and R: Big Data Baseball chapter 4