Describing and Summarizing Sports Data

DATA 299

Based on Ch. 2 in *Analytic Method in Sports, 2nd ed.*, Severini

Hitchman

Spring '25

Summarizing One Variable

- Visual displays help to reveal
  - the overall *shape* of the data
  - patterns within it, and exceptions to these patterns (outliers)
- A *summary statistic* is a number calculated from data that summarizing the data.
- Summary statistics help measure
  - the *center* - what is a "typical element"?
    - `mean(v)`
    - `median(v)`
  - the *spread* - how widely do values vary, and/or stray from center?
    - standard deviation `sd(v)`
    - coefficient of variation (standard deviation divided by mean)
    - Interquartile Range (IQR)

Histograms

- Histograms provide a view of the *data density*. It reveals the different values appearing in the data and how frequently they occur.
- Histograms reveal the *shape* of the data distribution.
- The chosen *bin width* can alter the story the histogram is telling.

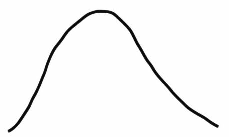
Key Shape Features

- *modality* - (number of peaks)
- *skewness* - (right, left, or symmetric?)
- *outliers* - unusual observations

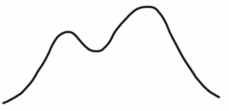
Commonly observed shapes of distributions

- modality

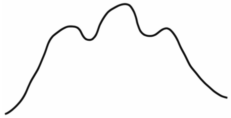
unimodal



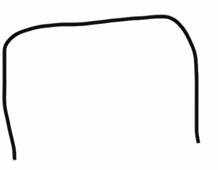
bimodal



multimodal

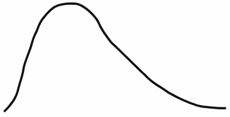


uniform

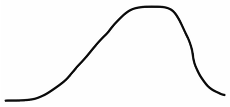


- skewness

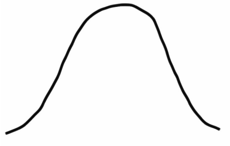
right skew



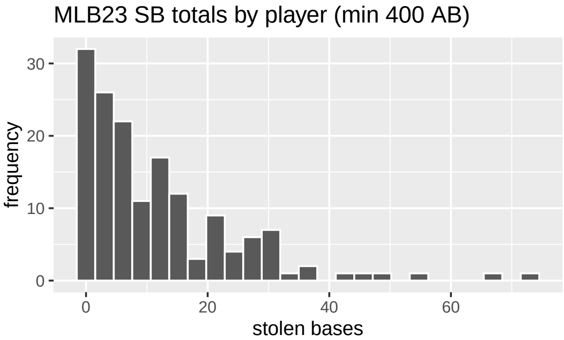
left skew



symmetric

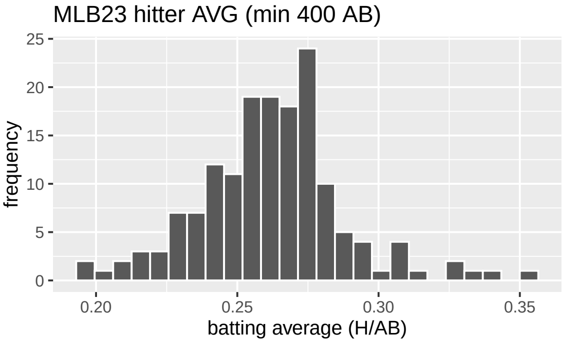


MLB Stolen bases



- unimodal
- skewed right
- some extreme values

MLB Batting Average



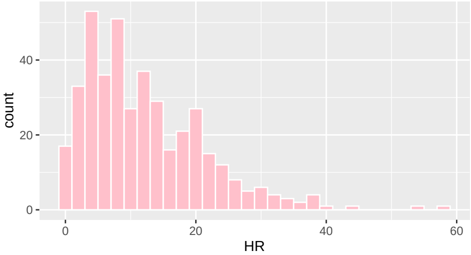
- unimodal
- distribution is fairly symmetric
- some extreme values on the high end

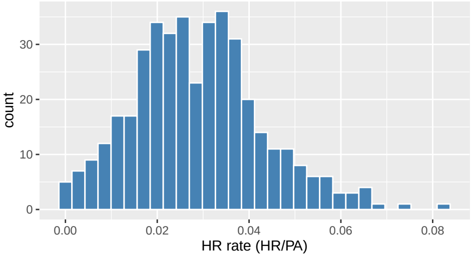
Guess the shape

Q: *In each case, based on what you know about the sport, decide on the likely shape of the distribution: uniform, symmetric, skewed left, skewed right*

- *the distribution of all PGA golf player earnings for the 2024 season*
- *The total number of points scored in each NFL game in the 2024 season*
- *The total number of goals scored in each game in the last women's world cup.*
- *The home runs hit by NCAA Div III baseball players in 2024.*
- *The home run rate (HR/PA) by NCAA Div III baseball players in 2024.*

MLB Home Run Count vs Rate

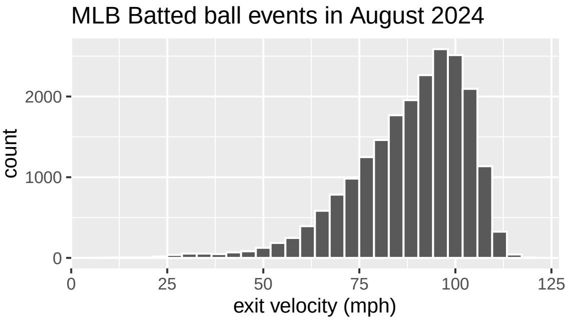

Home Runs hit by MLB players in 2024 (min 150 PA)


Home Runs Rates by MLB players in 2024 (min 150 PA

Skewed Left Distribution

Can you think of a sports statistic that likely has a skewed left shape?

MLB Batted ball velocity

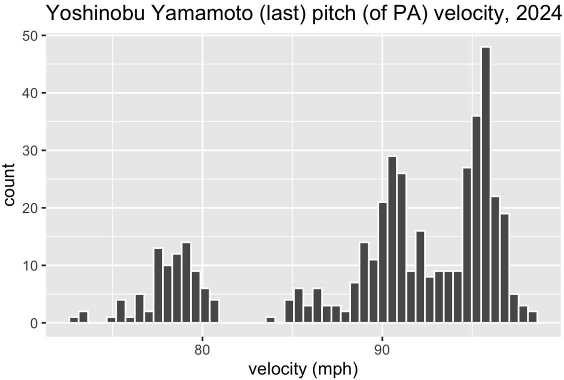


Classic skewed left shape! Why?

Multimodal Distribution

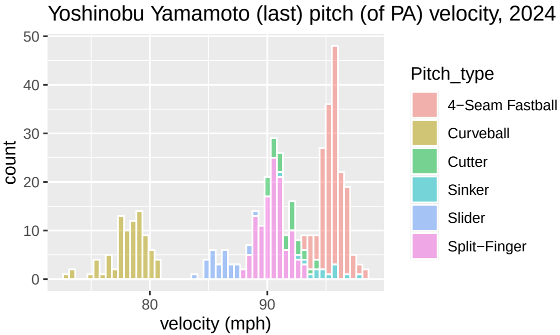
Can you think of a sports statistic that likely has a multimodal shape?

MLB pitch velocity for a pitcher?



Multimodal! Why?

MLB pitch velocity for a pitcher



Mean - One measure of the center of a distribution

The *mean*, denoted as $\bar{x}$, of a vector $v = c(x_1, x_2, \ldots, x_n)$ is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

In R, the mean value of a numeric vector `v` is `mean(v)`.

Median - A second measure of center

- The *median* is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

- If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- Since the median is the midpoint of the data, 50% of the values are below it, and the median is also called the 50^{th} *percentile*.

Hank Aaron's home runs by year

```
hank=c(13,27,26,44,30,39,40,34,45,44,24,32,
       44,39,29,44,38,47,34,40,20,12,10)
```

Finding the median by hand
How many seasons?

```
> length(hank)
[1] 23
```

The median will be the 12th value in a sorted list of HR counts:

```
> sort(hank)
[1]  10 12 13 20 24 26 27 29 30 32 34 34 38 39 39 40 40 44
44 44 44 45
[23] 47
> sort(hank)[12]
[1] 34
```

The median is $M = 34$ (which we can let R find for us: `median(hank)`)
To compare: `mean(hank)` $= 32.8$

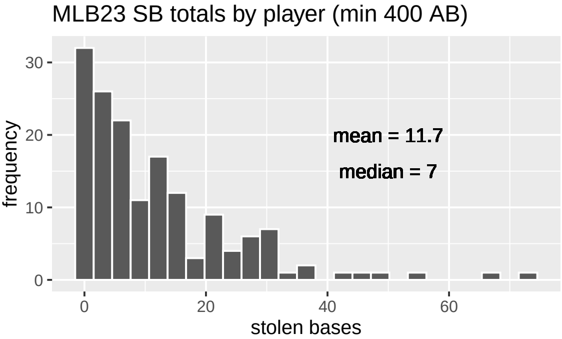
Comparing Mean and Median

The median is a more *robust* measure of center than the mean - it is less sensitive to extreme values.
Consider these two LEGO collections (number of pieces in each set)

Old collection	617	739	759	811	923	1015	1792
New collection	617	739	759	811	923	1015	4092

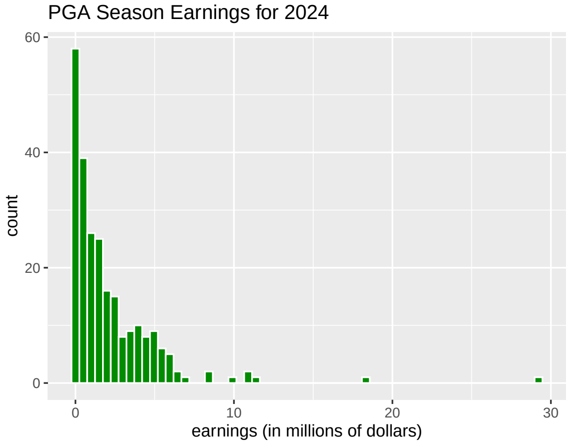
- The median of the new collection is the same as the old - 811.
- The mean increases dramatically from 950.9 to 1279.4.

MLB Stolen bases



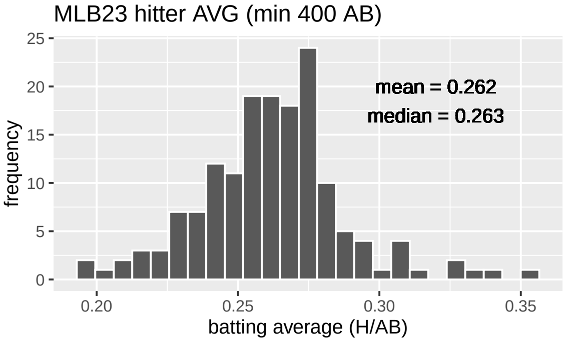
- The mean gets pulled toward the longer tail!
- distribution is skewed right, so mean is larger than median

Golf



- the mean earnings for 2024: 2.11 million dollars
- the median earnings for 2024: 1.23 million dollars
- which is a better measure of center?

MLB Batting Average



- symmetric distribution
- mean and median nearly identical

Measuring the spread of a distribution

Variation is

1. The *standard deviation*: a single number that captures how far the elements tend to be from the mean.
2. The *five number summary* is a set of 5 numbers that captures the spread and overall range of the data.
3. The five number summary is a more *robust* measure of spread than the standard deviation - it is less sensitive to extreme values, and it can reveal skewness.

Standard Deviation

The standard deviation of a set of values is a single number that captures how much the values tend to be from the mean.

Here are three data sets, and all of them have the same mean, $\overline{x} = 5$. Which has the greatest variance?

1. $[5, 5, 5, 5, 5, 5]$
2. $[4, 4, 5, 5, 6, 6]$
3. $[0, 0, 0, 10, 10, 10]$

Variance and Standard Deviation

- The *variance* of a data set with n values $x_1, x_2, \ldots, x_n$, denoted s^2, is:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

- The *standard deviation* s of a data set is the square root of the variance:

$$s = \sqrt{s^2}.$$

Example

The data set $[2, 3, 10]$ has mean $\overline{x} = (2+3+10)/3 = 5$. The standard deviation will be

$$s = \sqrt{\frac{(2-5)^2 + (3-5)^2 + (10-5)^2}{3-1}} = \sqrt{\frac{9+4+25}{2}} = \sqrt{19} \approx 4.36.$$

standard deviation in R

Of course, R loves to compute variance and standard deviation: Recall,

```
hank=c(13,27,26,44,30,39,40,34,45,44,24,32,
       44,39,29,44,38,47,34,40,20,12,10)

> var(hank)
[1] 125.06

> sd(hank)
[1] 11.18
```

Five Number Summary

The five number summary consists of the 5 statistics:

$$L \quad Q_1 \quad M \quad Q_3 \quad H$$

- L stands for 'low' - it is the minimum value.
- H stands for 'high' - it is the maximum value.
- M stands for median, as usual
- Q_1 stands for the first quartile, a number marking the 25% mark.
- Q_3 stands for the third quartile, a number marking the 75%mark.

The Interquartile Range (IQR)

The Interquartile Range

$$\text{IQR} = Q_3 - Q_1.$$

The IQR represents the spread of the middle 50% of the data.
Each component of the five number summary of a vector v can be found in R directly:

- L = `min(v)`
- Q_1 = `quantile(v,.25)`
- M = `median(v)`
- Q_3 = `quantile(v,.75)`
- H = `max(v)`

Five Number Summary in R

The `fivenum()` function in R returns a five number summary for a vector in one fell swoop.

```
> fivenum(hank)
[1] 10.0 26.5 34.0 42.0 47.0
```

The interquartile range of a vector *v* can also be found directly in R, via `IQR(v)`.

```
> IQR(hank)
[1] 15.5
```

Robustness with measures of spread

- The standard deviation is greatly influenced by outliers.
- The IQR is not.

Example

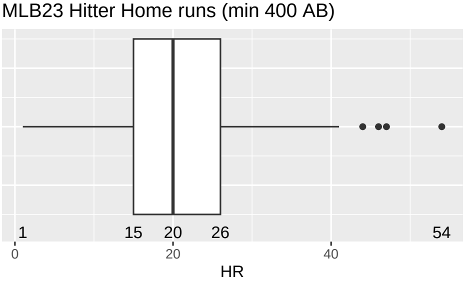
data set 1 : $4, 6, 6, 7, 7, 8, 8, 10, 11, 12$

data set 2 : $4, 6, 6, 7, 7, 8, 8, 10, 11, 22$

Standard deviations: $s_1 = 2.47$ and $s_2 = 5.02$, but their IQRs are equal.

Box Plots

A **box plot** is a pictorial representation related to the 5 number summary. A middle box represents the range from Q_1 to Q_3, with the median M drawn inside the box. Then whiskers run down to L and up to H, unless outliers are taken into account.



Whiskers and outliers in RStudio

- *Whiskers* of a box plot can extend up to 1.5 × IQR away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$

$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

In the previous slide:

$$\text{IQR} : 26 - 15 = 11$$

$$\text{max upper whisker reach} = 26 + 1.5 \times 11 = 42.5$$

$$\text{max lower whisker reach} = 15 - 1.5 \times 11 = -1.5$$

- A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

Q: *Why is it important to look for outliers?*

- *Identify extreme skew in the distribution.*
- *Identify data collection and entry errors.*
- *Provide insight into interesting features of the data.*

Coefficient of Variation (CV)

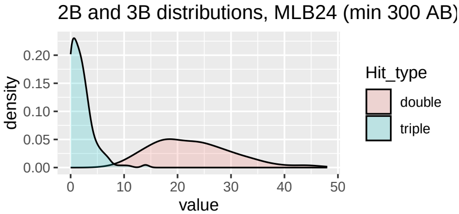
Question: Which sports league has the greatest variation in points/runs scored per game, NFL, MLB, or NBA?

- Data: All game scores for each league in 2024
- Compare standard deviation of scores in each league
- Does scale matter?
- We make variation "unitless" by considering *coefficient of variation*, which is standard deviation divided by mean.

Time permitting we will answer this question as a group using the `Margin of victory...` data set on our course resource page.

Coefficient of Variation (CV)

In baseball, is there more variation in batter doubles or batter triples? Consider all batters from 2024 with at least 300 at bats.



- 2B: mean $= 22.8$; sd $= 7.70$
- 3B: mean $= 2.05$; sd $= 2.23$
- $CV_{2B} = 0.34$; $CV_{3B} = 1.09$

The doubles distribution has a larger standard deviation, (a larger IQR as well), but the triples distribution has more variability relative to the average value of the dist'n.

Variation in categorical data

Question: How well does a pitcher mix their pitches?
First approximation: Consider only the frequency with which each was thrown (not taking sequence into account). Here are three pitchers from the 2024 Season

Pitcher	N	Fastball	Slider	Cutter	Curve	Change
Bailey Ober	2696	0.381	0.116	0.195	0.014	0.294
Cole Ragans	3118	0.418	0.133	0.108	0.104	0.237
Dylan Cease	3188	0.435	0.430	0.007	0.119	0.009

data: `https://www.fangraphs.com/` (leaders → pitching: 2024 → pitch level data → pitch type)
Observations?

Entropy

- *Entropy* is a measure of the variability of values occurring in a categorical variable.
- Entropy will be zero if there is no variability at all (if everyone has brown eyes in a population, the ‘eye color’ variable has entropy 0)
- Entropy will be maximized when every outcome is equally likely
- Entropy formula incoming!

Entropy formula

Suppose a categorical variable takes k possible outcomes, and $p_1, p_2, \ldots, p_k$ represent the proportion of the observations taking on these k outcomes, respectively. Then the entropy associated with the variable is

$$-\sum_{i=1}^{k} p_i \cdot ln(p_i),$$

where $0 \cdot ln(0)$ is interpreted as 0 (i.e., ignore outcomes that don't actually occur). Cole Ragans:

$$p_1 = .418, p_2 = .133, p_3 = .108, p_4 = .104, p_5 = .237,$$

and the entropy of these 3118 pitches is

-[.418 $\cdot$ ln(.418) + .133$\cdot$ ln(.133) + .108 $\cdot$ ln(.108) + .104 $\cdot$ ln(.104) + .237 $\cdot$ ln(.237)]=1.451.

Interpreting this number?

Maximum Entropy

The largest possible entropy value occurs when all k of the p_i are equal. In this case, each $p_i = 1/k$, and here's how the entropy formula evaluates:

$$\begin{aligned}
\text{max entropy} &= -[p_1 \cdot \ln(p_1) + p_2 \cdot \ln(p_2) + \cdots + p_k \ln(p_k)] \\
&= -k \cdot [\frac{1}{k} \ln(1/k)] \\
&= -\ln(1/k) \\
&= \ln(k).
\end{aligned}$$

In the case of 5 possible outcomes (as was the case with Cole Ragans' pitches), the maximum entropy occurs when each $p_i = 0.2$, and this maximum value is $\ln(5) = 1.609$.

Standardized Entropy

We define *standardized entropy* to be the entropy rescaled from 0 to 1, obtained by dividing entropy by $\ln(k)$.

```
std_entropy<-function(x){
  #x is an input vector of p_i>0
  k <- length(x)
  entropy <- -sum(x*log(x))
  max_entropy <- log(k)
  std_entropy <- entropy/max_entropy
  return(std_entropy)
}
```

Pitcher	N	FF	SL	CT	CU	CH	std_E
Ober	2696	0.381	0.116	0.195	0.014	0.294	.843
Ragans	3118	0.418	0.133	0.108	0.104	0.237	.904
Cease	3188	0.435	0.430	0.007	0.119	0.009	.656

Is entropy useful?

Is entropy useful in terms of evaluating performance, for instance?

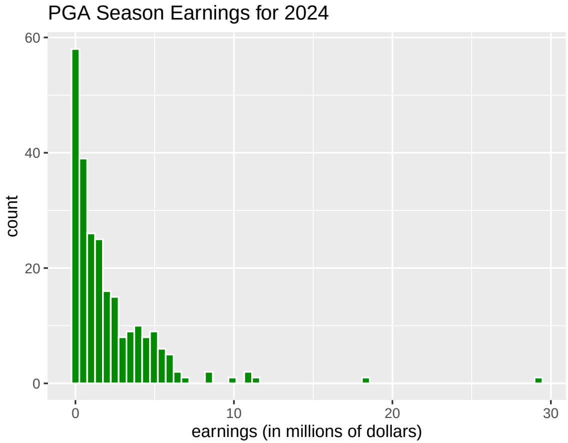
- Entropy is a rough measure of predictability.
- If a pitcher throws the same type of pitch every time (so entropy is zero), might batters have an easier time getting hits off that pitcher?
- Is entropy associated with pitching performance?
- Let's take a look (in R Studio)

Using Transformations

- Motivation: Improve measures of team and player performance.
- Strategy: Given stat X, perhaps a function of it creates a new stat $Y = f(X)$.
- Generally, we chose f to be a one-to-one function so order is preserved. Common fs:
  - $f(x) = \frac{1}{x}$
  - $f(x) = \log(x)$
  - $f(x) = e^x$

Golf

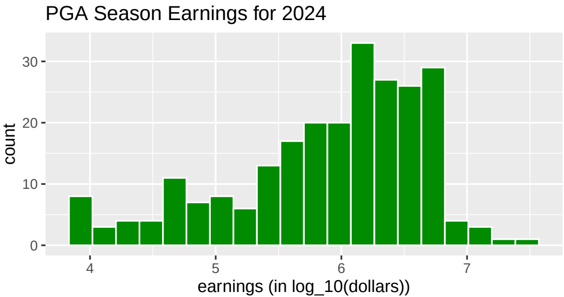
Recall earnings distribution with strong skewness



What if we transformed earnings using $f(x) = \log(x)$?

Golf

Resulting distribution is closer to symmetric, and order is preserved. Scottie Scheffler still on top!



Is this useful?

log transformations

If $Y = \log_{10}(X)$, an increase in Y by 1 corresponds to an increase in X by a factor of 10:

X	$Y = \log_{10}(X)$
10	1
100	2
1000	3
10000	4
100000	5
1000000	6

Comparing 3 golfers

	rank	player	money
13	13	Shane Lowry	6095881
45	45	Cam Davis	4039533
85	85	Adam Svensson	2014485

- If we compare players via difference in winnings, Cam Davis is equally close (in earnings) to Shane Lowry and Adam Svensson.
- If we compare via ratio of earnings, Davis is closer to Lowry, because he made about 2/3 what Lowry made, and Svensson made about 1/2 what Davis made.
- An analysis based on ratios (as opposed to differences) corresponds to comparing in terms of log-winnings.

Transforming to symmetric distributions

From text (pp. 29-30) regarding the usefulness of the two golf earnings plots we've seen (winnings vs log-winnings)

> *[T]he two histograms tell vastly different stories.*
> *The histogram of the raw winnings suggests that there was one outstanding golfer and a few very good golfers, and the vast majority of golfers performed relatively poorly compared to these few top golfers. The histogram of log-winnings suggests a distribution of performances that is more bell shaped . . . with the majority having an "average" performance and a few performing either very well or very poorly. For most, but not all, purposes, a performance measure based on log-winnings appears to be more useful.*

Rating Home Run Hitters!

Who is the better home run hitter?

player	home runs
A	30
B	10

Hah! You don't fool me, Hitchman. I know not to look at raw counts. How many opportunities did each hitter have?

Rating Home Run Hitters!

Who is the better home run hitter?

player	at bats	home runs
A	500	30
B	200	10

Two ideas: Calculate HR/AB and AB/HR.
Which measure is better if we're trying to assess performance?

Rating Home Run Hitters!

Who is the better home run hitter?

player	AB	HR	AB/HR	HR/AB
A	500	30	16.7	0.060
B	150	10	15.0	0.067

Either one suggests player B does better with their opportunities. Is one stat better to use?

With AB/HR, lower is better; With HR/AB, higher is better.

HR/AB or AB/HR?

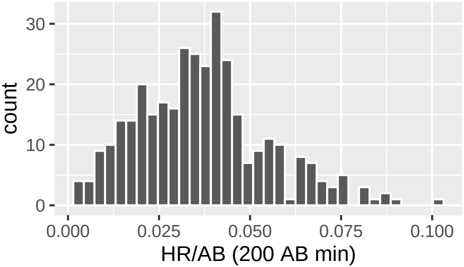
First note that we obtain AB/HR from HR/AB by applying the transformation $f(x) = 1/x$. This means if someone has a “better” (higher) HR/AB ratio than someone else, they also have a “better” (lower) AB/HR ratio.

Either one suggests player B does better with their opportunities. Is one stat better to use?

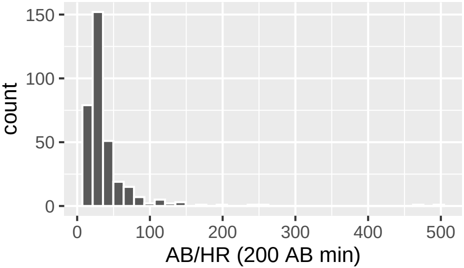
Let’s look at plots from Lahman hitting data from 2023.

```
hit <- Batting |>
  filter(yearID==2023,AB >= 200) |>
  select(playerID,AB,HR) |>
  mutate(HRperAB = HR/AB,
            ABperHR = AB/HR)
```

HR/AB or AB/HR?



count
30
20
10
0
0.000
0.025
0.050
0.075
0.100
HR/AB (200 AB min)



count
150
100
50
0
0
100
200
300
400
500
AB/HR (200 AB min)

HR/AB or AB/HR?

HR/AB appears to be better as a measure of home run hitting performance for several reasons.

- AB/HR is undefined if a batter has hit no home runs!
- Distribution looks much closer to symmetric, a distribution shape that often arises when measuring athletic performance.
- From the text, p. 29: “Measurements for which the interpretation of one unit is the same throughout the range of the variable tend to have a distribution that is roughly symmetric”
- Let’s look at several hypothetical hitters, each having 500 AB (almost a full season’s worth) (4*162 = 648, but who does that any more?)

HR/AB or AB/HR?

Each increase of .01 in HR/AB corresponds to 5 additional HR over 500 AB.

player	AB	HR	AB/HR	HR/AB
A	500	45	11.1	0.09
B	500	40	12.5	0.08
C	500	35	14.3	0.07
D	500	30	16.7	0.06
E	500	25	20.0	0.05
F	500	20	25.0	0.04
G	500	15	33.3	0.03
H	500	10	50	0.02
I	500	5	100	0.01
J	500	0	∞	0.00

HR/AB or AB/HR?

An increase of 5 in AB/HR corresponds to *different changes in HR totals over 500 AB.*

player	*AB*	*HR*	*AB/HR*	*HR/AB*
A	*500*	*50*	*10*	*0.1*
B	*500*	*33.3*	*15*	*0.067*
C	*500*	*25*	*20*	*0.05*
D	*500*	*20*	*25*	*0.04*
E	*500*	*16.7*	*30*	*0.033*
F	*500*	*14.3*	*35*	*0.029*
G	*500*	*12.5*	*40*	*0.025*
H	*500*	*11.1*	*45*	*0.022*
I	*500*	*10*	*50*	*0.02*