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
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

<table>
<thead>
<tr>
<th>Count</th>
<th>1B</th>
<th>2B</th>
<th>3B</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>90</td>
<td>38</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.56</td>
<td>0.24</td>
<td>0.01</td>
<td>0.19</td>
</tr>
</tbody>
</table>

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

\[ R: \text{barplot}(x) \]

\[ R: \text{pie}(x) \]
World's Most Accurate Pie Chart

- Pie I have eaten
- Pie I have not yet eaten
Describing quantitative variables

R: stem(team.data.2014$HR, scale = 2)

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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>59</td>
</tr>
<tr>
<td>11</td>
<td>178</td>
</tr>
<tr>
<td>12</td>
<td>233558</td>
</tr>
<tr>
<td>13</td>
<td>1246</td>
</tr>
<tr>
<td>14</td>
<td>267</td>
</tr>
<tr>
<td>15</td>
<td>0255567</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>19</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1</td>
</tr>
</tbody>
</table>

R: hist(x)

Histogram of team HR totals
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 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

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

Mean = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum 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?

```r
> 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 $\mu$, which is the Greek letter “mu”
Give the proper notation: $\mu$ 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

\[ R: \text{median}(x) \]
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

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

<table>
<thead>
<tr>
<th>Values</th>
<th>Deviations</th>
<th>Squared Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>20.91</td>
<td>437.19</td>
</tr>
<tr>
<td>35</td>
<td>1.91</td>
<td>3.64</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>54</th>
<th>35</th>
<th>23</th>
<th>28</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>23</td>
<td>30</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>