CS 149: Understanding Statistics Using Baseball
Outline for today

Introductions

Go over syllabus

The history of statistics in baseball

Baseball statistics and structured data
Introductions

About you:

• Name
• Preferred gender pronouns
• Year at Hampshire: Div I, II-first year, II-second year, III
• A bit about your background
  • E.g. experience with baseball, programming and Statistics
• Anything else you want to say

Is there any topic you are particularly interested in? i.e., why are you interested in this class?
General Information

My contact info:
Email: emeyers@hampshire.edu
Phone: x5500
Office: ASH 133

Drop-in office hours:
• Monday 2:30-3:30pm
• Tuesday 12:30-1:30
• By appointment

Website:
https://moodle.hampshire.edu/course/view.php?id=5532
Course objectives

1. To understand how to use **statistical methods and thinking** to make sense of baseball and other systems that involve random processes

2. To be able to use the R programming language to analyze data
Note: statistics vs. Statistics

**statistics**: are numerical summaries of data

**Statistics**: is the mathematics of collecting, organizing and interpreting data
Why use baseball to study Statistics?

High degree of randomness
  • Very good players hit safely 3 out of 10 times (ave = .300)
  • Bad players hit safely 2 out of 10 times (ave = .200)

Contains a rich structure that repeats, which makes it possible to isolate components and analyze them
  • Discrete events makes it relatively easy to analyze:
    • Pitches -> plate appearances -> innings -> games -> seasons -> $

Lots of data available
  • Data going back to 1871

Overall: Excellent system to practice using data to answer real questions
Background knowledge

No prerequisites

Familiarity with the basic rules of baseball will be useful
  • We will briefly go over the rules and watch an innings of baseball next class
    • We can also arrange additional time for students who want
Required Textbooks

Teaching Statistics Using Baseball (TSUB)
Big Data Baseball (BDB)
Highly recommended textbook

Analyzing Baseball Data with R
  • Available as an e-book through Hampshire
Other recommended books

Moneyball
The Sabermetric revolution
Topics covered

Visualizing and exploring a single batch of data
Comparing batches of data and standardization
Relationships between measured variables
Introduction to probability
Introduction to statistical inference
  • Hypothesis tests and confidence intervals
If there is time:
  • Modeling baseball games
  • Other sabermetric topics
Types of questions we will be trying to answer using statistics

How much more valuable is a home run compared to a single?

Which statistics best capture a baseball players ability
  • i.e., is on base percentage a better measure than batter average?

Who is the best baseball hitter of all time?

Are certain baseball players streaky or clutch hitters?

Did the Tigers make a good decision signing Miguel Cabrera for $292 million over the next 8 years?
Class assignments

1. Weekly worksheets that involve R and other problems
   • Assigned on Thursdays and they are due the following Sunday at midnight
   • Also class readings from Teaching Statistics Using Baseball

2. Weekly ‘quote and reactions’ to Big Data Baseball chapter

3. One ‘better know a player presentations’ (on Mondays)
   • 5-10 minute presentation about a baseball/softball player (or another topic related to the class)

4. A final project where you analyze related to baseball (or the class more generally) and write up a 5-10 page paper about the results

5. A class presentation about your final project
Policies

Very important to turn your work in on time!
  • Three strikes late policy
  • If you turn in three assignments late you don’t get an evaluation

Attend class

You can use a laptop for notes, but obviously do not use it for anything not related to the class

Check your Hampshire email and/or setup mail forwarding. I will send announcements to your Hampshire email account
Policies

Incompletes will only be given out in very exceptional circumstances

Special needs: please let me know if you have a disability or special need. Also it would be a good to talk to Aaron Ferguson – x5498

Academic dishonesty: you can work on the worksheets with others but the work you turn in needs to be your own (i.e., you need to understand the concepts).
To make this class a success...

Actively engaged in the class
  • Do the readings and join in the discussions

Community learning environment
  • Share interesting things you’ve learned
  • I want to learn from you too

Following your interests
  • I can adapt the class to people’s interests
Class survey

In order for me to get to know you and to better adjust the class to your interests, please fill out the survey
The history of baseball statistics
Early statistics

Henry Chadwick (1824-1908) created the first boxes core in the 1859 issue of Clipper.

- First to use K for strike out, said to have invented batting average and earned run average
- Did not record walks because he did not feel they reflected a batters skill
Classic statistics

Most prominent hitting statistics:
  • Batting average, RBIs, and home runs

Most prominent pitching statistics:
  • Wins, earned run average, strike outs
Sabermetrics

Questioned how useful traditional measures of performance, such as batting average or pitcher wins

Sabermetrics definitions:
1. The empirical or mathematical/statistical study of baseball
2. "the search for objective knowledge about baseball”
   - Bill James

Name comes from ‘Society for American Baseball Research’ (SABR), a group started in 1971
- Pre-computers, had to compile all information from box scores by hand since there was no encyclopedia that had game by game data

Sabermetrics first widely introduce to the public in 1982 with the publication of Bill James Baseball Abstract
Baseball data sets

Lahman Database: Season-by-season data

Retrosheet Game-by-Game data

Retrosheet Play-by-Play data

PITCHf/x: Pitch-by-Pitch data (location, pitch type)
A few prominent Sabermetric publications/websites

Society for American Baseball Research
Bill James Online
Baseball Analysts
Baseball Prospectus
Beyond the Box Score
Fan Graphs
The Hardball Times
Tango Tiger
Moneyball

Story about how the Billy Bean, the general manager of the A’s, was able to put together a top ranked team in 2002 on a tight budget by finding undervalued players using advanced statistics.

Some of the claims of the book might be exaggerated by the book had a bit impact on the expansion of major league clubs doing advanced data analyses.
Big Data Baseball

More recent sabermetric advances

• 2013 Pittsburgh Pirates
Increasing analytics to gain an edge

http://online.wsj.com/news/article_email/baseballs-science-experiment-1411135882-IMyQjAxMTA0NzE3OTIxNDkwWj
Common baseball statistics

Let’s look at some baseball cards...
Structured data
Lahman Database – Individual player yearly batting statistics

Cases

<table>
<thead>
<tr>
<th>playerID</th>
<th>yearID</th>
<th>stint</th>
<th>teamID</th>
<th>lgiID</th>
<th>G</th>
<th>G_batting</th>
<th>AB</th>
<th>R</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardsda01</td>
<td>2004</td>
<td>1</td>
<td>SFN</td>
<td>NL</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2006</td>
<td>1</td>
<td>CHN</td>
<td>NL</td>
<td>45</td>
<td>43</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2007</td>
<td>1</td>
<td>CHA</td>
<td>AL</td>
<td>25</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2008</td>
<td>1</td>
<td>BOS</td>
<td>AL</td>
<td>47</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2009</td>
<td>1</td>
<td>SEA</td>
<td>AL</td>
<td>73</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2010</td>
<td>1</td>
<td>SEA</td>
<td>AL</td>
<td>53</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2012</td>
<td>1</td>
<td>NYA</td>
<td>AL</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>aaronha01</td>
<td>1954</td>
<td>1</td>
<td>ML1</td>
<td>NL</td>
<td>122</td>
<td>122</td>
<td>468</td>
<td>58</td>
<td>131</td>
</tr>
</tbody>
</table>

Data taken from the Lahman Batting dataset
Example Dataset – Individual player yearly statistics

### Variables

<table>
<thead>
<tr>
<th>playerID</th>
<th>yearID</th>
<th>stint</th>
<th>teamID</th>
<th>lgID</th>
<th>G</th>
<th>G_batting</th>
<th>AB</th>
<th>R</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>aardsda01</td>
<td>2004</td>
<td>1</td>
<td></td>
<td>122</td>
<td>122</td>
<td>468</td>
<td>58</td>
<td>131</td>
</tr>
<tr>
<td>2</td>
<td>aardsda01</td>
<td>2006</td>
<td>1</td>
<td></td>
<td>122</td>
<td>122</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>aardsda01</td>
<td>2007</td>
<td>1</td>
<td></td>
<td>122</td>
<td>122</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>aardsda01</td>
<td>2008</td>
<td>1</td>
<td></td>
<td>122</td>
<td>122</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>aardsda01</td>
<td>2009</td>
<td>1</td>
<td></td>
<td>122</td>
<td>122</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>aardsda01</td>
<td>2010</td>
<td>1</td>
<td></td>
<td>122</td>
<td>122</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>aardsda01</td>
<td>2012</td>
<td>1</td>
<td></td>
<td>122</td>
<td>122</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>8</td>
<td>aaronha01</td>
<td>1954</td>
<td>1</td>
<td>ML1</td>
<td>NL</td>
<td>122</td>
<td>468</td>
<td>58</td>
<td>131</td>
</tr>
</tbody>
</table>
Example Dataset – Individual player yearly statistics

<table>
<thead>
<tr>
<th>playerID</th>
<th>yearID</th>
<th>stint</th>
<th>teamID</th>
<th>lgID</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardsda01</td>
<td>2004</td>
<td>1</td>
<td>SFN</td>
<td>NL</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2006</td>
<td>1</td>
<td>CHN</td>
<td>NL</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2007</td>
<td>1</td>
<td>CHA</td>
<td>AL</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2008</td>
<td>1</td>
<td>BOS</td>
<td>AL</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2009</td>
<td>1</td>
<td>SEA</td>
<td>AL</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2010</td>
<td>1</td>
<td>SEA</td>
<td>AL</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2012</td>
<td>1</td>
<td>NYA</td>
<td>AL</td>
</tr>
<tr>
<td>aaronha01</td>
<td>1954</td>
<td>1</td>
<td>ML1</td>
<td>NL</td>
</tr>
</tbody>
</table>

Cases

Variables
# Categorical and Quantitative Variables

Cases

<table>
<thead>
<tr>
<th>playerID</th>
<th>yearID</th>
<th>stint</th>
<th>teamID</th>
<th>lgID</th>
<th>G</th>
<th>G_batting</th>
<th>AB</th>
<th>R</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardsda01</td>
<td>2004</td>
<td>1</td>
<td>SFN</td>
<td>NL</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2006</td>
<td>1</td>
<td>CHN</td>
<td>NL</td>
<td>45</td>
<td>43</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2007</td>
<td>1</td>
<td>CHA</td>
<td>AL</td>
<td>25</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2008</td>
<td>1</td>
<td>BOS</td>
<td>AL</td>
<td>47</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2009</td>
<td>1</td>
<td>SEA</td>
<td>AL</td>
<td>73</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2010</td>
<td>1</td>
<td>SEA</td>
<td>AL</td>
<td>53</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aardsda01</td>
<td>2012</td>
<td>1</td>
<td>NYA</td>
<td>AL</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>aaronha01</td>
<td>1954</td>
<td>1</td>
<td>ML1</td>
<td>NL</td>
<td>122</td>
<td>122</td>
<td>468</td>
<td>58</td>
<td>131</td>
</tr>
</tbody>
</table>
Explanatory and Response Variables

Sometimes we use one variable (the explanatory variable) to understand/predict another variable (the response variable)
Another Dataset – 2014 Team statistics

Variables

<table>
<thead>
<tr>
<th>Cases</th>
<th>Tm</th>
<th>X.Bat</th>
<th>BatAge</th>
<th>R.G</th>
<th>G</th>
<th>PA</th>
<th>AB</th>
<th>R</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ARI</td>
<td>52</td>
<td>27.6</td>
<td>3.80</td>
<td>162</td>
<td>6089</td>
<td>5552</td>
<td>615</td>
<td>1379</td>
</tr>
<tr>
<td>2</td>
<td>ATL</td>
<td>39</td>
<td>26.8</td>
<td>3.54</td>
<td>162</td>
<td>6064</td>
<td>5468</td>
<td>573</td>
<td>1316</td>
</tr>
<tr>
<td>3</td>
<td>BAL</td>
<td>39</td>
<td>28.3</td>
<td>4.35</td>
<td>162</td>
<td>6130</td>
<td>5596</td>
<td>705</td>
<td>1434</td>
</tr>
<tr>
<td>4</td>
<td>BOS</td>
<td>50</td>
<td>29.2</td>
<td>3.91</td>
<td>162</td>
<td>6226</td>
<td>5551</td>
<td>634</td>
<td>1355</td>
</tr>
<tr>
<td>5</td>
<td>CHC</td>
<td>48</td>
<td>26.8</td>
<td>3.79</td>
<td>162</td>
<td>6102</td>
<td>5508</td>
<td>614</td>
<td>1315</td>
</tr>
<tr>
<td>6</td>
<td>CHW</td>
<td>36</td>
<td>27.7</td>
<td>4.07</td>
<td>162</td>
<td>6077</td>
<td>5543</td>
<td>660</td>
<td>1400</td>
</tr>
<tr>
<td>7</td>
<td>CIN</td>
<td>45</td>
<td>28.9</td>
<td>3.67</td>
<td>162</td>
<td>5978</td>
<td>5395</td>
<td>595</td>
<td>1282</td>
</tr>
<tr>
<td>8</td>
<td>CLE</td>
<td>43</td>
<td>28.5</td>
<td>4.13</td>
<td>162</td>
<td>6222</td>
<td>5575</td>
<td>669</td>
<td>1411</td>
</tr>
</tbody>
</table>
Next class – examining a single batch of data

Fill out survey online! (link is on Moodle)

Read chapter 1 of Big Data Baseball and post a quote and reaction to the Moodle forum by midnight on Sunday

Read chapter 1 of Teaching Statistics Using Baseball