Baseball is like a church, many attend but few understand.

Wes Westrum
Outline

- Data use example
- Sabermetrics
- Data sources
- Tools
- Resources
- Further analysis examples
This presentation is for folks who want to get data, crunch data and analyze data.

If you want to view the highlights, please watch SportsCenter on ESPN.
The presentation PowerPoint slides and all of the presentation references can be found at:

http://mercury.webster.edu/aleshunas
It’s been going on in baseball for 100 years. When pitchers make quality pitches, batters do not make good contact.

Tony La Russa
Batter Analysis

Batters with AVG > .299

Year


Lehman
George William “Bill” James (born October 5, 1949, in Holton, Kansas) is a baseball writer, historian, and statistician whose work has been widely influential. Since 1977, James has written more than two dozen books devoted to baseball history and statistics. His approach, which he termed sabermetrics in reference to the Society for American Baseball Research (SABR), scientifically analyzes and studies baseball, often through the use of statistical data, in an attempt to determine why teams win and lose. In 2006, Time magazine named him in the Time 100 as one of the most influential people in the world.
Society for American Baseball Research (SABR) was established in Cooperstown, New York in August, 1971. Their mission is to foster the study of baseball past and present, and to provide an outlet for educational, historical and research information about the game.

http://www.sabr.org/
Data Sources (pay)

- The Baseball Cube
- Baseball Info Solutions
- Baseball Prospectus
- Baseball Reference

- All of these sources offer good service, for a fee
Data sources (free)

- Society for American Baseball Research (SABR)
- Lahman database
- Retrosheet
- Baseball Index
- The Official MLB Website

- These are good low-cost data sources
Compiled annual data for batters, fielders, pitchers, managers and teams

Covers all major league baseball from 1870 to present (current to 2007)

MS Access DB - http://www.baseball1.com/

MySQL DB - http://www.baseball-databank.org/
Retrosheet Data

- Retrosheet is a non-profit corporation, incorporated in the State of Delaware in 1994.

- Retrosheet's work has three distinct aspects.
  - First is the collection of the game accounts, which have been obtained from several sources.
  - The second activity is the translation of these accounts to a unified, modern system which is essential since there are an extraordinary variety of scoring systems which have been used.
  - The final activity is the entry of the translated accounts into the computer.

- The first meeting of Retrosheet, Inc. was held in Arlington, TX on June 17, 1994.
Salary Analysis

SLN Team Salaries

Year

Salary

0
10000000
20000000
30000000
40000000
50000000
60000000
70000000
80000000
90000000
100000000
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
Salary Analysis (STL vs. CHC)

Team Salary Comparison (STL vs CHN)
Data Manipulation & Analysis Tools (pay)

- Microsoft Office
  - Excel (spreadsheet)
  - Access (database)
- SPSS, SAS (statistical analysis software)
- Oracle (database) *
Microsoft Excel & Access

- Familiar
- Established applications
- Reference materials
- User generated files
- Cost: $500 (retail – MS Office Professional)
GUI based statistical software
Industry leader
http://www.spss.com/
Cost: $639 (Base – Higher Ed, $1699 Commercial)
Data Manipulation & Analysis Tools (free)

- Open Office
  - Calc (spreadsheet)
  - Base (database)
- R, Octave, Sage (statistical and mathematical analysis)
- MySQL (database)
- Perl (programming language)
Open Office Calc & Base

- OpenOffice.org is a multiplatform and multilingual office suite and an open-source project
- Compatible with all other major office suites
- The product is free to download, use, and distribute
- http://www.openoffice.org/
R, Octave & Sage

- R is a free software environment for statistical computing and graphics

- GNU Octave is a high-level language, primarily intended for numerical computations
  - It is mostly compatible with Matlab
  - It is also freely redistributable software

- SAGE is a viable free open source alternative to Magma, Maple, Mathematica, and Matlab
MySQL Database

- The world's most popular open source database (free for personal use)
- http://www.mysql.com/
- Learning MySQL by Seyed Tahaghoghi & Hugh Williams
Perl Language

- Programming PERL by Larry Wall, Tom Christiansen & Jon Orwant

http://www.oreilly.com/
Which MLB manager has the most career wins?

We’ll use the Lahman database on MySQL
Salary Analysis - revisited

- Let’s redo the team salary analysis to show the steps in developing the chart
- We’ll add Houston to the mix, just to add additional interest
Retrosheet Download and Processing

- Let’s download a file from Retrosheet and process it.
  - Download
  - Bevent processing
  - Opening in a spreadsheet
  - Loading into MySQL
Double Play Analysis

- Which MLB batter hit into the most double plays in 2007?
- We’ll use Retrosheet data, again on MySQL
The goal, in baseball, is to score more runs than your opponent. You score runs by putting runners on base. Therefore it seems counter-intuitive for the defense to put a runner on base. Reasons for an intentional walk:
- Avoid a dangerous hitter
- Set up a double play situation
In 2006 NL play, 37.6% of IWs were followed by a run scored

An average NL batter had a .272 BA (>149 AB)

The best batter had a .344 BA

A hit resulted in a run 53.2%

Advantage = P(hit) * P(run)
  - Average hitter = (.272)*(.532) = 0.145
  - Best hitter = (.344)*(.532) = .183
  - Intentional walk = (1.000)*(.376) = 0.376
Further Reading

Baseball Hacks by Joseph Alder
Summary

- Data use example
- Sabermetrics
- Data sources
- Tools
- Resources
- Further analysis examples
You know how to pitch God?
Hard stuff inside, then down and away, and if you get it there you’ll get Him out. Even though He’ll know it’s coming. Or at least they say He knows.

Jim Lefebvre
References

Will, George (1990), Men at Work the Craft of Baseball, New York, Macmillan Publishing Company
Bill James Article,
http://en.wikipedia.org/wiki/Bill_James
My Web Site,
http://mercury.webster.edu/aleshunas