

Baseball Data at Season, Play-by-Play, and Pitch-by-Pitch Levels

Jim Albert
Bowling Green State University

Journal of Statistics Education Volume 18, Number 3 (2010)
www.amstat.org/publications/jse/v18n3/albert.pdf

Copyright © 2010 by Jim Albert all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Batting; Density estimate; Lowess smoother; Pitching; Plate appearance; Streakiness.

Abstract

Baseball provides a rich context to learn statistical concepts, and one can learn much about baseball players and teams though exploratory analyses. We describe three readily available extensive baseball datasets that describe baseball at the season, play-by-play, and pitch-by-pitch levels. We use R to illustrate some sample analyses with these datasets and provide a list of possible explorations for the student. There is a review of the literature and a description how these datasets can be used to communicate statistical concepts.

1. Introduction

Sports provides an interesting setting to illustrate exploratory and confirmatory statistical analyses of data, and the analysis of sports data is a popular project topic both at the undergraduate and graduate levels. One challenge in the development of projects is finding appropriate sports datasets that are relatively easy to access and

are sufficiently rich to address the sports questions of interest. Baseball is arguably the most statistical sport with regards to the availability of detailed data on players and teams. Baseball players are known and evaluated by an assortment of statistical measures and there is an ongoing effort to use data to better understand the game.

The purpose of this paper is to introduce three sophisticated baseball databases that are useful in investigating many interesting questions about the game. These databases represent different layers of baseball, from a coarse season description to a detailed pitch-by-pitch description. We observe baseball by seasons, next by plate appearances in games within seasons, and finally by pitches within a plate appearance. Lahman's database (Lahman 2010), described in Section 3, gives seasonal pitching and hitting data for all players in the history of professional baseball. The Retrosheet organization (Retrosheet 2010) is a grassroots effort to computerize play-by-play records for all baseball games in recent seasons and Section 4 describes the format of this play-by-play data for a particular season. Last, the PITCHf/x system (Nathan 2009) is a recent technology that computes detailed information such as the speed, movement, and location on all pitches thrown in a game. Section 5 describes one particular dataset giving information about all pitches thrown by twenty starting pitchers in the 2009 season.

Each section describes the source of the dataset and information about the variables. Each dataset is available both as a text file and an R worksheet and is easily input into most statistical packages. A sample analysis is provided in each section to illustrate inputting the data and performing some exploratory analyses. A number of interesting questions are posed in each section to motivate the interested student or researcher to perform his/her own baseball analysis. There is a review of literature that use these datasets and a description of how the baseball data can be used in an introductory statistics class.

2. Downloading Files

The reader is encouraged to download data files and documentation files prior to reading this paper. While the paper can be read without doing the analyses presented, many readers will want to do the analyses as they read the paper. If you would like to do the illustrated analyses as you read the paper, please download the following files from the *JSE* website by clicking on each file's link. Note that the files mlb_batting.dat, playbyplay2008.dat, and pitchfx.dat are quite large. (An alternative to downloading the eight files is to download a single R package `Baseball`. The appendix gives more details for accessing this package.)

Data File	Associated Documentation File
mlb_batting.dat	mlb_batting.txt
playbyplay2008.dat	playbyplay2008.txt
roster2008.dat	roster2008.txt
pitchfx.dat	pitchfx.txt

3. Lahman Baseball Database

3.1 Description

One of the most complete databases of baseball statistics is currently available on Sean Lahman's Baseball Archive website at www.baseball1.com. This database contains seasonal pitching, hitting, and fielding statistics for all players in Major League Baseball from the first professional league in 1871, to the formation of Major League Baseball in 1901, to the present day. In addition, this database includes a number of supplemental tables including All-Star game appearances, Hall of Fame voting data, managerial statistics, and batting and pitching statistics for players in the post-season.

These data are available as a single Microsoft Access or SQL database. Alternatively, the data is available as a collection of comma delimited files that are easily imported into a statistics package such as R. Here we focus on the use of two popular data files: Batting.csv that contains seasonal batting statistics for all players in Major League Baseball and Master.csv that contains biographical information about all of the players in the database.

The datafiles Batting.csv and Master.csv were imported into R and several editing changes were made to create a new data file "mlb_batting.dat". In the original batting dataset Batting.csv, a single row was used for each player's batting statistics for each season for each team played. If a player was traded once during the season, there would be two rows in the dataset corresponding to the player's two teams. For many uses, it is convenient to reformat the data so that the total statistics for a player in a season are represented by a single row. Also, since the Batting.csv file only contains the player id code, we used the Master.csv file to add the first and last names for each player to the new dataset.

The batting dataset is available as a text file "mlb_batting.dat" that can be read into R by use of the `read.table` function. This function reads the dataset and creates a data frame with the name `batting`, where the columns of the data frame correspond to the variables collected for each baseball hitter.

```
> batting = read.table("mlb_batting.dat", header = TRUE)
```

(Note that the data file must be in the same location as the R working directory. To find out the R working directory, use `getwd()`.) Table 1 gives all of the variables and descriptions for the data frame `batting`.

Table 1. Variables and descriptions of data frame `batting` created using the text file "mlb_batting.dat".

Variable	Description	Variable	Description
first.name	player's first name	cs	caught stealing
last.name	player's last name	bb	base on balls
name	player's id code	so	strikeouts
year	season	ibb	intentional base on balls
game	games played	hbp	hit by pitch
ab	at-bats	sh	sacrifice hits
r	runs scored	sf	sacrifice flies
h	hits.	gdp	grounded into double plays
x2b	doubles	age	player's age
x3b	triples	obp	on-base percentage
hr	home runs	slg	slugging percentage
rbi	runs batted in	ops	ops statistic
sb	stolen bases	pa	plate appearances

It is straightforward in R to access particular subsets of the data of interest. For example, suppose one is interested in accessing the batting statistics for Mark McGwire. One uses the `subset` function to access the rows of the dataset where the player's first name is "Mark" and the last name is "McGwire", and the subsetted data is stored in the data frame `mg`.

```
> mg=subset(batting,
+   batting$first.name=="Mark"&batting$last.name=="McGwire")
```

We display the year, number of at-bats, hits, and home runs for all of the seasons of McGwire's career.

```
> mg[,c("year","ab","h","hr")]
```

```
      year  ab   h hr
63783 1986  53  10  3
```

```
63784 1987 557 161 49
63785 1988 550 143 32
63786 1989 490 113 33
63787 1990 523 123 39
63788 1991 483  97 22
63789 1992 467 125 42
63790 1993  84  28  9
63791 1994 135  34  9
63792 1995 317  87 39
63793 1996 423 132 52
63794 1997 540 148 58
63795 1998 509 152 70
63796 1999 521 145 65
63797 2000 236  72 32
63798 2001 299  56 29
```

3.2 A Sample Analysis

Since the data frame `batting` contains all of the batting data for all years of Major League Baseball, an interesting exploration is to see how some standard measures of batting performance have changed over time. The traditional measure of batting performance is the batting average *avg* defined as the number of hits *h* divided by the number of official at-bats *ab*.

First we use the `aggregate` function to find the sum of the hits and sum of the at-bats for each year of baseball. There are three arguments to `aggregate`, the variables (`h` and `ab`) that we wish to work with, the list of grouping elements (`year`), and the summary statistic for each group (`sum`). We compute the season batting averages by dividing the yearly hits by the yearly at-bats and storing the vector result in the variable `avg`. We construct a plot of batting average against year and superimpose a lowess smoothing curve (using the `lines(lowess())` function) to pick up the general pattern. (The argument `f=1/8` controls the degree of smoothing of the curve.) The resulting graph is displayed in Figure 1.

```
> hits.ab=aggregate(batting[,c("h","ab")],
+   list(year=batting$year),sum,na.rm=TRUE)
> avg=hits.ab$h/hits.ab$ab
> plot(hits.ab$year, avg, xlab="Year", ylab="Batting Average")
> lines(lowess(hits.ab$year, avg, f = 1/8))
```

Note that the batting average of the MLB players went through some dramatic changes from 1876 through 1930, exhibited a relatively steady decrease from 1930 to 1965, and has shown a gradual increase in the period 1965 to 2009.

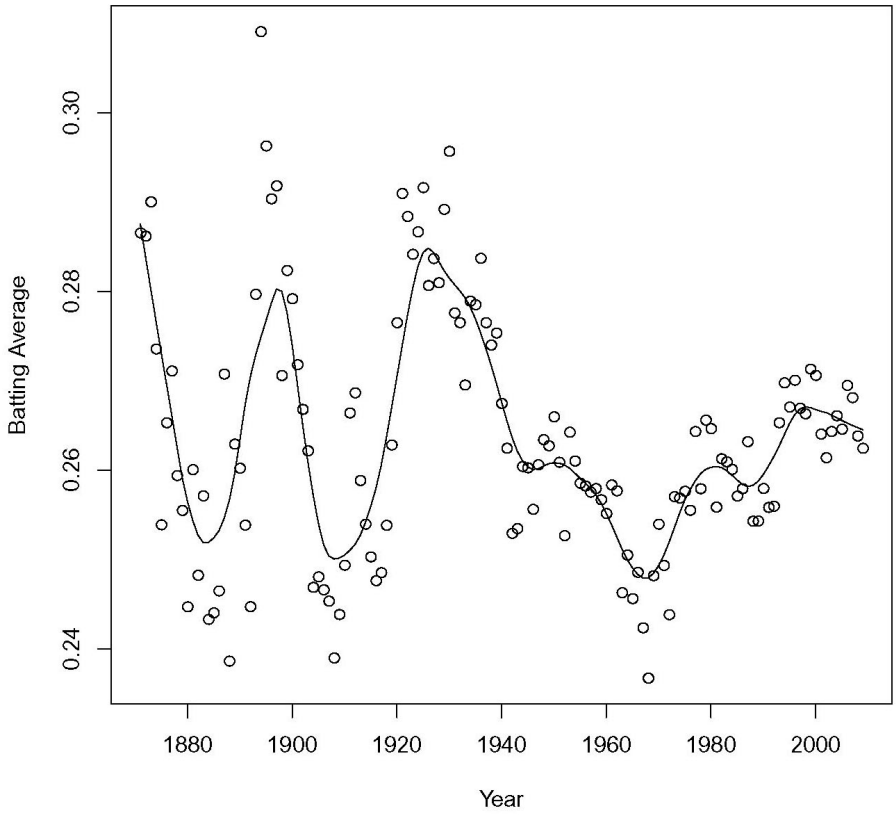


Figure 1. Time series plot of the batting average $avg = h/ab$. Although the batting average has gone through many flucuations over the years, this batting statistic has stabilized at about 0.265 in the last 10 seasons.

Other insights into the history of the game can be found using less familiar hitting measures. One negative measure of hitting is the strikeout rate found by dividing the count of strikeouts by the number of at-bats: $so.rate = so/ab$. Using the following R commands, we compute the strikeout rate for all seasons and graph the rate against the season in Figure 2.

```
> so.ab=aggregate(batting[,c("so","ab")],
+   list(year=batting$year),sum,na.rm=TRUE)
> so.rate=so.ab$so/so.ab$ab
> plot(so.ab$year, so.rate, xlab="Year", ylab="Strikeout Rate")
> lines(lowess(so.ab$year, so.rate, f=1/8))
```

Note that there was much variability in strikeout rates in the early years of baseball. The cluster of zeros in the plot can be explained by the fact that strikeouts were not recorded in the 1897–1909 seasons. However in recent years, the overall strikeout rate has shown a steady increase from 1920 to 1967, a decline from 1967 to 1980, and a second increase from 1980 to 2009.

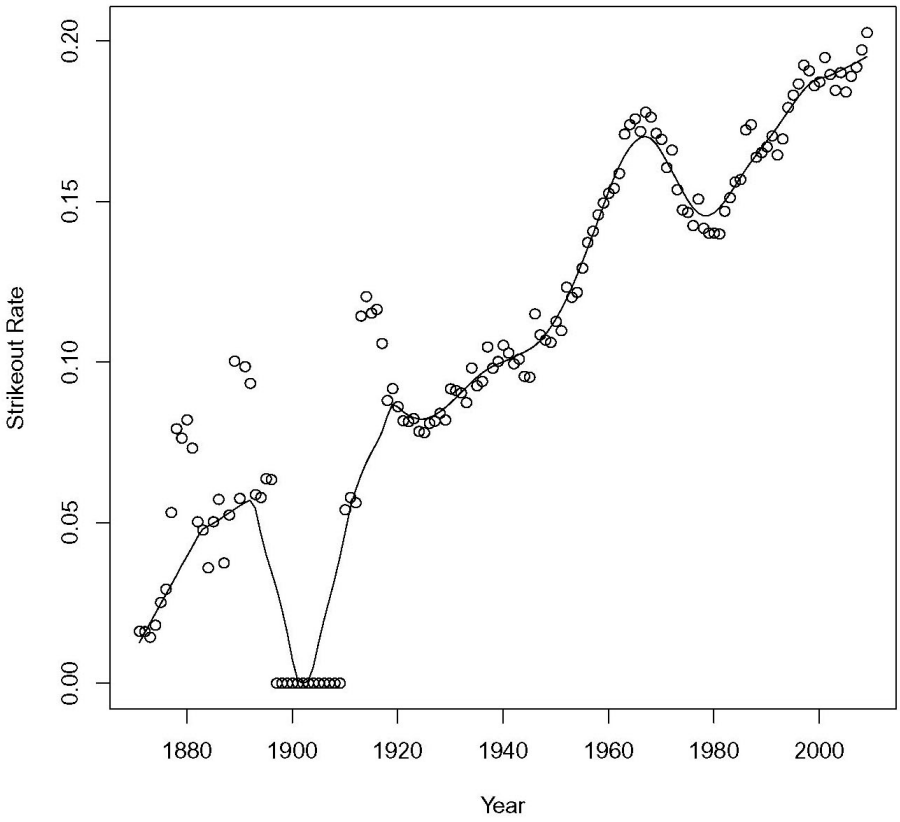


Figure 2. Time series plot of the strikeout rate so/ab. The tendency of a batter to strike out has increased from the seasons 1920 to 1967 and increased again from the seasons 1980 to 2009.

3.3 Possible Explorations

Many interesting questions can be addressed using this hitting dataset.

- One can look at the career trajectory of different players and see when each achieved his peak performance. At what age did Babe Ruth achieve a peak performance with respect to home run rate hr/ab? Did modern sluggers such as Mickey Mantle, Hank Aaron, Barry Bonds, and Mark McGwire peak at similar ages as Babe Ruth with respect to home run rate?

- One can compare the batting careers of two current players such as Derek Jeter and Alex Rodriguez. One can compare the season statistics of these two players by using different batting measures. Which player was generally more successful in getting on-base? Which player was more likely to get extra-base hits?

- Players from different eras such as Ted Williams and Tony Gwynn can be compared by use of suitable adjustments. Which achievement was more noteworthy—Ted Williams' .343 batting average in 1949 or Tony Gwynn's batting average of .358 in 1993? Comparing these two batting averages on face value is deceptive since the distribution of batting averages differs for the two seasons. One can put Williams and Gwynn's batting averages on the same scale by finding the mean and standard deviation of the batting averages of all players in the 1949 and 1993 seasons and computing the standardized scores of Williams' and Gwynn's batting averages using these summary statistics.

- Figure 1 and Figure 2 focused on the average performances of batting statistics over seasons. How has the variability of batting performances changed over seasons? It would be interesting to explore how the standard deviation of batting averages has changed over seasons. The pool of baseball talent has greatly increased over the history of baseball and this change in the pool of talent will impact the variability of player batting averages.

3.4 Literature and Use in Teaching

Season to season hitting baseball data has been used in the literature for different purposes. Berry (2000b) and Hoffman (1989) describe constructing good measures of hitting performance. Berry (2000a) and Wang (2007) discuss what hitting statistics are necessary for entrance in the Baseball Hall of Fame. By use of a season-by-season pitching dataset from the Lahman database, Bradlow, Jensen, Wolfers, and Wyner (2008) and Albert (2009a) focus on the career trajectories of pitchers. Schmotzer, Kilgo, and Switchenko (2009) use this data to investigate the effect of steroids on offensive performance.

In an introductory statistics class, this dataset can be used to look at the careers of individual players and to compare players. Albert (2002b) and Albert (2003), Chapters 1 and 2 uses data on individual players to illustrate exploratory methods on a single batch and in comparing batches. There are many measures for batting performance and one can explore the relationships between the different measures by scatterplots and a correlation analysis. Comparisons can also be made, in say home run hitting, from players from different eras. Albert (2003), Chapter 4, uses baseball

hitting from two consecutive seasons to demonstrate the regression effect, and uses batting averages of two players from different eras to illustrate the use of standardized scores.

4. Retrosheet Data

4.1 Description

The Retrosheet organization was founded in 1989 for the purpose of collecting play-by-play information about individual baseball games. Currently, one is able to download play-by-play data for individual seasons directly from the website www.retrosheet.org.

If one clicks on the 2008 season under the "Regular Season Event Files" at www.retrosheet.org/game.htm, one downloads a zip file containing sixty files. Thirty of the files contain the play-by-play records for all of the home games at each of the thirty home stadiums, and the remaining thirty files contain detailed information about the player rosters for the teams. One particular play-by-play file, say `2008cin.evn`, contains a large amount of information about the games in a format not convenient for the typical statistics package. One uses a special DOS program `bevent` (supplied at the Retrosheet website) to extract variables and put the dataset in a convenient format where each row of the new file corresponds to a particular baseball play.

All thirty files for the 2008 season were combined to create a single data frame for all baseball plays with 193,492 rows and 38 variables. The text file "playbyplay2008.dat" is read into R and saved in the data frame named `pbp` by use of the command

```
> pbp=read.table("playbyplay2008.dat",header=TRUE)
```

All of the team rosters were also combined and saved in a text file `roster2008.dat` that can be read into R and saved into the data frame `roster` by the command

```
> roster=read.table("roster2008.dat", header=TRUE)
```

This data file is helpful in finding the id codes for the batters and pitchers who appear in the play-by-play data frame.

Table 2 gives the variables and descriptions for the data frame `pbp`. Each row gives detailed information about a particular plate appearance during a game, including the names of the batter and the pitcher, and details about the game situation such as the inning, number of bats, and runners on base. A description of the result of the plate appearance, and how this result changes the number of outs and runners on base, is given.

Table 2. Variables and descriptions of play-by-play data frame `pbp` created from the text file "playbyplay2008.dat".

Variable	Description	Variable	Description
game.id	game id	def_pos	defensive position of batter
v_team	code for visiting team	batting_pos	position in batting order
inning	inning of game	event_code	numerical code for batting event
team_at_bat	id of team at bat	bevent_flag	end of batting appearance
outs	number of current outs	ab_flag	indicator of at-bat
balls	number of balls	hit_value	value of hit
strikes	number of strikes	sh_flag	sacrifice hit?
v_score	current visitor score	sf_flag	sacrifice fly?
h_score	current home score	outs_play	number of outs recorded
batter	id code of batter	rbi_play	number of rbi's credited
batter_hand	batter side	wp_play	wild pitch?
pitcher	id code of pitcher	pb_flag	passed ball?
pitcher_hand	pitcher side	nerrors	number of errors on play
b1_runner	code of runner on first	batter_dest	base reached by batter
b2_runner	code of runner on second	b1_runner_d	new runner on first
b3_runner	code of runner on third	b2_runner_d	new runner on second
event	code description of event	b3_runner_d	new runner on third
leadoff_flag	leadoff hitter?	date	date of game
pinchhit_flag	pinch hitter?	h_team	code for home team

To help in understanding the variables, consider one particular plate appearance during the 2008 season. On May 10, the Philadelphia Phillies slugger Ryan Howard hit a home run in the second inning in an away game against the San Francisco Giants pitcher Tim Lincecum. To learn about the circumstances of this particular home run, we find the corresponding line in the play-by-play database. We use the R `subset` function where we find the subset of the dataframe `pbp` where the date is "5100" (May 10), the visiting team was "PHI", the batter was "howar001" (code for Ryan Howard), and the inning was 2. This play line is stored in the vector `howard.hr` and we display several variables.

```
> howard.hr=subset(pbp,pbp$date=="5100"&pbp$v_team=="PHI"&
+   pbp$batter=="howar001"&pbp$inning==2)
```

```
> howard.hr[,c("inning","outs","balls","strikes",
+   "b1_runner","b2_runner","b3_runner","event")]
```

```
        inning outs balls strikes b1_runner b2_runner b3_runner  event
156072      2    0     1       0                                HR/7/L
```

We learn from these variables that

- As `inning = 2`, `outs = 0`, `balls = 1`, and `strikes = 0`, this home run was hit in the 2nd inning when there were no outs on a pitch count with 1 ball and 0 strikes.
- Since `b1_runner`, `b2_runner`, and `b3_runner` are all empty strings, all of the bases were empty, which indicates it was a solo home run. At the conclusion of the play, the bases remained empty.
- Since `event = "HR/7/L"`, we see the home run was hit to left field.

4.2 A Sample Analysis

Baseball fans are generally fascinated with the variability of hitting and pitching performances in short time periods. Announcers typically talk about the performance of hitters in the most recent at-bats; one may hear comments like "Jeter has not gotten a hit in his last 20 at-bats", indicating that Jeter went through a hitting slump. One can explore the pattern of hitting streaks (clusters of successes) and slumps (clusters of failures) by use of this Retrosheet data.

To begin, we find the id code for Derek Jeter by using the `roster` data frame.

```
>  with(roster,abbrev[first.name=="Derek"&last.name=="Jeter"])
```

```
[1] jeted001
1291 Levels: aardd001 aberr001 abreb001 accaj001 aceva001 ... zumaj001
```

We see the id code for Jeter is `"jeted001"`. Next, we use the `subset` function to select the plays where Derek Jeter was batting and an official at-bat was recorded. The Jeter data is stored in the data frame `jeter`.

```
> jeter=subset(pbp,pbp$batter=="jeted001"&pbp$ab_flag==TRUE)
```

We define a new vector `y`, where a component is equal to 1 or 0 depending if the hit value (number of bases of the base hit) of the at-bat was one or greater.

```
> y=ifelse(jeter$hit_value>0,1,0)
```

The vector `y` indicates if Jeter was successful or not in each of his 596 at-bats in the 2008 season.

To look at the pattern of Jeter's hitting over short time periods, we use the `filter` function to compute moving proportions of hits using a window of 20 at-bats. We can graph a time series line plot using `plot` with the `type="l"` argument to see the variability of Jeter's hitting over twenty at-bats.

```
> m.avg=filter(y,rep(1/20,20))
> plot(m.avg,type="l",ylab="Moving Batting Average")
```

We see some interesting patterns in this graph displayed in Figure 3. Jeter had a hitting slump at at-bat 180, but showed some streaks of successful hitting towards the end of the season (beginning at about at-bats 440 and 530).

4.3 Possible Explorations

The Retrosheet dataset gives detailed information about each batting play including information about the opposing team, the inning and pitch count, and runners on base. This play-by-play data invites a number of different explorations.

- As a followup to the study of Section 3.2, how can one quantify the streakiness of a baseball hitter? Using a streaky criterion such as the longest streak of consecutive outs, find the players who appear to be unusually streaky and the players who are unusually non-streaky or consistent.
- One can explore the pattern of run scoring by inning. Are there particular innings of the game when runs are more likely to be scored?
- The configuration of a ballpark and the weather of the city can have an impact on run scoring. How does run scoring vary between ballparks?

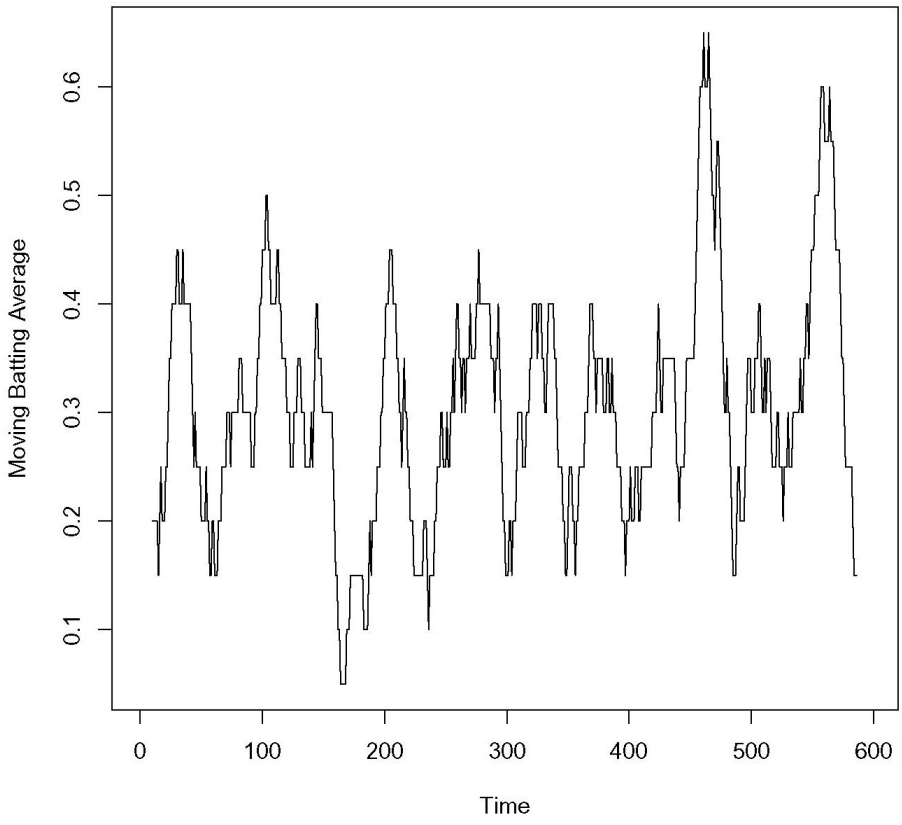


Figure 3. Moving batting average plot for Derek Jeter using a window of 20 at-bats for the 2008 baseball season. This graph is helpful in understanding periods during this season where Jeter was unusually "hot" or "cold".

- Each team plays half of its games at its home ballpark and half of its games away from home. Do teams generally score more runs at home? If so, are there particular teams that have an unusually large home advantage or home disadvantage in scoring runs?

- One can investigate how batters and pitchers perform in different situations. How do batters perform at home and away games, during different innings and base situations during a game, and against different pitchers? To understand these "situational effects", it is helpful to look at a particular effect, say home versus away, for all hitters. What is the general situational bias (e.g., how many batting average points do hitters perform better during home games?) and which players deviate from the general pattern?

4.4 Literature and Use in Teaching

There has been a large amount of literature on hitting streaks in baseball. There is much attention to Joe DiMaggio's 56-game hitting streak (Berry 1991; Chance 2009; Gould 1989; Rockoff and Yates 2009; and Warrack 1995). There have also been investigations in the overall pattern of streakiness of hitting for particular seasons (Albert 2008; Albright 1993). Albert (1998) looked at the pattern of home run hitting of Mike Schmidt over his career and Berry (1999) checks if home run hitters have periods where they are unusually successful.

This play-by-play data is helpful in looking for situational effects. Albert (1994) looked at player batting averages across a number of situations, such as home versus away games, against pitchers of the same arm (as the hitter) or the opposite arm, on grass versus turf fields, etc. Albert (2002a) looked specifically at the abilities of batters to do well in important situations during a game.

Albert (2003), Chapter 8, uses this play-by-play data to describe several problems in statistical inference for an introductory class. For example, batting data of players broken down by home and away games illustrates the notion of bias. There is evidence that the batting abilities of players are enhanced by a fixed amount by playing at home. In the analysis of streak data, one explains what it means (from a modeling perspective) for a player to display "consistent" hitting ability, and one investigates the use of different statistics, say the longest streak or run of consecutive hits, to provide evidence of "true" streakiness.

5. PITCHf/x Database

5.1 Description

During the 2007 baseball season, Major League Baseball (MLB) began a systematic effort to record detailed information about the pitches that are thrown. All baseball stadiums were equipped with video cameras that would track each pitched ball and determine its precise trajectory. From the measurements made from the cameras, one is able to learn about the speed of each pitch at its release point and at the point where it reaches home plate. Also one can measure the amount and angle of the movement of the pitch in the path from the pitcher's release point to crossing the plate. This technology is known as the PITCHf/x system. A good introduction to the PITCHf/x system and the associated variables is given in Nathan (2009).

This pitch-by-pitch data together with extra information such as the inning, batter,

and result of the plate appearance is available for free from the Major League Baseball website http://gd2.mlb.com/components/game/mlb/. The data is available as a collection of xml files, but they can be tedious to download many individual files without use of a scripting language such as perl.

The author was able to collect pitch-by-pitch data for 20 starting pitchers in the 2009 season. The list of pitchers is given in Table 3. Nine of these pitchers, labeled E for elite, are considered among the best pitchers since each received or was nominated for the Cy Young pitching award. Data were collected for each of the games that these pitchers played and the files were combined to create a single data file.

Table 3. List of starting pitchers for which pitch-by-pitch data is collected for the 2009 baseball season. Nine of these pitchers are labeled E (elite) since each was nominated or received the prestigious Cy Young award.

Zack Greinke (E)	Brett Anderson
Roy Halladay (E)	Bronson Arroyo
Danny Haren (E)	Scott Baker
Felix Hernandez (E)	Joe Blanton
Cliff Lee (E)	Scott Feldman
Tim Lincecum (E)	Gavin Floyd
C C Sabathia (E)	Cole Hamels
Justin Verlander (E)	Derek Lowe
Adam Wainwright (E)	Ricky Nolasco
Roy Oswalt	Andy Pettitte

The text file "pitchfx.dat" is read into R and saved in the data frame with name `pitchdata` using the `read.table` function:

```
> pitchdata=read.table("pitchfx.dat",header=TRUE)
```

Table 4 gives the list of variables and associated descriptions for the pitch-by-pitch dataset. Each row of the data contains information about a particular pitch, including its pitch type, speed, movement, and location of the pitch in the strike zone. In addition, there is information about the pitcher, the batter, and the outcome of the plate appearance of the batter. Each pitch results in a transition from a current pitch count, such as 2-1, to a new pitch count, such as 2-2, and the variables `count` and `new_count` give these "current" and "new" counts.

To get a feeling for the data, suppose we wish to focus on the pitch-by-pitch sequence of the first batter of the first game pitched by Roy Halladay during the 2009 season. By use of the `subset` function, we create a new data frame `halladay1` that looks at all pitches where `pitcher = "halladay"`, `game = 1` and `num = 1`.

Table 4. Variables and description of pitch-by-pitch data frame `pitchdata` using the text file "pitchfx.dat".

Variable	Description	Variable	Description
pitcher	name of pitcher	start_speed	starting speed of pitch
game	game number	end_speed	speed of pitch crossing plate
id	pitcher id number	sz_top	top of strike zone
inning	inning of game	sz_bot	bottom of strike zone
num	number of batter	pfx_x	deviation in horizontal direction
batter	batter id number	pfx_z	deviation in vertical location
stand	hitting side of batter	px	pitch location in x direction
b_height	height of batter	pz	pitch location in z direction
p_throws	throwing side of pitcher	pitch_type	pitch classification
des	play description	count	current pitch count
event	result of plate appearance	new.count	new pitch count
brief_event	brief description of result	value	pitch value
des2	pitch outcome	new.count.type	PA event or new count
type	ball, strike, or in-play?	count.adv	pitcher or batter or neutral count

```
> halladay1=subset(pitchdata,pitchdata$pitcher=="halladay"&
+   pitchdata$game==1&pitchdata$num==1)
```

We output several of the variables.

```
> halladay1[,c("des","des2","pitch_type","count","new.count")]
```

```
                            des          des2 pitch_type count new.count
22911 Curtis Granderson walks.          Ball          FC   0-0       1-0
22912 Curtis Granderson walks. Called Strike          FF   1-0       1-1
22913 Curtis Granderson walks.          Ball          CU   1-1       2-1
22914 Curtis Granderson walks.          Ball          FC   2-1       3-1
22915 Curtis Granderson walks.          Ball          FC   3-1      Walk
```

In this particular plate appearance where Curtis Granderson was the batter, the sequence of pitches was "Ball", "Strike", "Ball", "Ball", "Ball", resulting in a walk or base on balls. Using the description of pitch type codes in Table 5, we see the sequence of pitches was FC (a cutter), FF (4-seam fastball) , CU (curve ball), and two cutters. The `count` and `new.count` variables show the change in the pitch count for this batter.

Table 5. Descriptions of pitch type codes in the PitchF/X database.

Abbreviation	Pitch Type
CH	Changeup
CU	Curve ball
FA	Fastball
FC	Cutter
FF	Four-seam fastball
FS	Split-finger fastball
FT	Two-seam fastball
IN	Intentional Ball
PO	Pitchout
SI	Sinker
SL	Slider
UN	Unknown

5.2 A Sample Analysis

To illustrate using the pitch-by-pitch data, suppose we are interested in learning about the pitching tendencies of the National League (NL) Cy Young winner Tim Lincecum during the 2009 season.

We begin by using the `subset` function to create a new data frame `tim` containing the pitch data for Lincecum for all his games in the 2009 season.

```
> tim=subset(pitchdata, pitchdata$pitcher=="lincecum")
```

What type of pitches does Lincecum throw? We answer this question by constructing a frequency table of the variable `pitch_type` by use of the R function `table`.

```
> table(tim$pitch_type)
```

```
      CH   CU   FA   FC   FF   FS   FT   IN   PO   SI   SL   UN
  32 639  621    0    0 1877    0   25    4    0    0  241    0
```

We see that Lincecum primarily throws 4-seam fastballs (FF), then changeups (CH) and curve balls (CU), and then sliders (SL) and a few 2-seam fastballs (FT).

What were the outcomes of Lincecum's pitches at the end of the plate appearance? Using the R `subset` and `table` functions, we look at the subset of the pitches that result in a batting event and construct a frequency table of the variable `brief_event`.

```
> with(subset(tim,tim$new.count.type=="PA event"),
+  table(brief_event))


brief_event
 Bunt out     C/INT    Double     Error   Fly out Groundout
       20         0        32         5       118       195
      HBP     Homer Int. Walk  Line out   Pop out   Rnr out
        5        10         2        31        33         0
   Single Strikeout    Triple      Walk
      123       251         3        66
```

We see that Lincecum had a large number of strikeouts (251), many fly outs (118) and groundouts (195), and only 10 pitches were hit for home runs.

Baseball pitches can be distinguished by how fast they are thrown and by their movement. In the dataset, the movement of pitches is described by the variables `pfx_x` and `pfx_z` that give, respectively, the movement in the horizontal and vertical directions. (This movement is viewed from behind home plate and a value of `pfx_x` < 0 corresponds to a pitch that moves horizontally towards a right-handed batter and away from a left-handed batter.) The dataset also records two speeds for each pitch—the pitch as it leaves the pitcher's hand (variable `start_speed`) and the speed as it crosses the plate (variable `end_speed`). In the following R code, we use the functions plot and points to display the horizontal and vertical movement for all of the 4-seam fastballs, curve balls, and changeups of Tim Lincecum. The resulting display is shown in Figure 4. The shading of the points corresponds to the speed of the pitch over the plate, where darker shades refer to pitches thrown at a higher speed. The message from the graph is that Lincecum's fastballs break up and toward right-handed batters, and his curve balls break down and away from right-handed batters. Changeups resemble fastballs in that they have similar movement up and toward right-handed batters, but changeups are thrown at a slower speed.

```
> palette(gray(seq(0,.9,len=25)))
> scale.speed=function(x)
+   25-25*(x-53)/(93-53)
> tim.FF=subset(tim,tim$pitch_type=="FF")
> tim.CH=subset(tim,tim$pitch_type=="CH")
> tim.CU=subset(tim,tim$pitch_type=="CU")
> with(tim.FF,plot(pfx_x,pfx_z,
```

```
+   xlim=c(-13,13),ylim=c(-15,18),
+   col=scale.speed(end_speed),pch=1,
+   xlab="Horizontal Break",ylab="Vertical Break"))
> with(tim.CH,points(pfx_x,pfx_z,
+   col=scale.speed(end_speed),pch=2))
> with(tim.CU,points(pfx_x,pfx_z,
+   col=scale.speed(end_speed),pch=3))
> text(c(-5,-12,-12),c(-7,2,15),
+   c("Curveball","Changeup","Fastball"))
```

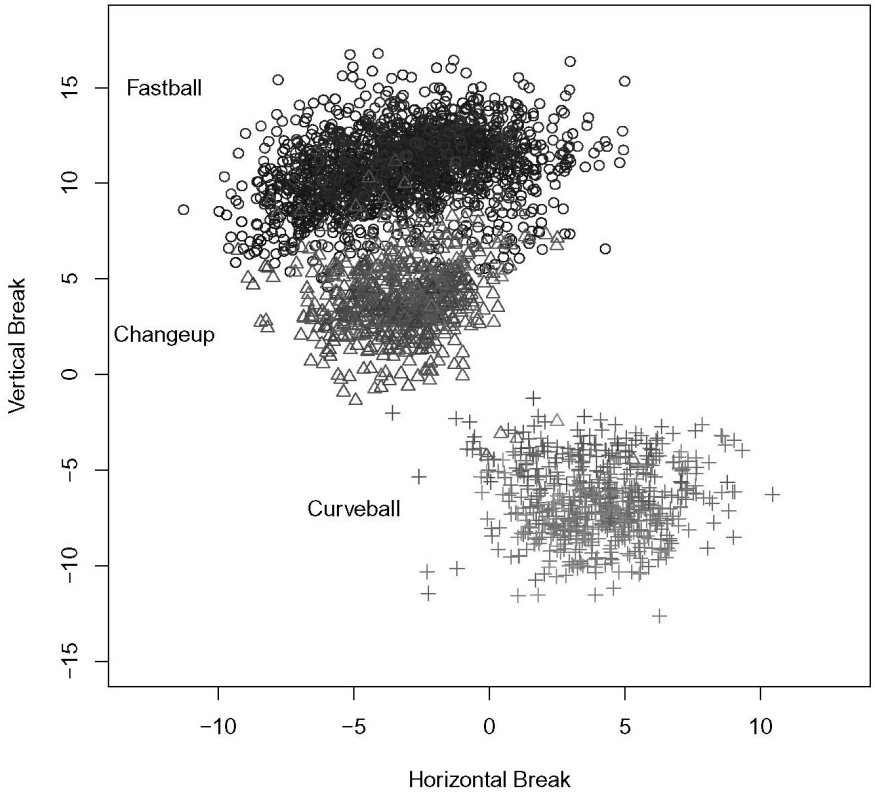


Figure 4. Movement of three types of pitches of Tim Lincecum for 2009 season where the speed of the pitch is indicated by the darkness of the plotting point where darker means faster. Four-seam fastballs, changeups and curve balls, corresponding to different plotting symbols, are clearly distinguished by movement and speed.

Figure 4 is helpful for understanding the movement of Lincecum's pitches, but is uninformative on the location of his pitches relative to the strike zone. A bivariate density estimate is an attractive way of summarizing the location of pitches. A short

R function `plot2D` is written to display a kernel density of these pitch locations; this function uses the function `bkde2D` in the `KernSmooth` R package. A box is drawn on the display corresponding to the location of the strike zone for an average hitter. (The width of the strike zone is approximately one foot on each side of the middle of the plate and the height is 1.67 feet to 3.5 feet for a batter of average height.)

```
> library(KernSmooth)
> plot2D=function(loc.pitches,...)
+ {
+ est=bkde2D(loc.pitches,bandwidth=c(0.3,0.3))
+ contour(est$x1, est$x2, est$fhat,
+   xlim=c(-2,2),ylim=c(0,5),
+   levels=seq(.05,.3,by=.05),
+   xlab="Horizontal Location",
+   ylab="Vertical Location",...)
+ lines(c(-1,1,1,-1,-1),
+                c(1.67,1.67,3.5,3.5,1.67),lwd=3)
+ }
```

A pitcher typically throws to different locations depending on the batting side of the hitters. The following code displays a density estimate of Lincecum's fastballs to right-handed hitters. (See Figure 5.)

```
> tim.FF.R=subset(tim.FF,tim.FF$stand=="R")
> plot2D(with(tim.FF.R,cbind(px,pz)))
```

Then, the following code constructs a density estimate of Lincecum's curveballs to right-handed hitters. (See Figure 6.)

```
> tim.CU.R=subset(tim.CU,tim.CU$stand=="R")
> plot2D(with(tim.CU.R,cbind(px,pz)))
```

Recall that the viewing perspective is behind home plate and a right-handed hitter would be located to the left of the strike zone. Figure 5 tells us that Lincecum generally throws his fastballs in the middle and high regions of the strike zone. If he throws outside the zone, it is likely the pitch will be high. In contrast, Figure 6 tells us that Lincecum's curve balls are generally thrown low in the strike zone. These graphs are consistent with the author's beliefs that effective fastballs are thrown high in the zone and effective curve balls are thrown low in the zone.

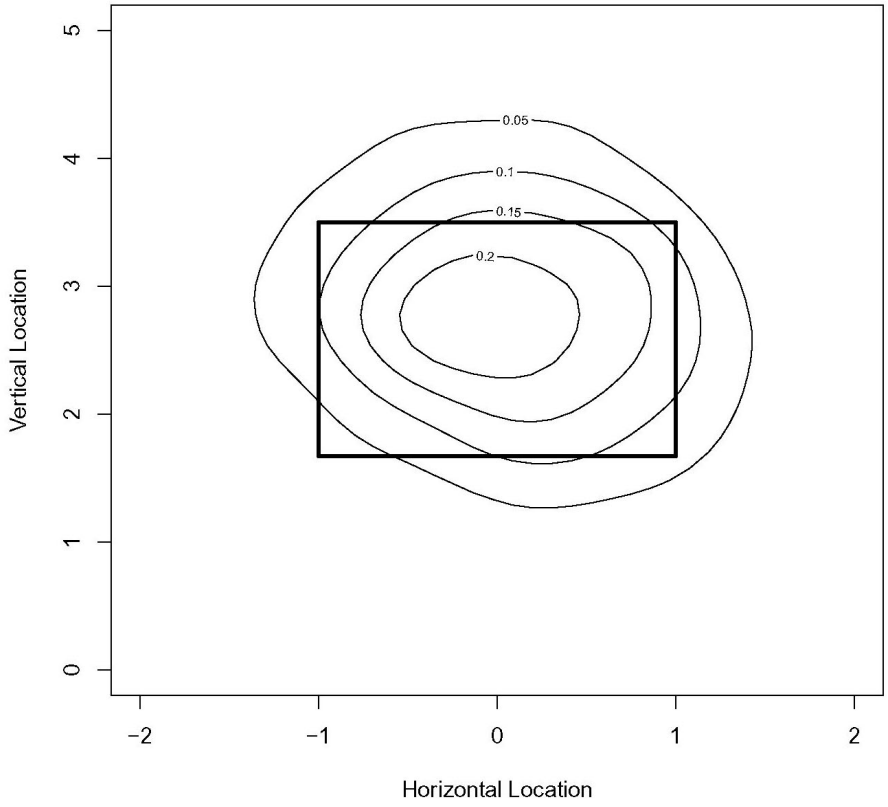


Figure 5. Density estimate of the location of Tim Lincecum's fastballs thrown to right-handed batters in the 2009 season. The rectangle corresponds to an "average" strike zone.

5.3 Possible Explorations

Using the pitch-by-pitch dataset, one can explore the pitching tendencies of the twenty pitchers.

- What pitches do these pitchers throw, what is the movement and speed of these pitches, and where are they thrown relative to the strike zone?
- Are particular pitches more successful in getting the batter to swing and miss?
- How do the pitchers differ with respect to pitch type and the speed that they throw the pitches?
- What pitching characteristics distinguish the nine "elite" pitchers from the remaining "non-elite" pitchers?

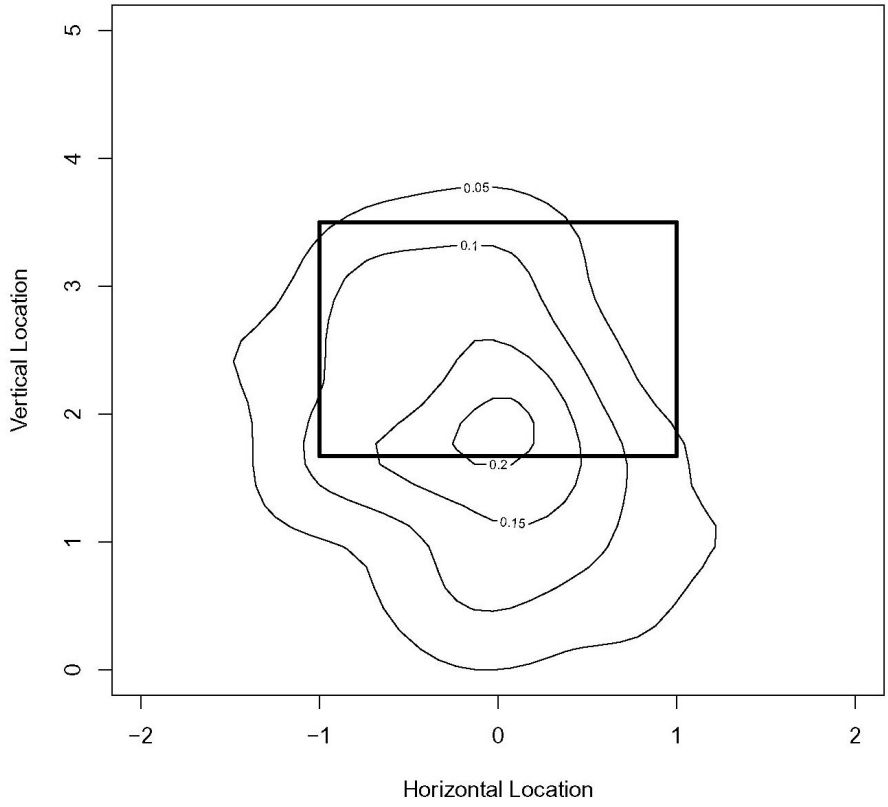


Figure 6. Density estimate of the location of Tim Lincecum's curveballs thrown to right-handed batters in the 2009 season. The rectangle corresponds to an "average" strike zone.

5.4 Literature and Use in Teaching

Since the PITCHf/x has only recently become available, there is a limited number of papers that have been published using this data. Albert (2009b) uses this data to model the probability that a swing of the batter results in a ball that is put into play, and uses this probability measure to compare pitchers. Albert (2010) focuses on the effect of the pitch count; for example, which pitchers are more likely to move from a 0-2 pitch count to an out? Brooks (2010) performs an exploratory analysis of the pitches thrown by any pitcher during any particular game and the Hardball Times (2010) and Baseball Prospectus (2010) regularly publish articles on the web that use the PITCHf/x to learn about pitcher tendencies.

6. Conclusions

Baseball is arguably the most sophisticated sport with regards to its use of data. While baseball provides a data-rich context for practicing statistical methods, the investigator can easily be overwhelmed by the sheer volume of data and tedium in extracting suitable datasets to address the questions of interest. The goal of this paper is to present three complete accessible datasets that describe baseball from three different perspectives—season to season, play-by-play, and pitch-by-pitch. There are many interesting questions about baseball that can be addressed and a variety of statistical concepts can be taught by use of these datasets. The hope is that the availability of these datasets will encourage statistics students to perform their own analyses, either as homework assignments or more extensive projects.

Appendix: Description of the Data Files

All of the text data files and documentation files are available on the JSE Data Archive, as well as the text file with the R script found in the paper. The R script is available at http://www.amstat.org/publications/jse/v18n3/Rscript.txt. In addition, these data files are available as R workspaces (with the Rdata extension). In addition, there is a R package `Baseball` that contains all three datasets, the associated documentation, and R demo scripts to implement the examples in the paper. The R workspaces and package `Baseball` can be found at the web site
http://bayes.bgsu.edu/baseball/threedatasets/

Lahman batting dataset

Description: Batting data for all players in professional baseball between 1871 and 2009.

Source: Sean Lahman's Baseball Archive at http://www.baseball1.com/

Data files: `mlb_batting.Rdata` (R workspace) and `mlb_batting.dat` (text, tab-delimited file)

Documentation files: `mlb_batting.txt`

To read the R workspace into R:

```
> load("mlb_batting.Rdata")
```

To read the text data file into R:

```
> batting = read.table("mlb_batting.dat", header = TRUE)
```

Retrosheet dataset

Description: play-by-play data for all plays in all games played in the 2008 season.

Source: The Retrosheet organization at http://www.retrosheet.org/

Data files: `playbyplay2008.Rdata` (R workspace) and `playbyplay2008.dat, roster2008.dat` (text, tab-delimited files)

Documentation files: `playbyplay2008.txt`

To read the R workspace into R:

```
> load("playbyplay2008.Rdata")
```

To read the text data files into R:

```
> pbp = read.table("playbyplay2008.dat", header = TRUE)
> roster = read.table("roster2008.dat",header = TRUE)
```

PITCHf/x database

Description: pitch-by-pitch data for twenty starting pitchers in the 2009 season

Source: MLB web pages http://gd2.mlb.com/components/game/mlb/year_2009/

Data files: `pitchfx.Rdata` (R workspace) and `pitchfx.dat` (text, tab-delimited files)

Documentation files: `pitchfx.txt`

To read the R workspace into R:

```
> load("pitchfx.Rdata")
```

To read the text data file into R:

```
> pitchfx = read.table("pitchfx.dat", header = TRUE)
```

References

Albert, Jim (1994), "Exploring baseball hitting data: What about those breakdown statistics?", *Journal of the American Statistical Association*, 89, 1066–1074.

Albert, Jim (1998), "The home-run hitting of Mike Schmidt," *Chance*, Vol. 11, No. 3, 3–11.

Albert, Jim (2002a) "Hitting with runners in scoring position," *Chance*, Vol. 15, No. 4, 8–16.

Albert, Jim (2002b), "A Baseball Statistics Course," *Journal of Statistics Education*, Vol. 10, No. 2. Available online at http://www.amstat.org/publications/jse/v10n2/albert.html

Albert, Jim (2003), *Teaching Statistics Using Baseball*, Mathematical Association of America.

Albert, Jim (2008), "Streaky Hitting in Baseball," *Journal of Quantitative Analysis in Sports*, 4.

Albert, Jim (2009a), "Is Roger Clemens' WHIP trajectory unusual?" *Chance*, Vol. 22, No. 2, 8–20.

Albert, Jim (2009b), "Exploring PITCHf/x data", *Proceedings of the 2nd International Conference on Mathematics in Sport*, The Institute of Mathematics and Its Applications.

Albert, Jim (2010), "Using the Pitch Count to Measure Pitching Performance," *Journal of Quantitative Analysis of Sports*, Vol. 6, Issue 4.

Albright, S. Christian (1993), "A statistical analysis of hitting streaks in baseball (Disc: p1184-1196)," *Journal of the American Statistical Association*, 88, 1175–1183.

Baseball Prospectus (2010), http://www.baseballprospectus.com/

Berry, Scott M. (1991), "The summer of '41: A probabilistic analysis of DiMaggio's "streak" and Williams's average of .406," *Chance*, Vol. 4, No. 4, 8–11.

Berry, Scott M. (1999), "Does "the zone" exist for home-run hitters?" *Chance*, Vol. 12, No. 1, 151–156.

Berry, Scott M. (2000a), "Modeling acceptance to the Major League Baseball Hall of Fame," *Chance*, Vol. 13, No. 1, 52–57.

Berry, Scott M. (2000b), "Modeling offensive ability in baseball." *Chance*, Vol. 13, No. 4, 56–59.

Bradlow, Eric T., Jensen, Shane T., Wolfers, Justin and Wyner, Abraham J. (2008), "A statistical look at Roger Clemens' pitching career," *Chance*, Vol. 21, No. 3, 24–30.

Brooks, Dan (2010), http://brooksbaseball.net/

Chance, Don M. (2009) , "What are the odds? – Another look at DiMaggio's streak," *Chance*, Vol. 22, No. 2, 33–42.

Gould, Stephen Jay (1989), "The streak of streaks," *Chance*, Vol. 2, No. 2, 10–16.

Hardball Times (2010), http://www.hardballtimes.com/

Hoffman, Tony (1989), "The search for the ultimate baseball statistic," *Chance*, Vol. 2, No. 3, 37–39.

Lahman, Sean (2010), *Sean Lahman's Baseball Archive*, http://www.baseball1.com.

Major League Baseball (2010), *Pitch F/X data files*, http://gd2.mlb.com/components/game/mlb.

Nathan, Alan (2009), "Tracking Baseball Pitches Using Video Technology: The PITCHf/x System," http://webusers.npl.illinois.edu/~a-nathan/pob/pitchtracker.html

Retrosheet (2010), http://www.retrosheet.org.

Rockoff, David and Yates, Philip (2009), "Chasing DiMaggio: Streaks in Simulated Seasons Using Non-Constant At-Bats," *Journal of Quantitative Analysis in Sports*, 5.

Schmotzer, Brian, Kilgo, Patrick D. and Switchenko, Jeff (2009), "The natural? – The effect of steroids on offensive performance in baseball," *Chance*, Vol. 22, No. 2, 21–32.

Wang, Steve C. (2007), "Teaching statistical thinking using the baseball hall of fame," *Chance*, Vol. 20, No. 2, 25–31.

Warrack, Giles (1995), "The great streak ," *Chance*, Vol. 8, No. 3, 41–43.

Jim Albert
Department of Math and Statistics
Bowling Green State University
Bowling Green, OH 43403-0001
E-mail: albert@bgnet.bgsu.edu
