AMS / MAA | TEXTBOOKS VOL 34

Teaching Statistics Using Baseball

Second Edition

Jim Albert

MAA PRESS
An Imprint of the AMERICAN MATHEMATICAL SOCIETY

10.1090/text/034

Teaching Statistics Using Baseball

Second Edition

Library of Congress Control Number: 2017931120

Print ISBN: 978-1-93951-216-1
Electronic ISBN: 978-1-61444-622-4

Printed in the United States of America

Current Printing (last digit):
10 9 8 7 6 5 4 3 2 1

Teaching Statistics Using Baseball

Second Edition

Jim Albert
Bowling Green State University

Published and Distributed by
The Mathematical Association of America

MAA TEXTBOOKS

Bridge to Abstract Mathematics, Ralph W. Oberste-Vorth, Aristides Mouzakitis, and Bonita A. Lawrence
Calculus Deconstructed: A Second Course in First-Year Calculus, Zbigniew H. Nitecki
Calculus for the Life Sciences: A Modeling Approach, James L. Cornette and Ralph A. Ackerman
Combinatorics: A Guided Tour, David R. Mazur
Combinatorics: A Problem Oriented Approach, Daniel A. Marcus
Common Sense Mathematics, Ethan D. Bolker and Maura B. Mast
Complex Numbers and Geometry, Liang-shin Hahn
A Course in Mathematical Modeling, Douglas Mooney and Randall Swift
Cryptological Mathematics, Robert Edward Lewand
Differential Geometry and its Applications, John Oprea
Distilling Ideas: An Introduction to Mathematical Thinking, Brian P. Katz and Michael Starbird
Elementary Cryptanalysis, Abraham Sinkov
Elementary Mathematical Models, Dan Kalman
An Episodic History of Mathematics: Mathematical Culture Through Problem Solving, Steven G. Krantz
Essentials of Mathematics, Margie Hale
Field Theory and its Classical Problems, Charles Hadlock
Fourier Series, Rajendra Bhatia
Game Theory and Strategy, Philip D. Straffin
Geometry Illuminated: An Illustrated Introduction to Euclidean and Hyperbolic Plane Geometry, Matthew Harvey
Geometry Revisited, H. S. M. Coxeter and S. L. Greitzer
Graph Theory: A Problem Oriented Approach, Daniel Marcus
An Invitation to Real Analysis, Luis F. Moreno
Knot Theory, Charles Livingston
Learning Modern Algebra: From Early Attempts to Prove Fermat's Last Theorem, Al Cuoco and Joseph J. Rotman
The Lebesgue Integral for Undergraduates, William Johnston
Lie Groups: A Problem-Oriented Introduction via Matrix Groups, Harriet Pollatsek
Mathematical Connections: A Companion for Teachers and Others, Al Cuoco
Mathematical Interest Theory, Second Edition, Leslie Jane Federer Vaaler and James W. Daniel
Mathematical Modeling in the Environment, Charles Hadlock
Mathematics for Business Decisions Part 1: Probability and Simulation (electronic textbook), Richard B. Thompson and Christopher G. Lamoureux
Mathematics for Business Decisions Part 2: Calculus and Optimization (electronic textbook), Richard B. Thompson and Christopher G. Lamoureux
Mathematics for Secondary School Teachers, Elizabeth G. Bremigan, Ralph J. Bremigan, and John D. Lorch
The Mathematics of Choice, Ivan Niven
The Mathematics of Games and Gambling, Edward Packel
Math Through the Ages, William Berlinghoff and Fernando Gouvea
Noncommutative Rings, I. N. Herstein

Non-Euclidean Geometry, H. S. M. Coxeter
Number Theory Through Inquiry, David C. Marshall, Edward Odell, and Michael Starbird
Ordinary Differential Equations: from Calculus to Dynamical Systems, V. W. Noonburg
A Primer of Real Functions, Ralph P. Boas
A Radical Approach to Lebesgue's Theory of Integration, David M. Bressoud
A Radical Approach to Real Analysis, 2nd edition, David M. Bressoud
Real Infinite Series, Daniel D. Bonar and Michael Khoury, Jr.
Teaching Statistics Using Baseball, 2nd edition, Jim Albert
Thinking Geometrically: A Survey of Geometries, Thomas Q. Sibley
Topology Now!, Robert Messer and Philip Straffin
Understanding our Quantitative World, Janet Andersen and Todd Swanson

MAA Service Center
P.O. Box 91112
Washington, DC 20090-1112
1-800-331-1MAA FAX: 1-240-396-5647

Contents

Preface to the First Edition

Over twenty years, this author has been in the enterprise of teaching introductory statistics to an audience that is taking the class to satisfy their mathematics requirement. This is a challenging endeavor because the students have little prior knowledge about the discipline of statistics and many of them are anxious about mathematics and computation. Statistical concepts and examples are usually presented in a particular context. However, one obstacle in teaching this introductory class is that we often describe the statistical concepts in a context, such as medicine, law, or agriculture that is completely foreign to the undergraduate student. The student has a much better chance of understanding concepts in probability and statistics if they are described in a familiar context.

Many students are familiar with sports either as a participant or a spectator. They know of the popular athletes, such as Tiger Woods and Barry Bonds, and they are generally knowledgeable with the rules of the major sports, such as baseball, football, and basketball. For many students sports is a familiar context in which an instructor can describe statistical thinking.

The goal of this book is to provide a collection of examples and exercises applying probability and statistics to the sport of baseball. Why baseball instead of other sports?

- Baseball is the great American game. Baseball is great in that it has a rich history of teams and players, and many people are familiar with the basic rules of the game. The popularity of baseball is reflected by the large number of movies that have been produced about baseball teams and players.
- Baseball is the most statistical of all sports. Hitters and pitchers are identified by their corresponding hitting and pitching statistics. For example, Babe Ruth is forever identified by the statistic 60, which was the number of home runs hit in his 1927 season. Bob Gibson is famous for his record low earned run average of 1.12 during the 1968 season. A flood of different statistical measures are used to rate players and salaries of players are determined in part by these statistics. There is an active effort among baseball writers to learn more about baseball issues by using statistics.
- A wealth of baseball data is currently available over the Internet. Player and team hitting and pitching statistics can be easily found. Comparisons between players of different eras can be made using a downloadable dataset that gives hitting and pitching data for all players who have ever played professional baseball.

This book is organized using the same basic organization structure presented in most introductory statistics texts. After an introductory chapter, there is a chapter on the analysis on a single batch of data, followed by chapters on the comparison of batches, and the analysis of relationships. There are chapters on introductory and more advanced topics in probability, followed by topics in statistical inference. Each chapter contains a number of essays or case studies that describe the analysis of statistical or probabilistic methods to particular baseball data sets. After the collection of case studies in each chapter, there is a set of activities and exercises that suggest further exploration of baseball datasets similar to the analysis presented in the case studies.

How can this book be used in teaching probability or statistics? We suggest several uses of this material.

- This book can be used as the framework for a one-semester introductory statistics class that is focused on baseball. Such a class has been taught at the author's home institution. This course covers the basic topics of a beginning statistics course (data analysis, introductory probability, and concepts of inference) using baseball as the primary source of applications. This course is suitable for students who are interested or curious about the game of baseball. It is also suitable for students with sports-related majors, such as sports management or sports medicine.
- This book can also be used as a resource for instructors who wish to infuse their present course in probability or statistics with applications from baseball. The material in this book has been presented at different levels to make it useable for introductory and more advanced courses. The case studies can be used by the instructor to present the particular topic within a baseball context and then the associated exercises and activities can be used for homework. The case studies can serve as useful springboards for undergraduate students who wish to do additional explorations on baseball data.

Acknowledgements

I am appreciative of the support given to this project by the Division of Undergraduate Education of the National Science Foundation and by my colleagues in the Department of Mathematics and Statistics at Bowling Green State University. The text was used for a number of experimental sections of MATH 115 Introduction to Statistics at Bowling Green State University and I am grateful for the valuable feedback from the students who enrolled in this course. In addition, Chris Andrews, Jay Bennett, Eric Bradlow, Jim Cochran, Joe Gallian, Carl Morris, Jerome Reiter, Ken Ross, Steve Samuels, Bob Wardrop, and Dex Whittinghill provided many helpful suggestions in reviewing the book. I thank the editors of the MAA publications for their support, particularly Zaven Karien and Dave Kullman. Last, but certainly not least, I thank my wife Anne, and children Lynne, Bethany, and Steven for their understanding and great patience during the completion of this text.

Jim Albert
December 2002

Preface to the Second Edition

The author has been gratified with the reception to the first edition of this text. It has been enjoyable teaching introductory statistics from a baseball viewpoint, as one can discuss statistical concepts in the context of current and historical players and teams.

Sports provide a wonderful context for teaching introductory statistics and other texts have been recently published using a sports theme. Rothman (2012) is a comprehensive statistics text with a baseball theme, and Tabor and Franklin (2011) is a statistics text with examples from a wide range of sports. There is a weath of publically available baseball data and Marchi and Albert (2013) describe the use of these datasets together with the open-source statistics software system R (R Core Team (2015)) to implement a variety of baseball studies.

Since the date of the first edition, there have been changes both in the players who play baseball, and also in the development and use of analytics. So this motivates revisions to the text reflecting these changes.

In this edition, many of the case studies and exercises have been revised to use data from current teams and players. I encourage the instructor to always use current season data since the modern players and teams are most familiar to students.

Also exercises have been added to the chapters reflecting some of the newer types of baseball data. The pitchFX system has been tracking the trajectories of baseball pitches since 2006 and we know more about the breaks, locations, and speeds of pitches. Websites such as `fangraphs.com` contain much of this pitchFX data for both batters and pitchers and this website also gives information about the location, speed, and type (flyball, pop-up, and grounders) of balls put in play. The tracking technology Statcast contains information about the speed of runners and fielders and velocities of balls coming off a bat.

To facilitate the use of the data described in the examples and exercises, all of the datasets are currently available using the StatCrunch statistical software system published by Pearson (in StatCrunch search using the acronym TSUB and the example or exercise number). In addition, the datasets are available as text files at https://bayesball.github.io/.

Jim Albert
December 2016

10.1090/text/034/01

1

An Introduction to Baseball Statistics

Leading Off

The baseball game has just started. The umpire has yelled "Play Ball!", and the batter at the top of the order is coming to bat. He's the leadoff hitter and his job is to help produce runs by getting on base. Who was the greatest leadoff hitter of all time? Most people believe that the best leadoff man was Rickey Henderson. Bill James, a leading baseball statistician, says in *The New Bill James Historical Baseball Abstract* that Henderson was

- the greatest base stealer of all time,
- the greatest power/speed combination of all time (except maybe Barry Bonds),
- the greatest leadoff man of all time,
- one of the top five players of all time in runs scored.

Moreover, James says: "You could find fifty Hall of Famers who, all taken together, don't own as many records, and as many important records, as Rickey Henderson."

Here's some biographical information about Rickey Henderson. He was born on Christmas Day, 1958, in Chicago, one of seven children. His family moved to Oakland, California when he was young and Henderson played baseball and football at Oakland Tech High School. When he graduated, he received many football scholarships and also was selected by the Oakland A's in the fourth round of the 1976 baseball draft. Although Henderson preferred football, his mother wanted him to play baseball, and Henderson agreed to go along with his mother's wishes. After a couple of years in the A's minor league organization, he was promoted to the Oakland major league team on June 23, 1979,[1] and immediately was a starter on the team. He has been a dominant player in the major leagues his entire career and was voted on the All-Star Team for the years 1980, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1990, and 1991.

How can we demonstrate that Rickey Henderson was indeed the best leadoff man in baseball? We look at his stats. Table 1.1 displays the season-to-season hitting statistics for Henderson for his 25 seasons in Major League Baseball.

[1] I have a kinship with Rickey Henderson since we both started professionally (me as a statistician and Henderson as a ballplayer) in 1979. I lasted longer than Henderson in the professional ranks.

Table 1.1. Batting statistics for Rickey Henderson's career

Year	Team	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO
1979	OAK	89	351	49	96	13	3	1	26	33	11	34	39
1980	OAK	158	591	111	179	22	4	9	53	100	26	117	54
1981	OAK	108	423	89	135	18	7	6	35	56	22	64	68
1982	OAK	149	536	119	143	24	4	10	51	130	42	116	94
1983	OAK	145	513	105	150	25	7	9	48	108	19	103	80
1984	OAK	142	502	113	147	27	4	16	58	66	18	86	81
1985	NYY	143	547	146	172	28	5	24	72	80	10	99	65
1986	NYY	153	608	130	160	31	5	28	74	87	18	89	81
1987	NYY	95	358	78	104	17	3	17	37	41	8	80	52
1988	NYY	140	554	118	169	30	2	6	50	93	13	82	54
1989	TOT	150	541	113	148	26	3	12	57	77	14	126	68
1990	OAK	136	489	119	159	33	3	28	61	65	10	97	60
1991	OAK	134	470	105	126	17	1	18	57	58	18	98	73
1992	OAK	117	396	77	112	18	3	15	46	48	11	95	56
1993	TOT	134	481	114	139	22	2	21	59	53	8	120	65
1994	OAK	87	296	66	77	13	0	6	20	22	7	72	45
1995	OAK	112	407	67	122	31	1	9	54	32	10	72	66
1996	SDP	148	465	110	112	17	2	9	29	37	15	125	90
1997	TOT	120	403	84	100	14	0	8	34	45	8	97	85
1998	OAK	152	542	101	128	16	1	14	57	66	13	118	114
1999	NYM	121	438	89	138	30	0	12	42	37	14	82	82
2000	TOT	123	420	75	98	14	2	4	32	36	11	88	75
2001	SDP	123	379	70	86	17	3	8	42	25	7	81	84
2002	BOS	72	179	40	40	6	1	5	16	8	2	38	47
2003	LAD	30	72	7	15	1	0	2	5	3	0	11	16

The stats that most people talk about are Henderson's career totals.

- He stole the most bases (1395) of any baseball player in history.
- He scored the most runs (2248) of any player in history.
- He received the most walks (2141) of any player in history.

Those are great achievements and the statistics are a measure of these achievements. But the goal of this book is to look deeper at baseball statistics.

There is a general confusion about the meaning of statistics. If you look at the American Heritage Dictionary of the English Language, you will find two very different definitions of statistics.

1. *Statistics* is a collection of numerical data.
2. *Statistics* is the mathematics of the collection, organization, and interpretation of numerical data.

Let's relate these two definitions to baseball. First, baseball statistics are the counts and measures that we use to evaluate players and teams—this refers to the first definition of statistics.

When the announcer on a television broadcast of a baseball game thanks the statistician, he or she is referring to the person who collects and gives baseball data to the announcers. Here the focus is on the data.

But this book concentrates on the second definition of statistics: how can we interpret or make sense of baseball stats? A professional statistician (to be distinguished from the person who is collecting the data) is concerned with how we can use data to learn about some underlying truth. In doing this, he or she has to think about several issues.

- How should the data be collected to make it useful in drawing conclusions?
- Once the data is collected, how do we organize and summarize it to learn about its general features?
- Last, how can we use the data to make our conclusions? (It turns out that probability or chance plays an important role in decision-making.)

It might be helpful to distinguish the two meanings by capitalization—in this section I will call numerical data "statistics," and the science of learning from data "Statistics."

The goal of this book is to introduce Statistical thinking and Statistical methods in the context of baseball. We introduce the chapters of this book by looking at the statistics of Rickey Henderson. Some questions will be raised in the following discussion and we'll continue our Statistical look at Rickey Henderson by "leadoff exercises" in each chapter.

Exploring a Single Batch of Baseball Data (Chapter 2)

The goal of a leadoff hitter is to get on base. The obvious measure of a player's ability to get on base is the on-base percentage (OBP), which is simply the fraction of plate appearances where the player gets on base. (A precise definition of OBP will be given later.) Here are Henderson's season OBPs for all 25 seasons in the majors:

0.338	0.420	0.408	0.398	0.414	0.399	0.419	0.358	0.423
0.394	0.411	0.439	0.400	0.426	0.432	0.411	0.407	0.410
0.400	0.376	0.423	0.368	0.366	0.369	0.321		

If we scan these numbers, we see variation—one season he had an OBP of .432 and another season his OBP was .366. A Statistician will try to make sense of these data by constructing an appropriate graph. Figure 1.1 shows a dotplot of the OBPs.

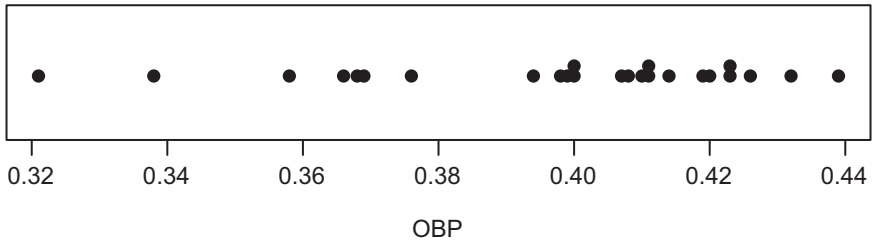


Figure 1.1. Dotplot of season on-base percentages for Rickey Henderson

We see from this graph that most of Henderson's season OBPs are between .400 and .420. There is a small cluster of values in the .340 to .380 range. Why? Was Henderson hurt these

particular seasons? Did these low OBPs correspond to the early or late periods of his career? (We'll answer these questions later.)

After we graph the data, we try to find suitable numbers to summarize the main features of the distribution of OBPs. Looking at the graph, it seems that .410 might be a representative season OBP for Henderson.

Comparing Batches and Standardization (Chapter 3)

We now have some handle on Henderson's on-base ability: he generally got on base about 40% of the time. But is an on-base percentage of .400 any good? How does his on-base performance compare with other players?

Another good leadoff hitter, a contemporary of Henderson, was Tim Raines. How effective was Tim in getting on base and how did he compare with Henderson?

Here are Tim's on-base percentages for the 1981–1998 seasons where he had at least 200 at-bats.

0.391	0.353	0.393	0.393	0.405	0.413	0.429	0.350	0.395
0.379	0.359	0.380	0.401	0.365	0.374	0.383	0.403	0.395

How does this batch of OBPs compare to the batch of OBPs for Henderson? In Chapter 3, we'll talk about ways of comparing two batches of data. We will discuss methods for graphing the two datasets and ways of stating Statistically (in this example) that Henderson is better at getting on base than Tim. This chapter will also show how we can judge the greatness of Henderson's on-base performance in the context of all players that particular year.

Relationships between Measurement Variables (Chapter 4)

When we look at Rickey Henderson's hitting statistics in Table 1.1, we notice that many of the statistics are related. For example, if a batter gets many doubles and triples, he will have a high slugging percentage, and a batter who rarely walks is likely to have a small on-base percentage. In Chapter 4, we discuss ways of looking at relationships between variables. To illustrate, consider the relationship between Henderson's count of doubles and his count of home runs for individual seasons. Doubles and home runs go hand-in-hand—one might think that if Henderson is hitting a lot of deep fly balls one season then he would have many doubles and home runs. We can graphically view the relationship between doubles and home runs by the scatterplot shown in Figure 1.2. There is a pattern in the scatterplot as the points drift from the lower left to the upper right sections—in seasons where Henderson hit many doubles, he tended also to hit many home runs.

In Chapter 4, we discuss ways of measuring the pattern of association in the scatterplot, and discuss how we can use a line to describe the relationship. The methods are helpful for finding a good measure of the batting performance of a player.

Introduction to Probability Using Tabletop Games (Chapter 5)

Fans love to play baseball games; currently millions of people are playing fantasy and simulation baseball. Before there were personal computers, Nintendo and fantasy baseball, there were a number of tabletop baseball games that were very popular among fans. In Chapter 5, we introduce the notion of a probability model by examining several tabletop baseball games. One

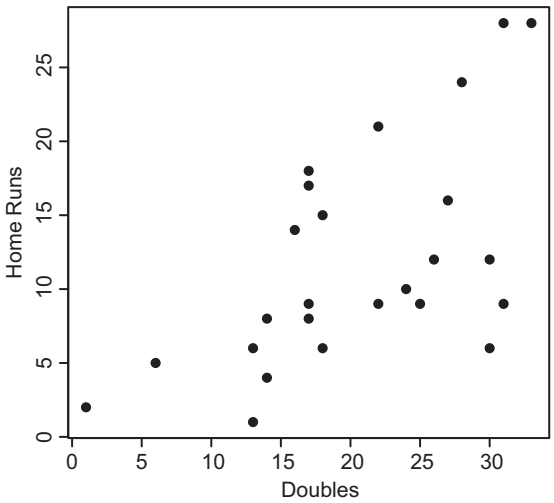


Figure 1.2. Scatterplot of count of doubles and count of home runs for all seasons of Rickey Henderson's career.

of the first games played by the author as a child was *All-Star Baseball* where the performance of a batter was represented by a random spinner with areas of the spinner corresponding to the different outcomes of a plate appearance. A Rickey Henderson spinner is shown in Figure 1.3 where the areas of the regions are computed using his batting statistics from the 1990 season. Note the large pie slice corresponding to a walk—this is a visual demonstration of Henderson's ability to draw walks this particular season.

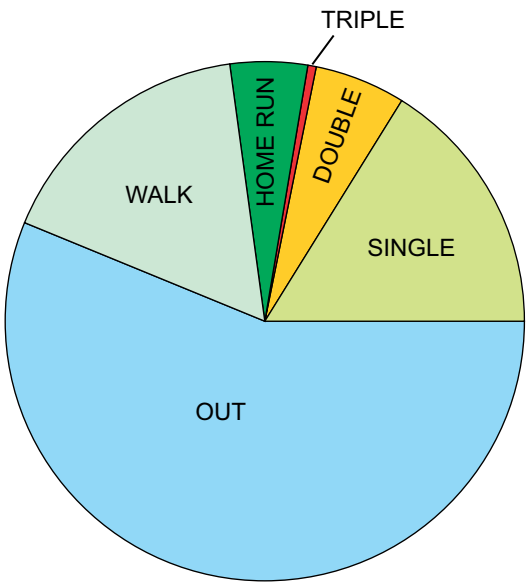


Figure 1.3. A Rickey Henderson spinner using his 1990 hitting statistics.

Probability Distributions and Baseball (Chapter 6)

Baseball has a very nice discrete structure that makes for convenient probability modeling. The basic event in baseball is the outcome of a plate appearance. Rickey Henderson comes to bat—either he will get on base or he won't. If you assume that Henderson has a probability, say .4, of getting on base for each plate appearance, and moreover, the chance of getting on base does not depend on how he hit in earlier games, then one can make reasonable predictions about the number of times he will get on base in a game, a week, or a month. In Chapter 6, we show how some popular probability distributions (such as the binomial and negative binomial) can be used to explain the random process of players getting on base, and then scoring runs.

Introduction to Statistical Inference (Chapter 7)

Rickey Henderson had a great ability to get on base. What does this mean? Well, to be an effective leadoff hitter, Henderson had to be very able at making the bat hit the ball. (Such a player is called a good contact hitter.) Also, he knew how to "work the count." This means that he would not swing at many bad pitches and would patiently wait for a good pitch to be thrown.

How do we know that Henderson had great leadoff hitter ability? We looked at his on-base percentages over his career and saw a pattern of high numbers. In other words, Henderson exhibited a high level of performance over many years, and there was no doubt that his performance reflected great skill. It was very unlikely that he had average ability to get on base and, by luck or chance variation, he happened to get high on-base percentages for all of those years.

In Chapter 7, we look at the connection between a player's hitting ability and the performance of the hitter over a season. Actually, if a player has an on-base percentage of OBP $= .400$ for only a single season, we really don't know if the player has a great ability to get on base. We need to see a pattern of great hitting performance over many seasons to properly gauge the player's hitting ability.

Topics in Statistical Inference (Chapter 8)

Making judgments about a player's hitting or pitching ability is relatively easy when you observe the player's performance over a 10–20 year career. But there are other aspects of ability that are much harder to detect. To illustrate, consider the following situational statistics for Henderson for the 1999 season. This season, Henderson's on-base percentage was .423. However, we learn that

- his OBP was .376 for home games and .462 for away games,
- his OBP was .490 for games played on artificial turf and .408 for games played on natural grass,
- his OBP was .472 for games played in a domed ballpark and .417 in open ballparks,
- his OBP was .509 when the pitch count reached 3-2, and .338 when the pitch count reached 1-2.

How do we make sense of these situational hitting stats? Is Henderson really better at getting on base when he is playing at home? Does Henderson really have better on-base ability when the game is played on turf? Is it meaningful that his on-base percentage is over .500 when the pitch count gets to 3 balls and 2 strikes? Does Henderson really have an ability to perform extra well in particular situations?

One main topic of Chapter 8 is to try to interpret the significance of situational hitting data. We will see that it is difficult to detect situational hitting ability and much of the variation in situational hitting statistics is attributable to chance or random variation.

Modeling Baseball Using a Markov Chain (Chapter 9)

We say that Rickey Henderson is the greatest leadoff hitter of all time, but we haven't talked about how well Henderson performed when he was the leadoff hitter in an inning. From the statistics in Table 1.1, we can compute Henderson had an OBP of .439 in the 1990 season; the OBP tells us the fraction of plate appearances that Henderson got on base for all situations. What was his on-base percentage in the innings when he was leading off?

Fortunately, this type of data is now readily available. Let's focus on the home games in Oakland in 1990. In these games, Henderson had 273 plate appearances, but only 108 of them occurred at the beginning of an inning. So Henderson only led off 108 times. How did he do in these lead-off opportunities? Table 1.2 shows the batting results. Note that he was out 63% of the time, so his on-base percentage when he actually was leading off the inning was .370. This on-base percentage appears to be a bit low, but one has be cautious about drawing a strong conclusion since this table summarizes the results of only 108 plate appearances.

Table 1.2. Batting results for Rickey Henderson's leadoff plate appearances at home during the 1990 season

Play	Count	Percentage
Out	68	63%
Walk	18	16.7%
Single	13	12.0%
Double	5	4.6%
Triple	1	0.9%
Home run	3	2.8%

We can also see how Henderson performed when the bases were empty with one out, or when the bases were loaded with two outs. We will use data such as this in Chapter 8 to construct a sophisticated probability model for the sequence of batting events in baseball. This model will be very useful for measuring the values of different types of hits such as a home run and for evaluating the worth of different baseball strategies such as a sacrifice bunt.

Some Basic Measures of Baseball Performance

The effectiveness of batters and pitchers is typically assessed by particular numerical measures. Here we define some basic measures for evaluating hitters and pitchers.

Measures for Batters

The classical measure of hitting effectiveness for a player is the batting average (AVG) that is computed by dividing the number of hits (H) by the number of at-bats (AB):

$$\text{AVG} = \frac{\text{H}}{\text{AB}}.$$

This statistic gives the proportion of time that a batter gets a hit among all at-bats. The batter with the highest batting average during a baseball season is called the batting champion that year. Batters are also evaluated on their ability to get singles (1B), doubles (2B), triples (3B), and home runs (HR). The slugging percentage (SLG) is an average of a player's "total bases" (TB) where the hits are weighted by means of the number of bases reached:

$$\text{SLG} = \frac{1 \times 1\text{B} + 2 \times 2\text{B} + 3 \times 3\text{B} + 4 \times \text{HR}}{\text{AB}}.$$

This measure reflects the ability of a batter to hit the ball a long distance. A third measure of hitting ability is the on-base percentage (OBP), which is defined as the proportion of plate appearances where the player gets on base.

$$\text{OBP} = \frac{\text{H} + \text{BB} + \text{HBP}}{\text{AB} + \text{BB} + \text{HBP} + \text{SF}}$$

In this formula, BB is the count of walks, HBP is the number of times the batter was hit by a pitch, and SF is the number of sacrifice flies.

Measures for Pitchers

A number of statistics are also used in the evaluation of pitchers. For a particular pitcher, one counts the number of games in which he was declared the winner (W) or loser (L) and the number of runs allowed. Pitchers are usually rated by means of the earned run average (ERA) the average number of earned runs (ER) allowed for a nine-inning game:

$$\text{ERA} = 9 \times \frac{\text{ER}}{\text{IP}}.$$

(In this formula, IP is the number of innings pitched.) Other statistics are useful in understanding pitching ability. A pitcher who can throw the ball very fast (such as Randy Johnson) can record a high number of strikeouts (SO). A pitcher who is "wild" or relatively inaccurate in his pitching will record a large number of walks (BB).

Better Measures of Hitting Ability

There is lively research to better interpret baseball statistics. Sabermetrics is the mathematical and statistical analysis of baseball records. (The term sabermetrics was first used by Bill James in honor of SABR, the Society of Baseball Research.) One interest of sabermetricians (the people who analyze baseball statistics) is to find good measures of hitting and pitching performance. Bill James compares in his *1982 Baseball Abstract* the batting records of two players, Johnny Pesky, who played from 1942 to 1954, and Dick Stuart, who played in the 1960s. Pesky was a batter who hit for a high batting average but hit few home runs. Stuart, in contrast, had a modest batting average, but hit a high number of home runs. Who was the more valuable hitter? James argues that a hitter should be evaluated by his ability to create runs for his team. From an empirical study of a large collection of team hitting data, he established the following runs created (RC) formula for predicting the number of runs scored in a season based on the number of hits, walks, at-bats, and total bases recorded in a season:

$$\text{RC} = \frac{(\text{H} + \text{BB}) \times \text{TB}}{\text{AB} + \text{BB}}.$$

This formula reflects two important aspects in scoring runs in baseball. The count of a team's hits and walks reflects the team's ability to get runners on base. The count of a team's total bases reflects the team's ability to move runners that are already on base. This runs created formula can be used at an individual level to compute the number of runs that a player creates for his team. In 1942, Johnny Pesky had 620 at-bats, 205 hits, 42 walks, and 258 total bases; according to the formula, he created 96 runs for his team. Dick Stuart in 1961 had 532 at-bats with 160 hits, 34 walks, and 309 total bases for 106 runs created. The conclusion is that Stuart in 1961 was a slightly better hitter than Pesky in 1942 since he created a few more runs for his team.

There are a number of alternative measures that have been proposed to evaluate hitters. Here we list some of the measures and they will be carefully compared and evaluated in later chapters. Since OBP is a measure of a player's ability to get on base and SLG is a measure of a player's ability to advance the runners, a simple measure OPS (for On-base percentage Plus Slugging percentage), adds the two measures:

$$\text{OPS} = \text{OBP} + \text{SLG}.$$

In the chapters to follow, we will investigate the goodness of the measures RC and OPS in predicting the number of runs scored by a team.

Baseball Data

Here we describe several common types of baseball data that we will analyze in this book.

Career statistics for a player

Probably the most familiar data set among baseball fans is the batting or pitching statistics of a player over the seasons of his career. Many fans have collected baseball cards, either as children or adults, and the back of these cards typically contain these career statistics. For each season, the card gives statistics similar to what is displayed for Rickey Henderson in Table 1.1.

Player statistics for a given season

Another informative data set is the collection of all batting and pitcher statistics for all players in a particular season. We will see that it is difficult to judge a single statistic, say Roger Maris' 61 home run season in 1961, by itself. (Roger Maris is famous since he set the single season record for home runs that particular year.) To understand the significance of this statistic, we need to look at it in the context of all player statistics for that season. Although the basic rules of baseball have not changed over the years, the athletic abilities of the players, the competitive balance, and the equipment and environment have changed, and these changes have had a substantive impact on the values of baseball statistics. To return to our Johnny Pesky and Dick Stuart comparison, it really is not fair to compare the runs created by Pesky and Stuart at face value, since there were many more runs scored in the 1961 season than the 1942 season.

Team statistics for a given season

Another interesting data set to work with is the batting and pitching statistics for all teams in a particular season. One general problem of interest is to find a suitable measure of the hitting ability of a player. Since the goal of batting is to score runs, one is interested in finding a hitting statistic that is useful in predicting the number of runs scored. But teams, not individual players, score runs, so one needs to look at team data in the development of good hitting measures.

Game logs

Baseball fans are fascinated with the day-to-day performance of their favorite players and teams. Teams and individual players go through good and bad periods and these performances in short time periods are often described in the media. So it is interesting to look at the performance of players and teams for each game of the baseball season.

Situational statistics

Baseball fans are also fascinated with the performances of teams and players in given situations. How does a batter perform at home and away games? How does he perform against different pitchers? How does he do at night and day games? How well does he hit when he swings at the first pitch, or when there are two strikes in the count? How well does a team perform when their ace pitcher is starting? These situational or breakdown statistics are now reported on all of the popular baseball news websites. One goal of this book is to try to make sense of the importance of these statistics.

Statistics through the years

Baseball has a fascinating history. It is fun to read about the great players of the past, including Ty Cobb, Walter Johnson, Shoeless Joe Jackson, Joe DiMaggio, and Babe Ruth. Also it is interesting to look at the great teams of all time, including the 1927 New York Yankees, the 1929 Philadelphia Athletics, the 1975 Cincinnati Reds, and the 1998 New York Yankees. One can explore this baseball history by means of baseball statistics. For example, we will look at home runs hit by teams in the years 1927, 1961, 1998, and 2001, and the differences that we see in this comparison tell us a lot about the relative difficulty of hitting a home run over time.

Miscellaneous statistics

Although we will focus much of our discussion on basic hitting and pitching statistics, there are many other associated baseball statistics that are fun to explore. These include:

- The salaries of the players.
- The attendance counts at the games.
- Statistics related to managerial strategy.
- The duration of the game.
- The number of pitches thrown.
- The time needed to complete the game.

Collecting baseball data

The collection of baseball data is much easier today than in the past due to the internet. There are many web sites that contain historical and current baseball data. Here we highlight several internet sources that are particularly convenient for downloading data that can be easily entered into a statistical computing package.

`www.baseball-reference.com`

This is a good site for historical data on teams and players. One can easily obtain career statistics for any historical player and team data is available for any past season. The data sets are stored as text files and it is relatively easy to import the data into a standard spreadsheet program such as Microsoft Excel.

`www.baseball1.com`

The Baseball Archive is an especially good site for downloading large collections of baseball data. In fact, one can download (either in text or Microsoft Access format) a single file that contains player and team statistics for all years in baseball history. Many of the data sets used in this book are taken from this database.

`www.retrosheet.org`

The Retrosheet organization is dedicated to collecting play-by-play baseball data. For each plate appearance in a given game, a data file will record the name of the hitter, the name of the pitcher, the game situation (score, runners on base, number of outs), the play, and other information. One can currently download this type of data for entire baseball seasons. Using these data, one can perform many interesting analyses. This type of data will be used in the Markov Chain modeling described in Chapter 9.

`www.fangraphs.com`

Fangraphs is a good place to visit to collect some of the more modern types of baseball data. For example, this site includes summaries of data collected from the PITCHf/x system, such as the number of pitches thrown of different types such as fastballs, curveballs, sliders, and changeups, and the velocities of these pitches. For batters, Fangraphs gives the percentages of batted balls that are line drives, popups and flyballs, and also the percentages of batted balls hit to the left, center, and right. This modern baseball data will be explored in some of the exercises.

10.1090/text/034/02

2

Exploring a Single Batch of Baseball Data

What's On-Deck?

In this chapter, we illustrate a number of graphs and summary statistics useful in exploring a single batch of data. In Case Study 2.1, we begin with batting statistics for the 30 Major League Teams for the 2014 season. We focus on the team home run numbers and use stemplots and five-number summaries to compare the home run production of the National League and American League teams. In Case Studies 2.2 and 2.3, we look at the career statistics for two current or future Hall of Famers, Derek Jeter and Randy Johnson. When looking at an individual's statistic, it is helpful to construct a graph of the statistic against time. The patterns in this time series plot are helpful for understanding how the player has matured as a baseball player. In Case Study 2.4, we study baseball attendance for the 30 teams in 2014. Although baseball teams would all like to make a profit, we will see a wide disparity in the teams' abilities to bring fans to the ballpark. We conclude in Case Study 2.5 by looking at statistics for managers. One basic play in baseball is the sacrifice bunt, and we will see that this is a popular strategy for some managers and a very unpopular move for other managers.

2.1 Looking at Teams' Offensive Statistics

Topics Covered: Stemplot, data distribution, five-number summary.

After the conclusion of a baseball season in November, teams begin to evaluate how well they performed during the season. How effective was a particular team, say the Phillies, in getting batters on base? Did the Phillies score a lot of runs this year? What teams were good and bad at hitting home runs? How many home runs were hit by a representative major league team this year?

Table 2.1 displays a number of offensive statistics for six major league teams for the 2014 season. Many of these statistics are counts, such as the number of hits, the number of doubles, the number of home runs, and so on. Other offensive statistics are derived measures of offensive performance, such as batting average (AVG), slugging percentage (SLG), and on-base percentage (OBP), which are computed from the count statistics.

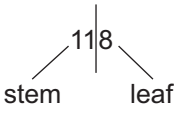
Home Run Totals

To explore the power hitting among the teams, let's look at the number of home runs hit by the 30 teams. The first step in exploring a single batch of data, such as these 30 home run totals,

Table 2.1. Batting statistics for six Major League Baseball teams in 2014

Team	AVG	OBP	SLG	R.G	H	2B	3B	HR	BB	SO
Arizona	.248	.302	.376	3.80	1379	259	47	118	398	1165
Atlanta	.241	.305	.360	3.54	1316	240	22	123	472	1369
Baltimore	.256	.311	.422	4.35	1434	264	16	211	401	1285
Boston	.244	.316	.369	3.91	1355	282	20	123	535	1337
Chicago (NL)	.239	.300	.385	3.79	1315	270	31	157	442	1477
Chicago (AL)	.253	.310	.398	4.07	1400	279	32	155	417	1362

is to draw a suitable graph. An effective graph that is easy to draw by hand is the stemplot. To construct a stemplot, we divide each home run total into two parts, called the stem and the leaf. For example, Arizona's home run total, 118, can be divided between the tens and units places (see below)



to get a stem of 11 and a leaf of 8. We write down all of the possible stems, and record each home run total by writing down its leaf on the line corresponding to the stem. If we do this for all 30 home run totals, we get the stemplot shown in Figure 2.1.

```
 9 | 5
10 | 59
11 | 178
12 | 233558
13 | 1246
14 | 267
15 | 0255567
16 | 3
17 | 7
18 | 6
19 |
20 |
21 | 1
```

Figure 2.1. Stemplot of team home run numbers from 2014 season.

The second line tells us that the totals 105, 109 were the number of home runs hit by two of the 30 teams. To study this distribution of home run totals, it may be helpful to flip this stemplot by a 90-degree turn, so that the small totals are on the left.

```
 9 | 5
10 | 59
11 | 178
12 | 233558
13 | 1246
14 | 267
15 | 0255567
16 | 3
17 | 7
18 | 6
19 |
20 |
21 | 1
```

What observations can we make from this data distribution?

1. First, we look for the general shape of these home run totals. Although there is a general mound shape, we see two clusters in these home run counts, one in the 120's, and a second in the 150's.
2. After we think of the general shape, we look for an "average" home run total. Since exactly half (15) of the totals fall below 135 and half fall above 135, we can regard 135 home runs as a measure of the center of the distribution. (We call 135 the median of the observations.)
3. Next, we look at the spread or variation in these home run totals. Here the spread of the totals is pretty large, the largest value 211 (number of home runs hit by Baltimore) is more than twice as large as the smallest value 95 (hit by Kansas City). But most of the home run totals fall between 110 and 157.
4. Last, we look for any unusual characteristics of the totals. There is clearly one large number that is separated from the rest—Baltimore hit a lot of home runs in 2014.

If you are a baseball fan, you are probably interested in the relative standing of your team in this distribution of home run totals. To better see the teams' relative standing, we can redraw this stemplot in Figure 2.2 using team labels instead of numerical leaves.

```
 9 | KCR
10 | STL SDP
11 | TEX TBR ARI
12 | MIA ATL BOS NYM PHI MIN
13 | CIN SFG LAD SEA
14 | CLE OAK NYY
15 | MIL WAS CHW DET LAA PIT CHC
16 | HOU
17 | TOR
18 | COL
19 |
20 |
21 | BAL
```

Figure 2.2. Stemplot of team home run numbers from 2014 season with teams identified.

My team, the Phillies, appears in the cluster of lower numbers. It is interesting to note that home run production is not always associated with winning and losing. Colorado, the team with the second highest home run total, finished next to last place in their division, whereas Kansas City, a team who appeared in the 2014 World Series, had the lowest team home run total.

This stemplot display motivates a follow-up question: which league hit more home runs in 2014? To compare the team totals for the two leagues, in Figure 2.3 we will put the leaves of the home run totals for the NL teams to the left of the stems and the leaves for the home run totals for the AL on the right.

Comparing the left and right stemplots, I would conclude that the American League teams tended to hit more home runs than the National League teams in 2014. Note that only one of

```
 National League     American League
                 |  9 | KCR
        SDP STL  | 10 |
            ARI  | 11 | TEX TBR
PHI NYM ATL MIA  | 12 | BOS MIN
    LAD SFG CIN  | 13 | SEA LAA
                 | 14 | CLE OAK NYY
CHC PIT LAA WAS MIL | 15 | CHW DET
                 | 16 | HOU
                 | 17 | TOR
             COL | 18 |
                 | 19 |
                 | 20 |
                 | 21 | BAL
```

Figure 2.3. Back-to-back stemplots of team home run numbers from the National and American Leagues for the 2014 season.

the 15 National League teams and three of the 15 American League teams hit more than 160 home runs. One possible explanation for this pattern is that the American League plays with the designated hitter who is a substitute hitter for the pitcher (in contrast to the National League where the pitcher comes to bat).

```
29 | 2
29 | 6
30 | 0022
30 | 5789
31 | 011144
31 | 6777
32 | 001234
32 | 7
33 | 013
```

Figure 2.4. Stemplot of the OBPs for the 2014 Major League teams.

On-Base Percentages

Actually, the number of home runs hit is not really a good reflection of a team's offensive production. A better indicator of run production is the team's on-base percentage (OBP) that measures the fraction of plate appearances in which the team gets on-base. In constructing a stemplot of the 30 team OBPs, we break the OBP between the 2nd and 3rd digits. Also we write two lines for each possible stem, where the leaves 0–4 are written on the first line and the leaves 5–9 on the second line. The completed stemplot is displayed in Figure 2.4. The third line tells us that two teams had an on-base percentage of .300 and two teams had an on-base percentage of .302.

What do we see in these displays?

1. The shape of these team OBPs is pretty symmetric.
2. An average OBP is about .311 and the team OBPs range from .292 (San Diego) to .333 (LA Dodgers).
3. One unusual feature that stands out is the three teams (Detroit, LA Dodgers, and Pittsburgh) in the last line with the largest values of OBP.

One can summarize these team OBPs by the computation of a median and quartiles. The median is the value that divides the data into a bottom half and a top half. Here we have 30 values—so the median is the average of the 15th and 16th largest OBPs:

$$\text{Median} = (.314 + .314)/2 = .314.$$

The quartiles divide the data into quarters. The lower quartile is the median of the lower half of the data, and the upper quartile is the median of the upper half. In our example, we can divide the team OBPs into an upper half of 15 values, and a lower half of 15 values. The median of the lower half is the 8th lowest value (.307), and the median of the upper half is the 8th largest value (.321). So

$$\text{lower quartile} = .307, \quad \text{upper quartile} = .321.$$

A five-number summary of these data is (lowest value, lower quartile, median, upper quartile, highest value) which is (.292, .307, .314, .321, .333). These five numbers divide the data roughly in quarters. So approximately $1/4$ or 25% of the team OBPs fall between .292 and .307, approximately 25% of the OBPs fall between .307 and .314, and so on.

Again, it's interesting to note the relative standing of the OBP of your favorite team. My team (the Phils) had a team OBP of .302 that puts it in the lower quarter of the dataset. It is likely that the Phillies would like to improve their offensive production during the 2014–15 off-season by signing free-agents or getting players by trade who appear to be effective in getting on-base.

2.2 A Tribute to Derek Jeter

Topics Covered: Dotplot, stemplot, time series plot, fitted line.

The 2014 baseball season was memorable for the retirement of one of baseball's most popular players, Derek Jeter. Jeter played shortstop for the New York Yankees for a 20-year period. Table 2.2 displays Jeter's batting statistics for his first six major league seasons. For all of the seasons of Jeter's career, we'll focus on a couple of interesting statistics—the number

Table 2.2. Derek Jeter's batting statistics for the first six seasons of his MLB career

Year	AB	R	H	2B	3B	HR	BB	SO	AVG	OBP	SLG	OPS
1995	48	5	12	4	1	0	3	11	.250	.294	.375	.669
1996	582	104	183	25	6	10	48	102	.314	.370	.430	.800
1997	654	116	190	31	7	10	74	125	.291	.370	.405	.775
1998	626	127	203	25	8	19	57	119	.324	.384	.481	.864
1999	627	134	219	37	9	24	91	116	.349	.438	.552	.989
2000	593	119	201	31	4	15	68	99	.339	.416	.481	.896

of home runs (HR) and the OPS statistic that is a good estimate of a player's overall hitting ability. In the following, we will not include the statistics for his 1995 rookie and his 2013 injury seasons since he only had 48 and 63 at-bats, but we'll analyze his batting data for the remaining 18 seasons.

Jeter's Home Runs

Scanning over Jeter's hitting statistics, we see that he displayed some power and hit a good number of home runs in his career. We graph his season by season home runs using a dotplot in Figure 2.5.

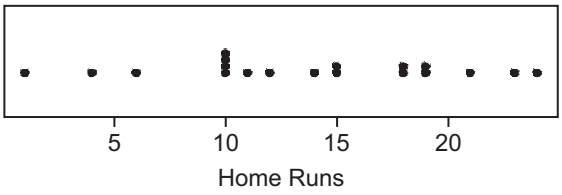


Figure 2.5. Dotplot of season home run numbers for Derek Jeter.

There appears to be much variation in these home run numbers—we see from the graph that the numbers range from 4 to 24. There is one cluster of values in the 10–15 range. The median number of home runs hit is 14.5 which seems to be a reasonable measure of the middle of this home run distribution.

Maybe some of the variation of Jeter's home run numbers can be explained by the age at which he hit them. A ballplayer generally improves in hitting ability in the early part of his career and declines in ability towards the end of his career. In Figure 2.6, we graph the home run count against the year.

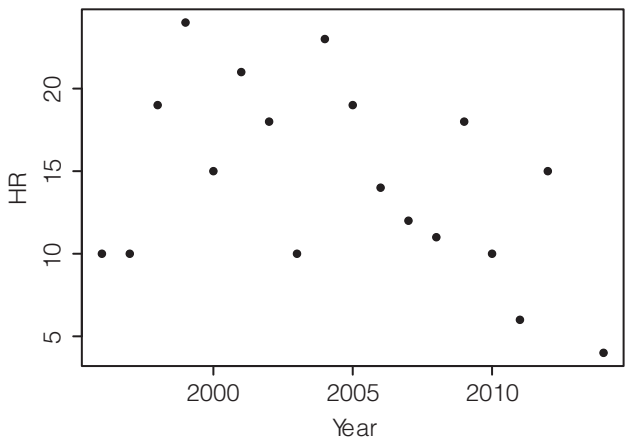


Figure 2.6. Time series plot of home run numbers for Derek Jeter.

We see that Jeter's home run numbers were in the 15–20 range during seasons around 2000, but it appears that his season home run counts generally decreased over the later years of his

career. We can summarize this decrease by drawing a line through the points in Figure 2.7. (We will discuss the use of one "best fitting" line in Chapter 4.)

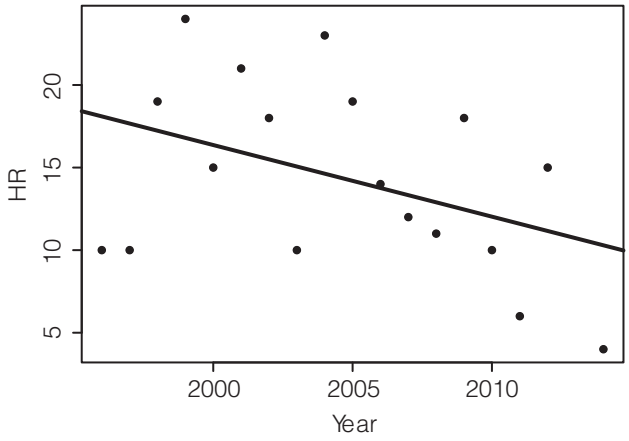


Figure 2.7. Time series plot of home run numbers for Derek Jeter with a "good fitting" line drawn on top.

The equation of this line is

$$\text{HR} = 883.68 - 0.4337 \times \text{Year}$$

So Jeter's home run count tended to decrease about .4 for each year of his career. There were some exceptions to this general pattern such as his 18 home runs hit in 2009 and his 15 home runs in 2012.

Jeter's Season OPS Values

Although Jeter's home runs decreased during his career, it's not clear that his hitting ability changed in a similar way. After all, good hitting is more than just hitting home runs. As explained in Chapter 1, a good estimate of hitting effectiveness is the OPS statistic. In Figure 2.8 we display Jeter's season OPS values for his 18 seasons using a stemplot.

```
6* | 1
6. |
7* | 14
7. | 7799
8* | 02344
8. | 5679
9* | 0
9. | 8
```

Figure 2.8. Stemplot of Jeter's season OPS values.

From the stemplot, we see that Jeter's OPS values are roughly bell-shaped centered about .83. Most of his season values fall between .770 and .890 with a few large extreme values. To see how Jeter's OPS values changed over the seasons, we construct the scatterplot in Figure 2.9 and add a smoothing curve to help identify the main pattern in the plot.

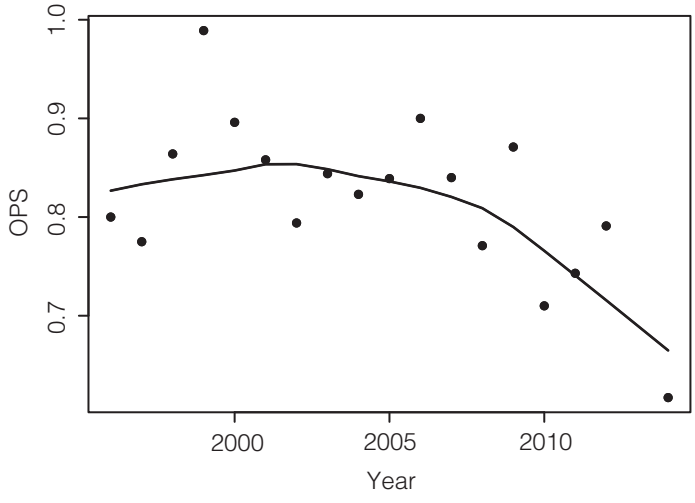


Figure 2.9. Time series plot of Jeter's OPS numbers

The general message from this graph is that Jeter's batting performance as measured by OPS was very consistent during the period 1996 through 2008, but then his performance dropped off in the final seasons of his career. He had an unusually high value in 1999 when his OPS was 0.989. Since OPS values in the $(0.8, 0.9)$ range were high (relative to other hitters), this figure reinforces the impression that Jeter was a consistently good hitter during his whole career.

2.3 A Tribute to Randy Johnson

Topics Covered: Stemplot, time series plot, summary statistics, comparison of distributions.

Randy Johnson was one of the greatest strikeout pitchers in modern baseball. He had a lifetime win/loss record of 303-166 and he won the Cy Young award as the best pitcher in the league for the 1995, 1999, 2000, 2001, and 2002 seasons. (In these seasons, a Cy Young award was given in each league.) We collect Johnson's pitching statistics for his 22 seasons in Major League Baseball; the statistics for the first six seasons are presented in Table 2.3. (One should note an unusual feature of innings pitched (IP) data typically presented. In 1989, Johnson pitched for 160 and $2/3$ of an inning—on Baseball-Reference this is represented as 160.2, but here we represent it as the more traditional decimal representation of 160.67.)

Table 2.3. Pitching statistics for Randy Johnson for the first six seasons of his career

Season	W	L	G	IP	H	R	SO	BB	ERA
1988	3	0	4	26	23	8	25	7	2.42
1989	7	13	29	160.67	147	100	130	96	4.82
1990	14	11	33	219.67	174	103	194	120	3.65
1991	13	10	33	201.33	151	96	228	152	3.98
1992	12	14	31	210.33	154	104	241	144	3.77
1993	19	8	35	255.33	185	97	308	99	3.24

Johnson's Strikeouts

The strikeout numbers, 25, 130, and so on, are hard to interpret since the number of innings pitched changes across seasons. A reasonable measure of strikeout ability that adjusts for the number of innings is the strikeout rate defined by

$$\text{SO Rate} = 9 \times \frac{\text{SO}}{\text{IP}}.$$

If we divide the count of strikeouts by the innings pitched, we get the number of strikeouts per inning. By multiplying the ratio SO/IP by 9, we get the number of strikeouts for a standard 9-inning game. A strikeout rate of 9 is a useful reference value, since it means that the pitcher struck out a batter per inning. (This particular rate value is quite rare.) Since there are 27 outs for a team during a game, a strikeout rate of 9 means that one third of the outs were strikeouts. If we compute the strikeout rate for all of Johnson's seasons, we get the table shown in Table 2.4. A stemplot of the rates is shown in Figure 2.10.

Table 2.4. Strikeout rates for Randy Johnson

Season	SO.Rate	Season	SO.Rate
1988	8.65	1999	12.06
1989	7.28	2000	12.56
1990	7.95	2001	13.41
1991	10.19	2002	11.56
1992	10.31	2003	9.87
1993	10.86	2004	10.62
1994	10.67	2005	8.41
1995	12.35	2006	7.55
1996	12.47	2007	11.43
1997	12.30	2008	8.46
1998	12.12	2009	8.06

```
   n: 22
 7* | 2
 7. | 59
 8* | 044
 8. | 6
 9* |
 9. | 8
10* | 13
10. | 668
11* | 4
11. | 5
12* | 01334
12. | 5
13* | 4
```

Figure 2.10. Stemplot of season strikeout rates for Randy Johnson.

Looking at the stemplot, we see that an average strikeout rate for Johnson is about 10.6. That means that he strikes out about one batter per inning. We see two clusters of strikeout rates. For a majority of his seasons, Johnson's strikeout rate was in the 10–12 range, although there were seven seasons where his rate was smaller than 9. To see when these high and low strikeout seasons occurred, we plot the rates against the season year in Figure 2.11. We place a horizontal line on our graph corresponding to the reference strikeout rate of 9.

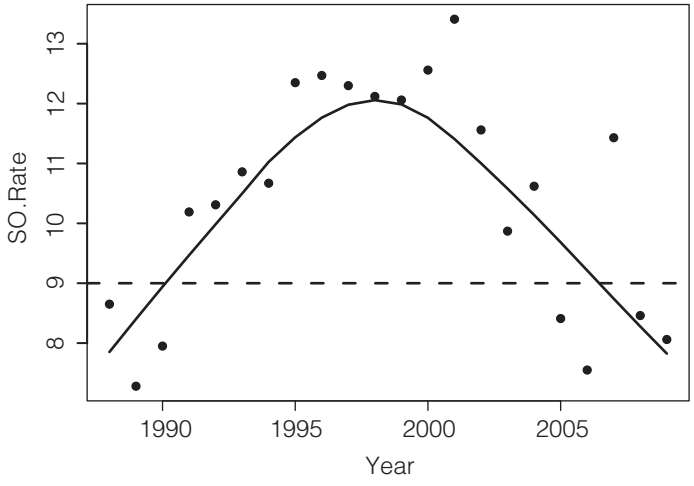


Figure 2.11. Time series plot of Johnson's season strikeout rates.

There is an interesting pattern in this plot of strikeout rates. We see that Johnson's strikeout rate exhibited a steady increase at the beginning of his career, hit a peak of about 12 around the year 1998, and then decreased steadily until his retirement in 2009. This is a common pattern of performance of baseball players.

Johnson's Win/Loss Record in 1995

Randy Johnson had a remarkable win/loss record in 1995—he won 18 and lost only two games. For this reason (and others), Johnson won the Cy Young award for the best pitcher in the American League in 1995. But was Johnson really the best AL pitcher in 1995?

One problem with a pitcher's win/loss record is that winning and losing is really a team accomplishment. A team will win games if it scores more runs than its opponent. This requires good pitching that allows few runs and good hitting that produces runs. It is possible that Johnson won so many games in 1995 because his team, the Mariners, scored a lot of runs when he was pitching. To see if this is true, we record the number of runs scored by the Mariners for every one of the 30 games that Johnson started in the 1995 season. The run numbers are shown below. We see that the Mariners scored three runs in the first game that Johnson started, 15 runs in the second game he started, and so on.

```
3  15  3  6  4  5  8  11  2  3  1  9  3  5  4
5   3  4  8  2  2  6   6  7  4  7  8  7  6  9
```

We display the runs scored for Johnson in Figure 2.12 using a stemplot.

We see that the distribution of runs scored is right skewed. We see that most of the runs scored are in the 3–7 range—it was unusual for the Mariners to score 0 or 1 runs, or to score more than 9 runs. The median runs scored for Johnson was 5 and the mean was 5.53 runs.

```
 1 | 0
 2 | 000
 3 | 00000
 4 | 0000
 5 | 000
 6 | 0000
 7 | 000
 8 | 000
 9 | 00
10 |
11 | 0
12 |
13 |
14 |
15 | 0
```

Figure 2.12. Stemplot of runs scored by the Mariners during games in which Johnson started in 1995.

```
   |  0 |00
  0|  1 |000000000000
 000|  2 |00000000000000
00000|  3 |000000000000
 0000|  4 |000000000000
  000|  5 |000000000000
 0000|  6 |0000000000000000
  000|  7 |00000
  000|  8 |00000
   00|  9 |0000000
     | 10 |0000000000
   0| 11 |000000
    | 12 |
    | 13 |
    | 14 |0
   0| 15 |000
```

Figure 2.13. Back-to-back stemplots of runs scored by the 1995 Mariners for games where Randy Johnson was the starter (LEFT) and games where Johnson was not the starter (RIGHT).

How does this distribution compare to the runs scored by the Mariners in the 1995 games where Johnson was not a starter? Figure 2.13 displays back-to-back stemplots of the runs scored for the two groups of games.

Comparing the two stemplots in Figure 2.13, we see that the Mariners tended to score about the same number of runs in games in which Johnson did not start than they did in games in which Johnson started. The median and mean runs scored for Mariners games when Johnson was not starting are 5 and 5.48, respectively. If we compare medians or means of the two groups, we see that the Mariners tended to average the same amount of runs when Johnson was starting or when Johnson was not starting. So Johnson did not receive unusually good run support from his team.

2.4 Analyzing Baseball Attendance

Topics Covered: Histogram, stemplot, data distribution.

Baseball is a business. Each of the thirty major league teams is privately owned and each wants to make a profit. Much of the revenue produced by the teams comes from the money that is made from ticket and concession sales. So it is desirable for a team to have high attendance at the games played at its ballpark. We will see in this study that some teams are very successful and other teams are not successful in getting fans to come to their games.

Table 2.5 shows the mean attendance at the home ballpark for all thirty major league teams in 2014. We note that the teams who played in the 2014 World Series, San Francisco (SFN) and Kansas City (KCA), had mean attendance of 41,589 and 24,154 per game, respectively. Are these large or small values? We can answer this question by looking at the distribution of mean attendance for all teams.

Table 2.5. Mean home attendance for all Major League Baseball teams in 2014

Team	Attendance	Team	Attendance	Team	Attendance
ANA	38,221	DET	36,015	PHI	29,924
ARI	25,602	HOU	21,628	PIT	30,155
ATL	29,065	KCA	24,154	SDN	27,103
BAL	30,426	LAN	46,696	SEA	25,486
BOS	36,495	MIA	21,386	SFN	41,589
CHA	20,381	MIL	34,536	SLN	43,712
CHN	32,742	MIN	27,785	TBA	17,858
CIN	30,576	NYA	41,995	TEX	33,565
CLE	17,746	NYN	26,528	TOR	29,327
COL	33,090	OAK	24,736	WAS	31,844

We first construct a histogram of the attendance for all teams. We chose to group the data using the bins $(15{,}000, 20{,}000]$, $(20{,}000, 25{,}000], \ldots, (45{,}000, 50{,}000]$ and count the number of values in each bin. The graph of the grouped data, called a histogram, is shown in Figure 2.14.

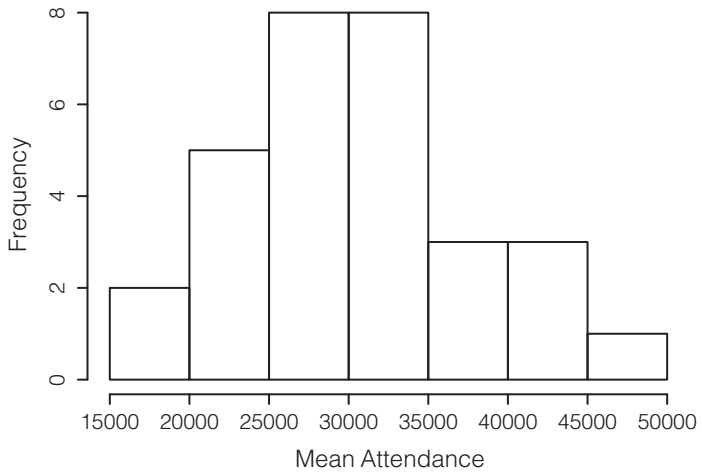


Figure 2.14. Histogram of average attendance for the Major League teams in 2014.

There are a number of interesting features about these data that we see from the histogram.

- There is a large range in the average attendance, from about 15,000 to 50,000.
- The shape of this distribution is symmetric about a typical average attendance of 30,000.
- There are two teams that had small average attendance under 20,000.

To look further at these teams' attendance, we construct a stemplot displayed in Figure 2.15. In addition, we construct a Cleveland-style dotplot in Figure 2.16—this is a display of labeled data where one lists the team names on the vertical axis and plots attendance values on the horizontal scale.

```
   n: 30
1. | 77
2* | 01144
2. | 55677999
3* | 00012334
3. | 668
4* | 113
4. | 6
```

Figure 2.15. Stemplot of average attendance of the Major League teams in 2014.

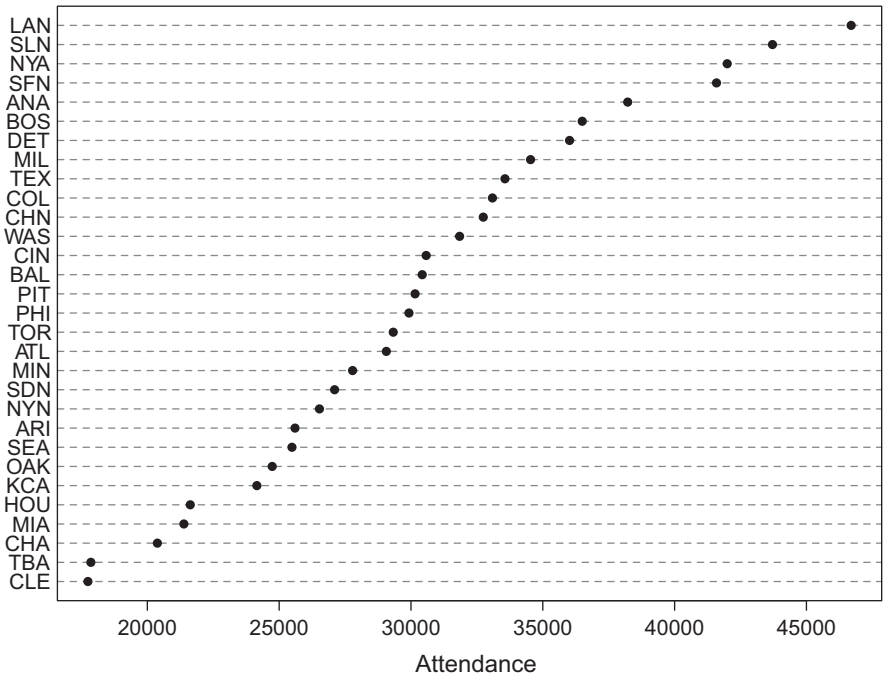


Figure 2.16. Cleveland-style dotplot of average attendance of the Major League teams in 2014.

The graphs in Figure 2.15 and Figure 2.16 give us additional insight about the attendance averages. The two weak attendance teams are Cleveland (CLE) and Tampa Bay (TBA). The team with the highest average attendance in 2014 was the Los Angeles Dodgers who had a successful 94-68 record in this season. But winning games and attendance don't always go hand in hand. Cleveland (CLE) had a 85-77 winning record but a small average attendance, but Colorado (COL) and Texas (TEX), two teams with weak 66-96 and 67-95 records, had large average attendance. It would be interesting for a future study to look into the impact of a number

of variables, such as the population of the surrounding region and the newness of the ballpark, on the average home attendance of a baseball team.

2.5 Manager Statistics: the Use of Sacrifice Bunts

Topics Covered: Dotplot, comparison of distributions.

In this chapter, we've looked at statistics for teams and for individual players. Here we'll look at statistics for baseball managers. This might surprise you—how can we measure characteristics of a manager? Managers do have different styles or strategies that they use in managing a game, and we can contrast managers by measuring how often particular strategies are used.

Here we focus on the use of the sacrifice bunt, one of the more popular strategic plays in baseball. Suppose there is a runner on first base with 0 or 1 out. Then the manager may elect to have the batter bunt, which is a short hit in the infield. If the bunt is successful, the runner on first advances to second. If the batter is out, this play is called a sacrifice bunt because the batter is sacrificing his at-bat to advance the runner to second. Is the sacrifice bunt a good strategy? It can be a good strategy when there is a relatively weak batter at the plate or the objective is to score a single run.

Table 2.6 gives the number of sacrifice bunts for all 30 Major League teams in the 2013 season.

Table 2.6. Successive count of sacrifice hits for all Major League teams in 2013

Team	SH	League	Team	SH	League
ANA	37	AL	MIL	77	NL
ARI	50	NL	MIN	29	AL
ATL	58	NL	NYA	36	AL
BAL	27	AL	NYN	53	NL
BOS	24	AL	OAK	21	AL
CHA	19	AL	PHI	57	NL
CHN	43	NL	PIT	62	NL
CIN	85	NL	SDN	52	NL
CLE	31	AL	SEA	26	AL
COL	65	NL	SFN	66	NL
DET	32	AL	SLN	56	NL
HOU	46	NL	TBA	24	AL
KCA	37	AL	TEX	45	AL
LAN	71	NL	TOR	29	AL
MIA	57	NL	WAS	68	NL

The dotplot in Figure 2.17 displays the number of sacrifice bunts for all teams. The median number of "sac-bunts" was 45.5. But we note a wide spread—the Chicago White Sox only had 19 sacrifice bunts all season and other teams sacrificed over 70 times (in a 162 game season).

How can we explain the wide spread in the number of sacrifice bunts between teams? Perhaps some managers don't think that the sacrifice bunt is a good strategy. Or possibly a

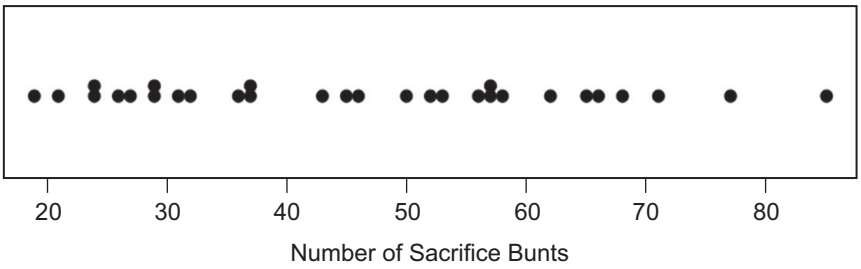


Figure 2.17. Dotplot of number of sacrifice bunt attempts by the Major League managers in the 2013 season.

manager of a team with strong pitching uses a sacrifice more often than a manager of a team with weak pitching.

Is it possible that the number of sacrifice bunts differs between National League and American League teams? We can check this by constructing two parallel dotplots in Figure 2.18—one for the NL teams and a second for the AL teams.

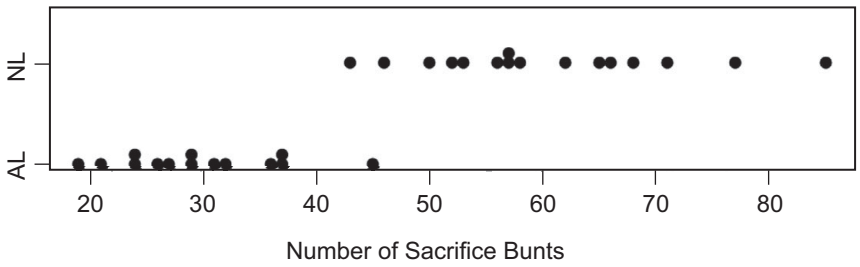


Figure 2.18. Parallel dotplots of number of sacrifice bunt attempts by National and American League managers.

We see in the Figure 2.18 display that there is quite a difference in the number of sacrifice bunts attempted in the two leagues. The median number of attempts for NL teams is 57.5 (about 1 in every 3 games) contrasted with 29 (about 1 in every 5 games) for the AL teams. There actually is one simple explanation for this discrepancy in sacrifice bunts—the designated hitter rule. Because National League pitchers (typically weak hitters) come to bat, it is common for them to attempt a sacrifice bunt when there is a runner on first base. They do this to avoid a double play and to advance the runner to second. In contrast, in the AL, the pitcher doesn't bat (the designated hitter does instead) and there is much less reason to perform this strategy.

The league policy does explain many of the differences in attempts that we see. However, even within one league, we still see large differences in sacrifice bunt attempts. It is interesting to note that the Chicago Cubs 43 sac-bunt attempts is a small value even compared to other NL teams, and the NL values ranged between 52 and 85. Why do these differences exist? Well, the number of sacrifice bunts attempted depends partly on the batting abilities of the players and the beliefs of the manager regarding the value of the sacrifice bunt. It would be interesting to see if there is any relationship between the number of sacrifice bunts and other batting statistics.

Also, one could see if particular managers have tendencies over many years to attempt a large or small number of sacrifice bunts.

2.6 Exercises

2.0. Table 2.7 displays the slugging percentages (SLGs) for Rickey Henderson for his first 23 years in the major leagues.

Table 2.7. Seasonal slugging percentages for Rickey Henderson

AGE	SLG	AGE	SLG	AGE	SLG	AGE	SLG
20	.336	26	.516	32	.423	38	.342
21	.399	27	.469	33	.457	39	.347
22	.437	28	.497	34	.474	40	.466
23	.382	29	.399	35	.365	41	.305
24	.421	30	.399	36	.447	42	.351
25	.458	31	.577	37	.344		

(a) Construct a stemplot of the slugging percentages.
(b) Compute a five-number summary of the SLGs. What was a representative slugging percentage for Henderson?
(c) Construct a time series plot of the SLG against his age. Comment on any pattern that you see in this plot.
(d) Based on your work in (c), do you have any explanations for the low slugging percentages that Henderson had in his career?

2.1. Table 2.8 gives the career batting statistics for the great Yankee player Joe DiMaggio.

Table 2.8. Career Statistics for Joe DiMaggio

Year	G	AB	H	HR	AVG	OBP	SLG
1936	138	637	206	29	0.323	0.352	0.576
1937	151	621	215	46	0.346	0.412	0.673
1938	145	599	194	32	0.324	0.386	0.581
1939	120	462	176	30	0.381	0.448	0.671
1940	132	508	179	31	0.352	0.425	0.626
1941	139	541	193	30	0.357	0.440	0.643
1942	154	610	186	21	0.305	0.376	0.498
1946	132	503	146	25	0.290	0.367	0.511
1947	141	534	168	20	0.315	0.391	0.522
1948	153	594	190	39	0.320	0.396	0.598
1949	76	272	94	14	0.346	0.459	0.596
1950	139	525	158	32	0.301	0.394	0.585
1951	116	415	109	12	0.263	0.365	0.422

(a) Construct a dotplot of DiMaggio's yearly batting averages on the number line below.

```
+---------+---------+---------+---------+---------+-AVG
.200      .240      .280      .320      .360      .400
```

(b) Describe the basic features of this distribution (shape, average value, spread, and unusual values).

(c) What proportion of the years was DiMaggio at least a .300 hitter?

(d) Construct a dotplot of DiMaggio's yearly hits on the number line below.

```
+---------+---------+---------+---------+---------+-H
100        120        140        160        180        200
```

(e) What was a median number of yearly hits for DiMaggio? Were there any unusually small or large hit values? What is a possible explanation for these unusual values?

2.2. (Continuation of Exercise 2.1)

(a) Construct a time series plot of DiMaggio's AVG values.

(b) From the graph, DiMaggio appeared to peak in AVG two times during his career. Which two years?

(c) From the graph, did you detect any gaps in DiMaggio's career? What is a possible explanation for these gaps?

(d) Generally, ballplayers mature slowly and reach their maximum performance during the middle of their careers. Do you see any evidence for this maturation in DiMaggio's AVG plot? Explain.

(e) Also, ballplayers generally decline in ability during the last part of their career. Do you see any decline in DiMaggio's AVG plot? Explain.

2.3. (Joe DiMaggio data from Exercise 2.1)

(a) Construct a stemplot of DiMaggio's on-base percentages (OBP) for his 13 seasons.

(b) Compute DiMaggio's mean OBP and his median OBP.

(c) Compare the mean and median that you computed in (b). Which is a better measure of a representative season OBP for DiMaggio?

2.4. (Joe DiMaggio data from Exercise 2.1)

A stemplot of DiMaggio's season HR numbers is shown below:

```
1 | 24
1 |
2 | 01
2 | 59
3 | 00122
3 | 9
4 |
4 | 6
```

(a) Find a five-number summary of the HR numbers.

(b) Based on your calculations in part (a), half of all the HR numbers are smaller than ____ and approximately the middle 50 percent of the season HRs are between ____ and ____ .

2.5. Table 2.9 shows the career hitting statistics for Babe Ruth, who was considered by ESPN to be the second greatest athlete of the 20th century (after Michael Jordan).

Table 2.9. Career Statistics for Babe Ruth

Year	G	AB	H	HR	AVG	OBP	SLG
1914	5	10	2	0	0.200	0.200	0.300
1915	42	92	29	4	0.315	0.376	0.576
1916	67	136	37	3	0.272	0.322	0.419
1917	52	123	40	2	0.325	0.385	0.472
1918	95	317	95	11	0.300	0.411	0.555
1919	130	432	139	29	0.322	0.456	0.657
1920	142	457	172	54	0.376	0.533	0.849
1921	152	540	204	59	0.378	0.512	0.846
1922	110	406	128	35	0.315	0.434	0.672
1923	152	522	205	41	0.393	0.545	0.764
1924	153	529	200	46	0.378	0.513	0.739
1925	98	359	104	25	0.290	0.393	0.543
1926	152	495	184	47	0.372	0.516	0.737
1927	151	540	192	60	0.356	0.486	0.772
1928	154	536	173	54	0.323	0.463	0.709
1929	135	499	172	46	0.345	0.430	0.697
1930	145	518	186	49	0.359	0.493	0.732
1931	145	534	199	46	0.373	0.495	0.700
1932	133	457	156	41	0.341	0.489	0.661
1933	137	459	138	34	0.301	0.442	0.582
1934	125	365	105	22	0.288	0.448	0.537
1935	28	72	13	6	0.181	0.359	0.431

(a) Construct a dotplot of the season home run numbers for Babe Ruth.
(b) Comment on the basic shape of the distribution of home run numbers.
(c) Can you think of some possible reasons why Ruth had so many seasons with a small number of home runs?
(d) Since the number of at-bats is not constant across seasons, it may be better to consider the home run rate, which is obtained by dividing the number of home runs by the number of at-bats. So, for example, Ruth's home run rate in 1915 is given by

$$\text{HOME RUN RATE} = \frac{4}{92} = .0435.$$

The home run rates for Ruth in his 22 seasons are given below:

```
   0  .043  .022  .016  .035  .067  .118  .109
.086  .079  .087  .070  .095  .111  .101  .092
.095  .086  .090  .074  .060  .083
```

(e) Construct a dotplot of the home run rates.

(f) Summarize the key features of the home run rates. Explain why it is better to look at the home run rates instead of the home run numbers.

2.6. (Continuation of Exercise 2.5) Consider again the home rate rates for Babe Ruth displayed in Exercise 2.5.
(a) Graph the home run rates with a time series plot.
(b) What pattern do you see in the graph you made in (a)? Draw a smooth curve over the points.
(c) Using the smooth curve you constructed in (b), what was Babe Ruth's peak year with respect to home run rate? Is this the same year as the year when he hit the greatest number of home runs?
(d) Repeat parts (a)–(c) using Ruth's batting average statistic.

2.7. (Babe Ruth data from Exercise 2.5)
(a) Construct a dotplot of Ruth's season SLG values.
(b) Find the mean and median SLG values.
(c) Which average computed in (b) seems to be most representative of Ruth's slugging ability? Explain.
(d) In the table of Ruth's data, Ruth's career slugging percentage is .69. Is this the same as the mean SLG value that you computed in (b)? If the numbers are different, can you explain why?

2.8. (Babe Ruth data from Exercise 2.5)
A stemplot of Ruth's season batting averages is shown below. (Here the stem is the first digit, so a batting average of .315 would have a stem of 3 and a leaf of 1.)

```
1 | 8
2 | 0
2 |
2 |
2 | 7
2 | 89
3 | 0011
3 | 222
3 | 4455
3 | 77777
3 | 9
```

(a) Compute a five-number summary.
(b) Looking back at the table of data in Exercise 2.3, find the years where Ruth's batting average was above the median.
(c) Find the years where Ruth's average was below the median.
(d) Based on (b) and (c), do you think the median is a representative AVG for Ruth?
(e) In the table of Ruth's data, Ruth's career AVG is equal to .342. Compare this value to the median and explain why they are different.

2.9. Table 2.10 shows the career pitching statistics for Hall of Famer Sandy Koufax. (Koufax is the only pitcher to make ESPN's list of the top 50 Greatest Athletes of the 20th century.)

Table 2.10. Pitching Statistics for Sandy Koufax

Year	G	W	L	IP	H	SO	BB	ERA
1955	12	2	2	41.70	33	30	28	3.02
1956	16	2	4	58.70	66	30	29	4.91
1957	34	5	4	104.30	83	122	51	3.88
1958	40	11	11	158.70	132	131	105	4.48
1959	35	8	6	153.30	136	173	92	4.05
1960	37	8	13	175.00	133	197	100	3.91
1961	42	18	13	255.70	212	269	96	3.52
1962	28	14	7	184.30	134	216	57	2.54
1963	40	25	5	311.00	214	306	58	1.88
1964	29	19	5	223.00	154	223	53	1.74
1965	43	26	8	335.70	216	382	71	2.04
1966	41	27	9	323.00	241	317	77	1.73

(a) Construct a stemplot of the season ERA values for Koufax (the breakpoint between the stem and the leaf will be at the decimal point).
(b) There are two clusters in this dataset. Identify the two clusters and explain which season years are identified with the two clusters.
(c) Construct a stemplot of Koufax winning percentages (PCT). Describe the features of this dataset. Would it be accurate to say that Koufax was a successful pitcher? Why? What proportion of seasons did Koufax have a winning record?

2.10. (Sandy Koufax data from Exercise 2.9.)
(a) Draw a stemplot with five leaves per stem for the season strikeout counts. Describe the main features of this dataset. Are there any SO counts that appear unusually small or large? Is there any explanation for these unusual values?
(b) Find the mean and median SO number.
(c) Which average in (b) is the best measure of the center of the distribution? Explain. (It might help to look at the stemplot of the season strikeout counts from Exercise 2.9 (c).)

2.11. (Sandy Koufax data from Exercise 2.9)
(a) Compute the mean and median of Koufax season ERAs.
(b) Suppose Koufax played an additional season and his ERA was 6.5. Compute the mean and median of Koufax's ERAs with this additional data value.
(c) Which average (mean or median) changed the most when you added this new ERA value? Explain why this average changed.
(d) Koufax's career ERA was 2.76. Explain why this career ERA is different from the averages you computed in part (a).

2.12. Table 2.11 gives career pitching statistics for the Hall of Famer Bob Feller. Note that there are no statistics given for the three-year period 1942–1944—Feller served in the Navy during World War II between 1942 and 1945.

Table 2.11. Pitching statistics for Bob Feller

Year	G	W	L	IP	H	BB	SO	ERA
1936	14	5	3	62.00	52	47	76	3.34
1937	26	9	7	148.67	116	106	150	3.39
1938	39	17	11	277.67	225	208	240	4.08
1939	39	24	9	296.67	227	142	246	2.85
1940	43	27	11	320.33	245	118	261	2.61
1941	44	25	13	343.00	284	194	260	3.15
1945	9	5	3	72.00	50	35	59	2.50
1946	48	26	15	371.33	277	153	348	2.18
1947	42	20	11	299.00	230	127	196	2.68
1948	44	19	15	280.33	255	116	164	3.56
1949	36	15	14	211.00	198	84	108	3.75
1950	35	16	11	247.00	230	103	119	3.43
1951	33	22	8	249.67	239	95	111	3.50
1952	30	9	13	191.67	219	83	81	4.74
1953	25	10	7	175.67	163	60	60	3.59
1954	19	13	3	140.00	127	39	59	3.09
1955	25	4	4	83.00	71	31	25	3.47
1956	19	0	4	58.00	63	23	18	4.97

(a) Construct a stemplot of Feller's strikeout counts for his 18 seasons. What is the basic shape of this distribution? Are there any unusually small or large SO counts?

(b) It is difficult to tell which years Feller was especially good at striking out batters since the innings pitched (IP) is not constant across years. To adjust for the innings pitched, one can compute the strikeout ratio, the number of strikeouts divided by the innings pitched. Here are the strikeout ratios for Feller's 18 seasons:

```
1.23  1.01  .87  .83  .82  .76  .82  .94  .66
 .59   .51  .48  .45  .42  .34  .42  .30  .31
```

Graph these SO ratios against year number. Describe any pattern that you see in the graph. What does that tell you about Feller's ability to strike out batters over time?

(c) Use Feller's lifetime average number of wins, losses, and strikeouts to fill in the missing years 1942–1944 and the partial year 1945. What are his new career totals in these categories?

2.13. (Bob Feller data from Exercise 2.12)

(a) Construct a dotplot of Feller's strikeout ratios for the 18 seasons. (These ratios are listed in part (b) of Exercise 2.12.)

(b) Compute the mean and median strikeout ratio. Explain why the mean is larger than the median for this dataset.

2.14. Figure 2.19 displays a histogram of the on-base percentages (OBP) for the 944 MLB players who batted during the 2014 baseball season.

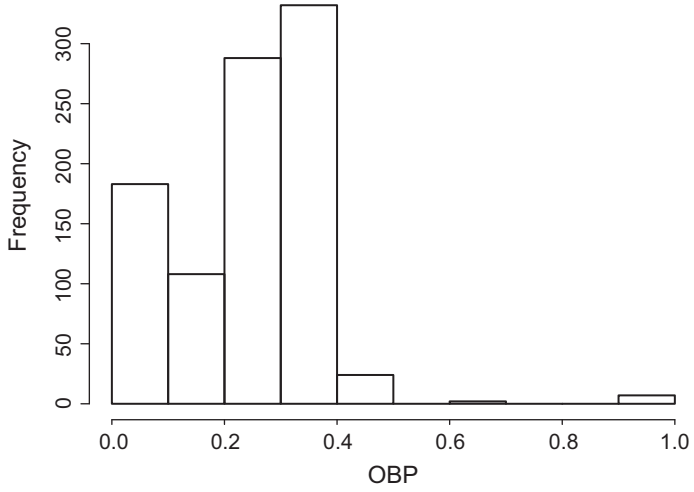


Figure 2.19. Histogram of OBPs of 2014 MLB players.

(a) Describe the basic shape of the histogram and any unusual features.
(b) What percentage of players had an OBP value between .3 and .4?
(c) What percentage of players had an OBP smaller than .2?
(d) There were a few players that had an OBP value of 1 during the 2014 season. (This means that they reached base 100% of the time.) This may seem surprising—can you offer any possible explanation for these large values?

2.15. (Exercise 2.14 continued.) If one graphs the OBPs for only those National League hitters who had at least 300 at-bats (AB), one obtains the histogram in Figure 2.20. (In the following, we will refer to the players with at least 300 AB as the "regulars.")

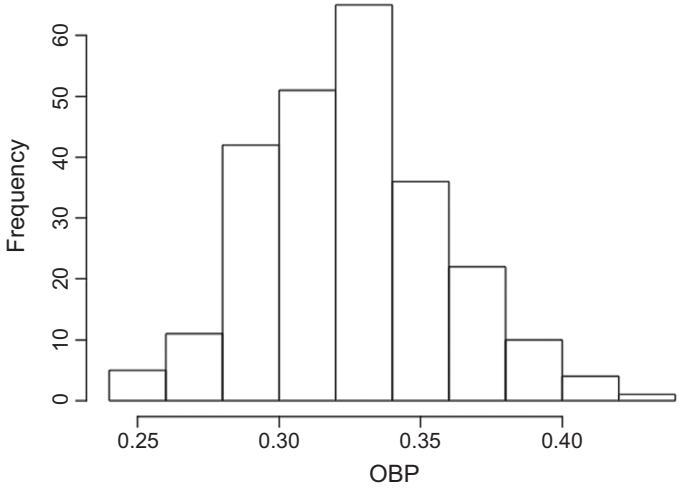


Figure 2.20. Histogram of OBPs of 2014 MLB players with at least 300 at-bats.

(a) Describe the shape of this histogram. Why does this histogram look so different from the histogram of the OBP of all NL players in Figure 2.18?
(b) What percent of NL regulars had an OBP value exceeding .3?
(c) What percent of NL regulars had an OBP value between .3 and .4?
(d) From the histogram, what OBP value would you consider outstanding? Why?

2.16. Figure 2.21 shows a histogram of the number of home runs for all regular players in the 2014 season with at least 300 at-bats.
(a) Describe the basic shape of the histogram and any unusual characteristics.
(b) What percentage of NL regular players hit fewer than ten home runs?
(c) What percentage of NL regular players hit more than 20 home runs?
(d) How many home runs did an NL regular player hit, on average?

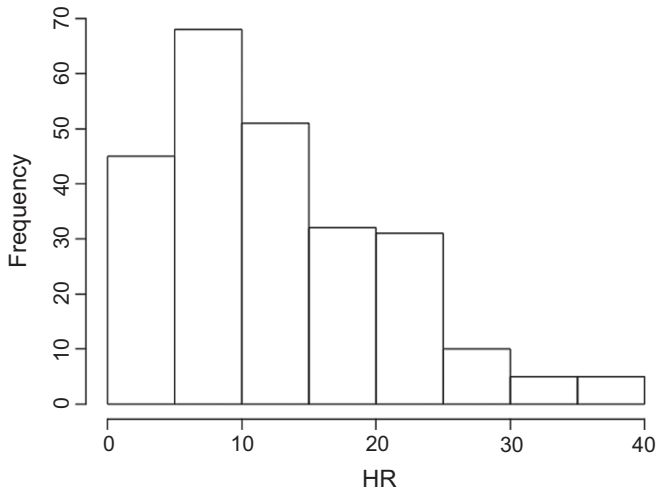


Figure 2.21. Histogram of count of home runs of 2014 MLB players with at least 300 at-bats.

2.17. (Exercise 2.16 continued.) One explanation for the wide variation of home run totals shown in Figure 2.21 is that players had different numbers of at-bats. If we divide a player's home run count by his number of at-bats, one obtains a player's home run rate. (A rate of .08 means a batter hit a home run in 8% of his at-bats.) Figure 2.22 shows a histogram of the home run rates for the 2014 MLB regulars.
(a) Describe the basic shape of this distribution of home run rates. Are there any unusual values?
(b) What is an average home run rate among these NL regulars?
(c) If a batter comes to bat 100 times, how many home runs do you expect him to hit?
(d) What percent of NL regulars had a home run rate under .02?
(e) A hitter is considered to be an unusually good home run hitter if his rate of hitting home runs exceeds 10% or .1. What percent of regular NL hitters fall into this classification? Can you guess which players are in this class?

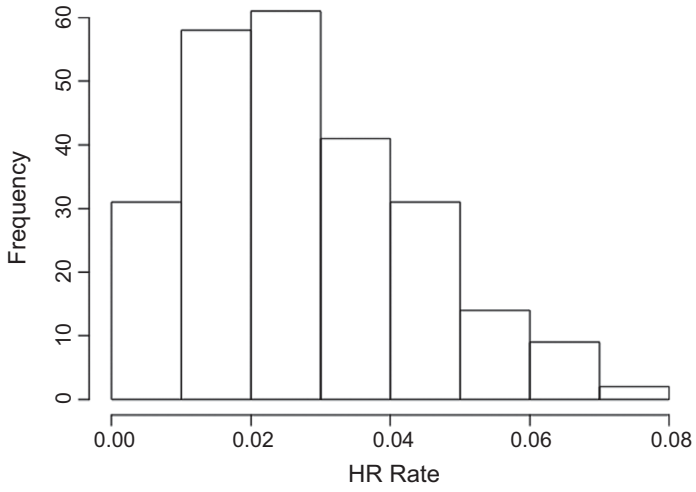


Figure 2.22. Histogram of count of home run rates of 2014 MLB players with at least 300 at-bats.

2.18. Table 2.12 displays the length (in minutes) of a sample of 60 baseball games in 2014.

Table 2.12. Length in minutes of a selection of 2014 baseball games

169	181	178	212	159	213	156	197	201	217
241	195	165	194	163	175	180	242	159	190
169	184	227	166	259	182	355	201	199	206
197	216	196	144	202	169	182	191	181	196
173	205	185	154	198	165	175	221	187	193
155	176	171	218	185	203	222	194	215	205

(a) Construct a stemplot of these game times.
(b) What is the general shape of the distribution of times? Are there any unusually short or long games?
(c) What is a median length of a baseball game?
(d) What fraction of games are over three hours?
(e) Why is there such a large spread of times? Can you think of some variables that might influence the length of a baseball game?

2.19. (Length of baseball games from Exercise 2.18.)
A stemplot of the times of games (in minutes) is shown below.

```
1 | 2: represents 12
leaf unit: 1
           n: 60
  14 | 4
  15 | 45699
  16 | 3556999
  17 | 135568
```

```
   18 | 011224557
   19 | 013445667789
   20 | 1123556
   21 | 235678
   22 | 127
   23 |
   24 | 12
   25 | 9
HI: 355
```

(a) Find the five-number summary of the lengths of the 2014 baseball games.
(b) Approximately the middle half of the game lengths is between ____ and ____ .
(c) The mean and standard deviation of the game lengths are 193.5 minutes and 31.6 minutes, respectively. (The standard deviation is a measure of spread defined in Chapter 3.) Find an interval defined by (mean − standard deviation, mean + standard deviation) .
(d) Find the proportion of games that fall in the interval you found in part (c).
(e) If the data is bell-shaped, then one would expect 68% of the data to fall in the interval in (c). Here you should find that the proportion in (d) is significantly larger than .68. Can you explain why?

2.20. For each of 60 baseball games in 2014, the total number of home runs was recorded—the numbers are shown in Table 2.13.

Table 2.13. Number of home runs hit in a sample of 2014 baseball games

1	1	2	1	3	1	1	0	1	0
1	0	0	0	3	2	2	1	1	2
1	1	5	1	1	1	1	1	2	1
2	4	1	0	1	0	0	5	3	3
2	3	3	4	2	0	0	4	3	3
2	1	2	2	1	1	4	1	3	2

(a) Construct a frequency table for the number of home runs.
(b) Graph the frequencies using a bar chart.
(c) Find the proportion of games where exactly two home runs are hit.
(d) Find the proportion of games where four or more home runs are hit.
(e) What is an average number of home runs hit during a game? Explain how you computed this average number.

2.21. Table 2.14 displays basic batting statistics for the 2014 American League teams.
(a) Draw a stemplot of the batting averages (AVG) for the 15 AL teams.
(b) Find the median batting average and find a team that has a batting average close to the median value.
(c) Find the team that has the highest batting average and the team with the lowest batting average.

Table 2.14. Batting statistics for the 2014 American League teams

Team	AVG	OBP	SLG	AB	R	H	2B	3B	HR
Baltimore	0.256	0.311	0.422	5596	705	1434	264	16	211
Boston	0.244	0.316	0.369	5551	634	1355	282	20	123
Chicago	0.253	0.310	0.398	5543	660	1400	279	32	155
Cleveland	0.253	0.317	0.389	5575	669	1411	284	23	142
Detroit	0.277	0.331	0.426	5630	757	1557	325	26	155
Houston	0.242	0.309	0.383	5447	629	1317	240	19	163
Kansas City	0.263	0.314	0.376	5545	651	1456	286	29	95
Los Angeles	0.259	0.322	0.406	5652	773	1464	304	31	155
Minnesota	0.254	0.324	0.389	5567	715	1412	316	27	128
New York	0.245	0.307	0.380	5497	633	1349	247	26	147
Oakland	0.244	0.320	0.381	5545	729	1354	253	33	146
Seattle	0.244	0.300	0.376	5450	634	1328	247	32	136
Tampa Bay	0.247	0.317	0.367	5516	612	1361	263	24	117
Texas	0.256	0.314	0.375	5460	637	1400	260	28	111
Toronto	0.259	0.323	0.414	5549	723	1435	282	24	177

(d) For a second statistic of your choice, construct a stemplot. Find a team that has an average value of the statistic. Also find a team that has the smallest value, and a team that has the largest value of the statistic that you chose.

2.22. (Batting statistics for 2014 AL teams from Exercise 2.21)
(a) Draw a stemplot of the HR numbers for the 15 AL teams.
(b) Find the five-number summary.
(c) Approximately ____ percent of the HR numbers are smaller than 155.
(d) Approximately ____ percent of the HR numbers are larger than 146.
(e) Find a team that has an HR number that is close to the average.

2.23. Table 2.15 gives the year of birth and earned run average (ERA) for all 57 pitchers who have been inducted into the Baseball Hall of Fame.
(a) Draw a stemplot of the ERAs of these Hall of Fame pitchers.
(b) Discuss the general shape of the distribution of ERAs and any unusual characteristics of this data.
(c) Find an average ERA and a Hall of Fame pitcher who has (approximately) this average ERA.
(d) Find the pitchers who have the lowest and highest ERAs.
(e) What might explain this wide range of ERAs if all of these pitchers are in the Hall of Fame?

2.24. (Exercise 2.23 continued.) Consider the years of birth of the pitchers in the Hall of Fame.
(a) Construct a frequency table of the years of birth, using the intervals in the table below. Place the counts and the proportions in the table.
(b) Construct a histogram from the above frequency table.
(c) Describe the basic shape of the distribution.

Table 2.15. Year of birth and ERA for pitchers who have been selected in the Baseball Hall of Fame

Pitcher	Birthyear	ERA	Pitcher	Birthyear	ERA
Grover Alexander	1887	2.56	Sandy Koufax	1935	2.76
Charles Bender	1884	2.46	Bob Lemon	1920	3.23
Mordecai Brown	1876	2.06	Juan Marichal	1937	2.89
Jim Bunning	1931	3.27	Richard Marquard	1889	3.08
Steve Carlton	1944	3.22	Christy Mathewson	1880	2.13
Jack Chesbro	1874	2.68	Joe McGinnity	1871	2.66
John Clarkson	1861	2.81	Hal Newhouser	1921	3.06
Stan Coveleski	1889	2.89	Charles Nichols	1869	2.95
W.A. Cummings	1848	2.78	Phil Niekro	1939	3.35
Jay Hanna Dean	1911	3.02	Satchel Paige	1906	3.29
Don Drysdale	1936	2.95	Jim Palmer	1945	2.86
Urban Faber	1888	3.15	Herb Pennock	1894	3.60
Bob Feller	1918	3.25	Gaylord Perry	1938	3.11
Rollie Fingers	1946	2.90	Eddie Plank	1875	2.35
Edward Ford	1928	2.75	Eppa Rixey	1891	3.15
Rube Foster	1888	2.36	Robin Roberts	1926	3.41
James Galvin	1856	2.87	Amos Rusie	1871	3.07
Bob Gibson	1935	2.91	Nolan Ryan	1947	3.17
Vernon Gomez	1908	3.34	Tom Seaver	1944	2.86
Burleigh Grimes	1893	3.53	Warren Spahn	1921	3.09
Robert Grove	1900	3.06	Don Sutton	1945	3.26
Jesse Haines	1878	3.64	Dazzy Vance	1891	3.24
Waite Hoyt	1899	3.59	George Waddell	1876	2.16
Carl Hubbell	1903	2.98	Ed Walsh	1881	1.82
Jim Hunter	1946	3.26	Mickey Welch	1859	2.71
Ferguson Jenkins	1943	3.34	Hoyt Wilhelm	1923	2.52
Walter Johnson	1887	2.16	Vic Willis	1876	2.63
Addie Joss	1880	1.89	Early Wynn	1920	3.54
Tim Keefe	1857	2.62	Cy Young	1867	2.63

(d) You should note that the Hall of Fame pitchers are not uniformly distributed over the six eras 1840–1859, 1860–1879, . . . , 1940–1959. Can you explain this pattern? Were pitchers better during particular years of baseball?

Interval for year of birth	Count	Proportion
1840 to 1859		
1860 to 1879		
1880 to 1899		
1900 to 1919		
1920 to 1939		
1940 to 1959		

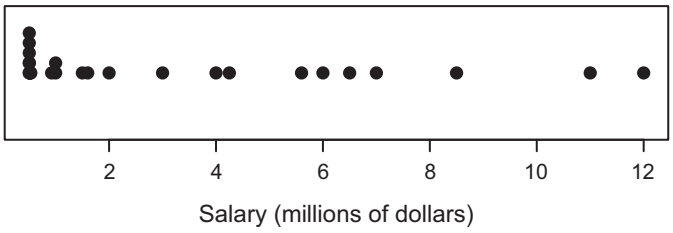
2.25. The salaries of the players on the 2014 Atlanta Braves are displayed in Table 2.16.

Table 2.16. Salaries of the 2014 Atlanta Braves

Name	Salary	Name	Salary
Freddie Freeman	$8,500,000	Eric Young	$1,000,000
Andrelton Simmons	$3,000,000	Josh Outman	$925,000
Nick Markakis	$11,000,000	Shelby Miller	$535,000
Julio Teheran	$1,000,000	Luis Avilan	$530,000
Trevor Cahill	$12,000,000	Chris Withrow	$522,500
Cameron Maybin	$7,000,000	Alex Wood	$520,000
Chris Johnson	$6,000,000	Shae Simmons	$508,750
Juan Uribe	$6,500,000	Mike Foltynewicz	$508,750
Jason Grilli	$4,250,000	Joey Terdoslavich	$507,500
Mike Minor	$5,600,000	Philip Gosselin	$507,500
Jonny Gomes	$4,000,000	Christian Bethancourt	$507,500
A.J. Pierzynski	$2,000,000	Jace Peterson	$507,500
Jim Johnson	$1,600,000	Todd Cunningham	$507,500
Kelly Johnson	$1,500,000		

(a) Construct a stemplot of the salaries.
(b) Describe the basic shape of the data. Are there any unusually small or large salaries?
(c) Find the median salary and a player who has this median salary.
(d) Can you explain why there are no salaries below $200,000?

2.26. (Exercise 2.25 continued). Collect from the internet the salaries of your favorite MLB in the current season.
(a) Answer questions (a), (b), (c) of Exercise 2.25 for this salary dataset.
(b) Compare the distributions of salaries of the 2014 Braves with your team.



2.27. (Exercise 2.25 continued) A dotplot of the salaries of the 2014 Atlanta Braves is shown above.
(a) Based on the shape of the distribution of these salaries, do you expect the mean to be larger than, equal to, or smaller than the median?
(b) Compute the mean and median of the salaries of the 2014 Atlanta Braves.
(c) Do the values you computed in (b) agree with your expectation in (a)?

2.28. In baseball, there are four famous (or notable) years with respect to home runs: 1927 when Babe Ruth hit 60 home runs, 1961 when Roger Maris hit 61, 1998 when Mark McGwire

hit 70 home runs, and 2001 when Barry Bonds hit 73 home runs. For the three years 1927, 1961, 1998, the number of home runs was collected for each player who had at least 400 at-bats. A player's home run count was classified into one of the six groupings, 0, 1–5, 6–10, 11–20, 21–40, 41 and more. The frequency table for the home runs for each of the four years is presented in Tables 2.17–2.20.

Table 2.17. Classification of home run numbers for 1927 players

Number of home runs	0	1–5	6–10	11–20	21–40	41 and more
Count	7	50	19	12	3	2
Proportion						

Table 2.18. Classification of home run numbers for 1961 players

Number of home runs	0	1–5	6–10	11–20	21–40	41 and more
Count	0	18	12	31	28	7
Proportion						

Table 2.19. Classification of home run numbers for 1998 players

Number of home runs	0	1–5	6–10	11–20	21–40	41 and more
Count	0	21	34	60	63	12
Proportion						

Table 2.20. Classification of home run numbers for 2001 players

Number of home runs	0	1–5	6–10	11–20	21–40	41 and more
Count	1	15	34	63	66	12
Proportion						

(a) For each table, find the proportion of players who fall in each of the six categories. Place the proportions in the rows labeled "Proportion."

(b) In 1927, what proportion of players hit more than 20 home runs? Repeat this computation for the 1961 and 1998 years. Place your answers in the table below.

YEAR	Proportion hitting more than 20 home runs
1927	
1961	
1998	
2001	

(c) Based on your computations in (a) and (b), would you say that in some years it was more difficult to hit a home run than in other years? Explain.

2.29. Table 2.21 gives average home ballpark attendance for each of the major league teams in 2014.

Table 2.21. Average attendance figures for all ML teams in 2014

American League		National League	
Team	Attendance	Team	Attendance
Los Angeles	38,221	Arizona	25,602
Baltimore	30,426	Atlanta	29,065
Boston	36,495	Chicago	32,742
Chicago	20,381	Cincinnati	30,576
Cleveland	17,746	Colorado	33,090
Detroit	36,015	Los Angeles	46,696
Houston	21,628	Miami	21,386
Kansas City	24,154	Milwaukee	34,536
Minnesota	27,785	New York	26,528
New York	41,995	Philadelphia	29,924
Oakland	24,736	Pittsburgh	30,155
Seattle	25,486	San Diego	27,103
Tampa Bay	17,858	San Francisco	41,589
Texas	33,565	St. Louis	43,712
Toronto	29,327	Washington	31,844

(a) Construct a dotplot of the home attendance averages for all AL teams. Comment on the basic shape of the data and note any unusual values.
(b) Repeat (a) for the home averages for the NL teams.
(c) Find the AL team that has the smallest average home attendance, the team with the largest average, and a team that has a median average.
(d) Repeat (c) for the NL teams.
(e) Can you explain the great variation in home attendance averages? Do you think the variation can be explained solely by the differences in the sizes of the cities?

2.30. (Ballpark attendance data from Exercise 2.29.)
(a) Find the five-number summary of the home average attendance for the AL teams.
(b) Find the five-number summary of the home average attendance for the NL teams.
(c) By comparing the medians you found in (a) and (b), which league tends to draw more fans at their home games?
(d) Which league seems to have more variation in home attendance among its teams? Which number should you be computing to measure the spread of the AL and NL datasets?

2.31. Table 2.22 gives the batting side (L = left, R = right, B = both sides) and the throwing hand (L = left, R = right) for 40 randomly selected ballplayers who were born in 1950 or later.
(a) Construct a frequency table for the batting side of the 40 players.
(b) Graph the proportions of left, right, and both sides batters using a bar chart.
(c) Repeat (a) and (b) for the throwing hand of the 40 players.

Table 2.22. Batting side and throwing hand for forty randomly selected players

Batting Side									
L	R	L	B	R	R	L	R	R	B
R	R	R	L	R	L	R	L	R	L
R	L	R	R	R	R	L	R	R	R
L	R	R	L	R	R	B	L	R	B
Throwing Hand									
R	R	R	R	R	R	R	R	L	L
R	R	R	L	R	R	R	R	R	R
R	L	R	R	R	R	R	R	R	R
R	L	R	R	L	L	R	R	R	R

(d) Do you think that the proportion of left-handed batters is less than, equal to, or greater than the proportion of lefties in the American population (which is about 10%)? Explain.
(e) Do you think that the proportion of left-handed throwers is less than, equal to, or greater than the proportion of lefties in the American population? Explain.

2.32. Table 2.23 displays the speed (in miles per hour) of 40 four-seam fastballs thrown by Zack Greinke during a game on July 19, 2015.

Table 2.23. Speeds of fastballs (in mph) thrown by Zack Greinke during a specific game during the 2015 season

92.3	94.3	93.0	93.4	93.4	94.1	94.4	93.3	93.5	93.5
94.3	93.7	92.7	94.1	93.0	92.4	92.0	92.5	92.9	91.3
92.9	92.6	92.4	92.9	91.8	91.6	93.4	93.0	93.5	93.0
91.9	93.5	92.4	93.6	94.0	92.2	92.9	93.3	93.2	93.3

(a) Construct a suitable graph of these pitch speeds.
(b) Find a five-number summary of these speeds.
(c) These speeds are listed in order of when they were thrown in the game—the first row contains the speeds of the first 10 pitches, the second row contains the speeds of the second 10, and so on. Construct a graph of the speeds against the order they were thrown—do you see any pattern?

2.33. Table 2.24 displays the pitch type of the first 40 pitches thrown by Zack Greinke during the game played on July 19, 2015. (FF is a four-seam fastball, SL is a slider, CH is a changeup, FT is a two-seam fastball, and CU is a curveball.)
(a) Construct a frequency table of the pitch types.
(b) Construct a suitable graph of these data.
(c) For each row of 10 pitches, compute the number of different pitch types that were thrown. What does this say about Greinke's pitching pattern during this game?

Table 2.24. Pitch types of the first 40 pitches thrown by Zack Greinke during a specific game during the 2015 season

FF	SL	FT	FF	FF	FF	CH	SL	FF	FF
FT	FF	CU	FT	CU	FF	CH	SL	FT	SL
FT	FF	CH	FF	FF	SL	FF	FT	FF	SL
FF	SL	FF	FF	SL	SL	FF	FT	CH	FT

2.34. Table 2.25 shows batted ball data for Albert Pujols for the seasons 2002 through 2015. The "Direction" statistics give the proportion of batted balls that were pulled, and hit in the center and opposite sides of the field. The "Hardness" statistics give the proportion of batted balls that were hit softly, medium hardness, or hard.

Table 2.25. Batted ball statistics for Albert Pujols for seasons 2002 through 2015

	Direction			Hardness		
Season	Pull	Center	Opposite	Soft	Medium	Hard
2002	0.474	0.265	0.261	0.126	0.601	0.273
2003	0.508	0.258	0.234	0.142	0.526	0.332
2004	0.490	0.277	0.233	0.097	0.552	0.352
2005	0.469	0.285	0.246	0.127	0.484	0.389
2006	0.453	0.340	0.207	0.078	0.580	0.342
2007	0.404	0.373	0.223	0.136	0.464	0.400
2008	0.475	0.318	0.207	0.155	0.416	0.429
2009	0.496	0.342	0.162	0.137	0.457	0.406
2010	0.507	0.348	0.145	0.161	0.416	0.424
2011	0.422	0.400	0.178	0.206	0.489	0.305
2012	0.520	0.317	0.164	0.142	0.523	0.335
2013	0.431	0.420	0.149	0.120	0.519	0.362
2014	0.510	0.298	0.193	0.145	0.494	0.361
2015	0.460	0.347	0.193	0.151	0.510	0.338

(a) Construct a suitable graph of the proportions of batted balls that were pulled for all seasons. Write a short paragraph describing the main features of this distribution.
(b) Construct a graph of the proportions of "hard" batted balls and summarize this distribution.
(c) Construct a graph of the proportion of "hard" batted balls against season. How has the proportion of hard batted balls changed over time?

Further Reading

Devore and Peck (2011) and Moore, McCabe and Craig (2012) provide good descriptions of the exploratory methods used in this chapter to describe a single batch of data. Chapter 2 of Albert and Bennett (2003) illustrates the use of data analysis methods on baseball data.

10.1090/text/034/03

3

Comparing Batches and Standardization

What's On-Deck?

In this chapter, we illustrate some basic data analysis tools for comparing two or more datasets. It is popular among baseball fans to make comparisons between individual players. In Case Study 3.1, we compare two popular sluggers in baseball, Albert Pujols and Manny Ramirez. By using parallel boxplots and time series plots, we compare the batting performance of the two players. In Case Study 3.2, we compare Robin Roberts and Whitey Ford, two Hall of Fame pitchers who pitched in the opening game of the 1950 World Series. When you think of great individual home run accomplishments, one naturally thinks of Babe Ruth, who hit 60 home runs in 1927, Roger Maris, who hit 61 in 1961, Mark McGwire, who hit 70 in 1998, and Barry Bonds, who hit 73 home runs in 2001. To better understand these hitting accomplishments, one should look at the general pattern of home run hitting during these four seasons, and Case Study 3.3 compares the team home run rates for these seasons. Case Study 3.4 looks at the slugging percentages of all players in the 2014 season that had at least 400 at-bats. We will see that the distribution of slugging percentages has a distinctive bell-shape and this pattern makes it easy to find intervals that contain a given percentage of the data. We conclude the chapter in Case Study 3.5 by looking at four of the highest season batting averages in recent baseball history. One can assess the greatness of each hitting accomplishment by looking at each batting average in the context of all batting averages for that particular season. By computing standardized scores, we can compare these hitting accomplishments and say which player had the best average relative to his peers.

3.1 Albert Pujols and Manny Ramirez

Topics Covered: Stemplot, five-number summary, time series plot, boxplot.

In this first case study, we statistically compare Albert Pujols and Manny Ramirez, two of the best sluggers in modern baseball history.

Let's make some initial comments about these two players:

- Both players were born in the Dominican Republic and are both known for their great batting skill and power.
- Ramirez debuted in Major League Baseball in 1993 and played parts of 19 seasons. Pujols debuted with the St. Louis Cardinals in 2001 and is currently (in 2015) in his 15th season playing for the Los Angeles Angels of Anaheim.

- Both players are especially proficient in hitting home runs. Ramirez hit 555 in his career and Pujols is approaching that career home run mark in the 2015 season.

Who is the Better Hitter?

Before we compare Pujols and Ramirez, some general truths about evaluating hitting should be stated.

1. The objective for a baseball team is to score RUNS.
2. How does a team score runs?
   - batters get ON-BASE
   - other batters ADVANCE the runners to home by hits, walks, HBP, or errors
3. A measure of a hitter's ability to get on-base is the On-Base Percentage (OBP):

$$\text{OBP} = \frac{\text{H} + \text{BB} + \text{HBP}}{\text{AB} + \text{BB} + \text{HBP} + \text{SF}}.$$

   The total number of plate appearances is PA = AB + BB + HBP + SF and the number of times on-base is H + BB + HBP. So OBP is the fraction of PAs that are on-base.
4. A measure of a hitter's ability to advance runners is the Slugging Percentage (SLG):

$$\text{SLG} = \frac{\text{TB}}{\text{AB}}.$$

   Here TB is total bases = # of Singles + 2 × (# of Doubles) + 3 × (# of Triples) + 4 × (# of Home Runs).
5. As discussed in Chapter 1, a simple measure that combines ON-BASE ability with ADVANCEMENT ability is the OPS statistic:

$$\text{OPS} = \text{OBP} + \text{SLG}.$$

   (For example, Ramirez in 2000 had an OBP of .457 and an SLG of .697, so his 2000 OPS was OPS = .457 + .697 = 1.154.)

Comparing OPS

Table 3.1 gives the season OPS values for Pujols and Ramirez for their careers (through the 2014 season).

We use back-to-back stemplots in Figure 3.1 to make a comparison—we break an OPS value like 1.106 into a stem of 11 and a leaf of 06. We use one-digit leaves, so the OPS value of 1.106 is represented by a 0 leaf on the 11 stem line.

We summarize each dataset by a five-number summary—this is

$$(LO, Q_L, M, Q_U, HI)$$

where LO, HI are the lowest and highest values in the dataset, Q_L, Q_U are the lower and upper quartiles, and M is the median.

We illustrate these calculations for Pujols' data. Here there are $n = 14$ values. The position of the median is $\text{pos}(M) = (14 + 1)/2 = 7.5$. The median is the average of the 7th and 8th values which is $(1.011 + 1.013)/2 = 1.012$. To find the quartiles, we divide the 14 remaining observations into two groups of 7, and find the median of the lower 7 and the median of the

Table 3.1. Career hitting statistics for Albert Pujols and Manny Ramirez

	Pujols			Ramirez		
Age	SLG	OBP	OPS	SLG	OBP	OPS
21	0.610	0.403	1.013			
22	0.561	0.394	0.955	0.521	0.357	0.878
23	0.667	0.439	1.106	0.558	0.402	0.960
24	0.657	0.415	1.072	0.582	0.399	0.981
25	0.609	0.430	1.039	0.538	0.415	0.953
26	0.671	0.431	1.102	0.599	0.377	0.976
27	0.568	0.429	0.997	0.663	0.442	1.105
28	0.653	0.462	1.114	0.697	0.457	1.154
29	0.658	0.443	1.101	0.609	0.405	1.014
30	0.596	0.414	1.011	0.647	0.450	1.097
31	0.541	0.366	0.906	0.587	0.427	1.014
32	0.516	0.343	0.859	0.613	0.397	1.010
33	0.437	0.330	0.767	0.594	0.388	0.982
34	0.466	0.324	0.790	0.619	0.439	1.058
35				0.493	0.388	0.881
36				0.601	0.430	1.031
37				0.531	0.418	0.949
38				0.460	0.409	0.870

```
1 | 2: represents 0.12, leaf unit: 0.01
  Pujols OPS    Ramirez OPS

                |  7* |
              96|  7. |
                |  8* |
               5|  8. |778
               0|  9* |4
              95|  9. |56788
             311| 10* |0113
               7| 10. |59
            1000| 11* |0
                | 11. |5
                | 12* |
```

Figure 3.1. Back-to-back stemplots for Albert Pujols and Manny Ramirez's season OPS values.

upper 7. We get

$$Q_L = .906, \quad Q_U = 1.101.$$

The smallest value is 0.767 and the largest is 1.114.

So Pujols' OPS values 5-number summary is $(0.767, 0.906, 1.012, 1.101, 1.114)$. In a similar fashion, we compute Ramirez's OPS values 5-number summary: $(0.870, 0.953, 0.982, 1.031, 1.154)$.

Essentially, a 5-number summary divides the data into quarters. For Pujols' data, we can say that (approximately) 25% of the data falls between 0.767 and 0.906, 25% falls between 0.906 and 1.012, and so on.

Before we graph these two datasets, we look for possible outliers. Using a standard rule, we say that an extreme observation is worthy of special attention if it falls outside one step from the lower and upper quartiles, where a step is defined to be

$$\text{STEP} = 1.5 \times (Q_U - Q_L).$$

To illustrate for Ramirez's data,

- $\text{STEP} = 1.5 \times (1.031 - 0.953) = 0.117$
- Outliers are observations that have values smaller than

$$Q_L - \text{STEP} = 0.953 - 0.117 = 0.836.$$

and larger than

$$Q_U + \text{STEP} = 1.031 + 0.117 = 1.148.$$

- Looking back at the stemplot of Ramirez's OPS values, we see that there is one outlier in this dataset at the high end.
- If one does a similar analysis for Pujols' OPS values, one finds that there are no outliers.

A boxplot is a graph of a five-number summary. One draws this by

- drawing a number line covering all of the values,
- drawing a box, where the ends of the box correspond to the quartiles Q_L and Q_U and a vertical line is drawn through the box at the median M,
- drawing lines (whiskers) out to the most extreme values that are not considered outliers,
- indicating outliers by plotting a special symbol.

Side-by-side boxplots are most useful in comparing datasets. In Figure 3.2 we show boxplots for the Pujols data and the Ramirez data on the same scale.

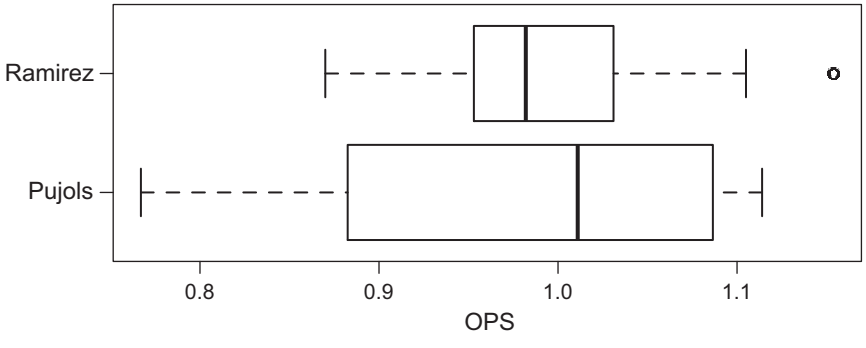


Figure 3.2. Parallel boxplots of the OPS season values for Albert Pujols and Manny Ramirez.

What have we learned in this comparision?

- On the average, Pujols is a slightly better hitter than Ramirez on the OPS scale. The median OPS for Pujols is 1.012 and the median OPS for Ramirez is 0.982, so Pujols is $1.012 - 0.982 = 0.030$ better using this measure.
- On the other hand, the spread or variability of Ramirez's OPS values is smaller than the spread of Pujols' OPS values. This is pretty obvious from the boxplot as the width of the box in the boxplot of Ramirez's OPS values is clearly smaller than the width of the box of the boxplot of Pujols' OPS values. One could conclude that Ramirez is a more consistent in the pattern of his hitting than Pujols.

Adjusting Comparison for Ages

But this comparison may be viewed as unfair, since Pujols is six years younger than Ramirez and has more baseball seasons left. Maybe the two players would be viewed differently after both have finished their careers.

Let's consider a different type of comparison that accounts for the ages of the two players.

What does it mean for Player A to have more ability than Player B? A popular way to model the relationship between ability and age is with the curve shown in Figure 3.3. In this curve, ability is low at the start of one's career, grows until mid-career, and then declines with advancing years. The peak ability, age at peak, the steepness of incline and decline, would be expected to vary from player-to-player.

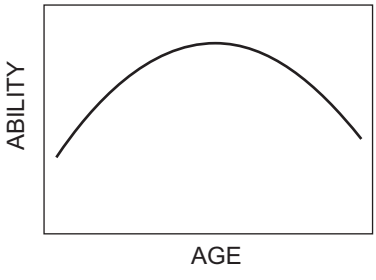


Figure 3.3. Expected ability curve for a Major League player.

Looking at the graph, we can distinguish between two measures of ability.

- **peak ability**—this is the player's ability at the top of the curve,
- **career ability**—this is the total accumulation of a player's ability over time—you can think of the area under the above curve as representing career performance.

So when you say Pujols is "better than" Ramirez, you should be clear what you're talking about. One player might play better at his peak, and the second player might be better in his total performance over his Major League career. We can look at Pujols and Ramirez's hitting performances by plotting the season OPS against age—Figure 3.4 graphs the OPS against age for both players.

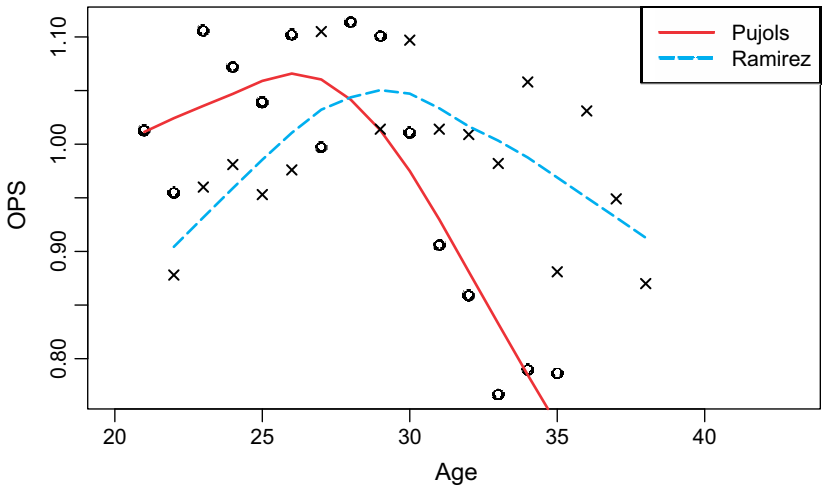


Figure 3.4. Time series plots of Albert Pujols' and Manny Ramirez's OPS values; the o's correspond to Pujols and the x's correspond to Ramirez. Smooth curves are drawn through the points to show the basic patterns.

Drawing separate smooth curves over the points to show the basic patterns, what do we see?

- It appears that Pujols peaked earlier than Ramirez—Pujols' peak in OPS is about at age 26 and Ramirez peaked at closer to 30 years.
- Ramirez's deterioration in batting performance was more gradual than Pujols. It is interesting that Pujols' drop-off in OPS occurred as he was changing teams from the Cardinals to the Angels.
- As I am writing, Pujols is having a much better 2015 hitting season, so perhaps it is premature to talk about Pujols' slump towards the end of his career.

3.2 Robin Roberts and Whitey Ford

Topics Covered: Stemplot, time series plot.

In this case study, we discuss two great pitchers, Robin Roberts and Whitey Ford. In Phillies history, there were four great seasons:

1. 1950 when the Whiz Kids won the NL pennant—they were a young team that won the pennant on the last game of the season. (They lost to the Yankees in the World Series in four games.)
2. 1980 when they won the World Series against the Royals.
3. 1993 when they won the NL pennant and lost in six games to Toronto. (Joe Carter who hit the series-ending home run comes to mind.)
4. 2008 when they won the World Series against the Rays.

We compare two pitchers, both in the Hall of Fame, who played in the 1950 World Series. Robin Roberts, born September 30, 1926, was a great pitcher for the 1950 Whiz Kids. He pitched for Philadelphia between 1948 and 1961. He won 286 games—many seasons he won

Table 3.2. Career pitching statistics for Robin Roberts

Year	Age	W	L	G	IP	H	HR	BB	SO	ERA	lgERA
1948	21	7	9	20	146.67	148	10	61	84	3.19	3.96
1949	22	15	15	43	226.67	229	15	75	95	3.69	3.96
1950	23	20	11	40	304.33	282	29	77	146	3.02	4.06
1951	24	21	15	44	315	284	20	64	127	3.03	3.84
1952	25	28	7	39	330	292	22	45	148	2.59	3.66
1953	26	23	16	44	346.67	324	30	61	198	2.75	4.20
1954	27	23	15	45	336.67	289	35	56	185	2.97	4.03
1955	28	23	14	41	305	292	41	53	160	3.28	3.96
1956	29	19	18	43	297.33	328	46	40	157	4.45	3.73
1957	30	10	22	39	249.67	246	40	43	128	4.07	3.80
1958	31	17	14	35	269.67	270	30	51	130	3.24	3.95
1959	32	15	17	35	257.33	267	34	35	137	4.27	4.11
1960	33	12	16	35	237.33	256	31	34	122	4.02	3.88
1961	34	1	10	26	117	154	19	23	54	5.85	4.07
1962	35	10	9	27	191.33	176	17	41	102	2.78	3.77
1963	36	14	13	35	251.33	230	35	40	124	3.33	3.52
1964	37	13	7	31	204	203	18	52	109	2.91	3.59
1965	38	10	9	30	190.67	171	18	30	97	2.78	3.42
1966	39	5	8	24	112	141	15	21	54	4.82	3.54

over 20 games. His pitching was notable for his great control. We will compare Roberts with a great Yankees pitcher, Whitey Ford. Ford, born October 21, 1928, is notable for having 10 World Series wins. He won the Cy Young award in 1961 with a record of 25-4. (He was overshadowed that year by Roger Maris who hit 61 home runs.) His career win/loss record was 236-106. Like Roberts, Ford was a young pitcher in 1950.

Comparing two pitchers is a little tougher than comparing two hitters since some of the standard pitching statistics are team-dependent. For example, if a pitcher has a great win/loss record such as 20-5, this could mean that he was an outstanding pitcher or it could mean that he was just a good pitcher whose team scored a lot of runs when he was pitching. Here we consider the ERA pitching statistic that is the mean number of runs earned by the opposing team in a 9-inning game. An "earned" run is one that is not produced by means of an error by the team's fielders. This is one of the most common measures that is used to rate pitchers. Tables 3.2 and 3.3 displays the season by season pitching statistics for both Roberts and Ford.

Figure 3.5 displays back-to-back stemplots of the ERAs for Ford and Roberts. In constructing the stemplot, we break an ERA such as 2.81 at the decimal point, so the stem is 2 and the (one-digit) leaf is 8.

What do we see in comparing Ford's and Roberts' ERAs?

Both pitchers had several seasons with ERAs in the 2.5–3.3 range. But Ford's worst season with respect to ERA was only 3.2 and he had five seasons where his ERA fell under 2.5. In contrast, Roberts' best season was 2.5 and he had six seasons where his ERA was over 4.0. Ford generally appears to be the better pitcher with respect to this pitching statistic.

Table 3.3. Career pitching statistics for Whitey Ford

Year	Age	W	L	G	IP	H	HR	BB	SO	ERA	lgERA
1950	21	9	1	20	112	87	7	52	59	2.81	4.30
1953	24	18	6	32	207	187	13	110	110	3	3.68
1954	25	16	8	34	210.67	170	10	101	125	2.82	3.42
1955	26	18	7	39	253.67	188	20	113	137	2.63	3.76
1956	27	19	6	31	225.67	187	13	84	141	2.47	3.87
1957	28	11	5	24	129.33	114	10	53	84	2.57	3.60
1958	29	14	7	30	219.33	174	14	62	145	2.01	3.54
1959	30	16	10	35	204	194	13	89	114	3.04	3.63
1960	31	12	9	33	192.67	168	15	65	85	3.08	3.60
1961	32	25	4	39	283	242	23	92	209	3.21	3.70
1962	33	17	8	38	257.67	243	22	69	160	2.90	3.73
1963	34	24	7	38	269.33	240	26	56	189	2.74	3.52
1964	35	17	6	39	244.67	212	10	57	172	2.13	3.62
1965	36	16	13	37	244.33	241	22	50	162	3.24	3.39
1966	37	2	5	22	73	79	8	24	43	2.47	3.33
1967	38	2	4	7	44	40	2	9	21	1.64	3.13

But we have to be careful about making this conclusion. Why? Well, the measurement of a pitcher's ability, such as an ERA, should be made in the context of the season and league and ballpark in which the player played. Ford and Roberts played during roughly the same seasons. But Ford pitched primarily in the American League, Roberts in the National League, and they pitched in different ballparks. Maybe Ford's ERAs looked better than Roberts' ERAs because of the differences in league and ballpark.

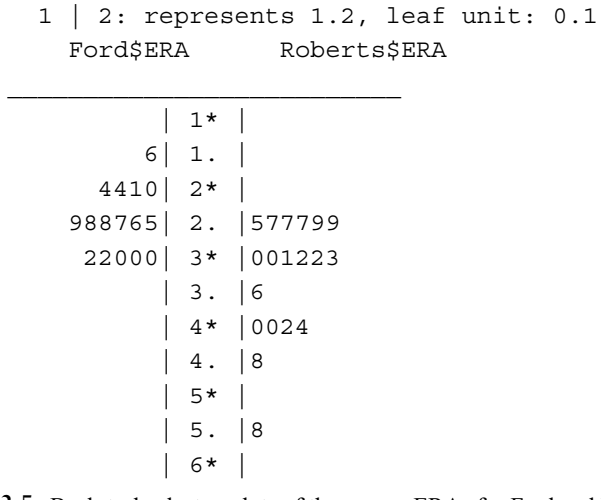


Figure 3.5. Back-to-back stemplots of the season ERAs for Ford and Roberts.

To illustrate how much ERAs can be different across seasons and leagues, Figure 3.6 displays a time series plot of the ERA for all pitchers. The season ERAs of the National League are plotted using a light line and the AL ERAs by a darker line. Note the great rises and falls in the mean ERA across seasons. In the "dead-ball" era in the early 1900s, pitching was dominant and the mean ERA was low. Then pitching got worse over time and the AL ERA peaked at a value close to 5.0 around the year 1940. In the late 1960s, pitching again was dominant—the league ERAs dipped to about 3.0—and recently the ERAs have been rising. Also, note that the NL and AL ERAs were quite different for some seasons. Since the season ERAs are so variable, it is reasonable to view a pitcher's season ERA in the context of the average ERA that particular year.

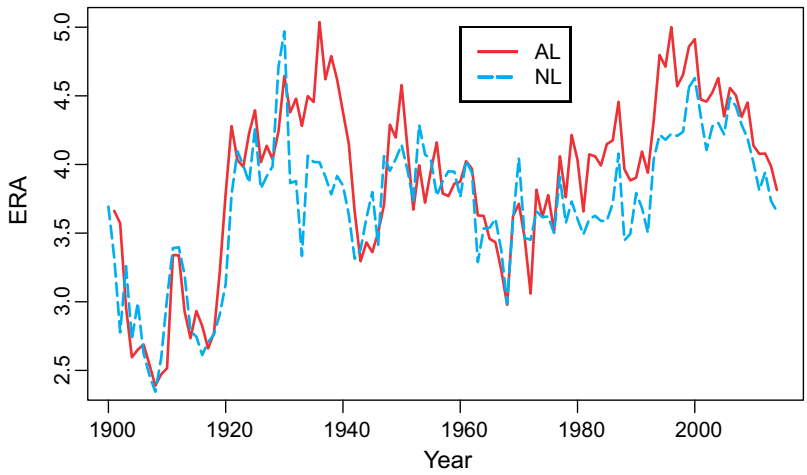


Figure 3.6. Time series plot of the season ERAs for all pitchers in the National and American Leagues.

Actually, the `baseball-reference.com` site goes one step further than a season adjustment. In the data tables shown above, the "lgERA" statistic is the average ERA of all pitchers in the same league and the same ballparks. We adjust a pitcher's ERA by computing

$$\text{ERA+} = \frac{\text{lgERA}}{\text{ERA}} \times 100,$$

which is called the "adjusted ERA".

To illustrate the computation and interpretation of ERA+, Robin Roberts' ERA in 1950 was 3.02. The average ERA in the National League in the ballparks that Roberts pitched was 4.06. So Roberts' adjusted ERA was

$$\text{ERA+} = \frac{4.06}{3.02} \times 100 = 135.$$

An adjusted ERA over 100 means that the pitcher performed above-average. Roberts had a good year in 1950—the league/ballpark average ERA was 35% higher than his ERA.

Suppose that we compare the adjusted ERAs of Roberts and Ford by stemplots in Figure 3.7.

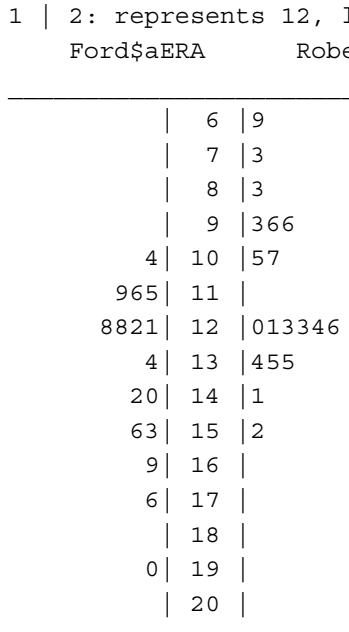


Figure 3.7. Back-to-back stemplots of the adjusted ERAs of Roberts and Ford.

We get some additional insight about the quality of these pitchers. Ford had an adjusted ERA of over 100 for every year that he pitched—that means that he was above average every year. In contrast, Roberts had 13 years where he was above-average and six years where he was below average. (It is interesting to note that five of his bad seasons occurred between 1956 and 1961.) Ford had three seasons where his adjusted ERA was 170 or higher. Ford does appear to be the superior pitcher in this study, but Ford pitched for a much stronger team (the Yankees) than Roberts (the Phillies), and it is possible that the difference in team strengths might explain part of the difference in ERAs.

3.3 Home Runs: A Comparison of Four Seasons

Topics Covered: Stemplot, five-number summary, time series plot, boxplot.

When a baseball fan thinks about home runs from a historical perspective, four seasons should come to mind: 1927, 1961, 1998, and 2001. The year 1927 was the year in which Babe Ruth hit 60 home runs and broke his own season record for home runs. This record stood for 34 years until 1961, when Roger Maris hit his 61 home runs. Another 37 years went by until Mark McGwire broke the season record with 70 home runs. It was generally thought that McGwire's record would last for a while, but Barry Bonds broke the record in 2001 (only three years later) with his great feat of 73 home runs.

Is Bonds really the greatest home run slugger of all time? Well, it is not clear. It is difficult to compare Bonds' accomplishment in 2001 with Ruth's accomplishment in 1927 since the difficulty of hitting a home run may have been different in the two years. Maybe we can better understand the magnitude of these players' accomplishments if we compare the home run hitting of all players for these four seasons.

Table 3.4. Number of home runs (HR) hit by every team in the 1927, 1961, 1998, and 2001 seasons

Year											
1927			1961			1998			2001		
HR	G	Rate	HR	G	Rate	HR	G	Rate	HR	G	Rate
158	155	1.02	240	163	1.47	198	162	1.22	212	162	1.31
56	155	0.36	180	163	1.10	198	163	1.21	164	162	1.01
29	157	0.18	149	163	0.91	134	161	0.83	214	162	1.32
51	156	0.33	138	163	0.85	115	162	0.71	139	162	0.86
36	153	0.24	150	161	0.93	165	162	1.02	152	162	0.94
26	153	0.17	112	163	0.69	207	162	1.28	203	161	1.26
55	155	0.35	167	161	1.04	205	162	1.27	198	161	1.23
28	154	0.18	189	162	1.17	221	163	1.36	195	162	1.20
54	156	0.35	90	162	0.56	214	162	1.32	136	162	0.84
84	153	0.55	119	161	0.74	111	162	0.69	121	162	0.75
109	155	0.70	158	154	1.03	201	162	1.24	169	162	1.04
74	153	0.48	157	154	1.02	147	162	0.91	199	162	1.23
29	153	0.19	183	155	1.18	234	161	1.45	158	162	0.98
39	154	0.25	188	155	1.21	149	162	0.92	246	162	1.52
37	155	0.24	103	155	0.66	166	162	1.02	208	162	1.28
57	155	0.37	128	154	0.83	212	163	1.30	199	162	1.23
			176	156	1.13	223	163	1.37	194	162	1.20
			103	155	0.66	138	162	0.85	209	162	1.29
						152	162	0.94	176	162	1.09
						107	163	0.66	161	162	0.99
						215	162	1.33	174	162	1.07
						136	162	0.84	164	162	1.01
						126	162	0.78	147	162	0.91
						147	162	0.91	166	162	1.02
						114	162	0.70	131	162	0.81
						167	162	1.03	208	162	1.28
						161	163	0.99	235	162	1.45
						159	162	0.98	206	162	1.27
						183	162	1.13	161	162	0.99
						159	162	0.98	213	162	1.31

Table 3.4 gives the number of home runs (HR) hit by every team in the 1927, 1961, 1998, and 2001 baseball seasons. The table also gives the number of games (G) for each team each year. Note that the number of games in a season has changed over the years—the 1927 teams and the 1961 teams had season lengths that were about 7–9 games shorter than the current seasons. To make a reasonable comparison, it is helpful to compute the home run rate (RATE)—the number of home runs hit divided by the number of games (G).

This home run rate is also given in the table. If a team's RATE $= 1$, then it hits, on average, one home run each game.

One effective way of graphically comparing the team home run rates across seasons is by means of parallel dotplots shown in Figure 3.8. What do we see when we compare team home run rates across these four seasons?

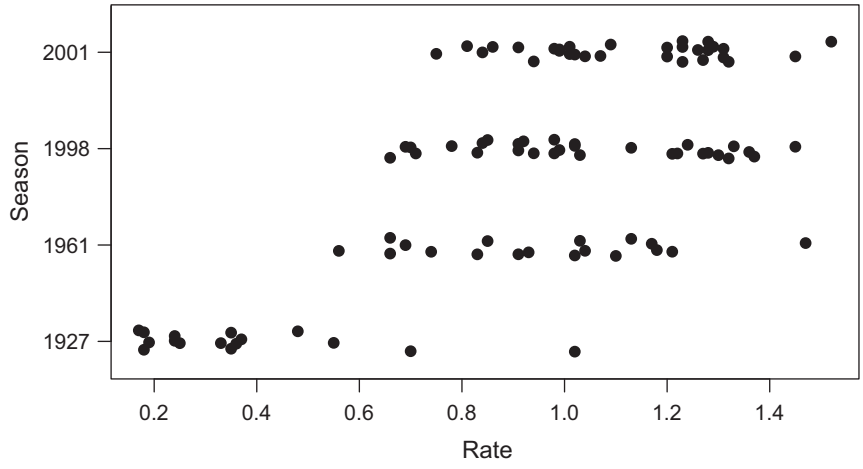


Figure 3.8. Parallel dotplots of the team home run rates for the four seasons.

- The pattern of home run hitting in 1927 is very interesting. Babe Ruth's team, the Yankees, averaged about a home run every game. But the Yankees were an outlier. A typical team in 1927 had a home run rate about .3—this average team hit a home run every three games. It is pretty clear that Ruth's 60 home run total was quite an accomplishment in the year 1927 where the rate of hitting home runs was small.
- The year 1961 was a sharp contrast to 1927 with respect to home run hitting. Most of the team home run rates are in the .8–1.1 range which was comparable to the home run rate of the great 1927 Yankees. It was much more common to hit a home run in 1961 when Roger Maris hit his 61 roundtrippers.
- We see that the years 1998 and 2001 saw even greater home run hitting than 1961. Looking at the dotplots in Figure 3.8, it appears that a typical home run rate has increased from 1961 to 1998 and from 1998 to 2001. Half of the 2001 baseball teams had home run rates of 1.19 or higher.

To get a better handle on the differences in home run hitting between seasons, we can compute summary statistics. Table 3.5 gives five-number summaries for each batch of team home run rates. Figure 3.9 draws parallel boxplots using these summaries.

Table 3.5. Five-number summaries of the team home rates for the four seasons

Year	N	LO	Q_L	Median	Q_U	HI
1927	16	0.16	0.20	0.34	0.45	1.01
1961	18	0.55	0.73	0.98	1.14	1.47
1998	30	0.65	0.85	1	1.27	1.45
2001	30	0.74	0.99	1.14	1.28	1.51

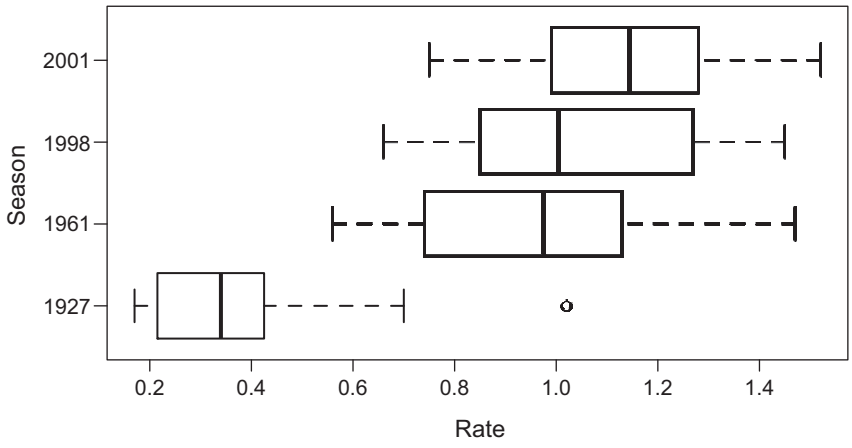


Figure 3.9. Parallel boxplots of the team home run rates for the four seasons.

This display reinforces the conclusions that were drawn earlier. Home runs were generally not hit in 1927 and there was a small spread in the team home run rates. The dot in the 1927 boxplot indicates the outlying team, the Yankees. From 1961 to 2001 there has been a gradual increase in the rate of hitting home runs. Comparing the median home run rate (.98) in 1961 with the median rate (1.14) in 2001, we see that a typical team in 2001 hit approximately $1.14 - .98 = .16$ more home runs than a typical team in 1961—that translates to about an additional home run each six games.

3.4 Slugging Percentages are Normal

Topics Covered: Stemplot, mean and standard deviation, normal distribution probabilities.

The standard deviation, denoted s, is a measure of dispersion about the center (mean) and has a useful interpretation when our data is bell-shaped. We illustrate the use of the empirical rule or "68-95-99.7 rule" that tells us about the fraction of observations that fall within three intervals. Let's look at all hitters in 2014 that had at least 400 at-bats—we can regard these hitters as the "regular" players, since they played most of the games during the season. Figure 3.10 displays a stemplot of the slugging percentages (SLG) for the 177 regular players. In this stemplot, a slugging percentage of .402 is divided between the first and second digits, so 4 is the stem and 0 is the leaf. Also the 10 possible leaf values is divided into five parts—leaves of 0 and 1 appear next to the "*" symbol, leaves of 2 and 3 appear next to the "t" symbol, the leaves of 4 and 5 appear next to "f", leaves of 6 and 7 appear next to "s", and leaves of 8 and 9 appear next to ".".

This set of slugging percentages looks pretty symmetric; most of the values are in the .350 to .450 range and there is a large spread—the best hitter in 2014 had a slugging percentage of .581 and the worst was only about .280. (Who are the players with the best and worst SLG values in 2014? It might be easier to guess the largest value rather than the smallest.) We use R to compute the mean and standard deviation for these SLGs. Rounding, we see that

$$\bar{x} = 0.416 \quad \text{and} \quad s = 0.060.$$

```
1 | 2: represents 0.12
leaf unit: 0.01
n: 177

 2. | 8
 3* | 0111
 3t | 222333333
 3f | 44444555555
 3s | 66666666666677777777777777777
 3. | 88888888888888888999999999999
 4* | 00000000000000111111111111
 4t | 22222333333
 4f | 44444444555555555555555
 4s | 6666677777
 4. | 888899999
 5* | 0001
 5t | 2222
 5f | 44455
 5s | 666
 5. | 8
```

Figure 3.10. Stemplot of slugging percentages for the 177 regular players in the 2014 season.

When the distribution of the data is bell-shaped (as it is here), we expect

- 68% of the data to fall within one standard deviation of the mean (that is, between $\bar{x} - s$ and $\bar{x} + s$),
- 95% of the data to fall within 2 standard deviations of the mean (between $\bar{x} - 2s$, $\bar{x} + 2s$),
- 99.7% of the data to fall within 3 standard deviations of the mean (between $\bar{x} - 3s$, $\bar{x} + 3s$).

Applying this 68-95-99.7 rule, we compute

$$\begin{aligned} \bar{x} - s \quad &\text{and} \quad \bar{x} + s \\ .416 - .060 \quad &\text{and} \quad .416 + .060 \\ .356 \quad &\text{and} \quad .476. \end{aligned}$$

We expect 68% of the SLGs to fall between .356 and .476.

Let's check how many actually fall between .356 and .476: there are 128 SLGs between .356 and .476 (using the original dataset) and the proportion of SLGs in this interval is $128/177 = .723$, which is pretty close to .68.

Next, we compute

$$\begin{aligned} \bar{x} - 2 \times s \quad &\text{and} \quad \bar{x} + 2 \times s \\ .416 - 2(.060) \quad &\text{and} \quad .416 + 2(.060) \\ .296 \quad &\text{and} \quad .536. \end{aligned}$$

We expect 95% of the SLGs to fall between .296 and .536. Checking the data again, we obtain 167 (out of 177) in this interval which corresponds to $167/177 = .943$, which is very close to .95.

Last, we compute

$$\begin{aligned} \bar{x} - 3 \times s \quad &\text{and} \quad \bar{x} + 3 \times s \\ .416 - 3(.060) \quad &\text{and} \quad .416 + .3(.060) \\ .236 \quad &\text{and} \quad .596. \end{aligned}$$

We expect practically all (99.7%) of the SLGs to fall between .236 and .596. In fact all fall in this range which is $177/177 = 1$ which is very close to .997, so again this rule seems to work.

Generally, most "derived" baseball statistics that are computed by a division or a multiplication will be bell-shaped and well-suited for applying the 68-95-99.7 rule. Examples of such derived statistics are AVG, SLG, OBP, and ERA.

3.5 Great Batting Averages

Topics Covered: Stemplot, mean and standard deviation, standardized score. In this case study, we examine Great Batting Averages. We identify four players who had unusually high batting averages (AVG) in recent history:

- 1994—Tony Gwynn (SD) hit .394
- 1980—George Brett (KC) hit .390
- 1977—Rod Carew (MIN) hit .388
- 1941—Ted Williams (BOS) hit .406

Now, on face value, any baseball statistics like these high averages are meaningless because we don't know the context in which these statistics were achieved. If I tell you that Joe Schmoe (a hypothetical player) batted .420, you should ask:

- When? You have to know when Joe Schmoe got this AVG. For example, batting averages in 1900 were generally much higher than batting averages in 1968.
- Where? It matters where Joe played. Some parks (like Coors Field) are easier to hit in, and others (like Dodger Stadium) are harder to hit in.
- How many at-bats? We'll talk about this in a later chapter, but it is easier to hit .420 based on 100 AB than 500 AB.

We can make better sense of a batting average, like Carew's .388 AVG in 1977, when we see it compared to the AVGs of all players in the year 1977. Let's focus on the 1977 hitters who had at least 400 at-bats. Figure 3.11 displays a stemplot of the 168 AVGs.

Here are some features of this dataset. (1) The distribution looks symmetric and bell-shaped. (2) A typical AVG is about .275. (3) Most of the averages fall between .230 and .300. (4) We can't miss the one large AVG at .388 which corresponds to Carew. We would like a measure of relative standing that tells us how good Carew's batting average is. We compute a standardized score of an average by subtracting the mean and then dividing by the standard deviation s. Calculations reveal that $\bar{x} = .27742$ and $s = .02717$ for these 1977 AVGs. So the standardized

```
1 | 2: represents 0.012
 leaf unit: 0.001
            n: 173

        20 | 06
        21 | 5
        22 | 018
        23 | 012335899
        24 | 01344666779
        25 | 0001111223345678899
        26 | 0111122333333456667788
        27 | 000111233344455557777899 99
        28 | 00011111222233444566677788999
        29 | 0001122235677888
        30 | 000022344777888
        31 | 00112556779
        32 | 02558
        33 | 557
        34 |
        35 |
        36 |
        37 |
        38 | 7
```

Figure 3.11. Stemplot of the batting averages of the 1977 Major League players with at least 400 at-bats.

score (or z score) for Carew's .388 is

$$z = \frac{.388 - .27742}{.02717} = 4.07 \quad \text{(WOW!)}.$$

Carew's z-score is 4.07 which means that his AVG was about four standard deviations above the mean. Let's compute the standardized score for the weakest hitter in 1977 who only batted .200. His z-score would be

$$z = \frac{.200 - .27742}{.02717} = -2.85.$$

In other words, his AVG was 2.85 standard deviations below the mean. If a player has a z-score of 0, this means that his AVG is equal to the mean 277. So

- A positive z-score corresponds to an AVG above the mean.
- A negative z-score corresponds to an AVG below the mean.
- A z-score of 0 corresponds to an AVG that is equal to the mean.

Recall, for bell-shaped data, practically all of the z-scores for the data will fall between -3 and 3. (This is a restatement of the 99.7 rule.) Using the concept of standardized scores, we can compare these great batting averages:

- We already saw that Carew's .388 corresponded to a z-score of 4.07.
- Tony Gwynn's .394—that year (1994), the mean $\bar{x} = .29311$ and $s = .03201$, so his z-score would be

$$z = \frac{.394 - .29311}{.03201} = 3.15.$$

- Ted Williams' .406—that year (1941), $\bar{x} = .28063$ and $s = .03281$, and

$$z = \frac{.406 - .28063}{.03281} = 3.82.$$

- George Brett's .390—that year (1980), $\bar{x} = .27882$ and $s = .02757$, and

$$z = \frac{.390 - .27882}{.02757} = 4.03.$$

Comparing the four, Carew's AVG was best (relative to his peers) since he had the largest standardized score. In this comparison, we see that batting averages have changed over the years. Table 3.6 shows the mean and standard deviation for five years.

Table 3.6. Mean and standard deviation of players' batting averages with at least 400 at-bats for five years

Year	Mean	Standard Deviation
1900	.295	.037
1941	.280	.033
1977	.277	.027
1980	.279	.028
1994	.293	.032

A couple of general comments from this table:

- Looking at the means, there doesn't seem to be any general trend in batting averages—they have gone up and down. Higher batting averages could mean better hitting or weaker pitching, or both.
- If you look at the standard deviations over years, you will see a general decreasing pattern. The standard deviation of AVGs today is smaller than it used to be. This means that the hitting abilities of current players are more similar than they used to be.

3.6 Exercises

3.0a. Rickey Henderson and Tim Raines were both great leadoff hitters who played in the 1980's and 1990's. Table 3.7 shows the career on-base percentage (OBP) and the slugging percentage (SLG) for each player—the seasons where Raines had fewer than 200 at-bats have been removed.

(a) Construct back-to-back stemplots of the OBPs for Henderson and Raines.
(b) Find five-number summaries of the two batches.
(c) Based on your work in (a) and (b), which hitter was more effective in getting on-base? On average, how much better was the superior player?

Table 3.7. On-base and slugging percentages for Rickey Henderson and Tim Raines for the seasons of their careers

Henderson			Raines		
Age	OBP	SLG	Age	OBP	SLG
20	.338	.336			
21	.420	.399	21	.391	.438
22	.408	.437	22	.353	.369
23	.398	.382	23	.393	.429
24	.414	.421	24	.393	.437
25	.399	.458	25	.405	.475
26	.419	.516	26	.413	.476
27	.358	.469	27	.429	.526
28	.423	.497	28	.350	.431
29	.394	.399	29	.395	.418
30	.411	.399	30	.379	.392
31	.439	.577	31	.359	.345
32	.400	.423	32	.380	.405
33	.426	.457	33	.401	.480
34	.432	.474	34	.365	.409
35	.411	.365	35	.374	.422
36	.407	.447	36	.383	.468
37	.410	.344	37	.403	.454
38	.400	.342	38	.395	.383
39	.376	.347			
40	.423	.466			
41	.368	.305			
42	.366	.351			

(d) Construct one time series plot of the slugging percentages of Henderson and Raines graphing against age. Compare the patterns of change (across time) for each player. Was one player consistently better than the other with respect to SLG? Did both players display the typical pattern in which one gets better in performance, hits a peak, and then declines towards the end of his career?

3.0b. If you look at Rickey Henderson's batting statistics in Table 1.1, we see that his highest season OBP was .439 in the 1990 season. How impressive was that .439 OBP? To check, Figure 3.12 displays a stemplot of the OBPs of all players that had at least 400 plate appearances in the 1990 season.

(a) Describe the general shape of the distribution of OBPs and discuss any unusual features.

(b) The mean and standard deviation of the OBPs are given by $\bar{x} = .337$ and $s = .036$, respectively. Find the standardized score of Henderson's OBP of .439.

(c) Find an interval that contains approximately 95% of the OBPs.

```
25 | 88
26 | 7
27 | 466999
28 | 2222589
29 | 0012236799
30 | 00112334444445557778
31 | 223466889
32 | 000122234466677888999
33 | 000002233345666778899
34 | 0001112222234677888899
35 | 001223466777899
36 | 00444466679
37 | 012345555566789
38 | 13455679
39 | 77
40 | 0457
41 | 377
42 |
43 | 9
44 | 1
```

Figure 3.12. Stemplot of the OBPs of all players with at least 400 plate appearances in 1990.

(d) How does Henderson's .439 OBP compare with the "greatness" of the batting averages discussed in Case Study 3.5. Is this study sufficient to convince you that Henderson was the greatest leadoff hitter of all time? Explain.

3.1. (Comparing batting averages across eras). The stemplots in Figure 3.13 display the batting averages of all players with at least 400 at-bats in the years 1897 and 1997:

(a) Find the five-number summaries of each dataset

(b) Which group of players tends to have the higher batting averages? Explain.

(c) Which group of players had the greater spread in batting averages? Explain.

3.2. (Continuation of Exercise 3.1.)

(a) For the 1897 batting averages, one can compute the mean and standard deviation to be .318 and .037, respectively. Find an interval of values where you expect 68% of the batting averages to fall.

(b) Count the number of players who had a batting average in the interval in (a).

(c) Find the proportion of players who had an average in the interval in (a). Is it close to what you expect?

(d) For the 1997 batting averages, the mean $= .281$ and the standard deviation $= .028$. Find an interval where you think 95% of the averages will fall.

(e) Find the proportion of 1997 players who had an average in the interval you found in (d). Check if it is close to what you expect.

```
1897  players                    1997 players

            0 | 20 | 2
              | 21 |
              | 22 | 389
              | 23 | 35799
              | 24 | 012334567889
              | 25 | 000000222234678
           76 | 26 | 01112333445555557789999
           73 | 27 | 0112333345555677789
   9997554322 | 28 | 00011223456666667788
          210 | 29 | 000001111234444556677888999
       876520 | 30 | 0223445678888
      8865320 | 31 | 01344889
      9998521 | 32 | 1337789
           75 | 33 | 022
        53000 | 34 | 7
        54322 | 35 |
           21 | 36 | 16
            7 | 37 | 1
           92 | 38 |
              | 39 |
              | 40 |
              | 41 |
            3 | 42 |
```

Figure 3.13. Stemplots of batting averages of all players with at least 400 at-bats from 1897 and 1997 seasons.

3.3. (Continuation of Exercise 3.1)

(a) In 1897, Willie Keeler won the batting crown with an average of .424. Using the mean and standard deviation of batting averages given in the previous exercise, find the standardized score for Keeler's average.

(b) In 1997, Tony Gwynn won the batting crown with an average of .372. Find the standardized score for Gwynn's average.

(c) Based on your computations in (a) and (b), which batter was better relative to his contemporaries? Why?

(d) In the book *Full House*, Stephen Jay Gould argues that ballplayers are generally getting less variable over time and it will be difficult for a player to hit for a .400 season batting average in the future. On the basis of your computations, do you agree with Gould's assertion? Why?

3.4. (Comparison of hit profiles) Table 3.8 gives the number of at-bats, hits, doubles, triples and home runs for the players during the 1900, 1945 and 1990 baseball seasons.

(a) For each year, compute the number of singles and put the numbers in the 1B column.

Table 3.8. Batting statistics for all players in the 1900, 1945 and 1990 baseball seasons

YEAR	AB	H	1B	2B	3B	HR
1900	39132	10925		1432	607	254
1945	84447	21977		3497	728	1007
1990	142768	36817		6526	865	3317

(b) For each year, compute the proportion of hits that are doubles. (For 1900, for example, you would divide the number of doubles 1432 by the number of hits 10925.) Put your answers in the table below.

	Proportion of hits that are			
YEAR	1B	2B	3B	HR
1900				
1945				
1990				

(c) For each year, compute the proportion of hits that are singles, proportion of hits that are triples, and the proportion of hits that are home runs.

(d) Based on the proportions you computed in (b) and (c), comment on how the game of baseball has changed from 1900 to 1990. Would you say that the game is more or less exciting than the games in the past? Why?

3.5. (Comparisons of rates of strikeouts and walks.) Table 3.9 gives the number of at-bats, strikeouts, and walks for the three baseball seasons 1900, 1945, and 1990.

Table 3.9. Strikeout and walk statistics for all players in the 1900, 1945 and 1990 baseball seasons

YEAR	AB	BB	SO
1900	39132	3034	2697
1945	84447	8295	8050
1990	142768	13852	24390

(a) For each season, compute the number of plate appearances by adding the at-bats to the walks. Put your answers in the PA column.

	Proportion of		
YEAR	PA	BB	SO
1900			
1945			
1990			

(b) For each year, compute the proportion of PAs that are strikeouts.

(c) For each year, compute the proportion of PAs that are walks.

(d) Explain how pitching has changed from 1900 to 1990 by looking at the proportions you computed in (b) and (c). Have pitchers become more or less dominating over the years? Do pitchers today have more or less control?

3.6. (Pujols vs Ramirez, continued.) Refer back to the Case Study 3.1 comparing the hitting of Albert Pujols and Manny Ramirez.

(a) Construct back-to-back stemplots of the season on-base percentages (OBP) for Pujols and Ramirez.

(b) Compare the two datasets. Which hitter generally has a higher OBP? Which hitter has the larger variation in OBP from season to season?

(c) Repeat (a) and (b) using the season slugging percentages (SLG).

(d) Based on your work, who is the more valuable hitter, Pujols or Ramirez? Explain why.

(e) Although Pujols may have better hitting statistics than Ramirez, some people still claim that Ramirez is the more valuable player? Why might they say that?

3.7. Table 3.10 gives the number of wins (W), losses (L), winning proportion (PCT), and Earned Run Average (ERA) for two great modern pitchers, Greg Maddux and Tom Glavine.

Table 3.10. Career pitching statistics for Greg Maddux and Tom Glavine

	Greg Maddux				Tom Glavine			
YEAR	W	L	PCT	ERA	W	L	PCT	ERA
1986	2	4	0.333	5.52				
1987	6	14	0.300	5.61	2	4	0.333	5.54
1988	18	8	0.692	3.18	7	17	0.291	4.56
1989	19	12	0.612	2.95	14	8	0.636	3.68
1990	15	15	0.500	3.46	10	12	0.454	4.28
1991	15	11	0.576	3.35	20	11	0.645	2.55
1992	20	11	0.645	2.18	20	8	0.714	2.76
1993	20	10	0.666	2.36	22	6	0.785	3.2
1994	16	6	0.727	1.56	13	9	0.590	3.97
1995	19	2	0.904	1.63	16	7	0.695	3.08
1996	15	11	0.576	2.72	15	10	0.600	2.98
1997	19	4	0.826	2.2	14	7	0.666	2.96
1998	18	9	0.666	2.22	20	6	0.769	2.47
1999	19	9	0.678	3.57	14	11	0.560	4.12
2000	19	9	0.678	3	21	9	0.700	3.40
2001	17	11	0.607	3.05	16	7	0.696	3.57

(a) Using back-to-back stemplots, compare the season winning percentages of Maddux and Glavine.

(b) Based on your comparison, which pitcher tended to win a greater percentage of games?

(c) Compare the season ERAs of Maddux and Glavine using stemplots.

(d) Which pitcher tended to have a lower season ERA?

3.8. Table 3.11 shows the slugging percentage (SLG) for the "regular" shortstops and "regular" 3rd basemen in 2014.

Table 3.11. Slugging percentages for regular shortstops and 3rd basemen in 2014

Shortstops		3rd Basemen	
Name	SLG	Name	SLG
Andrus	0.333	Beltre	0.492
Aybar	0.379	Carpenter	0.375
Castro	0.438	Castellanos	0.394
Cozart	0.300	Dominguez	0.330
Crawford	0.389	Donaldson	0.456
Desmond	0.430	Johnson	0.361
Escobar	0.377	Longoria	0.404
Escobar	0.340	McGehee	0.357
Hardy	0.372	Moustakas	0.361
Hechavarria	0.356	Rendon	0.473
Mercer	0.387	Sandoval	0.415
Peralta	0.443	Seager	0.454
Ramirez	0.408	Wright	0.374
Reyes	0.398		
Rollins	0.394		
Segura	0.326		
Simmons	0.331		

(a) Construct parallel boxplots of the slugging percentages of the shortstops and the 3rd basemen.
(b) Based on your work in (a), which type of player is more likely to be a slugger?
(c) On the average, how much superior is one group over the other group with respect to slugging percentage?
(d) Is the general pattern you found in (b) and (c) true for all players? Find some players that don't agree with the general pattern.

3.9. (Triple rates for 1899 and 1999 players.) Old-timers would argue that there is less use of speed in baseball today than in the old days. One indication of the lack of speed in baseball today is the relatively small number of triples hit. (There are other explanations for the small number of triples, including the liveliness of the ball and the change in ball park design.) For each of the "regular" players (with at least 400 at-bats) in 1899 and 1999, we compute the TRIPLE RATE $= 3B/AB$. Stemplots of the triple rates for each group of players are shown in Figure 3.14. For the stemplots, the break point is between the hundredths and thousandths places, so for a triple rate of .048, the stem is 4 and the leaf is 8.

(a) Compute five-number summaries of each dataset and graph parallel boxplots.
(b) Which group of players tended to get a higher rate of triples? What is the difference between the two average triple rates?
(c) Does your comparison support the claim that there is less speed in baseball nowadays? What other variables could you use to measure speed of ballplayers?

```
1899 players
0 | 114
0 | 6667999
1 | 00011222233334444
1 | 55555666788889999
2 | 01122233334
2 | 68
3 | 0012
3 | 7
4 | 23
1999 players
0 | 00000000000000000000000000001111111111111111111
0 | 22222222222333333333333333
0 | 444444444444445555555
0 | 666666666666677777777
0 | 8888888888888889999999
1 | 0000000000111
1 | 222222333
1 | 44444455
1 | 666
1 | 889
2 | 0
```

Figure 3.14. Stemplots of triple rates of players in 1899 and 1999 seasons with at least 400 at-bats.

3.10. Table 3.12 displays the total number of home runs (HR) hit by each of the 30 major league teams in 2014.

Table 3.12. Home run numbers of all Major League teams in 2014

Team	HR	Team	HR
BAL	211	ARI	118
BOS	123	ATL	123
CHA	155	CHN	157
CLE	142	CIN	131
DET	155	COL	186
HOU	163	LAN	134
KCA	95	MIA	122
LAA	155	MIL	150
MIN	128	NYN	125
NYA	147	PHI	125
OAK	146	PIT	156
SEA	136	SDN	109
TBA	117	SFN	132
TEX	111	SLN	105
TOR	177	WAS	152

(a) Using the stems shown below, construct back-to-back stemplots of the HR totals of the American League and the National League.

```
American League        National League

                              10
                              11
                              12
                              13
                              14
                              15
                              16
                              17
                              18
                              19
                              20
                              21
                              22
```

(b) Find five-number summaries of the two datasets.

(c) Using your work in (a) and (b), compare the two datasets. Did one league tend to hit more home runs than the other league? Which teams hit an unusually small or large number of home runs in 2014?

3.11. (Comparing two hitters.) Find two contemporary hitters who have each played ten or more major league seasons. Compare the two hitters on the basis of season data using one batting statistic such as AVG, SLG, HR, or OBP. You should (1) list the season data for both players, (2) compare the datasets using appropriate graphs, (3) compute summaries of both datasets, and (4) describe what you've learned from this comparison.

3.12. (Comparing two pitchers.) Find two contemporary pitchers who have each played ten or more major league seasons. Compare the two pitchers on the basis of season data using one pitching statistic such as ERA, PCT, SO, or BB. You should (1) list the season data for both players, (2) compare the datasets using an appropriate graph, (3) compute summaries of both datasets, and (4) describe what you've learned from this comparison.

3.13. (Comparing two teams.) Find two teams that you are interested in comparing with respect to some batting statistic, such as AVG, SLG, OBP, or HR. For each of the two teams, list all of the regular players on the team for a particular season (the ones with at least 400 at-bats) and the value of the batting statistic for each player. Compare the two batches of data using an appropriate graph, compute summaries of both batches, and describe which team appears to be superior from the viewpoint of this batting statistic.

3.14. (Comparing NL and AL teams.) Compare the NL and AL teams for any season with respect to a particular hitting, pitching, or fielding statistic. List the 30 teams and the value of the statistic for each team. Compare the NL and AL batches using a graph, compute statistics, and describe which league appears to be superior with respect to this statistic.

3.15. (Comparing players from different eras.) Compare baseball hitters or pitchers from two different years. Choose two years of interest (not too close together) and one baseball statistic that you are interested in. Randomly pick at least 30 "regular" players from each year, and list the name and the baseball statistic for each player. Compare the two years of data using an appropriate graph, compute summary statistics, and describe what you have learned in this comparison.

3.16. There has been much talk about the recent surge in home run hitting. How has the rate of home run hitting changed since 1990? From a baseball web site, find the total number of home runs and at-bats for each of the 30 teams for the current year and compute the home run rate for the teams. Find the home run rates for all the 1990 teams. (This data is available on the book website described in Appendix 2.) Comparing the 1990 and current datasets, has there been a significant change in the rate of home run hitting?

3.17. Figure 3.15 displays parallel boxplots of the speeds of four-seam fastballs thrown by Clayton Kershaw and Zack Greinke during the 2015 season.

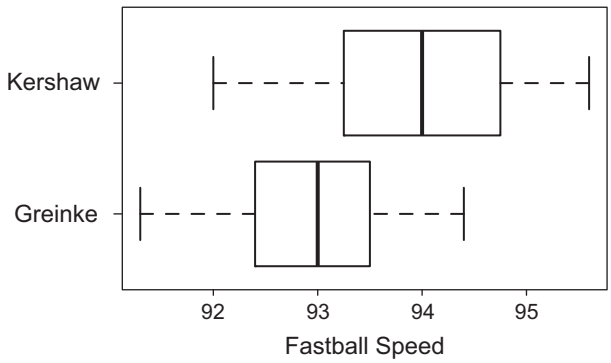


Figure 3.15. Speeds of four-seam fastballs (in miles per hour) by Clayton Kershaw and Zack Greinke during 2015 season.

(a) By reading from the graph, find the five-number summaries of the pitch speeds for the two pitchers.
(b) Which pitcher shows more variation in the fastball speeds? Explain what measure was helpful in answering this question.
(c) Which pitcher, on average, throws a faster fastball? By how many mph?

3.18. Table 3.13 displays the number of changeups (CH), curve balls (CU), four-seam fastballs (FF), two-seam fastballs (FT), and sliders (SL) thrown by Clayton Kershaw and Zack Greinke during a single game for each pitcher in the 2015 season.

Table 3.13. Frequency of pitches of different types of Clayton Kershaw and Zack Greinke during a game pitched by each pitcher in the 2015 season

	CH	CU	FF	FT	SL
Clayton Kershaw	2	20	48	0	26
Zack Greinke	18	12	43	23	23

(a) For each pitcher, compute the percentages of pitches of each type.
(b) Describe, using several sentences, how the pitch selection differs between the two pitchers.

3.19. Table 3.14 shows batted ball data for Ichiro Suzuki for 13 seasons of his career.

Table 3.14. Batted ball statistics for Ichiro Suzuki for seasons 2002 through 2015

	Direction			Hardness		
Season	Pull	Center	Opposite	Soft	Medium	Hard
2002	0.307	0.298	0.395	0.196	0.611	0.193
2003	0.383	0.270	0.347	0.176	0.601	0.223
2004	0.337	0.313	0.350	0.161	0.639	0.200
2005	0.375	0.280	0.345	0.171	0.593	0.237
2006	0.331	0.351	0.318	0.161	0.613	0.225
2007	0.293	0.379	0.328	0.201	0.506	0.293
2008	0.341	0.344	0.314	0.207	0.587	0.206
2009	0.319	0.333	0.349	0.198	0.615	0.187
2010	0.229	0.405	0.366	0.232	0.545	0.222
2011	0.299	0.392	0.310	0.323	0.558	0.119
2012	0.296	0.396	0.308	0.182	0.640	0.178
2013	0.299	0.426	0.275	0.245	0.578	0.176
2014	0.244	0.437	0.319	0.298	0.512	0.190
2015	0.227	0.382	0.391	0.290	0.628	0.082

(a) Suzuki has a very different pattern of batted balls than Albert Pujols whose batted ball statistics are displayed in Table 2.25 (Exercise 2.34). Use a suitable graph to compare the proportions of "Pull" batted balls for Suzuki and Pujols. Write a short paragraph explaining how the two batters differ.
(b) Do a similar comparison of Suzuki and Pujols using the proportion of "Hard" batted balls.

Further Reading

Devore and Peck (2011) and Moore, McCabe and Craig (2012) provide good descriptions of the exploratory methods used in this chapter to compare several batches of data. Chapter 2 of Albert and Bennett (2003) illustrates the use of these methods to compare batches of baseball data.

10.1090/text/034/04

4

Relationships Between Measurement Variables

What's On-Deck?

This chapter illustrates statistical methods for studying the association between different baseball statistics. A large number of statistics, such as hits, doubles, triples, home runs, batting average, slugging percentage, and on-base percentage are used to evaluate hitters. Case Study 4.1 explores the relationships between these different batting statistics for the teams in the 2014 baseball season. The basic tool used in this case study is the scatterplot, and the patterns in the scatterplots are informative in assessing the strength and direction of association between pairs of variables. The most valuable team batting statistic is the one that is most highly associated with runs scored per game. In Case Study 4.2, we rank the different batting statistics with respect to their correlation with runs scored. In Case Study 4.3, we go one step further and evaluate a number of batting statistics, such as batting average, slugging percentage, on-base percentage, and OPS (on-base percentage plus slugging percentage) by how close each statistic can be used to predict teams' runs scored per game. Multiple regression is a useful tool for finding the "best" batting statistic for predicting team runs based on the number of singles, doubles, triples, home runs, walks, and hit-by-pitch by the team. In Case Study 4.4, we find the best linear combination of these batting events using the least-squares criterion, and the weights of this linear combination are useful for comparing the worth of a single, double, triple, and home run from the standpoint of producing runs.

The remaining case studies illustrate some interesting relationships in baseball data. In Case Study 4.5 we illustrate Bill James' Pythagorean Formula which is an empirical relationship between the ratio of a team's wins and losses and the square of the ratio of a team's runs scored and runs allowed. In Case Study 4.6, we see that there is a natural tendency for a player's statistic one year to regress towards the average value of the statistic for the next year. Despite popular opinion, it is natural for a hot-hitting rookie to have a less-than-hot hitting performance the following year.

4.1 Relationships in Team Offensive Statistics

Topics Covered: Relationships between two measurement variables, scatterplot, looking for association.

Table 4.1. Team names, league, and batting statistics for the teams in the 2014 baseball season

Team	G	AB	R	H	2B	3B	HR	AVG	SLG	OBP
ARI	162	5552	615	1379	259	47	118	0.248	0.376	0.302
ATL	162	5468	573	1316	240	22	123	0.241	0.360	0.305
BAL	162	5596	705	1434	264	16	211	0.256	0.422	0.311
BOS	162	5551	634	1355	282	20	123	0.244	0.369	0.316
CHA	162	5543	660	1400	279	32	155	0.253	0.398	0.310
CHN	162	5508	614	1315	270	31	157	0.239	0.385	0.300
CIN	162	5395	595	1282	254	20	131	0.238	0.365	0.296
CLE	162	5575	669	1411	284	23	142	0.253	0.389	0.317
COL	162	5612	755	1551	307	41	186	0.276	0.445	0.327
DET	162	5630	757	1557	325	26	155	0.277	0.426	0.331
HOU	162	5447	629	1317	240	19	163	0.242	0.383	0.309
KCA	162	5545	651	1456	286	29	95	0.263	0.376	0.314
LAA	162	5652	773	1464	304	31	155	0.259	0.406	0.322
LAN	162	5560	718	1476	302	38	134	0.265	0.406	0.333
MIA	162	5538	645	1399	254	36	122	0.253	0.378	0.317
MIL	162	5462	650	1366	297	28	150	0.250	0.397	0.311
MIN	162	5567	715	1412	316	27	128	0.254	0.389	0.324
NYA	162	5497	633	1349	247	26	147	0.245	0.380	0.307
NYN	162	5472	629	1306	275	19	125	0.239	0.364	0.308
OAK	162	5545	729	1354	253	33	146	0.244	0.381	0.320
PHI	162	5603	619	1356	251	27	125	0.242	0.363	0.302
PIT	162	5536	682	1436	275	30	156	0.259	0.404	0.330
SDN	162	5294	535	1199	224	30	109	0.226	0.342	0.292
SEA	162	5450	634	1328	247	32	136	0.244	0.376	0.300
SFN	162	5523	665	1407	257	42	132	0.255	0.388	0.311
SLN	162	5426	619	1371	275	21	105	0.253	0.369	0.320
TBA	162	5516	612	1361	263	24	117	0.247	0.367	0.317
TEX	162	5460	637	1400	260	28	111	0.256	0.375	0.314
TOR	162	5549	723	1435	282	24	177	0.259	0.414	0.323
WAS	162	5542	686	1403	265	27	152	0.253	0.393	0.321

In this case study we discuss relationships in baseball hitting data. Table 4.1 lists batting statistics for all 30 teams for the 2014 baseball season. These data give G, AB, R, H, 2B, 3B, HR, RBI, AVG, TB, SLG, and OBP for all teams. When there are many variables measured on each team, we are interested in exploring relationships between them.

Relating Slugging Percentage and On-Base Percentage

A first step in examining the relationship between two numerical variables, say SLG and OBP, is to draw a scatterplot. This graph plots each pair of values on a grid, and we look for association by finding particular patterns in the display.

Note from Table 4.1 that Arizona (ARI) has an SLG value of .376 and an OBP value of .302. Here we've decided to use OBP on the vertical scale and SLG on the horizontal. (It would not have mattered if we had switched what variables were on which axis.) We plot (SLG=.376,

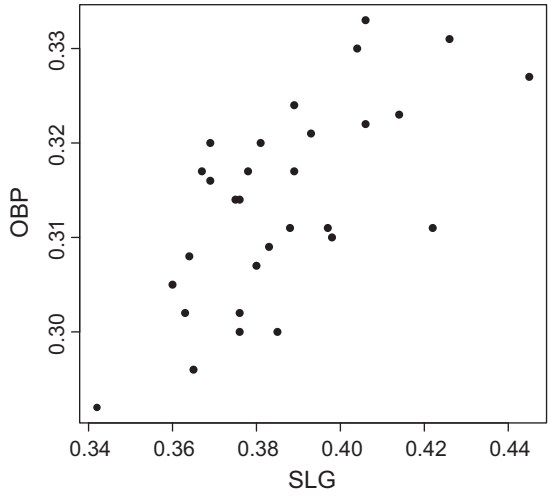


Figure 4.1. Scatterplot of slugging percentage and on-base percentage for the 2014 Major League teams.

OBP=.302) by plotting the point (.376, .302) on a Cartesian grid. We continue for the remaining 29 teams, getting the display shown in Figure 4.1.

We see a general tendency for the points to drift from (left, low) to (right, high), indicating that teams with a low OBP tend also to have a low SLG, and teams with high OBP tend to have a high SLG. This is a positive association in the scatterplot.

Relating Triples and Doubles

As a second example, we draw a labeled scatterplot, shown in Figure 4.2, of the number of triples and number of doubles hit by the 30 teams. Each plotting point has a two-letter abbreviation label for the team.

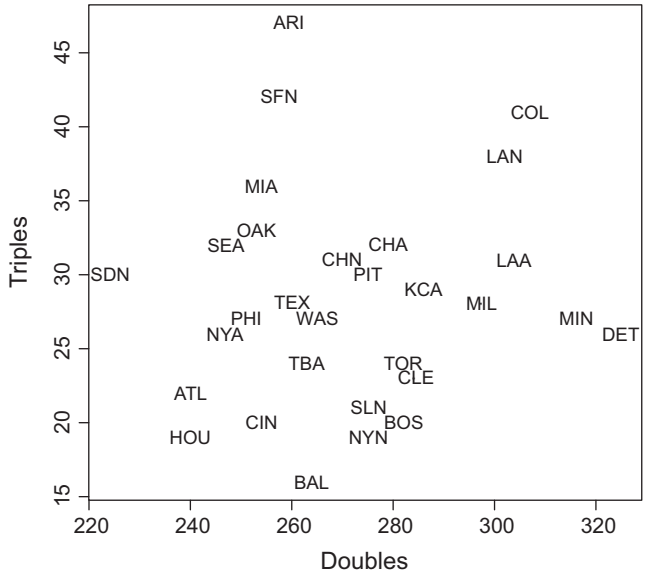


Figure 4.2. Scatterplot of numbers of doubles and triples for the 2014 Major League teams.

The team labels give some interesting information: Colorado (COL) was high in both triples and doubles and Houston (HOU) was low on both variables. Looking at the pattern in the graph, we don't see a very strong association between the numbers of doubles and triples hit. The points may be drifting slightly upward as one goes from left to right, but the positive relationship is clearly weaker than the relationship we saw above between OBP and SLG.

Relating Home Runs and Triples for Historical Teams

To illustrate a different type of relationship, 30 teams were randomly selected from the group of all major league baseball teams in the years from 1901 to 2013. For each team, we collected two hitting statistics:

- Home run rate (HR/G): the number of home runs hit per game. This statistic is computed by dividing the number of team home runs by the number of games played.
- Triple rate (3B/G): the number of triples hit per game. This is computed by dividing the number of triples by the number of games.

Table 4.2 gives the team names, the years, and the two hitting statistics.

Table 4.2. Team names, years, home runs per game, and triples per game for 30 historical baseball teams

Team	Year	HR Rate	T Rate	Team	Year	HR Rate	T Rate
WS1	1916	0.0023	0.0117	NYA	2001	0.0364	0.0036
CHN	1983	0.0254	0.0076	NY1	1942	0.0209	0.0067
ATL	1982	0.0265	0.0040	NYA	1939	0.0313	0.0104
CLE	1983	0.0157	0.0057	PHI	1927	0.0107	0.0087
SDN	2003	0.0231	0.0058	DET	1934	0.0135	0.0097
SDN	1984	0.0198	0.0076	WS1	1946	0.0112	0.0118
CIN	1902	0.0037	0.0157	SLA	1944	0.0137	0.0085
PHA	1934	0.0271	0.0094	NY1	1923	0.0156	0.0139
BAL	1987	0.0378	0.0036	SLN	1906	0.0020	0.0136
PIT	1972	0.0200	0.0086	CLE	1932	0.0144	0.0137
MON	1989	0.0182	0.0055	CHN	1999	0.0345	0.0064
PIT	1975	0.0251	0.0086	ML4	1973	0.0262	0.0072
BOS	1942	0.0196	0.0105	HOU	2007	0.0298	0.0054
LAN	2008	0.0249	0.0053	SDN	2000	0.0282	0.0067
DET	1943	0.0144	0.0088	PIT	1907	0.0038	0.0157

In the scatterplot in Figure 4.3, we plot the home run rate against the triple rate for the 30 teams. Here the points drift from top left to bottom right, which corresponds to a negative relationship in the graph. Teams that hit a high rate of home runs tended to hit a low rate of triples, and teams that hit a low rate of home runs tended to have a high triple rate. We can explain this relationship if we look at the years of the teams in the table. In Figure 4.4, we have redrawn the scatterplot, where the plotting symbol corresponds to the time when the team played.

Note that the points in the upper left portion of the scatterplot generally correspond to teams that played in the first quarter of the 20th century, and the points in the lower right section

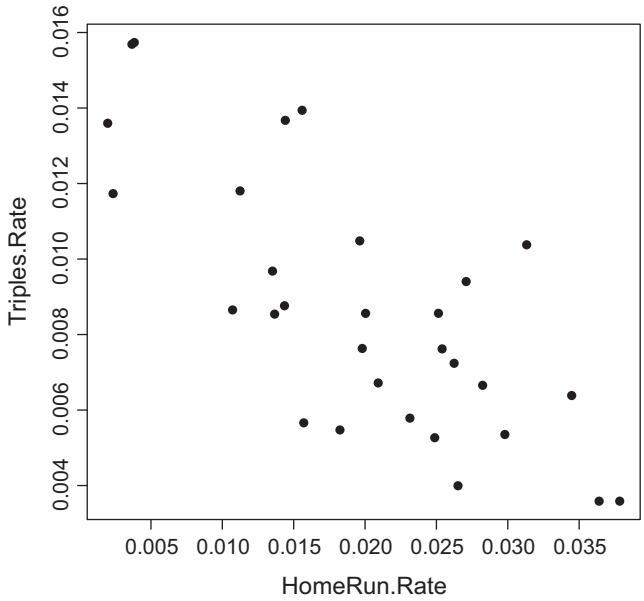


Figure 4.3. Scatterplot of home run rates against triple rates for 30 historical baseball teams.

correspond to teams that played in the most recent period from 1991–2014. In the early days of baseball, relatively few home runs were hit and it was important to use players' speed to hit triples to score runs. In recent years, speed has played much less of a factor in scoring runs, and the home run is an important offensive weapon. So the association structure in this scatterplot tells us how the game of baseball has changed in the last 100 years.

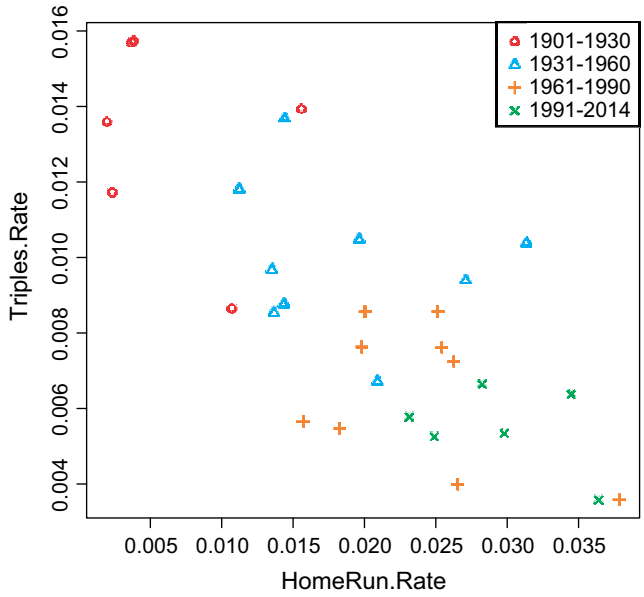


Figure 4.4. Scatterplot of home run rates against triple rates for 30 historical baseball teams where the plotting symbol gives the era of the team.

4.2 Runs and Offensive Statistics

Topics Covered: Scatterplots and looking for association, correlation. Let's talk more about baseball offensive statistics:

- Since the objective of a baseball team is to score more runs than its opponent, the most important offensive statistic is runs.
- The value of any offensive statistic depends on its relationship with runs.
- We are interested in finding the offensive statistic that has the strongest association with runs.

Here is a list of the basic offensive statistics.

H 2B 3B HR RBI AVG TB SLG OBP

How strong is the relationship of these statistics with runs scored (R) by a team? We make some preliminary observations based on our knowledge of baseball.

- Clearly, RBI will have a strong positive relationship with runs, since RBI counts the number of runs that are batted in by a team.
- Triples (3B), in contrast, seem to have a relatively weak relationship with runs. Teams that get a lot of triples don't necessarily score more runs. This is true since triples are a relatively rare event and a large number of triples might just be a reflection of the team's speed or the configuration of the ballpark.
- Home runs (HR) and doubles (2B) have stronger relationships with runs since they occur more frequently than triples and do typically result in runs. (Home runs actually cause at least one run to be scored.)
- TB (total bases) and SLG (slugging percentage) are essentially the same stat (you divide TB by AB to get SLG), so they both have about the same relationship with runs.
- AVG is probably less associated than OBP or SLG with runs. Unlike SLG, AVG doesn't give added weight to extra base hits (doubles, triples, and home runs) over singles; and unlike OBP, AVG doesn't reflect walks or HBP, which both contribute to runs scored by a team.

From this discussion, we get the following list of the statistics ranked with respect to our expected strength of the association with runs scored.

Statistic	Rank
RBI	High association with runs
OBP	
SLG, TB	
AVG	
HR	
2B	
3B	Low association with runs

At this point, we are thinking about association by means of the pattern that we see in the scatterplot. We can summarize these visual associations by means of correlations. Briefly, a correlation is a number between -1 and 1 that summarizes the straight-line assocation between

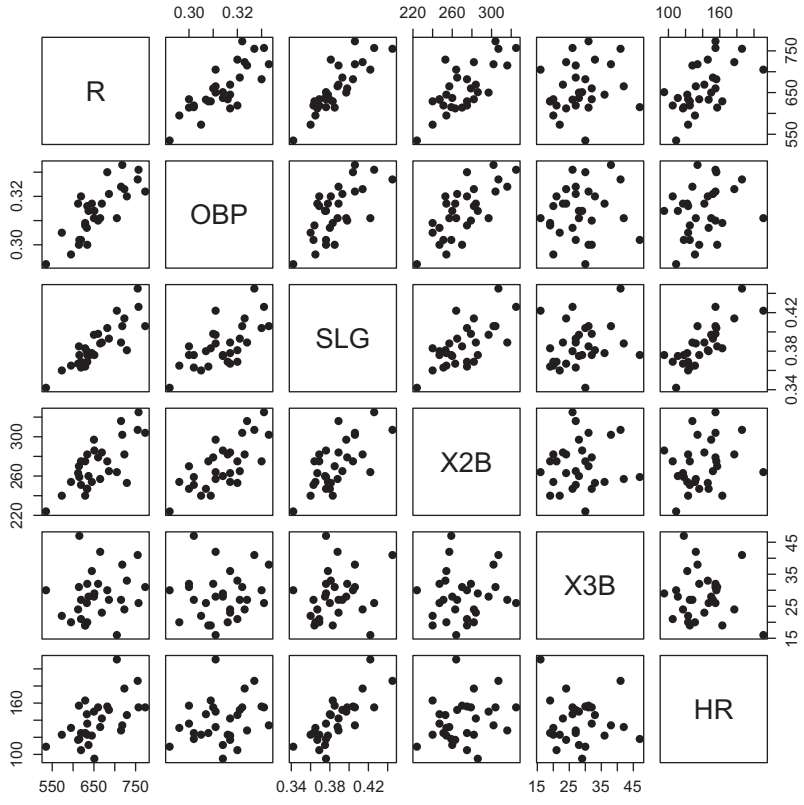


Figure 4.5. Scatterplot matrix of different offensive statistics for 30 Major League teams in the 2014 season.

two numerical variables. A stronger association corresponds to a correlation value that is closer to the extreme values -1 and 1.

We return to the team offensive statistics from the 2014 season discussed previously. We're interested in the offensive statistics that are most associated with runs scored. Look at the scatterplot matrix in Figure 4.5 and focus on the plots in the first column; these are the ones that plot runs (R) against OBP, SLG, 2B, 3B, HR.

Some comments:

- Runs are strongly positively associated with OBP and SLG, although the relationship with SLG looks a bit stronger.
- Runs are somewhat positively associated with HR; the association is less strong than the association with OBP and SLG.
- The relationship between Runs and 2B, and Runs and 3B looks weak.

We can support these comments by computing correlations. The matrix in Table 4.3 gives correlations between the different offensive team statistics. Using these correlation values, we rank the most valuable offensive statistics in Table 4.4.

Table 4.3. Correlation matrix between a number of different offensive team statistics

	R	OBP	SLG	2B	3B	HR
R	1	0.797	0.857	0.737	0.202	0.578
OBP	0.797	1	0.668	0.724	0.104	0.260
SLG	0.857	0.668	1	0.676	0.209	0.788
2B	0.737	0.724	0.676	1	0.090	0.238
3B	0.202	0.104	0.209	0.090	1	−0.094
HR	0.578	0.260	0.788	0.238	−0.094	1

Table 4.4. Correlations of different offensive statistics with runs scored, ordered by the size of the correlation

Statistic	Correlation with Runs
SLG	0.857
OBP	0.797
2B	0.737
HR	0.578
3B	0.202

Of the two "averages", while SLG is the best OBP is nearly as good. Of the three count variables, HR is by far the most strongly correlated with runs scored, but it lags well behind the two averages. Doubles (2B) and triples (3B) are inferior to HR, with 3B being virtually useless.

4.3 Most Valuable Hitting Statistics

Topics Covered: Linear regression, prediction.

There is a group of baseball fans who are members of SABR, the Society of American Baseball Research. These people are interested in the history of the game and write many books and articles about the game. Sabermetrics is the mathematical and statistical study of baseball records.

We focus on one of the most important topics to sabermetricians: What is the most valuable hitting statistic?

Who is the better hitter Mike Trout or Miguel Cabrera? This can be a difficult question to answer since batting has two different characteristics, the ability to get on-base and the ability to advance runners on base to score. Your decision about who is the better hitter may depend partly on how you value power hitting relative to on-base ability. We want to find statistics that can distinguish the "total" hitting accomplishments of Trout and Cabrera.

Let's review the basic hitting statistics recognized by Major League Baseball.

- AVG, the batting average, is the most commonly quoted hitting statistic of Major League Baseball. The player who wins the batting crown is the one with the highest batting average.
- OBP, the on-base percentage.
- SLG, the slugging percentage.

Starting in the 1960's, a large number of new statistics have been introduced as "improvements" to the basic three statistics:

- OPS, the sum of OBP and SLG.
- Runs created (RC) introduced by Bill James. It is defined as

$$\mathrm{RC} = \frac{(\mathrm{H} + \mathrm{BB})\mathrm{TB}}{\mathrm{AB} + \mathrm{BB}}.$$

- Total average (TA). It is like SLG, but it is the ratio of bases to the number of outs. (An approximate formula is given below which ignores rare events such as stealing, hit-by-pitches, and ground balls resulting in double plays.)

$$\mathrm{TA} = \frac{\mathrm{TB} + \mathrm{BB}}{\mathrm{AB} - \mathrm{H}}.$$

- Batter's runs average (BRA), the product of OBP and SLG.

How do we evaluate all of these hitting statistics?

First, we have to understand that the goal of hitting is to create runs, and a single player can't create a run (unless he hits a home run). Teams create runs and so we have to look at team data to understand the usefulness of these various statistics.

Let's consider team statistics for the 2014 American League teams displayed in Table 4.5. (We present data for only one league in this case study to make it easier to present the methodology for evaluating a hitting statistic.)

Table 4.5. Offensive statistics for the 2014 American League teams

Team	R	G	R/G	AVG	OBP	SLG	OPS	RC	TA	BRA
BAL	705	162	4.35	0.256	0.311	0.422	0.733	723	0.664	0.131
BOS	634	162	3.91	0.244	0.316	0.369	0.685	635	0.615	0.117
CHA	660	162	4.07	0.253	0.310	0.398	0.708	673	0.634	0.123
CLE	669	162	4.13	0.253	0.317	0.389	0.706	683	0.641	0.123
DET	757	162	4.67	0.277	0.331	0.426	0.757	790	0.698	0.141
HOU	629	162	3.88	0.242	0.309	0.383	0.692	636	0.624	0.118
KCA	651	162	4.02	0.263	0.314	0.376	0.690	646	0.603	0.118
LAA	773	162	4.77	0.259	0.322	0.406	0.728	731	0.665	0.131
MIN	715	162	4.41	0.254	0.324	0.389	0.713	693	0.652	0.126
NYA	633	162	3.91	0.245	0.307	0.380	0.687	632	0.613	0.117
OAK	729	162	4.50	0.244	0.320	0.381	0.701	668	0.644	0.122
SEA	634	162	3.91	0.244	0.300	0.376	0.676	604	0.593	0.113
TBA	612	162	3.78	0.247	0.317	0.367	0.684	632	0.614	0.116
TEX	637	162	3.93	0.256	0.314	0.375	0.689	633	0.607	0.118
TOR	723	162	4.46	0.259	0.323	0.414	0.737	735	0.680	0.134

The important statistic for a team is the number of runs scored. If we divide this number by the number of games, we get the RUNS/GAME statistic (or R/G). For example, Baltimore scored 705 runs in 162 games, so

$$\mathrm{R/G} = \frac{705}{162} = 4.35.$$

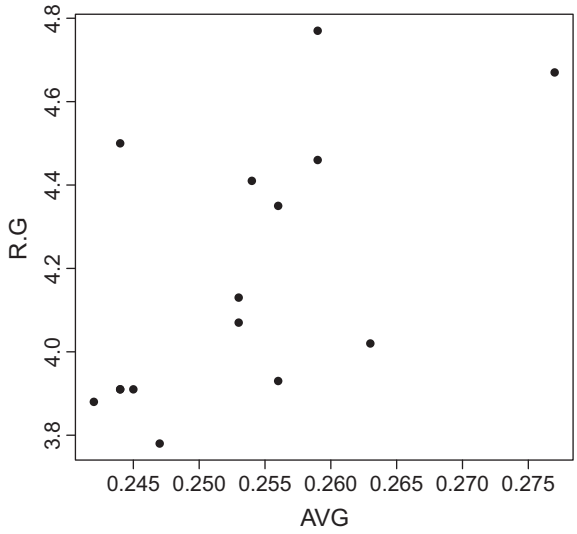


Figure 4.6. Scatterplot of batting average and runs scored per game for 2014 American League teams.

We are interested in seeing the effectiveness of different hitting statistics in predicting R/G. Let's start with the traditional AVG measure. How well can we predict a team's runs per game (R/G) using AVG? If we do a scatterplot of R/G against AVG (shown in Figure 4.6) we see a positive association—teams that hit for high averages tend to score more runs.

The standard "least-squares" fit to these data gives the relationship

$$\text{R/G} = 20.48 \times \text{AVG} - 1.00.$$

Is this a good fit? In other words, how close are the points in the scatterplot to the fitted line? We evaluate the goodness of fit of this prediction equation in several steps. The calculations are summarized in the following table.

1. First, we find the predicted values of R/G for each team. For example, we see that Baltimore had a batting average of .256. We would predict its R/G to be $\text{R/G} = 20.48(.256) - 1.00 = 4.24$. This value is placed in the "Predicted" column.
2. The residual is the difference between the actual R/G for a team and its predicted R/G. In the case of Anaheim, it actually scored an average of 4.35 runs per game and we predicted 4.24, so

$$\text{RESIDUAL} = \text{ACTUAL R/G} - \text{PRED R/G} = 4.35 - 4.24 = 0.11.$$

In Figure 4.7, we have drawn the least-squares line on the scatterplot. The vertical lines drawn from each point represent the residuals. (This least-squares line is actually the line that minimizes the sum of the squared residuals.) We have identified two teams with unusually large residuals. LAA and KC both had a season batting average close to 260. However, LAA's R/G in 2014 was much greater than the R/G predicted by using batting average and the residual is a large positive value. In contrast, KC scored a very small number of runs given its batting average and has a large negative residual.

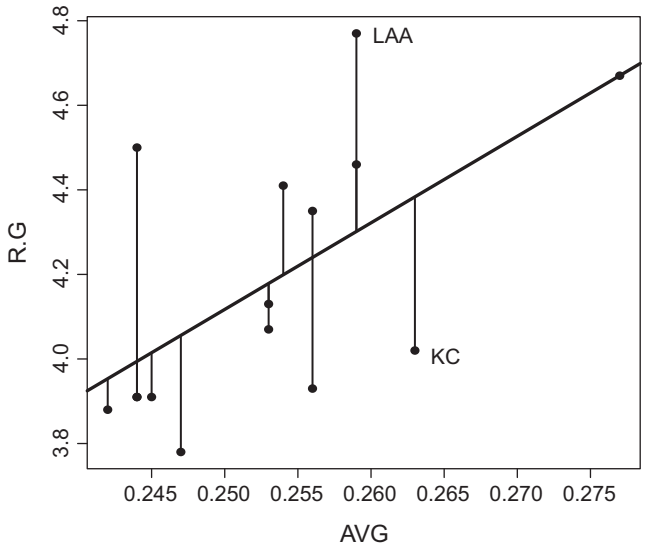


Figure 4.7. Scatterplot of batting average and runs scored per game for 2014 American League teams, with least-squares fit and residuals displayed.

3. We summarize the sizes of these residuals by use of a Root Mean Square Error (RMSE) criterion. Table 4.6 illustrates the calculation of the RMSE. We first square all the residuals and put the answers in the "Residual^2" column.

   The Sum of Squared Errors is the sum of values in the "Residual^2" column. Here it is 0.9092. The mean of the Squared Residuals, called Mean Square Error or MSE, is

$$\text{MSE} = 0.9092/15 = 0.0606.$$

   The Root Mean Square Error, RMSE, is the square root of MSE:

$$\text{RMSE} = \sqrt{0.0606} = 0.246.$$

   The RMSE is a measure of the size of the residuals from the model that uses AVG to predict R/G. Here is a nice interpretation of RMSE: Generally, if you graph all of the residuals, you will find them approximately normally distributed with mean 0 and standard deviation RMSE. Since we have a normal distribution, we can apply the 68-95-99.7 rule and say that 68% (roughly $2/3$) of the residuals fall between $-$RMSE and $+$RMSE To illustrate this interpretation, suppose we use AVG to predict R/G for the American League teams. We found that RMSE $= 0.246$. This means if you use AVG to predict R/G, then roughly $2/3$ (about 9) of the residuals will fall between -0.246 and $+0.246$.

   How good are other batting measures? Now let's try using the batting statistic OBP to predict R/G. The least-squares line is

$$\text{R/G} = 29.6(\text{OBP}) - 5.16.$$

   As we did above, we find the predicted values of R/G for each team and compute the residuals and squared residuals and put the results in Table 4.7. The sum of squared residuals is equal to .6798 and the RMSE is equal to

$$\text{RMSE} = \sqrt{\frac{0.6798}{15}} = .213.$$

Table 4.6. Calculation of the Root Mean Square Error criterion for the least-squares fit with AVG as the predicting variable

Team	AVG	R.G	Predicted	Residual	Residual^2
BAL	0.256	4.35	4.24	0.11	0.0121
BOS	0.244	3.91	3.99	−0.08	0.0064
CHA	0.253	4.07	4.18	−0.11	0.0121
CLE	0.253	4.13	4.18	−0.05	0.0025
DET	0.277	4.67	4.67	0	0.0000
HOU	0.242	3.88	3.95	−0.07	0.0049
KCA	0.263	4.02	4.38	−0.36	0.1296
LAA	0.259	4.77	4.30	0.47	0.2209
MIN	0.254	4.41	4.20	0.21	0.0441
NYA	0.245	3.91	4.01	−0.10	0.0100
OAK	0.244	4.50	3.99	0.51	0.2601
SEA	0.244	3.91	3.99	−0.08	0.0064
TBA	0.247	3.78	4.06	−0.28	0.0784
TEX	0.256	3.93	4.24	−0.31	0.0961
TOR	0.259	4.46	4.30	0.16	0.0256
Sum					0.9092

Let's summarize what we learned. The RMSE is a measure of the size of the residuals in our prediction. When we used AVG to predict Runs/Game, the RMSE was equal to 0.246; when we used OBP, the RMSE is equal to 0.213. This tells us that the residuals are generally much larger using AVG instead of using OBP as a predictor. This means that OBP is a much better predictor

Table 4.7. Calculation of the Root Mean Square Error criterion for the least-squares fit with OBP as the predicting variable

Team	OBP	R.G	Predicted	Residual	Residual^2
BAL	0.311	4.35	4.04	0.31	0.0961
BOS	0.316	3.91	4.19	−0.28	0.0784
CHA	0.310	4.07	4.01	0.06	0.0036
CLE	0.317	4.13	4.22	−0.09	0.0081
DET	0.331	4.67	4.63	0.04	0.0016
HOU	0.309	3.88	3.98	−0.10	0.0100
KCA	0.314	4.02	4.13	−0.11	0.0121
LAA	0.322	4.77	4.37	0.40	0.1600
MIN	0.324	4.41	4.43	−0.02	0.0004
NYA	0.307	3.91	3.92	−0.01	0.0001
OAK	0.320	4.50	4.31	0.19	0.0361
SEA	0.300	3.91	3.72	0.19	0.0361
TBA	0.317	3.78	4.22	−0.44	0.1936
TEX	0.314	3.93	4.13	−0.20	0.0400
TOR	0.323	4.46	4.40	0.06	0.0036
Sum					0.6798

of Runs Scored than AVG using the RMSE criterion. (In general, the smaller the RMSE, the closer the predictions.)

A Prediction Contest

We used each of our seven batting statistics to predict runs per game for teams for our American League data. Table 4.8 displays the corresponding values of RMSE:

Table 4.8. Values of the root mean squared criterion using each of seven batting statistics to predict the runs scored per game for the 2014 American League team data

Statistic	RMSE
AVG	0.246
OBP	0.213
SLG	0.199
OPS	0.159
TA	0.151
RC	0.147
BRA	0.153

We see that OPS (OBP + SLG), TA, RC, and BRA (OBP × SLG) are all pretty close at the top, OBP and SLG are a bit behind, and AVG is pulling up the rear. Of course, we just looked at 2014 data. Is it possible that another batting statistic would do better a different year?' Albert and Bennett (2001) looked at all years from 1952 to 1999. Each year, they used each statistic to predict runs per game for the teams data, and found RMSE for each statistic. Albert and Bennett's general conclusions were:

- AVG performs terribly in predicting runs scored. This measure should be thrown out. (But it probably won't be.)
- Both SLG and OBP are relatively mediocre measures.
- The "brainstorming" statistics TA, BRA, RC, OPS are all good measures and they are pretty close. You can't say for sure that one statistic in this group is better than another in the group.

Remember our motivation for finding the best hitting statistic? We want to use a good statistic to compare the relative worth of two hitters.

To illustrate the use of good hitting statistics to compare players, let's compare Mike Trout and Miguel Cabrera who were both candidates for Most Valuable Player (MVP) in the 2013 season. Table 4.9 gives the batting average and the six alternative batting measures for both Trout and Cabrera for this season.

Table 4.9. Batting measures for Mike Trout and Miguel Cabrera for the 2013 season

	AVG	OBP	SLG	OPS	TA	RC	BRA
Mike Trout	.323	.432	.557	.988	1.098	141	.241
Miguel Cabrera	.348	.442	.636	1.078	1.224	155	.281

Note from the table that, although Cabrera and Trout had similar on-base performances, Cabrera was substantially better than Trout using each of the four "new" batting statistics. Cabrera clearly had a much better hitting year in 2013.

4.4 A New Measure of Offensive Performance

Topics Covered: Multiple linear regression, root mean square error, least-squares criterion.

In the earlier case studies, we introduced a number of batting measures, some old and some new, for evaluating the value of a hitter. We used the 2014 American League team data to evaluate these different measures. Specifically, we used each measure to predict the average runs per game for the 15 AL teams, and compared the measures by computing the average size of a residual (RMSE). We found that many of the modern batting measures, such as OPS, and RC did substantially better compared to the traditional measures AVG and SLG for predicting run production.

There is actually a straightforward way of finding a "best" measure of hitting performance using the tool of multiple linear regression.

The basic batting counts are the numbers of singles, doubles, triples, and home runs (1B, 2B, 3B, and HR, respectively), and the number of walks (BB) and hit-by-pitch (HBP). The goal is to combine these different counts in some way to obtain an accurate prediction of the runs scored per game.

Many of the standard batting measures combine these batting counts in a linear way. For example, the batting average AVG can be expressed as

$$\text{AVG} = \frac{1}{\text{AB}}1\text{B} + \frac{1}{\text{AB}}2\text{B} + \frac{1}{\text{AB}}3\text{B} + \frac{1}{\text{AB}}\text{HR},$$

where singles, doubles, triples, and home runs are given equal weights. The slugging percentage SLG is also a linear measure of the form

$$\text{SLG} = \frac{1}{\text{AB}}1\text{B} + \frac{2}{\text{AB}}2\text{B} + \frac{3}{\text{AB}}3\text{B} + \frac{4}{\text{AB}}\text{HR},$$

where a hit is weighted by the number of bases. The on-base percentage, OBP, weights all on-base events (hits, walks, and hit-by-pitch) equally:

$$\text{OBP} = \frac{1}{\text{PA}}\text{BB} + \frac{1}{\text{PA}}\text{HPB} + \frac{1}{\text{PA}}1\text{B} + \frac{1}{\text{PA}}2\text{B} + \frac{1}{\text{PA}}3\text{B} + \frac{1}{\text{PA}}\text{HR},$$

where PA = AB + BB + HBP is the number of plate appearances.

Suppose that we consider an alternative batting measure, called OPTAVG (for optimal average), that combines all of the batting events in a linear way with arbitrary weights:

$$\text{OPTAVG} = w_0 + w_1 \times 1\text{B} + w_2 \times 2\text{B} + w_3 \times 3\text{B} + w_4 \times \text{HR} + w_5 \times \text{BB} + w_6 \times \text{HBP}.$$

Can we find values of the weights $w_0, \ldots, w_6$ that give a "best" batting measure?

Fortunately, there is an easy way to find optimal weights for this batting measure using a least-squares criterion. As in the earlier case study, suppose that we have team hitting data. For each team, we observe the runs scored per game (R/G) and the counts of the six batting events (1B, 2B, 3B, HR, BB, HBP). The goal is to predict R/G based on the linear measure OPTAVG. Using the least-squares criterion, we wish to find values of the weights $w_0, \ldots, w_6$ that minimize the sum of squared residuals. Values of these weights, called the least-squares estimates, are easily available using any standard statistical computing package. We apply this method to estimate R/G for the 30 Major League Teams in 2014. Using the R package, we find the regression equation to be

$$\begin{aligned}\text{R/G} = &-1.83144 + 0.00251 \times 1\text{B} + 0.00563 \times 2\text{B} + 0.00761 \times 3\text{B} \\ &+ 0.00725 \times \text{HR} + 0.00175 \times \text{BB} - 0.00092 \times \text{HBP}.\end{aligned}$$

Before we interpret the weights of this measure, let's explain why this is the best linear batting measure. Among all batting statistics that combine the different batting events (single, double, etc.) in a linear way, this measure will give the smallest value of RMSE, which is an average size of a residual when this measure is used as a predictor. Since this "best linear" measure has the smallest RMSE, it will have a smaller RMSE than other linear measures such as AVG, SLG, OBP, and OPS.

To demonstrate this fact, we also tried using AVG, SLG, OBP, OPS, and RC, each alone, as predictors for this 2014 Major League team dataset. Table 4.10 gives values of RMSE for all 30 Major League teams in 2014 using the best linear and the more familiar measures.

Table 4.10. Values of the root mean squared criterion using the best linear and other measures for the 2014 Major League team data

Statistic	RMSE
OPTAVG	0.116
AVG	0.205
OBP	0.206
SLG	0.176
OPS	0.144
TA	0.138
RC	0.129
BRA	0.141

We see that the best linear measure OPTAVG has an RMSE value, 0.116, that appears to be much smaller than its nearest competitor RC (0.129) and the popular OPS statistic (0.144). But this difference is a bit deceptive since we found the best weights using the same 2014 dataset—it is not clear that this best measure will also be good in predicting the runs scored per game for team data for 2015 year.

The weights of the best linear measure OPTAVG tell about the worth of each type of batting event. When we compute a batting average, we give each type of hit the same weight and we ignore walks and hit-by-pitches. A slugging percentage weights each hit by the number of bases

Table 4.11. Weights of batting events for the best linear measure OPTAVG

Event	Weight	Wt/Wt (Single)
Single	0.00251	1
Double	0.0056	2.25
Triple	0.00761	3.04
Home run	0.00725	2.90
Walk	0.00175	0.70
Hit by Pitch	−0.00092	−0.37

produced. What does our best linear measure do for weights? Table 4.11 shows the values of the weights for each base event, and, to make the comparison easy, it standardizes the weights by dividing each by the weight of a single.

Some of the values of these weights are a bit surprising. The weight for a home run is actually smaller than the weight given to a triple! A hit by pitch is given a negative weight and a triple and home run are given roughly the same weights.

How can we explain these weights?

The first comment is that we are using a relatively small amount of data and these unusual weight values are a consequence of the fact that we are fitting a relatively complicated model to a small dataset. We could get more reasonable estimates at these weights by using, say, 10 years of team data instead of just one year. Albert and Bennett (2001) (Chapter 5) do exactly that and get more reasonable weights for this linear batting measure.

However, some of these weights do make sense. It is well known by quantitative baseball people that a home run is not four times the worth of a single—its actual worth is closer to three times a single which is the value in this table. A walk and single should have weights of a similar size since all these events get the batter to first base. But the single is a more valuable event than a walk since the runners are typically advanced more than one base. In other words, the larger weight for a single reflects the fact that the single is more effective than a walk in producing runs.

4.5 How Important is a Run?

Topics Covered: Nonlinear regression, transformation, residuals.

In this chapter, we have discussed how it is important for a baseball team to score runs and we evaluated the goodness of different batting measures by their relationship with runs scored. But of course the objective of a baseball team is not solely to score runs—it wins games by scoring more runs than their opponent. That raises the interesting question. What is the importance of a single run towards the goal of winning a baseball game? If a player is responsible for scoring, say 20 runs, then how many wins for his team has he contributed? Bill James discovered a special relationship between the number of wins (W) and losses (L) for a baseball team and the number of runs scored (R) and number of runs allowed (RA). He called this relationship "The Pythagorean Method." This result says that the ratio between a team's wins and losses is approximately equal to the square of the ratio of runs scored to runs allowed.

That is, approximately,

$$\frac{W}{L} = \left(\frac{R}{RA}\right)^2.$$

If we take logs of both sides, we get the equivalent relationship

$$\log\left(\frac{W}{L}\right) = 2\log\left(\frac{R}{RA}\right).$$

(We take logs to convert a nonlinear equation in the runs ratio to a linear equation.) Can we demonstrate that the Pythagorean Method gives a good description of the relationship between the win/loss pattern and the runs scored/allowed for current Major League Baseball teams? To answer this question, we look at the relevant team statistics (wins, losses, runs scored, and runs allowed) for the 30 teams for the 2014 season. Looking at Table 4.12, we see that if a team has a winning record, then it generally scores more runs than its opponent. But there is one interesting exception—the New York Yankees (NYA) had a win/loss record of 84-78 but actually allowed $664 - 633 = 31$ more runs this season than they scored. (We suppose that the Yankees won a lot of close games in 2014.)

Table 4.12. Team statistics for the Major League teams in the 2014 season

Team	W	L	R	RA	Team	W	L	R	RA
ARI	64	98	615	742	MIL	82	80	650	657
ATL	79	83	573	597	MIN	70	92	715	777
BAL	96	66	705	593	NYA	84	78	633	664
BOS	71	91	634	715	NYN	79	83	629	618
CHA	73	89	660	758	OAK	88	74	729	572
CHN	73	89	614	707	PHI	73	89	619	687
CIN	76	86	595	612	PIT	88	74	682	631
CLE	85	77	669	653	SDN	77	85	535	577
COL	66	96	755	818	SEA	87	75	634	554
DET	90	72	757	705	SFN	88	74	665	614
HOU	70	92	629	723	SLN	90	72	619	603
KCA	89	73	651	624	TBA	77	85	612	625
LAA	98	64	773	630	TEX	67	95	637	773
LAN	94	68	718	617	TOR	83	79	723	686
MIA	77	85	645	674	WAS	96	66	686	555

To look for the Pythagorean relationship, we compute log(W/L) and log(R/RA) for all teams and construct a scatterplot of the two quantities in Figure 4.8.

We see a linear positive association in this graph, indicating that there is indeed a linear association between log(W/L) and log(R/RA). Next we want to fit a "best line" to this graph. It seems natural to restrict this line to pass through one point. If a team scores the same number of runs against its opponents (R = RA), then we expect the team to win half of its games (W = L). In other words, the point $(\log(R/RA), \log(W/L)) = (0, 0)$ should fall on the line.

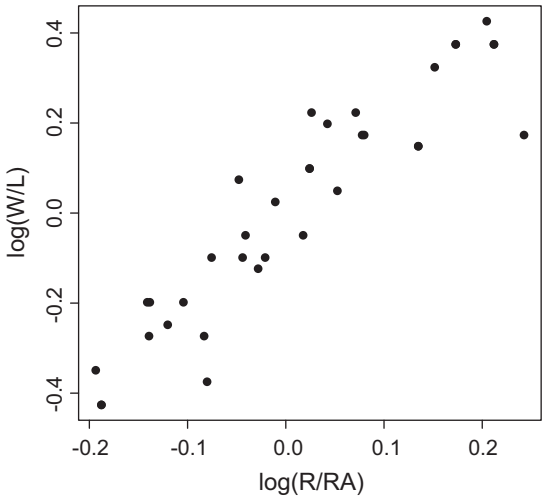


Figure 4.8. Scatterplot of log runs ratio against log of ratio of wins to losses for Major League team data from the 2014 season.

With this restriction, we look at line fits of the form

$$\log\left(\frac{\mathrm{W}}{\mathrm{L}}\right) = k \log\left(\frac{\mathrm{R}}{\mathrm{RA}}\right).$$

We choose k by using a least-squares criterion. It turns out that the sum of squared residuals is minimized when $k = 1.81$. Figure 4.9 shows this best line on the scatterplot and a display of

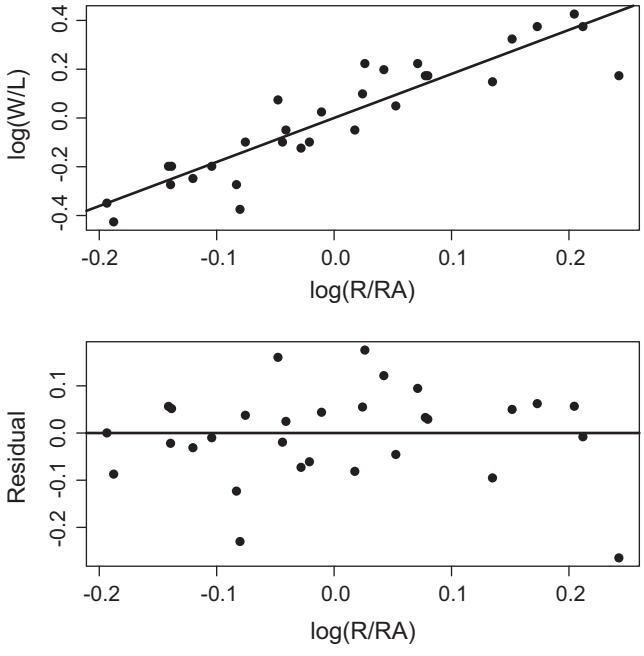


Figure 4.9. Least-squares fit (top) and residual plot (bottom) for (R/RA, W/L) data.

the corresponding residuals. We do not see any linear trend or any other pattern in the residual plot, so it appears that our fit is satisfactory.

So based on our analysis, we arrive at the relationship

$$\frac{W}{L} = \left(\frac{R}{RA}\right)^{1.81}.$$

which is pretty close to James' Pythagorean relationship which uses the power of 2. How useful is this rule in predicting a team's win numbers? To check the accuracy of this relationship in prediction, Table 4.13 gives the actual number of wins, the predicted number of wins (using the above model) and the residual (actual minus predicted). One can see from the table that 24 of the sizes of the 30 residuals are smaller than 4. This indicates that for 80% of the teams, we can predict the number of wins to within four games using this formula.

Table 4.13. Number of wins, predicted number of wins, and residuals using James' Pythagorean relationship

Team	W	expected	residual	Team	W	expected	residual
ARI	64	67.4	−3.4	MIL	82	80.2	1.8
ATL	79	78.0	1.0	MIN	70	74.9	−4.9
BAL	96	93.5	2.5	NYA	84	77.5	6.5
BOS	71	72.2	−1.2	NYN	79	82.3	−3.3
CHA	73	70.9	2.1	OAK	88	98.5	−10.5
CHN	73	70.7	2.3	PHI	73	73.4	−0.4
CIN	76	78.9	−2.9	PIT	88	86.7	1.3
CLE	85	82.8	2.2	SDN	77	75.5	1.5
COL	66	75.1	−9.1	SEA	87	90.8	−3.8
DET	90	86.2	3.8	SFN	88	86.8	1.2
HOU	70	70.9	−0.9	SLN	90	82.9	7.1
KCA	89	84.1	4.9	TBA	77	79.5	−2.5
LAA	98	95.8	2.2	TEX	67	67.0	0.0
LAN	94	92.0	2.0	TOR	83	84.8	−1.8
MIA	77	77.8	−0.8	WAS	96	96.3	−0.3

4.6 Baseball Players Regress to the Mean

Topics Covered: Least-squares regression, prediction.

Here we examine least-squares regression. Specifically, we describe the "regression effect" that is generally unknown to most people in baseball.

If you look up the word "regress" in the dictionary, it will tell you the word means to "go back". We will see that there is a general tendency for a player's baseball statistics from one year to the next to go back, or regress to the mean.

Many of you have heard about the so-called "sophomore slumps" in sports. This happens when someone does well in his/her rookie year and then slumps in the sophomore year. This tendency is commonly discussed among baseball people. Here we see that there is a natural tendency for best performing rookies to slump their sophomore years.

Table 4.14. 2013 and 2014 slugging percentages for 20 randomly selected players

Name	SLG.2013	SLG.2014	Improvement
Domonic Brown	0.494	0.349	−0.145
Melky Cabrera	0.360	0.458	0.098
Miguel Cabrera	0.636	0.524	−0.112
Alberto Callaspo	0.369	0.290	−0.079
Zack Cozart	0.381	0.300	−0.081
Chris Davis	0.634	0.404	−0.230
Matt Dominguez	0.403	0.330	−0.073
Brian Dozier	0.414	0.416	02
Conor Gillaspie	0.390	0.416	0.026
Alex Gordon	0.422	0.432	0.010
Chase Headley	0.400	0.372	−0.028
Aaron Hill	0.462	0.367	−0.095
Matt Holliday	0.490	0.441	−0.049
Brandon Moss	0.522	0.438	−0.084
Albert Pujols	0.437	0.466	0.029
Pablo Sandoval	0.417	0.415	−02
Denard Span	0.380	0.416	0.036
Ichiro Suzuki	0.342	0.340	−02
Mark Trumbo	0.453	0.415	−0.038
Jayson Werth	0.532	0.455	−0.077

We look at the SLG data for twenty randomly selected players who had at least 300 at-bats in the 2013 and 2014 seasons. Table 4.14 gives the 2013 and 2014 SLG values for these players. In addition, in the column labeled "Improvement", we show how many SLG points a player improved from 2013 to 2014; if this improvement is negative, this means that the player had a lower SLG in 2014.

Suppose that some player has a 500 slugging percentage in 2013. What do you predict his SLG value to be in 2014? (Assuming that you don't know his 2014 data yet.) If you don't know anything about the hitter, it would seem reasonable to predict his 2014 SLG also to be 500. Is this the best prediction? To investigate this, we first plot the players' 2014 SLG against the players' 2013 values in Figure 4.10.

We see a relatively strong positive association in this graph. Players that were good sluggers in 2013 (like Miguel Cabrera) tended to be good in 2014. Likewise, players with small SLG values in 2013 (Ichiro Suzuki comes to mind) tended to be poor also in 2004.

The least-squares line (plotted on the scatterplot) is

$$\text{SLG.2014} = 0.358(\text{SLG.2013}) + 0.242.$$

Let's illustrate using this least-squares line to predict a player's 2014 SLG. Let us select Chris Davis of the Orioles.

His 2013 SLG value was 0.634. Using the line, we would predict his 2014 SLG value to be

$$\text{SLG.2014} = 0.358 \times 0.634 + 0.242 = 0.469.$$

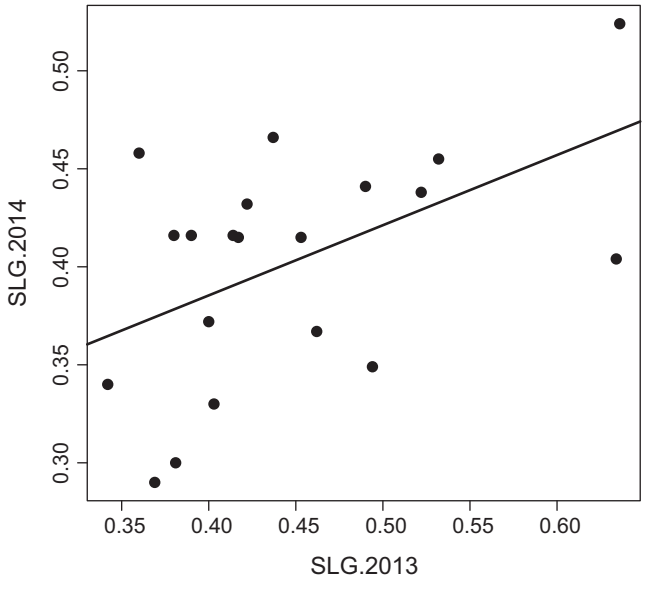


Figure 4.10. Scatterplot of 2013 and 2014 slugging percentages for 20 players. A least-squares line is shown on the graph.

Let's interpret this prediction:

1. In 2013, Davis' SLG value was 0.634. Since the mean SLG (among these 20 players) was 0.469, Chris was 165 points above average in 2013.
2. We would predict Davis' 2014 SLG value to be 0.469. Since the mean SLG in 2014 was 0.402, our prediction is 67 points above average.

So his performance in 2013 was 165 points above average, and we predict his 2014 performance to be 67 points above average. In other words,

> We predict that his 2014 slugging percentage will be closer to the mean than his 2013 slugging percentage.

This demonstrates a general tendency for a player's SLG to regress towards the mean.

Let's explain this regression effect a different way. Recall that we defined a player's improvement as

$$\text{IMPROVEMENT} = (\text{SLG.2014}) - (\text{SLG.2013}).$$

For example, Melky Cabrera's improvement would be $0.458 - 0.360 = 0.098$—he was a better slugger in 2014. Looking at the data, we see that

- the best improvement was Melky Cabrera at $+0.098$,
- the most negative improvement was Chris Davis at -0.230.

Suppose that we compute the improvement for all players and graph the Improvement against the 2013 SLG. The scatterplot is shown in Figure 4.11. We see that good players in 2013

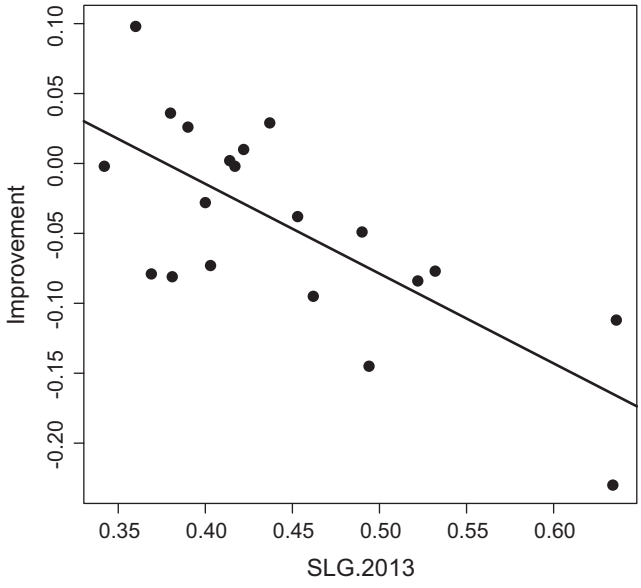


Figure 4.11. Scatterplot of 2013 slugging percentages and improvement in SLG from 2013 to 2014 for 20 players. A least-squares fit is drawn on the graph.

(with large SLGs) tend to have negative improvements, and poor players in 2013 (with small SLGs) tend to improve in 2014. This demonstrates a negative association in this graph.

The line in the graph is a least-squares fit. The correlation $r = -0.72$ and the least-squares line is

$$\text{IMPROVEMENT} = 0.242 - 0.39\ (\text{SLG.2013}).$$

This "regression effect" actually is very common. If you take statistics for a group of players for two consecutive years, you will find that a player's improvement is negatively associated with his first year performance. Players' statistics (from one year to the next) tend to move towards the average value.

4.7 Exercises

4.0. Table 4.15 shows career on-base percentage (OBP) and slugging percentage (SLG) for Rickey Henderson for the first 23 seasons of his career.

(a) Construct a scatterplot of Henderson's OBP and SLG values where OBP is the horizontal variable and SLG is the vertical variable.

(b) There are five unusual points on the left side of the plot that don't follow the general pattern. This is where Henderson had a small slugging percentage. Find the ages that correspond to these unusual points.

(c) Circle the point that corresponds to Henderson's best season with respect to both OBP and SLG. What was Henderson's age this particular year?

(d) If you ignore the five unusual points, what is the general pattern in the scatterplot?

Table 4.15. On-base and slugging percentages for Rickey Henderson for his first 23 seasons of his career

Age	OBP	SLG	Age	OBP	SLG
20	.338	.336	32	.400	.423
21	.420	.399	33	.426	.457
22	.408	.437	34	.432	.474
23	.398	.382	35	.411	.365
24	.414	.421	36	.407	.447
25	.399	.458	37	.410	.344
26	.419	.516	38	.400	.342
27	.358	.469	39	.376	.347
28	.423	.497	40	.423	.466
29	.394	.399	41	.368	.305
30	.411	.399	42	.366	.351
31	.439	.577			.345

4.1. Table 4.16 shows the batting average and number of runs per game for all of the National League teams in the 2014 season.

Table 4.16. Batting average and runs scored per game for 2014 National League teams

Team	AVG	R.G	Team	AVG	R.G
ARI	0.248	3.80	NYN	0.239	3.88
ATL	0.241	3.54	PHI	0.242	3.82
CHN	0.239	3.79	PIT	0.259	4.21
CIN	0.238	3.67	SDN	0.226	3.30
COL	0.276	4.66	SFN	0.255	4.10
LAN	0.265	4.43	SLN	0.253	3.82
MIA	0.253	3.98	WAS	0.253	4.23
MIL	0.250	4.01			

(a) On the grid in Figure 4.12, construct a scatterplot of Runs per Game (vertical axis) against Batting Average (horizontal).

(b) What does the scatterplot say about the relationship between Runs per Game and Batting Average? If you know that a team has a high batting average, what does that say about the number of runs it scores?

4.2. (Exercise 4.1 continued.) A least-squares fit to the (Runs per Game, Batting Average) data for the 2014 NL teams gives the relationship.

$$\text{RUNS} = -2.57 + 26.2\text{AVG}.$$

(a) Suppose that a team has a .250 batting average. How many runs do you predict the team will score in a game?

(b) Suppose that a team has a .260 batting average. Predict the number of runs the team will score.

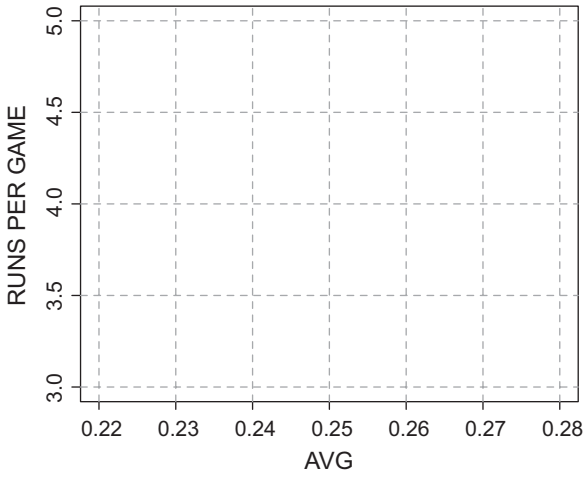


Figure 4.12

(c) Suppose a team hits for a 10 point higher batting average (say from .250 to .260). How many additional runs per game could we have predicted the team to score?

(d) Arizona in 2014 had a .248 batting average and scored an average of 3.80 runs per game. For Arizona, find the predicted runs scored and the residual. Can you explain why the residual is negative in this case?

4.3. Table 4.17 gives a number of pitching statistics for the 2014 National League teams. The abbreviations used in the table are

- SOR = number of pitching strikeouts per game
- BOR = number of pitching walks per game
- ERA = earned run average

Table 4.17. Team pitching statistics for 2014 National League teams

Team	SOR	BOR	ERA	PCT	HG	HRG	RG
ARI	7.89	2.90	4.26	0.40	9.06	0.73	4.58
ATL	8.03	2.91	3.38	0.49	8.45	0.76	3.69
CHN	8.09	3.11	3.91	0.45	8.63	0.97	4.36
CIN	7.96	3.13	3.59	0.47	7.91	0.81	3.78
COL	6.63	3.28	4.84	0.41	9.43	1.15	5.05
LAN	8.48	2.65	3.40	0.58	8.26	0.83	3.81
MIA	7.35	2.83	3.78	0.47	9.14	0.75	4.16
MIL	7.69	2.66	3.67	0.51	8.56	0.93	4.06
NYN	8.04	3.14	3.49	0.49	8.46	0.77	3.81
PHI	7.75	3.22	3.79	0.45	8.62	0.77	4.24
PIT	7.58	3.08	3.47	0.54	8.28	0.96	3.90
SDN	7.93	2.85	3.27	0.47	8.02	0.67	3.56
SFN	7.48	2.40	3.50	0.54	8.06	0.81	3.79
SLN	7.54	2.90	3.50	0.56	8.15	0.65	3.72
WAS	7.95	2.17	3.03	0.59	8.34	0.94	3.43

- PCT = winning percentage
- HG = number of hits allowed per game
- HRG = number of home runs allowed per game
- RG = number of runs allowed per game

(a) Circle the variables below that you believe have a positive association with runs allowed per game (RG).

SOR BOR ERA PCT HG HRG

(b) Circle the variables below that you believe have a negative association with runs allowed per game (RG).

SOR BOR ERA PCT HG HRG

(c) Do you think a team's pitched strikeouts per game (SOR) is related to a team's pitched walks per game (BOR)? Explain what type of relationship you would expect to find and why.

(d) What variable among the above do you think has the strongest relationship with a team's winning percentage (PCT)?

4.4. (2014 NL team pitching data) Figure 4.13 displays scatterplots of runs allowed per game (RG) and each of the variables SOR, BOR, ERA, PCT, HG, HRG.

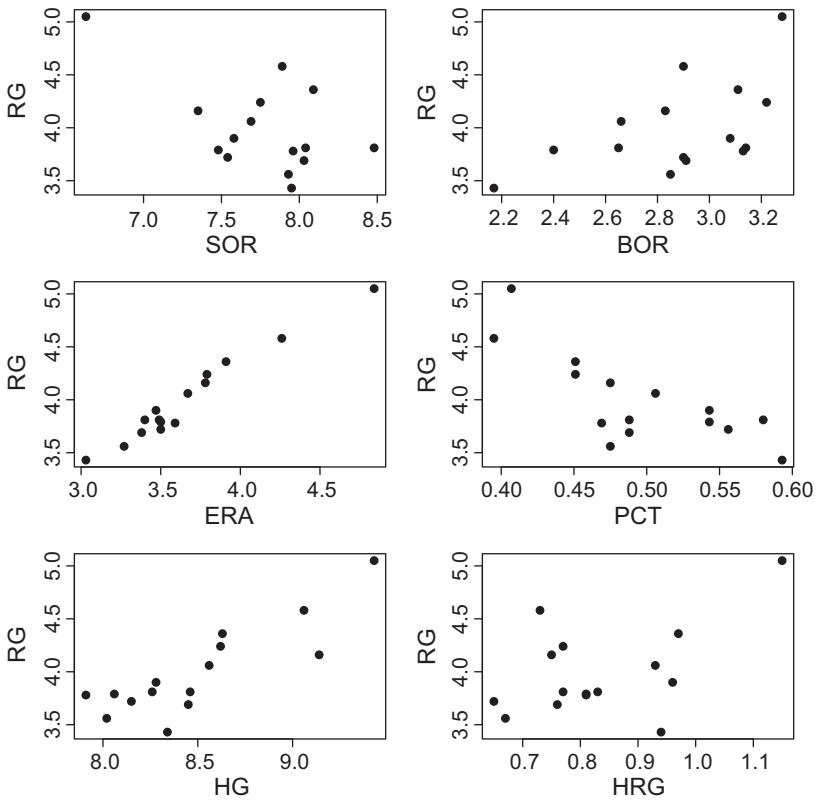


Figure 4.13. Scatterplots of runs allowed and different pitching statistics for 2014 National League teams.

(a) From looking at the scatterplots, list two variables that have a negative association with runs allowed per game.
(b) List two variables that have a positive association with runs allowed per game.
(c) List the variables in order in terms of their association with runs allowed, from most negatively associated to most positively associated.

4.5. (2014 NL team pitching data) Table 4.18 gives the correlations between all of the variables in Exercise 4.6 for the 2014 NL team pitching data. To help interpret the table, note that the correlation between the walk rate (BOR) and the strikeout rate (SOR) is −0.230. This means that there is a positive association between the strikeout and walk rates—teams that struck out a lot of batters tended also to walk a lot of batters, and likewise teams with small strikeout rates tended also to have small walk rates.

Table 4.18. Correlation matrix for a number of pitching variables for 2014 National League teams

	SOR	BOR	ERA	PCT	HG	HRG
BOR	−0.230					
ERA	−0.627	0.562				
PCT	0.320	−0.679	−0.781			
HG	−0.539	0.315	0.805	−0.626		
HRG	−0.413	0.097	0.453	−0.084	0.404	
RG	−0.567	0.543	0.980	−0.759	0.849	0.498

(a) From looking at the correlation matrix, which variable has the strongest relationship with the average runs per game allowed (RG)? Why do you think the correlation value is so close to one?
(b) Which variable has the weakest association with average runs per game? Does this variable have a positive or negative association with RG?
(c) What is the correlation between home runs allowed per game (HRG) and strikeouts pitched per game (SOR)? Explain in words what this correlation says about the relationship between HRG and SOR.
(d) By the correlation matrix, the correlation between home runs allowed per game and walks allowed per game is 0.097 Explain in words what this means about the relationship between home runs and walks allowed.

4.6. (2014 NL team pitching data). We saw earlier that there was a negative relationship between a team's winning percentage (PCT) and the team's earned run average (ERA). A least-squares line through the data has the equation

$$\text{PCT} = 0.880 - 0.105\text{ERA}.$$

(a) Suppose a team allows, on average, four earned runs per game. Use the line to predict its winning percentage.
(b) Suppose a team's ERA jumps from 4 to 5. Using the least-squares line, do you predict its winning percentage would increase or decrease? By how much?
(c) Arizona in 2014 had a 4.26 ERA and a .395 winning percentage. Find Arizona's predicted PCT and the residual. Can you explain why the residual is large in this case?

4.7. Two hundred teams were randomly selected from baseball history. For each team, the season and the number of triples hit per game were recorded. Figure 4.14 plots the triples per game (vertical) against the season (horizontal).

(a) Describe the basic pattern in the graph? Has the number of triples per game increased or decreased over time?

(b) Would it be appropriate to find a best line to these data? Why or why not?

(c) Can you offer any explanation for the pattern in this graph? In other words, why has the frequency of triples changed over the years?

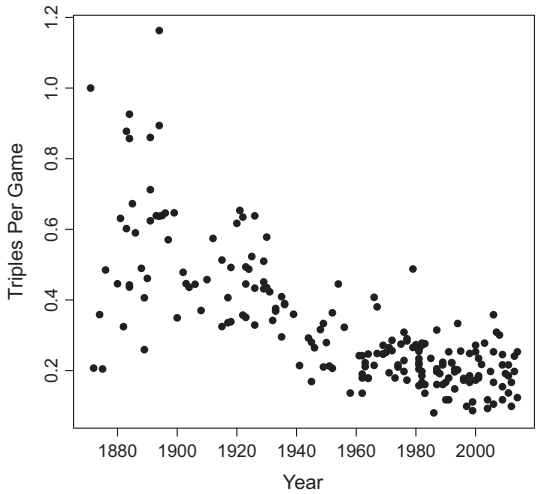


Figure 4.14. Scatterplot of triples per game against season for 200 randomly selected baseball teams.

4.8. Two hundred teams were randomly selected from baseball history. For each team, the season and the number of stolen bases (SB) per game were recorded. In Figure 4.15, we plot the SB per game (vertical) against the season (horizontal).

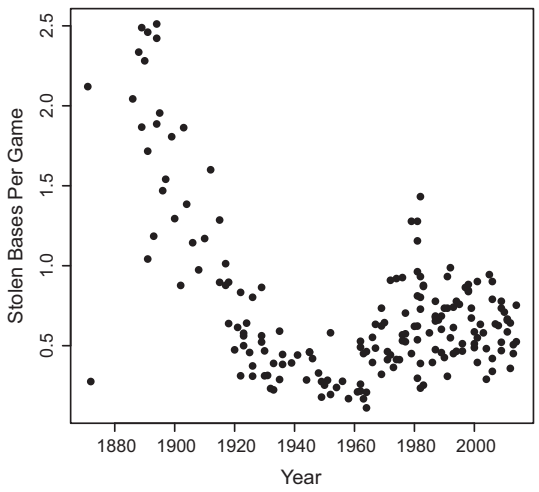


Figure 4.15. Scatterplot of stolen bases per game against season for 200 randomly selected baseball teams.

(a) Describe the basic pattern in the graph. Is the number of SBs per game increasing or decreasing over time?
(b) Is there a "straight-line" relationship between SB per game and year?
(c) What does this graph say about the importance of stolen bases in baseball today compared to the past?

4.9. For all pitchers in MLB history, we collected the year of birth and the throwing hand (left or right). For each birth year, the fraction of left-handed throwers is computed and Figure 4.16 displays a time series plot of the fraction as a function of the birth year.

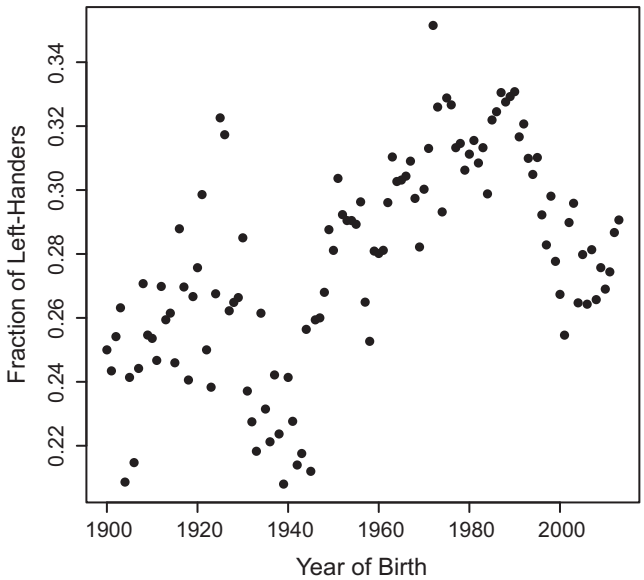


Figure 4.16. Time series plot of the fraction of left-handed throwers.

(a) Draw vertical lines on the scatterplot at the years 1900, 1920, 1940, 1960, 1980, and 2000. For each of the periods defined by the lines (1900–1920, 1920–1940, etc.), find the approximate average fraction of left-handers. Put your averages in the table below.

Era	Average Fraction of Left-Handed Pitchers
1900–1920	
1920–1940	
1940–1960	
1960–1980	
1980–2000	

(b) Based on your work, do you see any general pattern in the fraction of left-handers over time?

4.10. For each of a sample of 100 games played during the 2014 regular season, we recorded:

- TIME: the time of game (in minutes)
- RUNS: the number of runs scored in the game
- HITS: the number of hits in the game
- PITCHES: the number of pitches thrown in the game

We are interested in the relationship between the time of game and the three other variables. Scatterplots of TIME with RUNS, HITS, and PITCHES are displayed in Figure 4.17.

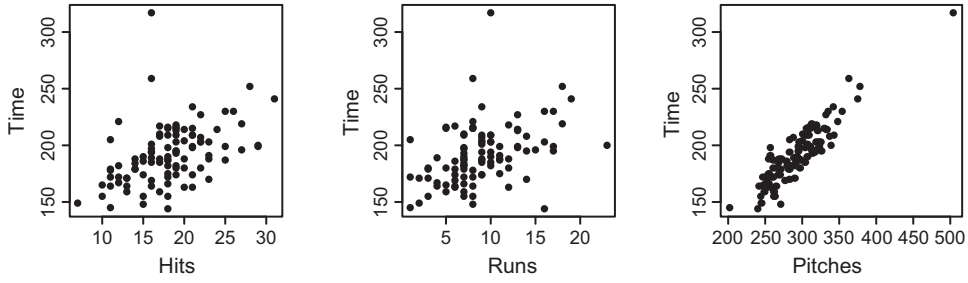


Figure 4.17. Scatterplots of time of game with runs scored, hits, and pitches for 100 games played during the 2014 season.

(a) From the scatterplots, describe each of the relationships (TIME and HITS, TIME and RUNS, TIME and PITCHES) as either positive, negative, or small. Which variable has the strongest relationship with time of game?

(b) The least-squares fit to the (TIME, PITCHES) data is

$$\text{TIME} = 14.2523 + 0.6061 \times \text{PITCHES}$$

Suppose that a game has 250 pitches. Use this least-square line to predict the length of the game (in minutes).

(c) Predict the length of the game if there are 350 pitches.

4.11. Two individual measures of pitching performance are the ERA (earned run average) and PCT (the percentage of pitching decisions won). A scatterplot of the ERA and PCT for Hall of Famer Walter Johnson for each of his 21 seasons in the major leagues is displayed in Figure 4.18.

(a) Describe the direction and strength of the association between PCT and ERA.

(b) Circle two points that correspond to two years where Johnson had the higher ERA and also the higher PCT in both years.

(c) Based on this graph, do you think that PCT is a good measure of a pitcher's performance? Why or why not?

(d) A famous pitcher who played primarily for Texas teams was inducted into the Hall of Fame but had a winning percentage that was close to 50%. Name this pitcher. Explain why he was elected in spite of his low winning percentage.

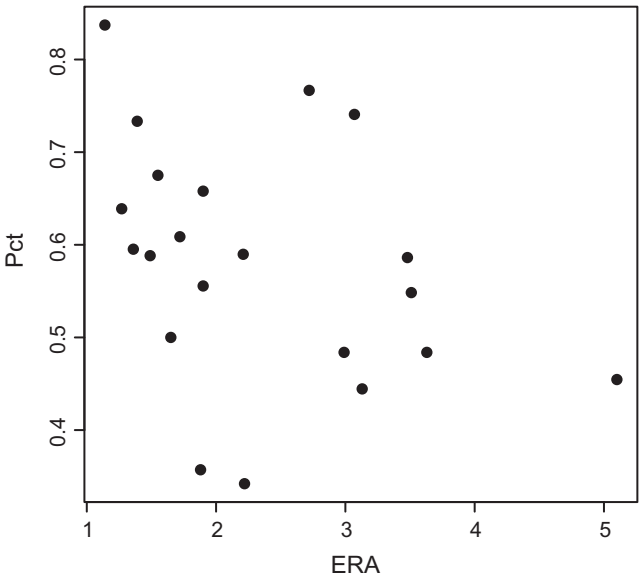


Figure 4.18. Scatterplot of season ERA and winning percentage for Walter Johnson.

4.12. Over the last hundred years, there has been a substantial growth in the United States population, and a corresponding large growth in the number of players and fans of Major League baseball. Table 4.19 gives the U.S. population and total MLB baseball attendance for the years 1900, 1910, . . . , 2010. The third column divides the baseball attendance by

Table 4.19. United States population, Major League baseball attendance, and ratio for the years 1900 through 2010

Year	Population	Baseball Attendance	Ratio = Attend. / Pop.
1900	76.2	1.8	0.024
1910	92.2	6.2	0.067
1920	106	9.1	0.086
1930	123.2	10.1	0.082
1940	132.2	9.8	0.074
1950	151.3	17.5	0.116
1960	179.3	19.9	0.111
1970	203.3	28.7	0.141
1980	226.5	43	0.190
1990	248.7	54.8	0.220
2000	281.4	71.4	0.254
2010	309.3	73.1	0.236

the population size. We see that in 2010, baseball attendance was equal to 23.6 per cent of the American population. Figure 4.19 displays a scatterplot of year and ratio.

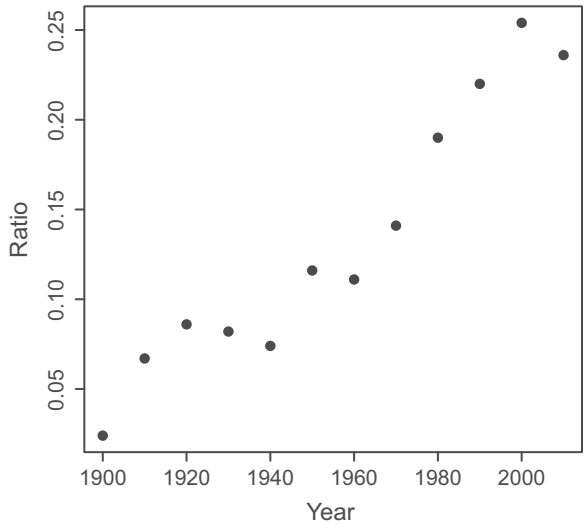


Figure 4.19. Time series plot of the ratio of Major League baseball attendance to population.

(a) Describe, using a few sentences, how the attendance/population ratio has changed over the years. Would it be accurate to say that the ratio has consistently increased from 1900 to 2010?
(b) Would it be reasonable to fit a least-squares line to these data? Why or why not?
(c) What would explain the dip in the ratio at the year 1940? (Hint: What was happening in the world scene at this time?)
(d) If one fits a least-squares line only for the 1940–2010 data, one obtains the equation

$$\text{RATIO} = -54869 + 02619 \times \text{YEAR}.$$

Use this equation to predict the attendance/population ratio for the year 2020.
(e) Would you expect the same straight-line relationship between RATIO and YEAR for the next 50 years? Why or why not?

4.13. Suppose you are interested in comparing the batting abilities of Jose Altuve and Nelson Cruz for the 2014 baseball season; the batting statistics for both players are given in the following table.

Player	AB	H	2B	3B	HR	BB	HBP	SF	1B	TB
Jose Altuve	660	225	47	3	7	36	5	5		
Nelson Cruz	613	166	32	2	40	55	5	5		

(a) Compute the number of singles (1B) and total bases (TB) for each player and place the values in the table.
(b) Compute the batting average (AVG), slugging percentage (SLG), on-base percentage (OBP), on-base plus slugging (OPS), and runs created (RC) measures for each player.
(c) Compare the two players with respect to ability to get on base, slugging ability, and total offensive contribution for this season.

(d) Altuve was the singles leader and Cruz was a home run leader in 2014. Which player would you rather have on your team? Why?

4.14. Find the batting statistics (AB, R, H, 2B, 3B, HR, BB) for the most recent year for two good hitters, one from the National League and one from the American League. Find the batting average (AVG), slugging percentage (SLG), on-base percentage (OBP), on-base plus slugging (OPS), and runs created (RC) measures for each player. Which player had a better offensive year? Why?

4.15. (Least-squares criterion) Suppose you are interested in a typical number of home runs hit by a National League team in 2014, and that you guess that the typical number is 150 home runs. We can evaluate the goodness of the guess 150 by means of the Root Mean Squared Error (RMSE) criterion. Table 4.20 summarizes the calculations of RMSE using the 2014 National League team home run data. (The "Residual" column contains the difference between the HR value and the guess, and the "Residual^2" column contains the square of the residual.) Fill in the missing cells in the table and compute the RMSE.

Table 4.20. Summary of RMSE calculations for the 2014 National League team home run data using the guess 150

Team	HR	Guess	Residual	Residual^2
ARI	118	150		
ATL	123	150	−27	729
CHN	157	150	7	49
CIN	131	150		
COL	186	150	36	1296
LAN	134	150	−16	256
MIA	122	150	−28	784
MIL	150	150	0	0
NYN	125	150		
PHI	125	150	−25	625
PIT	156	150		
SDN	109	150	−41	1681
SFN	132	150	−18	324
SLN	105	150	−45	2025
WAS	152	150	2	4
Sum				9819

4.16. (Exercise 4.15 continued.) In the previous exercise, it turns out that the best guess at this typical number of home runs is the mean, which is found by adding up all of the home run numbers and dividing by the number of teams (15). Here the mean is $\bar{x} = (118 + 123 + 157 + \cdots + 105 = 152)/15 = 135$.

(a) Compute the RMSE of the mean by completing Table 4.21.

(b) Compare the RMSE of the mean with the RMSE of the guess 150 that you found in Exercise 4.16. Which is a better estimate at a typical home run total?

Table 4.21. Summary of RMSE calculations for the 2014 National League team home run data using the mean estimate 135

Team	HR	Guess	Residual	Residual^2
ARI	118	135	−17	289
ATL	123	135	−12	144
CHN	157	135	22	484
CIN	131	135		
COL	186	135	51	2601
LAN	134	135	−1	1
MIA	122	135		
MIL	150	135		
NYN	125	135	−10	100
PHI	125	135	−10	100
PIT	156	135	21	441
SDN	109	135		
SFN	132	135	−3	9
SLN	105	135	−30	900
WAS	152	135	17	289
Sum				6444

4.17. (Exercise 4.1 continued.) The "AVG" and "R.G" columns of Table 4.22 give the batting averages and runs per game for the 2014 NL teams. The "Predicted" column gives the

Table 4.22. RMSE calculations using AVG to predict runs per game

Team	R.G	AVG	Predicted	Residual	Residual^2
ARI	3.80	0.248	4.24		
ATL	3.54	0.241	3.99		
CHN	3.79	0.239	4.18	−0.39	0.1521
CIN	3.67	0.238	4.18	−0.51	0.2601
COL	4.66	0.276	4.67	−0.01	0.0001
LAN	4.43	0.265	3.95		
MIA	3.98	0.253	4.38		
MIL	4.01	0.250	4.30	−0.29	0.0841
NYN	3.88	0.239	4.20	−0.32	0.1024
PHI	3.82	0.242	4.01	−0.19	0.0361
PIT	4.21	0.259	3.99	0.22	0.0484
SDN	3.30	0.226	3.99	−0.69	0.4761
SFN	4.10	0.255	4.06	0.04	0.0016
SLN	3.82	0.253	4.24	−0.42	0.1764
WAS	4.23	0.253	4.30	−0.07	0.0049

predicted runs per game using the least-squares fit $\text{RUNS} = -1.00 + 20.48\text{AVG}$. The "Residual" column gives the residual for each team, and the "Residual^2" column contains the squared value of the residual.

Table 4.23. RMSE calculations using SLG to predict runs per game

Team	R.G	SLG	Predicted	Residual	Residual^2
ARI	3.80	0.376	3.87		
ATL	3.54	0.360	3.66		
CHN	3.79	0.385	3.98		
CIN	3.67	0.365	3.73		
COL	4.66	0.445	4.76	−0.10	0.0100
LAN	4.43	0.406	4.26	0.17	0.0289
MIA	3.98	0.378	3.89	0.09	0.0081
MIL	4.01	0.397	4.14	−0.13	0.0169
NYN	3.88	0.364	3.71	0.17	0.0289
PHI	3.82	0.363	3.70	0.12	0.0144
PIT	4.21	0.404	4.23	−0.02	0.0004
SDN	3.30	0.342	3.43	−0.13	0.0169
SFN	4.10	0.388	4.02	0.08	0.0064
SLN	3.82	0.369	3.78	0.04	0.0016
WAS	4.23	0.393	4.09	0.14	0.0196

(a) Fill in the missing cells in the "Residual" and "Residual^2" columns.
(b) Compute the Root Mean Square (RMSE) value. This is a measure of the goodness of using AVG as a predictor of runs scored per game.

4.18. (Exercise 4.1 continued.) Suppose that the slugging percentage (SLG) is used, instead of AVG, to predict runs per game. The least-squares fit is given by

$$\text{RUNS} = -0.99 + 12.92\text{SLG}.$$

Table 4.23 contains the predicted runs per game, the residuals, and the residuals squared using SLG as a predictor.
(a) Fill in the missing cells in the "Residual" and "Residual^2" columns.
(b) Compute the RMSE value for the SLG fit. Compare this value with the RMSE value from the AVG fit that you computed in Exercise 4.17.
(c) Construct parallel boxplots of the residuals from the SLG fit and the residual from the AVG fit (Exercise 4.17). Looking at the boxplot display, which is the better predictor AVG or SLG? Why?

4.19. (Exercise 4.1 continued.) Suppose that we estimate the runs per game using the derived statistic OPS = OBP + SLG as a predictor. The least-squares fit is given by RUNS = $-2.56 + 9.38$OPS.
(a) Suppose a team has an OBP = .360 and SLG = .700. What is the value of the team's OPS?
(b) For each team in the 2014 NL, Table 4.24 computes the OPS statistic, the predicted runs per game (using the least-squares formula), the residual and the residual squared.
(c) Fill in the blank cells in the table.
(d) Compute the RMSE. Compare the RMSE value with the RMSE values using the predictors AVG and SLG. (Exercises 4.3 and 4.4.) Which is the best predictor among AVG, SLG, and OPS? Why?

Table 4.24. RMSE calculations using OPS to predict runs per game

Team	R.G	OPS	Predicted	Residual	Residual^2
ARI	3.80	0.678	3.80	0.00	0.0000
ATL	3.54	0.665	3.68	−0.14	0.0196
CHN	3.79	0.685	3.86	−0.07	0.0049
CIN	3.67	0.661	3.64		
COL	4.66	0.772	4.68		
LAN	4.43	0.739	4.37		
MIA	3.98	0.695	3.96		
MIL	4.01	0.708	4.08	−0.07	0.0049
NYN	3.88	0.672	3.74	0.14	0.0196
PHI	3.82	0.665	3.68	0.14	0.0196
PIT	4.21	0.734	4.32	−0.11	0.0121
SDN	3.30	0.634	3.39	−0.09	0.0081
SFN	4.10	0.699	4.00	0.10	0.0100
SLN	3.82	0.689	3.90	−0.08	0.0064
WAS	4.23	0.714	4.14	0.09	0.0081

4.20. Table 4.25 contains the proportion of batted balls that were "hard" and the slugging percentage (SLG) for 14 seasons of Albert Pujols career.

Table 4.25. Proportion of hard batted balls and the slugging percentages for 14 seasons of the career of Albert Pujols

Season	Hard	SLG	Season	Hard	SLG
2002	0.273	0.561	2009	0.406	0.658
2003	0.332	0.667	2010	0.424	0.596
2004	0.352	0.657	2011	0.305	0.541
2005	0.389	0.609	2012	0.335	0.516
2006	0.342	0.671	2013	0.362	0.437
2007	0.400	0.568	2014	0.361	0.466
2008	0.429	0.653	2015	0.338	0.527

(a) Before any exploration, do you believe there is a relationship between the two variables? Explain.
(b) Construct a scatterplot of "Hard" (horizontal) against SLG. Does the pattern in this plot confirm with your beliefs about the relationship in part (a)?
(c) Find a best (least-squares) fit to these data. If Pujols hits 40% hard batted balls next season, predict the value of his slugging percentage.

4.21. Table 4.26 contains the number of pitches and the game duration (in minutes) for 20 randomly selected games from the 2014 season.
(a) Construct a scatterplot of Pitches (horizontal) against Duration (vertical) to confirm that there is a strong relationship.

Table 4.26. Number of pitches and game duration in minutes for twenty games in the 2015 season

Game	Pitches	Duration	Game	Pitches	Duration
1	278	182	11	318	203
2	316	197	12	273	216
3	287	168	13	267	163
4	225	142	14	254	168
5	270	170	15	297	186
6	266	173	16	284	173
7	339	221	17	292	177
8	348	235	18	292	186
9	279	181	19	250	167
10	343	226	20	276	185

(b) A best-fitting line has the equation

$$\text{Duration} = -0.65 + 0.65 \times \text{Pitches}.$$

If a game has 50 more pitches, how much longer do you predict the game will be?

(c) Predict the length of a game with 250 pitches.

4.22. Table 4.27 presents a table which categorizes all balls put in play in the 2014 season by type of hit (popup, groundball, fly ball, and liner) and the outcome.

Table 4.27. All balls put in play in the 2014 season categorized by the type of batted ball and the outcome

	out	single	double	triple	home run
Popup	8823	146	42	1	0
Groundball	46003	13877	1073	70	0
Fly Ball	23111	957	1356	240	2873
Liner	11261	13440	5641	537	1313

(a) Find the proportion of batted balls that are popups, groundballs, fly balls, and liners.

(b) For each type of hit, find the proportion of outs. Which type of batted ball is most likely to result in an out?

(c) Which type of batted ball is most likely to be a home run? Support your answer by a suitable calculation.

Further Reading

Devore and Peck (2011) and Moore, McCabe and Craig (2012) describe basic descriptive tools for understanding their relationship between two measurement variables. Albert (1998), Bennett (1998), Thorn and Palmer (1985), and Albert and Bennett (2003), Chapters 6, 7, 8, describe a number of ways of measuring offensive performance. The mean squared error criterion for

evaluating a particular batting measure is used in Chapter 6 of Albert and Bennett (2003). The Pythagorean Formula for relating a team's win/losses with the number of runs scored/allowed is described in James (1982, 2001). A nice discussion of the regression-to-the-mean effect is given in Berry (1996).

10.1090/text/034/05

5

Introduction to Probability Using Tabletop Games

What's On-Deck?

In this chapter we introduce basic concepts of probability by using tabletop games. To begin, we introduce the relative frequency notion of probability by using Chris Davis' hitting log for the 2013 season. During a plate appearance, Davis can either hit a home run or not, and we assume that the chance that he hits a home run is p. We learn about the value of p by observing his hitting performance over a number of games and we can estimate his home run probability by calculating the relative frequency of home runs. In Case Study 5.2, we explore a tabletop dice game *Big League Baseball*, and find probabilities of different events, say home runs, singles, and outs, by finding probabilities of the sum of two rolls of the dice. The game *All-Star Baseball* is a more elaborate game in that the batting performance of each hitter is modeled using a separate spinner. In Case Study 5.3, we show how career statistics for a player can be used to compute probabilities that the player gets a single, double, etc. and these probabilities are used to compute areas of the random spinner. We conclude our discussion by describing a more sophisticated game, *Strat-O-Matic Baseball*, that uses four dice and models the abilities of each pitcher and each hitter by a separate card. We consider a classic matchup in this game Mark McGwire against Greg Maddux and show how one can compute the probability of McGwire hitting a home run using the theorem of total probabilities.

5.1 What is Chris Davis' Home Run Probability?

Topics Covered: Relative frequency interpretation of probability, law of large numbers.

In this case study, we begin our discussion on probability.

What Is a Probability?

First, we recognize that life is full of uncertain events. For example

- Who will win the next World Series?
- Will you retire before the age of 60?
- Will a Major League player ever break Joe DiMaggio's record of hitting in 56 consecutive games?

A probability is a way of measuring the uncertainty that we see. We can define a probability scale from 0 to 1 and any event can be assigned a number in this range.

- We assign a probability of 0 to an event that we are certain will not occur. For example, the probability that the Phillies will meet the Mets in the World Series is zero since the World Series matches the winners of the National League and American League playoffs and the Phillies and Mets both play in the National League.
- On the other hand, we assign a probability of 1 to an event that we are sure will occur, such as "a Major League Baseball game will finish in ten hours."
- What if we assign a probability of .5? Suppose we toss a coin. We give the event "heads" a probability of .5 if we think that the events "heads" and tails, or "not heads", have the same chance of occurring.

One way of thinking about probabilities is the following relative frequency interpretation.

An experiment is a process where the outcome (or result) is unknown. We let an event be a collection of outcomes, and we are interested in computing the probability of the event.

Say that we can repeat the experiment many times under similar conditions. (For example, if the experiment is tossing a coin, then suppose that we can toss the coin repeatedly under similar conditions.)

Then the Prob(event) is approximately the relative frequency of the event. If N is the number of replications, then

$$\text{Prob(event)} \approx \frac{\text{\# of times event occurs}}{N}.$$

To illustrate this interpretation, suppose we are interested in estimating the probability that Chris Davis in 2013 would hit a home run in a single plate appearance. We assume that Davis comes to bat multiple times during the season under the same conditions. (Note that this is a questionable assumption, but it simplifies our discussion.) For our purposes, there are two results of this plate appearance he either hits a home run or he doesn't, and the chance that he hits a home run in this particular season is measured by a probability.

We don't know the value of Davis' home run probability, but we can learn about it by watching Davis perform during the 2013 baseball season. Table 5.1 gives the number of plate appearances (PA) and the number of home runs (HR) Davis hit for each of the 160 games he played in 2013.

In the first game, Davis came to bat five times and had a single home run. At that point, the current estimate of Davis' home run probability is

$$\hat{p}_{\text{HR}} = \frac{1}{5} = .200.$$

Certainly, this is not a great estimate of Davis' home run probability since it is based on only five plate appearances. Next, suppose it is the middle of April and we ve now watched Davis play ten games. In games 1–10, he has 42 PA and 6 HR our new estimate of Davis' home run probability is

$$\hat{p}_{\text{HR}} = \frac{6}{42} = .143.$$

Table 5.1. Plate appearances and home runs for Chris Davis for each game played during the 2013 baseball season

Game	PA	HR	Game	PA	HR	Game	PA	HR	Game	PA	HR
1	5	1	41	5	1	81	4	2	121	5	0
2	4	1	42	4	0	82	4	1	122	5	1
3	4	1	43	5	1	83	4	0	123	5	0
4	5	1	44	4	0	84	4	1	124	4	0
5	4	0	45	4	1	85	4	0	125	4	1
6	4	0	46	4	1	86	3	0	126	5	0
7	4	0	47	5	1	87	4	1	127	4	0
8	4	1	48	4	0	88	4	0	128	4	0
9	4	1	49	5	0	89	5	0	129	3	0
10	4	0	50	5	0	90	4	0	130	4	1
11	4	0	51	4	1	91	4	0	131	5	0
12	4	0	52	4	2	92	3	1	132	4	0
13	4	0	53	3	0	93	4	1	133	4	0
14	4	0	54	4	0	94	4	1	134	4	0
15	4	0	55	4	0	95	4	1	135	5	0
16	4	0	56	4	1	96	4	0	136	5	0
17	4	1	57	4	0	97	5	0	137	4	0
18	4	0	58	5	0	98	5	0	138	3	0
19	4	0	59	5	0	99	5	0	139	4	1
20	4	0	60	4	0	100	4	0	140	5	0
21	5	0	61	3	0	101	4	0	141	4	0
22	5	1	62	5	0	102	4	0	142	4	0
23	4	0	63	3	0	103	4	0	143	4	1
24	4	0	64	4	0	104	4	0	144	4	0
25	5	1	65	4	1	105	4	0	145	4	0
26	3	0	66	7	0	106	4	1	146	5	1
27	5	0	67	4	1	107	4	0	147	4	0
28	4	0	68	4	0	108	4	1	148	4	0
29	4	0	69	4	1	109	4	1	149	4	1
30	2	0	70	4	1	110	4	0	150	6	0
31	4	0	71	4	0	111	4	0	151	4	0
32	4	0	72	5	2	112	4	0	152	8	0
33	4	0	73	4	1	113	5	1	153	4	0
34	4	1	74	4	0	114	5	0	154	4	0
35	6	0	75	4	0	115	5	1	155	5	1
36	4	0	76	4	0	116	5	0	156	5	0
37	4	1	77	4	1	117	4	1	157	4	0
38	4	0	78	4	0	118	4	1	158	3	1
39	5	0	79	4	0	119	6	0	159	4	0
40	4	0	80	3	0	120	4	0	160	1	0

This is likely a better estimate of Davis' ability since it is based on a greater number of plate appearances. Suppose that we compute Davis' current 2013 home run rate after each of his 153 games. Figure 5.1 plots the home run rate (our probability estimate) against the game number.

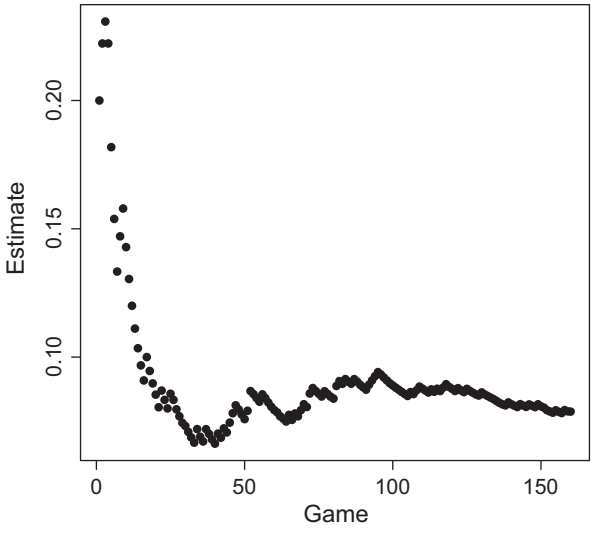


Figure 5.1. Plot of home run rate against game number for Chris Davis during the 2013 baseball season.

Note that for early game numbers, the probability estimate shows a lot of fluctuation. But after about game 80, the home run rate appears to settle down to about the value 0.08. This illustrates the relative frequency of probability. As Davis gets more and more plate appearances, the relative frequency of his home runs settles down and approaches his probability of a home run for 2013. After 160 games, he hit 53 home runs in 673 plate appearances for an estimate of $53/673 = 0.079$. This value is a good estimate of Davis' home run ability for the 2013 season.

5.2 *Big League Baseball*

Topics Covered: Sample space, experiments with equally likely outcomes, experiment of rolling two dice, finding probabilities of events.

Assigning probabilities is generally hard to do if we are not able to repeat the experiment many times. But probabilities can be assigned for simple experiments, such as those involving dice and cards, and we talk about several of these experiments here. We describe a simple tabletop game, *Big League Baseball*, made in the 1960's based on dice rolls.

Suppose We Roll a Fair Die

We first think of the possible outcomes of rolling a single die. There are six possibilities:

ROLL	1	2	3	4	5	6

We call the collection of all possible outcomes the sample space. To assign probabilities, we assume that each roll outcome has the same probability. (In other words, we assume that the

outcomes are equally likely. We want to assign a positive number to each outcome such that the sum of all the probabilities assigned will be equal to 1, since we are certain there will be one of the six outcomes in each roll. It should be clear that we should assign a probability of 1/6 to each outcome:

ROLL	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6

Next suppose we roll two dice. It will be convenient to distinguish the dice, so we will take one die to be brown and one to be orange. (Brown and orange are the colors of the author's university.)

1. How many outcomes are there in this experiment? We know from above that there are six possibilities for the result on the brown die. For each result on the brown die, say BROWN $=2$, there are six possible outcomes for the orange die. So the number of possible outcomes for two dice is
$$6 \times 6 = 36.$$
2. If we assume that each of the 36 possible rolls of two dice have the same probability, then (by the same logic as before) we assign a probability of $1/36$ to each outcome. So, for example,
$$\text{Prob}(5 \text{ on brown and } 2 \text{ on orange}) = 1/36.$$

Big League Baseball was a dice tabletop baseball game made by Sycamore Games of Lima, Ohio in the 1960's. This game is based on rolling 1 red die and 2 white dice. One first rolls the red die to get the pitch:

- If 1 or 6 is rolled, a fair ball is hit.
- If 2 or 3 is rolled, a ball is pitched
- If 4 or 5 is rolled, a strike is pitched.

Let's find the probabilities of some outcomes:

1. The probability of pitching a ball, Prob(ball), is the same as the probability of rolling a 2 or 3. We find the probability of a set of outcomes by adding the probabilities of the outcomes. So
$$\text{Prob(ball)} = \text{Prob}(2 \text{ or } 3) = \text{Prob}(2) + \text{Prob}(3) = 1/6 + 1/6 = 2/6.$$
2. The probability of pitching a ball or a strike is found the same way:
$$\text{Prob(ball or strike)} = \text{Prob}(2, 3, 4, 5) = \text{Prob}(2) + \text{Prob}(3) + \text{Prob}(4) + \text{Prob}(5) = 4/6.$$
3. What is the probability that a strike is not thrown? We note that a strike is not thrown if a ball is pitched or a fair ball is hit, so
$$\text{Prob(no strike)} = \text{Prob(ball or fair ball)} = \text{Prob}(1, 6, 2, 3) = 4/6.$$

In *Big League Baseball*, if a ball is in play, then two white dice are rolled. Table 5.2 shows the outcomes for each possibility of rolling two dice.

Table 5.2. Outcomes of the rolls of two dice

		2nd Die					
		1	2	3	4	5	6
1st Die	1	Single	Out	Out	Out	Out	Error
	2	Out	Double	Single	Out	Single	Out
	3	Out	Single	Triple	Out	Out	Out
	4	Out	Out	Out	Out	Out	Out
	5	Out	Single	Out	Out	Out	Single
	6	Error	Out	Out	Out	Single	Home Run

Since each cell in the table is assigned a probability of $1/36$, we can compute several probabilities of interest:

1. Prob(Home run) $= 1/36$. (There is just one way to roll a home run.)
2. Prob(Single). We see from the table that there are 7 ways of getting a single each outcome has probability $1/36$, so Prob(Single) $= 7/36$.
3. Prob(Hit) $= 10/36$. (From the table, we see 10 ways of getting a hit.)
4. Prob(Out) $= 24/36$. (It is easiest to note that there are 12 ways of getting on base by a hit or error, and therefore $36 - 12 = 24$ ways of getting an out.)

Is *Big League Baseball* a realistic baseball game? In other words, does this game provide a good representation of real baseball? Of course not—it would be insulting to the game to think that we could simulate real baseball by only using three dice. This game assumes many unrealistic things, including

- that all batters have the same ability,
- that all pitchers have the same ability,
- that it is equally likely to add a strike or a ball to the pitch count. (This is not realistic in real baseball, it is more common to pitch a strike than a ball.)

Although this game is a bit unrealistic, it is a nice first attempt to simulate a baseball game. The results of a game are partly due to chance variation, and this game introduces chance variation by the use of dice. This game provides a useful comparison to the more sophisticated baseball games that are described in the next two case studies.

5.3 *All-Star Baseball*

Topics Covered: Spinner as a randomization device, probabilities represented by areas of the spinner, multinomial experiment.

We next consider a more sophisticated baseball game, *All-Star Baseball*. This game was popular in the 1960's and 1970's and was played by the author when he was young. This game is based on using a spinner. Each hitter is represented by a circular spinner where different areas of the spinner correspond to different play outcomes. An *All-Star Baseball* spinner for the great baseball player Babe Ruth is displayed in Figure 5.2. The areas of the spinner correspond to the probabilities of the different outcomes. This randomization device has more flexibility than dice, since we can represent a greater range of probabilities in the outcomes.

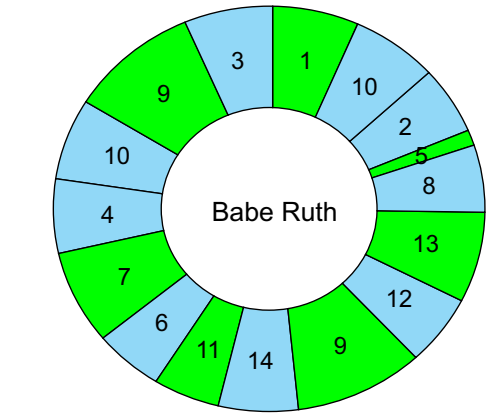


Figure 5.2. *All-Star Baseball* spinner for Babe Ruth.

We illustrate constructing a spinner for one of my favorite players, Mike Schmidt. By the way, it is doubtful that one spinner for an entire career is realistic, but we need to start somewhere. We start with a very simple model one spinner for an entire career, regardless of age, ballpark, opposing pitcher and defense. In later chapters, we will examine the benefits and drawbacks of using more complicated models. We make this spinner in two steps. First, we calculate approximate probabilities for the different outcomes (1B, 2B, 3B, HR, BB, Out) that can happen when Schmidt comes to bat. Then we shade regions on the spinner where the areas of the regions correspond to the probabilities. We start with Schmidt's career statistics shown below.

AB	R	H	2B	3B	HR	SO	BB	AVG	OBP	SLG
8352	1506	2234	408	59	548	1883	1507	.267	.380	.527

We want to classify all plate appearances (PAs) into the outcomes 1B, 2B, 3B, HR, BB, and Outs. (We ignore events like HBP and Sacrifices, since they are relatively insignificant compared to the other outcomes.) We put the counts we know from the career data in Table 5.3.

Table 5.3. Counts of career offensive statistics for Mike Schmidt

PLAY	COUNT
PA	
AB	8352
H	2234
1B	
2B	408
3B	59
HR	548
BB	1507
OUTS	

We complete Table 5.3 in Table 5.4, using the computations below.

- We get PAs by adding at-bats (AB) and walks (BB).

$$\text{PA} = \text{AB} + \text{BB} = 8352 + 1507 = 9859.$$

- We get singles (1B) by adding the doubles (2B), triples (3B), and home runs (HR), and subtracting the total from hits (H).

$$1\text{B} = \text{H} - (2\text{B} + 3\text{B} + \text{HR}) = 2234 - (408 + 59 + 548) = 1219.$$

- We get OUTS by subtracting hits (H) from at-bats (AB).

$$\text{OUTS} = \text{AB} - \text{H} = 8352 - 2234 = 6118.$$

We next change these counts to approximate probabilities in Table 5.5 by dividing each count (1B, 2B, 3B, HR, BB, OUTS) by the number of PAs. We check if we did this right by seeing if the sum of probabilities is equal to one.

Table 5.4. Counts of career offensive statistics for Mike Schmidt with PA, singles, and outs included

PLAY	COUNT
PA	9859
AB	8352
H	2234
1B	1219
2B	408
3B	59
HR	548
BB	1507
OUTS	6118

Table 5.5. Computation of event probabilities from Mike Schmidt's career offensive statistics

PLAY	COUNT	PROPORTION
PA	9859	xxxx
AB	8352	xxxx
H	2234	xxxx
1B	1219	0.124
2B	408	0.041
3B	59	0.006
HR	548	0.056
BB	1507	0.153
OUTS	6118	0.621
		1

To make our spinner, we start with a blank circle and shade areas of the regions of single (1B), double (2B), etc., that correspond to these probabilities. To make the construction process easier, one can divide the circular region into 36 equal regions. We convert the probabilities into number of regions by multiplying the probability by 36 and rounding the result to the nearest whole number:

$$\text{Number of Regions} = \text{round}\,(36 \times \text{probability}).$$

To illustrate, we multiply the probability of a single by 36 to get

$$\text{Number of regions} = 36 \times 0.124 = 4.46.$$

We round this to the nearest integer, getting four regions. If we do this calculation for all events, we get the region numbers in Table 5.6.

This didn't quite work, since the total number of regions is 35, not 36. So we make a small adjustment I changed the number of regions of 1B from 4 to 5 to make the sum of regions add

Table 5.6. Computation of spinner region numbers from Mike Schmidt's career offensive statistics

PLAY	COUNT	PROPORTION	REGIONS
PA	9859	xxxx	xxxx
AB	8352	xxxx	xxxx
H	2234	xxxx	xxxx
1B	1219	0.124	4
2B	408	0.041	1
3B	59	0.006	0
HR	548	0.056	2
BB	1507	0.153	6
OUTS	6118	0.621	22
		1	35

up to 36. (I adjusted the number of regions of 1B instead of HR, say, since this adjustment has a modest change in the single probability.)

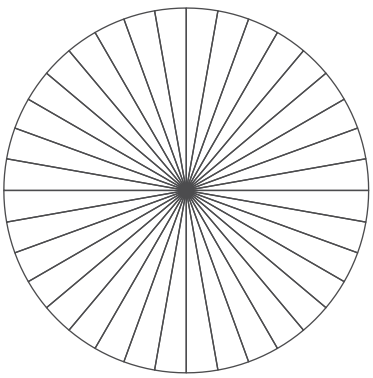


Figure 5.3. Blank spinner divided into 36 equal-size regions.

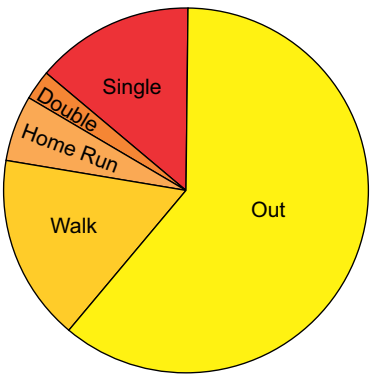


Figure 5.4. Spinner using Mike Schmidt's career hitting statistics.

We completed the calculations for our spinner and now can begin the construction process. We take a blank spinner, shown in Figure 5.3, and color-code the events according to the work in Table 5.6. So we color five regions black (corresponding to single), one region purple (corresponding to double), two regions light purple, six regions green (walk) and 22 regions white (out). When we are done we get the spinner in Figure 5.4. We will be using a spinner as a basic probability model in our study of inference in Chapter 7.

5.4 *Strat-O-Matic Baseball*

Topics Covered: Probabilities of the sum of two dice, theorem of total probabilities, conditional probability.

We conclude our discussion of tabletop games by briefly describing my favorite game, *Strat-O-Matic Baseball*. It's a game played with three dice (like *Big League Baseball*). The game is

much more realistic than the two previous games discussed, since it models the different abilities of pitchers as well as hitters. The *All-Star Baseball* game described in Case Study 5.3 doesn't take into account the abilities of pitchers.

We introduce this game by considering a classic matchup between Mark McGwire and Greg Maddux. (McGwire was a great power hitter during the so-called "Steroid Era" of baseball and Maddux was a great control pitcher in recent baseball history.) Each player is represented by a card—the Mark McGwire and Greg Maddux cards are shown in Tables 5.7 and 5.8. We first roll a single white die if the result is 1, 2, 3, we look at McGwire's card; otherwise we look at Maddux's card. Then we roll two red dice and observe the sum. The play is determined by reading the line (corresponding to the sum) on the pitcher's or hitter's card. Sometimes the outcome is not determined by the roll of the white and red dice and a twenty-sided die must be rolled to determine the play result.

Table 5.7. Mark McGwire's *Strat-O-Matic* card from the 1998 baseball season

1	2	3
2-lineout	2-flyball	2-WALK
3-strikeout	3-WALK	3-WALK
4-flyball	4-HOMERUN	4-strikeout
5-WALK	5-HOMERUN	5-strikeout
6-WALK	6-HOMERUN	6-strikeout
7-WALK	7-HOMERUN	7-strikeout
8-WALK	8-HOMERUN	8-strikeout
9-WALK	1-10	9-strikeout
10-flyball	flyball	10-strikeout
11-groundball	12-20	11-WALK
12-flyball	9-WALK	12-flyball
	10-DOUBLE	
	1-11	
	SINGLE	
	12-20	
	11-SINGLE	
	1-6	
	lineout	
	7-20	
	12-WALK	

Table 5.8. Greg Maddux's *Strat-O-Matic* card from the 1998 baseball season

4	5	6
2-lineout	2-flyball	2-strikeout
3-groundball	3-flyball	3-groundball
4-flyball	4-groundball	4-flyball
5-groundball	5-strikeout	5-HOMERUN
6-popout	6-strikeout	1
7-groundball	7-DOUBLE	flyball
8-flyball	1-7	2-20
9-strikeout	SINGLE	6-SINGLE
10-groundball	8-20	7-SINGLE
11-flyball	8-flyball	1-12
12-flyball	9-strikeout	lineout
	10-catcher	13-20
	11-groundball	8-strikeout
	12-groundball	9-strikeout
		10-groundball
		11-groundball
		12-lineout

Let's illustrate playing this game for two plate appearances.

1. For the first plate appearance, we roll a 2 on the white die and so we refer to the 2 column of McGwire's card. We roll the two red dice and get a 2 and 3 for a sum of 5. We look at the number 5 in the 2 column and read the result HOMERUN! Mac has hit a home run against Greg Maddux.
2. For the second plate appearance, we roll a 5 on the white die and we refer to the 5 column of Maddux's card. The roll of the red dice is 5 and 2 for a sum of 7. Looking at the 7 line, we see

that the result is DOUBLE if the roll of the twenty-sided die is between 1 and 7 and SINGLE if the die roll is between 8 and 20. We roll the twenty-sided die and get a 10—McGwire has hit a single against Maddux.

To really get an understanding how the game works, we need to calculate probabilities for the sum of two dice. As in Case Study 5.2, we distinguish between the two red dice. The 36 possible outcomes are displayed in Table 5.9.

Table 5.9. Outcomes of the rolls of two dice

		2nd Die					
		1	2	3	4	5	6
1st Die	1	x	x	x	x	x	x
	2	x	x	x	x	x	x
	3	x	x	x	x	x	x
	4	x	x	x	x	x	x
	5	x	x	x	x	x	x
	6	x	x	x	x	x	x

Since each possible outcome has the same chance, we assign a probability of $1/36$ to each outcome. So

$$\text{Prob(1st die is 4, 2nd die is 3)} = 1/36, \quad \text{Prob(1st die is 5, 2nd die is 6)} = 1/36.$$

We are interested in probabilities about the sum of the two dice.

First, we think of possible values of the sum. Table 5.10 shows the value for the sum for each possible rolls of the two dice. Looking at the table, we see that the possible sums are

$$2, 3, 4, \ldots, 12.$$

Table 5.10. Table of the sum of the rolls of two dice for all possible outcomes

		2nd Die					
		1	2	3	4	5	6
1st Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

We find the probabilities of the different sums by adding the probabilities of the individual outcomes for the two dice. For example, suppose we wish to compute the probability that the sum is 4. We first note that we can get a sum of 4 three ways:

- (1st die is 1, 2nd die is 3)
- (1st die is 2, 2nd die is 2)
- (1st die is 3, 2nd die is 1)

and since the probability of each outcome is $1/36$,

$$\text{Prob(sum is 4)} = 1/36 + 1/36 + 1/36 = 3/36.$$

If we continue this way, we obtain the probability table for the sum shown in Table 5.11.

Table 5.11. Probability distribution for the sum of two fair dice

Sum	Probability
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Now we can compute some probabilities of the *Strat-O-Matic* game:

Question: Looking at McGwire's card, if we roll a 1 on the white die, what is the probability that he walks?

Answer: Looking at McGwire's card, we see that he walks if the sum of the dice is 5, 6, 7, 8, 9 so

$$\text{Prob(walk)} = \text{Prob(sum is 5)} + \text{Prob(sum is 6)} + \cdots + \text{Prob(sum is 9)} = 24/36.$$

Question: If we roll a 3 on the white die, what is the probability that he strikes out?

Answer: $\text{Prob(strikeout)} = \text{Prob(sum is 4 through 10)} = 30/36$.

We use similar logic for finding probabilities off of Maddux's card. For example, if we roll a 4 on the white die (look at Maddux's card), we see that a fly ball results if we roll a sum equal to 4, 8, 11. So

$$\text{Prob(flyball)} = \text{Prob(sum is 4)} + \text{Prob(sum is 8)} + \text{Prob(sum is 11)} = 10/36.$$

Now, actually we are really interested in the probability that McGwire hits a home run if Mac is facing Greg Maddux. Figure 5.5 shows, using what is commonly called a tree diagram, all of the ways McGwire can hit a home run off of Maddux.

- He can hit a home run if one rolls a 2 on the white die (look at column 2 of McGwire's card), and rolls a sum of 4, 5, 6, 7, or 8 on the red dice. If one rolls an 8, then one uses the 20-sided die and a roll between 1 and 10 results in a home run.
- Mac also hits a home run if one rolls 6 on the white die (look at column 6 of Maddux's card), rolls a sum of 5 on the red dice, and rolls a 1 on the 20-sided die. In the diagram, the branches of the tree are labeled with the corresponding probabilities.

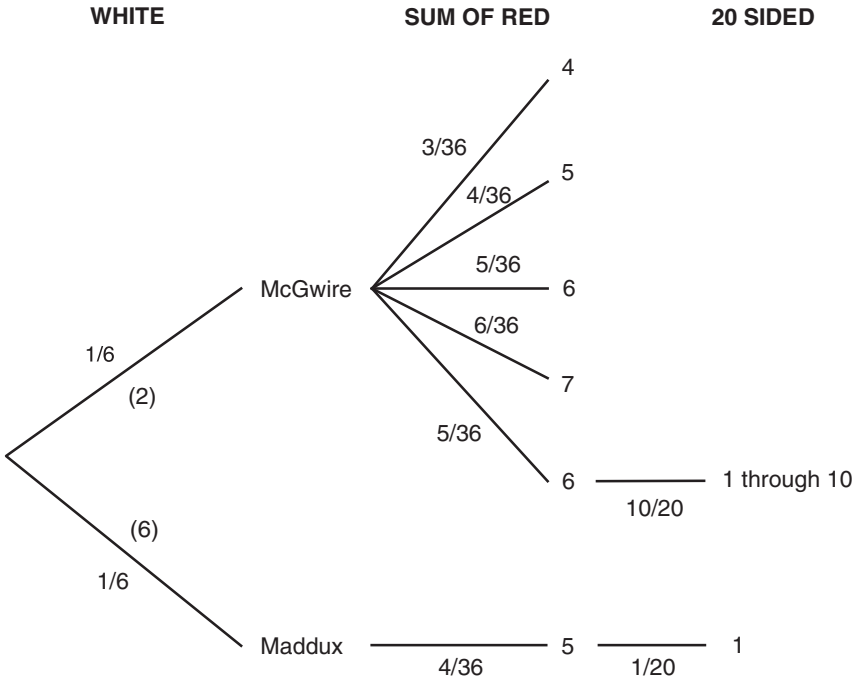


Figure 5.5. Tree diagram to illustrate the computation that Mark McGwire hits a home run against Greg Maddux.

By use of the multiplication rule, we multiply the probabilities across a particular branch to find the probability of a particular outcome of the white, red, and twenty-sided dice. For example, the probability

Prob(White die is 6 AND the sum of the red dice is 5 AND the twenty-sided die is 1)

is found by multiplying the conditional probabilities

$$\begin{aligned}\text{Prob(White die is 6)} &\times \text{Prob(Sum of red dice is 5 IF white die is 6)} \\ &\times \text{Prob(Twenty-sided die is 1 IF white die is 6 and sum of red dice is 5)} \\ &= (1/6) \times (4/36) \times (1/20).\end{aligned}$$

We find the probability of Mac hitting a home run off of Greg Maddux by

- multiplying the probabilities along each branch for each possible way of Mac hitting a home run,
- adding the products.

We then find the probability to be

$$\begin{aligned}\text{Prob(Home Run)} &= \left(\frac{1}{6} \times \frac{3}{36}\right) + \left(\frac{1}{6} \times \frac{4}{36}\right) + \left(\frac{1}{6} \times \frac{5}{36}\right) + \left(\frac{1}{6} \times \frac{6}{36}\right) \\ &\quad + \left(\frac{1}{6} \times \frac{5}{36} \times \frac{10}{20}\right) + \left(\frac{1}{6} \times \frac{4}{36} \times \frac{1}{20}\right) \\ &= 0.0958.\end{aligned}$$

5.5 Exercises

5.0. Table 5.12 shows the basic batting statistics for Rickey Henderson for the 1990 season.

Table 5.12. Batting statistics for Rickey Henderson for the 1990 season

AB	R	H	2B	3B	HR	RBI	BB	SO
489	119	159	33	3	28	61	97	60

Construct a random spinner for Henderson using his 1990 season. Compute the number of plate appearances in the 1990 season. Find the approximate probability that Henderson gets (1) a single, (2) a double, (3) a triple, (4) a home run, (5) a walk, and (6) an out. Make a spinner like the one described in Case Study 5.3 where the areas of the regions correspond to the probabilities of the batting events that you computed.

5.1. Table 5.13 shows the number of at-bats (AB) and hits (H) for Ichiro Suzuki during each month of the 2004 season.

Table 5.13. Number of hits and at-bats for Ichiro Suzuki for each month of the 2004 season

Month	H	AB
April	26	102
May	50	125
June	29	106
July	51	118
August	56	121
September	44	118
October	6	14

(a) Compute Suzuki's batting average at the end of April. Do you think this is a good estimate of Suzuki's ability to get a base hit?

(b) Compute Suzuki's batting average at the end of May. Do you think this average is a better estimate of Suzuki's "true" batting average than the value you computed in part (a)? Why?

(c) Compute Suzuki's batting average at the end of each month and at the end of the baseball season. Put your batting averages in the table below.

Month	Total H	Total AB	AVG
End of April			
End of May			
End of June			
End of July			
End of August			
End of October			

(d) Plot the batting averages against the month.

(e) What pattern do you see in your graph? What do you think is Suzuki's true batting average? Why?

5.2. Angel Pagan, the centerfielder for the 2012 San Francisco Giants, is coming to bat.

(a) Without using any data, guess at the probability that Pagan will get a triple on this at-bat.

(b) Table 5.14 shows the number of at-bats (AB) and the number of triples (3B) for Pagan for each month of the 2012 season. For each month, compute the proportion of triples for all games through that month and place your answers in the last column of the table.

Table 5.14. Number of at-bats and triples for Angel Pagan for each month of the 2012 season

Month	Triples	AB	Proportion of Triples
April	3	92	
May	1	104	
June	0	98	
July	1	81	
August	4	114	
September	6	107	
October	0	9	

(c) Based on your calculations, what is your best estimate of the probability that Pagan will get a triple? Explain why this is your best guess.

(d) Suppose a hitter gets no triples in 400 at-bats. Does this mean that the probability that he hits a triple is exactly zero? Why or why not?

5.3. If you had watched Mark Reynolds bat, you would have noticed that he tends to strike out a lot. After each game of the 2009 season, I computed the strikeout rate

$$\text{SO RATE} = \frac{\text{SO}}{\text{AB}}.$$

Figure 5.6 graphs Reynolds' strikeout rate (SO RATE) after each game of the 2009 season.

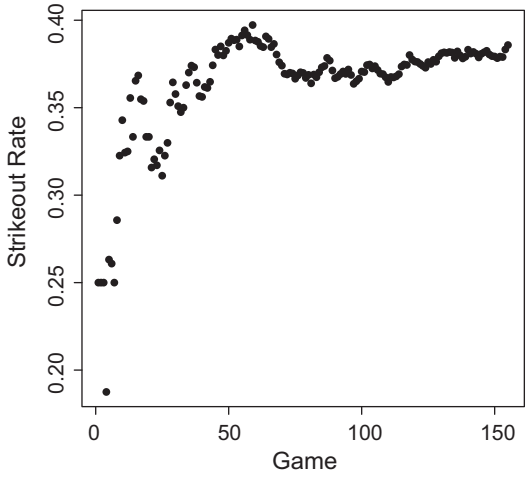


Figure 5.6. Plot of strikeout rate against game number of Mark Reynolds for the 2009 baseball season.

(a) Using the graph, estimate Reynolds' probability of striking out in an at-bat after 30 games.
(b) Estimate the chance of striking out in an at-bat after 50 games.
(c) What is your best estimate at Reynolds' strikeout probability based on this data? (Use the graph.)

5.4. Table 5.15 shows the number of innings pitched (IP) and the number of strikeouts for each game that Madison Bumgarner started during the 2014 season.

Table 5.15. Number of strikeouts and innings pitched for all games Madison Bumgarner started in the 2014 season

Game	SO	IP	Game	SO	IP	Game	SO	IP
1	3	4.00	12	10	7.00	23	2	4.00
2	10	6.33	13	5	8.00	24	10	9.00
3	7	6.00	14	5	7.00	25	5	8.00
4	6	4.33	15	9	7.00	26	9	7.00
5	6	8.00	16	7	8.00	27	12	7.00
6	5	5.00	17	3	6.00	28	13	9.00
7	9	6.00	18	6	5.00	29	7	6.00
8	8	8.00	19	3	7.00	30	0	6.00
9	5	5.00	20	5	6.33	31	9	7.00
10	6	6.00	21	7	6.00	32	6	6.00
11	10	7.00	22	6	8.00	33	5	7.33

(a) Define the strikeout rate per inning as SO.Rate $=$ SO/IP. Find Bumgarner's strikeout rate after the first game in the 2014 season.
(b) Find Bumgarner's strikeout rate after 11 games, 22 games, and at the end of the season. Put your answers in the table below.

	SO	IP	Strikeout Rate
After 11 Games			
After 22 Games			
After 33 Games			

Figure 5.7 plots Bamgarner's strikeout rate after each game pitched in the 2012 season.
(c) Describe the pattern of the graph from left to right. What does this pattern mean in terms of Bumgarner's pitching performance during the 2014 season?
(d) Suppose that Bumgarner pitched one more game in 2014 and strike out 16 batters in 8 innings. Without calculating anything, do you think Bumgarner's strikeout rate would go up? Now calculate what the strikeout rate would be. Is this more or less than you expected?

5.5. In Case Study 5.2, a basic tabletop baseball game, *Big League Baseball*, is described. This game was played on a computer for 1000 games. The number of home runs hit in each game was recorded and these data are summarized in Table 5.16.

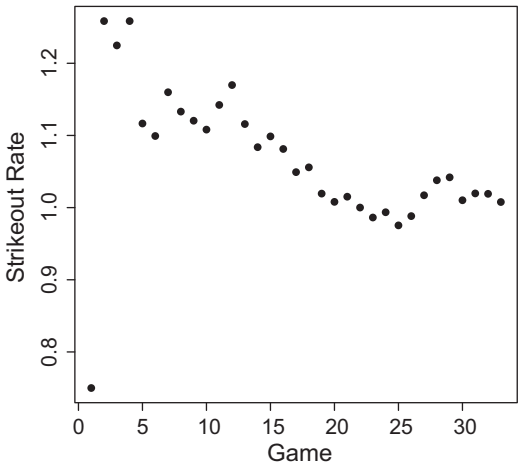


Figure 5.7. Plot of season strikeout rate against game number for Madison Bumgarner after each game he pitched in the 2014 season.

Table 5.16. Frequency table of the number of home runs hit in 1000 simulated games of *Big League Baseball*

HR	0	1	2	3	4	5	6	7
Count	165	289	263	170	75	29	5	4

(a) Find the probability that no home runs are hit in a game.
(b) Find the probability that between two and four home runs are hit.
(c) Find the probability that at least one home run is hit.
(d) If you play a game of *Big League Baseball*, what is the most likely number of home runs that will be hit?

5.6. (Exercise 5.5 continued.) For each of the 1000 games played of *Big League Baseball*, the number of pitches was recorded. Table 5.17 gives a grouped frequency table of these data.

Table 5.17. Grouped frequency table of the number of pitches thrown in 1000 simulated games of *Big League Baseball*

# Pitches	1–150	151–200	201–250	251–300	301–350	351–400	401–450
Count	20	661	278	29	8	3	1

(a) What is the probability that between 151 and 200 pitches are thrown in a game?
(b) What is the probability that at most 250 pitches are thrown?
(c) What would be a typical number of pitches thrown in a game?
(d) Would you be surprised to see a game with over 300 pitches thrown? Why?
(e) Can you think of circumstances where over 300 pitches are likely to to occur?

5.7. (Exercise 5.5 continued.) For each of the 1000 games played of *Big League Baseball*, the margin of victory (winning team score—losing team score) was recorded. A frequency table of these data is given in Table 5.18.

Table 5.18. Grouped frequency table of the number of pitches thrown in 1000 simulated games of *Big League Baseball*

Margin of victory	1	2	3	4	5	6	7	8	9	10 or more
Count	333	209	142	108	66	48	33	21	21	19

(a) What is the probability that a game of *Big League Baseball* is decided by one run?
(b) Suppose you define a blowout as a game where a team wins by six runs or more. What is the probability a game is a blowout?
(c) What is the probability a game is not a blowout?
(d) Is it unusual that a team would win by ten or more runs? Why?

5.8. (Exercise 5.5 continued.) Suppose you are interested in exploring the relationship between the number of runners to reach base and the runs scored in a half-inning of baseball. For 1000 games of *Big League Baseball*, we record for each half-inning

RUNNERS the total number of runners to reach base in the half-inning,

RUNS the number of runs scored.

Table 5.19 classifies 18,200 half-innings with respect to RUNNERS and RUNS.

Table 5.19. Two-way count table of the number of runners on base and the runs scored in half-innings from 1000 simulated games of *Big League Baseball*

	Runs								
Runners	0	1	2	3	4	5	6	>= 7	SUM
0	5481	0	0	0	0	0	0	0	5481
1	5441	539	0	0	0	0	0	0	5980
2	2408	896	244	0	0	0	0	0	3548
3	463	778	434	108	0	0	0	0	1783
4	13	215	369	176	52	0	0	0	825
5	0	3	81	159	71	19	0	0	333
6	0	0	1	35	66	35	11	0	148
7	0	0	0	0	15	32	14	2	63
8	0	0	0	0	0	6	11	6	23
9	0	0	0	0	0	0	3	4	7
10	0	0	0	0	0	0	0	6	6
11	0	0	0	0	0	0	0	3	3
SUM	13806	2431	1129	478	204	92	39	21	18200

(a) Find the probability that there are no runners on base in a particular half-inning.
(b) Find the probability that a team doesn't score any runs during their at-bat.
(c) Suppose that a team has one runner on base during their half-inning what is the probability that the runner scores?
(d) Suppose that a team has three runners on base find the probability that at least one run is scored.

5.9. (Exercise 5.8 continued.) For ten consecutive games played by the Phillies in 1999, the number of runners and the number of runs scored are recorded for each half-inning. Table 5.20 classifies all half-innings by RUNS and RUNNERS.

Table 5.20. Two-way count table of the number of runners on base and the runs scored in half-innings from ten consecutive Phillies games in 1999

	Runs							
RUNNERS	0	1	2	3	4	5	8	SUM
0	57	0	0	0	0	0	0	57
1	46	5	0	0	0	0	0	51
2	17	12	5	0	0	0	0	34
3	3	5	9	0	0	0	0	17
4	0	1	6	1	0	0	0	8
5	0	0	1	1	1	2	0	5
6	0	0	0	1	0	1	0	2
7	0	0	0	0	1	0	0	1
10	0	0	0	0	0	0	1	1
SUM	123	23	21	3	2	3	1	176

(a) Compute the same probabilities asked in parts (a)–(d) of Exercise 5.8.
(b) Compare your answers with those of Exercise 5.8. Are the results from the game *Big League Baseball* similar to those from real baseball?

5.10. In the game *Big League Baseball* described in Case Study 5.2, a red die is thrown to represent the pitch. The possible rolls of the die and the pitch results are shown in the table below.

Red die roll	Pitch result
1, 6	batter hits fair ball
2, 3	ball
4, 5	strike

(a) What is the probability that a pitch results in a strike?
(b) What is the probability that the pitch results in the batter hitting a fair ball?
(c) What is the probability that the pitch results in a ball or a strike?

5.11. (*Big League Baseball*, continued.) Suppose two pitches are thrown to a batter. Figure 5.8 displays a tree diagram that shows the possible outcomes, where a ball is denoted by B and a strike by S.
(a) Assign probabilities to each branch of the tree using the table of dice rolls in Exercise 5.10.
(b) Find the probability that two balls are thrown in a row. (To find a probability of Ball on first pitch and Ball on second pitch or Ball 1, Ball 2, you multiply the probabilities along the branches of the tree.)
(c) Find the probability that exactly one of the two pitches is a strike. (Find the probability of Ball 1, Strike 2 and Strike 1, Ball 2 and then add the probabilities of the two ways of getting one strike.)
(d) Find the probability that at least one of the pitches is a strike.

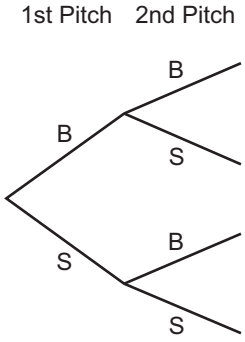


Figure 5.8. Outcomes of two pitches thrown to a batter.

5.12. (*Big League Baseball*, continued.) Use a tree diagram like the one in Figure 5.8 to answer the questions about what happens on three or more pitches.
(a) Find the probability that a batter strikes out on three consecutive pitches.
(b) Suppose the pitch count is 2-1 (that is 2 balls and 1 strike) after three pitches. List below all the possible pitch sequences (like BSB) which would result in a 2-1 count.
(c) Find the probability that the pitch count is 2-1 after three pitches. (Hint: Find the probability of each pitch sequence in (b) and then add the probabilities of all the sequences to obtain the probability you want.)
(d) Find the probability that a batter strikes out on four pitches.
(e) Find the probability that a batter strikes out on at most four pitches.

5.13. (*Big League Baseball*, continued.) If the roll of the red die is 1 or 6, then the batter hits a fair ball. Two white dice are rolled and the outcome of the play depends on the roll of the dice as shown in Table 5.21. Using the table,

Table 5.21. Play outcomes from the rolls of two dice from *Big League Baseball*

		White Die 2					
		1	2	3	4	5	6
White Die 1	1	Single	Out	Out	Out	Out	Error
	2	Out	Double	Single	Out	Single	Out
	3	Out	Single	Triple	Out	Out	Out
	4	Out	Out	Out	Out	Out	Out
	5	Out	Single	Out	Out	Out	Single
	6	Error	Out	Out	Out	Single	Home Run

(a) Find the probability that the batter gets a double.
(b) Find the probability that the batter gets a triple.
(c) Find the probability the batter gets a hit (single, double, triple, or home run).
(d) Find the probability that the batter gets on base. (Note that getting on base is different from getting a hit.)
(e) Find the probability that the batter gets on base as a result of an error.

5.14. (*Big League Baseball*, continued.) In the roll of the white dice, suppose that we consider only the outcomes that result in hits. The results of the rolls of the white dice that result in hits are shown in Table 5.22, all other results are left blank. Suppose that each hit shown in the table has the same probability.

(a) How many outcomes of the two dice result in a hit?
(b) Find the probability that a hit is a single.
(c) Find the probability that a hit is a double.
(d) Find the probability that a hit is not a single.

Table 5.22. Play outcomes from the rolls of two dice from *Big League Baseball* that result in hits

		White Die 2					
		1	2	3	4	5	6
White Die 1	1	Single					
	2		Double	Single		Single	
	3		Single	Triple			
	4						
	5		Single				Single
	6					Single	Home Run

5.15. Table 5.23 gives total at-bats, singles, doubles, etc. for all Major League games played in the 2014 season.

(a) Compute the number of plate appearances (PA) for the 2014 season.
(b) Find the probability that a hitter gets a single in a PA.
(c) Find the probability the hitter gets an extra base hit in a PA.
(d) In a PA, what is the most likely outcome: a hit, a strikeout, a walk, or an out? Explain.

Table 5.23. Total offensive statistics for all games played in the 2014 season

AB	1B	2B	3B	HR	BB	SO	HBP
165614	28423	8137	849	4186	14020	37441	1652

5.16. (2014 hitting data, continued.)

(a) Consider only the PAs that get on base (including home runs) in the 2014 offensive data. In the table below, put the number of hits of different types and the number of walks and HBPs, and find the proportion of each type.

On-base Profile

Type	Count	Proportion
Single		
Double		
Triple		
Home run		
Walk		
HBP		
TOTAL		

(b) Find the probability that an on-base event is for extra-bases (an extra-base is a hit for more than one base; it includes doubles, triples, and home runs).
(c) Find the probability that a person on-base has reached there by a walk or an HBP.
(d) Compare your proportions in the table above with the on-base profile from the tabletop game *Big League Baseball*. Are there any similarities? Any differences?

5.17. (*All-Star Baseball*) The spinner shown in Figure 5.9 was created using Ty Cobb's batting statistics for the year 1911. As in Case Study 5.3, the spinner has 36 areas of the same size. By spinning the spinner, one is simulating the result of a single plate appearance by Ty Cobb during the 1911 season.

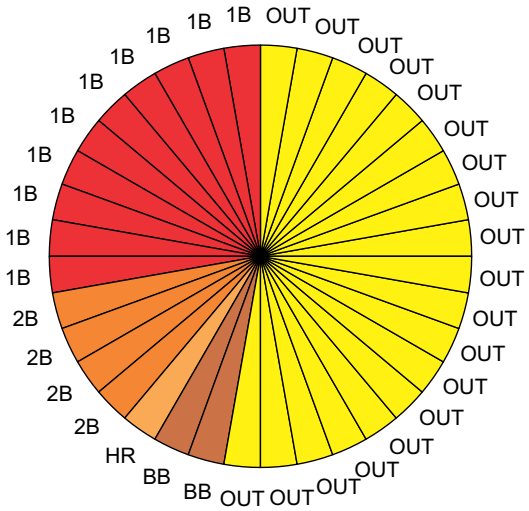


Figure 5.9. Spinner for Ty Cobb based on 1911 batting statistics.

(a) Find the probability that Cobb gets a home run on a single plate appearance.
(b) Find the probability that Cobb gets a single.
(c) Find the probability that Cobb gets on-base.
(d) Find the most likely outcome for Cobb on a single plate appearance.
(e) Find the probability that Cobb gets an extra base hit.

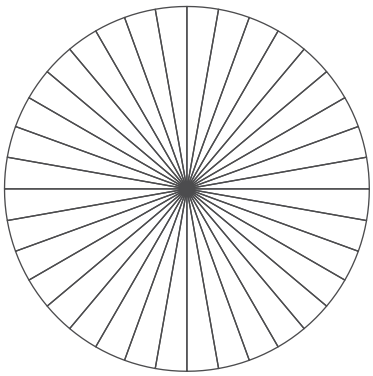
5.18. (Creating a random spinner for *All-Star Baseball* for a player from a season of batting data.) Table 5.24 gives basic hitting statistics for Mickey Mantle for the 1956 baseball season. For simplicity, we assume that plate appearances (PA) are recorded as either at-bats (AB) or walks (BB). (Hit by pitches and sacrifice hits are ignored) Also we assume at-bats that are not hits are either strikeouts (SO), or groundouts or flyouts.
(a) Compute the number of Plate Appearances (PA) by adding the AB and BB.
(b) Compute the number of singles (1B) by adding up doubles (2B), triples (3B), and home runs (HR), and subtracting this sum from the number of hits (H).
(c) Find the number of groundouts or flyouts by adding H and SO, and subtracting this sum from the number of at-bats (AB).

Table 5.24. Mickey Mantle's 1956 batting statistics

	PA	AB	H	1B	2B	3B	HR	BB	SO	Groundballs and Flyouts
Count		533	188		22	5	52	99	112	
Probability	xxx	xxx	xxx							
Regions = 36 × Prob	xxx	xxx	xxx							

(d) Find the proportion of PAs that are 1B, 2B, 3B, HR, BB, SO, and Groundouts or Flyouts put the results in the Probability row.

(e) Convert the probabilities to spinner regions by multiplying by 36 and rounding to the nearest whole number as was done in Case Study 5.3.



(f) Construct the random spinner (from the blank one shown below) using the Region values computed in part (e).

5.19. (Group activity) Suppose you are interested in playing a game of *All-Star Baseball* between the greatest players of the National League (NL) and the greatest players in the American League (AL). Using the career statistics for the 18 great players given in Table 5.25, construct a set of random spinners using the same method as Exercise 5.17.

5.20. (Probabilities of sum of two dice.) Suppose you toss two dice. Assume that the dice are distinguishable (think of one die as white and the other die as red) and there are 36 possible outcomes shown in Table 5.26. The outcomes are written in the table as roll of white, roll of red. So, for example, a 4,3 indicates that the roll of the white die was 4 and the roll of the red die was 3, a 3, 5 indicates that the white die was 3 and the red die was 5, and so on.

(a) List the outcomes in the table where the roll of the white die is equal to 2. (One of these outcomes is 2, 3.)

(b) List the outcomes where the roll of the white die is equal to the roll on the red die.

(c) List the outcomes where the roll of the white die is greater than the roll on the red die.

(d) Find the probabilities of the events

Table 5.25. Career batting statistics for 18 great players

Catcher	G	AB	H	2B	3B	HR	SO	BB
(NL) Johnny Bench	2158	7658	2048	381	24	389	1278	891
(AL) Yogi Berra	2120	7555	2150	321	49	358	414	704
FIrst Base	G	AB	H	2B	3B	HR	SO	BB
(AL) Lou Gehrig	2164	8001	2721	534	163	493	790	1508
(NL) Mark McGwire	1745	5830	1553	247	6	547	1461	1247
Second Base	G	AB	H	2B	3B	HR	SO	BB
(NL) Jackie Robinson	1382	4877	1518	273	54	137	291	740
(AL) Eddie Collins	2826	9949	3315	438	187	47	286	1499
Shortstop	G	AB	H	2B	3B	HR	SO	BB
(AL) Cal Ripken, Jr.	2848	10982	3046	580	44	415	1228	1095
(NL) Honus Wagner	2792	10430	3415	640	252	101	327	963
Third Base	G	AB	H	2B	3B	HR	SO	BB
(NL) Mike Schmidt	2404	8352	2234	408	59	548	1883	1507
(AL) Brooks Robinson	2896	10654	2848	482	68	268	990	886
Outfielders	G	AB	H	2B	3B	HR	SO	BB
(AL) Babe Ruth	2503	8399	2873	506	136	714	1330	2062
(AL) Ted Williams	2292	7706	2654	525	71	521	709	2019
(NL) Willie Mays	2992	10881	3283	523	140	660	1526	1464
(NL) Hank Aaron	3298	12364	3771	624	98	755	1383	1402
(AL) Joe DiMaggio	1736	6821	2214	389	131	361	369	790
(AL) Mickey Mantle	2401	8102	2415	344	72	536	1710	1733
(NL) Pete Rose	3562	14053	4256	746	135	160	1143	1566
(NL) Stan Musial	3026	10972	3630	725	177	475	696	1599

Table 5.26. Outcomes of rolling a white die and a red die

		Red Die					
		1	2	3	4	5	6
White Die	1	1, 1	1, 2	1, 3	1, 4	1, 5	1, 6
	2	2, 1	2, 2	2, 3	2, 4	2, 5	2, 6
	3	3, 1	3, 2	3, 3	3, 4	3, 5	3, 6
	4	4, 1	4, 2	4, 3	4, 4	4, 5	4, 6
	5	5, 1	5, 2	5, 3	5, 4	5, 5	5, 6
	6	6, 1	6, 2	6, 3	6, 4	6, 5	6, 6

i. roll of white die is equal to 2,
ii. roll of white die is equal to roll of red die,
iii. roll of white die is greater than roll of red die.

5.21. (*Strat-O-Matic* baseball) Suppose the 1998 Tony Gwynn is facing the 1998 Randy Johnson. (A great hitter against a great pitcher.) The Gwynn and Johnson *Strat-O-Matic* cards

are displayed in Table 5.27 and Table 5.28. Three dice, one red and two white, are tossed to determine the result. The red die determines which card to look at: if the red die is 1, 2, 3, Gwynn's card is used and if the die is 4, 5, 6, Johnson's card is used (these numbers are shown at the top of the cards). The sum of the two white dice determines which number is checked under the red dice number. Suppose that the roll of the red die is 3 and the sum of the white dice is 9. We check the number 9 under Gwynn's 3 column and read that Gwynn has walked on this particular at-bat.

Table 5.27. Tony Gwynn's *Strat-O-Matic* card from the 1998 baseball season

1	2	3
2-lineout	2-lineout	2-lineout
3-groundball	3-groundball	3-groundball
4-HOMERUN	4-groundball	4-flyball
5.HOMERUN	5.flyball	5.SINGLE
1-6	6-popout	6-SINGLE
DOUBLE	7-popout	7-groundball
7-20	8-lineout	8-groundball
6-DOUBLE	9-flyball	9-WALK
7-DOUBLE	10-groundball	10-SINGLE
1-7	11-groundball	1-4
SINGLE	12-lineout	lineout
8-20		5.20
8-SINGLE		11-groundball
9-groundball		12-foulout
10-SINGLE		
11-SINGLE		
12-popout		

Table 5.28. Randy Johnson's *Strat-O-Matic* card from the 1998 baseball season

4	5	6
2-flyball	2-strikeout	2-groundball
3-groundball	3-WALK	3-flyball
4-catcher	4-strikeout	4-groundball
5.strikeout	5.strikeout	5.strikeout
6-strikeout	6-DOUBLE	6-strikeout
7-WALK	1-13	7-SINGLE
8-strikeout	SINGLE	8-strikeout
9-strikeout	14-20	9-strikeout
10-groundball	7-groundball	10-groundball
11-strikeout	8-strikeout	11-groundball
12-flyball	9-HOMERUN	12-groundball
	1-15	
	DOUBLE	
	16-20	
	10-flyball	
	11-WALK	
	12-SINGLE	
	1-12	
	lineout	
	13-20	

(a) What is the chance that the Gwynn card is used (when the red die is rolled 1, 2, or 3)?
(b) Suppose that the roll of the red die is 2 (so the middle column of Gwynn's card is used). What rolls of the sum of the white dice will result in a home run?
(c) If the roll of the red die is 6, which rolls of the sum of the white dice will result in a strikeout for Gwynn?
(d) If the roll of the red die is 1, which rolls of the sum of the white die will result in a hit?

5.22. (*Strat-O-Matic* continued). The 1998 Tony Gwynn is facing the 1998 Randy Johnson. If the roll of the red die is 6, Table 5.29 gives the possible rolls of the sum of white dice, the probabilities, and the outcomes (copied from the Johnson card above).

(a) In this case (the roll of the red die is 6)
   i. What is the probability that Gwynn will single?
   ii. What is the probability that Gwynn will strikeout?
   iii. What is the probability that Gwynn will hit a groundball?

(b) If the red die is rolled 2, what is the probability that Gwynn will popout? (Look at the Gwynn card above.)

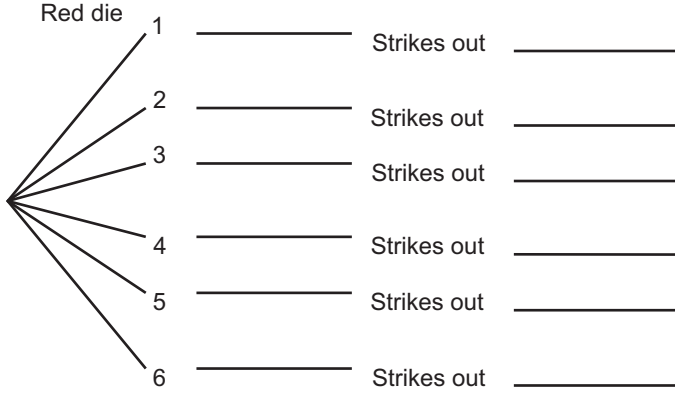
Table 5.29. Rolls of the sum of the white dice, probabilities, and outcomes from the "6" column of Johnson's card

Sum of white dice	Probability	Outcome
2	1/36	Groundball
3	2/36	Flyball
4	3/36	Groundball
5	4/36	Strikeout
6	5/36	Strikeout
7	6/36	Single
8	5/36	Strikeout
9	4/36	Strikeout
10	3/36	Groundball
11	2/36	Groundball
12	1/36	Groundball

5.23. (*Strat-O-Matic* continued). The 1998 Tony Gwynn is facing the 1998 Randy Johnson. Suppose that we are interested in the probability that Gwynn strikes out against Johnson. The event that Gwynn strikes out can be divided into six different events.

- The roll of the red die is 1 and Gwynn strikes out.
- The roll of the red die is 2 and Gwynn strikes out.
- The roll of the red die is 3 and Gwynn strikes out.
- The roll of the red die is 4 and Gwynn strikes out.
- The roll of the red die is 5 and Gwynn strikes out.
- The roll of the red die is 6 and Gwynn strikes out.

These events are represented as the six branches of the tree diagram below. To find the probability that Gwynn strikes out, we first find the probabilities of all of the subbranches



and then add the products of the probabilities along the subbranches to get the desired result. We outline these calculations below.

(a) First, find the probability that the roll of the red die is 1, 2, . . . , 6, and place these probabilities at the six branches under the label Red die in the diagram.

(b) If the roll of the red die is 1 (so that you look at Gwynn's card in the 1 column), find the probability that Gwynn strikes out. Place this probability above the first horizontal line in the second set of branches.

(c) If the roll of the red die is 2, find the probability of a strikeout. Similarly, find the probability of a strikeout if the red die falls 3, 4, 5, and 6. Place these five probabilities of a strikeout above the corresponding horizontal lines to the right of 2, 3, 4, 5, 6 in the diagram.

(d) The probability of a strikeout is found in two steps:
- Multiply (for each possible red die roll) the probability of the red die roll and the probability of a strikeout for that red die roll, placing them in the blank lines to the right of the word strikes out.
- Add the six products you found above.

You have now found the probability of Gwynn striking out.

5.24. (*Strat-O-Matic* continued.) Use the tree diagram method described in Exercise 5.23 to find the probability that Tony Gwynn gets a walk against Randy Johnson.

5.25. (Win Probabilities) Probabilities can be used to describe the certainty that a team will win at different times during a game. One assumes that the probability of the home team winning is 0.5 at the beginning of a game, and this probability will increase or decrease during the game based on the runs scored by either team. The table below gives the inning, score, and probability the home team wins for particular instances during the Royals at Astros playoff game on October 12, 2015. (The information was collected from the box score of this game posted at `baseball-reference.com`.)

	Score		
Inning	Royals	Astros	P(Astros win)
Start of game	0	0	0.50
End of bottom of 5th	2	3	0.67
End of bottom of 7th	2	6	0.97
End of bottom of 8th	7	6	0.30
End of top of 9th	9	6	0.04
End of game	9	6	0

(a) What team was likely to win at the end of the 7th inning?

(b) What team was likely to win at the end of the 8th inning? What happened during the 8th inning (runs scored) which caused a dramatic change in win probabilities?

(c) For a different baseball game of interest, create a similar table (using information from `baseball-reference.com`) displaying the probability the home team wins during different times during the game.

Further Reading

Basic concepts of probability are presented in Devore and Peck (2011), Moore, McCabe and Craig (2012), and Scheaffer and Young (2009). Albert and Bennett (2003), Chapter 1, discuss the probability models behind some popular tabletop baseball games, including *All-Star Baseball*, *Strat-O-Matic Baseball*, and *APBA Baseball*.

10.1090/text/034/06

6

Probability Distributions and Baseball

What's On-Deck?

In this chapter, we show how some basic discrete probability distributions can be used to model events in baseball. One nice feature of baseball is its discrete structure. A player comes to bat during an inning, and there is a result of this plate appearance which might be a hit, a walk, an out, an error, or a hit-by-pitch. The number of hits by a player in a given number of at-bats can be represented by a binomial distribution where the probability of a hit is the player's true batting average. We show in Case Study 6.1 that binomial probabilities can be very useful in predicting the number of games with 0 hits, 1 hit, and so on for a particular hitter. The remainder of the chapter focuses on modeling of run production for a team. To score runs, the batters need to get on base, and Case Study 6.2 considers a probability distribution for the number of batters that come to bat in a half-inning. If the result of each batting appearance is "out" or "get on base" and we are interested in the number of hitters B until three outs, then we can model B with a negative binomial distribution. If we can estimate a team's on-base probability, then we can find a probability distribution for the number of batters that come to bat during an inning. The second step of the run scoring process is advancing runners that get on-base. Suppose that we know the proportion of on-base events of a team that are walks, singles, doubles, triples, and home runs. (We call this set of proportions the team's on-base profile.) If we know that a team has a particular number of on-base events, then Case Study 6.3 shows how we can use a team's on-base profile to compute the probability that the team scores a given number of runs. By looking at the actual runs scored by the 2014 Boston Red Sox, we will see that this model does a reasonable job in explaining a team's run production.

6.1 The Binomial Distribution and Hits per Game

TOPICS COVERED: Binomial probabilities, independence, expected counts, simulation.

The binomial distribution is one of the most useful probability distributions in statistics. It's a most useful distribution since it is applicable in a wide variety of situations, including baseball. Suppose we have an experiment consisting of a sequence of n identical trials, where each trial can result in one of two possibilities, called a success and a failure. The chance of a success, p, is assumed constant for each trial, and the results of different trials are assumed independent. Then X, the number of successes in n trials, has a binomial probability function

given by

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

where $\binom{n}{x}$ is the binomial coefficient

$$\binom{n}{x} = \frac{n!}{x!(n-x)!},$$

x is an integer and the values of x can range from 0 to n. (The notation $n!$, called n factorial, is the product of integers $n \times (n-1) \times (n-2) \times \cdots \times 1$.)

Suppose a baseball hitter comes to bat n times during a game. (For this example, we will only consider official at-bats, where the hitter gets a hit or produces an out. A walk or hit-by-pitch or a sacrifice fly are not considered official at-bats.) For each at-bat, we will call a base hit a success, and an out a failure. Suppose that p, the probability of a hit remains constant for all at-bats, and the results of different at-bats are independent. Then X, the number of hits in n at-bats during a game, will have a binomial distribution with parameters n and p.

Before we try to fit a binomial distribution to hitting data, we should ask ourselves if the above assumptions make sense. One important assumption is that the probability of a hit for a batter doesn't change across a game. In our modeling, it is convenient to go one step further and assume that the probability of a hit for a batter doesn't change across a season. The probability p represents the hitting ability of the particular player, and so we are saying that a batter's ability stays relatively constant across games during the season. The second important assumption is that the chance of a player getting a hit doesn't depend on his performance in previous at-bats. This means that the player can't be streaky in his batting ability the result of a particular at-bat (hit or out) doesn't depend on how he did in his recent at-bats.

One might argue that the binomial assumptions are too simplistic you might think that a batter's ability does change over the season, and you may have heard that particular players are streaky who tend to go through stretches of good hitting and bad hitting. Also, a batter may have different abilities to hit against different pitchers and in different ballparks. But, although the assumptions seem a bit restrictive, the binomial model will be shown to give reasonable predictions of batting performance.

Let's illustrate the use of the binomial distribution to model the game-to-game hitting performance of Melky Cabrera. In the 2014 season, Cabrera got 171 hits in 568 official at-bats for a batting average of $171/568 = .301$. Table 6.1 categorizes the games where he had at least one official at-bat by the number of at-bats (AB) and the number of hits (H).

Let's focus on the games where Cabrera had exactly four at-bats. We notice a lot of variation in the number of game hits. Sixteen games Cabrera was hitless, 32 games Cabrera was 1 for 4, 22 games he was 2 for 4, three games he was 3 for 4, and he never went 4 for 4.

Can this variation in the number of hits be explained by the binomial distribution?

To find probabilities using the binomial formula, we need to specify n and p. Since we're only considering games where Cabrera has four (official) opportunities to bat, $n = 4$. A reasonable guess at p is .301, the batting average of Cabrera for the entire 2014 season. Here X denotes the number of hits for Cabrera for this four-AB game.

Table 6.1. Categorization of the 2014 game batting results of Melky Cabrera by the number of at-bats and the number of hits

		Hits					
		0	1	2	3	4	Total
	1	1	1	0	0	0	2
	2	3	1	0	0	0	4
Game	3	5	8	2	1	0	16
AB	4	16	32	22	3	0	73
	5	7	13	12	6	2	40
	6	0	0	3	0	0	3

Using these values of n and p, we compute the probabilities for the five possible values of X in Table 6.2. Also, since Cabrera has 73 of these games, we can also compute the expected number of games where he has different numbers of hits by multiplying these probabilities by 73.

Table 6.2. Probability and expected number of games (in 73 games) for Melky Cabrera to have different hit numbers using the binomial formula with $n = 4$ and $p = .3$

X	Binomial probability	Expected number	Observed
0	$\binom{4}{0}.301^0(1-.301)^{4-0} = .2387$	$73 \times .2387 = 17.4$	16
1	$\binom{4}{1}.301^1(1-.301)^{4-1} = .4112$	$73 \times .4112 = 30.0$	32
2	$\binom{4}{2}.301^2(1-.301)^{4-2} = .2656$	$73 \times .3656 = 19.4$	22
3	$\binom{4}{3}.301^3(1-.301)^{4-3} = .0762$	$73 \times .0762 = 5.6$	3
4	$\binom{4}{4}.301^4(1-.301)^{4-4} = .0082$	$73 \times .0082 = .6$	0

To see if the binomial distribution is a good fit to Cabrera's hitting data, we compare the expected counts using the binomial formula with the actual observed counts in the table. Comparing the two columns, we see that:

- Cabrera had more 1-hit and 2-hit games than we would expect.
- Cabrera had fewer no-hit and 3-hit games than we expect.

These observations don't necessarily mean that the binomial formula is a poor fit to Cabrera data. For example, we expect 50 heads in 100 tosses of a fair coin, but the probability of getting exactly 50 heads is very small. The question is if the differences between Cabrera's observed numbers and the expected numbers can be explained by chance, or if the differences really reflect a misfit of the binomial model.

To answer this question, we perform a simple simulation. We assume that Cabrera's hitting probability is $p = .301$ and we simulate many sequences of 73 four-AB games using the binomial model with $n = 4$ and $p = .301$. Each time, we simulate the season data, we keep track of the number of no-hit games, one-hit games, and so on. In Figure 6.1, we graph the numbers of game hits for the 100 simulated seasons using dots, and graph Cabrera's numbers by a solid line. Comparing Cabrera with the dotplots, we see that his numbers of 0-hit, 1-hit,

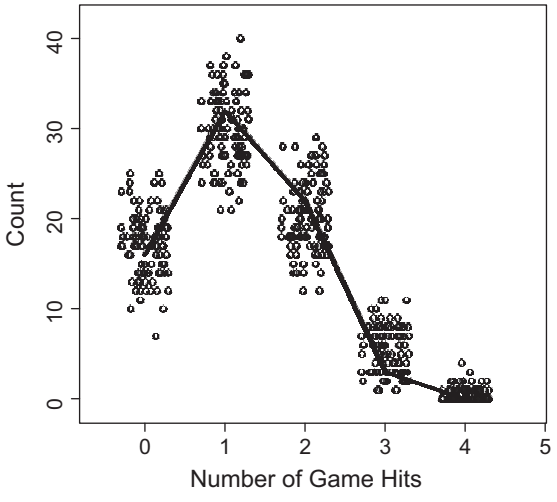


Figure 6.1. Graph of count of games (out of 73) with different number of game hits from the simulated binomial distribution with $n = 4$ and $p = .301$. The observed number of 0-hit, 1-hit, 2-hit, 3-hit, and 4-hit games for Melky Cabrera is displayed by a solid line.

2-hit, 3-hit, and 4-hit games are in the middle of the simulated distributions. For this particular player, we conclude that the binomial formula is a good fit for that game-to-game variation in hitting data.

What if we looked at a large number of batters and tried to fit a binomial distribution to the daily hit numbers for each player. What would we find? In my experience, the binomial probability distribution tends to be a pretty good fit to hitting data for most players. Although it might be insulting to say this to a baseball hitter, his variation of hitting performance across games resembles the same type of chance variation you get from tossing a coin many times where the probability of heads matches the batter's true average.

6.2 Modeling Runs Scored: Getting on Base

Topics Covered: Negative binomial distribution, expected counts, Pearson residuals.

In the next two case studies, we discuss the use of probability distributions to model the number of runs that a team scores.

How does a team score runs? It is a two-step process. First batters need to get on-base by means of hits, walks, errors, or hit-by-pitches. Second, once players get on-base, runs are generally scored by hits by other batters that move the base runners to home plate. Hitters are valuable if they are successful in getting on-base, or if they are effective in driving home runners. One measure of the ability of a batter to get on base is the on-base percentage (OBP). The ability to drive runners home is typically measured by the runs batted in or RBI.

Let R denote the number of runs scored by a particular team during a half-inning of a baseball game. We are interested in modeling the variation in the variable R by means of a probability distribution. We will do this in two steps:

1. We first construct a model for B, the number of players that come to bat during a half-inning.

2. We then find a suitable model for the runs scored (R) given that we know how many batters came to bat in the inning.

Here we focus on the number of hitters (B) that come to bat. The random variation in B can be modeled by a well-known probability distribution for experiments consisting of a sequence of yes or no outcomes (so-called Bernoulli trials).

During a half-inning, a number of players come to bat. Each player will either

OB: get on-base without creating an out

or

OUT: create an out.

Assume that the probability that any player creates an out is equal to p and the outcomes of different players in the inning are independent. Then the results of the batters in the inning can be regarded as independent Bernoulli trials with probability of creating an out p. Players will continue to come to bat until the number of OUTS, or the number of successes, is equal to 3. (Here we are referring to a "success" as getting an OUT.) If the batter results are independent Bernoulli trials, then the number of trials until the 3rd success, B, is distributed according to a negative binomial distribution where

$$\Pr(B = b) = \binom{b-1}{2} p^3 (1-p)^{b-3}, \quad b = 3, 4, \ldots$$

Let's apply this formula to model the number of batters per inning for the 2014 Houston Astros. To compute this negative binomial formula, we need only estimate p, the chance that a batter creates an out. For the 2014 Astros, the team on-base percentage is

$$\text{OBP} = .309$$

and so the probability that a player creates an out can be estimated by

$$p = 1 - .309 = .691.$$

In Table 6.3, we show a table of these negative binomial probabilities. In the 2014 baseball season, the Astros batted in 1450 innings. We can obtain the expected number of innings with three batters, with four batters, and so on by multiplying the negative binomial probabilities by 1450.

Table 6.3. Probability of having different numbers of batters in a half-inning from a negative binomial distribution where b is the number of batters until the third out with $p = .691$

b	3	4	5	6	7	8	9
Probability	.3299	.3059	.1890	.0973	.0451	.0195	.0080
Expected	478.4	443.3	274.1	141.1	65.4	28.3	11.6

Are these probabilities and expected counts a reasonable match to the actual on-base production of the Astros during the 2014 season? To check, Table 6.4 also includes the observed number of innings where the Astros had different numbers of batters. A standard way of

gauging the difference between the observed and expected counts is by means of the Pearson residual

$$r = \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}.$$

When the observed count is far from its corresponding expected count, the residual will be large. A value of the residual that is 4 or larger indicates a significant discrepancy between the observed count and the fitted count assuming the negative binomial model. Looking at Table 6.4, we see two large residuals corresponding to $B = 3$. and $B = 6$. We would expect (from the model) that the Astros would have three batters (that is, a one-two-three inning) for 478 innings when in actuality they had three batters for 554 innings, 76 more. Also, the model predicts that the Astros would have 141 six-batters innings when they actually had only 107 of these type of innings. Generally the negative binomial distribution does not appear to be a good match to the Astros on-base data.

Table 6.4. Probability of having different numbers of batters, the observed numbers from the 2014 Astros season, and the Pearson residuals comparing the observed and expected counts

b	3	4	5	6	7	8	9
Probability	.3299	.3059	.1890	.0973	.0451	.0195	.0080
Expected	478.4	443.3	274.1	141.1	65.4	28.3	11.6
Observed	554	430	257	107	60	27	13
Residual	11.96	0.41	1.06	8.23	0.45	0.06	0.17

Can we offer any explanation for the lack-of-fit of our model? We are assuming that the probability that a player gets on-base is constant for all players. We know that players are not equally proficient in getting on-base and the batters at the bottom of the order are weak hitters with relatively small values of p. So the Astros large number of three-batter innings may be a reflection of the innings where the bottom of the batting order is hitting.

6.3 Modeling Runs Scored: Advancing the Runners to Home

TOPICS COVERED: Multinomial probability distribution, independence, expected counts, Pearson residuals.

In the previous study, we focused on the number of hitters that come to bat during a particular half-inning. In this case study, we consider the problem of modeling the number of runs scored in a half-inning. The runs scored depends significantly on the number of players B that bat during an inning. So we focus on modeling, with a simple probability distribution, the number of runs scored conditional on the fact that exactly B batters come to bat.

If six hitters come to bat in an inning, we know that three hitters created outs (there are three outs in an inning), and so the remaining $6 - 3 = 3$ hitters get on-base. How many runs can the team score when 3 runners get on-base? Well, the team could leave the bases loaded and score no runs. Or maybe all of the runners will score, resulting in three runs. Actually, the number of runs scored can be 0, 1, 2, or 3.

The exact number of runs scored depends heavily on the type of hit or non-hit of the batters that reach base. There are five possibilities for this on-base event:

On-base event	Abbreviation
walk or hit-by-pitch.	W
single	1
double	2
triple	3
home run	H

(We place a walk and a hit-by-pitch in the same classification since both events have the same effect on runners on base.) Each team will tend to hit different proportions for these five events. Some teams will rely on power and hit a high fraction of doubles and home runs; other teams may rely on their ability to draw a walk and have a high proportion of walk/hit-by-pitch events. We let f_0, f_1, f_2, f_3, f_4 denote the probabilities that the on-base event of the team is a walk/hit-by-pitch, single, double, triple, and home run, respectively. We call the proportions $(f_0, f_1, f_2, f_3, f_4)$ the on-base profile of the team.

Suppose that you are given a particular sequence of on-base events. For example, suppose you are told that the first batter on-base gets a double, the next one singles, the next one draws a walk, and the last on-base person singles. (We abbreviate this sequence as "21W1".) Then by making some assumptions about runner advancement, we can figure out how many runs will score.

We will use the following runner advancement assumptions based on what is typical in a baseball game.

1. A single will move a runner from first base to third base, and score a runner from second or third base.
2. A double or a triple will score all runners from first, second, and third bases.
3. A home run will score all runners.
4. An out does not advance a runner and does not eliminate a runner. (That is, all outs are treated as strikeouts.)

Using these assumptions, we can compute the runs scored for any sequence of on-base events. To illustrate, Table 6.5 shows the base situation after every event in the sequence 21W1. For this particular sequence, a total of two runs were scored in the inning.

Table 6.5. Illustration of runs scored in the play sequence "21W1" using our advancement assumptions

Event	Description	Bases	Runs Scored
"2"	Double		0
"1"	Single		1
"W"	Walk		0
"1"	Single		1

Suppose that each on-base event can be one of the five possibilities with probabilities f_0, f_1, f_2, f_3, f_4, and the on-base events for different hitters are independent. Then we can compute the probability of any sequence of events by simply multiplying the corresponding probabilities. For example, the probability of the sequence "21W1" will be

$$\begin{aligned}\text{Prob}(\text{“21W1”}) &= \text{Prob}(2) \times \text{Prob}(1) \times \text{Prob}(W) \times \text{Prob}(1)\\ &= f_2 \times f_1 \times f_0 \times f_1.\end{aligned}$$

Let's return to our original question suppose a team has three on-base events. Note that at most three runs can score from three on-base sequences, since a runner can only score if he gets on base. What is the probability that the team scores 0, 1, 2 or 3 runs? We compute this by

- finding all three on-base event sequences that result in the particular number of runs scored,
- finding the probability of each on-base sequence by multiplying the probabilities as we did above,
- adding the probabilities of all of the on-base sequences.

We illustrate this computation for one case. What is the probability of scoring exactly one run if you have three on-base events?

It turns out there are 24 sequences of three on-base events that result in one run scored.

ww1	12w	23w	hww
w11	13w	3w1	hw1
w2w	2w1	31w	h1w
w3w	21w	311	h11
1w1	211	32w	h2w
111	22w	33w	h3w

So, the probability of scoring one run (given three people on-base) is

$$\begin{aligned}\text{Prob} = {} & f_0 f_0 f_1 + f_1 f_2 f_0 + f_2 f_3 f_0 + f_4 f_0 f_0\\ & + f_0 f_1 f_1 + f_1 f_3 f_0 + f_3 f_0 f_1 + f_4 f_0 f_1\\ & + f_0 f_2 f_0 + f_2 f_0 f_1 + f_3 f_1 f_0 + f_4 f_1 f_0\\ & + f_0 f_3 f_0 + f_2 f_1 f_0 + f_3 f_1 f_1 + f_4 f_1 f_1\\ & + f_1 f_0 f_1 + f_2 f_1 f_1 + f_3 f_2 f_0 + f_4 f_2 f_0\\ & + f_1 f_1 f_1 + f_2 f_2 f_0 + f_3 f_3 f_0 + f_4 f_3 f_0.\end{aligned}$$

Using this method, we can compute the probability of scoring any number of runs given that we know how many batters get on base.

Does this probability model explain the variation in the runs scored in an inning in baseball? To check, we look at the 2014 Boston Red Sox. To compute the probabilities of runs scored (given a particular number of runners on base), we need only to estimate the on-base profile for the Red Sox shown in Table 6.6.

Table 6.6. On-base profile of the 2014 Boston Red Sox

Event	Singles	Doubles	Triples	Home runs	Walks
Count	930	282	20	123	603
Proportion	.4750	.1440	.0102	.0628	.3080

We see that, of the on-base events of the Red Sox, about half (47.50%) were singles, 14% were doubles, 1% were triples, 6% were home runs, and 31% were walks. We can use this on-base profile to compute the chance that the Red Sox will score any number of runs given a particular number of base runners. Actually, it is a bit tedious to compute these probabilities using formulas—it is much simpler to simulate this process and use the simulated output to compute the probabilities.

We consider four scenarios one, two, three, and four runners on-base. Table 6.7 shows, in each case, the computation of

Table 6.7. Probability of scoring different numbers of runs from the model, the observed count of run numbers from the 2014 Red Sox season, and the Pearson residuals. These quantities are given for each of four runner on-base scenarios

One runner on-base		
	Runs Scored	
	0	1
Probability	0.938	0.062
Expected count	389.3	25.7
Observed	374	41
Residual	0.60	9.06

Two runners on-base			
	Runs Scored		
	0	1	2
Probability	0.660	0.277	0.063
Expected count	169.0	70.9	16.1
Observed	149	90	17
Residual	2.36	5.14	0.05

Three runners on-base				
	Runs Scored			
	0	1	2	3
Probability	0.202	0.459	0.276	0.063
Expected count	22.5	50.6	30.7	7.0
Observed	21	61	22	7
Residual	0.01	2.13	2.49	0.00

Four runners on-base				
	Runs Scored			
	1	2	3	4
Probability	0.203	0.456	0.277	0.063
Expected count	12.6	28.3	17.2	3.9
Observed	7	29	21	5
Residual	2.48	0.02	0.85	0.31

- the probability of scoring different numbers of runs from the model,
- the expected counts found by multiplying the probabilities by the number of times the Red Sox had that number of batters on base,

- the observed counts of run numbers from the 2014 Red Sox season,
- the Pearson residual

$$r = \frac{(o-e)^2}{e}$$

that compares the expected (e) and observed (o) numbers.

How do the expected counts from our model compare with the observed counts? Generally the model works very well, especially when there are three or four runners on-base in the inning. Looking at the largest Pearson residuals, it seems that the model is not working well in two situations: scoring one run when there is one runner ($r = 9.06$), and scoring one run when there are two runners in the inning ($r = 5.14$). The Red Sox actually scored one run 41 times where they had only one runner in the inning and we would predict 25.7 from the model. Also, when the Red Sox had two runners, they scored a run 90 times compared to 70.9 predicted from the model.

Why doesn't the model fit well in these two situations? Remember that the run probabilities are based on our model, which assumed that runner advancement was a function only of a team's on-base profile of walks and hits. There is no allowance for base stealing, sacrifice hits, or errors that are also helpful in advancing runners. Strategies like base stealing and sacrificing are often used in baseball when the game is close and the team is trying to score a single run. These other types of run advancement may explain why a team is more likely to score a run than what we predict.

But the model produces estimates that are generally close to the actual numbers of runs scored. This is a bit surprising since we are assuming that

- each player on the team has the same on-base profile (this clearly is not true),
- on-base outcomes by different players in an inning are independent,
- as we said above, the only way to advance a runner is by means of a hit or a walk.

We could change our model to make it more realistic. (Certainly people who construct baseball simulation games such as the ones discussed in Chapter 5 would want to design a model to make it as realistic as possible.) But the above model is attractive in that it is fairly simple and helps us understand the importance of on-base events towards the goal of scoring runs.

6.4 Exercises

6.0. Here we use the Rickey Henderson spinner that we constructed for the leadoff exercise of Chapter 5.

(a) Suppose that Henderson plays 50 games and each game he has five plate appearances. Spin the Henderson spinner a total of 250 times, keeping track of the number of times he gets on-base for each game. Record your counts in the following table.

Number of times on-base	0	1	2	3	4	5
Count						

(b) Find the probabilities that Henderson gets on-base 0, 1, 2, 3, 4, and 5 times, if the number of times on-base has a binomial distribution with 5 trials and probability of success p (Estimate p by the fraction of times he gets on-base for the 1990 season.)

(c) Find the expected number of games where Henderson gets on base $0, 1, \ldots, 5$ times. Compare your expected counts with the simulated counts using the spinner.

6.1. Barry Bonds had a remarkable 2001 season when he hit 73 home runs. Suppose we focus on the games in which Bonds had at least three plate appearances. Table 6.8 gives a frequency distribution of the number of home runs that Bonds hit in these games.

Table 6.8. Frequency distribution of the number of home runs hit by Barry Bonds in 2001 in all games with at least three plate appearances

Number of home runs hit	0	1	2	3
Count	87	50	8	2

Overall in these games he hit 72 home runs an average of .4898 home runs per game.

(a) If x denotes the number of home runs hit per game, estimate the probability that x is equal to 0, 1, 2, 3 using a Poisson density with mean equal to .4898.
A *Poisson density* with mean L has probabilities given by

$$\Pr(x) = \frac{e^{-L}L^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Put your probabilities in the table below.

Number of home runs hit x	0	1	2	3
Count	87	50	8	2
Probability				
Expected count				

(b) Find the expected number of games that Bonds would hit 0, 1, 2, 3 home runs put these expected counts in the same table.

(c) Use Pearson residuals to compare the observed and expected counts. Is the Poisson distribution a good fit to these data?

6.2. (Exercise 6.1 continued.) As an alternative to the Poisson distribution, perhaps the distribution of x can be modeled using a binomial distribution with $n = 4$ and $p = .1099$. (Here $n = 4$ represents a typical number of opportunities for Bonds to bat each game and p is the probability that he will hit a home run in a single plate appearance in that particular season. Find the binomial probabilities, expected counts, and Pearson residuals using this binomial fit. Comment on the suitability of the fit and contrast the fit with the Poisson fit in Exercise 6.1.

6.3. A baseball team typically experiences a number of winning and losing streaks during a season. Should we be surprised by these streaks? The 162 games played by the 2001 World Champion Arizona Diamondbacks were divided into 54 groups of three consecutive

games. In each group of three games, the number of wins w was recorded. A frequency distribution for w is shown in Table 6.9.

Table 6.9. Frequency distribution of games won in groups of three consecutive games by the Arizona Diamondbacks in 2001

w	0	1	2	3
Count	5	16	23	10

(a) Find the probabilities that Arizona wins 0, 1, 2, and 3 games in a three game period if w is given a binomial distribution with $n = 3$ and $p = .5679$. (In the 2001 season, the Diamondbacks had a 92-70 record for a winning fraction of $92/162 = .5679$.)
(b) Find the expected number of groups that Arizona wins 0, 1, 2, and 3 in a 162-game season assuming the binomial model.
(c) Compare the observed and expected counts by the computation of Pearson residuals. Does a binomial fit seem reasonable for these data?

6.4. Table 6.10 gives the distribution of game hits for all 4-AB games in the 2001 baseball season for Shawn Green and Brian Jordan

Table 6.10. Distribution of game hits in all 4-AB games in the 2001 season by Shawn Green (2001 AVG = .297) and Brian Jordan (2001 AVG = .295)

Number of hits	0	1	2	3	4
Count for Shawn Green	21	21	30	2	1
Count for Brian Jordan	16	35	12	6	1

As in Case Study 6.1, investigate if a binomial distribution is a reasonable model of the number of game hits for each player. In each case, use the season batting average (AVG) for the binomial probability of success.

6.5. Suppose that we are interested in the random variable X that is the number of the half-inning in which the first run is scored in a baseball game. In the game

		R	H	E	LOB
Texas	040 210 000	7	8	0	7
Anaheim	012 001 001	5	15	0	10

the first run scored was in the top of the 2nd inning, so $X = 3$. In the game

		R	H	E	LOB
Milwaukee	001100010	3	9	1	9
Houston	20600030x	11	11	0	5

the first run scored was in the bottom of the 1st, so $X = 2$. We recorded X for a sample of 150 games in the 2001 season and the observed frequency distribution of X is shown in Table 6.11.

Table 6.11. Frequency distribution of the number of half-innings until the first run scored for 150 games in the 2001 season

X	1	2	3	4	5	6
Count	39	39	22	14	9	4
X	7	8	9	10	11	19
Count	9	5	4	2	2	1

One can model X by a geometric distribution where p is the probability that a team scores in a single inning. Here X represents the number of trials (half-innings) until the first success (half-inning with a run scored). In a sample of 2400 half-innings from the 2001 baseball season, the offensive team scored in 703 of the half-innings, so the probability p can be estimated by $703/2400 = .297$.

(a) Find the probabilities that X takes on the values $1, 2, 3$, using the geometric probability distribution with formula

$$\Pr(X = x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

where $p = .297$.

(b) Find the expected number of games (out of 150) in which $X = 1, 2, 3, \ldots, 19$.

(c) Compare the observed and expected counts. Is the geometric distribution a good fit to these data?

6.6. Suppose a team has the on-base profile $(f_0, f_1, f_2, f_3, f_4)$, where the fractions f_0, f_1, f_2, f_3, f_4 represent the probabilities of the events walk (w), single (1), double (2), triple (3), and home run (h), respectively.

(a) Suppose that the team has three on-base events during a particular half-inning. Find the probability of not scoring any runs that half-inning. (Hint: Using the runner advancement rules described in this chapter, the sequences of events that will not produce a run are $\{www, w1w, 1ww, 11w, 2ww, 3ww\}$.)

(b) Suppose that you have 2 on-base events during a half-inning. Find the probability of scoring 0, 1, and 2 runs in that inning. (*Hint*: There are 25 possible sequences of two on-base events—six of these sequences will not score a run and five of these sequences will score exactly two runs.)

6.7. Table 6.12 gives the observed distribution of inning runs scored for 150 games in the 2001 season (the first through eighth innings).

Table 6.12. Frequency distribution of the number of runs scored in an inning for 150 games in the 2001 season

Runs scored	0	1	2	3	4	5	6 or more	Total
Count	1687	372	179	90	40	21	11	2400
Expected count								

Suppose that we apply the run scoring model described in the case studies to these data. For the 2001 teams, a typical on-base fraction is .3315, so we can estimate the probability of an out to be $p = 1 - .3315 = .6895$. Also for the 2001 teams, the on-base fractions

(of a walk, single, double, triple, and home run) are given by

$$f_0 = .2649, \quad f_1 = .4805, \quad f_2 = .1477, \quad f_3 = .0155, \quad f_4 = .0914.$$

By simulation, we used this model to estimate the probability that a team scores $0, 1, 2, \ldots$ runs in the half-inning. The probabilities from the model are displayed in Table 6.13.

Table 6.13. Probability distribution of the number of inning runs scored for using the run-scoring model

Runs scored	0	1	2	3	4	5	6 or more
Probability	.697	.149	.080	.043	.018	.009	.005

Use these probabilities to find the expected number of half-innings (out of 2400) where $0, 1, 2, \ldots$ runs would score. Put the expected counts in the first table. Comparing the observed and expected counts by the use of Pearson residuals, is the model effective in predicting the number of inning runs scored in major league games?

6.8. (Exercise 6.7 continued.) One assumption in our run scoring model is that the run production ability of the team doesn't change across innings. Are there particular innings during a game where a team is more or less likely to score runs? For our sample of 150 games during the 2001 season, the mean runs scored for each inning (first through eighth) was computed—a graph of the mean runs scored is shown in Figure 6.2. This figure illustrates that teams are indeed more or less likely to score in particular innings. Describe the variation that you see in the plot and explain why some innings are particularly good (or bad) for run production.

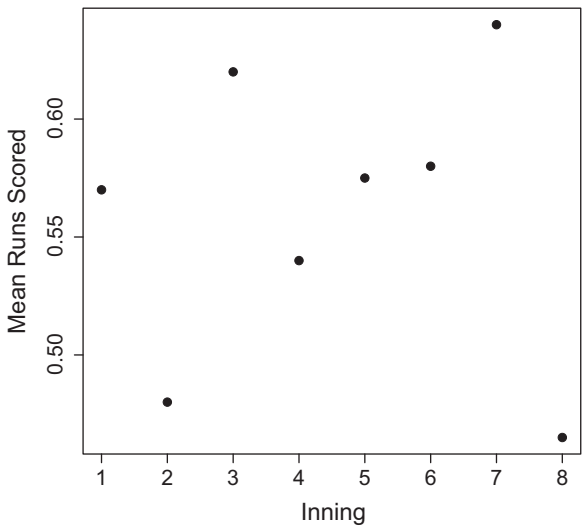


Figure 6.2. Plot of the mean number of runs scored in different innings for a sample of 150 games during the 2001 season.

6.9. In the 2001 baseball season, the Colorado Rockies were very effective in scoring runs and the Tampa Bay Devil Rays were relatively ineffective in scoring. Here are some hitting statistics for the two teams.

Team	H	2B	3B	HR	BB	OBP
COL	1663	324	61	213	511	.354
TBD	1426	311	21	121	456	.320

(a) Compare the on-base profile of the two teams.
(b) Compute the probability that the team scores at least one run with two base runners for the two teams.
(c) Compute the average number of runners per inning for the two teams.
(d) If you were the General Manager of Tampa Bay, what type of hitters would you try to sign for the following season? Explain.

6.10. Of all of the pitches thrown in the 2014 season, 26.4% ended the plate appearance with a ball hit in-play, a strikeout, a walk, or a hit-by-pitch. If X denotes the number of pitches thrown in a plate appearance, we can represent X with a geometric distribution where

$$\Pr(X = x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \ldots$$

where $p = .264$.
(a) Use this formula to find the probability the plate appearance ends in 1, 2, 3, 4, or more than 4 pitches. Put your probabilities in the following table.
(b) Of the 189,792 plate appearances in the 2014 season, find the expected number that end in 1, 2, 3, 4, and more than pitches.
(c) The actual number of plate appearances that end in 1, 2, 3, 4, or more than 4 pitches is displayed in the table. Compute Pearson residuals and conclude if the geometric distribution is a good fit to these data.

Number of Pitches	Probability	Expected	Observed
1			22248
2			31124
3			35032
4			36030
4 or more			65358

Further Reading

Basic discrete probability distributions, such as the binomial, negative binomial, and Poisson, are presented in Scheaffer and Young (2009). Mosteller (1952) uses probability modeling to understand the World Series baseball competition. D'Esopo and Lefkowitz (1977), Cover and Keilers (1977) and Albert and Bennett (2003), Chapter 8, discuss probabilistic models for run production.

10.1090/text/034/07

7

Introduction to Statistical Inference

What's On-Deck?

This chapter describes some fundamental notions about statistical inference in the context of baseball. One important idea in inference discussed in the introduction is the distinction between a player's ability and his performance. We are interested in a player's hitting ability, which can be measured by a probability p that represents the player's chance of getting on base in a single plate appearance. We don't know a player's ability p, but we learn about this value when we see the player perform in a series of games. In Case Study 7.2, we first consider the situation where you know a player's on-base probability p, and we look for basic patterns in his hitting performance in ten plate appearances. In Case Study 7.3, we use a simple simulation to describe how one can learn about a player's batting ability when one observes his performance in ten at-bats. We initially suppose that the player's on-base ability p is equally likely to be .2, .3, or .4. In the simulation, we choose a player's ability at random, and then simulate the process of having this player have ten plate appearances. We categorize the simulated players abilities and performances in a two-way table, and we perform inference by looking at the abilities of the players corresponding to a given number of on-base events in ten plate appearances. Case Study 7.4 illustrates the use of a basic formula for an interval estimate for a batter's ability p of a particular probability content. We use this interval formula in Case Study 7.5 to compare the hitting abilities of two great hitters, Wade Boggs and Tony Gwynn. We will see that it is difficult to distinguish the hitting abilities of two players based on only one season of data, but one can make a clearer distinction by looking at the pattern of hitting of the two players over all of the seasons of their careers.

7.1 Ability and Performance

Topics Covered: Distinction between ability and performance.

Recently a statistics class of 28 students played a spinner baseball game between Hall of Fame players of the American League and the National League. In this game, the AL defeated the NL 10-0. Why? I gave the students four possible explanations in a test question:

1. The AL was lucky the spins went the right way.
2. The AL players were better than the NL players.
3. The win was a result of luck AND the fact that the AL players were better.
4. There was some cheating going on.

How did the students answer this question? Here's the tally:

Answer	Tally
Luck	12
Ability	1
Luck & Ability	14
Cheating	1

Practically all of the students thought that luck played a role in the AL win.

Now what if I asked the students the same question regarding the Giants 2014 World Series win against the Royals. Would they also say that the Giants' win was partly due to luck? I'm guessing that most of the students would say that the Giants won because they were more skilled than the Royals. But actually luck or chance variation does play a big role in baseball games.

Why did the students think differently about the spinner game and the real World Series? Well, it is obvious that chance plays a big role in the spinner game since all of the outcomes depend on the spins of the spinners. We don't see any obvious dice or spinners in baseball games. But, if you think about it, there are a lot of chance elements in a baseball game (how the ball moves through the infield grass, how the bat hits the ball, how a player fields a groundball, etc.) that have random or unpredictable elements and this randomness can affect the outcome of the game.

> One primary role of the statistician is to understand how much of the variation in baseball data (and other data as well) can be explained by chance and how much is due to some real cause, like the skill of a player.

Baseball Hitting—Ability and Performance

It's helpful to look at the dictionary's definition of these two words.

- **Ability** is (1) the power to do or act (2) skill (3) power to do some special thing, natural gift, talent.
- **Performance** is (1) the act of carrying out, doing; performing (2) a thing performed; act; deed.

When we say a player has great ability to hit, we are talking about his skill or his gift to hit the ball. This player might have a great eye for the ball, have a nice swing, and make good contact with the ball. He may be a muscular individual with the capability to hit the ball a long way. In contrast, the performance of a player is actually how he hits during games. Here a batter's performance is simply the record of hitting that you observe in the box scores. If Albert Pujols hit two home runs today, he had a great batting performance.

These two words are connected. If a player has great batting ability, he will generally exhibit great batting performances. But it is important to distinguish ability and performance. Barry Bonds may be 1 for 10 in two playoff games. Although he had a weak batting performance, it doesn't mean that he's turned into a bad hitter. Likewise, a hitter who is 4-4 for one game isn't necessarily a much better hitter. He may actually be a mediocre hitter and happened to be lucky or fortunate that particular game.

We say that Mickey Mantle was a great hitter. Most baseball fans would agree that Mantle had great batting ability. Why? Because he had one great season? No. He had a number of great seasons. In other words, he exhibited a pattern of great batting behavior. Roger Maris, in contrast, had only a few great hitting seasons (especially 1961 when he hit 61 home runs). Since Maris didn't show a consistent behavior of great hitting for many seasons, there is some doubt that he really had great batting ability. In fact, some baseball people think that he was a bit lucky in 1961; he just happened to be the right hitter at the right time.

In statistical inference, the objective is to learn about a player's ability based on his performance in the field. We will see that it is difficult to learn a lot about a player's ability based on the performance of the player in a single season. In the next case study, we'll look at the relationship between performance and ability more carefully.

7.2 Simulating a Batter's Performance if His Ability is Known

Topics Covered: Random spinner as a probability model, simulation of a binomial experiment using dice.

We represent ability by means of a **probability model**. This is a simple randomization device with known properties. This model is said to be realistic if it generates baseball data similar to what is actually observed. A basic probability model is one that we have already seen, a spinner. Figure 7.1 displays a spinner that represents the outcome of a plate appearance:

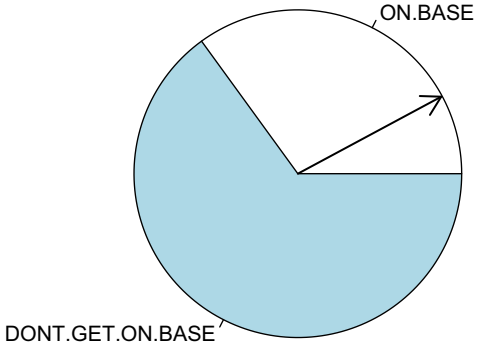


Figure 7.1. Random spinner to represent the outcome of a plate appearance.

In this spinner there are two possible outcomes: "on-base" and "not on-base". The ratio of the area of the on-base region to the area of the entire circle in the spinner represents the probability that the player will get on-base.

Another convenient probability model is based on a die. Imagine that you have a die with ten sides, labeled $0, 1, \ldots, 9$. Each side has the same chance of being rolled and each side has probability $1/10$. Then we can represent the result of a plate appearance by a roll of this die. Suppose that we decide to let

- rolls 1, 2, 3, 4 correspond to "on-base",
- rolls 0, 5, 6, 7, 8, 9 correspond to "not on-base".

The 10-sided die represents the player's ability to get on-base. A roll of this 10-sided die represents the performance of the player on a single at bat.

Let's say we specify a player's true ability. For example, for Mike Trout, we specify a probability of .400 of getting on base. What kind of performance can we expect, say in ten plate appearances?

In a statistics class, 10-sided dice were handed out to the students and a number of simulations were performed. Each simulation, we rolled the die ten times, representing Trout's ten PAs. Remember an on-base event corresponds to a 1, 2, 3, 4 on the die, and we keep track of the number of rolls that were either 1, 2, 3, 4. Each of the 30 students performed five simulations; the results of the 150 simulations are shown in Table 7.1. This table represents the performance of the batter over 150 10-plate appearance periods.

Table 7.1. Frequency distribution of number of times on-base using dice with 150 simulations

Number of Times on Base	Count	Probability
0	1	0.007
1	7	0.047
2	18	0.120
3	40	0.267
4	31	0.207
5	32	0.213
6	11	0.073
7	8	0.053
8	2	0.013
9	0	0
10	0	0

After we do this a sufficient number of times, we answer the following questions:

- What was the most likely number of times on-base for Trout?
  Here the most likely outcome in our 150 simulations was 3 times on-base; this occurred 40 times.
- What is the chance of this most likely number?
  The estimated probability of three times on-base is $40/150 = .267$.
- What is the probability that Trout will get on-base two or more times?
  In our 150 simulations, we see that Trout got on-base once or never 8 times. So, by subtraction, the frequency of 2 or more times is $150 - 8 = 142$. The estimated probability of 2 or more times is $142/150 = .947$.
- What is the probability that Trout will not get on-base during these ten PAs?
  In our 150 simulations, Trout didn't get on-base one time. So the estimated probability of not getting on-base is $1/150 = .007$.

In statistical inference, we're actually interested in looking at ability and performance the opposite way. We observe a hitter's performance. What does that tell us about a player's ability? This is what we'll talk about in the next section.

7.3 Learning About a Batter's Ability

Topics Covered: Modeling ability by using a spinner with probability p, simulating hitting data for a given ability, simulating abilities and performance, Bayes thinking, finding the most likely ability for a given performance.

We represent a player's ability by means of a probability model. Imagine that a player's talent to get on-base is represented by the spinner displayed in Figure 7.1. For a circle of area 1, the area of the on-base region is equal to the player's ability to get on-base. We call this area p—this is our measure of a player's hitting ability.

Suppose we know a player's hitting ability. Specifically, suppose we know his on-base probability p is equal to .4. This player comes to bat ten times during a doubleheader. How many times will he get on-base?

We did this simulation in a statistics class in the previous case study using dice. These probabilities aren't too precise since the simulation was performed only 150 times. On the computer I performed this experiment 1000 times. We assume that the player has an on-base probability of $p = .4$, and the results of 1000 doubleheaders were simulated where the player had ten opportunities to hit in each doubleheader. A frequency distribution of the times on-base is shown in Table 7.2.

Table 7.2. Frequency distribution of number of times on-base using a computer with 1000 simulations

# of Times on Base	Count	Probability
0	4	0.004
1	39	0.039
2	130	0.130
3	208	0.208
4	246	0.246
5	209	0.209
6	98	0.098
7	49	0.049
8	14	0.014
9	3	0.003
10	0	0
Sum	1000	1

Note that the most likely outcome for this hitter is four times on-base. This makes sense: if a player with probability of .4 of getting on-base has ten chances, one would expect him to get on-base $10(.4) = 4$ times.

Above we assumed that our player had an on-base probability of .4. What if his on-base probability was $p = .3$? On the computer I simulated the result of 1000 doubleheaders assuming $p = .3$. The results are shown in Table 7.3 and compared with the results when $p = .4$.

If we compare the results when $p = .4$ with the results when $p = .3$, we see some differences. If $p = .4$, the most likely outcome is four times on-base; in the case when $p = .3$ the most likely value is three times on-base.

Table 7.3. Frequency distribution of number of times on-base using computer with 1000 simulations for the cases where $p = .4$ and where $p = .3$

	$p = 0.4$		$p = 0.3$	
# of Times on Base	Count	Prob	Count	Prob
0	4	0.004	35	0.035
1	39	0.039	131	0.131
2	130	0.130	208	0.208
3	208	0.208	261	0.261
4	246	0.246	210	0.210
5	209	0.209	101	0.101
6	98	0.098	39	0.039
7	49	0.049	13	0.013
8	14	0.014	2	0.002
9	3	0.003	0	0
10	0	0	0	0
Sum	1000	1	1000	1

If a batter has a true .3 on-base probability, is it accurate to say that he's sure to get on-base three times (out of ten)? No. We see that the probability of three times on-base is only .261—it's actually more likely (.739) that he won't get on-base exactly three times.

In our simulations above, we assumed that we knew the player's ability (the value of p) and we looked at possible outcomes in a doubleheader (the batter's performance in ten plate appearances).

We actually want to solve the inverse problem. If a guy gets on-base four times (out of ten), what does that say about the guy's ability (value of p)? We solve this problem by an application of Bayesian thinking.

We use a simple simulation to see how we can learn about a batter's hitting ability. Suppose there is a manager named Casey who has a dugout of players who are equally divided between hitters of three abilities: the crummy hitters who have a true on-base probability of p of .2, the mediocre hitters who have $p = .3$, and the good hitters who have $p = .4$. Suppose Casey picks a player from the dugout at random and the player gets 10 chances to hit. Casey observes

- the player's ability (value of p),
- the player's performance the value of $x =$# of times on-base.

Three spinners were used in this simulation: one with an on-base probability of .2, another with an on-base probability of .3, and the third with an on-base probability of .4. We first choose a spinner at random. We rolled a single die; if the die roll was 1 or 2, we used the $p = .2$ spinner; if we rolled 3 or 4, we used the $p = .3$ spinner; if the roll was 5 or 6, we used the $p = .4$ spinner. We then spin the chosen spinner ten times.

In one particular simulation, we chose the bad spinner ($p = .2$). We spun it ten times with the results

NNNNNBNNNB

(N means not on-base, B means on-base.) So we observed $p = .2$ and $x = 2$. We repeat this simulation 1000 times we classify all of the results in Table 7.4 by ability (value of p) and performance (value of x):

Table 7.4. Two-way table of simulation results classified by the player's ability (value of p) and his performance (value of x)

		Performance (value of x)										
		0	1	2	3	4	5	6	7	8	9	10
Ability	0.2	38	105	113	61	22	5	0	0	0	0	0
(value	0.3	14	35	74	75	83	36	12	3	1	0	0
of p)	0.4	3	9	34	76	90	54	46	4	6	1	0

Before we observed any hitting data, what is the chance that the spinner chosen is a $p = .2$ spinner? Each spinner has the same chance of being chosen, so the probability that a $p = .2$ spinner is chosen is $1/3$. That is,

$$\text{Prob}(p = .2) = \text{Prob}(p = .3) = \text{Prob}(p = .4) = 1/3.$$

Now suppose that we observe four times on-base for our player? (That is, $x = 4$.) In Table 7.5, we focus on the column where $x = 4$. We see that we observed four on-base a total of $22 + 83 + 90 = 195$ times. Of these 195 times, 22 corresponded to a hitting probability of $p = .2$, 83 corresponded to a probability of $p = .3$, and 90 corresponded to a hitting probability of $p = .4$. Converting these counts to probabilities

		Count	Probability
Ability	0.2	22	$22/195 = 0.11$
(value	0.3	83	$83/195 = 0.43$
of p)	0.4	90	$90/195 = 0.46$

We see that this player is likely to be either a $p = .3$ hitter or a $p = .4$ hitter. The probability that he is a $p = .2$ hitter is only about 11%.

What if the hitter got on-base six times? (That is, you observed $x = 6$.) Now you would focus on the column of the table corresponding to $x = 6$.

		Count	Probability
Ability	0.2	0	$0/58 = 0$
(value	0.3	12	$12/58 = 0.21$
of p)	0.4	46	$46/58 = 0.79$

Here it is highly likely (probability .79) that $p = .4$. So if you observe six times on-base, you can conclude that the hitter is likely a true 40% hitter. Also you are pretty sure that the hitter is not a .2 hitter, since $\text{Prob}(p = .2) = 0$.

7.4 Interval Estimates for Ability

Topics Covered: Learning about a true ability (value of p) by means of an interval estimate, subjective interpretation of probability.

A Different Way of Thinking About a Probability

Al Gore said during a particular moment during the 2001 Presidential campaign that he had a 50% chance of winning the presidential election. What does 50% mean? It's a probability, but not in the relative frequency sense—the election is a one-time event and it doesn't make sense to imagine having many elections between Bush and Gore. Al Gore's 50% is a probability, but it represents Gore's degree of belief in winning the election. This interpretation of probability is subjective; different people can assign different probabilities to the event "Gore will win", since different people will have different opinions about the likelihood of the event. This interpretation of probability is relevant here. We will use probability to represent our beliefs about different batting abilities of a player.

Suppose we are interested in learning about the batting ability of Alex Rodriguez, who is called A-Rod. Recall that we represent the hitting ability of a player by a spinner, where there are two events, HIT and OUT, and the area of the HIT area represents the probability of a HIT. We denote this hitting probability by p this is a player's **true batting average**.

The first step in learning about A-Rod's hitting ability is to list some possible values of p. Let's assume that A-Rod's hitting probability could be

$$p = .1,\ .2,\ .3,\ .4,\ .5,\ .6,\ .7,\ .8,\ .9$$

and each of these nine values are equally likely. So Prob(A-Rod is a .100 hitter) = Prob(A-Rod is a .200 hitter) $= \cdots =$ Prob(A-Rod is a .900 hitter) $= 1/9$.

Now you are probably thinking these assumptions are silly. A-Rod can't be a .100 or .900 hitter, and even if there were nine possible batting abilities, it doesn't make sense to assign each value of p the same probability. (A-Rod is more likely to be a .300 or .400 hitter.) You're right that these are unrealistic assumptions, but it makes the calculations to follow easy to explain.

Next, we let A-Rod bat 20 times and we observe $x = \#$ of hits. Suppose we observe $x = 6$. (A-Rod gets six hits out of 20 at-bats.)

What can we say about A-Rod's true batting ability p? We use the simulation scheme that we introduced in Case Study 7.3 to learn about A-Rod's hitting ability.

1. We first choose an ability at random from the values $p = .1, .2, \ldots, .9$. Imagine nine spinners corresponding to the nine possible hitting probabilities and we choose one spinner at random.
2. We spin the chosen spinner (from part 1) 20 times and we count $x = \#$ of hits.

We repeat this process (choose a spinner and spin 20 times) a total of 10,000 times (on a computer).

The results are displayed in Table 7.5.

Remember that each time we do the simulation, we select an ability p (a spinner) and observe a hit number x. Look at the first count in the table, 136, in the upper left corner. This means that 136 times we chose the $p = .1$ spinner and observed no hits ($x = 0$).

Remember A-Rod got six hits (out of 20 AB). To see what we have learned about A-Rod's batting ability, we focus on the $x = 6$ column of the table, shown in Table 7.6.

We convert the counts to probabilities by dividing each count by the Total. These probabilities represent the likelihoods of A-Rod having different batting abilities.

Table 7.5. Two-way table of counts from simulation where there are nine possible abilities $p = .1, \ldots, .9$, and the batter comes to bat 20 times and gets x hits

		Ability (p)								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	0	136	17	0	0	0	0	0	0	0
	1	276	66	11	0	0	0	0	0	0
	2	311	140	32	4	0	0	0	0	0
	3	211	211	87	16	1	0	0	0	0
	4	104	211	156	42	5	0	0	0	0
	5	42	189	194	92	11	1	0	0	0
	6	7	128	212	137	37	4	1	0	0
	7	2	62	196	209	76	20	0	0	0
	8	0	24	134	190	121	44	3	0	0
	9	0	11	63	177	182	66	20	1	0
x	10	0	2	34	126	216	122	46	4	0
	11	0	2	9	82	176	175	69	11	0
	12	0	0	5	39	153	189	130	29	0
	13	0	0	2	19	87	169	168	56	0
	14	0	0	1	6	42	125	216	116	9
	15	0	0	0	0	16	79	208	223	28
	16	0	0	0	0	3	35	140	247	94
	17	0	0	0	0	0	11	61	250	224
	18	0	0	0	0	1	3	27	165	328
	19	0	0	0	0	0	2	13	68	293
	20	0	0	0	0	0	0	0	7	146

Table 7.6. Simulated values of ability p when the hitter gets $x = 6$ hits

	Ability (p)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Total
Count	7	128	212	137	37	4	1	0	0	526
Probability	.013	.243	.403	.260	.070	.008	.002	.000	.000	1

From the table we see

- Prob(A-Rod has a true .200 AVG) = 128 / 526 = .243.
- Prob(A-Rod's ability is between $p = .2$ and .4) = Prob($p = .2, .3, .4$) = .243 + .403 + .260 = .906.

We'd like to find an interval of ability values that are very likely. We call this interval a **probability interval** for the true batting average p.

To find a probability interval, we use the following table. In the left column, we put values of p, from most likely to least likely, and the second column contains the total probability content of these ability values.

1. We note from the table below that $p = .3$ is the most likely ability for A-Rod we put the value (.3) in the first column and the associated probability (.380) in the second column.

2. The next most likely value of p is $p = .4$ we put this value in the first column and the probability (.260) in the second column.
3. We continue doing this until the total probability (the sum of the probabilities) in the second column is a large number, say between 80 and 95 percent.

Values of p	Probability
.3	.403
.4	.260
.2	.243
TOTAL	.906

We see that the values .3, .4, .2 or the interval [.2, .4] is an 90.6% probability interval for A-Rod's ability p. The chance that A-Rod actually has one of these abilities is 90.6%. In other words, we are 90.6% confident that A-Rod's true batting average is between .2 and .4.

Actually this is very little information since we know nearly every player's batting average falls between .2 and .4. We really haven't learned much about ability based on only 20 AB. We need much more data to get a better handle on a player's ability.

There is a simpler recipe for constructing this interval estimate for a player's true batting average p.

1. We first compute a player's observed AVG $\hat{p} = x/\text{AB}$ (this is his reported batting average).
2. We compute the **standard error** (SE) which is a measure of the accuracy of $\hat{p}$ to estimate a true batting average p:

$$\text{SE} = \text{square root of } (\hat{p}(1-\hat{p})/\text{AB}).$$

3. A probability interval of content PROB will have the general form

$$(\hat{p} - z \times \text{SE},\ \hat{p} + z \times \text{SE}).$$

where z is a value of a standard normal density such that the upper tail area is equal to $(1 - \text{PROB})/2$.

Table 7.7 gives values of z for several choices of PROB.

Table 7.7. Value z of a standard normal density corresponding to various values of the probability content PROB

PROB	Upper tail area	z
.8	.1	1.28
.9	.05	1.645
.95	.025	1.96

We illustrate using this recipe. A-Rod got six hits in 20 at-bats, so

$$\hat{p} = \frac{6}{20} = .3.$$

The standard error is

$$\text{SE} = \text{square root of } (.3(1-.3)/20) = .102.$$

From Table 7.7, we see that if we are interested in a probability content of PROB = .900, we should use a value of $z = 1.645$. So a 90% interval estimate for A-Rod's batting ability p is

$$[.3 - 1.645 \times .102 \text{ to } .3 + 1.645 \times .102] = [.132, .468].$$

(This gives a similar answer to what we got using our simulation method.)

Generally, you learn more about a player's true batting average by observing more AB. Suppose A-Rod plays for LA next year and has a good season: 210 hits in 600 AB.

His observed batting average is $= 210/600 = .350$. We compute SE = square root of $(.35(1 - .35)/600) = .019$. A 90% probability interval for A-Rod's p would be

$$.350 - 1.645 \times .019 \text{ to } .350 + 1.645 \times .019 = (.319, .381).$$

So actually, A-Rod's batting ability in 2001 could be as low as .319 or as high as .381. So we don't learn as much about a player's true batting average as we might expect.

7.5 Comparing Wade Boggs and Tony Gwynn

Topics Covered: Interval estimates for a proportion, comparing proportions by use of interval estimates, time-series plots.

In this case study we compare two of the greatest "hitters in average", Wade Boggs and Tony Gwynn. Boggs played many years for the Red Sox. He was very effective in getting his bat on the ball (a so-called contact hitter) and won many batting crowns for the best batting average. (He's also known for his preference in diet: he had chicken before every game.) Gwynn played for the Padres his entire career and is also considered a great contact hitter. Michael Schell, in *Baseball's All-Time Greatest Hitters*, rates Gwynn as the best hitter of all time with respect to batting average.

In Table 7.8, I present the batting averages for Boggs and Gwynn for all years that they played in the major leagues. I give the ages for each player each season; we'll make comparisons of the players for different ages.

Let's focus on the batting averages when both players were 31. At this age, Gwynn had 168 hits in 530 at-bats for a batting average of .317; Boggs had 205 hits in 621 at-bats for an average of .330.

Who Was a Better Hitter at Age 31?

I didn't ask who performed better at age 31. That question is simple to answer: Boggs had the higher batting average (.330 compared to .317). I'm actually asking about the hitting abilities of Gwynn and Boggs. Can we say that Boggs had a better hitting ability this year?

Remember we represent a hitter's ability by means of a spinner where the HIT area is equal to p. So Gwynn and Boggs this year have abilities (or spinners) with hitting probabilities of p_G and p_B. (We call these the true batting averages.)

Can We Say That p_G Is Larger Than p_B?

We answer this question by computing probability intervals for the two true batting averages. We use 95% intervals (recall the formula from Case Study 7.4).

Table 7.8. Career batting statistics for Tony Gwynn and Wade Boggs

	Tony Gwynn			Wade Boggs		
Age	AB	H	AVG	AB	H	AVG
22	190	55	0.289			
23	304	94	0.309			
24	606	213	0.351	338	118	0.349
25	622	197	0.317	582	210	0.361
26	642	211	0.329	625	203	0.325
27	589	218	0.370	653	240	0.368
28	521	163	0.313	580	207	0.357
29	604	203	0.336	551	200	0.363
30	573	177	0.309	584	214	0.366
31	530	168	0.317	621	205	0.330
32	520	165	0.317	619	187	0.302
33	489	175	0.358	546	181	0.332
34	419	165	0.394	514	133	0.259
35	535	197	0.368	560	169	0.302
36	451	159	0.353	366	125	0.342
37	592	220	0.372	460	149	0.324
38	461	148	0.321	501	156	0.311
39	411	139	0.338	353	103	0.292
40	127	41	0.323	435	122	0.280
41	102	33	0.324	292	88	0.301

- Boggs: $\hat{p}_B = .330$ with a sample size $n_B = 621$; we compute a 95% probability interval for to be (.293, .369).
- Gwynn: $\hat{p}_G = .317$ with a sample size $n_G = 530$; we compute a 95% probability interval for to be (.278, .359).

Let's interpret what these mean. We are 95% confident that Boggs' true batting average (at age 31) is between .293 and .369; that is, the probability that falls between .293 and .369 is 95%. Likewise, we are pretty sure (with confidence .95) that Gwynn's true batting average that year is between .278 and .359. We have plotted the two intervals in Figure 7.2.

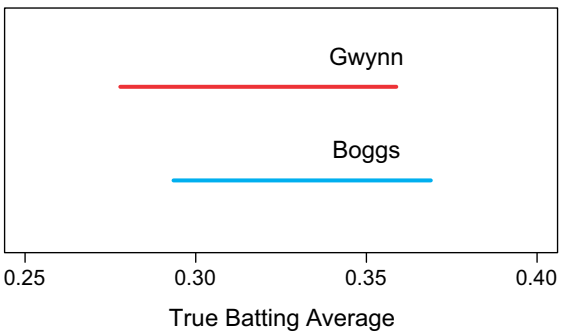


Figure 7.2. Display of probability intervals for Gwynn's and Boggs' true batting abilities.

These two intervals overlap so it is possible that Gwynn really had a higher batting ability that year and by chance or luck variation, Boggs just happened to perform better. So really a 13 point difference in season batting averages is not enough to say that one hitter is better than another hitter.

Of course, we know much more about these two hitters than their performance in a single year. In Figure 7.3, we plot the two players' batting averages across all years. (We are plotting AVG against AGE.)

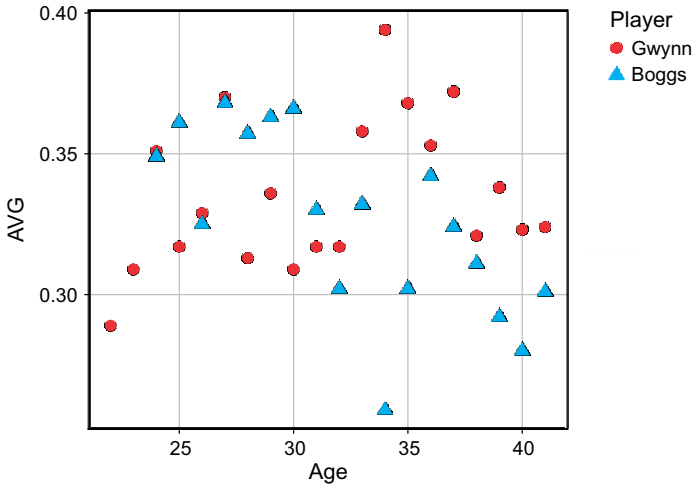


Figure 7.3. Graph of Tony Gwynn and Wade Boggs season batting averages against age.

This graph is hard to interpret. Why? First, you should notice a lot of up and down variation in both players' batting averages. This is very common. Let me explain why. Suppose that we assume that Gwynn is a true $p = .338$ hitter for all of the years of his career (note that .338 is Gwynn's career batting average). I simulated hitting data for Gwynn using his at-bat numbers. Figure 7.4 shows the plot of one simulation.

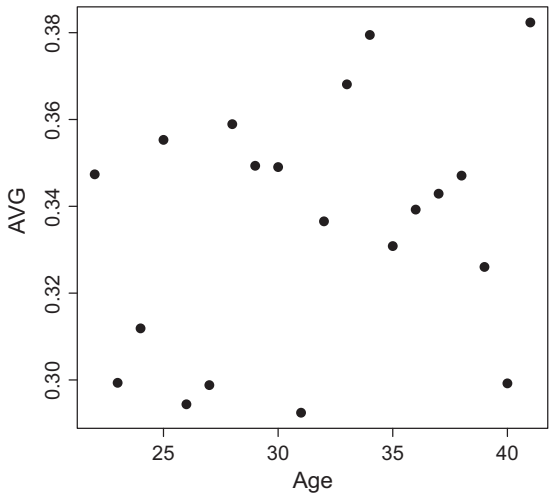


Figure 7.4. Simulated batting average for a true $p = .338$ hitter using the at-bat numbers of Tony Gwynn.

This is an interesting plot. We're assuming that Gwynn has the same batting ability across his whole career. His hitting probability is .338 his first year, .338 his second year, etc. But his actual batting performance (his season batting average) shows great variability. The first year he bats .374, the second year he bats .322 this is the natural variation even when Gwynn is assumed to have the same true batting average for all years. So you will typically see a lot of variation in season batting averages for any player that you look at.

Does this mean that we can't tell if Gwynn or Boggs is the better hitter? No—we can draw a conclusion by looking at the performance of both hitters for all years.

Let's look at Figure 7.3 again. Despite the great up and down fluctuation in the batting averages, we see some general patterns.

- Boggs's season batting averages were high relatively early in his career (pre 30). After 30 he appears to be a weaker hitter despite his good year at age 36.
- Gwynn, in contrast, has maintained a high (around .350) batting average for most of his career.
- Boggs and Gwynn are similar hitters before 32; after 32, Gwynn was consistently higher than Boggs in batting average

Based on the above observations, I think Gwynn has been the better hitter on the basis of batting average. You can only make this conclusion based on 20 years of data, not just a single season.

7.6 Exercises

7.0. In the following table, some batting statistics for Rickey Henderson for the 1990 and 1991 seasons are displayed.

Season	AB	BB	PA	OBP
1990	489	97		.439
1991	470	98		.400

(a) For each season, compute the number of (approximate) plate appearances by adding at-bats (AB) to walks (BB). Put these values in the table.
(b) Let denote Henderson's true OBP in 1990. Construct a 95% probability interval for using the number of plate appearances (PA) and his on-base fraction OBP.
(c) Construct a 95% probability interval for Henderson's true OBP in 1991.
(d) Based on your work in (b) and (c), can you say that Henderson's on-base ability was better in 1990 than 1991? Explain how you reached this conclusion.

7.1. Table 7.9 shows the batting average of twelve ballplayers for the years 2013 and 2014.
(a) Draw a scatterplot of these data, plotting the 2013 AVG on the horizontal axis and the 2014 AVG on the vertical axis.
(b) Comment on any relationship between the 2013 and 2014 AVGs that you see from the scatterplot.
(c) Guess at the value of the correlation coefficient.
(d) The least-squares line to these data is

$$\text{AVG.2014} = 0.190 + 0.269 \times \text{AVG.2013}.$$

Table 7.9. Batting averages for twelve players for the 2013 and 2014 seasons

Name	AVG.2013	AVG.2014
Adrian Beltre	0.315	0.324
Michael Bourn	0.263	0.257
Starlin Castro	0.245	0.292
Shin-Soo Choo	0.285	0.242
Zack Cozart	0.254	0.221
Josh Donaldson	0.301	0.255
Todd Frazier	0.234	0.273
Evan Longoria	0.269	0.253
Jed Lowrie	0.290	0.249
David Ortiz	0.309	0.263
Buster Posey	0.294	0.311
Jean Segura	0.294	0.246

Suppose that a player bats .320 in 2013. Use this line to predict his AVG in 2014.

(e) Explain why there is a positive correlation between a player's 2013 AVG and his 2014 AVG.

7.2. Consider the simple experiment of tossing a fair coin 20 times and recording the number of heads.

(a) What is the probability of a head on a single toss? Why?

(b) Toss a fair coin 20 times and record the number of heads. Put your sequence of Heads and Tails and the number of heads in the following table.

Toss	1	2	3	4	5	6	7	8	9	10
Result (H or T)										
Toss	11	12	13	14	15	16	17	18	19	20
Result (H or T)										

(c) Combine your result (Number of Heads) with the results of other students in your class. Plot the results using a dotplot on the number line below.

```
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
                  Number of Heads
```

(d) Describe the basic features of the dotplot of Number of Heads that you constructed in (c).

(e) Suppose that you get seven heads in 20 tosses. Since the fraction of heads is only $7/20 = .35$, does that mean that the coin is not fair? Why or why not?

7.3. Suppose that Dustin Pedroia is really a .300 hitter. That is, his true probability of getting a hit on a single at-bat is .3. We represent his batting ability by means of a spinner shown in Figure 7.5 with a HIT area of .3.

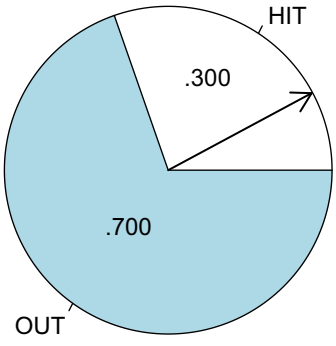


Figure 7.5. Spinner to simulate an at-bat for a hitter with a probability of a hit equal to $p = .3$.

Suppose Pedroia comes to bat 12 times over the weekend. We can simulate 12 at-bats for Pedroia by spinning the spinner 12 times and counting the number of times the spinner lands in the HIT region. We repeated this process 20 times and below are the number of hits that we observed for these 20 weekends.

4, 2, 5, 8, 1, 5, 5, 0, 2, 4, 5, 1, 2, 7, 3, 2, 3, 3, 2, 3

(a) Construct a frequency table of these hit numbers and put your counts in the following table.

Number of hits	0	1	2	3	4	5	6	7	8	9	10	11	12
Count													

(b) What is the most frequent number of hits that Pedroia gets?
(c) Find the probability that Pedroia gets five or more hits over the weekend.

7.4. (Exercise 7.3 continued.) Suppose that Pedroia really is a .400 hitter, which means that the chance that he gets a hit is .4. We think of a spinner with a HIT area of .4 (instead of .3 as in the previous exercise) and we simulate the results of 12 at-bats by spinning this .4 spinner 12 times. We repeated this (spinning the spinner 12 times) 20 times, obtaining the following number of hits.

5, 3, 7, 7, 1, 6, 4, 3, 2, 5, 2, 5, 6, 3, 7, 4, 6, 4, 6, 4

Construct a frequency table of these hit numbers using the table below and answer questions (b) and (c) from Exercise 7.3.

Number of hits	0	1	2	3	4	5	6	7	8	9	10	11	12
Count													

7.5. (Exercise 7.3 continued.) Is it possible to distinguish a true .300 hitter from a true .400 hitter on the basis of 12 at-bats? First, using a spinner, we let a true .300 hitter bat 12 times (a weekend of hitting), and we repeat this simulation for 1000 weekends to get 1000 hit numbers. Likewise, we let a true .400 hitter bat for 1000 weekends, obtaining 1000

simulated hit numbers. We construct count tables in Table 7.10 for the number of hits for both types of hitter.

Table 7.10. Simulated number of hits in 12 at-bats for true .300 and .400 hitters

	Number of hits										
	0	1	2	3	4	5	6	7	8	9	10
.300 hitter	11	65	166	229	241	161	85	34	8	0	0
.400 hitter	3	24	70	125	208	211	173	117	47	19	3

(a) Find the probability that a true .300 hitter will get exactly five hits during a weekend.
(b) Find the probability that a true .400 hitter will get exactly five hits during a weekend.
(c) Suppose that you don't know the batter's ability: he either could be a .300 or .400 hitter. Given that you observe this batter get exactly five hits over the weekend, use Table 7.10 to find the probability that the hitter has a .300 true batting average and a .400 true batting average.
(d) From your computation in (c), are you pretty sure that the hitter has a .400 batting average? Can you really learn much about a batter's ability on the basis of a weekend of hitting (12 at-bats)?

7.6. Suppose that a manager has 11 different types of hitters on his team. One player hits with a true probability of .200, another hits with a true probability of .210, the third player hits with a true probability of .220, . . . , and the last player hits with a true probability of .300. Consider this hypothetical experiment the manager chooses one player at random from the 11, and then this player comes to bat 20 times. Each time, the manager records

$$p = \text{the true batting average of the player selected,}$$
$$x = \text{the number of hits of this player in 20 at-bats.}$$

Suppose that this hypothetical experiment is repeated 10,000 times, obtaining 10,000 values of p (the true batting average) and x (the number of hits). These values are organized by means of the two-way table shown in Table 7.11.

(a) How many times was a player with a .22 ability chosen and four hits were observed? Find the probability that a player with a .22 ability is chosen and four hits are observed.
(b) What is the most likely number of hits observed? What is the probability of this number of hits?
(c) Suppose that a hitter only gets two hits (out of 20). What is the chance that he is a .200 ability hitter?
(d) If a hitter only gets two hits, find the probability that the hitters ability (p) is at most .25.
(e) If a hitter gets two hits, find the smallest interval of values that contains the ability p with a probability at least .9.

7.7. (Exercise 7.6 continued.)

(a) Suppose that the player gets five hits. Use the two-way table to find the probability that the player has a true batting average (p) greater than .25, and the probability

Table 7.11. Two-way table of simulated values of true batting average and number of hits in simulation

No. of hits	True Batting Average											Total
	0.2	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.3	
0	13	8	8	6	2	3	2	4	0	0	0	46
1	49	52	37	27	33	19	12	11	7	7	5	259
2	127	105	99	82	74	62	50	29	38	32	23	721
3	169	155	146	161	134	114	89	99	67	75	73	1282
4	190	180	201	167	220	175	155	181	138	149	109	1865
5	174	167	170	183	172	199	178	169	175	175	165	1927
6	99	101	114	133	149	155	160	172	167	175	181	1606
7	53	67	70	88	89	119	103	123	137	151	158	1158
8	19	27	32	34	45	69	54	67	64	108	119	638
9	7	7	13	16	14	21	31	44	51	44	57	305
10	2	4	3	8	6	9	14	18	24	26	25	139
11	0	0	2	2	1	2	2	7	5	12	9	42
12	0	0	1	1	0	1	1	0	1	3	1	9
13	0	0	0	0	0	0	0	0	1	1	0	2
14	0	0	0	0	0	0	0	0	0	1	0	1
Total	902	873	896	908	939	948	851	924	875	959	925	10000

that the player has a true batting average .25 or smaller. Put your answers in the table below.

Observe 5 hits in 20 at-bats

Event	Probability
Player's true average is greater than .25	
Player's true average is .25 or smaller	

(b) Repeat (a) assuming that the player gets 8 hits.

Observe 8 hits in 20 at-bats

Event	Probability
Player's true average is greater than .25	
Player's true average is .25 or smaller	

(c) Compare your answers to parts (a) and (b). In which case (observing five hits or observing eight hits) did you learn more about the player's true batting ability? Why?

7.8. Suppose Josh Donaldson has 50 at-bats at some point in the 2014 season and has 14 hits.
(a) Find Donaldson's current batting average (his average over the 50 at bats).
(b) Find a 95% probability interval for Donaldson's true batting average for the 2003 season.
(c) Is it possible that Donaldson has a true batting average of .320? Is it likely? Why?

7.9. Suppose you are following Paul Goldschmidt's batting average during the 2015 season.
(a) Suppose that Goldschmidt has 40 hits after 100 at-bats. Find a 95% probability interval for Goldschmidt's true 2003 batting average.
(b) Suppose that Goldschmidt has 80 hits after 200 at-bats. Find a 95% probability interval for Goldschmidt's batting average.
(c) Compare the two probability intervals you computed in (a) and (b) with respect to the center and length of the interval.

7.10. In 2014, Victor Martinez had 188 hits in 561 at-bats, and Michael Brantley had 200 hits in 611 at-bats.
(a) Find the 2014 AVGs for Martinez and Brantley.
(b) Find 95% probability intervals for Martinez and Brantley's true batting averages.
(c) Are you confident that Martinez really was a better hitter than Brantley during the 2014 season? (Use the probability intervals you calculated in (b) to answer the question.)

Further Reading

Albert and Bennett (2003), Chapter 3, introduce statistical inference in the context of baseball. A spinner is used to model a player's hitting ability, and they use a simulation, such as described in Case Study 7.3, to learn about a player's ability based on his batting performance. Introductory inference is described from a Bayesian perspective in Berry (1996) and Albert and Rossman (2001). Basic inferential methods for one proportion are contained in Devore and Peck (2011) and Moore, McCabe and Craig (2012).

10.1090/text/034/08

8

Topics in Statistical Inference

What's On-Deck?

In this chapter, we focus on two interesting statistical inferential topics related to baseball the interpretation of situational data and the search for true streakiness in baseball data. Today baseball hitting and pitching data is recorded in very fine detail, and it is popular to report the performance of hitters and pitchers in a large number of situations. For example, we record how a player hits in home and away games, against left and right-handed pitchers, during different months of the season, during different pitch counts, and against different teams. The reporting of this situational data raises an interesting question: how much of the variation in this data corresponds to real effects and how much of the variation is attributed to luck or chance variation? In Case Study 8.1, we look at the situational hitting data that is reported for a single player. When we look at the situational hitting data for the home vs. away situation for a number of players (Case Study 8.2), we will see some interesting effects. Some players will hit for a much higher average during home games and other players perform much better during away games. But when we graph the situational effects for a group of players during two consecutive years, we will see that there is no association. In other words, players don't appear to possess an ability to perform unusually well (or poorly) during home games.

In Case Studies 8.3 and 8.4, we describe some useful statistical models for situational data. One can represent the hitting abilities of players by means of a normal or bell-shaped curve. Many situations are in the "no effect" scenario here the player has the same probability of getting a hit in either situation. Other situations are so-called biases the situation, such as playing at home will add a constant number to every player's hitting probability. The most interesting situation can be regarded as an "ability effect" where the particular situation is to one player's advantage, and to another player's disadvantage. Generally speaking, most of the variation in the reported situational data is essentially noise or chance variation, and it is difficult to pick up real situational effects in baseball data.

In the last two case studies, we look at the general topic of streakiness in baseball hitting data. In Case Study 8.5, we look at a player, Michael Brantley, and discuss ways of measuring streakiness in his day-to-day hitting data. By looking at some of these streaky statistics, there may be some evidence that Brantley is genuinely a streaky hitter. But Case Study 8.6 shows that these patterns of streakiness are also common in results in tossing a coin many times. The conclusion from this brief study is that genuine streakiness in hitting data is difficult to detect.

8.1 Situational Hitting Statistics for Mike Trout

Topics Covered: Introduction to situational hitting data.

In this chapter, we first focus on situational statistics. These are currently very popular among baseball fans. The `fangraphs` website contains a remarkable collection of hitting and pitching situational stats and it's fun reading.

When you watch a baseball game, you'll hear the announcer say something like "Francisco Lindor has a .421 batting average when he is facing Cole Hamels at Progressive Field." You are supposed to be surprised by this statement. If you are watching Lindor bat against Hamels in Progressive Field, you might expect Lindor to get a hit.

But is this the right interpretation? How can we make sense of all of these "interesting" situational stats that we hear in the media?

Let's focus on Mike Trout, the Angels hitter, who had a great 2014 season. Here are Trout's basic hitting stats that we are familiar with.

	Avg	PA	OBP	SLG
Total	.287	705	.377	.561

Next, we see how Trout did against left-handed pitchers and right handers. Generally a hitter bats better against a pitcher who throws with an arm opposite from which he takes his batting stance. Managers believe in this effect and make substitutions based on this belief. It is surprising that, Trout, a right-handed hitter, had a better AVG and SLG against right-handers.

Handedness	PA	AVG	OBP	SLG
vs Left	176	.275	.386	.523
vs Right	529	.291	.374	.574

Next, we see how Trout batted at home and on the road. Generally it is believed that ballplayers perform better at home (more comfortable surroundings, loving fans, home cooking, etc.) Trout had a higher slugging percentage at home, although his batting average was smaller at home.

	PA	AVG	OBP	SLG
Home	349	.282	.378	.578
Away	356	.292	.376	.545

The next situation refers to the runners on base. We see that Trout was a better hitter (actually performed better) with men on base compared with the bases empty.

Runners on Base	PA	AVG	OBP	SLG
Bases Empty	408	.275	.360	.544
Men on Base	297	.306	.401	.587

Next we see how Trout hit each month of the baseball season. There is quite some variation here. He was hot in June and cold in August. How do we interpret these numbers? Was Trout a much better hitter in March/April than in May?

Month	PA	AVG	OBP	SLG
Mar/Apr	124	.321	.403	.596
May	118	.263	.356	.495
Jun	102	.361	.471	.759
Jul	126	.265	.341	.504
Aug	132	.254	.318	.492
Sept/Oct	103	.274	.398	.571

The next situations relate to the notion of leverage which is a measure of the importance of the particular game situation (inning, outs, and runners on base) towards the goal of winning the game. The batting outcome in a "low leverage" situation has little effect on the game outcome. In contrast, a hit in a "high leverage situation" may have a large influence on the game outcome. We see that Trout did especially well in high leverage situations.

	PA	AVG	OBP	SLG
Low Leverage	387	.276	.357	.501
Medium Leverage	250	.288	.380	.611
High Leverage	68	.358	.485	.755

Next, we see how Trout hit plate apperances that passed through different pitch counts. Generally players bat better when they have a pitch advantage (like 1-0, 2-0 or 3-0), and bat much worse when there are two strikes (the pitcher has the advantage). Trout's averages are consistent with this pattern.

Pitch Count	PA	AVG	OBP	SLG
Through 3-0	46	.375	.761	.563
Through 3-1	93	.280	.602	.560
Through 3-2	142	.290	.493	.630
Through 2-0	103	.385	.602	.600
Through 1-0	308	.312	.435	.611
Through 2-1	181	.350	.497	.686
Through 1-1	360	.273	.367	.562
Through 0-1	370	.268	.335	.539
Through 2-2	227	.260	.348	.505
Through 1-2	249	.211	.265	.404
Through 0-2	131	.216	.244	.440

Next, we see how Trout performed when he hit grounders, fly balls (flies), or line drives (liners). Trout had a very high slugging percentage when he hit a fly ball or a line drive.

Batted Ball Type	PA	AVG	OBP	SLG
Grounders	145	.352	.352	.400
Flies	202	.330	.317	1.021
Liners	81	.734	.716	1.038

Last, we see how Trout did on balls hit to the left, center, and right portions of the baseball field. We see that Trout hit better when the ball was hit to left and center—this is a typical pattern for a right-handed hitter.

Handedness	PA	AVG	OBP	SLG
as R to Left	158	.494	.487	.853
as R to Center	152	.473	.461	1.068
as R to Right	118	.228	.220	.412

To sum up, although Trout overall hit for a .287 batting average in the 2014 season, he appears to bat much better (or much worse) in particular situations. Specifically, we see that Trout

- batted 31 points better with men on base
- batted 10 points worse at home games,
- batted for a high average in June and March/April and a low average in August.
- hit for a .358 average in situations with high leverage

The main question we will address in the following case studies is:

Do these observed situational effects correspond to real effects?

For example, can we say that there is really an advantage to hitting in situations of high leverage? Maybe Trout has the same ability to get a base hit in low and high leverage situations and, by luck or chance variation, he happened to hit better in high leverage situations this year. Or maybe hitters have a general advantage in situations of high leverage. We will see in this chapter that much of the variation that we see in situational data such as Trout's batting averages can be explained by chance, and it is relatively difficult to pick out real situational effects.

8.2 Observed Situational Effects for Many Players

Topics Covered: Stemplot, scatterplot, relationship between two variables, distinction between ability and performance.

To start to get an understanding of the patterns of situational data, we look at Table 8.1 where we see the HOME on-base percentage and AWAY on-base percentage for a group of 20 hitters for the years 2013 and 2014. Let's focus first on the 2013 data. For each player, we compute the difference

$$\text{DIFF} = \text{HOME OBP in 2013} - \text{AWAY OBP in 2013}.$$

For example, Robinson Cano's OBP was .401 at home games and .366 at away games for a difference of DIFF $= .401 - .366 = .035$. Figure 8.1 displays a dotplot of the differences of the home and way OBPs in the 2013 season. What can we see in this dotplot?

- We can compute the median difference; it is $-.003$. This means, on average, that a hitter bats 3 points lower during HOME games than during AWAY games. This is a little surprising since we thought players hit better at home, on average.

Table 8.1. Home and away on-base percentages for a group of 20 hitters for the 2013 and 2014 seasons

	2013			2014		
Name	Home	Away	Diff	Home	Away	Diff
Pedro Alvarez	.282	.310	−.028	.315	.309	.006
Adrian Beltre	.375	.367	.008	.415	.364	.051
Carlos Beltran	.338	.340	−.002	.319	.280	.039
Michael Bourn	.314	.317	−.003	.340	.294	.046
Robinson Cano	.401	.366	.035	.377	.386	−.009
Matt Carpenter	.432	.353	.079	.353	.396	−.043
Alejandro De Aza	.338	.308	.030	.320	.308	.012
Ian Desmond	.361	.301	.060	.318	.308	.010
Josh Donaldson	.367	.400	−.033	.322	.361	−.039
Alcides Escobar	.252	.266	−.014	.324	.310	.014
Jedd Gyorko	.301	.301	0	.294	.264	.030
Eric Hosmer	.352	.355	−.003	.260	.370	−.110
Juan Lagares	.266	.297	−.031	.314	.327	−.013
James Loney	.290	.404	−.114	.352	.321	.031
Justin Morneau	.336	.312	.024	.363	.364	−.001
Buster Posey	.366	.377	−.011	.307	.417	−.110
Alex Rios	.309	.338	−.029	.319	.304	.015
Kyle Seager	.316	.360	−.044	.370	.301	.069
Jean Segura	.314	.344	−.030	.305	.274	.031
Ben Zobrist	.368	.341	.027	.379	.329	.050

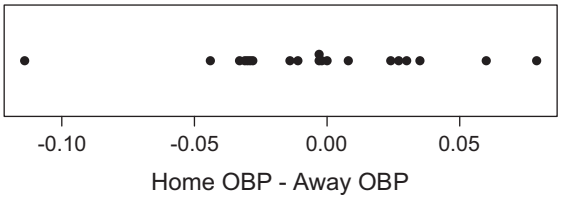


Figure 8.1. Dotplot of differences in home and away on-base percentages for 20 players in the 2013 season.

- However, there is a tremendous range in these difference values. One hitter had a −.114 difference—this player batted 114 points better on away games. In contrast, another hitter had a difference of .079. This player batted 79 points higher at home. Generally, situational data is interesting since one will typically see a great range of differences there will be many large positive values and large negative values.

Since we see many "extreme" home/away effects, it is tempting to try to explain why a particular player is doing so well (or so poorly) at home. But do players really have different abilities to use the home field advantage? Is it possible that one player uses the home field to great advantage in his hitting, while a second player has the same batting ability at home and

away games? We can learn about the presence of home/away batting abilities by looking at data for many players.

Demonstration that Players Have Different Batting Abilities

It is obvious to most baseball fans that players have different batting abilities. But how can we demonstrate this fact from hitting data? Let's look at the home hitting data in Table 8.1. For each player, we collect his 2013 home on-base percentage and his 2014 home on-base percentage. Figure 8.2 displays a scatterplot of the 2013 and 2014 home averages for the 20 players. Looking at the graph, we see that the points drift from the lower left to the upper right regions this means that the 2013 and 2014 batting averages are positively associated. (The correlation between the two variables is .34.) This means that players who hit well at home in 2013 tend also to hit well at home the following year. Likewise, a weak hitting season by a player in 2013 tends to be associated with a weak-hitting performance the following season. The explanation for the positive association in the graph is that players have different batting abilities—the better hitters correspond to points in the upper right section of the graph and the weak hitters correspond to points in the lower left section.

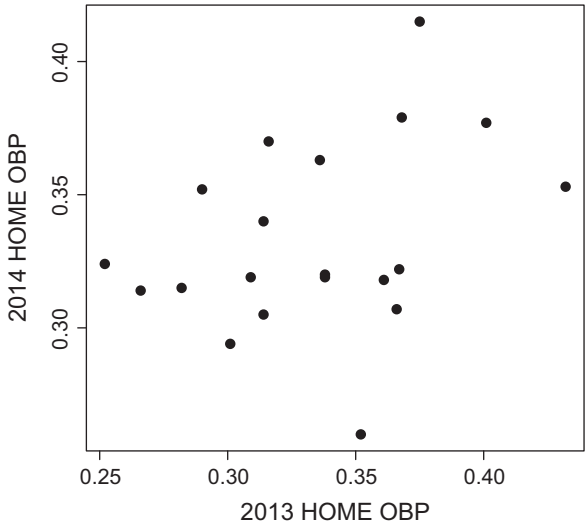


Figure 8.2. Scatterplot of 2013 and 2014 home on-base percentages for 20 players.

Do Players Have Different Home/Away Abilities?

We use the above strategy to search for situational abilities. Let's consider a particular ballplayer who bats 50 points better at home than away games. Is this real? In other words, does this player really have an extra ability to hit well during home games? We can see if this is a real effect by looking at next year's performance. If this player continues to hit much better at home, then we would be more confident that he has some extra home batting ability. If we have home/away batting data for a group of players for two consecutive years, we look for a general relationship between a player's home/away effect one year with the corresponding effect the next year.

In Table 8.1, we have computed the difference between the home and away on-base percentages (DIFF) for both the 2013 and 2014 baseball seasons for our 20 players. To see if there is a

general relationship between a player's 2013 DIFF and his 2014 DIFF, we construct a scatterplot in Figure 8.3.

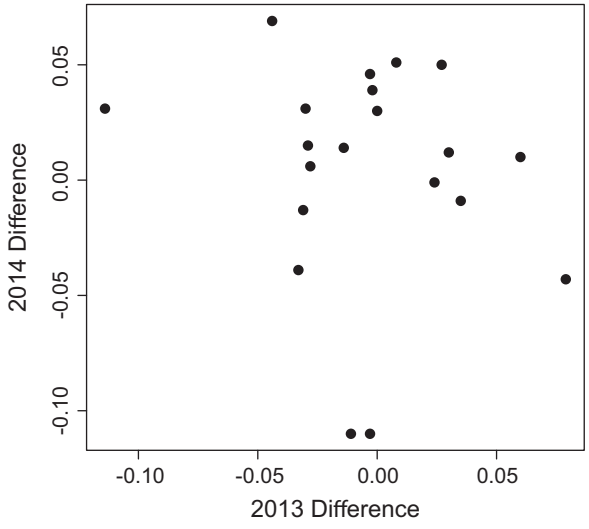


Figure 8.3. Scatterplot of differences in home and away on-base percentages for 20 players in 2013 and 2014 seasons.

In contrast to the pattern in Figure 8.2, we don't see a strong trend in this graph. (The correlation between the two variables in this case is actually the negative value $-.16$.) There does not appear to exist a tendency to show the same type of home/away effect for two consecutive years. This suggests that the home/away effect is not an ability characteristic. This means that a player who hits unusually well at home one particular year generally will not hit unusually well at home the next year likewise poor home hitters one year will not be poor hitters at home the following year. Remember that there will be a positive association between a player's batting average (or any other batting measure) for two consecutive years—this means that players have intrinsic batting abilities. But there is no general tendency for players to hit better (or worse) in the home/away situation for two consecutive years.

8.3 Modeling On-Base Percentages for Many Players

Topics Covered: Probabilities of hitting for many players, a random effects model, assessing the goodness of fit of a model, normal distribution.

In our earlier case studies:

- We distinguished between the **Ability** of a hitter from his **Performance**.
- We observe a hitter's performance during the season; we don't know a batter's ability, but we'll learn about it from the hitting data.
- We model a hitter's ability by a spinner. The spinner has two areas, "On-Base" and "Not On-Base", and the area of the On-Base region is given by p. This is our measure of ability.

```
1 | 2: represents 0.012
 leaf unit: 0.001
           n: 238
   24 | 046
   25 | 46
   26 | 358
   27 | 12388
   28 | 0113556666777778899
   29 | 1122233456777991010
   30 | 000001234555556678889999
   31 | 00112233344555567777788999
   32 | 000001111123334444444455566777888889
   33 | 01122233344556666667777899999
   34 | 0012222223345566777999
   35 | 111123445556689
   36 | 11222444579
   37 | 00123335577
   38 | 2235668
   39 | 456
   40 | 239
   41 | 0
   42 |
   43 | 2
```

Figure 8.4. Stemplot of the on-base percentages of all players with at least 300 at-bats in the 2014 season.

Now instead of one player, we consider all 238 players who were "regulars" (had at least 300 at-bats) in the 2014 MLB season. Figure 8.4 displays a stemplot of the on-base percentages.

We see that the on-base percentages are bell-shaped about the mean value of .326. The highest batting average was Troy Tulowitzki at .432 and the lowest was John Schierholtz at .240.

Can we construct a model for the true on-base percentages (the abilities) for these 238 hitters?

The simplest model I can think of is what we call the "One Spinner" model. Maybe all of the 238 players have the same batting ability. Each player is using the same spinner with on-base probability $p = .326$ (the average of all players). If you believe this, then the variation in the season on-base percentages that we observe in the stemplot above are simply due to chance variation. Players all have the same ability, but some like Tulowitzki are lucky and have high season averages; others, like Schierholtz, were unlucky and had low averages.

Now, if you know anything about baseball, you should be thinking that this is a crazy model—players do have different batting abilities. (In fact, we illustrated this fact in the last case study where we looked at batting averages of some players for two consecutive seasons.) I agree, but I want to demonstrate how a statistician checks the suitability of a probability model.

Simulating Hitting Data Using a "One Spinner" Model

To see if the "One Spinner" model is reasonable, we simulate hitting data from the model.

We imagine 238 spinners, each with on-base probability $p = .326$. We spin each spinner for a whole season using the same numbers of at-bats as the 2014 players and obtain 238 season on-base percentages. We did this one time. In the back to back stemplot display in

Figure 8.5, we have placed the simulated batting averages on the left and the actual 2014 batting averages on the right.

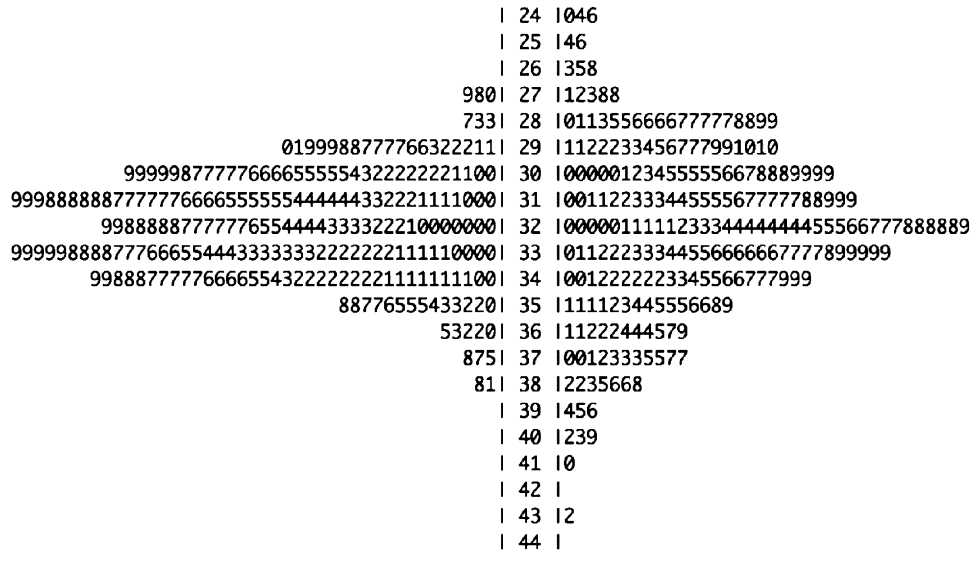


Figure 8.5. Back-to-back stemplots of simulated on-base percentages from one-spinner model (left) and 2014 MLB on-base percentages (right).

How does the simulated data compare with the actual data? There is a substantial difference: the simulated batting averages appear to have less variation or spread than the real averages.

We can measure spread of a dataset by the standard deviation s. For the dataset of real batting averages, we compute $s = .0329$ and for the simulated data, $s = .0217$. This confirms what we just said the simulated averages aren't as spread out as the real averages.

To see if this always happens, we did the simulation from the "One Spinner" model many times. In four additional simulation runs, we obtained

$$s = .0212, \quad s = .0212, \quad s = .0201, \quad s = .0208.$$

Each time, the standard deviation is smaller than the standard deviation (.0329) for the real data. So this model does not seem to generate data that is similar to the 2014 dataset of on-base percentages.

We conclude the "One Spinner" model isn't appropriate and consequently that batters have different abilities.

A Many Spinners Model

What is an alternative probability model that can represent hitting data in baseball? Here is a "Many Spinners" model that seems to better represent what is really going on.

- We first represent the hitting abilities of the 238 players by a normal curve shown in Figure 8.6 with an average of .326 and standard deviation of .025. In other words, the on-base probabilities of the players are variable according to a normal curve with mean .326.

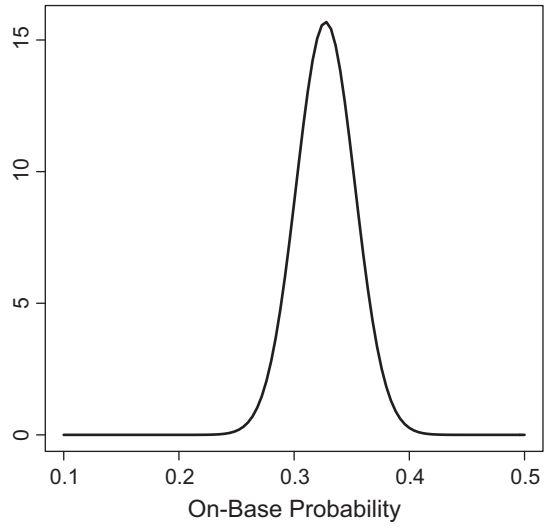


Figure 8.6. Bell-shaped curve to represent the on-base probabilities of the MLB players.

- We sample at random 238 on-base probabilities from this normal curve. Then we represent the abilities of the players using a set of spinners, where the "On-Base" areas are different for different players. One player might have a on-base area of $p = .310$, another may have a on-base area of $p = .330$, and so on.

We tried simulating data using this "Many Spinners" model. We first simulated a set of random hitting probabilities and then simulated hitting data using these probabilities. Figure 8.7

```
                                4| 24 |046
                             8760| 25 |46
                             7765| 26 |358
                         99996650| 27 |12388
                   987766443333210| 28 |0113556666777778899
          0101998888765554333321| 29 |1122233456777991010
            99999866554433332210| 30 |0000012345555566788899999
    99998877776665544443332222000| 31 |001122333445555677777888999
988877666655544333332222211100| 32 |0000011111233344444444555667778888889
  999988887766554443333222211111| 33 |011222333445566666677777899999
      998887655555533222222210| 34 |00122222223345566777999
                       99866411| 35 |111123445556689
                98866655543210000| 36 |11222444579
                     9776433321| 37 |00123335577
                       9977621| 38 |2235668
                          7620| 39 |456
                           000| 40 |239
                              | 41 |0
                            3| 42 |
                              | 43 |2
                              | 44 |
```

Figure 8.7. Back-to-back stemplots of simulated batting averages from many-spinners model (left) and 2014 MLB on-base percentages (right).

shows back-to-back stemplots of the simulated season on-base percentages and the actual 2014 on-base percentages.

Comparing the two stemplots, the distribution of the simulated data from the "Many Spinners" data does resemble the distribution of the actual 2014 on-base percentages. In particular, the standard deviations match up—the standard deviation of the simulated data is .0320, which is close to the standard deviation of the observed data .0329. So the "Many Spinners" model seems to be a good representation for the hitting abilities of many players. We will use this probability model in the following case study when we model situational hitting data.

8.4 Models for Situational Effects

Topics Covered: Probability models for situational hitting data, bias model, model with ability effects.

What are good models for situational data? We describe three basic models here that seem to describe the pattern in hitting data for all of the situations that are displayed in the `fangraphs.com` site.

Recall our basic probability model for batting data for a group of players described in Case Study 8.3. Players have different abilities (values of p) selected from a normal curve (with a mean of .326 and a standard deviation of .025), and batting data is found by spinning a bunch of spinners, where each spinner has a different hitting probability.

The "No Effect" Model

To describe the simplest situational hitting model, suppose we have five spinners with on-base probabilities of $p = .2, .3, .4, .5, .6$. These spinners represent the abilities for five baseball players. Suppose that we spin the spinners in the dark and the light.

Now, it is reasonable to think that the lightness of the room has no effect on the probability of getting a hit on a particular spinner. So if the first spinner has a hitting probability of $p = .2$ in the light, then this spinner will also have a hitting probability of $p = .2$ in the dark. Likewise, the .3 spinner will have a hitting probability of .3 both in the light and the dark, the .4 spinner will have the same hitting probability in the dark and in the light, etc. In this case there is no true situational effect so we call this the "no effect" model.

Now suppose we spin the $p = .2$ spinner 100 times in the dark and 100 times in the light. It is certainly possible that we'll get 22 on-base events in the dark and only 18 on-base events in the light for an observed situational effect of

$$\text{Observed Situational Effect} = 22/100 - 18/100 = .04$$

But this effect is just due to chance there is no true situational effect. An example of this "no effect" situation is before and after the All-Star Game. Generally, players appear to have the same batting abilities before and after the All-Star Game. Certainly, we will observe some players one season who bat better before the All-Star Game and we'll see other players who are "hot" after the All-Star Game. But most of the variation in the differences in batting averages before and after the All-Star Game is due to chance variation.

The "Bias" Model

The bias model is a different way of representing the hitting abilities of players in two situations. Here the situation changes the hitting probability by the same amount for all hitters.

Let's consider our dark and light example again. Let's suppose that it is hard to see the spinner in the dark, and the light increases the on-base probability by .05 for all hitters. So if one hitter has a dark on-base probability of $p = .3$, it will be $p = .3 + .05 = .35$ in the light. Another hitter with a dark on-base probability of $p = .4$ will have an on-base probability of $p = .4 + .05 = .45$ in the light.

The bias model seems appropriate for the home vs away hitting data. A particular ballpark has a positive or negative impact on a player's batting average. Coors Field is an obvious example of a ballpark that helps a player's batting ability, and Dodger Stadium is an example that hurts a player's batting ability. But the effect of the ballpark is to add a constant number to each player's batting probability. So Coors Field may add a positive number, say .10 to each player's true batting average. There is a situational effect here, but this effect is the same for all players.

A Simulation

I used a statistics computing package to illustrate what situational data looks like if there is a bias effect. I assumed

- there are 100 players,
- each player has 300 at-bats at HOME and 300 at-bats AWAY,
- there is a bias situational effect of twenty points due to playing at home so each player's HOME on-base probability is .020 higher than his AWAY on-base probability.

I had the program first simulate 100 "Away" true on-base percentages from a normal curve with mean .326 and standard deviation .025—these hitting probabilities are put in the "p.AWAY" column. We compute "Home" true batting averages in the "p.HOME" column by adding .020 to each probability in the "p.AWAY" column. We then simulated hits for 300 at-bats at the home games using the Home hitting probabilities—these numbers are put in the "h.HOME" column. Similarly, we simulated hits for 300 away at-bats using the Away hitting probability—these hit numbers are in the "h.AWAY" column. We compute on-base percentages at home and away (these OPB's are in the "OBP.HOME" and "OBP.AWAY" columns), and compute observed differences

$$\text{OBP(home)} - \text{OBP (away)}.$$

A part of the output of this simulation is shown in Table 8.2.

Table 8.2. Some situational hitting data assuming that the home ballpark adds .020 to the probability of getting on base for all players

	p.AWAY	p.HOME	OBP.AWAY	OBP.HOME	DIFF
1	.32	.34	.30	.30	.00
2	.36	.38	.37	.43	.06
3	.30	.32	.32	.34	.02
4	.35	.37	.37	.35	−.02
5	.35	.37	.36	.37	.01
6	.37	.39	.37	.37	.00

Figure 8.8 displays a dotplot of the observed situational effects for the 100 players contained in the DIFF column. The median value of DIFF is equal to .020 and the extreme values are −.083 and .093.

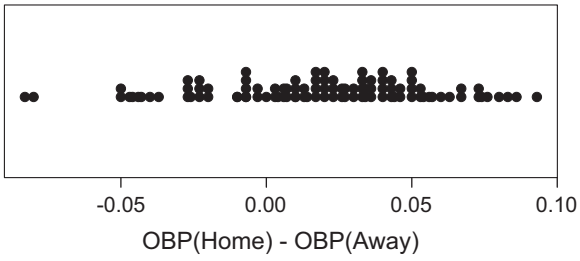


Figure 8.8. Dotplot of observed situational effects from simulated data where there is a bias situational effect.

Here we see that when there this is a bias situational effect, the observed effects can look interesting. One player in the simulation hit for 93 points better at home games and another hit for 83 points higher at away. This high variation is misunderstood by baseball fans. People think that some players (like the one that hit 93 points higher during home games) have a special ability to play better at home, when really the home ballpark has the same positive effect on all players.

The "Ability" model

The last model, the so-called "ability" model, is the most complicated description of situational data. This probability model says that there are real situational effects, and they are different among players. I'll illustrate this scenario using the following "Joe Cool" and "Harry Hyper" example—two players who react differently to a home crowd.

Joe Cool plays the same way everywhere he plays. So if Joe hits for a on-base probability of .350 at home, he'll also hit for a on-base probability of .350 on the road. Harry, in contrast, is very emotional (like Pete Rose) and feeds off of the enthusiasm of the home crowd. If Harry hits for a on-base probability of .325 on away games, he might hit 50 points higher, or .375, at home.

There are relatively few situations (among the ones that are discussed) that are ability effects. The one exception is the batting performance of the hitter under different pitch counts. Suppose one looks at two batting averages:

- the batting average when there are two strikes on the batter,
- the batting average when the batter is ahead in the count (that is, the pitch count is 3-0, 3-1, 2-0, 2-1, or 1-0).

If one looks at the situational hitting data for all players in this situation, there is evidence that pitch count is an ability effect.

This means that players have different abilities under various pitch counts. A big slugger (Ryan Howard comes to mind) is an ineffective hitter when he is behind in the count. When the count is 0 balls and 2 strikes, he's likely to strike out. In contrast, the good contact hitters (like

Ichiro Suzuki) are effective hitters even when they are behind in the count. Suppose we define the pitch count situational effect to be

$$\text{DIFF} = \text{OBP(ahead in the count)} - \text{OBP(behind in the count)}.$$

There is evidence to suggest that big sluggers (like Howard) have large values of this situational effect, and other hitters (the contact-hitter like Suzuki) have relatively small pitch count effects. How does one find a group ability effect in situational data? Recall our discussion about looking for abilities in hitting data in Case Study 8.2. Suppose you look at a group of players and compute their situational effect (say, on-base percentage ahead in the count minus on-base percentage behind in the count) for two consecutive years, say 2013 and 2014. Construct a scatterplot of the two years of situational effects. You have found a group ability situational effect if you find a positive association in the scatterplot between the 2013 effect and the 2014 effect. This means that players with high situational effects in 2013 will tend to have high situational effects in 2014. Also players who have low situational effects one year will tend to have low effects the second year. Remember that most situational effects in baseball hitting are "no effects" or "biases"—the pitch count is one of the few situations where players generally appear to have different abilities to take advantage of the situation.

8.5 Is Michael Brantley Streaky?

Topics Covered: Distinction between streaky performance and streaky ability, moving averages, runs.

In this case study we talk about streakiness or the hot hand. Do players really get hot and cold? Here I'm not talking about the hot and cold streaks that we observe in baseball data. I'm instead talking about a player's ability to "get into a groove" in hitting or pitching. Maybe during a certain week, the player's batting stroke feels just right or he sees the ball particularly well. Another week, he feels different maybe his batting stroke is out of sync or he has an injury that makes hitting more difficult.

Is there evidence for a player to be truly streaky, that is, have an ability to be hot and cold?

We focus on the Indians' player Michael Brantley. To look for streakiness in Brantley's hitting data, we focus on his game-to-game hitting data for the 2014 season.

Moving Average

One way of looking for streaky behavior is to compute a moving average—this is a short-term batting average using a window of a particular number of games.

Suppose that we wish to compute moving averages using a window of five games.

- The first moving average is the batting average for games 1 through 5.
- The second moving average is the batting average of games 2 through 6.
- The third moving average is the batting average of games 3 through 7.

Essentially we're just computing averages over short time intervals.

In Table 8.3, I compute moving averages for the 2014 Brantley data using a window of 9 games.

Table 8.3. Computation of moving averages for day-by-day batting data for Michael Brantley

Game	AB	H	9-Game AB	9-Game H	Moving AVG
1	4	2			
2	3	1			
3	4	1			
4	4	1			
5	4	1	34	10	.294
6	5	3	34	9	.265
7	4	0	35	9	.257
8	3	1	34	9	.265
9	3	0	33	9	.273
10	4	1	32	8	.250
11	4	1	32	7	.219
12	3	1	33	8	.242
13	3	1	33	7	.212
14	3	0	33	9	.273
15	5	2	33	9	.273

1. In games 1–9, Brantley had 34 AB and 10 H for an average of $10/34 = .294$. This moving average of .294 is put in the table across the average game number 5 (5 is the average of games $1, 2, \ldots, 9$).
2. In the table, we see a moving average of .250 across game 10. We look at the nine games centered about game 10 these are games 6 through 14. In this period, Brantley had 8 hits in 32 at-bats for a batting average of $8/32 = .25$.

If we graph the game numbers against the corresponding moving averages, the graph shown in Figure 8.9 is obtained.

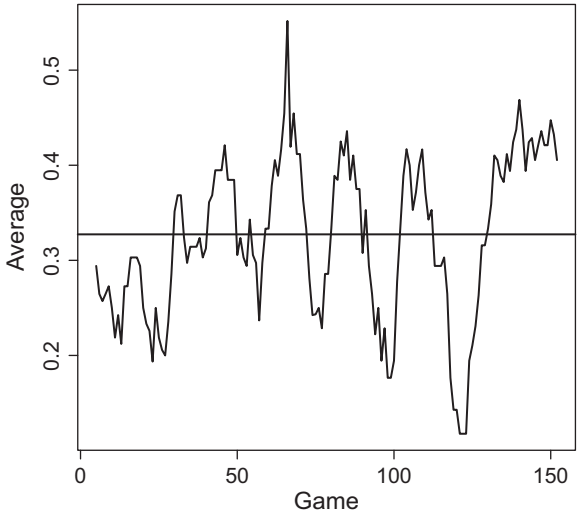


Figure 8.9. Moving average plot of Mike Brantley's batting average using a window of nine games.

This graph is a picture of the hot and cold periods of Brantley's hitting for the 2014 season. We see

- a hot period about game 70—during one nine game period, Brantley batted over .500,
- this hot period was surrounded by two cold periods where he batted close to .250,
- Brantley finished the season with a very cold period where his average was under .100 followed by a hot period where he batted over .4. So one indication of streakiness is the extreme hills and valleys that we see in the moving average plot.

Runs

Another way of measuring streakiness is based on runs of good and bad days. (Not to be confused with runs scored in a baseball game.)

Suppose we classify each day of Brantley's hitting as being either "Hot" or "Cold"—we say he is Hot if his day batting average is over .327 (his season avg); otherwise he is classified as Cold. Table 8.4 shows Brantley's batting results for the first 13 games and the hot and cold classification.

Table 8.4. Batting results for Michael Brantley for the first 13 games of the 2014 season and classification of the result in hot and cold states

Game	AB	H	Hot or Cold?
1	4	2	Hot
2	3	1	Hot
3	4	1	Cold
4	4	1	Cold
5	4	1	Cold
6	5	3	Hot
7	4	0	Cold
8	3	1	Hot
9	3	0	Cold
10	4	1	Cold
11	4	1	Cold
12	3	1	Hot
13	3	1	Hot

In Game 1, Brantley was $2/4 = .500$; this is larger than .327, so we call it a Hot day. In Game 2, he was $1/4 = .250$ which is under .300, so he is Cold that day.

Now we look for **Runs** in this sequence. A run is simply a consecutive sequence of Hot's or a sequence of Cold's. (This is different from the usual meaning of run in baseball.) Specifically, we count

1. the total number of runs ,
2. the length of the longest run (either Hot or Cold).

In the above sequence of 13 games, we see that

- total number of runs = 7,
- longest run is (Cold, Cold, Cold), so length of longest run = 3.

If the player is truly streaky, we expect to see

- a small number of runs,
- long runs of Hot's or Cold's, so the length of the longest run would be large.

This makes sense. If a player is streaky, then he will have long runs of hot games and long runs of cold games, so the number of runs in the sequence will be small.

Here we have discussed ways of detecting the observed streakiness that we see in baseball data. However, that does not mean that the player (or team) is truly streaky. In the next case study, we'll clearly make a distinction between a player's streaky ability (or lack of streaky ability) and his streaky performance during a baseball season. We will see that a fair die, an object with consistent ability, can exhibit very streaky performance.

8.6 A Streaky Die

Topics Covered: Moving averages, runs, simulation of dice tossing, patterns in dice tossing data.

In the previous case study, we looked for streaky behavior in the day-to-day hitting data for Michael Brantley's 2014 baseball season. Specifically, we looked at

- moving averages using a window of nine games we were looking for unusually high or low moving averages to say that Brantley is streaky,
- runs of good and bad hitting days here we looked at the total number of runs, and the length of the longest run—a small total number of runs and/or a long run (of hot or cold) games would indicate streakiness.

Using moving averages and runs, we saw some interesting features in Brantley hitting data:

- In one 9-game stretch, Brantley hit .552; in another 9-game stretch, he only hit .118.
- Brantley had a total of 78 runs (either runs of Hot or runs of Cold). His longest run had length 9. But are these really interesting? Do they mean that Brantley is truly a streaky hitter? In other words, can we tell from these data that Brantley has an ability to be streaky? Maybe he isn't really streaky and by luck or chance variation, we just happened to see some interesting streaky behavior.

Mr. Consistent

Let's consider this last question more carefully. Suppose that Brantley is truly not streaky. What would that mean?

The opposite of a streaky hitter is a consistent hitter. This type of hitter always gets a hit with the same probability no matter what the time or situation. This consistent hitter will get a hit with the same probability p in every at-bat in the season.

We can represent hitting of a consistent hitter by use of a die. Consider a 10-sided die with the sides 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Suppose we roll the die and a "hit" is recorded if the die rolls 1, 2, 3; otherwise, the hitter is out. Then the probability of a hit will be $3/10 = .3$. More

importantly, the chance of a hit is always .3 and the outcome of the die won't depend on what rolls occurred in the past.

We can use this die to simulate Brantley's hit outcomes if he were really a consistent hitter. We will simulate hitting data for all 162 games by working in groups—group 1 will simulate Brantley's hitting for games 1–18, group 2 will simulate Brantley's hitting for games 19–36, etc.

When we complete this activity we will see if this simulated data looks streaky. In particular, we will look at

- moving averages using a window width of 9 games,
- the number of runs and the length of the longest run.

What I think we'll discover is that this simulated data can look pretty streaky. Even when a hitter is truly consistent, there can be interesting patterns in the moving average graph that can be interpreted as streaky behavior.

Figure 8.10 shows the moving average graph (using a width of nine games) for the data that a statistics class generated using a 10-sided die.

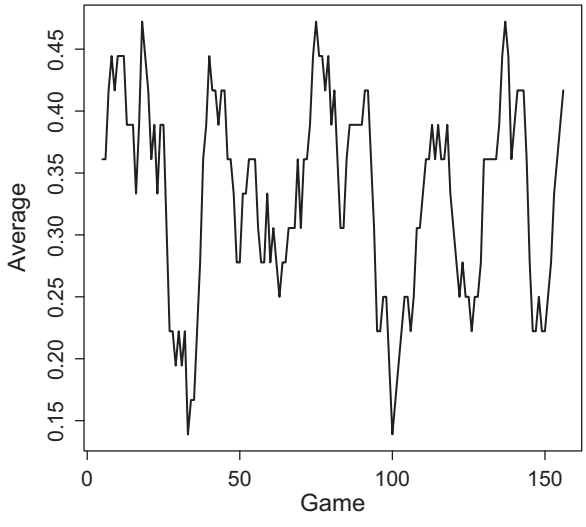


Figure 8.10. Moving average plot of simulated hitting data for a truly consistent hitter where the probability of a hit is equal to .3.

We see a lot of interesting patterns. Our hitter had an early slump around game 20 and had one significant hot streak where he batted in the .500 range. But remember this hitter is truly a consistent batter what we are observing is the streakiness that is inherent in chance variation.

To show you that this is not a fluke occurrence, I did our simulation five times on a computer. I'm using the same assumptions as above.

1. Our player is a truly consistent hitter with probability $p = .3$ of getting a hit on a single at-bat.
2. This player has exactly 4 at-bats in each game.

Figure 8.11 shows the moving average plots for my six simulations (I'm still using a window width of nine games).

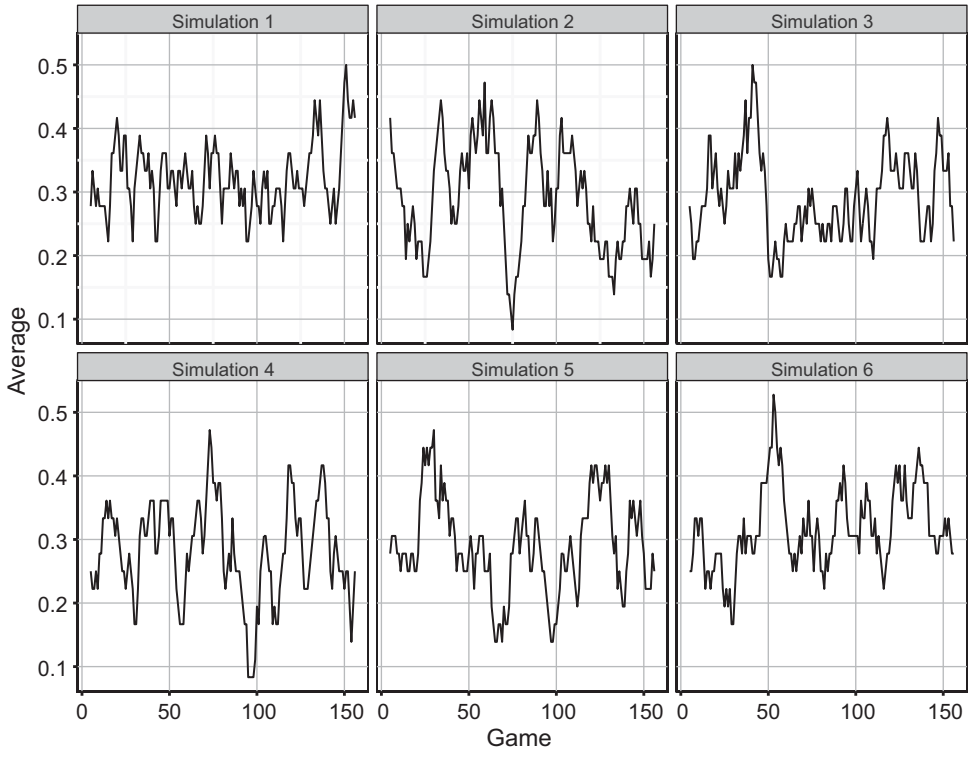


Figure 8.11. Moving average plots for six simulated datasets for a consistent hitter with a constant probability of .3 of getting a hit.

You should notice a lot of up-and-down behavior in the five moving average plots. So even if Brantley were a truly consistent hitter, he would likely have several interesting hot and cold streaks.

The moral of the story is that you should be cautious about interpreting streaky behavior of baseball players. Dice have a clearly consistent ability—just the opposite from true streakiness. But we've seen that dice can look streaky!

8.7 Exercises

8.0. Table 8.5 gives some situational OBPs for Rickey Henderson in the 1999 baseball season. This table shows how Henderson did for day and night games, for games home and away, for games played on grass and turf fields, for games played in domed and open ballparks, and against left- and right-handed pitchers.

(a) For each breakdown, compute the number of (approximate) plate appearances (PA) by adding at-bats (AB) to walks (BB).

(b) Let p_{day} denote Henderson's on-base probability when he is playing a day game and let p_{night} denote his on-base probability when he's playing at night. Using the data in

Table 8.5. Situational on-base percentages of Rickey Henderson for the 1990 season

Breakdown	AB	BB	PA	OBP
Day	144	21		.400
Night	294	61		.433
Home	200	34		.376
Away	238	48		.462
Grass	364	60		.408
Turf	74	22		.490
Dome	43	10		.472
Open	395	72		.417
vs. Right	333	54		.403
vs. Left	105	28		.481

the table, construct 90% probability intervals for p_{day} and p_{night}. Comparing the two intervals, can you conclude that Henderson really has a higher on-base probability at night games?

(c) Using the same method as in (b), compare the home/away OBPs, the grass/turf OBPs, the dome/open OBPs, and the right/left OBPs.

(d) Based on your work in (c) and (d), rank the five situations with respect to the most significant effect to the least significant effect.

8.1. Table 8.6 gives batting statistics for 14 randomly selected players in the 2013 season. The first three columns give the number of at-bats (AB), hits (H), and batting average (AVG)

Table 8.6. Home and away batting averages for 14 randomly selected players from the 2013 season

	Home			Away			
Name	AB	H	AVG	AB	H	AVG	Home—Away
David Ortiz	272	83	.305	246	77	.313	
Daniel Murphy	320	84	.262	338	104	.308	
Evan Longoria	291	75	.258	323	90	.279	
Khris Davis	72	22	.306	64	16	.250	
Christian Yelich	121	31	.256	119	38	.319	
Marlon Byrd	261	67	.257	271	88	.325	
Xander Bogaerts	22	5	.227	22	6	.273	
Marcell Ozuna	127	32	.252	148	41	.277	
Todd Frazier	264	70	.265	267	54	.202	
Matt Carpenter	311	112	.360	315	87	.276	
Andrelton Simmons	286	72	.252	320	78	.244	
Zack Cozart	289	68	.235	278	76	.273	
Martin Prado	292	85	.291	317	87	.274	
Jacoby Ellsbury	284	84	.296	293	88	.300	

for all games played at Home, and the next three columns give the same statistics for all games played Away from home.

(a) For each player, compute the difference in batting averages

$$\text{Difference} = \text{AVG (Home)} - \text{AVG (Away)}$$

and put the differences in the "Home − Away" column of Table 8.6.

(b) Draw a stemplot of the batting average differences that you computed in part (a).

(c) Looking at the stemplot constructed in part (b), what is the median value of the difference? Also find the smallest and largest differences, and find the players that had these extreme values.

(d) From your work, can you say that players generally hit better at home? If so, by how much on the average?

8.2. Table 8.7 gives 2014 batting statistics for 15 players before the All-Star break (months April, May, and June) and after the All-Star break (months July, August, and September).

(a) Repeat parts (a) – (c) of Exercise 8.1 for this dataset. Here the PRE-POST column of Table 8.7 will contain the differences in batting average

$$\text{AVG(Pre All-Star Game)} - \text{AVG(Post All-Star Game)}.$$

(b) From your work, can you say that players generally play better after or before the All-Star break? If so, how much on the average?

Table 8.7. Batting statistics for 15 players before and after the 2014 All-Star break

	Pre AS Game			Post AS Game			
Name	AB	H	AVG	AB	H	AVG	Pre—Post
David Ortiz	296	74	.250	222	62	.279	
Daniel Murphy	340	103	.303	256	69	.270	
Jose Abreu	272	76	.279	284	100	.352	
Evan Longoria	332	87	.262	292	71	.243	
Khris Davis	290	75	.259	211	47	.223	
Christian Yelich	255	66	.259	327	99	.303	
Marlon Byrd	310	83	.268	281	73	.260	
Xander Bogaerts	290	72	.248	248	57	.230	
Marcell Ozuna	291	77	.265	274	75	.274	
Todd Frazier	310	89	.287	287	74	.258	
Matt Carpenter	317	89	.281	278	73	.263	
Andrelton Simmons	291	72	.247	249	60	.241	
Zack Cozart	276	63	.228	230	49	.213	
Martin Prado	314	84	.268	222	67	.302	
Jacoby Ellsbury	302	87	.288	273	69	.253	

8.3. Table 8.8 displays 2014 batting statistics for 15 players when the batter had an initial ball (a 1-0 count) and when the batter had an initial strike (a 0-1 count).

Table 8.8. Batting statistics for 15 players with an initial ball (1-0 count) and an initial strike (0-1 count)

	Initial Ball			Initial Strike			
Name	AB	H	AVG	AB	H	AVG	Initial B—Initial S
David Ortiz	255	64	.251	207	53	.256	
Daniel Murphy	304	76	.250	206	67	.325	
Jose Abreu	263	78	.297	208	67	.322	
Evan Longoria	324	70	.216	205	56	.273	
Khris Davis	244	49	.201	199	53	.266	
Christian Yelich	306	64	.209	223	79	.354	
Marlon Byrd	315	78	.248	201	48	.239	
Xander Bogaerts	301	65	.216	194	50	.258	
Marcell Ozuna	298	74	.248	195	57	.292	
Todd Frazier	293	67	.229	226	70	.310	
Matt Carpenter	341	93	.273	226	60	.265	
Andrelton Simmons	244	63	.258	198	46	.232	
Zack Cozart	261	58	.222	184	39	.212	
Martin Prado	307	89	.290	199	52	.261	
Jacoby Ellsbury	264	64	.242	226	67	.296	

(a) Repeat parts (a)–(c) of Exercise 8.1 for this dataset. Here the "Initial B − Initial S" column of Table 8.8 will contain the differences in batting average

$$\text{AVG(Initial Ball)} - \text{AVG(Initial Strike)}.$$

(b) From your work, can you say that players generally hit better when they have an initial ball as opposed to an initial strike? If so, how much on the average?

8.4. Suppose that one is interested in how players hit on odd-numbered days (like April 7, May 13, June 9) as opposed to even-numbered days (like April 10, May 14, July 20). Here we are pretty sure that there is no true situational effect for any player. (Why would any player actually be a better or worse hitter on odd-numbered days?) We simulate the type of hitting data that one might see in this scenario by means of the following experiment. For each of 15 players, we first simulate their abilities (their hitting probabilities) using a normal curve with mean .276 and standard deviation .021. Here are the twenty simulated abilities:

.262 .282 .291 .293 .321 .227 .279 .248
.288 .309 .264 .238 .295 .309 .287

Then we simulate the situational data as follows. For each player, we simulate the results of 300 at-bats for the odd-numbered days using the above hitting probabilities, and then

simulate the results of 300 at-bats for the even-numbered days using the same set of hitting probabilities. We obtain the data in Table 8.9.

Table 8.9. Simulated batting statistics for 15 players on even-numbered and odd-numbered days

	Even # Days			Odd # Days			
Name	AB	H	AVG	AB	H	AVG	Even—Odd
Jim	300	82	.273	300	78	.260	
Pat	300	80	.267	300	87	.290	
Edsel	300	87	.290	300	94	.313	
Arjun	300	84	.280	300	72	.240	
Rich	300	97	.323	300	94	.313	
Pete	300	67	.223	300	59	.197	
Sam	300	89	.297	300	80	.267	
Dave	300	91	.303	300	61	.203	
Curt	300	87	.290	300	84	.280	
Ben	300	85	.283	300	89	.297	
Adam	300	80	.267	300	93	.310	
Dale	300	67	.223	300	72	.240	
Joe	300	72	.240	300	80	.267	
Dick	300	75	.250	300	92	.307	
Brad	300	90	.300	300	77	.257	

(a) Compute all of the batting average differences

$$\text{Difference} = \text{AVG(Even-Numbered Days)} - \text{AVG(Odd-Numbered Days)}$$

and put the differences in the "Even − Odd" column of Table 8.9.

(b) Graph the batting average differences using a stemplot.

(c) Find the average difference, the low and high differences, and find the batters who had these extreme values.

(d) Do these simulated data resemble any of the situational data described in the chapter? Explain.

8.5. Suppose that it is found out that some ballparks are using a special baseball that travels further when hit. Moreover, it is known that this special baseball will add 30 points to every player's hitting probability. Table 8.10 generates some simulated hitting data using this scenario.

To simulate these data, we first simulate hitting probabilities for the 15 players using a normal curve with mean .261 and standard deviation .021. These probabilities correspond to the abilities of the hitters playing with the Usual ball.

.290 .279 .267 .252 .249 .264 .244
.276 .287 .244 .294 .295 .279 .242 .288

To obtain the hitting probabilities for the players with the Special ball, we add 30 points to each of the Usual hitting probabilities.

.320 .309 .297 .282 .279 .294 .274
.306 .317 .274 .324 .325 .309 .272 .318

Then we simulate hitting data for 300 at-bats using the Usual probabilities and for 300 at-bats using the Special probabilities—we show the data in Table 8.10.

Table 8.10. Simulated hitting data for 15 players using a special baseball and the usual baseball when there is a true situational bias

	Usual Baseball			Special Baseball			
Name	AB	H	AVG	AB	H	AVG	Special—Usual
Larry	300	89	.297	300	81	.270	
Moe	300	90	.300	300	98	.327	
Curly	300	85	.283	300	79	.263	
Harold	300	95	.317	300	74	.247	
Harvey	300	72	.240	300	85	.283	
Lee	300	74	.247	300	97	.323	
Charles	300	72	.240	300	81	.270	
Bill	300	77	.257	300	105	.350	
Bob	300	83	.277	300	101	.337	
Rick	300	81	.270	300	78	.260	
Britt	300	89	.297	300	94	.313	
Clark	300	86	.287	300	99	.330	
Randy	300	85	.283	300	91	.303	
Tom	300	86	.287	300	75	.250	
Gene	300	87	.290	300	102	.340	

(a) Compute all of the batting average differences

$$\text{Difference} = \text{AVG(Special Ball)} - \text{AVG(Usual Ball)}$$

and put the differences in the "Special − Usual" column of Table 8.10.

(b) Graph the batting average differences using a stemplot.

(c) Find the average difference, the low and high differences, and find the batters who had these extreme values.

(d) Do these simulated data resemble any of the situational data described in the chapter? Explain.

8.6. (Exercise 8.5 continued.) Suppose that players have different abilities to use the "Special Ball" that was discussed in Exercise 8.5. For the first five players listed in Table 8.11 (the first group), the special ball adds 50 points to the "Usual Ball" hitting probability. For the next five players in the table (the second group), the special ball adds 30 points to the player's hitting probability, and for the final five players (the third group), the special ball

only adds ten points to the hitting probability. Below we give the hitting probabilities for the 15 players. Table 8.11 shows hitting data simulated from this model.

Players	Hitting Probabilities for Usual Ball		Hitting Probabilities for Special Ball
First Group	.282, .248, .291 .310, .261	Add 50 Points	.332, .298, .341 .360, .311
Second Group	.243, .246, .288 .303, .290	Add 30 Points	.273, .276, .318 .333, .320
Third Group	.251, .276, .273 .268, .290	Add 10 Points	.261, .286, .283 .278, .300

(a) Compute the differences in batting average (Special Ball − Usual Ball) and put the differences in the last column of Table 8.11.
(b) Construct a stemplot of the differences.
(c) Compute the average batting average difference, and find the smallest and largest differences.
(d) Compare the batting average differences with the differences found in Exercise 8.5. Can you offer any explanation for the different patterns that you see in stemplots from Exercise 8.5 and this exercise?

Table 8.11. Simulated hitting data for 15 players using a special baseball and the usual baseball when there are ability situational effects

		Usual Baseball			Special Baseball			
		AB	H	AVG	AB	H	AVG	Special—Usual
First group	Larry	300	90	.300	300	91	.303	
	Moe	300	77	.257	300	91	.303	
	Curly	300	99	.330	300	95	.317	
	Harold	300	88	.293	300	95	.317	
	Harvey	300	82	.273	300	106	.353	
Second group	Gene	300	67	.223	300	73	.243	
	Bill	300	86	.287	300	82	.273	
	Bob	300	93	.310	300	105	.350	
	Rick	300	92	.307	300	115	.383	
	Lynn	300	74	.247	300	95	.317	
Third group	Jean	300	86	.287	300	82	.273	
	Joe	300	93	.310	300	73	.243	
	Bret	300	84	.280	300	79	.263	
	Britt	300	79	.263	300	73	.243	
	Clark	300	86	.287	300	81	.270	

8.7. In Exercise 8.1, we looked at home and away batting averages for 14 players during the 2013 season. Table 8.12 gives home and away averages for the same 14 players, but for the following (2014) season.

Table 8.12. Home and away batting averages for 14 randomly selected players from the 2014 season

	Home			Away			
Name	AB	H	AVG	AB	H	AVG	Home—Away
David Ortiz	250	67	.268	268	69	.257	
Daniel Murphy	276	69	.250	320	103	.322	
Evan Longoria	303	79	.261	321	79	.246	
Khris Davis	243	57	.235	258	65	.252	
Christian Yelich	273	81	.297	309	84	.272	
Marlon Byrd	299	78	.261	292	78	.267	
Xander Bogaerts	255	66	.259	283	63	.223	
Marcell Ozuna	278	80	.288	287	72	.251	
Todd Frazier	299	85	.284	298	78	.262	
Matt Carpenter	287	68	.237	308	94	.305	
Andrelton Simmons	261	69	.264	279	63	.226	
Zack Cozart	255	60	.235	251	52	.207	
Martin Prado	261	73	.280	275	78	.284	
Jacoby Ellsbury	264	80	.303	311	76	.244	

(a) Compute the differences in batting averages AVG(home) − AVG(away) and place your work in Table 8.12.

(b) Construct a stemplot of the batting average differences. Based on this graph, would you say that there is a general tendency for players to bat better at home? Why or why not?

(c) Suppose that some batters hit unusually well at home, and other batters actually hit better on the road. If players did have different abilities to bat at home games versus away games, then one would expect to see a relationship between the batting average difference (Home − Away) for 2013 and the batting average difference for 2013. Using the data from Table 8.6 and Table 8.12, construct a scatterplot of the 2013 and 2014 differences.

(d) Based on your scatterplot drawn in (c), do you see a relationship between a player's batting average difference for 2013 and 2014? Interpret what this means in terms of one's ability to hit well at home versus away games.

8.8. Situational data are also recorded for pitchers. Table 8.13 displays pitching statistics for 20 pitchers for the three-year period 1997–1999. For each pitcher, the table gives

- Throws: the throwing hand of the pitcher,
- Batting Average—Left: the batting average of left-handed hitters against the pitcher,
- Batting Average—Right: the batting average of right-handed hitters against the pitcher,
- Batting Average—Pitch 1–15: the batting average of the hitters on pitches 1–15 during a game,
- Batting Average—Pitch 46–60: the batting average of the hitters on pitches 46–60 during a game,
- Batting Average—Pitch 91–105: the batting average of the hitters on pitches 91–105 during a game.

(a) If a pitcher is right-handed, then who will be a more effective hitter a left-hander or a right-hander? Why?
(b) If a pitcher is left-handed, then who will be a more effective hitter a left-hander or a right-hander?
(c) Define an opposite-side hitter as one who bats at the opposite side from the arm of the pitcher. (Likewise, define a same-side hitter as a hitter who bats from the same side as the arm of the pitcher.) For each pitcher, compute the difference

$$\text{DIFF1} = \text{AVG(opposite-side hitter)} - \text{AVG(same-side hitter)}.$$

Put your batting average differences in the "DIFF1" column of Table 8.13.
(d) Graph the differences and find the average value. Conclude what you have learned about the effectiveness of opposite-side hitters compared to same-side hitters.

8.9. (Exercise 8.8 continued.) Table 8.13 also shows the batting average of hitters during different pitch counts of each pitcher.
(a) How do you expect a starting pitcher to perform at the beginning of a game? Do you think pitchers need to warm-up and don't reach full effectiveness until they have pitched a few innings?

Table 8.13. Situational statistics for 20 pitchers for the three-year period 1997–1999

		Batting Average		DIFF1	Batting Avg. Pitch			
Pitcher	Throws	Left	Right	Opp—Same	1–15	46–60	91–105	DIFF2
John Smoltz	Right	.252	.227		.293	.216	.260	
Alan Ashby	Right	.271	.250		.266	.279	.284	
Alex Fernandez	Right	.284	.203		.283	.212	.208	
Andy Pettite	Right	.295	.266		.255	.287	.302	
Bartolo Colon	Right	.267	.248		.248	.282	.219	
Chuck Finley	Left	.236	.249		.226	.294	.254	
Charles Nagy	Right	.291	.291		.294	.298	.272	
David Cone	Right	.237	.219		.212	.260	.273	
Doug Drabek	Left	.282	.292		.260	.261	.254	
Darryl Kile	Right	.284	.251		.299	.296	.267	
Greg Maddux	Right	.244	.254		.284	.245	.216	
Kevin Brown	Right	.249	.216		.259	.264	.221	
Randy Johnson	Left	.178	.213		.216	.221	.258	
Curt Schilling	Right	.244	.220		.250	.210	.232	
Tom Glavine	Left	.250	.251		.283	.256	.224	
Terry Mulholland	Left	.273	.270		.253	.269	.266	
Pedro Martinez	Right	.210	.193		.171	.204	.221	
Todd Stottlemyre	Right	.279	.216		.259	.242	.300	
Wilson Alvarez	Right	.258	.238		.291	.244	.224	
Mike Mussina	Left	.244	.251		.257	.246	.238	

(b) How do you expect a starting pitcher to perform at the end of the game after he has thrown 90 pitches? Is fatigue a factor for a starting pitcher?
(c) If a starting pitcher gets tired, what effect would that have on the batting average of opposing hitters?
(d) To see how a pitcher performs in the middle of the game as opposed to the beginning, we can compute the difference

$$\text{DIFF2} = \text{AVG(pitches 46–60)} - \text{AVG(pitchers 1–15)}.$$

For each pitcher, compute this difference and put the values in the "DIFF2" column of Table 8.13.
(e) Graph the differences you computed in part (d) and compute an average value. Can you detect a general tendency for pitchers to tire out between the beginning and middle parts of the game?

8.10. (Exercise 8.8 continued.)
If one thought that pitchers generally tire out during a game, then one would expect the opposing batting average to be larger for pitch count 91–105 than for pitch count 46–60.
(a) In Table 8.13, count the number of pitchers who pitch better on pitches 91–105 (compared to pitches 46–60), and count the number of pitchers who do better on pitches 46–60 put the counts in the table below. Next, find the proportion in each group and put the values in the "Proportion" column

	Count	Proportion
Pitchers who do better on pitches 91–105		
Pitchers who do better on pitches 46–60		

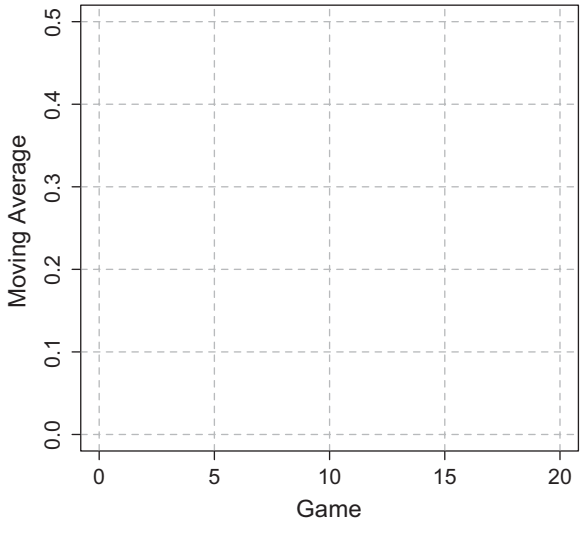
(b) What have you learned from the proportions you computed in (a)?
(c) Suppose that you now look at how batters perform during different pitch counts. If batters generally have a smaller AVG for pitches 91–105, does that mean that pitchers tend to get stronger during a baseball game? Explain why this might be a wrong conclusion.

8.11. (Exercise 5-1 continued.) Table 8.14 shows Ichiro Suzuki's daily batting record for the first 20 games in the 2004 baseball season. To see how Suzuki's performs in short periods, one can compute moving averages of five games. For example, in games 1 through 5, Table 8.14 shows that Suzuki had seven hits in 22 at-bats for a batting average of $7/22 = .318$. In the next group of five games (2 through 6), Suzuki was 8 for 23 for a batting average of $8/23 = .348$.
(a) For each group of five games, find the number of at-bats, hits, and five-game moving batting average. Put the answers in the table.
(b) Graph the moving averages against the middle-game number on the grid below. (For example, you would graph the first moving average .318 against the mid-game number 3, the moving average .348 against the mid-game number 4, and so on.)

Table 8.14. Computation of moving averages for Ichiro Suzuki's 2004 daily batting record

Game	AB	H	5 Games	5-Game AB	5-Game H	Moving AVG
1	4	1				
2	5	1				
3	4	0	1 through 5	22	7	$7/22 = .318$
4	5	3	2 through 6	23	8	$8/23 = .348$
5	4	2	3 through 7			
6	5	2	4 through 8			
7	5	2	5 through 9			
8	5	2	6 through 10			
9	5	0	7 through 11			
10	5	0	8 through 12			
11	4	1	9 through 13			
12	4	3	10 through 14			
13	6	2	11 through 15			
14	3	0	12 through 16			
15	3	1	13 through 17			
16	4	0	14 through 18			
17	4	2	15 through 19			
18	4	1	16 through 20			
19	5	1				
20	4	1				

(c) Describe the pattern that you see in the moving average plot. Are there any periods that Suzuki appears to be hot or cold?



8.12. (Exercise 8.11 continued.) Table 8.15 shows Suzuki's day-to-day batting performance for the first 20 games of the 2004 season. Suzuki's batting average for the whole season

Table 8.15. Ichiro Suzuki's 2004 daily batting record for the first 20 games

Game	AB	H	Game AVG	Hot or Cold	Game	AB	H	Game AVG	Hot or Cold
1	4	1	$1/4 = .250$	Cold	11	4	1		
2	5	1	$1/5 = .200$	Cold	12	4	3		
3	4	0			13	6	2		
4	5	3			14	3	0		
5	4	2			15	3	1		
6	5	2			16	4	0		
7	5	2			17	4	2		
8	5	2			18	4	1		
9	5	0			19	5	1		
10	5	0			20	4	1		

was .372. We say that Suzuki is hot for a particular game if his game batting average is over .372; otherwise we say that Suzuki is cold. For example, on game 1, his game average was $1/4 = .250$. Since this is smaler than .300, we say that he was hot in game 1.

(a) Complete Table 8.15. For each game, compute the game batting average and classify the game as hot or cold.

(b) From the sequence of Hot and Cold games, find
   i. the longest run of Hot games,
   ii. the longest run of Cold games,
   iii. the total number of runs.

8.13. (Exercise 8.11 continued.) For the game-to-game batting data for Ichiro Suzuki, count the number of games that Suzuki had 0 hits, 1 hit, etc. Put your counts in the table below.

Number of hits	0	1	2	3
Count				

8.14. Table 8.16 shows the game results for the New York Yankees and the Atlanta Braves for the first thirty games of the 2015 season.

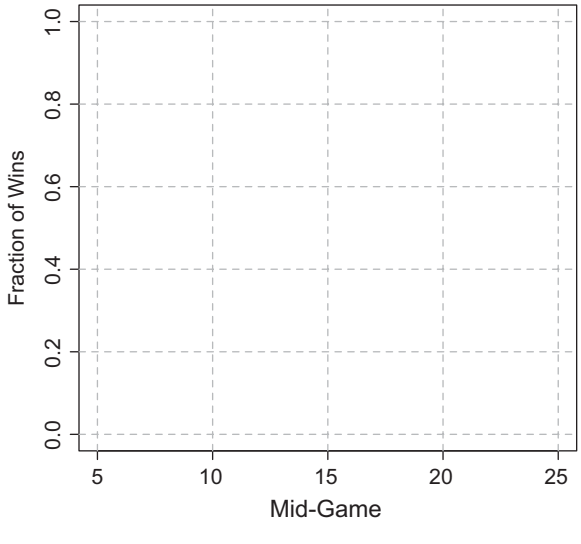
Table 8.16. Game results for New York Yankees and Atlanta Braves for first thirty games of the 2015 season

Yankees															
Games 1–15:	L	W	L	L	L	W	W	L	L	W	W	W	L	W	W
Games 16–30:	W	W	L	W	W	W	L	W	W	W	L	W	L	W	W
Braves															
Games 1–15:	W	W	W	W	W	L	W	L	L	W	L	W	L	L	L
Games 16–30:	L	W	L	W	L	L	L	W	L	W	L	W	W	L	L

(a) Compute all moving fractions of wins using a width of 9 games for both teams. That is, find the proportion of wins in games 1–9, in games 2–10, . . . , in games 22–30. Put your results in the table below

Games	Mid game	Yankees Fraction of wins	Braves Fraction of wins	Games	Mid game	Yankees Fraction of wins	Braves Fraction of wns
1–9	5			12–20	16		
2–10	6			13–21	17		
3–11	7			14–22	18		
4–12	8			15–23	19		
5–13	9			16–24	20		
6–14	10			17–25	21		
7–15	11			18–26	22		
8–16	12			19–27	23		
9–17	13			20–28	24		
10–18	14			21–29	25		
11–19	15			22–30	26		

(b) Graph the moving fractions against the mid-game for the two teams on the axes below.



(c) Comment on the pattern of moving-fractions that you see in the graph. Did the Yankees or Braves seem unusually hot or cold during these first 30 games?

8.15. (Exercise 8.14 continued.) For each of the win/loss sequences of the Yankees and Braves shown in Table 8.16, find (1) the longest run of wins or losses, and (2) the total number of runs. Put your answers in the table below.

Team	Longest run of wins or losses	Number of runs
Yankees		
Braves		

8.16. (Exercise 8.12 continued.) Suppose Ichiro Suzuki is truly a consistent hitter. The chance that he gets a hit on a single at-bat is .372 (his true batting average), and the results of different at-bats are independent. Suppose we simulate data from this model using Suzuki's actual number of at-bats for the 20 games. The results of this simulation are placed in Table 8.17.

Table 8.17. Simulated data for 20 games assuming Suzuki is a consistent hitter with hit probability of .372.

Game	AB	H	Game AVG	Hot or Cold	Game	AB	H	Game AVG	Hot or Cold
1	4	2	$2/4 = .500$	Hot	11	4	2		
2	5	1			12	4	0		
3	4	1			13	6	1		
4	5	2			14	3	1		
5	4	2			15	3	1		
6	5	2			16	4	2		
7	5	1			17	4	1		
8	5	1			18	4	2		
9	5	3			19	5	1		
10	5	3			20	4	3		

(a) As in Exercise 8.12, classify the hitting of each game as either hot or cold, depending if the game batting average is larger or smaller than .372. Put your results in the table.

(b) Compute
   i. the longest run of Hot games,
   ii. the longest run of Cold games,
   iii. the total number of runs.

(c) Compare the results of these simulated data (the longest run of Hot, the longest run of Cold, and the total number of runs) with the results using Suzuki's actual data in Table 8.15. Is there any evidence that Suzuki is a streaky hitter?

8.17. (Exercise 8.16 continued.) In Exercise 8.16, we assumed that Suzuki was really a consistent hitter with probability of .372 of getting a hit, and simulated 20 games of hitting assuming this consistent model. Suppose we repeat this experiment 20 times. For each experiment, we simulate 20 games of hitting (assuming Suzuki is really a consistent hitter), and then classify each game as either hot or cold. The results are shown in Table 8.18.

(a) For each experiment, find the longest run of either Hot or Cold. Put the longest run in the LONGEST RUN column of the table.

Table 8.18. Simulated data for twenty experiments, where one experiment consists of 20 games assuming Suzuki is a consistent hitter with hit probability of .372. Suzuki's performance in each game is classified as hot (H) or cold (C)

Game																				Longest
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	run
C	H	H	H	H	H	H	C	H	H	H	H	C	H	H	H	H	H	C	H	
C	H	H	C	C	H	C	H	C	C	C	H	H	C	H	H	C	C	C	H	
H	H	C	C	C	H	H	H	C	H	C	C	C	C	C	H	H	H	C	H	
H	H	H	H	C	H	C	C	C	C	H	C	C	C	C	H	C	C	H	C	
H	C	H	H	H	C	C	H	H	C	H	C	C	H	C	C	C	H	H	H	
C	C	H	C	C	H	C	H	C	H	H	H	H	C	C	H	C	H	H	H	
C	H	C	C	H	C	H	C	C	H	H	C	C	H	C	C	H	H	H	H	
H	C	C	H	C	C	C	H	H	C	C	C	H	C	H	C	H	C	H	H	
C	H	H	C	H	H	H	H	C	C	H	C	H	C	C	H	H	C	H	C	
H	H	H	C	C	H	H	C	H	C	H	C	H	C	C	H	H	H	H	H	
C	H	C	H	C	C	H	H	C	H	H	C	H	C	C	C	C	H	C	H	
H	C	C	H	H	C	H	H	H	H	C	C	C	C	C	C	H	C	H	C	
C	H	H	H	H	H	H	H	H	C	C	H	C	C	H	H	H	C	H	C	
H	H	C	H	H	H	C	C	H	H	H	C	C	C	C	H	H	H	C	H	
H	C	C	H	H	C	H	H	C	C	C	H	H	C	H	H	H	H	C	C	
H	C	H	C	H	C	H	H	H	H	C	C	H	H	C	C	H	C	C	C	
C	C	H	H	C	H	C	H	C	H	C	H	C	C	C	H	H	C	H	H	
C	C	C	H	H	H	C	H	C	H	H	H	H	C	C	C	C	C	H	C	
H	H	C	H	H	C	C	C	H	H	H	H	C	C	C	H	H	H	C	C	
H	H	H	C	H	H	H	H	H	C	H	C	H	C	H	C	H	H	H	C	

(b) Construct a dotplot of the longest run values using the number line below.

```
+---+---+---+---+---+---+---+---+---+---+--
0   1   2   3   4   5   6   7   8   9   10  LONGEST RUN
```

(c) For Suzuki's data in Table 8.15, the longest run (of either Hot or Cold) was equal to ____. Does this value seem unusually small or large if Suzuki was really a consistent hitter? (Compare Suzuki's value with the longest run values plotted in part (b).)

8.18. Was Chris Davis a streaky home run hitter in 2013? Using Davis' home run log from Case Study 5-1, one can construct a moving average plot of his home run rate. The moving averages are shown in Figure 8.12 using a window of ten games. We see that Davis showed some streaky behavior—his 10-game home run rate was over .15 about games 50 and 70 and his home run rate dropped to 0 about game 100.

How would Davis perform if he were truly a consistent home run hitter? In 2013, Davis hit 53 home runs in 673 plate appearances for a season rate of $53/673 = .0789$. Suppose that the probability that Davis hits a home run in a single plate appearance is $p = .0789$, and the results of different plate appearances are independent. Using this consistent model, four seasons of home run hitting were simulated using the same

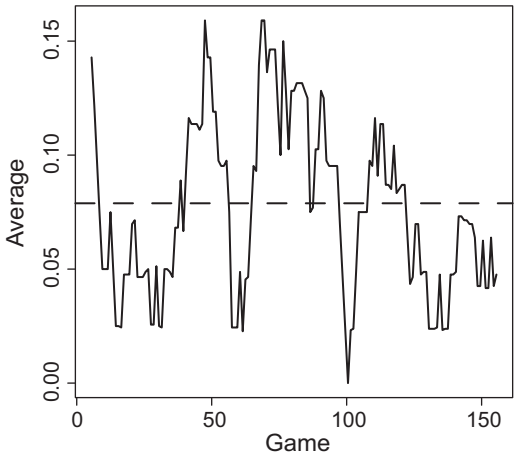


Figure 8.12. Moving average plot of 2013 Chris Davis' home run rates using a window of ten games.

game-to-game plate appearances as Davis. Moving average plots of the home run rates for the four simulations are shown in Figure 8.13.

Discuss any unusual features of each of these four simulations. Do you think that Davis' moving average plot is different, with respect to streakiness, from the plots of the four simulated consistent hitters? Do you think that Davis was a truly streaky home run hitter?

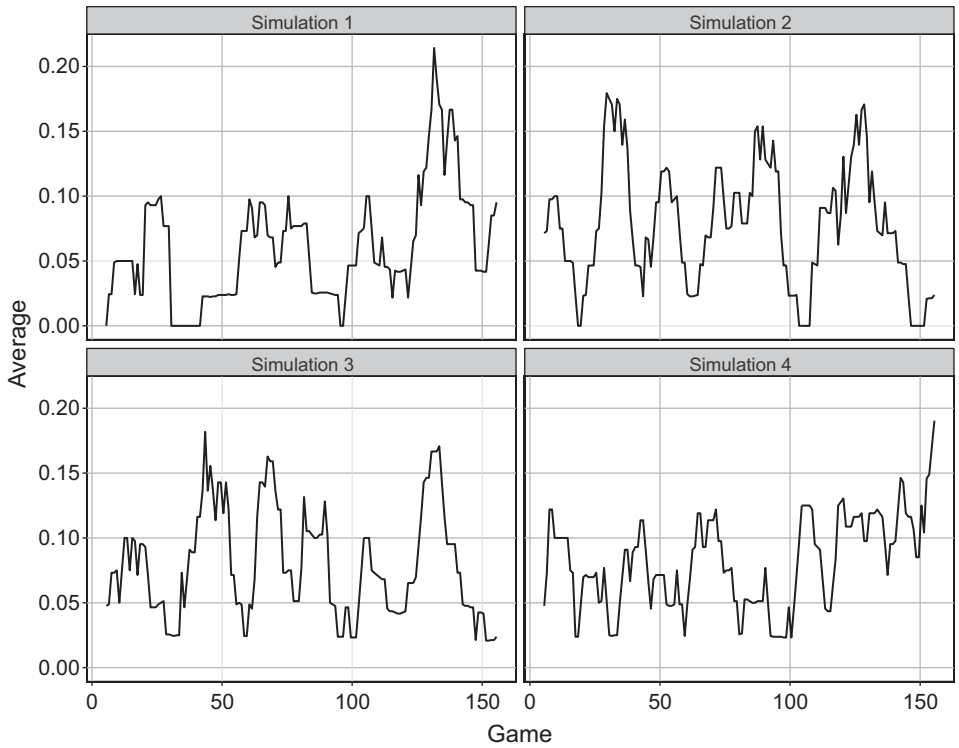


Figure 8.13. Moving average plot of home run rates of simulated data using a consistent model with home run probability .0789.

8.19. On the `baseball-reference.com` web site, one can explore the patterns of hot streaks and slumps of any Major League baseball team in history. (Look for the streaks link on the web site.) The Oakland Athletics had an interesting pattern of wins and losses during the 2002 season. Figure 8.14 displays a moving average of the winning proportion using a window of 20 games.

(a) Looking at Figure 8.14, describe the slumps and hot streaks of Oakland during this season.

(b) Oakland concluded this season with a 103-59 win/loss record with a winning percentage of 63.6%. Simulate sequences of wins and losses for an entire 162 game season assuming Oakland is a truly consistent team with probability of winning each game equal to .636. By comparing the simulated sequences with Figure 8.14, can you say that Oakland was truly a streaky team in the 2002 season?

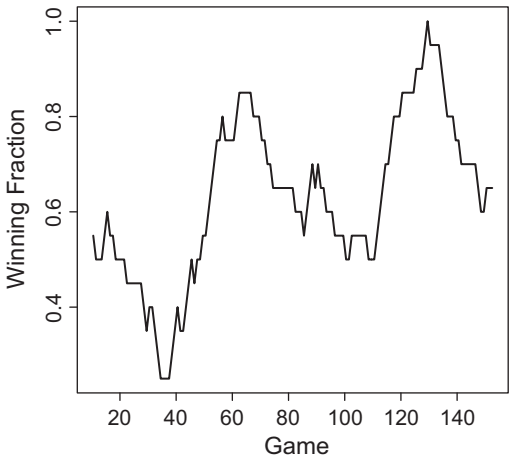


Figure 8.14. Moving average plot of winning fraction of the 2002 Oakland Athletics using a window of 20 games.

(c) Choose another team in history that had a reputation for being unusually streaky during a particular season. By using the simulation method described in part (b), decide if the pattern of wins and losses for this team was different from that of a truly consistent team.

Further Reading

Situational baseball data for batters and pitchers is presented in many internet sites such as FanGraphs. Chapter 4 of Albert and Bennett (2003) describes the different probability models that can be used for situational hitting data. Based on an analysis of situational data for the 1998 season, they classify the different true situational effects as "no effects", biases, or ability effects. Gilovich et al (1985) describe the tendency of people to misinterpret the inherent streaky nature of random sequences. Albert and Bennett (2003), Chapter 5 discuss hitters who are genuinely consistent or streaky, and describe how one can perform inference about streakiness.

10.1090/text/034/09

9

Modeling Baseball Using a Markov Chain

What's On Deck?

In this chapter, we introduce a special probability model, called a Markov Chain, to represent the sequence of plays in a baseball game. The state of an inning is defined by the number of outs and the runners on the three bases. A half-inning of baseball can be regarded as a sequence of states until there are three outs. One can represent the random movement between states by a Markov Chain, where one moves from one state to another state with a given probability. In Case Study 9.1, we introduce a Markov Chain using an example of a person traveling between cities, and in Case Study 9.2 we extend the basic structure of a Markov Chain to a baseball game. To specify a Markov Chain, one needs only to specify the probabilities that control the movement between states, and we use actual baseball game data to estimate these probabilities. The remainder of the chapter shows the usefulness of this probability model to answer questions about baseball. Using the model, one can estimate the number of players that come to bat during the inning, and compute the probability that a team will score at least one run in an inning. One of the most useful computations is the expected number of runs scored from a particular state in an inning. Using these expected numbers of runs scored, one can assess the value of a particular hit, such as a home run. In addition, one can use these expected numbers of runs scored to judge the value of a particular batting play. Baseball fans are interested in the value of particular baseball strategies, such as a sacrifice bunt and a steal, and this probability model is useful in seeing if a particular strategy is helpful towards the general goal of scoring runs.

9.1 Introduction to a Markov Chain

Topics Covered: Transition probabilities, absorbing states, expected number of visits to different states, matrices, matrix multiplication and inversion.

In this chapter, we will show how a baseball game can be modeled using a special probability model called a Markov Chain. In this first case study, we give a gentle introduction to Markov Chains and the remainder of the case studies will show the application of this idea to baseball.

Suppose a baseball fan is planning an interesting (one might call it bizarre) trip to New York City. He will start from San Francisco (SF). The next day, he will either stay in SF another night or he will fly to St. Louis (STL); he is equally likely to stay or fly to STL. If he does go to STL, then he will stay in STL another night with probability .4, or fly back to SF with probability .3, or fly to Chicago (CHI) with probability .3. If he is in CHI, the next day he is

equally likely to stay another night, fly to SF, fly to STL, or fly to New York City (NY). Once he arrives at New York City, he will stay there—his trip is over.

Obviously the exact trip of this fan is random. All he knows for sure is that he will eventually arrive in New York City. But what is the probability that he will arrive in NY in exactly two days? In three days? How long will it take this fan, on average, to get to NY? And how many days can this fan expect to stay in SF, STL, and CHI?

The map in Figure 9.1 shows the possible city-to-city connections of our traveler. An arrow from one city to another city indicates the particular trip is possible and the number label is the probability of this trip. An arrow from a city back to the same city corresponds to a night where the traveler stays over.

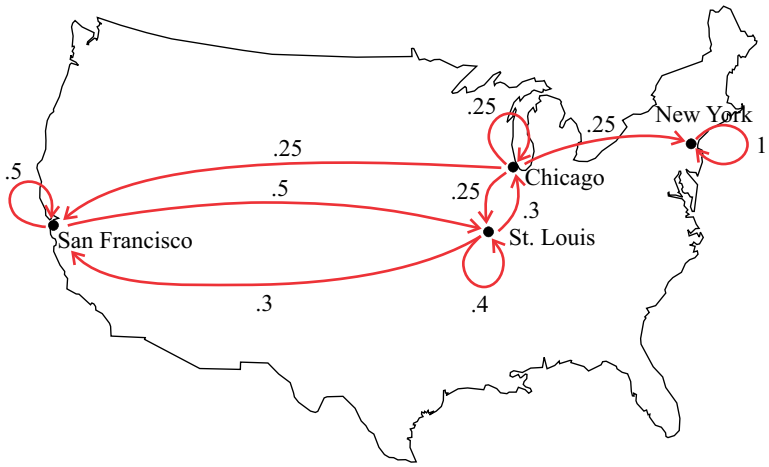


Figure 9.1. City-to-city connections of a traveler in trip from San Francisco to New York City.

A Markov Chain describes movement between a number of locations, called states. Here there are four states for our traveler, SF, STL, CHI, and NY. Given that the traveler is in a particular state (city), then we will move to another state the next day with specific probabilities. The probability that he will move to another city depends only on his current location and not on the previous cities he visited. (This is a special property of a Markov Chain.) Given the information above, we know if the person is currently in SF, then he will travel to

SF STL CHI NY

with respective probabilities

.5 .5 0 0.

If he is currently in STL, then he travels to these four cities the next day with respective probabilities

.3 .4 .3 0.

These probabilities are called transition probabilities—they describe the likelihoods of moving between various states in the Markov Chain. We summarize all of these transition probabilities by means of a transition matrix P shown in Table 9.1.

Table 9.1. Matrix of transition probabilities P for the traveler example

	SF	STL	CHI	NY
SF	0.50	0.50	0.00	0.00
STL	0.30	0.40	0.30	0.00
CHI	0.25	0.25	0.25	0.25
NY	0.00	0.00	0.00	1.00

The first row of this matrix P gives the transition probabilities of the traveler starting from SF, the second row gives the probabilities starting from STL, and so on.

One special feature of this particular Markov Chain is that once the traveler arrives at New York City, he will remain there. We call the state NY an absorbing state—as indicated in the transition matrix, the probability of remaining in an absorbing state is one.

There are a number of nice results about Markov Chains that will make it easy to answer the questions posed above.

What is the Probability of Reaching States After a Specific Number of Moves?

The matrix P gives the probabilities of reaching various states in exactly one move. We can obtain the probabilities of reaching states in two moves by squaring the matrix P:

$$P = P \times P = \begin{bmatrix} .5 & .5 & 0 & 0 \\ .3 & .4 & .3 & 0 \\ .25 & .25 & .25 & .25 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} .5 & .5 & 0 & 0 \\ .3 & .4 & .3 & 0 \\ .25 & .25 & .25 & .25 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} .4 & .45 & .15 & 0 \\ .345 & .385 & .195 & .075 \\ .2625 & .2875 & .1375 & ..3125 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The first row of this matrix, $[.4,\ .45,\ .15,\ 0]$, gives the probability that, starting at SF, we will be in the respective cities SF, STL, CHI, NY in exactly two days. So he will be in Chicago in two days with probability .15. Note that the probability that he'll be in NY in two days is 0.

If we multiply the matrix P additional times, we obtain the probabilities of being in the states after more than two days. If we multiply

$$P^6 = P \times P \times P \times P \times P \times P,$$

we obtain the six-step transition probabilities

$$P^6 = \begin{bmatrix} .323 & .361 & .154 & .163 \\ .293 & .328 & .140 & .240 \\ .219 & .244 & .104 & .433 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The first row gives the probabilities of being in the four states, starting in SF, in exactly six days. We see that if we start in SF, the chance that we will be in NY in six days is .163.

How Long Will One Stay in the States?

If we remove the row and column corresponding to the single absorbing state from the matrix P, we have the matrix Q of transition probabilities

$$Q = \begin{bmatrix} 0.5 & 0.5 & .0 \\ 0.3 & 0.4 & 0.3 \\ 0.25 & 0.25 & 0.25 \end{bmatrix}$$

To find the expected number of times that the process visits the states before being absorbed, we simply compute the matrix $E = (I - Q)^{-1}$, where I is the 3 by 3 identity matrix. Here

$$E = \left(I - \begin{bmatrix} 0.5 & 0.5 & .0 \\ 0.3 & 0.4 & 0.3 \\ 0.25 & 0.25 & 0.25 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 10 & 10 & 4 \\ 8 & 10 & 4 \\ 6 & 6.67 & 4 \end{bmatrix}$$

Let's interpret this matrix. The first row, $[\,10 \quad 10 \quad 4\,]$, gives the expected number of visits to the cities SF, STL, and CHI if we start at San Francisco. So in the trip, the person will be in SF, on average, ten days (including the starting day), STL for ten days, and CHI for four days. Likewise, the second and third rows of the matrix E give the expected number of visits if we start in St Louis and Chicago, respectively.

The matrix E tells us how the traveler will do on average. But actually there is a lot of variation in the length of the trip. To see this, I had the computer simulate 1000 trips originating from San Francisco using the matrix of transition probabilities. Figure 9.2 is a histogram of the lengths (in days) of the 1000 trips. The distribution is right-skewed where the length ranged from three to 142 days. The mean length of a trip was 23.31 days. This is reasonable—looking at the first row of the matrix E, we see that we will spend a total of $10 + 10 + 4 = 24$ days in our journey.

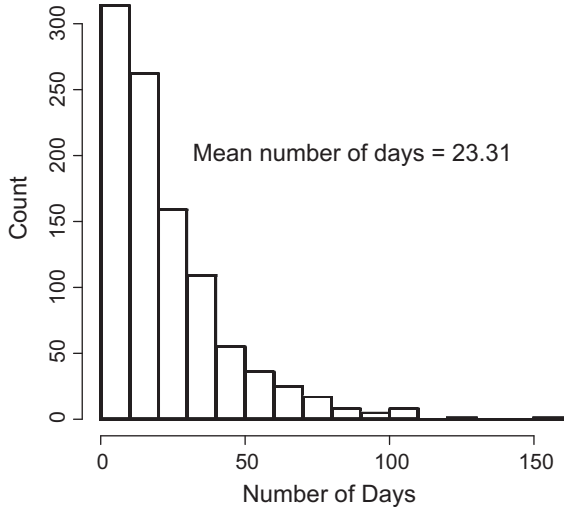


Figure 9.2. Histogram of lengths of 1000 trips from San Francisco simulated from the Markov Chain.

How Many Frequent Flyer Miles?

Suppose our traveler plans on flying and he is interested in how many frequent flyer miles he will accumulate in his trip. First we record in a table the distances between all of these cities as displayed in Table 9.2.

Table 9.2. Distances between all cities in the traveler example

	SF	STL	CHI	NY
SF	0	2062	2133	2908
STL	2062	0	294	953
CHI	2133	294	0	790
NY	2908	953	790	0

Then we find the expected length of a one-day trip from each city. If we start in SF, then we will travel 0, 2062, 2133, 2908 miles with respective probabilities 0.5, 0.5, 0, 0. So the expected length of a one-day trip from SF is

$$0 \times 0.5 + 2062 \times 0.5 + 2133 \times 0 + 2908 \times 0 = 1031.$$

Similarly, we find the expected length of a one-day trip from STL and CHI to be 709.8 and 804.3 miles, respectively.

Last, using a standard result for Markov Chains, we can compute the expected length of the trip from each starting city by multiplying our matrix E by the column of expected length of one-day trips.

$$\begin{bmatrix} 10 & 10 & 4 \\ 8 & 10 & 4 \\ 6 & 6.67 & 4 \end{bmatrix} \times \begin{bmatrix} 1031 \\ 709.8 \\ 804.3 \end{bmatrix} = \begin{bmatrix} 20,625 \\ 18,563 \\ 14,135 \end{bmatrix}$$

The product vector gives the expected length in miles of the total trip starting from each of the three cities. So if we start at San Francisco, we can expect to log 20,625 frequent flyer miles before getting to New York City. We will travel, on average, 18,563 miles if we start our trip at St. Louis.

9.2 A Half-inning of Baseball as a Markov Chain

Topics Covered: State of an inning in a baseball game, transitions between states, batting plays, runs scored.

In the first case study, we described the basic structure of a Markov Chain. How can this model be applied to baseball?

We focus on the run production of a baseball team during its half-inning at-bat. A state will describe the current runners on bases situation and the number of outs. Each base (first, second, and third) can either be occupied or not. There are a total of eight possible base situations that will be graphically represented in Table 9.3 using symbols on a diamond shape.

Table 9.3. Possible runner situations in a baseball game

Empty	1st	2nd	3rd	1st, 2nd	1st, 3rd	2nd, 3rd	Loaded
◇	◇	◇	◇	◇	◇	◇	◇

When a batter comes up, the number of outs at any one time during an inning can either be 0, 1, or 2. So when a batter comes to bat, there are eight possible base situations and three different out situations, and so there are $8 \times 3 = 24$ different base/out situations. If we add the final situation, three outs, to this list, we have a total of $24 + 1 = 25$ possible states. In Table 9.4 all of the states are presented by use of a table classified by number of outs and the base situation.

Table 9.4. Diagram of all 25 possible states of an inning defined by the runners on base situation and the number of outs

	Bases Situation							
								
0 outs								
1 out								
2 outs								
3 Outs								

A player comes to bat when the inning is in a particular state, say runners on 1st and 2nd with one out. There will be a batting play, such as a hit, out, or walk, that will change the inning state. For example, suppose a batter comes up with a runner on 1st and one out

( , 1 out)

He singles into right field, moving the runner from 1st to 3rd. The new inning state is

( , 1 out)

We can view a half-inning of baseball as a sequence of changes in state, starting with (no runners on base, 0 out) and ending with (3 outs). (To simplify the following discussion, we will ignore non-batting plays, such as steals and balks, that can also change the state of an inning. A more realistic model would include these non-batting plays in the analysis.)

A Markov Chain can be used to model the change in states during an inning. We assume that the probability of moving to a particular (bases, outs) state only depends on the current state and not on any earlier state. We let p_{ij} denote the probability of moving from the ith state to the jth state. We let P denote the matrix of transition probabilities with 25 rows and 25 columns. Note that the (3 outs) situation is an absorbing state in the Markov Chain since one cannot leave this state once it is entered. In other words, the inning is over when there are three outs.

When there is a change in the (bases, outs) state, runs can score. Suppose that there are R_{before} runners on base and O_{before} outs before the batting play, and R_{after} runners and O_{after} outs after the play. Then the number of runs scored in this transition would be

$$\text{Runs Scored} = (R_{\text{before}} + O_{\text{before}} + 1) - (R_{\text{after}} + O_{\text{after}}).$$

Let's illustrate this computation for one batting play. Suppose there are runners on 1st and 2nd with one out. The batter doubles, scoring both runners, leaving a runner on 2nd with one out. Table 9.5 verifies that the Runs Scored formula gives that two runs scored on this play.

Table 9.5. Illustration of the number of runs scored for one batting play

Situation before Play	$(R_{\text{before}}, O_{\text{before}})$	Play	Situation after Play	$(R_{\text{after}}, O_{\text{after}})$	Runs Scored
 , 1 out	(2, 1)	Double	 , 1 out	(1, 1)	2

9.3 Useful Markov Chain Calculations

Topics Covered: Computation of probabilities by use of frequencies, matrix computations, expected number of visits to states, expected runs scored in a half-inning, computation of event probabilities by simulation.

To use this Markov Chain model, we need to estimate the matrix of transition probabilities P. We use play-by-play data for all the Major League teams in the 2014 season to estimate this matrix.

Let's illustrate computing the transition probabilities starting from the bases empty, no outs state. All half-innings begin with this particular state, although this state may happen more than once in a particular inning. In the 2014 season, there were 45,312 instances where this state occurred. There are five possible transitions from this state, depending on the batting play:

- The batter can hit a home run, and the state remains at bases empty and no outs.
- The batter can get out, and the state changes to bases empty and one out.
- The batter can get to first base by a single, walk, hit-by-pitch, or an error and the state changes to runner on first and no outs.
- The batter can hit a double, and the state changes to runner on second with no outs.
- The batter can hit a triple, and the state changes to runner on third base with no outs.

Table 9.6 below shows the five possible transitions from (bases empty, no outs) and the frequency with which each transition occurred. Note that the most common transition is an out this event happened 31,059 times out of a total of 45,312 transitions and so the probability of this transition is estimated to be $31{,}059/45{,}312 = 0.6854$. In a similar fashion, we can estimate the probability of the four other possible transitions.

Table 9.6. All possible transitions from the no-runners, no outs state. For each state, the number of runs scored, the number of times this transition occurred, and the corresponding probability are given

Batting Play	End State	Runs Scored	Count	Probability
Home run	 , 0 out	1	1149	.0254
Out	 , 1 out	0	31059	.6854
Single or walk	 , 0 out	0	10663	.2353
Double	 , 0 out	0	2205	.0487
Triple	 , 0 out	0	236	.0052
Total			45312	1.000

Suppose instead that the current state is two outs with a runner on first. In this case, there are more possible transitions types depending on the advancement of the runner from first base. Table 9.7 shows all of the possible transitions from (runner on first, two outs), the frequency of each transition, and the corresponding probability. We see that the most likely play is an out that results in the third out the probability of this transition is $7988/11,630 = 0.6868$. Other possible transitions, ordered in terms of their likelihood of occurring, are

- a single, walk, or hit-by-pitch that result in runners on 1st and 2nd,
- a single where the runner on first advances to 3rd base,
- a home run that clears the bases,
- a double that scores the baserunner on 1st,
- a double that advances the baserunner to 3rd base,
- a triple that scores the baserunner on 1st,
- a single that scores the baserunner on 1st.

We see that the single scoring the runner on first is quite an unusual play—it happened only 10 times that particular season.

Table 9.7. All possible transitions from the (runner on first, two outs) state. For each state, the number of runs scored, the number of times this transition occurred, and the corresponding probability are given

Batting Play	End State	Runs Scored	Count	Probability
Home run	◇, 2 outs	2	282	.0242
Triple	◇, 2 outs	1	78	.0067
Double	◇, 2 outs	1	255	.0219
Double	◇, 2 outs	0	248	.0213
Single	◇, 2 outs	1	10	.0009
Single	◇, 2 outs	0	593	.0510
Single or Walk	◇, 2 outs	0	2176	.1871
Out	3 outs	0	7988	.6868
Total			11630	1.000

Suppose that we compute the probabilities of all possible transitions starting from each of the 24 possible initial states. We can then construct the matrix P (of dimension 25 by 25) that contains the transition probabilities for all (runners on base, number of outs) states. Given this transition matrix, we can perform several matrix calculations, such as done in the first case study, to find some quantities of interest.

Reaching Various States After a Given Number of Batters

Suppose three players come to bat at the beginning of an inning. What will be the state of the inning after these three plate appearances? What's the chance that there will be at least one runner in scoring position?

Remember the matrix P gives the probabilities of one-step transitions. To obtain the probabilities of three-step transitions, we simply multiply the matrix by itself three times:

$$P^3 = P \times P \times P.$$

Table 9.8 gives the first row of the matrix, displaying it in a familiar two-way format where rows correspond to the number of outs and the columns represent the runners situation.

Table 9.8. Transition probabilities of being in different states after three plays starting from the (no outs, no runners) state

	Bases Situation							
	◇							
0 outs	.0022	.0030	.0021	.0005	.0074	.0041	.0030	.0109
1 out	.0159	.0229	.0160	.0064	.0825	.0283	.0190	.0000
2 outs	.0411	.2441	.0789	.0300	.0000	.0000	.0000	.0000
3 outs				.3818				

These probabilities represent the chances of being in different inning states after three batters starting from the (no outs, no runners) state. We see that the most likely state after three at-bats is (3 outs) with a probability of .3818. The next most likely state is (2 outs, runner on first) with a probability of .2441. Several three-batter movements are impossible. For example, we see from the table that the probability of getting to (2 outs, runners on first and second) from three batters has a probability of 0.

What is the chance of having a runner in scoring position (that is, a runner on 2nd or 3rd base) after three at-bats? We look at the above table and sum the probabilities over the ◇,  ,  ,  ,  ,  states where one runner or more are in scoring position. So the probability of runners in scoring position after 3 batters is equal to $.0021 + .0064 + \cdots + .0000 + .0000 = .2891$.

How Many Batters?

As before, let Q denote the submatrix of P found by deleting the one row and one column corresponding to the absorbing (3 outs) state. The matrix $E = (I - Q)^{-1}$ contains the expected times that the inning will be in each state starting with each of the 24 possible beginning states. Suppose that we are currently at the (no runners on, no outs) state that begins the inning. The first row of the matrix E will give the expected number of visits to all the states (before absorption) given that one starts in this (no runners on, no outs) state. This first row of E is displayed in a convenient table format in Table 9.9.

Table 9.9. The expected number of visits to each state starting from the (no runners on, no outs) state

	Bases Situation								
	◇								
0 outs	1.037	0.249	0.059	0.008	0.065	0.021	0.013	0.016	
1 out	0.745	0.302	0.104	0.030	0.113	0.047	0.031	0.039	
2 outs	0.583	0.311	0.127	0.051	0.141	0.066	0.033	0.045	
SUM									4.236

Of course, since we are starting at the (no runners on, no outs) state, we'll visit this state at least once this table tells us that we will visit it, on the average, 1.037 times. Also, we will visit the (runner on first, no outs) state an average of .249 times, the (runner on second, no outs) state an average of .059 times, and so on.

If we sum all of these expected state counts, we will obtain the expected number of state visits before absorption. In baseball lingo, this sum will be the expected number of batters before the inning is over. Here the sum is 4.236, which means that, on average, there will be 4.236 batters in the remainder of the inning starting with the (no runners on, no outs) state.

In a similar fashion, one can use the matrix E to compute the expected number of batters in the remainder of the inning starting from each of the 24 possible states. Table 9.10 shows the "expected number of batters" matrix.

Table 9.10. Expected number of batters in the remainder of the inning starting from each possible state

	Bases Situation							
	◇	◇	◇	◇	◇	◇	◇	◇
0 outs	4.236	4.022	4.316	4.317	3.994	4.105	4.370	4.010
1 out	2.844	2.683	2.910	2.982	2.648	2.687	3.007	2.662
2 outs	1.446	1.447	1.497	1.513	1.407	1.424	1.494	1.389

A manager can use this matrix in strategic decisions during a game. For example, suppose there are bases loaded with one out. Looking at this table, we see that, on average, there will be 2.662 batters in the remainder of this inning. A manager can use this to plan his batting lineup; for example, it might help him make a decision regarding the use of a pinch-hitter.

Expected Runs in the Remainder of the Inning

The expected number of visits matrix E can be used to compute the expected number of runs in the remainder of the inning starting from each state. This computation is similar to the computation for the expected number of frequent flyer miles for our first example. Let $R_{\text{one step}}$ denote the column vector that contains the expected number of runs that will be scored in a single batting play starting from each of the 24 states. Let's illustrate the computation of the first element of $R_{\text{one step}}$. Suppose that one starts in the (no runners on, no outs) state. In one batting play, only 0 and 1 runs can score. The probability of scoring one run is the chance of hitting a home run that is estimated to be .0254, and so the chance of scoring no runs is estimated to be $1 - .0254 = .9746$. The expected number of runs scored in one batting play is therefore

$$1 \times .0254 + 0 \times .946 = .0254.$$

If we do this computation for each of the 24 states, we obtain the vector $R_{\text{one step}}$ that is displayed in matrix form in Table 9.11.

The vector of expected number of runs scored, denoted by R, is found by multiplying the number of visits matrix E by $R_{\text{one step}}$:

$$R = E \times R_{\text{one step}}.$$

Table 9.11. The expected number of runs scored in a single batting play starting at each possible state. The vector contains these values

	Bases Situation							
	◇							
0 outs	.025	.072	.149	.491	.201	.713	.603	.843
1 out	.021	.076	.170	.476	.248	.543	.639	.814
2 outs	.022	.078	.175	.223	.245	.271	.317	.505

Note that R is a column vector with 24 entries, where the entries correspond to the expected number of runs in the remainder of the inning starting from each of the 24 states. This vector R is displayed, in matrix form, in Table 9.12.

Table 9.12. The expected number of runs scored in the remainder of the inning starting at each possible state

	Bases Situation							
	◇							
0 outs	0.44	0.81	1.06	1.29	1.36	1.73	1.90	2.21
1 out	0.22	0.46	0.61	0.90	0.83	1.06	1.33	1.49
2 outs	0.08	0.19	0.28	0.33	0.38	0.40	0.47	0.63

This fundamental matrix, often called the "run expectancy matrix", is useful for many purposes in baseball research. Starting an inning in the (no runners, no outs) state, we see from Table 9.12 that a team will score, on average, .44 runs. In contrast, when there are bases loaded with one out, there is a good potential to score runs; using the table, we see that, on average, 1.49 runs will score from this state. This run expectancy matrix gives the potential for a team scoring runs starting from each (bases, outs) situation.

Computing Other Event Probabilities

Once the Markov Chain model is defined by means of the transition probability matrix P, one can compute the probability of any event of interest by simulating the chain many times. For example, suppose a manager is interested in the probability that the team will score at least one run if there is a runner on 3rd base with one out. On a computer, one can simulate the batting plays in the remainder of the inning by using the transition probability matrix, starting with the row of the matrix corresponding to the (3rd base, one out) state, and stopping when the (3 outs) state has been reached. One records if one or more runs were scored in this simulated inning. Then one repeats this simulation process for a large number of innings, each time recording if one or more runs occurred. Then the probability of scoring at least one run, Prob(at least one run), can be approximated by

$$\frac{\#\text{ of innings where at least one run was scored}}{\#\text{ of innings simulated}}.$$

We actually simulated the Markov Chain starting from each of the 24 possible inning states. For each starting state, we simulated the remainder of the inning 10,000 times. From each state,

we computed the probability that at least one run is scored in the remainder of the innings. The estimated probabilities are displayed in Table 9.13.

Table 9.13. The probability of scoring at least one run in the remainder of the inning starting from each possible state

	Bases Situation							
	◇	◇	◇	◇	◇	◇	◇	◇
0 outs	.246	.391	.595	.827	.589	.865	.852	.842
1 out	.138	.236	.380	.635	.386	.605	.674	.643
2 outs	.059	.101	.196	.244	.206	.241	.217	.273

When the inning starts (no outs, no runners on), the probability the team will score at least one run is .246. This chance of scoring drops to .059 when there are two outs in the inning. A team is most likely to score when the bases are loaded with no outs—the chance of scoring is .842.

9.4 The Value of Different On-base Events

Topics Covered: Value of a batting event defined by expected runs, mean value of a particular type of hit.

Using the run expectancy matrix R defined in the previous case study, we can define the value of a batting event. Suppose a batter comes to bat in a particular state (runners on base and number of outs) where the expected number of runs scored in the remainder of the inning is R_{before}. After the batter event, there is a new (runners on base, outs) situation with an expected number of runs scored called R_{after}. Then the value of this batting event is

$$\text{VALUE} - R_{\text{after}} - R_{\text{before}} + (\text{runs scored on play}).$$

Values of a Terrible Play and a Great Play

Let us illustrate this formula for two extreme cases that correspond to the least and most valuable batting plays. Suppose that the bases are loaded with no outs. The batter hits a sharp grounder to third; the third baseman touches third base, throws it quickly to the second baseman who touches second base and throws it to first, completing an (unusual) triple play. Clearly this is a bad play for the hitter (and the team)—the question is how bad?

When the batter came to the plate, the run potential (bases loaded, no outs) is 2.21 runs. After the play, there are three outs and a run potential of zero. No runs scored on this play. The value of this plate appearance is

$$\text{VALUE} = (0 - 2.21) + 0 = -2.21.$$

So this bad hitting play essentially cost the team about two runs. This is the worst possible play where the worth is defined in terms of this value measure.

Let's contrast this with a great batting play. The bases are loaded with two outs and the batter hits a deep fly that goes over the center field fence it's a grand slam! When this batter came to bat, the run potential of (bases loaded, two outs) is .63 runs. After the play, the bases are empty (with 2 outs) and the run potential is .08 runs. Four runs scored on this play. The value

of this plate appearance is

$$\text{VALUE} = (.08 - .63) + 4 = 3.45.$$

One might think the value of this home run would be four runs after all, four runs scored on this play. But, by clearing the bases, the batter has decreased the run potential in the remainder of the inning, and the value measure adjusts for this decrease.

The Value of a Home Run

Using the notion of value, we can measure the effectiveness of different types of batting plays, such as hits, walks, sacrifice flies, and outs. Here we focus on the biggest hit, the home run. There were a total of 4186 home runs hit in the MLB in 2014. But these home runs were not equally valuable—certainly a home run hit with runners on base is more valuable than a home run hit with the bases empty. In Table 9.14, we classify all of the home runs by the bases situation and the number of outs. In each cell, the number on top is the value of the home run in that particular situation and the number in parentheses is the number of home runs hit in that situation.

Table 9.14. Value and the number of home runs that occur in all possible situations using 2014 data

	Bases Situation							
	◇	◇	◇	◇	◇	◇	◇	◇
0 outs	1	1.63	1.38	1.08	2.07	1.67	1.56	2.10
	(1142)	(257)	(69)	(9)	(51)	(27)	(6)	(13)
1 out	1	1.76	1.62	1.29	2.40	2.14	1.88	2.73
	(687)	(306)	(135)	(37)	(111)	(39)	(23)	(34)
2 outs	1	1.89	1.78	1.74	2.70	2.67	2.58	3.45
	(561)	(280)	(114)	(54)	(117)	(43)	(34)	(37)

Note that most of the home runs were hit with the bases empty in fact, 2390 (57%) of the home runs hit were solo shots and the value of each of these home runs is 1. In contrast, only 13% of the home runs were hit with two or more runners on base. The values of these "two runners or more on base" home runs vary from 1.56 (runners on 2nd and 3rd and no outs) to 3.45 (bases loaded with 2 outs).

To measure the value of a home run, we average all of the values for the 4186 home runs hit in the 2014 season. To find this average, we first total all of the values of the home runs in Table 9.14, and divide this total by the number of home runs. In Table 9.15, we find the total value in each situation, and show the sum of these totals in the lower right cell of the table. The average value of a home run is then equal to

$$\text{Avg value of home run} = \frac{1142 \times 1 + 257 \times 1.63 + \cdots + 37 \times 3.45}{4186}$$
$$= \frac{5861.98}{4186} = 1.40.$$

This average value of a home run may seem small, since we credit a home run with four bases in the computation of a slugging percentage. But this average of 1.40 runs represents a

typical run value for this type of hit. In the exercises, we will investigate the run values for other types of batting plays.

9.5 Answering Questions About Baseball Strategy

Topics Covered: Evaluating the worth of a play by use of expected runs.

The 2001 World Series between the New York Yankees and the Arizona Diamondbacks is considered one of the most exciting World Series in history. We focus on Game 4 that involved repeated use of a well-known baseball strategy. The question we want to address is whether this strategy really helps in scoring runs.

In the top of the first inning, the lead off hitter for Arizona, Tony Womack, singled to center. Then Craig Counsell, the second batter, was instructed to hit a sacrifice bunt. The bunt was effective—Counsell was thrown out at first and Womack advanced to second base. The inning finished without Womack scoring. In the top of the third inning, the same situation happened. Womack opened the inning with a walk and Counsell again sacrificed with a bunt to move Womack to second. (Again Womack didn't score.) In the top of the 5th inning, Womack started the inning with a double. Counsell again hit a sacrifice bunt, moving Womack to third. The next hitter, Luiz Gonzalez, hit a fly ball, but Womack was thrown out at home plate, ending the inning.

It appears that the Arizona manager, Bob Brenly, likes to play the sacrifice bunt. Counsell was instructed to sacrifice his at-bat (and get an out) in order to advance the runner one additional base. Is the sacrifice bunt a smart play in baseball?

We have the tools to evaluate the effectiveness of this play in creating runs for a team. As in evaluating the home run, we use the run potential matrix that gives the average number of runs scored in each bases/outs situation.

Sacrifice Bunt to Move Runner from First to Second

We first look at the situation that occurred in the first and third innings of the World Series game. Arizona has a runner on first with no outs. Looking at our run expectancy matrix, this situation has a potential of .81 runs. If Counsell hits a successful sacrifice bunt, the batter is out, but the runner moves to second base. The run potential of (runner on 2nd, one out) is .61 runs. So the value of the sacrifice bunt in this situation is

$$\text{VALUE} = 0.61 - 0.81 = 0.20$$

So Arizona really has hurt themselves on the average the team has decreased their run production by .2 runs.

But this calculation is assuming that we are primarily interested in scoring as many runs as possible. Maybe the Arizona manager wants to score one run or more and thinks that he has put the team in a better position to score (that is, get one run or more) using the sacrifice bunt. Using the Markov Chain model, we showed earlier how we could simulate the probability of scoring in the remainder of the inning from each of the 24 starting states. From Table 9.13, we obtain

$$\text{Prob(scoring with runner on 1st and no outs)} = .391,$$

$$\text{Prob(scoring with runner on 2nd and 1 out)} = .380.$$

Even from this perspective, Arizona has decreased their chances of scoring slightly by using the sacrifice bunt.

Let's return to the situation where these sacrifice bunts occurred. Counsell was instructed to sacrifice in the 1st and 3rd innings of the game. In these early innings, it would seem that the goal of a team would be to score as many runs as possible. Also, Craig Counsell is one of the better hitters on the Diamondbacks and he is likely to get a base hit that would greatly increase Arizona's run potential. So it would seem that the sacrifice hit was not an effective strategy in the situations where it was used.

9.6 Exercises

9.0. Rickey Henderson is considered the greatest leadoff hitter, but how often did he actually lead off? Let's focus on Henderson's 273 plate appearances at home games for the 1990 baseball season we record the state of the inning (outs and runners) for each plate appearance. Table 9.15 classifies these 273 plate appearances with respect to the number of outs (0, 1 and 2) and the eight possible runner situations.

Table 9.15. Count of number of plate appearances in different runner and out situations for home games for the 1990 season

	Bases Situation							
	◇	◇	◇	◇	◇	◇	◇	◇
0 outs	109	17	0	0	2	2	0	0
1 out	29	13	8	1	3	2	0	2
2 outs	37	18	11	4	6	5	2	2
Total								273

(a) What fraction of times did Henderson actually bat with the bases empty and no outs? (These were essentially the times when he was a leadoff batter.)

(b) What fraction of times did Henderson bat with exactly one runner on base? With exactly two runners on base? When the bases were loaded?

(c) Suppose that you were able to find a similar classification of plate appearances for Barry Bonds for the 1990 season. Would you expect Bonds to have similar fractions of plate appearances with the bases empty, one runner, two runners, and three runners as Rickey Henderson? Explain.

9.1. A baseball fan has been celebrating the recent success of his team at location D. He takes a random walk down the street hoping to arrive at his home (location H). From location D, he is sure to go to location J in the next minute. From location J, he is equally likely in the next minute to return to D or go ahead to location K. From location K, he is equally likely the nesxt minute to go home (location H) or back to location J. Once he is home, he will remain there with probability 1. Figure 9.3 shows the four locations, arrows to show the possible transitions, and the transition probabilities. A Markov Chain with states D, J, K, H and transition matrix P shown in Table 9.16 can represent this random walk. (Note that H is an absorbing state in the chain.) The matrices P^2, P^3, and P^4 are also shown below.

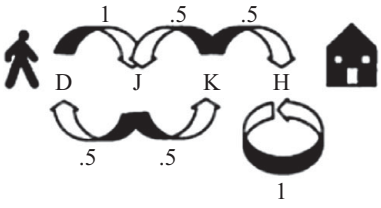


Figure 9.3. States and possible moves in a random walk from location D to home (H).

Table 9.16. Transition probability matrix for the random walk

$$P = \begin{array}{c} \\ \text{D} \\ \text{J} \\ \text{K} \\ \text{H} \end{array} \begin{array}{c} \begin{array}{cccc} \text{D} & \text{J} & \text{K} & \text{H} \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$P^2 = \begin{bmatrix} .5 & 0 & .5 & 0 \\ 0 & .75 & 0 & .25 \\ .25 & 0 & .25 & .50 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad P^3 = \begin{bmatrix} 0 & .75 & 0 & .25 \\ .375 & 0 & .375 & .25 \\ 0 & .375 & 0 & .625 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P^4 = \begin{bmatrix} .375 & 0 & .375 & .25 \\ 0 & .5625 & 0 & .4375 \\ .1875 & 0 & .1875 & .6250 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

(a) One possible path of our baseball fan is DJDJKH. Find the probability of this path using the transition probabilities.
(b) If the fan starts at D, is it possible for our fan to return to D after three minutes? (Recall that each step takes one minute.) Why or why not?
(c) Using the given matrices, find the probability that the fan (starting at D) will arrive home in four minutes.
(d) If the fan starts at J, find the probability that he returns to J in two minutes.
(e) If the fan starts at J, where is his most likely location in three minutes?

9.2. (Exercise 9.1 continued.) Let Q denote the matrix obtained by deleting the last row and last column corresponding to the absorbing state from the transition matrix P. The fundamental matrix is displayed below.

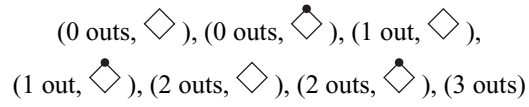
$$E = (I - Q)^{-1} = \begin{bmatrix} 3 & 4 & 2 \\ 2 & 4 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

(a) If the fan starts at D, how many minutes does the typical fan expect to spend at location J before he arrives home?
(b) If the fan starts at J, how many minutes will the typical fan spend, on average, at location D before he gets home?

(c) Starting from D, how many minutes will the fan take, on average, before arriving home?

(d) Is it possible for the fan to take 20 minutes to arrive home (starting from D)? Why or why not?

9.3. Consider the following simplified version of baseball. When a player comes to bat, he gets out with probability .6, he hits a double with probability .3, and he hits a home run with probability .1. No other batting events besides outs, doubles, and home runs are possible. There are only seven possible (outs, bases) situations in this game given by



(a) Fill in the transition probability matrix P below. The impossible transitions are indicated by zeros in the matrix.

Transition matrix P

	0 outs	0 outs	1 out	1 out	2 outs	2 outs	3 outs
	◇	◇	◇	◇	◇	◇	
0 outs, ◇				0	0	0	0
0 outs, ◇				0	0	0	0
1 out, ◇	0	0				0	0
1 out, ◇	0	0			0		0
2 outs, ◇	0	0	0	0			
2 outs, ◇	0	0	0	0			
3 outs	0	0	0	0	0	0	

(b) For each of the possible transitions, find the number of runs scored on the play and place in the table below. The impossible transitions are crossed out in the table.

Runs matrix

	0 outs	0 outs	1 out	1 out	2 outs	2 outs	3 outs
	◇	◇	◇	◇	◇	◇	
0 outs, ◇				xxx	xxx	xxx	xxx
0 outs, ◇				xxx	xxx	xxx	xxx
1 out, ◇	xxx	xxx				xxx	xxx
1 out, ◇	xxx	xxx			xxx		xxx
2 outs, ◇	xxx	xxx	xxx	xxx			
2 outs, ◇	xxx	xxx	xxx	xxx			
3 outs	xxx	xxx	xxx	xxx	xxx	xxx	

9.4. (Exercise 9.3 continued.) Consider a Markov Chain model for the simplified game of baseball with transition matrix found in Exercise 9.3.

(a) Find the two-step probability matrix. Using this matrix, find the probability that there will be a runner on 2nd and one out after two players have batted in the inning.

(b) Compute the expected number of visits matrix E. Using this matrix, find the average number of batters in a half-inning of baseball.

(c) If the current state of the inning is one out with a runner on 2nd, find the expected number of batters in the remainder of the inning.

(d) Find the run potential vector . This vector gives the expected number of runs in the remainder of the inning starting at each possible state.

(e) Suppose that a player comes to bat with a runner on 2nd with one out. He hits a double. Using the run potential vector R, find the value of this play.

9.5. Consider again the run potential matrix shown in Table 9.17 that gives the expected runs in the remainder of the inning for each of the 24 possible outs/bases situations.

Table 9.17. The expected number of runs scored in the remainder of the inning starting at each possible state

	Bases Situation							
								
0 outs	0.44	0.81	1.06	1.29	1.36	1.73	1.90	2.21
1 out	0.22	0.46	0.61	0.90	0.83	1.06	1.33	1.49
2 outs	0.08	0.19	0.28	0.33	0.38	0.40	0.47	0.63

Using this matrix, find the value of the following batting plays.

(a) There are runners on the corners (first and third) with one out. The batter hits a double, scoring both runners.

(b) There is a runner on 1st with no outs. The batter hits a grounder, which is converted to a double play, getting both the runner and the batter out.

(c) The bases are loaded with no outs. The batter hits a grand slam home run. Compare with the value of the grand slam with two outs.

9.6. What is the value of a single when there are runners on 1st and 2nd with no outs? The value of this hit depends on the advancement of the runners. In the 1987 National League, a single occurred in this situation (runners on 1st and 2nd with no outs) a total of 136 times. Table 9.18 shows the three types of run advancement and the count of each type.

Table 9.18. Three types of runner advancement and the count of each type when there is a single with runners on 1st and 2nd and no outs

Starting state	Final state	Count	Runs scored	Value
 , no outs	 , no outs	43		
	 , no outs	35		
	 , no outs	58		

(a) For each final state, compute the number of runs scored on the play. Put these values in the "Runs scored" column of the table.
(b) Using the run potential matrix in Table 9.12 and the runs scored from (a), find the value of each transition. Put the values in the "Value" column of the table.
(c) Use the previous calculations to find the mean value of a single when runners are on 1st and 2nd with no outs.

9.7. Suppose that the batter hits a single with the bases loaded and one out. How important is this play? The value depends on the advancement of the runners. This play occurred 88 times in the 1987 National League. Table 9.19 below shows the possible advancement of the runners and the number of times each type of advancement occurred.

Table 9.19. Five types of runner advancement and the count of each type when there is a single with the bases loaded with one out

Starting state	Final state	Count	Runs scored	Value
 , 1 out	 , 1 out	2		
	 , 1 out	34		
	 , 1 out	14		
	 , 1 out	5		
	 , 1 out	33		

(a) For each final state, compute the number of runs scored on the play. Put these values in the "Runs scored" column of the table.
(b) Using the run potential matrix and the runs scored from (a), find the value of each transition. Put the values in the "Value" column of the table.
(c) Use the previous calculations to find the mean value of a single when the bases are loaded with one out.

9.8. How valuable is a walk? The value of this play depends on the beginning (runners, outs) state. In Table 9.20, the value of the walk is shown for each of the 24 possible states and the number of times that the state occurred (using 2014 MLB data) is displayed in parentheses. For example, we see that there was a walk with the bases empty and no outs

Table 9.20. Values and the number of walks that occur in all possible situations

	Bases Situation							
								
0 outs	0.37	0.56	0.32	0.40	0.97	0.57	0.47	1.00
	(3025)	(566)	(268)	(46)	(136)	(50)	(78)	(32)
1 out	0.24	0.36	0.22	0.16	0.67	0.41	0.15	1.00
	(2214)	(683)	(657)	(247)	(289)	(119)	(287)	(84)
2 outs	0.11	0.19	0.08	0.07	0.25	0.24	0.14	1.00
	(1973)	(791)	(1022)	(402)	(401)	(200)	(312)	(138)

3025 times. The value of this particular play (moving from  , 0 outs to  , 0 outs) is .37 runs.

(a) In what situation(s) is a walk most valuable? When is a walk least valuable?
(b) Explain why a walk has a value of one run when the beginning state is bases loaded .
(c) In what situation is a walk most likely to occur? When is a walk least likely to occur?
(d) Find the average value of a walk over all situations.

9.9. In Game 7 of the 2014 World Series, Hunter Pence and Brandon Belt were starters for the San Francisco Giants. The tables below show how the players did in all of their plate appearances in that particular game. Table 9.21 gives the inning in which the player batted, the before and after game states and the batting play.

Table 9.21. Results of all plate appearances of Hunter Pence and Brandon Belt in Game 7 of the 2014 World Series

Hunter Pence

Inning	Before state	Play	After state	Value
2nd	 , 0 outs	Single	 , 0 outs	
4th	 , 0 outs	Single	 , 0 outs	
6th	 , 0 outs	Groundout double play	 , 2 outs	
8th	 , 2 outs	Groundout	3 outs	

Brandon Belt

Inning	Before state	Play	After state	Value
2nd	 , 0 outs	Single	 , 0 outs	
4th	 , 0 outs	Flyball	 , 1 out	
6th	 , 2 outs	Single	 , 2 outs	
8th	 , 0 outs	Groundout	 , 1 out	

(a) For each batting play of each player, find the value using the run potential matrix in Table 9.12. Record the values in the "Value" column of the tables.
(b) What was the most valuable batting play for each hitter?
(c) Which player had the better batting performance in this particular game? Explain.

9.10. (Value of a sacrifice bunt.) Suppose that there is a runner on 2nd base with no outs. Is it a good play to have the batter hit a sacrifice bunt to move the runner from 2nd to 3rd? (Use the run potential matrix in Table 9.12 in your explanation. Alternately, you can use the matrix that gives the probability of scoring at least one run in all situations.)

9.11. (Value of a steal.) Suppose the first batter in the inning gets a walk. He is thinking about stealing 2nd base.

(a) Suppose that the runner attempts to steal 2nd base and is successful. Find the value of this play. (It should be positive.)

(b) Suppose that the runner attempts to steal 2nd base and is unsuccessful. (The catcher throws him out.) Find the value of this play. (This value should be negative.)
(c) Suppose that this particular runner is successful in stealing 2nd base 70% of the time. Does the team benefit (in terms of runs scored) by having this player steal? Explain.
(d) Let the probability that the runner is successful in stealing 2nd base be equal to p. Find the value of p such that the value of the stealing play is equal to zero. (So if the runner has a success rate that is larger than p, it would benefit the team to have him attempt the steal of 2nd base.)

9.12. Table 9.22 displays data about the changes in pitch count using data from the 2014 season. The column "End" refers to the end of the plate appearance by a ball put in play, a walk, or a strikeout. The row "0-0", this table shows there were 93,466 occurrences of "adding a strike" (to count 0-1), 74,098 occurrences of "adding a ball" (to count 1-0), and 22,438 occurrences to "End." For two-strike outs, like 0-2, 1-2, 2-2, and 3-2, it is possible to transition to the same count. For example, we see there were 8643 transitions from a count of 0-2 to 0-2.

Table 9.22. Number of transitions in pitch count using data from the 2014 season

Count	Add Strike	Add Ball	Same	End
0-0	93466	74098	0	22438
0-1	37710	37850	0	17886
0-2	0	20546	8643	17164
1-0	35860	25000	0	13238
1-1	32424	24731	0	16555
1-2	0	24889	14394	28081
2-0	12257	8081	0	4662
2-1	17010	10459	0	9519
2-2	0	14273	16073	25826
3-0	4567	0	0	3514
3-1	7116	0	0	7910
3-2	0	0	9417	23186

(a) Explain why the transitions in pitch counts can be viewed as a Markov Chain. Give the possible states, and write down several rows of the transition probability matrix P.
(b) Find the probability that the count will be 0-2 after two pitches.
(c) Find the probability the count will be 1-2 after three pitches.
(d) Using matrix operations, find the average number of pitches in a plate appearance.

Further Reading

A good description of the discrete Markov Chain probability model is contained in Kemeny and Snell (1976). Pankin (1987) and Bukiet and Palacios (1997) describe the use of Markov

Chains to model baseball. Lindsey (1963), using actual game data, found the distribution of runs scored in the remainder of the inning starting from each possible bases occupied/outs situation. Lindsey's run production data is used to estimate the value of different types of base hits in Albert and Bennett (2003), Chapter 7.

10.1090/text/034/10

A

An Introduction to Baseball

A.1 The Game of Baseball

Baseball is one of the most popular games in the United States; it is often called the national pastime. The game evolved out of various ball-and-stick games played in many areas of the world, including the Russian game of lapta and the English game of rounders. It became a popular sport in the eastern United States in the mid-1800s. Professional baseball started near the end of the 19th century; the National League was founded in 1876 and the American League in 1900. Currently in the United States, there are 30 professional teams in the American and National Leagues and millions of people watch games in ballparks and on television.

Baseball is a game between two teams of nine players each, played on an enclosed field. A game consists of nine innings. Each inning is divided into two halves; in the top half of the inning, one team plays defense in the field and the second team plays offense, and in the bottom half, the teams reverse roles. The fielding positions for the nine players playing defense are a catcher, first baseman, second baseman, shortstop, third baseman, left fielder, center fielder and right fielder. Figure A.1 shows a diagram of a baseball field and shows the fielding positions for the defensive players. The corners of the diamond shape are the locations of the four bases, home base, first base, second base, and third bases, that play an important part in the game. This figure also shows the location of one offensive player, the batter, at home base.

The team that is batting during a particular half-inning, the offensive team, is trying to score runs. A player from the offensive team begins by batting at home base. A run is the score made by this player who advances from batter to runner and touches first, second, third, and home bases in that order. A team wins a game by scoring more runs than its opponent at the end of nine innings. There are some exceptions to the nine-inning game. A game that is tied after nine innings continues into extra innings until one team has won, and a game may be shortened due to inclement weather.

A basic play in baseball consists of a player on the defensive team, called a pitcher, throwing a spherical ball (called a pitch) toward the batter. The batter is attempting to strike or hit the pitch using a smooth round stick called a bat. After a number of thrown pitches, the batter will either be put out or become a runner on one of the bases. The batter may be put out in several ways:

- He hits a fly ball (a ball in the air) that is caught by one of the fielders.
- He hits a ball in fair territory (explained below) and first base is tagged before the batter reaches first base.
- A third strike (explained below) is caught by the catcher.

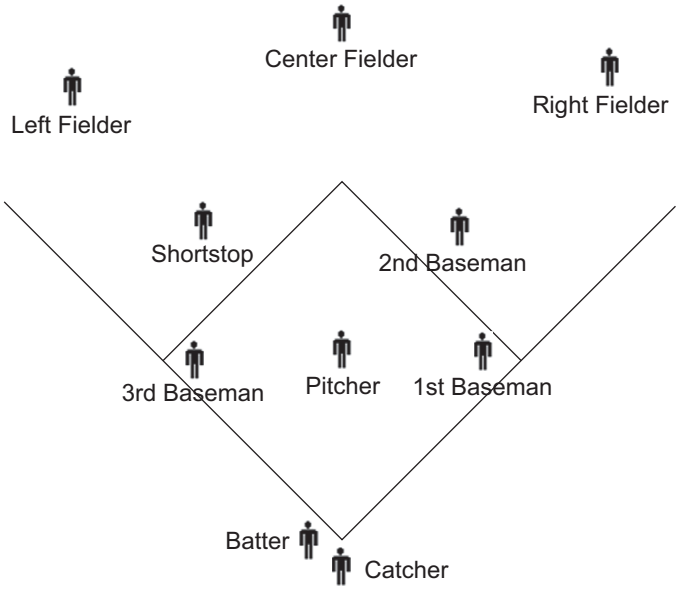


Figure A.1. Diagram of baseball field and bases and location of nine defensive players and batter.

Fair territory is the part of the playing field between the line from home plate to first base and the line from home plate to third base. Foul territory is the region of the field outside of fair territory. A strike is a pitch that is struck at by the batter and missed, or is not struck by the batter and passes through a region called the strike zone. A ball is a pitch that is not struck at by the batter and does not enter the strike zone in flight.

A hitter can advance to a runner and reach base safely by:

- Receiving four pitches that are balls. In this case, the batter receives a walk or base-on-balls and can advance to first base.
- Hitting a ball in fair territory that is not caught by a fielder or thrown to first base before the runner reaches first base. There are different types of hits depending on the advancement of the runner on the play. A single is a hit where the runner reaches first base, a double is a hit where the runner reaches second base, a triple is a hit where a runner reaches third base, and a home run is a big hit (usually over the outfield fence) where the runner advances around all bases safely.

A.2 One Half-Inning of Baseball

In a half-inning of baseball, the nine players on the offensive team will come to bat in sequence. The players will continue to bat in the inning until three outs are made. To get a flavor of how baseball is played, let us revisit the last game played in the 2014 Major League Baseball season. The Kansas City Royals and San Francisco Giants were playing the final game of the World Series – the winner of this game would be declared the best team of the season. We focus on the top half of the 4th inning of the game where the Giants were batting and the game score was tied at 2-2. The pitcher for the Royals at the beginning of the inning was Jeremy Guthrie.

Figure A.2 displays diagrams of the runner and batter situations for each of the six players that came to bat this particular half-inning

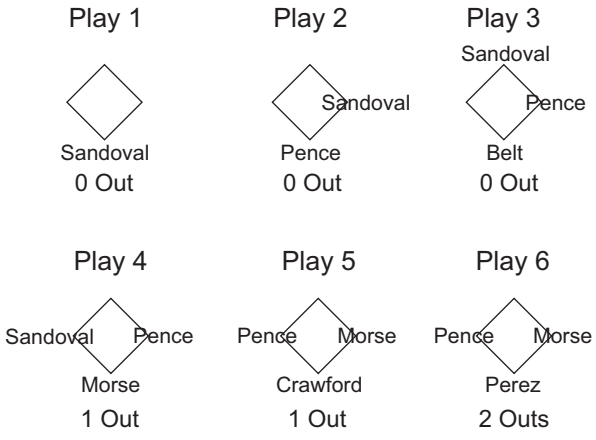


Figure A.2. Batter and runner diagrams for each of six San Francisco Giant players that came to bat in the top of the fourth inning of the final game of the 2014 World Series.

Batter 1: The inning started with no runners on base and no outs. The first batter for the Giants was Pablo Sandoval. Sandoval hit a grounder to second-base for a single. The Giants now have a runner on first base with no outs.

Batter 2: The next Giants batter Hunter Pence singles to center field. Sandoval advances to second base. The situation is now runners on first and second with no outs.

Batter 3: Brandon Belt batted next for the Giants. Belt hits a deep fly that is caught in left field for an out. On the fly ball, Sandoval is able to advance to third base, so the Giants have runners on first and third with one out.

Batter 4: Michael Morse, the next batter, hits a linear to right field for a single. Sandoval scores from third base and Pence advances to third base. Moore is credited with a run batted in as one run scored on the basis of his hit. Now the Giants again have runners on first and third base with one out.

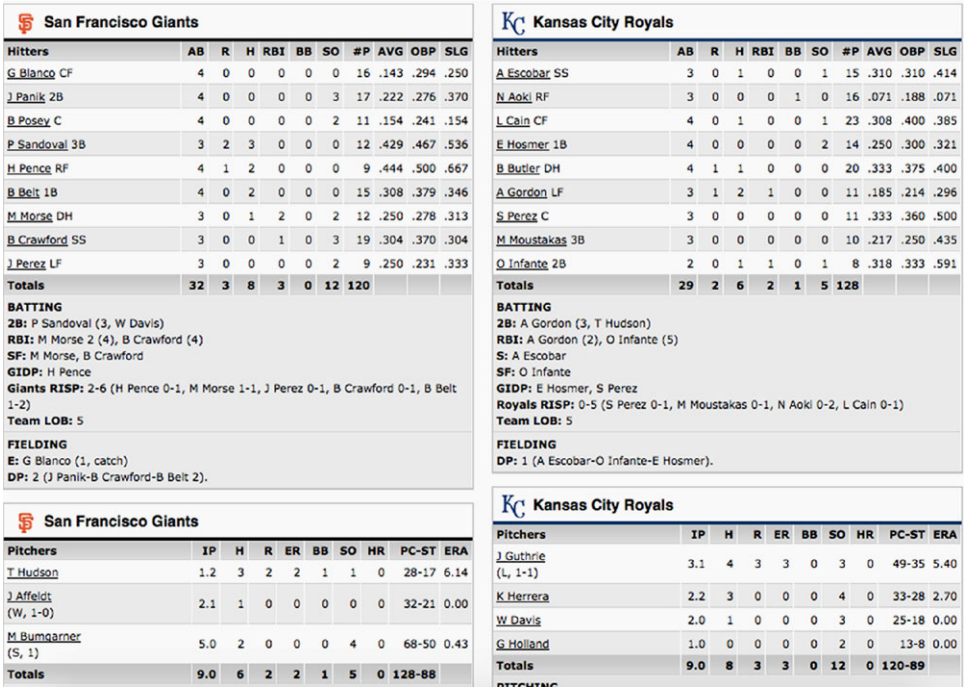
Batter 5: The next hitter, Brandon Crawford, strikes out, and the runners remain on 1st and 3rd with two outs.

Batter 6: Juan Perez, the next batter, hits a grounder to shortstop who throws to first base in time for the third out.

With the run scored in the top of the fourth inning, the Giants took a 3-2 lead that would eventually be the final score of the game—the Giants became the 2014 MLB champions.

A.3 The Boxscore: A Statistical Record of a Baseball Game

One notable aspect of the game of baseball is the wealth of numerical information that is recorded about the game. A boxscore is a statistical record of a particular game. Figure A.3



San Francisco Giants

Hitters	AB	R	H	RBI	BB	SO	#P	AVG	OBP	SLG
G Blanco CF	4	0	0	0	0	0	16	.143	.294	.250
J Panik 2B	4	0	0	0	0	3	17	.222	.276	.370
B Posey C	4	0	0	0	0	2	11	.154	.241	.154
P Sandoval 3B	3	2	3	0	0	0	12	.429	.467	.536
H Pence RF	4	1	2	0	0	0	9	.444	.500	.667
B Belt 1B	4	0	2	0	0	0	15	.308	.379	.346
M Morse DH	3	0	1	2	0	2	12	.250	.278	.313
B Crawford SS	3	0	0	1	0	3	19	.304	.370	.304
J Perez LF	3	0	0	0	0	2	9	.250	.231	.333
Totals	32	3	8	3	0	12	120			

BATTING
2B: P Sandoval (3, W Davis)
RBI: M Morse 2 (4), B Crawford (4)
SF: M Morse, B Crawford
GIDP: H Pence
Giants RISP: 2-6 (H Pence 0-1, M Morse 1-1, J Perez 0-1, B Crawford 0-1, B Belt 1-2)
Team LOB: 5

FIELDING
E: G Blanco (1, catch)
DP: 2 (J Panik-B Crawford-B Belt 2).

Kansas City Royals

Hitters	AB	R	H	RBI	BB	SO	#P	AVG	OBP	SLG
A Escobar SS	3	0	1	0	0	1	15	.310	.310	.414
N Aoki RF	3	0	0	0	1	0	16	.071	.188	.071
L Cain CF	4	0	1	0	0	1	23	.308	.400	.385
E Hosmer 1B	4	0	0	0	0	2	14	.250	.300	.321
B Butler DH	4	1	1	0	0	0	20	.333	.375	.400
A Gordon LF	3	1	2	1	0	0	11	.185	.214	.296
S Perez C	3	0	0	0	0	0	11	.333	.360	.500
M Moustakas 3B	3	0	0	0	0	0	10	.217	.250	.435
O Infante 2B	2	0	1	1	0	1	8	.318	.333	.591
Totals	29	2	6	2	1	5	128			

BATTING
2B: A Gordon (3, T Hudson)
RBI: A Gordon (2), O Infante (5)
S: A Escobar
SF: O Infante
GIDP: E Hosmer, S Perez
Royals RISP: 0-5 (S Perez 0-1, M Moustakas 0-1, N Aoki 0-2, L Cain 0-1)
Team LOB: 5

FIELDING
DP: 1 (A Escobar-O Infante-E Hosmer).

San Francisco Giants

Pitchers	IP	H	R	ER	BB	SO	HR	PC-ST	ERA
T Hudson	1.2	3	2	2	1	1	0	28-17	6.14
J Affeldt (W, 1-0)	2.1	1	0	0	0	0	0	32-21	0.00
M Bumgarner (S, 1)	5.0	2	0	0	0	4	0	68-50	0.43
Totals	9.0	6	2	2	1	5	0	128-88	

Kansas City Royals

Pitchers	IP	H	R	ER	BB	SO	HR	PC-ST	ERA
J Guthrie (L, 1-1)	3.1	4	3	3	0	3	0	49-35	5.40
K Herrera	2.2	3	0	0	0	4	0	33-28	2.70
W Davis	2.0	1	0	0	0	3	0	25-18	0.00
G Holland	1.0	0	0	0	0	2	0	13-8	0.00
Totals	9.0	8	3	3	0	12	0	120-89	

Figure A.3. Boxscore of Game 7 of the 2014 World Series between the Giants and the Royals.

displays a boxscore for the last game in the 2014 World Series. We will define particular baseball events and the associated notation by using this boxscore.

When a player comes to bat during an inning, he is making a plate appearance (PA). What can happen during this plate appearance? The batter can get a hit (H) and there are four possible hits: a single (1B), a double (2B), a triple (3B), and a home run (HR). The batter may get a walk (base-on-balls abbreviated BB) by receiving four pitched ball—she advances to first base. Also, the batter can advance to first base when he is hit by a pitch (HBP). The player might create an out. Some outs like a sacrifice bunt (SH) and a sacrifice bunt (SB) advance runners on base. Finally, the player might reach base by an error by a fielder (E).

An official at-bat (AB) is a plate appearance excluding walks, hit-by-pitches, sacrifice flies and sacrifice hits. The top half of the boxscore lists all of the batters for both teams. For each player, the boxscore first gives his fielding position. For example, we see that Blanco was the center fielder (CF) for the Giants in this game. Then the boxscore lists

- AB: the number of at-bats of the player in the game,
- R: the number of runs scored by the player,
- H: the number of hits by the player,
- RBI: the number of runs batted in by the player,
- BB: the number of walks by the player,
- SO: the number of strikeouts by the player,
- # P: the number of pitches received by the player,
- AVG: the World Series batting average of the player

- OBP: the WS on-base percentage of the player,
- SLG: the WS slugging percentage of the player.

Under the basic batting table, the boxscore lists some special events not included in the table. Under BATTING, the boxscore lists the players who hit doubles (2B), triples (3B) and home runs (HR). In this listing, the inning in which the hit occurred and the opposing pitcher are recorded. So the listing

2B P. Sandoval (3, W. Davis)

means that Pedro Sandoval hit a double in the third inning against Wade Davis. In addition, this section gives specific information about RBI, SF, GIDP (ground out in a double play), and RISP. The Giants were 2-6 in RISP which indicates that two of the six runners in scoring position (either 2nd or 3rd base) eventually scored.

Below the batting tables is a table of statistics for the pitchers in the game. For each pitcher, the table gives the number of innings pitched (IP), the number of hits (H) and runs (R) allowed. Next it shows the number of earned runs (ER) allowed—these are runs allowed by the pitcher not due to errors by the fielders of the team. Next the table gives the number of walks (BB) and strikeouts (SO) and home runs (HR) allowed by each pitcher. The table gives the number of pitches (PC) and the number of strikes thrown (ST). The last number, the earned run average (ERA), gives the average number of earned runs allowed by this pitcher for nine innings for all games played in the series.

In addition to this batting and pitching information, the boxscore gives a line summary of the runs and hits scored in the game.

```
-------------------------------------------------
    Giants     - 020 100 110  --  3
    Royals     - 020 000 000  --  2
-------------------------------------------------
```

Each column of numbers corresponds to the number of runs scored by the two teams in a given inning. We see that the Giants and Royals scored two runs in the second inning, and the Giants scored one run in the top of the fourth that become the winning run in the game. Last, the boxscore gives some other information about the game (not shown in Figure A.3). This includes the names and positions of the umpires, the elapsed time (T) of the game, the ballpark attendance, and some data on the weather during the game.

Bibliography

[1] Albert, Jim (2008), "Streaky hitting in baseball," *Journal of Quantitative Analysis in Sports*, 4, 1.

[2] _____ (1998), "Sabermetrics." In *Encyclopedia of Statistical Sciences* (edited by S. Kotz, C. B. Read and D. L. Banks). New York: John Wiley.

[3] Albert, James and Rossman, Allan (2001), *Workshop Statistics: Discovery with Data, A Bayesian Approach*, Emeryville, CA: Key College.

[4] Albert, Jim and Bennett, Jay (2003), *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*, Springer, New York.

[5] Bennett, Jay (1998), "Baseball." In *Statistics in Sport* (edited by Jay Bennett), London; New York: Arnold Publishers.

[6] Berry, Donald A. (1996), *Statistics: A Bayesian Perspective,* Belmont, CA: Wadsworth Publishing.

[7] Bukiet, Bruce, Harold, Elliotte, and Palacios, Jose (1997), "A Markov Chain Approach to Baseball," *Operations Research*, Vol. 45, No. 1, pp. 14–23.

[8] Cover, Thomas M. and Keilers, Carroll, W. (1977), "An Offensive Earned Run Average for Baseball," *Operations Research*, 25, pp. 729–740.

[9] D'Esopo, D. A. and Lefkowitz, B. (1977), "The Distribution of Runs in the Game of Baseball." In *Optimal Strategies in Sports* (edited by S. P. Ladany and R. E. Machol), pp. 55–62. New York: North-Holland.

[10] Devore, Jay and Peck, Roxy (2011), *Statistics: The Exploration and Analysis of Data*, Pacific Grove, CA: Brooks/Cole.

[11] Gilovich, Thomas, Vallone, Robert, and Tversky, Amos (1985), "The hot hand in basketball: On the misperception of random sequences," *Cognitive Psychology*, 17, 295–314.

[12] James, Bill (1982), *The Bill James Baseball Abstract,* New York: Ballantine Books.

[13] _____ (1997), *The Bill James Guide to Baseball Managers,* New York: Scribners.

[14] _____ (2001), *The New Bill James Historical Baseball Abstract,* New York: The Free Press.

[15] Katz, Stanley M., "Study of 'The Count," *1986 Baseball Research Journal* (#15), pp. 67–72.

[16] Kemeny, John G. and Snell, J. Laurie (1976). *Finite Markov Chains,* New York: Springer-Verlag.

[17] Lindsey, George R. (1963), "An Investigation of Strategies in Baseball," *Operations Research*, vol. 11, pp. 477–501.

[18] *Major League Handbook* (2001), Lincolnwood, IL: Sports Team Analysis & Tracking Systems.

[19] Marchi, Max and Albert, Jim (2013), *Analyzing Baseball Data with R*, Chapman and Hall.

[20] Moore, David, McCabe, George and Craig, Bruce (2012), *Introduction to the Practice of Statistics (4th edition),* New York: W. H. Freeman Company.

[21] Mosteller, Frederick (1952), "The World Series Competition," *Journal of the American Statistical Association,* 47, 355–380.

[22] Neft, David S., and Cohen, Richard M. (1997), *The Sports Encyclopedia: Baseball,* New York: St. Martin's Press.

[23] Pankin, Mark D. (1987), "Baseball as a Markov Chain" In *The Great American Baseball Stat Book*, pp. 520–524.

[24] R Core Team (2015), *R: A Language and Environment for Statistical Computing*, Vienna, Austria, https://www.R-project.org/.
[25] Rothman, Stanley (2012), *Sandlot Stats: Learning Statistics with Baseball*, John Hopkins University Press.
[26] Schell, Michael (1999), *Baseball's All-Time Best Hitters: How Statistics Can Level the Playing Field,* Princeton, NJ: Princeton University Press.
[27] Scheaffer, Richard L. and Young, Linda J. (2009), *Introduction to Probability and its Applications*, Duxbury Press.
[28] StatCrunch (2015), Pearson Education.
[29] Tabor, Josh and Franklin, Chris (2011), *Statistical Reasoning in Sports*, New York: W. H. Freeman Company.
[30] Thorn, John and Palmer, Palmer (1985), *The Hidden Game of Baseball,* New York: Doubleday.
[31] ——— (eds) (1989), *Total Baseball*. New York: Warner Books.

Index

AMS / MAA | TEXTBOOKS

Teaching Statistics Using Baseball is a collection of case studies and exercises applying statistical and probabilistic thinking to the game of baseball. Baseball is the most statistical of all sports since players are identified and evaluated by their corresponding hitting and pitching statistics. There is an active effort by people in the baseball community to learn more about baseball performance and strategy by the use of statistics. This book illustrates basic methods of data analysis and probability models by means of baseball statistics collected on players and teams. Students often have difficulty learning statistics ideas since they are explained using examples that are foreign to the students. The idea of the book is to describe statistical thinking in a context (that is, baseball) that will be familiar and interesting to students.

The book is organized using a same structure as most introductory statistics texts. There are chapters on the analysis on a single batch of data, followed with chapters on comparing batches of data and relationships. There are chapters on probability models and on statistical inference. The book can be used as the framework for a one-semester introductory statistics class focused on baseball or sports. This type of class has been taught at Bowling Green State University. It may be very suitable for a statistics class for students with sports-related majors, such as sports management or sports medicine. Alternately, the book can be used as a resource for instructors who wish to infuse their present course in probability or statistics with applications from baseball.

The second edition of *Teaching Statistics* follows the same structure as the first edition, where the case studies and exercises have been replaced by modern players and teams, and the new types of baseball data from the PitchFX system and fangraphs.com are incorporated into the text.