

Can Baseball Be Used to Teach Statistics?

Reviewed by Mason A. Porter

Teaching Statistics Using Baseball

Jim Albert

Mathematical Association of America, 2003

US\$56.95, 304 pages

ISBN-13:978-0883857274

“Statistics are the lifeblood of baseball. In no other sport are so many available and studied so assiduously by participants and fans. Much of the game’s appeal, as a conversation piece, lies in the opportunity the fan gets to back up opinions and arguments with convincing figures, and it is entirely possible that more American boys have mastered long division by dealing with batting averages than in any other way.”

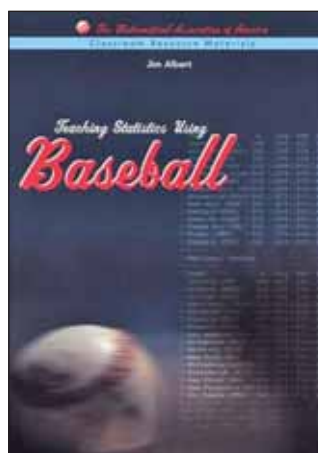
—Leonard Koppett, *A Thinking Man’s Guide to Baseball*, 1967 [7]

Starting Lineups

Much of my initial inclination towards mathematics arose from my passionate interest in baseball. As the above quote by Hall of Fame sportswriter Leonard Koppett indicates, I am hardly the only person who has discovered mathematics and statistics through such means.¹ It probably ran a bit deeper in my case than in most—as a child, I actually used to search for errors on the backs of baseball cards just so I could send letters to the companies that produced them to ask them about their errors. (On one occasion, one of them actually wrote back to inform me that the error was in one

Mason A. Porter is a University Lecturer in the Oxford Centre for Industrial and Applied Mathematics at the University of Oxford and a Tutorial Fellow of Somerville College. His email address is porterm@maths.ox.ac.uk.

¹More recent editions of Koppett’s book [7] have more gender-neutral titles, but I wanted to cite the original version to give some indication of how long baseball and statistics have been married.



of the counts rather than in the computation.) It should thus come as no surprise that I am fascinated by the idea of using baseball to help teach subjects like statistics and probability.

Teaching Statistics Using Baseball by Jim Albert [1] is certainly founded on a solid idea: Given the long marriage between statistics and base-

ball, the wealth of baseball fans in the U.S., and the desperate need to develop clever ways to ensure that students learn statistics, why not use baseball to help teach the subject?

Albert’s intended audience is students with little mathematics background—e.g., his book does not assume any knowledge of calculus—and he has used it as a textbook for a course that gives a gentle introduction to statistics for students with an established interest in baseball. I am skeptical about using Albert’s book as the main text for a course, as I think that it would work only for a very specific niche audience. This is indicated explicitly in the book’s preface, though I believe that this niche is significantly smaller than the author seems to think. Albert’s book offers much better value as a source for exercises and project ideas for statistics and modeling courses with a broader scope.² (It might also be nice for self-study.)

²Indeed, it offers enough value that my review shouldn’t make anybody reminisce too much about the time that Tommy Lasorda gave his opinion of Dave Kingman’s performance.

Albert's book is intentionally introductory, but I think that it is too gentle—to the point of coming across as condescending. (For example, do the students who are going to use this book really need to be informed that the square A^2 of a matrix A is equal to $A \times A$ and then immediately be shown the same level of detail for the cube of a matrix? I think that college students ought to be given more credit than this!) My own introductory formal experience in statistics was as a teacher's assistant in a rigorous, calculus-based course taught by Gary Lorden at Caltech using portions of the book by Ross [10]. (I had also previously taken courses in probability.) Although Albert's book is intended for an audience with considerably less preparation, I nevertheless think the book's hand-holding approach goes too far and actually damages the facility of weaker students to learn the material. Several important concepts, such as correlation, are used without *ever* being defined with mathematical formulas. Others, such as standard deviation, are mentioned in the exercises without any definition or description until one or more chapters later. Albert also seems to eschew the idea of developing concepts further with the exercises, which essentially ask the same questions over and over again. A smattering of more advanced exercises, discussions, and appendices would have been nice, and I don't think that the introductory nature of the book would have been ruined one bit by including such components. There are also flaws with the baseball discussion, including an incomplete definition of a balk that includes only one of the ways that a balk can occur [4]. Frustratingly, Albert often includes raw speculation in his baseball-related conclusions in a book that specifically introduces tools that are meant to be used to tackle such questions more seriously. The book would also have benefited from including pointers to a wider variety of supplementary reading material.

Albert's book does cover a good selection of topics—including data batches, standardization, relationships between measurement variables, probability distributions, statistical inference, Markov chain modeling, etc. It also has a nice accompanying website that includes helpful tidbits such as solutions to odd-numbered exercises and *vital* information such as the data sets that are necessary for the end-of-chapter exercises. The introductory chapter presents a good roadmap for the organization of the remainder of the book and the format of the subsequent chapters. This chapter also presents some basic baseball statistics to get students warmed up and indicates some online resources (especially retrosheet.org and baseball-reference.com) that are indispensable sources for baseball statistics and research. Each of the subsequent chapters begins with a summary of the ensuing contents and then proceeds through several case studies, numerous exercises, and (in

my opinion, unsatisfactorily minimal) suggestions of other reading materials. The first exercise (or two) in each chapter is a “leadoff exercise” about Rickey Henderson, though I was unable to find substantive differences between these exercises and the regular ones. I think that the book's case-study approach has the potential to be extremely valuable. Unfortunately, the variable quality of the case studies curtails their utility. A few of them are legitimately fantastic, as they cover salient topics—such as whether hitters can have inherent differences in their ability to be streaky—that remain active areas of modern baseball research that students can already investigate in an introductory course.³ Other case studies don't have any meat, and several of them contain material that would be better placed as intermediate discussions between case studies.

Unfortunately, Albert's book has some notable and unfortunate gaps. Despite the book's premise, it inexplicably has almost no discussion of sabermetrics, the quantitative study of baseball [4]. I find it astounding, for example, that the book develops interesting considerations similar to concepts from sabermetrics such as Runs to End of Inning (RUE)—which attempts to give a value to each outcome of a plate appearance based on the number of runs expected to score before the inning ends—without even a mention of any of the modern sabermetric statistics (or “Jamesian” statistics, as I like to call them) [5, 11–13]. Instead, Albert seems to be satisfied with mentioning OPS (on-base plus slugging percentage), which was already commonplace by the time his book was published. Defensive statistics, which have now become much more sophisticated than they were for decades precisely because of statistical research, were *never* examined even once in the book. Also omitted were important statistical topics such as Monte Carlo simulations, which are now used in the Diamond Mind simulations of baseball seasons [14]. A brief discussion with pointers to appropriate references that build on germane concepts discussed in the book would have been welcome, and such topics could have at least been introduced in the exercises at the end of each chapter. I think that Albert missed a wonderful teaching opportunity by not doing this.

Play by Play

In this section, I'll give some more details about the topics covered in the book. These topics are introduced in a reasonable order, although I am more familiar with expositions in which concepts

³Some of the case studies—including one that poses the question of whether or not Roger Clemens deserved his Cy Young Award in 2001—are particularly meaningful to me, as they remind me of specific, passionate arguments in which I have participated.

from basic probability are introduced earlier in the game.

Chapter 2 is concerned with exploring single batches of baseball data. Stem plots, distributions, and five-number summaries (smallest value, lower quartile, median, upper quartile, and largest value) are introduced in a case study about teams' offensive statistics. Time series, fitting, histograms, dot plots, and comparisons of distributions are then introduced in subsequent case studies. One of my favorite moments in reading the book was going through Case Study 2.3 about Roger Clemens because after illustrating his miraculously increased strikeout rate at an advanced stage of his career, it contains the following lovely statement: *It is not clear if this pattern in the plot corresponds to any change in Roger's physical condition or his style of pitching, but it may deserve further investigation.* Given the strong evidence—revealed years after the release of Albert's book—that Clemens used steroids during this period of his career, Albert's comment seems to be a rather prescient statement indeed. (For similar reasons, I was also amused by Case Study 3.1, which compares the hitting statistics of Ken Griffey Jr. and Barry Bonds.) This also underscores the importance of statistical analysis in finding apparent anomalies that might suggest something interesting that awaits discovery and is a point that can be made to students who are learning introductory statistics from this book.

Chapter 3 discusses standardization and the comparison of data batches. In addition to the application of topics from Chapter 2, the author introduces new topics such as box plots, mean, standard deviation, the normal distribution, and standardized (z) score. I especially liked Case Study 3.5, in which Albert examines the relative greatnesses of high batting averages in different seasons. I have a quibble, however, that first shows up in this chapter and is then repeated elsewhere—namely, that the author claims that certain skewed distributions (such as slugging percentages in 1999 in Case Study 3.4) are normal when perhaps that is not a good enough model. I think that the idea of skewness should have been discussed at appropriate points in the book.

Chapter 4 discusses the relationships between measurement variables. The new topics introduced include scatter plots, correlation, linear and non-linear regression, root mean square error, least-squares fitting, and residuals. I especially like Case Study 4.4, which concerns the creation of a new measure of offensive performance using multiple regression, and Case Study 4.6, which examines regression to the mean in player performance (which sabermetricians and sabermetrics-friendly baseball writers such as Rob Neyer stress repeatedly in their analyses and outcome predictions).

Chapter 5 gives an introduction to probability using tabletop games such as *Strat-O-Matic*

Baseball. The new topics introduced in this chapter include the relative frequency interpretation of probability, the law of large numbers, sample spaces, finding probabilities of events, randomization devices, multinomial experiments, and conditional probability. Although tabletop games provide a reasonable way to introduce fundamental concepts from probability, in 2003 (when the book was published) this comes across as rather anachronistic. I suspect that almost none of the students who might use Albert's book have ever played any of these games, and I wonder if that might prove problematic when teaching from this chapter. Additionally, some aspects of the chapter aren't well thought out; for example, it is silly to have exercises that ask students to actually construct spinners to represent events of different probability. I think that college students—even ones who are learning basic statistics as part of their general education—should be given more credit than that.

Chapter 6 discusses probability distributions and baseball. The topics that it covers include binomial and multinomial distributions, independence, expected counts, simulation, and Pearson residuals. Case Study 6.3's discussion of modeling runs scored made me think of the RUE measure employed by sabermetricians (see my earlier discussion).

Chapter 7 provides an introduction to statistical inference. The topics it covers include the distinction between ability and performance, modeling of ability and simulating the data produced by such a model, Bayes' rule, finding the most likely ability for a given performance, interval estimates, and subjective interpretation of probability. I really like Case Studies 7.2 and 7.3, which concern the simulation of a batter's performance if ability is known (7.2) and then attempting to learn a batter's ability based on performance (7.3). I also like Case Study 7.5, which compares the hitting performances of Wade Boggs and Tony Gwynn, because it entails the use of quantitative analysis to improve the level of sophistication of familiar water-cooler arguments.

Chapter 8 discusses topics in statistical inference. It builds on ideas introduced in previous chapters and also covers topics such as situational hitting data, goodness of fit, models with bias and/or ability effects, streakiness (and runs in the statistical sense), and moving averages. A couple of the case studies are concerned with ideas and baseball players that particularly interest me. For example, Case Study 8.1 discusses the situational hitting statistics of Todd Helton, who plays his home games in Coors Field, and offers a well-known (to baseball fans) situation in which there is a large disparity between home and road hitting statistics. (This disparity is a major issue when trying to discern Helton's ranking among the all-time greatest hitters [11].) Case Study 8.5

asks whether John Olerud (who is one of my favorite players) is streaky, and the subject of hitting streaks (and the existence or nonexistence of “hot streaks” more generally) is of course an active area of research [2, 8].

Chapter 9 discusses modeling baseball using Markov chains. The topics it covers include transition probabilities, absorbing states, matrices, matrix multiplication and inversion, computation of event probabilities using simulation, expected numbers of events, and values of batting events. As one might guess from my comments in the introduction, I like Case Studies 9.3 and 9.4, as the former includes a discussion of the expected numbers of runs in the remainder of an inning (though without the discussion of such quantities that have been studied by sabermetricians) and the latter discusses the values of different on-base events. Case Study 9.5 is also very interesting, as it discusses a frequently debated topic: the worth (or lack thereof) of sacrifice bunts.

Albert’s book also includes two appendices: Appendix A provides an introduction to baseball, though I think that any reader who requires it likely won’t be very interested in learning statistics using this book; and Appendix B discusses the data sets used in the book and more generally how to obtain baseball data online. Appendix B is crucial to teaching statistics using this book, because of course students need to possess data in a usable format to do numerical computations. The website retrosheet.org is mentioned prominently, and it remains an essential tool for everybody who does research using baseball statistics. One of my favorite sites, baseball-reference.com (the brainchild of former mathematics professor Sean Forman), is also mentioned prominently, and it is now much more expansive than it was in 2003 when Albert’s book came out. (In particular, the site now has search features that can be used in conjunction with homework problems in a course that uses Albert’s book.) Crucially, Albert’s book has an accompanying website that has errata, solutions to the odd-numbered exercises, and (most importantly) data sets that go with the book’s exercises. Appendix B also includes a welcome discussion of data format and formatting, although I am admittedly a bit perturbed by Albert’s characterization of MATLAB as his “favorite graphing package”. (Given that I am an applied mathematician with a healthy interest in computation, such a pithy description of this wonderful computational tool gives me the shakes.) Unfortunately, Albert’s book fails to mention any of the numerous sabermetrics blogs and other websites, which were already prominent when the book came out and have continued to multiply during the past few years. Appendix B would have been a logical place to discuss the ones available in 2003. Some of the very nice ones currently available are Baseball

Think Factory (baseballthinkfactory.org/), The Hardball Times (hardballtimes.com), and a blog associated with *The Book: Playing the Percentages in Baseball* (insidethebook.com/ee/).

The Final Score

In concluding this review, I should probably answer the question I posed in my title. Statistics is of course crucial to baseball, but that does not imply that one can make baseball the central point of an effective introduction to the subject. I think that Albert’s book can provide a successful starting point for students who are not very mathematically inclined, but it seems to me that many students will become frustrated by the book’s overabundance of hand-holding and lack of mathematical meat. (Where are the formulas and advanced exercises that introduce new concepts, and why can’t at least some of the derivations be included in appendices?) Teaching an introductory statistics course using only baseball (via this book, for example) reduces the audience to a relatively small niche of people who are already interested in baseball and either want or (more likely) are forced to learn some statistics. The book does attempt to introduce baseball to people who aren’t already baseball fans, but I just can’t see how that’s going to be effective. I think that most people who don’t already like baseball are simply not going to appreciate learning statistics from this book, and I would also argue that being a baseball fan is a necessary but not sufficient condition to be a member of the book’s target audience. On the other hand, this book has interesting exercises and case studies that can be used effectively to develop exercises and projects in statistics courses with a broader scope. In my opinion, this is by far the most valuable usage of Albert’s book.

Mathematical and statistical research on baseball continues to flourish, and many interesting insights have been published since this book appeared. Tools such as Diamond Mind and CHONE are founded on ideas that rely on fundamentals that are discussed in Albert’s book. (In fact, a discussion of Diamond Mind might be rather appropriate for Chapter 5, and a discussion of CHONE could be added when regression is introduced.) Obviously, there are also numerous interesting academic and sabermetric studies. For example, Samuel Arbesman and Steve Strogatz recently used Monte Carlo simulations to examine the likelihood of Joe DiMaggio’s 56-game hitting streak (as well as which players were most likely to have such a streak, its length, and when it occurred) [2]. Trent McCotter built on this work and used permutation tests to conclude that hitting streaks are not mere byproducts of randomness [8]. Other researchers have borrowed tools from statistical physics and network analysis to study baseball problems [3, 9]. For instance, my collaborators and I recently

examined the properties of the bipartite batter-pitcher matchup graphs from 1954–2008 for both individual seasons and the entire time period [11]. We looked at random walks on these graphs to develop rank orderings of hitters and pitchers using the head-to-head matchup as the fundamental quantity of interest, and we found interesting connections between the rankings and structural network properties (such as betweenness centrality) of baseball players. Naturally, the sabermetrics community also continues to produce numerous interesting insights. The possibilities for analysis have expanded even further with an abundance of new data, such as extractions from *Pitch f/x*, which can be used to determine the speed of a thrown baseball at its release time and the time history of the baseball's location (up to an accuracy of one inch) during its flight to the batter (see, e.g., [6] for one of myriad examples of research that uses *Pitch f/x*).

Despite my reservations about Albert's book, I need to give him a lot of credit: Reviewing his book has compelled me to think much more deeply about teaching statistics than I ever did before (whether or not one chooses to do so using baseball), and his book definitely merits a perusal by people in the mathematics teaching profession. I know that my students at Oxford would be turned off by Albert's overly gentle approach, but I do believe that a market exists for his book. However, I don't think that it's a good textbook and feel it would serve better as a source of interesting exercises and projects for broader-minded statistics courses.

Acknowledgements

I thank Joe Blitzstein, Sean Forman, Allyn Jackson, Trent McCotter, Tom MacCarone, Jesús Rodríguez, Serguei Saavedra, and Steve Strogatz for useful comments.


References

- [1] J. ALBERT, *Teaching Statistics Using Baseball*, Mathematical Association of America, Washington, DC, 2003.
- [2] S. ARBESMAN and S. H. STROGATZ, *A Monte Carlo approach to Joe DiMaggio and streaks in baseball*, arXiv:0807.5082, 2008.
- [3] E. BEN-NAIM, F. VAZQUEZ, and S. REDNER, Parity and predictability of competitions, *Journal of Quantitative Analysis in Sports* 2 (2006), 1.
- [4] P. DICKSON, *The Dickson Baseball Dictionary, Third Edition*, W. W. Norton & Company, New York, NY, 2009.
- [5] B. JAMES, *Win Shares*, STATS Publishing Inc., Northbrook, IL, 2002.
- [6] J. KALK, *A first look at the 2008 Pitch f/x data*, <http://www.hardballtimes.com/main/article/a-first-look-at-the-2008-pitchf-x-data/>, 2008.
- [7] L. KOPPETT, *A Thinking Man's Guide to Baseball*, Dutton, New York, NY, 1967.
- [8] T. MCCOTTER, Hitting streaks don't obey your rules: Evidence that hitting streaks aren't just by-products

of random variations, *The Baseball Research Journal* 37 (2009), 62–70.

- [9] A. M. PETERSEN, W.-S. JUNG, and H. E. STANLEY, On the distribution of career longevity and the evolution of home run prowess in professional baseball, *Europhysics Letters* 83 (2008), 50010.
- [10] S. M. ROSS, *Introduction to Probability and Statistics for Engineers and Scientists*, John Wiley & Sons, New York, NY, 1987.
- [11] S. SAAVEDRA, S. POWERS, T. MCCOTTER, M. A. PORTER, and P. J. MUCHA, Mutually antagonistic interactions in baseball networks, *Physica A* 389 (2010), 1131–1141.
- [12] T. TANGO, M. LICHTMAN, and A. DOLPHIN, *The Book: Playing the Percentages in Baseball*, Potomac Books, Dulles, VA, 2007.
- [13] J. THORN and P. PALMER, *The Hidden Game of Baseball: A Revolutionary Approach to Baseball and Its Statistics*, Doubleday Books, New York, NY, 1984.
- [14] THE HARDBALL TIMES WRITERS, *The Hardball Times Baseball Annual 2009*, ACTA Sports, Skokie, IL, 2008.

CENTRE INTERFACULTAIRE BERNOULLI
CIB


ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

CALL FOR PROPOSALS

The Bernoulli Center (CIB), funded jointly by the Swiss National Science Foundation and the Swiss Federal Institute of Technology in Lausanne, has started its activity in March 2002.

Its mission is to support research in mathematics and its applications, to organize and host thematic programs, to provide a supportive and stimulating environment for researchers, and to launch and foster collaborations between mathematicians working in different areas as well as mathematicians and other scientists.

The CIB launches a call for proposals of four one-semester programs during the **period July 1, 2012 - June 30, 2014**. A thematic program consists of a six months period (January 1 - June 30 or July 1 - December 31) of concentrated activity in a specific area of current research interest in the mathematical sciences. In exceptional cases, one year and three month programs will also be considered.

Those who are interested in organizing a program at the CIB should submit a **two page letter of intent by December 1, 2010**. This letter should give the names of the organizers, of the potential visitors, and outline the program. For more details see <http://cib.epfl.ch/>

École Polytechnique Fédérale de Lausanne
 Centre Interfacultaire Bernoulli
 SB CIB-GE
 AAC034 (Bâtiment AAC)
 Station 15
 CH-1015 Lausanne