Chapman & Hall/CRC

Handbooks of Modern Statistical Methods

Handbook of Statistical Methods and Analyses in Sports

Edited by

Jim Albert
Mark E. Glickman
Tim B. Swartz
Ruud H. Koning

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC

Handbooks of Modern Statistical Methods

Handbook of Statistical Methods and Analyses in Sports

Edited by

Jim Albert
Bowling Green State University, Ohio, USA

Mark E. Glickman
Harvard University, Cambridge, Massachusetts, USA

Tim B. Swartz
Simon Fraser University, Burnaby, British Columbia, Canada

Ruud H. Koning
University of Groningen, The Netherlands

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **Informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20161109

International Standard Book Number-13: 978-1-4987-3736-4 (Hardback)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Handbook of Statistical Methods and Analyses in Sports

Chapman & Hall/CRC Handbooks of Modern Statistical Methods

Series Editor

Garrett Fitzmaurice
Department of Biostatistics
Harvard School of Public Health
Boston, MA, U.S.A.

Aims and Scope

The objective of the series is to provide high-quality volumes covering the state-of-the-art in the theory and applications of statistical methodology. The books in the series are thoroughly edited and present comprehensive, coherent, and unified summaries of specific methodological topics from statistics. The chapters are written by the leading researchers in the field, and present a good balance of theory and application through a synthesis of the key methodological developments and examples and case studies using real data.

The scope of the series is wide, covering topics of statistical methodology that are well developed and find application in a range of scientific disciplines. The volumes are primarily of interest to researchers and graduate students from statistics and biostatistics, but also appeal to scientists from fields where the methodology is applied to real problems, including medical research, epidemiology and public health, engineering, biological science, environmental science, and the social sciences.

Published Titles

Handbook of Mixed Membership Models and Their Applications
Edited by Edoardo M. Airoldi, David M. Blei,
Elena A. Erosheva, and Stephen E. Fienberg

Handbook of Statistical Methods and Analyses in Sports
Edited by Jim Albert, Mark E. Glickman,
Tim B. Swartz, and Ruud H. Koning

Handbook of Markov Chain Monte Carlo
Edited by Steve Brooks, Andrew Gelman,
Galin L. Jones, and Xiao-Li Meng

Handbook of Big Data
Edited by Peter Bühlmann, Petros Drineas,
Michael Kane, and Mark van der Laan

Published Titles Continued

Handbook of Discrete-Valued Time Series
Edited by Richard A. Davis, Scott H. Holan, Robert Lund, and Nalini Ravishanker

Handbook of Design and Analysis of Experiments
Edited by Angela Dean, Max Morris, John Stufken, and Derek Bingham

Longitudinal Data Analysis
Edited by Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs

Handbook of Spatial Statistics
Edited by Alan E. Gelfand, Peter J. Diggle, Montserrat Fuentes, and Peter Guttorp

Handbook of Cluster Analysis
Edited by Christian Hennig, Marina Meila, Fionn Murtagh, and Roberto Rocci

Handbook of Survival Analysis
Edited by John P. Klein, Hans C. van Houwelingen, Joseph G. Ibrahim, and Thomas H. Scheike

Handbook of Spatial Epidemiology
Edited by Andrew B. Lawson, Sudipto Banerjee, Robert P. Haining, and María Dolores Ugarte

Handbook of Missing Data Methodology
Edited by Geert Molenberghs, Garrett Fitzmaurice, Michael G. Kenward, Anastasios Tsiatis, and Geert Verbeke

Handbook of Neuroimaging Data Analysis
Edited by Hernando Ombao, Martin Lindquist, Wesley Thompson, and John Aston

Preface

A strong relationship has always existed between sports and the statistics that are used to measure player and team performance. Many interesting questions about sports have led to serious research in sports statistics. Comprehensive surveys of statistics in sports research have been provided by books such as *Management Science in Sports* (1976), *Optimal Strategies in Sports* (1977), and *Statistics in Sport* (1998). The American Statistical Association created a section on Statistics in Sports in 1992, the International Statistical Institute created a Sports Statistics Committee in 1993, and the journal *Chance* has devoted a regular column to sports statistics.

In the approximate 20 years since the publication of *Statistics in Sport*, there has been a remarkable change in both the accumulation of sports data and the opportunities to address sports questions using statistical methods. Researchers and sports professionals have seen a recent explosion in the proliferation of data collected on a variety of aspects of sports. Motion-tracking technology has permitted the accumulation of detailed information on player-level dynamics, and large data archives for sports have become much more accessible worldwide. For example, a baseball researcher has opportunities to explore season data by Lahman database (`www.seanlahman.com`), game and play data using Retrosheet (`www.retrosheet.org`), and pitch-by-pitch data through the PitchFX system (`www.sportvision.com/baseball/pitchfx`). This accumulation of data has led to the development of new statistical methodologies. As an example, the traditional measurement of fielding in baseball is the fielding percentage, the fraction of plays by a particular fielder that is successful. With the development of new tracking systems, one can now measure the movement of a fielder toward a ball that is hit in his direction and obtain a much better measure of fielding performance that accounts for the range of a player.

The opportunities for statistical research in sports have correspondingly grown immensely, as reflected in new dedicated journals to the subject area. Two examples are the *Journal of Quantitative Analysis in Sports* founded in 2006 and the *Journal of Sports Analytics* founded in 2014. In addition, meetings focusing on statistics in sports such as MathSport International and the New England Symposium on Statistics in Sports are regularly scheduled events. Also, opportunities exist to present research on statistics in sports at industry professional meetings such as the MIT Sloan Sports Analytics Conference and the SABR Analytics Conference.

Volumes such as *Statistical Thinking in Sports* (2007, ed. Albert and Koning) and *Anthology of Statistics in Sports* (2005, ed. Albert, Bennett, and Cochran) consist of general arrays of statistical articles but are not designed to provide the reader with a complete survey of the state-of-the-art methods in statistics in sports. The general aim of this handbook is to provide a basic reference for statistical researchers and students with an interest in sports applications to learn about the fundamental background, problems, and ongoing challenges in statistical methods in sports. The chapters in this book provide both overviews of statistical methods in sports and in-depth treatment of critical problems and challenges confronting statistical research in sports. This handbook intends to provide the reader the necessary background to conduct serious statistical analyses for sports applications and to appreciate scholarly work in this expanding area.

This handbook should be of interest for three types of readers. First, the handbook can serve as the basis for a graduate course or seminar in statistical methods in sports. Using the

handbook in this fashion can take advantage of connections between the methods typically used in a sports context with the methods that graduate students are learning in theoretical coursework. Second, the handbook can serve as a reference for statistical practitioners in professional sports who may not be aware of the breadth of statistical issues and problems in their area, or who may simply want a refresher in the problem areas they are likely to encounter. Finally, the handbook can provide statistical researchers who are interested in delving more into sports applications the requisite background to produce sound scholarly work that is set in a proper context. The handbook is organized by major sport (baseball, American football, basketball, hockey, and soccer) followed by a section on other sports.

The four chapters on baseball provide a general description of measures of player performance and situational and streakiness effects. The chapter by Ben Baumer and Pamela Badian-Pessot describes the wide range of measures proposed for evaluating batters and base runners. Carson Sievert and Brian Mills discuss, in their chapter, traditional measures of pitching performance and explore the opportunities for further insight about pitching using pitch-level data from the PITCHf/x system. Measuring fielding performance has been one of the more challenging problems in baseball analysis. The chapter by Mitchel Lichtman describes a plethora of fielding measures that have been proposed and the use of modern technology systems such as Statcast and Fieldf/x that can construct improved defensive metrics. In sports, fans are fascinated with streaky and "clutch" performances of players and teams, and Jim Albert, in his chapter, describes the use of statistical models to detect and estimate the size of situational and streaky effects.

The topics of the American football chapters address the issues of evaluating college talent, measuring player and team abilities, and decision-making within a game. The section on football starts with a chapter by Mark Glickman and Hal Stern describing methods for estimating NFL team abilities based on game outcomes. This is followed by a chapter describing the methods of evaluating NFL quarterbacks and placekickers by Drew Pasteur and John David. The next chapter by Julian Wolson, Vittorio Addona, and Rob Schmicker is devoted to forecasting the success of NFL players based on college performance. The final chapter in this section by Keith Goldner presents a discussion of quantitative methods for making optimal strategic decisions within a football game.

The four chapters on basketball analytics cover a range of topics about player abilities and game outcomes. The chapter by James Lackritz describes the ongoing controversy surrounding streak shooting and methods for detecting the hot hand in NBA basketball. This is followed by a chapter by Jeremias Engelmann on the history and current usage of plus/minus methods for evaluating player contributions from possession-based data. Brian Skinner and Matthew Goldman present the basics of optimal strategy in basketball in their chapter, with a focus on optimizing when players should take shots, at what point in the shot clock a team should take a shot, and under what circumstances teams should try high-risk tactics. Finally, the basketball section is concluded by a chapter by Luke Bornn, Daniel Cervone, Alexander Franks, and Andrew Miller that explores the current state of the art of analyzing player tracking data to measure both offensive and defensive player abilities.

The next group of chapters concerns hockey analytics. A chapter by Andrew Thomas develops the structure for an NHL match simulator using Poisson/Exponential models. The approach is comprehensive in that it takes into account various game situations (e.g., penalties, score, etc.) that affect the play of the game. In the next chapter, Bobby Gramacy, Matt Taddy, and Sen Tian use regularized logistic regression to assess player performance. Their approach may be seen as a generalization and an improvement of traditional plus/minus methods. A chapter by Michael Schuckers surveys the statistics used

to evaluate goaltending. Some of the statistics take into account the detailed aspects of goaltending, including the type of shots and the location of shots. In the final hockey chapter, Peter Tingling looks at drafting with a specific focus on the nuances of the NHL draft.

Soccer is an example of a low-scoring sport. Compared to, say, baseball, relatively few performance measures are available, and for that reason, focus has been on models for outcomes and the effect of interventions on these outcomes. The chapter by Phil Scarf and Jose Rangel Sellitti discusses recent developments in the literature on score modeling, focusing on the dependence between the number of goals scored by the home team and the number scored by the away team. At a slightly more abstract level, information about scores results in information about team quality. Ruud Koning, in his chapter, discusses a range of models that have been proposed to measure team quality. At a more disaggregate level, the chapter by Ian McHale and Samuel Relton describes the analysis of individual player ratings using data that have become available only recently. The soccer section concludes with chapters on intervention issues in the quantitative analysis of soccer. The chapter by Martin van Tuijl discusses whether dismissing a coach midseason results in better performance. Another issue explored by the chapter by Rob Simmons is whether referees are biased or impartial judges of the game, a topic that is relevant to other sports as well.

The final section of the handbook contains three contributions related to other sports. The first chapter by Mark Broadie and Bill Hurley investigates the two major research areas in golf. They look at the use of detailed ShotLink data in golf analytics and the age-old problem of handicapping in golf. In the second chapter, Tim Swartz surveys statistical research in cricket, the second most popular sport in the world. One take-away from this chapter is that the game has been underexplored and that opportunities exist in cricket analytics. The final chapter by Elmer Sterken discusses the issue whether inequality in performance in Olympic sports between top athletes increases or decreases over time. As discussed in this chapter, a general observation is that the performance of athletes tends to improve, but once a sport has reached maturity, improvement can level off.

We thank the chapter writers for their important contributions and our many colleagues who have inspired us to create a handbook that we hope sets a standard for researchers and practitioners in statistics in sports.

Jim Albert
Mark E. Glickman
Tim B. Swartz
Ruud H. Koning

Contributors

Vittorio Addona
Department of Mathematics, Statistics, and Computer Science
Macalester College
Saint Paul, Minnesota

Jim Albert
Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, Ohio

Pamela Badian-Pessot
School of Operations Research and Information Engineering
Cornell University
Ithaca, New York

Benjamin S. Baumer
Program in Statistical & Data Sciences
Smith College
Northampton, Massachusetts

Lucas M. Besters
Department of Economics
Tilburg University
Tilburg, the Netherlands

Luke Bornn
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, British Columbia, Canada

Mark Broadie
Graduate School of Business
Columbia University
New York, New York

Babatunde Buraimo
Department of Economics, Finance, and Accounting
University of Liverpool Management School
Liverpool, United Kingdom

Daniel Cervone
Center for Data Science
New York University
New York, New York

John A. David
Department of Applied Mathematics
Virginia Military Institute
Lexington, Virginia

Jeremias Engelmann
Consultant
ESPN.com
Heidelberg, Germany

Alexander Franks
Department of Statistics
University of Washington
Seattle, Washington

Mark E. Glickman
Department of Statistics
Harvard University
Cambridge, Massachusetts

Matthew Goldman
Microsoft Research
Redmond, Washington

Keith Goldner
numberFire
New York, New York

Robert B. Gramacy
Department of Statistics
Virginia Tech
Blacksburg, Virginia

William J. Hurley
Department of Mathematics and Computer Science
Royal Military College of Canada
Kingston, Ontario, Canada

Ruud H. Koning
Department of Economics, Econometrics and Finance
University of Groningen
Groningen, the Netherlands

James Lackritz
MIS Department
College of Business Administration
San Diego State University
San Diego, California

Mitchel Lichtman
Baseball Analyst
Canandaigua, New York

Ian G. McHale
Centre for Sports Business
Salford Business School
University of Salford
Manchester, United Kingdom

Andrew Miller
Department of Computer Science
Harvard University
Cambridge, Massachusetts

Brian M. Mills
Department of Tourism, Recreation, and Sport Management
University of Florida
Gainesville, Florida

R. Drew Pasteur
Department of Mathematics and Computer Science
The College of Wooster
Wooster, Ohio

José Sellitti Rangel Jr.
Salford Business School
University of Salford
Manchester, United Kingdom

Samuel D. Relton
School of Mathematics
University of Manchester
Manchester, United Kingdom

Phil Scarf
Salford Business School
University of Salford
Manchester, United Kingdom

Michael E. Schuckers
Department of Mathematics, Computer Science, and Statistics
St. Lawrence University
Canton, New York

Robert Schmicker
Department of Biostatistics
University of Washington
Seattle, Washington

Dirk Semmelroth
Department of Management
University of Paderborn
Paderborn, Germany

Carson Sievert
Department of Statistics
Iowa State University
Ames, Iowa

Rob Simmons
Department of Economics
Lancaster University Management School
Lancaster, United Kingdom

Brian Skinner
Department of Physics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Elmer Sterken
Institute of Economics, Econometrics, and Finance
University of Groningen
Groningen, the Netherlands

Hal S. Stern
Department of Statistics
University of California, Irvine
Irvine, California

Tim B. Swartz
Department of Statistics and
Actuarial Science
Simon Fraser University
Burnaby, British Columbia, Canada

Matt Taddy
Microsoft Research New England
Cambridge, Massachusetts
and
Booth School of Business
University of Chicago
Chicago, Illinois

Andrew C. Thomas
Minnesota Wild
National Hockey League
Saint Paul, Minnesota

Sen Tian
Stern School of Business
New York University
New York, New York

Peter M. Tingling
Beedie School of Business
Simon Fraser University
Vancouver, British Columbia, Canada

Jan C. van Ours
Department of Applied Economics
Erasmus University Rotterdam
Rotterdam, the Netherlands
and
Department of Economics
University of Melbourne
Parkville, Australia

Martin A. van Tuijl
Department of Economics
Tilburg University
Tilburg, the Netherlands

Julian Wolfson
Division of Biostatistics
School of Public Health
University of Minnesota
Minneapolis, Minnesota

Contents

1

Evaluation of Batters and Base Runners

Benjamin S. Baumer and Pamela Badian-Pessot

CONTENTS

Since Henry Chadwick began publishing boxscores around the turn of the twentieth century (Schwarz, 2005a), people have been interested in evaluating the performance of baseball players. In particular, the offensive contributions made by position players—in the form of both batting and baserunning—are the most obviously varied and carefully studied contributions. In this chapter, we will catalog the most enduring sabermetric models for evaluating batters and base runners. Our approach is model centric, in that we will attempt to categorize metrics based on the type of model on which they are built. It should not be surprising that over time these models have become more sophisticated, both in terms of the model complexity and the rigor with which any parameters are estimated.

The fundamental challenge in evaluating batters and base runners is that run scoring in baseball is the result of interdependent actions among teammates and opponents. While separating the individual contributions of these team actions is considered easier in baseball than in many other sports, it is far from trivial. For example, consider an inning in which the first batter leads off with a walk, steals second base, advances to third base on a groundout, and then scores on a sacrifice fly. What is unambiguous is that the team scored one run. What is debatable is how much of that run is attributable to each player. Does the second batter make a positive contribution by advancing the runner, even though he made an out? How much credit did the first player accrue though baserunning? The tools developed in this chapter will help us answer these questions.

One common technique that will prime the search for models is

- To recognize that neither the runs scored statistic (R), which gives full credit for the team run to the player who crossed the plate, nor the runs batted in statistic (RBI), which gives full credit for the team run to the player who drove in the run, are reasonable ways to apportion the run
- To find a model for R that works well for teams
- To apply that model to individual players

Later in this chapter, we will show how this method can be used to evaluate the accuracy of offensive metrics.

In Section 1.1, we motivate this line of inquiry, outline some basic tools necessary to understand these models, and define our notation. In Section 1.2, we explore models that can be computed using data that are aggregated at the seasonal level. These are the simplest but most numerous and storied models for offensive performance. Models that require more detailed play-by-play data are explored in Section 1.3. Predictive models, including Bayesian models, are discussed in Section 1.4. We conclude the chapter in Section 1.5 by pointing toward some open problems.

1.1 Basic Tools

1.1.1 Overview

Several comprehensive assessments of offensive players have advanced the field of sabermetrics. Many have drawn inspiration from the work of Bill James interspersed in his books (James, 1986; James and Henzler, 2002). Perhaps the first book-length treatise was Cook (1964). This was followed 20 years later by Thorn and Palmer (1984), an important book that was reprinted in 2015 (Thorn and Palmer, 2015). The article-length analysis of Bennett and Flueck (1983) was greatly expanded by Albert and Bennett (2003) into what remains probably the best place to start reading about sabermetrics. The more recent book by Tango et al. (2007) not only focuses more on in-game strategy, but also includes some measures of offensive assessment. Methods for analyzing baseball data using the statistical computing environment R—as we do in this chapter—are explicated by Marchi and Albert (2013).

1.1.2 Expected Run Matrix

One of the fundamental tools in sabermetric analysis is the *expected run matrix*, which gives the expected number of runs scored in the remainder of an inning, given that the inning is currently in one of the 24 (*base*, *out*) states. Throughout this chapter, we will use the notation $\mathbf{R}$ to refer to this 8×3 matrix, and the notation $\mathbf{r}$ to refer to the corresponding vector of length 24. Specifically, the notation $\mathbf{R}_{23,2}$ indicates the value of the expected run matrix when runners are on second and third with two outs.* The complete expected run matrix for 2013 is presented in Table 1.1.

1.1.3 Notation

In this chapter, we denote the value of player i's batting and baserunning contributions as y_i. Typically, but not always, player value is measured in the units of runs. Since, as we discussed earlier, there is no way to *directly* measure the run value of these contributions for individual players, we consider y_i to be unknown. The goal of this chapter is to describe models for y_i in a coherent fashion. The estimates for y_i will be denoted by $\hat{y}_i$. Note that from the example at the beginning of this chapter, each of the three offensive players has a y_i—we just don't know what they are.

An example may make this clearer. It is an undisputed fact that the St. Louis Cardinals scored $y_{STL} = 798$ runs in 1987. This number represents the total batting and baserunning contributions of the entire team. However, we don't know how many of those 798 runs are attributable to each player. We know that Ozzie Smith scored 104 of those runs and drove in another 75, but neither of those numbers represents y_{Smith}. Furthermore, neither is

* There are several commonly used notations for referring to the configuration of base runners indicated by *base*. The most intuitive is a notation like `23` (or `×23`), which indicates that there are runners on second and third. We will use this notation in the text of this chapter. However, this notation is very inconvenient computationally. In our computations, we use the following notation: imagine each of the three bases as a binary digit that can either be unoccupied (0) or occupied (1). Then the binary string 110 indicates that runners are on second and third. This has the decimal equivalent of 6. Thus, the notations `x23`, 110, and 6 all refer to having runners on second and third. We trust the reader will be able to keep this clear in context.