Statistical Thinking in Sports

Contents

1	Intro	duction		1		
	Jim Albert and Ruud H. Koning					
	1.1	Introdu	action	1		
		1.1.1	Patterns of world records in sports (2 chapters)	2		
		1.1.2	Competition, rankings, and betting in soccer (3 chapters).	2		
		1.1.3	An investigation into some popular baseball myths (3 chap-			
			ters)	3		
		1.1.4	Uncertainty of attendance at sports events (2 chapters)	4		
		1.1.5	Home advantage, myths in tennis, drafting in hockey pools, American football	4		
	1.2	Wahsit	te	5		
				5		
	Kelei	ence .		3		
2	Mod	elling th	ne development of world records in running	7		
	Gera	rd H. Ku	per and Elmer Sterken			
	2.1	Introdu	action	7		
	2.2 Modelling world records					
		2.2.1	Cross-sectional approach	10		
		2.2.2	Fitting the individual curves	11		
	2.3	Selecti	on of the functional form	12		
		2.3.1	Candidate functions	12		
		2.3.2	Theoretical selection of curves	17		
		2.3.3		18		
		2.3.4	The Gompertz curve in more detail	18		
	2.4	Runnir	ng data	23		
	2.5	Results	s of fitting the Gompertz curves	23		
	2.6	Limit values of time and distance				
	2.7	Summary and conclusions				
	Refe	rences .		29		
3	The j	physics a	and evolution of Olympic winning performances	33		
	Ray S	Stefani				
	3.1	Introdu	action	33		
	3.2	Runnir	ng events	34		

		3.2.1 The physics of running	34		
		3.2.2 Measuring the rate of improvement in running	37		
		3.2.3 Periods of summer Olympic history	38		
		3.2.4 The future of running	40		
	3.3	Jumping events	40		
		3.3.1 The physics of jumping	40		
		3.3.2 Measuring the rate of improvement in jumping	43		
		3.3.3 The future of jumping	44		
	3.4	Swimming events	46		
		3.4.1 The physics of swimming	46		
		3.4.2 Measuring the rate of improvement in swimming	47		
		3.4.3 The future of swimming	49		
	3.5	Rowing	49		
		3.5.1 The physics of rowing	49		
		3.5.2 Measuring the rate of improvement in rowing	50		
		3.5.3 The future of rowing	52		
	3.6	Speed skating	53		
		3.6.1 The physics of speed skating	53		
		3.6.2 Measuring the rate of improvement in speed skating	54		
		3.6.3 Periods of winter Olympic history	55		
		3.6.4 The future of speed skating	57		
	3.7	A summary of what we have learned	57		
	Refer	rences	59		
4		petitive balance in national European soccer competitions	63		
		o Haan, Ruud H. Koning, and Arjen van Witteloostuijn			
	4.1	Introduction	63		
	4.2	Measurement of competitive balance	64		
	4.3	Empirical results	67		
	4.4	Can national competitive balance measures be condensed?	72		
	4.5	Conclusion	74		
	Refer	rences	74		
5	Statio	stical analysis of the effectiveness of the FIFA World Rankings	77		
		IcHale and Stephen Davies	• •		
		Introduction	77		
	5.2	FIFA's ranking procedure	78		
	5.3	Implications of the FIFA World Rankings			
	5.4				
	5.5	Preliminary analysis	80		
	5.5	5.5.1 Team win percentage, in and out of own confederation	80		
		5.5.2 International soccer versus domestic soccer	82		
		2.2.2 International society versus defines a contest contest.	02		

Ta	ble of	Contents	V				
	5.6	Forecasting soccer matches	84				
	5.7	Using the FIFA World Rankings to forecast match results	84				
		5.7.1 Reaction to new information	85				
		5.7.2 A forecasting model for match result using past results	86				
	5.8	Conclusion	89				
	Refe	rences	89				
6	Fore	casting scores and results and testing the efficiency of the fixed-odd	s				
	betti	betting market in Scottish league football					
	Stepl	hen Dobson and John Goddard					
	6.1	Introduction	91				
	6.2	Literature review	92				
	6.3	Regression models for goal scoring and match results	95				
	6.4	Data and estimation results	97				
	6.5	The efficiency of the market for fixed-odds betting on Scottish league					
		football	102				
	6.6	Conclusion	107				
	Refe	rences	107				
7	Hitting in the pinch						
	Jim A	Jim Albert					
	7.1	Introduction	111				
	7.2	A breakdown of a plate appearance: four hitting rates	112				
	7.3	Predicting runs scored by the four rates	114				
	7.4	Separating luck from ability	114				
	7.5	Situational biases	117				
	7.6	A model for clutch hitting	124				
	7.7	Clutch stars?	125				
	7.8	Related work and concluding comments	127				
	Refe	rences	133				
8	Does	s momentum exist in a baseball game?	135				
	Rebe	ecca J. Sela and Jeffrey S. Simonoff					
	8.1	Introduction	135				
	8.2	Models for baseball play	136				
	8.3	Situational and momentum effects	138				
	8.4	Does momentum exist?	140				
		8.4.1 Modeling transition probabilities	140				
		8.4.2 Modeling runs scored	144				
	8.5	Rally starters and rally killers	149				
	8.6	Conclusions	150				
	Refe	rences	151				

y		ence about batter-pitcher matchups in baseball from small	
	samp	les	153
	Hal S. Stern and Adam Sugano		
	9.1	Introduction	153
	9.2	The batter-pitcher matchup: a binomial view	154
	9.3	A hierarchical model for batter-pitcher matchup data	155
		9.3.1 Data for a single player	155
		9.3.2 A probability model for batter-pitcher matchups	156
		9.3.3 Results - Derek Jeter	158
		9.3.4 Results - multiple players	160
	9.4	Batter-pitcher data from the pitcher's perspective	160
		9.4.1 Results - a single pitcher	161
		9.4.2 Results - multiple players	163
	9.5	Towards a more realistic model	163
	9.6	Discussion	164
	Refer	rences	165
10		ome uncertainty measures: how closely do they predict a close	
	game		167
		tunde Buraimo, David Forrest, and Robert Simmons	
	10.1	Introduction	167
	10.2	Measures of outcome uncertainty	169
	10.3	Data	171
	10.4	Preliminary analysis of the betting market	172
	10.5	Model	173
	10.6	Out-of-sample testing	175
	10.7	Concluding remarks	176
	Refer	rences	177
11		mpact of post-season play-off systems on the attendance at regula	
		on games	179
		Bojke	
	11.1		179
	11.2	Theoretical model of the demand for attendance and the impact of	
		play-off design	181
	11.3	Measuring the probability of end-of-season outcomes and game sig-	
		nificance	183
	11.4	The data: the 2000/01 English Football League 2nd tier	185
	11.5	Statistical issues in the measurement of the determinants of atten-	
		dance	190
		11.5.1 Skewed, non-negative heteroscedastic data	190

Tal	ble of (Contents		vii
		11.5.2	Clustering of attendance within teams and unobserved het-	
			erogeneity	192
		11.5.3	Multicollinearity	192
		11.5.4	Final statistical model	193
	11.6		estimation	194
		11.6.1	Choice of explanatory variables	194
		11.6.2	Regression results	195
	11.7		pact of the play-off system on regular league attendances .	197
	11.8		sions	199
				201
	110101	ciices		201
12			t and interpretation of home advantage	203
	•	Itefani		
	12.1		ction	203
	12.2		ring home advantage	204
	12.3		union, soccer, NBA	207
	12.4	Austral	ian rules football, NFL, and college football	211
	12.5		ockey and MLB baseball	212
	12.6	Can ho	me advantage become unfair?	214
	12.7	Summa	nry	214
	Refer	ences		215
13	Myth	s in Ten	nis	217
	Jan Magnus and Franc Klaassen			
	13.1	Introduction		
	13.2		ta and two selection problems	218
	13.3		e myths	221
	10.0	13.3.1	A player is as good as his or her second service	223
		13.3.2	Serving first	224
		13.3.3	New balls	226
	13.4		g mood	229
	13.1	13.4.1	At the beginning of a final set, both players have the same	22)
		13.4.1	chance of winning the match	230
		13 / 2	In the final set the player who has won the previous set has	230
		13.4.2	the advantage	231
		13.4.3	After breaking your opponent's service there is an increased	231
		13.4.3		222
		12 4 4	chance that you will lose your own service	232
		13.4.4	After missing break points in the previous game there is an	222
	12.5	D'.	increased chance that you will lose your own service	233
	13.5	Big poi		234
		13.5.1	The seventh game	234
		13.5.2	Do big points exist?	235

viii	i	Statistical Thinking in Sp	orts
	13.6 Refer	13.5.3 Real champions	237 238 239
14	Back	to back evaluations on the gridiron	241
	David	l J. Berri	
	14.1	Why do professional team sports track player statistics?	241
	14.2	The NFL's quarterback rating measure	242
	14.3	The Scully approach	243
	14.4	Modeling team offense and defense	244
	14.5	Net Points, <i>QB Score</i> , and <i>RB Score</i>	252
	14.6	Who is the best?	253
	14.7	Forecasting performance in the NFL	254
	14.8	Do different metrics tell a different story?	259
	14.9	Do we have marginal physical product in the NFL?	260
	Refer	ences	261
15	Optir	nal drafting in hockey pools	263
Amy E. Summers, Tim B. Swartz, and Richard A. Lockhart			
	15.1	Introduction	263
	15.2	Statistical modelling	264
		15.2.1 Distribution of points	264
		15.2.2 Distribution of games	266
	15.3	An optimality criterion	268
	15.4	A simulation study	269
	15.5	An actual Stanley Cup playoff pool	273
	15.6	Discussion	276
	Refer	ences	276
Re	ferenc	es	277
Lis	st of au	uthors	291
Ind	lex		295

Introduction

Jim Albert

Bowling Green State University

Ruud H. Koning

University of Groningen

1.1 Introduction

Sports have taken an ever more prominent position in society. An increasing number of people watch sports events on television and more people see live sports in stadiums and arenas. The economic value of franchises, broadcasting rights, and merchandising has grown. Books on sports appear on the *New York Times* bestseller list (for example, Lewis, 2004). Besides this increasing interest from the general public, scientists have also taken on sports for their research agenda. Traditionally, research from physiology and medicine has been used to improve performance in sports. Currently researchers from economics, statistics, sociology, and law are working in the sports field.

In this volume, the focus is on the statistical analysis of sports. There has always been a close connection between sports and statistics. In most sports, players and teams are measured by various statistics, and these statistics are used to provide rankings of players and teams. A recent popular phenonema is fantasy sports where participants draft teams of players and games are won and lost on the basis of actual statistical information.

One reason for the close connection between statistics and sports is probably the abundant data that is available on sports. Scores are kept and individual performance is measured and tracked over time. The advent of the internet has perhaps helped to distribute these data to an ever wider group of researchers. Also, sport is a convenient and familiar context to use in teaching or in demonstrating a new statistical method.

In this volume, we illustrate a number of different models for sports data, such as time series, linear regression, ordered probit regression, factor analysis, and generalized linear models. A variety of distributions are used to model the variable of interest such as binomial, Gamma, Poisson, and others. Parameters are estimated by least squares, maximum likelihood, and Bayesian methods and used in simulation models. Indeed, sports provide a very broad area of application of statistical thinking.

1.1.1 PATTERNS OF WORLD RECORDS IN SPORTS (2 CHAPTERS)

One fascinating subject for study is the pattern of world records in sports over time. Gerald Kuper and Elmer Sterken, in "Modelling the development of world records in running", provide an interesting look into the pattern of world records of metric running events for men and women. As the authors explain, there are interesting questions associated with this data. Can one estimate the ultimate human performance in these events? What is the impact of technology innovations on the pattern of records? Is the pattern of world records similar for different running lengths, and will women outperform men in the future? The authors provide a comprehensive survey of the use of different parametric families to model world records and give some interesting conclusions.

Ray Stefani in "The physics and evolution of Olympic winning performances" takes a "holistic" view of the pattern of winning performances over time in a variety of Olympic sports. To understand the changes in the winning performance in a given sport, say swimming, one should understand the factors influencing the power of a swimmer, such as the size and fitness of the athletes, the arm and leg positioning techniques, coaching, and the quality of the venue and equipment. Although one cannot directly compare the time of a swimming event with the height of a pole vault, Stefani defines a dimensionless measure, percent improvement per Olympiad, to describe the change in winning performances. This measure is used to contrast the evolution of winning performances in a wide variety of sports.

1.1.2 COMPETITION, RANKINGS, AND BETTING IN SOCCER (3 CHAPTERS)

Soccer is the most popular sport in the world, as judged by the number of people playing or watching the sport. Besides being interesting as a sport, it has also become an economic activity of some significance. For example, anti-trust regulators watch the sale of television rights, and the European Commission is involved in setting up a system of transfer fees. Also, just as in many other sports, betting on soccer matches has become increasingly popular. Despite ongoing commercialization of soccer, one would almost forget that it is a game; and organizing leagues want to know which team is best. These issues are addressed in Chapters 4, 5, and 6.

Marco Haan, Ruud Koning, and Arjen Van Witteloostuijn look at the development of competitive balance over time in their chapter "Competitive balance in national European soccer competitions." They do so for seven different countries. First, they discuss different dimensions of competitive balance and propose empirical measures that capture these dimensions. Then they proceed to examine whether balance has changed over time, in particular, they investigate the popular belief that competitive balance has worsened over time. Finally, noting the lack of agreement on a single measure of competitive balance in soccer, they use a factor model to see whether seven different indicators can be reduced to one factor. It turns out that the predominant factor can be interpreted as contemporaneous competitive balance.

Soccer is played at different levels: club teams play in national leagues and international tournaments as the Champions League, and national teams play every four years to win the World Cup. Ranking of club teams is relatively easy, considering the

Introduction 3

number of games they play. It is much harder to rank national teams as, in a given year, they play only a limited number of games, against a selected set of opponents. Still, the world soccer federation FIFA publishes a ranking of national teams, that is updated frequently. What is the quality of this ranking? This issue is addressed by Ian McHale and Stephen Davies in "Statistical analysis of the effectiveness of the FIFA World Rankings." They conclude that the FIFA World Ranking does not use all past information efficiently.

Betting on sport results is a hobby for one, and a way of earning a living for others. To what extent are betting markets efficient? Stephen Dobson and John Goddard's "Forecasting scores and results and testing the efficiency of the fixed-odds betting market in Scottish league football" examine different betting strategies using two different statistical forecast models: a goals-based model and a results-based model. These forecasting models are capable of eliminating almost all of the bookmakers over-round.

1.1.3 AN INVESTIGATION INTO SOME POPULAR BASEBALL MYTHS (3 CHAPTERS)

Baseball has been called the most statistical sport in the sense that more numerical information is collected about this game than any other. For a given baseball play, such as a batted ball hit into center field for a single, many associated variables will be recorded about the event including the inning, runners on base, the players on the field, and the exact location of the hit in the field. Websites such as The Baseball Archive (www.baseball1.com) and Retrosheet (www.retrosheet.org) provide extensive datasets on historical players and teams and play-by-play game results. The easy availability of this data invites interesting analyses by researchers that are reflected in the three baseball papers in this volume.

All three articles investigate the validity of popular myths in baseball. "Hitting in the pinch" by Jim Albert investigates the popular belief that particular ballplayers have the ability to perform better in important or "clutch" situations during a game. In his paper, Albert shows that the ability to hit well can depend on the runners on base and the number of outs in an inning. But there is little evidence to suggest that particular players have the ability to do better in important situations. Another popular belief in sports is the importance of momentum during a game. If particular players perform well during a game, many people believe that this will motivate other players to also perform well, causing the team to rally. Rebecca Sela and Jeffrey Simonoff take a statistical view of this issue in "Does momentum exist in a baseball game?" They begin with a Markov Chain model for baseball, where the probability of a movement from one state (say no runners on and one out) to another state (runner on first and one out) only depends on the beginning state. They consider a more sophisticated model where the probability of a movement can depend on various "momentum" variables. The authors find little statistical evidence of momentum effects, especially from a predictive viewpoint. Hal Stern and Adam Sugano in "Inference about batter-pitcher matchups in baseball from small samples" investigate the final myth, the importance of batter/pitcher matchup data. Baseball managers will often make decisions on the basis of how particular batters perform against particular pitchers. The problem is that one has many samples of large number of batter/pitcher matchups, and one is likely to see extreme sample outcomes by chance. Stern and Sugano suggest modeling this type of data by a hierarchical model; this model will allow one to see how a batter's ability can vary depending on the quality of the pitcher.

1.1.4 Uncertainty of attendance at sports events (2 chapters)

Professional sports is a business and all teams wish to have high attendance at their games. A popular assumption is that audiences will be attracted to games where the outcome is very uncertain. Babatude Buraimo, David Forrest, and Robert Simmons, in "Outcome uncertainty measures: how closely do they predict a close game?" note that there is little empirical support for this assumption. They suggest that one problem is that although there are several measures of outcome uncertainty used in the literature, it is unclear whether any of these measures are actually good predictors of close contests. This paper defines several measures of outcome uncertainty and finds that they only explain a small amount of the variation in game results of Spanish football.

Attendance at sports events is also a central theme in the article "The impact of post-season play-off systems on the attendance at regular season games" by Chris Bojke. Many different sports have introduced play-off systems; one benefit of these systems is that they enhance attendance by increasing the time that teams are in contention for the league championship. Unfortunately, there is little research on the impact of the play-off design on attendance and Bojke presents a statistical model to understand this relationship. By fitting the model to English football data, one is able to measure the impact of a particular play-off system on attendance. Even though his application is to soccer, the methodology can be applied to modeling attendance for other sports as well.

1.1.5 HOME ADVANTAGE, MYTHS IN TENNIS, DRAFTING IN HOCKEY POOLS, AMERICAN FOOTBALL

One is familiar with the saying "there is no place like home", and this statement is especially true for sports competitions. For all sports, the team playing in its home field or area generally has an advantage. Ray Stefani's article "Measurement and interpretation of home advantage" explores the home-field advantage for a number of sports. A mathematical model is used to quantify home field advantage for a particular sport and the size of the estimated home advantage is shown to differ greatly between sports. Stefani describes the physiological, psychological and tactical factors implicit in home-field advantage and argues that player fatigue, especially in a continuous-action sport, plays an important role in home advantage.

Jan Magnus and Franc Klaassen provide a nice survey of their tennis research in "Myths in tennis." Tennis is an international sport watched by fans all over the world and the television commentators have strong views about competition. In particular, commentators believe a player has an advantage if he or she serves first, and top play-

Introduction 5

ers have a special ability to perform well in the important points in a match. This article explores these beliefs by use of four years of point-by-point tennis data from the Wimbledon tennis championships. Many interesting conclusions are reached about tennis matches and most of the beliefs held by television competitors are shown to be false.

David Berri's article "Back to back evaluations on the gridiron" describes the evaluation of player performance in American football. Baseball is a relatively easy sport in evaluating player performance since the game is essentially a confrontation between a single pitcher and a single batter. Football is fundamentally different from baseball in that the performance of a particular player such as a quarterback is highly dependent on the performance of his teammates on the field. Berri applies the regression methodology of Scully and Blass to develop a measure of marginal performance for football players. This methodology leads to some interesting measures of player performance; in particular, Berri's measure of a quarterback's performance is likely superior to the official quarterback rating system of the National Football League. But these measures do not appear to show consistency over time, suggesting that the statistics collected by professional football are not useful for measuring the productivity of individual players.

Amy Summers, Tim Swartz, and Richard Lockhart in "Optimal drafting in hockey pools" considers how one can be successful in a hockey fantasy league. In this game, participants draft players from the 16 National Hockey League teams that have qualified for the Stanley Cup Playoffs. The winner of the league is the one whose players have accumulated the largest number of points. An interesting statistical model is devised for the number of points scored by hockey players, and this is used to find an optimal selection strategy. One attractive by-product of this model is that a player can make intelligent draft choices in real time.

1.2 Website

This book is intended for sports enthusiasts, with some background in statistics. They can be students, teachers, researchers, but also practitioners or (sport) policy makers. To promote more statistical thinking in sports, we have made a website www.statistical-thinking-in-sports.com that has additional material as appendices, references, tables, and other data. Feel free to use the information provided there, but do send a copy of your paper or project for inclusion on the website if you use material from that website.

REFERENCES

Lewis, M. (2004). Moneyball: The Art of Winning an Unfair Game. New York: W.W. Norton & Company.