



Statistics in Sports



Newsletter of the Statistics in Sports Section of the American Statistical Association. Spring 1999 Vol 1 No. 1

The View From The Chair On The Bench

Bob Wardrop, 1999 SIS Section Chair,
Department of Statistics, University of Wisconsin—Madison

I am very excited about this first of what I am sure will be many newsletters. The Section is deeply indebted to Jim Albert and Scott Berry for their efforts in starting this endeavor! I encourage you to become involved, as a contributor or a reviewer.

As a teacher, I find sports examples to be invaluable. Nearly all of my students are involved in some sort of sports activity. I have my students perform small projects, and a majority freely select a project with a sports theme. Applications of Statistics to sports show my students that Statistics can be relevant in their lives! Sports examples are good for showing the interplay of theory and applications. Theory provides formulas, but knowledge of the field of application is necessary for deciding whether a formula is appropriate and, if it is, for interpreting its answer.

As a profession we continually need to address the issue of our relevance. One of the best ways to do this is by demonstrating that we can “learn the hidden truths” in data that are of interest to the public. And, clearly, the public is interested in sports data. Thus, I believe that one of the best ways to educate the public to the importance of Statistics is through high-quality sports research.

The good news is that much good research is being conducted, including work by some of the most highly respected statisticians. The bad news is that there are two serious barriers to academic statisticians doing sports research—funding and recognition. I will not try to address funding, nor will I attempt to pretend to understand issues that nonacademics face. I can only dream of a day when tenure is awarded on the basis of outstanding research in sports applications of Statistics. In the meantime, defend the value of sports research when your colleagues demean it. Work to ensure that quality sports research will be viewed positively when salaries are set and promotions given. Take your role as a referee very seriously; if we allow shoddy sports research to be published, we only provide fuel for our detractors.

Sports Statistics News From Down Under



Steve Clarke and David Dyte, School of Mathematical Sciences, Swinburne University of Technology

We cannot claim this to be the news Australia wide, but at least from Swinburne University. But we reckon this to be the centre for Sports Statistics in Australia. Swinburne has been involved in Sports Statistics since Steve Clarke began providing computer tips for Australian rules football to a daily newspaper in 1981. Currently these tips appear in *The Australian Financial Review* and in the last few years they have also obtained coverage on TV.

CONTINUED ON PAGE 2



Where Should I Send That Sports Article?

Hal Stern, Department of Statistics, Iowa State University (Editor, *Chance*)

I expect that many members of the section have completed a small research project or data analysis related to sports and wondered whether it might be published in a refereed journal. Articles applying probability and statistics to sports have appeared in *The American Statistician*, *JASA*, *Applied Statistics*, and *Chance*, and more. This note contains some thoughts about whether and where to submit sports articles. As the current editor of *Chance*, I would like to take the opportunity to explain the kinds of articles that *Chance* would like to publish.

CONTINUED ON PAGE 2

Sports Statistics News From Down Under (Cont.)

We also have a significant input into articles now written on sport in the Australian Financial Review each Weekend Edition. During the Football season, they are usually on Australian Rules, and relate to official Australian Football League performance statistics collected by Champion data. The articles often cover topics in which we have research interest, such as team ratings and home advantage. In the off season, the articles range over many sports, and are not afraid to discuss statistical principles. For example one on football covered regression to the mean, one on cricket dynamic programming. The articles have discussed some of our research in sports such as golf, cricket, basketball, baseball, soccer, tennis, and football. Needless to say, our work doesn't have to be world shaking - the day to day nature of the daily press has meant some of our research support has been done in half a day.

David Dyte was appointed as a Sports Statistician here in April to assist Steve. His arrival has seen an increase in our Web presence. We now have 3 alternative computer football predictions, and a similar system for basketball. Some soccer tips (provided by Ray Stefani from California State University) are on the way. We have also produced material relating to tennis and cricket, including a playable cricket simulator. Check out our Website www.swin.edu.au/sport.

We have a couple of graduate students: Steve Clarke completed his PhD in 1998; Andrew Patterson is doing an MSc on Player performance statistics in Australian football; Paul Allsopp is doing a PhD in home ground advantage in cricket, David Dyte is about to start a PhD in cricket simulation and modeling. We are about to launch a new undergraduate course in Computing and Applied Statistics, which includes a substantial sports statistics component. Our published papers tend to appear in the OR journals and conferences, so may be missed by ASA members.

We are very interested in expanding activities here, and would certainly be keen to hear from anyone wishing to visit. We do have a great conference planned in 2000 in Sydney, Australia on "Mathematics and Computers in Sport" About half the papers at this meeting are statistical in nature.

Where Should I Send That Sports Article (Cont.)

Any publication will include sports articles only to the extent that they fit the mission of the publication. Thus, *JASA* would be expected to publish sports articles only if they represent an unusual data set (Albright's streak hitting in baseball article in 1993) or a novel method of analysis. In fact most traditional statistics journals would have similar policies. This means that articles which apply traditional statistical techniques (linear regression, logistic regression, etc.) to sports data will likely not be accepted. It is important to note that this is not necessarily any prejudice against sports, those journals would not accept standard analyses of data from any field. The difficulty of working with sports data is that there is at the moment no real subject-matter journal to which material can be sent.

This brings us to two topics of important current interest: the role of *Chance* and the possibility of a new sports journal. The *Chance* magazine mission is to publish interesting applied work that will showcase the role of statistics in a variety of fields. The two key features that we look for in a *Chance* article are: (1) the article should address an interesting and important question; (2) the article should make appropriate use of statistics. An article that compares the performance of two specific baseball players is not likely to be accepted because the question is so narrow. A similar article that showed how baseball performance can be measured might well be acceptable. At the current time, *Chance* receives many more submissions related to sports than to any other field. *Chance* will continue publishing sports articles but must try to maintain a balance among the areas of application that it covers.

The other idea that is discussed on occasion is the creation of a sports statistics journal. Could such a journal work? There appear to be a lot of willing authors from within the field of statistics. However would editorial or refereeing work for such a journal be valued by academic statistics departments as professional service? In some ways it might be best if there were a statistics corner in some existing sports science journal. The section website has an extensive list of possible journals (thanks to Bob Schutz).

The Value of Home Runs: McGwire Versus Sosa in 1998



Jay Bennett, Bellcore

In 1998, baseball fans thrilled to the assault by Mark McGwire and Sammy Sosa on the seasonal home run record. Roger Maris' record of 61 home runs lasted for almost 37 years, longer than the tenure of the previous record holder, Babe Ruth. The excitement built while McGwire broke the record with his 62nd home run on September 8 and then withstood a final charge by Sosa (who also broke Maris' mark) to post a new season record of 70 home runs.

As exciting as the race was, I still felt that there was something lost. Home runs were being counted as in a home run derby with no consideration given to the value of these home runs within the context of the game. Ultimately, the power of a home run is reflected in how much it alters the course of the game to produce victory.

We know that McGwire has the new record for the most home runs in a season, but were these home runs more valuable than those of the runner-up Sosa? Sammy Sosa played on the Chicago Cubs, a wild card team, while McGwire's St. Louis Cardinals did not make the playoff cut. Did McGwire hit more home runs, but ones with less value than Sosa's? In fact, using play-by-play data, we can get answers to these questions.

The most obvious comparison of home run value is the number of runs scored. Table 1 presents the number of home runs hit by each player with respect to the number of runners on base. A quick calculation from the table shows that 118 runs scored from McGwire's 70 HRs (1.69 R/HR) while 108 scored from Sosa's 66 HRs (1.64 R/HR). However, a goodness of fit test indicates no significant difference between the distributions of the players ($p = 0.56$).

Table 1. Number of Home Runs by McGwire and Sosa for Different Numbers of Runners on Base

	Number of Base Runners			
	0	1	2	3
McGwire	33	28	7	2
Sosa	37	19	7	3
Total	70	47	14	5

The value of a home run is not just the number of runs it produces, but is also a function of the situation in which it is struck. The situation or state of the game can be expressed as the inning, score, and outs as well as the base situation. Analyses of innings, outs and score provide little support of any difference between McGwire and Sosa. For example, Table 2 presents the distribution of their HRs with respect to inning. A goodness of fit test provides little support for a significant difference between the two players ($p = 0.21$). We note, however, that McGwire and Sosa both hit 66 HRs in the first 9 innings; the difference in their totals is that McGwire hit 4 HRs in extra innings while Sosa hit none.

Table 2. Number of Home Runs by McGwire and Sosa in Different Innings

Batter	Inning			
	1-3	4-6	7-9	10+
McGwire	19	24	23	4
Sosa	22	25	19	0
Total	41	49	42	4

One metric which encapsulates the main characteristics of the state of the game is the probability of victory given the inning, score, outs, and base situation. When a player comes to bat, the game is in State A which can be assigned a probability P_A of victory for his team. The effects of the at bat change the game to State B which can be assigned a probability P_B of victory. The value of the player's at bat with respect to probability of victory is then $P_B - P_A$. As an example, let's look at that historic 62nd home run hit by McGwire on September 8. When McGwire came to bat in the bottom of the fourth inning off Steve Trachsel of the Chicago Cubs, his Cardinals trailed the Cubs 2-0. With bases empty and 2 outs, the probability of a Cardinal victory was $P_A = 0.249$. His historic solo home run made the game 2-1, raising the probability of victory to $P_B = 0.368$. The increase of $P_B - P_A = 0.119$ in the probability of victory is very close to the median value for all of McGwire's home runs in 1998. The total value of McGwire's home runs would be the sum of these changes in probability given the game context in which each home run was hit. For more details on how these probabilities may be used to assess the value of player performance, see Bennett and Flueck (1984), Bennett (1992, 1993, 1998), and the SIS web site (www.stat.duke.edu/~box/sis/).

CONTINUED ON PAGE 4

The Value of Home Runs: McGwire Versus Sosa in 1998 (Cont.)

Figure 1 presents a boxplot of the distribution of the home run values for McGwire and Sosa. The distributions of HR values clearly are skewed. One home run definitely catches the eye. McGwire HR#38 was a game-winner in extra-innings. What made the HR especially valuable was that McGwire hit it in the bottom of the 11th inning with his team *trailing* Houston by a run. With one out and a runner on first base, the Cardinals had a 0.204 chance of victory. The play won the game giving McGwire a 0.796 value HR.

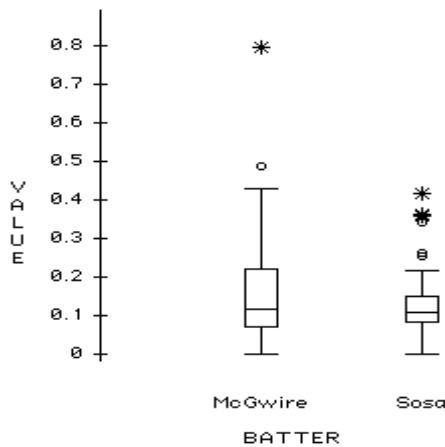


Figure 1. Distributions of Home Run Values for McGwire and Sosa

Table 3 presents summary statistics for the HR value distributions. Not only did McGwire hit more HRs than Sosa but the average value of McGwire's HRs was greater than that for Sosa's HRs. When 1,000 random samples (size 70, without replacement from the values of all 136 HRs) were each summed, only 47 of the 1,000 sums were greater than the actual sum of values for McGwire's HRs (11.155). This test indicates that McGwire's HRs had significantly greater value than Sosa's at the $p = .047$ level of significance.

Table 3. Summary Statistics for the Value of Home Runs by McGwire and Sosa

Batter	Count	Mean	Median	Std. Dev.
McGwire	70	0.159	0.120	0.142
Sosa	66	0.127	0.111	0.086
Total	136	0.144	0.115	0.119

Much of the difference between the HR value distributions lies in McGwire's extra inning HRs. If one looks only at the value of McGwire's home runs through the first nine innings the distribution is much closer to that of Sosa's HRs.

In summary, in 1998,

- Mark McGwire not only hit more HRs than Sammy Sosa but on average each of his HRs was more valuable than Sosa's to team victory. The difference in HR value is significant at the .05 level. However, no statistically significant difference was found in the distributions of HRs over various game situations. The dependence of the 1998 Cardinals on McGwire's HRs is reflected in their winning percentage: about 60% when McGwire hit a HR and less than 50% when he didn't. On the other hand, the Cubs had a similar winning percentage whether or not Sosa hit a home run.

- Both McGwire and Sosa hit the same number of HRs in the first nine innings. The difference in their seasonal total counts was McGwire's 4 extra inning HRs. These four extra inning HRs also account for much of the difference in distributions of HR value between McGwire and Sosa. McGwire's average HR value remains slightly higher than Sosa's even after deleting the extra-inning HRs.

References

- Bennett, J. (1993). "Did Shoeless Joe Jackson Throw the 1919 World Series?" *The American Statistician*, 47, 4, 241-250.
- Bennett, J. (1994). "MVP, LVP, and PGP: A Statistical Analysis of Toronto in the World Series," In *1994 Proceedings of the Section on Statistics in Sports*, American Statistical Association, 66-71.
- Bennett, J. (1998), "Baseball," in *Statistics in Sport*, ed. J. Bennett, London: Arnold, 25-64.
- Bennett, J. M. and Flueck, J. A. (1984). "Player Game Percentage". In *1984 Proceedings of the Social Statistics Section*, American Statistical Association, 378-380.

Statistics in Sports Website

<http://www.stat.duke.edu/~box/sis/>

Jim Box

Duke University Medical Center



The Statistics in Sports website has been operational since May of 1996, and since that time there have been over 10,000 visits to the main page. There are six main sections to the website:

Section Information – This area contains the section charter, a listing of officers, a news section, and information on the section's activities at the JSM

Publications – This area contains links to recently published articles concerning statistical analyses of sports and a listing of journals that have published sports-related articles.

Sports Data – This is an area with links to searchable databases and other sites that have data from 12 different sports.

Sports Websites – The most popular page on our site – it has links to the official homepages of most major sports leagues [men and women], the network's sports pages, sports servers such as ESPN.com and CNN/SI, and links to 69 newspaper sports pages across North America

Statistics on the Web – provides links to academic sites around the world, the ISDS statistics server, and a page of companies that employ statisticians.

Member's Forum – This is the place for SIS members to display their work. For the third year, Jay Bennett has published his PGP analysis of the World Series MVP race [surprisingly, it was the Padres' Tony Gwynn this year]. Recently we were linked to Scott Berry's analysis of the home run race [using Bayesian models]. This space provides an excellent opportunity for members to share their work.

Comments, suggestions and submissions are readily accepted – contact SIS webmaster Jim Box [jim.box@duke.edu] to give your feedback.

Sports Sessions at the 1999 JSM in Baltimore



Robin Lock, 1999
SIS Program Chair
Mathematics Department
St. Lawrence University

The Statistics in Sports portion of the program for the 1999 Joint Statistical Meetings in Baltimore is taking shape. Two invited paper sessions have been organized and additional sessions with contributed papers will be arranged after the Feb. 1 deadline for abstracts.

The first of the invited sessions will focus on education themes, "Using Sports Activities, Examples, and Data in Teaching Statistics". Karla Ballman from Macalaster College will open the session by discussing some of her experiences with "Sports-based Classroom Activities". University of Wisconsin-Madison's Bob Wardrop will focus on "Sports Examples for the Classroom". The SIS webmaster, Jim Box, will show us some of the "Sports Data Resources on the WWW". This session is scheduled for 4:00 p.m. on Sunday August 8th.

The second invited paper session deals with "Statistical Methods for Rating and Comparing Players". We'll kick off that session with Christopher White and Scott Berry (both of Texas A&M) presenting a paper on "Rating NFL Football Quarterbacks with a Markov Chain Model". Batting in the number two slot, Carl Morris from Harvard University will tell us about some "Better Ways to Evaluate Hitters" in baseball. For our closer, we'll bring in Bellcore's Jay Bennett to discuss "A Multidimensional Rating for Starting Pitchers". This session is scheduled for 10:30 a.m. on Wednesday, August 11th.

In addition to our two allotted invited sessions, we have at least one bonus session. Scott Berry, Shane Reese, and Patrick Larkey were drafted by JASA and awarded a special slot for their featured applied paper "Bridging Different Eras in Sports". Along with the SIS invited paper sessions and the special JASA session, we will have one or more contributed paper sessions, the annual SIS business meeting, the SIS Luncheon, and a short walk to watch the Orioles play the Tigers at Camden Yards. More meeting information, full abstracts, and an updated schedule of sessions can be found at the ASA's website <http://www.amstat.org/meetings/jsm/1999/index.html>.

Book Reviews



Statistics in Sport, Jay Bennett (ed.) Arnold Applications of Statistics Series. London. 1998

Statistica e Sport: non solo numeri. Antonio Mussino (ed.) Societo Stampa Sportiva. Roma. 1997. (in Italian)

Donald Guthrie
Department of Psychiatry
University of California, Los Angeles

These two recent books are collections of papers concerned with the statistics of sports. Both are carefully edited and produced, and each gives interesting insight into applications of statistics. The somber and boring cover found on most statistics books has been replaced by color photographs of Pete Sampras following through on a serve (Bennett), and the start of the Rome Marathon (Mussino). These have caught the eyes of many of my colleagues, and they foreshadow the vitality of the works contained within the covers. For the ASA-SIS readership, Bennett is the more readable, being in English; I must confess that I read very little Italian and therefore my review of Mussino must be somewhat superficial.

As is the tradition in English language publications on statistics of sport, the emphasis in Bennett is on modeling the sports as they are played. Part I includes nine chapters on individual sports -football, baseball, basketball, cricket, soccer, football, golf, ice hockey, tennis and athletics. Each chapter begins with a brief outline of the rules and conduct of competition and includes a summary of statistical literature on that sport. Reading through, one gains an appreciation for differences in quantification. Soccer football data, for example, are largely based on match outcomes, and lead to models of such issues as home ground advantage and effect of red cards. Baseball data, at the opposite extreme, is exquisitely detailed and isolated situations may be studied. Basketball has been the focus of the "hot hand" issue, and it provides the vehicle for that continuing debate. It seems curious to me that golf has apparently not been so analyzed. Strings of

birdies or bogies seem to appear in nearly every televised tournament. Each sport seems to have led to its own set of statistical models, indeed it seems that the models may tend to exploit the data which happen to be available rather than to develop knowledge of the sports. For example, there are few discussions of the effect of rules on outcomes.

Part II of Bennett consists of four chapters -design of tournaments, graphical displays, gambling and predicting outcomes, and hierarchical modeling. Rather than using statistical methods to interpret sports data, these papers use sports data to illustrate general statistical methodology.

Mussino's volume, in contrast to Bennett's, illustrates a broader, societal focus on participation which seems to prevail in international sports statistics research. The papers were presented in Rome in 1996 as the report of the Scientific Commission of the Italian Statistics Society on "The Statistical Analysis of Sports." In addition to the usual analysis of outcomes, papers deal with such topics as the implication of recent demographic tendencies for the Italian sport system, economic aspects and sports in the context of everyday life. The range and depth of coverage is commendable. Indeed, one paper on outcomes written by Mussino et al. contains one of the most complete bibliographies on sports statistics I have seen.

Both volumes are carefully planned and executed. The papers in Bennett are easily readable by anyone with some background in statistics and interest in sports. While I cannot comment on the quality of the prose in Mussino, it is apparent that the papers are very carefully done. They are illustrated beautifully, and their scope is comprehensive. Combined, they represent the scope of international activities. Relatively little appears in ASA programs and publications about participation, but relatively little appears in ISI programs and publications about factors affecting outcomes.

With the proviso that one should be able to read the language, I can recommend both to anyone interested in quantitative interpretation of sports. Both volumes include papers with extensive bibliographies (in the language of the source references), and therefore provide excellent starting points for broadening of your interests.



What's New?

Scott M. Berry, Department of Statistics, Texas A&M University

What's new? Well, this column is new. Here I will provide information on recent research in sports statistics. This may be formal papers appearing in refereed journals or web sites which report on interesting analyses. I try to keep up with the sports research being done, but I can not catch it all. Please let me know of research being done on statistics in sports and send me copies of papers, preprints, and web sites.

Recent Articles:

Chance is always a good source for sports articles. Hal Stern wrote some great articles in his "A Statistician Reads the Sports Page," column. Hal's article in the most recent *Chance* (Vol. 11, No. 4) investigates how accurate the posted betting odds are in horse racing, American football, and baseball. All these posted odds are a reflection of the betting public, and in each case there is very good agreement between the betting public's odds and the results of the competition. Strangely enough all sports betters think they are smarter than average—Garrison Keillor would be proud!

In the same issue of *Chance* (Vol. 11, No. 4) Christine Anderson-Cook and Tim Thornton have an article in which they redefine the way ice hockey (National Hockey League) special teams are classified. They model goal scoring as exponential and model the parameters for each team while they are on a power play or while they are short-handed. They present some differences in team ratings from the standard success proportions in which the NHL reports.

Like *Chance*, *The Statistician* (JRSS-D), is a great source for sports articles, especially if you are a cricket and/or soccer (oops,...football) fan. In the most recent issue (Volume 47, Part 3, 1998), Howard Grubb has an article analyzing the times for athletic (running) events. He looks at the average speed for the runners for different length races. Parametric models on the speed vs. distance relationship are used to predict lower bounds on world records. Mark Dixon and Michael Robinson develop Poisson models for goal times in professional football (soccer) matches. Offensive, defensive, and home field parameters are estimated for each team. They also present a birth process model for scoring which allows the probability of scoring a goal to depend on the current score.

The electronic journal *Journal of Statistics Education* (<http://www.stat.ncsu.edu/info/jse/>) often has sports article that are useful in teaching. In the current issue, Jeff Simonoff provides an exploratory view of the 1998 McGwire and Sosa homerun data. The dataset is easily downloadable and Simonoff suggests a number of different ways for a student to analyze the data.

Web Sites:

In keeping with ice hockey, Robin Lock has a web site <http://it.stlawu.edu/~chodr/> which rates NCAA Division I Men's hockey teams. He uses a Poisson distribution to model scoring and each team is categorized by their offensive and defensive ability. He updates the site frequently throughout the season and even predicts the scores for games in the upcoming week.

Jeff Sagarin's computer rankings for NCAA football teams (www.usatoday.com/sports/sagarin.htm) were used to determine the Bowl Championship Series rankings. While he is nationally known for his football rankings, he has rankings for many sports (both player and team).

An internet source for interesting discussion of statistical ideas is Rob Neyer's column at ESPNnet at <http://espn.go.com/mlb/columns/neyer/>. His comparisons of players and strategies are very well done.



Statistics in Sports



INSIDE:

The View From The Chair on the Bench 1

Sports Statistics News From Down Under 1

Where Should I Send That Sports Article? 1

The Value of Home Runs: McGwire Vs. Sosa 3

Statistics in Sports Website 5

Sports Sessions at the 1999 JSM 5

Book Review 6

What's New? 7

Newsletter Editors

Jim Albert
Bowling Green State Univ.
albert@math.bgsu.edu

Scott Berry
Texas A&M Univ.
berry@stat.tamu.edu

1999 SIS Officers

Chair:
Robert Wardrop
University of Wisconsin
wardrop@stat.wisc.edu

Program Chair:
Robin Lock
St. Lawrence Univ.
rlock@vm.stlawu.edu

Secretary/Treasurer:
William Kaigh
Univ of Texas-El Paso
willamk@laguna.epcc.edu

Publications Officer:
Jim Gentle
George Mason Univ.
gentle@gmu.edu

Council of Sections Rep:
Katherine Hansen Simonson
Sandia National Laboratories
kmsimon@sandia.gov

Section Webmaster:
Jim Box
Duke Medical Center
jim.box@duke.edu