Newsletter of the Statistics in Sports Section
of the American Statistical Association

Volume 2, Issue 1

Spring 2000

STATISTICS IN SPORTS

THE VIEW FROM THE CHAIR ON THE BENCH

Jim Albert, 2000 SIS Section Chair

In last year's newsletter, Bob Wardrop talked about the use of sports examples in the teaching of statistics. At Bowling Green, our department teaches an intro statistics class to hundreds of students each semester. One difficulty in teaching this course is that students generally do not relate to many of the examples discussed in textbooks, and so they don't see the value of statistical methodology in these applications. Many students are interested in sports, either as observers or participants, and they can make sense of statistical concepts when applied to illustrations in sports. This fall I'm trying something new. I am teaching one section of our intro course entirely from a baseball perspective. In this "Baseball Statistics" class, I will cover all of the usual topics (data analysis, probability, and inference) using examples from baseball. Hopefully the students will pick up statistical concepts while they are learning how to properly interpret baseball data. And certainly the course will be a lot of fun, which is something that can't be said about the usual statistics class.

I encourage you to attend the SIS Annual Section Meeting in Indianapolis in August. There will be many items on the agenda, including the future of our Proceedings and discussion about a Statistics in Sports journal. I believe that a journal would greatly stimulate research in sports statistics, and our section should provide leadership in the development of this journal.

INSIDE:

STATISTICS IN SPORTS AT JSM 2000

Bob Umholtz, 2000 SIS Program Chair

SIS has two invited paper sessions this year. The first session, **Simulation and Modeling in Baseball**, is scheduled for **Monday, Aug 14 from 2:00 – 3:50 pm**. The authors will discuss several interesting applications of simulation, including assessing the power of baseball schedules to discriminate among evenly matched teams, prediction of final season batting averages based on early season averages, and prediction of baseball outcomes based on player statistics.

The second session, scheduled for **Wednesday, Aug 16 from 8:30 – 10:20 am**, also focuses on simulation and modeling, but examines a wider variety of sports. Modeling, prediction, and resampling of football games and analysis of inter-rater agreement in boxing matches will be discussed.

There will be one **Regular Contributed Paper Session**, scheduled on **Tuesday, Aug 15 from 2:00 – 3:50**, as well as one poster session (date and time to be determined). Nine speakers will discuss rating, ranking, and modeling in professional and college sports. The presentations will cover a wide range of team sports, including baseball, football, basketball, soccer, and hockey.

David Harville, this year's SIS Section Award recipient, will be the featured speaker at the **Statistics in Sports Luncheon** on **Monday, Aug 14 from 12:30 – 2:00**. David will discuss selection and seeding of college basketball and football teams for postseason competition from the statistician's perspective.

THE MAINSTREAM SPORTS MEDIA

Scott Berry, Texas A&M

Over the last year I have had some interesting experiences with the media and non-statistical outlets. Being an assistant professor in the sheltered world of academia this has been an enlightening experience for me. In this article I describe some of what I have learned about the media and general public.

In the September, 1999 issue of *The Journal of the American Statistical Association*, Shane Reese, Pat Larkey and I published a paper comparing players from different eras in baseball, golf, and hockey. We set up models to account for year, age, individual ability, and stadium effects (round difficulty in golf and home field effects in baseball). By estimating these effects we were able to discuss the relative ability of players, and estimate how different players would do in different eras.

Various newspapers and talk shows took an interest. I did interviews for about 10 radio talk shows and gave phone interviews for about a dozen newspaper reporters. Interestingly, I got calls from Europe, the United States, and Canada. Almost without exception Europe called to discuss golf, the United States called to discuss baseball, and Canada called to discuss hockey. The interviews on radio talk shows were quite interesting--I had very little idea what questions the hosts were going to ask. On one show I was asked to, briefly explain my "formula"--invariably the media assumed I had a simple formula to plug numbers into. I gave the best 3 minute description of what we had done--on a very understandable level--I even thought I was quite eloquent. The response from the host was... "oh, so you used a *computer*." The ideas of how I did it had no meaning--all that mattered was that I used the magic box on my desk! In one example, with a Vancouver sports radio show, I spent 30 minutes on the line and even answered questions from callers. The *National Post*, a nationwide newspaper in Canada had run a story about the article, with the headline "Scientists Prove Lemieux is the Real Great One." Well, despite the fact that both Lemieux and Gretzky are from Canada, Canadians were very upset. I tried to explain to one caller that we are claiming that given both players at their peak, in the same season, with the same teammates, Lemieux would score about 7 points more in a season than Gretzky--we were not saying that Lemieux had accomplished more than Gretzky in his career. I said it was quite the opposite, that Gretzky may be the most accomplished player in any sport. The caller responded that my computer couldn't understand "heart"--it couldn't measure the will-to-win for a player. I learned for the most part that I couldn't argue with hosts or callers--I had to state my points and be happy with that. On the same show the host said to me "How can the top 25 hockey players of all-time not include Bobby Orr?" Here I faced a dilemma--Bobby Orr was a defenseman--and we didn't include defensemen in the study. I also didn't want my list of "Best scorers at the peak of their career" to be confused for the "best hockey players of all-time." I chose to explain that we only studied forwards--defensemen were not allowed. I thought the host was pleased with that answer. After I hung up and listened over the internet, the host commented "Wow, a list of the top 25 players of all-time and they don't even consider defensemen!" I found that callers and hosts had their own agendas and beliefs and I could rarely change any minds--and I could never get the last word!

Credibility is also a strange thing. I sound quite young (thank goodness they can not see me, I look 18), and I think that this hurts my credibility. The letters "Ph.D." and "Professor" seem to bring instant credibility to me--and thus hosts would always mention that I had a Ph. D. or that I am an assistant professor. I got a call--and this was a bad side effect of the media attention--at 6:00 am--from a radio station in Winnipeg, Canada. They asked me some questions about the study, but were generally disbelieving. The host finally asked me his big worry-- "What does some guy in the middle of Texas know about hockey anyway?" I explained that I grew up in Minnesota and played hockey for 25 years. "Oh, well, that's essentially a province of Canada." I was instantly okay, and the study made more sense. I don't think the skepticism is necessarily bad. I teach my students to be very skeptical of "studies" they hear about. It was just strange for me to experience the defensive side for the first time.

In many ways radio shows are easier than speaking to writers. On the radio I get to say everything in my own words. It is much more difficult when speaking to writers because I could be "half-quoted" or crucial information would be left out of the article. Any time I made a statement that had a condition to it--invariably that condition would not be in the article. I gave an interview to a reporter for the Fort Worth Star-Telegram. He asked me about the golf study--and in particular--he

asked me about Tiger Woods. The data we used for the study went through 1997, the year Tiger won the Masters by 12 shots, as a 21-year-old. My favorite aspect of the model we used was that it had individual aging functions for each player--and these were hierarchical. If Tiger Woods aged like the average 21-year-old he would be 2 shots better than any player in history by the time he is at his peak. We concluded that it was much more likely that he was closer to his peak than the average 21-year-old. I thought this was the most incredible result from the model. What did the article have to say about this: "Woods is as good as he'll ever be." Surely part of this misinterpretation problem is my fault. I realized that I was not talking to a Ph.D. in statistics, but I generally overestimated the writer's understanding of what I had said. I learned to ask for a copy of the article before it is submitted. Not all were willing and able, but some did.

While some of the attention was nice, and I would be lying if I said I did it all for the good of statistics, I do have the good of the science of statistics on my mind. It is nice to have the opportunity to explain random variability to the media. In my first radio interview I was asked to predict the number of home runs Mark McGwire would hit in the 1999 season. I said 59. I was then asked "when you do those analyses and it says 59... does that take the fun out of the season for you?"

I am now experiencing another side of the media. I frequently write some short pieces for *ESPN the Magazine*. It started as calculating probabilities for certain events related to the theme of the *Pulse* section within the magazine. For example, could I find the probability that Mark McGwire breaks the 70 home run plateau? In the second article they asked me to find the probability that a woman plays in the NBA in the next 10 years. I think they were surprised that I refused to answer that one! While I enjoyed the attention, I didn't want to belittle my profession by answering such questions that were clearly 100% non-empirical. It would do harm to statistics if I did answer such questions, claiming in any way that it was a statistical analysis. It would also belittle the amount of effort and modeling I use to answer the other questions.

In this effort, it was important to read anything that is attributed entirely to me. During the 1999 baseball season, 10 players with 10 or more seasons of 300 or more at-bats, were on pace to post their highest batting average for their career. Interestingly, only one player in history has had their highest single season average in their last (Dots Miller). They asked me to find the probability that each of these players would have their highest average in the 1999 season and I did a simple model based analysis. When I saw the *Chances* section, it read that I had calculated the probability that each player would have their highest average--and RETIRE! When I told them I had not done that, they said... "well can you fix that, and by the way, by tomorrow at noon?"

The editors of *ESPN The Magazine* seem to like the stuff I have done, and feel statistics is important, and that the future of sports journalism will contain more "analysis." But, I still don't know the impact the stuff I have done with ESPN and the general media will have--if any--on statistics. Does our field become more or less respected? I do know that it helps me with teaching--especially in the motivation for some students--but also in research. If I have calculated a p-value of .02 and I am explaining it to a journalist, and it is not clear what it means--is it my fault, the journalist's fault, or possibly the p-values fault? Every statistician will tell you that answering the important questions is a main goal of statistics...is my research addressing the important questions? I will also be asked a very logical question, like, "What is the probability that we see another home run hitter of McGwire's ability in the next 20 years?" I usually have to say that this is too hard--is it? Am I working on the right set of problems? I realize this is extrapolation, but many interesting and important questions are--should I be doing different research? In a clinical trial should I be asking "what is the probability the next patient survives?"

So what have I learned that I can pass on to you? First, you have to be prepared to summarize things, neatly, and correctly. I found this to be very difficult--and also very frustrating. In any analysis there are certain assumptions that are important and they greatly affect the conclusions. Invariably the sports media is "headline driven." They want to say... "the statistician found that this player is the best." While I want to explain how it was found, how best is defined, and what are the measures of uncertainty, the media is for the most part happy with the headline. It is frustrating not being able to have more time to explain things more clearly, maybe I can get a few people to read the original study, or better yet become interested in this interesting field of statistics. You should *strongly* encourage anyone who writes about your work to let you read it--it will benefit both of you. Beware of someone who calls you and says "I have an *easy* question for you..." It is bound to be far from an easy question!

INTERNATIONAL SOCCER RANKS AND RATINGS

Albyn C. Jones, Reed College

Introduction

Going in to the 1998 World Cup in France, the US clearly was faced with a tougher challenge to advance to the elimination round than it had faced at home in 1994, if only because the increased tournament size (32 teams instead of 24) meant that only the top two teams from each 4 team bracket would advance. Furthermore the US was in a bracket with two highly regarded teams: Germany and Yugoslavia. The US soccer press, and probably the US coach, Steve Sampson, thought that the US was a clear favorite over Iran, the fourth team in the bracket. Here are a couple of typical quotes from the soccer press, in this case Jerry Langdon of the Gannett News Service, (June 11, 1998):

"Iran is the fourth team in Group F, and widely seen as an underdog in all three games."

"Iran barely qualified as the fourth representative from the Asia region, and appears over its head."

Apparently as a result of this analysis, Sampson restructured the team from a `4-4-2' lineup, with 4 defenders, 4 midfielders, and two attackers, to a more defensively oriented `3-6-1' alignment. The idea was to increase the chance of getting a draw against one of the stronger teams. The expected win against Iran, together with a lucky win or draw against one of the favored teams, could put the US into the second round. Following a humbling 2-0 loss to Germany, the 2-1 loss to Iran eliminated the US. The final loss to Yugoslavia was no surprise at that point, nor was the June 29 Soccer America headline: Steve Sampson resigns.

In fact, Sampson was dealt a losing hand, and did about as well as could be expected with it. Why the aura of disgrace? It stems from inflated expectations - the US soccer community was simply expecting too much from a mediocre team. What was the source of these expectations? One source was the official FIFA/Coca Cola rankings. FIFA stands for `Federation Internationale de Football Association', the international soccer governing body; you probably know what Coca Cola stands for. The May 1998 rankings show the US in 11th place, not far behind Yugoslavia (8), and ahead of traditional European soccer powers Italy (14), Spain (15), and France (17), and the Netherlands (25). Iran (42) was way down the list, apparently undeserving of their slot in the 32 team tournament.

Ranking and Rating systems

Ranking and rating systems have numerous uses. Perhaps the most basic use is to give participants a yard-stick with which to measure their performance; additionally they provide a rational basis for invitations to championship events, or for seeding or pairing contestants in tournaments.

In the following discussion, I will distinguish between ranking and rating systems. A rating system yields predictions either for the probability that one contestant beats another, or for the expected score. A ranking system simply produces a rank ordering of the contestants. Rating systems yield ranks as a by-product. There are numerous rating systems in use in various games and sports. Perhaps the best known is the Elo rating system developed for chess (Elo, 1978, Glickman and Jones, 1998) which is based on the Bradley-Terry, or logistic model.

The FIFA ranking system

The FIFA system awards points to teams based on their results against other teams. The system was apparently designed to mimic the usual league standings, which are based on 2 or 3 points for a win, 1 for a tie, and 0 for a loss. The number of points depends on the following factors:

- the match result (win, loss, or tie)
- number of goals scored
- home/visitor
- match importance
- strength of opponent
- opponent's region

Most of these enter additively, but match importance (from 1.0 for a friendly match to 2.0 for a World Cup match) and opponent's region (1.0 for Europe and South America, dropping to .84 for Oceania) enter multiplicatively. The points earned for a win, loss, or

tie depend on the difference in team strengths as measured by previous results, then points are added for goals scored, subtracted for goals conceded, and added for the visiting team. If negative, the total is set to zero. Finally, the match importance factor and the opponent's regional factor multiply the total.

Results are combined in a yearly average as follows: for a team playing N games, compute the total points (T_A) for all games, compute the point total (T_B) for the best 7 results in the year, and then combine

$$\text{Points} = 7\ T_A /N + T_B$$

Finally, the current points are averaged with the previous 7 years, giving each year weights $1 - j/8$ where j ranges from 0 for the current year to 7 for the earliest year included. This averaging is done to reduce the effect of team activity; until recently the ranks were based on the simple total T_A, so the most active teams had the highest totals. The current FIFA algorithm reduces this effect, but does not eliminate it. A team that plays 20 matches and wins 10 of those may easily receive more points than a team that plays 7 matches and wins 5, in spite of achieving a much lower proportion of wins. In 1998 FIFA was still using the simple total within years. For more details, see the online documents at the FIFA website (http://www.fifa.com/).

Even though the rankings depend on recent years as well as the current year, the effect of current activity is visible in the May 1998 FIFA ranks: the US (ranked 11th) played 12 matches through May of 1998, while Iran (42) played 5, France (17) played 6, and Spain (15) and Italy (14) played only 2.

Leaving aside several interesting questions about how various values and weights were chosen, the single most important feature of the FIFA system is that it is not based on a probability model linking results to team strength, and thus has no justification based on principles of statistical inference. The FIFA system does not generate predictions, except based on ranks: there is no implied metric.

A Rating System

For the past several years I have been rating international soccer teams using a Bayesian dynamic generalized linear model (http://www.reed.edu/~jones/ratings/). The model represents trinomial outcomes (win, loss, draw) using a standard parametrization:

$$\textit{logit}(\text{Prob}(A \text{ beats } B) = \exp(\beta\ (\alpha\ NS + R_A - R_B - \delta))$$

where A is the home team, B is the visiting team, R_A and R_B are the ratings for A and B respectively. α represents the home field advantage, NS is the neutral site indicator variable, δ determines the probability of a tie, which will depend on the rating difference, and β scales the ratings so that if $\delta = 0$, then a 100 point rating difference yields odds of 2:1 at a neutral site.

Based on this rating system, The US men's team was ranked essentially even with Iran at the end of 1997 (Iran was 41st, the US 42nd). Both Germany and Yugoslavia were in the top 10, with ratings suggesting respective odds of roughly 8:1 and 4:1 against the US. Not only was the US not a clear favorite against Iran, but they had only about a 15% chance of advancing to the second round, and nearly 1 chance in 3 of losing all 3 matches. In other words, the actual outcome (losing all three matches), while disappointing, should not have been a surprise.

REFERENCES

Elo, Arpad E. (1978), "The Rating of Chessplayers, Past and Present", Arco Publishing, Inc: New York.

Joe, H. (1990), "Extended Use of Paired Comparison Models, with Application to Chess Rankings", *Applied Statistics*, 39, 85-93.

Glickman, Mark E. (1995), "A Comprehensive Guide to Chess Ratings", *American Chess Journal*, 3, 59-102.

Glickman, Mark E. and Jones, Albyn (1998), "The United States Chess Federation rating system: current issues and recent developments.", Proceedings of the American Statistical Association, Section on Statistics in Sports.

Glickman, Mark E. (1998), Parameter estimation in large dynamic paired comparison experiments, Submitted to *Applied Statistics*.

Glickman, Mark E. (1993), Paired Comparison Models with Time Varying Parameters, Ph.d. thesis, Harvard University, Department of Statistics.

Glickman, et al (1994), USCF Ratings Committee Report, unpublished.

INTERNATIONAL MEETING OF COMPASS

Don Guthrie, UCLA

In November 1999 I was honored to be invited to attend a meeting in Rome of COMPASS-99 (COrdinated Monitoring of PArticipation in SportS), the second of a series of conferences dedicated to the development of a uniform scheme for the measurement of participation in sports and related physical activities. The goals of COMPASS are the specification of reporting standards for surveys of activity in the European nations, and formulation of cross-national comparisons of the levels and structures of sports participation. The meeting was organized by Professor Antonio Mussino of the University of Rome (La Sapienza), and supported by the Italian Olympic Committee (CONI), the Italian National Statistical Institute (ISTAT), and UK Sport and Sport England. About forty participants represented most of the EU nations. Observers came from the Czech Republic, Japan and the US (me).

The agenda began with a summary of results obtained from recent national surveys. An initial questionnaire was sent to 44 member countries of the Council of Europe. Nineteen responded that they had collected or were collecting appropriate survey data, sixteen responded that they did not have data, and nine did not respond. Detailed cross-national comparisons were made of participation among seven countries–-Finland, Ireland, Italy, Netherlands, Spain, Sweden and the UK. These results required close cooperation among the statistical offices and the major sports agencies in these countries.

Participation was measured on a seven point scale from "competitive", through "intensive" to "non-participant with other physical activities" and "no physical activities". Italy and Spain have the lowest level of participation, followed by UK, Netherlands and Ireland, with Sweden and Norway the highest. Two striking findings from these data distinguish the Nordic countries, the continuing participation by older people and the higher participation levels of women. One might argue that the most prevalent sports activities in the Nordic countries are more accessible to older people and women, but it is also apparent that these countries encourage these participants. For example, only 23% of Finnish women were classified in the lowest two categories of activity, whereas 62% of Irish women and 88% of Italian women were. The delegate from Portugal presented similar data from a recent survey indicating results consistent with the cross-national comparisons with age and gender trends similar to those in Italy and Spain. Data presented by the Japanese observers suggest that women participate in near parity with men, but these data were not gathered in a comparable form for comparison with the European countries.

The present national systems, while often useful for gross comparisons, lack consistency of terminology and sampling methods, therefore the long-term aim of COMPASS is the 'harmonization' of sports participation surveys in Europe. Uniform reporting involves several careful definitions to maintain validity across nations and cultures. Considerable attention was paid, for example, to the definition of what constitutes sport activity. At one extreme was the detailed tabulation by Statistics Finland of every level from international football competition through showshoeing to mushroom gathering as activities. Vigorous discussion led to more definite specification of the levels of participation, and to the definition of analysis categories. Working groups were appointed to recommend decisions for future COMPASS meetings.

Being the only American at the meeting gave me the opportunity to observe several things. First, one is impressed with the intensity and depth of involvement of each of the participating countries. For example, the CONI contributed very substantial financial and logistic support. Also, the European Union central office in Brussels sent a delegate who clearly reflected the importance of the project there. Second, one gains a feeling of cooperation that might be hard to find in the U.S. I was asked informally if the U.S. had such a program, and I had to admit that I know of none since the decline in the President's Council on Fitness. Indeed, it is hard to imagine this level of cooperation among the states. Perhaps the COMPASS model will facilitate our development of similar surveys. Finally, I was surprised by the level of encouragement given within the countries. In the UK a significant portion of the proceeds of the national lottery are dedicated to development of sports programs and facilities. These funds permit the development of sports programs and facilities from organized football leagues to recreational parks and trails. I realize that we also support these activities, but it seems that our organization is less centralized and coordinated. Without a doubt, the participation patterns across the

COMPASS (CONT.)

several nations in COMPASS are primarily influenced by cultural and geographic considerations. It would be interesting to see whether the COMPASS model, applied in the U.S., would produce meaningful contrasts with the European countries. The developed categories are general enough to be applicable and the methodology seems tractable for a national survey of our activity.

Finally, I should acknowledge the cordial hospitality of the hosts and organizers. The program was tightly structured, but time was left for social and cultural activities. My brief talk on baseball statistics was well received, but perhaps best understood by the Japanese visitors. The meeting was held at the Olympic training facility, and it was enlightening to note that the first field one sees at the entrance is a baseball diamond!

WHAT'S NEW?

Scott Berry, Texas A&M

In this column I highlight some of the research and innovations related to statistics and sports. The internet has had a huge effect on statistics in sports--not so much in new techniques or problems, but in the access to data sources. The internet has also made the presentation of analyses easier. One such interesting presentation is at Brigham Young University-- statweb.byu.edu/larc/weblarc.exe. This site has team rankings for college football, college basketball, the NHL, and the NBA. Their ranking method is a "Bayesian" alteration of the Bradley-Terry model. One of the fascinating aspects of this site is that it automatically updates its data every night! Some sort of messenger program--neat! The best site for sports statistics is cnnsi.com. They have the stats broken down in many ways. Everyone interested in the hot hand effect will find game-by-game stats for the NHL, MLB, etc.

Chance is still the best source for statistics in sports papers. Each issue contains a column titled, "A Statistician Reads the Sports Pages." There are also featured articles. One such article is by Wainer, Njue, and Palmer, "Assessing Time Trends in Sex Differences in Swimming & Running" (Vol. 13, No. 1, 2000; with discussion). They fit models which have women trailing men by a certain percentage and a shift in years. The models seem to fit pretty well, and seem to have a lot of face validity. Rather than gender affecting race times, Martin and Buoncristiani (Vol. 12, No. 4,1999) study the effect of temperature on Marathon race times. They find an optimal temperature of 55 degrees (F) for men, and interestingly the women's optimal temperature may be less than that--they give interesting biological reasons why--stay tuned for more.

Stern and Carlin (Vol. 12, No. 3, 1999) discuss the issue of a college football playoff system. They discuss what a possible system would look like, and how many teams are needed to properly ensure the best team wins. They conclude the top 8 teams have about a 90% chance of winning a 32-team playoff...It sounds so much more exciting than the current 2-team playoff. Why don't we have that yet? In the same issue Blumstein and Benedict discuss the rates of criminal violence for NFL players compared to the general public. Wow, with the problems Lewis and Carruth have had this is a very interesting look at a serious problem for the NFL. They find evidence that NFL players have a lower arrest rate than the general population (controlling for race and age).

Magnus and Klaassen (The Statistician, Vol. 48, No. 2, 1999) have two papers discussing certain effects in tennis--do new balls matter and is it an advantage to serve first. They have an amazing data set of 4 years worth of point-by-point data at Wimbledon. In The American Statistician (Vol. 53, No. 2) Smith and Schwertman analyze the NCAA basketball championship tournament. They model the point spreads in games based on their seeding in the tournament.

Michael Schell has a book out, “Baseball's All-Time Best Hitters", published by University Press, 1999. He analyzes batting averages over time, addressing the question ``Who is baseball's best batting-average hitter of all-time?" (He concludes it is Tony Gwynn)

For those of you looking for a statistics in sports journal--not yet--but the "Journal of Sports Economics" is a new endeavor for Kahane (Cal State) and Idson (Columbia). I look forward to interesting articles. If the journal enjoys great success it may possibly lead to a statistics in sports journal?

ASA Section on Statistics in Sports

Newsletter Editor:
Jim Albert
Bowling Green State University

SIS Website

http://www.amstat.org/sections/SIS

2000 SIS Officers