A Graduate Course in Statistics in Sport

Swartz, Tim
Simon Fraser University, Department of Statistics and Actuarial Science
8888 University Drive
Burnaby BC V5A1S6, Canada
E-mail: tim@stat.sfu.ca

1. Introduction

There have been various efforts over the years in teaching introductory courses in statistics based on examples from sport (Reiter 2001; Albert 2002; Kvam and Sokal 2004; Lock 2006). However, to my knowledge, there are only two graduate level courses in statistics departments that have a primary focus on sport. STAT832 (Statistics in Sports) at the University of Nebraska-Lincoln is a two-credit course which is offered every other year. It is described as a course which introduces statistical methodology useful for analyzing sports-related data. This paper describes aspects of a full-credit special topics course STAT890 (Statistics in Sport) that was first taught at Simon Fraser University during the summer semester of 2004. A description of STAT890 is provided by navigating to the Teaching link at www.stat.sfu.ca/∼tim.

2. Rationale

In contemplating a graduate course on statistics in sport, I wanted students to gain experience in three areas:

1. statistical modelling
2. the reading of scientific papers
3. statistical methodology

These three areas might be explored adequately by a type of Data Analysis course that is common in many statistics graduate programmes. However, it is my belief that a focus on sports related topics can provide an enjoyable and direct route to experience in the above three areas as now argued.

2.1 Statistical modelling

In order to propose realistic statistical models, it seems a prerequisite to have at least some understanding of the underlying physical processes in a system. Even in the case of black box methodologies such as neural networks, one's level of understanding must be sufficient to collect meaningful covariates.

A feature of sports related topics is that we often have immediate intuition of the underlying physical processes. For example, contrast a typical sporting event (say, basketball) with some biological phenomenom (say, cancer). In basketball, we typically know what to measure in terms of performance (e.g. score differential) whereas in cancer research, it is not clear what should be measured as there is uncertainty whether malignant stem cells are the root cause of cancerous growth (O'Brien, Pollett, Gallinger and Dick 2006). Most students have at least a vague understanding of what covariates are important in basketball (e.g. assists, rebounds) and they do not require hours of instruction before models can be proposed. We contrast this with the enormous amount of background instruction needed in basic genetics and physiology to have an appreciation of the issues related to cancer research. Generally speaking, our potential to model sporting events is due to the simplicity of games (e.g.

well defined rules and activities of short duration) when contrasted with the complexity of biological systems.

Statistical modelling is facilated when there is a scientific question of interest, and sport is full of such questions. For example,

- Who is the better player?
- Which lineup is preferred?
- What is the probability of an event occuring?
- What is an optimal strategy in a given situation?

are questions that could apply to a number of sports.

Another factor that facilitates statistical modelling in sport is the wealth of data that is available in almost every sport imaginable. For example, ball-by-ball results of international one-day cricket matches for over 100 years are recorded in the CricInfo website (www.cricinfo.org).

2.2 The reading of scientific papers

My experience is that students in graduate programmes in statistics often have a difficult time reading scientific papers. Yet, it is clear that reading papers is a valuable activity and forms an important component of graduate level training. To that end, I believe that it is helpful if a student's introduction to reading articles begins with papers that are relatively easy to read. Statistics in sport papers often satisfy this requirement. A statistics in sport paper typically begins with a question that is easily understood, and often, standard techniques are utilized to explore the question of interest.

In STAT890, the papers listed below formed a reading list for the course. The volume of papers exceeded the amount that I have asked of students in other graduate courses yet the students were able to complete their reading assignment without great difficulty. They found this to be an enjoyable activity and gave them confidence in reading papers.

- Verducci, T. (2004). Welcome to the new age of information. Sports Illustrated, April 5, 50-62.
- McKeon, J. (2004). This is the ultimate? Bull! Sports Illustrated, April 5, 67.
- Blount, R. Jr. (2004). As so often happens. Sports Illustrated, April 5, 68-73.
- Beaudoin, D. and Swartz, T.B. (2003). The best batsmen and bowlers in one-day cricket. South African Statistical Journal, 37, 203-222.
- Insley, R., Mok, L. and Swartz, T.B. (2004). Practical results related to sports gambling. The Australian and New Zealand Journal of Statistics, 46, 219-232.
- Tversky, A. and Gilovich, T. (1989). The cold facts about the "hot hand" in basketball. Chance, New Directions for Statistics and Computing, 2 (1), 16-21.
- Larkey, D., Smith, R.A. and Kadane, J.B. (1989). It's okay to believe in the hot hand. Chance, New Directions for Statistics and Computing, 2 (4), 22-30.
- Tversky, A. and Gilovich, T. (1989). The "hot hand": Statistical reality or cognitive illusion? Chance, New Directions for Statistics and Computing, 2 (4), 31-34.
- Hooke, R. (1989). Basketball, baseball and the null hypothesis. Chance, New Directions for Statistics and Computing, 2 (4), 35-37.

- Wardrop, R.L. (1995). Simpson's paradox and the hot hand in basketball. The American Statistician, 49 (1), 24-28.
- Dorsey-Palmateer and Smith, G. (2004). Bowlers' hot hands. The American Statistician, 58 (1), 38-45.
- Duckworth, F.C. and Lewis, A.J. (1998). A fair method for resetting the target in interrupted one-day cricket matches. Journal of the Operational Research Society, 49, 220-227.
- de Silva, B.M., Pond, G.R. and Swartz, T.B. (2001). Estimation of the magnitude of victory in one-day cricket. Australian and New Zealand Journal of Statistics, 43 (3), 259-268.
- Beaudoin, D. and Swartz, T.B. (2003). The best batsmen and bowlers in one-day cricket. South African Statistical Journal, 37, 203-222.
- Swartz, T.B., Gill, P.S., Beaudoin, D. and de Silva, B.M. (2004). Optimal batting orders in one-day cricket. Computers and Operations Research, to appear.
- Bingham, D.R. and Swartz, T.B. (2000). Equitable handicapping in golf. The American Statistician, 54 (3), 170-177.
- Stern, H.S. and Wilcox, W. (1997). Shooting darts. In the column, A Statistician Reads the Sports Pages, Chance, 10 (3), 16-19.
- Stern, H.S. (1998). Football strategy: Go for it! In the column, A Statistician Reads the Sports Pages, Chance, 11 (3), 20-24.
- Berry, S.M. (2000). My triple crown. In the column, A Statistician Reads the Sports Pages, Chance, 13 (3), 56-61.
- Berry, S.M. (2001). How Ferocious is Tiger? In the column, A Statistician Reads the Sports Pages, Chance, 14 (3), 51-56.
- Simon, G.A. and Simonoff, J.S. (2002). Were the 1996-2000 Yankees the best baseball team ever? Chance, 15 (1), 23-29.
- Berry, S.M., Reese, C.S. and Larkey, P.D. (1999). Bridging different eras in sports (with discussion). Journal of the American Statistical Association, 94, 661-686.
- Gill, P.S. (2000). Late game reversals in professional basketball, football and hockey. The American Statistician, 54, 94-99.
- Stern, H.S. (1994). A brownian motion model for the progress of sports scores. Journal of the American Statistical Association, 89, 1128-1134.

2.3 Statistical methodology

The lecture component of the course involved discussion of the papers in section 2.2 with extra attention given to unfamiliar methodologies. Below is the course outline for the first 8 weeks of the course with the technical content covered during the lectures. The students appeared to gain an appreciation for the methodologies as they saw relevance to interesting problems.

- May 03
  - introduction to the course

  - provision of reading materials
  - introduction to one-day cricket (for future lectures)

- May 10 and May 17
  - sports gambling
    * lines, odds: American and European
    * parlays, teasers, futures, etc
    * vigorish, middling, scalping (i.e. arbitrage)
    * testing for profitable systems and the issue of multiple comparisons
    * money management: fixed wagers, fixed percentage wagers, Kelly system
  - technical material in lecture: expectations, hypothesis testing, bootstrapping, Gauss-Seidel algorithm

- May 17
  - the hot hand: does it exist?
  - technical material in lecture: hypothesis testing, simulation

- May 24
  - Victoria Day holiday (class cancelled)

- May 31 and June 07
  - one day cricket
    * the Duckworth/Lewis method
    * quantifying the margin of victory in matches
    * performance measures for batting and bowling
    * optimal batting orders
  - technical material in lecture: regression, simulated annealing, goodness-of-fit, log-linear models, Markov chain Monte Carlo, delta method

- June 14
  - handicapping in golf; issues of fairness
  - modelling the outcomes of major league baseball games
  - technical material in lecture: simulation, order statistics, distribution theory, CART

- June 21
  - questions of interest from "A Statistician Reads the Sports Pages"
    * where to aim in darts?
    * under what circumstances should you kick on fourth down in American football?
    * when do you pull the goalie in hockey?
    * how many majors will Tiger Woods win in his career?
  - technical material in lecture: goodness-of-fit, dynamic programming, distribution theory

- June 28
  - comparing performances from different eras

- comebacks in team sports
- technical material in lecture: density estimation, logistic regression, hierarchical models, distribution theory, Brownian motion, probit regression

3. Further Remarks

The remaining weeks in the course involved student presentations. This was a major component of the course, and involved a research project in statistics in sport. I asked students to begin with a sports question that caught their fancy. They were asked to collect data and utilize sensible statistical analyses. Classmates were encouraged to think critically about the presentations and provide classroom discussion. Students worked in teams.

My impression was that this was also an enjoyable experience and provided some students with their first foray into problem solving using statistics. In fact, one of the problems was subsequently developed into a publishable paper (Summers, Swartz and Lockhart 2007).

REFERENCES

Albert, J. (2002). "A baseball statistics course". *Journal of Statistics Education* [Online], 10(2).

Kvam, P.H. and Sokol, J. (2004). "Teaching statistics with sports examples". *Informs Transactions on Education* [Online], 5(1).

Lock, R.H. (2006). "Teaching an introductory statistics class based on sports examples". *The International Association for Statistical Education* [Online], Session 5F at ICOTS-7.

O'Brien, C.A. Pollett, A. Gallinger, S. and Dick, J.E. (2006). "A human colon cancer cell capable of initiating tumour growth in immunodeficient mice". *Nature*, advance online publication, 19 November.

Reiter, J. (2001). "Motivating students' interest in statistics through sports". *ASA 2001 Proceedings of the Section on Statistics in Sports*, Alexandria, VA: American Statistical Association.

Summers, A.E. Swartz, T.B. and Lockhart, R.A. (2007). "Optimal drafting in hockey pools", To appear in *Statistical Thinking in Sports* (editors R.H. Koning and J. Albert), CRC Press.

ABSTRACT

This paper describes a graduate course in statistics in sport that was first offered in the summer semester of 2004 at Simon Fraser University. In this paper, we describe the topics that were covered in the course and we provide a rationale for introducing such a course. Specifically, we suggest that the course provides an immediate avenue to statistical modelling, an easy introduction to the reading of scientific papers and an introduction to an array of statistical methodology. It is further argued that the course provides an enjoyable experience in advanced statistical training.