9-2004

Teaching Statistics with Sports Examples

Paul Kvam
University of Richmond, pkvam@richmond.edu

Joel Sokol

Teaching Statistics with Sports Examples

Paul H. Kvam
Joel Sokol
School of Industrial and Systems Engineering
Georgia Institue of Technology
pkvam@isye.gatech.edu
jsokol@isye.gatech.edu

Abstract

Class material for introductory and advanced statistics can be colorfully illustrated by using appropriate data and examples from sports. Specific methods, including statistical graphics (e.g., boxplots), ball-and-urn probabilities, and statistical regression are demonstrated. Examples are drawn from popular American sports such as baseball, basketball, soccer and American football. Classroom feedback indicates that that most students enjoy sports examples as a way to learn abstract concepts using familiar, recreational settings.

Editor's note: This is a pdf copy of an html document which resides at http://archive.ite.journal.informs.org/Vol5No1/KvamSokol/

1. Introduction

Modern statistics education has emphasized the application of tangible and interesting examples to motivate students learning about statistical concepts. Introductory texts aimed at special audiences (e.g., business students, epidemiology students, or engineering students) feature problems and illustrations relevant to those audiences, complementing course material from related classes. The current textbook (Hayter, 2002) used for the Georgia Institute of Technology's introductory statistics course in the School of Industrial and Systems Engineering includes a strong emphasis on science and engineering; more than half of the exercises in the text are simple and illustrative examples that are related to topics studied by engineering undergraduates.

So why should one consider teaching statistics using sports examples? Clearly, an introductory course that is dominated by such examples is inappropriate for students who will apply statistical methods in business, science, or engineering. Most sports examples found in the statistics literature are based on sports that are mainly popular in North America or Europe, the most commonly cited topic being baseball. While American and European instructors might be familiar with such sports examples, an increasing proportion of students in western universities are not from western countries, and have less experience with these sports.

In our experience, however, when it comes to choosing projects for various data analyses (regression, contingency tables, analysis of variance), the most popular themes, year after year, are sports related. We're surprised to find students from China or India eager to analyze attendance data for Atlanta Braves home games or apply goodness-of-fit tests to National Collegiate Athletic Association (NCAA, 2003a) college basketball outcomes. While engineering examples have a clear purpose in teaching students in our College of Engineering, sports examples seem to bring an added level of excitement to the classroom experience.

Introductory statistical techniques lend themselves to endless applications in sports, especially baseball, where statistics are collected on almost all aspects of player performance. Albert (2002), a professor at Bowling Green State University, outlines a basic statistics course that can be taught entirely through baseball examples. Simonoff (1998) focused on the home run race between Sammy Sosa and Mark McGwire during the 1998 baseball season, and utilized both introductory statistics (graphs, categorical data analysis, analysis of variance) along with more advanced methods (logistic regression and smoothing methods).

The statistics literature features several more sports examples; in general, they are used to motivate or illustrate new and advanced methods of statistical inference, e.g., Cochran (2002), Samaniego and Watnik (1997), Harville and Smith (1994), Crowder, et al.

  ISSN: 1532-0545

(2002), Gill (2000). For its eight most published sports topics, the Current Index to Statistics (CIS) lists 230 articles that appeared in statistics-related journals between 1960 and 2002. Figure 1 charts the frequency of the eight sports in the database; although many of the international journals in the CIS are published outside the United States, baseball still dominates the list. This is partly due to baseball's close affinity with statistics and statistical analysis, and because so much statistical information about baseball is readily available on the Internet. Another reason U.S. sports dominate the literature is because mostly U.S. authors are submitting sports-related research papers to refereed journals (case in point: peruse the author list of this special issue!). The modest goal of this article is to show different ways sports examples can be used to illustrate simple statistical methods or to motivate project work in an introductory class. Examples are limited to the sports seen in Figure 1 as most frequently published on, notably baseball, (American) football, basketball, and soccer.

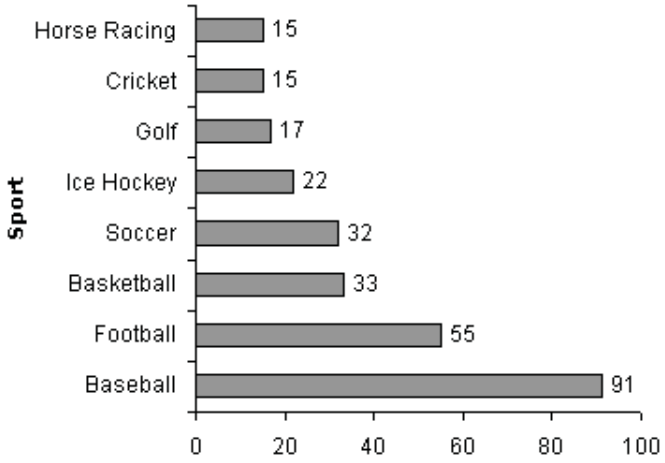

Figure 1: No. Articles in CIS.

2. NBA Draft Lottery: A Tiring Exercise in Probability

For teaching elementary probability, a colorful substitute to the standard ball-and-urn examples can be found in the National Basketball Association (NBA) draft-order determination held in spring before the summer draft. Prior to 1985, the last-place finishers in each of the two conferences would flip a coin to determine which team picked first and which picked second. A lottery system started in 1985 prevented the teams with the worst records from automatically receiving the first two picks, so that teams would not intentionally lose games to ensure that top draft pick. Each of the seven teams that failed to make the playoffs had an equal chance of drafting first. The first year proved to be memorable as the New York Knicks received the first pick (with a one in seven chance) and selected Patrick Ewing weeks later on draft day.

After a few seasons, critics pointed out that the first selection in the draft generally had not gone to the worst or even second worst team in the league. In response, the draft lottery changed in 1990 to a weighted probability system. Since then, the NBA draft lottery has provided probability and statistics instructors with non-trivial alternatives to the bland ball-in-urn homework problems seen in most introductory textbooks.

In the 1990 draft, the eleven worst teams participated in the lottery and the ith "best" team (of the 11) would receive a weight of wi = i. Although this change made the worst team eleven times more likely to receive the number one pick than the 11th worst team, luck came to the Orlando Magic in 1993 (the 11th worst team) when they received the first pick with the highly unlikely chance of 1/(1+2+...+11)= 0.0152.

Critics again demanded a change in the system, perhaps not fully understanding the rarity of occurrence for the 1993 draft outcome, and this "catastrophic error" rate changed from 0.015 to 0.005. The prerequisite for understanding the draft lottery probabilities evolved even more in subsequent years. Fourteen numbered balls were placed in a drum, and four were chosen without replacement (14-choose-4 = 1001 ways). One thousand combinations were assigned to the 11 lottery teams, with 250 of the combinations belonging to the worst team and 5 to the best (one combination was left over; drawing it would lead to a re-drawing).

In 1995, the lottery brought in two more teams and reassigned some of the 1000 combinations, keeping 250 for the worst team and reducing the chances for the 2nd to 6th worst teams. Each augmentation provides different probability distributions for the lottery teams, and each one offers interesting insights to probability students computing and comparing the probabilities associated with lottery ranking. The NBA has posted several web pages associated with the draft lottery and the history of lottery picks and probabilities (see National Basketball Association, 2003a-c).

3. Statistical Graphics

Graphs in statistics, including bar charts (e.g., Figure 1), pie charts and histograms, represent a broad interface between statistics and the general public. Statistical graphics are mandatory in the print media, and it is now commonplace to see a political candidate use statistical charts to support their point of view, especially in debates. Ross Perot used charts in his presidential bid in 1992. Dennis J. Kucinich, during his 2004 campaign for the Democratic presidential nomination, actually came to a National Public Radio debate prepared with a pie chart to argue his point about the Pentagon budget (to show the other candidates, he claimed).

Television, magazines, and newspapers all rely on charts to communicate data. The USA Today relies on charts to communicate anything from national trends to entertaining trivia. Occasionally, bar charts are used in the sports pages. While sports examples can easily be used to motivate bar charts, there are less common sports examples that show more powerfully how statistical graphics can communicate information. In fact, sports provide numerous examples for illustrating statistics with pie charts, scatter plots, Pareto charts, bubble charts, surface plots and box plots.

3.1. Uses and Misuses of Statistical Graphics

Statistical lies are most frequently committed in graphical form, where the eyes can be more easily deceived by spurious trends suggested in a picture. A common abuse is manipulating scales on charts and graphs by truncating, censoring or transforming the axis values. Figure 2 shows two different charts showing an increase in average attendance at NCAA Women's Soccer[(1)] games between 1998 and 2003. The (blue) chart on the right is the default Microsoft Excel chart; many statistical software packages, in fact, will restrict both axes to a small set of values that contains the data, which helps the reader focus on chart differences more clearly. However, it also removes the scale of difference from the picture, which has potential to mislead readers who pay little attention to the axis labels.

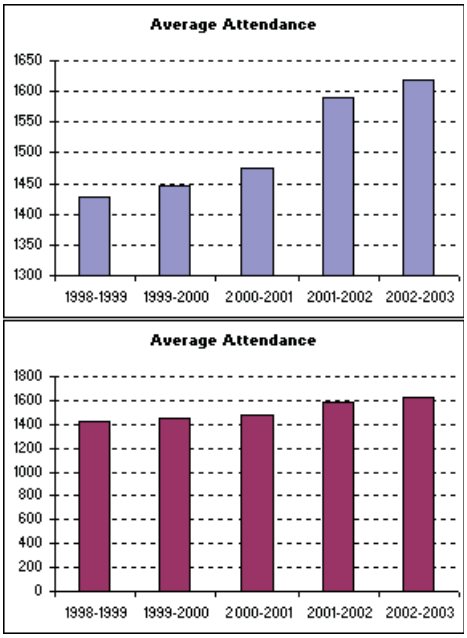

Figure 2: Two different charts showing average attendance at NCAA Women's Soccer (season) matches.

The reader's sense of proportion can be manipulated further with image-based charts, which are standard in publications such as USA Today. As an example, Figure 3 below graphs the season wins for the New England Patriots using clip-art in place of vertical bars. While the height of the football icons corresponds to the information the graph is meant to communicate, the size of the footballs does not; the Patriots improved 56% in wins between 2002 and 2003, but the increase in area of the football icons is over 140%.

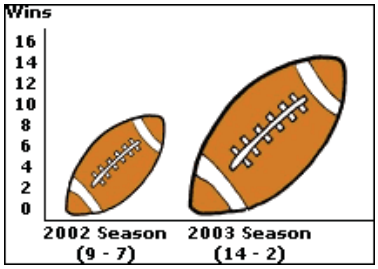

Figure 3: Regular season wins for the New England Patriots, 2002-2003.

(1) http://archive.ite.journal.informs.org/Vol5No1/KvamSokol/soccerAttendance.xls

  ISSN: 1532-0545

3.2. Boxplots

Below is an example of how a box plot[(2)] can summarize salary differences in Major League Baseball (MLB) for the 2003 season. In this case, outlying data points (Alex Rodriguez - Texas, Carlos Delgado - Toronto) draw attention away from the bars, and a plot without plotted outliers (an option in most statistical packages) can show more with respect to team salary quartiles.

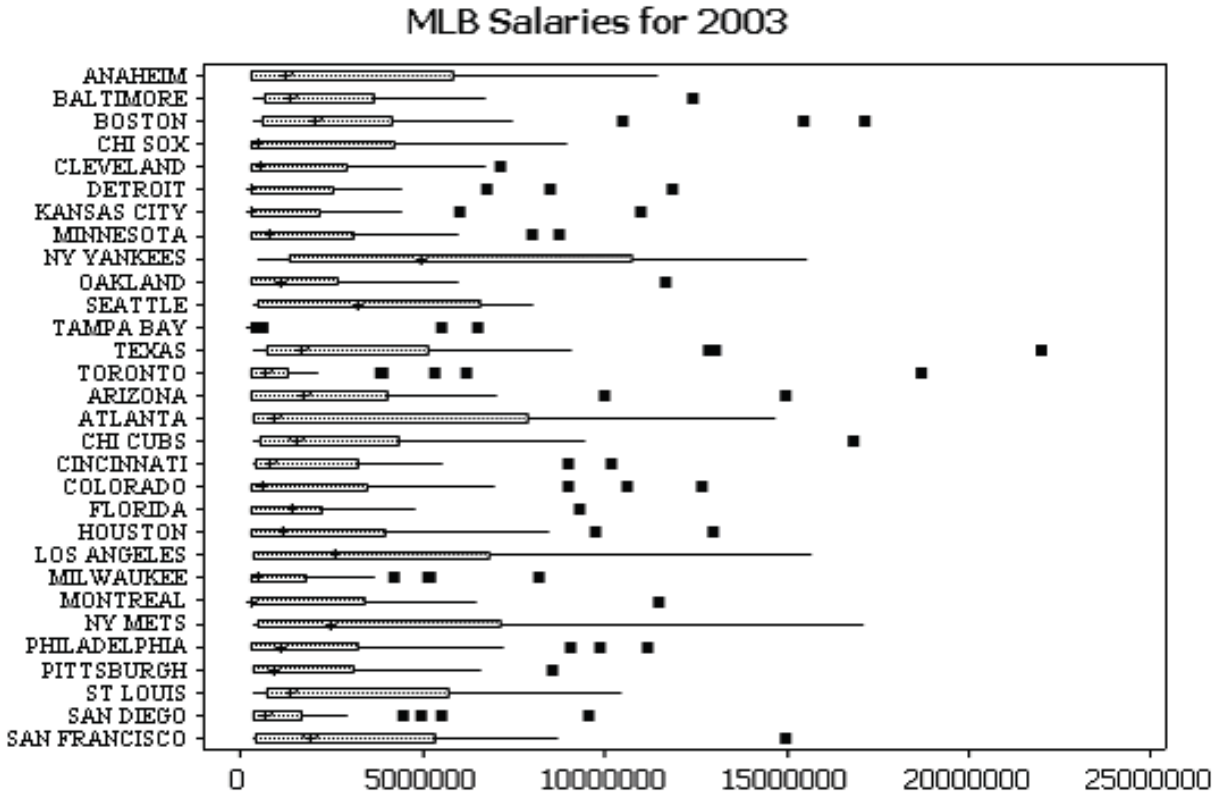

Figure 4: Box plot for player salaries of MLB teams in 2003.

3.3. Graphical Summary for Basketball Games

Innovative plots have been developed for special sets of data. Westfall (1990) presented a simple, yet revealing graphical summary of a basketball game by plotting the point difference between the two teams' scores across time. In basketball, perhaps more than any other of the mainstream American sports, the game is difficult to summarize in a simple box score. Figure 5 below shows the summary of the February 1, 2004 NBA game between the Minnesota Timberwolves and the Philadelphia 76ers. Minnesota won the game 106-101. The box score, shown in Table 1, fails to summarize what happened in the game: Minnesota overcame an 18-point deficit and pulled ahead for the first time late in the game. Students can learn about the power of statistical graphics through such novel uses of charts. We note that this type of chart can also be used in a stochastics course to illustrate the idea of one-dimensional random walks with varying step sizes.

Table 1: Box score for NBA game between Minnesota and Philadelphia, 2/1/2004

	1	2	3	4	**Total**
Philadelphia	36	27	17	21	101
Minnesota	23	29	27	27	106

(2) http://archive.ite.journal.informs.org/Vol5No1/KvamSokol/MLBSalaries.xls

  ISSN: 1532-0545

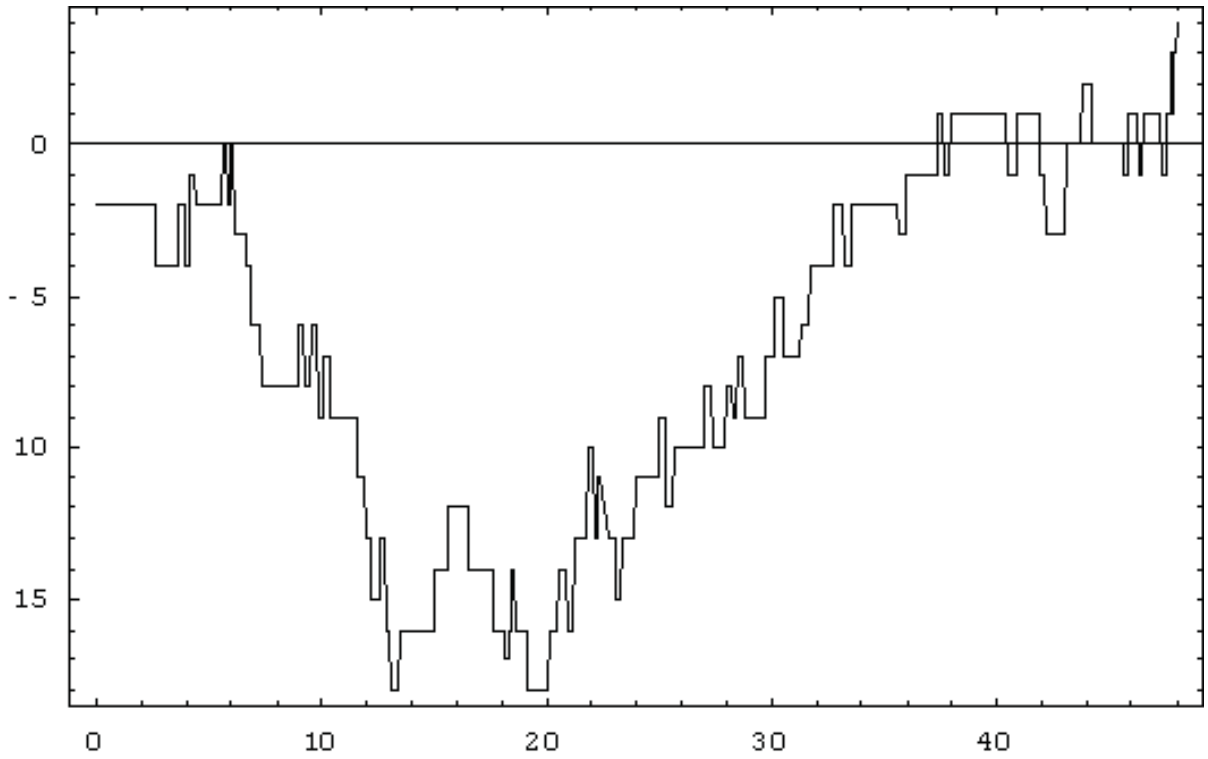

Figure 5: Point difference in NBA game between Minnesota Timberwolves and Philadelphia 76ers, 2/1/2004.

4. Teaching Simpson's Paradox with Sports Statistics

Simpson's paradox occurs with categorical data that has three variables when an association between two of the variables is consistent across all of the levels of the third variable, but is completely different if one aggregates over the third.

The paradox is best described using a pair of two-by-two contingency tables, and baseball presents many examples of Simpson's paradox. The three variables, each at two levels, are player (two batters), batting outcome (hit or out), and batting situation (runners in scoring position or not). Table 2 below shows one of 56 pairings in which this paradox took place in the 2003 MLB season. It shows how Dustan Mohr (Minnesota Twins and San Francisco Giants) had a higher batting average (hits per at-bat) than Darin Erstad (Anaheim Angels) in both batting situations when examined separately, but overall Erstad had a higher batting average than Mohr. The key to the paradox, of course, is that the proportions being compared are based on different sample sizes. In this case, Erstad appeared with runners in scoring position a smaller proportion of the time (20%) than did Mohr (28%). (The reason for the disparity in at-bats with runners in scoring position is that Mohr generally batted after more players who were likely to get on base; see Sokol (2003) for more discussion of the effect of batting order placement.) Other pairings that illustrate Simpson's paradox include Carl Everett vs. Hideki Matsui, Jose Reyes vs. Carlos Beltran, and Frank Thomas vs. Josh Phelps.

Table 2: Simpson's Paradox in MLB batting averages

	Runners in Scoring Position		**No Runners in Scoring Pos.**		**Overall**	
Mohr	Erstad	Mohr	Erstad	Mohr	Erstad	
Hits	19	9	68	56	87	65
At Bats	97	50	251	208	348	258
PCT	0.196	0.180	0.271	0.269	0.250	0.252

5. Regression Analysis

Student projects involving large sets of real data are a vital part of effective statistics classes. Projects are ideal for teaching linear regression because students have a high degree of freedom to select their own models to characterize the relationship between the response and the regressors.

One of the richest examples we have found for use in a statistics class is the problem of modeling a baseball player's value based on their individual statistics. For each player, batter or pitcher, there are dozens of potential regressor variables to consider in the model. The Microsoft Excel file MLB.xls[3] contains the 2003 MLB batting statistics for 336 major league batters and lists 23 basic statistics (more refined databases have many more statistics to consider). We used a "fantasy league value" as the response of interest. This fantasy league value, from The Sporting News 2003 Fantasy Players Guide, is related to player performance via statistics such as hits, RBI, runs, home runs, stolen bases, but the functional link cannot easily be characterized in a linear or nonlinear regression because many other variables influence the response. Other variables that influence fantasy value are age, team, position, injury history, and consensus findings from scouting reports. Up-to-date data sets can be obtained from many on-line sources such as ESPN.com[4]. Historical data (every player, every season) can be downloaded from the The Baseball Archive[5].

Students usually work in pairs, and with so many possible regression models, it is possible that no two groups arrive at the same model. As instructors, we could not help but notice that students who knew the most about baseball did not derive the best fitting model. Often, a pair of students knowing little about the nuances of the game would garner the best model (with a small number of regressors) relying entirely on empirical results of the data to guide their model selection. Some baseball fans, on the other hand, tended to interject regressors they subjectively preferred but were not optimal variables to add into the regression model. More advanced students can consider categorical (or nominal) inputs (e.g., player's team) to form general linear models, regression diagnostics, and variable transformations to improve model fit.

6. Logistic Regression Analysis

Examples from sports can also be used to teach more advanced regression techniques such as logistic regression. Examples of logistic regressions are usually limited to biostatistics and other life sciences, but the following example, which examines the effects of home court advantage in college basketball, shows how statistics can be used to provide students with new insights into a familiar problem.

Many NCAA basketball conferences play full or partial home-and-home round-robin schedules, so that the conference teams play each other twice during the season, once at each school. Using data collected from the 1999-2000 season through the 2002-2003 season, we seek to answer the question "Given that team A beat team B at home (or on the road) by X points, how likely are they to win the return match on the road (or at home)?"

College students, especially those at a school like Georgia Tech with a major basketball program, often give a question like this much more passionate thought than it might deserve (especially when asked close to NCAA tournament selection time), so it might make capturing their attention an easier task. However, answering the question might not be as easy as they would expect, because the model is more complex than they first imagine - in addition to modeling binomial data by linking the success probability to the observed point difference, students observe grossly unequal sample sizes; that is, there are very few observations of extreme cases because few teams ever win or lose a game by more than 40 points. Figure 6 shows the observed probability of winning a road game given the previously observed point spread in the home game (blue bars) along with the estimated probability based on the logistic regression model (white bars) with

$$P(\text{Win} \mid \text{point spread} = x) = \frac{e^{-(ax+b)}}{1+e^{-(ax+b)}}$$

(3) http://archive.ite.journal.informs.org/Vol5No1/KvamSokol/MLB_Regressiondata2003.xls
(4) http://sports.espn.go.com/mlb/stats/batting?league=mlb
(5) http://www.baseball1.com/statistics/

  ISSN: 1532-0545

where (a,b) are estimated as (-0.6228, 0.0292) with standard errors (0.0231, 0.0017). Figure 7 charts the number of observations collected at the respective point spreads. This data was collected from the daily college basketball scoreboard pages at Yahoo.com. Unfortunately, there is no easy way to download and parse all of the scores; we wrote our own C and Unix C-shell code, specialized for our system, to compile the data.

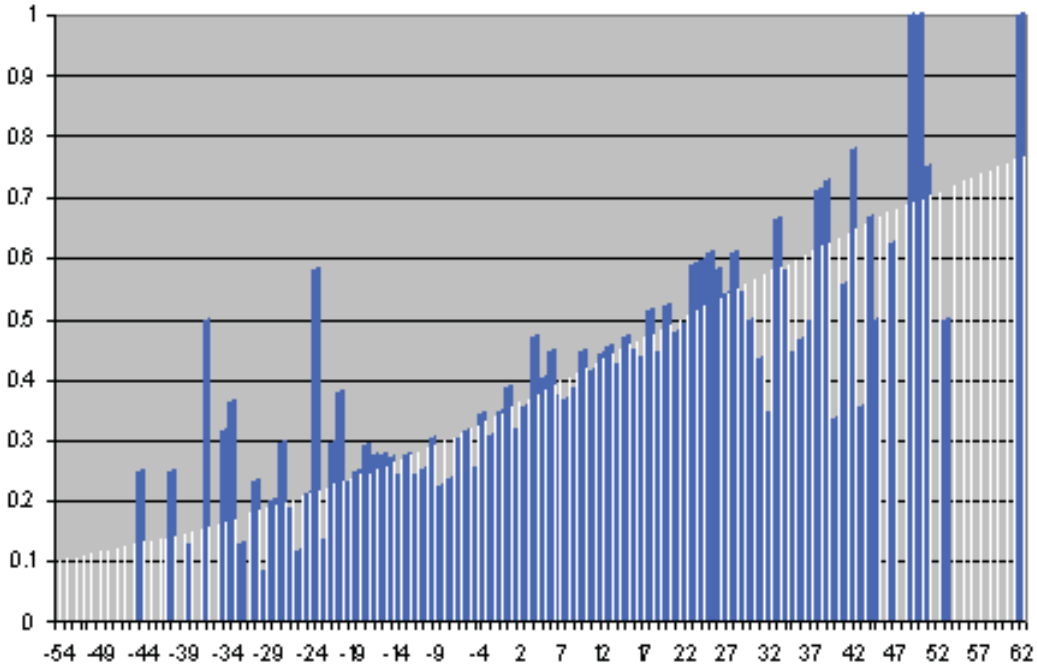

Figure 6: Observed win probabilities (blue) and logistic regression estimates (white) for home games at a given point spread.

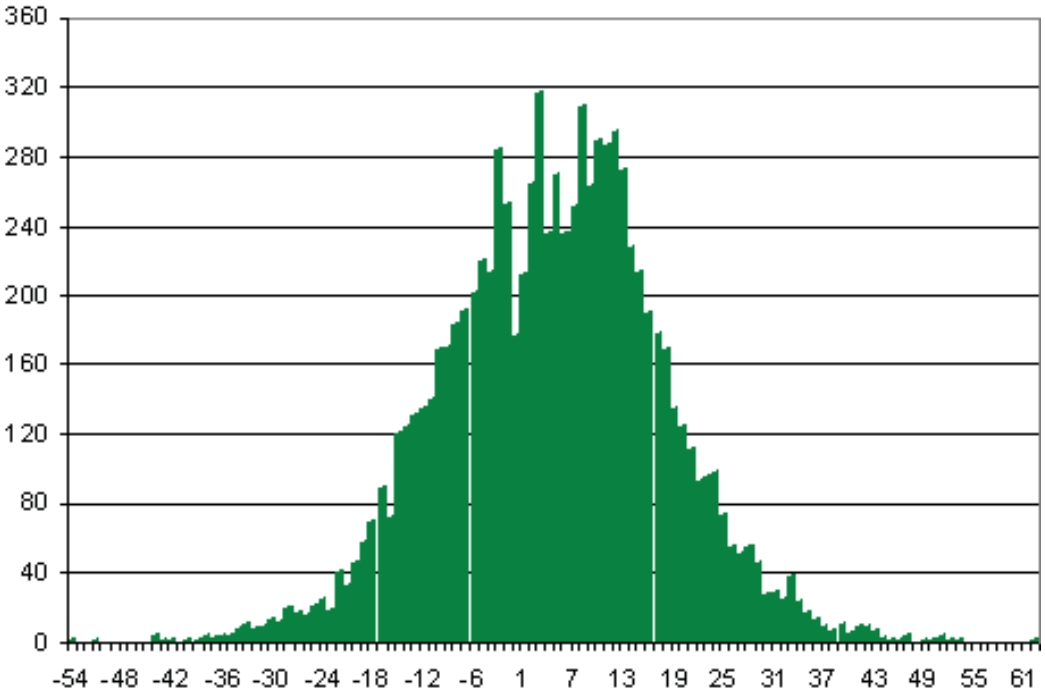

Figure 7: Number of games at various point spreads.

  ISSN: 1532-0545

A benefit of using this example to teach statistics is that, in addition to learning more about statistics, students also see how properly applied statistical methods can give sports fans a new understanding of an old problem. In this case, they can see for themselves that home-court advantage, usually valued at 3-5 points (see, for example, Sagarin (2004)), is probably really worth about 10-12 points. The model indicates that a team needs to win a home game by 20-24 points, or twice the home-court advantage, in order to have an approximately 50% chance of beating that same opponent in a road game. (Mathematically, suppose h is the value of home-court advantage. If Team A is p points more skillful than Team B, then we would expect Team A to win at home by p+h points and have a p-h point-differential on the road. Therefore, when we observe (see Figure 6) that p-h is approximately zero (a 50% chance of winning the road game) when p+h is in the 20-24 range, it is easy to deduce that h must be between 10 and 12.)

The moral of the story? In addition to having learned (and gained an appreciation for) statistical methods, students now know it's worthwhile walking across campus to the basketball arena when that "unbeatable" opponent comes to play; with a 12-point advantage, who knows what might happen!

7. Regression Vs. Linear Programming

Statistical regression methods are also often used to obtain relative ratings of sports teams. In statistics classes (and in optimization classes), power-rating (a widely-used measure for predicting a game's point differential; see, for example, Sagarin (2003)) examples from sports can help teach students this use of regression as well. For example, in college football many conferences are too large for full round-robin play. The conference winner is still determined by won-lost record within the conference, but some teams play more difficult schedules than others. In the 1999 Big Ten example below, for example, students might wonder whether Wisconsin's easier schedule led to their finishing with a better record than Michigan and/or Michigan State.

Table 3: Results of play in the Big Ten Conference, 1999 (winner's score is listed first).

Loser → / Winner ↓	Wisconsin	Michigan	Michigan State	Penn State	Minnesota	Illinois	Purdue	Ohio State	Indiana	Northwestern	Iowa	**Overall Won-Lost Record**
Wisconsin	---	---	40-10	---	20-17	---	28-21	42-17	59-0	35-19	41-3	**7-1**
Michigan	21-16	---	---	31-27	---	---	38-12	24-17	34-31	37-3	---	**6-2**
Michigan State	---	34-31	---	35-28	---	27-10	---	23-7	---	34-0	49-3	**6-2**
Penn State	---	---	---	---	---	27-7	31-25	23-10	45-24	---	31-7	**5-3**
Minnesota	---	---	---	24-23	---	37-7	---	---	44-20	33-14	25-21	**5-3**
Illinois	---	35-29	---	---	---	---	---	46-20	---	29-7	40-24	**4-4**
Purdue	---	---	52-28	---	33-28	---	---	---	30-24	31-23	---	**4-4**
Ohio State	---	---	---	---	20-17	---	25-22	---	---	---	41-11	**3-5**
Indiana	---	---	---	---	---	34-31	---	---	---	34-17	38-31	**3-5**
Northwestern	---	---	---	---	---	---	---	---	---	---	23-21	**1-7**
Iowa	---	---	---	---	---	---	---	---	---	---	---	**0-8**

This is an interesting example that makes students think about the relative benefits of different statistical models. If the power ratings are defined using a linear programming approach (where the error in a prediction is defined as its absolute difference from the observation), then Michigan and Michigan State are much closer to Wisconsin. On the other hand, if the power ratings are defined using a linear regression with the error defined as the squared difference, then Wisconsin has a much larger advantage.

  ISSN: 1532-0545

Table 4: Power ratings calculated using two simple regression models.

Team (Record)	Linear Programming Power Rating	Team (Record)	Regression Power Rating
Wisconsin (7-1)	17.2	Wisconsin (7-1)	19.6
Michigan State (6-2)	14.2	Michigan (6-2)	9.4
Michigan (6-2)	11.2	Michigan State (6-2)	8.2
Minnesota (5-3)	8.2	Penn State (5-3)	7.8
Penn State (5-3)	7.2	Minnesota (5-3)	5.5
Purdue (4-4)	1.2	Purdue (4-4)	2.6
Ohio State (3-3)	-1.8	Illinois (4-4)	-2.4
Illinois (4-4)	-2.8	Ohio State (3-5)	-2.6
Indiana (3-5)	-13.8	Indiana (3-5)	-10.2
Northwestern (1-7)	-19.8	Northwestern (1-7)	-17.0
Iowa (0-8)	-20.8	Iowa (0-8)	-19.9

We describe both of these models in more detail in the appendix.

8. Classroom Experience

In this section, we describe our experiences with using these examples in the classroom. Because our experience covers multiple courses, we first describe the courses in which we have used this material, and where those courses fit into the curriculum.

All of the courses in which we have used this material are in the School of Industrial and Systems Engineering (ISyE) at Georgia Tech. The undergraduate courses are both required for and restricted to ISyE majors, so we have a relatively homogeneous set of students. The graduate-level course is required for ISyE students who are pursuing a Master's degree in Operations Research (MSOR), and is taken by most ISyE students who are pursuing a Master's degree in Industrial Engineering (MSIE). The course also attracts first-year ISyE PhD students who may not have seen mathematical programming in their undergraduate curricula, as well as Master's and PhD students from other disciplines whose research relates to optimization.

We have used this material in the following set of courses:

ISyE 2027, Probability with Applications: A sophomore-level course covering conditional probability, probability distributions and Poisson processes. Basic calculus is required.

ISyE 2028, Basic Statistical Methods: A sophomore-level course covering parameter estimation, statistical decision-making, and analysis and modeling of relationships between variables. Students taking this course have already seen basic calculus and probability, but may not have taken any other ISyE courses.

ISyE 3039, Methods of Quality Improvement: A junior-level course covering design of experiments, measurement, statistical process analysis and control, and acceptance sampling. Students in this course must have already taken statistics (see ISyE 2028 above) and stochastics.

ISyE 4231, Engineering Optimization: A senior-level course covering optimization modeling and solution techniques, mathematical programming, and network and graph models. Students in this course are usually near the end of their ISyE curriculum, all have taken ISyE 2028, and most have taken ISyE 3039.

ISyE 6669, Deterministic Optimization: A Master's-level course covering linear, discrete, and nonlinear optimization models, algorithms, and computations. The students in this course have a nominal requirement of ISyE 4231 (or an equivalent course from their undergraduate institution), but many or most actually take the course without having taken the prerequisite. This course also attracts Master's and PhD students from other disciplines, giving us a very diverse set of student backgrounds.

ISyE 6739, Basic Statistical Methods: A Master's-level (service) course intended for graduate students who want an overview of basic tools for probability and statistics, and covers most of the material in courses ISyE 2027-2028.

In all of the undergraduate courses, most of our students are American and have at least a basic understanding of the sports involved. Even so, there are always some who are unfamiliar with even the basic

  ISSN: 1532-0545

rules of the games, either from lack of exposure (in the case of many foreign exchange students) or lack of interest. Interestingly, as we noted in Section 5, those students who are most familiar with the sports in question are not necessarily the ones who do the best analysis. Often, they bring their own learned biases to the analysis, whereas students who are unfamiliar with the application can approach the problem with a fresh perspective. (We note that we have observed the same phenomenon with other, non-sports applications as well; for example, students who have co-op or internship experience in logistics sometimes get bogged down in minor details, e.g., where the truck driver will stop for lunch, and miss the overall analytical benefit of using a mathematical model.) Moreover, even in the graduate course where the majority of students may not be from the US, students have no difficulty understanding the concept of the underlying model of sports or games, just as they can quickly pick up models of factories despite generally having little to no experience inside of one.

Before enrolling in the probability and statistics courses, students have taken a year of basic calculus. Fundamental probability is covered in ISyE 2027 (and ISyE 6739), where the lottery example is used to illustrate basic counting problems. Statistical graphics are a core subject for ISyE 2028, and also introduced in more abbreviated form in ISyE 6739. Both of these courses finish the term with a regression project, and every usage of the MLB data set has proved successful. More than textbook data sets, the MLB example introduces students to the gray issues of over-fitting versus parsimony. In final course evaluations, the baseball project receives more praise than any other specific project or homework assignments.

In the optimization courses, we teach students who are further along in their curriculum; almost all of the students in ISyE 4231 are seniors, and students in ISyE 6669 are all Master's or PhD-level. Therefore, the more advanced examples described in this paper are useful for two reasons. First, the students are more advanced and can understand more complex models. Second, students in both courses will have previously seen statistical concepts such as regression: undergraduates will have already taken ISyE 2028, and graduate students should have seen regression as undergraduates. Therefore, the football power-rating example of Section 7 has the benefit of showing linkage between optimization and statistics. From this exercise, students learn that standard least-squares regression can be formulated as a convex quadratic program or a linear program (see Appendix), and also that statistical parameter estimation in general is a type of optimization problem. We find that the students enjoy this type of example quite a bit, because it makes the curriculum seem more unified rather than a set of unrelated methodologies.

9. Conclusion

In this paper, we have described several ways in which introductory and advanced statistical concepts can be illustrated using examples from sports. Based on student feedback, we find that most students enjoy sports examples. The fact that the abstract concepts they learn can be applied in recreational ways often gets them thinking about other real-life situations, not just traditional industrial engineering applications, where statistics can be useful. In fact, when local television and radio stations reported on the success of a predictive model (Kvam and Sokol, 2004) we created based partially on the logistic regression example of Section 6, we even had several students approach us asking if we would supervise them in independent research on these topics.

Overall, we have had a lot of success using these and other sports examples in the classroom. We find that students are very receptive to the application of statistics to sports, even if they are not sports fans themselves, and that they enjoy seeing how the material they learn can be applied in settings other than those of traditional industrial engineering. Educationally, we have observed that the students' enjoyment leads to increased interest in the material and therefore, we hope, increased learning.

  ISSN: 1532-0545

References

Albert, J. (2002), "A Baseball Statistics Course," Journal of Statistics Education, Vol. 10, No. 2.

Cochran, J. (2002), "Data Management, Exploratory Data Analysis, and Regression Analysis with 1969-2000 Major League Baseball Attendance," Journal of Statistics Education, Vol. 10, No. 2.

Crowder, M., Dixon, M., Ledford, A. and Robinson, M. (2002), "Dynamic Modelling and Prediction of English League Football Matches for Betting," The Statistician, Vol. 51, No. 2, pp. 157-168.

Gill, P.S. (2000), "Late-game Reversals in Professional Basketball, Football, and Hockey," The American Statistician, Vol. 54, No. 2, pp. 94-99.

Harville, D. A. and Smith, M. H. (1994), "The Home Court Advantage: How large is it, and does it vary from team to team?," The American Statistician, Vol. 48, No. 1, pp. 22-28.

Hayter, Anthony J. (2002), Probability and Statistics for Engineers and Scientists, 2nd edition. Duxbury Press.

Kvam, P. and Sokol, J.S. (2004), "A Successful Logistic Regression/Markov Chain Model for NCAA Basketball," Working paper, School of Industrial and Systems Engineering, Georgia Institute of Technology.

National Basketball Association (2003a), Evolution of the Draft Lottery, http://www.nba.com/history/draft_evolution.html

National Basketball Association (2003b), Year by Year Lottery Picks, http://www.nba.com/history/lottery_picks.html

National Basketball Association (2003c), Year by Year Lottery Probabilities, http://www.nba.com/history/lottery_probabilities.html

Sagarin, J. (2003), Jeff Sagarin NCAA Football Ratings, http://www.usatoday.com/sports/sagarin/fbt03.htm

Sagarin, J. (2004), Jeff Sagarin NCAA Basketball Ratings, http://www.usatoday.com/sports/sagarin/bkt0304.htm

Samaniego, F.J. and Watnik, M.R. (1997), "The Separation Principle in Linear Regression," Journal of Statistics Education, Vol. 5, No. 3.

Simonoff, J.S. (1998), "Move Over, Roger Maris: Breaking Baseball's Most Famous Record," Journal of Statistics Education, Vol. 6, No. 3.

Sokol, J.S. (2003), "A Robust Heuristic for Batting Order Optimization Under Uncertainty," Journal of Heuristics, Vol. 9, pp. 353-370.

Westfall, P. H. (1990) "Graphical Presentation of a Basketball Game," The American Statistician, Vol. 44, No. 1, pp. 35-38.

  ISSN: 1532-0545

Appendix

1. Power Rating Models

In Section 7, we refer to two different statistical models that can be used to determine power ratings. In this appendix, we describe each model mathematically. For both models, we define the following notation:

G set of games played, where each game $g \in G$ is an unordered set of teams $\{i,j\}$.

p_{ig} points scored by team i in game g

r_i power rating assigned to team i

Both models use the standard technique of minimizing a function of the total error in their predictions. We define the error e_g for a single game $g = \{ i, j \}$ to be the difference in predicted point spread and actual point spread, or $e_g = (r_i - r_j) - (p_{ig} - p_{jg})$.

The only difference between the two models is that one finds ratings that minimize the total absolute error $\sum_{g \in G} |e_g|$ while the other finds ratings that minimize the total squared error $\sum_{g \in G} e^2_g$.

1.1. Linear Programming Using Absolute Error

We can formulate the problem of finding ratings to minimize the total absolute error as a linear program:

$$\text{Minimize} \quad \sum_{g \in G} a_g \tag{1}$$

$$\text{Subject to} \quad e_g = (r_i - r_j) - (p_{ig} - p_{jg}) \qquad \forall\, g = \{i,j\} \in G, \tag{2}$$

$$a_g \geq e_g \qquad \forall\, g = \{i,j\} \in G, \tag{3}$$

$$a_g \geq -e_g \qquad \forall\, g = \{i,j\} \in G. \tag{4}$$

In this model, the variables a_g denote the absolute error for game g. Constraints (3) and (4) ensure that $a_g \geq |e_g|$ for each game g, while the minimization objective ensures that one constraints (3) and (4) will be binding at optimality for each game g, and thus $a_g = |e_g|$.

1.2. Mathematical Programming Using Squared Error

We can formulate the standard regression problem of minimizing the squared error as a the following mathematical program:

$$\text{Minimize} \quad \sum_{g \in G} e^2_g \tag{5}$$

  ISSN: 1532-0545

Subject to $e_g = (r_i - r_j) - (p_{ig} - p_{jg}) \quad \forall\, g = \{i,j\} \in G,$ (6)

This mathematical program is similar to the linear program in Section 9.1. In fact, it is well-known that we can solve this problem by substituting (6) into the objective for e_g, setting partial derivatives taken with respect to each ri equal to zero, and solving the resulting system of linear equations; therefore, this model too can be optimized using linear programming software.

  ISSN: 1532-0545