USING SPORTS DATA TO MOTIVATE STATISTICAL CONCEPTS: EXPERIENCES FROM A FRESHMAN COURSE

Vittorio Addona
Mathematics, Statistics, and Computer Science, Macalester College, United States of America
addona@macalester.edu

We discuss observations from teaching a freshman course: Statistical Analysis of Sports and Games. In many respects, this is a standard first college statistics course. Analyzing sports data, however, generates student interest in statistical ideas. Moreover, quantitative analysis in sports has become a serious research field, and many professional teams now employ statisticians. It is important for students to realize that academic and non-academic opportunities exist beyond the course. There also remain issues that need to be addressed before sports statistics courses become commonplace. They should: (1) appeal to both male and female students, (2) have a broad focus and not be too baseball-centric, (3) be primarily about statistics, not sports, and (4) have access to more appropriate textbooks. At the core of any statistics course is a desire to answer questions in meaningful ways. We offer ideas on how this can best be accomplished in this context.

INTRODUCTION

Successful statistics courses are the result of an appropriate mixing of many ingredients, perhaps the most important of which is student interest. When students are engaged in class material, they are motivated, and motivated students ultimately take more away from courses, and tend to perform better than they otherwise would. There are many ways to generate student interest. One of which is by providing an environment to illustrate ideas that is of relevance to the students outside of their academic life. Even when attempting to avoid the context-less examples of the past, introductory statistics courses have too often resorted to contrived examples of little bearing to the students. Using a sports theme to introduce statistical concepts provides a setting which is of interest to a wide array of students, while maintaining pertinent substance.

Every incoming student at Macalester College is enrolled in a "first-year seminar" (FYS) during the fall semester of their freshman year. These are regular, but often thematic, courses that are typically limited to 16 students (limits were raised to 17 students for 2009 only). This paper presents examples from the FYS entitled *Statistical Analysis of Sports and Games* (SASG), which was offered for the first time in Fall 2009. In many respects, SASG was taught as a standard first college course in statistics (i.e. content focused on descriptive and inferential methods, and extended through multiple regression). At the termination of the course, students were ready to proceed with any flavor of "second course" in statistics. It is crucial that the purpose of the course is made clear to the students: this is *not* a sports discussion class, nor a class exclusively on sabermetrics, but a statistics class with examples drawn from the sporting world to include as wide a variety of sports as possible, and have broad appeal. A necessary condition for a successful sports statistics course is that students need not be "super fans" to enjoy it.

SASG had 17 students, including 6 females and 3 international students. This is slightly higher than the college wide proportion of international students, but represents a substantially lower percentage of female students. We feel that the percentage of females will increase once the course is better understood on campus as gender neutral. There was no calculus prerequisite. The textbook for SASG was *Statistics, 11th Edition,* by McClave and Sincich (2009). The course included a data analysis project ideal for exploring individual sporting interests. Biweekly in-class labs allowed the students to reflect on their understanding of recent material. The course software was the R language (R Development Core Team, 2009).

The remainder of this paper is organized as follows: In the next section, we provide four examples used in SASG. For each example, we indicate what statistical topic(s) it addresses and, whenever appropriate, we mention issues that arose in the class discussion. We then conclude with some limitations to the course as it was taught in Fall 2009, and offer some final remarks.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

SOME EFFECTIVE EXAMPLES

Example 1: Birthdays of National Hockey League (NHL) Players

Data was gathered on every player who played in the National Hockey League's (NHL's) regular season through the 2008-09 season. Of primary interest was the birthday of each player, and this was obtained for 6,391 of the 6,407 eligible players. This in-class example was used after introducing discrete distributions, including the Binomial distribution. The students are shown the graph presented in Figure 1, a plot of the birth month frequencies, and asked to comment.

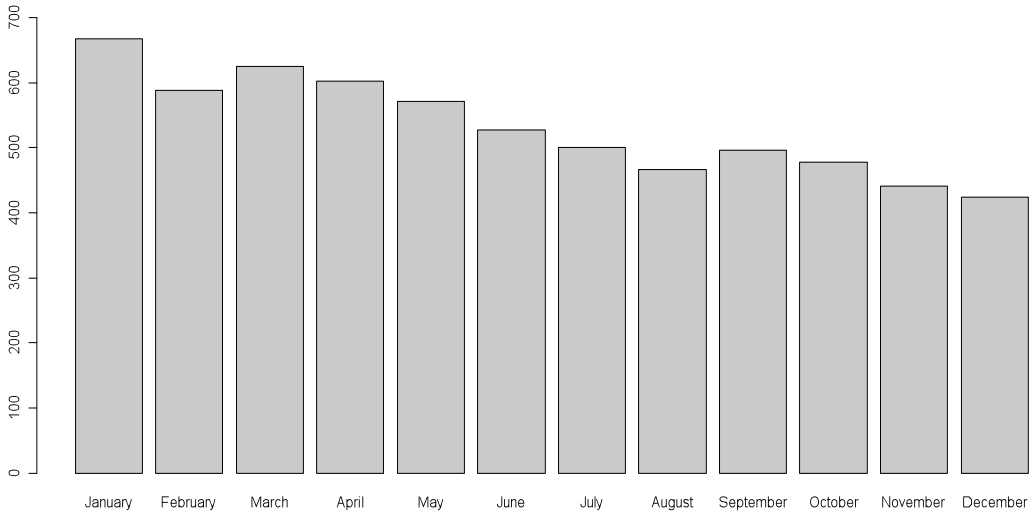

Figure 1. Absolute frequency of the birth month of NHL players through the 2008-09 season

Someone inevitably observes that there is a downward trend, or that January has many more observations than December. The students are asked to explain why this graph is surprising. In their own words, they explain that they had expected to see a uniform distribution. The class can discuss whether the graph *needs* to look *perfectly* uniform, or how "non-uniform" would it have to be before they began to suspect that this represented a real phenomenon. The professor can then introduce the goodness-of-fit test, although we allude to inference in a slightly different way, by performing a small simulation. At this point, the students have seen the *rbinom* function in R, which simulates the flip of a coin. Here, we simply have 6,391 coin flips. How do we know the chance of "heads"? For us, "heads" is "born in January", and we do not, a priori, know the probability of this event. We *believe*, however, that it should be, approximately, 31/365. This is purely an assumption (our H_0). Now, under this H_0, we can replay history 1,000,000 times, say, by typing: JanDist = rbinom(1000000, size=6391, prob=31/365). This will produce a graph similar to the one in Figure 2, the sampling distribution for the number of January births.

How compatible is the *observation* of 668 births in January with Figure 2? Or, how likely is it that we observe (at least) 668 births in January, *if* H_0 is true? The answer is "very unlikely", as seen from Figure 2, or by determining how many of the 1,000,000 replications are at least 668 (easily done in R with one command line). This corresponds precisely to the concept of a p-value.

Some students may question the assumption of uniform births and this can lead to an excellent discussion (e.g., what data could we obtain to form a more reasonable H_0?). Having rejected H_0, however, leads to a more intriguing topic for debate: *why* has this happened? The phenomenon is known as the *relative age effect* (RAE) (Barnsley, Thompson & Barnsley, 1985). Page constraints prevent a deeper study of the RAE here, but it states that the oldest children in any age category enjoy more success. The RAE has been observed in other sports (e.g. Thompson, Barnsley & Stebelsky, 1991; Barnsley, Thompson & Legault, 1992), with respect to academic performance (Barnsley, 1988), and in the business world (Du, Gao & Levi, 2009). A popularized version of some RAE research is given by Gladwell (2008).

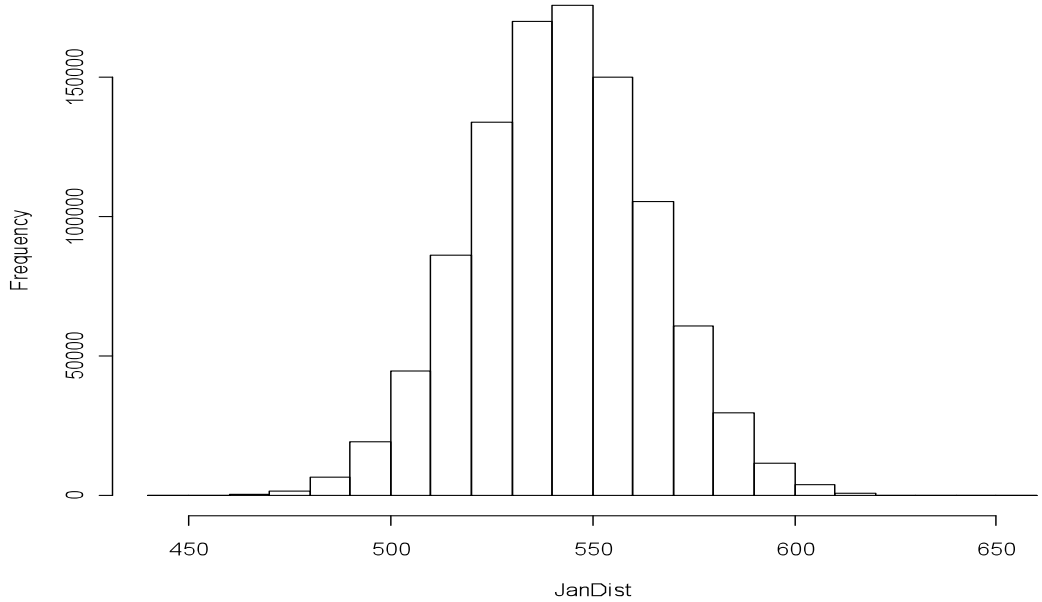

Figure 2. Sampling distribution, under H_0, for the number of January births

Example 2: Graduation Rates of Student Athletes

This example was used on an assignment which covered Simpson's Paradox. Students read the paper *Research Note: Athletic Graduation Rates and Simpson's Paradox* (Matheson, 2007). This paper addresses the commonly held belief that student athletes underperform academically, particularly in "big money" sports (i.e., men's basketball and football) at Division 1 schools. Upon further reflection, an in-class discussion of the paper would have been more suitable, as there are several excellent, but subtle, points that the students might miss.

As a metric of academic success, Matheson chooses 6-year graduation rate. An overall comparison shows underperformance by athletes. So, is conventional wisdom correct? What are potential confounding factors? The author points out that the racial makeup of the athlete population is quite different from that of the general student population (e.g., 26.8% of the athlete population is African-American, compared to 9.3% of the general population). Once we account for 'race', nearly the entire gap in graduation rate disappears. It is thus not that *athletes* are graduating at lower rates, but that African-American students have lower graduation rates, and represent a larger than proportional share of the athlete population. The graduation rate issue amongst African-Americans is well known, and it deserves undivided attention. We should not confound one matter with another.

Matheson presents an important example on a topic of societal concern, and helps refocus our attention by using the crucial method of covariate adjustment. It illustrates how easily anyone can be fooled by facts and figures presented in a certain manner. It makes the students realize that they need to question results, and think about other possible explanations. For example, the discussion might continue by considering factors which could bring the results of this paper into doubt. Do athletes merely graduate at similar levels as non-athletes (after controlling for race) because they tend to major in disciplines which have higher graduation rates? Do athletes get special treatment, or private tutoring, which allows them to succeed more frequently? Both are excellent questions which should be pondered, even if they may not have definitive answers.

Other sentences in the paper warrant mention, but due to space constraints, we only consider one. Matheson illustrates a typical piece of evidence provided in favor of the notion that athletes underperform: "Only 4 of the 64 teams [in the 2005 NCAA men's basketball tournament] had graduated all of their players over the past year". The students are asked to criticize this sentence. It seems to indicate a low graduation rate because 4/64 = 6.25% seems low. But, this says that 6.25% of teams had a 100% graduation rate! Stated in this manner, it becomes unclear whether

6.25% is low or not. Some ambiguity must also be cleared up: we assume that "all of their players" means the 5 or so *seniors* on the roster. If so, how does the 6.25% compare to the proportion of "100% graduation rates" we would observe in randomly selected sets of 5 non-athlete students? This is an easy question to answer using the Binomial distribution, or through a simple simulation. It turns out that 6.25% is almost identical to the corresponding value in the general population. Example 1 and 2 draw on sport contexts, but are, at their core, about something of broader appeal.

Example 3: Why We Don't Go For It

This example stems from the article *Why We Don't Go For It*, by Shankar Vedantam, published in the Washington Post on June 18, 2007. The article details two typical decisions made in the National Basketball Association (NBA) which, Vedantam argues, are indefensible under the scrutiny of objective reasoning. One scenario relies on a logical argument to determine whether coaches should remove a player from the game because they have a few early fouls. We will not discuss this piece further in this paper, but it is a useful discussion to have with students, since it makes them think about a problem differently from the way they have been trained to think by sports commentators, coaches, and athletes.

The other example can be summarized as follows: *Suppose that in the final few seconds of a basketball game, Player A can take a 2-point shot to tie the game. He is relatively open, but he decides instead to pass the ball to an open teammate for a 3-point shot, to win the game. The teammate misses the shot, and the team loses the game. Player A is criticized for not taking the safer 2-point shot. Did Player A make the wrong decision?*

The solution involves a simple application of conditional probability. If Player A takes the shot then *two* events must occur for his team to win: he must make the shot *and* his team must outscore their opponents in overtime (OT). Thus, we might write P(win) = P(Player A makes shot)*P(team outscores opponents in OT) = 0.5*0.5 = 0.25. Here, of course, we are assuming what we feel are reasonable values. The students were asked to fill in these two probabilities. They suggested that P(Player A makes shot) be between 0.4 and 0.5. This depends on how open, and where on the court, he is. For P(team outscores opponents in OT), there will inevitably be talk of momentum gathered by Player A's team if they tie the game. Who the home team is might matter, along with a host of other factors (e.g. who has fouled out), but if the two teams are tied, it seems reasonable to select 0.5. In class, a student commented "So, are we just making these numbers up?" This is interesting since, if the book had declared "the chance that Player A makes the 2-point shot is 0.46", no student would ever question it. There are at least two points to be made in response to this remark: (1) we are not "making up" the numbers, so much as we are trying to use our knowledge to obtain reasonable estimates of these probabilities, and (2) anyone can choose values that they feel are appropriate and potentially arrive at a different conclusion.

We complete the solution by considering the chance that the team wins if player A passes the ball. In that case, the chance of winning is simply the chance that his teammate makes the shot: P(win) = P(teammate makes shot) = 0.30, say (again, this depends on the player). For the values we have chosen, Player A did not make the wrong decision. Only at this point should the students receive Vedantam's article, which was about Lebron James (Player A) who passed to his teammate Donyell Marshall for the 3-point shot to give the Cleveland Cavaliers the win against the Detroit Pistons in the 2007 Eastern Conference Finals. Marshall missed the shot, and James was unfairly criticized. The class discussion can move to the psychology of why James was criticized even though the evidence suggests that he made the right decision. The most important reason for the criticism seems to have been the outcome. Had Marshall made his shot, James would have been praised for a "gutsy" decision. Clearly, we should not judge a decision based on its outcome. Vedantam also argues that human psychology is such that we want to put off a potentially bad outcome (losing in regulation time) for as long as possible, even when that decision is irrational. One student suggested another reason for the criticism: In his words, James is "the money". He was criticized because he is the star player. He should not be passing up opportunities for shots when the game is on the line. This is, of course, a ridiculous reason for criticism. A fan of the Cavaliers would want the team to do whatever gives them the best chance to win the game, be it in regulation time or OT, with James or Marshall taking the final shot. This example implores the students to answer a pertinent question in an objective, quantitative, fashion.

Example 4: Interpreting Regression Coefficients Using NBA Player Data

Data was assembled on 331 NBA players for the 2008-09 season. Admittedly, this dataset requires some knowledge of basketball. Information on the players included conventional statistics (e.g. points per game (PPG), minutes per game (MPG)), and new metrics, like *adjusted plus-minus* (ADJPM). We discuss the shortfalls of using conventional statistics to evaluate players. They do not capture many of the things which happen that help a team win. For example, a player who "boxes out" helps his teammates obtain a rebound, but gets no numerical credit for his action. Or, a player who manages to place a hand in his opponent's face during a shot lowers his opponent's chance of scoring, but this does not appear in the box score. This helps the students understand the importance of using appropriate metrics in any analysis.

Here, we only consider building models for PPG. First, consider univariate models for PPG which use offensive and defensive rebounds (OFFREB and DEFREB), respectively. Both show significant *positive* relationships. But, if we fit a multivariate model which uses both DEFREB and OFFREB, then both variables retain significance, except that OFFREB has a *negative* coefficient. Is this contradictory? No, it simply requires knowledge of how to interpret the coefficients from a multivariate model. The coefficient on OFFREB represents a *partial* change in PPG, holding DEFREB constant. Why is the coefficient on OFFREB positive in the univariate model? It is because OFFREB and DEFREB are quite highly correlated. Once DEFREB is accounted for, however, we see that players with higher OFFREB tend to score less. This is an example which illustrates a crucial point to the students: the context set by variables makes a huge difference to the meaning of coefficients in a model.

Another question arises from a plot of PPG vs. MPG, shown in Figure 3. Is there any curvature noticeable in this relationship? Indeed, the answer is 'yes', and this can be confirmed by testing a quadratic model term. But is this curvature explainable? Although PPG is not the perfect measure of a player, it does tell us something about his value. Moreover, playing time is not delegated randomly; coaches decide who plays more than others. If the relationship between PPG and MPG were linear, this would indicate ineffective decisions regarding playing time. In other words, if the relationship were linear, then the only difference in PPG between a player who plays 10 MPG, and one who plays 30 MPG, would be amount of playing time. This should not be, as the player who plays 30 MPG will tend to be the better scorer (this is *why* he is getting more minutes!). This example is valuable because it extends beyond the usual mechanical questions asked, and requires the students to think carefully about the system of variables being examined.

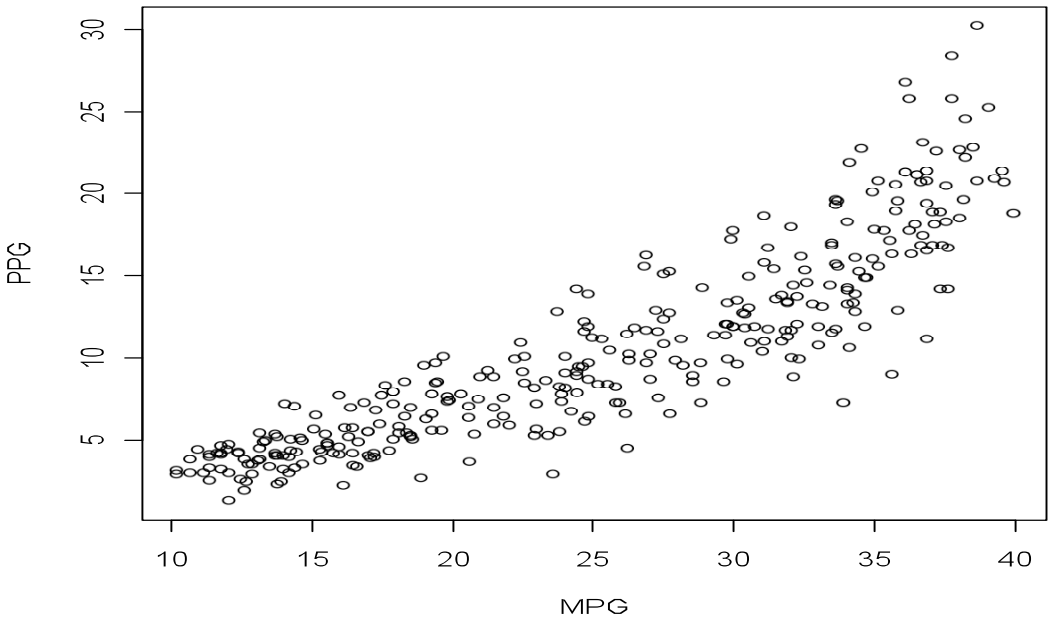

Figure 3. Scatterplot of PPG vs. MPG for 331 NBA players from the 2008-09 season

CONCLUSION

SASG was 60-65% lecture based. Ideally, that value would be reduced to roughly 50%. Two plans of action can be used to achieve this goal: (1) increase the number of in-class labs, from 7 to 10, say, and (2) Organize 1-2 guest talks. The professor should contact quantitatively inclined individuals associated with sports teams, or other individuals, suitable for giving a lecture. For SASG, two such people were found, a coach on Macalester's football team, and a graduate student who had worked for STATS (owners of the world's largest collection of sports data). By the time these individuals were contacted, however, it was late in the semester, and organizing a talk did not seem viable. We realize that, in the first teaching of a class, a lot of material must be developed, and some compromises need to be made. With regards to the lecture/lab ratio, a certain amount of lecturing is necessary, but it has become increasingly clear that when students carry out procedures on their own, the concepts register with them much more quickly and effectively.

Sports examples were selected from McClave and Sincich (2009), but it is not a sports themed book. There are many books on how statistics are used in sports (Albert & Koning, 2007; Winston, 2009; Albert, Bennett & Cochran, 2005), but there is a shortage of suitable *textbooks* for an *introductory* course. One exception is *Teaching Statistics Using Baseball* (Albert, 2003), which would be perfect for a baseball-focused course, like the one Albert has taught (Albert, 2002). Its focus, however, is too narrow for SASG. The development of appropriate introductory sports statistics textbooks will allow SASG-like courses to thrive at many colleges.

As we have illustrated, sports statistics courses need not be restricted to baseball examples. Some baseball data was used successfully, but in order to maximize the appeal of the course, it should only represent one portion of the many topics that can be covered. Using sports data to motivate statistical thinking is an effective avenue for engaging students. The possibilities for such a course are numerous, and we have only scratched the surface in this paper.

REFERENCES

Albert, J. (2002). A Baseball Statistics Course. *Journal of Statistics Education, 10*(2).

Albert, J. (2003). *Teaching Statistics Using Baseball*. Washington, DC: The Mathematical Association of America.

Albert, J., Bennett, J., & Cochran, J. J. (Eds.) (2005). *Anthology of Statistical in Sports*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Albert, J., & Koning, R. H. (Eds.) (2007). *Statistical Thinking in Sports*. Boca Raton, FL: Chapman & Hall.

Barnsley, R. H. (1988). Birthdate and Performance: The Relative Age Effect. *Paper presented at the Annual Meeting of the Canadian Society of Education*, Windsor, Ontario, June 1988.

Barnsley, R. H., Thompson, A. H., & Barnsley, P. E. (1985). Hockey success and birthdate: The relative age effect. *Canadian Association for Health, Physical Education, and Recreation, 51*, 23-28.

Barnsley, R. H., Thompson, A. H., & Legault, P. (1992). Family Planning: Football Style. The Relative Age Effect in Football. *International Review for the Sociology of Sport, 27*(1), 77-86.

Du, Q., Gao, H., & Levi, M. D. (2009). Born Leaders: The Relative-Age Effect and Managerial Success. Online: *http://ssrn.com/abstract=1365006*.

Gladwell, M. (2008). *Outliers: The Story of Success.* New York: Little, Brown and Company.

Matheson, V. A. (2007). Research Note: Athletic Graduation Rates and Simpson's Paradox. *Economics of Education Review, 26*(4), 516-520.

McClave, J. T., & Sincich, T. (2009) *Statistics* (11th edition). Upper Saddle River, NJ: Prentice Hall.

R Development Core Team (2009). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Online: http://www.R-project.org.

Thompson, A. H., Barnsley, R. H., & Stebelsky, G. (1991). "Born to Play Ball" The Relative Age Effect and Major League Baseball. *Sociology of Sport Journal*, *8*, 146-151.

Vedantam, S. (2007, June 18). Why We Don't Go For It. *The Washing Post*. Online: www.washingtonpost.com/wp-dyn/content/article/2007/06/17/AR2007061700968.html.

Winston, W. L. (2009). *Mathletics*. Princeton, NJ: Princeton University Press.