TEACHING THE FUNDAMENTALS OF STATISTICS WITH SPORTS DATA: SHOULD TEAMS WALK OR PITCH TO BARRY BONDS?

Jerome P. Reiter
Duke University, United States
jerry@stat.duke.edu

Sports data are commonly used to present topics from introductory statistics, such as exploratory data analysis and probability. They also can illustrate more subtle and complex statistical issues, such as selecting appropriate variables, making casual inferences from observational data, and specifying appropriate inferential populations. In this paper, I discuss how sports data can be used to engage students on such fundamental aspects of data analysis. I frame the discussion around the question posed in the title, a question which has generated much debate among baseball enthusiasts.

INTRODUCTION

Although introductory statistics courses come in a variety of flavors, almost all include lessons on study design, exploratory data analysis, and the concept of variability when making inferences. It is challenging to find real-life, interesting examples that are simple enough to illustrate these topics simultaneously. Often, teachers are forced to use examples homogenized for educational purposes. For example, the examples pretend the data were collected in a simple random sample when in fact they were not; they disregard complications due to missing data, noncompliance in experiments, and measurement error; they gloss over the scientific aspects of the subject that led the researchers to collect the data in the first place; or, they leave out discussions of costs and benefits associated with decisions. Such homogenization detaches the teaching of introductory statistics from what is done in practice, which could leave students ill-prepared to analyze data outside of the classroom setting.

Complex sports examples represent opportunities for students to analyze genuine data with minimal homogenization. The data collection is usually transparent; missing values, non-compliance, and measurement error are rarely problematic; the substantive questions can be clearly explained without extensive investment in the science of the problem; and, consequences of decisions are well defined and immediate. Complex sports examples often raise fundamental questions in statistical inference like those faced by statistical researchers in other areas but often not adequately covered in exercises and examples in text books. Should the data be treated as a census or a sample? Are the sampled observations truly independent? What variables are most appropriate for addressing the question of interest? What lurking variables exist, and how can they be dealt with? In addition to such fundamental questions, complex sports examples allow students to investigate univariate and multivariate relationships, as many teachers have discovered with simple sports examples.

In this paper, I discuss the utility of complex sports examples for teaching the entire process, including subtle aspects, of introductory statistics. I do so by illustrating an analysis of a question in baseball strategy: do opposing teams fare better when they walk Barry Bonds, or should they pitch to him? A proper data analysis of this question requires many thoughtful decisions, even though ultimately a sensible analysis can be done by comparing means and proportions. The decisions are ones that students, with guidance from teachers, can make on their own, thereby mimicking the entire process of data analysis.

I do not review the rules of baseball here, primarily to save space. For a summary of the game, see Albert’s (2003) text on teaching statistics using baseball and the introductory material in the anthology of sports articles edited by Albert, Bennett, and Cochran (2005).

DESCRIPTION OF ILLUSTRATIVE EXAMPLE

From 2001 through 2003, Barry Bonds of the San Francisco Giants strung together arguably the greatest individual seasons in the history of major league baseball. He set major league season records for home runs (73), walks (198), slugging percentage (86.3%), and on-base percentage (60.9%). He won the Most Valuable Player all three years, the first major league player ever to win three consecutively. His performance was remarkable by any standard.

Particularly surprising are the inflated walk totals in 2001 through 2003: no other players come close to those totals. They reflect increased use of a strategy for dealing with Bonds' prodigious ability to hit home runs, namely not to let him have a chance to hit the baseball and instead to walk him. Why risk pitching to a player who averaged one home run every 6.5 at bats in 2001 and every 8.7 at bats in 2002 and 2003? However, walking Bonds is not failsafe. Putting him on base via walk could actually help the Giants, since having runners on base greatly increases a team's chance of scoring runs. Plus, as prolific a batter as he is, over his career Bonds has made an out roughly 70% of the times he is not walked. Why not pitch to him since outs are the most likely outcome?

Thus, we are confronted with an interesting question of baseball strategy: Is it better to walk Barry Bonds or to pitch to him? To answer this straightforward causal question, students must deal with fundamental statistical issues frequently glossed over in the standard text book exercises or examples, but often present in genuine analyses. We now discuss these issues.

DEFINING TREATMENT AND OUTCOME VARIABLES

In any causal study, the first step is to define the treatment and outcome variables. Unfortunately, most text book exercises and examples take this step for students, which circumvents an important part of the analysis process. For this research question, the treatments include pitching to Bonds and walking Bonds. However, the treatment "walking Bonds" is ambiguous. One might consider only intentional walks, when the pitchers deliberately throw four balls far away from the strike zone. Or, one might include intentional walks and walks when the pitcher is trying to throw strikes but fails to do so for four pitches. Students undertaking this analysis must choose a definition. Their choice has implications for sample size (the number of intentional walks is lower than the number of total walks), ease of getting the data (some data sources do not distinguish between unintentional and intentional walks), and interpretability of results (intentional walks are an obvious reflection of the decision to walk Bonds, but what about "unofficial" intentional walks where the pitcher purposefully throws just outside the strike zone on four pitches?). Other treatment definitions are possible, each with different implications for analysis.

Students undertaking this analysis also must specify the outcome variable. Should it be a game outcome measure, such as whether the Giants won the game or how many runs they scored? Or should it be an inning-specific measure, such as some function of the number of runs scored in the inning when Bonds appears at the plate? Students also must relate their selected outcome measures to opposing managers' desiderata. For example, a manager seeking the strategy that minimizes the chance of any runs from scoring in an inning is concerned with the dichotomous outcome of zero runs or at least one run, whereas the manager seeking the strategy that minimizes the numbers of runs in an inning is concerned with the number of runs as the outcome. This is essentially specifying an appropriate loss function based on the costs of actions, a topic rarely encountered by introductory statistics students.

Finally, for outcomes based on runs in an inning, students need to be careful about how they count runs. They should include all batters who cross home base in the inning after the first pitch to Bonds, since effectively that first pitch initiates the treatment. What happens before Bonds steps to the plate is pre-treatment and so not part of the outcome. This distinction is an opportunity for teachers to discuss subtle differences between outcomes and background variables.

COLLECTING THE DATA

After deciding on treatment and outcome definitions (of course students can utilize multiple ones), students must collect data. This study requires observational data: decisions to walk or pitch to Bonds are made by baseball managers without random assignment. The data collection therefore involves investigation of various data sources, generally on the web. Students probably will have to enter data manually, which gives them a sense of the type of cost-sample size tradeoffs that data collectors have to consider.

Because this is an observational study, the walk-innings could differ from the hit-innings in ways that affect the chances of scoring runs. Students therefore need to decide which variables might be confounders before collecting the data, as do researchers in genuine studies. It can be costly (in terms of time if entering data manually or money of purchasing data from a company) to obtain additional variables after assembling the original dataset.

Some data may be missing, for example because of inactive web links. With sports data collected off the web, missing values are usually missing completely at random, so that ignoring the missing data does not bias inferences. Teachers can use the presence of missing data to talk about different types of missing data and their effect on inferences, yet students still can perform valid inferences without pretending no data were missing.

WHAT INFERENCES CAN BE MADE

Sports data on individuals can be obtained for entire seasons, as is the case for the plate appearances of Barry Bonds in 2001, 2002, and 2003. These are censuses; we have each one of Bonds' appearances in these seasons. This seems to make inferential statistics inapplicable. On the other hand, the plate appearances can be conceived as realizations of treatment applications from some hypothetical super-population of Bonds' plate appearances. This permits inferences to be drawn about that super-population with the usual techniques.

Can inferences be made, and if so to what population? Students are not likely to solve these dilemmas on their own. Teachers can use this opportunity to discuss foundational issues in statistical inference, which can be challenging to motivate outside the context of practical applications. As an example of such issues, in randomized experiments, it is possible to obtain inferences about the study population, but these inferences may be hard to generalize unless the study population is representative of broader populations. The same logic can be applied to observational studies.

DEALING WITH CONFOUNDING IN OBSERVATIONAL STUDIES

Even assuming Bonds' plate appearances in 2001 through 2003 are representative of the super-population of potential plate appearances, it remains challenging to estimate the causal effect of walking Bonds because the collected data are observational. Students need to check whether the distributions of the causally-relevant variables—which they specified and collected—in the walk-innings differ from those in the pitch-innings. If the distributions differ substantially, they need to adjust for differences to obtain reasonable inferences. At this stage, teachers can encourage students to learn about methods of creating matched treated and control group in a way that balances background variables (see Reiter, 2000 for an example) or learn about regression modeling. Alternatively, teachers can encourage students to focus on subsets of the data space (e.g., only plate appearances when no one is on base) with good balance in the background variables. If no reasonable balance can be found, teachers can recommend that students report the unhappy truth: no reliable conclusions can be made from the data. Although this is disheartening, it also mimics reality. Sometimes the data just don't permit one to answer the question. Arguably, this lesson is not driven home hard enough in introductory statistics courses.

Comparing distributions of causally-relevant background variables in the walk-innings and pitch-innings requires exploratory data analysis skills. Here, the analyses have an explicit purpose, namely to check distributional balance. This is an improvement over exploratory data analyses that simply ask students to describe distributions with no apparent objective other than to practice using statistical methods.

When asking students to compare distributions, I have found that many of them ask a difficult question: how large of a difference is enough to be a serious problem? Here there are few hard and fast rules, which is important to explain to students. Not everything in statistics is easily solved. One approach is to compute the regression (if appropriate) of the outcome on the predictor in the treated group, and predict the average outcome using the control group mean. The difference between the treated group's mean and the control group's predicted mean gives some sense of the effect of the difference in the distributions of the background variables. (I have not evaluated the merits of this approach; it should be treated with a healthy dose of skepticism.)

ADDRESSING THESE ISSUES IN THE BONDS DATA

Having raised many questions that students can grapple with, I now briefly present my answers to these questions. A more detailed analysis of these data can be found in Reiter (2004).

I define walks to include intentional and unintentional walks and hit by pitches. I do this primarily to increase the sample size in the walk-innings group. The result of walks, whether intentional or not, is identical: Bonds is given first base without hitting the baseball. Hence, it is not unreasonable to lump all walks together as one treatment. Furthermore, many unintentional walks and hit by pitches are in fact intentional walks in disguise, as pitchers decide to go through the motions of pitching to Bonds without really trying to make good pitches to him.

For the outcome variable, I use whether or not the Giants score at least one run in the inning after the first pitch to Bonds. For analyses involving averages, see Reiter (2004). I do not use game-level outcomes. These are too blunt to detect treatment effects, since in many games Bonds walks and hits at least once. They also are in a sense inconsistent with the treatment, since walks are fundamentally an appearance-level measurement whereas game outcomes are not.

I retrieved data from the web site http://www.cbs.sportsline.com, which has links to pitch-by-pitch game logs for each game in 2001, 2002, and 2003. Data for a few games were unavailable because of invalid web links; these games are excluded from the analyses. This should not skew results since these games are missing completely at random, that is, they are missing for reasons unrelated to the variables measured.

Using my knowledge of baseball, I decided that game situation could be a confounding variable. The Giants, and any team, are more likely to score runs when there are players at second or third base, and teams also are more likely to walk Barry Bonds, or any player, in the same situation so as to set up force plays if necessary. Table 1 displays the number of walks and plate appearances by Bonds in 2001, 2002, and 2003 in each of twenty-four game situations. The twenty-four situations are obtained by crossing the three possible out values and the eight possible configurations of players on base. As I expected, Bonds is walked with relatively high frequency when there are two outs and runners in scoring position (on second or third), which suggests confounding. To control for this, I analyzed the data separately in each game situation. I also decided to use only None On and First Only categories, since the other game situations have few observations.

Another potential confounding variable is the quality of the opposing pitcher. One might expect high quality pitchers to be willing to take the risk of pitching to Bonds, whereas low quality pitchers would not be willing to do so. Since low quality pitchers tend to give up more runs than high quality pitchers, this could affect the comparisons. A pitcher's quality can be measured by his earned run average (ERA) over his career, which roughly equals the total number of all runs allowed by the pitcher divided by the number of innings he has pitched. After performing side-by-side box plots within each game situation, I found that the distributions of ERA are very similar in the innings when Bonds walks and the innings when he does not walk (Reiter, 2004). The comparisons are fair with respect to pitcher quality.

Table 1: Bonds' plate appearances in 2001, 2002, and 2003. Entries are walks/appearances.

		None On	First Only	Second Only	Third Only	First + Second	First + Third	Second + Third	Bases Loaded
Zero Outs	2001	20/111	3/30	1/9	0/1	1/14	1/7	1/3	0/1
	2002	32/116	7/25	2/4	1/2	2/9	1/3	2/2	0/3
	2003	27/150	4/23	2/5	0/0	4/14	0/0	4/4	0/4
One Outs	2001	17/88	19/84	12/23	3/12	5/15	2/5	2/2	0/3
	2002	22/86	13/45	15/21	12/13	4/18	2/7	4/4	0/4
	2003	12/60	11/41	5/10	1/1	2/19	4/8	9/10	1/5
Two Outs	2001	41/151	14/34	13/18	5/7	9/15	4/9	1/1	1/5
	2002	36/117	16/48	14/21	1/6	5/15	7/11	1/1	0/4
	2003	15/76	17/52	16/26	8/11	9/17	1/3	5/6	0/2

Since the primary potential confounders are no longer an issue, I proceeded to compare the outcomes in the walk-innings and pitch-innings. Figure 1 displays the percentages of innings in which the Giants score at least one run after Bonds comes to the plate for the None On and First Only situations. Within each situation, the left bar shows the percentage when Bonds is walked, and the right bar shows the percentage when Bonds is pitched to. Each bar represents the combined percentage obtained by pooling the three years of data. The annual percentages are above the bars, going from 2001 at the top to 2003 at the bottom. For example, in 2001 the Giants scored in 6 of the 20 innings in which Bonds was walked with none on and no outs; in 2002 they scored in 17 of 32 such innings; and, in 2003 they scored in 14 of 27 such innings. Thus, the Giants scored when Bonds was walked with none on and no outs a combined (6+17+14)/(20+32+27) = 47% of the time.

Pooling the data across years simplifies comparisons of the strategies. Additionally, the combined percentages are based on larger numbers of innings than the annual percentages, which improves our ability to differentiate the effectiveness of the strategies. A drawback to pooling the data is that it masks any differences across years. This issue can be discussed with students.

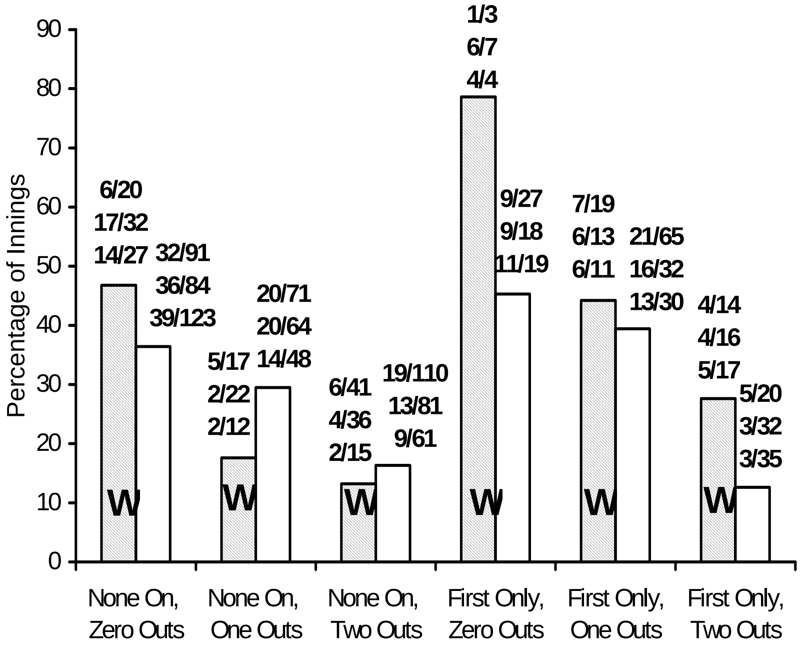

Figure 1: Percentage of innings Giants score at least one run after first pitch to Bonds. Bars with a "W" show innings in which Bonds walked. The top fraction above each bar is the percentage of innings with at least one run scored for 2001; the middle fraction is the same percentage for 2002; and, the bottom fraction is the same percentage for 2003.

At first glance, the combined percentages suggest competitive advantages for each strategy. With none on and at least one out, walking Bonds seems more effective than pitching to him. This may be because in these situations, the risk that Bonds hits a home run outweighs the risk that he scores when put on first base. With none on and no outs, walking Bonds seems less effective than pitching to him. This suggests that avoiding Bonds' home run power is outweighed by beginning an inning with a free pass. With one man on base, pitching to Bonds seems to be the better strategy. The walk advances the runner on first to scoring position, and it may be that the risk of that runner scoring outweighs the risk of Bonds driving in the runner from first.

These percentages are based on a limited number of plate appearances. Suppose there is no difference in the true probabilities of the Giants scoring when Bonds walks or hits. Could these apparent differences be plausibly explained by random chance? To answer this question, we conceive of a hypothetical population of Bonds' plate appearances under the same conditions that existed in 2001 through 2003, and we consider the plate appearances in 2001, 2002, and 2003 a random sample from this hypothetical population. Under this framework, the answer to our question is "not likely" for some situations and "entirely plausible" for others. When we combine the three years of data, the p-values for two-tailed statistical hypothesis tests are small for None On Zero Outs (p-value = .08), for None On One Outs (p-value = .04), and for First Only Two Outs (p-value = .04). The p-values associated with the tests for None On Two Outs and for First Only One Outs are both much greater than .10, indicating walking and pitching to Bonds in these situations plausibly could be equally effective. There is little data for First Only Zero Outs, although a simple examination of the differences suggests pitching to Bonds is a better option than walking him in that scenario.

CONCLUDING REMARKS

This analysis encompasses all the steps students must take to address a genuine question, without hiding the complications. Other complex sports examples, especially causal questions based on strategy, afford similar opportunities. For example, what is the advantage to winning a coin toss in overtime in football? Does fouling a poor free throw shooter in basketball result in fewer points on average than letting him or her shoot field goals? Students can be highly creative in forming these questions, especially those who appreciate sports.

I believe it is important that examples match statistical practice as closely as possible. We should not be finding data to illustrate techniques; we should be finding techniques to analyze data. If we as statistical educators emphasize the entire process of statistical analysis and eliminate as much homogenization of examples as possible, I firmly believe we will see students' knowledge and appreciation of statistics deepen.

REFERENCES

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

Albert, J., Bennett, J., and Cochran, J. J. (Eds.) (2005). *Anthology of Statistics in Sports.* ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

Reiter, J. P. (2000). Using statistics to determine causal relationships. *The American Mathematical Monthly*, 107, 24-32.

Reiter, J. P. (2004). Should teams walk or pitch to Barry Bonds? *Baseball Research Journal*, 32, 63-69.