An Application of Statistics: Using the "Moneyball" Story in a Basic Statistics Course

Samuel Wathen; Wall College of Business; Coastal Carolina University; Conway, South Carolina, 29526
Dennis Rauch, Coastal Carolina University; Wall College of Business; Coastal Carolina University; Conway, South Carolina, 29526

ABSTRACT

The purpose of this paper is to provide a teaching guide to a set of supplementary materials for a beginning statistics course using the book "Moneyball" by Michael Lewis. It provides a great business context for using statistics to gain a better perspective of the world. The book helps to illustrate the importance of measurement, data collection, hypothesis testing, model building, application of the scientific method, and, most importantly, developing a conversation of questions, hypotheses, results of analyses, model building, that is pertinent to the management of a business in a very competitive environment.

INTRODUCTION

A first course in statistics can be a challenge for the student and instructor. One of the instructors' challenges is student lack of motivation or interest to learn about statistics that is, perhaps, related to the question of relevancy. However, a general knowledge of statistics, probability, and the scientific method, as they relate to decision making and policy making is a very real value-added product of a statistics course.

Unfortunately, many courses are designed to create the knowledge of basic statistics, their calculation and interpretation. Indeed, the difficulty is relating the basic and simple, elementary statistics to something that is relevant. As instructors, we struggle to find a context that makes the statistics meaningful. One classic article [1] is "The Median is not the Message" by Stephen Jay Gould. The article clearly, at least to these authors, uses the positively skewed distribution of lifespan after a cancer diagnosis to develop a conversation about using the distribution to explore variance and the reasons for the variance in the distribution. The statistics provide an objective basis to continue a conversation about life span and life after a diagnosis of cancer. Our careers as instructors of statistics are made easier by articles like these.

This paper explores another source of applied statistics that is interesting, relevant, and supported by widely available data [2]. The source is the book "Moneyball" by Michael Lewis [3]. This paper will explore and propose a guide to using "Moneyball" and the Major League Baseball data as supporting materials for a first course in statistics. There are other supporting materials that will also be noted in this paper.

"Moneyball": The book

"Moneyball" (MB) has several themes and stories. They include a brief biography of Billy Beane's baseball career, a comparison of a traditional approach and a new data based approach to managing a professional baseball team, and stories of several professional baseball players whose careers were extended because of a new data based approach to professional baseball. However, the theme that is most useful to a beginning course in statistics encompasses the essence of our discipline as an extraordinary application to the evolution of an entertainment industry, professional baseball.

A major challenge of using MB in a beginning statistics course is that all students are not familiar with baseball. So the challenge is to define baseball in a larger context, which is the competitive environment of an organization, corporation, or company. And how an organization can remain competitive with limited resources. This burden falls on the instructor. The challenge is to use the book in the broader context while limiting baseball jargon to the very necessary references. The authors do not claim to know exactly how to face this challenge but have explored the challenge in the Spring Semester of 2013. So, in the context of organizations seeking to be successful with limited resources in a competitive environment, a class discussion is useful to establish the basis for using MB. The following is a proposed sequence of questions for a classroom discussion that can frame the context of the book.

- What defines organizational success? Making a profit.
- How do organizations make a profit? (Revenue-costs) > 0
- How is revenue generated? Customers' decision to purchase a product or service.
- How are costs determined? Player salaries.
- What attracts customers? Winning baseball teams
- How do baseball teams win games?

So how do baseball teams win games? What are the hypotheses? What is the science? As the baseball industry has evolved, the marketplace for talented baseball players has become very competitive. Players sign contracts worth millions of dollars. And the teams with the largest budgets to spend on players are able to buy the best talent.

So is it correct that an organization must simply hire the most expensive players to win the most games? Here is an opportunity to test a hypothesis, to start the scientific method, and to show the relevancy of knowing how to apply statistics. The proposed hypothesis is: The teams with the highest payroll win the most games. Using the Major League Baseball (MLB) data, students can do the analysis. All analyses in this paper are based on the 2011-2012 seasons.

			WINS				
Payroll (Millions $)	**50-59**	**60-69**	**70-79**	**80-89**	**90-99**	**100-110**	**Grand Total**
35-52	1		2	1	1		5
52-69	1		4	3	3		11
69-86		3	4	1	5		13
86-103		3	2	4	1		10
103-120		2	1	3	3		9
120-137			2	1	1		4
137-154			1				1
154-171				1	1		2
171-188		1		1		1	3
188-205					2		2
Grand Total	**2**	**9**	**16**	**15**	**17**	**1**	**60**

So the process of discovery starts. Using elementary statistics, students can summarize the data and start a conversation about this industry. Specifically, do the statistics indicate that the organizations with the highest payroll are more likely to be successful, that is, win the most games? It is clear that the

hypothesis is incorrect. The null hypothesis is rejected and this result should stimulate further inquiry. So, what is the continuing conversation? How are the seventeen organizations with varying payrolls successful, that is how do they win more than ninety baseball games? How does an organization recruit, hire, and train employees that lead to success?

So now the fun begins. And this is really part of statistics, measurement. How does an organization measure employee potential and talent? And what measures of an employee are most directly related to organizational success? What employee actions are most highly correlated with winning games? How does an organization select employees?

Measuring employee talent in baseball has been, and probably continues to be a subjective assessment based on limited, biased, qualitative data. A scout or interviewer travels and observes potential employees. Paul DePodesta, a Harvard graduate and the current vice president of player development and scouting for the New York Mets, …"was fascinated by irrationality, and the opportunities it created in human affairs for anyone who resisted it. … the market for baseball players … was far more interesting than anything Wall Street offered. There was, for starters, the tendency of everyone who actually played the game to generalize wildly from his personal experience. People always thought their own experience was typical when it wasn't. There was also a tendency to be overly influenced by a guy's most recent performance: what he did last was not necessarily what he would do next. Thirdly—but not lastly—there was the bias toward what people saw with their own eyes, or thought they had seen. The human mind plays tricks on itself when it relied exclusively on what it saw… There was a lot you couldn't see when you watched a baseball game" [3, p.18].

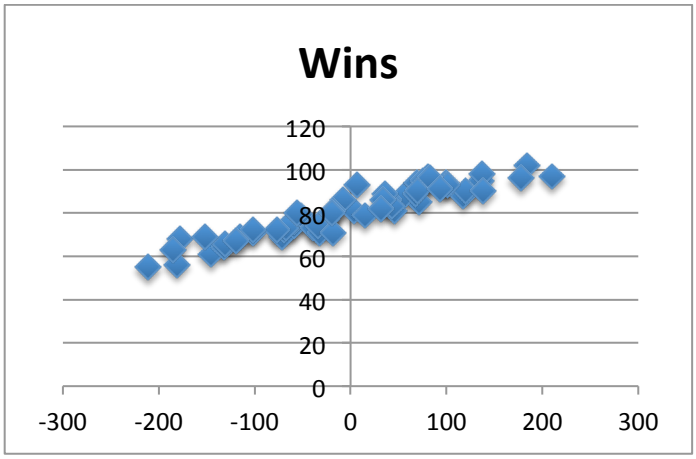
However, observers of baseball organizations have been recording employee performance data for some time. Some of the performance measures are hits, outs, homeruns, fielding errors and many more. Still the question is, what measures of performance are most closely associated with winning games.

Enter Bill James who authored the Baseball Almanac. Mr. James was not a statistician; he was a writer and his topic was the baseball industry. He was the science behind the industry. He wanted some one to prove their hypotheses, their hunches, and their guesses about baseball. Conjecture was his enemy. James' approach to baseball was that "… everything from on-field strategies to player evaluation was better conducted by scientific investigation—hypotheses tested by analysis of historical statistical baseball data—than by reference to the collective wisdom of old baseball men. By analyzing baseball statistics you could see through a lot of baseball nonsense. For instance, when baseball managers talked about scoring runs, they tended to focus on team batting average, but if you ran the analysis you could see that the number of runs a team scored bore little relation to that team's batting average" [3, pp 56-57]. Based on MLB data for the 2011-2012 seasons, the correlation coefficient is 0.767 or a R^2 of 59%. Conversations about the industry needed to be based on premises substantiated by data and statistics. And the conversation was about how a baseball organization can be successful. Selecting players based on batting average was not the best measure of employee performance.

Managing for Success

In the baseball industry, the number of wins defines success in a regular season. Therefore how do employees generate wins? Bill James was not a statistician, but he had data and experimented with performance statistics to develop interesting predictors of "number of wins". One premise that he tested is that if a team scored runs, and more runs than the opponent, the team would win games. Can

this premise be verified? Again, we can turn to the data. Plotting wins in a season against runs differential we get:



For this relationship, the correlation coefficient is 0.94, the coefficient of determination is 88%, and the equation is Wins = 80.93 + 0.1078(runs scored-runs scored against). So the statistics support and prove a point. Winning teams score runs, and they score more runs than their opponents. In addition, the results of the statistical analysis show that, on average, a team needs to score at least 84 more runs than their opposition over a season to win 90 or more games.

So the conversation continues. What employee actions generate runs? Further, how does an organization assemble a team to generate runs? What organizational policies and guidelines should be established to be successful?

Again enter Bill James, measurement, and the controversy of the "walk". Traditionally, walks were considered a non-event. It was not counted or measured or considered a contribution to producing runs and being successful. Batting average does not account for walks. Mr. James experimented scientifically to arrive at a model that would evolve into decision-making policy. The model was:

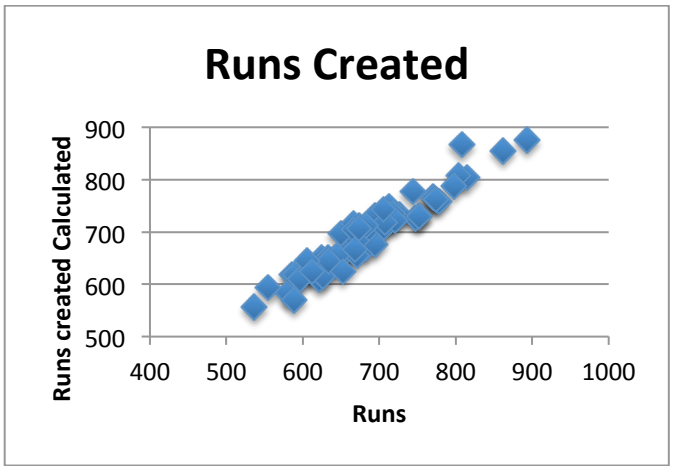
Runs Created = (hits + walks)*total bases/(at bats plus walks).

There was nothing elegant or principled in the way he went about solving the problem. He apparently did not know about regression analysis. He simply tried various equations or models until he found one that closely calculated a team's season run totals. Crude as it was, the equation could fairly be described as a scientific hypothesis: a model that would predict the number of runs a team would score given its walks, singles, doubles, etc. He could test the model using data from past seasons and compare the results to the runs the team scored that season. If the actual number of runs scored by the 1975 Boston Red Sox differed dramatically from the predicted number, his model was clearly false. If they were identical, James was probably onto something. As it turned out, James was onto something. His model came far closer, year in and year out, to describing the run totals of every big league baseball team than anything the teams themselves had come up with.

That, in turn, implied that professional baseball people had a false view of their offenses. It implied, specifically, that they didn't place enough value on walks and extra base hits, which featured

prominently in the "Runs Created" model, and placed too much value on batting average and stolen bases, which James didn't even bother to include.

When applied to the data (MLB data for the 2011-2012 seasons), the correlation between Bill James formula and actual runs is 0.96. For the 2011 and 2012 data, the formula overestimates by about +10 runs on average. Clearly, the formula proposed a conversation about how to be successful in creating runs and therefore winning games.



“But once again, the details of James's equation didn't matter all that much. He was creating opportunities for scientists as much as doing science himself. Other, more technically skilled people would soon generate closer approximations of reality. What mattered was (a) it was a rational, testable hypothesis; and (b) James made it so clear and interesting that it provoked a lot of intelligent people to join the conversation.” [3,p. 78] "The fact that the formulas work with the accuracy that they do is a way of saying there are essentially stable relationships between batting average, home runs, walks, other offensive elements—and runs," wrote James

As a result of applied statistical analysis, the way that organizations competed changed. It became clear that offense in terms of getting on base with hits or walks, and collecting bases through extra base hits was the clear criterion for deciding on the characteristics of employees that would make a baseball team successful.

Summary

The purpose of this paper is to provide a teaching guide to a set of supplementary materials for a beginning statistics course using the book “Moneyball” by Michael Lewis. It provides a great business context for using statistics to gain a better perspective of the world. The book helps to illustrate the importance of measurement, data collection, hypothesis testing, model building, application of the scientific method, and, most importantly, developing a conversation of questions, hypotheses, results of analyses, model building, that is pertinent to the management of a business in a very competitive environment.

We hope that this paper inspires statistics instructors to broaden their course objectives beyond the calculation and interpretation of statistics and show their students the real meaning and relevancy of statistics.

In addition to the book, "Moneyball" story has been made into a movie that is a good tool for generating student interest. For additional interest, to validate the realism of the movie, instructors can use the following true and false reference for the movie: www.mercurynews.com/news/ci_18937797. Instructors and students can access an interview with Billy Beane at: "'Moneyball': Tracking Down How Stats Win Games." *Fresh Air from WHYY.* September 23, 2011 (originally broadcast on May 28, 2003). [audio: 31 min 53 sec]. Another excellent reference is an interview with Bill James: "The Man Behind the 'Moneyball' Sabermetrics." *Talk of the Nation.* September 26, 2011. [audio: 17 min 6 sec]. We believe that all these references and sources support the discussion and conversation to support the use of "Moneyball" in a beginning statistics course.

References

[1] Gould, Stephen Jay. The Median Isn't the Message. 1985. Discover, 6(June), 40-42.

[2] mlb.mlb.com/stats

[3] Lewis, Michael. Moneyball. 2011. W.W. Norton & Company.

[4] James, Bill. 1977-1988. Annual Editions of the Baseball Abstract.