Statistics and the College Football Championship

Hal S. Stern

Hal S. Stern is Professor, Department of Statistics, Donald Bren School of Information and Computer Sciences, University of California, Irvine. The author thanks the TAS editor, Jim Albert, for encouraging this article and the many people with whom I've discussed this subject. Notable among the latter group are David Harville and Carl Morris. Thanks also to Haynes Philips for providing a copy of the Dickinson article and the discussants for pointing out a factual error or two in the original version.

Published online: 01 Jan 2012.

To cite this article: Hal S Stern (2004) Statistics and the College Football Championship, The American Statistician, 58:3, 179-185, DOI: 10.1198/000313004X2098

To link to this article: http://dx.doi.org/10.1198/000313004X2098

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://amstat.tandfonline.com/page/terms-and-conditions
General

Statistics and the College Football Championship

Hal S. STERN

Editor’s Note: Presently there is a controversy associated with the identification of a champion in United States college football. Teams cannot play a complete round robin and the best team is decided on the basis of a complicated Bowl Championship Series (BCS) rating system and the results of postseason “bowl” games. Because the rating problem is essentially a statistical problem, The American Statistician invited Hal Stern to describe the history of the college football championship and explain the contributions that statistical thinking might make. To respond to Stern’s article, we invited several authors of the BCS computer ratings systems and several statisticians familiar with ranking systems.

—James Albert, Editor, The American Statistician

The U.S. college football champion is determined each winter by the Bowl Championship Series (BCS), a set of four college football games and an associated ranking system that helps to determine the participants in the four games. One game each winter (the specific game rotates among the four participating games) hosts the national championship game between the top two teams in the BCS ranking. The BCS has been a controversial system since its implementation prior to the 1998 season with the most recent 2003 season producing a disputed championship, the very thing the BCS system was developed to avoid. This current article reviews the history of the college football national championship, the rise of the BCS, the BCS ranking system, and the contributions that statistical thinking can make toward improving the BCS. Though the problem of optimally ranking sports teams is a difficult one there is clearly room for improvement in the present system!

KEY WORDS: Quantitative reasoning; Rankings; Ratings; Sports.

1. ON SPORTS AND CHAMPIONSHIPS

Professional sports leagues in the United States tend to be organized as collections of small (four to eight teams) divisions that play unbalanced schedules (more games against opponents within the division than against other teams). This promotes strong local rivalries but implies that one cannot just compare records at the end of the season to determine a champion. Instead the regular season is used mainly to identify participants for a post-season championship tournament. It is the post-season tournament’s champion (e.g., the World Series winner in baseball and the Super Bowl winner in football) that is universally recognized as that season’s champion. Division or conference titles earned during the year are celebrated but not with the same level of enthusiasm as the championship.

This is of course not the only model for professional sports. For example, football (soccer) leagues in Europe are generally organized as 18- to 20-team leagues that play a balanced regular season schedule composed of home-and-away games with each team in the league. The regular season championship in such leagues is highly valued. There are also tournaments and competitions, both intranational (e.g., the FA Cup in England) and international (e.g., Europe’s Champions League), considered as separate events. The tournaments resemble the American post-season tournaments. Tournament titles are valued but not necessarily more than the regular season title.

American college sports (except football) have evolved a kind of compromise of the two models. University teams are organized in athletic conferences. The conference regular season title is usually based on a balanced schedule, though this is not always true now that conferences have grown in size. Regular season conference champions along with some of the best nonchampions are invited to a post-season national championship tournament. As in professional sports the post-season championship tournaments have become the most sought-after title, the one that is highly recognized. Unlike the professionals it seems that regular season conference championships remain highly valued.

College football’s most competitive level (Division I-A) is unique among American sports. There is no post-season championship tournament. (There are championship tournaments in the football divisions for smaller schools). Instead for many years the college football national champion has been crowned based on the final votes of two polls (the Associated Press poll of sportswriters and the USA Today/ESPN poll of coaches). In recognition of the fact that the championship is settled by polls rather than on the field, the college football title has long been known as the “mythical” national championship.

The remainder of this article considers the way in which college football currently selects its national champion and how ideas from the field of statistics might improve the process.

Hal S. Stern is Professor, Department of Statistics, Donald Bren School of Information and Computer Sciences, University of California, Irvine (E-mail: sternh@uci.edu). The author thanks the TAS editor, Jim Albert, for encouraging this article and the many people with whom I’ve discussed this subject. Notable among the latter group are David Harville and Carl Morris. Thanks also to Haynes Phillips for providing a copy of the Dickinson article and the discussants for pointing out a factual error or two in the original version.
2. AMERICAN COLLEGE FOOTBALL—PAST AND PRESENT

It is natural to start with a brief history of the search for a college football national champion.

2.1 The Early Days

A first key point is that there has long been interest in knowing which college football team is the nation’s best. This is not a modern phenomenon. The earliest recognized rating system designed to address the topic was developed by Frank Dickinson of the University of Illinois in 1924. Teams were awarded points for winning games, with the number of points awarded determined by the quality of the opponent. Dickinson stopped producing his ratings in 1940, citing his age, a declining interest in college football, and the large number of alternative systems that had been developed (Dickinson 1941). Thus, by 1940 there is evidence of great national interest in identifying the best college football team in the land.

2.2 The Rise of the Polls

In 1936 the Associate Press (AP) news agency began a national poll of sportswriters to identify the top teams. In 1950 United Press International (UPI), a competing news agency, developed a poll of coaches (currently run by USA Today newspaper and the ESPN cable television network). The two polls quickly became the most highly regarded source of information regarding the top teams. The top-ranked poll team (assuming the two polls agreed) was recognized by all as the national champion. The two polls agreed most of the time during these early years (with disagreements in 1954, 1957, and 1965).

During this era a number of college football “bowl” games began to develop. (The first bowl game, the Rose Bowl, actually dates back to the early 1900s but all the others began in 1935 or later). The bowl games are played after the regular season in the second half of December and the early part of January. Top teams are invited to play one additional game against another top team, usually from another region of the country. The games are major tourist attractions for the host cities. There are currently 28 bowl games each year. Initially the bowl games were viewed as separate from the business of identifying a national champion. But first the AP (in 1968) and then the UPI (in 1974) decided to select their champion after the bowl games were played. This raised the importance of the bowl games.

The two polls remain the final arbiters of the “mythical” college football championship to this day. In response to a couple of controversial split championships around 1990, conferences, teams, and bowls began to formulate alliances and coalitions designed to ensure that the top two teams would meet in one of the bowls allowing for an unambiguous champion. There were some successes during this period but still a number of “failures” where two different champions were crowned (due to the absence of a particular conference, team, or bowl from the arrangement).

2.3 The Retirement of Tom Osborne in 1997 and the Start of the BCS

In 1997 there were two undefeated teams at the end of the regular season, Michigan and Nebraska. Prior to the bowl games Michigan was ranked ahead of Nebraska in both polls. Michigan was committed to playing against a West Coast team in the Rose Bowl and thus not available to meet Nebraska in a championship game. Two things happened during the bowl season that year: (1) Tom Osborne, Nebraska’s long time coach, announced prior to the bowl games his intention to retire at the end of the season; and (2) Nebraska and Michigan both won their bowl games. When the polls were announced after the bowl games the writers had maintained Michigan as the top team but the coaches selected Nebraska, jumping them over Michigan despite the fact that both teams won. The resulting controversy over the split championship was one of the factors that led to the creation of the Bowl Championship Series or BCS. This agreement involved all of the major conferences and bowls and finally “guaranteed” (or so we thought) that the two best teams would meet each year to determine a unique national champion.

2.4 Why is There No Tournament?

Though not the main point of the article it is worth a short paragraph to consider the question of why there is no tournament for the top tier of college football. Among the reasons given by opponents of a tournament are: (1) the danger of injury to players in the additional games; (2) the pressure that additional games would put on the student-athletes’ academic programs; (3) a desire to avoid having the college football season run into a second semester; (4) a sense that the occasional split championship controversy is good for the game; and (5) recognition that the bowls have been good allies to college football. Though the issues raised on this list are valid concerns, none would seem to pose an insurmountable obstacle. In particular, the decision of many college football teams and conferences to add preliminary games in August and conference championships in December, as well as the experience of other college football divisions, seem to suggest that limiting the scope of the season is not in reality a major concern. It is similarly hard to argue that the presence of tournaments in other sports has decreased their popularity or reduced the joy fans have in arguing about who the true best team really was. Finally, it should be noted that every tournament plan that has been mentioned would try to make use of some of the existing bowl games while leaving the others no worse off than they are now. What then can be the reason for the absence of a tournament? One possibility is that having a championship tournament sponsored by college football’s lead organization (the NCAA) would require the member schools of the football-dominant BCS conferences to split the many of millions of dollars of revenue generated by such an event more equally with their fellow institutions of higher learning than is required by the current system.

3. THE BOWL CHAMPIONSHIP SERIES (BCS)

The BCS resulted from negotiations among six major athletic conferences (plus the University of Notre Dame), four major bowl games, and the television networks. The eight slots in the four participating bowl games would be earned by the champions of the six major conferences plus two at-large teams. A BCS rating system was developed to assist in the selection of the teams. The actual selection of teams is guided by a number of rules, including special rules that govern Notre Dame’s participation. The most important rule for our purposes is the
rule which stipulates that the two most highly ranked teams in the BCS rating system compete in the bowl game designated as hosting that year’s national championship game.

### Table 1. Final (pre-bowl) 2003 Bowl Championship Series (BCS) Rankings. Only the top ten teams shown. Details for computing the various components are provided in the accompanying text. SOS stands for “strength of schedule.”

|------|-------------|------------------|----------|-----------------------------------------------|----------------|----------|------------|------------|-----------|-------------|------|-------|
| 1    | Oklahoma    | (3,3)            | 3        | (1,1,2,5,1,1)                                | 1.17           | 11       | 0.44       | 1          | 5.61      | -0.5   | 5.11
| 2    | LSU         | (2,2)            | 2        | (2,2,2,1,2,2,2)                              | 1.83           | 29       | 1.16       | 1          | 5.99      | 5.99
| 3    | USC         | (1,1)            | 1        | (3,3,3,1,4,3)                                | 2.67           | 37       | 1.48       | 1          | 6.15      | 6.15
| 4    | Michigan    | (4,4)            | 4        | (7,4,6,5,3,5,5)                              | 4.67           | 14       | 0.56       | 2          | 11.23     | -0.6  | 10.63
| 5    | Ohio St     | (7,6)            | 6.5      | (6,6,4,4,8,6,7)                              | 5.50           | 7        | 0.28       | 2          | 14.28     | 14.28
| 6    | Texas       | (5,5)            | 5        | (5,9,8,7,4,8,10)                             | 6.83           | 20       | 0.80       | 2          | 14.63     | -0.1  | 14.53
| 7    | Florida St  | (9,8)            | 8.5      | (8,8,5,7,7,6)                                | 6.83           | 15       | 0.60       | 2          | 17.93     | 17.93
| 8    | Tennessee   | (6,7)            | 6.5      | (10,7,1,11,9,10,11)                          | 9.50           | 46       | 1.84       | 2          | 19.84     | -0.2  | 19.64
| 9    | Miami(Fl)   | (10,9)           | 9.5      | (9,5,7,9,10,11,9)                            | 8.17           | 13       | 0.52       | 2          | 20.19     | -0.4  | 19.79
| 10   | Kansas St   | (8,10)           | 9        | (16,12,12,12,6,13,13)                        | 11.33          | 10       | 0.40       | 3          | 23.73     | -1.0  | 22.73

3.1 Design of the BCS Rating System

The BCS rating system consisted initially of four elements: polls, computer ratings, a measure of schedule strength, and team record. A team’s rating is the sum of four numbers, one corresponding to each element. The average rank of the team in the coaches’ and writers’ polls is the first number used in the BCS rating. Because the writers were reluctant to play too large a role in determining the two teams to compete for the championship, the BCS organizers also included a number of computer ratings. A couple of points about terminology are needed here. The reference to these as computer ratings is a bit misleading: the ratings in question are determined by mathematically based algorithms for rating football teams. Because of the large number of teams involved the systems are implemented on computers, but that is not really the key feature. There is also some ambiguity in the terms ranking and ratings. Typically the term ratings is used when there is information contained in the numerical scores assigned to the teams and the term rankings is used when the primary goal is only to put the teams in order. We try to follow this distinction but it is difficult given the variation in usage throughout the literature. The number of quantitative rating systems included has varied from three to eight with the 2003 BCS including seven ratings: those of Anderson & Hester, Billingsley, Colley, Massey, New York Times, Sagarin, and Wolfe. These are discussed further below. The average rank of each team in the seven computer ratings (actually the worst ranking is dropped for each team) is the second number used in the BCS ranking system. The teams’ schedules are ranked for difficulty (using a measure that combines the records of each team’s opponents and the records of their opponents’ opponents) and that ranking divided by 25 is the third number in the BCS sum. Finally teams receive one point for each game that they lose. The system has been modified over time to correct perceived inequities. We return to this point below, for now we note only that the current version of the BCS includes one additional factor besides the four described thus far, a quality win deduction that removes points from a team’s BCS score for wins over other highly ranked BCS teams. The top ten teams in the final 2003 regular season BCS standings along with the different elements of the system are presented in Table 1.

3.2 A System Without a Guiding Principle

From its beginning the BCS ranking system has been hurt by the fact that its main objective is poorly defined. It is designed to select the “top two teams” but this term lacks a formal definition. The relative weights of the included elements were determined using trial and error to ensure that the BCS would have selected the top teams (as defined by an informal consensus) in earlier years. This is not a terrible approach, but it is worth pointing out that there are likely many combinations of the BCS ranking system elements that would pass such a test. How to choose among them? This is where it would be helpful to have in mind a particular set of principles or goals. For example, one might elect to identify the two teams judged most likely to defeat all the others at the end of the season. Other possible goals are described later. Without a specific goal it is all too easy to overreact to short-term fluctuations and random events. The events that have unfolded since the formation of the BCS in 1998 have exposed this as a major weakness in the BCS system.

There is a joke told in which an insurance company is described as an automobile traveling on the highway at high speed. The chief executive officer has the steering wheel, the marketing department is stepping on the gas, the chief financial officer is applying the brakes, and the actuary is in the backseat screaming directions using a map made by looking out the rear window. The events that have unfolded since the formation of the BCS in 1998 have exposed this as a major weakness in the BCS system.

There is a joke told in which an insurance company is described as an automobile traveling on the highway at high speed. The chief executive officer has the steering wheel, the marketing department is stepping on the gas, the chief financial officer is applying the brakes, and the actuary is in the backseat screaming directions using a map made by looking out the rear window. The events that have unfolded since the formation of the BCS in 1998 have exposed this as a major weakness in the BCS system.

- 2000: Undefeated Oklahoma was an obvious choice for the championship game. There were several once-beaten teams that were candidates to play Oklahoma. The BCS system selected once-beaten Florida State even though Florida State had lost to once-beaten Miami during the season. This bothered many fans (who were not quite so bothered that once-beaten Miami had lost to once-beaten Washington during the same season). It is at this point that a quality win bonus was added to the BCS. After the initial BCS rankings are developed each year an adjustment is

The American Statistician, August 2004, Vol. 58, No. 3 181
made for wins against other top BCS teams. Had the quality win bonus been in effect for 2000, Miami would have been chosen. This would turn out to be a theme for the BCS—the current year’s system would definitely have gotten last year’s decision correct!

- 2001: The following year it was undefeated Miami that was the obvious choice. This time the BCS selected once-beaten Nebraska over once-beaten Oregon despite the fact that Nebraska had lost badly in their last regular season game. Popular perception this time was that the computer ratings paid too much attention to the large margin of victory in Nebraska’s early season triumphs while not putting enough value on Oregon’s steady but unspectacular performances. It was decided that the computer ratings should not use the scores of the games.

- 2003: Controversy of a different type arose during this most recent season. There were no undefeated teams, but there were three once-beaten teams from strong conferences. How to choose among them? Let’s examine the data in Table 1. The human polls agreed that USC and LSU were the top two teams. The computers were equally insistence that Oklahoma and LSU were the top two teams. Oklahoma and LSU were selected to play in the national championship game which was won by LSU. The final coaches poll, as per the BCS agreement, elevated LSU to the top ranking. The writers’ poll, not obligated by the BCS agreement, maintained USC as its top ranked team. The BCS, a system expressly created to identify a single champion, had failed to do so. What next? It seems clear that another “tweak” of the rules is in order. The plans for the 2004 season have not yet been announced though a couple of possibilities include: (1) only using the full BCS formula if the polls do not agree on the top two teams; (2) only considering teams that won their conference championship (this would have eliminated Oklahoma in 2003 and Nebraska in 2001). When asked for his opinion about the outcome in 2003, BCS coordinator Mike Tranghese offered that “People know I’m an anti-computer person” hinting that the weight of the computer ratings would be reduced.

4. THE ROLE OF STATISTICAL THINKING

The problem facing college football is to find a method for selecting the top two teams at the end of the football season. Each team plays between 11 and 13 games against a wide range of competition. The wide range of competition means that just looking at the records of the various teams is not sufficient. A team with one loss in a powerful conference may well be stronger than a team with no losses that plays in a weaker conference. Choosing the top two teams in such an environment is clearly a difficult task. It is basically impossible unless one can define what one means by the “top two teams.” As mentioned earlier, one possible goal would be to identify the two teams that would be predicted to defeat all others in a post-season tournament. This is not, however, the only possible goal. Morris (1978) proposed that it would be ideal to try and predict what would have happened if a complete round-robin schedule (each team playing every other team) were to be played. Games that were played could perhaps be combined with statistical predictions for the unplayed parts of the round-robin. Another plausible goal is to try, as Dickinson did, to recognize the two teams that have “achieved” the most during the season. This of course requires a formal approach to measuring achievement! It is clear that different goals might lead us to identify different teams at the end of the year. A team that lost two games early in the year without its star player but won every subsequent game by 40 points might be the team most likely to defeat all the others at the end of the season but clearly did not achieve as much as an undefeated team in a reasonably strong conference.

The situation facing college football is not all that different from the one that faces a data analyst trying to identify a one-number summary of a distribution. The mean, median, and mode all have clear and compelling arguments for being the best summary. The best choice in a particular context depends on the objectives of the person who wants the summary. To help decide we often ask about the loss that will be incurred by choosing a wrong summary. Though formal specification of a loss function is unlikely in the college football world, a full discussion of the ramifications of different rating systems would clearly seem valuable. Some of the problems faced by the BCS (what to do if the polls and computers disagree, what game information should be used in the computer ratings, what to do if team A is ranked above team B but team B has defeated team A during the season) could have and should have been anticipated in advance. The following section provides a short discussion of the computer rating piece of the BCS; this discussion is intended to serve as an illustration of one area in which statistical expertise would prove valuable to BCS organizers.

5. COMPUTER RATINGS AND THE BCS

It is disappointing, if not surprising, that most of the blame for 2003’s BCS fiasco is being placed on the computer ratings. The distinction between human rankings and computer-generated mathematically based ratings is somewhat artificial. The computers are only systematically and emotionlessly applying a series of human-generated rules to determine the ratings of the teams. There is a great deal of research demonstrating that human opinions tend to be clouded by a number of consistent biases. There is a bias in favor of recent events, a bias in favor of familiar people/teams, and many other biases. In choosing the top teams it would seem desirable to have rules applied consistently instead of based on which games a particular poll participant happened to watch or remember hearing about. Despite this there is clear evidence that many reporters and college football officials are much more comfortable with the human polls than with the mathematically based ratings. For the moment we put this preference for human polls aside and consider the computer ratings part of the BCS in more detail.

5.1 A Ratings Primer

An entire article on just rating methods would be possible (and even quite interesting). Such an article would not in my opinion address the true difficulties that plague the BCS. Instead, we settle here for a short primer on statistical methods for rating teams to illustrate some of the issues that BCS organizers need to address. In particular, a range of rating methods that are based on different amounts of information are considered. Several references to the team rating literature appear below; interested readers are directed to Mease (2003) and Massey (2004) and the references therein for more information.
Table 2. Comparison of the Proportion of Correct Predictions Using Four Different Prediction Methods. The column labeled “Home” gives results for a method that always predicts the home team will win; “LS(1/− 1)” makes predictions based on least squares ratings with 1/− 1 game outcomes; “LS(scores)” makes predictions based on least squares ratings with actual game outcomes; “Oddsmaker” reflects the predictions of professional oddsmakers. Least squares ratings are used to predict the final two-thirds of each season with the first one-third of the season used to initialize the rankings.

<table>
<thead>
<tr>
<th>Sport</th>
<th>Years</th>
<th>Home</th>
<th>LS (1/− 1)</th>
<th>LS (scores)</th>
<th>Oddsmaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro football</td>
<td>1988–1993</td>
<td>0.58 (1342)</td>
<td>0.63 (862)</td>
<td>0.65 (862)</td>
<td>0.67 (1303)</td>
</tr>
<tr>
<td>Pro basketball</td>
<td>1985–1986</td>
<td>0.66 (1886)</td>
<td>0.69 (1255)</td>
<td>0.70 (1255)</td>
<td>0.71 (1827)</td>
</tr>
<tr>
<td>Pro baseball</td>
<td>1986</td>
<td>0.53 (3884)</td>
<td>0.56 (643)</td>
<td>0.56 (643)</td>
<td>0.55 (938)</td>
</tr>
<tr>
<td>College football</td>
<td>1992–1996</td>
<td>0.62 (3038)</td>
<td>0.69 (2104)</td>
<td>0.73 (2104)</td>
<td>0.75 (1551)</td>
</tr>
<tr>
<td>College basketball</td>
<td>1995</td>
<td>0.68 (2258)</td>
<td>0.72 (1724)</td>
<td>0.72 (1724)</td>
<td>0.75 (2068)</td>
</tr>
<tr>
<td>College hockey</td>
<td>1991–1992</td>
<td>0.62 (1405)</td>
<td>0.72 (998)</td>
<td>0.73 (998)</td>
<td>—</td>
</tr>
</tbody>
</table>

NOTE: Numbers in parentheses are number of games predicted.

One of the simplest rating methods and a useful starting point for any discussion of college sports is the ratings percentage index (or RPI). This is a device used by the NCAA (college sports’ governing body) to rank teams in a variety of sports. The RPI is a weighted average of a team’s winning percentage (with weight 0.25), the average winning percentage of its opponents (with weight 0.50), and the average of its opponents’ opponents winning percentage (with weight 0.25). The RPI trades off the two key aspects of a team’s performance: the achievement of the team (winning percentage) and the quality of the opposition. (The last two components of the RPI are the schedule strength measure used by the BCS.) The RPI is simple and easy to interpret. Of course, the RPI does have some weaknesses; for one thing it is possible to win a game and have your rating go down! Another problem is that the measure of opponent’s strength (their record) is clearly not the ideal measure. The RPI also does not take the site of the game into account.

At a next level of sophistication, but still ignoring the scores of games, there are a variety of paired-comparison methods. This includes, for example, the Bradley-Terry model (Bradley and Terry 1952) which postulates the existence of team strength parameters 𝜋𝑖 such that the probability that team 𝑖 defeats team 𝑗 is 𝑝𝑖𝑗 = 𝜋𝑖/(𝜋𝑖 + 𝜋𝑗). There are a number of methods for determining optimal values of 𝜋𝑖. One common approach is to maximize the likelihood of the observed game outcomes. There is a small problem with undefeated or winless teams but this can be easily handled in a variety of ways (see, e.g., Mease 2003). This approach improves upon the RPI by measuring opponents’ strength implicitly rather than with a simple measure like winning percentage. There are a wide variety of related methods available because the method of paired comparisons (comparing a large number of treatments two at a time) is used in a variety of fields of application. David (1988) is a comprehensive reference on paired comparisons. The majority of the BCS methods currently in use are of this type.

It is natural to ask what can be done if we allow ourselves to make use of game scores. One answer can be found in the long history of rating teams using least-squares methods (see, e.g., Leake 1976; Stefani 1977; Harville 1977; Stern 1995). The basic idea is to choose ratings 𝑅𝑖 for each team to minimize the difference between the observed outcome 𝑌𝑖𝑗 when team 𝑖 plays team 𝑗 and the outcome predicted by the ratings, that is, to minimize ∑games(𝑌𝑖𝑗 − (𝑅𝑖 − 𝑅𝑗))². Such methods can also be viewed as maximum likelihood ratings under a normal model for game outcomes. This approach can be easily modified to incorporate a home field advantage and to downweight the influence of large values of 𝑌𝑖𝑗 (see, e.g., Harville 2003). Both modifications are important. The least squares idea can also be used without the scores by defining the game outcome 𝑌𝑖𝑗 as 1 or −1 depending on which team wins the game.

It is possible to increase further the amount of information used in our rating system by including results from previous seasons. This is done by Harville (1980) and Glickman and Stern (1998). Such methods begin each season by recognizing that the final ratings of the previous year contain some information about the quality of the teams for the coming year. Though this is undeniably true one can easily imagine that this would be an undesirable feature for a system whose goal is to unbiasedly select the top teams for the current year.

5.2 The Value of Information

One obvious way that statistics and statisticians might be of assistance to the BCS process is by evaluating the many rating methods and identifying their strengths and weaknesses. The discussion of the previous section begins this process. To illustrate how statistics can inform our choice of rating methods, let us suppose for the moment that we agree the goal of a rating system is to accurately predict future games. We repeat here that this is only one possible goal and not even the goal that most people would use to describe the BCS’s mission. Given that goal, however, it is possible to apply statistical thinking to assess the various methods.

Table 2 gives the proportion of game outcomes predicted correctly in a variety of sports by a sequence of increasingly sophisticated methods. The first column gives the proportion of games correctly predicted knowing only the site of the game; that is, this method always predicts the home team will win. This is clearly a crude system because it does not use any information about the relative strengths of the competing teams. The fact that the proportion of correct predictions is always greater than 0.5 reinforces the notion that the site of the game is an important piece of information. (Interestingly, this important piece of information is excluded from most of the BCS computer ratings.) The second column is a form of compromise between least-squares and paired comparisons methods. It uses the least-squares idea, but replaces the score of each game by +1 or −1 depending upon whether a team won or lost its game. This column also uses the
site of the game. Note that using the results of past games improves the accuracy of our predictions by a fair amount as would be expected. In these experiments no least squares predictions were attempted until one-third of the season was completed; this initial portion of each season was used to initialize the ratings. Then each week’s games were used to update the ratings before predicting the next week’s games. The third column indicates the value of adding the scores to our least squares rating method. In each sport there is some benefit to incorporating the score of the game if we would like to predict future outcomes. The biggest increase is in college football! The final column of Table 2 may be thought of as an upper bound; it reflects the frequency with which professional oddsmakers’ predictions are found to be correct. Least squares ratings do quite well, near but not quite at our upper bound. The RPI is not included in Table 2. Separate experiments in college basketball and college hockey (not shown) suggest that the RPI (69% correct predictions) is less effective than least squares ratings.

Table 3 provides information about the value of information from previous seasons. Again, the aim here is not to argue that such information should be used but only that statistical thinking allows one to provide objective data about the value of such information. The results in Table 3 are based on a crude version of the state-space model for rating teams used by Glickman and Stern (1998). In the crude version of the model applied here the data for each year are fit by a normal-theory least squares model and the parameters/ratings for the different years are modeled as a random walk with between-season variance $\sigma^2_s$. The value of $\sigma^2_s$ is fixed at a number of different values for the prediction experiments in the table. The resulting models are applied to forecast games in the second-half of nine professional football seasons in the late 1980s and early 1990s. When $\sigma^2_s$ is small, the team ratings are assumed to vary only a little from year-to-year. When $\sigma^2_s = \infty$, the ratings from one year have no effect on the predictions the next year. The data show that some sharing of information from one year to the next (i.e., values of $\sigma^2_s$ in between the two extremes) improves predictive performance. A complete data analysis in Glickman and Stern (1998) finds that the best estimate for $\sigma^2_s$ is in the same range (6 to 9) that provides optimal predictions in Table 3. It should be emphasized again that there is no claim that prediction is an appropriate goal for the BCS. Tables 2 and 3 merely demonstrate how statistical thinking can be used to support good decision making in the context of a specified goal.

### 5.3 Computer Ratings in the BCS

The previous section provides a quick (but hopefully clear) discussion of a number of quantitative rating methods. It is interesting to briefly review what we know about the BCS components in the context of that discussion. The information provided here was obtained by visiting the relevant Web sites (pointers to the various rating sites can be found at the Massey or Billingsley Web sites; www.mratings.com and www.cfrc.com, respectively).

Sagarin’s BCS rating is based on the chess rating system developed by Elo (1978). Billingsley is a long-time observer of college football who has developed a series of rules for updating a team’s rating based on its most recent result. Interestingly, Billingsley’s method does use information from the previous seasons to initialize the ratings whereas most (all?) of the other methods do not! Wolfe’s ratings are described as being based on the Bradley-Terry method. Colley’s method is described in great detail on his Web site (www.colleyrankings.com); it turns out to be closely related to least-squares ratings with $1/\sigma^2 - 1$ outcomes though that is not the motivation presented. Colley’s method also includes a bit of what statisticians generally refer to as shrinkage estimation. Massey’s ratings are well-documented; they represent a variation on the Bradley-Terry paired comparison idea. (Incidentally, both Massey and Sagarin maintain ratings with and without score information on their Web sites). The Anderson & Hester and the New York Times ratings provide very few details. In the case of the New York Times it has been reported that even the people who produce the ratings each week don’t know the algorithm. If true, this is disturbing especially given that many observers find the week-to-week changes of the New York Times’ ratings to be quite unusual. The week-to-week changes in the New York Times’ results suggest that they may be giving extremely large influence to recent results compared with the other ratings.

Given that scores are not allowed to be used in the BCS computer ratings, each of the systems now in use basically represents a formal quantitative combination of team performance and schedule strength, which is based on win/loss information. The variation we see among the seven included ratings is due to choices made about: whether to include/exclude site information, whether schedule strength should be measured by examining the records/ranking of teams when the game was played or revised to account for later performances, whether recent games are weighed more than early games, and a variety of other ad hoc choices (constants, shrinkage). It is interesting to note that the pre-BCS system using only polls was often criticized for not considering schedule strength. The BCS ranking includes schedule strength implicitly within each computer rating and then again explicitly as an additional term. Could the schedule strength pendulum have swung too far in the opposite direction? Is it possible that the BCS now over corrects for schedule strength? The argument here is that it would be beneficial if BCS organizers were aware of the issues involved so that an informed choice could be made.

<table>
<thead>
<tr>
<th>Prediction method</th>
<th>Percent correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oddsmaker</td>
<td>66.3</td>
</tr>
<tr>
<td>State-space ($\sigma^2_s = \infty$)</td>
<td>63.6</td>
</tr>
<tr>
<td>State-space ($\sigma^2_s = 100$)</td>
<td>63.9</td>
</tr>
<tr>
<td>State-space ($\sigma^2_s = 9$)</td>
<td>64.9</td>
</tr>
<tr>
<td>State-space ($\sigma^2_s = 6.25$)</td>
<td>64.9</td>
</tr>
<tr>
<td>State-space ($\sigma^2_s = 1$)</td>
<td>62.0</td>
</tr>
</tbody>
</table>
6. SUMMARY: WHAT CAN STATISTICIANS DO?

I take the goal of this article as being to review the BCS and to ask how statistics might be used to improve the situation. The latter question is interesting to think about.

The BCS is not the first time that statisticians have been excluded from the discussion of a policy question of great public interest, and it probably won’t be the last time this happens either. Also the BCS is surely not the most important policy question currently lacking statistical input. Why worry about it at all? One reason is that the BCS raises familiar questions about how statistics and statisticians can get themselves invited to the table when policies based on quantitative data are discussed. Moreover, the BCS raises those questions in front of a relatively large audience of sports fans. As a discipline, statistics focuses on the collection, analysis, interpretation, and reporting of data. These are the key elements facing the BCS organizers and thus statistical input is essential.

One natural reaction is to suggest that statisticians should be put in charge of developing the mathematically based rating systems that should be included in the BCS ranking system. However, I think it would be a mistake for statisticians to claim a unique ability to develop the ideal rating system. I’ve argued that the best rating system must depend on the goal for which the system is being developed. It is unlikely that we could agree as a discipline on the correct goal and besides that’s not necessarily our decision anyway.

Instead perhaps we can better frame the public debate by reinforcing some of the very basic ideas from our field. We should be asking the questions that drive the public discussion such as:

- What is the objective of the BCS rankings? The discussion of ratings systems in Section 5 focuses on prediction as the objective. A recent article by Mease (2003) in this same journal focuses on how closely a computer rating system mimics the human polls. This seems to match what the BCS folks would have liked . . . at least for this year! The bottom line is that for the BCS to succeed a goal or objective must be identified.

- What sources of information should be included in the BCS rankings? The discussion of mathematically based rating systems in Section 5 identifies a number of key issues in this vein. These issues clearly need to be discussed in the context of “computer” ratings but they should also be discussed in the context of the human polls. Who votes in these polls and what information is provided to the voters?

- How sensitive is the BCS method to slight changes or variations in the data? At the least answering this question would require that the details of the existing rating systems be revealed to a group of experts able to judge their sensibility.

If we can get sportswriters and fans asking the above questions, then statistics will have contributed mightily to reforming the BCS.

[Received May 2004. Revised June 2004.]

REFERENCES


Discussion

Kenneth Massey

1. THE BCS CHALLENGE

The Bowl Championship Series has granted me the singular experience of being an academic suddenly thrust into the world of jocks, executives, and media. I produce the Massey Ratings, a computer-implemented mathematical algorithm for ranking college football teams that has been a component of the BCS formula since 1999. As a college football fan, I am honored to be part of the system, but as a mathematician this role has more often been a source of frustration. In his article, Hal Stern discussed the instabilities and weakly defined objectives of the current BCS formula. Reactionary patches have been successful only in hindsight, and recent debacles have many fans justifiably upset with the BCS process.

Kenneth Massey is Visiting Assistant Professor, Mathematics Department, Hollins University, Roanoke, VA 24020 (E-mail: kmassey@hollins.edu).