The Journal of Gambling Business and Economics (2009) 3 2 63-87

SEEDING IN THE NCAA MEN'S BASKETBALL TOURNAMENT: WHEN IS A HIGHER SEED BETTER?

Sheldon H. Jacobson Ph.D[*]
University of Illinois
Department of Computer Science
Simulation and Optimization Laboratory
201 N. Goodwin Avenue (MC258)
Urbana, Illinois 61801, USA
shj@uiuc.edu

Douglas M. King, M.S
University of Illinois
Department of Industrial and Enterprise Systems Engineering
Simulation and Optimization Laboratory
104 S. Mathews Avenue (MC238)
Urbana, Illinois 61801, USA
dmking@uiuc.edu

ACKNOWLEDGMENTS

The computational results reported were obtained with support from the Simulation and Optimization Laboratory in the Department of Computer Science at the University of Illinois (at Urbana-Champaign).

ABSTRACT

A number of methods have been proposed for predicting game winners in the National Collegiate Athletic Association's (NCAA) annual men's college basketball championship tournament. Since 1985, more than 70% of the teams in the fourth, fifth, and sixth rounds of the tournament have been high-seeded teams (i.e., teams assigned seeds of one, two, or three); a method that can accurately compare two such teams is often necessary to predict games in these rounds. This paper statistically analyzes tournaments from 1985 to 2009. A key finding is that there is an insignificant difference between the historical win percentages of high-seeded teams in each of the fourth, fifth, and sixth tournament rounds, which implies that choosing the higher seed to win games between these seeds does not provide accurate predictions in these rounds, and alternate predictors or methods should be sought. Implications on gambling point spreads are discussed.

Keywords: Statistical Hypothesis Testing, Sports Predictions, Sports Betting, NCAA Basketball.

INTRODUCTION

The annual men's college basketball championship tournament sponsored by the National Collegiate Athletic Association (NCAA), herein referred to as

[*] Corresponding author

the tournament, attracts considerable attention from the general public, resulting in a phenomenon referred to as *March Madness*. Opportunities to gamble on the outcome of tournament games are plentiful, with an estimated US$2.25B being wagered on the 2007 Final Four through illegal sources alone (McCarthy, 2007). In addition to traditional sports betting venues, office pools are a popular way to participate. In an office pool, a bettor must predict the winners of all tournament games, including games in the second and later rounds, whose participants have not yet been decided. Points are allocated based on the number of correct predictions, and bettors with the highest point totals are declared the winners. Based on a survey of betting behavior in twenty-four office pools for the 1993 tournament, Metrick (1996) finds that favorites are overbacked (i.e., more favored to win than in the Las Vegas odds) in the surveyed pools. A review of recent sports gambling trends and related issues is presented by Claussen and Miller (2001).

Regardless of the betting scenario, the goal in tournament-related wagers is to correctly forecast the result of one or more tournament games; a simple method for choosing winners is attractive, particularly to neophytes who may not be familiar with the relative strengths of the tournament participants. One such method is to choose the team with the better (i.e., numerically lower) seed as the winner for each game. A selection committee assigns these seeds, which are then used to structure the sequence and location of tournament games (NCAA, 2006). Seeding information is available to all potential bettors before making their tournament predictions.

Recent attention has been placed on the abilities of neophytes to forecast the results of sporting events. Surveys testing the recognition heuristic – wherein an individual chooses a recognized alternative in lieu of an unrecognized alternative – have shown that, under certain conditions, predictions made using this heuristic can be more accurate than those made by an individual with expert knowledge of both alternatives (Goldstein and Gigerenzer, 2002). This heuristic performed better than chance when used to predict outcomes in the 2004 European Soccer Championship and 2003 Wimbledon, though in the former case, the predictions did not perform as well as those made using other rankings; this discrepancy is attributed to the higher accuracy of predictions made by soccer rankings, not lower accuracy of recognition-based predictions (Pachur and Biele, 2007, Serwe and Frings, 2006). By evaluating the quality of predictions that are made using predictors other than recognition in the basketball games in the tournament, potential bettors will be more capable of determining the best predictor (i.e., published rankings vs. recognized teams) to use in their predictions.

HISTORY OF TOURNAMENT

The tournament was first held in 1939 with a field of eight teams (NCAA, 2007a). The number of teams participating in the tournament has increased since that time; a summary of the number of participating teams each year is given in Table 1. Since 1985, the beginning of the so-called *modern era*, the NCAA has used a sixty-four team, six-round tournament format. Though a 65th team was added in 2001 to guarantee space for thirty-four at-large teams in the tournament in addition to thirty-one conference champions, the field is reduced to sixty-four teams before the formal start of the tournament by having the two worst-ranked participants take part in a "play-in" game, where the winner moves on to the field of sixty-four.

Once the play-in game has occurred, the tournament is structured as follows: each team is seeded from one to sixteen by the selection committee and assigned to one of four tournament regions. Each seed value is assigned to four teams, and each of these teams is placed into a different region (the winner of the play-in game is given a seed of sixteen). Seed values of three or less are herein referred to as *high seeds*, while seed values of four or more are referred to as *low seeds*. Similarly, when comparing two seed values, the seed value closer to one is referred to as the *higher* seed, while the other seed is the *lower* seed. Higher seed values are assigned to better teams, as decided by the selection committee. When assigning teams to each region, efforts are made to keep higher-seeded teams close to their "natural area of interest," and to avoid having two teams from the same conference meet before the fourth round (NCAA, 2006). The deliberations of the selection committee are closed to the public; a model for predicting the teams it will select to participate in the tournament is proposed by Coleman and Lynch (2001).

In each region, first round match-ups are created by having the k^{th}-best seed in each region play the k^{th}-worst seed in the region (i.e., seed k plays seed $17 - k$, for $k = 1,2,\ldots,8$). The tournament follows a single-elimination format; the winner of each game advances to the next round, while the loser is eliminated. Subsequent rounds are structured in a similar manner. Under the assumption that the higher-seeded team wins each first round match-up (i.e., seeds one through eight advance), seed k will play seed $9 - k$ (for $k = 1,2,3,4$), in the second round. If one of the top eight seeds loses in the first round, then the winning team plays the same opponent that the top eight seed would have played (e.g., if seed twelve beats seed five and seed four beats seed thirteen in the first round, then seed twelve plays seed four in the second round); Harbaugh and Klumpp (2005) discuss the likelihood of such upsets in a two-round tournament. The tournament bracket used for each region can be found in Figure 1. Once the four regional champions have been determined, these four teams (the *Final Four*) meet in two final rounds to crown the NCAA basketball national champion.

A number of methods for predicting winners in the tournament have been proposed in the literature. Several studies use seeding alone to predict the outcome of the regional tournaments. Schwertman et al. (1991) proposes three probability models for estimating $P(j,k)$, the probability that a team with seed j defeats a team with seed k. These models provide a good fit to the observed regional tournament results from 1985 – 1990 using a chi-squared goodness of fit test. A follow-up study by Schwertman et al. (1996) proposes eight additional models that use seed as the only predictor, and extends the analysis of the models to consider winners of individual games, rather than choosing regional tournament winners only. Smith and Schwertman (1999) extend this work using regression models for predicting margin of victory for regional tournament games based on the participating teams' seeds. Boulier and Stekler (1999) use a Probit model to estimate the probability of each team winning a game based on the seeds of the two teams. Based on regional tournament results from 1985 to 1995, this study finds that seeds are a good predictor of game outcomes. Caudill (2003) uses the maximum score estimator to predict regional tournament game outcomes, and observes that this method performs slightly better than that of Boulier and Stekler (1999). Though all of these studies suggest that the seed is a good predictor of performance in the regional tournaments (i.e., the first four rounds of the championship tournaments), several issues remain unexplored. While the consensus among these studies is that games between two teams with a small difference in seed are less predictable based on seed than those with a large seed difference, it is possible that the level of predictability in these games changes with the tournament round, as the potential seed match-ups in each round can differ (e.g., whether games between an eight seed and a nine seed in the first round and games between a one seed and a two seed in the fourth round are equally predictable). Moreover, these studies do not consider games in the final two rounds of the tournament in their analyses, so the conclusions drawn can only be made for the regional tournaments, and not the final two rounds of the championship tournament.

Additional studies investigate predictors other than seeds. These predictors include won/loss records and margin of victory for season games, Vegas point spreads, and the Sagarin (2007), Massey (2000), and RPI ratings (NCAA, 2005). Carlin (1996) predicts the outcome of the 1994 tournament using Vegas point spreads from the first round games and Sagarin ratings, finding that this method performs better than the model proposed in Schwertman et al. (1991). Kaplan and Garstka (2001) consider the case of the office pool, under different point systems. A Markov probability model is proposed to determine winners based on several predictors; this model performs better than picking the seeds as the office pool point structure becomes more complex. Harville (2003) proposes a modified least squares model to predict score differences in tournament games based on score differences of games in the regular season; this model is able to correctly predict the winner of 76.3% of the games in the 1999 – 2000 postseason (which include games in the National Invitational Tournament, in addition to those in the championship tournament). Kvam and Sokol (2006) use a Markov chain/logistic regression model for predicting game winners using margin of victory from regular-season games in 1999 to 2005. This model is shown to outperform several other prediction methods when the goal is to pick the largest number of game winners. All these studies suggest that varying the number and type of predictors can lead to models with more predictive power. As the complexity of a model increases, however, the model becomes more difficult to use in practice, thereby making it less attractive to the general public, particularly when other models, such as choosing the higher seed to win each game, can be more easily implemented and are thought to perform well.

Picking the team with the better seed to win each game tends to give good picks in the early rounds of the tournament, when seed differences in a particular game are more likely to be large. In the later rounds of the tournament, however, the remaining teams tend to be those with seed three or better, resulting in smaller seed differences and, presumably, more evenly matched teams. Since 1985, 144 of 200 (72%) teams in the Elite Eight, 79 of 100 (79%) teams in the Final Four, and 22 of 25 (88%) tournament champions have had high seeds. Regardless of its seed, any team appearing in the Elite Eight has won three consecutive games, indicating that the team is capable of winning against quality competition. Therefore, as the tournament progresses, a team's seed may have less predictive value than in earlier rounds.

To investigate this hypothesis, this paper investigates the difference in performance of high-seeded teams in each round of the tournament by comparing their historical won/loss records and win percentages (herein referred to as win proportions) in the modern era. Analysis is restricted to high seeds due to the paucity of game data involving low seeds in the later rounds of the tournament. Statistical hypothesis testing suggests that these data do not provide sufficient evidence that high seeds perform differently in the fourth, fifth, and sixth rounds of the tournament; each seed's ability to win games is statistically indistinguishable from the others. Analysis of gambling point spreads indicates that, in games between high-seeded teams in the fourth, fifth, and sixth rounds of the tournament, bettors do not show significant bias (at the $\alpha = 0.05$ level) toward choosing the higher-seeded team to win in games between two high seeds, suggesting that bettors use additional predictors to place their bets.

This paper is organized as follows: the Data Sets section describes the tournament data used in the analysis. The Results section presents the results of applying statistical hypothesis tests to this data. The Implications section discusses the implications of these results. The Conclusions section summarizes key conclusions of the analysis. An appendix describing the statistical hypothesis tests used in the analysis is included.

DATA SETS

This section describes the tournament game data that are used in the analysis in this paper. Games that take place during the sixty-four team bracket modern era of the tournament are considered. The play-in game is not included, as it is often considered separate from the remaining tournament games, with the winning teams "playing into" the tournament. With twenty-five years of data and sixty-three games per year, this study collects data for a total of 63×25 = 1,575 games.

Input (predictor) and output (game result) data are gathered for each game. To analyze the performance of high seeds in each round of the tournament, seed and round data are the only predictors recorded. These predictors are readily available; the NCAA publishes each team's seed, as well as the round of each tournament game, on its tournament bracket. Output data are denoted by the winning seed of each game. Therefore, the winning seed, losing seed, and tournament round describe each game.

Description of Game Data

Let G be the set of all 1,575 tournament games in the modern era, whose results are reported by the NCAA (2007a, 2007b, 2008, 2009). A summary of the high seeds' records against different opponents is given in Table 2. Each game is labeled with the ordered three-tuple

$$g = (r,w,l) \in G, \tag{1}$$

where r is the tournament round, w is the seed of the winning team, and l is the seed of the losing team. The following three subsets of G will be used within the analysis.

A_r is the set of games involving at least one high seed in round r
D_r is identical to A_r, with games involving two identical seeds removed
B_r is the set of games involving one high seed and one low seed in round r

Using set notation, these sets are defined (for r = 1, 2, ...,6) as:

$$\begin{aligned} A_r &\equiv \{g = (r',w,l) \in G \mid r' = r, \min(w,l) \le 3\} \subseteq G, \\ D_r &\equiv \{g = (r',w,l) \in G \mid r' = r, \min(w,l) \le 3, w \ne l\} \subseteq A_r, \\ B_r &\equiv \{g = (r',w,l) \in G \mid r' = r, \min(w,l) \le 3, \max(w,l) \ge 4\} \subseteq A_r. \end{aligned} \tag{2}$$

Define the number of wins and losses for seed n, given the set of games, X, as:

$$\begin{aligned} W(n,X) &\equiv |\{g = (r,w,l) \in X \mid w = n\}|, \\ L(n,X) &\equiv |\{g = (r,w,l) \in X \mid l = n\}|, \end{aligned} \tag{3}$$

where X is one of the sets defined in (2). Similarly, define the number of times a team with seed n appears in games in X as:

$$N(n,X) = W(n,X) + L(n,X). \tag{4}$$

Note that (4) counts any game between two teams with seed n as two games. In each regional tournament (rounds one through four), only one n-seeded team can appear in each game, while in the National Semi-Final round and the National Championship game (rounds five and six), it is possible for an n-seeded team to play another n-seeded team. The observed proportion of an n-seeded team's wins, given a set of games, is:

$$P_W(n,X) = W(n,X) / N(n,X), \tag{5}$$

The proportion of wins in a set of games is used to represent a seed's performance in those games, and will be used in the ensuing analysis.

Sample Size and Independence

There are several issues that must be addressed when applying statistical hypothesis tests (as described in the appendix) to tournament results. One such issue deals with sample size. While there are 1,575 total games in G, this number shrinks dramatically when subsets are taken based on round or seed information, as with the subsets described in (2). For example, out of the 25 games in round six, only 11 involve a team with seed two. As these sample sizes decrease, the normal approximation to the binomial distribution becomes less applicable, making the results of some tests questionable at best. When such cases arise (most often in rounds five and six), more weight should be given to Fisher's exact test when drawing conclusions regarding test results, since this test does not require the underlying components of the test statistic to be normally distributed.

All of the statistical hypothesis tests used in this study require samples to be drawn randomly from their population. By treating all modern era tournament results as a sample, these results can be assumed to represent a random sample from the population of all possible tournament results. That is, it is assumed that, for each pair of seeds that play in a tournament game in a particular round, the observed frequency of each seed's wins is representative of the frequency that the seed, playing this round, will win a game against the same seed in future tournaments. It is assumed that each seed's probability of winning against another seed in a particular round (without other information about the teams) is constant. Under this assumption, a seed's probability of winning against a pool of opponent seeds in a given round is constant and can be computed using the law of total probability.

This paper compares won/loss records of high-seeded teams, either directly or using win proportions. By using won/loss data, the outcome of a single game produces two observations: a win for the winning team, and a loss for the losing team. In games between two high-seeded teams, both of these observations are counted as part of the high-seeded teams' won/loss records. These games are herein referred to as *double-counted games*. Since one team must win and the other must lose, the two observations are not independent. This fact cannot be avoided, and its implications on the statistical hypothesis tests will be discussed. By design, no double-counted games can occur in the first two rounds of the tournament, as the bracket structure prevents two high seeds from meeting before the third round. Furthermore, comparisons are only made within a given round (e.g., comparing the win proportions of seeds one and two in the fourth round), and not between rounds (e.g., comparing the win proportion of seed one in the fourth and fifth rounds). By restricting the analysis to such comparisons, concerns regarding the independence of games between rounds are avoided.

RESULTS

This section reports the results of using statistical hypothesis tests to compare the won/loss records of high-seeded teams in the modern era of the tournament. In particular, these tests investigate whether a seed's win proportion differs significantly from ½ (a toss-up is defined as a game where either participant is equally likely to win), or whether the high seeds' performances (as measured by won/loss records or win proportions) differ significantly. Note that proportions are referred to using the letter q. Results of all statistical hypothesis tests were computed with Microsoft Excel 2002 or Matlab 6.5 (Release 13) software packages.

Testing for Toss-Ups in the Performance of Individual High Seeds

To determine whether a high seed's win proportion differs significantly from ½ during a particular round of the tournament, a t-test is performed under the null hypothesis, H_0: $q = ½$, with the alternative hypothesis either that the seed wins more than half of its games (i.e., H_A: $q > ½$) or that the seed wins fewer than half of its games (i.e., H_A: $q < ½$). Both of these alternative hypotheses are tested for seeds two and three, while only the first is tested for seed one. The second alternative hypothesis is omitted for seed one, as this seed cannot play a team with a higher seed; a seed one is not expected to win substantially fewer than half of its games, while seeds two and three can play seeds higher than themselves, and therefore could win substantially fewer than half of their games. The win proportion for seed x in round r is given as $P_W(x, D_r)$; games between two identically seeded teams are not included in this computation, as including these games would drive the proportion closer to ½. Moreover, using seeds alone, all games between identically seeded teams must be toss-ups, since the two teams are considered indistinguishable. The p-values from these tests are reported in Table 3. At $\alpha = 0.05$, the designated level of significance of each test, these results indicate that the games involving seed-two teams do not differ significantly from toss-ups from the fourth round onward, and games involving seed-three teams do not differ significantly from toss-ups from the third round onward. Games involving the first seed-one teams do not differ significantly from toss-ups in the fifth round only; the *p*-value for the sixth round is borderline, falling just below the significance of the test (it is rounded up to 0.050 in Table 3). To demonstrate the interpretation of these results, consider the *p*-values for seed two. Under the alternative hypothesis, H_A: $q > ½$, the *p*-values for these *t*-tests are all below $\alpha = 0.05$ for the first three rounds of the tournament. Therefore, the data suggest that seed-two teams win substantially more than half of their games in these rounds.

The same analysis is applied to the performance of seed two in the fourth, fifth, and sixth tournament rounds. The *p*-values for the one-sample *t*-test in these rounds (0.615, 0.500, and 0.815, respectively) are all significantly greater than $\alpha = 0.05$; there is insufficient evidence to reject the null hypothesis in favor of the alternative, and the data suggest that seed-two teams do not win more than half of their games in these rounds. Under the alternative hypothesis, H_A: $q < ½$, the *p*-values for all six rounds (1.000, 1.000, 1.000, 0.385, 0.500, and 0.185, respectively) are greater than the significance of the test; there is insufficient evidence to reject the null hypothesis in favor of the alternative, and the data do not suggest that seed-two wins fewer than half of its games in any round. Conclusions regarding the other high seeds are made using the same reasoning. Note that in the modern era, teams given a seed of one have won all of their first round games, so the statistical hypothesis test could not be conducted for this case, as sample variance could not be computed. Moreover, the sample size for the analysis of seed three in the sixth round, with a won/loss record of 2-4, does not satisfy the requirements to use a normal approximation to the binomial distribution. Therefore, conclusions regarding seed three in that round should be made with caution for this and other tests, and are presented here for completeness.

In general, the *p*-values reported in Table 3 suggest that, for seeds two and three, there is a cutoff round; before this cutoff, the data suggest that the seed wins more (or fewer) than half of their games, but in the cutoff round and later rounds, there is insufficient evidence that the seed wins more (or fewer) than half of its games. For seeds two and three these cutoffs occur in the fourth and third tournament rounds, respectively. The behavior of seed one is more varied; its games do not differ significantly from toss-ups in the fifth round, but it wins more than half of its games in the fourth and sixth rounds. However, the *p*-value in the sixth round is borderline, falling just below the significance of the test, making this result less conclusive. Furthermore, if analysis for seed one in round four is restricted to games in which its opponent is a seed two or a seed three, its won/loss record is 28-25; when testing this win proportion with the hypotheses H_0: $q = ½$ and H_A: $q > ½$, the resulting *p*-value is 0.341, which suggests that the null hypothesis is not rejected in favor of the alternative hypothesis; when seed one plays another high seed, the results of these games do not differ significantly from toss-ups. These data suggest that there may be some similarity in how the high seeds perform in the fourth, fifth, and sixth rounds of the tournament. While the tests performed in this section compare their performance against a given win proportion (i.e., winning half of their games), the remaining results presented in this paper make direct comparisons between the performance of the high seeds, rather than comparing both to a given standard.

Comparing the Win Proportions and Won/Loss Records for the All High Seeds

When using multiple *t*-tests, the Bonferroni effect (Wright, 1992) increases the level of significance, and hence, may provide misleading conclusions. To circumvent this fact, two methods were used to analyze all three high seeds simultaneously. Analysis of variance (ANOVA) was used to compare the win proportions of the three seeds, while a two-way contingency table was used to test for independence between seed and win proportion.

The win proportions for the top three seeds in a particular round can be compared using ANOVA, under the null hypothesis that all three win proportions are equal and the alternative hypothesis that at least one differs from the others. ANOVA requires that the variance of win proportion be equal for all three high seeds. The Bartlett (Mendenhall and Sincich, 1995) test was used to determine whether the three population proportions have equal variance (Table 4). This test requires that data be normally distributed; all three high seeds satisfy the requirements for a normal approximation to the binomial in all rounds, with the exception of seed one in the first round and seed three in the sixth round; analysis in those rounds is provided here for completeness. The results of the Bartlett test suggest that these variances are not equal in the first two rounds, but there is insufficient evidence to reject the hypothesis of equal variances in rounds three through six, for which ANOVA was conducted.

As with the *t*-tests, the win proportion used in ANOVA for seed x in round r is given as $P_W(x, D_r)$. The *p*-values of these tests are given in the ANOVA1 column of Table 5. These *p*-values (0.000, 0.399, 0.775, and 0.142, respectively) indicate that at least one team's performance (as estimated by its win proportion) differs from the others in round three, but there is insufficient evidence to suggest that the win proportion for any of the three seeds differs from the others in the fourth, fifth, and sixth rounds. In the fifth and sixth rounds, additional tests were run, with win proportions estimated by $P_W(x,A_r)$. This estimate includes games where two teams with the same seed play in the same game, and the resulting *p*-value for each case is given in parentheses in Table 5. The *p*-values obtained by using $P_W(x, A_r)$ in rounds five and six (0.809 and 0.261, respectively) do not change the conclusions drawn, as they are both above the significance of the test. Therefore, the performances of high-seeded teams are statistically indistinguishable based on seed value in rounds four, five, and six, but statistically distinguishable (at α = 0.05) in round three.

Additional ANOVA tests were conducted using $P_W(x, B_r)$ to compare the win proportions for a high seed in each tournament round, when the opposing team has a low seed. This analysis avoids consideration of double-counted games, and therefore avoids the previously-discussed issue regarding independence of observations stemming from these games. In this case, a seed's performance is estimated by its ability to win games against low seeds. This test could not be completed for the final round of the tournament, however, as neither seed two nor seed three have played low seeds in this round in the modern era. Examining the won/loss records of games between a high-seed team and a low-seed team in rounds four and five (Table 2) shows that their sample sizes may be too small to justify a normal approximation to the binomial distribution; their analysis is provided here for completeness. The *p*-values of these tests are given in the ANOVA2 column of Table 5, and indicate that there is insufficient evidence to suggest that the win proportions for all three seeds differ after the second round, when only games against seeds four or higher are considered. High seeds are statistically indistinguishable based on their seed value when playing low seeds after the second round of play.

A two-way contingency table was also used to compare the won/loss records of the three high seeds. The null hypothesis for such a test is independence between the factors, while the alternative hypothesis is dependence between the factors. Independence between these two factors suggests that proportion of wins and losses for a seed does not differ substantially between the seeds, while dependence indicates that there are substantial differences in the win proportion. The form of a contingency table for this test is depicted in Table 6. Won/loss data for round *r* were taken from either D_r, A_r, or B_r, as for the ANOVA tests. Both chi-squared and Fisher tests were conducted on the data for each round. The *p*-values of tests on D_r and A_r (provided in parentheses for rounds five and six) are given in columns Chi-Squared1 and Fisher1 of Table 5 for the chi-squared and Fisher tests, respectively, while the results of tests on B_r are given in columns Chi-Squared2 and Fisher2. The conclusions for these tests are the same as those for ANOVA; there is insufficient evidence that the win proportions for high seeds differ after the third round when games between two distinct seeds are considered and the second round when games between a high seed and a low seed are considered. The Fisher test is given particular attention, as its usefulness isn't limited by the small sample sizes noted earlier. Therefore, high-seeded teams' win proportions are statistically indistinguishable based on seed value in the fourth, fifth and sixth rounds, but not in the first three rounds; the high seeds' abilities to win games are statistically equivalent in the fourth, fifth, and sixth rounds.

As stated earlier, the treatment of double-counted games must be dealt with to ensure independence of observations. Previously, some analysis has addressed this issue by omitting these games and considering only games in which a high seed plays a low seed. One additional method is proposed for treating double-counted games. Rather than counting both outcomes of each double-counted game, only one of these outcomes is chosen, each with probability ½; either the win or the loss is chosen randomly. Each choice is made independently. Given a set of games, X, the double-counted games it contains can be divided into six sets of indistinguishable games, each containing games where two distinct high seeds play one another:

$$\begin{aligned} D_1(X) &= \{g \in X \mid g = (r,w,l), w = 1, l = 2\}, \\ D_2(X) &= \{g \in X \mid g = (r,w,l), w = 1, l = 3\}, \\ D_3(X) &= \{g \in X \mid g = (r,w,l), w = 2, l = 1\}, \\ D_4(X) &= \{g \in X \mid g = (r,w,l), w = 2, l = 3\}, \\ D_5(X) &= \{g \in X \mid g = (r,w,l), w = 3, l = 1\}, \\ D_6(X) &= \{g \in X \mid g = (r,w,l), w = 3, l = 2\}. \end{aligned} \tag{6}$$

Let the cardinality of these sets be given by:

$$\begin{aligned} d_i(X) &\equiv |D_i(X)|, i \in \{1,2,...,6\}, \\ d(X) &\equiv \sum_{i=1}^{6} d_i(X). \end{aligned} \tag{7}$$

The probability of choosing $z_i \le d_i(X)$ observations to be counted as wins for a particular $i \in \{1,2,...,6\}$ is:

$$p_i(X, z_i) = \binom{d_i(X)}{z_i} 0.5^{z_i} 0.5^{d_i(X)-z_i} = \frac{1}{2^{d_i(X)}} \binom{d_i(X)}{z_i}. \tag{8}$$

It then follows that the probability of choosing such a z_i for every i is:

$$p(X, z_1, z_2, ..., z_6) = \prod_{i=1}^{6} p_i(X, z_i) = \frac{1}{2^{d(X)}} \prod_{i=1}^{6} \binom{d_i(X)}{z_i}. \tag{9}$$

Let the p-value of the contingency table formed by the observations in X, after choosing a z_i for every i, be $b(X|z_1,z_2,...,z_6)$, where the p-value is calculated using Fisher's exact test, given in (A11). By the law of total probability, the p-value for the table formed by observations in X is:

$$p(X) = \sum_{k_1=1}^{d_1(X)} \sum_{k_2=1}^{d_2(X)} \cdots \sum_{k_6=1}^{d_6(X)} b(X|k_1,k_2,...,k_6)p(X,k_1,k_2,...,k_6) \qquad (10)$$
$$= \sum_{k_1=1}^{d_1(X)} \sum_{k_2=1}^{d_2(X)} \cdots \sum_{k_6=1}^{d_6(X)} \frac{b(X|k_1,k_2,...,k_6)}{2^{d(X)}} \prod_{i=1}^{6} \binom{d_i(X)}{k_i}.$$

Using this approach, the *p*-values given by $p(D_r)$ for each of rounds three through six are 0.010, 0.317, 0.624, and 0.461, respectively. Therefore, the resulting conclusions are consistent with those made earlier using ANOVA, chi-squared, and Fisher tests, with the win proportions of high seeds differing in round three, but becoming statistically indistinguishable in rounds four, five, and six, implying that they are equally capable of defeating their opponents in those rounds.

IMPLICATIONS

Through statistical hypothesis testing, this paper has shown that there is insufficient evidence to suggest that the performance of a high-seeded team (as estimated by its historical won/loss record) in the Elite Eight round, the National Semi-Final round, and the National Championship game is dependent upon the team's seeding. This result implies that these seeds' abilities to defeat opponents are statistically equivalent during these rounds. Therefore, choosing the higher seed to win in this situation may not be a better method than choosing a winner randomly. This conclusion is empirically consistent with past tournament results in the modern era. For example, in the Elite Eight, the fifty-three matches pairing a seed one with either a seed two or seed three have resulted in a record or 28-25, nearly an even split. Similarly, the twenty games pairing two distinct high seeds in the Final Four have led to ten wins for the team with the higher seed. One could expect to correctly select the same number of winners by choosing the winner randomly.

If high seeds do, indeed, perform comparably in these rounds, and bettors still assume that higher seeds win games, then betting histories should reflect this bias. For the purposes of gambling, point spreads are set, and continuously adjusted, such that half of the bets are placed on the favorite covering the spread. Given the point spread for a game, the favorite is said to have *covered the spread* if it wins by a number of points greater than the point spread. If there is bias toward betting on the higher seed in games between two high seeds in the fourth, fifth, or sixth rounds of the tournament, then point spreads in those games should be more generous to the lower seed, making it more difficult for the favorite to cover the spread. There is evidence that betting on the underdog in sporting events is a favorable wagering policy (i.e., statistical hypothesis testing rejects the null hypothesis of a 50% success rate in favor of a larger success rate). For example, betting on an underdog that is currently on a one- or two-game losing streak was found to be a favorable wagering policy in National Football League (NFL) games taking place between the fall of 1985 and the fall of 1997 (Woodland and Woodland, 2000). Betting on "big underdogs" (i.e., underdogs in games with large point spreads) was found to be a favorable wagering policy for National Basketball Association (NBA) games when applied to all games between the 1995-1996 and 2001-2002 seasons (Rodney and Weinbach, 2005). A blanket rule of "bet on all underdogs" also led to significantly more than 50% success (statistically) in NFL games played during the 1998-2002 seasons, though this result does not hold when additional successful bets are required to offset transaction costs (Kochman and Goodwin, 2004).

This hypothesis was investigated using point spreads published in the Chicago Sun-Times (1985-2009) for all modern era games in the fourth, fifth, and sixth rounds of the tournament. Of the games between two distinct high seeds, the number of games when either the lower seed wins or the higher seed fails to cover the spread was counted. These counts were 30 of 53 games (57%) in the fourth round, 12 of 20 (60%) in the fifth round, and 5 of 12 (42%) in the sixth round. These proportions, q, were tested with a one-sample t-test under the null hypothesis, H_0: $q = ½$, with the alternative hypothesis, H_A: $q > ½$, to determine whether such an outcome occurred in significantly more than half of these games. The resulting p-values were 0.168, 0.186, and 0.715 for the fourth, fifth, and sixth rounds, respectively. At the $\alpha = 0.05$ level of significance, the null hypothesis is rejected in each case. These results suggest that bettors do not show significant bias toward higher-seeded teams when making wagers on games between two high seeds.

It should be noted that these results do not suggest that the results from these games are unpredictable. Indeed, several methods have been shown to effectively pick game winners using predictors other than seed (e.g., Carlin, 1996, Kaplan and Garstka, 2001, Kvam and Sokol, 2006). To a neophyte who may not be familiar with these prediction methods or their predictors, basing predictions on team recognition rather than seed could yield better results without requiring additional effort in making these predictions, as the recognition heuristic has been shown to perform well in predicting some kinds of sporting events (Serwe and Frings, 2006, Pachur and Biele, 2007).

The conclusions drawn in this paper were restricted by the relatively few years of tournament results using the 64-team bracket; as future tournaments take place, the results of those tournaments may provide evidence that seeding does affect a high seed's win proportion in the later rounds of the tournament.

CONCLUSIONS

In the modern era of the tournament, at least 70% of the teams appearing in each round following the Sweet Sixteen have been seeded three or higher. Predicting winners for games during these rounds requires a method for accurately differentiating between these high-seeded teams. One common method is to predict the team with the higher seed as the winner.

Through several types of statistical hypothesis tests, this paper has shown that, in rounds four, five, and six of the NCAA men's college basketball championship tournament, the win proportions for teams with high seeds (i.e., seed values of three or less) are statistically indistinguishable. This finding contradicts the common sentiment that higher seeds perform better than lower seeds, and indicates that the simple prediction method of favoring the higher seed to win in games between two high seeds cannot be expected to perform better than choosing the winner randomly in these rounds. To investigate this claim, analysis using point spreads was conducted. These tests do not suggest that bettors undervalue the lower seed in these games.

One question that can be posed is why the win proportions of high seeds are distinguishable in the first three rounds of the tournament, but indistinguishable in the last three rounds. One possible explanation deals with the opposing seeds that each team can face. While it is possible for a seed to play against any other seed in the fifth or six rounds, this is not true in the earlier rounds. For example, consider the difference in opponents for seeds one and two. Seed one can only play seed sixteen in the opening round, and will play either seed eight or nine in the second round, while a seed two will play seed fifteen in the opening round, and either seed seven or ten in the second round. As the rounds progress, each team can face increasingly skilled competition. Seed two, however, will tend to face this competition earlier than seed one. In the third round, a seed two can play a seed three, while, in the worst case, a seed one will face a seed four. Therefore, in the early rounds, seed two will likely face more skilled competition and, consequently, will be more likely to lose than a one seed will. As weaker teams are eliminated, however, the teams remaining in the later rounds are more likely to be more evenly matched. While a seed two may be more likely to lose in the earlier rounds of the tournament than a seed one, the analysis in this paper shows that, given that either seed makes it to the later rounds, both perform equally well once they reach such rounds.

Due to the relatively small number of tournament games using the 64-team bracket, the conclusions of this study are limited by these sample sizes. These limitations are more evident in later rounds of the tournament, with only 100, 50, and 25 games taking place in rounds four, five, and six, respectively. As this paper focuses on high-seeded teams, some of these games were not used in the analysis presented. By analyzing the win proportions and won/loss records of high seeds, any game between two such seeds will be counted twice in some of the analysis: as a win for the winning team, and as a loss for the losing team. Therefore, independence of observations was another issue that was recognized and addressed.

As future tournaments take place, additional data will be available that can be used to update this analysis, and to analyze seed performance in different ways. In this paper, seed performance was evaluated by grouping together all games involving that seed, or by dividing these games in a coarse manner (i.e., by measuring performance against a particular pool of opponent seed values). As additional samples are available, it may be possible to accurately estimate particular seed vs. seed win proportions in a particular round (e.g., the proportion of time that a seed one defeats a seed four in the Final Four), which will allow testing for additional conclusions.

It is emphasized that these results are not meant to suggest that games between two distinct high seeds in the final three rounds of the tournament are unpredictable, but rather suggest that, despite the attraction and ease of favoring the higher seed to win each game, alternative predictors should be used to select these game winners.

APPENDIX

Numerous statistical hypothesis tests can be used to draw conclusions about the parameters of a distribution, or to test for independence between factors in an experiment. This appendix discusses statistical hypothesis tests that can be used to analyze tournament data. Results for these tests are given as p-values representing the probability of observing a test statistic value that is at least as contradictory to the null hypothesis, under the assumption that the null hypothesis is true. A p-value below the significance, $\alpha = 0.05$, of the test indicates that the null hypothesis should be rejected. Unless otherwise noted, all information in this appendix is based on Mendenhall and Sincich (1995).

Student's *t*-test for Population Proportion

The t-test, based on the Student t-distribution (Student, 1908), is a small-sample alternative to the z-test for population means. This test can be applied to population proportions by treating samples as binary variables that are equal to one when a success condition is met and a zero when it is not. Relevant null and alternative hypotheses for testing a population proportion are:

$$\begin{aligned} H_0 &: q = q_0 \\ H_{A1} &: q \neq q_0 \\ H_{A2} &: q > q_0 \\ H_{A3} &: q < q_0, \end{aligned} \tag{A1}$$

where q_0 is the hypothesized proportion of successes in the population. Given a sample of n observations, the test statistic is computed as:

$$t^* = \frac{\hat{q} - q_0}{\hat{q}(1-\hat{q})/\sqrt{n}}, \tag{A2}$$

where $\hat{q}$ is the proportion of successes in the sample. Under the three alternative hypotheses in (A1), the p-value is given by:

$$\begin{aligned} p_1 &= 2P(t > |t^*|), \\ p_2 &= P(t > t^*), \\ p_3 &= P(t < t^*). \end{aligned} \tag{A3}$$

The t-distribution in (A3) has $(n - 1)$ degrees of freedom. The t-distribution assumes that samples are drawn from a normal population. Binary data for proportions are assumed to be draw randomly from a binomial population, which can be approximated by a normal population when sample size is large enough (i.e., $n\hat{q} \geq 4, n(1-\hat{q}) \geq 4$).

ANOVA for Population Means

Analysis of variance (ANOVA) compares the means of two or more populations. ANOVA is often preferred to the *t*-test when comparing many population means, since all means can be compared using one test, thereby avoiding Bonferroni effects, such as those described by Wright (1992), that arise from using many *t*-tests. As with the *t*-test, a population proportion can be represented by the population mean of a binary outcome. In ANOVA, the null and alternative hypotheses for a test comparing proportions at *k* factor levels are:

$$H_0: q_1 = q_2 = \ldots = q_k \tag{A4}$$

H_A: At least one of the *k* proportions differs from the others.

where q_i is the proportion of the i^{th} population. This study uses the ANOVA model for completely randomized designs, whose test statistic is computed using the equations:

$$\begin{aligned}
SS_{Total} &= \sum_{i=1}^{k} ns_i - \frac{\left(\sum_{i=1}^{k} ns_i\right)^2}{n} \\
SS_{Treatment} &= \sum_{i=1}^{k} \frac{ns_i^2}{n_i} - \frac{\left(\sum_{i=1}^{k} ns_i\right)^2}{n} \\
SS_{Error} &= SS_{Total} - SS_{Treatment} \\
F^* &= \frac{SS_{Treatment}/(k-1)}{SS_{Error}/(n-k)},
\end{aligned} \tag{A5}$$

where ns_i is the number of successes observed in population *i*, n_i is the number of observations taken from population *i*, and *n* is the total number of observations. Then, F^* follows an *F*-distribution with $(k - 1)$ and $(n - k)$ degrees of freedom. Under this distribution, the *p*-value of the test, using the hypotheses in (A4), is:

$$b = P(F > F^*). \tag{A6}$$

To use ANOVA, each population must have a normal distribution, and all populations must have equal variance.

Contingency Table Tests for Factor Independence

A two-factor contingency table with two factors is used to tests for independence between the two factors. This table is an n x m matrix, where element o_{ij} represents the number of observations with level i of the first factor, and level j of the second. For K total observations, let R_i ($i = 1,2,\ldots,n$) be the matrix row sums, such that R_i/K represents the marginal distribution of the first factor levels. Similarly, let C_j ($j = 1,2,\ldots m$) be the column sums, such that C_j/K represents the marginal distribution of the second factor levels. Therefore, if the two factors are independent, then $e_{ij} = (R_iC_j)/K$ is the expected number of observations with level i of the first factor, and level j of the second. If observations are made independently, then:

$$X^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(e_{ij} - o_{ij}\right)^2}{e_{ij}} \tag{A7}$$

has a chi-squared distribution with $(n - 1)(m - 1)$ degrees of freedom, and the p-value is $P(\chi^2 > X^2)$, under the null hypothesis of factor independence. This test is typically used when a large number of samples are available.

Fisher's exact test is a small-sample alternative to the chi-squared test (Mehta and Patel, 1983). This test treats the observations in the contingency table as a hypergeometric distribution, where the row and column sums are identical to those of the observed table. Thus, the set of alternative tables that could be observed are:

$$Y_T = \left\{ \{y_{ij}\}_{i\in\{1,2,\ldots,n\},\, j\in\{1,2,\ldots,m\}} \left| \sum_{j=1}^{m} y_{ij} = R_i\,(i = 1,2,\ldots,n), \sum_{i=1}^{n} y_{ij} = C_j\,(j = 1,2,\ldots,m) \right. \right\}, \tag{A8}$$

where R_i and C_j are the row and column sums of the observed table. Under the hypergeometric distribution with independent observations, the probability of a particular table, T, is

$$P(T) = \frac{\left(\prod_{i=1}^{n} R_i!\right)\left(\prod_{j=1}^{m} C_j!\right)}{K!\prod_{i=1}^{n}\prod_{j=1}^{m} t_{ij}}. \tag{A9}$$

For any table T, let X be the set of tables that are less likely to be observed, where:

$$X = \{y \in Y_T \mid P(y) \leq P(T)\}. \tag{A10}$$

The *p*-value for the test, given table *T*, is:

$$p = \sum_{x \in X} P(x). \tag{A11}$$

REFERENCES

B L Boulier and H O Stekler 'Are Sports Seedings Good Predictors?: An Evaluation' *International Journal of Forecasting* (1999) 15 83-91.

B P Carlin 'Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information' *The American Statistician* (1996) 50(1) 39-43.

S B Caudill 'Predicting Discrete Outcomes with the Maximum Score Estimator: The Case of the NCAA Men's Basketball Tournament' *International Journal of Forecasting* (2003) 19 313-317.

Chicago Sun-Times (1985-2009). NCAA Tournament Point Spread Data. Various dates.

C L Claussen, and L K Miller 'The Gambling Industry and Sports Gambling: A Stake in the Game?' *Journal of Sport Management* (2001) 15 350-363.

J Coleman and A K Lynch 'Identifying the NCAA Tournament "Dance Card"'.*Interfaces* (2001) 31(3) 76-86.

D G Goldstein and G Gigerenzer 'Models of Ecological Rationality: The Recognition Heuristic' *Psychological Review* (2002) 109(1) 75-90.

R Harbaugh and T Klumpp 'Early Round Upsets and Championship Blowouts' *Economic Inquiry* (2005) 43(2) 316-329.

D A Harville 'The Selection or Seeding of College Basketball or Football Teams for Postseason Competition' *Journal of the American Statistical Association* (2003) 98(461) 17-27.

E H Kaplan and S J Garstka 'March Madness and the Office Pool' *Management Science* (2001) 47(3) 369-382.

L Kochman and R Goodwin 'Underdog's are Man's Best Friend: A Test of Football Market Efficiency' *Journal of Sports Economics* (2004) 5(4) 387-391.

P Kvam and J S Sokol 'A Logistic Regression/Markov Chain Model for NCAA Basketball' *Naval Research Logistics* (2006) 53(8) 788-803.

K Massey 'Massey Ratings Description' (2000) http://www.masseyratings.com/theory/massey.htm [July 22nd 2007]

M McCarthy 'Gambling Madness Can Snag Court Fans' (2007) http://www.usatoday.com/sports/2007-03-27-gambling-cover_n.htm [June 21st 2007]

C R Mehta and N R Patel 'A Network Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables' *Journal of the American Statistical Association* (1983) 78(382) 427-434.

W Mendenhall and T Sincich. *Statistics for Engineering and the Sciences*. (Upper Saddle River, NJ, Prentice Hall, 1995)

A Metrick 'March Madness? Strategic Behavior in NCAA Basketball Tournament Betting Pools' *Journal of Economic Behavior & Organization* (1995) 30 159-172.

NCAA 'Men's Basketball Committee Approves RPI Modifications, January 3, 2005' (2005) http://www.ncaa.org/ [August 6th 2007]

NCAA 'Principles and Procedures for Establishing the Men's Bracket, Jan. 23, 2006' (2006) http://www.ncaasports.com/basketball/mens/story/9183455 [June 21st 2007]

NCAA *Official 2007 NCAA Men's Final Four Records Book* (Indianapolis, IN, The National Collegiate Athletic Association, 2007a).

NCAA 'Scoreboard' (2007b) http://www.ncaasports.com/basketball/mens/scoreboard [August 2nd 2007]

NCAA 'Scoreboard' (2008) http://www.ncaa.com/basketball-mens/default.aspx?id=114&date=20080407 [April 8th 2008]

NCAA 'Scoreboard' (2009) http://www.ncaa.com/teams/m-baskbl/scoreboards/top25/top25-scoreboard-day-main.html [April 22nd 2009]

T Pachur and G Biele 'Forecasting From Ignorance: The Use and Usefulness of Recognition in Lay Predictions of Sports Events' *Acta Psychologica* (2007) 125 99-116.

J P Rodney and A P Weinbach 'Bettor Misperceptions in the NBA: The Overbetting of Large Favorites and the "Hot Hand"' *Journal of Sports Economics* (2005) 6(4) 390-400.

J Sagarin 'Jeff Sagarin Computer Ratings' (2007) http://www.usatoday.com/sports/sagarin.htm [July 22nd 2007]

N C Schwertman, T A McCready and L Howard 'Probability Models for the NCAA Regional Basketball Tournaments' *The American Statistician* (1991) 45(1) 35-38.

N C Schwertman, K L Schenk and B C Holbrook 'More Probability Models for the NCAA Regional Basketball Tournaments' *The American Statistician* (1996) 50(1), 3438.

S Serwe and C Frings 'Who Will Win Wimbledon? The Recognition Heuristic in Predicting Sports Events' *Journal of Behavioral Decision Making* (2006) 19 321-332.

T Smith and N C Schwertman 'Can the NCAA Basketball Tournament Seeding Be Used to Predict Margin of Victory?' *The American Statistician* (1999) 53(2) 94-98.

Student 'The Probable Error of a Mean' *Biometrika* (1908) 6(1) 1-25.

B M Woodland and L M Woodland 'Testing Contrarian Strategies in the National Football League' *Journal of Sports Economics* (2000) 1(2) 187-193.

S P Wright 'Adjusted P-Values for Simultaneous Inference' *Biometrics* (1992) 48(4) 1005-1013.

Figure 1: Tournament bracket used for each region

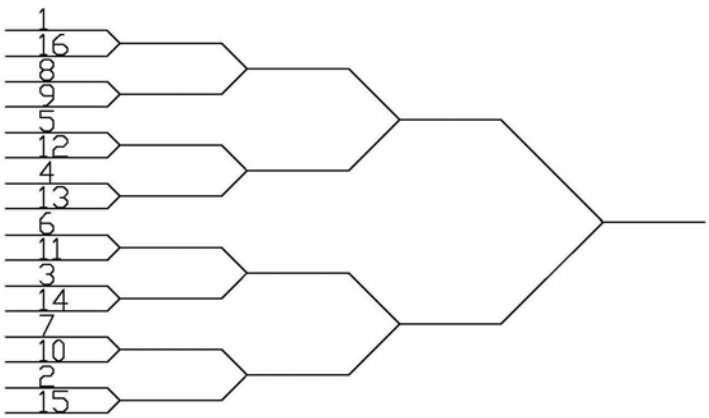

Table 1: Number of participating teams

Years	# of Teams
1930-1950	8
1951-1952	16
1953-1974	22-25
1964-1978	32
1979	40
1980-1984	48
1985-2009	64-65

Table 2: Summary of game records for high seeds (1, 2, and 3)

		Record in games against			
Round	**Seed**	**Seed 1**	**Seed 2**	**Seed 3**	**Seed 4-16**
1	1	*	*	*	100-0
	2	*	*	*	96-4
	3	*	*	*	85-15
2	1	*	*	*	88-12
	2	*	*	*	64-32
	3	*	*	*	52-33
3	1	*	*	*	73-15
	2	*	*	21-12	25-6
	3	*	12-21	*	13-6
4	1	*	18-17	10-8	16-4
	2	17-18	*	*	5-6
	3	8-10	*	*	5-2
5	1	11-11	5-4	2-4	6-1
	2	4-5	1-1	3-2	3-3
	3	4-2	2-3	NA	2-0
6	1	5-5	5-1	1-0	4-3
	2	1-5	NA	3-2	NA
	3	0-1	2-3	1-1	NA

* – match-up is impossible in this round; NA – match-up is possible, but has not occurred in the modern era

Table 3: Results from testing H_0: $q = ½$ for each high seed against the specified alternative hypotheses, given as p-values (q is the seed's win proportion)

	Seed 1	**Seed 2**		**Seed 3**	
Round	**H_A: q > 0.5**	**H_A: q > 0.5**	**H_A: q < 0.5**	**H_A: q > 0.5**	**H_A: q < 0.5**
1	*	0.000	1.000	0.000	1.000
2	0.000	0.000	1.000	0.019	0.981
3	0.000	0.000	1.000	0.609	0.391
4	0.039	0.615	0.385	0.422	0.578
5	0.198	0.500	0.500	0.205	0.795
6	0.050	0.815	0.185	0.787	0.213

Table 4: Results from testing for equal variance of the win proportion, q, for all high seeds, given as p-values

Round	**Bartlett's Test**
1	*
2	0.000
3	0.061
4	0.985
5	0.994 (0.004)
6	0.978 (1.000)

Table 5: Results from comparing performance of all three high seeds in each round, given as *p*-values

Round	ANOVA 1	ANOVA 2	Chi-Squared1	Chi-Squared2	Fisher1	Fisher2
1	*	*	0.000	0.000	0.000	0.000
2	*	*	0.000	0.000	0.000	0.000
3	0.000	0.355	0.000	0.350	0.000	0.374
4	0.399	0.146	0.393	0.138	0.389	0.152
5	0.775(0.809)	0.269	0.764(0.803)	0.229	0.830(0.823)	0.277
6	0.142(0.261)	*	0.133(0.247)	*	0.136(0.276)	*

Table 6: Format of a two-way contingency table for a set of games, X

	Seed 1	Seed 2	Seed 3
Wins	$W(1,X)$	$W(2,X)$	$W(3,X)$
Losses	$L(1,X)$	$L(2,X)$	$L(3,X)$