ISSN 1750-9823 (print)
International Journal of Sports Science and Engineering
Vol. 04 (2010) No. 04, pp. 239-245

A Simple and Effective Method to Predict Seeded Tournament Outcomes

S. J. Robinson

Department of Chemistry & Physics, Belmont University, Nashville, TN 37212

(*Received October 26, 2010, accepted November 19, 2010*)

Abstract. Predicting the outcomes of sporting events has always been an attractive yet elusive endeavor. Much work has been done in previous decades to improve models that depend on specific team strengths, such as point-spread data. Others work on the premise of a lack of knowledge about which teams are playing in the tournament, and a popular method within this latter category relies on seed numbers. In order to capture relevant historical data, the model presented here builds on seed number models by including the overall winning percentage of those seeds. This best-fit approach 1) provides reasonably good matches with empirical probabilities associated with individual games and overall tournament results and 2) gives insight into unexpected results and likely future behavior.

Keywords: basketball, tournament, prediction

1. Introduction

The possible outcomes of a college or professional sports matchup are universally and invariably discussed much more than the actual outcome; after all, the uncertainty itself is a major cause of a fan's interest. But uncertainty and curiosity also have large economic impacts in both direct forms (e.g., ticket sales, gambling) and indirect forms (e.g., advertising, merchandise). Thus, numerous studies have been done over the last few decades to better predict the outcomes of games and tournaments. See, for example, Mosteller [1], Stern [2], or Clair [3]. In particular, the National Collegiate Athletic Association (NCAA) men's basketball tournament draws special attention each year. Each fan's completed pre-tournament bracket is an attempt to forecast the winner, and several authors have worked to improve forecasting abilities in this tournament with the aid of computer models, including Schwertman, McCready and Howard [4], Carlin [5], and Niemi, Carlin, and Alexander [6].

This being said, clearly the best way to determine the winner of the NCAA tournament is to examine teams on a case-by-case basis and look at factors such as record, experience, talent, injuries, individual matchups, nearby fan base, etc., and this can all be done very effectively by humans (e.g., Las Vegas has been doing this quite well for decades). But if one wishes to take a probabilistic look at future outcomes of the NCAA tournament (e.g., could a 12-seed possibly win the tournament in the next 50 years?), team information is, of course, unavailable. This leaves only one predictor of the behavior of tournaments in which the teams aren't known yet: seeding (which includes the seeds themselves and the probable behavior of those seeds). Several good papers have been written on the subject of tournament seeding, notably Boulier and Stekler [7], Smith and Schwertman [8], and Jacobson and King [9]. Schwertman, Schenk, and Holbrook [10] detailed eleven different models and their results using seeds and functions of seeds. The model presented herein builds on these types of models by inserting an additional component of historical behavior into the prediction of both individual games and overall tournament results. It will be shown that a simple and intuitive formula can be quite effective in modeling tournament outcomes, with effectiveness determined as the model's ability to blindly "predict" the past (i.e, without knowing the teams involved).

2. Tournament Format

The NCAA tournament is comprised of a field of 68 teams, which, after four play-in games, is reduced to 64 (2^6) teams seeded in four regions from 1 to 16; this has been the format since 1985. Teams play each other in a single-elimination format according to their regional seeds (see Figure 1). 31 conference champions receive automatic bids to the tournament, while a committee appoints the remaining teams and

seeds all teams. The best teams, as judged by the committee, receive higher seeds (1, 2, 3,…). A team must win its first four games to reach the semifinals (i.e., the "Final Four") while two more wins yield a championship. The committee has recently had good success in seeding; that is, the teams that were "supposed" to win did. For example, in 2008, the four 1-seeds reached the Final Four, and in 2009, 15 of the remaining 16 teams after the second round were top-four seeds.

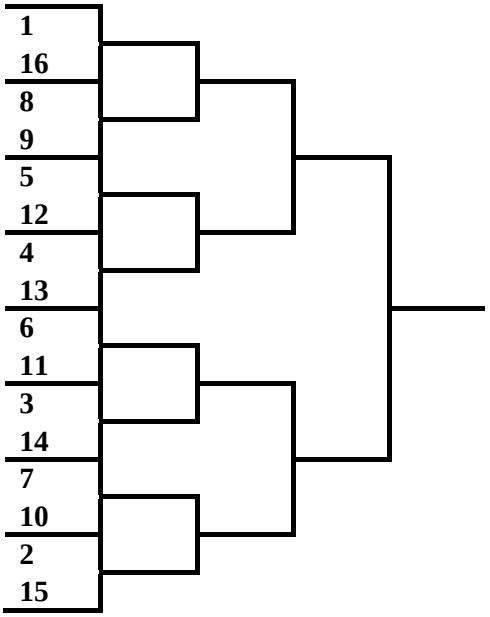

Fig. 1: Regional bracket. The winner of each region plays in a four-team final to determine the champion.

Table 1. Actual overall winning percentages by seed since 1985.

seed	W	L	pct.
1	356	88	.802
2	253	100	.717
3	191	101	.654
4	152	103	.596
5	123	104	.542
6	127	103	.552
7	86	104	.453
8	68	103	.398
9	61	104	.370
10	68	104	.395
11	51	104	.329
12	54	104	.342
13	26	104	.200
14	18	104	.148
15	4	104	.037
16	0	104	.000

As seen in Table 1, the highest seeds, as expected, have performed much better overall than the lowest seeds. There are a few exceptions, notably in the poorer-than-expected play of 5-seeds and the better-than-expected play of 10- and 12-seeds. Many of those filling out brackets pick the 5-12 matchups as probable candidates for upsets, and for good reason: the 12-seed has won this matchup 35 of 104 times through 2010, or an average of 1.35 of 4 times per tournament. Keeping in mind that the tournament committee believes that 5-seeds are the 17–20th best teams in the nation, while the 12-seeds are the 45–48th best teams, the number of 5-12 upsets is remarkable. One explanation is that mediocre teams are often undeservedly given a 5-seed because they play in a well-known conference. Another is that this matchup, compared to the 3-14 and 4-13 first-round games, is at the point where the increased effort accompanying the "underdog mentality" is enough to overcome the difference in talent between teams. Regardless of the reason, as subjective as it may be, such consistent behavior should not be ignored when predicting future behavior.

Stern [2] effectively used a margin-of-victory normal distribution to calculate long-term results, but this has its limitations, especially in the NCAA tournament. Many of the first-round games have a clear winner by the end of the first half of play, and higher-seeded teams let their starters rest. This naturally deflates margins of victory and can skew probabilities. Thus, using the data from Table 1 gives two advantages. First, it objectively views wins and losses apart from in-game data. Second, it provides enough data for statistical significance. Using the history of individual seed matchups would obviously give perfect agreement with the actual data, but because of the enormous number of possibilities, 26 years (and perhaps 1000 years) is not enough time to establish a statistically significant history for most matchups. For example, two 4-seeds have never played each other and the 3-5 matchup has only been played twice. Clearly, those individual matchup histories should have no bearing on predicting such games which might occur in the future.

In addition to historical data, one may consider the seed numbers themselves. Certainly, there is no way to quantify the desire, confidence, or basketball maturity of any team. However, if one trusts the ability of the experts on the selection committee, the subjective elements of a team are taken into account during seeding. Thus, the seeds themselves may give us more information than is readily available on the surface.

In summary, winning percentage provides a statistically significant, empirical, and objective set of data while seed number provides a subjective yet quantitative predictor of game outcomes. The combination of the two should yield a simple and effective way to predict a tournament in which the participants are unknown.

3. A Simple Probabilistic Model

There are three main criteria that any such model should meet. First, it should make sense. That is, if $p_{i,j}$ represents the probability of seed i beating seed j, the following criteria must be met:

$$0 < p_{i,j} < 1 \tag{1}$$

$$p_{i,j} = 1 - p_{j,i} \tag{2}$$

$$p_{i,i} = 0.5 \tag{3}$$

We may also add the non-binding general principle of

$$p_{i,j} < p_{i,j+1} \tag{4}$$

Second, the model should be able to predict individual matchups within reason. If the model predicts that a 1-seed should beat a 4-seed 73% of the time, the history between those teams over a statistically significant time period should fit that prediction relatively well. Finally, the model should be able to predict the probability of each seed winning a championship. If the model works for individual games and is run enough times, clear patterns for differently seeded champions should emerge, and those should also match historical results.

Consider a simple example of ratios from the natural world. If a total voltage V is applied across two resistors R and R' in series, the voltage drop across resistor R will be in proportion to the total resistance of the series resistors:

$$\frac{V_R}{V} = \frac{R}{R + R'} = \frac{1}{1 + R'/R} \tag{5}$$

The probabilistic model discussed here follows this simple pattern in that the probability should fall more heavily on the higher seeded teams and those teams that have historically done well. With S as seed number, W as historic winning percentage, H representing the higher seed, and L representing the lower seed, one can create a general formula and rely on regression to determine the best-fitting parameters. Here, two models are presented on the principle of ratios. Both give the probability of a higher seed beating a lower seed:

$$p_1 = A + \sum_{i=1}^{5} \left[\frac{B_i}{1 + (S_H/S_L)^i} + \frac{C_i}{1 + (W_L/W_H)^i} \right] \tag{6}$$

$$p_2 = a + \frac{b}{1 + (S_H/S_L)^c} + \frac{d}{1 + (W_L/W_H)^f} \tag{7}$$

Determining the best fit was a two-step process. First, using game-by-game results, where at least four games have been played for a certain matchup, χ^2-reduction was used to find the optimal range for these parameters. Then the parameters were finely tuned to best match overall tournament results by reducing the average difference between predicted and empirical probabilities. For p_1, this yield nonzero coefficients of

$B_1 = 0.635$, $C_1 = 0.174$, and $C_2 = 0.190$; for p_2, $b = 0.617$, $c = 0.964$, $d = 0.383$, and $f = 1.82$. Both give quite similar per-game and championship probabilities; thus, for brevity, only the latter will be discussed in the remainder of the article. It will be shown shortly that this model, involving only two terms, is quite effective in predicting tournament outcomes. Table 2 shows $p_{i,j}$ for individual matchups based on data through 2010. For example, there is a 62.9% chance that a 3-seed will beat a 6-seed and an 82.5% chance that a 2-seed will beat an 11-seed.

Table 2. Predicted probabilities of a higher-seeded team beating a lower-seed team for the NCAA men's basketball tournament.

	1															
1	.500	**2**														
2	.619	.500	**3**													
3	.685	.575	.500	**4**												
4	.731	.631	.559	.500	**5**											
5	.766	.676	.607	.550	.500	**6**										
6	.778	.694	.629	.573	.524	.500	**7**									
7	.818	.742	.681	.628	.581	.557	.500	**8**								
8	.843	.774	.717	.667	.621	.598	.542	.500	**9**							
9	.859	.795	.741	.693	.649	.627	.572	.530	.500	**10**						
10	.857	.795	.743	.696	.653	.631	.576	.534	.504	.500	**11**					
11	.881	.825	.778	.734	.693	.672	.620	.580	.550	.546	.500	**12**				
12	.882	.828	.782	.739	.699	.678	.626	.586	.556	.552	.506	.500	**13**			
13	.924	.879	.840	.804	.771	.749	.710	.677	.651	.644	.606	.599	.500	**14**		
14	.938	.898	.862	.830	.800	.779	.747	.719	.696	.687	.655	.646	.563	.500	**15**	
15	.956	.921	.890	.863	.838	.817	.796	.777	.760	.746	.730	.718	.696	.673	.500	**16**
16	.960	.927	.898	.872	.848	.827	.808	.791	.775	.760	.747	.734	.722	.711	.701	.500

3.1. Results

There are a few important points to make regarding these predictions. First, note that every number in the table is at least 50% and same-seeded matchups are exactly 50%. This is consistent with the sensible idea that higher-seeded teams *should* beat lower-seeded teams. Second, the probabilities do not distinguish between matchups with a many-game history and those with few-to-no actual games played. As mentioned above, this is important in blind prediction. Finally, as an example, although a 16-seed has never beaten a 1-seed in 100 games played, the probability is certainly not zero, and the results confirm that.

Actual vs. predicted data for all matchups played is shown in Figure 2. A linear regression is applied to the data, which shows a relationship of $y = 1.002x$, while a perfect fit would yield $y = x$. There are a few anomalies, but as can be seen, the model works well. Still, while individual matchups are interesting, how well a seed does throughout the tournament is more intriguing.

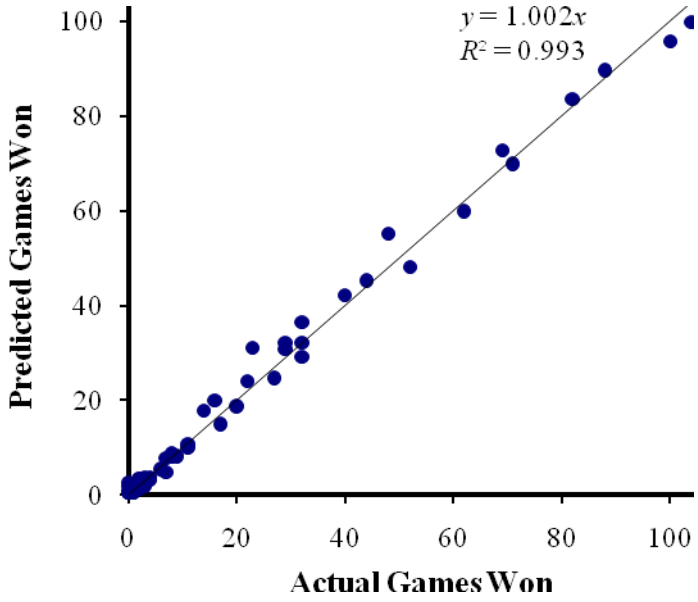

Fig. 2. Predicted vs. actual games won. The line and accompanying equation represent a linear regression. A perfect fit would yield $y = x$.

To predict the probabilities of a seed's advancement, the probabilities of a given seed winning in a given round were calculated using the formulation of Searls [11]. That is, the probabilities of a seed being in a round, its opponent being in that round, and the higher seed winning that matchup are all taken into account. This, along with a comparison with historical results, is shown in Table 3, where P_4 represents the probability of a seed winning its region (winning four games) and P_6 the probability of winning the championship (winning six games). Note that these probabilities are not directly comparable. For example, we see that the probability of a 1-seed winning the tournament is greater than the probability of a 1-seed winning its region. This strange result is simply because there can be more than one 1-seed in the Final Four (there have been an average of 1.8 per year), but only one per region.

Table 3. The probability of a seed winning in each round. The last four columns represent the predicted and actual numbers of regional and overall championships for each seed through 2010.

Seed	Rd 1	Rd 2	Rd 3	Rd 4	Rd 5	Rd 6	$104P_4$	actual	$26P_6$	actual
1	9.60E-1	8.17E-1	6.28E-1	4.31E-1	5.27E-1	6.15E-1	44.9	46	16.0	16
2	9.21E-1	7.04E-1	4.54E-1	2.25E-1	2.24E-1	2.11E-1	23.4	22	5.49	4
3	8.62E-1	5.84E-1	2.98E-1	1.28E-1	1.10E-1	8.84E-2	13.3	13	2.30	3
4	8.04E-1	4.88E-1	1.69E-1	7.68E-2	5.82E-2	4.08E-2	7.99	9	1.06	1
5	6.99E-1	3.59E-1	1.11E-1	4.53E-2	3.05E-2	1.89E-2	4.71	6	0.49	0
6	6.72E-1	2.87E-1	1.13E-1	3.64E-2	2.33E-2	1.37E-2	3.79	3	0.36	1
7	5.76E-1	1.73E-1	6.88E-2	1.91E-2	1.04E-2	5.17E-3	1.98	0	0.13	0
8	5.30E-1	9.65E-2	3.79E-2	1.15E-2	5.54E-3	2.42E-3	1.19	3	0.06	1
9	4.70E-1	7.83E-2	2.87E-2	7.98E-3	3.53E-3	1.41E-3	0.83	0	0.04	0
10	4.24E-1	1.05E-1	3.47E-2	7.88E-3	3.48E-3	1.39E-3	0.82	0	0.04	0
11	3.28E-1	9.25E-2	2.31E-2	4.53E-3	1.72E-3	5.87E-4	0.47	2	0.02	0
12	3.01E-1	9.85E-2	1.75E-2	4.16E-3	1.56E-3	5.26E-4	0.43	0	0.01	0
13	1.96E-1	5.51E-2	6.99E-3	1.24E-3	3.35E-4	7.91E-5	0.13	0	< 0.01	0
14	1.38E-1	3.60E-2	5.79E-3	6.90E-4	1.58E-4	3.14E-5	0.07	0	< 0.01	0
15	7.91E-2	1.78E-2	2.73E-3	2.51E-4	4.46E-5	6.79E-6	0.03	0	< 0.01	0
16	3.98E-2	8.63E-3	1.37E-3	1.55E-4	2.55E-5	3.60E-6	0.02	0	< 0.01	0

There have been 104 regional winners since 1985 and 26 overall champions, so the probabilities were multiplied by those numbers and compared to historical results. This comparison allows us to see which seeds have unexpectedly won regions or the overall tournament. The predictions match the data surprisingly

well, so there are few anomalies, exceptions being that the 7-seed seed has never won its region, the 8- and 11-seeds have 3 and 2 regional championships respectively, and the 6- and 8-seeds have each won championships (Kansas in 1988 and Villanova in 1985 respectively). When the probabilities are plotted vs. seed in Figure 3, lower seeds continually have smaller probabilities than the next-highest seed except for the 10- and 12-seeds, giving evidence that those teams, while not having won a championship, have continually done better than they "should." The greater downward slope seen for the 13–16 seeds indicates relatively poorer performances in past tournaments. This is no surprise, given the fact that these teams generally represent the champions of lesser-known conferences that are untested against the best teams in the country.

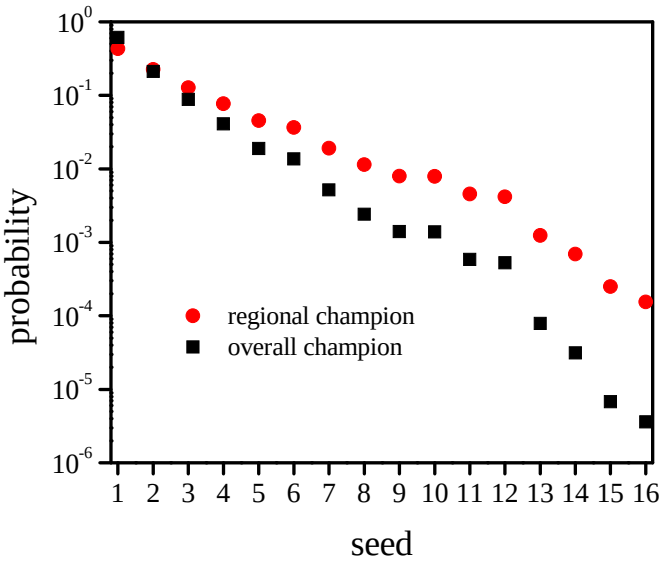

Fig. 3. The probabilities of a seed winning its region and being the overall champion.

3.2. Conclusion

Human behavior, including the game of basketball, is an inexact science, but computational modeling is being done effectively even on a social level. The best way to do this is by taking into account both the history and expectations of an event, which is done for the model presented here. Meanwhile, the simplicity of the model has two main practical benefits. First, if large numbers of simulations are done (perhaps by implementing the model on a website with many users or as done with this model to confirm probabilities), a simple model is less resource-consuming (i.e., computing cycles, physical memory, time, etc.) than a complex one. Second, simplicity would benefit any instructor who wished to teach modeling and/or probability to his/her students using popular examples such as the one contained herein. Even with the prospect of the NCAA tournament expanding to a greater number of teams, the model is flexible enough to easily adapt. In any case, it has been shown that a straightforward probabilistic analysis of a complex problem can be quite powerful and useful if applied to the right situations.

4. References

[1] F. Mosteller. The World Series Competition. *Journal of the American Statistical Association*. 1952, **47** (259): 355-380.

[2] H.S. Stern. On the Probability of Winning a Football Game. *The American Statistician*. 1991, **45** (3): 179-183.

[3] B. Clair and D. Letscher. Optimal Strategies for Sports Betting Pools. *Operations Research*. 2007, **55** (3): 1163-1177.

[4] N.C. Schwertman, T.A. McCready, and L. Howard. Probability models for the NCAA regional basketball tournament. *The American Statistician*. 1991, **45** (1)**:** 35–38.

[5] B.P. Carlin. Improved NCAA basketball tournament modeling via point spread and team strength information. *The American Statistician*. 1996, **50** (1): 39-43.

[6] J.B. Niemi, B.P. Carlin, and J.M. Alexander. Contrarian Strategies for NCAA Tournament Pools: A Cure for March Madness? *Chance*. 2008, **21**(1): 35-42.

[7] B.L. Boulier and H.O. Stekler. Are sports seedings good predictors?: An evaluation. *International Journal of*

Forecasting. 1999, **15**(1): 83-91.

[8] 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.

[9] S.H. Jacobson and D.M. King. Seeding in the NCAA Men's Basketball Tournament: When is a Higher Seed Better? *The Journal of Gambling Business and Economics.* 2009, **3** (2): 63-87.

[10] 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): 34-38.

[11] D.T. Searls. On the Probability of Winning with Different Tournament Procedures. *American Statistical Association Journal.* 1963, **58** (304): 1064-1081.