



## Seed distributions for the NCAA men's basketball tournament

Sheldon H. Jacobson<sup>a,\*</sup>, Alexander G. Nikolaev<sup>1,b</sup>, Douglas M. King<sup>2,a</sup>, Adrian J. Lee<sup>3,c</sup>

<sup>a</sup> University of Illinois, Urbana, IL, USA

<sup>b</sup> University of Buffalo (SUNY), Buffalo, NY, USA

<sup>c</sup> Central Illinois Technology and Education Research Institute, Springfield, IL, USA

### ARTICLE INFO

#### Article history:

Received 14 August 2010

Accepted 14 February 2011

Processed by B. Lev

#### Keywords:

NCAA basketball  
March Madness  
Bracketology  
Tournament seeds  
Geometric distribution

### ABSTRACT

*Bracketology*, the art of successfully picking all the winners in the National Collegiate Athletic Association's (NCAA) annual men's Division I college basketball championship tournament, has become a favorite national activity. In spite of the challenges and uncertainty faced in this endeavor, patterns exist in how the seeds appear in each round, particularly the later rounds. This paper statistically analyzes tournaments from 1985 to 2010, finding that the distribution of seeds that win in the rounds beyond the Sweet Sixteen can be modeled as a truncated geometric random variable. This model allows one to consider any set of seeds in each tournament round and compute the probability that these seeds would win in that round; this methodology can evaluate the likelihood of each seed combination in each tournament round, based on past tournament history. Finally, each tournament from 1985 through 2010 is analyzed using this model to assess its likelihood and measure the probability of its occurrence.

© 2011 Elsevier Ltd. All rights reserved.

### 1. Introduction and background

The National Collegiate Athletic Association (NCAA) Division I men's basketball tournament draws an enormous amount of national media and fanatical interest. This single-elimination tournament, commonly referred to as *March Madness*, brings together teams to compete for the national championship; some teams receive automatic bids, while others are selected "at-large" by a selection committee. Given the popularity of sports betting, this tournament has been estimated to draw billions of dollars of illegal wagers on the outcomes of the games [1].

A popular activity is to create a *bracket*, where an individual predicts the outcome of every tournament game, including future games whose participants have not yet been decided. While the  $10^{20}$  possible brackets make it highly unlikely that an individual will successfully predict every game (i.e., a perfect bracket), competitions abound on numerous web sites (such as [www.cbssportsline.com](http://www.cbssportsline.com) and [www.espn.com](http://www.espn.com)) where contestants compete to come the closest. Furthermore, academic researchers have

studied different ways to assign teams to regions, assess the field, and pick winners [e.g., 2–5].

Jacobson and King [6] provide a brief history of the tournament and its evolution into the current format and structure. The first tournament was held in 1939, with eight teams participating [7]. Since that time, there have been several occasions when the number of teams participating has been increased; see Table 1 in Ref. [6] for specific details. The *modern era* for the tournament began in 1985, when the format of 64 teams playing in six rounds was introduced. In 2001, a 65th team was added to allow space for 34 at-large teams in addition to the 31 automatic bid conference champions, with the field reduced to 64 teams by having the two weakest teams in the field play prior to the formal start of the tournament (termed the *play-in game*). Beginning with the 2011 tournament, the NCAA expanded the field to 68 teams, with the field being reduced to 64 teams through four such play-in games (coined the *first four*). Once the field of 64 teams is determined, the tournament is structured into four *regions* (designated as East, Midwest, South, and West), with 16 teams assigned to each region. The tournament lasts six rounds of single-elimination games; the losing team in each game is eliminated from the tournament, while the winner advances to the next round. Since half of the remaining teams are eliminated from the tournament in each round, the teams appearing in the third round are commonly called the *Sweet Sixteen*, while those appearing in the fourth and fifth rounds are the *Elite Eight* and *Final Four*, respectively. In the first four rounds, teams only play other teams in the same region to determine four regional

\* Correspondence to: University of Illinois at Urbana-Champaign, Department of Computer Science, Urbana, IL 61801, USA. Tel.: +1 217 244 7275.

E-mail addresses: [shj@illinois.edu](mailto:shj@illinois.edu) (S.H. Jacobson), [anikolaev@buffalo.edu](mailto:anikolaev@buffalo.edu) (A.G. Nikolaev), [dming@illinois.edu](mailto:dming@illinois.edu) (D.M. King), [ajlee@citeri.org](mailto:ajlee@citeri.org) (A.J. Lee).

<sup>1</sup> Tel.: +716 645 4710.

<sup>2</sup> Tel.: +217 333 2731.

<sup>3</sup> Tel.: +217 553 0511.

champions (i.e., the Final Four). These regional champions compete in the final two rounds to crown a National Champion.

In each region, the selection committee seeds the 16 teams from one (the highest seeded team in a region) to sixteen (the lowest seeded team in a region). Therefore, there are a total of four teams seeded No. 1, four teams seeded No. 2, and so forth, with four teams seeded No. 16. These seeds also determine the structure of each region's bracket; for example, seed No.  $n$  plays seed No.  $17 - n$ ,  $n = 1, 2, \dots, 8$ , in the first round of the tournament. Assuming that the selection committee has done a good job in seeding the teams, then each team's seed provides a simple metric for comparing the relative strength of teams, and hence, provides a quantitative measure to predict which team will win. It is reasonable to expect that the performance of each seed is monotonically non-increasing with the seed value, where performance is measured as the probability of seeds winning in each round. However, given the uncertainty inherent in any given game, upsets occur with great regularity [8]. For example, in 12 of the 26 tournaments from 1985 to 2010, two or more teams seeded No. 11 or lower (worse) have reached the Sweet Sixteen (i.e., the final 16 teams remaining in the tournament). Note that for some seed combinations, actual performance in the tournament violates this monotonicity. For example, No. 9 seeds have a winning record over No. 8 seeds in the first round (56-48), although this difference is not statistically significant at the  $\alpha = .05$  level.

The distribution of seeds in each round provides one way to assess the frequency of upsets in a given year. For example, if the teams seeded one through four win both their games in the first two rounds, then the Sweet Sixteen would be made up of the top sixteen seeded teams (i.e., four sets of teams seeded one through four). As likely as this may seem, it has never occurred. In fact, since 1985, the year in which this was closest to occurring was 2009, when 14 of these 16 teams reached the Sweet Sixteen. On the other hand, since 1985, the smallest such number has been seven of these 16 teams, which occurred on three occasions (1986, 1990, and 2000). On average, slightly fewer than 10 of the top 16 teams reach the Sweet Sixteen. A natural question to ask is: can some probability distribution model the frequency with which each seed wins in each round of play?

To answer this question, the entire data set from the 26 completed tournaments (1985–2010) is analyzed to identify such a distribution. This distribution can be used to retrospectively analyze the past 26 tournaments and determine the frequency of rare seed appearances in each round. The distribution can also be used by sports fans and bracketology pundits to prospectively evaluate brackets and determine their likelihood of occurring. Therefore, the distribution cannot be used directly to design a bracket, but rather, can provide a means to evaluate the likelihood of a given bracket.

The paper is organized as follows. Section 2 discusses the truncated geometric distribution and investigates which rounds of the tournament it can model. Section 3 shows how this distribution can be used to compute the probabilities of different seed combinations advancing in these rounds. Section 4 reports computational results with these models and retrospectively evaluates each of the 26 sets of team seeds that reached the Final Four from 1985 to 2010. Section 5 provides concluding comments and a discussion of the limitations of the results.

**2. Truncated geometric distribution and seeds in each round**

The geometric distribution is a (non-negative) discrete random variable, defined as the number of independent and identically distributed Bernoulli random variables (with probability  $p$ ) until

the first success occurs. Therefore, if  $Y$  is a distributed geometric random variable with probability  $p$ , then its probability mass function is given by  $P\{Y=k\} = (1-p)^{k-1}p$ ,  $k = 1, 2, \dots$ . This section investigates the relationship between seed performance and the geometric distribution.

Theorem 1 provides necessary and sufficient conditions on these Bernoulli random variables such that the associated random variable follows a geometric distribution.

**Theorem 1.** (Shishebor and Towhidi [9]): Let  $X_1, X_2, \dots$  be an arbitrary sequence of Bernoulli trials. Define  $Z$  to be the number of these Bernoulli trials until the first success occurs. Then  $Z$  is a distributed geometric random variable with probability  $p$  if and only if

$$P\{X_i = 1 \mid \sum_{h=1,2,\dots,i-1} X_h = 0\} = p \text{ for all } i = 1, 2, \dots \quad (1)$$

While Theorem 1 aids in verifying whether a geometric distribution provides a good model for game outcomes, this analysis is complicated by two factors: first, the set of integers from which these outcomes can be chosen varies over the course of the tournament (e.g., seed No. 1 can only play seed No. 16 in the first round, and can only play either seeds No. 8 or No. 9 in the second round) and, second, each outcome can take on at most 16 different values (whereas the geometric distribution can generate outcomes that are arbitrarily large). The first issue can be mitigated by treating each distinct set of possible outcomes as a separate geometric distribution, while truncating the distribution can alleviate the second.

The potential winning seeds of each game are defined by a particular set for each tournament round. For example, in the first round, the seed match-ups are predetermined: a No.  $n$  seed plays a No.  $17 - n$  seed,  $n = 1, 2, \dots, 8$ , with the winner advancing to round 2. Therefore, for  $r = 1$ , there are eight non-overlapping sets of potential winners; in general, there are  $2^{4-r}$  non-overlapping sets of  $2^r$  potential winners for rounds  $r = 1, 2, 3$ , and one set of 16 Bernoulli random variables for each round  $r = 4, 5, 6$ . Table 1 lists all the seeds contained in each set; define the  $i$ th element (seed) in set  $j$  in round  $r$  as  $t_{ij,r}$  (e.g.,  $t_{2,3,2} = 6$ ).

Define the random variable  $Z_{j,r}$  as the winning seed in set  $j$  in round  $r$ . The goal is to obtain an expression for the probability that the  $i$ th seed in  $j$  in round  $r$  wins,

$$P\{Z_{j,r} = t_{ij,r}\}, \quad i = 1, 2, \dots, \min\{2^r, 16\}, \quad j = 1, 2, \dots, \max\{2^{4-r}, 1\}, \quad r = 1, 2, \dots, 6.$$

Moreover, if Eq. (1) holds for set  $j$  in round  $r$ , then  $Z_{j,r}$  is distributed as a geometric random variable with probability  $p_{j,r} = P\{X_{ij,r} = 1 \mid \sum_{h=1,2,\dots,i-1} X_{hj,r} = 0\}$ , and  $X_{hj,r}$  are Bernoulli random

**Table 1**  
Possible seeds in each set.

Round $r$	Set $j$	Possible seeds ( $t_{ij,r}$ )
1	1	(1,16)
	2	(2,15)
	3	(3,14)
	4	(4,13)
	5	(5,12)
	6	(6,11)
	7	(7,10)
	8	(8,9)
2	1	(1,8,9,16)
	2	(2,7,10,15)
	3	(3,6,11,14)
	4	(4,5,12,13)
3	1	(1,4,5,8,9,12,13,16)
	2	(2,3,6,7,10,11,14,15)
4, 5, 6	1	(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16)

**Table 2**  
 $p_{j,r}$  Estimates.

Round (r)	Set (j)	i	( $p_{j,r}$ )
1	1,2,3,4,5,6,7,8	1	(1.00, .961, .846, .788, .663, .683, .596, .461)
2	1,2,3,4	1	(.875, .644, .510, .423)
		2	(.692, .486, .725, .633)
3	1,2	1	(.721, .462)
		2	(.483, .464)
		3	(.467, .433)
		4	(.750, .353)
4	1	1	(.433)
		2	(.390)
		3	(.361)
		4	(.391)
		5	(.429)
		6	(.375)
5	1	1	(.481)
		2	(.407)
		3	(.500)
6	1	1	(.615)
		2	(.400)
		3	(.500)

variables for element (seed)  $h$  in set  $j$  to win in round  $r$ , for  $j=1,2,\dots, \max\{2^{4-r},1\}$ ,  $r=1,2,\dots,6$ .

Since the number of seeds is finite, the probability mass function of the geometric random variable must be altered so the probabilities for the desired set of outcomes sum to one (resulting in a truncated geometric random variable). Therefore, for set  $j$  in round  $r$

$$P\{Z_{j,r} = t_{i,j,r}\} = k_{j,r} p_{j,r} (1 - p_{j,r})^{i-1}, \tag{2}$$

for  $i=1,2,\dots, \min\{2^r,16\}$ ,  $j=1,2,\dots, \max\{2^{4-r},1\}$ ,  $r=1,2,\dots,6$ , where the coefficients are  $\kappa_{j,r} = 1/(1 - (1 - p_{j,r})^{2^r})$  for  $r=1,2,3$  and  $j=1,2,\dots, 2^{4-r}$ , and  $\kappa_{j,r} = 1/(1 - (1 - p_{j,r})^{16})$  for  $r=4,5,6$  and  $j=1$ .

The validity of the result in (2) depends on (1) being satisfied in each set, in each round. To empirically assess whether (1) is satisfied, the actual NCAA basketball tournament data for the 26 tournaments from 1985 to 2010 can be used and grouped based on the winning seeds in each round (see the discussion in Section 1 on the play-in game introduced in 2001 and the first four games introduced in 2011). Table 2 reports estimates for  $p_{j,r}$ , using data for the 26 tournaments from 1985 to 2010. These values for  $p_{j,r}$  are reported as  $2^{4-r}$ -tuple vectors for all the sets of potential winners, for  $r=1,2,3$ .

For a geometric distribution to fit the data for a given set in a given round, the corresponding  $p_{j,r}$  values should be identical for each seed considered (i.e., over all  $i$ ). The results in Table 2 suggest that for the later rounds (particularly the Elite Eight and beyond), the (truncated) geometric distribution may be a reasonable fit for the seed distributions.

### 3. Computing the probability of the seed combinations in each round

The results in Section 2 provide empirical validation for when the truncated geometric distribution may be used to compute the probability that a seed wins in a particular round. In this section, these results are used to obtain a closed form expression for the probability that a particular set of  $2^{6-r}$  seeds win in a particular round  $r$ .

Define  $R(r) = 2^{6-r}$  to be the number of teams that win in round  $r$ . For  $r=1,2,3,4$ , since there are four identical regions, the vector of winning seeds can be broken up into four equal parts, with  $(u_{m,1}, u_{m,2}, \dots, u_{m,R(r)/4})$  representing the seeds of the teams that win in round  $r$  from region  $m=1,2,3,4$ . Moreover, in each

of these rounds, each region can be decomposed into  $2^{4-r}$  (non-overlapping) sets, with  $2^r$  possible values for  $u_{m,n}$ ,  $n=1,2,\dots,R(r)/4$ . For  $r=5$ ,  $R(5)$  is a pair of seeds, labeled  $(u_a, u_b)$  that must come from distinct regions. For  $r=6$ ,  $R(6)$  is a single seed, labeled  $(u_c)$ .

In general, the seeds that win in any given round are dependent. For example, no more than four of any seed can win in any round. Therefore, there cannot be five No. 1 seeds winning in round  $r=3$ . However, for each value of  $r$ , the seeds that win are based on Bernoulli random variables decomposed into non-overlapping sets  $j$ , where within each set, the only restriction on these random variables is that the  $p_{j,r}$  are equal to the conditional probability expression given by (1). Moreover, since the 64 team bracket is decomposed into four identical (and independent) regions, then there are four independent geometric random variables for each set, for  $r=1,2,3,4$ , two independent geometric random variables for  $r=5$ , and one geometric random variables for  $r=6$ . Therefore, the set and region structure is such that for round  $r=1,2,3,4$

$$P\{u_{1,1}, u_{1,2}, \dots, u_{1,R(r)/4}, u_{2,1}, u_{2,2}, \dots, u_{2,R(r)/4}, u_{3,1}, u_{3,2}, \dots, u_{3,R(r)/4}, u_{4,1}, u_{4,2}, \dots, u_{4,R(r)/4}\} = \prod_{j=1,2,\dots,2^{4-r}} M_j(u_{1,j}, u_{2,j}, u_{3,j}, u_{4,j})$$

$$P_{m=1,2,3,4} P\{Z_{j,r} = u_{m,j,r}\}, \tag{3}$$

where the coefficient  $M_j(u_{1,j}, u_{2,j}, u_{3,j}, u_{4,j})$  is the number of distinct permutations that the four seeds can assume in set  $j$  across the four regions (i.e., a multinomial coefficient over four items). For  $r=5$

$$P\{u_a, u_b\} = B(u_a, u_b) P\{Z_{1,5} = u_a\} P\{Z_{1,5} = u_b\}, \tag{4}$$

where the coefficient  $B(u_a, u_b)$  is the number of distinct permutations that the two seeds can assume (i.e., a binomial coefficient over two items). For  $r=6$

$$P\{u_c\} = P\{Z_{1,6} = u_c\}. \tag{5}$$

### 4. Computational results

This section uses the results from the 26 tournaments from 1985 to 2010 to compute the distribution of the seeds that win in each round. The results are reported in two sections: (i) the second round, the Sweet Sixteen and the Elite Eight ( $r=2,3,4$ ), and (ii) the National Semifinals and National Championship game ( $r=5,6$ ). The observed number of times that each seed has won in a given round will be compared to the expected number of times that the seeds should occur based on the truncated geometric distribution.

The estimated values for  $p_{j,r}$  are computed by the method of moments. In particular, if  $Y(N,p)$  is a truncated geometric random variable with parameter  $p$  and  $N$  observations, then  $E(Y(N,p)) = (1/p) - N(1-p)^N / (1 - (1-p)^N)$ . For each set  $j$  in each round  $r$ , this expression is equal to the average seed position (within the set) based on the data for the 26 tournaments from 1985 to 2010; denote these computed values as  $A_{j,r}$ . Therefore, for each set  $j$  and round  $r$ , the goal is to solve for the value of  $p_{j,r}$  such that  $E(Y(\min\{2^r,16\}, p_{j,r})) = A_{j,r}$ . Given that the expression for  $E(Y(N,p))$  is monotonically decreasing in  $p$  (since its first derivative is negative for  $0 < p < 1$ ), a simple (iterative) bisection algorithm is used to solve for an estimate for  $p_{j,r}$ . In particular, starting with the two values 0 and 1 for  $p_{j,r}$ , the interval is successively halved until the desired value for  $p_{j,r}$  is obtained. All these values are reported in Table 3 (to three decimal places).

#### 4.1. The second round, the Sweet Sixteen, and the Elite Eight ( $r=2,3,4$ )

Upsets occur with great frequency in the early rounds, resulting in numerous Sweet Sixteens and Elite Eights containing low seeded teams. In particular, there have been 36 teams seeded

No. 11 or lower and 50 teams seeded between No. 7 to No. 10 reaching the Sweet Sixteen from 1985 through 2010. There have also been 5 teams seeded No. 11 or lower and 20 teams seeded between No. 7 to No. 10 reaching the Elite Eight from 1985 through 2010. Therefore, lower seeded teams appear regularly in such rounds.

Table 4 reports the number of times that the different seeds won in the second round (and hence, reached the Sweet Sixteen), won in the Sweet Sixteen (and hence, reached the Elite Eight) and won in the Elite Eight (and hence, reached the Final Four), from 1985 through 2010, as well as the expected number of times that each seed should occur, assuming that this random variable follows a truncated geometric distribution. For the second round, the square of the difference between the expected and observed values, divided by the expected value, is also reported, to provide a standardized measure for the difference between the expected and the observed values (i.e., the smaller the value, the smaller the difference); this value is labeled as  $\delta_n$  for seed  $n$ . The values in Columns 2, 3, and 4 in Table 4 suggest that the truncated geometric distribution is a poor model for seed appearances in the Sweet Sixteen. This was also corroborated by the  $p_{i,r}$  values not being the same for different values for  $i$  in Table 2.

While upsets have occurred regularly, some lower seeds have not won frequently in past tournaments. Therefore, it may be prudent and practical to pool such seeds when estimating the probability associated with any set of seeds in any round. If the three lowest seeds in each set are pooled (i.e., (8, 9, 16) for set 1, (7, 10, 15) for set 2, (6, 11, 14) for set 3, and (5, 12, 13) for set 4), then the expected number of times that each of these groups of pooled seeds should have reached each of these rounds is 14.70, 38.73, 44.21, and 51.96, respectively, with 13, 37, 51, and 60 of these seeds actually reaching the Sweet Sixteen. On the other hand, if the two highest seeds in each set are pooled, and the two lowest seeds in each set are pooled (i.e., (1, 8) and (9, 16) for set 1, (2, 7) and (10, 15) for set 2, (3, 6) and (11, 14) for set 3, and (4, 5)

and (12, 13) for set 4), then the expected number of times that each of these pairs of groups of pooled seeds should have reached each of these rounds is (101.95, 2.05), (90.49, 13.51), (86.60, 17.40) and (80.30, 23.70), respectively, with (100, 4), (85, 19), (90, 14), and (82, 22) of these seed pairs actually reaching the Sweet Sixteen. This suggests that pooling seeds that are less represented may provide a more accurate approach to compute the expected number of such seeds that reach this round. Of course, since pooling can be done rather arbitrarily, some reasonable basis for such pooling should be provided.

Columns 5 and 6 in Table 4 report the number of times that the different seeds won in the Sweet Sixteen (and hence, reached the Elite Eight) from 1985 through 2010, as well as the expected number of times that each seed should occur, assuming that this random variable follows a truncated geometric distribution. The  $X^2$  test statistic for set 1 is 12.56, resulting in a  $p$ -value between .05 and .10 (suggesting a questionable fit to the truncated geometric distribution), while the  $X^2$  test statistic for set 2 is 5.54, resulting in a  $p$ -value  $> .50$  (suggesting a very strong fit to the truncated geometric distribution). Once again, the small sample sizes for lower seeds make it difficult to accurately predict how often each seed is expected to reach the Sweet Sixteen. If the five lowest seeds in each set are pooled (i.e., 8, 9, 12, 13, 16 for set 1 and 7, 10, 11, 14, 15 for set 2), then the expected number of times that each of these groups of pooled seeds should have reached these rounds are 3.28 and 16.14, respectively, with 8 and 17 of these seeds actually reaching the Elite Eight. With this pooling, the  $X^2$  test statistic for set 1 is 10.22, resulting in a  $p$ -value  $< .020$  (suggesting a questionable fit to the truncated geometric distribution), while the  $X^2$  test statistic for set 2 is 0.14, resulting in a  $p$ -value  $> .98$  (suggesting a very strong fit to the truncated geometric distribution). These results are consistent with conclusions drawn from the values reported in Table 2; set  $i=1$  did not appear to strongly follow a truncated geometric distribution, while set  $i=2$  did appear to follow a truncated geometric distribution.

Columns 7 and 8 in Table 4 report the number of times that different seeds won in the Elite Eight (and hence, reached the Final Four) from 1985 through 2010, as well as the expected number of times that each seed should occur, assuming that this random variable follows a truncated geometric distribution. These results suggest a good fit for this data to a truncated geometric random variable. Using a  $X^2$  goodness of fit test, the resulting test statistic value (with 15 degrees of freedom) is 19.22 with a  $p$ -value  $> .20$ . Note that the two No. 11 seed teams that reached

**Table 3**  
 $p_{i,r}$  Values based on the method of moments.

Round ( $r$ )	Set ( $j$ )	( $p_{i,r}$ )
2	1,2,3,4	(.858, .614, .552, .457)
3	1,2	(.684, .455)
4	1	(.400)
5	1	(.456)
6	1	(.510)

**Table 4**  
Seed frequency of teams who reach the Sweet Sixteen, the Elite Eight, and the Final Four.

Seed ( $n$ )	Sweet Sixteen appearances	Sweet Sixteen expected frequency	$\delta_n$	Elite Eight appearances	Elite Eight expected frequency	Final Four appearances	Final Four expected frequency
1	91	89.30	.03	75	71.12	45	41.61
2	67	65.27	.05	48	47.70	23	24.97
3	53	59.79	.77	26	25.99	13	14.98
4	44	52.04	1.24	14	22.49	9	8.99
5	38	28.27	3.35	7	7.11	6	5.39
6	37	26.80	3.87	13	14.16	3	3.24
7	18	25.22	2.07	6	7.72	0	1.94
8	9	12.65	1.05	6	2.25	3	1.16
9	4	1.79	2.72	1	0.71	0	0.70
10	19	9.75	8.79	7	4.21	0	0.42
11	12	12.02	.00	4	2.29	2	0.25
12	18	15.35	.46	1	0.22	0	0.15
13	4	8.34	2.26	0	0.07	0	0.09
14	2	5.39	2.13	0	1.25	0	0.05
15	0	3.77	3.77	0	0.68	0	0.03
16	0	0.25	.25	0	0.02	0	0.02

the Final Four contribute 12.15 of the 19.22 value to the test statistic value. Since the expected number of seed appearances was below 5 for teams seeded No. 6 or lower, if all these teams are pooled, the expected number of times that this pooled group of seeds should have reached this round are 8.06, with eight of these seeds actually doing so. Moreover, the resulting  $X^2$  goodness of fit test statistic value (with five degrees of freedom) is 0.76 with a  $p$ -value  $> .975$ , which suggests a strong fit to the truncated geometric distribution.

4.2. The National Semifinals and National Championship games ( $r=5,6$ )

There are two winners in the National Semifinals and one winner in the National Championship game. Columns 2 and 3 in Table 5 report the number of times that each seed has won in the National Semifinals (and hence, reached the National Final game), from 1985 through 2010, as well as the expected number of times that each seed should have won, assuming that this random variable follows a truncated geometric distribution.

Using a  $X^2$  goodness of fit test, the resulting test statistic value (with 15 degrees of freedom) is 4.77 with a  $p$ -value  $> .99$ . Since the expected number of seed appearances was below 5 for teams seeded No. 4 or lower, if all these teams are pooled, the expected number of times that this pooled group of seeds should have reached this round are 8.37, with eight of these seeds actually doing so, resulting in  $X^2$  a test statistic value of 0.50 (with 3 degrees of freedom), with a  $p$ -value  $> .91$ .

Columns 4 and 5 in Table 5 report the number of times that each seed has won the National Championship, from 1985 through 2010, as well as the expected number of times that each seed should have won, assuming that this random variable follows a truncated geometric distribution. The  $X^2$  goodness of fit test statistic value (with 15 degrees of freedom) is 13.00 with a  $p$ -value  $> .60$ . This is due to the No. 8 team winning the National Championship (which contributes 9.18 to the test statistic). Since the expected number of seed appearances was below 3 for teams seeded No. 4 or lower, if all these teams are pooled, the expected number of national champions in this pooled group of seeds are 3.06, with 3 of these seeds actually doing so, resulting in a  $X^2$  test statistic value of 1.54 (with 3 degrees of freedom), with a  $p$ -value  $> .67$ .

Table 6 reports the probability of each of the Final Four seeds occurring for the 26 tournaments between 1985 and 2010. The reciprocal of this is also reported as the expected frequency (as

**Table 5**  
Seed frequency of teams who reach the National Finals game and who win the National Championship.

Seed (n)	National Final appearances	National Final expected frequency	National Champions	National Champion expected frequency
1	25	23.71	16	13.26
2	11	12.90	4	6.50
3	8	7.02	3	3.19
4	2	3.82	1	1.56
5	3	2.08	0	.77
6	2	1.13	1	.38
7	0	.61	0	.18
8	1	.33	1	.09
9	0	.18	0	.044
10	0	.10	0	.022
11	0	.05	0	.011
12	0	.03	0	.005
13	0	.02	0	.003
14	0	.01	0	.001
15	0	.00	0	.001
16	0	.00	0	.000

measured in (tournament) years) between the occurrence of such seed combinations. As noted previously, since the expected number of seed appearances was below five for teams seeded No. 6 or lower (the actual data shows that 96 of the 104 teams have been seeded Nos. 1, 2, 3, 4, or 5), such sparse data suggest that the probabilities computed for Final Four seed combinations teams seeded No. 6 or lower may be subject to greater error.

The seed combinations with the highest probabilities are 1123, 1112, and 1122, which should each occur on average once every 15, 16, and 18 years, respectively. As the case may be, 1123 occurred three times, 1112 occurred once, and 1122 occurred once, from 1985 through 2010, which are consistent with their expected frequency, given the small sample sizes. The rarest seed combinations that have occurred are 1588 (with an expected frequency of once every 32,015 years) and 234(11) (with an expected frequency of once every 5762 years). Note that seed combination 234(11) has also been the only Final Four seed combination not containing a No. 1 seed. Lastly, given that a No. 16 seed has never won a game, the probability that all four No. 16 seeds reach the Final Four is  $1.25 \times 10^{-15}$  or an expected frequency of once every  $8 \times 10^{14}$  years. To put this number into perspective, there are approximately  $10^{11}$  stars in the Milky Way galaxy and the US national debt (in US\$, as of February 2011) was  $1.4 \times 10^{13}$ .

Given the prevalence of teams seeded No. 1, Table 7 provides the probabilities and the expected number of occurrences (from 1985 to 2010) of Final Four seed combinations based on the number of No. 1 seeds. The rarest combination occurs with all four No. 1 seeds reaching the Final Four. The second rarest is when zero No. 1 seeds reach the Final Four. Both events have occurred exactly once (2008 and 2006, respectively). The most common event is when exactly two No. 1 seeds reach the Final Four, which occurs on average slightly more than once every three years; this has in fact occurred in twelve of the past 26 tournaments.

Clearly, No. 1 seeds dominate the Final Four. In fact, the probability of two or more No. 1 seeds reaching the Final Four

**Table 6**  
Probability of Final Four seeding (1985–2010).

Year	Seeds	Probability	Expected frequency (years)
1985	1,1,2,8	.0052	194
1986	1,1,2,11	.0011	896
1987	1,1,2,6	.014	70
1988	1,1,2,6	.014	70
1989	1,2,3,3	.024	42
1990	1,3,4,4	.0052	194
1991	1,1,2,3	.066	15
1992	1,2,4,6	.0062	161
1993	1,1,1,2	.062	16
1994	1,2,2,3	.040	25
1995	1,2,2,4	.024	42
1996	1,1,4,5	.0086	116
1997	1,1,1,4	.022	45
1998	1,2,3,3	.024	42
1999	1,1,1,4	.022	45
2000	1,5,8,8	$3.12 \times 10^{-5}$	32015
2001	1,1,2,3	.066	15
2002	1,1,2,5	.024	42
2003	1,2,3,3	.024	42
2004	1,2,2,3	.040	25
2005	1,1,4,5	.0086	116
2006	2,3,4,11	$1.74 \times 10^{-4}$	5763
2007	1,1,2,2	.055	18
2008	1,1,1,1	.026	39
2009	1,1,2,3	.066	15
2010	1,2,5,5	.0031	322

**Table 7**  
Probability of Final Four seed combinations.

Scenario	Probability	Expected number of occurrences	Number of times actually occurred
Exactly zero No. 1 seeds	.130	3.4	1
Exactly one No. 1 seeds	.346	9.0	10
Exactly two No. 1 seeds	.346	9.0	11
Exactly three No. 1 seeds	.154	4.0	3
Exactly four No. 1 seeds	.026	0.7	1

is .525 (slightly more than once every 2 years). Such Final Four combinations have occurred 15 times since 1985. Lastly, when there is exactly one team seeded No. 3 or lower in the Final Four, the unconditional probability that the three other teams are seeded 111 (i.e., a 111X seed combination with  $X \geq 3$ ) is significantly lower than these three teams being seeded 112 (.051 versus .166). In fact, the 112X seed combination (with  $X \geq 3$ ) has occurred eight times since 1985, while the 111X seed combination (with  $X \geq 3$ ) has occurred just twice since 1985. Similar seed combination oddities can be computed using the model presented.

## 5. Summary and conclusions

The NCAA Division I men's basketball tournament provides a cornucopia of challenges to those who search for patterns in how seeds advance through the tournament. This paper shows that a truncated geometric distribution can be used to compute the probability that a set of seeds will win in a given round. Using data from the 1985 to 2010 tournaments, this distribution appears to be most valid for the Elite Eight, the National Semifinals, and the National Championship Game rounds. There is also some indication that this distribution may also be reasonable in the Sweet Sixteen.

The results in this paper are not designed to help budding bracketologists create winning brackets, but rather, evaluate brackets and determine the likelihood that the combination of

seeds in each round (particularly the last three rounds) will occur. With each passing tournament, additional data points will become available, which will further refine the estimates and provide a more useful tool for sports fans. Until then, March Madness will continue to capture the interest of the nation through both the excitement that it inspires in basketball fans and the cornucopia of statistics that it generates.

## Acknowledgements

The authors would like to thank the Associate editor and three anonymous referees for their helpful comments and suggestions, resulting in a significantly improved manuscript. The author would also like to thank Ammar Rizwan and Emon Dai for their comments and feedback on this research, and their efforts in creating the website, <http://bracketodds.cs.illinois.edu>, which transitions this research into an interactive tool. The computational results were obtained with support from the Simulation and Optimization Laboratory in the Department of Computer Science at the University of Illinois at Urbana-Champaign.

## References

- [1] Horowitz I. Seeds, spreads, and synchronicity in the big dance betting market. *International Journal of Applied Decision Sciences* 2010;3(1):1–14.
- [2] Fearnhead P, Taylor BM. Calculating strength of schedule, and choosing teams for March Madness. *American Statistician* 2010;64(2):108–15.
- [3] Koenker R, Bassett Jr. GW. March Madness, Quantile Regression Bracketology, and the Hayek Hypothesis. *Journal of Business and Economic Statistics* 2010;28(1):26–35.
- [4] Kvam P, Sokol J. A logistic regression/Markov chain model for NCAA Basketball. *Naval Research Logistics* 2006;53(8):788–803.
- [5] Smith JC, Fraticelli BMP, Rainwater C. A bracket assignment problem for the NCAA Men's Basketball Tournament. *International Transactions in Operational Research* 2006;13(3):253–71.
- [6] Jacobson SH, King DM. Seeding in the NCAA Men's Basketball Tournament: when is a higher seed better? *Journal of Gambling Business and Economics* 2009;3(2):63–87.
- [7] NCAA. Official 2007 NCAA Men's Final Four Records Book. Indianapolis, IN: The National Collegiate Athletic Association; 2007.
- [8] Baumann R, Matheson VA, Howe CA. Anomalies in tournament design: the madness of March Madness. *Journal of Quantitative Analysis in Sports* 2010;6(2). Article 4.
- [9] Shishebor Z, Towhidi M. On the generalization of negative binomial distribution. *Statistics and Probability Letters* 2004;66:127–33.