Seed Distributions for the NCAA Men’s Basketball Tournament: Why it May Not Matter Who Plays Whom*

Sheldon H. Jacobson
Department of Computer Science
University of Illinois at Urbana-Champaign
shj@illinois.edu
https://netfiles.uiuc.edu/shj/www/shj.html

* Joint work with, Alexander G. Nikolaev, Adrian, J. Lee, Douglas M. King
NCAA Men’s Basketball Tournament

• National Collegiate Athletic Association (NCAA) Men’s DI College Basketball Tournament (aka March Madness)
  – First held in 1939 with 8 teams
  – Since 1985, 64 teams participate annually
    • Increased to 68 teams with four play-in games (2011)

• Popularity of gambling on tournament games
  – Estimated $2.25B (US) wagered on 2007 Final Four through illegal channels alone
  – Common types of gambling: traditional (single game) and office pool (entire tournament bracket)
  – Goal: Forecast the winners of one or more tournament games
Predicting Game Winners

Models have been proposed to forecast game winners (e.g., binary win/lose, final score difference)

• Predictors:
  – Outcomes of season games (winner, score)
  – Las Vegas odds
  – Other rankings (RPI, Sagarin, Massey, Pomeroy)

• Useful to the general public?
  – Difficult to gather relevant predictor data and implement the model
  – Simple alternatives are attractive
Tournament Structure

• Selection committee
  – Chooses 37 “at large” participants (31 conference champions)
  – Creates 4 regions of 16 teams each (plus 4 play-in game teams)
  – Assigns an integer seed to each team in each region, with values from 1 (best) to 16 (worst)
  – Several issues unrelated to team skill are considered (geography, conference affiliation) when placing teams in regions

• Format of the bracket in each region
  – Single elimination
  – First round: seed $k$ plays seed $17-k$
  – Later rounds: opponents determined by results of earlier rounds
<table>
<thead>
<tr>
<th>ROUND 1</th>
<th>ROUND 2</th>
<th>ROUND 3</th>
<th>ROUND 4</th>
<th>REGIONAL WINNER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed 1</td>
<td></td>
<td></td>
<td></td>
<td>{1, 16}</td>
</tr>
<tr>
<td>Seed 16</td>
<td>{1, 16}</td>
<td></td>
<td></td>
<td>{1, 8, 9, 16}</td>
</tr>
<tr>
<td>Seed 8</td>
<td></td>
<td></td>
<td>{1, 8, 9, 16}</td>
<td></td>
</tr>
<tr>
<td>Seed 9</td>
<td></td>
<td>{8, 9}</td>
<td></td>
<td>{1, 4, 5, 8, 9, 12, 13, 16}</td>
</tr>
<tr>
<td>Seed 5</td>
<td></td>
<td></td>
<td>{5, 12}</td>
<td></td>
</tr>
<tr>
<td>Seed 12</td>
<td></td>
<td></td>
<td></td>
<td>{1, 2, …, 16}</td>
</tr>
<tr>
<td>Seed 4</td>
<td></td>
<td>{4, 13}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seed 13</td>
<td></td>
<td></td>
<td>{4, 13}</td>
<td></td>
</tr>
<tr>
<td>Seed 6</td>
<td></td>
<td></td>
<td>{6, 11}</td>
<td></td>
</tr>
<tr>
<td>Seed 11</td>
<td></td>
<td></td>
<td>{3, 6, 11, 14}</td>
<td></td>
</tr>
<tr>
<td>Seed 3</td>
<td></td>
<td>{3, 14}</td>
<td></td>
<td>{1, 2, 3, 6, 7, 10, 11, 14, 15}</td>
</tr>
<tr>
<td>Seed 14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seed 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seed 10</td>
<td></td>
<td></td>
<td>{7, 10}</td>
<td></td>
</tr>
<tr>
<td>Seed 2</td>
<td></td>
<td></td>
<td></td>
<td>{2, 7, 10, 15}</td>
</tr>
<tr>
<td>Seed 15</td>
<td>{2, 15}</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

List of possible seeds in each game of the regional tournaments

(C)2011 Jacobson NCAA BB March Madness
If “best” teams win in the first round, seed \( k \) plays seed 9-\( k \) in the second round.
If “best” teams win in the second round, seed $k$ plays seed $5-k$ in the third round.
<table>
<thead>
<tr>
<th>ROUND 1</th>
<th>ROUND 2</th>
<th>ROUND 3</th>
<th>ROUND 4</th>
<th>REGIONAL WINNER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seed 1</td>
<td>Seed 1</td>
<td>Seed 1</td>
<td>Seed 1</td>
<td>Seed 1</td>
</tr>
<tr>
<td>Seed 16</td>
<td>Seed 8</td>
<td>Seed 8</td>
<td>Seed 5</td>
<td></td>
</tr>
<tr>
<td>Seed 9</td>
<td>Seed 4</td>
<td>Seed 4</td>
<td>Seed 4</td>
<td></td>
</tr>
<tr>
<td>Seed 6</td>
<td>Seed 13</td>
<td>Seed 3</td>
<td>Seed 3</td>
<td>Seed 2</td>
</tr>
<tr>
<td>Seed 7</td>
<td>Seed 10</td>
<td>Seed 7</td>
<td>Seed 2</td>
<td></td>
</tr>
<tr>
<td>Seed 2</td>
<td>Seed 15</td>
<td>Seed 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(©2011 Jacobson NCAA BB March Madness)
The Final Four

- Four regional winners meet in two more rounds
- Two identical seeds can play in a single game
- Any seed can play against any seed (in theory)

ROUND 5  ROUND 6  TOURNAMENT CHAMPION

Reg1 Winner

Reg2 Winner

Reg3 Winner

Reg4 Winner
Is It Best To Pick the Better Seed?

• One way to forecast winners: **Pick the better seed**
  – Simplicity of this method makes it attractive
  – Does it provide good predictions?

• Selection committee tends to assign better seeds to better teams

• When seed differences are large, games tend to be more predictable (and hence, fewer upsets)
Predictions by Round

• As the tournament progresses, seed differences tend to be smaller
  – 70% in round 4 (Elite Eight) have been seeded No. 3 or better
  – 76% in round 5 (National Semi-final) have been seeded No. 3 or better
  – 83% in round 6 (National Final) have been seeded No. 3 or better
  – 89% of tournament champions have been seeded No. 3 or better

• Other indicators of success?
  – To appear in the \( r^{th} \) round, a team must have won its preceding \( r-1 \) games
  – Teams with worse seeds tend to face more skilled competition earlier in the tournament

• Are seed less informative as tournament progresses?
  – Jacobson and King (2009) focus on the top three seeds.
Goals of the Study

- Compare historical performance of the seed distributions in each round.
- Model the seed distributions in each round
- Comparisons model with statistical hypothesis testing
  - $X^2$ Goodness-of-fit
- Data Sources
Statistical Hypothesis Testing

Requirements

• A sufficient number of samples
  – 1,638 total games (63 games over 26 years)
    • Play-in and First Four games not included
  – When subsets are taken based on seeds and rounds, sample sizes drop dramatically

• A random sample. To this effect, assume:
  – Historical data are a representative sample of each seed’s performance
  – Each seed has a constant probability of winning against any other seed in a specified round
The Math Behind The Numbers
Geometric Distribution

• Common (nonnegative) discrete random variable.

• Defined as the number of independent and identically distributed Bernoulli random variables (with probability $p$) until the first success occurs.

• If $Y$ is distributed geometric with probability $p$, then
  
  $P\{Y=k\} = (1-p)^{k-1}p, \; k=1,2,\ldots$
Key Theorem*

Let $X_1, X_2, \ldots$ be an arbitrary sequence of Bernoulli trials. Let $Z$ be the number of these Bernoulli trials until the first success. Then $Z$ is a geometric random variable with probability $p$ iff

$$P\{X_i = 1 \mid \sum_{h=1,2,\ldots,i-1} X_h = 0\} = p \text{ for all } i = 1,2,\ldots.$$

**Implication:** Provides a N&S condition for a geometric RV.

**Intuition:** If the first $i-1$ seed positions have not advanced to the next round (i.e., won), then the probability that the $i$th seed position advances is $p$, the same value for all seed positions $i$.

* Shishebor and Towhidi (2004)
Sets of Seeds in Each Round

- Possible seeds defined by *sets of seeds* in each round
  - First round: Seed No. $n$ plays Seed No. $17-n$, $n = 1,2,...,8$

- Rounds $r = 1,2,3$:
  - $2^{4-r}$ non-overlapping sets of $2^r$ possible winners
    - $r = 1$: $\{1,16\}$, $\{2,15\}$, $\{3,14\}$, $\{4,13\}$, $\{5,12\}$, $\{6,11\}$, $\{7,10\}$, $\{8,9\}$
    - $r = 2$: $\{1,8,9,16\}$, $\{2,7,10,15\}$, $\{3,6,11,14\}$, $\{4,5,12,13\}$
    - $r = 3$: $\{1,4,5,8,9,12,13,16\}$, $\{2,3,6,7,10,11,14,15\}$

- Rounds $r = 4,5,6$:
  - One set of 16 possible winners

Define $Z_{j,r}$ as the $j^{th}$ set in the $r^{th}$ round
Define $t_{i,j,r}$ as the $i^{th}$ element in set $Z_{j,r}$
Truncated Geometric Distribution

Truncate the geometric distribution (finite number of seeds)

- Ensure that discrete probabilities sum to one

For set $j$ in round $r$,

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

- Coefficients:
  - $\kappa_{j,r} = 1/(1-(1-p_{j,r})^{2^r})$ for set $j = 1,2,\ldots, 2^{4-r}$ in round $r = 1,2,3$
  - $\kappa_r = 1/(1-(1-p_{r,1})^{16})$ for round $r = 4,5,6$ (only one set $j = 1$).

Important Note:

$p_{j,r}$ must be estimated for each position in each set in each round.
Geometric Distribution Validation: Values for $p_{j,r}$

<table>
<thead>
<tr>
<th>Round $r$</th>
<th>Set $j$</th>
<th>Position $i$</th>
<th>$p_{j,r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1,2,3,4</td>
<td>1</td>
<td>(.875, .644, .510, .423)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>(.692, .486, .725, .633)</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>(.433)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>(.390)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>(.361)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>(.391)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>(.429)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>(.375)</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>(.615)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>(.400)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>(.500)</td>
</tr>
</tbody>
</table>
Probability of Seed Combinations

\[ R(r) = 2^{6-r} = \text{number of teams that win in round } r = 1, 2, \ldots, 6. \]
- Teams that advance to the next round

Given that there are four nonoverlapping regions, there are
- four independent geometric rv’s for each set in round \( r = 1, 2, 3, 4, \)
- two independent geometric rv’s for \( r = 5, \)
- one geometric rv’s for \( r = 6 \)

Probability of seed combinations in a round are computed by taking the product of
- Probabilities of each seed appearing in that round
- Number of distinct permutations that the four seeds can assume in set \( j \) in round \( r \) across the four regions
Estimates for $p_{j,r}$

- Estimates for $p_{j,r}$ computed by method of moments
- $Y(n,p)$ truncated geometric with parameter $p$ and $n$
  \[ E(Y(n,p)) = \frac{1}{p} - \frac{n(1-p)^n}{1-(1-p)^n} \]
- Iterative bisection algorithm used to solve for an estimate of $p_{j,r}$ using the average seed position over the past 26 tournaments in each set ($j$) within each round ($r$)

<table>
<thead>
<tr>
<th>Round $r$</th>
<th>Set $j$</th>
<th>$p_{j,r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 (Elite Eight)</td>
<td>1,2</td>
<td>(.684, .455)</td>
</tr>
<tr>
<td>5 (National Finals)</td>
<td>1</td>
<td>(.456)</td>
</tr>
</tbody>
</table>

(C)2011 Jacobson
NCAA BB March Madness
The Final Four
## Seed Frequency in Final Four

<table>
<thead>
<tr>
<th>Seed n</th>
<th>No. Times Actually Appeared</th>
<th>Expected No. Times Should Appear</th>
<th>$\delta_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>25.0</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>15.0</td>
<td>0.26</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>9.0</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>5.4</td>
<td>0.07</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>3.2</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1.2</td>
<td>2.89</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.7</td>
<td>0.09</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.0</td>
<td>0.05</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0.0</td>
<td>0.02</td>
</tr>
</tbody>
</table>

$\delta_i = \frac{(45 - 41.6)^2}{41.6} = 0.28$
### Final Four Seed Combinations

- Compute probability of Final Four seed combinations
- Reciprocal is expected frequency between occurrences

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Probability</th>
<th>Expected # Occurrences</th>
<th># Actual Occurrences</th>
<th>Expected Frequency (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero No. 1 Seeds</td>
<td>0.130</td>
<td>3.4</td>
<td>1</td>
<td>7.70</td>
</tr>
<tr>
<td>One No. 1 Seed</td>
<td>0.346</td>
<td>9.0</td>
<td>10</td>
<td>2.89</td>
</tr>
<tr>
<td>Two No. 1 Seeds</td>
<td>0.346</td>
<td>9.0</td>
<td>11</td>
<td>2.89</td>
</tr>
<tr>
<td>Three No. 1 Seeds</td>
<td>0.154</td>
<td>4.0</td>
<td>3</td>
<td>6.49</td>
</tr>
<tr>
<td>Four No. 1 Seeds</td>
<td>0.026</td>
<td>0.7</td>
<td>1</td>
<td>38.46</td>
</tr>
</tbody>
</table>

(C)2011 Jacobson

NCAA BB March Madness
# Most Likely Final Four Seed Combinations

<table>
<thead>
<tr>
<th>Seeds</th>
<th>Actual Occurrences (Tournament Year)</th>
<th>Probability</th>
<th>Expected Frequency (in Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,1,2</td>
<td>1991, 2001, 2009</td>
<td>0.066</td>
<td>15</td>
</tr>
<tr>
<td>1,1,2</td>
<td>1993</td>
<td>0.062</td>
<td>16</td>
</tr>
<tr>
<td>1,2,2</td>
<td>2007</td>
<td>0.055</td>
<td>18</td>
</tr>
<tr>
<td>1,2,3</td>
<td>1994, 2004</td>
<td>0.040</td>
<td>25</td>
</tr>
<tr>
<td>1,2,3</td>
<td>1989, 1998, 2003</td>
<td>0.024</td>
<td>42</td>
</tr>
<tr>
<td>1,5,8,8</td>
<td>2000</td>
<td>0.0000312</td>
<td>32015</td>
</tr>
</tbody>
</table>

* Compiled based on data from 1985-2010 tournaments

(C)2011 Jacobson

NCAA BB March Madness
## Final Four Seed Combination Odds

<table>
<thead>
<tr>
<th>Seed Description</th>
<th>Probability</th>
<th>Expected Frequency (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One or More 16</td>
<td>0.000756</td>
<td>1307</td>
</tr>
<tr>
<td>One or More 15 or 16</td>
<td>0.002037</td>
<td>491</td>
</tr>
<tr>
<td>One or More 14, 15, or 16</td>
<td>0.004152</td>
<td>241</td>
</tr>
<tr>
<td>One or More 13, 14, 15, or 16</td>
<td>0.007665</td>
<td>130</td>
</tr>
<tr>
<td>One or More 12, 13, 14, 15, or 16</td>
<td>0.013493</td>
<td>74</td>
</tr>
<tr>
<td>One or More 11, 12, 13, 14, 15, or 16</td>
<td>0.023137</td>
<td>43</td>
</tr>
<tr>
<td>All 16</td>
<td>1.34E-15</td>
<td>747 trillion</td>
</tr>
<tr>
<td>No teams 1, 2, or 3</td>
<td>0.00220</td>
<td>454</td>
</tr>
</tbody>
</table>

* Compiled based on data from 1985-2010 tournaments

(C)2011 Jacobson  NCAA BB March Madness
2011 Final Four

Odds against any 3,4,8,11 seeds in the Final Four:
121,000 to 1

Odds against UConn, UKentucky, Butler, VCU in the FF:
2.9 Million to 1

Probability of UConn (#3) winning the NC: 0.0306

Number of ESPN Brackets: 5.9 Million

Number who chose UConn: 279,308

Expected number picking UConn, assuming all No. 3 seeds are equally likely: 181,000
2011 Final Four

Probability of UKentucky (#4) winning the NC: .0150
Number of ESPN Brackets that chose UKentucky: 107,249
Expected number picking UKentucky, assuming all No. 4 seeds are equally likely: 89,000

Probability of Butler (#8) winning the NC: .00347
Number of ESPN brackets that chose Butler: 4,325
Expected number picking Butler, assuming all No. 8 seeds are equally likely: 5,100

Probability of VCU (#8) winning the NC: .000102
Number of ESPN brackets that chose VCU: 1,023
Expected number picking VCU, assuming all No. 11 seeds are equally likely: 600
Conclusions and Limitations

• Truncated geometric distribution used to compute probability of seed combinations in each round
  – Distribution fits closest (via $X^2$ goodness of fit test) in later rounds of tournament (Elite Eight and onwards)

• Rule changes may impact seed winning probabilities over time
  – Introduction of 35 second clock
  – Expansion of three point arc
  – Selection committee criteria changes

• Distribution parameters, $p_{j,r}$, must be updated annually following each year’s tournament
March Madness
Let the games begin!

http://bracketodds.cs.illinois.edu

Website Developers: Ammar Rizwan and Emon Dai
(Students, Department of Computer Science, University of Illinois at Urbana-Champaign)
Website Functionality

Uses model to odds against seed combinations in

- Elite Eight
- Final Four
- National Finals
- National Championship

Allows one to
- Compare the relative likelihood of seed combinations
- Compute conditional probabilities of seed combinations in the final two rounds.

Note: Model can do much more than the web site functionality.
Thank you

http://bracketodds.cs.illinois.edu

Sheldon H. Jacobson, Ph.D.
https://netfiles.uiuc.edu/shj/www/shj.html
(217) 244-7275
Skype: sheldon.jacobson1
shj@illinois.edu