Data-Driven Advice for Filling Out a Winning NCAA Tournament Bracket

Dale L. Zimmerman

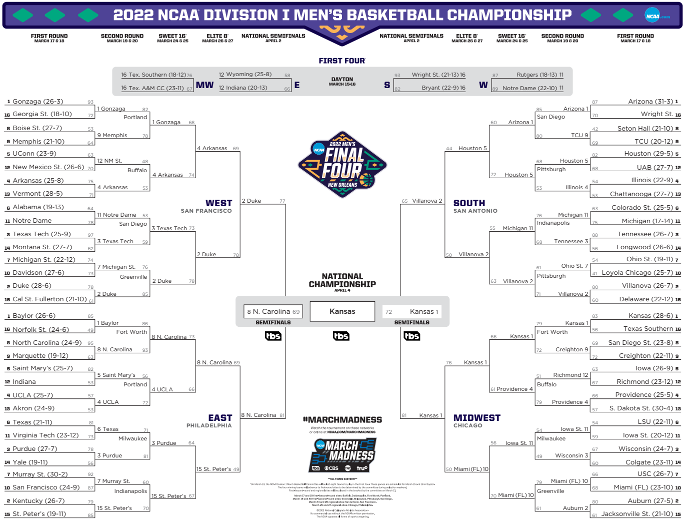
Department of Statistics and Actuarial Science
University of Iowa

March 7, 2023

Published articles related to today's chat

- Harville, D.A. and Smith, M.H. (1994). The Home-Court Advantage in Basketball: How Large Is It and Does it Vary from Team to Team? *The American Statistician*, **48**, 22–28.
- Carlin, B.P. (1996). Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information. *The American Statistician*, **50**, 39–43.
- Zimmerman, D.L., Zimmerman, N.D., and Zimmerman, J.T. (2021). March Madness "Anomalies": Are They Real, and If So, Can They Be Explained? *The American Statistician*, **75**, 207–216.

NCAA Tournament Bracket (from 2022)



Some elementary probability calculations

- Probability of filling out a perfect bracket, if we flip a fair coin to choose the winner of each game:

$$0.5^{63} \doteq 1.08 \times 10^{-19}$$

- Probability of picking the winners of all first-round games, if we flip a fair coin to choose the winner of each game:

$$0.5^{32} \doteq 2.33 \times 10^{-10}$$

The seeds supplied by the NCAA Selection Committee give us some additional info that we might use to increase these probabilities.

March Madness, Round-2 appearances by seed, 1985-2019 (out of 140)

Seed	Round-2 appearances	1st-round win prob.
1	139	
2	132	
3	119	
4	111	
5	90	
6	88	
7	85	
8	68	
9	72	
10	55	
11	52	
12	50	
13	29	
14	21	
15	8	
16	1	

March Madness, Round-2 appearances by seed, 1985-2019 (out of 140)

Seed	Round-2 appearances	1st-round win prob.
1	139	0.993
2	132	0.943
3	119	0.850
4	111	0.793
5	90	0.643
6	88	0.629
7	85	0.607
8	68	0.486
9	72	
10	55	
11	52	
12	50	
13	29	
14	21	
15	8	
16	1	

Some more elementary probability calculations

Let us suppose that these empirical probabilities are the true probabilities of each seed winning in the first round. Then:

- Probability of picking the winners of all first-round games, if we always pick the higher-seeded team to win, is

$$\{[139 \times 132 \times 119 \times 111 \times 90 \times 88 \times 85 \times 68]/140^8\}^4 \doteq 0.000032$$

- This is $137,186$ times larger than if we use the coin-flipping strategy.
- Can we do even better by using something more refined than seeds to pick winners?
- Consider using a rating system for "team strength." Several proprietary rating systems exist (Kenpom, NET, Sagarin, Torvik).

Quantifying team strength

A well-known statistical approach for ranking and game prediction methods for sports teams is based on the following assumptions:

- The ith team's strength in a given season (t) can be represented by a parameter θ_{it}
- Game outcomes (difference in score, y_{ijk}) between teams i and j depend on their team strengths only via $\theta_{it} - \theta_{jt}$ (and on a home-court advantage parameter)

Then act as though

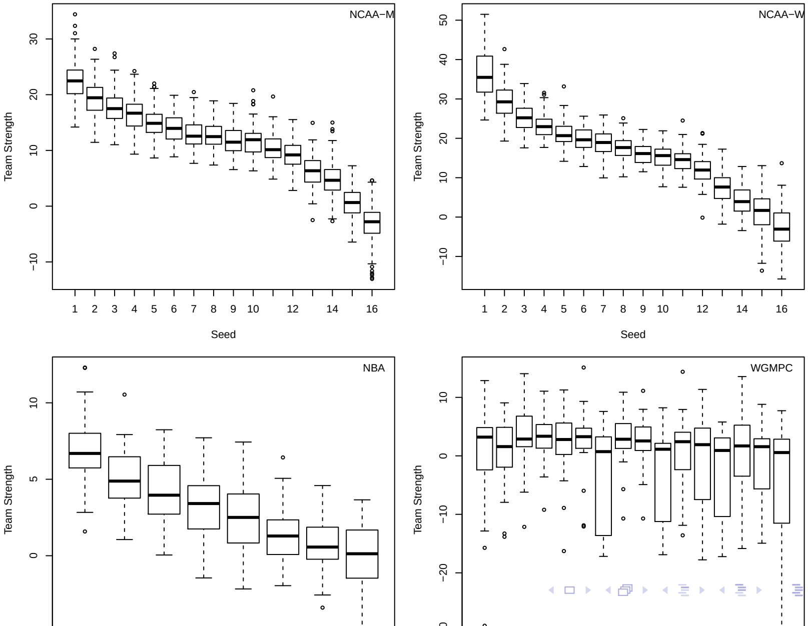
$$y_{ijk} = \begin{cases} H + \theta_i - \theta_j + e_{ijk} & \text{if } x_{ijk} = 1 \\ \theta_i - \theta_j + e_{ijk} & \text{if } x_{ijk} = 0 \end{cases}$$

where H is the home court advantage (for the given season), the e_{ijk}'s are uncorrelated random errors having mean 0 and common variance σ^2 (for that season), and $\sum_{i=1} \theta_i = 0$.

Quantifying team strength

- We can estimate the θ_{it}'s and the home-court advantage parameter by fitting this model using standard regression methodology.
- The estimates are very highly correlated with the Sagarin ratings ($r > 0.995$).
- We can also use these estimates to estimate the win probability of one of the teams in a given match-up.

Team strength by seed, 1985-2019



Estimating NCAA tournament game win probabilities from team strengths

$$
\begin{aligned}
P(\text{Team } i \text{ beats Team } j) &= P(y_{ijk} > 0) \\
&= P(\theta_i - \theta_j + e_{ijk} > 0) \\
&= P(e_{ijk} > \theta_j - \theta_i) \\
&= P\left(\frac{e_{ijk}}{\sigma} < \frac{\theta_i - \theta_j}{\sigma}\right) \\
&= \Phi\left(\frac{\theta_i - \theta_j}{\sigma}\right)
\end{aligned}
$$

where for the last two steps we added an assumption that the errors are normally distributed (Φ is the normal cdf).

Estimating NCAA tournament game win probabilities from team strengths, continued

This last quantity, though unknown, may be well-estimated by

$$\Phi\left(\frac{\hat{\theta}_i - \hat{\theta}_j}{\hat{\sigma}}\right)$$

Example — First-round match-up between Iowa State (5-seed) and Nevada (12-seed) in 2017:

$$\Phi\left(\frac{\hat{\theta}_{ISU,2017} - \hat{\theta}_{NEV,2017}}{\hat{\sigma}_{2017}}\right) = \Phi\left(\frac{20.051 - 9.826}{10.487}\right) = 0.835$$

Upsets

- Define an upset w.r.t. seed as a lower-seeded team beating a higher-seeded team. On average, there were 8.25 such upsets in the first round (out of 32 possible) from 1985-2019.
- Define an upset w.r.t. strength as a weaker team beating a higher-seeded team. On average, there were 6.4 such upsets in the first round from 1985-2019.
- Thus, using team strength, perhaps you can improve slightly upon a first-round strategy of picking only higher-seeded teams to win.
- It may still be beneficial to choose some upsets (of either kind) if you want to set your bracket apart from others in a pool.

March Madness, Round-2 and Sweet 16 appearances by seed, 1985-2019

Seed	Round-2 appearances	Sweet 16 appearances
1	139	120
2	132	89
3	119	74
4	111	66
5	90	47
6	88	42
7	85	27
8	68	13
9	72	7
10	55	23
11	52	22
12	50	21
13	29	6
14	21	2
15	8	1
16	1	0

The middle-seed anomaly

- Refers to the fact that 10-, 11-, and 12- seeds make it to the Sweet 16 much more than 8- and 9-seeds, and almost as often as 7-seeds.
- Largely due to 10-, 11-, and 12-seeds performing very well in the second round (relative to their team strengths).
- It suggests that it is not a bad strategy (especially to set your bracket apart from others) to ride, all the way to the Sweet 16, whichever 10-, 11-, and 12-seeds you pick to win their first-round games.
- The middle-seed anomaly disappears after the Sweet 16, so don't ride them any farther.