TREES, PROBABILITY AND PREDICTION

Steve Krevisky
Middlesex Community College
USA

In basic Statistics classes, we are often interested in "Tree Diagrams", which provide a visual way for our students to compute how many ways various events can occur. One special example of this isthe US National Collegiate Association of America (NCAA) Basketball tournament, which takes place in March of each year. Fans get caught up in " March Madness," and enjoy trying to predict the "Final Four". In this paper, we discuss many aspects of this tournament, including sharing of what the Tree diagram looks like, various probabilities of what different teams will do, and making predictions about what will happen in the First Round and beyond.

The NCAA Tournament consists of 4 Regions, with 16 teams in each region. The Regions are named East, South, Midwest and West. Within each region, the teams are seeded from 1 to 16, with number 1 being the highest seed (strongest team) and number 16 being the lowest seed (weakest team). The NCAA Selection Committee chooses and seeds teams for this tournament based upon their won-lost records, strength of schedule, opponents' strength of schedule, and so forth. All of this is put together in a mathematical formula called the RPI Rating. These selections can be quite controversial, and it's fun to speculate and *predict* who will win.

While the strongest teams (highest seeds) do have an advantage overall, in terms of the likelihood of winning, the First Round can produce many *upsets*. In each region, the number 1 seed plays the 16 seed, the 2 seed plays 15, 3 plays 14, and so forth. In what follows, I will present *probabilities* based upon 17 years of data, from 1985 to 2001 (2002 data to be presented at the conference).

We present some preliminary *probabilities* as follows:

1. Over 17 NCAA Tournaments, a number 1 seed has won the championship 10 times, so that P(# 1 seed winning the championship) = 10 / 17.
2. In March, prior to the start of the tournament, there is a final, pre-tournament poll of the top 25 teams. The number 1 ranked team in this poll rarely wins the tournament!! It has happened only 3 times! Hence, P(# 1 ranked team in final pre-tournament poll winning the championship) = 3 / 17.
3. Only twice in 17 tournaments has a seed lower than 4 won the championship. Thus, P(seed lower than 4 winning the championship) = 2 / 17.
4. Just 3 teams ranked out of the top 10 in the final pre-tournament poll have won the Championship. Therefore, P(team ranked out of top ten in poll winning the championship) = 3 / 17.

Next, while number 1 seeds have good winning chances, they often fail to make the Final Four, which consists of one team coming out of each of the 4 regions. Each region has a number 1 seed, so there are 4 number 1 seeds each year. In 17 years of data, at least one number 1 seed has failed to make the Final Four each year-no exceptions through 2001 !

We analyze this as follows:

5. P(all 4 # 1 seeds reaching the Final Four) = 0 / 17 = 0. In other words, this event has never happened!
6. P(3 of the 4 # 1 seeds reaching the Final Four) = 3 / 17. This has happened 3 times.
7. P(2 of the 4 # 1 seeds reaching the Final Four) = 7 / 17. This has happened 7 times.
8. P(1 of the 4 # 1 seeds reaching the Final Four) = 7 / 17. This seemingly unlikely event has happened 7 times!!
9. P(0 of the 4 # 1 seeds reaching the Final four) = 0 / 17 = 0. This event has never happened!

Low seeds, especially 10 or lower (11 to 16), rarely go deep into the tournament. Only once in 17 years has a double digit seed reached the Final Four (in 1986).

10. P(double-digit seed reaching the Final Four) = 1 / 17.

Much of the fun of the NCAA Tournament is predicting *first round upsets*, they happen every year! I define an upset as occurring when a lower seed beats a higher seed, provided that the difference between the seeds is at least 5. Therefore, 11 beating 6, 12 beating 5, and so forth, would classify as upsets. I view the 8 vs. 9 and 7 vs. 10 games as toss-ups, although if you include 10 beating 7, the upset potential jumps considerably. It's interesting to note that 9 seeds have a winning record vs. 8 seeds in first round play! There are 32 first round games. If we remove the four 8 vs. 9 games (one in each of 4 Regions), then there are 28 first round games with upset potential. We note the following:

11. P(10 seed beating a 7 seed) = 28 / 68, which is close to 50 % !

If you remove these 4 games, then there are 24 possible first round games which qualify as upsets by my definition above. My research shows that since 1985, when the NCAA Tournament went to its current format with 64 teams (actually, now there are 65 teams because of the play-in game, which started in 2001), at least one upset has occurred every year-no exceptions!! We analyze this as follows:

Over 17 years, there are four 11 vs. 6 games (one in each region), four 12 vs. 5 games, and so forth, so that there have been 68 of 11 against 6 games, 68 of 12 vs. 5 games, and so on, over these years.

12. P(11 beating a 6) = 21 / 68
13. P(12 beating a 5) = 20 / 68
14. P(13 beating a 4) = 14 / 68
15. P(14 beating a 3) = 13/ 68
16. P(15 beating a 2) = 4 / 68
17. P(16 beating a 1) = 0 / 68 = 0 (never happened in men's tournament).

In the 2001 tournament, there were 7 upsets, plus the 10 seeds won 2 of their 4 games. We further note that a 10, 11 and 12 seed all reached the Sweet 16, meaning that these seeds won also in the second round in order to advance as they did. We particularly note that the 12 seed has an interesting tendency to reach the Sweet 16, because 12 seeds have the following:

18. P(12 seed winning in second round) = 11 / 20, which is 55 % !

We might speculate that the lower seeds, who have won conference championships or Conference tournaments from so-called mid-level or lower level conferences, have a lot to prove in the first round against mid-level teams from the big-time conferences, who might not have been that good of a team, and who did not take their "lesser" opponent seriously. Therefore, when you or your students try to *predict* the outcome of a tournament of this nature, which is single-elimination (regardless of the sport), consider the *following*:

- At least one number 1 seed has failed to make the Final four each year!
- The team ranked number 1 in the final pre-tournament poll in March has rarely won the NCAA Championship!
- A 16 seed has never beaten a 1 seed in the 17 years of the tourney.
- There has been at least one upset every year, and sometimes as many as 7 in the first Round (even more if you include the 7-10 game)!
- Low seeds have reached the Sweet 16 by winning first and second round games, but only one low seed (a number 11) has reached the Final Four (in 1986)!

Therefore, you should look to predict a number of first round upsets, taking into account the teams, their conferences, how well they played in their last 10 games, how they did in their conference tournament, and so forth. Also, be prepared for many top Seeds to fall before the Final Four (many of these teams may have been overrated). You and your students should enjoy this, look for other patterns and possible predictions, and modify this for other sports and tournaments.

NOTES

1. Official 2002 NCAA Men's Final 4 Tournament Record Book, by Gary Johnson. Published by the NCAA.
2. College Basketball Magazine, 2002. Published by the Sporting News.