In this department *The American Statistician* publishes articles, reviews, and notes of interest to teachers of the first mathematical statistics course and of applied statistics courses. The department includes the Accent on Teaching Materials section; suitable contents for the section are described under the section heading. Articles and notes for the department, but not intended specifically for the section, should be useful to a substantial number of teachers of the indicated types of courses or should have the potential for fundamentally affecting the way in which a course is taught.

Probability Models for the NCAA Regional Basketball Tournaments

NEIL C. SCHWERTMAN, THOMAS A. McCREADY, and LESLEY HOWARD*

The study of mathematics can be motivated and enhanced when the concepts are reinforced with interesting and timely illustrations. Athletics in general and postseason competition in particular afford such opportunities to demonstrate, in the classroom, the application of probabilistic concepts. With the general strong interest in athletics it is not surprising that such competitions have been analyzed statistically. Mosteller (1952), for example, analyzed the baseball world series competition by studying the effectiveness of a seven-game series (first team to win four) in deciding which team is stronger as measured by each team's probability of winning any one of the games. Searls (1963) investigated other postseason competitions by considering four types of tournaments, namely, single elimination, single elimination with replication (first team to win two), double elimination, and double elimination with replication. Another athletic competition analysis was done by Moser (1982), who investigated the effects of post position in the game of Jai Alai. The professional ice hockey tournament (Stanley Cup) was the focus of a paper by Monahan and Berger (1977) in which three pairings of the eight competing teams were considered: (a) the most commonly used pairings, (b) maximizing the probability that the strongest team wins, and (c) maximizing the probability that the strongest two teams meet in the finals. The point rating of each team is incorporated in the calculation of the probability of winning any game. A more recent study by Schwertman and Howard (1990) used a similar scheme to develop a probabilistic analysis of the Australian Football League Grand Final Series competition. David (1959) and Glenn (1960) provided other interesting studies of athletic competition.

Each spring the National Collegiate Athletic Association (NCAA) organizes four regional collegiate basketball tournaments as the first of two steps in the determination of a "national champion" collegiate basketball team. The tournaments in each region are planned so that in the first round or series of games the strongest of the 16 teams plays the weakest, the second strongest plays the second weakest, etcetera. Then, in the second set of games, the winners of the first set of games play each other and this continues until the regional champion is decided. The pattern of the tournament is predetermined before the tournament begins in that pairings of the winners are already decided. Figure 1 describes the pattern currently in use with (1) indicating the team evaluated as the strongest, (2) the second strongest, etcetera, and (16) indicating the weakest.

First, we will determine the teams that must be played for a team to become the regional champion. In Round 1 (Games 1–8, see Fig. 1) there are 2^8 possible sets of winners. In Round 2 (Games 9–12, see Fig. 1) there are 2^4 possible sets of winners, in Round 3 (Games 13 and 14) there are 2^2 possible sets of winners, and in Game 15 there are two possible outcomes. Hence $2^8 \cdot 2^4 \cdot 2^2 \cdot 2 = 2^{15}$ possible outcomes for the 15 games. To compute the probability of any team, say the strongest team, (1), winning the tournament, we must consider all possible opponents in the games they must win. For Team (1) to be regional champions they must win in Games 1, 9, 13, and 15. Their possible opponents in these games are [(16)], [(8), (9)], [(4), (13), (5), (12)], and [(2), (15), (7), (10), (3), (14), (16), (11)], respectively. We need consider only $1 \cdot (2) \cdot (2)^2 \cdot (2)^3 \cdot = 2^6$ sets of opponents that Team (1) must defeat. For example, one possible sequence of opponents is (16), (8), (5), (3). Let $P_k(i, j)$ be the probability that seed (team) i defeats seed (team) j in the kth game from Figure 1. Then the probability that Team (1) wins by this path is $P_1(1, 16) \cdot P_9(1, 8)P(\text{Team (8) plays in Game 9}) \cdot P_{13}(1, 5)P(\text{Team (5) plays in Game 13})P_{15}(1, 3)P(\text{Team (3) plays in Game 15})$.

$$P(\text{Team (8) plays in Game 9}) = P_2(8, 9)$$

$$P(\text{Team (5) plays in Game 13})$$
$$= P_4(5, 12)[P_{10}(5, 4)P_3(4, 13) + P_{10}(5, 13)\,P_3(13, 4)]$$

$$P(\text{Team (3) plays in Game 15})$$
$$= P_7(3, 14)[P_{12}(3, 6)P_8(6, 11) + P_{12}(3, 11)\,P_8(11, 6)]$$
$$\{P_{14}(3, 2)\,P_5(2, 15)[P_{11}(2, 7)\,P_6(7, 10)$$

*Neil C. Schwertman is Professor of Statistics and Thomas A. McCready is Professor and Chairman, both in the Department of Mathematics and Statistics, California State University, Chico, CA 95929. Lesley Howard is Tutor, Department of Mathematics and Computing, Deakin University, Victoria 3217, Australia.

$$
\begin{aligned}
&+ P_{11}(2, 10)\, P_6(10, 7)] \\
&+ P_{14}(3, 15)\, P_5(15, 2)[P_{11}(15, 7)\, P_6(7, 10) \\
&+ P_{11}(15, 10)\, P_6(10, 7)] \\
&+ P_{14}(3, 7)\, P_6(7, 10)[P_{11}(7, 2)\, P_5(2, 15) \\
&+ P_{11}(7, 15)\, P_5(15, 2)] \\
&+ P_{14}(3, 10) P_6(10, 7)[P_{11}(10, 2)\, P_5(2, 15) \\
&+ P_{11}(10, 15)\, P_5(15, 2)]\}.
\end{aligned}
$$

The preceding computation is for 1 of 64 different sets of opponents Team (1) might have to play to win the regional tournament. Clearly, such computation by hand would be quite tedious. However, a fairly simple computer program can be written to do the computation.

Calculating the probability of any team winning the regional tournament requires the evaluation of the probability for each game played. Obviously, many factors influence the outcome of a particular game and hence the accuracy of the probabilities. To compute the probability of each of the teams eventually winning the regional tournament, it is necessary to make the same simplifying assumption made by most of the other authors cited—that the games are independent. Intangible effects that can affect the independence of each game are very difficult to measure and, consequently, we will assume independence of each contest. Although the independence assumption may alter the exact probabilities somewhat, we nevertheless still have useful models that should provide a reasonable approximation to the actual probabilities as well as a practical and interesting probability exercise.

The simplest model for describing the probability of a team winning the regional tournament would be to further assume that each team has an equal chance of winning any contest. Then the probability would be the same for each of the 16 teams; that is, the probability of winning the regional tournament would be 1/16 for all teams.

There are, however, 64 teams participating, and the tournament organizers attempt to ensure that no region has weaker teams than the others. Therefore, one would expect a substantial range in the overall quality of the teams. Instead of assuming that in each game each team has an equal chance of winning, a more realistic probability of the outcome should incorporate the relative strength of the teams.

One technique for including the relative strength of the teams in the computation of $P_k(i, j)$ is to define some function of $P(i)$ and $P(j)$ where $P(i)$ and $P(j)$ are the proportion of wins for the season for the ith and jth teams, respectively. This approach assumes that the two teams play equally rigorous schedules, which is generally not the case. For example, in recent years the Big East and Atlantic Coast conferences have been exceptionally strong, with many excellent teams, and hence even the best teams in the conference may have several losses. These teams, if playing in most other conferences, would have nearly perfect seasons. An alternative method for determining the probability for each contest, which is not directly influenced by the difficulty of the schedule, would be to use a "Delphi" approach, employing a panel of experts who determine the relative strength. Fortunately, this is done when the NCAA organizes the tournament. Their panel of experts not only selects the teams but decides the pairings (seeding) based on the consensus of their relative strength.

We will consider three probability models that incorporate

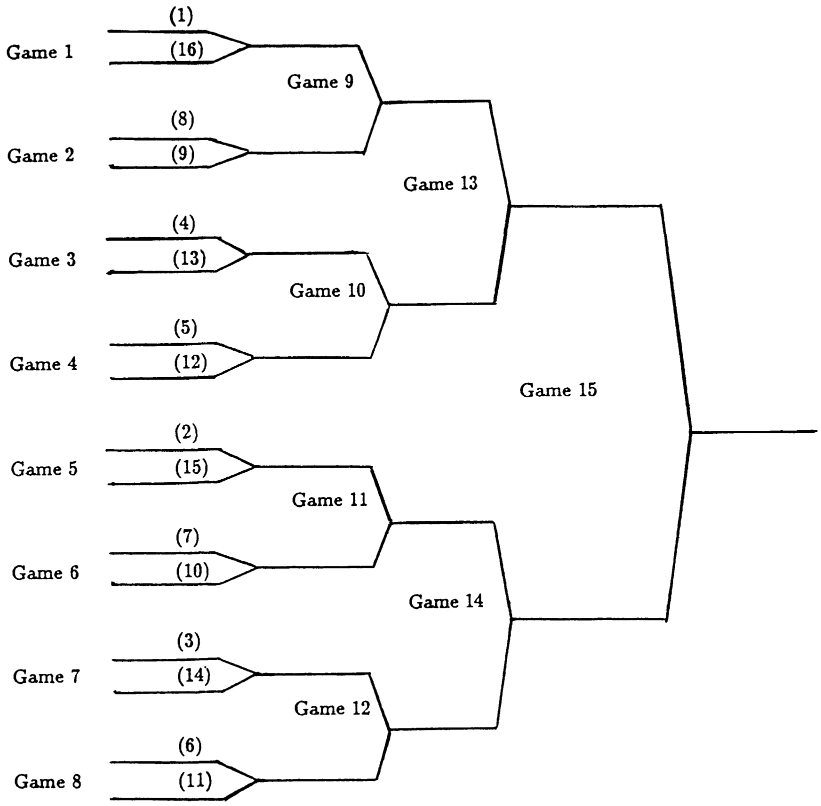

Figure 1. NCAA Regional Basketball Tournament Pairings.

Table 1. *Probabilities of Winning the NCAA Regional Basketball Tournament*

Seed position	Model 1, $P(i, j) = j/(i+j)$	Model 2, $P(i, j) = .5 + 1/32(j-i)$	Model 3, $P(i, j) = .5 + .2813625(S(i) - S(j))$
1	.5192184	.274773059	.458828968
2	.2160397	.208339722	.188134282
3	.1067895	.154288031	.110316885
4	.056686348	.111044558	.067982598
5	.033561502	.080092151	.046971729
6	.021798624	.057816726	.036253863
7	.014006608	.039791195	.025971459
8	.0087412084	.025514550	.015486031
9	.0061364875	.017127023	.011373606
10	.0048488911	.012358736	.011278378
11	.0036734916	.008261224	.009039390
12	.002667062	.005023861	.006028946
13	.0020387262	.002952291	.004539589
14	.0016381438	.001666639	.004046039
15	.0012518217	.000759986	.002837565
16	.00090356213	.000190245	.000910673

the relative strengths of the teams as measured by their seeding by the tournament committee. The first is a simple method that computes the probability of winning in a particular game as the opponents' seeding divided by the sum of its own seeding plus that of the opponents; that is, $P(i, j) = j/(i + j)$. For example, in Game 1 (see Fig. 1) between the strongest Team (1) and the weakest Team (16), the probability of the strongest winning is (16)/[(1) + (16)] = 16/17. It may be of interest to note that, if $P(i, j)$ were placed in a 16 × 16 matrix P with i, jth element $P(i, j)$, then P is the matrix product of the finite Hilbert matrix H with D a diagonal matrix where ith diagonal element $d_{ii} = i$. Thus the i, jth element of the Hilbert matrix, $h_{ij} = 1/(i + j)$, and $P = H \cdot D$. [See Kato (1957, pp. 73–81) for more details about the Hilbert matrix.] Tabulation of the probability for each team using this model is given in Table 1.

The previous method yields probabilities for each contest that heavily favor the number one seed. For example, $P(1, 2) = 2/3$ and $P(2, 1) = 1/3$; that is, seed 1 is twice as likely to defeat seed 2 as is the reverse. But $P(12, 13) = 13/25$ and $P(13, 12) = 12/25$, yet in both cases the teams are seeded in consecutive order.

The second model assumes linearity based on the difference in the seeding of the two teams. That is,

$$P(i, j) = .5 + \Delta(j - i), \qquad (1)$$

where $1 \le i \le 16$, $1 \le j \le 16$, and $0 \le \Delta \le 1/30$. This model assures that $0 \le P(i, j) \le 1$ and that $P(i, j) + P(j, i) = 1$. It remains to determine an appropriate Δ.

For the NCAA regional tournament with 16 teams, the range of probabilities is $.5 - 15\Delta \le P(i, j) \le .5 + 15\Delta$. Clearly, for smaller Δ the range of probabilities spans a smaller interval, and for $\Delta = 1/30$ the first seed defeats the sixteenth seed with near certainty. It seems reasonable to assume that teams have, roughly, a uniform distribution in strengths and that the resulting probabilities have some equally spaced pattern on the interval [0, 1]. If we further assume that indeed there is some positive probability that the sixteenth seed defeats the first seed ($P(16, 1) > 0$) while maintaining the linearity or equal spacing pattern of the probability, it seems reasonable to assign $P(16, 1) = \Delta$, $P(16, 2) = 2\Delta$, etcetera. Then $P(16, 1) = .5 - 15\Delta = \Delta$ or $\Delta = 1/32$. This value of Δ maintains the uniformity in spacing over the interval [0, 1] while permitting a positive probability of each team in each contest beating the other. The tabulation of the probabilities of winning the regional tournament for each seed using this model and Δ are given in Table 1.

The third model is similar to the second except, instead of a uniform distribution of strengths, the strengths of all 292 Division I basketball teams are assumed to be normally distributed. It is further assumed that the 64 teams that are selected to compete in the four regional tournaments are the 64 strongest of the 292 and the first seeds in the four regional tournaments are the 4 strongest teams, the next seeds are the next group of 4 strongest teams, and so forth. Figure 2 shows the assumed distribution of team strengths with the top 22% selected for the tournament competition.

Observe that, generally, the teams selected for the fourteenth, fifteenth, and sixteenth seed positions (to the left in the shaded area) are much closer in relative strength than those teams selected for the first, second, or third seed positions (to the right in the shaded area). It then follows that matches between teams with higher seed numbers (13, 14, 15, or 16) should be more even and have probabilities closer to .5 than matches between teams with lower seed numbers (1, 2, 3, 4). This phenomenon is observed in the first model but not in the second.

We now use the method of normal scoring to establish the strength of the teams. The strength of the first seed is the z score for the percentile for the median of the top group

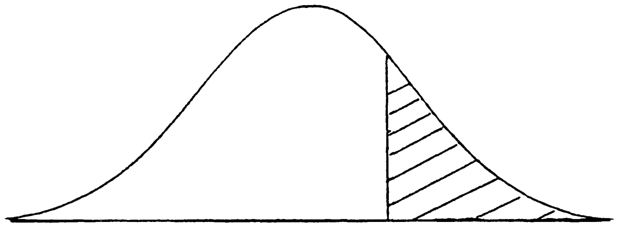

Figure 2. Distribution of Strengths of Division I Basketball Teams for Model III.

Table 2. Goodness-of-Fit Analysis

Group (seed no.)	Observed	Expected numbers		
		Model I	Model II	Model III
1	10	12.461242	6.5945542	11.011896
2	5	5.1849528	5.0001528	4.5152232
3	3	2.562948	3.702912	2.6476056
4 or more	6	3.7908572	8.7023832	5.8252752
		$\chi^2_{(3)}$ = 1.854641	$\chi^2_{(3)}$ = 2.73119	$\chi^2_{(3)}$ = .197177
	P values	.603	.435	.978

of four. For example, for the first seed (Teams 289, 290, 291, 292, in order of strength) the percentile is (using continuity correction for $N = 292$) 290.5/292.5 = 99.316% with corresponding z score 2.466. Similarly, for the second seed (second group of four), 286.5/292.5 = 97.9487% with z score 2.044, etcetera. The z score measures of strength then are 2.466, 2.044, 1.823, 1.666, 1.542, 1.438, 1.348, 1.267, 1.194, 1.127, 1.064, 1.006, .951, .898, .848, and .800 for seed positions 1 to 16, respectively. To incorporate these measures of relative strength into probabilities for all possible contestants, we modify the procedure provided by Equation (1) by defining $P(i, j) = .5 + \Delta(S(i) - S(j))$, where $S(i)$ and $S(j)$ are the measures of relative strength provided by the normal scoring procedure. If this model is to span the same range of probabilities as probability model 2, then $P(1, 16) = 31/32 = .5 + \Delta(2.466 - .800)$ and $\Delta = .2813625$. The tabulation of the probabilities of each seed winning the regional tournament using this model and Δ are given in Table 1.

The goodness-of-fit test for the model was done using the usual chi-squared statistic. Actual data for the six years that NCAA has used this format were obtained from the NCAA office. In six years, 10 first seeds, 5 second seeds, 3 third seeds, 2 fourth seeds, 2 sixth seeds, 1 eighth seed, and 1 eleventh seed have won their region. The seeds were partitioned into four groups: (1), (2), (3), (4 or more). Table 2 contains the observed and expected numbers for this partition and the computed chi-squared value with 3 df and the p values. Although Models I and III provided similar probabilities, those from Model II were quite different. The chi-squared test, however, failed to reject any of the models as being unsuitable at the customary significance levels. This is not surprising, since with such small samples goodness-of-fit tests generally have little power. The chi-squared value, however, does provide some indication of the relative adequacy of the three models. Model III, ($\chi^2_{(3)} = .197177$), had the smallest chi-squared value and, therefore, was the model most compatible with the empirical data.

The NCAA regional basketball tournament provides an interesting opportunity to apply probabilistic concepts. The probabilistic Model III seems to provide an exceptionally fine fit to the actual data and should be of interest to many students without requiring a strong mathematical or statistical background. These probability models illustrate the use of the multiplication principle in computing the probability of each path to the regional championship, as well as the additive property of mutually exclusive events (paths). The general interest in the NCAA basketball tournament provides ample motivation to the student for this recreational and educational probability exercise.

[*Received January 1990. Revised May 1990.*]

REFERENCES

David, H. A. (1959), "Tournaments and Paired Comparisons," *Biometrika*, 46, 139–49.

Glenn, W. A. (1960), "A Comparison of the Effectiveness of Tournaments," *Biometrika*, 47, 253–262.

Kato, T. (1957), "On the Hilbert Matrix," *Proceedings of the American Mathematical Society*, 8, 73–81.

Monahan, J. P., and Berger, P. D. (1977), "Playoff Structures in the National Hockey League," in *Optimal Strategies in Sports*, eds. S. P. Ladany and R. E. Machol, Amsterdam: North-Holland, pp. 123–128.

Moser, L. E. (1982), "A Mathematical Analysis of the Game of Jai Alai," *The American Mathematical Monthly*, 89, 292–300.

Mosteller, F. (1952), "The World Series Competition," *Journal of the American Statistical Association*, 47, 355–380.

Schwertman, N. C., and Howard, L. (1990), "Probability Models for the Australian Football League GRAND FINAL SERIES," *The American Mathematical Society Gazette*, 17(4), 89–94.

Searls, D. T. (1963), "On the Probability of Winning With Different Tournament Procedures," *Journal of the American Statistical Association*, 58, 1064–1081.