A JUSTIFICATION OF THE $\log 5$ RULE FOR WINNING PERCENTAGES

STEVEN J. MILLER

Abstract. Let p and q denote the winning percentages of teams A and B. The following formula has numerically been observed to provide a terrific estimate of the probability that A beats B: $(p-pq)/(p+q-2pq)$. In this note we provide a justification for this observation.

1. Introduction

In 1981, Bill James introduced the $\log 5$ method to estimate the probability that team A beats team B, given that A wins $p\%$ of its games and B wins $q\%$ of theirs. He estimates this probability as

$$\frac{p - pq}{p + q - 2pq}. \tag{1.1}$$

See [?, Ti] for some additional remarks. This formula has many nice properties:

(1) The probability A beats B plus the probability B beats A adds to 1.
(2) If $p = q$ then the probability A beats B is 50%.
(3) If $p = 1$ and $q \neq 0, 1$ then A always beats B.
(4) If $p = 0$ and $q \neq 0, 1$ then A always loses to B.
(5) If $p > 1/2$ and $q < 1/2$ then the probability A beats B is greater than p.
(6) If $q = 1/2$ then the probability A wins is p (and similarly if $p = 1/2$ then B wins with probability q).

In the next section we provide a justification for this estimate.

2. Justification of the $\log 5$ method

When we say A has a winning percentage of p, we mean that if A were to play an average team many times, then A would win about $p\%$ of the games (for us, an average team is one whose winning percentage is .500). Let us image a third team, say C, with a .500 winning percentage. We image A and C playing as follows. We randomly choose either 0 or 1 for each team; if one team has a higher number then they win, and if both numbers are the same then we choose again (and continue indefinitely until one team has a higher number than the other). For A we choose 1 with probability p and 0 with probability $1 - p$, while for C we choose 1 and 0 with probability $1/2$. It is easy to see that this method yields A beating C exactly $p\%$ of the time.

Date: April 20, 2008.
2000 *Mathematics Subject Classification.* 46N30 (primary), 62F03, 62P99 (secondary).
Key words and phrases. sabermetrics, $\log 5$ rule.

The probability that A wins the first time we choose numbers is $p\cdot 1/2$ (the only way A wins is if we choose 1 for A and 0 for C, and the probability this happens is just $p \cdot 1/2$). If A were to win on the second iteration then we must have either chosen two 1's initially (which happens with probability $p\cdot 1/2$) or two 0's initially (which happens with probability $(1-p)\cdot 1/2$), and then we must choose 1 for A and 0 for B (which happens with probability $p \cdot 1/2$. Continuing this process, we see that the probability A wins on the n^{th} iteration is

$$\left(p\cdot\frac{1}{2}+(1-p)\cdot\frac{1}{2}\right)^{n-1}\cdot\left(p\cdot\frac{1}{2}\right) \ = \ \frac{p}{2^n}. \tag{2.1}$$

Summing these probabilities gives a geometric series:

$$\sum_{n=1}^{\infty}\frac{p}{2^n} \ = \ p, \tag{2.2}$$

proving the claim.

Imagine now that A and B are playing. We choose 1 for A with probability p and 0 with probability $1-p$, while for B we choose 1 with probability q and 0 with probability $1-q$. If in any iteration one of the teams has a higher number then the other, we declare that team the winner; if not, we randomly choose numbers for the teams until one has a higher number.

The probability A wins on the first iteration is $p\cdot(1-q)$ (the probability that A is 1 and B is 0). The probability that A neither wins or loses on the first iteration is $(1-p)(1-q)+pq = 1-p-q+2pq$ (the first factor is the probability we chose 0 twice, while the second is the probability we chose 1 twice). Thus the probability A wins on the second iteration is just $(1-p-q+2pq)\cdot p(1-q)$; see Figure 1.

Continuing this argument, the probability A wins on the n^{th} iteration is just

$$(1-p-q+2pq)^{n-1}\cdot p(1-q). \tag{2.3}$$

Summing[1] we find the probability A wins is just

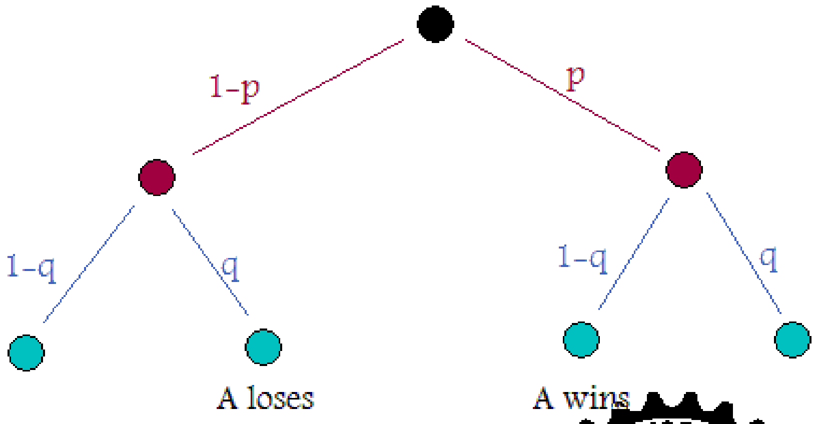
$$\begin{aligned}\sum_{n=1}^{\infty}(1-p-q+2pq)^{n-1}\cdot p(1-q) \ &= \ p(1-q)\sum_{n=0}^{\infty}(1-p-q+2pq)^n \\ &= \ \frac{p(1-q)}{1-(1-p-q+2pq)} \\ &= \ \frac{p(1-q)}{p+q-2pq}. \end{aligned}\tag{2.4}$$

It is illuminating to write the denominator as $p(1-q)+q(1-p)$, and thus the formula becomes

$$\frac{p(1-q)}{p(1-q)+q(1-p)}. \tag{2.5}$$

[1]To use the geometric series formula, we need to know that the ratio is less than 1 in absolute value. Note $1-p-q+2pq = 1-p(1-q)-q(1-p)$. This is clearly less than 1 in absolute value (as long as p and q are not 0 or 1). We thus just need to make sure it is greater than -1. But $1-p(1-q)-q(1-p) > 1-(1-q)-(1-p) = p+q-1 > -1$. Thus we may safely use the geometric series formula.

FIGURE 1. Probability tree for A beats B in one iteration.



This variant makes the extreme cases more apparent. Further, there are only two ways the process can terminate after one iteration: A wins (which happens with probability $p(1-q)$) or B wins (which happens with probability $(1-p)q$). Thus this formula is the probability that A won given that the game was decided in just one iteration.

REFERENCES

[Al] Baseball Almanac, `http://baseball-almanac.com`.

[Fe1] W. Feller, *An Introduction to Probability Theory and its Applications, Vol. I.*, third edition. Wiley, New York 1968.

[Fe2] W. Feller, *An Introduction to Probability Theory and its Applications, Vol. II.*, third edition, Wiley, New York 1971.

[Ja] B. James, *Baseball Abstract 1983*, Ballantine, 238 pages.

[Ti] T. Tippet, *May the best team win... at least some of the time.* `http://www.diamond-mind.com/articles/playoff2002.htm`.

DEPARTMENT OF MATHEMATICS, BROWN UNIVERSITY, 151 THAYER STREET, PROVIDENCE, RI 02912

E-mail address: `sjmiller@math.brown.edu`