Math Club Talk: Sabermetrics

I. Introduction

1. Definition of sabermetrics:
   - "The study of baseball statistics." *Wolfram Mathworld*
   - "The search for objective knowledge about baseball." *Bill James*
   - "I do not start with numbers any more than a mechanic starts with a monkey wrench. I start with the game, with the things that I see there and the things that people say there. And I ask: is it true? Can you validate it? Can you measure it? How does it fit with the rest of the machinery? And for those answers I go to the record books. . . . What is remarkable to me is that I have so little company. Baseball keeps copious records, and people talk about them and argue about them and think about them a great deal. Why doesn't anybody use them? What doesn't anybody say, in the face of this contention or that one, 'Prove it' "? *Moneyball* pg. 75
2. Bill James: changing the face of the game
   - Two of James' most notable contributions: Pythagorean W-L and log5 (will look at log5 later)
   - Revolutionized baseball with questioning, seen in this quotation about HOFer Jim Rice: "Virtually all sportswriters, I suppose, believe that Jim Rice is an outstanding player. If you ask them how they know this, they'll tell you that they just know; I've seen him play. That's the difference in a nutshell between knowledge and bullshit; knowledge is something that can be objectively demonstrated to be true, and bullshit is something that you just 'know.' If someone can actually demonstrate that Jim Rice is a great ballplayer, I'd be most interested to see the evidence." *Bill James*
   - This gave a motivation for sabermetrics: the numbers would reveal meaningful information that could not be observed by day to day observation
   - Teams now pay strict attention to the numbers, and general managers have moved from employing "baseball people" to those who had the knowledge and creativity to look to adopt these strategies
3. Uses of sabermetrics

– Assessing past performance

– Predicting future performance

– Finding skills undervalued by the market (like Billy Beane and the Oakland A's in *Moneyball*)

II. Basic Tools of Sabermetrics

1. The old "slashline stats": batting average (AVG), Home runs (HR), Runs Batted In (RBI)

$$\text{Batting Average} = \frac{\#\text{ hits}}{\#\text{ at bats}}$$

2. The new "slashline": Batting average, On Base Percentage (OBP), Slugging Percentage (SLG)
   - On Base Percentage: "measure of how often you don't make an out", league average is approximately .330

$$\text{On Base Percentage} = \frac{\#\text{ hits} + \#\text{ walks} + \#\text{ hit-by-pitches}}{\#\text{ plate appearances}}$$

   - Slugging Percentage: "measure of how hard and far you hit the ball", league average is approximately .420

$$\text{Slugging Percentage} = \frac{\#\text{ total bases}}{\#\text{ at bats}}$$

2. Example: In 2009 AL MVP Joe Mauer hit .365/.444/.587, compared to the league average of .265/.330/.420. That's pretty good.

III. Moneyball

1. While the rest of baseball was still obsessed with AVG, HR, and RBIs, the Oakland A's focused on OBP:

   "Anything that increases that offense's chances of making an out is bad; anything the decreases it is good. And what is on base percentage? Simply yet exactly put, it is the probability that the batter will not make an out. When we state it that way, it becomes, or should become, crystal clear that the most important isolated offensive statistic is the on base percentage. It measures the probability that the batter will not be another step toward the end of the inning." *Moneyball* pg. 58

2. The fact that the league undervalued OBP so much made it that much more valuable for the small market Oakland A's. A typical year from 1999 to 2006: the Yankees payroll was $200 million, and the A's payroll was$40 million. Over that same time, the Yankees record was 777-515 (.601), and the A's record was 751-544 (.580). So, despite the fact that the Yankees spent approximately $1.2 billion more over that span, they only won 16 more games than the A's.

3. A notable example:

	Old Slashline (AVG/HR/RBI)	New Slashline (AVG/OBP/SLG)
Eric Chavez	.241/22/72	.241/.351/.435
Alex Rodriguez	.290/35/121	.290/.392/.523

   The value in Eric Chavez is revealed in his new slashline: while his batting average is fairly low, his on base percentage (which is much more important), is above league average. In short, Eric Chavez is much more valuable to a team like the A's because he is undervalued, and thus makes a small fraction of Rodriguez's $25.2 million.

4. The success of the A's with the *Moneyball*-strategy would not last forever: since 2006 they are 226-259 (.466). They are still looking, it seems, for another undervalued stat.

IV. Pythagorean W-L Record

1. The Pythagorean W-L record is the expected value for a teams winning percentage, discovered by Bill James:

$$\text{Expected Winning \%} = \frac{\text{Runs Scored}^2}{\text{Runs Scored}^2 + \text{Runs Allowed}^2}$$

2. Bill James derived this formula as a result of trial and error and experimentation. However, Professor Stephen Miller of Williams College showed that, under reasonable statistical assumptions, the formula can be shown to follow mathematically with a 3-parameter Weibull distribution.

3. How is this useful? Both for assessing past performance and prediction of future performance.

4. Example: the 2007 Seattle Mariners were 88-74, with an expected record of 79-83 (RS= 794, RA= 813). They ended 2008 with a record of 61-101.

5. Example: the 2009 Atlanta Braves were 86-76, with an expected record of 91-71 (RS= 735, RA= 641). Had they actually finished 91-71, they would have made the playoffs.

6. Applications to other sports only involves changing the exponent:
   baseball: a more accurate exponent is 1.82
   basketball: 14
   football: 2.37
   hockey: 1.86

V. The Log5 Method

1. Another discovery by Bill James, the probability that team A beats team B is

$$\frac{a(1-b)}{a(1-b)+b(1-a)}$$

where a and b are the respective winning percentages of teams A and B.

2. Theoretical justification: this is a special case of Bayes' Theorem from probability. There is, however, another theoretical justification by simulation (also devised by Stephen Miller).

   Choose either 1 or 0 for both teams A and B:
   – 1 with probability a, b (respectively)
   – 0 with probability $1-a$, $1-b$ (respectively)

   The team with the larger number wins; if tied then replay until there is a winner.

P(team A wins the ith iteration)=

1st: $a(1-b)$
2nd: $\big((1-a)(1-b)+ab\big)\big(a(1-b)\big)$
.
.
.
nth: $(1-a-b+2ab)^{n-1}\big(a(1-b)\big)$

To find the probability that A wins, we sum over all possible iterations:

$$\begin{aligned} P(A \text{ wins}) &= \sum_{n=1}^{\infty} (1-a-b+2ab)^{n-1}\big(a(1-b)\big) \\ &= a(1-b)\sum_{n=1}^{\infty} (1-a-b+2ab)^{n-1} \\ &= \frac{a(1-b)}{1-(1-a-b+2ab)} \qquad (**) \\ &= \frac{a(1-b)}{a(1-b)+b(1-a)} \end{aligned}$$

$(**)$ It is easy to check that $|1-a-b+2ab| < 1$

□