The Probability of a Statistical Oddity

David E Dobbs

Introduction: A Paradoxical Loss

The accompanying table summarizes the batting performances of four Major League Baseball players during a recent season.

After the first half of the season, the race for the batting average title was not close. The preseason favourite, Player A (with a batting average of 0.295 after 81 games), trailed a journeyman, Player B, who had a batting average of 0.320 after 81 games. Also contending was Player C, whose batting average after 81 games was 0.294. After midseason, Player A took some time off to nurse an old injury. Following some rest and recuperation, Player A returned for the last 23 games of the season, amassing a second-half batting average of 0.333 and an overall batting average of 0.3025 (rounded off as 0.303) for the entire season. When the results came in, Player A had overcome the lead that Player B had enjoyed at midseason. Player A reasoned, "Yes, Player B beat me during the first half of the season, but I beat him more decisively during the second half of the season." In addition, Player A was not surprised that he had beaten Player C. Player A reasoned, "After all, I beat Player C in each half of the season, so of course I beat him overall for the entire season." Player A prepared to be awarded yet another batting title. But it was not to be.

The actual winner of the batting title was Player D. This young professional had been brought up too quickly from the minor leagues at the beginning of the season. Because Player D had a batting average of only 0.240 after 23 games, he was sent back to the minor leagues until midseason so that he could work on hitting a curve ball. Player D returned to the major leagues after the 81 game of the season. Hitting lead-off, he compiled 325 at-bats and a batting average of 0.320 for the second half of the season. This effort produced a batting average of 0.305 for Player D for the entire season.

Player A was dismayed and confused. How, he wondered, could he have lost the batting title to the unheralded Player D? Player A reasoned, "I beat Player D in each half of the season, so how and when did he end up beating me overall for the entire season? What are the chances that this kind of paradoxical loss could happen?"

In this article, I answer the questions raised by Player A. By doing so, I also provide enrichment material for mathematics classes at various levels. Table 1 can be used to help beginning students practise

Table 1

Player A	**At-Bats**	**Hits**	**Batting Average**
First half of the season	325	96	0.295(38...)
Second half of the season	75	25	0.333(3...)
Totals for the entire season	400	121	0.3025 (recorded as 0.303)
Player B	**At-Bats**	**Hits**	**Batting Average**
First half of the season	200	64	0.320
Second half of the season	200	52	0.260
Totals for the entire season	400	116	0.290
Player C	**At-Bats**	**Hits**	**Batting Average**
First half of the season	160	47	0.294
Second half of the season	240	67	0.279(16...)
Totals for the entire season	400	114	0.285
Player D	**At-Bats**	**Hits**	**Batting Average**
First half of the season	75	18	0.240
Second half of the season	325	104	0.320
Totals for the entire season	400	122	0.305

calculating averages (technically called *means* in statistics). This activity also points out how a player's overall batting average is weighted toward the batting average that he obtained during the half-season in which he had the greater number of at-bats. This activity also gives students practical experience with rounding off to three decimal places. It is important to realize that someone's batting average for the entire season is not simply the arithmetic mean of his batting averages for the two half-seasons. In effect, the table entry "Totals for the entire season" treats the season as a whole because the batting average for the entire season for each player (A, B, C or D) is calculated as [total number of hits]/400.

We require more than elementary algebra to answer the questions raised by Player A. He asked about the chances of a paradoxical loss (that is, the chance of losing the batting title to a player whom one had beaten in each half of the season). First, I made some reasonable assumptions, which are specified below. Then I found the chance (or probability) by using analytic geometry. I compared the area of a certain rectangle with the area lying between a horizontal line and a branch of a certain rectangular hyperbola. To identify the parameters involved, I used the SOLVER feature on a Texas Instruments (TI) graphing calculator. This part of the analysis would fit well in a precalculus class.

Finding the area between the line and the hyperbola depends on two mathematical matters. The first determines that the graph of the hyperbola is rising—that is, described by an increasing function. I give three proofs of this fact in Theorem I, and each proof is designed for an audience at a different level. Proof 1 can be given to an algebra class; it reinforces the long division (or synthetic division) of polynomials and the core meaning of fractions. Proof 2 involves inequalities, and thus could be given to a precalculus class. Finally, Proof 3 involves the sign of a derivative, and thus may be the method of choice for a calculus class.

The second mathematical matter involves calculating the area between the line and the hyperbola. This is the only part of the analysis that requires calculus. After this area was described by a definite integral, I calculated that integral by using the numerical integrator (fnInt) on a TI graphing calculator. In this way, the entire experience reinforces two important technological functions—SOLVER and fnInt—on a TI calculator.

I carried out the above analysis three times to answer several precise questions that are suggested by Player A's queries. The first result is that there is a probability of only 0.00635 (that is, 0.635 per cent) that Player A could, paradoxically, be beaten by a player (like Player D) with a first half-season batting average of only 0.240 and a second half-season batting average of less than 0.333. The second result is that there is a probability of about 0.0368 (that is, 3.68 per cent) that Player A could, paradoxically, be beaten by a player (like Player C) with a first half-season batting average of 0.294 and a second half-season batting average of less than 0.333. Finally, I found that there was a probability of 0.3679 (that is, 36.79 per cent) that Player A could be beaten (not necessarily paradoxically) by a player with a first half-season batting average of 0.294.

To make the analysis more realistic, I only considered players who had between 50 and 350 at-bats in each half-season and (like Players A, B, C and D) at least 400 at-bats for the entire season. I also assumed that no player would have a batting average exceeding 0.500 during any half-season. Even more realistic analysis is possible, but at the cost of considering the calculus of functions of several variables and using computer technology to evaluate various multiple integrals. This is explained, along with some philosophical musing, in the closing comments. As a final consolation for Player A, I will also provide a theorem explaining that paradoxical losses are not possible for players with the same number of at-bats in each half-season.

The Probability of Losing Paradoxically as in the Above Example

Next, I determined the probable paradox that Player A will lose to a random player, such as Player G who (like Player D) has a batting average of 0.240 for the first half of the season and less than 0.333 for the second half of the season. The next table summarizes the batting performance of such a Player G.

Table 2

Player G	**At-Bats**	**Hits**	**Batting Average**
First half of the season	x	$0.240\,x$	0.240
Second half of the season	$400 - x$	$(400 - x)$	$r(<.333)$
Totals for the entire season	400	$.240x + (400 - x)r$	$.240x + (400 - x)r$ 400

Note that I indicated the randomness of Player G by letting x denote the number of at-bats for Player G during the first half of the season. Now, Player A will only lose to Player G if the batting average of Player G for the entire season exceeds that of Player A—that is, only if

$$\frac{.240x+(400-x)r}{400} > .3025.$$

An equivalent inequality is

$$r > \frac{121-.240x}{400-x}.$$

The modern graphical approach to solving inequalities requires that we understand the graph of the equation $r=\frac{121-.240x}{400-x}$ in the xr-plane. As will be shown in Proof 1 of Theorem 1, this equation in the xr-plane can also be expressed as $r=.240+\frac{25}{400-x}$ by long division, from which is obtained the equivalent equation $(x-400)(r-0.2400)=-25$. By translating the x- and r-axes 400 units to the right and 0.240 units upwards, a new x^*r^*-coordinate system is obtained with an origin at the point formerly known as (400, 0.240). The equation $(x-400)(r-0.2400)=-25$ can now be expressed as the equivalent equation $x^*r^*=-25$. Because the product of the variables is a nonzero constant, the graph is a rectangular hyperbola.

The above graph (of a hyperbola) is rising for the relevant values of x (namely, for $50 \le x \le 350$), and this is established in Theorem 1, Proof 1. As is explained in the introduction, the three proofs of Theorem 1 are designed for classes with varying backgrounds. (A fourth proof, which is more geometric, follows by noticing that for $c < 0$, the left-hand branch of a rectangular hyperbola $x^*r^* = c$ is rising.)

Theorem I

The function $r=f(x)=\frac{121-.240x}{400-x}=.240+\frac{25}{400-x}$ is increasing for $0 < x < 400$. More generally, if a, b and c are positive constants, then the function given by $y=a+\frac{b}{c-x}$ is increasing for $0 < x < c$.

Proof 1

Using long division (or synthetic division), observe that when $0.240x - 121$ is divided by $x - 400$, the quotient is 0.240 and the remainder is −25. Thus,

$$\frac{121-.240x}{400-x}=\frac{.240x-121}{x-400}=.240+\frac{25}{400-x}.$$

This expression takes the asserted form, with $a = 0.240$, $b = 25$ and $c = 400$. To prove that this expression is increasing, increase x (strictly between 0 and c, and note the effect). As x increases, $400 - x$ decreases (but remains positive)—and so the ratio $\frac{25}{400-x}$ of positive quantities increases, a situation that is unaltered by the addition of the constant 0.240. Thus, this first proof is based on the behaviour of a fraction that has a constant positive numerator and a varying positive denominator. (Students may have developed this understanding as a result of studying instances of inverse proportionality, such as Boyle's Law in chemistry.)

Proof 2

Using the rules for operating on inequalities, find that if $0 < x_1 < x_2 < c$, then $a+\frac{b}{c-x_1} < a+\frac{b}{c-x_2}$. (One could work directly with the original form of $f(x)$ as well. I leave the details of that similar proof to the reader.) Observe that $0 < c - x_2 < c - x_1$. Multiply this inequality with the positive number $\frac{b}{(c-x_2)(c-x_1)}$, to obtain $\frac{b}{c-x_1} < \frac{b}{c-x_2}$. The inequality is unaltered if the constant a is added to both sides, thus completing the second proof.

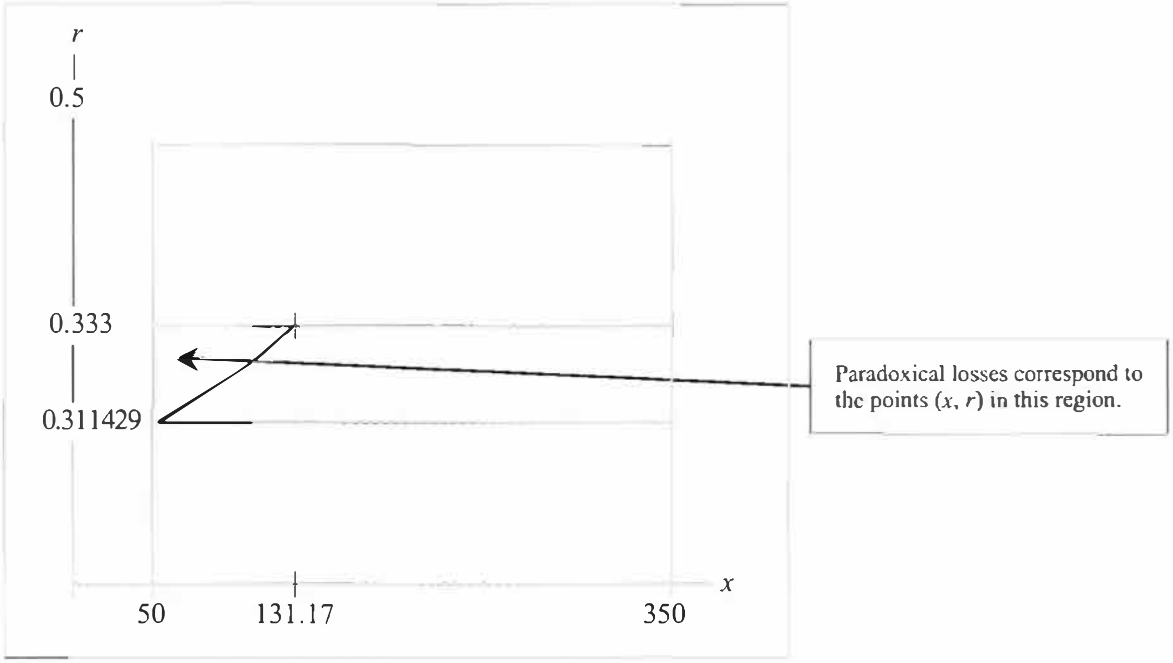
Proof 3

For this calculus-based proof, find that the function given by $y=a+\frac{b}{c-x}$ has a positive derivative. Using the formulas of different calculus, that derivative is found to be $\frac{b}{(c-x)^2} > 0$.

The next figure graphs the feasible points (x,r) that have been discussed, as well as the subset of points that describe a paradoxical loss. This subset consists of the points (x,r) that lie above the rising graph of the hyperbola and below the horizontal line $r = 0.333$. Using the evalF function on a TI graphing calculator, the hyperbola is found to intersect the vertical line $x = 50$ at $r=.240+\frac{25}{400-50}\approx .311429$. Moreover, by using the SOLVER function on a TI calculator, the hyperbola is found to lie above the horizontal line $r = 0.333$ for $x > 131.1728$. Thus, paradoxical losses correspond to the points (x,r) such that $50 \le x \le 131$ and $.311429 \le r < .333$. Theorem 1 justifies the appearance of Figure 1.

Next, determine the probable paradox of Player A losing to someone with a batting average of only 0.240 for the first half of the season (such as Player D). To determine the probability of such a paradoxical loss, the ratio of two areas must be found. (This approach to calculating probability through so-called "geometrical

Figure 1



Paradoxical losses correspond to the points (x, r) in this region.

methods" likely originated in the classic textbook *Higher Algebra* (Hall and Knight 1960, 401–02, especially the example on page 402). This approach is justified if all the feasible points are deemed to be equally likely. In this case, one takes a uniform probability density function, and the usual calculus-based method for computing probability relative to a continuous distribution reduces to taking a ratio of areas. Specifically, the probability in question is the ratio obtained by dividing the area that is above the hyperbola and below the line $r = 0.333$ by the area of the rectangle that encloses all the feasible points. The area of that rectangle is (base)(height) = $(350 - 50)(0.5 - 0) = 150$. The other area is found by this article's only essential use of calculus—the definite integral

$$\int_{50}^{131} .333 - \frac{121 - .240x}{400 - x} dx \approx 0.95245562796.$$

Accordingly, the probability of Player A losing to someone with a first half-season batting average of 0.240 is approximately

$$\frac{0.95245562796}{150} \approx 0.006349704186 \approx 0.635\%.$$

The low value of this probability, which is less than 1 per cent, somewhat justifies the skepticism (if not the disappointment) that Player A felt when informed of his defeat by Player D. The next two sections will show that one should be less sceptical that foes who are more formidable than Player D could also defeat Player A.

We have already seen that Player D can defeat Player A with a paradoxical loss. The above work helps construct and identify another example of a victor: Player D*. The following table describing the batting performance of Player D* can be obtained by taking $50 \leq x \leq 131$ in the above data for Player G.

Although the mathematical analysis led to a maximum value of x that was slightly greater than 131, it is desirable to have an example of the above phenomenon, along with a table of batting average performances that lists a whole number of at-bats and a whole number of hits for each half of the season. The preceding table displays the example (Player D*) with the largest integral value (namely, 104) for x.

The Probability of Losing Paradoxically to a More Worthy Opponent

Next, calculate the probably paradox of Player A losing to someone like Player C, who had a batting average of 0.294 during the first half of the season. The same reasoning from the preceding section can be used with the following changes. First, replacc 0.240 with 0.294. The result is the function $r = f(x) = \frac{121 - .294x}{400 - x} = .294 + \frac{3.4}{400 - x}$ instead of $r = f(x) = .240 + \frac{25}{400 - x}$. Replace the number 0.311429 from Figure 1 with 0.3037 . . ., and replace 131.18 with 312.8205 (rounded down to 312). As before, the probability is the ratio of two arcas, the denominator is still 150 and thc numerator is given by thc definite integral

$$\int_{50}^{312} .333 - \frac{121 - .294x}{400 - x}\,dx \approx 5.52397244398.$$

Hence, the probability of Player A losing to somcone with a first half-season batting average of 0.294 is approximately

$$\frac{5.52397244398}{150} \approx 0.03682648296 \approx 3.68\%.$$

An example of someone (Player D**) who can inflict a paradoxical defeat on Player A after recording a batting average of 0.294 during the first half of the season is described in the next table.

Although the mathematical analysis in this section led to a value of x that was slightly greater than 312, it is desirable to have an example of the above phenomenon with a table of batting averages that lists a whole number of at-bats and a whole number of hits for each half-season. The preceding table displays such an example, with x taking the largest possible integral value closest to 200 (that is, 197).

The Probability of Losing to a More Worthy Opponent Whom One Had Defeated During the First Half-Season

Next, calculate the probability that Player A will lose (not necessarily paradoxically) to someone who had a batting average of 0.294 during the first half-season. The same reasoning from the preceding section can be used with the following changes. Because nonparadoxical losses are allowed, the maximum permissible batting average for the second half of the season is 0.500. This changes the integrand. In addition, the upper limit of integration changes from 312 to 350 because the (rising) graph of $r = f(x) = \frac{121 - .294x}{400 - x} = .294 + \frac{3.4}{400 - x}$ remains below the horizontal line $r = 0.500$, the point being that $.294 + \frac{3.4}{400 - 350} \approx .362 < .500$. As before, the probability is thc ratio of two areas, the denominator is still 150 and the numerator is given by the definite integral

$$\int_{50}^{350} .500 - \frac{121 - .294x}{400 - x}\,dx \approx 55.1839054932.$$

Hence, the probability of Player A losing (not necessarily paradoxically) to someone with a first half-season batting average of 0.294 is approximately

$$\frac{55.1839054932}{150} \approx 0.367892703288 \approx 36.79\%.$$

Closing Comments

A more realistic analysis of the above probabilities may require the consideration of more than just areas. In the above analysis, it was assumed that all points (x,r) in the feasible region were equally probable. This means that constant (or uniform) probability-density functions were implicitly used. An analysis

Table 3

Player D*	At-Bats	Hits	Batting Average
First half of the season	104	25	0.240
Second half of the season	296	97	0.328
Totals for the entire season	400	122	0.305

Table 4

Player D**	At-Bats	Hits	Batting Average
First half of the season	197	58	0.294(416...)
Second half of the season	203	64	0.315(27...)
Totals for the entire season	400	122	0.305

of batting-average data from major league baseball may show that this assumption is inappropriate. In that case, certain points would have to be weighted more heavily than others by using nonconstant (that is, nonuniform) probability-density functions. Such nonuniform density functions would arise as factors in integrands figuring in the numerators (and implicitly in the denominators) of refinements of the above probability calculations.

A more realistic analysis would allow for competition between players who have a different number of at-bats for the entire season. More independent variables would need to be introduced to address such considerations. The resulting graphs would not be planar and the resulting calculations would involve multiple (or iterated) integrals. Computer technology would be needed for the analyses and the calculating of the more realistic probabilities. From a qualitative point of view, these more realistic modelling activities would likely lead to the same conclusions that were drawn from the more accessible work described above.

One role of mathematics and philosophy is to attempt to resolve a paradox by clarifying the (often unstated) assumptions that underlie the paradox's original formulation. Consider, for instance, the paradox of Zeno that claims that flight is impossible. Zeno reasoned that at any given instant an arrow cannot be in motion. Cameras seem to support this view by presenting images of objects that are momentarily frozen in one place. Zeno reasoned that, because time is nothing more than a succession (or set, as we might say now) of instants, there is no time during which motion is possible. The flaw in Zeno's argument is that moving objects can, in fact, have nonzero instantaneous velocities. This great insight of differential calculus was the key concept that defeated Zeno's paradox. What insight, if any, can defeat the paradox expressed by Player A?

Player A, like every student of mathematics, needs to understand that batting averages are global entities that summarize performances over extended periods of time. Player A was surprised at losing to Player D because he had not accounted for competition from a player who had a different number of at-bats than he did during each half-season. I hope that the following result will be of some consolation to Player A (and the reader). The result justifies the intuition of anyone who shared Player A's confusion; Theorem II shows that paradoxical losses are impossible on a level playing field—that is, when both players have the same number of at-bats in each half of the season.

Theorem 2

Two players cannot be involved in a paradoxical loss if they have the same number of at-bats in each half of the season.

Proof

Suppose that Players E and F each have b at-bats during the first half of the season and B at-bats during the second half of the season. Suppose that Player E has h_1 hits during the first half of the season and h_2 hits during the first half of the season (and Player F, respectively, has H_1 and H_2). Finally, suppose that Player E has a lower batting average than Player F during each half-season—that is,

$$\frac{h_1}{b} < \frac{H_1}{b} \text{ and } \frac{h_2}{B} < \frac{H_2}{B}.$$

Player E therefore has a lower batting average than Player F for the entire season—that is, that

$$\frac{h_1 + h_2}{b + B} < \frac{H_1 + H_2}{b + B}.$$

Now, since $b > 0$ and $B > 0$, it follows that $h_1 < H_1$ and $h_2 < H_2$. Therefore, $h_1 + h_2 < H_1 + H_2$. Because $b + B > 0$, the required inequality follows and the proof is complete.

Reference

Hall, H S, and S R Knight. 1960. *Higher Algebra*. 4th ed. New York, NY: Macmillan.

David E Dobbs received his bachelor of arts (majoring in mathematics, statistics and philosophy) and his master of arts (majoring in geometry and minoring in logic) from the University of Manitoba in his native Winnipeg. He received his doctorate in algebraic geometry from Cornell University. He has been on staff at UCLA, Rutgers University and the University of Tennessee, Knoxville, where he has held the position of professor since 1980. He enjoys teaching at all levels, and has directed seven PhD dissertations and several dozen senior and masters theses. His main research interest is in commutative algebra, where he has written eight books and more than 200 papers.