

# *Journal of Quantitative Analysis in Sports*

---

*Volume 4, Issue 2*

2008

*Article 3*

---

## Why On-Base Percentage is a Better Indicator of Future Performance than Batting Average: An Algebraic Proof

Ben S. Baumer\*

\*CUNY Graduate School and University Center, [bbaumer@gc.cuny.edu](mailto:bbaumer@gc.cuny.edu)

Copyright ©2008 The Berkeley Electronic Press. All rights reserved.

# Why On-Base Percentage is a Better Indicator of Future Performance than Batting Average: An Algebraic Proof

Ben S. Baumer

## Abstract

Batting Average (*AVG*) and On-Base Percentage (*OBP*) are two of the most commonly cited statistics in baseball. Existing research has demonstrated that for a team, *OBP* is more closely correlated to runs scored than is *AVG*, and secondly, for players, *OBP* is more closely correlated over time than is *AVG*. We offer an algebraic explanation for the latter phenomenon. Specifically, we will prove that batting average depends more heavily upon a particularly unpredictable variable, hits per balls in play (*HPBP*), than does *OBP*. This result will explain why for both batters and pitchers, on-base percentage is a better indicator of future performance than batting average.

**KEYWORDS:** baseball, dips, hpbp, proof, batting average, obp, math, algebra, calculus

## 1 Introduction

### 1.1 Background

One of the more recent chapters in the storied relationship between statistics and baseball has been the realization that the game's most well-known batting statistic, batting average (*AVG*), does not measure offensive prowess nearly as accurately as many people had previously believed. A newer statistic, on-base percentage (*OBP*), has now gained currency in the mainstream. On-base percentage was used by Brooklyn Dodgers statistician Allan Roth in the early 1950s (5), was famously touted by members of the Oakland Athletics front office in the early 2000s (3), and is now understood widely enough to be included on many television broadcasts underneath a player's name alongside his so-called "Triple Crown" stats: batting average, home runs (*HR*), and runs batted in (*RBI*).

Numerous researchers have demonstrated that in general, the correlation coefficient between a team's *batting average* over an entire season and the number of runs that the team scores is lower than the correlation coefficient between a team's *on-base percentage* over an entire season and the number of runs that the team scores<sup>1</sup>. This suggests that for an individual player, on-base percentage is a better measure of offensive prowess than is batting average. A separate, but related question, is whether on-base percentage or batting average is a better indicator of future offensive performance. The simplest way to address this question is to examine the consistency of these statistics for individual players over time. Today, it is well known that the correlation coefficient between players' on-base percentages in consecutive seasons is generally higher than the correlation coefficient between players' batting averages in consecutive seasons<sup>2</sup>. In this paper, we will explain why this phenomenon exists. To this end, we will approximate both *OBP* and *AVG* with functions of just four variables, and then prove that batting average is more heavily dependent upon one of those variables, hits per balls in play (*HPBP*), than is on-base percentage. As hits per balls in play happens to be considerably less predictable than the other three variables, we trace the inconsistency of batting average over time (relative to *OBP*) to its stronger dependence upon hits per balls in play.

### 1.2 The Hits per Balls in Play Breakthrough

The choice of the particular four variables we will use is motivated by a theory known to the sabermetric community as Defense Independent Pitching Statistics

---

<sup>1</sup>See, for example, Albert(1).

<sup>2</sup>For example, in our sample the correlation coefficient for *OBP* was 0.596, but was just 0.367 for *AVG*. Further explication as to the derivation of these numbers will follow.

Statistic	Batters ( $N = 715$ )	Pitchers ( $N = 587$ )
$\overline{BIP}$	0.854	0.760
$\overline{SO}$	0.824	0.764
$\overline{BB}$	0.757	0.626
$\overline{HR}$	0.730	0.349
$\overline{HBP}$	0.643	0.406
$\overline{HPBP}$	0.334	0.195

Table 1: Table of Year-to-Year Correlation Coefficients for selected frequencies

(DIPS). As a graduate student, McCracken noticed that the rate at which pitchers allowed hits on balls in play fluctuated wildly from season-to-season, and his research suggested that this variable behaved almost randomly (4). Subsequent research by many others, including James (2) and Tippett (6), has corroborated the central thesis of McCracken’s paper, but has argued persuasively that the behaviour of *HPBP* is not entirely random. In particular, Tippett’s evidence suggested that pitchers who threw a high percentage of knuckleballs tended to have lower *HPBP* rates, even when factors such as defense and era were controlled for. While there remains considerable debate over the extent to which certain pitchers (through the use of pitch type, movement, velocity, and location) can influence the rate of hits on balls put in play against them, McCracken’s primary contention that pitchers have far less control over hits per balls in play than previously believed is now widely accepted (at least among sabermetricians).

On the contrary, McCracken noted that the frequency of events in which the ball was *not* put in play against any given pitcher were quite stable. In Table 1, we list the correlation coefficients for consecutive season pairs of several frequencies for both batters and pitchers<sup>3</sup>. While McCracken’s observation about the volatility of *HPBP* was limited to pitchers, the data in Table 1 suggest that for batters as well as pitchers, *HPBP* is much less consistent across seasons than the frequency of walks, strikeouts, and home runs<sup>4</sup>.

<sup>3</sup>This was done by taking all consecutive year pairs in which the batter or pitcher had at least 250 plate appearances (or batters faced) in *both* seasons. Unless otherwise noted, all data is from STATS, Inc., and covers the 2003-2006 regular seasons. The  $\overline{bar}$  notation denotes frequency with respect to plate appearances, while the abbreviations used for the statistics listed will be explained precisely below.

<sup>4</sup>If you’re thinking that the correlation coefficient for  $\overline{HR}$  among pitchers is low, note that these numbers do not include corrections for home ballpark, which would improve this figure considerably. Also, the important thing is just that for both batters and pitchers, the correlation coefficient for *HPBP* is much lower than it is for the other statistics.

## 2 A Mathematical Framework

### 2.1 Basic Definitions

We begin by outlining the basic equations that govern the space of baseball. The fundamental event, from our perspective, is the plate appearance. The plate appearance will ideally cover all instances in which a batter faces a pitcher and an outcome is reached. (This excludes such events as when a runner is caught stealing for the third out of an inning, or catcher interference calls.)

**Definition 1.** *Total plate appearances (TPA) is the sum of at-bats (AB), walks (BB), hit-by-pitches (HBP), sacrifice flies (SF), and sacrifice hits (SH)<sup>5</sup>. That is,*

$$(1) \quad TPA = AB + BB + HBP + SF + SH$$

**Definition 2.** *Hits (H) are exactly those singles (1B), doubles (2B), triples (3B), and home runs (HR). Thus,*

$$H = 1B + 2B + 3B + HR$$

**Definition 3.** *Balls in play (BIP) includes all outcomes of a plate appearance in which the ball is struck by the batter and the opposing defense is potentially involved in a non-trivial way. This excludes strikeouts (SO) and home runs. That is,*

$$(2) \quad BIP = TPA - BB - SO - HBP - HR$$

The first two definitions are standard. The last definition was used by McCracken in his original publication of DIPS (4). The use of "nontrivial" excludes the putouts recorded by catchers on almost all strikeouts, including those which were foul-tipped. Third strikes which elude the catcher are sufficiently rare to be ignored, and in any case would not be considered a ball in play since they were not struck by the batter. Foul balls and would-be home runs are only included if they are caught. A home run is not considered a ball in play since (in almost all cases) no fielder had a chance to make a play on the ball. Note that we have attempted to eliminate from our definition of *BIP* the rather arbitrary designation of at-bat (*AB*). We should also note that all quantities considered to this point are non-negative integers. By combining equations (1) and (2), we see that

$$AB = BIP + HR + SO - SF - SH$$

---

<sup>5</sup>More commonly known as sacrifice bunts.

and rearranging (2) shows that

$$(3) \quad TPA = BIP + HR + SO + BB + HBP$$

The major outcomes of a plate appearance are listed in Figures 1 and 2 with their relative frequencies. The outcome  $O$  simply counts all generic balls in play that are not hits or sacrifices.

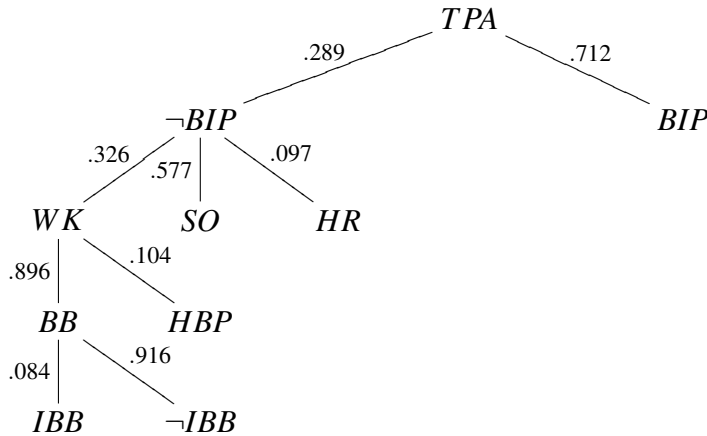


Figure 1: Tree diagram for balls not in play, with relative frequencies

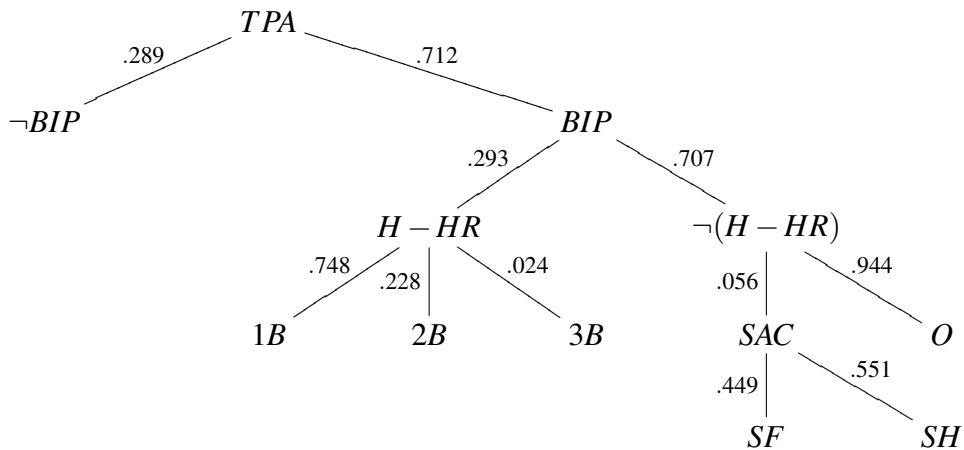


Figure 2: Tree diagram for balls in play, with relative frequencies

We will focus on the five major mutually independent outcomes of a plate appearance present in equation (3). Let  $S = \{BIP, HR, SO, BB, HBP\}$  be the set of all such outcomes. For any  $s \in S$ , let  $\bar{s} = \frac{s}{TPA}$ , be the frequency of  $s$  with respect to

plate appearances and observe that  $\bar{s} \in [0, 1]$ . In particular, note that

$$\begin{aligned}\overline{BIP} &= \frac{BIP}{TPA} \\ &= \frac{TPA - BB - SO - HBP - HR}{TPA} \\ &= 1 - \overline{BB} - \overline{SO} - \overline{HBP} - \overline{HR}\end{aligned}$$

In Table 1 above, we listed the year-to-year correlation coefficients for the members of  $S$  for both batters and pitchers. Note that (perhaps contrary to popular belief),  $\overline{HBP}$  for batters has a relatively high year-to-year correlation coefficient and is thus generally regarded as a skill.

Since the value of  $\overline{HBP}$  is generally small, and each  $HBP$  has exactly the same consequences as a  $BB$ , we define their sum  $WK = BB + HBP$  and write:

$$\overline{BIP} = 1 - \overline{WK} - \overline{SO} - \overline{HR}$$

Similarly, we group the relatively infrequent sacrifices together by defining  $SAC = SF + SH$ , with the idea that in both cases, the batter was content in making an out in order to advance a runner. Thus, we observe that

$$AB = TPA - BB - HBP - SF - SH \Rightarrow \overline{AB} = 1 - \overline{WK} - \overline{SAC}$$

The concept of hits per balls in play is central to our analysis, and we define it here:

**Definition 4.** *Hits per balls in play is the percentage of balls in play that fall for hits*

$$HPBP = \frac{H - HR}{BIP}$$

## 2.2 Batting Average

We have now defined all of the quantities we will need. Let  $X = \{\overline{WK}, \overline{SO}, \overline{HR}, HPBP\} = \{w, x, y, z\}$  be the set of the four variables to which we previously alluded, and let  $V = [0, 1]^4$ . Our goal is simply to express batting average solely in terms of the members of  $X$ . That is, we seek a function on  $V$  that computes batting average, which we must first define.

**Definition 5.** *Batting average is the quotient of hits and at-bats. That is,  $AVG = H/AB$ .*

Next, we perform a simple (albeit convoluted), algebraic manipulation:

$$\begin{aligned}
 AVG &= \frac{H}{AB} = \left(\frac{TPA}{AB}\right) \left(\frac{H}{TPA}\right) \\
 &= \frac{1}{AB} \cdot \frac{H - HR + HR}{TPA} \\
 &= \frac{1}{1 - \overline{WK} - \overline{SAC}} \cdot \left[\frac{H - HR}{TPA} + \overline{HR}\right] \\
 &= \frac{1}{1 - \overline{WK} - \overline{SAC}} \cdot \left[\left(\frac{H - HR}{BIP}\right) \left(\frac{BIP}{TPA}\right) + \overline{HR}\right] \\
 &= \frac{1}{1 - \overline{WK} - \overline{SAC}} \cdot \left[HPBP \cdot \overline{BIP} + \overline{HR}\right] \\
 &= \frac{1}{1 - \overline{WK} - \overline{SAC}} \cdot \left[HPBP(1 - \overline{WK} - \overline{SO} - \overline{HR}) + \overline{HR}\right] \\
 &\approx \frac{1}{1 - \overline{WK}} \cdot \left[HPBP(1 - \overline{WK} - \overline{SO} - \overline{HR}) + \overline{HR}\right]
 \end{aligned}$$

where the last approximation is justified by the observation that  $\overline{SAC}$  rates are very low for non-pitchers. Thus, we can define a function  $f : V \rightarrow [0, 1]$  that approximates batting average, as

$$f(w, x, y, z) = \frac{1}{1 - w} \cdot \left(z(1 - w - x - y) + y\right) = z \cdot \left(1 - \frac{x + y}{1 - w}\right) + \frac{y}{1 - w}$$

Calculating the partial derivatives of  $f$ , we see that:

$$\begin{aligned}
 \frac{\partial f}{\partial w} &= \frac{y - z(x + y)}{(1 - w)^2} & \frac{\partial f}{\partial x} &= -\frac{z}{1 - w} \\
 \frac{\partial f}{\partial y} &= \frac{1 - z}{1 - w} & \frac{\partial f}{\partial z} &= 1 - \frac{x + y}{1 - w}
 \end{aligned}$$

Note that if  $w = x = y = 0$ , then  $f = z$  and so  $f_z = 1$ . This simply states that if you never walked, struck out, or homered, your batting average would equal your hits per balls in play rate. Also, note that in each case, the change in  $AVG$  is proportional to the inverse of 1 minus the walk rate (or the square of this). This may seem counter-intuitive, since walks do not directly factor into the calculation of batting average as it is defined, but our observation is a consequence of the fact that at-bats are partially defined in terms of those plate appearances that are not walks. Thus, while it is true that a walk will not raise your batting average, a player who walks more frequently will, under certain conditions, have a lower batting average! As such, we can make the following observation:

**Observation 1.** If  $\frac{1-z}{z} < \frac{x}{y}$ , then  $f_w < 0$ .

*Proof.* Since the denominator of  $f_w$  is always positive, it follows that

$$f_w < 0 \iff y - z(x+y) < 0 \iff y(1-z) < xz \iff \frac{1-z}{z} < \frac{x}{y}$$

□

This is interesting only because it has been the case lately that almost all batters meet the condition required in the observation. Specifically, of the 715 batters whose statistics qualified for inclusion in Figure 1, 675 (94.4%) satisfied the condition of the observation in the first year, and 672 (94.0%) satisfied the condition in the second year. [Those batters who did not satisfy the condition were primarily those with low values of  $z$  and/or low ratios of  $x$  to  $y$  (i.e. - a low *HPBP* rate, but a relatively many *HR* as compared to *SO*.)] Thus, in practice, almost all batters will lower their batting average by increasing their walk rate.

### 2.3 On-Base Percentage

Next, we will perform a similar exercise with on-base percentage, which we first define.

**Definition 6.** *On-base percentage is the quotient of outcomes in which the batter earns at least first base by those plate appearances in which he attempted to reach first. Thus,*

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF} = \frac{H + WK}{TPA - SH}$$

The calculation to express *OBP* as a function on  $V$  follows more easily than it did for batting average:

$$\begin{aligned} OBP &= \frac{H + WK}{TPA - SH} \\ &= \frac{TPA}{TPA - SH} \cdot \left[ \frac{H - HR}{TPA} + \overline{HR} + \overline{WK} \right] \\ &= \frac{1}{1 - \overline{SH}} \cdot \left[ \frac{H - HR}{BIP} \cdot \overline{BIP} + \overline{HR} + \overline{WK} \right] \\ &= \frac{1}{1 - \overline{SH}} \cdot \left[ \overline{HPBP} \cdot \overline{BIP} + \overline{HR} + \overline{WK} \right] \\ &= \frac{1}{1 - \overline{SH}} \cdot \left[ \overline{HPBP}(1 - \overline{WK} - \overline{SO} - \overline{HR}) + \overline{HR} + \overline{WK} \right] \\ &\approx \overline{HPBP}(1 - \overline{WK} - \overline{SO} - \overline{HR}) + \overline{HR} + \overline{WK} \end{aligned}$$

Our approximation here is even more justified than in the first case, since  $\overline{SH} \leq \overline{SAC}$ , and indeed, for many of the game's best hitters,  $\overline{SH} = 0$ . Thus *OBP* can be written as a function  $g : V \rightarrow [0, 1]$ , where  $g(w, x, y, z) = z(1 - w - x - y) + (w + y)$ . Again, we have the four partial derivatives:

$$\begin{aligned} \frac{\partial g}{\partial w} &= 1 - z & \frac{\partial g}{\partial x} &= -z \\ \frac{\partial g}{\partial y} &= 1 - z & \frac{\partial g}{\partial z} &= 1 - x - w - y \end{aligned}$$

Note that by definition  $0 \leq z \leq 1$ , but in practice, for even a modest number of plate appearances,  $z < 0.5$ , so that in general,  $1 - z > z$ . Thus,  $|g_w| = |g_y| > |g_x|$ , and thus you will raise your *OBP* more by walking or homering than you will reduce it by not striking out. Recalling multivariable calculus, for each frequency  $a \in X$ , the partial derivative of  $g$  with respect to  $a$  reflects the rate of change of  $g$  with respect to  $a$ . Thus,  $g_z$  tells us that the more you walk, homer, or strike out, the less your *OBP* will change relative to small changes in your *HPBP*. This is an important fact, since the major lesson from DIPS theory is that *HPBP* is a far more volatile variable than  $\overline{WK}$ ,  $\overline{SO}$ , or  $\overline{HR}$ .

Note the similarities between the expressions for  $f$  and  $g$ . In fact, we have that  $g - w = z(1 - w - x - y) + y$ , and so

$$f = \frac{g - w}{1 - w} \Rightarrow AVG \approx \frac{OBP - \overline{WK}}{1 - \overline{WK}}$$

### 3 Conclusions

#### 3.1 Some Implications

We are now ready to prove our main result.

**Theorem 1.** *If  $w < 1$ , then  $f_z \geq g_z$ .*

*Proof.* Under our assumption,  $0 \leq w < 1 \Rightarrow 0 < 1 - w \leq 1 \Rightarrow 1 \leq \frac{1}{1-w} < \infty$ . The result then follows directly from the partial derivatives calculated above, since

$$\frac{\partial f}{\partial z} = 1 - \frac{x+y}{1-w} = \frac{1}{1-w}(1-x-w-y) = \frac{1}{1-w} \cdot \frac{\partial g}{\partial z} \Rightarrow \frac{\partial f}{\partial z} \geq \frac{\partial g}{\partial z}$$

□

The interpretation here is that for any batter who does not walk every time he comes to the plate, small changes in his *HPBP* will induce smaller changes in his *OBP* than it will his *AVG*. Furthermore, for any batter who has ever walked, the inequality becomes strict, as the following corollary shows.

**Corollary 2.** *If  $0 < w < 1$ , then  $f_z > g_z$ .*

This result, while somewhat trivial, is importantly different from the statistical analysis that has been performed to date, in that it proves something specific about the mathematical relationship between *HPBP* and *AVG* and *OBP*. In particular, it provides compelling evidence that the reason that year-to-year correlations for *OBP* are higher than those for *AVG* is a simple consequence of two things: the mathematical definitions of the two formulas; and the volatility of *HPBP*. While statistical analysts had previously observed that players' *OBPs* were generally more consistent over time than their batting averages, McCracken was the first to isolate the *HPBP* variable. We have taken our inspiration from him and shown definitively that close approximations of the formulas for *OBP* and *AVG* can be constructed using only four variables, one of which is significantly less predictable than the others, and that *OBP* is less sensitive to changes in that variable (*HPBP*). This is the reason that *OBP* tends to be more consistent over time.

### 3.2 Extensions

The implications of Theorem 1 probably contribute nothing new to those familiar with modern thinking about baseball statistics. If there is value in this result beyond a mathematical explanation of something that statistical analysts have already observed, it is a framework for analyzing baseball players. Let  $V$  be the space of quadruples  $(w, x, y, z)$ , where  $w, x, y, z \in [0, 1]$ . Note that  $V$  is not vector space, since it is not closed under addition, but if we view  $V$  as a subset of  $\mathbb{R}^4$ , we can safely perform vector operations. Any batter or pitcher can be modelled by a vector in  $V$ . For example, David Wright walked in 9.6% of his plate appearances from 2003-2006, struck out in 16.7%, homered in 4.1%, and was hit by a pitch in another 0.9%. His *HPBP* rate was 0.331. We could thus define the vector  $v_{DW} = (0.105, 0.167, 0.041, 0.331) \in V$ , and use it to model Wright. We could approximate his batting average, as  $f(v_{DW}) = 0.300$  (his actual batting average was .303), and his on-base percentage, as  $g(v_{DW}) = 0.373$  (his actual *OBP* was .374). We could define the NL average batter over this period as the vector  $v_{NL} = (0.096, 0.172, 0.027, 0.291)$ , and calculate  $\|v_{DW} - v_{NL}\|_p$  for any norm  $p$  of our choice. This would give us a measure of how differently Wright performed over this time period from the NL average batter. We could similarly compare two different players, or the same player over different time intervals. We could use  $V$  as a setting for a Markov model. Further analysis is beyond the scope of this paper, but there are a few quick observations we can make using this framework.

The sabermetrician, in particular, will appreciate the following:

**Observation 2.** Let  $A = (w_a, x_a, y_a, z) \in V$  and  $B = (w_b, x_b, y_b, z) \in V$  be two batters. If  $w_a + x_a + y_a \geq w_b + x_b + y_b$ , then  $g_z(A) \leq g_z(B)$ .

*Proof.* This is immediate, since:

$$g_z(A) = 1 - x_a - w_a - y_a = 1 - (w_a + x_a + y_a) \leq 1 - (w_b + x_b + y_b) = g_z(B)$$

□

The implication here is that since *HPBP* is such a volatile variable, a batter can insulate himself from the vagaries of chance by relying more upon those outcomes that do *not* result in a ball in play. Perhaps more importantly for those of us in front offices, the implication is that players with a higher frequency of walks, home runs, and strikeouts, are, *ceteris paribus*, more (easily) predictable. That is, other things being equal, their future *OBP* is more likely to reflect their recent *OBP*.

A similar statement can be made for batting average, but with stronger conditions.

**Corollary 3.** With  $A$  and  $B$  as above, if, additionally,  $w_a \leq w_b$ , then  $f_z(A) \leq f_z(B)$ .

*Proof.* Using the fact that  $f_z = \frac{g_z}{1-w}$ , we get that

$$f_z(A) = \frac{g_z(A)}{1 - w_a} \leq \frac{g_z(B)}{1 - w_b} = f_z(B)$$

Since, by Observation 2,  $g_z(A) \leq g_z(B)$ , and by assumption,  $w_a \leq w_b \Rightarrow 1 - w_a \geq 1 - w_b$ . □

Note that while Theorem 1 was a result about the impact of *HPBP* upon two different *statistics* (*OBP* and *AVG*), these last two statements are results about the impact of *HPBP* upon the statistics of two different *players*.

## References

- [1] ALBERT, JIM AND BENNETT, JAY. *Curveball: Baseball, Statistics, and the Role of Chance in the Game*. 2003. Copernicus Books, New York.
- [2] JAMES, BILL. *The New Bill James Historical Baseball Abstract*. 2001. Free Press, New York.
- [3] LEWIS, MICHAEL. *Moneyball: The Art of Winning an Unfair Game*. 2003. W. W. Norton & Co., New York.

- [4] MCCRACKEN, VÖRÖS . *Pitching and Defense: How Much Control Do Hurlers Have?*. 23 January 2001. Baseball Prospectus, [<http://baseballprospectus.com/article.php?articleid=878>]
- [5] SCHWARZ, ALAN. *The Numbers Game: Baseball's Lifelong Fascination with Statistics*. 2004. St. Martin's Press, New York.
- [6] TIPPETT, TOM. *Can Pitchers Prevent Hits on Balls in Play?*. 21 July 2003. Diamond Mind Baseball, [<http://www.diamond-mind.com/articles/ipavg2.htm>]