SPORTS

Is an .833 Hitter Better Than a .338 Hitter?

Jesse FREY

This article considers the problem of using batting average alone to estimate a baseball player's chance of getting a hit. This problem differs from typical proportion estimation problems because we know only the observed proportion of successes rather than both the number of successes and the number of trials. Our information is also restricted because the observed proportion of successes is reported to only three decimal places. We solve this problem in the context of present-day major league baseball by first developing a model for the joint distribution of hits, at bats, and chance of getting a hit. We then treat that model as a prior distribution and update the prior based on the observed batting average. One interesting result is that among batting averages likely to occur in practice, .334 leads to the highest posterior mean for true ability.

KEY WORDS: Baseball; Batting average; Bayesian methods.

1. INTRODUCTION

Batting average, defined as the ratio $b \equiv h/a$ of a player's hits h to his at bats a, is one of the oldest and perhaps the best known of all baseball statistics. Though it is now recognized to be of less value in assessing the quality of a player's contributions than on-base percentage, slugging average, or a myriad of other statistics [see, e.g., the recent bestseller *Moneyball* (Lewis 2003)], batting average still has special significance for baseball fans and players. For example, a batting average of .300 tends to be regarded as good almost regardless of how meager a player's other contributions are, and a .400 average (not achieved in major league baseball since 1941) continues to be a goal that some of the greatest players aspire to. One batting-average-related convention that is important for this article is that batting average is always reported to exactly three decimal places, with rounding used to determine the value in the third decimal place.

To a statistician, a batting average is an estimate of a player's true chance p of getting a hit. However, it is an estimate that fails to take into account important auxiliary information such as the player's past history and the distribution of true abilities in the full population of baseball players. Thus, provided that the values a and h that led to the batting average in question are available, a number of improved estimates are possible. A simple Bayesian approach might consist of putting a prior distribution on p, modeling the distribution of batting average given p and a, and updating the distribution of p based on the observed batting average. One might also follow the strategy of Efron and Morris (1975), who used an empirical Bayes approach to estimate true abilities for selected players on the basis of the first 45 at bats of the 1970 season. For such approaches to work, however, not just the batting average, but both a and h, must be known.

This article considers the problem of using batting average to estimate a player's true chance p of getting a hit when the values h and a are unavailable. This problem differs from typical proportion estimation problems because we know only the observed proportion of successes rather than both the number of successes and the number of trials. It is also a censored data problem because batting average is reported in an interval-censored form. This problem might arise because of space considerations (e.g., space on a scoreboard) or because of poor record-keeping. Work on this problem was motivated by a trip to the local minor-league ballpark in June 2004. Finding that batting average was the only statistical information presented on the scoreboard and then discovering that one of the home team's players entered the game with an .857 average, the author wondered what the appropriate inference was.

Although we focus on batting average in this article, similar problems may also arise in analyzing sample surveys because of the need to preserve confidentiality. In a highly stratified survey in which some of the in-stratum sample sizes are small, reporting only the sample proportion rather than the number of successes and the number of trials may help in preserving anonymity for respondents.

To see that this problem may lead to surprising conclusions, consider the following example. Suppose that we are told that Player A and Player B, selected at random, posted batting averages of .833 and .338, respectively, in a recent season. Our initial reaction might be to think that Player A has the better true chance of getting a hit. Once we consider, however, that players who play regularly tend to post batting averages between .200 and .400, we realize that Player A has probably batted only a handful of times. In fact, the single most likely scenario is that he has collected five hits in six at bats. To achieve a batting average of .338, however, Player B must have batted at least 65

Jesse Frey is Assistant Professor, Department of Mathematical Sciences, Villanova University, Villanova, PA 19085 (E-mail: *jesse.frey@villanova.edu*). The author thanks editor Peter Westfall and three anonymous reviewers for helpful comments that have improved the article. He also thanks Steve MacEachern for comments on an earlier version of the article.

 DOI: 10.1198/000313007X192437

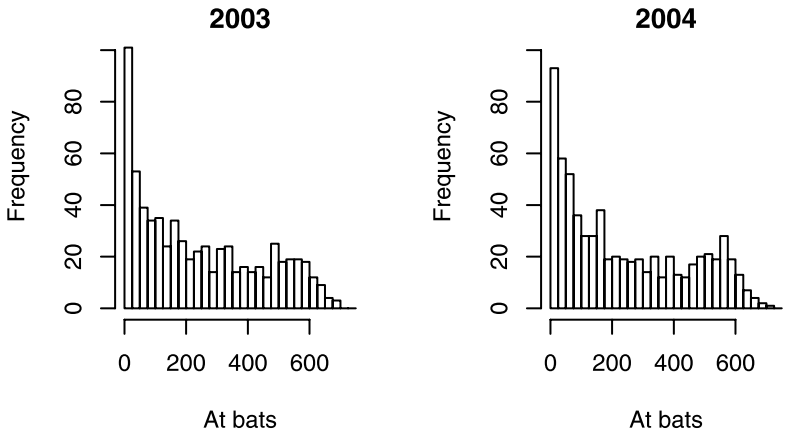

Figure 1. Histograms of player at bats in 2003 and 2004.

times, which in itself is an indication of quality. That .338 is also a higher-than-usual batting average provides additional evidence that it is Player B who has the better true chance of getting a hit. In fact, we find in this article that in the context of present-day major league baseball, .338 is among the most impressive of all batting averages in the sense that it leads to a high posterior mean for true ability. Among batting averages likely to occur in practice, .334 is the batting average that leads to the highest posterior mean for true ability.

We treat this problem in a Bayesian fashion, using data from the 2003 and 2004 major league baseball seasons to produce our prior. In Section 2, we develop a model for the joint distribution of at bats, hits, and true ability. The details of this model are specific to the context we consider, but the general approach is adaptable to other settings. In Section 3, we use this model as a prior distribution, and we update it based on the observed batting average to make inference on the distribution of true ability. We summarize the results using both tables and graphs. In Section 4, we address modeling concerns related to selection bias. Section 5 summarizes our results.

2. THE PROPOSED MODEL

In this section, we develop a model for the joint distribution of the random variables at bats (A), true ability (P), and hits (H). We first find a model for the marginal distribution of at bats. We then model the conditional distribution of hits given both at bats and true ability. Finally, we model the conditional distribution of true ability given at bats. We provide empirical and theoretical justifications for each modeling step, and we conclude the section with an informal test whose outcome suggests that our model captures the important features of the data. Two key features of the data that are captured by the model are that the distribution of at bats is right-skewed and that players with more at bats tend to be better than players with fewer at bats.

The data that we used, which was drawn from more complete data given in the Lahman (2004) database, consists of at bat and hit totals for all nonpitchers who batted at least once during the 2003 and 2004 seasons. Players who had at least one at bat for more than one team in a particular season appear in the data multiple times, once for each team.

To model the marginal distribution of at bats, we first made histograms of player at bat totals for the 2003 and 2004 seasons. Figure 1, which presents these histograms side by side, shows that the marginal distribution of at bats was roughly the same for the two seasons. It also suggests that the marginal distribution for at bats may be well approximated by a fairly simple curve. We constructed a probability mass function for at bats by first constructing a probability density supported on the real numbers, then interpreting the probability assigned to the interval $[a-1, a]$ as the probability of exactly a at bats.

Based on Figure 1, we took this probability density function to be composed of three parts, namely a decreasing quadratic portion supported on [0, 243], a flat portion supported on [243, 486], and a decreasing linear portion supported on [486, 729]. The break points 243, 486, and 729 were chosen both to agree with the histograms and to be easily interpretable in terms of the number of games, 162, in a full season. For example, the flat portion supported on [243, 486] includes players who batted between 1.5 and 3.0 times per game. Specifically, we took the probability of a player having exactly a at bats to be given by

$$d(a) \equiv \int_{a-1}^{a} f(t)dt, \tag{1}$$

where f is the density function

$$f(x) = \begin{cases} \dfrac{1}{850.5}\left(1 + \dfrac{3(x-243)^2}{243^2}\right), & 0 \le x < 243, \\ \dfrac{1}{850.5}, & 243 \le x < 486, \\ \dfrac{1}{850.5}\left(1 - \dfrac{(x-486)}{243}\right), & 486 \le x \le 729, \\ 0, & \text{otherwise}, \end{cases}$$

which agrees closely with the histograms in Figure 1. One possible objection to this model for at bats is that since there is no theoretical bound on the number of at bats a player could have in a season, we should not use a model with only finite support. However, no batter has ever had more than 705 at bats in a season, and if higher totals were to become common, that would

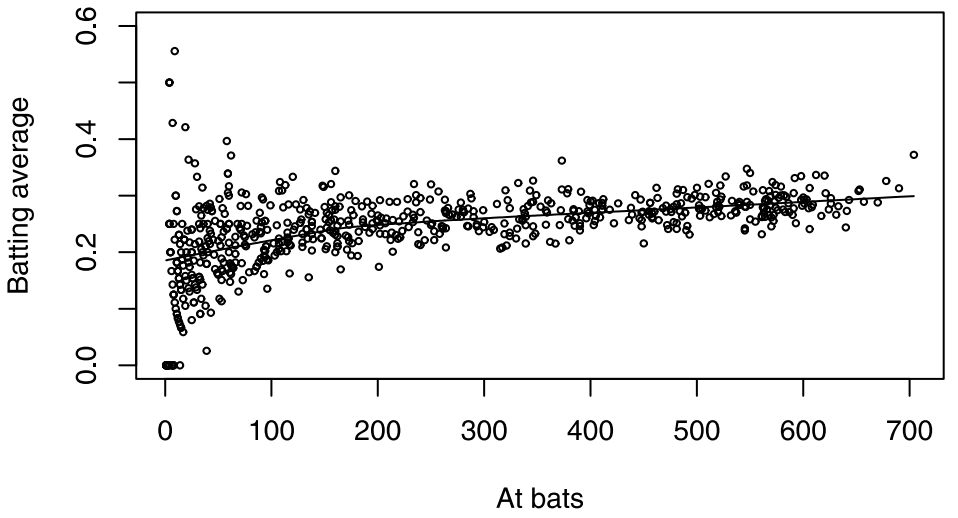

Figure 2. Plot of batting average versus at bats for 2004. The extra line is a lowess fit.

indicate a corresponding increase in batting averages that would also require us to adjust the other parts of our model.

We next turned to the problem of modeling the distribution of hits given at bats and true ability. It is clear that there would be computational advantages if the distribution of H given $A = a$ and $P = p$ were simply a Binomial(a, p) distribution. However, there are reasons to doubt that such a model is appropriate. Interestingly, there are both arguments that H should show extrabinomial variability and arguments that H should show subbinomial variability.

The argument that H should show extrabinomial variability rests on the observation that players do not have the same chance of success in each at bat. This occurs because, among other differences, some pitchers are better than others, some ballparks offer better hitting conditions than others, and some configurations of baserunners are more conducive to hitting than others. If a hitter's chance of success in a given at bat is independently drawn from some distribution with mean p, then the distribution of hits given at bats will show extrabinomial variability. In such a case we might want to try a model such as the one proposed by Rudolfer (1990).

The argument that H should show subbinomial variability rests on the observation that while players do not have the same chance of success in each at bat, many of the different situations that players encounter occur with orderly frequencies rather than at random. Some ballparks offer better hitting conditions than others, but the number of games a full-time player gets in each ballpark is fixed in advance. Since each game features roughly the same number of at bats (typically three to five), the situation is similar to one in which a player has a fixed number of at bats at each of several different chances of success. Such a set-up leads to subbinomial variability. In such a case we might use a model such as the one proposed by Kupper and Haseman (1978).

What seems to be the case, however, is that the conditional distribution of H given $A = a$ and $P = p$ shows neither extrabinomial nor subbinomial variability. Instead, the factors described in the last two paragraphs essentially cancel each other out, leading to a level of variability which is well captured by a binomial model. Since baseball has been featured in a number of articles in the statistical literature, the question of how to model hits given at bats and ability has arisen before. Berry, Reese, and Larkey (1999) used a binomial model for hits given at bats and true ability in their examination of changing skill levels in baseball across time. Casella and Berger (1994) modeled at bats as independent Bernoulli trials in their study of how to estimate a binomial success probability on the basis of a certain kind of selectively reported data. Albright (1993), in analyzing consecutive runs of successes and failures for batters in the 1987 to 1990 seasons, found no evidence that players are more or less streaky than would be expected if at bats are in fact independent Bernoulli trials. In the next paragraph, we consider one additional piece of evidence for this position.

One way to examine whether the distribution of hits given at bats and true ability has binomial variability is to look at data for consecutive years. A player's true ability is unlikely to change much in one year, meaning that the difference in batting average between two years can be used to assess the variability in batting average in a single year. Suppose that a player collects h_i hits in a_i at bats for consecutive years $i = 1, 2$. One estimate of his true ability is his overall success rate

$$\hat{p} = \frac{h_1 + h_2}{a_1 + a_2}.$$

Since a Binomial(a_i, p) distribution has variance $a_i\, p(1-p)$, the variance of batting average h_i/a_i is $p(1-p)/a_i$ under a binomial model, and the variance of the difference $h_1/a_1 - h_2/a_2$ is the sum $p(1-p)/a_1 + p(1-p)/a_2$. We looked at the group of players who had at least 300 at bats *for the same team* in each of 2003 and 2004, and we considered the normalized value

$$\frac{\dfrac{h_3}{a_3} - \dfrac{h_4}{a_4}}{\sqrt{\hat{p}(1-\hat{p})\left(a_3^{-1} + a_4^{-1}\right)}},$$

where $\hat{p} = \frac{h_3+h_4}{a_3+a_4}$ and the subscripts 3 and 4 correspond to 2003 and 2004, respectively. If the conditional distribution of hits given at bats and true ability does in fact show binomial variability, these values should be distributed like independent standard

normal random variables. A normal probability plot of these values was constructed, and this plot showed no noticeable departure from binomial variability. We thus model the conditional distribution of H given $A = a$ and $P = p$ as Binomial(a, p).

The final piece needed for our model is the conditional distribution of true ability given at bats. Since true ability p is a probability, the natural family to turn to is the Beta family of distributions. To simplify the modeling process, we proceed in two steps, first modeling the mean of ability given at bats, then considering the appropriate level of variability. Thus, rather than modeling the usual parameters α and β for the Beta family, we model the parameters μ and c, where $\mu = \frac{\alpha}{\alpha+\beta}$ is the mean and $c = \alpha + \beta$ is a mass parameter controlling the amount of spread around the mean. Figure 2 gives a plot of batting average versus at bats for the 2004 season. This plot, which is similar to the one obtained for the 2003 season, also include a lowess fit (see Cleveland 1979), and this lowess fit suggests that the mean of ability given at bats can be well approximated by a piecewise linear function of at bats. Specifically, it appears that two linear pieces are needed, one to account for the fairly steep grade from 0 to about 162 at bats and the second to account for the more gradual increase from about 162 at bats onward. Thus, we modeled the mean of ability as

$$\mu(a) = \begin{cases} .202 + a/3000, & a \le 162, \\ .256 + .00008(a - 162), & a > 162, \end{cases} \quad (2)$$

where the break point 162 was chosen both to agree with Figure 2 and to be easily interpretable in terms of the 162-game season.

In modeling the mass parameter c, we run into one theoretical concern. Since better players play more, we expect the distribution of P for batters with, say, 500 at bats to be stochastically larger than the distribution for batters with, say, 200 at bats. If we allow the mass parameter c to be a function of the number of at bats a, however, such a stochastic ordering need not hold. Thus, we elected to look for a single mass parameter c that will be applied for all values of a. To find this value c, we first plotted the variability of batting average as a function of at bats. For at bat bins $[290, 310), [310, 330), \ldots, [690, 710)$, we computed within-bin empirical standard deviations based on combined data from both the 2003 and 2004 seasons. These empirical standard deviations are plotted against their midpoints $300, 320, \ldots, 700$ in Figure 3. Figure 3 also gives curves representing the theoretical standard deviation (as a function of a) for various choices of the mass parameter c. These curves are drawn using the easily-shown fact that the variance of batting average given at bats is given by

$$\text{var}\left(\frac{h}{a}\right) = \frac{1}{a}\mu(a) + \left(1 - \frac{1}{a}\right)\left(\frac{c\mu(a)^2 + \mu(a)}{c+1}\right) - \mu(a)^2.$$

Based on the results shown in Figure 3, the value $c = 900$ was chosen.

At this point, our model is complete. As an informal test of this model, we simulated several seasons worth of data. Specifically, since there were 670 (a, h) pairs in the data for the 2004 season, we took each season to consist of 670 independently simulated values for the triplet (a, p, h). We then made plots of batting average h/a against at bats a for these 670 values. Figure 4 gives plots for four simulated seasons together with plots of the actual data for 2003 and 2004. The strong similarity between the simulated and actual data suggests that the model captures the important features of the data.

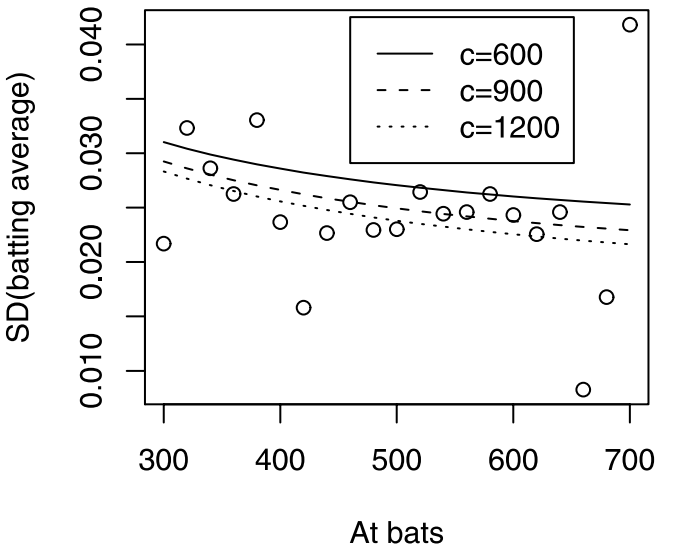

Figure 3. Empirical bin-by-bin standard deviations for batting average. The curves give theoretical standard deviations for three choices of the mass parameter c.

3. FINDING THE POSTERIOR DISTRIBUTION

In order to estimate p on the basis of batting average alone, we need to be able to compute the distribution of true ability given batting average. This calculation is facilitated by the fact that the Binomial and Beta distributions are conjugate distributions (see, e.g., Berger 1985). Suppose that a player has a at bats and h hits. Since the player has a at bats, the prior distribution of P given $A = a$ is a Beta$(c\mu(a), c(1 - \mu(a)))$ distribution. Updating this on the basis of h hits in a at bats, we find that the posterior distribution is a Beta$(c\mu(a) + h, c(1 - \mu(a)) + a - h)$ distribution. In our model, this Beta$(c\mu(a) + h, c(1 - \mu(a)) + a - h)$ distribution is the conditional distribution of P given a and h.

When we know only the batting average $B \equiv B(a, h)$ rather than the full pair (a, h), the updating process is somewhat more complicated. Instead of simply finding the distribution of P conditional on a single (a, h) pair, we need to find the distribution of P conditional on a collection of (a, h) pairs, namely all those pairs that lead to the batting average of interest. Given that the batting average arose with a certain number a of at bats, the posterior distribution of P is a Beta distribution as described above. Thus, the marginal posterior distribution of P given the batting average is a mixture of Beta distributions, where the mixing weights are the conditional probabilities that the batting average of interest arises in each particular way. Since a satisfies $1 \le a \le 729$, a given batting average can arise in only a finite number of ways, meaning that the mixture is a finite mixture. Let $(a_1, h_1), \ldots, (a_K, h_K)$ be a list of all the ways in which the batting average b can arise. Then the posterior distribution of P

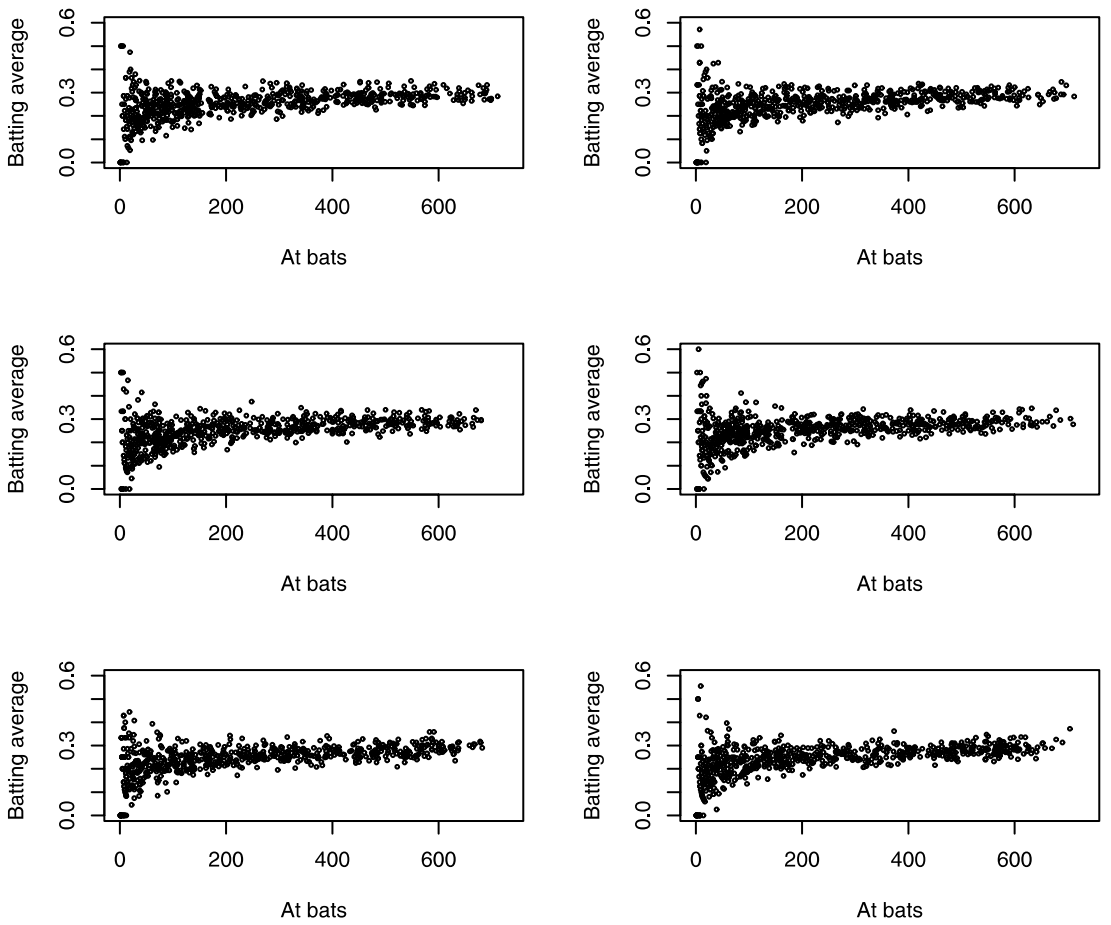

Figure 4. Batting average versus at bats for 2003 (lower left), 2004 (lower right), and for four datasets simulated using the model from Section 2.

given $B = b$ is the mixture

$$\sum_{i=1}^{K}\left(\frac{P(h_i \text{ hits and } a_i \text{ at bats})}{\sum_{j=1}^{K} P(h_j \text{ hits and } a_j \text{ at bats})}\right) \times \text{Beta}(c\mu(a_i)+h_i, c(1-\mu(a_i))+a_i-h_i). \quad (3)$$

The probabilities $P(h \text{ hits and } a \text{ at bats})$ may be computed using the fact that given that the number of at bats is $A = a$, the number of hits h has a Beta $-$ binomial distribution. Specifically, the probability of h hits and a at bats is given by

$$\begin{aligned}
&P(H=h, A=a) = P(A=a)P(H=h|A=a)\\
&= d(a)P\,(H=h|h \sim \text{Binomial}(a,\, p)\\
&\quad \text{and } p \sim \text{Beta}(c\mu(a), c(1-\mu(a))))\\
&= d(a)\int_{p=0}^{1}\left[\binom{a}{h} p^h(1-p)^{a-h}\right]\\
&\quad \times \frac{\Gamma(c)}{\Gamma(c\mu(a))\Gamma(c(1-\mu(a)))} p^{c\mu(a)-1}(1-p)^{c(1-\mu(a))-1}dp.\\
&= d(a)\left(\frac{\Gamma(a+1)\Gamma(c)}{\Gamma(h+1)\Gamma(a+1-h)\Gamma(c\mu(a))\Gamma(c(1-\mu(a)))}\right)\cdot\\
&\quad \int_{p=0}^{1} p^{h+c\mu(a)-1}(1-p)^{a-h+c(1-\mu(a))-1}dp\\
&= d(a)\left(\frac{\Gamma(a+1)\Gamma(c)\Gamma(h+c\mu(a))\Gamma(a-h+c(1-\mu(a)))}{\Gamma(h+1)\Gamma(a+1-h)\Gamma(c\mu(a))\Gamma(c(1-\mu(a)))\Gamma(a+c)}\right), \quad (4)
\end{aligned}$$

where $\Gamma(\alpha) = \int_{x=0}^{\infty} x^{\alpha-1}e^{-x}dx$ and $d(\cdot)$ is defined in Equation (1).

Using Equations (3) and (4), we can compute the posterior distribution for P given the batting average b, and similar considerations allow computation of the posterior distribution of, for example, A given b. These posterior distributions can be summarized in a number of ways, but one obvious summary consists of reporting the posterior mean and standard deviation. Figures 5 and 6 give the posterior mean and standard deviation of true ability P as a function of the batting average b. Rather than plot a point for every single batting average, a point is plotted only for those batting averages b whose probability $P(B = b)$ is greater than one in a billion, the others being essentially impossible to see in practice. These figures show that the batting averages corresponding to the highest posterior means lie mainly in the range .200 to .400 of ordinarily observed batting averages. These batting averages are also, however, the least informative in the sense that the posterior standard deviation for P tends to be large.

The ten single batting averages (with probabilities larger than one in a billion) with the highest and lowest posterior means are given in Table 1. What is apparent from Table 1 is that a batting average cannot correspond to a high posterior mean for P unless a large number of at bats are required to achieve that batting average. We also see, however, that a batting average b achievable in fewer at bats than another batting average may still correspond to a higher posterior mean for P provided that b is

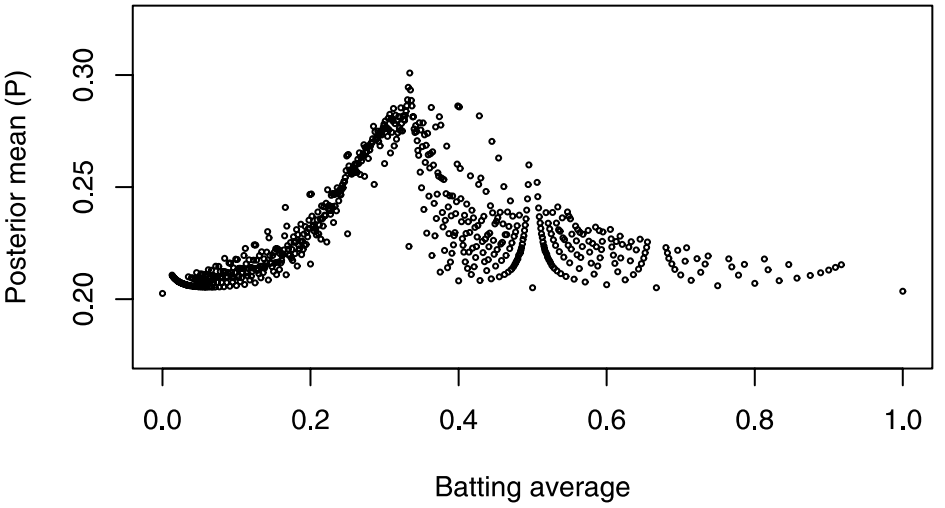

Figure 5. Posterior mean for P as a function of batting average.

high enough. The values corresponding to the lowest posterior means for P are those batting averages that are either achievable in just a few at bats or both achievable in a small number of at bats and extremely unlikely to be achieved in a large number of at bats.

4. SELECTION BIAS

In our modeling in Section 2, we behaved as if the mean batting average for all players with a at bats were an unbiased estimate of the true ability level for players who batted a times. That assumption may be reasonable if a is large, but for players who batted only a few times, we might worry about selection bias. Specifically, players who perform well in just a few at bats are likely to be given more opportunities, while players who perform poorly in just a few at bats are likely to see their opportunities reduced or even eliminated. Thus, we might expect the observed batting averages for players who batted just a few times to underestimate their true ability levels. To explore the extent to which this sort of bias might affect our results, we considered several alternatives to the mean model given in (2). These alternatives were obtained by making different choices for the slope of the mean as a function of a on the interval [0, 162]. In the most extreme alternative considered, we modeled the mean of ability as a linear function of a, in effect taking the higher observed slope on the interval [0, 162] to be entirely due to selection bias.

The alternatives considered led to only minimal changes in the list of batting averages corresponding to the very highest posterior means for P. Posterior means for batting averages above .400 increased in a relative sense, however, and the list of batting averages corresponding to the very lowest posterior means changed noticeably, with the list becoming dominated by batting averages that are both low and achievable in a low number of at bats.

5. SUMMARY

Using Bayesian methods, we have treated the problem of using batting average alone to estimate a baseball player's chance

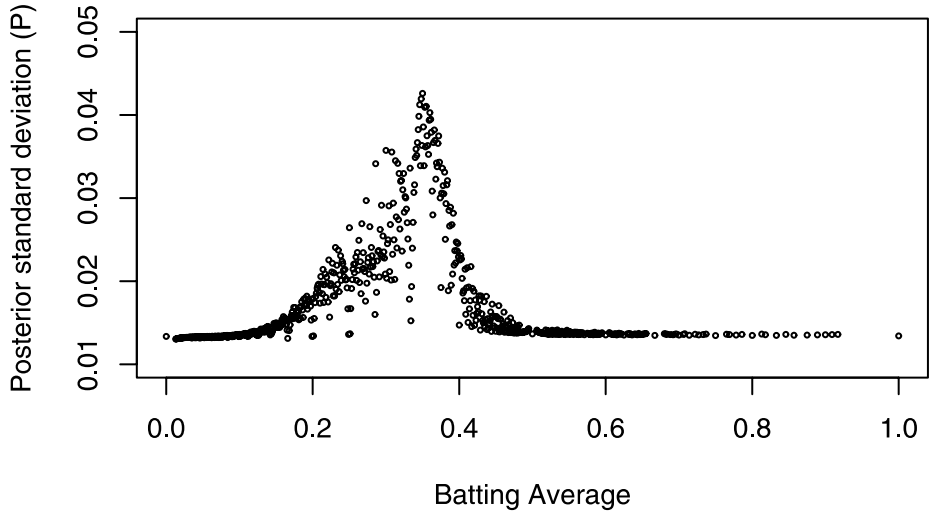

Figure 6. Posterior standard deviation for P as a function of batting average.

Table 1. Batting averages corresponding to the highest and lowest posterior means for p. Attention is restricted to batting averages b with at least a one-in-a-billion chance of occurring.

b	Posterior mean for P	Posterior SD for P	Posterior mean for A	b	Posterior mean for P	Posterior SD for P	Posterior mean for A
.334	.301	.015	505.3	.000	.203	.013	4.7
.332	.295	.018	441.6	1.000	.204	.013	1.3
.335	.293	.019	419.7	.500	.205	.014	4.6
.331	.289	.022	398.8	.667	.205	.013	3.6
.336	.289	.024	390.1	.067	.205	.013	18.6
.330	.286	.025	390.7	.059	.205	.013	19.6
.337	.286	.027	375.5	.063	.205	.013	19.3
.399	.286	.023	230.1	.071	.205	.013	18.3
.401	.286	.023	224.2	.056	.205	.013	20.2
.363	.286	.031	317.8	.053	.205	.013	20.9

of getting a hit. We have shown that in this restricted information setting, the most impressive batting averages are not the highest batting averages, but the batting averages that are both high and impossible to attain in a small number of at bats. We have also shown that no matter how high it may be, a batting average that is attainable in just a handful of at bats is evidence not that a player has a good chance of getting a hit, but that he has a poor chance. We have found that in the context of present-day major league baseball, .334 is the batting average that corresponds to the highest posterior mean for true ability. Though our modeling process involved the use of data specific to the context we considered, the general approach is adaptable not only to different baseball contexts, but also to any situation in which, because of space considerations, a poorly designed data collection process, or a data-reporting process designed to preserve confidentiality, a probability must be estimated without knowing the number of trials or the number of successes.

[Received October 2006. Revised December 2006.]

REFERENCES

Albright, S. C. (1993), "A Statistical Analysis of Hitting Streaks in Baseball," *Journal of the American Statistical Association*, 88, 1175–1183.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

Berry, S. M., Reese, C. S., and Larkey, P. D. (1999), "Bridging Different Eras in Sports," *Journal of the American Statistical Association*, 94, 661–676.

Casella, G., and Berger, R. L. (1994), "Estimation With Selected Binomial Information or Do you Really Believe Dave Winfield is Batting .471?" *Journal of the American Statistical Association*, 89, 1080–1090.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Efron, B., and Morris, C. (1975), "Data Analysis using Stein's Estimator and its Generalizations," *Journal of the American Statistical Association*, 70, 311–319.

Kupper, L. L., and Haseman, J. K. (1978), "The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments," *Biometrics*, 34, 69–76.

Lahman, S. (2004), "The Lahman Baseball Database." Available online at *http://www.baseball1.com*.

Lewis, M. (2003), *Moneyball: The Art of Winning an Unfair Game*, New York: W. W. Norton.

Rudolfer, S. M. (1990), "A Markov Chain Model of Extrabinomial Variation," *Biometrika*, 77, 255–264.