Which On-Base Percentage Shows the Highest True Ability of a Baseball Player?

January 31, 2018

Abstract

This paper looks at the true on-base ability of a baseball player given their on-base percentage. Unlike a typical proportion estimation problem where one is able to make an analysis based on knowing the number of successes and number of trials, this problem only gives us the proportion of successes rounded to three decimal places. This paper presents two different Bayesian models for finding the highest true on-base ability. The first Bayesian model considers modeling the distribution of times on-base, plate appearances and probability of being on-base as a prior distribution and then finding a posterior distribution given an observed on-base percentage. The second Bayesian model looks into the distribution of hits, walks, hit-by-pitches, plate appearances and probability of getting on-base and then creating a simulation of baseball players to see their true ability given a particular on-base percentage. After implementing the two models, it turns out that the highest reasonable on-base percentage that shows a player's true ability is around .4.

I. Introduction

On-base percentage (OBP), defined as the ratio $OBP \equiv \frac{H+BB+HBP}{AB+BB+HBP+SF}$ where H, BB, HBP, AB, and SF is a player's hits, walks, hit-by-pitches, at bats and sacrifice flies respectively, is one of the main statistics used to measure a baseball player's offensive capacity. According to the bestseller *Moneyball* (Lewis 2003), among the different statistics used to measure a player's hitting ability, OBP is one of the better measurements.

However, one of the things that can be deceiving about OBP is the true ability of a player given their OBP. It may be that the player's OBP was achieved with a very small sample size. For example, just by solely looking at the numbers, one would probably assume that a player with .833 OBP is better than a player with a .398 OBP. However, it is very much possible for a player to achieve a .833 OBP just by getting on-base 5 out of 6 times and more likely than not, the player probably got a .833 OBP in this fashion. However, in order for a player to get a .398 OBP, one would need to bat at least 161 times, which shows a high degree of skill since this percentage is sustained upon a high amount of batting opportunities.

In this paper, we will look at which OBP's show the highest amount of true ability p, where a higher true ability indicating a better player. The way that we will approach this problem is by turning real world baseball data into models by using random variables and probability distribution functions of these random variables. These probability density functions would retain the important details that the real world data contains. A random variable X is a variable whose possible values are outcomes of random events. We use an uppercase letter X to refer to the random variable itself and we use a lowercase letter x to refer to the possible value that the random variable can take on. Probability density functions are real valued functions $f_{x_1,x_2,...,x_n}(x_1, x_2, ..., x_n)$ such that $f_{x_1,x_2,...,x_n}(x_1, x_2, ..., x_n) > 0$ for all random variables $X_1, X_2, ..., X_n$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{x_1,x_2,...,x_n}(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n = 1$. With probability density functions, we can create conditional distribution functions which are probability density functions with extra information given with one or more of the variables in the original distribution function. We can also create marginal distribution functions which are probability density functions that isolate one of the variables through aggregating the other variables. The models presented will use conditional distributions and marginal distributions to fit the real world data in an appropriate way.

Sections 2 and 3 will present two different models to approach the same problem. The goal is to create such a model that if we made a simulation out of this model, we would obtain fake data that would be reasonably similar to the real world baseball data. Section 4 will present the necessary methods to calculate the posterior true ability P given the OBP information as well as the information that we created with both our models in sections 2 and 3. Section 5 will give the results of our calculations as well as analyze some of the results' highlights. All of the major computations were handled by R, a statistical program. The data that we used was taken from *www.baseballreference.com* where we recorded the offensive statistics of non-pitchers who had at least one at bat during the 2015 and 2016 seasons (which were the two most recent complete data available). 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. There were about 700 players who met this criterion per season so we had about 1400 players' statistics to work with.

II. The First Proposed Model

The first model will use Bayesian statistics to model the true ability of a player given an OBP thorough a formula that measures the true ability in relation to information relevant to on-base percentage. Bayesian statistics studies the probabilities of events occurring given one has relevant information about the event before making the actual calculation. The probability calculated without the information is called the prior probability and the probability calculated using the updated information is called the posterior probability. In this case, we will use plate appearances, times on base, and on-base percentage as the relevant information and we will use true ability as the prior/posterior probability. We will use true ability for as a prior probability for building the model and we will use true ability as a posterior probability for determining which on-base percentages measure the better players' offensive skill.

We will aggregate the OBP formula to $OBP = \frac{\text{times on base}}{\text{plate appearances}}$ where times-on-base $= H + BB + HBP$ and plate appearances $= AB + BB + HBP + SF$ and find the joint distribution of the random variables TOB (times-on-base), P (prior true ability), and PA (plate appearances). We first find the marginal distribution of PA. Then we model the conditional distribution of P given PA. After that, we model the conditional distribution of TOB given P and PA.

i. Plate Appearances

To model the marginal distribution of PA, we first made histograms of the player plate appearances totals for the 2015 and 2016 seasons. Figure 1 presents the histograms of the plate appearances for the 2015 season, 2016 season, and the two seasons combined.

By approximating the histograms with a combination of lines and curves, we created a probability density function for the plate appearances. The reason that we do this task is to convert the plate appearances data from a practical histogram model to a theoretical probability density function model. The probability density function will allow us to make calculations when

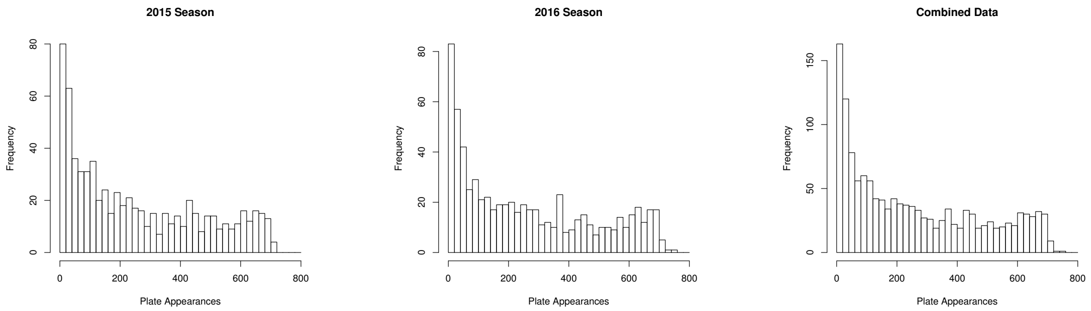

Figure 1: *Histogram of Plate Appearances for the 2015, 2016, and Combined seasons respectively*

this distribution is combined with the other distributions in the model. Notice that all three histograms were relatively similar, which allows us to form a density function that will capture the main properties of each of the histograms. The goal of the density function is to create a function that will approximate the probability of each plate appearance occurring. While one could do this by simply counting the frequencies of a plate appearance occurring and dividing by the number of players, there are a lot of variance with each of the probabilities since the probabilities would vary a lot from year to year, thus making for an unreliable model. Making a probability density function using lines and curves will allow us to retain the general pattern that is observed with the baseball data over the course of multiple seasons, rather than baseball data just for a particular year. With turning a probability distribution function into a function that approximates the probabilities of individual plate appearances occurring, we will use the fact that $P(PA = pa) = P(pa \leq PA < pa + 1)$. This is reasonable to do since plate appearances are only integers and partitioning the domain of f(x), which is the real line, into discrete parts allows us to have the distribution $d(pa)$ to have the domain consisting of integers.

$$d(pa) \equiv \int_{pa}^{pa+1} f(t)dt \tag{1}$$

where f is the density function

$$f(x) = \begin{cases} 0.004779146e^{-.014285714x} + .001013758 & \text{if } 0 \leq x < 280 \\ 0.00083273 & \text{if } 280 \leq x < 600 \\ 0.001122375 & \text{if } 600 \leq x < 700 \\ -0.00000501319544x + 0.0038100285 & \text{if } 700 \leq x < 760 \end{cases}$$

Even though one could theoretically have an unlimited number of plate appearances in a season, in practice, none of the plate appearances have surpassed 760, so for the sake of this model, we assume that $x < 760$.

ii. True Ability

To model the conditional distribution of prior true ability P of a player getting on-base given only PA and nothing else, we turned to the Beta family of distributions. The Beta distribution is a probability density function such that X is between 0 and 1 and α and β are greater than 0. With this distribution, $f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(\alpha) = (\alpha - 1)! = \int_0^\infty x^{\alpha-1}e^{-x}dx$. Frey (2007) found that one can find the true ability for batting average by modeling the mean of ability ($\mu(a)$) given at bats and the one of the parameters of variability of ability (c) given at bats. Then one can set $\mu = \frac{\alpha}{\alpha+\beta}$ and $c = \alpha + \beta$ since these two parameters are sufficient to create a beta distribution model. We can do likewise for OBP by substituting at bats with plate appearances. One can find μ by doing a lowess fit of a plot of on-base percentages versus plate appearances and creating the equation of the lowess fit line similar to what we did with the plate appearances model. Figure 2 contains the lowess fit plot. This equation is meant to be an approximation of the lowess fit in Figure 2.

From this process it follows that

$$\mu(pa) = \begin{cases} 0.000261494x + .253448276 & \text{if } pa \leq 194 \\ 0.000078736x + .288965517 & \text{if } pa > 194 \end{cases}$$

With finding c, we have to account for the fact that better players play more. As a result of this fact, the distribution of p for players with a higher number plate appearances will be stochastically larger than those with a lower number of plate appearances. What Frey (2007) did for finding

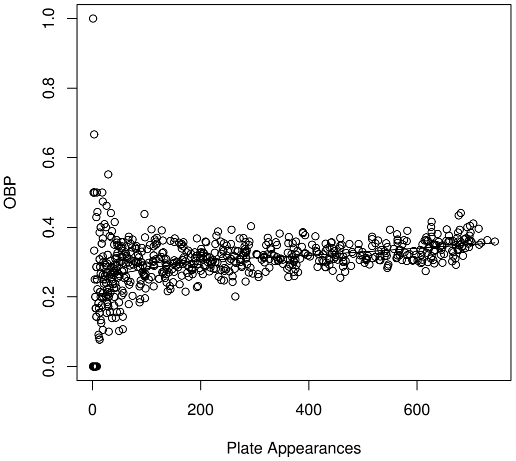

Figure 2: *Plot of Plate Appearances vs OBP's for the 2016 season. The extra line is a lowess fit.*

the variability of batting average was to first plot the variability of batting average as a function of at bats. For at bat bins [290,310), [310,330),...[690,710), he computed the within-bin standard deviations and plotted them as a function of at bats, thus creating a scatter plot of points. We can do likewise for finding the true ability for OBP. Once one has made the plot, one can find c by using the function

$$var(\frac{tob}{pa}) = \frac{1}{pa}\mu(pa) + (1 - \frac{1}{pa})(\frac{c\mu(pa)^2 + \mu(pa)}{c+1}) - \mu(pa)^2$$

. This variance function will act as contours on the graph of standard deviations with varying curves such that every point on a curve will have the same value of c with the locations of the curves changing as a function of c (i.e., with this function, as c increases or decreases the curve will go down or up relative to the points on the graph respectively). All we need to do is to find the curve such that half of the points fall above the curve and half the points fall below it and mark the curve's corresponding c value. After doing this procedure, we end up having c being equal to 625. Figure 3 shows the standard deviation versus plate appearances plot with the middle curve being best representation the variability in the true ability. At this point, we have established a relationship between the prior true ability and plate appearances.

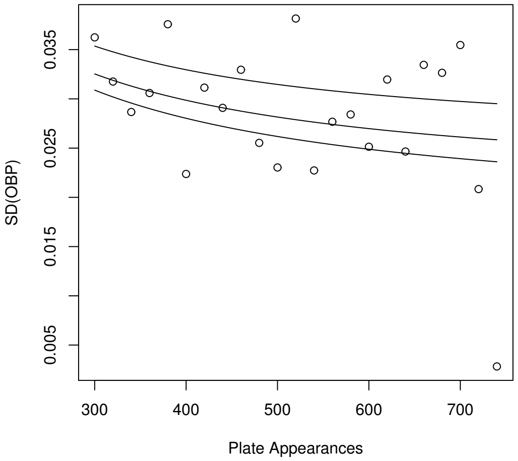

Figure 3: *Empirical bin-by-bin standard deviations for OBP. In the top, middle, and bottom curve, c=400, 625 and 900 respectively. We will use c=625 for our model.*

iii. Times on Base

Finally, we need to model the distribution of TOB given PA and P. The most natural model for this would be to see if TOB were distributed as Binomial(pa,p). The binomial distribution $B(n,p)$ is a probability mass function (probability distribution function for discrete random variables) that models the number of successes or failures of n independent identical trials where each trial only has an outcome of a success with probability p or failure with probability $1-p$. The probability mass function for this distribution is $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$ where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ and n is an integer between 0 and n inclusively. In this model, we are trying to see if we are use each plate appearance as a trial and the prior true ability as the probability of getting on-base since we have found a model for plate appearances and prior true ability in the earlier sections. Frey (2007) showed that one can see if hits (h) were Binomial(at bats(a), prior true ability(p)) by selecting players with at least 300 at bats for the same team for two consecutive years and seeing if

$$\frac{(\frac{h_1}{a_1} - \frac{h_2}{a_2})}{\sqrt{\hat{p}(1-\hat{p})(a_1^{-1} + a_2^{-1})}} \sim N(0,1)$$

where h_1, h_2, a_1, a_2 are the hits and at bats of two consecutive seasons respectively and $\hat{p} = \frac{h_1+h_2}{a_1+a_2}$. The premise behind picking players' data from two consecutive seasons is the fact that a player's skill isn't going to change very much over this time frame which makes the trials more "independent" and "identical". We can do likewise for OBP by substituting a with pa and h with tob. After carrying out this process by looking at a histogram and calculating the mean and standard deviation, it turns out that the times on base does follow a binomial distribution. At this point, we have established a relationship between TOB with P and PA.

iv. Checking the model

After completing our model, we decided to create a simulation of 700 players using the model and compare its scatter plot of OBP to plate appearances to that of the real data. For each player, we first found a random value of pa subject to its probability density function, then found a random value of p using its conditional Beta distribution with the random value of pa, and then found a random number of tob by picking a random number from it's corresponding binomial distribution with the random value of pa and p. Then we computed the OBP for each simulated player using the random values of pa and tob. Then we made a scatter plot with the players' OBP and plate appearances to compare it to the scatter plot of the original data. Figure 4 shows the scatter plots of the data alongside with the original data. Notice that the simulated data plot is very similar to that of the real data so this model captures the essence of the baseball data.

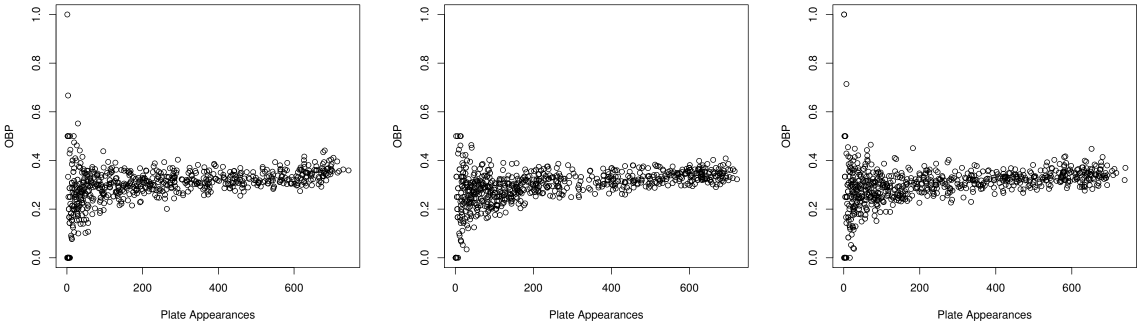

Figure 4: *Plate Appearances by OBP Plots. The left-most one is with the 2016 data and the other two are created by simulation with the first model.*

III. The Second Proposed Model

The second model will use a more specific model to better fit the OBP data and will use a simulation to find the true ability given the information presented in the model. We will look at each of the numerator components of OBP and find the joint distribution of the variables H, BB, HBP, P, and PA. We try to find the marginal distribution of PA, conditional distributions of P_H,P_{BB}, and P_{HBP} given PA and then the conditional distribution of H, BB, and HBP given P_H, P_{BB}, P_{HBP} and PA.

i. Plate Appearances

In this model, the model for PA is the exact same model as the model used for PA in section 2.1 so we omit the details.

ii. True Ability

Unlike section 2.2 where we only needed to find the prior true ability of a player getting on-base given only PA, in this model, we need to find the prior true abilities for a player to get a hit (P_H), walk (P_{BB}), hit-by-pitch (P_{HBP}), and out ($1 - P_H - P_{BB} - P_{HBP}$) given only PA. Since we are looking at probabilities such that the sum of probabilities of the four components adds up to 1 (reaching on-base by fielders' choice or by an error is counted as an out according to the definition of on-base percentage), the most natural distribution to turn to would be the Dirichlet distribution. The Dirichlet distribution is the multi-variable equivalent of the Beta distribution where $n = 2$ of the Dirichlet distribution is the Beta distribution. The probability density function of the Dirichlet distribution is $f_{x_1,x_2,...,x_n}(x_1, x_2, ..., x_n) = \frac{1}{B(\alpha_1,\alpha_2,...,\alpha_n)}\Pi_{i=1}^{n} x_i^{\alpha_i - 1}$ where $B(\alpha_1, \alpha_2, ..., \alpha_n) = \frac{\Pi_{i=1}^{n}\Gamma(\alpha_i)}{\Gamma(\Sigma_{i=1}^{n}\alpha_i)}$ We can use the same process of finding the distribution as in Section 2.2 by finding the means of the different abilities and the variability of the abilities. In this case, we set $\mu_1 = \frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3+\alpha_4}$, $\mu_2 = \frac{\alpha_2}{\alpha_1+\alpha_2+\alpha_3+\alpha_4}$, $\mu_3 = \frac{\alpha_3}{\alpha_1+\alpha_2+\alpha_3+\alpha_4}$, and $\mu_4 = 1 - \mu_1 - \mu_2 - \mu_3$ where μ_1,μ_2,μ_3, and μ_4 are the means corresponding to hits, walks, hit-by-pitches and outs respectively. In this case, c will be equal to $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$. Using the same process as what we used in section 2.2, it follows that

$$\mu_1(pa) = \begin{cases} 0.000125969x + .187209302 & \text{if } pa \leq 230 \\ 0.000070543x + .2 & \text{if } pa > 230 \end{cases}$$

$$\mu_2(pa) = \begin{cases} 0.000217241x + .04137931 & \text{if } pa \leq 167 \\ 0.000015517x + .075172414 & \text{if } pa > 167 \end{cases}$$

$$\mu_3(pa) = \begin{cases} 0.000030864x & \text{if } pa \leq 232 \\ 0.000003292x + .006419753 & \text{if } pa > 232 \end{cases}$$

In this case, μ_4 is omitted since it is just a function of the other 3 functions and carries little benefit in the model. Since the process of finding c is the exact same process as in section 2.2, we omit the details with the process and say that we will use $c = 625$ as was used in the previous model. Now we have a connection between the prior true ability with plate appearances.

iii. Hits, Walks, Hit-By-Pitches

Finally, we need to model the distribution of hits (H), walks (BB), and hit-by-pitches (HBP) given PA, P_H, P_{BB}, and P_{HBP}. Since we are modeling plate appearances as independent events that only lead to outcomes H, BB, HBP, and Outs, the most natural distribution to turn to would be the multinomial distribution. The multinomial distribution is the multi-variable analogue of the binomial distribution. Like the binomial distribution, the multinomial distribution measures the outcomes of n identical, independent trials. However, unlike the binomial distribution where each trial only has outcomes success and failure with probabilities p and $1 - p$ respectively, the multinomial distribution trials have outcomes x_1 with probability p_1, x_2 with probability p_2,...,x_k with probability p_k where x_i's are integers with $\Sigma_{i=1}^{k} x_i = n$ and p_i's are probabilities such that $\Sigma_{i=1}^{k} p_i = 1$. The probability mass function for this distribution is $\frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} p^2 x_2 ... p_k^{x_k}$. Notice that for $k = 2$, the multinomial distribution is the binomial distribution. However, in order for H, BB and HBP to be modeled this way, we need to have hits to be modeled as

$$\frac{(\frac{h_1}{pa_1} - \frac{h_2}{pa_2})}{\sqrt{\hat{p_h}(1 - \hat{p_h})(pa_1^{-1} + pa_2^{-1})}} \sim N(0,1)$$

where $\hat{p_h} = \frac{h_1+h_2}{pa_1+pa_2}$. Notice that this modeled function is analogous to the one in section 2.3. We would also need walks and hit-by-pitches to be verified in a similar fashion. The way that one verifies this is the same as the technique used in the first model so we omit the details. After carrying out this process by looking at the histograms of H, BB, and HBP and calculating their means and standard deviations, it turns out that each of the formulas are normal, which means that hits, walks and hit-by-pitches follows as a multinomial distribution.

iv. Checking the Model

After completing our model, we decided to create a simulation of 700 players using the model and compare its scatter plot of OBP to plate appearances to that of the real data. For each player, we first found pa using its probability density function, then found a random value of p_h, p_{bb}, p_{hbp}, and p_{out} using the Dirichlet distribution, and then found the number of hits, walks, hit-by-pitches and outs by picking a random number from it's corresponding multinomial distribution. We then computed each simulated player's OBP by the formula $\frac{\text{hits+walks+hit-by-pitches}}{\text{plate appearances}}$ and made a scatter plot similar to that in the first model. Figure 5 shows the scatter plots of the data alongside with the real data. Notice again that the simulated data plot is very similar to that of the real data so this model captures the essence of the baseball data.

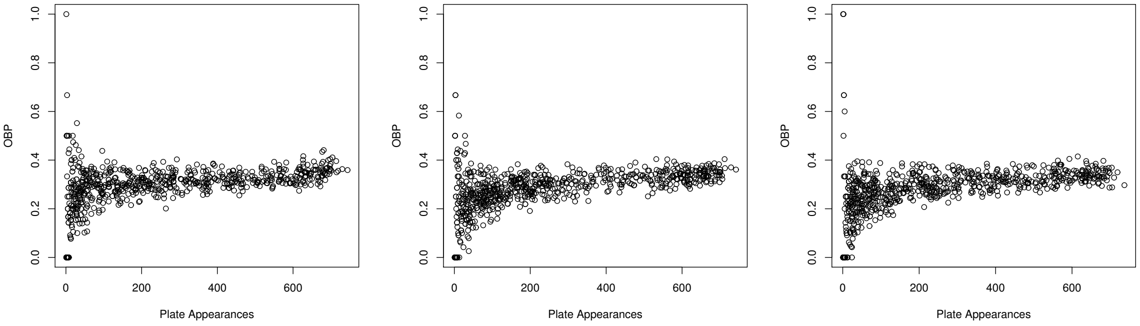

Figure 5: *Plate Appearances by OBP Plots. The left-most one is with the 2016 data and the other two are created by simulation with the second model.*

IV. Finding the Posterior Distribution

i. Using the First Model

In order to find the posterior true ability P based on OBP alone, we need to find the true ability given OBP. We can use the fact that P(prior true ability) $\sim Beta(\alpha, \beta)$ and $X|p$(prior true ability) $\sim Binomial(n, p) \implies P$(posterior true ability)$|x \sim Beta(\alpha + x, \beta + n - x)$ where $A \sim b$ means that random variable A has distribution b. Frey (2007) found that for batting averages, one can find the posterior distribution conditional on a given batting average by listing all of the ways (a_1,h_1), (a_2,h_2),...(a_k,h_k) a batting average b can occur and summing a mixture of their corresponding distributions, which, in this case, each one is a Beta distribution. Frey (2007) also showed that the posterior p given a batting average b is the mixture

$$\sum_{i=1}^{K} \left(\frac{P(h_i \text{ hits and } a_i \text{ at bats})}{\sum_{j=1}^{K} P(h_j \text{ hits and } a_j \text{ at bats})} \right) \times \text{Beta}(c\mu(a_i) + h_i, c(1 - \mu(a_i)) + a_i - h_i) \tag{2}$$

where

$$P(h_i \text{ hits and } a_i \text{ at bats}) = d(a) \left(\frac{\Gamma(a+1)\Gamma(c)\Gamma(h + c\mu(a))\Gamma(a - h + c(1 - \mu(a)))}{\Gamma(h+1)\Gamma(a+1-h)\Gamma(c\mu(a))\Gamma(c(1-\mu(a)))\Gamma(a+c)} \right). \tag{3}$$

In equation (3), $d(a)$ is defined in equation (1). We can do the same thing with OBP by substituting h with tob and a with pa. Due to the similarity the formula is retained. Using the formula, we can find the posterior p for each OBP. However, we only considered the OBP of those such that $P(OBP = obp) > \frac{1}{10{,}000{,}000}$ since other OBP's are pretty much impossible to see in practice.

ii. Using the Second Model

Unfortunately, with the second model, one cannot use the same Bayesian techniques for the second model due to too much storage needed to accomplish this task. Doing this kind of calculation would require a 14 billion entry matrix and unfortunately R can only hold 2 billion at most. Instead, we simulated some fake baseball data using the second model similar to the simulation used to test the model. We made 10,000,000 fake players each with their own h, bb, hbp, pa, p_h,

p_{bb}, and p_{hbp}, similar to how we checked the second model by simulating fake players in section 3.4. We then computed each player's OBP, and summed each player's p_h, p_{bb}, and p_{hbp} to find each player's overall true ability. We then did an aggregate mean on the true ability by OBP. This will give us the posterior mean P for the players. One thing to take into account with using this model is that we will need to watch for simulation errors. However with 10,000,000 players being simulated, the issue of simulation error is probably negligible since the sample size is really high.

V. Results

i. Using the First Model

Table 1 gives the posterior true abilities and posterior standard deviations for the corresponding top ten OBP's and the corresponding bottom ten OBP's using the first model. The posterior true ability tells us what the real value of a player's probability of getting on-base given the player's documented OBP. One interesting observation is that the top 10 posterior true abilities had OBP's that were all above .35 with posterior true abilities around .34-.36 and the bottom 10 posterior true abilities had OBP's that were less than .1 with posterior true abilities around .25. This observation makes intuitive sense since even without applying the analysis of this paper, among the laypersons, a player with an OBP of more than .35 is considered to be talented and a player with an OBP of less than .1 would be considered to be terrible. Another observation to make is that the top 10 posterior true abilities needed at least 100 plate appearances in order for such values to exist. This also lines up with intuition as better players will have more opportunities to contribute offensively. One more interesting thing to notice is that the bottom 10 posterior true abilities require more than 100 plate appearances which suggests that only having a high number of plate appearances isn't sufficient for being a good player. An interesting comparison to make is with Frey (2007)'s data for batting average. While his top 10 posterior true abilities contained a similar pattern of having batting averages above .3, his bottom 10 posterior true abilities only contained batting averages that can be achieved with a low number of at bats (Frey (2007) lacks low batting averages for his bottom 10 posterior true abilities other than .000.). The posterior standard deviations show how much a player's posterior true ability can vary. The fact that the posterior deviations are between .015 and .04 for the top 10 OBP's suggest that there is a reasonable amount of variation with posterior true abilities.

OBP	Posterior True Ability *P*	Posterior SD for *P*	*OBP*	Posterior True Ability *P*	Posterior SD for *P*
0.428	0.362	0.031	0.022	0.249	0.019
0.401	0.356	0.026	0.023	0.249	0.019
0.399	0.355	0.026	0.024	0.249	0.019
0.445	0.349	0.038	0.025	0.250	0.019
0.416	0.347	0.038	0.039	0.250	0.019
0.384	0.346	0.029	0.026	0.250	0.019
0.398	0.345	0.033	0.041	0.250	0.019
0.402	0.345	0.035	0.027	0.250	0.019
0.374	0.343	0.026	0.028	0.250	0.019
0.376	0.343	0.027	0.052	0.250	0.019

Table 1: *OBP corresponding to the highest and lowest posterior true abilities p. We only consider OBP with at least a one in 10,000,000 chance of occurring.*

ii. Using the Second Model

Table 2 gives posterior true abilities, posterior standard deviations, and average plate appearances for the corresponding top ten OBP's and the corresponding bottom ten OBP's using the second model (We were able to find the posterior mean for plate appearances with this model). A similar analysis regarding the usage of the first model can be said with the second model since the general patterns of both tables are relatively similar (i.e., the top 10 OBP's are above .35 and the bottom 10 OBP's are mostly below .1 as well as the relatively similar posterior standard deviations). Unlike the posterior true abilities in the first model, the second model have lower posterior true abilities for the bottom 10 OBP. Another thing to notice is that the average plate appearances for the top 10 OBP's are around 400 to 500 while the bottom ten OBP's are less than 65. One standout contrast though is that the top OBP's for both models are similar while the bottom ten OBP's are different. A standout figure that appears with this model is the OBP value of .889. Though this value is significantly greater than .35, it can easily be achieved with 8 times-on-base out of 9 plate appearances which does not really indicate the batter is a good player. This example shows that only having a high valued OBP isn't sufficient enough for determining a player's offensive capacities.

VI. Conclusion

When it comes to finding the best batters, it isn't always the case that a batter with a higher OBP alone implies a better hitting. We need a player to have both a high number of plate appearances

OBP	*P*	SD for *P*	Posterior Mean for *PA*	*OBP*	*P*	SD for *P*	Posterior Mean for *PA*
0.445	0.366	0.035	449.5	0.052	0.217	0.000	58.0
0.428	0.365	0.029	493.9	0.022	0.219	0.017	45.5
0.401	0.357	0.024	519.6	0.016	0.221	0.000	61.0
0.399	0.357	0.024	521.4	0.889	0.225	0.006	9.0
0.416	0.353	0.036	458.9	0.026	0.226	0.014	38.4
0.402	0.350	0.033	474.3	0.028	0.226	0.021	36.0
0.398	0.350	0.031	481.4	0.058	0.228	0.016	54.1
0.384	0.348	0.027	503.7	0.024	0.228	0.017	41.6
0.403	0.344	0.040	446.5	0.000	0.229	0.017	4.1
0.376	0.344	0.025	500.0	0.047	0.229	0.015	43.0

Table 2: *OBP corresponding to the highest and lowest posterior true abilities p.*

and a high OBP. Both of the models used show that the top OBP's in terms of posterior true abilities are around the .4 values, which is a reasonable conclusion to make. However, one of strange results is that the bottom 10 OBP's for both models yields values that require a reasonably high number of plate appearances as well. This result is counter-intuitive to the notion that worse players tend to receive less opportunities to go to the plate as they are less reliable for contributing to the team's offense. It also runs a bit contradictory to Frey (2007)'s results on the optimal batting average. Nevertheless, this analysis shows us that we can make a model out of limited real data and apply simulations along with Bayesian analysis to make a practical conclusion. With knowing only a subset of the full baseball data, we can conclude that a good baseball player will have an OBP of .445, .421, .401 and other values that are high in magnitude and are only achievable through a high number of plate appearances.

VII. Acknowledgements

I would like to acknowledge Dr. Jesse Frey for guiding me with thoughts and insights as I worked out the details of this paper as well as answering questions that I had regarding the work that he has done with this field of study.

References

[1] Jesse Frey "Is an .833 Hitter Better Than a .338 Hitter?" *The American Statistician*, vol. 61, no. 2, May 2007, pp. 105–111.

[2] Lewis M. 2003. *Moneyball: The Art of Winning an Unfair Game.* New York: W. W. Norton.

[3] "MLB Stats, Scores, History, & Records." *Baseball-Reference.com*, www.baseball-reference.com/.