A Look Into Empirical Bayes

Jas Kainth

06/10/2020

Contents

Introduction

In this paper, we will take a look into Empirical Bayes. Empirical Bayesian models are a form of approximation methods to more exact methods. This leads to much controversy, however, they work well when there are large datasets. Empirical Bayes offer "shortcuts" to regular Bayesian statistics and are often used when we need to perform estimations thousands of times. Full Bayesian methods are useful when performance is less important than accuracy (which is the case for many studies). We will look at Empirical Bayesian methods which are introduced in the *Introduction to Empirical Bayes* textbook by *David Robinson*. However, rather than just learning the concepts, we will compare them to the frequentist inference and discuss which method gives us a better result. For the examples, we will be using the same dataset in the textbook which is a baseball dataset from the `Lahman` package. All of the code will be available in the .Rmd file. The main statistic which we will be taking a look at is batting average (number of hits/number of at-bats) which means we will be taking a look at the binomial distribution.

What is Empirical Bayes?

Empirical Bayes are a method of estimation for a hierarchical Bayesian model. the usefulness of these estimation methods can be seen when discussing the Bayes' Theorem. Let us assume that our observed data $y = \{y_1, y_2, \ldots, y_n\}$ are generated from a set of parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$. Then using the general form of Bayes' Theorem Posterior Probability $= \frac{\text{Likelihood} \cdot \text{Prior Probability}}{\text{Marginal Likelihood}}$, where in our model the posterior is $p(\theta|y)$, the likelihood is $p(y|\theta)$, the prior is $p(\theta)$ and the marginal likelihood is $p(y)$

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta) \cdot p(\theta)}{p(y)} \\ &= \frac{p(y|\theta) \cdot p(\theta)}{\int_\Theta p(y|\gamma) \cdot p(\gamma) d\gamma} \end{aligned}$$

Where

$$p(\theta) = \int p(\theta|\eta) \cdot p(\eta) d\eta$$

using a hierarchical model. The problem arises when trying to calculate either of the integrals (especially the first one). The integral must be evaluated by numerical methods. Stochastic methods (like MCMC) may also be used.

Empirical Bayes can give you quick estimates for questions rather than going through those long and tedious calculations. For example, assume you want to sign a new player to your team, and you are indecisive on two players; player 1 who has 10 AB (At bats) and 4 H (hits) and player 2 who has 200 AB and 60 H. Due to the difference in the sample sizes, this question may be difficult to answer. Player 1 has a batting average of 0.4 and player 2 has a batting average of 0.3 we can be a lot more sure that player 2 is closer to their actual respective average than player 1. This is where we can use our *prior* information to help us.

Prior Distribution

A Prior Distribution is a constraint that we can add to our model which takes our prior belief of the data into account. For instance, if we know that most batting averages are around 0.21 to 0.36 then we can set our prior to convey that message to our model. However, suppose we had no prior knowledge on the batting averages, then we can set a prior that conveys no information, like $Unif(0, 1)$. There are multiple different priors that we can set, but for this, we will be using the Beta distribution.

An Introduction to the Beta Distribution

The Beta distribution is used as a prior for when we want to set a prior on a probability. Looking at baseball hits, our batting average, $\mu = \frac{H}{AB}$, is a probability and that is why the beta distribution will be extremely helpful. Below, are some examples of Beta distributions.

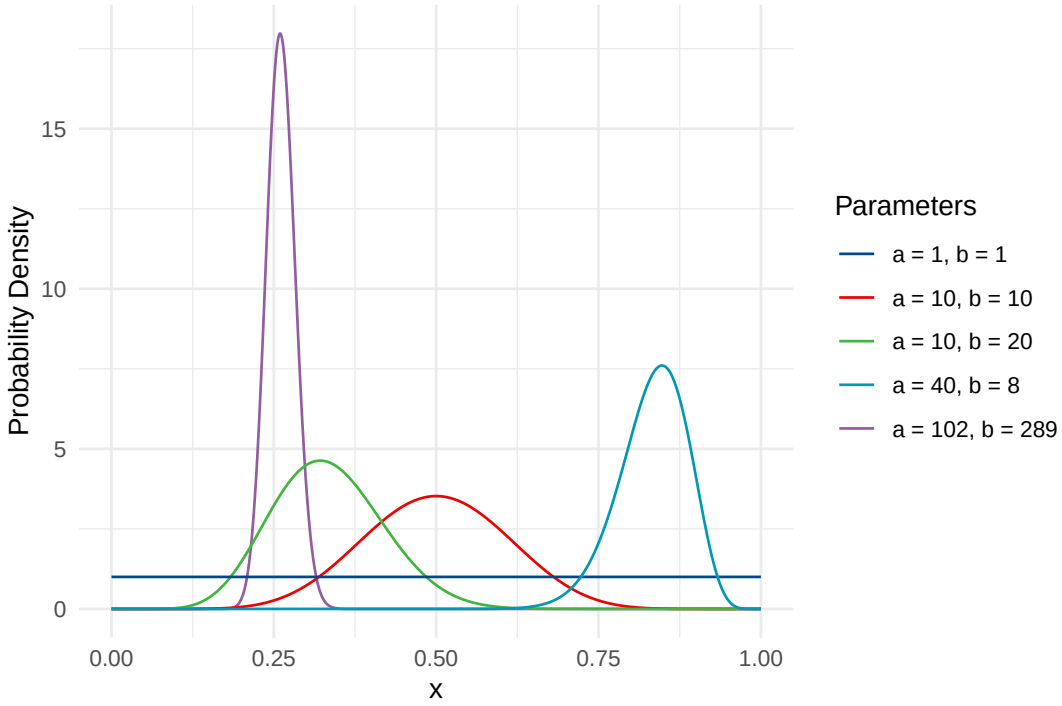

Figure 1: Beta Distribution with Different Parameter Values

When calculation the posterior density, with our choice of prior it can make the calculations tedious (and sometimes not possible analytically). However, there is a class of priors, called conjugate priors. For this class, when we multiply the prior to the likelihood, the posterior is the same distribution as the prior. Note the parameters do not have to be the same only the distribution family. Our likelihood, for the baseball batting average, is a binomial and if we set the prior to a beta distribution, then the posterior is also a beta distribution.

Proof that the beta distribution is a conjugate prior to the binomial distribution.

Proof. First, we note that the Beta distribution takes on the following form

$$\text{Beta(a,b)} = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\text{B(a,b)}} \propto \theta^{a-1}(1-\theta)^{b-1}$$

where $\text{B(a,b)} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, so is therefore indepdent of the parameter θ. Now, we set our prior to the prior distribution to the beta distribution and set the likelihood to the binomial distribution. Then, our posterior is

$$\begin{aligned} p\left(\theta|y\right) &= \frac{p\left(y|\theta\right) \cdot p\left(\theta\right)}{p(y)} \\ &\propto p\left(y|\theta\right) \cdot p\left(\theta\right) \\ &\propto \theta^{x}(1-\theta)^{N-x} \cdot \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{x+a-1}(1-\theta)^{N+b-x-1} \\ &= \theta^{a'-1}(1-\theta)^{b'-1} \end{aligned}$$

where $p(y|\theta) = \binom{N}{x}\theta^x(1-\theta)^{N-x} \propto \theta^x(1-\theta)^{N-x}$, $a' = a + x$ and $b' = N + b - x$. So, we see that the final formula takes on the form of the Beta distribution. We could do the full calculation to show that the constants line up, but since we know that probability distributions are normalized and the prior and likelihood were valid probability distributions, the posterior must also be a valid probability distribution and therefore we can conclude that the posterior is a beta distribution. ■

Given our results above, we also find an interesting relationship for our data. Given a and b are the original parameters, we interpret x as the number of hits from the batter and N is the number of at-bats. This then gives us an easy and convenient way to update our information. The mean of the beta distribution is $E[X] = \frac{a}{a+b}$ so given our interpretation, we find that the new, updated mean of a player is $\frac{\alpha+H}{\alpha+\beta+AB}$ where H is the number of hits and AB is the number of at-bats. (The mode is $\frac{\alpha+H-1}{\alpha+\beta+AB-2}$ but this is close to the mean for large α and β).

However, for this model, it made sense to set the prior as a beta. It should not become a common practice to set the prior to the conjugate prior since the results can sometimes be nonsensical. There are software programs now that can give the posterior even if the prior is not a conjugate prior by using highly accurate and quick estimation techniques like the `INLA` package (Integrated Nested Laplace Approximation).

Finding the parameters for our Prior

To find the parameters of the beta distribution, we will plot the

Table 1: Parameter values for the baseball model

Parameter	Values
Alpha	101.8625
Beta	289.3795

We can see how this distribution looks in Figure 1. Figure 2 shows how well this fits our data.

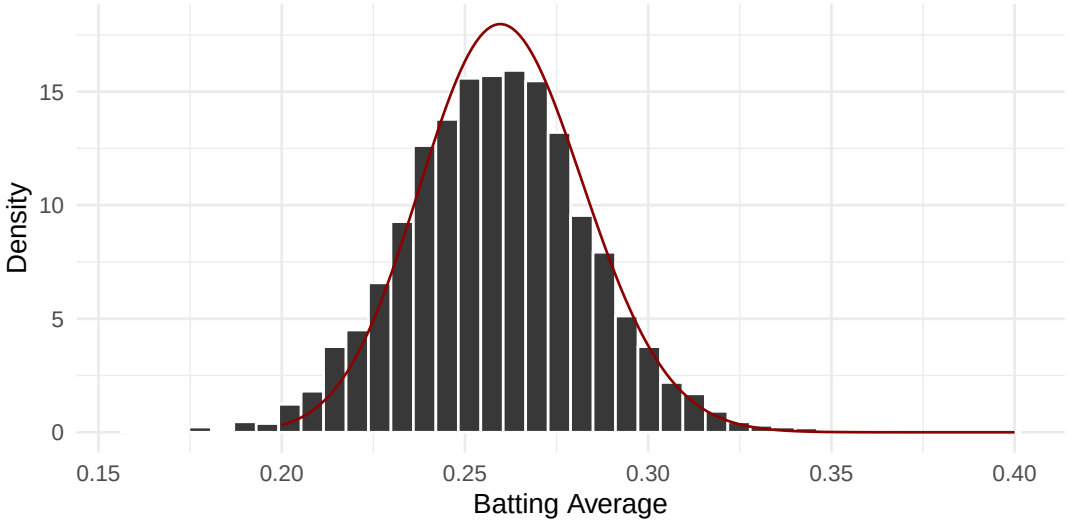

Figure 2: Density Plot of Batting Averages and Fit of Beta Distribution

It seems the beta distribution fits the model pretty well. Of course, this is a prior distribution so it doesn't have to fit it exactly, but it does tend to follow the general trend.

Estimations

Now for our first main difference between the two inference results. We return to the question that was brought up before, who is the better player to pick between player 1 who has 10 AB and 4 H or player 2 who has 200 AB and 60 H. The regular average would say that it is player 1 since they have a higher batting average (although they would hopefully still be a bit skeptical). But, with our prior distribution, we can come up with a much better estimate. We brought up our updated distribution when we are introduced the beta distribution. The average is $\frac{\alpha+H}{\alpha+\beta+AB}$ and now that we have our values of α and β we can go ahead and calculate these averages. For player one, we get $\frac{102+4}{289+102+10} \approx 0.2643$ and for player two we get $\frac{102+60}{289+102+200} \approx 0.2741$. With the updated means from the prior distribution, we see that even though the actual batting average is higher for player 1, the posterior mean is actually higher for player 2.

This brings us to the topic of shrinkage. When there is lacking evidence (in this case a low AB), the adjusted mean will tend to go towards the prior mean. This is shown in Figure 3.

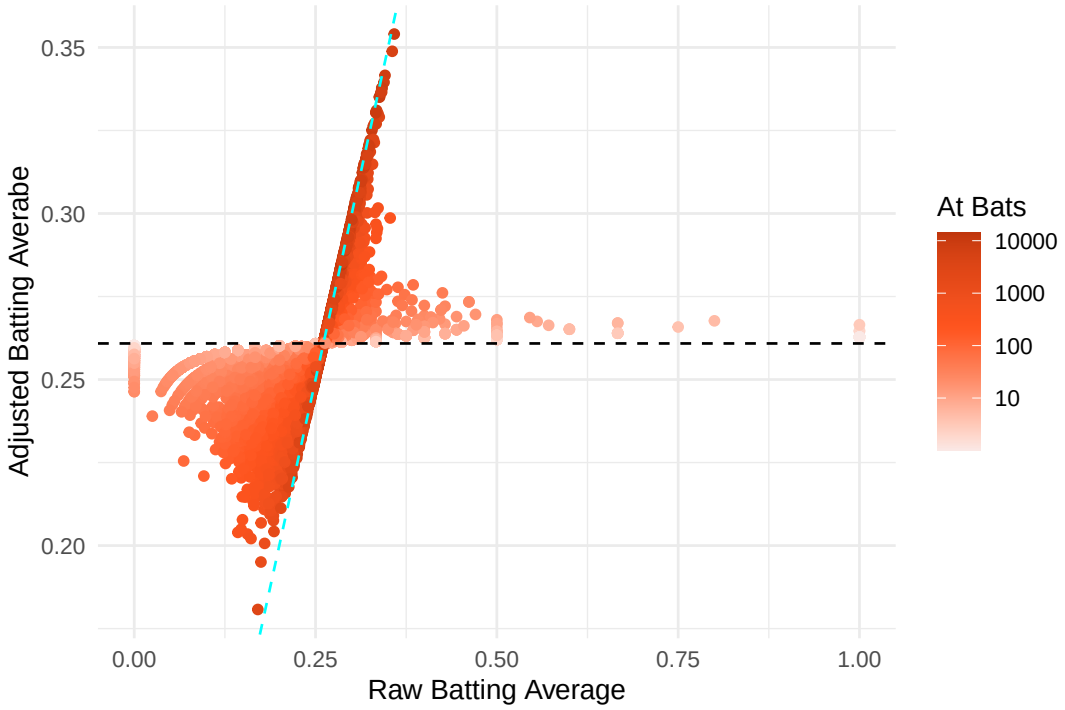

Figure 3: Raw Batting Averages vs Adjusted Batting Averages

The points lined up against the cyan line, which represents $x = y$, is for the points that don't get adjusted a lot. These are the dark orange points, where the AB for each player is high. Then, we can see The black line, which is a straight line at the prior average. The points that are hovering near those points are the light orange points, the ones that don't have a lot of AB. So, we move the points that don't have a lot of evidence to support their average towards the prior mean, whereas if the batter has a lot of AB, then we have enough information to believe their average and therefore they don't move a lot.

Credible Interval vs. Confidence Interval

The difference in the definition of credible interval and confidence interval is small but can lead to a vastly different interval, especially with small sample sizes. Credible intervals incorporate the information we give them model through the prior, whereas the confidence interval is completely based on the data. The main difference between credible intervals and confidence intervals is for a credible interval the bounds are fixed and the estimated parameter is random whereas for a confidence interval the bounds are random and the estimated parameter is fixed. If we take a 95% interval, we create the confidence interval the following way: 95% Confidence Interval $\approx$ Point Estimate $\pm$ 2 $\cdot$ Standard Deviation. Therefore, we say that with large enough trials, we expect this interval to contain the true parameter approximately 95% of the time. For the credible interval, we can choose any interval we want, as long as that interval contains 95% of the probability of the distribution. For example, for a 95% credible interval, we can choose the lower bound to be the 1^{st} quantile and the upper bound to be the 96^{th} quantile. However, it is general practice to create the interval so the lower bound is the $\frac{\alpha}{2}$ quantile and the upper bound is the $100 - \frac{\alpha}{2}$ quantile. So for a 95% credible interval, we use the 2.5 and 97.5 quantiles. The way to interpret this is that with large enough trials, we expect the random parameter to be found in this interval about 95% of the time; a slight difference from the confidence interval interpretation but it lets us interpret intervals more intuitively.

Figure 4 shows the difference between credible intervals and confidence intervals for 20 players. The figure is arranged from the least AB to the most, and we see that the confidence interval is extremely wide for the first player but tends to approach the credible interval as the number of AB increasea. This is because credible intervals incorporate information from the prior, so it is more narrow. Also, there is a larger deviation in the point estimate with lower AB, which is due to shrinkage.

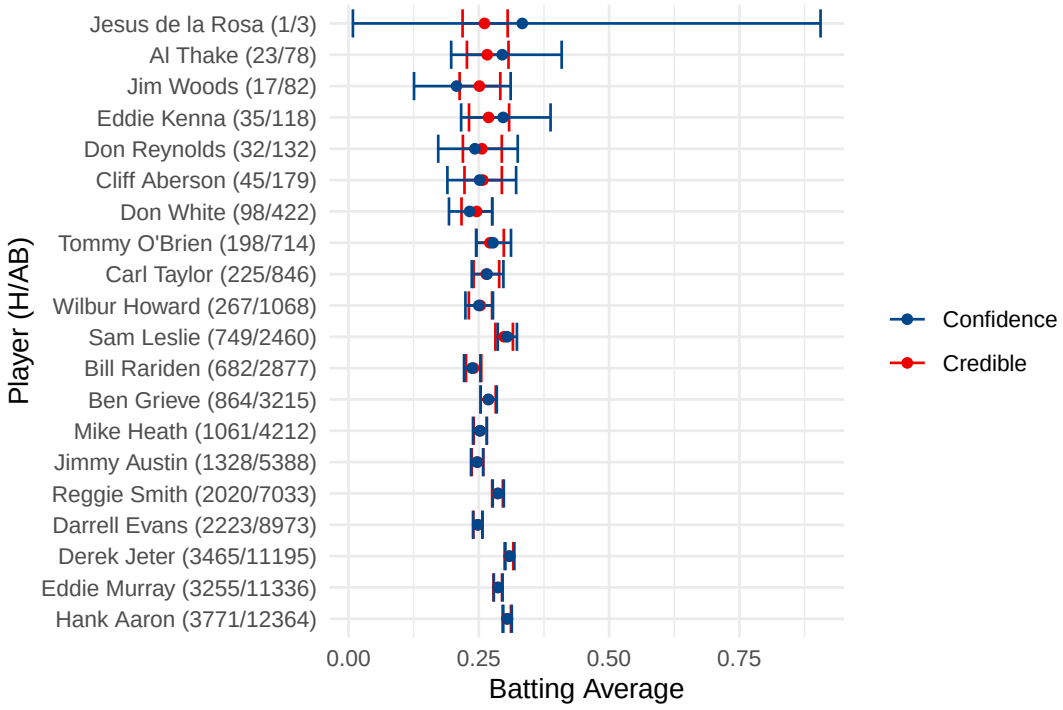

Figure 4: Comparing Credible Interval to Confidence Interval

Acknowledgments

To gain a deeper understanding of Empirical Bayes, I read the textbook Introduction to Empirical Bayes by David Robinson which is a fantastic textbook since it introduces concepts via examples rather than strictly mathematical formulas. If you are interested, it can be purchased *here*.