# Empirical Bayes

## Introduction

*Its approximations become inaccurate when you have only a few observations.*

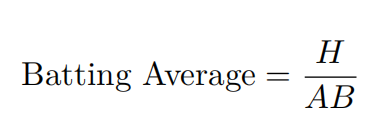
Full Bayesian methods that use **Markov Chain Monte Carlo** (MCMC) are useful when *performance is less important than accuracy*, such as analyzing a scientific study. However, production systems often need to perform estimation in a fraction of a second, and run them thousands or millions of times each day.

**Bayesian statistics** is a way of modeling our prior expectations against the current data observed.

**Empirical Bayes estimation**, where a beta distribution fit on all observations is then used to improve each individually. As long as you have a lot of examples, you don’t need to bring in prior expectations.

## Batting average rate

**Batting average** is the number of hits (H) divided by the number of at-bats (AB):



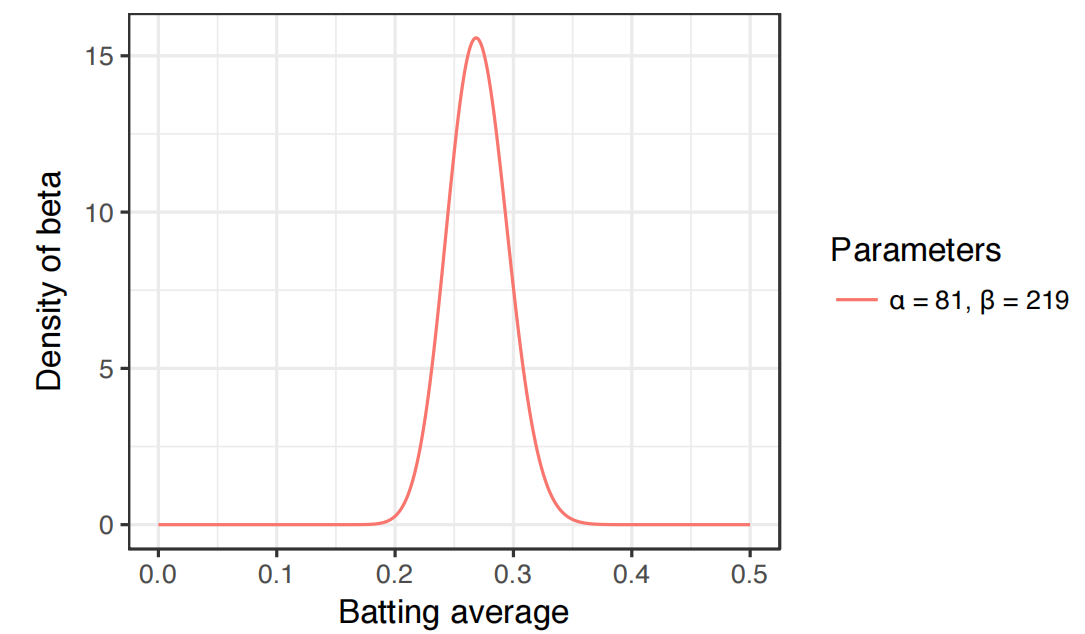
.270 (27%) is considered a typical batting average, while .300 (30%) is considered an excellent one.

The number of hits a player gets out of his at-bats is an example of a **binomial distribution**, which models a count of successes out of a total and the best way to represent the prior expectations is with the beta distribution.

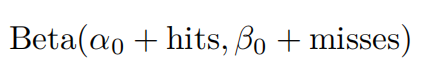
## Beta distribution

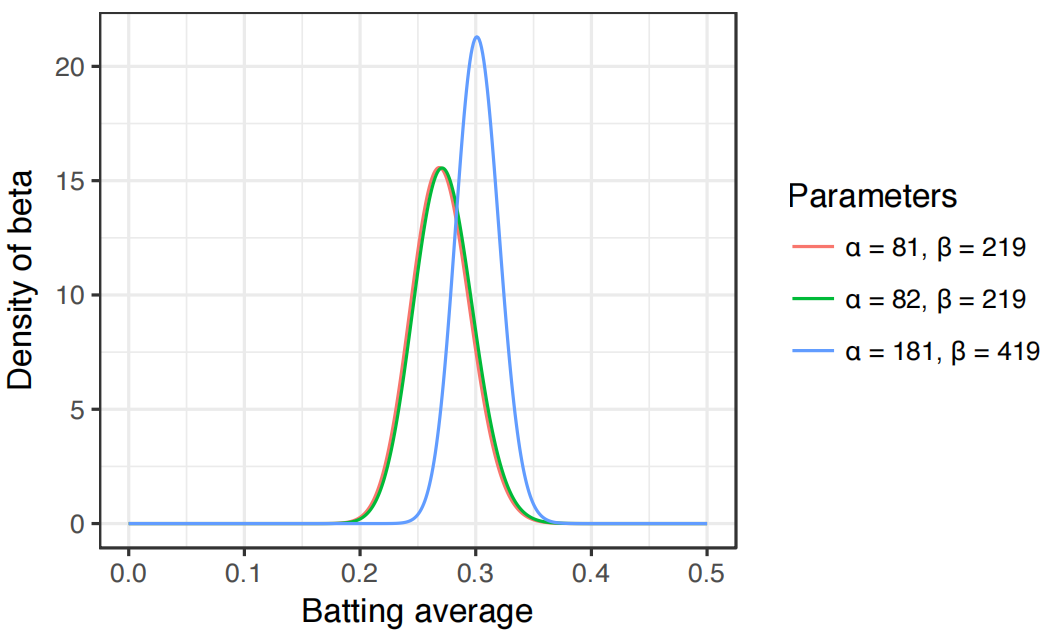
This is the Bayesian philosophy in a nutshell: we start with a prior distribution, see some evidence, then update to a posterior distribution.

The following distribution represents the historical batting average of players.

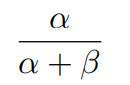


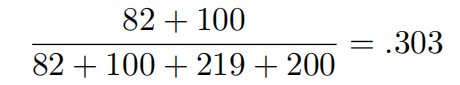
Imagine the player gets a single hit. His record for the season is now “1 hit; 1 at bat.” We have to then update our probabilities- we want to shift this entire curve over just a bit to reflect our new information. According to that our new parameters will be:



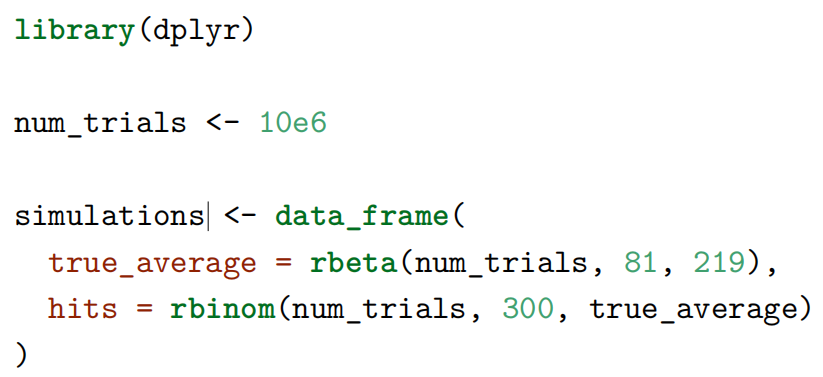


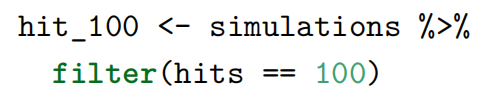
The Expected value of the blue curve can be calculated using the next formula:

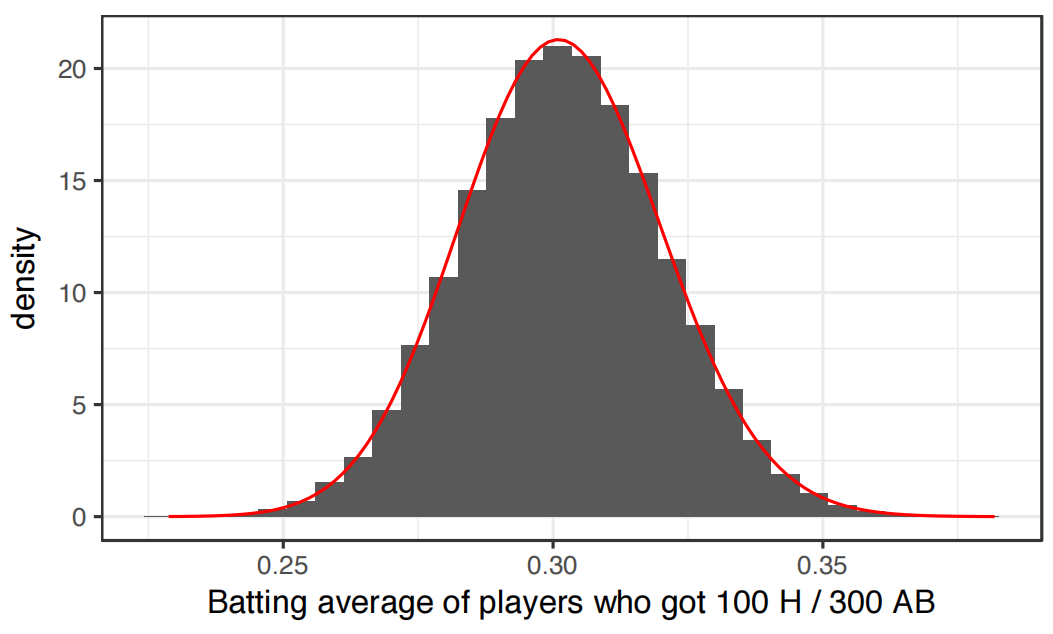




Let’s the this with a simulation.

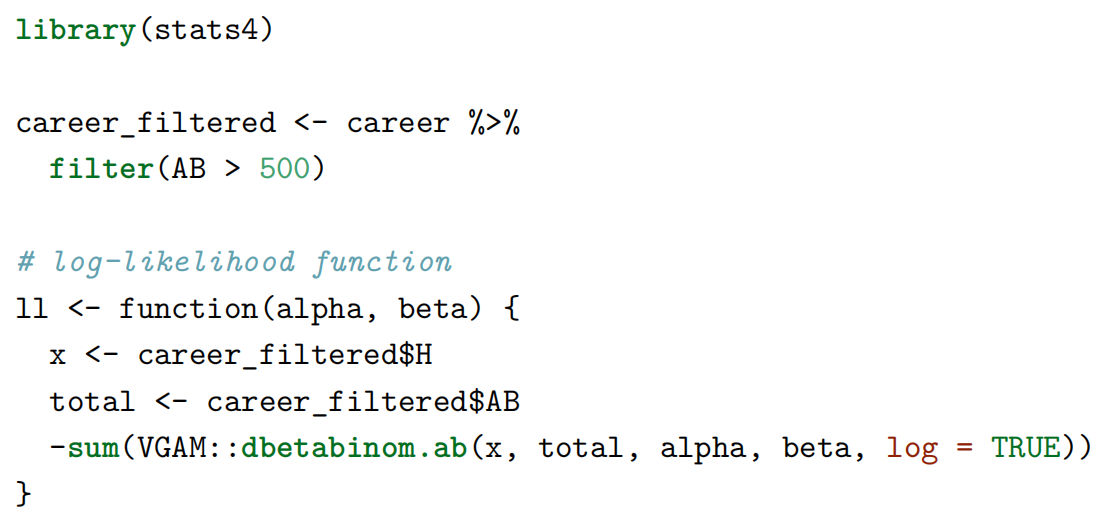


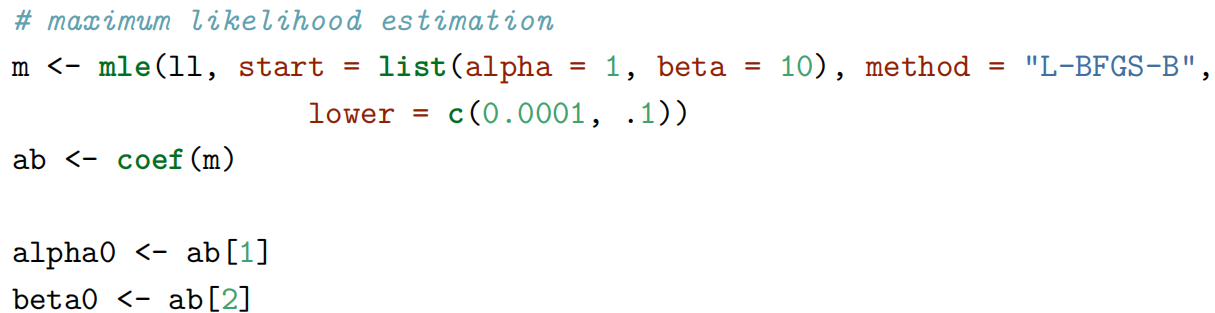


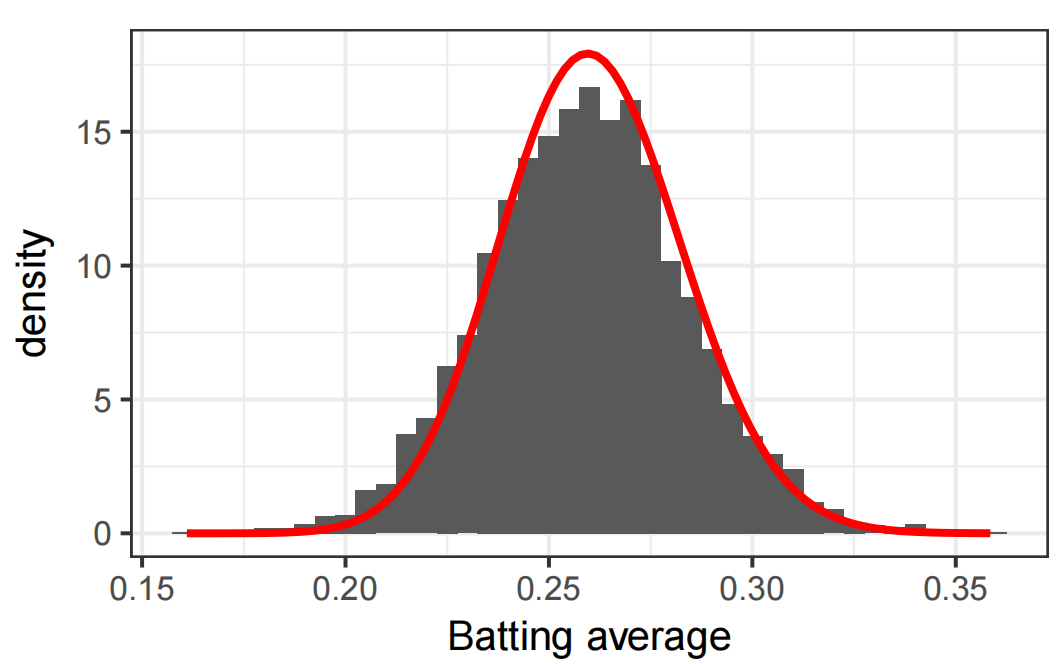


## Empirical Bayesian shrinkage towards a Beta prior

1. Estimate the beta distribution using prior data







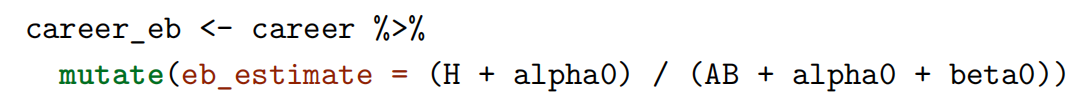
1. Use that distribution as a prior for each individual estimate by adding α0 to the number of hits, and α0 + β0 to the total number of at-bats.

A picture containing chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated



1. Based on Empirical Bayes Estimate we can select what are the best and worst batters?

Table

Description automatically generated

## Computing a credible interval using the empirical Bayes Method

1. Estimate new parameters for each batter.

Text

Description automatically generated

1. Calculate interval of 95% using qbeta function.



Text

Description automatically generated with medium confidence

Chart

Description automatically generated

Chart

Description automatically generated

We can compare these results with the intervals produced by **binom.test**

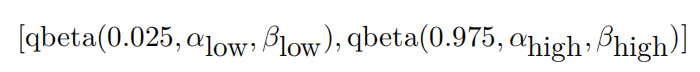
A picture containing chart

Description automatically generated

**binom.test** uses Clopper-Person intervals and it’s more conservative (wider) than Jeffreys one.

Text, table

Description automatically generated



## False discovery rate control

This method solves problems associated with multiple testing.

Despite a batter has an expected value of 0.30 there is probability that the true batting average is bellow that number. If we check the distribution on of Hank Aaron with **pbeta** function with alpha1 and beta1, we will see that he has 17% of having a batting average bellow. This probability is known as **Posterior Error Probability**, or **PEP**

Chart, line chart

Description automatically generated

Notice that crossover point: to have a PEP less than 50%, you need to have a shrunken batting average greater than .300.

Chart, line chart

Description automatically generated