Bayesball

Empirical Bayesian Estimation of Batting Averages

Nick Sun

Oregon State University

June 3, 2020

Outline

Batting Averages

- No number is more important or ubiquitous in sabermetrics than batting average. The formula is simply $\frac{\text{Hits}}{\text{At-bats}}$.
- Batting average is calculated cumulatively over the season, starting at .000. This makes the variance of early season batting average calculations quite large.
- We can think of these batting averages as an estimation of p, a batter's true probability of getting a hit in an at-bat.
- Is there a "better" $\hat{p}$ than this?

Empirical Bayesian Estimation

- One of the first models we learned this quarter was a binomial model with a beta conjugate prior. Perhaps we can use this for modelling p?
- If we have a prior $\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$, then our posterior distribution will be:

$$\pi(p|AB, H) \propto p^{\alpha+H-1}(1-p)^{\beta+AB-H-1}$$

where H is the number of hits and AB is the number of at-bats.

Empirical Bayesian Estimation

I had three goals with this analysis:

1. Explore some ways of finding a prior $\pi(p)$
2. Compare the MSE of empirical Bayesian estimation and normal batting average calculations as a season progressed
3. Compare credible and confidence intervals

We will answer these questions using simulations.

Credible Intervals vs. Confidence Intervals

Bayesian estimation allows us to make bayesian credible intervals which we can compare against frequentist confidence intervals.

- **Bayesian Credible interval**: `qbeta(c(.025, .975), alpha0 + cumH, beta0 + cumAB - cumH)`
- **Jeffrey's interval**: `qbeta(c(.025, .975), .5 + cumH, .5 + cumAB - cumH)`
- **Clopper-Pearson interval**: `qbeta(.025, cumH, cumAB - cumH + 1), qbeta(.975, cumH + 1, cumAB - cumH)`

where *cumH* and *cumAB* are the cumulative hits and at-bats at any given point in the season.[3]

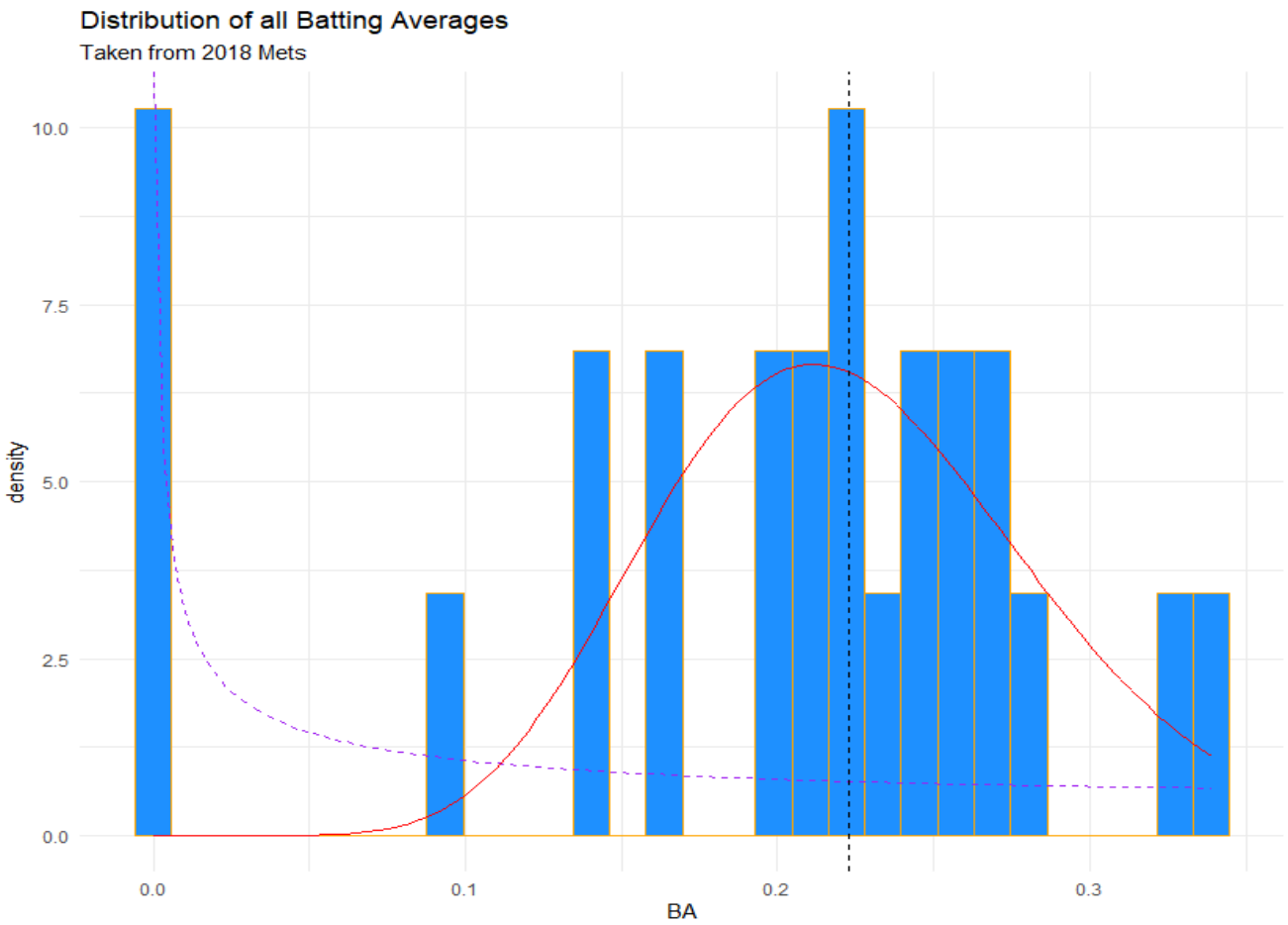
Meet our Subject

- Michael Conforto (aka "Scooter") is a promising right fielder for the New York Mets. He's also from good ol' OSU.
- His young career so far has been somewhat inconsistent, fluctuating between average to All-Star.[2]

Season	AB	H	Batting Average
2015	174	47	.270
2016	304	67	.220
2017	373	104	.279*
2018	543	132	.243
2019	549	141	.257

Prior

A reasonable place to start with a prior is at team level. Below is a density histogram of all batting averages from the 2018 New York Mets.[1]

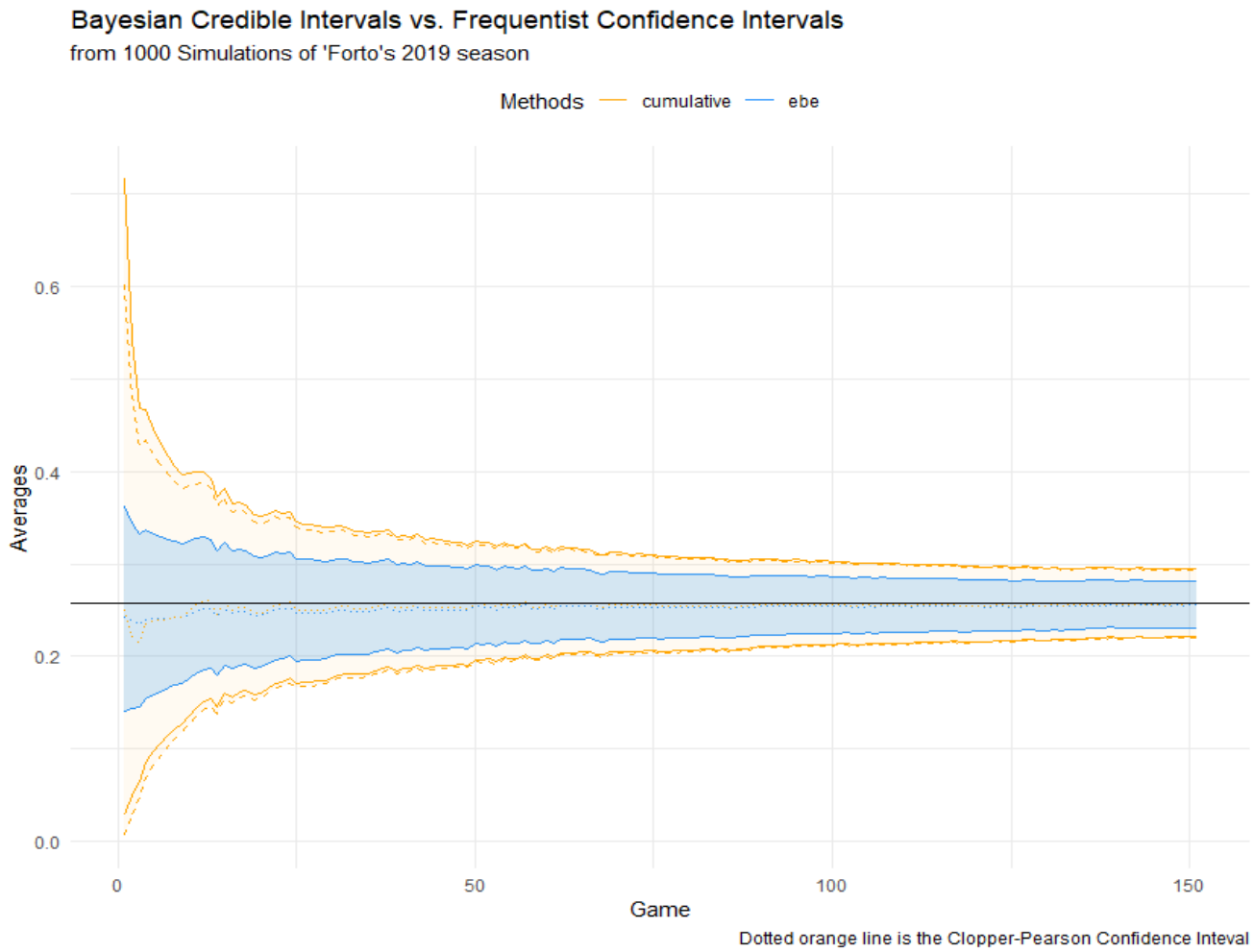


Prior

- It is known that the mean of a beta distribution is given by $\frac{\alpha}{\alpha+\beta}$.
- We can therefore try "shifting" the mean of this Mets distribution by changing the value of α to Michael Conforto's batting average in 2018
- This will also change the variance of the beta prior, but this was the method I found which worked the best.
- In 2018, Conforto had a batting average of .243 which gave me a prior of Beta(11.74, 36.76)

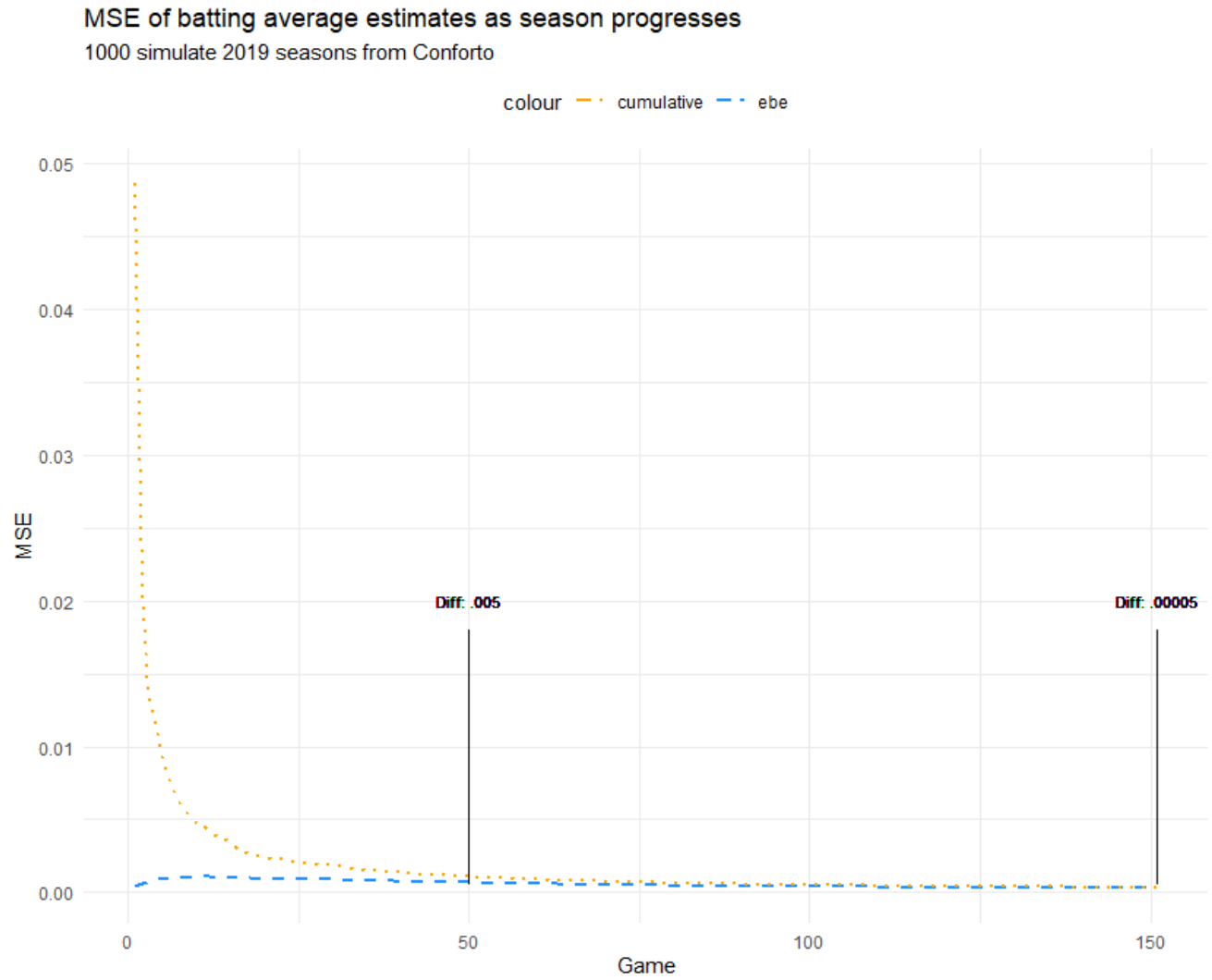
Simulating the 2019 season

Michael Conforto had a .257 batting average in 2019. If we treat this as the true value of p, we can simulate 1000 2019 seasons using game logs and check the MSE of our estimates at each game in the season.



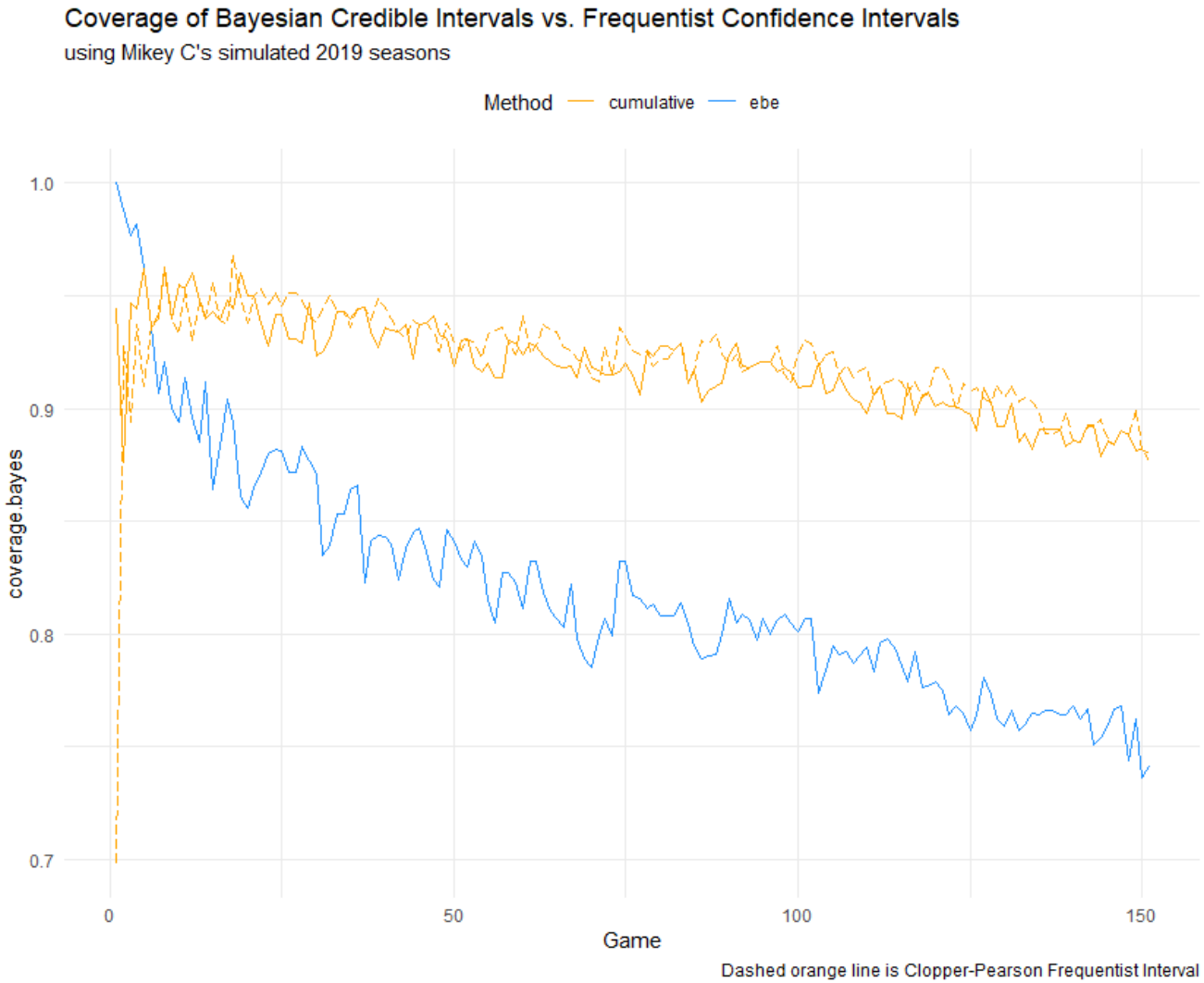
MSE of the 2019 simulations

MSE is calculated as the average of the squared deviations from the true value of p.



Inteval Coverage

For each game, I calculated the proportion in which p fell within each of the three intervals.

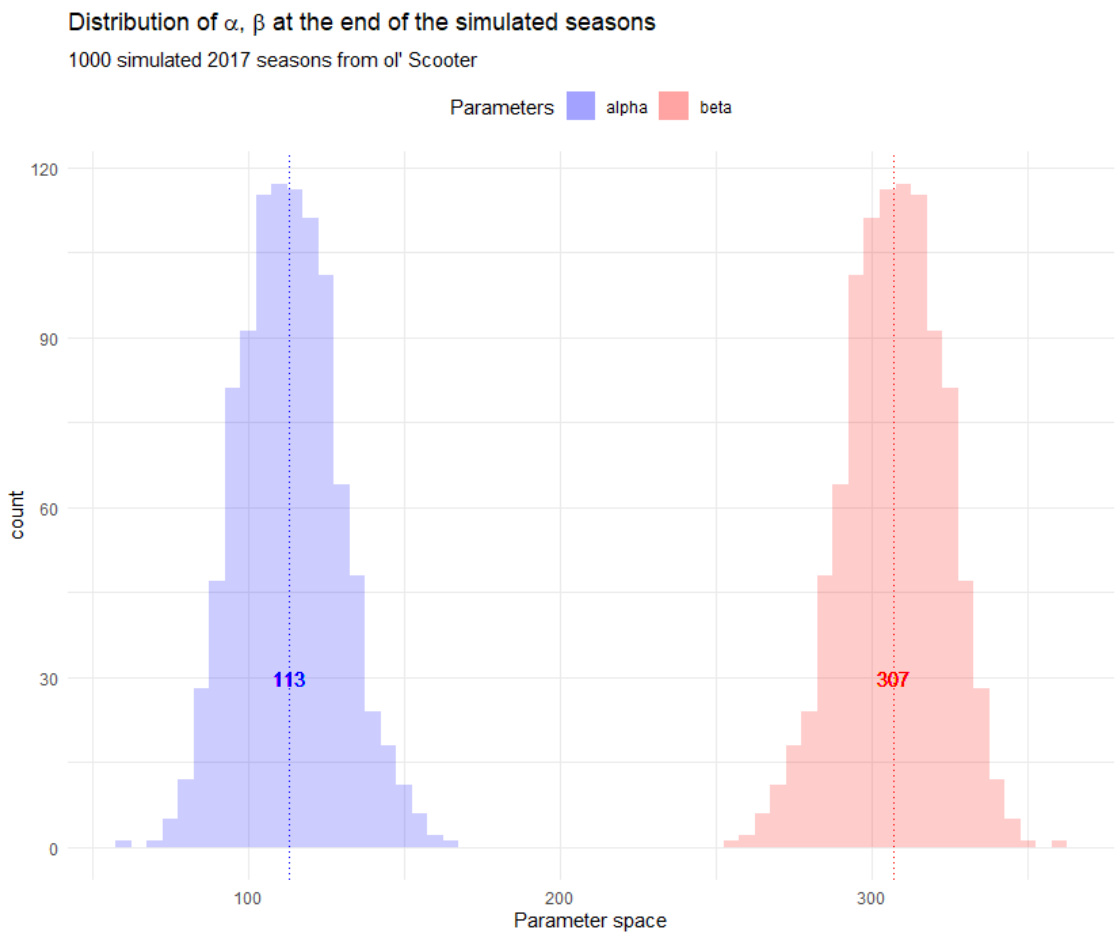


New simulation: p is drawn randomly

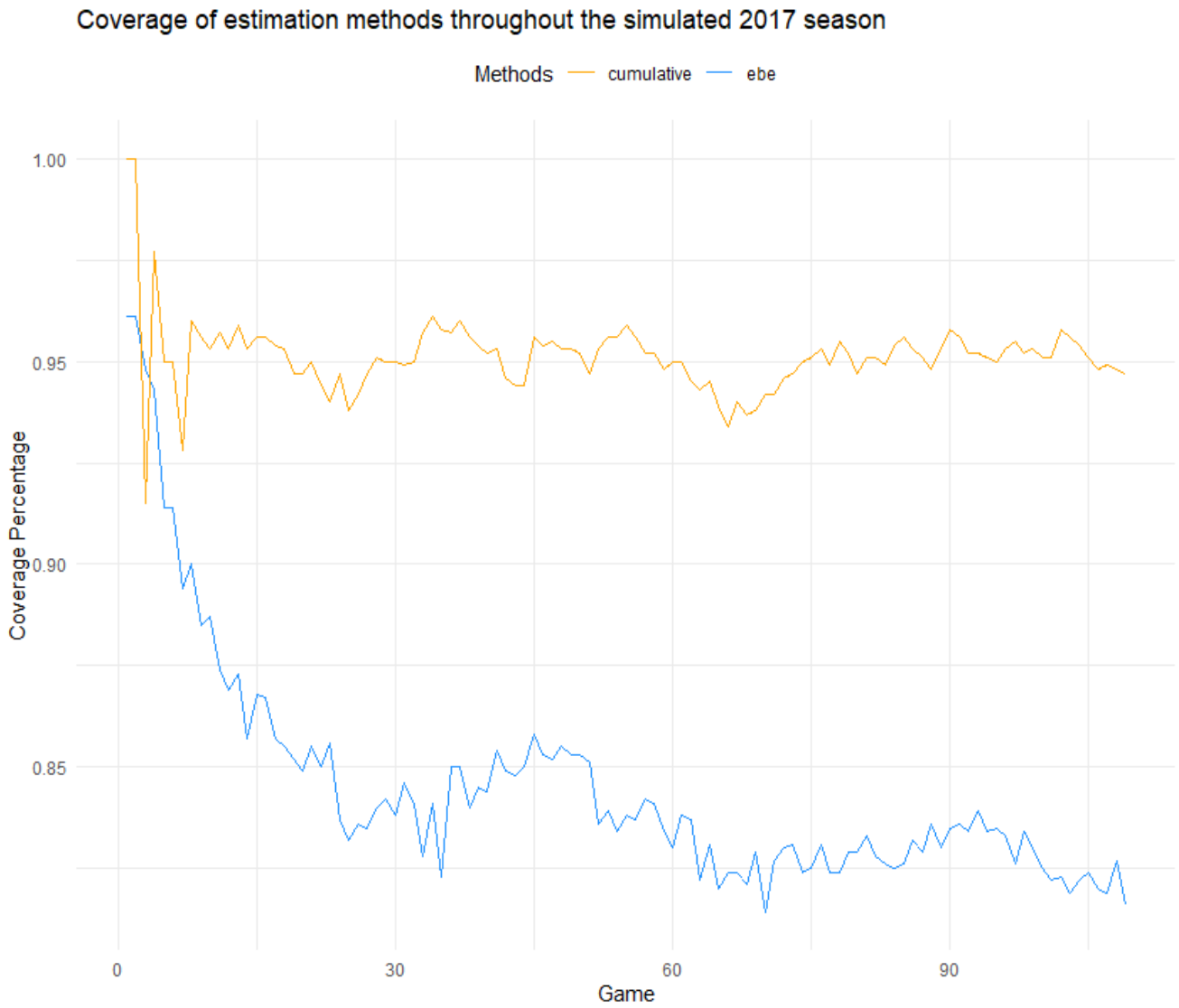
- Instead of having a fixed constant p for every simulated season, we will draw random a p from a Beta(45.173, 117.316) distribution, representing the batting average distribution of all Mets in 2017 and adjusted to have a mean of .279
- The prior mean will be equal to Conforto's 2016 performance of .220

Posterior distribution of parameters

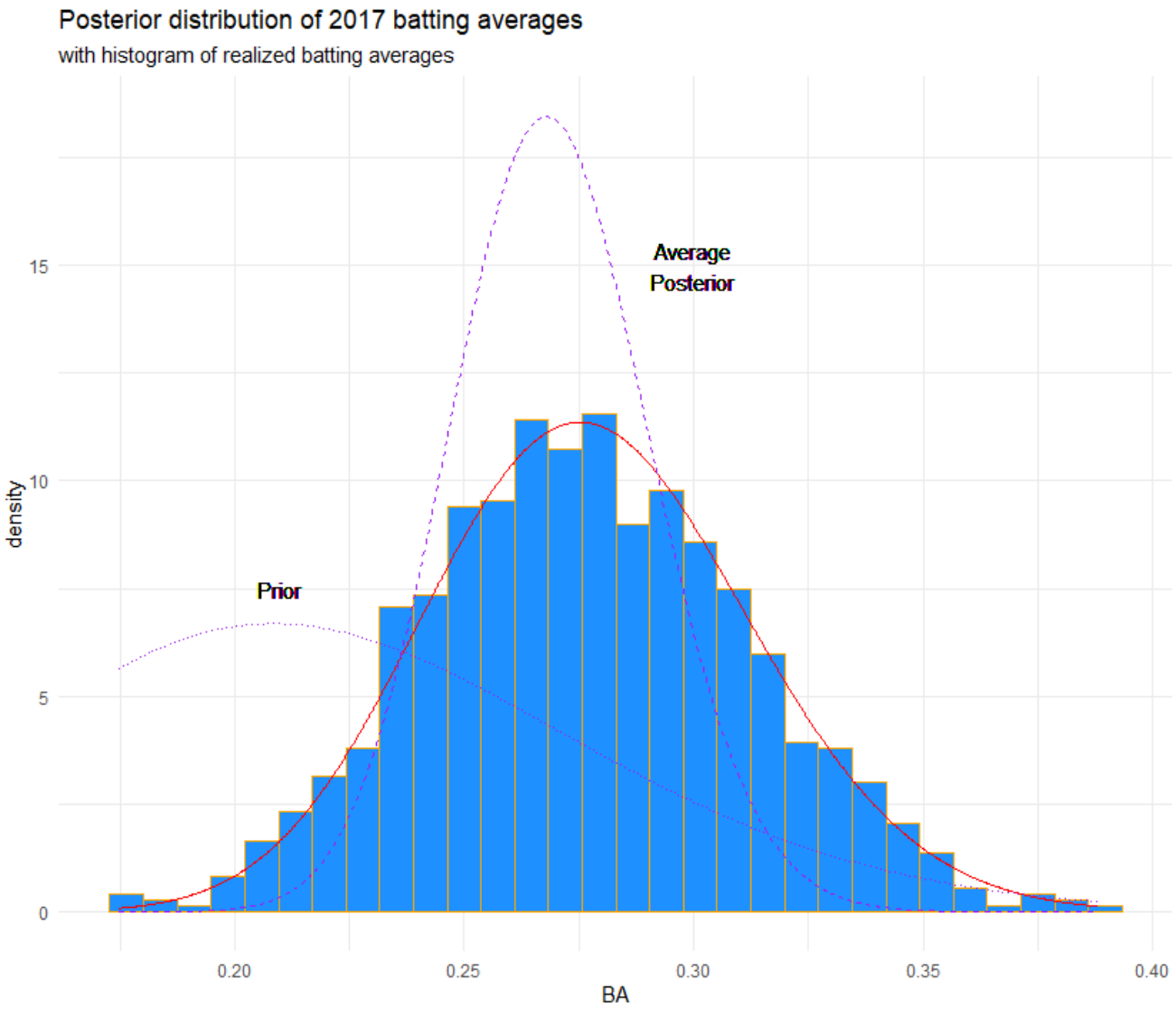
Conforto played 109 games in 2017. We can plot α and β for all 1000 simulated posterior distributions.



Interval Coverage

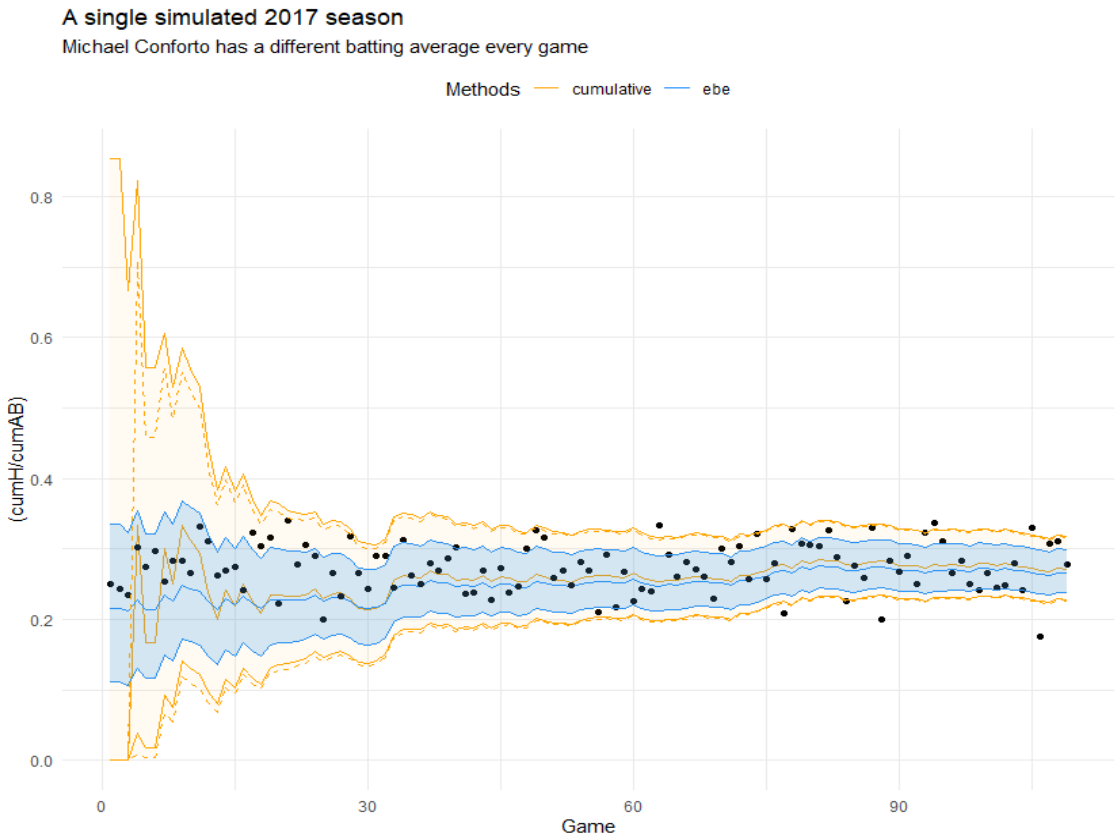


Posterior Distribution

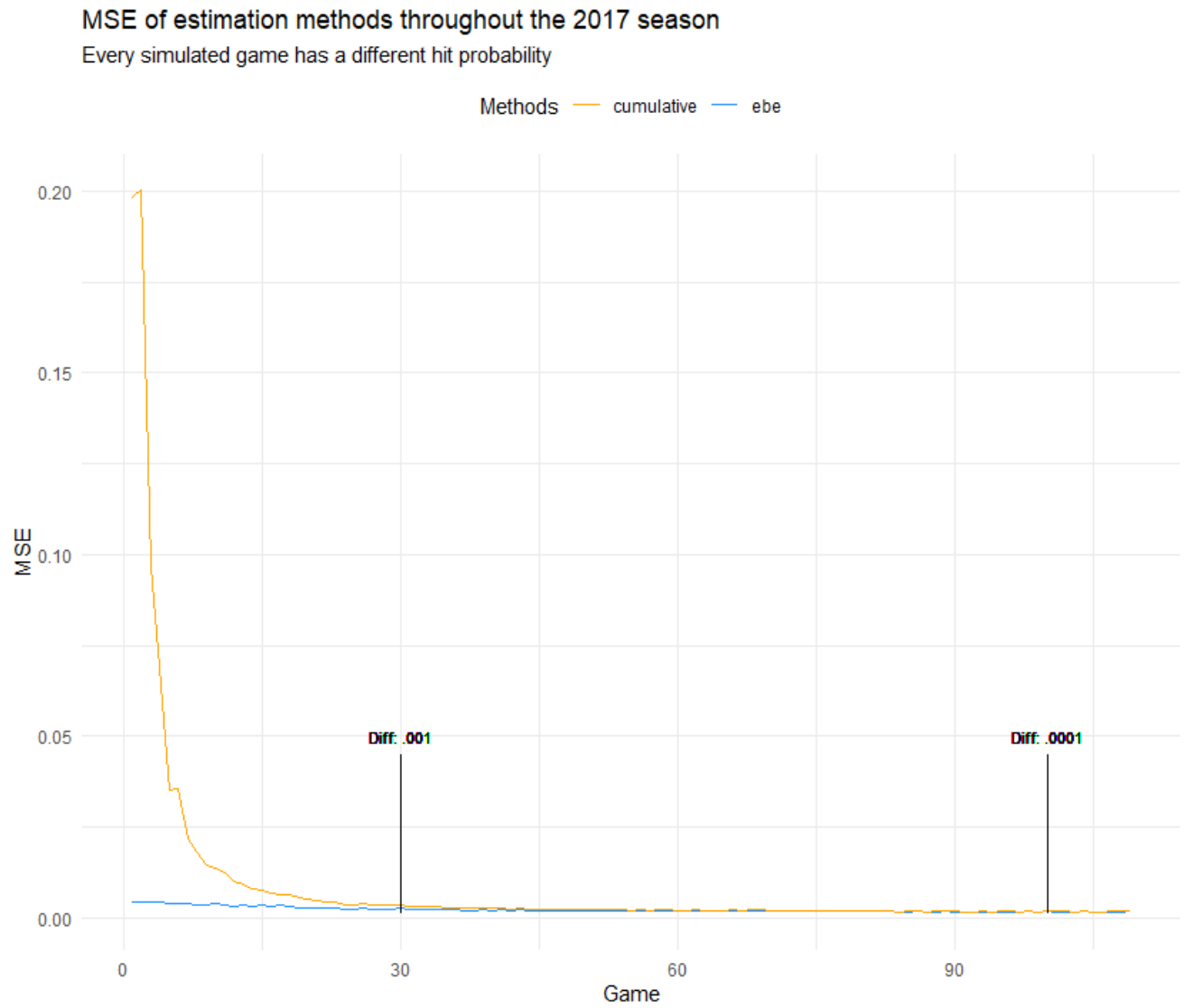


One last simulation

Instead of having a single p for an entire season, what if p was changed every day? This is likely a more realistic model of MLB players hitting ability. Again we sample random p from Beta(45.173, 117.316), but this time p is a vector instead of a scalar.



MSE comparison



Conclusion

- Empirical Bayesian estimation with a prior based on the last season's performance had a lower MSE in all simulations and all time points, but especially in the beginning of the season
- Bayesian credible intervals are narrower than frequentist intervals owing to the lower variance of the posterior distribution as α and β grow
- Both Bayesian and Frequentist intervals are too wide to be of much practical use (average width being around .1 for Bayesian intervals)

Reference


Sean Lahman. *Lahman package*. 2020. URL: https://cran.r-project.org/web/packages/Lahman/Lahman.pdf.


Sports Reference LLC. *Michael Conforto Game Logs*. 2020. URL: https://www.baseball-reference.com/players/c/confomi01.shtml.


David Robinson. *Introduction to Empirical Bayes: Examples from Bayesian Statistics*. 2016.