# Empirical Bayes

**Q1** Suppose Mookie Betts' batting average midway through the season is .300. Using no other information, predict his end-of-season batting average.

**Model** mid-season batting average is

$$\frac{H}{N} \sim \frac{1}{N} \cdot \text{Binomial}(N, p)$$

$H = \#$ hits, $N = \#$ at-bats

As discussed previously, the MLE of a binomial is

$$\hat{p}_{MLE} = \frac{H}{N}$$

and this is our prediction.

Concretely, we predict Mookie's end-of-season batting average to be .300.

Yesterday we used a prior, but we can't do that here since we have no other information.

**Q2** Suppose we know each player's batting average midway through the 2023 season. Using no information from any previous season, i.e. only using these 2023 mid-season batting averages, predict each player's end-of-season batting average.

One approach:

Notation
player index $i$
player $i$'s # hits $H_i$
player $i$'s # at-bats $N_i$
player $i$'s batting average $\frac{H_i}{N_i}$

Model
$$\frac{H_i}{N_i} \sim \frac{1}{N_i} \text{Binomial}(N_i, p_i)$$

MLE
$$\hat{p}_i^{(MLE)} = \frac{H_i}{N_i}$$

Can we do better??

## Previous idea: Shrinkage

But, we only have access to batting averages from 2023 and nothing else.
So, we can't shrink to prior information from previous seasons.
What can we shrink to??

Well, we have the midseason batting average of each baseball player, so perhaps we can pool information and shrink to the overall mean batting average.

Idea: to predict the batting average of Mookie Betts, use the batting averages of every other player!

Insight: Mookie Betts is a baseball player
use that information!

**Notation** player index $i$
player $i$'s # hits $H_i$
player $i$'s # at-bats $N_i$
player $i$'s batting average $X_i = \frac{H_i}{N_i}$

**Model** $X_i \sim \frac{1}{N_i} \text{Binomial}(N_i, p_i)$

Remove players from the dataset with small $N_i$ (say, $N_i < 25$).
Since $N_i$ large,

$$X_i \approx \mathcal{N}\left(p_i, \frac{p_i(1-p_i)}{N_i}\right)$$

by Central Limit Theorem.

## Simplification

$p_i$ unknown, so variance $\sigma_i^2 = \frac{p_i(1-p_i)}{N_i}$ unknown

Much easier to work with known variance

so let's just assume for simplicity

that $\sigma_i^2 = \frac{C}{N_i}$ for some known

constant $C$.

We'll use $C = 0.035$, treat this as a tuning hyperparameter... (tune using a previous season).

In practice, we use a <u>variance stabilizing transform</u> $h(X)$ which transforms the batting average so that it has a known variance...

Now, $X_i \sim N(\mu, \sigma_i^2)$ $\sigma_i^2$ known...

## Parametric Bayesian Model

Bayesian: treat the parameter $\mu_i$ as having a probability distribution

$$\boxed{\begin{aligned} X_i &\sim N(\mu_i, \sigma_i^2) \\ \mu_i &\sim N(\mu, \tau^2) \end{aligned}}$$

<u>PRIOR:</u> $\mu_i \sim N(\mu, \tau^2)$

insight: each player $i$ is a baseball player
so, let's pool information!
there is a mean $\mu$ across all baseball players $\mu_i$
and some s.d. $\tau$

$\mu_i = P_i$ unknown $\quad \sigma_i = \frac{\sqrt{C}}{\sqrt{N_i}}$ known

$\mu, \tau$ unknown

**MLE:** ignore the prior, only use the information relevant to player $i$

$$\boxed{\hat{\mu}_i^{(MLE)} = X_i}$$

observed mid-season batting average

**Bayesians:** use the PRIOR!

The Bayesian estimate is the posterior mean, i.e. the mean of the dist of $\mu_i$ after seeing the data $\{X_i\}$,

$$\hat{\mu}_i^{(Bayes)} = \mathbb{E}(\mu_i \mid X_i)$$

need the posterior dist $\mu_i \mid X_i$,

$$\mathbb{P}(\mu_i \mid x_i) = \frac{\mathbb{P}(x_i \mid \mu_i)\,\mathbb{P}(\mu_i)}{\mathbb{P}(x_i)} \quad \text{Bayes Rule}$$

$$\propto \underbrace{\mathbb{P}(x_i \mid \mu_i)}_{N(\mu_i, \sigma_i^2)}\ \underbrace{\mathbb{P}(\mu_i)}_{N(\mu, \tau^2)}$$

since $\mathbb{P}(x_i)$ has no $\mu_i$ term

$$= \mathbb{P}\left(\mathcal{N}(\mu_i, \sigma_i^2) = X_i\right) \cdot \mathbb{P}\left(\mathcal{N}(\mu, \tau^2) = \mu_i\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(X_i-\mu_i)^2}{\sigma_i^2}\right) \cdot \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\frac{(\mu_i-\mu)^2}{\tau^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\cancel{\frac{X_i^2}{\sigma_i^2}} - \frac{2\mu_i X_i}{\sigma_i^2} + \frac{\mu_i^2}{\sigma_i^2} + \frac{\mu_i^2}{\tau^2} - \frac{2\mu_i\mu}{\tau^2} + \cancel{\frac{\mu^2}{\tau^2}}\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\mu_i^2\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right) - 2\mu_i\left(\frac{X_i}{\sigma_i^2}+\frac{\mu}{\tau^2}\right)\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right)\left[\mu_i^2 - 2\mu_i\frac{\left(\frac{X_i}{\sigma_i^2}+\frac{\mu}{\tau^2}\right)}{\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right)}\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right)\left[\mu_i - \frac{\left(\frac{X_i}{\sigma_i^2}+\frac{\mu}{\tau^2}\right)}{\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right)}\right]^2\right)$$

$$\Longrightarrow \mu_i \mid X_i \sim \mathcal{N}\left(\frac{\left(\frac{X_i}{\sigma_i^2}+\frac{\mu}{\tau^2}\right)}{\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right)}, \frac{1}{\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right)}\right)$$

$\Longrightarrow$ Posterior Mean

$$\boxed{\hat{\mu}_i^{(\text{Bayes})} = \frac{\frac{X_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} = \mu + \frac{\tau^2}{\tau^2 + \sigma_i^2}(X_i - \mu)}$$

$X_i = \frac{H_i}{N_i}$, $\sigma_i^2 = \frac{C}{N_i}$

$$\Longrightarrow \boxed{\hat{\mu}_i^{(\text{Bayes})} = \frac{\frac{H_i}{C} + \frac{\mu}{\tau^2}}{\frac{N_i}{C} + \frac{1}{\tau^2}}}$$

* Looks a lot like $\frac{W + W'}{W + L + W' + L'}$ from last lecture!

Problem: $\mu$ and $\tau$ are unknown quantities

Empirical Bayes: Estimate $\mu$ and $\tau$ from observed data!

$$\boxed{\hat{\mu}_i^{(EB)} = \frac{\frac{H_i}{C} + \frac{\hat{\mu}}{\hat{\tau}^2}}{\frac{N_i}{C} + \frac{1}{\hat{\tau}^2}}}$$

If $\tau^2 = 0$, $\quad \mu_i \sim N(\mu, \tau^2) \approx N(\mu, 0) = \mu$

$$\Longrightarrow \hat{\mu}_i = \hat{\mu}$$

If $\tau^2 = \infty$, $\quad \mu_i \sim N(\mu, \infty) \approx \text{Uniform}(-\infty, \infty)$

and $\hat{\mu}_i = \frac{H_i}{N_i} = X_i = \hat{\mu}_i^{(MLE)}$

Else, $\hat{\mu}_i = \hat{\mu} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2}(X_i - \hat{\mu})$

is closer to $\hat{\mu}$ if $\frac{\hat{\tau}^2}{\hat{\tau}^2+\sigma_i^2}$ is small — large $\sigma_i^2$, small $N_i$

is closer to $X_i$ if it is large — small $\sigma_i^2$, large $N_i$

* use $\hat{\mu}^{(MLE)}$, $\hat{\tau}^{2(MLE)}$ to estimate $\mu, \tau^2$

Model $\begin{cases} X_i \sim N(\mu_i, \sigma_i^2) \\ \mu_i \sim N(\mu, \tau^2) \end{cases}$

Get Rid of $\mu_i$

Marginal distribution $\quad X_i \sim N(\mu, \tau^2 + \sigma_i^2)$

By Bayes Rule.

Log-Likelihood

$$L(\mu,\tau^2) = \log \mathbb{P}(X \mid \mu,\tau^2)$$

$$= \log \prod_{i=1}^{n} \mathbb{P}(X_i \mid \mu,\tau^2) \quad \text{by independence}$$

$$= \sum_{i=1}^{n} \log \mathbb{P}\left(\mathcal{N}(\mu,\tau^2+\sigma_i^2) = X_i\right)$$

$$= \sum_{i=1}^{n} \log \left[ \frac{1}{\sqrt{2\pi(\tau^2+\sigma_i^2)}} e^{-\frac{(X_i-\mu)^2}{2(\tau^2+\sigma_i^2)}} \right]$$

$$= \sum_{i=1}^{n} \left\{ \log\left(\frac{1}{\sqrt{2\pi(\tau^2+\sigma_i^2)}}\right) - \frac{(X_i-\mu)^2}{2(\tau^2+\sigma_i^2)} \right\}$$

$$\propto -\frac{1}{2}\sum_i \log(\tau^2+\sigma_i^2) - \frac{1}{2}\sum_i \left(\frac{(X_i-\mu)^2}{\tau^2+\sigma_i^2}\right)$$

MLE $(\mu,\tau^2) = \underset{\mu,\tau^2}{\operatorname{argmin}}\, L(\mu,\tau^2) \Longrightarrow$ solve $\frac{dL}{d\mu}=0, \frac{dL}{d\tau^2}=0.$

$$\frac{\partial L}{\partial \mu} = \frac{1}{2}\sum_i \frac{2(X_i-\mu)}{\tau^2+\sigma_i^2} = 0$$

$$\Longrightarrow \mu = \frac{\sum_i X_i/(\tau^2+\sigma_i^2)}{\sum_i 1/(\tau^2+\sigma_i^2)}$$

$$\frac{\partial L}{\partial \tau^2} = -\frac{1}{2}\sum_i \frac{1}{\tau^2+\sigma_i^2} + \frac{1}{2}\sum_i \frac{(X_i-\mu)^2}{(\tau^2+\sigma_i^2)^2} = 0$$

$$\Longrightarrow \quad \sum_i \frac{(X_i-\mu)^2}{(\tau^2+\sigma_i^2)^2} = \sum_i \frac{1}{\tau^2+\sigma_i^2}$$

<u>Problem</u> $\mu$ is in terms of $\tau^2$ and $\tau^2$ is in terms of $\mu$, but we need both...

<u>Iteratively Solve for $\hat{\mu}, \hat{\tau}^2$:</u>

$$\hat{\mu} = \frac{\sum X_i/(\hat{\tau}^2+\sigma_i^2)}{\sum 1/(\hat{\tau}^2+\sigma_i^2)}$$

$$\sum \frac{(X_i-\hat{\mu})^2}{(\hat{\tau}^2+\sigma_i^2)^2} = \sum \frac{1}{\hat{\tau}^2+\sigma_i^2}$$

$\mu^{(0)}$ = educated guess

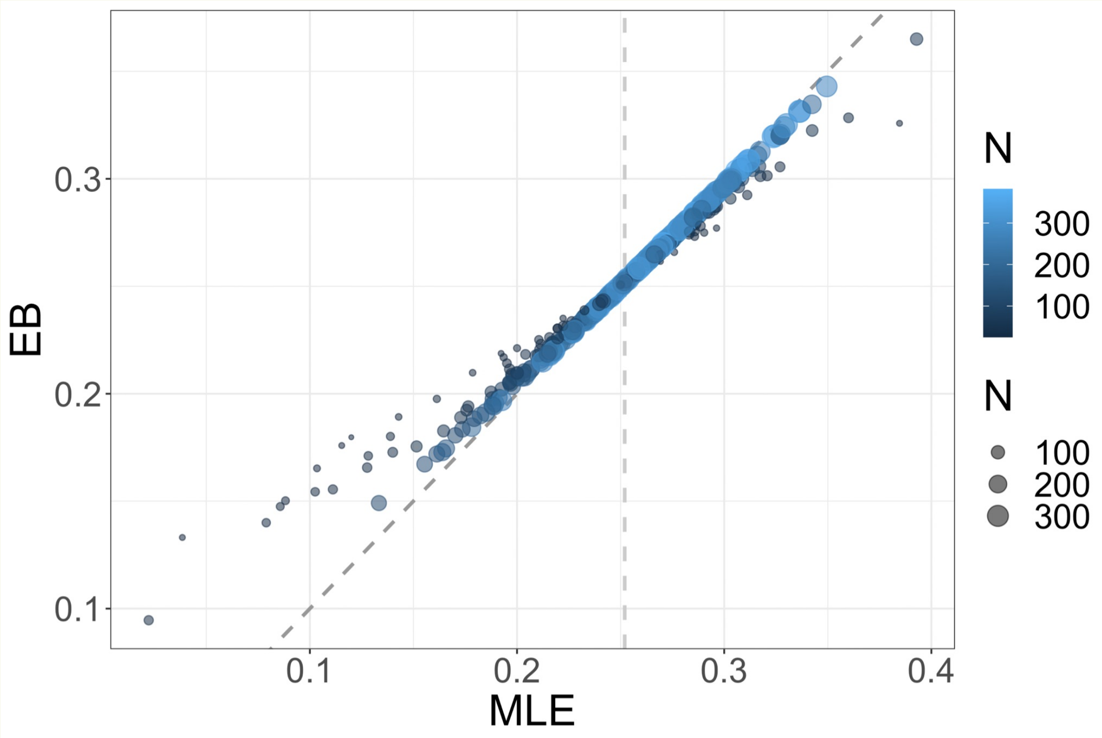
$\tau^{2(0)}$ = educated guess

while not $|\mu^{(t)} - \mu^{(t-1)}| \leq \delta$ and $|\tau^{2(t)} - \tau^{2(t-1)}| < \delta$:

$$\mu^{(t)} \leftarrow \frac{\sum_i X_i/(\tau^2+\sigma_i^2)}{\sum_i 1/(\tau^2+\sigma_i^2)}$$

$\tau^{2(t)} \leftarrow$ the solution of $\sum_i \frac{(X_i-\mu)^2}{(\tau^2+\sigma_i^2)^2} = \sum_i \frac{1}{\tau^2+\sigma_i^2}$

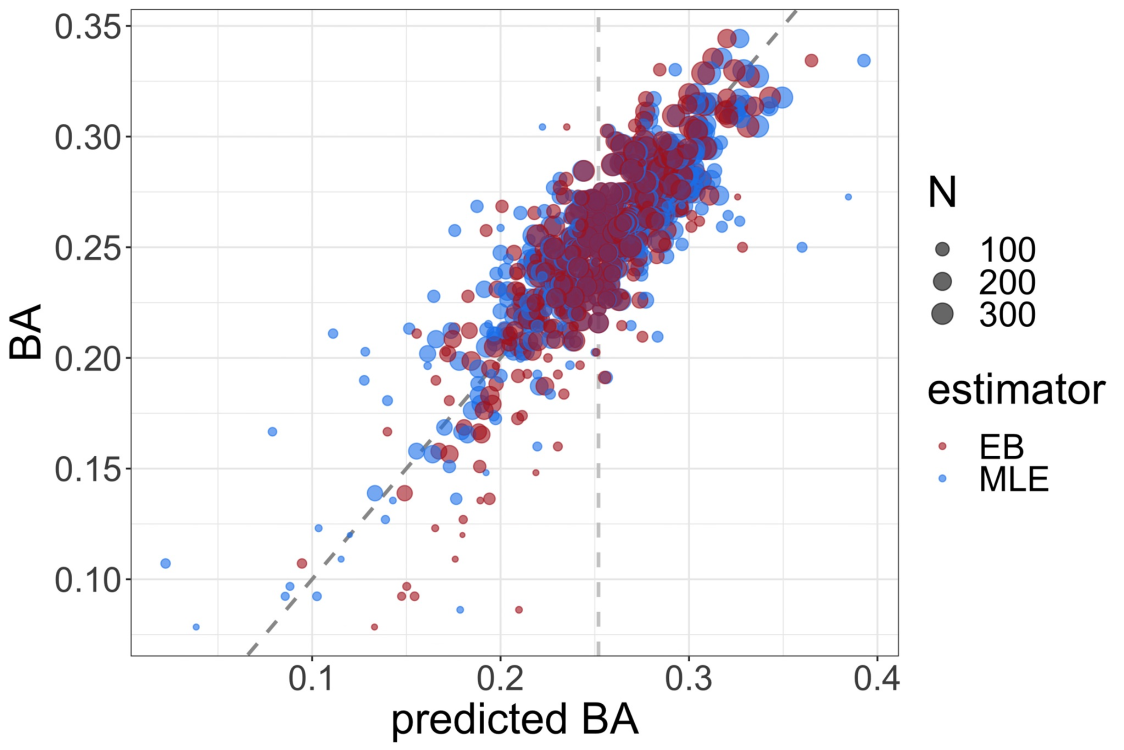
(use UniRoot in R)




* Players with smaller N have $\hat{\mu}^{EB}$ shrunk towards the overall mean, while players with larger N have $\hat{\mu}^{(EB)} = \hat{\mu}^{(MLE)}$ their mid-szn BA

| rmse_MLE | rmse_EB |
|---|---|
| 0.02629828 | 0.02383808 |

* $\hat{\mu}^{(EB)}$ predict better the actual end-of-szn BA!




* There really isn't a huge difference b/t the predictions though...

Takeaway

- shrinkage towards the overall mean helps prediction when have smaller sample sizes
- sharing information helps!

# Shrinkage, with Access to Previous Seasons' Data

Parametric Empirical Bayes Estimator:

$$\hat{\mu}_i^{(PEB)} = \hat{\mu} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2}(X_i - \hat{\mu})$$

General Shrinkage Estimator:

$$\hat{\mu}_i = \hat{\mu} + \beta (X_i - \hat{\mu})$$

<u>Procedure</u> $\{X_i\}$ data

$\hat{\mu}$ = overall mean (sample mean of data from the season of datapoint $i$

$\hat{\mu}_i$ = response column
= known end-of-szn batting avg of datapoint $i$ (player $i$) from a previous szn

Estimate $\beta$ using Regression, call it $\hat{\beta}$

Then, for future prediction,

$$\hat{\mu}_i = \hat{\mu} + \hat{\beta}(X_i - \hat{\mu}).$$