# Lecture 9: Bayesian Estimation & Empirical Bayes

## Big Data and Machine Learning for Applied Economics
## Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

September 8, 2020

# Announcement

- **Next Thursday September 10, I'll be teaching the class**
- **Problem Set 1 is due next Tuesday September 15 at 11:00**
- At some point over the weekend I'll send what points everyone should present
- Assignment would be based on the groups created on `Github`
- You should consider class presentations as mini-seminars, just 2-5 minutes using one or two transparencies
- Attempt to make a concise interpretation of the relevant material, making effective use of supporting numerical and graphical evidence.

# Agenda

1 Bayes Theorem

2 A Simple Covid Example

3 Empirical Bayes
- Batting Averages
- Predicting Batting Averages

4 Further Readings

# Bayes Theorem

$$\pi(\theta|X) = \frac{f(X|\theta)p(\theta)}{m(X)} \tag{1}$$

with $m(X)$ is the marginal distribution of $X$, i.e.

$$m(X) = \int f(X|\theta)p(\theta)d\theta \tag{2}$$

It is important to note that Bayes' theorem does not tell us what our beliefs should be, it tells us how they should change after seeing new information.

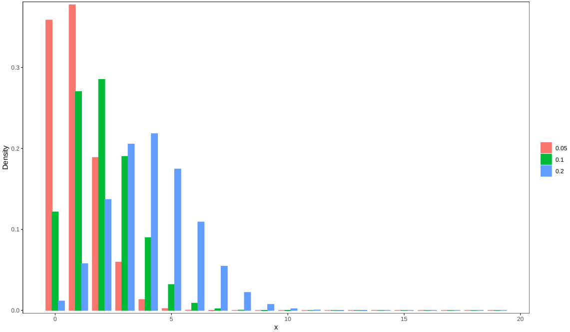
# A Simple Covid Example

- Suppose we are interested in the prevalence of COVID in a small city. The higher the prevalence, the more public health precautions we would recommend be put into place.
- A small random sample of 20 individuals from the city will be checked for the presence of the virus.
- Interest is in $\theta$, the fraction of infected individuals in the city. Roughly speaking, the parameter space includes all numbers between zero and one. The data $X$ records the total number of people in the sample who are infected.
- Before the sample is obtained the number of infected individuals in the sample is unknown.

# A Simple Covid Example

If the value of $\theta$ were known, a reasonable sampling model would be

$$X|\theta \sim Binomial(20, \theta) \tag{3}$$




$$Pr(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{4}$$

$$Pr(X = 0) = \binom{20}{0} 0.05^0 (1-0.05)^{20-0} \approx 0.36 \tag{5}$$

# A Simple Covid Example

## Prior distribution

- Other studies from various parts of the country indicate that the infection rate in comparable cities ranges from about 0.05 to 0.20, with an average prevalence of 0.10.
- We will therefore use a prior distribution $p(\theta)$

$$\theta \sim Beta(a,b) \tag{6}$$

where the density of a Beta takes the form of

$$p(\theta) = \frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)}\theta^{a-1}\left(1-\theta\right)^{b-1} \tag{7}$$

with $a = 2$ and $b = 20$. Note that

$$E(\theta) = \frac{a}{a+b} = 0.09 \tag{8}$$

$$Pr(0.05 < \theta < 0.20) = 0.66 \tag{9}$$

# A Simple Covid Example

## Posterior distribution

$$\pi(\theta|X) = \frac{f(X|\theta)p(\theta)}{m(X)} \tag{10}$$

$$\pi(\theta|X) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\frac{1}{m(x)} \tag{11}$$

The marginal

$$\begin{aligned}
m(x) &= \int f(X|\theta)p(\theta)d\theta &(12)\\
&= \int_0^1 \binom{n}{x}\theta^{x}(1-\theta)^{n-x} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}d\theta &(13)\\
&= \binom{n}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \theta^{x+a-1}(1-\theta)^{n-x+b-1}d\theta &(14)
\end{aligned}$$

# A Simple Covid Example

## Posterior distribution

### The marginal (cont)

$$m(x) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} \int_0^1 \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)} \theta^{x+a-1}(1-\theta)^{n-x+b-1} d\theta \tag{15}$$

$$= \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} \tag{16}$$

### The posterior

$$\pi(\theta|X) = \frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)} \theta^{x+a-1}(1-\theta)^{n-x+b-1} \tag{17}$$

$$\sim Beta(a+x, b+n-x) \tag{18}$$

# A Simple Covid Example

With the posterior we can calculate then any moment of the posterior distribution. For example suppose that for our study none of the sample of individuals is infected (x=0). Then the posterior is

$$\pi(\theta|X=0) \sim Beta(2,40) \tag{19}$$

$a=2, b=20, n=20$. Then

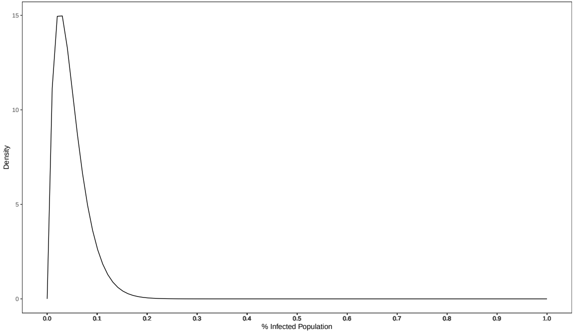
$$\begin{aligned}
E(\theta|X=0) &= \frac{a+x}{a+b+n} && (20)\\
&= \frac{n}{a+b+n}\frac{x}{n} + \frac{a+b}{a+b+n}\frac{a}{a+b} && (21)\\
&= \frac{n}{a+b+n}\bar{x} + \frac{a+b}{a+b+n}\theta_{prior} && (22)\\
&= \frac{n}{a+b+n}0 + \frac{a+b}{a+b+n}\frac{2}{22} && (23)\\
&= 0.048 && (24)
\end{aligned}$$

# A Simple Covid Example

Since we have the full distribution we could calculate for example:

$$mode(\theta|X) = 0.025 \tag{25}$$

$$Pr(\theta < 0.10|X = 0) = 0.93 \tag{26}$$



# Bayes Theorem

**Conjugate Priors. Basic idea:**

- $X \sim D(\theta)$ and $\theta \sim P(\lambda) \rightarrow \theta|X \sim P(\lambda')$
- $X \sim Bernoulli(\theta)$ and $\theta \sim Beta(a,b) \rightarrow \theta|X \sim Beta(a',b')$
- $X \sim N(\mu,\sigma)$ and $\theta \sim N(\mu_0,\sigma_0) \rightarrow \theta|X \sim N(\mu',\sigma')$

# Batting Averages

We can model

$$\text{Batting Average} \sim Binomial(n, \theta) \tag{27}$$

- where $n$ is the times at bat and $\theta$ is the proportion of successes
- We use a conjugate prior for simplicity

$$p(\theta) \sim Beta(\alpha_0, \beta_0) \tag{28}$$

The posterior is:

$$\pi(\theta) \sim Beta(\alpha_0 + hits, \beta_0 + N - hits) \tag{29}$$

# Batting Averages

Using last class data:

```
## # A tibble: 6 x 4
##   name                H    AB average
##   <chr>           <int> <int>   <dbl>
## 1 Hank Aaron       3771 12364  0.305
## 2 Tommie Aaron      216   944  0.229
## 3 Andy Abad           2    21  0.0952
## 4 John Abadie        11    49  0.224
## 5 Ed Abbaticchio    772  3044  0.254
## 6 Fred Abbott       107   513  0.209
```

# Batting Averages

We are using batting averages to assess who are the best and worst batters

- Best?

```
## # A tibble: 6 x 4
##   name                H    AB average
##   <chr>           <int> <int>   <dbl>
## 1 Roe Skidmore        1     1       1
## 2 Charlie Snow        1     1       1
## 3 Matt Tupman         1     1       1
## 4 Allie Watt          1     1       1
## 5 Al Wright           1     1       1
## 6 George Yantz        1     1       1
```

# Batting Averages

We are using batting averages to assess who are the best and worst batters

- Worst?

```
## # A tibble: 6 x 4
##   name                  H    AB average
##   <chr>             <int> <int>   <dbl>
## 1 Frank Abercrombie     0     4       0
## 2 Horace Allen          0     7       0
## 3 Pete Allen            0     4       0
## 4 Walter Alston         0     1       0
## 5 Bill Andrus           0     9       0
## 6 Wyman Andrus          0     4       0
```

# Batting Averages

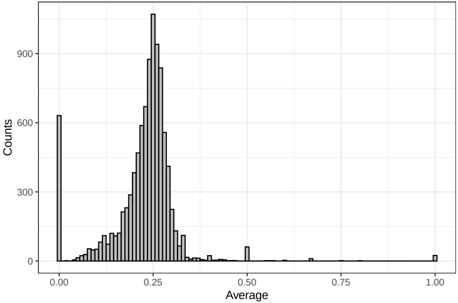
Question: Can we use Bayesian stats to get a better estimate?

$$X \sim Beta(\alpha_0, \beta_0) \tag{30}$$

- We don't know $\alpha_0$ and $\beta_0$. We could use the fact that most batting averages are between .210 and .360. Select $\alpha_0$ and $\beta_0$ accordingly.
- Or we can use Empirical Bayes: estimate these parameters from the data

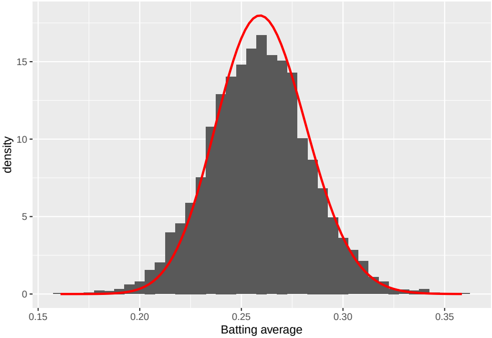
# Batting Averages

Histogram of batting averages



# Batting Averages

Restrict our sample to those data points that are informative (individuals that have gone at bat at least 500 times)



# Batting Averages

How we find the parameters that find the red line $\rightarrow$ MLE! We know that

$$f(x_i|\alpha_0,\beta_0) = \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} x_i^{\alpha_0-1}(1-x_i)^{\beta_0-1} \tag{31}$$

The log likelihood

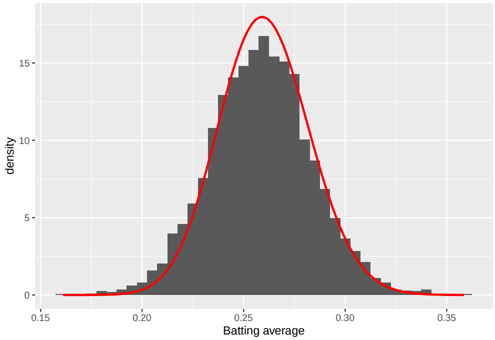
$$l(\alpha_0,\beta_0|X) = n.log(\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}) + \sum_{i=1}^{n}((\alpha_0-1)log(x_i)+(\beta_0-1)log(1-x_i)) \tag{32}$$

In R

```
# log-likelihood function
ll <- function(alpha, beta) {
-sum(VGAM::dbetabinom.ab(x, total, alpha, beta, log = TRUE))
}
# maximum likelihood estimation
m <- mle(ll, start = list(alpha = 1, beta = 10),
method = "L-BFGS-B", lower = c(0.0001, .1))
ab <- coef(m)
```

# Batting Averages

```
alpha0 <- ab[1]
101.7319
beta0 <- ab[2]
289.046
```



# Batting Averages

We can use the estimated average based on the posterior mean

$$E(\theta|X) = \frac{\alpha + hits}{\alpha + \beta + N} \tag{33}$$

- And ask again: who are the best batters by this improved estimate?

```
## # A tibble: 5 x 5
##   name                    H    AB average eb_estimate
##   <chr>               <int> <int>   <dbl>       <dbl>
## 1 Rogers Hornsby       2930  8173   0.358       0.354
## 2 Shoeless Joe Jackson 1772  4981   0.356       0.349
## 3 Ed Delahanty         2597  7510   0.346       0.342
## 4 Billy Hamilton       2164  6283   0.344       0.339
## 5 Willie Keeler        2932  8591   0.341       0.338
```

# Batting Averages

We can use the estimated average based on the posterior mean

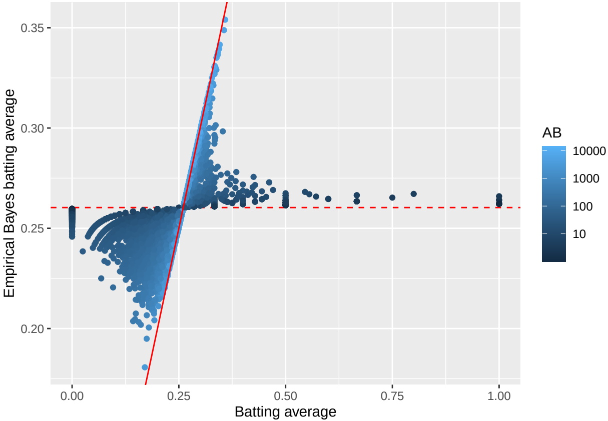
$$E(\theta|X) = \frac{\alpha + hits}{\alpha + \beta + N} \quad (34)$$

- Who are the *worst* batters?

```
## # A tibble: 5 x 5
##   name                H    AB average eb_estimate
##   <chr>           <int> <int>   <dbl>       <dbl>
## 1 Bill Bergen       516  3028   0.170       0.181
## 2 Ray Oyler         221  1265   0.175       0.195
## 3 Henry Easterday   203  1129   0.180       0.201
## 4 John Vukovich      90   559   0.161       0.202
## 5 George Baker       74   474   0.156       0.203
```

# Batting Averages

We can see how EB changed all of the batting average estimates:



# Predicting Batting Averages

- Now supposed you want to know the end of season final batting average of players, after observing them their 45 first times at bat.

| Player | Observed | Final |
|---|---|---|
| 1 | 0.395 | 0.346 |
| 2 | 0.355 | 0.279 |
| 3 | 0.313 | 0.276 |
| 4 | 0.291 | 0.266 |
| 5 | 0.247 | 0.271 |
| 6 | 0.224 | 0.266 |
| 7 | 0.175 | 0.318 |

# Predicting Batting Averages

- Recall that we can think each time at bat can be thought as a binomial trial, with $\theta$ the probability of success equal to the player's true batting average.
- With 45 trials, we can "reasonably" use a Normal Approximation.

$$X_i \sim N(\theta_i, \sigma^2) \tag{35}$$

where

- $\theta_i$ is the true batting average for player $i$
- $\sigma^2$ is the known variance that equals $(0.0659)^2$

We are going to use also a normal prior

$$\theta_i \sim N(\mu, \tau^2) \tag{36}$$

# Predicting Batting Averages

With this model the posterior mean for $\theta_i$ is $E(\theta_i|X_i)$

$$E(\theta_i|X_i) = \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}X_i \tag{37}$$

Note that the marginal of $X_i$

$$m(X_i) \sim N(\mu, \sigma^2+\tau^2) \;\; i = 1, \ldots, n \tag{38}$$

with these we can construct estimates of $E(\theta_i|X_i)$, note that

$$E(\bar{X}) = \mu \tag{39}$$

$$E\left[\frac{(n-3)\sigma^2}{\sum(X_i-\bar{X})^2}\right] = \frac{\sigma^2}{\sigma^2+\tau^2} \tag{40}$$

# Predicting Batting Averages

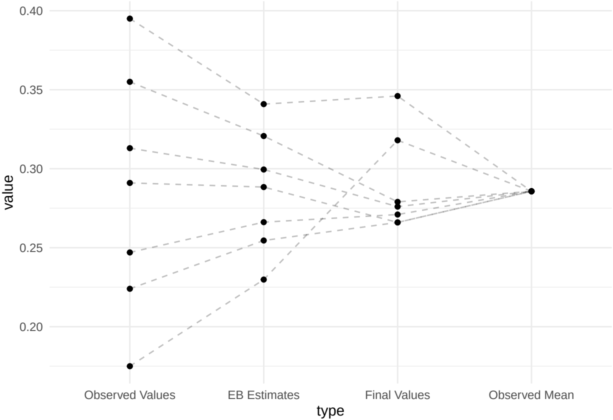
The empirical Bayes estimator of $\theta_i$ is then

$$\delta(X_i) = \left[\frac{(n-3)\sigma^2}{\sum((X_i-\bar{X})^2}\right]\bar{X} + \left[1 - \frac{(n-3)\sigma^2}{\sum((X_i-\bar{X})^2}\right]X_i \tag{41}$$

| Player | Observed | Final | Empirical Bayes |
|---|---|---|---|
| 1 | 0.395 | 0.346 | 0.341 |
| 2 | 0.355 | 0.279 | 0.321 |
| 3 | 0.313 | 0.276 | 0.299 |
| 4 | 0.291 | 0.266 | 0.288 |
| 5 | 0.247 | 0.271 | 0.266 |
| 6 | 0.224 | 0.266 | 0.255 |
| 7 | 0.175 | 0.318 | 0.230 |

- RMSE Observed 6.861903
- RMSE EB 3.918203

# Predicting Batting Averages



# Review & Next Steps

- Recap Bayesian
- Empirical Bayes Examples
- Next couple of classes we are going to focus on
    - Concepts underlying spatial data: points, lines, polygons, reference systems
    - Plotting and describing spatial data
    - Econometric models for spatial data
- Questions? Questions about software?

# Further Readings


- Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2, pp. 337-472). Pacific Grove, CA: Duxbury. Chapter 7
- Casella, G. (1985). An introduction to empirical Bayes data analysis. The American Statistician, 39(2), 83-87.
- Robinson, D. (2017). Introduction to Empirical Bayes: Examples from Baseball Statistics. 2017.