

MAD-CB



Empirical Bayes

Baseball de Novo!?!?

- Mais estatístico dos esportes

Baseball de Novo!?!?

- Mais estatístico dos esportes
- Sabermetrics

Baseball de Novo!?!?

- Mais estatístico dos esportes
- Sabermetrics
- Moneyball

Análise Bayesiana

- Depois da minha apresentação da teorema de Bayes
- Deixei de falar disso
- Boa série de posts num blog mais um novo livro
- Oferece um novo approach a análise Bayesiana
- Baseada em beisebol
- Esta apresentação uma adaptação do post e do livro

Blog Post e Livro

- Blog Post: “Understanding empirical Bayes estimation (using baseball statistics)” de David Robinson do blog *Variance Explained*, 10 de outubro 2015
- Livro: David Robinson, **Introduction to Empirical Bayes: Examples from Baseball Statistics** (San Francisco: Gumroad Publishers, 2016). Pode obter (de graça) através do site:
<http://store.varianceexplained.org>.



Empirical Bayes

- Técnica Bayesiana bem útil para estudos modernos de dados biomédicos.
 - ▶ Microarray, sequenciamento, etc.
 - ▶ Estudos com grandes quantidades de dados
- Quando tem muitas observações, só tem uma pequena diferença entre modelos tradicionais Bayesianos e a aproximação de Empirical Bayes (EB)
- Quando tem poucas observações, aproximações de EB podem errar
- EB um atalho para entender métodos Bayesianos
 - ▶ Métodos Bayesianos tradicionais
 - ▶ São difíceis a entender
 - ▶ São custosos em termos de tempo de computação

Data

- Vamos examinar a média de rebatidas, um índice de habilidade de rebatedores
- A habilidade fundamental (?)
- Base de dados Lahman
- Tem a história completa de todos as vezes ao bastão (at-bats)
- Eliminar arremessadores – rebatedores horríveis

Nossa Pergunta

Quem são os melhores rebatedores na historia de beisebol americano?

Média de Rebatidas

- Número de rebatidas (H) dividido pelas vezes ao bastão (AB)

$$BA = \frac{H}{AB}$$

- Tem muito outras medidas de habilidade de rebatedores, mas esse é a mais tradicional - Em MLB, .270 (27%) é considerado uma boa média e .300 excelente - Como padrão, pensamos que a maioria das médias ficam entre .210 e .360 para uma temporada

Distribuição de Rebatidas

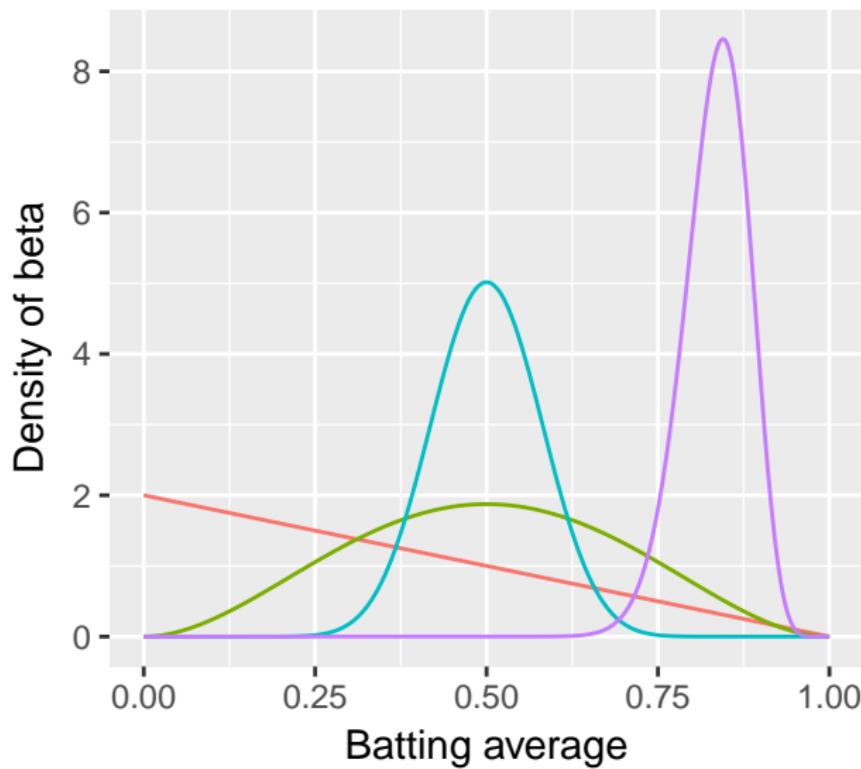
- Rebatidas são exemplos de ‘sucessos’ de uma distribuição *binomial*
 - ▶ Vezes ao bastão = trials
- Guardar na cabeça a pergunta seguinte: Quem é o melhor rebatedor?
 - ▶ José que tem 4 rebatidas em 10 vezes ao bastão?
 - ▶ Pedro que tem 100 rebatidas em 300 vezes ao bastão?
- Distribuição *binomial* tem forte relação com a distribuição *beta*
 - ▶ *Beta* é priori conjugada (“conjugate prior”) da *binomial*

Ponto Fundamental da Teorema de Bayes

- Você pode ter uma estimativa da distribuição e valor da variável antes de medi-la
 - ▶ Priori
- Medida dos dados permite que você pode mudar a conclusão sobre valor
 - ▶ Posteriori
- Distribuições da Priori e da Posteriori devem ser da mesma família
- Como *beta* (priori) e *binomial* (posteriori)

Distribuição Beta

- Tem dois parâmetros: α e β
- Pode assumir formas diferentes baseados nos valores dos parâmetros
- Todos as formas têm valores entre 0 e 1
- Ideal para análise de probabilidades (ou BA)

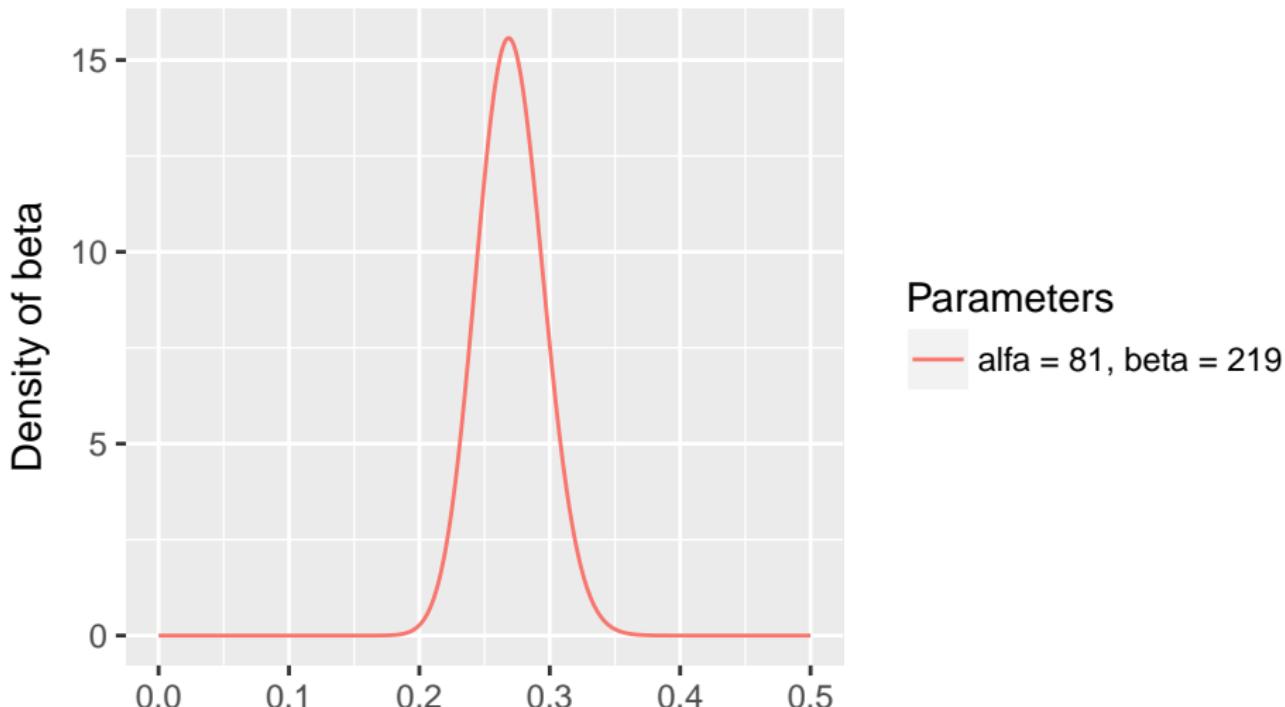


Parameters

- $\text{alfa} = 1$, $\text{beta} = 2$
- $\text{alfa} = 3$, $\text{beta} = 3$
- $\text{alfa} = 20$, $\text{beta} = 20$
- $\text{alfa} = 50$, $\text{beta} = 10$

Aplicar a Distribuição Beta a Média das Rebatidas

- Sabemos que os jogadores rebatem no intervalo de .210 - .360
- Podemos modelar esta condição com uma distribuição Beta(81, 219)



Quando Uma Temporada Inicia . . .

- O que seria nossa expectativa da BA de um rebatedor?
 - ▶ Que cairia na distribuição das médias do ano anterior
 - ▶ Que estaria a média de desta distribuição
- Este é a estimativa *priori* da BA
- Média de uma distribuição Beta:

$$\mu_{\text{Beta}} = \frac{\alpha}{\alpha + \beta} = \frac{81}{81 + 219} = 0.270$$

Atualização

- Voltaremos ao José e Pedro
- Depois de um tempo, queremos determinar a média de rebatidas justa deles
- Eles dois começam a temporada com uma expectativa priori de 0.270
- Agora temos nova informação sobre cada um:
 - ▶ José: 4 de 10 (0.400)
 - ▶ Pedro: 100 de 300 (0.333)
- Podemos calcular a nova distribuição Beta com a nova informação
 - ▶ Processo de atualização

Formula para Atualização de Beta

$$\text{Beta}(\alpha_o + H, \beta_o + (AB - H))$$

- Em termos gerais,
 - ▶ α ajustado pelos sucessos
 - ▶ β ajustado pelos falhas (ou tentativas menos sucessos)
- Para recalcular a média da nova Beta, só precisa substituir os novos valores para α e β

Quem É Melhor? – José?

- BA Posteriori de José

$$\text{Beta}(81 + 4, 219 + 6) = \text{Beta}(85, 225)$$

$$\text{BA}_{\text{José}} = \frac{85}{85 + 225} = 0.2741935$$

Quem É Melhor? – Pedro?

- BA Posteriori de Pedro

$$\text{Beta}(81 + 100, 219 + 200) = \text{Beta}(181, 419)$$

$$\text{BA}_{Pedro} = \frac{181}{181 + 419} = 0.3016667$$

Encolhimento

- Percebem que as duas BA's agora são menores que a BA “instantâneo” (.400 e .333)
- José caiu muito mais: .400 -> .274
- Pedro menos: .333 -> .302
- Com a distribuição Beta, tem uma regressão à média
- Os casos com menos informação (4 de 10) tem regressão maior
- Os com mais (100 de 300) tem menos
- Formalmente: o processo de mudar nossas estimativas na direção da média

Mais Um Experimento com Distribuições Priori e Posteriori

- As distribuições conjugadas – Beta e Binomial
- Criamos um universo com 1 milhão de rebatedores que vivem num mundo com uma distribuição priori de rebatidas de Beta(81, 219)
- Vamos dar para eles 300 vezes na bastão (como Pedro)
- A distribuição Beta dará para nos a média de rebatidas exata
- A Binomial contará quantos rebatidas eles conseguem em 300 AB (trials)

Código para Simulação

```
trials <- 10e5
set.seed(42)
sims <- data_frame(
  ba_verd = rbeta(trials, 81, 219), ## priori com beta
  hits = rbinom(trials, 300, ba_verd)
)
```

Data_Frame (Tibble) sims

```
sims
```

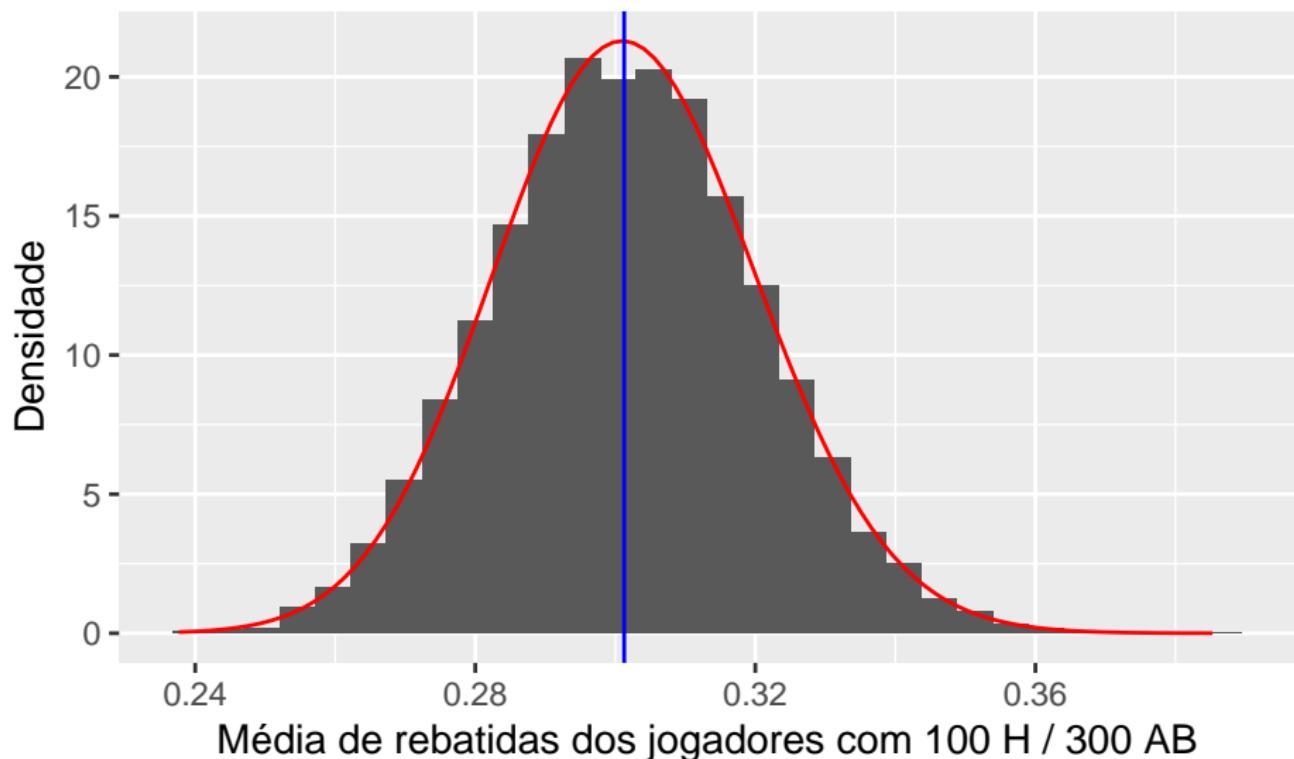
```
## # A tibble: 1,000,000 × 2
##       ba_verd   hits
##       <dbl> <int>
## 1 0.2537397     79
## 2 0.2807047     96
## 3 0.2890551     93
## 4 0.2819510     93
## 5 0.2669373     83
## 6 0.3208808     98
## # ... with 1e+06 more rows
```

Comparação com Pedro (Sr. 100/300)

- Quantos jogadores teve o mesmo # de rebatidas que Pedro e o que eram as médias deles?

```
hit_100 <- sims %>%
  filter(hits == 100)
med_BA_H100 <- median(hit_100$ba_verd)
# Histograma
dens <- function(x) dbeta(x, 81 + 100, 219 + 200)
h100hist <- ggplot(hit_100, aes(ba_verd))
h100hist <- h100hist + geom_histogram(aes(y = ..density..), bins = 30)
h100hist <- h100hist + stat_function(color = "red", fun = dens)
h100hist <- h100hist +
  labs(x = "Média de rebatidas dos jogadores com 100 H / 300 AB")
h100hist <- h100hist + labs(y = "Densidade")
h100hist <- h100hist + geom_vline(aes(xintercept = med_BA_H100), color = "blue")
```

Histograma



Valor Mediana dos Rebatedores com 100 Rebatidas

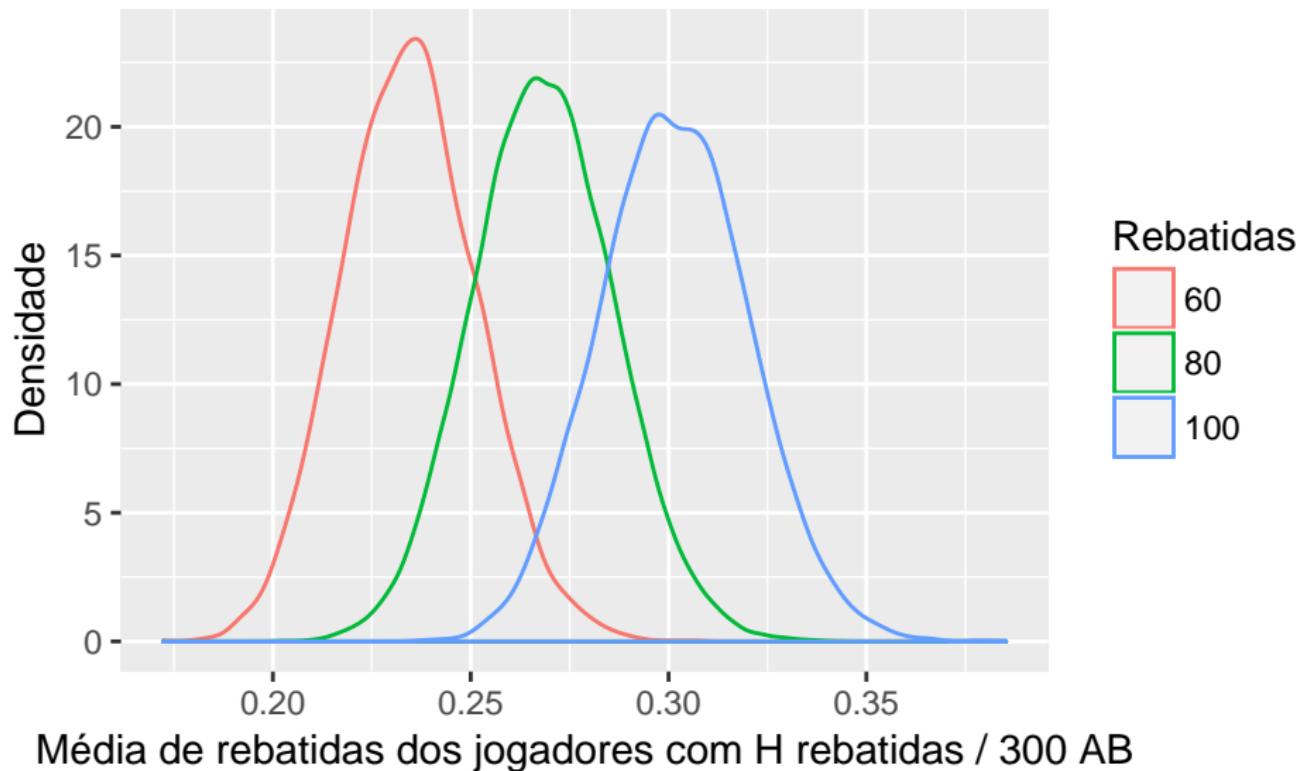
- Valor mediana = 0.301
- Valor do Pedro = 0.302

O Que Está Fazendo a Atualização Bayesiana?

- Responde a pergunta: baseado na informação apriori, qual tipo de rebatedor consegue 100 rebatidas?
- Ou, em termos gerais:
 - ▶ Baseado na informação apriori, qual tipo de caso conseguiria esses resultados?

Caso de Rebatedores com 60 ou 80 Rebatidas Invés de 100

```
hist2 <- sims %>%
  filter(hits %in% c(60, 80, 100)) %>%
  ggplot(aes(ba_verd, color = factor(hits))) +
  geom_density() +
  labs(x = "Média de rebatidas dos jogadores com H rebatidas / 300 AB",
       color = "Rebatidas") +
  labs(y = "Densidade")
```



Resultado da Simulação

- Forma das distribuições posteriores são parecidas
- Deslocam para refletir a nova evidência
- Análise Bayesiana coloca os dados sob estudo no contexto da toda a história da fenômeno sendo investigado
- VSS: Pode ficar enganado pela incerteza quando tem poucos dados (4 de 10)

Método Empirical Bayes

- Na simulação, aprendemos de usar a distribuição Beta para representar suas expectativas apriori
- Usar atualização com nova evidência para fazer sua estimativa mais precisa
- Com Empirical Bayes, vamos estender este conceito para usar a distribuição Beta sobre todos os dados para melhorar cada observação

Back to Baseball – Quem É o Melhor Rebatedor desde 1890?

- Extrair e limpar os dados de Lahman base de dados de rebatidas
- Tirar arremessadores da base

Preparação dos Dados – 1

```
carreira <- Batting %>%
  filter(AB > 0 & yearID > 1890) %>%
  anti_join(Pitching, by = "playerID") %>%
  group_by(playerID) %>%
  summarize(H = sum(H), AB = sum(AB)) %>%
  mutate(average = H / AB)
```

Preparação dos Dados – 2 – Incluir Nomes

```
carreira <- Master %>%
 tbl_df() %>%
  select(playerID, nameFirst, nameLast, finalGame) %>%
  unite(name, nameFirst, nameLast, sep = " ") %>%
  inner_join(carreira, by = "playerID") %>%
  mutate(finalYear = substr(finalGame, 1, 4)) %>%  select(-finalGame)
```

Exploração dos Dados – 1

carreira

```
## # A tibble: 8,642 × 6
##   playerID      name     H     AB   average finalYear
##   <chr>        <chr> <int> <int>    <dbl>    <chr>
## 1 aaronha01   Hank Aaron  3771 12364  0.3049984   1976
## 2 aaronto01   Tommie Aaron   216   944  0.2288136   1971
## 3 abadan01    Andy Abad      2     21  0.0952381   2006
## 4 abbated01   Ed Abbaticchio  772  3044  0.2536137   1910
## 5 abbotfr01   Fred Abbott    107   513  0.2085770   1905
## 6 abbotje01   Jeff Abbott    157   596  0.2634228   2001
## # ... with 8,636 more rows
```

Exploração dos Dados – 2 – Quem São os Melhores?

```
head(arrange(carreira, desc(average)), 10)
```

```
## # A tibble: 10 × 6
##   playerID          name     H     AB average finalYear
##   <chr>            <chr> <int> <int>    <dbl>    <chr>
## 1 banisje01        Jeff Banister  1      1       1  1991
## 2 bassdo01         Doc Bass     1      1       1  1918
## 3 birasst01        Steve Biras   2      2       1  1944
## 4 burnscb01        C. B. Burns   1      1       1  1902
## 5 gallaja01        Jackie Gallagher 1      1       1  1923
## 6 gleasro01         Roy Gleason   1      1       1  1963
## 7 hopkimi01         Mike Hopkins  2      2       1  1902
## 8 kuczest01         Steve Kuczek   1      1       1  1949
## 9 liddeda01         Dave Liddell  1      1       1  1990
## 10 lindsch02        Charlie Lindstrom 1      1       1  1958
```

Jeff Bannister – Acho que Não

- Atual treinador de Texas Rangers

Jeff Bannister – Acho que Não

- Atual treinador de Texas Rangers
- 1 AB em 1991

Jeff Bannister – Acho que Não

- Atual treinador de Texas Rangers
- 1 AB em 1991
- Conseguiu uma rebatida

Jeff Bannister – Acho que Não

- Atual treinador de Texas Rangers
- 1 AB em 1991
- Conseguiu uma rebatida
- Único jogo para os Pirates nos major leagues

Jeff Bannister – Acho que Não

- Atual treinador de Texas Rangers
- 1 AB em 1991
- Conseguiu uma rebatida
- Único jogo para os Pirates nos major leagues
- **Sortudo**

Refinar os Dados

- Os dados podem refletir melhor jogadores com carreiras longas

Refinar os Dados

- Os dados podem refletir melhor jogadores com carreiras longas
- Um rebatedor que joga uma temporada inteira teria > 300 ABs

Refinar os Dados

- Os dados podem refletir melhor jogadores com carreiras longas
- Um rebatedor que joga uma temporada inteira teria > 300 ABs
- Revisar nosso data_frame para ter um mínimo de 300 AB's

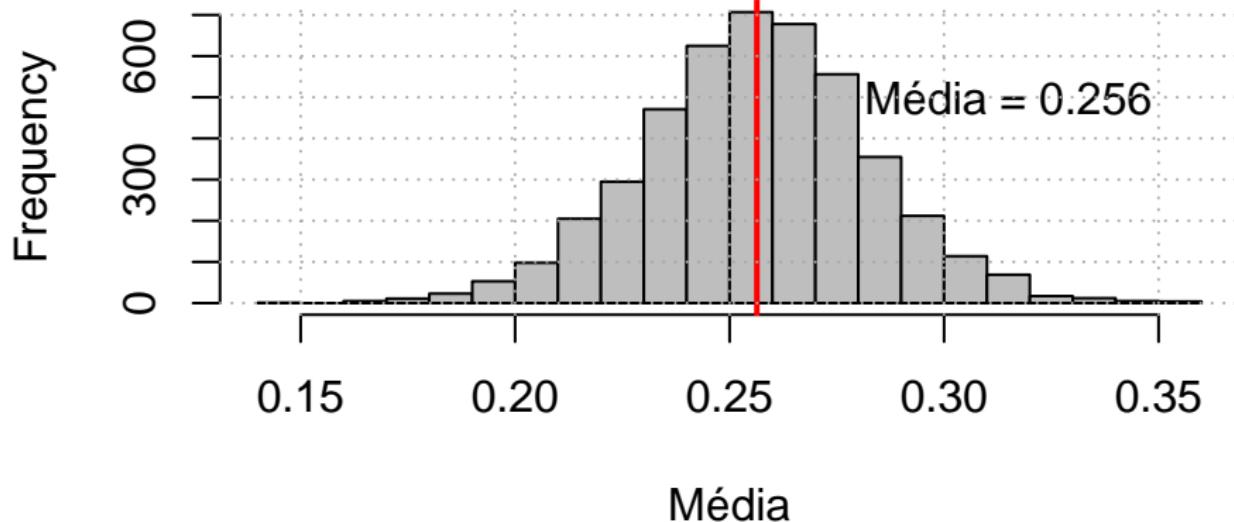
Código para a Revisão dos Dados

```
carrfiltrado <- carreira %>% filter(AB >= 300)
head(carrfiltrado)
```

```
## # A tibble: 6 × 6
##   playerID          name     H     AB    average finalYear
##   <chr>            <chr> <int> <int>      <dbl>      <chr>
## 1 aaronha01      Hank Aaron  3771 12364  0.3049984  1976
## 2 aaronto01      Tommie Aaron   216   944  0.2288136  1971
## 3 abbated01     Ed Abbaticchio   772  3044  0.2536137  1910
## 4 abbotfr01      Fred Abbott    107   513  0.2085770  1905
## 5 abbotje01      Jeff Abbott    157   596  0.2634228  2001
## 6 abbotku01      Kurt Abbott    523  2044  0.2558708  2001
```

Histograma da Distribuição das Médias de Rebatidas

Histograma das Médias de Rebatidas



Passo 1 de Empirical Bayes: Estimar Uma apriori Usando Todos os Dados

- Uma método Bayesiana pura estimaria uma distribuição priori usando outra informação, não dados
- Empirical Bayes é uma aproximação aos métodos Bayesianos mais ortodoxos
- Queremos determinar a distribuição Beta que replica esses dados
 - ▶ Um novo α_0 e β_0
 - ▶ Os *hiperparâmetros* de nosso modelo

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

Pode Escolher Os Hiperparâmetros com Probabilidade Máxima

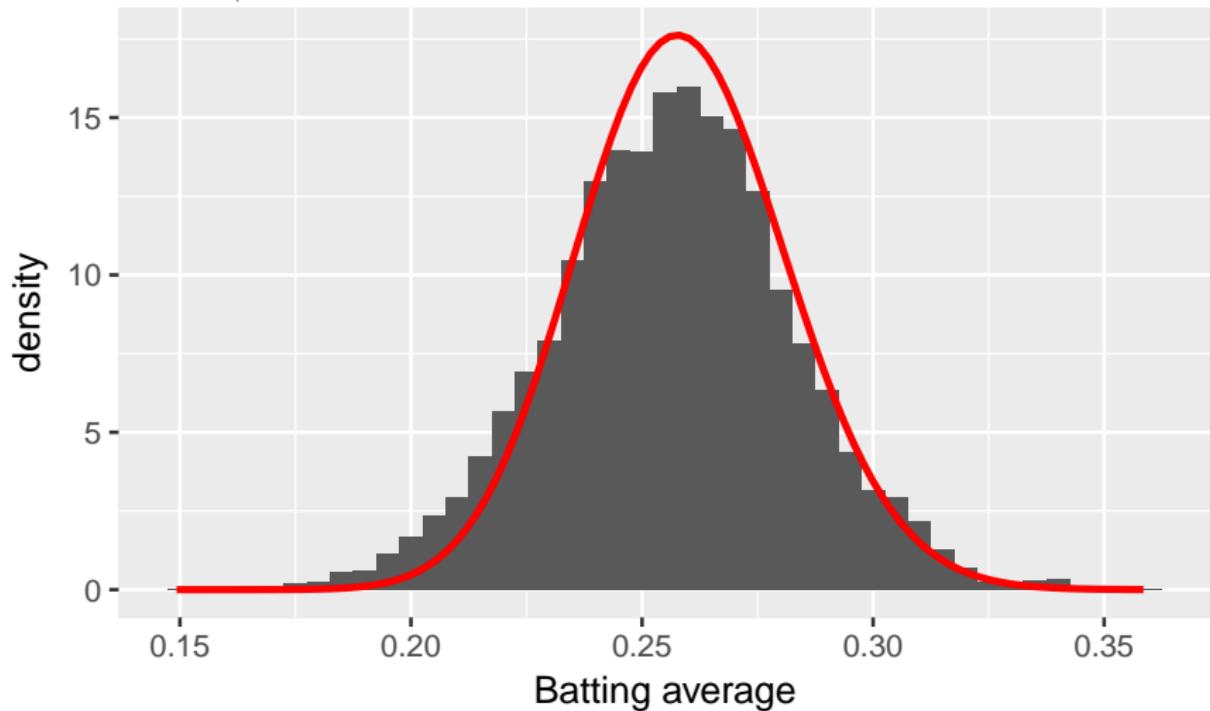
- Probabilidade máxima = maximum likelihood
- Método para achar os parâmetros para maximizar a probabilidade de gerar a distribuição empírica
- Porque estamos usando as distribuições Beta e Binomial
 - ▶ Usaremos a distribuição betabinom para calcular a probabilidade máxima
 - ▶ Vem do pacote VGAM (que deveria ter no seu computador)
- Função para maximizar a probabilidade é `mle` que faz parte do pacote `stats4`
 - ▶ Um grupo das funções programado num outra sub-sistema de R chamado S4

Calculo de α_0 e β_0

```
# log-likelihood function
ll <- function(alpha, beta) {
  x <- carrfiltrado$H
  total <- carrfiltrado$AB
  -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)
alpha0 <- ab[1]
beta0 <- ab[2]
```

Novos Hiperparâmetros

- $\alpha = 96.818; \beta = 277.062$



Parece Mais como Beisebol Verdadeiro

- Replica bem os dados
- Média histórica de 0.256
- Começamos com 8642 jogadores
- Agora: 4516 jogadores com um mínimo de 300 AB's
- Redução de 47.74 porcentagem

Passo 2 – Use a Nova Distribuição como Priori para Cada Rebatedor

- Mesmo processo que fizemos com a simulação anterior
- Estimar a média de todos os individuais
 - ▶ Começar com priori universal
 - ▶ Fazer a atualização baseado no desempenho de cada jogador
- Chamaremos esta estimativa a “estimativa de empirical Bayes” (*EEB*)
- Formula:

$$EEB = \frac{H + \alpha}{AB + \alpha + \beta}$$

```
ebb <- carreira %>%
  mutate(eb_est = (H + alpha0) / (AB + alpha0 + beta0))
```

Nova Lista dos Melhores Rebatedores

```
showebb <- ebb %>%
  select(-playerID) %>%
  arrange(desc(eb_est)) %>%
  slice(1:10)
```

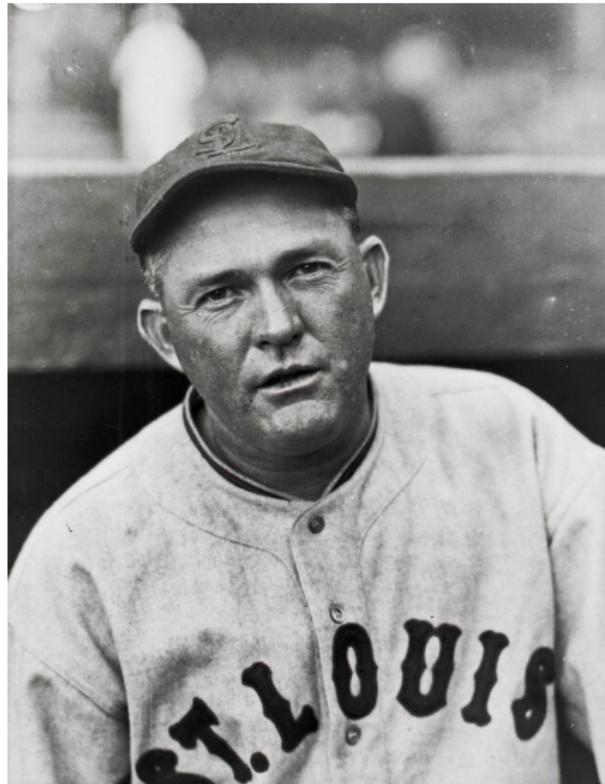
Table 1: Melhores Rebatedores por EBB

name	H	AB	average	finalYear	eb_est
Rogers Hornsby	2930	8173	0.3584975	1937	0.3541430
Ed Delahanty	2305	6452	0.3572536	1903	0.3518693
Shoeless Joe Jackson	1772	4981	0.3557519	1920	0.3489934
Billy Hamilton	1802	5109	0.3527109	1901	0.3463176
Willie Keeler	2932	8591	0.3412874	1910	0.3378537
Harry Heilmann	2660	7787	0.3415950	1932	0.3378089
Bill Terry	2193	6428	0.3411637	1936	0.3366448
Lou Gehrig	2721	8001	0.3400825	1939	0.3364607
Nap Lajoie	3242	9589	0.3380957	1916	0.3351258
Tony Gwynn	3141	9288	0.3381783	2001	0.3351126

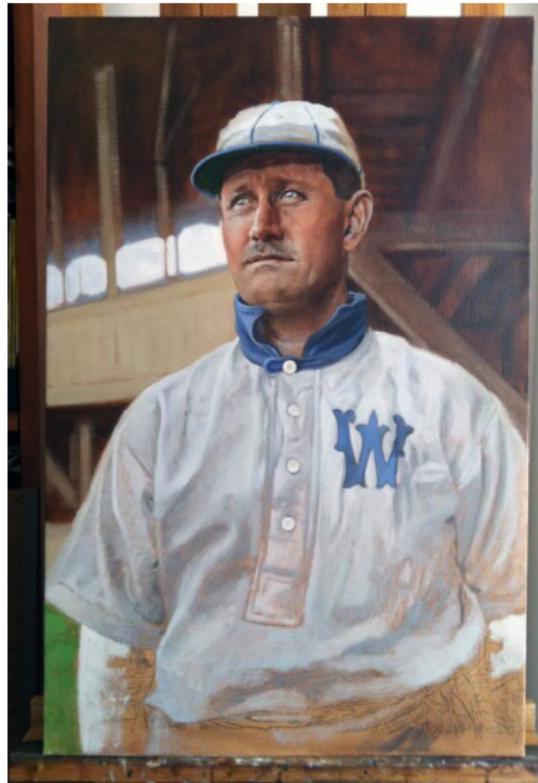
Comentários

- Rogers Hornsby geralmente considerado o melhor rebatedor na historia do beisebol
- Ed Delahanty produto de Seculo 19 – desfecho interessante na carreira (e vida)
- Shoeless Joe Jackson podia ter ido para cima ou para baixo
 - ▶ Mandado embora do beisebol depois de escândalo “Black Sox” de 1919
- Único jogador moderno na lista - Tony Gwynn, aposentado em 2001
- Omissão interessante da lista: Ted Williams
- Último jogador para rebater 0.400 (1952 e 1953)
- “Cientista” de rebater - o “Perfesser”

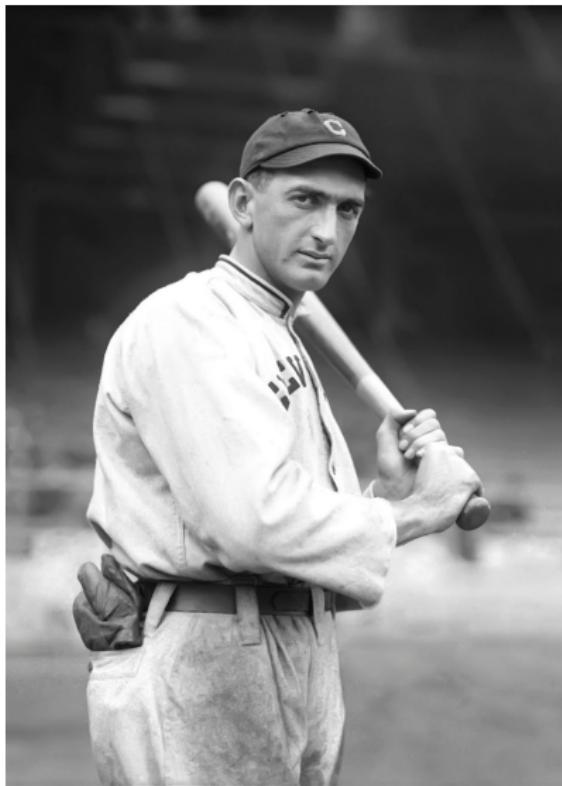
Rogers Hornsby



“Big” Ed Delahanty



Shoeless Joe Jackson

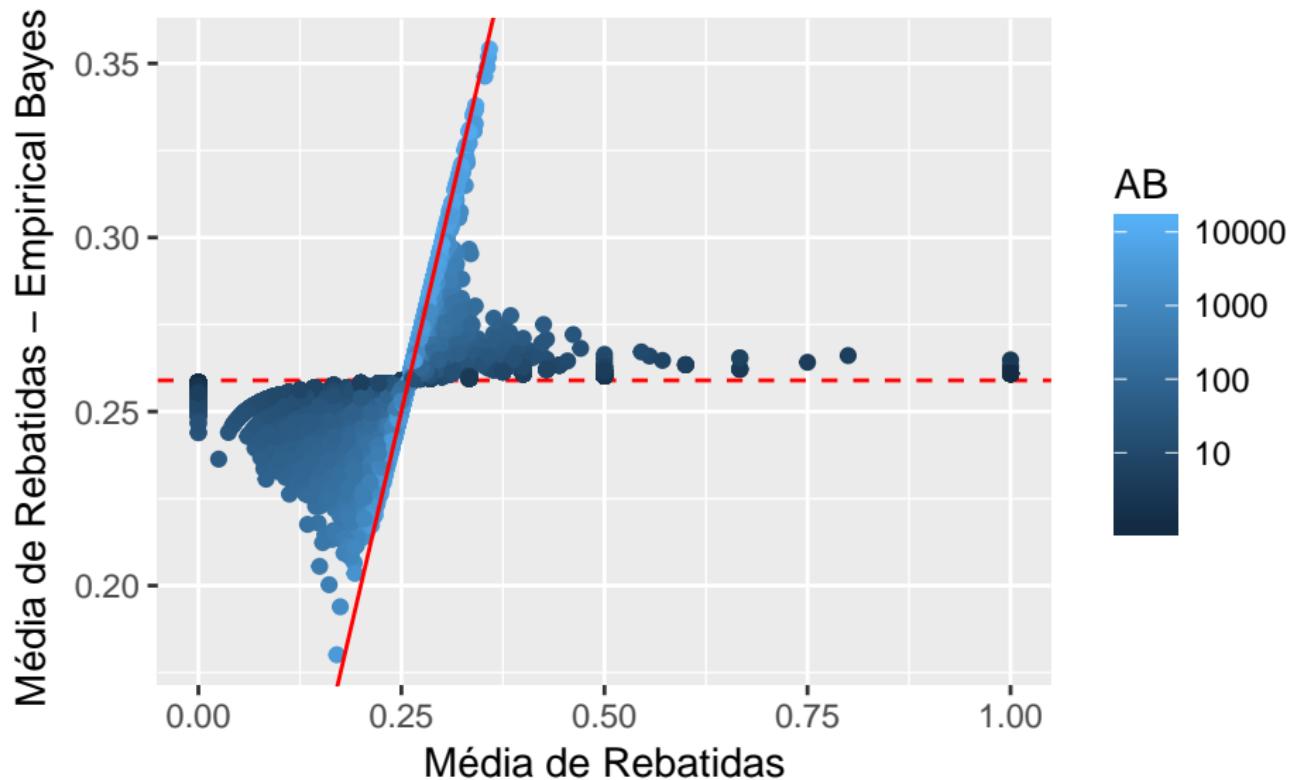


Resultado de Empirical Bayes

- Mudou as médias de todos os jogadores
- Scatter plot mostra as mudanças

Código para Scatter Plot

```
ebscat <- ggplot(ebb, aes(average,
                           eb_est, color = AB)) +
  geom_hline(yintercept = alpha0 / (alpha0 + beta0),
              color = "red", lty = 2) +
  geom_point() +
  geom_abline(color = "red") +
  scale_colour_gradient(trans = "log", breaks = 10 ^ (1:5)) +
  xlab("Média de Rebatidas") +
  ylab("Média de Rebatidas - Empirical Bayes ")
```



Como Interpretar o Gráfico

- A linha vermelha tracejada é a média apriori $y = \frac{\alpha_0}{\alpha_0 + \beta_0} = 0.259$
- A linha vermelha sólida mostra onde $x = y$
- Os pontos mais claros são mais próximos à linha xy
 - ▶ Eles são os rebatedores com mais informação, mais ABs
 - ▶ Eles sofrem encolhimento muito menor
- Os pontos mais escuras são aqueles com poucos ABs (como Jeff Bannister)
 - ▶ Eles sofrem muito mais encolhimento que os rebatedores com experiência

Limitações deste Modelo

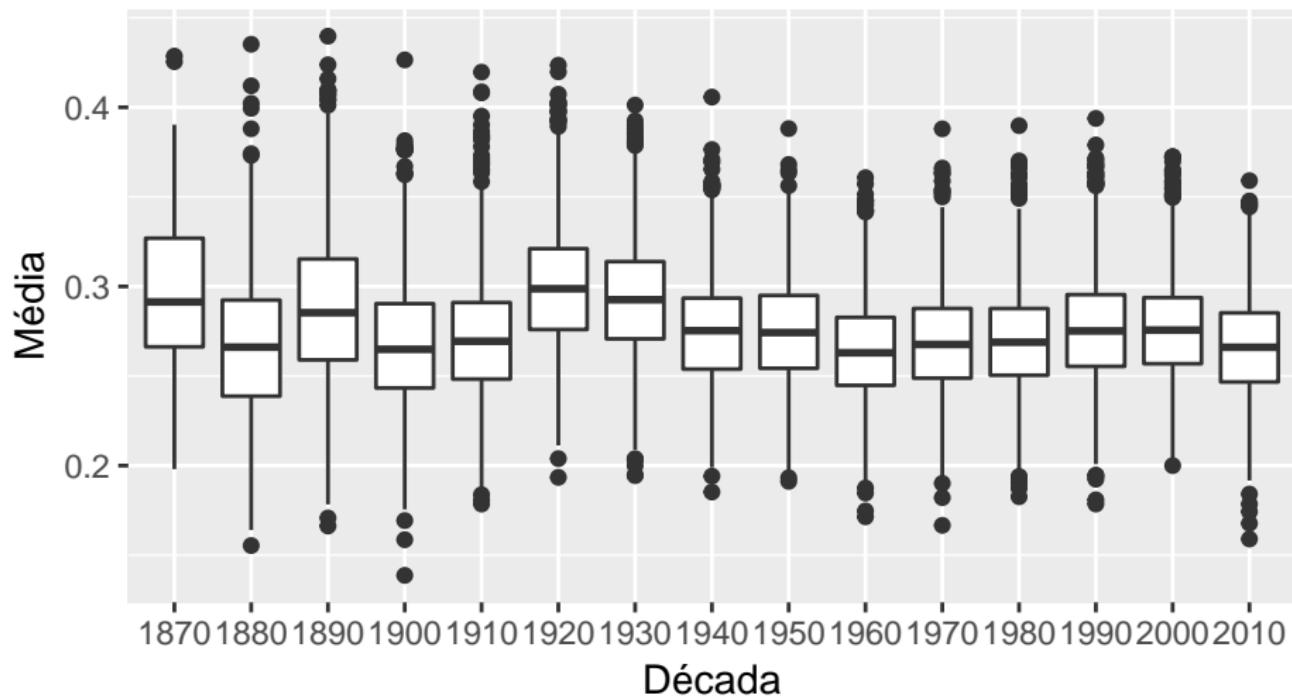
- Nosso experimento assume que rebater em beisebol sempre teve as mesmas restrições, regras, e parâmetros
- Beisebol está sempre evoluindo, regras e equipamentos mudando
- Condicionamento dos jogadores muito diferente do que até 30 anos atrás
- Tivemos a época da bola morta— “dead ball”, época de expansão, época de PEDs, época de arremessadores com velocidade fenomenal
- Podemos considerar as médias de rebatidas por década, podemos ver que estimativas precisam ser ajustadas para tomar conta dessas diferenças.

Data Frame das Decadas

```
decadas <- Batting %>% mutate(decada = floor(yearID/10)*10) %>%
  mutate(Average = H / AB) %>%
  filter(AB > 300)
```

Graph of Batting through the Decades

Média de Rebatidas por Década



Modelos Hierárquicos Bayesianos

- Esta questão de década pode ser tratado por um modelo Empirical Bayes mais sofisticado
- Modelos desenvolvidos em sequência
- Frequentemente usados no estudo de regulação epigenética da expressão genética

Recomendação Final

- Curso de MIT - The Analytics Edge MIT 15.071x
 - ▶ Próxima sessão começa 06/06/17
 - ▶ <https://www.edx.org/course/analytics-edge-mitx-15-071x-3>
 - ▶ * * * * * Eu fiz; amei

That's all Folks!



kalilak