Lecture 5: Hierarchical Modeling

Professor Alexander Franks

Announcements

- Reading: Chapter 5 of BDA

Comparing Multiple Related Groups

- Hierarchy of nested populations
- Models which account for this are called *hiearchical* or *multi-level* models

Some examples:

- Patient outcomes within several different hospitals
- People within counties in the United States (e.g. Asthma mortality example)
- Athlete performance in sports
- Genes within a group of animals

Eight schools example

- A study was performed for the Educational Testing Service (ETS) to evaluate the effects of coaching programs on SAT preparation
- Each of eight different schools used a short-term SAT prep coaching program
- Compute the average SAT score in those who did take the program minus those that did not participate in the program
- We observe the average difference varies by school. What accounts for these differences?

Eight schools example

sampling

- Interested in “real” differences due to training
- Want to reduce effect of chance variability
- How do we estimate the effect of the program in each of the schools?

Have: $Y_1, \ldots Y_8$ is the estimate of the causal effect.

Want: $\mu_1, \ldots \mu_8$

Eight schools example

- Consider two extremes:
  - Estimate the effect of the program in every school independently
    - A separate prior distribution for each school effect
  - Or assume the effect is the same in every school
    - Combine all the data
  - A compromise between the above 2 options?

Eight Schools Example

known

$$y_j \sim N(\theta_j, \sigma_j^2)$$

- y_j is the observed effects of the program in school j
  - Based on a sample of test scores from those in the program and those not in the program
- θ_j are the true *unknown* effects of the program in school j
- Assume variances, σ_j^2, are *known*
  - e.g. determined by the number of students in the sample

Eight Schools Example

```r
J <- 8
y = c(28,  8, -3,  7, -1,  1, 18, 12)
sigma <- c(15, 10, 16, 11,  9, 11, 10, 18)
```

- Assuming the effect of the program on each school is identical.
- What are the chances of seeing a value as large as 28?
- As small as -3?

Eight Schools Example

- Assume the effect of the program on each school is identical, i.e. $\theta_j = \mu$
- Assume a flat prior on μ, what is $p(\mu \mid y_1, \ldots, y_8, \sigma_1, \ldots \sigma_8)$?

$P(\mu) \propto \text{const.}$

$$\propto L(\mu) \propto \prod^{8} e^{-\frac{(y_i-\mu)^2}{2\sigma_i^2}}$$

$$\propto \exp\left(-\sum^{8} \frac{1}{2\sigma_i^2}\left(\mu^2 - 2y_i\mu\right)\right)$$

$$\propto \exp\left(-\tfrac{1}{2}\sum \frac{1}{\sigma_i^2}\left(\mu - \frac{\sum y_i}{\sum 1/\sigma_i^2}\right)^2\right)$$

$$N\left(\sum w_i y_i, \frac{1}{\sum w_i}\right)$$

$$w_i = \frac{1/\sigma_i^2}{\sum 1/\sigma_j^2}$$ <u>Fisher Weighting</u>

$$\hat{\mu}_{MLE} = \sum w_i y_i$$

Eight Schools Example

```r
## Compute the precision frome each school
prec <- 1/sigma^2

## global estimate is a weighted vareage
mu_global <- sum(prec * y / sum(prec))
mu_global
```

```
[1] 7.685617
```

$$7.69 = \hat{\mu} = \hat{\theta}_1 = \dots = \hat{\theta}_8$$

Eight Schools Example

- Assume the effect of the program on each school is identical, i.e. $\theta_j = \mu$
- What are the chances of school 1 having an effect large as 28 (given $\sigma_1 = 15$)?
- Y_3 as small as -3 (given $\sigma_3 = 16$)?

Posterior Prediction Under Complete Pooling

```r
prec <- 1/sigma^2

## global estimate is a weighted average
mu_global <- sum(prec * y / sum(prec))

print(sprintf("mu is %f", mu_global))
```

$= 7.69$

```
[1] "mu is 7.685617"
```

```r
1 - pnorm(28, mean=mu_global, sd=sqrt(1/sum(1/sigma^2)+sigma[1]^2))
```

```
[1] 0.09560784
```

```r
pnorm(-3, mean=mu_global, sd=sqrt(1/sum(1/sigma^2)+sigma[3]^2))
```

```
[1] 0.2587447
```

$$\mu \sim N(\mu_{global}, 1/\Sigma w_i)$$

$$y_i | \mu \sim N(\mu, \sigma_i^2)$$

$$\tilde{y} | y \sim N(\mu_{global}, 1/\Sigma w_i + \sigma_i^2)$$

Eight Schools Example

$$y_j \sim N(\theta_j, \sigma_j^2)$$

- θ_j are the true unknown effects of the program in school j
- y_j is the observed effects of the program in school j
  - Based on a sample of test scores from those in the program and those not in the program
  - Number of people in the sample determine the magnitude of σ_j^2

Eight Schools Example

How do we estimate θ_j?

- Assume effects are totally independent: $\hat{\theta}_j^{(MLE)} = y_j$ is the MLE
- Assume effects are identical: $\hat{\theta}_j^{(pool)} = \frac{\sum_i \frac{1}{\sigma_i^2} y_i}{\sum \frac{1}{\sigma_i^2}}$
  - Same effect for all schools: estimate using a weighted average of the observed effects

Eight Schools

```r
theta_j_mle <- y
theta_j_mle
```

```
[1] 28  8 -3  7 -1  1 18 12
```

```r
theta_j_pooled <- rep(sum(1/sigma^2 * y) / sum(1/sigma^2), J)
theta_j_pooled
```

```
[1] 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617
```

- Compromise: $\hat{\theta}_j^{\text{shrink}} = w\underbrace{\theta_j^{\text{MLE}}}_{= y_j} + (1-w)\underbrace{\theta^{pooled}}_{\mu_{global}}$

Eight schools example

Add a *shared* normal prior distribution to θ_j

$$\theta_i \sim N(\mu, \tau^2)$$ — Shared Prior

$$y_j \sim N(\theta_j, \sigma_j^2)$$

$p(\mu) \propto$ const (hyperprior)

- The global mean, μ, is also an unknown parameter. What prior should we choose?
- τ^2 determines how much weight weight we put on the independent estimate vs the pooled estimate.
- A 9-parameter posterior:

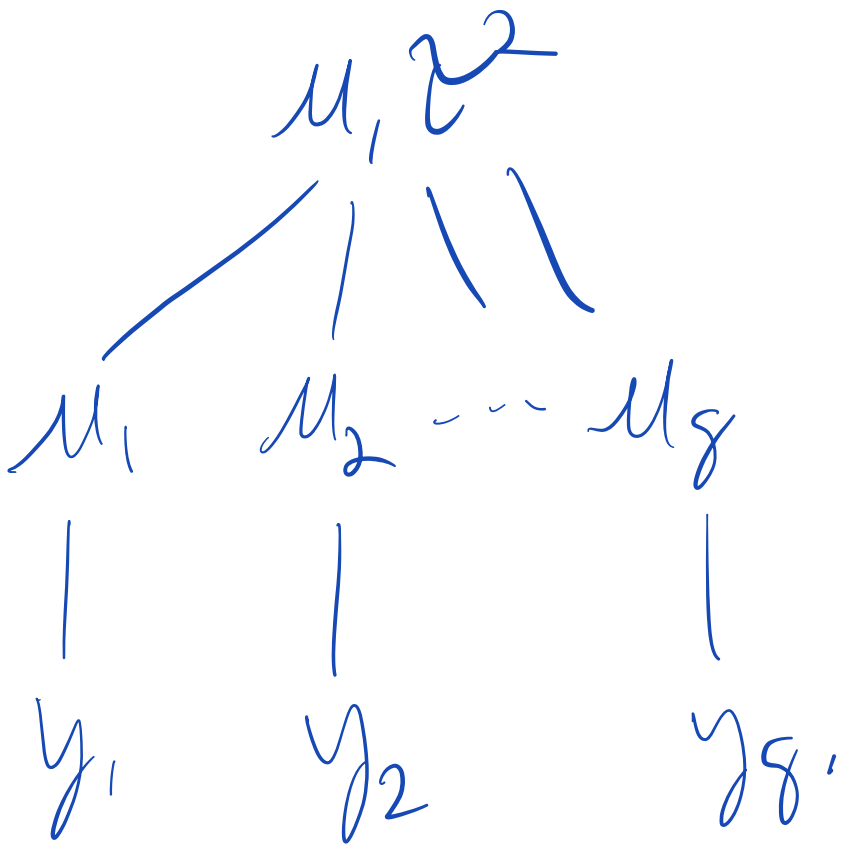
$$p(\mu, \theta_1, \ldots, \theta_8 \mid y_1, \ldots, y_8, \sigma_1, \ldots, \sigma_8, \tau^2)$$

If μ & τ^2 are known.

$$P(\theta_1 \mid y_1, \cancel{y_2, \ldots, y_8}, \mu, \tau^2)$$

$$\sim N\left(w_1 y_1 + (1-w)\mu, \frac{1}{1/\sigma_1^2 + 1/\tau^2}\right)$$

$$w = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\tau^2}$$



Intuition for shrinkage

$\theta_j \sim N(\mu, \tau^2)$, $\epsilon_j \sim N(0, \sigma_j^2)$

- $Y_j = \theta_j + \epsilon_j$
  - For simplicity assume $Var(\epsilon_j) = \sigma^2$ for all j
  - θ_j represents true effect in school j (signal)
  - $Var(\hat{\theta}_j) = \tau^2$ represents how much the true effects vary across schools
    - ϵ_j is sampling variability (noise, chance variation)
- $\hat{\theta}_{MLE} = Y_j$

Intuition for shrinkage

- Consequence: the observed outcomes always have higher variance than the signal, i.e. $\mathrm{Var}(Y_j) > \mathrm{Var}(\theta_j)$
- Intuition: reduce the variance by shrinking estimates to a common mean!
- The variance of the shrunken estimates should be close to τ^2

Reminder

$$\begin{cases} y_j \sim N(\theta_j, \sigma_j^2) \\ \theta_j \sim N(\mu, \tau^2) \end{cases}$$

known

"shrinkage variance"

MLE: $\hat{\theta}_{MLE} = (y_1, \ldots, y_r)$

Eight schools example

Questions:

- Is the training program effective in school j?
  - What is $P(\theta_j > 0 \mid y)$?
- On average (over all schools) is the training program effective?
  - What is $P(\mu > 0 \mid y)$? $E[\mu \mid y]$ (avg benefit)
- Will the training program be effective in a new school?
  - What is $P(\theta_{J+1} > 0 \mid y)$? $p(\theta_{J+1} \mid \mu, \tau, y_1, \ldots y_n)$

Eight schools example

Comments:

- The global average, μ, is a parameter so also has uncertainty
- How dow we determine how much to shrink, e.g. how do we determine τ^2?
- What σ_j^2 were also unknown?

Eight schools example

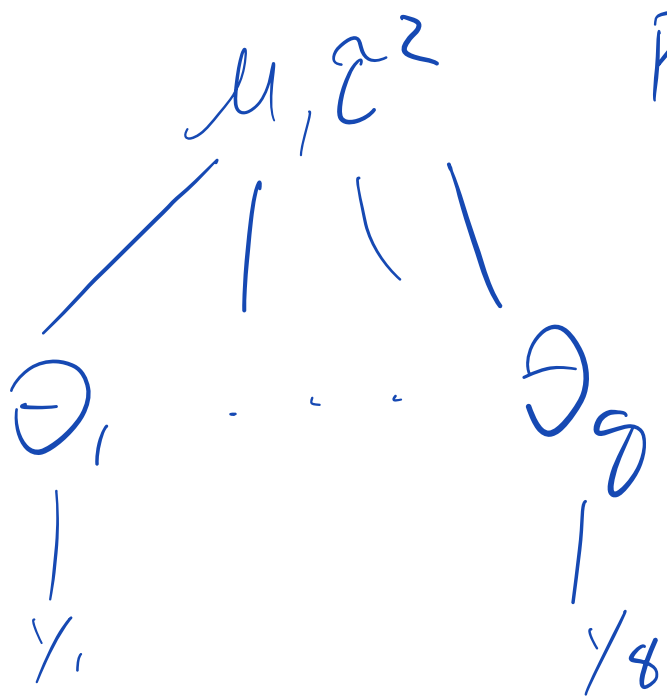
- If τ^2 is large, the prior for θ_j is not very strong
  - If $\tau^2 \to \infty$ equivalent to the no pooling model
- If τ^2 is small, we assume a priori that θ_j are very close
  - if $\tau^2 \to 0$ equivalent to the complete pooling model, $\theta_j = \mu$

Inference

$$P(\mu, \theta_1, \ldots \theta_8 \mid y_1, \ldots y_8, \sigma_i^2)$$

$$\|$$

- Factorize the density into tractable components
  - $p(\mu \mid y_1, \ldots, y_8, \tau^2)$ $\quad P(\mu \mid \tau^2, y's) \prod_{i=1}^{8} P(\theta_i \mid \mu, \tau^2, y_i)$
  - $p(\theta_i \mid \mu, y_i, \tau^2)$
- Later: MCMC or other approximate methods



$$P(\theta_i \mid y_i, \mu) \rightarrow N\left(w_i y_i + (1-w_i)\mu, \frac{1}{1/\sigma_i^2 + 1/\tau^2}\right)$$

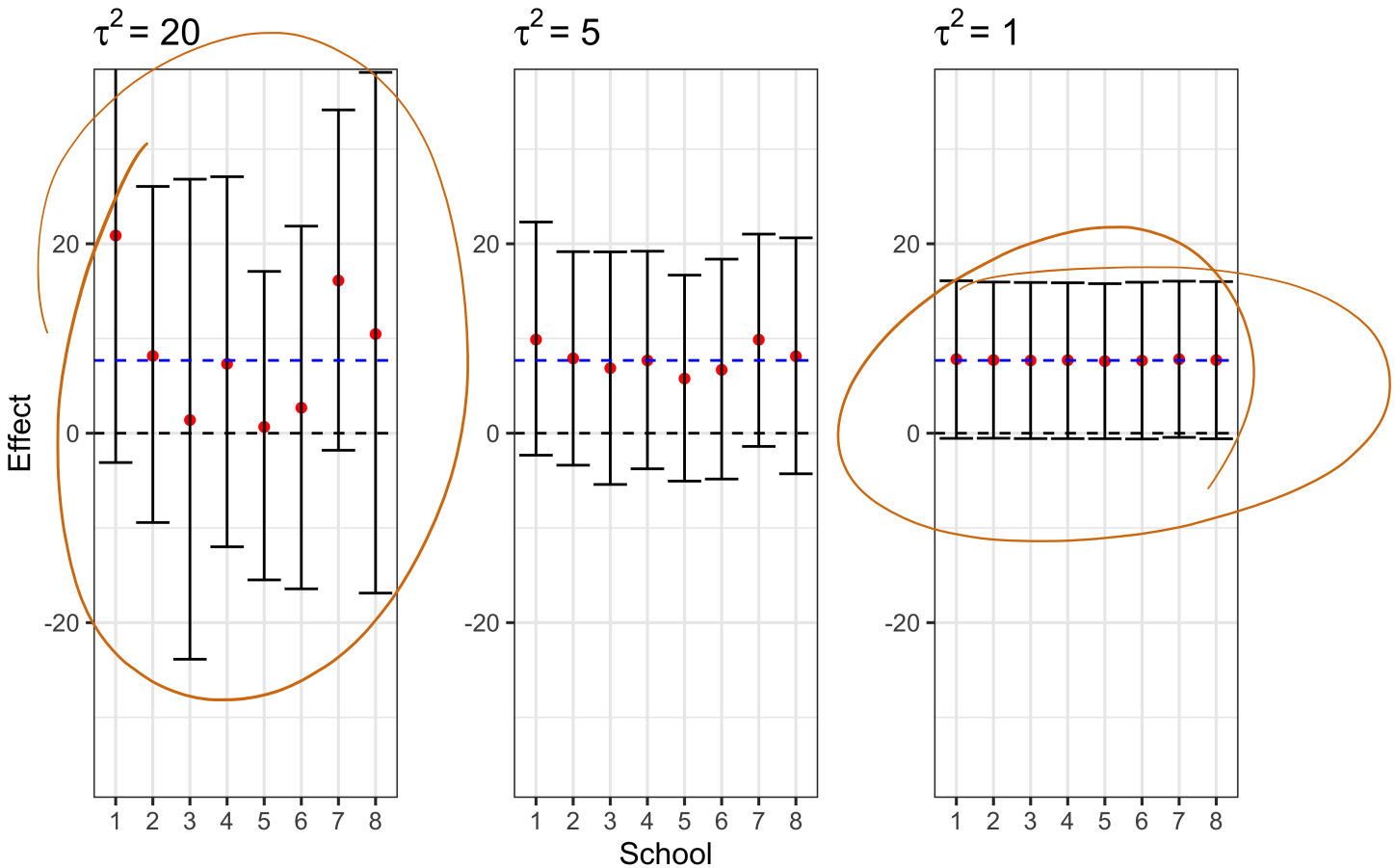
$$w_i = \frac{1/\sigma_i^2}{1/\sigma_i^2 + 1/\tau^2}$$

$$P(\mu \mid y_1, \ldots, y_8, \tau^2)$$

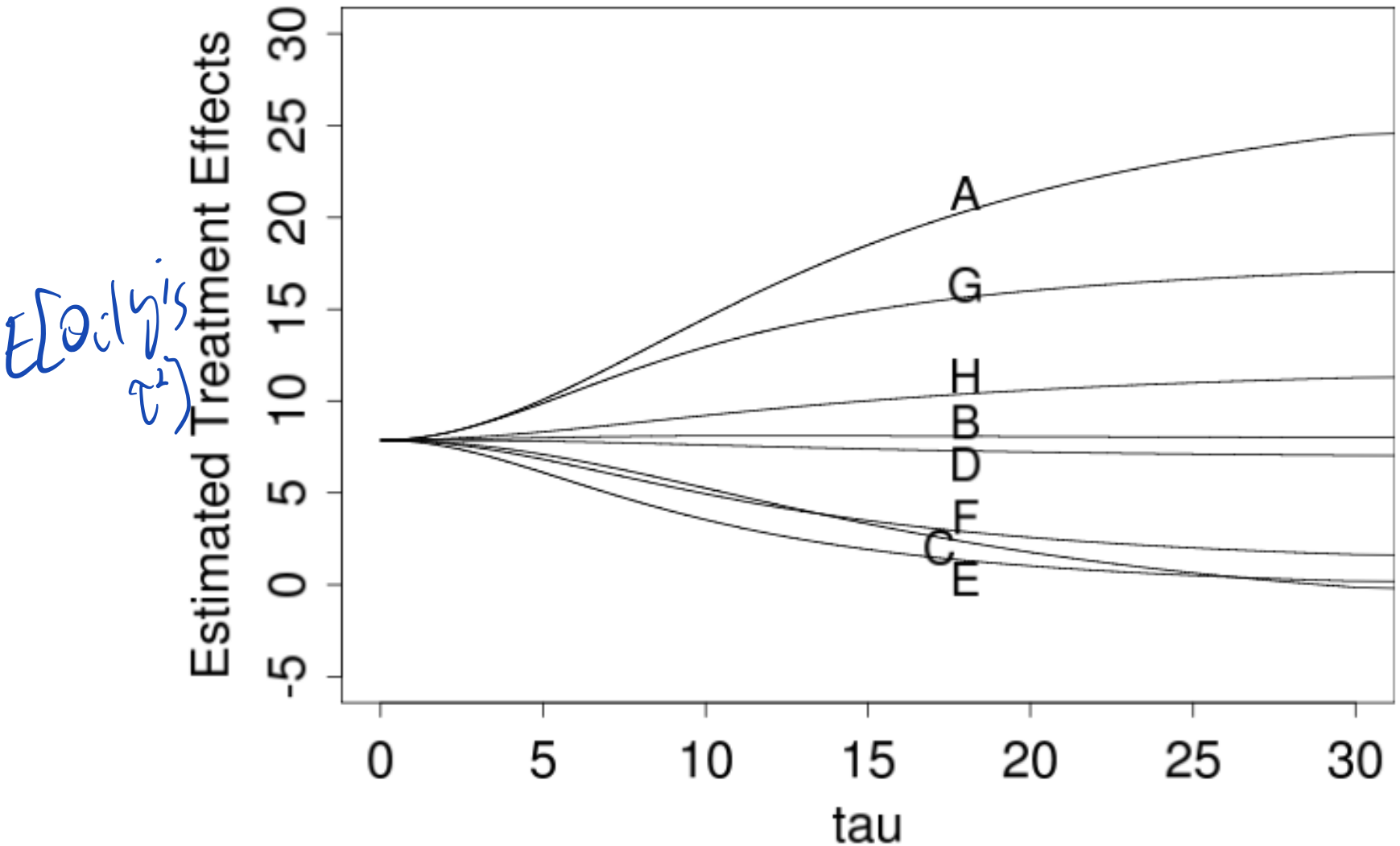
$$\sim N\left(\sum \frac{1}{\sigma_i^2} y_i, \frac{1}{\sum 1/\sigma_i^2}\right)$$

$$y_i \sim N(\mu, \sigma_i^2 + \tau^2)$$

Eight Schools example



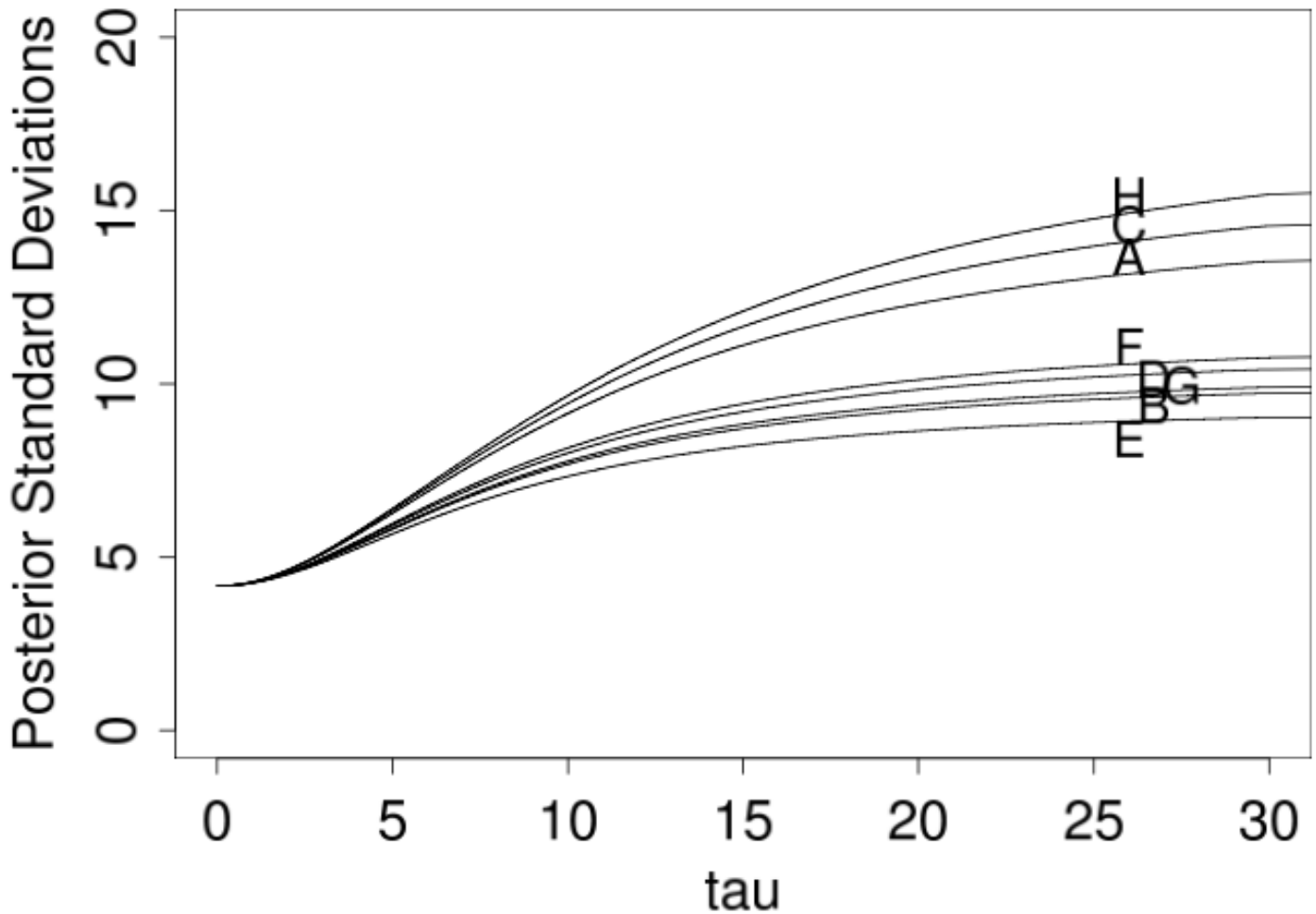
The impact of τ



E[Y(1)-Y(0)|τ²]

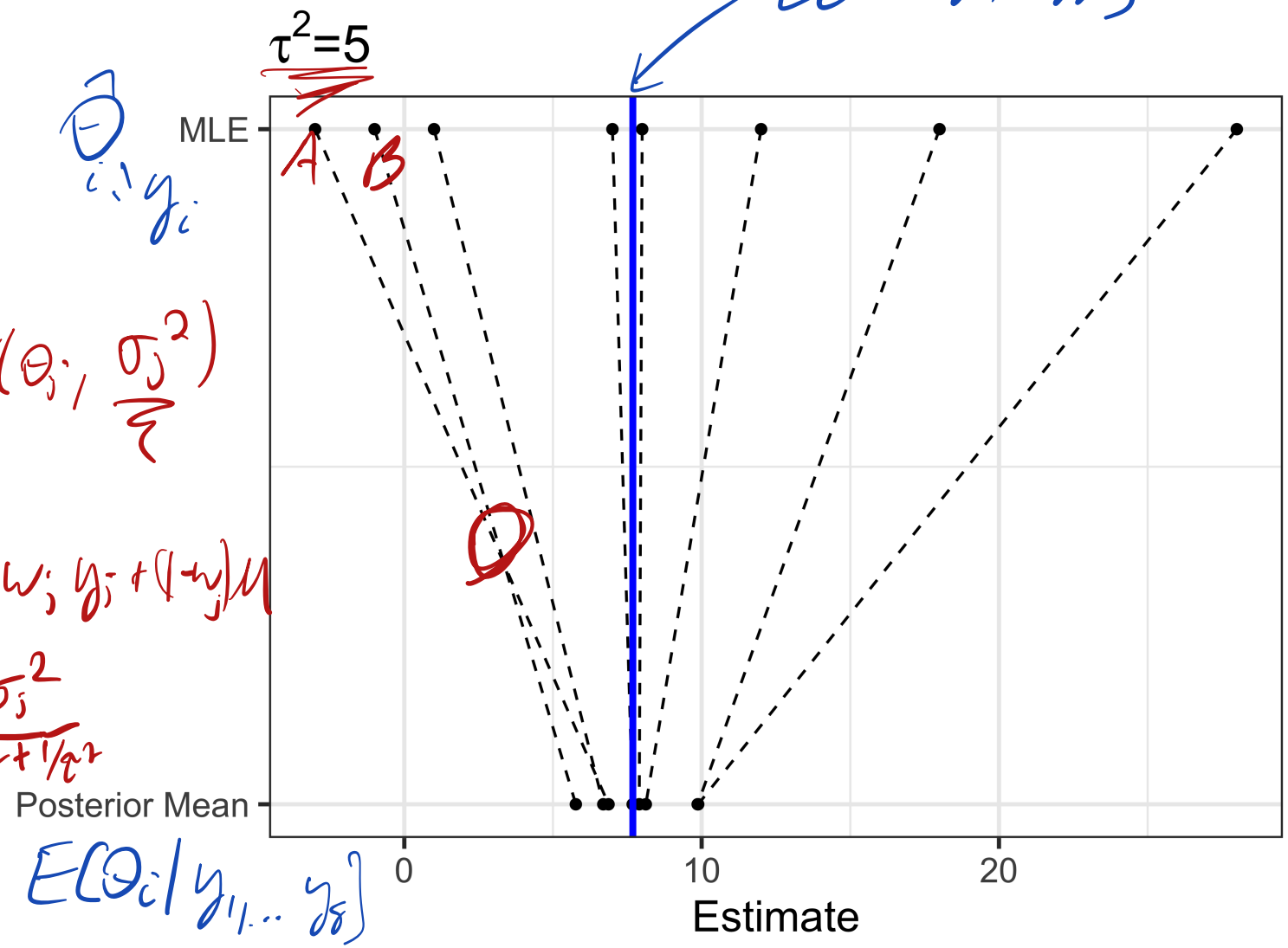
The impact of τ

$SD(\theta_i \mid y)$



MLE vs Posterior Mean

$E[\mu \mid y_1, \ldots y_n]$



$\theta_i \mid y_i$

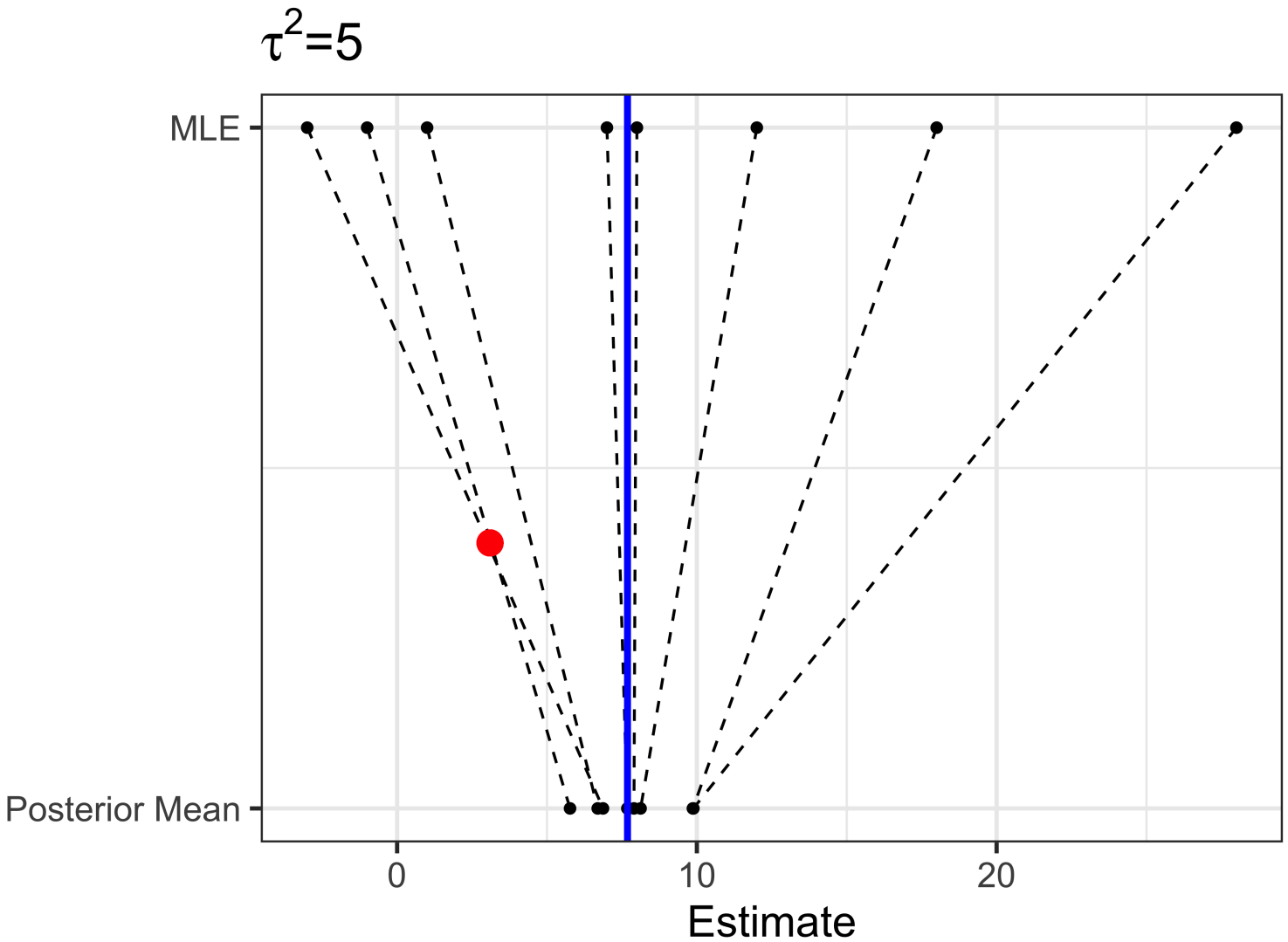
$y_j \sim N(\theta_j, \sigma_j^2)$

$E[\theta_j \mid y_j, \mu] = w_j y_j + (1-w_j)\mu$

$w_j = \dfrac{1/\sigma_j^2}{1/\sigma_j^2 + 1/\tau^2}$

$E[\theta_i \mid y_1, \ldots y_8]$

MLE vs Posterior Mean



Inference for τ

- Can infer τ (don't need to set τ as a hyperparameter)
- How?

$P(\mu, \tau)$ (hyperprior)

Where do we get info about τ^2?

Assume $\sigma_j^2 = \sigma^2$ (known)

$Y_i = \theta_i + \varepsilon \Rightarrow Y_i \sim N(0, \tau^2 + \sigma^2)$

$\theta_i \sim N(\mu, \tau^2)$

$\varepsilon \sim N(0, \sigma^2)$

crude estimate:

$$\frac{\sum_j (y_j - \bar{y})^2}{n-1} - \sigma^2 \approx \tau^2$$

Weak and noninformative Priors on τ

Read Book on priors for τ.

Be careful

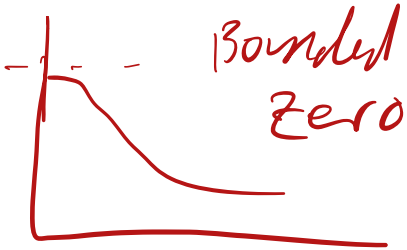
- Consider limits of proper priors
  - Uniform[0, A] as $A \to \infty$ (ok for $J > 2$)
    Unif$[0, \infty]$
  - Inverse-Gamma(ϵ, ϵ) as $\epsilon - > 0$ (improper posterior!)
- Uniform on log(τ) (improper posterior)
  - $p(y \mid \tau) \to const$ as $\tau \to 0$
- Half-Cauchy prior distribution on τ^2
  - Recommended by Gelman et al

$P(\mu, \sigma^2) \propto 1/\sigma^2$

$\Rightarrow P(\log \sigma) \propto$ const.

$P(\tau \mid y) \propto L(\tau) P(\tau)$

$L(\tau)$

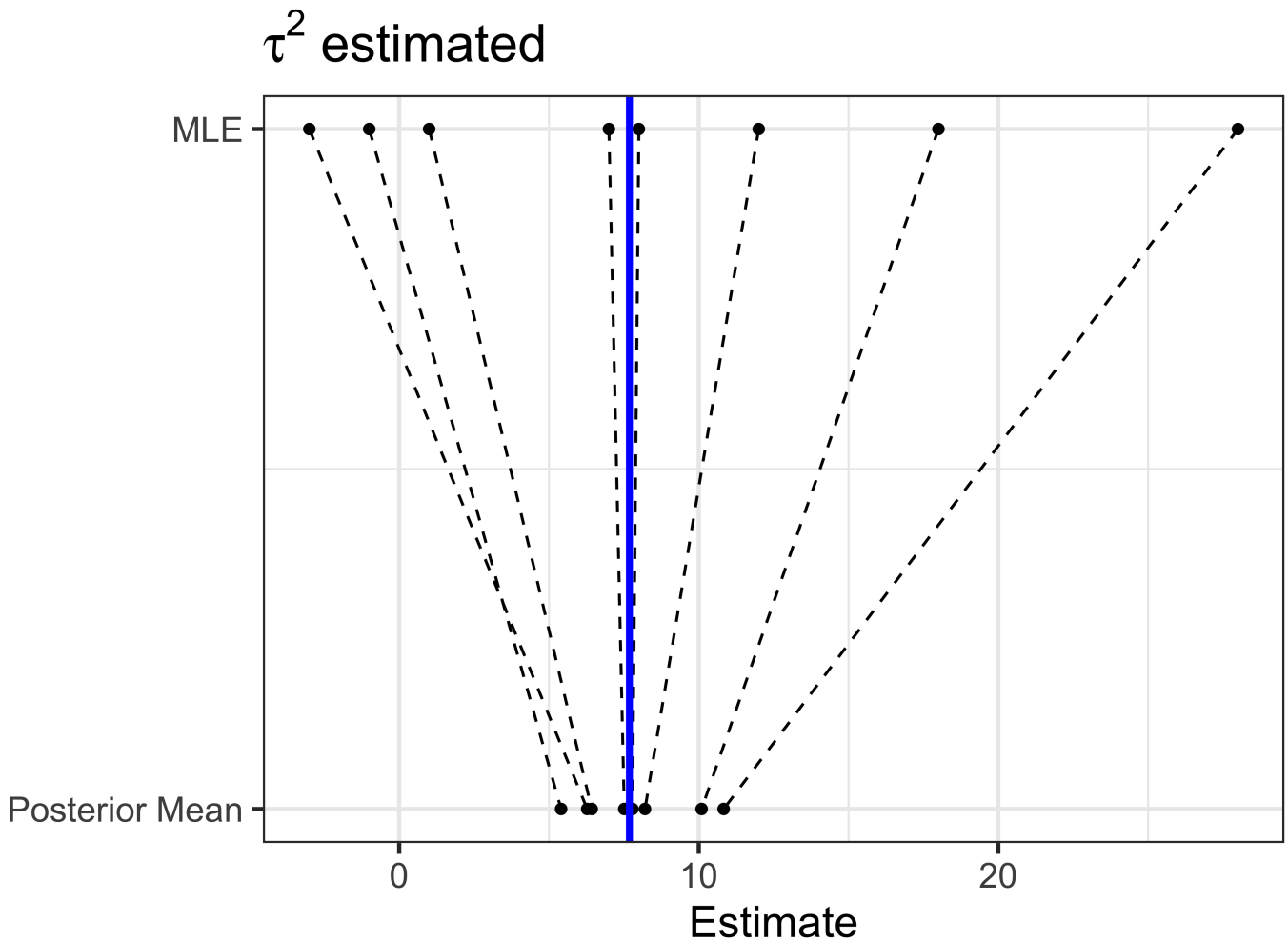


Bounded at zero.

$\int_0^K \text{const} \times \frac{1}{\tau} d\tau$

not finite.

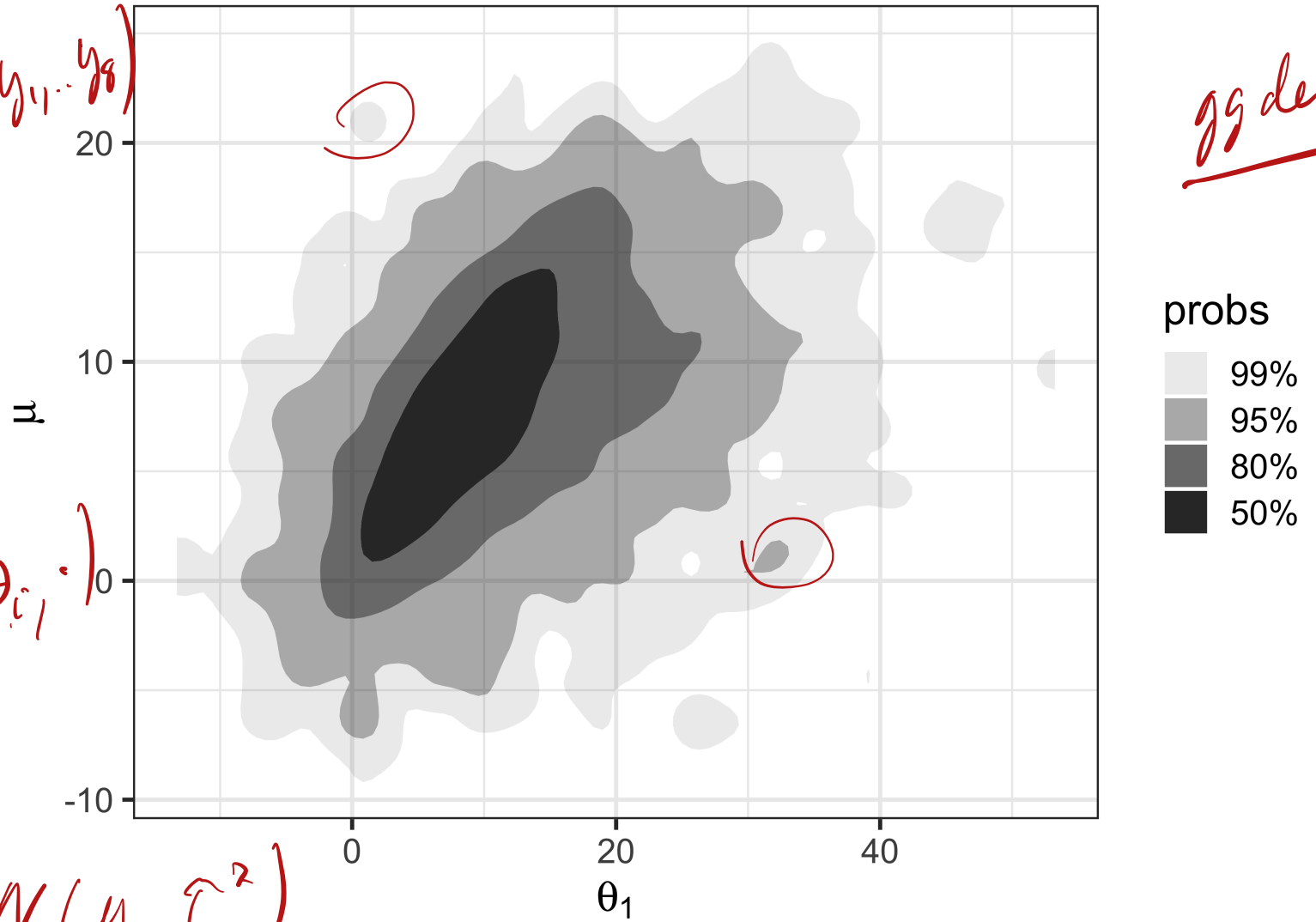
MLE vs Posterior Mean



Posterior mean of $\tau = 5.6422886$

Eight Schools Density Plot: θ_1 vs μ

$P(\mu, \theta_1 \mid y_1, \ldots, y_8)$

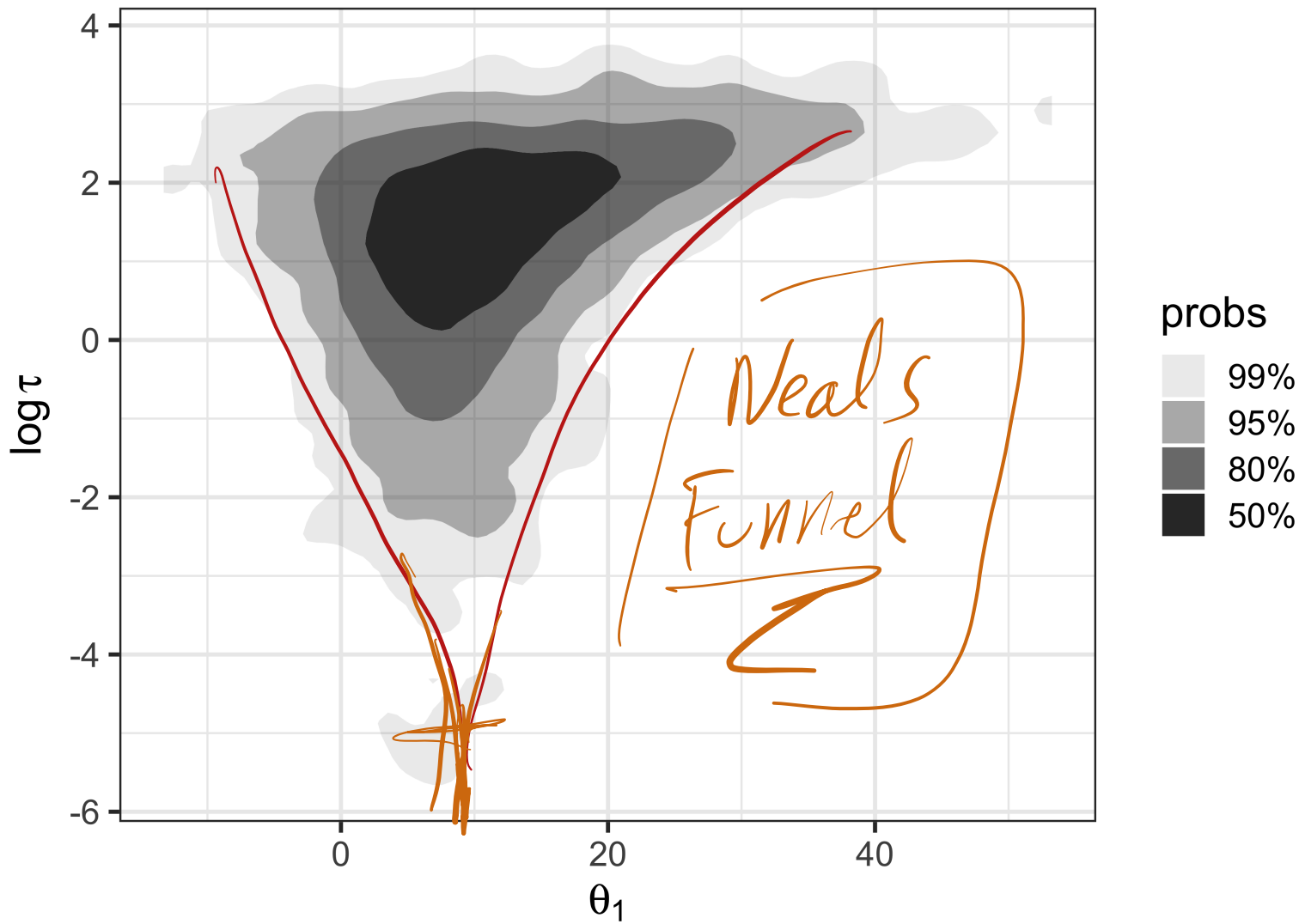


gg density

$y_i \sim N(\theta_i, \cdot)$

$\theta_i \sim N(\mu, \tau^2)$

Eight Schools Density Plot: θ_1 vs $log(\tau)$



Sampling variance

What if the sampling variances are unknown?

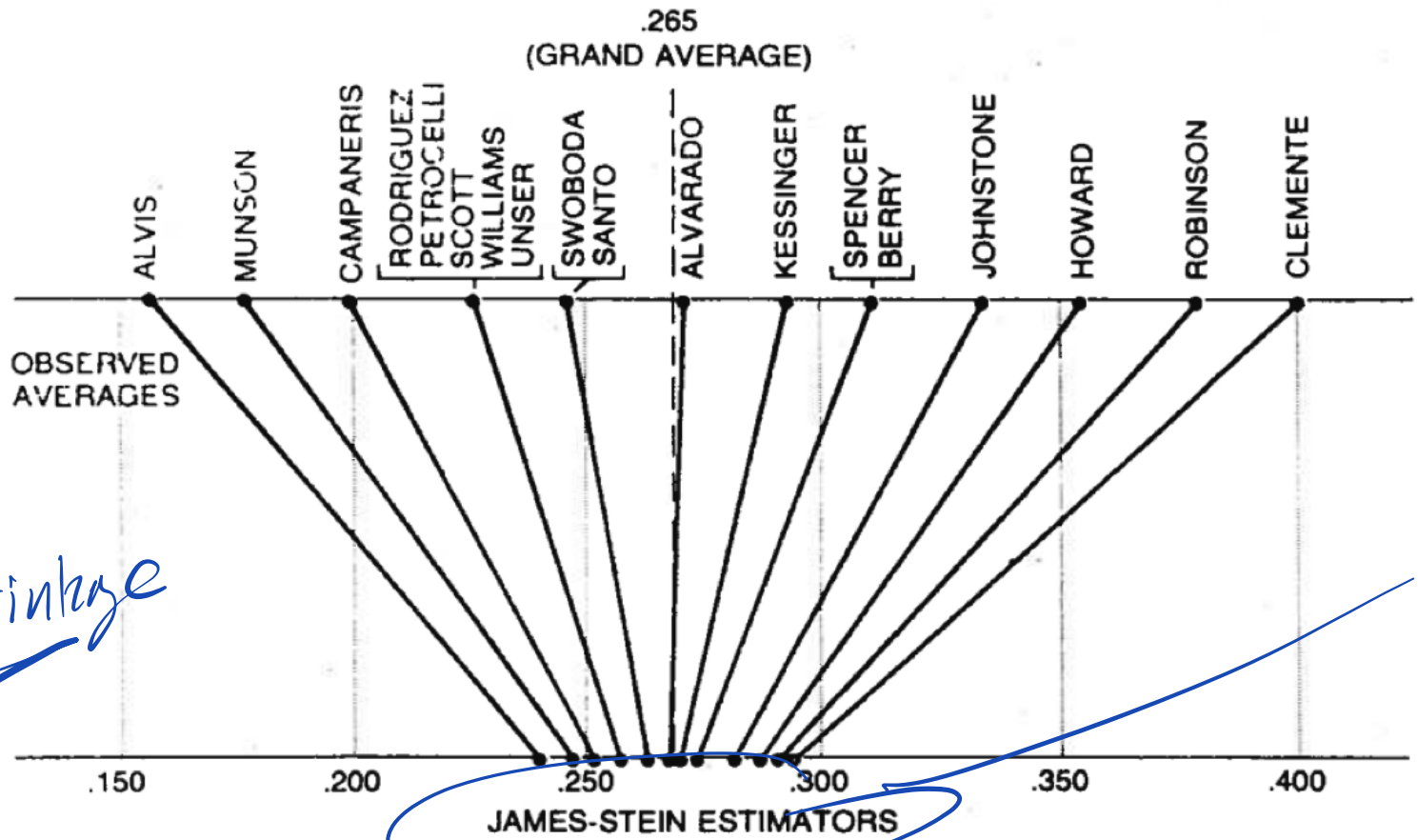
Hierarchical modeling in sports

1. 1970 Batting Averages for 18 Major League Players and Transformed Values X_i, θ_i

i	Player	Y_i = batting average for first 45 at bats (1)	p_i = batting average for remainder of season (2)	At bats for remainder of season (3)	X_i (4)	θ_i (5)
1	Clemente (Pitts, NL)	.400	.346	367	−1.35	−2.10
2	F. Robinson (Balt, AL)	.378	.298	426	−1.66	−2.79
3	F. Howard (Wash, AL)	.356	.276	521	−1.97	−3.11
4	Johnstone (Cal, AL)	.333	.222	275	−2.28	−3.96
5	Berry (Chi, AL)	.311	.273	418	−2.60	−3.17
6	Spencer (Cal, AL)	.311	.270	466	−2.60	−3.20
7	Kessinger (Chi, NL)	.289	.263	586	−2.92	−3.32
8	L. Alvarado (Bos, AL)	.267	.210	138	−3.26	−4.15
9	Santo (Chi, NL)	.244	.269	510	−3.60	−3.23
10	Swoboda (NY, NL)	.244	.230	200	−3.60	−3.83
11	Unser (Wash, AL)	.222	.264	277	−3.95	−3.30
12	Williams (Chi, AL)	.222	.256	270	−3.95	−3.43
13	Scott (Bos, AL)	.222	.303	435	−3.95	−2.71
14	Petrocelli (Bos, AL)	.222	.264	538	−3.95	−3.30
15	E. Rodriguez (KC, AL)	.222	.226	186	−3.95	−3.89
16	Campaneris (Oak, AL)	.200	.285	558	−4.32	−2.98
17	Munson (NY, AL)	.178	.316	408	−4.70	−2.53
18	Alvis (Mil, NL)	.156	.200	70	−5.10	−4.32

y_i y_i'

Hierarchical modeling in sports



Shrinkage

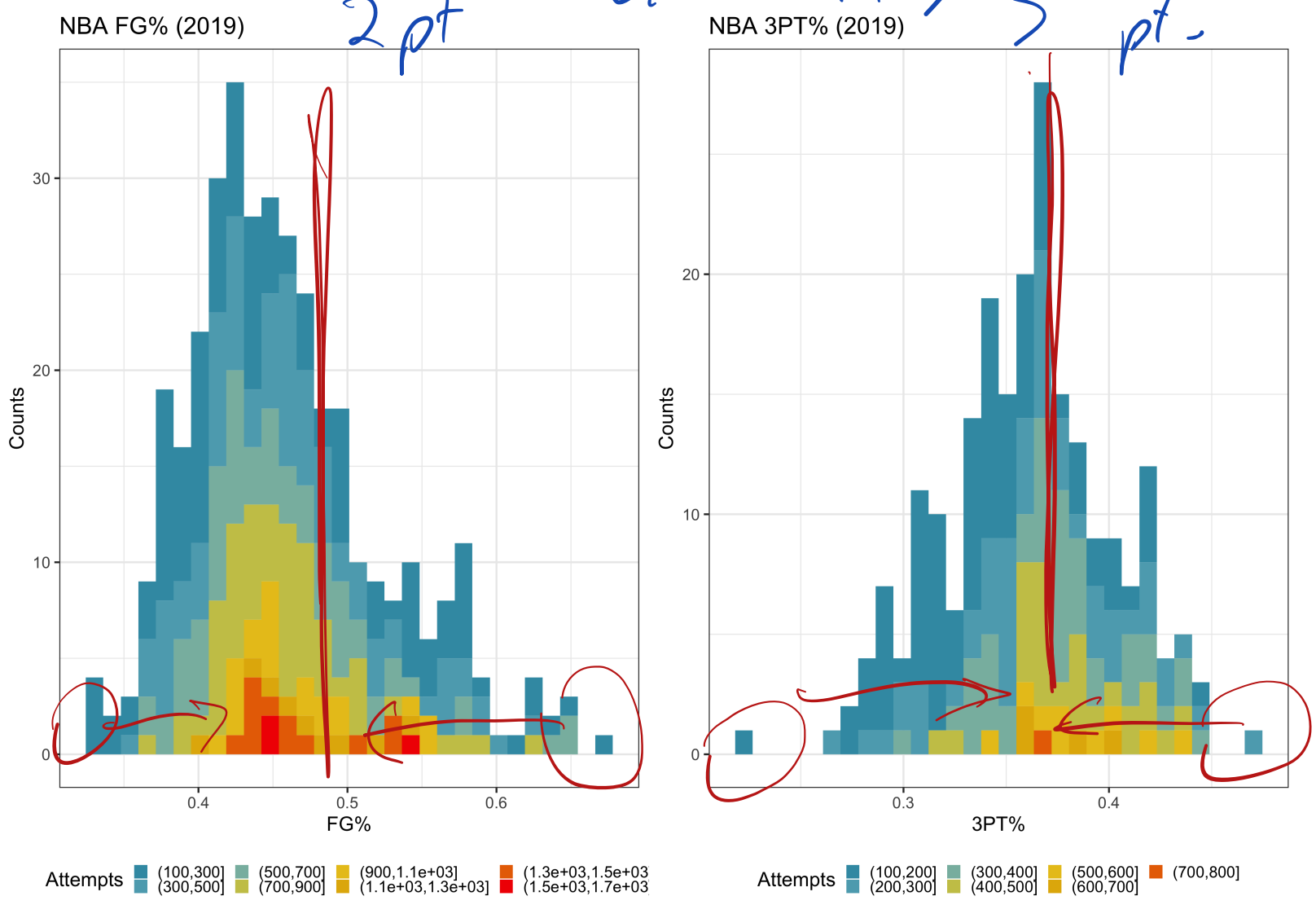
Empirical Bayes.

JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

Basketball Example

$y_i \sim Bin(n_i, \theta_i)$

$\theta_i \sim Beta(a, b)$



A Beta/Binomial Hierarchical Model

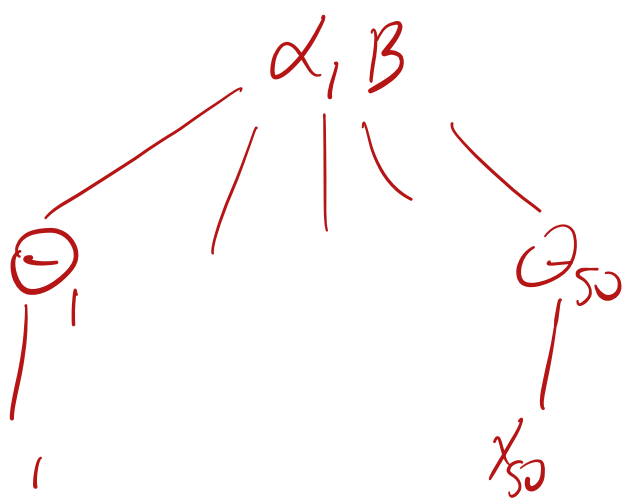
- If we know α and β then the player effects are independent:

$\theta_i \sim Beta(\alpha, \beta)$, $y_i \sim Bin.$

$$p(\alpha, \beta, \theta_1, \ldots, \theta_{50} \mid (y_1, n_1), \ldots (y_{50}, n_{50})) =$$

$$\underbrace{p(\alpha, \beta, \mid (y_1, n_1), \ldots (y_{50}, n_{50}))}_{\text{Marginal Post. (Annoying)}} \prod_i \underbrace{p(\theta_i \mid (y_i, n_i), \alpha, \beta)}_{\text{Easy}}$$

Monte Carlo procedure:



Uniform on $\left(\frac{\alpha}{\alpha+\beta}, \frac{1}{\sqrt{\alpha+\beta}}\right)$

mean

model "sample size"

$$P(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$$

$$P(\alpha, \beta \mid y_1, \ldots y_{50}) \propto \int P(\alpha, \beta, \theta_1, \ldots \theta_{50} \mid y_1, \ldots y_{50})\, d\theta$$

$$= \int \left(\prod P(\theta_i \mid \alpha, \beta, y_i) \right) P(\alpha, \beta)\, d\theta$$

$$= \underbrace{(\alpha+\beta)^{-5/2}}_{P(\alpha,\beta)} \int_0 \underbrace{\prod_{i=1}^{50} P(\theta_i \mid \alpha, \beta, y_i)}_{P(\theta_1 \ldots \theta_{50} \mid \alpha, \beta, y_1, \ldots y_{50})}\, d\theta$$

$$= (\alpha+\beta)^{-5/2} \prod^{50} \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\theta_i^{y_i+\alpha-1} (1-\theta)^{n_i-y_i+\beta-1}}_{\text{Prop. to } \mathrm{Beta}(y_i+\alpha,\ n_i-y_i+\beta)}\, d\theta_i$$

$$= (\alpha+\beta)^{-5/2} \prod^{50} \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_i)\Gamma(\beta+n_i-y_i)}{\Gamma(\alpha+\beta+n)} \right)$$

- Sample $\alpha, \beta \mid y$
- Sample $\theta_i \mid \alpha, \beta, y$

y

Modeling Three Point Shooting

$$p(\alpha, \beta | y_1, \ldots, y_{50})$$

```r
## Evaluate the log marginal posterior
marginal_posterior <- function(pars) {
  alpha <- pars[1]
  beta <- pars[2]

  ## log posterior
  -5/2*log(alpha+beta) +
    sum(lgamma(alpha + beta) +
    lgamma(alpha + y) + lgamma(beta+n-y) -
    lgamma(alpha) - lgamma(beta) - lgamma(alpha+beta+n))
}

res <- optim(c(50, 50), function(x) -1*marginal_posterior(x))
bball_map <- res$par
```

$p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$

$L(\alpha, \beta)$

The MAP estimate is ($\hat{\alpha}$ 114.4, $\hat{\beta}$ 196.6)

Empirical Bayes

- An **empirical Bayes** solution is to find a MAP estimate of α and β and condition on them as prior parameters.
- Why is this not "true" Bayes?
- When will empirical Bayes be a reasonable approximation to the full Bayes solution?

ignores uncertainty in (α, β)

Modeling Three Point Shooting

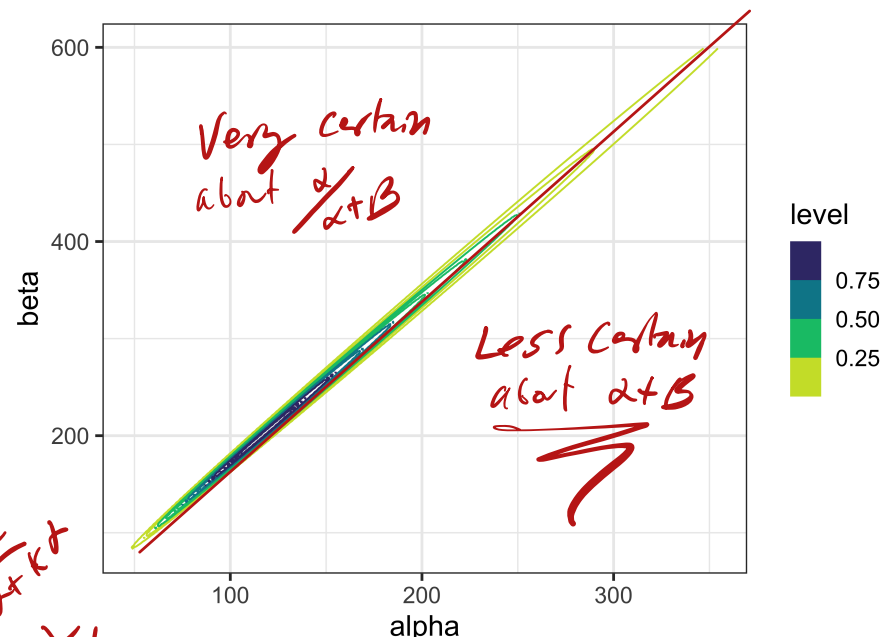
```r
alpha_beta_grid <- expand.grid(seq(25, 600, by=2), seq(25, 600, by=2))
marginal_post <- apply(alpha_beta_grid, 1, marginal_posterior)
marginal_post <- marginal_post - max(marginal_post)

tibble(alpha_beta_grid) %>% rename(alpha=Var1, beta=Var2) %>%
  ggplot() +
  geom_contour(aes(x=alpha, y=beta, z=exp(marginal_post), colour=..level..)
  colorspace::scale_color_binned_sequential("Viridis")
```

$\theta_i \sim Beta(\alpha, \beta)$

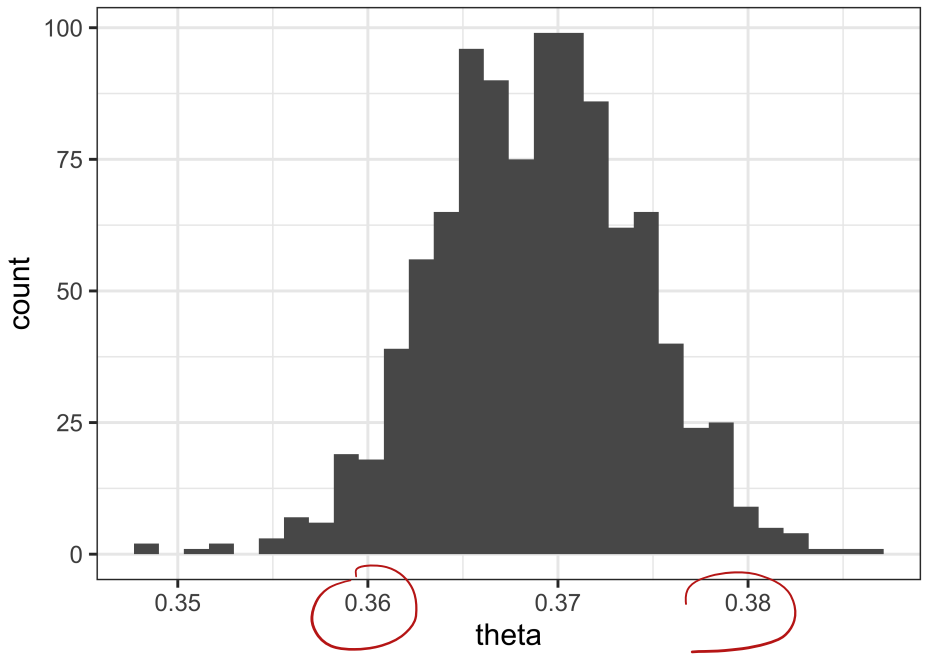
$\frac{\beta}{\alpha} = K$

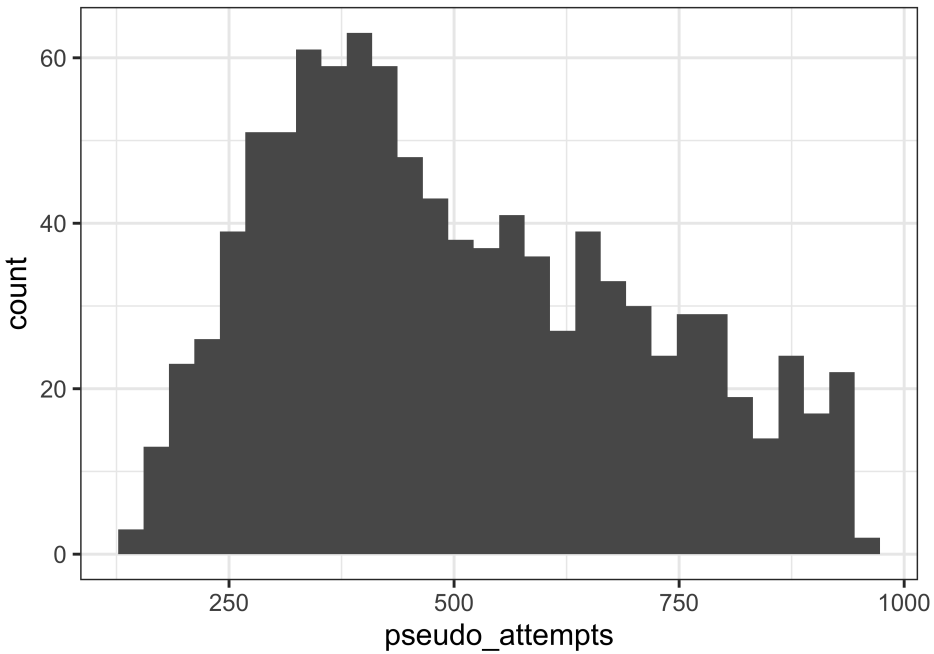
$\beta = K\alpha$

$\Rightarrow \frac{\alpha}{\alpha+\beta} = \frac{\alpha}{\alpha+K\alpha} = \frac{1}{1+K}$



Summaries of the Marginal Posterior





$P(\alpha/\alpha+\beta \mid y's)$

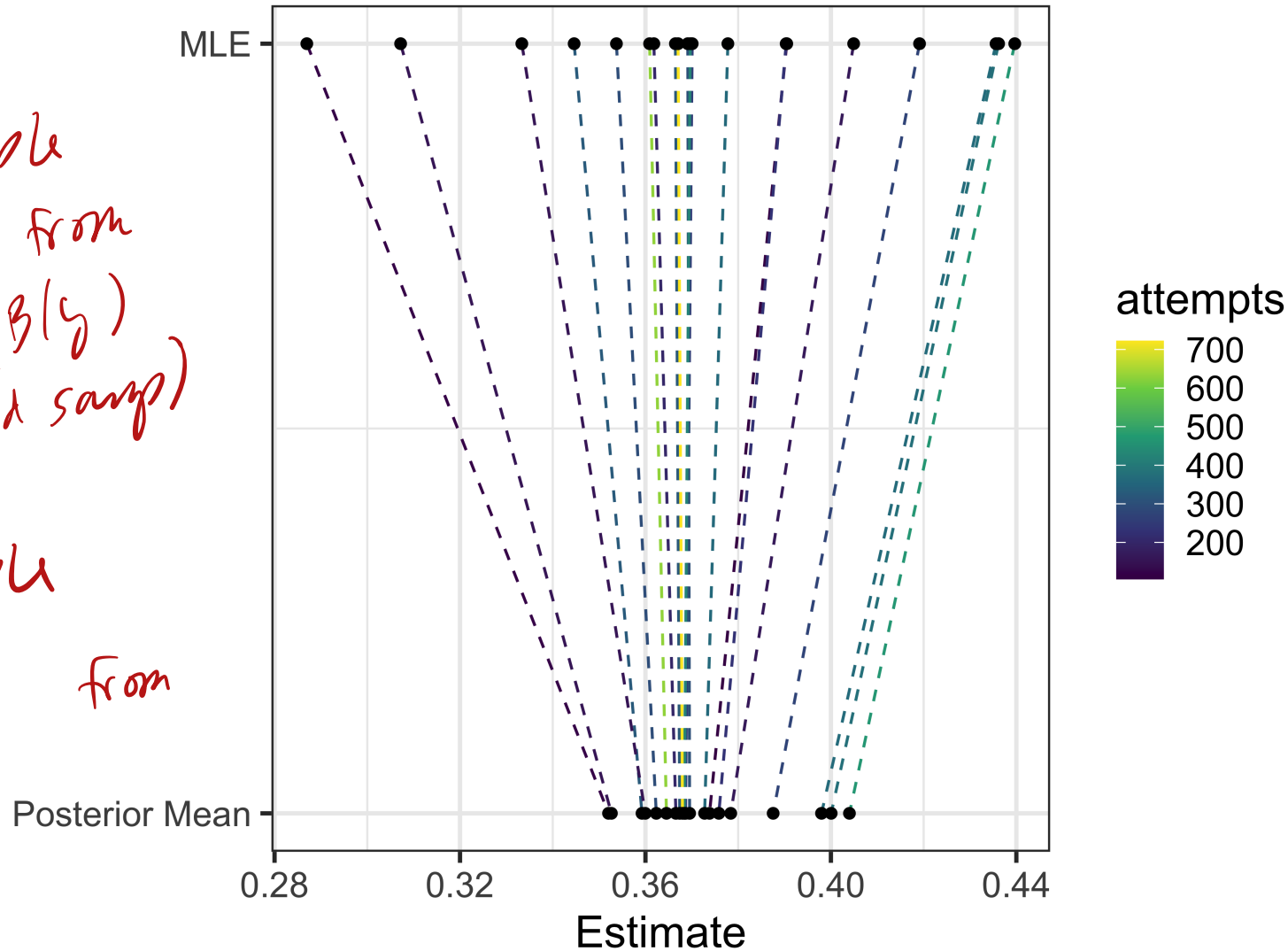
$[.36, .38]$

$P(\alpha+\beta \mid y)$

Sample theta's

1. sample α, β from $P(\alpha, \beta | y)$ (grid samp)

2. sample $\theta_i | \alpha, \beta$ from Beta



Hierarchical Modeling and Independence

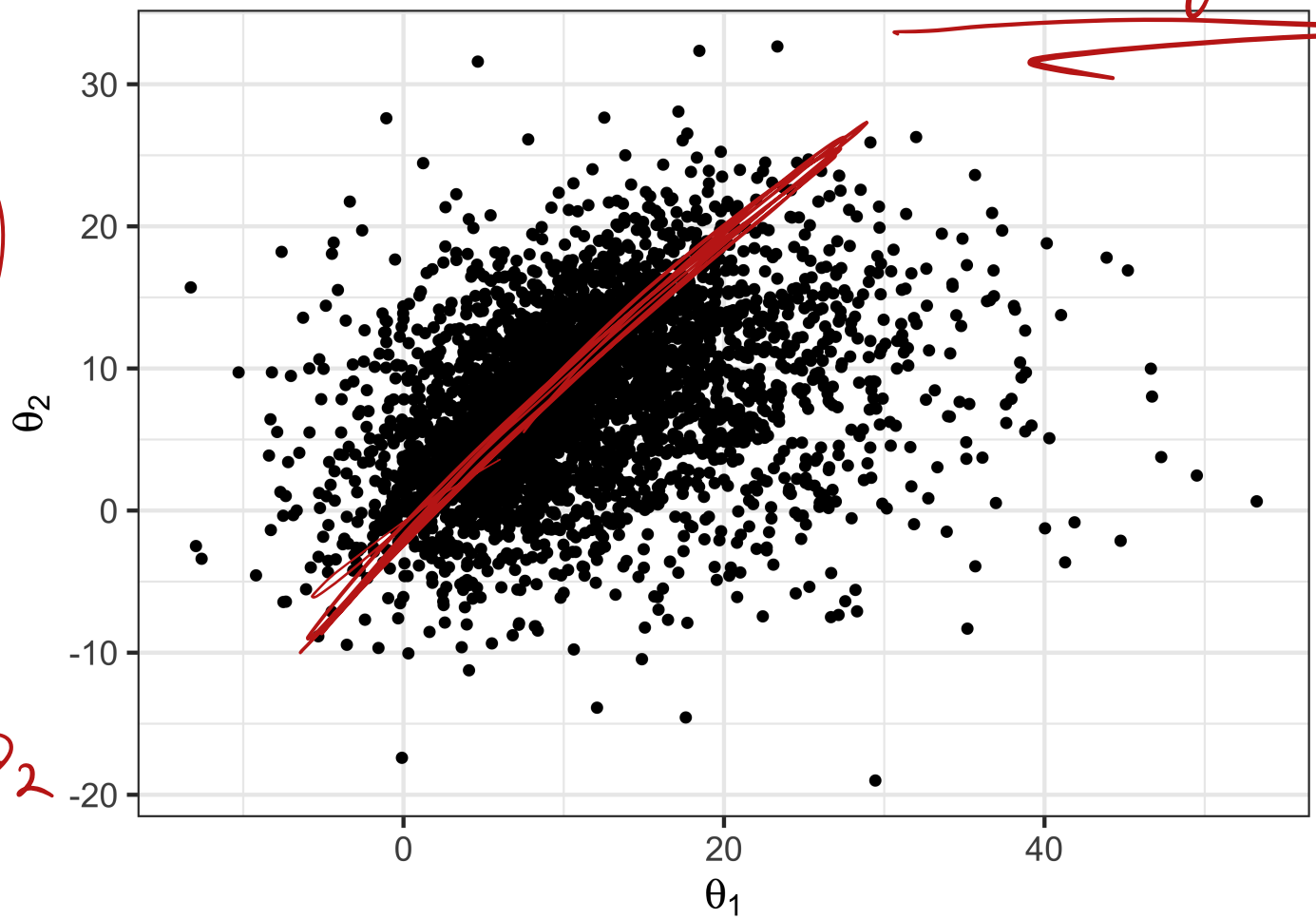
- The shooting skills θ_j are *not* independent!
- E.g. if we know θ_1 is really large, it means θ_j is also more likely to be large
- However, before we see data the distributions of θ_j's, are indistinguishable
  - Don't know which θ's will be large and which small

Hierarchical Modeling and Independence

Back to eight schools…

Prior Dependence.

$p(\theta_1, \theta_2)$



μ

$\theta_1 \quad \theta_2$

Exchangeability

An exchangeable sequence of random variables is a finite (or infinite) sequence $X_1, X_2, \ldots X_n$ of random variables such that for any finite permutation π of the indices 1, 2, 3, …, , the joint probability distribution of the permuted sequence

$$X_{\pi(1)}, X_{\pi(2)}, X_{\pi(3)}, \ldots$$

is the same as the joint probability distribution of the original sequence.

$$P(X_1, \ldots X_n) \stackrel{d}{=} P(X_{\pi(1)}, \ldots X_{\pi(n)})$$

Exchangeability

- Consider a set of J experiments in which experiment j has data y_j and parameter θ_j
- The J experiments are related, but no information in the indices that distinguish θ_j

  a priori
- Assume an exchangeable prior distribution: $p(\theta_1, \ldots \theta_J) = p(\theta_{\pi_1}, \ldots \theta_{\pi_J})$ where π is any permutation of the indices $1 \ldots J$
- Equivalent to a prior assumption about symmetry among the parameters $(\theta_1, \ldots \theta_J)$

Exchangeability

- Example: All i.i.d random variables are exchangeable
- Example: Bernoulli random variables conditional on the sum

$$X_i \sim \mathrm{Bern}(\theta)$$

$$\sum X_i = K$$

$$P(X_1, \ldots X_n \mid \text{sum}) = P(X_{\pi(1)}, \ldots X_{\pi(n)})$$

- Example: multivariate normal with common mean μ and equi-correlation

$$(X_1, \ldots X_p) \sim N_p\left(0, \begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{pmatrix}\right)$$

Non-exchangeable random variables

- Time series data (closer in time -> more correlated)
- Spatial data (closer in space -> more correlated)
- Typically, ignorance implies exchangeability
- When exchangeability doesn't hold, can often assume conditional exchangeability

Mixture of i.i.d random variables

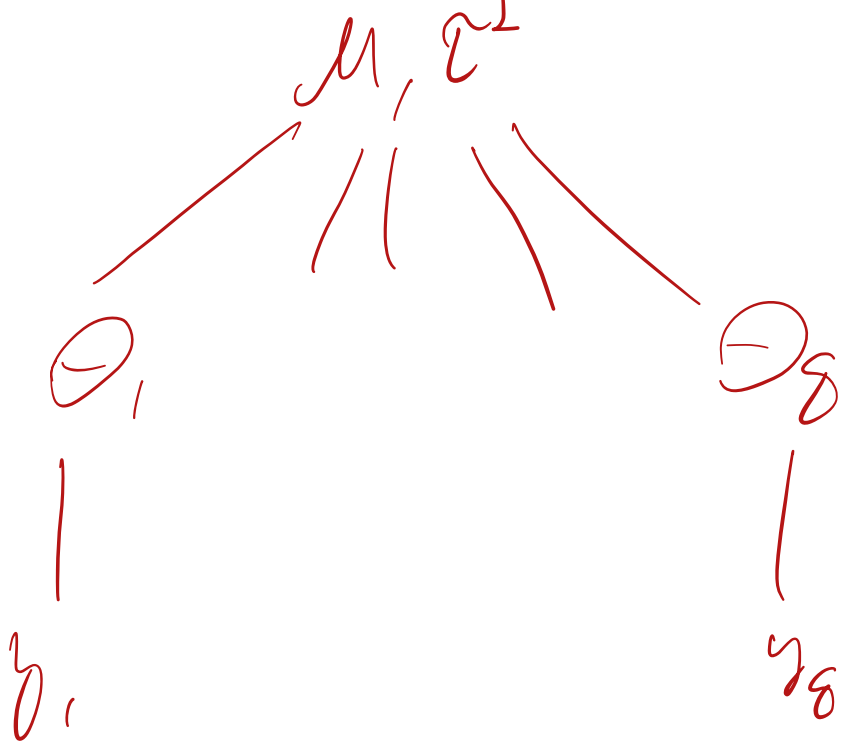
- Conditionally i.i.d random variables are exchangeable:
  $p\left(\theta_1, \ldots, \theta_J \mid \phi\right) = \prod_{j=1}^{J} p(\theta_j \mid \phi)$
- Mixtures of i.i.d random variables:

$$p\left(\theta_1, \ldots, \theta_J\right) = \int \left\{ \prod_{j=1}^{J} p\left(\theta_j | \phi\right) \right\} p(\phi) d\phi$$

  - If ϕ were known, θ_j would be i.i.d.
  - Since ϕ is not known, integrate over uncertainty
  - θ_j are a mixture of i.i.d random variables
  - Dependent but exchangeable

de Finetti's theorem

- **Theorem:** As $J \to \infty$ all exchangeable distributions can be expressed as a mixture of independent and identical distributions
- For a finite J, no guarantee that the variables can be represented as a mixture of i.i.d random variables
- But, if variables are exchangeable usually can be modeled as *approximately* a mixture of i.i.d
- Mixture's of i.i.d random variables $\to$ a hierarchical model is reasonable



- Theta's exchang.
- θ's indep. given μ, τ^2.
- $p(\theta_1, \ldots \theta_8) = \int \left(\prod p(\theta_i \mid \mu, \tau^2)\right) \times p(\mu, \tau^2)\, d\mu\, d\tau$