Bayesian Statistics

Hierarchical models

Nan Lin
Department of Mathematics
Washington University in St. Louis

Hierarchical model

- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters

- Why go hierarchical?
  - Non-hierarchical models with few parameters generally don't fit the data well.
  - Non-hierarchical models with many parameters then to fit the data well, but have poor predictive ability (overfitting)
  - Hierarchical models can often fit data with a small number of parameters but can also do well in prediction.

A simple example

- Observed datum is X.
  - $X|\theta \sim N(\theta, 1)$
- First stage prior: $\theta|\tau^2 \sim \mathrm{N}(0, \tau^2)$
- Second stage prior: $\tau^2 \sim \pi$, where π is completely specified.

- Possibly more levels
  - Second stage prior: $\tau^2|\alpha \sim gamma(\alpha, 1)$
  - Third stage prior: $\alpha \sim Exp(1)$

Connection between Hierarchical and Empirical Bayes methods

- Example: Suppose that $X_i|\theta_i \sim N(\theta_i, 1)$ are independent, and $\theta_1, \dots, \theta_n$ are i.i.d. $N(\mu, \sigma^2)$, where μ and σ^2 are known.
- Empirical Bayes: μ and σ^2 are estimated from $X_1, \dots, X_n$
  - Joint density of $X_1, \dots, X_n$ and $\theta_1, \dots, \theta_n$

$$f(X_1, \dots, X_n, \theta_1, \dots, \theta_n) = \prod_{i=1}^{n} \phi(x_i - \theta_i) \times \sigma^{-n} \prod_{i=1}^{n} \phi\left(\frac{\theta_i - \mu}{\sigma}\right)$$

  - Marginal likelihood of $X_1, \dots, X_n$

$$f(X_1, \dots, X_n) = \sigma^{-n} \prod_{i=1}^{n} \int \phi(x_i - \theta_i) \phi\left(\frac{\theta_i - \mu}{\sigma}\right) d\theta_i$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(\sigma^2 + 1)}} \phi\left(\frac{x_i - \mu}{\sqrt{\sigma^2 + 1}}\right)$$

- That is, marginally, $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2 + 1)$
- Then the MLEs of μ and σ^2 are $\bar{X}$ and $\hat{\sigma}^2 = \max(0, S^2 - 1)$
- This results in Bayes estimates of θ_i as

$$\hat{\theta}_{i,EB} = Bx_i + (1 - B)\,\bar{x}$$

where $B = \hat{\sigma}^2/(\hat{\sigma}^2 + 1)$

- Empirical Bayes uses the hierarchy idea:
  - First stage: $X_i|\theta_i \sim N(\theta_i, 1)$
  - Second stage: $\theta_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$

- Hierarchical Bayes uses a second level prior for $(\mu, \sigma^2) \sim \pi_2$
  - Joint posterior

$$f(\boldsymbol{\theta}, \mu, \sigma^2 | \boldsymbol{x}) \propto f(x|\theta, \mu, \sigma^2)\pi(\theta|\mu, \sigma^2)\pi_2(\mu, \sigma^2)$$
$$= \prod_i^n \phi(x_i - \theta_i) \times \prod_i^n \frac{1}{\sigma}\phi\left(\frac{\theta_i - \mu}{\sigma}\right) \times \pi_2(\mu, \sigma^2)$$

  - $\pi_2(\mu, \sigma^2) = \pi_2(\mu|\sigma^2)p(\sigma^2)$
    - $\pi_2(\mu|\sigma^2) \propto 1$
    - $p(\sigma^2) \propto \sigma^{-2}$ yields an improper posterior
    - $p(\sigma^2) \propto 1$ yields a proper posterior

Hierarchical linear model

- The classical linear model

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times k}\boldsymbol{\beta}_{k\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

- Hierarchical linear model

$$Y|X,\beta,\Sigma \sim N(X\beta,\Sigma)$$
$$\beta|X_\beta,\alpha,\Sigma_\beta \sim N(X_\beta\alpha,\Sigma_\beta)$$
$$\alpha|\alpha_0,\Sigma_\alpha \sim N(\alpha_0,\Sigma_\alpha)$$

Simple random effects model

- J groups
- Data in group j: $Y_{1j}, \dots, Y_{n_j j}$
- $Y_{ij}|\beta_j, \sigma^2 \sim N(\beta_j, \sigma^2)$ independent, $j = 1, \dots, J$, $i = 1, \dots, n_j$
- Random effects: $\beta_j|\alpha, \sigma_\beta^2 \sim N(\alpha, \sigma_\beta^2)$ i.i.d.
- $\alpha \sim N(\alpha_0, \sigma_\alpha^2)$

Exchangeability

- Random effects: $\beta_j|\alpha, \sigma_\beta^2 \sim N(\alpha, \sigma_\beta^2)$ i.i.d.
- $\beta_1, \dots, \beta_J$ are *exchangeable* if $f(\beta_1, \dots, \beta_J)$ is invariant to permutations of indexes $j = 1, \dots, J$
  - E.g. if $J = 3$, then the distributions of $f(\beta_1, \beta_2, \beta_3)$, $f(\beta_1, \beta_3, \beta_2)$, $f(\beta_2, \beta_1, \beta_3)$, $f(\beta_2, \beta_3, \beta_1)$, $f(\beta_3, \beta_1, \beta_2)$, $f(\beta_3, \beta_2, \beta_1)$ are all of the same form.
- Exchangeability → identically distributed (i.e. same marginal distribution)
- Exchangeability $\not\rightarrow$ independent, e.g. $N(\mu\mathbf{1}, \Sigma)$, where off-diagonal elements in Σ are all ρ.
- Ignorance→Exchangeability
  - If no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution

De Finetti's theorem

- In general, we may treat β_j as independent draws from a population distribution governed by some unknown parameter ϕ

$$f(\theta|\phi) = \prod_{j=1}^{J} f(\theta_j|\phi)$$

$$f(\theta) = \int \left[\prod_{j=1}^{J} f\left(\theta_j|\phi\right)\right] \pi(\phi) d\phi$$

- $f(\theta)$ is a mixture of iid distributions, and exchangeable
- De Finetti's theorem: As $J \to \infty$, any "suitable well-behaved" exchangeable distribution on $\beta_1, \dots, \beta_J$ can be written in the iid mixture form.

Intraclass correlation (ICC)

$$\text{Cov}(Y_{ij}, Y_{rs}) = E(Y_{ij}Y_{rs}) - E(Y_{ij})E(Y_{rs}).$$

$$\begin{aligned} E(Y_{ij}) &= E[E(Y_{ij}|\beta_j)] \\ &= E(\beta_j) \\ &= E[E(\beta_j|\alpha)] \\ &= E(\alpha) = \alpha_0, \end{aligned} \qquad \begin{aligned} E(Y_{ij}Y_{rs}) &= E[E(Y_{ij}Y_{rs}|\boldsymbol{\beta})] \\ &= E(\beta_j\beta_s) \\ &= E[E(\beta_j\beta_s|\alpha)]. \end{aligned}$$

if $j \neq s$, then

$$E[E(\beta_j\beta_s|\alpha)] = E(\alpha^2) = \alpha_0^2 + \sigma_\alpha^2,$$

and if $j = s$

$$E[E(\beta_j\beta_s|\alpha)] = E(\alpha^2 + \sigma_\beta^2) = \alpha_0^2 + \sigma_\alpha^2 + \sigma_\beta^2.$$

- Then we have $Cov(Y_{ij}, Y_{rs}) = \begin{cases} \sigma_\alpha^2, & j \neq s \\ \sigma_\alpha^2 + \sigma_\beta^2, & j = s \end{cases}.$

- Next, we need the unconditional variance

$$\begin{aligned} \mathrm{Var}(Y_{ij}) &= \mathrm{Var}\left(E(Y_{ij}|\beta_j)\right) + E\left(\mathrm{Var}(Y_{ij}|\beta_j)\right) \\ &= \mathrm{Var}(\beta_j) + \sigma^2 \\ &= \mathrm{Var}(\alpha) + \sigma_\beta^2 + \sigma^2 \\ &= \sigma_\alpha^2 + \sigma_\beta^2 + \sigma^2. \end{aligned}$$

- Therefore, the intraclass correlation is

$$\mathrm{Corr}(Y_{ij}, Y_{rs}) = \begin{cases} \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\beta^2+\sigma^2}, & j \neq s \\ \frac{\sigma_\alpha^2+\sigma_\beta^2}{\sigma_\alpha^2+\sigma_\beta^2+\sigma^2}, & j = s. \end{cases}$$

- Frequentist random effect model
  - $\sigma_\alpha = 0$ and $\sigma_\beta > 0$

Extensions

- Random effects with multiple populations
  - Suppose the components of β fall into K groups. The β_j's in group k are a random sample from $N(\alpha_k, \sigma_{\beta k}^2)$.
- "Many regression" model
  - $Y_i | X_i, \beta_i, \sigma_i^2 \sim N\left(X_i \beta_i, \sigma_i^2 I\right), i = 1, \dots, k$
  - $\beta_i \sim N(\alpha, D)$ i.i.d.
  - $\sigma_i \sim g(\cdot; \gamma)$ i.i.d.
  - $\alpha \sim N(\alpha_0, \Sigma_0)$
  - $\gamma \sim \pi_1$

Hierarchical models for meta-analysis

- If there are several studies that address the same research question, one might be interested in combining the information from the individual studies in order to draw an 'overall' conclusion
- The studies can be thought of a belonging to a population of studies addressing the same question, and the combining individual studies in order to learn about the whole is referred as '*meta-analysis*'.

Rat tumor example

- 71 different groups rats. (one current + 70 historical)
- Interest in the rate of endometrial stromal polyps in different groups. The number of rats varies from group to group.
- Data model: $y_i =$ number of tumors in group i

$$y_i|\theta_i \overset{ind}{\sim} Bin(n_i, \theta_i) \quad i = 1, \ldots, 71$$

- θ_i =tumor rate in group i

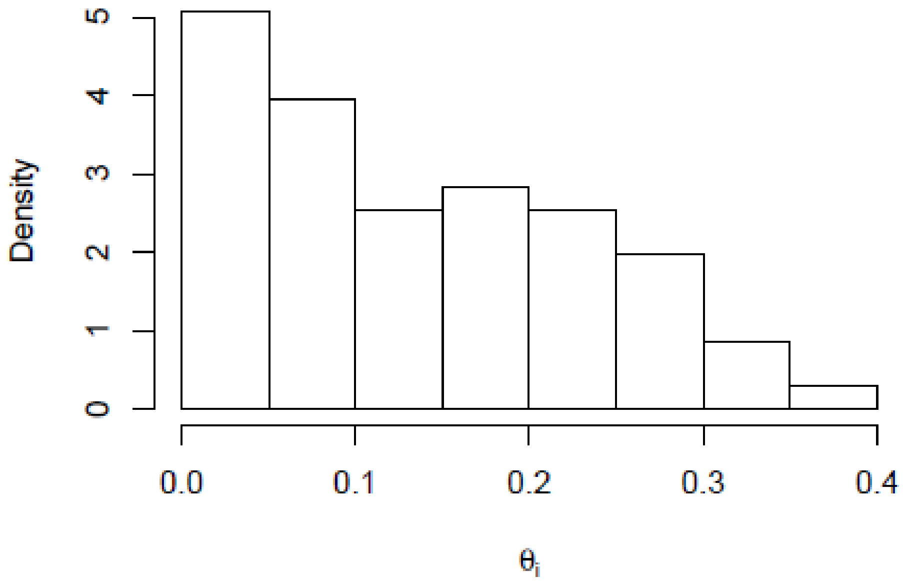
$$\theta_i \overset{ind}{\sim} Beta(\alpha, \beta)$$

- Prior of population parameter

$$\alpha, \beta \sim p(\alpha, \beta).$$

- Data
  - Current experiment: 4/14
  - Historical data

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24



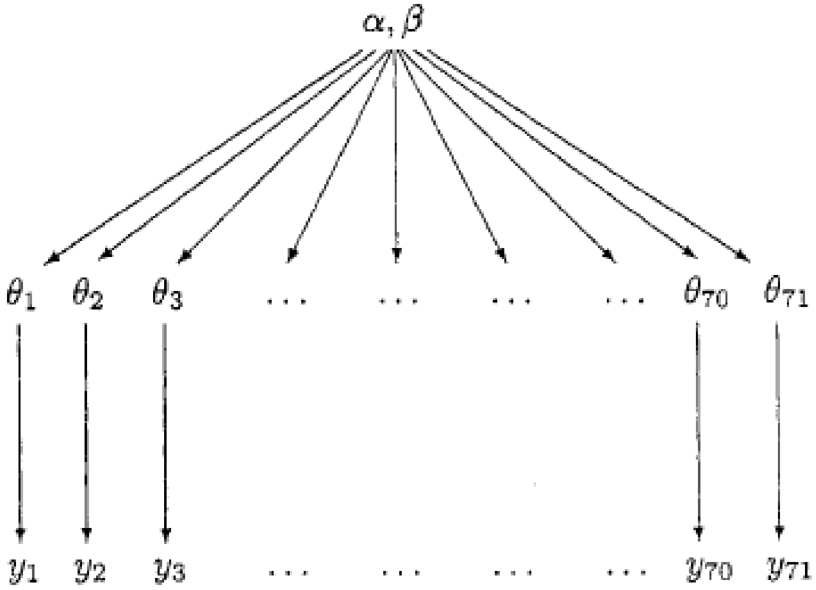
Exchangeability in rat tumor example

- Although the individual θ_j's differ, it might be perfectly acceptable to consider them as if drawn from a common distribution
- In the rat example, we have no preferences for different orderings of the theta's.

$$\alpha, \beta \sim p(\alpha, \beta).$$

$$\theta_i \overset{ind}{\sim} Beta(\alpha, \beta)$$

$$y_i|\theta_i \overset{ind}{\sim} Bin(n_i, \theta_i) \quad i = 1, \ldots, 71$$



α, β
θ_1 θ_2 θ_3 ... ... θ_{70} θ_{71}
y_1 y_2 y_3 ... ... y_{70} y_{71}

Summary

- $y|\theta \sim f(y|\theta)$
- $\theta|\phi \sim f(\theta|\phi)$
- $\phi \sim f(\phi)$: hyper-prior

- Posterior distribution
  $$f(\theta,\phi|y) \propto f(\theta,\phi)f(y|\theta,\phi) \propto f(\phi)f(\theta|\phi)f(y|\theta)$$

- Predictive distribution
  1. $\tilde{y}$ for an existing θ_j
  2. $\tilde{y}$ for an new θ_j

Computation with hierarchical models

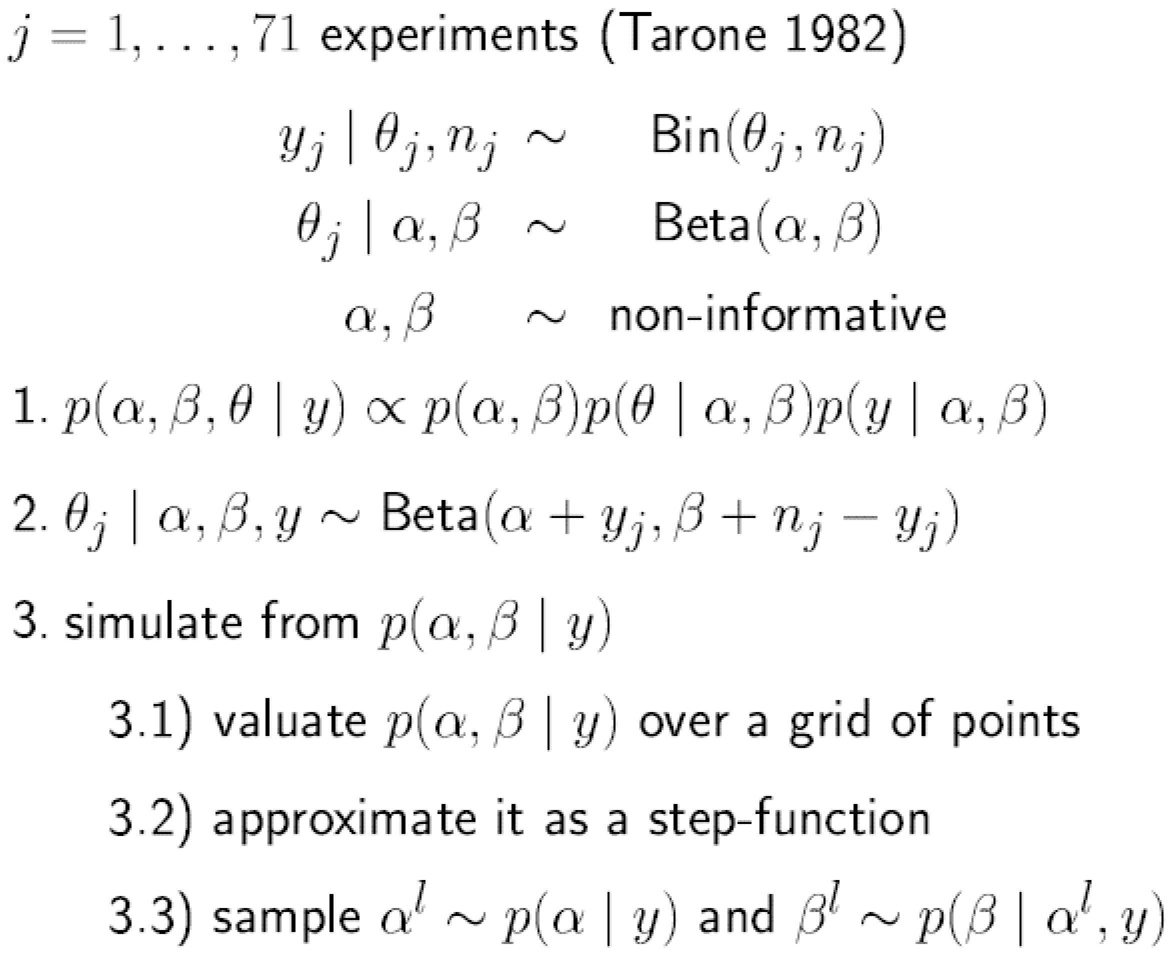
- θ: parameter of interest
- ϕ: nuisance parameter

- If $f(\theta|\phi)$ is taken as a conjugate prior, the conditional posterior $f(\theta|y,\phi)$ is easy to obtain
- To obtain the marginal posterior of $f(\phi|y)$, one need to solve the integral $f(\phi|y) = \int f(\theta,\phi|y)d\theta$
  - But this can be difficult in general

Computation with hierarchical models

1. draw $\phi^* \sim p(\phi \mid y)$
2. draw $\theta^* \sim p(\theta \mid \phi^*, y)$
3. if the factorization $p(\theta \mid \phi, y) = \prod p(\theta_j \mid \phi, y)$ holds, then the components θ_j can be drawn independently, one at a time
4. draw $\tilde{y} \sim p(y \mid \theta^*)$
5. repeat the steps L times in order to obtain a set of L draws

Bayesian analysis of the rat tumor example



$j = 1, \ldots, 71$ experiments (Tarone 1982)

$$\begin{aligned} y_j \mid \theta_j, n_j &\sim \quad \text{Bin}(\theta_j, n_j) \\ \theta_j \mid \alpha, \beta &\sim \quad \text{Beta}(\alpha, \beta) \\ \alpha, \beta &\sim \quad \text{non-informative} \end{aligned}$$

1. $p(\alpha, \beta, \theta \mid y) \propto p(\alpha, \beta) p(\theta \mid \alpha, \beta) p(y \mid \alpha, \beta)$

2. $\theta_j \mid \alpha, \beta, y \sim \text{Beta}(\alpha + y_j, \beta + n_j - y_j)$

3. simulate from $p(\alpha, \beta \mid y)$

   3.1) valuate $p(\alpha, \beta \mid y)$ over a grid of points

   3.2) approximate it as a step-function

   3.3) sample $\alpha^l \sim p(\alpha \mid y)$ and $\beta^l \sim p(\beta \mid \alpha^l, y)$

Analytics

$$p(\theta, \alpha, \beta|y) \propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta)$$

$$\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}$$

$$p(\theta|\alpha, \beta, y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$$

$$p(\alpha, \beta|y) = \frac{p(\theta, \alpha, \beta|y)}{p(\theta|\alpha, \beta, y)} \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$$

Set up a noninformative prior

- let $\gamma_1 = \frac{\alpha}{\alpha+\beta}$
- let $\delta_1 = (\alpha + \beta)^{-1/2}$
- $p(\gamma, \delta) \propto$ constant

this prior leads a proper posterior, and yields:

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

Computing the marginal posterior density

- Contour plot of the unnormalized marginal posterior $p(\gamma, \delta \mid y)$
- draw 1000 random samples from the joint posterior $p(\alpha, \beta, \theta_1, \ldots, \theta_{71} \mid y)$ as follows:

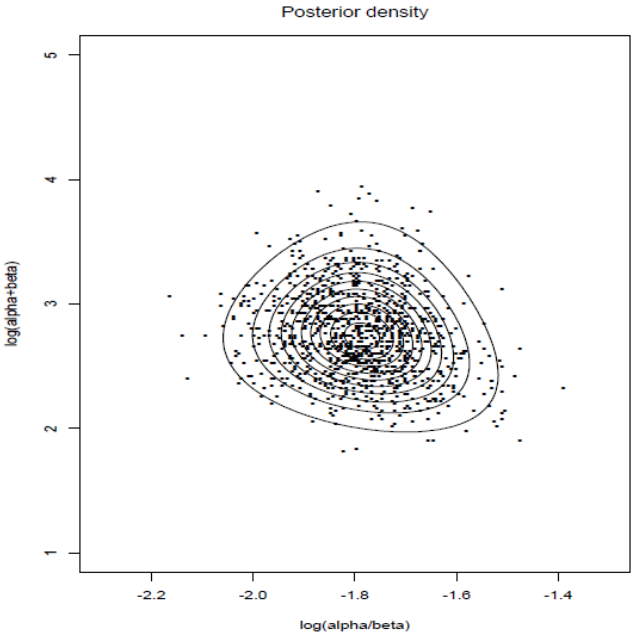
1) simulate γ^l, δ^l from $p(\gamma, \delta \mid y)$ with the discrete-grid sampling procedure

   [1.1] for $l = 1, \ldots 1000$

   [1.2] transform $\gamma^l, \delta^l \rightarrow \alpha^l, \beta^l$

2) for each l draw $\theta^l \sim$ Beta$(\alpha^l + y_j, \beta^l + n_j - y_j)$

3) displays the results, for example histogram of $ED50 = \alpha/\beta$



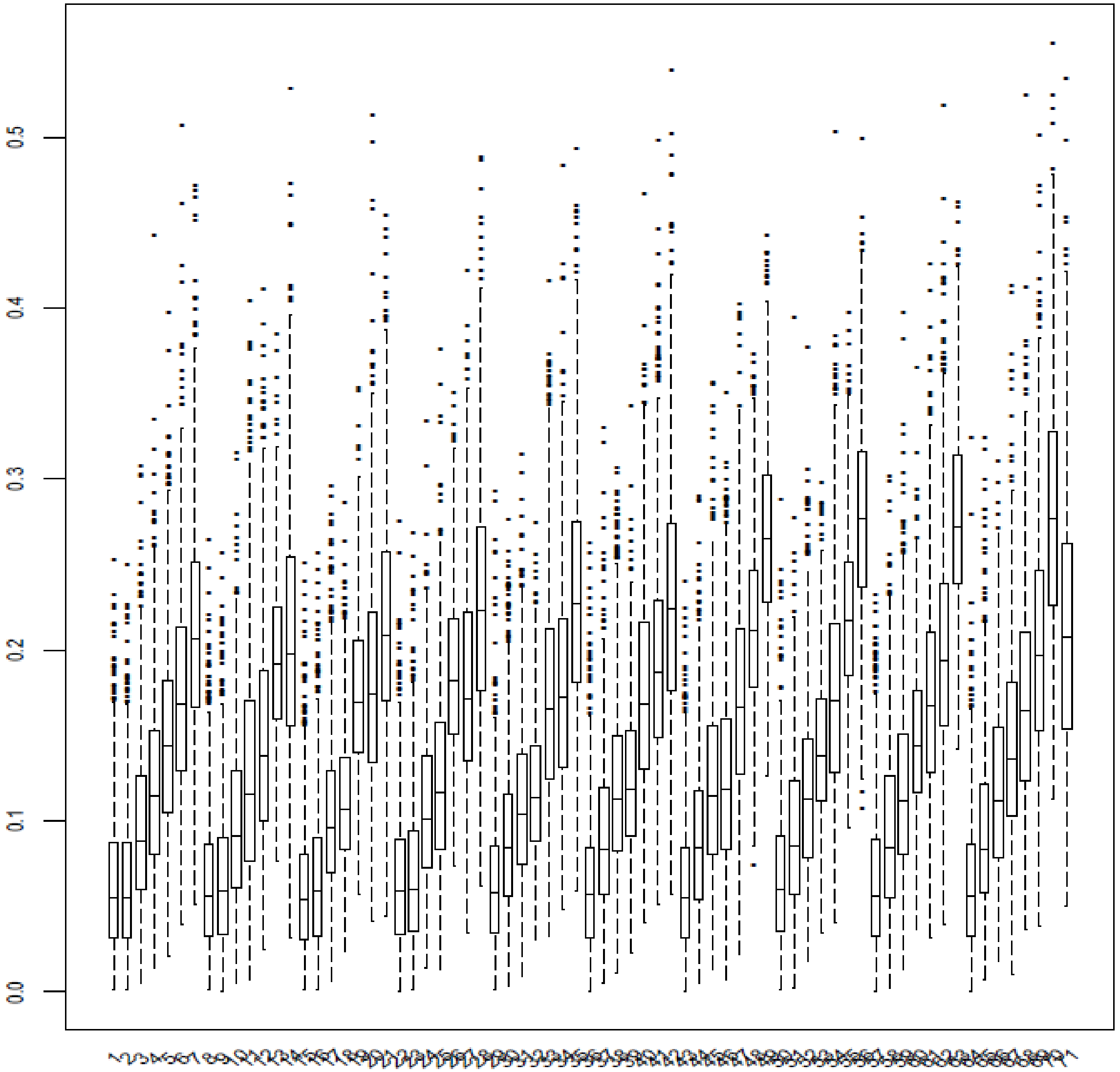


Figure 1: Marginal posterior distributions of $\theta_1, \ldots, \theta_7 1$

SAT coaching example

- Goal: The Educational Testing Service (ETS) wants to analyze the effects of special coaching programs on SAT-V scores
- Separate randomized controlled experiments were performed at eight high schools
- For each school, the estimated coaching effect and its standard error were obtained.

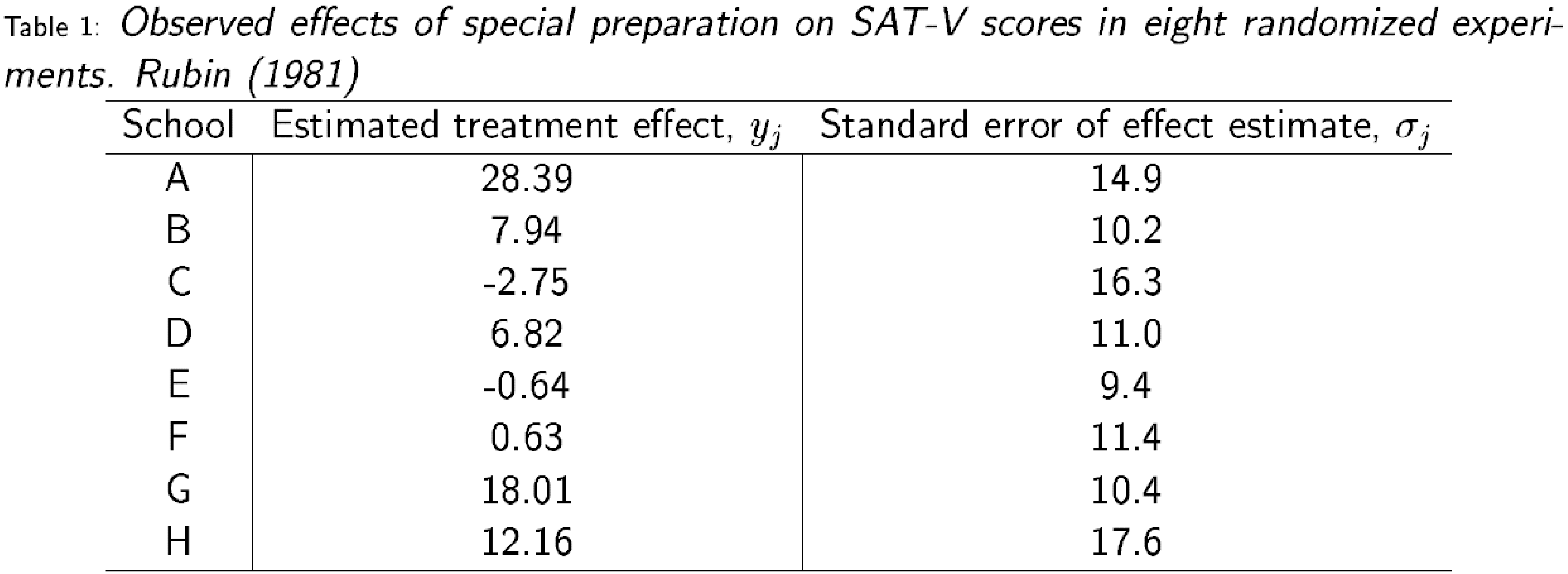
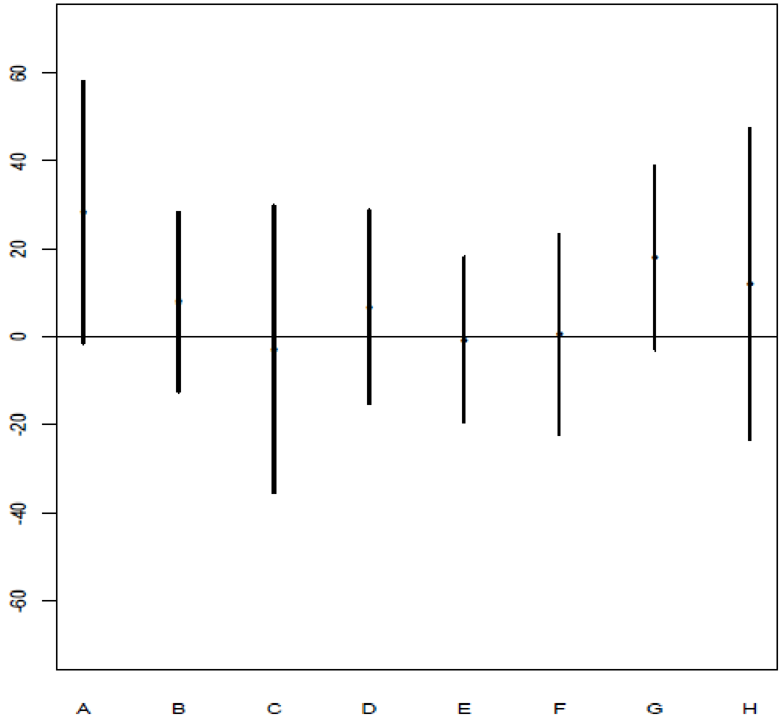


Table 1: *Observed effects of special preparation on SAT-V scores in eight randomized experiments. Rubin (1981)*

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28.39	14.9
B	7.94	10.2
C	-2.75	16.3
D	6.82	11.0
E	-0.64	9.4
F	0.63	11.4
G	18.01	10.4
H	12.16	17.6

Separate Estimates

- Treating experiments at each school separately, and applying the simple normal analysis, yields 95% posterior credible intervals that all overlap
- Difficulty
  - For example, from $\theta_A|y \sim N(28.4, 14.9^2)$ we will conclude that $P(\theta_A > 28.4|y_A) = \frac{1}{2}$. However, from data on the other seven schools, this looks a doubtful statement.



Pooled estimate

- If we assume that $\theta_1 = \cdots = \theta_8$, then between School A and School C, we have $P(\theta_A - \theta_C < 0|y) = \frac{1}{2}$, which is difficult to justify from the table.

Hierarchical Model

- The quantities of interest are the θ_j: Average "true" effects of coaching programs.

- Data y_j: separate estimated treatment effects for each school.

- The standard errors σ_j are assumed known (large samples).

- This is a randomized experiment with large samples, no outliers, so we appeal to the central limit theorem:

$$y_j|\theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$$

The non-hierarchical methods

1. Each school is analyzed separately

2. All schools are pooled

- The hierarchical model provides a compromise

Random-effect ANOVA with known variance

$p(\mu,\tau) \propto 1$

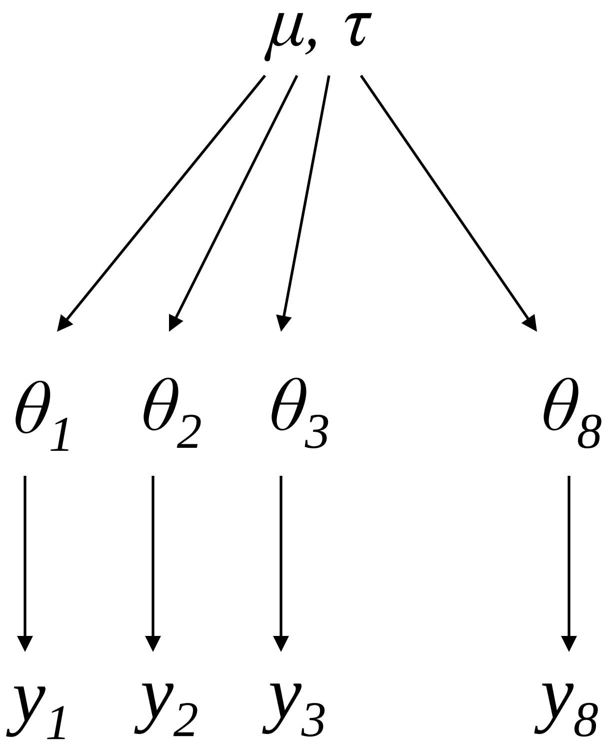
μ: overall treatment effect

τ: heterogeneity among schools

$\theta_j \sim N(\mu, \tau)$

θ_j effect at school *j*

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



μ, τ

θ_1 θ_2 θ_3 θ_8

y_1 y_2 y_3 y_8

Computation

The joint posterior distribution:

$$
\begin{aligned}
&p(\theta,\mu,\tau|y)\\
&\quad\propto p(\mu,\tau)p(\theta|\mu,\tau)p(y|\theta)\\
&\quad\propto \prod_{j=1}^{J} N\left(\theta_j|\mu,\tau^2\right)\prod_{j=1}^{J} N\left(y_j|\theta_j,\sigma_j^2\right)\\
&\quad\propto \tau^{-J}\exp\left[-\frac{1}{2}\sum_j\frac{1}{\tau^2}(\theta_j-\mu)^2\right]\exp\left[-\frac{1}{2}\sum_j\frac{1}{\sigma_j^2}(y_j-\theta_j)^2\right]
\end{aligned}
$$

Factors depending only on y and $\{\sigma_j\}$ treated as constant.

Conditional posterior dist'n of θ given μ, τ, y

- Treat (μ, τ) as fixed in the previous expressions.
- Given (μ, τ), the J separate parameters θ_j are independent in their posterior distribution since they appear in different factors in the likelihood (which factors into J components).
- $\theta_j | y, \mu, \tau \sim N(\hat{\theta}_j, V_j)$ with

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\, y_j + \frac{1}{\tau^2}\, \mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

Marginal posterior dist'n of μ, τ given y

To derive $p(\mu, \tau|y)$, think of inference about (μ, τ) directly.

Prior distribution: $p(\mu, \tau) \propto 1$.

Data distribution: $p(y|\mu, \tau) = \prod_{j=1}^{J} N(y_j|\mu, \sigma_j^2 + \tau^2)$

$$
\begin{aligned}
p(\mu, \tau|y) &\propto \prod_{j=1}^{J} N(y_j|\mu, \sigma_j^2 + \tau^2) \\
&\propto \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(y_j - \mu)^2}{2(\sigma_j^2 + \tau^2)} \right)
\end{aligned}
$$

Posterior distribution of μ given τ, y

Instead of sampling (μ, τ) on a grid, factor the distribution: $p(\mu, \tau|y) = p(\tau|y)p(\mu|\tau, y)$.

$p(\mu|\tau, y)$ is obtained by looking at $p(\mu, \tau|y)$ and thinking of τ as known. With a uniform prior for $\mu|\tau$, the log posterior is quadratic in μ and therefore normal:

$$p(\mu|\tau, y) \propto \prod_{j=1}^{J} N\left(y_j|\mu, \sigma_j^2 + \tau^2\right)$$

This is a normal sampling distribution with a noninformative prior density on μ.

The mean and variance are obtained by considering group means y_j as J independent estimates of μ with variance $\sigma_j^2 + \tau^2$.

Result: $\mu|\tau, y \sim N(\hat{\mu}, V_\mu)$ with

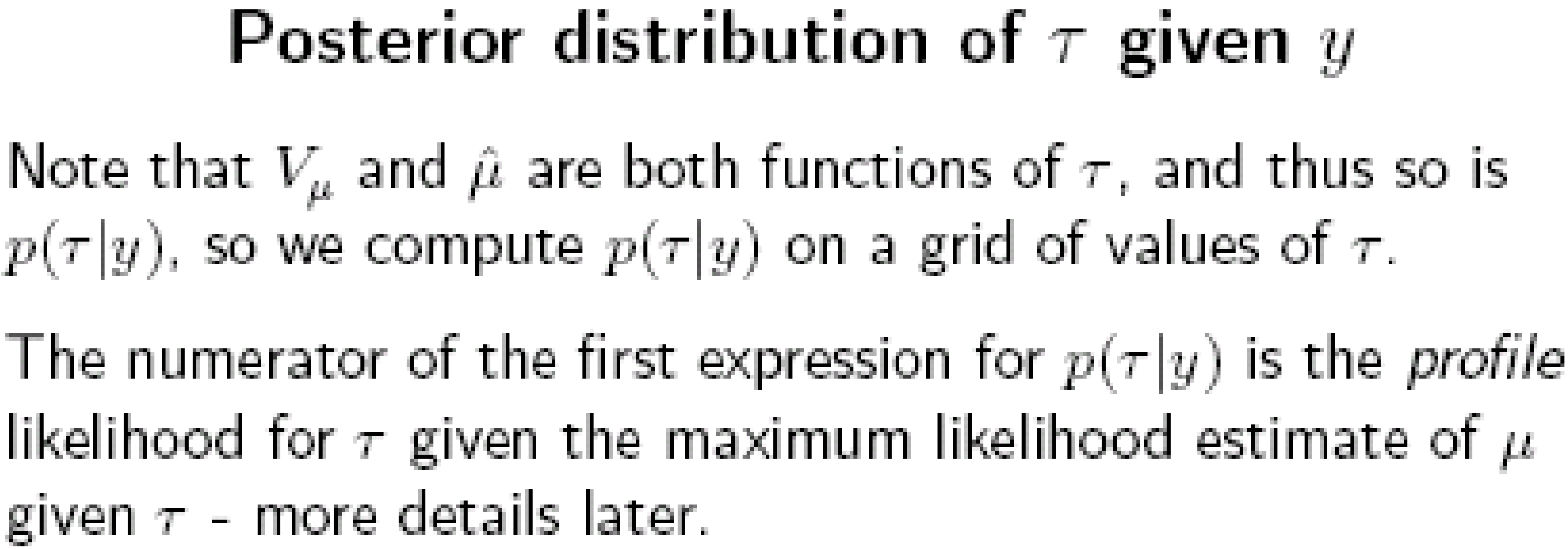
$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2} y_j}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}} \quad \text{and} \quad V_\mu = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}$$

Posterior distribution of τ given y

We could integrate $p(\mu, \tau|y)$ over μ?

It is easier to use identity $p(\tau|y) = p(\mu, \tau|y)/p(\mu|\tau, y)$ (which holds for all μ), and evaluate at $\mu = \hat{\mu}$:

$$\begin{aligned} p(\tau|y) &\propto \frac{\prod_{j=1}^{J} N(y_j|\hat{\mu}, \sigma_j^2 + \tau^2)}{N(\hat{\mu}|\hat{\mu}, V_\mu)} \\ &\propto V_\mu{}^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(y_j - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)} \right) \end{aligned}$$



Posterior distribution of τ given y

Note that V_μ and $\hat{\mu}$ are both functions of τ, and thus so is $p(\tau|y)$, so we compute $p(\tau|y)$ on a grid of values of τ.

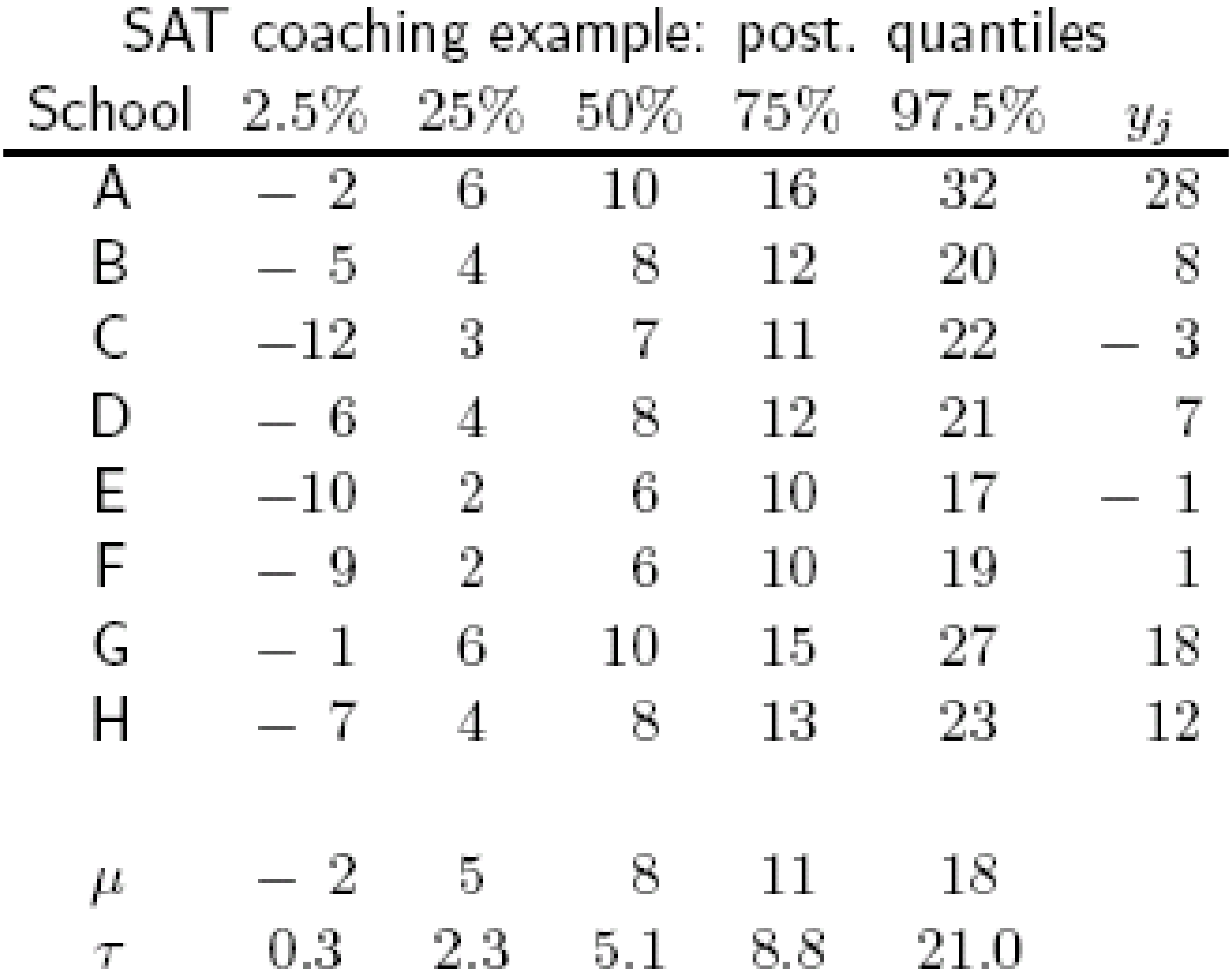
The numerator of the first expression for $p(\tau|y)$ is the *profile* likelihood for τ given the maximum likelihood estimate of μ given τ - more details later.

Normal-normal model computation: Summary

To simulate from joint posterior distribution $p(\theta, \mu, \tau | y)$:

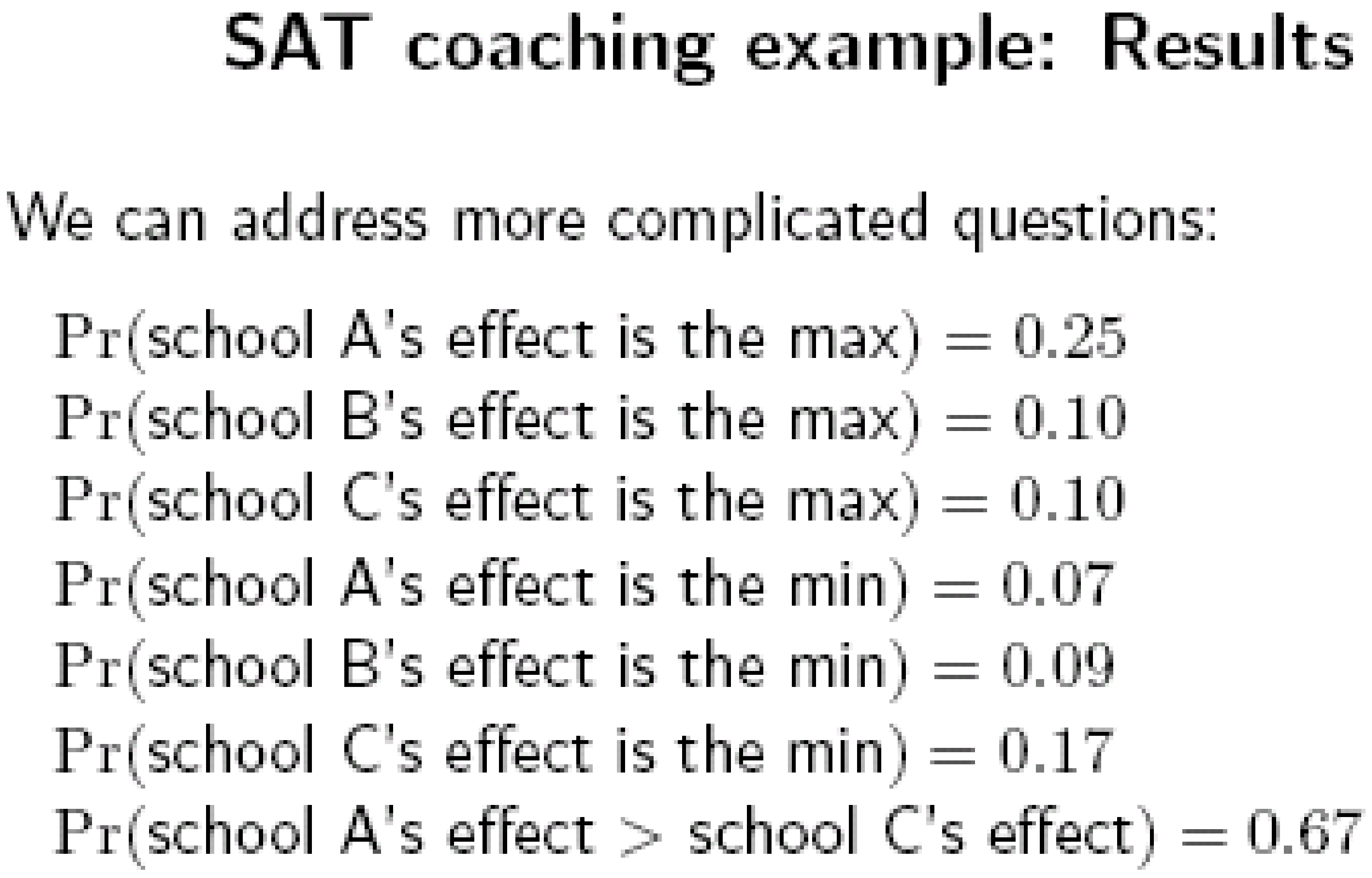
1. Draw τ from $p(\tau | y)$ (grid approximation)
2. Draw μ from $p(\mu | \tau, y)$ (normal distribution)
3. Draw $\theta = (\theta_1, \dots, \theta_J)$ from $p(\theta | \mu, \tau, y)$ (independent normal distribution for each θ_j)

Apply these ideas to SAT coaching data; repeat 1000 times to obtain 1000 simulations.



SAT coaching example: post. quantiles

School	2.5%	25%	50%	75%	97.5%	y_j
A	-2	6	10	16	32	28
B	-5	4	8	12	20	8
C	-12	3	7	11	22	-3
D	-6	4	8	12	21	7
E	-10	2	6	10	17	-1
F	-9	2	6	10	19	1
G	-1	6	10	15	27	18
H	-7	4	8	13	23	12
μ	-2	5	8	11	18	
τ	0.3	2.3	5.1	8.8	21.0	

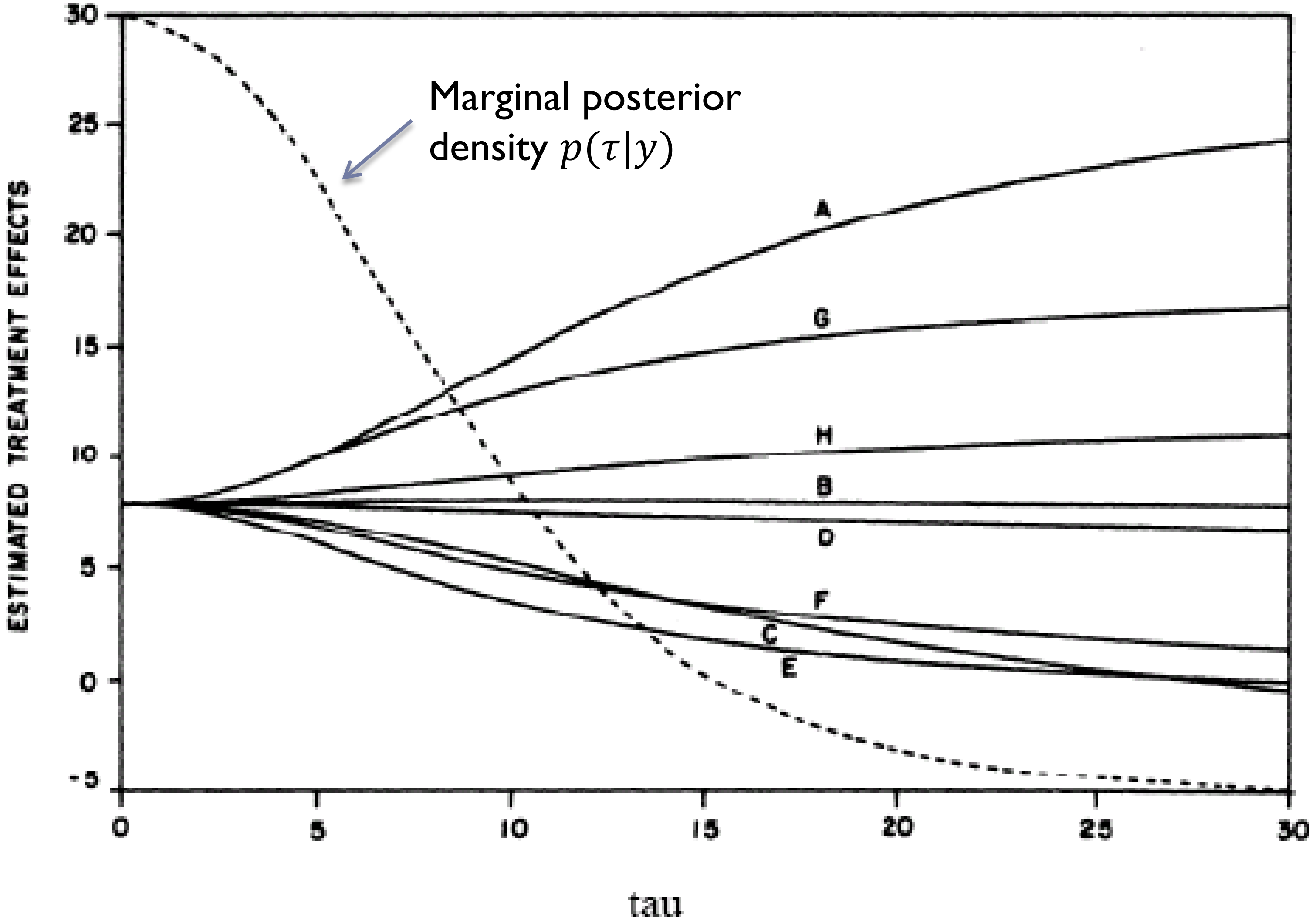


SAT coaching example: Results

We can address more complicated questions:

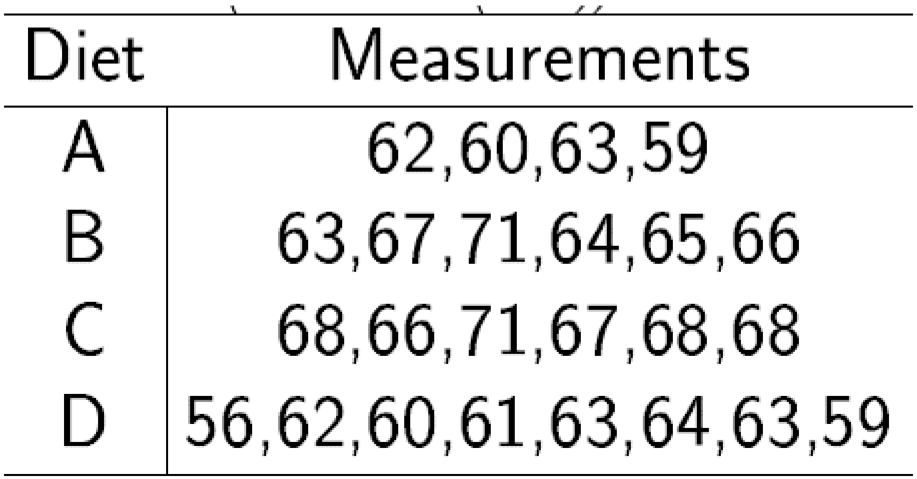
$$\begin{aligned}
\Pr(\text{school A's effect is the max}) &= 0.25\\
\Pr(\text{school B's effect is the max}) &= 0.10\\
\Pr(\text{school C's effect is the max}) &= 0.10\\
\Pr(\text{school A's effect is the min}) &= 0.07\\
\Pr(\text{school B's effect is the min}) &= 0.09\\
\Pr(\text{school C's effect is the min}) &= 0.17\\
\Pr(\text{school A's effect} > \text{school C's effect}) &= 0.67
\end{aligned}$$

Estimated School Effects $E(\theta_j|\tau, y)$



The DIET example

- Coagulation time (in seconds) for blood drawn from 24 animals randomly allocated to four different diets. (Box and Hunter, 1978)
- Data



Diet	Measurements
A	62,60,63,59
B	63,67,71,64,65,66
C	68,66,71,67,68,68
D	56,62,60,61,63,64,63,59

Random-effect ANOVA with unknown variance

- $y_{ij} \sim N(\theta_j, \sigma^2), i = 1, \ldots, n_j, \sum_j n_j = n$
- $\theta_j \sim N(\mu, \tau^2), j = 1, \ldots, J$
- $\mu \sim N(0,100)$
- $\sigma^2 \sim Inv - gamma(a_1, b_1)$
- $\tau^2 \sim Inv - gamma(a_2, b_2)$
- What is $\boldsymbol{\theta}|y$?
  - Integration over σ^2 is not easy
- We can derive
  - $\boldsymbol{\theta}|\sigma^2, \mu, \tau^2, y$
  - $\sigma^2|\theta, \mu, \tau^2, y$
  - $\mu|\theta, \sigma^2, \tau^2, y$
  - $\tau^2|\theta, \sigma^2, \mu, y$