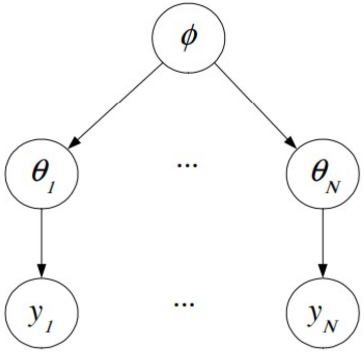
# Normal Hierarchical Models




Statistics Foundations

~~Antonio Máximo Ponce de León Pérez~~
Jorge III Altamirano Astorga
Ángel Rafael Ortega Ramirez

# Motivation – Introduction

- Many stat applications involve hierarchical data, so hierarchical models are more appropriate, as it’s possible to structuresome dependence.
- Having insufficient parameters, they tend to overfit.
- Can be used for “meta-analysis”: used for research in order to understand a relationship between different related experiments.

# Hierarchical Models

- Some authors mention coin the term *Empirical Bayes* to the analysis using the data to estimatate prior parameters.
- Exchangeability: if no information is given to distinguish any of the θj. Then, no order or grouping of the parameters can be made. Ignorance of this info implies exchangeability.

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

$$p(\theta) = \int [\prod_{j=1}^{J} p(\theta_j|\phi)] p(\theta_j|phi)$$

# Hierarchical Models

$$\mathrm{p}(\phi, \theta) = p(\phi)p(\theta|\phi)$$

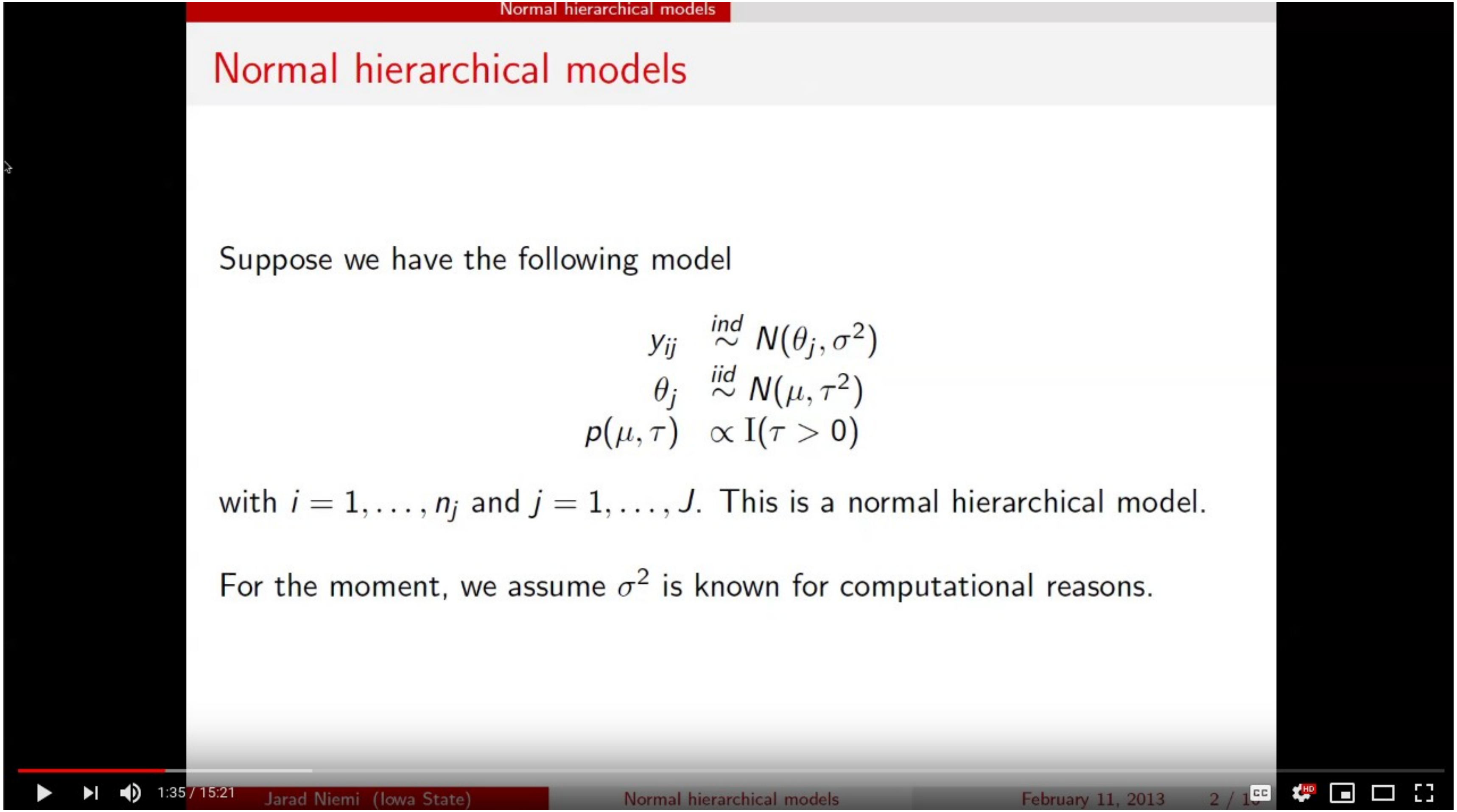
$$Joint\ Posterior\ Distribution:$$

$$p(\phi, \theta|y) \propto p(\phi, \theta)p(y|\phi, \theta) = p(\phi, \theta)p(y|\theta)$$

$$Hyperprior\ Distribution\ for\ \phi:$$

- May use a diffuse distribution, if little is known.
- Should result in a posterior dist. that is proper.
- Should at least constrain the hyper params into a finite region.

# Normal Hierarchical Models



# Hierarchical Normal Distributions

$$\mathrm{p}(\phi|y) = \frac{p(\phi,\theta|y)}{p(\theta|\phi,y)}$$

- The marginal posterior distribution of phi can be computed algebraically using the conditional probability formula.
- The denominator has a normalizing factor that depends on phi and theta.
- Care must be taken to make sure that the proportionality constant is actually a constant.
- Many times, a conjugate hierarchical scheme assumes normal sampling and normally distributed latent effects.

# Estimating an Exchangeable Set of Params for a Normal Model

- We will show a simple normal hierarchical model: one way normal random effects model.
- Different mean for each group or experiment.
- Known variance, this assumption can be an adequate approximation at the sampling level.
- $\text{Data: } \mathrm{y}_i|\theta_j \sim N(\theta_j, \sigma^2) \;\; for \;\; i = 1, ..., n_j; j = 1, ...J$

$$Likelihood: \; \overline{y}_{.j}\theta_j = N(\theta_j, \sigma_j^2)$$

$$Analysis \; of \; variance: \; \overline{y_{.j}} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

# Estimating an Exchangeable Set of Params for a Normal Model

- Pooled estimate: $\overline{y_{..}} = \frac{\sum_{j=1}^{J} \frac{\overline{y_{.j}}}{\sigma_j^2}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}}$
- Traditionally, the analysis was to test differences among means. Choosing between the lesser.
- Alternatively we can use both: weighted combination: $\hat{\theta}_j = \lambda_j \overline{y_{.j}} + (1 - \lambda_j)\overline{y_{..}}$

# Normal Hierarchical Model

- $\mathrm{p}(\theta_1, ..., \theta_J|\mu, \tau) = \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2)$

  $p(\theta_1, ..., \theta_J) = \int \prod_{j=1}^{J} [N(\theta_j|\mu, \tau^2)] p(\mu, \tau) d(\mu, \tau)$
- We can assign a noninformative uniform hyperprior distribution to mu, given tau:

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

# Joint Posterior Distribution

- Combining the sampling distribution for the observable $y_{ij}$ we can express:

$$\mathrm{p}(\theta, \mu, \tau | y) \propto p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta)$$
$$\propto p(\mu, \tau) \prod_{j=1}^{J} N(\theta_j | \mu, \tau^2) \prod_{j=1}^{J} N(\overline{y_{.j}} | \theta_j, \sigma^2)$$

- We can say that theta parameters are independent of the prior distribution, then:

$$\theta_j | \mu, \tau, y \sim N(\hat{\theta}_j, V_j) \ldots \ being\ proper$$
$$where, \hat{\theta}_j = \frac{\overline{y_{.j}} \sigma^{-2} + \mu \tau^{-2}}{\sigma_j^{-2} + \tau^{-2}}, V_j = \frac{1}{\sigma_j^{-2} + \tau^{-2}}$$

# Marginal Posterior Distribution (Hyperparameters)

- Prev slide is only a partial solution. As mu and tau are unknown.
- In the normal model discussed the marginal likelihood has a simple form (which is not the case in other dist.).

$$\overline{y_{.j}}|\mu, \tau \sim N(\mu, \sigma_j^2 + \tau^2)$$

$$p(\mu, \tau|y) \propto p(\mu, \tau) \prod_{j=1}^{J} N(\overline{y_{.j}}|\mu, \sigma_j^2 + \tau^2)$$

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{\overline{y_{.j}}}{\sigma_j^2 + \tau^2}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}}, V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}$$

# Posterior distribution of τ

- We can now get it:
- This holds true as any value of μ, so all the factors of μ must cancel when simplification is done.

$$\mathrm{p}(\tau|y) = \frac{p(\mu,\tau|y}{p(\mu|\tau,y)}$$

$$\propto \frac{p(\tau)\prod_{j=1}^{J} N(\overline{y_{.j}}|\mu\sigma_j^2+\tau^2)}{N(\mu|\hat{\mu},V_\mu)}$$

- If we set μ to hat(μ)

$$\mathrm{p}(\tau|y) \propto \frac{p(\tau)\prod_{j=1}^{J} N(\overline{y_{.j}}|\hat{\mu}\sigma_j^2+\tau^2)}{N(\hat{\mu}|\hat{\mu},V_\mu)}$$

$$\propto p(\tau)V_\mu^{\frac{1}{2}}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-\frac{1}{2}}exp[-\frac{(\overline{y_{.j}}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}]$$

# Prior distribution of τ

- Prior distribution for τ should be assigned: a diffuse noninf. prior will be used for convenience.
- It should be a finite integral: $p(\tau) \propto 1$
- Other priors can be defined:
  - $p(log\tau) \propto 1$
  - Inverse chi-squared: being a natural choise for variance parameters.

# Posterior Predictive Distributions

- There are 2 scenarios
  - Taking the future data from current batches
    $$\theta = (\theta_1, ..., \theta_j)$$
  - Future data from future batches ($y_{new}$)
    $$\tilde{\theta} = (\tilde{\theta}_1, ..., \tilde{\theta}_{\tilde{J}})$$
    - Steps to follow in this scenario:
      (i) Draw (μ, τ) from the posterior, (ii) draw Ĵ new params from the population distribution p(tilde{θ}$_J$|μ, τ), (iii) draw ŷ given tilde{θ} from the distribution

# Example

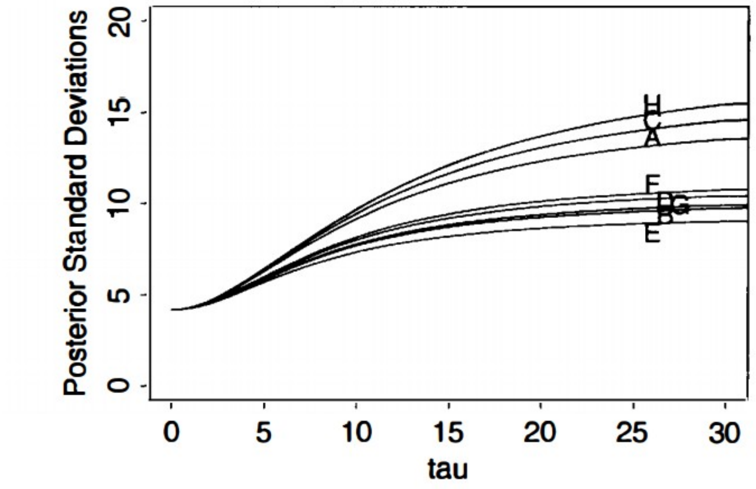
- SAT-Verbal is applied to 8 different schools with a coaching program.
  - Variable of interest: score, with values 200-800, mean 500, standard deviation 100
- Separate estimates: It’s statistically difficult to distinguish between experiments: yielding 95% posterior intervals overlapping.

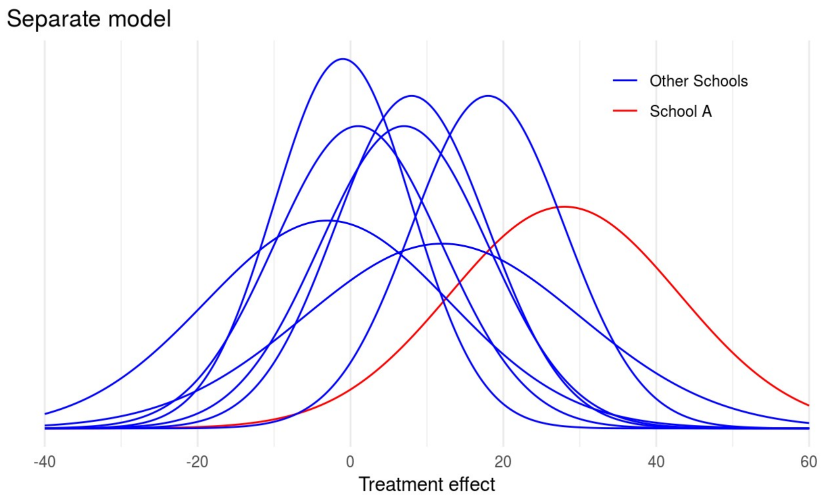
# Example:

- Overlapping in the normal non pooled simple model suggests that we’re estimating the same quantity.
- Pooled estimate, in contrast, hypothesize that all experiments have the same effect and produce independent estimates of this common effect.

# Example

- Anyway, the posterior mode of τ is on the boundary of its parameter space.
- The same for the joint posterior modal estimate.



# Meta-Analysis

- It's increasingly popular to summarize and integrate the findings of research studies.
- It's a method for combining several parallel data sources.
- We introduce 0 subscripts for control groups, and 1 in treatment groups.

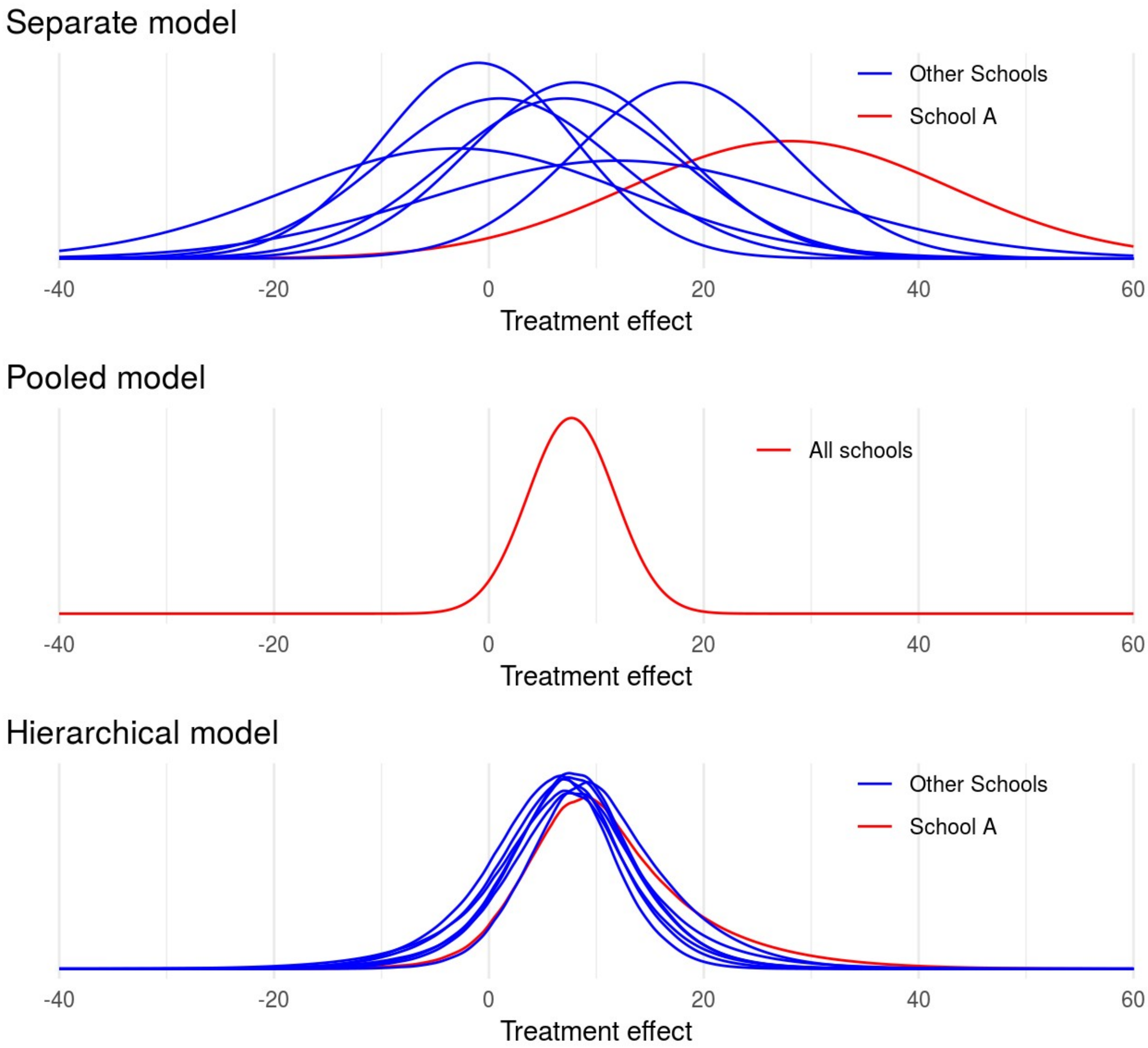
# Example: Hierarchical Normal Model

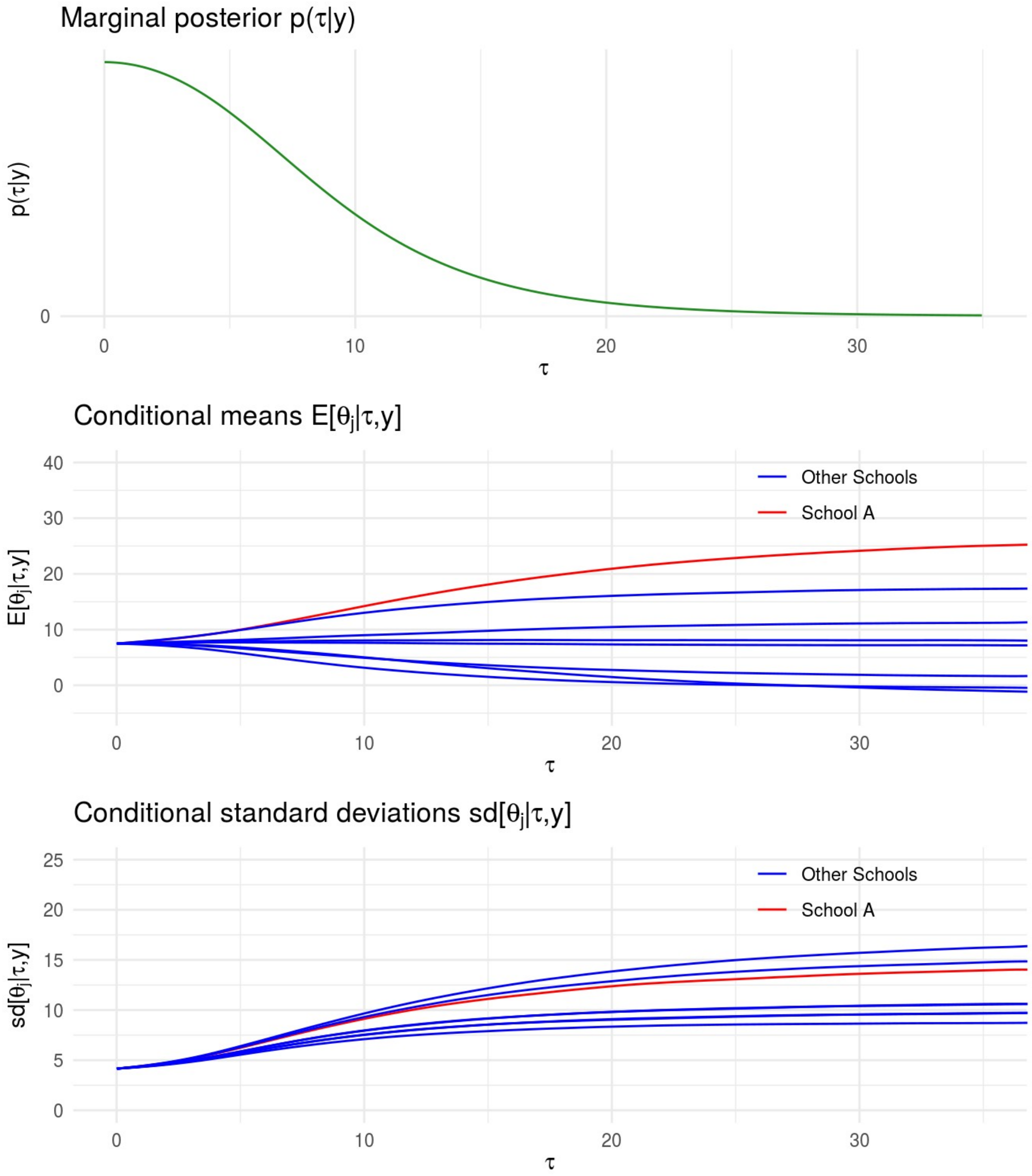
- $\sigma_j^2 = \frac{1}{y_{1j}} + \frac{1}{n_{1j}-y_{1j}} + \frac{1}{y_{0j}} + \frac{1}{n_{0j}-y_{0j}} \qquad Appr\ sampling\ variance$

$y_j = log(\frac{y_{ij}}{n_{1j}-y_{1j}}) - log\frac{y_{0j}}{n_{0j}-y_{0j}} \qquad empirical\ logits$

$y_j|\theta_j, \sigma_j^2 \sim N(\theta_j, \sigma_j^2), j = 1, \ldots, J$

- This method has a marginal posterior that peaks at nonzero value, which is plausible.
- The Expected value of μ is shrinked, compared with the non-hierarchical model. But the variance is accounted for, as it’s uncertain in the estimation.


Separate model
Other Schools
School A
-40
-20
0
20
40
60
Treatment effect
Pooled model
All schools
-40
-20
0
20
40
60
Treatment effect
Hierarchical model
Other Schools
School A
-40
-20
0
20
40
60
Treatment effect


Marginal posterior p(τ|y)
p(τ|y)
0
0
10
20
30
τ
Conditional means E[θj|τ,y]
40
30
20
10
0
E[θj|τ,y]
Other Schools
School A
0
10
20
30
τ
Conditional standard deviations sd[θj|τ,y]
25
20
15
10
5
0
sd[θj|τ,y]
Other Schools
School A
0
10
20
30
τ

# Example 2

- $y_{ij}, i = 1, \ldots, n_j,\ j = 1, \ldots, J, n = \sum_{j=1}^{J} n_j$
- Independently normally distributed, for each J groups. With n total number of observations.
- Data variance ($\sigma^2$) is unknown
- The group means, we assume, follow a Normal Distribution; with unknown $\mu$, $\tau^2$
- Prior for $(\mu, log\sigma, \tau), \sigma > 0, \tau > 0$

$$p(\mu, log\sigma, log\tau) \propto \tau$$

# Example 2

- If we assign a uniform prior for τ, it will result in an improper posterior.
- The joint posterior density of all parameters is:

$$p(\theta, , log\ \sigma, log\ \tau|y) \propto \tau \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2) \prod_{j=1}^{J} \prod_{i=1}^{n_j} N(y_i j|\theta_j, \sigma^2)$$

- P 326 – Gelman

# References


- Carlin, J., Gelman, A., & Stern, H., et al. (2003). Bayesian data analysis. Chapman and Hall/CRC.
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). The Bugs Book: A Practical Introduction to Bayesian Analysis (1st ed.).