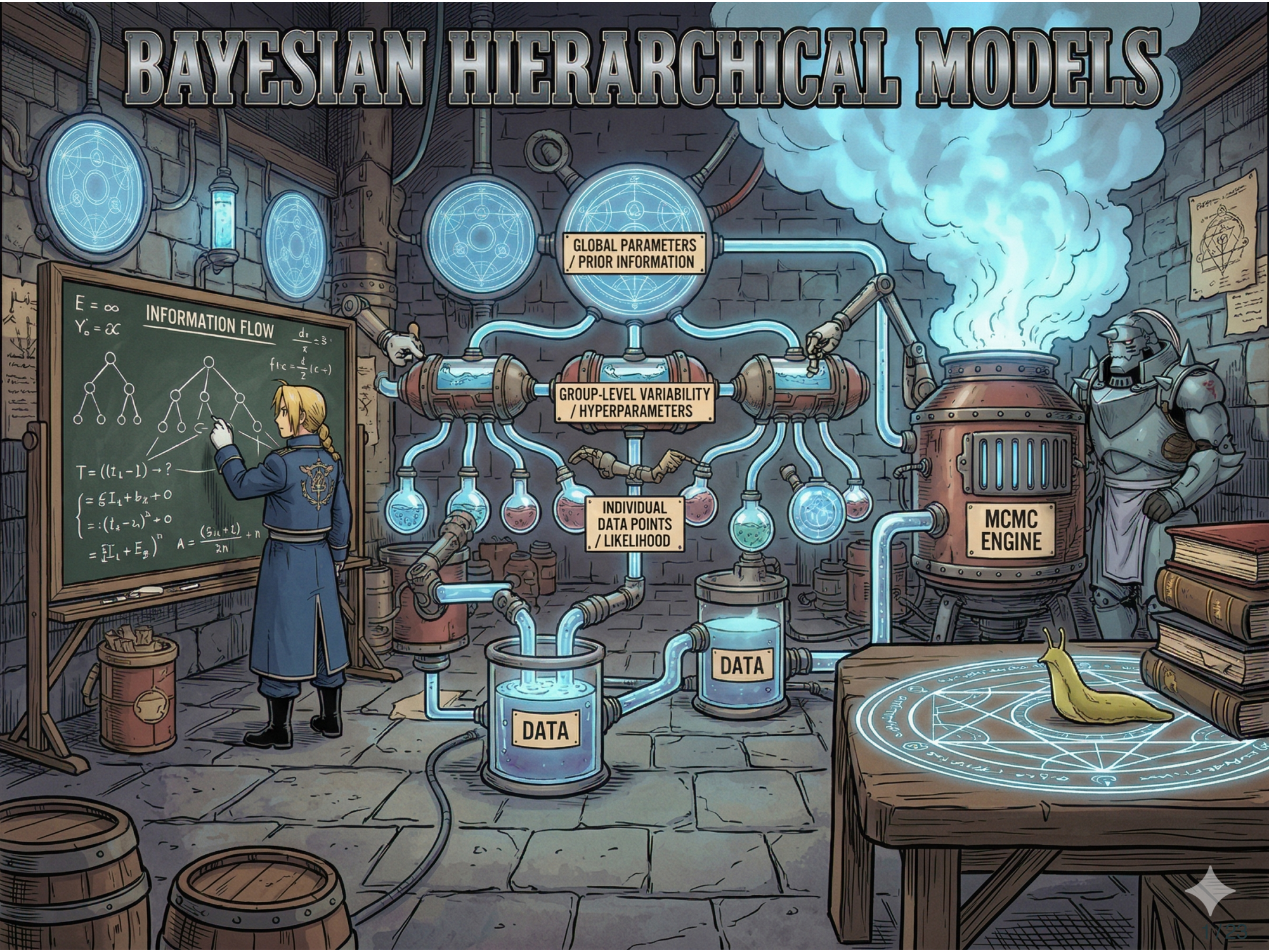

BAYESIAN HIERARCHICAL MODELS
GLOBAL PARAMETERS / PRIOR INFORMATION
GROUP-LEVEL VARIABILITY / HYPERPARAMETERS
INDIVIDUAL DATA POINTS / LIKELIHOOD
INFORMATION FLOW
MCMC ENGINE
DATA
DATA

Learning Objectives

- Understand the motivation of shrinkage estimation
- Develop hierarchical models for continuous and categorical data
- Make inference on both population and group-level parameters
- Apply hierarchical models to real datasets

James Stein's Paradox

Setup

Consider:

$$\theta_i \sim_{iid} N(0,1), \quad i = 1, \ldots, m$$
$$x_i|\theta_i \sim_{indep} N(\theta_i, 1), \quad i = 1, \ldots, m$$

What is a good estimator for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$?

An intuitive answer is the MLE $\hat{\boldsymbol{\theta}} = \boldsymbol{x}$, since all the X_i's are independent.

Surprisingly, James and Stein (1961) showed this is inadmissible and proposed

$$\hat{\boldsymbol{\theta}}^{JS} = \left(1 - \frac{m-2}{\|\boldsymbol{x}\|_2^2}\right)\boldsymbol{x}$$

which satisfies: when $m \geq 3$, for all $\boldsymbol{\theta}$,

$$\mathbb{E}(\|\hat{\boldsymbol{\theta}}^{JS} - \boldsymbol{\theta}\|_2^2) < \mathbb{E}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2)$$

The MLE is inadmissible!

James Stein's Paradox

```
m <- 10
risk.x <- risk.JS <- 0
for(rep in 1:1000){
  theta <- rnorm(m, 0, 1)
  x <- rnorm(m, theta, 1)
  JS <- (1 - (m - 2) / sum(x^2)) * x
  risk.x <- risk.x + mean((x - theta)^2)
  risk.JS <- risk.JS + mean((JS - theta)^2)
}
c(risk.x / 1000, risk.JS / 1000)
```

- The form of the JS estimator will appear often in our discussion of hierarchical models
- If we want to minimize MSE for multiple measurements, combine all measurements
- JS estimator shrinks each component of $\boldsymbol{x}$ toward 0. There is nothing special about 0. Can construct variants:
  $\hat{\boldsymbol{\theta}}^{JS} = (1 - \frac{m-2}{\|\boldsymbol{x}-\boldsymbol{\theta}_0\|_2^2})(\boldsymbol{x} - \boldsymbol{\theta}_0) + \boldsymbol{\theta}_0$
- JS estimator itself is not admissible either!

Consider the example presented in the textbook about combining information from educational testing experiments in eight schools.
A study was performed to analyze the effects of special coaching programs on SAT-V scores in 8 high schools. The observed effects of special preparation are estimates based on separate analyses for the eight school experiments. The effects, are labeled as $\bar{y}_{\cdot j}$. Over 30 students were tested on each school.

School	A	B	C	D	E	F	G	H
$\bar{y}_{\cdot j}$	28.39	7.94	-2.75	6.82	-0.64	0.63	18.01	12.16
σ_j	14.9	10.2	16.3	11.0	9.4	11.4	10.4	17.6

A statistical model

- We have $J = 8$ independent experiments, each of which estimates a θ_j from n_j independent data points $y_{ij}, i = 1, ..., n_j$.
- What estimates might be reasonable for $\theta = (\theta_1, ..., \theta_J)$?
- The classical statistics approach is to perform an analysis of variance (ANOVA) to test for differences among the means. Assume for now that $n_j = n$ and $\sigma_j = \sigma$.

Variability	Degrees of Freedom	Sum of Squares	Mean Squares
Between	$J-1$	$\sum_i \sum_j (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$	$SS/(J-1)$
Within	$J(n-1)$	$\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot j})^2$	$SS/(J(n-1))$
Total	$Jn-1$	$\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2$	$SS/(Jn-1)$

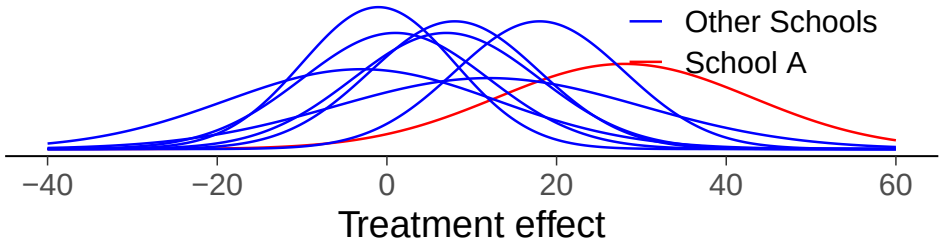
- Then we test whether the ratio $MS_{between}/MS_{within}$ is significantly greater than 1. If yes, then the data favor $\hat{\theta}_j = \bar{y}_{\cdot j}$. Otherwise, $\hat{\theta}_j = \bar{y}_{\cdot\cdot}$.

"I went off on binary conceptions of the world and said there was no way I was swallowing some symbolic reduction of my life."
(Matrix Resurrection, 2022)

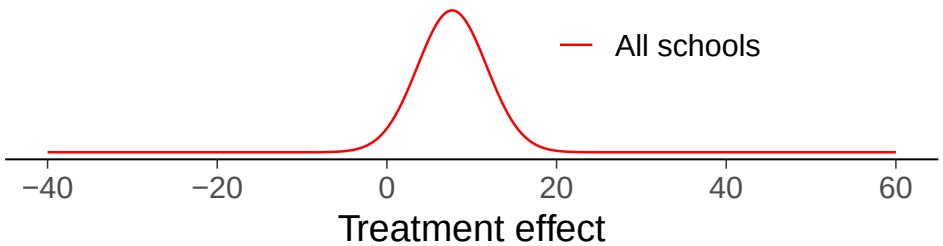
- Usually, neither complete pooling ($\hat{\theta}_j = \bar{y}_{..}$) nor no pooling $\hat{\theta}_j = \bar{y}_{.j}$ is ideal.
- From a Bayesian perspective, complete pooling is equivalent to assuming $\theta_j = \theta$ and using a uniform prior on θ. This ignores any between group variability, and **can be biased** if data are not collected by simple random sampling.
  - The overall mean is 7.9. Do we believe that effect in school A has 50% chance of being below 7.9?
- No pooling is equivalent to using independent uniform priors on θ_j. This could lead to **unstable** estimates when data are sparse.
  - Do we believe the effect in school A has 50% chance of being above 28.39?
- Rather, we want to couple the different groups by assuming the parameters are linked through a common probability distribution.

Results

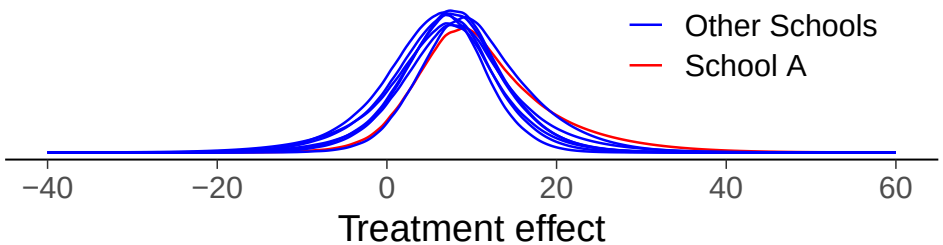
Separate model



Pooled model



Hierarchical model



"And the woman with the pills laughed 'cause I was missing the point. What point?
The choice is an illusion. You already know what you have to do."
(Matrix Resurrection, 2022)

Model Structure

Assume we know $\sigma_j^2 = \sigma^2/n_j$:

$$\bar{y}_{\cdot j}|\theta_j \sim N(\theta_j, \sigma_j^2)$$
$$\theta_j|\mu, \tau^2 \sim N(\mu, \tau^2)$$

with priors $p(\mu) \propto 1$ and $p(\tau) \propto 1$.

- Level 1: Observed individual school effects $\bar{y}_{\cdot j}$ vary around school-specific mean θ_j
- Level 2: School-specific mean θ_j follow a population distribution with mean μ and variance τ^2 (between-school variability)
- Level 3: Priors on hyperparameters (μ, τ)
- The joint posterior is

$$p(\theta, \mu, \tau|y) \propto p(\mu, \tau) \prod_j N(\theta_j; \mu, \tau^2) \prod_j N(\bar{y}_{\cdot j}; \theta_j, \sigma_j^2)$$

Posterior for School Effects

Conditional Posterior

Given μ, τ, y, the posterior for school effect θ_j is:

$$\theta_j | \mu, \tau, y \sim N(m_j, V_j)$$

where

$$m_j = \frac{\bar{y}_{\cdot j}/\sigma_j^2 + \mu/\tau^2}{1/\sigma_j^2 + 1/\tau^2}, \quad V_j = \frac{1}{1/\sigma_j^2 + 1/\tau^2}$$

The posterior mean m_j is a weighted average of:

- Observed estimate: $\bar{y}_{\cdot j}$ (precision $1/\sigma_j^2$)
- Prior mean: μ (precision $1/\tau^2$)

More precise schools (small σ_j) stay closer to observed data; uncertain schools shrink toward μ.

Posterior marginals

- The normal-normal model makes it easy to identify posterior marginal distributions.
- Integrating out θ_j, we get the marginal posteriors for μ and τ

$$p(\mu, \tau|y) \ \propto \ p(\mu, \tau) \prod_j N(\bar{y}_{\cdot j}; \mu, \sigma_j^2 + \tau^2)$$

- Uniform prior for μ here is reasonable since the data is very informative about the mean. Let us say $p(\mu, \tau) \ \propto \ p(\tau)$.
- Then we have

$$\mu|\tau, y = N(\mu_0, V_0)$$

$$\mu_0 = \frac{\sum_j \frac{1}{\sigma_j^2+\tau^2} \bar{y}_{\cdot j}}{\sum_j \frac{1}{\sigma_j^2+\tau^2}}, \quad V_0^{-1} = \sum_j \frac{1}{\sigma_j^2 + \tau^2}$$

Prior choices for τ

- For $p(\tau|y)$, we observe that it can be written as $p(\tau|y) = p(\mu, \tau|y)/p(\mu|\tau, y)$.
- Since it holds for any μ, i.e., the RHS can simplify to an expression without μ, so we can plug in $\mu = \mu_0$ for a simple evaluation

$$p(\tau|y) \propto p(\tau)V_0^{1/2}\prod_j(\sigma_j^2+\tau^2)^{-1/2}\exp(-\frac{(\bar{y}_{\cdot j}-\mu_0)^2}{2(\sigma_j^2+\tau^2)})$$

- What are the reasonable priors for τ, or equivalently $\log(\tau)$ or τ^2?

1. Improper uniform prior on τ, i.e. $\tau \sim Unif(0, A)$ as $A \to \infty$.

$$p(\tau) \propto 1 \quad or \quad p(\tau^2) \propto 1/\tau$$

2. Improper uniform prior on $\log(\tau)$, i.e. $\log(\tau) \sim Unif(-A, A)$ as $A \to \infty$.

$$p(\log(\tau)) \propto 1 \quad or \quad p(\tau) \propto 1/\tau \quad or \quad p(\tau^2) \propto 1/\tau^2$$

3. Inverse Gamma prior on τ^2. To achieve non-informativeness, say, let $\tau^2 \sim$ Inv-Gamma(ϵ, ϵ) with small ϵ values such as $1, 0.01, 0.001, ...$

Prior choices for τ

- For option 1 and 2, the first thing we need to check is that the posterior is proper. Since the posterior of μ and θ are proper, we only need to check $p(\tau|y)$ is proper, i.e., is $p(\tau|y)$ bounded by an integrable function for $\tau \in (0, \infty)$?

$$p(\tau|y) \propto p(\tau)\Big(\sum_j \frac{1}{\sigma_j^2+\tau^2}\Big)^{-1/2}\prod_j(\sigma_j^2+\tau^2)^{-1/2}\exp\Big(-\frac{(\bar{y}_{\cdot j}-\mu_0)^2}{2(\sigma_j^2+\tau^2)}\Big)$$

$$= p(\tau)\Big(\sum_j\prod_{k\neq j}(\sigma_j^2+\tau^2)\Big)^{-1/2}\prod_j\exp\Big(-\frac{(\bar{y}_{\cdot j}-\mu_0)^2}{2(\sigma_j^2+\tau^2)}\Big)$$

- First, notice that when $\tau \to 0$, the colored terms converge to a constant. Thus $p(\tau) \propto 1/\tau$ leads to an improper posterior as $1/\tau$ is not integrable for any interval $(0, A)$.
- On the large τ side, when $\tau > 1$, the term $\Big(\sum_j \prod_{k\neq j}(\sigma_j^2+\tau^2)\Big)^{-1/2}$ is bounded above by

$$\Big(\sum_j\prod_{k\neq j}(\tau^2)\Big)^{-1/2} = (J\tau^{2(J-1)})^{-1/2} \propto 1/\tau^{J-1}$$

- We can bound $\prod_j \exp\Big(-\frac{(\bar{y}_{\cdot j}-\mu_0)^2}{2(\sigma_j^2+\tau^2)}\Big)$ by $\exp\Big(-\frac{A}{\tau^2}\Big)$ for some small enough constant A.

Prior choices for τ

- As for the Inv-Gamma(ϵ, ϵ) prior, it turns out the posterior is very sensitive to the choice of ϵ.

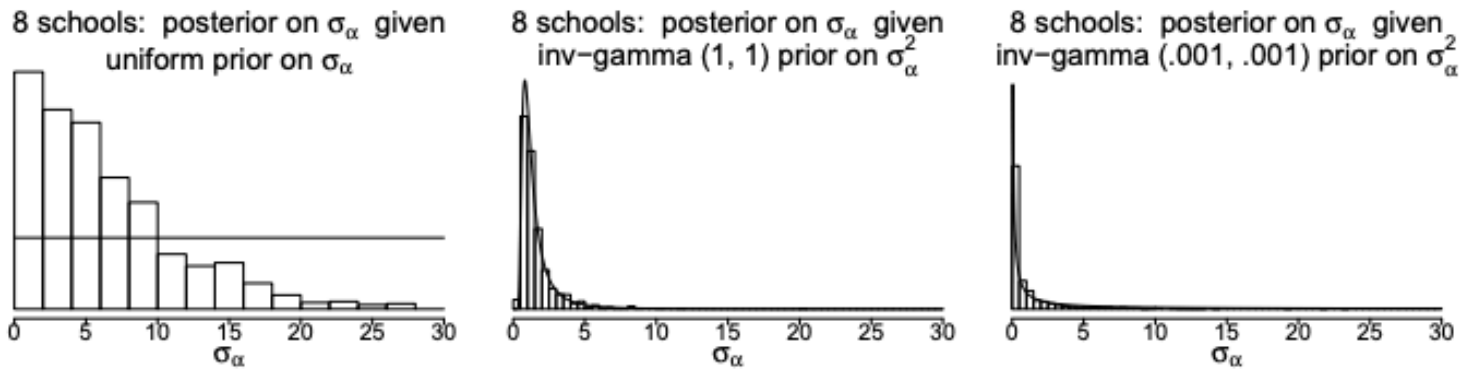
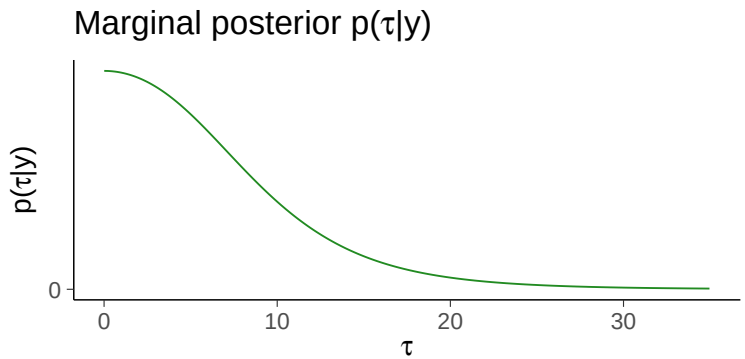
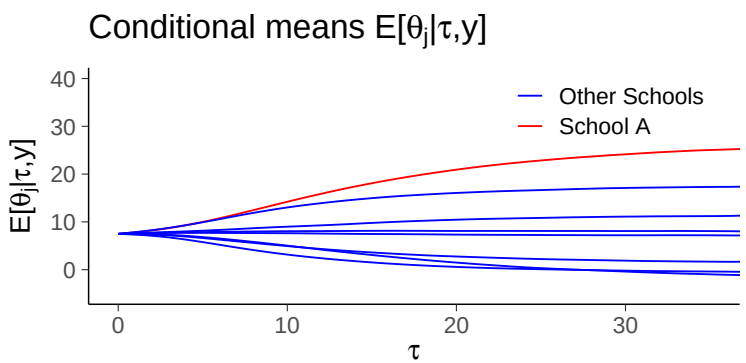
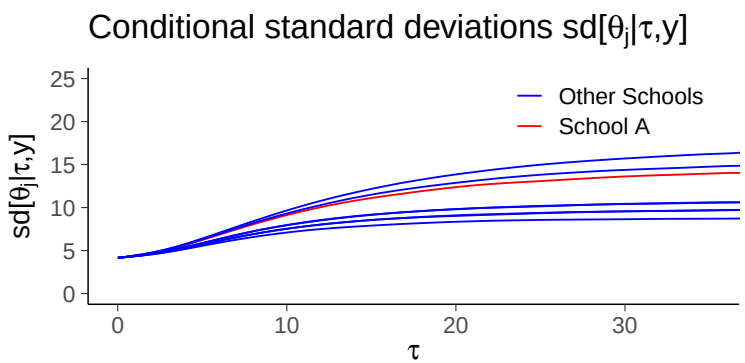


Figure 1: Histograms of posterior simulations of the between-school standard deviation, σ_α, from models with three different prior distributions: (a) uniform prior distribution on σ_α, (b) inverse-gamma$(1, 1)$ prior distribution on σ_α^2, (c) inverse-gamma$(0.001, 0.001)$ prior distribution on σ_α^2. Overlain on each is the corresponding prior density function for σ_α. (For models (b) and (c), the density for σ_α is calculated using the gamma density function multiplied by the Jacobian of the $1/\sigma_\alpha^2$ transformation.) In models (b) and (c), posterior inferences are strongly constrained by the prior distribution. Adapted from Gelman et al. (2003, Appendix C).

Results under prior option 1: $p(\tau) \propto 1$







Prior choices for τ

- Thus for non-informative priors, the uniform prior on τ is preferred in this model, as J is reasonably large for us to estimate τ.
- In practice, there are more considerations:
- Do I have enough groups (J) to estimate the variance?
  - If not, maybe we should use weakly informative priors.
  - Gelman's 2006 paper recommend half-Cauchy prior for τ (we will see more details in future homework problems).
  - An increasingly popular class of prior is the so-called penalized complexity (or PC) priors. The framework puts a constant decay prior on the KL divergence of the assumed model from a baseline model, and allows a simple specification with two interpretable parameters $p(\tau > U) = \alpha$. The PC prior is $\tau \sim Exp(-\log(\alpha)/U)$, or a type-2 Gumbel distribution on $1/\tau^2$.
- Do I have prior information on the variance (e.g., 5 is impossibly large)?
  - If you draw from the prior $p(\tau)$, what fractions of them fall into a reasonable range?
- How sensitive is my posterior to my prior specification?
- ...

Hierarchical model for binomial data

- Consider a survey where you collect binary individual responses y_{ik}, for the sampled individuals k in the i-th region.
- If we take $y_i = \sum_{k \in n_i} y_{ik}$, the obvious sampling model is:

$$y_i | \theta_i \sim \text{Binomial}\left(n_i, \theta_i\right).$$

- The MLE of θ without borrowing information leads to $\widehat{\theta}_i = y_i / n_i$. Notice that the standard error does not exist if $y_i = n_i$ or 0.
- A very simple hierarchical model for θ_i is to let

$$\theta_i | a, b \sim \text{Beta}(a, b),$$

- This leads to the posterior $\theta_i | y_i, a, b \sim \text{Beta}(a + y_i, b + n_i - y_i)$.
- The posterior mean estimate is

$$\begin{aligned} \widehat{\theta}_i &= \frac{a + y_i}{a + b + n_i} \\ &= \underbrace{\frac{a}{a+b}}_{\text{Prior Mean}} \frac{a+b}{a+b+n_i} + \underbrace{\frac{y_i}{n_i}}_{\text{Observed}} \frac{n_i}{a+b+n_i} \end{aligned}$$

- The posterior variance is usually smaller than the prior variance (but depending on specific values of a, b, n, and y).

Binomial GLMM Model

- The beta prior model is computationally convenient but is not very flexible.
- A Binomial GLMM random effect model is given by:

$$\begin{aligned} y_i|\theta_i &\sim \text{Binomial}(n_i, \theta_i) \\ \log\left(\frac{\theta_i}{1-\theta_i}\right) &= \beta_0 + \delta_i \\ \delta_i|\sigma_\delta^2 &\sim \text{N}(0, \sigma_\delta^2) \end{aligned}$$

  where δ_i are area-specific random effects that capture the residual or unexplained (logit) risk in area $i, i = 1, \ldots, n$.
- It is straightforward to add area-level covariates to this model via

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_1 + \delta_i.$$

  Note such models can be easily fitted using many automatic Bayesian computation software (e.g., Stan, INLA, JAGS, NIMBLE, ...)

Small area estimation of diabetes risk

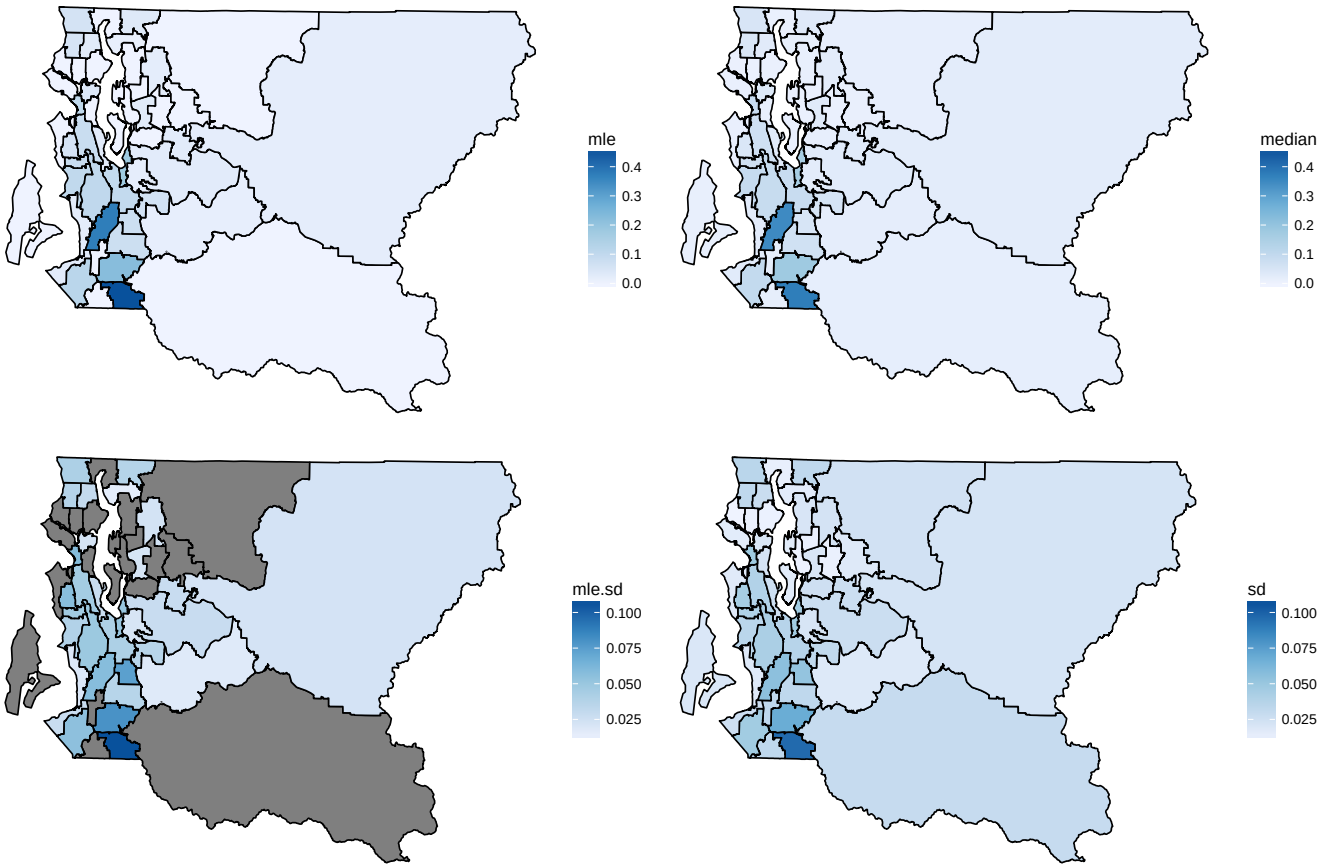


Figure 1: Top row: Estimates of area proportions with diabetes via MLE's (left) and posterior medians (right). Bottom row: Uncertainty of estimates with standard errors (left) and posterior standard deviations (right).

Small area estimation of diabetes risk

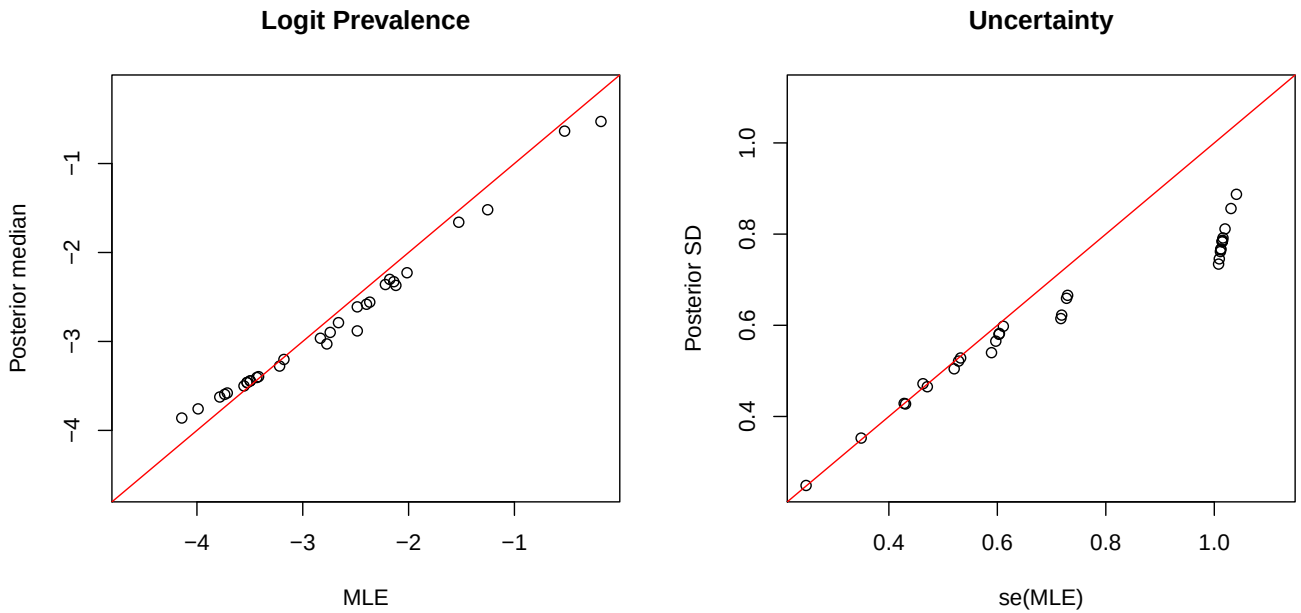


Figure 2: Comparison of area averages: Posterior medians versus MLEs on the logistic scale (left). Posterior standard deviations versus standard errors associated with the MLEs on the logistic scale (right).

Prior choices for logistic and other generalized linear (mixed) models

- This example is one particular case of generalized linear mixed model (GLMM). If we do not have the random effect δ_i, it falls back into the generalized linear model (GLM).
- Prior choices for generalized linear (mixed) models can usually be tricky and it is still on-going research as the model becomes more complicated. See Canvas reference page for Gelman et al (2008) and Fong, Rue, and Wakefield (2010) for more discussions.
- One useful tool to parameterize prior information is by checking the prior quantiles of an interpretable quantity.
- For example, consider the model

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + x_i\beta_1 + \delta_i$$

- The prior for β_0 is usually less consequential as even with a small amount of data, the overall intercept should be estimatable.
- How about the priors for the regression coefficient β_1 and the random effect δ_i?

Prior choices for logistic and other generalized linear (mixed) models

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + x_i\beta_1 + \delta_i$$

- Suppose we want to model $\beta_1 \sim N(0, \sigma^2)$.
- Notice that $\exp(\beta_1)$ is the prior odds ratio between subjects with $x = 1$ and $x = 0$.
- If we have prior information that such odds ratio is unlikely to exceed a certain value (say, 3), we can use this information to choose σ^2

```
> b = rnorm(1e5, 0, 0.668)
> quantile(exp(b), c(0.5, 0.95))
      50%       95%
0.9988231 2.9932162
```

Prior choices for logistic and other generalized linear (mixed) models

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + x_i\beta_1 + \delta_i$$

- Similarly, suppose we want to borrow information on δ_i and treat it as a random effect, with $\delta_i \sim_{iid} N(0, \tau^2)$ and $\tau^2 \sim$ Inv-Gamma(a, b).
- We can do the same thing for a and b. Say if we expect the odds ratio between two subjects has prior 95% range on $[0.1, 10]$

```
> tau2 <- 1/rgamma(1e5, 0.5, 0.0164)
> delta <- rnorm(1e5, 0, tau2^.5)
> quantile(exp(delta), c(0.025, 0.5, 0.975))
     2.5%       50%     97.5%
 0.102827  1.000780 10.526572
```

- The values $\sigma = 0.668, a = 0.5, b = 0.0164$ comes from analytically derived marginal distributions. See Fong, Held and Wakefield (2010) for details.