Hierarchical models II

Dr. Olanrewaju Michael Akande

March 25, 2020

Announcements

- Review changes to syllabus.
- Any concerns from the lab meetings?
- Going forward, there will be 5 minute breaks (roughly) halfway through each class meeting.
- Some initial details on final exam.

Outline

- Hierarchical modeling of means recap
- Hierarchical modeling of means and variances
- Gibbs sampler
- ELS data

Regular univariate normal model

- Recall that if we assume

$$y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \ldots, n,$$

and set our priors to be

$$\pi(\mu) = \mathcal{N}\left(\mu_0, \gamma_0^2\right).$$
$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right),$$

then we have

$$\pi(\mu, \sigma^2 | Y) \propto \left\{\prod_{i=1}^{n} p(y_i | \mu, \sigma^2)\right\} \cdot \pi(\mu) \cdot \pi(\sigma^2).$$

Full conditionals

- So that

$$\pi(\mu|\sigma^2, Y) = \mathcal{N}\left(\mu_n, \gamma_n^2\right).$$

where

$$\gamma_n^2 = \frac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\gamma_0^2}}; \qquad \mu_n = \gamma_n^2 \left[\frac{n}{\sigma^2}\bar{y} + \frac{1}{\gamma_0^2}\mu_0\right],$$

- and

$$\pi(\sigma^2|\mu, Y) = \mathcal{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right),$$

where

$$\nu_n = \nu_0 + n; \qquad \sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0 \sigma_0^2 + \sum_{i=1}^{n}(y_i - \mu)^2\right].$$

Hierarchical modeling of means recap

- We've looked at the hierarchical normal model of the form

$$y_{ij}|\theta_j, \sigma^2 \sim \mathcal{N}\left(\theta_j, \sigma^2_j\right); \quad i = 1, \ldots, n_j$$
$$\theta_j|\mu, \tau^2 \sim \mathcal{N}\left(\mu, \tau^2\right); \quad j = 1, \ldots, J.$$

- The model gives us an extra hierarchy through the prior on the means, leading to sharing of information across the groups, when estimating the group-specific means.

- As before, first set $\sigma^2_j = \sigma^2$ for all groups, to simplify posterior inference. We will revisit this today.

- Thus, we only have two variance terms, σ^2 and τ^2, to inform us on the within-group variation and between-group variation respectively.

Hierarchical normal model recap

- Standard semi-conjugate priors as before:

$$\pi(\mu) = \mathcal{N}\left(\mu_0, \gamma_0^2\right)$$
$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$
$$\pi(\tau^2) = \mathcal{IG}\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right).$$

  with

  - μ_0: best guess of average of school averages
  - γ_0^2: set based on plausible ranges of values of μ
  - τ_0^2: best guess of the (scaled) variance of school averages
  - η_0: set based on how tight prior for τ^2 is around τ_0^2
  - σ_0^2: best guess of the (scaled) variance of individual test scores around respective school means
  - ν_0: set based on how tight prior for σ^2 is around σ_0^2.

Posterior inference recap

- The resulting posterior is therefore:

$$
\begin{aligned}
\pi(\theta_1,\ldots,\theta_J,\mu,\sigma^2,\tau^2|Y) &\propto p(y|\theta_1,\ldots,\theta_J,\mu,\sigma^2,\tau^2) \\
&\quad \times p(\theta_1,\ldots,\theta_J|\mu,\sigma^2,\tau^2) \\
&\quad \times \pi(\mu,\sigma^2,\tau^2) \\
\\
&= p(y|\theta_1,\ldots,\theta_J,\sigma^2) \\
&\quad \times p(\theta_1,\ldots,\theta_J|\mu,\tau^2) \\
&\quad \times \pi(\mu)\cdot\pi(\sigma^2)\cdot\pi(\tau^2) \\
\\
&= \left\{\prod_{j=1}^{J}\prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma^2)\right\} \\
&\quad \times \left\{\prod_{j=1}^{J} p(\theta_j|\mu,\tau^2)\right\} \\
&\quad \times \pi(\mu)\cdot\pi(\sigma^2)\cdot\pi(\tau^2)
\end{aligned}
$$

Full conditional for grand mean

Recap

- $$\pi(\mu|\theta_1,\ldots,\theta_J,\sigma^2,\tau^2,Y) \propto \left\{\prod_{j=1}^{J} p(\theta_j|\mu,\tau^2)\right\} \cdot \pi(\mu).$$
- This looks like the full conditional distribution from the one-sample normal case, so that

$$\pi(\mu|\theta_1,\ldots,\theta_J,\sigma^2,\tau^2,Y) = \mathcal{N}\left(\mu_n,\gamma_n^2\right) \quad \text{where}$$

$$\gamma_n^2 = \frac{1}{\dfrac{J}{\tau^2}+\dfrac{1}{\gamma_0^2}}; \qquad \mu_n = \gamma_n^2\left[\frac{J}{\tau^2}\bar{\theta} + \frac{1}{\gamma_0^2}\mu_0\right]$$

and $\bar{\theta} = \frac{1}{J}\sum_{j=1}^{J}\theta_j$.

Full conditionals for group means

recap

- $$\pi(\theta_j|\theta_{-j},\mu,\sigma^2,\tau^2,Y) \propto \left\{\prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma^2)\right\} \cdot p(\theta_j|\mu,\tau^2)$$

- Those terms include a normal density for θ_j multiplied by a product of normal densities in which θ_j is the mean, again mirroring the one-sample case, so you can show that

$$\pi(\theta_j|\theta_{-j},\mu,\sigma^2,\tau^2,Y) = \mathcal{N}\left(\mu_j^\star,\tau_j^\star\right) \quad \text{where}$$

$$\tau_j^\star = \frac{1}{\frac{n_j}{\sigma^2}+\frac{1}{\tau^2}}; \qquad \mu_j^\star = \tau_j^\star\left[\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\mu\right]$$

Full conditionals for within-group variance recap

- $$\pi(\sigma^2|\theta_1,\ldots,\theta_J,\mu,\tau^2,Y) \propto \left\{\prod_{j=1}^{J}\prod_{i=1}^{n_j} p(y_{ij}|\theta_j,\sigma^2)\right\}\cdot\pi(\sigma^2)$$

- We can take advantage of the one-sample normal problem, so that our full conditional posterior is

$$\pi(\sigma^2|\theta_1,\ldots,\theta_J,\mu,\tau^2,Y) = \mathcal{IG}\left(\frac{\nu_n}{2},\frac{\nu_n\sigma_n^2}{2}\right) \quad \text{where}$$

$$\nu_n = \nu_0 + \sum_{j=1}^{J} n_j; \qquad \sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2\right].$$

Full conditionals for across-group variance recap

- $$\pi(\tau^2|\theta_1,\ldots,\theta_J,\mu,\sigma^2,Y) \propto \left\{\prod_{j=1}^{J} p(\theta_j|\mu,\tau^2)\right\} \cdot \pi(\tau^2)$$

- Again, we have

$$\pi(\tau^2|\theta_1,\ldots,\theta_J,\mu,\sigma^2,Y) = \mathcal{IG}\left(\frac{\eta_n}{2}, \frac{\eta_n \tau_n^2}{2}\right) \quad \text{where}$$

$$\eta_n = \eta_0 + J; \qquad \tau_n^2 = \frac{1}{\eta_n}\left[\eta_0 \tau_0^2 + \sum_{j=1}^{J}(\theta_j - \mu)^2\right].$$

Hierarchical modeling of means and variances

- Often researchers emphasize differences in means. However, variances can be very important.
- If we think means vary across groups, why shouldn't we worry about variances also varying across groups?
- In that case, we have the model

$$
\begin{aligned}
y_{ij}|\theta_j, \sigma^2 &\sim \mathcal{N}\left(\theta_j, \sigma_j^2\right); \quad i = 1, \ldots, n_j \\
\theta_j|\mu, \tau^2 &\sim \mathcal{N}\left(\mu, \tau^2\right); \quad j = 1, \ldots, J,
\end{aligned}
$$

- However, now we also need a prior on all the σ_j^2's that lets us borrow information about across groups.

Full conditionals

- Notice that our prior won't affect the full conditions for μ and τ^2 since those have nothing to do with all the σ_j^2's.

- The full conditional for each θ_j, we have

$$\pi(\theta_j|\theta_{-j}, \mu, \sigma_1^2, \ldots, \sigma_J^2, \tau^2, Y) \propto \left\{ \prod_{i=1}^{n_j} p(y_{ij}|\theta_j, \sigma_j^2) \right\} \cdot p(\theta_j|\mu, \tau^2)$$

  with the only change from before being σ_j^2.

- That is, those terms still include a normal density for θ_j multiplied by a product of normals in which θ_j is the mean, again mirroring the previous case, so you can show that

$$\pi(\theta_j|\theta_{-j}, \mu, \sigma_1^2, \ldots, \sigma_J^2, \tau^2, Y) = \mathcal{N}\left(\mu_j^\star, \tau_j^\star\right) \quad \text{where}$$

$$\tau_j^\star = \frac{1}{\dfrac{n_j}{\sigma_j^2} + \dfrac{1}{\tau^2}}; \qquad \mu_j^\star = \tau_j^\star \left[\frac{n_j}{\sigma_j^2} \bar{y}_j + \frac{1}{\tau^2} \mu \right]$$

How about within-group variances?

- Now we need to find a semi-conjugate prior for the σ_j^2's. Before, with one σ^2, we had

$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right),$$

  which was nicely semi-conjugate.

- That suggests that maybe we should start with.

$$\sigma_1^2, \ldots, \sigma_J^2 | \nu_0, \sigma_0^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

- However, if we just fix the hyperparameters ν_0 and σ_0^2 in advance, the prior on the σ_j^2's does not allow borrowing of information across other values of σ_j^2, to aid in estimation.

- Thus, we actually need to treat ν_0 and σ_0^2 as parameters in a hierarchical model for both means and variances.

How about within-group variances?

- Before we get to the choice of the priors for ν_0 and σ_0^2, we have enough to derive the full conditional for each σ_j^2. This actually takes a similar form to what we had before we indexed by j, that is,

$$\pi(\sigma_j^2 | \sigma_{-j}^2, \theta_1, \ldots, \theta_J, \mu, \tau^2, \nu_0, \sigma_0^2, Y) \propto \left\{ \prod_{i=1}^{n_j} p(y_{ij} | \theta_j, \sigma_j^2) \right\} \cdot \pi(\sigma_j^2 | \nu_0, \sigma_0^2)$$

- This still looks like what we had before, that is, products of normals and one inverse-gamma, so that

$$\pi(\sigma_j^2 | \sigma_{-j}^2, \theta_1, \ldots, \theta_J, \mu, \tau^2, \nu_0, \sigma_0^2, Y) = \mathcal{IG}\left(\frac{\nu_j^\star}{2}, \frac{\nu_j^\star \sigma_j^{2(\star)}}{2} \right) \quad \text{where}$$

$$\nu_j^\star = \nu_0 + n_j; \qquad \sigma_j^{2(\star)} = \frac{1}{\nu_j^\star} \left[\nu_0 \sigma_0^2 + \sum_{i=1}^{n_j} (y_{ij} - \theta_j)^2 \right].$$

Remaining hyper-priors

- Now we can get back to priors for ν_0 and σ_0^2. Turns out that a semi-conjugate prior for σ_0^2 (see question 2 on homework 2) is a gamma distribution. That is, if we set

$$\pi(\sigma_0^2) = \mathcal{G}a\left(a, b\right),$$

then,

$$\begin{aligned}\pi(\sigma_0^2|\theta_1, \ldots, \theta_J, \sigma_1^2, \ldots, \sigma_J^2, \mu, \tau^2, \nu_0, Y) &\propto \left\{\prod_{j=1}^{J} p(\sigma_j^2|\nu_0, \sigma_0^2)\right\} \cdot \pi(\sigma_0^2)\\ &\propto \mathcal{IG}\left(\sigma_j^2; \frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \cdot \mathcal{G}a\left(\sigma_0^2; a, b\right)\end{aligned}$$

- Recall that
  - $\mathcal{G}a(y; a, b) \equiv \dfrac{b^a}{\Gamma(a)} y^{a-1} e^{-by}$, and
  - $\mathcal{IG}(y; a, b) \equiv \dfrac{b^a}{\Gamma(a)} y^{-(a+1)} e^{-\dfrac{b}{y}}$.

Remaining hyper-priors

- So $\pi(\sigma_0^2 | \theta_1, \ldots, \theta_J, \sigma_1^2, \ldots, \sigma_J^2, \mu, \tau^2, \nu_0, Y)$

$$\propto \left\{ \prod_{j=1}^{J} p(\sigma_j^2 | \nu_0, \sigma_0^2) \right\} \cdot \pi(\sigma_0^2)$$

$$\propto \mathcal{IG}\left(\sigma_j^2; \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \cdot \mathcal{G}a\left(\sigma_0^2; a, b\right)$$

$$= \left[\prod_{j=1}^{J} \frac{\left(\frac{\nu_0 \sigma_0^2}{2}\right)^{\left(\frac{\nu_0}{2}\right)}}{\Gamma\left(\frac{\nu_0}{2}\right)} (\sigma_j^2)^{-\left(\frac{\nu_0}{2}+1\right)} e^{-\frac{\nu_0 \sigma_0^2}{2(\sigma_j^2)}} \right] \cdot \left[\frac{b^a}{\Gamma(a)} (\sigma_0^2)^{a-1} e^{-b\sigma_0^2} \right]$$

$$\propto \left[\prod_{j=1}^{J} (\sigma_0^2)^{\left(\frac{\nu_0}{2}\right)} e^{-\frac{\nu_0 \sigma_0^2}{2(\sigma_j^2)}} \right] \cdot \left[(\sigma_0^2)^{a-1} e^{-b\sigma_0^2} \right]$$

$$\propto \left[(\sigma_0^2)^{\left(\frac{J\nu_0}{2}\right)} e^{-\sigma_0^2 \left[\frac{\nu_0}{2} \sum_{j=1}^{J} \frac{1}{\sigma_j^2} \right]} \right] \cdot \left[(\sigma_0^2)^{a-1} e^{-b\sigma_0^2} \right]$$

Remaining hyper-priors

- That is, the full conditional is

$$
\begin{aligned}
\pi(\sigma_0^2 | \cdots\cdots) &\propto \left[\left(\sigma_0^2\right)^{\left(\frac{J\nu_0}{2}\right)} e^{-\sigma_0^2 \left[\frac{\nu_0}{2} \sum_{j=1}^{J} \frac{1}{\sigma_j^2} \right]} \right] \cdot \left[(\sigma_0^2)^{a-1} e^{-b\sigma_0^2} \right] \\
&\propto \left[\left(\sigma_0^2\right)^{\left(a + \frac{J\nu_0}{2} - 1\right)} e^{-\sigma_0^2 \left[b + \frac{\nu_0}{2} \sum_{j=1}^{J} \frac{1}{\sigma_j^2} \right]} \right] \\
&\equiv \mathcal{G}a\left(\sigma_0^2; a_n, b_n\right),
\end{aligned}
$$

where

$$
a_n = a + \frac{J\nu_0}{2}; \quad b_n = b + \frac{\nu_0}{2} \sum_{j=1}^{J} \frac{1}{\sigma_j^2}.
$$

Remaining hyper-priors

- Ok that leaves us with one parameter to go, i.e., ν_0. Turns out there is no simple conjugate/semi-conjugate prior for ν_0.
- Common practice is to restrict ν_0 to be an integer (which makes sense when we think of it as being degrees of freedom, which also means it cannot be zero). With the restriction, we need a discrete distribution as the prior with support on $\nu_0 = 1, 2, 3, \ldots$.
- **Poll question: Can we use either a binomial or a Poisson prior on for ν_0?**
- A popular choice is the geometric distribution with pmf $p(\nu_0) = (1-p)^{\nu_0 - 1} p$.
- However, we will rewrite the kernel as $\pi(\nu_0) \propto e^{-\alpha\nu_0}$. How did we get here from the geometric pmf and what is α?

Final full conditional

- With this prior, $\pi(\nu_0|\theta_1,\ldots,\theta_J,\sigma_1^2,\ldots,\sigma_J^2,\mu,\tau^2,\sigma_0^2,Y)$

$$\propto \left\{\prod_{j=1}^{J} p(\sigma_j^2|\nu_0,\sigma_0^2)\right\}\cdot\pi(\nu_0)$$

$$\propto \mathcal{IG}\left(\sigma_j^2;\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right)\quad\cdot\ e^{-\alpha\nu_0}$$

$$= \left[\prod_{j=1}^{J}\frac{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\left(\frac{\nu_0}{2}\right)}}{\Gamma\left(\frac{\nu_0}{2}\right)}\left(\sigma_j^2\right)^{-\left(\frac{\nu_0}{2}+1\right)}e^{-\frac{\nu_0\sigma_0^2}{2(\sigma_j^2)}}\right]\cdot e^{-\alpha\nu_0}$$

$$\propto \left[\left(\frac{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\left(\frac{\nu_0}{2}\right)}}{\Gamma\left(\frac{\nu_0}{2}\right)}\right)^{J}\cdot\left(\prod_{j=1}^{J}\frac{1}{\sigma_j^2}\right)^{\left(\frac{\nu_0}{2}-1\right)}\cdot e^{-\nu_0\left[\frac{\sigma_0^2}{2}\sum_{j=1}^{J}\frac{1}{\sigma_j^2}\right]}\right]\cdot e^{-\alpha\nu_0}$$

Final full conditional

- That is, the full conditional is

$$\pi(\nu_0 | \cdots\cdots) \propto \left[\left(\frac{\left(\frac{\nu_0 \sigma_0^2}{2}\right)^{\left(\frac{\nu_0}{2}\right)}}{\Gamma\left(\frac{\nu_0}{2}\right)} \right)^J \cdot \left(\prod_{j=1}^{J} \frac{1}{\sigma_j^2} \right)^{\left(\frac{\nu_0}{2} - 1\right)} \cdot e^{-\nu_0 \left[\alpha + \frac{\sigma_0^2}{2} \sum_{j=1}^{J} \frac{1}{\sigma_j^2}\right]} \right],$$

  which is not a known kernel and is thus unnormalized (i.e., does not integrate to 1 in its current form).

- This sure looks like a lot, but it will be relatively easy to compute in R.

- Now, technically, the support is $\nu_0 = 1, 2, 3, \ldots$, however, we can compute this to compute the unnormalized distribution across a grid of ν_0 values, say, $\nu_0 = 1, 2, 3, \ldots$ for some large K, and then sample.

Final full conditional

- One more thing, computing these probabilities on the raw scale can be problematic particularly because of the product inside. Good idea to transform to the log scale instead.
- That is,

$$\pi(\nu_0|\cdots\cdots) \propto \left[\left(\frac{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\left(\frac{\nu_0}{2}\right)}}{\Gamma\left(\frac{\nu_0}{2}\right)}\right)^J \cdot \left(\prod_{j=1}^{J}\frac{1}{\sigma_j^2}\right)^{\left(\frac{\nu_0}{2}-1\right)} \cdot e^{-\nu_0\left[\alpha+\frac{\sigma_0^2}{2}\sum_{j=1}^{J}\frac{1}{\sigma_j^2}\right]}\right]$$

$$\begin{aligned}\Rightarrow \ln\pi(\nu_0|\cdots\cdots) \propto & \left(\frac{J\nu_0}{2}\right)\ln\left(\frac{\nu_0\sigma_0^2}{2}\right) - J\ln\left[\Gamma\left(\frac{\nu_0}{2}\right)\right] \\ & + \left(\frac{\nu_0}{2}-1\right)\left(\sum_{j=1}^{J}\ln\left[\frac{1}{\sigma_j^2}\right]\right) \\ & - \nu_0\left[\alpha + \frac{\sigma_0^2}{2}\sum_{j=1}^{J}\frac{1}{\sigma_j^2}\right]\end{aligned}$$

ELS DATA

- Finally, enough math and some data!
- We have data from the 2002 Educational Longitudinal Survey (ELS). This survey includes a random sample of 100 large urban public high schools, and 10th graders randomly sampled within these high schools.

```
Y <- as.matrix(dget("http://www2.stat.duke.edu/~pdh10/FCBS/Inline/Y.school.mathscore"))
dim(Y)
```

```
## [1] 1993    2
```

```
head(Y)
```

```
##      school mathscore
## [1,]      1     52.11
## [2,]      1     57.65
## [3,]      1     66.44
## [4,]      1     44.68
## [5,]      1     40.57
## [6,]      1     35.04
```

```
length(unique(Y[,"school"]))
```

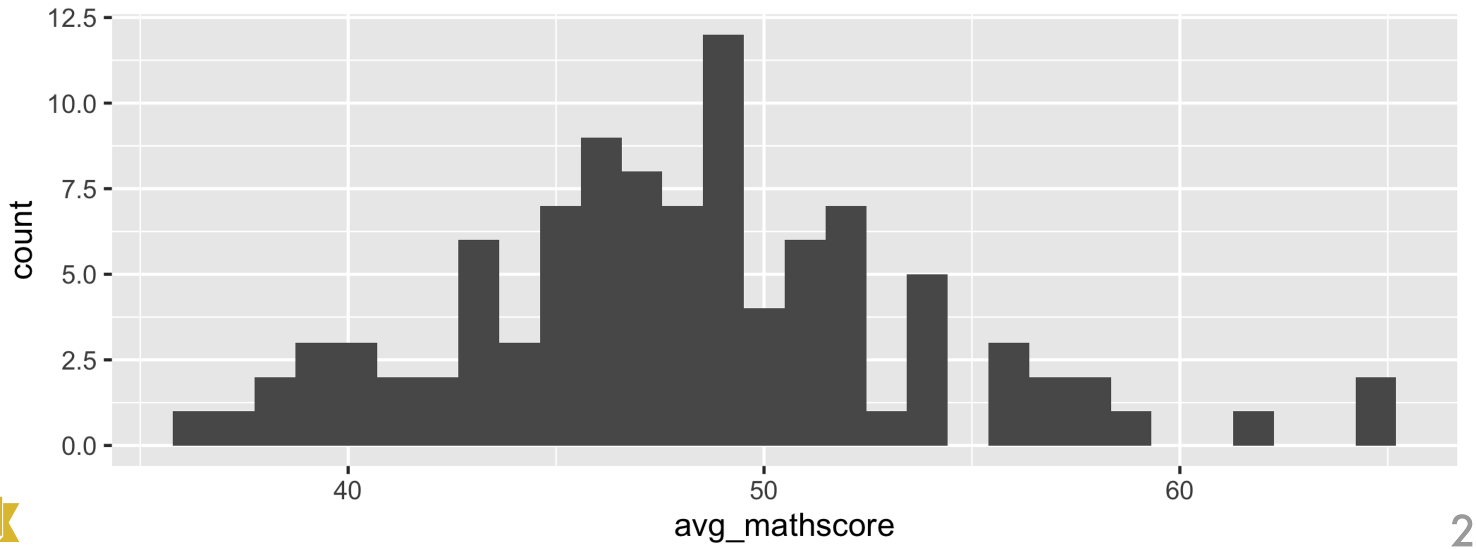
```
## [1] 100
```

ELS DATA

First, some EDA:

```
Data <- as.data.frame(Y)
Data$school <- as.factor(Data$school)
Data %>%
  group_by(school) %>%
  na.omit()%>%
  summarise(avg_mathscore = mean(mathscore)) %>%
  dplyr::ungroup() %>%
  ggplot(aes(x = avg_mathscore))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

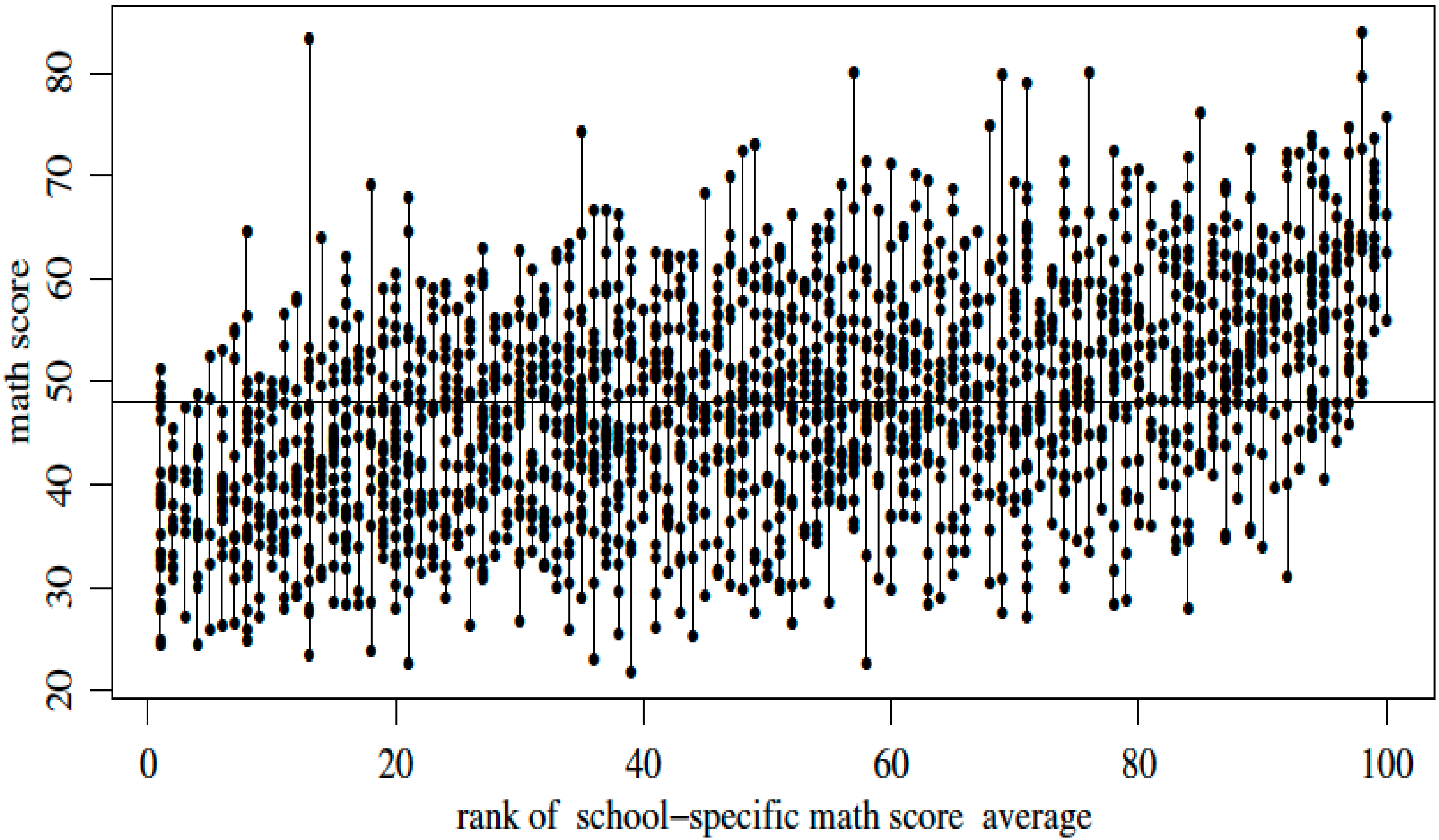


ELS DATA

There does appear to be school-related differences in means and in variances, some of which are actually related to the sample sizes.

ELS DATA

Consider the math scores of these children:



ELS HYPOTHESES

- Investigators may be interested in the following:
  - Differences in mean scores across schools
  - Differences in school-specific variances
- How do we evaluate these questions in a statistical model?

Hierarchical model

- We can write out the full model we've been describing as follows.

$$y_{ij}|\theta_j,\sigma^2 \sim \mathcal{N}\left(\theta_j,\sigma_j^2\right); \quad i=1,\ldots,n_j$$

$$\theta_j|\mu,\tau^2 \sim \mathcal{N}\left(\mu,\tau^2\right); \quad j=1,\ldots,J$$

$$\sigma_1^2,\ldots,\sigma_J^2|\nu_0,\sigma_0^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right)$$

$$\mu \sim \mathcal{N}\left(\mu_0,\gamma_0^2\right)$$

$$\tau^2 \sim \mathcal{IG}\left(\frac{\eta_0}{2},\frac{\eta_0\tau_0^2}{2}\right).$$

$$\pi(\nu_0) \propto e^{-\alpha\nu_0}$$

$$\sigma_0^2 \sim \mathcal{G}a\left(a,b\right).$$

- Now, we need to specify hyperparameters. That should be fun!

Prior specification

- This exam was designed to have a national mean of 50 and standard deviation of 10. Suppose we don't have any other information.
- Then, we can specify

$$\mu \sim \mathcal{N}\left(\mu_0 = 50, \gamma_0^2 = 25\right)$$

$$\tau^2 \sim \mathcal{IG}\left(\frac{\eta_0}{2} = \frac{1}{2}, \frac{\eta_0 \tau_0^2}{2} = \frac{100}{2}\right).$$

$$\pi(\nu_0) \propto e^{-\alpha\nu_0} \propto e^{-\nu_0}$$

$$\sigma_0^2 \sim \mathcal{G}a\left(a = 1, b = \frac{1}{100}\right).$$

- Are these prior distributions overly informative?

Full conditionals (recap)

- $$\pi(\theta_j | \cdots\cdots) = \mathcal{N}\left(\mu_j^\star, \tau_j^\star\right) \quad \text{where}$$
  $$\tau_j^\star = \frac{1}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}}; \qquad \mu_j^\star = \tau_j^\star \left[\frac{n_j}{\sigma_j^2}\bar{y}_j + \frac{1}{\tau^2}\mu\right]$$
- $$\pi(\sigma_j^2 | \cdots\cdots) = \mathcal{IG}\left(\frac{\nu_j^\star}{2}, \frac{\nu_j^\star \sigma_j^{2(\star)}}{2}\right) \quad \text{where}$$
  $$\nu_j^\star = \nu_0 + n_j; \qquad \sigma_j^{2(\star)} = \frac{1}{\nu_j^\star}\left[\nu_0\sigma_0^2 + \sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2\right].$$
- $$\pi(\mu | \cdots\cdots) = \mathcal{N}\left(\mu_n, \gamma_n^2\right) \quad \text{where}$$
  $$\gamma_n^2 = \frac{1}{\frac{J}{\tau^2} + \frac{1}{\gamma_0^2}}; \qquad \mu_n = \gamma_n^2 \left[\frac{J}{\tau^2}\bar{\theta} + \frac{1}{\gamma_0^2}\mu_0\right]$$

Full conditionals (recap)

- $$\pi(\tau^2|\cdots\cdots) = \mathcal{IG}\left(\frac{\eta_n}{2}, \frac{\eta_n\tau_n^2}{2}\right) \quad \text{where}$$
  $$\eta_n = \eta_0 + J; \qquad \tau_n^2 = \frac{1}{\eta_n}\left[\eta_0\tau_0^2 + \sum_{j=1}^{J}(\theta_j - \mu)^2\right].$$
- $$\begin{aligned}\ln\pi(\nu_0|\cdots\cdots) \propto & \left(\frac{J\nu_0}{2}\right)\ln\left(\frac{\nu_0\sigma_0^2}{2}\right) - J\ln\left[\Gamma\left(\frac{\nu_0}{2}\right)\right] \\ & + \left(\frac{\nu_0}{2} - 1\right)\left(\sum_{j=1}^{J}\ln\left[\frac{1}{\sigma_j^2}\right]\right) \\ & - \nu_0\left[\alpha + \frac{\sigma_0^2}{2}\sum_{j=1}^{J}\frac{1}{\sigma_j^2}\right]\end{aligned}$$
- $$\pi(\sigma_0^2|\cdots\cdots) = \mathcal{G}a\left(\sigma_0^2; a_n, b_n\right) \quad \text{where}$$
  $$a_n = a + \frac{J\nu_0}{2}; \quad b_n = b + \frac{\nu_0}{2}\sum_{j=1}^{J}\frac{1}{\sigma_j^2}.$$

Side notes

- Obviously, as you have seen in the lab, we can simply use Stan (or JAGS, BUGS) to fit these models without needing to do any of this ourselves.
- The point here (as you should already know by now) is to learn and understand all the details, including the math!

Gibbs sampler

```
#Data summaries
J <- length(unique(Y[,"school"]))
ybar <- c(by(Y[,"mathscore"],Y[,"school"],mean))
s_j_sq <- c(by(Y[,"mathscore"],Y[,"school"],var))
n <- c(table(Y[,"school"]))


#Hyperparameters for the priors
mu_0 <- 50
gamma_0_sq <- 25
eta_0 <- 1
tau_0_sq <- 100
alpha <- 1
a <- 1
b <- 1/100


#Grid values for sampling nu_0_grid
nu_0_grid<-1:5000


#Initial values for Gibbs sampler
theta <- ybar
sigma_sq <- s_j_sq
mu <- mean(theta)
tau_sq <- var(theta)
nu_0 <- 1
sigma_0_sq <- 100
```

Gibbs sampler

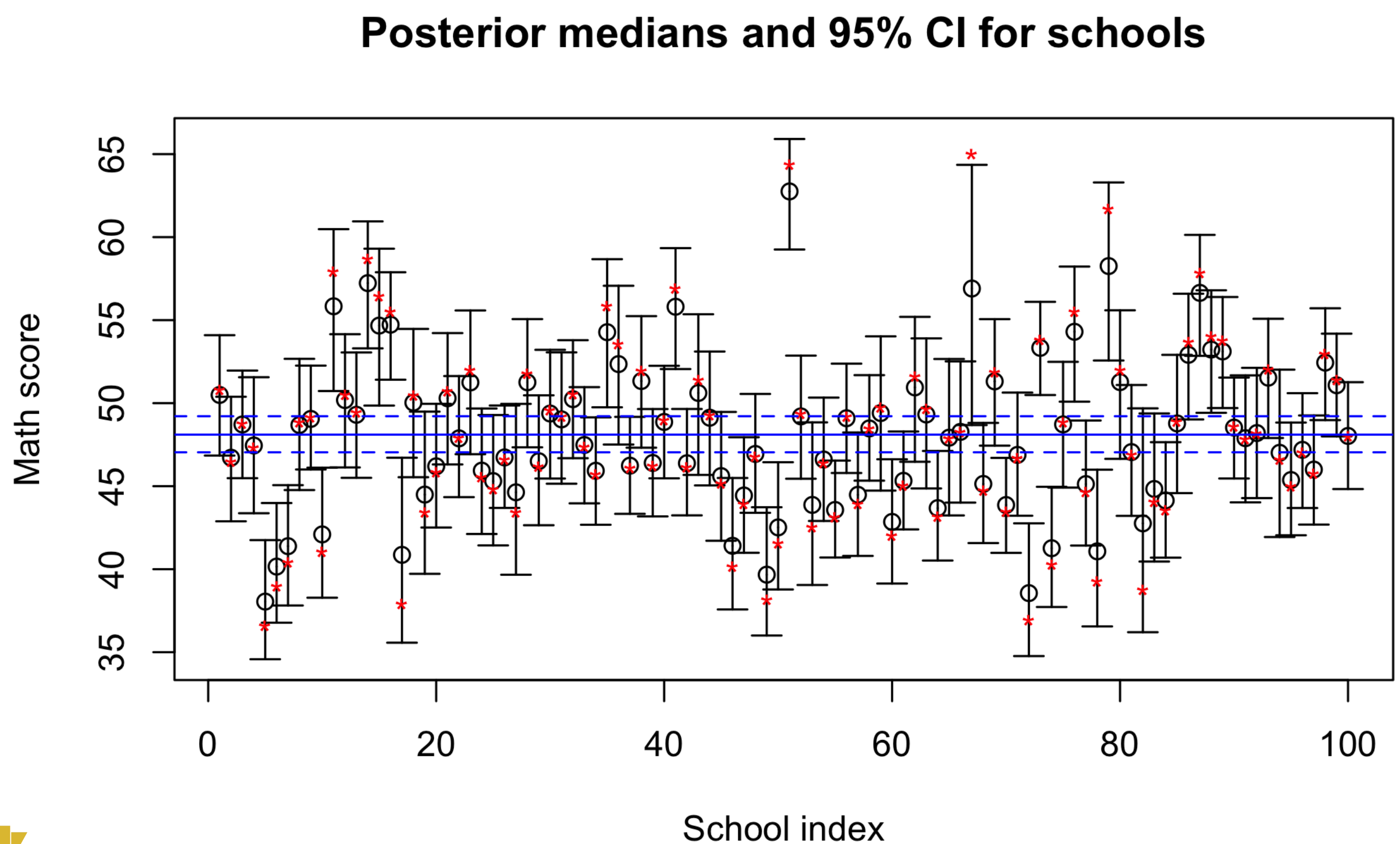
```r
#first set number of iterations and burn-in, then set seed
n_iter <- 10000; burn_in <- 0.3*n_iter
set.seed(1234)


#Set null matrices to save samples
SIGMA_SQ <- THETA <- matrix(nrow=n_iter, ncol=J)
OTHER_PAR <- matrix(nrow=n_iter, ncol=4)


#Now, to the Gibbs sampler
for(s in 1:(n_iter+burn_in)){

  #update the theta vector (all the theta_j's)
  tau_j_star <- 1/(n/sigma_sq + 1/tau_sq)
  mu_j_star <- tau_j_star*(ybar*n/sigma_sq + mu/tau_sq)
  theta <- rnorm(J,mu_j_star,sqrt(tau_j_star))

  #update the sigma_sq vector (all the sigma_sq_j's)
  nu_j_star <- nu_0 + n
  theta_long <- rep(theta,n)
  nu_j_star_sigma_j_sq_star <-
    nu_0*sigma_0_sq + c(by((Y[,"mathscore"] - theta_long)^2,Y[,"school"],sum))
  sigma_sq <- 1/rgamma(J,(nu_j_star/2),(nu_j_star_sigma_j_sq_star/2))
  #update mu
  gamma_n_sq <- 1/(J/tau_sq + 1/gamma_0_sq)
  mu_n <- gamma_n_sq*(J*mean(theta)/tau_sq + mu_0/gamma_0_sq)
  mu <- rnorm(1,mu_n,sqrt(gamma_n_sq))
```

Gibbs sampler

```
  #update tau_sq
  eta_n <- eta_0 + J
  eta_n_tau_n_sq <- eta_0*tau_0_sq + sum((theta-mu)^2)
  tau_sq <- 1/rgamma(1,eta_n/2,eta_n_tau_n_sq/2)

  #update sigma_0_sq
  sigma_0_sq <- rgamma(1,(a + J*nu_0/2),(b + nu_0*sum(1/sigma_sq)/2))

  #update nu_0
  log_prob_nu_0 <- (J*nu_0_grid/2)*log(nu_0_grid*sigma_0_sq/2) -
    J*lgamma(nu_0_grid/2) +
    (nu_0_grid/2-1)*sum(log(1/sigma_sq)) -
    nu_0_grid*(alpha + sigma_0_sq*sum(1/sigma_sq)/2)
  nu_0 <- sample(nu_0_grid,1, prob = exp(log_prob_nu_0 - max(log_prob_nu_0)) )
  #this last step substracts the maximum logarithm from all logs
  #it is a neat trick that throws away all results that are so negative
  #they will screw up the exponential
  #note that the sample function will renormalize the probabilities internally


  #save results only past burn-in
  if(s > burn_in){
    THETA[(s-burn_in),] <- theta
    SIGMA_SQ[(s-burn_in),] <- sigma_sq
    OTHER_PAR[(s-burn_in),] <- c(mu,tau_sq,sigma_0_sq,nu_0)
  }
}
colnames(OTHER_PAR) <- c("mu","tau_sq","sigma_0_sq","nu_0")
```

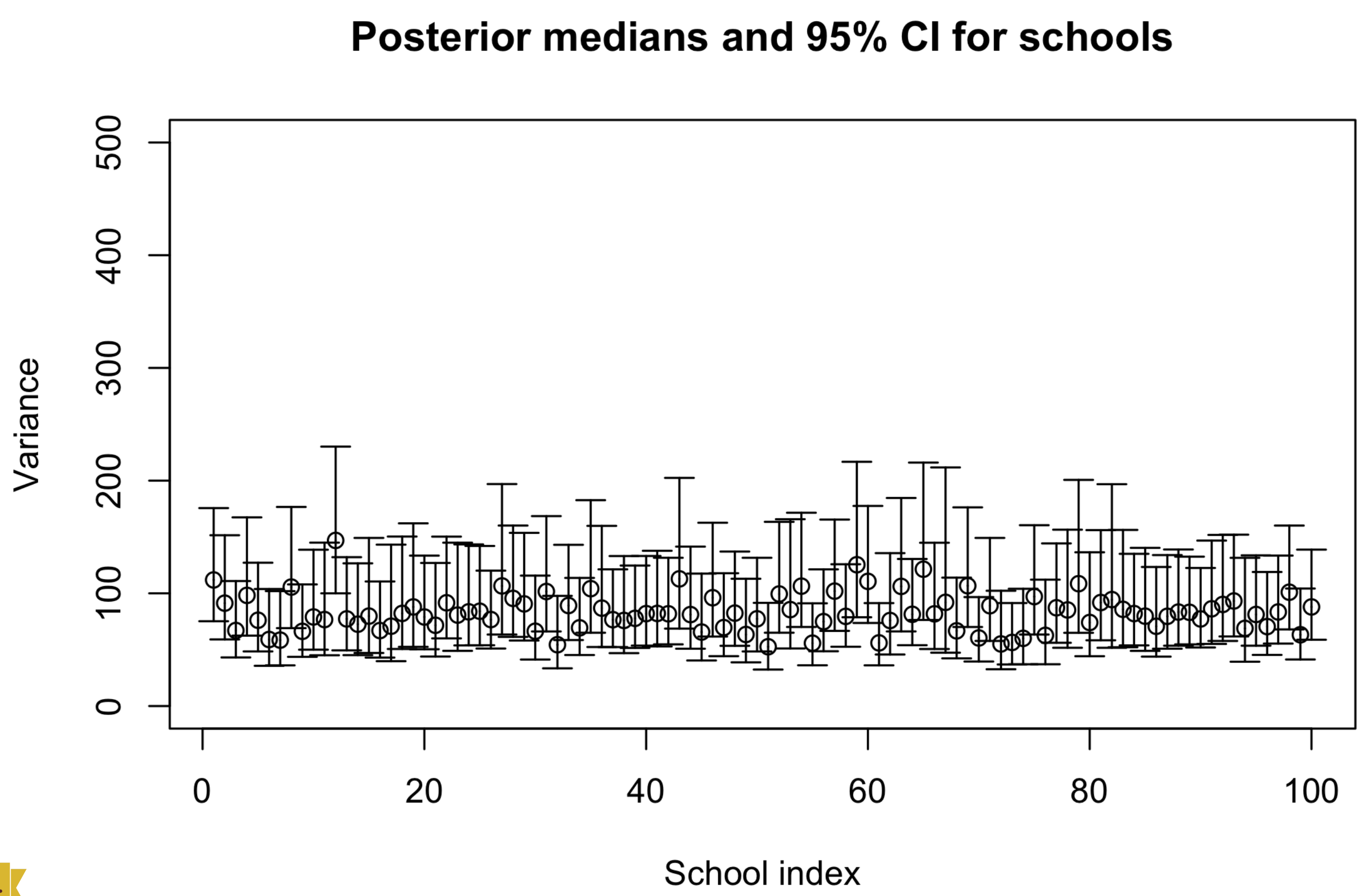
Posterior inference

The blue lines indicate the posterior median and a 95% for μ. The red asterisks indicate the data values $\bar{y}_j$.



Posterior inference

Posterior summaries of σ_j^2.



Posterior inference

Shrinkage as a function of sample size.

```
##    n Sample group mean Post. est. of group mean Post. est. of overall mean
## 1 31          50.81355                 50.49363                   48.10549
## 2 22          46.47955                 46.71544                   48.10549
## 3 23          48.77696                 48.71578                   48.10549
## 4 19          47.31632                 47.44935                   48.10549
## 5 21          36.58286                 38.04669                   48.10549
```

```
##     n Sample group mean Post. est. of group mean Post. est. of overall mean
## 15 12          56.43083                 54.67213                   48.10549
## 16 23          55.49609                 54.72904                   48.10549
## 17  7          37.92714                 40.86290                   48.10549
## 18 14          50.45357                 50.03007                   48.10549
```

```
##     n Sample group mean Post. est. of group mean Post. est. of overall mean
## 67  4          65.01750                 56.90436                   48.10549
## 68 19          44.74684                 45.13522                   48.10549
## 69 24          51.86917                 51.31079                   48.10549
## 70 27          43.47037                 43.86470                   48.10549
## 71 22          46.70455                 46.88374                   48.10549
## 72 13          36.95000                 38.55704                   48.10549
```

How about non-normal models?

- Suppose we have $y_{ij} \in \{0, 1, \ldots\}$ being a count for subject i in group j.
- For count data, it is natural to use a Poisson likelihood, that is,

  $$y_{ij} \sim \text{Poisson}(\theta_j)$$

  where each $\theta_j = \mathbb{E}[y_{ij}]$ is a group specific mean.
- When there are limited data within each group, it is natural to borrow information.
- How can we accomplish this with a hierarchical model?
- See homework 6 for a similar setup!