STA 360/602L: Module 5.5

Hierarchical normal modeling of means and variances (illustration)

Dr. Olanrewaju Michael Akande

STA 602L

ELS 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)
```

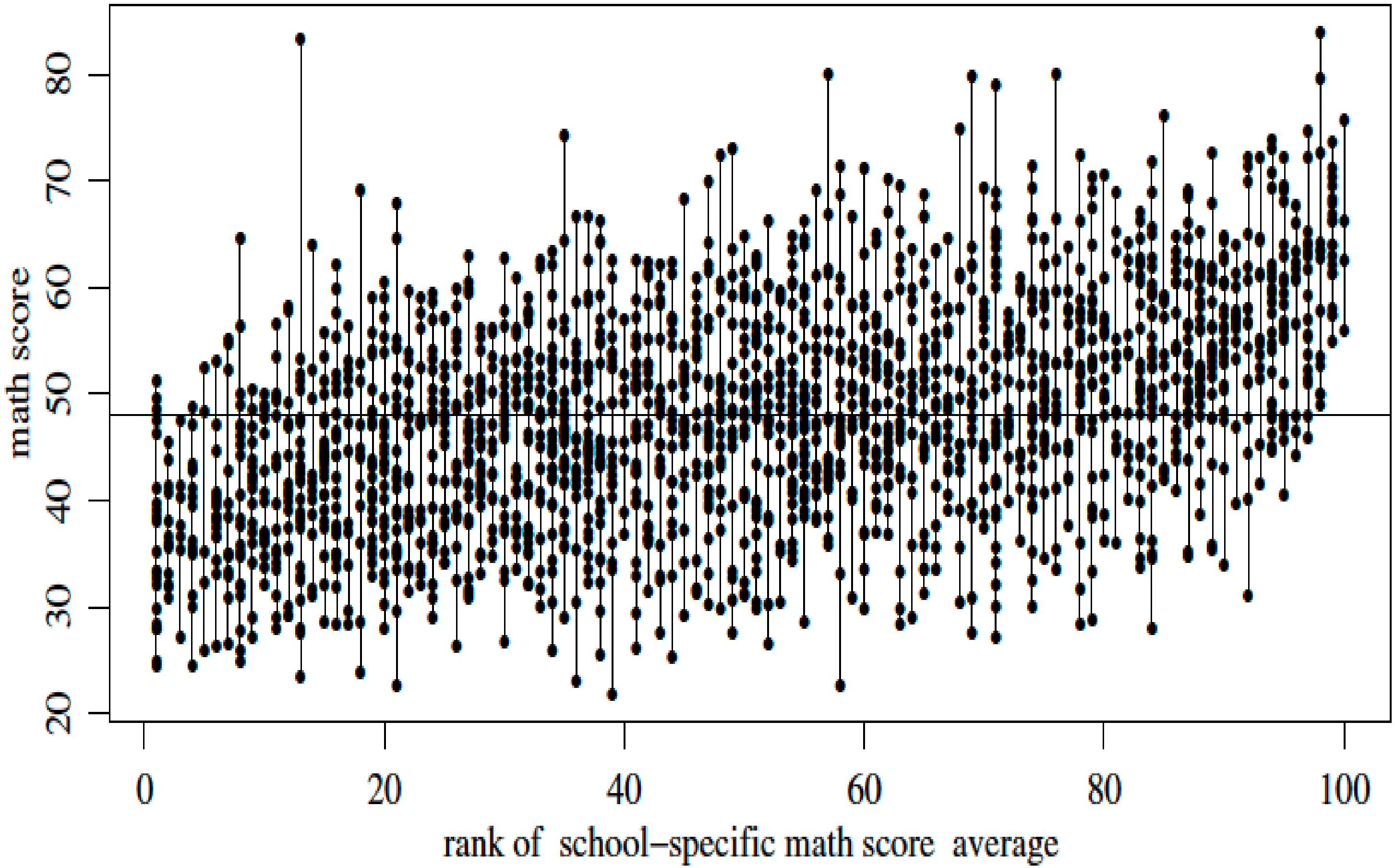
```
##      school mathscore
## [1,]      1     52.11
## [2,]      1     57.65
## [3,]      1     66.44
## [4,]      1     44.68
## [5,]      1     40.57
## [6,]      1     35.04
```

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

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

STA 602L

ELS DATA

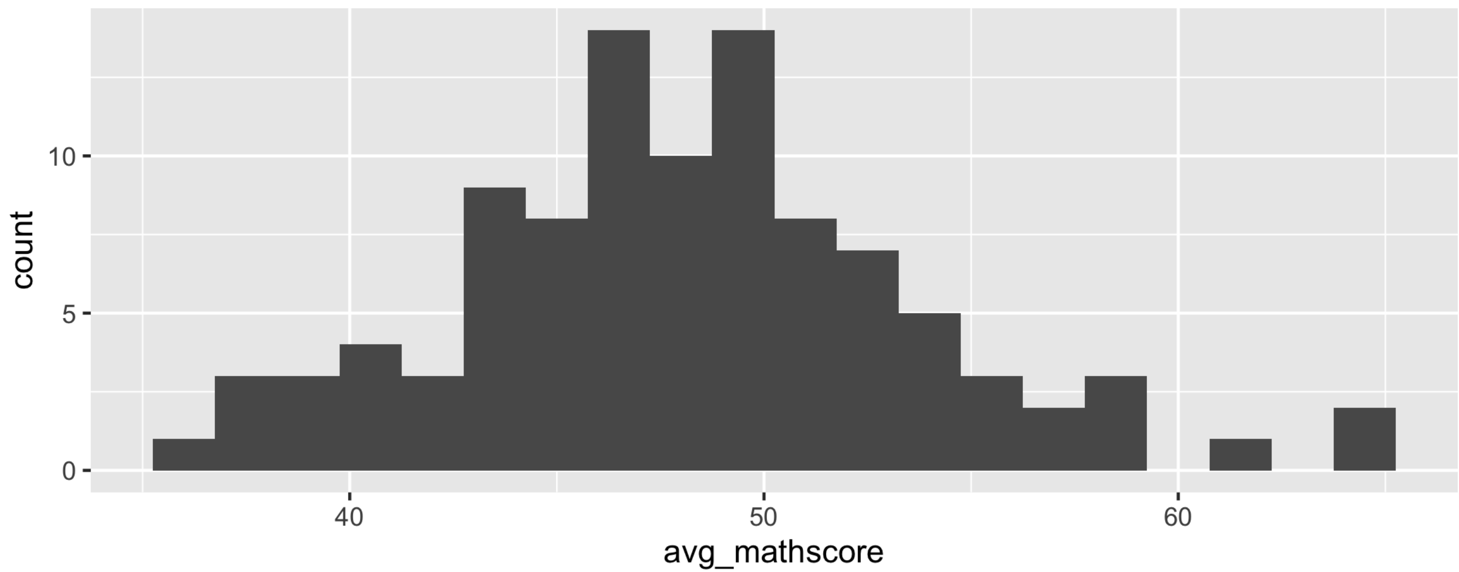

math score
20
30
40
50
60
70
80
0
20
40
60
80
100
rank of school–specific math score average

STA 602L

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(binwidth=1.5)
```

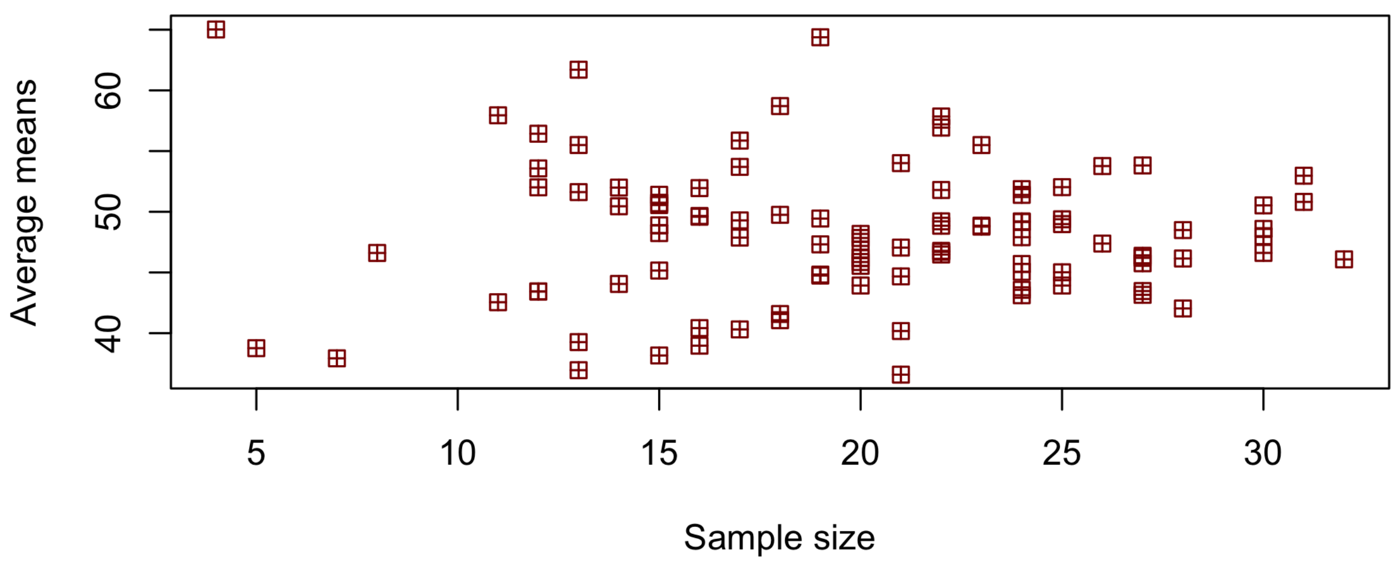


STA 602L

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.

```
plot(c(table(Data$school)),c(by(Data$mathscore,Data$school,mean)),
     ylab="Average means",xlab="Sample size",col="red4",pch=12)
```



STA 602L

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 model described in the previous module as:

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

$$\begin{aligned}
\theta_j|\mu, \tau^2 &\sim \mathcal{N}\left(\mu, \tau^2\right); \quad j = 1, \ldots, J \\
\sigma_1^2, \ldots, \sigma_J^2|\nu_0, \sigma_0^2 &\sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)
\end{aligned}$$

$$\begin{aligned}
\mu &\sim \mathcal{N}\left(\mu_0, \gamma_0^2\right) \\
\tau^2 &\sim \mathcal{IG}\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right).
\end{aligned}$$

$$\begin{aligned}
\pi(\nu_0) &\propto e^{-\alpha\nu_0} \\
\sigma_0^2 &\sim \mathcal{G}a\left(a, b\right).
\end{aligned}$$

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

STA 602L

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?

STA 602L

Full conditionals (recap)

- $$\pi(\theta_j|\cdots\cdots) = \mathcal{N}\left(\mu_j^\star, \tau_j^\star\right) \quad \text{where}$$

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

- $$\pi(\sigma_j^2|\cdots\cdots) = \mathcal{IG}\left(\frac{\nu_j^\star}{2}, \frac{\nu_j^\star \sigma_j^{2(\star)}}{2}\right) \quad \text{where}$$

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

- $$\pi(\mu|\cdots\cdots) = \mathcal{N}\left(\mu_n, \gamma_n^2\right) \quad \text{where}$$

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

STA 602L

Full conditionals (recap)

- $$\pi(\tau^2 | \cdots\cdots) = \mathcal{IG}\left(\frac{\eta_n}{2}, \frac{\eta_n \tau_n^2}{2}\right) \quad \text{where}$$

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

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

- $$\pi(\sigma_0^2 | \cdots\cdots) = \mathcal{G}a\left(\sigma_0^2; a_n, b_n\right) \quad \text{where}$$

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

STA 602L

Side note

- 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

```r
#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
```

STA 602L

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))
```

STA 602L

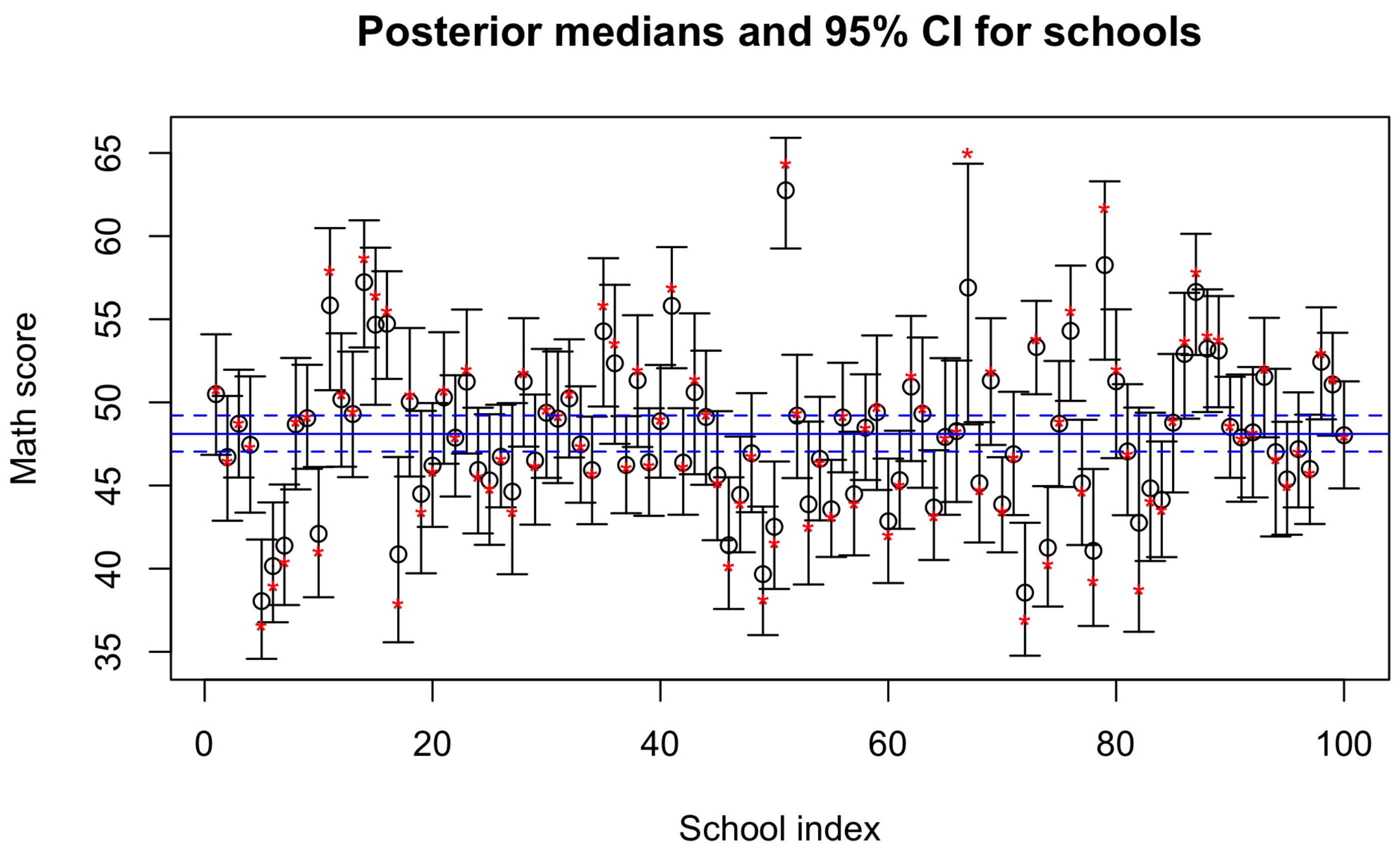
Gibbs sampler

```
  #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")
```

STA 602L

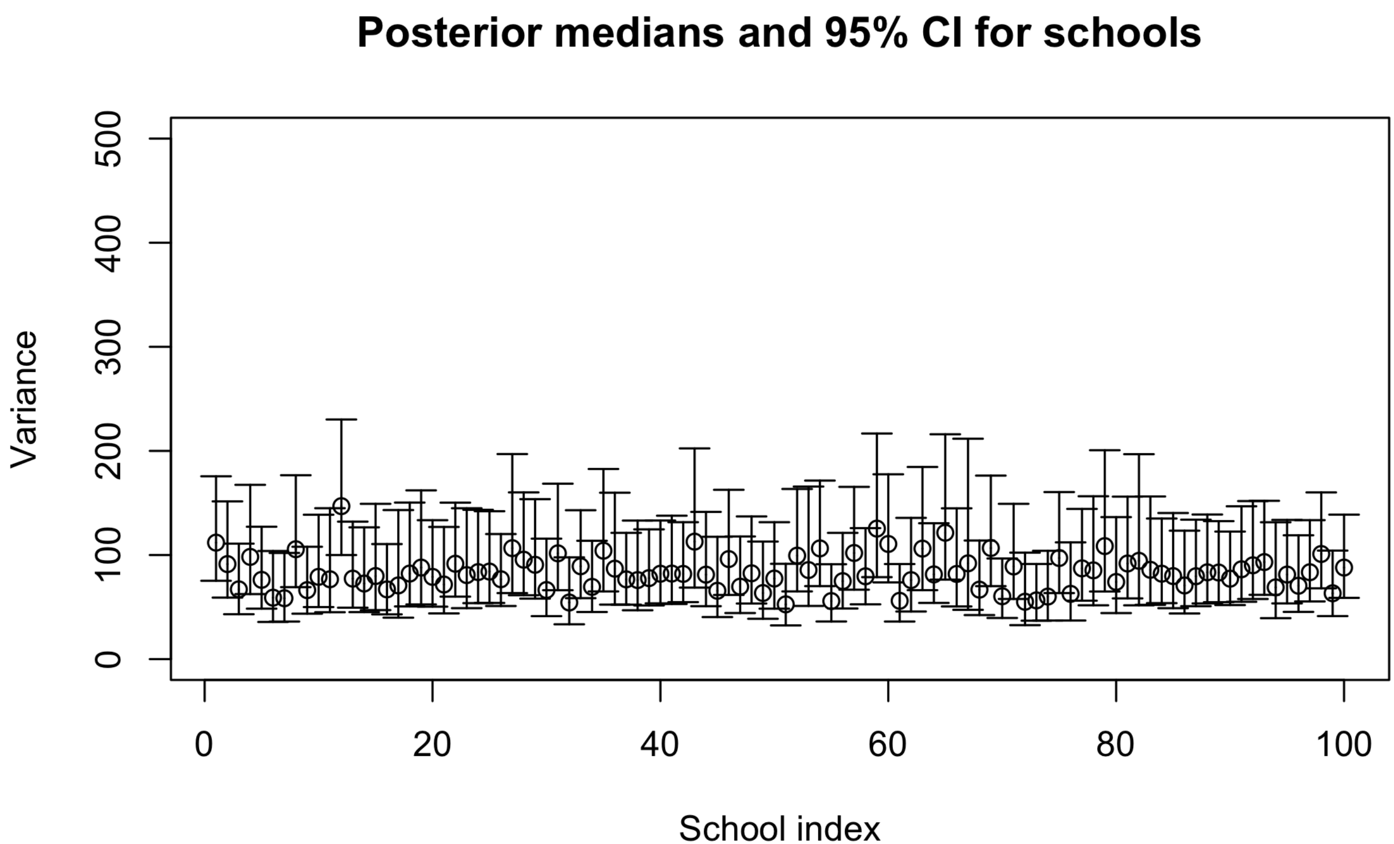
Posterior inference

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.



STA 602L

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
```

STA 602L

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?
- We can assume all the θ_j's come from the same distribution, then place priors on the parameters of the distribution.
- See homework for a similar setup!

STA 602L

What's next?

Move on to the readings for the next module!

STA 602L