## Hierarchical Modeling

This chapter focuses on comparison of means across groups and more generally Bayesian hierarchical modeling. Hierarchical modeling is defined by datasets with a multilevel structure, such as:

- 
- 

The most basic form of this type of data consists of two-levels, groups and individuals within groups.

Recall, observations are exchangeable if $p(y_1, \ldots, y_n) =$
Consider where $Y_1, \ldots, Y_n$ are test scores from randomly selected students from a given STAT 216 instructor/course. If exchangeability holds for these values, then:

$$\begin{aligned} \phi &\sim p(\phi), \\ Y_1, \ldots, Y_n|\phi &\sim \end{aligned}$$

The exchangeability can be interpreted that the random variables are independent samples from a population with a parameter, $\phi$. For instance in a normal model, $\phi = \{\theta, \sigma^2\}$ and the data are conditionally independent from a normal distribution $N(\theta, \sigma^2)$.

In a hierarchical framework this can be extended to include the group number:

$$Y_{1,j}, \ldots, Y_{n_j,j}|\phi_j \sim$$

The question now is how to we characterize the information between $\phi_1, \ldots, \phi_m$?

Is it reasonable to assume that the values are independent, that is does the information from $\phi_i$ tell you anything about $\phi_j$?

Now consider the groups as samples from a larger population, then using the idea of exchangeability with group-specific parameters gives:

$$\phi_1, \ldots, \phi_m | \psi \sim$$

This is similar to the idea of a random effect model and gives the following hierarchical probability model:

$$\begin{aligned} y_{1,j}, \ldots, y_{n_j,j} | \phi_j &\sim p(y|\phi_j) \\ \phi_1, \ldots, \phi_m | \psi &\sim p(\phi|\psi) \\ \psi &\sim p(\psi) \end{aligned}$$

The distributions $p(y|\phi)$ and $p(\phi|\psi)$ represent sampling variability:

- $p(y|\phi)$ represents

- $p(\phi|\psi)$ represents

**Hierarchical normal model**

The hierarchical normal model is often used for modeling differing means across a population.

$$\begin{aligned} \phi_j &= \{\theta_j, \sigma^2\}, \\ \psi = \{\mu, \tau\}, p(\theta_j|\psi) &= \end{aligned}$$

Note this model specification assumes constant variance for each within-group model, but this assumption can be relaxed.

This model contains three unknown parameters that need priors, we will use the standard semi-conjugate forms:

$$\begin{aligned} \sigma^2 &\sim \\ \tau^2 &\sim \\ \mu &\sim \end{aligned}$$

Given these priors, we need to derive the full conditional distributions in order to make draws from the posterior distribution. Note the joint posterior distribution, can be expressed as:

$$\begin{aligned} p(\tilde{\theta}, \mu, \tau^2, \sigma^2 | \tilde{y}_1, \ldots, \tilde{y}_n) &\propto p(\mu, \tau^2, \sigma^2) \times p(\tilde{\theta} | \mu, \tau^2, \sigma^2) \times p(\tilde{y}_1, \ldots, \tilde{y}_m | \tilde{\theta}, \mu, \tau^2, \sigma^2) \\ &\propto p(\mu) p(\sigma^2) p(\tau^2) \times \end{aligned}$$

.

- **Sampling** $\mu$: $p(\mu|-) \propto p(\mu) \prod_{j=1}^{m} p(\theta_j | \mu, \tau^2)$. This is a familiar setting with two normal models, hence, the posterior is also a normal distribution.
  - $\mu|- \sim$

- **Sampling** $\tau^2$: $p(\tau^2|-) \propto p(\tau^2) \prod_{j=1}^{m} p(\theta_j | \mu, \tau^2)$. Again this is similar to what we have seen before.
  - $\tau^2|- \sim$

Now what about $\theta_1, \ldots, \theta_m$?

- **Sampling** $\theta_1, \ldots, \theta_m$. Consider a single $\theta_j$, then $\theta_j|- \propto$
  - $\theta_j|- \sim$

- **Sampling** $\sigma^2$: $p(\sigma^2|-) \propto p(\sigma^2) \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma)$.
  - $\sigma^2|- \sim$ .

## Data Example

Consider the dataset outline in Chapter 8, that focuses on math tests scores for students spread across 100 schools. Using the Gibbs sampling procedure described above we can fit this model, code courtesy of textbook.

```
> Y.school.mathscore<-dget("http://www.stat.washington.edu/~hoff/Book/Data/data/Y.school.mathscore")
>
> Y <- Y.school.mathscore
> 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
> ### weakly informative priors
> nu.0<-1
> sigmasq.0<-100
> eta.0<-1
> tausq.0<-100
> mu.0<-50
> gammasq.0<-25
> ###
>
> ### starting values
> m <- length(unique(Y[,1])) # number of schools
> n<-sv<-ybar<-rep(NA,m)
> for(j in 1:m)
+ {
+   ybar[j]<-mean(Y[Y[,1]==j,2])
+   sv[j]<-var(Y[Y[,1]==j,2])
+   n[j]<-sum(Y[,1] ==j)
+ }
> theta<-ybar
> sigma2<-mean(sv)
> mu<-mean(theta)
> tau2<-var(theta)
> ###
>
> ### setup MCMC
> set.seed(1)
> S<-5000
> THETA<-matrix( nrow=S,ncol=m)
> MST<-matrix( nrow=S,ncol=3)
> ###
>
```

```
> ### MCMC algorithm
> for(s in 1:S)
+ {
+
+   # sample new values of the thetas
+   for(j in 1:m)
+   {
+     vtheta<-1/(n[j]/sigma2+1/tau2)
+     etheta<-vtheta*(ybar[j]*n[j]/sigma2+mu/tau2)
+     theta[j]<-rnorm(1,etheta,sqrt(vtheta))
+   }
+
+   #sample new value of sigma2
+   nun<-nu0+sum(n)
+   ss<-nu0*sigmasq.0;
+   for(j in 1:m){
+     ss<-ss+sum((Y[[j]]-theta[j])^2)
+   }
+   sigma2<-1/rgamma(1,nun/2,ss/2)
+
+   #sample a new value of mu
+   vmu<- 1/(m/tau2+1/gammasq.0)
+   emu<- vmu*(m*mean(theta)/tau2 + mu.0/gammasq.0)
+   mu<-rnorm(1,emu,sqrt(vmu))
+
+   # sample a new value of tau2
+   etam<-eta.0+m
+   ss<- eta.0*tausq.0 + sum( (theta-mu)^2 )
+   tau2<-1/rgamma(1,etam/2,ss/2)
+
+   #store results
+   THETA[s,]<-theta
+   MST[s,]<-c(mu,sigma2,tau2)
+ }
```

Now consider the following plot that contains the posterior distribution for school 46 and school 48, along with the data points for each plotted along the bottom. Note the large circle represents the sample mean for each school. Comment on the differences between the sample means and the means of the posterior distributions. Why does this happen and is it a good thing?

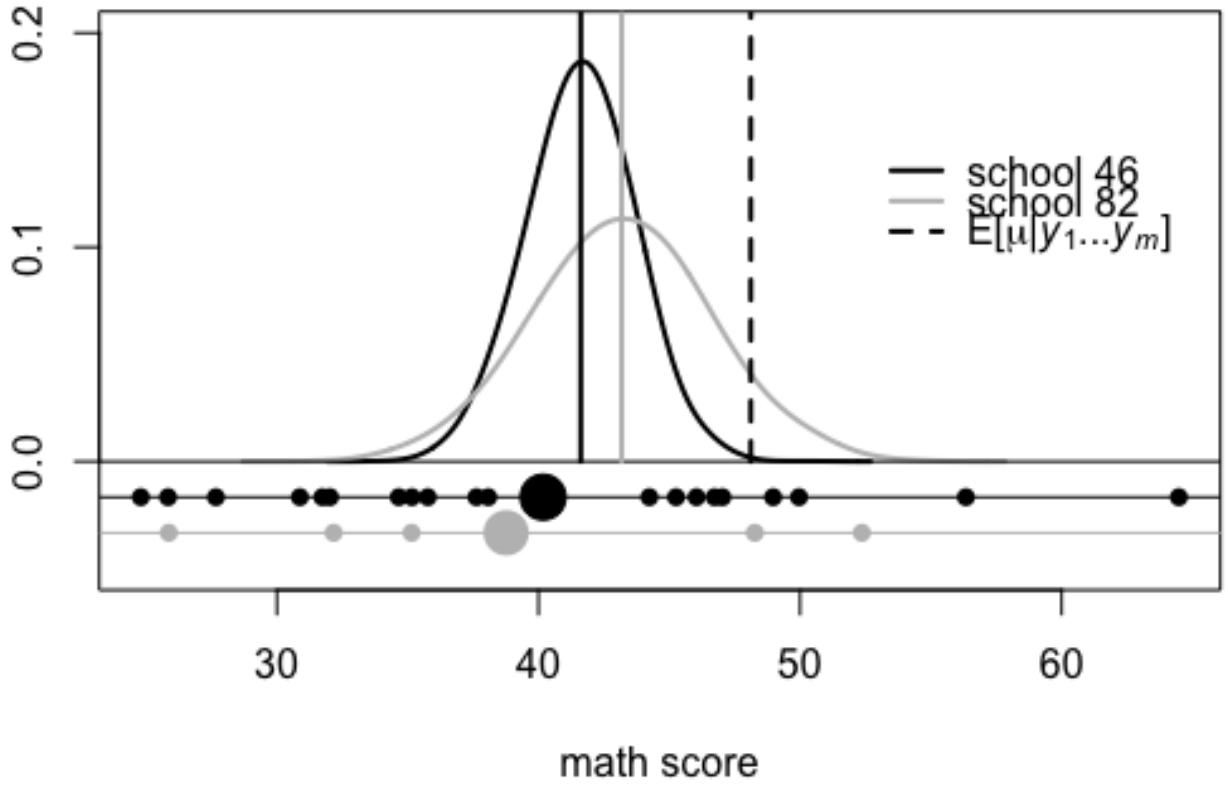


Figure 1: Posterior distributions and data points for schools 46 and 82.

## Shrinkage

Recall the posterior mean can be represented as a weighted average, specifically in this case:

$$E[\theta_j|\tilde{y}_j, \mu, \tau^2, \sigma^2] = \tag{1}$$

In this case $\mu$ and $\tau^2$ are not chosen parameters from prior distributions, but rather they come from the between group model. So the posterior means for test scores at each school are pulled from the sample mean toward the overall group mean across all of the schools. This phenemenon is known as

Schools with more students taking the exam see less shrinkage, as there is more weight on the data given more observations. So the figure we discussed before, shows more shrinkage for school 82 as there were fewer observations.

So what about shrinkage, does it make sense? Is it a good thing?

We will soon see that it is an extremely powerful tool and actually dominates the unbiased estimator (the MLE for each distribution) in some settings. This suprising result is commonly known as Stein's Paradox.

### Hierarchical Modeling of Means and Variances

The model we just described and fit was somewhat restrictive in that each school was known to have a common variance. It is likely that schools with a more heterogenous mix of students would have greater variance in the test scores. There are a couple of solutions, the first involves a set of i.i.d. priors on each $\sigma_j^2$

$$\sigma_1^2, \ldots, \sigma_m^2 \sim \text{i.i.d.}$$

however, this results in a full conditional distribution for $\sigma_j$ that only takes advantage of data from school $j$. In other words no information from the other schools is used to estimate that variance.

Another option is to consider $\nu_0$ and $\sigma_0^2$ as parameters to be estimated in the hierarchical model. A common prior for $\sigma_0^2$ would be $p(\sigma_0^2) \sim Gamma(a, b)$. Unfortunately, there is not a semi-conjugate prior distribution for $\nu_0$. The textbook suggests a geometric distribution, where $p(\nu_0) \propto \exp(-\alpha\nu_0)$. Then the full conditional distribution allows a sampling procedure that enumerates over the domain of possible values. This procedure allows shrinkage for the variance terms as well. It is worth noting, that pooling variances is a common way to tackle this particular problem.