Part 7:

Hierarchical Modeling

Nested data

It is common for data to be nested: i.e., observations on subjects are organized by a hierarchy

Such data are often called hierarchical or multilevel

For example,

- patients within several hospitals
- genes within a group of animals, or
- people within counties within regions within countries

Two groups

The simplest type of multilevel data has 2 levels, in which

- one level consists of groups
- and the other consists of units within groups

In this case, we denote $y_{i,j}$ as the data on the i^{th} unit within group j

Hierarchical model

The sampling model should reflect/acknowledge the hierarchy so that we may distinguish between

- within-group variability, and
- between-group variability

One typically uses the following hierarchical model, for $j = 1, \ldots, m$, with n_j observations in each group

$$\{Y_{1,j}, \ldots, Y_{n_j,j} | \theta_j\} \overset{\text{iid}}{\sim} p(Y|\theta_j) \quad \text{(within-group sampling variability)}$$

$$\{\theta_1, \ldots, \theta_m | \phi\} \overset{\text{iid}}{\sim} p(\theta_j|\phi) \quad \text{(between-group sampling variability)}$$

$$\phi \sim p(\phi) \quad \text{(prior distribution)}$$

Variability accounting

It is important to recognize that the distributions $p(y|\theta)$ and $p(\theta|\phi)$ both represent sampling variability among populations of objects:

- $p(y|\theta)$ represents variability among measurements within a group
- $p(\theta|\phi)$ represents variability across groups

In contrast, $p(\phi)$ represents information about a single fixed but unknown quantity

These are both sampling distributions; the data are used to estimate θ and ϕ; but $p(\phi)$ is not estimated

Hierarchical normal model

A popular model for describing the heterogeneity of means across several populations is the hierarchical normal model, in which the within- and between-group sampling models are both normal:

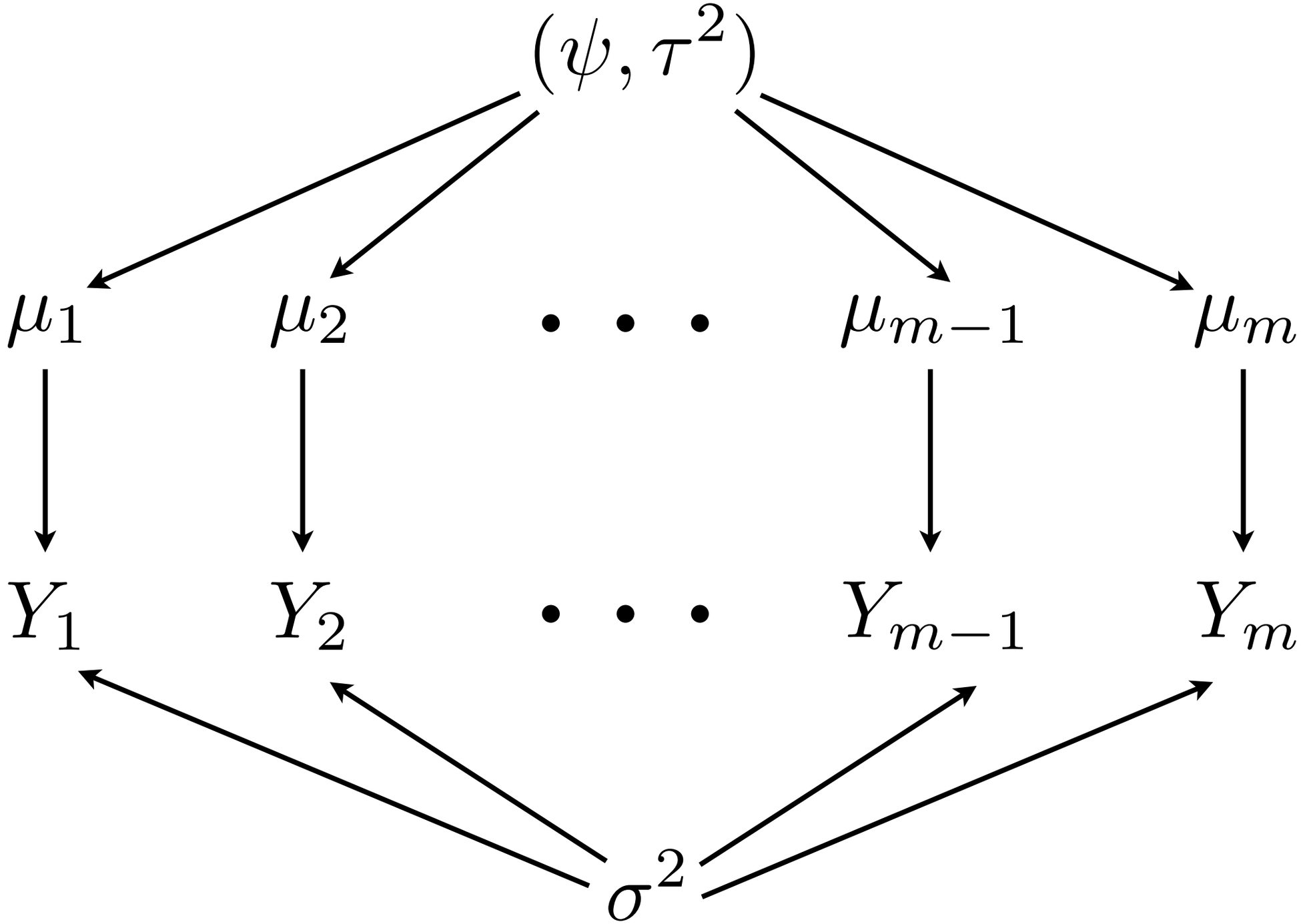
$$\theta_j = (\mu_j, \sigma^2),\ p(y|\theta_j) = \mathcal{N}(\mu_j, \sigma^2) \quad \text{(within-group model)}$$

$$\phi = (\psi, \tau^2),\ p(\theta_j|\phi) = \mathcal{N}(\psi, \tau^2) \quad \text{(between-group model)}$$

Note that $p(\theta|\phi)$ only describes heterogeneity across group means, and not any heterogeneity in group-specific variances

The within-group sampling variability σ^2 is assumed to be constant across groups

Hierarchical diagram



The priors

The fixed but unknown parameters in the models are ψ, τ^2 and σ^2

For convenience we will use the standard semi-conjugate normal and IG prior for these parameters

$$\sigma^2 \sim \mathrm{IG}(\nu_0/2, \nu_0\sigma_0^2/2)$$
$$\tau^2 \sim \mathrm{IG}(\eta_0/2, \eta_0\tau_0^2/2)$$
$$\psi \sim \mathcal{N}(\psi_0, \gamma_0^2)$$

Posterior inference

The full set of unknown quantities in our system include the group-specific means $\{\mu_1, \ldots, \mu_m\}$, the within-group sampling variability σ^2 and the mean and variance (ψ, τ^2) of the population group-specific means

Posterior inference for these parameters can be made by GS which approximates the joint posterior distribution

$$p(\mu_1, \ldots, \mu_m, \psi, \tau^2, \sigma^2 | y_1, \ldots, y_m)$$

GS proceeds by iteratively sampling each parameter from its full conditional distribution

Full conditionals

- Deriving the full conditional distributions in this highly parameterized system may seem like a daunting task
- But it turns out that we have already worked out all of the necessary technical details
- All that is required is that we recognize certain analogies between the current model and the univariate normal model

Posterior essence

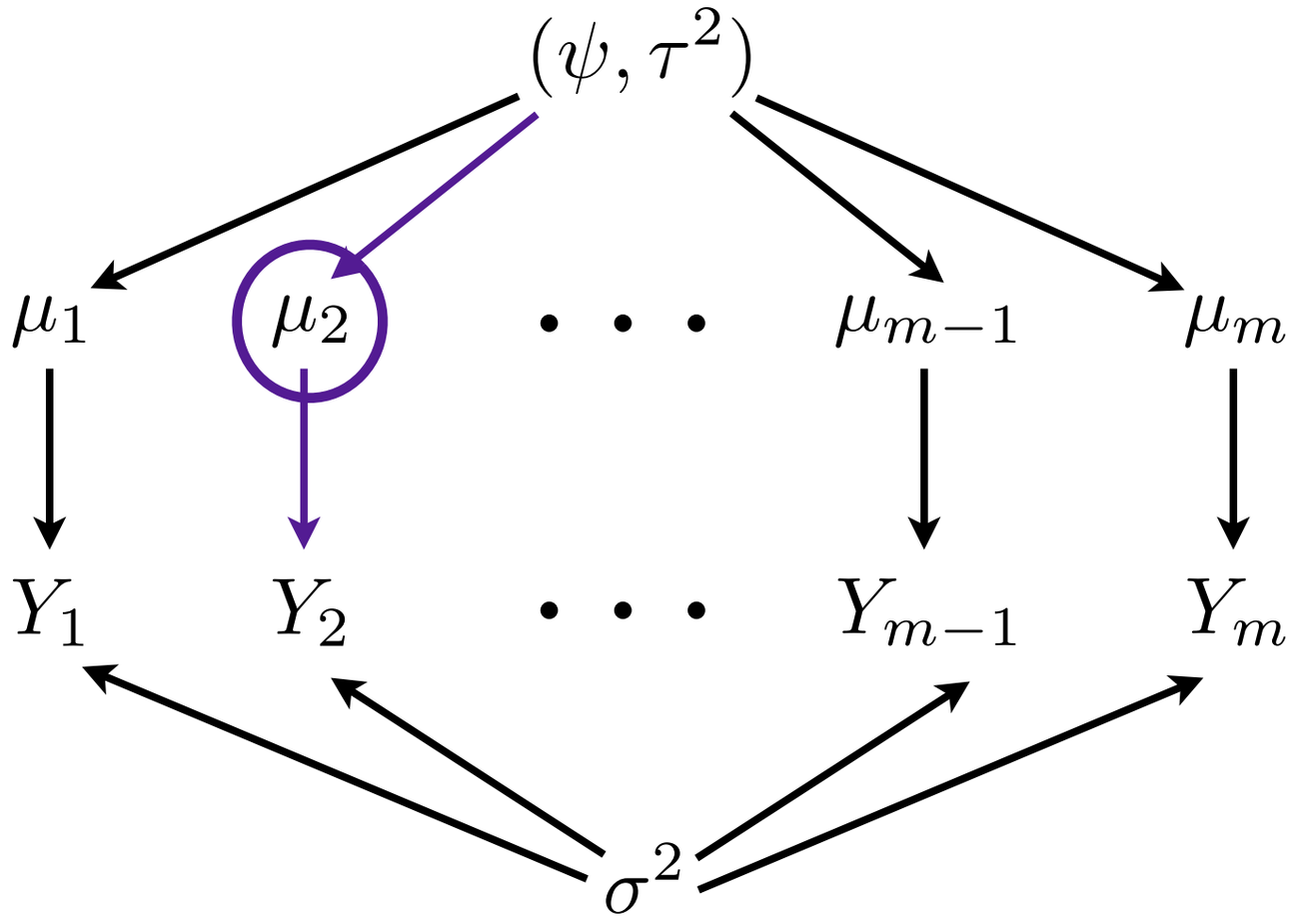
The following factorization of the posterior distribution will be useful

$$p(\mu_1, \ldots, \mu_m, \psi, \tau^2, \sigma^2 | y_1, \ldots, y_m)$$
$$\propto p(y_1, \ldots, y_m | \mu_1, \ldots, \mu_n, \sigma^2, \psi, \tau^2) \times p(\mu_1, \ldots, \mu_m | \psi, \tau^2) \times p(\psi, \tau^2, \sigma^2)$$
$$= \left\{ \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{i,j} | \mu_j, \sigma^2) \right\} \times \left\{ \prod_{j=1}^{m} p(\mu_j | \psi, \tau^2) \right\} \times p(\psi) p(\tau^2) p(\sigma^2)$$

This term is the result of an important conditional independence feature of our model

This relates back to our diagram ...

Conditional independence



The existence of a path from (ψ, τ^2) to each Y_j indicates that these parameters provide information about Y_j

but only indirectly through μ_j

Conditional on $\{\mu_1, \ldots, \mu_m, \psi, \tau^2, \sigma^2\}$ the random variables $Y_{1,j}, \ldots, Y_{n_j,j}$ are independent with a distribution that depends only on μ_j and σ^2

Full conditional for (ψ, τ^2)

As a function of ψ and τ^2, the posterior distribution is proportional to

$$\prod_{j=1}^{m} p(\mu_j|\psi, \tau^2)p(\psi)p(\tau^2)$$

And so the full conditional distributions of ψ and τ^2 are also proportional to this quantity

$$p(\psi|\mu_1, \ldots, \mu_m, \tau^2, \sigma^2, y_1, \ldots, y_m) = \prod p(\mu_j|\psi, \tau^2)p(\psi)$$

$$p(\tau^2|\mu_1, \ldots, \mu_m, \psi, \sigma^2, y_1, \ldots, y_m) = \prod p(\mu_j|\psi, \tau^2)p(\tau^2)$$

Full conditional for (ψ, τ^2)

These conditionals are exactly the full conditional distributions from the one-sample normal problem

Therefore, by analogy:

$$\{\psi|\mu_1,\ldots,\mu_m,\tau^2\} \sim \mathcal{N}\left(\frac{m\bar{\mu}/\tau^2+\psi_0/\gamma_0^2}{m/\tau^2+1/\gamma_0^2}, [m/\tau^2+1/\gamma_0^2]^{-1}\right)$$

$$\{\tau^2|\mu_1,\ldots,\mu_m,\psi\} \sim \text{IG}\left(\frac{\eta_0+m}{2}, \frac{\eta_0\tau_0^2+\sum(\mu_j-\psi)^2}{2}\right)$$

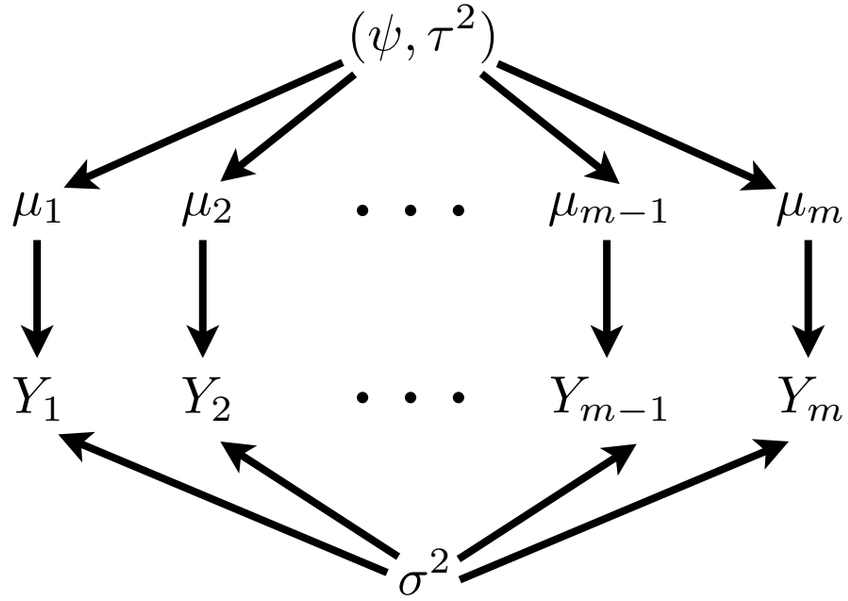
Full conditional for μ_j

Likewise, the full conditional distribution of μ_j must be proportional to

$$p(\mu_j|\psi, \tau^2, \sigma^2, \mu_{(-j)}, y_1, \ldots, y_n)$$

$$\propto \prod_{i=1}^{n_j} p(y_{i,j}|\mu_j, \sigma^2) \times p(\mu_j|\psi, \tau^2)$$

μ_j is conditionally independent of $\{\mu_k, y_k\}$ for $k \neq j$



While there is a path from each μ_j to every other μ_k, the paths go through (ψ, τ^2) or σ^2

Full conditional for μ_j

We can think of this as meaning that the μ's contribute no information about each other beyond that contained in ψ, τ^2 and σ^2

The terms in our conditional are

$$\prod_{i=1}^{n_j} p(y_{i,j}|\mu_j, \sigma^2) \times p(\mu_j|\psi, \tau^2)$$

(product of normals) (normal)

Mathematically, this is the same setup as our normal model. So by analogy, the full conditional distribution is

$$\{\mu_j|\sigma^2, \psi, \tau^2, y_{1,j}, \ldots, y_{n_j,j}\} \sim \mathcal{N}\left(\frac{n_j\bar{y}_j/\sigma^2 + \psi/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, [n_j/\sigma^2 + 1/\tau^2]^{-1}\right)$$

Full conditional for σ^2

By similar arguments, σ^2 is conditionally independent of $\{\psi, \tau^2\}$ given $\{y_1, \ldots, y_m, \mu_1, \ldots, \mu_m\}$

The derivation of the full conditional of σ^2 is similar to that in the one-sample normal model, except that now we have information about σ^2 from m separate groups

$$p(\sigma^2|\mu_1, \ldots, \mu_m, y_1, \ldots, y_m)$$

$$\propto \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{i,j}|\mu_j, \sigma^2) \times p(\sigma^2)$$

$$\propto (\sigma^2)^{\sum n_j/2} e^{-\frac{\sum\sum(y_{i,j}-\mu_j)^2}{2\sigma^2}} \times (\sigma^2)^{-\nu_0/2+1} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}}$$

Full conditional for σ^2

Adding powers of σ^2 and collecting terms in the exponent, we recognize that this is an IG:

$$\{\sigma^2|\mu, y\} \sim \text{IG}\left(\frac{1}{2}\left[\nu_0 + \sum_{j=1}^{m} n_j\right], \frac{1}{2}\left[\nu_0\sigma_0^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j} - \mu_j)^2\right]\right)$$

Note that $\sum\sum(y_{i,j} - \mu_j)^2$ is the sum of squared residuals across all groups, conditional on the within-group means

So the conditional distribution concentrates probability around a pooled-sample estimate of the variance

Example: Math scores in US public schools

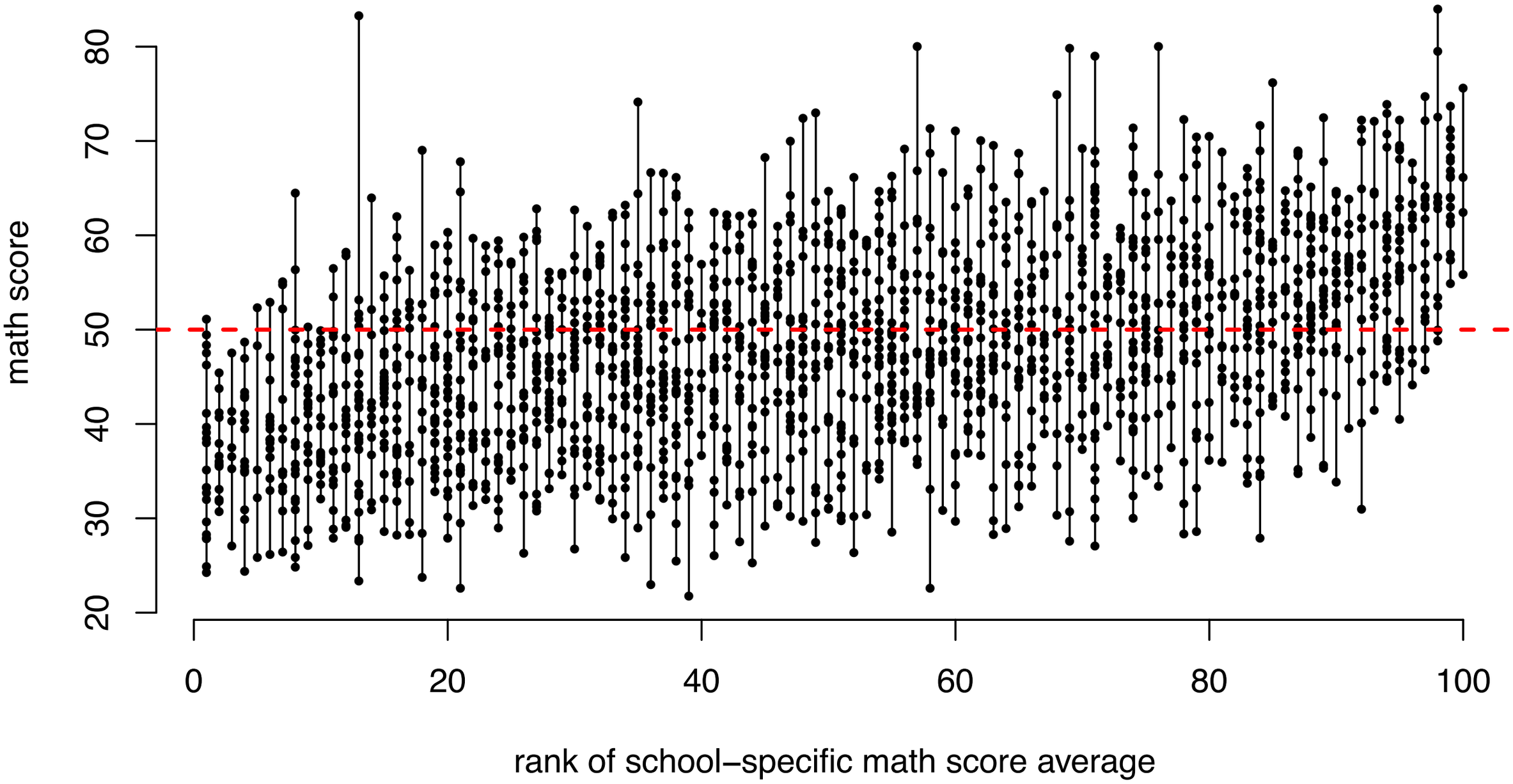
Consider data that is part of the 2002 Educational Longitudinal Study (ELS), a survey of students from a large sample of schools in the United States

The data consist of math scores of 10th grade students at 100 different urban public high schools with a (10th grade) enrollment of 400+

The scores are based on a national exam, standardized to produce a nationwide mean of 50 and standard deviation of 10

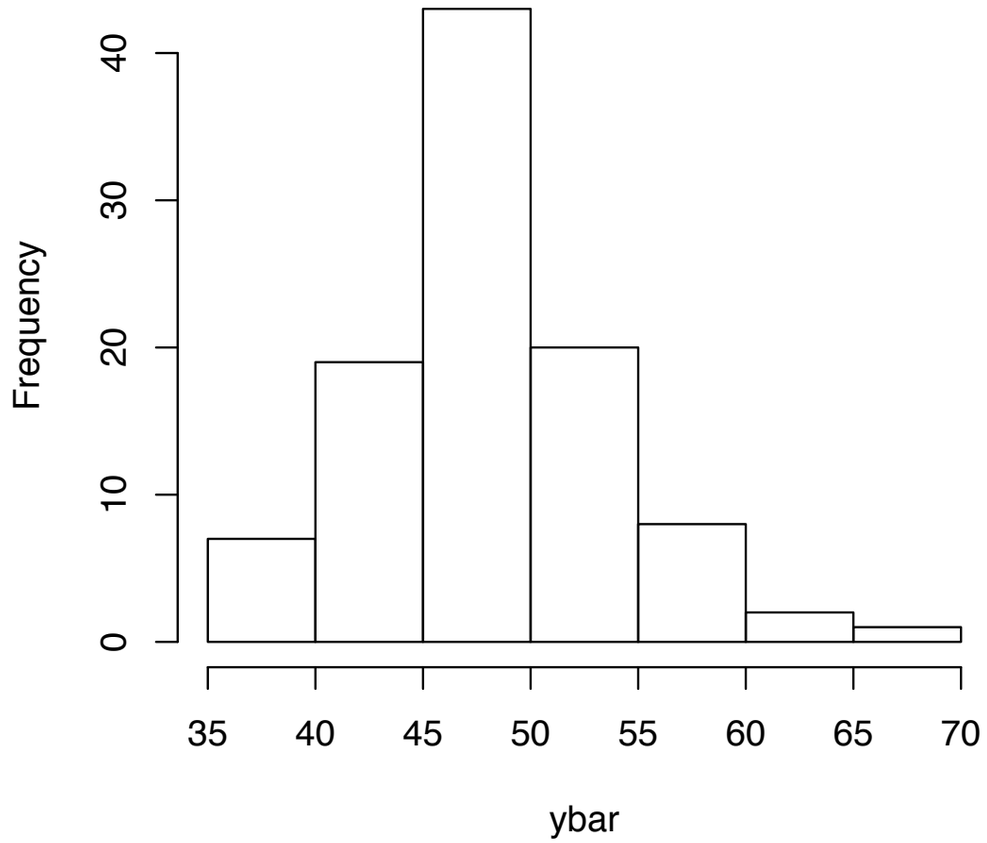
Example: the data

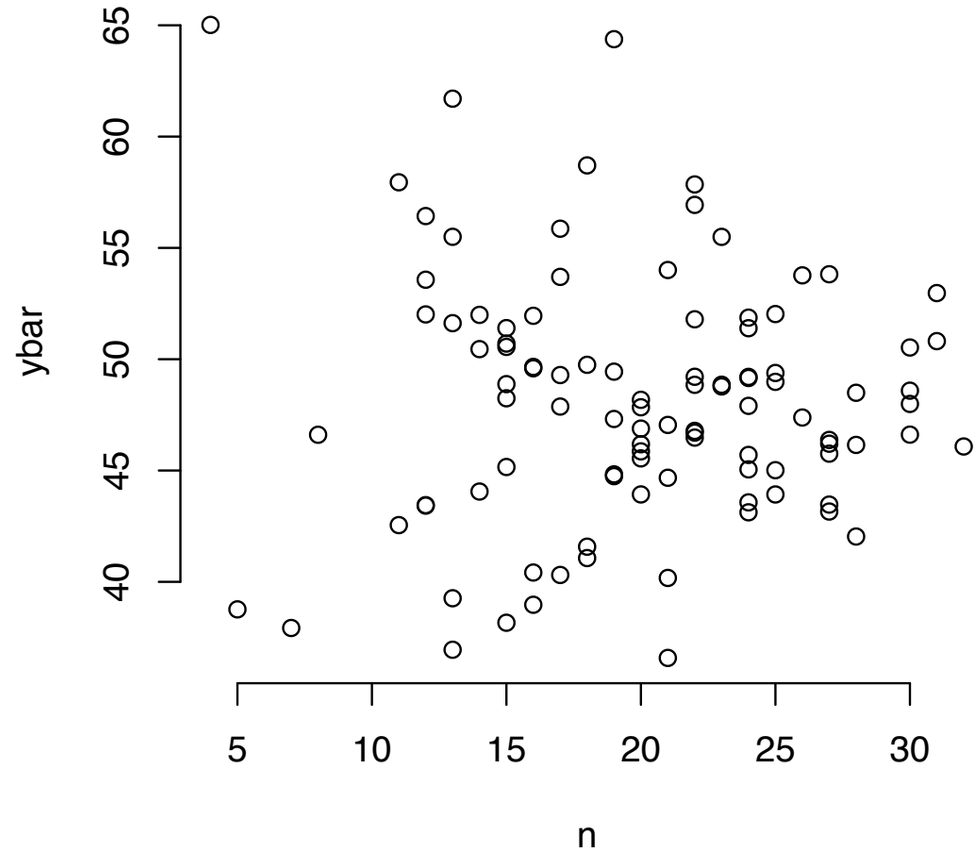
Scores for students within the same school plotted along a common vertical bar:



Example: the data

The range of average scores (36, 65) is quite large





Extreme sample averages occur for schools with small sample sizes

This is a common relationship in hierarchical datasets

Example: priors

The prior parameters that we need to specify are

$$(\nu_0, \sigma_0^2) \text{ for } p(\sigma^2)$$
$$(\eta_0, \tau_0^2) \text{ for } p(\tau^2), \text{ and}$$
$$(\psi_0, \gamma_0^2) \text{ for } p(\psi)$$

The exam was designed to give a variance of 100, both within and between schools

- so the within-school variance should be at most 100, which we take as σ_0^2
- this is likely an overestimate, so we take a weak prior concentration around this value with $\nu_0 = 1$

Example: priors, ctd.

Similarly, the between-school variance should not be more than 100, so we likewise take $(\eta_0, \tau_0^2) = (1, 100)$

Finally, the nationwide mean over all schools is 50

Although the mean for large urban public schools may be different than the nationwide average, it should not differ by too much

We shall take $\psi_0 = 50$ and $\gamma_0^2 = 25$, so that the prior probability that $\psi \in (40, 60)$ is about 95%

Example: Gibbs sampling

Posterior approximation may now proceed by GS

Given a current state of the unknowns

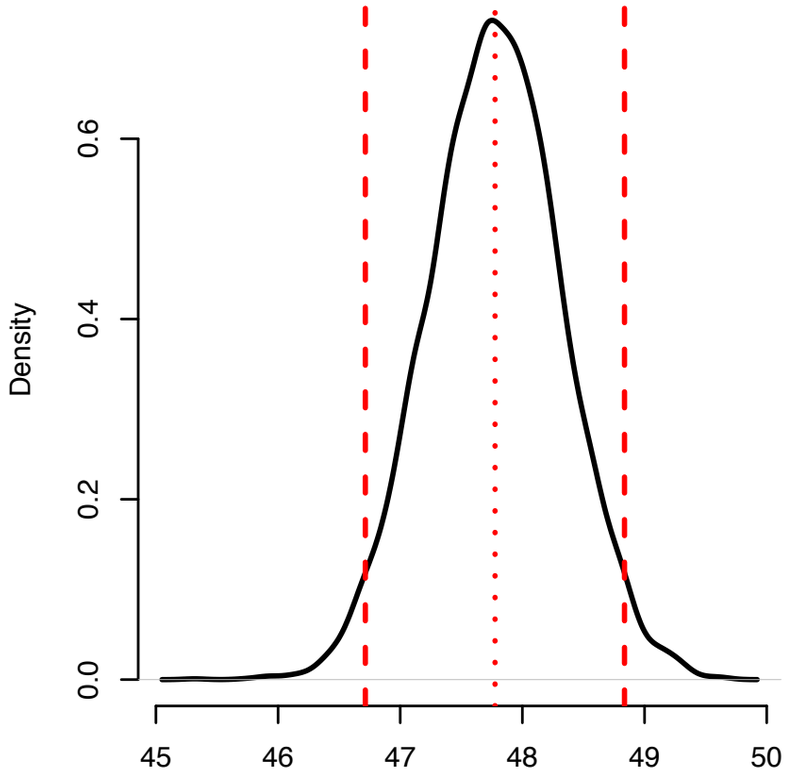
$$\{\mu_1^{(s)}, \ldots, \mu_m^{(s)}, \psi^{(s)}, \tau^{2(s)}, \sigma^{2(s)}\}$$

a new state may be generated as follows

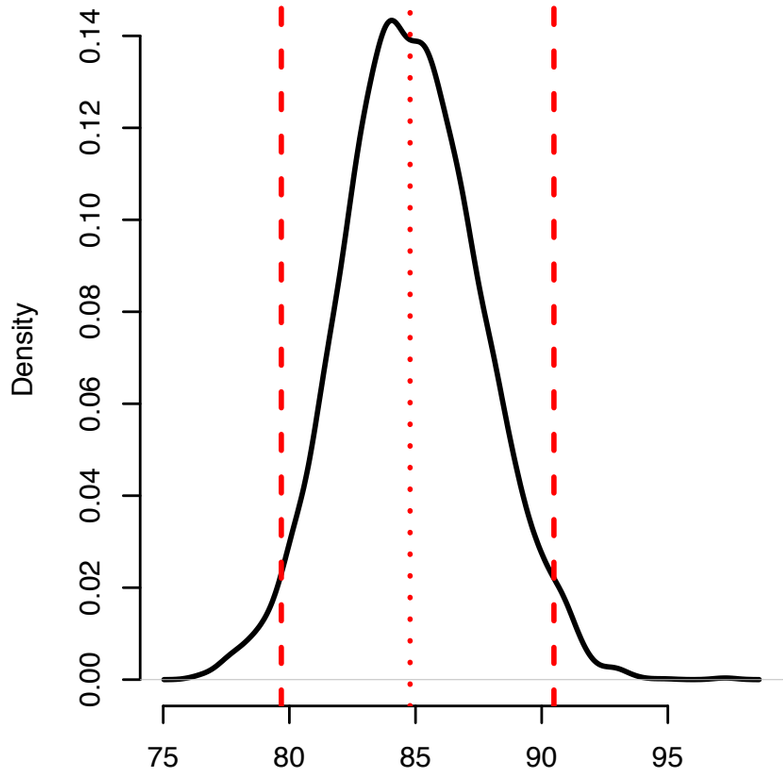
1. sample $\psi^{(s+1)} \sim p(\psi|\mu_1^{(s)}, \ldots, \mu_m^{(s)}, \tau^{2(s)})$
2. sample $\tau^{2(s+1)} \sim p(\tau^2|\mu_1^{(s)}, \ldots, \mu_m^{(s)}, \psi^{(s+1)})$
3. sample $\sigma^{2(s+1)} \sim p(\sigma^2|\mu_1^{(s)}, \ldots, \mu_m^{(s)}, y_1, \ldots y_m)$
4. sample for each $j \in \{1, \ldots, m\}$ sample

$$\mu_j^{(s+1)} \sim p(\mu_j|\psi^{(s+1)}, \tau^{2(s+1)}, \sigma^{2(s+1)}, y_j)$$

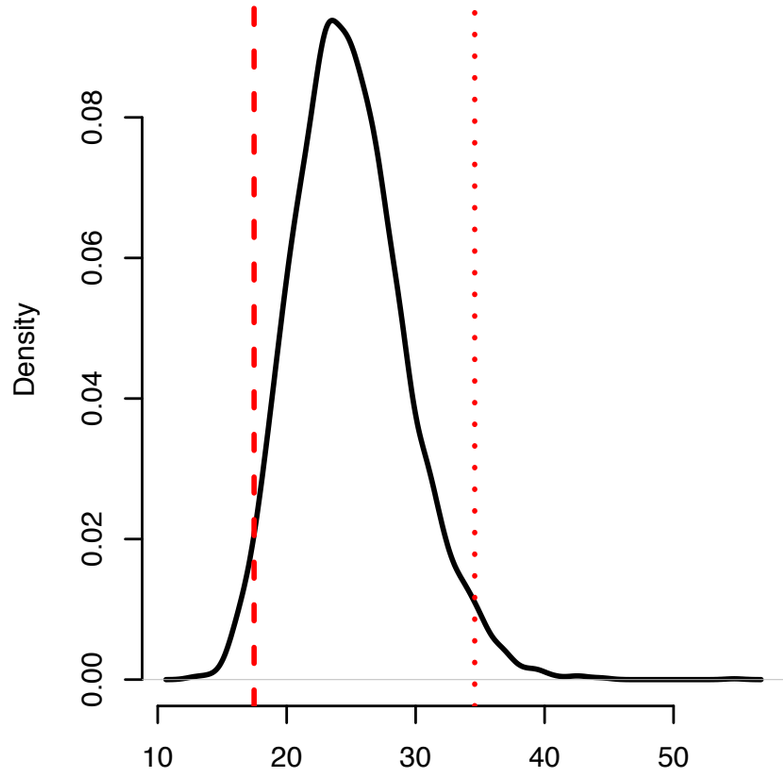
Example: Posterior summaries



$\bar{\psi} = 47.78$



$\bar{\sigma} = 9.21$



$\bar{\tau} = 4.97$

- 95% of the scores within a school are within $4 \times 9.21 \approx 37$ points of each other
- whereas, 95% of the average school scores are within $4 \times 4.97 \approx 20$ points of each other

Example: Shrinkage

One of the motivations behind hierarchical modeling is that information can be shared across groups

Recall that, conditional on ψ, τ^2, σ^2 and the data, the expected value of μ_j is a weighted average of $\bar{y}_j$ and ψ

$$\mathbb{E}\{\mu_j | y_j, \psi, \tau^2, \sigma^2\} = \frac{\bar{y}_j n_j/\sigma^2 + \psi/\tau_2}{n_j/\sigma^2 + 1/\tau^2}$$

As a result, the expected value of μ_j is pulled a bit from $\bar{y}_j$ towards ψ by an amount depending upon n_j

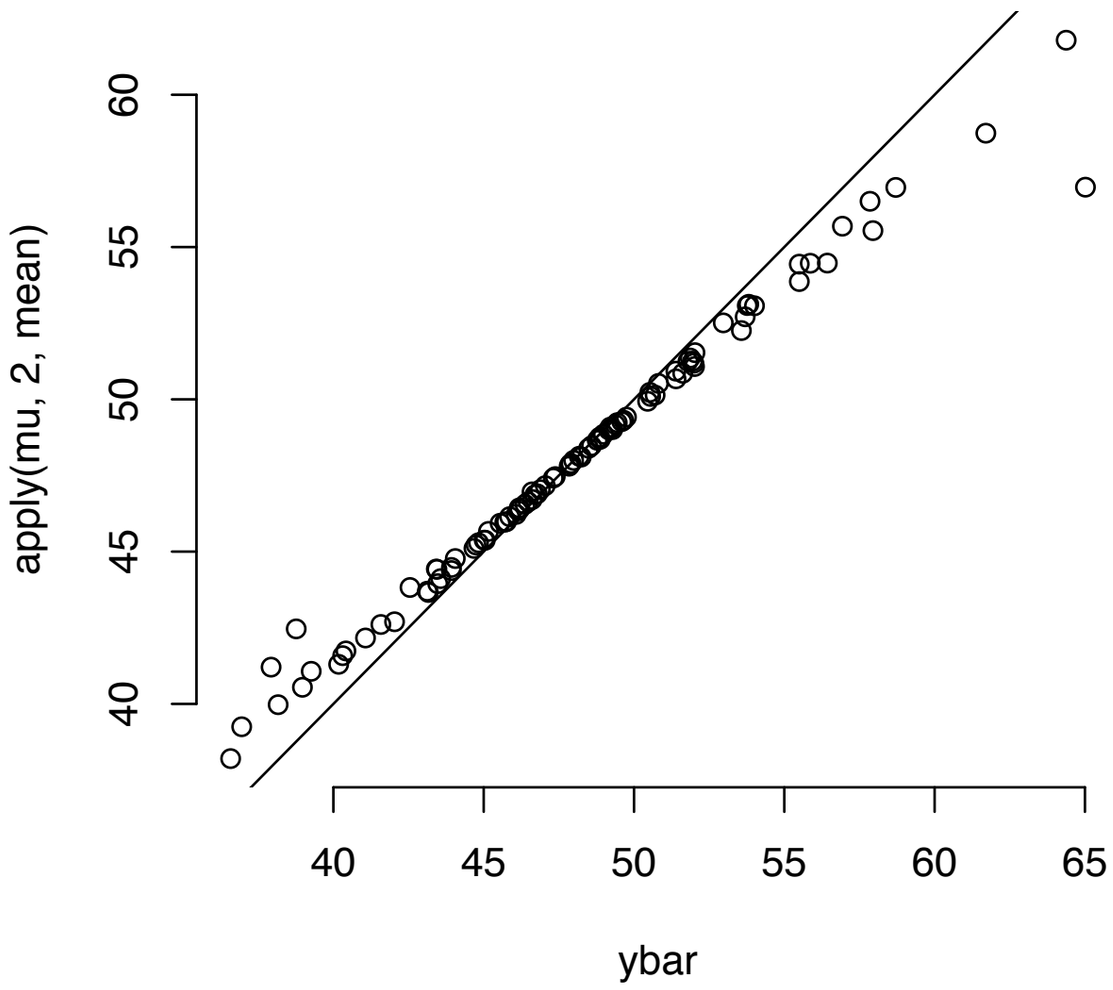
This effect is called shrinkage

Example: Shrinkage

Consider the relationship between $\bar{y}_j$ and

$$\bar{\mu}_j = \mathbb{E}\{\mu_j | y_j, \psi, \tau^2, \sigma^2\}$$

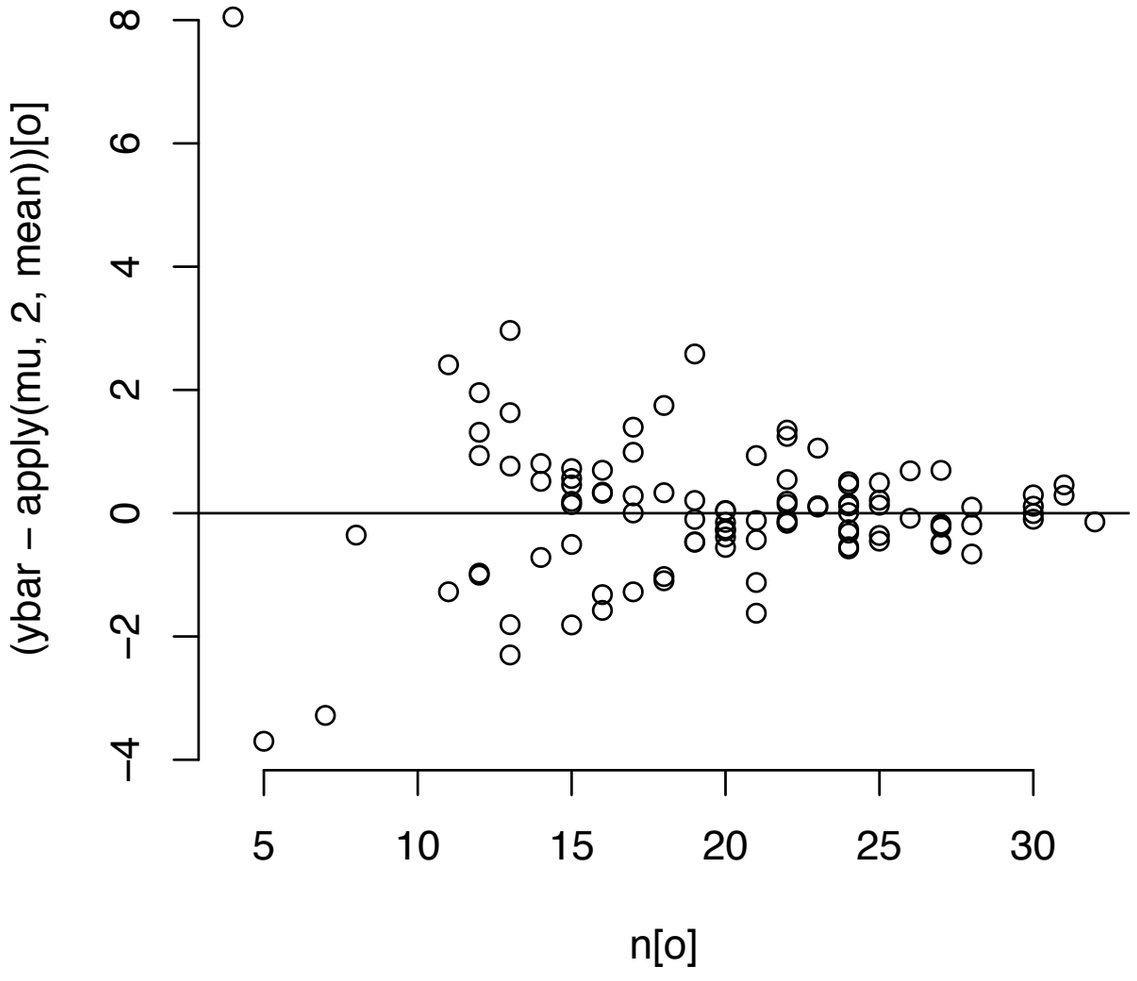
for $j = 1, \ldots, m$ obtained in via our MCMC method



Notice that the relationship follows a line with a slope that is less than one, indicating that high values of $\bar{y}_j$ correspond to slightly less high values of $\bar{\mu}_j$ and vice-versa for low values

Example: Shrinkage

It is also interesting to observe the shrinkage as a function of the group-specific sample size



Groups with low sample sizes get shrunk the most, where as groups with large sample sizes hardly get shrunk at all

This makes sense:

The larger the sample size the more information we have for that group, and the less information we need to borrow from the rest of the population

Example: Ranking schools

Suppose our task is to rank the schools according to what we think their performances would be if every student in each school took the math exam

In this case, it makes sense to rank the schools according to the school-specific posterior expectations $\{\bar{\mu}_1, \ldots, \bar{\mu}_m\}$

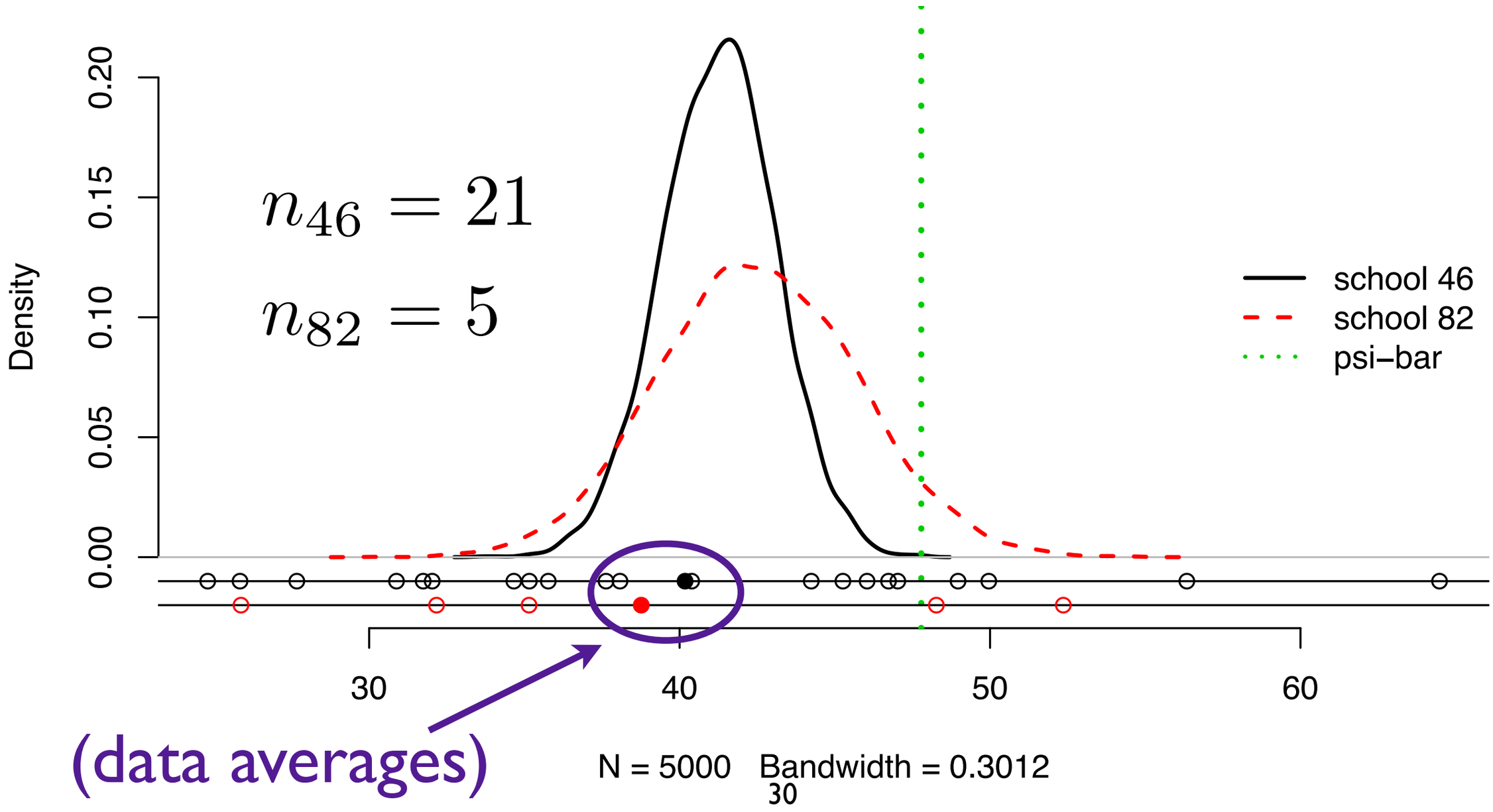
Alternatively, we could ignore the results of the hierarchical model and just use the school-specific averages $\{\bar{y}_1, \ldots, \bar{y}_m\}$

The two methods will give similar, but not identical rankings

Example: Ranking schools

Consider the posterior distributions of μ_{46} and μ_{82}

Both schools have exceptionally low sample means, in the bottom 10% of all schools



(data averages)

Example: Ranking schools

So school 82 is ranked lower than 46 by the data averages, but higher by the posterior group-means

Does this make sense?

The posterior density for school 46 is more peaked because it was derived from a much larger sample size

Therefore our degree of certainty about μ_{46} is much higher than that for μ_{82}

How does this translate into lack of faith in the data averages if we are going to justify using the posterior means instead?

Example: Hypothetical absence

Suppose that on the day of the exam the student who got the lowest exam score in school 82 did not come to class

Then the sample mean for school 82 would have been 41.99 rather than 38.76, a change of more than three points (and a change in ranking)

In contrast, if the lowest performing student in school 46 had not shown up, then $\bar{y}_{46}$ would have been 40.9 as opposed to 40.18, a change of only 3/4

Example: Hypothetical absence

In other words, the low value of the sample mean for school 82 can be explained by either μ_{82} being very low

... or just the possibility that a few of the 5 sampled students were among the poorer performing students in the school

In contrast for school 46, this latter possibility cannot explain the low value of the sample mean, because of the large sample size

Therefore it makes sense to shrink the expectation of school 82 towards the population expectation $\bar{\psi}$ by a greater amount than for 46

Example: Is it fair?

To some people this reversal of rankings may seem strange or "unfair"

While "fairness" may be debated, the hierarchical model reflects the objective fact that there is more evidence that μ_{46} is exceptionally low than there is evidence that μ_{82} is exceptionally low

There are many other real-life situations where differing amounts of evidence results in a switch of ranking

Hierarchical binomial model

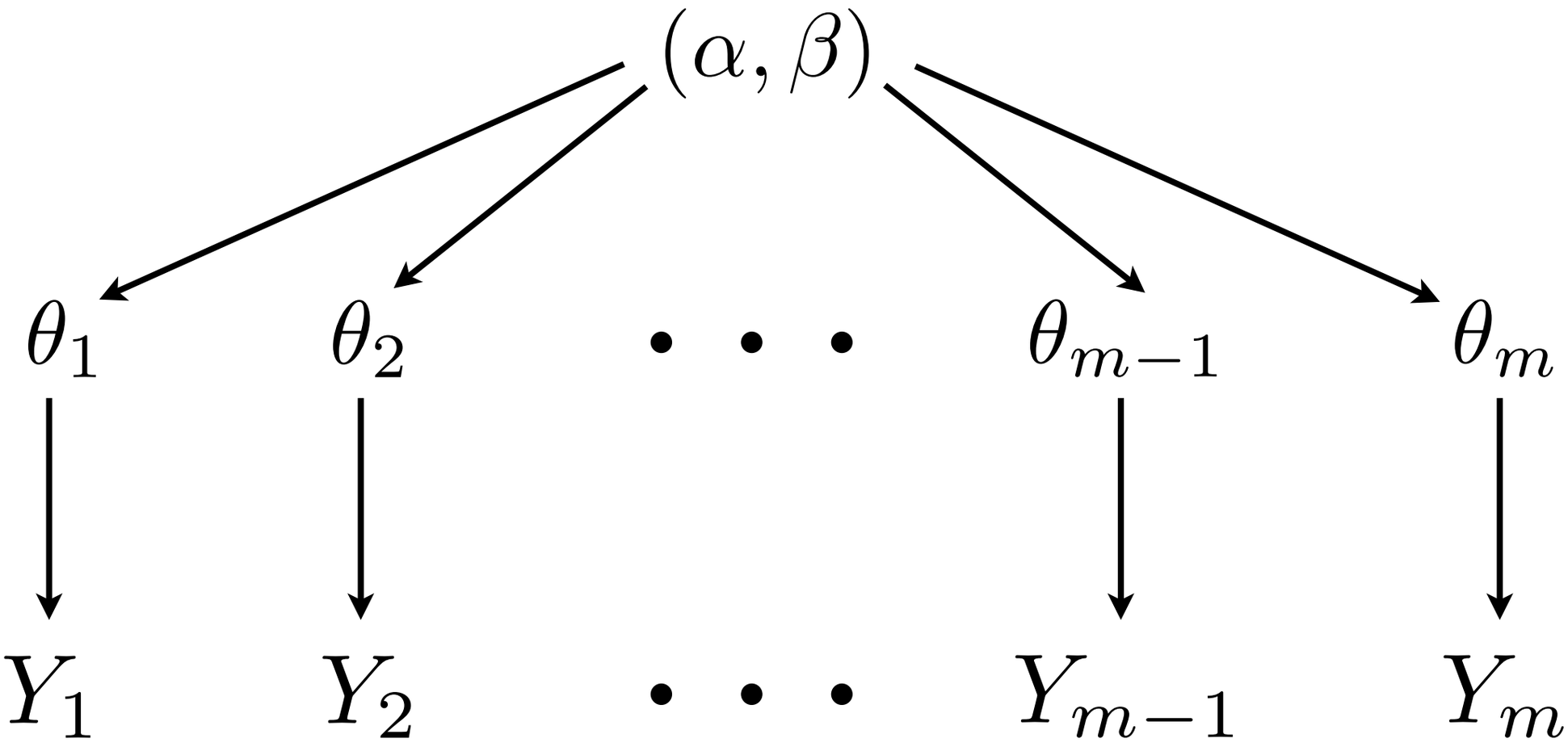
Another commonly used hierarchical model is the Beta-binomial model, where

$$Y_j \sim \text{Bin}(n_j, \theta_j)$$
$$\theta_j \sim \text{Beta}(\alpha, \beta)$$
$$(\alpha, \beta) \sim p(\alpha, \beta)$$

Conditional on α and β the posterior conditional for θ_j is the familiar Beta distribution

$$\theta_j | Y_j, \alpha, \beta \sim \text{Beta}(\alpha + y_j, \beta + n_j - y_j)$$

Hierarchical diagram



As usual, we treat $n_1, \ldots, n_m$ as known

Hyperprior

Unfortunately, there is no (semi-) conjugate prior for α and β

However, it is possible to set up a non-informative hyperprior that is dominated by the likelihood and yields a proper posterior distribution which leads to a convenient Metropolis-within-Gibbs sampling method

A hyperprior choice

A reasonable choice of diffuse hyperprior for the Beta-binomial hierarchical model is uniform on

$$\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right)$$

A "change of variables" shows that this implies the following prior on the original scale

$$p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$$

Is proper as long as $0 < y_j < n_j$ for at least one experiment j

The posterior marginal

How can we calculate the posterior marginal distribution?

(cond prob.) (Bayes' rule)

$$p(\alpha,\beta|y) = \frac{p(\alpha,\beta,\theta|y)}{p(\theta|\alpha,\beta,y)} \propto \frac{p(y|\alpha,\beta,\theta)p(\alpha,\beta,\theta)}{p(\theta|\alpha,\beta,y)}$$

$$= \frac{p(y|\theta)p(\theta|\alpha,\beta)p(\alpha,\beta)}{p(\theta|\alpha,\beta,y)}$$ (cond indep. & cond prob.)

$$= \frac{\prod_{j=1}^{m}\text{Bin}(y_j;n_j,\theta_j)\times\prod_{j=1}^{m}\text{Beta}(\theta_j;\alpha,\beta)\times p(\alpha,\beta)}{\prod_{j=1}^{m}\text{Beta}(\theta_j;\alpha+y_j,\beta+n_j-y_j)}$$

(...)
$$\propto p(\alpha,\beta)\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right]^m\prod_{j=1}^{m}\frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$$

Metropolis-within-Gibbs

We may sample from the marginal $p(\alpha, \beta|y)$ by generating $(\alpha^{(s+1)}, \beta^{(s+1)})$ from $(\alpha^{(s)}, \beta^{(s)})$ via

1. propose $\alpha' \sim \text{Unif}(\alpha^{(s)}/2, 2\alpha^{(s)})$
2. if $u < \dfrac{p(\alpha', \beta^{(s)}|y)}{p(\alpha^{(s)}, \beta^{(s)}|y)} \times \dfrac{\alpha^{(s)}}{\alpha'}$ for $u \sim \text{Unif}(0, 1)$

   then $\alpha^{(s+1)} \leftarrow \alpha'$, else $\alpha^{(s+1)} \leftarrow \alpha^{(s)}$
3. propose $\beta' \sim \text{Unif}(\beta^{(s)}/2, 2\beta^{(s)})$
4. if $u < \dfrac{p(\alpha^{(s+1)}, \beta'|y)}{p(\alpha^{(s+1)}, \beta^{(s)}|y)} \times \dfrac{\beta^{(s)}}{\beta'}$ for $u \sim \text{Unif}(0, 1)$

   then $\beta^{(s+1)} \leftarrow \beta'$, else $\beta^{(s+1)} \leftarrow \beta^{(s)}$

Completing the MC

Conditional on samples $(\alpha^{(s)}, \beta^{(s)})$ we may complete the MC method for sampling from the joint posterior distribution by sampling from the conditional(s)

$$\theta_j^{(s)} \sim \text{Beta}(\alpha^{(s)} + y_j, \beta^{(s)} + n_j - y_j)$$

for $j = 1, \ldots, m$

We may also obtain samples from the posterior predictive $p(\tilde{y}|y) = p(\tilde{y}_1, \ldots, \tilde{y}_m | y_1, \ldots, y_m)$ as

$$\tilde{y}_j^{(s)} \sim \text{Bin}(n_j, \theta_j^{(s)})$$

for $j = 1, \ldots, m$

Example: risk of tumors in a group of rats

In the evaluation of drugs for possible clinical application, studies are routinely performed on rodents

In a particular study, the aim is to estimate the probability of a tumor in a population of female rats "F344" that receive a zero-dose of the drug (control group)

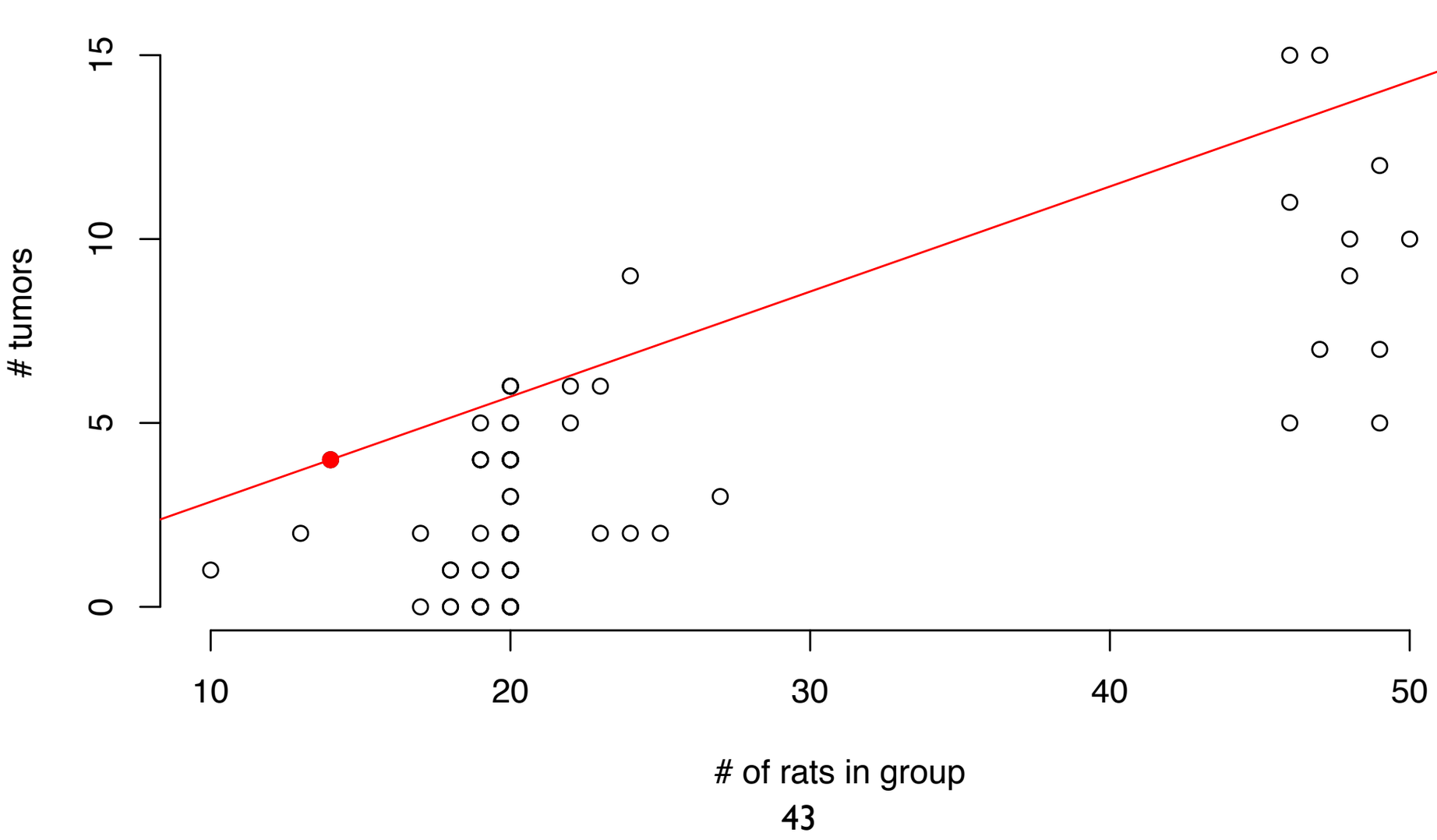
The data show that 4/14 rats developed endometrial stromal polyps (a kind of tumor)

Typically, the mean and standard deviation of underlying tumor risks are not available to form a prior

Example: prior *data*

Rather, historical *data* are available on previous experiments on similar groups of rats

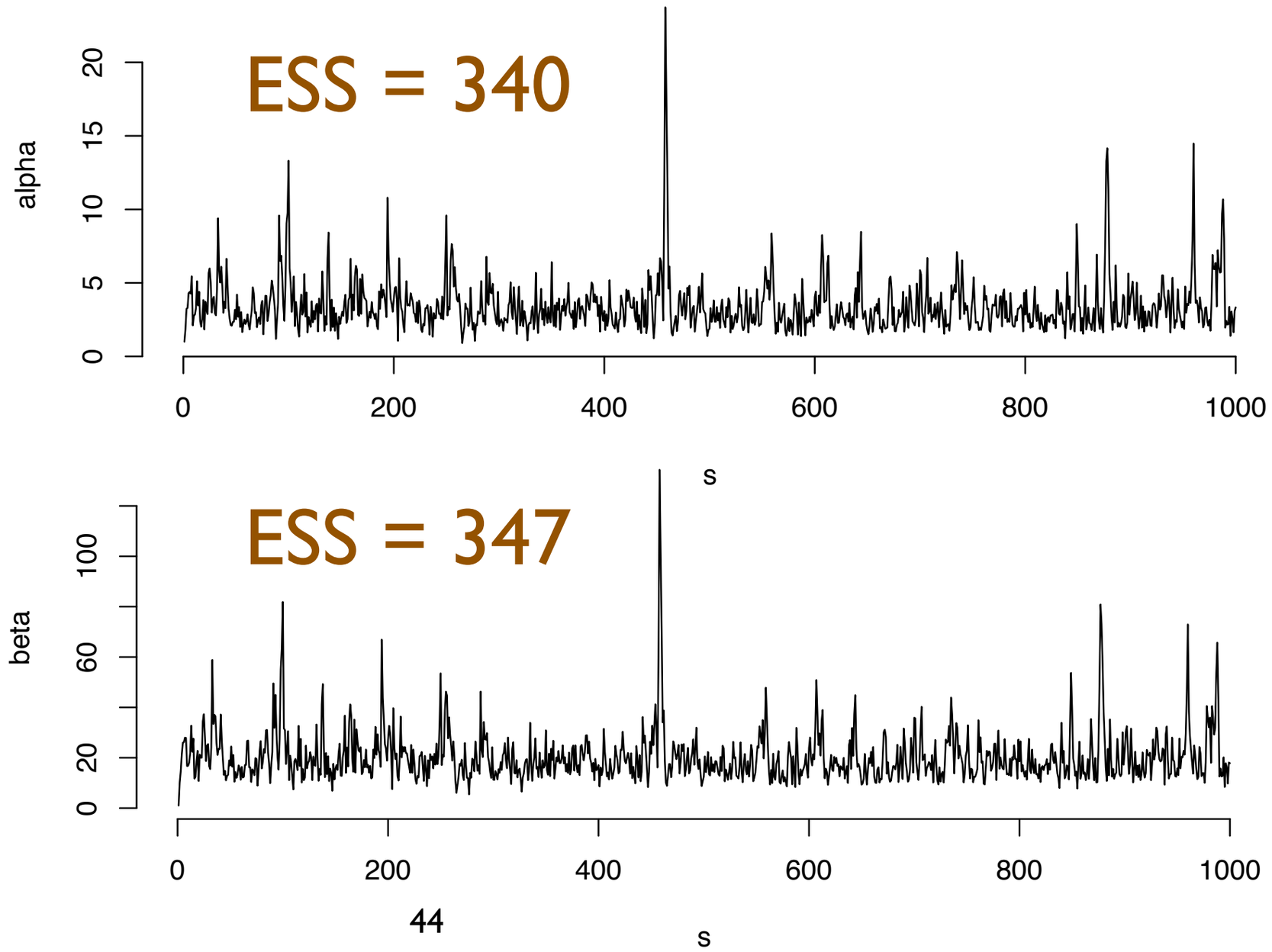
Tarone (1982) provides data on the observations of tumor incidence in 70 groups of rats



Example: Bayesian analysis

We shall model the $m = 71$ rat tumor data with a hierarchical Beta-binomial sampling model with MC(MC) inference, as just described

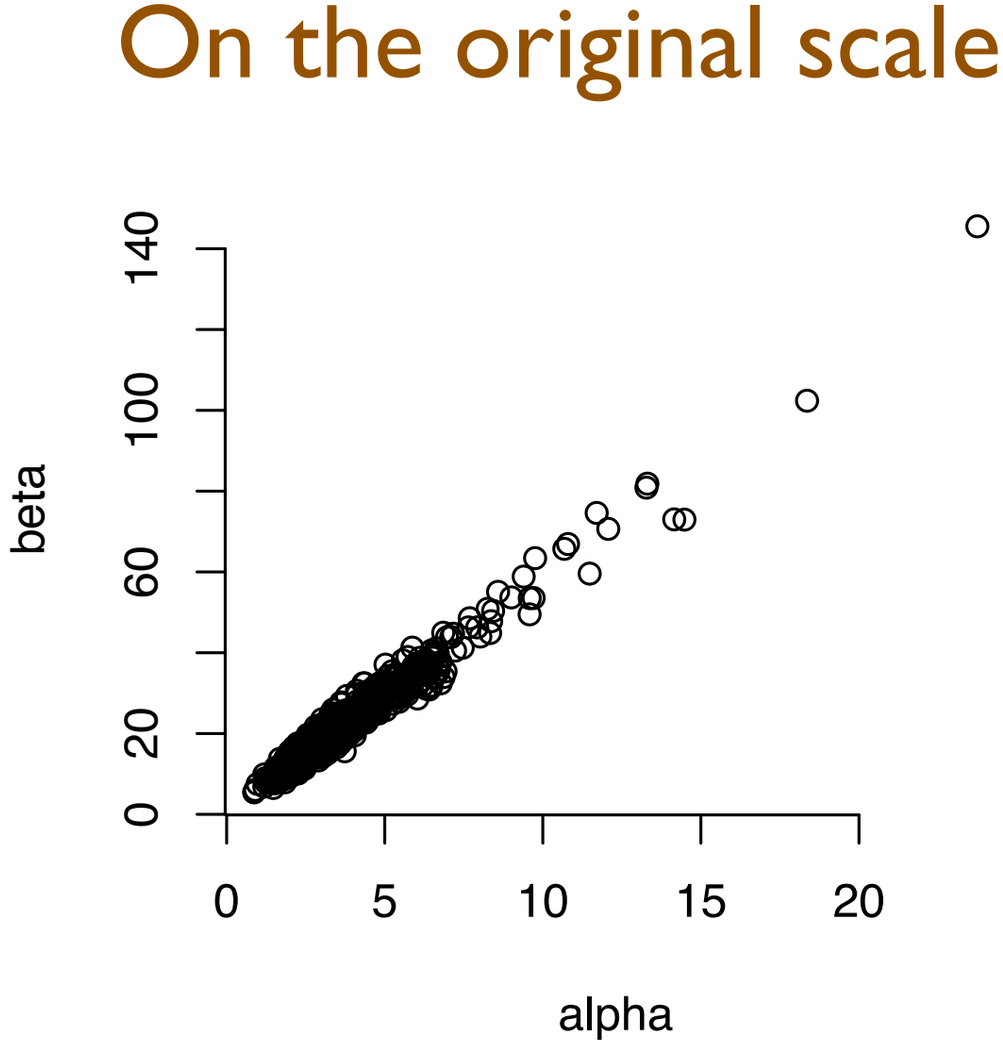
First we must obtain samples from the marginal posterior of (α, β)

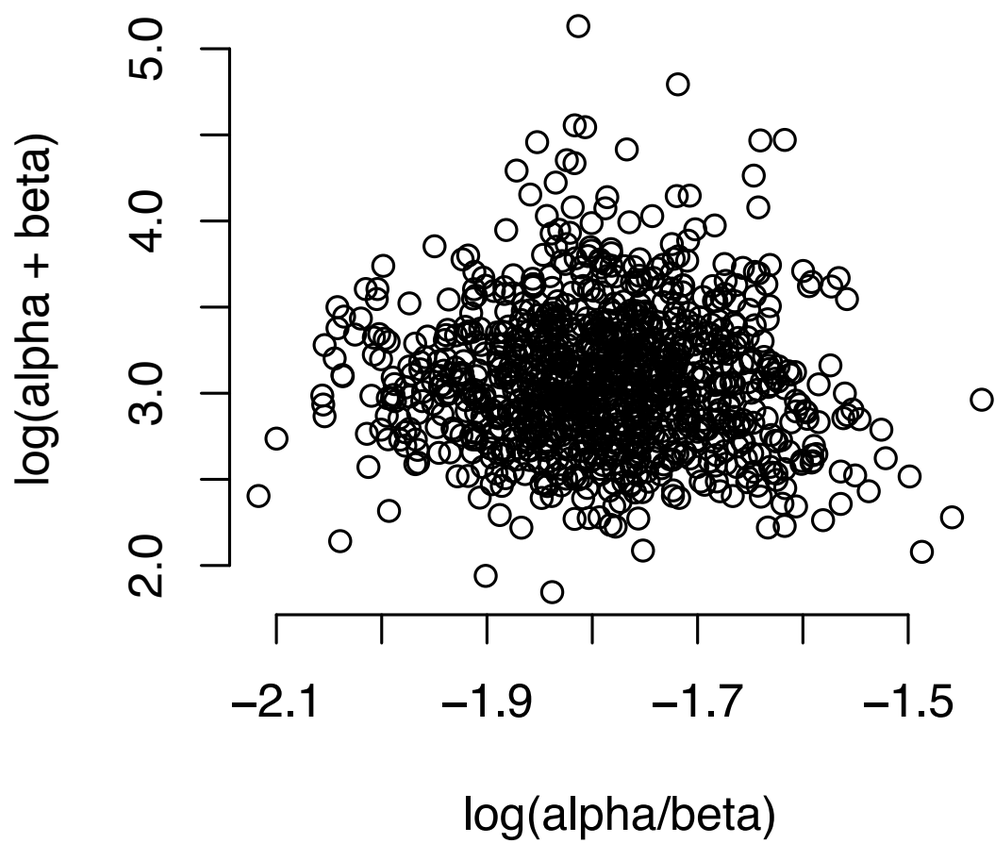


Example: The posterior marginal

Once we have determined that the mixing is good, and we think the chain has achieved stationarity we can inspect the marginal posterior in a number of ways

On the original scale

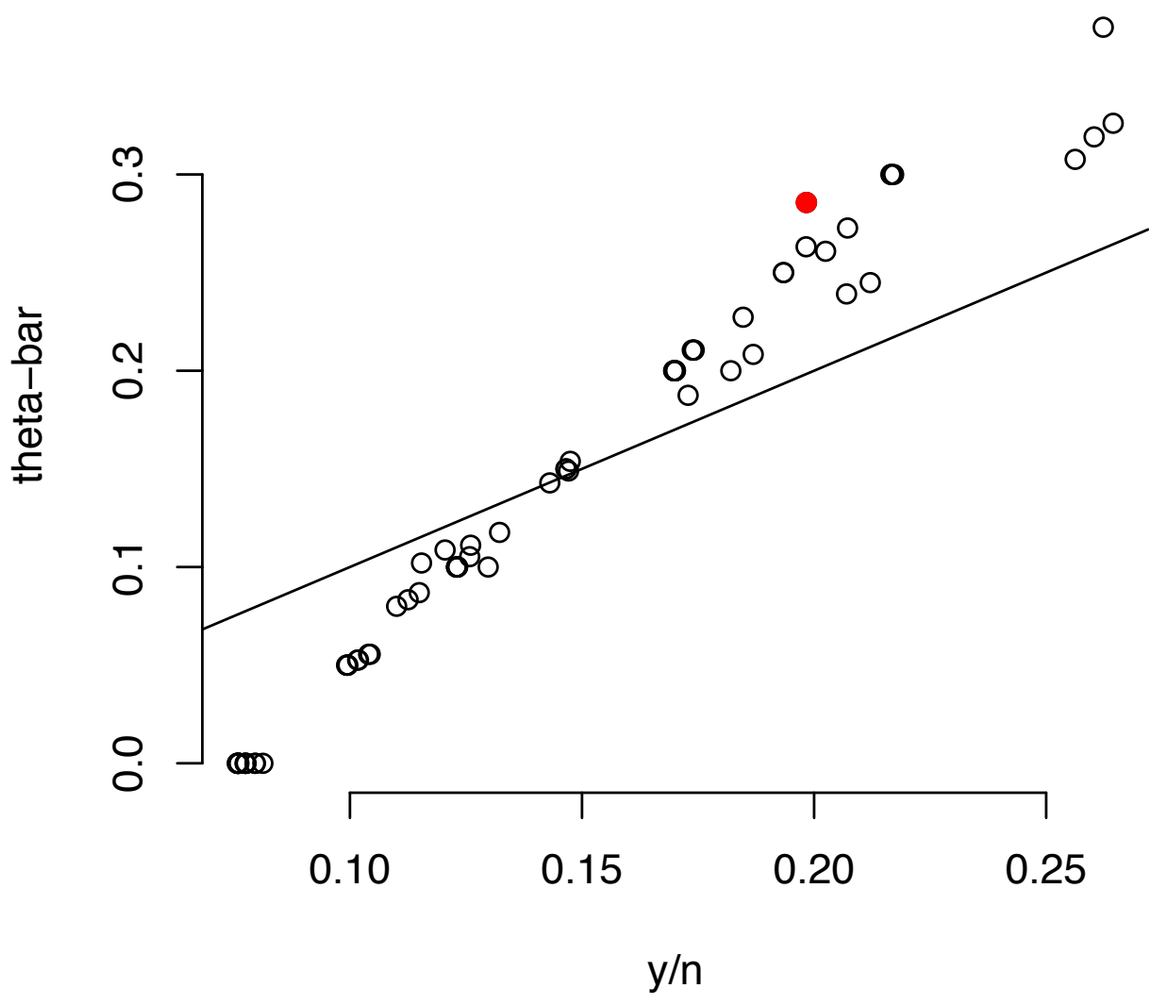


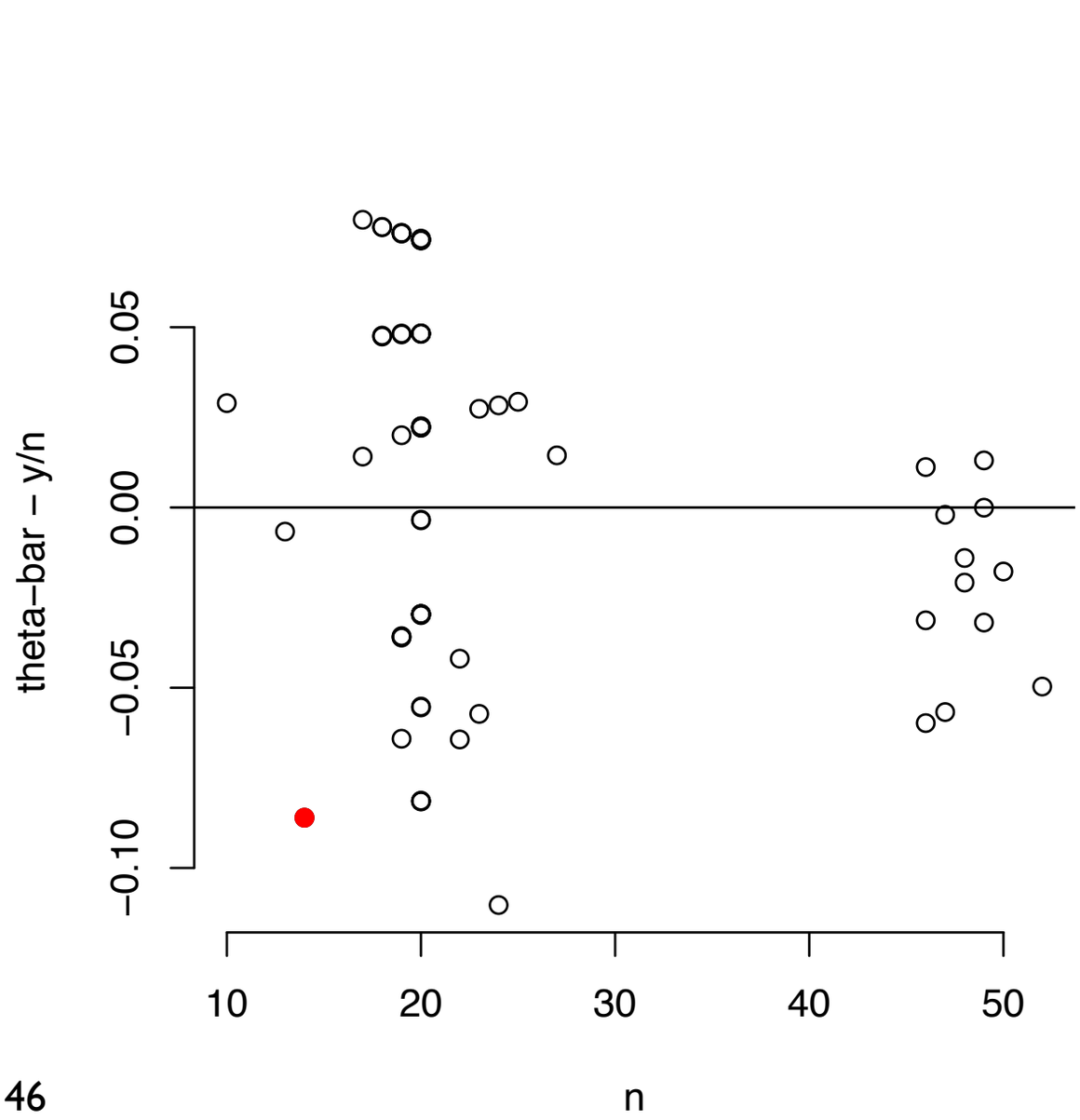


A sensible transformation

Example: Mean Shrinkage

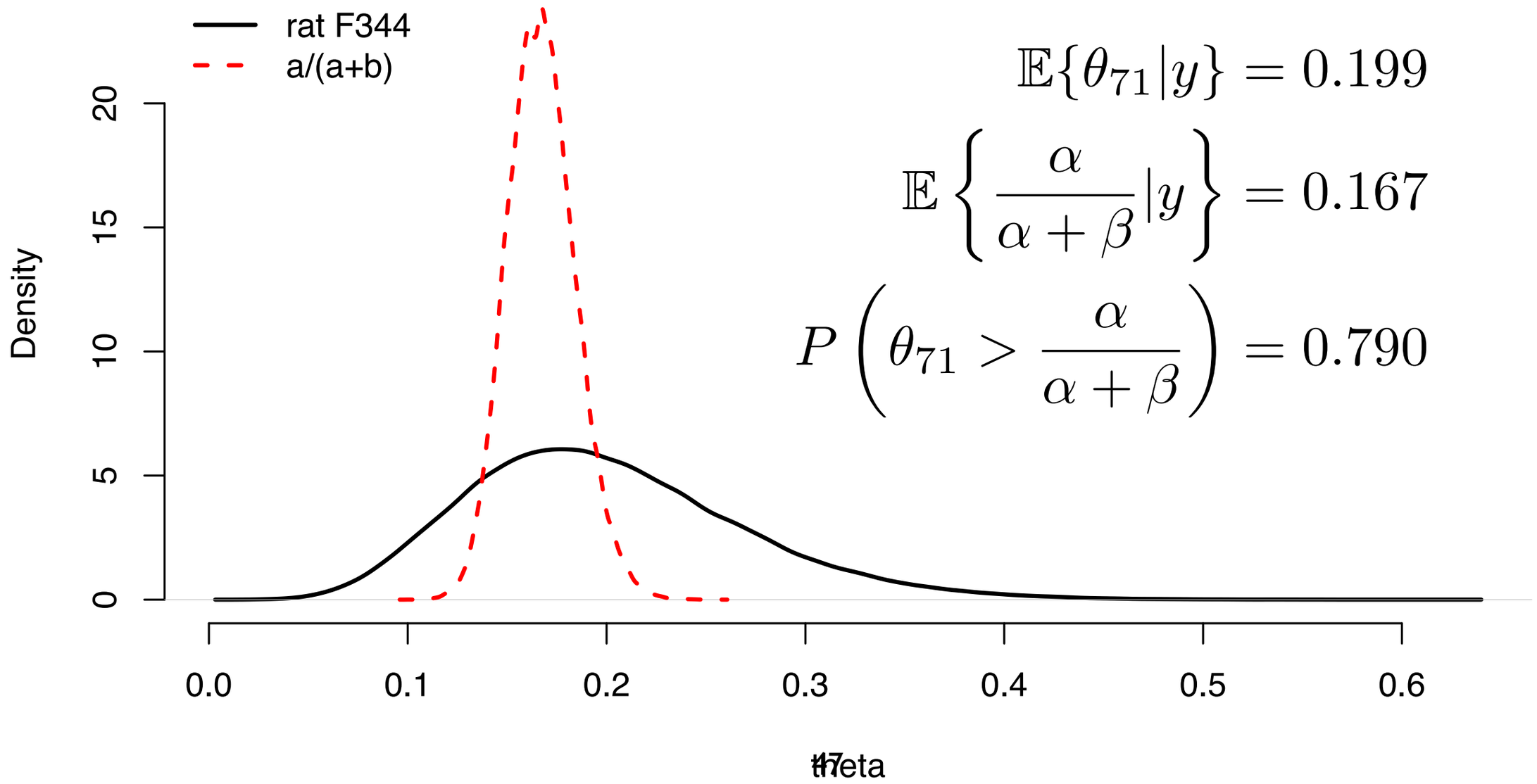
As in the normal hierarchical model, we can assess the amount of shrinkage in the group-specific means, which may be obtained by direct MC, in a number of ways





Example: Rat group F344

We can now present the posterior distribution for for our 71st rat group, and compare it to the population mean of tumor rates in the 70 "prior" rat groups



$$\mathbb{E}\{\theta_{71}|y\} = 0.199$$

$$\mathbb{E}\left\{\frac{\alpha}{\alpha+\beta}|y\right\} = 0.167$$

$$P\left(\theta_{71} > \frac{\alpha}{\alpha+\beta}\right) = 0.790$$

In summary

We have seen how hierarchical models may be used to

- model data which are nested or have a natural hierarchy
- pool information about groups of similar populations so that smaller groups may borrow information from larger ones (i.e., shrinkage)
- provide an efficient way of using "prior data" in an appropriate way