Bayesian Hierarchical Models

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
- sites within counties within regions within countries

Why groups?

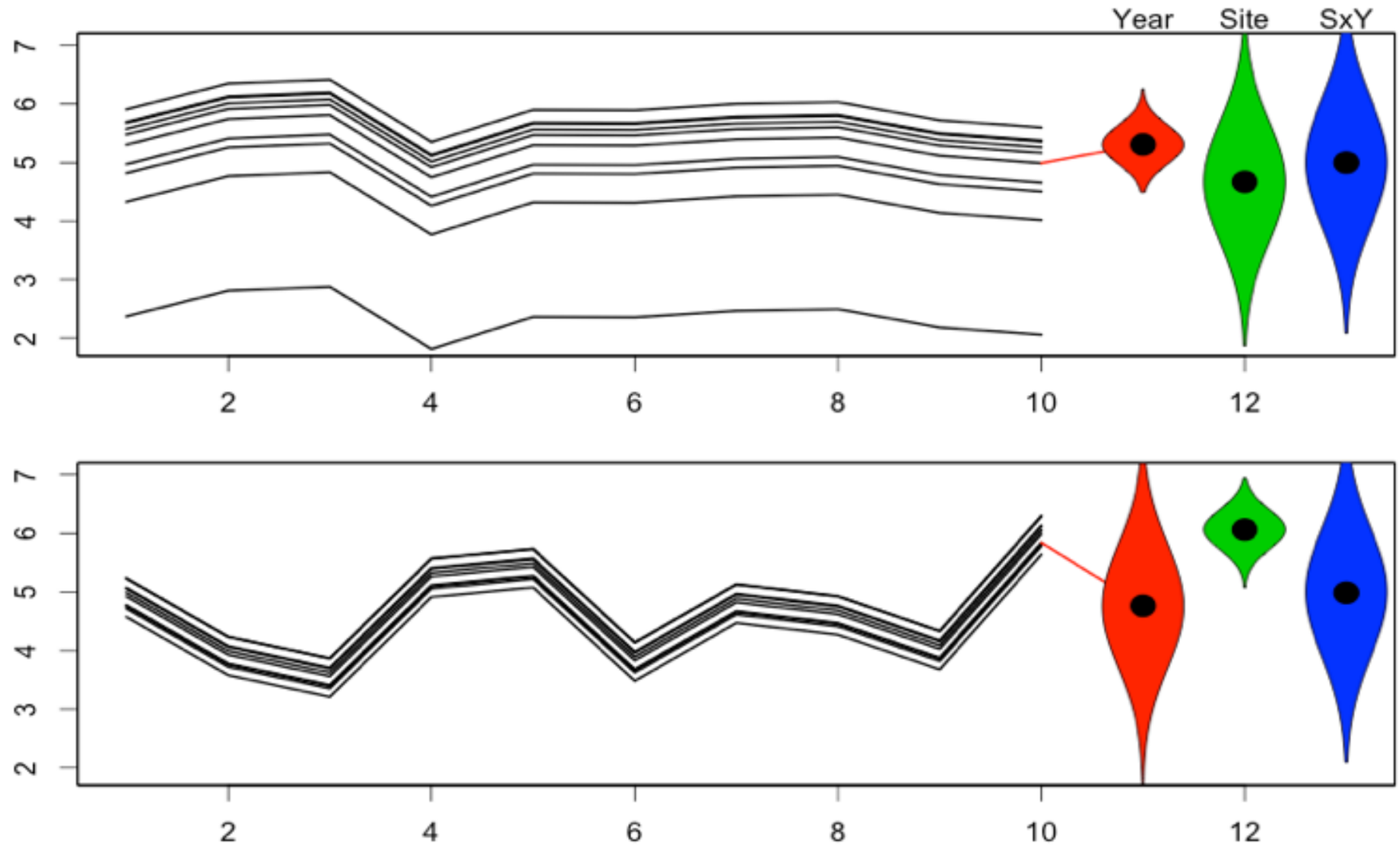
We are usually interested in these groupings/nested structures because the indicate where we think variability may come in.

- For example: Plots, Blocks, Years, Individuals
- we often use this to account for a lack of independence between samples within a group

Usually we need some replication within a group (or we won't be able to estimate the variance...)

Why groups?

Accounting for this nesting/group structure can have a big impact on our inferences about what is going on in a system:



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.

We want to explicitly model the variation between these in order to properly partition and identify the sources of randomness in our system.

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

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

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

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

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

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

$p(\phi)$ represents information about a single unknown quantity

Hierarchical normal model

We use this to model the heterogeneity of means across several populations so that the within- and between-group sampling models are both normal:

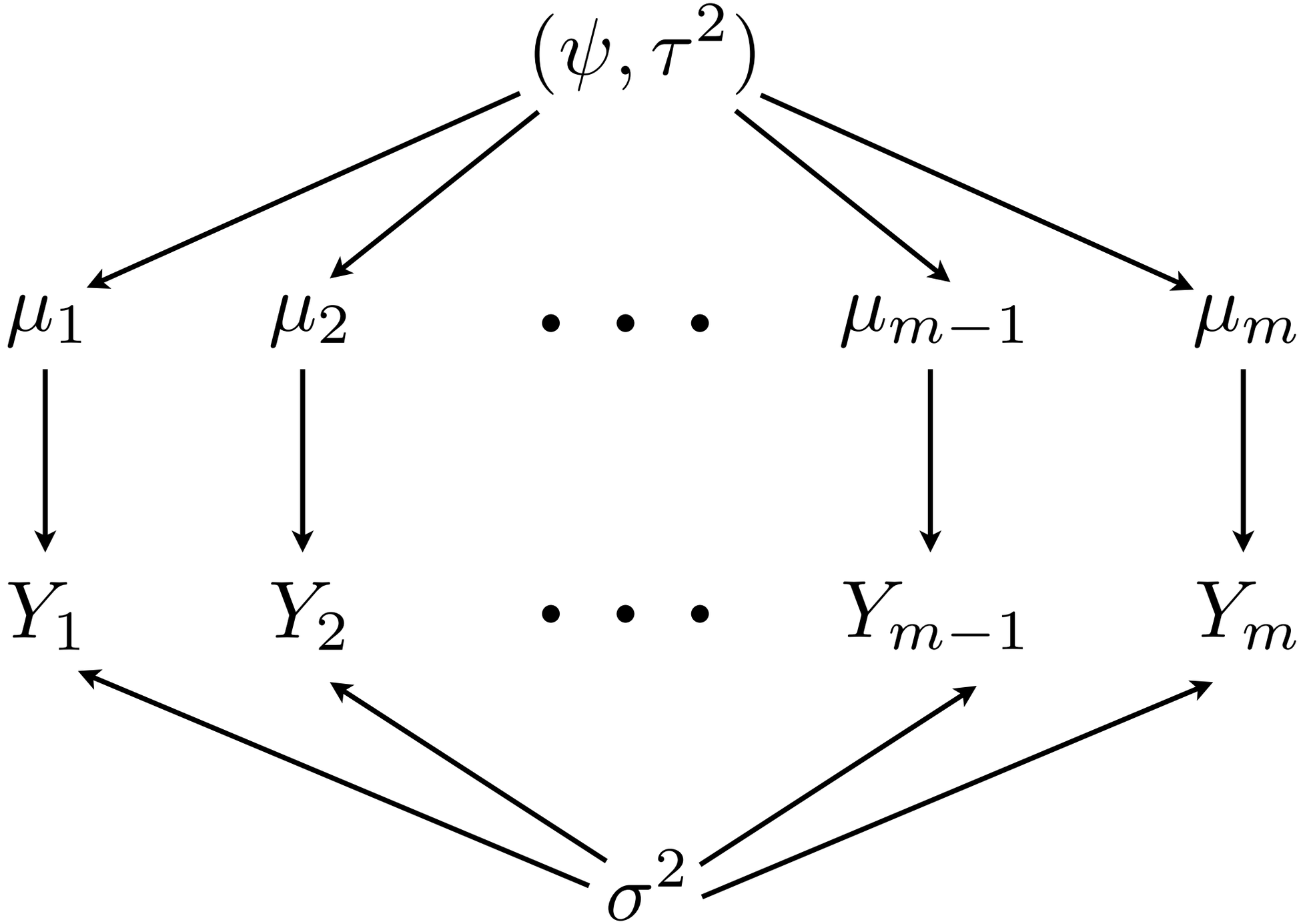
$$p(y|\theta_j) = \mathcal{N}(\mu_j, \sigma^2) \quad \text{(within-group model)}$$

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

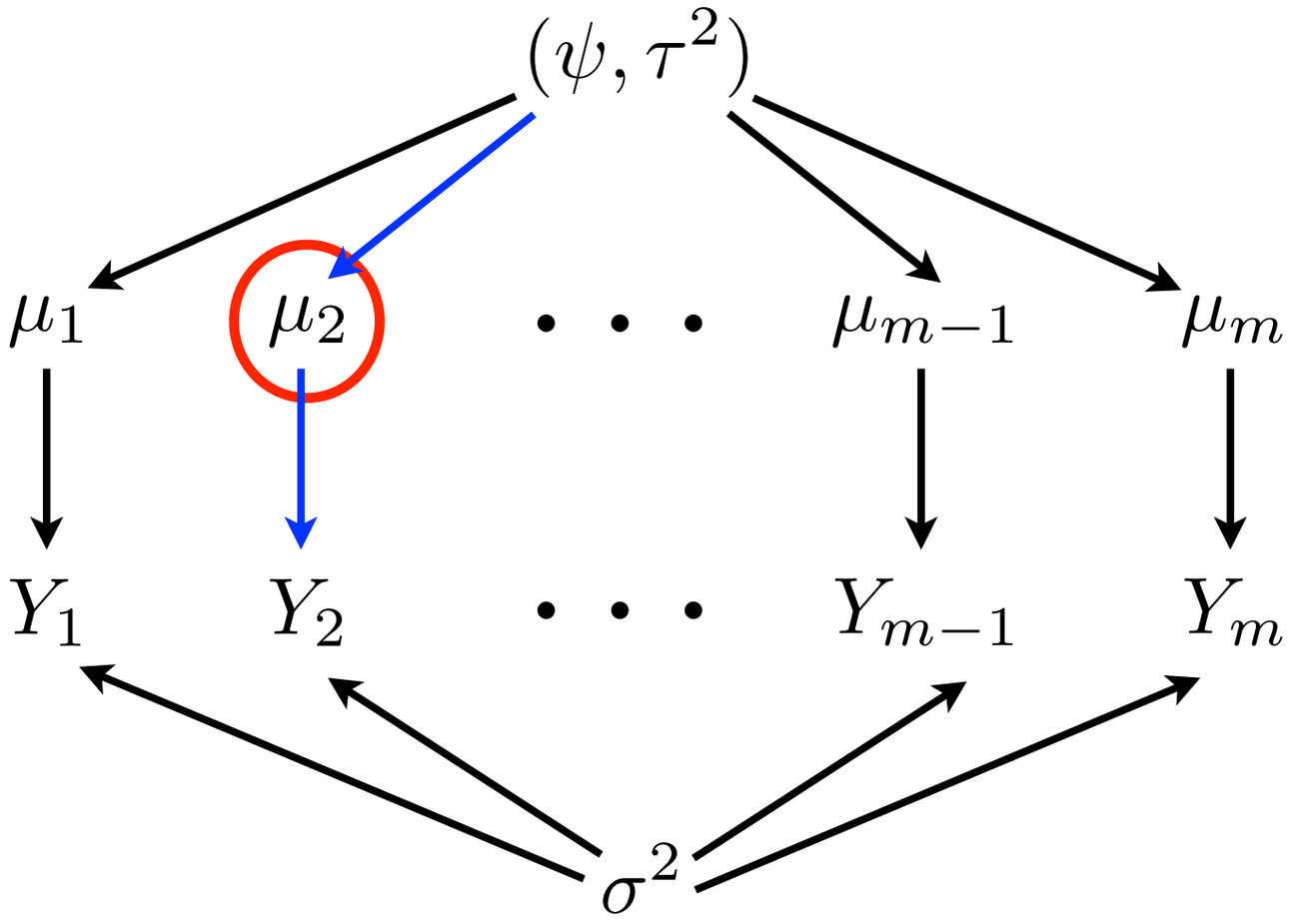
Note that $p(\theta|\phi)$ only describes heterogeneity across group means, and **not** any heterogeneity in group-specific variances

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

Hierarchical diagram



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

Hierarchical normal model: priors

The "fixed" but unknown parameters in the models are ψ, τ^2 and σ^2 (these are shared among all the data).

The most common prior choice would be the "semi-conjugate" normal and Inverse-Gamma priors:

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

Same as putting a gamma prior on the precision instead!

Then turn the Bayesian crank...

Example: Math scores in US public schools

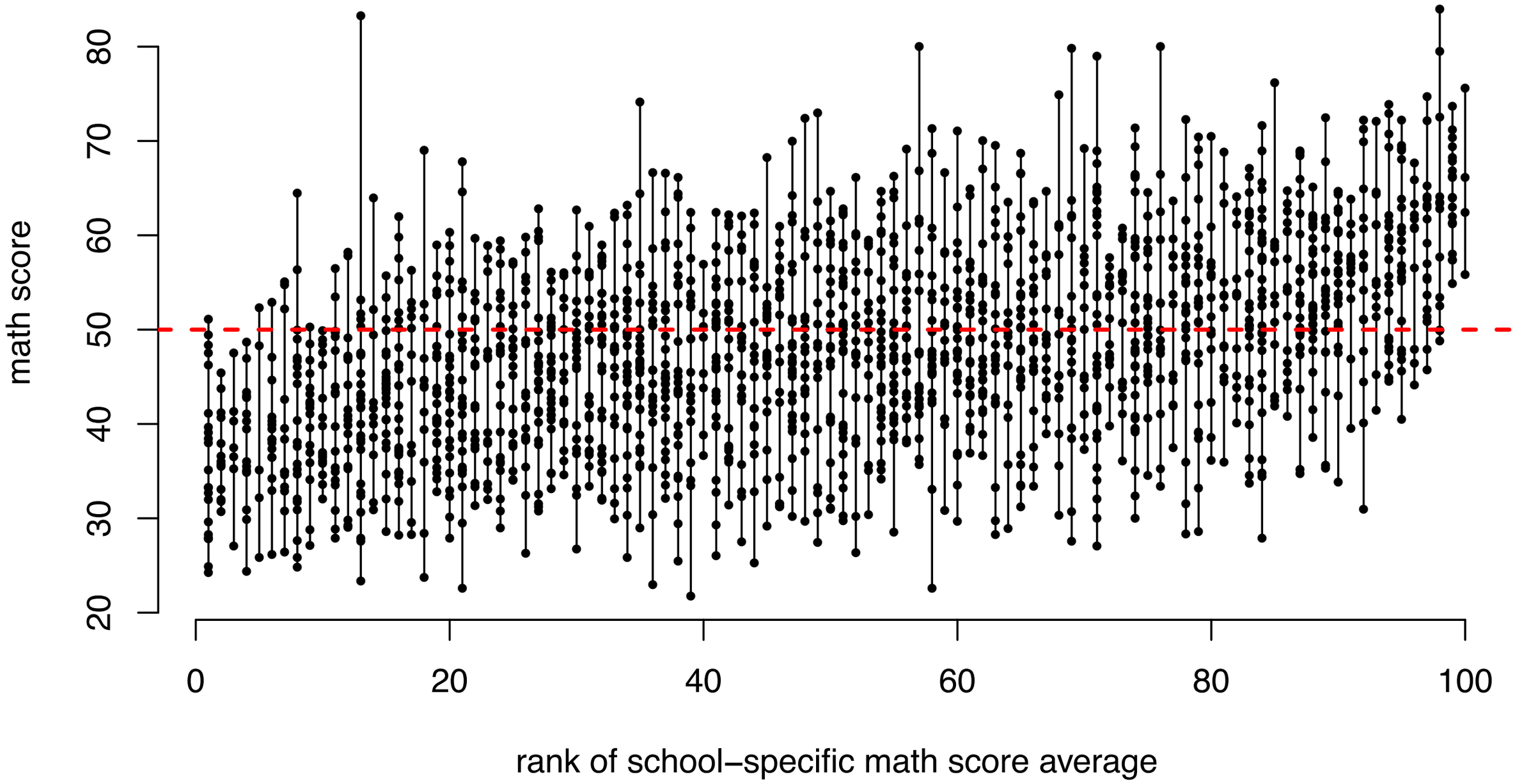
Consider data that is part of the 2002 Educational Longitudinal Study (ELS), a survey of students from a large sample of schools in the United States.

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

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

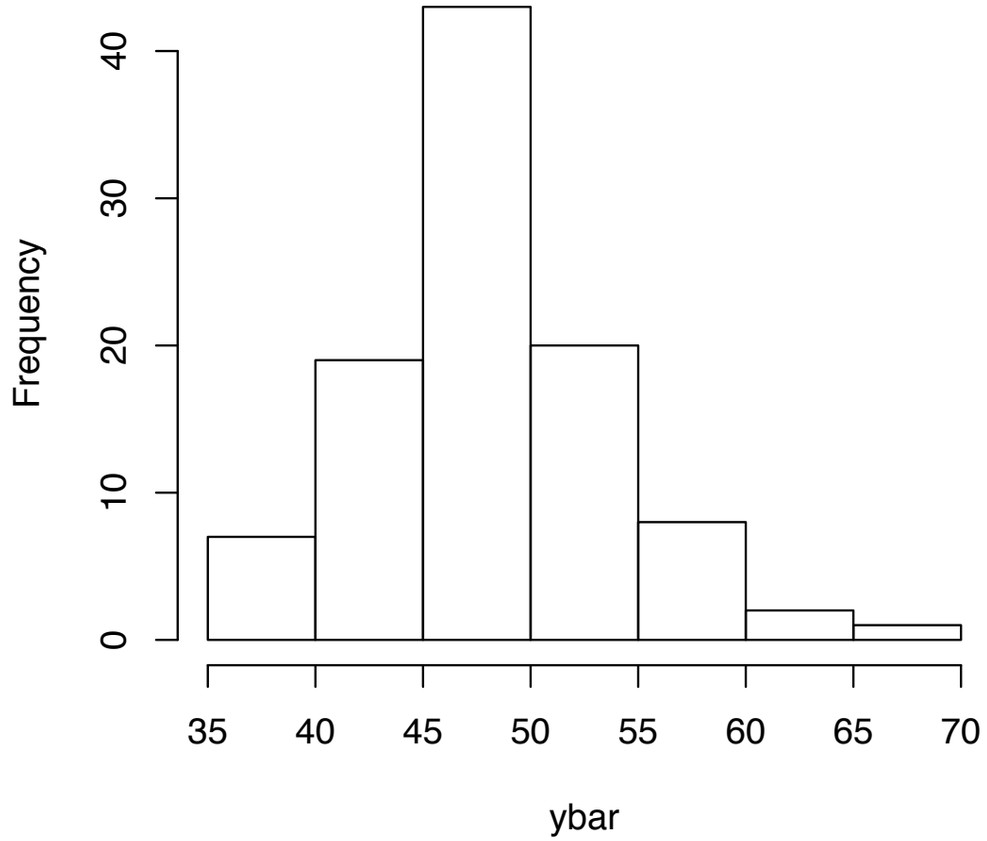
Example: the data

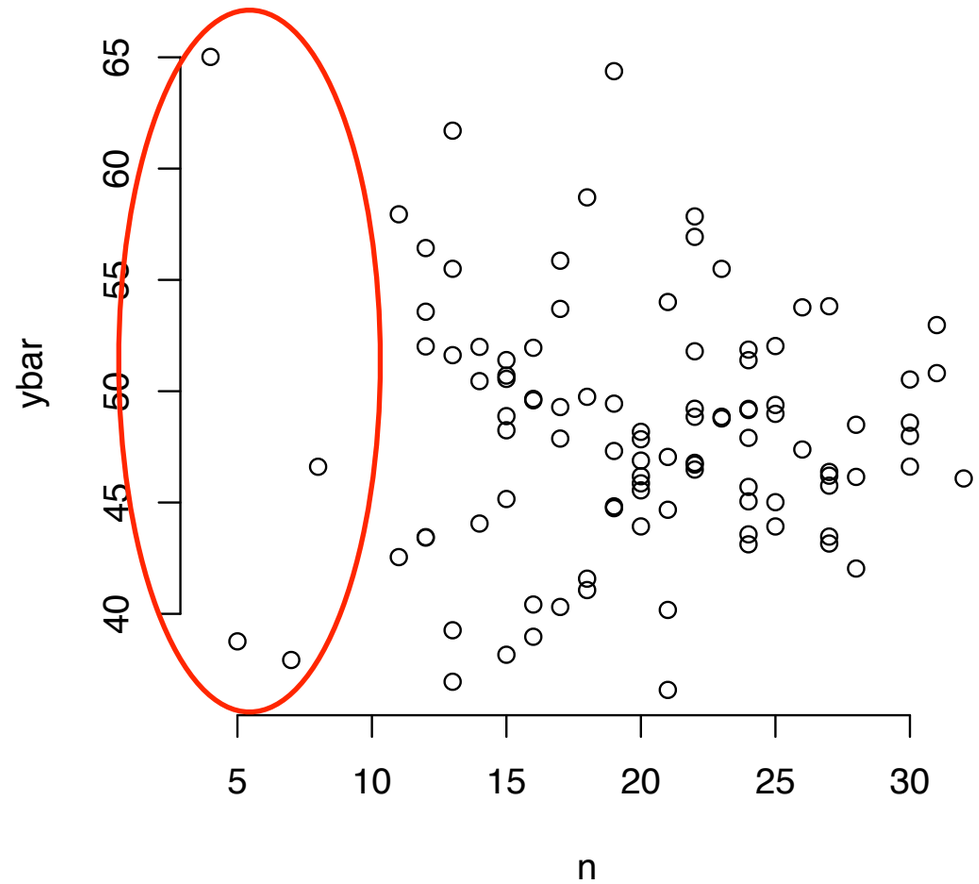
Scores for students within the same school plotted along a common vertical bar:



Example: the data

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

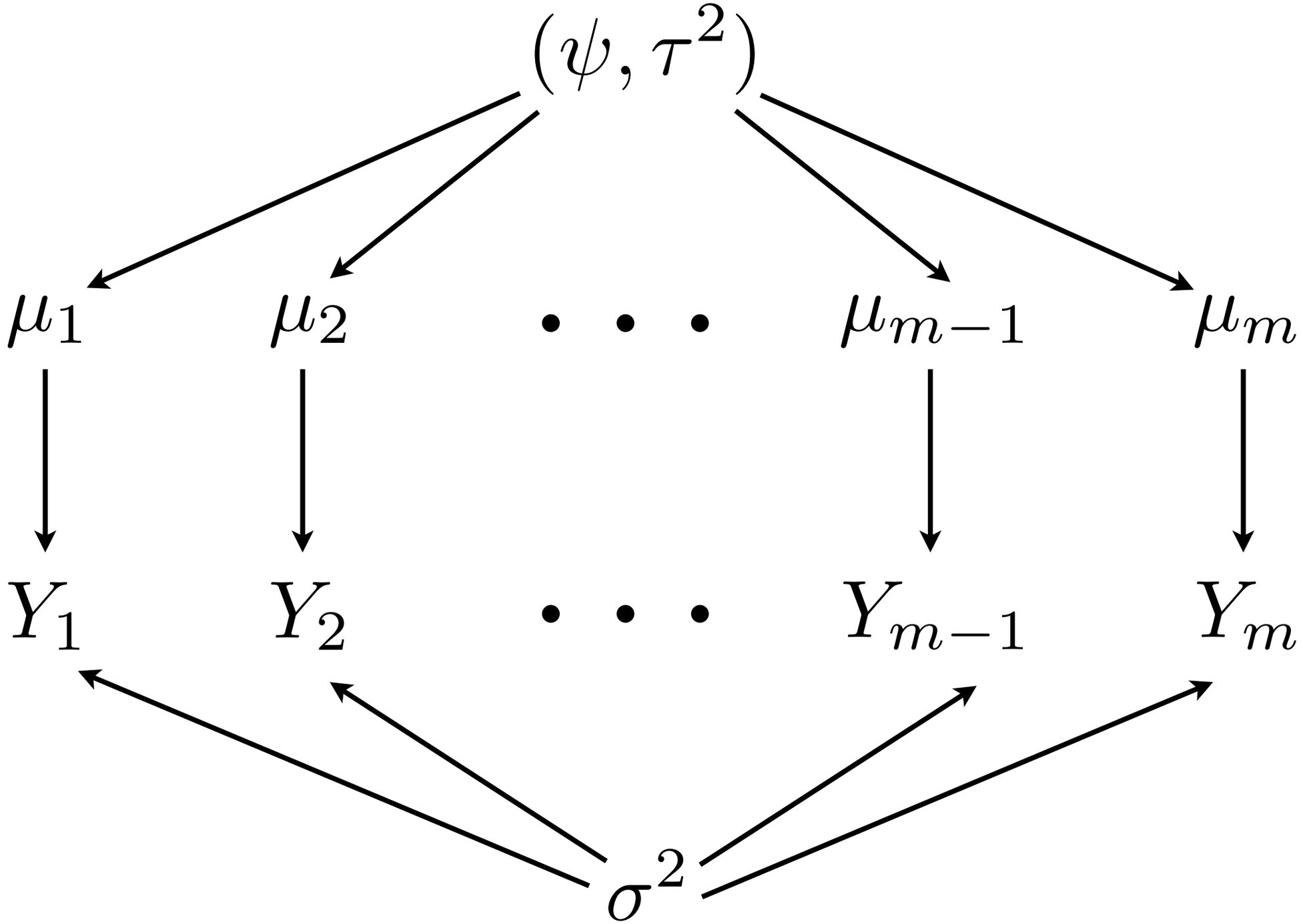




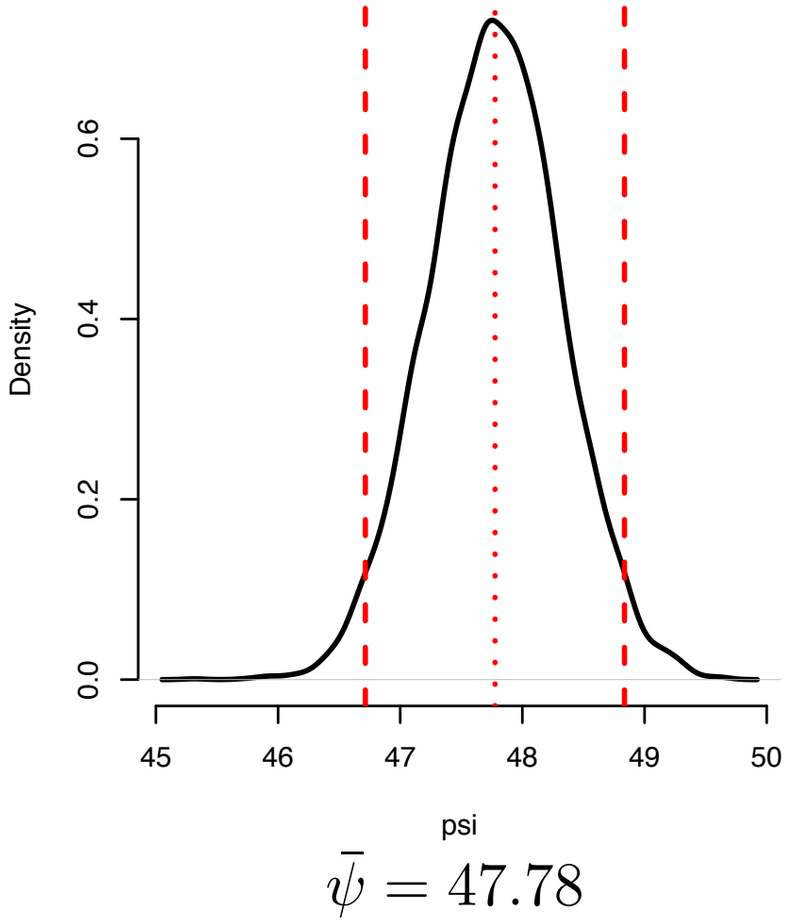
Extreme sample averages occur for schools with small sample sizes

This is a common relationship in hierarchical datasets

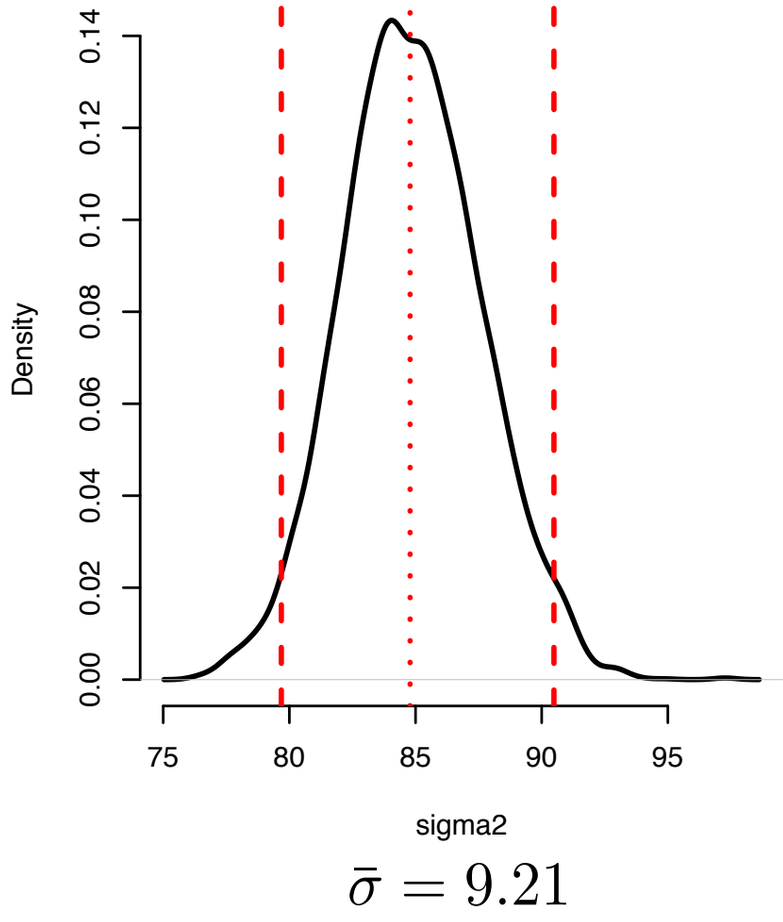
Hierarchical diagram



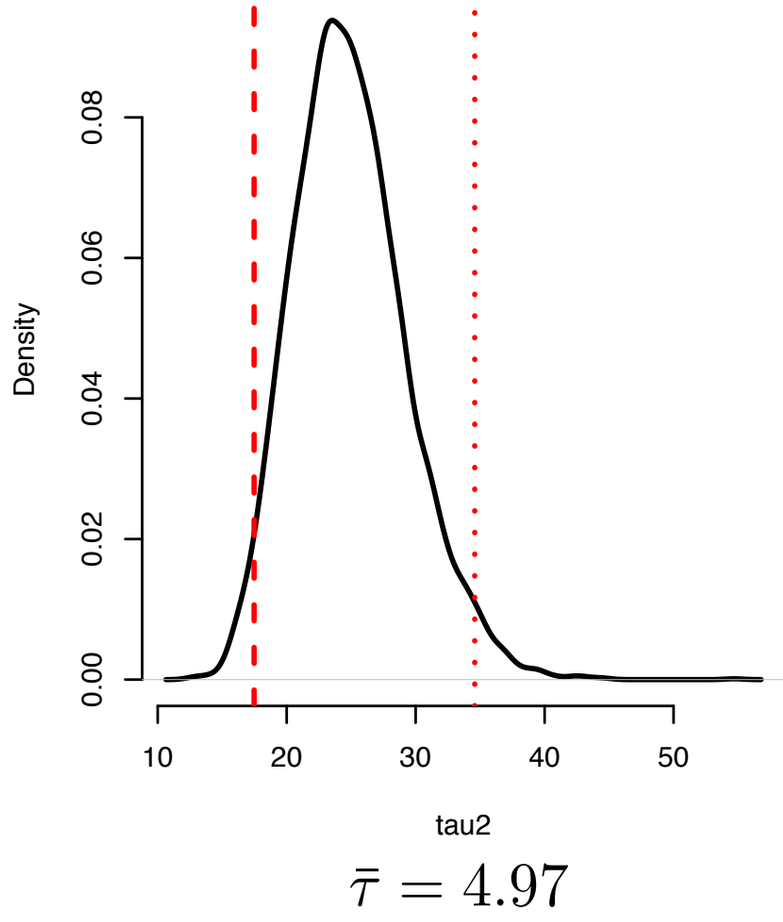
Example: Posterior summaries



$\bar{\psi} = 47.78$



$\bar{\sigma} = 9.21$



$\bar{\tau} = 4.97$

- 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

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.

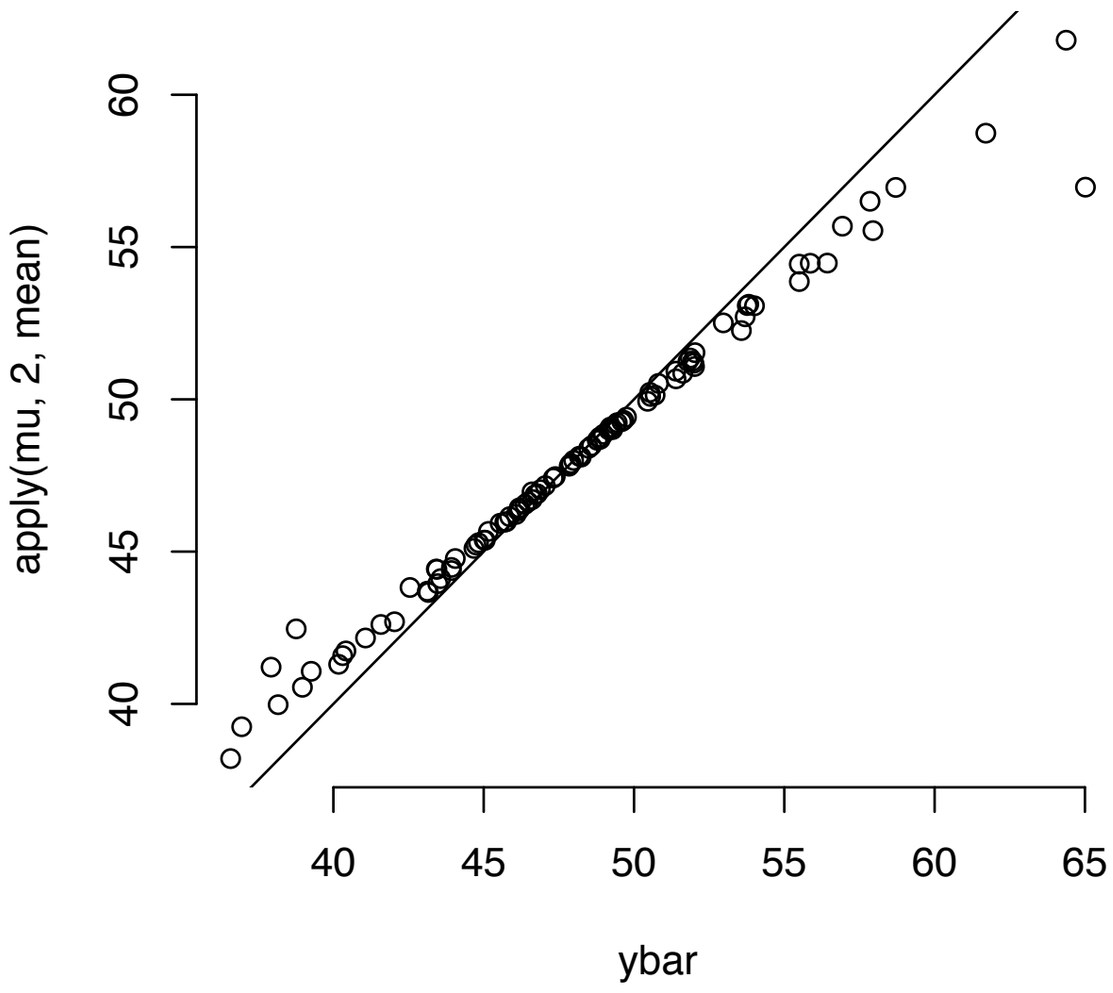
* Note: we get this formula assuming the normal model with the semi-conjugate priors

Example: Shrinkage

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

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

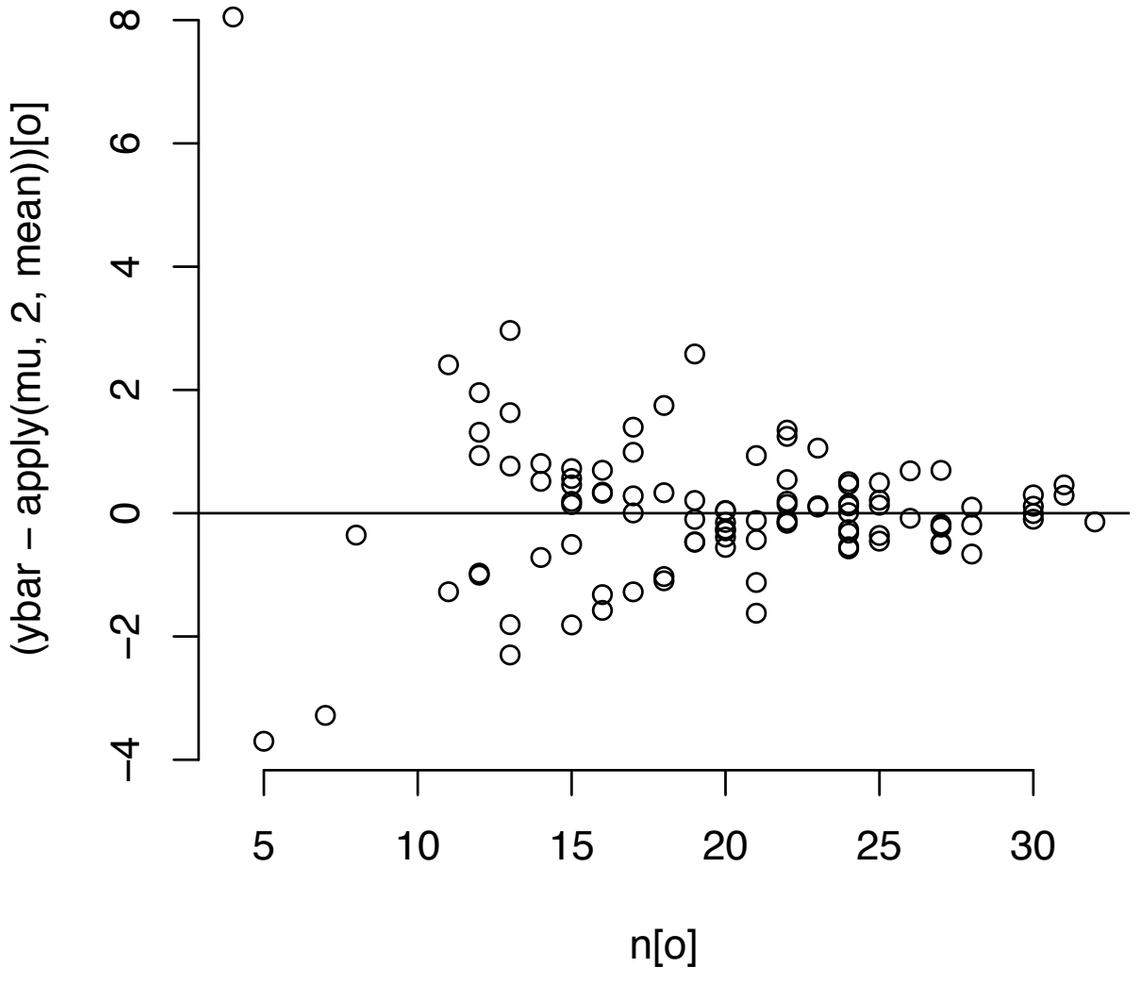
for $j = 1, \ldots, m$ obtained in via our MCMC method



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

Hierarchical binomial model

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

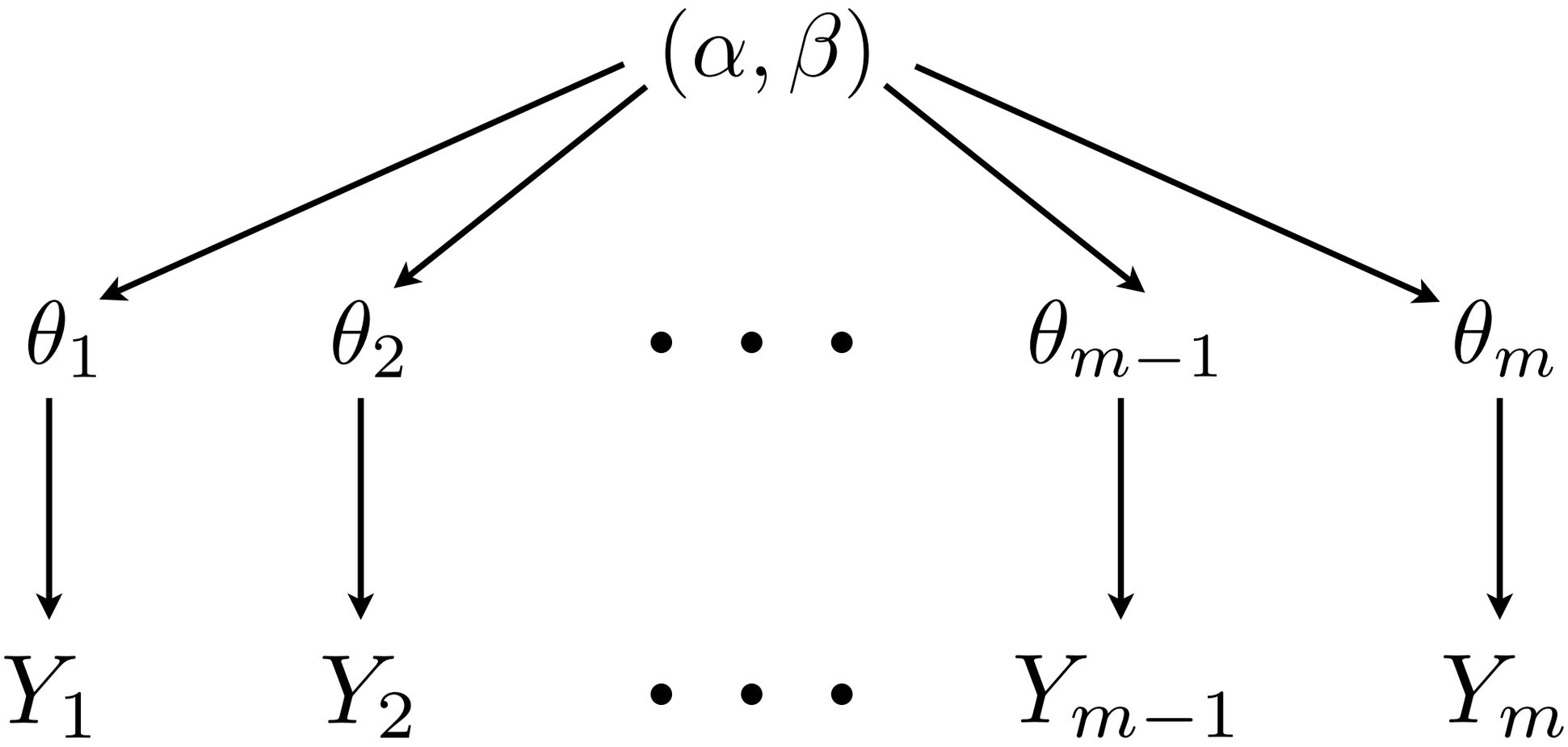
$$Y_j \sim \text{Bin}(n_j, \theta_j)$$
$$\theta_j \sim \text{Beta}(\alpha, \beta)$$
$$(\alpha, \beta) \sim p(\alpha, \beta)$$

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

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

This should look familiar!

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 sampling method:

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

Using JAGS/rjags, however, you can choose whatever, as long as the priors are restricted to values ≥ 0.

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)

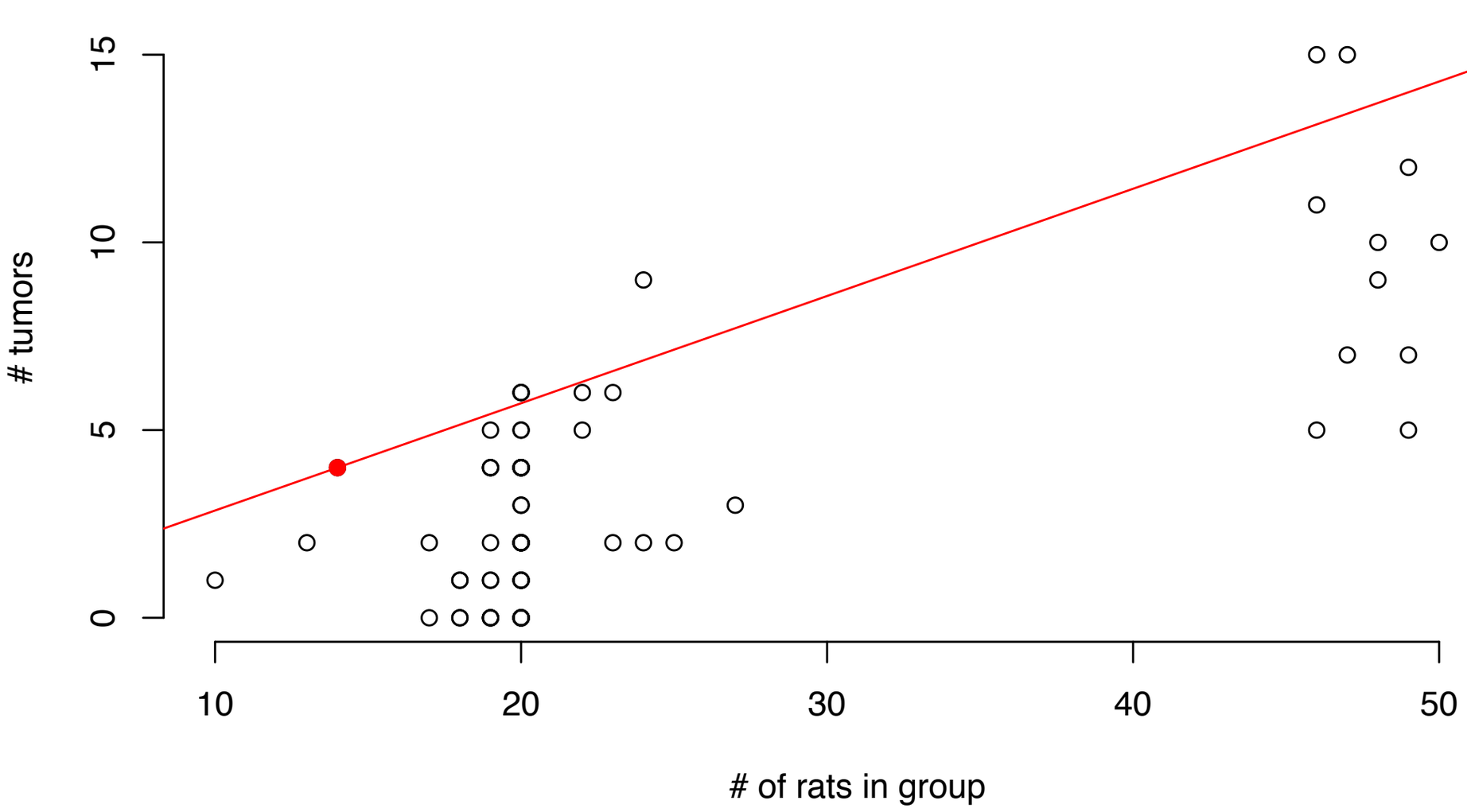
The data show that 4/14 rats developed endometrial stromal polyps (a kind of tumor)

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

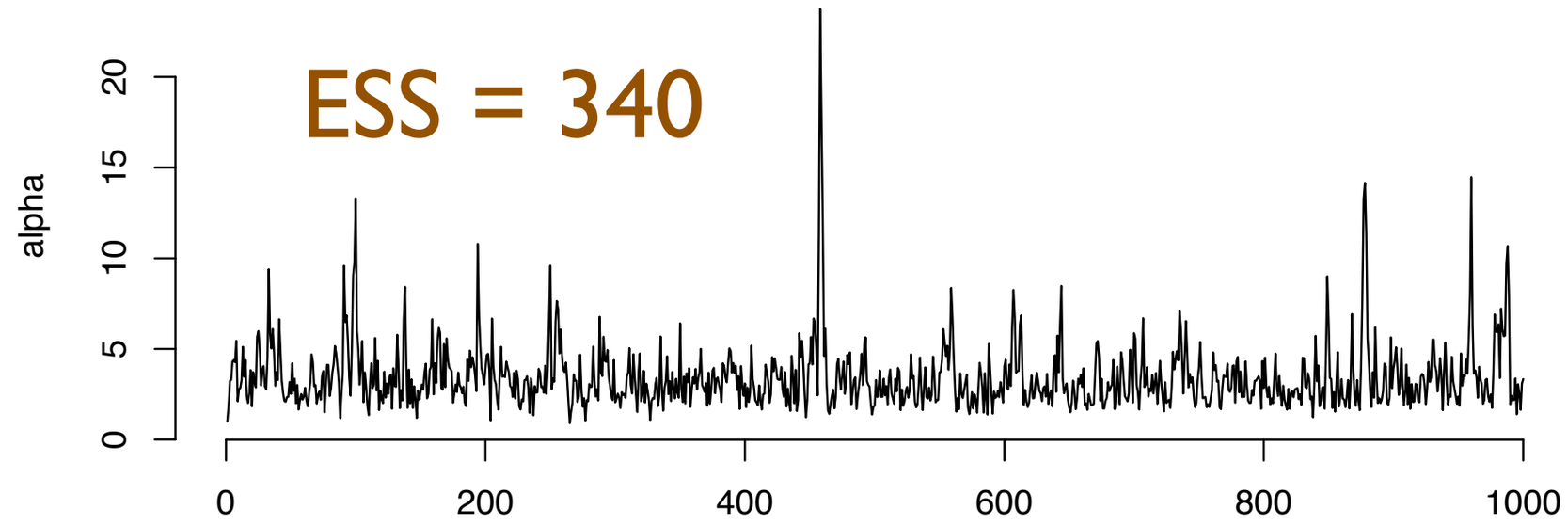
Tarone (1982) provides data on the observations of tumor incidence in 70 groups of rats

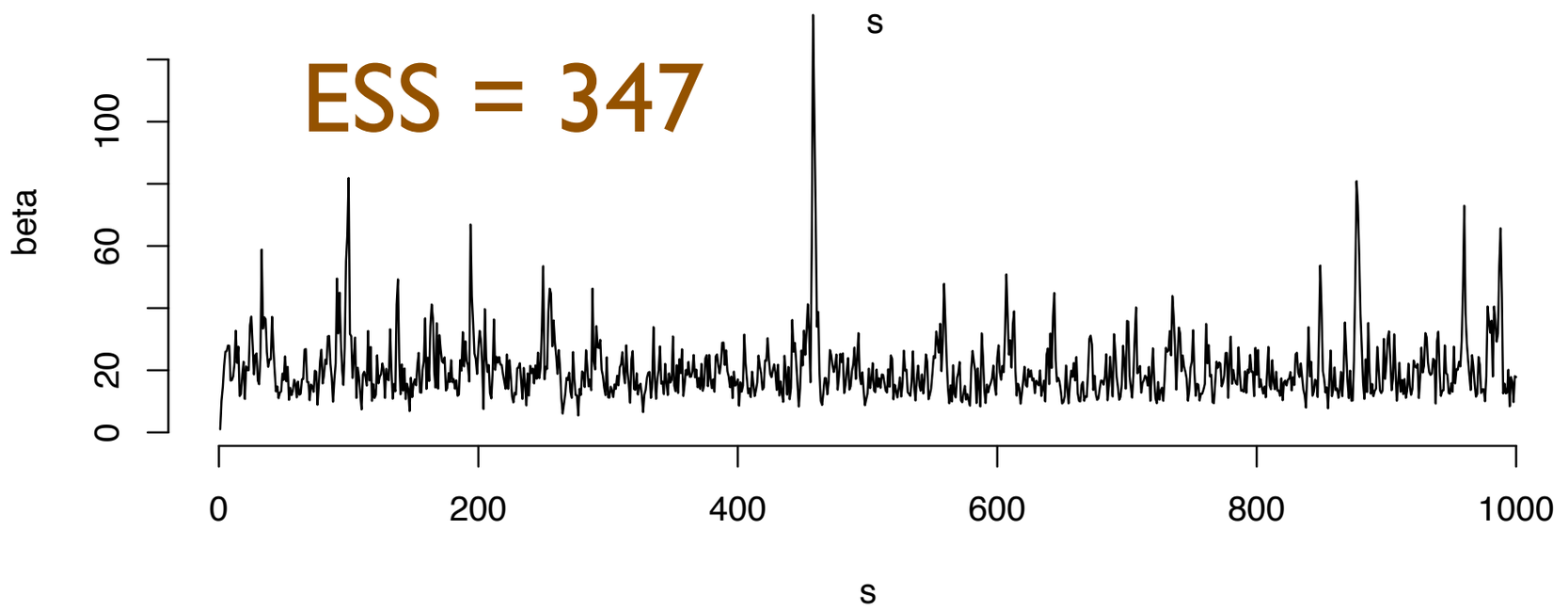


Example: Bayesian analysis

We model the $m = 71$ rat tumor data with the hierarchical Beta-binomial sampling model + joint prior

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

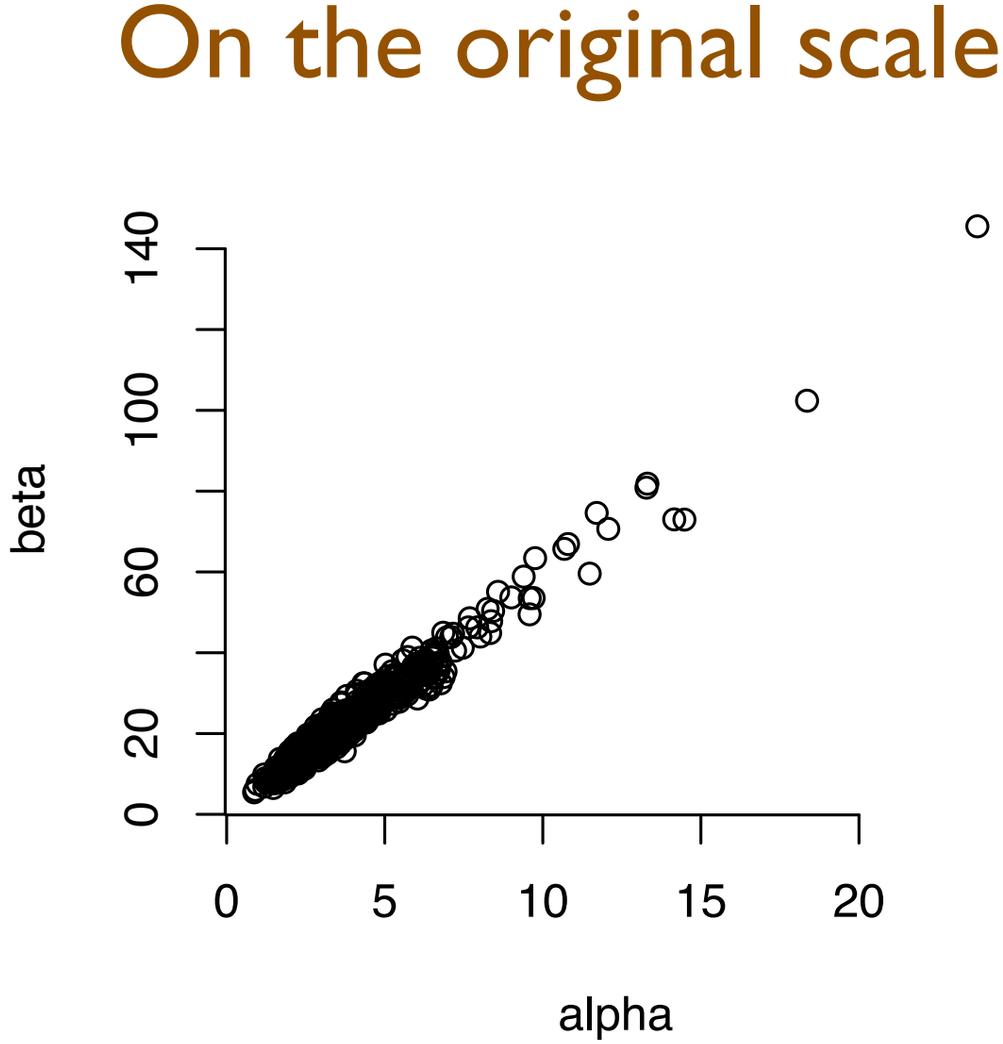


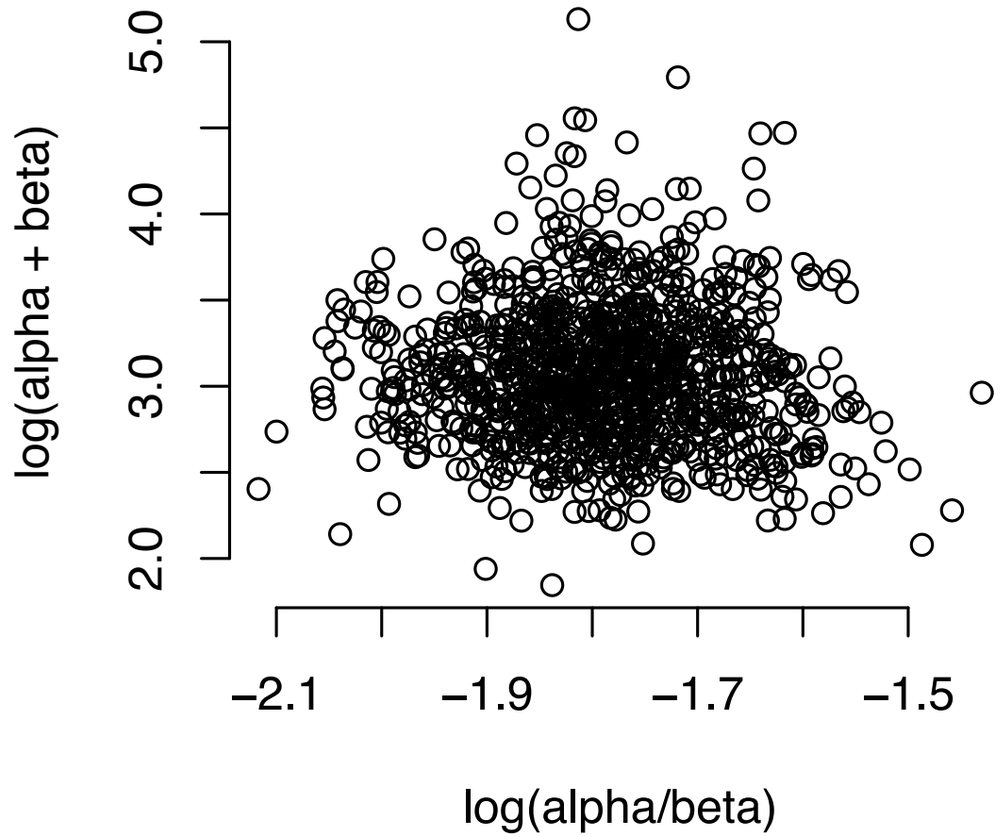


Example: The posterior marginal

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

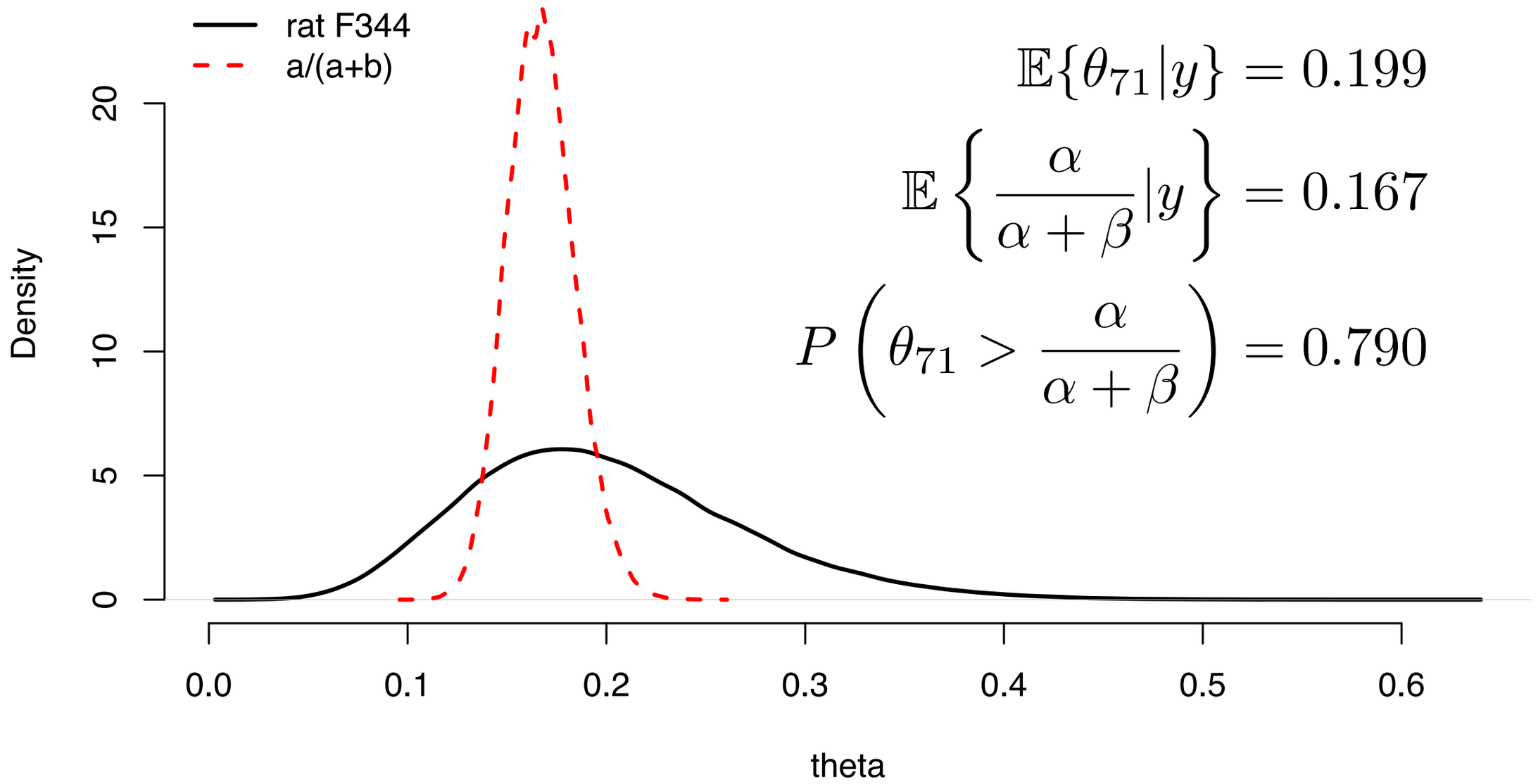




A sensible transformation

Example: Rat group F344

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



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

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

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

We have seen how Bayesian 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

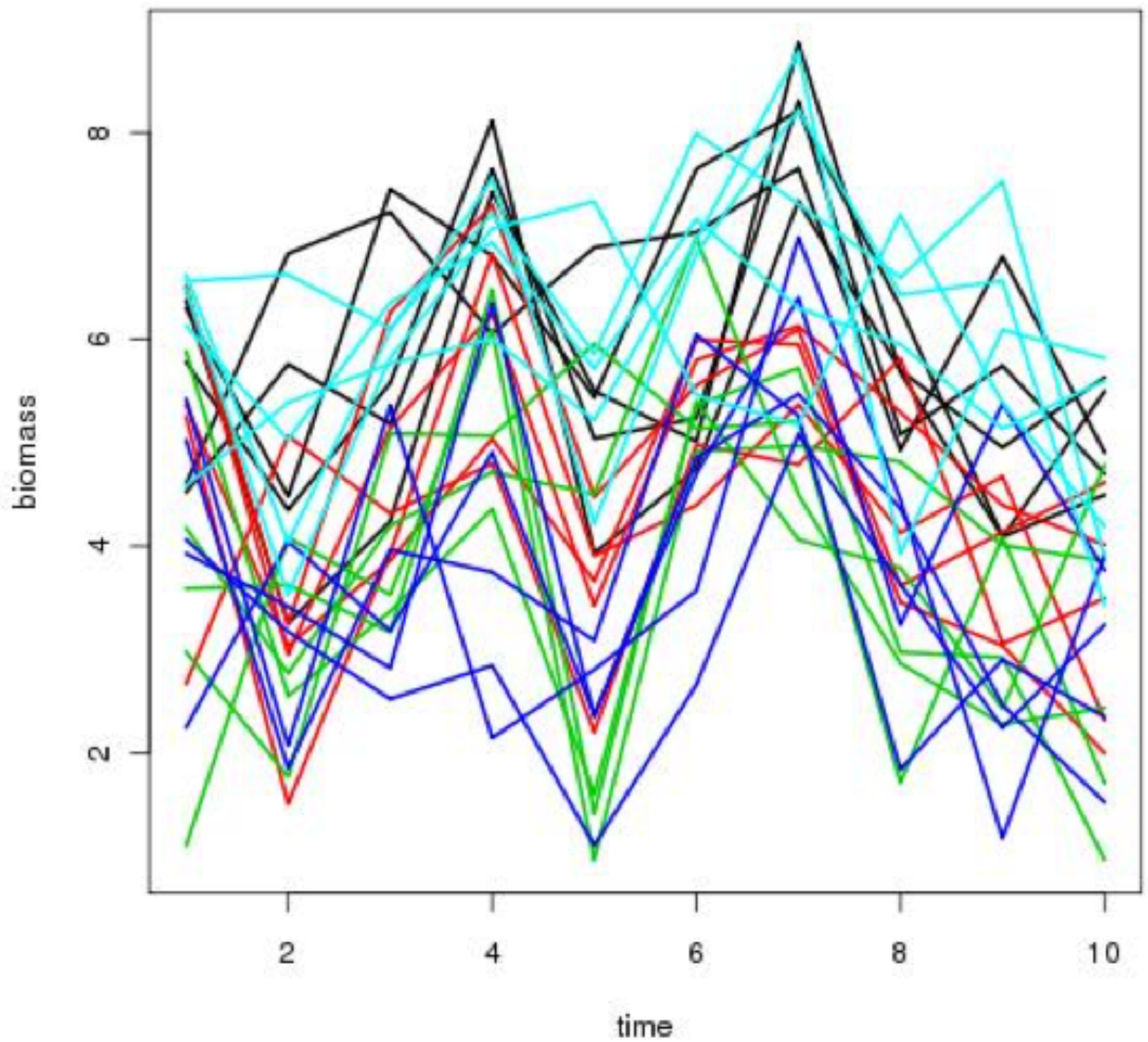
Building an Hierarchical model

When you build a HM remember:

1. Always start simple and add complexity
2. Think about at what levels you expect observations to "group" or where the variability and randomness comes from
3. Use your posterior samples carefully to obtain inferences for the level of your hierarchy that you're most interested in.

But what about forecasting?!?!?

Example: Biomass by Block and Time



Note that the following code is illustrative and does not necessarily represent exact JAGS syntax

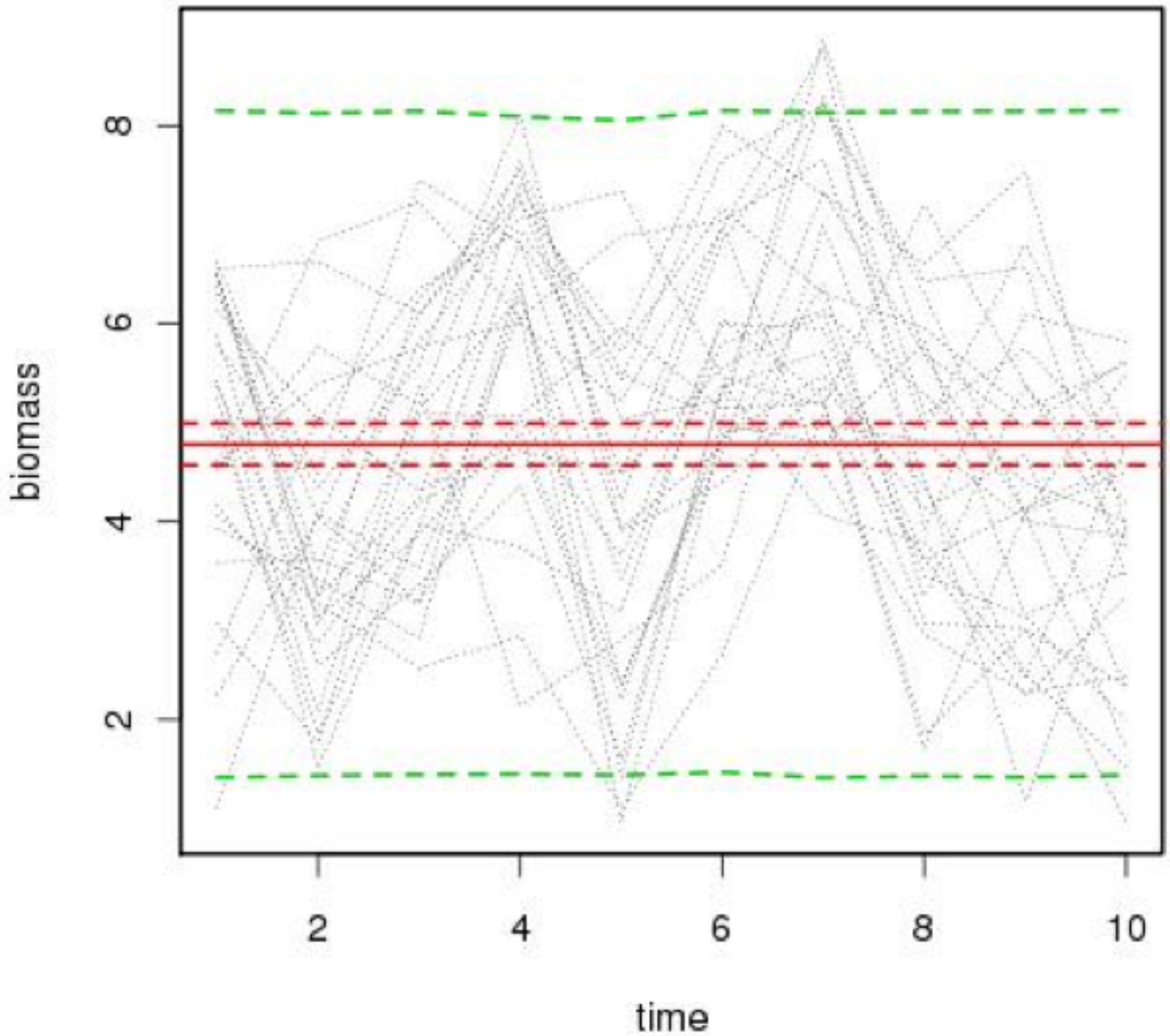
Model 1: Global Mean

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

```
model{
  ## priors
  mu ~ dnorm(0, 0.001)
  sigma ~ dgamma(0.001, 0.001)

  ## the likelihood
  ## remember to loop over all obs!
  for(t in 1:nt){ ## time
    for(b in 1:nb){ ## blocks
      for(i in 1:nrep){ ## obs in a block
        x[t,b,i] ~ dnorm(mu, sigma)
      }
    }
  }
}
```

Model 1: Global Mean



Model 2: Time Varying Mean

$$X_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(\alpha_t, \sigma^2)$$

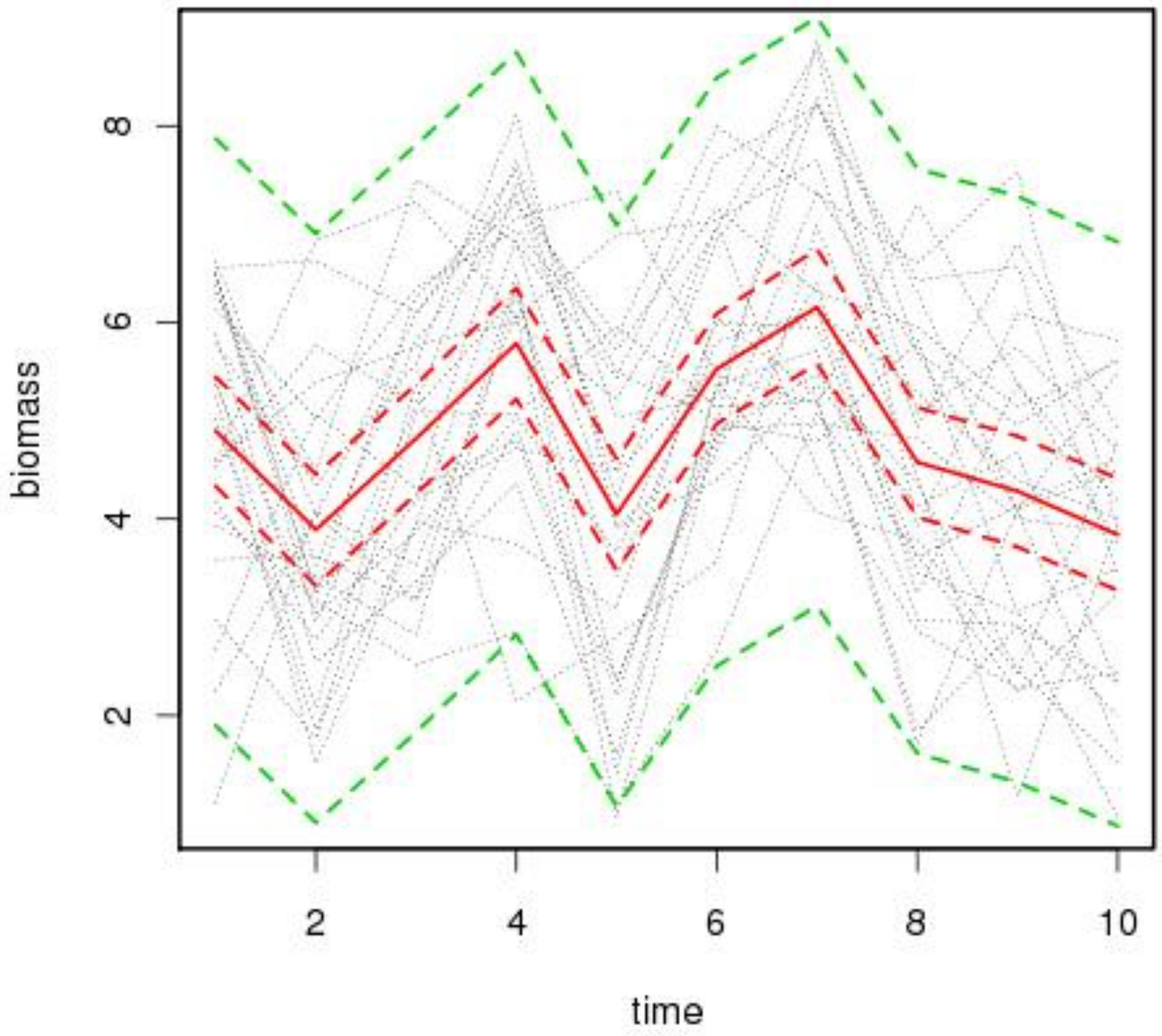
$$\alpha_t \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau^2)$$

```
model{
  ## priors
  sigma ~ dgamma(0.001, 0.001)

  ## hyperpriors
  mu ~ dnorm(0, 0.001)
  tau ~ dgamma(0.001, 0.001)

  ## the likelihood
  ## remember to loop over all obs!
  for(t in 1:nt){ ## time
    alpha.t[t] ~ dnorm(mu, tau) ## random draw for each time
    for(b in 1:nb){ ## blocks
      for(i in 1:nrep){ ## obs in a block
        x[t,b,i] ~ dnorm(alpha.t[t], sigma) ## time dept alpha
      }
    }
  }
}
```

Model 2: Time Varying Mean



Model 3: Grouping by Block

$$X_{i,b} \overset{\text{iid}}{\sim} \mathcal{N}(\alpha_b, \sigma^2)$$

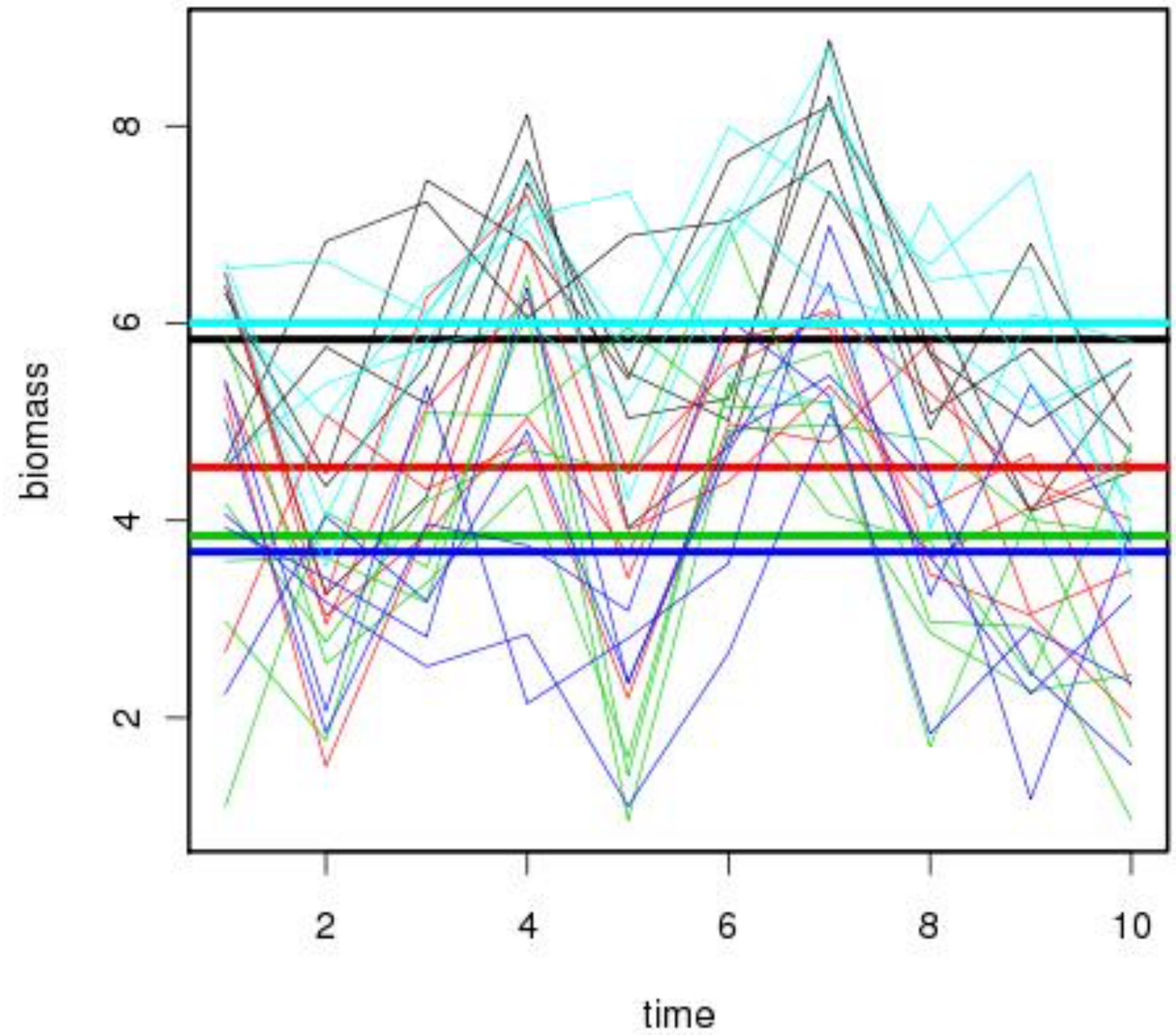
$$\alpha_b \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau_b^2)$$

```
model{
  ## priors
  sigma ~ dgamma(0.001, 0.001)

  ## hyperpriors
  mu ~ dnorm(0, 0.001)
  tauB ~ dgamma(0.001, 0.001)

  ## the likelihood
  ## remember to loop over all obs!
  for(b in 1:nb){ ## blocks
    alpha.b[b] ~ dnorm(mu, tauB) ## random draw for each block
    for(t in 1:nt){ ## time
      for(i in 1:nrep){ ## obs in a block
        x[t,b,i] ~ dnorm(alpha.B[b], sigma) ## block dept alpha
      }
    }
  }
}
```

Model 3: Grouping by Block



Model 4: Grouping by Block and Time!

$$X_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(\alpha_b + \alpha_t, \sigma^2)$$

$$\alpha_b \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau_b^2)$$

$$\alpha_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_t^2)$$

Model 5: Grouping by Block and Time + trend!!

$$X_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(\alpha_b + \alpha_t + \beta t, \sigma^2)$$

$$\alpha_b \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau_b^2)$$

$$\alpha_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_t^2)$$

This process can go on to infinity…

…so remember models represent hypotheses…

1. The proper accounting of uncertainty can be just as important for making valid inference from your model as the process model and covariates
2. BHM allows you to take into account unmeasured covariates/processes
3. But have your hypotheses worked out before you do extensive model building — don't go fishing!

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 Gibbs Sampling (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

WinBUGs, JAGS....

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

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

There are many other possibilities.

A hyperprior choice

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

It may be shown that it gives a "roughly uniform" prior distribution to the log standard deviation of the resulting $\theta_j \sim \text{Beta}(\alpha, \beta)$ sampling model

It may also be shown that the posterior marginal

$$p(\alpha, \beta | y)$$

is proper with this choice as long as $0 < y_j < n_j$ for at least one experiment j