Chapter 5

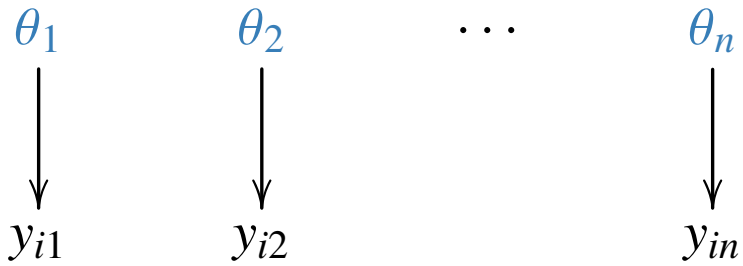
- 5.1 Lead-in to hierarchical models
- 5.2 Exchangeability (useful concept)
- 5.3 Bayesian analysis of hierarchical models (we use Stan/brms for computation)
- 5.4 Hierarchical normal model (we use Stan/brms for computation)
- 5.5 Example: parallel experiments in eight schools (useful discussion on benefits of hierarchical model)
- 5.6 Meta-analysis (can be skipped)
- 5.7 Weakly informative priors for hierarchical variance parameters

Hierarchical model

- In simple model: posterior for the parameters
- In hierarchical model: posterior for the prior parameters

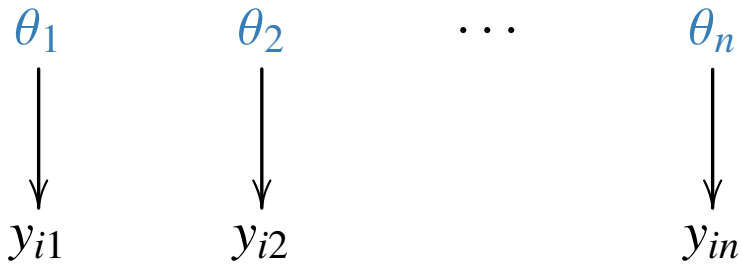
Hierarchical model

- Example: CVD treatment effectiveness
  - in hospital j the survival probability is θ_j
  - observations y_{ij} tell whether patient i survived in hospital j

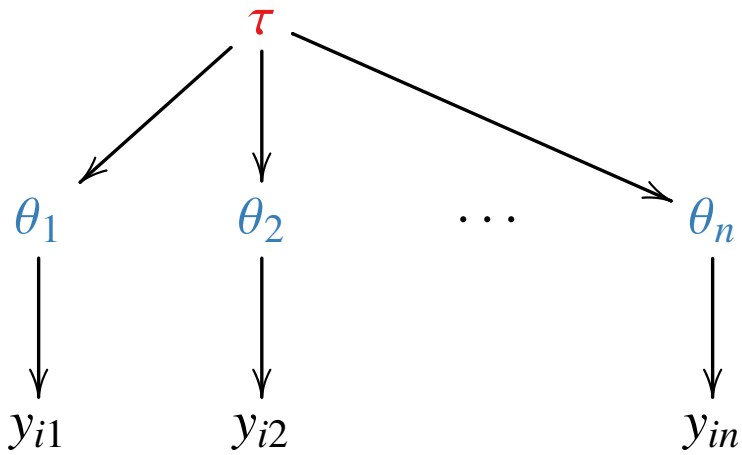


Hierarchical model

- Example: CVD treatment effectiveness
  - in hospital j the survival probability is θ_j
  - observations y_{ij} tell whether patient i survived in hospital j



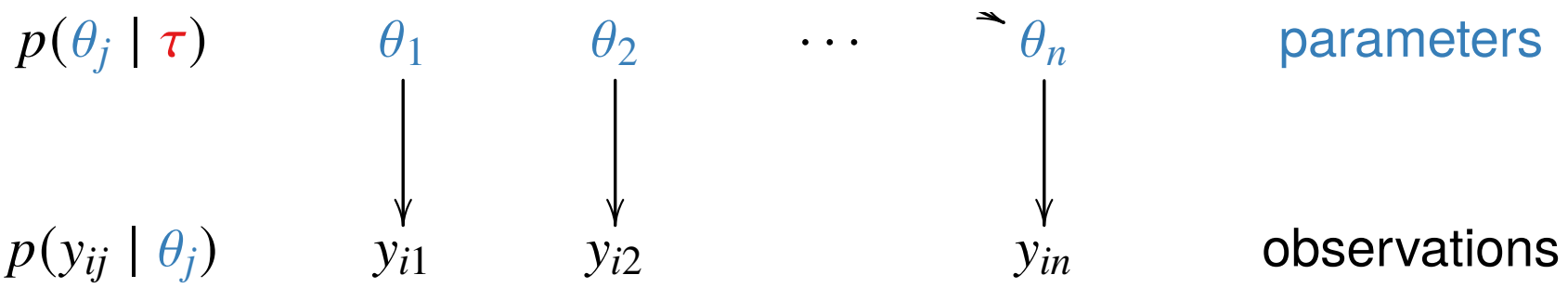
  - sensible to assume that θ_j are similar



  - natural to think that θ_j have common population distribution
  - θ_j is not directly observed and the population distribution is unknown

Hierarchical model: terms

Level 1: observations given parameters $p(y_{ij} \mid \theta_j)$



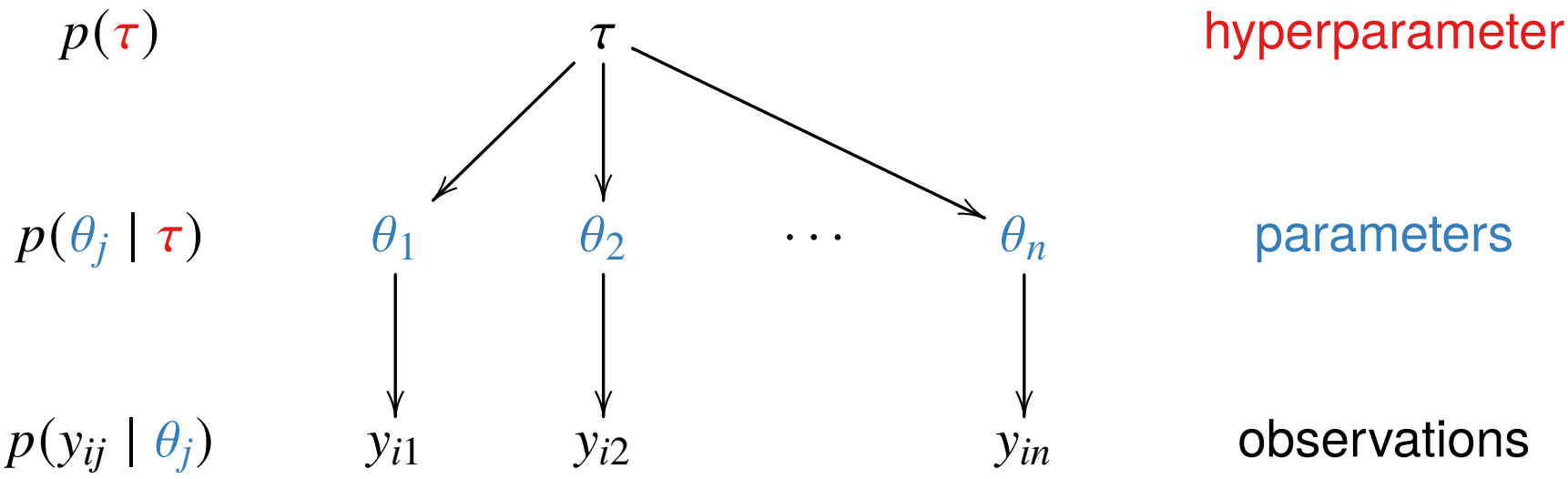
Joint posterior

$$\begin{aligned} p(\theta, \tau \mid y) &\propto p(y \mid \theta, \tau) p(\theta, \tau) \\ &\propto p(y \mid \theta) p(\theta \mid \tau) p(\tau) \end{aligned}$$

Hierarchical model: terms

Level 1: observations given parameters $p(y_{ij} \mid \theta_j)$

Level 2: parameters given hyperparameters $p(\theta_j \mid \tau)$

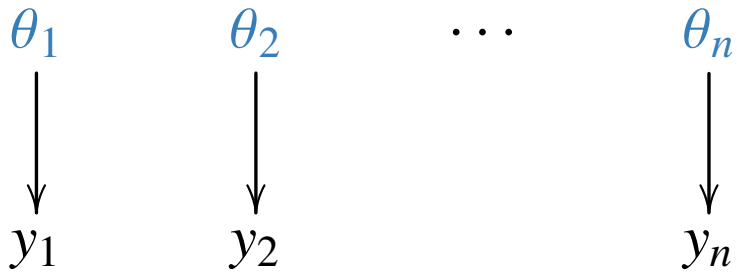


Joint posterior

$$\begin{aligned} p(\theta, \tau \mid y) &\propto p(y \mid \theta, \tau) p(\theta, \tau) \\ &\propto p(y \mid \theta) p(\theta \mid \tau) p(\tau) \end{aligned}$$

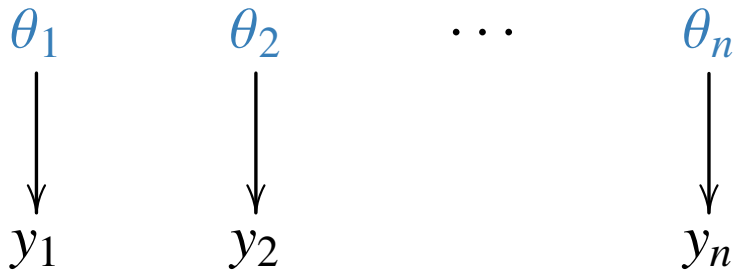
Compare

- "Separate model" (model with separate/independent effects)

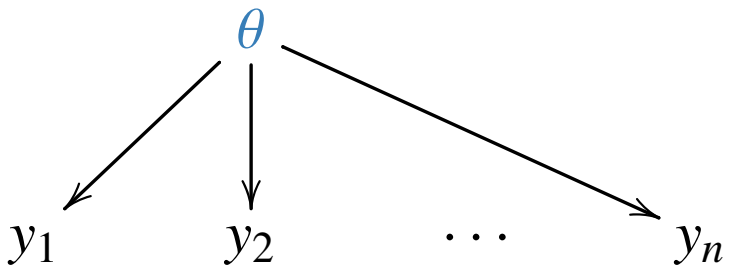


Compare

- "Separate model" (model with separate/independent effects)

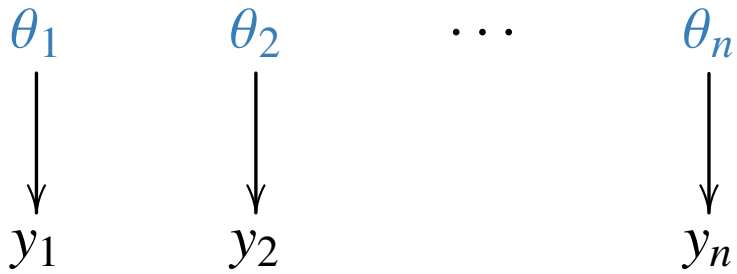


- "Joint model" (model with a common effect / pooled model)

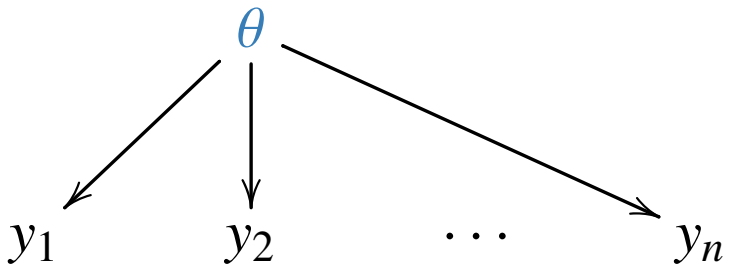


Compare

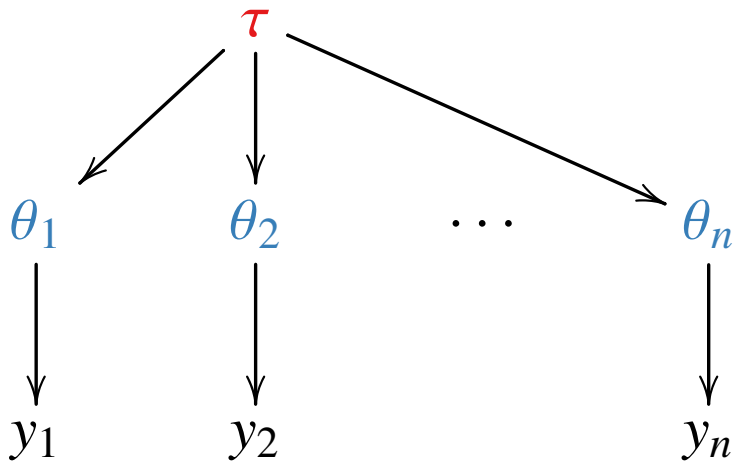
- "Separate model" (model with separate/independent effects)



- "Joint model" (model with a common effect / pooled model)



- Hierarchical model



Hierarchical binomial model: rats

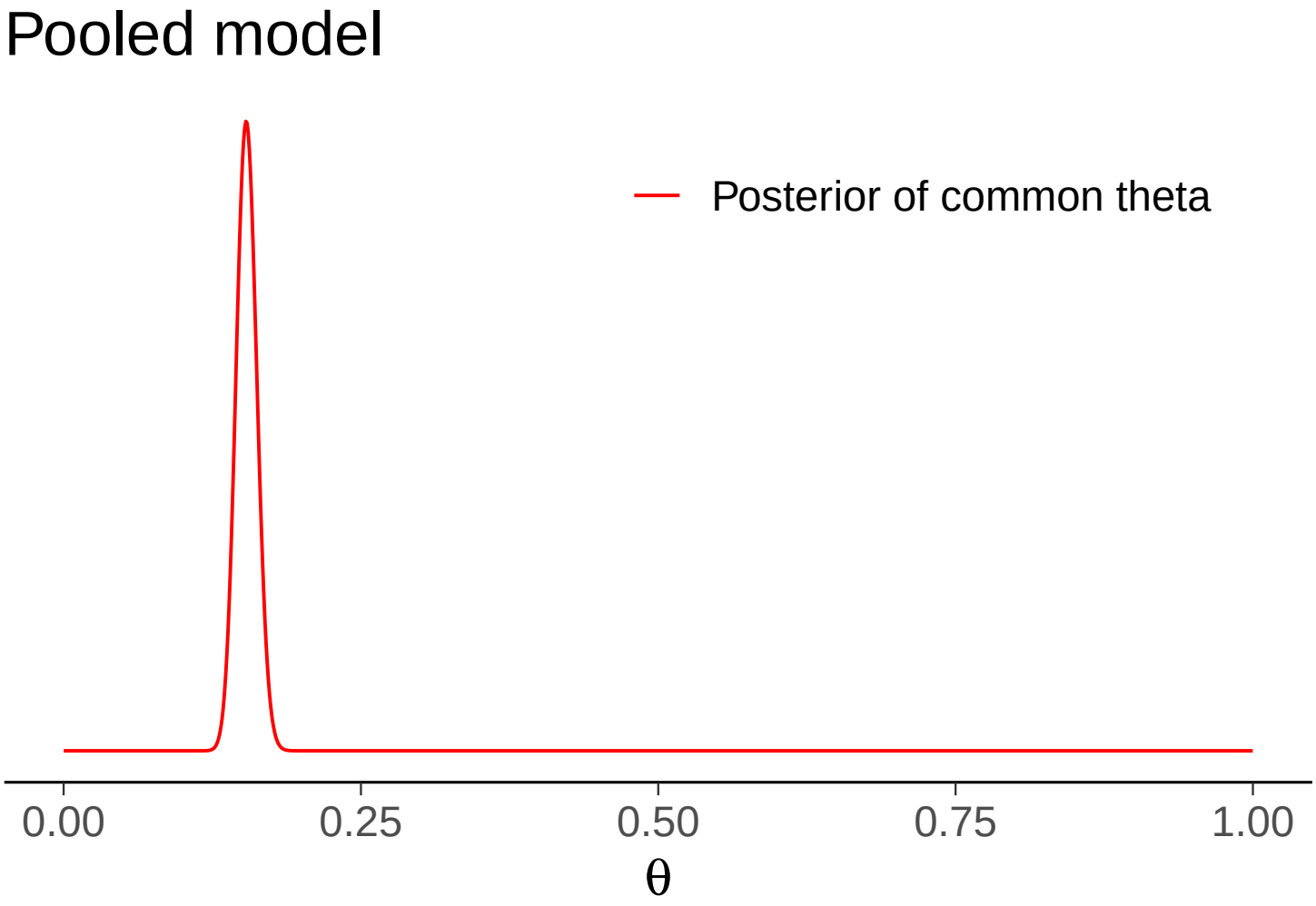
- Medicine testing
- Type F344 female rats in control group given placebo
  - count how many get endometrial stromal polyps
  - familiar binomial model example

Hierarchical binomial model: rats

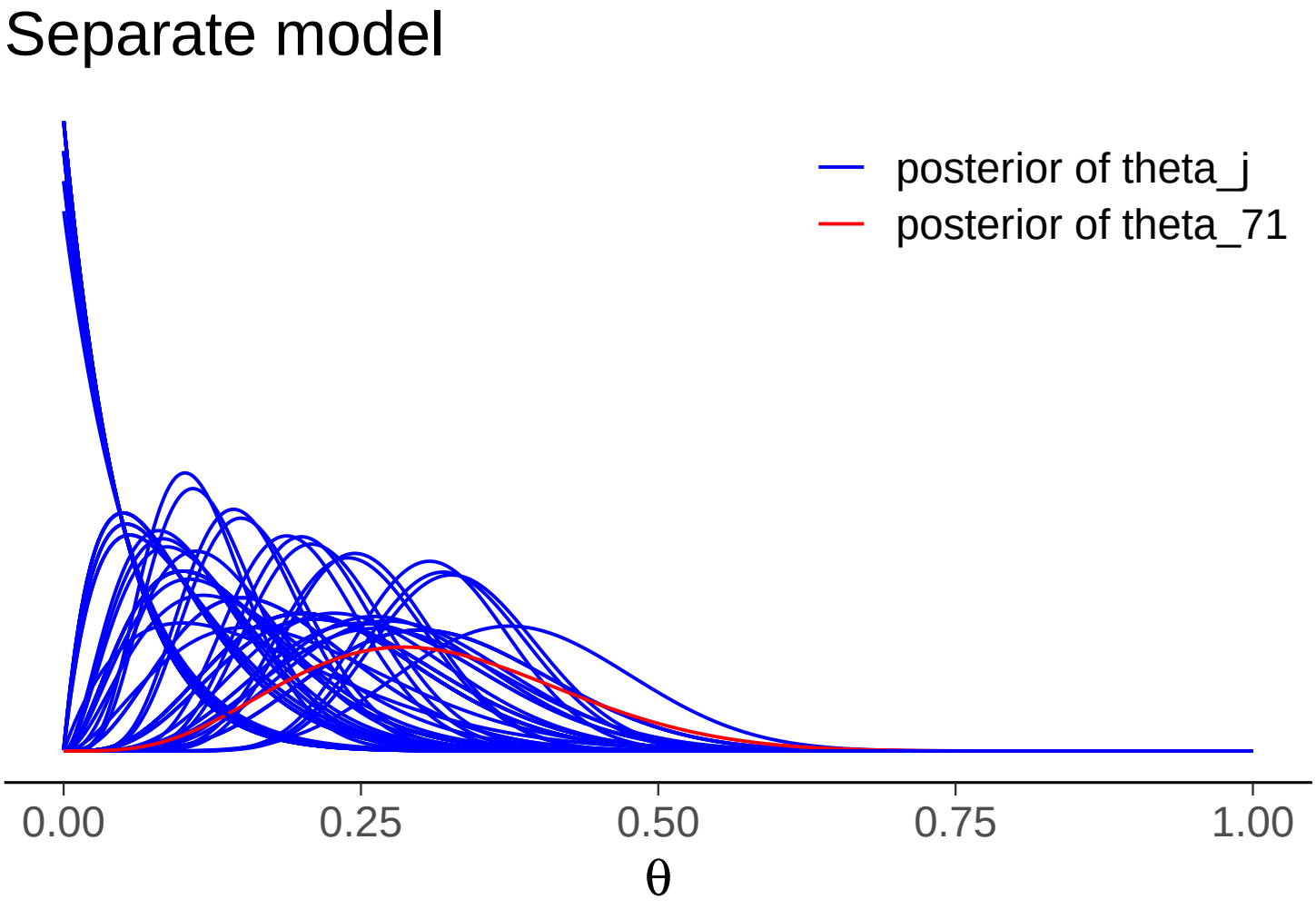
- Medicine testing
- Type F344 female rats in control group given placebo
  - count how many get endometrial stromal polyps
  - familiar binomial model example
- Experiment has been repeated 71 times

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/46	15/47	9/24
4/14									

Hierarchical binomial model: rats



Hierarchical binomial model: rats

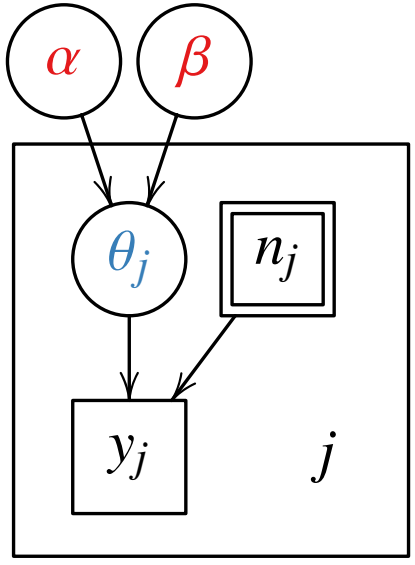


Hierarchical binomial model: rats

- Hierarchical binomial model for rats
  prior parameters α and β are unknown

$$\theta_j \mid \alpha, \beta \sim \text{Beta}(\theta_j \mid \alpha, \beta)$$

$$y_j \mid n_j, \theta_j \sim \text{Bin}(y_j \mid n_j, \theta_j)$$



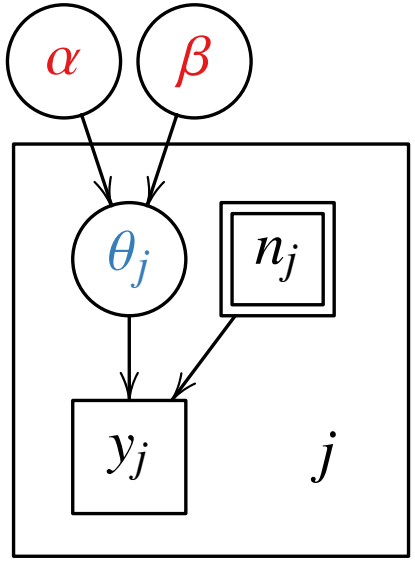
- Joint posterior $p(\theta_1, \ldots, \theta_J, \alpha, \beta \mid y)$
  - multiple parameters

Hierarchical binomial model: rats

- Hierarchical binomial model for rats
  prior parameters α and β are unknown

$$\theta_j \mid \alpha, \beta \sim \text{Beta}(\theta_j \mid \alpha, \beta)$$

$$y_j \mid n_j, \theta_j \sim \text{Bin}(y_j \mid n_j, \theta_j)$$

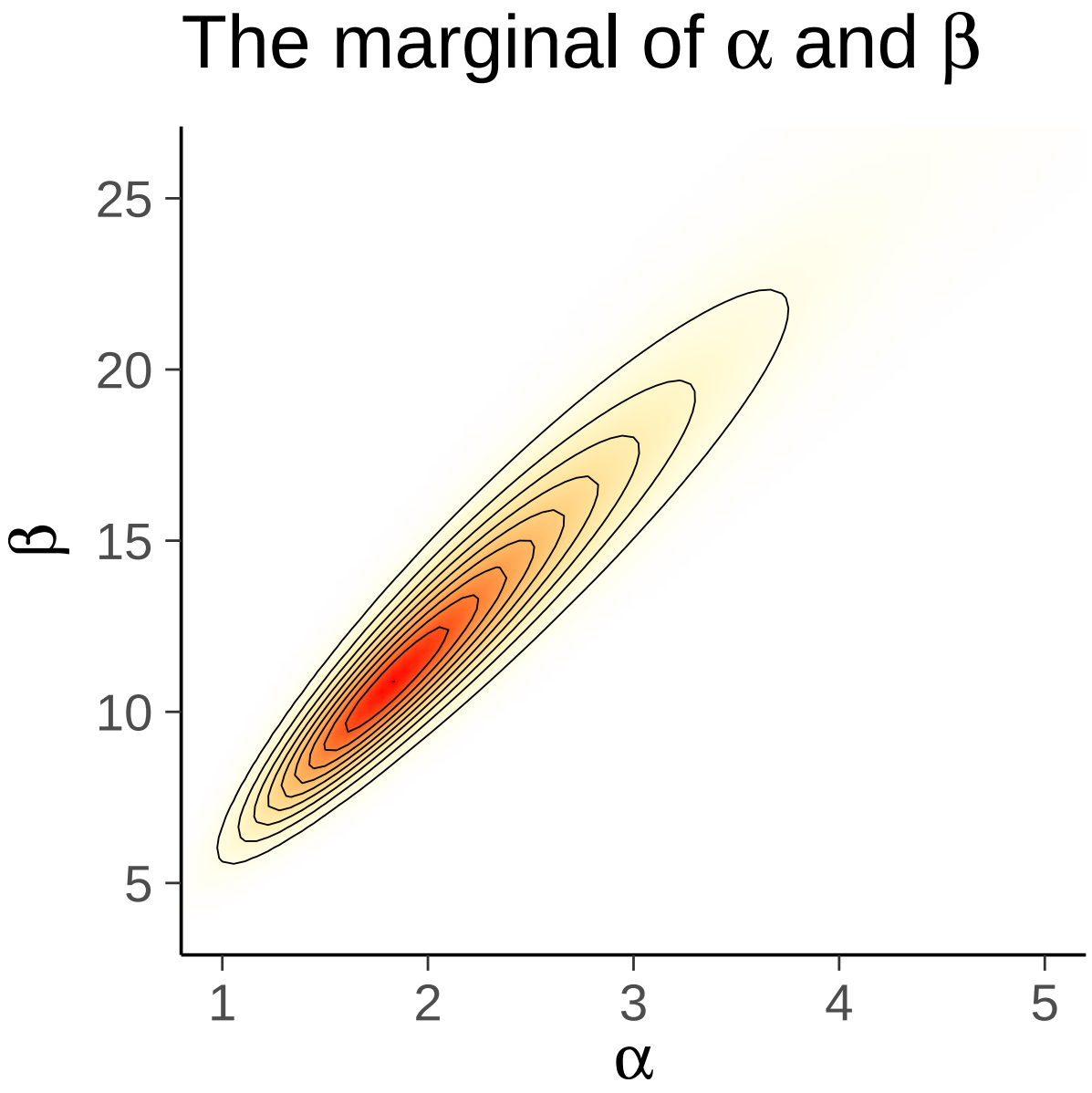


- Joint posterior $p(\theta_1, \ldots, \theta_J, \alpha, \beta \mid y)$
  - multiple parameters
  - factorize $\prod_{j=1}^{J} p(\theta_j \mid \alpha, \beta, y) p(\alpha, \beta \mid y)$

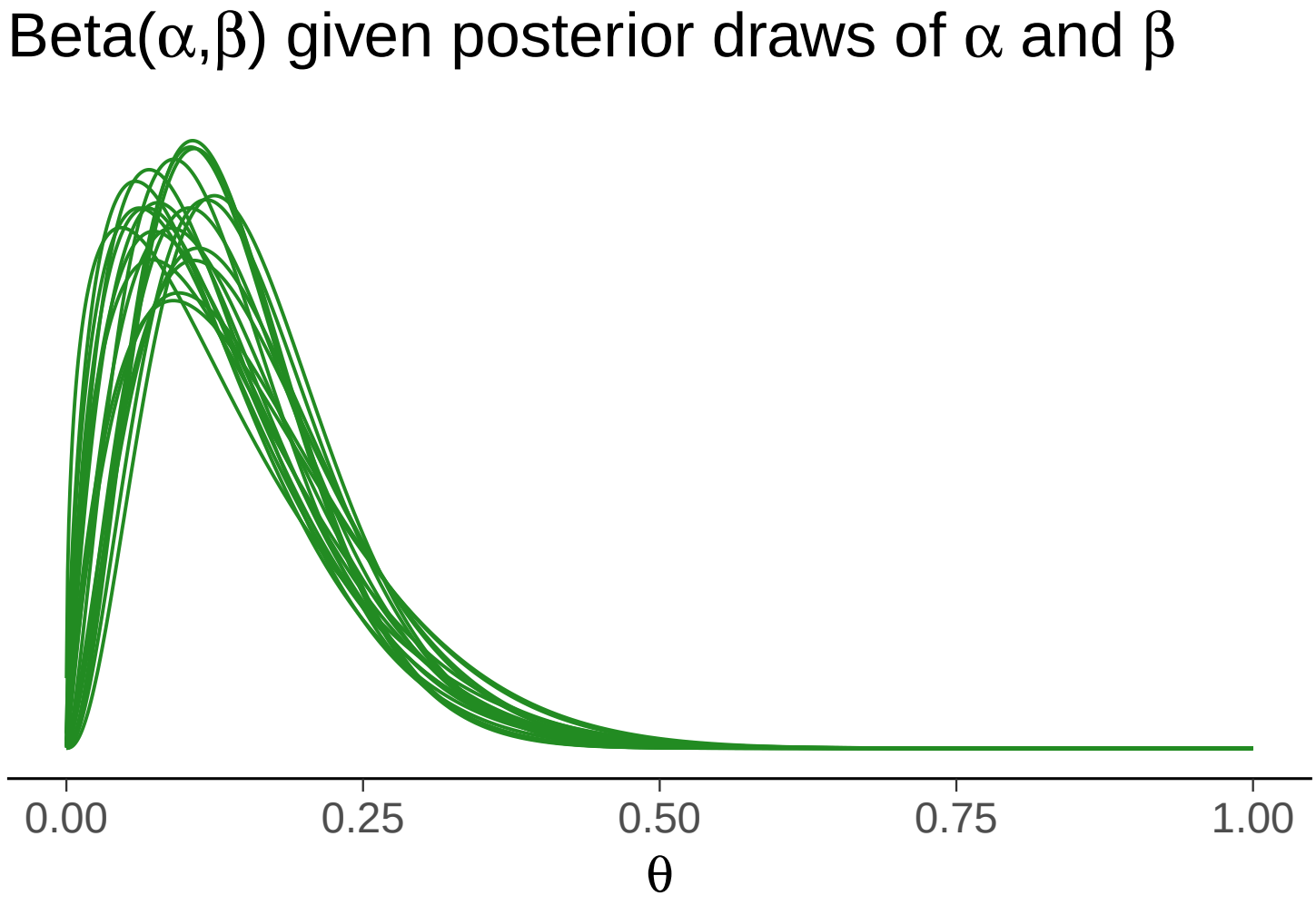
Hierarchical binomial model: rats

- Population prior $\mathrm{Beta}(\theta_j \mid \alpha, \beta)$
- Hyperprior $p(\alpha, \beta)$?
  - α, β both affect the location and scale
  - BDA3 has $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$
    - diffuse prior for location and scale (BDA3 p. 110)
- demo5_1

Hierarchical binomial model: rats

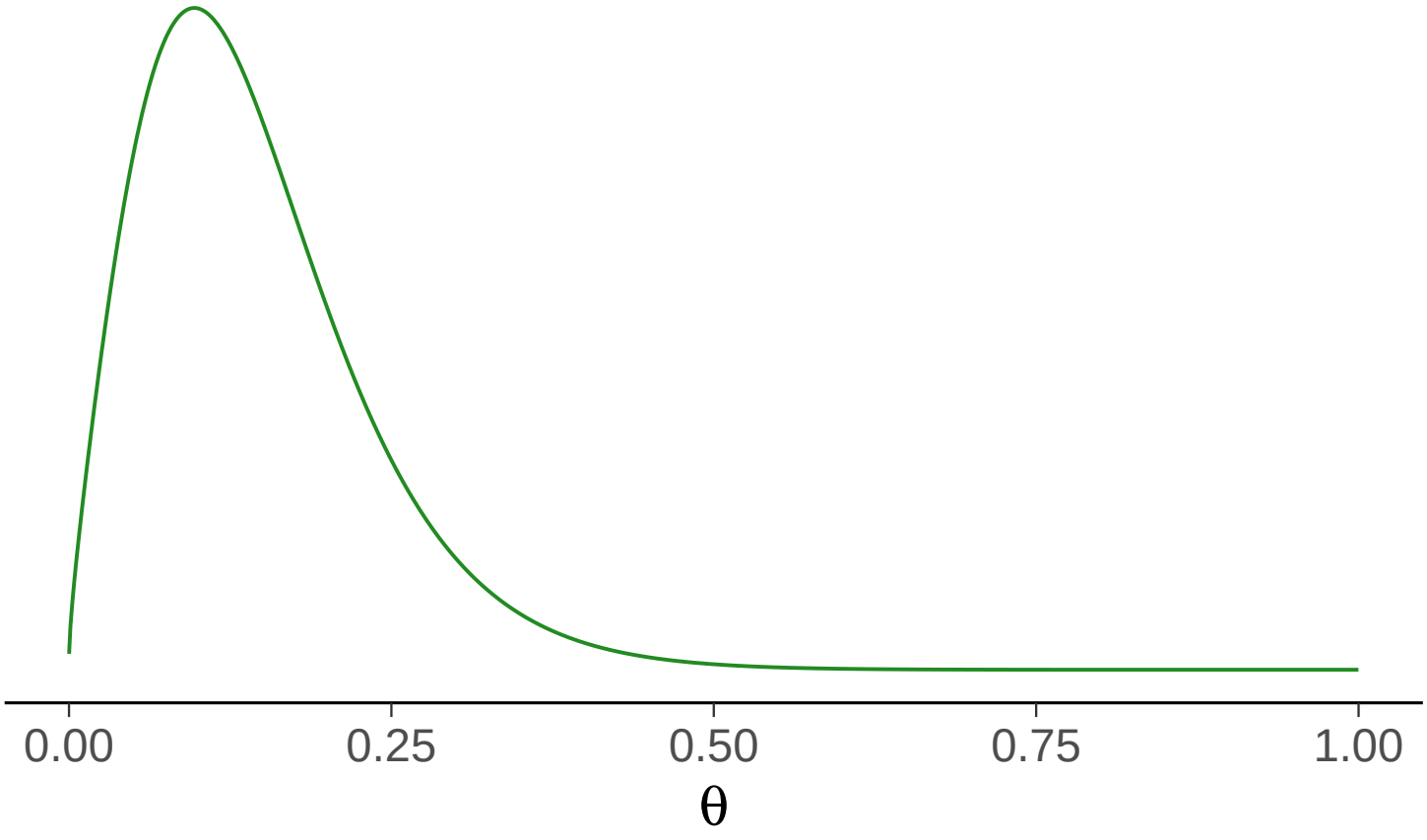


Hierarchical binomial model: rats

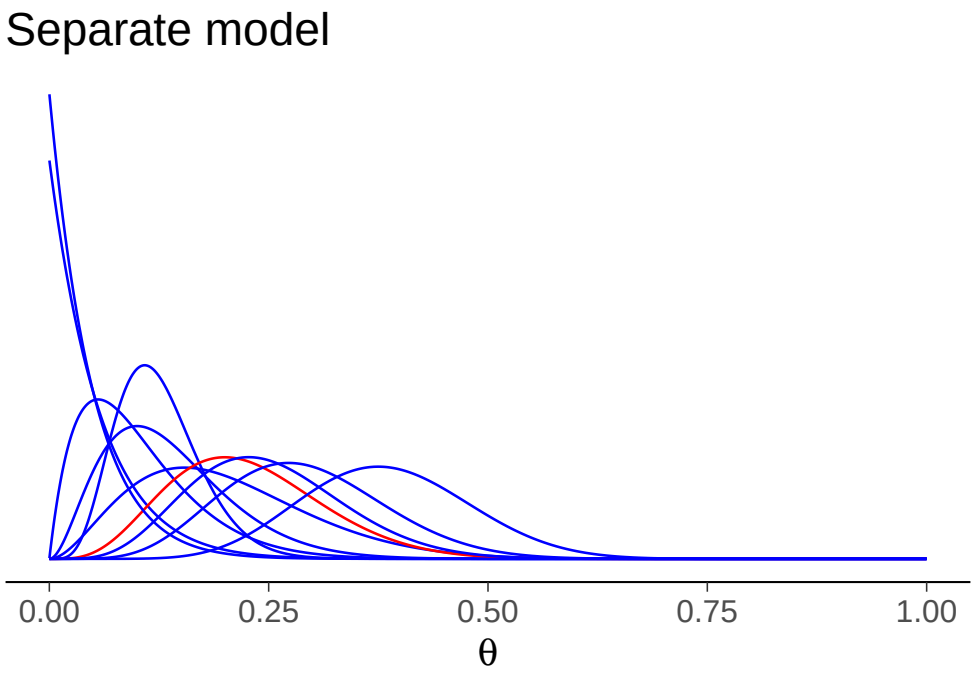


Hierarchical binomial model: rats

Population distribution (prior) for θ_j

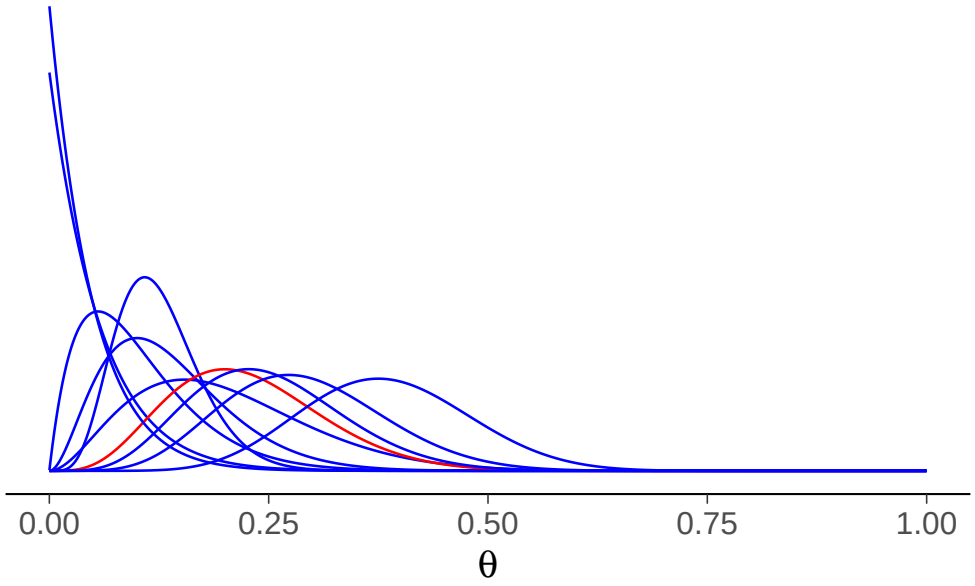


Hierarchical binomial model: rats

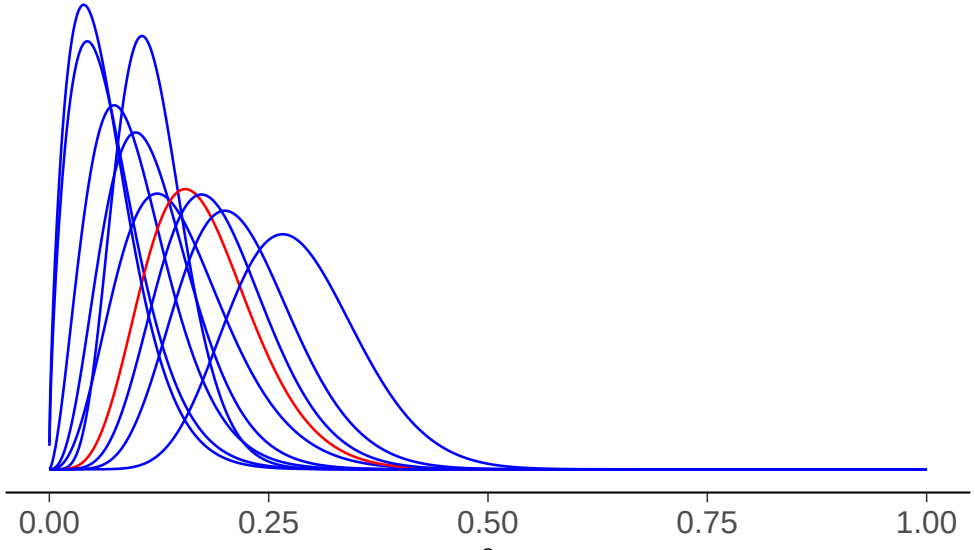


Hierarchical binomial model: rats

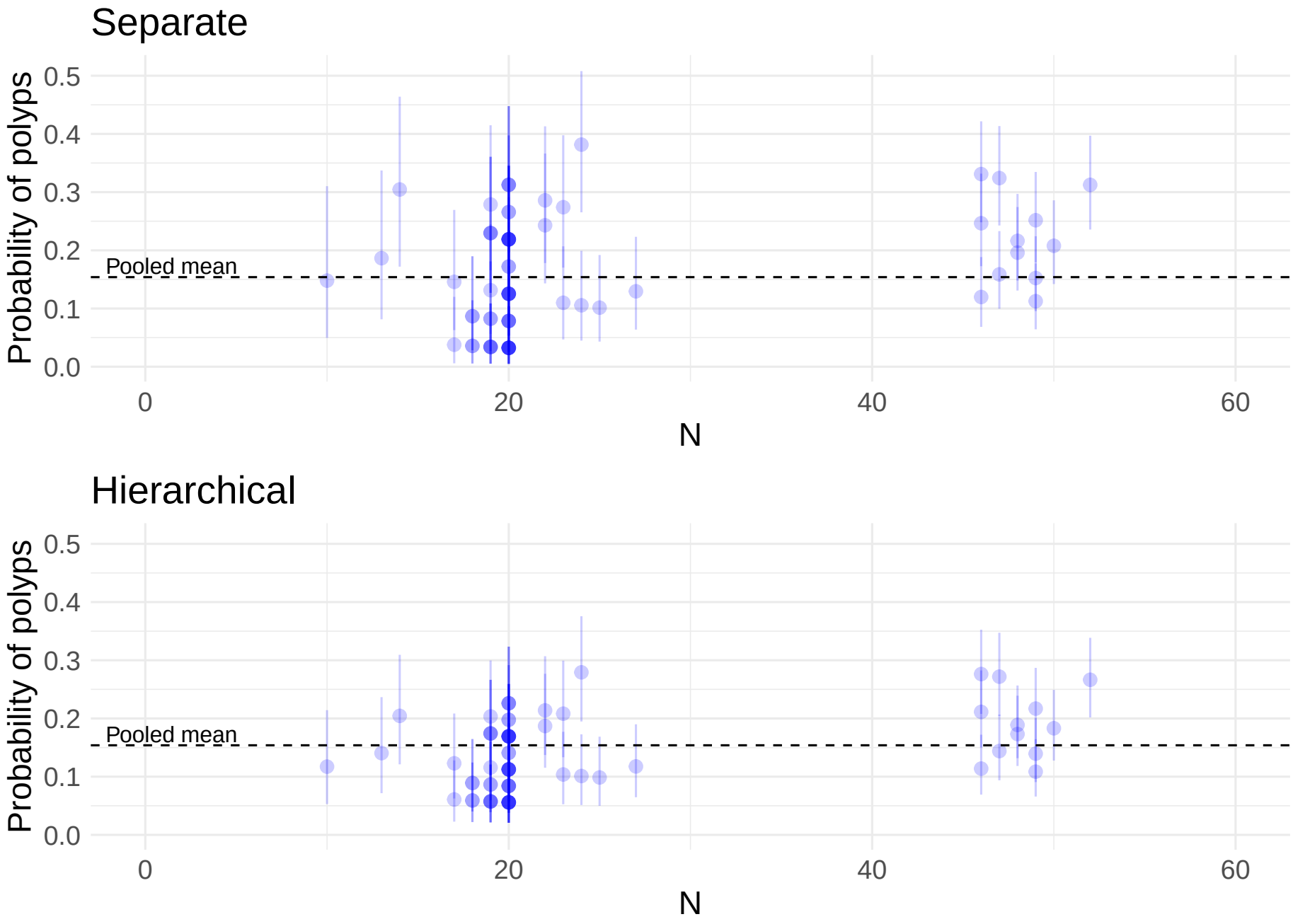
Separate model



Hierarchical model

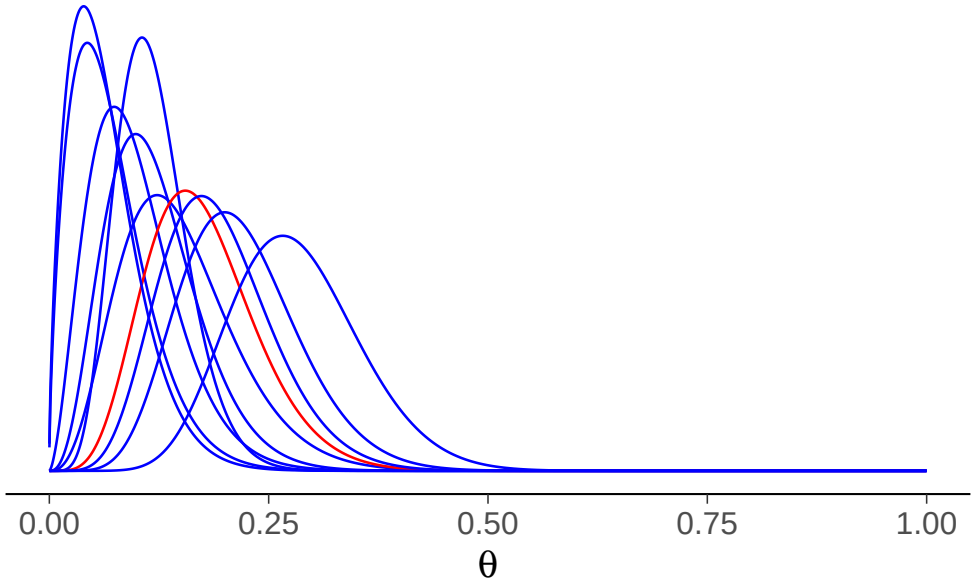


Hierarchical model and group size: Rats

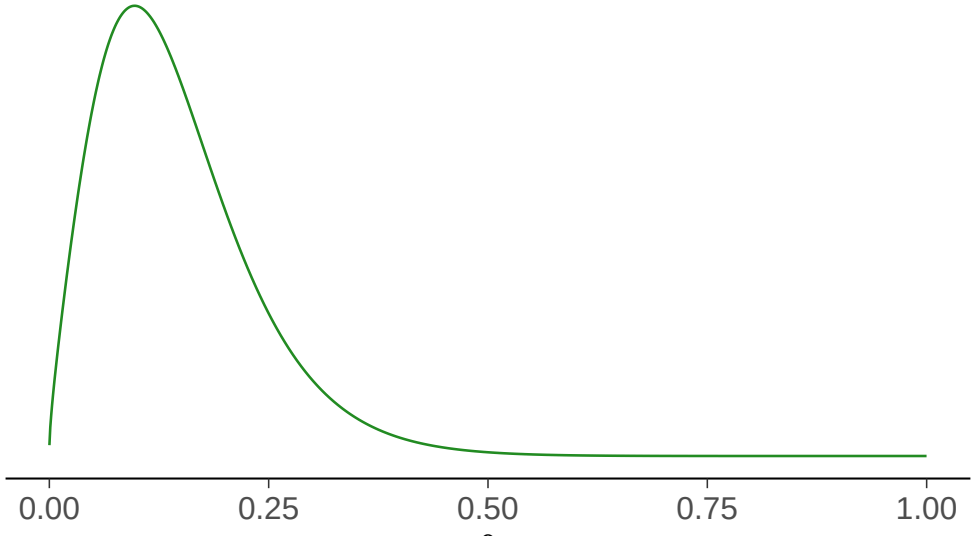


Hierarchical binomial model: rats

Hierarchical model



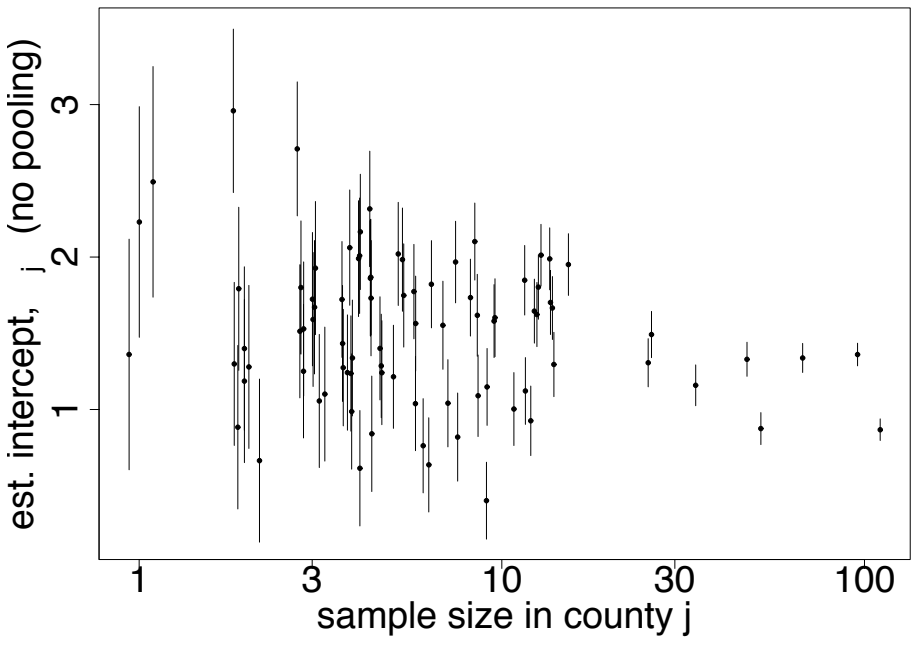
Population distribution (prior) for θ_j



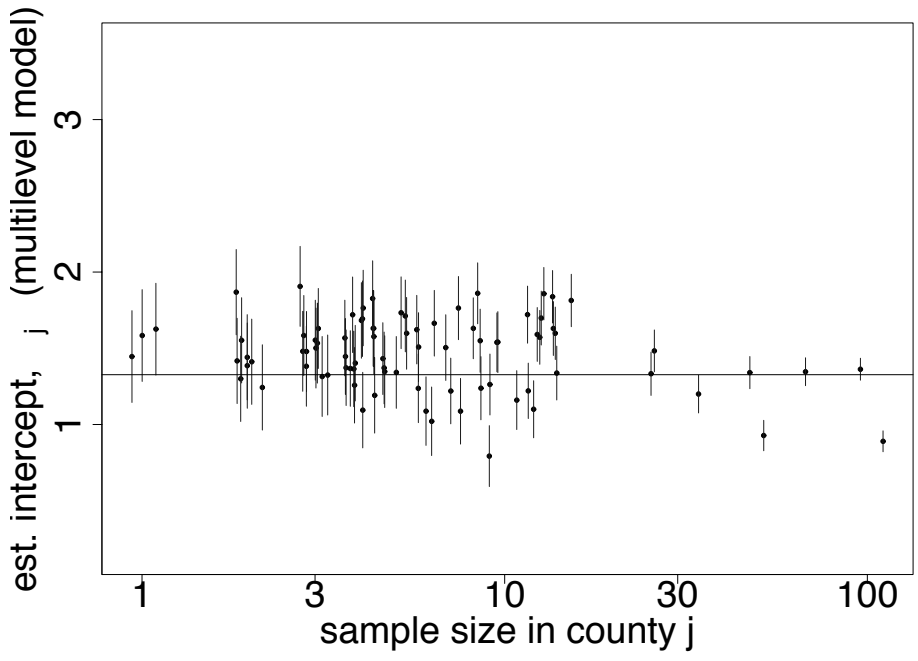
Hierarchical model and group size: Radon

919 home radon levels in 85 counties in Minnesota:

Separate



Hierarchical



Diet effect on chicken weights (at age 12 days)

- A typical treatment effect analysis
- Models
  - a separate model, in which each diet is modeled individually
  - a pooled model, in which all measurements are combined and there is no distinction between diets
  - a hierarchical model

Stan vs brms

```
model {
  // Priors
  for (diet in 1:N_diets) {
    mean_diet[diet] ~ normal(0, 10);
    sd_diet[diet] ~ exponential(1);
  }

  // Observation model / likelihood
  for (obs in 1:N_observations) {
    weight[obs] ~ normal(mean_diet[diet_idx[obs]],
                         sd_diet[diet_idx[obs]]);
  }
}
```

Stan vs brms

```
model {
  // Priors
  for (diet in 1:N_diets) {
    mean_diet[diet] ~ normal(0, 10);
    sd_diet[diet] ~ exponential(1);
  }

  // Observation model / likelihood
  for (obs in 1:N_observations) {
    weight[obs] ~ normal(mean_diet[diet_idx[obs]],
                         sd_diet[diet_idx[obs]]);
  }
  // Best practice would be to write the likelihood
  // without the for loop as
  // weight ~ normal(mean_diet[diet_idx],
  //                 sd_diet[diet_idx]);
}
```

Stan vs brms

```
model {
  // Priors
  for (diet in 1:N_diets) {
    mean_diet[diet] ~ normal(0, 10);
    sd_diet[diet] ~ exponential(1);
  }

  // Observation model / likelihood
  for (obs in 1:N_observations) {
    weight[obs] ~ normal(mean_diet[diet_idx[obs]],
                         sd_diet[diet_idx[obs]]);
  }
}

brm(weight ~ 1 + (1 | Diet),
```

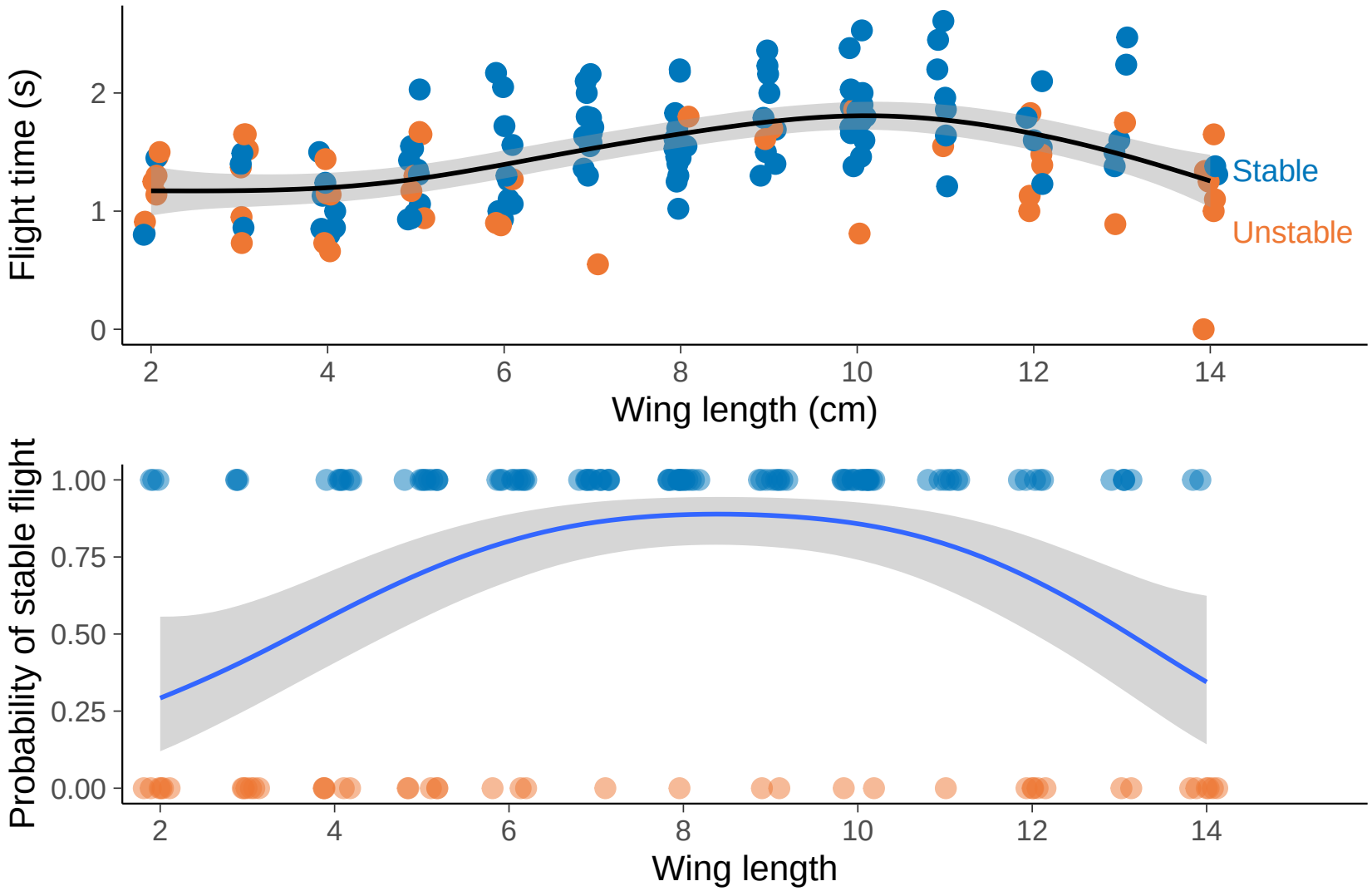
Stan vs brms

```
model {
  // Priors
  for (diet in 1:N_diets) {
    mean_diet[diet] ~ normal(0, 10);
    sd_diet[diet] ~ exponential(1);
  }

  // Observation model / likelihood
  for (obs in 1:N_observations) {
    weight[obs] ~ normal(mean_diet[diet_idx[obs]],
                         sd_diet[diet_idx[obs]]);
  }
}
```

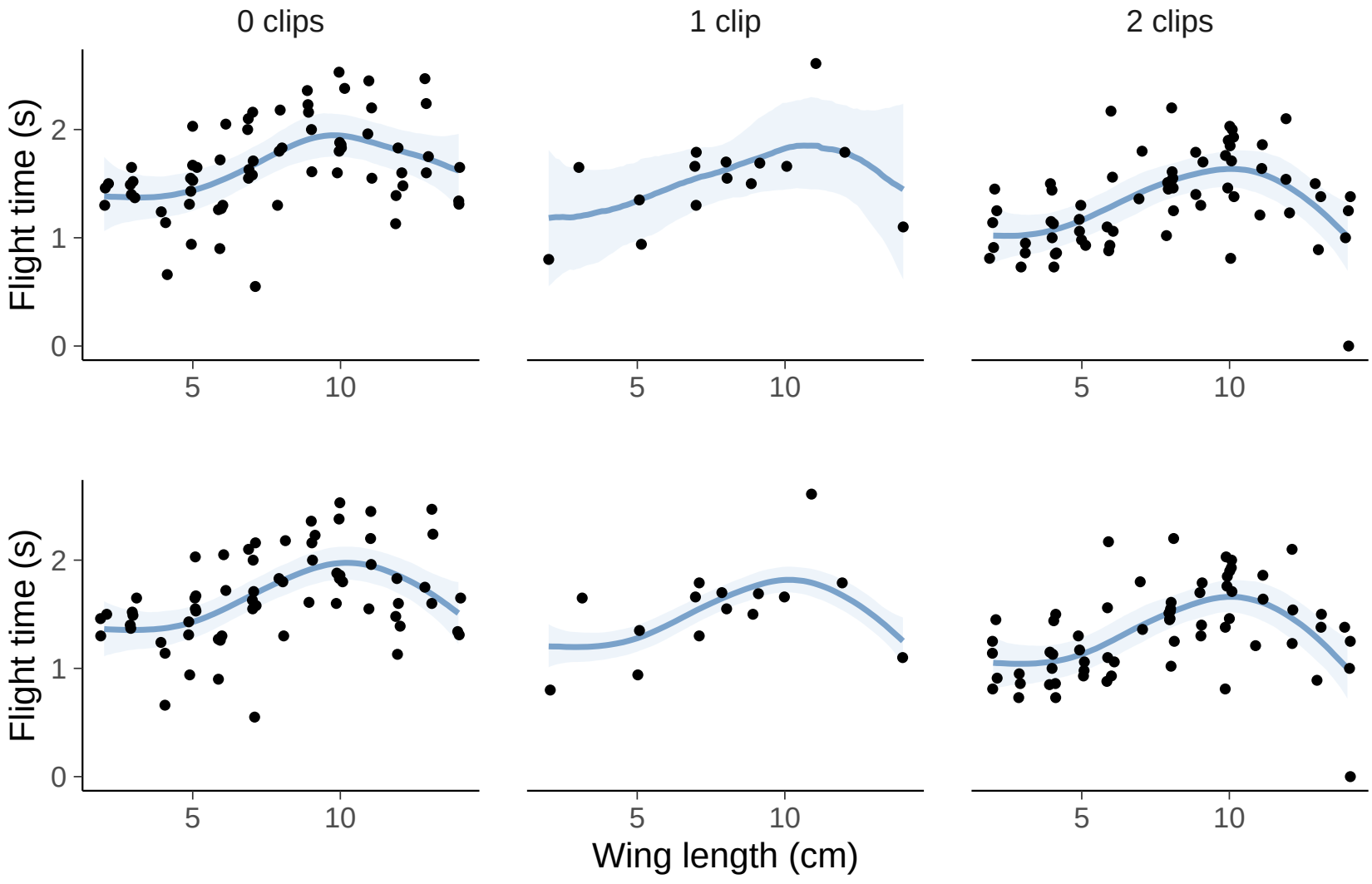
```
brm(weight ~ 1 + (1 | Diet), data=Chick12,
    prior=c(prior(normal(0,1), class="Intercept"), # p(mu_0)
            prior(normal(0,1), class="sigma"),     # p(sigma)
            prior(normal(0,1), class="sd"))        # p(tau)
```

Paper helicopters



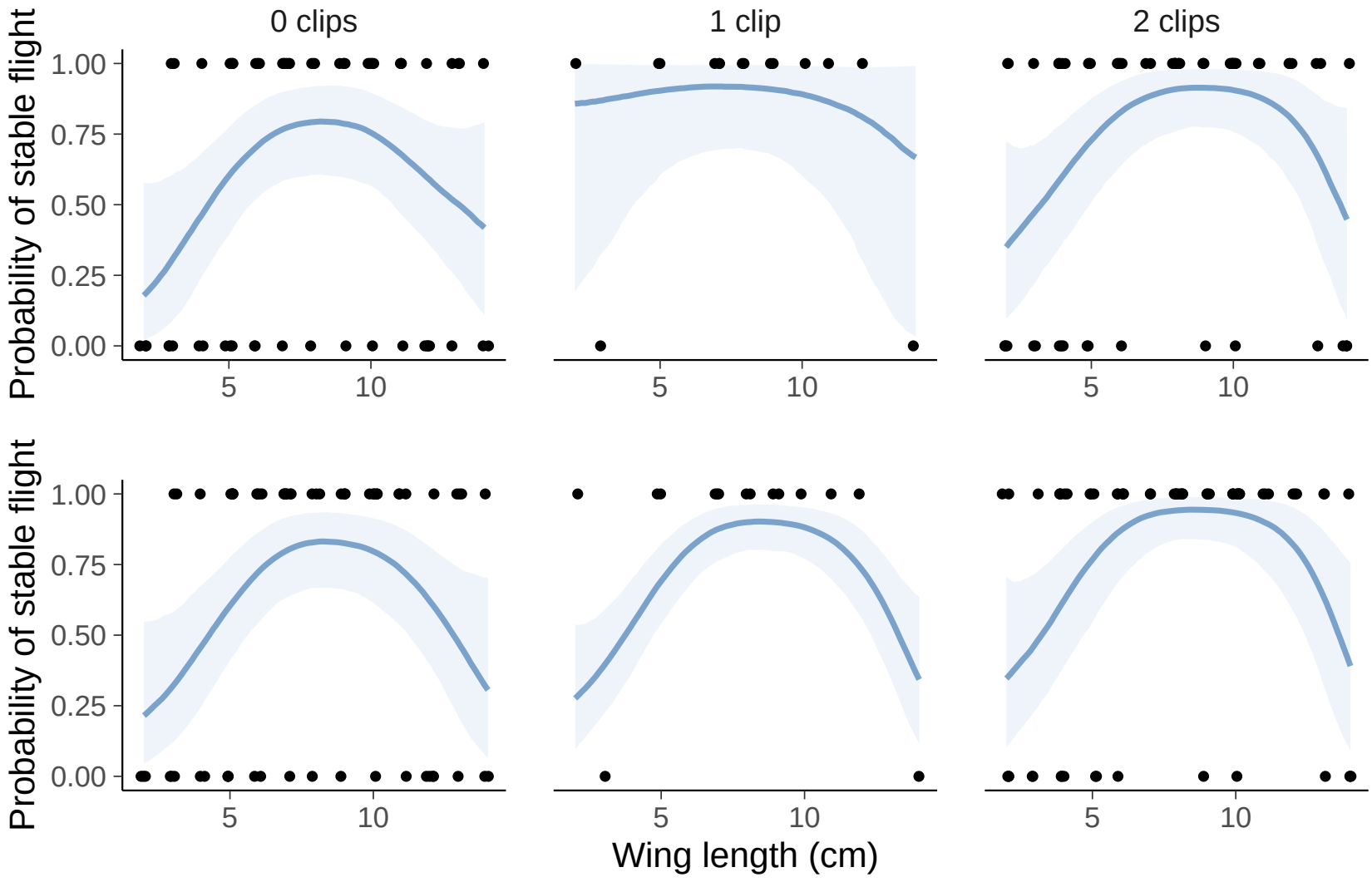
Paper helicopters: flight time

Separate model vs. hierarchical model



Paper helicopters: stability

Separate model vs. hierarchical model



Paper helicopters: brms

Flight time

```
flight_time ~ s(wing_length) + s(wing_length, by = nclips)
```

Paper helicopters: brms

Flight time

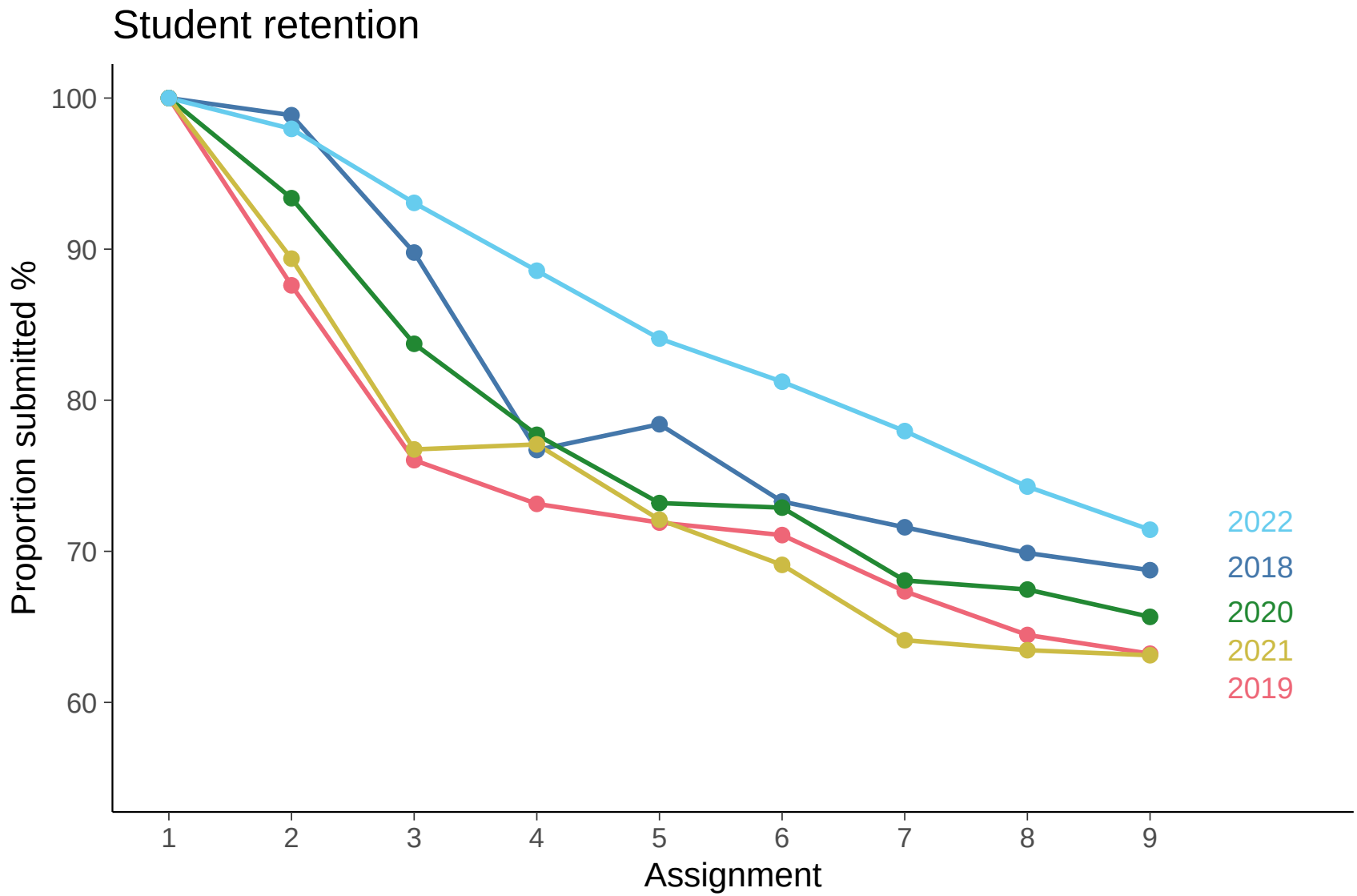
```
flight_time ~ s(wing_length) + s(wing_length, by = nclips)
```

Stability

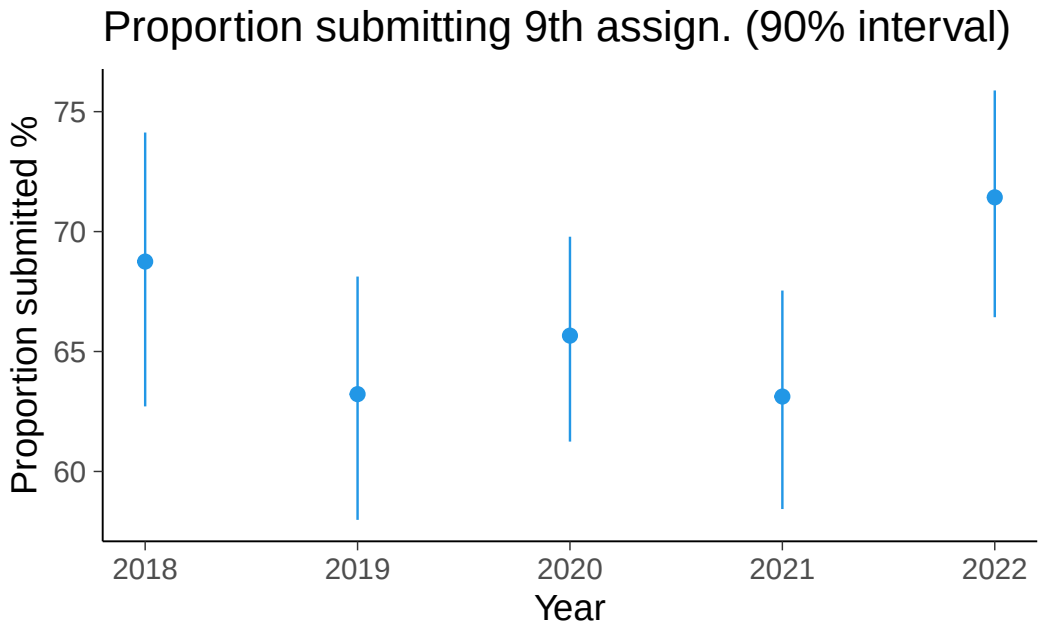
```
stable_flight ~ s(wing_length) + s(wing_length, by = nclips)
family = bernoulli()
```

Student retention

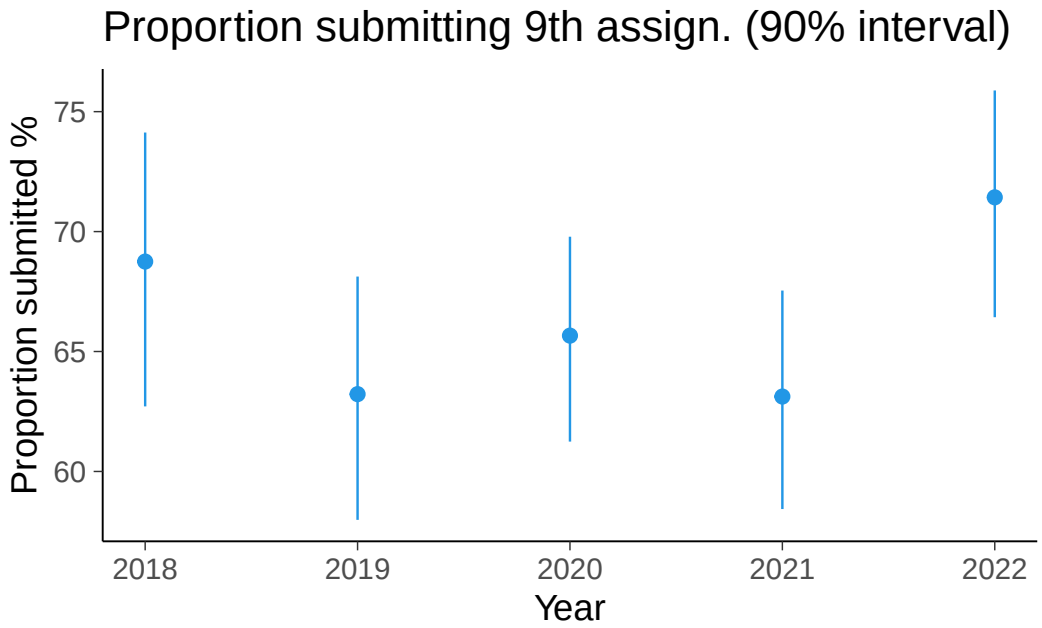
Was year 2022 better than earlier year?

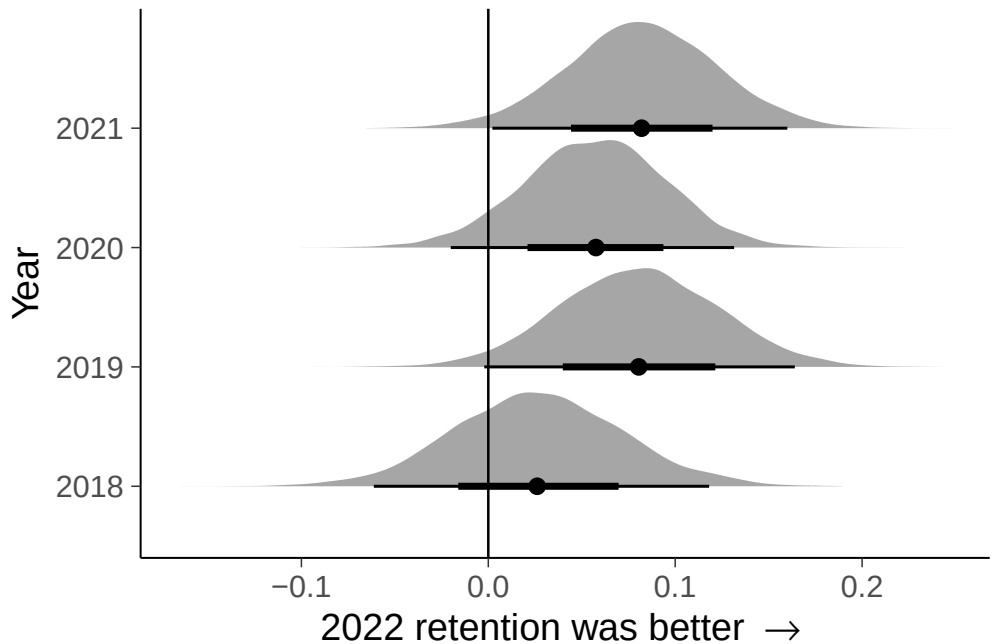


Student retention separate model

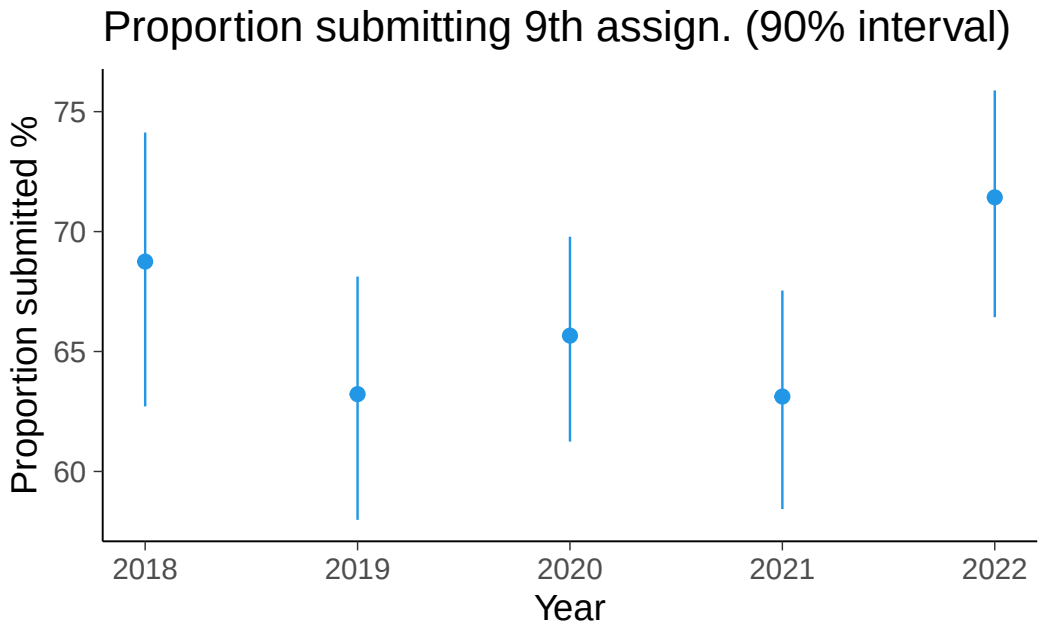


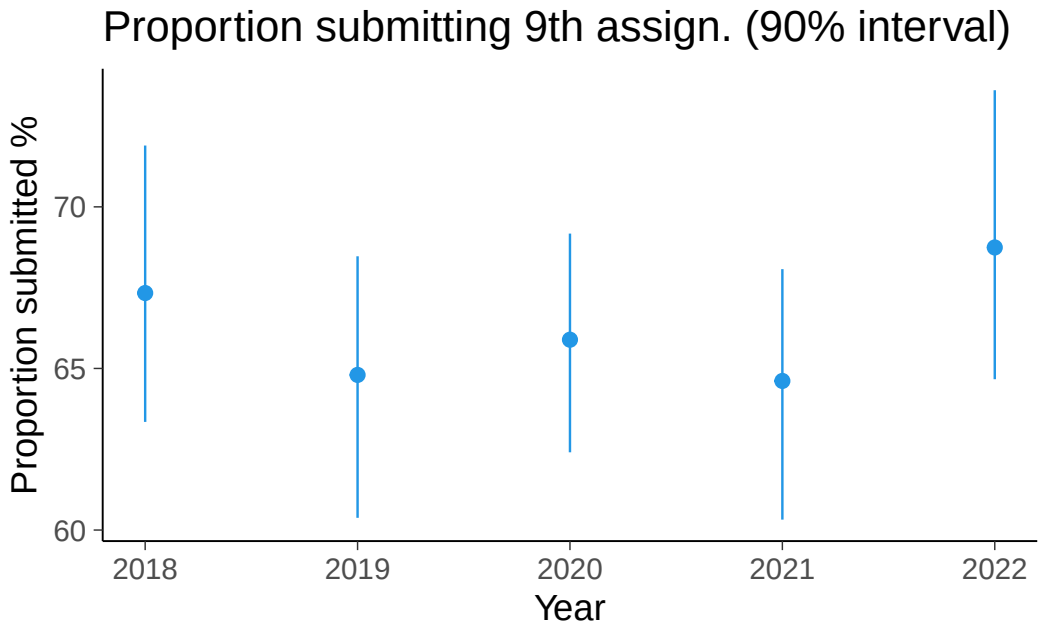
Student retention separate model



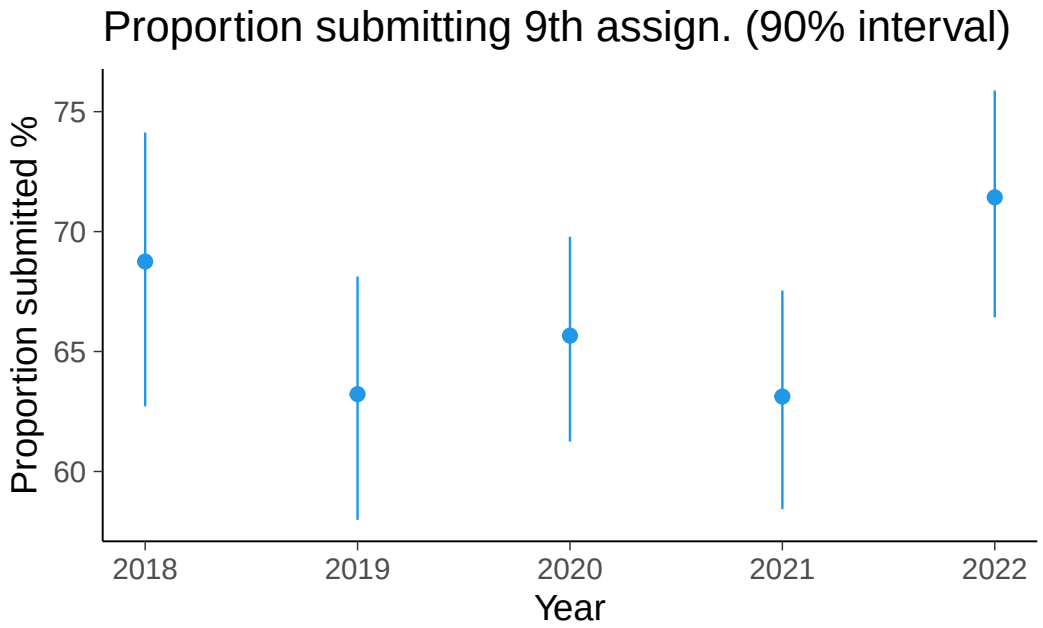


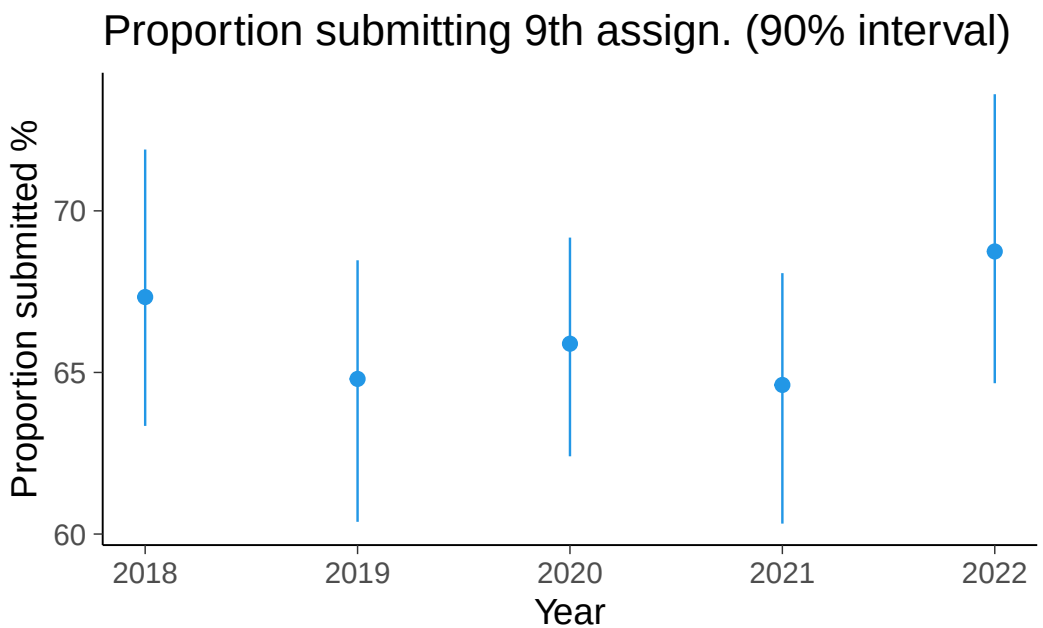
Student retention separate vs hierarchical model





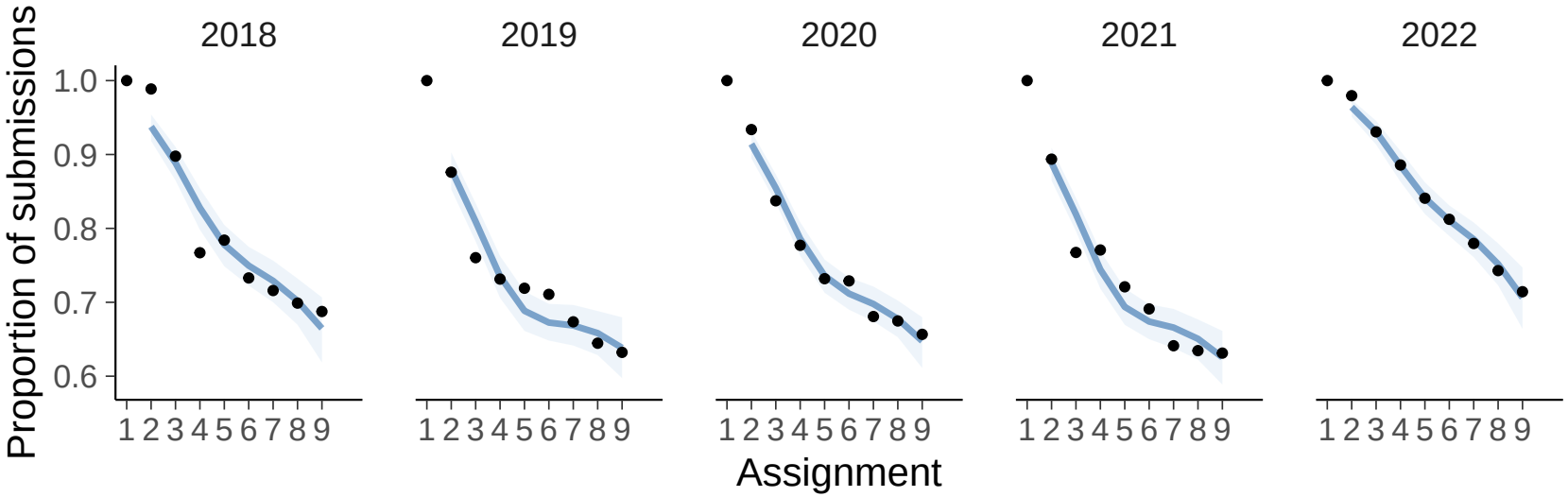
Student retention separate vs hierarchical model





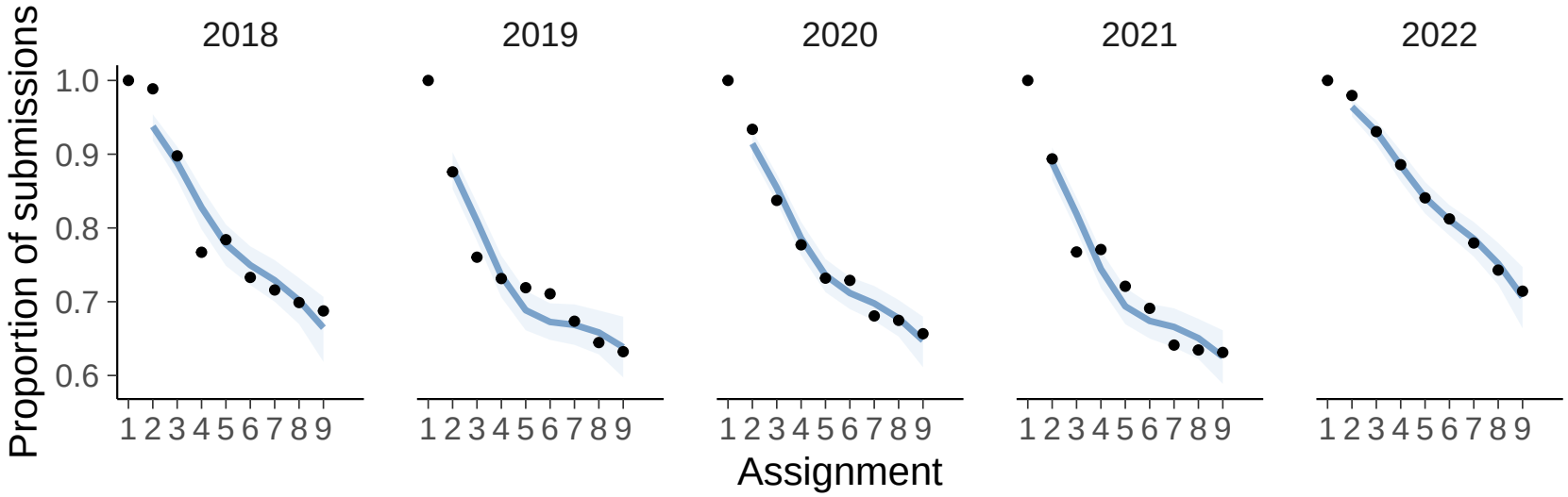

```
nstudents | trials(nstudents1) ~ 1 + (1 | year), family=binomial()
```

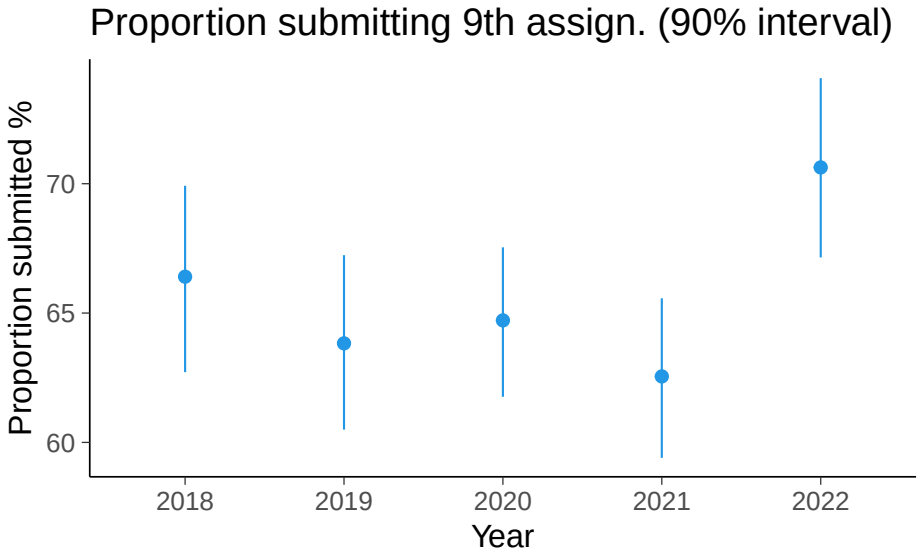
Student retention latent spline model



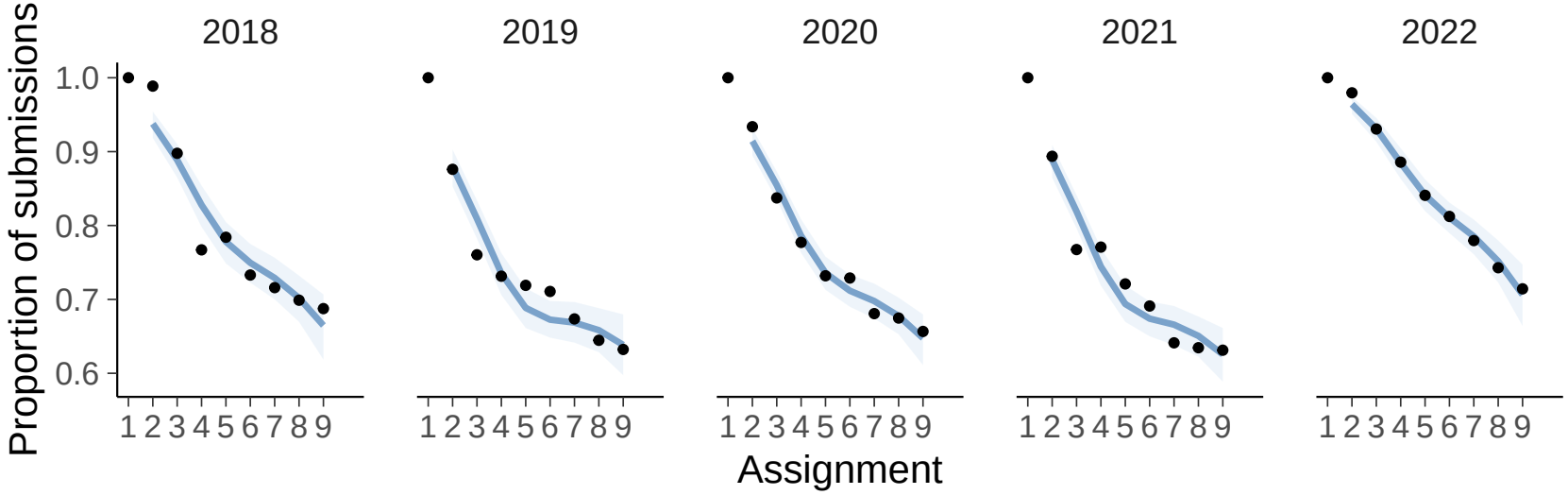

```
nstudents | trials(nstudents1) ~ s(assignment, k=4) + (assignment | year),
family=binomial()
```

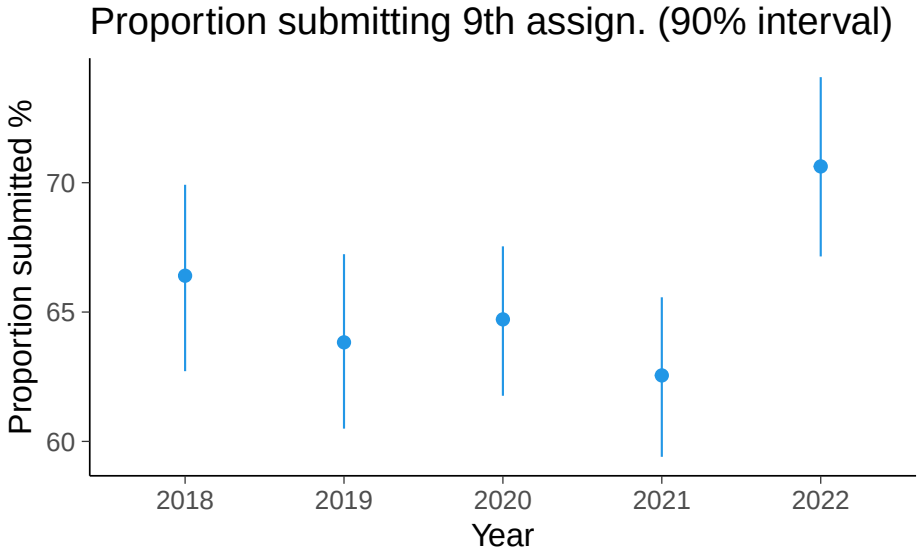
Student retention latent spline model

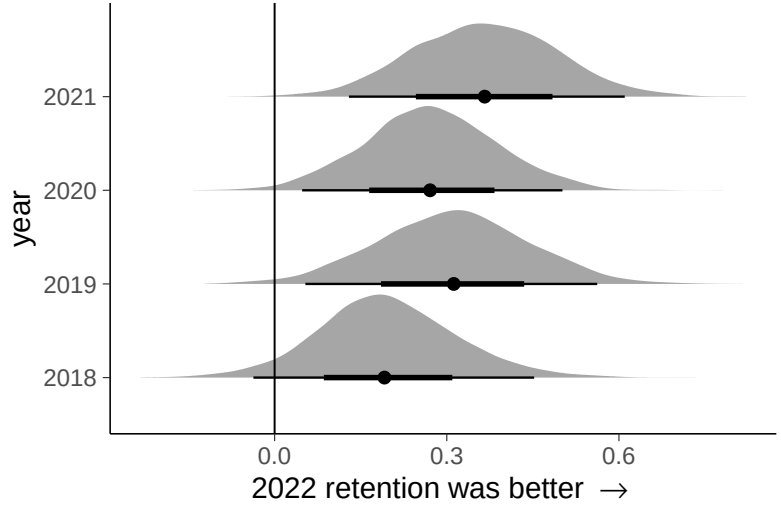




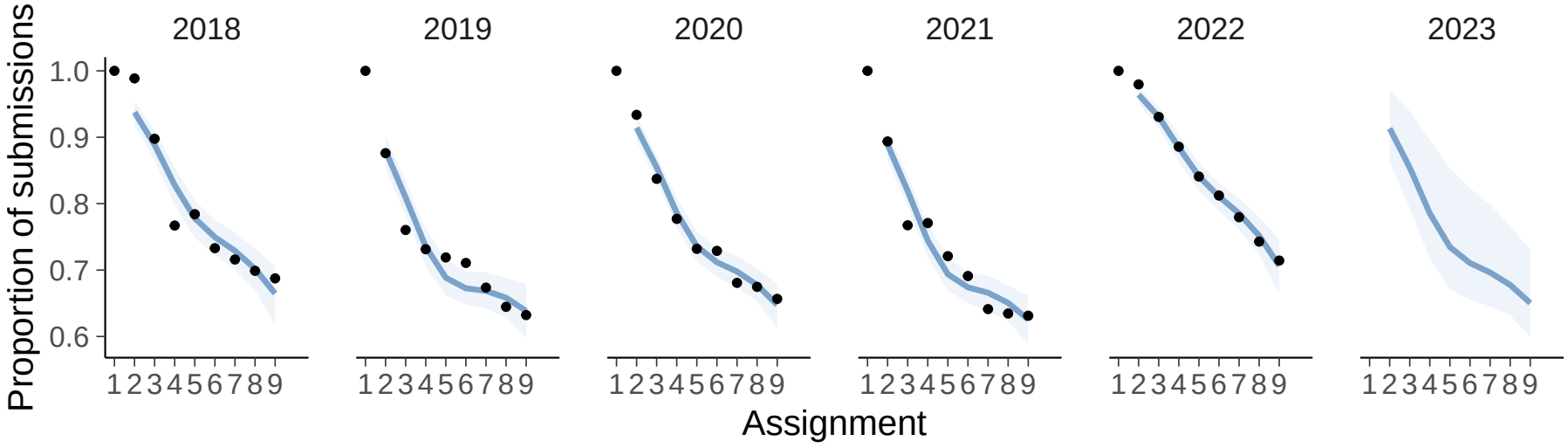
Student retention latent spline model



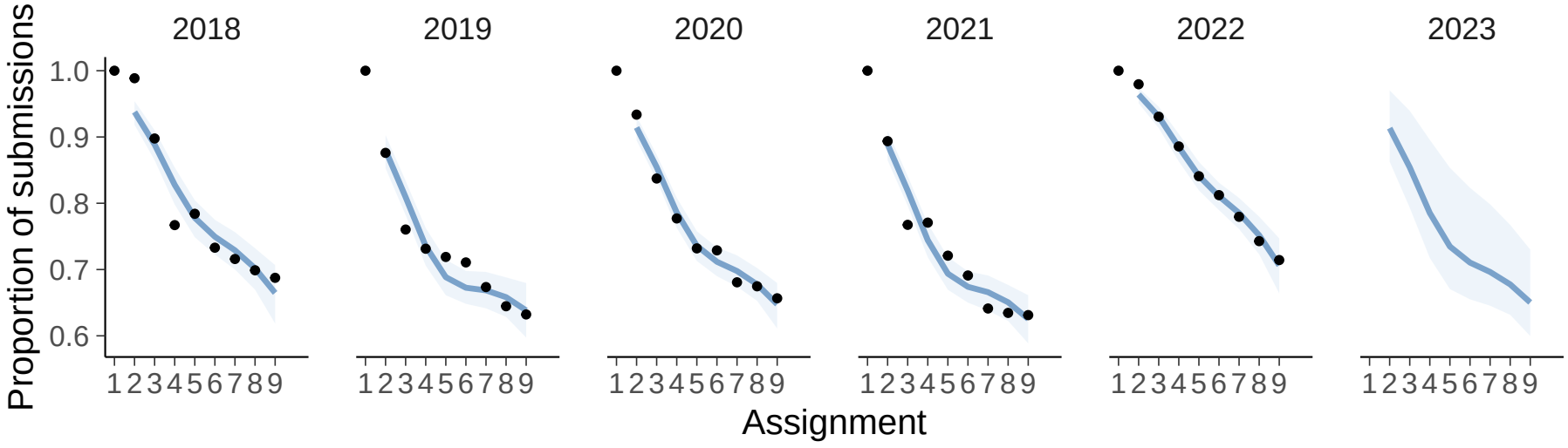


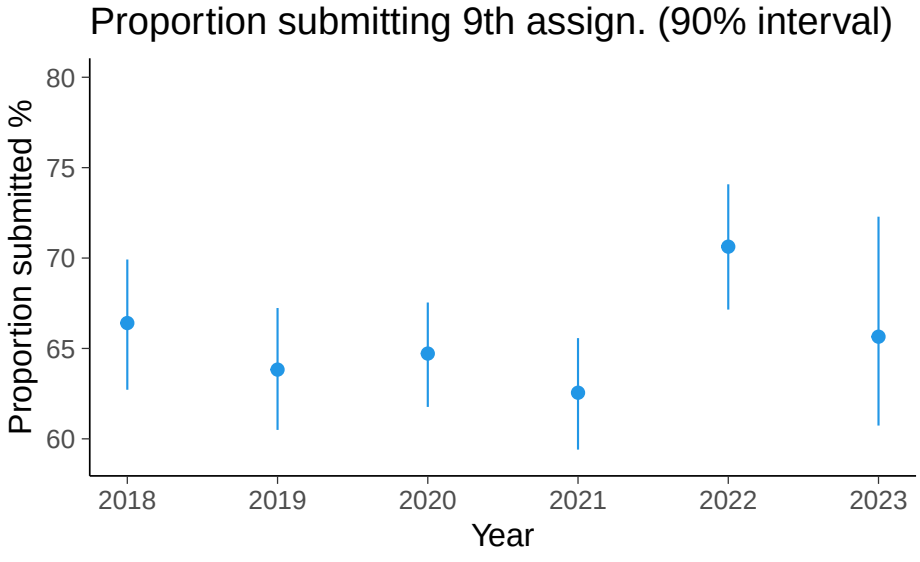


Student retention latent spline model, year 2023?

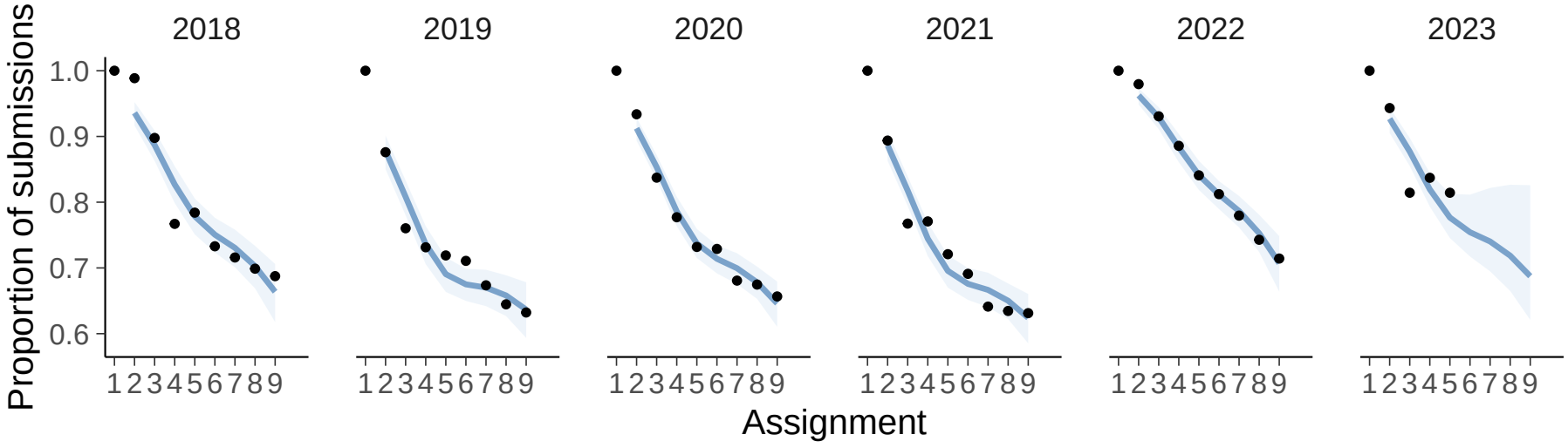


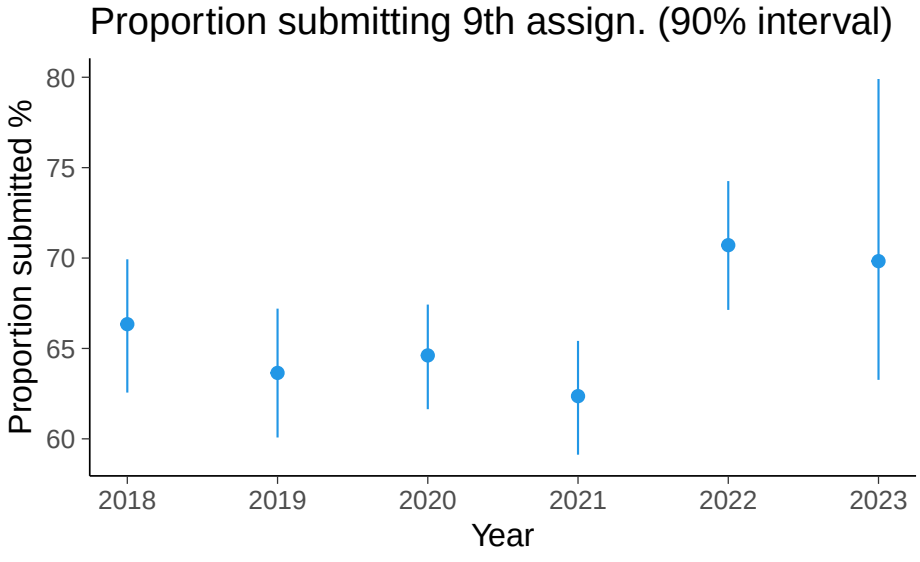
Student retention latent spline model, year 2023?



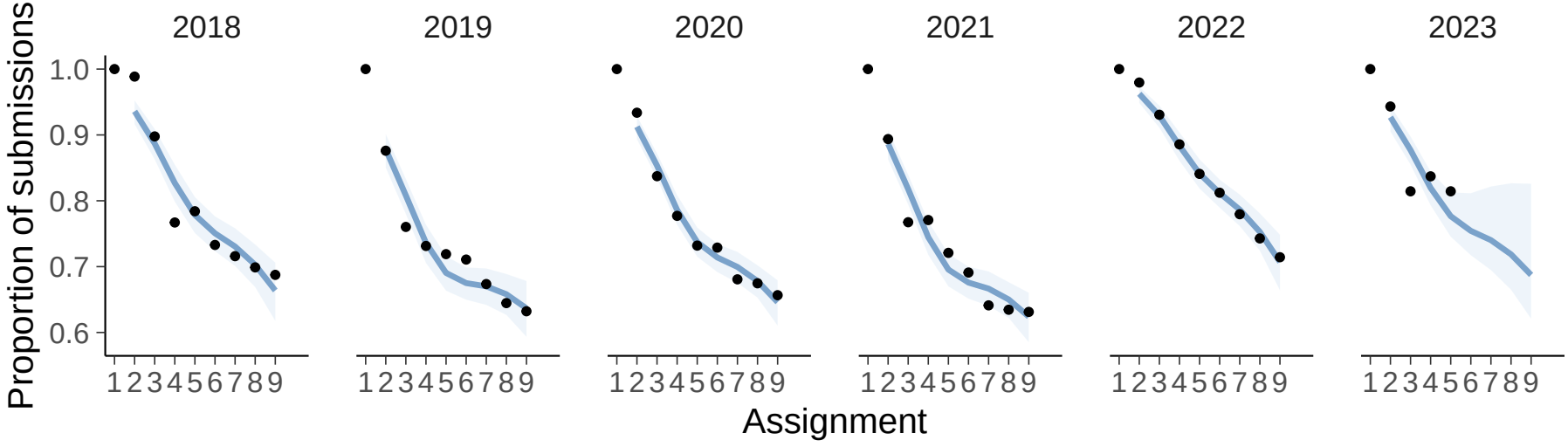


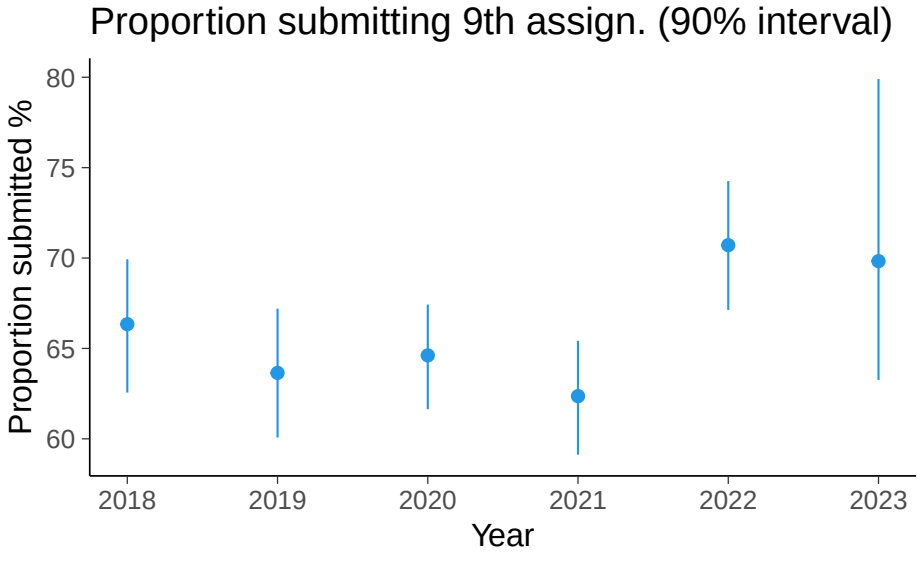
Student retention latent spline model, year 2023?

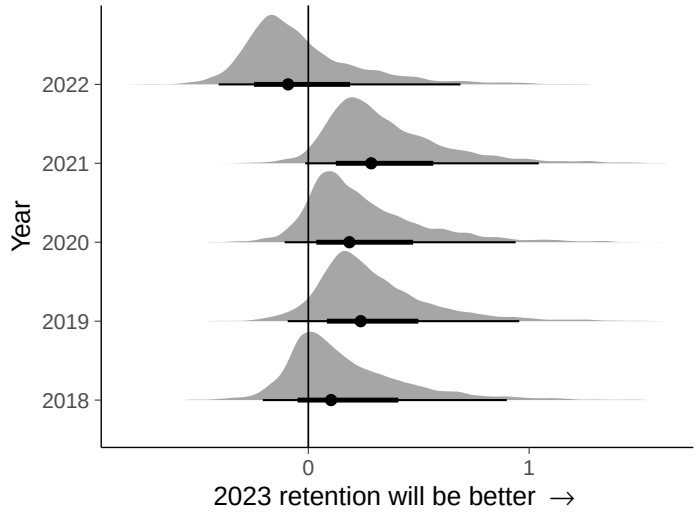




Student retention latent spline model, year 2023?







Centered vs non-centered parameterization

HMC divergences are more likely when using hierarchical models

Centered vs non-centered parameterization

Hierarchical model code from the course demos

```
data {
  int<lower=0> N; // number of observations
  int<lower=0> K; // number of groups
  array[N] int<lower=1, upper=K> x; // discrete group indicators
  vector[N] y; // real valued observations
}
```

Centered vs non-centered parameterization

Hierarchical model code from the course demos

```
data {
  int<lower=0> N; // number of observations
  int<lower=0> K; // number of groups
  array[N] int<lower=1, upper=K> x; // discrete group indicators
  vector[N] y; // real valued observations
}

parameters {
  real mu0; // prior mean
  real<lower=0> sigma0; // prior std constrained to be positive
  vector[K] mu; // group means
  real<lower=0> sigma; // common std constrained to be positive
}
```

Centered parameterization

Hierarchical model code from the course demos

```
data {
  int<lower=0> N; // number of observations
  int<lower=0> K; // number of groups
  array[N] int<lower=1, upper=K> x; // discrete group indicators
  vector[N] y; // real valued observations
}

parameters {
  real mu0; // prior mean
  real<lower=0> sigma0; // prior std constrained to be positive
  vector[K] mu; // group means
  real<lower=0> sigma; // common std constrained to be positive
}

model {
  mu0 ~ normal(10, 10);      // weakly informative prior
  sigma0 ~ normal(0, 10);    // weakly informative prior
  mu ~ normal(mu0, sigma0); // population prior with unknown param
  sigma ~ lognormal(0, .5); // weakly informative prior
  y ~ normal(mu[x], sigma); // observation model / likelihood
}
```

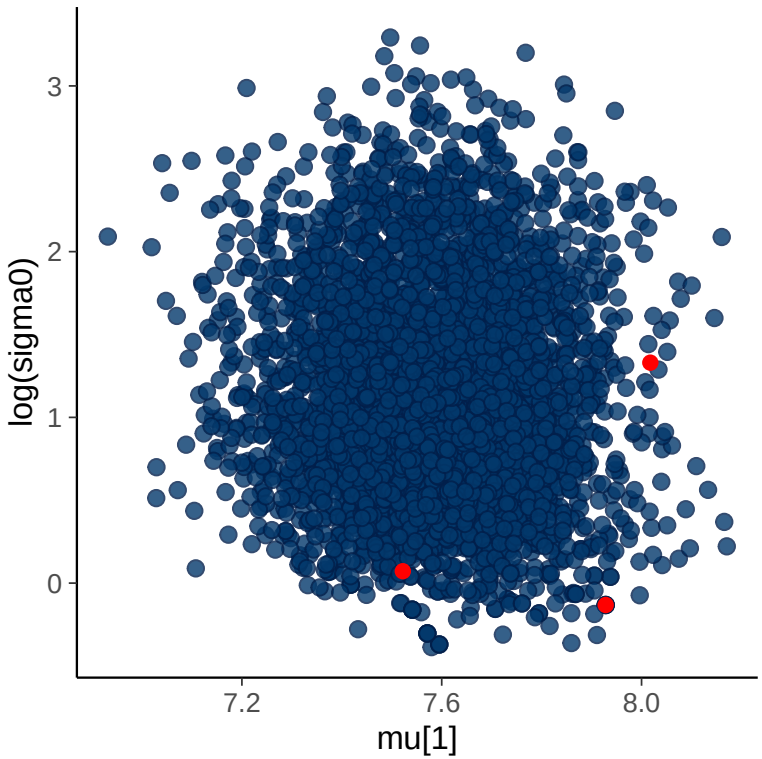
Centered parameterization

First data with many observations per group: 3 summer months with each having 71 observations.

Centered parameterization

First data with many observations per group: 3 summer months with each having 71 observations.

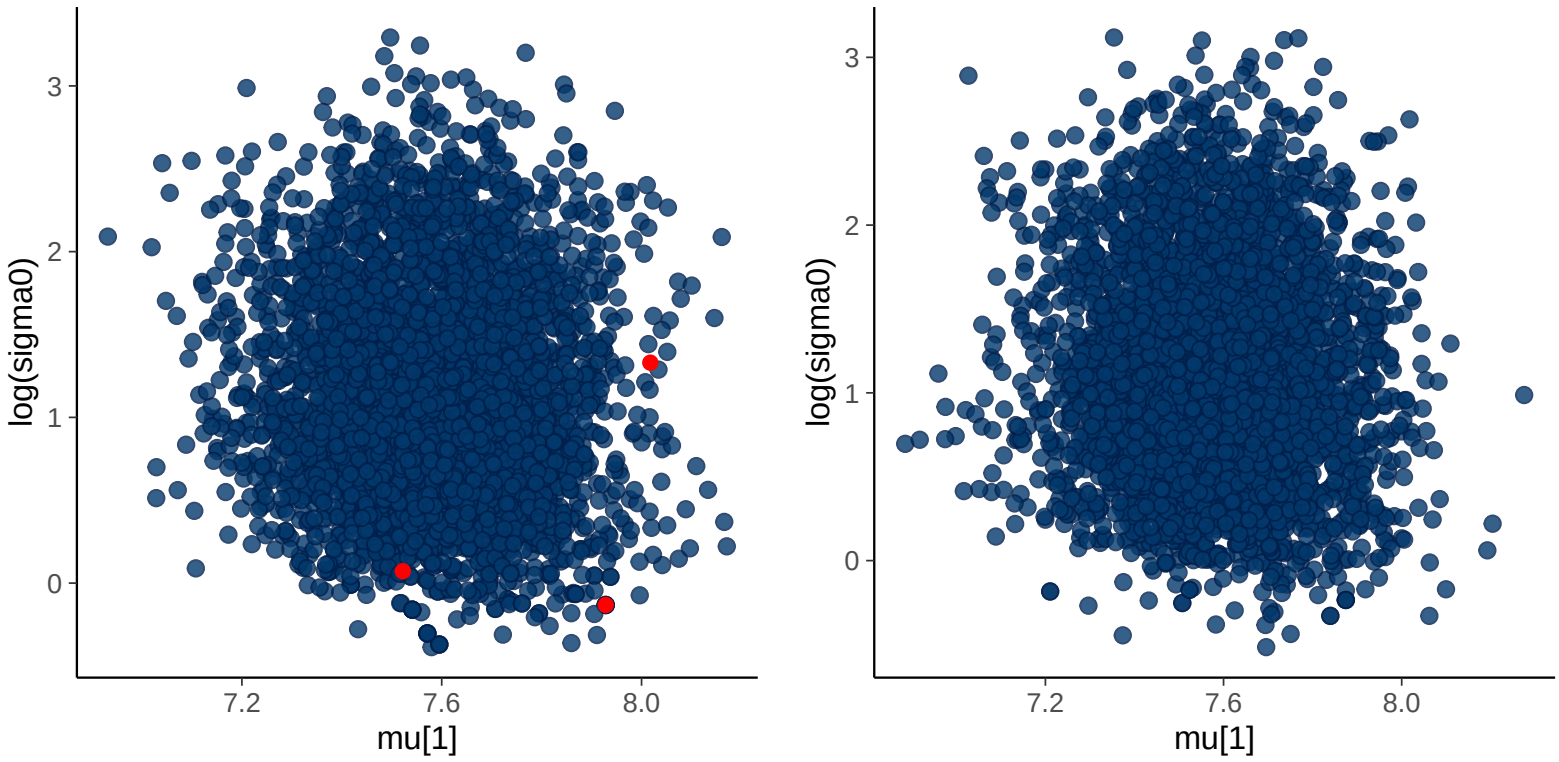
A few divergences that are not clustered.



Centered parameterization

First data with many observations per group: 3 summer months with each having 71 observations.

And decreasing step size a little helps.



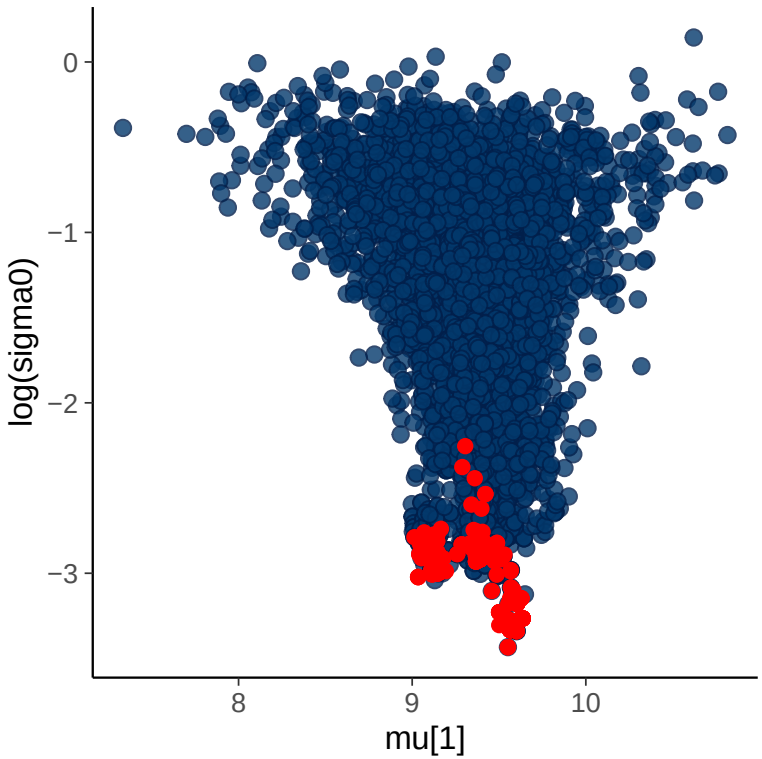
Centered parameterization

Second data with a few observations per group: 71 years with each having 3 observations.

Centered parameterization

Second data with a few observations per group: 71 years with each having 3 observations.

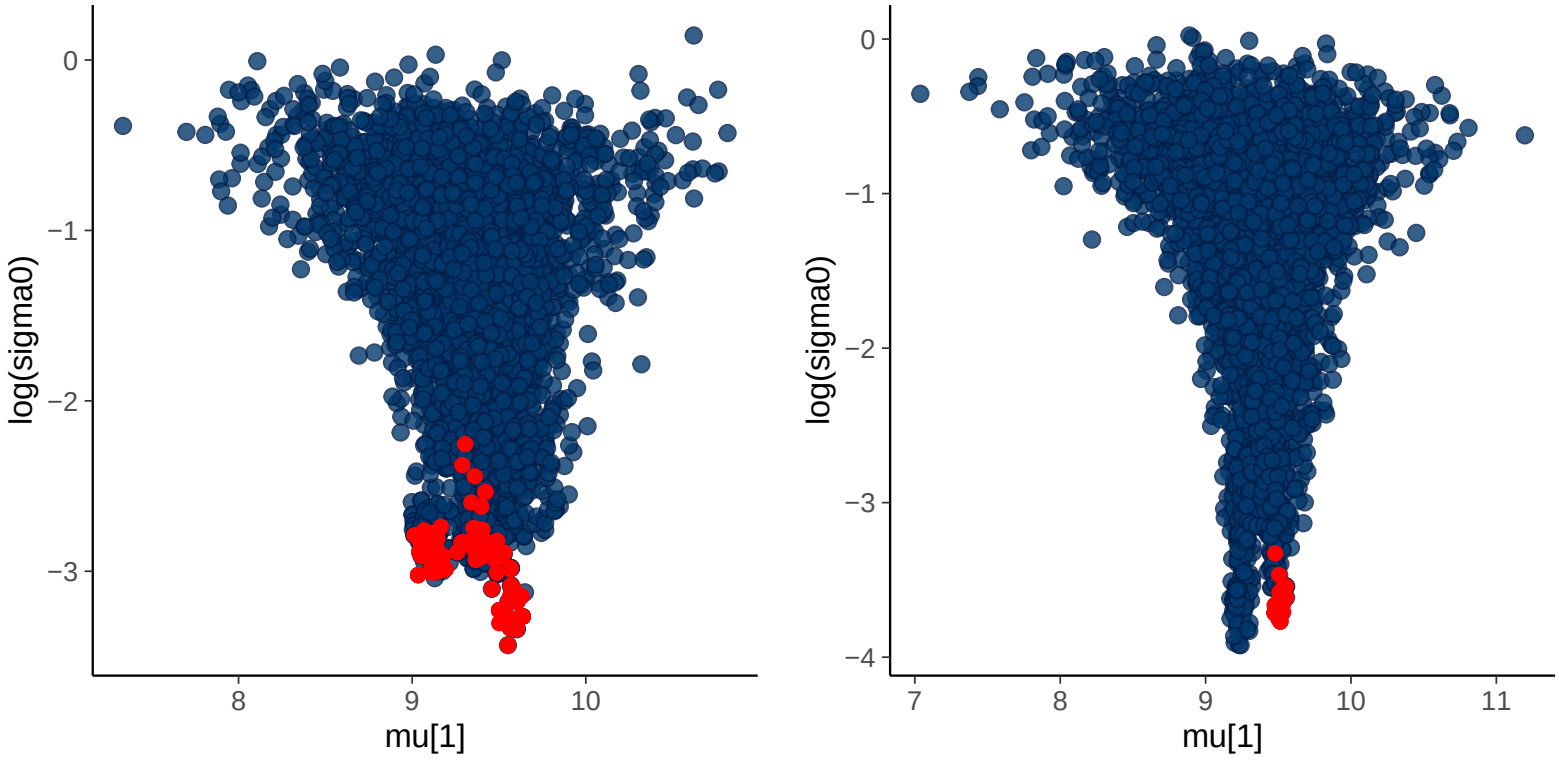
Many divergences that are clustered.



Centered parameterization

Second data with a few observations per group: 71 years with each having 3 observations.

And decreasing step size doesn't remove the problem.



Non-centered parameterization

Transformation

```
parameters {
  real mu0; // prior mean
  real<lower=0> sigma0; // prior std constrained to be positive
  vector[K] z; // latent variable
  real<lower=0> sigma; // common std constrained to be positive
}

transformed parameters {
  vector[K] mu = mu0 + sigma0 * z; // group means
}

model {
  mu0 ~ normal(10, 10);       // weakly informative prior
  sigma0 ~ normal(0, 10);     // weakly informative prior
  z ~ normal(0, 1);           // unit normal
  sigma ~ lognormal(0, .5); // weakly informative prior
  y ~ normal(mu[x], sigma); // observation model / likelihood
}
```

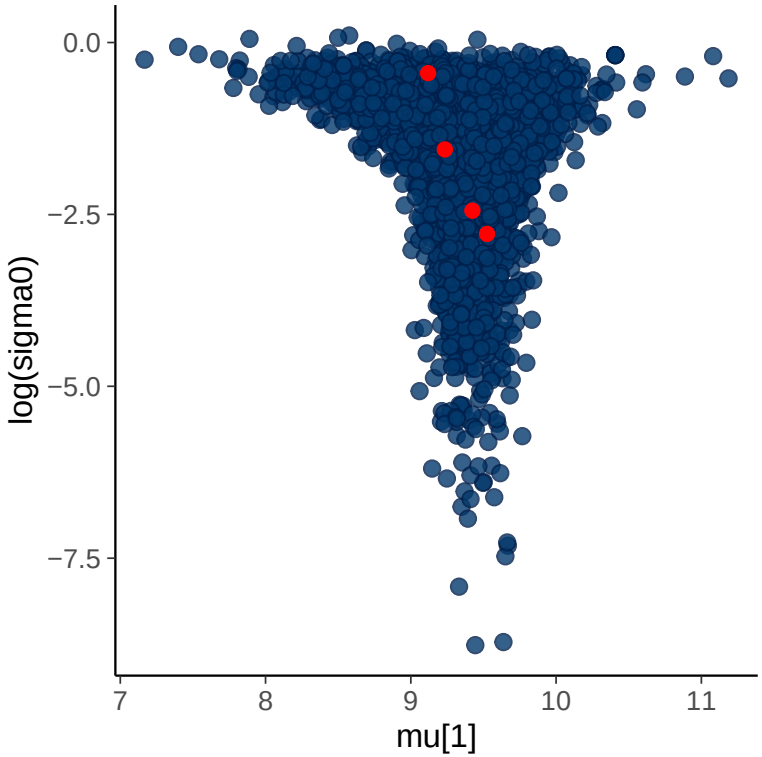
Non-centered parameterization

Second data with a few observations per group: 71 years with each having 3 observations.

Non-centered parameterization

Second data with a few observations per group: 71 years with each having 3 observations.

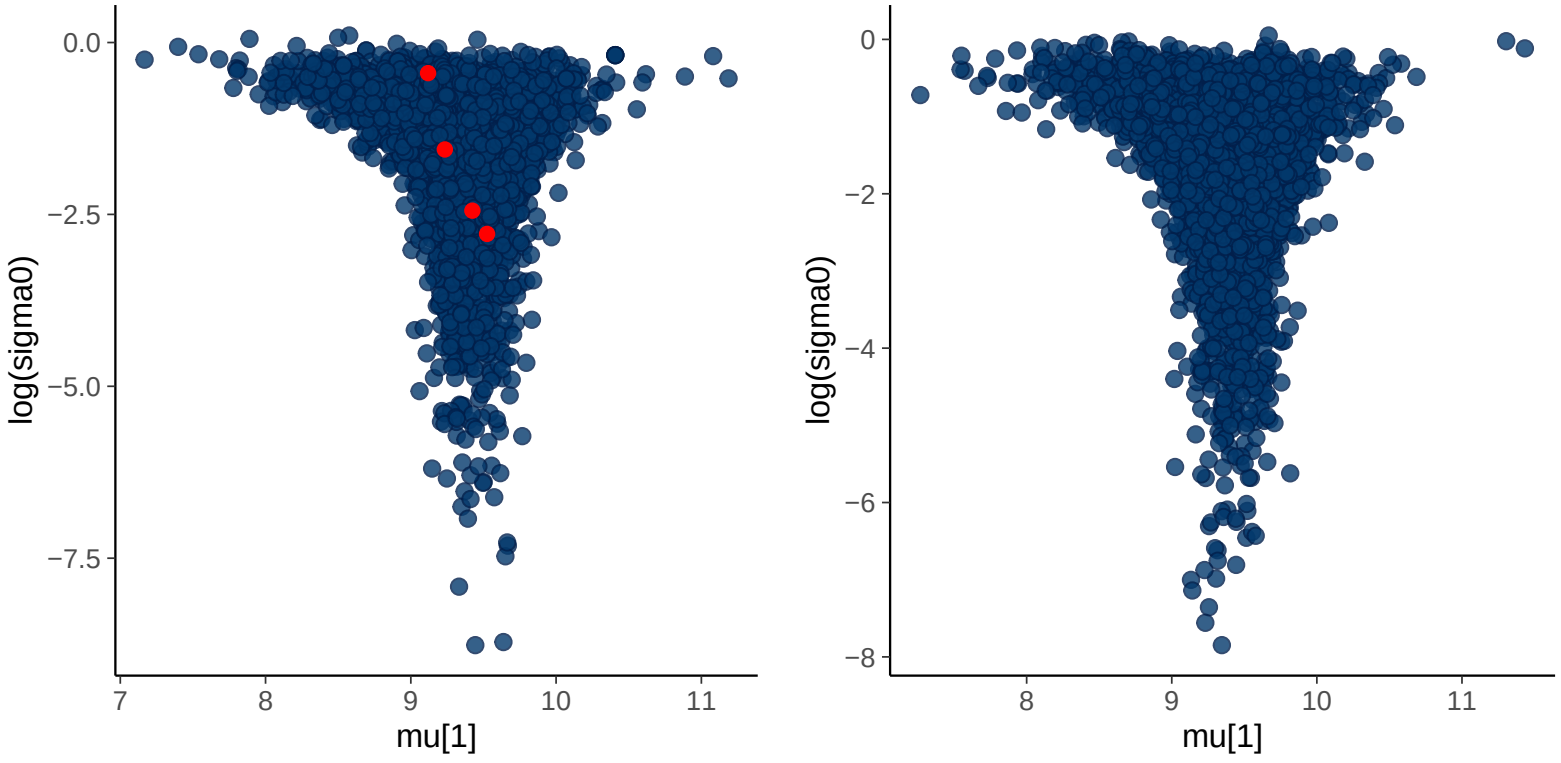
A few divergences that are not clustered.



Non-centered parameterization

Second data with a few observations per group: 71 years with each having 3 observations.

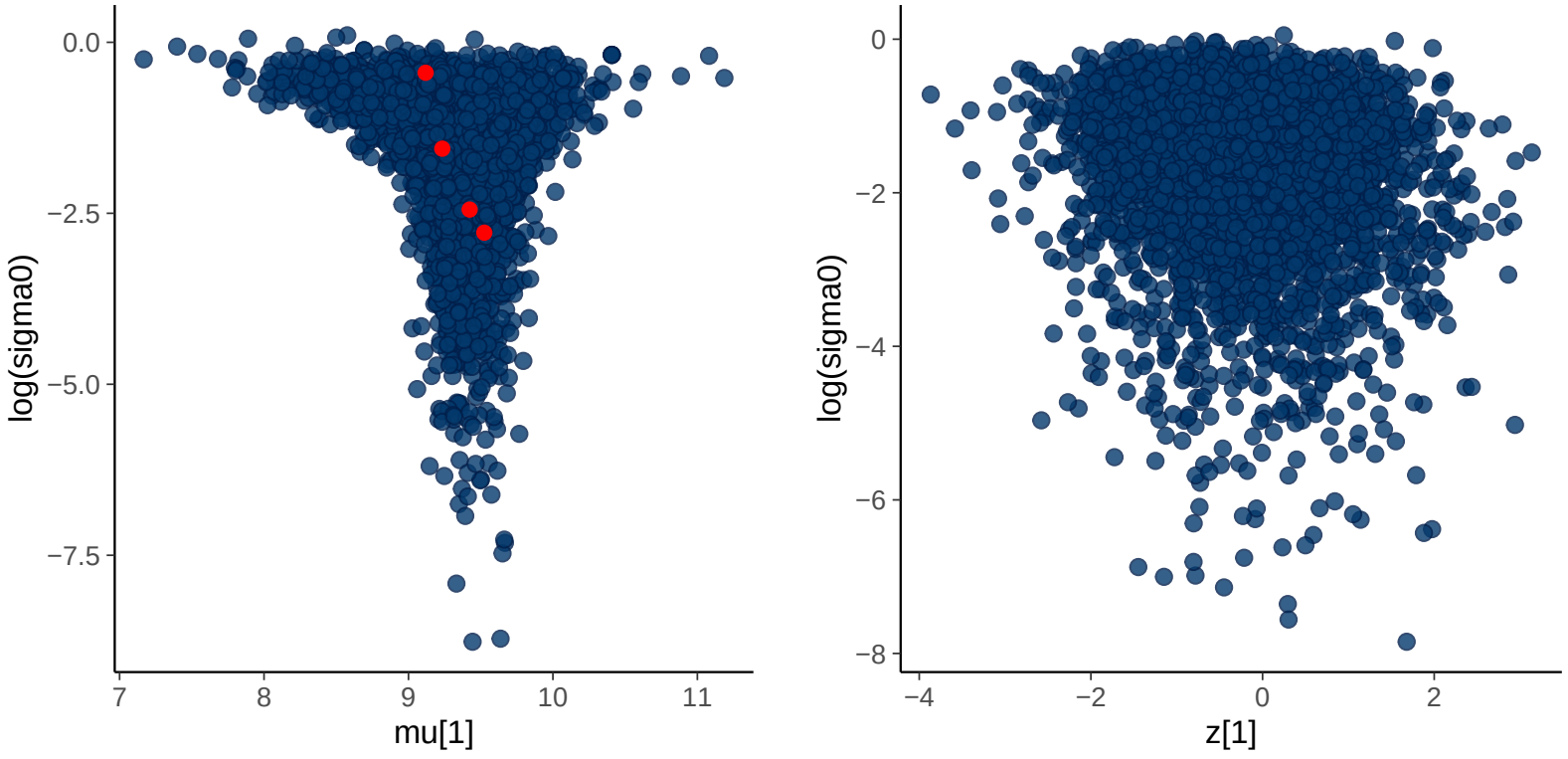
And decreasing step size a little helps.



Non-centered parameterization

Second data with a few observations per group: 71 years with each having 3 observations.

Because we're actually sampling `z` and not `mu`



Non-centered parameterization

No free lunch

- non-centered parameterization is good when likelihood is weak
- non-centered parameterization is bad when likelihood is strong

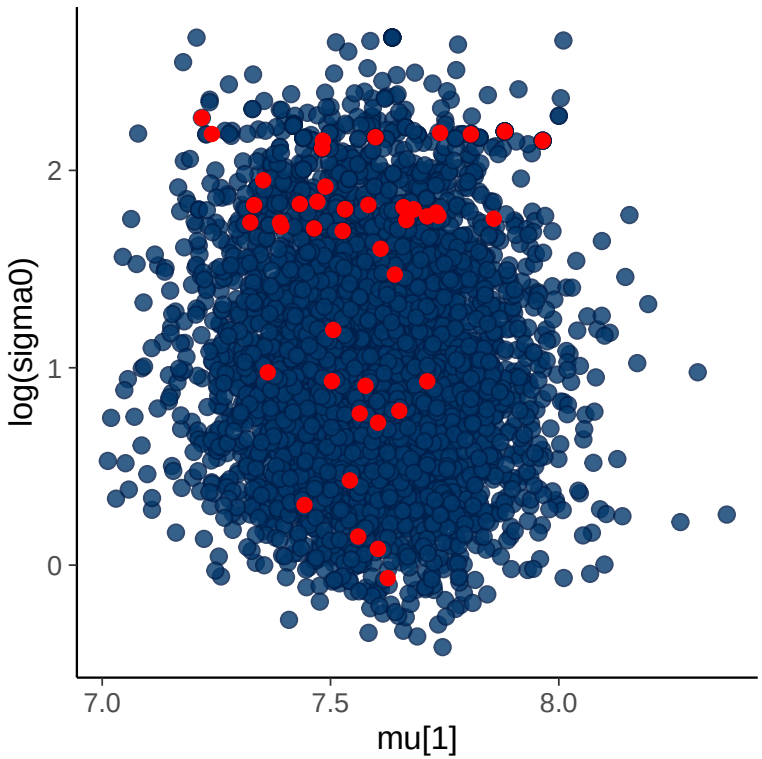
Non-centered parameterization

First data with many observations per group: 3 summer months with each having 71 observations.

Non-centered parameterization

First data with many observations per group: 3 summer months with each having 71 observations.

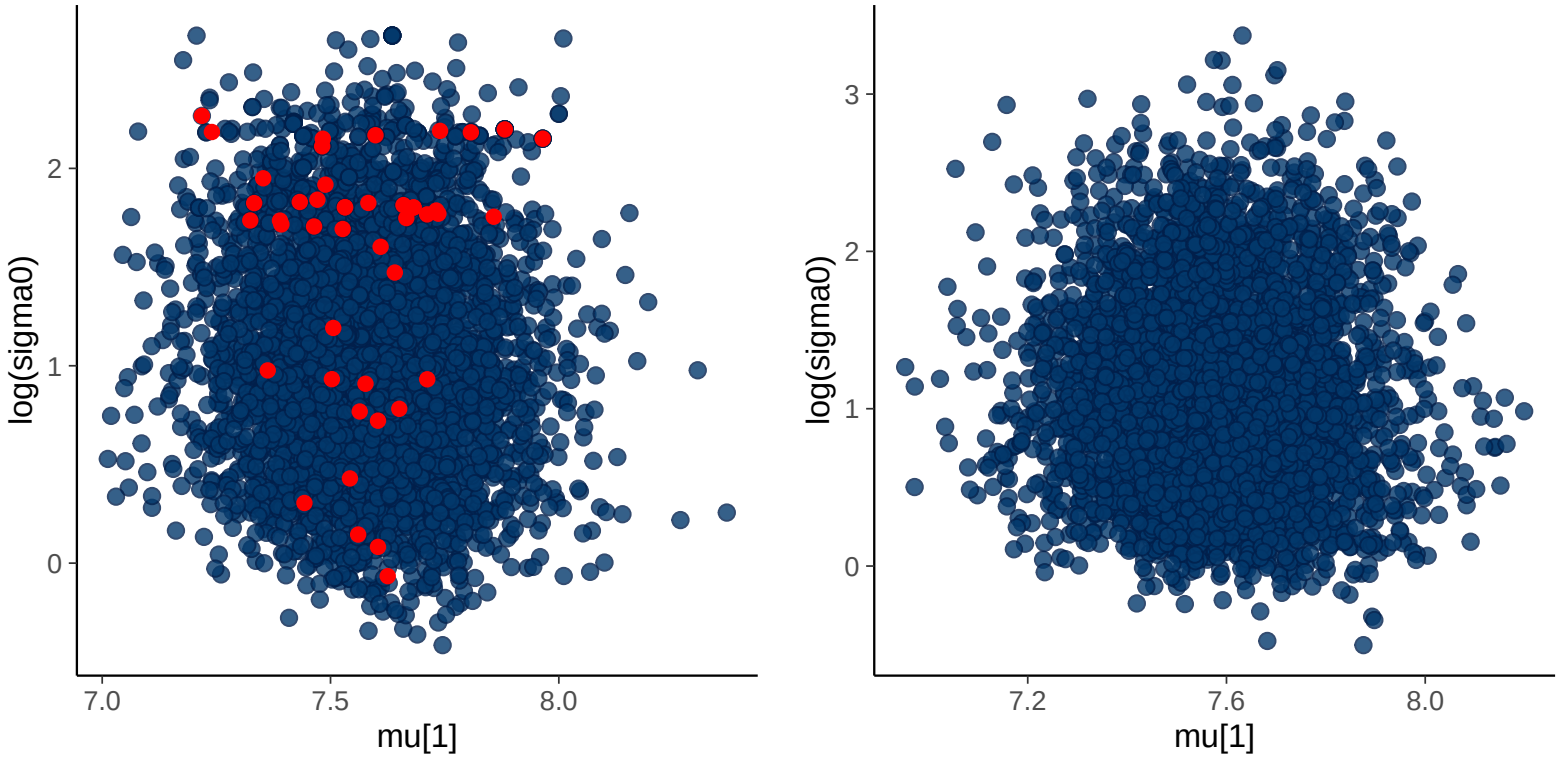
Many divergences that are not clustered.



Non-centered parameterization

First data with many observations per group: 3 summer months with each having 71 observations.

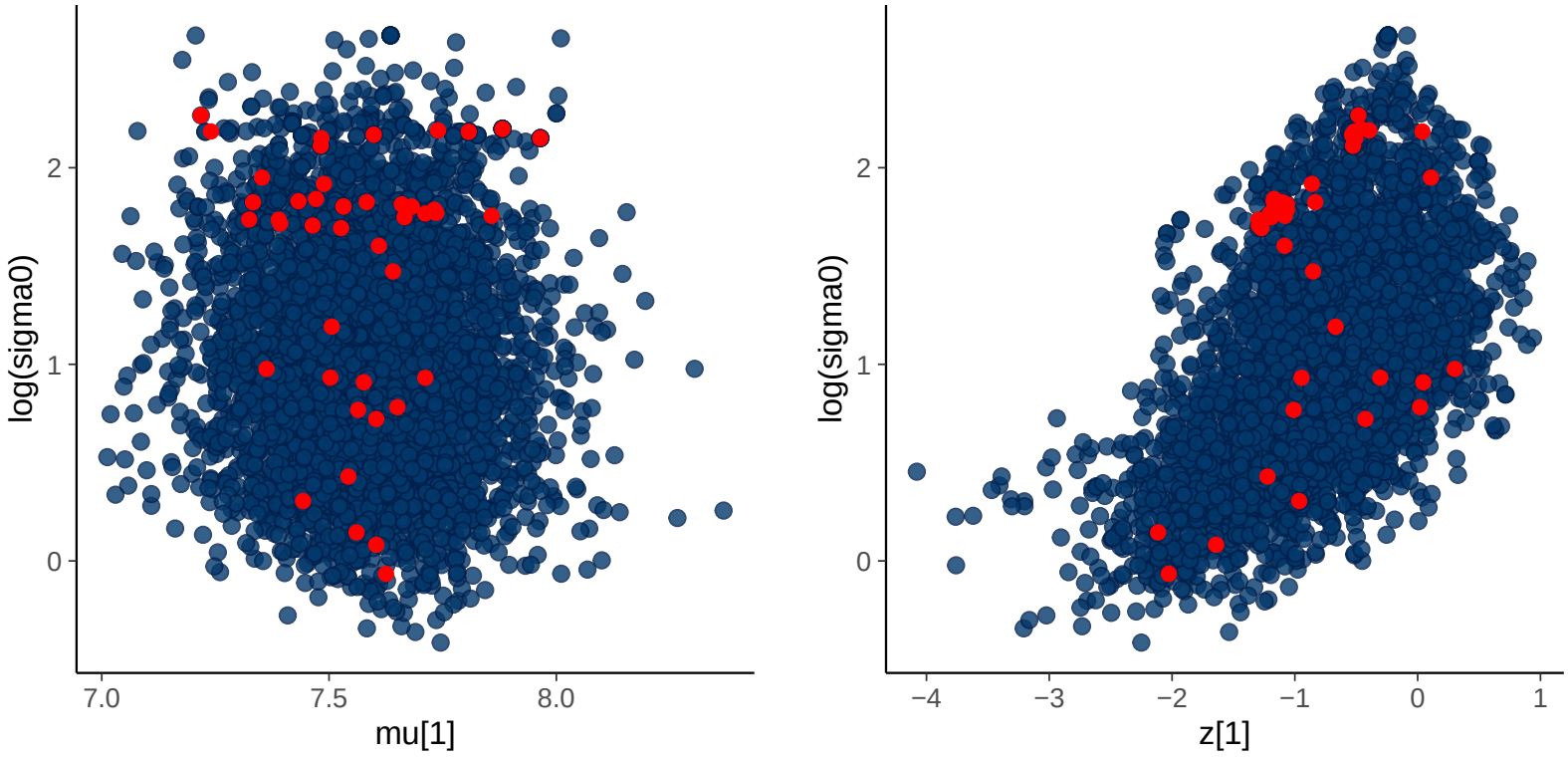
But decreasing step size a lot helps.



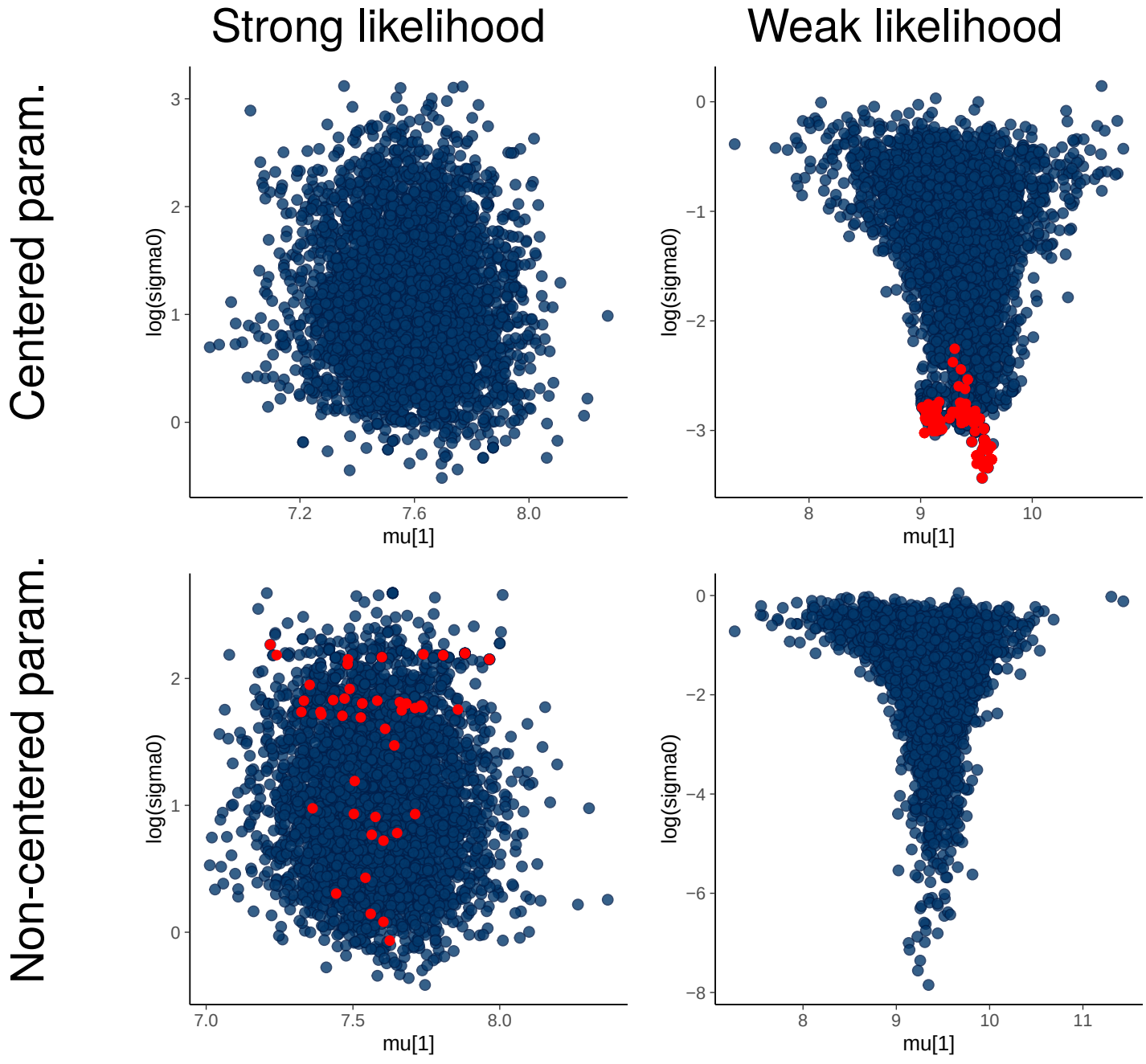
Non-centered parameterization

First data with many observations per group: 3 summer months with each having 71 observations.

Now the posterior for z is problematic.



Centered vs. non-centered parameterization



brms and rstanarm

- brms and rstanarm use non-centered parameterization
  - as hierarchical models and Bayesian inference is most useful when likelihood is weak
- There can be need for both centered and non-centered parameterization in the same model
  - automation not easy, but research goes on

Exchangeability

- Justifies why we can use
  - a joint model for data
  - a joint prior for a set of parameters
- Less strict than independence

Exchangeability

- *Exchangeability*: Parameters $\theta_1, \ldots, \theta_J$ (or observations $y_1, \ldots, y_J$) are exchangeable if the joint distribution p is invariant to the permutation of indices $(1, \ldots, J)$
- e.g.

$$p(\theta_1, \theta_2, \theta_3) = p(\theta_2, \theta_3, \theta_1)$$

- Exchangeability implies symmetry: If there is no information which can be used *a priori* to separate θ_j form each other, we can assume exchangeability. ("Ignorance implies exchangeability")

Exchangeability

- Exchangeability does not mean that the results of the experiments could not be different
  - e.g. if we know that the experiments have been in two different laboratories, and we know that the other laboratory has better conditions for the rats, but we do not know which experiments have been made in which laboratory
  - a priori experiments are exchangeable
  - model could have unknown parameter for the laboratory with a conditional prior for rats assumed to come form the same place (clustering model)

Exchangeability and additional information

- Example: bioassay
  - y_i number of dead animals are not exchangeable alone

Exchangeability and additional information

- Example: bioassay
  - y_i number of dead animals are not exchangeable alone
  - x_i dose is additional information

Exchangeability and additional information

- Example: bioassay
  - y_i number of dead animals are not exchangeable alone
  - x_i dose is additional information
  - (x_i, y_i) exchangeable and logistic regression was used

$$p(\alpha, \beta \mid y, n, x) \propto \prod_{i=1}^{n} p(y_i \mid \alpha, \beta, n_i, x_i) p(\alpha, \beta)$$

Hierarchical exchangeability

- Example: hierarchical rats example
  - all rats not exchangeable

Hierarchical exchangeability

- Example: hierarchical rats example
  - all rats not exchangeable
  - in a single laboratory rats exchangeable

Hierarchical exchangeability

- Example: hierarchical rats example
  - all rats not exchangeable
  - in a single laboratory rats exchangeable
  - laboratories exchangeable

Hierarchical exchangeability

- Example: hierarchical rats example
  - all rats not exchangeable
  - in a single laboratory rats exchangeable
  - laboratories exchangeable
  - $\rightarrow$ hierarchical model

Partial or conditional exchangeability

- Conditional exchangeability
  - if y_i is connected to an additional information x_i, so that y_i are not exchangeable, but (y_i, x_i) exchangeable use joint model or conditional model $(y_i \mid x_i)$.

Partial or conditional exchangeability

- Conditional exchangeability
  - if y_i is connected to an additional information x_i, so that y_i are not exchangeable, but (y_i, x_i) exchangeable use joint model or conditional model $(y_i \mid x_i)$.
- Partial exchangeability
  - if the observations can be grouped (a priori), then use hierarchical model

Exchangeability

- The simplest form of the exchangeability (but not the only one) for the parameters θ conditional independence

$$p(x_1, \ldots, x_J \mid \theta) = \prod_{j=1}^{J} p(x_j \mid \theta)$$

Exchangeability - Counter example

- A six sided die with probabilities $\theta_1, \ldots, \theta_6$
  - without additional knowledge $\theta_1, \ldots, \theta_6$ exchangeable
  - due to the constraint $\sum_{j=1}^{6} \theta_j$, parameters are not independent and thus joint distribution can not be presented as iid

Exchangeability

- See more examples in the BDA3 notes - Exchangeability vs. independence