Applied Bayesian Modeling module 7:
Bayesian multilevel models

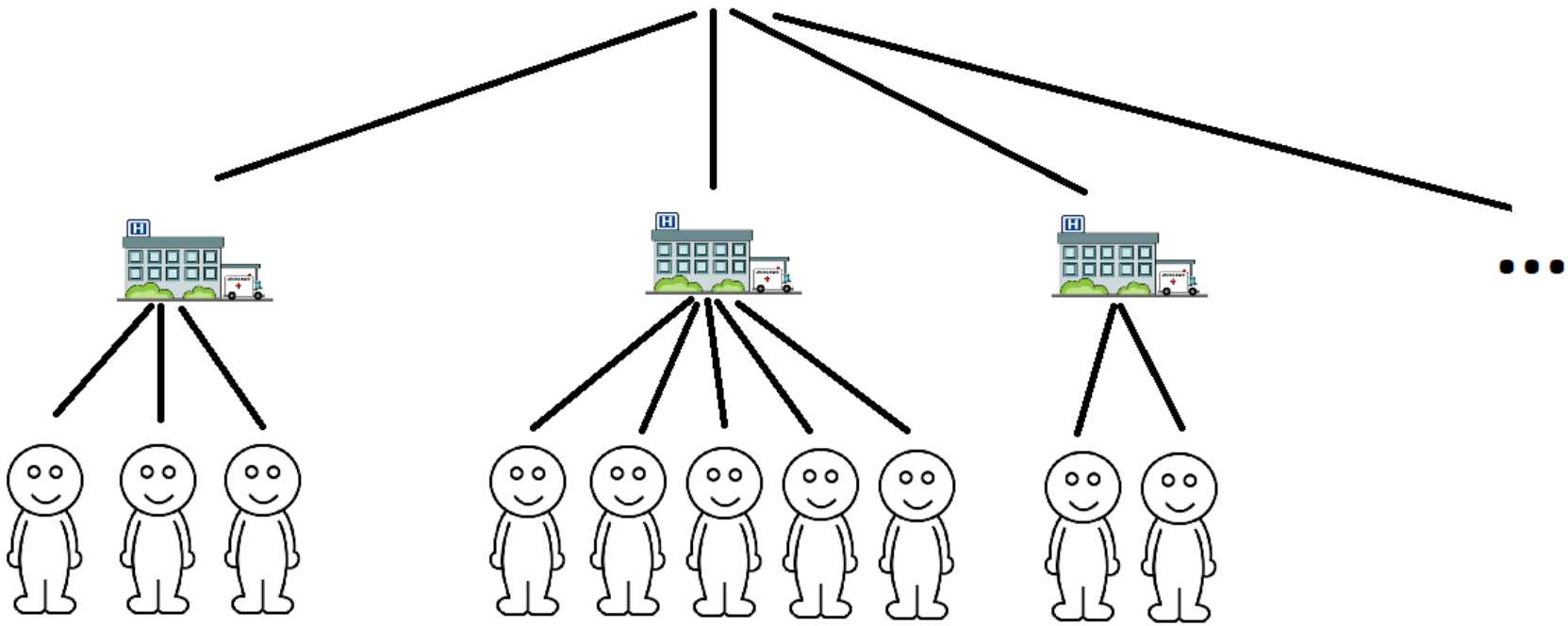
Leontine Alkema, lalkema@umass.edu
Fall 2022

Summary so far

- Suppose that $y_i|\mu, \sigma^2 \sim N(\mu_i, \sigma^2)$
- Part I: Bayesian inference for $\mu_i = \mu$ with σ^2 known
  - Bayesian inference: use probability statements/densities to reflect a state of knowledge
  - Derive the posterior $p(\mu|\boldsymbol{y})$ using Bayes rule: $p(\mu|\boldsymbol{y}) \propto p(\mu)p(\boldsymbol{y}|\mu)$; for $y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$ and $\mu \sim N(m_0, s_{\mu 0}^2)$; then $\mu|\boldsymbol{y}, \sigma^2 \sim N\left(\frac{m_0/s_{\mu 0}^2 + n \cdot \bar{y}/\sigma^2}{1/s_{\mu 0}^2 + n/\sigma^2}, \frac{1}{1/s_{\mu 0}^2 + n/\sigma^2}\right)$
  - Summarize the posterior into point estimates and credible intervals
- Part II: Sampling-based approaches to Bayesian inference, motivated by estimating μ and σ
  - Use an MCMC algorithm to obtain a sample from the posterior distribution of interest; just make sure to check trace plots, Rhat, and effective sample size.
  - Use stan/brms to fit your own Bayesian models
- Part III: Bayesian multilevel/hierarchical models
  - Take account of data hierarchies in specification of μ_i

Multilevel models

- Multilevel models, also referred to as hierarchical models, are commonly used for estimating parameters in settings where there is a hierarchy of nested populations.
- Simplest set-up: two levels, in which one level consist of the groups and the other of units within groups, e.g.
  - patients' health outcomes, where patients are organized in hospitals

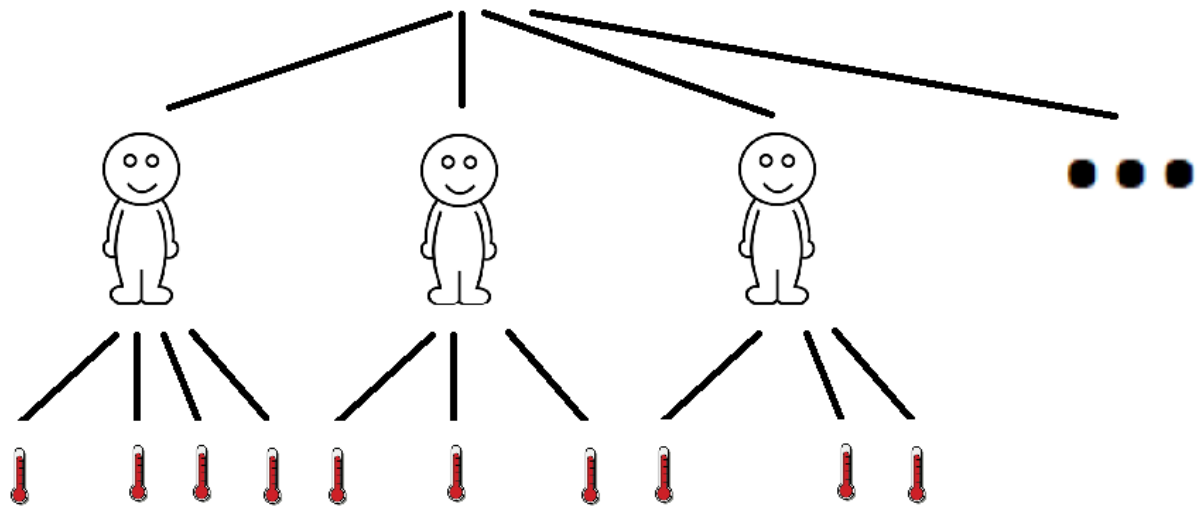


Hierarchy of nested populations

- Examples with more levels:
  - patients' health outcomes, where patients are organized in hospitals in regions within countries
  - student test scores, where students are organized in schools, which are organized in school districts in states within countries
  - maternal mortality within countries within subregions within the world

Other settings where hierarchical models can be used

- Repeated measurements on the same unit of observation, e.g.
  - repeated observations on the same patient over time, for several patients.
- Non-nested structures, when units of observation are characterized by overlapping categories of attributes, e.g.
  - persons by job and state (job status and state are non-nested grouping variables).

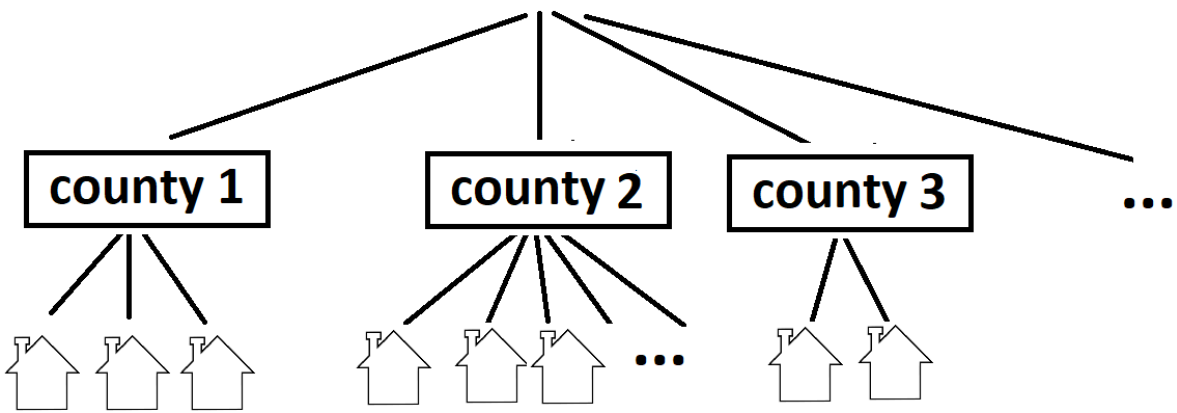


Outline

- This module:
  - Introduction to Bayesian multilevel models:
    a 2-level hierarchical model for estimating group means using normal distributions
  - Alternative ways to write the same model, Bayesian multilevel models vs traditional/frequentist mixed effects models
- Next module:
  - Predictions (for yet-to-be-sampled units or group-level parameters)
  - Bayesian multilevel regression models

Motivating example: Radon measurement (GH Ch.12)

- Data: radon measurements in houses in counties in Minnesota
- Hierarchy: houses observed in counties.
- Q: How to estimate the expected radon level for each county?



Notation

- Note: Some references use alternative notation, we start with this notation because it is the easiest notation to learn main concepts
- unit $i = 1, \ldots, n$ refer to the smallest items of measurement, here household
- outcome y_i, here log(radon) for house $i = 1, 2, \ldots, n$,
- groups are indexed by $j = 1, \ldots, J$, here referring to counties
- index $j[i]$ denotes the county for house i, see toy example below
- $\bar{y}_j = 1/n_j \sum_{i \in G_j} y_i$, the sample mean in the j-th group, with G_j the set of indices for county j and $n_j =$ sample size in group j.

Set-up: $n = 5,\ J = 2$

i	j[i]	y_i	
1	1	y_1	$\bar{y}_1$
2	1	y_2	
3	2	y_3	$\bar{y}_2$
4	2	y_4	
5	2	y_5	

Questions to be answered

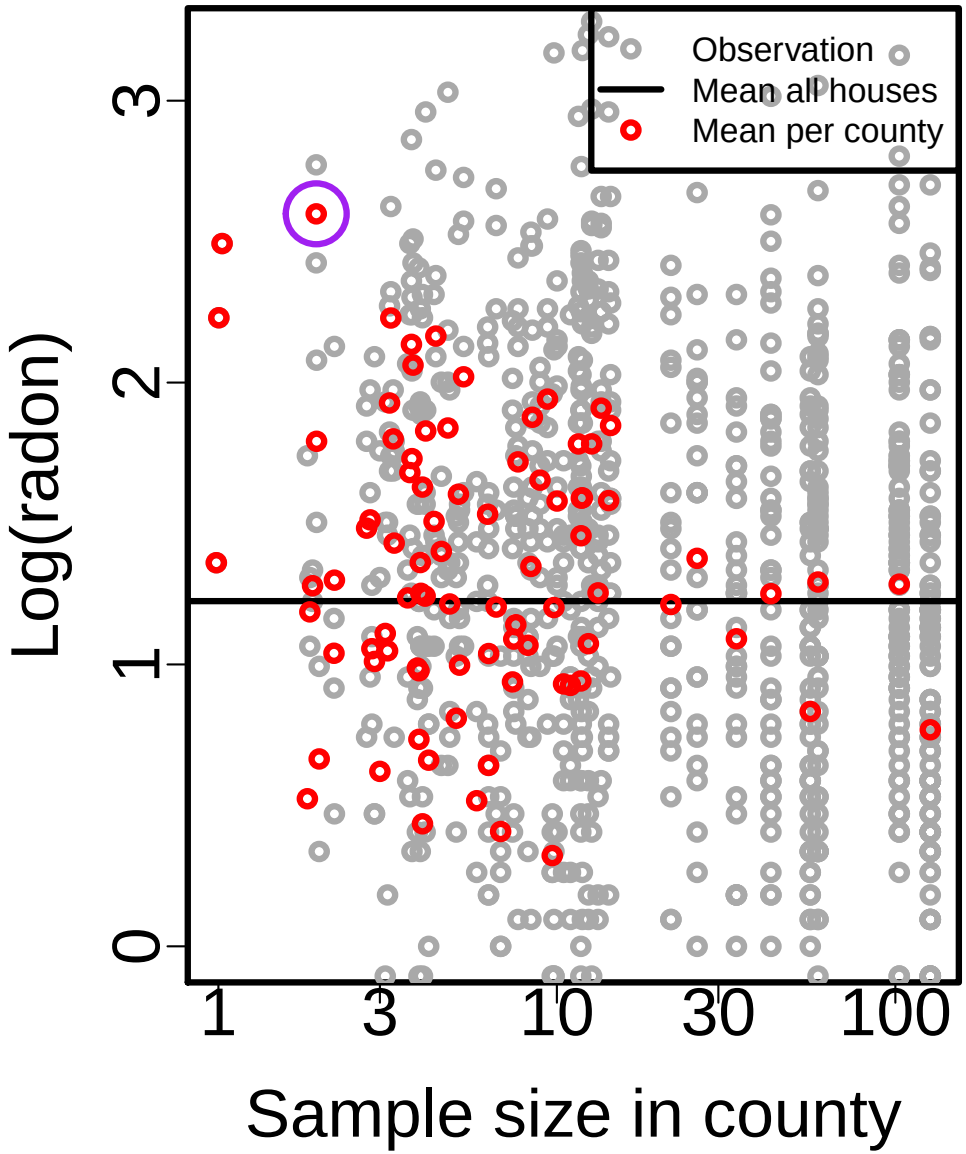
- Let log(radon) be outcome variable y_i measured for house i in county $j[i]$.
- Questions:
  - How to estimate the expected radon level for each county.
  - How to predict the radon level for a not-yet-sampled house i in county $j[i]$?
- Let's assume the y_i's are normally distributed and conditionally independent, $y_i|\mu_i, \sigma_y^2 \sim N(\mu_i, \sigma_y^2)$,
  - introducing subscript y in σ_y because we will use various σ's
- How to estimate group means; what expression to use for μ_i?

Modeling options for estimating mean radon levels

Option 1: Estimate the county-level mean for each county, using only the data from that county.

- Model: $\mu_i = \alpha_{j[i]}^{nopool}$, mean log-radon level in county $j[i]$
- For each j, estimate α_j^{nopool} using data from county j only.
  $\Rightarrow$ Use a vague prior, e.g. $\alpha_j \sim N(0, \text{large variance})$, then $\hat{\alpha}_j^{nopool} \approx \bar{y}_j$, the county sample mean.
- This is referred to as the no pooling (of information across counties) model.
- Cons for counties with small samples sizes:
  - county means are based on very limited information and thus highly uncertain.
  - we may be making the counties look more different than they actually are.

Radon data (grey) and county sample means (red)



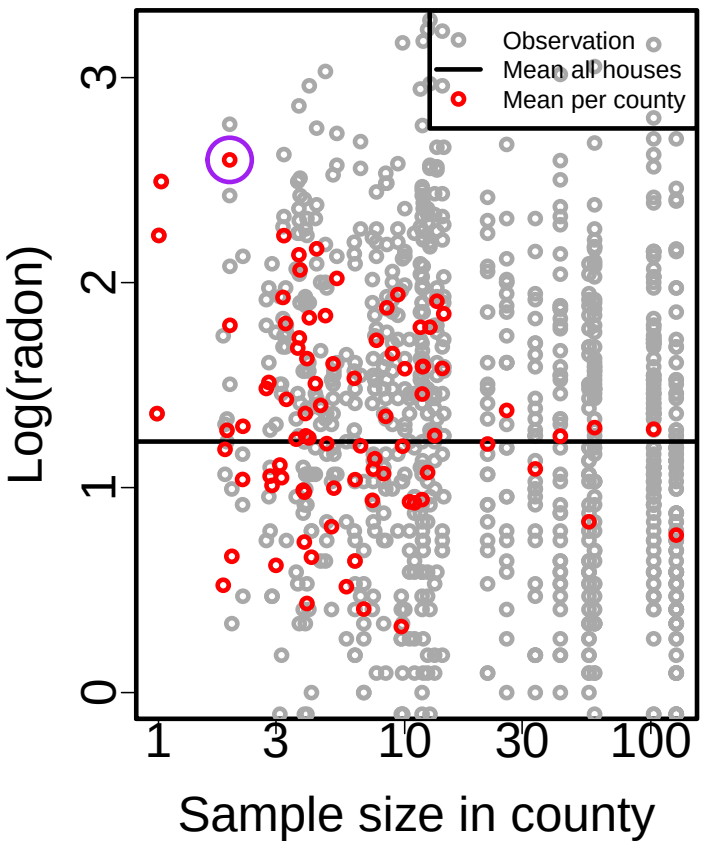
Note the county indicated with purple circle around its sample mean: is radon really that high?

Modeling options for estimating mean radon levels

Option 2: Estimate each county mean using all data in the state

- Model: $\mu_i = \mu$, the state mean
- Estimate with vague prior on μ: $\hat{\mu} = \bar{y} = 1/n \sum_{i=1}^{n} y_i$, the state sample mean.
- This is called the complete pooling (of information across counties) model.
- Con: Ignores across-county variance, may thus result in inaccurate estimates for counties with outlying levels.

Modeling options for estimating mean radon levels: overview



Options for μ_i in $y_i|\mu_i, \sigma_y^2 \sim N(\mu_i, \sigma_y^2)$:

- Option 1: $\mu_i = \alpha_{j[i]}^{nopooling} \sim$ vague prior
  no pooling, use county-specific means, results in the estimation of some county means based on very limited information.
- Option 2: $\mu_i = \mu$,
  complete pooling, use the state mean, understates (ignores) across-county variance, and may thus result in inaccurate estimates for counties with outlying levels.
- Option 3: partial pooling, with a hierarchical model.

Hierarchical/multilevel model

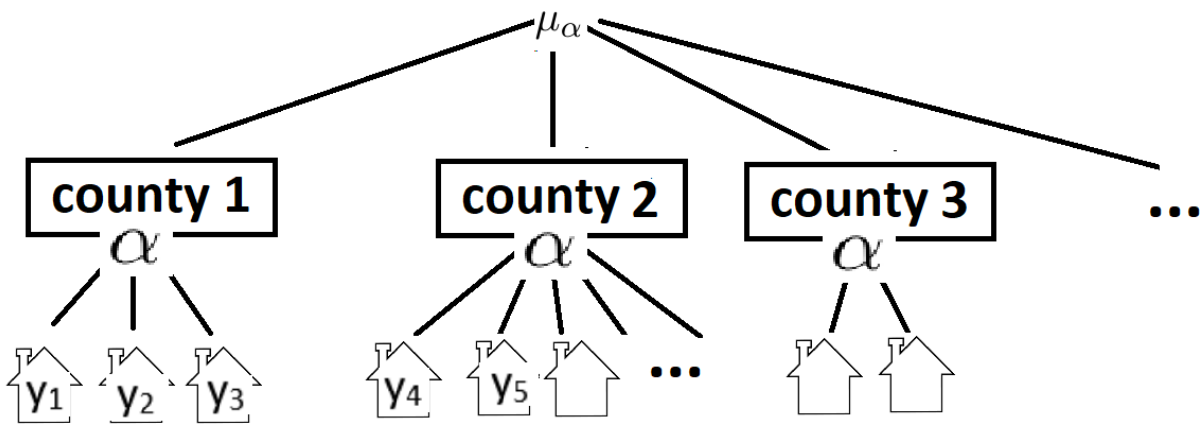
- A hierarchical model for estimating county-level radon is as follows:

$$y_i|\alpha_{j[i]}, \sigma_y \overset{i.i.d}{\sim} N(\alpha_{j[i]}, \sigma_y^2),$$

where

$$\alpha_j|\mu_\alpha, \sigma_\alpha^2 \overset{i.i.d}{\sim} N(\mu_\alpha, \sigma_\alpha^2),$$

  - α_j represents the county-specific mean,
  - μ_α the mean of the county radon levels, and σ_α^2 the between-county variance.
- Because of the hierarchical set-up, the resulting estimates for the county means are in-between the no-pooling and complete-pooling estimates (details to follow).



Full Bayesian model

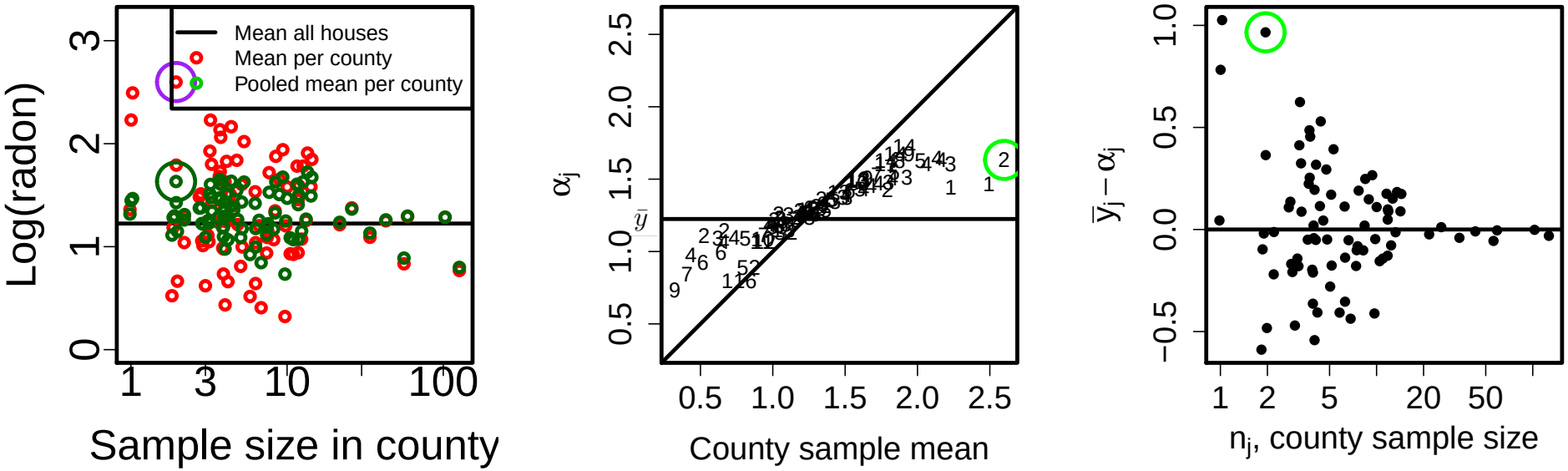
- To fully specify a Bayesian multilevel model that includes

$$y_i|\alpha_{j[i]}, \sigma_y \overset{i.i.d}{\sim} N(\alpha_{j[i]}, \sigma_y^2), \tag{1}$$

$$\alpha_j|\mu_\alpha, \sigma_\alpha \overset{i.i.d}{\sim} N(\mu_\alpha, \sigma_\alpha^2), \tag{2}$$

  we need to set priors for the model parameters $\sigma_y, \mu_\alpha, \sigma_\alpha$.
- We use defaults implemented in brms

Results: Partial pooling in the radon data



- The plots show the Bayesian estimates for the partially pooled means α_j in the multilevel model.
- Finding:
  - partially pooled mean is in between the county sample mean and the state level mean;
    we say that the partially pooled mean is *shrunk* from the county sample mean towards the state level mean
  - the extent of shrinkage of county means from sample mean towards the group mean decreases with sample size.

Partial pooling: more detail

- We can see what's going on with partially pooled means in more detail by obtaining the conditional distr. $\alpha_j|\boldsymbol{y}, \mu_\alpha, \sigma_y, \sigma_\alpha$.
- For the multilevel model

$$y_i|\alpha_{j[i]}, \sigma_y \overset{i.i.d}{\sim} N(\alpha_{j[i]}, \sigma_y^2), \text{ with } \alpha_j|\mu_\alpha, \sigma_\alpha^2 \overset{i.i.d}{\sim} N(\mu_\alpha, \sigma_\alpha^2),$$

the conditional distribution for the j-th state mean is given by:

$$\begin{aligned} \alpha_j|\boldsymbol{y}, \mu_\alpha, \sigma_y, \sigma_\alpha &\sim N(m, v), \\ v &= \left(n_j/\sigma_y^2 + 1/\sigma_\alpha^2\right)^{-1}, \\ m &= v \cdot \left(\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha\right) = \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}, \end{aligned}$$

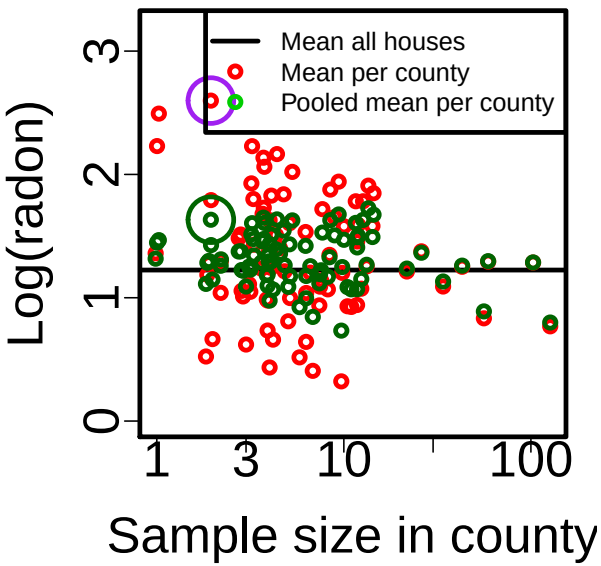
where n_j is the number of observations (houses) in county j.

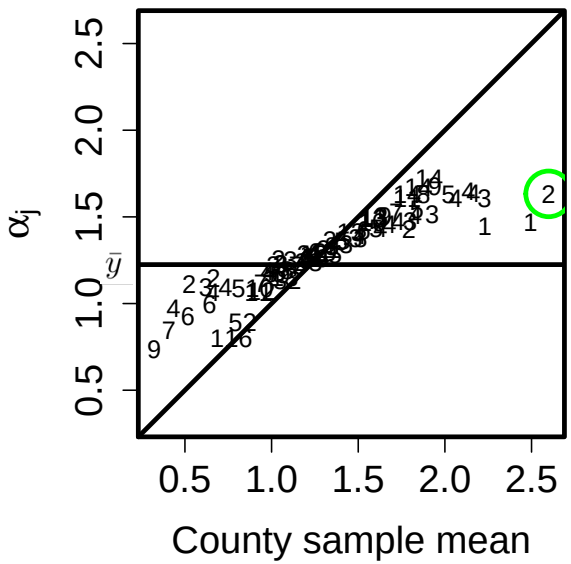
- Really? How to verify?
  - Similar to module 4 (obtaining posterior for normal-normal set-up)
  - Use Bayes rule to express full conditional into knowns pdfs:

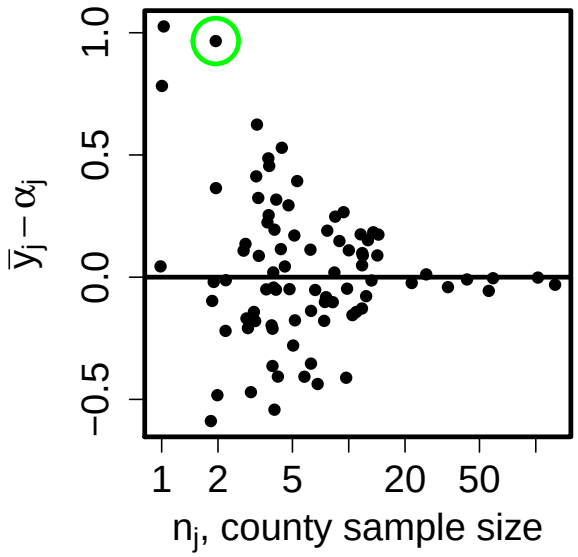
$$p(\alpha_j|\boldsymbol{y}, \mu_\alpha, \sigma_y, \sigma_\alpha) \propto p(\boldsymbol{y}|\alpha_j, \sigma_y)p(\alpha_j|\mu_\alpha, \sigma_\alpha).$$

Interpretation of the conditional mean for α_j

- Given the data $\boldsymbol{y}$, and parameters $\mu_\alpha, \sigma_y, \sigma_\alpha$, the expected value of α_j is given by $m_j = w_j \bar{y}_j + (1 - w_j)\mu_\alpha$, which is a weighted average of
  - county mean $\bar{y}_j$, and "mean of county means" μ_α,
  - with weight $w_j = \frac{n_j}{\sigma_y^2} / \left(\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2} \right)$ increasing with n_j, the number of observations in county j.
- Hence the partially pooled estimate α_j is shrunk from the county mean towards the state mean for counties with small sample size.







Hierarchical models: summary so far

- When working with data y_i on units that are organized in groups, the following hierarchical model can be used to estimate group means α_j for groups $j = 1, 2, \ldots, J$:

$$y_i|\alpha_{j[i]}, \sigma_y \overset{i.i.d}{\sim} N(\alpha_{j[i]}, \sigma_y^2), \qquad \alpha_j|\mu_\alpha, \sigma_\alpha^2 \overset{i.i.d}{\sim} N(\mu_\alpha, \sigma_\alpha^2),$$

  where μ_α, the mean of the group means, and σ_α^2, the between-group variance, are estimated.
- The resulting mean estimates $\hat{\alpha}_j$ are in-between the no-pooling and complete-pooling estimates and referred to as partially pooled estimates:
  - the partially pooled mean is *shrunk* from the county sample mean towards the state level mean
  - the extent of shrinkage of county means from sample mean towards the group mean decreases with sample size.
- Remainder of this slide set: alternative way of writing the same model (helps to interpret `brm` model output), Bayesian vs traditional/frequentist approach

Two ways to specify the same model for group-level parameters

- Two ways to write a multilevel model for group mean α_j:
  Approach A (centered parametrization)

$$\alpha_j | \mu_\alpha, \sigma^2_\alpha \sim N(\mu_\alpha, \sigma^2_\alpha).$$

  Approach B (non-centered parametrization)

$$\begin{aligned} \alpha_j &= \mu_\alpha + \eta_j, \\ \eta_j | \sigma^2_\alpha &\sim N(0, \sigma^2_\alpha). \end{aligned}$$

- `brm` (and `lmer`) fits report estimates associated with approach B, with any group-level parameters that are estimated with a hierarchical model assumed to have mean zero.

Bayesian multilevel models vs traditional "mixed" models

- In traditional/frequentist inference, group-level parameters can also be assigned a distribution

$$\begin{aligned} \alpha_j &= \mu_\alpha + \eta_j, \\ \eta_j|\sigma_\alpha^2 &\sim N(0, \sigma_\alpha^2). \end{aligned}$$

  - The η_j's are commonly referred to as random effects
  - Models with random and fixed effects are referred to as mixed effects models.
- Multilevel models fit naturally into a Bayesian framework, where all parameters are random.
  - Gelman et al use the term "modeled parameters" for those parameters (because they are assigned a model as opposed to a prior).
  - In the context of hierarchical models, parameters that do not vary across groups are referred to as hyperparameters
  - In module 8 we will discuss how to obtain a density for modeled parameters, including for yet-to-be-sampled groups, that includes uncertainty in hyperparameters

Model fitting using `brm`

- Function call: `fit <- brm(y ~ (1|county), ...)`
  - in the formula, add "`|grouping variable`" to any covariate (here the intercept) for which you would like to estimate group specific parameters, using a hierarchical model
- Output summary includes info on group-level effects ($\hat{\sigma}_\eta = 0.32$), pop-level effects ($\hat{\mu}_\alpha = 1.31$), and family-specific parameters (here $\hat{\sigma}_y = 0.81$).

```
## Group-Level Effects:
## ~county (Number of levels: 85)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.32      0.05     0.23     0.42 1.00      951     1539
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     1.31      0.05     1.21     1.41 1.00     1237     1450
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.81      0.02     0.77     0.85 1.00     3042     1566
##
```