

# Chapter 5

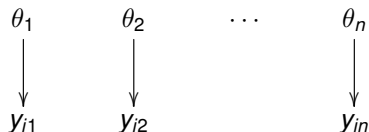
- 5.1 Lead-in to hierarchical models
- 5.2 Exchangeability (useful concept)
- 5.3 Bayesian analysis of hierarchical models (but we use Stan for computation)
- 5.4 Hierarchical normal model (but we use Stan for computation)
- 5.5 Example: parallel experiments in eight schools (uses hierarchical normal model, part of assignment, but we use Stan for computation)
- 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

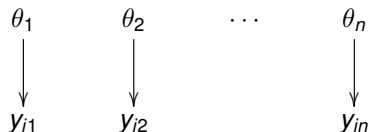
# Hierarchical model

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

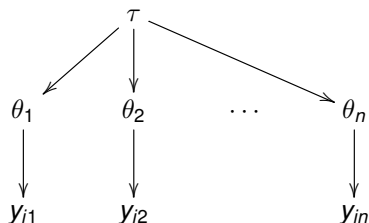


# Hierarchical model

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



- sensible to assume that  $\theta_j$  are similar



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

# Hierarchical model: terms

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



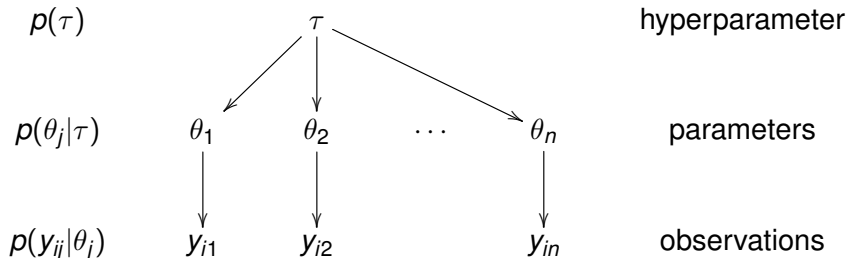
Joint posterior

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

# Hierarchical model: terms

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

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

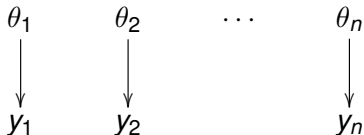


Joint posterior

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

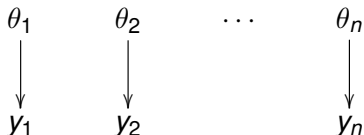
# Compare

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

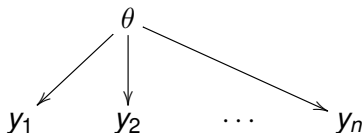


# Compare

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



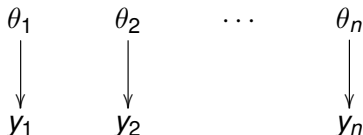
- "Joint model" (model with a common effect / pooled model)



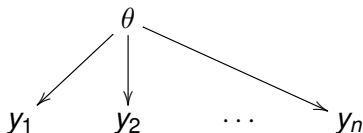


# Compare

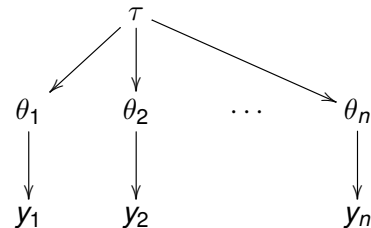
- "Separate model" (model with separate/independent effects)



- "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

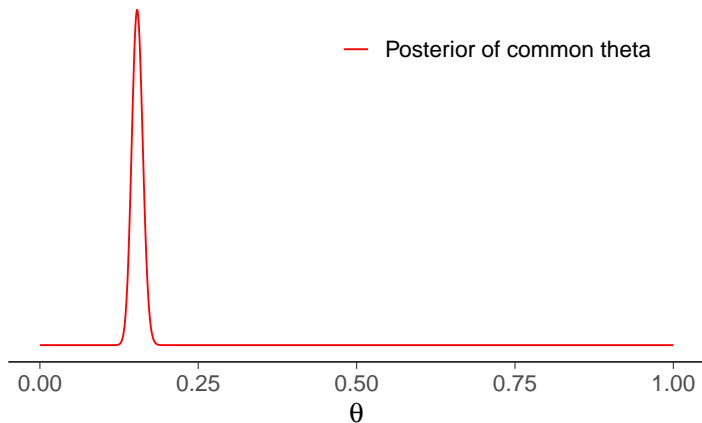
# Hierarchical binomial model: rats

- 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									

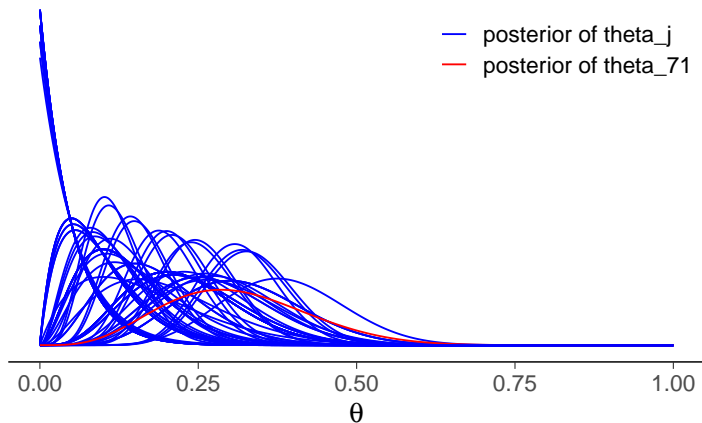
# Hierarchical binomial model: rats

Pooled model



# Hierarchical binomial model: rats

Separate model



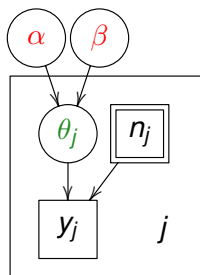
# Hierarchical binomial model: rats

- Hierarchical binomial model for rats  
prior parameters  $\alpha$  and  $\beta$  are unknown

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

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

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

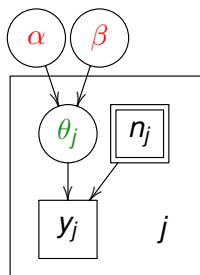


# Hierarchical binomial model: rats

- Hierarchical binomial model for rats  
prior parameters  $\alpha$  and  $\beta$  are unknown

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

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



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

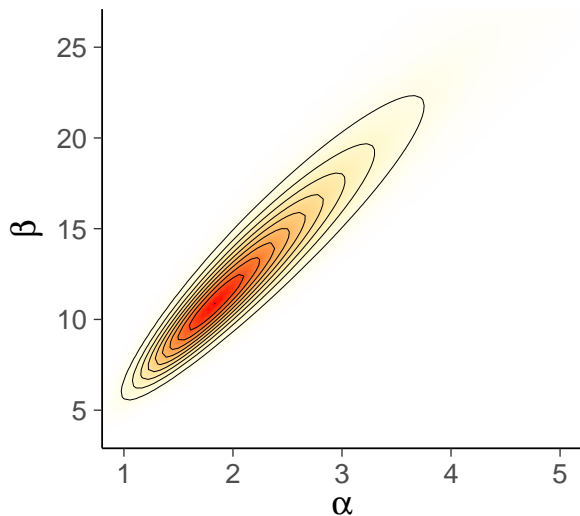
# Hierarchical binomial model: rats

- Population prior  $\text{Beta}(\theta_j | \alpha, \beta)$
- Hyperprior  $p(\alpha, \beta)$ ?
  - $\alpha, \beta$  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



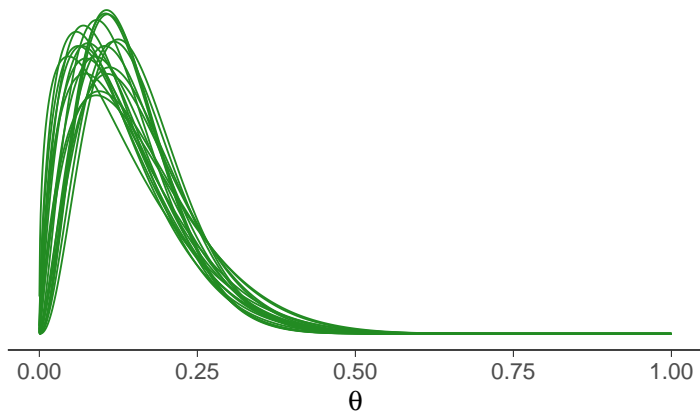
## Hierarchical binomial model: rats

The marginal of  $\alpha$  and  $\beta$



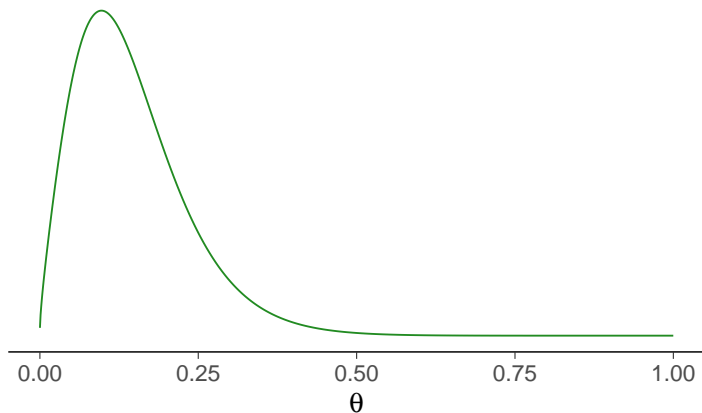
# Hierarchical binomial model: rats

Beta( $\alpha, \beta$ ) given posterior draws of  $\alpha$  and  $\beta$



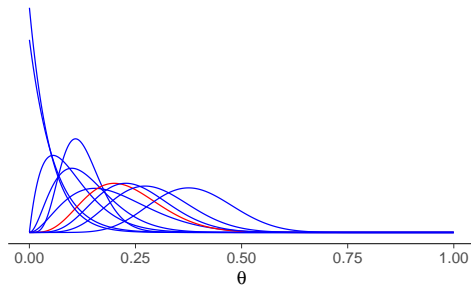
# Hierarchical binomial model: rats

Population distribution (prior) for  $\theta_j$



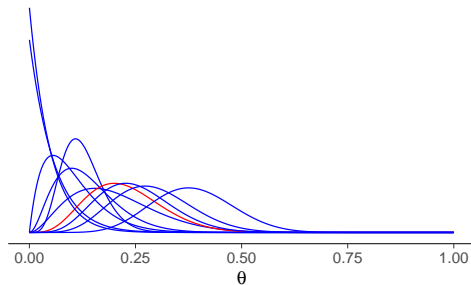
# Hierarchical binomial model: rats

Separate model

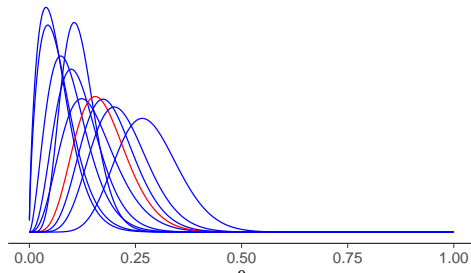


# Hierarchical binomial model: rats

Separate model

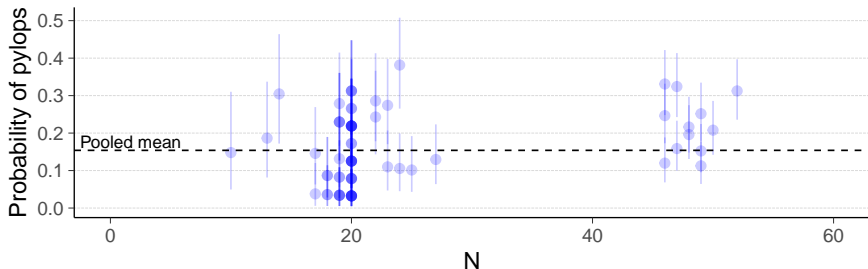


Hierarchical model

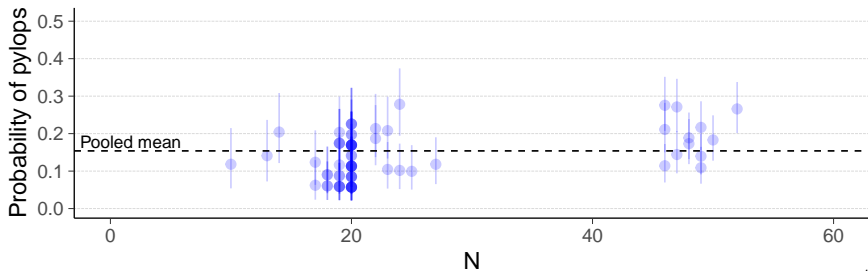


# Hierarchical model and group size: Rats

Separate

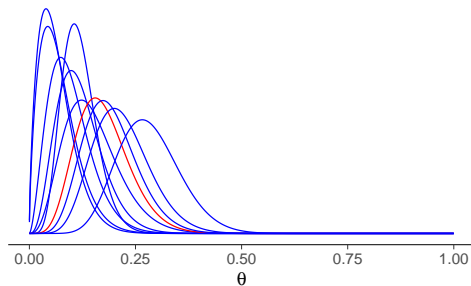


Hierarchical

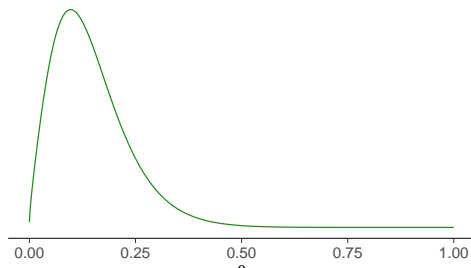


# Hierarchical binomial model: rats

Hierarchical model



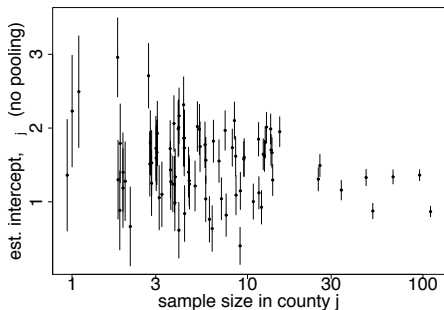
Population distribution (prior) for  $\theta_j$



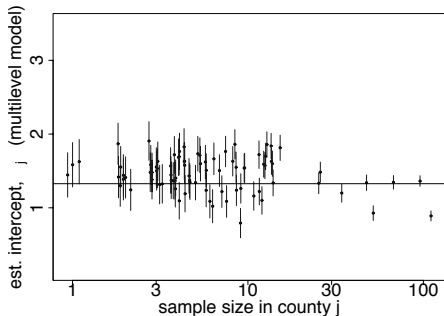
# Hierarchical model and group size: Radon

919 home radon levels in 85 counties in Minnesota:

Separate



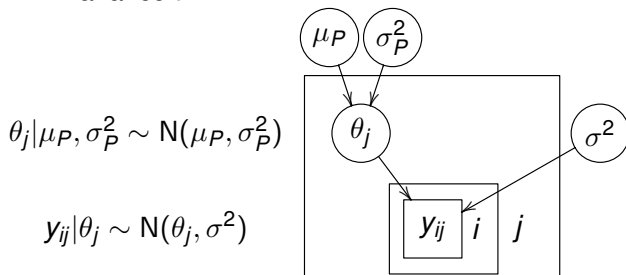
Hierarchical





# Hierarchical normal model: factory

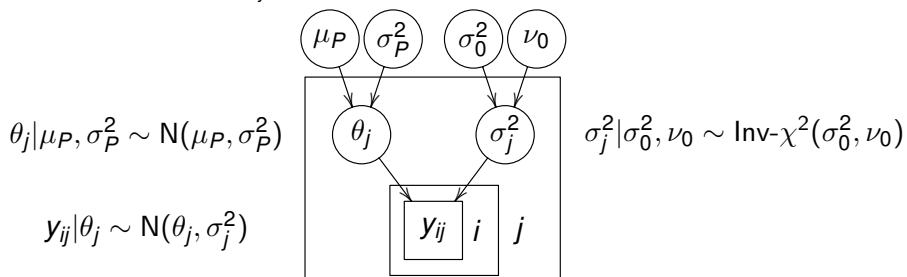
- Factory has 6 machines which quality is evaluated
- Assume hierarchical model
  - each machine has its own (average) quality  $\theta_j$  and common variance  $\sigma^2$



- Can be used to predict the future quality produced by each machine and quality produced by a new similar machine

# Hierarchical normal model: factory

- Factory has 6 machines which quality is evaluated
- Assume hierarchical model
  - each machine has its own (average) quality  $\theta_j$  and own variance  $\sigma_j^2$



- Can be used to predict the future quality produced by each machine and quality produced by a new similar machine

# Hierarchical normal model: 8 schools

- Example: SAT coaching effectiveness
  - in USA commonly used Scholastic Aptitude Test (SAT) is designed so that short term practice should not improve the results significantly
  - schools have anyway coaching courses
  - test the effectiveness of the coaching courses

# Hierarchical normal model: 8 schools

- Example: SAT coaching effectiveness
  - in USA commonly used Scholastic Aptitude Test (SAT) is designed so that short term practice should not improve the results significantly
  - schools have anyway coaching courses
  - test the effectiveness of the coaching courses
- SAT
  - standardized multiple choice test
  - mean about 500 and standard deviation about 100
  - most scores between 200 and 800
  - different topics, e.g., V=Verbal, M=Mathematics
  - pre-test PSAT

# Hierarchical normal model: 8 schools

- Effectiveness of the SAT coaching
  - students had made pre-tests PSAT-M and PSAT-V
  - part of students were coached
  - linear regression was used to estimate the coaching effect  $y_j$  for the school  $j$  (could be denoted with  $\bar{y}_{.j}$ , too) and variances  $\sigma_j^2$
  - $y_j$  approximately normally distributed, with variances assumed to be known based on about 30 students per school
  - data is group means and variances (not personal results)

# Hierarchical normal model: 8 schools

- Effectiveness of the SAT coaching
  - students had made pre-tests PSAT-M and PSAT-V
  - part of students were coached
  - linear regression was used to estimate the coaching effect  $y_j$  for the school  $j$  (could be denoted with  $\bar{y}_{.j}$ , too) and variances  $\sigma_j^2$
  - $y_j$  approximately normally distributed, with variances assumed to be known based on about 30 students per school
  - data is group means and variances (not personal results)

• Data:	School	A	B	C	D	E	F	G	H
	$y_j$	28	8	-3	7	-1	1	18	12
	$\sigma_j$	15	10	16	11	9	22	20	28

# Hierarchical normal model for group means

- $J$  experiments, unknown  $\theta_j$  and known  $\sigma^2$

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2), \quad i = 1, \dots, n_j; \quad j = 1, \dots, J$$

- Group  $j$  sample mean and sample variance

$$\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

$$\sigma_j^2 = \frac{\sigma^2}{n_j}$$

# Hierarchical normal model for group means

- $J$  experiments, unknown  $\theta_j$  and known  $\sigma^2$

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2), \quad i = 1, \dots, n_j; \quad j = 1, \dots, J$$

- Group  $j$  sample mean and sample variance

$$\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

$$\sigma_j^2 = \frac{\sigma^2}{n_j}$$

- Use model

$$\bar{y}_{.j}|\theta_j \sim N(\theta_j, \sigma_j^2)$$

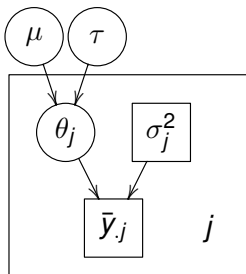
this model can be generalized so that,  $\sigma_j^2$  can be different from each other for other reasons than  $n_j$



# Hierarchical normal model for group means

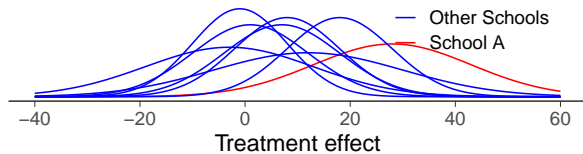
$$\theta_j | \mu, \tau \sim N(\mu, \tau)$$

$$\bar{y}_{.j} | \theta_j \sim N(\theta_j, \sigma_j^2)$$



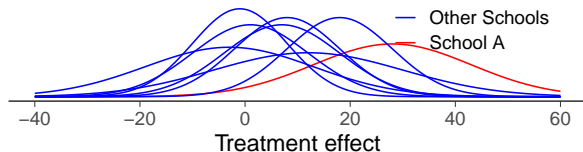
# Hierarchical normal model: 8 schools

Separate model

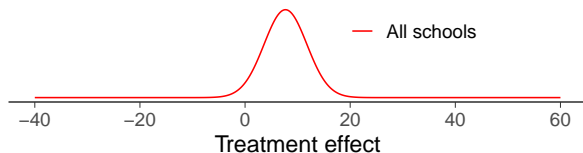


# Hierarchical normal model: 8 schools

Separate model

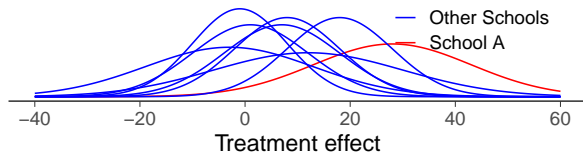


Pooled model

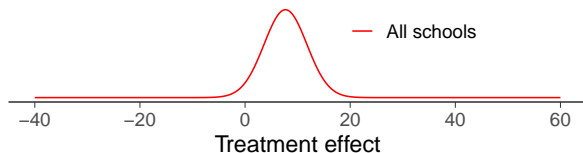


# Hierarchical normal model: 8 schools

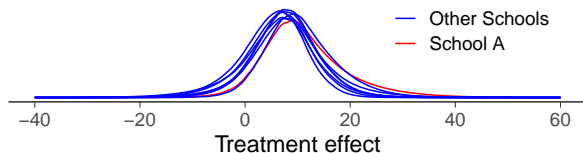
Separate model



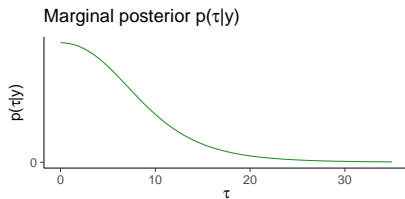
Pooled model



Hierarchical model

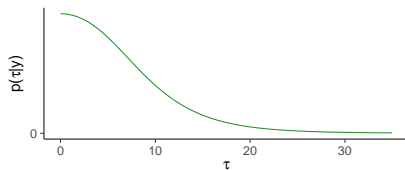


# Hierarchical normal model: 8 schools

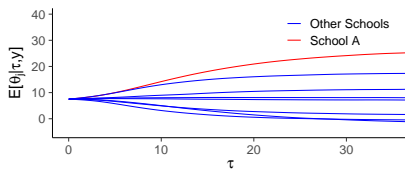


# Hierarchical normal model: 8 schools

Marginal posterior  $p(\tau|y)$

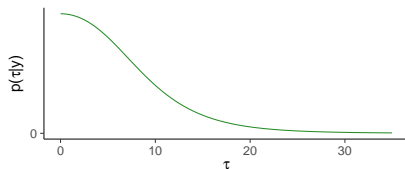


Conditional means  $E[\theta_i|\tau, y]$

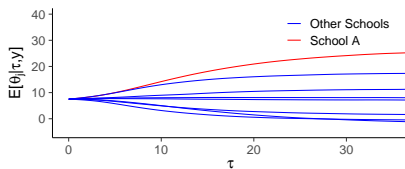


# Hierarchical normal model: 8 schools

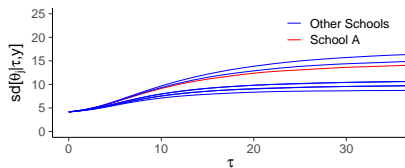
Marginal posterior  $p(\tau|y)$



Conditional means  $E[\theta_j|\tau, y]$



Conditional standard deviations  $sd[\theta_j|\tau, y]$



# 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  $\theta_1, \dots, \theta_J$  (or observations  $y_1, \dots, y_J$ ) are exchangeable if the joint distribution  $p$  is invariant to the permutation of indices  $(1, \dots, J)$
- 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  $\theta_j$  from 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 from 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

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

# 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
  - → 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  $(y_i|x_i)$ .

# 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  $(y_i|x_i)$ .
- 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  $\theta$  conditional independence

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

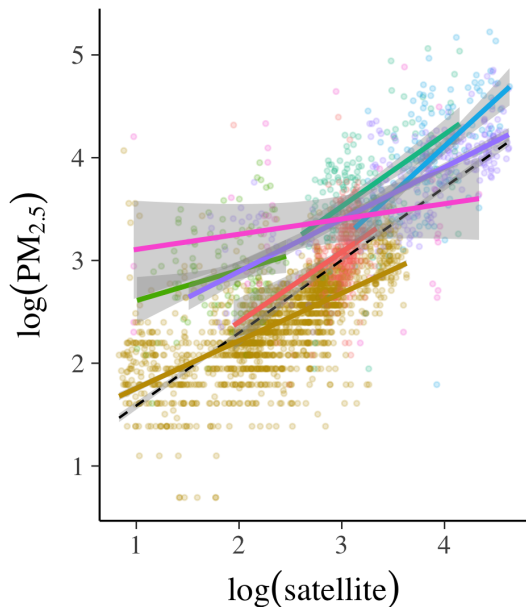
# Exchangeability - Counter example

- A six sided die with probabilities  $\theta_1, \dots, \theta_6$ 
  - without additional knowledge  $\theta_1, \dots, \theta_6$  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

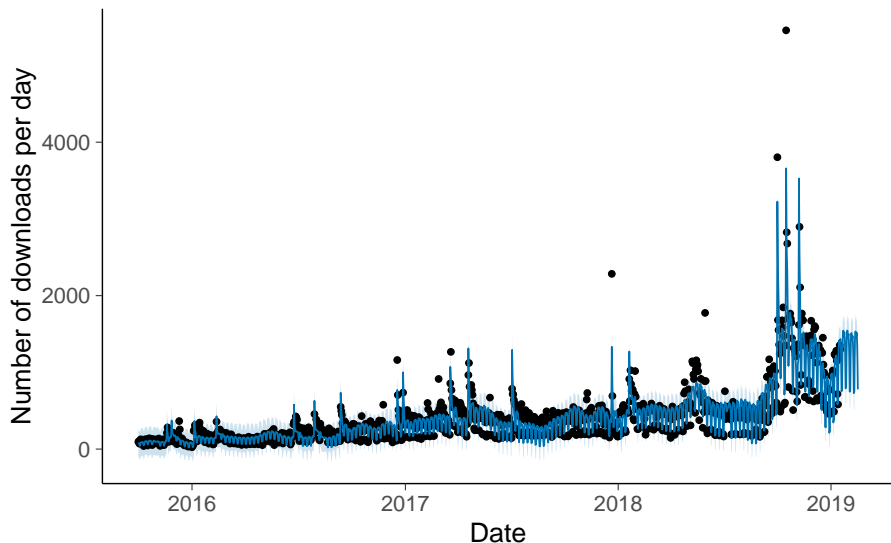
- See more examples in the BDA\_notes\_ch5.pdf

# Hierarchical linear model



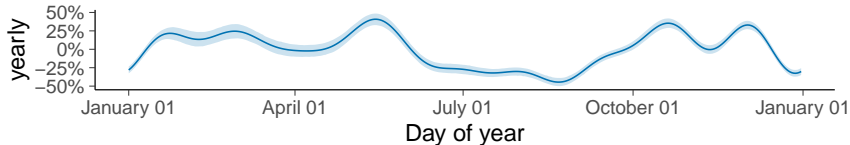
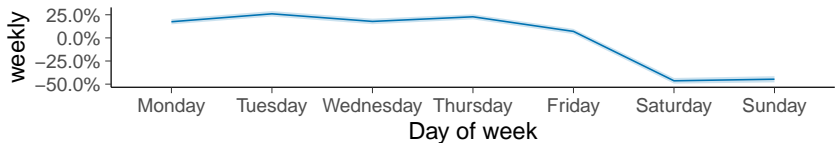
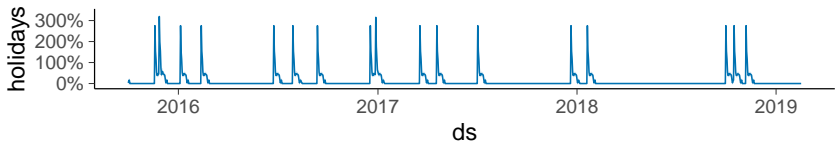
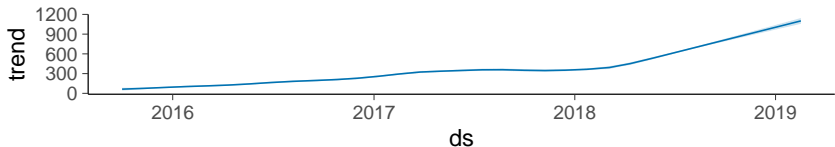
# Hierarchical models and smooths

RStan downloads per day from RStudio CRAN mirror





# Hierarchical models and smooths

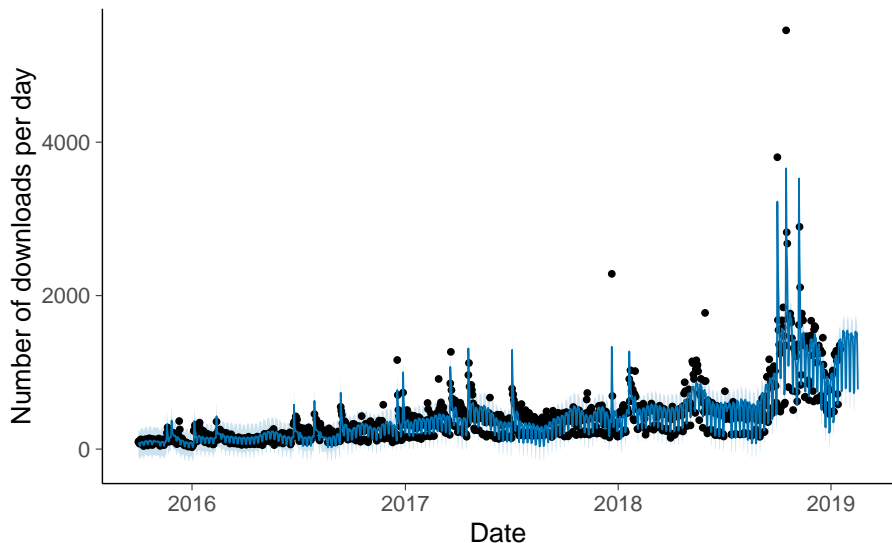


# Hierarchical models and Bayesian deep learning

- Prior scale on each weight layer is a hyperparameter
  - connections between priors and drop-out / stochastic optimization
  - connections between posterior draws and big networks / ensembles

# Hierarchical models and smooths

RStan downloads per day from RStudio CRAN mirror



# Hierarchical models and smooths

