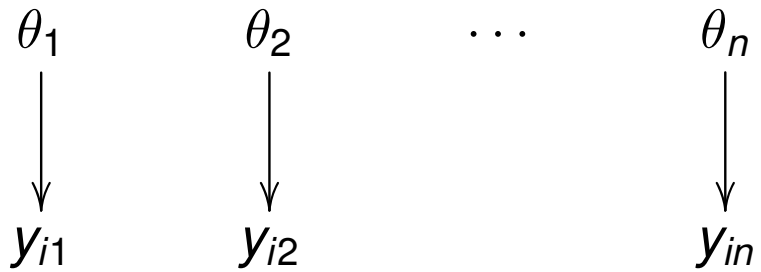
# Chapter 5

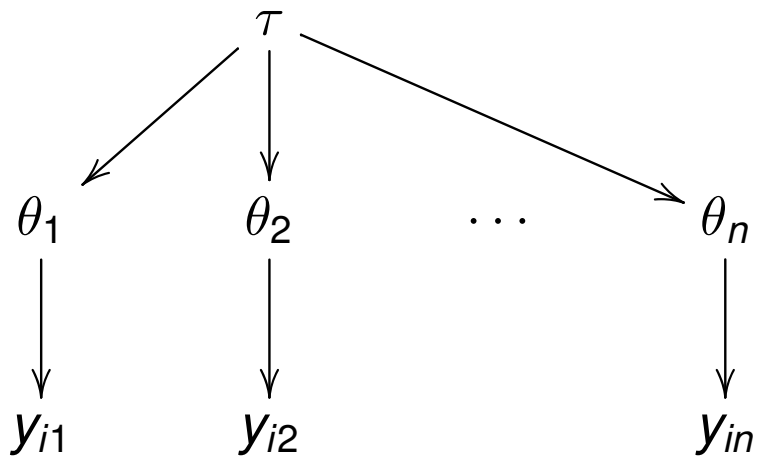
- 5.1 Lead-in to hierarchical models
- 5.2 Exchangeability (important and difficult concept)
- 5.3 Bayesian analysis of hierarchical models
- 5.4 Hierarchical normal model
- 5.5 Example: parallel experiments in eight schools (uses hierarchical normal model, part of exercises)
- 5.6 Meta-analysis (can be skipped)
- 5.7 Weakly informative priors for hierarchical variance 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 $j$




  - sensible to assume that $\theta_j$ are similar

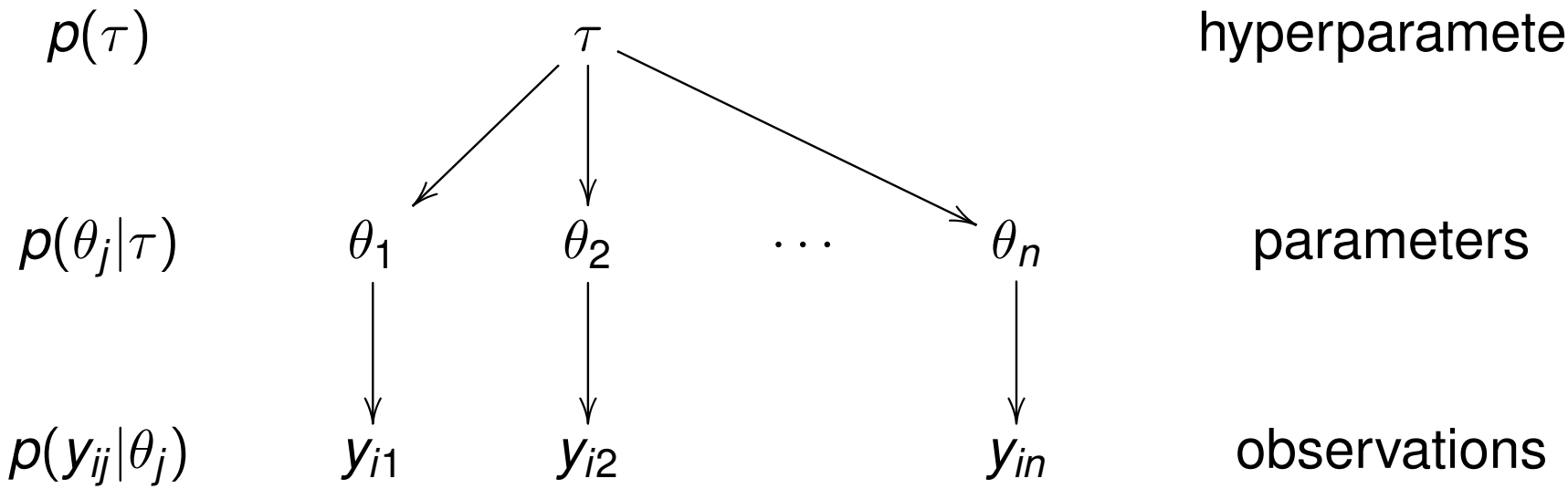



  - natural to think that $\theta_j$ have common population distribution
  - $\theta_j$ is not directly observed and the population distribution is unknown

# Hierarchical model: terms

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

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

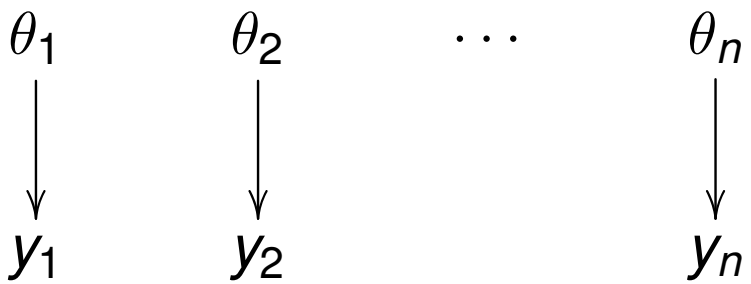



Joint posterior

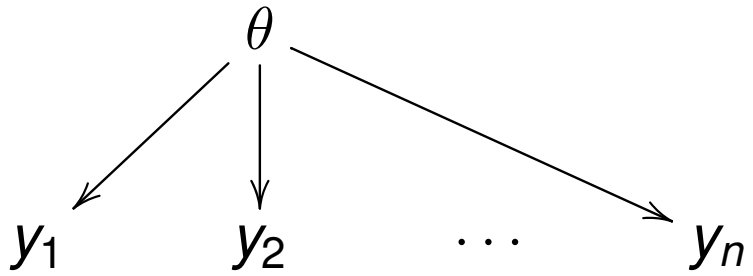
$$\begin{aligned} p(\theta, \tau|y) &\propto p(y|\theta, \tau)p(\theta, \tau) \\ &\propto p(y|\theta)p(\theta|\tau)p(\tau) \end{aligned}$$

# Compare

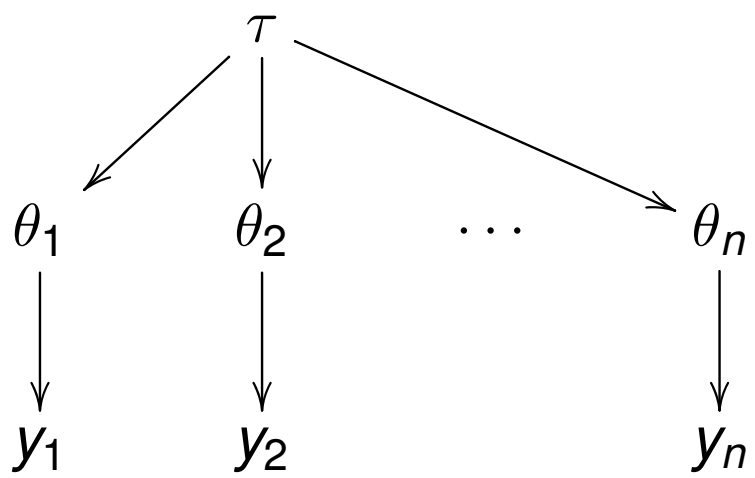
- "Separate model" (model with separate/independent effects)




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




- Hierarchical model

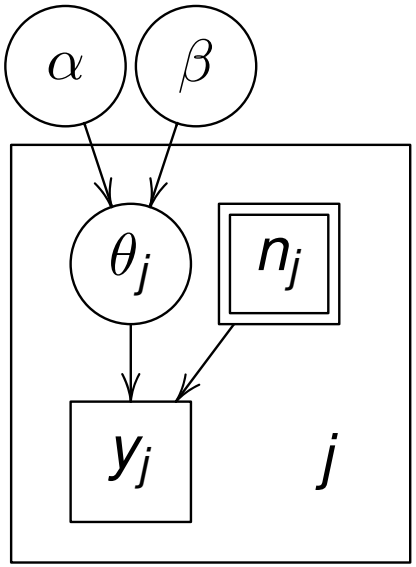


# Hierarchical model: rats

- Hierarchical model for rats

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

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



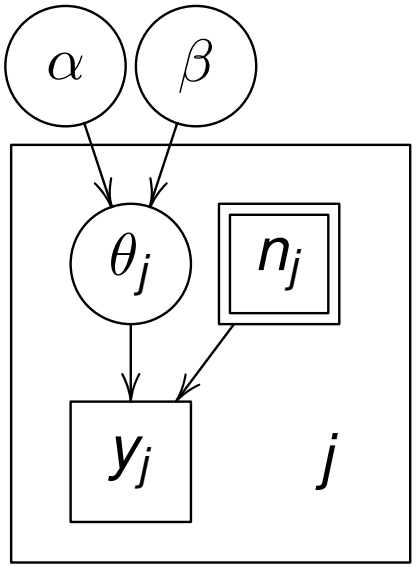

- Joint posterior $p(\theta_1,\ldots,\theta_J,\alpha,\beta|y)$
  - multiple parameters

# Hierarchical model: rats

- Hierarchical model for rats

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

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




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

# Hierarchical model: rats

- Population prior 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 s. 110)
- demo5_1.m

# Hierarchical model: computation

- Factorize joint posterior

$$p(\theta, \tau|y) = p(\theta|\tau, y)p(\tau|y)$$

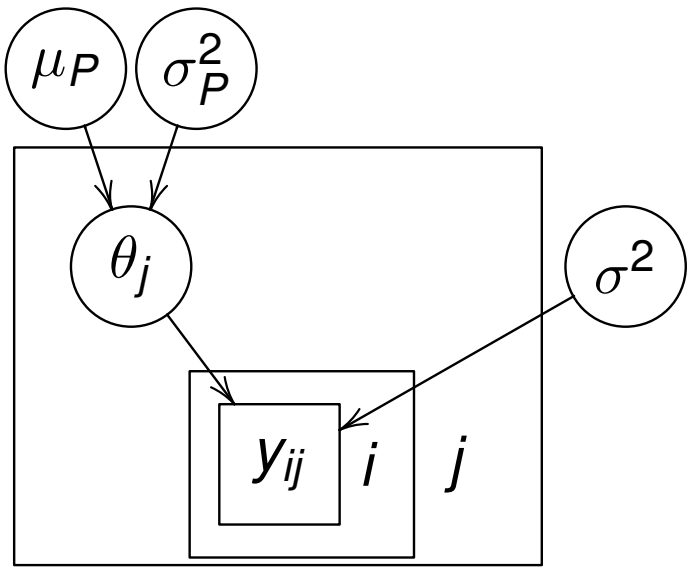
- Sample
  1. sample $\tau^{(t)}$ from $p(\tau|y)$
  2. sample $\theta^{(t)}$ from $p(\theta|\tau^{(t)}, y)$
  3. if needed sample $y^{(t)}$ from $p(y|\theta^{(t)})$
     - repeat $L$ times to get $L$ samples

# 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 $\sigma^2$

$$\theta_j|\mu_P, \sigma_P^2 \sim \mathrm{N}(\mu_P, \sigma_P^2)$$

$$y_{ij}|\theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$$



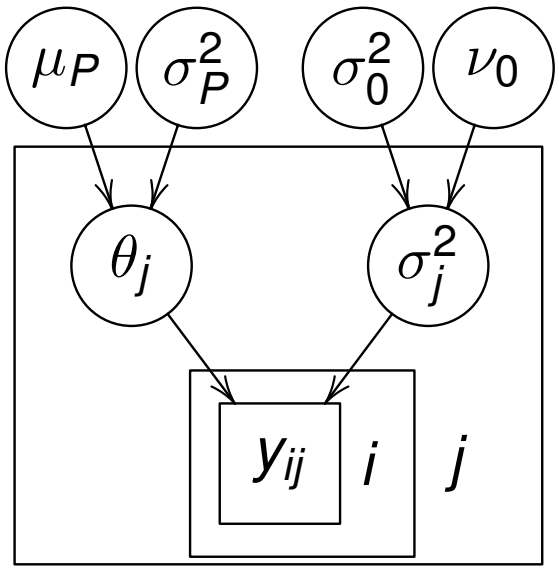

- Can be used to predict the future quality produced by each machine and quality produced by a new similar machine

# Hierarchical normal model: factory

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

$$\theta_j|\mu_P, \sigma_P^2 \sim \mathrm{N}(\mu_P, \sigma_P^2)$$

$$y_{ij}|\theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$$




$$\sigma_j^2|\sigma_0^2, \nu_0 \sim \text{Inv-}\chi^2(\sigma_0^2, \nu_0)$$

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

# Hierarchical normal mixture model for group means

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

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

- Group $j$ sample mean and sample variance

$$\begin{aligned} \bar{y}_{.j} &= \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \\ \sigma_j^2 &= \frac{\sigma^2}{n_j} \end{aligned}$$

# Hierarchical normal mixture model for group means

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

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

- Group $j$ sample mean and sample variance

$$\begin{aligned} \bar{y}_{.j} &= \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \\ \sigma_j^2 &= \frac{\sigma^2}{n_j} \end{aligned}$$

- Use model

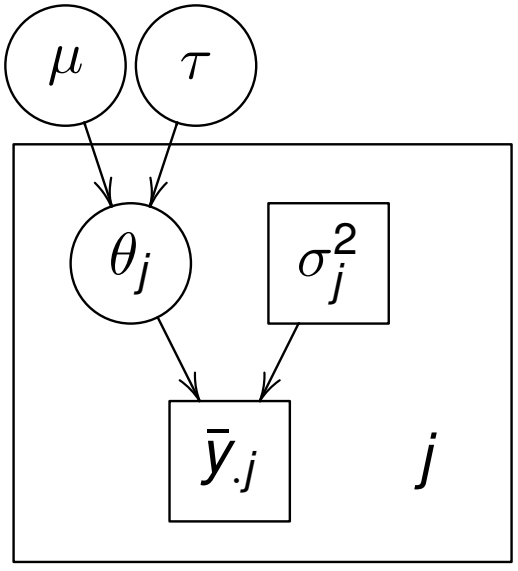
$$\bar{y}_{.j}|\theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2)$$

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

# Hierarchical normal mixture model for group means

$$\theta_j | \mu, \tau \sim \mathsf{N}(\mu, \tau)$$

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



# Hierarchical normal mixture model for group means

- Model

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

  - can also be used if means $\bar{y}_{.j}$ are assumed to be normally distributed, although $y_{ij}$ is not assumed to be normally distributed

# Hierarchical normal prior

- Semi-conjugate prior

$$p(\theta_1, \ldots, \theta_J | \mu, \tau) = \prod_{j=1}^{J} \mathrm{N}(\theta_j | \mu, \tau^2)$$

  - if $\tau \to \infty$, then we get a separate model (*erillismalli*), that is, each $\theta_j$ is estimaied separately using non-informative prior
  - if $\tau \to 0$, then we get a pooled model (*yhteismalli*), that is, $\theta_j = \mu$ ja $\bar{y}_{.j} | \mu \sim \mathrm{N}(\mu, \sigma_j^2)$

# Hierarchical normal model – hyperprior

- Model

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

- Semi-conjugate prior

$$p(\theta_1, \ldots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} \mathrm{N}(\theta_j|\mu, \tau^2)$$

- Hyperprior

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

  - uniform prior for $\mu$ ok
  - prior for $\tau$ need to be selected more carefully to get a proper posterior
  - $p(\tau) \propto 1/\tau$ would produce improper posterior
  - if $J > 4$, $p(\tau) \propto 1$ is ok non-informative prior
  - if $J \leq 4$, half-Cauchy is reasonable weakly-informative prior

# Hierarchical normal model – factorization

- Factorize joint posterior

$$p(\theta, \mu, \tau|y) \propto p(\theta|\mu, \tau, y)p(\mu, \tau|y)$$

- Conditional posterior of $\theta_j$

$$\theta_j|\mu, \tau, y \sim \mathrm{N}(\hat{\theta}_j, V_j)$$

where $\hat{\theta}_j$ and $V_j$ same as for $J$ independent normal distributions given informative conjugate prior

  - that is, precisions weighted mean of data and prior

# Hierarchical normal model – factorization

- Hyperparameter marginal posterior

$$p(\mu, \tau|y) \propto p(\mu, \tau) \prod_{j=1}^{J} \mathsf{N}(\bar{y}_{.j}|\mu, \sigma_j^2 + \tau^2)$$

- Further factorization

$$p(\mu, \tau|y) = p(\mu|\tau, y)p(\tau|y)$$

where

$$p(\mu|\tau, y) = N(\hat{\mu}, V_\mu)$$

where $\hat{\mu}$ is precisions weighted mean of $\bar{y}_{.j}$ and $V_\mu$ is the total precision

- Final part

$$p(\tau|y) = \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)}$$

has no closed form, but it is easy to sample from univariate distribution, for example, using inverse cdf

# Hierarchical normal model – factorization

- Easy to sample using the factorization

$$p(\theta, \mu, \tau|y) \propto p(\tau|y)p(\mu|\tau, y)p(\theta|\mu, \tau, y)$$

- See "Computation" BDA3 s. 118

# Hierarchical normal model – factorization

- 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 – factorization

- 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
  - standardised 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 – factorization

- 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 shool $j$ (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 – factorization

- 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 shool $j$ (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 |

- demo5_2.m

# 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 $\theta_1, \ldots, \theta_J$ are exchangeable if the joint distribution $p$ is invariant to the permutation of indeces $(1, \ldots, J)$
- 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 $\theta_j$ 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 laboratorioes, and we know that the other laboratoriy 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

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

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi)$$

- marginal distribution of $\theta$

$$p(\theta) = \int \left[ \prod_{j=1}^{J} p(\theta_j|\phi) \right] p(\phi) d\phi$$

- mixture of iid distributions

# Exchangeability

- Let $(X_n)_{n=1}^{\infty}$ to be an infinite sequence of exchangeable random variables. De Finetti's theorem then says that there is some random variable $\theta$ so that $X_i$ are conditionally independent given $\theta$, and joint density for $X_1, \ldots, X_J$ can be written in the *iid mixture* form

$$p(x) = \int \left[ \prod_{j=1}^{J} p(x_j|\theta) \right] p(\theta) d\theta$$

- Counter example: A six sided die with probabilities (a finite sequence!) $\theta_1, \ldots, \theta_6$
  - 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 mixture

# Exchangeability

- See examples in the comments5.pdf

# Exchangeability and additional information

- Example: bioassay
  - $x_i$ dose
  - $y_i$ number of dead animals
  - $(x_i, y_i)$ 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: CVD treatment effectiveness
  - all patients not exchangeable
  - in a single hospital patients exchangeable
  - hospitals exchangeable
  - $\rightarrow$ hierarchical model

# Partial or conditional exchangeability

- Partial exchangeability
  - if the observations can be grouped (a priori), then use hierarchical model
- 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 $(y_i|x_i)$.