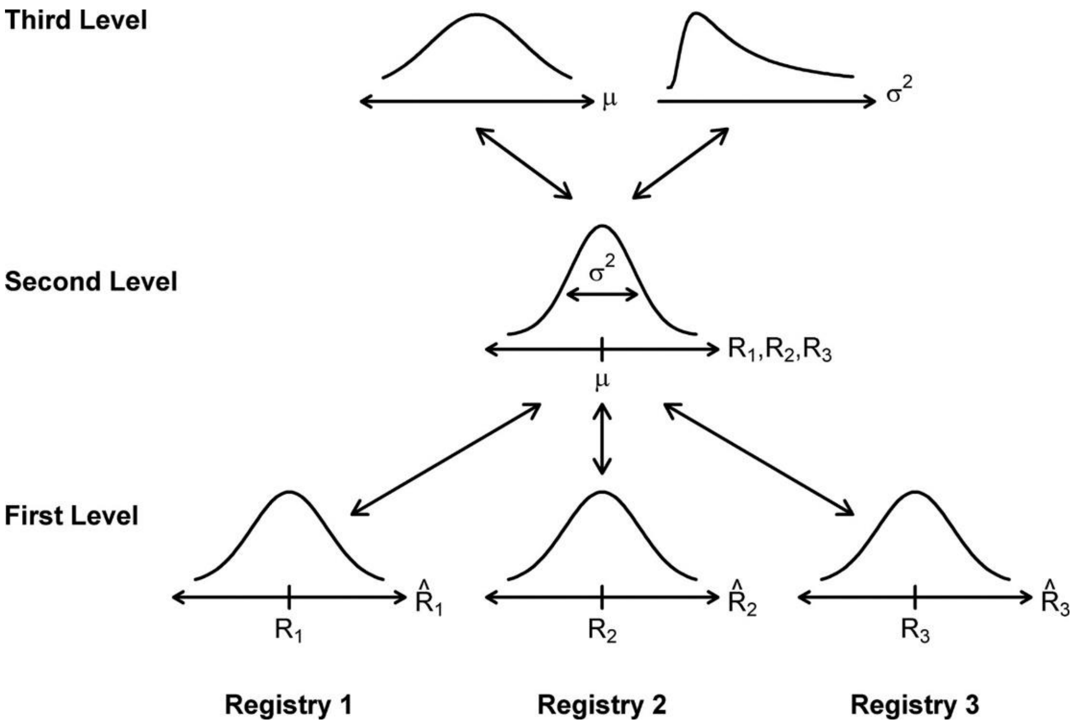
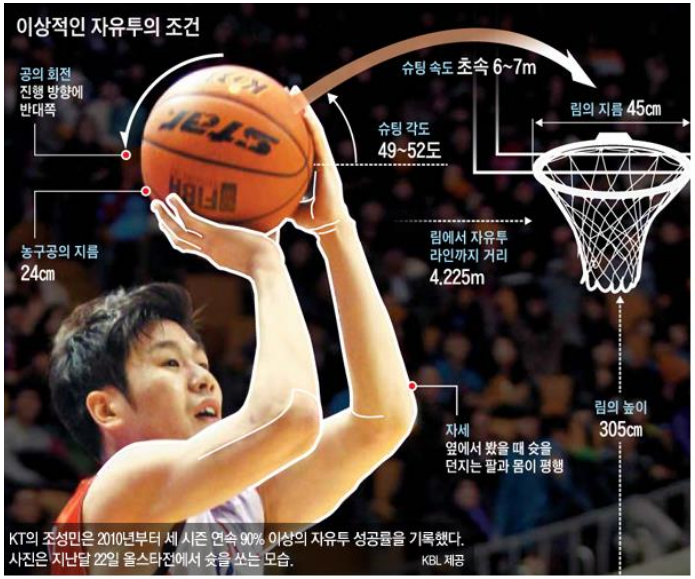
L4. Hierarchical Bayesian models



Why Hierarchical models?



KT의 조성민은 2010년부터 세 시즌 연속 90% 이상의 자유투 성공률을 기록했다.
사진은 지난달 22일 올스타전에서 슛을 쏘는 모습. KBL 제공

Seasons	Made	Attempts
2012-2013	25	46
2013-2014	41	93
2014-2015	93	176
2015-2016	79	120

Is his free-throw percentage higher this year than past years?

Why Hierarchical models?

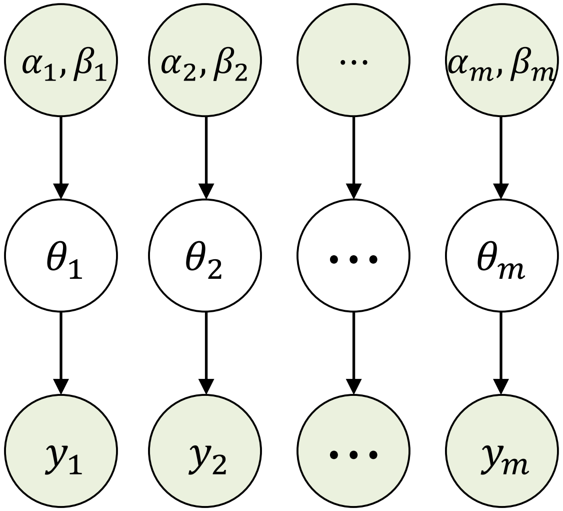
Likelihood :

$$Y_i \sim \mathrm{Bin}(n_i, \theta_i) \rightarrow p(y_i|\theta_i) = \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i}$$

$$p(y|\theta) = \prod_{i=1}^{m} p(y_i|\theta_i) = \prod_{i=1}^{m} \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i}$$

Prior:

$$p(\theta) = \prod_{i=1}^{m} p(\theta_i) = \prod_{i=1}^{m} \mathrm{Beta}(\alpha_i, \beta_i) = \prod_{i=1}^{m} \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \theta_i^{\alpha_i-1}(1-\theta_i)^{\beta_i-1}$$



$$\theta = (\theta_1, \dots, \theta_m)$$

$$y = (y_1, \dots, y_m)$$

Why Hierarchical models?

Likelihood :

$$Y_i \sim \mathrm{Bin}(n_i, \theta_i) \rightarrow p(y_i|\theta_i) = \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i}$$

$$p(y|\theta) = \prod_{i=1}^{m} p(y_i|\theta_i) = \prod_{i=1}^{m} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n_i - y_i}$$

Prior:

$$p(\theta) = \prod_{i=1}^{m} p(\theta_i) = \prod_{i=1}^{m} \mathrm{Beta}(\alpha_i, \beta_i) = \prod_{i=1}^{m} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} \theta_i^{\alpha_i - 1} (1-\theta_i)^{\beta_i - 1}$$

Posterior :

$$\begin{aligned} P(\theta|y) &\propto P(y|\theta) p(\theta) \\ &= \prod_{i=1}^{m} p(y_i|\theta_i) \prod_{i=1}^{m} p(\theta_i) \\ &= \prod_{i=1}^{m} p(y_i|\theta_i)\, p(\theta_i) \\ &= \prod_{i=1}^{m} \mathrm{Beta}(\theta_i | \alpha_i + y_i, \beta_i + n_i - y_i) \end{aligned}$$

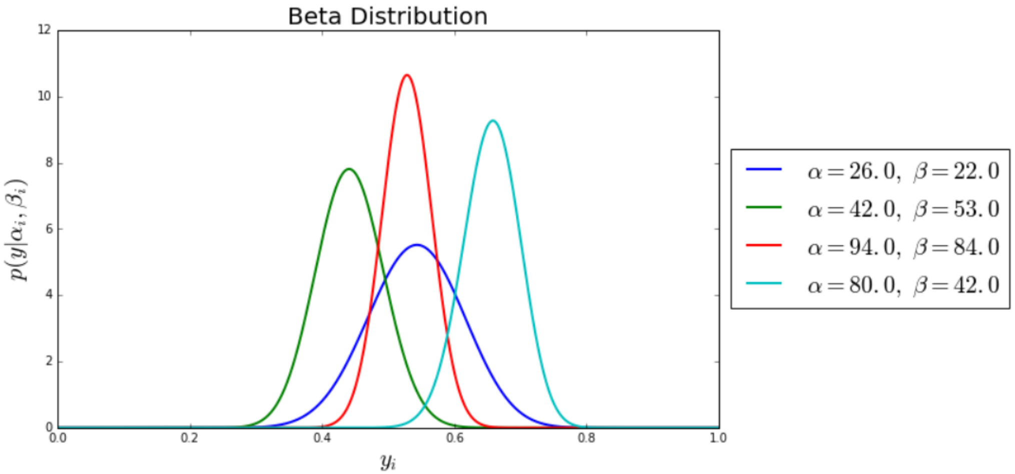
So the posterior for each θ_i is exactly the same as if we treated each season independently

Why Hierarchical models?

Assuming $\theta_i \sim \mathrm{Beta}(\alpha_i = 1, \beta_i = 1)$ for all i, results in independent Beta posterior

Seasons	Made	Attempts
2012-2013	25	46
2013-2014	41	93
2014-2015	93	176
2015-2016	79	120

Seasons	α_i	β_i
2012-2013	26	22
2013-2014	42	53
2014-2015	94	84
2015-2016	80	42



Is there any way to use the data from the previous seasons for estimating the success probability for the current season?

Why Hierarchical models?

Likelihood :

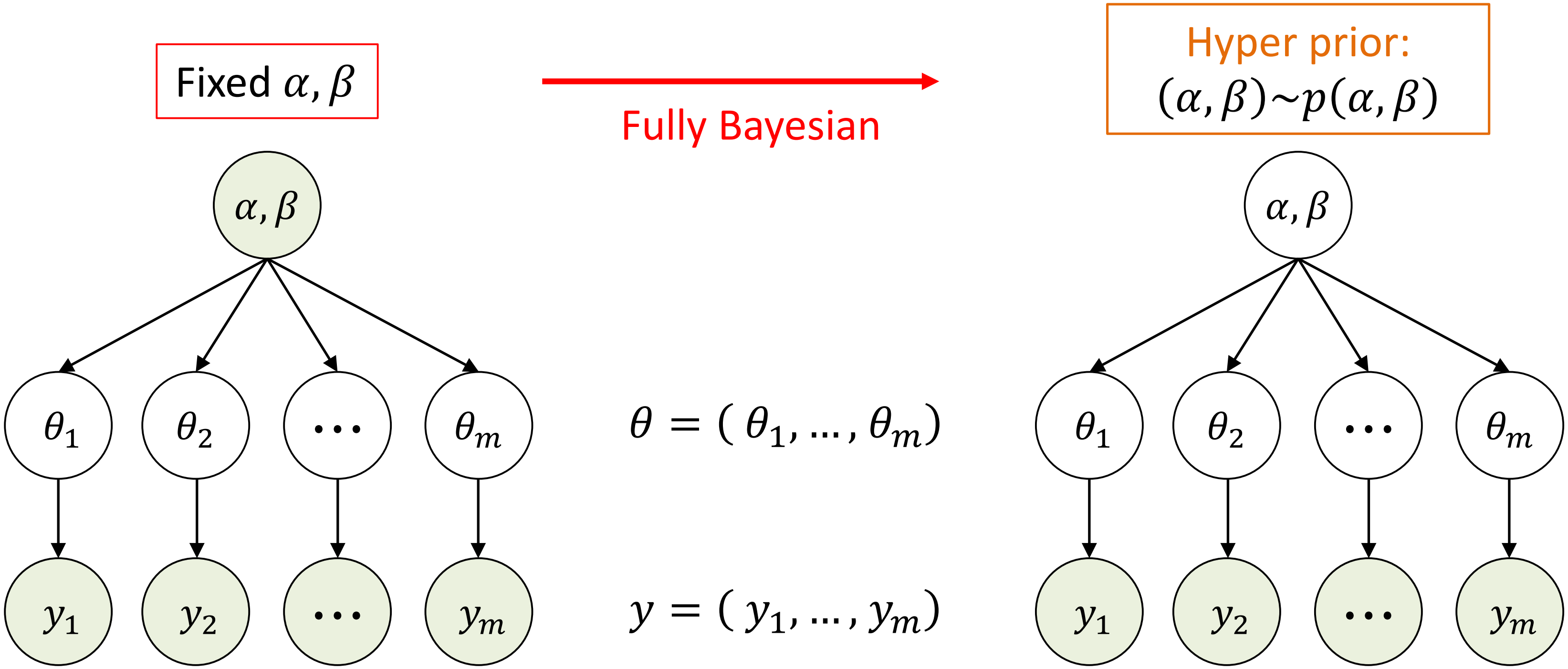
$$Y_i \sim \mathrm{Bin}(n_i, \theta_i) \rightarrow p(y_i|\theta_i) = \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i}$$

$$p(y|\theta) = \prod_{i=1}^{m} p(y_i|\theta_i) = \prod_{i=1}^{m} \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i}$$

Prior:

$\theta_i \sim p(\theta|\alpha, \beta)$

$p(\theta_i|\alpha, \beta) = \mathrm{Beta}(\theta_i|\alpha, \beta)$



Fixed α, β

Fully Bayesian

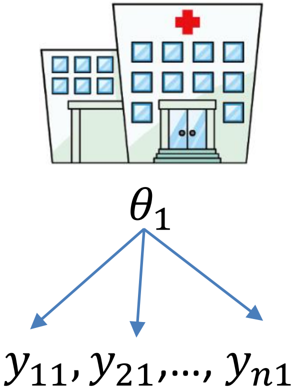
Hyper prior:
$(\alpha, \beta) \sim p(\alpha, \beta)$

$\theta = (\theta_1, \dots, \theta_m)$

$y = (y_1, \dots, y_m)$

Why Hierarchical models?

Survival probability of cardiac patients θ_1

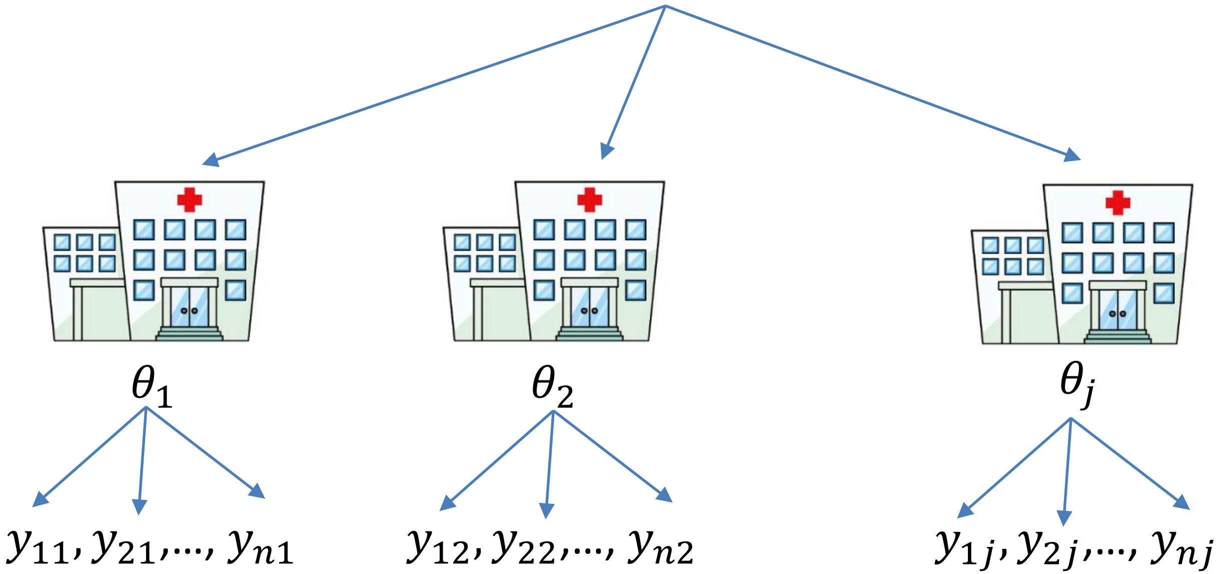


θ_1

$y_{11}, y_{21}, \dots, y_{n1}$

Why Hierarchical models?

Survival probability of cardiac patients $\theta_j \sim$ population distribution



θ_1 θ_2 θ_j

$y_{11}, y_{21}, \ldots, y_{n1}$ $y_{12}, y_{22}, \ldots, y_{n2}$ $y_{1j}, y_{2j}, \ldots, y_{nj}$

- Population distribution is used to structure some dependence into the parameters, thereby avoiding problems of overfitting
- It is natural to model such a problem hierarchically, with observable outcomes modeled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters, known as hyper-parameters
- Such hierarchical thinking helps in understanding multi-parameter problems and also plays an important role in developing computational strategies

Why Hierarchical models?

Nonhierarchical models

- With too small parameters, a model cannot fit large data set
  →Large Bias
- With too many parameters, a model over fit data set
  →Poor generalization

Hierarchical models:

- Can have enough parameters to fit the data well
- Uses population distribution to structure some dependencies into the parameters
  - ✓ Prior knowledge can be encoded into hierarchical structure
  - ✓ Advantageous when only a small data set is available
- Avoid problems of overfitting

Motivating example : New drug test

- Current experimental result

4 success from 14 tests

- What is the probability of success?

$$\frac{4}{14} = 28.6\%$$

- It seems that we only have very small data set

Motivating example : New drug test

- Current experimental result

4 success from 14 tests

- Historical experimental results

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	2/10	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	15/47	15/46	9/24

✓ The observed sample mean of 70 values $\frac{y_j}{n_j}$ = 0.136

✓ The observed sample standard deviation : 0.103

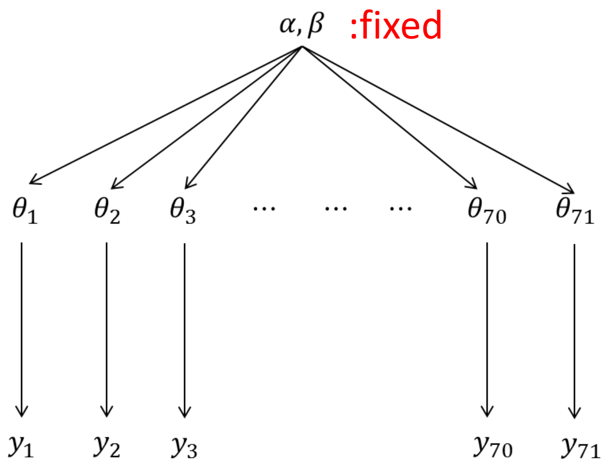
Motivating example : Drug test

- Historical experimental results

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	2/10	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	15/47	15/46	9/24

✓ The observed sample mean of 70 values $\frac{y_j}{n_j}$ = 0.136

✓ The observed sample standard deviation : 0.103



- Assume a success rate θ for each experiment follows Beta distribution

$$\theta \sim \mathrm{Beta}(\alpha, \beta)$$

$$\mathrm{E}(\theta) = \frac{\alpha}{\alpha + \beta}$$

$$\mathrm{var}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Estimated parameters
$(\alpha, \beta) = (1.4, 8.6)$

- Now, we have prior distribution that is empirically estimated from data

$$p(\theta) = \mathrm{Beta}(\theta | 1,4,8.6)$$

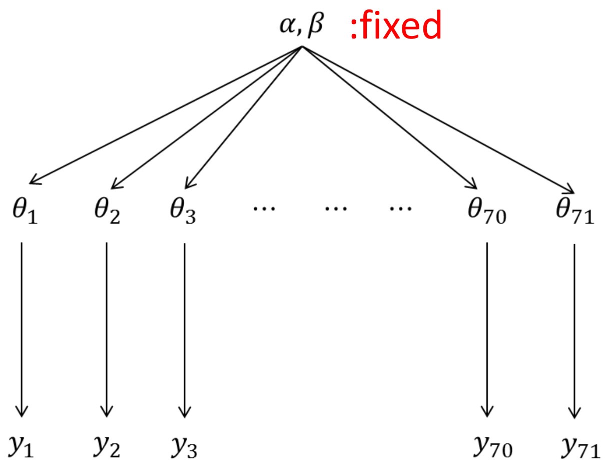
Motivating example : Drug test

- Historical experimental results

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	2/10	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	15/47	15/46	9/24

✓ The observed sample mean of 70 values $\frac{y_j}{n_j}$ = 0.136

✓ The observed sample standard deviation : 0.103



- Now, we have prior distribution that is empirically estimated from data

$$p(\theta) = \mathrm{Beta}(\theta|1.4{,}8.6)$$

- Likelihood of the current observation (4 success from 14 tests)

$$p(y|\theta) = \mathrm{Bin}(4|14, \theta)$$

- Posterior distribution of the current experiment results

$$p(\theta|y) = \mathrm{Beta}(5.4, 18.6)$$

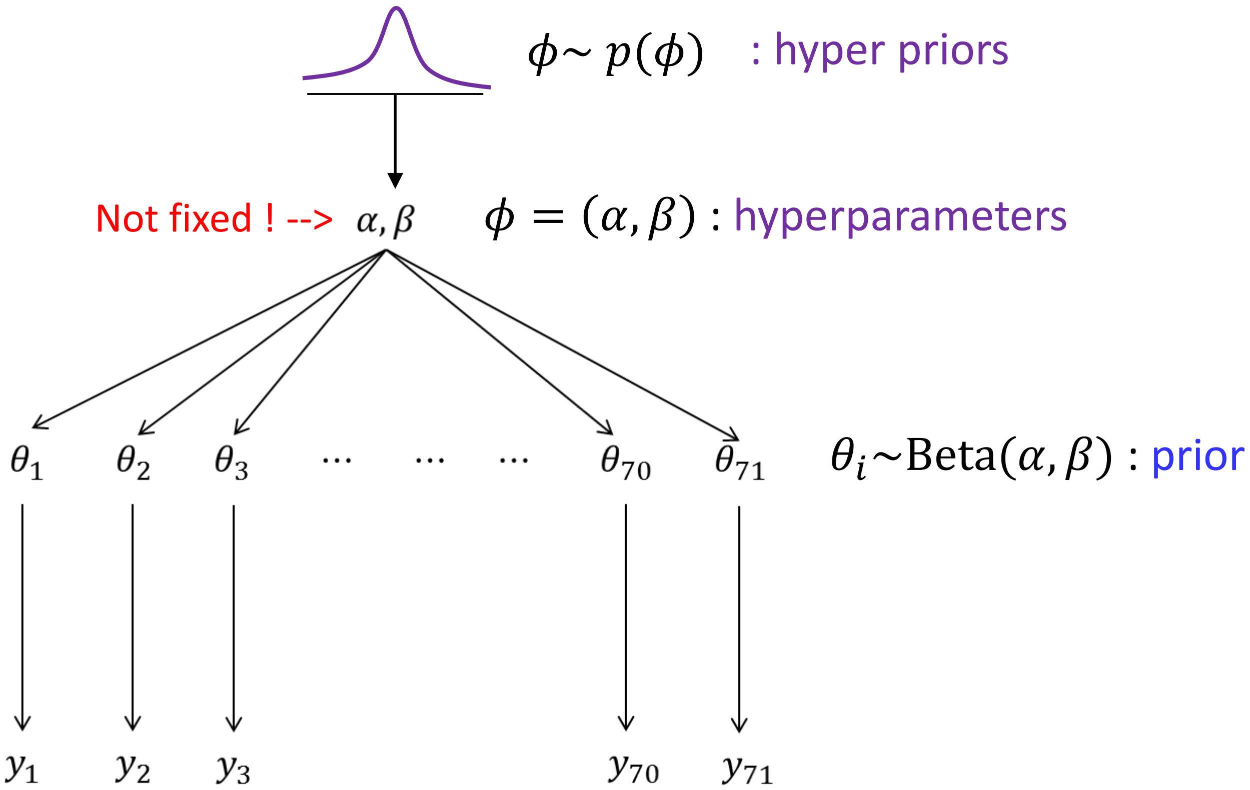
$$\mathrm{E}(\theta|y) = 0.223 < \frac{y_{71}}{n_{71}} = \frac{4}{14} = 0.286$$

$$\mathrm{var}(\theta|y) = 0.083$$

- This is an empirical Bayes analysis

→ The point estimation on α, β is arbitrary, and the point estimates ignore some uncertainties

The full Bayesian treatment of the hierarchical model



$\phi \sim p(\phi)$: hyper priors

Not fixed ! --> α, β $\phi = (\alpha, \beta)$: hyperparameters

θ_1 θ_2 θ_3 $\cdots$ $\cdots$ $\cdots$ θ_{70} θ_{71} $\theta_i \sim \text{Beta}(\alpha, \beta)$: prior

y_1 y_2 y_3 y_{70} y_{71}

The key characteristics of hierarchical Bayesian model is that ϕ is not known and thus has its own prior distribution $p(\phi)$

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

- This model include the uncertainty in hyperparameters ϕ
- The hyper parameter ϕ affects y only through parameters

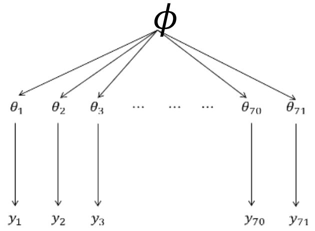
Exchangeability and hierarchical models

- In order to create a joint probability model for all the parameters $\theta = \left(\theta_1, \dots, \theta_J\right)$, we use the crucial idea of exchangeability

- The parameters $\left(\theta_1, \dots, \theta_J\right)$ are *exchangeable* in their joint distribution if $p\left(\theta_1, \dots, \theta_J\right)$ is invariant to permutations of the indexes $(1, \dots, J)$

- The simplest form of an exchangeable distribution has each of the parameters θ_j as an independent sample from a prior (or population) distribution governed by some unknown parameter vector ϕ; thus,

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



- In general, ϕ is unknown, so the distribution for θ must average over the uncertainty in ϕ :

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

- *De Finetti's theorem* said any suitably well-behaved exchangeable distribution on $\left(\theta_1, \dots, \theta_J\right)$ can be expressed as a mixture of independent and identical distributions
- Statistically, the mixture model characterizes parameters θ as drawn from a common 'superpopulation' that is determined by the unknown hyperparameters, ϕ

Exchangeability and hierarchical models

Assume $(y_1, y_2, \dots)$ are infinitely exchangeable, then by de Finetti's theorem for the $(y_1, \dots, y_n)$ that you actually observed, there must exist

- A parameter θ
- A distribution $p(y|\theta)$ such that $y_j \sim p(y|\theta)$ (independent draws)
- A distribution $p(\theta)$

Assume $(\theta_1, \theta_2, \dots)$ are infinitely exchangeable, then by de Finetti's theorem for the $(\theta_1, \dots, \theta_n)$ that you actually observed, there exists

- A parameter ϕ
- distribution $p(\theta|\phi)$ such that $\theta_j \sim p(\theta|\phi)$ (independent draws)
- A distribution $p(\phi)$

Although hierarchical models are typically written using the conditional independence notation, the assumptions underlying the model are exchangeability and functional forms for the priors

Inferencing Bayesian Models

- Bayes' Theorem is usually expressed very simply in the unscaled form:
  - *posterior* proportional to *prior* times *likelihood*:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{K} \approx p(y|\theta)p(\theta) = g(\theta|y)$$

(unscaled Posterior distribution)

- $K = \iiint p(y|\theta_1, \theta_2, \dots, \theta_p)p(\theta_1, \theta_2, \dots, \theta_p)d\theta_1 \dots d\theta_p$
- The closed form for high-dimensional integral exists only for particular cases, (e.g., likelihood is in exponential family and prior is conjugate prior to likelihood)

- The difficulty in computing K has restricted the easy implementation of Bayesian statistics

- Without knowing the value for K, we only know the shape of $p(\theta|D)$:
  - Can: find a mode, compute relative values at any two location
  - Can't: compute a probability (density), compute moments, conduct inferences

To conduct meaningful inferences, we need to compute or approximate the posterior $p(\theta|D)$

How to approximate $p(\theta\,|\,\mathrm{y})$?

1. Analytical computation

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)\,d\theta}$$

2. Numerical Approximation (integration):

$$\int p(y|\theta)p(\theta)\,d\theta \approx \frac{1}{S}\sum_{s=1}^{S} w_s\, p(y|\theta^s)p(\theta^s)$$

- with the weight w_s corresponding to the volume of space represented by the point θ^s

3. Direct Sampling (from unscaled posterior $g(\theta|y) = p(y|\theta)p(\theta)$)

- Rejection sampling
- Importance sampling

4. Markov Chain Monte Carlo (MCMC) sampling (from unscaled posterior $g(\theta|y) = p(y|\theta)p(\theta)$)

- Metropolis Sampling
- Metropolis Hasting sampling
- Gibbs sampling→ good for hierarchical model

5. Distributional Approximation

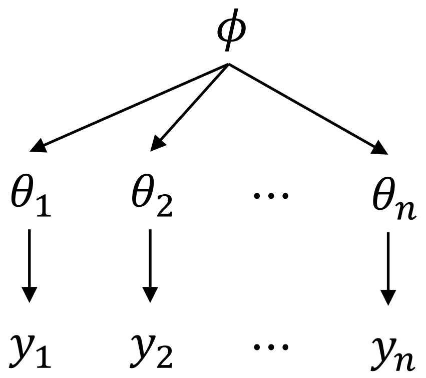
- Laplace approximation
- Variational inference

Gibbs Sampling Algorithm

Algorithm

```
    Initialize $\boldsymbol{\theta}^{(0)}$
    **for** iteration $i = 1, \ldots do$
            $\boldsymbol{\theta}_1^{(i)} \sim P\left(\boldsymbol{\theta}_1 \middle| \boldsymbol{\theta}_2^{(i-1)}, \boldsymbol{\theta}_3^{(i-1)}, \ldots, \boldsymbol{\theta}_J^{(i-1)}\right)$
            $\boldsymbol{\theta}_2^{(i)} \sim P\left(\boldsymbol{\theta}_2 \middle| \boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_3^{(i-1)}, \ldots, \boldsymbol{\theta}_J^{(i-1)}\right)$
            $\boldsymbol{\theta}_3^{(i)} \sim P\left(\boldsymbol{\theta}_3 \middle| \boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}, \ldots, \boldsymbol{\theta}_J^{(i-1)}\right)$
            $\vdots$
            $\boldsymbol{\theta}_J^{(i)} \sim P\left(\boldsymbol{\theta}_J \middle| \boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}, \ldots, \boldsymbol{\theta}_{J-1}^{(i)}\right)$
    **end for**
```

Procedure of inferencing for Bayesian hierarchical model



$\phi \sim p(\phi)$
$\theta_i \sim p(\theta|\phi)$
$y_i \sim p(y|\theta_i)$

- only ϕ has a prior that is set
- $\theta = (\theta_i, \dots, \theta_n)$ and ϕ are parameters
- y_i is observed

- The joint posterior distribution of interest in hierarchical models is

$$p(\phi, \theta|y) \propto p(y|\phi, \theta)p(\phi, \theta) = p(y|\theta)p(\theta|\phi)p(\phi)$$

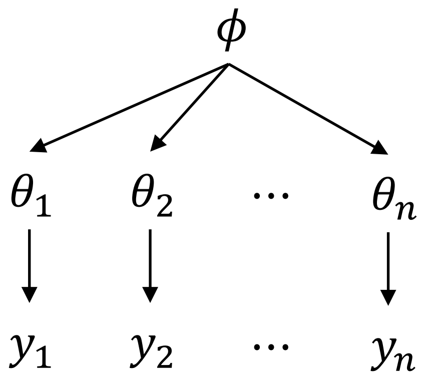
- We may be focused on *marginal posteriors*

$$p(\theta|y) = \int_{\phi} p(\phi, \theta|y)d\phi \qquad \text{or} \qquad p(\phi|y) = \int_{\theta} p(\phi, \theta|y)d\theta$$

This integral is hard to compute due to multiple parameters involved in θ

- Two posterior predictive distributions are of interest
  - ✓ The distribution of future observations $\tilde{y}$ corresponding to an existing θ_j
  - ✓ The distribution of future observations $\tilde{y}$ corresponding to future θ_j *drawn from hyper-prior*

Procedure of inferencing for Bayesian hierarchical model



$\phi \sim p(\phi)$
$\theta_i \sim p(\theta|\phi)$
$y_i \sim p(y|\theta_i)$

- only ϕ has a prior that is set
- $\theta = (\theta_i, \ldots, \theta_n)$ and ϕ are parameters
- y_i is observed

We present an approach that combines **analytical** and **sampling methods** to obtain simulations from the joint posterior distribution, $p(\phi, \theta|y)$, in the case that the population distribution $p(\theta|\phi)$ is conjugate to the likelihood $p(y|\theta)$

Analytical method : to compute an unscaled form of posterior

Sampling method : to draw sample to compute the approximate posterior

Procedure of inferencing for Bayesian hierarchical model

Analytical

Step 1: Write the *joint posterior density* $p(\phi, \theta|y)$

$$p(\phi, \theta|y) \propto p(y|\theta)p(\theta|\phi)p(\phi) \quad \text{(un-normalized form)}$$

Step 2: Determine analytically *the conditional posterior density* $p(\theta|\phi, y)$

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

- ✓ $p(\theta|\phi, y)$ is a distribution on θ given ϕ and the fixed data y
- ✓ when ϕ is fixed→ single level Bayesian approach can be used, thus easy for conjugate model
- ✓ Conditional posterior distribution is a product of conjugate posterior densities for the components θ_j

Step 3: Obtain *marginal posterior distribution* $p(\phi|y)$ and estimate ϕ

$$p(\phi|y) = \int_\theta p(\phi, \theta|y)d\theta \qquad \text{or} \qquad p(\phi|y) = \frac{p(\phi, \theta|y)}{p(\theta|\phi, y)}$$

Brute force approach

Drawing simulations from the posterior distribution

Sampling

By factorization: $p(\phi, \theta|y) = p(\theta|\phi, y)p(\phi|y)$ (Not Bayesian factorization)

Step 1: Draw the vector of hyperparameters ϕ from its marginal posterior distribution, $p(\phi|y)$

Step 2: Draw the parameter vector θ from its conditional posterior distribution

$$p(\theta|\phi, y) = \prod_{j=1}^{J} p(\theta_j \underbrace{|\phi, y)}_{\text{(fixed)}}$$

(The components θ_j can be drawn independently, one at a time)

Step 3: Draw predictive values $\hat{y}$ from the posterior predictive distribution given the drawn θ

$$\hat{y} \sim p(\hat{y}|\theta)$$

or draw future observations $\tilde{y}$ corresponding to future θ_j drawn from hyper-prior ϕ

Repeat L times, and compute posterior distribution of any estimand or predictive quantity of interest

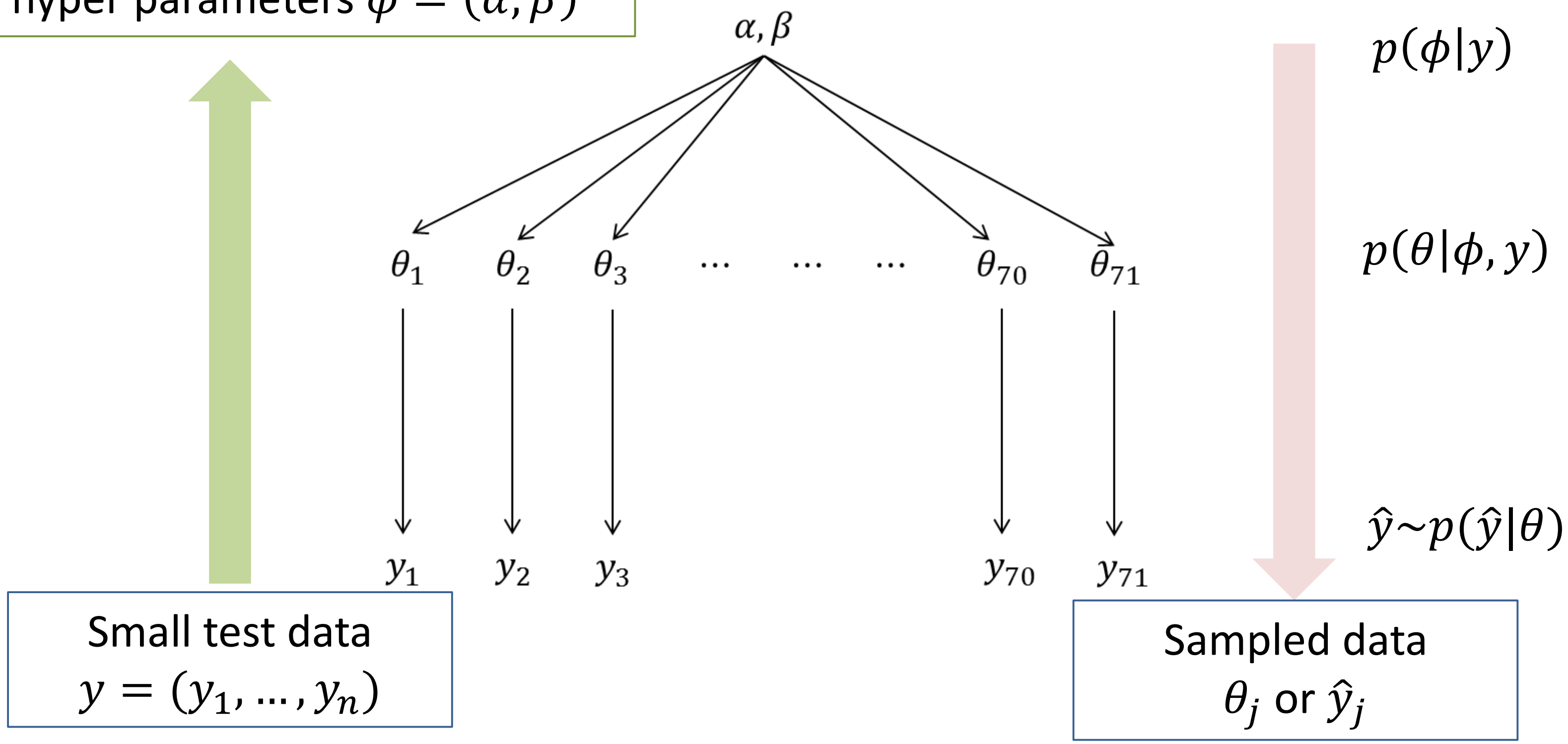
Bayesian analysis of conjugate hierarchical models-Procedure : Rat tumor example

- $\phi \sim p(\phi)$: Hyper prior
- $\theta_i \sim \mathrm{Beta}(\alpha, \beta)$: Prior
- $y_i \sim \mathrm{Bin}\left(n_j, \theta_j\right)$: Likelihood

Analytical approach

$p(\phi|y)$

Construct posterior on
the hyper parameters $\phi = (\alpha, \beta)$



Small test data
$y = (y_1, \ldots, y_n)$

Sampled data
θ_j or $\hat{y}_j$

Simulational approach

Bayesian analysis of conjugate hierarchical models-Procedure : Rat tumor example

Analytical

- Models are given:

$$(\alpha,\beta)\sim p(\alpha,\beta) \qquad \text{: hyper prior}$$

$$\theta_j\sim \mathrm{Beta}(\alpha,\beta) \qquad \text{: Prior}$$

$$y_j\sim \mathrm{Bin}(n_j,\theta_j) \qquad \text{: sampling distribution}$$

- **Step 1** (joint posterior distribution):

$$\begin{aligned} p(\theta,\alpha,\beta|y) &\propto p(\alpha,\beta)p(\theta|\alpha,\beta)\,p(y|\theta,\alpha,\beta) \\ &\propto p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\prod_{j=1}^{J}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j} \end{aligned}$$

- **Step 2** (conditional posterior distribution):

$$p(\theta|\alpha,\beta,y)=\prod_{j=1}^{J}p(\theta_j|\alpha,\beta,y_j)=\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}\theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$$

(Given the hyper-parameters (α,β), it is just a single layer Bayesian posterior)

- **Step 3** (marginal posterior distribution):

$$p(\alpha,\beta|y)=\frac{p(\theta,\alpha,\beta|y)}{p(\theta|\alpha,\beta,y)}\propto p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha)}\frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)} \qquad \text{(Unscaled form)}$$

Bayesian analysis of conjugate hierarchical models-Procedure : Rat tumor example

Sampling Approach

- Models are given:

$(\alpha, \beta) \sim p(\alpha, \beta)$: hyper prior

$\theta_j \sim \mathrm{Beta}(\alpha, \beta)$: Prior

$y_j \sim \mathrm{Bin}(n_j, \theta_j)$: sampling distribution

Compute the posterior distribution using sampling methods

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

Repeat L times $(i = 1, \dots, L)$

Sample $\left(\alpha^{(i)}, \beta^{(i)}\right)$ from

$$p(\alpha, \beta | y) = \frac{p(\theta, \alpha, \beta | y)}{p(\theta | \alpha, \beta, y)} \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\alpha)} \frac{\Gamma(\alpha + n_j)\Gamma(\beta + n_j - y_j)}{\Gamma(\alpha, \beta + n_j)}$$

Sample $\theta^{(i)}$ from

$$p\left(\theta | \alpha^{(i)}, \beta^{(i)}, y\right) = \prod_{j=1}^{J} p\left(\theta_j | \alpha^{(i)}, \beta^{(i)}, y_j\right) = \prod_{j=1}^{J} \frac{\Gamma\left(\alpha^{(i)} + \beta^{(i)} + n_j\right)}{\Gamma\left(\alpha^{(i)} + y_j\right)\Gamma\left(\beta^{(i)} + n_j - y_j\right)} \theta_j^{\alpha^{(i)} + y_j - 1} \left(1 - \theta_j\right)^{\beta^{(i)} + n_j - y_j - 1}$$

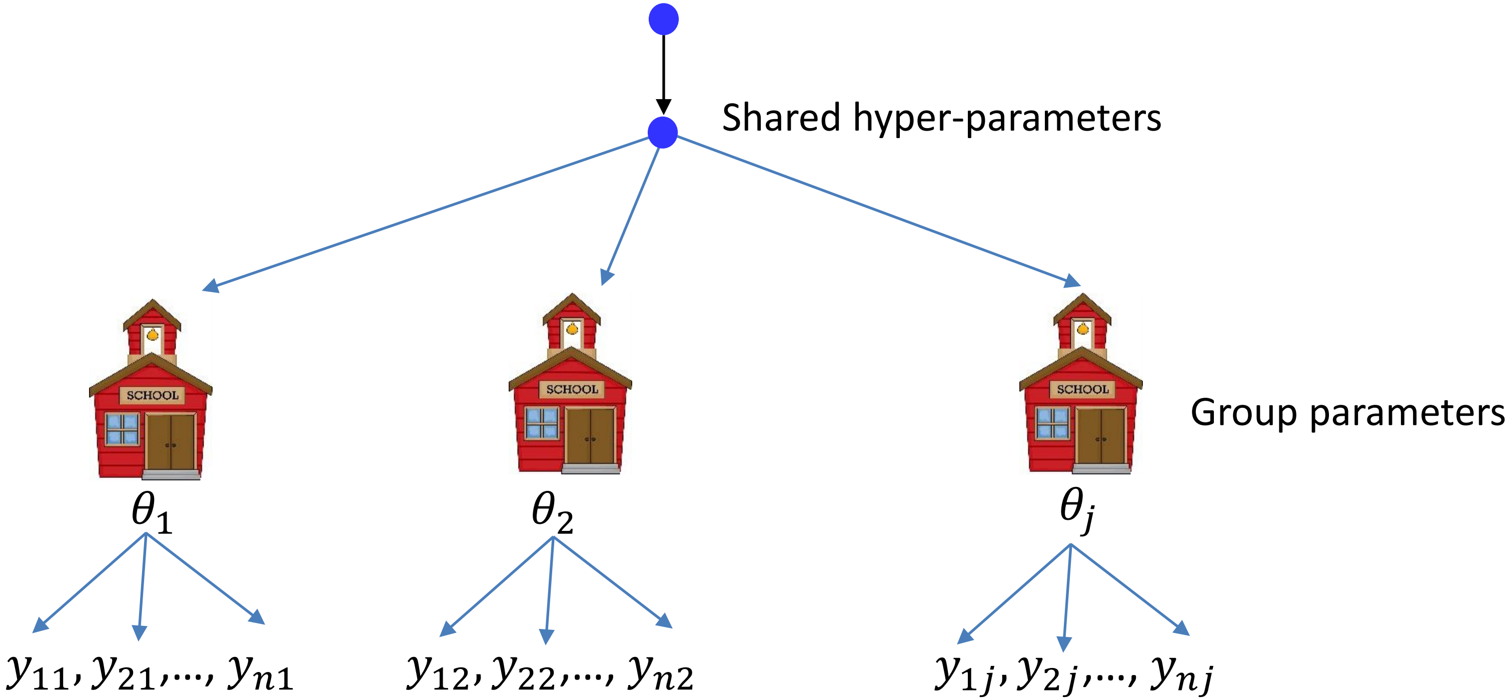
$$\theta^{(i)} = \left(\theta_1^{(i)}, \dots, \theta_J^{(i)}\right)$$

We can then have L samples : $\left(\alpha^{(1)}, \beta^{(1)}, \theta^{(1)}\right), \dots, \left(\alpha^{(L)}, \beta^{(L)}, \theta^{(L)}\right)$

Bayesian analysis of conjugate hierarchical models-Procedure : Rat tumor example

Jupyter Demo Simulation

Normal model with exchangeable parameters



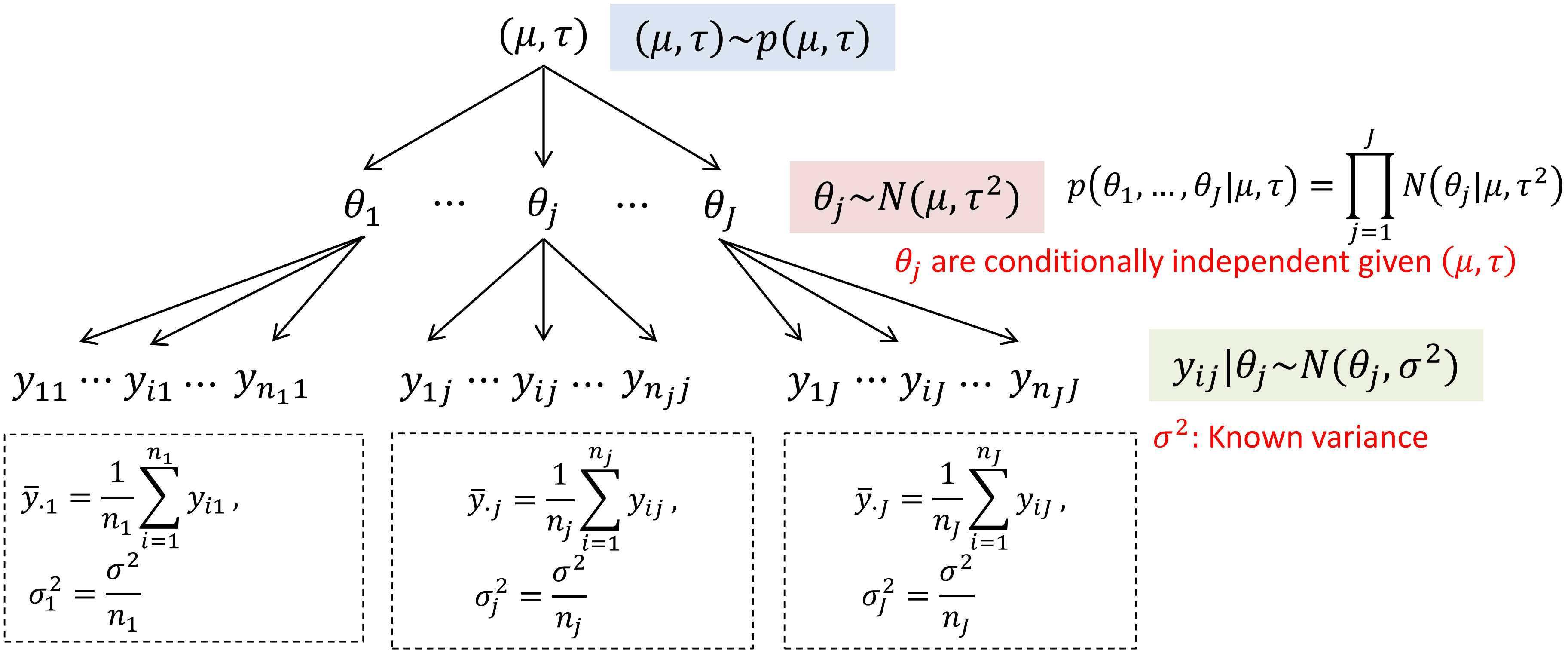
Shared hyper-parameters

Group parameters

θ_1 θ_2 θ_j

$y_{11}, y_{21}, \ldots, y_{n1}$ $y_{12}, y_{22}, \ldots, y_{n2}$ $y_{1j}, y_{2j}, \ldots, y_{nj}$

Normal model with exchangeable parameters



(μ,τ)

$(\mu,\tau) \sim p(\mu,\tau)$

$\theta_1 \quad \cdots \quad \theta_j \quad \cdots \quad \theta_J$

$\theta_j \sim N(\mu,\tau^2)$

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

θ_j are conditionally independent given (μ,τ)

$y_{11} \cdots y_{i1} \cdots y_{n_1 1} \qquad y_{1j} \cdots y_{ij} \cdots y_{n_j j} \qquad y_{1J} \cdots y_{iJ} \cdots y_{n_J J}$

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

σ^2: Known variance

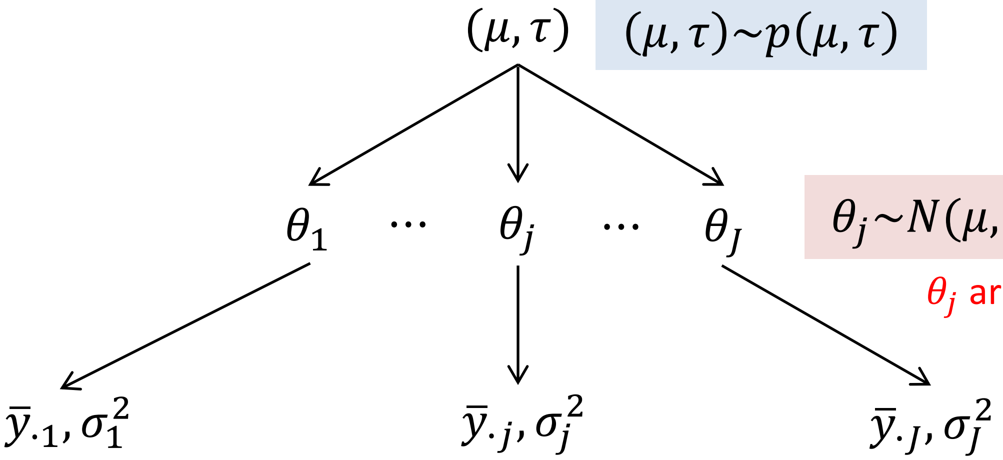
$$\bar{y}_{\cdot 1} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{i1}, \qquad \sigma_1^2 = \frac{\sigma^2}{n_1}$$

$$\bar{y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}, \qquad \sigma_j^2 = \frac{\sigma^2}{n_j}$$

$$\bar{y}_{\cdot J} = \frac{1}{n_J}\sum_{i=1}^{n_J} y_{iJ}, \qquad \sigma_J^2 = \frac{\sigma^2}{n_J}$$

- Observed data are normally distributed with a different mean θ_j for each ‘group’ with known observation variance σ^2
- Normal population distribution for the group mean $\theta_j \sim N(\mu,\tau^2)$

Normal model with exchangeable parameters



(μ, τ)

$\theta_1 \quad \cdots \quad \theta_j \quad \cdots \quad \theta_J$

$\bar{y}_{\cdot 1}, \sigma_1^2 \qquad \bar{y}_{\cdot j}, \sigma_j^2 \qquad \bar{y}_{\cdot J}, \sigma_J^2$

$(\mu, \tau) \sim p(\mu, \tau)$

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

θ_j are conditionally independent given (μ, τ)

$\bar{y}_{\cdot j} | \theta_j \sim N(\theta_j, \sigma_j^2)$ with $\sigma_j{}^2 = \frac{\sigma^2}{n_j}$

Distribution on sufficient statistics

- Observed data are normally distributed with a different mean for each ‘group’ with known observation variance
- Normal population distribution for the group mean
- Write the likelihood for each θ_j using the sufficient statistics, $\bar{y}_{\cdot j}$

$$y_{ij} | \theta_j \sim N(\theta_j, \sigma^2) \rightarrow p(y_{.j} | \theta) = \prod_{i=1}^{n_j} N(y_{ij} | \theta_j, \sigma_j^2) = \prod_{i=1}^{n_j} \frac{1}{(2\pi\sigma_j^2)^{1/2}} \exp\left(-\frac{(y_{ij} - \theta_j)^2}{2\sigma_j^2}\right) = \frac{1}{(2\pi\sigma_j^2)^{n_j/2}} \exp\left(\sum_{i=1}^{n_j} -\frac{(y_{ij} - \theta_j)^2}{2\sigma_j^2}\right)$$

$$y_{.j} = (y_{1j}, \dots, y_{n_j, j})$$

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

Normal model with exchangeable parameters

- Models are given:
  - $(\mu, \tau) \sim p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$: hyper prior
  - $\theta_j \sim N(\mu, \tau^2) \rightarrow p(\theta|\mu, \tau) = \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2)$: Prior

    (parameters are conditionally independent given hyper parameters
  - $y_{.j} \sim N(\theta_j, \sigma_j^2) \rightarrow p(y|\theta) = \prod_{j=1}^{J} N(\bar{y}_{.j}|\theta_j, \sigma_j^2)$: sampling distribution

    (data are conditionally independent given parameter)

Analytical

- $p(\theta, \mu, \tau|y)$: The joint posterior distribution
- $p(\theta|\mu, \tau, y)$: The conditional posterior distribution
- $p(\mu, \tau|y) \rightarrow p(\mu\ |\tau, y)p(\tau|y)$: The marginal posterior distribution

Sampling

- To derive the posterior computational methods, we factorize the posterior as :

$$\text{(1)} \longrightarrow \text{(2)} \longrightarrow \text{(3)}$$

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

Procedure of inferencing for Bayesian hierarchical model

Analytical

Step 1: Write the ***joint posterior density*** $p(\mu, \tau, \theta | y)$

$$p(\mu, \tau, \theta | y) \propto p(\mu, \tau)\, p(\theta | \mu, \tau) p(y | \theta)$$ (un-normalized form)

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

$\left(\sigma_j^2 = \frac{\sigma^2}{n_j} \text{ is known}\right)$

Step 2: Determine analytically *the* ***conditional posterior density*** $p(\theta | \mu, \tau, y)$

$$p(\theta | \mu, \tau, y) = \prod_{j=1}^{J} p(\theta_j | \mu, \tau, y) = \prod_{j=1}^{J} N(\theta_j | \hat{\theta}_j, V_j)$$

Recall posterior of Gaussian Prior + Gaussian Likelihood

$$\boxed{\frac{1}{\sigma_j^2} = \frac{n}{\sigma^2}}$$

$\theta_j \sim N(\mu, \tau^2)$ + $y_{ij} \sim N(\theta_j, \sigma^2)$ = $\theta_j | \mu, \tau, y \sim N(\hat{\theta}_j, V_j)$

Prior **sampling distribution** **posterior**

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2} \bar{y}_{.j} + \frac{1}{\tau^2} \mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \qquad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

Given hyper parameters, it requires one layer posterior distribution

Procedure of inferencing for Bayesian hierarchical model

Analytical

Step 3: Obtain ***marginal posterior distribution*** $p(\mu,\tau|y)$

$$p(\mu,\tau|y) = \int_\theta p(\mu,\tau,\theta|y)d\theta \qquad \text{or} \qquad p(\mu,\tau|y) = \frac{p(\mu,\tau,\theta|y)}{p(\theta|\mu,\tau,y)}$$

For the hierarchical normal model, we can simply consider the information supplied by the data about the hyper parameters directly

$$\begin{aligned} p(\mu,\tau|y) &\propto p(\mu,\tau)p(y|\mu,\tau) \\ &= p(\mu,\tau)\prod_j^J N(\bar{y}_i|\mu,\sigma_j^2+\tau^2) \end{aligned} \qquad \because p(\bar{y}_i|\mu,\tau) \sim N(\bar{y}_i|\mu,\sigma_j^2+\tau^2)$$

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

Step 3-1: posterior distribution of μ given τ

$$p(\mu|\tau,y) = N(\mu|\hat{\mu},\, V_\mu) \qquad \hat{\mu} = \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2+\tau^2}\bar{y}_{\cdot j}}{\sum_{j=1}^J \frac{1}{\sigma_j^2+\tau^2}} \quad \text{and} \quad V_\mu^{-1} = \sum_{j=1}^J \frac{1}{\sigma_j^2+\tau^2}$$

Step 3-2: posterior distribution of τ

$$\begin{aligned} p(\tau|y) = \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} &= \frac{p(\mu,\tau)\prod_j^J N(\bar{y}_i|\mu,\sigma_j^2+\tau^2)}{N(\mu|\hat{\mu},\, V_\mu)} \\ &\propto p(\tau)V_\mu^{\frac{1}{2}}\prod_{j=1}^J(\sigma_j^2+\tau^2)^{-1/2}\exp\left(-\frac{(\bar{y}_{\cdot j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}\right) \end{aligned}$$

Assuming:
$p(u,\tau) = p(\tau)p(\mu|\tau) = p(\tau)$
with $p(u|\tau) = 1$

Computation of Posterior

Sampling

Different factorization for simulation:

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

$$p(\theta,\mu,\tau|y) = \underset{(3)}{p(\theta|\mu,\tau,y)}\;\underset{(2)}{p(\mu|\tau,y)}\;\underset{(1)}{p(\tau|y)}$$

$(3) \longleftarrow (2) \longleftarrow (1)$

Step 1

$$p(\tau|y) = \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} \propto \frac{p(\tau)\prod_{j=1}^{J} N\left(\mu|\theta_j,\sigma_j^2+\tau^2\right)}{N\left(\widehat{\mu},V_\mu\right)} \propto p(\tau)V_\mu^{\frac{1}{2}}\prod_{j=1}^{J}\left(\sigma_j^2+\tau^2\right)^{-1/2}\exp\left(-\frac{\left(\bar{y}_{.j}-\hat{\mu}\right)^2}{2\left(\sigma_j^2+\tau^2\right)}\right)$$

Step 2

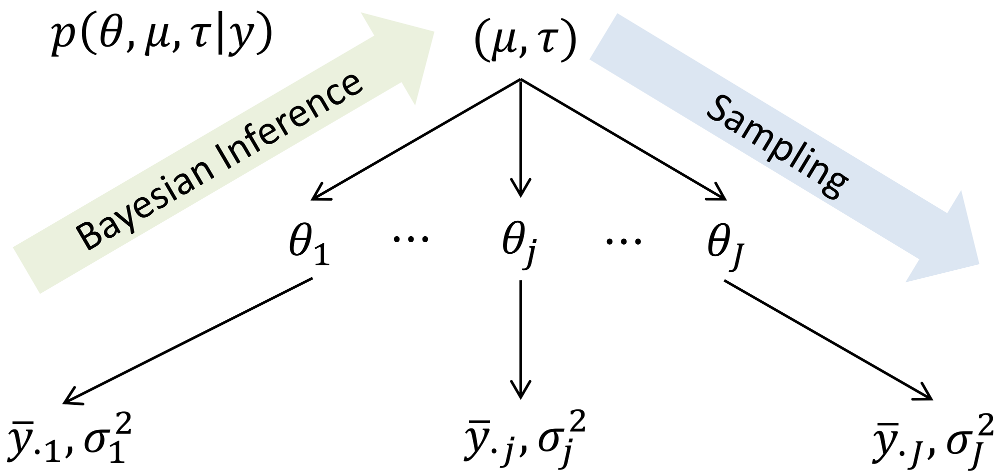
$$p(\mu|\tau,y) = \mathrm{N}\left(\mu|\widehat{\mu},V_\mu\right) \qquad \widehat{\mu} = \frac{\sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}\bar{y}_{.j}}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}} \quad \text{and} \quad V_\mu = \frac{1}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}}$$

Step 3

$$p\left(\theta_j|\mu,\tau,y\right) \sim N\left(\hat{\theta}_j,V_j\right) \qquad \hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_{.j}+\frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}$$

Posterior Predictive distribution

Sampling using given the posterior distribution of the parameters



$p(\tau|y)$

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

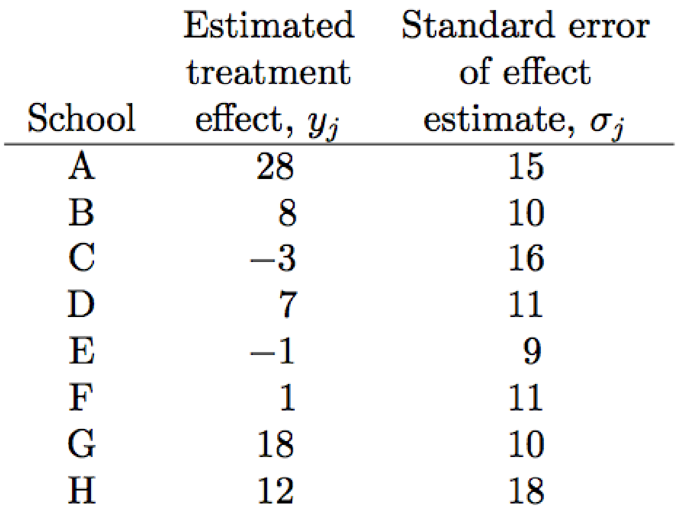
$p(\theta_j|\mu,\tau,y)$

$p(y_{ij}|\theta_j)$

- Future data $\tilde{y}$ from the current set of batches, with means $\theta = (\theta_1,\dots,\theta_J)$
  1. First draw $\theta = (\theta_1,\dots,\theta_J)$ from $p(\theta,\mu,\tau|y)$ and sample data $\tilde{y}$ using $p(y_{ij}|\theta_j)$

- Future data $\tilde{y}$ from $\tilde{J}$ future batches, with means $\tilde{\theta} = (\tilde{\theta}_1,\dots,\tilde{\theta}_{\tilde{J}})$
  1. First draw (μ,τ) from $p(\theta,\mu,\tau|y)$
  2. Second draw $\tilde{J}$ new parameters $\tilde{\theta} = (\tilde{\theta}_1,\dots,\tilde{\theta}_{\tilde{J}})$ from $p(\tilde{\theta}_j|\mu,\tau)$
  3. Third, draw $\tilde{y}$ given $\tilde{\theta}$ from the data distribution $p(y_{ij}|\tilde{\theta}_j)$

Example: Parallel experiments in eight schools



School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18

- A study was performed for the Educational Testing Service to analyze the effects of special coaching programs on test scores in eight schools.

- There was no prior reason to believe that any of the eight programs was more effective than any other or that some were more similar in effect to each other than to any other

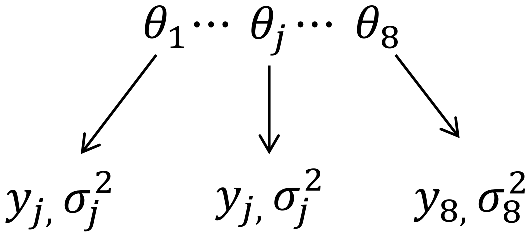
- The estimates y_j and σ_j are obtained by independent experiments

Example: Parallel experiments in eight schools

Independent estimate

- The estimates y_j are obtained by independent experiments and have approximately normal distribution
- it is difficult statistically to distinguish between any of the experiments.
  - ✓ 95% posterior intervals all overlap substantially
- Each estimation is based on a small size of subset.

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18

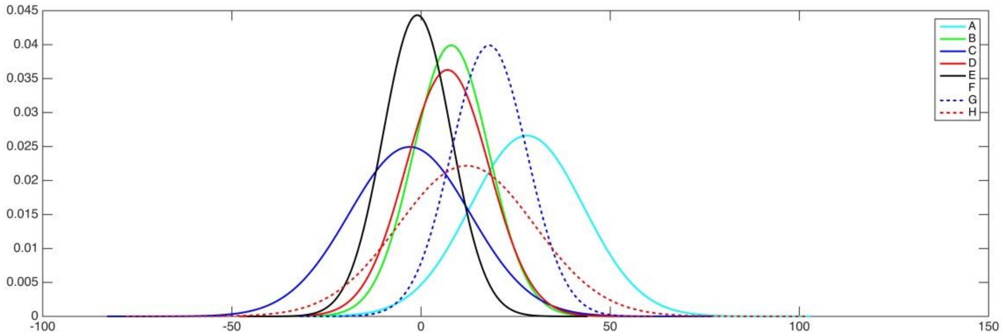


$\theta_1 \cdots \theta_j \cdots \theta_8$

y_j, σ_j^2 $\quad$ y_j, σ_j^2 $\quad$ y_8, σ_8^2

$y_j \sim N(\theta_j, \sigma_j^2)$

σ^2: Known variance



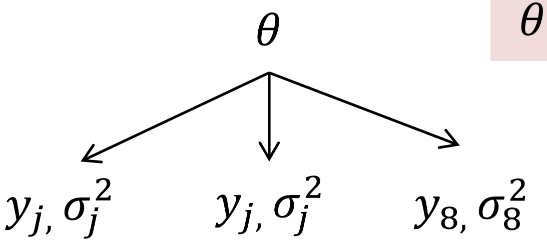
$p(A > 28|\theta) = 0.5?$

Example: Parallel experiments in eight schools

Pooled estimate

- All experiments have the same effect and produce independent estimates of this common effect
- It is difficult to account for the different effects from the 8 schools.

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18



$\theta \sim N(\mu, \tau^2)$

$y_j \sim N(\theta, \sigma_j^2)$

σ^2: Known variance

$$\mu = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2} y_j}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}} = 7.7$$

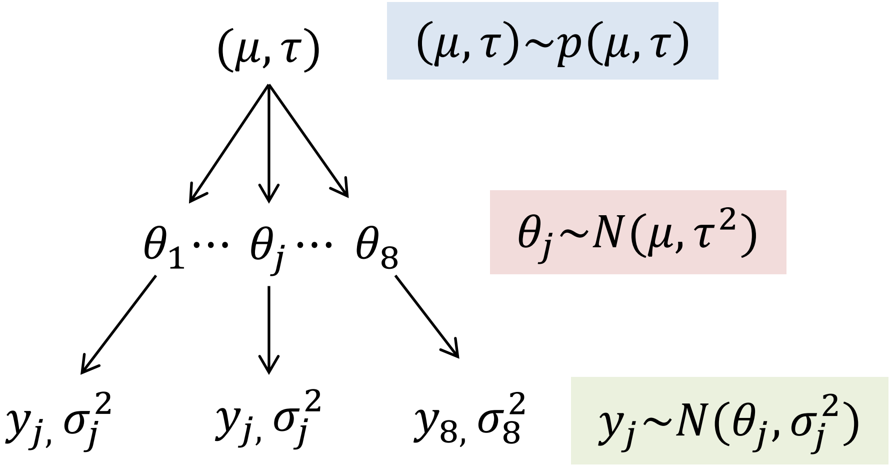
$$\tau^2 = \frac{1}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2}} = 16.6, \sigma = 4.1$$

$p(A < 7.7|\theta) = 0.5?$

Example: Parallel experiments in eight schools

Hierarchical Bayesian Model

We would like a compromise that combines information from all eight experiments without assuming all the θ_j's to be equal.



$(\mu, \tau) \sim p(\mu, \tau)$

$\theta_j \sim N(\mu, \tau^2)$

$y_j \sim N(\theta_j, \sigma_j^2)$

σ^2: Known variance

Example: Parallel experiments in eight schools

Sampling procedure

$$p(\tau|y) \propto p(\tau) V_{\mu}^{\frac{1}{2}} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(-\frac{(\bar{y}_{\cdot j} - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)}\right)$$

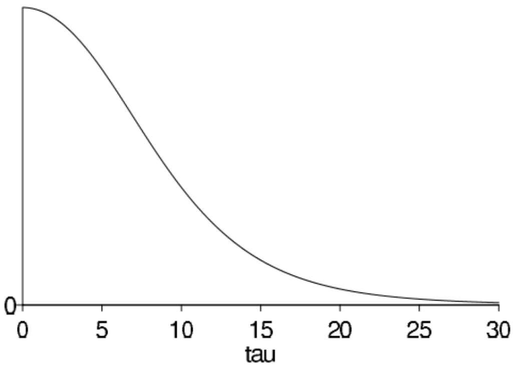
Sample τ given data y

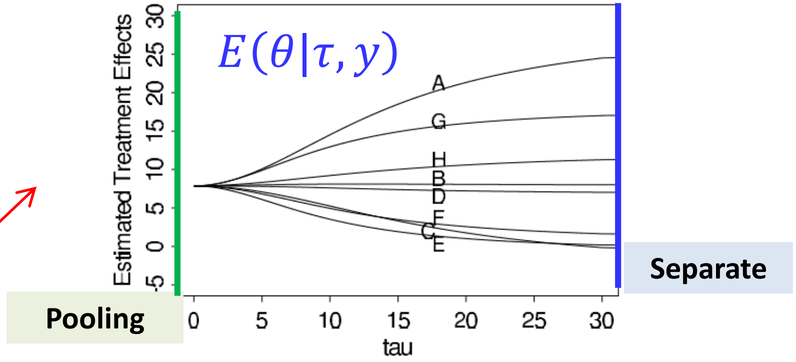
$$p(\theta|\tau, y) = \int_{\mu} p(\theta, \mu|\tau, y) d\mu$$

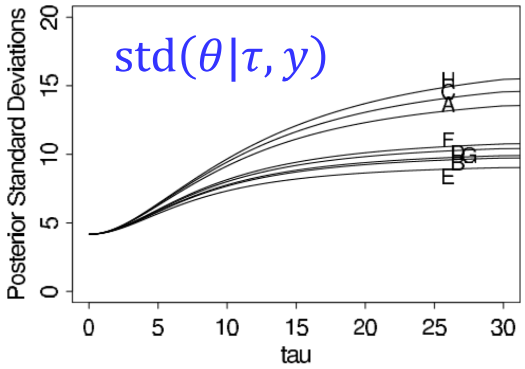
$$= \int_{\mu} p(\theta|\mu, \tau, y) p(\mu|\tau, y) d\mu$$

Sample θ given τ and data y

Marginal posterior density $p(\tau|y)$

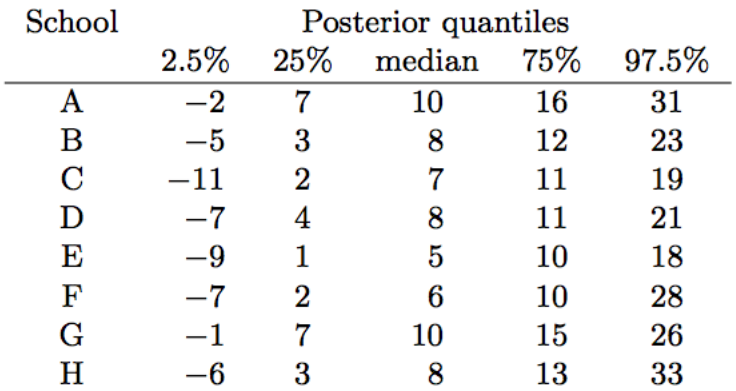




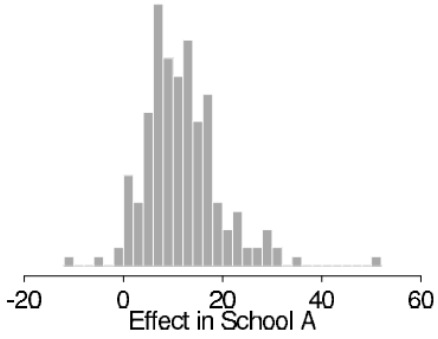


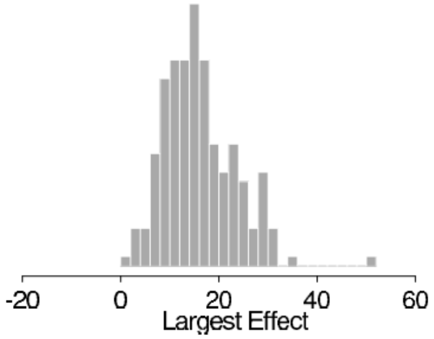
Example: Parallel experiments in eight schools

Results from the posterior distribution estimated using the sampled data



School	Posterior quantiles				
	2.5%	25%	median	75%	97.5%
A	−2	7	10	16	31
B	−5	3	8	12	23
C	−11	2	7	11	19
D	−7	4	8	11	21
E	−9	1	5	10	18
F	−7	2	6	10	28
G	−1	7	10	15	26
H	−6	3	8	13	33





- The Bayesian probability that the effect in school A is as large as 28 points:

  Individual test: 50% → Bayesian : less than 10%

- What is the maximum $\{\theta_j\}$?

- What is $\mathrm{Pr}(\theta_3 > \theta_5 | \mathrm{y})$?

Hierarchical model is flexible enough to adapt to the data, thereby providing posterior inferences that account for the partial pooling as well as uncertainty in the hyper parameters