Lecture 5

Hierarchical Models

Textbook Ch5

邓婉璐

wanludeng@tsinghua.edu.cn

Review: Basic Scheme

- Modeling
  - ✓ Sampling distribution
  - ✓ Prior

Y i.i.d. — Single parameter model, Multi-parameter model

Y 不同来源 — Hierarchical model

(X,Y)

- Inference (for posterior and based on posterior)
  - ✓ Analytic inference
  - ✓ Inference based on simulation
  - ✓ Auxiliary tool: Asymptotics
- Model checking / Comparison

Outline

- Introduction – Why we need a hierarchical model
- Definition – How we build a hierarchical model
- Inference – How we analyze a hierarchical model
  - Binomial model
  - Normal model
- Application – Illustration with two examples on real data
  - Rat tumor (binomial model)
  - ETS test scores (normal model)

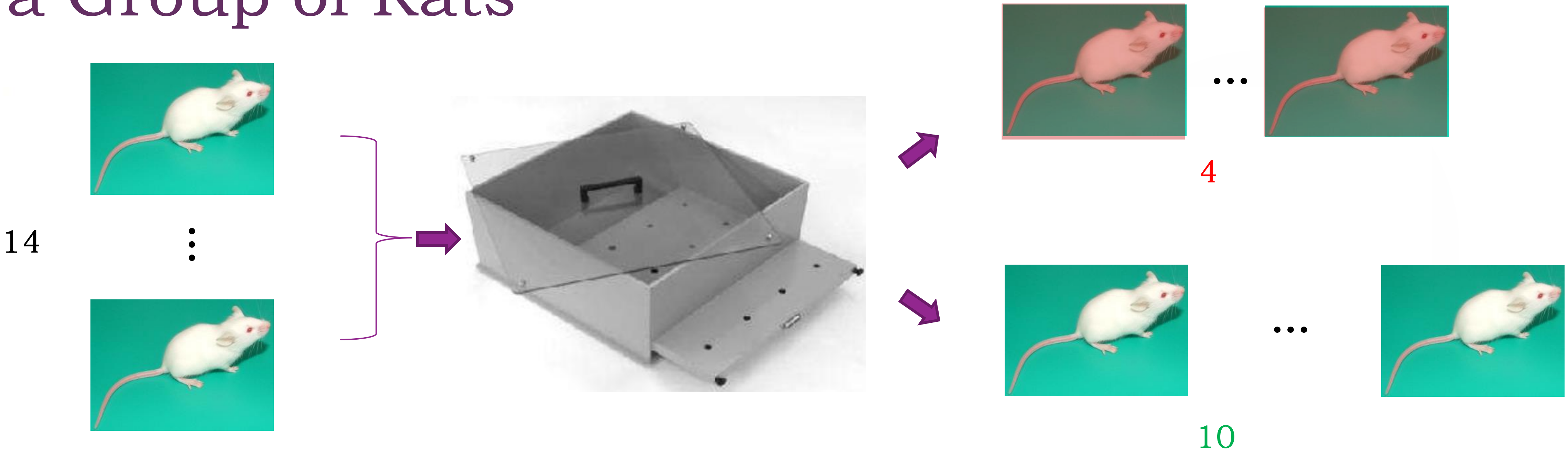
Objectives for Today

- 理解可交换性假设，了解并能够判断适用层次模型的情景
- 掌握构造层次模型的基本框架，明确选择超先验的准则，通过经典模型了解选择合理超先验的方式
- 给定超先验、先验、抽样模型下，掌握如何进行后验的推断和新观测的预测

How to find a suitable prior distribution in complicated case?

– Parameterization
– Motivating Examples

Example: Estimating the Risk of Tumor in a Group of Rats



Current experiment:
4/14

Results from a new experiment.
Wants to estimate the **risk of tumor**

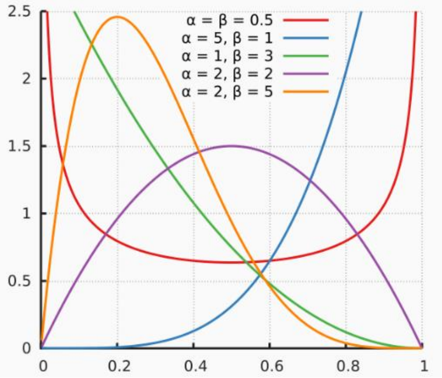
Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of* $\frac{y_j}{n_j}$: *(number of rats with tumors)/(total number of rats).*

Recall: Binomial Example

Likelihood: $p(y|\theta) \propto \theta^{a}(1-\theta)^{b}$

Prior: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \Longleftrightarrow \theta \sim \text{Beta}(\alpha, \beta)$

- Hyper-parameters
- Control the shape of prior



Posterior: $p(\theta|y) \propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$$
$$= \text{Beta}(\theta|\alpha+y, \beta+n-y)$$

Posterior mean: $E(\theta|y) = \dfrac{\alpha+y}{\alpha+\beta+n}$

Posterior variance: $var(\theta|y) = \dfrac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = \dfrac{E(\theta|y)[1-E(\theta|y)]}{\alpha+\beta+n+1}$

Limiting behavior of posterior: $\left(\dfrac{\theta-\text{E}(\theta|y)}{\sqrt{\text{var}(\theta|y)}}\middle| y\right) \sim N(0,1)$

Example: Estimating the Risk of Tumor in a Group of Rats

70 historical experiments

Previous experiments:

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	15/47	15/46	9/24

Current experiment: 4/14

Posterior distribution with fixed prior $\text{Beta}(\alpha, \beta)$

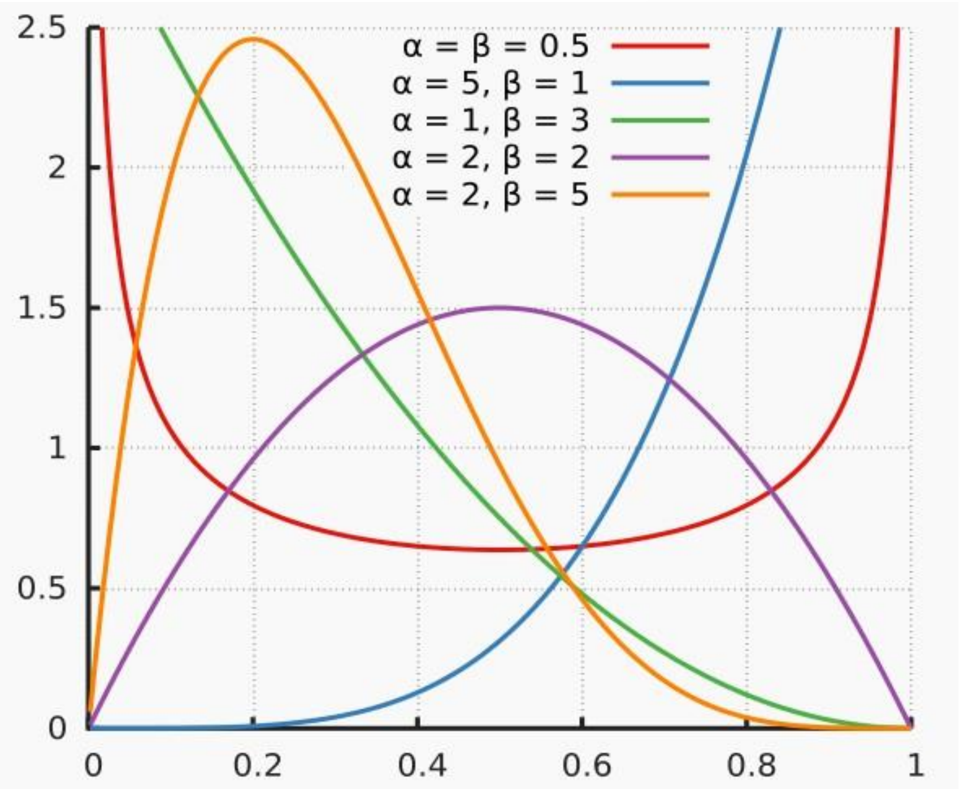
$\text{Beta}(\alpha + 4, \beta + 10)$

Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of* $\frac{y_j}{n_j}$*: (number of rats with tumors)/(total number of rats).*

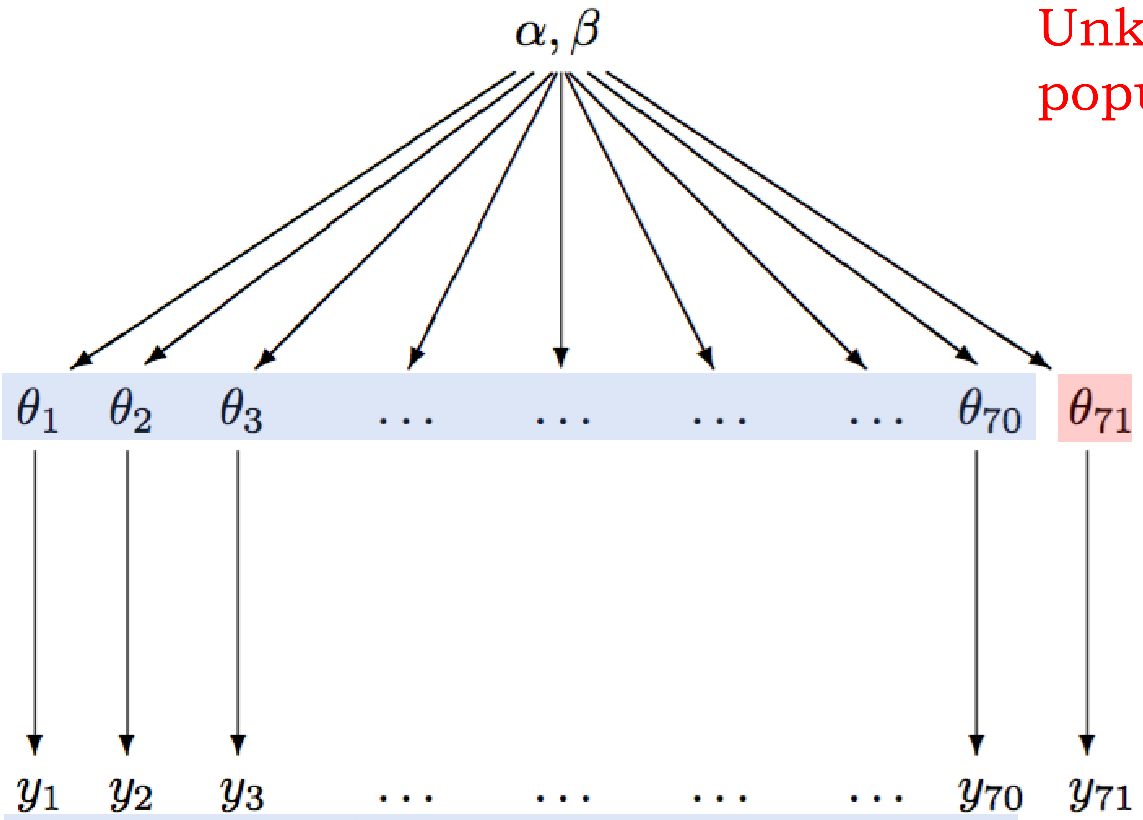
Recall: $Beta(\alpha, \beta)$

Notation	Beta(α, β)
Parameters	α > 0 shape (real) β > 0 shape (real)
Support	$x \in (0, 1)$
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$
CDF	$I_x(\alpha, \beta)$
Mean	$\boxed{\mathrm{E}[X] = \frac{\alpha}{\alpha+\beta}}$ $\mathrm{E}[\ln X] = \psi(\alpha) - \psi(\alpha+\beta)$ (see digamma function and see section: Geometric mean)
Median	$I_{\frac{1}{2}}^{[-1]}(\alpha, \beta)$ (in general) $\approx \frac{\alpha - \frac{1}{3}}{\alpha+\beta-\frac{2}{3}}$ for $\alpha, \beta > 1$
Mode	$\frac{\alpha - 1}{\alpha+\beta-2}$ for α, β >1
Variance	$\boxed{\mathrm{var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$ $\mathrm{var}[\ln X] = \psi_1(\alpha) - \psi_1(\alpha+\beta)$ (see trigamma function and see section: Geometric variance)

Beta density functions



Hierarchical Model



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	15/47	15/46	9/24

4/14

Unknown hyper-parameters describe the population distribution of θs

Key idea:
assume unknown parameters of different experiments are *iid* samples from a common population

$$\frac{\alpha}{\alpha+\beta} = E(\theta) = 0.136,$$

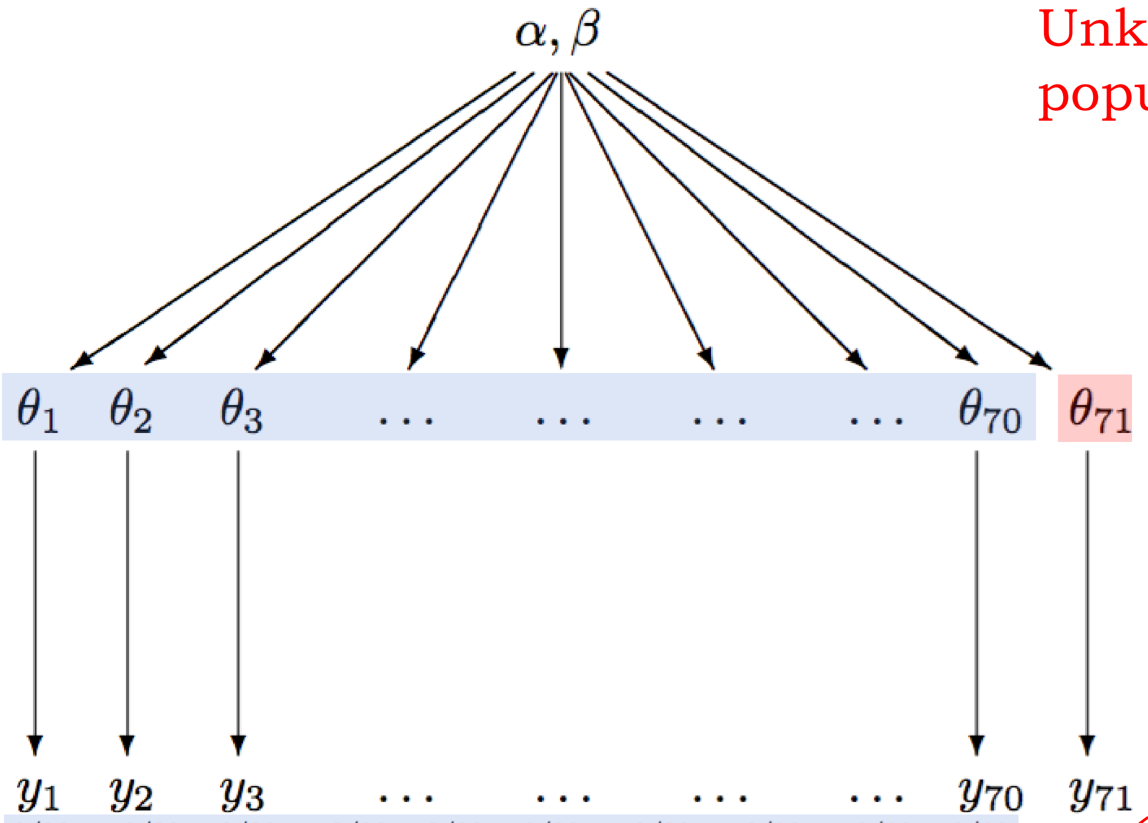
$$\alpha+\beta = \frac{E(\theta)[1-E(\theta)]}{\mathrm{var}(\theta)} - 1 = 10.076$$

$$\hat{\theta}_j = \frac{y_j}{n_j}, j = 1, \dots, 70.$$

mean = 0.136
standard deviation = 0.103

The First-Thought Solution

Unknown hyper-parameters describe the population distribution of θs



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	15/47	15/46	9/24

4/14

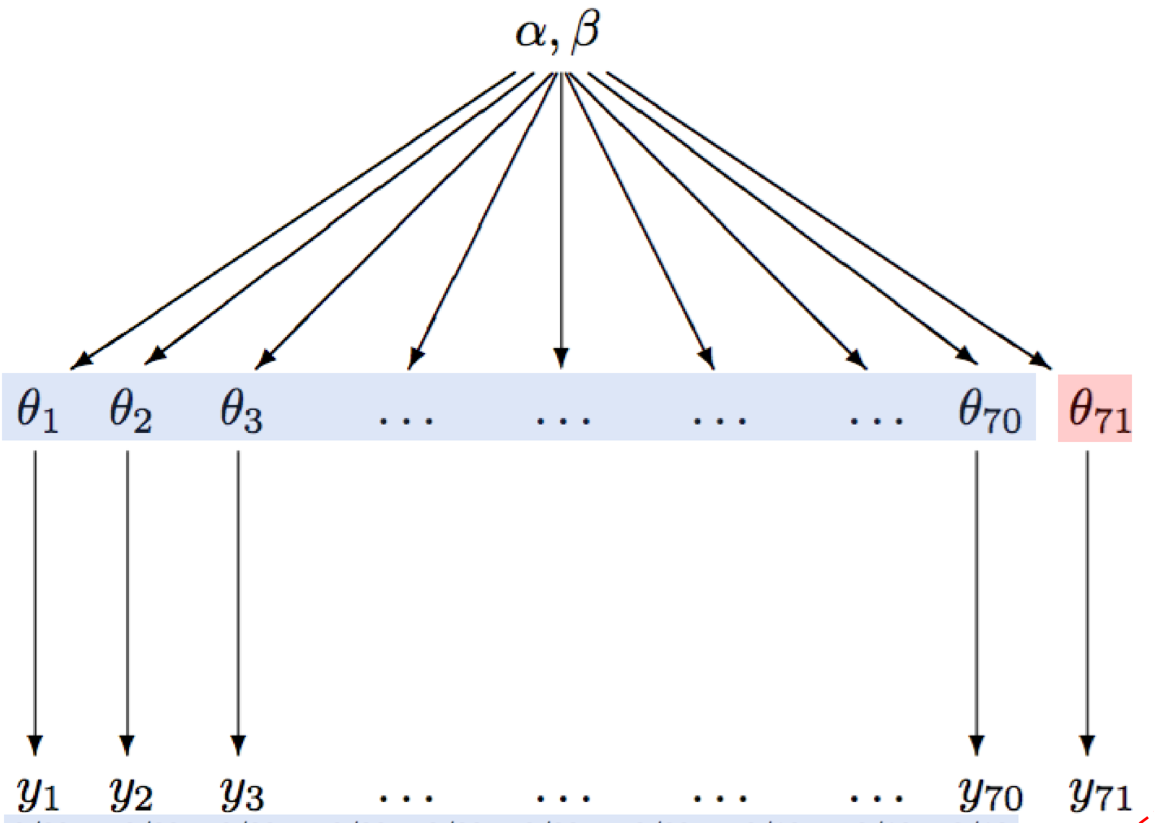
Key idea:
assume unknown parameters of different experiments are *iid* samples from a common population

Posterior
Beta(5.4, 18.6)

(α, β) is $(1.4, 8.6)$

mean = 0.136
standard deviation = 0.103

The First-Thought Solution



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	15/47	15/46	9/24

4/14

mean = 0.136
standard deviation = 0.103

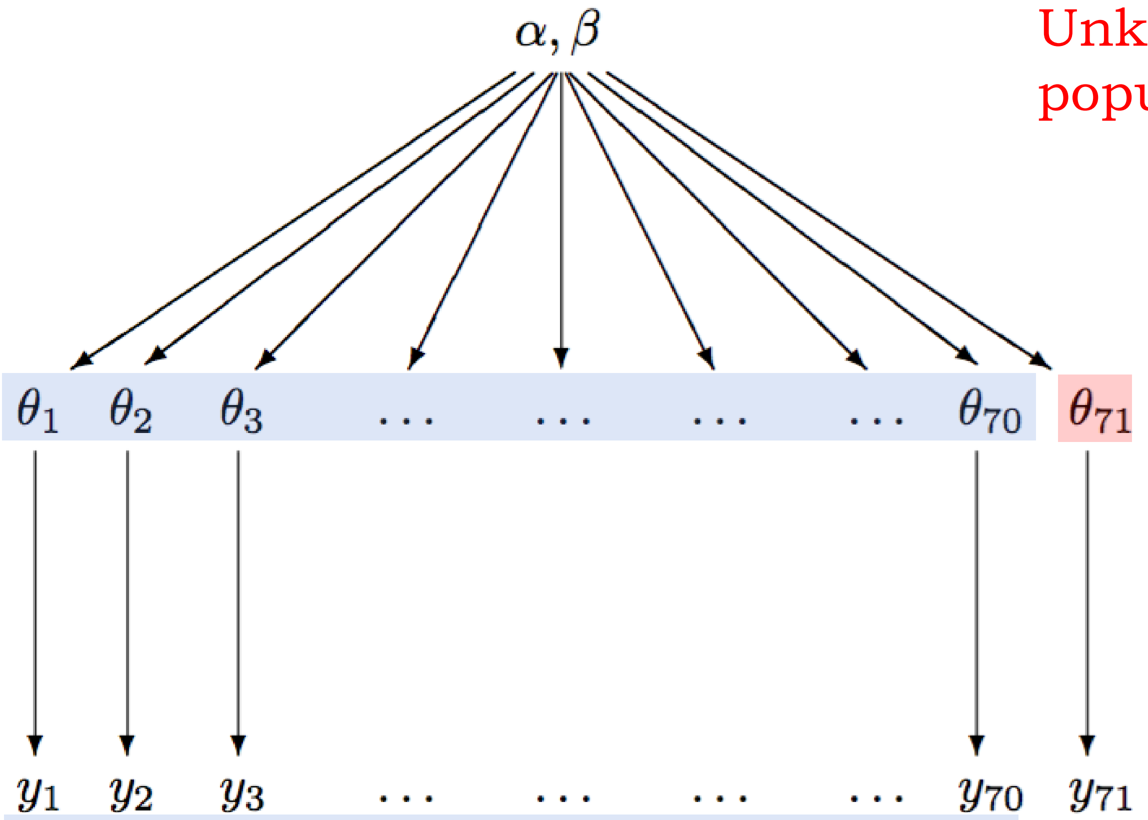
(α, β) is $(1.4, 8.6)$

Posterior
Beta$(5.4, 18.6)$

Problems of this solution:

- This is not a Bayesian calculation because it is not based on any specified full probability model. The point estimate for α and β seems arbitrary, and ignores some posterior uncertainty.
- If we wanted to use the estimated prior distribution for inference about the first 70 experiments, then the data would be used twice.

A Full Bayesian Model



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	15/47	15/46	9/24

4/14

Unknown hyper-parameters describe the population distribution of θs

Key idea:

- Treat hyper-parameters as random variables
- Assign a prior distribution to them

Prior

$$p(\alpha,\beta,\theta_1,\theta_2,\ldots,\theta_{71}) = p(\alpha,\beta)p(\theta_1,\theta_2,\ldots,\theta_{71}|\alpha,\beta)$$

Posterior $p(\alpha,\beta,\theta_1,\theta_2,\ldots,\theta_{71}|y)$

$$\propto p(\alpha,\beta,\theta_1,\ldots,\theta_{71})p(y_1,\ldots,y_{71}|\theta_1,\theta_2,\ldots,\theta_{71},\alpha,\beta)$$

$$= p(\alpha,\beta,\theta_1,\ldots,\theta_{71})p(y_1,\ldots,y_{71}|\theta_1,\theta_2,\ldots,\theta_{71})$$

A typical multi-parameter case!

Shall we build a model for y with α,β directly?

Hierarchical Models

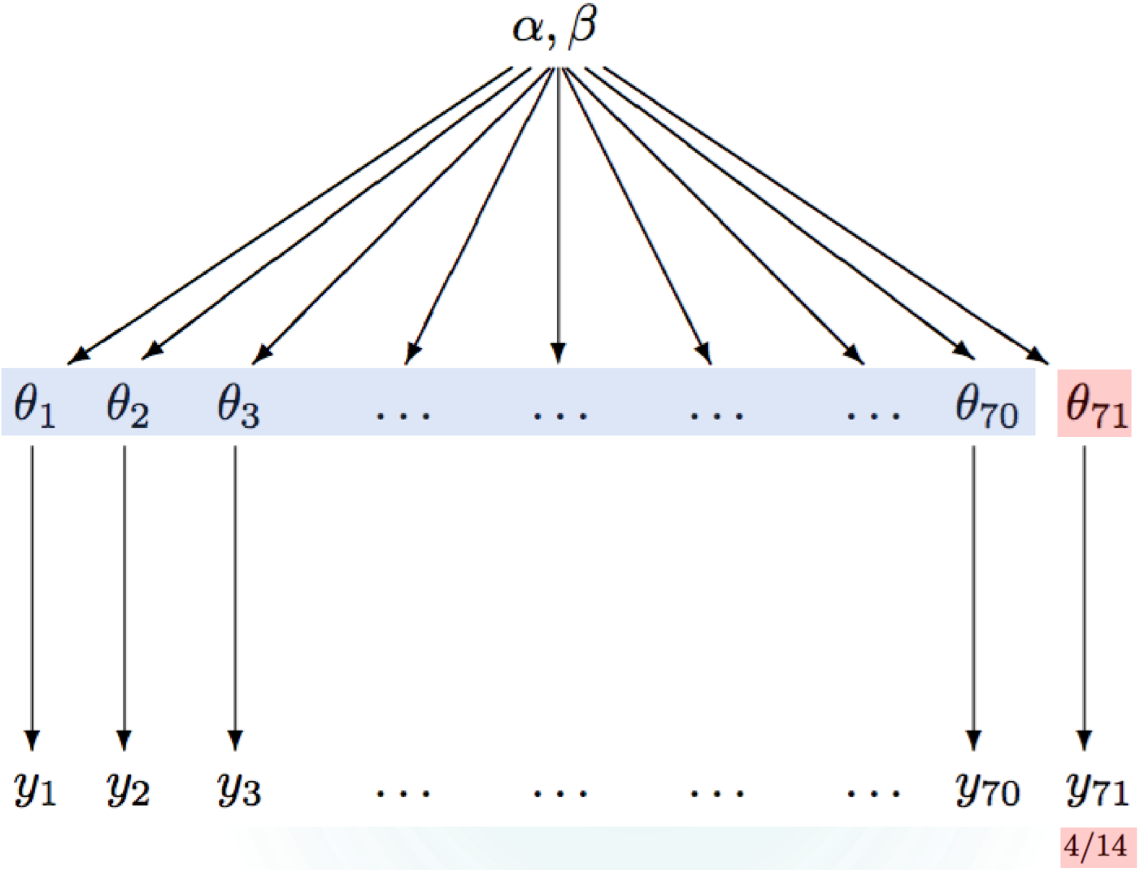
- Powerful technique for describing complex models. Idea is to break the model down into smaller easier understood pieces, which when put together describes the joint distribution of all data and parameters.
  - ✓ Note 1: actually all of the models we have seen so far have been hierarchical, but most only had two levels to the hierarchy.
  - ✓ Note 2: there may be a hierarchical structure within each piece.

- Why go hierarchical?
  - ✓ Non-hierarchical models with few parameters generally don’t fit the data well.
  - ✓ Non-hierarchical models with many parameters then to fit the data well, but have poor predictive ability (overfitting)
  - ✓ Hierarchical models can often fit data with a small number of parameters but can also do well in prediction.
  - ✓ Hierarchical models with more parameters than data points can be useful and can give reasonable answers

How to build a hierarchical model?

- Assumption: Exchangeability
- Full Bayesian Hierarchical Model

Artificial Example



70 historical experiments

0/20	1/20	0/20	0/20	1/20	0/21	0/20
1/20	0/19	1/18	2/21	0/19	1/20	1/19
0/21	0/20	0/20	0/20	0/20	0/19	0/20
1/20	1/19	1/18	1/20	1/19	1/20	1/18
2/22	1/20	2/20	0/19	0/20	1/21	0/20
19/21	17/19	17/18	19/20	17/19	16/20	17/18
20/20	18/20	20/20	16/20	19/21	18/21	16/20
19/22	17/19	17/18	18/20	16/19	16/20	18/18
20/23	18/20	20/20	19/20	19/21	18/21	19/20
19/20	17/19	17/19	17/20	18/19	19/20	16/18

Key idea:
assume unknown parameters of different experiments are *i.i.d.* samples from a common population

Exchangeability

- If no information—other than the data y—is available to distinguish any of the θ_j's from any of the others, and no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution.
- This symmetry is represented probabilistically by exchangeability: the joint distribution $p(\theta_1,\ldots,\theta_J)$ is invariant to permutations of the indexes $(1,\ldots,J)$.

– For example, in the rat tumor example, we have no prior reason to assume that $\theta_{70} < \theta_{71}$ is more likely than $\theta_{70} > \theta_{71}$. In fact, for the information given, the order that the groups are listed in is meaningless. So for this problem, it seems reasonable to have the distribution on the θ_j's be exchangeable.

Exchangeability

- If no information—other than the data y—is available to distinguish any of the θ_j's from any of the others, and no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution.
- This symmetry is represented probabilistically by exchangeability: the joint distribution $p(\theta_1, \ldots, \theta_J)$ is invariant to permutations of the indexes $(1, \ldots, J)$.

– If there is information in the indices about the distributions, exchangeability is usually not reasonable. Suppose that different pure-bred rat strains were used for groups 50 to 71 than those used for groups 1 to 49. Then exchanging indices 49 and 50 would not be reasonable (probably).

Exchangeability

- Note that exchangeability does not imply independence

  - For example, the multivariate normal model $y \sim N_d(\mu_1, \Sigma)$, where $Var(y_j) = \sigma^2$ for all i and $Corr(y_i, y_j) = \rho \neq 0$ for all i and j.
  - It is exchangeable, but obviously not independent.

  - Exchangeability implies the marginal distributions for each component are the same (identically distributed), but nothing about independence. In fact the dependence between the different components must be the same.
  - However all *i.i.d.* models are exchangeable.

Exchangeability

- The simplest way to introduce symmetry:

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

de Finetti’s Theorem

In the limit as $J \to \infty$, any suitably well-behaved exchangeable distribution on $(\theta_1, \dots, \theta_J)$ can be expressed as a mixture of independent and identical distributions.

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

A theoretical support to the hierarchical model

- Note: the conditional independence of the θ_j's holds in many hierarchical model, e.g. the rat tumor example. It will also be useful when we introduce Gibbs sampling.

Example

A hypothetical case
假想一位研究员想了解6个国家在某阶段的新冠肺炎得病率(rate in 10,000,000 population). Call these $\theta_1, \ldots, \theta_6$. What can you say about θ_6, the rate in the eighth province, before he / she get any data?

- **Scenario I**: we have **NO** information to distinguish any of the 6 countries at all.
  - We have to model the 6 rates exchangeably.
  - For example, randomly sample 5 countries: 6363, 8212, 2102, 128, 739
- **Scenario II**: we know that the 6 countries are: 美国, 英国，意大利，韩国，伊朗，中国, but selected in a random order.
  - The 6 rates should still be modeled exchangeably.
  - However, our prior distribution for the 6 rates may have to change, e.g. to priors with heavy tails.

Example

A hypothetical case

假想一位研究员想了解6个国家在某阶段的新冠肺炎得病率(rate in 10,000,000 population). Call these $\theta_1, \ldots, \theta_6$. What can you say about θ_6, the rate in the eighth province, before he / she get any data?

- **Scenario III**: we observed data of other 5 countries except for 中国.
  - Even before seeing the 5 observed values, we cannot assign an exchangeable prior distribution to the set of 6 rates any more
  - Once we see the 5 observed values, a reasonable posterior distribution for θ_6 plausibly should have most of its mass below the smallest observed rate, i.e., $p(\theta_6 < \min(\theta_1, \ldots, \theta_5) | \theta_1, \ldots, \theta_5)$ should be large.
  - Actually the observed rates for the 6 countries 美国, 英国，意大利，韩国，伊朗，中国 are 6363, 8212, 2102, 128, 739, 0.97. (based on the data in January, 2022)

Fully Bayesian Hierarchical Models

Suppose we have the following hierarchical model:

$$\vec{y}_i|\theta,\phi \sim p(\vec{y}_i|\theta_i),$$
$$\vec{\theta}|\phi \sim p\left(\vec{\theta}\middle|\phi\right) = \prod_i p(\theta_i|\phi),$$
$$\phi \sim p(\phi)$$

The joint prior is

$$p(\vec{\theta},\phi) = p(\phi)p\left(\vec{\theta}\middle|\phi\right)$$

The joint posterior is

$$\begin{aligned} &p\left(\vec{\theta},\phi|y\right) \\ &\propto p(\phi)p\left(\vec{\theta}\middle|\phi\right)p(y|\theta,\phi) \\ &= p(\phi)p\left(\vec{\theta}\middle|\phi\right)p(y|\theta) \\ &= p(\phi)p\left(\vec{\theta}\middle|\phi\right)\prod_i p(\vec{y}_i|\theta_i) \end{aligned}$$

How to analyze the full Bayesian hierarchical model?

– Binomial Model

– Normal Model

Inference of interest

The posterior

$$p(\vec{\theta}|y) \longrightarrow p(\vec{\theta}, \phi|y) \longrightarrow p(\vec{\theta}, \phi) = p(\phi)p(\vec{\theta}|\phi)$$

Posterior predictive distributions. There are two situations of interest:

1. $\tilde{y}$ for an existing θ_j
2. $\tilde{y}$ for a new θ_j

Fully Bayesian Analysis of Conjugate Hierarchical Models

Three steps for analytical analysis:

1. Write the joint posterior density, $p(\theta, \phi|y)$, in unnormalized form as a product of the hyperprior distribution $p(\phi)$, the population distribution $p(\theta|\phi)$, and the likelihood $p(y|\theta)$.
2. Determine analytically the conditional posterior density of θ given the hyperparameters ϕ; for fixed observed y, this is a function of ϕ, $p(\theta|\phi, y)$.
3. Estimate ϕ using the Bayesian paradigm; that is, obtain its marginal posterior distribution, $p(\phi|y)$.

Inverse step 3 and 2 to draw samples from the joint posterior

Two ways to get marginal posterior:

1. Bruce force integration:

$$p(\phi|y) = \int p(\theta, \phi|y) d\theta.$$

High dimensional integration is usually difficult

2. Conditional probability formula:

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}.$$

The normalizing constant depends on ϕ, and can be difficult to calculate

Example I: Binomial Model

Hierarchical model for the rat tumors data

$$y_j \sim \mathrm{Bin}(n_j, \theta_j) \qquad \theta_j \sim \mathrm{Beta}(\alpha, \beta) \qquad j = 1, \ldots, J,\ J = 71$$

# of tumor case — Known sample size — Unknown hyper-parameters — size of experiment population

Joint posterior:

Prior for the hyper-parameters (to be specified)

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

Conditional posterior:

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

Beta density

Marginal posterior:

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

Not a standard density

What kind of prior do you prefer? What’s your rule?

Potential Ways to Specify Hyper-Prior

Marginal posterior:

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

Goes to 1 when α & β go to infinity

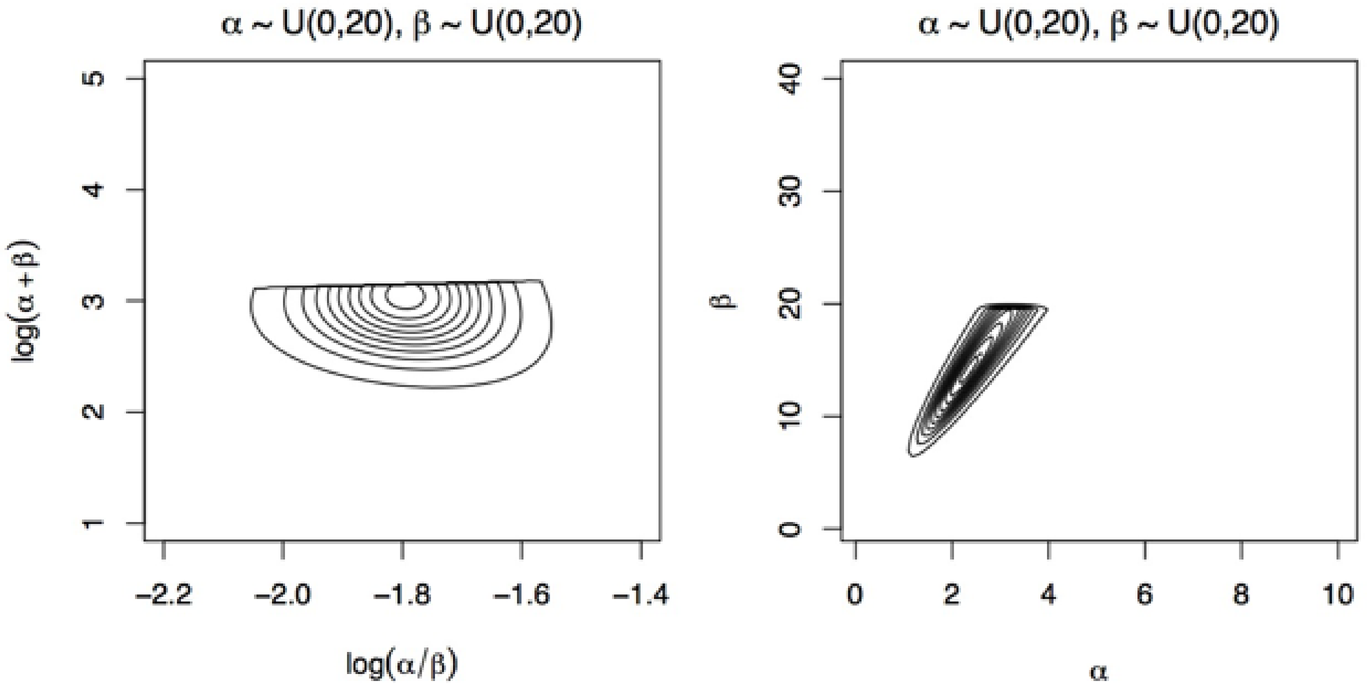
No a proper density
(integral=∞)

$p(\alpha,\beta)$ must have a very light tail to make $p(\alpha,\beta|y)$ proper

Possibility 1. Improper uniform prior $p(\alpha,\beta) \propto 1$ ✗

Potential Ways to Specify Hyper-Prior

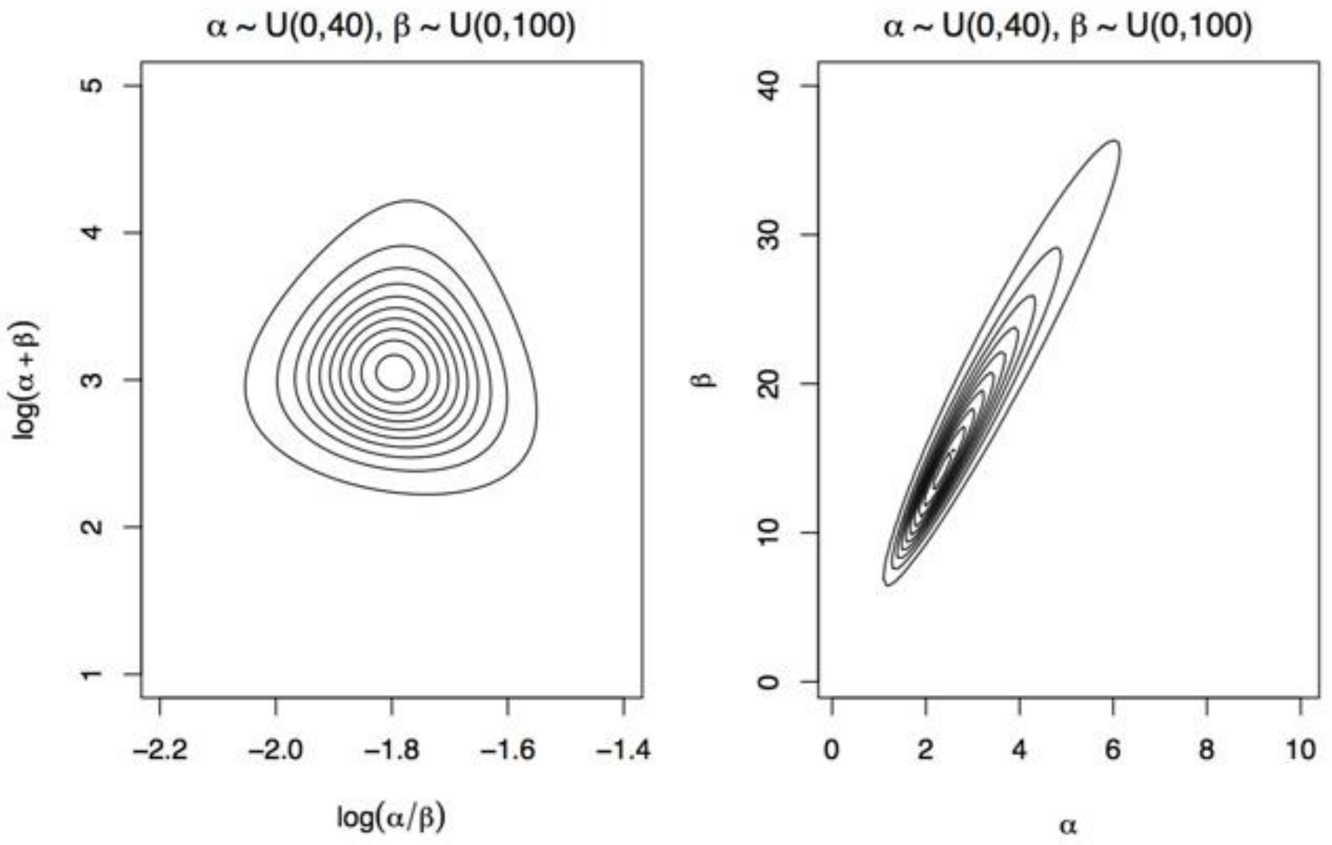
- Let's try an independent prior on α and β: $\alpha \sim Unif(0{,}20), \beta \sim Unif(0{,}20)$
- This gives $p(\alpha, \beta|y) \propto I(\alpha \leq 20) I(\beta \leq 20) \prod_{j=1}^{J} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}$



- Note that the posterior gets clipped due to the upper limits on α and β.

Potential Ways to Specify Hyper-Prior

- So naive implementation of uniform priors seem work.
- Let’ s extend those limits so that they match the data better



Potential Ways to Specify Hyper-Prior

Marginal posterior:

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

Possibility 1. Restricted uniform prior $\alpha \sim \mathcal{U}(0,20), \beta \sim \mathcal{U}(0,20)$

$p(\alpha,\beta)$ must have a very light tail to make $p(\alpha,\beta|y)$ proper

Possibility 2. Improper uniform prior $p(\frac{\alpha}{\alpha+\beta}, \alpha+\beta) \propto 1$ ✗

Transformation of parameters:

"sample size" of additional data in prior

$$(\alpha,\beta) \longrightarrow (\log(\tfrac{\alpha}{\beta}), \log(\alpha+\beta))$$

$$\log\left(\frac{\alpha}{\beta}\right) = logit\left(\frac{\alpha}{\alpha+\beta}\right)$$

"prior mean"

Advantage of the reparameterization: "prior sample size" & "prior mean" are separated, and can be assign prior distribution independently

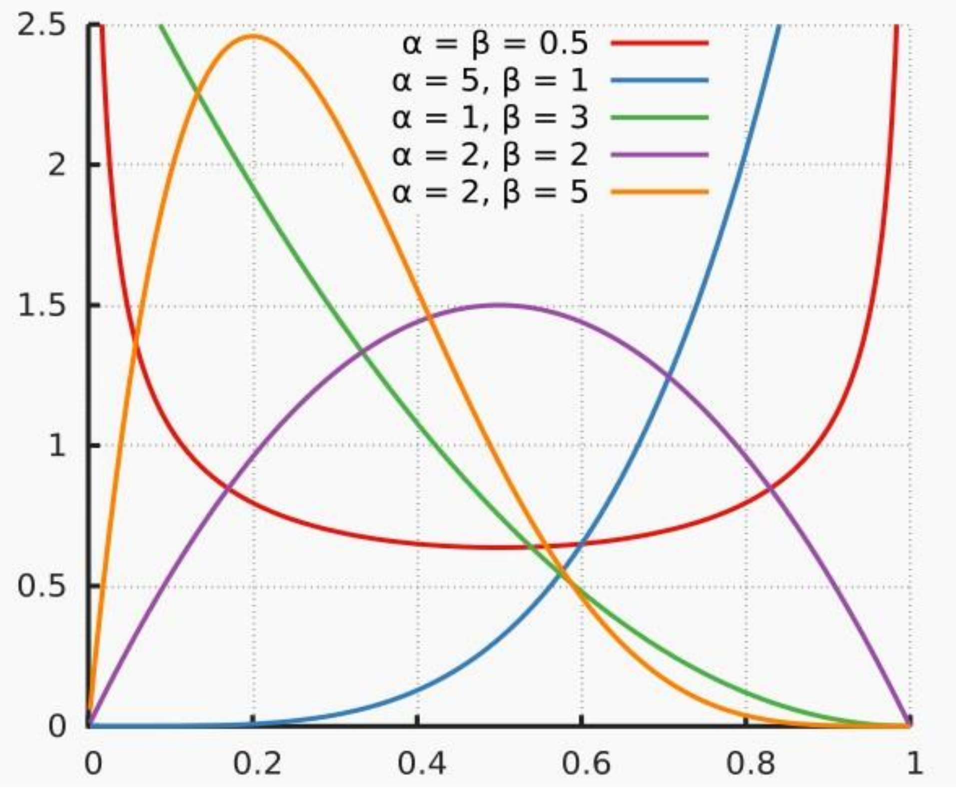
Possibility 3. Improper uniform prior $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)) \propto 1$ ✗

Possibility 4. A diffuse hyperprior density uniform on $(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2})$

Recall: $Beta(\alpha, \beta)$

Notation	Beta(α, β)
Parameters	α > 0 shape (real) β > 0 shape (real)
Support	$x \in (0,1)$
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$
CDF	$I_x(\alpha,\beta)$
Mean	$\boxed{\mathrm{E}[X] = \frac{\alpha}{\alpha+\beta}}$ $\mathrm{E}[\ln X] = \psi(\alpha) - \psi(\alpha+\beta)$ (see digamma function and see section: Geometric mean)
Median	$I_{\frac{1}{2}}^{[-1]}(\alpha,\beta)$ (in general) $\approx \frac{\alpha-\frac{1}{3}}{\alpha+\beta-\frac{2}{3}}$ for $\alpha, \beta > 1$
Mode	$\frac{\alpha-1}{\alpha+\beta-2}$ for α, β >1
Variance	$\boxed{\operatorname{var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$ $\operatorname{var}[\ln X] = \psi_1(\alpha) - \psi_1(\alpha+\beta)$ (see trigamma function and see section: Geometric variance)

Beta density functions



Potential Ways to Specify Hyper-Prior

Marginal posterior:

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

Possibility 1. Restricted uniform prior $\alpha \sim \mathcal{U}(0,20), \beta \sim \mathcal{U}(0,20)$

$p(\alpha,\beta)$ must have a very light tail to make $p(\alpha,\beta|y)$ proper

Possibility 2. Improper uniform prior $p(\frac{\alpha}{\alpha+\beta}, \alpha+\beta) \propto 1$ ✗

Transformation of parameters:

"sample size" of additional data in prior

$$(\alpha,\beta) \longrightarrow (\log(\tfrac{\alpha}{\beta}), \log(\alpha+\beta))$$

$\log\left(\frac{\alpha}{\beta}\right) = logit\left(\frac{\alpha}{\alpha+\beta}\right)$ "prior mean"

Advantage of the reparameterization: "prior sample size" & "prior mean" are separated, and can be assign prior distribution independently

Possibility 3. Improper uniform prior $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)) \propto 1$ ✗

Possibility 4. A diffuse hyperprior density uniform on $(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2})$ ✓

$$p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2} \quad \text{or} \quad p\left(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)\right) \propto \alpha\beta(\alpha+\beta)^{-5/2}$$

Possibility 5, 6, ⋯

Computing the Marginal Posterior Density of the Hyperparameters

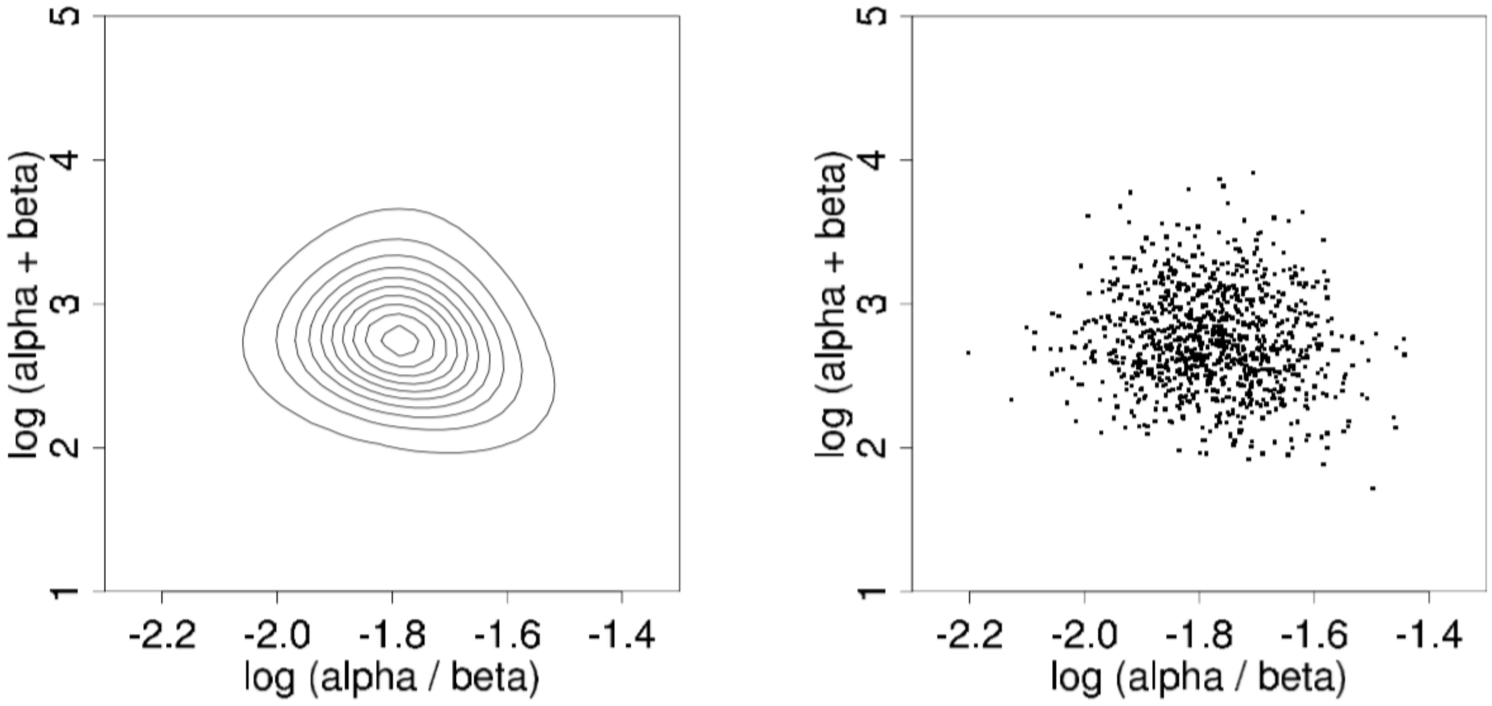
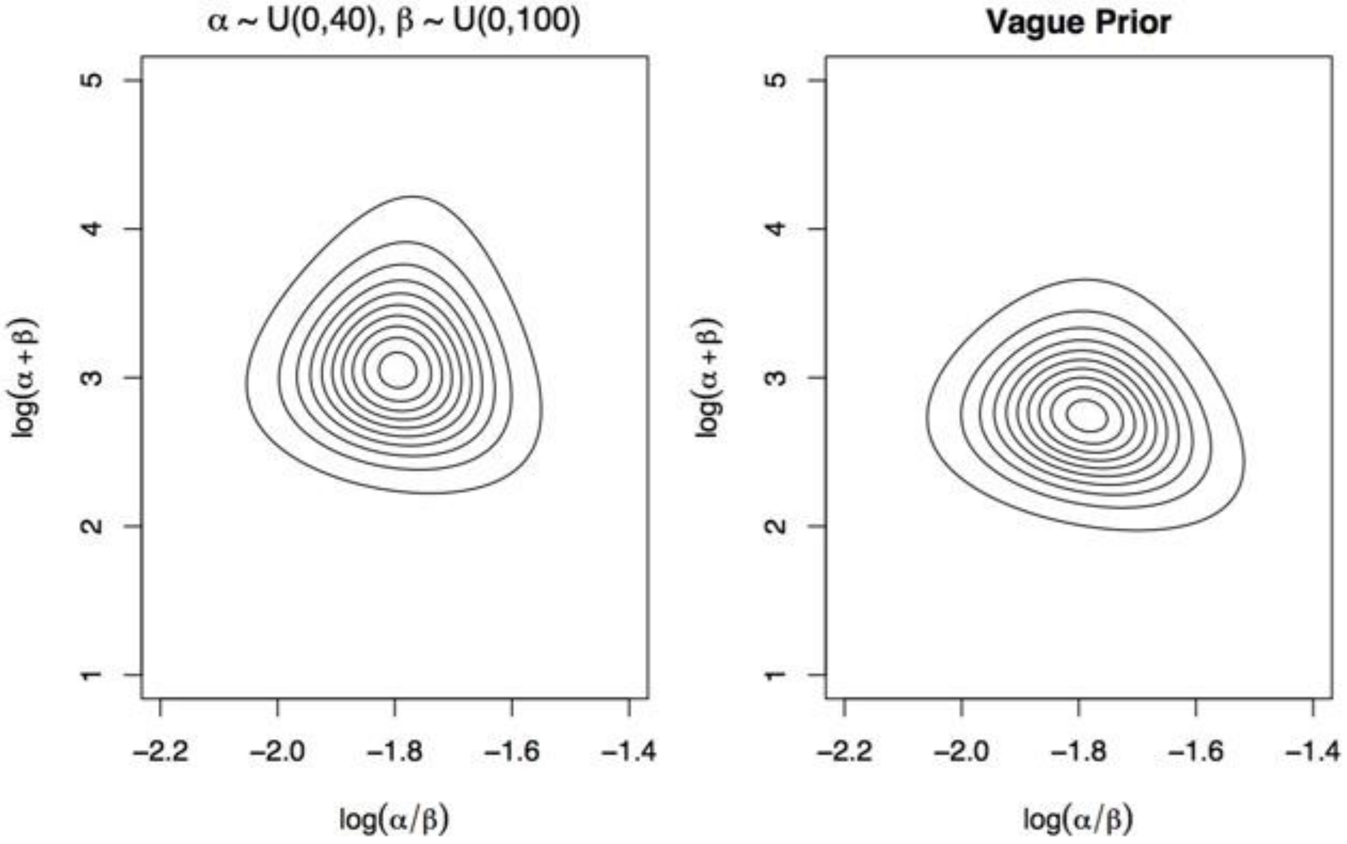


Figure 5.3 *(a) Contour plot of the marginal posterior density of* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ *for the rat tumor example. Contour lines are at* $0.05, 0.15, \ldots, 0.95$ *times the density at the mode. (b) Scatterplot of 1000 draws* $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ *from the numerically computed marginal posterior density.*

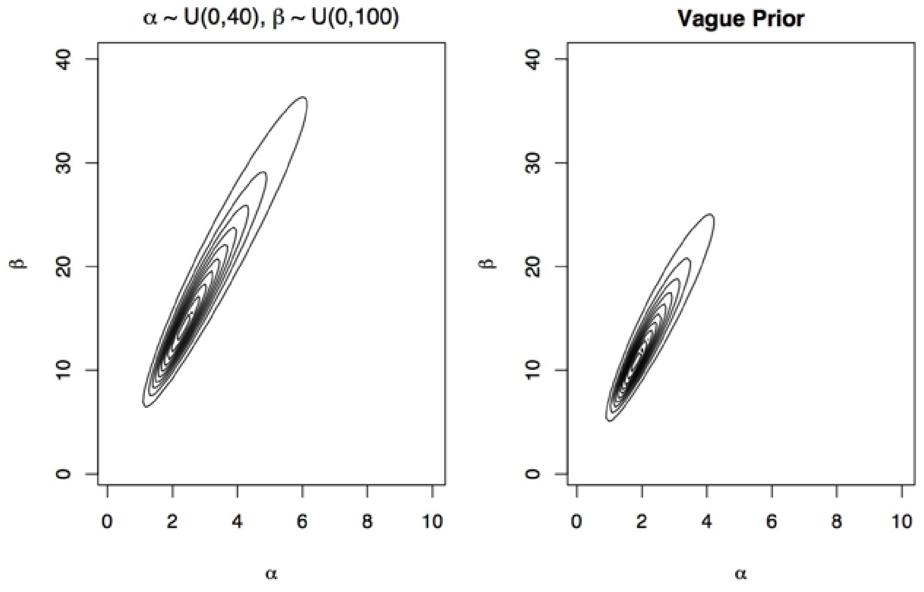
Comparing different Hyper-Priors

- Now lets compare the marginal posterior under this prior with the posterior under the vague prior suggested by the book.



Comparing different Hyper-Priors

- So as expected, the vague prior pulls $\alpha+\beta$ down.



- The posterior means of α and β (as calculated by simulation) are

Prior	Vague	$\alpha \sim U(0,20), \beta \sim U(0,20)$	$\alpha \sim U(0,40), \beta \sim U(0,100)$
α	2.398	2.482	3.448
β	14.291	14.805	20.649

Inference for Parameter θ

- Now we are really interested in the θ_j, the tumor rates in the different groups. So we want to determine

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

- This does not have a nice closed form as integrating α and β is ugly, so we have to use simulation.

- How to do the simulation?

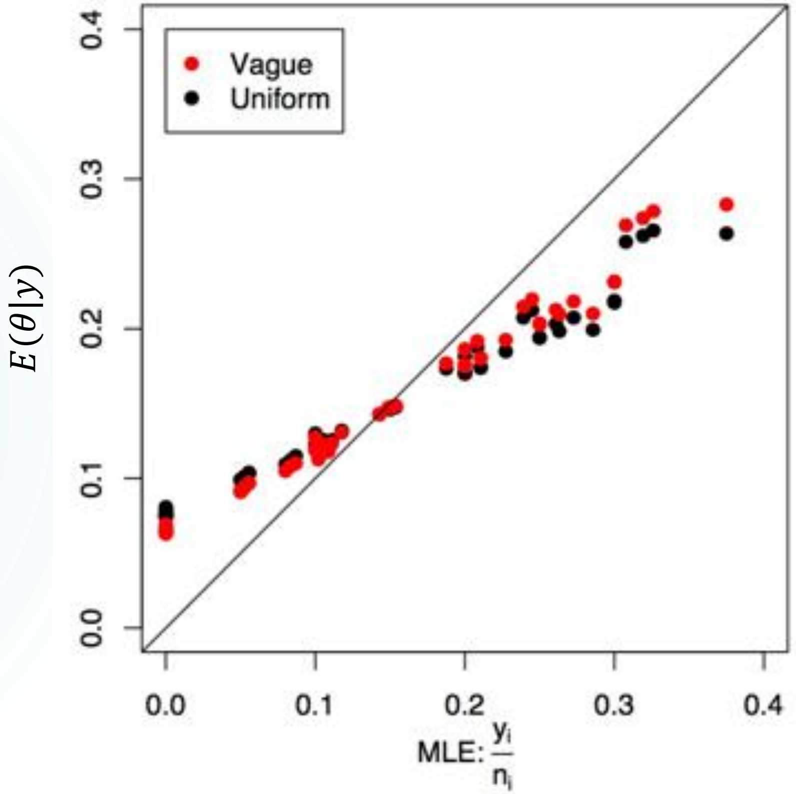
Detailed Procedure for Simulation

- Draw random samples (e.g. 1000) from the joint posterior distribution of $(\alpha, \beta, \theta_1, \dots, \theta_J)$

1. Simulate 1000 draws of $\left(\log\left(\frac{\alpha}{\beta}\right), \log(\alpha+\beta)\right)$ from their posterior distribution displayed in Figure 5.3, using the discrete-grid sampling procedure introduced in Lec 4.
2. For $l = 1, \dots, 1000$:
   a) Transform the lth draw of $\left(\log\left(\frac{\alpha}{\beta}\right), \log(\alpha+\beta)\right)$ to the scale (α, β) to yield a draw of the hyperparameters from their marginal posterior distribution
   b) For each $j = 1, \dots, J$, sample θ_j from its conditional posterior distribution, $\theta_j | \alpha, \beta, y \sim Beta(\alpha + y_j, \beta + n_j - y_j)$.

Inference for Parameter θ

- In this case the $\alpha \sim Unif(0{,}40), \beta \sim Unif(0{,}100)$ prior shrinks the estimates more than the vague prior, though they are shrinking to about the same place.



Inference for Parameter θ

- This is supported by the posterior means (as calculated by simulation) for

Prior	Vague	$\alpha \sim Unif(0,40), \beta \sim Unif(0,100)$
$\frac{\alpha}{\alpha+\beta}$	0.144	0.143
$\alpha+\beta$	16.689	24.097

- For a new group

$$E(\theta|y) = E\big(E(\theta|\alpha,\beta,y)\big) = E\left(\frac{\alpha}{\alpha+\beta}\middle|y\right)$$

- So the two priors seem to be shrinking to the same place.

Inference for Parameter θ

- For an observed group j,

$$E(\theta_j|y) = E\left(E(\theta_j|\alpha,\beta,y)\right) = E\left(\frac{\alpha+y_j}{\alpha+\beta+n_j}\middle|y\right)$$

- Note that

$$\frac{\alpha+y_j}{\alpha+\beta+n_j} = \frac{\alpha+\beta}{\alpha+\beta+n_j}\frac{\alpha}{\alpha+\beta} + \frac{n_j}{\alpha+\beta+n_j}\frac{y_j}{n_j}$$

- So this agrees with more shrinking for the uniform prior as the effective sample size from the prior component $(\alpha+\beta)$ is larger for the uniform prior.

How to analyze the full Bayesian hierarchical model?

– Binomial Model

– Normal Model

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

Can this be just by chance?

Table 5.2 Observed effects of special preparation on SAT-V scores in eight randomized experiments. Estimates are based on separate analyses for the eight experiments.

Pooled estimate

Posterior mean = 7.7

Posterior variance $\left(\sum_{j=1}^{8} \frac{1}{\sigma_j^2}\right)^{-1} = 16.6$

Stand error = $\sqrt{16.6} = 4.1$

95% posterior interval $[-0.5, 15.9]$

or approximately $[8 \pm 8]$

Remark:

- Assume the data in Table 5.2 are eight normally distributed observations with known variances.
- Use a noninformative prior distribution

Difficulties with the Separate and Pooled Estimates

Separate Estimate

- The effect in school A is estimated as 28.4 with a standard error of 14.9
- Prob(the true effect in A is more than 28.4) = 0.5

Pooled Estimate

- The effect in school A is estimated as 7.7 with a standard error of 4.1
- Prob (the true effect in A is less than 7.7) = 0.5
- Prob (the true effect in A is less than the true effect in C) = 0.5

Neither estimate is fully satisfactory, and we would like a compromise that combines information from all eight experiments without assuming all the θ_j's to be equal.

Example II: Normal Model

Data structure of a hierarchical normal model

Unknown group mean Known variance Group sample size

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

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

Joint posterior:

Prior for the hyper-parameters (to be specified)

$$\begin{aligned} p(\theta, \mu, \tau|y) &\propto p(\mu, \tau) p(\theta|\mu, \tau) p(y|\theta) \\ &\propto p(\mu, \tau) \prod_{j=1}^{J} \mathrm{N}(\theta_j|\mu, \tau^2) \prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j}|\theta_j, \sigma_j^2), \end{aligned}$$

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

Conditional posterior:

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

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\overline{y}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}.$$

Marginal posterior:

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

Specifying Hyper-Prior

Marginal posterior: $p(\mu,\tau|y) \propto p(\mu,\tau)\prod_{j=1}^{J} \mathrm{N}(\overline{y}_{.j}|\mu,\sigma_j^2+\tau^2)$

Noninformative uniform hyperprior distribution to μ given τ:

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

Alternative form of the marginal posterior:

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

$$\mu|\tau,y \sim \mathrm{N}(\hat{\mu},V_\mu)$$

$$\hat{\mu} = \frac{\sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}\overline{y}_{.j}}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}} \quad \text{and} \quad V_\mu^{-1} = \sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}$$

$$\begin{aligned} p(\tau|y) &= \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} \\ &\propto \frac{p(\tau)\prod_{j=1}^{J}\mathrm{N}(\overline{y}_{.j}|\mu,\sigma_j^2+\tau^2)}{\mathrm{N}(\mu|\hat{\mu},V_\mu)} \end{aligned}$$

holds for any μ

$$\mu = \hat{\mu}$$

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

Potential ways to specify

1. $p(\tau) \propto 1$ ✓
2. $p(\log\tau) \propto 1$ ✗
3. scaled inverse-χ^2 ✓

Detailed Procedure for Simulation

- Draw random samples from the joint posterior distribution

1. Sample τ_k from $p(\tau|y)$
2. Sample μ_k from $p(\mu_k|\tau_k, y) = N(\mu_k|\hat{\mu}_k, V_{\mu_k})$ where

$$\hat{\mu}_k = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau_k^2} \bar{y}_{\cdot j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau_k^2}} \quad \text{and} \quad V_{\mu_k}^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau_k^2}$$

Detailed Procedure for Simulation

3. Sample θ_k from $p(\theta_k|\mu_k,\tau_k,y)$. In this case, the individual components are conditionally independent given μ_k,τ_k and y, giving

$$\theta_{j,k} \sim N(\hat{\theta}_{j,k}, V_{j,k})$$

where

$$\hat{\theta}_{j,k} = \frac{\frac{1}{\sigma_j^2}\bar{y}_{\cdot j} + \frac{1}{\tau_k^2}\mu_k}{\frac{1}{\sigma_j^2} + \frac{1}{\tau_k^2}} \qquad V_{j,k} = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau_k^2}}$$

Results from Hierarchical Model

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

Table 5.2 Result from separate estimate

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

Table 5.3: Summary of 200 simulations of the treatment effects $\theta_j|y$ in the eight schools.

Results from Hierarchical Model

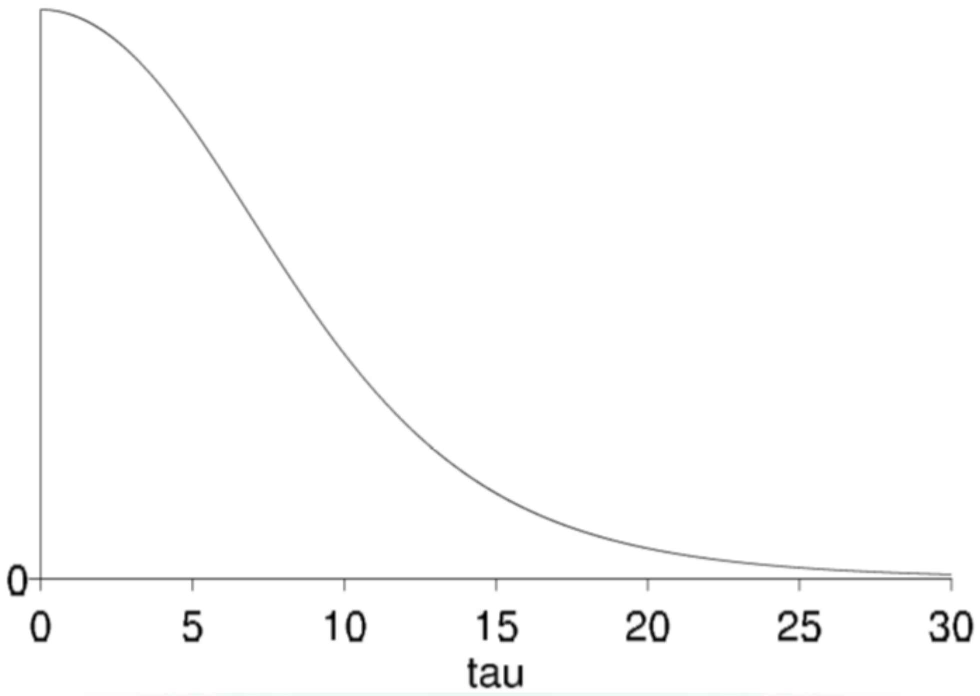
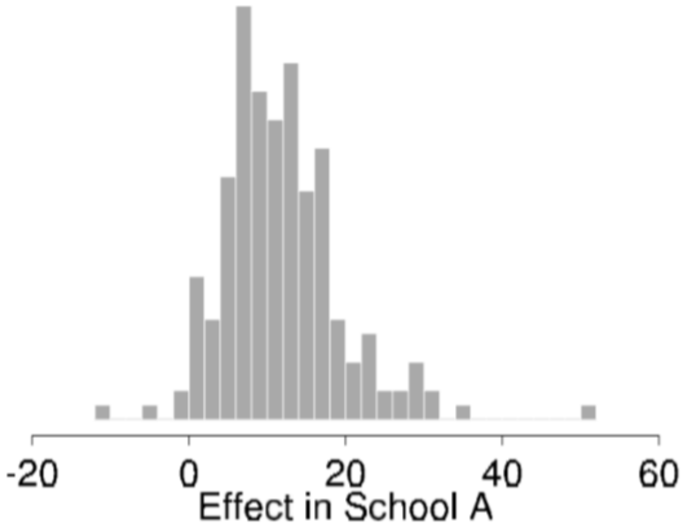
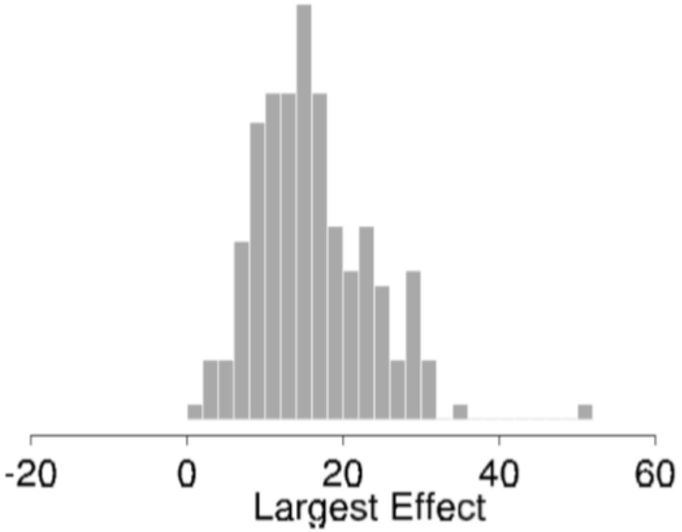


Figure 5.5 Marginal posterior density $p(\tau|y)$





Results from Hierarchical Model

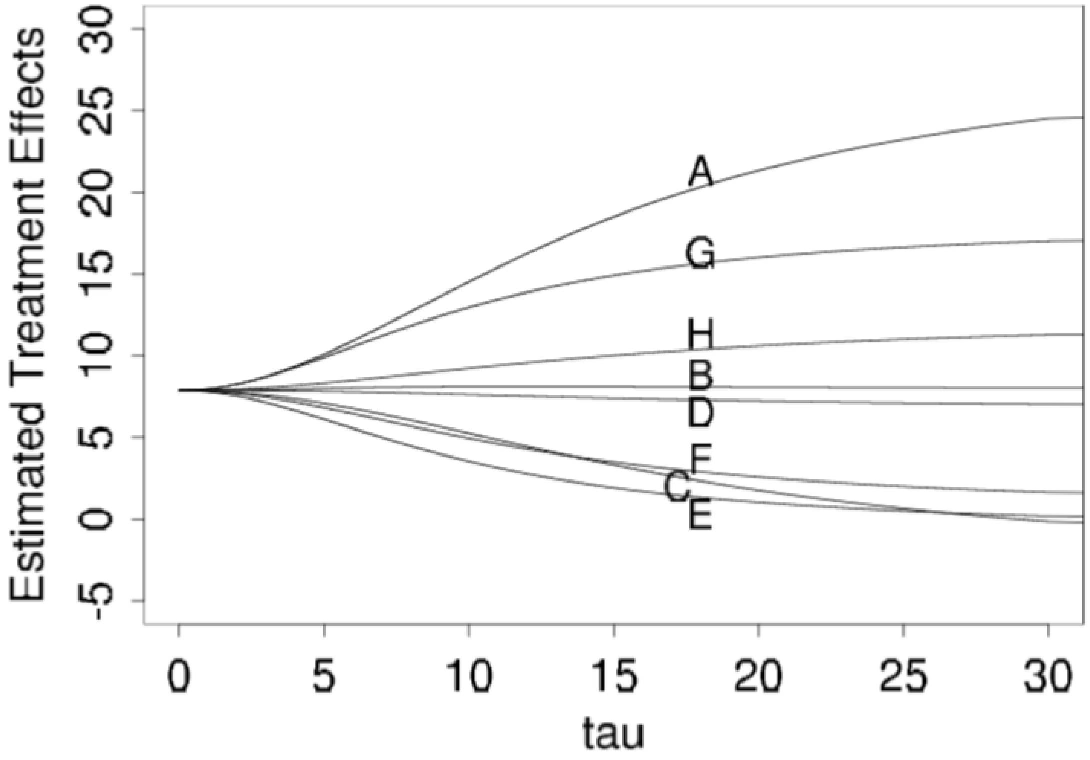


Figure 5.6 Conditional posterior means of treatment effects $E(\theta_j|\tau, y)$

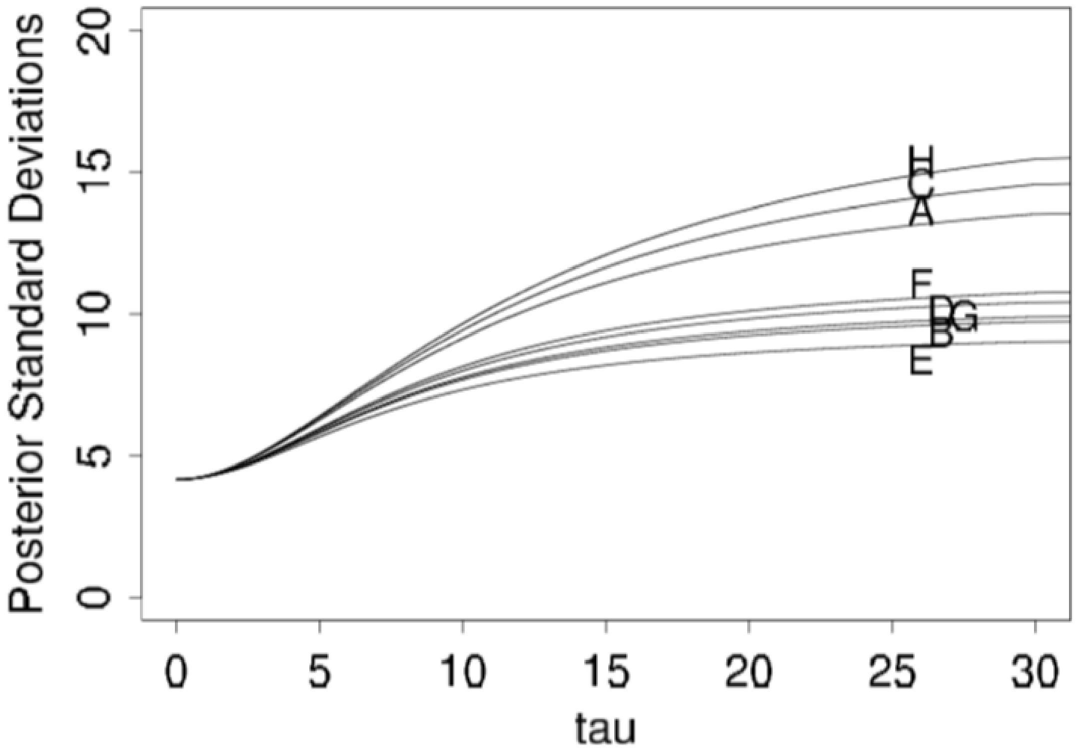


Figure 5.7 Conditional posterior standard deviations of treatment effects $sd(\theta_j|\tau, y)$

Detailed Procedure for Simulation

- There are two situations where the posterior predictive distribution may need to be calculated. These can be fit into the simulations already done

1. $\tilde{y}$ from a group j already observed.
   - Sample $\tilde{y}_{j,k}$ from $N(\theta_{j,k}, \sigma^2)$
   - If m observation are needed draw m values of $\tilde{y}$ from the above distribution.

2. $\tilde{y}$ from a new group $\tilde{j}$.
   - Sample $\theta_{\tilde{j},k}$ from $N(\theta|\mu_k, \tau_k^2)$ (draw from prior for θ, not the posterior)
   - Sample $\tilde{y}_{\tilde{j},k}$ from $N(\theta_{\tilde{j},k}, \sigma^2)$. Similarly to above if m samples are needed.

- The key difference is do we need to draw a new θ or use one we already have. The second situation will lead to more variable samples as there is less information about the corresponding θ in this case.

Summary

Key Points for Today

- Empirical Bayesian vs Full Bayesian
- Exchangeability
  - ✓ Assumption, not necessarily i.i.d., conditional independent is fine.
- Inference
  - ✓ Goal: posterior, prediction for an existing / a new group
  - ✓ To achieve the goal above, we derive joint / marginal / conditional posterior.
  - ✓ Choose a suitable prior for hyper-parameter.
  - ✓ Simulation procedure

Inference of interest

The posterior

$$p(\vec{\theta}|y) \longrightarrow p(\vec{\theta},\phi|y) \longrightarrow p(\vec{\theta},\phi) = p(\phi)p(\vec{\theta}|\phi)$$

Posterior predictive distributions. There are two situations of interest:

1. $\tilde{y}$ for an existing θ_j
2. $\tilde{y}$ for a new θ_j

Fully Bayesian Analysis of Conjugate Hierarchical Models

Three steps for analytical analysis:

1. Write the joint posterior density, $p(\theta, \phi|y)$, in unnormalized form as a product of the hyperprior distribution $p(\phi)$, the population distribution $p(\theta|\phi)$, and the likelihood $p(y|\theta)$.
2. Determine analytically the conditional posterior density of θ given the hyperparameters ϕ; for fixed observed y, this is a function of ϕ, $p(\theta|\phi, y)$.
3. Estimate ϕ using the Bayesian paradigm; that is, obtain its marginal posterior distribution, $p(\phi|y)$.

Inverse step 3 and 2 to draw samples from the joint posterior

Two ways to get marginal posterior:

1. Bruce force integration:

$$p(\phi|y) = \int p(\theta, \phi|y) d\theta.$$

2. Conditional probability formula:

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}.$$