# Hierarchical models

InSung Kong

October 20, 2020

Seoul National University

# Table of Contents

# Table of Contents

# Exchangeability

- The samples $(y_1, ..., y_n)$ are **exchangeable** if $p(y_1, ..., y_n)$ is invariant to permutations of the index $(1, ..., n)$
- **Exchangeability** is weaker condition than i.i.d(identical, independent distribution)
- Example : Sampling without replacement is not i.i.d, but exchangeable.

# Exchangeability

- The parameters $(\theta_1, ..., \theta_J)$ are **exchangeable** if $p(\theta_1, ..., \theta_J)$ is invariant to permutations of the index $(1, ..., J)$

- Example : $X \sim \eta_1 N(\mu_1, \sigma_1^2) + \eta_2 N(\mu_2, \sigma_2^2) + \eta_3 N(\mu_3, \sigma_3^2)$, where $\eta_1 + \eta_2 + \eta_3 = 1$

# Table of Contents

# Why hierarchical model?

## What will you answer to this question?

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

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).*

- The tumor probabilities $\theta$ vary because of differences in rats and experimental conditions among the experiments.
- What's tumor probabilities of 71's experiments' conditions??

# Why hierarchical model?

## Frequentist

- Frequentist 1 : $\theta_{71} = \frac{4}{14}$, since all experiments were done at different environment

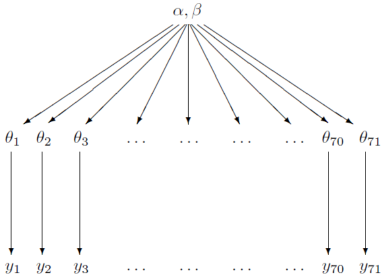
- Frequentist 2 : $\theta_{71} = \frac{\sum y_j}{\sum n_j}$, since all experiments are same essentially

## Frequentist - problem

- Too extreme(Too strong assumption)

# Why hierarchical model?

## Half Bayesian




- Assume $y_j \sim Bin(n_j, \theta_j)$, and $\theta_j$ are unknown. i.e. random.
- Make prior distribution with $(n_j, y_j), j = 1, ..., 70$. Then, update $\theta$'s distribution with $y_{71}$

# Why hierarchical model?

## Half Bayesian

- Because beta distribution is conjugate prior, assume $\theta \sim beta(\alpha, \beta)$.

- Since observed sample mean and standard deviation of the 70 values $\frac{y_j}{n_j}$ are 0.136 and 0.103, we can estimate $(\hat{\alpha}, \hat{\beta}) = (1.4, 8.6)$

- $y_{71} = 4$ and $n_j = 14$, so posterior distribution about $\theta_{71}$ become beta(5.4, 18.6)

# Why hierarchical model?

## Half Bayesian - problem

- If we want inference about first 70 experiments, data would be used twice
- The point estimate for $\alpha$ and $\beta$ seems dogmatic, and using any point estimate for $\alpha$ and $\beta$ necessarily ignores some posterior uncertainty.

The analysis using the data to estimate the prior parameters, called **empirical Bayes**, can be viewed as an approximation to the **complete hierarchical Bayesian** analysis.

# Table of Contents

# Hierarchical model - General

Observation y, parameter $\theta$, hyperparameter $\phi$

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 $\theta$ given the hyperparameters $\phi$; for fixed observed y, this is a function of $\phi$, $p(\theta|\phi, y)$.

3. Estimate $\phi$ using the Bayesian paradigm; that is, obtain its marginal posterior distribution, $p(\phi|y)$.

# Hierarchical model - General

## Drawing simulations from the posterior distribution

4. Draw $\phi$ from $p(\phi|y)$.
   - If $\phi$ is low-dimensional, the methods discussed in Chapter 3 can be used
   - If $\phi$ is high-dimensional, more sophisticated methods such as described in Part III may be needed.

5. Draw $\theta$ from $p(\theta|\phi, y)$

6. If desired, draw predictive values $\tilde{y}$ from the posterior predictive distribution given the drawn $\theta$.

# Table of Contents

# Hierarchical model - Binomial model

Complete Hierarchical Bayesian

- Assume $y_j \sim Bin(n_j, \theta_j)$, and $\theta_j \sim Beta(\alpha, \beta)$, and $\theta, \alpha, \beta$ are all random.

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

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

# Hierarchical model - Binomial model

## Complete Hierarchical Bayesian

- Since $\frac{p(\theta,\alpha,\beta|y)}{p(\theta|\alpha,\beta,y)} = \frac{p(\theta,\alpha,\beta,y)}{p(y)}\frac{p(\alpha,\beta,y)}{p(\theta,\alpha,\beta,y)}$,

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

- Choose hyperprior : noninformative & proper posterior. e.g.

$$p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$$

# Hierarchical model - Binomial model

## Drawing simulations from the posterior distribution

4. Draw $\phi$ from $p(\phi|y)$.
   - If $\phi$ is low-dimensional, the methods discussed in Chapter 3 can be used
   - If $\phi$ is high-dimensional, more sophisticated methods such as described in Part III may be needed.

5. Draw $\theta$ from $p(\theta|\phi, y)$

6. If desired, draw predictive values $\tilde{y}$ from the posterior predictive distribution given the drawn $\theta$.

# Table of Contents

# Hierarchical model - Normal model

## The data structure

- With known $n_j$, $\sigma^2$ and unknown(=random) $\theta_j$,

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

- Let $\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$ and $\sigma_j^2 = \sigma^2 / n_j$. Then

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

- Let $\bar{y}_{..} = \frac{\sum_{j=1}^{J} n_j \bar{y}_{.j}}{\sum_{j=1}^{J} n_j}$

# Hierarchical model - Normal model

Frequentist

- For estimate $\theta_{J+1}$, there are two choices
  1. $\hat{\theta}_{J+1} = \bar{y}_{.j+1}$
  2. $\hat{\theta}_{J+1} = \bar{y}_{..}$

- Using ANOVA(analysis of variance), we can decide which estimate to use.

# Hierarchical model - Normal model

## Hierarchical model

- We assume $\theta_j$ are drawn from a normal distribution $N(\mu, \tau)$:

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

- We assign a noninformative uniform hyperprior distribution to $\mu$, given $\tau$:

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

# Hierarchical model - Normal model

The joint posterior distribution

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. ...
3. ...

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

# Hierarchical model - Normal model

The conditional posterior distribution of the normal means, given the hyperparameters

1. ...
2. Determine analytically the conditional posterior density of $\theta$ given the hyperparameters $\phi$; for fixed observed y, this is a function of $\phi$, $p(\theta|\phi, y)$.
3. ...

Since $\bar{y}_{.j} \sim \mathrm{N}\left(\theta_j, \sigma_j^2\right)$ and $\theta_j \sim N(\mu, \tau)$

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

where

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

# Hierarchical model - Normal model

The marginal posterior distribution of the hyperparameters

1. ...
2. ...
3. Estimate $\phi$ using the Bayesian paradigm; that is, obtain its marginal posterior distribution, $p(\phi|y)$.

Since $p(\mu, \tau \mid y) \propto p(\mu, \tau) p(y \mid \mu, \tau)$
and $\bar{y}_{.j} \mid \mu, \tau \sim \mathrm{N}\left(\mu, \sigma_j^2 + \tau^2\right)$,

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

# Hierarchical model - Normal model

## Drawing simulations from the posterior distribution

4. Draw $\phi$ from $p(\phi|y)$.
   - If $\phi$ is low-dimensional, the methods discussed in Chapter 3 can be used
   - If $\phi$ is high-dimensional, more sophisticated methods such as described in Part III may be needed.
5. Draw $\theta$ from $p(\theta|\phi, y)$
6. If desired, draw predictive values $\tilde{y}$ from the posterior predictive distribution given the drawn $\theta$.

- At binomial model, we get $p(\alpha, \beta|y) \propto$ ..., but we had to do some complex work to get sample from posterior distribution

- (Unfortunately), we can do something more in normal model

# Hierarchical model - Normal model

Drawing simulations from the posterior distribution

From

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

, assume $\tau$ is known and $p(\mu \mid \tau) \propto 1$ We can find that

$$\mu \mid \tau, y \sim \mathrm{N}\left(\hat{\mu}, V_\mu\right)$$

where

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

# Hierarchical model - Normal model

Drawing simulations from the posterior distribution

So,

$$p(\tau \mid y) = \frac{p(\mu, \tau \mid y)}{p(\mu \mid \tau, y)}$$
$$\propto \frac{p(\tau) \prod_{j=1}^{J} \mathrm{N}\left(\bar{y}_{.j} \mid \mu, \sigma_j^2 + \tau^2\right)}{\mathrm{N}\left(\mu \mid \hat{\mu}, V_\mu\right)}$$

and this identity must hold for any value of $\mu$. So let set $\mu$ to $\hat{\mu}$.

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

## Drawing simulations from the posterior distribution

4. Draw $\phi$ from $p(\phi|y)$
   - Simulating $\tau$ using inverse cdf method(section 1.9), with

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

   - Simulating $\mu$ with

$$\mu \mid \tau, y \sim \mathrm{N}\left(\hat{\mu}, V_\mu\right)$$

where

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

# Hierarchical model - Normal model

Drawing simulations from the posterior distribution

5. Draw $\theta$ from $p(\theta|\phi, y)$
   - simulating $\theta$ using

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

where

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

6. If desired, draw predictive values $\tilde{y}$ from the posterior predictive distribution given the drawn $\theta$.