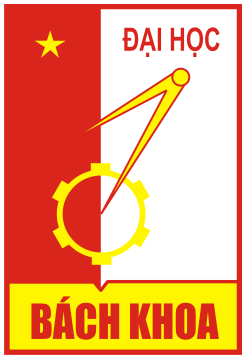
ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN - TIN



MÔ HÌNH PHÂN CẤP CHO BÀI TOÁN SO SÁNH NHIỀU GIÁ TRỊ TRUNG BÌNH

ĐỒ ÁN II

Chuyên ngành: TOÁN ỨNG DỤNG

Chuyên sâu: Các phương pháp ngẫu nhiên

Giảng viên hướng dẫn: **TS. ĐỖ VĂN CƯỜNG**

Sinh viên thực hiện: **TẠ DUY HẢI**

MSSV: **20206197**

Lớp: **CTTN Toán Tin - K65**

HÀ NỘI – 2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

(a) Mục tiêu: .....

. . . . . .

. . . . . .

. . . . . .

(b) Nội dung: .....

. . . . . .

. . . . . .

. . . . . .

2. Kết quả đạt được: .....

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

3. Ý thức làm việc của sinh viên: .....

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Hà Nội, ngày tháng năm 2023

Giảng viên hướng dẫn

TS. Đỗ Văn Cường

Mục lục

Bảng ký hiệu và chữ viết tắt

X	Biến ngẫu nhiên
S_X	Không gian mẫu của biến ngẫu nhiên X
$\mathbf{D}$	Mẫu ngẫu nhiên
$\mathbf{d}$	Mẫu
$i.i.d$	Độc lập cùng phân phối
$p(X_1, ..., X_n)$	Hàm mật độ đồng thời
$E(X)$	Kỳ vọng của biến ngẫu nhiên X
$V(X)$	Phương sai của biến ngẫu nhiên X
$p(X)$	Hàm khối xác suất hoặc hàm mật độ xác suất
MCMC	Markov chain Monte Carlo

Danh sách bảng

Danh sách hình vẽ

Mở đầu

Bài toán so sánh giá trị trung bình là một vấn đề quan trọng trong lĩnh vực phân tích thống kê. Sự quan tâm đặt ra từ nhu cầu thực tế trong việc đánh giá sự khác biệt giữa các nhóm hoặc điều kiện khác nhau, mang lại cái nhìn sâu sắc về tính chất và xu hướng của dữ liệu. Việc so sánh giá trị trung bình không chỉ giúp chúng ta hiểu rõ hơn về sự đa dạng trong mẫu dữ liệu, mà còn cung cấp cơ sở khoa học để đưa ra các quyết định trong nhiều lĩnh vực ứng dụng.

Đối tượng nghiên cứu trong báo cáo tập trung vào dữ liệu điểm thi của học sinh tại một trường trung học phổ thông trên địa bàn Hà Nội. Phạm vi nghiên cứu sẽ được định hình để xây dựng một mô hình so sánh giữa các nhóm lớp học sinh, với sự tập trung vào việc hiểu rõ hơn về sự biến động và khác biệt trong kết quả học tập của các đối tượng nghiên cứu thông qua bài toán kiểm định. Báo cáo trình bày hai cách tiếp cận của hai trường phái Frequentist và Bayesian tương ứng với hai mô hình ANOVA cổ điển và ANOVA phân cấp.

Nghiên cứu cũng có ý nghĩa thực tiễn khi cung cấp thông tin hữu ích cho học sinh và phụ huynh. Bằng cách hiểu rõ hơn về sự chênh lệch giữa các nhóm học sinh, những người quan tâm đến giáo dục có thể đưa ra quyết định thông minh hơn về hướng nghiệp và hỗ trợ cá nhân hóa quá trình học tập. Ngoài ra, đối với những người quản lý và quyết định chính sách giáo dục, mô hình so sánh này có thể là nguồn thông tin quý báu, giúp họ hiểu rõ hơn về cơ hội và thách thức trong việc cải thiện chất lượng giáo dục. Từ đó, những quyết định lớn hơn về chính sách giáo dục có thể được đưa ra dựa trên cơ sở dữ liệu và kết quả của mô hình.

Lời cảm ơn

Báo cáo này được thực hiện và hoàn thành tại Đại học Bách Khoa Hà Nội, nằm trong nội dung học phần Đồ Án II. Em xin được gửi lời cảm ơn chân thành tới người thầy đã hướng dẫn em - TS. Đỗ Văn Cường. Sự kiện em và thầy làm việc với nhau xảy ra một cách ngẫu nhiên, nhưng em thấy rất vui và thích thú với chủ đề thầy gợi ý. Cảm ơn thầy đã cho em cơ hội thực hiện đề tài này cùng những lời khuyên và hướng dẫn trong suốt quá trình làm đồ án. Những kiến thức mà thầy truyền đạt đã giúp em có thêm góc nhìn trong lĩnh vực phân tích thống kê. Chúc thầy có thật nhiều sức khỏe, luôn hạnh phúc, gặt hái được nhiều thành công trong lĩnh vực giảng dạy và nghiên cứu.

Em cũng muốn gửi lời cảm ơn đến TS. Nguyễn Phương Thùy - giảng viên chủ nhiệm của lớp em. Cô đã tận tình hướng dẫn em trong suốt quá trình học tập. Những chỉ bảo, hướng dẫn của cô đã giúp báo cáo này hoàn thiện hơn. Cảm ơn cô đã giúp thời sinh viên của em có nhiều kỷ niệm đáng nhớ. Hy vọng em và cô sớm có cơ hội gặp lại nhau trong tương lai.

Cuối cùng, em xin gửi lời cảm ơn đến tập thể lớp CTTN- Toán Tin- K65, những người bạn đã đồng hành với em từ những ngày đầu tiên bước chân vào đại học. Nhờ các bạn mà quá trình học tập của em bớt tẻ nhạt, những kiến thức học hỏi từ các bạn đã giúp đỡ em rất nhiều. Chúc các bạn thành công với con đường đã chọn.

Em xin trân trọng cảm ơn!

Hà Nội, ngày tháng năm 2024

Tác giả đồ án

Tạ Duy Hải

Chương 1

Kiến thức chuẩn bị

Chương này trình bày một số khái niệm cơ bản về xác suất và quá trình ngẫu nhiên trong [1], [3], [4], [5], [7].

1.1 Độc lập có điều kiện

Định nghĩa 1.1.1 *Cho vector tham số Θ và dãy các biến ngẫu nhiên $X_1,...,X_n$. Ta nói dãy biến $X_1,...,X_n$ là độc lập có điều kiện với Θ nếu với mọi tập $\{B_1,...,B_n\}$ ta có*

$$P\left(X_1 \in B_1,...,X_n \in B_n|\Theta\right) = P\left(X_1 \in B_1|\Theta\right) \times \cdots \times P\left(X_n \in B_n|\Theta\right).$$

Với dãy $X_1,...,X_n$ là biến ngẫu nhiên liên tục, ta có thể viết dưới dạng

$$\begin{aligned} p\left(x_1,...,x_n|\Theta\right) &= p_{X_1}\left(x_1|\Theta\right) \times \cdots \times p_{X_n}\left(x_n|\Theta\right) \\ &= \prod_{i=1}^{n} p_{X_i}(x_i|\Theta). \end{aligned}$$

Nếu dãy các biến $X_1,...,X_n$ đều là kết quả của cùng một quá trình hay đều là kết quả của một phép thử dưới các điều kiện giống nhau thì hàm mật độ xác suất của các biến là đồng nhất, ta có thể viết

$$p\left(x_1,...,x_n|\Theta\right) = \prod_{i=1}^{n} p\left(x_i|\Theta\right).$$

Trong trường hợp này, dãy biến $X_1,...,X_n$ gọi là độc lập cùng phân phối có điều kiện với Θ. Để ngắn gọn, ta dùng ký hiệu

$$X_1,...,X_n|\Theta \sim \text{i.i.d.} \quad p(x|\Theta)$$

1.2 Một vài phân phối thông dụng

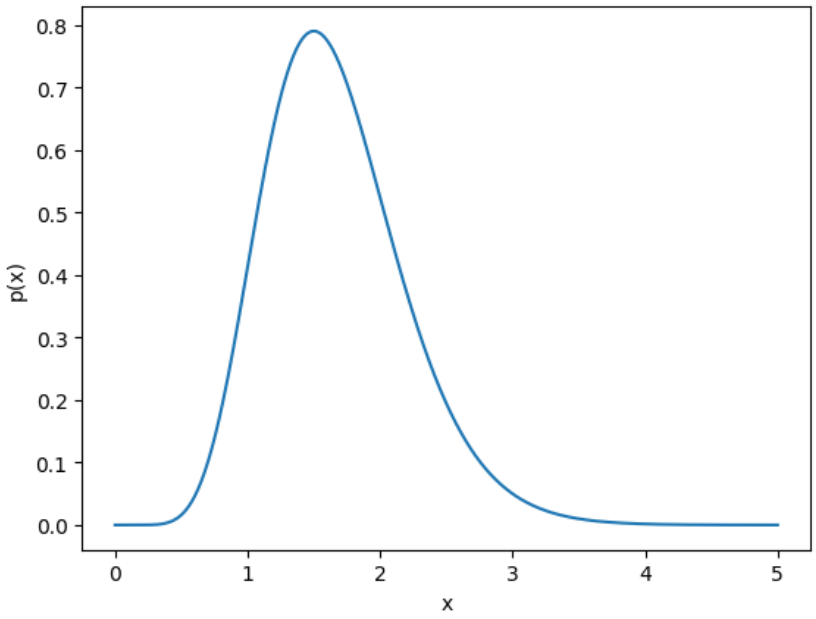
1.2.1 Phân phối Gamma

Định nghĩa 1.2.1 *Biến ngẫu nhiên X được gọi là tuân theo phân phối Gamma với hai tham số dương* α, β *nếu nó có hàm mật độ xác suất*

$$p(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), x > 0, \alpha, \beta > 0.$$

Ký hiệu $X \sim \text{Gamma}(\alpha, \beta)$, kỳ vọng và phương sai của X như sau:

$$E(X) = \frac{\alpha}{\beta},$$
$$V(X) = \frac{\alpha}{\beta^2}.$$



Hình 1.1: Đồ thị hàm mật độ của $X \sim \text{Gamma}(10, 6)$.

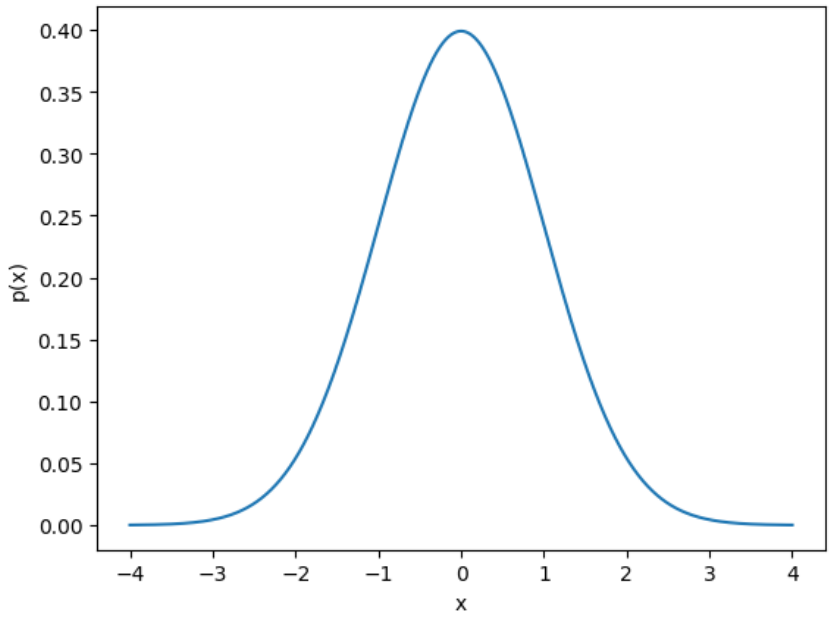
1.2.2 Phân phối chuẩn

Định nghĩa 1.2.2 *Biến ngẫu nhiên liên tục X được gọi là tuân theo phân phối chuẩn với tham số* μ, σ^2 *nếu nó có hàm mật độ xác suất*

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2, \ x \in \mathbb{R}.$$

Ký hiệu $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, kỳ vọng và phương sai của X là:

$$E\left(X\right) = \mu,$$
$$V\left(X\right) = \sigma^2.$$



Hình 1.2: Đồ thị hàm mật độ của $X \sim \mathcal{N}\left(0,1\right)$.

1.2.3 Phân phối khi bình phương

Định nghĩa 1.2.3 *Biến ngẫu nhiên X được gọi là tuân theo phân phối khi bình phương với n bậc tự do, nếu X có hàm mật độ xác suất*

$$p\left(x\right) = \frac{\left(\frac{x}{2}\right)^{\frac{n}{2}-1}}{2\Gamma\left(\frac{n}{2}\right)} \exp -\frac{x}{2},\ x > 0,\ n \in \mathbb{N}^*.$$

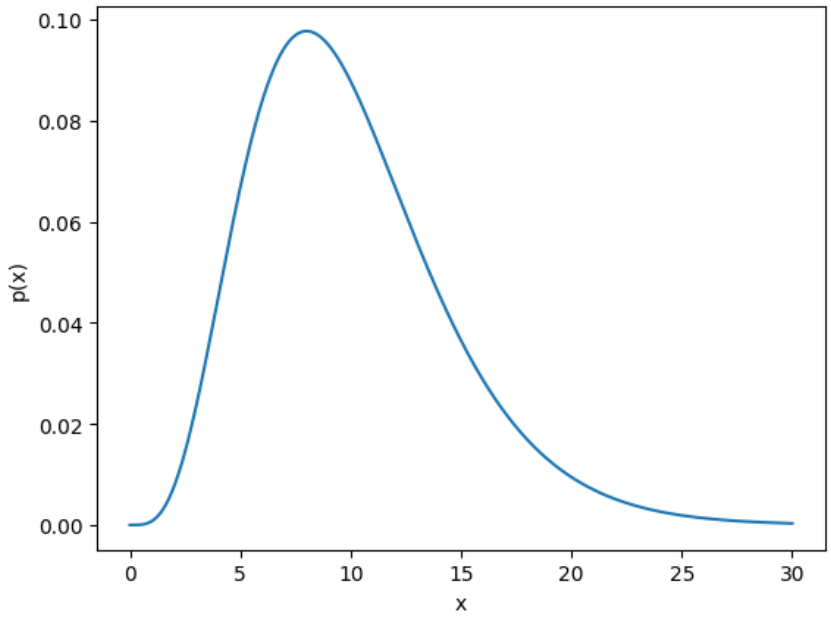
Ký hiệu $X \sim \chi^2_n$. Kỳ vọng và phương sai của X:

$$E\left(X\right) = n,$$
$$V\left(X\right) = 2n.$$

Định nghĩa 1.2.4 *Nếu dãy biến $X_1, X_2, \ldots, X_n$ độc lập cùng phân phối chuẩn $\mathcal{N}\left(0,1\right)$ thì*

$$U_n = \sum_{i=1}^{n} X_i^2$$

có phân phối khi bình phương với n bậc tự do, hay $U_n \sim \chi^2_n$.



Hình 1.3: Đồ thị hàm mật độ của $X \sim \chi^2_{10}$.

1.2.4 Phân phối Fisher

Định nghĩa 1.2.5 *Biến ngẫu nhiên X tuân theo phân phối Fisher với (n,m) bậc tự do nếu hàm mật độ xác suất có dạng*

$$p(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)\left(\frac{n}{m}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{x^{\frac{n}{2}-1}}{\left(1+\frac{n}{m}x\right)^{\frac{n+m}{2}}},\ x>0, m,n \in \mathbb{N}^*.$$

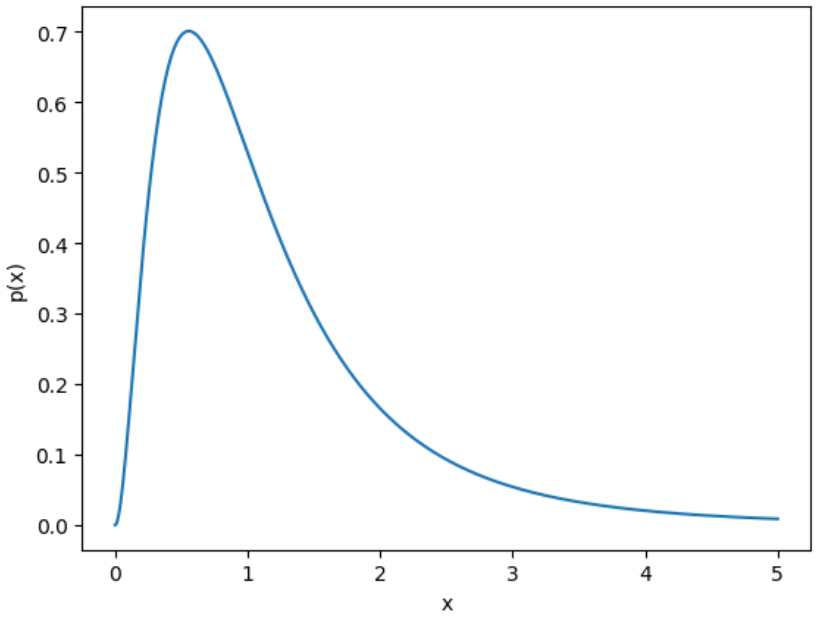
Ký hiệu $X \sim F(n,m)$, kỳ vọng và phương sai của X là:

$$E(X) = \frac{m}{m-2}, m>2,$$
$$V(X) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}, m>4.$$

Định nghĩa 1.2.6 *Cho hai biến ngẫu nhiên độc lập $X \sim \chi^2_n$ và $Y \sim \chi^2_m$, biến ngẫu nhiên $F_{n,m}$ định nghĩa bởi*

$$F_{n,m} = \frac{X/n}{Y/m}$$

tuân theo phân phối Fisher với (n,m) bậc tự do, hay $F_{n,m} \sim F(n,m)$.



Hình 1.4: Đồ thị hàm mật độ $X \sim F_{6,10}$.

1.3 Định lý Bayes

Định lý 1.3.1 *Gọi* $\Omega = \{\boldsymbol{d} = (x_1, x_2, ..., x_n)\}$ *là không gian mẫu.* $\mathbb{P} = \left\{P_{\Theta,\boldsymbol{D}} | \Theta \in S_\Theta\right\}$ *là một mô hình với tham số* Θ *và mẫu ngẫu nhiên* $\boldsymbol{D}$*. Kí hiệu* $\pi(\Theta)$ *là phân phối tiên nghiệm của* Θ*,* $f(\boldsymbol{D}|\Theta)$ *là hàm mật độ xác suất của* $\boldsymbol{D}$*,* $p(\Theta|\boldsymbol{D})$ *là phân hối hậu nghiệm của* Θ *khi có* $\boldsymbol{D}$*. Định lý Bayes phát biểu rằng*

$$p(\Theta|\boldsymbol{D}) = \frac{\pi(\Theta) \times f(\boldsymbol{D}|\Theta)}{\int_{S_\Theta} \pi(\Theta) \times f(\boldsymbol{D}|\Theta)\, d\Theta}.$$

$f(\mathbf{D}|\Theta)$ được gọi là hàm hợp lý, phần mẫu số được gọi là hằng số chuẩn hóa. Định lý thường viết dưới dạng

$$p(\Theta|\mathbf{D}) \propto \pi(\Theta) \times f(\mathbf{D}|\Theta).$$

1.4 Xích Markov và phương pháp lấy mẫu Gibbs

1.4.1 Xích Markov

Một cách tổng quan, xích Markov hay quá trình Markov là một quá trình ngẫu nhiên mô tả chuỗi các sự trạng thái mà nếu biết trạng thái hiện tại, thì trạng thái

trong quá khứ và trạng thái ở tương lai là độc lập với nhau. Có nhiều loại xích hay quá trình Markov, trong phần này, tác giả giới thiệu về `xích Markov thuần nhất với không gian trạng thái liên tục`. Để tiện trình bày, tác giả viết tắt là xích Markov.

Định nghĩa 1.4.1 *Cho S_X là không gian trạng thái liên tục. Dãy biến ngẫu nhiên $\{X_n\}_{n\in\mathbb{N}}$ được gọi là xích Markov nếu*

$$\begin{aligned} &P\left(X_n \in A_n | X_{n-1} \in A_{n-1}, X_{n-2} \in A_{n-2}, \ldots, X_0 \in A_0\right) \\ &= P\left(X_n \in A_n | X_{n-1} \in A_{n-1}\right). \end{aligned}$$

Trong đó $A_i \subset S_X$ với mọi $i = \overline{0,n}$.

Do không gian S_X là liên tục, $P(x,y) = P(X_n = y | X_{n-1} = x) = 0$ nên việc định nghĩa ma trận xác suất chuyển như trong không gian rời rạc là không khả thi. Thay vào đó, xích Markov sử dụng mật độ chuyển để đặc trưng khả năng chuyển trạng thái của xích.

Định nghĩa 1.4.2 *Xác suất chuyển $P(x,y)$ của xích Markov được định nghĩa bởi*

$$P(x,y) = P(X_n \leq y | X_{n-1} = x) = P\left(X_1 \leq y | X_0 = x\right), x,\ y \in S_X.$$

Mật độ chuyển của xích xác định bởi công thức

$$p(x,y) = \frac{\partial P(x,y)}{\partial y}, x,\ y \in S_X.$$

Với mỗi thời điểm $n \in \mathbb{N}$, xích có thể nhận bất cứ giá trị trong S_X. Với không gian trạng thái S_X rời rạc, ta dùng phân phối $\lambda_n = \left\{\lambda_n^i\right\}_{i\in S_X}$ để đặc trưng cho khả năng mà xích ở trạng thái i nào đó tại thời điểm n. Tương tự với S_X liên tục, ta sử dụng mật độ $\lambda_n = \lambda_n(x)$.

Định nghĩa 1.4.3 *Xích Markov tại thời điểm n, $n \in \mathbb{N}$ có mật độ $\lambda_n(x)$ với $x \in S_X$ thỏa mãn*

$$\begin{aligned} &\lambda_n(x) \geq 0, \\ &\int_{S_X} \lambda_n(x)\, dx = 1. \end{aligned}$$

Với mật độ chuyển $p(x,y)$ và mật độ $\lambda_n(x)$ tại thời điểm n, mật độ của xích tại thời điểm $n+1$ xác định bởi

$$\lambda_{n+1}(x) = \int_{S_X} p(y,x)\,\lambda_n(y)\,dy, x \in S_X.$$

Nhìn chung với mỗi thời điểm n, mật độ của xích là khác nhau. Nếu mật độ của xích không đổi theo thời gian, ta gọi đó là mật độ dừng hay mật độ cân bằng.

Định nghĩa 1.4.4 *Xích Markov với mật độ chuyển $p(x,y)$ nhận $\lambda(x)$ với $x \in S_X$ là mật độ dừng, nếu điều sau thỏa mãn*

$$\lambda(y) = \int_{S_X} \lambda(x)\,p(x,y)\,dx.$$

1.4.2 Phương pháp lấy mẫu Gibbs

Lấy mẫu Gibbs là một thuật toán Markov chain Monte Carlo (MCMC) dùng để thu thập chuỗi các quan sát sao cho chuỗi "xấp xỉ" một phân phối xác suất khó để lấy mẫu trực tiếp.

Định lý 1.4.1 (xem [1]) *Cho không gian trạng thái $S_{\boldsymbol{X}}$ và một phân phối mà ta quan tâm đặc trưng bởi hàm mật độ $\lambda(\boldsymbol{x})$ với $\boldsymbol{x} \in S_{\boldsymbol{X}}$, $\boldsymbol{x} = (x_1,\ldots,x_n)$. Xét xích Markov $\{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$ với mật độ xác suất chuyển cho bởi*

$$p(\boldsymbol{x},\boldsymbol{y}) = \prod_{i=1}^{n} \lambda\left(y_i | x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_n\right).$$

Thì $\lambda(\boldsymbol{x})$ là mật độ dừng của xích $\{\boldsymbol{X}_n\}_{n\in\mathbb{N}}$.

Chứng minh. Xét tích phân

$$\begin{aligned}
&\int_{S_{\mathbf{X}}} \prod_{i=1}^{n} \lambda\left(y_i | y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_n\right) \lambda(\mathbf{x})\,d\mathbf{x} \\
&= \int_{S_{x_1}} \cdots \int_{S_{x_n}} \prod_{i=1}^{n} \lambda\left(y_i | y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_n\right) \lambda(\mathbf{x})\,dx_1\ldots dx_n \\
&= \int_{S_{x_2}} \cdots \int_{S_{x_n}} \prod_{i=1}^{n} \lambda\left(y_i | y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_n\right) \lambda(x_2,\ldots,x_n)\,dx_2\ldots dx_n \\
&= \int_{S_{x_2}} \cdots \int_{S_{x_n}} \prod_{i=2}^{n} \lambda\left(y_i | y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_n\right) \lambda(y_1,x_2,\ldots,x_n)\,dx_2\ldots dx_n
\end{aligned}$$

$$= \int_{S_{x_3}} \cdots \int_{S_{x_n}} \prod_{i=3}^{n} \lambda\left(y_i | y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_n\right) \lambda\left(y_1, y_2, \ldots, x_n\right) dx_3 \ldots dx_n$$

$$\ldots$$

$$= \int_{S_{x_n}} \lambda\left(y_n | y_1, \ldots, y_{n-2}, y_{n-1}\right) \lambda\left(y_1, \ldots, y_{n-2}, y_{n-1}, x_n\right) dx_n$$

$$= \lambda\left(y_n | y_1, \ldots, y_{n-2}, y_{n-1}\right) \lambda\left(y_1, \ldots, y_{n-2}, y_{n-1}\right)$$

$$= \lambda\left(y_1, \ldots, y_n\right) = \lambda\left(\mathbf{y}\right).$$

Vậy $\lambda\left(\mathbf{x}\right)$ là mật độ dừng của xích. □

Quay lại với phương pháp lấy mẫu Gibbs, giả sử phân phối mà ta muốn lấy mẫu có hàm mật độ $\lambda\left(\mathbf{x}\right)$ phức tạp với $\mathbf{x} \in S_{\mathbf{X}}$. Nếu hàm mật độ có điều kiện $\lambda\left(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_n\right)$ lấy mẫu dễ dàng, ta có thể sử dụng thuật toán Gibbs theo các bước sau

- Bước 1: Chọn giá trị ban đầu bất kỳ $\mathbf{y}_1$, khởi tạo biến đếm $j = 1$.
- Bước 2: Tạo mẫu mới $\mathbf{y}_{j+1}$ bằng cách lấy mẫu từng thành phần

$$\begin{aligned}
y_1^{(j+1)} &\sim \lambda\left(y_1 | y_2^{(j)}, \ldots, y_n^{(j)}\right), \\
y_2^{(j+1)} &\sim \lambda\left(y_2 | y_1^{(j+1)}, y_2^{(j)} \ldots, y_n^{(j)}\right), \\
&\vdots \\
y_n^{(j+1)} &\sim \lambda\left(y_n | y_1^{(j+1)}, \ldots, y_{n-1}^{(j+1)}\right).
\end{aligned}$$

- Bước 3: cho $j = j + 1$ và quay lại bước 2 cho tới khi số vòng lặp đủ lớn.

Xích Markov mà ta xây dựng từ thuật toán Gibbs có mật độ dừng chính là mật độ $\lambda\left(\mathbf{x}\right)$ mà ta đã chứng minh ở định lý 1.4.1. Số vòng lặp đủ lớn được xác định tùy vào yêu cầu của bài toán cần giải quyết cũng như quá trình thực nghiệm.

1.5 Xấp xỉ tích phân

Đặc điểm nổi bật của biến ngẫu nhiên là phân phối xác suất, đặc trưng đơn giản nhất từ phân phối là kỳ vọng và phương sai. Trong một số trường hợp, chỉ cần biết hai thông tin này mà không cần nắm rõ phân phối của biến ngẫu nhiên là gì. Vì vậy, việc xác định kỳ vọng và phương sai của biến ngẫu nhiên là vô cùng quan trọng. Xét

biến ngẫu nhiên X liên tục với hàm mật độ $p(x)$, bằng công cụ giải tích, kỳ vọng và phương sai xác định bởi

$$E(X) = \int_{S_X} xp(x)\,dx,$$
$$V(X) = \int_{S_X} (x - E(X))^2 p(x)\,dx.$$

Với các phân phối có dạng hàm phức tạp hoặc tập giá trị S_X là nhiều chiều, việc tính đúng các tích phân trên gặp nhiều khó khăn, thay vào đó các phương pháp xấp xỉ được ưu tiên sử dụng trong trường hợp này.

1.5.1 Xấp xỉ Monte Carlo

Biến ngẫu nhiên X có hàm $p(x)$ phức tạp, tuy nhiên lấy mẫu ngẫu nhiên từ $p(x)$ dễ dàng, sử dụng xấp xỉ Monte Carlo là lựa chọn hợp lý để tính gần đúng tích phân, dựa vào luật số lớn yếu.

Định lý 1.5.1 *Cho $X_1, \ldots, X_n$ là các biến ngẫu nhiên độc lập cùng phân phối với kỳ vọng μ và phương sai σ^2, khi đó biến ngẫu nhiên*

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

hội tụ theo xác suất tới μ khi n lớn hay

$$\lim_{n\to\infty} p(|\bar{X}_n - \mu| < \varepsilon) = 1$$

với ε dương bé tùy ý.

Như vậy khi n lớn, đa số các giá trị của biến ngẫu nhiên $\bar{X}$ sẽ thuộc đoạn $(\mu - \varepsilon,\ \mu + \varepsilon)$. Xét phương sai

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

khi n lớn thì $V(\bar{X})$ sẽ tiến về 0, hay các giá trị của biến $\bar{X}$ có độ tập trung cao quanh giá trị μ. Đây cũng là ý tưởng để thực hiện xấp xỉ Monte Carlo, xét dãy quan sát $x_1, \ldots, x_n$ được lấy mẫu độc lập với phân phối $p(x)$ thì

$$\int_{S_X} h(x)\,p(x)\,dx = E(h((X)) \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i),$$

với trung bình bình phương sai số là

$$MSE_{MC} = E\left(\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i) - E(h(X))\right)^2\right)$$
$$= V\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right) = \frac{V(h(X))}{n}.$$

1.5.2 Xấp xỉ Markov chain Monte Carlo

Trong trường hợp biến ngẫu nhiên X có dạng hàm $p(x)$ phức tạp và lấy mẫu ngẫu nhiên gặp khó khăn, phương pháp MCMC có thể xử dụng để xấp xỉ tích phân. Thay vì sử dụng các quan sát được lấy mẫu độc lập từ phân phối $p(x)$, MCMC sử dụng các quan sát là các trạng thái tạo bởi một xích Markov với phân phối (mật độ) dừng là $p(x)$ để xấp xỉ tích phân

$$\int_{S_X} h(x)\, p(x)\, dx = E(h((x)) \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i),$$

Không như xấp xỉ Monte Carlo, trung bình bình phương sai số MCMC xác định bởi

$$MSE_{MCMC} = E\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i) - E(h(X))\right)^2$$
$$= \frac{1}{n^2}E\left(\sum_{i=1}^{n} h(X_i) - nE(h(X))\right)^2$$
$$= MSE_{MC} + \frac{1}{n^2}\sum_{i\neq j} E\left[\left(h(X_i) - E(h(X))\right)\left(h(X_j) - E(h(X))\right)\right].$$

Số hạng thứ hai khác không do sự tương quan của các quan sát tạo bởi xích Markov và thường lớn hơn không nên sai số của MCMC lớn hơn xấp xỉ Monte Carlo. Như vậy với kích thước mẫu lớn trong MCMC không đảm bảo thu được xấp xỉ đủ tốt. Nếu xích có tính ergodic, điều này được đảm bảo.

Định nghĩa 1.5.1 *Xích Markov $\{X_n\}_{n\in\mathcal{N}}$ có tính chất ergodic với mật độ dừng $\lambda(x)$ nếu*

$$P\left(\lim_{t\to\infty}\frac{1}{t-1}\sum_{i=0}^{t} f(X_i) = \int_{S_X} f(x)\,\lambda(x)\,dx\right) = 1,$$

với f là hàm bị chặn bất kỳ.

Chương 2

Bài toán so sánh nhiều giá trị trung bình

Chương này trình bày bài toán so sánh nhiều nhóm thông qua giá trị trung bình, phương pháp ANOVA và mô hình phân cấp để giải quyết bài toán. Các kiến thức trong chương này lấy từ [2], [3], [6], [8].

2.1 Bài toán

Cho n nhóm, mỗi nhóm đều có thuộc tính η cần điều tra. Cho $\mathbf{x}_1, \ldots, \mathbf{x}_n$ là các mẫu đo thuộc tính η của từng cá thể được lấy ra từ các nhóm tương ứng với kích cỡ $n_1, n_2, \ldots, n_n$. Với dữ liệu này, hãy kiểm tra xem các nhóm có khác nhau không dựa trên thuộc tính η.

2.2 ANOVA một nhân tố

Phân tích phương sai (analysis of variance) viết tắt là ANOVA là lớp phương pháp thống kê dùng để kiểm tra sự khác biệt giữa các nhóm thông qua phương sai hay độ biến đổi của dữ liệu. ANOVA là một công cụ mạnh mẽ được sử dụng rộng rãi, phương pháp này xác định liệu sự khác biệt giữa các nhóm do thành phần nhiễu ngẫu nhiên hay do bản thân các nhóm có sự khác biệt. ANOVA có nhiều mô hình khác nhau, để giải bài toán vừa nêu, tác giả trình bày mô hình ANOVA một nhân tố. Một số giả định về dữ liệu khi áp dụng mô hình này gồm:

- Mẫu $\mathbf{x}_i$ có kích thước n_i được lấy từ nhóm i với kỳ vọng μ_i, $i = \overline{1,n}$. Các mẫu là độc lập với nhau.

- Phương sai của thuộc tính η trong mỗi nhóm đều bằng σ^2.
- Thuộc tính η trong mỗi nhóm đều tuân theo phân phối chuẩn.

Cặp giả thuyết - đối thuyết tương ứng với bài toán

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_n, \text{ với } H_1: \ \mu_j \neq \mu_i, \ \ i, j \text{ nào đó.} \tag{2.1}$$

Trong mô hình ANOVA một nhân tố, giá trị trung bình của nhóm i được biểu diễn dưới dạng

$$\mu_i = \mu + \tau_i. \tag{2.2}$$

Trong đó μ_i là trung bình nhóm i, μ là trung bình giữa các nhóm và τ_i là ảnh hưởng của nhóm i. Gọi X_{ji} là biến ngẫu nhiên thứ j của mẫu $\mathbf{X}_i$ lấy từ nhóm i có biểu diễn

$$X_{ji} = \mu + \tau_i + \varepsilon_{ji}. \tag{2.3}$$

Với biến ngẫu nhiên độc lập $\varepsilon_{ji} \sim \mathcal{N}\left(0, \sigma^2\right)$, do đó $X_{ji} \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$. Với mỗi quan sát x_{ji} tương ứng, có biểu diễn

$$x_{ji} = \overline{x} + \left(\overline{x}_i - \overline{x}\right) + \left(x_{ji} - \overline{x}_i\right). \tag{2.4}$$

Số hạng $\overline{x}$ gọi là trung bình mẫu giữa các nhóm, ước lượng của μ; số hạng thứ hai là ước lượng ảnh hưởng τ_i của nhóm thứ i và số hạng cuối là ước lượng của sai số ε_{ji}. Xét biểu thức (2.4)

$$\begin{aligned} x_{ji} &= \overline{x} + \left(\overline{x}_i - \overline{x}\right) + \left(x_{ji} - \overline{x}_i\right) \\ \Leftrightarrow x_{ji} - \overline{x} &= \left(\overline{x}_l - \overline{x}\right) + \left(x_{ji} - \overline{x}_i\right) \\ \Leftrightarrow \left(x_{ji} - \overline{x}\right)^2 &= \left(\overline{x}_l - \overline{x}\right)^2 + \left(x_{ji} - \overline{x}_i\right)^2 + 2\left(\overline{x}_l - \overline{x}\right)\left(x_{ji} - \overline{x}_l\right) \end{aligned}$$

Chú ý rằng $\sum_{j=1}^{n_i}\left(x_{ji} - \overline{x}_i\right) = 0$, lấy tổng các quan sát trong nhóm i

$$\sum_{j=1}^{n_i}\left(x_{ji} - \overline{x}\right)^2 = n_i\left(\overline{x}_i - \overline{x}\right)^2 + \sum_{j=1}^{n_i}\left(x_{ji} - \overline{x}_i\right)^2.$$

Tiếp tục lấy tổng các nhóm từ biểu thức trên

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(x_{ji} - \overline{x}\right)^2 &= \sum_{i=1}^{n} n_i\left(\overline{x}_i - \overline{x}\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(x_{ji} - \overline{x}_i\right)^2 \\ \Leftrightarrow \sum_{i=1}^{n}\sum_{j=1}^{n_i} x_{ji}^2 &= \overline{x}^2\sum_{i=1}^{n} n_i + \sum_{i=1}^{n} n_i\left(\overline{x}_i - \overline{x}\right)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(x_{ji} - \overline{x}_i\right)^2. \end{aligned}$$

Như vậy tổng bình phương các quan sát có thể chia ra làm 3 số hạng. Số hạng thứ nhất là tổng bình phương trung bình giữa các nhóm ss_{mean}, số hạng thứ hai là tổng bình phương ảnh hưởng của các nhóm ss_{tr}, số hạng cuối tổng bình phương sai số của toàn bộ quan sát ss_{res}.

Quay trở lại với cặp giả thuyết đối thuyết (2.1), với mẫu $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ đã có, tính được 3 số hạng trên một cách dễ dàng. Một cách trực quan, càng có cơ sở bác bỏ H_0 và chấp nhận H_1 nếu ss_{tr} càng lớn so với ss_{res}, hay tổng bình phương ảnh hưởng của các nhóm có ảnh hưởng đáng kể so với tổng bình phương sai số.

Định lý 2.2.1 (xem [8]) *Cho $X_1, \ldots, X_n$ là các biến ngẫu nhiên độc lập cùng phân phối $\mathcal{N}(\mu, \sigma^2)$. Khi đó biến ngẫu nhiên*

$$T = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

tuân theo phân phối khi bình phương với $n-1$ bậc tự do.

Chứng minh. Chọn ma trận trực giao $\mathbf{A}$, thỏa mãn

$$a_{11} = a_{12} = \cdots = a_{1n} = \frac{1}{\sqrt{n}}.$$

Xây dựng $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ là kết quả của ánh xạ tuyến tính trực giao $\mathbf{Y} = \mathbf{AX}$. Suy ra $Y_1 = \dfrac{1}{\sqrt{n}}(X_1 + \cdots + X_n) = \sqrt{n}\bar{X}$. Do $\mathbf{A}$ trực giao nên

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{X}^T\mathbf{A}^T\mathbf{AX} = \mathbf{X}^T\mathbf{X}.$$

Suy ra

$$\begin{aligned}
\sum_{i=1}^{n} Y_i^2 &= \sum_{i=1}^{n} X_i^2 \\
&= \sum_{i=1}^{n} (X_i - \bar{X} + \bar{X})^2 \\
&= \sum_{i=1}^{n} (X_i - \bar{X})^2 + n\bar{X}^2 \\
&= \sum_{i=1}^{n} (X_i - \bar{X})^2 + Y_1^2.
\end{aligned}$$

Điều này tương đương với

$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{2.5}$$

Xét

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu)^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2\\ &= \sum_{i=2}^{n}Y_i^2+n(\bar{X}-\mu)^2\\ &= \sum_{i=2}^{n}Y_i^2+\left(Y_1-\sqrt{n}\mu\right)^2.\end{aligned}$$

Hàm mật độ đồng thời của $\mathbf{X}=\{X_1,\ldots,X_n\}$ cho bởi

$$p(x_1,\ldots,x_n)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

Xác suất của $\mathbf{X}$ được tính dựa trên hàm mật độ đồng thời tính bởi

$$\begin{aligned}F(\mathbf{X}\in B)=P(\mathbf{X}\in B) &= \int_B p(\mathbf{x})\,d\mathbf{x}\\ &= \int_B\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]d\mathbf{x}.\end{aligned}$$

Trong đó $B\subset S_X$. Đổi sang biến $\mathbf{Y}$ với $\mathbf{Y}=\mathbf{AX}$

$$\begin{aligned}F\left(\mathbf{Y}\in B'\right) &= \int_{\mathbf{Y}\in B'}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=2}^{n}y_i^2+\left[y_1-\sqrt{n}\mu\right]^2\right)\right]\mid\mathbf{J}\mid d\mathbf{y}\\ &= \int_{\mathbf{Y}\in B'}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=2}^{n}y_i^2+\left[y_1-\sqrt{n}\mu\right]^2\right)\right]d\mathbf{y}.\end{aligned}$$

Với B' là ảnh của B qua ánh xạ $\mathbf{A}$, $\mathbf{J}$ là định thức ma trận Jacobi với $\mathbf{J}=|\mathbf{A}^T|$. Vậy các biến ngẫu nhiên $Y_1,\ldots,Y_n$ là độc lập, trong đó $Y_2,\ldots,Y_n$ cùng phân phối chuẩn $\mathcal{N}\left(0,\sigma^2\right)$. Kết hợp định nghĩa 1.2.4 và đẳng thức (2.5) suy ra biến ngẫu nhiên

$$T=\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2=\sum_{i=2}^{n}\left(\frac{Y_i}{\sigma}\right)^2$$

tuân theo phân phối khi bình phương với $n-1$ bậc tự do. □

Nhận xét 2.2.1 *Nếu cho* $\sqrt{n_i}X_i\sim\mathcal{N}\left(\sqrt{n_i}\mu,\sigma^2\right),\ i=\overline{1,n}$ *và*

$$\bar{X}=\frac{\sum_{i=1}^{n}n_iX_i}{\sum_{i=1}^{n}n_i}$$

định lý trên vẫn đúng. Ma trận trực giao $\boldsymbol{A}$ *lúc này thỏa mãn*

$$a_{1i} = \frac{\sqrt{n_i}}{\sqrt{\sum_{i=1}^{n} n_i}},\ i = \overline{1,n}.$$

Với $Y_1 = \sqrt{\sum_{i=1}^{n} n_i}\bar{X}$*, các bước chứng minh còn lại tương tự trên.*

Định lý 2.2.2 (xem [6]) *Cho* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *là các véc tơ mẫu ngẫu nhiên với kích cỡ tương ứng* $n_1, \ldots, n_n$ *lấy từ n nhóm độc lập với nhau và các điều kiện của ANOVA một nhân tố được thỏa mãn. Xét cặp giả thuyết đối thuyết*

$$H_0:\ \mu_1 = \mu_2 = \cdots = \mu_n,\ \text{với}\ H_1:\ \mu_j \neq \mu_i,\ \ i, j \text{ nào đó}.$$

Đặt $SS_{tr} = \sum_{i=1}^{n} n_i \left(\overline{X}_i - \overline{X}\right)^2$, $SS_{res} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \left(X_{ji} - \overline{X}_i\right)^2$*. Nếu giả thuyết* H_0 *đúng thì biến ngẫu nhiên*

$$F = \frac{SS_{res}/(n-1)}{SS_{res}/\left(\sum_{i=1}^{n} n_i - n\right)}$$

tuân theo phân phối Fisher với $\left(n-1, \sum_{i=1}^{n} n_i - n\right)$ *bậc tự do. Ta bác bỏ* H_0 *chấp nhận* H_1 *với mức ý nghĩa* α *nếu*

$$F > F_{n-1,\sum_{i=1}^{n} n_i - n}(\alpha).$$

Trong đó $F_{n-1,\sum_{i=1}^{n} n_i - n}(\alpha)$ *là phân vị phải của phân phối Fisher mức* α *với* $\left(n-1, \sum_{i=1}^{n} n_i - n\right)$ *bậc tự do.*

Chứng minh. Giả sử H_0 đúng, ta có

$$\mu_1 = \mu_2 = \cdots = \mu_n.$$

Khi đó $X_{ji} \sim \mathcal{N}\left(\mu, \sigma^2\right)$, với $i = \overline{1,n}$ và $j = \overline{1,n_i}$. Như vậy

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ji}$$

tuân theo phân phối chuẩn $\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n_i}\right)$ hay $\sqrt{n_i}\bar{X}_i \sim \mathcal{N}\left(\sqrt{n_i}\mu, \sigma^2\right)$. Mà

$$\bar{X} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i} X_{ji}}{\sum_{i=1}^{n} n_i} = \frac{\sum_{i=1}^{n} n_i \bar{X}_i}{\sum_{i=1}^{n} n_i}$$

nên theo định lý 2.2.1 và nhận xét 2.2.1 biến ngẫu nhiên

$$T_{num} = \sum_{i=1}^{n} n_i \left(\frac{\bar{X}_i - \bar{X}}{\sigma}\right)^2$$

tuân theo phân phối khi bình phương với $n-1$ bậc tự do. Tương tự biến ngẫu nhiên

$$T_{den} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(\frac{X_{ji} - \bar{X}_i}{\sigma} \right)^2$$

tuân theo phân phối khi bình phương với $\sum_{i=1}^{n} n_i - n$ bậc tự do. Từ định nghĩa 1.2.6 biến ngẫu nhiên

$$F = \frac{T_{num}/(n-1)}{T_{den}/(\sum_{i=1}^{n} n_i - n)} = \frac{\mathrm{SS}_{res}/(n-1)}{\mathrm{SS}_{res}/(\sum_{i=1}^{n} n_i - n)}$$

có phân phối Fisher với $(n-1, \sum_{i=1}^{n} n_i - n)$ bậc tự do.

Quay lại với bài toán kiểm định giả thuyết, với mức ý nghĩa α ta muốn

$$P(\text{Bác bỏ } H_0 \mid H_0 \text{ đúng}) \leq \alpha.$$

Như đã nói trên, ta càng có cơ sở bác bỏ H_0 và chấp nhận H_1 nếu tỉ số giữa SS_{tr} và SS_{res} càng lớn, hay lớn hơn một số C nào đó. Như vậy để kiểm định ta cần tìm số C sao cho

$$\begin{aligned} P\left(\frac{\mathrm{SS}_{tr}/(n-1)}{\mathrm{SS}_{res}/(\sum_{i=1}^{n} n_i - n)} > C \mid H_0 \text{ đúng} \right) &\leq \alpha \\ \Leftrightarrow P(F > C \mid H_0 \text{ đúng}) &\leq \alpha \\ \Leftrightarrow C &\geq F_{n-1, \sum_{i=1}^{n} n_i - n}(\alpha). \end{aligned}$$

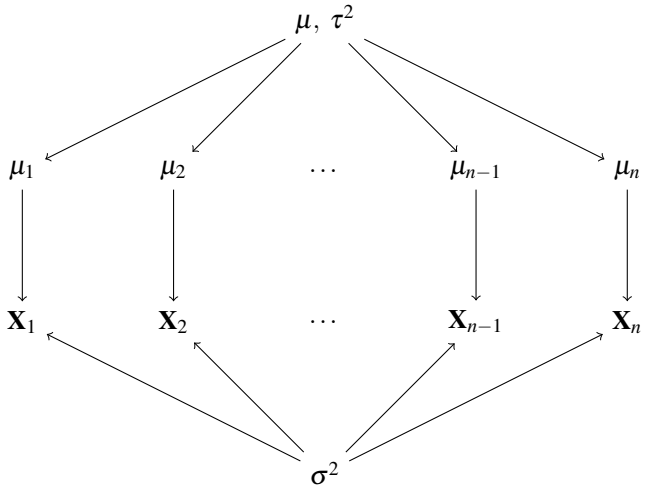
Vậy ta chọn $C = F_{n-1, \sum_{i=1}^{n} n_i - n}(\alpha)$. Như vậy ta bác bỏ H_0 chấp nhận H_1 nếu $F > F_{n-1, \sum_{i=1}^{n} n_i - n}(\alpha)$. □

Thông thường, chỉ số $p-$ giá trị được sử dụng để đặc trưng mức độ bác bỏ giả thuyết. $p-$ giá trị trong ANOVA được xác định bởi

$$p - \text{giá trị} = P\left(F_{n-1, \sum_{i=1}^{n} n_i - n} > \frac{\mathrm{SS}_{tr}/(n-1)}{\mathrm{SS}_{res}/(\sum_{i=1}^{n} n_i - n)} \mid H_0 \text{ đúng} \right).$$

Nói cách khác, $p-$ giá trị là ngưỡng α cao nhất để bác bỏ giả thuyết H_0, $p-$ giá trị càng nhỏ chứng tỏ H_1 càng đáng tin cậy.

Phương pháp ANOVA một nhân tố dễ dàng kiểm tra xem có sự khác nhau giữa các nhóm hay không, tuy nhiên lại không chỉ rõ sự khác biệt giữa từng nhóm. Để so sánh hai nhóm, có thể dùng ANOVA để tính toán dữ liệu của hai nhóm này và đưa ra kết quả. Việc này có thể "vô tình" bỏ phí các thông tin hữu ích trong dữ liệu của các nhóm khác, đặc biệt trong trường hợp giữa các nhóm có những điểm chung nào đó.



Hình 2.1: Mô hình phân cấp ANOVA.

2.3 Mô hình phân cấp

Mô hình phân cấp là các mô hình thống kê mà các tham số của nó được chia thành nhiều cấp độ, phù hợp với dữ liệu phân cấp. Dữ liệu trong bài toán so sánh nhiều giá trị trung bình đã nêu trên cũng có cấu trúc phân 2 cấp. Cấp 1 là dữ liệu của từng nhóm, cấp 2 là dữ liệu của từng cá thể trong nhóm. Mỗi nhóm được đặc trưng bởi bộ tham số, giữa các bộ tham số lại có sự "liên hệ với nhau", điều này giúp các nhóm chia sẻ thông tin trong quá trình tính toán. Cụ thể, xét mô hình ANOVA một nhân tố với các giả sử được thỏa mãn được thỏa mãn

$$X_{ji} = \mu_i + \varepsilon_{ji}$$

với $\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right)$. Nhóm i đặc trưng bởi tham số μ_i, các quan sát X_{ji} là độc lập có điều kiện với tham số μ_i và σ^2

$$\left\{X_{1i}, \ldots, X_{n_i i} \mid \mu_i, \sigma^2\right\} \sim \text{ i.i.d. } \mathcal{N}\left(\mu_i, \sigma^2\right).$$

Tương tự, các tham số $\mu_1, \ldots, \mu_n \,|\mu, \tau^2 \sim$ i.i.d. $p\left(\mu_i \,|\mu, \tau^2\right)$. Ngoài ra các tham số μ, τ^2 và σ^2 cũng là các biến ngẫu nhiên. Hình 2.1 mô tả tổng quan mô hình phân cấp cũng như mức tác động của các tham số tới các mẫu ngẫu nhiên. Có thể tưởng tượng rằng dữ liệu thu thập bằng cách chọn một cách ngẫu nhiên các nhóm từ một tổng thể lớn, sau đó lại với mỗi nhóm lại chọn ra từng cá thể. Điều này khá hợp lý trong thực tế vì số lượng nhóm có thể rất nhiều và chỉ có "ngẫu nhiên" các nhóm để quan sát.

2.3.1 Phấn phối tiên nghiệm và hàm hợp lý

Phân phối tiên nghiệm của các tham số μ, τ^2, σ^2 là

$$\mu \sim \mathcal{N}\left(\mu_0, \gamma_0^2\right), \tag{2.6}$$

$$\frac{1}{\tau^2} \sim \text{Gamma}\left(\alpha_0, \beta_0\right), \tag{2.7}$$

$$\frac{1}{\sigma^2} \sim \text{Gamma}\left(\alpha_1, \beta_1\right). \tag{2.8}$$

Các tham số $\mu_1, \ldots, \mu_n$ là độc lập có điều kiện với tham số μ, τ^2 tuân theo phân phối chuẩn

$$\left\{\mu_i \mid \mu, \tau^2\right) \sim \mathcal{N}\left(\mu, \tau^2\right).$$

Với các mẫu ngẫu nhiên $\mathbf{X}_1, \ldots, \mathbf{X}_n$, hàm hợp lý xác định bởi

$$\begin{aligned}
&p\left(\mathbf{X}_1, \ldots, \mathbf{X}_n \mid \mu_1, \ldots, \mu_n, \mu, \tau^2, \sigma^2\right) \\
&= \left\{\prod_{i=1}^{n} \prod_{j=1}^{n_i} p\left(X_{ji} \mid \mu_i, \sigma^2\right)\right\} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{\sum_{i=1}^{n} n_i}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(X_{ji} - \mu_i\right)^2\right]\right)
\end{aligned}$$

2.3.2 Tính toán phấn phối hậu nghiệm

Áp dụng công thức Bayes trong định lý 1.3.1, công thức hậu nghiệm có dạng

$$\begin{aligned}
&p\left(\mu_1, \ldots, \mu_n, \mu, \tau^2, \sigma^2 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n\right) \\
&\propto p\left(\mu_1, \ldots, \mu_n, \mu, \tau^2, \sigma^2\right) p\left(\mathbf{x}_1, \ldots, \mathbf{x}_n \mid \mu_1, \ldots, \mu_n, \mu, \tau^2, \sigma^2\right) \\
&\propto p\left(\mu\right) p\left(\tau^2\right) p\left(\sigma^2\right) \left\{\prod_{i=1}^{n} p\left(\mu_i \mid \mu, \tau^2\right)\right\} \left\{\prod_{i=1}^{n} \prod_{j=1}^{n_i} p\left(x_{ji} \mid \mu_i, \sigma^2\right)\right\}.
\end{aligned}$$

Hàm mật độ đồng thời của các tham số trong mô hình là hàm phức tạp kéo theo khó khăn trong xác định mật độ xác suất biên của từng tham số, khó xác định kỳ vọng, phương sai hay khoảng tin cậy. Các điều này cản trở việc suy luận giải quyết bài toán. Thay vào đó nên sử dụng các phương pháp để "xấp xỉ" mật độ hâu nghiệm, ví dụ như phương pháp lấy mẫu Gibbs. Yêu cầu của phương pháp này đó là xác định các hàm mật độ biên có điều kiện đầy đủ là dễ dàng lấy mẫu Với sự lựa chọn phân phối tiên nghiệm và hàm hợp lý như trên, yêu cầu này được thỏa mãn.

Phân phối hậu nghiệm có điều kiện đầy đủ của μ

Trước tiên biến đổi đẳng thức

$$\begin{aligned}
&p\left(\mu_1,\ldots,\mu_n,\mu,\tau^2,\sigma^2 \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)\\
&= p\left(\mu \mid \mu_1,\ldots,\mu_n,\tau^2,\sigma^2\mathbf{x}_1,\ldots,\mathbf{x}_n\right) p\left(\mu_1,\ldots,\mu_n,\tau^2,\sigma^2 \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right).
\end{aligned}$$

Hay $p\left(\mu \mid \mu_1,\ldots,\mu_n,\tau^2,\sigma^2,\mathbf{x}_1,\ldots,\mathbf{x}_n\right) \propto p\left(\mu_1,\ldots,\mu_n,\mu,\tau^2,\sigma^2 \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)$. Thu được kết quả trên do những số hạng không chứa μ được coi là hằng số. Sử dụng kỹ thuật này và tiếp tục biến đổi

$$\begin{aligned}
&p\left(\mu \mid \mu_1,\ldots,\mu_n,\tau^2,\sigma^2,\mathbf{x}_1,\ldots,\mathbf{x}_n\right)\\
&\propto p\left(\mu_1,\ldots,\mu_n,\mu,\tau^2,\sigma^2 \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)\\
&\propto p(\mu)\prod_{i=1}^{n} p\left(\mu_i \mid \mu,\tau^2\right)\\
&\propto \exp\left(-\frac{1}{2\gamma_0^2}\left[\mu^2-2\mu\mu_0\right]\right)\times\exp\left(-\frac{1}{2\tau^2}\left[n\mu^2-2\sum_{i=1}^{n}\mu_i\mu\right]\right)\\
&\propto \exp\left(-\frac{1}{2\gamma_0^2\tau^2}\left[\left(\tau^2+n\gamma_0^2\right)\mu^2-2\mu\left(\mu_0\tau^2+\gamma_0^2\sum_{i=1}^{n}\mu_i\right)\right]\right)\\
&\propto \exp\left(-\frac{\tau^2+n\gamma_0^2}{2\tau^2\gamma_0^2}\left[\mu^2-2\mu\frac{\tau^2\mu_0+\gamma_0^2\sum_{i=1}^{n}\mu_i}{\tau^2+n\gamma_0^2}\right]\right).
\end{aligned}$$

Như vậy

$$\left\{\mu|\mu_1,\ldots,\mu_n,\tau^2,\sigma^2,\mathbf{x}_1,\ldots,\mathbf{x}_n\right\} \sim \mathcal{N}\left(\frac{\tau^2\mu_0+\gamma_0^2\sum_{i=1}^{n}\mu_i}{\tau^2+n\gamma_0^2}, \frac{\tau^2\gamma_0^2}{\tau^2+n\gamma_0^2}\right).$$

Phân phối hậu nghiệm của điều kiện đầy đủ của τ^2

Áp dụng kỹ thuật tương tự trên thu được

$$\begin{aligned}
&p\left(\tau^2 \mid \mu_1,\ldots,\mu_n,\mu,\sigma^2,\mathbf{x}_1,\ldots,\mathbf{x}_n\right)\\
&\propto p\left(\mu_1,\ldots,\mu_n,\mu,\tau^2,\sigma^2 \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)\\
&\propto p\left(\tau^2\right)\prod_{i=1}^{n} p\left(\mu_i \mid \mu.\tau^2\right)\\
&\propto \left(\frac{1}{\tau^2}\right)^{\alpha_0-1}\times\exp\left(-\frac{\beta_0}{\tau^2}\right)\times\left(\frac{1}{\tau^2}\right)^{\frac{n}{2}}\times\exp\left(-\frac{1}{2\tau^2}\left[\sum_{i=1}^{n}\left(\mu_i-\mu\right)^2\right]\right)\\
&\propto \left(\frac{1}{\tau^2}\right)^{\alpha_0+\frac{n}{2}-1}\times\exp\left(-\frac{1}{2\tau^2}\left[2\beta_0+\sum_{i=1}^{n}\left(\mu_i-\mu\right)^2\right]\right).
\end{aligned}$$

Nói cách khác

$$p\left(\frac{1}{\tau^2}\mid \mu_1,\ldots,\mu_n,\mu,\sigma^2,\mathbf{x}_1,\ldots,\mathbf{x}_n\right)$$
$$\sim \text{Gamma}\left(\frac{2\alpha_0+n}{2},\frac{2\beta_0+\sum_{i=1}^{n}\left[\mu_i-\mu\right]^2}{2}\right).$$

Phân phối hậu nghiệm có điều kiện đầy đủ của σ^2

Biến đổi tương tự thu được

$$\begin{aligned}
&p\left(\sigma^2\mid \mu_1,\ldots,\mu_n,\mu,\tau^2,\mathbf{x}_1,\ldots,\mathbf{x}_n\right)\\
&\propto p\left(\mu_1,\ldots,\mu_n,\mu,\tau^2,\sigma^2\mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)\\
&\propto p\left(\sigma^2\right)\prod_{i=1}^{n}\prod_{j=1}^{n_i}p\left(x_{ji}\mid \mu_i,\sigma^2\right)\\
&\propto \left(\frac{1}{\sigma^2}\right)^{\alpha_1-1}\times\exp\left(\frac{-\beta_1}{\sigma^2}\right)\times\left(\frac{1}{\sigma^2}\right)^{\frac{\sum_{i=1}^{n}n_i}{2}}\times\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[x_{ji}-\mu_i\right]^2\right)\\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{\beta_1+\sum_{i=1}^{n}n_i}{2}-1}\times\exp\left(-\frac{1}{2\sigma^2}\left[2\beta_1+\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(x_{ji}-\mu_i\right)^2\right]\right).
\end{aligned}$$

Hay

$$p\left(\frac{1}{\sigma^2}\mid \mu_1,\ldots,\mu_n,\mu,\tau^2,\mathbf{x}_1,\ldots,\mathbf{x}_n\right)$$
$$\sim \text{Gamma}\left(\frac{2\alpha_1+\sum_{i=1}^{n}n_i}{2},\frac{2\beta_1+\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(x_{ji}-\mu_i\right)^2}{2}\right).$$

Phân phối hậu nghiệm có điều kiện đầy đủ của μ_i

Biến đổi như trên thu được

$$\begin{aligned}
&p\left(\mu_i\mid \mu_1,\ldots,\mu_{i-1},\mu_{i+1},\ldots,\mu_n,\mu,\tau^2,\sigma^2\right)\\
&\propto p\left(\mu_1,\ldots,\mu_n,\mu,\tau^2,\sigma^2\mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)\\
&\propto p\left(\mu_i\mid \mu,\tau^2\right)\prod_{j=1}^{n_i}p\left(x_{ji}\mid \mu_i,\sigma^2\right)\\
&\propto \exp\left(-\frac{1}{2\tau^2}\left[\mu_i^2-2\mu_i\mu\right]\right)\times\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}\left[\mu_i^2-2x_{ji}\mu_i\right]\right)\\
&\propto \exp\left(-\frac{\sigma^2+n_i\tau^2}{2\tau^2\sigma^2}\left[\mu_i^2-2\mu_i\frac{\tau^2\sum_{j=1}^{n_i}x_{ji}+\sigma^2\mu}{\sigma^2+n_i\tau^2}\right]\right).
\end{aligned}$$

Vậy

$$p\left(\mu_i \mid \mu_1,\ldots,\mu_{i-1},\mu_{i+1},\ldots,\mu_n,\mu,\tau^2,\sigma^2,\mathbf{x}_1,\ldots,\mathbf{x}_n\right) \sim \mathcal{N}\left(\frac{\tau^2\sum_{j=1}^{n_i}x_{ji}+\sigma^2\mu}{\sigma^2+n_i\tau^2},\frac{\sigma^2\tau^2}{\sigma^2+n_i\tau^2}\right).$$

2.3.3 Lấy mẫu Gibbs và kiểm định giả thuyết

Như đã nói ở trên, phân phối hậu nghiệm có điều kiện đầy đủ của các tham số là phân phối Gamma và phân phối chuẩn, các phân phối này đều dễ dàng lấy mẫu khi sử dụng các ngôn ngữ lập trình như `python` hoặc `R`. Thuật toán lấy mẫu Gibbs thực hiện như sau:

- Bước 1: Chọn bộ giá trị ngẫu nhiên cho bộ tham số $\left(\mu^{(1)},\tau^{(1)},\sigma^{(1)},\mu_1^{(1)},\ldots,\mu_n^{(1)}\right)$, khởi tạo biến đếm $j=1$.
- Bước 2: Lấy mẫu $\mu^{(j+1)} \sim p\left(\mu \mid \left(\tau^{(j)}\right)^2,\left(\tau^{(j)}\right)^2,\mu_1^{(j)},\ldots,\mu_n^{(j)},\mathbf{x}_1,\ldots,\mathbf{x}_n\right)$.
- Bước 3: Lấy mẫu $\frac{1}{\left(\tau^{(j+1)}\right)^2} \sim p\left(\frac{1}{\tau^2}\mid\mu^{(j+1)},\left(\sigma^{(j)}\right)^2,\mu_1^{(j)},\ldots,\mu_n^{(j)},\mathbf{x}_1,\ldots,\mathbf{x}_n\right)$.
- Bước 4: Lấy mẫu $\frac{1}{\left(\sigma^{(j+1)}\right)^2} \sim p\left(\frac{1}{\sigma^2}\mid\mu^{(j+1)},\left(\tau^{(j+1)}\right)^2,\mu_1^{(j)},\ldots,\mu_n^{(j)},\mathbf{x}_1,\ldots,\mathbf{x}_n\right)$.
- Bước 5: Lần lượt lấy mẫu $\mu_i^{(j+1)} \sim p\left(\mu_i \mid \mu^{(j+1)},\left(\tau^{(j+1)}\right)^2,\left(\sigma^{(j+1)}\right)^2,\mu_1^{(j+1)},\ldots,\mu_{i-1}^{(j+1)},\mu_{i+1}^{(j)},\ldots,\mu_n^{(j)},\ldots,\mathbf{x}_n\right)$.
- Bước 6: Đặt $j=j+1$ và quay lại bước 2.

Thông thường, thuật toán kết thúc khi j vượt ngưỡng nào đó và thu được nhiều véc tơ tham số. Trong thuật toán, các giá trị tham số mới nhất được sử dụng để lấy mẫu cho tham số khác, điều này làm cho thông tin giữa các nhóm được chia sẻ với nhau. Cụ thể, $\mathbf{x}_i$ dùng để cung cấp thông tin về μ_i, các $\mu_1,\ldots,\mu_n$ được sử dụng để lấy mẫu cho μ,τ,σ và các tham số này lại cung cấp thông tin cho việc lấy mẫu μ_i, hay thông tin giữa các nhóm được chia sẻ để suy ra những đặc trưng của từng nhóm. Các véc tơ

tham số sẽ được dùng để sử dụng tính toán các đặc trưng như kỳ vọng, phương sai và kiểm định giả thuyết. Giả sử thuật toán kết thúc thu được n véc tơ tham số, các đặc trưng của tham số được xấp xỉ bởi MCMC, ví dụ với biến ngẫu nhiên μ

$$E(\mu \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{1}{n} \sum_{i=1}^{n} \mu^{(i)}.$$

Quay trở lại với cặp giả thuyết đối thuyết

$$H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_n, \text{ với } H_1 : \ \mu_j \neq \mu_i, \ \ i, j \text{ nào đó.}$$

Với mức ý nghĩa α, giả thuyết H_0 bị bác bỏ và H_1 được chấp nhận nếu tồn tại i, j thỏa mãn

$$P([\mu_i - \mu_j] > 0 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) > \alpha$$
$$\text{hoặc } P([\mu_i - \mu_j] < 0 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) > \alpha.$$

Biểu thức xác xuất tính bởi xấp xỉ MCMC, ví dụ

$$P([\mu_i - \mu_j] > 0 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{1}{n} \sum_{m=1}^{n} I_{(\mu_i - \mu_j) > 0} \left(\mu_i^{(m)} - \mu_j^{(m)} \right).$$

Để tiện so sánh với $p-$ giá trị, $p_{Bayes}-$ giá trị được tính bởi

$$p_{Bayes}-\text{giá trị} = 1 - \max\left(P([\mu_i - \mu_j] > 0 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n), P([\mu_i - \mu_j] < 0 \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) \right).$$

Gọi $\lambda(\Theta)$ là phân phối hậu nghiệm của mô hình phân cấp với $\Theta = (\mu, \tau, \sigma, \mu_1, \ldots, \mu_n)$. Định lý 2.3.1 cung cấp cơ sở cho sự hội tụ của quá trình xấp xỉ MCMC.

Định lý 2.3.1 (xem [2]) *Xét xích Markov tạo bởi thuật toán Gibbs trong mô hình ANOVA phân cấp với phân phối hậu nghiệm $\lambda(\Theta)$, với n nhóm tương ứng với kích cỡ $n_1, \ldots, n_n$. Đặt $m = \min\{n_1, \ldots, n_n\}$, $M = \max\{n_1, \ldots, n_n\}$. Nếu $m \geq \left(\sqrt{5} - 2\right) M$ và tham số α_0 trong (2.7) thỏa mãn $\alpha_0 > (3n-2)/(2n-2)$ thì với mọi $\Theta \in S_\Theta$*

$$\sup_{\mathbf{y} \in S_\Theta} |P^t(\Theta, \mathbf{y}) - \lambda(\mathbf{y})| \leq M(\Theta) r^t.$$

Trong đó $M(\Theta)$ là hàm dương bị chặn trên tập S_Θ, r là hằng số thỏa $0 < r < 1$.

Chương 3

Ứng dụng

Phần này trình bày các kết quả của mô hình ANOVA truyền thống và ANOVA phân cấp trên hai tập dữ liệu mô phỏng và thực tế. Với dữ liệu mô phỏng, chúng ta sẽ thực hiện nhiều kịch bản về phân phối tiên nghiệm và kích cỡ dữ liệu, phân tích kết quả và rút ra nhận xét về sự ảnh hưởng của các thành phần trong mô hình cũng như so sánh kết quả kiểm định giả thuyết giữa hai mô hình. Những nhận xét này góp phần quan trọng vào việc sử dụng mô hình với dữ liệu thực tế.

3.1 Nghiên cứu mô phỏng

Dữ liệu được sinh ra trong phần này được thiết lập trong bảng 3.1 và bảng 3.2.

Bảng 3.1: Thiết lập các tham số của dữ liệu mô phỏng.

Tham số	**Thiết lập**
μ	10
τ^2	1
σ^2	1
Số nhóm	20
μ_i	$\mathcal{N}(10,1)$

Bảng 3.2: Tham số đặc trưng của từng nhóm.

Tham số	Giá trị
μ_0	10.49671415
μ_1	10.49671415
μ_2	10.64768854
μ_3	11.52302986
μ_4	9.76584663
μ_5	9.76586304
μ_6	11.57921282
μ_7	10.76743473
μ_8	9.53052561
μ_9	10.54256004

Tham số	Giá trị
μ_{10}	9.53658231
μ_{11}	9.53427025
μ_{12}	10.24196227
μ_{13}	8.08671976
μ_{14}	8.27508217
μ_{15}	9.43771247
μ_{16}	8.98716888
μ_{17}	10.31424733
μ_{18}	9.09197592
μ_{19}	8.5876963

Với bộ dữ liệu mô phỏng này, ta sẽ sử dụng mô hình ANOVA phân cấp với hai kịch bản tiên nghiệm trong bảng 3.3.

Bảng 3.3: Các tham số cho từng kịch bản.

(a) Tham số cho kịch bản 1.

Tham số	Giá trị
μ_0	1
γ_0^2	1
α_0	2
β_0	2
α_1	2
β_1	2

(b) Tham số cho kịch bản 2.

Tham số	Giá trị
μ_0	5
γ_0^2	2
α_0	1.6
β_0	1.6
α_1	1.6
β_1	1.6

Sau khi mô phỏng dữ liệu, sử dụng thuật toán lấy mẫu Gibbs với 10000 vòng lặp để xác định các đặc trưng như kỳ vọng và phương sai của các tham số và kiểm định giả thuyết, sau đó so sánh các đặc điểm này với mô hình ANOVA cổ điển. Để kiểm tra sự hội tụ của xấp xỉ MCMC, ta sử dụng đồ thị vết để đánh giá. Mô hình ANOVA phân cấp được sử dụng với hai trường hợp là dữ liệu có 3 nhóm và dữ liệu có 20 nhóm, mỗi nhóm có 70 quan sát. Do càng nhiều nhóm thì khả năng bác bỏ H_0 càng cao và để làm nổi bật ưu thế của mô hình phân cấp, cả hai loại mô hình chỉ xét cặp giả thuyết

đối thuyết:

$$H_0 : \ \mu_1 = \mu_2, \text{ với } H_1 : \ \mu_1 \neq \mu_2. \tag{3.1}$$

3.1.1 Kịch bản 1

Bảng 3.4: Đặc trưng của các tham số với kịch bản 1 và 3 nhóm.

(a) Ước lượng kỳ vọng của các tham số cấp 1.

Tham số	**Bayes**
μ	1.796300
τ^2	48.340470
σ^2	1.102482

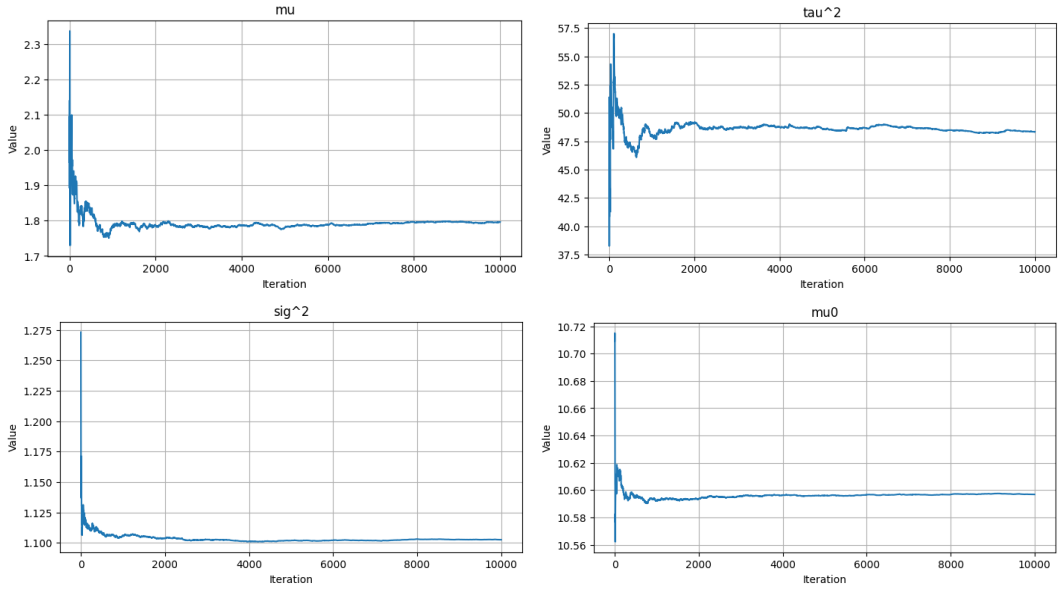
(b) Ước lượng kỳ vọng các tham số cấp 2.

Tham số	**Bayes**	**MLE**
μ_0	10.596812	10.601099
μ_1	10.532208	10.532679
μ_3	10.753042	10.754418

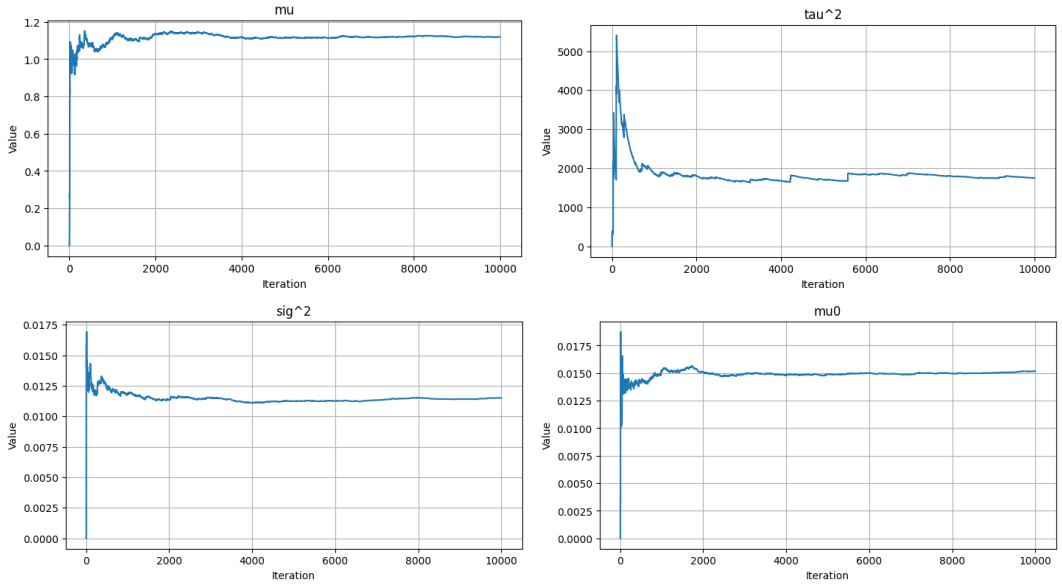
(c) Ước lượng phương sai các tham số.

Tham số	**Phương sai**
μ	1.119466
τ^2	1750.403185
σ^2	0.011498
μ_0	0.015176

Kết quả trong bảng 3.4 có vẻ không hợp lý khi mà ước lượng của μ và τ^2 lệch nhiều so với thông số đã thiết lập. Do phân phối tiên nghiệm của μ có độ phân tán thấp, các giá trị của biến ngẫu nhiên tập trung trong khoảng $(-2;4)$ và số nhóm để suy luận về μ chỉ có 3 nên phân phối hậu nghiệm của μ không thay đổi nhiều. Điều đó kéo theo kỳ vọng τ^2 phải lớn để cân bằng giữa μ và $\mu_1,\ \mu_2,\ \mu_3$ cũng như phương sai của τ^2, μ lớn. Tuy nhiên nếu chỉ ước lượng tham số cấp 2 và σ^2, các kết quả thu được khá tốt do các tham số này phụ thuộc chủ yếu vào số lượng quan sát trong từng nhóm do mỗi nhóm có 70 quan sát.



(a) Đồ thị vết kỳ vọng của các tham số.



(b) Đồ thị vết phương sai của các tham số.

Hình 3.1: Đồ thị vết với kịch bản 1 và 3 nhóm.

Để mô tả quá trình hội tụ của xấp xỉ MCMC, hình 3.1 bao gồm đồ thị vết của tham số μ, τ^2, σ^2, do vai trò của các nhóm là tương đương nên chỉ cần xét đồ thị vết cùa μ_0. Trong khoảng 2000 vòng lặp đầu tiên, đồ thị vết của các nhóm giao động nhiều do quá trình lấy mẫu Gibbs chưa ổn định, nhưng vòng lặp tiếp theo đồ thị dần ổn định chứng tỏ xấp xỉ MCMC đã hội tụ.

Bảng 3.5: Đặc trưng của các tham số với kịch bản 1 và 20 nhóm.

(a) Ước lượng kỳ vọng các tham số cấp 1.

Tham số	Bayes
μ	9.383498
τ^2	1.237073
σ^2	0.978773

(b) Ước lượng phương sai các tham số.

Tham số	Phương sai
μ	0.121474
τ^2	0.659492
σ^2	0.001400
μ_0	0.013656

(c) Ước lượng kỳ vọng các tham số cấp 2.

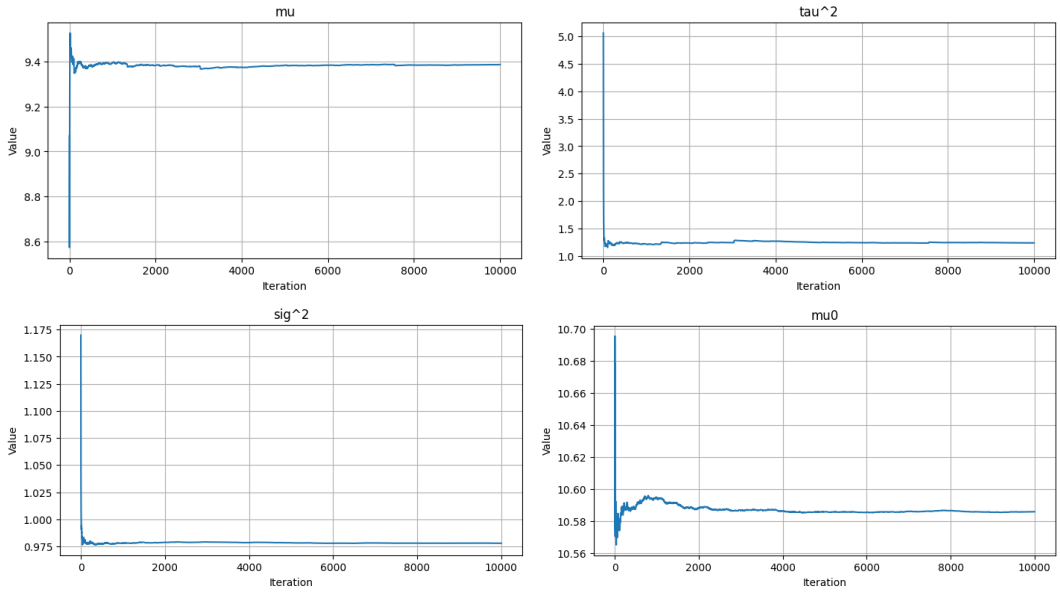
Tham số	Bayes	MLE
μ_0	10.585715	10.601099
μ_1	10.519333	10.532679
μ_2	10.735894	10.754418
μ_3	11.529131	11.554501
μ_4	9.807341	9.813113
μ_5	9.679049	9.680669
μ_6	11.381020	11.407143
μ_7	10.811568	10.829249
μ_8	9.521564	9.522698
μ_9	10.550118	10.564881

Tham số	Bayes	MLE
μ_{10}	9.495278	9.497040
μ_{11}	9.540574	9.543363
μ_{12}	10.230643	10.241216
μ_{13}	8.291711	8.276475
μ_{14}	8.539328	8.527350
μ_{15}	9.412385	9.410763
μ_{16}	9.016200	9.010055
μ_{17}	10.241475	10.251189
μ_{18}	9.212759	9.209883
μ_{19}	8.828689	8.821940

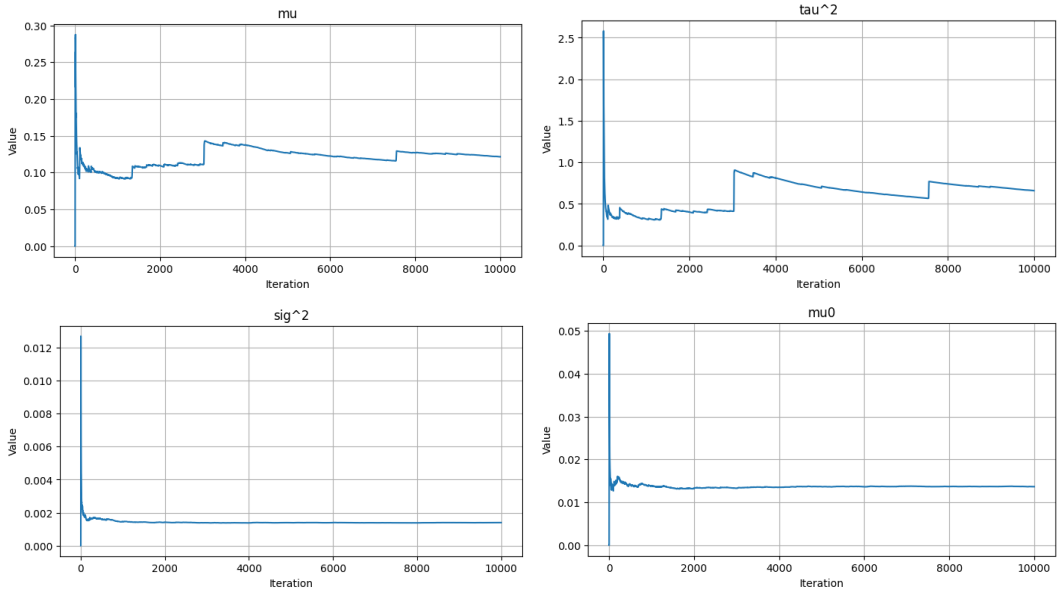
Khi số lượng nhóm được tăng lên là 20, kết quả thu được lưu trong bảng 3.5. Kỳ vọng của μ và τ^2 lúc này khá gần giá trị đúng là 10 và 1. Khi số lượng nhóm nhiều, phân phối hậu nghiệm của μ đã được điều chỉnh đáng kể để phù hợp với dữ liệu từ các nhóm kéo theo kỳ vọng của τ^2 cũng được xấp xỉ tốt. Các kỳ vọng xấp xỉ MCMC của từng nhóm nhìn chung gần giá trị đúng hơn khi so với ước lượng bằng giá trị trung bình. Số lượng nhóm lớn, hình 3.2 cho thấy quá trình hội tụ của xấp xỉ kỳ vọng nhanh nhưng xấp xỉ phương sai của μ, τ^2 có vẻ hội tụ chưa đủ tốt. Do phân phối tiên nghiệm và phân phối hậu nghiệm có sự thay đổi đáng kể, nên giá trị xấp xỉ có sự biến động lớn và cần số vòng lặp nhiều hơn để có kết quả tốt.

Quay trở lại bài toán kiểm định (3.1), kết quả được mô tả tại bảng 3.6, mô hình phân cấp cho thấy ưu thế của mình khi tận dụng được thông tin giữa các nhóm trong quá trình ước lượng, từ đó nâng cao khả năng bác bỏ H_0 so với mô hình ANOVA thông thường. Việc lựa chọn tiên nghiệm không quá ảnh hưởng đến quá trình kiểm định do số lượng quan sát trong mỗi nhóm nhiều.

H



(a) Đồ thị vết của kỳ vọng các tham số.



(b) Đồ thị vết của phương sai các tham số.

Hình 3.2: Đồ thị vết với kịch bản 1 và 20 nhóm.

Bảng 3.6: Các trường hợp kiểm định của kịch bản 1.

Số nhóm	p **– giá trị**	p_{Bayes} **– giá trị**
3	0.230130	0.1046
20	0.230130	0.0937

3.1.2 Kịch bản 2

Bảng 3.7: Đặc trưng các tham số với kịch bản 2 và 3 nhóm.

(a) Ước lượng kỳ vọng của các tham số cấp 1.

Tham số	**Bayes**
μ	9.087979
τ^2	3.896592
σ^2	1.106656

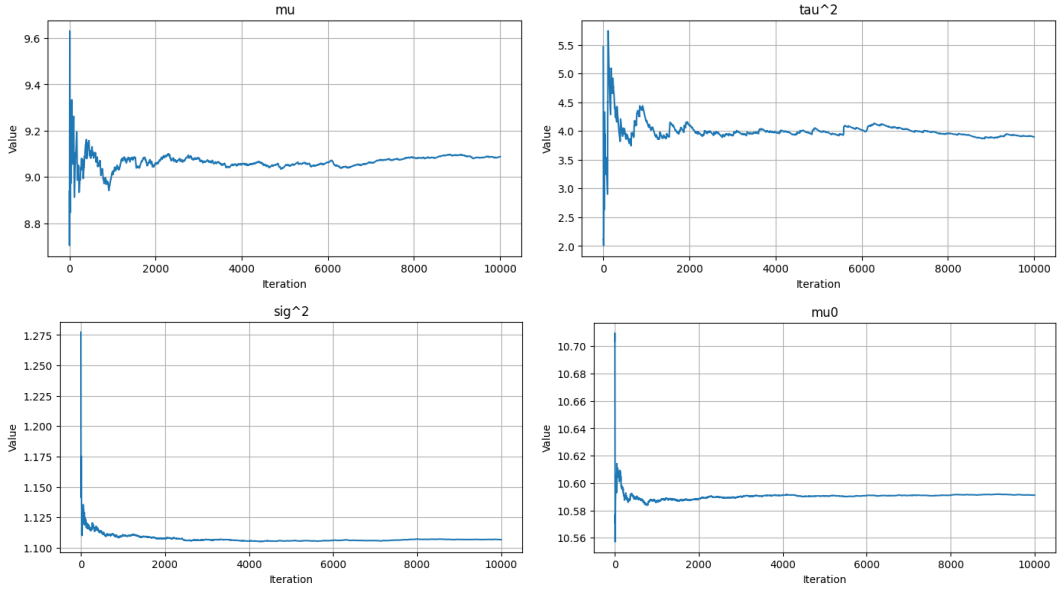
(b) Ước lượng kỳ vọng các tham số cấp 2.

Tham số	**Bayes**	**MLE**
μ_0	10.591117	10.601099
μ_1	10.527025	10.532679
μ_3	10.744668	10.754418

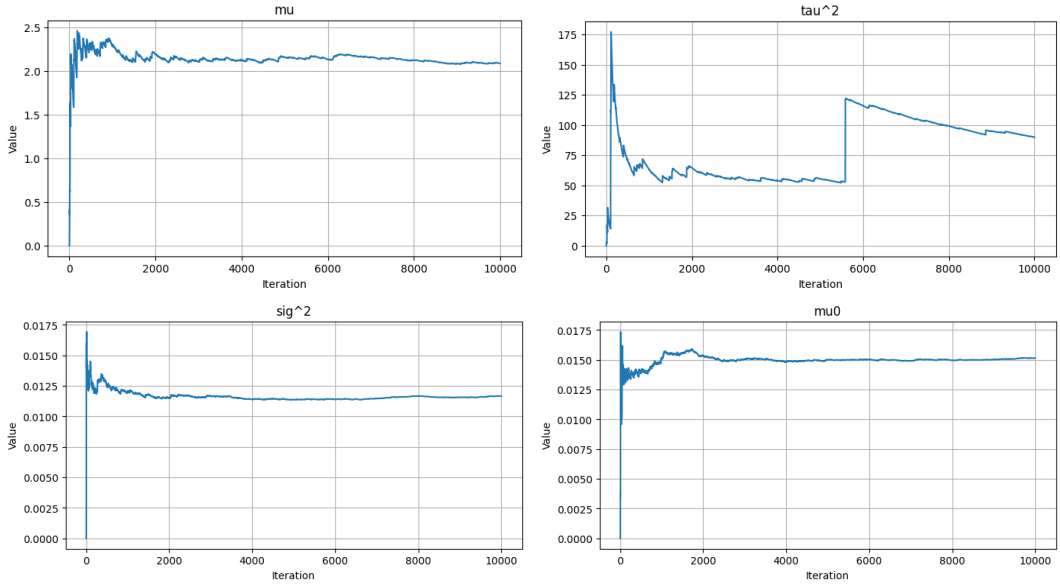
(c) Ước lượng phương sai các tham số.

Tham số	**Phương sai**
μ	2.087898
τ^2	89.895457
σ^2	0.011658
μ_0	0.015166

Kết quả trong bảng 3.7 trong khá hợp lý mặc dù không gần với giá trị đúng, chỉ sử dụng 3 nhóm nhưng tiên nghiệm của μ trong kịch bản 2 không quá tập trung dẫn tới phân phối hậu nghiệm thay đổi đáng kể để phù hợp dữ liệu cũng như ước lượng phương sai của μ lớn hơn so với kịch bản 1. Các đặc trưng của τ^2 cũng bớt bất thường do ước lượng kỳ vọng của μ lớn. Kỳ vọng của các tham số cấp 2 bé hơn so với kịch bản 1, nguyên nhân là do sự đóng góp đáng kể của μ, τ^2 trong công thức phân phối hậu nghiệm của chúng. Hình3.3 mô tả quá trình hội tụ của xấp xỉ MCMC. Đồ thị vết phương sai của τ^2 có một chỗ tăng quá nhanh mặc dù vị trí tương ứng trong đồ thị vết của μ không biến động nhiều là do quá trình lấy mẫu Gibbs đã lấy giá trị ở mức phân vị phải quá cao. Điều này khiến số lượng mẫu cần lớn hơn để thu được kết quả tốt.



(a) Đồ thị vết kỳ vọng của các tham số.



(b) Đồ thị vết phương sai của các tham số.

Hình 3.3: Đồ thị vết với kịch bản 2 và 3 nhóm.

Khi số nhóm tăng lên 20, kết quả các đặc trưng được lưu trong bảng 3.8, các đặc trưng lúc này đã chính xác hơn. Số nhóm tăng cũng khiến phương sai các của biến ngẫu nhiên giảm bởi sự trao đổi thông tin trong quá trình lấy mẫu Gibbs. Hình 3.4 mô tả quá trình hội tụ của xấp xỉ MCMC.

Bảng 3.8: Đặc trưng của các tham số với kịch bản 2 và 20 nhóm.

(a) Ước lượng kỳ vọng của các tham số cấp 1.

Tham số	**Bayes**
μ	9.786208
τ^2	0.939051
σ^2	0.978410

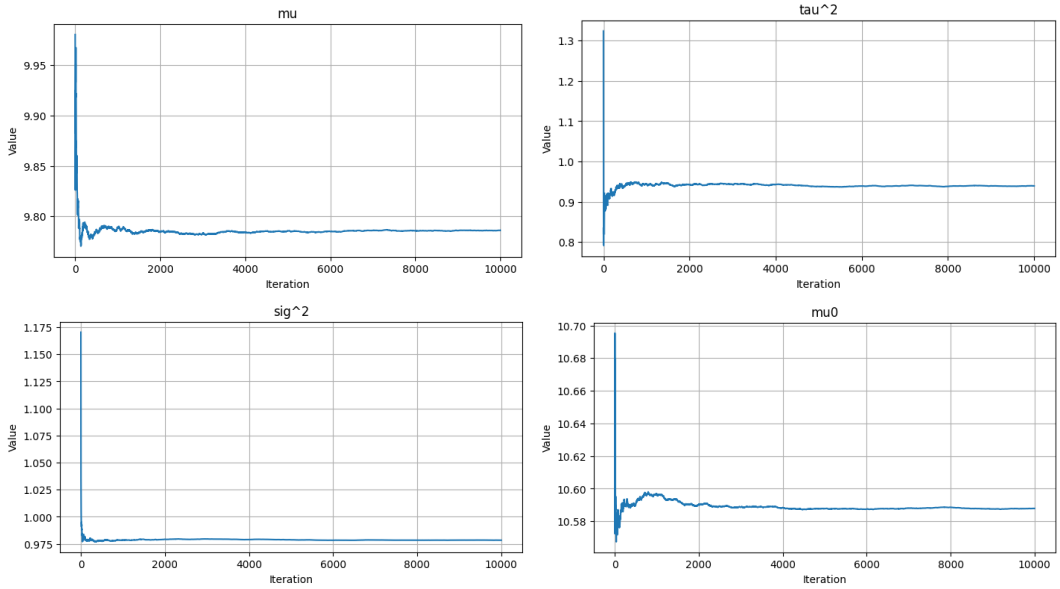
(b) Ước lượng phương sai các tham số.

Tham số	**Phương sai**
μ	0.048089
τ^2	0.102864
σ^2	0.001403
μ_0	0.013629

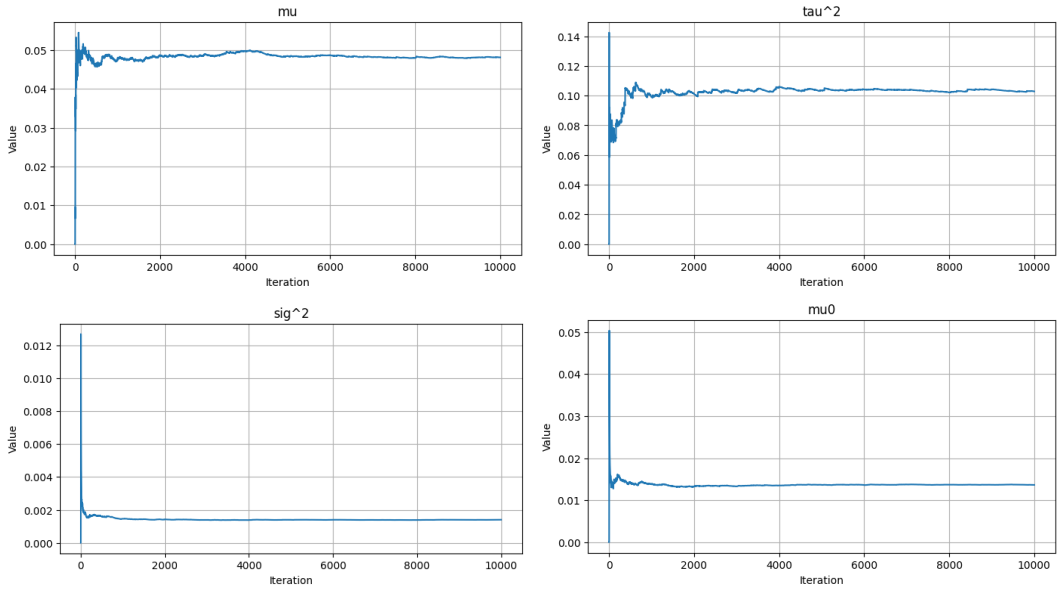
(c) Ước lượng kỳ vọng của các tham số cấp 2.

Tham số	**Bayes**	**MLE**
μ_0	10.587892	10.601099
μ_1	10.521702	10.532679
μ_2	10.737627	10.754418
μ_3	11.528563	11.554501
μ_4	9.811781	9.813113
μ_5	9.683867	9.680669
μ_6	11.380880	11.407143
μ_7	10.813087	10.829249
μ_8	9.526840	9.522698
μ_9	10.552398	10.564881

Tham số	**Bayes**	**MLE**
μ_{10}	9.500626	9.497040
μ_{11}	9.545790	9.543363
μ_{12}	10.233849	10.241216
μ_{13}	8.300565	8.276475
μ_{14}	8.547462	8.527350
μ_{15}	9.417980	9.410763
μ_{16}	9.022947	9.010055
μ_{17}	10.244654	10.251189
μ_{18}	9.218929	9.209884
μ_{19}	8.835977	8.821940



(a) Đồ thị vết kỳ vọng của các tham số.



(b) Đồ thị vết phương sai của các tham số.

Hình 3.4: Đồ thị vết với kịch bản 2 và 20 nhóm

Kết quả về kiểm định giả thuyết được lưu trong bảng 3.9, phân phối tiên nghiệm của μ, τ^2, σ^2 không ảnh hưởng nhiều tới kết quả do các tham số này đều tác động đồng thời tới các tham số cấp 2 trong quá trình lấy mẫu.

Bảng 3.9: Các trường hợp kiểm định của kịch bản 2.

Số nhóm	p **– giá trị**	p_{Bayes} **– giá trị**
3	0.230130	0.1071
20	0.230130	0.0939

3.2 Dữ liệu thực tế

Dữ liệu được sử dụng trong phần này là điểm thi thử lớp 12 các môn học của một trường trung học phổ thông ở thành phố Hà Nội. Dữ liệu bao gồm điểm của 644 học sinh thuộc 15 lớp. Xét yêu cầu so sánh điểm trung bình môn toán giữa các lớp của ngôi trường này, ta chỉ sử dụng trường điểm toán và trường thông tin về lớp để xây dựng mô hình. Coi các yêu cầu của mô hình ANOVA một nhân tố thỏa mãn.

Sau khi tiền xử lý dữ liệu, số lượng học sinh trong từng lớp được mô tả tại bảng 3.10.

Bảng 3.10: Số lượng học sinh của từng lớp.

Lớp	Số lượng
12A1	45
12A2	42
12A3	44
12A4	41
12D1	46
12D2	45
12D3	45

Lớp	Số lượng
12D4	44
12D5	44
12D6	41
12D7	40
12D8	42
12D9	41
12D10	42
12D11	39

Các tham số tiên nghiệm của mô hình ANOVA phân cấp có trong bảng 3.11.

Bảng 3.11: Tham số tiên nghiệm với dữ liệu thực tế.

Tham số	Giá trị
μ_0	7
γ_0	2
α_0	1.6
β_0	1.6
α_1	1.6
β_1	1.6

Thực hiện thuật toán lấy mẫu Gibbs với cỡ mẫu 10000, các đặc trưng thu được trong bảng 3.12. Số lượng lớp và số học sinh nhiều nên kết quả các ước lượng khá tốt với phương sai nhỏ. Quá trình hội tụ của xấp xỉ MCMC được mô tả tại hình 3.5. Nhìn chung chất lượng các lớp không có sự chênh lệch nhiều với điểm Toán, tuy nhiên trình

độ học sinh trong từng lớp lại chệnh lệch lớn với phương sai $\sigma^2 = 1.044800$. Như vậy để cải thiện điểm số cũng như thành tích chung của nhà trường, các thầy cô cần quan tâm đến tất cả các lớp học để cải thiện điểm số.

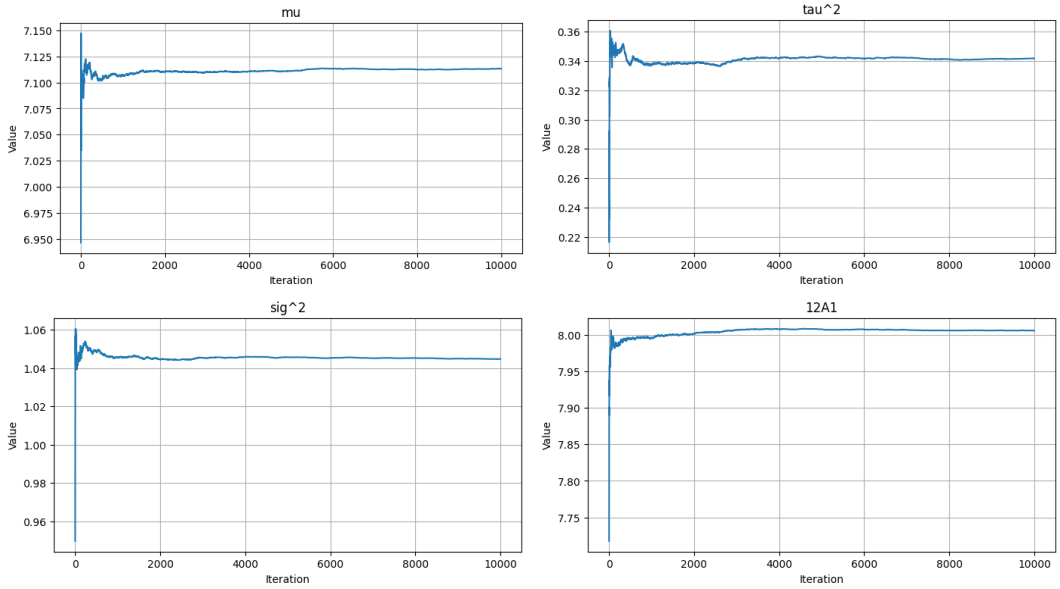
Bảng 3.12: Đặc trưng của các tham số với dữ liệu thực tế.

(a) Ước lượng kỳ vọng và phương sai của tham số cấp 1.

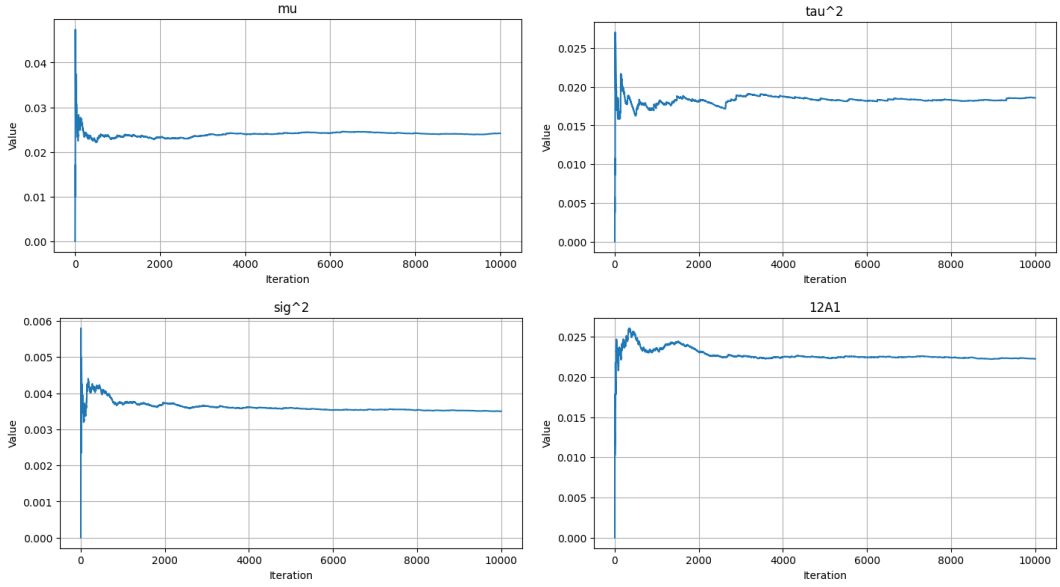
Tham số	**Kỳ vọng**	**Phương sai**
μ	7.113251	0.024147
τ^2	0.341780	0.018582
σ^2	1.044800	0.003499

(b) Ước lượng kỳ vọng và phương sai của tham số cấp 2.

Tham số	**MLE**	**Bayes**	**Phương sai**
μ_{12A1}	8.071111	8.005758	0.022265
μ_{12A2}	7.214286	7.209075	0.023517
μ_{12A3}	7.277273	7.264773	0.021977
μ_{12A4}	7.146341	7.143811	0.024002
μ_{12D1}	7.482609	7.454402	0.021278
μ_{12D2}	6.751111	6.775827	0.021361
μ_{12D3}	6.617778	6.653863	0.022114
μ_{12D4}	6.968182	6.981139	0.022138
μ_{12D5}	6.904545	6.920078	0.021928
μ_{12D6}	7.419512	7.395100	0.023600
μ_{12D7}	6.990000	7.000432	0.024115
μ_{12D8}	7.114286	7.114130	0.023296
μ_{12D9}	7.258537	7.245320	0.023680
μ_{12D10}	6.747619	6.775062	0.023352
μ_{12D11}	6.728205	6.760426	0.024933



(a) Đồ thị vết kỳ vọng.



(b) Đồ thị vết phương sai.

Hình 3.5: Đồ thị vết với dữ liệu thực tế.

Điểm số quyết định bởi nhiều yếu tố, hai lớp có lực học như nhau nhưng đến lúc thi điểm số có thể chênh lệch do các yếu tố khách quan như tâm lý, thời tiết, hay một số học sinh bỏ thi. Việc quyết định xem hai lớp có khác nhau về lực học không có thể ảnh hưởng nhiều đến kế hoạch giảng dạy của nhà trường. Ví dụ về kiểm định điểm trung bình môn toán của hai lớp 12D8 và 12D2 có khác nhau hay không, xét cặp giả

thuyết đối thuyết

$$H_0 : \ \mu_{12D8} = \mu_{12D2}, \text{ với } H_1 : \ \mu_{12D8} \neq \mu_{12D2}.$$

Bằng xấp xỉ MCMC, $P\left(\mu_{12D8} - \mu_{12D2} < 0 | \text{dữ liệu điểm toán}\right) = 0.0544$ trong khi $p-$ giá trị của phương pháp ANOVA cổ điển là 0.112872. Như vậy theo góc nhìn của mô hình ANOVA phân cấp, khả năng bác bỏ H_0 là cao hơn. Nói cách khác có đủ cơ sở thống kê với mức ý nghĩa 6% rằng hai lớp 12D8 và 12D2 có lực học môn toán khác nhau.

Kết luận

Đồ án đã đạt được mục tiêu đề ra

Đồ án này đã trình bày những kiến thức cơ bản về hai loại mô hình ANOVA. Qua đó cho ta thấy một phương pháp tiếp cận mới cho bài toán so sánh nhiều giá trị trung bình nói riêng và ước lượng tham số nói chung. Với mô hình phân cấp, ta có thể tích hợp các kiến thức tiên nghiệm vào mô hình ước lượng. Lợi thế của mô hình phân cấp là sự chia sẻ thông tin giữa các nhóm, từ đó có thểm nhiều bằng chứng để kiểm định giả thuyết hơn so với ANOVA truyền thống. Tuy nhiên, mô hình phân cấp có tính toán phức tạp cũng như yêu cầu khối lượng tính toán lớn để thu được kết quả tốt.

Kết quả của đồ án

Đồ án đã trình bày kiến thức cũng như các bước cơ bản để xây dựng cũng như tính toán hai loại mô hình ANOVA. Cụ thể:

1. Trình bày lý thuyết cơ bản có liên quan đến mô hình như các hàm phân phối, định lý Bayes, độc lập cùng phân phối có điều kiện, xấp xỉ tích phân.
2. Trình bày mô hình ANOVA theo trường phái Frequentist và Bayesian.
3. Thử nghiệm mô hình trên số liệu mô phỏng và thực tế.

Kỹ năng đạt được

1. Biết tìm kiếm, đọc, dịch tài liệu chuyên ngành liên quan đến nội dung đồ án.
2. Biết tổng hợp các kiến thức đã học và kiến thức trong tài liệu tham khảo để viết báo cáo đồ án.

3. Chế bản đồ án bằng LaTeX, viết chương trình tính toán của mô hình nhị thức sử dụng ngôn ngữ `Python`.
4. Biết tóm tắt nội dung đồ án và biết trình bày một báo cáo khoa học.

Hướng phát triển của đồ án trong tương lai

Trong quá trình làm đồ án, tác giả thấy có một số hướng để phát triển đề tài như sau:

1. Bài toán so sánh nhiều véc tơ trung bình.
2. Mô hình có nhiều nhân tố.
3. Mô hình với phương sai giữa các nhóm khác nhau.

Tài liệu tham khảo

[1] Dani Gamerman and Hedibert F. Lopes, *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference, Second Edition*,Chapman & Hall, 2006.

[2] Hobert, James P and Geyer, Charles J, "Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model", *Journal of Multivariate Analysis*, Vol 67, 1988, 414-430.

[3] Peter D.Hoff, *A First Course in Bayesian Statistical Methods*, Springer Science & Business Media, 2009.

[4] Robert V. Hogg and Joseph W. McKean and Allen T. Craig, *Introduction to Mathematical Statistics*, Pearson, 2019.

[5] Roberts, Gareth O, and Jeffrey S. Rosentha, "General state space Markov chains and MCMC algorithms", *Probability Survey*, Vol 1, 2004, 20-71.

[6] Richard Johnson and Dean Wichern, *Applied Multivariate Statistical Analysis*, Pearson New International Edition, 2014.

[7] Sean Meyn and Richard L. Tweedie, *Markov Chains and Stochastic Stability*, Cambridge University Press, 2009.

[8] S.C. Gupta and V.K. Kapoor, *Fundamentals of Mathematical Statistics*, SULTAN CHAND & SONS, Delhi, 2000.