BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

BÁO CÁO HỌC THUẬT

KHOẢNG TIN CẬY TRONG THỐNG KÊ BAYES

TS. Nguyễn Thị Hằng

Hà Nội, tháng 06 – 2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

BÁO CÁO HỌC THUẬT

KHOẢNG TIN CẬY TRONG THỐNG KÊ BAYES

Xác nhận của bộ môn

Hà Nội, tháng 06 - 2023

MỤC LỤC

LỜI MỞ ĐẦU

Trong những năm gần đây, phương pháp thống kê Bayes đang được sử dụng nhiều hơn trong các lĩnh vực từ khảo cổ học đến tính toán. Suy luận Bayes là phương pháp kết hợp thông tin thu thập được từ dữ liệu thực nghiệm với những thông tin có từ trước đó.

Nếu như suy luận thống kê cổ điển coi tham số như là một số giá trị cố định chưa biết, thì với thống kê Bayes, nó lại là biến ngẫu nhiên (theo nghĩa là ta có thể đưa nó về một phân bố xác suất thể hiện sự chắc chắn về giá trị thực của tham số). Đây chính là sự khác biệt cơ bản giữa hai cách tiếp cận. Có hai lý do để thống kê bayes hiệu quả hơn thống kê cổ điển:

(1) là các kết luận Bayes được thiết lập có điều kiện (thông tin tiên nghiệm) dựa trên mẫu dữ liệu thu thập được.

(2) là với quan điểm Bayes hoàn toàn hợp lý khi nói về xác suất để tỉ lệ rơi vào khoảng ước lượng (vì tham số là biến ngẫu nhiên).

Sau đây ta sẽ cụ thể hóa suy luận Bayes trong các bài toán ước lượng.

Hà Nội, tháng 06-2023

TS. Nguyễn Thị Hằng

CHƯƠNG 1. CƠ SỞ LÝ THUYẾT

1.1. Tổng quan về thống kê Bayes

Thống kê Bayes càng ngày càng phổ biến trong xây dựng các mẫu thống kê cho các vấn đề trong thực tế. Trong những năm gần đây, phương pháp thống kê Bayes đang được sử dụng nhiều hơn trong các lĩnh vực từ khảo cổ học đến tính toán. Suy luận Bayes là phương pháp kết hợp thông tin thu thập được từ dữ liệu thực nghiệm với những thông tin có từ trước đó.

Cách suy luận thống kê cổ điển mà ta đã biết là chỉ dựa trên một mẫu ngẫu nhiên, nhưng ý nghĩa của kết quả suy luận lại không phải như vậy. Ví dụ cho một mẫu chuẩn với phương sai đã biết, khoảng tin cậy 95% cho trung bình tập hợp μ là khoảng $\left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}}; \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right)$, điều này có nghĩa là khi các mẫu giá trị được lấy lặp lại từ tập hợp, ít nhất 95% các khoảng ngẫu nhiên chứa giá trị trung bình đúng μ. Kết luận này hoàn toàn khó kiểm tra lại mức độ đúng (có những bài toán khó có thể lấy được nhiều mẫu giá trị từ thực tế) và vì vậy, nó không phù hợp với thực tế lắm, hơn nữa trong suy luận cổ điển, ta không sử dụng đến những thông tin tiên nghiệm có từ trước. Vấn đề đặt ra là làm thế nào chỉ với một bộ mẫu dữ liệu và sử dụng thông tin tiên nghiệm để có thể ước lượng giá trị của tham số tại thời điểm thu thập dữ liệu đó? Thống kê Bayes là một trong những câu trả lời cho câu hỏi này.

Nếu như suy luận thống kê cổ điển coi tham số như là một số giá trị cố định chưa biết, thì với thống kê Bayes, nó lại là biến ngẫu nhiên (theo nghĩa là ta có thể đưa nó về một phân bố xác suất thể hiện sự chắc chắn về giá trị thực của tham số). Đây chính là sự khác biệt cơ bản giữa hai cách tiếp cận. Có 2 lý do để thống kê bayes hiệu quả hơn thống kê cổ điển: (1) là các kết luận Bayes được thiết lập có điều kiện (thông tin tiên nghiệm) dựa trên mẫu dữ liệu thu thập được, (2) là với quan điểm Bayes hoàn toàn hợp lý khi nói về xác suất để tỉ lệ rơi vào khoảng ước lượng (vì tham số là biến ngẫu nhiên).

1.2. Ước lượng Bayes

Cơ sở của phương pháp Bayes là định lý Bayes, cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên *A* với điều kiện sự kiện liên quan *B* đã xảy ra:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Giả sử $X = (X_1, \dots X_n)$ là mẫu ngẫu nhiên với không gian mẫu $(\mathcal{X}, \mathcal{A}, \mathcal{P} = \{P_\theta, \theta \in \Theta\})$. Giả sử $P_\theta \ll \mu$ (σ-hữu hạn) (tức là nếu $\mu(A) = 0$ thì $P_\theta(A) = 0$ với $A \in \mathcal{A}$), $p(x, \theta) = \frac{dP_\theta(x)}{d\mu(x)}$.

Nếu coi θ là biến ngẫu nhiên Θ với phân phối $\Lambda(\theta)$ thì ta có thể xem $p(x, \theta)$ như hàm mật độ của X với điều kiện $\Theta = \theta$ đã cho. Khi đó ta viết $p(x|\theta)$.

Định nghĩa 1.2.1. (Phân phối tiên nghiệm) Nếu $\Lambda \ll \nu$ trên không gian tham số $\mathcal{O}$ thì $\lambda(\theta) = \frac{d\Lambda}{d\nu}$ là mật độ của Θ. Hàm phân phối Λ hoặc mật độ λ của biến ngẫu nhiên Θ được gọi là phân phối tiên nghiệm của Θ (nếu mật độ $\lambda(\theta|\eta)$ phụ thuộc vào tham số η khác thì η được gọi là siêu tham số).

Định nghĩa 1.2.2 (Phân phối hậu nghiệm) Phân phối có điều kiện của Θ khi $X = x$ đã cho (lưu ý $x = (x_1, \dots, x_n)$) được gọi là phân phối hậu nghiệm, ký hiệu là $\lambda(\theta|x)$, được xác định bởi

$$\lambda(\theta|x) = \frac{p(x|\theta)\lambda(\theta)}{\int_{\mathcal{O}} p(x|\theta)\lambda(\theta)d\nu(\theta)}$$

trong đó $m_\Lambda(x) = \int_{\mathcal{O}} p(x|\theta)\lambda(\theta)d\nu(\theta) = \int_{\mathcal{O}} p(x|\theta)d\Lambda(\theta)$ được gọi là phân phối biên duyên. Trong trường hợp θ là rời rạc, dấu tích phân sẽ được thay bằng tổng.

Bây giờ giả sử hàm tham ẩn cần được ước lượng là $g(\theta)$ và ước lượng là $\hat{g}(X)$ với hàm tổn thất không âm $L\big(g(\theta), \hat{g}(X)\big)$. Khi đó: $R(g, \hat{g}) = EL\big(g(\theta), \hat{g}(X)\big) = \int_{\mathcal{X}} L\big(g, \hat{g}(x)\big)p(x|\theta)d\mu(x)$ được gọi là độ rủi ro (độ mạo hiểm).

Định nghĩa 1.2.3. (Độ rủi ro hậu nghiệm) Độ rủi ro hậu nghiệm của ước lượng $\hat{g}(X)$ được xác định bởi

$$R(g,\hat{g}|x) = E\{L(g(\theta),\hat{g}(X))|x\} = \int_{O} L(g(\theta),\hat{g}(x))\lambda(\theta|x)d\nu(\theta)$$

Định nghĩa 1.2.4. (Ước lượng Bayes) Ước lượng $\hat{g}_\Lambda(X)$ làm cực tiểu độ rủi ro hậu nghiệm, nghĩa là: $R(g,\hat{g}_\Lambda|X) = \inf\limits_{\tilde{g}} R(g,\tilde{g}|X)$ được gọi là ước lượng Bayes của $g(\theta)$.

Định nghĩa 1.2.5. (Độ rủi ro trung bình) Độ rủi ro của ước lượng $\hat{g}$ của $g(\theta)$ khi cho phân phối tiên nghiệm Λ của Θ: $R(\hat{g},\Lambda) = \int_O R(g,\hat{g})d\Lambda(\theta)$ được gọi là độ rủi ro trung bình.

Nhận xét 1.2.6. Ước lượng Bayes làm cực tiểu độ rủi ro trung bình.

Thật vậy, ta có thể viết lại độ rủi ro trung bình dưới dạng

$$\begin{aligned}
R(\hat{g},\Lambda) &= \int_{O} \int_{\mathcal{X}} L(g,\hat{g})p(x|\theta)d\mu(x)\,\lambda(\theta)d\nu(\theta)\\
&= \int_{O} \int_{\mathcal{X}} L(g,\hat{g})m_\Lambda(x)\lambda(\theta|x)d\mu(x)\,d\nu(\theta)\\
&= \int_{\mathcal{X}} m_\Lambda(x)d\mu(x) \int_{O} L(g,\hat{g})\lambda(\theta|x)d\nu(\theta)\\
&= \int_{\mathcal{X}} m_\Lambda(x)R(g,\hat{g}|x)d\mu(x)
\end{aligned}$$

Ký hiệu G là lớp các hàm đo được $\tilde{g}$ là ước lượng của $g(\theta)$. Khi đó theo bổ đề Fatou:

$$\begin{aligned}
\inf_{\tilde{g}\in G} R(\tilde{g},\Lambda) \geq \int_{\mathcal{X}} m_\Lambda(x) \inf_{\tilde{g}\in G} R(g,\tilde{g}|x)d\mu(x) = \int_{\mathcal{X}} m_\Lambda(x)R(g,\hat{g}_\Lambda|x)d\mu(x)\\
= R(\hat{g}_\Lambda,\Lambda) \geq \inf_{\tilde{g}\in G} R(\tilde{g},\Lambda)
\end{aligned}$$

Suy ra $R(\hat{g}_\Lambda,\Lambda) = \inf\limits_{\tilde{g}\in G} R(\tilde{g},\Lambda)$.

Định lý 1.2.7. Với các ký hiệu như trên, giả sử có tồn tại ước lượng $\varphi(X)$ với độ rủi ro hữu hạn.

a. Nếu $L(g, \hat{g}) = (\hat{g} - g)^2$ thì $\hat{g}_\Lambda(x) = E(g(\Theta)|x)$

b. Nếu $L(g, \hat{g}) = w(\theta)(\hat{g} - g)^2$ thì $\hat{g}_\Lambda(x) = \frac{E[w(\Theta)g(\Theta)|x]}{E[w(\Theta)|x]}$

c. Nếu $L(g, \hat{g}) = |\hat{g} - g|$ thì $\hat{g}_\Lambda(x)$ là median của phân phối có điều kiện của Θ với điều kiện $X = x$ đã cho.

Chứng minh.

a. Ta có $E(L(g, \hat{g})|x) = E[(\hat{g} - g)^2|x]$. Xét hàm lồi chặt $\rho(t) = t^2$, với giả thiết tồn

b. tại ước lượng $\varphi(X)$ để độ rủi ro hữu hạn, khi đó $E(L(g, \hat{g})|x)$ có giá trị cực tiểu duy nhất. Hơn nữa

$$E(L(g, \hat{g})|x) = E[(\hat{g} - g)^2|x] = E(g^2|x) + \hat{g}^2 - 2\hat{g}E(g|x)$$

Suy ra $\min_{\hat{g}} E(L(g, \hat{g})|x) \Leftrightarrow \min_{\hat{g}}\left(\hat{g}^2 - 2\hat{g}E(g|x)\right) = -\left(E(g|x)\right)^2$ khi $\hat{g} = E(g|x)$. Như vậy $\hat{g}_\Lambda(x) = E(g(\Theta)|x)$ (đpcm).

c. Chứng minh hoàn toàn tương tự phần a.

d. $E(L(g, \hat{g})|x) = E[|\hat{g} - g||x]$. Xét hàm lồi $\rho(t) = |t|$, khi đó $E[|\hat{g} - g||x]$ lấy giá trị nhỏ nhất trên tập đóng. Mặt khác:

$$E[|\hat{g} - g||x] = \begin{cases} E(|g - m||x) + 2\displaystyle\int_m^{\hat{g}} (\hat{g} - g)d\Lambda(\theta) \text{ nếu } g > m \\ E(|g - m||x) + 2\displaystyle\int_{\hat{g}}^{m} (g - \hat{g})d\Lambda(\theta) \text{ nếu } g < m \end{cases}$$

trong đó m là median của $\lambda(g(\Theta)|x)$. Rõ ràng hai tích phân trên đều dương nếu $\hat{g}$ không trùng với m, còn khi $\hat{g}$ trùng với m thì cách tích phân đó bằng 0. Vậy $E(L(g, \hat{g})|x)$ đạt giá trị nhỏ nhất khi $\hat{g}_\Lambda(x)$ là median của $\lambda(g(\Theta)|x)$.

Ta có thể tóm tắt quá trình tìm ước lượng Bayes cho tham số θ như sau:

1.3.Phân phối tiên nghiệm

Một câu hỏi đặt ra khi xây dựng phân phối hậu nghiệm là chọn phân phối tiên nghiệm như thế nào là hợp lý? Có rất nhiều cách xây dựng phân phối tiên nghiệm, để phục vụ cho mục đích so sánh với thống kê tần suất, luận văn xin đề cập đến hai cách tiếp cận: tiên nghiệm liên hợp (khi có ít thông tin tiên nghiệm) và tiên nghiệm thiếu thông tin (suy ra trực tiếp từ phân phối mẫu).

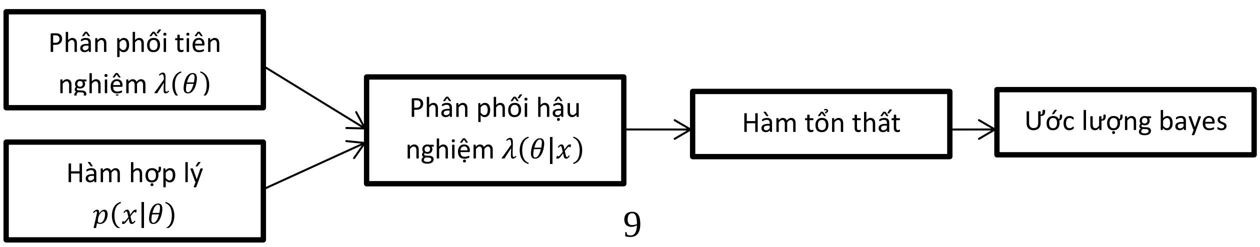
Định nghĩa 1.3..1. (Tiên nghiệm liên hợp) (christan P.robert) Họ $\mathcal{F}$ các phân phối xác suất trên không gian tham $\mathcal{O}$ được gọi là liên hợp cho hàm hợp lý $p(x|\theta)$ nếu phân bố hậu nghiệm $p(\theta|x)$ cũng thuộc $\mathcal{F}$.

Nếu hàm hợp lý là họ mũ thì sẽ tồn tại tiên nghiệm liên hợp với nó. Điều này sẽ giúp ta tìm được xác suất hậu nghiệm mà không gặp khó khăn trong xử lý tích phân.

Một loại tiên nghiệm khác cũng khá thuận tiện trong việc tính toán, đó là tiên nghiệm Jeffreys. Tuy tiên nghiệm này không phải là liên hợp nhưng nó có đặc điểm nổi bật là bất biến đối với các phép đổi biến số. Luận văn xin đề cập đến định nghĩa cho tiên nghiệm này trong trường hợp tham số đơn.

Định nghĩa 1.3.2. (Tiên nghiệm Jeffreys) Tiên nghiệm Jeffreys cho tham số θ được xác định bởi: $\lambda_J(\theta) \propto \sqrt{I(\theta)}$, trong đó $I(\theta)$ là lượng thông tin Fisher của θ.

Sử dụng định nghĩa này, ta dễ dàng chỉ ra được tiên nghiệm Jeffreys bất biến đối với phép đổi biến song ánh φ nào đó. Thật vậy:



$$\lambda_J(\varphi) = \lambda_J(\theta)\left|\frac{d\theta}{d\varphi}\right| \propto \sqrt{I(\theta)\left(\frac{d\theta}{d\varphi}\right)^2} = \sqrt{E\left[\left(\frac{d\ln L}{d\theta}\right)^2\right]\left(\frac{d\theta}{d\varphi}\right)^2}$$

$$= \sqrt{E\left[\left(\frac{d\ln L}{d\theta}\frac{d\theta}{d\varphi}\right)^2\right]} = \sqrt{E\left[\left(\frac{d\ln L}{d\varphi}\right)^2\right]} = \sqrt{I(\varphi)}$$

a. *Phân phối nhị thức $B(n,\pi)$*

- *Tiên nghiệm liên hợp:* Hàm hợp lý của quan sát Y, tổng số "thành công" trong n phép thử, khi cho trước p là $p(x|\pi) = C_n^x \pi^x (1-\pi)^{n-x}$ với $0 \le \pi \le 1$. Dễ thấy $p(x|\pi)$ có dạng hàm $beta(x+1, n-x+1)$. Vậy tiên nghiệm liên hợp sẽ là $\lambda(\pi) = beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}$

Khi đó phân phối hậu nghiệm: $\lambda(\pi|x) = beta(a+x, b+n-x)$

- *Tiên nghiệm Jeffreys:* Lượng thông tin Fisher cho tham số π là

$$I(\pi) = -E_\pi\left[\frac{d^2 \log p(x|\pi)}{d\pi^2}\right] = \frac{n\pi}{\pi^2} + \frac{n - n\pi}{(1-\pi)^2} = \frac{n}{\pi(1-\pi)}$$

Vậy tiên nghiệm Jeffreys cho π là $\lambda_J(\pi) = \sqrt{I(\pi)} \propto \pi^{-\frac{1}{2}}(1-\pi)^{-\frac{1}{2}} = beta\left(\frac{1}{2},\frac{1}{2}\right)$

Tương tự, ta có được các kết quả tương ứng cho các phân bố quen thuộc sau.

b. *Phân phối Poisson*

Giả sử $x = (x_1, \dots, x_n)$ là mẫu ngẫu nhiên rút từ phân bố Pois(μ), khi đó hàm hợp lý sẽ là $p(x|\mu) = \prod_{i=1}^n p(x_i|\mu) \propto \mu^{\sum x_i} e^{-n\mu} = gamma(r,v)$ với $r = \sum x_i + 1, v = n$

- Tiên nghiệm liên hợp: $\lambda(\mu) = gamma(a,b) \Rightarrow \lambda(\mu|y) \propto gamma(r', v')$ với $r' = a + \sum x_i, v' = b + n$.

- Tiên nghiệm Jeffreys: $\lambda_J(\mu) \propto \frac{1}{\sqrt{\mu}} \Rightarrow \lambda(\mu|y) \propto gamma(r', v')$ với $r' = \frac{1}{2} + \sum x_i, v' = n$

c. *Phân phối chuẩn (với giá trị trung bình μ chưa biết)*

Giả sử $x = (x_1, \dots, x_n)$ là mẫu ngẫu nhiên rút từ phân bố chuẩn $N(\mu, \sigma^2)$, khi đó hàm hợp lý là $p(x|\mu) \propto \exp\left[-\frac{1}{\frac{2\sigma^2}{n}}(\bar{x} - \mu)^2\right]$

- Tiên nghiệm liên hợp: $\lambda(\mu) \sim N(m, s^2) \Rightarrow \lambda(\mu|x) \propto N(m', s'^2)$ với $m' = \frac{\sigma^2 m + s^2 y}{\sigma^2 + s^2}, s'^2 = \frac{\sigma^2 s^2}{\sigma^2 + s^2}$.
- Tiên nghiệm Jeffreys: $\lambda_J(\mu) = 1$.

CHƯƠNG 2. KHOẢNG TIN CẬY TRONG THỐNG KÊ BAYES

2.1. Khoảng tin cậy Bayes

Định nghĩa 2.1.1. Khoảng tin cậy Bayes $100(1-\alpha)\%$ là khoảng (a,b) sao cho:

$$P(a<\theta<b|x)=\int_a^b \lambda(\theta|x)d\nu(\theta)\geq 1-\alpha$$

trong đó $\lambda(\theta|x)$ là phân phối xác suất hậu nghiệm của θ với dữ liệu $x=(x_1,\dots,x_n)$ (nếu θ là rời rạc thì thay dấu tích phân bằng tổng).

Với thống kê Bayes, do tham số là biến ngẫu nhiên nên hoàn toàn hợp lý khi ta nói về phân bố xác suất của tham số. Nghĩa là, nếu (a,b) là khoảng tin cậy Bayes $100(1-\alpha)\%$ cho tham số θ thì xác suất để θ nằm trong khoảng (a,b) là $1-\alpha$, thống kê tần suất không cho ta kết luận này.

Theo định nghĩa, ta có thể tìm được rất nhiều khoảng tin cậy Bayes. Luận văn xin đề cập đến hai loại chính: khoảng tin cậy Bayes đối xứng và khoảng tin cậy Bayes chứa mật độ hậu nghiệm cao nhất (khoảng HPD).

2.2. Khoảng tin cậy Bayes đối xứng

Cách dễ nhất để tính khoảng tin cậy Bayes cho tham số là sử dụng lượng quantile hậu nghiệm, thường cho khoảng tin cậy đối xứng (equal-tail interval). Để có khoảng tin cậy Bayes đối xứng $100(1-\alpha)\%$, ta tiến hành tìm các số $\theta_{\frac{\alpha}{2}}<\theta_{1-\frac{\alpha}{2}}$ thỏa mãn:

i. $P\left(\theta<\theta_{\frac{\alpha}{2}}\middle|X=x\right)=\frac{\alpha}{2}$

ii. $P\left(\theta>\theta_{1-\frac{\alpha}{2}}|X=x\right)=\frac{\alpha}{2}$

Tức là, các số $\theta_{\frac{\alpha}{2}},\theta_{1-\frac{\alpha}{2}}$ lần lượt là $\frac{\alpha}{2}$ và $\left(1-\frac{\alpha}{2}\right)$ quantile hậu nghiệm của θ. Khi đó:

$$P\left(\theta \in \left[\theta_{\frac{\alpha}{2}}, \theta_{1-\frac{\alpha}{2}}\right] \middle| X = x\right) = 1 - P\left(\theta \notin \left[\theta_{\frac{\alpha}{2}}, \theta_{1-\frac{\alpha}{2}}\right] \middle| X = x\right)$$

$$= 1 - P\left(\theta < \theta_{\frac{\alpha}{2}} \middle| X = x\right) - P\left(\theta > \theta_{1-\frac{\alpha}{2}} | X = x\right) = 1 - \alpha$$

Ví dụ 2.2.1. (Peter D.Hoff) Vào những năm 1990, Khảo sát xã hội đã thu thập thông tin về trình độ học vấn và số con của 155 phụ nữ ở độ tuổi 40, để tìm hiểu xem liệu trình độ học vấn có ảnh hưởng đến số con sinh ra hay không. Ta gọi $X_1 = (X_{11}, X_{12}, \dots, X_{1n_1})$ và $X_2 = (X_{21}, X_{22}, \dots, X_{2n_2})$ là dữ liệu cho số con của nhóm phụ nữ có trình độ học vấn dưới và ngang (hoặc trên) đại học. Giả sử $X_1 \sim \text{Pois}(\theta_1), X_2 \sim \text{Pois}(\theta_2)$. Theo khảo sát ta có số liệu như sau:

- Học vấn dưới đại học: $n_1 = 111, \sum_{i=1}^{n_1} X_{1i} = 217, \overline{X_1} = 1.95$
- Học vấn ngang hoặc trên đại học: $n_2 = 44, \sum_{i=1}^{n_2} X_{2i} = 66, \overline{X_2} = 1.50$

Hàm hợp lý tương ứng là: $p(x_1|\theta_1) \propto \theta^{\sum_{i=1}^{n_1} x_{1i}} e^{-n_1\theta_1}$ và $p(x_2|\theta_2) \propto \theta^{\sum_{i=1}^{n_2} x_{2i}} e^{-n_2\theta_2}$

Giả sử phân phối tiên nghiệm cho θ_1 và θ_2 đều là gamma$(\alpha = 2, \beta = 1) = \frac{1}{2}\theta e^{-\theta}$, khi đó ta có phân phối hậu nghiệm là:

$$p(\theta_1|x) \sim \text{gamma}(2 + 217, 1 + 111) = \text{gamma}(219, 112)$$

$$p(\theta_2|x) \sim \text{gamma}(2 + 66, 1 + 44) = \text{gamma}(68, 45)$$

Dựa vào phân phối hậu nghiệm này, ta dễ dàng tìm được khoảng tin cậy Bayes đối xứng 95% cho θ_1 và θ_2:

```
> qgamma(c(.025,.975),219,112)
[1] 1.704943 2.222679
> qgamma(c(.025,.975),68,45)
[1] 1.173437 1.890836
```

Như vậy, ta có thể kết luận rằng, theo dữ liệu thu thập được thì có tới 95% khả năng phụ nữ có trình độ học vấn ngang (hoặc trên) đại học sẽ có ít hơn 2 con.

2.3. Ứng dụng khoảng tin cậy Bayes HPD

Có thể thấy rằng, thuận lợi của việc tính khoảng tin cậy Bayes đối xứng (dựa trên giá trị quantile) là công việc rất dễ dàng (có thể tính tay hoặc sử dụng phương pháp Monte Carlo xích Markov), hơn nữa có nét giống cách xây dựng khoảng tin cậy tần suất. tuy nhiên, kết quả này chỉ hữu ích khi phân bố hậu nghiệm của tham số là đối xứng. Trong trường hợp phân bố không đối xứng, lựa chọn tối ưu hơn là khoảng tin cậy Bayes HPD (khoảng chứa mật độ hậu nghiệm cao nhất).

Định nghĩa 2.3.1. Khoảng HPD $100(1-\alpha)\%$ cho tham số θ là tập con $R(k_\alpha)$ của không gian tham O được xác định bởi: $R(k_\alpha) = \{\theta: \lambda(\theta|x) \geq k_\alpha\}$, trong đó k_α là số lớn nhất sau cho $\int_{\theta:\lambda(\theta|x)\geq k_\alpha} \lambda(\theta|x)dv(\theta) \geq 1-\alpha$.

Số k_α có thể coi như đường thẳng song song với trục tham số, có giao điểm với đường mật độ hậu nghiệm của θ sao cho diện tích khoảng giữa hai điểm đó là $(1-\alpha)$. Theo Box&Tiao (1992), khoảng HPD như thế có hai tính chất:

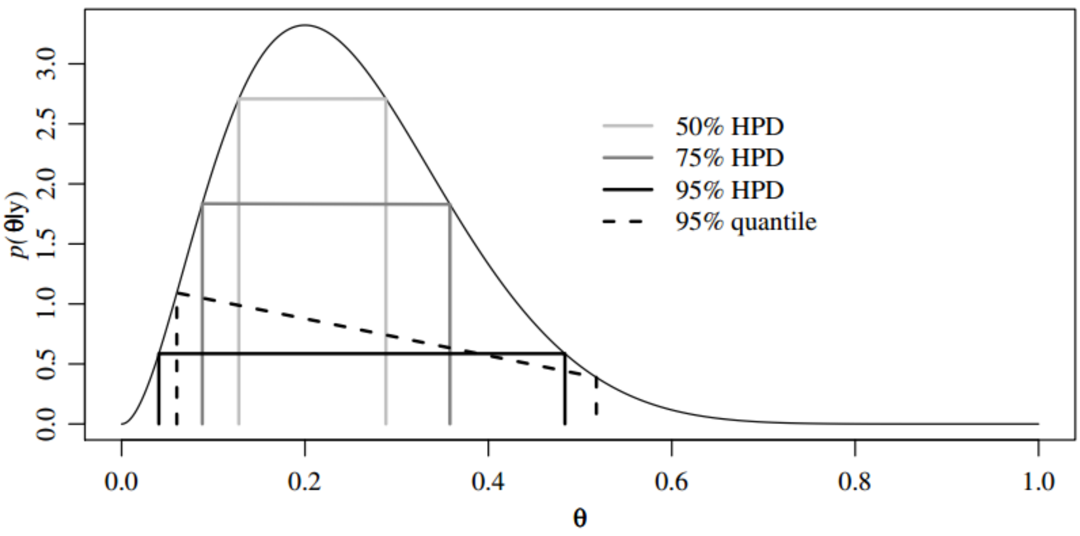
- Mật độ mọi điểm nằm trong khoảng HPD đều lớn hơn các điểm nằm ngoài khoảng này: nếu $\theta_a \in C$ và $\theta_b \notin C$ thì $p(\theta_a|X=x) > p(\theta_b|X=x)$.
- Với xác suất cho trước (chính là $1-\alpha$), khoảng có độ dài ngắn nhất.

Như vậy trong trường hợp phân bố đối xứng thì khoảng HPD trùng với khoảng tin cậy Bayes đối xứng.

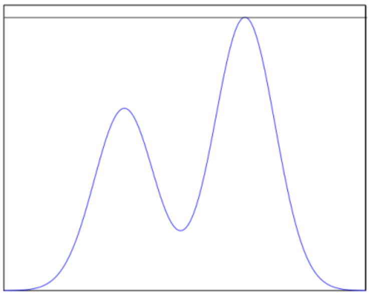
Ví dụ 2.3.1. . (Peter D.Hoff) Giả sử trong số 10 phép thử độc lập của biến ngẫu nhiên X tuân theo phân bố Bernoulli(θ), ta quan sát được $X = 2$. Sử dụng phân bố tiên nghiệm cho θ là phân phối đều (tức là phân bố beta$(1,1)$), ta nhận được phân bố hậu nghiệm cho θ là: $p(\theta|X=x) \sim$ beta$(1+2, 1+8)$. Bằng phần mềm R, ta tính được khoảng tin cậy Bayes đối xứng 95% cho θ là $(0.06, 0.52)$:

```
> qbeta(c(.025,.975),3,9)
[1] 0.06021773 0.51775585
```

Còn khoảng tin cậy HPD 95% tương ứng là (0.04,0.48), hẹp hơn so với khoảng đối xứng 95%:



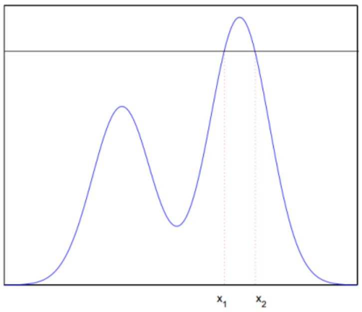
Hình ảnh trên cho ta so sánh giữa khoảng tin cậy Bayes đối xứng 95% (dựa trên giá trị quantile) và khoảng HPD 95% cũng như các khoảng HPD khác. Dễ thấy với khoảng đối xứng 95%, vẫn còn các điểm nằm ngoài khoảng mà có mật độ lớn hơn các điểm trong khoảng, nhưng với khoảng HPD 95%, mọi điểm nằm ngoài khoảng đều có mật độ nhỏ hơn các điểm trong khoảng. Điều này chính là ưu điểm của khoảng HPD.

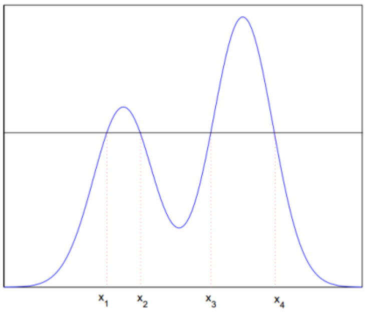


Trong trường hợp phân bố hậu nghiệm không là unimodal, khoảng HPD cho tham số có thể bị tách thành nhiều khoảng, ta gọi là vùng HPD. Cách tìm vùng HPD cho phân bố multimodal có thể được miêu tả như sau:

- Xét phân bố bimodal như hình vẽ (trường hợp unimodal xác định tương tự), sử dụng đường song song với trục giá trị tham số để xác định giá trị mode hậu nghiệm.

- Tịnh tiến dần đường thẳng này xuống dưới giá trị mode hậu nghiệm, xác định các giao điểm với phân bố hậu nghiệm x_1 và x_2, tính diện tích phần dưới đường phân bố và giới hạn bởi 2 điểm x_1, x_2, tịnh tiến đường thẳng đến khi diện tích này đạt giá trị $(1-\alpha)$ thì dừng lại. Khi đó đoạn (x_1, x_2) tương ứng chính là khoảng HPD.
- Chú ý với phân bố multimodal, khoảng HPD có thể không liên tục (có thể gồm các khoảng rời nhau (x_1, x_2) và (x_3, x_4)), khi đó ta có vùng HPD cho tham số.





Tuy nhiên, một nhược điểm của khoảng HPD là rất khó tìm được bằng biến đổi giải tích mà thường phải dùng phương pháp số, đưa ra kết quả xấp xỉ (trừ trường hợp phân bố hậu nghiệm rất đơn giản, ví dụ như phân bố chuẩn tắc). Tanner (1996) đưa ra thuật toán Monte Carlo để tính toán các giới hạn của vùng HPD (trường hợp phân bố multimodal), nhưng yêu cầu đánh giá phân bố hậu nghiệm biên duyên tương đối phức tạp và thiên về tin học. Chen&Shao (1998) đã đưa ra phương pháp Monte Carlo đơn giản để ước lượng khoảng HPD bằng hai cách, tạo mẫu xích Markov từ phân bố hậu nghiệm biên duyên của tham số và phương pháp lấy mẫu trọng số.

KẾT LUẬN

Báo cáo đã trình bày một số khái niệm về thống kê Bayes và bài toán ước lượng tham số. Cung cấp các khái niệm về phân phối tiên nghiệm, phân phối hậu nghiệm, độ rủi ro hậu nghiệm, độ rủi ro trung bình Bayes, cách tìm khoảng tin cậy Bayes.

Việc tìm khoảng tin cậy Bayes đối xứng là công việc tương đối đơn giản (dựa trên giá trị quantile có thể tính tay hoặc sử dụng phương pháp Monte Carlo xích Markov), hơn nữa có nét giống cách xây dựng khoảng tin cậy tần suất, kết quả này là hữu ích khi phân bố hậu nghiệm của tham số là đối xứng. Trong trường hợp phân bố không đối xứng, lựa chọn tối ưu hơn là khoảng tin cậy Bayes HPD (khoảng chứa mật độ hậu nghiệm cao nhất), trong báo cáo cũng cách tìm khoảng tin cậy Bayes HPD này.

TÀI LIỆU THAM KHẢO

[1]. Đào Hữu Hồ, 1999. *Xác suất thống kê*. In lần thứ 3, Nhà xuất bản Đại học Quốc gia Hà Nội.

[2]. Nguyễn Văn Hữu, Đào Hữu Hồ, Hoàng Hữu Như, 2004. *Thống kê toán học*. Nhà xuất bản Đại học Quốc gia Hà Nội.

[3]. Nguyễn Viết Phú, Nguyễn Duy Tiến, 2004. *Cơ sở lý thuyết xác suất.* Nhà xuất bản Đại học Quốc gia Hà Nội.

[4]. Tống Đình Quỳ, 2003. *Giáo trình Xác suất thống kê*, Nhà xuất bản Đại học Quốc gia Hà Nội.

[5]. E.L. Lehmann, Joseph P. Romano, (2005). *Testing Statistical Hypotheses*. Springer, USA.

[6]. McCulloch J., 1996. Financial applications of stable districbutions, in: Handbook of Statistics, Vol. 14 (eds. G. Maddala and C. Rao) (Elsevier Science/North-Holland, Amsterdam), 393-397.

[7]. Nolan J., 2002. *Maximum likelihood estimaton and diagnostics for stable distributions*. American University, Washington.

[8]. Nolan J., 2005. *Stable Distributions Models for Heavy Tailed Data*. American University, Washington.