ỨNG DỤNG ƯỚC LƯỢNG BAYESIAN PHÂN TÍCH VỀ TỶ LỆ NGHÈO CỦA CÁC TỈNH Ở VIỆT NAM

Dominique Haughton[*]
Nguyễn Phong[**]

Vấn đề ước lượng tham số nhị thức (binomial parameter) đã thu hút sự quan tâm của các nhà thống kê và những người khác trong lĩnh vực ước lượng tỷ trọng. Mọi người đều biết rằng khi ước lượng các số tỷ trọng thường cần một cỡ mẫu lớn để đạt được độ chính xác chấp nhận được của ước lượng. Các ước lượng cỡ mẫu thường dựa vào cách tính toán cổ điển về khoảng tin cậy, đôi khi được điều chỉnh theo các thiết kế mẫu đặc biệt. Công trình nghiên cứu gần đây của Brown (2001) đã tập trung sự chú ý vào những khiếm khuyết của những khoảng tin cậy này, đặc biệt vào thực tế là trong nhiều trường hợp "khoảng tin cậy 95%" có độ bao phủ ít hơn 95%.

Ước lượng tỷ lệ nghèo là ước lượng tham số nhị thức vì tỷ lệ nghèo nói chung được xác định là tỷ trọng hộ có mức chi tiêu dùng bình quân đầu người hàng năm thấp hơn chuẩn nghèo cho trước. Trong phần lớn nội dung của bài viết này, chúng tôi giả sử rằng chuẩn nghèo này là không ngẫu nhiên và việc xác định hộ nghèo hay không nghèo được thực hiện chính xác. Chúng ta sẽ thảo luận những hàm ý của việc xác định không chính xác chuẩn nghèo trong phần sau của bài viết.

Ước lượng tỷ lệ nghèo cho các tỉnh của Việt Nam rất phù hợp với phân tích Bayesian: thông tin tiên nghiệm (ví dụ ý kiến chuyên gia về tỷ lệ nghèo) thường đã có sẵn và cỡ mẫu ở cấp tỉnh có xu hướng khá nhỏ do điều tra lớn thường tốn kém và bị ảnh hưởng bởi sai số phi chọn mẫu. Những nhà thống kê chọn mẫu và những người khác có liên quan đến thiết kế và phân tích những cuộc điều tra như vậy (ở Việt Nam hoặc ở nơi khác) cho đến nay không sử dụng phân tích Bayesian với tỷ lệ nghèo (xem Glewwe và Yansaneh về một phân tích đặc thù trong lĩnh vực này). Trong bài này chúng tôi sẽ chỉ ra rằng mức độ chính xác của các ước lượng sẽ cao hơn khi sử dụng những thông tin tiên nghiệm hợp lý và chuẩn nghèo đã biết. Chúng tôi sẽ minh hoạ kết quả này bằng một mẫu ở thành thị giàu hơn (thành phố Hồ Chí Minh) và một mẫu ở nông thôn nghèo hơn (Nghệ An). Tuy nhiên, để đảm bảo có được kết quả này, một điều cần lưu ý là khi sự xác định nghèo/không nghèo sai xảy ra, mà trong thực tế rất có thể xảy ra, độ bao phủ trung bình của khoảng xác suất rộng 4 điểm % không đạt tới 0,95, ngay cả với cỡ mẫu lớn, trong khi nó có thể đạt tới 0,95 đối với khoảng xác suất rộng 8 điểm %.

Ước lượng Bayesian về tỷ lệ nghèo khi chuẩn nghèo đã biết

Ở khu vực thành thị của thành phố Hồ Chí Minh, mẫu trong Khảo sát Mức sống Dân cư Việt Nam năm 1998 có 433 hộ, trong đó có 2 hộ nghèo. Các tính toán có quyền số thông thường (frequentist) (theo quyền số chọn mẫu) cho tỷ lệ nghèo là 0,00462, với độ lệch chuẩn là 0,00334 (hệ

[*] Đại học Bentley, Hoa Kỳ, [**] Tổng cục Thống kê - Việt Nam

số biến thiên khoảng 0,7). Để thực hiện phân tích Bayesian, chúng tôi sử dụng hỗn hợp các phân phối bêta làm tiên nghiệm cho tỷ lệ nghèo chưa biết, được gợi ý trong kỹ thuật của Nandram và Sedransk. Điều này đã được chứng minh trong công trình nghiên cứu của Dalal và Hall (1983), rằng bất kỳ tiên nghiệm nào cũng có thể là xấp xỉ bởi hỗn hợp như vậy. Sau đó chúng tôi áp dụng công thức dạng đóng của Nandram và Sedransk đối với giá trị trung bình hậu nghiệm và độ lệch chuẩn hậu nghiệm của tỷ lệ nghèo đối với thiết kế mẫu chùm hai bước. Trong trường hợp của nước ta, chúng tôi giả sử rằng xã/phường được chọn ngẫu nhiên, sau đó hộ được chọn ngẫu nhiên từ xã/phường. Trong thực tế có thêm một bước trong thiết kế mẫu: thôn/cụm được chọn ngẫu nhiên từ xã/phường, sau đó hộ được chọn ngẫu nhiên từ thôn/cụm. Chúng tôi dự kiến sẽ đề cập đến vấn đề chọn mẫu chùm ba bước trong tương lai, trong đó sẽ không có sẵn công thức dạng đóng nào đối với giá trị trung bình và độ lệch chuẩn hậu nghiệm của tỷ lệ nghèo. Mặc dù vậy, chúng tôi không nghĩ rằng việc thêm một bước chọn mẫu nữa sẽ tạo ra sự khác biệt lớn và phân tích này sẽ xấp xỉ với thực tế. Tiếp theo, chúng tôi sử dụng WINBUGS để mô phỏng phân phối hậu nghiệm, với các lệnh (code) được công bố trong Congdon (2001; ví dụ 5.18 trang 196). Ngoài số liệu về hộ nghèo/không nghèo từ những xã/phường được điều tra, phân tích này còn sử dụng số liệu về số hộ trong từng xã/phường đó tương ứng của khu vực thành thị thành phố Hồ Chí Minh và khu vực nông thôn Nghệ An. Mô hình trong phân tích này tính tỷ lệ nghèo riêng của từng xã/phường và sau đó kết hợp những tỷ lệ nghèo này vào tỷ lệ nghèo chung của tỉnh.

Trong Bảng 1 và Hình 1, chúng tôi trình bày kết quả từ bốn tiên nghiệm khác nhau cho khu vực thành thị của thành phố Hồ Chí Minh. Trong Bảng 2 và Hình 2, chúng tôi trình bày kết quả từ hai tiên nghiệm khác nhau cho khu vực nông thôn của Nghệ An. Các giá trị trung bình và độ lệch chuẩn hậu nghiệm là của tỷ lệ nghèo chung của toàn khu vực (khu vực thành thị của thành phố Hồ Chí Minh và khu vực nông thôn Nghệ An). Hỗn hợp các phân phối bêta được sử dụng làm tiên nghiệm cho véc tơ θ của N tỷ lệ nghèo của N xã/phường được Nandram và Sedransk đưa ra như sau:

$$\pi(\theta \mid \tau) = \sum_{r=1}^{R} \omega_r B(a_r, \tau - a_r)^{-N} \prod_{k=1}^{N} \theta_k^{a_r - 1}(1-\theta_k)^{\tau - a_r - 1},$$

Trong đó θ_k là tỷ lệ nghèo của tỉnh thứ k, B là ký hiệu chỉ hàm Bêta. Giá trị của w_r, a_r và τ phải được lựa chọn khi suy ra tiên nghiệm. Lưu ý rằng các giá trị trung bình của các phân phối bêta trong hỗn hợp là a_r/τ, và do đó giá trị của τ điều khiển độ lệch chuẩn của các phân phối bêta: τ càng cao, độ lệch chuẩn càng nhỏ.

Hai tiên nghiệm đầu tiên cho khu vực thành thị của thành phố Hồ Chí Minh được căn cứ một cách lỏng lẻo vào tỷ lệ nghèo và độ lệch chuẩn của chúng ở các tỉnh Việt Nam được Baulch và các đồng sự ước lượng, sử dụng số liệu từ Tổng điều tra Dân số và Nhà ở năm 1999 và các phương trình hồi quy dựa trên các số liệu của Khảo sát mức sống dân cư. Những ước lượng này được sử dụng để định nghĩa 4 beans tập trung vào các giá trị được trình bày trong cột "Giá trị trung bình" trong Bảng 1 cho riêng 4

cấu thành và các xác suất tiên nghiệm của mỗi bin tương ứng là 0,07; 0,43; 0,43; 0,07. Lưu ý rằng giá trị bằng 4 của R được chọn phần nào mang tính áp đặt nhằm tiện lợi và linh hoạt trong tính toán. Tiên nghiệm 1 và 2 khác nhau bởi giá trị của τ, và vì vậy cũng bởi độ lệch chuẩn. Các cấu thành trong tiên nghiệm 2 ít tách rời nhau, như trong Hình 1. Kết quả từ cả hai tiên nghiệm là gần nhau, tỷ lệ nghèo hậu nghiệm khoảng 0,01 với độ lệch chuẩn khoảng 0,005, là một sự cải thiện (hệ số biến thiên khoảng 0,5) so với ước lượng thông thường. Hình 1 cho thấy hai mật độ hậu nghiệm từ tiên nghiệm 1 và 2 gần nhau, và cho phần lớn xác suất hậu nghiệm thành hai bộ phận tương ứng với xã giàu hơn và xã nghèo hơn. Tiên nghiệm 3 tương ứng với tiên nghiệm từ ý kiến chuyên gia (của một trong số các tác giả) rằng "chúng tôi chắc chắn đến 95% rằng tỷ lệ nghèo của khu vực thành thị của thành phố Hồ Chí Minh nằm trong khoảng 0,01 đến 0,03". Giống như với tiên nghiệm 1 và 2, 4 bin cũng được tạo ra cho tiên nghiệm 3, tập trung vào giá trị được đưa ra trong Bảng 1 và với độ rộng phù hợp với tiên nghiệm theo ý kiến của chuyên gia. Những thống kê tóm tắt của tỷ lệ nghèo hậu nghiệm rất gần với kết quả theo tiên nghiệm 1 và 2. Tiên nghiệm 4 là tiên nghiệm rất tản mạn, và trong trường hợp này, tỷ lệ nghèo hậu nghiệm không chính xác (độ lệch chuẩn là 0,008) như dự kiến.

Trong trường hợp này, chúng tôi có cả các biểu thức dạng đóng của giá trị trung bình và độ lệch chuẩn hậu nghiệm, lẫn khả năng sử dụng WINBUGS để tạo ra một mẫu từ hậu nghiệm. Kết quả từ cả hai phân tích này dự kiến là gần nhau, và thực tế đúng như vậy. Chúng tôi lưu ý ở đây rằng chúng tôi phát hiện ra nếu các cấu thành bêta quá tách rời hoặc nếu một trong các cấu thành quá gần 0 thì chuỗi MCMC trong WINBUGS có thể bị "tắc" ở một cấu thành và cho giá trị trung bình hậu nghiệm không đúng. Vấn đề này trong thực tế không gây ngạc nhiên cho những tác giả của WINBUGS (N. Best, personal communication), và có thể được khắc phục bằng cách kiểm tra những kết quả của WINBUGS dựa vào công thức dạng đóng dùng cho thiết kế mẫu chùm hai bước đối với một tiên nghiệm đã cho, và sau chuyển sang những thiết kế điều tra phức tạp hơn nếu cần thiết.

Đối với khu vực nông thôn tỉnh Nghệ An có 225 hộ trong mẫu, trong đó có 110 hộ nghèo. Các ước lượng có quyền số thông thường cho tỷ lệ nghèo là 0,489 với độ lệch chuẩn là 0,104. Tiên nghiệm 1 lại căn cứ một cách lỏng lẻo vào các ước lượng của Baulch và các đồng sự; đưa ra giá trị trung bình hậu nghiệm của tỷ lệ nghèo là 0,5 với độ lệch chuẩn là 0,05, là một sự cải thiện về độ chính xác so với phân tích thông thường. Tiên nghiệm 2 căn cứ vào tỷ lệ nghèo là 0,2 theo ước tính của Bộ Lao động Thương binh Xã hội để tạo ra 4 bin với độ rộng như trong tiên nghiệm 1. Tỷ lệ nghèo tiên nghiệm là 0,2 có thể quá thấp, và thật thú vị khi xem phân tích Bayesian sử dụng số liệu để sửa thông tin tiên nghiệm này: chuỗi MCMC tập trung gần như riêng vào một bộ phận cao hơn để đưa ra giá trị trung bình hậu nghiệm của tỷ lệ nghèo là 0,42 với độ lệch chuẩn khoảng 0,01.

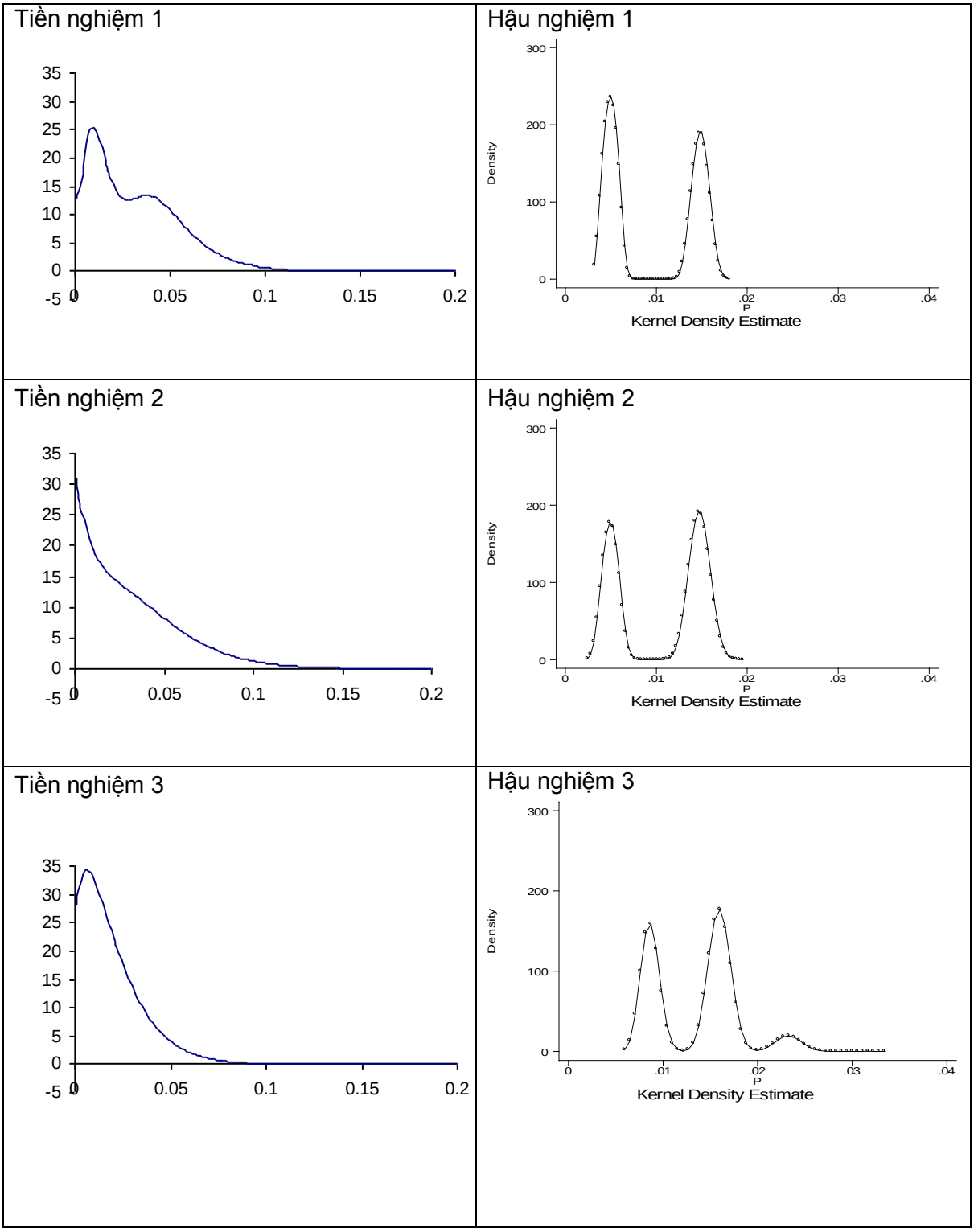
BẢNG 1: GIÁ TRỊ TRUNG BÌNH VÀ ĐỘ LỆCH CHUẨN TIÊN NGHIỆM VÀ HẬU NGHIỆM CỦA KHU VỰC THÀNH THỊ CỦA THÀNH PHỐ HỒ CHÍ MINH

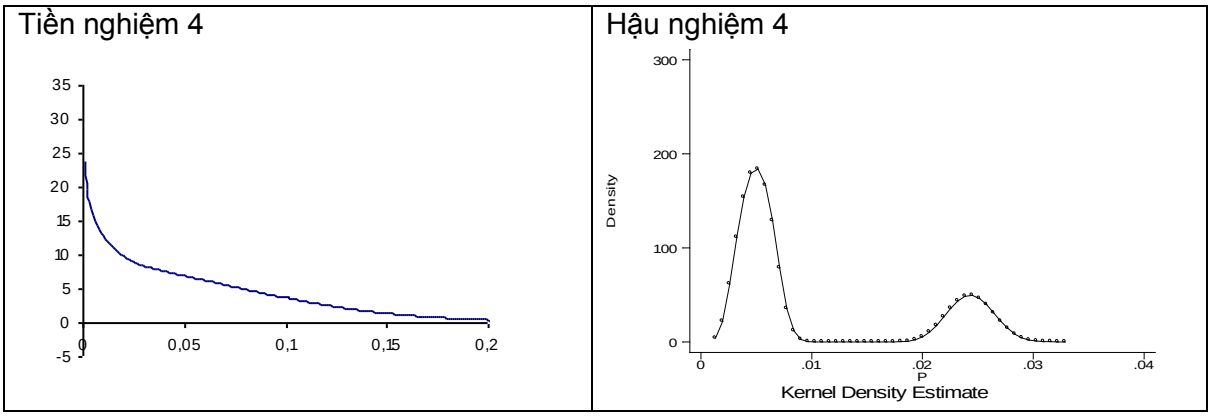
W_i		**Tiên nghiệm 1 τ=200**		**Tiên nghiệm 2 τ=80**		**Tiên nghiệm 3 τ=80**		**Tiên nghiệm 4 τ=40**	
		Giá trị TB	***Độ lệch chuẩn***	***Giá trị TB***	***Độ lệch chuẩn***	***Giá trị TB***	***Độ lệch chuẩn***	***Giá trị TB***	***Độ lệch chuẩn***
0,07	Cấu thành 1	0,005	0,005	0,005	0,008	0,009	0,010	0,005	0,011
0,43	Cấu thành 2	0,015	0,009	0,015	0,014	0,016	0,014	0,025	0,024
0,43	Cấu thành 3	0,045	0,015	0,045	0,023	0,024	0,017	0,080	0,042
0,07	Cấu thành 4	0,075	0,019	0,075	0,029	0,031	0,019	0,140	0,054
	Tổng	0,031	0,023	0,031	0,027	0,020	0,017	0,055	0,051
		Giá trị TB hậu nghiệm	***Độ lệch chuẩn hậu nghiệm***	***Giá trị TB hậu nghiệm***	***Độ lệch chuẩn hậu nghiệm***	***Giá trị. TB hậu nghiệm***	***Độ lệch chuẩn hậu nghiệm***	***Giá trị TB hậu nghiệm***	***Độ lệch chuẩn hậu nghiệm***
	Dạng đóng	0,009872	0,004982	0,010765	0,004911	0,013684	0,004561	0,008841	0,007801
	Winbugs	0,009664	0,004964	0,010611	0,004910	0,013530	0,004508	0,010130	0,008632

BẢNG 2: GIÁ TRỊ TRUNG BÌNH VÀ ĐỘ LỆCH CHUẨN TIÊN NGHIỆM VÀ HẬU NGHIỆM CỦA KHU VỰC NÔNG THÔN TỈNH NGHỆ AN

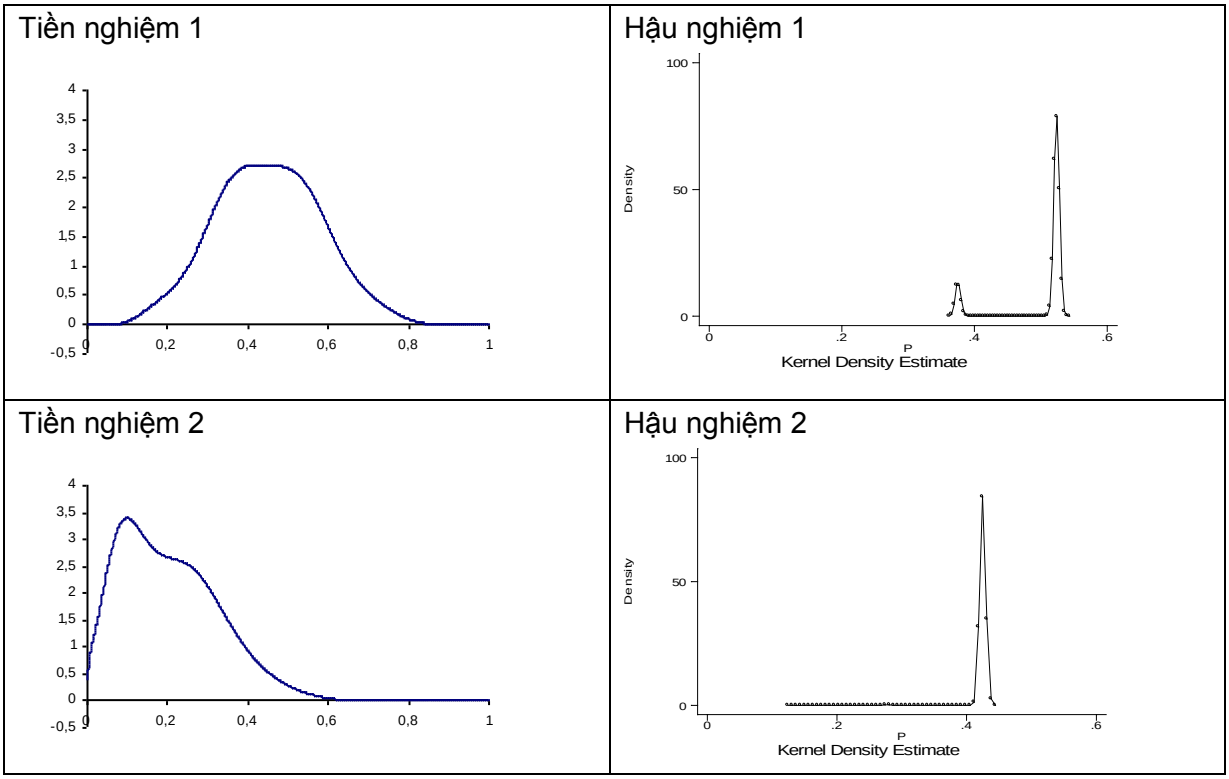
w_i		**Tiên nghiệm 1 τ=40**		**Tiên nghiệm 2 τ=30**	
		Giá trị trung bình	***Độ lệch chuẩn***	***Giá trị trung bình***	***Độ lệch chuẩn***
0,07	Cấu thành 1	0,225	0,065	0,050	0,039
0,43	Cấu thành 2	0,375	0,076	0,125	0,059
0,43	Cấu thành 3	0,525	0,078	0,275	0,080
0,07	Cấu thành 4	0,675	0,073	0,425	0,089
	Tổng	0,450	0,133	0,205	0,122
		Giá trị trung bình hậu nghiệm	***Độ lệch chuẩn hậu nghiệm***	***Giá trị trung bình hậu nghiệm***	***Độ lệch chuẩn hậu nghiệm***
	Dạng đóng	0,499810	0,055138	0,424697	0,008203
	Winbugs	0,503400	0,051560	0,424500	0,009934

HÌNH 1: MẬT ĐỘ KERNEL TIỀN NGHIỆM VÀ HẬU NGHIỆM;
KHU VỰC THÀNH THỊ TP. HỒ CHÍ MINH





HÌNH 2: MẬT ĐỘ KERNEL TIỀN NGHIỆM VÀ HẬU NGHIỆM; KHU VỰC NÔNG THÔN NGHỆ AN



Ước lượng Bayesian với trường hợp có sự phân loại sai

Chúng ta xét tình huống có sai số trong việc xác định hộ nghèo và hộ không nghèo (có sự phân loại sai). Nguyên nhân là do trong thực tế khó tính chính xác chuẩn nghèo mà phần lớn là do khó thu thập chính xác số liệu về giá của các hàng hoá cơ bản. Vấn đề xác định hộ nghèo cũng tương tự như trong chẩn đoán bệnh dựa trên những xét nghiệm không hoàn hảo. Ở đây chúng tôi sử dụng công trình nghiên cứu của Rahme và đồng sự ứng dụng vào lĩnh vực y tế trong đó thực hiện xác định cỡ mẫu

Bayesian cho tham số nhị thức với giả thiết có sự phân loại sai. Trong phạm vi vấn đề này, việc kiểm tra đối với tình trạng nghèo có độ nhạy (xác suất của một hộ nghèo được phân loại là nghèo) và sự đặc trưng (xác suất của một hộ không nghèo được phân loại là không nghèo), với phân phối tiên nghiệm bêta dựa theo Rahme và đồng sự, và tỷ lệ nghèo cũng được cho phân phối tiên nghiệm bêta. Minh hoạ cách tiếp cận này với khu vực nông thôn Nghệ An, với định nghĩa phân phối tiên nghiệm bêta cho tỷ lệ nghèo có các tham số α=70,32 và β=77,1 dựa trên cơ sở của các ước lượng cho tỷ lệ nghèo và độ lệch chuẩn của nó trong công trình nghiên cứu của Baulch và đồng sự. Lấy các phân phối bêta làm những tiên nghiệm cho độ nhạy và sự đặc trưng của phân loại nghèo/không nghèo từ ý kiến cho rằng độ nhạy trung bình sự đặc trưng cùng là khoảng 0,95 và chúng ta chắc chắn 95% rằng độ nhạy và sự đặc trưng nằm trong khoảng 0,9 đến 1. Ý kiến này cho tham số bêta những giá trị nêu trong bảng 3.

Bảng 3 trình bày các độ bao phủ trung bình của khoảng xác suất đối với hai độ rộng khoảng khác nhau và ba cỡ mẫu khác nhau, được tính toán từ chương trình S-plus và được Rahme và đồng sự công bố. Rõ ràng là độ bao phủ không đạt đến 0,95 đối với độ rộng bằng 4 điểm %, ngay cả với những cỡ mẫu rất lớn. Độ bao phủ như vậy là khả thi với khoảng có độ rộng là 0,08 với những cỡ mẫu lớn. Tuy nhiên, chúng tôi lưu ý rằng những kỹ thuật của Rahme và đồng sự giả thiết rằng các mẫu là độc lập và được tạo ra từ cùng một phân phối (independent identically distributed sample - mẫu iid), vì vậy tình hình có thể xấu đi khi sử dụng một thiết kế điều tra phức tạp hơn.

BẢNG 3: ĐỘ BAO PHỦ TRUNG BÌNH CỦA KHOẢNG XÁC SUẤT ĐỐI VỚI TỶ LỆ NGHÈO CỦA KHU VỰC NÔNG THÔN NGHỆ AN VỚI GIẢ THIẾT MẪU IID.

$\alpha_{sens}=\alpha_{spec}$=71,25; $\beta_{sens}=\beta_{spec}$=3,75; α=70,32; β=77,1		
Độ rộng của khoảng	**Cỡ mẫu**	**Độ bao phủ xác suất**
0,04	1000	0,6439
0,04	2000	0,6924
0,04	3000	0,6995
0,08	1000	0,9261
0,08	2000	0,9471
0,08	3000	0,9587

Tài liệu tham khảo

Baulch, B. & N. Minot (2002). The Spatial Distribution of Poverty in Vietnam and the Potential for Targeting. *World Bank working paper* 2829.

Brown, L. D., T. Tony Cai & A. DasGupta (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2), 101-133.

Congdon, P (2001). *Bayesian Statistical Modelling*. Wiley.

.....