Ước lượng theo phương pháp hệ thống đôi đánh giá Tổng điều tra dân số và nhà ở

Đồng Bá Hướng *(dịch)* **(*)**

Để ước tính về mức độ đầy đủ của các kết quả giữa hai cuộc điều tra, chúng ta cần phải sử dụng một *mô hình thống kê.* Mô hình này sẽ chỉ ra các "sai sót" trong một cuộc tổng điều tra do nguyên nhân nào như trong việc ước tính cho một biến ngẫu nhiên nào đó và từ đó có thể xây dựng được một hàm ước tính hợp lý cho mức khai báo thiếu trong tổng điều tra. Bài này sẽ giới thiệu về một phương pháp đơn giản cho ước tính theo hệ thống đôi, sau khi đã điều chỉnh các biến ước lượng để cho phép tính toán số trường hợp khai báo sai trong tổng điều tra và đo lường các vấn đề xảy ra trong phúc tra.

1. Lập mô hình cho khai báo thiếu (sót) trong tổng điều tra

Mô hình đánh giá mức độ đầy đủ đầu tiên đã được xây dựng để nghiên cứu sinh trắc học cho việc ước tính quy mô dân số "đóng". Khái niệm dân số "đóng" theo sinh trắc học là dân số mà trong đó cơ cấu dân số ít bị thay đổi trong khoảng thời gian nghiên cứu: không có sinh, không có chết và không có nhập cư, xuất cư. Các ứng dụng đầu tiên của kỹ thuật này, trong sinh trắc học, như ước tính "bắt - thả - bắt lại" dùng cho mục đích ước tính quy mô dân số của loài vật hoang dã. Ví dụ đưa ra dưới đây chỉ ra mô hình ban đầu được hình thành như thế nào.

Tiến hành xem xét một hồ nước kín (nghĩa là không có lối cho nước thoát ra và cũng không có lối cho nước chảy vào). Chúng ta muốn ước tính xem có bao nhiêu cá trong hồ này. Giả định là cá cùng chủng loại và được phân bổ ngẫu nhiên trong hồ. Chúng ta lấy một cái lưới, và đánh bắt cá. Trong lần đánh bắt đầu tiên, cố gắng bắt được càng nhiều cá càng tốt. Đánh dấu số cá bắt được (N1) và lại thả xuống hồ. Việc đánh dấu phải đảm bảo không làm bị thương hay gây hại đến cá. Sau một khoảng thời gian vừa đủ cho số cá được thả đó có thể trở lại phân bổ ngẫu nhiên trong hồ (lưu ý: không quá lâu đến mức mà số cá đó có thể sinh con hoặc bị chết). Tiến hành bắt cá lần 2 và số cá bắt được ký hiệu là N2. Trong lần đánh bắt lần 2 này, đếm số cá bị bắt lại (đã bị đánh dấu trong lần 1), được gọi là M. Như vậy chúng ta có ba con số, và các lần đánh bắt cá có thể được thể hiện qua Biểu 1 dưới đây:

Biểu 1. *Phân bổ của các lần đánh bắt trong nghiên cứu theo phương pháp "bắt - thả - bắt lại"*

Lần đánh bắt 1	Lần đánh bắt 2		
	Tổng	Số cá bắt được	Số cá không bắt được
Tổng	N_t	N_2	
Bắt được	N_1	M	
Không bắt được			

(*) Vụ trưởng Vụ Thống kê Dân số và Lao động

Trong đó:

N_1: Số cá bắt được trong lần 1

N_2: Số cá bắt được trong lần 2

M: Số cá bắt được trong cả hai lần

N_t: Tổng số cá trong hồ

Mục đích của bài toán này là để ước tính ra số cá có trong hồ (N_t), nhưng số cá chưa bị bắt trong hồ của cả 2 lần cũng là số chưa biết. Để lập mô hình toán cho quá trình bắt cá, ta coi việc bắt được một con cá là một biến cố tĩnh hoặc cụ thể hơn là một biến cố Bernoulii. Điều đó có nghĩa là, các con số về số cá bắt được M, N_1, N_2 được hiểu là các biến ngẫu nhiên. Nếu xác suất để mỗi một con cá có thể bắt được trong lần 1 là "a", và "b" là cho lần 2, và hai lần bắt cá này được thực hiện hoàn toàn độc lập với nhau thì các giá trị kỳ vọng của các biến số ngẫu nhiên M, N_1, và N_2 có thể được tính như sau:

(1.1) $E(M) = abNt$ (dựa theo tính độc lập của các biến cố)

(1.2) $E(N_1) = aNt$

(1.3) $E(N_2) = bNt$

Tham số ước tính có thể thu được bằng việc thế các giá trị quan sát được vào các giá trị kỳ vọng của M, N_1, và N_2 và thay thế phương trình (1.2) và (1.3) vào phương trình (1.1) ta được:

$$(1.4)\ M = abN_t = aN_t * bN_t/N_t = N_1 * N_2/N_t$$

Sắp xếp lại phương trình (1.4) ta được:

$$(1.5)\ N_t = N_1 * N_2/M$$

2. Mô hình áp dụng cho các cuộc tổng điều tra dưới các "điều kiện lý tưởng"

Trong ứng dụng để đánh giá cho một cuộc tổng điều tra, về lý thuyết là hoàn toàn tương tự như nghiên cứu "bắt – thả - bắt lại" đã được trình bày ở phần 1, ngoại trừ việc ước tính cho "dân số của con người" cần phải có một số giả định khác biệt chút ít.

Giả sử rằng lần bắt đầu tiên là tổng điều tra và lần bắt thứ hai là phúc tra. Gọi N_c (thay thế N_1) là số tổng điều tra, và N_p (thay thế cho N_2) là tổng mẫu đã được gia quyền của phúc tra. Np là tổng số người (hoặc đơn vị) sẽ được "**tìm thấy**" nếu tổng thể dàn mẫu cũng được sử dụng cho lấy mẫu phúc tra. Tất nhiên, dàn lấy mẫu của tổng điều tra là phải độc lập với dàn mẫu của phúc tra. Để ước tính, tốt nhất là phải có một dàn mẫu đầy đủ nhằm giảm thiểu độ biến thiên của ước lượng về số người (đơn vị) "không được tìm thấy" trong cả hai nguồn số liệu trên. Biến ngẫu nhiên M được coi là số người (đơn vị) "được điều tra" trong phúc tra mà cũng đã được điều tra trong tổng điều tra. Phép ước lượng được thực hiện giống như trong phương trình (1.5).

Lưu ý rằng có một số giả định phải được thoả mãn trong khi áp dụng mô hình ước tính dân số "con người". Mô hình giả định rằng việc "số trùng khớp" về số người (đơn vị) giữa hai nguồn số liệu là được thực hiện hoàn toàn đúng đắn, nghĩa là không có một quan sát nào bị "lặp lại" (không có người nào bị điều tra trùng) trong cả hai nguồn số liệu, và hai nguồn số liệu này là hoàn toàn độc lập. Giả định thứ 2 phải được tính đến trong quá trình thiết kế thông qua việc giữ cho "phúc tra" phải độc lập với "tổng điều tra". Vì vậy, chọn mẫu phúc tra phải độc lập với chọn mẫu của tổng điều tra. Có lẽ lựa chọn tốt nhất

cho một dàn mẫu sẽ là việc cập nhật lại mới nhất danh sách các đơn vị điều tra trong tổng điều tra trước *hoặc* là một mẫu địa bàn được thiết kế qua việc lập danh sách mới nhất về các địa bàn điều tra.

Các yêu cầu khác để đảm bảo tính độc lập cho hai nguồn số liệu là việc tuyển chọn điều tra viên cho phúc tra phải là những người mới (không phải là điều tra viên trong tổng điều tra) và việc xử lý kết quả phúc tra phải hoàn toàn độc lập với quá trình xử lý kết quả tổng điều tra nhằm loại bỏ "lỗi gây nhiễu".

Mẫu phúc tra được sử dụng để ước tính số bị "bỏ sót" của tổng điều tra sẽ được xem như là "mẫu P" trong thảo luận dưới đây nhằm phân biệt nó với "mẫu ước tính" được rút ra từ tổng điều tra, "mẫu E".

3. Mô hình áp dụng cho các cuộc tổng điều tra dưới các điều kiện "ít hoàn hảo hơn"

Một số loại lỗi mà có thể ảnh hưởng tới thủ tục ước tính đã giới thiệu ở trên:

- Kết quả tổng điều tra có thể có các trường hợp khai báo trùng.
- Các đơn vị nhà ở được lập danh sách trong tổng điều tra có thể được kê khai đúng nhưng lại không được phân định đúng theo khu vực địa lý.
- Các thành viên trong một đơn vị nhà ở có thể được điều tra ở nơi khác hoặc có thể không là đối tượng điều tra của cuộc tổng điều tra (ví dụ những người thuộc diện không phải điều tra).
- Các thành viên trong một đơn vị nhà ở có thể không được kê khai đầy đủ (thông tin định danh của cá nhân đó không đầy đủ).

Bất kỳ một lỗi nào kể trên đều có thể dẫn đến sự sai lệch trong ước tính về quy mô dân số. Để đo lường hoặc tính được ảnh hưởng của từng nhân tố này, một mẫu thứ 2 được thiết kế để tính các hệ số điều chỉnh trong quá trình ước tính. Mẫu thứ 2 hay còn gọi là mẫu E, được chọn trực tiếp từ tổng điều tra cho cùng "địa bàn" đó và sử dụng phương pháp phân tầng giống như trong mẫu được trình bày ở phần trước.

Tiến hành điều tra lại các hộ trong mẫu E và thu thập một số chỉ tiêu. Ngay khi đến đơn vị nhà ở được chọn điều tra, kiểm tra thông tin định danh (vị trí địa lý) và so sánh với thông tin đã được ghi cho đơn vị nhà ở đó trong tổng điều tra. Nếu khác biệt, phải kiểm tra lại trường hợp này (thường được kiểm tra bởi một nhóm thứ 3) và quyết định xem liệu việc ghi mã định danh ban đầu trong tổng điều tra là đúng hay sai. Lý do kiểm tra lại việc ghi mã định danh là vì các đơn vị nhà ở bị ghi sai mã định danh trong tổng điều tra sẽ không chắc được "**tìm thấy**" khi kết quả so sánh với phúc tra là trùng khớp nếu như số thông tin nghiên cứu bị hạn chế. Vì thế, sẽ xảy ra trường hợp đơn vị nhà ở đó có được kê khai đúng trong tổng điều tra nhưng sẽ không một thành viên nào được "tìm thấy". Điều này dẫn đến tổng dân số Nc mặc dù đúng nhưng số "trùng khớp" M sẽ rất nhỏ. Phương pháp để điều chỉnh tổng "trùng khớp" M được giới thiệu ở phần 4.

Khi tiến hành điều tra theo mẫu này, điều tra viên sẽ biết về từng đối tượng hiện cư trú trong đơn vị nhà ở được chọn mẫu là liệu người đó đã được kê khai trong tổng điều tra thực tế có "tồn tại hay không" (đôi khi có một số trường hợp kê khai "không đúng" trong tổng điều tra do lỗi của điều tra viên) và cũng sẽ biết được rằng liệu người đó có thuộc diện được kê khai trong đơn vị nhà ở đó trong tổng điều tra hay không. Và ta sẽ có được con số về số người không phải kê khai trong đơn vị nhà ở đó theo từng hộ. Trong trường hợp khai báo "**thừa**" trong

tổng điều tra, những người không thuộc diện kê khai sẽ bị trừ đi trong tổng dân vì thực tế họ không tồn tại. Trong trường hợp kê khai sai, nếu cuộc phỏng vấn trong phúc tra được tiến hành đúng theo các quy định trong tổng điều tra, thì cuộc phỏng vấn phúc tra đó hoặc sẽ không kê khai cho cá nhân đó (không thuộc diện điều tra) hoặc sẽ phải kê khai họ ở một địa chỉ khác, và trong cả hai trường hợp sẽ không có sự "trùng khớp". Trong trường hợp này, tổng dân số Nc sẽ không đúng và sẽ không có "trường hợp nào trùng khớp".

Hai hoạt động tiếp sau sẽ được thực hiện ở đơn vị xử lý kết quả phúc tra. Để nghiên cứu về tỷ lệ trùng khớp của phúc tra, các trường hợp mẫu được lấy từ tổng điều tra phải được so sánh với tất cả các trường hợp khác trong tổng điều tra trong cùng một địa bàn điều tra. So sánh này phải được tiến hành cho từng người đã được liệt kê trong đơn vị mẫu đã được thẩm định trong phúc tra nhằm xác định xem liệu có bất kỳ trường hợp kê khai trùng nào không trong cùng địa bàn đó. Kê khai trùng trong tổng điều tra sẽ làm cho tổng dân Nc bị cường điệu lên mặc dù số trường hợp trùng khớp khi so với kết quả trong mẫu phúc tra, M, là đúng.

Điểm cuối cùng, có những trường hợp trong tổng điều tra không có đủ nguồn thông tin cần thiết để so sánh với mẫu phúc tra. Mặc dù có được kê khai trong tổng điều tra (có thể do kết quả làm sạch số liệu cho nhiều thông tin bị sót) dẫn đến tổng Nc là đúng, nhưng độ trùng khớp không hoàn toàn và tổng M sẽ bị ước tính thiếu.

Có bốn tổng số mà có thể thu được từ mẫu thứ 2 này và từ các file máy tính của các câu trả lời. Đó là:

G: số người được phân định sai vị trí địa lý trong tổng điều tra

E: Số người bị kê khai sai trong tổng điều tra (khai thừa hoặc không thuộc đối tượng điều tra)

D: Số trường hợp kê khai trùng trong tổng điều tra

I: Số người tuy được kê khai trong tổng điều tra nhưng không có đủ nguồn thông tin phù hợp để chạy so sánh (bằng máy tính)

4. Ước tính quy mô dân số và tỷ lệ kê khai sót thuần

Phương trình (1.5) đưa ra ước lượng quy mô dân số khi nguồn số liệu được sử dụng trong nghiên cứu "bắt - bắt lại" là không có lỗi "sai số". Khi tính thêm yếu tố "lỗi", tổng số dân phải được điều chỉnh để loại bỏ các trường hợp khai báo trùng, khai báo thừa, và các trường hợp mà không thể chạy so sánh. Thông số ước tính cuối cùng cho quy mô dân số sẽ là:

$$(1.6)\quad N_t = N_p * (N_c - G - E - D - I)/M$$

Tỷ lệ kê khai sót thuần, Rn, có thể được tính bằng tỷ lệ giữa tổng số dân được kê khai trong tổng điều tra so với tổng dân số ước tính:

$$(1.7)\quad R_N = N_c/N_t = [M/N_p] / [(N_c-G-E-D-I)/N_c]$$

Nguồn:

Tài liệu đào tạo thống kê, Văn phòng Tổng điều tra dân số Hoa Kỳ Bộ Thương mại Hoa Kỳ, tháng 8/1985.

"Evaluating Censuses of Population and Housing"