KINH NGHIỆM THIẾT KẾ MẪU CHỦ CỦA PHILIPPINES

(Tiếp theo)

(4). Xác định số PSU; SSU; USU

Dựa trên thiết kế của điều tra MS năm 2003, số lượng USU được xác định bởi công thức:

$$b_{opt} = \sqrt{\frac{C_\alpha (1 - roh)}{c \times roh}} \qquad (4)$$

Trong đó:

b_{opt}: Tổng số hộ gia đình tối ưu nhất của một PSU;

C_α: Khinh phí tăng thêm 1 PSU;

c: Kinh phí khi tăng thêm 1 hộ gia đình;

roh: Hệ số tương quan vùng, được xác định $roh = \frac{deff - 1}{\bar{b} - 1}$, $\bar{b}$ là số hộ trung bình của địa bàn, được lấy từ điều tra năm 1997.

Trên cơ sở tính toán, quyết định chọn 12 hộ trên 1 địa bàn đối với những vùng NCR, và 16 hộ đối với ngoài NCR. Nguyên nhân chính cho sự khác biệt đó là chi phí thu thập dữ liệu. Sau khi tính được số hộ cho từng PSU, thì tính được số PSU bằng cách lấy tổng số hộ chia cho số hộ của 1 PSU.

(5). Điều chỉnh cỡ mẫu

Như trước đây đề cập, cỡ mẫu và phân bổ mẫu được tính chưa xét đến những điều không đáp ứng được, bản báo cáo tình hình, dân số dự kiến tăng lên giữa năm 2000 và 2003. Việc điều chỉnh cỡ mẫu cuối cùng đã được đưa ra bằng cách điều chỉnh số hộ gia đình lựa chọn trong khi số PSU được cố định (tức là PSU được xác định từ phân bổ ban đầu). Số hộ điều chỉnh (c) cho mỗi khu vực được xác định như sau:

$$c = \frac{\frac{n}{\bar{m}}}{a(N^* / N)} \qquad (5)$$

Trong đó:

c: Cỡ mẫu điều chỉnh;

$\bar{m}$: Tỷ lệ trả lời trong cuộc điều tra trước;

a: Số đơn vị mẫu (Số PSU);

N*: Số hộ gia đình được điều chỉnh

$$(N^* = N(\bar{\phi})(1 - r)^3)$$

$\bar{\phi}$: Tỷ lệ bản báo cáo tình hình của điều tra được ước tính;

$(1-r)^3$: Tổng số hộ gia đình tăng lên trong 3 năm;

r: Tỷ lệ số hộ gia đình tăng lên giữa hai cuộc điều tra dân số.

Từ đó đưa ra bảng cỡ mẫu ban đầu và sau khi điều chỉnh. Cỡ mẫu ban đầu là 43.882 hộ và cỡ mẫu điều chỉnh là 45.842 hộ. Cỡ mẫu được tính cấp khu vực (vùng, miền).

(6). Cách chọn mẫu

Để có thể kiểm soát về cỡ mẫu con (cỡ mẫu được thống nhất trong những đơn vi chọn mẫu chính) và đảm bảo rằng mỗi đơn vị có cơ hội lựa chọn như nhau. Ta tiến hành chọn mẫu PSU, SSU, USU như sau:

Chọn PSU theo phương pháp lựa chọn với xác suất tỷ lệ thuận với quy mô ước tính (PPES). Chọn SSU (ở Philippines là địa bàn) cũng được lựa chọn bằng phương pháp PPES và USU (hộ gia đình) trong địa bàn lựa chọn bằng phương pháp chọn mẫu hệ thống.

- Tuy nhiên, trước khi lựa chọn PSU, cần kiểm tra mẫu PSU "quá cỡ". Một PSU được gọi là quá cỡ nếu xác suất lựa chọn lớn hơn 1. Như vậy, PSU này được gọi là "đơn vị chọn mẫu chính - tự đại diện (Self-Representing PSUs) viết tắt là SR-PSUs. Mỗi SR-PSUs được coi như một phần của mẫu chủ (MS) và được coi như một tầng riêng biệt. SR-PSUs được xác định bằng cách kiểm tra quy mô của PSU. Có nghĩa là, nếu quy mô của PSU vượt quá điểm cắt thì PSU đó được loại ra khỏi khung và coi như là một tầng riêng biệt. Trong một vùng, PSU được phân loại là SR nếu quy mô của PSU đó đáp ứng:

$$M_{h\alpha} > \frac{\sum_{\alpha} M_{h\alpha}}{a_h} \times 0.7 \qquad (6)$$

$M_{h\alpha}$: Là số hộ của PSU thứ α của vùng h;

$a_{h:}$ Số PSU của vùng h;

Sau khi các SR được loại ra khỏi khung thì các quy trình lặp lại để xác định các SR còn lại. Các PSU còn lại sau khi lọc hết SR-PSU được gọi là các PSU không tự đại diện, viết tắt là NSR-PSU.

Cách lựa chọn các SR-PSU

- Tất cả các SR-PSU có trong mẫu được xem như một tầng riêng biệt

- Về phương diện này, cho phép ước lượng giá trị phương sai của các ước tính đã tạo ra tối thiểu cho 2 EA lựa chọn bằng phương pháp PPES cho mỗi SR-PSU.

- Mỗi EA, các hộ gia đình được chọn bằng phương pháp hệ thống

- Số EA trong mỗi PSU được xác định bởi:

$$b = \frac{f \times M_{\alpha}}{c} \qquad (7)$$

Trong đó:

b: Số EA được lựa chọn trong mỗi SR- PSU;

f=n/N: Tỷ lệ chọn mẫu mong muốn trong vùng

n: Cỡ mẫu của vùng;

N: Tổng số hộ gia đình trong vùng;

M_{α}: Tổng số hộ gia đình của SR-PSU thứ α ;

c: Số mong muốn, HHs/PSU.

Chú ý: Trong một số trường hợp, b<2, lúc này làm tròn thành 2, điều này giúp cho cỡ mẫu của của EA (HHs/EA) nhỏ hơn số mong muốn. Khi đó có biến động trong cỡ mẫu của PSU. Mong muốn này làm cho thiết kế epsem trong mỗi vùng thay đổi tổng cỡ mẫu của SR-PSU.

Cách lựa chọn các NSR-PSU

Tronh mỗi vùng, các NSR-PSU được phân tầng theo tỉnh/HUCs/ICC[1]. Mục đích chính của việc phân tầng này là mẫu quan sát được đảm bảo ở mỗi tỉnh, giúp ước tính ở cấp tỉnh (sử dụng ước tính trực

[1] *Highly Urbanized Cities (HUC) là các thành phố đô thị hóa cao và Independent Chartered Cities (ICC) là các thành phố điều lệ*

tiếp hoặc gián tiếp). Số lượng mẫu NSR-PSU cho từng tỉnh/ HUCs/ICC được xác định như sau:

- Tổng số mẫu của khu vực đã được phân bổ theo tỷ lệ (tổng số hộ gia đình) ở mỗi tỉnh/HUCs/ICC. Phương pháp này của phân bổ đảm bảo giúp duy trì thống nhất quyền số trong từng vùng (Chọn mẫu epsem). Như vậy số PSUs được xác định bằng cách chia cỡ mẫu cho tổng số hộ gia đình của một PSU

- Số NSR-PSU của mỗi tầng là Tỉnh/ HUCs/ICC được tính bằng lấy tổng số PSU trừ đi số SR PSU, có sự khác biệt là nó được điều chỉnh, làm tròn tăng lên hoặc giảm xuống sao cho là bội của 4. Mục đích của điều chỉnh này để tạo điều kiện lấy mẫu xoay vòng. Như vậy điều chỉnh sẽ ảnh hưởng đến cỡ mẫu cuối cùng được rút ra trong NSR PSU lấy vào mẫu.

- NSR-PSU trong tỉnh / HUC / ICC được tiếp tục phân tầng dựa trên lựa chọn các biến phân tầng để thực sự cải thiện trên hiệu quả của thiết kế mẫu.

Trong mỗi NSR vừa chọn, đơn vị chọn mẫu cấp 2 (các EA) được chọn bằng phương pháp PPES. Ở mỗi EA, hộ gia đình chọn bằng phương pháp chọn mẫu hệ thống.

(7). Tiêu chí để phân tầng chọn mẫu

Trong thiết kế nhiều tầng, tổng phương sai được tạo thành từ nhiều thành phần. Mỗi phương sai thành phần là số đo của một đóng góp của một giai đoạn lấy mẫu cụ thể để có phương sai tổng thể. Như đã được biết đến, phương sai của thành phần của giai đoạn đầu tiên của chọn mẫu đóng góp cho tổng thể thường lớn nhất trong các thành phần, đặc biết nếu tương quan nội nhóm là tích cực.

- Trong điều kiện này, phân nhóm các PSU tương tự nhau vào cùng tầng cải thiện độ chính xác của các ước lượng mà không ảnh hưởng đến sự tăng lên của mẫu tổng thể.

- Tiêu chí lựa chọn các biến để phân tầng

+ Phải liên quan, tương quan với các biến quan tâm - tỷ lệ nghèo;

+ Biến có sẵn với các đơn vị mẫu chính (PSUs);

+ Đo được hoặc có sai số không đáng kể;

- Phân tầng dựa vào danh sách sẵn có:

+ Danh sách các cơ sở năm 2000;

+ Tổng điều tra dân số Dân số năm 2000 - Mẫu 2 (Bảng hỏi thường cho hộ), Mẫu 5 (Đặc điểm làng);

+ Các khoản thu nhập (COA).

- Phân tầng biến lựa chon: Các biến sau đây thường được lấy làm tiêu chí phân tầng:

+ Tỷ lệ nhà giàu trong PSU;

+ Thu nhập thành phố thể hiện đơn vị bình quân đầu người, dựa báo cáo thu nhập năm 2001;

+ Tỷ lệ hộ gia đình trong PSU tham gia vào nông nghiệp;

- Thứ tự của phân tầng được xác định dựa trên mức độ liên quan của các biến với tỷ lệ nghèo đói. Sử dụng dữ liệu từ năm 2000 FIES cho mục đích này. Sử dụng thống kê Chi-bình phương Person, mục hiệu chỉnh thứ 2 Rao and Scott (1984) để xác định mực độ liên quan.

- Tổng số của các tầng hiển nhiên (SR-PSU) (1 hoặc 2 biến) được xác định cho mỗi tỉnh (cỡ mẫu /không là tầng xác định) là bội của 4.

- Việc xác định tầng hiển nhiên được thực hiện với 1 biến tại một thời điểm và ranh giới của nó được

xác định bởi sự cân bằng tổng số hộ mỗi tầng. (Điều này được sử dụng để đảm bảo một thiết kế epsem với số lượng bằng nhau của các PSU).

- Mẫu PPEs (mẫu hệ thống) của NSR-PSU đã được lựa chọn trong mỗi tầng là tường minh.

- Sau khi tầng tường minh được xác lập (với 1 hoặc 2 biến), PSUs được ngầm phân tầng cùng với sự liên quan với các biến phân tầng còn lại.

(8). Xác suất lựa chọn mẫu

Với thiết kế như vậy, phương trình tính xác suất lựa chọn cho mỗi vùng cho bởi công thức như sau:

$$P(h\alpha\beta\gamma) = \frac{a_h M_{h\alpha}}{\sum_h M_{h\alpha}} \times \frac{b_h M_{h\alpha\beta}}{M_{h\alpha}} \times \frac{c_h}{M_{h\alpha\beta}} = \frac{n_h}{N_h} \quad (8)$$

Trong đó:

a_h: Số PSUs được lựa chọn ở tầng thứ h;

b_h: Số mẫu của mỗi địa bàn (EA);

c_h: Số mẫu hộ;

$M_{h\alpha}$: Tổng số hộ của đơn vị chọn mẫu cấp 1 thứ α;

$M_{h\alpha\beta}$: Tổng số hộ của địa bàn β của đơn vị chọn mẫu cấp 1 thứ α;

n_h: Tổng số mẫu của tầng h;

N_h: Tổng số hộ của tầng h ($= \sum_h M_{h\alpha}$).

Phương trình lựa chọn này cho thấy rằng các thiết kế mẫu cho một khu vực là epsem. Phương trình cũng được sử dụng để xác định số lượng đơn vị được lựa chọn tại mỗi giai đoạn lấy mẫu. Tuy nhiên, do làm tròn số thập phân, phương trình lựa chọn như vậy có thể không được thỏa mãn mà sẽ cho kết quả của một thiết kế không epsem. Để buộc một thiết kế epsem, một số điều chỉnh được thực hiện trên ước tính các biện pháp cho mẫu. Đầu tiên làm tròn lên với khoảng lấy mẫu, ($M_{h\alpha\beta}/c_h$) để có một số tự nhiên, gọi là $M'_{h\alpha\beta}$. Như vậy,

$$M'_{h\alpha} = \sum_\beta M'_{h\alpha\beta} \quad (9)$$

Ta có phương trình tương đương:

$$P(h\alpha\beta\gamma) = \frac{a_h M'_{h\alpha}}{\sum_h M'_{h\alpha}} \times \frac{b_h M'_{h\alpha\beta}}{M'_{h\alpha}} \times \frac{1}{M'_{h\alpha\beta}} = \frac{n_h}{N_h} \quad (10)$$

Dựa vào phương trình lựa chọn này cách thức lấy mẫu được triển khai như sau:

+ Lựa chọn PSU a_h với PP $M'_{h\alpha}$, chú ý rằng mỗi SR - PSU là 1 tầng

+ Trong mỗi PSU được chọn, chọn các b_h EA với xác xuất $M'_{h\alpha\beta}$, chú ý rằng với NSR - PSU thì b_h=1 và b_h>=2 với SR- PSU.

+ Trong mỗi EA, chọn số hộ với một khoảng $M'_{h\alpha\beta}$

(9). Quy trình ước lượng

Hầu hết các ước lượng được kỳ vọng tạo ra từ các cuộc điều tra là đại lượng: tổng thể/lĩnh vực; trung bình/tỷ trọng; tỷ lệ.

Cho địa bàn/vùng thì tổng thể dân số được ước lượng bởi công thức sau:

$$\hat{Y} = \hat{Y}_{SR} + \hat{Y}_{NSR} \quad (11)$$

Trong đó:

$\hat{Y}$: Ước lượng tổng thể;

$\hat{Y}_{SR}$: Ước lượng tổng thể từ đơn vị chọn mẫu cấp 1 tự đại diện;

$\hat{Y}_{NSR}$: Ước lượng tổng thể từ đơn vị chọn mẫu cấp 1 không tự đại diện.

Và ta có $\hat{Y}_{SR} = \sum_{\alpha}\sum_{\beta}\sum_{\gamma} w_{\alpha\beta\gamma} y_{\alpha\beta\gamma}$

$y_{\alpha\beta\gamma}$: Quan sát của hộ gia đình thứ γ của địa bàn thứ β của đơn vị chọn mẫu cấp 1 tự đại diện thứ α

$w_{\alpha\beta\gamma}$: Quyền số điều tra của hộ gia đình thứ γ của địa bàn thứ β của đơn vị chọn mẫu cấp 1 tự đại diện thứ α

Chú ý rằng trọng số điều tra được định nghĩa là tạo bởi từ 3 quyền số: (1) Nghịch đảo của xác suất lựa chọn; (2) Quyền số không trả lời; Quyền số do phân tầng. Đối với mỗi SR-PSU, thì (1) được xác định:

$$w_{1\alpha\beta\gamma} = \frac{M_{\alpha}}{b_{\alpha}M_{\alpha\beta}} \times \frac{M_{\alpha\beta}}{c_{\alpha}} = \frac{M_{\alpha}}{b_{\alpha}c_{\alpha}} = \frac{M'_{\alpha}}{b_{\alpha}M'_{\alpha\beta}} \times M'_{\alpha\beta} = \frac{n}{N}$$

$$\hat{Y}_{NSR} = \sum_{h}\sum_{\alpha}\sum_{\beta}\sum_{\gamma} w_{h\alpha\beta\gamma} y_{h\alpha\beta\gamma}$$

$y_{h\alpha\beta\gamma}$: Quan sát của hộ gia đình thứ γ của địa bàn thứ β của đơn vị chọn mẫu cấp 1 tự đại diện thứ α trong tầng thứ h.

$w_{\alpha\beta\gamma}$: Quyền số điều tra của hộ gia đình thứ γ của địa bàn thứ β của đơn vị chọn mẫu cấp 1 tự đại diện thứ α trong tầng thứ h.

Tương tự như vậy, trọng số của điều tra dựa trên nghịch đảo của xác suất lựa chọn.

$$w_{1h\alpha\beta\gamma} = \frac{M_{h\alpha}}{a_h M_{h\alpha}} \times \frac{M_{h\alpha}}{M_{h\alpha\beta}} \times \frac{M_{h\alpha\beta}}{c_h} = \frac{\sum_h M_{h\alpha}}{a_h c_h} = \frac{\sum_h M'_{h\alpha}}{a_h M'_{h\alpha}} \times \frac{M'_{h\alpha}}{M'_{h\alpha\beta}} \times M'_{h\alpha\beta} = \frac{n}{N}$$

Trong trường hợp cho tỷ lệ (hoặc trung bình/tỷ trọng), được ước lượng như sau:

$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\hat{Y}_{SR} + \hat{Y}_{NSR}}{\hat{X}_{SR} + \hat{X}_{NSR}}$ nghĩa là ước lượng tỷ lệ được kết hợp

Phương sai được ước lượng như sau:

$$v(\hat{Y}) = v(\hat{Y}_{SR} + \hat{Y}_{NSR}) = v(\hat{Y}_{SR}) + v(\hat{Y}_{NSR})$$

$$v(\hat{R}) = \hat{R}^2 \left\{ \frac{v(\hat{Y})}{\hat{Y}^2} + \frac{v(\hat{X})}{\hat{X}^2} - 2\frac{cov(\hat{Y}, \hat{X})}{\hat{Y}\hat{X}} \right\}$$

Phương sai được ước lượng nhờ phần mềm ứng dụng trong thống kê.

Tài liệu tham khảo:

1. Arturo Y.Pacificador, JR, "The 2003 Master Sample Design", Bài trình bày tại khóa tập huấn chọn mẫu điều tra cho cán bộ TCTK Việt Nam năm 2013;

2. Hans Pettersson, "Báo cáo Mẫu tổng thể cho điều tra hộ gia đình tại Việt Nam", Báo cáo từ chuyến công tác tới Tổng cục Thống kê Việt Nam năm 2009;

3. David J. Megill, International Sampling Consultant; Nguyen Manh The, National Consultant; Nguyen Thi Thanh Mai, National Consultant GSO, UNDP, "Assessment of Sampling Frames for GSO National Household Surveys in Viet Nam, and Initial Planning for New Master Sample", Thực hiện năm 2013.

Vũ Vân Anh (lược dịch và tổng hợp)