MỘT SỐ KINH NGHIỆM THIẾT KẾ VÀ QUẢN LÝ MẪU ĐIỀU TRA
(tiếp theo)

*Đồng Bá Hướng**

3. Thiết kế mẫu tự gia quyền

Không làm giảm tính tổng quát,[1] phần này trình bày phương pháp thiết kế mẫu tự gia quyền hai giai đoạn là loại mẫu điều tra phổ biến và hiệu quả nhất hiện nay.

Thiết kế *mẫu tự gia quyền 2 giai đoạn* với chọn đơn vị điều tra (PSU) ở giai đoạn 1 và chọn USU (trong dân số là "chọn hộ") ở giai đoạn 2, mỗi hộ sau khi đã chọn sẽ điều tra tất cả các nhân khẩu trong hộ.

Gọi a_h là số PSU cần phải chọn của tầng h, M_{hi} là quy mô đã ước lượng của đơn vị điều tra (PSU) thứ i thuộc tầng h, $\sum M_{hi}$ là tổng quy mô của tầng h (đo bằng dân số hoặc số hộ) - đã được xác định trước theo *phương pháp phân bổ tỷ lệ nghịch*.

Giai đoạn 1 (chọn địa bàn): Các PSU được chọn theo *phương pháp chọn hệ thống PPS* với xác suất chọn bằng nhau (và bằng $f_h = n_h/N_h$) thông qua sử dụng khoảng cách chọn $I_{1h} = \sum M_{hi}/a_h$. Mỗi PSU đã chọn sẽ điều tra tất cả nhân khẩu trong địa bàn (mẫu chùm cả khối). Với các điểm mẫu phân bố trong khoảng I_{1h} này, một địa bàn (PSU) có quy mô M_{hi} sẽ có xác suất chọn là M_{hi}/I_{1h}. Vì vậy, xác suất chọn mẫu giai đoạn 1 của PSU thứ i thuộc tầng h là:

$$P_{1hi} = M_{hi}/I_{1h} = a_h \times M_{hi}/\sum M_{hi} \qquad (3.1)$$

Trong đó M_{hi} là số hộ khi lập bảng kê của PSU thứ i thuộc tầng h.

Giai đoạn 2 (chọn USU hay chọn hộ): Gọi P_{2hi} là xác suất chọn hộ trong PSU thứ i thuộc tầng h, do mỗi PSU sau khi đã chọn ở giai đoạn 1 sẽ chọn đúng 20 hộ, khi đó:

$$P_{2hi} = 20/M^*_{hi} \qquad (3.2)$$

Trong đó M^*_{hi} là số hộ đã cập nhật đến thời điểm điều tra của PSU thứ i thuộc tầng h.

Việc chọn các hộ theo phương pháp chọn hệ thống với các xác suất chọn bằng nhau, và khoảng cách chọn hộ I_{2hi} trong PSU thứ i của tầng h là:

$$I_{2hi} = 1/P_{2hi} = M^*_{hi}/20 \qquad (3.3)$$

Đối với mỗi PSU, một bảng kê hộ đã được lập trước khi tiến hành điều tra, và khoảng cách chọn hộ (I_{2hi}) được sử dụng để chọn cụ thể các hộ điều tra.

Để bảo đảm mẫu tự gia quyền trong nội bộ mỗi tầng, xác suất chung của mỗi tầng qua 2 giai đoạn chọn mẫu phải hoàn toàn giống nhau đối với mỗi hộ thuộc mẫu:

$$f_{hij} = P_{1hi} \times P_{2hi} = M_{hi} \times 20a_h/M^*_{hi} \times \sum M_{hi} \qquad (3.4)$$

---

* Ủy viên BCH Hội Thống kê Việt Nam

[1] Mẫu 1 giai đoạn chỉ bao gồm giai đoạn chọn PSU, còn mẫu nhiều hơn 2 giai đoạn thì bổ sung thêm các giai đoạn trước khi chọn PSU (như chọn tỉnh, huyện, xã,…), trong mỗi giai đoạn đó đều sử dụng phương pháp chọn mẫu hệ thống PPS tương tự như quy trình chọn PSU.

Quyền số thiết kế cơ bản của hộ j đã chọn thuộc PSU thứ i của tầng h là:

$$W_{hij} = 1/f_{hij} = M^*_{hi} \times \sum M_{hi} / M_{hi} \times 20a_h \qquad (3.5)$$

Để bù trừ việc phân bổ mẫu không theo tỷ lệ thuận giữa các tầng, *quyền số thiết kế cơ bản* của mỗi phần tử (ở đây là mỗi hộ) đã chọn thuộc PSU thứ i của tầng h được tính như sau:

$$W_{hij} = f/f_{hij} \qquad (3.6)$$

Trong đó: f_{hij} là tỷ lệ chọn mẫu chung của tầng h, f là tỷ lệ chọn của tổng thể mẫu: f=n/N (n = số hộ đã chọn cho điều tra, N = số hộ đã ước lượng đến năm điều tra của cả nước).

Chú ý:

a) Sử dụng các quyền số W_{hij} theo công thức (3.6), tổng quy mô mẫu sau khi gia quyền sẽ đúng bằng tổng quy mô mẫu trước khi gia quyền (n). Đây là phương pháp gia quyền đã và đang được áp dụng phổ biến trên thế giới hiện nay.

Tuy nhiên, nếu muốn *suy rộng theo quy mô tổng thể chung*, phải nhân quyền số W_{hij} (ở công thức 3.6) với tỷ số N/n hay 1/f, khi đó:

$$W_{hij} = f/f_{hij} \times 1/f = 1/f_{hij}. \qquad (3.7)$$

Tính quyền số theo công thức (3.7) là phương pháp *suy rộng mẫu* phổ biến ở nước ta hiện nay. Khi ước lượng mẫu, chúng tôi thường sử dụng công thức này.

b) Các quyền số W_{hij} phải được tạo thành một biến riêng và tích hợp vào từng bản ghi cá nhân để phục vụ cho việc tổng hợp, phân tích số liệu và tính sai số mẫu cho cuộc điều tra.

4. Ước lượng mẫu và sai số mẫu

4.1 Ước lượng mẫu

Thiết kế điều tra cung cấp các ước lượng riêng cho hai tầng thứ cấp là thành thị và nông thôn. Hầu hết các ước lượng của cuộc điều tra là các ước lượng về tỷ lệ, song các nhà quản lý thường yêu cầu các ước lượng về tổng số cho các tầng.

▪ *Các quyền số thiết kế:*

Cho dù trong nội bộ mỗi tầng đã thiết kế được một mẫu tự gia quyền, song xét trong toàn bộ tổng thể mẫu thì mẫu đã thiết kế thường không phải là mẫu tự gia quyền (do đã phân bổ mẫu với tỷ lệ khác nhau giữa các tầng). *Các quyền số thiết kế* được sử dụng để bù trừ sự khác nhau về xác suất (tỷ lệ) chọn mẫu của các tầng. Các quyền số thiết kế phải được tính cho mỗi PSU đã chọn vào mẫu và quyền số của mỗi PSU được sử dụng như là hệ số suy rộng cho mỗi PSU.

Quyền số của PSU thứ i thuộc tầng h bằng tỷ lệ nghịch của xác suất chọn PSU, trong đó, theo kết quả đã nêu trên, quyền số thiết kế (dùng để suy rộng) của hộ thứ j trong PSU thứ i thuộc tầng h là:

$$W_{hij} = M^*_{hi} \times \sum M_{hi} / M_{hi} \times 20a_h \qquad (4.1)$$

Thừa số M^*_{hi}/M_{hi} trong phương trình (4.1) là tỷ lệ giữa số hộ đã cập nhật đến thời điểm điều tra chia cho số hộ đã ghi trong bảng kê của PSU thứ i, thừa số này chỉ ra mức độ ảnh hưởng của chất lượng lập bảng kê địa bàn trước khi tiến hành điều tra.

▪ *Phương pháp ước lượng mẫu:*

Ước lượng tổng số của một đặc trưng y của tầng h theo công thức sau:

$$Y_h = \sum_i \sum_j w_{hij} y_{hij} \qquad (4.2)$$

$$i = 1, 2, \ldots, a_h$$

$$j = 1, 2, \ldots, n_{hi}$$

Trong đó:

Y_h = đặc trưng y của tầng h

y_{hij} = đặc trưng bất kỳ của nhân khẩu k trong hộ j của PSU i thuộc tầng h

a_h = số PSU đã chọn thuộc tầng h

n_{hi} = số hộ mẫu trong PSU thứ i

w_{hij} = $1/f_h$

Ước lượng Y riêng cho khu vực thành thị và nông thôn của 63 tầng (tỉnh/thành phố) tính bằng tổng các ước lượng tương ứng cho mỗi khu vực của các tầng, nghĩa là:

$$Y = \sum Y_h \quad (h = 1, 2,\ldots, 63) \qquad (4.3)$$

Hầu hết các ước lượng của cuộc điều tra được tính dưới dạng các số bình quân hoặc các tỷ lệ. Ước lượng số bình quân của tầng h có dạng một tỷ lệ và xác địnhtheo công thức sau:

$$r_h = \frac{Y_h}{X_h} = \frac{\sum_i \sum_j w_{hij} y_{hij}}{\sum_i \sum_j w_{hij} x_{hij}} \qquad (4.4)$$

Trong đó:

y_{hij}, a_h, n_{hi}, w_{hij} đã được xác định ở trên

X_{hij} = 1 khi: i = 1, 2, …, a_h

j = 1, 2, …., n_{hi}

Số bình quân của tổng thể cũng có dạng tỷ lệ, ký hiệu là r, được ước lượng theo công thức sau:

$$r_h = \frac{\sum_h \sum_i \sum_j w_{hij} y_{hij}}{\sum_h \sum_i \sum_j w_{hij} x_{hij}} \qquad (4.5)$$

Trong đó X_{hij} được xác định như trong phương trình (4.4).

4.2 Sai số mẫu

Sai số mẫu được đo bằng *sai số chuẩn* của một chỉ tiêu thống kê cụ thể (số bình quân, tỷ lệ), sai số chuẩn bằng căn bậc hai của phương sai. Sai số chuẩn được dùng để tính các khoảng tin cậy mà trong đó chứa giá trị thực của tổng thể. Ví dụ, đối với mỗi chỉ tiêu thống kê tính được từ số liệu điều tra mẫu, giá trị của chỉ tiêu thống kê sẽ rơi vào một dãy số liệu trong phạm vi ±2×sai số chuẩn của chỉ tiêu thống kê với xác suất tin cậy 95% của tất cả các mẫu có thể chọn được theo quy mô và phương pháp thiết kế mẫu đã xác định.

Nếu một mẫu các đối tượng điều tra đã được chọn theo phương pháp chọn mẫu ngẫu nhiên đơn giản, ta có thể sử dụng công thức tính trực tiếp sai số mẫu, đó là $S/\sqrt{n}$ và $\sqrt{(pq/n)}$ để ước lượng sai số chuẩn của chỉ tiêu bình quân và tỷ lệ tương ứng. Tuy nhiên, do mẫu điều tra thường được thiết kế theo mẫu điều tra phân tầng-hệ thống nhiều giai đoạn, nên phải sử dụng công thức phức tạp hơn.

Phần mềm máy tính dùng để tính sai số mẫu là mô-đun sai số mẫu của ISSA. Mô-đun này sử dụng phương pháp tuyến tính hóa Taylor để ước lượng phương sai cho các ước lượng điều tra, đó là các ước lượng tỷ lệ (cho chỉ tiêu trung bình hay tỷ lệ). Phương pháp tuyến tính hóa Taylor coi mỗi số phần trăm hay số trung bình như là một tỷ số r = y/x, trong đó y biểu thị cho giá trị tổng số mẫu của biến Y, x biểu thị cho tổng số các trường hợp trong nhóm hay phân nhóm đang xem xét.

▪ *Phương sai của tỷ lệ r của tổng thể mẫu*, với sai số chuẩn bằng căn bậc hai của phương sai:

$$Var(r) = SE^2(r) = \frac{1-f}{x^2}\sum_{h=1}^{H}\left[\frac{m_h}{m_h-1}\left(\sum_{i=1}^{m_h} z_{hi}^2 - \frac{z_h^2}{m_h}\right)\right] \quad (4.6)$$

trong công thức trên: $z_{hi} = y_{hi} - r.x_{hi}$, và $z_h = y_h - r.x_h$

trong đó:

- h biểu thị tầng thay đổi từ 1 đến H
- m_h là số chùm (PSU) đã chọn trong tầng h
- y_{hi} là tổng giá trị của biến y trong chùm (PSU) thứ i của tầng h

$$y_{hi} = \sum_j w_{hij} y_{hij}$$

- x_{hi} là tổng các trường hợp trong chùm (PSU) thứ i của tầng h

$$x_{hi} = \sum_j w_{hij} x_{hij}$$

- y_h được tính theo công thức (4.2)

$$x_h = \sum_i \sum_j w_{hij} x_{hij}$$

- r được tính theo công thức (4.5)

$$X^2 = \left(\sum_h \sum_i \sum_j w_{hij} x_{hij}\right)^2$$

- W_{hij} = quyền số của mỗi phần tử (nhân khẩu) thuộc đơn vị mẫu cuối cùng đã chọn (công thức 3.7) (trong dân số, đó là hộ mẫu đã chọn)
- f = phân số mẫu chung, thường quá nhỏ có thể bỏ qua.

▪ *Phương sai của tỷ lệ r_h của tầng h* được ước lượng theo công thức sau đây:

$$Var(r_h) = SE^2(r_h) = \frac{1-f_h}{x_h^2}.\frac{m_h}{m_h-1}.\sum_i\left(z_{hi}^2 - \frac{z_h^2}{m_h}\right) \quad (4.7)$$

Ngoài sai số chuẩn (se) của ước lượng r, các tham số quan trọng khác là: hiệu quả thiết kế (deft), sai số chuẩn tương đối (se/r), khoảng tin cậy 95% (r±2se),...

Tính đúng sai số mẫu cho những thiết kế mẫu phức tạp đòi hỏi phải hiểu rõ các tầng và các đơn vị mẫu cơ bản mà mọi phần tử mẫu nằm trong đó. Vì vậy vấn đề quan trọng là những thông tin này phải được ghi vào bản ghi số liệu cho từng cá nhân, nếu không thì các chương trình như ISSA, CLUSTERS, STATA,… sẽ không sử dụng được.

5. Nhận xét và khuyến nghị

i) Tăng cường đào tạo, chia sẻ và tổng kết kinh nghiệm thiết kế và quản lý điều tra mẫu

Đến nay, mỗi năm Tổng cục Thống kê triển khai gần 30 cuộc điều tra kinh tế-xã hội với quy mô khác nhau. Với khối lượng điều tra khổng lồ như vậy, chúng ta chưa có điều kiện tổ chức nghiên cứu đánh giá đầy

đủ cả mặt ưu và nhược điểm, dẫn đến những khác biệt khá lớn giữa các đơn vị tham mưu cho Tổng cục cả về công tác thiết kế và tổ chức thực hiện các cuộc điều tra.

Để thực hiện có hiệu quả Chiến lược phát triển Thống kê Việt Nam thời kỳ 2011-2020 và tầm nhìn đến năm 2030, chúng ta cần tranh thủ mọi nguồn lực để đẩy mạnh công tác đào tạo thống kê chất lượng cao, thiết lập phương pháp luận thống kê chuẩn; xây dựng và quản lý có hiệu quả quy trình thiết kế, phát triển bảng câu hỏi, giám sát điều tra, xử lý và công bố số liệu.

ii) Xây dựng và định kỳ cập nhật dàn mẫu

Muốn thiết kế một mẫu điều tra tốt phải có dàn mẫu chuẩn. Dàn mẫu là một danh sách các đơn vị mẫu cơ bản (PSU), trong đó mỗi PSU phải có một số đo thể hiện quy mô của nó.

Điều tra thống kê thường áp dụng phương pháp thiết kế mẫu phân tầng-hệ thống. Trong nội bộ mỗi tầng, để mẫu có sai số thấp và hiệu quả cao, quy mô của các đơn vị mẫu cơ bản trong mỗi tầng cần có độ đồng đều cao. Ví dụ, trong điều tra mẫu về dân số-lao động, quy mô địa bàn điều tra cần xấp xỉ nhau với quy mô trung bình 500 nhân khẩu (100±20 hộ). Khi lấy thôn/ấp/bản/tổ dân phố làm đơn vị phân chia địa bàn, mỗi thôn/ấp/bản/tổ dân phố có quy mô lớn (trên 200 hộ) cần chia tách thành các địa bàn điều tra nhỏ hơn, song các địa bàn này phải có ranh giới rõ ràng nhằm phục vụ khâu chọn hộ và giúp điều tra viên tiếp cận hộ dễ dàng.

Trong mấy năm gần đây, tình trạng sử dụng chung một dàn mẫu chủ, thậm chí lấy mẫu điều tra của đơn vị này làm dàn mẫu để chọn một mẫu con cho đơn vị khác trong cùng ngành Thống kê, dẫn đến sử dụng trùng lặp đơn vị điều tra. Tình trạng này gây ra hiệu ứng nhàm chán cho các đối tượng điều tra, ảnh hưởng đến chất lượng thông tin được cung cấp. Để khắc phục, Tổng cục cần giao cho một đơn vị chủ trì phối hợp với các đơn vị liên quan thiết kế và quản lý một số dàn mẫu chủ dành riêng cho từng đơn vị nghiệp vụ. (Ví dụ dàn mẫu 100% và dàn mẫu chủ 15% trong kho dữ liệu Tổng điều tra dân số và nhà ở 2009 có thể thiết kế được nhiều dàn mẫu chủ để giao cho các đơn vị nghiệp vụ sử dụng.)

Do mức biến động dân số cao, cơ cấu và phân bố dân cư luôn thay đổi, tình hình phát triển kinh tế-xã hội diễn ra sôi động trong cả nước và từng địa phương, vì vậy định kỳ Tổng cục Thống kê cần tổ chức cập nhật dàn mẫu để phục vụ tốt công tác thiết kế mẫu điều tra./.

TÀI LIỆU THAM KHẢO

1. Tổng cục Thống kê (2006-2011). Báo cáo "những kết quả chủ yếu" của điều tra biến động dân số và KHHGĐ và Báo cáo điều tra lao động và việc làm.

2. Macro International Inc. (1996). Sampling Manual for Demographic and Health Surveys (Phase III).

3. Chris Scott and Truly Harpham (1975). Sample Design.

4. Vijay Verma, Christopher Scott and Colm O'muircheartaigh (1980). Sample Design and Sampling Errors for the World Fertility Surveys.

5. Janet L. Peacock (Prof.) and Philip J. Peacock (Dr.) (2011). Oxford Handbook of Medical Statistics.

6. D. N. Elhance and Veena Elhance (1956, 1992). Fundamentals of Statistics.

7. Báo cáo chuyến công tác của các chuyên gia mẫu đến từ UNSD, Macro International Inc., ILO.