CHƯƠNG 2
THIẾT KẾ, ƯỚC LƯỢNG VÀ ĐÁNH GIÁ KẾT QUẢ MẪU

1. Cỡ mẫu

Trong Tổng điều tra dân số và nhà ở năm 2009, ngoài việc điều tra toàn bộ, một số chỉ tiêu còn được tiến hành điều tra mẫu. Mẫu điều tra được thiết kế nhằm: (1) mở rộng nội dung điều tra; (2) nâng cao chất lượng điều tra, nhất là đối với những câu hỏi nhạy cảm và phức tạp; và (3) tiết kiệm kinh phí tổng điều tra. Để nâng cao hiệu quả và độ tin cậy của số liệu Tổng điều tra, quy mô mẫu là 15% tổng số dân của cả nước. Mẫu điều tra trong cuộc Tổng điều tra là mẫu chùm cả khối, được thiết kế theo phương pháp phân tầng-hệ thống một giai đoạn. Việc chọn mẫu được thực hiện theo hai bước: *Bước 1*, chọn phân tầng để xác định quy mô mẫu của từng huyện/quận/thị xã/thành phố trực thuộc tỉnh. *Bước 2*, chọn độc lập và hệ thống từ dàn mẫu địa bàn của mỗi huyện/quận/thị xã/thành phố trực thuộc tỉnh để xác định các địa bàn điều tra cụ thể.

Cỡ mẫu của điều tra chọn mẫu trong hai cuộc Tổng điều tra 1989 và 1999 tương ứng là 5% và 3%, chỉ đại diện cho cấp tỉnh; các chỉ tiêu điều tra mẫu chỉ là các câu hỏi về lịch sử sinh của phụ nữ 15-49 tuổi và các trường hợp chết của hộ trong 12 tháng trước điều tra. Trong Tổng điều tra năm 2009, ngoài hai chỉ tiêu nói trên, nhiều chỉ tiêu khác cũng được điều tra mẫu. Điều tra mẫu sẽ đưa ra số liệu đại diện cho đến cấp huyện.

Khi tính toán cỡ mẫu và phân bổ mẫu đã tính đến số sự kiện cần thu thập đối với các chỉ tiêu số trẻ em sinh, số người chết trong vòng 12 tháng trước thời điểm điều tra, số người thất nghiệp ở khu vực thành thị, v.v..; đồng thời cũng đảm bảo khả năng so sánh kết quả giữa các huyện, quận trong phạm vi một tỉnh, thành phố và giữa các tỉnh, thành phố với nhau.

2. Phân tầng và phân bổ mẫu cho các tầng

Để đảm bảo mức độ đại diện mẫu cho từng huyện, quận trong cả nước; do quy mô dân số phân bổ không đồng đều giữa các huyện, quận và các tỉnh, thành phố; Ban Chỉ đạo Trung ương quyết định phân bổ mẫu trực tiếp cho 682/684 huyện/quận (không tính 2 huyện đảo) trong cả nước theo 2 bước:

Bước 1: Xác định tỷ lệ mẫu $f^{(v)}$ cho 3 vùng gồm:

- Vùng 1: gồm 132 quận/thị xã/thành phố thuộc tỉnh;
- Vùng 2: gồm 294 huyện đồng bằng, ven biển;
- Vùng 3: gồm 256 huyện miền núi, hải đảo.

Bước 2: Phân bổ mẫu cho các huyện/quận trong mỗi vùng dựa trên cơ sở tỷ lệ mẫu của mỗi vùng đã được xác định ở bước 1. Áp dụng ***phương pháp phân bổ mẫu nghịch đảo.*** Theo phân bổ này, số lượng đơn vị mẫu của các huyện, quận có qui mô nhỏ được tăng lên đủ đảm bảo mức độ đại diện.

Công thức được sử dụng để tính tỷ lệ chọn mẫu cho từng huyện/quận trong từng vùng như sau:

$$f_i = \frac{a_i \times f_1}{1 + (a_i - 1) \times f_1}$$

Trong đó:
- i là số thứ tự của huyện quận ($i = 2,3,\ldots m_v$.);
- 1 là huyện/quận thứ nhất trong vùng;
- m_v là số đơn vị huyện/quận trong vùng (v=1,2,3);
- $a_i = N1/Ni$; Ni là dân số của huyện/quận thứ i;
- f_1 là tỷ lệ chọn mẫu của huyện, quận thứ nhất được tính theo công thức sau:

$$f_1 = \frac{\mathbf{f^{(v)}} \left(1 + \sum_{i=2}^{mv} a_i^{-1} \right)}{(m_v - 1)}$$

Kết quả xác định tỷ lệ mẫu của các vùng $\mathbf{f^{(v)}}$ theo phương pháp nêu trên được gia quyền, quyền số là dân số bình quân một đơn vị huyện của từng vùng ước tính đến 01 tháng 4 năm 2009 như sau: $f^{(1)}$ =13,11%; $f^{(2)}$ = 13,16% và $f^{(3)}$ = 22,68%. Kết quả phân bổ mẫu cho các huyện/quận và tỉnh/thành phố nêu tại Phụ lục 1.

3. Đơn vị và phương pháp chọn mẫu

Đơn vị chọn mẫu là địa bàn điều tra đã được phân định trong bước phân chia địa bàn điều tra. Dàn mẫu là danh sách các địa bàn điều tra được lập tuần tự theo

danh mục các đơn vị hành chính cấp xã trong từng huyện/quận/thị xã/thành phố thuộc tỉnh. Như vậy cả nước sẽ có 682 dàn mẫu (682 tầng).

Ban Chỉ đạo tỉnh, thành phố chịu trách nhiệm chọn ra các địa bàn điều tra mẫu theo phương pháp chọn mẫu ngẫu nhiên hệ thống như sau:

Bước 1: Lấy tổng số địa bàn điều tra trong huyện, quận chia cho số địa bàn điều tra cần chọn mẫu để xác định khoảng cách chọn (ký hiệu là "k", *lấy 1 số lẻ thập phân*).

Bước 2: Chọn số thứ tự đầu tiên (giả sử là số "b", điều kiện: $b \leq k$), ứng với địa bàn đầu tiên được chọn. Các địa bàn tiếp theo được chọn ứng với các số: $b_i = b + i \times k$; ở đây i = 1,2,3,...và dừng lại khi chọn đủ số địa bàn mẫu cần thiết.

4. Phương pháp ước lượng và suy rộng mẫu

Quyền số chung có thể được tính toán dựa vào xác suất/quyền số sau:

1) Quyền số thiết kế (quyền số cơ bản): dựa vào xác suất;
2) Hệ số hiệu chỉnh quyền số do thay đổi số hộ, khác biệt trong số hộ trung bình một địa bàn hoặc thay đổi tổng số địa bàn do mất đi mà không chọn thay thế;
3) Hệ số hiệu chỉnh quyền số theo cơ cấu giới tính, thành thị/nông thôn của tổng thể nghiên cứu (gia quyền).

Ký hiệu:

W_{1hji} - Quyền số thiết kế (quyền số cơ bản) của địa bàn *j*, tầng *h*;

W_{2hji} - Hệ số hiệu chỉnh quyền số do số hộ (dân số) thay đổi;

W_{3hji} - Hệ số hiệu chỉnh quyền số theo quy mô địa bàn trung bình của tầng *h*;

W_{4hji} - Hệ số hiệu chỉnh quyền số do số địa bàn điều tra thay đổi;

W_{5hji} - Hệ số hiệu chỉnh quyền số theo cơ cấu tổng thể nghiên cứu;

W_{hji} - Quyền số mẫu đối với hộ hoặc dân số nam/nữ của địa bàn *j* tầng *h*.

Xác định quyền số cơ bản

Giả sử a_h là số địa bàn điều tra được chọn trong tầng *h* và N_h là tổng số địa bàn của tầng *h*. Do mẫu được chọn độc lập ở từng tầng theo phương pháp ngẫu nhiên hệ thống, nên xác suất chọn cơ bản được tính theo công thức sau: $P_{1hji} = \frac{a_h}{N_h}$ và quyền số cơ bản (quyền số thiết kế) của địa bàn *j* thuộc tầng *h* là nghịch đảo của xác suất chọn, được tính như sau:

$$W_{1hji} = \frac{1}{P_{1hji}} = \frac{N_h}{a_h} \approx \frac{M_h}{\sum m_{hj}}$$

Trong đó, M_h là tổng số hộ (dân số) của tầng h và $\sum m_{hj}$ là tổng số hộ (dân số) của các địa bàn đã chọn điều tra của tầng h.

Xác định hệ số hiệu chỉnh quyền số do thay đổi số hộ (dân số) và số địa bàn

a) Hiệu chỉnh quyền số do thay đổi số hộ (dân số):
Giả sử m_{hj} là tổng số hộ (dân số) khi lập bảng kê của địa bàn *j* của tầng *h* và m^*_{hj} là tổng số hộ (dân số) khi điều tra của địa bàn *j* của tầng *h*. Hệ số hiệu chỉnh do thay đổi số hộ (dân số) được tính theo công thức sau:

$$W_{2hji} = \frac{1}{P_{2hji}} = \frac{m_{hj}}{m^*_{hj}}$$

b) Các địa bàn của Tổng điều tra dân số 2009 được phân chia với quy mô không đều nhau khoảng 100 hộ hoặc 500 nhân khẩu/địa bàn (cộng/trừ 20 hộ), nên cần phải xác định hệ số điều chỉnh quy mô hộ/dân số của các địa bàn Tổng điều tra dân số 2009 về quy mô hộ/dân số trung bình của tầng đó. Giả sử $\overline{m}_{hj}$ là tổng số hộ (dân số) trung bình của địa bàn thuộc tầng *h* và và hệ số hiệu chỉnh do thay đổi số hộ (dân số) được tính theo công thức sau :

$$W_{3hji} = \frac{1}{P_{3hji}} = \frac{\overline{m}_{hj}}{m_{hj}}$$

c) Hiệu chỉnh quyền số do thay đổi số địa bàn:

Điều tra mẫu 01 tháng 4 năm 2009 quy định: nếu địa bàn nào đã được chọn mà trong quá trình hiệu chỉnh sơ đồ - bảng kê phát hiện đã bị giải toả hoặc mất đi thì được phép thay thế bằng 1 địa bàn liền kề, không thay đổi tổng số địa bàn đã được chọn. Nên:

$$W_{4hji} = \frac{1}{P_{4hji}} = 1$$

Xác định hệ số hiệu chỉnh quyền số theo cơ cấu tổng thể (gia quyền theo tỷ trọng dân số nghiên cứu)

Số hộ/dân số có đến 01 tháng 4 năm 2009 sử dụng để gia quyền được ước tính dựa vào số liệu sơ bộ của Tổng điều tra theo thành thị/nông thôn và giới tính cho 63 tỉnh/thành phố, nên có thể gia quyền theo tỷ trọng dân số thành thị/nông thôn và dân số nam/nữ. Giả sử m^*_{hji} là tổng số hộ (dân số nam/nữ) khi điều tra của địa bàn *j* tầng *h*; $m^{*'}_{hji}$ là tổng số hộ (dân số nam/nữ) hiệu chỉnh theo tỷ trọng thành thị/nông thôn và nam nữ của địa bàn *j* tầng *h* và tính theo công thức:

$$m^{*'}_{hji} = m^*_{hj} \text{ x } \frac{M^*_{hi}}{M^*_h}$$

Trong đó:

m^*_{hj} số hộ (dân số nam/nữ) thu được từ điều tra mẫu của địa bàn *j* tầng *h*;

M^*_{hi} số hộ (dân số nam/nữ) chia theo thành thị/nông thôn sơ bộ đến 01 tháng 4 năm 2009 của tầng *h*; (i = 1 – thành thị ; i = 2 – nông thôn)

M^*_h số hộ (dân số) sơ bộ đến 01 tháng 4 năm 2009 của tầng h.

Hệ số hiệu chỉnh theo cơ cấu tổng thể của dân số (số hộ) ước lượng đến 01 tháng 4 năm 2009 được xác định như sau :

$$W_{5hji} = \frac{1}{P_{5hji}} = \frac{m^{*'}_{hji}}{m^*_{hji}} \text{x} \frac{M^*_h}{M_h} = \frac{m^*_{hj}}{m^*_{hji}} \text{x} \frac{M^*_{hi}}{M^*_h} \text{x} \frac{M^*_h}{M_h} = \frac{m^*_{hj}}{m^*_{hji}} \text{x} \frac{M^*_{hi}}{M_h}$$

Vì phân bổ mẫu là không tỷ lệ thuận đối với các tổng thể nghiên cứu, nên các quyền số mẫu sẽ được tính cho tất cả các phân tích sử dụng số liệu của điều tra mẫu trong Tổng điều tra dân số và nhà ở năm 2009 nhằm đảm bảo tính đại diện

thực tế của mẫu. Quyền số mẫu đối với mỗi hộ (hoặc dân số loại *i*) của địa bàn *j* thuộc tầng *h* là nghịch đảo của xác suất chọn:

$$W_{hji} = 1/P_{hji} = W_{1hji} \times W_{2hji} \times W_{3hji} \times W_{4hji} \times W_{5hji}$$

$$W_{hji} = 1/P_{hji} = \frac{M_h}{\sum m_{hj}} \times \frac{m_{hj}}{m^*_{hj}} \times \frac{\overline{m}_{hj}}{m_{hj}} \times \frac{m^*_{hj}}{m^*_{hji}} \times \frac{M^*_{hi}}{M_h} = \frac{\overline{m}_{hj}}{\sum m_{hj}} \times \frac{M^*_{hi}}{m^*_{hji}}$$

Văn phòng Ban Chỉ đạo Tổng điều tra dân số và nhà ở Trung ương đã phối hợp chặt chẽ với Trung tâm Tin học Thống kê Trung ương lập trình, tính toán cụ thể và kiểm tra chính xác các quyền số trên cho tất cả 30720 địa bàn mẫu của Tổng điều tra dân số và nhà ở năm 2009.

5. Phương pháp tính sai số mẫu

Các ước lượng từ điều tra mẫu bị ảnh hưởng của hai loại sai số: (1) sai số phi mẫu, và (2) sai số mẫu. Sai số phi mẫu là kết quả của các sai sót trong khi thực hiện thu thập và xử lý số liệu, như chọn sai ngôi nhà, chọn không đúng hộ, đối tượng điều tra không hiểu đúng câu hỏi cả từ phía điều tra viên và phía đối tượng điều tra, nhập tin sai. Mặc dù có nhiều cố gắng được thực hiện trong quá trình tiến hành điều tra nhằm giảm thiểu sai số loại này, nhưng sai số phi mẫu là không thể tránh khỏi và rất khó đánh giá về mặt thống kê.

Mặt khác, sai số mẫu có thể đánh giá được về mặt thống kê. Mẫu các đối tượng điều tra trong Tổng điều tra chỉ là một trong nhiều mẫu có thể được lựa chọn từ cùng một tổng thể nghiên cứu, sử dụng cùng một phương pháp thiết kế mẫu và cỡ mẫu đã định. Mỗi một trong các mẫu đó có thể cho kết quả khác với kết quả của mẫu thực tế đã chọn. Sai số mẫu là số đo sự biến thiên giữa tất cả các mẫu có thể có. Mặc dù mức độ biến thiên không thể biết được một cách chính xác, song nó có thể ước lượng được từ kết quả điều tra.

Sai số mẫu thường được đo bằng *sai số chuẩn* đối với một chỉ tiêu thống kê cụ thể (giá trị trung bình, phần trăm, …), sai số chuẩn chính là căn bậc hai của phương sai. Sai số chuẩn có thể sử dụng để tính khoảng tin cậy mà trong đó chứa giá trị đúng của tổng thể. Ví dụ, đối với một chỉ tiêu thống kê bất kỳ được tính từ điều tra mẫu, thì giá trị thống kê thực sẽ rơi vào trong khoảng cộng hoặc trừ hai lần

sai số chuẩn của chỉ tiêu đó với độ tin cậy 95 phần trăm của tất cả các mẫu có thể với cùng quy mô và cùng kiểu thiết kế mẫu.

Nếu đơn vị mẫu được chọn theo mẫu ngẫu nhiên đơn giản, thì mẫu đó có thể sử dụng các công thức trực tiếp để tính sai số mẫu. Tuy nhiên, mẫu của Tổng điều tra được thiết kế phân tầng, do đó phải dùng công thức phức tạp hơn. Phần mềm máy tính sử dụng để tính sai số mẫu cho các thiết kế dạng phân tầng có thể dùng một mô-đun tính sai số mẫu ISSA hoặc chương trình STATA. Các chương trình này sử dụng phương pháp tuyến tính hóa Taylor để ước lượng phương sai cho các ước lượng giá trị trung bình, tỷ trọng của các cuộc điều tra mẫu.

Phương pháp tuyến tính hóa Taylor xem chỉ tiêu phần trăm hoặc trung bình như là một ước lượng tỷ số, $r = y/x$, trong đó y là tổng giá trị mẫu của biến y, và x là số lượng các sự kiện trong nhóm hoặc nhóm con nghiên cứu. Phương sai của r được tính bằng công thức dưới đây, trong đó sai số chuẩn bằng căn bậc hai của phương sai:

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

trong công thức này:

$$z_{hi} = y_{hi} - rx_{hi}\text{, và} \quad z_h = y_h - rx_h$$

trong đó:

h - biểu thị tầng thay đổi từ 1 đến H,
m_h - là tổng số các địa bàn điều tra đã chọn trong tầng h,
y_{hi} - tổng các giá trị gia quyền của biến y của địa bàn i, trong tầng h,
x_{hi} - tổng số các sự kiện đã gia quyền của địa bàn i, tầng h, và
f - là tỷ lệ chọn mẫu chung, nếu giá trị này quá nhỏ thì có thể bỏ qua.

Sai số mẫu của Tổng điều tra được tính toán cho một số biến lựa chọn cần thiết nhất. Kết quả được trình bày trong phụ lục cho toàn quốc, thành thị và nông thôn, cho 6 vùng kinh tế xã hội và 63 tỉnh/thành phố. Với mỗi biến, các giá trị thống kê (R), sai số chuẩn (SE), sai số chuẩn tương đối (SE/R) và khoảng tin cậy 95 phần trăm (R±2SE) được đưa ra ở Phụ lục 4.

Khoảng tin cậy (ví dụ, như khi tính cho chỉ tiêu *tỷ số giới tính khi sinh*) có thể được giải thích như sau: tỷ số giới tính khi sinh tính chung từ mẫu của toàn

quốc là 110,5 trẻ em trai trên 100 trẻ em gái và sai số chuẩn là 0,54. Do đó, muốn có độ tin cậy là 95%, cộng và trừ hai lần sai số chuẩn đối với ước lượng, tức là, 110,5 ± 2 x 0,54. Với xác suất cao (95 phần trăm) thì tỷ số giới tính khi sinh của toàn quốc sẽ nằm trong khoảng 109,5 và 111,6 trẻ em trai/100 trẻ em gái.