MỘT SỐ KINH NGHIỆM THIẾT KẾ VÀ QUẢN LÝ MẪU ĐIỀU TRA

*Đồng Bá Hướng**

Chiến lược phát triển Thống kê Việt Nam giai đoạn 2011-2020 và tầm nhìn đến năm 2030 đã chỉ rõ: "*Áp dụng đồng bộ phương pháp thống kê tiên tiến và tăng cường sử dụng công nghệ hiện đại,… phấn đấu đến năm 2020 Thống kê Việt Nam đạt trình độ khá và đến năm 2030 đạt trình độ tiên tiến trong khu vực*".

Trong gần 30 năm qua, thông qua hợp tác đa phương và song phương, Tổng cục Thống kê đã áp dụng có hiệu quả mẫu tự gia quyền được thiết kế theo phương pháp phân tầng - hệ thống nhiều giai đoạn - là phương pháp thiết kế mẫu tiên tiến đã được khẳng định trong Hội nghị lần thứ 42 của Viện Thống kê Quốc tế tổ chức tại Ireland cuối năm 2011.

Bài báo này trình bày khái quát phương pháp thiết kế mẫu nói trên nhằm góp phần nhỏ bé vào tiến trình thực hiện Chiến lược về phát triển Thống kê Việt Nam. Mặc dù đôi chỗ chúng tôi có sử dụng mẫu và dàn mẫu thống kê dân số - lao động để minh họa, song phương pháp luận trình bày ở đây hoàn toàn có thể áp dụng để thiết kế các mẫu điều tra khác.

1. Một số khái niệm cơ bản

- *Mẫu tự gia quyền* là loại mẫu mà mỗi đơn vị mẫu cuối cùng được chọn với *xác suất chung* bằng nhau và khác 0. Trong mẫu tự gia quyền, xác suất chung có thể khác nhau giữa các tầng nhưng không khác nhau trong nội bộ một tầng. *Quyền số mẫu* bằng tỷ lệ nghịch của xác suất chọn mẫu. Như vậy, mẫu tự gia quyền luôn luôn có các quyền số bằng nhau.

Mẫu tự gia quyền có nhiều ưu điểm về kỹ thuật khi ước lượng mẫu và tính sai số mẫu, nó đòi hỏi phải tính các quyền số cho mỗi tầng.

- *Đơn giản hóa thiết kế:* Thiết kế mẫu càng đơn giản và trực tiếp thì càng tốt. Nguyên tắc này đòi hỏi cần tránh những phương pháp thiết kế quá nhiều giai đoạn và đáp ứng đa mục tiêu.

- *Cỡ mẫu (hay quy mô mẫu):* Chọn cỡ mẫu là sự cân đối giữa nhu cầu phân tích với khả năng tổ chức thực hiện và tài chính cho phép. Cỡ mẫu ở đây là nói đến quy mô mẫu mục tiêu. Các yếu tố làm giảm phạm vi điều tra như thiếu sót trong khâu vẽ sơ đồ/lập bảng kê, tỷ lệ không tiếp cận được hoặc không trả lời khi điều tra đều làm giảm quy mô mẫu mục tiêu với một mức độ nhất định. Vì vậy, cần phải tăng quy mô mẫu khi chọn để bảo đảm quy mô mẫu mục tiêu.

Thực tế thường áp dụng 3 cách xác định cỡ mẫu sau đây:

* Ủy viên BCH Hội Thống kê Việt Nam

a) Căn cứ vào độ tin cậy mẫu:

	Ước lượng giá trị trung bình	*Ước lượng tỷ trọng (tỷ suất, tỷ lệ)*
Độ tin cậy 90%:	$n= 1,64^2\times 4SD^2/d^2$	$n= 1,64^2\times 4\ p(1-p)/d^2$
Độ tin cậy 95%:	**$n= 1,96^2\times 4SD^2/d^2$** (1.1)	**$n= 1,96^2\times 4\ p(1-p)/d^2$** (1.2)
Độ tin cậy 99%:	$n= 2,58^2\times 4SD^2/d^2$	$n= 2,58^2\times 4\ p(1-p)/d^2$
Trong đó: ▪ 1,64; 1,96; 2,58 tương ứng là các điểm sai số hai phía 10%, 5%, và 1% của phân phối chuẩn; ▪ d là độ rộng mong muốn của khoảng tin cậy; ▪ SD là độ lệch chuẩn của giá trị trung bình cần ước lượng; ▪ p là tỷ trọng (tỷ suất, tỷ lệ) của dân số cần ước lượng.		

b) Căn cứ vào các nghiên cứu hiện thời:

Theo khuyến nghị của Liên hợp quốc, để bảo đảm đủ thông tin cho phân tích nhiều biến số, các cuộc điều tra mẫu quốc gia cần giới hạn phạm vi mẫu trong khoảng 2.000-8.000 hộ. Kinh nghiệm thực tế của các nước cho thấy, các mẫu điều tra cần có quy mô 5.000-6.000 hộ/tầng nhằm bảo đảm phân tích theo các biến chính.

c) Xác định cỡ mẫu theo phương pháp phân bổ tỷ lệ nghịch với quy mô dân số:

- Tỷ lệ chọn mẫu của tầng thứ nhất:

$$f_1 = \frac{f\left(1+\sum_{h=2}^{63} a_h^{-1}\right)}{63} \qquad (1.3)$$

Trong đó: $a_h = N_1/N_h$; N_h = tổng dân số (hoặc số hộ) của tầng thứ h

f là tỷ lệ mẫu chung = n/N (N là dân số cả nước, n là tổng quy mô mẫu).

- Tỷ lệ chọn mẫu của 62 tầng còn lại:

$$f_h = \frac{a_h \cdot f_1}{1+(a_h-1)\,f_1} \quad (h=2,3,\ldots,63) \qquad (1.4)$$

- Quy mô mẫu của tầng h: $n_h = f_h\times N_h$; (h=1,2,3,…,63).

▪ *Dàn chọn mẫu:* Dàn chọn mẫu được khai thác từ một dàn mẫu/dàn mẫu chủ hay danh sách các đơn vị mẫu cơ bản có sẵn của một cuộc tổng điều tra gần nhất, hoặc một mẫu lớn đã thực hiện trước đó.

Trong dàn mẫu, một số đơn vị mẫu cơ bản có quy mô quá lớn cần chia thành các đơn vị mẫu có quy mô chuẩn trước khi lập danh sách các đơn vị điều tra, mỗi đơn vị điều tra phải có một bản sơ đồ hoặc bản đồ có ranh giới rõ ràng.

▪ *Phân tầng:* Phân tầng là một quá trình mà, dựa trên những tiêu chuẩn nhất định, tổng thể chung được chia thành các nhóm hoặc các tầng càng đồng chất càng tốt. Mục đích của phân tầng là:

Thứ nhất, nhằm giảm sai số mẫu. Trong mẫu phân tầng, sai số mẫu phụ thuộc vào sự biến thiên của tiêu thức nghiên cứu trong nội bộ một tầng nhưng không phụ thuộc vào sự biến thiên giữa các tầng. Vì lý do này, người ta xây dựng các tầng bảo đảm có mức biến thiên nội bộ thấp.

Thứ hai, khi có sự khác nhau đáng kể giữa các nhóm (ví dụ giữa thành thị và nông thôn), phân tầng cho phép chọn cách phân bổ mẫu và thiết kế mẫu một cách riêng biệt cho mỗi nhóm (thành thị, nông thôn).

▪ *Quy mô "lấy" mẫu trên một chùm:*

Số lượng hộ tối ưu được lựa chọn trong một chùm, hay còn được gọi là quy mô "lấy" mẫu/chùm, phụ thuộc vào biến số nghiên cứu. Ví dụ, khi ước lượng tỷ lệ sử dụng biện pháp tránh thai và các yếu tố

tác động của nó, các biến số chính liên quan có xu hướng "tập trung" cao, vì vậy quy mô chùm trung bình tối ưu thường từ 15-20 phụ nữ trong độ tuổi sinh đẻ/chùm. Tuy nhiên, các biến số khác về mức sinh có mức độ "phân tán" cao, nên quy mô "lấy" mẫu tối ưu còn cao hơn nhiều. Theo khuyến nghị của Liên hợp quốc, một chùm cần "lấy" khoảng 30-40 hộ ở khu vực nông thôn và 20-30 hộ ở khu vực thành thị. Trong thống kê dân số-lao động, "chùm" được hiểu là địa bàn điều tra.

▪ *Một số ký hiệu*: PSU = đơn vị mẫu cơ bản (trong thống kê dân số, đó là địa bàn điều tra); USU = đơn vị mẫu cuối cùng hay đơn vị mẫu cấp hai (trong thống kê dân số, đó là "hộ").

2. Các kỹ thuật chọn mẫu

- Chọn mẫu không theo tỷ trọng giữa các tầng:

Có hai lý do phải chọn mẫu không theo tỷ trọng giữa các tầng:

Thứ nhất, hiệu quả đầu tư tăng lên nếu tăng tỷ lệ mẫu trong các tầng có phương sai lớn và chi phí thấp hơn; vì thế tỷ lệ chọn mẫu có thể được điều chỉnh nhằm đạt thiết kế tối ưu.

Thứ hai, người lập kế hoạch điều tra muốn báo cáo các phát hiện cho nhóm chiếm tỷ lệ nhỏ trong toàn bộ tổng thể. Nếu sử dụng một tỷ lệ chọn mẫu cố định (chọn tỷ lệ thuận với quy mô) thì những tầng có quy mô nhỏ sẽ được phân bổ một mẫu nhỏ. Sai số mẫu tăng lên theo sự giảm đi của cỡ mẫu, và kết quả là, do cố định tổng cỡ mẫu, sai số mẫu của các tầng nhỏ này sẽ lớn tới mức không thể chấp nhận được.

Khi tăng mẫu cho một tầng có quy mô nhỏ sẽ làm giảm sai số mẫu của các ước lượng. Đây là chiến lược có giá trị đối với thiết kế mẫu vì đã cung cấp đủ mẫu cho các tầng khác nhau. Vì các vùng/tỉnh khác nhau về quy mô, nên phương pháp phân bổ trên dẫn đến phân số mẫu khác nhau giữa các vùng/tỉnh - đó là phương pháp *phân bổ tỷ lệ nghịch với quy mô.*

- Gia quyền mẫu:

Nếu một mẫu có quy mô n được chọn từ một tổng thể có quy mô N, sử dụng phương pháp thiết kế xác suất bằng nhau với xác suất chọn là n/N, khi đó tổng thể chung được ước lượng bằng cách nhân tổng thể mẫu với N/n. Thừa số N/n gọi là *hệ số suy rộng*.

Thay vì ước lượng tổng số, người ta có thể muốn ước lượng một số bình quân, một tỷ suất, một tỷ trọng hay một tỷ lệ (sau đây gọi tắt là "tỷ lệ"). Vì các quyền số xuất hiện cả ở tử số và mẫu số của ước lượng cuối cùng, nên mọi thừa số chung được khử khỏi quyền số. Theo đó, các số bình quân hay tỷ lệ có thể được tính trực tiếp từ mẫu: giá trị mẫu cho ta một ước lượng trực tiếp của giá trị tổng thể. Hầu hết các con số trong điều tra là các số bình quân hay tỷ lệ, các con số này không đòi hỏi hệ số suy rộng và không cần đến các quyền số khi đã sử dụng thiết kế mẫu xác suất bằng nhau. Tuy nhiên, nếu xác suất chọn mẫu thay đổi giữa các tầng (hoặc giữa các đơn vị điều tra), thì các quyền số phải được áp dụng riêng cho mỗi tầng (hoặc mỗi đơn vị điều tra) có xác suất chọn mẫu khác nhau.

Phương pháp chuẩn để đưa các quyền số vào phân tích là đưa "biến quyền số" vào mỗi bản ghi cá nhân. Khi một biến quyền số được đưa vào theo cách này, người ta có thể trực tiếp sử dụng nó để thực hiện mọi phân tích theo yêu cầu.

- Chọn mẫu hệ thống:

Chọn mẫu hệ thống là chọn các đơn vị mẫu tại một khoảng cách cố định của một dàn mẫu, bắt đầu từ một điểm được xác định ngẫu nhiên.

So với phương pháp chọn ngẫu nhiên đơn giản, chọn mẫu hệ thống có 3 ưu điểm: (1) dễ thực hiện; (2) dễ điều chỉnh kết quả đã chọn; (3) nếu dàn mẫu được xếp thứ tự theo một tiêu thức nào đó, phương pháp này cho ta biết cấp độ phân tầng tương ứng với biến số làm nền cho dàn mẫu. Trên thực tế,

hầu hết các dàn mẫu đều được xếp thứ tự với một mức độ nhất định.

Chọn mẫu hệ thống thường được tiến hành như sau: Giả sử có một khoảng cách đo bằng số nguyên (hay số thập phân) là I, đầu tiên xác định một số ngẫu nhiên R sao cho R≤ I. Các đơn vị mẫu được chọn là các số R, R+I, R+2I, v.v… cho đến hết danh sách (dàn mẫu). Nếu thiết kế xác định được số đơn vị cần chọn, thì khoảng cách I được tính bằng N/n sau khi đã làm tròn số, trong đó N là số đơn vị trong danh sách và n là số đơn vị được chọn. Mặt khác, nếu thiết kế xác định được phân số mẫu hay xác suất chọn mẫu f, khi đó khoảng cách được tính bằng I = 1/f.

Việc chọn mẫu theo khoảng cách tính bằng số thập phân được tiến hành như sau:

i. Tính khoảng cách I làm tròn đến một số thập phân.

ii. Xác định một số ngẫu nhiên R nằm giữa 1 và 10*I và đặt dấu thập phân trước chữ số cuối cùng của R.

iii. Tính dãy các số mẫu: R, R+I, R+2I, …

iv. Phần nguyên của mỗi số mẫu chính là đơn vị mẫu được chọn.

- Thiết kế mẫu tự gia quyền theo phương pháp PPS:

Giai đoạn 1 sẽ chọn các đơn vị điều tra (PSU) theo xác suất tỷ lệ thuận với quy mô đã ước lượng của PSU. Vì vậy, nếu ước lượng quy mô đơn vị A lớn gấp 10 lần đơn vị B, thì xác suất chọn đơn vị A sẽ cao gấp 10 lần đơn vị B. Phương pháp chọn mẫu này được gọi là *chọn mẫu với xác suất tỷ lệ thuận với quy mô*, hay *mẫu PPS*.

Phương pháp PPS để chọn PSU trong giai đoạn 1 như trên sẽ cho một mẫu "lệch" về phía các đơn vị lớn. Một phương pháp điều chỉnh mức độ "lệch" là sử dụng hệ thống nghịch đảo tại giai đoạn 2 (chọn USU, trong dân số đó là chọn "hộ"), nghĩa là chọn mẫu với xác suất tỷ lệ nghịch với quy mô. Điều này có nghĩa là tỷ lệ chọn USU ở đơn vị A sẽ thấp hơn ở đơn vị B 10 lần, do vậy khử được độ "lệch" đã phát sinh trong giai đoạn 1 (chọn PSU).

Qua hai giai đoạn (chọn PSU và chọn USU), một USU (hộ) cụ thể của đơn vị A bây giờ có cùng một xác suất chọn như đơn vị B. Theo phương pháp thiết kế mẫu này, nếu các ước lượng quy mô đã sử dụng *ở giai đoạn 1* (chọn PSU) luôn đúng bằng số USU trong từng PSU (*chọn cả khối*) hay hoàn toàn tỷ lệ thuận với số USU đó, thì *ở giai đoạn 2* (chọn USU) người ta chọn được một số lượng USU cố định trong mỗi đơn vị PSU.

Sự hấp dẫn của mẫu PPS là khối lượng công việc cho mỗi PSU là như nhau (vì là một mẫu tự gia quyền). Công tác tổ chức điều tra thực địa sẽ thuận lợi hơn do khối lượng công việc đồng đều ở tất cả các đơn vị điều tra. Ưu điểm này càng có ý nghĩa khi mẫu được phân bổ cho các đơn vị hành chính thường có quy mô dân số rất khác biệt nhau. Một ưu điểm quan trọng hơn của mẫu PPS là sai số mẫu giảm đi khi ước lượng các chỉ tiêu tổng số. Nói chung, loại mẫu này làm tăng hiệu quả mẫu khi ước lượng các số bình quân và tỷ lệ.

Chọn mẫu PPS được thực hiện cho tầng h như sau:

+ Lập danh sách các đơn vị điều tra (PSU) có quy mô ước lượng M_{hi} của mỗi PSU.

+ Cộng dồn các giá trị M_{hi} và ghi tương ứng vào dòng của mỗi PSU. Kiểm tra rằng số cộng dồn cuối cùng phải bằng tổng các giá trị M_{hi}, hay $\sum M_{hi}$.

+ Nếu a_h là số PSU cần chọn, tính khoảng cách chọn mẫu $I_h = \sum M_{hi}/a_h$.

+ Chọn một số ngẫu nhiên R nằm giữa 1 và I_h.

+ Tính các số mẫu R, R+ I_h, R+2 I_h,…, R+(a_h-1) I_h.

+ Đối với mỗi số mẫu, tìm số cộng dồn đầu tiên của M_{hi} bằng hoặc lớn hơn số mẫu. Đơn vị tương ứng đó sẽ được chọn.

(Còn tiếp)

Tài liệu tham khảo **(xem trang 44)**

(tiếp theo trang 21)

Tài liệu tham khảo:

1. Tổng cục Thống kê (2006-2011). Báo cáo điều tra biến động dân số và KHHGĐ và Báo cáo điều tra lao động và việc làm hàng năm.

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 and Philip J.Peacock (2011). Oxford Handbook of Medical Statistics.

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

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