NHỮNG YÊU CẦU VỀ DỮ LIỆU PHỤC VỤ ĐÁNH GIÁ THIẾT KẾ MẪU TRONG CÁC CUỘC ĐIỀU TRA CHỌN MẪU HỘ GIA ĐÌNH

Tóm tắt:

Đánh giá thiết kế mẫu là một trong những nội dung để đánh giá chất lượng cuộc điều tra chọn mẫu. Ở nhiều quốc gia, đặc biệt là những quốc gia có ít kinh nghiệm khi tiến hành điều tra hộ gia đình, các sổ ghi chép và những báo cáo điều tra thường cung cấp dữ liệu đặc tả (metadata) rất hạn chế. Điều này làm xuất hiện những sai số trong việc phân tích điều tra, vì vậy, trong các cuốn cẩm nang về điều tra nêu bật tầm quan trọng việc lưu trữ các bản ghi chi tiết về dữ liệu đặc tả, nó giúp việc phân tích được thực hiện đúng và đưa ra các biện pháp để đánh giá thiết kế mẫu. Bài viết này giới thiệu một số yêu cầu về dữ liệu phục vụ đánh giá thiết kế mẫu và được giới thiệu trong cuốn Sổ tay "Thiết kế điều tra chọn mẫu hộ gia đình" của Liên hợp quốc.

1. Dữ liệu khi xây dựng phương án chọn mẫu và thực hiện phương án chọn mẫu

Bất cứ một cuộc điều tra chọn mẫu nào cũng cần có phương án chọn mẫu (thiết kế mẫu). Chuyên gia về kỹ thuật chọn mẫu không chỉ có nhiệm vụ cung cấp dữ liệu trong khi xây dựng phương án chọn mẫu mà còn phải lưu trữ dữ liệu khi triển khai thực hiện cuộc điều tra đó. Phương án chọn mẫu thường đòi hỏi phải phù hợp với công việc thực địa ở các giai đoạn khác nhau, lường trước những tình huống phát sinh trong quá trình điều tra. Điều quan trọng là ghi lại từng bước tất cả các hoạt động đã xảy ra trong quá trình thực hiện phương án chọn mẫu, để đảm bảo việc thực hiện là đúng với thiết kế. Trong quá trình triển khai, phương án chọn mẫu thay đổi, dù chỉ là những thay đổi nhỏ có thể sẽ nghiêm trọng hơn cả việc cung cấp tất cả các sai lệch từ mẫu. Thông tin về những thay đổi trong quá trình thực hiện này thật sự cần thiết ở giai đoạn phân tích về sau. Trong trường hợp bắt buộc, phương án chọn mẫu vẫn phải thay đổi nhưng cần được lưu ý khi xây dựng phương án cho các điều tra trong tương lai.

2. Dữ liệu về đánh mã cho các đơn vị chọn mẫu

Trong từng giai đoạn của thiết kế mẫu, xác định các đơn vị được chọn vào mẫu phải gắn với việc đánh mã rõ ràng và duy nhất. Khi đó phải thiết lập các mã cho các đơn vị lấy mẫu ban đầu, thứ hai, thứ ba và cấp cuối cùng (phụ thuộc vào thiết kế mẫu bao nhiêu giai đoạn). Thông thường giai đoạn đầu tiên mã gồm bốn chữ số là đủ và mã có ba chữ số cho các giai đoạn còn lại. Các tên miền địa lý phải được ghi nhãn đúng cách. Ngoài ra, các mã hành chính xác định địa lý, cấu trúc hành chính của các khu vực chứa đơn vị chọn mẫu là một phần của quy trình ghi mã.

Ví dụ: Giả sử, một thiết kế mẫu gồm hai giai đoạn, với 1200 đơn vị chọn mẫu ban đầu (PSU). PSU được định nghĩa là 1 đơn vị địa bàn trong Tổng điều tra. *Giai đoạn thứ nhất*, chọn 600 mẫu cho mỗi tầng là nông thôn và thành thị. Dễ dàng nhất để đánh mã cho các PSU là từ 1 đến 1200 (việc đánh mã như này cũng sử dụng để lựa chọn các PSU phục vụ việc tính toán phương sai mẫu). Do đó, nếu các PSU khu vực nông thôn được lựa

chọn trước thì chúng sẽ được đánh mã từ 0001 đến 0600 trong khi đó các PSU khu vực thành thị được đánh mã từ 0601 đến 1200. Cách mã hóa như vậy có hai ưu điểm: (1) Mỗi PSU được đánh số và xác định duy nhất; (2) Các nhà phân tích có thể ngay lập tức nhận ra PSU thành thị hay nông thôn dựa vào mã của nó. *Giai đoạn thứ hai* của chọn mẫu, mỗi PSU chọn 20 hộ gia đình để phỏng vấn. Ở giai đoạn này, tất cả các hộ gia đình đã được liệt kê sẽ được cấp một mã số gồm ba chữ số (hoặc bốn chữ số nếu một số địa bàn điều tra có hơn 999 hộ), một lần nữa mã số được đánh theo thứ tự chúng được liệt kê. Các hộ được chọn vào mẫu sẽ giữ lại mã số được đánh mã theo cách này. Cuối cùng, những mã hành chính được chỉ định khi cần thiết. Do đó, một hộ gia đình được chọn vào mẫu có thể được mã hoá là 09 003 008 0128 080. Mã này được hiểu như sau: Đó là hộ gia đình thứ 80 được liệt kê (được chọn để phỏng vấn) trong PSU 0128 thuộc xã 008 của quận 003 tỉnh 09. Hơn nữa, nhìn mã số của PSU ngay lập tức cho biết hộ gia đình là thuộc khu vực nông thôn. Nếu cuộc điều tra thu thập thông tin về các thành viên của hộ gia đình, mỗi người trong số họ sẽ có một mã duy nhất gồm hai chữ số từ 01 đến 99.

Tóm lại, việc đánh mã phù hợp là điều thật sự cần thiết, lý do rõ nhất là: (1) Để kiểm soát chất lượng. Nhiệm vụ được phân công cho điều tra viên và bảng hỏi được ghi tại nơi điều tra sẽ được kiểm tra lại theo danh sách để đảm bảo rằng tất cả các hộ gia đình được chọn vào mẫu đều được thực hiện đúng; (2) Cách đánh mã duy nhất này có giá trị vô giá cho các cán bộ thực hiện xử lý dữ liệu bởi vì các bảng biểu có thể lập được theo khu vực địa lý.

3. Dữ liệu về xác suất chọn mẫu

Một nội dung thông tin thường bị bỏ qua trong tài liệu mẫu là tính toán xác suất chọn mẫu cho các đơn vị mẫu ở các giai đoạn khác nhau. Hoặc nếu có thông tin, nó thường chỉ giới hạn ở quyền số mẫu chung (đã được tính toán từ xác suất chung) cho từng mẫu.

Một chi tiết đặc biệt quan trọng trong các tài liệu hướng dẫn có đề cập đến là khi phải lấy mẫu con phát sinh tại thực địa. Nó xảy ra khi một mẫu chùm quá lớn hoặc khi có nhiều hơn một hộ gia đình trong cùng một chỗ ở (khi chỗ ở là đơn vị lập danh sách). Việc ghi lại cẩn thận tỷ lệ mẫu con là rất cần thiết. Xác suất lựa chọn của mẫu chùm (đã thay đổi) và xác suất chọn hộ gia đình có thể tính toán một cách chính xác bởi các cán bộ chọn mẫu và do đó quyền số được điều chỉnh một cách chính xác.

Ghi lại xác suất chọn mẫu ở từng giai đoạn là rất hữu ích. Ví dụ, xác suất lựa chọn mỗi PSU là khác nhau khi sử dụng bất cứ cách chọn mẫu pps (phương pháp chọn mẫu xác suất tương ứng với quy mô). Điều này đúng ngay cả khi thiết kế mẫu chung là mẫu tự cân đối quyền số. Nếu không xác định được xác suất lựa chọn PSU thì không thể tính đúng quyền số nếu các PSU này cần được lấy mẫu con cho các cuộc điều tra tiếp theo.

4. Dữ liệu về tỷ lệ trả lời và tỷ lệ bao phủ ở các giai đoạn chọn mẫu khác nhau

Là một phần của quá trình đánh giá, để kiểm tra việc thực hiện điều tra mẫu cần cung cấp thông tin cho người sử dụng về tỷ lệ trả lời và tỷ lệ bao phủ. Thông tin càng nhiều và chi tiết, sẽ hữu ích cho việc đánh giá. Hơn nữa, không những phải cung cấp thông tin về tỷ lệ trả lời (hoặc bổ sung, tỷ lệ không trả lời, mà còn tổng hợp các lý do không trả lời. Không trả lời thường gồm các lý do sau: Không có người ở nhà; đơn vị nhà ở thiếu (mất mẫu); từ chối trả lời; nghỉ ngơi tạm thời (nghỉ phép, v.v...).

Ngoài tỷ lệ trả lời, tỷ lệ bao phủ mẫu là một nội dung được chú ý khi chọn mẫu ở các giai đoạn khác nhau. Có trường hợp các chùm chọn vào mẫu nhưng không thể phỏng vấn vì gặp phải vấn đề an ninh, xung đột hoặc rối loạn dân sự hoặc khó có khả năng tiếp cận do địa hình hoặc do thời tiết. Khi đó, sẽ phải lựa chọn các chùm thay thế. Cần thiết một quy trình thay thế nghiêm túc bởi vì những cư dân của chùm thay thế hầu hết khác những người trong chùm được thay thế. Khi có thay thế được thực hiện, đội ngũ điều tra cần phải ghi lại số lượng và vị trí của các chùm đó. Đặc biệt quan trọng là phải cung cấp một số thông tin về mức độ bao phủ của mẫu được thay thế, bằng cách đưa ra các ước tính số người trong tổng thể mục tiêu trong các khu vực mà các chùm được thay thế đại diện.

Các phức tạp sẽ được giảm, nếu xác định được những chùm khó tiếp cận trước khi lựa chọn mẫu. Những đối tượng xác định này nên được loại trừ khỏi phạm vi điều tra trước khi lấy mẫu và các báo cáo điều tra nên đề cập rõ ràng về các khu vực này không "đại diện" bằng mẫu.

5. Quyền số: Quyền số cơ sở, điều chỉnh không trả lời và điều chỉnh khác

Tính toán quyền số của các cuộc điều tra hộ gia đình nói chung bao gồm ba loại: Quyền số cơ sở[6] (còn gọi là quyền số thiết kế); quyền số điều chỉnh không trả lời; quyền số điều chỉnh sau phân tầng.

Trong nhiều trường hợp chỉ sử dụng quyền số cơ sở (loại thứ nhất), trong khi ở những trường hợp khác, quyền số cơ sở được điều chỉnh bằng một yếu tố bổ sung để phản ánh việc không trả lời bảng hỏi điều tra (quyền số điều chỉnh không trả lời, loại thứ hai). Chỉ một số rất ít trường hợp, quyền số có thể phản ánh một yếu tố khác, có hoặc không có điều chỉnh không trả lời, nhằm điều chỉnh phân bố tổng thể dựa trên kết quả từ mẫu để khẳng định phù hợp sự phân bố từ một nguồn dữ liệu độc lập. Loại cuối cùng thường được gọi là quyền số sau phân tầng. Không phải trường hợp nào cũng tính quyền số này. Nó chỉ được tính khi hai điều kiện được đáp ứng: Mẫu phải là mẫu tự gia quyền và số liệu chỉ ở dạng chỉ tiêu tỷ lệ (phần trăm, tỷ số hoặc tỷ lệ so với tổng số ước lượng hoặc so với số tuyệt đối).

Khi sử dụng quyền số thống kê, cần thiết phải ghi chép các tính toán một cách cẩn thận. Như đã đề cập trước đó, các quyền số (hoặc xác suất chọn mẫu) ở mỗi giai đoạn lựa chọn phải được tính toán và ghi lại. Ngoài ra, cần đo quyền số riêng ở mỗi giai đoạn của hoạt động dữ liệu, nghĩa là: (1) Quyền số cơ sở, (2) Quyền số thiết kế sau khi nhân với hệ số điều chỉnh không trả lời và (3) cuối cùng là nhân hệ số điều chỉnh sau phân tầng.

6. Những dữ liệu về kinh phí thực hiện

Mặc dù các cuộc điều tra hộ gia đình thường được cấp ngân sách rất đầy đủ, nhưng cần giữ các hồ sơ về chi tiêu thực tế của các hoạt động khác nhau trong quá trình thực hiện điều tra. Thông tin về kinh phí sẽ hữu ích cho hoạt động chọn mẫu, đặc biệt hữu ích đối với thiết kế mẫu chủ, cũng như để xây dựng phương án chọn mẫu của các cuộc điều tra trong tương lai.

Các hoạt động lấy mẫu phải được giám sát cẩn thận về các chi phí bao gồm:

1. Lương cho hoạt động thiết kế mẫu bao gồm cả phí cho bất kỳ một tư vấn từ chuyên gia bên ngoài;

[6] *Quyền số cơ sở là nghịch đảo của xác suất chọn mẫu của đơn vị chọn mẫu cuối cùng.*

2. Chi phí thực địa để cập nhật dàn mẫu bao gồm công cho người thực hiện và chi phí chuẩn bị các tài liệu (bản đồ, danh sách...);

3. Chi phí về công nghệ thông tin để chuẩn bị dàn mẫu phục vụ chọn mẫu của các PSU;

4. Chi phí cho người thực hiện chọn mẫu các PSU (nếu không được thực hiện bằng máy tính);

5. Chi phí thực địa để tiến hành hoạt động niêm yết ở các đơn vị lấy mẫu ở giai đoạn gần cuối, bao gồm công cho người thực hiện và chuẩn bị các tài liệu.

6. Chi phí cho người thực hiện chọn mẫu hộ gia đình.

Như vậy thông tin kinh phí là nhân tố quan trọng trong việc đánh giá thiết kế mẫu.

7. Sai số chọn mẫu

Phần lớn các mục đã đề cập ở trên rất hữu ích cho việc đánh giá thiết kế mẫu và quá trình thực hiện điều tra cũng như để xử lý các kết quả điều tra. Thông tin về tỷ lệ phản hồi được sử dụng để đánh giá kết quả điều tra, trong khi chi phí lấy mẫu có thể được sử dụng để đánh giá hiệu quả của thiết kế mẫu và phục vụ cho các cuộc điều tra trong tương lai.

Tuy nhiên, một thành phần quan trọng hơn cả của đánh giá mẫu là tính toán sai số chọn mẫu cho các chỉ tiêu chính của cuộc điều tra. Một trong những đặc điểm phân biệt một mẫu xác suất là bản thân mẫu đó có thể được sử dụng để tính toán sai số chuẩn. Chỉ cần tính toán sai số chuẩn cho các chỉ tiêu quan trọng, được quan tâm trong cuộc điều tra vì nó không thực tế và cũng không cần thiết phải tính toán cho tất cả các chỉ tiêu. Các sai số chuẩn là thông tin để người dùng đánh giá độ tin cậy của ước lượng điều tra và xây dựng các khoảng tin cậy xung quanh các ước lượng điểm.

Các sai số chuẩn cũng có thể được dùng để đánh giá thiết kế mẫu. Một thống kê nữa đặc biệt hữu ích để đánh giá thiết kế mẫu là hiệu quả thiết kế mẫu (viết tắt deff[7], hoặc chính xác hơn là giá trị deft, là căn bậc hai của deff). Giá trị deft được tính toán đơn giản khi biết sai số chuẩn. Deff được tính bằng cách chia sai số chuẩn được tính toán (cho mỗi chỉ tiêu cụ thể) cho sai số chuẩn có được từ một mẫu ngẫu nhiên đơn giản có cùng cỡ mẫu, cụ thể là pq/n, trong đó p là tỷ lệ ước lượng; q = 1 - p và n là kích thước mẫu. Tính giá trị này nhằm xác nhận hoặc bác bỏ các hiệu quả thiết kế đã được giả định khi mẫu đang được thiết kế, vì các giá trị deff (hoặc deft) thực tế không thể có cho đến khi cuộc điều tra được tiến hành, các dữ liệu được xử lý và các sai số chuẩn được tính toán.

Các nhà thống kê chọn mẫu có thể sử dụng các hiệu quả thiết kế để đánh giá xem các cỡ mẫu của chùm có hợp lý cho các chỉ tiêu quan trọng và có động tác khắc phục nếu cần. Ví dụ, nếu deft lớn hơn nhiều so với tính toán đối với một số chỉ tiêu, thì trong tương lai mẫu cho một cuộc điều tra có thể được thiết kế để sử dụng các kích thước của chùm nhỏ hơn.

Như vậy để đánh giá thiết kế mẫu cho cuộc điều tra chọn mẫu không chỉ có một số thông số tính toán từ mẫu (sai số mẫu, giá trị deff...) mà quan trọng là lưu trữ đầy đủ càng chi tiết càng tốt các dữ liệu từ khi xây dựng phương án mẫu, trong quá trình thực hiện phương án mẫu đến khi công bố thông số có được từ mẫu điều tra.

Vân Anh (lược dịch và tổng hợp)

Nguồn: United Nations New York, 2008, Designing Household Survey Samples, Practical Guidelines, Series F No.98

[7] *Design effect*