

MỘT SỐ VẤN ĐỀ VỀ KIỂM TRA TÍNH ĐẠI DIỆN CỦA MẪU

Phạm Sơn
Viện Khoa học Thống kê

Hiện nay trong công tác thống kê phương pháp điều tra chọn mẫu đang được áp dụng khá phổ biến và đã đáp ứng được phần nào nhu cầu của các cơ quan quản lý, hoạch định chính sách cũng như các cơ quan nghiên cứu. Tuy nhiên một vấn đề bức xúc đặt ra là: mẫu được chọn có tính đại diện đến đâu. Thường câu hỏi này sau khi kết thúc điều tra mới được công bố, thậm chí nhiều cuộc điều tra do ngành Thống kê tiến hành cũng không đánh giá phạm vi sai số. Do vậy một vấn đề đặt ra là bằng cách nào để trả lời hay nói một cách khác chúng ta khẳng định với các đơn vị tiến hành điều tra chọn mẫu rằng “Mẫu được chọn có khả năng đáp ứng được yêu cầu đặt ra”. Với mục đích đó, bài viết sẽ trình bày một số suy nghĩ về kiểm tra tính đại diện trong quá trình xây dựng mẫu.

I. Những vấn đề chọn mẫu cần quán triệt trong điều tra chọn mẫu

1. Xác định cỡ mẫu

Trong lý thuyết điều tra chọn mẫu đã định ra nhiều công thức xác định cỡ mẫu cho các hình thức tổ chức chọn mẫu khác nhau như: chọn giản đơn không hoàn lại, chọn giản đơn có hoàn lại, chọn xác suất đều, chọn xác suất không đều, chọn hệ thống, chọn phân tổ, chọn cả khối... Dù chọn theo hình thức nào, thì cỡ mẫu (n) đều phụ thuộc vào các yếu tố sau:

- Lược đồ chọn (ký hiệu là P)

- Quy mô của tổng thể (ký hiệu là N) và người ta cũng chứng minh được rằng cỡ mẫu tỷ lệ thuận với quy mô tổng thể.

- Mức độ phức tạp của hiện tượng nghiên cứu được đo bằng độ lệch chuẩn (s) hoặc phương sai (σ) của chỉ tiêu dùng làm căn cứ xác định cỡ mẫu.

- Mức độ sai số cho phép (ký hiệu là Δ).

- Xác suất tin cậy (hệ số tin cậy - ký hiệu là t).

Một cách tổng quát, cỡ mẫu được xác định theo công thức:

$$n = f(P, N, \sigma, \Delta, t) \quad (1)$$

Trong các giáo trình lý thuyết thống kê, hoặc lý thuyết điều tra chọn mẫu đã đưa ra nhiều công thức tính cỡ mẫu cụ thể (xem [1], [2], [3]) để cho cán bộ thực hành nghiên cứu ứng dụng. Ở đây, sẽ phân tích kỹ về bản chất của cỡ mẫu (n) được xác định theo công thức (1).

a. Khối lượng mẫu chỉ là một ước lượng

Chúng ta biết rằng, để xác định cỡ mẫu cho một cuộc điều tra chọn mẫu cụ thể cần phải có thông tin tiên nghiệm về 5 yếu tố trên, trong đó chỉ có lược đồ chọn là tương đối ổn định và tùy thuộc vào người thiết kế mẫu, 4 yếu tố còn lại suy cho cùng chỉ là những ước lượng.

Trước hết, độ lệch chuẩn hay phương sai của tổng thể hầu như không có và phần nhiều “vay mượn” của cuộc điều tra trước

hoặc các cuộc điều tra tương tự. Trong trường hợp không có những phương sai có sẵn, cơ quan tiến hành thiết kế mẫu thường tổ chức điều tra nhỏ để ước lượng phương sai (σ). Tiếp đến, phạm vi sai số (Δ), chỉ đưa ra mức. Chẳng hạn, trong điều tra năng suất lúa người ta quy định mức sai số khoảng 0,2 – 0,3 tấn/ha.

Yếu tố độ tin cậy (t), cũng được xác định tùy theo tính chất của từng cuộc điều tra và cuối cùng quy mô của tổng thể càng khó xác định chính xác mà chỉ là con số “ước lượng”. Do hầu hết “đầu vào” đưa vào công thức (1) đều là ước lượng, nên đầu ra (cỡ mẫu n) cũng chỉ là một ước lượng mà thôi.

b. Cỡ mẫu xác định theo công thức trên mới phản ánh một phần cơ bản hiện tượng nghiên cứu

Chúng ta biết rằng, để phản ánh các hiện tượng kinh tế xã hội cần một hệ thống chỉ tiêu, chẳng hạn để đánh giá mức sống các tầng lớp dân cư số chỉ tiêu thống kê thường lên đến vài trăm. Nhưng khi xác định cỡ mẫu theo công thức (1) chỉ dựa vào phương sai của một chỉ tiêu được gọi là cơ bản, còn hầu hết các chỉ tiêu khác chưa được phản ánh vào mẫu vì nhiều lý do, cả vấn đề học thuật lẫn thực tế. Về học thuật, cho đến nay chúng ta mới có cỡ mẫu tối ưu cho một chỉ tiêu, còn tối ưu cho nhiều chỉ tiêu chưa được đầu tư nghiên cứu. Theo tài liệu Hội nghị chọn mẫu quốc tế được tổ chức tại Vác-xô-vi tháng 8 năm 1968, mới khuyến cáo đây là một hướng nghiên cứu cần quan tâm. Song cho đến nay hầu như chưa có công trình nào nghiên cứu vấn đề này được công bố. Hướng thứ hai chọn ra một số chỉ tiêu chủ yếu và tiến hành thu thập tính toán để ước lượng phương sai của các chỉ tiêu đó, sau đó lấy chỉ tiêu có phương sai lớn nhất để

đưa vào xác định cỡ mẫu. Cách làm này, mới khắc phục được một phần bài toán nêu ra, bởi lẽ nó thực ra cũng chỉ tối ưu theo một chỉ tiêu, chứ chưa phải là tối ưu cho nhiều chỉ tiêu.

Từ một số điểm trình bày trên, chúng ta dễ dàng thấy rằng, dù có dùng bất cứ công thức xác định cỡ mẫu nào và với công sức bỏ ra nhiều bao nhiêu thì kết quả cũng chỉ là ước lượng để các cơ quan tổ chức điều tra chọn mẫu tham khảo và có quyết định phù hợp. Cũng vì lẽ đó, trên thực tế người ta căn cứ vào tiềm lực tài chính và yêu cầu cần đáp ứng để chọn cỡ mẫu cho phù hợp.

Còn đối với các cuộc điều tra chọn mẫu được lặp lại nhiều lần, hướng cơ bản là sử dụng số liệu của các điều tra trước để giảm cỡ mẫu (xem [4]).

2. Phân bố mẫu

Như chúng ta đều biết với cùng một cỡ mẫu n như nhau, nhưng việc phân bố cỡ mẫu vào các bộ phận (các tổ) khác nhau sẽ cho những kết quả hoàn toàn khác nhau. Trong lý thuyết điều tra chọn mẫu vấn đề này được các nhà chọn mẫu học đầu tư nghiên cứu và đã định ra một số phương pháp phân bố mẫu khác nhau, trong đó phổ biến nhất gồm:

a. Phân bố mẫu tỷ lệ thuận với quy mô tổng thể. Cách phân bố này có ưu điểm là những bộ phận có quy mô lớn, số lượng mẫu sẽ rơi vào nhiều và việc ước lượng các tham số mẫu, cũng như các tham số của tổng thể nghiên cứu thuận lợi.

b. Phân bố cỡ mẫu theo phương pháp cân bằng quyền số. Đây là một biến thể của phương pháp phân bố mẫu tỷ lệ với quy mô của tổng thể và ở nước ta trong một số cuộc

tổng điều tra dân số 1979 và 1989 đã áp dụng cho mẫu điều tra sinh tử (xem [5], [6]).

c. Phân bổ mẫu theo phương pháp Maitra (xem [7]) ưu điểm lớn nhất của phương pháp phân bổ mẫu này là những tổ có quy mô nhỏ có được “tiếng nói” quan trọng trong tổng thể mẫu.

d. Phân bổ tối ưu Neyman (xem [1]) phương pháp phân bổ mẫu này vừa chú ý đến 2 yếu tố: quy mô tổ và tính chất biến động của chỉ tiêu nghiên cứu. Cách phân bổ mẫu này cho ta sai số nhỏ nhất.

Ngoài ra, trong thực tiễn công tác thống kê của các nước, cũng như của nước ta các nhà chọn mẫu học đưa ra một số phương pháp phân bổ mẫu khác nhau (xem [1], [2], [3], [4] và [5]).

Cũng như cỡ mẫu, việc phân bổ mẫu cho các bộ phận chỉ dựa vào mấy yếu tố sau:

- Quy mô từng bộ phận trong tổng thể (Ni) với $N = \sum_{i=1}^k N_i$;

- Tính chất phức tạp của chỉ tiêu nghiên cứu trong từng bộ phận được đo bằng độ lệch chuẩn hay phương sai (σ_i ; $i= 1,2,3,\dots,k$)

Trong 4 phương pháp phân bổ mẫu thông dụng có 3 phương pháp (a, b và c) chỉ chú ý đến quy mô, không chú ý đến tính chất của chỉ tiêu nghiên cứu, còn phương pháp d - Phân bổ tối ưu Neyman có sử dụng cả 2 yếu tố, được tính theo công thức

$$N_i = \frac{n\sigma_i N_i}{\sum \sigma_i N_i} \quad (2)$$

Từ những điều trình bày trên có thể đưa ra một số nhận xét sau:

Thứ nhất, bất kỳ phương pháp phân bổ mẫu nào chỉ tập trung vào quy mô của các

bộ phận cấu thành tổng thể sẽ gặp nhiều hạn chế vì trên thực tế quy mô của các bộ phận cấu thành lên tổng thể có sự khác biệt nhau rất lớn. Thí dụ trong điều tra doanh nghiệp có những xí nghiệp số lượng lao động, vốn, tài sản và giá trị sản xuất lớn gấp hàng chục lần, thậm chí hàng trăm lần so với các xí nghiệp khác. Hiển nhiên, có những tổng thể như dân số, số lượng đơn vị trong các bộ phận cấu thành xem ra khá giống nhau. Nhưng nếu xét kỹ cũng có nhiều vấn đề phải tìm hiểu nghiên cứu.

Từ phân tích trên đây thấy rằng, việc phân bổ mẫu theo tiêu chí quy mô còn bộc lộ nhiều hạn chế và do đó kết quả phân bổ mẫu tuy hết sức quan trọng, nhưng chưa giải quyết tốt vấn đề tính đại diện của mẫu.

Thứ hai, phân bổ mẫu kết hợp giữa quy mô và tính chất hiện tượng nghiên cứu, tuy đã khắc phục phần nào hạn chế của phương pháp phân bổ mẫu dựa vào quy mô, song trên thực tế để có số liệu tính toán độ lệch chuẩn, hay phương sai của các bộ phận gặp hết sức khó khăn. Do vậy, trong thực tiễn công tác thống kê của nước ta, cũng như nhiều nước trên thế giới phương pháp này ít được áp dụng.

Từ hai vấn đề chủ chốt của lý thuyết mẫu (cỡ mẫu và phân bổ mẫu), chúng ta thấy rằng việc xây dựng mẫu phụ thuộc nhiều vào phương pháp chọn. Phải chăng, vì lý do đó mà nhiều nhà thống kê học gọi tắt lý thuyết mẫu bằng tên “chọn mẫu” và trong lý thuyết chọn mẫu đã đưa ra nhiều hình thức tổ chức chọn mẫu khác nhau (xem [1], [2], [3]).

(còn nữa)

Tài liệu tham khảo:

[1] PGS.TS. Tăng Văn Khiên, Lý thuyết điều tra chọn mẫu, NXB Thống kê, Hà Nội - 2003.

[2] Phạm Sơn, Một số vấn đề lý luận và thực tiễn áp dụng phương pháp điều tra chọn mẫu trong thực tiễn công tác thống kê nước ta, Thông tin Khoa học Thống kê số 3 - 1992.

[3] SIAP, Tài liệu tập huấn điều tra chọn mẫu (Sample survey), Tokyo - 1990

[4] Leslie Kish, Survey Sampling, P.H.John Weley, 1995.

[5] Houston Dominique và Nguyễn Phong, Điều chỉnh cỡ mẫu điều tra mức sống dân cư Việt Nam, theo số liệu 2001-2002, Thông tin Khoa học Thống kê số 3-2003.