Xã hội học, số 2 - 1989

MỘT SỐ VẤN ĐỀ VỀ SAI SỐ TRONG NGHIÊN CỨU BẰNG MẪU

NGUYỄN THỊ PHƯƠNG

Thông thường các cuộc nghiên cứu xã hội học thực nghiệm được tiến hành trên một bộ phận nào đó của tập hợp tổng, nghĩa là nghiên cứu trên tập hợp mẫu. Để tránh những sai số lớn, việc chọn mẫu cần phải được tiến hành một cách nghiêm túc khoa học cụ thể phải tuân theo một số nguyên tắc nhất định là:

1 - Tránh độ chệch trong biện pháp lựa chọn.

2 - Bảo đảm độ chính xác tối đa với cùng một phí tổn nhất định.

Người ta gọi độ lệch của các cấu trúc thống kê của mẫu số với cấu trúc của tổng thể là sai số của mẫu. Có nhiều sai số gặp phải khi sử dụng phương pháp mẫu, song chung quy có thể xếp thành hai loại cơ bản: *sai số ngẫu nhiên và sai số hệ thống*.

Trước khi nghiên cứu các sai số của mẫu, cần tìm hiểu việc lập mẫu, thông qua đó sẽ thấy những sai số thường xuất hiện như thế nào. Khi tổ chức các thủ tục chọn trực tiếp các đối tượng, phải luận chứng cấu trúc của mẫu trên quan điểm các nhiệm vụ nghiên cứu. Và không thể đại diện cho tập hợp tổng theo vô số các tính chất của nó nên cấu trúc của mẫu hoàn toàn được qui định bởi đặc tính của các giả thiết nghiên cứu: Cấu trúc của mẫu dựa trên cơ sở các giả thiết cơ bản, nên cần phải có thông tin xã hội về tập hợp tổng.

Chẳng hạn: Khi nghiên cứu tập hợp các khán giả điện ảnh, nhà xã hội học có thể đưa ra giả thiết về sự phụ thuộc thị hiếu của người xem vào trình độ học vấn của họ. Nếu biết rằng tập hợp tổng cần nghiên cứu có 15% người có trình độ đại học 40% có trình độ trung học và 45% chưa tốt nghiệp trung học, thì phải giữ tỷ lệ đó trong mẫu.

Việc lập mẫu với tư cách là mô hình của tổng thể phải dựa trên việc tính đến các mối quan hệ và liên hệ thực tạo nên cấu trúc của tổng thể. Nếu cầu trúc của mẫu không tương ứng với những đặc trưng trong cấu trúc tổng thể mà người nghiên cứu quan tâm thì dẫn đến sai số và từ đó sẽ có những kết luận sai lầm.

Như đã nói trên, có hai loại sai số ngẫu nhiên và sai số hệ thống. Sai số ngẫu nhiên là những sai số xảy ra do những vi phạm ngẫu nhiên trong các thủ tục thu thập thông tin. Nghĩa là những sai số thường mắc phải về mặt kỹ thuật trong quá trình điều tra thực địa. Có hai nguyên nhân chính dẫn đến sai số ngẫu nhiên.

Một là: do sai lệch về đặc trưng của phân phối mẫu so với phân phối của tổng thể, do sự khác nhau về kích thước của hai tập hợp đó (mẫu và tổng). Khi tính đến các chỉ tiêu của mẫu phải để ý tới sai số này.

Trong phương pháp mẫu không cho lời giải vạn năng đối với vấn đề dung lượng của mẫu. Nói chung dung lượng của mẫu phụ thuộc hai yêu tố: *Mức độ đồng nhất các tập hợp tổng và mức độ chính xác cần thiết của các kết quả rút ra ở mẫu*. Mẫu càng lớn tính đại diện càng cao, càng nhỏ tính đại diện càng ít.

Bản quyền thuộc Viện Xã hội học www.ios.org.vn

Ví dụ: Cán bộ điều tra tới nhà một số gia đình được chọn vào mẫu để tìm hiểu những công việc trong hợp tác xã, những vấn đề sản xuất ảnh hưởng tới cuộc sống hàng ngày của nông dân. Song đã không thu được thông tin chính xác, vì người được hỏi không còn sản xuất trong hợp tác xã. Nguyên nhân dẫn đến sai sót lý do ngay từ khi chọn mẫu, người ta đã không để ý tới tính đồng nhất trong mẫu. Ở nông thôn không phải người nào cũng tham gia sản xuất ở hợp tác xã. Những người này hoặc đã thôi sản xuất trực tiếp, về già, bỏ đi buôn, trông nhà, trông cháu, nghỉ ngơi... hoặc làm các công việc khác mà trong hộ khẩu không nói tới. Như vậy những người này không mang đặc trưng của mẫu mà ta cần chọn là phải sản xuất trực tiếp trong hợp tác xã. Với cách chọn như vậy đặc trưng của phân phối mẫu không nằm trong đặc trưng của tổng thể, vì thế sai số có thể tăng lên.

Hai là: do độ lệch không điều khiển được so với mẫu đã kế hoạch hóa trước, tức là những sai số quan sát và những sai số của thủ tục thu thập thông tin.

Những yếu tố thường dẫn đến các sai lầm đó là: các đơn vị nghiên cứu đã được qui định trong kế hoạch lấy mẫu bị thay bằng các đơn vị khác để đạt hơn, nhưng lại không có giá trị đầy đủ theo quan điểm của kế hoạch lấy mẫu đã được dự định trước. Chẳng hạn: Cần trưng cầu ý kiến ở một khu đông dân cư, ta qui định cứ 15 hộ chọn một hộ vào mẫu, số của căn hộ được xác định theo một thủ tục chọn lọc hệ thống. Khi cán bộ đi phỏng vấn không gặp ai trong căn hộ đã chọn, liền phỏng vấn luôn căn hộ bên cạnh. Kết quả trong mẫu điều tra có quá nhiều người về hưu, nhiều gia đình đông người vì những gia đình này thường có người ở nhà. Như vậy những người được điều tra ít đại diện cho những người độc thân, hộ ít người, vì những người này thường ít ở nhà. Như vậy những người được điều tra ít đại diện cho những người độc thân, hộ ít người, vì những người này thường ít ở nhà mà số gia đình này lại chiếm số đông trong khu.

Do đó cần phải cử những cán bộ có trình độ chuyên môn cần thiết tham gia điều tra. Mặt khác sau mỗi cuộc điều tra phải có sự kiểm tra kĩ lưỡng các phiếu thu được công việc này do những cán bộ có trình độ cao nhiều kinh nghiệm đảm nhận. Việc chọn mẫu càng tuân thủ đúng những qui tắc là: mẫu đủ lớn, đại diện cho tổng v.v.. càng tránh được những sai số. Nhất là khi tính chỉ tiêu của mẫu phải chú ý tới sai số ngẫu nhiên, số đo của sai số này bằng số đo của sai số đại diện.

Công thức tính sai số đại diện như sau:

+ Trong mẫu lặp:

$$M = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Trong đó:

M: sai số đại diện

σ^2: phương sai dấu hiệu trong tổng

n: dung lượng của mẫu

Qua công thức trên ta thấy:

M nhỏ, khi σ nhỏ (nghĩa là tổng càng thuần nhất), hoặc n thật lớn (dung lượng lớn).

Trong thực hành tính toán theo công thức trên, muốn tính được M phải tính được σ, tuy nhiên thường người ta không biết được σ. Vì vậy để làm một ước lượng của phương sai của tổng, người ta sử dụng phương sai mẫu đã hiệu đính.

Bản quyền thuộc Viện Xã hội học www.ios.org.vn

$$\partial^2 = s^2 \times \frac{n}{n-1}$$

s: phương sai mẫu

Đối với mẫu có dung lượng n > 100 thì $\frac{n}{n-1} \sim 1$ khi đó $\sigma^2 \sim s^2$ nên M = $\frac{s}{\sqrt{n-1}}$, nhưng vì n >100, nên $\sqrt{n-1} \sim \sqrt{n}$, suy ra M ~ $\frac{s}{\sqrt{n}}$

+ Trong mẫu không lặp:

M = $\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$

Trong đó N: dung lượng tổng khi n ~ N thì $\left(1-\frac{n}{N}\right) \sim 0$ khi đó M ~ 0

n rất nhỏ so với N: $\left(1-\frac{n}{N}\right) \sim 0$ khi đó M ~ $\frac{s}{\sqrt{n}}$ giống mẫu lặp, như vậy việc xác định sai số đại diện rất quan trọng đối với mẫu nào (có dung lượng nhỏ vì khi đó sai số đại diện M ~ $\frac{s}{\sqrt{n}}$ là đại lượng đáng kể, tính được, có thể ảnh hưởng đến chất lượng thông tin thu được.

Thực tế điều tra xã hội học là sử dụng mẫu không lặp, nhưng khi tính toán vẫn sử dụng công thức như mẫu lặp vì những lý do trên.

+ Trong phép lấy mẫu nhiều nấc, công thức tính sai số đại diện có dạng như sau:

Đối với mẫu hai nấc, ta thường lấy mẫu ngẫu nhiên nghiêm ngặt ở nấc thứ nhất, nấc thứ hai là mẫu xác xuất tỉ lệ. Khi có mẫu dung lượng đủ lớn công thức tính sai số hai nấc là:

$$M = \sqrt{\frac{s_1^2}{n_1} + \sum\left(\frac{s_2^2}{n_2}\right) \times \frac{1}{n_1}}$$

Trong đó: s_1^2: phương sai của những đơn vị ở cấp chọn thứ 1

s_2^2 : phương sai của những đơn vị ở cấp chọn thứ 2

n_1: số lượng những đơn vị ở cấp chọn thứ 1

n_2: số lượng những đơn vị ở cấp chọn thứ 2.

Trong công, thứ sẽ chứa hai nguồn sai số đại diện:

Số hạng thứ nhất $\frac{s_1^2}{n_1}$ chỉ phương sai gây ra do sự lập nấc đầu tiên của phép lấy mẫu - là phương sai các trung bình nhóm (các phần), là nguồn đầu tiên của các sai số ngẫu nhiên (chỉ phần của tổng thể chưa được cho vào mẫu).

- Số hạng thứ hai $\frac{s_2^2}{n_1} \times \frac{1}{n_1}$ chỉ phương sai trong nhóm gắn liền với việc tổ chức nấc hai của phép lấy mẫu, nguồn thứ hai của sai số ngẫu nhiên. Phương sai trong nhóm được tính trong mỗi đơn vị của nấc đầu, sau khi đã chọn từ nó các đơn vị thuộc nấc thứ hai. Những sai số này được xác định bằng

Bản quyền thuộc Viện Xã hội học www.ios.org.vn

cách đối chiếu mẫu đã lập được trong thực tế với kế hoạch lập mẫu đó. Những sai số này có thể giảm bớt được bằng cách dùng thủ tục "hiệu chỉnh" mẫu, tức là bằng việc tổ chức thu thập thông tin bổ sung những thông tin không đầy đủ.

Ngoài nhưng sai số do vi phạm ngẫu nhiên còn những sai số không mang đặc tính ngẫu nhiên, do việc tái tạo không đầy đủ trong mẫu những phân bố chính có thể

Bản quyền thuộc Viện Xã hội học www.ios.org.vn

không mang đặc tính ngẫu nhiên. Sai số thuộc loại này là sai số hệ thống. Sai số hệ thống dẫn đến việc làm sai lệch trong mẫu những phân bố chính: hoặc nâng cao hoặc hạ thấp quá mức những giá trị tổng thể. Những sai số hệ thống có thể làm cho kết quả của toàn bộ cuộc điều tra mất giá trị.

Những nguyên nhân gây ra sai số hệ thống là: mẫu lập không thích ứng với nhiệm vụ nghiên cứu.

Ví dụ: Nghiên cứu thị hiếu của thanh niên thành phố, song chỉ điều tra bộ phận sinh viên đại học, là những người có thành phần xuất thân khác nhau, như từ nông thôn ra, bộ đội về học tiếp, học sinh phổ thông lên... Những người này chiếm số lượng khá lớn trong toàn bộ thanh niên thành phố, nhưng không đại diện cho các tầng lớp thanh niên thành phố mà ta dự định sẽ chọn nghiên cứu. Nếu cứ tiếp tục tiến hành điều tra mẫu lập ra sẽ không thích ứng với nhiệm vụ. Nguy hiểm hơn là có trường hợp số đơn vị trong mẫu thiếu, nên người đi điều tra chọn thêm một số thanh niên ở nơi khác bù vào. Lượng thông tin do số người này cung cấp sẽ hoàn toàn không thích hợp với yêu cầu nghiên cứu, vì những thanh niên này chỉ đại diện cho địa bàn của họ.

Do vậy, khi lập mẫu phải quan tâm tới tính đầy đủ của mẫu, nghĩa là tất cả các phần tử của tổng thể, đều được đại diện trong cơ sở của mẫu.

Do không biết đặc tính phân phối trong tổng thể, trong thủ tục lấy mẫu có thể làm sai lệch những phân phối đó (nhất là trong tổng thể không đồng nhất về mặt thống kê). Chẳng hạn: Khi dùng cách lấy mẫu ở trên cơ sở bảng lương, trong đó cá nhân được xếp theo thứ tự lương tăng dần. Trong trường hợp nếu mẫu chứa phần đầu của bảng, gồm những người có mức lương thấp (những người mới vào nghề, trình độ chuyên môn lao động thấp...) sẽ làm hạ mức lương trung bình trong mẫu. Nếu lấy mẫu ở phần cuối của bảng, sẽ dẫn đến nâng cao giá trị của đại lượng đó. Vì vậy khi lấy mẫu, nếu không chú ý tính đồng nhất trong tổng, sẽ dẫn đến sai số có hệ thống.

Sai số hệ thống còn xuất hiện khi sử dụng mẫu tự phát, loại mẫu này đại diện cho tổng thể nào nhiều khi không biết, nên không nắm được đặc tính phân phối của tổng cần nghiên cứu.

Ví dụ: Tiến hành trưng cầu ý kiến nhờ các bảng kê khai phát trên đài truyền thanh hay trên vô tuyến truyền hình hoặc đăng trên báo. Các thính giả hoặc độc giả của một hệ thống (kênh) thông tin quần chúng, đóng vai trò là tổng thể cần nghiên cứu. Nhưng không biết đặc trưng của tổng thể và thường không biết quy mô của tổng thể đó, nên không thể xác định được chất lượng của mẫu. Lúc đó mẫu chưa chắc đã đại diện cho tổng thể, chưa nói có khi mẫu là bức tranh méo mó của tổng.

Khi tổ chức lấy mẫu phân nhóm đại diện, việc lấy mẫu từ các nhóm điển hình của tổng phải được tiến hành tỉ lệ với quy mô của chúng trong toàn tổng thể.

Ví dụ: Trong xã hội tỉ lệ công nhân viên trí thức là khác nhau, nhưng dùng mẫu điển hình ta đều lấy mỗi tầng lớp đó 1000 người để nghiên cứu, khi tổng kết sẽ cho một sai số lớn, đó là sai số hệ thống.

Khi lấy mẫu cơ học, trong tổng có thể tồn tại một trật tự sắp xếp nhất định các cá thể, theo trật tự tăng hay giảm của một dấu hiệu nào đó, cũng dễ dẫn đến nguy cơ phạm phải sai số hệ thống. Do vậy khi lấy mẫu hệ thống phải tính đến khả năng có sự sắp xếp hệ thống trong bảng danh sách các đơn vị, mà sự sắp xếp này có thể trùng với giá trị của khoảng chọn. Chẳng hạn, khi lập cơ sở của mẫu để trưng cầu ý kiến của công nhân thuộc một phân xưởng của nhà máy. Khoảng cách chọn có thể trùng với số công nhân trong các đơn vị sản xuất, mà người đứng đầu danh sách các đội thường là

Bản quyền thuộc Viện Xã hội học www.ios.org.vn

đội trưởng. Khi chọn khoảng lấy mẫu trùng với số người của đội, mẫu có thể hoàn toàn bị chệch, nghĩa là mẫu có thể gồm hoặc các đồng chí đội trưởng hoặc các đồng chí có bậc lương cao hoặc toàn những người có lương thấp. Trong trường hợp này, người nghiên cứu đã chọn các phần tử của mẫu quá điển hình.

Sai số hệ thống còn có thể do nguyên nhân sau:

Chọn lọc một cách có ý thức các phần tử thuộc tổng thể, để thuận tiện và có lợi cho việc tiến hành thu thập thông tin, nhưng các phần tử này lại không đại diện cho tổng nói chung. Ví dụ: Trong cuộc điều tra thanh niên công nhân của một thành phố, do một số nguyên nhân (có thể để dễ tiếp cận với các khách thể, điều kiện đi lại thuận tiện, học do không để ý...) phần lớn các nhà máy được chọn trong mẫu là các nhà máy xí nghiệp công nghiệp nhẹ. Do tính chất công việc của những nhà máy này, tỉ lệ nữ trong mỗi nhà máy rất lớn, nên thông tin thu được từ cách chọc như vậy là do đại đa số nữ cung cấp. Kết quả đã nâng tỷ trọng nữ trong mẫu lên rất cao. Vì thế mẫu này không đại diện cho tổng thể cần nghiên cứu là thanh niên công nhân nói chung.

Muốn tránh được những sai số khi chọn mẫu để đảm báo chất lượng thông tin thu được có độ tin cậy cao, người nghiên cứu phải chú ý tới các yếu tố sau:

- Lập mẫu phải được thực hiện bằng mọi phương pháp ngẫu nhiên, nghĩa là việc lựa chọn không bị ảnh hưởng một cách có ý thức, hay không ý thức bởi sự lựa chọn của con người.

- Cơ sở mẫu (danh sách, mục lục hoặc các ghi chép khác về tổng thể, được dùng làm cơ sở cho sự lựa chọn) phải bao quát đầy đủ và đáng tin cậy các cá thể của tập hợp tổng.

- Mọi bộ phận của tổng (những khách thể trong mẫu) luôn có khả năng được chọn và chịu cộng tác với những cán bộ nghiên cứu.

Bất cứ yếu tố nào trong các yếu tố trên bị vi phạm, đều có thể dẫn đến sai số hệ thống. Nếu mẫu lấy từ một cơ sở mẫu không thích hợp, thì dù có tăng dung lượng lên cũng sẽ không sửa được tính không đại diện của nó, hoặc không loại bỏ được độ chệch trong các đặc điểm của một số mẫu đã được chọn như vậy.

Để tránh sai số hệ thống người ta thường dùng phương pháp lấy mẫu ngẫu nhiên để nhận được mẫu đại diện. Theo các định luật của toán học, phương pháp lấy mẫu ngẫu nhiên sẽ không xuất hiện các sai số hệ thống dĩ nhiên và điều kiện là phép lấy mẫu như vậy có thể thực hiện được). Mặt khác trong phép lấy mẫu ngẫu nhiên chỉ có sai số ngẫu nhiên, mà những sai số này sẽ giảm khi tăng dung lượng của mẫu.

Trên đây là những vấn đề chung nhất về sai số trong nghiên cứu bằng mẫu. Thực tế trong nghiên cứu xã hội học thực nghiệm, nó còn những biến thái khác nhau tùy theo từng đề tài cụ thể. Chỉ có tổ chức nghiêm ngặt, luận chứng đầy đủ các vấn đề, mới tránh được những sai số đáng tiếc có thể xảy ra.

TÀI LIỆU THAM KHẢO CHÍNH

1. Phương pháp thống kê toán phân tích số liệu trong nghiên cứu xã hội học. Viện hàn lâm khoa học Liên Xô, 1980. Tiếng Nga, Bản dịch ở phòng tư liệu Viện xã hội học.

2. Lê Văn Phong, Thống kê toán tập 4. Trường Đại học kinh tế quốc dân, Hà Nội.

3. Sách công tác của nhà xã hội học, chương V, VI, VII. Viện Hàn lâm khoa học Liên Xô, 1976. Tiếng Nga, bản dịch ở phòng tư liệu Viện Xã hội học.

4. C.A. Boser và Katon, Phương pháp điều tra trong điều tra nghiên cứu xã hội, tập 1 và tập 3. Tiếng Anh, 1977. Bản dịch ở phòng tư liệu Viện Xã hội học.

Bản quyền thuộc Viện Xã hội học www.ios.org.vn