VẤN ĐỀ CHỌN CÁC ĐƠN VỊ LỚN VÀO MẪU VÀ TÍNH ĐẠI DIỆN CỦA MẪU

Phạm Thành Đạo(*)

1. Từ thực tiễn công tác thống kê

Trong điều tra thu thập số liệu thống kê, nhiều trường hợp chúng ta thu thập số liệu của các đơn vị điển hình (vì những đơn vị lớn này có vai trò quan trọng trong lĩnh vực nghiên cứu). Chẳng hạn trong điều tra công nghiệp toàn quốc, yêu cầu có tính nguyên tắc là phải có số liệu của thành phố Hà Nội và thành phố Hồ Chí Minh, là những địa phương có số lượng doanh nghiệp nhiều, chiếm giá trị sản xuất công nghiệp đáng kể và có nhiều đơn vị lớn. Nhưng nếu chúng ta chọn các đơn vị lớn vào mẫu với cách suy rộng các đặc trưng của tổng thể qua số đơn vị của tổng thể thì số liệu bị đẩy lên quá cao không chấp nhận được. Đứng trước vấn đề này thống kê các nước và chúng ta đã chọn cách giải quyết như sau:

- Chia tổng thể chung thành 3 nhóm: Các đơn vị lớn, các đơn vị vừa và các đơn vị nhỏ. Đối với các đơn vị lớn điều tra toàn bộ, đối với các đơn vị vừa điều tra với tỷ lệ cao hơn, đối với các đơn vị nhỏ điều tra với tỷ lệ thấp, còn mẫu vẫn phải rải theo các yêu cầu đại diện phù hợp.

- Điều tra các đơn vị lớn để tính các chỉ số thành phần và dùng một hệ thống quyền số để tính các chỉ số tổng hợp như phương pháp điều tra chỉ số sản xuất công nghiệp hàng tháng của Vụ thống kê Công nghiệp và xây dựng, nghiên cứu triển khai với sự giúp đỡ của JICA.

Đây là giải pháp tốt, tuy vậy các nhà nghiên cứu luôn đặt câu hỏi hiện còn có giải pháp nào khác không? Bài viết đề cập tới câu hỏi này.

Ta hãy bắt đầu bằng việc tham khảo định lý sau:

Trong lý thuyết xác suất thống kê toán có một định lý nói rằng nếu các biến ngẫu nhiên gốc tuân theo phân phối chuẩn $X_i \in N(m, \sigma^2)$ (i =1, 2, 3.....n) thì biến ngẫu nhiên $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ có phân phối xấp xỉ phân phối chuẩn $N(m, \frac{\sigma^2}{n})$ [1]. Nghĩa là:

Với một mẫu ngẫu nhiên $X_1, X_2, ... X_n$, và ta dùng $\overline{X}$ để ước lượng cho m thì 68% các mẫu ngẫu nhiên có chênh lệch giữa $\overline{X}$ với m nhỏ hơn $\sqrt{\frac{\sigma^2}{n}}$ nghĩa là sai số của $\overline{X}$ là $\sqrt{\frac{\sigma^2}{n}}$, hoặc có 95% các mẫu ngẫu nhiên có sai số nhỏ hơn $2\sqrt{\frac{\sigma^2}{n}}$, hoặc có 99% các mẫu ngẫu nhiên có sai số nhỏ hơn $3\sqrt{\frac{\sigma^2}{n}}$,... Như vậy, trong số các mẫu ngẫu nhiên có 68% các mẫu ngẫu nhiên "tốt" mà nếu ta chọn được một trong số này để điều tra thì sai số chỉ là $\sqrt{\frac{\sigma^2}{n}}$ (sai số bình quân

(*) Viện Khoa học Thống kê

chọn mẫu) còn nếu ta không có cách nào để chọn được một mẫu ngẫu nhiên "tốt" như vậy thì ta đành phải chấp nhận mẫu của ta nằm trong số 95% hoặc 99% các mẫu ngẫu nhiên ít tốt hơn có sai số là $2\sqrt{\frac{\sigma^2}{n}}$, hoặc $3\sqrt{\frac{\sigma^2}{n}}$,v.v... Vấn đề tiếp theo là làm thế nào để nhận biết được một mẫu ngẫu nhiên là "tốt" đến mức nào.

2. Tính đại diện của mẫu

Trong điều tra thống kê chúng ta thường yêu cầu: "mẫu được rải phải bảo đảm tính đại diện thì mới suy rộng được" vậy tính đại diện là gì? Để nghiên cứu vấn đề này ta dùng tổng thể chung đã có đầy đủ số liệu để có thể đánh giá được ngay sai số của mẫu được chọn.

Giả sử ta đã có đầy đủ số liệu một chỉ tiêu nào đó của toàn bộ tổng thể chung X_1, X_2, X_3,.....X_N, từ đây ta chọn ra các mẫu ngẫu nhiên có n đơn vị, số lượng các mẫu ngẫu nhiên có thể chọn ra được là rất lớn (C_N^n mẫu) ta hãy phân loại các mẫu này theo mức độ "tốt, xấu" theo quan điểm sai số. Ta dễ dàng nhận thấy:

- Chọn mẫu hệ thống là cách lấy mẫu theo khoảng cách đều trên cơ sở các đơn vị tổng thể được sắp xếp theo thứ tự giảm dần của chỉ tiêu điều tra sẽ là mẫu có sai số nhỏ nhất. Ta nói rằng mẫu này là một mẫu đại diện "cao" vì nó rải đều ở tất cả các mức của chỉ tiêu dùng để ước lượng giá trị trung bình.

- Mẫu có sai số lớn nhất là mẫu chọn lấy n đơn vị lớn nhất hoặc nhỏ nhất của tổng thể chung. Sai số của 2 mẫu này là rất lớn nếu ta suy rộng mẫu bằng số đơn vị của tổng thể chung (N).

Vấn đề đáng quan tâm là đưa được các đơn vị lớn vào mẫu. Vì đó là mẫu chứa đựng nhiều nhất thông tin thống kê nhận được từ n đơn vị điều tra nên là mẫu đáng được quan tâm nhất. Nhưng trở ngại lại là: mẫu này có sai số rất lớn nếu ta suy rộng bằng số đơn vị tổng thể.

Tính đại diện của mẫu gắn chặt với việc mẫu được rải đều ở tất cả các mức của chỉ tiêu dùng để ước lượng giá trị trung bình tính từ mẫu. Mức độ "đều" càng cao thì tính đại diện càng cao.

Như vậy việc chọn các đơn vị lớn vào mẫu làm cho tính đại diện của mẫu không được đảm bảo nếu ta suy rộng bằng số đơn vị tổng thể. Còn nếu chúng ta không suy rộng mẫu bằng số đơn vị tổng thể mà bằng số nhân khẩu hoặc số lao động đối với hộ gia đình; Số lao động hoặc, Tổng giá trị tài sản hoặc, Doanh thu đối với các đơn vị kinh tế; Số lao động hoặc công xuất tầu thuyền trong điều tra đánh bắt thủy sản thì việc lựa chọn các đơn vị điển hình đưa vào mẫu không mâu thuẫn với tính đại diện của mẫu, nghĩa là một mẫu điển hình vẫn là một mẫu đại diện.

Ví dụ nếu ta dùng số lao động để suy rộng thì yêu cầu của tính đại diện là mẫu cần được rải đều ở các mức của năng suất lao động (được hiểu là chỉ tiêu cần điều tra chia cho lao động) ở các mức này ta có thể chọn các đơn vị lớn nhất vào mẫu. Hay nói cách khác việc chọn các đơn vị lớn nhất vào mẫu không mâu thuẫn trực tiếp với tính đại diện của mẫu.

Việc sử dụng các chỉ tiêu khác ngoài chỉ tiêu "số đơn vị tổng thể" để suy rộng còn đem lại cho ta 2 ưu việt:

- Phương sai của ước lượng số bình quân trong trường hợp này nhỏ hơn nhiều phương sai của ước lượng số bình quân trong trường hợp suy rộng theo số đơn vị tổng thể, nên phạm vi sai số cũng sẽ nhỏ hơn nhiều lần

- Ta đưa được các đơn vị điển hình vào mẫu nên tỷ lệ chỉ tiêu cần điều tra rơi vào mẫu cao: khoảng 30 % đối với cỡ mẫu 5% và khoảng 60% đối với cỡ mẫu 15%, kết quả là số liệu điều tra chắc chắn hơn nhiều.

Chi tiết về cách lựa chọn chỉ tiêu để suy rộng và phương pháp suy rộng đã giới thiệu trong bài "Điều tra chọn mẫu trong thống kê" - Thông tin khoa học thống kê số 2 năm 2005■

Tài liệu tham khảo

1. Tống Đình Quỳ - Giáo trình xác suất thống kê – Nhà xuất bản Giáo dục Hà Nội 1999.

2. Phạm Thành Đạo - Điều tra chọn mẫu trong thống kê - Thông tin khoa học thống kê số 2 năm 2005.