Sai số phi chọn mẫu

Nguyễn Thái Hà (dịch)

Bên cạnh sai số chọn mẫu có liên quan đến quá trình chọn mẫu, một cuộc điều tra còn có thế mắc một số loại sai số khác. Các sai số đó được gọi là sai số phi chọn mẫu.

Sai số phi chọn mẫu được xác định là các sai số phát sinh trong tất cả các hoạt động điều tra (ngoài sai số chọn mẫu). Không giống như sai số chọn mẫu, các sai số phi chọn mẫu xuất hiện trong cả các cuộc điều tra chọn mẫu và tổng điều tra.

Sai số phi chọn mẫu có thể được phân ra thành hai nhóm đó là sai số ngẫu nhiên và sai số hệ thống.

- Sai số ngẫu nhiên là các sai số không thể lường trước mà do kết quả của việc ước tính. Các sai số này thường bị triệt tiêu khi sử dụng mẫu điều tra đủ lớn. Tuy nhiên, khi những sai số này có tác dụng, thì thường dẫn đến phương sai của các đặc trưng được quan tâm (tức là như sự khác nhau giữa các đơn vị tổng thể càng lớn, thì quy mô mẫu yêu cầu để đạt được mức độ tin cậy cụ thể càng phải lớn).

- Sai số hệ thống là những sai số có xu hướng tích tụ trên toàn bộ mẫu. Ví dụ như nếu có sai sót trong thiết kế phiếu điều tra, thì điều này có thể gây ra vấn đề với câu trả lời của đối tượng điều tra và đến lượt nó có thể tạo ra các lỗi trong quá trình xử lý, v.v... Các dạng loại sai số này thường dẫn đến sai lệch ở các kết quả cuối cùng.

Sai số phi chọn mẫu rất khó đo lường nếu không muốn nói là không thể đo lường được. Do các sai số ngẫu nhiên có xu hướng tự triệt tiêu, nên các sai số hệ thống là nguyên nhân chính cần phải quan tâm. Không giống như phương sai mẫu, độ trệch do các sai số hệ thống gây ra không thể giảm được bằng cách tăng quy mô mẫu.

Các đặc trưng của sai số phi chọn mẫu:

- Có thể xảy ra trong tất cả các hoạt động của quá trình điều tra (trừ chọn mẫu);

- Tồn tại trong cả điều tra chọn mẫu và tổng điều tra; và

- Khó đo lường.

Sai số phi chọn mẫu có thể xảy ra do các vấn đề về: phạm vi, trả lời, không trả lời, xử lý số liệu, ước tính và phân tích. Dưới đây giải thích cho từng loại sai số.

Sai số phạm vi

Sai số phạm vi xảy ra khi có sự bỏ sót, tính trùng hoặc đưa sai về các đơn vị vào tổng thể hoặc mẫu. Các lỗi bỏ sót thường liên quan đến thiếu phạm vi, tính trùng và đưa sai các đơn vị vào tổng thể hoặc mẫu thường gọi là quá phạm vi. Các sai số này thường là do những khiếm khuyết trong dàn mẫu điều tra như: không chính xác, không đủ, không đúng, trùng, và lỗi thời. Sai số phạm vi cũng có thể xẩy ra trong các bước điều tra tại thực địa (như cuộc điều tra được thực hiện nhưng điều tra viên đã bỏ sót một số người hoặc một số hộ gia đình).

Sai số do trả lời

Sai số do trả lời xảy ra do kết quả số liệu theo yêu cầu, cung cấp, nhận được hoặc ghi chép không đúng. Các sai số do trả lời có thể xảy ra vì phiếu điều tra có khiếm khuyết, vì người phỏng vấn, người trả lời hoặc do quá trình điều tra.

- Thiết kế phiếu kém: Vấn đề cốt yếu là các câu hỏi điều tra chọn mẫu và tổng điều tra được dùng từ ngữ cẩn thận để tránh hiểu sai lệch. Nếu các câu hỏi bị hiểu sai hoặc nhầm lẫn thì cuối cùng những câu trả lời có thể sai lệch.

- Chệch do phỏng vấn: Điều tra viên có thể ảnh hưởng đến việc người trả lời, trả lời các câu hỏi điều tra. Điều này có thể xảy ra khi điều tra viên quá thân mật, quá cách biệt hoặc gợi ý cho người trả lời như thế nào. Để ngăn chặn điều này, điều tra viên phải được tập huấn để có được sự trung lập trong suốt quá trình phỏng vấn. Họ cũng phải chú ý tới cách họ hỏi từng câu hỏi. Nếu điều tra viên thay đổi các câu hỏi được diễn đạt, thì nó có thể ảnh hưởng đến câu trả lời của những người trả lời.

- Các sai số do trả lời: Người trả lời cũng có thể đưa ra các câu trả lời sai. Một số nguyên nhân người trả lời đưa ra câu trả lời sai có thể là do hồi tưởng sai, các xu hướng mang tính phóng đại hoặc xem nhẹ và đưa ra những câu trả lời mang tính xã hội nhiều hơn.

- Sai số do quy trình điều tra: Sai số cũng xuất hiện cùng với quy trình điều tra. Việc sử dụng những người trả lời được ủy quyền (người trả lời khác mà không phải là đối tượng điều tra) hoặc thiếu kiểm soát các bước điều tra là một số nguyên nhân làm tăng khả năng sai số do trả lời.

Sai số do không trả lời

Sai số do không trả lời là kết quả của việc không nhận được các câu trả đầy đủ cho các câu hỏi của phiếu điều tra. Có hai loại sai số do không trả lời: không trả lời toàn bộ hoặc không trả lời một phần.

- Sai số do không trả lời toàn bộ: Sai số này có thể xảy ra khi không điều tra được một số đơn vị điều tra đã được chọn mẫu. Nguyên nhân của loại sai số này có thể là do đối tượng điều tra không có nhà hoặc tạm thời vắng mặt, người trả lời không có khả năng hoặc từ chối tham gia điều tra, hoặc đối tượng được chọn để điều tra rơi vào nhà bỏ trống/nhà không có người ở. Nếu có một số lượng đáng kể đối tượng không trả lời trong cuộc điều tra, thì kết quả điều tra có thể bị sai lệch vì các đặc trưng của các đối tượng không trả lời có thể khác với đối tượng đã tham gia.

- Sai số do không trả lời một phần: Loại sai số này là do thông tin thu thập được từ đối tượng điều tra không đầy đủ. Đối với một số người, có thể một vài câu hỏi là khó hiểu. Để giảm dạng sai lệch này, cần phải quan tâm trong khâu thiết kế và thử nghiệm phiếu điều tra. Chiến lược hiệu đính và tính toán có thể giúp giảm thiểu sự sai lệch này.

Sai số do xử lý

Các sai số do xử lý đôi khi xuất hiện trong quá trình chuẩn bị các file dữ liệu cuối cùng. Ví dụ, sai số có thể xảy ra trong khi đánh mã, sao chép, hiệu đính và nhập số liệu vào máy tính. Sai lệch do người đánh mã thường là do tập huấn đánh mã kém hoặc hướng dẫn không đầy đủ, thực hiện công tác đánh mã không thống nhất (do mệt mỏi, ốm đau), sai số do nhập dữ liệu sai, hoặc do máy móc trục trặc (một số sai sót là do lỗi của chương trình máy tính). Vấn đề sai sót cũng xảy ra tương tự khi sao chép số liệu. Đôi khi các lỗi là do nhận diện không chính xác trong quá trình hiệu đính. Ngay cả khi đã phát hiện ra các lỗi thì chúng có thể bị sửa không đúng do quy trình tính toán kém.

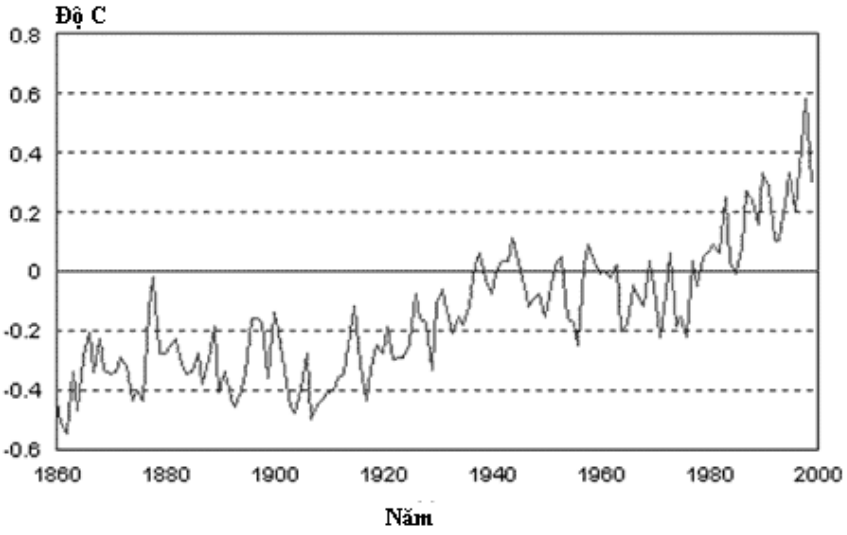
Sai số ước lượng

Cơ quan Thống kê Canada và các tổ chức thu thập số liệu khác có nhiều cố gắng trong việc thiết kế và giám sát các cuộc điều tra để

chúng càng có ít sai sót càng tốt. Nếu sử dụng phương pháp ước lượng không phù hợp thì sai sót có thể vẫn xẩy ra, cho dù mức độ hoàn hảo của việc điều tra đã thực hiện trước khi ước lượng thế nào.

Dưới đây là một ví dụ về ước lượng không phù hợp. Chúng ta biết rằng sự ấm lên toàn cầu là vấn đề có rất nhiều tranh cãi. Để đo lường chính xác hiện tượng này, cần phải có được một sự chấp nhận chung về "nhiệt độ trung bình toàn cầu". Đồ thị dưới đây miêu tả những điểm nổi bật về số liệu thay đổi nhiệt độ, cho thấy nhiệt độ trung bình toàn cầu tăng từ 0,3 đến 0,6 độ trong khoảng thời gian 140 năm.

Đồ thị. ***Thay đổi nhiệt độ toàn cầu, từ năm 1860 đến năm 2000***



Số liệu thống kê để có được tập số liệu này được thu thập từ nhiều trạm quan trắc thời tiết trên toàn thế giới. Trong trường hợp này, tổng thể là các số thống kê về thời tiết, mà từ đó điều tra mẫu có thể được tiến hành.

Một số nhà khoa học nêu vấn đề về độ chính xác của đồ thị trên vì họ cảm thấy là những số liệu ước tính từ điều tra mẫu bị sai lệch.

Một số nhà khoa học tranh luận rằng những số đo về nhiệt đó phải phản ánh tỷ số giữa diện tích rộng lớn của trái đất và diện tích rộng lớn của mặt nước. Ví dụ như nếu diện tích đất đai bằng nửa diện tích mặt nước (biển và đại dương) thì số các điểm quan trắc trên mặt nước sẽ phải gấp đôi so với số điểm trên mặt đất. Thực tế, đồ thị chỉ có một vài số thống kê được lấy từ các điểm quan trắc trên mặt nước, ngược lại, phần lớn các số thống kê được lấy từ các trạm quan trắc thời tiết trên mặt đất.

Tại sao các ước lượng từ mẫu lại có thể trệch? về bản chất nhiệt độ trên mặt đất có xu hướng cao hơn nhiệt độ của mặt nước, do hiện tượng được biết với tên gọi "hiệu ứng đảo nhiệt đô thị[1]". Nếu mẫu điều tra được lấy quá nhiều ở đất liền và khi ước lượng lại không xét đến vấn đề này (một số nhà khoa học đã khẳng định), thì các kết quả này có thể không phản ánh đúng nhiệt độ trung bình toàn cầu.

Sai số phân tích

Sai số phân tích bao gồm bất kỳ sai số nào xảy ra khi sử dụng sai các công cụ phân tích hoặc khi các kết quả sơ bộ được sử dụng thay cho kết quả cuối cùng. Các sai sót xảy ra trong quá trình xuất bản số liệu này cũng được xem là sai số phân tích. ■

Nguồn: Non-sampling error

http://www.statcan.gc.ca/edu/power-pouvoir/ch6/nse-endae/5214806-eng.htm

---

(1) "Urban heat island effect"