Những lưu ý khi thu thập, sử dụng và phân tích dữ liệu

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

06/06/2023

Nội dung

- Nguồn gốc và loại hình dữ liệu quyết định khả năng ứng dụng
- Thiết kế lấy mẫu và tính đại diện của dữ liệu:
  - Sử dụng đúng các loại dữ liệu lấy mẫu có phân tầng/phân nhóm/có quyền số (stratified, clustered random sampling, with sample weights)
- Điều tra dữ liệu sơ cấp và những quy tắc cần tuân thủ:
  - Cỡ mẫu bao nhiêu là đủ, Type 2 error, sức mạnh thống kê, quy tắc chung của lấy mẫu
  - Vấn đề lựa chọn và chệch do lấy mẫu (selection bias)
  - Quy tắc đạo đức (IRB) khi phỏng vấn điều tra sơ cấp
- Những sai lầm dễ gặp phải: Data mining dữ liệu, ngoại suy (extrapolation), và hiệu lực (validity)

Dữ liệu quyết định mô hình và khả năng ứng dụng

- Đơn vị/cấp độ đo lường
- Đặc tính của lấy mẫu: ngẫu nhiên hay thuận tiện, tính đại diện
- Cấu trúc dữ liệu, tần suất thu thập, số lượng quan sát,
- Dữ liệu có phù hợp với mục đích nghiên cứu không: Vấn đề nhận diện trong mô hình kinh tế lượng (identification problem).

Thiết kế lấy mẫu và tính đại diện

Các bộ dữ liệu điều tra chuyên nghiệp có thiết kế mẫu phức tạp, do đó yêu cầu phải hiểu về cách thức lấy mẫu để sử dụng đúng:

- Các thiết kế lấy mẫu nhiều giai đoạn (multi-stage survey design) được áp dụng để giảm thiểu số lượng mẫu phải điều tra, tăng độ chính xác, và đảm bảo điều tra được các phần tử khó điều tra (ví dụ người dân tộc thiểu số).
- Các thông số về quá trình lấy mẫu gồm có phân tầng (stratification), phân cụm (cluster), đơn vị điều tra chính (primary sampling unit - PSU), và quyền số điều tra (sampling weights).
- Cấp độ đại diện của dữ liệu: cấp độ quốc gia hay vùng/khu vực.

So sánh tổng thu nhập trong năm của hộ gia đình từ dữ liệu thô trong điều tra VHLSS 2020 và ước tính với quyền số điều tra

```
. svy: mean hhInc2020
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =         1        Number of obs    =         9,389
Number of PSUs   =     3,130        Population size  =    26,742,164
                                    Design df        =         3,129
```

with survey design

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
hhInc2020	272339.7	9412.441	253884.6	290794.9

```
. sum hhInc2020
```

raw sample mean

Variable	Obs	Mean	Std. Dev.	Min	Max
hhInc2020	9,389	254879.2	755898.4	5240	5.26e+07

Điều tra dữ liệu sơ cấp

Phục vụ cho cả nghiên cứu định tính và định lượng.

- Lấy mẫu ngẫu nhiên hay lấy mẫu thuận tiện: tùy thuộc vào mục đích nghiên cứu, tuy nhiên đại đa số các điều tra sơ cấp đều không đảm bảo quá trình lấy mẫu ngẫu nhiên.
- Chỉ áp dụng khi cần phải tính toán thống kê mẫu để đưa ra các nhận định định lượng: Số mẫu tối thiểu là bao nhiêu?

Độ chính xác $\rightleftarrows$ **Số mẫu** $\rightleftarrows$ **Chi phí điều tra**

Công thức để tính toán cỡ mẫu điều tra

- Tùy vào mục đích nghiên cứu và đặc thù dữ liệu cần thu thập có các công thức tính cỡ mẫu tối thiểu khác nhau.
- Với nghiên cứu đánh giá tác động chính sách, tính toán cỡ mẫu điều tra rất phức tạp.

Tham khảo: Phương pháp ước tính cỡ mẫu cho một nghiên cứu y học của Nguyễn Văn Tuấn.

Tính cỡ mẫu tối thiểu trong đánh giá tác động chính sách

- Sức mạnh thống kê (statistical power): khả năng phát hiện được tác động khi trên thực tế có tác động. Nếu chương trình can thiệp có tác động nhưng không phát hiện được thì có nghĩa là chúng ta mắc sai lầm loại 2. Do đó, sức mạnh thống kê chính là xác suất không mắc sai lầm loại 2.
- Nhân tố nào ảnh hưởng đến sức mạnh thống kê?
  - Tác động càng lớn càng dễ phát hiện
  - Cỡ mẫu càng lớn càng dễ phát hiện
  - Mẫu càng đồng nhất càng dễ phát hiện
  - Mức chấp nhân sai lầm

Công thức tính cỡ mẫu tối thiểu với thiết kế can thiệp lâm sàng truyền thống

Với thiết kế này, quá trình ngẫu nhiên hóa can thiệp được thực hiện ở cấp độ cá nhân:

$$N > \frac{\sigma^2}{\left(ETE \Big/ \left(\left(t_{1-\kappa} + t_{\alpha/2} \right) \sqrt{\frac{1}{p(1-p)}} \right) \right)^2}$$

- ETE là tác động can thiệp kỳ vọng (expected treatment effect)
- $t_{t-k} = 0.84$ và $t_{\alpha/2} = 1.96$
- p là tỷ lệ nhóm can thiệp trong mẫu
- σ^2 là phương sai của kết quả can thiệp

Từ công thức trên, chúng ta tính được tác động tối thiểu có thể phát hiện được (minimum detectable effect - MDE) như sau:

$$MDE > \left(t_{1-\kappa} + t_{\alpha/2}\right)\sqrt{\frac{1}{p(1-p)}}\sqrt{\frac{\sigma^2}{N}}$$

- MDE hàm ý rằng với một số mẫu và tỷ lệ nhóm can thiệp và đối chứng cho trước, tác động tối thiểu có thể phát hiện được là bao nhiêu.
- Giả dụ ngân sách cho thử nghiệm một loại vaccine là USD 100k. Ngân sách này chỉ đủ để thử nghiệm can thiệp trên 1000 người. Giả dụ chương trình can thiệp sử dụng 2 nhóm treatment và control với tỷ lệ bằng nhau. Với các thông số σ giả định trước, tác động tối thiểu có thể phát hiện được là 100 iu/ml. Nếu vaccine tạo ra kháng thể nhưng không mạnh (ví dụ 50 iu/ml) thì chương trình nghiên cứu này có thể không phát hiện được. Khi này nhóm nghiên cứu có thể phải tăng cỡ mẫu, do đó chi phí đánh giá tốn kém hơn.

Công thức tính cỡ mẫu tối thiểu hoặc MDE với thiết kế nhóm (cluster)

Áp dụng khi can thiệp xảy ra ở cấp độ nhóm (tất cả các thành viên trong nhóm đều được hưởng lợi hoặc không):

$$MDE > \left(t_{1-\kappa} + t_{\alpha/2}\right)\sqrt{\frac{1}{p(1-p)J}}\sigma\sqrt{\rho + \frac{1-\rho}{n}}$$

- J là số cluster (giả sử clusters đồng nhất)
- ρ là hệ số tương quan nội nhóm (intra-cluster correlation - ICC)
- n là số quan sát trong mỗi cluster

Một số lưu ý khi sử dụng dữ liệu trong nghiên cứu

- Data mining, lạm dụng mô hình và thuật toán để bù đắp thiếu hụt về cơ sở lý thuyết và hàm ý chính sách.
- Giới hạn ứng dụng và khả năng ngoại suy từ kết quả nghiên cứu:
  - Xuất phát từ dữ liệu
  - Xuất phát từ phương pháp nghiên cứu: Hiệu lực nội tại, hiệu lực nội tại địa phương, và hiệu lực ngoại vi
  - Xuất phát từ mô hình và những giả định đi kèm

Những lưu ý khi thu thập và sử dụng dữ liệu

- Các vấn đề về chất lượng dữ liệu: Các vấn đề về chất lượng dữ liệu có thể phát sinh khi dữ liệu không chính xác, không đầy đủ hoặc không đáng tin cậy, điều này có thể dẫn đến kết luận không chính xác hoặc sai lệch. Những vấn đề này có thể do nhiều yếu tố gây ra, chẳng hạn như lỗi khi nhập dữ liệu, định dạng dữ liệu không nhất quán hoặc dữ liệu bị thiếu.

- Thiên lệch trong dữ liệu: Thiên lêch trong dữ liệu có thể xảy ra khi dữ liệu không phản ánh chính xác quần thể hoặc hiện tượng đang được nghiên cứu. Điều này có thể dẫn đến kết luận sai lệch hoặc không chính xác và có thể do nhiều yếu tố gây ra, chẳng hạn như sai lệch lấy mẫu, sai lệch do quá trình tự lựa chọn mẫu hoặc sai số đo lường.

- Dữ liệu bị hạn chế: khi dữ liệu có sẵn để phân tích không đủ để trả lời các câu hỏi nghiên cứu hoặc đưa ra kết luận chắc chắn. Điều này có thể là do thiếu dữ liệu trong quá trình thu thập hoặc không thể tiếp cận được dữ liệu.

- Độ phức tạp của dữ liệu: Làm việc với dữ liệu phức tạp có thể là một thách thức, đặc biệt khi dữ liệu lớn hoặc có nhiều chiều. Điều này có thể gây khó khăn cho việc phân tích và giải thích dữ liệu và có thể yêu cầu các công cụ và kỹ thuật phức tạp để quản lý và xử lý dữ liệu một cách hiệu quả.

- Quyền riêng tư và bảo mật dữ liệu: Quyền riêng tư và bảo mật dữ liệu là mối quan tâm khi làm việc với dữ liệu cá nhân hoặc nhạy cảm, vì yêu cầu phải bảo vệ quyền riêng tư của cá nhân và tuân thủ các luật và quy định có liên quan. Đây có thể là một vấn đề phức tạp, đặc biệt khi làm việc với dữ liệu điều tra hoặc dữ liệu nội bộ.

- Cần hiểu và giải quyết những vấn đề tiềm ẩn liên quan đến dữ liệu để cải thiện độ chính xác và độ tin cậy của phân tích dữ liệu và đảm bảo rằng kết luận dựa trên dữ liệu chất lượng cao.