PHƯƠNG PHÁP NGHIÊN CỨU TRONG TÂM LÝ HỌC

Phạm Thị Thủy Tiên
T10/2019
Tien.phamthithuy2386@hoasen.edu.vn

Bài 6: Chọn mẫu (Sampling)

Tổng quan

1) Các vấn đề trong chọn mẫu
   - Tính khái quát, tính đại diện
   - Khoảng tin cậy, sai số biên
   - Sai số do chọn mẫu và sai số không do chọn mẫu
2) Các khái niệm liên quan đến chọn mẫu
   - Phân phối xác suất và phân phối chọn mẫu
   - Phân phối chuẩn
3) Cách xác định cỡ mẫu
4) Các phương án chọn mẫu
   - Chọn mẫu xác xuất
   - Chọn mẫu phi xác suất

Tính khái quát và tính đại diện

1) Tính khái quát (generalizability) – Mức độ mà kết quả của nghiên cứu có thể được suy ra từ mẫu cho quần thể
   - Còn gọi là độ ngoại hiệu lực
2) Tính đại diện (representation) - Mức độ mà mẫu có thể đại diện cho quần thể
   - Đặc điểm nhân khẩu học
   - Đặc điểm cá nhân
   - Đặc điểm cụ thể

Mối đe dọa đối với độ ngoại hiệu lực (threats to external validity)

Tính khái quát (Generalizability)

History threat

kết quả nghiên cứu không thể khái quát hóa cho các thời điểm khác

Setting threat (artificiality)

kết quả nghiên cứu không thể khái quát hóa cho các bối cảnh khác; mang tính giả tạo (artificiality)

Selection threat

sai lệch trong chọn mẫu

Suy luận thống kê: Từ mẫu đến quần thể

- Là tập hợp có N phần tử, là mối quan tâm của nhà nghiên cứu
- Có nhiều biến ta không quan sát
-Các thông số (parameter) như:
  μ – (population) mean
  σ –standard error
  β – regression coefficients

-Là tập hợp n phần tử được chọn ngẫu nhiên từ quần thể
-Các biến giải thích và biến kết quả quan sát được
-Các chỉ số thống kê (statistic) như:
  M - sample mean
  SD - standard deviation

Inference – Suy luận

Quần thể - Population

Mẫu – Sample

Một vài khái niệm

- **Element/unit**: các đơn vị riêng lẻ của quần thể (nhóm/ thành phố/ cá nhân/ quốc gia,…)
- **Stratum**: Nhóm có cùng một đặc tính (v.d. nam/nữ trong một trường đại học, v.v…)
- **Census**: tất cả thành viên của quần thể được thu thập thông tin (v.d khảo sát quốc gia của chính phủ)
- **Sampling frame**: danh sách của tất cả các cá thể trong một quần thể
  - có thể liên lạc được
  - nhưng có thể ko có đủ tất cả thành tố của một quần thể

Phân phối xác suất và phân phối chọn mẫu

Phân phối xác suất - Probability distribution
- Statistic (M, SD,…)

Phân phối chọn mẫu - Sampling distribution
- Parameter (μ, σ, …)

Phân phối chuẩn (normal distribution)

- Là một dạng của phân phối xác suất; rất phổ biến và quan trọng trong khoa học tự nhiên và xã hội
- Hình dạng chính xác của phân phối chuẩn được quyết định bởi giá trị trung bình và độ lệch chuẩn
- Cho phép ta tính được phân bố xác suất của các giá trị quan sát của mẫu (probability distribution)

Đặc điểm quan trọng:

- 68% các quan sát nằm trong khoảng 1 độ lệch chuẩn so với giá trị trung bình
- 95% các quan sát nằm trong khoảng 2 độ lệch chuẩn so với giá trị trung bình
- 97% các quan sát nằm trong khoảng 3 độ lệch chuẩn so với giá trị trung bình

Ví dụ: Phân phối chuẩn của IQ

IQ
Mean = 100
SD = 15

Chỉ có khoảng 2.1% dân số có IQ bằng hoặc lớn hơn 130
Khoảng 2/3 dân số có IQ từ 85 đến 115

Suy luận thống kê: Từ mẫu đến quần thể

- Khi phân phối xác suất của một biến số tuân theo quy luật phân phối chuẩn thì phân phối chọn mẫu (sampling distribution) của giá trị trung bình (và các chỉ số thống kê khác) cũng tuân theo luật phân phối chuẩn.

- Trong trường hợp phân phối xác suất của biến không là phân phối chuẩn thì, theo định lý giới hạn trung tâm (**Central limit theorem**), ta biết rằng, khi cỡ mẫu tăng, phân phối chọn mẫu cũng sẽ trở thành phân phối chuẩn.

Khoảng tin cậy (Confidence Interval – CI)

- Chúng ta biết giá trị trung bình của mẫu. Nhưng giá trị trung bình này sẽ khác nhau giữa các mẫu khác nhau mà ta rút ra từ quần thể
- Chúng ta không biết giá trị trung bình thực sự của quần thể
- Nhưng ta có thể biết độ tin cậy của giá trị trung bình của mẫu trong việc dự báo giá trị trung bình của quần thể
- Có thể tính toán khoảng giá trị mà giá trị trung bình của quần thể có thể rơi vào

=> Khoảng tin cậy – Confidence interval (CI)

Khoảng tin cậy (tt)

- Nếu ta lấy mẫu ngẫu nhiên 100 lần từ quần thể, thì 90 lần ta sẽ tìm được giá trị trung bình dao động trong khoảng giữa hai giá trị nhất định
- Nói cách khác, chúng ta dám chắc rằng 90% khả năng là giá trị trung bình thực sự của quần thể sẽ dao động trong một khoảng nhất định
- Ví dụ: Trong 90% lần lấy mẫu, ta tính được rằng tỉ lệ % cử tri bầu cho ứng cử viên A dao động trong khoảng 8% của các giá trị ước tính từ mẫu.

Sample 1
A: 60%
B: 40%

Sample 2
A: 61%
B: 39%

Sample X
A: 56%
B: 44%

Sai số do chọn mẫu (sampling error)

- Như vậy, nếu kết quả điều tra cho thấy có 60% cử tri bầu cho A, ta có tự tin rằng 90% khả năng là tỉ lệ bầu cử thực tế sẽ dao động trong khoảng +/-8%, tức là 52% đến 68%. Đây gọi là khoảng tin cậy.
- Ngược lại, trong 10% số lần lấy mẫu, tỉ lệ tìm được sẽ vượt ra ngoài 8% tỉ lệ bầu cử thực sự này
- 8% này gọi là **sai số biên** (margin of error), hay còn gọi là sai số do chọn mẫu (sampling error)
- Có thể tính toán được sai số chọn mẫu
- Cho ta đánh giá độ tin cậy và chính xác của các chỉ số thống kê của mẫu mà ta dùng để suy luận về quần thể

Cỡ mẫu (sample size)

- Quy tắc chung: cỡ mẫu càng lớn càng tốt
- Mục đích: Giảm sai số do chọn mẫu (sampling error)
- Ba mươi (N = 30) mỗi nhóm đối với nghiên cứu can thiệp/bán can thiệp
- 10% - 20% của quần thể đối với nghiên cứu mô tả
- Khi quần thể càng lớn, thì tỉ lệ % của quần thể mà ta cần lấy mẫu để đảm bảo tính đại diện càng nhỏ
- Nếu quần thể có khoảng 100 thành tố, dùng toàn bộ quần thể
- Nếu quần thể có khoảng 500 thành tố, lấy mẫu 50%
- Nếu quần thể có khoảng 1.500 thành tố, lấy mẫu 20%
- Nếu quần thể lớn hơn 5.000 thành tố, lấy mẫu 400 trở lên

Sai số không do chọn mẫu

- Là sự chênh lệch giữa mẫu và quần thể gây ra bởi các yếu tố khác, không phải do chọn mẫu
  - Thiên kiến trong chọn mẫu (sampling bias)
  - Không phản hồi (non-response)
  - Khung mẫu không hoàn thiện (incomplete sampling frame)
  - Sai sót trong thu thập dữ kiện
  - Sai sót trong quy trình nghiên cứu
  - Sai sót trong phân tích dữ kiện

=> Có thể ngẫu nhiên hoặc không ngẫu nhiên

- Khó đánh giá

Chọn mẫu xác suất (probability sampling)

- Là phương án chọn mẫu sao cho: với từng thành tố trong quần thể, xác suất được chọn là cố định và khác 0
- Cần thiết phải có khung mẫu (sampling frame) để tính toán được xác suất chọn mẫu
- Ưu điểm:
  - Giảm thiểu các mối đe dọa đối với độ ngoại hiệu lực của nghiên cứu
  - Giảm thiên kiến trong chọn mẫu (selection/ sampling biases)
  - Chọn mẫu thể hiện đúng tỉ lệ xuất hiện trong quần thể
  - Mẫu chọn được mang tính đại diện cho quần thể

Các phương án chọn mẫu xác suất

- Chọn mẫu ngẫu nhiên đơn giản
- Chọn mẫu ngẫu nhiên hệ thống
- Chọn mẫu xác suất phức tạp
  - Chọn mẫu ngẫu nhiên phân tầng (stratified random sampling)
  - Chọn mẫu chùm (multi-stage cluster sampling)

Chọn mẫu ngẫu nhiên đơn giản (Simple random sampling)

- Xác suất được chọn vào mẫu của mỗi thành tố trong khung mẫu là độc lập và bằng nhau (independent and equal probability of being sampled)
- Có thể dùng máy tính để tạo ra các dãy số ngẫu nhiên

RANDOM.ORG

Generate 100 random integers (maximum 10,000).

Each integer should have a value between 1 and 120000

Get Sequence Reset Form

Here are your random numbers:

51618	44403	20251	37787	35446
43912	28450	46403	112105	48210
70888	39548	6761	99072	19729

.....

Chọn mẫu ngẫu nhiên hệ thống (systematic sampling)

- Mục tiêu: lấy mẫu ngẫu nhiên
- Chọn ngẫu nhiên một thành tố đầu tiên trong khung mẫu
- Sau đó lấy mẫu theo quy ước
- Như vậy, mỗi thành tố có cùng xác suất được chọn, nhưng xác suất này không còn độc lập nữa!

- Có thể dẫn tới sai lầm và mẫu được chọn không ngẫu nhiên!
- Trừ khi ta có thể chắc rằng khung mẫu được sắp xếp một cách hoàn toàn ngẫu nhiên, nếu không thì chọn mẫu hệ thống không còn được coi là chọn mẫu xác suất -> non-probability sampling!

Chọn mẫu ngẫu nhiên phân tầng (stratified random sampling)

- Đầu tiên, ta phân chia quần thể thành từng nhóm với đặc tính tương đồng nhau nhau gọi là strata. Sau đó, ta lấy mẫu ngẫu nhiên cho từng nhóm stratum này. Các phần tử được chọn được tập hợp tại thành mẫu cuối cùng (final sample)

Chọn mẫu chùm

- Chọn mẫu theo giai đoạn sao cho các cá thể được chọn nằm gần nhau một cách tương đối về mặt vị trí địa lý
- Ví dụ:
  - Khu vực sinh sống
  - Các quận huyện hoặc thành phố trong một nước
  - Trường học
  - Lớp học
- Thường dùng trong các cuộc điều tra khảo sát vì tính thực tế và tính hiệu quả cao với chi phí tiết kiệm
- Thường gồm nhiều giai đoạn (multi-stage cluster sampling)

Chọn mẫu chùm (tt)

Phân tầng hay chùm?

*Sự khác nhau giữa chọn mẫu chùm và mẫu phân tầng

-> Chọn mẫu phân tầng là khi ta sử dụng tất cả các nhóm của quần thể. Trong khi đó, chọn mẫu chùm chỉ sử dụng một mẫu nhất định của các chùm trong quần thể.

Phương pháp	Mô tả	Ưu điểm	Nhược điểm
Chọn mẫu ngẫu nhiên đơn giản	Mỗi phần tử của quần thể có cùng xác suất được chọn.	Mẫu phản ánh trung thực về quần thể.	-Cần xác định được khung mẫu. - Tốn kém chi phí
Chọn mẫu ngẫu nhiên chùm	Chia quần thể thành nhiều chùm đó, sau đó tiến hành chọn chùm bằng cách lấy mẫu ngẫu nhiên.	-Không cần khung mẫu -Thích hợp trong những trường hợp mà nếu sử dụng cách lấy mẫu đơn giản sẽ khá tốn kém.	Với cùng một sai số biên nhất định thì phương pháp này cần cỡ mẫu lớn hơn so với chọn mẫu ngẫu nhiên đơn giản.
Chọn mẫu ngẫu nhiên phân tầng	Chia quần thể thành nhóm, sau đó tiến hành chọn mẫu ngẫu nhiên trong từng nhóm đó.	Mẫu sẽ có đủ các đối tượng thuộc mỗi nhóm mà bạn muốn so sánh.	-Cần xác định được khung mẫu -Có đủ thông tin rằng đối tượng nào của quần thể thuộc về nhóm nào.

Chọn mẫu phi xác suất

- Không theo nguyên tắc của chọn mẫu ngẫu nhiên
- Có ích khi ta không thể xác định được quần thể và khung mẫu
- Có 4 kỹ thuật phổ biến:
  - Chọn mẫu tiện lợi (convenience sampling)
  - Chọn mẫu lan tỏa (snowball sampling)
  - Chọn mẫu theo mục đích (purposive sampling)
  - Chọn mẫu theo quota (quota sampling)

Chọn mẫu tiện lợi

- Convenience sampling
- Là dạng đơn giản nhất của chọn mẫu phi xác suất
- Sai lầm trong chọn lựa -> Không thể khái quát hóa
- Nhưng vẫn được dùng thường xuyên
  - Khi quần thể mang tính toàn thể (universal)
  - Không thể xác định khung mẫu
  - Chí phí cơ hội
  - Khi câu hỏi nghiên cứu đi tìm quan hệ nhân quả (chú trọng internal validity)

Chọn mẫu lan tỏa

- Snowball sampling, là một loại cụ thể hơn của chọn mẫu tiện lợi
- Tuyển dụng một số ít người tham gia -> những người này giới thiệu thêm người tham gia khác -> tiếp tục như vậy đến khi đủ số lượng mẫu
- Có ích trong trường hợp quần thể khó tiếp cận, thuộc các nhóm kín, v.v…
- Có tất cả các nhược điểm như chọn mẫu tiện lợi

Chọn mẫu theo mục đích

- Purposive sample, là một tập hợp con (subset) không mang tính đại diện của một quần thể, và được hình thành để phục vụ một mục tiêu nghiên cứu cụ thể
- Các thành tố được chọn lựa sẽ có các đặc trưng điển hình mà nhà nghiên cứu quan tâm
- Thường dùng trong nghiên cứu định tính
- Có các nhược điểm của chọn mẫu phi xác suất khác

Chọn mẫu theo quota

- Quota sampling
- Tương tự với chọn mẫu phân tầng
- Khác nhau:
  - Chọn mẫu phân tầng: nhóm/ strata -> chọn mẫu ngẫu nhiên từ các strata
  - Chọn mẫu theo quota: phân nhóm -> chọn mẫu tiện lợi từ các nhóm
- Có các nhược điểm của chọn mẫu phi xác suất khác
- Nếu khái quát hóa thì phải hết sức cẩn thận!

Thực hành

Bài tập tổng hợp: câu hỏi nghiên cứu, thiết kế nghiên cứu, các loại hiệu lực và các mối đe doạ đối với độ hiệu lực của nghiên cứu.

SV mang theo file mềm tờ bài tập tổng hợp để làm tại lớp.