Estimation and Inference

October 19, 2023

```
[ ]: from IPython.display import Image
```

1 Sampling Methods

Tiết kiệm tiền bạc và thời gian là hai lý do cơ bản khiến phân tích viên cần thực hiện lấy mẫu để trả lời các câu hỏi về một tổng thể. Cụ thể hơn:

1. Trong một số trường hợp, chúng ta không thể biểu diễn hết tất cả phần tử của một tổng thể
2. Trong một số trường hợp khác, việc biểu diễn tất cả phần tử của một tổng thể không đạt được hiệu quả kinh tế

Có hai loại phương pháp lấy mẫu:

1. **Probability Sampling**: Lấy mẫu xác suất là một phương pháp lấy mẫu với nguyên tắc: xác suất xuất hiện trong mẫu là như nhau đối với tất cả các phần tử của tổng thể. Do đó, loại phương pháp này có thể tạo ra mẫu có tính đại diện cho tổng thể
2. **Non-Probability Sampling:** Lấy mẫu phi xác suất là phương pháp lấy mẫu phụ thuộc vào các yếu tố khác ngoài việc xem xét xác suất, chẳng hạn như nhận định của phân tích viên hay sự thuận tiện khi tiếp cận dữ liệu. Do đó, tồn tại nguy cơ việc lấy mẫu phi xác suất tạo ra mẫu không mang tính đại diện cho tổng thể

Nói chung, trong điều kiện các yếu tố khác tương tự, lấy mẫu xác suất có thể mang lại độ chính xác và độ tin cậy cao hơn so với lấy mẫu phi xác suất

Khi phân tích viên lấy mẫu, họ phải xây dựng **kế hoạch lấy mẫu** (`sampling plan`). Kế hoạch lấy mẫu là tập hợp các quy tắc được sử dụng để chọn mẫu

Sai số lấy mẫu (`sampling error`) đề cập đến sự chênh lệch giữa số liệu thống kê (trung bình, phương sai…) giữa mẫu và tổng thể

Phân phối mẫu (`random distribution`) của một thống kê là phân phối của tất cả các giá trị phân biệt có thể có mà số liệu thống kê có thể giả định khi được tính toán từ các mẫu có cùng kích thước được rút ngẫu nhiên từ cùng một tổng thể

1.1 Simple Random Sampling

Mẫu ngẫu nhiên đơn giản (`simple random sample`) là một tập hợp con của một tổng thể được tạo ra theo cách mà mỗi phần tử của tổng thể có xác suất được chọn vào tập hợp con đó như nhau

Mẫu ngẫu nhiên đơn giản là loại mẫu cơ bản mà từ đó chúng ta có thể đưa ra kết luận hợp lý về mặt thống kê về tổng thể

1.2 Systematic Sampling

Lấy mẫu có hệ thống là phương pháp lấy mẫu mà ở đó phân tích viên sẽ đưa một phần tử từ tổng thể vào mẫu sau mỗi k phần tử cho đến khi đạt được kích thước mẫu mong muốn

Phương pháp này thường được sử dụng khi chúng ta không thể đánh dấu và/hoặc xác định rõ ràng được các phần tử từ tổng thể

Mẫu được sinh ra từ phương pháp này tiệm cận mức ngẫu nhiên

1.3 Stratified Random Sampling

Lấy mẫu phân tổ là một phương pháp lấy mẫu trong đó:

- Tổng thể được chia thành các nhóm, gọi là các tổ (`strata`). Các tổ này được hình thành từ một hoặc nhiều điều kiện
- Các mẫu ngẫu nhiên đơn giản (đôi khi là mẫu ngẫu nhiên hệ thống) được thành lập từ từng tổ (`stratum`) với kích thước của từng mẫu tỷ lệ thuận với kích thước của tổ tương ứng (có thể không tuân theo tỷ lệ trong một số trường hợp). Hợp của các mẫu này thành lập nên một mẫu lấy theo tổ

Lấy mẫu ngẫu nhiên theo tổ đảm bảo rằng mọi vùng được quan tâm trong tổng thể đều được thể hiện trong mẫu. Bên cạnh đó, ước lượng số liệu thống kê từ lấy mẫu thống kê có độ chính xác cao hơn so với lấy mẫu ngẫu nhiên đơn giản

1.4 Cluster Sampling

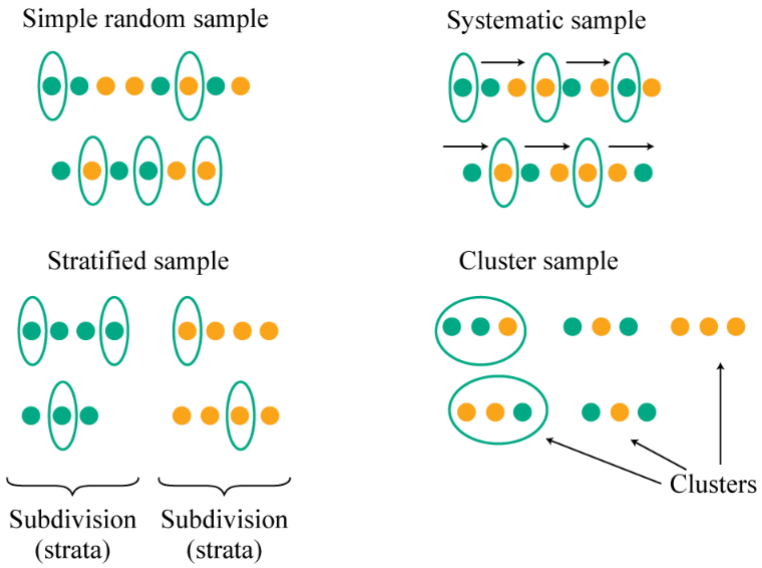
Lấy mẫu theo cụm là moojt phương pháp lẫy mẫu, trong đó:

- Tổng thể được chia thành các cụm (cluster), và mỗi cụm về cơ bản là một đại diện của toàn bộ tổng thể
- Một hoặc một số cụm được chọn thông qua lấy mẫu ngẫu nhiên đơn giản, tạo thành mẫu theo cụm
  - *Nếu tất cả phần tử trong các cụm đều được chọn, kế hoạch lấy mẫu này được gọi là lấy mẫu theo cụm một giai đoạn*
  - *Nếu chỉ có một mẫu con được chọn ngẫu nhiên từ mỗi cụm, kế hoạch lấy mẫu này được gọi là lấy mẫu theo cụm hai giai đoạn*

Ưu điểm của lấy mẫu theo cụm chính là tiết kiệm thời gian và chi phí, vì vậy phương pháp này thường được sử dụng trong khảo sát với quy mô lớn; bất chấp mẫu theo cụm có độ chính xác thấp hơn so với mẫu có cùng kích thước được hình thành từ các phương pháp khác bởi vì các cụm được chọn có thể không đại diện tốt cho toàn bộ tổng thể

[ ]:
```
# So sánh các phương pháp lấy mẫu xác suất
Image(filename = "Pictures/01.png")
```

[ ]:



1.5 Non-Probability Sampling

Các phương pháp lấy mẫu phi xác suất không dựa trên một quy trình chọn mẫu cố định, mà phụ thuộc vào khả năng chọn mẫu của phân tích viên. Có hai phương pháp lấy mẫu phi xác suất chúng ta thường gặp:

- **Lấy mẫu thuận tiện** (`convenience sampling`) là phương pháp lấy mẫu trong đó các phần tử được lựa chọn vào mẫu hay không phụ thuộc vào khả năng tiếp cận của phân tích viên đối với bản thân các phần tử đó
- **Lấy mẫu phán đoán** (`judgemental sampling`) là phương pháp lấy mẫu trong đó các phần tử được chọn lọc thủ công dựa trên kiến thức và phán đoán chuyên môn của phân tích viên

Cả hai phương pháp này đều có tỷ lệ lấy mẫu chính xác nhất định bởi vì các mẫu sinh ra có thể bị sai lệch bởi phần tử ngoại mẫu, sự thiên vị của phân tích viên... và do đó không nhất thiết đại diện cho toàn bộ tổng thể. Tuy nhiên, chúng vẫn được sử dụng trong một số trường hợp bởi những lợi ích các phương pháp này mang lại, nhất là trong trường hợp thời gian và kinh phí bị hạn chế

2 Central Limit Theorem and Inference

Nội dung định lý: Giả định một tổng thể được mô tả bởi bất kỳ phân phối xác suất nào, có trung bình μ và phương sai σ^2. Phân phối mẫu của trung bình mẫu $\overline{X}$ được xác định từ các mẫu ngẫu nhiên kích thước n từ tổng thể ban đầu sẽ có *phân phối tiệm cận chuẩn* với trung bình và phương sai phân phối mẫu lần lượt là μ và σ^2/n trong trường hợp n đủ lớn

3 Bootstrapping and Empirical Sampling Distributions

3.1 Bootstrap

Lấy mẫu lại (`resampling`) là phương pháp được sử dụng để suy luận thống kê đối với số liệu thống kê tổng thể thông qua việc lặp lại việc tạo ra mẫu từ mẫu ban đầu. **Bootstrap**, một trong những phương pháp lấy mẫu lại phổ biến nhất, sử dụng mô phỏng máy tính để đưa ra suy luận thống kê mà không sử dụng phân phối thống kê như z hay t

Trong bootstrap, phân tích viên liên tục tạo mẫu bootstrap có cùng kích thước từ mẫu gốc. Ý tưởng của bootstrap là lấy mẫu có hoàn lại, nghĩa là một phần tử được lấy ra từ mẫu ban đầu, ghi nhận, sau đó trả lại mẫu gốc và quy trình lấy mẫu tiếp tục. Với đặc điểm này, một phần tử nào đó trong mẫu ban đầu có thể xuất hiện nhiều lần, trong khi một số phần tử khác hoàn toàn không xuất hiện trong mẫu bootstrap (bất chấp kích thước của các mẫu là bằng nhau)

Phân tích viên có thể xây dựng phân phối mẫu với các mẫu bootstrap, và phân phối này tiệm cận với phân phối mẫu thực. Phân tích viên ước tính sai số chuẩn của trung bình mẫu bằng công thức dưới đây:

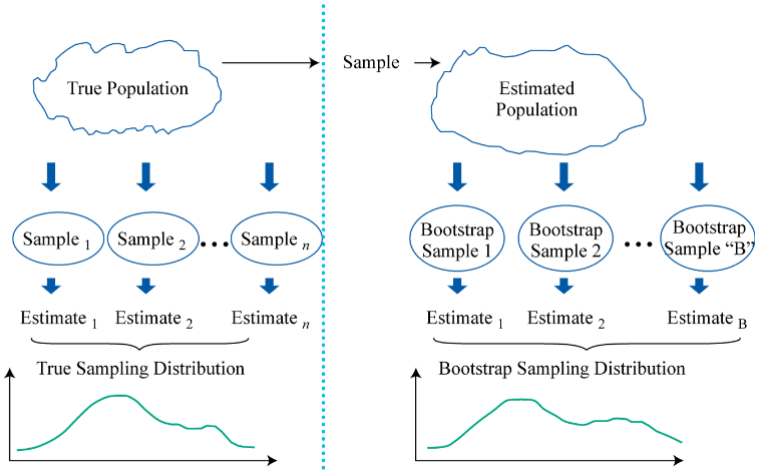
$$s_{\overline{X}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \overline{\theta}\right)^2}$$

trong đó, $s_{\overline{X}}$ là ước lượng sai số chuẩn trung bình mẫu; B là số lượng mẫu bootstrap, $\hat{\theta}_b$ là giá trị trung bình của mẫu thứ b; $\overline{\theta}$ là giá trị trung bình của hợp tất cả các mẫu bootstrap

Bootstrap là công cụ mạnh và được sử dụng rộng rãi nhất trong suy luận thống kê. Phương pháp này có thể được sử dụng để ước lượng sai số chuẩn của trung bình mẫu, sai số chuẩn và/hoặc khoảng tin cậy cho các tham số tổng thể khác, chẳng hạn như *trung vị*.

[ ]:
```
# Bootstrap
Image(filename = "Pictures/02.png")
```

[ ]:



3.2 Jackknife

Jackknife là một kỹ thuật lấy mẫu lại khác và cũng được sử dụng để suy luận thống kê cho các số liệu thống kê tổng thể. Các mẫu Jackknife được tạo ra bằng cách loại bỏ một phần tử duy nhất từ mẫu ban đầu. Vì vậy, Jackknife có một số đặc điểm:

- Số lần lấy mẫu khi thực hiện Jackknife đúng bằng kích thước của mẫu ban đầu
- Kết quả của Jackknife gần như tương tự đối với mỗi lần thực hiện lại

Jackknife được sử dụng để giảm độ lệch của phương pháp ước lượng cũng như tìm sai số chuẩn và khoảng tin cậy của các ước lượng