Sampling and Estimation

September 18, 2023

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

1 Point Estimates of the Population Mean

1.1 Point Estimators

Một khái niệm quan trọng được giới thiệu ở đây là thống kê mẫu là các biến ngẫu nhiên, bởi vì chúng được xem như các hàm liên quan đến các kết cục ngẫu nhiên.

- Các hàm được sử dụng để tính giá trị trung bình của mẫu và tất cả các thống kê mẫu là các ví dụ về hàm ước lượng (`estimator`). Một hàm ước lượng sinh ra một phân phối mẫu
- Giá trị cụ thể được tính toán từ mẫu bằng các hàm ước lượng được gọi là ước lượng (`estimate`). Bản thân một ước lượng là một hằng liên quan đến một mẫu nhất định và do đó không có phân phối mẫu

Ví dụ: Giá trị trung bình mẫu với một mẫu cụ thể, được sử dụng làm ước lượng cho trung bình tổng thể được gọi là **ước lượng điểm** của trung bình tổng thể. Rõ ràng, hàm xác định giá trị trung bình mẫu có thể trả về các kết quả khác nhau với mỗi lần lấy mẫu lại khi các mẫu được xét đến là khác nhau dù cùng được rút ra từ một tổng thể.

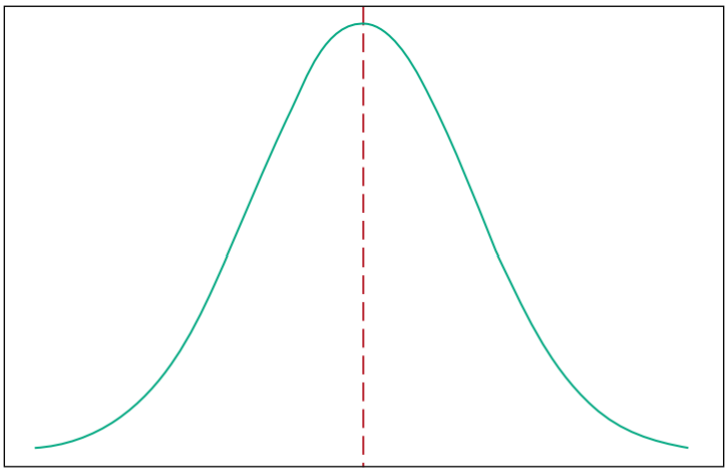
Khi ứng dụng, chúng ta có thể có nhiều lựa chọn hàm ước lượng khác nhau. việc lựa chọn hàm ước lượng phù hợp phụ thuộc vào việc hàm nào thỏa mãn nhiều tính chất thống kê mong muốn hơn. Ba tính chất đó là: tính không chệch (`unbiasedness`), tĩnh hiệu quả (`efficiency`) và tính vững (`consistency`)

1.2 Unbiasedness

Định nghĩa tính không chệch: Một hàm ước lượng được coi là không chệch khi giá trị kỳ vọng ($\overline{X}$) bằng với giá trị thực của tham số được ước tính (μ)

[ ]:
```
# Unbiasedness
Image(filename = "Pictures/01.png")
```

[ ]:



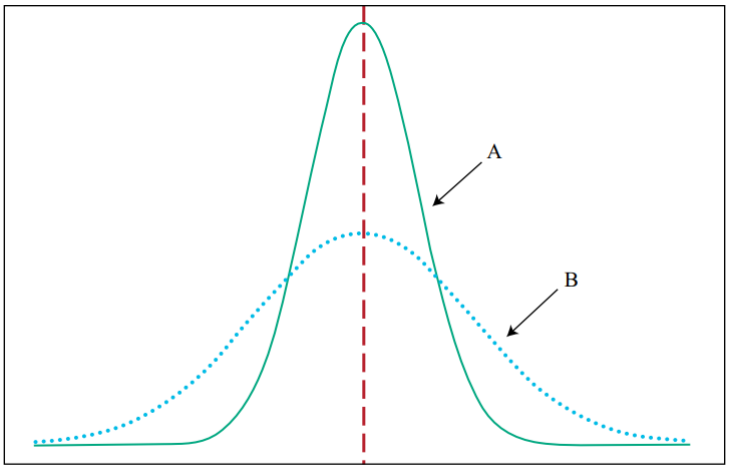
$\bar{X} = \mu$

1.3 Efficency

Định nghĩa tính hiệu quả: Một hàm ước lượng không chệch được coi là hiệu quả nếu không có bất kỳ hàm ước lượng không chệch nào khác của cùng tham số đó có phân phối mẫu với phương sai nhỏ hơn nó

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

[ ]:



$\bar{X}_A = \bar{X}_B = \mu$

1.4 Consistency

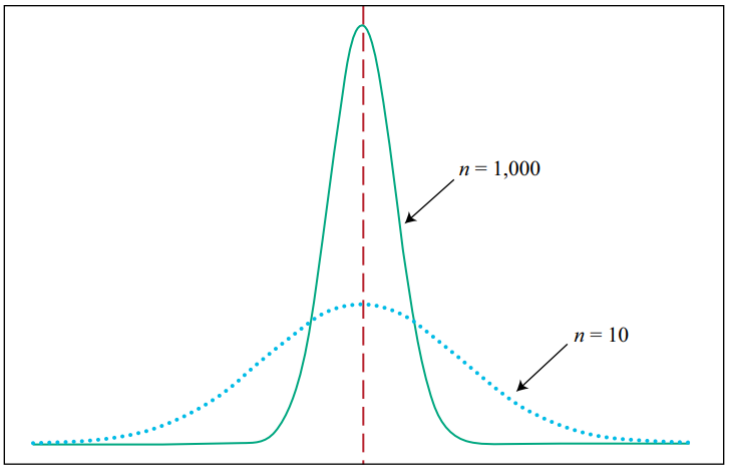
Định nghĩa về tính vững: Một hàm ước lượng được cho là vững nếu xác suất giá trị kỳ vọng gần với giá trị thực của tham số được ước lượng tăng lên khi cỡ mẫu tăng

Về mặt kỹ thuật, chúng ta có thể định nghĩa rằng một hàm ước lượng được cho là vững khi phân phối lấy mẫu của nó tập trung vào giá trị thực của tham số được ước lượng khi cỡ mẫu tăng lên. Điều này có nghĩa là:

- Đối với một hàm ước lượng không chệch mang tính vững, sai số chuẩn $\sigma/\sqrt{n}$ tiến về 0, có nghĩa là phân phối mẫu tập trung ở lân cận giá trị trung bình tổng thể

[ ]:
```
# Consistency
Image(filename = "Pictures/03.png")
```

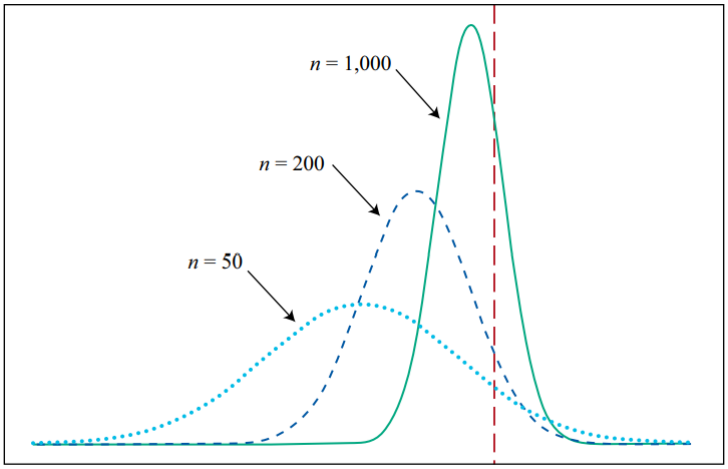
[ ]:



- Đối với một hàm ước lượng chệch nhưng mang tính vững, độ chính xác của ước lượng này sẽ được cải thiện khi cỡ mẫu tăng lên

[ ]:
```
# Consistency
Image(filename = "Pictures/04.png")
```

[ ]:



Nói tóm tại, với một hàm ước lượng mang tính vững, việc cố gắng tăng kích thước mẫu giúp ước lượng trở nên chính xác hơn. Điều này rất quan trọng trong thế giới dữ liệu lớn, khi nguồn dữ liệu ngày càng được mở rộng, cũng như tính sai lệch của ước lượng (được đo bằng độ chệch ước lượng phương sai) trở nên không đáng kể khi kích thước mẫu là rất lớn

2 Confidence Intervals for the Population Mean and Sample Size Selection

2.1 Confidence Intervals

Định nghĩa khoảng tin cậy: Khoảng tin cậy $1-\alpha$ là một phạm vi mà tại đó người ta khẳng định rằng xác suất giá trị thực của tham số được ước tính nằm trong khoảng giá trị này là $1-\alpha$

Khoảng tin cậy $100(1-\alpha)\%$ được xây dựng như sau:

Ước lượng điểm $\pm$ Hệ số tin cậy $\times$ Độ lệch chuẩn

Ba cách xây dựng khoảng tin cậy thường được sử dụng trong thực tế bao gồm:

1. Khoảng tin cậy cho trung bình tổng thể - Tổng thể có phân phối chuẩn và đã biết trước phương sai

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

2. Khoảng tin cậy cho trung bình tổng thể dựa trên phân phối t

$$\overline{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

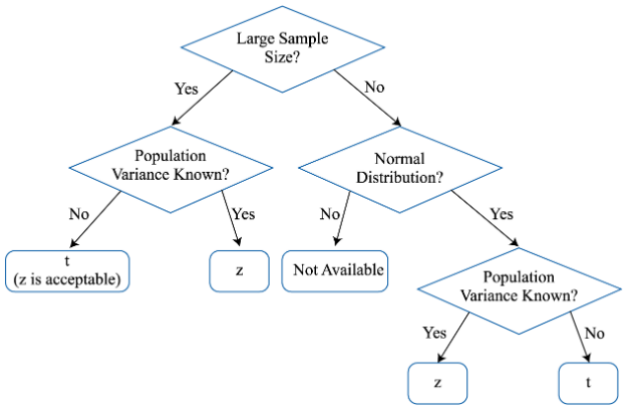
3. Khoảng tin cậy cho trung bình tổng thể - z-alternative

$$\overline{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Việc lựa chọn cách thức xác định khoảng tin cậy dựa trên một số điều kiện cụ thể, được trình bày trong sơ đồ dưới đây:

[ ]:
```
# Lựa chọn cách thức xác định trung bình tổng thể
Image(filename = "Pictures/05.png")
```

[ ]:



2.2 Sample Size Selection

Độ rộng của khoảng tin cậy bị ảnh hưởng bởi hai yếu tố:

1. **Hệ số tin cậy**

Hệ số này được xác định thông qua hai yếu tố khác: hệ số t hoặc z và mức ý nghĩa α

2. **Kích thước mẫu**

Chúng ta biết rằng, sai số chuẩn của trung bình mẫu tỷ lệ nghịch với căn bậc hai của kích thước mẫu ($s_{\overline{X}} = s/\sqrt{n}$). Điều này có nghĩa rằng sai số chuẩn giảm, từ đó độ rộng khoảng tin cậy cũng giảm khi cỡ mẫu tăng. Do đó, chúng ta kỳ vọng rằng việc tăng kích thước mẫu mang đến khả năng ước tính chính xác tốt hơn

Trên lý thuyết, cỡ mẫu lớn hơn rất tốt. Tuy nhiên, trong thực tế, có hai vấn đề cần được xem xét nếu muốn thực hiện tăng cỡ mẫu: **(1)** độ chính xác; **(2)** rủi ro lấy mẫu từ nhiều tổng thể; **(3)** sự đánh đổi giữa chi phí và độ chính xác với các cỡ mẫu khác nhau

3 Sampling-Related Biases

3.1 Data Snooping Bias

Thiên kiến khai phá dữ liệu (`data snooping bias`) mô tả việc lặp đi lặp lại quá trình khai thác dữ liệu từ cùng một tập dữ liệu để tìm kiếm các mẫu dữ liệu có liên quan đến kết quả mong muốn.

Nếu dữ liệu được khai phá đủ kỹ lưỡng, sẽ luôn có các mẫu mà các dữ liệu cụ thể phù hợp với mẫu hình ngay cả khi đó vốn là một sự kiện ngẫu nhiên. Hiện tượng này gọi là *phù hợp quá mức* (`overfitting`). Thiên kiến khai phá dữ liệu có thể dẫn đến việc ước lượng bị chệch.

Phương pháp phát hiện thiên kiến khai phá dữ liệu:

1. Chia dữ liệu thành 3 tập riêng biệt:
   - Tập đào tạo: xây dựng mô hình
   - Tập xác thực: đánh giá sự phù hợp và điều chỉnh thông số kỹ thuật của mô hình

- Tập thử nghiệm: đánh giá sự phù hợp của mô hình cuối cùng, cung cấp một thử nghiệm ngoài mẫu (`out-of-sample test`)

2. Kết luận thiên kiến khai phá dữ liệu tồn tại nếu thử nghiệm ngoài mẫu trả về ý nghĩa thống kê không đáng kể

Thiên kiến khai phá dữ liệu có thể tồn tại nếu:

1. "Đào" quá nhiều: thử nghiệm quá nhiều biến và/hoặc mẫu dữ liệu nhằm thu được kết quả mong muốn
2. Không có câu chuyện: không có lý do kinh tế rõ ràng ủng hộ tác động của biến số

3.2 Sample Selection Bias

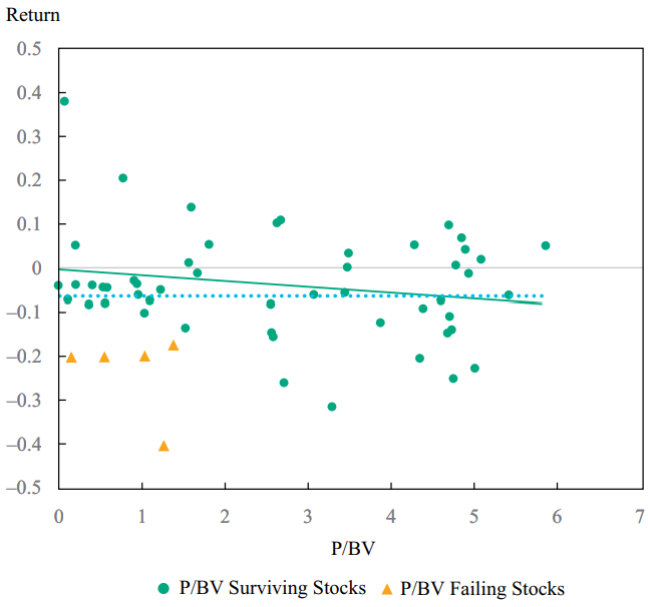
Thiên kiến lựa chọn mẫu (`sample selection bias`) mô tả việc lựa chọn mẫu không đại diện cho tổng thể, từ đó dẫn đến các kết quả phân tích bị sai lệch và không thể đại diện cho tổng thể

Một trong những thiên kiến lựa chọn mẫu thường gặp chính là *thiên vị kẻ sống sót* (`survivorship bias`), khi phân tích viên chỉ chú ý đến những thực thể còn tồn tại sau một quá trình chọn lọc nào đó mà bỏ qua những thực thể đã bị loại (ví dụ như các mẫu thường bỏ qua doanh nghiệp đã bị hủy niêm yết)

Một dạng khác của thiên kiến lựa chọn mẫu là *thiên vị lấp đầy* (`backfill bias`). Điều này xảy ra khi một đơn vị chéo mới được thêm vào mẫu, và dữ liệu lịch sử của nó được "backfilled"

[ ]:
```
# Thiên vị kẻ sống sót
Image(filename = "Pictures/06.png")
```

[ ]:



3.3 Look-Ahead Bias

Thiên kiến tiên liệu (`look-ahead bias`) xảy ra nếu phân tích viên sử dụng thông tin không có sẵn trong quá trình nghiên cứu. Thiên kiến tiên liệu thường xuất hiện trong các kịch bản có thể xảy ra

3.4 Time-Period Bias

Thiên kiến khoảng thời gian (`time-period bias`) mô tả việc lựa chọn các quan sát chỉ thuộc một khoảng thời gian nhất định. Điều này có thể khiến các kết quả phân tích bị chệch và không thê