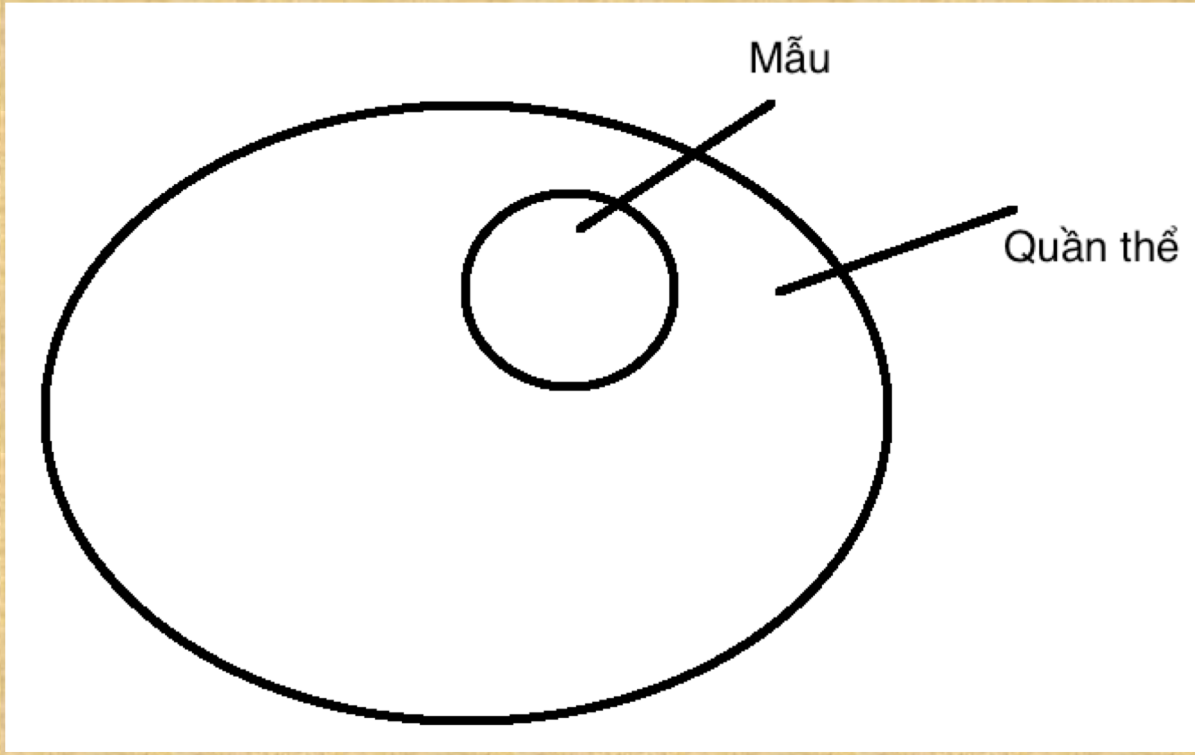
Giới thiệu về thống kê và Khoảng tin cậy

PGS.TS. Lê Sỹ Vinh

Quần thể và mẫu



- Quần thể (population): Tập hợp tất cả các đối tượng mà chúng ta muốn tiến hành nghiên cứu.
- Mẫu (sample): Một tập hợp con các đối tượng trong quần thể mà chúng ta tiến hành thu thập dữ liệu.

Ví dụ

- Khi tiến hành nghiên cứu số lượng bia trung bình 1 người đàn ông VN uống 1 năm, quần thể chúng ta quan tâm nghiên cứu là toàn bộ đàn ông VN.

- Để tiến hành nghiên cứu số lượng bia trung bình 1 người đàn ông VN uống 1 năm, người ta có thể chọn ngẫu nhiên một mẫu gồm 1000 người đàn ông ở các tỉnh, các độ tuổi khác nhau.

Lưu ý: Số phần tử trong tập mẫu gọi là kích thước mẫu.

Mẫu ngẫu nhiên/mẫu bị thiên lệch

- Để tập mẫu phản ánh được tổng thể, tập mẫu cần được lấy ngẫu nhiên từ quần thể.
- Mẫu bị thiên lệch (biased sample) sẽ làm cho kết quả thống kê thu được từ mẫu không phản ánh được bản chất của quần thể thể.

Ví dụ: Để thống kế số lượng bia trung bình 1 người đàn ông VN uống, người ta tiến hành lấy mẫu như sau:

- Chọn ngẫu nhiên 1000 người đàn ông uống bia tại quán bia Lan Chín, Cầu Giấy vào 4 ngày thứ bảy của tháng 6.
- Chọn ngẫu nhiễn 1000 người đàn ông uống bia ở 20 quán bia khác nhau tại Hà Nội vào các ngày bất kì từ tháng 6 đến tháng 10.
- Chọn ngẫu nhiễn 1000 người đàn ông uống bia ở 20 quán bia khác nhau tại 10 tỉnh/thành phố vào các ngày bất kì từ tháng 1 đến tháng 12.

Mẫu ngẫu nhiên/mẫu bị thiên lệch

Để điều tra mức lương ra trường trung bình của sinh viên Trường ĐHCN. Tiến hành lấy mẫu 100 sinh viên như sau:

- Chọn 50 sinh viên khoa cơ, 50 sinh viên khoa CNTT.
- Chọn ngẫu nhiên 100 sinh viên ra trường đang làm việc tại Hà Nội.
- Chọn ngẫu nhiên 100 sinh viên ra trường đang làm việc tại 5 công ty tại Hà Nội.
- Chọn ngẫu nhiên 100 sinh viên mới ra trường, trong đó có 70 sinh viên có điểm học trung bình >2.75.

Phân bố của mẫu và định lí giới hạn trung tâm
Central limit theory

Giả sử S={X1, X2,…,Xn} là một mẫu, hay một dãy các biến ngẫu nhiên độc lập có cùng phân bố với kì vọng μ và phương sai σ^2.
Trung bình cộng

$$\bar{x} = (X1 + X2 + \cdots . + Xn) / n$$

$\bar{x}$

Theo luật số lớn $\bar{x}$ sẽ tiến gần đến μ theo xác suất. $\bar{x}$ có phân bố chuẩn với kì vọng μ và phương sai σ^2/n.

Ước lượng kì vọng và phương sai quần thể từ tập mẫu với kích thước mẫu >=30

Giả sử S={X1, X2,…,Xn} là một mẫu, kì vọng μ và phương sai σ^2 của quần thể có thể được ước lượng như sau:

- Ước lượng kì vọng của quần thể

  $\mu \cong \overline{x} = (X1+X2+\ldots+Xn)/n$

- Ước lượng phương sai của quần thể với điều kiện (n>=30)

$$\sigma^2 \cong s^2 = \frac{\sum(xi - \overline{x})^2}{n-1}$$

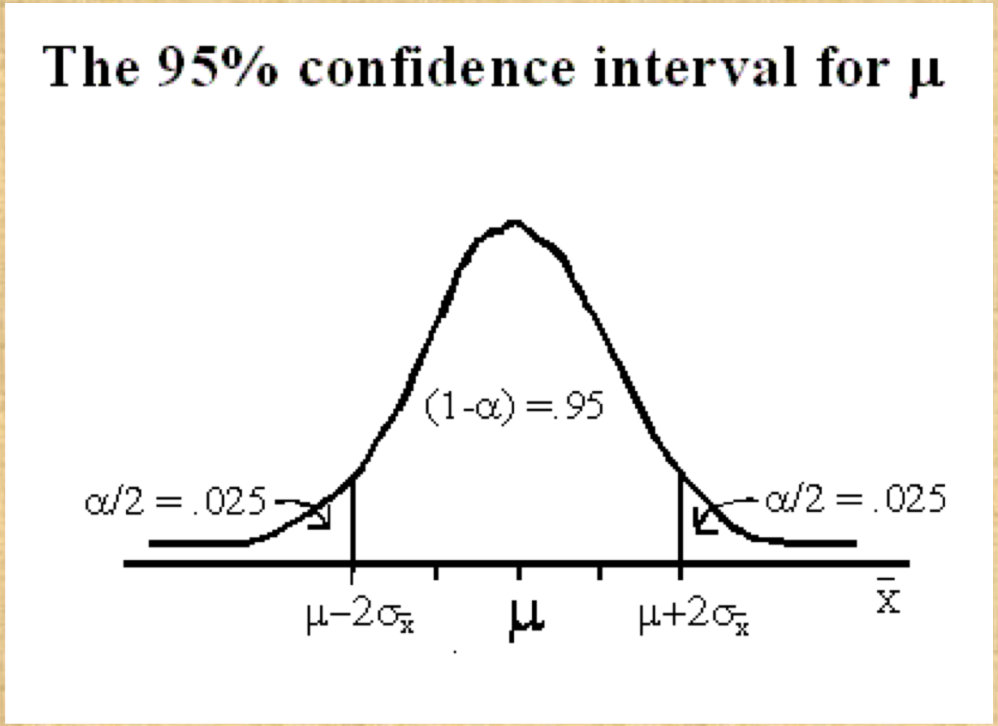
Khoảng tin cậy

Giả sử S={X1, X2,…,Xn} là một mẫu (n>=30), kì vọng μ của quần thể

$$\mu \cong (X1+X2+\ldots+Xn)/n$$

Câu hỏi: Ước lượng khoảng tin cây cho kì vọng μ?

Hay ta muốn tìm 1 đoạn [a,b] để μ thuộc đoạn trên với xác suất β%.



Khoảng tin cậy

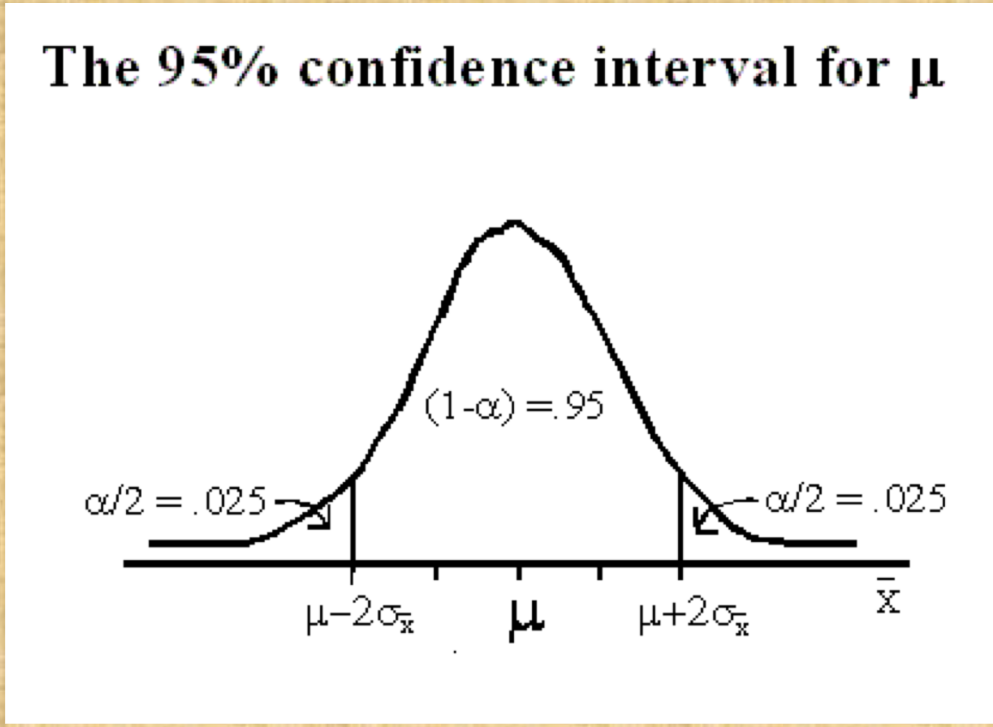
Đoạn [a, b] sẽ có dạng

$$[\ \bar{x} - u_\beta \sigma_{\bar{x}}\ ,\ \bar{x} + u_\beta \sigma_{\bar{x}}\ ]$$

Trong đó u_β là số lần độ lệch chuẩn; $\sigma^2_{\bar{x}} = \sigma^2/n$ là phương sai của $\bar{x}$.

Ví dụ

- β = 90%, thì u_β= 1.64
- β = 95%, thì u_β= 1.96
- β = 98%, thì u_β= 2,33
- β = 99%, thì u_β= 2,58



Ví dụ 1

Chiều cao trung bình của một tập 30 sinh viên ĐHCN có phân bố như sau: 10 sinh viên 162cm; 10 sinh viên 166cm; 10 sinh viên 170cm.

1. Tính khoảng tin cậy chiều cao trung bình sinh viên ĐHCN với độ tin cậy 90%.
2. Tính khoảng tin cậy chiều cao trung bình sinh viên ĐHCN với độ tin cậy 95%.
3. Tính khoảng tin cậy chiều cao trung bình sinh viên ĐHCN với độ tin cậy 99%.
4. Tính khoảng tin cậy chiều cao trung bình sinh viên ĐHCN với độ tin cậy 80%.

Ví dụ 2

Cân nặng trung bình của một tập mẫu gồm X sinh viên ĐHCN là 60 kg với độ lệch chuẩn là 6kg. Hãy ước lượng khoảng tin cậy cân nặng của sinh viên ĐHCN với độ tin cậy 95%:

1. Khi kích thước mẫu là 30
2. Khi kích thước mẫu là 50
3. Khi kích thước mẫu là 100
4. Khi kích thước mẫu là 200

Ví dụ 2A

Cân nặng trung bình của 64 sinh viên ĐHCN là 55 kg với độ lệch chuẩn là 5kg. Để nhà Trường thiết kế thang máy, bạn hãy ước lượng khoảng tin cậy tổng cân nặng của 12 sinh viên ĐHCN với độ tin cậy 99%:

Xác định kích thước mẫu

- Với một độ tin cậy β% cho trước, khoảng tin cậy [a,b] phụ thuộc vào kích thước mẫu. Kích thước mẫu càng lớn thì khoảng tin cậy càng hẹp và ngược lại.
- Câu hỏi: Giả sử muốn ước lượng μ với sai số không quá ε cho trước với độ tin cậy β, thì chúng ta phải tiến hành lấy tối thiểu bao nhiêu mẫu?

$$|\bar{x} - \mu| \leq u_\beta \frac{\sigma}{\sqrt{n}}$$

$$hay$$

$$u_\beta \frac{\sigma}{\sqrt{n}} \leq \varepsilon$$

$$\Rightarrow n \geq \left(\frac{\sigma u_\beta}{\varepsilon} \right)^2$$

Ví dụ 3

Biết rằng độ lệch chuẩn về chiều cao của người lớn Việt Nam là 10cm.

1. Tính số sinh viên phải lấy mẫu để tính chiều cao trung bình sinh viên ĐHCN với sai số không quá 1cm với độ tin cậy 90%.
2. Tính số sinh viên phải lấy mẫu để tính chiều cao trung bình sinh viên ĐHCN với sai số không quá 2cm với độ tin cậy 99%.
3. Tính số sinh viên phải lấy mẫu để tính chiều cao trung bình sinh viên ĐHCN với sai số không quá 1cm với độ tin cậy 95%.
4. Tính số sinh viên phải lấy mẫu để tính chiều cao trung bình sinh viên ĐHCN với sai số không quá 1cm với độ tin cậy 99%.

Ví dụ 4

Một trường đại học tiến hành một nghiên cứu xem trung bình một sinh viên tiêu hết bao nhiêu tiền điện thoại một tháng. Một tập mẫu ngẫu nhiên các sinh viên được chọn và kết quả như sau: 10 bạn 50k; 15 bạn 75k; 10 bạn 100k; 5 bạn 200k.

1. Bạn hãy tính khoảng tin cậy 95% cho số tiền gọi điện thoại trung bình hàng tháng của 1 sinh viên.

2. Bạn hãy tính khoảng tin cậy 99% cho số tiền gọi điện thoại trung bình hàng tháng của 1 sinh viên.

Ước lượng kì vọng và phương sai quần thể từ tập mẫu với kích thước mẫu <30

Trung bình mẫu:

$\bar{x} = (x1+x2+\ldots+xn)/n$

Phương sai mẫu:

$$s^2 = \frac{\sum(xi - \bar{x})^2}{n-1}$$

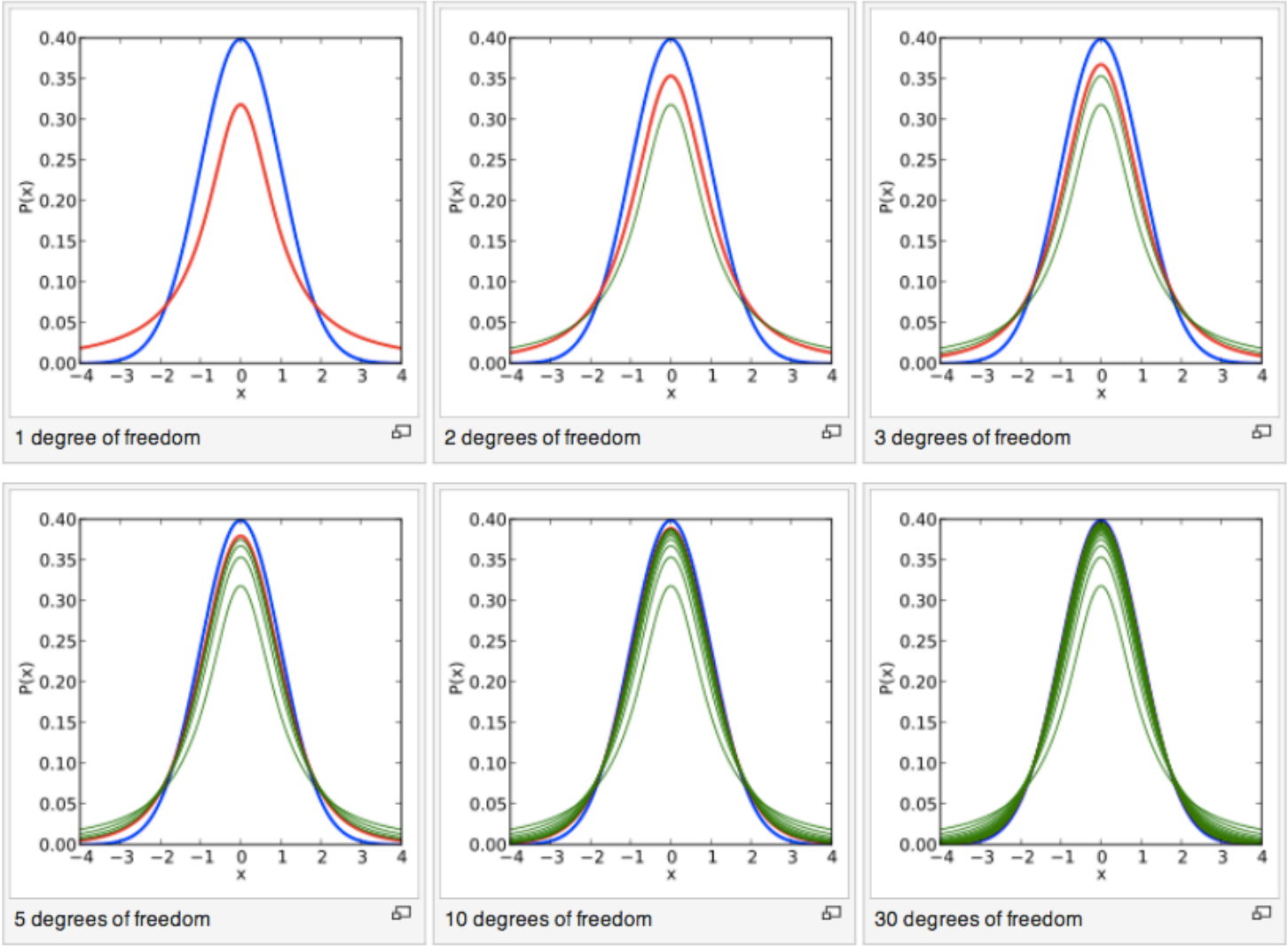
Khi kích thước mẫu nhỏ (n<30), thì $\bar{x}$ có phân bố Student (t-distribution) với (n-1) bậc tự do:

- Kì vọng μ
- Phương sai $\sigma^2_{\bar{x}} = s^2/n$.

Phân bố Student

Density of the *t*-distribution (red) for 1, 2, 3, 5, 10, and 30 degrees of freedom compared to the standard normal distribution (blue).

Previous plots shown in green.



Ví dụ 5

Để ước lượng chiều cao trung bình của nữ sinh ĐHCN, một tập mẫu ngẫu nhiên gồm 20 nữ sinh viên được chọn với kết quả như sau:

5 bạn 155cm; 8 bạn 160; 5 bạn 165cm; 2 bạn 170cm.

1. Hãy tìm khoảng tin cậy chiều cao với độ tin cậy $\beta=90\%$.
2. Hãy tìm khoảng tin cậy chiều cao với độ tin cậy $\beta=95\%$.
3. Hãy tìm khoảng tin cậy chiều cao với độ tin cậy $\beta=99\%$.

Khoảng tin cậy cho tỉ lệ

Nghiên cứu một quần thể mà mỗi cá thể có thể có hoặc không có một thuộc tính A nào đó.

- P là tỉ lệ cá thể có thuộc tính A trong quần thể
- f = k/n là tỉ lệ (tần suất) cá thể có thuộc tính A trong mẫu nghiên cứu

Câu hỏi: Ước lượng khoảng tin cậy cho tỉ lệ p dựa vào tần suất f.

Định lí: Tần suất f là một ĐLNN có phân bố xấp xỉ phân bố chuẩn với kì vọng $Ef = p$ và phương sai $Df = p(1-p)/n$ với điều kiện $np>5$ và $n(1-p)>5$.

Do không biết p, cho nên Df có thể được xấp xỉ bằng

$$Df = f(1-f)/n$$

với điều kiện $nf>10$ và $n(1-f)>10$.

Ví dụ 7

Trước ngày bầu cử tổng thống, Cơ quan khảo sát lấy ngẫu nhiên 100 người để hỏi ý kiến thì có 60 người ủng hộ ông Biden. Tìm khoảng tin cậy tỉ lệ cử tri bỏ phiếu cho ông Biden?

1. Với độ tin cậy 90%
2. Với độ tin cậy 95%
3. Với độ tin cậy 99%

Ví dụ 8

Một mẫu ngẫu nhiên gồm 100 người dùng xe máy có 30 người dùng xe Honda. Tìm khoảng tin cậy cho tỉ lệ người dùng xe Honda với:

1. Độ tin cậy 90%
2. Độ tin cậy 95%
3. Độ tin cậy 99%

Ví dụ 9

Kiểm tra ngẫu nhiên 300 người có 10 người mắc bệnh tim. Tìm khoảng tin cậy cho tỉ lệ người mắc bệnh tim trong toàn dân số với:

1. Độ tin cậy 90%
2. Độ tin cậy 95%
3. Độ tin cậy 99%