Khoảng tin cậy

Hoàng Văn Hà
University of Science, VNU - HCM
hvha@hcmus.edu.vn

Mục lục

Mục tiêu

Sau khi hoàn thành chương này, sinh viên có thể:

- Phân biệt được một ước lượng điểm (point estimate) và một ước lượng khoảng tin cậy (confidence interval estimate).
- Xây dựng và giải diễn giải được một ước lượng khoảng tin cậy cho trung bình của một tổng thể sử dụng các phân phối Z (phân phối chuẩn tắc) và phân phối t (phân phối Student).
- Xây dựng và giải diễn giải được một ước lượng khoảng tin cậy cho tỷ lệ của một tổng thể.

Bài toán ước lượng

Bài toán ước lượng

Các tham số đặc trưng của tổng thể như trung bình, tỷ lệ, phương sai, ...được sử dụng rất nhiều trong những phân tích kinh tế xã hội và các lĩnh vực khác. Tuy nhiên, các tham số đặc trưng này thường là chưa biết. Vì vậy đặt ra vấn đề cần ước lượng chúng bằng phương pháp mẫu.

Phát biểu bài toán

Giả sử ta cần khảo sát một đặc tính X thuộc một tổng thể xác định. Đại lượng ngẫu nhiên X có phân phối $F(x;\theta)$ trong đó tham số θ chưa biết. Hãy ước lượng tham số θ.

Các định nghĩa

Một ước lượng (estimator) của một tham số (của tổng thể) là

- một biến ngẫu nhiên có giá trị phụ thuộc vào thông tin của mẫu,
- giá trị của nó là một xấp xỉ cho tham số chưa biết của tổng thể.

Một giá trị cụ thể của biến ngẫu nhiên này gọi là một giá trị ước lượng (estimate).

Xét đại lượng ngẫu nhiên X có phân phối $F(x;\theta)$ với tham số θ chưa biết. Chọn mẫu ngẫu nhiên cỡ n từ X: $X_1, \ldots, X_n$.

- Thống kê $\hat{\Theta} = h(X_1, \ldots, X_n)$ gọi là một ước lượng điểm cho θ.
- Với một mẫu thực nghiệm $(x_1, \ldots, x_n)$, ta gọi $\hat{\theta} = h(x_1, \ldots, x_n)$ là một giá trị ước lượng điểm cho θ.

Ước lượng điểm: ví dụ

Gọi X = chiều cao của một sinh viên của một trường đại học (Đv: cm). Giả sử X tuân theo phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$ với trung bình μ và phương sai σ^2. Ta cần ước lượng trung bình tổng thể μ.

- Chọn một mẫu ngẫu nhiên cỡ n $(X_1, \ldots, X_n)$, thì trung bình mẫu

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

  là một ước lượng điểm cho trung bình tổng thể μ.

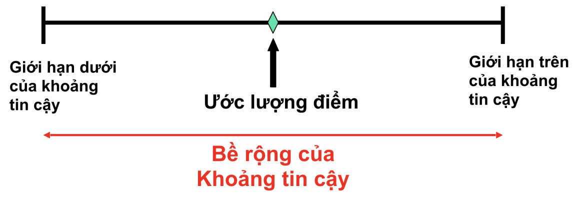
- Đo thực tế chiều cao của $n = 4$ sinh viên, thu được $x_1 = 160$, $x_2 = 168$, $x_3 = 155$ và $x_4 = 170$. Ta có

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{160 + 168 + 155 + 170}{4} = 165.75$$

  là một giá trị ước lượng điểm cụ thể cho μ.

Ước lượng điểm và ước lượng khoảng

- Một ước lượng điểm (point estimate) là một giá trị đơn,
- một khoảng tin cậy (confidence interval) cung cấp thông tin bổ sung về sự biến thiên của một ước lượng điểm tương ứng.



Khoảng tin cậy

- Đối với bài toán ước lượng điểm, một câu hỏi được đặt ra là: làm sao xác định được độ không chắc chắn của một ước lượng điểm cho một tham số của tổng thể?
- Để trả lời cho câu hỏi trên, ta sử dụng ước lượng khoảng tin cậy (confidence interval estimate)

Khoảng tin cậy và độ tin cậy

Định nghĩa 1

Cho $0 < \alpha < 1$, *một khoảng* $[L, U]$ *được gọi là một khoảng tin cậy* $100 \times (1-\alpha)\%$ *cho tham số* θ *nếu*

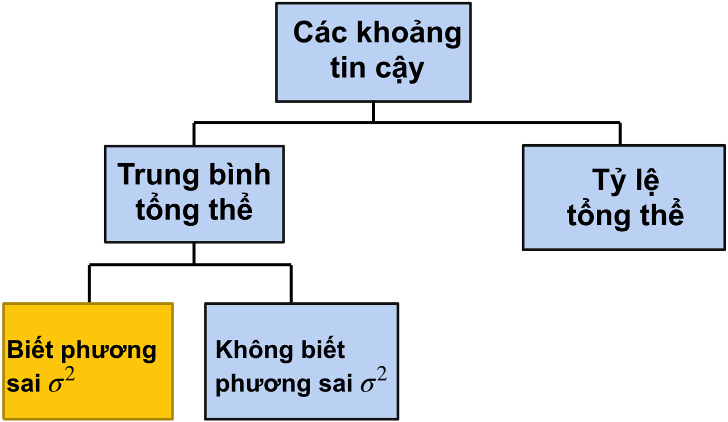
$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha.$$

Đại lượng $1-\alpha$ *được gọi là độ tin cậy (confidence level) của khoảng này.*

Ý nghĩa:

- Nếu lặp lại nhiều lần việc lấy mẫu từ một tổng thể, thì sẽ có $100 \times (1-\alpha)\%$ số khoảng được tính toán theo cách này sẽ chứa giá trị thực của tham số θ.
- Khoảng tin cậy được tính theo cách này được biểu diễn là: $L \leq \theta \leq U$ với độ tin cậy $100 \times (1-\alpha)\%$.

Khoảng tin cậy



Khoảng tin cậy cho kỳ vọng: TH biết σ^2

- **Các giả định:**
  - Mẫu ngẫu nhiên $X_1, \ldots, X_n$ được chọn từ tổng thể có phân phối chuẩn, $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$.
  - Biết phương sai σ^2.
  - Nếu tổng thể không tuân theo phân phối chuẩn, cần chọn cỡ mẫu lớn.
- **Công thức xác định khoảng tin cậy:**

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

trong đó $\bar{x}$ là trung bình mẫu, σ là độ lệch tiêu chuẩn của tổng thể (đã biết) và $z_{\alpha/2}$ là phân vị trên mức $\alpha/2$ của phân phối chuẩn tắc $\mathcal{N}(0, 1)$.

  - Tìm $z_{\alpha/2}$: tra bảng phân phối chuẩn tắc Z.

Sai số/Độ chính xác

- Khoảng tin cậy

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

- có thể được viết lại dưới dạng $\bar{x} \pm \varepsilon$ trong đó ε được gọi là sai số của ước lượng, hay còn được gọi là độ chính xác

$$\varepsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

- Chiều dài của khoảng tin cậy: $2 \times \varepsilon$.

Sai số/Độ chính xác

- Sai số

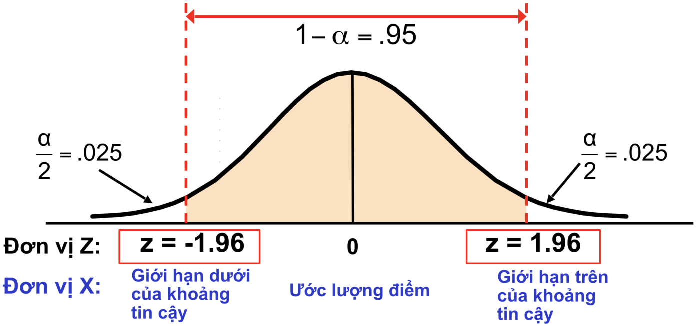
$$\varepsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

  có thể giảm được nếu
  - giảm độ lệch chuẩn của tổng thể ($\sigma \downarrow$),
  - tăng cỡ mẫu ($n \uparrow$),
  - giảm độ tin cậy, $1-\alpha \downarrow$.
- Câu hỏi: Với độ tin cậy $100(1-\alpha)\%$ Cần chọn cỡ mẫu bao nhiêu để thu được sai số ε^* như mong muốn?

$$n = \left(\frac{z_{\alpha/2}\times\sigma}{\varepsilon^*}\right)^2$$

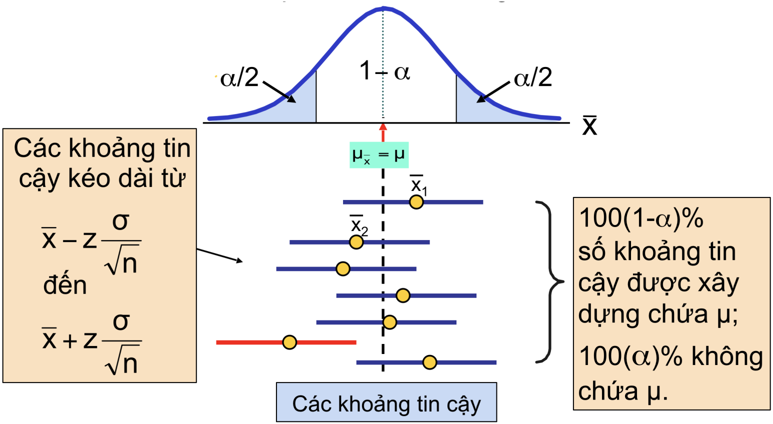
Tìm $z_{\alpha/2}$

- Xét một khoảng tin cậy 95% ($\alpha = 5\%$):



- Tra bảng phân phối chuẩn tắc Z: $z_{\alpha/2} = z_{0.025} = 1.96$.
- Cách kí hiệu khác: $z_{1-\alpha/2}$ (phân vị dưới - lower percentile).

Khoảng tin cậy cho kỳ vọng: tóm tắt



Khoảng tin cậy cho kỳ vọng: ví dụ

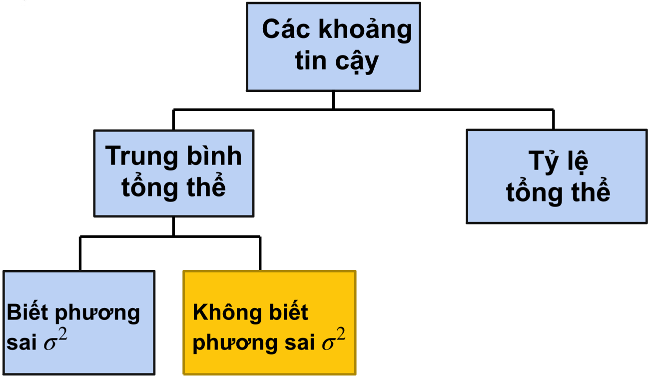
Ví dụ 1

Biết rằng tuổi thọ của các bóng đèn do một công ty sản xuất tuân theo phân phối chuẩn với độ lệch chuẩn là 40 giờ. Khảo sát một mẫu gồm 30 bóng đèn tính được tuổi thọ trung bình $\bar{x} = 780$ giờ.

a) *Tìm khoảng tin cậy 95% cho tuổi thọ trung bình của các bóng đèn do công ty này sản xuất.*

b) *Nếu muốn sai số ước lượng không quá 5 giờ, thì phải quan sát ít nhất bao nhiêu bóng đèn?*

Khoảng tin cậy cho kỳ vọng: TH không biết σ^2



Khoảng tin cậy cho kỳ vọng: TH không biết σ^2

- KTC cho kỳ vọng μ khi biết σ^2: $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
- Khi không biết σ, ta thay thế σ bằng độ lệch chuẩn mẫu S.
- Ta sử dụng phân phối Student t thay vì phân phối chuẩn.

Phân phối Student

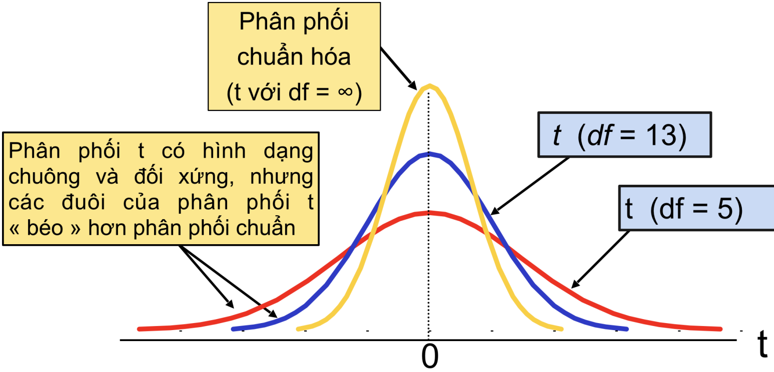
- Xét mẫu ngẫu nhiên $X_1, \ldots, X_n$ được chọn từ tổng thể có phân phối chuẩn có trung bình μ,
- trung bình mẫu $\bar{X}$ và độ lệch chuẩn S.
- Khi đó, biến ngẫu nhiên

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

có phân phối Student với $n - 1$ bậc tự do.

Phân phối Student

- $T \to \mathcal{N}(0,1)$ khi n tăng.



Khoảng tin cậy cho kỳ vọng: TH không biết σ^2

- **Các giả định:**
  - Phương sai của tổng thể σ^2 không được biết,
  - Tổng thể tuân theo phân phối chuẩn,
  - Nếu tổng thể không tuân theo phân phối chuẩn, sử dụng cỡ mẫu lớn,
- **Công thức xác định khoảng tin cậy:**

$$\bar{X} - t_{\alpha/2}^{n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}^{n-1}\frac{S}{\sqrt{n}},$$

trong đó $t_{\alpha/2}^{n-1}$ là phân vị (trên) mức $\alpha/2$ của phân phối Student với $n-1$ bậc tự do $t_{\alpha/2}^{n-1}$ thỏa

$$\mathbb{P}\left(T_{n-1} > t_{\alpha/2}^{n-1}\right) = \frac{\alpha}{2}.$$

Sai số

- Sai số (độ chính xác) của ước lượng:

$$\varepsilon = t_{\alpha/2}^{n-1}\frac{S}{\sqrt{n}}.$$

- Xác định cỡ mẫu n để có được sai số ε^* với độ tin cậy $100(1-\alpha)\%$ mong muốn:

$$n = \left(\frac{z_{\alpha/2} \times S}{\varepsilon^*}\right)^2$$

- **Chú ý:** khi cỡ mẫu n lớn, $t_{\alpha/2}^{n-1} \approx z_{\alpha/2}$.

Khoảng tin cậy cho kỳ vọng: ví dụ

Ví dụ 2

Đo chỉ số IQ của các sinh viên trong một trường đại học, khảo sát 18 *sinh viên thu được kết quả sau:*

130	122	119	142	136	127
120	152	141	132	127	118
150	141	133	137	129	142

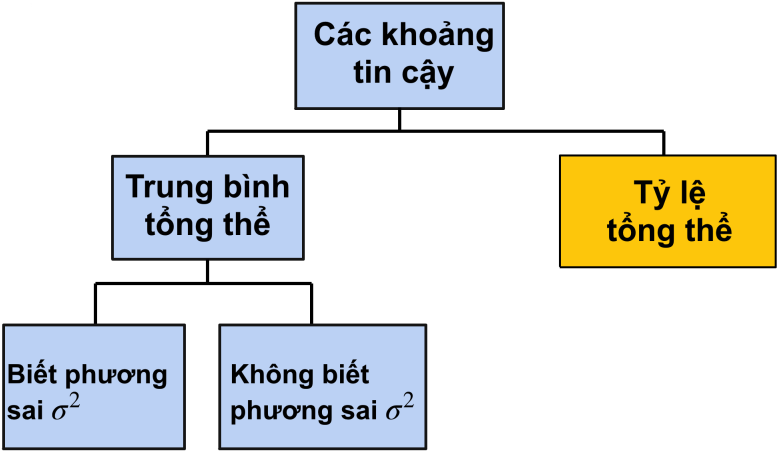
Biết rằng chỉ số IQ của sinh viên tuân theo phân phối chuẩn.

(a) *Vẽ đồ thị stem & leaf cho dữ liệu trên.*

(b) *Lập khoảng tin cậy* 99% *cho chỉ số IQ trung bình.*

(c) *Với độ tin cậy* 99%, *nếu muốn sai số ước lượng bằng* 1.5 *thì phải khảo sát thêm bao nhiêu sinh viên nữa?*

Khoảng tin cậy cho tỷ lệ



Khoảng tin cậy cho tỷ lệ

- Gọi p là tỷ lệ của những phần tử thỏa một đặc tính $\mathcal{A}$ của tổng thể, mà ta quan tâm.
  - Ví dụ: tỷ lệ những người ủng hộ một ứng viên A trong một cuộc bầu cử, tỷ lệ những người thuận tay trái, tỷ lệ sản phẩm hỏng trong một nhà máy sản xuất, ...
- **Câu hỏi:** lập khoảng tin cậy cho tỷ lệ p với độ tin cậy $100(1-\alpha)\%$.

Khoảng tin cậy cho tỷ lệ

Các bước thực hiện:

- Khảo sát n phần tử, đặt

$$Y_i = \begin{cases} 1, & \text{nếu phần tử thứ } i \text{ thỏa tính chất } \mathcal{A}, \\ 0, & \text{nếu không,} \end{cases}$$

  Ta có $Y_i \sim B(p)$, $i = 1, \dots, n$.
- Đặt $X = Y_1 + \ldots + Y_n = \sum_{i=1}^{n} Y_i$ thì $X =$ tổng số phần tử thóa tính chất $\mathcal{A}$ trong n phần tử khảo sát và $X \sim B(n, p)$.
- Tỷ lệ mẫu được tính bởi

$$\hat{P} = \frac{X}{n}.$$

Phân phối của tỷ lệ mẫu

- Nhắc lại: với n đủ lớn, ta có thể xấp xỉ phân phối của $\hat{P}$ bởi phân phối chuẩn với kỳ vọng $\mu_{\hat{P}} = p$ và độ lệch chuẩn

$$\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$$

- Khi đó, nếu ta đặt

$$Z = \frac{\hat{P} - \mu_{\hat{P}}}{\sigma_{\hat{P}}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{X - np}{\sqrt{np(1-p)}},$$

  thì $Z \sim \mathcal{N}(0,1)$.

- Do đó, ta sử dụng phân phối chuẩn tắc $\mathcal{N}(0,1)$ khi tính khoảng tin cậy cho tỷ lệ.

Khoảng tin cậy cho tỷ lệ

- **Giả định:** cỡ mẫu n đủ lớn.
- **Công thức tính khoảng tin cậy:**

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

trong đó $z_{\alpha/2}$ là phân vị trên mức $\alpha/2$ của phân phối chuẩn tắc, $\hat{P}$ là tỷ lệ mẫu.

Khoảng tin cậy cho tỷ lệ

- Độ chính xác (sai số) của khoảng tin cậy:

$$\varepsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

- Câu hỏi: Với độ tin cậy $100(1-\alpha)\%$, chọn cỡ mẫu n bằng bao nhiêu để có được sai số ε^* (cho trước) như mong muốn?
  - Dùng công thức

$$n = \left(\frac{z_{\alpha/2}}{\varepsilon^*}\right)^2 \hat{p}(1-\hat{p}).$$

Khoảng tin cậy cho tỷ lệ

Trường hợp muốn tính cỡ mẫu khi chưa có một ước lượng $\hat{p}$ của p:

- Ta có $p(1-p) \leq 0.25$ với $0 < p < 1$.
- Do vậy, nếu ta muốn sai số trong việc ước lượng p bởi $\hat{p}$ bé hơn ε^* với độ tin cậy ít nhất là $100(1-\alpha)\%$ thì cần chọn cỡ mẫu bằng

$$n = \left(\frac{z_{\alpha/2}}{\varepsilon^*}\right)^2 \times 0.25.$$

Khoảng tin cậy cho tỷ lệ: ví dụ

Ví dụ 3

Trong một nhà máy, ở khâu kiểm tra chất lượng sản phẩm, người ta lấy ngẫu nhiên 100 sản phẩm trong một lô hàng thì phát hiện được 20 sản phẩm kém chất lượng.

- *Hãy tìm KTC 95% cho tỷ lệ sản phẩm kém chất lượng của mỗi lô hàng.*
- *Với độ tin cậy 99%, nếu muốn độ chính xác bằng 0.04 thì phải kiểm tra bao nhiêu sản phẩm?*

Ví dụ 4

Trong một khu rừng nguyên sinh, người ta theo dõi một loài chim bằng cách đeo vòng cho chúng. Thực hiện đeo vòng cho 1000 con. Sau một thời gian bắt lại 200 con thì thấy có 40 con có đeo vòng. Hãy ước lượng số chim trong khu rừng đó với độ tin cậy 95%.

Bài tập

Bài tập 1

Đem cân một số trái cây vừa thu hoạch, ta thu được kết quả sau:

X (g)	200 - 210	210 - 220	220 - 230	230-240	240-250
Số trái	12	17	20	18	15

(a) Tính trung bình mẫu và độ lệch chuẩn mẫu.

(b) Tìm khoảng tin cậy 99% cho trọng lượng trung bình của trái cây.

(c) Với độ tin cậy 99%, nếu muốn sai số ước lượng không quá $E = 2$ g thì phải quan sát ít nhất bao nhiêu trái?

Bài tập

Bài tập 2

Đo chiều cao X (đv: cm) của một nhóm thanh nhiên ở một khu vực, ghi nhận được

X (cm)	140-145	145-150	150-155	155-160	160-165	165-170
Số người	1	3	7	9	5	2

a. Sử dụng 1 loại đồ thị thích hợp để chứng tỏ rằng số liệu mẫu chọn từ một biến ngẫu nhiên có phân phối chuẩn.

b. Tính các tham số mẫu.

c. Ước lượng chiều cao trung bình của thanh nhiên khu vực này với độ tin cậy 99%. Nếu muốn sai số ước lượng bằng 1 (cm) thì phải khảo sát thêm bao nhiêu người.

d. Những thanh niên có chiều cao trên 160 (cm) được xếp loại sức khỏe loại A. Hãy tìm khoảng tin cậy cho tỷ lệ thanh niên có sức khỏe loại A với độ tin cậy 95%.

Bài tập

Bài tập 3

Một loại thuốc mới đem điều trị cho 50 người bị bệnh B, kết quả có 40 người khỏi bệnh.

(a) Ước lượng tỷ lệ người khỏi bệnh khi dùng thuốc với độ tin cậy 95% và 99%.

(b) Nếu muốn sai số ước lượng không quá 0.02 ở độ tin cậy 95% thì phải khảo sát ít nhất bao nhiêu trường hợp.