Chương 5. Lý thuyết mẫu

Nguyễn Minh Trí

Ngày 27 tháng 3 năm 2025

Mục lục

5.1 Các khái niệm cơ bản của thống kê

Định nghĩa 5.1. Thống kê (statistics) là khoa học về thu thập, tổ chức, tóm tắt và phân tích thông tin để rút ra kết luận hoặc trả lời câu hỏi. Ngoài ra, thống kê còn là việc cung cấp thước đo độ tin cậy trong bất kỳ kết luận nào.

- **Thống kê mô tả** (descriptive statistics): Các phương pháp bao gồm chủ yếu là tổ chức, tóm tắt và trình bày dữ liệu dưới dạng bảng, đồ thị và biểu đồ.
- **Thống kê suy luận** (inferential statistics): Các phương pháp rút ra kết luận và đưa ra quyết định về tổng thể từ các mẫu.

Định nghĩa 5.2. Tổng thể (population) là tập hợp tất cả các đối tượng hoặc phép đo mà người thu thập quan tâm.

Ví dụ 5.3. Giả sử chúng ta muốn nghiên cứu chiều cao của tất cả nam sinh viên tại một trường đại học nào đó.

- **Tổng thể** là tập hợp các chiều cao đo được của tất cả sinh viên nam trong trường đại học.
- Tổng thể **không** phải là tập hợp tất cả sinh viên nam trong trường đại học.

Trong thực tế, ta khó thể có được thông tin của toàn bộ tổng thể. Mục tiêu chính của thống kê là thu thập và nghiên cứu một tập hợp con của tổng thể, được gọi là **mẫu**, để đưa ra thông tin về một số đặc điểm của tổng thể.

Định nghĩa 5.4. Mẫu (sample) là một tập hợp con của dữ liệu được chọn từ tổng thể. **Kích thước mẫu** là số phần tử của tập hợp con đó.

Ví dụ 5.5. Chúng ta muốn ước tính tỷ lệ phần trăm sản phẩm lỗi được sản xuất tại một nhà máy trong một ngày. Trong trường hợp này, "tất cả các sản phẩm được sản xuất trong ngày" là **tổng thể**. Chọn ngẫu nhiên 100 sản phẩm để kiểm tra số sản phẩm lỗi. Khi đó ta có một mẫu là "100 sản phẩm" và kích thước mẫu là 100.

Định nghĩa 5.6. Một mẫu được chọn sao cho mọi phần tử của tổng thể đều có cơ hội được chọn như nhau và độc lập với nhau được gọi là **mẫu ngẫu nhiên đơn giản** (simple random sample).

Ví dụ 5.7. Có 100 quả bóng trong một cái hộp. Lấy ra 10 quả bóng mà không cần nhìn vào trong hộp. Khi đó 10 quả bóng được lấy ra là một mẫu ngẫu nhiên.

Chú ý:

- Việc chọn mẫu luôn được xem là chọn không hoàn lại, tức là một phần tử nào đó không được chọn nhiều hơn 1 lần trong một mẫu.
- Mẫu phải được chọn ngẫu nhiên và kích thước mẫu đủ lớn.

5.2 Biểu diễn đồ họa của dữ liệu

Định nghĩa 5.8. 1. Cho một mẫu có kích thước n và các giá trị của dấu hiệu X mà ta muốn nghiên cứu là $x_1 < x_2 < \ldots < x_m$. Số lần lặp lại k_i của x_i được gọi là tần số của x_i. Bảng phân bố tần số

X	x_1	x_2	$\cdots$	x_m
Tần số	k_1	k_2	$\cdots$	k_m

2. Tần suất f_i của giá trị x_i :

$$f_i = \frac{k_i}{n}$$

Bảng phân bố tần suất

X	x_1	x_2	$\cdots$	x_m
Tần suất	f_1	f_2	$\cdots$	f_m

Ví dụ 5.9. Kiểm tra 80 hộp (mỗi hộp chứa 100 chip bán dẫn) để tìm số lượng chip bị lỗi trong mỗi hộp. Số chip bị lỗi trong mỗi hộp như sau

1	3	4	7	2	7	5	5
2	2	4	2	5	4	3	2
2	7	1	3	3	2	5	0
0	1	2	5	5	4	1	3
3	2	6	3	8	2	2	3
1	6	3	4	1	2	5	3
1	3	3	3	2	1	2	5
5	4	1	4	3	1	0	3
2	1	2	4	4	5	3	3
4	0	5	2	5	6	2	1

Số chip bị lỗi	Tần số	Tần suất
0	4	0,05
1	12	0,15
2	18	0,225
3	17	0,2125
4	10	0,125
5	12	0,15
6	3	0,0375
7	3	0,0375
8	1	0,0125
≥ 9	0	0
Tổng	80	1

Người ta thường xác định một số khoảng $C_1, C_2, \ldots, C_m$ sao cho mỗi giá trị mà X nhận được chỉ thuộc một khoảng nào đó. Các khoảng này được gọi là các lớp ghép của X.

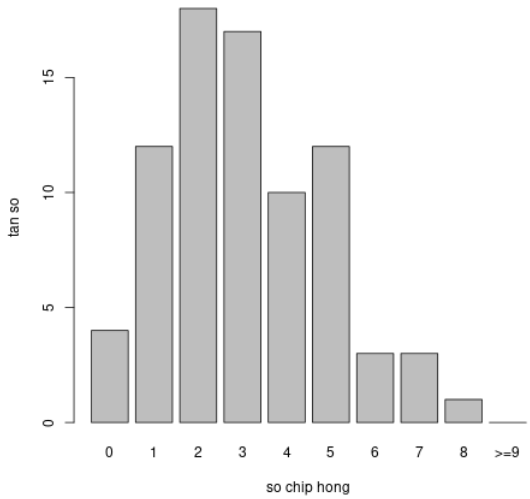
Ví dụ 5.10. Một mẫu về chiều cao của 40 sinh viên được trình bày trong bảng phân bố lớp ghép sau:

Khoảng	Tần số	Tần suất
(146; 151]	4	0,1
(151; 156]	2	0,05
(156; 161]	6	0,15
(161; 166]	10	0,25
(166; 171]	12	0,3
(171; 176]	6	0,15

Định nghĩa 5.11. Biểu đồ gồm các cột có chiều cao biểu thị tần số (tần suất) tương ứng của các loại đối tượng được gọi là biểu đồ cột (bar chart).

Ví dụ 5.12. Số chip bị lỗi trong mỗi hộp như sau

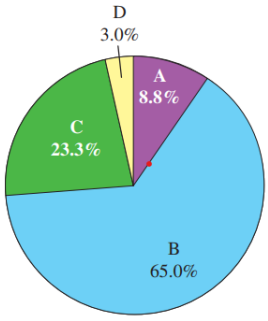
Số chip bị lỗi	Tần số	Tần suất
0	4	0,05
1	12	0,15
2	18	0,225
3	17	0,2125
4	10	0,125
5	12	0,15
6	3	0,0375
7	3	0,0375
8	1	0,0125
≥ 9	0	0
Tổng	80	1



Định nghĩa 5.13. Một hình tròn được chia thành các phần biểu thị tỷ lệ phần trăm tổng thể hoặc một mẫu thuộc các danh mục khác nhau được gọi là biểu đồ tròn (pie chart).

Ví dụ 5.14. Số chip bị lỗi

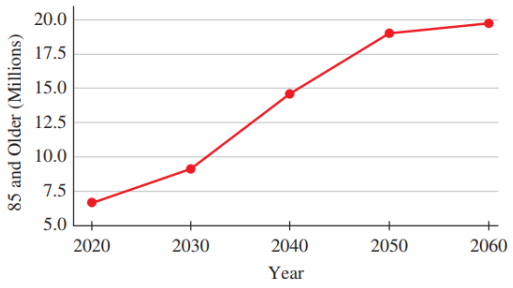
Số chip bị lỗi	Tần số	Tỉ lệ %
0	4	5%
1	12	15%
2	18	22,5%
3	17	21,25%
4	10	12,5%
5	12	15%
6	3	3,75%
7	3	3,75%
8	1	1,25%
≥ 9	0	0%
Tổng	80	100%



Định nghĩa 5.15. Biểu đồ đường (line chart) là một loại biểu đồ thống kê trong đó mỗi điểm dữ liệu (giá trị và tần số/tần suất) được biểu diễn bằng một điểm trên đồ thị và những điểm dữ liệu liên tiếp được kết nối bằng một đường thẳng.

Ví dụ 5.16. Số chip bị lỗi

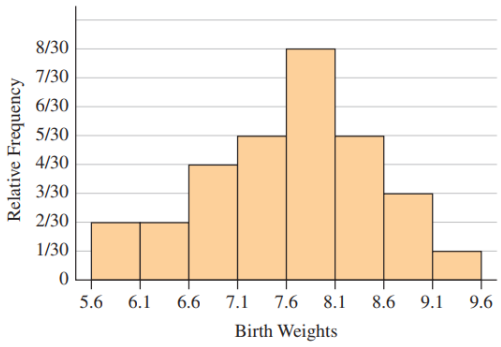
Số chip bị lỗi	Tần số	Tỉ lệ %
0	4	5%
1	12	15%
2	18	22,5%
3	17	21,25%
4	10	12,5%
5	12	15%
6	3	3,75%
7	3	3,75%
8	1	1,25%
≥ 9	0	0%
Tổng	80	100%



Định nghĩa 5.17. Tổ chức đồ (histogram) là biểu đồ trong đó các lớp ghép được đánh dấu trên trục hoành và tần số/tần suất/tỷ lệ phần trăm được biểu thị bằng độ cao trên trục tung. Trong một tổ chức đồ, các cột được vẽ liền kề nhau mà không có bất kỳ khoảng trống nào.

Ví dụ 5.18. Chiều cao của 40 sinh viên

Khoảng	Tần số
(155; 160]	8
(160; 165]	10
(165; 170]	15
(170; 175]	5
(175; 180]	2



5.3 Các số đo mô tả

- Một tổng thể có kích thước N : $v_1, v_2, \ldots, v_N$
- Một mẫu có kích thước n nhận các giá trị $x_1, \ldots, x_n$.

Định nghĩa 5.19. 1. **Trung bình tổng thể**

$$\mu = \frac{1}{N}\sum_{i=1}^{N} v_i.$$

2. **Phương sai tổng thể**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \mu)^2.$$

3. **Độ lệch chuẩn tổng thể**
$$\sigma = \sqrt{\sigma^2}.$$

4. **Trung bình của một mẫu** (sample mean)
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

5. **Phương sai mẫu** (sample variance), ký hiệu s^2,
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

6. **Độ lệch chuẩn mẫu** (sample standard deviation)
$$s = \sqrt{s^2}.$$

Ví dụ 5.20. So sánh giá cà phê tại 4 cửa hàng tạp hóa được chọn ngẫu nhiên ở Thủ Đức cho thấy các mức tăng so với tháng trước là 12, 15, 17 và 20 nghìn đồng cho một túi 1 kg. Tìm trung bình, phương sai của mẫu này.

Giải.

- Trung bình mẫu
$$\overline{x} = \frac{12+15+17+20}{4} = 16 \text{ (nghìn đồng)}$$

- Phương sai mẫu
$$\begin{aligned} s^2 &= \frac{1}{3}\sum_{i=1}^{4}(x_i - \overline{x})^2 \\ &= \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3} = \frac{34}{3} \end{aligned}$$

Định lí 5.21. Nếu s^2 là phương sai của một mẫu có kích thước n thì
$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

Định nghĩa 5.22. Giả sử các giá trị của mẫu được sắp xếp từ nhỏ đến lớn. Trung vị mẫu (median) là một số m thỏa mãn
$$m = \begin{cases} x_{(n+1)/2}, & n \text{ lẻ} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & n \text{ chẵn} \end{cases}$$

Ví dụ 5.23. Các số liệu thu được như sau:

a. 100 45 60 130 30

b. 100 45 60 130 30 70

Tìm trung vị mẫu.

Giải. a. Sắp xếp dữ liệu theo thứ tự

$$30 \quad 45 \quad 60 \quad 100 \quad 130$$

Khi đó trung vị mẫu là 60.

b. Sắp xếp dữ liệu theo thứ tự

$$30 \quad 45 \quad 60 \quad 70 \quad 100 \quad 130$$

Khi đó trung vị mẫu là $\dfrac{60+70}{2} = 65$.

Định nghĩa 5.24. Cho một phân bố lớp ghép với m khoảng $C_1, C_2, \ldots, C_m$. Giả sử x_i là trung điểm (tâm) của khoảng C_i và k_i là tần số của khoảng C_i với $i = 1, 2, \ldots, m$. Khi đó trung bình mẫu $\overline{x}$ và phương sai mẫu s^2 được xác định bởi

$$\overline{x} = \frac{\sum_{i=1}^{m} k_i x_i}{\sum_{i=1}^{m} k_i}; \quad s^2 = \frac{1}{\sum_{i=1}^{m} k_i - 1} \sum_{i=1}^{m} k_i (x_i - \overline{x})^2$$

Ví dụ 5.25. Chiều cao của 40 sinh viên

Khoảng	Tần số
$(155; 160]$	8
$(160; 165]$	10
$(165; 170]$	15
$(170; 175]$	5
$(175; 180]$	2

- Trung bình mẫu

$$\overline{x} = \frac{1}{40}(8.157,5 + 10.162,5 + 15.167,5 + 5.172,5 + 2.177,5) = 165,375$$

- Phương sai: $s^2 = 30,625$.

5.4 Phân phối mẫu

Định nghĩa 5.26. Cho các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ nhận giá trị từ một tổng thể. Tập hợp các biến ngẫu nhiên $\{X_1, X_2, \ldots, X_n\}$ tạo thành một mẫu ngẫu nhiên có kích thước n nếu

1. Các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ có phân phối giống nhau.
2. Các biến ngẫu nhiên $X_1, X_2, \ldots, X_n$ độc lập với nhau.

Ví dụ 5.27. Cho một tổng thể gồm 6 phần tử $\{2, 4, 6, 6, 7, 8\}$.

- Mẫu ngẫu nhiên gồm 3 phần tử $\{X_1, X_2, X_3\}$

- Mẫu ngẫu nhiên này có thể nhận các giá trị $\{2,4,6\},\{2,6,7\},\{4,6,8\}$.

Định nghĩa 5.28. Một hàm số được tính từ mẫu ngẫu nhiên $\{X_1, X_2, \ldots, X_n\}$ được gọi là một thống kê (statistic).

Ta có một số thống kê

- $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ (trung bình mẫu ngẫu nhiên)
- $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2$ (phương sai mẫu ngẫu nhiên)
- Đặt X là số phần tử có đặc điểm * mà ta quan tâm trong một mẫu có kích thước. Đặt biến ngẫu nhiên
$$\hat{P} = \frac{X}{n}$$
và gọi là tỉ lệ mẫu ngẫu nhiên.

Khi chọn một mẫu từ một tổng thể, các số đo mô tả tính được từ mẫu đó được gọi là các giá trị thống kê (statistic). Các giá trị thống kê thay đổi theo các mẫu khác nhau mà ta chọn, do đó chúng là các giá trị của một biến ngẫu nhiên (thống kê). Phân phối xác suất của các thống kê được gọi là các phân phối mẫu (sampling distribution)

Định nghĩa 5.29. Phân phối mẫu (sampling distribution) của một thống kê là phân phối xác suất của một thống kê.

Ví dụ 5.30. Cho một tổng thể gồm 6 phần tử $\{2,4,6,6,7,8\}$. Xét tất cả các mẫu có 2 phần tử được chọn ngẫu nhiên (không chọn lại). Tìm phân phối mẫu của trung bình mẫu.

Giải. Có tất cả 15 mẫu ngẫu nhiên có kích thước bằng 2

Mẫu	Trung bình mẫu		Mẫu	Trung bình mẫu		Mẫu	Trung bình mẫu
2,4	3		4,6	5		6,7	6,5
2,6	4		4,6	5		6,8	7
2,6	4		4,7	5,5		6,7	6,5
2,7	4,5		4,8	6		6,8	7
2,8	5		6,6	6		7,8	7,5

Phân phối mẫu của trung bình mẫu (kích thước mẫu bằng 2)

$\overline{X}$	3	4	4,5	5	5,5	6	6,5	7	7,5
$P(\overline{X} = \overline{x}_i)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{2}{15}$	$\frac{1}{15}$

- Trung bình của $\overline{X}$
$$E(\overline{X}) = 3.\frac{1}{15} + 4.\frac{2}{15} + \cdots + 7,5.\frac{1}{15} = 5,5 = \mu \text{ (trung bình tổng thể)}$$
- Phương sai của $\overline{X}$
$$\mathrm{Var}(\overline{X}) = \frac{1}{15}(3-5,5)^2 + \frac{2}{15}(4-5,5)^2 + \cdots + \frac{1}{15}(7,5-5,5)^2 = \frac{47}{30}$$

5.5 Phân phối mẫu của trung bình mẫu

Các phân phối mẫu có thể bao gồm vô hạn mẫu có kích thước nào đó. Mọi thống kê mẫu đều có phân phối mẫu.

Định lí 5.31. Cho $\{X_1, X_2, \ldots, X_n\}$ là một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể vô hạn có trung bình là μ và phương sai là σ^2. Khi đó

$$E(\overline{X}) = \mu \text{ và } \text{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Nhận xét: Nếu tổng thể có phân phối chuẩn $N(\mu, \sigma^2)$ thì phân phối mẫu của trung bình có phân phối chuẩn $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Định lí 5.32. Cho $\{X_1, X_2, \ldots, X_n\}$ là một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể có kích thước N với trung bình tổng thể là μ và phương sai tổng thể là σ^2. Khi đó

$$E(\overline{X}) = \mu \text{ và } \text{Var}(\overline{X}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1}.$$

Ví dụ 5.33. (Xem **Ví dụ 5.30**) Cho một tổng thể gồm 6 phần tử $\{2, 4, 6, 6, 7, 8\}$. Xét tất cả các mẫu có 2 phần tử được chọn ngẫu nhiên (chọn không hoàn lại). Xác định độ lệch chuẩn của trung bình mẫu.

Giải. Ta có $N = 6; n = 2$, trung bình tổng thể $\mu = 5,5$ và phương sai tổng thể

$$\sigma^2 = \frac{1}{6}\left((2-5,5)^2 + (4-5,5)^2 + \cdots + (8-5,5)^2\right) = \frac{47}{12}.$$

Phương sai của trung bình mẫu

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1} = \frac{47/12}{2}.\frac{6-2}{6-1} = \frac{47}{30}.$$

Nhận xét:

- Nếu tổng thể không có phân phối chuẩn thì theo Định lý giới hạn trung tâm, phân phối mẫu của trung bình mẫu $\overline{X}$ sẽ xấp xỉ phân phối chuẩn khi kích thước mẫu n đủ lớn ($n \geq 30$).
- Nếu tổng thể có phân phối chuẩn hay không chuẩn thì phân phối mẫu của trung bình mẫu là xấp xỉ phân phối chuẩn khi kích thước mẫu lớn hơn hoặc bằng 30.

Trường hợp tổng thể có phân phối chuẩn nhưng không biết độ lệch chuẩn.

Định lí 5.34. Nếu $\overline{X}$ là trung bình của mẫu ngẫu nhiên có kích thước n lấy từ một tổng thể có phân phối chuẩn với trung bình tổng thể là μ và phương sai mẫu ngẫu nhiên S^2 thì

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

là biến ngẫu nhiên có phân phối Student với bậc tự do $n-1$.

Ví dụ 5.35. Giả sử lượng chất rắn lơ lửng trong nước thải của một công ty có phân phối chuẩn với trung bình 40 mg/l. Người ta lấy ngẫu nhiên 20 mẫu nước thải và thấy rằng độ lệch chuẩn của mẫu này là $s = 9,4$ mg/l. Xác suất lượng chất thải trung bình của 20 mẫu này nhỏ hơn 46 mg/l là bao nhiêu?

Giải. Theo đề bài, ta có $\mu = 40; n = 20$ và $s = 9,4$. Đặt $\overline{X}$ là trung bình của mẫu ngẫu nhiên. Khi đó

$$\begin{aligned} P(\overline{X} < 46) &= P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} < \frac{46 - 40}{9,4/\sqrt{20}}\right) \\ &= P(T < 2,855) \\ &= 1 - P(T \geq 2,855) \approx 1 - 0,01 = 0,99. \end{aligned}$$

5.6 Phân phối mẫu của tỉ lệ mẫu

Định nghĩa 5.36. Trong một tổng thể có kích thước N, có y phần tử có tính chất $\mathcal{P}$ mà ta quan tâm. Giả sử X là số phần tử có tính chất $\mathcal{P}$ trong một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể. Khi đó tỉ lệ tổng thể

$$p = \frac{y}{N}$$

và tỉ lệ mẫu ngẫu nhiên (sample proportion)

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

Tỉ lệ mẫu ngẫu nhiên $\hat{P}$ là một thống kê và do đó nó có phân phối mẫu.

Ví dụ 5.37. Cho một tổng thể gồm 6 phần tử $\{2, 4, 6, 6, 7, 8\}$. Xét tất cả các mẫu có 2 phần tử được chọn ngẫu nhiên (chọn không hoàn lại). Giả sử ta quan tâm các phần tử nhỏ hơn 5. Tìm phân phối mẫu của tỉ lệ mẫu của mẫu ngẫu nhiên có 2 phần tử.

Giải. Có tất cả 15 mẫu ngẫu nhiên đơn giản có kích thước bằng 2

Mẫu	Tỉ lệ mẫu		Mẫu	Tỉ lệ mẫu		Mẫu	Tỉ lệ mẫu
2,4	1		4,6	1/2		6,7	0
2,6	1/2		4,6	1/2		6,8	0
2,6	1/2		4,7	1/2		6,7	0
2,7	1/2		4,8	1/2		6,8	0
2,8	1/2		6,6	0		7,8	0

phân phối mẫu của tỉ lệ mẫu ngẫu nhiên (kích thước mẫu bằng 2)

$\hat{P}$	0	1/2	1
$P(\hat{P} = \hat{p}_i)$	$\frac{6}{15}$	$\frac{8}{15}$	$\frac{1}{15}$

Khi đó

$$E(\hat{P}) = 0.\frac{6}{15} + \frac{1}{2}.\frac{8}{15} + 1.\frac{1}{15} = \frac{1}{3}$$

và

$$\text{Var}(\hat{P}) = \frac{6}{15}(0 - \frac{1}{3})^2 + \frac{8}{15}.(\frac{1}{2} - \frac{1}{3})^2 + \frac{1}{15}.(1 - \frac{1}{3})^2 = \frac{4}{45}.$$

Định lí 5.38. Cho p là tỉ lệ của một tổng thể và $\hat{P}$ là tỉ lệ mẫu ngẫu nhiên có kích thước n. Khi đó

$$E(\hat{P}) = p$$

và

- nếu kích thước tổng thể là N hữu hạn thì

$$\text{Var}(\hat{P}) = \frac{N-n}{N-1}.\frac{p(1-p)}{n}.$$

- nếu kích thước tổng thể là vô hạn thì

$$\text{Var}(\hat{P}) = \frac{p(1-p)}{n}.$$

Ví dụ 5.39. (**Ví dụ 5.37**) Cho một tổng thể gồm 6 phần tử $\{2, 4, 6, 6, 7, 8\}$. Xét tất cả các mẫu có 2 phần tử được chọn ngẫu nhiên (chọn không hoàn lại). Giả sử ta quan tâm các phần tử nhỏ hơn 5. Tìm phân phối mẫu của tỉ lệ mẫu ngẫu nhiên có 2 phần tử.

- Trung bình của phân phối mẫu của tỉ lệ mẫu

$$E(\hat{P}) = \frac{1}{3} = \frac{2}{6} = p.$$

- Phương sai phân phối mẫu của tỉ lệ mẫu

$$\text{Var}(\hat{P}) = \frac{N-n}{N-1}\frac{p(1-p)}{n} = \frac{6-2}{6-1}.\frac{1/3(1-1/3)}{2} = \frac{4}{45}.$$

Định lí 5.40. Nếu kích thước mẫu n đủ lớn thì tỉ lệ mẫu ngẫu nhiên $\hat{P}$ có phân phối chuẩn $N\left(p, \dfrac{p(1-p)}{n}\right)$.

Nhận xét: Định lý 5.40 được áp dụng khi $np, n(1-p) \geq 5$.

Ví dụ 5.41. Chọn ngẫu nhiên 270 người từ một thành phố để ước tính tỉ lệ người không sử dụng mạng xã hội. Người ta thấy rằng tỉ lệ người dân không sử dụng mạng xã hội của thành phố này là 20%. Xác suất tỉ lệ mẫu này từ 16% đến 24% là bao nhiêu?

Giải. Ta có tỉ lệ tổng thể $p = 0,2$ và kích thước mẫu $n = 270$. Vì $np = 54, n(1-p) = 270(1-0,2) = 216 > 5$ nên $\hat{P} \sim N\left(p, \dfrac{p(1-p)}{n}\right)$. Phương sai của phân phối mẫu của $\hat{P}$ là

$$\text{Var}(\hat{P}) = \frac{p(1-p)}{n} = \frac{0,2(1-0,2)}{270} = 0,00059.$$

Khi đó

$$\begin{aligned} P(0,16 \leq \hat{P} \leq 0,24) &= P\left(\frac{0,16-0,2}{\sqrt{0,00059}} \leq \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{0,24-0,2}{\sqrt{0,00059}}\right) \\ &= P\left(-1,67 \leq Z \leq 1,67\right) \\ &= 0,905 \end{aligned}$$

Bài tập

Bài tập 5.1. Cho một tổng thể gồm 5 phần tử $\{1, 2, 3, 4, 5\}$. Xét tất cả các mẫu ngẫu nhiên (không hoàn lại) có kích thước bằng 3 từ tổng thể. Tìm phân phối mẫu của trung bình mẫu. Sau đó, tìm trung bình và phương sai của trung bình mẫu ngẫu nhiên có kích thước bằng 3. (Đáp số: $E(\overline{X}) = 3, \text{Var}(\overline{X}) = 1/3$)

Bài tập 5.2. Giả sử điểm chỉ số IQ của người có phân phối chuẩn với trung bình 100 và độ lệch chuẩn 15. Chọn một mẫu ngẫu nhiên đơn giản từ 10 người. Tính xác suất điểm chỉ số IQ trung bình của 10 người này lớn hơn 110. (Đáp số: 0,0175)

Bài tập 5.3. Một nhà sản xuất cầu chì tuyên bố rằng với mức quá tải 20% cầu chì sẽ nổ. Chọn một mẫu ngẫu nhiên gồm 16 cầu chì trong số này đã bị quá tải 20% và thấy rằng thời gian chúng bị nổ có độ lệch chuẩn mẫu là 0,9 phút. Giả sử thời gian cầu chì bị nổ khi bị quá tải 20% có phân phối chuẩn với trung bình là 10 phút. Tính xác suất thời gian nổ trung bình của 16 cầu chì được chọn nhiều hơn 10,4 phút. (Đáp số: 0,05)

Bài tập 5.4. Người ta ước tính rằng 43% số người tốt nghiệp đại học tin thấy rằng giỏi tiếng Anh là một điều quan trọng. Chọn ngẫu nhiên một mẫu gồm 80 người đã tốt nghiệp đại học. Tính xác suất có hơn một nửa của mẫu này có niềm tin như trên. (Đáp số: 0,102)

Chương 6. Lý thuyết ước lượng

Nguyễn Minh Trí

Ngày 27 tháng 3 năm 2025

Mục lục

- Các giá trị trung bình, phương sai, độ lệch chuẩn và trung vị của tổng thể được gọi là các **tham số** (parameters).
- Các giá trị trung bình, phương sai, độ lệch chuẩn và trung vị của mẫu được gọi là các **thống kê** (statistics).

6.1 Ước lượng điểm

Định nghĩa 6.1. 1. Một ước lượng điểm là một giá trị dùng để ước lượng một tham số.

2. Một ước lượng khoảng là một khoảng giá trị dùng để ước lượng một tham số.

Ví dụ 6.2.
- Nếu nói chiều cao trung bình của sinh viên nam Trường Đại học Công nghệ Thông tin là 174 cm thì đó là một **giá trị ước lượng điểm**.
- Nếu nói chiều cao trung bình đó nằm trong khoảng từ 159 cm đến 169 cm hay 164 ± 5 cm thì ta đã có một **ước lượng khoảng.**

Định nghĩa 6.3. Ước lượng điểm (point estimator) của tham số tổng thể là một biến ngẫu nhiên phụ thuộc vào thông tin mẫu, giá trị của nó cho ta một sự xấp xỉ của tham số chưa biết này. Một giá trị cụ thể của biến ngẫu nhiên đó được gọi là **giá trị ước lượng điểm** (point estimate).

Ví dụ 6.4.
- Trung bình mẫu ngẫu nhiên $\overline{X}$ là ước lượng điểm (point estimator) của trung bình tổng thể μ.
- Độ lệch chuẩn mẫu ngẫu nhiên S^2 là ước lượng điểm của độ lệch chuẩn tổng thể σ^2.
- Tỷ lệ mẫu ngẫu nhiên $\hat{P}$ là ước lượng điểm của tỷ lệ tổng thể p.

Các giá trị cụ thể của $\overline{X}, S^2, \hat{P}$ được gọi là các giá trị ước lượng điểm (point estimate). Ký hiệu

$$\theta : \text{tham số của tổng thể mà ta quan tâm}$$
$$\hat{\theta} : \text{thống kê mẫu hoặc ước lượng điểm của } \theta$$

Định nghĩa 6.5. Thống kê mẫu $\hat{\theta}$ được gọi là một ước lượng không lệch (unbiased estimator) của tham số tổng thể θ nếu

$$E(\hat{\theta}) = \theta.$$

Ví dụ 6.6. Trong chương 5, ta đã có:

- Trung bình mẫu ngẫu nhiên $\overline{X}$ là ước lượng không lệch của trung bình tổng thể μ vì $E(\overline{X}) = \mu$.
- Tỉ lệ mẫu ngẫu nhiên $\hat{P}$ ước lượng không lệch của tỷ lệ tổng thể p vì $E(\hat{P}) = p$.

6.2 Ước lượng khoảng

- Vì một ước lượng điểm không thể cung cấp chính xác giá trị của tham số tổng thể nên ta thường dùng ước lượng khoảng (interval estimate).
- Ước lượng khoảng cung cấp thông tin về mức độ gần của ước lượng điểm do mẫu cung cấp với giá trị của tham số tổng thể.
- Ước lượng khoảng được xây dựng sao cho khi lấy mẫu lặp lại nhiều lần thì một tỷ lệ lớn các khoảng này sẽ bao quanh tham số tổng thể mà chúng ta đang quan tâm. Tỷ lệ này là độ tin cậy (confidence level) và khoảng được tạo ra được gọi là khoảng tin cậy (confidence interval).

Định nghĩa 6.7.

1. **Độ tin cậy** (confidence level), ký hiệu $1 - \alpha$, của ước lượng khoảng của một tham số là xác suất khoảng ước lượng chứa tham số. Giả sử một số lượng lớn mẫu được chọn và quá trình ước lượng trên cùng một tham số được lặp lại.

2. **Khoảng tin cậy** (confidence interval) là một khoảng ước lượng cụ thể của một tham số tương ứng với độ tin cậy đã cho.

Ví dụ 6.8. Khi nói **khoảng tin cậy** của chiều cao trung bình của sinh viên các trường đại học tại TPHCM là $[155; 175]$ với **độ tin cậy** 95% có nghĩa là xác suất khoảng $[155; 175]$ chứa trung bình tổng thể là 95%.

Bài toán. Gọi θ là tham số mà ta quan tâm. Tìm khoảng ước lượng của θ với độ tin cậy $1 - \alpha$. Tức là, ta cần tìm một đoạn $[a, b]$ sao cho

$$P(a \leq \theta \leq b) = 1 - \alpha.$$

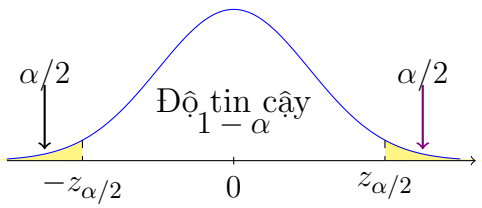
6.2.1 Ước lượng khoảng cho trung bình tổng thể khi biết σ

Bài toán 1. Giả sử rằng thời gian mua sắm của khách hàng tại một trung tâm thương mại có phân phối chuẩn với độ lệch chuẩn tổng thể là 20 phút. Chọn ngẫu nhiên 64 người đã mua sắm ở trung tâm đó. Người ta thấy rằng thời gian mua sắm trung bình của 64 người này là 75 phút. Tìm thời gian mua sắm trung bình của khách hàng tại trung tâm này với độ tin cậy 95%.

Dạng bài toán: Ước lượng trung bình tổng thể μ khi biết σ và độ tin cậy $1-\alpha$.

Cho $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên có kích thước n được lấy từ một tổng thể có phân phối chuẩn $N(\mu, \sigma^2)$ trong đó **đã biết** σ. Giả sử ta cần tìm khoảng tin cậy $1-\alpha$ của trung bình tổng thể.

- Từ mẫu đã cho, ta tính được trung bình mẫu $\overline{x}$.
- Cần tìm ε sao cho
$$P(\overline{x} - \varepsilon \le \mu \le \overline{x} + \varepsilon) = 1 - \alpha.$$
$$P\left(\frac{-\varepsilon}{\sigma/\sqrt{n}} \le \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \le \frac{\varepsilon}{\sigma/\sqrt{n}}\right) = 1 - \alpha$$
- Trong chương 5, ta đã biết biến ngẫu nhiên
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$
có phân phối chuẩn chuẩn tắc.
- Đặt $z_{\alpha/2} = \dfrac{\varepsilon}{\sigma/\sqrt{n}}$ và cần tìm $z_{\alpha/2}$ sao cho
$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$
- Với độ tin cậy $1-\alpha$ thì giá trị $z_{\alpha/2}$ thỏa mãn
$$\Phi(z_{\alpha/2}) = 1 - \alpha/2.$$



- Tìm được $z_{\alpha/2}$ bằng bảng phụ lục A4.
- Trung bình tổng thể sẽ thuộc khoảng (khoảng tin cậy của trung bình tổng thể với độ tin cậy $1-\alpha$)
$$\left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- Số $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ được gọi là **sai số của ước lượng** hoặc **độ chính xác** của ước lượng.

Bài toán 1. Giả sử rằng thời gian mua sắm của khách hàng tại một trung tâm thương mại có phân phối chuẩn với độ lệch chuẩn tổng thể là 20 phút. Chọn ngẫu nhiên 64 người đã mua sắm ở trung tâm đó. Người ta thấy rằng thời gian mua sắm trung bình của 64 người này là 75 phút. Tìm thời gian mua sắm trung bình của khách hàng tại trung tâm này với độ tin cậy 95%.

Giải Bài toán 1.

- Ta có $\overline{x} = ...; n = ...$ và $\sigma = ...;$
- Độ tin cậy $1 - \alpha = 95\%$. Suy ra $\alpha = ...$ và $z_{\alpha/2} = ...$
- Độ chính xác $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} = ...\dfrac{.....}{.....} = ...$
- Khoảng tin cậy của trung bình tổng thể với độ tin cậy 95% là

$$[.....].$$

Ví dụ 6.9. Độ tuổi của 50 kỹ sư IT được chọn ngẫu nhiên (tại một thành phố nọ) là 41,4. Tìm khoảng tin cậy 99% cho độ tuổi trung bình của tất cả kỹ sư IT. Giả sử độ lệch chuẩn theo độ tuổi của toàn bộ kỹ sư IT là 12,1 năm.

Giải. .....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Ví dụ 6.10. Thu nhập trung bình hàng tháng của 30 người dân trong một thành phố là 31,2 triệu đồng. Tìm khoảng tin cậy 90% của thu nhập trung bình hàng tháng của toàn thành phố. Biết thu nhập trung bình hàng tháng có phân phối chuẩn và có độ lệch chuẩn 3.4.

Giải. .....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

6.2.2 Ước lượng khoảng cho trung bình tổng thể khi chưa biết σ

Trường hợp 1: Kích thước mẫu $n \geq 30$.

1. Xác định $\overline{x}, s$ là trung bình và độ lệch chuẩn của một mẫu cụ thể
2. Đổi biến $Z = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, khi đó $Z \sim N(0;1)$.
3. Tra bảng $A4$, tìm $z_{\alpha/2}$.
4. Khoảng tin cậy của μ với độ tin cậy $1-\alpha$ là

$$[\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}].$$

Trường hợp 2: Kích thước mẫu $n < 30$ và tổng thể có phân phối chuẩn

1. Đổi biến $T = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, khi đó $T \sim St(n-1)$.
2. Tra bảng $A5$ dòng $n-1$, tìm $t_{\alpha/2}$ thỏa mãn $P(T > t_{\alpha/2}) = \dfrac{\alpha}{2}$.
3. Khoảng tin cậy của μ với độ tin cậy $1-\alpha$ là

$$[\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}].$$

Cách tìm t_β trong bảng A5: bậc tự do của phân phối Student là $n-1$. Cột đầu tiên bên trái của bảng A5 là cột bậc tự do, hai hàng đầu tiên bên trên là giá trị của β. Số nằm ở vị trí của giao của hàng tương ứng với bậc tự do $n-1$ và cột tương ứng với β là giá trị của t_β.

Ví dụ 6.11. Tìm giá trị $t_{0,005}$ với bậc tự do 17. Theo bảng A5, ta có $t_{0,005} = 2,898$.

ν	α, the right-tail probability									
(d.f.)	.10	.05	.025	.02	.01	.005	.0025	.001	.0005	.0001
1	3.078	6.314	12.706	15.89	31.82	63.66	127.3	318.3	636.6	3185
2	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60	70.71
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92	22.20
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610	13.04
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869	9.676
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959	8.023
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408	7.064
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041	6.442
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781	6.009
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587	5.694
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437	5.453
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318	5.263
13	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221	5.111
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140	4.985
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073	4.880
16	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015	4.790
17	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965	4.715
18	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922	4.648
19	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883	4.590
20	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850	4.539

Ví dụ 6.12. Theo một thống kê cho thấy số thời gian (tính bằng phút) dành cho việc đọc và trả lời email của 7 nhân viên văn phòng của một công ty trong 1 ngày là

$$54{,}6 \quad 59 \quad 60{,}9 \quad 63{,}1 \quad 71{,}6 \quad 84{,}4 \quad 99{,}3$$

Giả sử thời gian đọc và trả lời email của nhân viên văn phòng có phân phối chuẩn. Tính khoảng thời gian đọc và trả lời email trung bình của nhân viên văn phòng của công ty này với độ tin cậy 99%.

Giải.

- Trung bình mẫu: $\overline{x} =$ .....
- Độ lệch chuẩn mẫu: $s =$ .....
- Tìm $t_{\alpha/2}$ với độ tin cậy $1-\alpha = 0,99$ và bậc tự do 6. Ta có $t_{\alpha/2} =$ .....
- Khoảng tin cậy cần tìm

  .....

  .....

  .....

Ví dụ 6.13. Kiểm tra tuổi thọ (tính bằng giờ) của 50 bóng đèn do nhà máy A sản xuất, người ta được bảng số liệu sau

Tuổi thọ	(3350;3450]	(3450;3550]	(3550;3650]	(3650;3750]
Số bóng đèn	10	20	12	8

a. Ước tính tuổi thọ trung bình của các bóng đèn do nhà máy A sản xuất với độ tin cậy 97%.

b. Dựa vào mẫu trên để ước lượng tuổi thọ trung bình của các bóng đèn do nhà máy A sản xuất có độ chính xác 59,02 giờ thì phải đảm bảo độ tin cậy là bao nhiêu?

c. Dựa vào mẫu trên, nếu muốn ước lượng tuổi thọ trung bình của các bóng đèn do nhà máy A sản suất có độ chính xác nhỏ hơn 40 giờ với độ tin cậy 98% thì cần phải kiểm tra tối thiểu bao nhiêu bóng đèn?

Giải. a. (Kích thước mẫu bằng 50 và chưa biết độ lệch chuẩn tổng thể)

- Trung bình mẫu: $\overline{x} =$ .....
- Độ lệch chuẩn mẫu: $s =$ .....
- Độ tin cậy $1-\alpha = 0,97$. Suy ra $\Phi(z_{\alpha/2}) = 1-\alpha/2 = 0,985$. Do đó $z_{\alpha/2} =$ .....
- Độ chính xác:
$$z_{\alpha/2}\frac{s}{\sqrt{n}} = 2,17\frac{217,3683}{\sqrt{50}} = ...$$
- Khoảng tin cậy của tuổi thọ trung của bóng đèn với độ tin cậy 97% là

.....

b. Ta có độ chính xác bằng ..... giờ, tức là

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

Suy ra

$$z_{\alpha/2} = 59,02\frac{\sqrt{n}}{s} = ...$$

Do đó

$$\Phi(z_{\alpha/2}) = \Phi(.....) = 1-\frac{\alpha}{2}.$$

Trang bảng A4, ta có $\Phi(.....) =$ ..... và do đó $\alpha =$ .....

Như vậy, độ tin cậy là .....

c. Ta có độ chính xác nhỏ hơn 40 giờ với độ tin cậy 98%, tức là

$$z_{\alpha/2}\frac{s}{\sqrt{n}} < 40.$$

Suy ra

$$\sqrt{n} > z_{\alpha/2}\frac{s}{40}$$

Vì $1-\alpha = 0,98$ nên $\Phi(z_{\alpha/2}) = 1-\frac{\alpha}{2} = 0,99$. Suy ra $z_{\alpha/2} =$ .....Do đó

$$\sqrt{n} > z_{\alpha/2}\frac{s}{40} = 2,33\frac{217,3683}{40} = ...$$

Như vậy, $n >$ ..... và do đó cần khảo sát ít nhất..... bóng đèn.

Ví dụ 6.14. Một thống kê cho thấy chi phí (tính bằng triệu) của các mẫu quảng cáo 30-giây trên một số đài truyền hình được cho như sau

14 55 105 9 15 66 23 30 50 22 12 13 54 73 55 41 78

Giả sử chi phí cho một video quảng cáo 30-giây có phân phối chuẩn. Ước tính chi phí trung bình cho một quảng cáo 30-giây trên truyền hình với độ tin cậy 90%.

Giải. .....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

6.2.3 Ước lượng tỉ lệ của tổng thể

Bài toán 2. Trong mẫu 500 trang web mới được đăng ký trên Internet, có 24 trang web ẩn danh (tức là họ che giấu tên và thông tin liên lạc của mình). Xây dựng khoảng tin cậy 95% cho tỷ lệ tất cả các trang web mới ẩn danh.

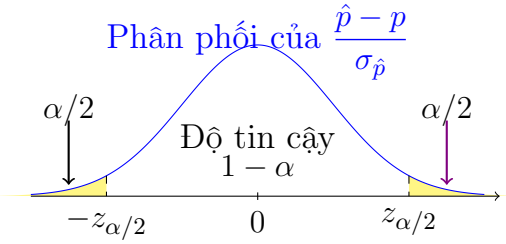
- p : tỉ lệ tổng thể (tỉ lệ phần tử có tính chất $\mathcal{P}$ trong tổng thể)
- $\hat{P}$: tỉ lệ mẫu ngẫu nhiên (tỉ lệ phần tử có tính chất $\mathcal{P}$ trong mẫu)
- Nếu $np, n(1-p) \geq 5$ thì tỉ lệ mẫu ngẫu nhiên $\hat{P}$ có phân phối chuẩn $N(p; \sigma^2_{\hat{p}})$ với $\sigma^2_{\hat{P}} = \dfrac{p(1-p)}{n}$.
- Biến ngẫu nhiên $\dfrac{\hat{P}-p}{\sigma_{\hat{P}}}$ có phân phối chuẩn chuẩn tắc.

Từ một mẫu đã chọn, khoảng ước lượng của p có dạng $[\hat{p}-\varepsilon;\hat{p}+\varepsilon]$. Khi đó

$$P(\hat{p}-\varepsilon \le p \le \hat{p}+\varepsilon) = 1-\alpha$$

$$P\left(\frac{-\varepsilon}{\sigma_{\hat{p}}} \le \frac{\hat{p}-p}{\sigma_{\hat{p}}} \le \frac{\varepsilon}{\sigma_{\hat{p}}}\right) = 1-\alpha.$$

Đặt $z_{\alpha/2} = \dfrac{\varepsilon}{\sigma_{\hat{p}}}$



- Vì p chưa biết (ta cần ước lượng) nên khi n đủ lớn, ta có thể thay p trong $\sigma_{\hat{p}}$ bởi giá trị của tỉ lệ mẫu $\hat{p}$.
- Với độ tin cậy $1-\alpha$, khoảng tin cậy chứa tỉ lệ tổng thể là

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

  trong đó $\Phi(z_{\alpha/2}) = 1 - \dfrac{\alpha}{2}$ (xem phụ lục A4).

- Độ chính xác (sai số) là $z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

Theo bất đẳng thức Cauchy, ta có độ chính xác (sai số)

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

Do đó, sai số tối đa trong ước lượng tỉ lệ tổng thể là $\dfrac{z_{\alpha/2}}{2\sqrt{n}}$.

Bài toán 2. Trong mẫu 500 trang web mới được đăng ký trên Internet, có 24 trang web ẩn danh (tức là họ che giấu tên và thông tin liên lạc của mình). Xây dựng khoảng tin cậy 95% cho tỷ lệ tất cả các trang web mới ẩn danh.

Giải Bài toán 2. Theo đề bài

- Tỉ lệ mẫu cụ thể .....
- Kích thước mẫu $n =$ .....
- Độ tin cậy $1-\alpha =$ ..... suy ra $1 - \dfrac{\alpha}{2} =$ ..... Do đó $z_{\alpha/2} =$ .....
- Độ chính xác (sai số) là

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

- Khoảng tin cậy .....

Như vậy có từ 70,68% đến 89,32% cử tri ủng hộ ứng viên A.

Ví dụ 6.15. Lấy ngẫu nhiên 200 sản phẩm trong một kho hàng để kiểm tra thì thấy có 21 sản phẩm có lỗi.

a. Với độ tin cậy 95%, hãy ước lượng tỉ lệ sản phẩm lỗi của cả kho hàng.

b. Dựa vào mẫu trên, để ước tính tỉ lệ sản phẩm bị lỗi có độ chính xác là 0,035 thì độ tin cậy bằng bao nhiêu?

c. Dựa vào mẫu trên, nếu muốn ước lượng tỉ lệ sản phẩm bị lỗi với độ chính xác nhỏ hơn 0,01 với độ tin cậy 93% thì cần kiểm tra ít nhất bao nhiêu sản phẩm.

Giải. .....

Bài tập

Bài tập 6.1. Trọng lượng của những vĩ thuốc do một công ty sản xuất có phân phối chuẩn với độ lệch chuẩn 0,038mg. Một mẫu ngẫu nhiên gồm 10 vĩ thuốc có trọng lượng trung bình 4,87mg. Hãy ước lượng trọng lượng trung bình của các vĩ thuốc do công ty sản xuất với độ tin cậy 95%.

Bài tập 6.2. Do đường kính trung bình của một mẫu ngẫu nhiên gồm 100 vòng bi do một máy sản suất trong một tuần có đường kính trung bình 0,824 cm và độ lệch chuẩn mẫu hiệu chỉnh là 0,042 cm. Hãy tìm khoảng tin cậy của tất cả các vòng bi với độ tin cậy 96%.

Bài tập 6.3. Để nghiên cứu khối lượng rác sinh hoạt thải ra trong một ngày tại một thành phố, người ta khảo sát ngẫu nhiên 400 gia đình. Kết quả khảo sát như sau

Khối lượng (kg/ngày)	0,5	1,5	2,5	3,5	4,5	5,5	6,5	7,5
Số gia đình	10	35	86	132	78	31	18	10

a. Hãy ước tính khối lượng rác trung bình thải ra của toàn bộ các hộ gia đình trong 1 ngày với độ tin cậy 99%. Biết rằng khối lượng rác thải ra có phân phối chuẩn và có độ lệch chuẩn là 0,75 kg.

b. Với mẫu khảo sát trên, nếu ước lượng lượng rác thải hàng ngày này với độ chính xác nhỏ hơn 0,08kg/ngày và độ tin cậy 95% thì cần khảo sát tối thiểu bao nhiêu gia đình?

Bài tập 6.4. Khảo sát giá (triệu đồng) của 10 loại laptop có RAM 8G tại một số cửa hàng kinh doanh online ta được bảng số liệu sau

18,5 27,5 26,4 17,9 28 27 14,5 22 24 28

Giả sử giá tiền của các laptop có RAM 8G có phân phối chuẩn. Hãy ước tính giá trung bình của các loại laptop có RAM 8G với độ tin cậy 95%.

Bài tập 6.5. Một tỉnh nọ có 1 triệu thanh niên trên 18 tuổi. Người ta khảo sát ngẫu nhiên 20 000 thanh niên của tỉnh này về trình độ học vấn thì thấy có 12575 thanh niên đã tốt nghiệp THPT. Hãy ước tính tỉ lệ thanh niên tốt nghiệp THPT của tỉnh này với độ tin cậy 95%.

Bài tập 6.6. Lấy ngẫu nhiên 200 sản phẩm trong một kho hàng để kiểm tra thì thấy có 21 sản phẩm có lỗi.

a. Dựa vào mẫu trên, để ước tính tỉ lệ sản phẩm bị lỗi có độ chính xác là 0,035 thì độ tin cậy bằng bao nhiêu?

b. Dựa vào mẫu trên, nếu muốn ước lượng tỉ lệ sản phẩm bị lỗi với độ chính xác nhỏ hơn 0,01 với độ tin cậy 93% thì cần kiểm tra ít nhất bao nhiêu sản phẩm.

Bài tập 6.7. Một nhà sản xuất bóng đèn tuyên bố rằng tuổi thọ của các bóng đèn của họ được phân phối chuẩn với giá trị trung bình là 60.000 giờ và độ lệch chuẩn là 4.000 giờ. Một mẫu ngẫu nhiên gồm 16 bóng đèn có tuổi thọ trung bình là 58.500 giờ. Nếu tuyên bố của nhà sản xuất là đúng thì xác suất giá trị trung bình mẫu là 58.500 hoặc thấp hơn là bao nhiêu? (0,0668)

Bài tập 6.8. Người ta ước tính rằng 43% sinh viên tốt nghiệp ngành công nghệ thông tin tin rằng một khóa học về lập trình Python là rất quan trọng để có thể tìm được việc làm tại các công ty lớn. Tìm xác suất để hơn một nửa mẫu ngẫu nhiên gồm 80 sinh viên tốt nghiệp ngành công nghệ thông tin có niềm tin này. (0,102)

Bài tập 6.9. Một mẫu ngẫu nhiên gồm 270 laptop được lấy từ một lượng lớn các laptop cũ để ước tính tỷ lệ laptop có ổ cứng bị lỗi. Nếu trên thực tế, 20% laptop có ổ cứng bị lỗi thì xác suất để tỷ lệ mẫu nằm trong khoảng từ 16% đến 24% là bao nhiêu? (0.905)

Bài tập 6.10. Thăm dò 500 người dân tại thành phố Hồ Chí Minh về việc xây lại Dinh Độc lập, có 380 người không đồng ý.

a. Hãy ước lượng tỉ lệ người dân TPHCM không đồng ý xây lại Dinh Độc Lập với độ tin cậy 95%.

b. Nếu muốn độ chính xác của ước lượng này là 3% thì độ tin cậy bằng bao nhiêu?

c. Nếu muốn độ chính xác của ước lượng này nhỏ hơn 3% với độ tin cậy 99% thì cần khảo sát ít nhất bao nhiêu người?

Chương 7. Kiểm định giả thuyết

Nguyễn Minh Trí

Ngày 31 tháng 3 năm 2025

Mục lục

Mục tiêu của thống kê là đưa ra suy luận về các tham số tổng thể chưa biết dựa trên thông tin có trong dữ liệu mẫu. Những suy luận này được diễn đạt theo một trong hai cách: như ước lượng các tham số tương ứng hoặc như các kiểm định giả thuyết về giá trị của chúng.

Theo nhiều cách, quy trình chính thức để kiểm định giả thuyết tương tự như phương pháp khoa học. Nhà khoa học quan sát thế giới tự nhiên, đề xuất các giả thuyết, sau đó kiểm tra giả thuyết này bằng quan sát. Giả sử nhà khoa học đặt ra giả thuyết liên quan đến một hoặc nhiều tham số tổng thể. Sau đó, nhà khoa học ấy lấy mẫu từ tổng thể và so sánh các quan sát của mình với giả thuyết. Nếu các quan sát không phù hợp với giả thuyết thì nhà khoa học bác bỏ nó. Nếu các quan sát phù hợp thì nhà khoa học kết luận rằng giả thuyết là đúng hoặc mẫu không phát hiện ra sự khác biệt giữa các giá trị thực và giá trị giả thuyết của các tham số tổng thể.

Ví dụ, một nhà nghiên cứu y khoa có thể đưa ra giả thuyết rằng một loại thuốc mới có hiệu quả hơn loại thuốc khác trong việc chống lại một căn bệnh. Để kiểm tra giả thuyết của mình, ông ấy chọn ngẫu nhiên những bệnh nhân bị nhiễm bệnh và chia ngẫu nhiên họ thành hai nhóm. Thuốc mới A được dùng cho những bệnh nhân trong nhóm đầu tiên, và thuốc cũ B được dùng cho những bệnh nhân trong nhóm thứ hai. Sau đó, dựa trên số lượng bệnh nhân trong mỗi nhóm phục hồi sau căn bệnh, nhà nghiên cứu phải quyết định xem loại thuốc mới có hiệu quả hơn loại thuốc cũ hay không.

7.1 Các khái niệm

Bài toán. Một công ty phần mềm tuyên bố rằng thời gian trung bình để tải một trang web trên nền tảng của họ không vượt quá 2 giây. Để kiểm tra tuyên bố này, công ty quyết định thu thập một mẫu ngẫu nhiên gồm 50 trang web chạy trên nền tảng của họ, đo lường thời gian tải của từng trang và tính toán thời gian tải trung bình từ dữ liệu mẫu.

Dựa trên kết quả thu thập được, công ty có đủ bằng chứng cho tuyên bố của họ có đúng hay không với một mức ý nghĩa thống kê nhất định.

Định nghĩa 7.1. Giả thuyết thống kê là một dự đoán về một tham số của tổng thể.

Khi kiểm tra một giả thuyết về một tham số θ nào đó, có hai giả thuyết liên quan: giả thuyết do người thử nghiệm đề xuất và phủ định giả thuyết này. Giả thuyết trước, được ký hiệu là H_1, được gọi là *đối thuyết* (alternative hyperthesis) hoặc *giả thuyết nghiên cứu*; giả thuyết sau được ký hiệu là H_0 và được gọi là *giả thuyết* (null hyperthesis). Mục đích của kiểm định là để quyết định xem bằng chứng có xu hướng bác bỏ *giả thuyết* hay không.

Định nghĩa 7.2. 1. **Giả thuyết** (null hypothesis), kí hiệu H_0, là một giả thuyết thống kê nói rằng **không có sự khác biệt** giữa một tham số và một giá trị cụ thể hoặc không có sự khác biệt giữa hai tham số.

2. **Đối thuyết** (alternative hypothesis), ký hiệu H_1, là một giả thuyết thống kê cho biết **có sự khác biệt** giữa một tham số và một giá trị cụ thể, hoặc có sự khác biệt giữa hai tham số.

- Khi kiểm tra một giả thuyết liên quan đến giá trị của một số tham số θ, phát biểu về sự bằng nhau sẽ luôn được bao gồm trong H_0. Tức là, H_0 xác định một giá trị số cụ thể. Giá trị này được ký hiệu là θ_0.

- Vì giả thuyết nghiên cứu là H_1 nên chúng ta hy vọng rằng bằng chứng sẽ dẫn đến việc bác bỏ H_0 và do đó sẽ ủng hộ H_1

Ví dụ 7.3. Các kỹ sư đường bộ đã phát hiện ra rằng có nhiều yếu tố ảnh hưởng đến hiệu suất của các biển báo phản quang trên đường bộ. Một trong số đó là sự căn chỉnh đèn pha của ô tô. Người ta cho rằng hơn 50% xe ô tô trên đường có đèn pha không đúng hướng. Nếu lập luận này có thể được hỗ trợ về mặt thống kê thì một chương trình kiểm tra mới nghiêm ngặt hơn sẽ được đưa vào hoạt động.

Đặt p là tỷ lệ ô tô đang hoạt động có đèn pha không đúng hướng. Vì chúng ta muốn ủng hộ rằng $p > 0,5$ nên lập luận này được coi là đối thuyết H_1. Khi đó giả thuyết là phủ định của H_1. Khi đó, ta có giả thuyết và đối thuyết như sau

$$\text{Giả thuyết } H_0 : p \leq 0,5 \text{ và đối thuyết } p > 0,5.$$

Chú ý rằng phát biểu liên quan đến đẳng thức (dấu bằng) luôn xuất hiện trong *giả thuyết* H_0.

Để tiến hành kiểm định giả thuyết, ta sẽ chọn một mẫu ngẫu nhiên và thu được các dữ liệu từ mẫu này. Dựa vào mẫu thu thập được, ta sẽ đưa ra quyết định. Kết quả của mỗi kiểm định là có *đủ bằng chứng để bác bỏ* H_0 **hoặc** *không có đủ bằng chứng để bác bỏ* H_0.

Quyết định được đưa ra bằng cách quan sát giá trị của một số thống kê có phân phối xác suất đã biết với giả định rằng giá trị θ_0 là giá trị của tham số tổng thể θ. Một thống kê như vậy được gọi là *giá trị kiểm định thống kê* (test statistic). Nếu giá trị kiểm định thống kê giả định một giá trị hiếm khi thấy khi $\theta = \theta_0$ và có xu hướng cho thấy đối thuyết là đúng thì chúng ta bác bỏ H_0 để ủng hộ H_1; nếu giá trị quan sát được là giá trị thường xảy ra theo giả định $\theta = \theta_0$ thì chúng ta không bác bỏ giả thuyết H_0. Điều này có nghĩa là khi kết thúc bất kỳ nghiên cứu nào, chúng ta sẽ phải rơi vào một trong những tình huống sau:

1. Ta sẽ bác bỏ H_0 trong khi H_0 đúng. Ta xem đây là *sai lầm loại I.*

2. Ta sẽ đưa ra kết luận đúng về việc bác bỏ H_0 khi H_1 đúng.

3. Ta sẽ đưa ra kết luận không bác bỏ H_0 khi H_1 đúng. Ta xem đây là *sai lầm loại II*.

4. Ta sẽ đưa ra kết luận không bác bỏ H_0 khi H_0 đúng.

Ví dụ 7.4. Xét bài toán kiểm định giả thuyết trong Ví dụ 7.3.

$$\text{Giả thuyết } H_0 : p \leq 0,5 \text{ và đối thuyết } p > 0,5.$$

- Ta mắc sai lầm loại I, tức là H_0 đúng nhưng ta bác bỏ H_0. Khi đó, ta kết luận rằng phần lớn các xe đang chạy trên đường có đèn pha không đúng hướng. Tuy nhiên, trong thực tế điều này không đúng. Sai lầm này có thể dẫn đến việc triển khai chương trình kiểm tra nhưng không cần thiết.

- Nếu sai lầm loại II xảy ra nếu ta không bác bỏ H_0 khi H_1 đúng. Trong trường hợp này sẽ dẫn đến việc không triển khai các chương trình kiểm tra đèn xe, nhưng thực tế là cần cần triển khai chương trình kiểm tra đèn xe.

Chú ý rằng, mặc dù ta làm điều gì thì sai lầm vẫn có thể xảy ra. Đó là điều không thể tránh được. Do đó, công việc của các nhà thống kê là thiết kế các phương pháp để xác định liệu rằng nên bác bỏ hay không bác bỏ H_0 sao cho xác suất mắc sai lầm là nhỏ nhất có thể.

Trong phần này, chúng ta sẽ xem xét phương pháp kiểm định giả thuyết để đưa ra quyết định bác bỏ hay không bác bỏ H_0. Kiểm định giả thuyết bao gồm một thủ tục trong đó các giá trị của kiểm định thống kê được thiết lập trước khi tiến hành thí nghiệm. Các giá trị kiểm định thống kê này dẫn đến việc bác bỏ giả thuyết H_0. Các giá trị này tạo nên một vùng được gọi là *miền tới hạn* hoặc *miền bác bỏ* của phép thử.

Định nghĩa 7.5. *Giá trị tới hạn* là giá trị phân chia giữa miền bác bỏ và miền không bác bỏ H_0.

Xác suất giá trị kiểm định thống kê sẽ rơi vào vùng được gọi là *mức ý nghĩa*, được ký hiệu là α. Nếu điều này xảy ra thì sai lầm loại I đã xảy ra. Nghĩa là, trong kiểm định giả thuyết, α là xác suất xảy ra sai lầm loại I.

Định nghĩa 7.6. *Mức ý nghĩa* của kiểm định giả thuyết là xác suất xảy ra sai lầm loại I.

Quy trình kiểm định giả thuyết đòi hỏi phải quyết định mức ý nghĩa α trước khi thu thập dữ liệu và tính giá trị kiểm định thống kê. Khi mức ý nghĩa được xác định thì vùng bác bỏ cũng sẽ được thiết lập. Nếu giá trị kiểm định thống kê rơi vào vùng bác bỏ thì ta sẽ bác bỏ H_0, ngược lại, ta không bác bỏ H_0.

Các dạng toán kiểm định thường gặp: Ta kí hiệu θ là một tham số của tổng thể.

Kiểm định 2 phía: (two-tailed test)

$$\text{Giả thuyết } H_0 : \theta = \theta_0 \text{ và đối thuyết } H_1 : \theta \neq \theta_0.$$

Kiểm định 1 phía trái: (left-tailed test)

$$\text{Giả thuyết } H_0 : \theta = \theta_0 \text{ và đối thuyết } H_1 : \theta < \theta_0.$$
$$\text{Giả thuyết } H_0 : \theta \geq \theta_0 \text{ và đối thuyết } H_1 : \theta < \theta_0.$$

Kiểm định 1 phía phải: (right-tailed test)

Giả thuyết $H_0 : \theta = \theta_0$ và đối thuyết $H_1 : \theta > \theta_0$.

Giả thuyết $H_0 : \theta \leq \theta_0$ và đối thuyết $H_1 : \theta > \theta_0$.

Ví dụ 7.7. 1. Mức độ chấp nhận được tối đa đối với việc tiếp xúc với bức xạ vi sóng ở Hoa Kỳ là trung bình 10 microwatt trên một cm^2. Người ta lo ngại rằng một truyền hình lớn có thể gây ô nhiễm không khí gần đó bằng cách đẩy mức bức xạ vi sóng lên trên giới hạn an toàn.

Ta kiểm định: Giả thuyết $H_0 : \mu \leq 10$ và đối thuyết $H_1 : \mu > 10$.

2. Một hệ thống máy tính hiện có 10 thiết bị đầu cuối và sử dụng một máy in duy nhất. Thời gian xử lý trung bình của hệ thống là 15 phút. Mười thiết bị đầu cuối mới và một máy in thứ hai được thêm vào hệ thống. Chúng ta muốn xác định xem thời gian xử lý trung bình có bị ảnh hưởng hay không.

Ta kiểm định: Giả thuyết $H_0 : \mu = 15$ và đối thuyết $H_1 : \mu \neq 15$.

3. Một công nhân sản xuất gạch thấy rằng số lượng gạch làm ra trong 1 giờ giảm khi áp dụng quy trình sản xuất mới. Trước đây, trung bình công nhân làm được 35 viên gạch trong một giờ.

Ta kiểm định: Giả thuyết $H_0 : \mu = 35$ và đối thuyết $H_1 : \mu < 35$.

4. Một nhân viên của một nhà hàng nói rằng thời gian trung bình khách phải chờ để được phục vụ của nhà hàng họ là không quá 10 phút.

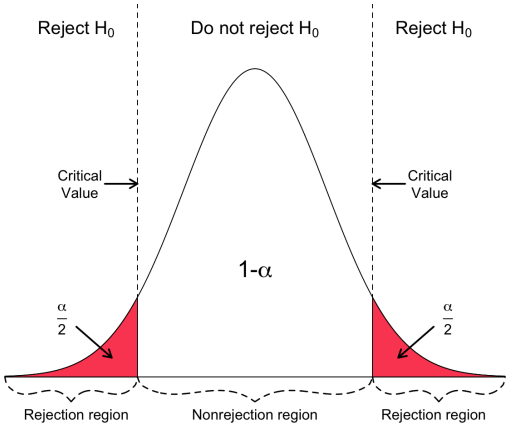
Ta kiểm định: Giả thuyết $H_0 : \mu \leq 10$ và đối thuyết $H_1 : \mu > 10$.

7.2 Kiểm định giả thuyết về trung bình

- Đặt μ là trung bình của tổng thể.
- Giả sử tổng thể có phân phối chuẩn $N(\mu; \sigma^2)$ hoặc kích thước mẫu $n \geq 30$.

Bài toán 1. Ta kiểm định

$$H_0 : \mu = \mu_0 \text{ và } H_1 : \mu \neq \mu_0.$$



Với mức ý nghĩa α, đặt $z_{\alpha/2}$ (được gọi là *giá trị tới hạn*) là giá trị thỏa mãn

$$P(|Z| > z_{\alpha/2}) = \alpha$$

hay

$$\boxed{\Phi(z_{\alpha/2}) = P(Z \leq z_{\alpha/2}) = 1 - \frac{\alpha}{2}}$$

Bài toán 2. Ta kiểm định

$$H_0 : \mu = \mu_0 \text{ và } H_1 : \mu > \mu_0.$$

Với mức ý nghĩa α, giá trị tới hạn z_α thỏa mãn

$$P(Z > z_\alpha) = \alpha$$

hay

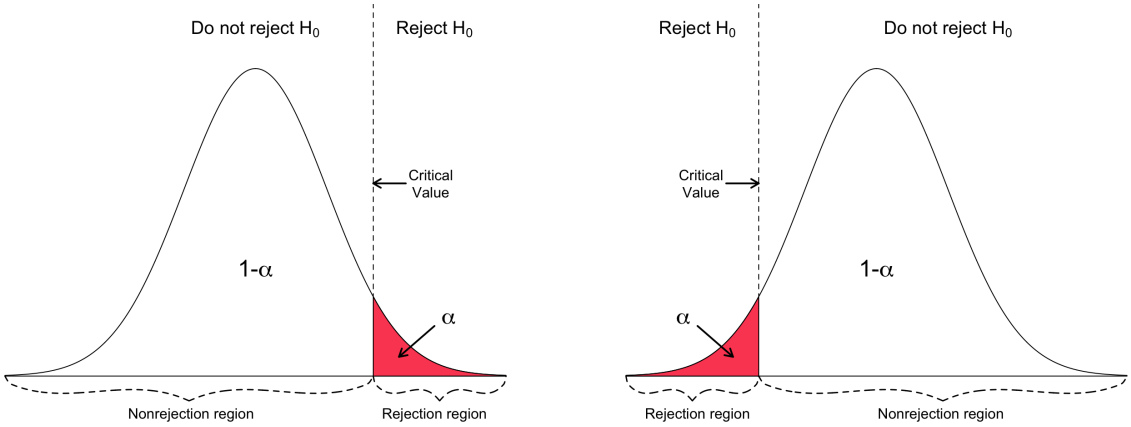
$$\boxed{\Phi(z_\alpha) = P(Z \leq z_\alpha) = 1 - \alpha}$$

Bài toán 3. Với mức ý nghĩa α, ta kiểm định:

$$H_0 : \mu = \mu_0 \text{ và } H_1 : \mu < \mu_0.$$

Đặt z_α là giá trị thỏa mãn

$$\boxed{\Phi(z_\alpha) = P(Z < z_\alpha) = \alpha}$$



Tính giá trị kiểm định z : Cho $\overline{x}$ là trung bình mẫu, n là kích thước mẫu, s là độ lệch chuẩn mẫu.

Trường hợp 1. σ đã biết.	**Trường hợp 2.** σ chưa biết và kích thước mẫu $n \geq 30$.
Tính $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	Tính $z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

Kiểm định	Bác bỏ H_0	Chấp nhận H_0
$H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0.$	$\lvert z\rvert \geq z_{\alpha/2}$	$\lvert z\rvert < z_{\alpha/2}$
$H_0 : \mu = \mu_0; H_1 : \mu > \mu_0.$	$z \geq z_\alpha$	$z < z_\alpha$
$H_0 : \mu = \mu_0; H_1 : \mu < \mu_0.$	$z \leq z_\alpha$	$z > z_\alpha$

Trường hợp 3: Với σ chưa biết và $n < 30$, tổng thể có phân phối chuẩn.

- Với mức ý nghĩa α, đặt $t_{\alpha/2,n-1}$ và $t_{\alpha,n-1}$ (các giá trị tới hạn) là các số thực thỏa mãn (xem bảng A5)

$$P(T > t_{\alpha,n-1}) = \alpha;\ P(T > t_{\alpha/2,n-1}) = \frac{\alpha}{2}.$$

- Tính giá trị kiểm định

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

trong đó $\overline{x}$ là trung bình mẫu cụ thể.

- Đưa ra quyết định (bác bỏ/không bác bỏ H_0)

Kiểm định	Bác bỏ H_0	Chấp nhận H_0
$H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0.$	$\lvert t\rvert \geq t_{\alpha/2}$	$\lvert t\rvert < t_{\alpha/2}$
$H_0 : \mu = \mu_0; H_1 : \mu > \mu_0.$	$t \geq t_\alpha$	$t < t_\alpha$
$H_0 : \mu = \mu_0; H_1 : \mu < \mu_0.$	$t \leq -t_\alpha$	$t > -t_\alpha$

Ví dụ 7.8. Theo báo cáo "Thị trường IT Việt Nam - Developers Recruitment State 2021" do TopDev công bố cho biết, tính đến quý II/2021, kỹ sư trí tuệ nhân tạo (AI) và máy học (Machine Learning) là vị trí có mức lương trung bình hàng tháng cao nhất trong các kỹ sư IT, đạt 3054 USD (khoảng 70 triệu đồng). Một cuộc khảo sát 30 kỹ sư trí tuệ nhân tạo tốt nghiệp từ một trường đại học X cho thấy họ có mức lương trung bình là 3105 USD/tháng. Hãy kiểm tra kết luận nói rằng các kỹ sư trí tuệ nhân tạo của trường X có mức thu nhập trung bình lớn hơn 3054 USD/tháng với mức ý nghĩa 0,05. Giả sử thu nhập của các kỹ sư trí tuệ nhân tạo có phân phối chuẩn với độ lệch chuẩn tổng thể là 120 USD.

Giải.

- Gọi μ thu nhập trung bình của các kỹ sư trí tuệ nhân tạo
- Ta kiểm định: Giả thuyết $H_0 : \mu = 3054$ và đối thuyết $H_1 : \mu > 3054$
- Theo đề bài, trung bình mẫu là $\overline{x} = 3105$, cỡ mẫu $n = 30$ và độ lệch chuẩn tổng thể $\sigma = 120$
- Vì mức ý nghĩa $\alpha = 0,05$ và kiểm định một phía bên phải nên giá trị tới hạn z_α thỏa $\Phi(z_\alpha) = 1 - \alpha = 0,95$. Suy ra $z_\alpha = 1,65$.
- Tính giá trị kiểm định

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{3105 - 3054}{120/\sqrt{30}} = 2,34.$$

- Vì $z = 2,34 > 1,65$ nên bác bỏ H_0.
- Ta có đủ bằng chứng để đồng ý với tuyên bố lương trung bình của các kỹ sư trí tuệ nhân tạo nhiều hơn 3054 USD/tháng.

Ví dụ 7.9. Một nhà nghiên cứu nói rằng trung bình giá tiền của một đôi giày thể thao nam là ít hơn 80 USD. Chọn ngẫu nhiên 36 đôi giày thể thao nam để khảo sát giá, ta được giá tiền trung bình của 36 đôi giày này là 75 USD. Giả sử giá giày có phân phối chuẩn với độ lệch chuẩn là 19,2 USD. Với mức ý nghĩa 10%, ta có đủ bằng chứng để ủng hộ tuyên bố của nhà nghiên cứu này không?

Giải.

- Gọi μ giá trung bình của một đôi giày thể thao nam.
- Ta điểm định: Giả thuyết $H_0 : \mu = 80$ và đối thuyết $H_1 : \mu < 80$
- Theo đề bài, trung bình mẫu là $\overline{x} = 75$, cỡ mẫu $n = 36$ và độ lệch chuẩn tổng thể $\sigma = 19,2$
- Vì mức ý nghĩa $\alpha = 0,1$ và kiểm định một phía bên trái nên giá trị tới hạn z_α thỏa mãn $\Phi(z_\alpha) = \alpha = 0,1$. Suy ra $z_\alpha = -1,28$.
- Tính giá trị kiểm định

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{75 - 80}{19,2/\sqrt{36}} = -1,56.$$

- Vì $z = -1,56 < -1,28$ nên bác bỏ H_0.
- Ta có đủ bằng chứng để đồng ý với tuyên bố giá tiền trung bình của một đôi giày thể thao nam ít hơn 80 USD.

Ví dụ 7.10. Một nhà nghiên cứu nói rằng trung bình giá tiền của một đôi giày thể thao nam là ít hơn 80 USD. Chọn ngẫu nhiên 16 đôi giày thể thao nam để khảo sát giá, ta được kết quả sau (USD/đôi)

60 70 75 55 80 55 50 40 70 50 95 120 90 75 85 80

Giả sử giá giày có phân phối chuẩn. Với mức ý nghĩa 10%, ta có đủ bằng chứng để ủng hộ tuyên bố của nhà nghiên cứu này không?

Giải.

- Gọi μ giá trung bình của một đôi giày thể thao nam.
- Ta điểm định: Giả thuyết $H_0 : \mu = 80$ và đối thuyết $H_1 : \mu < 80$
- Theo đề bài, trung bình mẫu là $\overline{x} = 71,875$, cỡ mẫu $n = 16$ và độ lệch chuẩn mẫu $s = ...$
- Vì mức ý nghĩa $\alpha = 0,1$ và kiểm định một phía bên trái nên giá trị tới hạn $t_\alpha = ...$ (bậc tự do 15).
- Tính giá trị kiểm định

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{.....}{...../\sqrt{....}} = ...$$

- Vì ..... nên .....
- .....

Ví dụ 7.11. Hóa đơn tiền điện trung bình đối với các đơn vị sản xuất tại Thành phố Hồ Chí minh dự kiến sẽ tăng trong năm năm 2025 so với năm 2024. Một phần là do nhu cầu và giá điện tăng. Hóa đơn tiền điện trung bình của các đơn vị sản xuất năm 2024 ở Thành phố Hồ Chí Minh ước tính là 10,38 triệu đồng. Một mẫu ngẫu nhiên gồm 12 đơn vị sản xuất đã được chọn và hóa đơn tiền điện tháng 3 năm 2025 đã được thu thập với số tiền trung bình phải trả là 11,72 triệu và độ lệch chuẩn là 2,05 triệu. Có bằng chứng nào cho thấy hoá đơn tiền điện sản xuất trung bình thực sự trong tháng 3 năm 2025 ở Thành phố Hồ Chí Minh lớn hơn 10,38 triệu không? Giả sử tiền điện có phân phối chuẩn và mức ý nghĩa 5%.

Giải.

- Gọi μ tiền điện sản xuất trung bình thực sự trong tháng 3 năm 2025
- Kiểm định: Giả thuyết $H_0 : \mu =$ ..... và đối thuyết $H_1 : \mu$.....
- Trung bình mẫu là ..... cỡ mẫu $n =$ ..... và độ lệch chuẩn mẫu là .....
- Mức ý nghĩa ..... suy ra giá trị tới hạn..... (xem bảng A5)
- Giá trị kiểm định thống kê

  .....

  .....
- Vì ..... nên ..... H_0.
- dotfill .....

  .....

7.3 Kiểm định giả thuyết về tỉ lệ

Bài toán. Một người ăn kiêng nói rằng có 60% số người không ăn bánh ngọt. Một cuộc khảo sát 200 người, ta thấy có 128 người nói rằng họ không ăn bánh ngọt. Với mức ý nghĩa 5%, ta có đủ bằng chứng để bác bỏ tuyên bố của người ăn kiêng này không?

- Gọi p (tương ứng $\hat{p}, \hat{P}$) là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong tổng thể (trong mẫu cụ thể, mẫu ngẫu nhiên).
- Kiểm định giả thuyết
$$H_0 : p = p_0$$
- Chọn một mẫu ngẫu nhiên có kích thước n. Ta biết rằng nếu $np \geq 5$ và $n(1-p) \geq 5$ thì phân phối lấy mẫu của F xấp xỉ phân phối chuẩn.
- Với mức ý nghĩa α, đặt các giá trị tới hạn là z_α (đối với kiểm định 1 phía), $z_{\alpha/2}$ (đối với kiểm định 2 phía).

- Tính giá trị kiểm định thống kê

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Với mức ý nghĩa α.

Kiểm định	Bác bỏ H_0	Chấp nhận H_0
$H_0 : p = p_0; H_1 : p \neq p_0.$	$\lvert z\rvert \geq z_{\alpha/2}$	$\lvert z\rvert < z_{\alpha/2}$
$H_0 : p = p_0; H_1 : p > p_0.$	$z \geq z_\alpha$	$z < z_\alpha$
$H_0 : p = p_0; H_1 : p < p_0.$	$z \leq z_\alpha$	$z > z_\alpha$

Giải Bài toán 4

- Gọi p là tỉ lệ người không ăn bánh ngọt.
- Ta kiểm định: Giả thuyết $H_0 : p = 60\%$ và đối thuyết $H_1 : p \neq 60\%$
- Tỉ lệ mẫu là $\hat{p} = \dfrac{128}{200} = 0,64$ và cỡ mẫu $n = 200$.
- Vì mức ý nghĩa $\alpha = 0,05$ và kiểm định hai phía nên giá trị tới hạn $z_{\alpha/2} = 1,96$. $(\Phi(z_{\alpha/2}) = 1 - \dfrac{\alpha}{2})$
- Tính giá trị kiểm định

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0,64 - 0,6}{\sqrt{0,6(1-0,6)/200}} = 1,15.$$

- Vì $\lvert z\rvert = 1,15 < 1,96$ nên không bác bỏ H_0.
- Ta không có đủ bằng chứng để bác bỏ phát biểu rằng có 60% người không ăn bánh ngọt.

Ví dụ 7.12. Một lập trình viên nói rằng có hơn 25% các lập trình viên đã học ngôn ngữ lập trình Python. Một cuộc khảo sát 200 lập trình viên tại một thành phố nọ, người ta thấy có 63 lập trình viên đã học Python. Với mức ý nghĩa 5%, hãy kết luận về nhận định của lập trình viên trên.

Giải. .....

.....

.....

.....

.....

.....

.....

.....

.....

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Bài tập

Bài tập 7.1. Phòng hành chính của một bệnh viện tuyên bố rằng thời gian chờ trung bình của bệnh nhân để được điều trị tại khoa cấp cứu là 25 phút. Một mẫu ngẫu nhiên gồm 16 bệnh nhân được điều trị tại khoa cấp cứu của bệnh viện này đã đưa ra thời gian chờ trung bình là 27,5 phút với độ lệch chuẩn là 4,8 phút. Sử dụng mức ý nghĩa 1%, hãy kiểm tra xem thời gian chờ trung bình tại khoa cấp cứu có khác 25 phút không. Giả sử rằng thời gian chờ của tất cả bệnh nhân tại khoa cấp cứu này có phân phối chuẩn.

Bài tập 7.2. Giả sử thời gian xem tivi mỗi ngày của một người có phân phối chuẩn với trung bình là 151 phút và độ lệch chuẩn là 30 phút vào năm 2011. Năm 2024, một mẫu gồm 120 người cho thấy họ xem tivi mỗi ngày trung bình 162 phút. Với mức ý nghĩa 10%, ta có đủ bằng chứng để khẳng định rằng thời gian xem tivi trung bình của một người năm 2024 nhiều hơn 151 phút không?

Bài tập 7.3. Một nhân viên bán hàng tại một cửa hàng laptop nói rằng tuổi thọ trung bình của laptop hiệu Z không quá 30500 giờ. Một cuộc khảo sát 40 laptop hiệu Z cho thấy tuổi thọ trung bình của chúng là 30456 giờ. Biết rằng độ lệch chuẩn tổng thể là 1684 giờ. Với mức ý nghĩa 10%, phát biểu của nhân viên bán hàng có chấp nhận được không?

Bài tập 7.4. Một công ty máy tính mới đây đã giới thiệu một sản phẩm phần mềm mới tuyên bố rằng thời gian trung bình để học cách sử dụng phần mềm này không quá 2 giờ đối với những người có phần quen thuộc với máy tính. Một mẫu ngẫu nhiên gồm 12 người như vậy đã được chọn. Dữ liệu sau đây cho biết thời gian (tính bằng giờ) mà những người này mất để học cách sử dụng phần mềm này.

1,75 2,25 2,40 1,90 1,50 2,75 2,15 2,25 1,80 2,20 3,25 2,60

Kiểm tra ở mức ý nghĩa 1% xem tuyên bố của công ty có đúng không. Giả sử rằng thời gian mà tất cả những người có phần quen thuộc với máy tính dành để học cách sử dụng phần mềm này có phân phối chuẩn.

Bài tập 7.5. Hiệu trưởng một trường ở TPHCM tin rằng chỉ số IQ trung bình của học sinh trường này thấp hơn 105. Theo kinh nghiệm trước đây, chỉ số IQ có phân phối chuẩn với độ lệch chuẩn là 10. Một mẫu ngẫu nhiên gồm 10 học sinh được chọn từ trường này và chỉ số IQ của họ được quan sát như sau

95 91 110 93 133 119 113 107 110 89

Với mức ý nghĩa $\alpha = 0,01$, ta có đủ bằng chứng để ủng hộ phát biểu của hiệu trưởng không?

Bài tập 7.6. Một quy trình sản xuất các chai dầu gội đầu, khi vận hành chính xác, sẽ tạo ra các chai có trọng lượng trung bình là 200 gam. Một mẫu ngẫu nhiên gồm chín chai từ một lần sản xuất duy nhất mang lại các trọng lượng hàm lượng sau (tính bằng gam):

201,4 109,7 109,7 200,6 200,8 200,1 190,7 200,3 200,9

Giả sử rằng trọng lượng của các chai dầu gội này có phân phối chuẩn. Hãy kiểm tra ở mức 5% đối với giả thiết rằng quy trình đang vận hành chính xác.

Bài tập 7.7. Một báo cáo cho thấy rằng có ít hơn 78% sinh viên sử dụng Google Translate khi đọc các trang web bằng tiếng Anh. Chọn ngẫu nhiên 143 sinh viên tại một trường đại học và người ta thấy có 100 sinh viên sử dụng Google Translate khi đọc các trang web tiếng Anh. Với mức ý nghĩa 5%, ta có đủ bằng chứng để ủng hộ báo cáo trên không?

Bài tập 7.8. Một mẫu gồm 202 giảng viên kinh tế được hỏi liệu có nên có môn học ngoại ngữ chuyên ngành cho sinh viên chuyên ngành kinh doanh hay không. Trong số các giảng viên này, 140 người cảm thấy cần có khóa học ngoại ngữ. Kiểm tra giả thuyết rằng có 75% các giảng viên kinh tế đều có quan điểm này với mức ý nghĩa 5%.

Bài tập 7.9. Một công ty cung cấp dịch vụ internet nói rằng có hơn 35% khách hàng của họ gặp sự cố về đường truyền trong một năm. Một nhóm gồm 120 khách hàng được chọn và người ta thấy rằng có 45 khách hàng gặp sự cố về đường truyền. Với mức ý nghĩa 1%, ta có đủ bằng chứng để ủng hộ tuyên bố của công ty cung cấp dịch vụ internet không?

Chương 8. Hồi quy tuyến tính

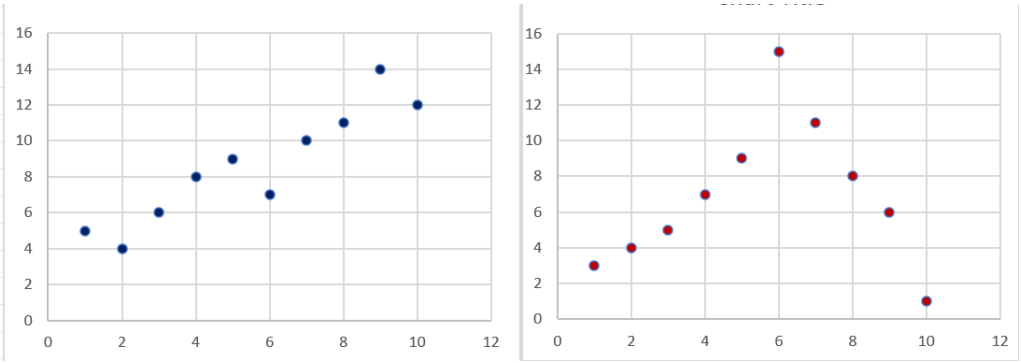
Nguyễn Minh Trí

Ngày 1 tháng 4 năm 2025

Mục lục

8.1 Hệ số tương quan

Cho hai biến (X, Y) lần lượt nhận các giá trị $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Biểu diễn các điểm giá trị trên mặt phẳng, ta được một biểu đồ mà ta gọi là **biểu đồ phân tán** (scatter diagram). Mỗi điểm (x_i, y_i) được gọi là một điểm dữ liệu.



Nhận xét:

- Hình bên trái: Các điểm dữ liệu được phân bố xung quanh một đường thẳng.
- Hình bên phải: Các điểm dữ liệu không được phân bố xung quanh một đường thẳng.

Ta nói hai biến của hình bên trái có quan hệ tuyến tính mạnh.

Định nghĩa 8.1. Số đo xác định độ mạnh của quan hệ tuyến tính giữa hai biến được gọi là hệ số tương quan (correlation coefficient).

- Hệ số tương quan của hai biến X, Y được xác định bởi

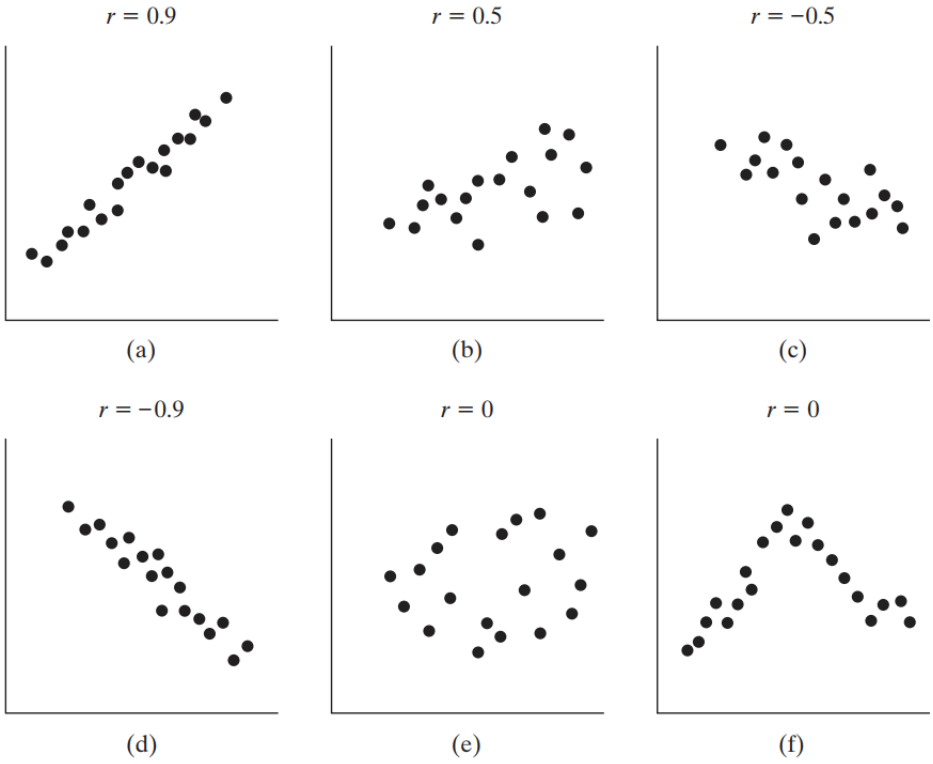
$$r = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2)(\overline{y^2} - \overline{y}^2)}}$$

hay

$$r = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i).(\sum_{i=1}^{n} y_i)}{\sqrt{(n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2)(n(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2)}}$$

trong đó n là số cặp điểm dữ liệu.

- Ta có $-1 \leq r \leq 1$.
- Nếu $0,8 \leq |r| \leq 1$ thì ta nói X, Y có tương quan tuyến tính mạnh.
- Nếu $|r| < 0,8$ thì ta nói X, Y có tương quan tuyến tính yếu.
- Nếu r gần bằng 1 thì ta nói có sự tương quan tuyến tính thuận giữa X và Y, tức là nếu X tăng thì Y tăng.
- Nếu r gần bằng -1 thì ta nói có sự tương quan tuyến tính nghịch giữa X và Y, tức là nếu X tăng thì Y giảm.



Hệ số tương quan của các điểm dữ liệu.[1]

Ví dụ 8.2. Điểm số môn Xác suất thống kê và số buổi vắng của 7 sinh viên được cho bên dưới

Số buổi vắng (X)	6	2	15	9	12	5	8
Điểm (Y)	8,2	8,6	4,3	7,4	5,8	9,0	7,8

Tìm hệ số tương quan giữa số buổi nghỉ học và điểm môn Xác suất thống kê.

[1]Nguồn ảnh: "Richard A. Johnson (2018). *Miller And Freund's Probability And Statistics For Engineers, 9th edition*, Pearson."

Giải. Ta có

$$\overline{xy} = \frac{6.8,2 + 2.8,6 + 15.4,3 + 9.7,4 + 12.5,8 + 5.9,0 + 8.7,8}{7} = 53,5$$
$$\overline{x} = \frac{6 + 2 + 15 + 9 + 12 + 5 + 8}{7} = 8,14$$
$$\overline{y} = \frac{8,2 + 8,6 + 4,3 + 7,4 + 5,8 + 9,0 + 7,8}{7} = 7,3$$
$$\overline{x^2} = \frac{6^2 + 2^2 + 15^2 + 9^2 + 12^2 + 5^2 + 8^2}{7} = 82,71$$
$$\overline{y^2} = \frac{8,2^2 + 8,6^2 + 4,3^2 + 7,4^2 + 5,8^2 + 9,0^2 + 7,8^2}{7} = 55,7$$

Do đó, hệ số tương quan là

$$r = \frac{53,5 - 8,14.7,3}{\sqrt{(82,71 - 8,14^2)(55,7 - 7,3^2)}} = \frac{-5,992}{6,296} = -0,9517.$$

Có một sự tương quan tuyến tính mạnh giữa số buổi vắng và số điểm. Nếu số buổi vắng càng nhiều thì số điểm càng thấp.

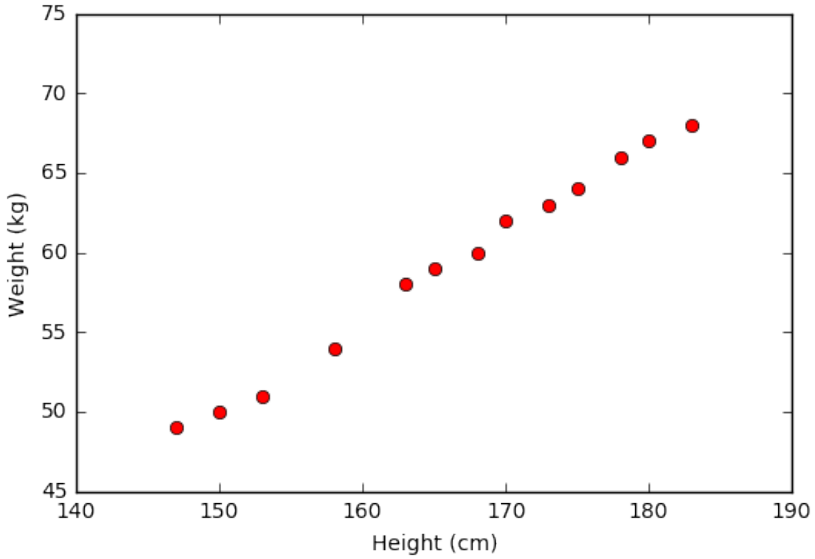
8.2 Hồi quy tuyến tính đơn

Bài toán. Bảng dữ liệu về chiều cao và cân nặng của 15 người:

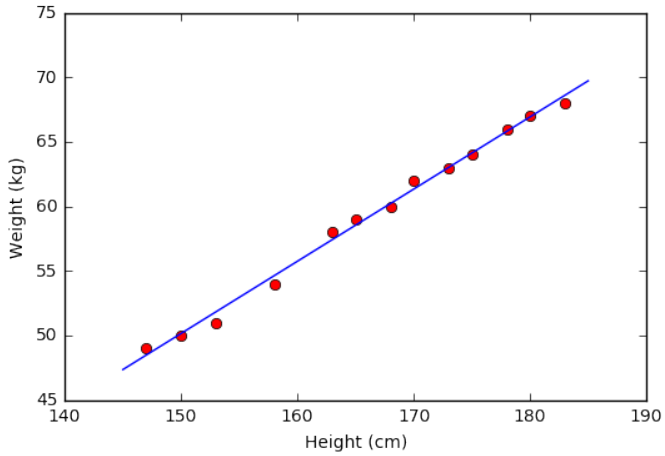
Chiều cao (cm)	Cân nặng (kg)	Chiều cao (cm)	Cân nặng (kg)
147	49	168	60
150	50	170	72
153	51	173	63
155	52	175	64
158	54	178	66
160	56	180	67
163	58	183	68
165	59		

Có thể dự đoán cân nặng của một người dựa vào chiều cao của họ không?

Biểu diễn các dữ liệu trên dưới dạng đồ thị như sau



Ta thấy rằng các điểm dữ liệu không nằm trên một đường thẳng nhưng chúng có thể được phân bố xung quanh một đường thẳng.



- Các điểm dữ liệu nằm khá gần đường thẳng (phương trình $y = a + bx$).
- Ta có thể đưa ra mối liên hệ giữa cân nặng và chiều cao như sau

$$\text{cân nặng} = b \times \text{chiều cao} + a.$$

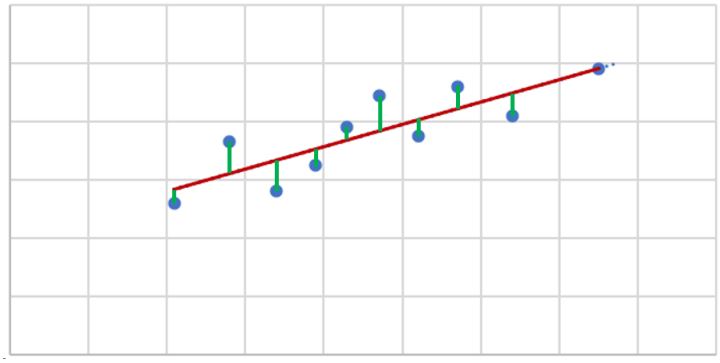
- Bằng các công cụ tính toán, chúng ta sẽ tính được a, b. Khi đó đường thẳng có phương trình $y = a + bx$ được gọi là đường thẳng hồi quy (regression line).
- Sử dụng mô hình này, ta có thể dự đoán cân nặng của một người có chiều cao 155cm, 160 cm hoặc 171cm.
- Mô hình trên là mô hình *hồi quy tuyến tính đơn*.

Bài toán. Xây dựng một mô hình toán học hay một hàm số mà có thể dùng để dự đoán giá trị của một biến dựa vào một hay một số biến được gọi là phân tích hồi quy (regression analysis).

- Mô hình hồi quy đơn giản nhất được gọi là hồi quy đơn (simple regression) liên quan đến hai biến trong đó một biến được dự đoán dựa vào một biến khác.
- Trong hồi quy đơn, biến được dự đoán được gọi là biến phụ thuộc (dependent variable) và ký hiệu là y. Biến mà ta dùng để dự đoán biến y được gọi là biến độc lập (independent variable/regressor variable/predictor variable) và kí hiệu là x.
- Giả sử mối quan hệ giữa biến độc lập và biến phụ thuộc được cho bởi một hàm số $y = f(x)$ trong đó f là hàm số mà ta chưa biết.
- Nếu $f(x) = a + bx$ thì đường thẳng được xác định bởi $y = a + bx$ được gọi là đường thẳng hồi quy.
- Dùng phương pháp bình phương tối tiểu (method of least squares) để tìm các giá trị a và b.
- Giả sử ta có một mẫu gồm n điểm dữ liệu được xác định bởi các cặp giá trị $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

- Tìm một mô hình đường thẳng
$$y = a + bx$$
sao cho các điểm dữ liệu là "gần nhất" với đường thẳng này.
- Tập hợp các điểm (x_i, y_i) được biểu diễn trên mặt phẳng tọa độ được gọi là **biểu đồ phân tán** (Scatter diagram).
- Từ biểu đồ phân tán sẽ cho ta dự đoán được hàm số $f(x)$.

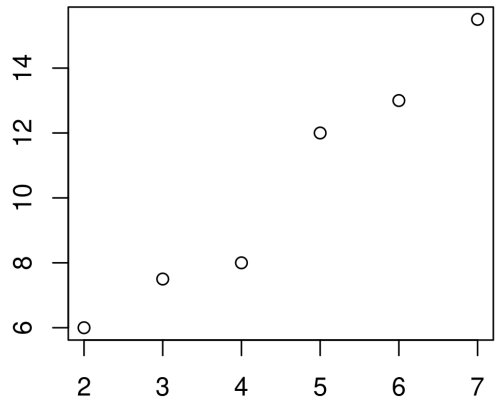
Với mỗi điểm (x_i, y_i), ta có
$$\hat{y}_i = a + bx_i$$
và
$$y_i - \hat{y}_i = y_i - (a + bx_i)$$



- Khi đó tổng bình phương của các độ lệch giữa giá trị của y_i và $\hat{y}_i$ được kí hiệu
$$L = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$
- Đường thẳng $y = a + bx$ được gọi là "gần nhất" với các điểm dữ liệu đã cho nếu L có giá trị nhỏ nhất.
- Khi đó
$$b = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2} \text{ và } a = \overline{y} - b\overline{x}.$$

Ví dụ 8.3. Tìm đường thẳng hồi quy biểu thị mối liên hệ giữa số tiền lương làm theo giờ y (trăm nghìn đồng) và số năm kinh nghiệm x dựa theo bảng dữ liệu sau

x	2	3	4	5	6	7
y	6	7,5	8	12	13	15,5



Giải. Dựa vào dữ liệu, ta tính được các giá trị

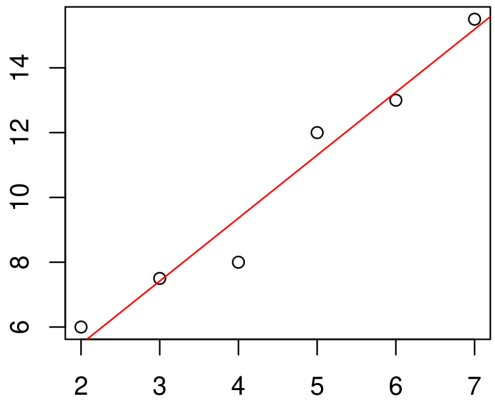
$$\overline{x} = 4,5 \quad \overline{y} = 10,33 \quad b = 1,943$$

và

$$a = \overline{y} - b\overline{x} = 10,33 - 1,943.4,5 = 1,5865$$

Như vậy, phương trình đường thẳng hồi quy tuyến tính là

$$y = 1,5865 + 1,943x$$



Dùng Microsoft Excel để tìm đường thẳng hồi quy

- Tạo bảng dữ liệu trong Microsoft Excel
- Tạo biểu đồ phân tán: Chọn bảng dữ liệu → **Insert** → **Charts** → **All Charts** → **X Y (Scatter)** → **OK**
- Tạo đường thẳng hồi quy: Nhấp vào ⊞ bên góc phải của Chart vừa hiện ra → **Chart Elements**, chọn **Trendline**
- Hiện phương trình đường thẳng hồi quy: Bên cạnh **Trendline** → ▶ **More Options**
- Trong bảng **Format Trendline**, chọn  , kéo xuống bên dưới và chọn **Display Equation on chart**.

Ví dụ 8.4. Điểm số môn Xác suất thống kê và số buổi vắng của 7 sinh viên được cho bên dưới

Số buổi vắng (x)	6	2	15	9	12	5	8
Điểm (y)	8,2	8,6	4,3	7,4	5,8	9,0	7,8

a. Tìm phương trình đường thẳng hồi quy tuyến tính thể hiện mối liên hệ giữa số điểm số và số buổi vắng học.

b. Dự đoán số điểm số của sinh viên chỉ vắng 1 buổi học.

Giải. a. Dựa vào dữ liệu, ta tính được các giá trị

$$\overline{x} = 8,142857 \quad \overline{y} = 7,3 \quad b = -0,3622$$

và

$$a = \overline{y} - b\overline{x} = 7,3 - (-0,3622).8,142857 = 10,2493$$

Như vậy, phương trình đường thẳng hồi quy tuyến tính là

$$y = 10,2493 - 0,3622x$$

b. Nếu $x = 1$ thì $y = 9,8871$. Do đó nếu sinh viên vắng một buổi học thì điểm số của sinh viên có thể đạt được là 9,8871 điểm.

Ví dụ 8.5. Bảng khảo sát doanh thu bán hàng online Y và chi phí quảng cáo online X (trong 15 phút) của 7 cửa hàng được cho như sau: Đơn vị tính là trăm nghìn đồng

Doanh số bán hàng	368	340	665	954	331	556	376
Chi phí quảng cáo	1,7	1,5	2,8	5	1,3	2,2	1,3

a. Tính hệ số tương quan và nhận xét về tính tuyến tính của X và Y (mạnh hay yếu).

b. Viết phương trình hồi quy tuyến tính của Y theo X. Dự đoán doanh số bán hàng khi chi phí quảng cáo online trong 15 phút là 4 trăm nghìn đồng.

Giải. a. Hệ số tương quan

$$r = 0,9804402$$

Doanh số bán hàng và chi phí quảng cáo có tương quan tuyến tính mạnh.

b. Đặt x là chi phí quảng cáo và y doanh thu bán hàng. Ta tính được

$$\overline{y} = 512,8571; \quad \overline{x} = 2,257143; \text{ và } b = 171,5$$

và

$$a = \overline{y} - b\overline{x} = 125,8$$

Phương trình đường thẳng hồi quy tuyến tính

$$y = 125,8 + 171,5x.$$

Như vậy, khi $x = 4$ thì $y = 811,8$. Tức là nếu chi phí quảng cáo trong 15 phút là 4 trăm nghìn đồng thì doanh thu bán hàng đạt được là 81 180 000 đồng.

Ví dụ 8.6. Quan sát ở một mẫu, ta có bảng thống kê sau (X: chiều cao; Y: trọng lượng)

Y \ X	$150-155$	$155-160$	$160-165$	$165-170$	$170-175$
50	5				
55	2	11			
60		3	15	4	
65			8	17	
70			10	6	7
75					12

Lập phương trình đường thẳng hồi quy tuyến tính của Y theo X.

Giải. .....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Bài tập

Bài tập 8.1. Cho hai biến x, y có các giá trị tương ứng như sau

x	2,4	2,7	5,6	2,6	2,1	3,3	6,6	5,7
y	25,3	14,3	151,6	91,1	80	49	173	95,8

Hai biến x, y có quan hệ tuyến tính không?

Bài tập 8.2. Lợi nhuận của 7 công ty cho thuê xe là Y (tỉ USD) trong 1 năm và số lượng xe cho thuê X (nghìn chiếc) được cho như sau

Công ty	Số xe (X)	Lợi nhuận (Y) (tỉ USD)
A	630	7
B	290	3,9
C	208	2,1
D	191	2,8
E	134	1,4
F	85	1,5

a. Tính hệ số tương quan giữa số xe cho thuê và lợi nhuận hàng năm.

b. Viết phương trình hồi quy tuyến tính của Y theo X. Dự đoán lợi nhuận trong một năm của một công ty có 200 000 xe cho thuê.

Bài tập 8.3. Thời gian sử dụng liên tục của 8 loại điện thoại Y (giờ) và số mAh X ghi trên pin của điện thoại được khảo sát như sau

Điện thoai	Số mAh (X)	Thời gian sử dụng (Y) (giờ)
A	2800	3,8
B	3000	3,9
C	3700	4,2
D	4000	3,8
E	4300	4,1
F	5000	5
G	5000	4.8
H	6000	4,9

a. Tính hệ số tương quan giữa số mAh trên pin và thời gian sử dụng.

b. Viết phương trình hồi quy tuyến tính của Y theo X. Dự đoán thời gian sử dụng của một loại pin điện thoại có 6550 mAh.

Bài tập 8.4. Quan sát ở một mẫu, ta có bảng thống kê sau (X: đường kính (cm); Y: chiều cao (m))

Y \ X	$20-22$	$22-24$	$24-26$	$26-28$	$28-30$
3	2				
4	5	3			
5		11	8	4	
6			15	17	
7			10	6	7
8					12

Lập phương trình đường thẳng hồi quy tuyến tính của Y theo X.