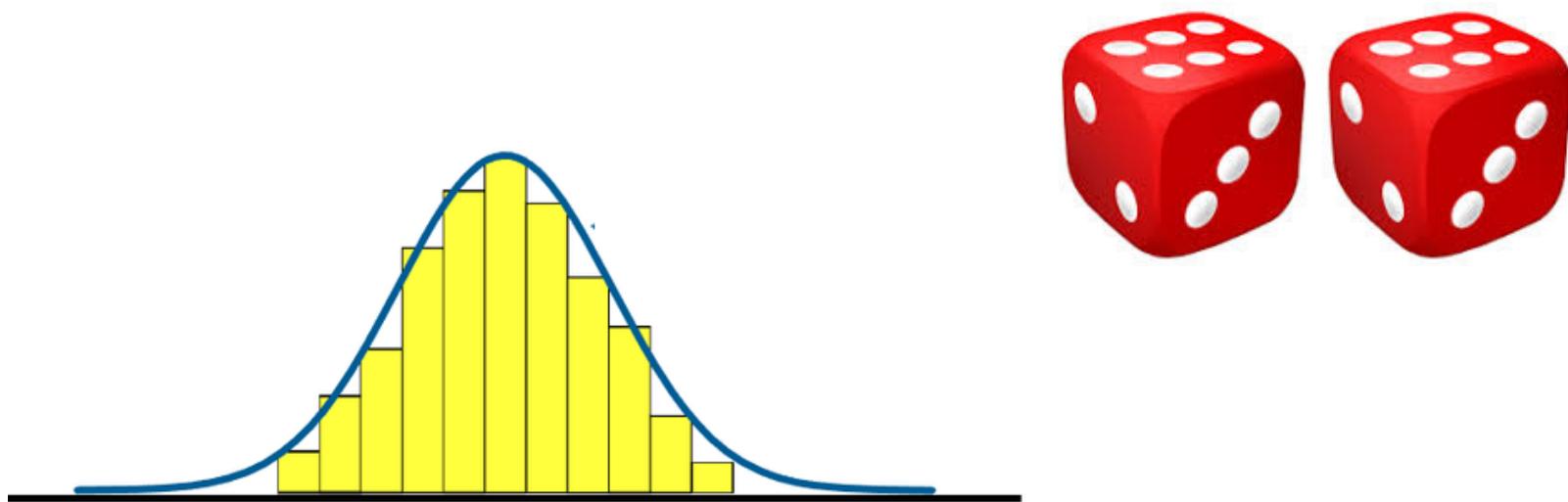
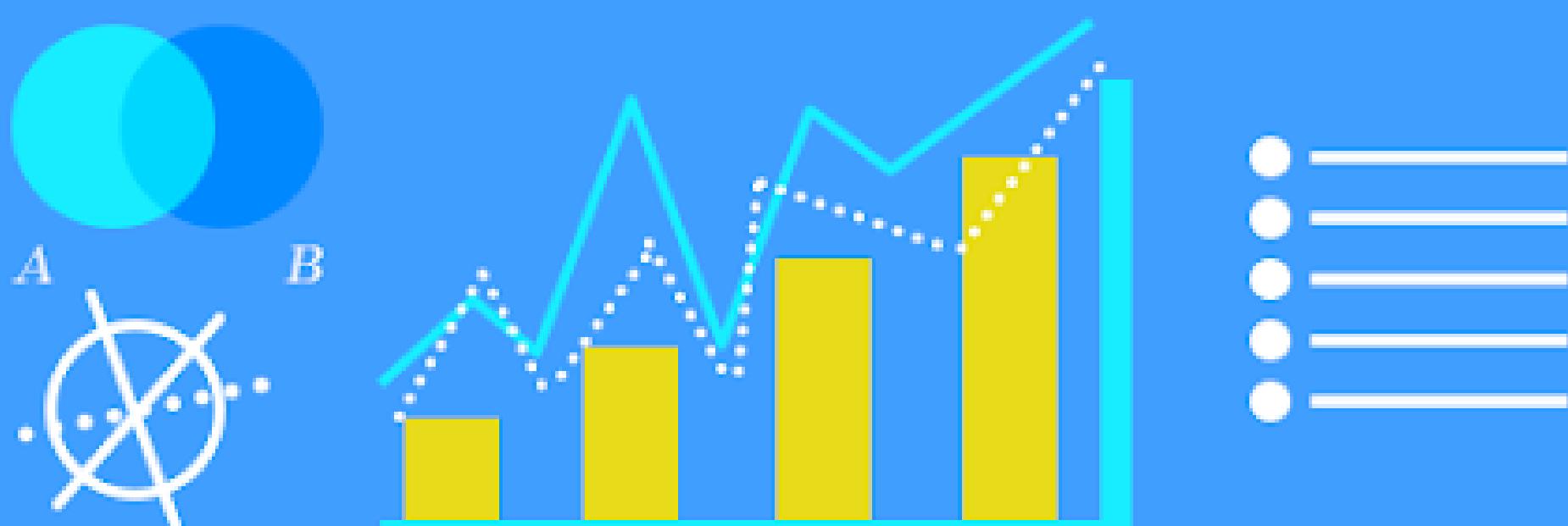
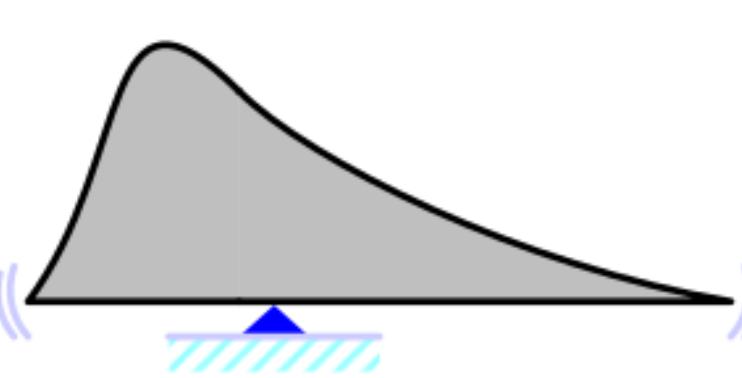
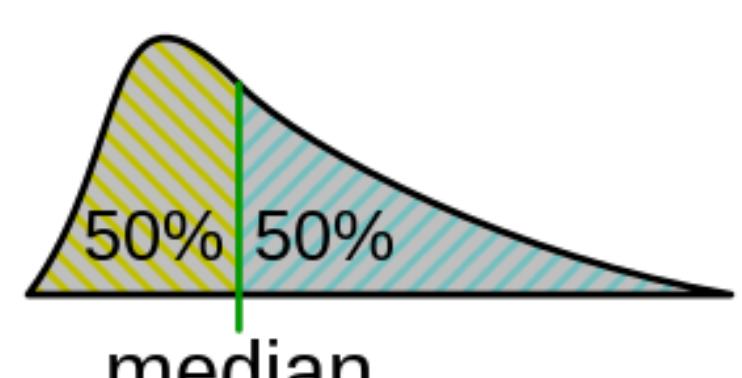
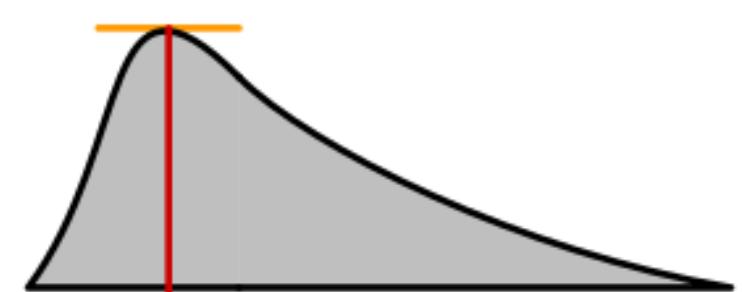


STATISTICS AND PROBABILITY



TÀI LIỆU HỌC TẬP

XÁC SUẤT THỐNG KÊ



Lớp.....

Họ và tên:

MSSV:



Chương 4. Vectơ ngẫu nhiên

Nguyễn Minh Trí

Trường Đại học Công nghệ Thông tin

Ngày 15 tháng 4 năm 2023

4.1 Vectơ ngẫu nhiên rời rạc

Dịnh nghĩa 4.1

- Cho X, Y là các biến ngẫu nhiên. Khi đó (X, Y) được gọi là một **vectơ ngẫu nhiên**.
- Nếu X, Y là các biến ngẫu nhiên rời rạc (liên tục) thì (X, Y) được gọi là **vectơ ngẫu nhiên rời rạc (liên tục)**.
- Cho X, Y là các biến ngẫu nhiên rời rạc. Xác suất đồng thời của X, Y (joint probability) được xác định bởi

$$P(x, y) = P(X = x, Y = y).$$

Bảng phân phối xác suất đồng thời

$X \backslash Y$	y_1	y_2	\dots	y_m
x_1	p_{11}	p_{12}	\dots	p_{1m}
x_2	p_{21}	p_{22}	\dots	p_{2m}
\dots	\dots	\dots	\dots	\dots
x_n	p_{n1}	p_{n2}	\dots	p_{nm}

trong đó $p_{ij} = P(X = x_i, Y = y_j)$.

4. Phân phối của từng biến X, Y được gọi là **phân phối xác suất thành phần** (marginal probability distributions) được xác định như sau

$$P_X(x_i) = P(X = x_i) = \sum_{j=1}^m P(x_i, y_j)$$

$$P_Y(y_j) = P(Y = y_j) = \sum_{i=1}^n P(x_i, y_j)$$

Bảng phân phối xác suất của X

X	x_1	x_2	\cdots	x_n
$P(X = x_i)$	p_{1*}	p_{2*}	\cdots	p_{n*}

trong đó $p_{i*} = p_{i1} + p_{i2} + \cdots + p_{im}$.

Bảng phân phối xác suất của Y

Y	y_1	y_2	\cdots	y_m
$P(Y = y_j)$	p_{*1}	p_{*2}	\cdots	p_{*m}

trong đó $p_{*j} = p_{1j} + p_{2j} + \cdots + p_{nj}$.

Kỳ vọng (trung bình) thành phần của X, Y

$$E(X) = \sum_{i=1}^n x_i p_{i*}, \quad E(Y) = \sum_{j=1}^m y_j p_{*j}.$$

Định nghĩa 4.2 Cho X, Y là các biến ngẫu nhiên rời rạc. Xác suất của X khi đã biết $Y = y_j, P_Y(y_j) > 0$:

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P_Y(y_j)}.$$

Xác suất của Y khi đã biết $X = x_i, P_X(x_i) > 0$:

$$P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P_X(x_i)}.$$

Bảng phân phối xác suất của X với điều kiện $Y = y_j$

X	x_1	x_2	\cdots	x_n
$P(X = x_i Y = y_j)$	$\frac{p_{1j}}{p_{*j}}$	$\frac{p_{2j}}{p_{*j}}$	\cdots	$\frac{p_{nj}}{p_{*j}}$

Kỳ vọng của X với điều kiện $Y = y_j$

$$E(X|Y = y_j) = x_1 \frac{p_{1j}}{p_{*j}} + x_2 \frac{p_{2j}}{p_{*j}} + \cdots + x_n \frac{p_{nj}}{p_{*j}}.$$

Định nghĩa 4.3 Hai biến ngẫu nhiên rời rạc X, Y là **độc lập** nếu

$$P(x, y) = P_X(x)P_Y(y)$$

với mọi x, y .

Ví dụ 4.4 Gọi X là số lần một máy PC gấp trực trặc: 1, 2 hoặc 3 lần vào bất kỳ ngày nào. Gọi Y là số lần một kỹ thuật viên được gọi đến sửa. Phân phối xác suất đồng thời của X, Y như sau

\backslash	X	1	2	3
Y				
1	0,05	0,05	0,1	
2	0,05	0,1	0,35	
3	0	0,2	0,1	

a. Tìm phân phối thành phần của X, Y

b. Tính $P(Y = 3|X = 2)$.

Giải. a. Tìm phân phối thành phần của X .

$$\begin{aligned} P(X = 1) &= P(X = 1, Y = 1) + P(X = 1, Y = 2) + P(X = 1, Y = 3) \\ &= 0,05 + 0,05 + 0 = 0,1 \end{aligned}$$

$$\begin{aligned} P(X = 2) &= P(X = 2, Y = 1) + P(X = 2, Y = 2) + P(X = 2, Y = 3) \\ &= 0,05 + 0,1 + 0,2 = 0,35 \end{aligned}$$

$$\begin{aligned} P(X = 3) &= P(X = 3, Y = 1) + P(X = 3, Y = 2) + P(X = 3, Y = 3) \\ &= 0,1 + 0,35 + 0,1 = 0,55 \end{aligned}$$

Bảng phân phối xác suất của X là

X	1	2	3
$P(X = x)$	0,1	0,35	0,55

Tương tự, bảng phân phối xác suất của Y là

Y	1	2	3
$P(Y = y)$	0,2	0,5	0,3

b. Tính $P(Y = 3|X = 2)$.

$$P(Y = 3|X = 2) = \frac{P(Y = 3, X = 2)}{P(X = 2)} = \frac{0,2}{0,35} = 0,5714$$

Ví dụ 4.5 Một chương trình bao gồm hai mô-đun. Đặt X là số lỗi trong mô-đun 1 và Y là số lỗi trong mô-đun 2 có xác suất đồng thời như sau $P(0,0) = P(0,1) = P(1,0) = 0,2; P(1,1) = P(1,2) = P(1,3) = 0,1; P(0,2) = P(0,3) = 0,05$.

- Tìm phân phối xác suất thành phần của X .
- Tìm phân phối của tổng số lỗi trong chương trình.
- Các lỗi trong hai mô-đun có xảy ra độc lập hay không?
- Giả sử chương trình có lỗi. Tính xác suất mô-đun 1 có lỗi.
- Giả sử mô-đun 1 có lỗi. Tính xác suất mô-đun 2 có lỗi.

Giải. Bảng phân phối xác suất đồng thời của như sau

		Y	0	1	2	3
		X	0	0,2	0,2	0,05
		1	0,2	0,1	0,1	0,1

4.2 Vectơ ngẫu nhiên liên tục

Định nghĩa 4.6 Cho X, Y là các biến ngẫu nhiên liên tục.

1. **Hàm mật độ xác suất đồng thời** (joint probability density function) của hai biến ngẫu nhiên là một hàm $f(x, y) \geq 0$ thỏa mãn

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1.$$

2. Hàm mật độ xác suất thành phần (marginal probability density function) của X và Y được lần lượt xác định như sau

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy; \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

3. **Hàm phân phối xác suất đồng thời** (joint probability density function) của hai biến ngẫu nhiên

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du.$$

Định lý 4.7 Cho X, Y là các BNN liên tục. Khi đó

1. $P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f(x, y) dx dy$
2. $F(a, b) = P(X \leq a, Y \leq b) = \int_{-\infty}^b \int_{-\infty}^a f(x, y) dx dy$
3. $f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$

Định lý 4.8 Cho X, Y là các BNN liên tục. Các điều sau là tương đương

1. X, Y là độc lập
2. $f(x, y) = f_X(x).f_Y(y), \forall x, y$
3. $F(x, y) = F_X(x).F_Y(y), \forall x, y.$

Ví dụ 4.9 Cho hàm mật độ xác suất đồng thời của các BNN X, Y như sau

$$f(x, y) = \begin{cases} cx(x - y), & 0 < x < 2, -x < y < x \\ 0, & \text{các trường hợp khác} \end{cases}$$

- a. Tìm c .
- b. Tìm hàm mật độ thành phần của X và Y .

Giải. a. Ta có

$$\begin{aligned} 1 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy \\ &= \int_0^2 \int_{-x}^x cx(x - y) dy dx \\ &= \int_0^2 2cx^3 dx \\ &= 8c \end{aligned}$$

Suy ra $c = \frac{1}{8}$.

- b. Tìm hàm mật độ thành phần của X .

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f(x, y) dy \\ &= \int_{-x}^x \frac{1}{8} x(x - y) dy = \frac{x^3}{4} \end{aligned}$$

Như vậy,

$$f_X(x) = \begin{cases} \frac{x^3}{4}, & 0 < x < 2 \\ 0, & \text{các trường hợp khác} \end{cases}$$

Tìm hàm mật độ thành phần của Y .

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{+\infty} f(x, y) dx \\ &= \int_0^2 \frac{1}{8} x(x - y) dx = \frac{1}{3} - \frac{y}{4} \end{aligned}$$

Như vậy,

$$f_Y(y) = \begin{cases} \frac{1}{3} - \frac{y}{4}, & -x < y < x \\ 0, & \text{các trường hợp khác} \end{cases}$$

Ví dụ 4.10 Cho X, Y là các biến ngẫu nhiên liên tục có hàm mật độ xác suất đồng thời như sau

$$f(x, y) = \begin{cases} 6e^{-2x-3y}, & x > 0, y > 0 \\ 0, & \text{các trường hợp còn lại} \end{cases}$$

- a. Tính $P(2 < X < 3, 1 < Y < 2)$.
- b. Tìm hàm phân phối xác suất đồng thời của X và Y .
- c. Tìm hàm mật độ thành phần của X và Y .
- d. X và Y có độc lập không?

Giải. a. Tính $P(1 < X < 2, 2 < Y < 3)$.

$$P(1 < X < 2, 2 < Y < 3) = \int_2^3 \int_1^2 6e^{-2x-3y} dx dy = \dots$$

- b. Tìm hàm phân phối xác suất đồng thời của X và Y .

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du$$

Với $x > 0, y > 0$, ta có

$$\begin{aligned} F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du = \int_0^x \int_0^y 6e^{-2u-3v} dv du \\ &= (1 - e^{-2x})(1 - e^{-3y}) \end{aligned}$$

Như vậy,

$$F(x, y) = \begin{cases} (1 - e^{-2x})(1 - e^{-3y}), & x > 0, y > 0 \\ 0, & \text{các trường hợp còn lại} \end{cases}$$

c. Tìm hàm mật độ thành phần của X . Với $x > 0$,

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_0^{+\infty} 6e^{-2x-3y} dy = \dots$$

Suy ra

$$f_X(x) = \begin{cases} \dots, & x > 0 \\ 0, & \text{các trường hợp còn lại} \end{cases}$$

Tìm hàm mật độ thành phần của Y . Với $y > 0$,

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx = \int_0^{+\infty} 6e^{-2x-3y} dx = \dots$$

Suy ra

$$f_Y(y) = \begin{cases} \dots, & y > 0 \\ 0, & \text{các trường hợp còn lại} \end{cases}$$

d. Ta thấy

$$f_X(x)f_Y(y) = \begin{cases} \dots, & x > 0, y > 0 \\ 0, & \text{các trường hợp còn lại} \end{cases}$$

Suy ra

$$f_X(x)f_Y(y) = f(x, y)$$

Như vậy $X, Y \dots$

Ví dụ 4.11 Cho X, Y là các biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} 4xy & , 0 < x < 1, 0 < y < 1 \\ 0, & \text{trường hợp khác} \end{cases}$$

Tính $P(X < Y)$.

Giải.

$$\begin{aligned} P(X < Y) &= \int_0^1 \int_0^y 4xy \, dx \, dy \\ &= \int_0^1 \left(2x^2 y \Big|_0^y \right) \, dy \\ &= \int_0^1 2y^3 \, dy = \frac{1}{2} \end{aligned}$$

Ví dụ 4.12 Cho hàm mật độ xác suất đồng thời của các BNN X, Y như sau

$$f(x, y) = \begin{cases} Ce^{-x}e^{-2y}, & x > 0, y > 0 \\ 0, & \text{các trường hợp khác} \end{cases}$$

- a. Tìm C .
- b. Tính $P(X > 1, Y < 1)$.
- c. Tính $P(X < Y)$.
- d. Tính $P(X < a)$.

Giải.

.....

.....

.....

.....

.....

.....

Định nghĩa 4.13 Cho X, Y là các biến ngẫu nhiên liên tục.

1. **Hàm mật độ xác suất có điều kiện** của X khi đã biết $Y = y$

$$f_X(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)}, & f_Y(y) > 0 \\ 0, & \text{các trường hợp khác} \end{cases}$$

2. **Hàm mật độ xác suất có điều kiện** của Y khi đã biết $X = x$

$$f_Y(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)}, & f_X(x) > 0 \\ 0, & \text{các trường hợp khác} \end{cases}$$

3. Trung bình thành phần của X, Y lần lượt là

$$E(X) = \int_{-\infty}^{+\infty} xf_X(x)dx = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x,y)dxdy$$

$$E(Y) = \int_{-\infty}^{+\infty} yf_Y(y)dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yf(x,y)dxdy$$

Ví dụ 4.14 Đặt biến ngẫu nhiên X biểu thị thời gian cho đến khi máy chủ kết nối với máy của bạn (tính bằng mili giây) và đặt Y biểu thị thời gian cho đến khi máy chủ ủy quyền cho bạn với tư cách là người dùng hợp lệ (tính bằng mili giây). Giả sử X, Y là các biến ngẫu nhiên liên tục có hàm mật độ xác suất đồng thời như sau

$$f(x, y) = \begin{cases} 6 \cdot 10^{-6} e^{-0,001x-0,002y}, & 0 < x < y \\ 0, & \text{các trường hợp khác} \end{cases}$$

- a. Tìm $f_Y(y|x)$.
- b. Tìm $P(Y > 2000|X = 1500)$ và $E(Y|X = 1500)$.

Giải. a. Tìm hàm mật độ thành phần của X . Với $x > 0$, ta có

$$\begin{aligned} f_X(x) &= \int_x^{+\infty} f(x, y) dy \\ &= \int_x^{+\infty} 6 \cdot 10^{-6} e^{-0,001x-0,002y} dy \\ &= 6 \cdot 10^{-6} \cdot e^{-0,001x} \left(\frac{e^{-0,002y}}{-0,002} \Big|_x^{+\infty} \right) \\ &= 0,003 \cdot e^{-0,003x} \end{aligned}$$

Hàm mật độ có điều kiện của Y . Với $0 < x < y$, ta có

$$\begin{aligned} f_Y(y|x) &= \frac{f(x, y)}{f_X(x)} \\ &= \frac{6 \cdot 10^{-6} e^{-0,001x-0,002y}}{0,003 \cdot e^{-0,003x}} \\ &= 0,002 \cdot e^{0,002x-0,002y} \end{aligned}$$

- b. Tìm $P(Y > 2000|X = 1500)$.

$$\begin{aligned} P(Y > 2000|X = 1500) &= \int_{2000}^{+\infty} f_Y(y|x) dy \\ &= \int_{2000}^{+\infty} 0,002 \cdot e^{0,002 \cdot 1500 - 0,002y} dy \\ &= 0,002 \cdot e^3 \left(\frac{e^{-0,002y}}{-0,002} \Big|_{2000}^{+\infty} \right) \\ &= 0,368 \end{aligned}$$

$$\begin{aligned}
E(Y|X=1500) &= \int_{1500}^{+\infty} y f_Y(y|X=1500) dy \\
&= \int_{1500}^{+\infty} y 0,002 \cdot e^{0,002 \cdot 1500 - 0,002y} dy \\
&= 0,002 \cdot e^3 \int_{1500}^{+\infty} y e^{-0,002y} dy \\
&= 0,002 \cdot e^3 \left(y \frac{e^{-0,002y}}{-0,002} \Big|_{1500}^{+\infty} - \int_{1500}^{+\infty} \frac{e^{-0,002y}}{-0,002} dx \right) \\
&= 0,002 \cdot e^3 \left(\frac{1500 \cdot e^{-3}}{0,002} + \frac{e^{-3}}{0,002 \cdot 0,002} \right) \\
&= 2000
\end{aligned}$$

Ví dụ 4.15 Cho X, Y là các biến ngẫu nhiên liên tục có hàm mật độ xác suất đồng thời như sau

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y) & , 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & , \text{các trường hợp còn lại} \end{cases}$$

a. Tìm hàm mật độ có điều kiện $f_Y(y|x)$.

b. Tính $P(\frac{1}{4} < Y < 1 | X = \frac{3}{4})$.

Giải. a. Hàm mật độ thành phần của X . Với $0 \leq x \leq 1, 0 \leq y \leq 1$

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \dots$$

.....

.....

.....

.....

Tìm hàm mật độ điều kiện $f_Y(y|x)$.

$$f_Y(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)}, & 0 \leq y \leq 1 \\ 0, & \text{các trường hợp khác} \end{cases}$$

Dịnh lý 4.16

Cho X, Y là các biến ngẫu nhiên và một hàm $h(X, Y)$. Kỳ vọng của hàm $h(X, Y)$, ký hiệu là $E(h(X, Y))$, được xác định như sau

- Nếu X, Y là các biến ngẫu nhiên rời rạc thì

$$E(h(X, Y)) = \sum_x \sum_y h(x, y)P(x, y)$$

- Nếu X, Y là các biến ngẫu nhiên liên tục có hàm mật độ đồng thời $f(x, y)$ thì

$$E(h(X, Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x, y)f(x, y)dxdy$$

Ví dụ 4.17 Cho X, Y là các biến ngẫu nhiên liên tục có hàm mật độ đồng thời xác định như sau

$$f(x, y) = \begin{cases} 24xy, & 0 \leq x \leq 1; 0 \leq y \leq 1; x + y \leq 1 \\ 0, & \text{các trường hợp khác} \end{cases}$$

và hàm $h(X, Y) = 0,75 + 0,75X + 1,5Y$. Tính $E(h(X, Y))$.

Giải. Ta có

$$\begin{aligned} E(h(X, Y)) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x, y)f(x, y)dxdy \\ &= \int_0^1 \int_0^{1-y} (0,75 + 0,75x + 1,5y)24xydxdy \\ &= 1,65. \end{aligned}$$

4.3 Hiệp phuong sai và hệ số tương quan

Định nghĩa 4.18 Cho X, Y là các biến ngẫu nhiên.

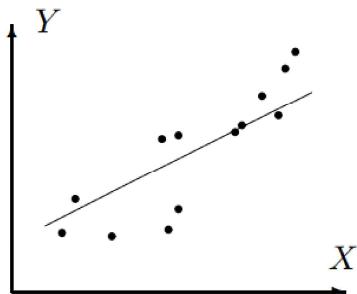
1. **Hiệp phuong sai** (covariance) của X và Y , ký hiệu $\text{Cov}(X, Y)$,

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y).$$

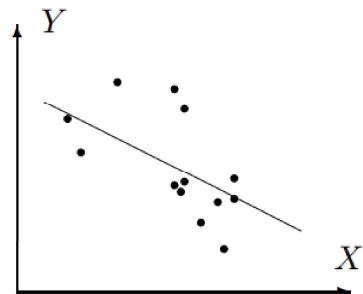
2. **Hệ số tương quan** (Correlation coefficient) của X, Y

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}}$$

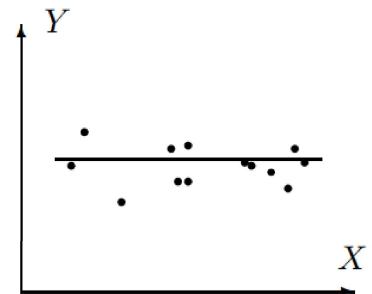
- $\text{Cov}(X, Y) > 0$: Nếu X tăng thì Y tăng; nếu X giảm thì Y giảm.
- $\text{Cov}(X, Y) < 0$: Nếu X tăng thì Y giảm; nếu X giảm thì Y tăng.
- $\text{Cov}(X, Y) = 0$: Ta nói X, Y không tương quan.
- Nếu $|\rho| = 1$ thì ta nói các điểm (x_i, y_j) nằm trên một đường thẳng.
- Nếu ρ gần 1 thì ta nói X, Y có tương quan dương mạnh.
- Nếu ρ gần -1 thì ta nói X, Y có tương quan âm mạnh.
- Nếu ρ gần 0 thì ta nói X, Y có tương quan yếu hoặc không tương quan.



(a) $\text{Cov}(X, Y) > 0$



(b) $\text{Cov}(X, Y) < 0$



(c) $\text{Cov}(X, Y) = 0$

- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\rho(X, Y) = \rho(Y, X)$
- $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$
- $\rho(aX + b, cY + d) = \rho(X, Y)$
- Nếu X, Y độc lập thì $\text{Cov}(X, Y) = 0$.

Ví dụ 4.19 Cho X, Y là các biến ngẫu nhiên liên tục và hàm mật độ xác suất đồng thời

$$f(x, y) = \begin{cases} 2, & x + y \leq 1, x > 0, y > 0 \\ 0, & \text{các trường hợp khác} \end{cases}$$

Tính $\text{Cov}(X, Y)$ và $\rho(X, Y)$.

Giải.

$$E(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf(x, y)dxdy = \int_0^1 \int_0^{1-y} 2xydxdy$$

$$= \dots$$

$$E(X) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x, y)dxdy$$

$$= \int_0^1 \dots$$

$$E(Y) = \dots$$

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{\dots}{\dots} - \frac{\dots}{\dots} = \frac{\dots}{\dots}$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f_X(x)dx$$

$$= \dots$$

$$E(Y^2) = \dots$$

$$V(X) = E(X^2) - E(X)^2 = \frac{\dots}{\dots} - \frac{\dots}{\dots} = \frac{\dots}{\dots}$$

$$V(Y) = \dots$$

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\dots}{\dots}$$

Ví dụ 4.20 Cho X, Y là các biến ngẫu nhiên liên tục có hàm mật độ xác suất đồng thời như sau

$$f(x,y) = \begin{cases} x^2 + \frac{xy}{3} & , 0 \leq x \leq 1, 0 \leq y \leq 2 \\ 0 & , \text{ các trường hợp còn lại} \end{cases}$$

- a. Tìm hàm mật độ có điều kiện $f_X(x|y)$.
 b. Tính $E(X|Y = \frac{1}{2})$.

Giải.

BÀI TẬP

Bài 4.1 Cho hai biến ngẫu nhiên X, Y có phân phối xác suất đồng thời như sau

	Y	1	2	3	4
X		0	0,06	0,06	0,1
1		0,1	0,1	0,04	0,04
2		0,4	0,1	0	0

a. X và Y có độc lập không? Vì sao?

b. Tính xác suất $P(X + Y \leq 3)$

c. Tính $P(X > 1|Y = 2)$

Bài 4.2 Cho X, Y là các biến ngẫu nhiên rời rạc có phân phối xác suất đồng thời như sau

	Y	0	1	2
X		0,1	0,04	0,02
0		0,08	0,2	0,06
1		0,06	0,14	0,3

a. Tính $P(X \leq 1, Y \leq 1)$.

b. Tính $P(X > 0, Y > 0)$

c. Tìm bảng phân phối xác suất thành phần của X và Y .

d. X, Y có độc lập không? Tại sao?

Bài 4.3 Một hộp có 3 bi đỏ, 2 bi vàng và 3 bi xanh. Lấy ngẫu nhiên 2 bi từ hộp. Gọi X là số bi đỏ và Y là số bi vàng trong 2 bi lấy ra.

a. Lập bảng phân phối xác suất đồng thời của X và Y .

b. Tính $P(X + Y \leq 1)$. (ĐS: 9/14)

c. Tìm các phân phối xác suất thành phần của X và Y .

Bài 4.4 Cho X, Y là các biến ngẫu nhiên liên tục có hàm mật độ đồng thời như sau

$$f(x, y) = \begin{cases} kx^3y^2, & 0 \leq x \leq 2, 0 \leq y \leq 2 \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm k .

b. Tìm hàm mật độ thành phần của X và Y .

c. Tìm hàm phân phối xác suất thành phần của X .

Bài 4.5 Cho X, Y là hai biến ngẫu nhiên có hàm mật độ xác suất đồng thời

$$f(x, y) = \begin{cases} Cxy, & x \in [0; 2]; y \in [1; 3] \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm C .

b. Tính $P(X \leq 1, Y > 2)$.

c. Tính $P(X \leq 1|Y > 2)$

Bài 4.6 Giả sử X, Y là tuổi thọ trung bình của 2 thiết bị có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} e^{-(x+y)}, & 0 < x; 0 < y \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm hàm mật độ thành phần của X và Y .

b. Tìm $E(X + Y), V(X + Y)$.

Bài 4.7 Giả sử X, Y là hai biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} e^{-y(x+1)}, & 0 \leq x; 0 \leq y \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm hàm mật độ thành phần của X và Y .

b. X, Y có độc lập không?

c. Tìm $P(0 < X < 1 | Y = 2)$.

Bài 4.8 Giả sử X, Y là hai biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} kx, & 0 < y < x < 1 \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm k .

b. Tìm hàm mật độ thành phần của X và Y .

c. X, Y có độc lập không?

Bài 4.9 Giả sử X, Y là hai biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} ke^{-x-y}, & 0 < y < x \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm k .

b. Tìm hàm mật độ thành phần của X và Y .

c. X, Y có độc lập không?

Bài 4.10 Cho X, Y là các biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} C(x^2 + y), & -1 \leq x < 1, 0 \leq y \leq 1. \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm C .

b. Tìm các hàm mật độ thành phần của X và Y . Các biến ngẫu nhiên X, Y có độc lập không?

c. Tính $P(Y < 0, 6)$ và $P(Y < 0, 6 | X < 0, 5)$.

Bài 4.11 Giả sử X, Y là hai biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} x + y, & 0 \leq x \leq 1; 0 \leq y \leq 1 \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm $P(X < \frac{1}{2}, Y > \frac{1}{4})$.

b. Tìm $P(X + Y < 1)$.

c. Tìm hàm mật độ thành phần của X và Y .

Bài 4.12 Giả sử X, Y là hai biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} cxy, & 0 \leq y \leq x; 0 \leq x \leq 2 \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm c .

b. Tìm hàm mật độ thành phần của X và Y . Hai biến X và Y có độc lập không?

c. Tìm $P(X < \frac{1}{2}, Y < \frac{3}{4})$.

d. Tìm $P(X \leq \frac{1}{2} | Y < \frac{3}{4})$.

Bài 4.13 Giả sử X, Y là hai biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} cx^2y, & 0 < x^2 \leq y < 1, x > 0 \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm c .

b. Tìm $P(0 < X < \frac{3}{4}, \frac{1}{4} \leq Y < 1)$.

Bài 4.14 Giả sử X, Y là hai biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} x + y, & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm hàm mật độ thành phần của X, Y

b. X và Y có độc lập không?

c. Tìm hiệp phương sai và hệ số tương quan của X và Y .

Bài 4.15 Giả sử X, Y là hai biến ngẫu nhiên liên tục có hàm mật độ đồng thời

$$f(x, y) = \begin{cases} k(x + y), & 0 \leq x \leq 2, 0 \leq y \leq 2 \\ 0, & \text{các trường hợp khác} \end{cases}$$

a. Tìm k .

b. Tìm các hàm mật độ có điều kiện $f_X(x|y), f_Y(y|x)$

c. Tìm $P(0 \leq Y \leq 0,5|X = 1)$.

Chương 5. Lý thuyết mẫu

Nguyễn Minh Trí

Trường Đại học Công nghệ Thông tin

Ngày 16 tháng 4 năm 2023

5.1 Tổng thể và mẫu

Ví dụ 5.1 Ta muốn biết thu nhập trung bình trong năm 2021 của giáo viên đang giảng dạy ở Thành phố Hồ Chí Minh. Ta sẽ lập danh sách tất cả các giáo viên đang dạy ở thành phố Hồ Chí Minh và ghi lại thu nhập của từng người trong năm 2021.

Tuy nhiên, việc thu thập dữ liệu của tất cả các giáo viên tốn rất nhiều thời gian, công sức. Do đó, người ta có thể chọn ra một nhóm giáo viên (ta sẽ gọi là **mẫu**) trong toàn bộ giáo viên (gọi là **tổng thể**) để điều tra. Bằng các phương pháp của ngành thống kê, người ta có thể đưa ra được mức thu nhập trung bình của toàn bộ giáo viên trên địa bàn Thành phố Hồ Chí Minh.

Chúng ta quan tâm đến tổng thể, nhưng tổng thể có thể khó hoặc không thể thống kê được. Thay vào đó, **người ta cố gắng mô tả hoặc dự đoán các đặc điểm của tổng thể trên cơ sở thông tin thu được từ mẫu đại diện của tổng thể đó.**

Định nghĩa 5.2

1. Một **tổng thể** (population) là tập hợp tất cả các đối tượng có chung một tính chất mà ta quan tâm. Số phần tử của tổng thể được gọi là *kích thước tổng thể*.
2. Việc chọn từ tổng thể một tập con nào đó được gọi là **phép lấy mẫu**.
3. Một **mẫu** (sample) là một tập con của tổng thể. Số phần tử của mẫu được gọi là *kích thước mẫu*.

Ví dụ 5.3 Một công ty sản xuất chip máy tính đóng gói mỗi hộp gồm 100 chip. Người ta muốn khảo sát tỉ lệ chip bị lỗi trong một lô hàng gồm 1000 hộp của công ty. Chọn ngẫu nhiên 80 hộp chip để kiểm tra. Ta có **tổng thể** là 1000 hộp chip và **mẫu** là 80 hộp chip được kiểm tra.

Một trong những nhiệm vụ quan trọng của thống kê là xây dựng các phương pháp cho phép rút ra các kết luận hoặc đưa ra dự báo về toàn bộ tổng thể dựa trên một mẫu. Do đó, **vấn đề lấy mẫu là một việc vô cùng quan trọng**.

Nếu những suy luận từ mẫu đối với tổng thể là hợp lệ thì ta phải lấy những mẫu đại diện cho tổng thể. Vì một số lý do chủ quan nên việc chọn mẫu có thể dẫn đến nhiều sai lệch trong việc suy luận.

Dịnh nghĩa 5.4 Một mẫu là ngẫu nhiên (random sample) nếu trong phép lấy mẫu đó, mỗi phần tử được chọn một cách độc lập và có xác suất được chọn như nhau.

Yêu cầu khi lấy mẫu:

- Mẫu được chọn là mẫu ngẫu nhiên.
- Kích thước mẫu đủ lớn.

Nếu kích thước mẫu càng lớn thì thông tin suy luận về tổng thể càng đáng tin cậy và có ý nghĩa.

Ví dụ 5.5 Khảo sát chiều cao trung bình của 1 triệu dân của một thành phố nọ.

- Nhóm 1 khảo sát một mẫu ngẫu nhiên gồm 50 người và tìm được chiều cao trung bình của 50 người này là 164 cm. Nhóm 2 khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm. Khi đó suy luận về chiều cao trung bình của tổng thể của nhóm 2 đáng tin cậy hơn nhóm 1.
- Một nhóm khảo sát khảo sát một mẫu ngẫu nhiên gồm 2000 người và tìm được chiều cao trung bình của 2000 người này là 164,6 cm. Khi đó nhóm đưa ra các kết luận
 1. Chiều cao trung bình của tổng thể là 164,6 cm.
 2. Chiều cao trung bình của tổng thể là từ 164 cm đến 165 cm.
 3. Chiều cao trung bình của tổng thể là từ 163,5 cm đến 165,5 cm.
 4. Chiều cao trung bình của tổng thể là từ 130 cm đến 185 cm.

Trong 3 suy luận đầu tiên, ta thấy suy luận nào **đáng tin cậy hơn?**
Kết luận 3 và 4, thông tin nào **có giá trị hơn?**

Dịnh nghĩa 5.6

1. **Thống kê mô tả** (Descriptive Statistics) là các phương pháp sử dụng để tóm tắt hoặc mô tả một tập hợp dữ liệu, một mẫu nghiên cứu dưới dạng số hay biểu đồ trực quan.
2. **Thống kê suy luận** (Inferential statistics) bao gồm các phương pháp được sử dụng để suy luận về các đặc điểm tổng thể từ thông tin có trong một mẫu được lấy từ tổng thể này.

5.2 Biểu diễn mẫu

Một số cách biểu diễn mẫu như sau:

- Mẫu dạng điểm
- Mẫu dạng tần số (tần suất)
- Mẫu dạng khoảng
- Mẫu dạng biểu đồ

Định nghĩa 5.7 1. Cho một mẫu có kích thước n , các giá trị của dấu hiệu X mà ta muốn nghiên cứu là $x_1 < x_2 < \dots < x_m$. 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	\dots	x_m
Tần số	k_1	k_2	\dots	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	\dots	x_m
Tần suất	f_1	f_2	\dots	f_m

Ví dụ 5.8 Kiểm tra 80 hộp (mỗi hộp chứa 100 chip bán dẫn) để tìm số lượng chip bị lỗi trong mỗi hộp.

1	3	4	7	2	7	5	5	2	2	4	2	4	3	2
2	7	1	3	3	2	5	0	0	1	2	5	5	4	1
3	2	6	3	8	2	2	3	1	6	3	4	1	2	5
1	3	3	3	2	1	2	5	5	4	1	4	3	1	0
2	1	2	4	4	5	3	3	4	0	5	2	5	6	2
5	3	3	3	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, \dots, 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 5.9 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

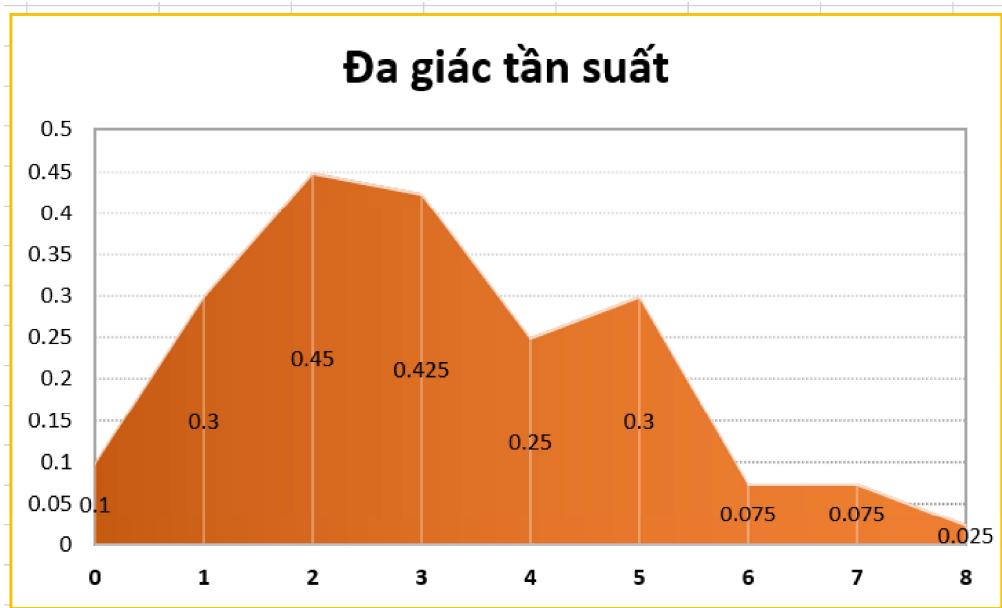
Trong một mẫu, dấu hiệu điều tra X có bảng phân bố tần số và tần suất

x_i	x_1	x_2	\dots	x_m
Tần số	k_1	k_2	\dots	k_m
Tần suất	f_1	f_2	\dots	f_m

Trên mặt phẳng toạ độ, nối điểm $(x_i; k_i)$ với điểm $(x_{i+1}; k_{i+1})$ bởi đoạn thẳng, với $i = 1, \dots, m - 1$ ta được **biểu đồ tần số hình gãy**.

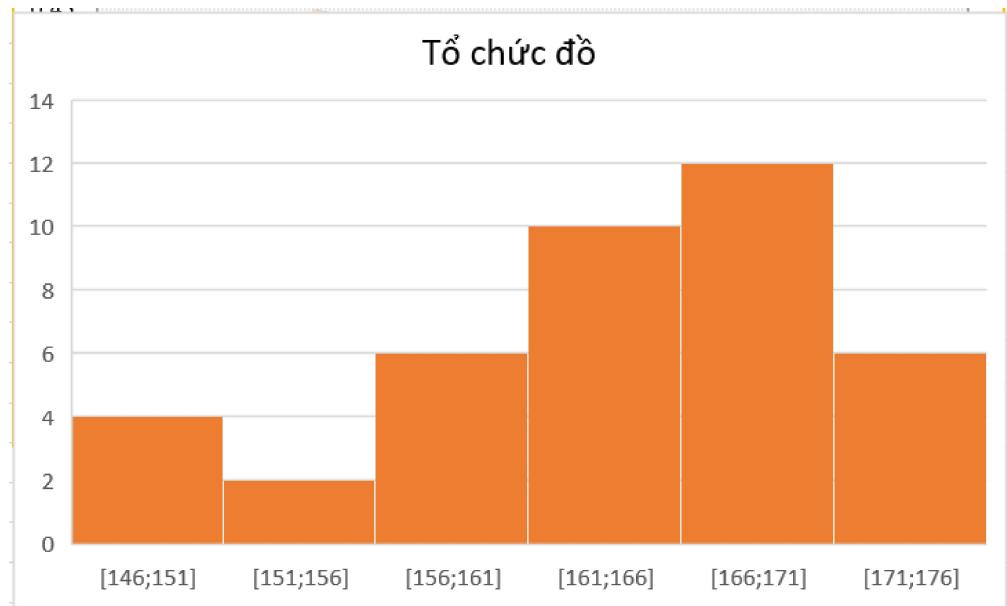


Nối các điểm $(x_i; f_i)$ với $(x_{i+1}; f_{i+1})$ bởi đoạn thẳng, với $i = 1, 2, \dots, k - 1$ ta được đường gấp khúc được gọi là **biểu đồ đa giác tần suất**

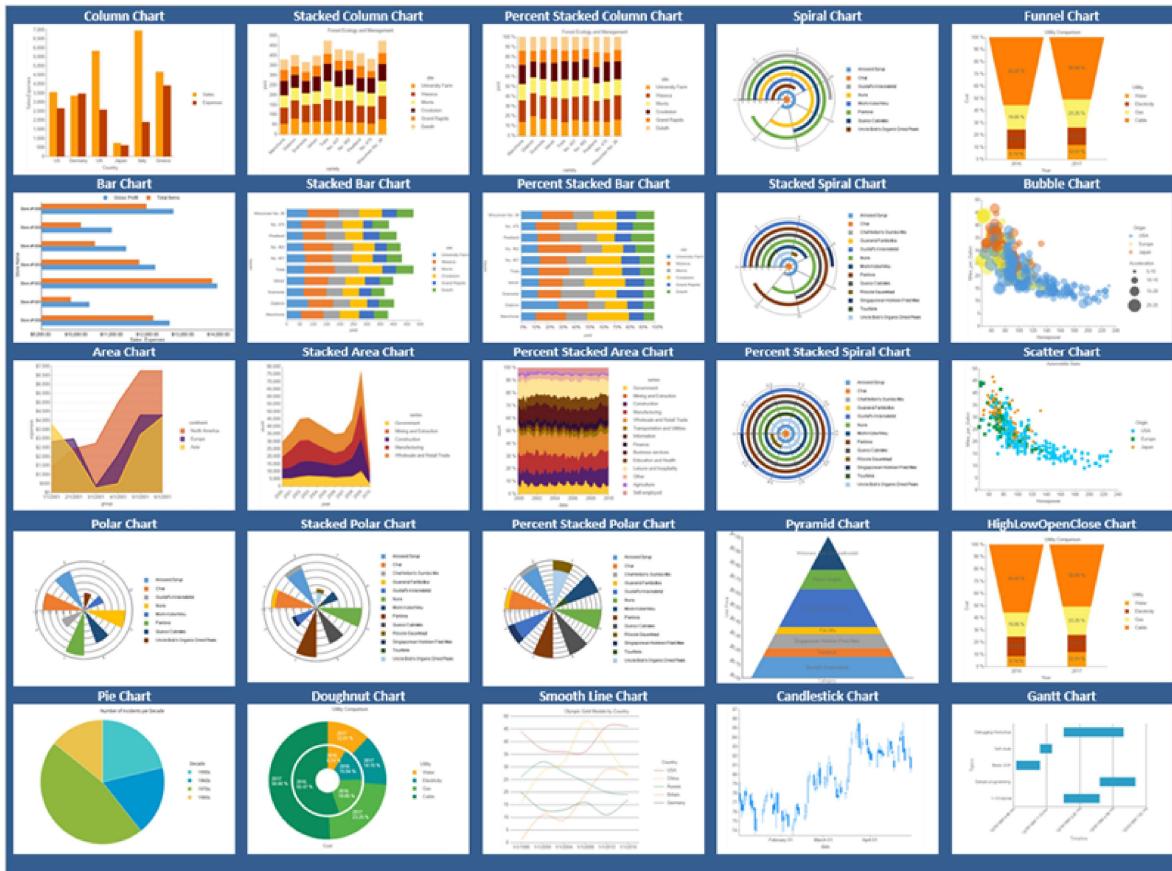


Đối với bản phân bố lớp ghép, người ta thường dùng **tổ chức đồ** (histogram) để biểu diễn.

Ta có tổ chức đồ tần số của Ví dụ 5.9



Người ta có thể dùng các biểu đồ khác nhau để biểu diễn dữ liệu thu được.



Nguồn: Internet.

5.3 Các số đo của mẫu

- Một tổng thể có kích thước N : v_1, v_2, \dots, v_N
- Các mẫu ngẫu nhiên có kích thước n . Ta xem mỗi phần tử trong mẫu tương ứng với một biến ngẫu nhiên X_i với $i = 1, 2, \dots, n$.
- Một mẫu có kích thước n nhận các giá trị x_1, \dots, x_n .

Định nghĩa 5.10

1. Trung bình tổng thể

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

2. Trung bình mẫu ngẫu nhiên

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

3. Trung bình của một mẫu cụ thể

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Nếu có một phân bố lớp ghép với m khoảng C_1, C_2, \dots, C_m và tần số của khoảng C_i là k_i thì trung bình mẫu \bar{x} được xác định bởi

$$\bar{x} = \frac{\sum_{i=1}^m k_i x_i}{\sum_{i=1}^m k_i}$$

trong đó x_i là trung điểm (tâm) của khoảng C_i .

Ví dụ 5.11 Một công ty đánh giá năng suất của mỗi nhân viên theo thang điểm từ 1 đến 5. Giả sử công ty có $N = 10000$ nhân viên trong đó 300 nhân viên được xếp hạng 1, 700 được xếp hạng 2, 4000 được xếp hạng 3, 4000 được xếp hạng 4 và 1000 được xếp hạng 5. Chọn ngẫu nhiên 10 nhân viên để phỏng vấn về mức độ hài lòng đối với công việc. Xếp hạng năng suất của 10 nhân viên được chọn là

$$x_1 = 2, x_2 = x_3 = x_4 = 3, x_5 = x_6 = x_7 = x_8 = 4, x_9 = x_{10} = 5$$

- a. Đặt X là xếp hạng năng suất của một nhân viên được chọn ngẫu nhiên. Hãy tính giá trị trung bình của X .
- b. Tính giá trị trung bình mẫu của xếp hạng năng suất của 10 nhân viên được chọn

Giải. Xếp hạng năng suất của các nhân viên như sau

$$\begin{aligned} v_i &= 1, i = 1, 2, \dots, 300 \\ v_i &= 2, i = 301, 302, \dots, 1000 \\ v_i &= 3, i = 1001, 1002, \dots, 5000 \\ v_i &= 4, i = 5001, 5002, \dots, 9000 \\ v_i &= 5, i = 9001, 9002, \dots, 10000 \end{aligned}$$

a. Giá trị trung bình của hạng năng suất của nhân viên

$$\mu = \frac{1}{10000} \sum_{i=1}^{10000} v_i = 3,47$$

b. Giá trị trung bình mẫu của hạng năng suất của 10 nhân viên được chọn

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 3,7$$

Nhận xét: Trung bình của mẫu \bar{x} nói chung là *xấp xỉ* trung bình tổng thể μ .

Định nghĩa 5.12 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 5.13 Một mẫu gồm 5 sinh viên đại học trả lời câu hỏi "Bạn đã dành bao nhiêu thời gian, tính bằng phút, cho trang mạng xã hội ngày hôm qua?" Các số liệu thu được như sau:

100 45 60 130 30

Tìm trung bình mẫu và trung vị mẫu.

Giải. Trung bình mẫu

$$\bar{x} = \frac{100 + 45 + 60 + 130 + 30}{5} = 73(\text{phút})$$

Sắp xếp dữ liệu theo thứ tự

30 45 60 100 130

Khi đó trung vị mẫu là 60 (phút).

Định nghĩa 5.14

1. **Phương sai tổng thể** (population variance), ký hiệu σ^2 ,

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

2. **Phương sai mẫu ngẫu nhiên** (sample variance), ký hiệu S^2 ,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

3. **Phương sai mẫu cụ thể** (sample variance), ký hiệu s^2 ,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

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

Định lý 5.15 Nếu s^2 là phương sai của một mẫu có kích thước n thì

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right)$$

Ví dụ 5.16 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 phương sai của mẫu này.

- Trung bình mẫu

$$\bar{x} = \frac{12 + 15 + 17 + 20}{4} = 16(\text{nghìn đồng})$$

- Phương sai mẫu

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

5.4 Phân phối lấy mẫu

- Ta đã biết một biến ngẫu nhiên là một mô tả bằng số về kết quả của một phép thử.
- Khi chọn một mẫu từ một tổng thể, các số đo mô tả tính được từ mẫu đó được gọi là **các thống kê** (statistic).
- Các thống kê thay đổi theo các mẫu khác nhau mà ta chọn, do đó chúng là các biến ngẫu nhiên.
- Phân phối xác suất của các thống kê được gọi là các **phân phối lấy mẫu** (sampling distribution)

Định nghĩa 5.17 Phân phối lấy mẫu của một thống kê là phân phối xác suất với tất cả các giá trị của thống kê mà nó là kết quả khi các mẫu ngẫu nhiên có kích thước n được lấy lặp lại nhiều lần từ một tổng thể.

Ví dụ 5.18 Phân phối lấy mẫu của \bar{X} là phân phối xác suất của tất cả các giá trị của trung bình mẫu \bar{X} .

Ví dụ 5.19 Thời gian làm việc của 6 nhân viên tại một văn phòng như sau

2 4 6 6 7 8

Trung bình số năm làm việc của 6 nhân viên là

$$\bar{x} = \frac{1}{6}(2 + 4 + 6 + 6 + 7 + 8) = 5,5$$

Tính trung bình số năm làm việc của 2 nhân viên được chọn ngẫu nhiên từ tổng thể 6 nhân viên. Có tất cả 15 mẫu ngẫu nhiên có kích thước bằng 2

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

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

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

Kỳ vọng của trung bình mẫu

$$E(\bar{X}) = 3\frac{1}{5} + 4\frac{2}{5} + 4,5\frac{1}{5} + 5\frac{3}{5} + 5,5\frac{1}{5} + 6\frac{2}{5} + 6,5\frac{2}{5} + 7\frac{2}{5} + 7,5\frac{1}{5} = 5,5$$

Định lý 5.20

- Trung bình (kỳ vọng) của \bar{X}

$$E(\bar{X}) = \mu.$$

- Phương sai của \bar{X}

Kích thước tổng thể hữu hạn	Kích thước tổng thể vô hạn
$\sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1}} \left(\frac{\sigma}{\sqrt{n}} \right)$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

- Nếu tổng thể có phân phối chuẩn thì các mẫu ngẫu nhiên sẽ có phân phối chuẩn. Do đó phân phối trung bình mẫu sẽ có phân phối chuẩn.
- Nếu tổng thể không có phân phối chuẩn thì theo Định lý giá trị trung tâm phân phối trung bình mẫu sẽ xấp xỉ phân phối chuẩn khi kích thước các mẫu đủ lớn.

Các nhà thống kê thấy rằng khi kích thước mẫu $n \geq 30$ thì có thể áp dụng định lý giới hạn trung tâm.

Dịnh nghĩa 5.21 Cho một tổng thể có kích thước N trong đó có k phần tử của tổng thể có tính chất \mathcal{P} mà ta quan tâm. Tỉ lệ tổng thể (population proportion) có tính chất \mathcal{P} là

$$p = \frac{k}{N}.$$

- Tỉ lệ mẫu (sample proportion) \bar{p} là một công thức ước lượng của tỉ lệ tổng thể p . Công thức tính tỉ lệ mẫu

$$\bar{p} = \frac{x}{n}$$

trong đó x là số phần tử có tính chất \mathcal{P} mà ta quan tâm, n là kích thước mẫu.

Ví dụ 5.22 Một cuộc khảo sát 1200 sinh viên ngành công nghệ thông tin sắp tốt nghiệp đại học, người ta thấy có 552 sinh viên sẽ tiếp tục học để nâng cao trình độ sau khi tốt nghiệp đại học. Khi đó, tỉ lệ mẫu của những sinh viên được khảo sát mà có dự định học tiếp sau khi tốt nghiệp đại học là

$$\bar{p} = \frac{552}{1200} = 0,46$$

- Từ khảo sát này, ta có thể ước tính tỉ lệ sinh viên ngành công nghệ thông tin sẽ tiếp tục học để nâng cao trình độ sau khi tốt nghiệp đại học là 46%.
- Người ta có thể tiến hành một khảo sát thứ 2 với 1200 sinh viên ngành công nghệ thông tin về việc tiếp tục học để nâng cao trình độ sau khi tốt nghiệp đại học. Tỉ lệ ở cuộc khảo sát thứ 2 có thể khác lần 1.
- Như vậy, tỉ lệ mẫu \bar{p} thay đổi tùy theo các mẫu. Do đó \bar{p} là một biến ngẫu nhiên và nó có phân phối xác suất.

Dịnh nghĩa 5.23 Phân phối lấy mẫu của \bar{p} là phân phối xác suất của tất cả các giá trị có thể có của tỉ lệ mẫu \bar{p} .

- Để biết được hình dạng, trung bình, độ lệch của phân phối lấy mẫu \bar{p} , ta tiến hành khảo sát trên nhiều mẫu có kích thước 1200 (sinh viên ngành công nghệ thông tin).
- Khi đó, ta thu được một danh sách các tỉ lệ mẫu. Bằng cách minh họa bằng tổ chức đồ, ta có thể thấy được hình dạng của phân phối của tỉ lệ mẫu.
- Ta cũng có thể tính được trung bình tỉ lệ mẫu, phương sai tỉ lệ mẫu.
- Trung bình (kỳ vọng) của \bar{p}

$$E(\bar{p}) = p$$

- Phương sai của \bar{p}

Kích thước tổng thể hữu hạn	Kích thước tổng thể vô hạn
$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p(1-p)}{n}}$	$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

Ta đã biết phân phối nhị thức $B(n; p)$ xấp xỉ phân phối chuẩn $N(np; np(1-p))$ khi

$$np \geq 5 \text{ và } n(1-p) \geq 5.$$

Sử dụng phần mềm R trong thống kê

R là một hệ thống phân tích dữ liệu, một máy tính phức tạp và một ngôn ngữ lập trình hướng đối tượng. Nó cung cấp một môi trường để phân tích thống kê và đồ họa.

R được tải xuống từ trang Web CRAN (Comprehensive R Archive Network) và được tải xuống bằng cách tiến hành như sau:

- Vào trang web <http://cran.r-project.org/>
- Chọn phiên bản tải về phù hợp với hệ điều hành.
- Chọn gói "base"
- Tải chương trình về và cài đặt

Chương 6. Lý thuyết ước lượng

Nguyễn Minh Trí

Trường Đại học Công nghệ Thông tin

Ngày 15 tháng 4 năm 2023

6.1 Ước lượng điểm

- Các giá trị trung bình, phương sai, độ lệch chuẩn và trung vị của tổng thể được gọi là các **tham số** (parameter).
- Các tham số đã biết μ và σ của phân phối chuẩn; p của phân phối nhị thức.
- Các giá trị trung bình, phương sai, độ lệch chuẩn và trung vị của mẫu được gọi là các **thống kê** (statistic). Các thống kê đóng vai trò là nguồn thông tin về một tham số.

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 **ước lượng điểm**.
- Nếu nói chiều cao trung bình đó nằm trong khoảng từ 159 cm đến 169 cm hay 164 ± 5 cm. Khi đó ta đã có một **ước lượng khoảng**.

Định nghĩa 6.3 Công thức 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).

- Ta tính các số đo mô tả từ mẫu được gọi là các **các thống kê** (statistic) và dùng chúng để ước tính giá trị của các tham số tổng thể.

Ví dụ 6.4 Ta xem trung bình mẫu \bar{x} là công thức ước lượng điểm (point estimator) của trung bình tổng thể μ , độ lệch chuẩn mẫu s^2 là công thức ước lượng điểm của độ lệch chuẩn tổng thể σ^2 và tỷ lệ mẫu \bar{p} là công thức ước lượng điểm của tỷ lệ tổng thể p . Các giá trị cụ thể của \bar{x}, s^2, \bar{p} được gọi là các giá trị ước lượng điểm (point estimate).

Ta ký hiệu chung như sau

θ : tham số của tổng thể mà ta quan tâm

$\hat{\theta}$: thống kê mẫu hoặc công thức ước lượng điểm của θ

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

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

6.2 Ước lượng khoảng

- Một công thức 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.
- Dạng tổng quát của ước lượng khoảng như sau:

Giá trị ước lượng điểm \pm độ sai số

Định nghĩa 6.6

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 6.7 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%.

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

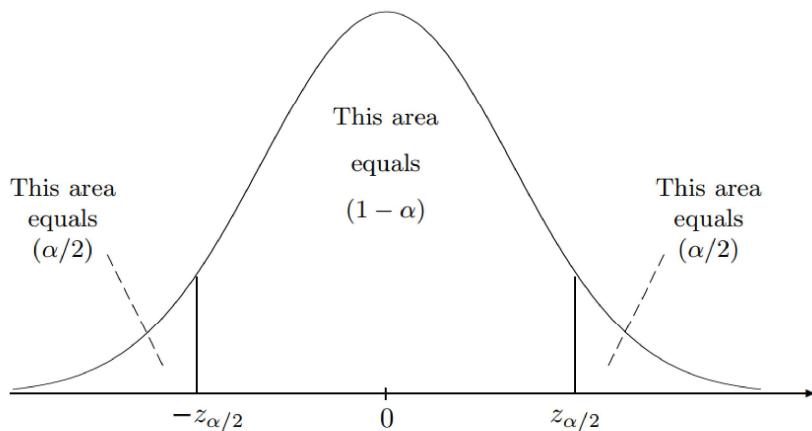
Bài toán. Giả sử rằng thời gian mua sắm của khách hàng tại một trung tâm thương mại có phân phối chuẩn với độ lệch chuẩn tổng thể là 20 phút. Chọn ngẫu nhiên 64 người đã mua sắm ở trung tâm đó. Người ta thấy rằng thời gian mua sắm trung bình của 64 người này là 75 phút. Tìm thời gian mua sắm trung bình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

- Nếu tổng thể có phân phối chuẩn hoặc kích thước mẫu $n \geq 30$ (tổng thể không có phân phối chuẩn) thì $\bar{X} \approx N(\mu; \frac{\sigma^2}{n})$.
- Đổi biến

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}},$$

khi đó $Z \sim N(0; 1)$.



Với độ tin cậy $1 - \alpha$ thì giá trị $z_{\alpha/2}$

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

Tìm được $z_{\alpha/2}$ bằng bảng A4. Do đó

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

Trung bình tổng thể sẽ thuộc khoảng (khoảng tin cậy của trung bình tổng thể với độ tin cậy $1 - \alpha$)

$$\boxed{[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]}$$

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

Giải Bài toán 1.

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

Ví dụ 6.8 Chọn ngẫu nhiên 30 người để kiểm tra thời gian sử dụng chiếc điện thoại di động đầu tiên. Người ta thấy rằng thời gian sử dụng trung bình của 30 người này là 5,6 năm. Giả sử thời gian sử dụng chiếc điện thoại di động đầu tiên có phân phối chuẩn với độ lệch chuẩn là 0,8 năm. Tính khoảng thời gian trung bình sử dụng chiếc điện thoại đầu tiên với độ tin cậy 99%.

Giải.

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

Ví dụ 6.9 Thu nhập trung bình hàng tháng của 30 hộ dân trong một thành phố được cho như sau (đơn vị triệu đồng)

12.23	16.56	4.39	2.89	1.24	2.17	13.19	9.16	1.42	73.25
1.91	14.64	11.59	6.69	1.06	8.74	3.17	18.13	7.92	4.78
16.85	40.22	2.42	21.58	5.01	1.47	12.24	2.27	12.77	2.76

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 14.405.

Giải.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

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

Bài toán 2. Thu nhập trung bình hàng tháng của **30** hộ dân trong một thành phố được cho như sau (đơn vị triệu đồng)

12.23	16.56	4.39	2.89	1.24	2.17	13.19	9.16	1.42	73.25
1.91	14.64	11.59	6.69	1.06	8.74	3.17	18.13	7.92	4.78
16.85	40.22	2.42	21.58	5.01	1.47	12.24	2.27	12.77	2.76

Tìm khoảng thu nhập trung bình hàng tháng của toàn thành phố với độ tin cậy 90%.

Bài toán 3. Theo một thống kê cho thấy số thu nhập của 7 công nhân trong năm 2021 của một công ty được cho như sau (đơn vị triệu đồng)

54,6	59	60,9	63,1	71,6	84,4	99,3
------	----	------	------	------	------	------

Giả sử thu nhập trong năm 2021 của công ty có phân phối chuẩn. Tính khoảng thu nhập trung bình của công ty này với độ tin cậy 99%.

- Cho một tổng thể có phân phối chuẩn với trung bình μ .
- Cho \bar{X} là công thức trung bình của mẫu ngẫu nhiên và \bar{x}, s lần lượt là giá trị của trung bình và độ lệch chuẩn của một mẫu có kích thước n .
- Biến ngẫu nhiên

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

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

- khoảng tin cậy của μ với độ tin cậy $(1 - \alpha)$ là

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

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

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

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

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

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

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

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

ν (d.f.)	α , the right-tail probability									
	.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 6.11 Theo một thống kê cho thấy số thu nhập của 7 công nhân trong năm 2021 của một công ty được cho như sau (đơn vị triệu đồng)

54,6 59 60,9 63,1 71,6 84,4 99,3

Giả sử thu nhập trong năm 2021 của công ty có phân phối chuẩn. Tính khoảng thu nhập trung bình của công ty này với độ tin cậy 99%.

Giải.

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

$$\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}$$

$$70,414 - 3,707 \frac{16,103}{\sqrt{7}} \leq \mu \leq 70,414 + 3,707 \frac{16,103}{\sqrt{7}}$$

$$47,852 \leq \mu \leq 92,976$$

Ví dụ 6.12 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	3300	3500	3600	4000
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 xuất 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: $\bar{x} = \dots$
- Độ lệch chuẩn mẫu: $s = \dots$
- Độ tin cậy $1 - \alpha = 0,97$. Suy ra $\Phi(z_{\alpha/2}) = 1 - \alpha/2 = 0,985$. Do đó $z_{\alpha/2} = \dots$
- Độ chính xác:

$$z_{\alpha/2} \frac{s}{\sqrt{n}} = 2,17 \frac{217,3683}{\sqrt{50}} = \dots$$

- Khoảng tin cậy của tuổi thọ trung của bóng đèn với độ tin cậy 97% là \dots

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

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

Suy ra

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

Do đó

$$\Phi(1,92) = \Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}.$$

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

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

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

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

Ví dụ 6.13 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 55 165 9 15 66 23 30 150 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 4. Thăm dò ý kiến của 100 cử tri được chọn ngẫu nhiên tại một thành phố cho thấy có 80% trong số cử tri này ủng hộ ứng viên A. Với độ tin cậy 98%, hãy ước lượng tỉ lệ của tất cả các cử tri ủng hộ ứng viên A tại thành phố này.

- p : tỉ lệ tổng thể (tỉ lệ phần tử có tính chất \mathcal{P} trong tổng thể)
- f : tỉ lệ mẫu cụ thể (tỉ lệ phần tử có tính chất \mathcal{P} trong mẫu)
- Khi $nf > 5$ và $n(1 - f) > 5$ thì tỉ lệ mẫu ngẫu nhiên sẽ xấp xỉ phân phối chuẩn $N(f; \frac{f(1 - f)}{n})$.
- Với độ tin cậy $1 - \alpha$, khoảng tin cậy chứa tỉ lệ tổng thể là

$$[f - z_{\alpha/2} \sqrt{\frac{f(1 - f)}{n}}; f + z_{\alpha/2} \sqrt{\frac{f(1 - f)}{n}}]$$

trong đó $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$ (xem bảng A4).

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

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

$$z_{\alpha/2} \sqrt{\frac{f(1-f)}{n}} \leq z_{\alpha/2} \frac{1}{2\sqrt{n}}.$$

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

Giải Bài toán 4. Theo đề bài

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

$$z_{\alpha/2} \sqrt{\frac{f(1-f)}{n}} = 2,33 \frac{0,4}{10} = 0,0932.$$

- Khoảng tin cậy $[0,7068; 0,8932]$.

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

Ví dụ 6.14 Lấy ngẫu nhiên 200 sản phẩm trong một kho hàng để kiểm tra thì thấy có 21 sản phẩm có lỗi.

- Với độ tin cậy 95%, hãy ước lượng **tỉ lệ sản phẩm lỗi** của cả kho hàng.
- 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?
- 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 6.1 Trọng lượng của những vĩ thuốc do một công ty sản xuất có phân phối chuẩn với độ lệch chuẩn $0,038\text{mg}$. Một mẫu ngẫu nhiên gồm 10 vĩ thuốc có trọng lượng trung bình $4,87\text{mg}$. Hãy ước lượng trọng lượng trung bình của các vĩ thuốc do công ty sản xuất với độ tin cậy 95%.

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

Bài 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 $0,75\text{ kg}$.

b. Với mẫu khảo sát trên, nếu ước lượng lượng rác thải hàng ngày này với độ chính xác nhỏ hơn $0,8\text{kg/ngày}$ và độ tin cậy 95% thì cần khảo sát tối thiểu bao nhiêu gia đình?

Bài 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

Hãy ước tính giá trung bình của các loại laptop có RAM 8G với độ tin cậy 95%.

Bài 6.5 Một tỉnh nợ 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 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 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 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 6.9 Một mẫu ngẫu nhiên gồm 270 laptop được lấy từ một lượng lớn các laptop cũ để ước tính tỉ lệ laptop có ổ cứng bị lỗi. Nếu trên thực tế, 20% laptop có ổ cứng bị lỗi thì xác suất để tỉ lệ mẫu nằm trong khoảng từ 16% đến 24% là bao nhiêu? (0.905)

Bài 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

Trường Đại học Công nghệ Thông tin

Ngày 16 tháng 4 năm 2023

7.1 Các khái niệm

Bài toán. Một hiệu trưởng của một trường THPT tại TPHCM đọc báo thấy rằng điểm trung bình của bài thi Đánh giá năng lực đợt 1 của Đại học Quốc gia TPHCM năm 2022 là 646,1 điểm. Hiệu trưởng nói rằng điểm trung bình của tất cả các học sinh của trường thi đánh giá năng lực lớn hơn 646,1. Sau đó, một phóng viên chọn ngẫu nhiên 50 học sinh của trường đã thi Đánh giá năng lực và thấy rằng điểm trung bình của nhóm học sinh này là 665. Như vậy, có đủ căn cứ để chấp nhận phát biểu của hiệu trưởng không?

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

Định nghĩa 7.2

1. **Giả thuyết** (null hypothesis), ký 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.

Kết quả của mỗi kiểm định là *chấp nhận H_0 hoặc bác bỏ H_0 và chấp nhận H_1* .

Các dạng toán kiểm định thường gặp:

Kiểm định 2 phía: Giả thuyết $H_0 : \theta = \theta_0$ và đối thuyết $H_1 : \theta \neq \theta_0$.

Kiểm định 1 phía trái: Giả thuyết $H_0 : \theta = \theta_0$ và đối thuyết $H_1 : \theta < \theta_0$.

Kiểm định 1 phía phải: Giả thuyết $H_0 : \theta = \theta_0$ và đối thuyết $H_1 : \theta > \theta_0$.

Ví dụ 7.3

1. Một nhà nghiên cứu nói rằng những trẻ em uống ít nhất 1 ly sữa mỗi ngày khi trưởng thành sẽ có chiều cao lớn hơn 170cm.

Ta kiểm định: Giả thuyết $H_0 : \mu = 170$ và đối thuyết $H_1 : \mu > 170$.

2. Một giám đốc của một doanh nghiệp thấy rằng sau dịch Covid-19 mức lương trung bình của công nhân toàn công ty có thay đổi. Mức lương trung bình trước dịch Covid-19 là 8,2 triệu đồng/tháng.

Ta kiểm định: Giả thuyết $H_0 : \mu = 8,2$ và đối thuyết $H_1 : \mu \neq 8,2$.

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 $H_0 : \mu = 10$ và đối thuyết $H_1 : \mu > 10$.

Các sai lầm trong kiểm định giả thuyết

- Sai lầm loại 1: H_0 đúng nhưng bác bỏ H_0
- Sai lầm loại 2: H_0 sai nhưng chấp nhận H_0

Trong thực tế, sai lầm loại 1 là nguy hiểm hơn, do đó ta thiết kế mô hình kiểm định sao cho xác suất sai lầm loại 1 bị chặn bởi một số rất nhỏ α .

Định nghĩa 7.4 Số α được gọi là **mức ý nghĩa** của kiểm định nếu α là xác suất ta bác bỏ H_0 khi H_0 đúng.

7.2 Kiểm định giả thuyết về trung bình

Giả sử biến ngẫu nhiên X có phân phối chuẩn $N(\mu; \sigma^2)$ trong đó μ là trung bình của tổng thể.

Bài toán 1. Ta kiểm định

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

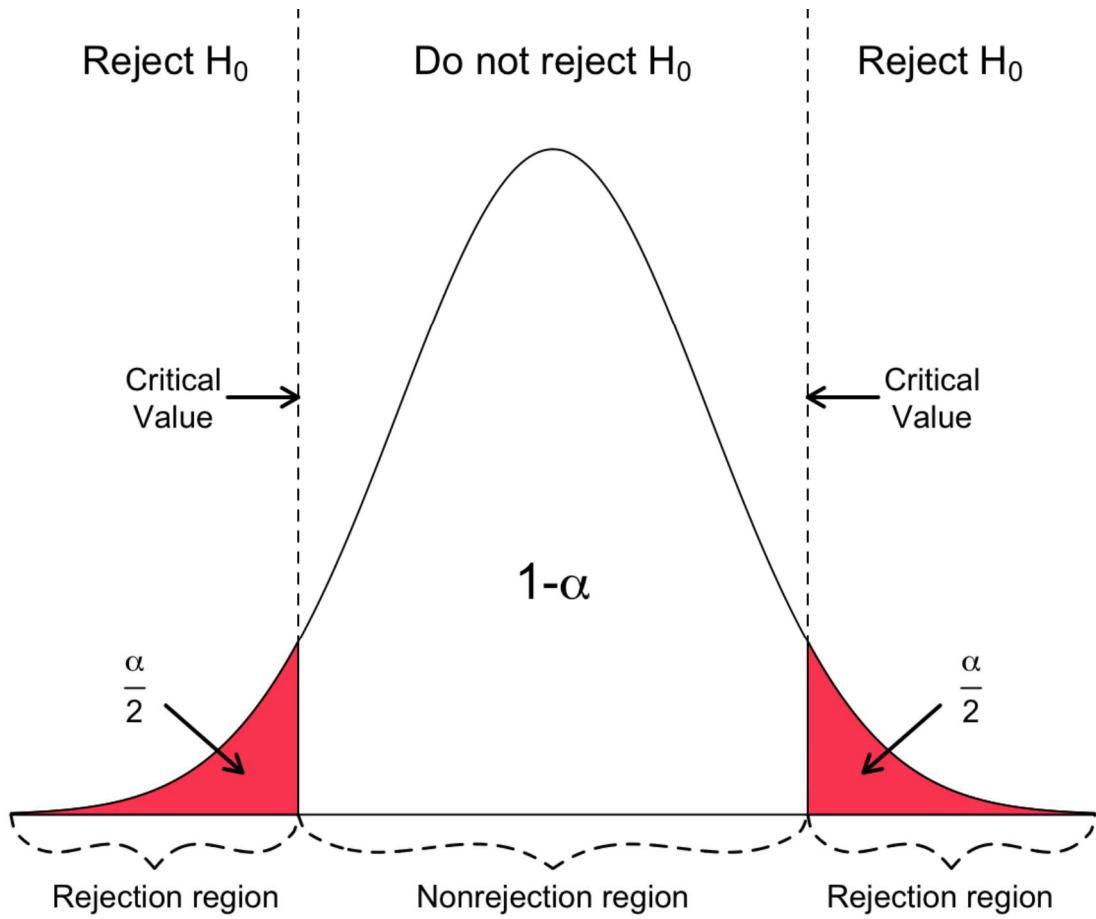
Đặt $Z = \frac{\bar{X} - \mu_0}{\sigma}$ là biến ngẫu nhiên có phân phối chuẩn $N(0; 1)$.

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

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

hay

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



Bài toán 2. 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

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

hay

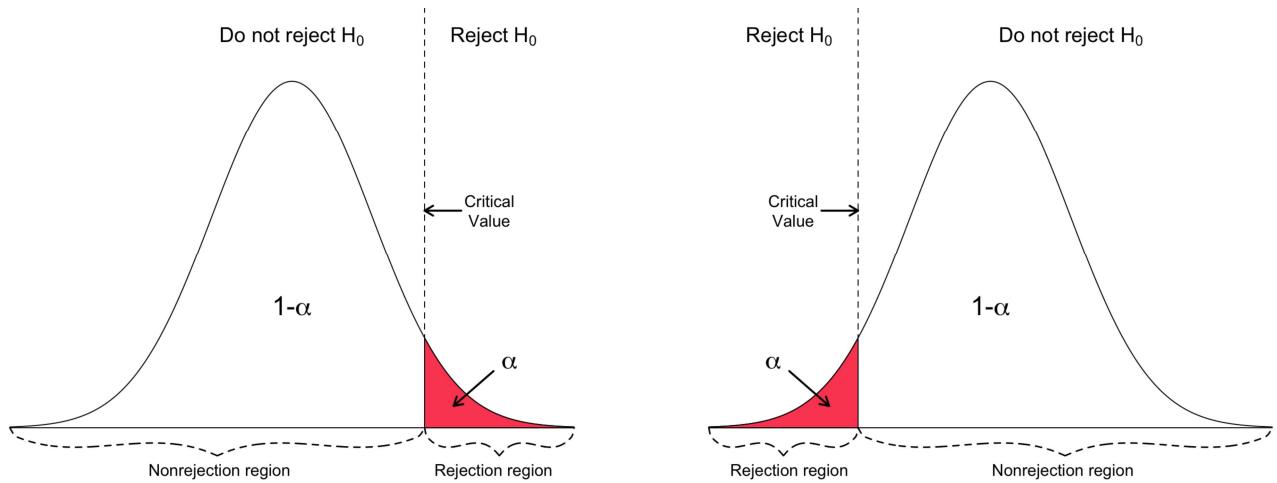
$$\boxed{\Phi(z_\alpha) = P(Z \leq z_\alpha) = 1 - \alpha}$$

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

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



Cho \bar{x} là trung bình mẫu, n là kích thước mẫu, s là độ lệch chuẩn mẫu.

Trường hợp 1. σ đã biết.

$$\text{Tính } z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Trường hợp 2. σ chưa biết và kích thước mẫu $n \geq 30$.

$$\text{Tính } z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Kiểm định	Bắc bỏ H_0	Chấp nhận H_0
$H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0.$	$ z \geq z_{\alpha/2}$	$ z < 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. Đặt

$$T = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \sim St(n)$$

trong đó \bar{X} là trung bình mẫu ngẫu nhiên, n là cỡ mẫu và s là độ lệch chuẩn mẫu.

Với mức ý nghĩa α , đặt $t_{\alpha/2}$ và t_α là các số thực thỏa mãn (xem bảng A5)

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

Đặt

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

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

Kiểm định	Bắc bỏ H_0	Chấp nhận H_0
$H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0.$	$ t \geq t_{\alpha/2}$	$ t < 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 7.5 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 $\bar{x} = 3105$, cỡ mẫu $n = 30$ và độ lệch chuẩn tổng thể $\sigma = 120$
- Vì mức ý nghĩa $\alpha = 0,05$ nên $z_\alpha = 1,65$.
- Đặt

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

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

Ví dụ 7.6 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 kết quả sau (USD/đôi)

60	70	75	55	80	55	50	40	80
70	50	95	120	90	75	85	80	60
110	65	80	85	85	45	75	60	90
90	60	95	110	85	45	90	70	70

Giả sử giá giày có phân phối chuẩn với độ lệch chuẩn là 19,2 USD. Tuyên bố của nhà nghiên cứu có chấp nhận được không với mức ý nghĩa 10%?

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 $\bar{x} = 75$, cỡ mẫu $n = 36$ và độ lệch chuẩn tổng thể $\sigma = 19,2$
- Vì mức ý nghĩa $\alpha = 0,1$ nên $z_\alpha = -1,28$.

- Đặt

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

- Vì $z = -1,56 < -1,28$ nên bác bỏ H_0 .
- Ta đồng ý với nhận xét giá tiền trung bình của một đôi giày thể thao nam ít hơn 80 USD.

Ví dụ 7.7 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. Tuyên bố của nhà nghiên cứu có chấp nhận được không với mức ý nghĩa 10%?

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 $\bar{x} = 71,875$, cỡ mẫu $n = 16$ và độ lệch chuẩn mẫu $s = \dots\dots\dots$
- Vì mức ý nghĩa $\alpha = 0,1$ nên $t_\alpha = \dots\dots\dots$ (bậc tự do 15).
- Đặt

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

- Vì $\dots\dots\dots < t_\alpha$ nên $\dots\dots\dots$
- $\dots\dots\dots$

7.3 Kiểm định giả thuyết về tỉ lệ

Bài toán 4. 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 thể bác bỏ tuyên bố của người ăn kiêng này không?

- Gọi p (f, F) 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 .
- Với n đủ lớn, biến ngẫu nhiên

$$Z = \frac{F - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim N(0; 1)$$

Đặt

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

Với mức ý nghĩa α .

Kiểm định	Bắc bỏ H_0	Chấp nhận H_0
$H_0 : p = p_0; H_1 : p \neq p_0.$	$ z \geq z_{\alpha/2}$	$ z < 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 $f = \frac{128}{200} = 0,64$ và cỡ mẫu $n = 200$.
- Vì mức ý nghĩa $\alpha = 0,05$ nên $z_{\alpha/2} = 1,96$.
- Đặt

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

- Vì $|z| = 1,15 < 1,96$ nên chấp nhận H_0 .
- Ta đồng ý với phát biểu rằng có 60% người không ăn bánh ngọt.

Ví dụ 7.8 Một giáo viên nói rằng lương trung bình của giáo viên tại TPHCM ít hơn 16 triệu đồng/tháng trong năm 2021. Chọn ngẫu nhiên 8 giáo viên thì thấy lương hàng tháng (đơn vị là triệu đồng) của họ trong năm 2021

16 15,6 16 15,5 17 15,5 16 15,5

Giả sử lương của giáo viên có phân phối chuẩn. Với mức ý nghĩa 10%, tuyên bố của giáo viên đó có chấp nhận được không?

Giải.

- Gọi μ tiền lương trung bình hàng tháng của giáo viên trong năm 2021.
 - Kiểm định: Giả thuyết $H_0 : \mu = \dots$ và đối thuyết $H_1 : \mu \dots$
 - Trung bình mẫu là $\bar{x} = \dots$ cỡ mẫu $n = 8$ và độ lệch chuẩn mẫu là $s = \dots$
 - Mức ý nghĩa $\alpha = 0,1$ suy ra (xem bảng A5)
 - Đặt

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

- Vì $t = \dots$ nên $\dots H_0$.
 - Ta \dots với tuyên bố tiền lương trung bình trong một tháng của giáo viên ít hơn 16 triệu đồng.

Ví dụ 7.9 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 7.1 Giám đốc một công ty phát biểu rằng thu nhập trung bình của công nhân trong công ty của ông là hơn 6,7 triệu đồng/tháng. Khảo sát ngẫu nhiên 40 công nhân, người ta thấy rằng thu nhập trung bình của họ là 7,25 triệu đồng/tháng và độ lệch chuẩn mẫu là 1,02 triệu đồng. Với mức ý nghĩa 1%, phát biểu của giám đốc có chấp nhận được không?

Bài 7.2 Một nhà báo nói rằng học phí trung bình của 4 năm học đại học của một sinh viên hơn 11,4 triệu đồng. Cô ấy chọn ngẫu nhiên 36 ngành học 4 năm tại các trường đại học và nhận thấy mức học phí trung bình của 36 ngành này là 11,9 triệu đồng. Biết độ lệch chuẩn của tổng thể là 1,318 triệu đồng. Với mức ý nghĩa 5%, phát biểu của nhà báo đó có chấp nhận được không?

Bài 7.3 Một nhân viên bán hàng tại một cửa hàng laptop nói rằng tuổi thọ trung bình của laptop hiệu Z là 30000 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 7.4 Một báo cáo cho biết rằng trung bình số lần mua hàng online của một phụ nữ trong một tháng là 5,8 lần. Một nhà nghiên cứu chọn ngẫu nhiên 20 phụ nữ và thu được bảng số liệu về số lần mua hàng online trong một tháng như sau

3	2	1	3	7	2	9	4	6	6
8	0	5	6	4	2	1	3	4	1

Giả sử số lần mua hàng online của phụ nữ có phân phối chuẩn. Với mức ý nghĩa 5%, hãy kết luận về báo cáo trên.

Bài 7.5 Một công ty cung cấp dịch vụ internet nói rằng có 40% khách hàng của họ gặp sự cố về đường truyền trong một năm. Một nhóm gồm 100 khách hàng được chọn và người ta thấy rằng có 37 khách hàng gặp sự cố về đường truyền. Với mức ý nghĩa 1%, hãy kết luận về tuyên bố của công ty cung cấp dịch vụ internet.

Bài 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 thuyết rằng quy trình đang vận hành chính xác.

Bài 7.7 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.

Bài 7.8 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%, hãy kết luận về nhận định của báo cáo trên.

Bài 7.9 Market Research, Inc., muốn biết liệu người mua có nhạy cảm với giá của các mặt hàng được bán trong siêu thị hay không. Một mẫu ngẫu nhiên gồm 802 người mua sắm đã được thu thập và 378 người trong số những người mua sắm ở siêu thị đó có thể đưa ra giá chính xác của một mặt hàng ngay sau khi đặt nó vào giỏ hàng của họ. Kiểm định ở mức 7% giả thuyết cho rằng ít nhất một nửa số người mua sắm có thể đưa ra mức giá chính xác.

Chương 8. Tương quan và hồi quy tuyến tính

Nguyễn Minh Trí

Trường Đại học Công nghệ Thông tin

Ngày 16 tháng 4 năm 2023

8.1 Các khái niệm

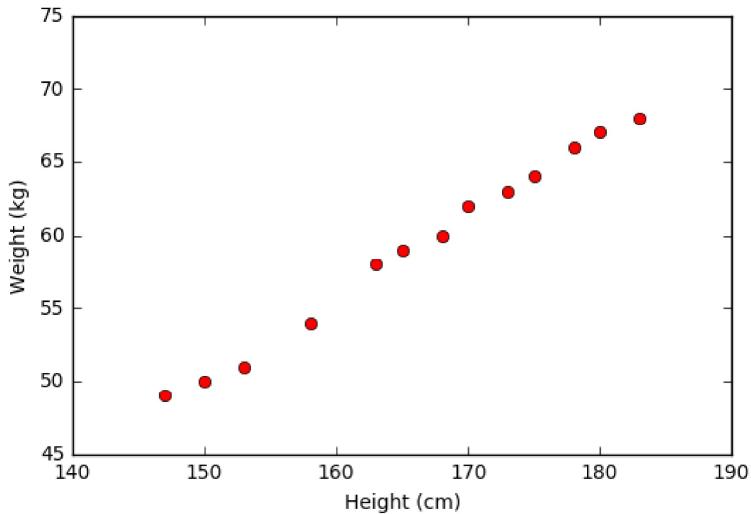
Trong Machine Learning, một trong những thuật toán quan trọng nhất là Thuật toán Hồi quy tuyến tính (Linear Regression) thuộc nhóm *Học có giám sát* (*Supervised Learning*).

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

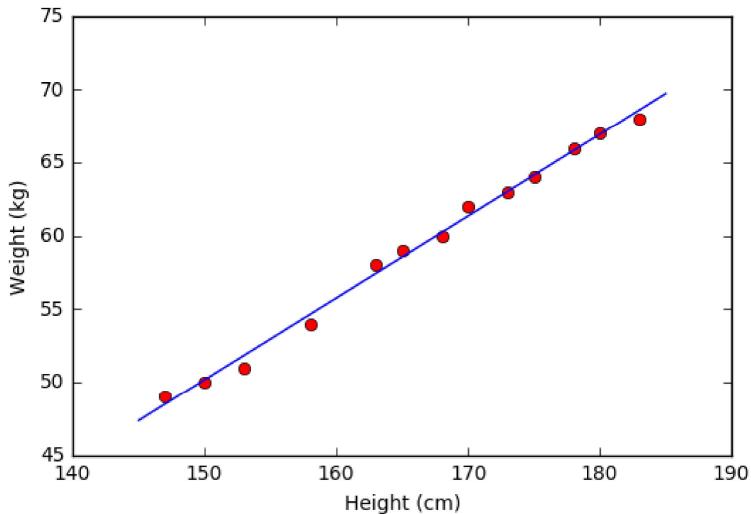


Ta thấy rằng dữ liệu được sắp xếp gần như theo một đường thẳng. Do đó mô hình Hồi quy tuyến tính (Linear Regression) nhiều khả năng sẽ cho kết quả tốt. Ta có thể đưa ra mối liên hệ giữa cân nặng và chiều cao như sau

$$\text{cân nặng} = B \times \text{chiều cao} + A.$$

Bằng các công cụ tính toán, chúng ta sẽ tính được A, B .

Khi đó, các điểm dữ liệu nằm gần đường thẳng mà ta dự đoán.



Sử dụng mô hình này, ta có thể dự đoán cân nặng của một người có chiều cao 155cm, 160 cm hoặc 171cm.

8.2 Hệ số tương quan

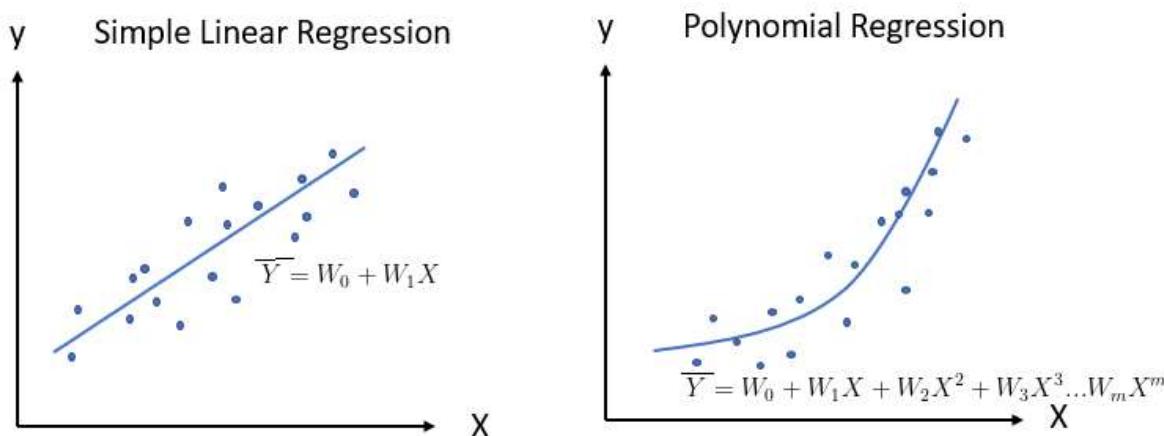
Xét vectơ ngẫu nhiên (X, Y) và tập n giá trị cụ thể $(x_1, y_1), \dots, (x_n, y_n)$. Các cặp giá trị này được gọi là **dữ liệu thực nghiệm**.

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).

Có nhiều kiểu phụ thuộc giữa hai biến ngẫu nhiên X và Y nhưng phổ biến nhất là dạng phụ thuộc hàm số $Y = f(X)$. Một trong những hàm đơn giản nhất là hàm số bậc nhất $Y = aX + b$ hay dạng tuyến tính.

Đường cong phù hợp là một đường cong xấp xỉ tốt nhất (ít sai lệch nhất) với các điểm dữ liệu đã cho.

- Nếu đường cong phù hợp là một đường thẳng thì ta có một **quan hệ tuyến tính** (linear relation) giữa hai biến ngẫu nhiên.
- Nếu đường cong phù hợp **không** là một đường thẳng thì ta có một **quan hệ phi tuyến tính** giữa hai biến ngẫu nhiên.



Bài toán.

1. Có một quan hệ tuyến tính hoặc phi tuyến tính giữa hai biến ngẫu nhiên không?
2. Nếu có một quan hệ tuyến tính (phi tuyến tính) giữa hai biến ngẫu nhiên thì có thể biểu diễn mối quan hệ này dưới dạng một hàm số không?

Ta cần một số đo để đo mức độ chặt chẽ trong quan hệ tuyến tính giữa hai biến ngẫu nhiên.

Định nghĩa 8.1 Hệ số tương quan mẫu của hai biến ngẫu nhiên X, Y

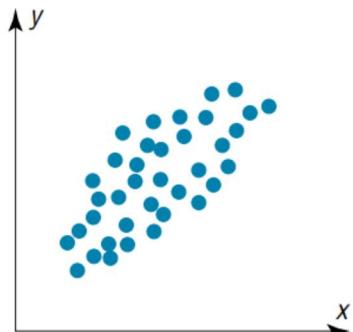
$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

Hay

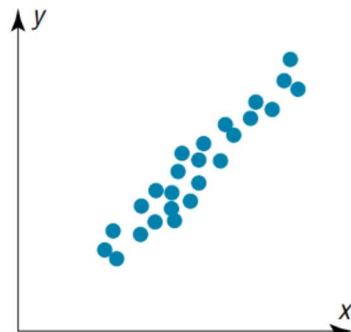
$$r = \frac{n(\sum_{i=1}^n x_i y_i) - (\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n y_i)}{\sqrt{(n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2)(n(\sum_{i=1}^n y_i^2) - (\sum_{i=1}^n y_i)^2)}}$$

trong đó n là số cặp điểm dữ liệu thực nghiệm.

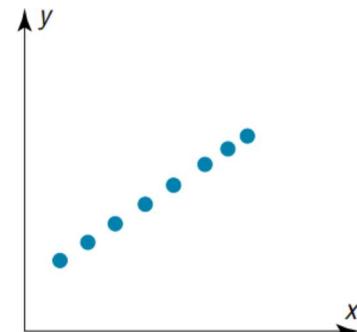
- Ta có $-1 \leq r \leq 1$.
- Nếu $0,8 \leq |r| \leq 1$ thì ta nói X, Y có tương quan tuyến tính mạnh.
- Nếu $|r| < 0,8$ thì ta nói X, Y có tương quan tuyến tính yếu.
- Nếu r gần bằng 1 thì ta nói có sự tương quan tuyến tính thuận giữa X và Y .
- Nếu r gần bằng -1 thì ta nói có sự tương quan tuyến tính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



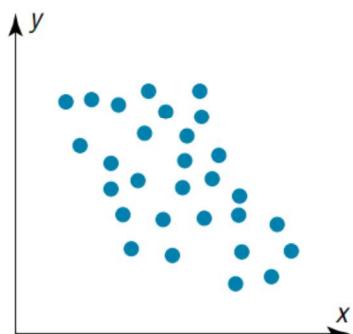
(a) $r = 0.50$



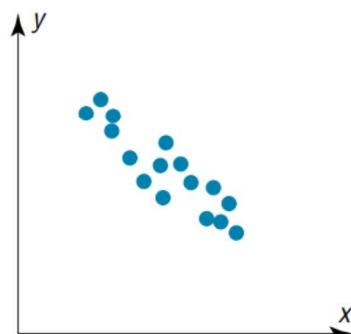
(b) $r = 0.90$



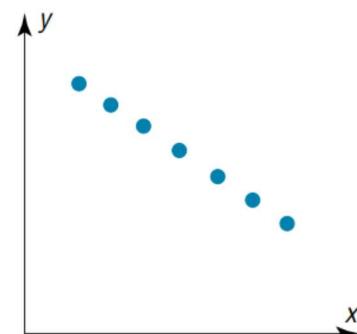
(c) $r = 1.00$



(d) $r = -0.50$



(e) $r = -0.90$



(f) $r = -1.00$

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.

Ta có

$$\bar{xy} = \frac{6 \cdot 8,2 + 2 \cdot 8,6 + 15 \cdot 4,3 + 9 \cdot 7,4 + 12 \cdot 5,8 + 5 \cdot 9,0 + 8 \cdot 7,8}{7} = 53,5$$

$$\bar{x} = \frac{6 + 2 + 15 + 9 + 12 + 5 + 8}{7} = 8,14$$

$$\bar{y} = \frac{8,2 + 8,6 + 4,3 + 7,4 + 5,8 + 9,0 + 7,8}{7} = 7,3$$

$$\bar{x^2} = \frac{6^2 + 2^2 + 15^2 + 9^2 + 12^2 + 5^2 + 8^2}{7} = 82,71$$

$$\bar{y^2} = \frac{8,2^2 + 8,6^2 + 4,3^2 + 7,4^2 + 5,8^2 + 9,0^2 + 7,8^2}{7} = 55,7$$

Do đó, hệ số tương quan là

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

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.

8.3 Hồi quy

Bài toán. Ta muốn khảo sát xem số buổi nghỉ học có ảnh hưởng đến điểm thi cuối kỳ của môn xác suất thống kê. Nếu biết số buổi nghỉ học thì ta có thể dự đoán điểm thi cuối kỳ được không?

- Mục đích của hồi quy là dự đoán một đại lượng này từ các đại lượng khác.
- Nếu biến Y được ước lượng từ biến X bằng một biểu thức $Y = f(X)$ thì biểu thức này được gọi là **phương trình hồi quy** của Y theo X .
- Đường cong biểu diễn đường $Y = f(X)$ được gọi là **đường cong hồi quy** của Y theo X .
- Đường thẳng biểu diễn đường $Y = A + BX$ (phương trình hồi quy tuyến tính) được gọi là **đường thẳng hồi quy** của Y theo X .

Trong việc nghiên cứu mối liên hệ giữa hai biến:

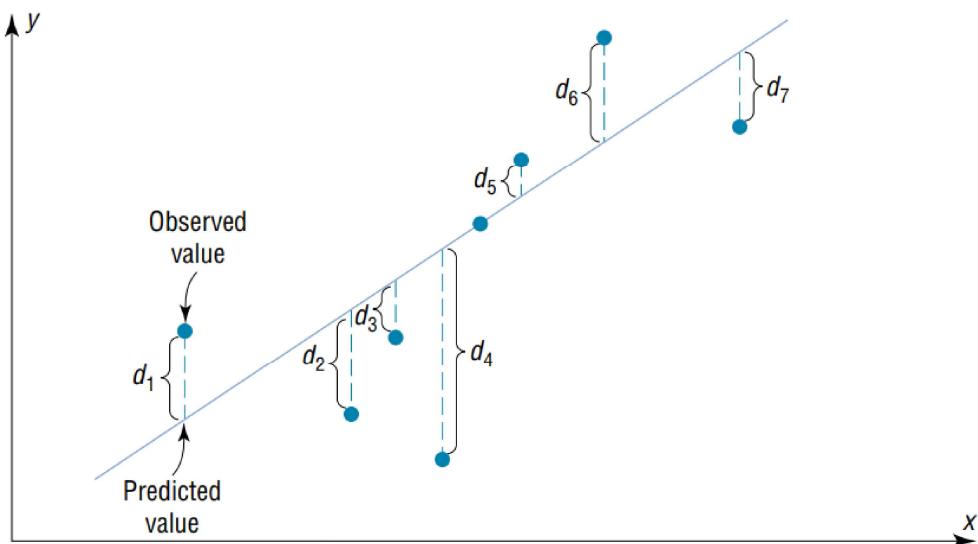
1. Thu thập dữ liệu và xây dựng biểu đồ phân tán
2. Tính hệ số tương quan r
3. Kiểm tra sự tương quan tuyến tính giữa hai biến
4. Nếu $|r|$ gần bằng 1 thì ta sẽ xác định đường thẳng hồi quy (regression line) (đường thẳng phù hợp nhất).
5. Đường thẳng hồi quy giúp các nhà nghiên cứu có thể nhìn thấy xu hướng và đưa ra các dự báo.

Cho các điểm $(x_1, y_1), \dots, (x_n, y_n)$, ta sẽ tìm phương trình đường thẳng (**hồi quy tuyến tính**) $Y = A + BX$, sao cho

$$\sum_{i=1}^n (y_i - (A + BX_i))^2$$

là nhỏ nhất.

Phương pháp trên được gọi là phương pháp *bình phương cực tiểu* (method of least squares).



Khi đó

$$B = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - \bar{x}^2} \text{ và } A = \bar{y} - B\bar{x}.$$

Ví dụ 8.3 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 phương trình đường thẳng hồi quy tuyến tính và dự đoán điểm của sinh viên chỉ vắng 1 buổi học.

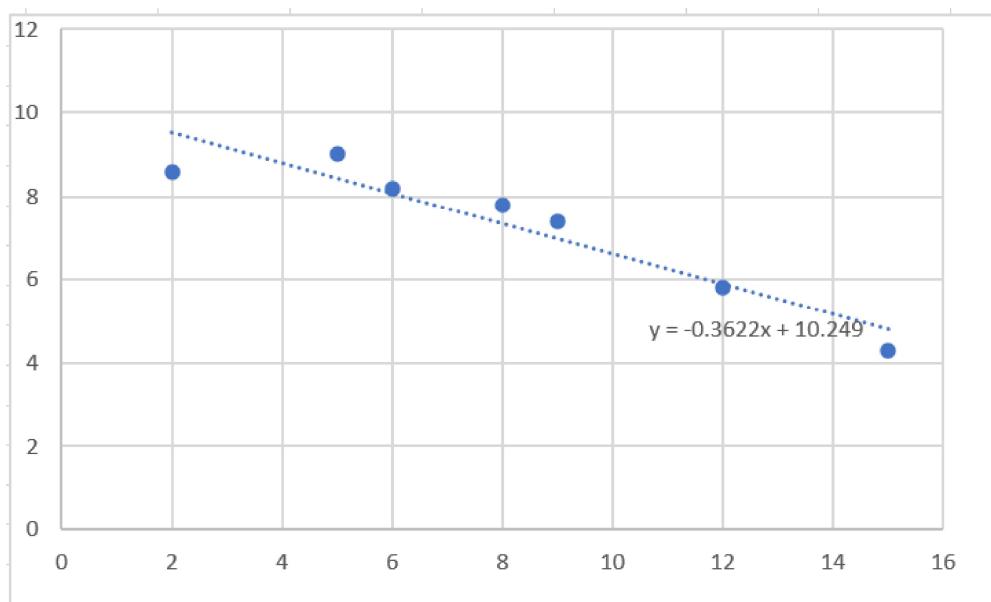
Giải.

- Phương trình hồi quy tuyến tính $Y = A + BX$.
- $B = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - \bar{x}^2} = \dots\dots\dots$
- $A = \bar{y} - B\bar{x} = \dots\dots\dots$
- Phương trình đường thẳng hồi quy tuyến tính cần tìm là $\dots\dots\dots$
- Khi $X = 1$ thì $Y = \dots$

Giải. Dùng máy tính CASIO fx-570VN-PLUS

- $\boxed{SHIFT} \rightarrow \boxed{MODE} \rightarrow \boxed{\triangledown} \rightarrow$ chọn STAT (trên màn hình - phím 4)
- Màn hình xuất hiện **Frequency**, chọn **OFF**
- $\boxed{SHIFT} \rightarrow \boxed{MODE} \rightarrow \boxed{\text{Data}} (\text{phím } \boxed{2})$
- Nhập dữ liệu cột X : $\boxed{6} \boxed{=} \boxed{2} \boxed{=}$...
- Nhập dữ liệu cột Y : $\boxed{8.2} \boxed{=} \boxed{8.6} \boxed{=}$...
- \boxed{ON}
- $\boxed{SHIFT} \rightarrow \boxed{1} \rightarrow \boxed{\text{Reg}} (\text{phím } \boxed{5})$
- Chọn **A** (phím $\boxed{1}$) $\boxed{=}$
- \boxed{ON}
- $\boxed{SHIFT} \rightarrow \boxed{1} \rightarrow \boxed{\text{Reg}} (\text{phím } \boxed{5})$
- Chọn **B** (phím $\boxed{2}$) $\boxed{=}$

Khi đó $A = 10,2493$ và $B = -0,3722$. Đường thẳng hồi quy tuyến tính là $Y = 10,2493 - 0,3622X$.



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.

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**.

Một vài lưu ý

- Đường thẳng hồi quy tuyến tính theo phương pháp bình phương tối thiểu luôn đi qua điểm (\bar{x}, \bar{y})
- Khi tính toán cần xác định rõ biến độc lập và biến phụ thuộc
 - ▶ Phương trình hồi quy tuyến tính của Y theo X
 - ▶ Phương trình hồi quy tuyến tính của X theo Y

$$Y = A + BX$$

▶ Phương trình hồi quy tuyến tính của X theo Y

$$X = A + BY$$

Ví dụ 8.4 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 USD

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 USD.

Giải.

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

BÀI TẬP

Bài 8.1 Lợi nhuận của 7 công ty cho thuê xe 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

- Tính hệ số tương quan giữa số xe cho thuê và lợi nhuận hàng năm.
 - 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 8.2** PCWorld đã đánh giá bốn đặc điểm thành phần cho 10 máy tính xách tay siêu di động: tính năng, hiệu suất, thiết kế và giá cả. Mỗi đặc điểm được đánh giá bằng thang điểm 0-100. Xếp hạng tổng thể, được gọi là PCW World Rating, sau đó được phát triển cho từng máy tính xách tay. Bảng sau đây cho thấy xếp hạng tính năng và PCW World Rating cho 10 máy tính xách tay (trang web PC World, ngày 5 tháng 2 năm 2009).

Model	Hạng tính năng (X)	PCW World Rating (Y)
Thinkpad X200	87	83
VGN-Z598U	85	82
U6V	80	81
Elitebook 2530P	75	78
X360	80	78
Thinkpad X300	76	78
Ideapad U110	81	77
Micro Express JFT2500	73	75
Toughbook W7	79	73
HP Voodoo Envy133	68	72

- 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 PCW World Rating của một laptop mới mà nó có hạng tính năng là 70.
- Bài 8.3** Một nghiên cứu đã được thực hiện về lượng đường được chuyển đổi trong một quy trình nhất định ở các nhiệt độ khác nhau. Dữ liệu được mã hóa và ghi lại như sau:

Nhiệt độ (X)	Đường được chuyển đổi (Y)	Nhiệt độ (X)	Đường được chuyển đổi (Y)
1,0	8,1	1,6	8,6
1,1	7,8	1,7	10,2
1,2	8,5	1,8	9,3
1,3	9,8	1,9	9,2
1,4	9,5	2,0	10,5
1,5	8,9		

- a. Tìm đường thẳng hồi quy tuyến tính.
- b. Ước tính lượng đường chuyển đổi được tạo ra khi nhiệt độ được mã hóa là 1,75.
- Bài 8.4** Thời gian sử dụng liên tục của 8 loại điện thoại Y (giờ) và số mAh X (nghìn chiếc) ghi trên pin của điện thoại được khảo sát như sau

Điện thoại	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.