

# NHẬP MÔN KHOA HỌC DỮ LIỆU

## KIẾN THỨC CƠ SỞ

ThS. Vũ Hoài Thư



# Nội dung

## 1 Ôn tập về đại số tuyến tính

- Ma trận
- Định thức
- Ma trận nghịch đảo

## 2 Ôn tập về xác suất

- Sự kiện ngẫu nhiên và quan hệ giữa các sự kiện
- Xác suất có điều kiện và quy tắc Bayes
- Biến ngẫu nhiên và phân loại biến ngẫu nhiên
- Phân phối chuẩn

## 3 Ôn tập về thống kê

- Khoảng tin cậy
- Kiểm định giả thuyết thống kê

# Ôn tập về đại số tuyến tính

# Định nghĩa ma trận

Ma trận cỡ  $m \times n$  là một bảng số hình chữ nhật có  $m$  hàng và  $n$  cột, được ký hiệu là:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Trong đó  $a_{ij}$  là phần tử nằm trên hàng thứ  $i$ , cột thứ  $j$  của ma trận  $A$ .

- Ma trận  $A$  còn được ký hiệu gọn là  $A = (a_{ij})_{m \times n}$
- Nếu  $m = n$  thì  $A$  được gọi là *ma trận vuông cấp  $n$* . Khi đó  $A = (a_{ij})_{n \times n}$
- Nếu  $n = 1$  thì  $A$  được gọi là *ma trận cột*, hay một *vector*

# Ma trận vuông

Cho ma trận vuông  $A = (a_{ij})_{n \times n}$ . Đường chéo  $a_{11}, a_{22}, \dots, a_{nn}$  được gọi là đường chéo chính.

- Ma trận vuông mà mọi phần tử nằm phía dưới đường chéo chính đều bằng 0 (tức là  $a_{ij} = 0, \forall i > j$ ) gọi là ma trận tam giác trên.
- Ma trận vuông mà mọi phần tử nằm phía trên đường chéo chính đều bằng 0 (tức là  $a_{ij} = 0, \forall i < j$ ) gọi là ma trận tam giác dưới.
- Ma trận vuông mà mọi phần tử không nằm trên đường chéo chính đều bằng 0 (tức là  $a_{ij} = 0, \forall i \neq j$ ) gọi là ma trận đường chéo, ký hiệu  $\text{diag}(a_{11}, a_{22}, \dots, a_{nn})$ .
- Ma trận đường chéo mà mọi phần tử của đường chéo chính đều bằng 1 được gọi là ma trận đơn vị, ký hiệu là  $I$ .

# Các phép toán với ma trận

## Phép cộng hai ma trận

Cho hai ma trận cùng cỡ  $A = (a_{ij})_{m \times n}$ ,  $B = (b_{ij})_{m \times n}$ . Tổng của hai ma trận  $A$  và  $B$ , ký hiệu  $A + B$ , là ma trận xác định bởi:

$$A + B = (a_{ij} + b_{ij})_{m \times n}$$

## Phép nhân ma trận với một số

Cho ma trận  $A = (a_{ij})_{m \times n}$  và số thực  $k$ . Tích của ma trận  $A$  với số  $k$ , ký hiệu  $kA$ , là ma trận xác định bởi:

$$kA = (ka_{ij})_{m \times n}$$

# Các phép toán với ma trận

## Phép nhân hai ma trận

Cho ma trận  $A = (a_{ij})_{m \times p}$ ,  $B = (b_{ij})_{p \times n}$ . Tích của ma trận  $A$  và  $B$ , ký hiệu  $AB$ , là ma trận  $AB = (c_{ij})_{m \times n}$  xác định bởi:

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \dots + a_{ip}b_{pj} = \sum_{k=1}^p a_{ik}b_{kj}$$

## Ma trận chuyển vị

Ma trận thu được từ ma trận  $A$  bằng cách chuyển các hàng thành các cột (các cột chuyển thành các hàng) được gọi là *ma trận chuyển vị* của ma trận  $A$ , ký hiệu là  $A^T$ .

# Định nghĩa

## Ma trận con

Xét ma trận vuông cấp  $n$ :  $A = (a_{ij})_{nn}$ . Phần phụ đại số của  $a_{ij}$ , ký hiệu là  $A_{ij}$  được xác định như sau:

$$A_{ij} = (-1)^{i+j} \det(M_{ij})$$

Trong đó  $M_{ij}$  là ma trận thu được từ ma trận  $A$  bằng cách bỏ đi hàng thứ  $i$  và cột thứ  $j$ .

## Định thức

Định thức của ma trận  $A$ , ký hiệu  $\det(A)$  hoặc  $|A|$ , được định nghĩa theo quy nạp như sau:

- $A$  là ma trận cấp 1:  $A = (a_{11})$  thì  $\det(A) = a_{11}$ .
- $A$  là ma trận cấp  $n$  thì:

$$\det(A) = a_{11}A_{11} + a_{12}A_{12} + \dots + a_{1n}A_{1n}$$



# Tính chất của Định thức

- (i)  $\det(A) = \det(A^T)$
- (ii) Nếu ma trận  $B$  thu được từ ma trận  $A$  bởi phép đổi chỗ hai hàng cho nhau thì  $\det(B) = -\det(A)$
- (iii) Một định thức có hai hàng (hoặc hai cột) như nhau thì bằng 0.
- (iv) Khai triển định thức theo hàng  $i$ :  
$$\det(A) = a_{i1}A_{i1} + a_{i2}A_{i2} + \dots + a_{in}A_{in}$$

Khai triển định thức theo cột  $j$ :

$$\det(A) = a_{1j}A_{1j} + a_{2j}A_{2j} + \dots + a_{nj}A_{nj}$$

# Tính chất của Định thức (Tiếp)

- (v) Một định thức có một hàng (hay một cột) toàn là số không thì bằng 0.
- (vi) Khi nhân các phần tử của một hàng (hay một cột) với cùng một số  $k$  thì được một định thức mới bằng định thức cũ nhân với  $k$ .
- (vii) Một định thức có hai hàng (hay hai cột) tỷ lệ thì bằng không.
- (viii) Khi tất cả các phần tử của một hàng (hay một cột) có dạng tổng của hai số hạng thì định thức có thể phân thành tổng của hai định thức, chẳng hạn:

$$\begin{vmatrix} a_{11} & a_{12} + a'_{12} \\ a_{21} & a_{22} + a'_{22} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} + \begin{vmatrix} a_{11} & a'_{12} \\ a_{21} & a'_{22} \end{vmatrix}$$

# Tính chất của Định thức (Tiếp)

- (ix) Nếu định thức có một hàng (hay một cột) là tổ hợp tuyến tính của các hàng khác (hay các cột khác) thì định thức ấy bằng không.
- (x) Khi cộng bội  $k$  của một hàng vào một hàng khác (hay bội  $k$  của một cột vào một cột khác) thì được một định thức mới bằng định thức cũ.

## Ví dụ

Tính định thức của ma trận:

$$A = \begin{bmatrix} -1 & 1 & 2 & 0 \\ 3 & -1 & 0 & 4 \\ -2 & 0 & 5 & -2 \\ 0 & 3 & 6 & 1 \end{bmatrix}$$

# Định nghĩa ma trận nghịch đảo

## Ma trận nghịch đảo

Giả sử  $A$  là ma trận vuông cấp  $n$ ,  $I$  là ma trận đơn vị cùng cấp. Ma trận  $A$  được gọi là khả nghịch nếu tồn tại ma trận vuông  $B$  sao cho  $AB = BA = I$ . Khi đó  $B$  được gọi là ma trận nghịch đảo của ma trận  $A$ , ký hiệu là  $A^{-1}$ .

## Ma trận phụ hợp

Xét ma trận  $A = (a_{ij})_{n \times n}$ . Đặt  $A_{ij} = (-1)^{i+j} \det(M_{ij})$ , trong đó  $M_{ij}$  là ma trận con ứng với phần tử  $a_{ij}$ . Khi đó  $P_A = (A_{ij})_{n \times n}$  được gọi là ma trận phụ hợp của ma trận  $A$ .

# Định lý

## Định lý

Ma trận vuông  $A$  có ma trận nghịch đảo  $A^{-1}$  khi và chỉ khi  $\det(A) \neq 0$ . Khi đó:

$$A^{-1} = \frac{1}{\det(A)} P_A^T = \frac{1}{\det(A)} \begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \dots & \dots & \dots & \dots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{bmatrix}$$

## Ví dụ

Tìm ma trận nghịch đảo của ma trận:

$$A = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 0 & 5 \\ 3 & 1 & 2 \end{bmatrix}$$

# Bài tập 1

Viết chương trình Python thực hiện các yêu cầu sau:

- (a) Kiểm tra một ma trận bất kỳ có phải là ma trận tam giác trên, dưới, đường chéo không?
- (b) Tính tổng hai ma trận, nhân ma trận với một số, nhân hai ma trận?
- (c) Tính định thức của một ma trận, tính ma trận nghịch đảo của một ma trận?

**Chú ý:** Mỗi yêu cầu thực hiện theo hai cách:

- Cách 1: code trực tiếp theo định nghĩa, định lý, cách xây dựng.
- Cách 2: sử dụng thư viện từ python.

# Ôn tập về xác suất

# Các định nghĩa cơ bản

- (i) Phép thử là việc thực hiện một thí nghiệm hoặc quan sát theo dõi một hiện tượng thực tế trong từng trường hợp cụ thể với những điều kiện cho trước.
- (ii) Hiện tượng có thể xảy ra hoặc không xảy ra trong kết quả của phép thử được gọi là sự kiện hay biến cố.
- (iii) Sự kiện ngẫu nhiên là sự kiện có thể xảy ra hoặc không xảy ra khi phép thử được thực hiện và ký hiệu là  $A, B, C, \dots$
- (iv) Trong một phép thử, tập hợp tất cả các kết quả có thể xảy ra được gọi là không gian mẫu và được ký hiệu là  $\Omega$ .



# Quan hệ giữa các sự kiện

- (i) Hai sự kiện  $A$  và  $B$  được gọi xung khắc với nhau nếu chúng không đồng thời xảy ra trong cùng một phép thử, ký hiệu  $A \cap B = \emptyset$ .
- (ii) Họ các sự kiện  $A_1, A_2, \dots, A_n$  được gọi là xung khắc từng đôi nếu một sự kiện bất kỳ trong họ xảy ra khi các sự kiện còn lại không xảy ra, tức là  $A_i \cap A_j = \emptyset, \forall i \neq j$ .
- (iii) Hai sự kiện  $A$  và  $B$  được gọi là độc lập với nhau nếu sự kiện này xảy ra hay không xảy ra không làm ảnh hưởng tới khả năng xảy ra của sự kiện kia. Ngược lại thì chúng phụ thuộc.

## Quan hệ giữa các sự kiện (Tiếp)

- (vi) Họ các sự kiện  $A_1, A_2, \dots, A_n$  được gọi là độc lập từng đôi nếu  $A_i, A_j$  độc lập với mọi  $i \neq j$ .
- (vii) Nhóm  $n$  sự kiện  $A_1, A_2, \dots, A_n$  được gọi là nhóm đầy đủ các sự kiện nếu nhất định phải xảy ra một và chỉ một sự kiện trong các sự kiện ấy sau phép thử. Tức là  $A_i \cap A_j = \emptyset, \forall i \neq j$  và  $A_1 \cup A_2 \cup \dots \cup A_n = \Omega$ .

# Xác suất có điều kiện

## Xác suất có điều kiện

Xác suất của sự kiện  $A$  trong điều kiện biết sự kiện  $B$  đã xảy ra được gọi là xác suất của  $A$  với điều kiện  $B$ . Ký hiệu là  $P(A|B)$ . Khi đó:

$$P(A|B) = \frac{P(AB)}{P(B)}, \text{ với } P(B) > 0$$

## Ví dụ

Một ngân hàng đề thi có cấu trúc như sau:

	Dễ	Khó
Lý thuyết	20	10
Bài tập	30	40

Bốc ngẫu nhiên một câu hỏi. Nếu biết bốc được câu bài tập, tính xác suất bốc được câu dễ.

# Công thức nhân xác suất

Cho  $A$  và  $B$  là hai sự kiện bất kỳ. Khi đó:

$$P(AB) = P(A).P(B|A) = P(B).P(A|B)$$

- Nếu  $A$  và  $B$  độc lập thì  $P(AB) = P(A).P(B)$ .
- Nếu  $A$  và  $B$  xung khắc thì  $P(AB) = 0$ .
- Mở rộng cho tích  $n$  sự kiện bất kỳ  $A_1, A_2, \dots, A_n$ :

$$P(A_1A_2\dots A_n) = P(A_1).P(A_2|A_1)\dots P(A_n|A_1A_2\dots A_{n-1})$$

# Ví dụ

Có 4 que thăm, trong đó có 3 que thăm dài bằng nhau và 1 que thăm ngắn hơn. Bốn người lần lượt lên rút ngẫu nhiên một que thăm. Tính xác suất người thứ  $i$  rút được thăm ngắn ( $i = 1, 2, 3, 4$ ).

# Công thức cộng xác suất

- Nếu  $A$  và  $B$  là hai sự kiện bất kỳ thì:

$$P(A + B) = P(A) + P(B) - P(AB)$$

- Nếu  $A$  và  $B$  là hai sự kiện xung khắc thì:

$$P(A + B) = P(A) + P(B)$$

- Nếu  $A, B, C$  là ba sự kiện bất kỳ thì:

$$P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

- Nếu  $\overline{A}$  là sự kiện đối ngẫu của  $A$  thì  $P(\overline{A}) = 1 - P(A)$

# Ví dụ

Một lớp có 100 sinh viên, trong đó có 30 sinh viên giỏi tiếng Anh, 40 sinh viên giỏi xác suất, 10 sinh viên giỏi cả tiếng Anh lẫn xác suất. Chọn ngẫu nhiên một sinh viên trong lớp. Tìm xác suất để sinh viên đó giỏi ít nhất 1 trong 2 môn trên.

# Công thức xác suất đầy đủ

Cho các sự kiện  $A_1, A_2, \dots, A_n$  là một nhóm đầy đủ các sự kiện của một phép thử ngẫu nhiên,  $B$  là một sự kiện ngẫu nhiên bất kỳ. Khi đó:

$$P(B) = \sum_{i=1}^n P(A_i B) = \sum_{i=1}^n P(A_i) P(B|A_i)$$

## Ví dụ

Có 3 hộp đựng bi. Hộp I đựng 4 bi trắng và 6 bi đen, hộp II đựng 3 bi trắng và 6 bi đen, hộp III đựng 5 bi trắng và 3 bi đen. Lấy ngẫu nhiên một hộp, rồi từ hộp đó lấy ngẫu nhiên một viên bi. Tìm xác suất để được bi trắng.



# Công thức xác suất Bayes

Cho các sự kiện  $A_1, A_2, \dots, A_n$  là một nhóm đầy đủ các sự kiện và  $B$  là một sự kiện bất kỳ nào đó. Khi đó xác suất có điều kiện của sự kiện  $A_k, k = 1, 2, \dots, n$  với điều kiện  $B$  được xác định bởi:

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^n P(A_i)P(B|A_i)}$$

## Ví dụ

Hai xạ thủ bắn mỗi người một phát đạn vào mục tiêu. Xác suất bắn trúng của từng người lần lượt là 0.6 và 0.7. Xác suất mục tiêu bị tiêu diệt khi trúng 1 phát đạn là 0.5 và khi trúng 2 phát là 0.8.

- Tính xác suất mục tiêu bị tiêu diệt.
- Giả sử mục tiêu bị tiêu diệt. Tính xác suất chỉ có một xạ thủ bắn.

# Định nghĩa

## Biến ngẫu nhiên

*Biến ngẫu nhiên* là biến nhận giá trị là các số thực phụ thuộc vào kết cục của một phép thử ngẫu nhiên.

Ta thường dùng các chữ cái hoa  $X, Y, Z, \dots$  để chỉ các biến ngẫu nhiên và các chữ cái thường  $x, y, z, \dots$  để chỉ các giá trị cụ thể mà biến ngẫu nhiên đó nhận.

## Phân loại biến ngẫu nhiên

Biến ngẫu nhiên được phân làm hai loại:

- a) **Biến ngẫu nhiên rời rạc:**  $X$  là biến ngẫu nhiên rời rạc nếu tập giá trị của nó là tập hợp hữu hạn hoặc vô hạn đếm được phần tử.
- b) **Biến ngẫu nhiên liên tục:**  $X$  là biến ngẫu nhiên liên tục nếu tập giá trị của nó lấp đầy một khoảng trên trục số.

# Ví dụ

- a) Gọi  $X$  là tổng số chấm xuất hiện khi tung đồng thời hai con xúc xắc thì  $X$  là một biến ngẫu nhiên rời rạc có thể nhận các giá trị từ 2 đến 12.
- b) Tại một bến xe buýt cứ 20 phút lại có một chuyến xe. Một người tới bến xe tại một thời điểm nào đó. Gọi  $Z$  là thời gian người đó phải chờ xe. Khi đó  $Z$  là một biến ngẫu nhiên liên tục có thể nhận các giá trị thuộc  $[0, 20)$ .

# Phân phối chuẩn

- Biến ngẫu nhiên  $X$  được gọi là tuân theo phân phối chuẩn với tham số  $\mu, \sigma^2$ , ký hiệu là  $X \sim N(\mu, \sigma^2)$ , nếu hàm mật độ xác suất của  $X$  có dạng:

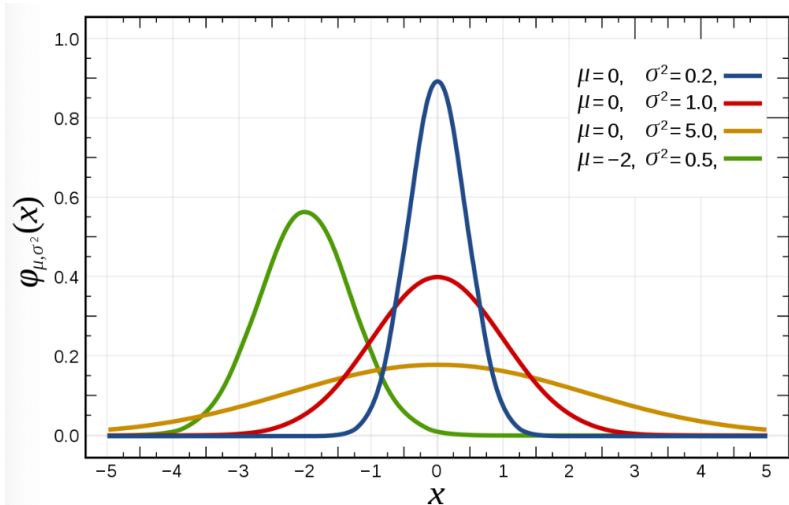
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \forall x \in \mathbf{R}$$

- Kỳ vọng và phương sai của biến ngẫu nhiên  $X$  tuân theo luật phân phối chuẩn là:

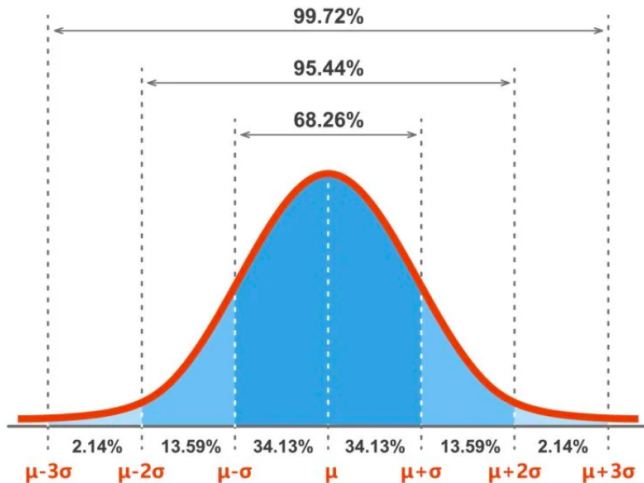
$$E(X) = \mu, \quad D(X) = \sigma^2$$

- Độ lệch chuẩn là  $\sigma(X) = \sqrt{D(X)} = \sigma$

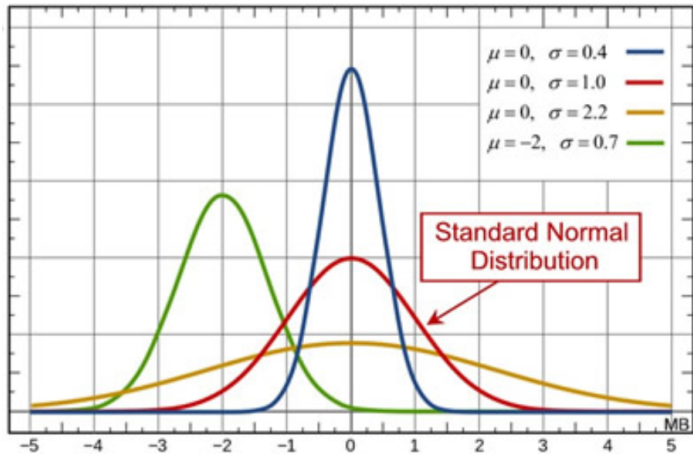
# Phân phối chuẩn (Tiếp)



# Phân phối chuẩn (Tiếp)



# Phân phối chuẩn tắc



## Phân phối chuẩn tắc (Tiếp)

- Phân phối chuẩn  $N(\mu, \sigma^2)$  với  $\mu = 0$  và  $\sigma = 1$  gọi là phân phối chuẩn tắc  $N(0, 1)$ .
- Hàm mật độ xác suất của biến ngẫu nhiên  $U$  có phân phối chuẩn tắc là:

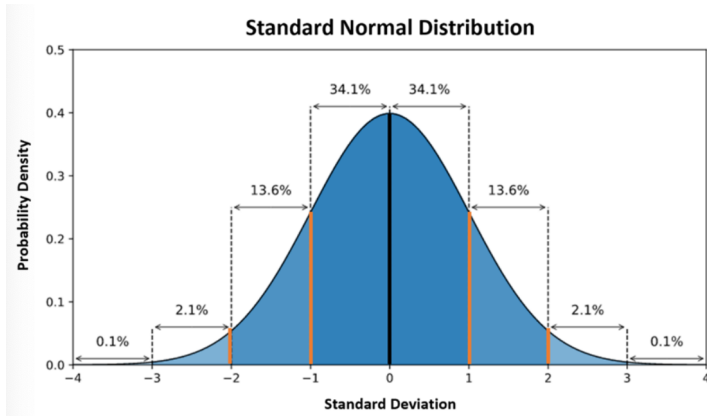
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbf{R}$$

- Hàm phân phối xác suất của biến ngẫu nhiên  $U$  có phân phối chuẩn tắc là:

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt, \quad x \in \mathbf{R}$$



# Phân phối chuẩn tắc (Tiếp)



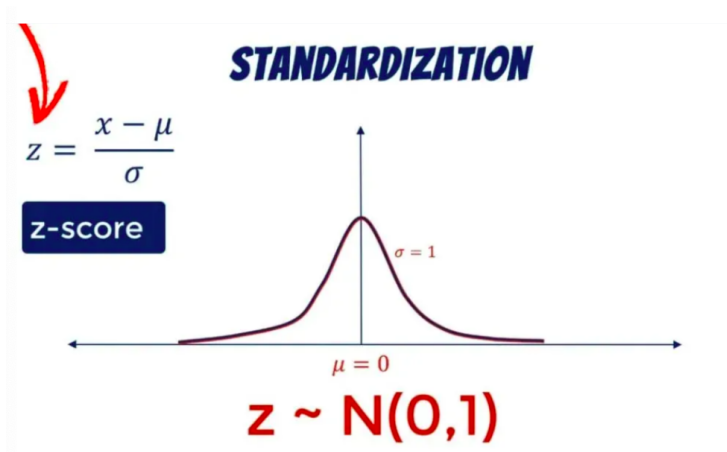
# Chuẩn hoá dữ liệu

## STANDARDIZATION

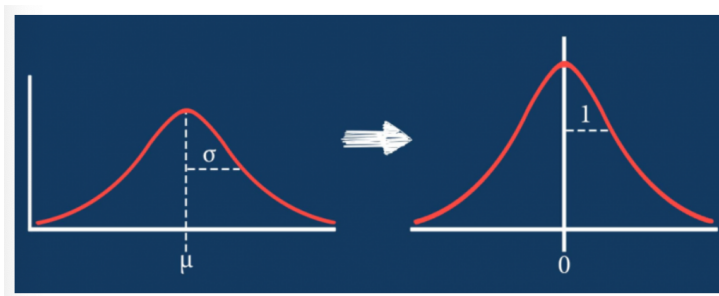
$$\sim (\mu, \sigma^2) \longrightarrow \sim (0, 1)$$

$$\frac{x - \mu}{\sigma}$$

# Chuẩn hoá dữ liệu (Tiếp)



# Chuẩn hoá dữ liệu (Tiếp)



## Chuẩn hoá dữ liệu (Tiếp)

Nếu  $X$  là biến ngẫu nhiên có phân phối chuẩn  $N(\mu, \sigma^2)$  thì:

$$Z = \frac{X - \mu}{\sigma}$$

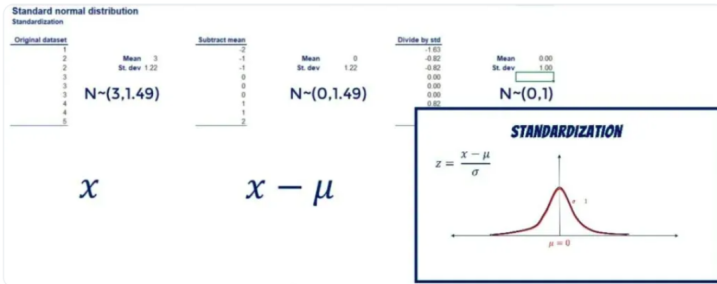
là biến ngẫu nhiên có phân phối chuẩn tắc  $N(0, 1)$ .

- Nếu  $Z \sim N(0, 1)$  thì  $P(a \leq Z < b) = \Phi(b) - \Phi(a)$ .
- Nếu  $X \sim N(\mu, \sigma^2)$  thì:

$$P(a \leq Z < b) = F(b) - F(a) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

- Hàm  $\Phi(x)$  đối xứng qua đường thẳng  $x = 0$  nên  $\Phi(-x) = 1 - \Phi(x)$

# Chuẩn hoá dữ liệu (Tiếp)



# Bảng Z (Z Table)

Bảng phân phối chuẩn tắc (bảng Z, hay bảng  $\Phi(z)$ ) cho biết xác suất:  $\Phi(z) = P(Z \leq z)$ , với  $Z \sim N(0, 1)$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

# Ví dụ

Giả sử độ dài một chi tiết máy tuân theo phân bố chuẩn với giá trị trung bình là 20cm và độ lệch chuẩn là 0,5cm. Tính xác suất khi chọn ngẫu nhiên ra một chi tiết thì độ dài của nó:

1. Lớn hơn 20cm
2. Bé hơn 19,5cm
3. nằm trong khoảng 19cm đến 21cm.



## Bài tập 2

- (1) Tính xác suất để chọn được 1 số trong đoạn từ 1 đến 10000 thỏa mãn điều kiện không chia hết cho bất kỳ số nào trong ba số 4, 6, và 9?
- (2) Xét một lô sản phẩm có số lượng rất lớn, trong đó, số sản phẩm do phân xưởng I sản xuất chiếm 20%, phân xưởng II sản xuất chiếm 30%, phân xưởng III sản xuất chiếm 50%. Xác suất phế phẩm của phân xưởng I là 0,001; phân xưởng 2 là 0,005; phân xưởng III là 0,006. Lấy ngẫu nhiên 1 sản phẩm của lô hàng. Tìm xác suất để sản phẩm đó là phế phẩm. Nêu ý nghĩa của xác suất đó?

## Bài tập 2 (Tiếp)

- (3) Một nhà máy sản xuất bóng đèn có tỉ lệ bóng đèn tốt là 90%. Trước khi xuất ra thị trường mỗi bóng đèn đều được kiểm tra chất lượng. Vì sự kiểm tra không tuyệt đối nên một bóng đèn tốt có xác suất 0,9 được công nhận tốt, còn một bóng đèn hỏng có xác suất 0,95 bị loại bỏ. Tính tỉ lệ bóng qua được kiểm tra chất lượng mà lại là bóng hỏng.
- (4) Thời gian cho đến khi sạc lại pin cho máy tính xách tay trong các điều kiện thông thường (thời gian sử dụng pin) có phân phối chuẩn với giá trị trung bình là 260 phút và độ lệch chuẩn là 50 phút.
- a) Tính xác suất để thời gian sử dụng pin là hơn bốn giờ?
  - b) Thời gian sử dụng pin tối thiểu bằng bao nhiêu để có xác suất là 95%?
  - c) Vẽ đồ thị của phân phối chuẩn đã cho sử dụng thư viện của python?

## Bài tập 2 (Tiếp)

- (5) Giả sử mẫu dữ liệu sau: 13, 16, 19, 22, 23, 38, 47, 56, 58, 63, 65, 70, 71, được trích ra từ một biến ngẫu nhiên có phân phối chuẩn với kỳ vọng là 43,15 và độ lệch tiêu chuẩn là 22,13. Chuẩn hoá mẫu dữ liệu đã cho.

# Ôn tập về thống kê

# Khoảng tin cậy

## Định nghĩa

Giả sử chưa biết đặc trưng  $\theta$  nào đó của biến ngẫu nhiên  $X$ . Ước lượng khoảng của  $\theta$  là chỉ ra một khoảng số  $(g_1, g_2)$  nào đó chứa  $\theta$ , tức là có thể ước lượng  $g_1 < \theta < g_2$ .

## Ước lượng bằng khoảng tin cậy

Với  $\alpha > 0$  khá bé, ta tìm được  $P(g_1 < \theta < g_2) = 1 - \alpha = \beta$  thì ta kết luận: với độ tin cậy  $1 - \alpha = \beta$ , tham số  $\theta$  nằm trong khoảng  $(g_1, g_2)$ . Khi đó:

- a)  $(g_1, g_2)$  được gọi là khoảng tin cậy của  $\theta$  với độ tin cậy  $\beta = 1 - \alpha$ .
- b)  $1 - \alpha = \beta$  được gọi là độ tin cậy của ước lượng.
- c)  $l = g_2 - g_1$  được gọi là độ dài khoảng tin cậy.

# Bài toán ước lượng bằng khoảng tin cậy

Giả sử biến ngẫu nhiên  $X$  tuân theo luật phân phối chuẩn  $N(\mu, \sigma^2)$  với kỳ vọng  $E(X) = \mu$  chưa biết. Hãy ước lượng  $E(X)$ . Bài toán được xét trong 3 trường hợp:

- Trường hợp 1: đã biết phương sai  $V(X) = \sigma^2$
- Trường hợp 2: chưa biết phương sai  $V(X) = \sigma^2$  và kích thước mẫu  $n > 30$
- Trường hợp 3: chưa biết phương sai  $V(X) = \sigma^2$  và kích thước mẫu  $n < 30$

# Bài toán ước lượng bằng khoảng tin cậy (Tiếp)

**Trường hợp đã biết phương sai  $V(X) = \sigma^2$**

Khoảng tin cậy đối xứng của  $E(X)$  là

$$\left( \bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

Trong đó:

- $\bar{x}$  là trung bình mẫu,  $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$
- $z_{\frac{\alpha}{2}}$  được xác định từ bảng giá trị hàm phân phối chuẩn tắc từ 2

hệ thức  $\Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$

# Bài toán ước lượng bằng khoảng tin cậy (Tiếp)

**Trường hợp chưa biết phương sai  $V(X) = \sigma^2$  và kích thước mẫu  $n > 30$**

Khoảng tin cậy đối xứng của  $E(X)$  là

$$\left( \bar{x} - z_{\alpha} \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha} \frac{s}{\sqrt{n}} \right)$$

Trong đó:

- $\bar{x}$  là trung bình mẫu,  $s = \sqrt{s^2}$ ,  $s^2$  là phương sai mẫu hiệu chỉnh

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

- $z_{\frac{\alpha}{2}}$  được xác định từ bảng giá trị hàm phân phối chuẩn tắc từ 2

hệ thức  $\Phi(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$



# Bài toán ước lượng bằng khoảng tin cậy (Tiếp)

**Trường hợp chưa biết phương sai  $V(X) = \sigma^2$  và kích thước mẫu  $n < 30$**

Khoảng tin cậy đối xứng của  $E(X)$  là

$$\left( \bar{x} - t_{\frac{\alpha}{2}(n-1)} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}(n-1)} \frac{s}{\sqrt{n}} \right)$$

Trong đó:

- $\bar{x}$  là trung bình mẫu,  $s = \sqrt{s^2}$ ,  $s^2$  là phương sai mẫu hiệu chỉnh

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

- $t_{\frac{\alpha}{2}(n-1)}$  được xác định từ bảng phân phối Student với  $n-1$  bậc tự do.

# Ví dụ

Giả sử rằng tuổi thọ của một loại bóng đèn hình TV có độ lệch chuẩn bằng 500, nhưng chưa biết trung bình. Ngoài ra, tuổi thọ của loại bóng đèn đó tuân theo luật phân phối chuẩn. Khảo sát trên một mẫu ngẫu nhiên gồm 15 bóng loại trên, người ta tính được tuổi thọ trung bình là 8900 giờ. Hãy tìm khoảng tin cậy 95% cho tuổi thọ trung bình của loại bóng đèn hình nói trên.

# Cách tính khoảng tin cậy trong Python

- Với dữ liệu đã cho xác định sử dụng hàm của phân phối chuẩn (`norm.interval()` function) hay phân phối t (`t.interval()` function).
- Xác định kích thước mẫu, độ tin cậy.
- Sử dụng thư viện `scipy.stats` trong Python.

Ví dụ: Ước lượng bằng khoảng tin cậy 95% chiều cao trung bình của một tổng thể cây dựa vào mẫu quan sát gồm độ cao của 22 cây sau: 11, 12, 12.5, 13, 13, 15.5, 16, 17, 22, 23, 25, 26, 27, 28, 28, 29, 30, 32, 33, 33, 34, 34.

# Cách tính khoảng tin cậy trong Python

Ví dụ: Ước lượng bằng khoảng tin cậy 95% chiều cao trung bình của một tổng thể cây dựa vào mẫu quan sát gồm độ cao của 22 cây sau: 11, 12, 12.5, 13, 13, 15.5, 16, 17, 22, 23, 25, 26, 27, 28, 28, 29, 30, 32, 33, 33, 34, 34.

```
1 import numpy as np
2 import scipy.stats as st
3
4 #define sample data
5 data = [11, 12, 12.5, 13, 13, 15.5, 16, 17, 22, 23, 25, 26, 27, 28, 28, 29,
6         30, 32, 33, 33, 34, 34]
7
8 #create 95% confidence interval for population mean weight
9 a = st.t.interval(alpha=0.95, df=len(data)-1,
10                  loc=np.mean(data), scale=st.sem(data))
```

# Cách tính khoảng tin cậy trong Python

Ví dụ: Ước lượng bằng khoảng tin cậy 95% chiều cao trung bình của một tổng thể cây dựa vào mẫu quan sát gồm độ cao của 33 cây sau: 11, 11, 11.5, 11.5, 11.5, 12, 12, 12, 12.5, 13, 13, 15.5, 16, 17, 22, 23, 25, 26, 27, 28, 28, 29, 30, 32, 33, 33, 34, 34, 35, 35, 35.5, 35.5, 36.

```
1 import numpy as np
2 import scipy.stats as st
3
4
5 #define sample data
6 data = [11, 11, 11.5, 11.5, 11.5, 12, 12, 12, 12.5, 13, 13, 15.5, 16, 17, 22,
7         23, 25, 26, 27, 28, 28, 29, 30, 32, 33, 33, 34, 34, 35, 35, 35.5, 35.5, 36]
8
9 #create 95% confidence interval for population mean weight
10 a = st.norm.interval(alpha=0.95, loc=np.mean(data), scale=st.sem(data))
```

# Giả thuyết không, giả thuyết đối

- (i) Giả sử cần nghiên cứu tham số  $\theta$  của biến ngẫu nhiên  $X$  và có cơ sở nào đó để nêu lên giả thuyết  $\theta = \theta_0$ . Giả thuyết này ký hiệu là  $H_0$ , còn gọi là giả thuyết cần kiểm định hay giả thuyết không (null hypothesis).
- (ii) Mệnh đề đối lập với giả thuyết  $H_0$  ký hiệu là  $H_1$ , còn gọi là đối thuyết (alternative hypothesis). Dạng tổng quát nhất của  $H_1$  là  $\theta \neq \theta_0$ . Trong nhiều trường hợp giả thuyết đối được phát biểu cụ thể là  $H_1 : \theta > \theta_0$  hoặc  $H_1 : \theta < \theta_0$ .

# Bài toán Kiểm định Giả thuyết thống kê

Giả sử biến ngẫu nhiên gốc  $X$  trong tổng thể có phân phối chuẩn  $N(\mu, \sigma^2)$ , trong đó  $E(X) = \mu$  chưa biết nhưng có cơ sở để nêu lên giả thuyết  $H_0 : \mu = \mu_0$  với  $\mu_0$  là tham số đã biết. Với mức ý nghĩa  $\alpha$  hãy kiểm định giả thuyết này với các đối thuyết  $H_1 : \mu \neq \mu_0$  hoặc  $\mu > \mu_0$  hoặc  $\mu < \mu_0$ . Bài toán được xét trong 3 trường hợp:

- Trường hợp 1: đã biết phương sai  $V(X) = \sigma^2$
- Trường hợp 2: chưa biết phương sai  $V(X) = \sigma^2$  và kích thước mẫu  $n > 30$
- Trường hợp 3: chưa biết phương sai  $V(X) = \sigma^2$  và kích thước mẫu  $n < 30$

# Trường hợp đã biết phương sai $V(X) = \sigma^2$

- Bước 1: Đặt giả thuyết  $H_0$  và đối thuyết  $H_1$ .
- Bước 2: Xác định miền bác bỏ  $H_0$  là  $W_\alpha$  theo bảng sau:

$H_0$	$H_1$	Miền bác bỏ $W_\alpha$
$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, +\infty)$
$\mu = \mu_0$	$\mu > \mu_0$	$(z_\alpha, +\infty)$
$\mu = \mu_0$	$\mu < \mu_0$	$(-\infty, -z_\alpha)$

Trong đó  $z_\alpha, z_{\frac{\alpha}{2}}$  được xác định từ bảng giá trị hàm 2 phân phối chuẩn tắc  $\phi(x)$ .



# Trường hợp đã biết phương sai $V(X) = \sigma^2$ (Tiếp)

- Bước 3: Tính giá trị quan sát:

$$z_{qs} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$$

- Bước 4: Xem  $z_{qs}$  có thuộc  $W_\alpha$  hay không để kết luận.
  - Nếu  $z_{qs} \in W_\alpha$  thì bác bỏ giả thuyết  $H_0$ .
  - Nếu  $z_{qs} \notin W_\alpha$  thì chưa có cơ sở để bác bỏ giả thuyết  $H_0$ .

# Trường hợp chưa biết phương sai, cỡ mẫu $n \geq 30$

- Bước 1: Đặt giả thuyết  $H_0$  và đối thuyết  $H_1$ .
- Bước 2: Xác định miền bác bỏ  $H_0$  là  $W_\alpha$  theo bảng sau:

$H_0$	$H_1$	Miền bác bỏ $W_\alpha$
$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, +\infty)$
$\mu = \mu_0$	$\mu > \mu_0$	$(z_\alpha, +\infty)$
$\mu = \mu_0$	$\mu < \mu_0$	$(-\infty, -z_\alpha)$

Trong đó  $z_\alpha, z_{\frac{\alpha}{2}}$  được xác định từ bảng giá trị hàm 2 phân phối chuẩn tắc  $\Phi(x)$ .

# Trường hợp chưa biết phương sai, cỡ mẫu $n \geq 30$ (Tiếp)

- Bước 3: Tính giá trị quan sát:

$$z_{qs} = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$$

- Bước 4: Xem  $z_{qs}$  có thuộc  $W_\alpha$  hay không để kết luận.
  - Nếu  $z_{qs} \in W_\alpha$  thì bác bỏ giả thuyết  $H_0$ .
  - Nếu  $z_{qs} \notin W_\alpha$  thì chưa có cơ sở để bác bỏ giả thuyết  $H_0$ .

# Trường hợp chưa biết phương sai, cỡ mẫu $n < 30$

- Bước 1: Đặt giả thuyết  $H_0$  và đối thuyết  $H_1$ .
- Bước 2: Xác định miền bác bỏ  $H_0$  là  $W_\alpha$  theo bảng sau:

$H_0$	$H_1$	Miền bác bỏ $W_\alpha$
$\mu = \mu_0$	$\mu \neq \mu_0$	$(-\infty, -t_{\alpha/2}^{(n-1)}) \cup (t_{\alpha/2}^{(n-1)}, +\infty)$
$\mu = \mu_0$	$\mu > \mu_0$	$(t_{\alpha}^{(n-1)}, +\infty)$
$\mu = \mu_0$	$\mu < \mu_0$	$(-\infty, -t_{\alpha}^{(n-1)})$

Trong đó  $t_{\alpha/2}^{(n-1)}, t_{\alpha}^{(n-1)}$  được xác định từ bảng phân phối Student.

# Trường hợp chưa biết phương sai, cỡ mẫu $n < 30$ (Tiếp)

- Bước 3: Tính giá trị quan sát:

$$z_{qs} = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$$

- Bước 4: Xem  $z_{qs}$  có thuộc  $W_\alpha$  hay không để kết luận.
  - Nếu  $z_{qs} \in W_\alpha$  thì bác bỏ giả thuyết  $H_0$ .
  - Nếu  $z_{qs} \notin W_\alpha$  thì chưa có cơ sở để bác bỏ giả thuyết  $H_0$ .

# Ví dụ

Một hãng bảo hiểm thông báo rằng số tiền trung bình hãng chi trả cho khách hàng bị tai nạn ô tô là 8500 USD. Để kiểm tra lại, người ta kiểm tra ngẫu nhiên hồ sơ chi trả của 25 khách hàng thì thấy số tiền trung bình chi trả là 8900 USD. Giả sử số tiền chi trả tuân theo luật phân phối chuẩn với độ lệch chuẩn là 2600 USD. Hãy kiểm định lại thông báo của hãng bảo hiểm trên với mức ý nghĩa 5%.

# Kiểm định giả thuyết thống kê bằng p-value

P-value là mức ý nghĩa nhỏ nhất có thể dẫn đến việc bác bỏ giả thuyết  $H_0$  với dữ liệu đã cho.

## Trường hợp đã biết phương sai

- Bước 1: Đặt giả thuyết  $H_0$  và đối thuyết  $H_1$ .
- Bước 2: Xác định giá trị quan sát:

$$z_{qs} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$$

# Trường hợp đã biết phương sai

- Bước 3: Xác định p-value

$H_0$	$H_1$	P-value
$\mu = \mu_0$	$\mu \neq \mu_0$	$2[1 - \Phi(z_{qs})]$
$\mu = \mu_0$	$\mu > \mu_0$	$1 - \Phi(z_{qs})$
$\mu = \mu_0$	$\mu < \mu_0$	$\Phi(z_{qs})$

- Bước 4: Kết luận:

- Nếu p-value  $< 0.05$ , bác bỏ giả thuyết  $H_0$ .
- Ngược lại: Chưa đủ điều kiện bác bỏ giả thuyết  $H_0$ .

Các trường hợp còn lại tương tự trường hợp đã biết phương sai, chỉ thay giá trị quan sát tương ứng.



## Bài tập số 3

1. Một tổng thể  $X$  có phân phối chuẩn. Quan sát một mẫu ngẫu nhiên kích thước 25 người ta tính được trung bình là 15 và độ lệch chuẩn là 3. Hãy ước lượng kỳ vọng của  $X$  bằng khoảng tin cậy 95%.
2. Trọng lượng của một sản phẩm theo quy định là 6kg. Sau một thời gian sản xuất, người ta nghi ngờ trọng lượng của sản phẩm giảm đi. Bởi vậy, người ta tiến hành kiểm tra 121 sản phẩm và tính được trung bình mẫu là 5.975kg và phương sai mẫu hiệu chỉnh là  $5.7596\text{kg}^2$ . Với mức ý nghĩa 5%, hãy kết luận về nghi ngờ nói trên.

## Bài tập số 3 (Tiếp)

3. Một công ty sản xuất hạt giống tuyên bố rằng một loại giống mới của họ có năng suất trung bình là 21.5 tạ/ha. Gieo thử hạt giống mới này tại 16 vườn thí nghiệm và thu được kết quả:

19.2, 18.7, 22.4, 20.3, 16.8, 25.1, 17.0, 15.8, 21.0, 18.6, 23.7, 24.1, 23.4, 19.8, 21.7, 18.9.

Dựa vào kết quả này hãy xác nhận xem quảng cáo của công ty có đúng không với mức ý nghĩa  $\alpha = 5\%$ . Biết rằng năng suất giống cây trồng là một biến ngẫu nhiên tuân theo luật phân phối chuẩn.

# Bài tập Nhóm số 1

- Tìm hiểu thư viện để giải bài toán kiểm định giả thuyết thống kê trong Python?
- Tìm dữ liệu và đặt đầu bài cho dữ liệu và giải bằng chương trình Python với 3 nội dung sau:
  1. Chuẩn hoá dữ liệu tìm được.
  2. Ước lượng kỳ vọng bằng khoảng tin cậy
  3. Kiểm định giả thuyết thống kê cho kỳ vọng.

## Ví dụ về đề bài cho bài tập nhóm

Theo tổng điều tra dinh dưỡng năm 2019-2020 của Bộ Y tế, Chiều cao trung bình của nam giới trưởng thành Việt Nam là 168,1cm. Tuy nhiên, nhóm 1 nghi ngờ rằng con số này có cải thiện đáng kể so với trước đây. Bởi vậy, nhóm 1 tiến hành khảo sát 100 thanh niên nam giới và tính được trung bình mẫu là  $X\text{cm}$  và phương sai mẫu hiệu chỉnh là  $Y\text{cm}^2$ . Nhóm 1 sẽ tiến hành các thí nghiệm để kết luận về nghi ngờ nói trên, với mức ý nghĩa 5%.

Các yêu cầu cần thực hiện:

- Bước 1: Khảo sát để thu thập dữ liệu mẫu (Dữ liệu cần có tên, tuổi của người được khảo sát)
- Bước 2: Chuẩn hoá dữ liệu thu được bằng z-score
- Bước 3: Ước lượng kỳ vọng bằng khoảng tin cậy
- Bước 4: Kiểm định giả thuyết thống kê vừa nêu cho kỳ vọng với mức ý nghĩa 5%.

# Yêu cầu bài tập Nhóm số 1

- Mỗi nhóm tạo 1 folder drive để lưu các bài tập nhóm. Sau đó gửi link folder qua email [thuvh@ptit.edu.vn](mailto:thuvh@ptit.edu.vn). Tiêu đề email: "KHDL-Nhóm lớp-STT Nhóm đăng ký". Ví dụ:  
**KHDL-N19-Nhóm 1**
- Các nhóm đẩy kết quả của các bài tập nhóm lên link drive trước khi hết hạn. Các nhóm nộp bài quá deadline: trừ 1 điểm/ 1 ngày nộp muộn.
- Bảo vệ kết quả bài tập nhóm của nhóm mình.
- **Deadline bài tập Nhóm số 1: 12h ngày 30/9/2025**