

Xác suất và thống kê

☑ Trắc nghiệm 8 ⌵ PDF bài

Hai model được chấm trên cùng 200 ảnh validation. Model A đúng 84%, model B đúng 86%. Nhóm chọn B. Một tuần sau, trên bộ test kín, A lại cao hơn B. Không ai làm sai gì cả: 200 ảnh là một mẫu ngẫu nhiên, và con số đo trên mẫu ngẫu nhiên thì dao động. Để biết một chênh lệch như 2% là thật hay chỉ do may rủi, cần vài khái niệm xác suất và thống kê: xác suất có điều kiện, Bayes, kỳ vọng, phương sai, phân phối chuẩn và sai số chuẩn. Mỗi khái niệm đều được kiểm bằng mô phỏng: cho máy tung xúc xắc hay rút mẫu hàng vạn lần rồi đếm.

MỤC TIÊU

- Tính xác suất, xác suất có điều kiện; kiểm hai biến cố có độc lập không.
- Dùng công thức Bayes.
- Tính trung bình, median, phương sai, độ lệch chuẩn, z-score.
- Tính sai số chuẩn của accuracy để biết chênh lệch nào là nhiều.

Xác suất là tỉ lệ khi lặp lại nhiều lần

Xác suất của một biến cố là một số từ 0 đến 1. Khi mọi kết quả đều khả năng như nhau, xác suất là **số kết quả thuận lợi chia cho tổng số kết quả**. Xúc xắc cân đối ra mặt 5 hoặc 6 có xác suất $2/6 = 1/3$. Xác suất biến cố **không** xảy ra là 1 trừ xác suất nó xảy ra.

Cách hiểu thứ hai hữu ích hơn khi lập trình: xác suất là tỉ lệ biến cố xảy ra nếu lặp lại phép thử rất nhiều lần. Cho máy tung xúc xắc một trăm nghìn lần rồi đếm:

Python

```
import numpy as np

rng = np.random.default_rng(0)
rolls = rng.integers(1, 7, size=100_000)    # số nguyên từ 1 đến 6
print(round((rolls >= 5).mean(), 4), round(1 / 3, 4))
print(round(rolls.mean(), 3))
```

KẾT QUẢ

0.3326 0.3333

3.498

Tỉ lệ đếm được khớp $1/3$ tới hai chữ số thập phân. Dòng thứ hai là trung bình các mặt, gần 3.5; mục kỳ vọng sẽ giải thích con số này. Một biểu thức so sánh như `rolls >= 5` cho mảng True/False, và `.mean()` của nó chính là tỉ lệ True, mẹo đã dùng ở bài 0.3.

HỎI NHANH

Rút ngẫu nhiên một lá từ bộ bài 52 lá. Xác suất **không** rút được lá cơ (13 lá) là bao nhiêu?

A $1/4$

B $13/52$ nhân $39/52$

C $3/4$

D $1/13$

Xác suất có điều kiện và độc lập

Một lớp 40 học sinh: 25 bạn làm bài tập đều, trong đó 20 bạn đạt; 15 bạn không làm đều, trong đó 6 bạn đạt. Chọn ngẫu nhiên một bạn. Xác suất bạn ấy đạt là $26/40 = 0.65$. Nhưng nếu đã biết bạn ấy làm bài tập đều, ta chỉ đếm **trong nhóm 25 bạn ấy**: $20/25 = 0.8$.

Đó là **xác suất có điều kiện**, viết $P(A | B)$, đọc “xác suất của A khi đã biết B ”:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}.$$

$A \cap B$ là “cả A lẫn B cùng xảy ra”. Ở ví dụ, $P(\text{đạt} \cap \text{làm đều}) = 20/40$ và $P(\text{làm đều}) = 25/40$, chia ra được 0.8. Mẫu số đổi từ cả lớp sang nhóm điều kiện.

Hai biến cố **độc lập** khi biết cái này không đổi xác suất của cái kia: $P(A | B) = P(A)$, tương đương

$$P(A \cap B) = P(A) P(B).$$

Ở lớp trên, $0.8 \neq 0.65$, nên đạt và làm bài tập đều không độc lập. Hai lần tung xúc xắc thì độc lập: xác suất cả hai lần ra 6 là $1/6 \cdot 1/6 = 1/36$. Chỉ được nhân hai xác suất khi đã có lý do để tin chúng độc lập.

Python

```
pass_ = np.array([1] * 20 + [0] * 5 + [1] * 6 + [0] * 9)      # đạt hay không
homework = np.array([1] * 25 + [0] * 15)                      # làm bài tập đều hay không
p_pass = pass_.mean()
p_pass_given_hw = pass_[homework == 1].mean()
p_both = (pass_ & homework).mean()
print(p_pass, p_pass_given_hw, p_both / homework.mean())
```

KẾT QUẢ

0.65 0.8 0.8

`pass_[homework == 1]` lọc đúng nhóm điều kiện rồi lấy tỉ lệ, cách viết thẳng theo nghĩa “chỉ đếm trong nhóm”. Cột cuối tính theo công thức và ra cùng số.

HỎI NHANH

Trong lớp trên, xác suất một bạn làm bài tập đều khi đã biết bạn ấy đạt, $P(\text{làm đều} | \text{đạt})$, là bao nhiêu?

A 20/25

B 20/26

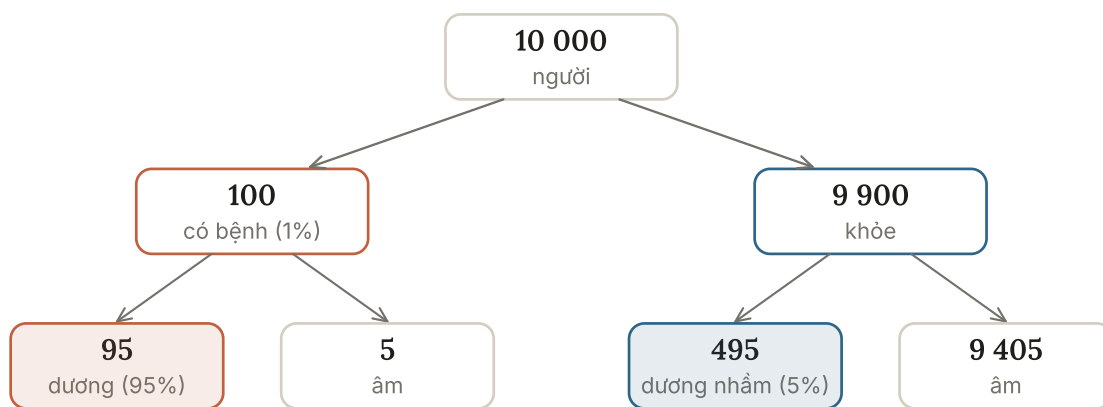
C 25/40

D 20/40

Bayes: đảo chiều điều kiện

Một xét nghiệm phát hiện được 95% người có bệnh, và báo nhầm 5% người khỏe. Bệnh này chỉ 1% dân số mắc. Một người xét nghiệm dương tính. Xác suất người ấy có bệnh là bao nhiêu?

Nhiều người trả lời 95%. Con số ấy là $P(\text{dương} \mid \text{bệnh})$, còn câu hỏi là $P(\text{bệnh} \mid \text{dương})$, điều kiện nằm ở phía ngược lại. Cách dễ nhất là đếm trên 10 000 người. 100 người có bệnh, 95 người trong số đó dương tính. 9900 người khỏe, 5% của họ là 495 người dương tính nhầm. Tổng cộng 590 người dương tính, chỉ 95 người thật sự có bệnh: $95/590 \approx 0.16$.



dương tính: $95 + 495 = 590$ người, có bệnh $95/590 \approx 0.16$

Trong 590 người dương tính, chỉ 95 người có bệnh.

Viết thành công thức, đó là **định lý Bayes**:

$$P(B \mid D) = \frac{P(D \mid B) P(B)}{P(D)}, \quad P(D) = P(D \mid B)P(B) + P(D \mid \neg B)P(\neg B).$$

Ở đây B là “có bệnh”, D là “dương tính”, $\neg B$ là “không có bệnh”. Tử số $0.95 \cdot 0.01 = 0.0095$; mẫu số $0.0095 + 0.05 \cdot 0.99 = 0.059$. Chia ra được 0.161, khớp phép đếm.

Chuyện này cũng xảy ra với model. Khi trường hợp cần tìm rất hiếm, một model bất được gần hết ca thật và ít báo nhầm vẫn có thể có phần lớn cảnh báo là nhầm, vì nhóm bình thường quá đông. Bài 1.5 sẽ đo điều này bằng precision.

Python

```
p_b = 0.01          # tỉ lệ mắc bệnh
p_d_b = 0.95        # dương tính khi có bệnh
p_d_not_b = 0.05    # dương tính nhầm khi khỏe
p_d = p_d_b * p_b + p_d_not_b * (1 - p_b)
print(round(p_d, 4), round(p_d_b * p_b / p_d, 4))

sick = rng.random(1_000_000) < p_b
positive = np.where(sick, rng.random(1_000_000) < p_d_b, rng.random(1_000_000) < p_d_not_b)
print(round(sick[positive].mean(), 4))
```

KẾT QUẢ

```
0.059 0.161
0.1609
```

Mô phỏng một triệu người cho tỉ lệ bệnh thật trong nhóm dương tính khoảng 0.161, khớp công thức tới hai chữ số. `np.where(điều kiện, a, b)` chọn phần tử của a ở chỗ điều kiện đúng và của b ở chỗ sai.

HỎI NHANH

Giữ nguyên xét nghiệm (bắt 95% người bệnh, báo nhầm 5% người khỏe), nhưng dùng ở một bệnh viện nơi một nửa số người đến khám có bệnh. Một người dương tính có bệnh với xác suất bao nhiêu?

- A 0.16, vì xét nghiệm không đổi
- B 0.5, bằng tỉ lệ mắc bệnh
- C 0.475
- D 0.95

Kỳ vọng và trung bình

Kỳ vọng của một đại lượng ngẫu nhiên là trung bình các giá trị, mỗi giá trị nhân với xác suất của nó:

$$\mathbb{E}[X] = \sum_x x \cdot P(X = x).$$

Xúc xắc cân đối: $\mathbb{E}[X] = (1 + 2 + \dots + 6)/6 = 3.5$, khớp con số 3.498 ở mục đầu. Kỳ vọng không nhất thiết là một giá trị có thể xảy ra; không lần tung nào ra 3.5. Nó là trung bình **về lâu dài**.

Một vé số giá 10 000 đồng, trúng 1 000 000 đồng với xác suất 0.005. Tiền lời kỳ vọng mỗi vé là $0.005 \cdot 1\,000\,000 - 10\,000 = -5000$ đồng: mua càng nhiều, càng gần chắc chắn lỗ.

Kỳ vọng có tính **tuyến tính**: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$, kể cả khi X và Y không độc lập. Tung hai xúc xắc, kỳ vọng tổng là $3.5 + 3.5 = 7$ mà không cần liệt kê 36 trường hợp.

Trên dữ liệu thật, ta không biết xác suất, chỉ có các giá trị đã quan sát. **Trung bình mẫu** $\bar{x} = \frac{1}{n} \sum x_i$ là ước lượng của kỳ vọng. **Median** là giá trị đứng giữa khi sắp xếp; nó ít bị kéo bởi một vài giá trị cực lớn, như bài 0.3 đã thấy với giờ tự học lệch phải.

HỎI NHANH

Tung hai xúc xắc cân đối rồi lấy mặt thứ nhất trừ mặt thứ hai. Kỳ vọng của hiệu là bao nhiêu?

A 0

B 3.5

C 7

D Không tính được nếu không liệt kê 36 trường hợp

Phương sai và độ lệch chuẩn

Hai nhóm học sinh có điểm $[6, 7, 7, 8, 7]$ và $[4, 9, 5, 10, 7]$. Trung bình đều bằng 7, nhưng nhóm sau phân tán hơn nhiều. Đo độ phân tán bằng **phương sai**: trung bình bình phương khoảng cách tới trung bình,

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2], \quad \sigma = \sqrt{\text{Var}(X)}.$$

σ là **độ lệch chuẩn** (standard deviation), cùng đơn vị với dữ liệu nên dễ đọc hơn phương sai. Nhóm đầu có độ lệch $-1, 0, 0, 1, 0$, bình phương có trung bình $2/5 = 0.4$, nên $\sigma \approx 0.63$. Nhóm sau có phương sai $26/5 = 5.2$, $\sigma \approx 2.28$.

z-score của một giá trị cho biết nó cách trung bình bao nhiêu lần độ lệch chuẩn: $z = (x - \mu)/\sigma$. Điểm 10 ở nhóm sau có $z = 3/2.28 \approx 1.32$; ở nhóm đầu, điểm 10 có $z \approx 4.7$, tức là bất thường hơn nhiều. Chuẩn hóa dữ liệu ở bài 1.4 làm đúng phép này cho từng cột.

Python

```
a = np.array([6, 7, 7, 8, 7], dtype=float)
b = np.array([4, 9, 5, 10, 7], dtype=float)
for g in (a, b):
    print(g.mean(), round(g.var(), 2), round(g.std(), 2), round((10 - g.mean()) / g.std(),
2))
print(np.var(a, ddof=1))
```

KẾT QUẢ

```
7.0 0.4 0.63 4.74
7.0 5.2 2.28 1.32
0.5
```

NumPy mặc định chia cho n . Dòng cuối dùng `ddof=1` để chia cho $n - 1$, ra $2/4 = 0.5$; pandas mặc định tính `std()` theo cách này. Chia cho $n - 1$ cho kết quả sát hơn khi dùng một mẫu nhỏ để đoán độ phân tán của cả tổng thể. Khi n lớn, hai cách gần như bằng nhau; chỉ cần nhớ để không ngạc nhiên khi NumPy và pandas in hai số khác nhau.

HỎI NHANH

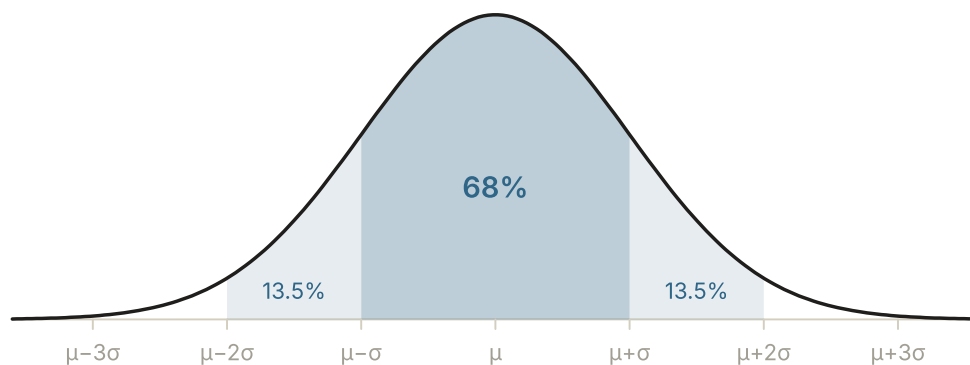
Thêm 10 vào mọi điểm của một nhóm. Trung bình và độ lệch chuẩn thay đổi thế nào?

- A Cả hai tăng thêm 10
- B Trung bình không đổi, độ lệch chuẩn tăng 10
- C Trung bình tăng 10, độ lệch chuẩn tăng $\sqrt{10}$
- D Trung bình tăng 10, độ lệch chuẩn không đổi

Phân phối chuẩn

Chiều cao, sai số đo, điểm thi của nhiều người thường có histogram hình chuông: dày ở giữa, mỏng dần hai bên, đối xứng. Hình chuông ấy là **phân phối chuẩn** (normal distribution), xác định bởi hai số: trung bình μ và độ lệch chuẩn σ .

Quy tắc dễ nhớ nhất là **68-95-99.7**: khoảng 68% giá trị nằm trong $\mu \pm \sigma$, 95% trong $\mu \pm 2\sigma$, 99.7% trong $\mu \pm 3\sigma$. Một giá trị có $|z| > 3$ là hiếm, đáng kiểm xem có phải lỗi nhập liệu không.



Phân phối chuẩn: 68% giá trị trong $\mu \pm \sigma$, 95% trong $\mu \pm 2\sigma$.

Phân phối chuẩn xuất hiện nhiều vì một lý do sâu hơn: **trung bình của nhiều đại lượng ngẫu nhiên độc lập có phân phối gần chuẩn**, dù từng đại lượng có phân phối gì. Đó là định lý giới hạn trung tâm. Ngay cả trung bình của 30 lần tung xúc xắc, vốn có phân phối đều, cũng cho histogram hình chuông. Mục tiếp theo dùng đúng kết quả này cho accuracy.

Python

```
z = rng.normal(0, 1, size=100_000)
print(round((np.abs(z) < 1).mean(), 3), round((np.abs(z) < 2).mean(), 3))
means = rng.integers(1, 7, size=(100_000, 30)).mean(axis=1)
print(round(means.mean(), 3), round(means.std(), 3), round(np.sqrt(35 / 12 / 30), 3))
print(round((np.abs(means - 3.5) < means.std()).mean(), 3))
```

KẾT QUẢ

```
0.682 0.954
3.5 0.311 0.312
0.691
```

Dòng đầu: 68.2% và 95.4% số ngẫu nhiên chuẩn nằm trong một và hai độ lệch chuẩn. Dòng hai: một trăm nghìn lần lấy trung bình 30 mặt xúc xắc; trung bình của chúng quanh 3.5, độ lệch chuẩn 0.311, gần đúng bằng $\sigma/\sqrt{30} \approx 0.312$ với $\sigma^2 = 35/12$ là phương sai một xúc xắc. Dòng ba: khoảng 69% các trung bình nằm trong một độ lệch chuẩn, gần 68% của phân phối chuẩn. Phần lệch nhỏ đến từ việc trung bình 30 mặt chỉ nhận một số giá trị rời rạc.

HỎI NHANH

Điểm một kỳ thi có phân phối gần chuẩn, trung bình 6.0, độ lệch chuẩn 1.5. Khoảng bao nhiêu phần trăm thí sinh có điểm từ 3.0 đến 9.0?

A 68%

B 95%

C 99.7%

D 50%

Sai số chuẩn: điểm validation dao động bao nhiêu

Quay lại câu hỏi đầu bài. Một model có accuracy **thật** $p = 0.85$ trên toàn bộ ảnh có thể gặp. Ta chỉ chấm trên $n = 200$ ảnh rút ngẫu nhiên. Mỗi ảnh đúng hay sai giống một lần tung đồng xu lệch, nên accuracy đo được là trung bình của 200 biến 0/1. Theo mục trước, nó có phân phối gần chuẩn quanh p , với độ lệch chuẩn

$$SE = \sqrt{\frac{p(1-p)}{n}}.$$

Độ lệch chuẩn của một con số ước lượng gọi là **sai số chuẩn** (standard error). Với $p = 0.85$, $n = 200$: $SE = \sqrt{0.1275/200} \approx 0.025$. Theo quy tắc 95%, accuracy đo được thường nằm trong khoảng 0.85 ± 0.05 , tức từ 0.80 tới 0.90. Hai model chênh 2 điểm phần trăm trên 200 ảnh là chênh lệch nằm gọn trong nhiều.

Muốn SE nhỏ hơn thì cần n lớn hơn, nhưng SE chỉ giảm theo $\sqrt{n}$: gấp 10 lần dữ liệu thì SE chỉ còn khoảng một phần ba.

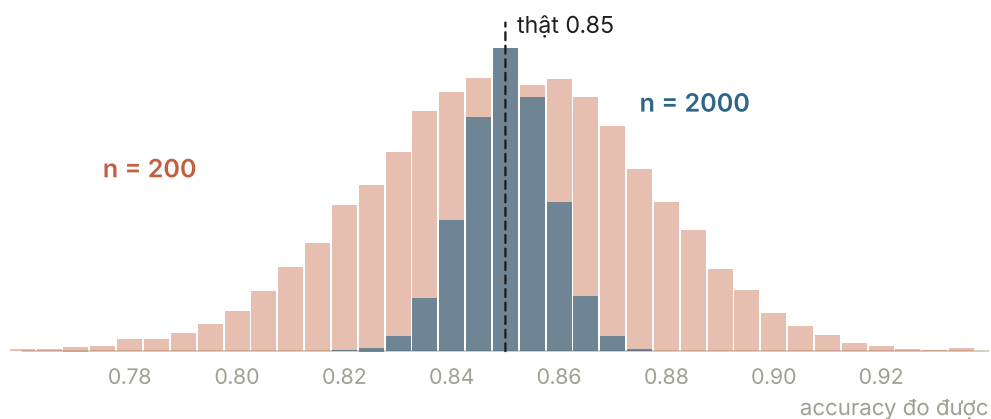
Python

```
p_true = 0.85
for n in [200, 2000]:
    acc = rng.binomial(n, p_true, size=10_000) / n      # 10 000 bộ validation, mỗi bộ n ảnh
    lo, hi = np.percentile(acc, [2.5, 97.5])
    print(n, round(acc.std(), 4), round(np.sqrt(p_true * (1 - p_true) / n), 4), lo, hi)
```

KẾT QUẢ

```
200 0.025 0.0252 0.8 0.895
2000 0.008 0.008 0.8345 0.866
```

`rng.binomial(n, p)` đếm số lần thành công trong n lần thử, mỗi lần xác suất p ; chia cho n được accuracy của một bộ validation. Mười nghìn bộ cho thấy với 200 ảnh, 95% số đo nằm từ 0.80 tới 0.895, rộng gần 10 điểm phần trăm. Với 2000 ảnh, khoảng ấy chỉ còn từ 0.8345 tới 0.866.



Cùng một model: chấm trên 200 ảnh dao động rộng, trên 2000 ảnh hẹp hơn.

Hai cách giảm nhiễu khi so model: dùng validation lớn hơn hoặc cross-validation (bài 1.3); và so hai model **trên cùng các ảnh**, đếm những ảnh mà hai model trả lời khác nhau, thay vì so hai con số tổng.

HỎI NHANH

Model đạt accuracy 0.90 trên 100 ảnh validation. Sai số chuẩn của con số này khoảng bao nhiêu?

A 0.09

B 0.003

C 0.03

D 0.9

Khi đi thi

- Viết rõ điều kiện nằm ở phía nào: $P(A | B)$ khác $P(B | A)$. Khi rồi, đếm trên 10 000 mẫu như mục Bayes.
- NumPy chia n , pandas chia $n - 1$ khi tính std.

- Hai model chênh nhau ít hơn hai lần $\sqrt{p(1-p)/n}$ thì chưa nói được model nào tốt hơn.

Tóm tắt

- Xác suất là tỉ lệ khi lặp lại nhiều lần; đếm hoặc mô phỏng đều được.
- $P(A | B) = P(A \cap B)/P(B)$: chỉ đếm trong nhóm điều kiện; độc lập khi $P(A \cap B) = P(A)P(B)$.
- Bayes $P(B | D) = P(D | B)P(B)/P(D)$ đảo chiều điều kiện; tỉ lệ nền $P(B)$ quyết định nhiều.
- Kỳ vọng là trung bình có trọng số xác suất, có tính tuyến tính; median ít bị giá trị cực đoan kéo.
- Phương sai là trung bình bình phương độ lệch, σ là căn của nó; z-score $= (x - \mu)/\sigma$.
- Phân phối chuẩn: 68-95-99.7; trung bình của nhiều biến độc lập gần chuẩn.
- Accuracy trên n mẫu có sai số chuẩn $\sqrt{p(1-p)/n}$, giảm theo $\sqrt{n}$.

Kiểm tra cuối bài

Mười câu dưới đây bao quát cả bài. Làm hết rồi xem điểm; câu nào sai thì đọc lại mục tương ứng.

CÂU 1

Trung bình của dãy 2, 4, 6, 8, 10 là bao nhiêu?

A 5

B 6

C 7

D 8

ĐỀ THI NOAI 2026 Singapore · Vòng sơ loại · Câu 57

CÂU 2

Median của dãy 1, 3, 5, 7, 9 là bao nhiêu?

A 3

B 7

C 9

D 5

ĐỀ THI NOAI 2026 Singapore · Vòng sơ loại · Câu 58

CÂU 3

Phương sai của dãy 0, 0, 0, 0 là bao nhiêu?

A 0

B 1

C 4

D Không xác định

ĐỀ THI NOAI 2026 Singapore · Vòng sơ loại · Câu 59

CÂU 4

Tung một xúc xắc sáu mặt cân đối. Xác suất ra mặt từ 5 trở lên là bao nhiêu?

A $1/6$

B $1/3$

C $1/2$

D $2/3$

ĐỀ THI NOAI 2026 Singapore · Vòng sơ loại · Câu 60

CÂU 5

Với phân phối chuẩn, khoảng bao nhiêu phần trăm giá trị nằm trong một độ lệch chuẩn quanh trung bình?

A 50%

B 95%

C 68%

D 99%

ĐỀ THI NOAI 2026 Singapore · Vòng sơ loại · Câu 61

CÂU 6

Hai biến cố A và B độc lập khi nào?

A Khi $P(A | B) = 0$

B Khi $P(A \cup B) = P(A) + P(B)$

C Khi A và B không thể cùng xảy ra

D Khi $P(A \cap B) = P(A)P(B)$

ĐỀ THI NOAI 2026 Singapore · Vòng sơ loại · Câu 62

CÂU 7

Định lý Bayes dùng để tính gì?

A $P(A | B)$ từ $P(B | A)$

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

C $P(A) \cdot P(B)$

D Trung bình và phương sai

ĐỀ THI NOAI 2026 Singapore · Vòng sơ loại · Câu 63

CÂU 8

Độ lệch chuẩn là gì?

A Bình phương của phương sai

B Tổng các độ lệch so với trung bình

C Trung bình trị tuyệt đối của độ lệch

D Căn bậc hai của phương sai

ĐỀ THI NOAI 2026 Singapore · Vòng sơ loại · Câu 64

CÂU 9

Một tập dữ liệu có trung bình 10, độ lệch chuẩn 2. z-score của giá trị 14 là bao nhiêu?

A 2

B 1

C 4

D 7

ĐỀ THI NOAI 2026 Singapore · Vòng sơ loại · Câu 65

CÂU 10

Trên 500 ảnh validation, model A đạt 0.80 và model B đạt 0.82. Sai số chuẩn của mỗi số khoảng 0.018. Nhóm nên kết luận thế nào?

A B chắc chắn tốt hơn vì 0.82 lớn hơn 0.80

B A tốt hơn vì B có thể đang overfit validation

C Chênh lệch còn nằm trong nhiễu, chưa kết luận được

D Hai model như nhau, vì chênh lệch nhỏ hơn 0.05

Đọc thêm

- Seeing Theory — [A visual introduction to probability and statistics](#). Trang tương tác của Đại học Brown: xác suất, Bayes, phân phối, định lý giới hạn trung tâm bằng hình động.
- Khan Academy — [Statistics and probability](#). Bài giảng và bài tập từ trung bình, phương sai tới xác suất có điều kiện.
- 3Blue1Brown — [Bayes theorem, the geometry of changing beliefs](#). Bayes bằng hình vẽ, đúng ví dụ xét nghiệm bệnh hiếm.

- Downey — [Think Stats](#). Sách miễn phí, thống kê bằng Python và mô phỏng, rất gần cách làm của bài.