Xác suất (probability)

Quách Đình Hoàng

6/10/2019

Xác suất (probability)

- Mục tiêu cuối cùng của phân tích dữ liệu là rút ra kết luận đáng tin cậy về quần thể dựa trên những gì ta đã khám phá trên mẫu.
- Để hiểu làm được việc đó, ta cần hiểu xác suất. Nó là nền tảng cơ bản cho các phương pháp suy luận thống kê.
- Ví dụ: Chọn ngẫu nhiên 1000 sinh viên nam và đo chiều cao ta thu được chiều cao trung bình là 1.7 m. Số liệu này có đúng cho toàn bộ sinh viên nam của trường?
  - Ta không chắc vì một mẫu ngẫu nhiên khác sẽ cho ra kết quả khác.
  - Nhưng ta có thể sử dụng xác suất để mô tả khả năng số liệu trên mẫu của ta nằm trong mức độ chính xác mong muốn.

Xác suất (probability)

- Xác suất là một cách để đo lường sự không chắc chắn.
- Gọi A là điều ta muốn tìm xác xuất, A được gọi là sự kiện (event).
- $P(A)$ thể hiện xác suất của sự kiện A, nó đo khả năng sự kiện A xảy ra.
- $0 \leq P(A) \leq 1, \forall A$, xác suất của sự kiện A bất kỳ luôn nằm ở giữa 0 và 1.

Xác định xác suất

- Có hai cách để xác định xác suất: lý thuyết (theoretical / classical) và thực nghiệm (empirical / observational).
- Phương pháp lý thuyết sử dụng bản chất của tình huống để xác định xác suất.
  - Tung một đồng xu cân bằng, $P(H) = P(T) = 0.5$.
- Phương pháp thực nghiệm dùng các thử nghiệm để tạo ra các kết quả không thể dự đoán trước
  - Tung đồng xu cân bằng 10000 lần và ghi lại kết quả để tính $P(H)$ và $P(T)$.
  - Ta khó thể đạt chính xác $P(H) = P(T) = 0.5$ như phương pháp lý thuyết nhưng thường cũng khá gần với nó.
  - Trong nhiều trường hợp, ta khó thể tìm được xác suất lý thuyết và phải dùng xác suất thực nghiệm để thay thế.

Không gian mẫu (sample space)

- Thí nghiệm ngẫu nhiên (random experiment): là một thí nghiệm tạo ra kết quả không thể dự đoán trước
  - Ví dụ: tung đồng xu và ghi lại kết quả, chọn ngẫu nhiên một SV và ghi lại ngày sinh, ...
- Mỗi thí nghiệm ngẫu nhiên có một tập hợp các kết quả có thể xảy ra được gọi là không gian mẫu (sample space).
  - Ta không chắc chắn về kết quả nào sẽ nhận được từ thí nghiệm nhưng chắc chắn về các kết quả có thể xảy ra.
  - Tình huống thường gặp là tất cả các kết quả trong không gian mẫu đều có khả năng xảy ra như nhau
- Một sự kiện (event) là một tập con các kết quả của không gian mẫu.
- Khi một sự kiện được xác định, chúng ta có thể tính xác suất (probability) của nó.

Ví dụ

- Thí nghiệm (Experiment)
  - Tung một đồng xu cân bằng 3 lần
- Không gian mẫu (Sample space)
  - S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
- Sự kiện (Event)
  - E = "Có 2 mặt H" = {HHT, HHT, HTH, THH}
- Hàm xác suất (Probability function)
  - Mỗi kết quả (outcome) có xác suất là 1/8

Outcome	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Probability	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Hàm xác suất

- Không gian mẫu rời rạc (discrete sample space)
  - Là một không gian mẫu mà các kết quả có thể liệt kê được, nó có thể hữu hạn hoặc vô hạn.
  - Ví dụ: {H, T}, {1, 2, 3, ⋯}
- Hàm xác suất (probability function)
  - Hàm xác suất P trên một không gian mẫu rời rạc Ω là hàm gán cho mỗi outcome ω một giá trị $P(\omega)$ gọi là xác suất của ω. Hàm P thỏa:

$$\begin{cases} 0 \leq P(\omega) \leq 1 \\ \sum_{i=1}^{n} P(\omega_i) = 1, \Omega = \{\omega_1, \omega_2, \dots \omega_n\} \end{cases}$$

- Xác suất của sự kiện
  - Xác suất của sự kiện E là tổng xác suất của tất cả các outcome trong E.

$$P(E) = \sum_{\omega \in E} P(\omega)$$

Các qui tắc xác suất cơ bản

1. $0 \leq P(A) \leq 1$, với mọi sự kiện A
2. $P(S) = 1$, với S là không gian mẫu của thí nghiệm
3. $P(\overline{A}) = 1 - P(A)$
4. $P(A \cup B) = P(A) + P(B)$, nếu $A \cap B = \emptyset$
5. $P(A \cap B) = P(A) * P(B)$, nếu A và B là độc lập
   - Hai sự kiện A và B là độc lập (independent) nếu A xảy ra không ảnh hưởng đến xác suất B xảy ra và ngược lại.
6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, với mọi sự kiện A, B

Xác suất có điều kiện (conditional probability)

- Xác suất có điều kiện của sự kiện B cho trước A, ký hiệu $P(B|A)$, được định nghĩa:

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

- Nếu hai sự kiện A và B là độc lập thì

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

.

  - Điều kiện tương đương: $P(B|A) = P(B|\overline{A})$

Luật xác suất toàn phần

- Luật nhân (multiplication rule)

$$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$$

- Luật xác suất toàn phần (Law of total probability)
  - Cho $A_1, A_2, \cdots, A_n$ là một phân hoạch (partition) của không gian mẫu Ω, tức $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$ và $A_i \cap A_j = \emptyset, \forall i, j$. Khi đó, với mọi sự kiện B:

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

  - Áp dụng luật nhân ta có thể viết lại

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

Sự độc lập (independence)

- Hai sự kiện được gọi là độc lập (independent) nếu việc biết sự kiện này xảy ra không ảnh hưởng đến xác suất của sự kiện còn lại.
- Định nghĩa: A độc lập với B nếu

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

- Hệ quả: A độc lập với B thì:

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

Định lý Bayes (Bayes' theorem)

- Định lý Bayes: Cho hai sự kiện A và B, khi đó:

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

- Ví dụ: Xét thí nghiệm tung 3 đồng xu cân bằng
  - Gọi A là sự kiện đồng xu thứ nhất có mặt sấp
  - B là sự kiện cả 3 đồng xu đều mặt sấp
  - Khi đó: $P(A) = 1/2, P(B) = 1/8, P(A|B) = 1$
  - Dùng định lý Bayes ta có thể tính $P(B|A)$ như sau:

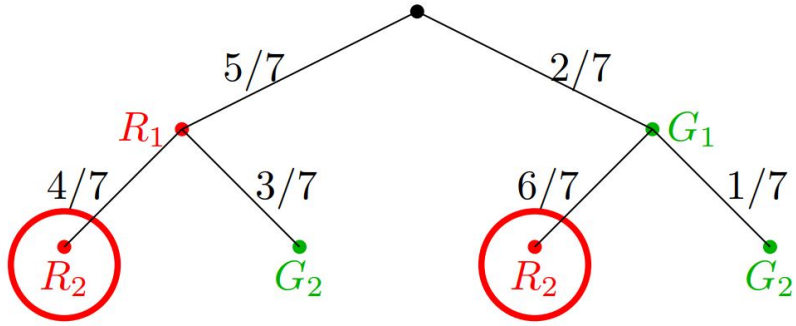
$$P(B|A) = \frac{P(B) \times P(A|B)}{P(A)} = \frac{1/8 \times 1}{1/2} = \frac{1}{4}$$

Ví dụ

- Có 5 bi đỏ và 2 bi xanh trong hộp. Chọn ra ngẫu nhiên một viên và thay thế bằng viên có màu kia rồi sau đó chọn ra ngẫu nhiên một viên thứ hai.
  1. Xác suất viên thứ hai có màu đỏ
  2. Xác suất viên thứ nhất có màu đỏ cho biết viên thứ hai có màu đỏ.

Ví dụ

- Có 5 bi đỏ và 2 bi xanh trong hộp. Chọn ra ngẫu nhiên một viên và thay thế bằng viên có màu kia rồi sau đó chọn ra ngẫu nhiên một viên thứ hai.
  1. Xác suất viên thứ hai có màu đỏ
  2. Xác suất viên thứ nhất có màu đỏ cho biết viên thứ hai có màu đỏ.

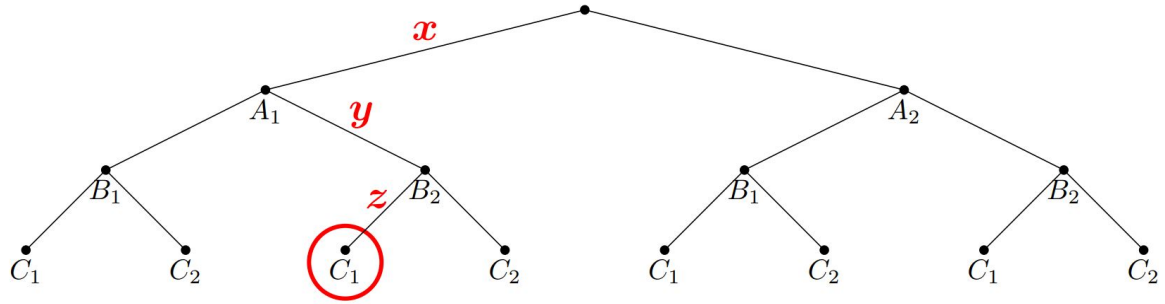


$$P(R_2) = \frac{5}{7} \times \frac{4}{7} + \frac{2}{7} \times \frac{6}{7} = \frac{32}{49}$$

$$P(R_1|R_2) = \frac{P(R_1 \cap R_2)}{P(r_2)} = \frac{20/49}{32/49} = \frac{20}{32}$$

Ví dụ

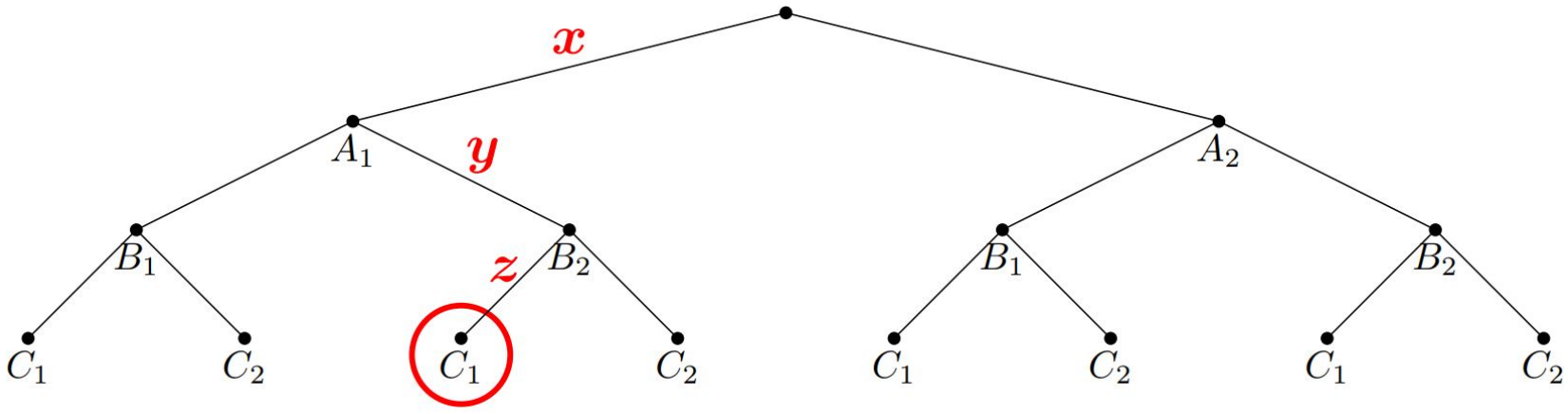
Cho cây xác suất



1. Xác suất mà x thể hiện là gì?
2. Xác suất mà y thể hiện là gì?
3. Xác suất mà z thể hiện là gì?
4. Nút được bao quanh bởi đường tròn thể hiện sự kiện gì?

Ví dụ

Cho cây xác suất



1. Xác suất mà x thể hiện là gì? $\rightarrow P(A_1)$
2. Xác suất mà y thể hiện là gì? $\rightarrow P(B_2|A_1)$
3. Xác suất mà z thể hiện là gì? $\rightarrow P(C_1|A_1 \cap B_2)$
4. Nút được bao quanh bởi đường tròn thể hiện sự kiện gì? $\rightarrow A_1 \cap B_2 \cap C_1$

Ví dụ

Monty Hall problem

- Một cửa có quà (C), 2 cửa không có (G).
- Người chơi chọn một cửa
- Monty mở cửa không có quà trong hai cửa còn lại
- Người chơi được cho phép đổi sự lựa chọn nếu muốn.

Đâu là chiến lược tốt nhất của người chơi?

1. Giữ nguyên
2. Đổi
3. Đổi hay không cũng vậy

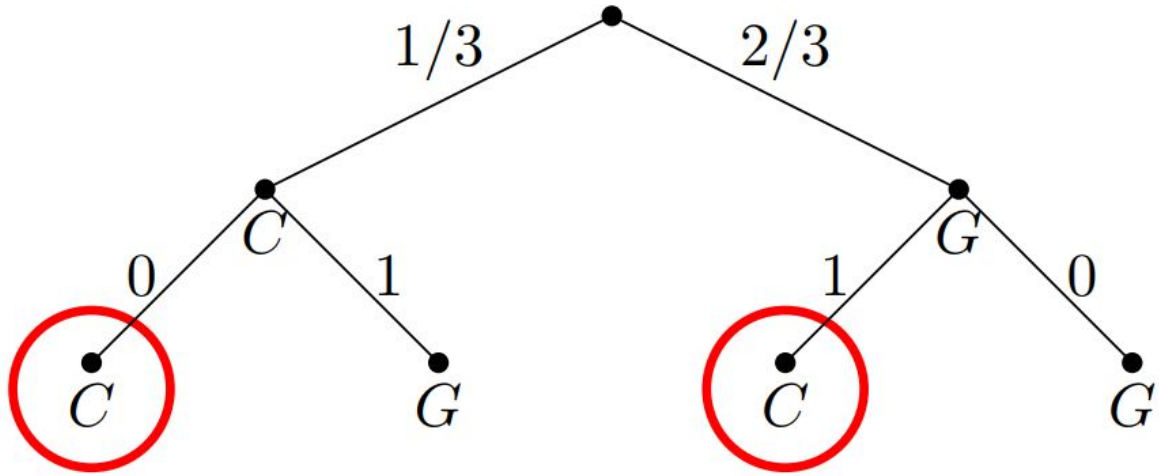
Ví dụ

Monty Hall problem

- Một cửa có quà (C), 2 cửa không có (G).
- Người chơi chọn một cửa
- Monty mở cửa không có quà trong hai cửa còn lại
- Người chơi được cho phép đổi sự lựa chọn nếu muốn.

Đâu là chiến lược tốt nhất của người chơi?

1. Giữ nguyên
2. Đổi
3. Đổi hay không cũng vậy



Ví dụ

- $P(HIV) = 0.0001 \quad P(\overline{HIV}) = 0.9999$
- $P(DT|\overline{HIV}) = 0.01 \quad P(DT|HIV) = 0.99$

$P(HIV|DT) = ?$

Ví dụ

- $P(HIV) = 0.0001 \quad P(\overline{HIV}) = 0.9999$
- $P(DT|\overline{HIV}) = 0.01 \quad P(DT|HIV) = 0.99$

$$P(HIV|DT) = \frac{P(DT|HIV)P(HIV)}{P(DT)}$$

$$P(HIV|DT) = \frac{P(DT|HIV)P(HIV)}{P(DT|HIV)P(HIV) + P(DT|\overline{HIV})P(\overline{HIV})}$$

$$P(HIV|DT) = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} \approx 0.01$$

Biến ngẫu nhiên rời rạc

- Cho Ω là không gian mẫu. Một biến ngẫu nhiên rời rạc (discrete random variable) là hàm

$$X : \Omega \to \mathbb{R}$$

  nhận các giá trị rời rạc.
- Ví dụ: Xét thí nghiệm tung 2 con xúc xắc
  - $\Omega = \{(\omega_1, \omega_2) | \omega_1, \omega_2 \in \{1, 2, 3, 4, 5, 6\}\}$

$$\begin{aligned} &X : \Omega \to R \\ &\omega = (\omega_1, \omega_2) \mapsto X(\omega) = X(\omega_1, \omega_2) = \omega_1 + \omega_2 \end{aligned}$$

Probability mass function (pmf)

- Probability mass function (pmf) của một biến ngẫu nhiên rời rạc (discrete random variable) X là hàm

$$p(a) = P(X = a)$$

  - Khi cần nhấn mạnh biến X ta viết $p_X(a)$.
- Tính chất
  - $0 \leq p(a) \leq 1, \forall a$
  - $p(a) = 0$ khi X không bao giờ nhận giá trị a.
- Khi ta viết $X \leq a$, ta đang nói đến tập tất cả các outcome ω sao cho $X(\omega) \leq a$.

$$X \leq 4 = \{(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)\}$$

Cumulative distribution function (cdf)

- Hàm phân bố tích lũy (cumulative distribution function - cdf) của một biến ngẫu nhiên rời rạc (discrete random variable) X là hàm

$$F(a) = P(X \leq a)$$

- Tính chất
  - $F(a) \leq F(b), \forall a \leq b$
  - $0 \leq F(a) \leq 1, \forall a$
  - $\lim_{x \to \infty} F(x) = 1, \lim_{x \to -\infty} F(x) = 0$
- Ví dụ: Xét thí nghiệm tung hai xúc xắc và X là biến ngẫu nhiên mô tả giá trị của mặt lớn hơn trong hai xúc xắc.

Value	a	1	2	3	4	5	6
pmf	p(a)	1/36	3/36	5/36	7/36	9/36	11/36
cdf	F(a)	1/36	4/36	9/36	16/36	25/36	36/36

Các phân bố rời rạc phổ biến

- Phân bố Bernoulli (Bernoulli distribution)
  - Phân bố Bernoulli mô hình hóa một phép thử (trial) trong một thí nghiệm (experiment) có thể dẫn đến thành công (success) hoặc thất bại (failure).
  - Biến ngẫu nhiên X có phân bố Bernoulli với tham số p, $X \sim Bernoulli(p)$, nếu
    1. X chỉ nhận hai giá trị là 1 (success) hoặc 0 (failure)
    2. $P(X = 1) = p$ và $P(X = 0) = 1 - p$.
  - Nếu X có (tuân theo/được sinh từ) phân bố Bernoulli với tham số p, ta viết $X \sim Bernoulli(p)$.
  - Ví dụ: Xét thí nghiệm tung đồng xu với xác suất mặt sấp là p, X là biến ngẫu nhiên nhận giá trị 1 nếu đồng xu có mặt sấp, và 0 nếu ngược lại. Khi đó, $X \sim Bernoulli(p)$.

value	a	0	1
pmf	p(a)	1 - p	p
cdf	F(a)	1 - p	1

Các phân bố rời rạc phổ biến

- Phân bố nhị thức (Binomial distribution)
  - Phân bố nhị thức $Binomial(n, p)$ mô hình hóa số lần thành công (success) trong n phép thử Bernoulli(p) độc lập.
  - Ví dụ: Xét thí nghiệm tung đồng xu với xác suất mặt sấp là p, X là biến ngẫu nhiên mô tả số lần đồng xu có mặt sấp trong n lần tung. Khi đó, $X \sim Binomial(n, p)$

value	a	k (k = 0, 1, 2, ..., n)
pmf	p(a)	$\left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$

Các phân bố rời rạc phổ biến

- Phân bố hình học (geometric distribution)
  - Phân bố geometric mô hình hóa số lần thất bại trước khi gặp thành công (hoặc ngược lại) trong một chuỗi các phép thử Bernoulli (Bernoulli trial).
  - Biến ngẫu nhiên X có phân bố hình học với tham số p, $X \sim geometric(p)$, nếu
    1. X nhận các giá trị 0, 1, 2, ...
    2. Hàm pmf của X được xác định bởi $p(k) = P(X = k) = (1 - p)^k p$.
- Phân bố đều (uniform distribution)
  - Phân bố đều mô hình thí nghiệm mà tất cả outcome đều có xác xuất như nhau.
  - Biến ngẫu nhiên X có phân bố đều với tham số N, $X \sim uniform(N)$, nếu X nhận các giá trị $1, 2, \cdots, N$ với cùng xác suất $1/N$.

Biến ngẫu nhiên liên tục

- Một biến ngẫu nhiên liên tục (continuous random variable) nhận một miền các giá trị liên tục (ví dụ: $[a, b], [0, \infty]$).
- Một biến ngẫu nhiên X là liên tục nếu có hàm $f(x)$ sao cho với mọi giá trị a, b, ta có

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

  - Hàm $f(x)$ được gọi là hàm mật độ xác xuất (probability density function - pdf).
- Hàm mật độ xác suất luôn thỏa mãn các tính chất sau:
  1. $f(x) \geq 0$
  2. $\int_{-\infty}^{\infty} f(x)dx = 1$
- Chú ý: Hàm mật độ xác suất không phải là hàm xác suất.

Cumulative distribution function (cdf)

- Hàm phân bố tích lũy (cumulative distribution function - cdf) của một biến ngẫu nhiên liên tục (continuous random variable) X là hàm

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

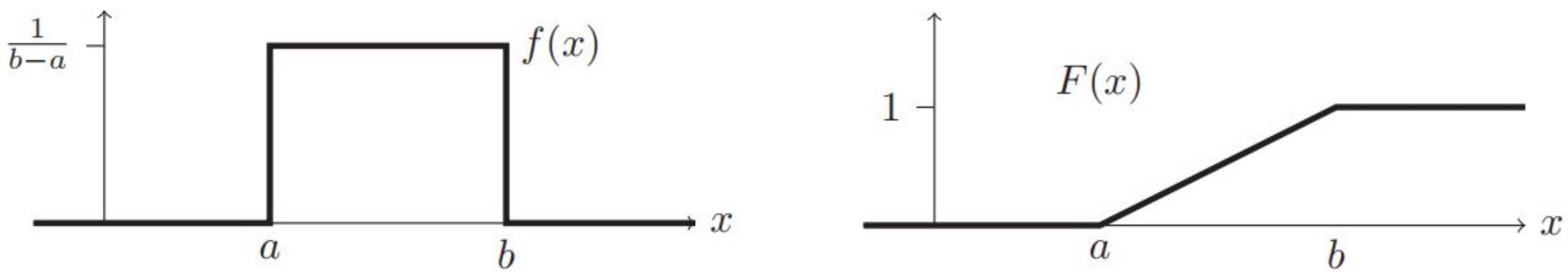
  Trong đó, $f(x)$ là hàm mật độ xác suất của X.
- Ta thường gọi X có phân bố $F(x)$ thay vì X có hàm phân bố tích lũy $F(x)$

Tính chất của hàm phân bố

1. $F(x) = P(X \leq x)$
2. $0 \leq F(x) \leq 1$
3. Nếu $a \leq b$ thì $F(a) \leq F(b)$
4. $\lim\limits_{x \to \infty} F(x) = 1, \lim\limits_{x \to -\infty} F(x) = 0$
5. $P(a \leq X \leq b) = F(b) - F(a)$
6. $F'(x) = f(x)$

Các phân bố liên tục phổ biến

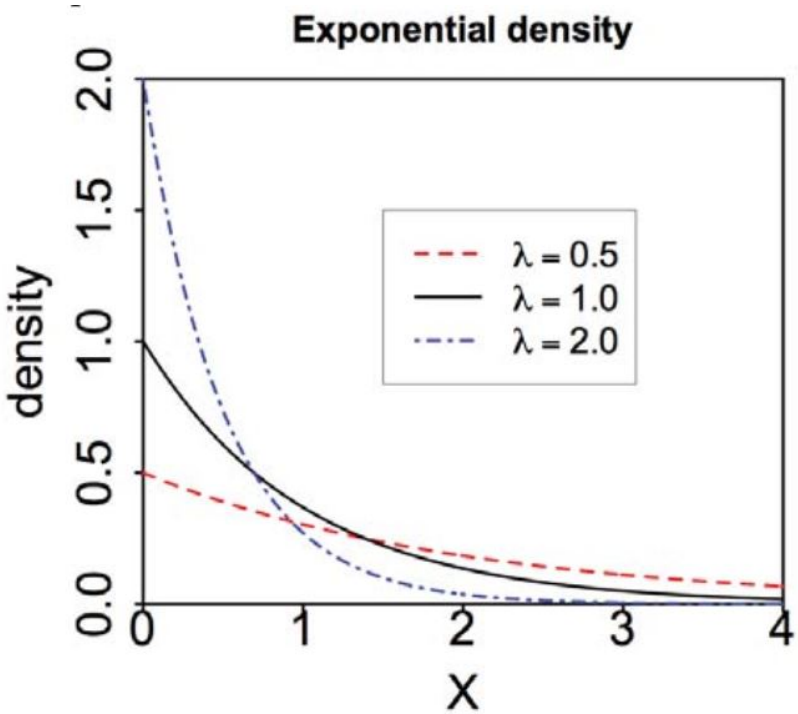
- **Phân bố đều (Uniform distribution)**
  1. Tham số (parameter): a, b
  2. Miền giá trị (range): $[a, b]$
  3. Ký hiệu (notation): $uniform(a, b)$ hay $U(a, b)$
  4. Hàm mật độ (pdf): $f(x) = \frac{1}{b-a}, a \leq x \leq b$
  5. Hàm phân bố (cdf): $F(x) = \frac{x-a}{b-a}, a \leq x \leq b$
  6. Mô hình (model): Tất cả outcome trong range $[a, b]$ có xác suất bằng nhau.

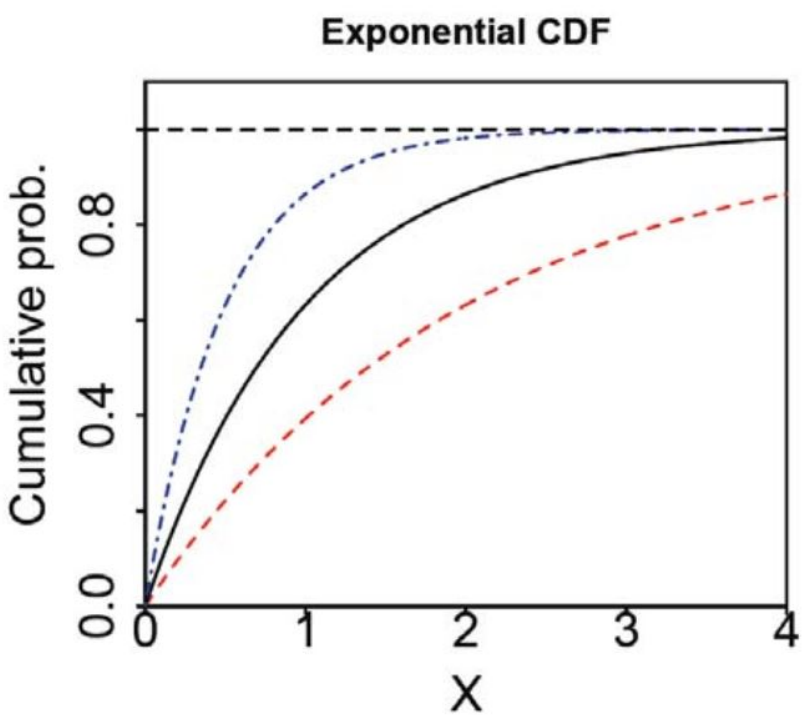


pdf and cdf for uniform(a,b) distribution.

Các phân bố rời rạc phổ biến

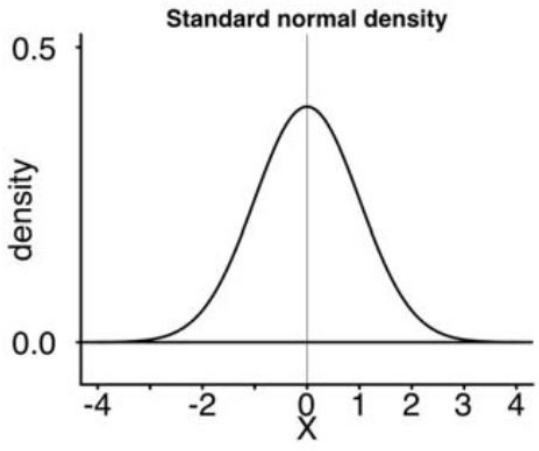
- Phân bố mũ (Exponential distribution)
  1. Tham số (parameter): λ
  2. Miền giá trị (range): $[0, \infty)$
  3. Ký hiệu (notation): $exponential(a, b)$ hay $exp(a, b)$
  4. Hàm mật độ (pdf): $f(x) = \lambda e^{-\lambda x}, x \geq 0$
  5. Hàm phân bố (cdf): $F(x) = 1 - e^{-\lambda x}, x \geq 0$
  6. Mô hình (model): Thời gian chờ cho đến khi một quá trình liên tục thay đổi trạng thái.

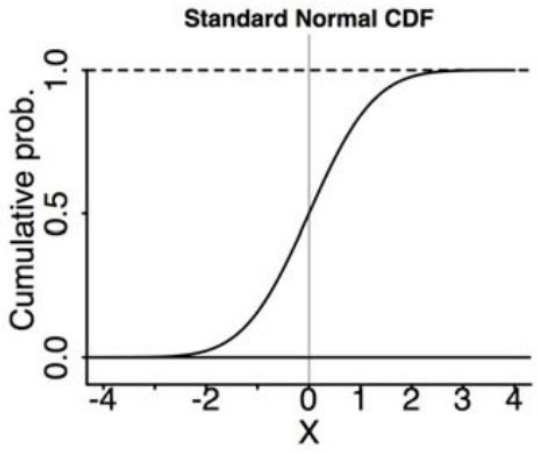


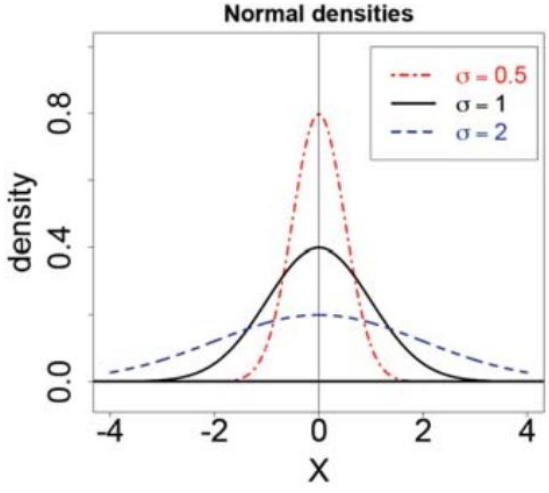


Các phân bố liên tục phổ biến

- Phân bố chuẩn (Normal distribution)
  1. Tham số (parameter): μ, σ
  2. Miền giá trị (range): $(-\infty, \infty)$
  3. Ký hiệu (notation): $normal(\mu, \sigma^2)$ hay $N(\mu, \sigma^2)$
  4. Hàm mật độ (pdf): $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
  5. Hàm phân bố (cdf): Không có công thức (sử dụng bảng để tính F(x))
  6. Mô hình (model): Sai số đo đạc, IQ, chiều cao, ...

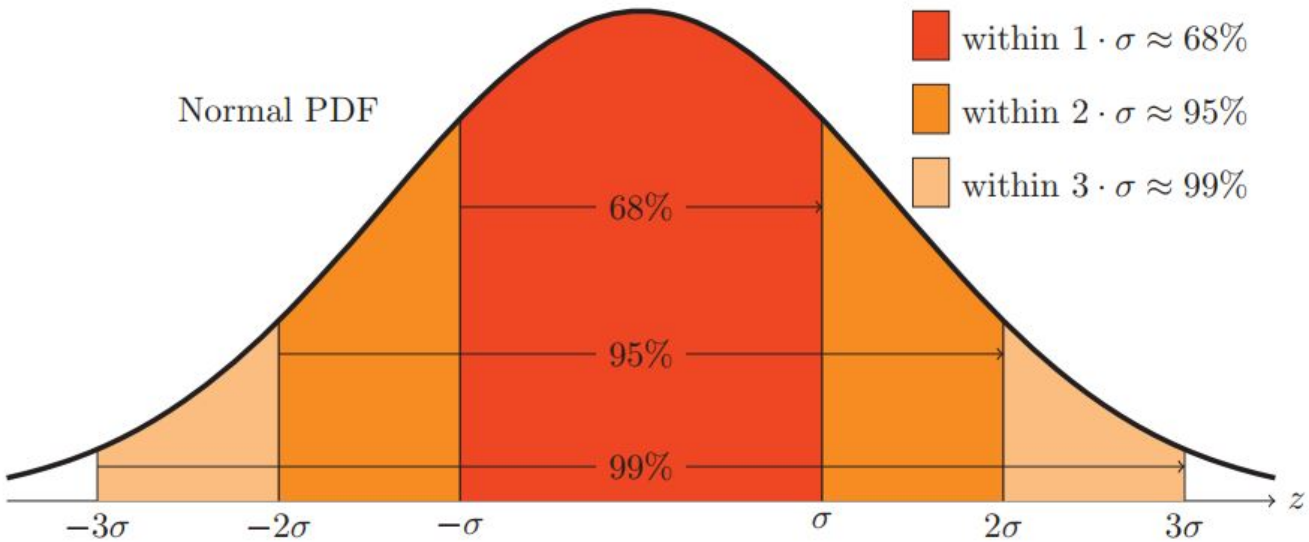






Các phân bố liên tục phổ biến

- Phân bố chuẩn (Normal distribution)
  - Phân bố chuẩn với $\mu = 0, \sigma = 1$ được gọi là standard normal distribution, $N(0, 1)$. Các ký hiệu sau được dành riêng cho standard normal distribution
    - Z để chỉ biến ngẫu nhiên có standard normal distribution,
    - $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ để chỉ standard normal density function,
    - $\Phi(z)$ để chỉ standard normal cumulative distribution function.



Phân bố liên hợp (joint distribution)

Thực tế, ta thường quan tâm đến mối quan hệ của nhiều biến ngẫu nhiên với nhau.

Trường hợp rời rạc

- Gọi X và Y là hai biến ngẫu nhiên rời rạc,
  - X nhận các giá trị $\{x_1, x_2, \cdots, x_n\}$,
  - Y nhận các giá trị $\{y_1, y_2, \cdots, y_m\}$.
  - (X, Y) sẽ nhận các giá trị $\{(x_1, y_1), (x_1, y_2), \cdots, (x_n, y_m)\}$.
- Joint probability mass function (joint pmf) của X và Y là hàm $p(x_i, y_j)$ mô tả xác xuất của $X = x_i$ và $Y = y_j$.
- Tính chất của joint pmf
  1. $0 \leq p(x_i, y_j) \leq 1$
  2. Tổng xác xuất là một, tức

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) = 1$$

Phân bố liên hợp (joint distribution)

- Bảng xác suất liên hợp của X và Y

X/Y	y_1	...	y_j	...	y_m
x_1	$p(x_1, y_1)$	...	$p(x_1, y_j)$	...	$p(x_1, y_m)$
...	...	...	...	...	...
x_i	$p(x_i, y_1)$	...	$p(x_i, y_j)$	...	$p(x_i, y_m)$
...	...	...	...	...	...
x_n	$p(x_n, y_1)$	...	$p(x_n, y_j)$	...	$p(x_n, y_m)$

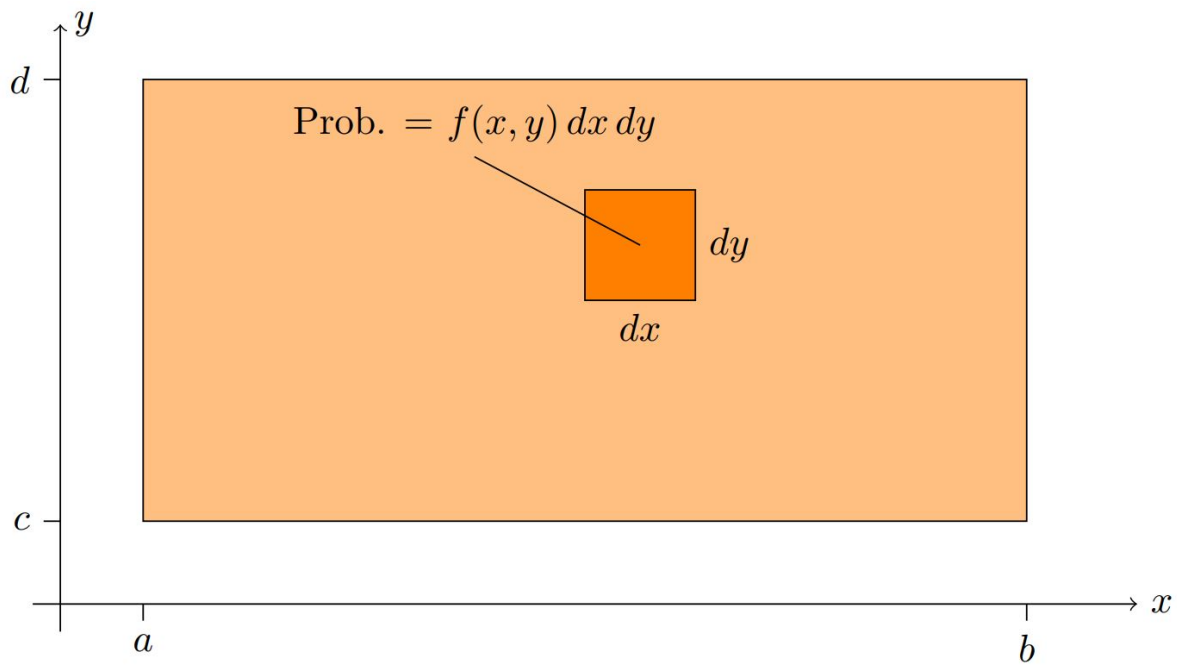
- Ví dụ: Tung hai đồng xu, X, Y lần lượt là giá trị của đồng xu thứ nhất và thứ hai.

X/Y	Head	Tail
Head	1/4	1/4
Tail	1/4	1/4

Phân bố liên hợp (joint distribution)

Trường hợp liên tục

- Cho X và Y là hai biến ngẫu nhiên liên tục,
  - X nhận các giá trị trong $[a, b]$, Y nhận các giá trị trong $[c, d]$.
  - (X, Y) sẽ nhận các giá trị trong $[a, b] \times [c, d]$.
- Joint probability density function (joint pdf) của X và Y là hàm $f(x, y)$ cho mật độ xác xuất tại (x, y). Nghĩa là xác suất (X, Y) nằm trong hình chữ nhật bao quanh điểm (x, y) với chiều rộng dx và chiều cao dy sẽ là $f(x, y)dxdy$.



Joint probability density function (joint pdf)

- Tính chất của joint pdf
  1. $f(x, y) \geq 0$
  2. Tổng xác xuất là một, tức

$$\int_c^d \int_a^b f(x, y) dx dy = 1$$

- Chú ý
  - Joint pdf $f(x, y)$ không phải là một hàm xác suất nên có thể nhận giá trị lớn hơn 1.
  - Tích phân bội (double integral) về mặt khái niệm là tương tự double sum.

Joint cumulative distribution function (Joint cdf)

- Joint cdf của X và Y được định nghĩa là

$$F(x, y) = P(X \leq x, Y \leq y)$$

  - "$X \leq x, Y \leq y$" là viết tắt của "$X \leq x$ và $Y \leq y$".

- Cho X và Y là hai biến ngẫu nhiên liên tục có joint pdf $f(x, y)$ trên miền $[a, b] \times [c, d]$. Joint cdf được định nghĩa là

$$F(x, y) = \int_c^y \int_a^x f(u, v) du dv$$

- Cho X và Y là hai biến ngẫu nhiên rời rạc có joint pmf $p(x_i, y_j)$. Joint cdf được định nghĩa là

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p(x_i, y_j)$$

Tính chất của joint cdf

1. $F(x, y)$ là hàm không giảm (khi x hay y tăng thì $F(x, y)$ tăng hoặc không đổi)
2. $$\lim_{(x,y)\to(-\infty,-\infty)} F(x, y) = 0$$
3. $$\lim_{(x,y)\to(\infty,\infty)} F(x, y) = 1$$
4. $$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y)$$

Phân bố cận biên (Marginal distribution)

- Khi X và Y là các biến ngẫu nhiên có phân bố chung (jointly-distributed random variables), ta có thể muốn chỉ xem xét một trong số chúng, ví dụ X. Ta gọi pmf (hay pdf hay cdf) của X mà không có Y là marginal pmf (hay pdf hay cdf).
- Trường hợp rời rạc
  - Gọi X và Y là các biến ngẫu nhiên rời rạc có phân bố chung có joint pmf là $p(x_i, y_j)$.
  - Marginal pmf của X và Y được định nghĩa là

$$p_X(x_i) = \sum_j p(x_i, y_j), \quad p_Y(y_j) = \sum_i p(x_i, y_j)$$

  - Marginal cdf được định nghĩa là

$$F_X(x) = \sum_{x_i \le x} \sum_j F(x_i, y_j) \quad F_Y(y) = \sum_i \sum_{y_i \le y} F(x_i, y)$$

Phân bố cận biên (Marginal distribution)

- Trường hợp liên tục
  - Gọi X và Y là các biến ngẫu nhiên liên tục có phân bố chung có joint pdf là $f(x, y)$ trên miền $[a, b] \times [c, d]$.
  - Marginal pdf của X và Y được định nghĩa là

$$f_X(x) = \int_c^d f(x, y)dy, \quad f_Y(y) = \int_a^b f(x, y)dx$$

    - So với marginal pmf thì marginal pdf thay tổng (sum) bởi tích phân (integral).
  - Marginal cdf của X và Y được định nghĩa là

$$F_X(x) = F(x, d), \quad F_Y(y) = F(b, y)$$

    - Khi $d \to \infty$ hay $b \to \infty$, thì

$$F_X(x) = \lim_{y\to\infty} F(x, y), \quad F_Y(y) = \lim_{x\to\infty} F(x, y)$$

Sự độc lập (independence)

- Hai biến ngẫu nhiên X và Y có phân bố chung là độc lập (independent) nếu joint cdf bằng tích của các marginal cdf.

$$F(x, y) = F_X(x)F_Y(y)$$

- Với trường hợp rời rạc điều này tương đương với

$$p(x_i, y_j) = p_X(x_i)p_Y(y_j)$$

- Với trường hợp liên tục điều này tương đương với

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

Conditional probability/distribution

- Conditional probability/distribution cho phép ta khám phá việc biết các giá trị của một biến ảnh hưởng đến niềm tin (belief) của ta về các biến khác như thế nào.
- Các ký hiệu
  - $p(x, y) = P(X = x, Y = y)$ là xác suất liên hợp (joint probability) của biến ngẫu nhiên X bằng x và biến ngẫu nhiên Y bằng y.
  - $p(x|y) = P(X = x|Y = y)$ là xác suất có điều kiện (conditional probability) của biến ngẫu nhiên X bằng x cho trước biến ngẫu nhiên Y bằng y.
  - $p_X(x) = P(X = x)$ hay $p_Y(y) = P(Y = y)$ là xác xuất cận biên (marginal probability) của X bằng x hay Y bằng y.

Conditional probability/distribution

- Mối quan hệ giữa joint probability và conditional probability

$$P(X = x, Y = y) = P(X = x) \times P(Y = y|X = x)$$

$$P(X = x, Y = y) = P(Y = y) \times P(X = x|Y = y)$$

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

- Mối quan hệ giữa joint probability và marginal probability

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

$$P(Y = y) = \sum_{x} P(X = x, Y = y)$$

Conditional probability/distribution

- Cho 3 biến ngẫu nhiên X_1, X_2, X_3,

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) =$$
$$P(X_1 = x_1)P(X_1 = x_1|X_2 = x_2)P(X_3 = x_3|X_1 = x_1, X_2 = x_2)$$

- Cho n biến ngẫu nhiên $X_1, X_2, \cdots, X_n$,

$$P\left(\bigcap_{i=1}^{n}(X_i = x_i)\right) = P(X_1 = x_1)\prod_{i=2}^{n} P\left(X_i = x_i|\bigcap_{j=1}^{i-1}(X_j = x_j)\right)$$

Independent and identically distributed (IID)

- Khi các biến ngẫu nhiên là độc lập (independent) và có phân bố giống nhau (identically distributed) ta gọi chúng là độc lập và có phân bố giống nhau (independent and identically distributed (IID).
  - I: independent
  - ID: identically distributed
- Ví dụ: Tung đồng xu cân bằng 100 lần. Gọi X_i là biến ngẫu nhiên có giá trị 1 nếu đồng xu thứ i mặt sấp và 0 nếu mặt ngửa. Khi đó,
  - X_i là độc lập và có phân bố giống nhau (IID), ta viết
  - X_i có phân bố Bernoulli với xác suất $p = 0.5$, $X_i \sim Bernoulli(0.5), IID$