Biến ngẫu nhiên

Biến ngẫu nhiên rời rạc - liên tục

PiMA 2019

Phuc H. Lam

16/07/2019

Nội dung

Sơ lược

Phân phối xác suất

Biến ngẫu nhiên rời rạc (Discrete random variable) có miền giá trị là một tập hữu hạn hoặc vô hạn đếm được.

Phân phối xác suất cho biết xác suất tổng (là 1) được phân phối như thế nào cho các giá trị khác nhau của biến ngẫu nhiên X.

Phân phối xác suất

Ví dụ 1

Tung một xúc xắc 6 mặt đồng chất, ta thu được các kết quả $\{1, 2, \ldots, 6\}$.

Biến ngẫu nhiên X: $X(''k'') = k$, $\forall k = \overline{1; 6}$.

Phân phối xác suất: $P(X = k) = \dfrac{1}{6}$, $\forall k = \overline{1; 6}$.

Ví dụ 2

Gọi X là số lần liên tiếp tung một đồng xu để ra mặt ngửa.

Biến ngẫu nhiên X: $X(k) = k$, $\forall k \in \mathbb{N}^*$.

Phân phối xác suất: $P(X = k) = \dfrac{1}{2^k}$, $\forall k \in \mathbb{N}^*$.

Hàm khối xác suất

Hàm khối xác suất (probability mass function - pmf)

$$p_X(x) = \begin{cases} P(X = x) & (x \in D) \\ 0 & (x \in \mathbb{R} \setminus D), \end{cases}$$

với D là tập các giá trị biến rời rạc X có thể nhận được.

Hàm khối xác suất thoả 2 điều kiện:

$$\begin{cases} 0 \leq p_X(x) \leq 1, \ \forall x \in \mathbb{R} \\ \sum\limits_{x \in D} p_X(x) = 1. \end{cases}$$

Hàm khối xác suất

Ví dụ 1

Tung xúc xắc 6 mặt, đồng chất.

$$\Rightarrow p_X(x) = \begin{cases} \frac{1}{6} & (x \in \{1, 2, \ldots, 6\}), \\ 0 & (\text{còn lại}). \end{cases}$$

Ở đây, $\sum_{x=1}^{6} p_X(x) = \sum_{x=1}^{6} \frac{1}{6} = 1$.

Ví dụ 2

$p_X(x)$ là hàm khối xác suất cho số lần tung liên tiếp một đồng xu đồng chất để ra được mặt ngửa.

$$\Rightarrow p_X(x) = \begin{cases} \frac{1}{2^x} & (x \in \mathbb{N}^*), \\ 0 & (\text{còn lại}). \end{cases}$$

Ở đây, $\sum_{x \in \mathbb{N}^*} p_X(x) = \sum_{x=1}^{\infty} \frac{1}{2^x} = 1$.

Hàm phân phối tích luỹ

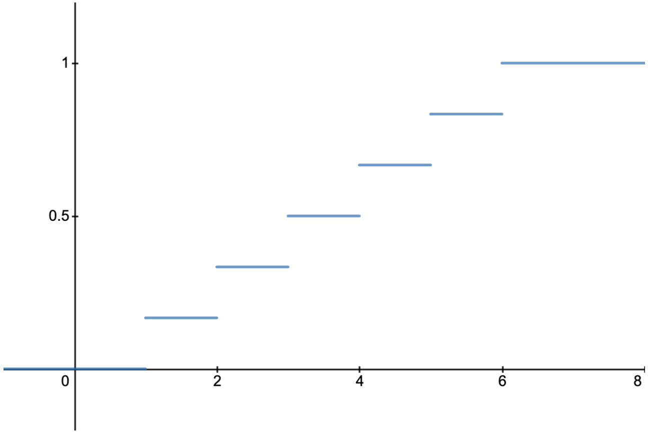
Hàm phân phối tích luỹ (cumulative distribution function - cdf)

$$F_X(x) = P(X \leq x) = \sum_{y:y \leq x} p_X(x)$$

Ngoài ra, với biến ngẫu nhiên X nhận giá trị nguyên, ta có:

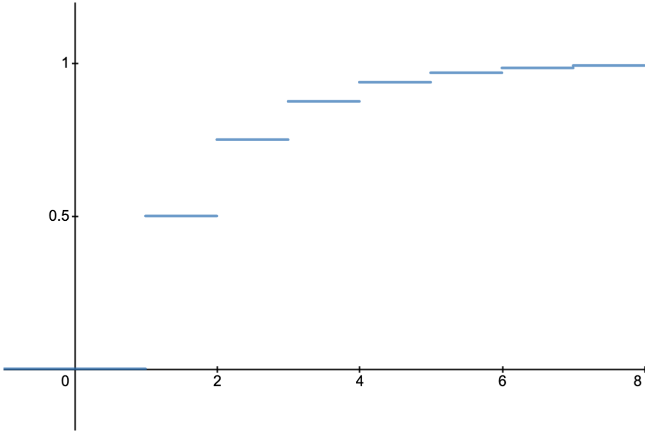
- $F_X(x) = F_X(\lfloor x \rfloor)$
- $P(X < x) = P(X \leq x) - P(X = x) = F_X(x) - p_X(x)$
- $P(a \leq X \leq b) = F_X(b) - F_X(a-1) \forall a, b \in Z$
- $p_X(a) = F_X(a) - F_X(a-1) \forall a \in Z$

Hàm phân phối tích luỹ



Hình: Hàm phân phối tích luỹ cho ví dụ 1

Hàm phân phối tích luỹ



Hình: Hàm phân phối tích luỹ cho ví dụ 2

Giá trị kỳ vọng

Giá trị kỳ vọng (hay giá trị trung bình) của biến ngẫu nhiên X (expectation)

$$E[X] = \sum_{x \in D} x \cdot p(x),$$

với D là tập giá trị biến ngẫu nhiên X có thể nhận được.

Ký hiệu khác: μ_X, μ.

Ngoài ra, ta còn quan tâm đến kỳ vọng của hàm số liên quan tới biến ngẫu nhiên đó:

Giá trị kỳ vọng của hàm số h(x) của biến ngẫu nhiên X

$$E[h(X)] = \sum_{x \in D} h(x) \cdot p(x),$$

Giá trị kỳ vọng

Ví dụ 1

Tung xúc xắc 6 mặt, đồng chất.

Giá trị kỳ vọng:

$$E[X] = \sum_{k=1}^{6} k \cdot p_X(k) = \frac{1}{6} \sum_{k=1}^{6} k = 3.5$$

Ví dụ 2

Chọn một số nguyên dương k bất kỳ với xác suất $\frac{1}{2^k}$.

Giá trị kỳ vọng:

$$E[X] = \sum_{k=1}^{\infty} k \cdot p_X(k) = \sum_{k=1}^{\infty} \frac{k}{2^k} = 2$$

Giá trị kỳ vọng

- Với a, b là các hằng số:

$$\begin{aligned} E[aX+b] &= \sum_{x\in D}(ax+b)\cdot p(x) \\ &= a\sum_{x\in D} x\cdot p(x) + b\sum_{x\in D} p(x) \\ &= \boxed{aE[X]+b} \end{aligned}$$

- Với X, Y là các biến ngẫu nhiên rời rạc:

$$\begin{aligned} E[aX+bY] &= \sum_{x\in D_X, y\in D_Y}(ax+by)\cdot p(x,y) \\ &= \sum_{x\in D_X} ax\left(\sum_{y\in D_Y} p(x,y)\right) + \sum_{y\in D_Y} by\left(\sum_{x\in D_X} p(x,y)\right) \\ &= a\sum_{x\in D_X} x\cdot p(x) + b\sum_{y\in D_Y} y\cdot p(y) \\ &= \boxed{aE[X]+bE[Y]} \end{aligned}$$

Phương sai - độ lệch chuẩn

Để đo mức độ phân tán của dữ liệu, ta dùng đến **phương sai** và **độ lệch chuẩn**.

Phương sai của biến ngẫu nhiên X (variance)

$$\mathrm{Var}[X] = \sum_{x \in D} (x - E[X])^2 p(x) = E[(X - E[X])^2]$$

Ký hiệu khác: σ_X^2, σ^2.

Độ lệch chuẩn (standard deviation)

$$\sigma_X = \sqrt{\sigma_X^2}$$

Phương sai (và độ lệch chuẩn càng lớn) $\rightarrow$ phân phối càng phân tán ra xa so với trung bình.

Phương sai - độ lệch chuẩn

Cách tính khác của phương sai:

$$\begin{aligned}
\mathrm{Var}[X] &= \sum_{x\in D}(x - E[X])^2 \cdot p(x) \\
&= \sum_{x\in D} x^2 \cdot p(x) - 2E[X]\sum_{x\in D} x \cdot p(x) + E[X]^2 \sum_{x\in D} p(x) \\
&= E[X^2] - 2E[X] \cdot E[X] + E[X]^2 \\
&= \boxed{E[X^2] - E[X]^2}
\end{aligned}$$

Phương sai - độ lệch chuẩn

Phương sai của hàm số h(x) của biến ngẫu nhiên X

$$\text{Var}[h(X)] = \sum_{x\in D}(h(x) - E[h(X)])^2 p(x)$$

Với a, b là các hằng số:

$$\begin{aligned}
\text{Var}[aX+b] &= \sum_{x\in D}(ax+b-E[aX+b])^2 p(x) \\
&= \sum_{x\in D}(ax+b-(aE[X]+b))^2 p(x) \\
&= \sum_{x\in D} a^2(x-E[X])^2 p(x) \\
&= a^2 \sum_{x\in D}(x-E[X])^2 p(X) \\
&= \boxed{a^2\text{Var}[X]}
\end{aligned}$$

Phân phối Bernoulli

Một thí nghiệm có 2 kết quả "thành công", "thất bại"; xác suất thành công p; biến ngẫu nhiên X thoả $X(\text{"thành công"}) = 1$ và $X(\text{"thất bại"}) = 0$.

Phân phối Bernoulli (Bernoulli distribution)

Biến ngẫu nhiên X được gọi là có phân phối Bernoulli nếu hàm khối xác suất $p_X(x)$ thoả:

$$p_X(x) = \begin{cases} p & (x = 1) \\ 1 - p & (x = 0) \\ 0 & (\text{còn lại}) \end{cases}$$

Ký hiệu: $X \sim B(1, p)$, với $p \in (0, 1)$.

Đặc trưng

- Kỳ vọng: $E[X] = p$
- Phương sai: $\mathrm{Var}[X] = p(1 - p)$

Phân phối Bernoulli

Ví dụ 3

Cho một câu hỏi trắc nghiệm 5 câu; có đúng 1 đáp án chính xác.

Gọi X là số câu trả lời đúng. $\Rightarrow \begin{cases} p_X(1) = P(X = 1) = 0.2 \\ p_X(0) = P(X = 0) = 0.8 \end{cases}$

Ở đây, $X \sim B(1, 0.2)$.

Phân phối nhị thức (Binomial distribution)

Một thí nghiệm có 2 kết quả "thành công", "thất bại" được lặp lại n lần độc lập; xác suất thành công p ở mỗi thí nghiệm.
Xác suất để thành công k lần là $\binom{n}{k} p^k (1-p)^{n-k}$.

Phân phối nhị thức

Biến ngẫu nhiên X được gọi là có phân phối nhị thức nếu hàm khối xác suất $p_X(x)$ thoả:

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & (x = 0, 1, 2, \ldots, n) \\ 0 & (\text{còn lại}) \end{cases}$$

Ký hiệu: $X \sim B(n,p)$, với $p \in (0,1)$.

Đặc trưng

- Kỳ vọng: $E[X] = np$
- Phương sai: $\mathrm{Var}[X] = np(1-p)$

Phân phối nhị thức

Ví dụ 4

Cho một bài trắc nghiệm gồm 40 câu trắc nghiệm như ở Ví dụ 3.
Gọi X là số câu trả lời đúng. Khi đó, $X \sim B(40, 0.2)$.

Sơ lược

Phân phối xác suất

Biến ngẫu nhiên liên tục (Continuous random variable) thoả 2 điều kiện:

- Miền giá trị là một (tập hợp) các đoạn trên trục số
- $P(X = c) = 0 \; \forall c$.

Phân phối xác suất cho biết xác suất tổng (là 1) được phân phối như thế nào cho các "đoạn" giá trị khác nhau của biến ngẫu nhiên X.

Hàm mật độ xác suất

Hàm mật độ xác suất (probability density function - pdf)

$$f_X(x) := P(a \leq X \leq b) = \int_a^b f(x)dx \ \forall a \leq b \in \mathbb{R}$$

Hàm mật độ xác suất thoả 2 điều kiện:

$$\begin{cases} 0 \leq f_X(x) \leq 1 \ \forall x \in \mathbb{R} \\ \displaystyle\int_{-\infty}^{\infty} f(x)dx = 1 \end{cases}$$

Hàm mật độ xác suất

Ví dụ 5

Chọn X ngẫu nhiên trong đoạn $[0; 2\pi]$.

$$\Rightarrow f_X(x) = \begin{cases} \dfrac{1}{2\pi} & (x \in [0; 2\pi]) \\ 0 & (\text{còn lại}) \end{cases}$$

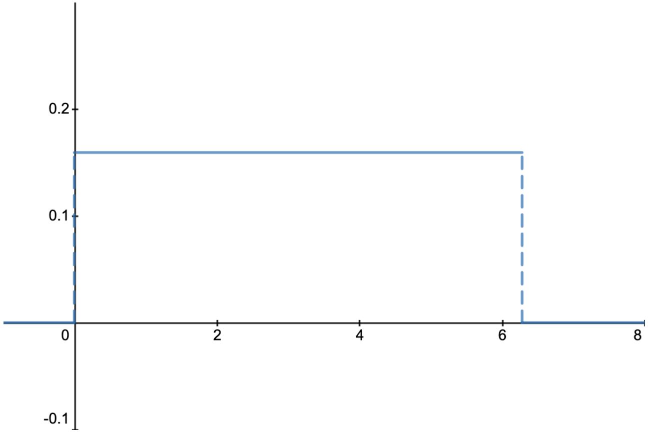
Ví dụ 6

Cho X có hàm mật độ

$$f_X(x) = \begin{cases} ax^n & (x \in [0; 1]) \\ 0 & (\text{còn lại}) \end{cases}$$

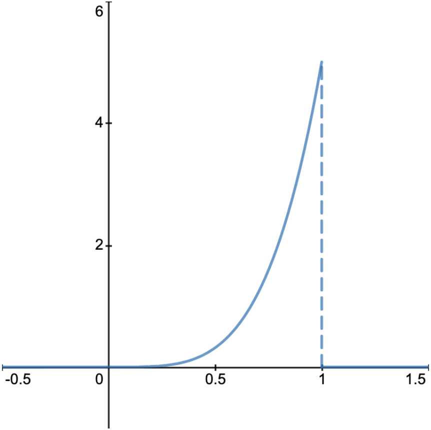
Khi đó, từ điều kiện thứ 2 của hàm mật độ xác suất, tính được $a = n + 1$.

Hàm mật độ xác suất



Hình: Hàm mật độ xác suất cho ví dụ 5

Hàm mật độ xác suất



Hình: Hàm mật độ xác suất cho ví dụ 6, với $n = 4$

Hàm phân phối tích luỹ

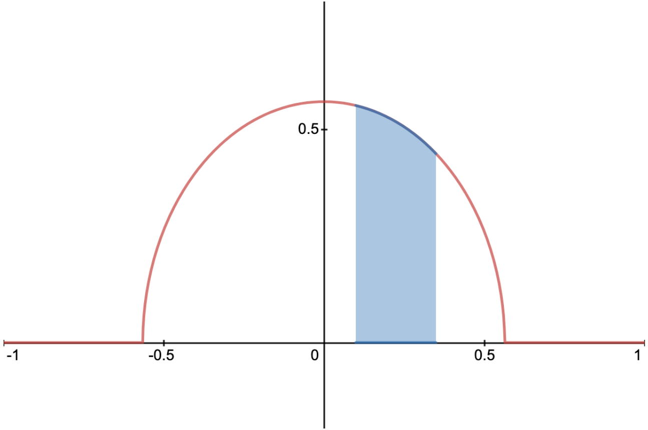
Hàm phân phối tích luỹ (cumulative distribution function - cdf)

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Như vậy, ta có: $\boxed{f_X(x) = F'_X(x)}$.

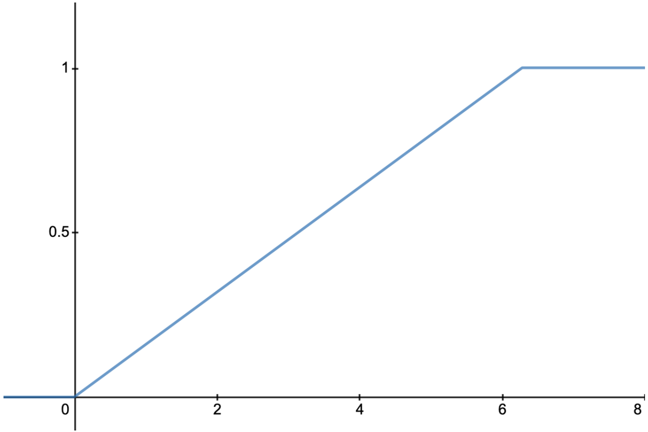
Khác với biến rời rạc, ta luôn có
$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$
$= F_X(b) - F_X(a) = \boxed{\int_a^b f_X(x)dx} \; \forall a < b \in \mathbb{R}$.

Hàm phân phối tích luỹ



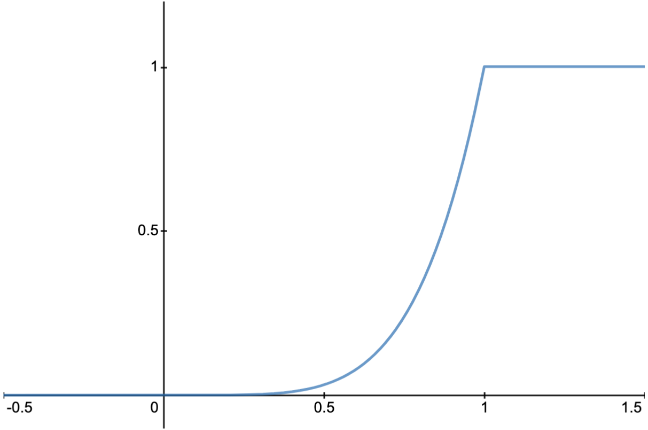
Hình: Ý nghĩa hình học của hàm phân phối tích luỹ

Hàm phân phối tích luỹ



Hình: Hàm phân phối tích luỹ cho ví dụ 5

Hàm phân phối tích luỹ



Hình: Hàm phân phối tích luỹ cho ví dụ 6, với $n = 4$

Đặc trưng

Giá trị kỳ vọng

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$$

Ngoài ra, với hàm số liên quan tới biến ngẫu nhiên,

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f_X(x)dx$$

Ký hiệu khác: μ_X, μ.

Đặc trưng

Phương sai

$$\text{Var}[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx = E[(X - E(X))^2]$$

Ký hiệu khác: σ_X^2, σ^2

Độ lệch chuẩn

$$\sigma_X = \sqrt{\sigma_X^2}$$

Đặc trưng

Các công thức sau vẫn đúng đối với biến ngẫu nhiên X, Y liên tục:

- $E[aX + bY] = aE[X] + bE[Y]$
- $E[aX + b] = aE[X] + b$
- $\mathrm{Var}[X] = E[X^2] - E[X]^2$
- $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$

Đặc trưng

Phân vị thứ 100p (pth quantile)

Cho $p \in [0; 1]$. Phân vị thứ 100p (ký hiệu $\eta(p)$) của X là giá trị thoả

$$p = F_X(\eta(p))$$

Hoặc ta cũng có thể viết $\eta(p) = F_X^{-1}(p)$.

$F_X(x)$ là hàm liên tục, tăng từ 0 tới 1, nên giá trị $\eta(p)$ luôn tồn tại.

So sánh: Rời rạc - liên tục

Phân phối	Rời rạc	Liên tục
Miền giá trị	Hữu hạn hoặc vô hạn đếm được	Vô hạn không đếm được
Hàm	Hàm khối xác suất $$\sum_{x\in D} p_X(x) = 1$$	Hàm mật độ xác suất $$\int_{-\infty}^{\infty} f_X(x)dx = 1$$
Hàm phân phối tích luỹ	$F_X(x) = \sum_{y:\ y\leq x} p_X(y)$	$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$
Giá trị kỳ vọng	$E[X] = \sum_{x\in D} x \cdot p_X(x)$	$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$
Phương sai	$\sum_{x\in D} (x - E[X])^2 p_X(x) = E\left[(X - E(X))^2\right]$	$\int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)dx = E\left[(X - E(X))^2\right]$

Sơ lược

Phân phối đều (continuous uniform distribution/ rectangular distribution)

Chọn một giá trị bất kỳ trên đoạn $[a; b]$.

Phân phối đều

Biến ngẫu nhiên liên tục X có phân phối đều trên đoạn $[a; b]$ (ký hiệu $X \sim U(a, b)$) nếu hàm mật độ xác suất $f_X(x)$ thoả

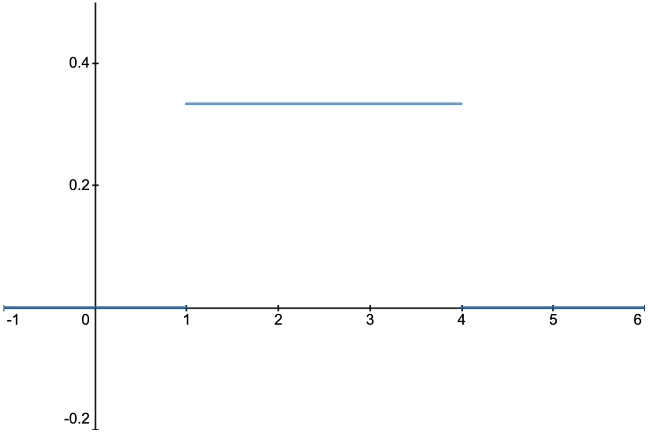
$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & (x \in [a; b]) \\ 0 & (\text{còn lại}) \end{cases}$$

Từ đó, ta tính được hàm phân phối tích luỹ của X như sau:

Hàm phân phối tích luỹ của $X \sim U(a, b)$

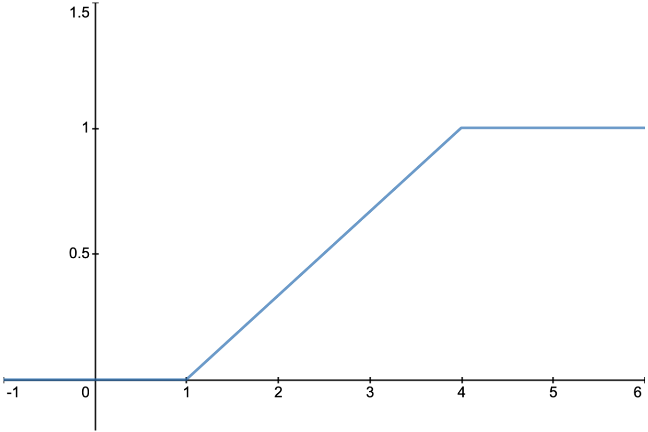
$$F_X(x) = \begin{cases} 0 & (x < a) \\ \dfrac{x-a}{b-a} & (x \in [a; b]) \\ 1 & (x > b) \end{cases}$$

Phân phối đều



Hình: Hàm mật độ xác suất với $a = 1$, $b = 4$

Phân phối đều



Hình: Hàm phân phối tích luỹ với $a = 1$, $b = 4$

Đặc trưng của phân phối đều

Đặc trưng của phân phối đều

- Kỳ vọng: $E[X] = \dfrac{a+b}{2}$
- Phương sai: $\mathrm{Var}[X] = \dfrac{(b-a)^2}{12}$

Phân phối chuẩn

Phân phối chuẩn (Gauss) (normal distribution)

Hàm mật độ xác suất:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Ký hiệu: $X \sim N(\mu, \sigma^2)$, với $\sigma > 0$.

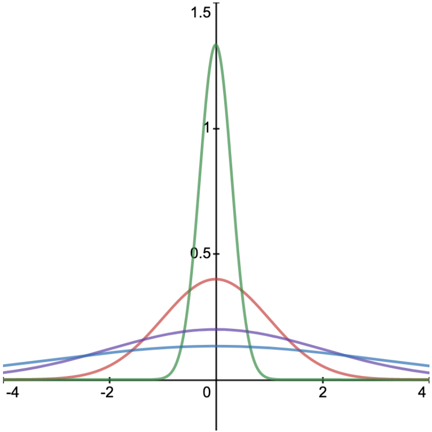
Với $f_X(x)$ như trên, ta luôn luôn có: $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Đặc trưng của phân phối chuẩn

Cho $X \sim N(\mu, \sigma^2)$. Khi đó:

- Kỳ vọng: $E[X] = \mu$
- Phương sai: $\text{Var}[X] = \sigma^2$ (Độ lệch chuẩn: σ)

Phân phối chuẩn



Hình: Phân phối chuẩn với $\mu = 0$; $\sigma = 3$; $\sigma = 2$; $\sigma = 1$; $\sigma = 0.3$

Phân phối chuẩn tắc

Phân phối chuẩn tắc

$Z \sim N(0,1)$ được gọi là phân phối chuẩn tắc, với

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Ký hiệu hàm phân phối tích luỹ của Z là

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

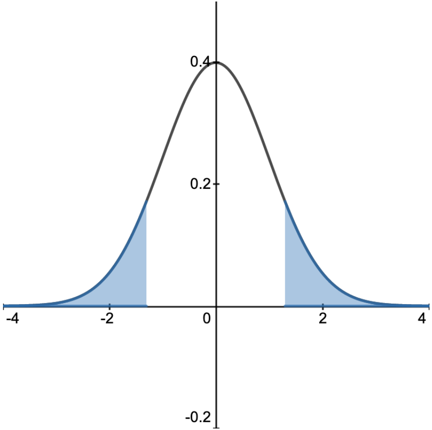
Kết quả: Với $X \sim N(\mu, \sigma^2)$, ta có: $\boxed{aX + b \sim N(a\mu + b, a^2\sigma^2)}$.

$\dfrac{X-\mu}{\sigma} \sim N(0,1)$. (Phân phối chuẩn $\rightarrow$ Phân phối chuẩn tắc)

$X \sim N(\mu, \sigma)$. Suy ra $P(a \leq X \leq b) = P\left(\dfrac{a-\mu}{\sigma} \leq \dfrac{X-\mu}{\sigma} \leq \dfrac{b-\mu}{\sigma}\right)$

$= \Phi\left(\dfrac{b-\mu}{\sigma}\right) - \Phi\left(\dfrac{a-\mu}{\sigma}\right)$

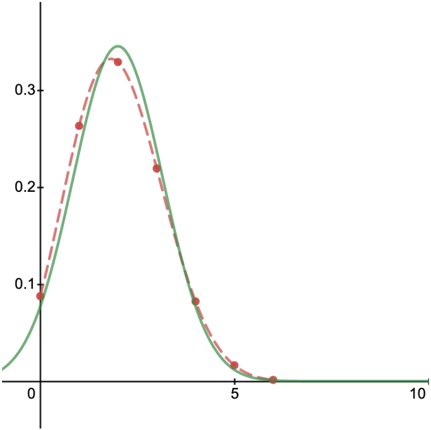
Phân phối chuẩn tắc



Hình: $\boxed{\Phi(x) + \Phi(-x) = 1}$

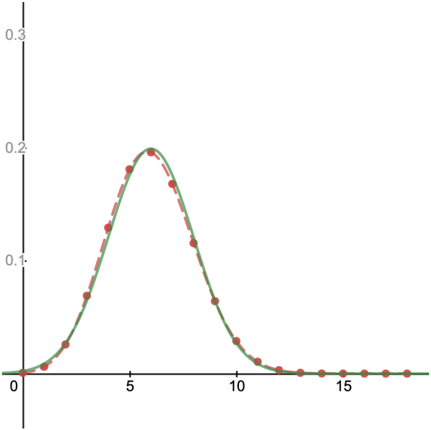
$$P(-a \leq x \leq a) = \Phi(a) - \Phi(-a) = 2\Phi(a) - 1$$

Ứng dụng của phân phối chuẩn



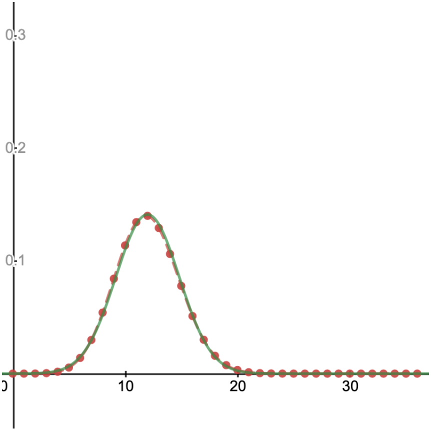
Hình: Phân phối $B(6, \frac{1}{3})$ và $N(6 \cdot \frac{1}{3}, 6 \cdot \frac{1}{3} \cdot \frac{2}{3})$

Ứng dụng của phân phối chuẩn



Hình: Phân phối $B(18, \frac{1}{3})$ và $N(18 \cdot \frac{1}{3}, 18 \cdot \frac{1}{3} \cdot \frac{2}{3})$

Ứng dụng của phân phối chuẩn



Hình: Phân phối $B(36, \frac{1}{3})$ và $N(36 \cdot \frac{1}{3}, 36 \cdot \frac{1}{3} \cdot \frac{2}{3})$

Ứng dụng của phân phối chuẩn

Vấn đề: với n lớn, việc tính $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ (và xét $B(n,p)$) trở nên khó khăn.

Ước lượng phân phối nhị thức

$X \sim B(n,p)$. Nếu n đủ lớn và p không quá gần 0 hoặc 1 thì
$B(n,p) \approx N(np, np(1-p))$.

Xấp xỉ thường tốt khi $np(1-p) > 10$, hoặc khi cả $np > 5$ và $n(1-p) > 5$.

Ứng dụng của phân phối chuẩn

Ví dụ 5

Bài toán: Tung đồng xu cân đối 10000 lần. Tính xác suất để số lần xuất hiện mặt ngửa nằm trong đoạn $[4850; 5100]$.

Giải: Gọi $X \sim B(10000, 0.5)$ là số lần xuất hiện mặt ngửa.
Khi đó $X \approx N(10000 \cdot 0.5, 10000 \cdot 0.5 \cdot (1 - 0.5)) \sim N(5000, 50^2)$

$$\begin{aligned} P(4850 \leq X \leq 5100) &= P\left(\frac{4850 - 5000}{50} \leq \frac{X - 5000}{50} \leq \frac{5100 - 5000}{50}\right) \\ &= P\left(-3 \leq \frac{X - 5000}{50} \leq 2\right) \\ &= \Phi(2) - \Phi(-3) \approx 0.9759 \end{aligned}$$

Xác suất chính xác: $P(4850 \leq X \leq 5100) = \sum_{k=4850}^{5100} \binom{10000}{k} 2^{-10000} \approx 0.9765$

Ứng dụng của phân phối chuẩn

Ước lượng phân phối nhị thức trên là một trường hợp đặc biệt của **Định lý giới hạn trung tâm** (Central Limit Theorem):

Định lý giới hạn trung tâm

Cho $X_1, X_2, \cdots$ là các biến ngẫu nhiên được định nghĩa trên cùng một không gian xác suất, có cùng phân phối, độc lập lẫn nhau, cùng có kỳ vọng μ và phương sai σ^2 hữu hạn.
Xét tổng $S_n = \sum_{i=1}^{n} X_i$. Khi đó

$$S_n \sim N(n\mu, n\sigma^2)$$

khi n tiến về vô cùng.

Câu hỏi: các biến X_i trong trường hợp phân phối nhị thức có phân phối gì?

Ứng dụng của phân phối chuẩn

Hiệu chỉnh liên tục (continuity correction)

Khi dùng phân phối chuẩn xấp xỉ biến X rời rạc nhận giá trị nguyên và $a \in Z$:

$$\begin{cases} P(X \leq a) = P(X \leq a + 0.5) \\ P(X \geq a) = P(X \geq a - 0.5) \end{cases}$$

Xấp xỉ này tốt hơn khi sử dụng mút a.

Ứng dụng của phân phối chuẩn

Ví dụ 6

Xác suất một học sinh rớt môn là 0.8. Xét một lớp có 64 học sinh. Gọi X là số học sinh rớt môn (như vậy, $X \sim B(64, 0.8)$). Tính xác suất để có không quá 50 học sinh rớt môn!

- Tính theo phân phối nhị thức: $P(X \leq 50) = \sum_{i=1}^{50} 0.8^i 0.2^{64-i} \approx 0.4019$
- Xấp xỉ bằng phân phối chuẩn: $X \approx N(64 \cdot 0.8, 64 \cdot 0.8 \cdot 0.2) = N(51.2, 10.24^2) \Rightarrow$ $P(X \leq 50) \approx \Phi(\dfrac{50 - 51.2}{10.24}) \approx 0.3557$
- Xấp xỉ bằng phân phối chuẩn (có hiệu chỉnh liên tục): $P(X < 50.5) \approx \Phi(\dfrac{50.5 - 51.2}{10.24}) \approx 0.4129$

Nhận xét: hiệu chỉnh liên tục cho xấp xỉ tốt hơn.