Biến ngẫu nhiên rời rạc

Bài 4

Linh Truong

Sep, 2024

Mục tiêu

- Giới thiệu về biến ngẫu nhiên
- Giới thiệu một số phân phối xác suất của biến ngẫu nhiên rời rạc

Nội dung

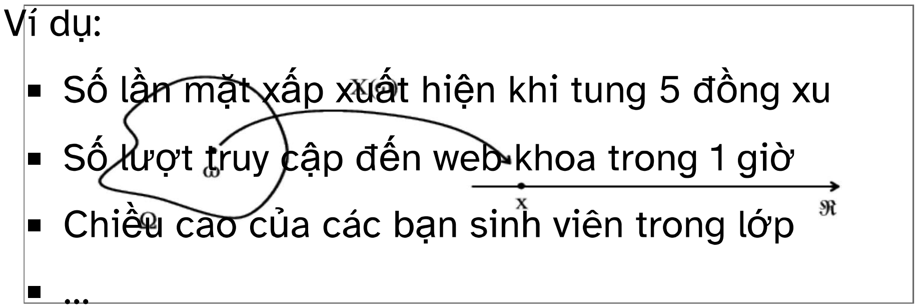
- Biến ngẫu nhiên
- Biến ngẫu nhiên rời rạc
- Phân phối Bernoulli
- Phân phối Nhị thức
- Phân phối Poisson

Biến ngẫu nhiên

- **Định nghĩa**: Giả sử có một phép thử có không gian mẫu là Ω. **Một biến ngẫu nhiên** (rv) là một hàm số học của các kết quả của phép thử ω

$$\begin{aligned} X: \Omega &\to \mathbb{R} \\ \omega &\mapsto X = X(\omega) \end{aligned}$$

- Ví dụ:
  - Số lần mặt xấp xuất hiện khi tung 5 đồng xu
  - Số lượt truy cập đến web khoa trong 1 giờ
  - Chiều cao của các bạn sinh viên trong lớp
  - ...



Biến ngẫu nhiên

TKUD - 2024

Ví dụ

Phân loại các biến ngẫu nhiên dưới đây (rời rạc hay liên tục)?

Mô tả	Range	Type
X, số lần mặt sấp xuất hiện khi tung một đồng xu n lần	$\{0, 1, \ldots, n\}$	rời rạc
N, số lượng em bé sinh trong một năm	$\{0, 1, \ldots, n\}$	rời rạc
B, thời gian (giây) đợi xe buýt kế tiếp	$[0, \infty)$	liên tục
H, chiều cao của một người	$(0, \infty)$	liên tục
C, nhiệt độ (Celcisus) của nước	$(0, 100)$	rời rạc

Hàm độ lớn xác suất (PMF)

- **Hàm độ lớn xác suất (PMF)** của một biến ngẫu nhiên rời rạc X dùng để tính xác suất của mọi giá trị mà biến ngẫu nhiên X có thể nhận được. Nghĩa là:

$$\begin{aligned} p_X : \Omega_X &\rightarrow [0,1] \\ k &\mapsto p_X(k) = \mathbb{P}(X = k) \end{aligned}$$

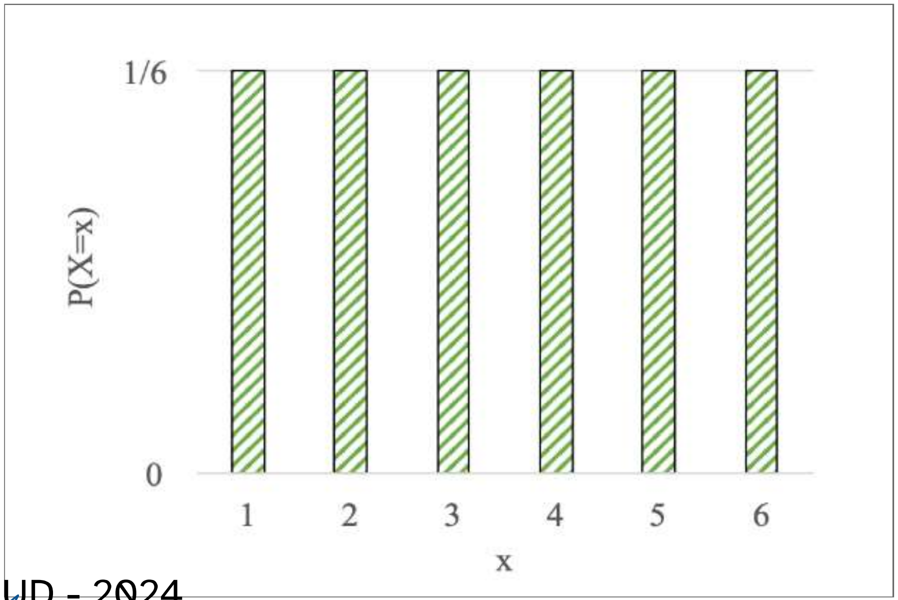
- Ký hiệu:có thể dùng ký hiệu $\mathbb{P}(y)$ thay cho $\mathbb{P}(Y = y)$
- Có thể biểu diễn PMF bằng:: dạng đồ thị, dạng bảng, dạng biểu thức

Ví dụ

Tung một con xúc sắc cân đối. Biến ngẫu nhiên X: số nút thu được

x	P(X = x)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

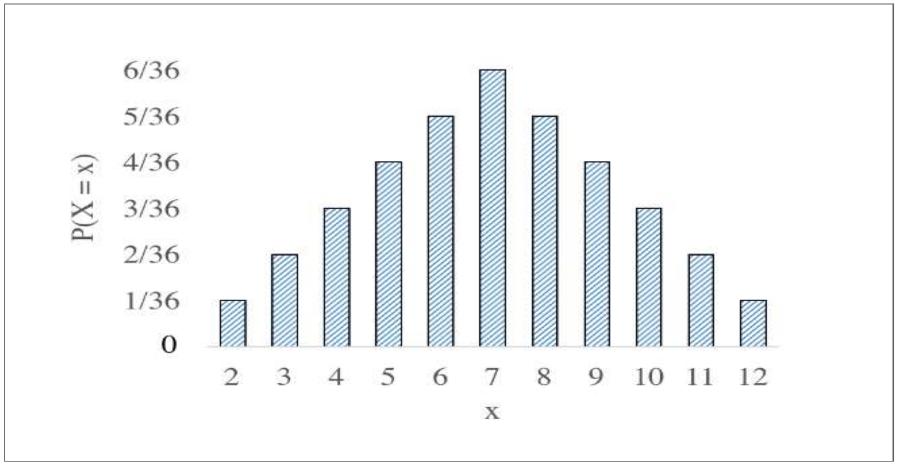
$$\mathbb{P}(x) = \frac{1}{6}, \quad \text{nếu} \quad 1 \leq x \leq 6$$



Ví dụ:

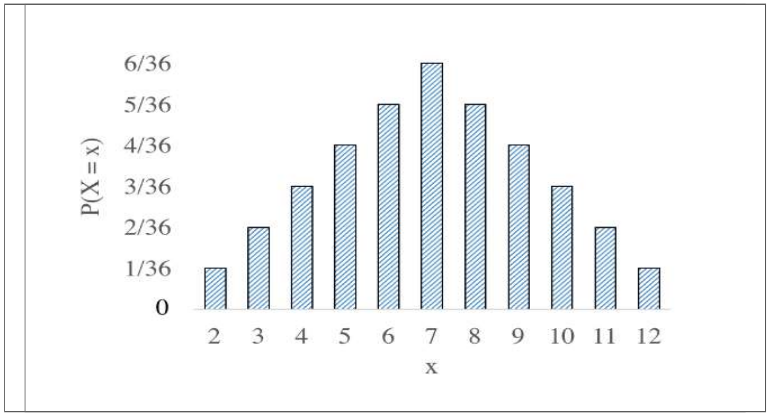
Tung hai con xúc sắc cân đối. Biến ngẫu nhiên X: tổng số nút thu được

$$\mathbb{P}(y) = \begin{cases} \frac{y-1}{36}, & \text{nếu} \quad 1 \leq y \leq 7 \\ \frac{13-y}{36}, & \text{nếu} \quad 8 \leq x \leq 12 \end{cases}$$



Tính chất

- Tổng các xác suất: gọi k là các giá trị mà X có thể nhận được $(k \in \Omega_X)$. Ta có: $\sum_k \mathbb{P}(X = k) = 1$



Hàm phân phối tích lũy (CDF)

- Hàm phân phối tích lũy (**cumulative distribution function - CDF**) của biến ngẫu nhiên X là hàm F(x) được định nghĩa

$$\begin{aligned} F_X &: \mathbb{R} \to [0,1] \\ & x \mapsto F_X(x) = \mathbb{P}(X \leq x) \end{aligned}$$

Ví dụ

Giả sử ta có một biến ngẫu nhiên rời rạc X có thể nhận các giá trị với xác suất tương ứng như sau:

x	1	2	3	4
$p_X(k)$	0.2	0.4	0.3	0.1

CDF của X như sau:

$$F(x) = \begin{cases} 0 & \text{khi } x < 1, \\ 0.2 & \text{khi } 1 \leq x < 2, \\ 0.6 & \text{khi } 2 \leq x < 3, \\ 0.9 & \text{khi } 3 \leq x < 4, \\ 1 & \text{khi } x \geq 4. \end{cases}$$

Kỳ vọng

- **Kỳ vọng (expected value hay expectation)** trong thống kê có thể được hiểu là giá trị trung bình của tất cả các giá trị mà biến ngẫu nhiên có thể nhận được. Kỳ vọng của biến ngẫu nhiên X, được ký hiệu là $E[X]$ hoặc μ_X

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

Tính chất

- **Kỳ vọng của một hằng số**: với c là một hằng số ta có
  $E[c] = c$
- **Tính tuyến tính của kỳ vọng**: giả sử có hai biến ngẫu nhiên thuộc cùng một không gian mẫu ($X, Y \in \Omega$) và các hằng số $a, b, c \in \mathbb{R}$, thì
  $E[X + Y] = E[X] + E[Y]$
  $E[aX] = aE[X]$
  $E[\sum_i^n X_i] = \sum_i^n E[X_i]$

TKUD - 2024

Tính chất

- **Luật của nhà thống kê vô thức** (The Law Of Unconcious of Statiscian - **LOTUS**): cho biến ngẫu nhiên $X \in \Omega_X$ và hàm số $g : D \to \mathbb{R}$, với $\Omega_X \subseteq D$
$$E[g(X)] = \sum_x g(x)\mathbb{P}(X = x)$$
- *Lưu ý:* $E[g(X)] \neq g(E[X])$.
- Ví dụ: $E[X^2] \neq (E[X])^2$ hay $E[log(X)] \neq log(E[X])$

Phương sai

- **Phương sai (variance)** của một biến ngẫu nhiên đo độ trải (độ phân tán) của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình (mean).

  $$Var(X) = E[(X - E[X])^2] = E[(X - \mu_X)^2]$$

- Có thể dùng công thức tương đương nhưng đơn giản hơn để tính phương sai dưới đây:

  $$Var(X) = E[X^2] - (E[X])^2 = E[X^2] - \mu_X^2]$$

Độ lệch chuẩn

- ***Độ lệch chuẩn (standard deviation)*** của biến ngẫu nhiên X được tính:

$$\sigma_X = std(X) = \sqrt{Var(X)}$$

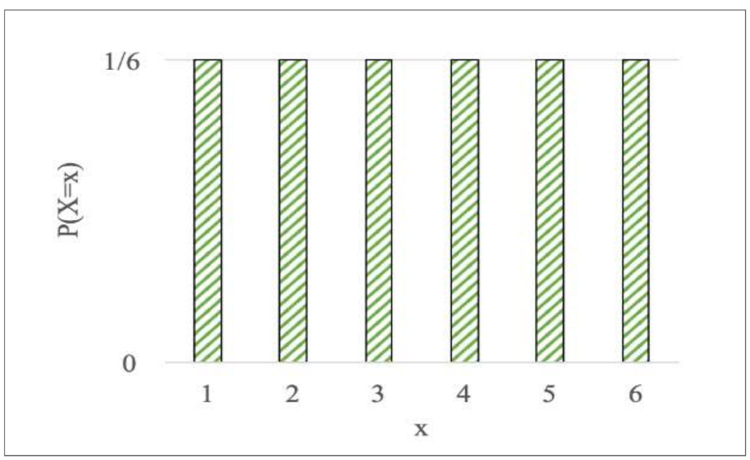
- Độ lệch chuẩn cũng dùng để đo độ trải, độ phân tán của biến ngẫu nhiên, nhưng độ lệch chuẩn có cùng đơn vị với biến ngẫu nhiên

Phân phối xác suất

- Phân phối xác suất (PPXS) của biến ngẫu nhiên có thể hiểu đơn giản là các giá trị mà biến ngẫu nhiên có thể nhận được và xác suất tương ứng với các giá trị đó.
- Ví dụ: Gọi X là số nút thu được khi tung xúc sắc

x	P(X = x)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

$$\mathbb{P}(x) = \frac{1}{6}, \quad \text{nếu} \quad 1 \leq x \leq 6$$



TKUD - 2024

Phân phối Bernoulli (Bernoulli Distribution)

- Mô tả: Biến ngẫu nhiên X được gọi là có phân phối Bernoulli nếu nó đại diện cho khả năng xảy ra của một sự kiện với xác suất p trong một phép thử. Khi đó, X chỉ có thể nhận hai giá trị: 1 nếu sự kiện xảy ra, và 0 nếu ngược lại.
- Ví dụ:
  - sự kiện mặt sấp xuất hiện khi tung đồng xu
  - sự kiện chọn ngẫu nhiên một người có chơi game Wukong
  - sự kiện tuần sau có bài kiểm tra...

Bernoulli Distribution

Notation	$X \sim \mathcal{B}er(p)$
Parameters	p
Support	$\{0, 1\}$
PMF	$\mathbb{P}(X = x) = \begin{cases} p & \text{, nếu } x = 1 \\ 1 - p & \text{, nếu } x = 0 \end{cases}$
PMF (smooth)	$\mathbb{P}(X = x) = p^x(1 - p)^{1-x}$
Expectation	$E[X] = p$
Variance	$Var[X] = p(1 - p)$

TKUD - 2024

Phân phối Nhị thức (Binomial Distribution)

- Mô tả: một phép thử được thực hiện n lần độc lập nhau, với mỗi lần xác suất để một sự kiện E xảy ra là p. Biến ngẫu nhiên X đại diện cho số lần sự kiện E xảy ra. Khi đó, ta nói biến ngẫu nhiên X có phân phối nhị thức.
- Ví dụ:
  - số lần thu được mặt sấp khi tung đồng xu n lần
  - số số 1 trong một dãy n bits được phát sinh ngẫu nhiên
  - số lần sút phạt đền thành công của đội tuyển Argentina trong một lượt 5 cú sút?

Phân phối Nhị thức

Notation	$X \sim \mathcal{B}in(n, p)$
Parameters	$n \in \{0, 1, \ldots\}$, số lần thử
	$p \in [0..1]$, xác suất sự kiện xảy ra trong một lần thử
Support	$\{1, 2, \ldots, n\}$
PMF	$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{1-x}$
Expectation	$E[X] = np$
Variance	$Var[X] = np(1-p)$

Phân phối Nhị thức

- Điều kiện:
  - Các lần thực hiện phép thử phải giống nhau
  - Xác suất sự kiện xảy ra ở mỗi phép thử đều như nhau p
  - Kết quả ở mỗi lần thực hiện phép thử không bị ảnh hưởng bởi những lần khác

Ví dụ

Cho biến ngẫu nhiên $X \sim \mathcal{B}in(5, 0.6)$. Tính $\mathbb{P}(X = 2) = ?$

$$\mathbb{P}(X = 2) = \binom{5}{2} 0.6^2 0.4^3 = 0.2304$$

```python
from scipy import stats

n, p = 5, 0.6
x = 2

p_x = stats.binom.pmf(x, n, p)
print(f'P(X = {x}) = {p_x}')
```

```python
from scipy import stats

n, p = 5, 0.6
x = 2

# define a random variable
rv = stats.binom(n, p)

p_x = rv.pmf(x)
print(f'P(X = {x}) = {p_x}')
```

Phân phối Poisson (Poisson Distribution)

- Mô tả: Biến ngẫu nhiên có phân phối Poisson cho biết số lần xảy ra sự kiện xảy ra trong một khoảng thời gian (hoặc không gian) cố định. Biến ngẫu nhiên có phân phối Poisson dựa trên giả định rằng số lần xảy ra sự kiện trung bình trong một khoảng thời gian (không gian) là độc lập nhau và không đổi.
- VD:

Poisson Distribution

Notation	$X \sim \mathcal{P}oi(\lambda)$
Parameters	$\lambda \in \{0, 1, \ldots\}$
Support	$\{0, 1, \ldots\}$
PMF	$\mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$
Expectation	$E[X] = \lambda$
Variance	$Var[X] = \lambda$

VD

- Một quán cà phê nhận được trung bình 5 khách hàng mỗi giờ. Số lượng khách đến trong một giờ được mô hình theo phân phối Poisson. Tính XS để trong một giờ có đúng 3 khách hàng đến quán?

$$\mathbb{P}(X=3)=\frac{e^{-\lambda}\lambda^{x}}{x!}=\frac{e^{-5}5^{3}}{3!}=0.0842$$

```
from scipy import stats

ld = 5 # lambda
x = 2

p_x = stats.poisson.pmf(x, ld)
print(f'P(X = {x}) = {p_x}')
```

```
from scipy import stats

ld = 5 # lambda
x = 2

rv = stats.poisson(ld) # define
p_x = rv.pmf(x)
print(f'P(X = {x}) = {p_x}')
```

TKUD - 2024

TÓM TẮT

Qua bài giảng này chúng ta đã biết được:

- Khái niệm về biến ngẫu nhiên, và hai loại biến ngẫu nhiên
- Một số đặc trưng thường dùng trong thống kê của biến ngẫu nhiên như: trung bình (**mean**), phương sai (**variance**), độ lệch chuẩn (**standard deviation**)
- Hàm độ lớn xác suất (**PMF)** và một số PPXS tiêu biểu của biến ngẫu nhiên rời rạc