BIẾN NGẪU NHIÊN VÀ PHÂN BỐ XÁC SUẤT

Nguyễn Văn Hạnh

AI Academy Vietnam

Tháng 7 năm 2021

Nội dung

1. Khái niệm Biến ngẫu nhiên
2. Phân bố của biến ngẫu nhiên: Hàm phân bố, hàm mật độ xác suất
3. Các số đặc trưng của biến ngẫu nhiên
4. Phương pháp nghịch đảo mô phỏng biến ngẫu nhiên rời rạc và Thực hành trên Python
5. Bài tập thực hành

Khái niệm về biến ngẫu nhiên

- Định nghĩa: Biến ngẫu nhiên là một kết quả số của một phép thử [1].
- Ví dụ: Khi gieo hai con xúc sắc, gọi X, Y lần lượt là số chấm xuất hiện trên mặt của con thứ nhất và thứ hai thì X, Y là hai biến ngẫu nhiên vì nó là kết quả có kiểu số. Do đó thì hàm $X + Y, 2XY, sin(XY)$ cũng là các biến ngẫu nhiên [1].
- Biến ngẫu nhiên được ký hiệu bởi chữ cái in hoa X, Y và các giá trị có thể nhận của nó ký hiệu bởi chữ cái in thường x, y. Sự kiện được viết dưới dạng toán học $X = x, X < x, X > x, X \leq X, x_1 \leq X < x_2$.
- Các giá trị có thể nhận của biến ngẫu nhiên là đếm được thì biến ngẫu nhiên được gọi là rời rạc. Các giá trị mà biến nhiên vô hạn không đếm được gọi là biến ngẫu nhiên liên tục [1].

Hàm trọng số- Probability mass function- PMF

- Biến ngẫu nhiên rời rạc X có miền giá trị có thể nhận $(x_1, \cdots, x_n)$
- Hàm trọng số của biến ngẫu nhiên rời rạc ký hiệu là

$$p_X(x) = P(X = x), \forall x \in \mathbb{R}$$

- Ý nghĩa: Hàm trọng số thể hiện khả năng xảy ra tại một điểm x.
- Bảng phân phối xác suất

$X = x$	x_1	$\cdots$	x_n
$p_x(x)$	$p_X(x_1)$	$\cdots$	$p_X(x_n)$

- Các tính chất: $p_X(x) \geq 0, \forall x \in \mathbb{R}; \sum_{i=1}^{n} p_X(x_i) = 1.$

Hàm phân phối xác suất (Cumulative distribution function -CDF)

- Hàm phân phối xác suất của biến ngẫu nhiên X là hàm được xác định bởi công thức
$$F_x(x) = P(X \leq x), \forall x \in \mathbb{R}$$
- Ý nghĩa: Hàm phân phối xác suất là xác suất của sự kiện bnn X nhận giá trị từ $-\infty$ tới x. Khi có hàm phân phối ta thực hiện với hàm giải tích thay vì làm với các phép toán với sự kiện.
- Các tính chất
  - $F_X(-\infty) = 0; F_X(+\infty) = 1$
  - $P(X \leq a) = F_X(a); P(X > a) = 1 - F_X(a);$
  - $P(a < X \leq b) = F_X(b) - F_X(a).$
- X là biến ngẫu nhiên rời rạc thì $F_X(x) = \sum_{x_i < x} P_X(x_i)$

Ví dụ 1: Gieo một con xúc sắc [1 tr 46]

- Gieo một con xúc sắc. X là số chấm xuất hiện
- Các giá trị X có thể nhận $S = \{1, 2, 3, 4, 5, 6\}$
- Hàm trọng số $P_X(x) = \begin{cases} 1/6; x \in S \\ 0; x \notin S \end{cases}$
- Hàm phân phối xác suất $F_X(x) = \begin{cases} 0; x < 1 \\ 1/6; 1 \leq x < 2 \\ 2/6; 2 \leq x < 3 \\ 3/6; 3 \leq x < 4 \\ 3/6; 3 \leq x < 4 \\ 4/6; 4 \leq x < 5 \\ 5/6; 5 \leq x < 6 \\ 1; x \geq 60; x \notin S \end{cases}$

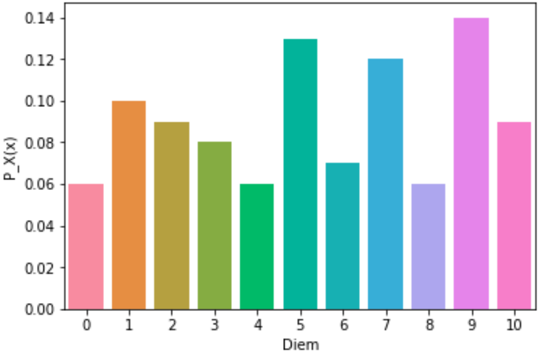
Hàm trọng số

```
import numpy as np
# Tạo ngẫu nhiên ra 100 con điểm nhận giá trị từ 0 đến 10
diem = np.random.randint(0, 11, 100)
bien_ngau_nhien, tan_so = np.unique(diem, return_counts=True)
pmf = tan_so / len(gia_tri)
#Bảng phân phối xác suất
np.column_stack((bien_ngau_nhien, pmf))
```

```
array([[ 0.  ,  0.11],
       [ 1.  ,  0.1 ],
       [ 2.  ,  0.07],
       [ 3.  ,  0.06],
       [ 4.  ,  0.11],
       [ 5.  ,  0.1 ],
       [ 6.  ,  0.1 ],
       [ 7.  ,  0.13],
       [ 8.  ,  0.07],
       [ 9.  ,  0.1 ],
       [10.  ,  0.05]])
```

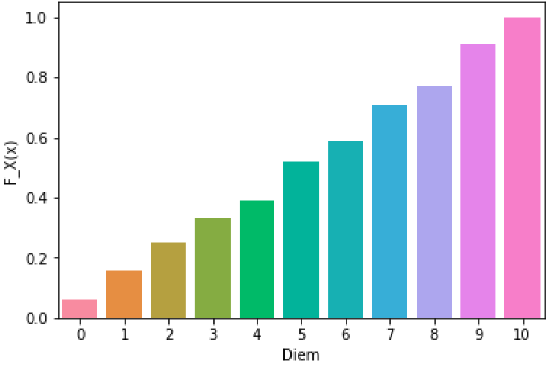
Đồ thị hàm trọng số

```
import seaborn as sns
PMF=sns.barplot(bien_ngau_nhien, pmf)
PMF.set(xlabel='Diem', ylabel='P_X(x)')
plt.show()
```



Hàm phân phối xác suất

```
CDF=sns.barplot(bien_ngau_nhien, cdf)
CDF.set(xlabel='Diem', ylabel='F_X(x)')
plt.show()
```



Tính xác suất

```
# Xác suất >=4
1-cdf[4]
```

```
0.61
```

```
# Xác suất điểm giỏi
1-cdf[8]
```

```
0.22999999999999998
```

```
# Xác suất điểm trung bình
cdf[7]-cdf[5]
```

```
0.19000000000000006
```

Hàm mật độ (Density probability function- PDF)

- X là biến ngẫu nhiên liên tục thì $P(X = x) = 0 \forall x \in \mathbb{R}$.
- Ta cần hàm số đo đo khả năng xảy ra tại lân cận một điểm $P(x \leq X < x + \Delta x) = F_X(x + \Delta x) - F_X(x)$
- Ý nghĩa: Hàm mật độ tương tự như hàm trọng số
- Định nghĩa:

Hàm mật độ-PDF

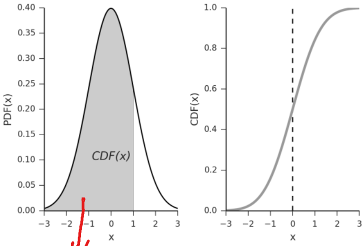
Hàm mật độ của biên ngẫu nhiên X ký hiệu là

$$f_X(x) = lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} = F_X^{'}(x)$$

VINBIGDATA VINGROUP vds AI Academy Vietnam

Tính chất hàm mật độ

- $f_X(x) \geq 0; \forall x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f_X(x)dx = 1$
- $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$
- $P(X < a) = \int_{-\infty}^{a} f_X(t)dt$
- $P(a \leq X < b) = F_X(b) - F_X(a)$
- $P(a \leq X < b) = \int_{a}^{b} f_X(t)dt$

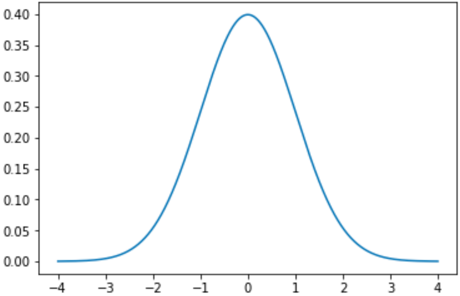


$= \int_{-\infty}^{1} (\ ) dx$

Hàm mật độ

```python
from scipy.stats import norm
import numpy as np
import matplotlib.pyplot as plt

x= np.arange(-4,4,0.001)
plt.plot(x, norm.pdf(x))
plt.show()
```



Tính chất hàm mật độ

```
# Xác suất tại lân cận điểm 1.5
norm.pdf(1.5)
```

```
0.12951759566589174
```

```
# Xác suất từ 1 đến 1.5 F_X(1.5)-F_X(1)
norm.cdf(1.5)-norm.cdf(1)
```

```
0.09184805266259899
```

```
#Tính xác suất theo hàm mật độ
import scipy.integrate as integrate

result = integrate.quad(lambda x: norm.pdf(x), 1, 1.5)
result[0]
```

```
0.09184805266259899
```

```
# Tính xác suất từ 0 đến vô cùng
result1 = integrate.quad(lambda x: norm.pdf(x), 0, np.inf)
```

```
result1[0]
```

```
0.4999999999999999
```

Biến ngẫu nhiên độc lập

- Định nghĩa [1 tr 47]: Hai biến ngẫu nhiên X, Y được gọi là độc lập nếu bất cứ tập giá trị nào của X, Y là I, J ta có

$$P(X \in I \cap Y \in J) = P(X \in I).P(Y \in J)$$

- Ví dụ X, Y là doanh thu của công ty A, B độc lập với nhau, xác suất để doanh thu trên 10 tỷ của A là 0.4, B là 0.3. Tìm xác suất để công ty A, B đều có doanh thu trên 10 tỷ.
- $P(X > 10 \cap Y > 10) = P(X > 10).P(Y > 10) = 0.4 \times 0.3 = 0.12$

Các số đặc trưng của biến ngẫu nhiên: Kỳ vọng

Định nghĩa kỳ vọng [1 tr 51]

Thực hiện một phép thử lặp đi lặp lại với biến ngẫu nhiên X. Giá trị kỳ vọng của X là giá trị trung bình của X khi chúng ta lặp lại một phép thử giống nhau nhiều lần.

- Khi thực hiện phép thử n lần giống nhau ta có $X_1, \cdots, X_n$ là các giá trị của X, khi đó kỳ vọng là

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n}$$

- Khi số lượng phép thử lớn ta có $\lim_{n\to\infty} \frac{\#(X=x_i)}{n} = P_X(x_i)$
- Kỳ vọng $\boxed{E(X) = \sum_{x_i} x_i.P_X(x_i)}$, X- rời rạc

Tính chất kỳ vọng

- c là hằng số thì $E(c) = c$
- X, Y là hai biến ngẫu nhiên thì $E(X + Y) = E(X) + E(Y)$
- c là hằng số thì $E(cX) = cE(X)$
- $c_1, \cdots, c_n$ là hằng số $X_1, \cdots, X_n$ là biến ngẫu nhiên thì $E(c_1X_1 + \cdots + c_nX_n) = E(c_1X_1) + \cdots + E(c_nX_n)$
- $Y = g(X)$, với X rời rạc thì $E(Y) = E[(g(X)] = \sum_x g(x).P_X(x)$
- X, Y là hai biến ngẫu nhiên độc lập thì $E(X.Y) = E(X).E(Y)$

Ví dụ: Dự báo nhu cầu sản phẩm tr 58 [1]

- Dự báo là một lĩnh vực trụ cột của khoa học dữ liệu. Khi chúng ta nghiên cứu mô hình dựa vào dữ liệu.
- Gọi $D_i, i = 1, 2, \cdots$ là số lượng mặt hàng bán được ngày thứ $1, 2, \cdots$.
- Giả sử rằng nếu một ngày bán được là 1, hoặc 2 mặt hàng thì ngày tiếp theo sẽ bán được 1, 2 và 3 mặt hàng với xác suất là $1/3$. Nhưng nếu nhu cầu cao (ngày bán được 3 mặt hàng) thì số lượng mặt hàng ngày tiếp theo bán được $1, 2, 3$ sẽ lần lượt là $0.2, 0.2, 0.6$ tương ứng.
- Gọi M là số lượng sản phẩm sẽ bán được của ngày mai, biết ngày hôm nay bán được 3 sản phẩm, tìm số lượng sản phẩm kỳ vọng của ngày mai. Tìm hàm trọng số của N là số lượng mặt hàng bán được cho 2 ngày tiếp theo?

Ví dụ: Dự báo nhu cầu sản phẩm tr 58 [1]

- M nhận giá trị trong tập $\{1, 2, 3\}$
- $E(M) = 1 \times 0.2 + 2 \times 0.2 + 3 \times 0.6 = 2.4$
- Xác suất tương ứng $P_N(1) = P(N = 1) = P(M = 1)P(N = 1|M = 1) + P(M = 2).P(N = 1|M = 2) + P(M = 3).P(N = 1|M = 3) = 0.2 \times \frac{1}{3} + 0.2 \times \frac{1}{3} + 0.6 \times 0.2 = 0.2533$
- $P_N(2) = P(N = 2) = 0.2533$
- $P_N(3) = P(N = 3) = 0.4934$
-

$N = x$	1	2	3
$p_N(x)$	0.2533	0.2533	0.4934

Các số đặc trưng của biến ngẫu nhiên: Phương sai

- Kỳ vọng $E(X) = \begin{cases} \sum_x x.P_X(x), \text{X- rời rạc} \\ \int_{-\infty}^{\infty} x f_X(x)dx, \text{X- liên tục.} \end{cases}$
- $Y = (X - EX)^2$
- Phương sai $Var(X) = E(Y) = E[(X - EX)^2] =$
  $\begin{cases} \sum_x (x - EX)^2.P_X(x), \text{X- rời rạc} \\ \int_{-\infty}^{\infty} (x - EX)^2 f_X(x)dx, \text{X- liên tục.} \end{cases}$
- Ý nghĩa của Phương sai là đo sự sai khác của biên ngẫu nhiên với kỳ vọng.

Tính chất phương sai

- c là hằng số $Var(c) = 0$
- $Var(X) = E(X^2) - [E(X)]^2$
- $Var(cX) = c^2 V(X)$
- $Var(X + c) = Var(X)$
- Bất đẳng thức Chebychev: X có kỳ vòng $EX = \mu$, phương sai $Var(X) = \sigma^2$ thì

$$P(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2}$$

Một số đặc trưng khác

- Độ lệch tiêu chuẩn (standard deviation) $std(X) = \sqrt{Var(X)}$
- Skewness đo tính đối xứng của phân phối

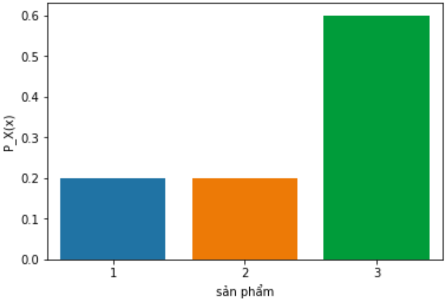
$$Skewness = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

- Mode là giá trị mà X đạt được xác suất lớn nhất
  $P_X(X = ModeX) \geq P_X(x), \forall x \in \mathbb{R}$
- Phân vị thứ $q \in [0, 1]$ ký hiệu x_q là giá trị mà X có thể nhận sao cho $P(X < x_q) = q$. Trung vị Med_X là phân vị $q = 0.5$
  $P(X < Med_X) = 0.5$

quantitle

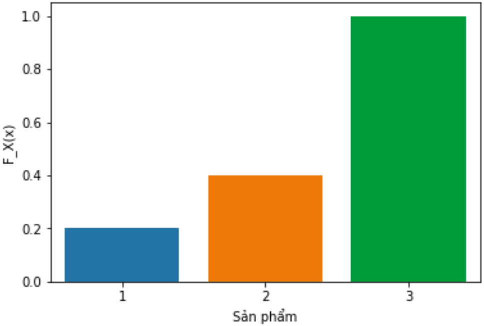
Phương pháp nghịch đảo mô phỏng biến ngẫu nhiên rời rạc: Biến ngẫu nhiên sản phẩm PMF

```
#pmf của một biến ngẫu nhiên số lượng sản phẩm
import seaborn as sns
sanpham=np.arange(1, 4)
pmf=[0.2, 0.2, 0.6]
PMF=sns.barplot(sanpham,pmf )
PMF.set(xlabel='sản phẩm', ylabel='P_X(x)')
plt.show()
```



Biến ngẫu nhiên sản phẩm - CDF

```
#CDF của biến ngẫu nhiên số lượng sản phẩm
cdf=np.cumsum(pmf)
CDF=sns.barplot(sanpham, cdf)
CDF.set(xlabel='Sản phẩm', ylabel='F_X(x)')
plt.show()
```



Phương pháp nghịch đảo mô phỏng biến ngẫu nhiên sản phẩm

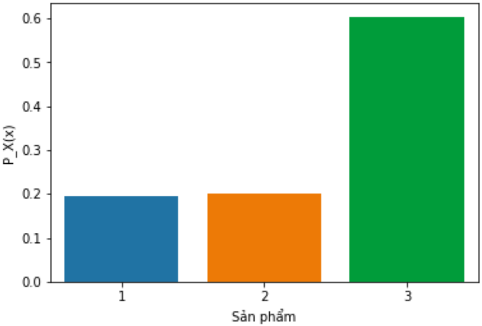
```
#Mô phỏng lại một biến ngẫu nhiên số lượng sản phẩm
from random import choices

solan=10000
X=choices(sanpham,pmf,k=solan)
bien_ngau_nhien, tan_so = np.unique(X, return_counts=
pmf = tan_so / len(X)
np.column_stack((bien_ngau_nhien, pmf))
```

```
array([[1.      , 0.1958],
       [2.      , 0.2009],
       [3.      , 0.6033]])
```

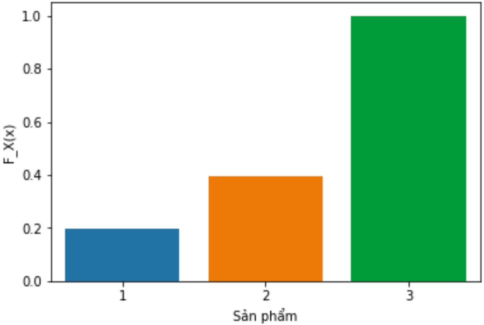
Phương pháp nghịch đảo mô phỏng biến ngẫu nhiên sản phẩm-PMF

```python
#Hàm trọng số thực nghiệm của số lượng sản phẩm
import seaborn as sns
PMF=sns.barplot(bien_ngau_nhien, pmf)
PMF.set(xlabel='Sản phẩm', ylabel='P_X(x)')
plt.show()
```



Phương pháp nghịch đảo mô phỏng biến ngẫu nhiên sản phẩm-CDF

```
#Hàm phân phối thực nghiệm của biến số lượng sản phẩm
cdf=np.cumsum(pmf)
CDF=sns.barplot(sanpham, cdf)
CDF.set(xlabel='Sản phẩm', ylabel='F_X(x)')
plt.show()
```



Phương pháp nghịch đảo mô phỏng biến ngẫu nhiên sản phẩm-CDF

```
print('Kỳ vọng:',np.mean(X))
print('Phương sai:', np.var(X))
print('Độ lệch tiêu chuẩn:', np.std(X))
print('Giá trị nhỏ nhất:', np.min(X))
print('Giá trị lớn nhất:', np.max(X))
print('Phân vị 0.25:',np.quantile(X,0.25) )
```

```
Kỳ vọng: 2.4
Phương sai: 0.6391999999999999
Độ lệch tiêu chuẩn: 0.7994998436522673
Giá trị nhỏ nhất: 1
Giá trị lớn nhất: 3
Phân vị 0.25: 2.0
```

Bài tập thực hành 1

- Cho biến ngẫu nhiên X nhận các giá trị $[10, 20, 100, 1000]$ với pmf cho trước $[0.4, 0.3, 0.2, 0.1]$
- Vẽ đồ thị PMF, vẽ đồ thị hàm CDF
- Tìm xác suất để biến ngẫu nhiên nhận giá trị nhỏ hơn 50
- Tìm giá trị trung bình, phương sai, độ lệch tiêu chuẩn, trung vị, skewness

Bài tập thực hành 2

- Lấy lại dữ liệu về hoa diên vỹ ở buổi 1, có y là vec tơ nhận các giá trị loài hoa
- Tìm bảng trọng số của y, y_train, y_test
- Vẽ đồ thị hàm trọng số PMF của của y, y_train, y_test
- Vẽ đồ thị hàm phân phối xác suất CDF của y, y_train, y_test

Tài liệu tham khảo

1. Norman Matloff, Probability and Statistics for Data Sciences, Taylor & Francis Group (2020)