MỘT SỐ PHÂN BỐ XÁC SUẤT QUAN TRỌNG

Nguyễn Văn Hạnh

AI Academy Vietnam

Tháng 7 năm 2021

Nội dung

1. Phân bố rời rạc: phân bố nhị thức, phân bố Poisson
2. Phân bố liên tục: phân bố chuẩn chuẩn, phân bố Student, phân bố Fisher
3. Phương pháp nghịch đảo mô phỏng biến ngẫu nhiên liên tục và Thực hành trên Python

Phân phối nhị thức: Định nghĩa

- Xét một phép thử có sự kiện A xảy ra với xác suất p. Gọi X là biến ngẫu nhiên nhận giá trị 1 nếu A xảy ra và 0 nếu A không xảy ra.
- Hàm trọng số của X là:

$$P_X(x) = \begin{cases} p; x = 1 \\ 1 - p; x = 0 \\ 0; \text{nếu trái lại} \end{cases}$$

- $E(X) = p; Var(X) = p(1 - p)$
- Định nghĩa: Tiến hành phép thử trên n lần độc lập với nhau. Gọi $X_1, \cdots, X_n$ là biến ngẫu nhiên (định nghĩa như X) của phép thử thứ $1, \cdots, n.$; Gọi $Y = X_1 + \cdots + X_n$ thì Y được gọi là biến ngẫu nhiên nhị thức ký hiệu $\mathcal{B}(n, p)$.. Y là số lần xảy ra sự kiện A trong dãy n phép thử.

Hàm trọng số và các giá trị đặc trưng của biến ngẫu nhiên tuân theo phân phối Nhị thức

- Y có thể nhận giá trị $\{0, \cdots, n\}$
- Hàm trọng số
  $P_Y(y) = P(Y = y) = C_n^y p^y (1-p)^{n-y}, \forall y \in \{0, \cdots, n\}.$
- $E(Y) = E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = np$
- $Var(Y) = Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n) = np(1-p)$

Ví dụ

- Một đề thi trắc nghiệm có 30 câu. Một người đi thi không biết gì chọn ngẫu nhiên 1 trong 4 phương án cho mỗi câu. Tìm xác suất người này được diểm qua nghĩa là trả lời được ít nhất 12 câu đúng. Tìm kỳ vọng và độ lệch tiêu chuẩn cho số câu trả lời đúng.
- A là sự kiện chọn được phương án đúng cho mỗi câu, $P(A) = p = 0.25$.
- $n = 30$ Gọi X là số câu trả lời đúng của người đi thi thì X có phân phối nhị thức $\mathcal{B}(30, 0.25)$.
- $P(X \geq 12) = 1 - P(X < 12) = 1 - \sum_{x=0}^{11} C_{30}^{x} 0.25^{x}.75^{30-x} = 0.0506$
- $E(X) = 30 \times 0.25 = 7.5$
- $std(X) = \sqrt{30 \times 0.25 \times 0.75} = 2.3717$

Tính phân phối Nhị thức

```
from scipy import stats
X= stats.binom(30, 0.25) # Khai báo biến nhị thức
```

```
# Xác suất số lượng câu trên 12
1-X.cdf(11)
```

```
0.05065828005155159 6
```

```
X.mean()# Kỳ vọng
```

```
7.5
```

```
X.std()# Độ lệch tiêu chuẩn
```

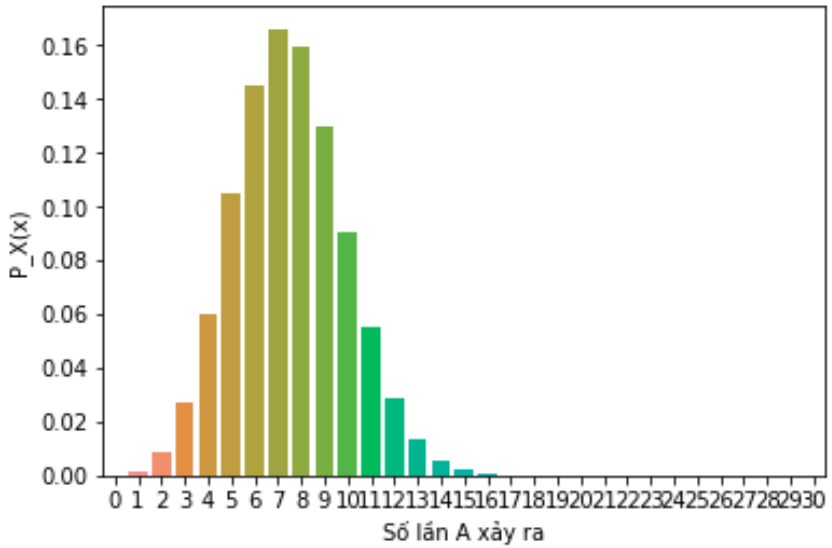
```
2.3717082451262845
```

PMF của biến ngẫu nhiên Nhị thức

```
from scipy import stats
X= stats.binom(30, 0.25) # Khai báo biến nhị thức
nhi_thuc=np.arange(0,31)
```

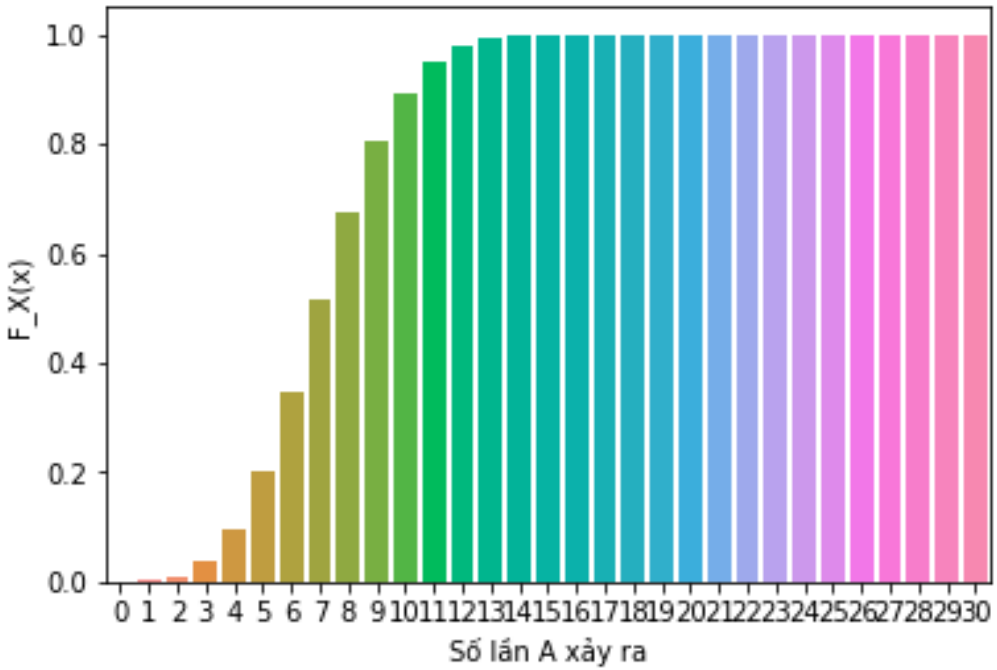
```
pmf=X.pmf(nhi_thuc)
```

```
import seaborn as sns
PMF=sns.barplot(nhi_thuc, pmf)
PMF.set(xlabel='Số lần A xảy ra', ylabel='P_X(x)')
plt.show()
```



CDF của biến ngẫu nhiên Nhị thức

```
cdf=X.cdf(nhi_thuc)
CDF=sns.barplot(nhi_thuc, cdf)
CDF.set(xlabel='Số lần A xảy ra', ylabel='F_X(x)')
plt.show()
```



Phân phối Poisson: Định nghĩa

- Gọi X là biến ngẫu nhiên tuân theo luật phân phối Poisson với tham số λ.
- Hàm trọng số của X là:

$$P_X(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!}, \forall x = 0, 1, \cdots 0; \text{nếu trái lại} \end{cases}$$

- $E(X) = \lambda; Var(X) = \lambda$

Ví dụ

- Số lượng khách hàng tới một cửa hàng giả sử tuân theo biến ngẫu nhiên có phân phối Poisson với tham số $\lambda = 3$ khách một giờ.
- Tìm xác suất trong một giờ có ít nhất 1 khách đến cửa hàng?
- Tìm xác suất trong 8 tiếng có 20 khách tới cửa hàng
- Tìm kỳ vọng, phương sai?

Giải:

- Gọi X là số lượng khách hàng tới một siêu thị trong 1 giờ
  $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - e^{-\lambda}\frac{\lambda^x}{x!} = 1 - e^{-3}\frac{3^0}{0!} =$
- Gọi Y là số lượng khách hàng tới của hàng trong 8 tiếng. Ta có
  $P(Y = 20) = e^{-24}24^{20}20!$; $Y = 8X$

$\Rightarrow E(Y) = E(8X) = 8 \times E(X) = 8 \times \lambda = 8 \times 3 = 24$

Tính toán ví dụ

```
X= stats.poisson(3)#Khai báo biến ngẫu nhiên X có phân phối Poisson có tham số lambda=3
```

```
X.mean()# Kỳ vọng số khách hàng tới cửa hàng
```

```
3.0
```

```
X.var()# Phương sai số khách hàng tới cửa hàng
```

```
3.0
```

```
1-X.pmf(0)# Xác suất có ít nhất một khách hàng tới cửa hàng
```

```
0.950212931632136
```

```
X.cdf(5)-X.cdf(1)# Có từ 2 đến 5 khách hàng tới siêu thị
```

```
0.49289197684185293
```

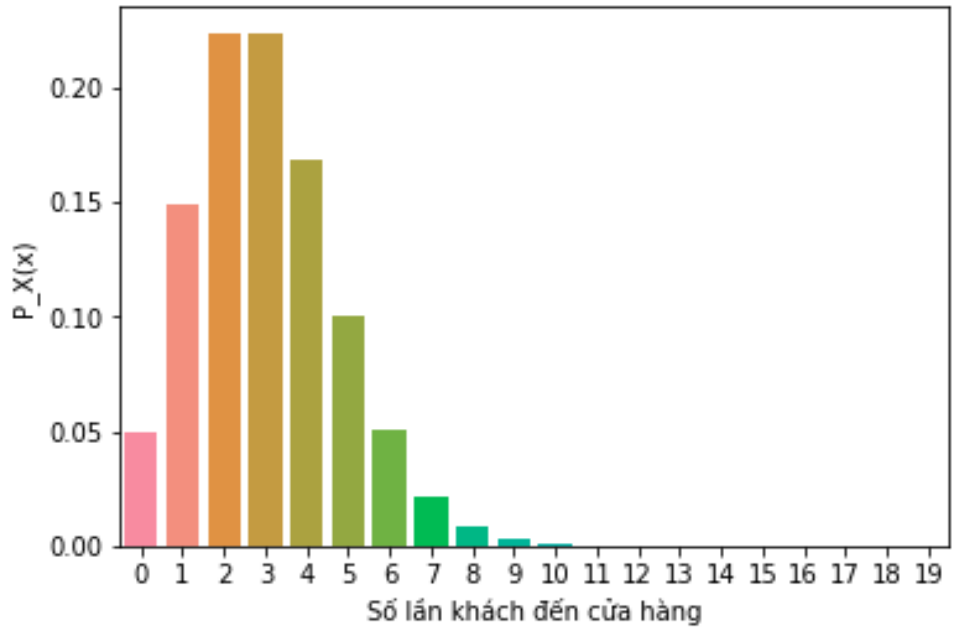
```
Y=stats.poisson(24)# Khai báo tham số Poisson có tham số gồm 24 khách hàng trong 8 tiếng
```

```
Y.pmf(20)# Xác suất có đúng 20 khách hàng tới siêu thị trong 8 tiếng
```

```
0.06237817316656375
```

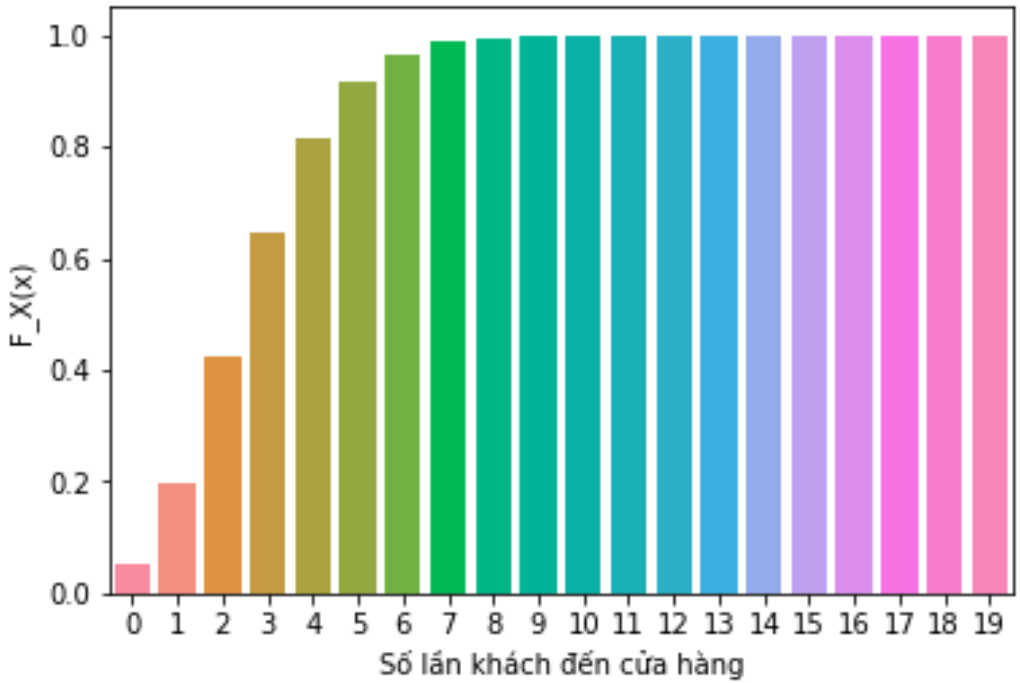
PMF của biến ngẫu nhiên tuân theo luật phân phối Poisson

```
# Vẽ đồ thị hàm trọng số của biến ngẫu nhiên X
import seaborn as sns
poisson=np.arange(0,20)
pmf=X.pmf(poisson)
PMF=sns.barplot(poisson, pmf)
PMF.set(xlabel='Số lần khách đến cửa hàng', ylabel='P_X(x)')
plt.show()
```



CDF của biến ngẫu nhiên thuân theo luật phân phối Poisson

```
#Vẽ đồ thị hàm phân phối xác suất của X
cdf=X.cdf(poisson)
CDF=sns.barplot(poisson, cdf)
CDF.set(xlabel='Số lần khách đến cửa hàng', ylabel='F_X(x)')
plt.show()
```



Định nghĩa phân phối chuẩn tắc

Phân phối chuẩn tắc

Định nghĩa: Biến ngẫu nhiên X được gọi là tuân theo luật phân phối chuẩn tắc, ký hiệu $X \sim Norm(0,1)$ nếu X có hàm mật độ:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \forall x \in \mathbb{R}$$

- $f_X(x)$ là hàm chẵn
- $f_X(x) \geq 0$,
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- $E(X) = \int_{-\infty}^{\infty} x f_X(x)dx = 0$; $Var(X) = E(X^2) - [E(X)]^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx = 1$

Định nghĩa phân phối chuẩn tắc

- Hàm phân phối xác suất $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ không có dưới dạng hàm hiện mà phải tính toán xấp xỉ.
- Vì biến ngẫu nhiên tuân theo luật phân phối chuẩn tắc thông dụng nên ta có ký hiệu đặc biệt là $\Phi(x) = F_X(x)$
- Do tính chất đối xứng
$$\Phi(-a) = \int_{-\infty}^{-a} f_X(t)dt = 1 - \int_{-\infty}^{a} f_X(t)dt = 1 - \Phi(a)$$
- Tính chất $P(X < a) = \Phi(a); P(X \geq a) = 1 - \Phi(a); P(a < X \leq b) = \Phi(b) - \Phi(a)$.

Quy tắc 1,2,3 sigma của phân phối chuẩn tắc

- Cho biến ngẫu nhiên $X \sim Norm(0,1)$.
- $P(-1 < X < 1) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 0.6827$
- $P(-2 < X < 2) = \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 = 0.9545$
- $P(-3 < X < 3) = \Phi(3) - \Phi(-3) = 2\Phi(3) - 1 = 0.9973$
- Phân vị mức $q = 0.9; 0.95; 0.99$

```
from scipy.stats import norm
2*norm.cdf(1)-1#Quy tắc 1 sigma
```

```
0.6826894921370859
```

```
2*norm.cdf(2)-1#Quy tắc 2 sigma
```

```
0.9544997361036416
```

```
2*norm.cdf(3)-1# Quy tắc 3 sigma
```

```
0.9973002039367398
```

```
norm.ppf(0.9)# phân vị mức 0.9
```

```
1.6448536269514722
```

```
norm.ppf(0.95)# phân vị mức 0.95
```

```
1.6448536269514722
```

```
norm.ppf(0.99)# phân vị mức 0.99
```

```
2.3263478740408408
```

PDF của biến ngẫu nhiên thuân theo luật phân phối chuẩn

- Biến ngẫu nhiên Y được gọi là tuân theo luật phân phối chuẩn với tham số $\mu, \sigma(\sigma \neq 0)$ ký hiệu $Y \sim Norm(\mu, \sigma)$ nếu $X = \frac{Y-\mu}{\sigma} \sim Norm(0, 1)$
- $Y = \sigma Y + \mu$ ta có $E(Y) = \mu$; $Var(Y) = Var(\sigma X + \mu) = \sigma^2$.
- $F_Y(x) = P(Y < x) = P(\sigma X + \mu < x) = P(X < \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$
- Tính chất
  - $P(Y < a) = \Phi(\frac{a-\mu}{\sigma})$;
  - $P(Y \geq a) = 1 - \Phi(\frac{a-\mu}{\sigma})$;
  - $P(a < Y \leq b) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$.

Ví dụ

- Giả sử chiều cao của nam thanh niên ở nước A tuân theo luật phân phối chuẩn với kỳ vọng 164cm với độ lệch tiêu chuẩn 8cm. Chọn ngẫu nhiên một nam sinh viên. Tìm xác suất sinh viên này:
- có chiều cao cao hơn 180cm
- Tìm chiều cao mà xác suất để chiều cao thanh niên nhỏ hơn con số này là 90%.

```
mu=164; sigma=8
1-norm.cdf((180-mu)/sigma)
```

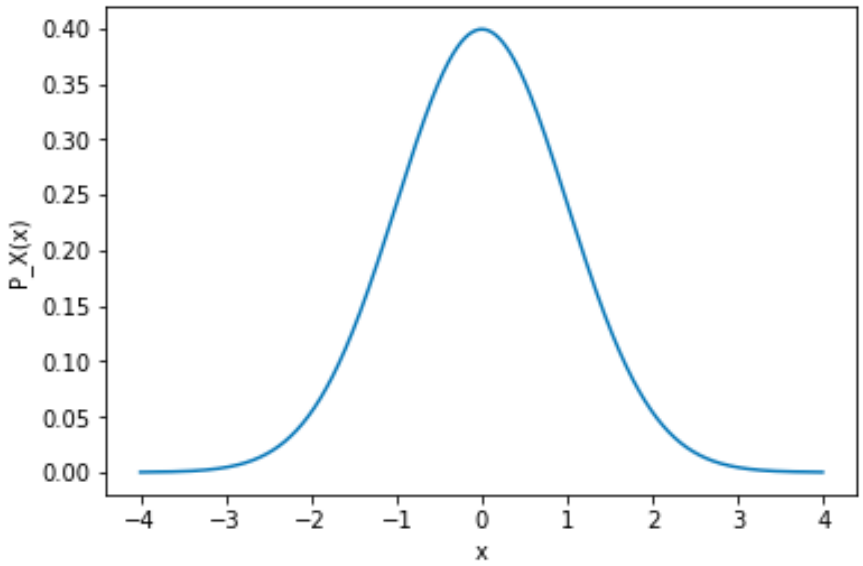
```
0.022750131948179207
```

```
norm.ppf(0.9, mu, sigma)
```

```
174.2524125243568
```

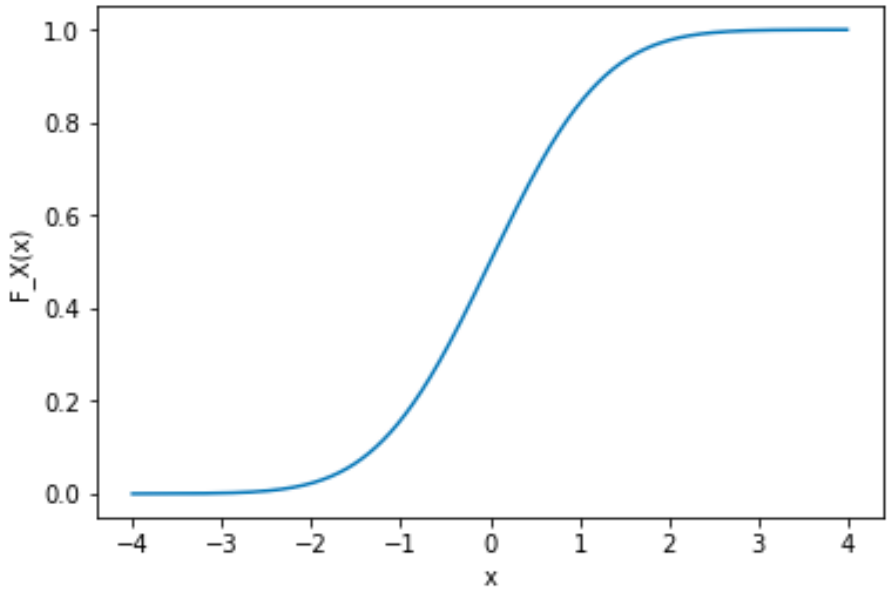
PDF của biến ngẫu nhiên thuân theo luật phân phối chuẩn tắc

```
# Vẽ đồ thị hàm trọng số của biến ngẫu nhiên X
import matplotlib.pyplot as plt
from scipy import stats
chuan=np.arange(-4,4,0.01)
#Khai báo biến ngẫu nhiên phân phối chuẩn tắc
pdf=X.pdf(chuan)
# Vẽ hàm mật độ chuẩn tắc
PDF=plt.plot(chuan, pdf)
plt.xlabel('x')
plt.ylabel('P_X(x)')
plt.show()
```



CDF của biến ngẫu nhiên thuân theo luật phân phối chuẩn tắc

```
import matplotlib.pyplot as plt
chuan=np.arange(-4,4,0.01)
#Hàm phân phối xác suất
cdf=X.cdf(chuan)
# Vẽ đồ thị hàm phân phối xác suất
CDF=plt.plot(chuan, cdf)
plt.xlabel('x')
plt.ylabel('F_X(x)')
plt.show()
```

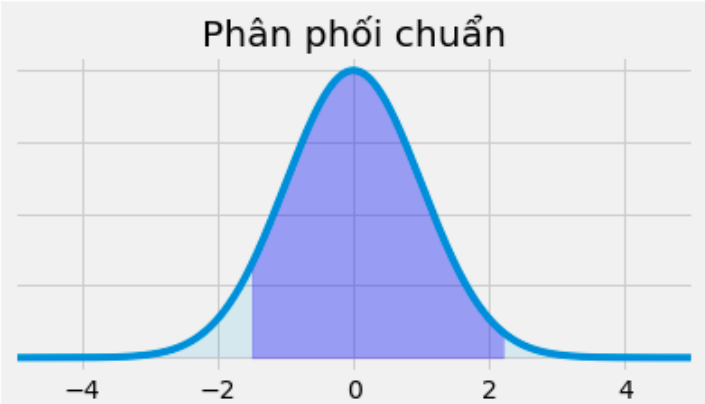


Xác suất trong khoảng a, b

```
#Vẽ xác suất từ điểm a đến điểm b
a=-1.5; b=2.2
x = np.arange(a, b, 0.001) # Miền giá trị của x nằm trong a,b
x_all = np.arange(-5, 5, 0.001) # Toàn bộ miền giá trị của X
y = norm.pdf(x,0,1) #Véc tơ giá trị pdf nằm trong a, b
y2 = norm.pdf(x_all,0,1)# Véc tơ giá trị tất cả các điểm
```

```
# Vẽ đồ thị xác suất tuân theo luật phân phối chuẩn
fig, ax = plt.subplots(figsize=(6,3))
plt.style.use('fivethirtyeight')
ax.plot(x_all,y2)

ax.fill_between(x,y,0, alpha=0.3, color='b')
ax.fill_between(x_all,y2,0, alpha=0.1)
ax.set_xlim([-5,5])
ax.set_xlabel('x')
ax.set_yticklabels([])
ax.set_title('Phân phối chuẩn')
plt.show()
```



Ý nghĩa phân phối chuẩn

- Phân phối chuẩn được Gauss phát minh năm 1809 nên cũng có khi nó được mang tên là phân phối Gauss.
- Bnn tuân theo phân phối chuẩn nhận giá trị trên cả trục số, tuy nhiên có thể xấp xỉ một số biến ngẫu nhiên không nhận tất cả các giá trị trên $\mathbb{R}$ theo phân phối chuẩn, đó là do qui tắc $3-\sigma$, tức là nếu ta có xác suất X rơi vào miền có xác suất bằng 0,9974 rất gần 1, nên hầu hết người ta chỉ cần quan tâm đến các giá trị trong lân cận $3-\sigma$ của kỳ vọng.
- Phân phối chuẩn chiếm vị trí quan trọng trong lý thuyết xác suất, là vị trí trung tâm trong các kết luận thống kê sau này.
- Trong thực tế, ví dụ trong lĩnh vực kinh tế, khoa học xã hội, ... nhiều phân phối không giống phân phối chuẩn, nhưng phân phối của trung bình cộng đối với mỗi trường hợp lại có thể xem là phân phối chuẩn miễn là cỡ mẫu n đủ lớn.

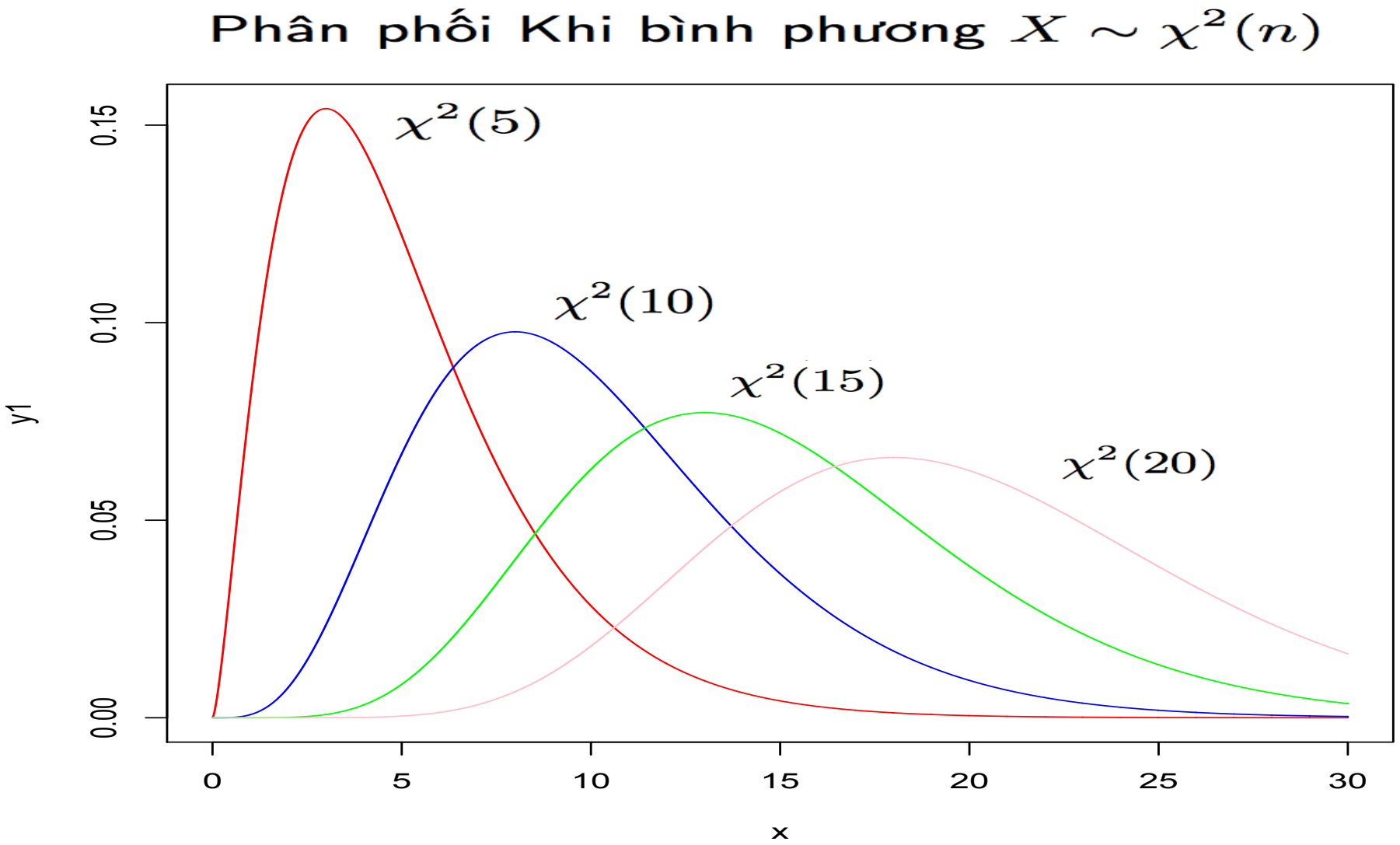
Phân phối Khi bình phương

Definition

Giả sử $X_i, (i = 1, 2, \ldots, n)$ là các biến ngẫu nhiên độc lập cùng phân phối chuẩn tắc. Biến ngẫu nhiên $Y = \sum_{i=1}^{n} X_i^2$ được gọi là tuân theo phân phối Khi bình phương với n bậc tự do.
Ký hiệu: $Y \sim \chi^2(n)$

Các tham số đặc trưng

- $EY = n$
- $VY = 2n$


Phân phối Khi bình phương $X \sim \chi^2(n)$
$\chi^2(5)$
$\chi^2(10)$
$\chi^2(15)$
$\chi^2(20)$
y1
0.00
0.05
0.10
0.15
0
5
10
15
20
25
30
x

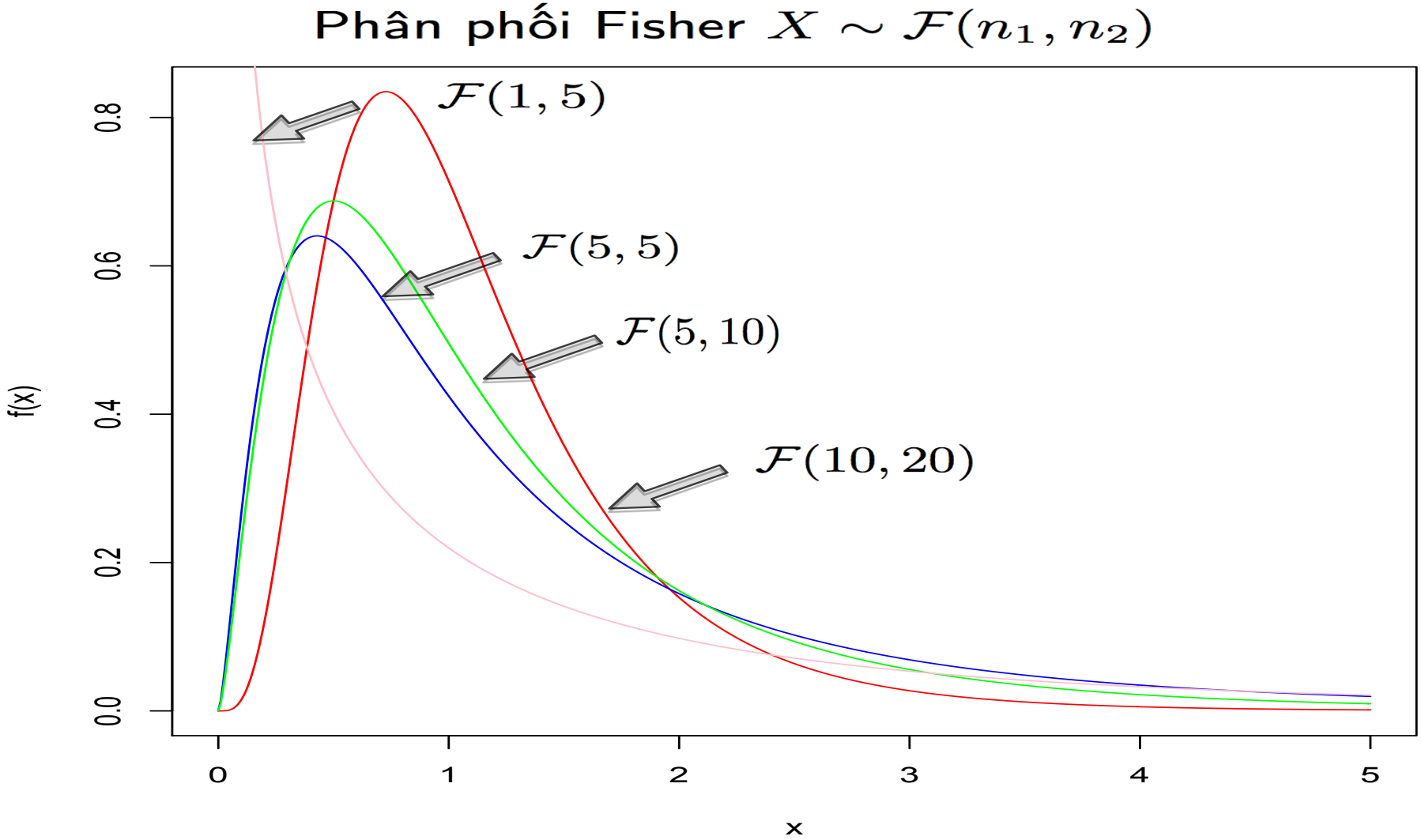
Phân phối Fisher

Definition

Giả sử hai biến ngẫu nhiên $\chi^2_{n_1} \sim \chi^2(n_1), \chi^2_{n_2} \sim \chi^2(n_2)$. Khi đó biến ngẫu nhiên F xác định bởi công thức sau:

$$F = \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$$

được gọi là tuân theo phân phối Fisher với bậc tự do n_1, n_2.
Ký hiệu: $F \sim \mathcal{F}(n_1, n_2)$


Phân phối Fisher $X \sim \mathcal{F}(n_1, n_2)$
$\mathcal{F}(1, 5)$
$\mathcal{F}(5, 5)$
$\mathcal{F}(5, 10)$
$\mathcal{F}(10, 20)$
f(x)
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
x

Phân phối Student

Definition

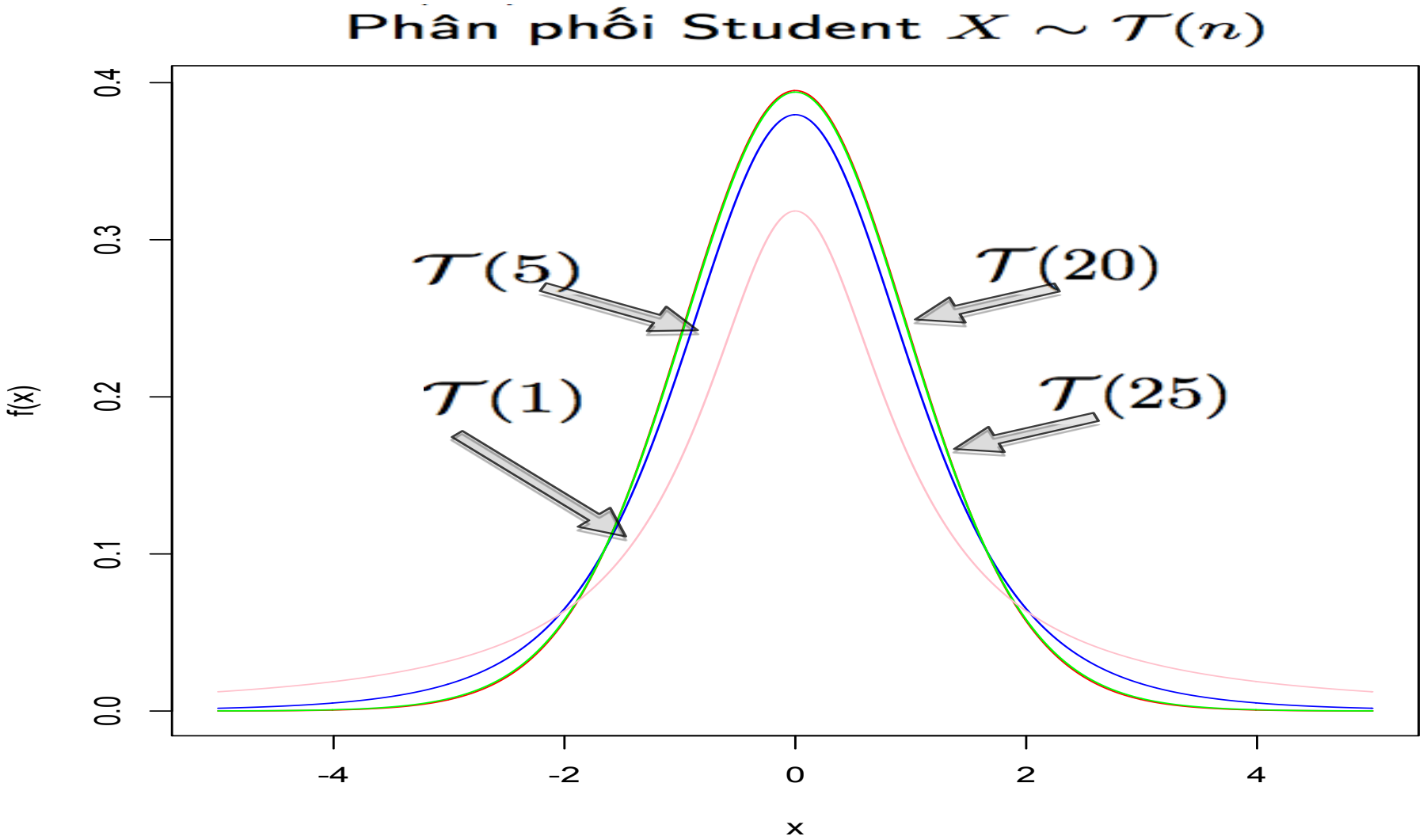
Giả sử $X \sim N(0;1)$ và $Y \sim \chi^2(n)$ là hai biến ngẫu nhiên độc lập. Khi đó:

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

được gọi là tuân theo phân phối Student với n bậc tự do.
Ký hiệu: $T \sim T(n)$

Các tham số đặc trưng

- $ET = 0$
- $VT = \frac{n}{n-2}$


Phân phối Student $X \sim \mathcal{T}(n)$
$\mathcal{T}(5)$
$\mathcal{T}(20)$
$\mathcal{T}(1)$
$\mathcal{T}(25)$
f(x)
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

Chú ý

- Phân phối Student có cùng dạng và tính đối xứng như phân phối chuẩn nhưng nó phản ánh tính biến đổi của phân phối sâu sắc hơn. Phân phối chuẩn không thể dùng để xấp xỉ phân phối khi mẫu có kích thước nhỏ. Trong trường hợp này ta dùng phân phối Student.
- Khi bậc tự do n tăng lên ($n > 30$) thì phân phối Student tiến nhanh về phân phối chuẩn. Do đó khi n $>$ 30 ta có thể dùng phân phối chuẩn thay thế cho phân phối Student.

Phương pháp nghịch đảo mô phỏng biến ngẫu nhiên chuẩn-CDF

```
#Mô phỏng 1000 biến ngẫu nhiên có phân phối chuẩn
mu=164; sigma=8
x = norm.rvs(mu, sigma, size=1000, random_state=123)
```

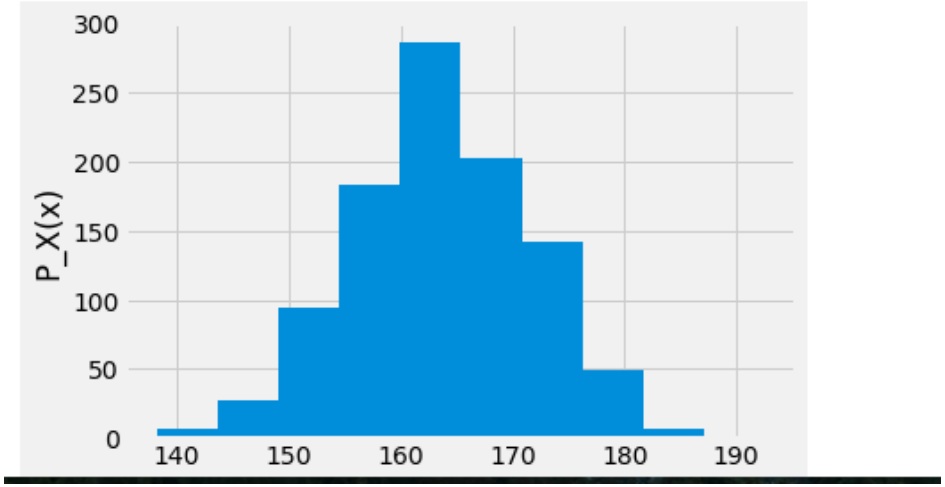
```
np.mean(x)#kỳ vọng
```

```
163.68348691135367
```

```
np.std(x)# độ lệch chuẩn
```

```
8.006300300129867
```

```
plt.hist(x)
plt.xlabel('x')
plt.ylabel('P_X(x)')
plt.show()
```



Các phân phối thông dụng trong pyhon của thư viện scipy.stats

- binom(*args, **kwds) A binomial discrete random variable.
- poisson(*args, **kwds) A Poisson discrete random variable.
- randint(*args, **kwds) A uniform discrete random variable.
- uniform(*args, **kwds) A uniform continuous random variable.
- norm(*args, **kwds) A normal continuous random variable.
- chi2(*args, **kwds) A chi-squared continuous random variable.
- t(*args, **kwds) A Student's t continuous random variable.
- f(*args, **kwds) An F continuous random variable.

Tất cả các phân phối của thư viện scipy.stats sẽ có các hàm:

- rvs() : tạo biến ngẫu nhiên
- pmf()/pdf: hàm trọng số (bnn rời rạc); hàm mật độ (bnn liên tục)
- cdf(): hàm phân phối
- ppf() : lấy quantile
- median(): trung vị
- mean(): kỳ vọng
- var() phương sai
- std(): độ lệch tiêu chuẩn

Bài tập thực hành 1

- Tạo ra 5000 biến ngẫu nhiên tuân theo luật phân phối Posion với tham số 30
- Tìm xác suất để biến ngẫu nhiên nhận giá trị là 35
- Tìm xác suất để biến ngẫu nhiên có giá trị nhỏ hơn hoặc bằng 15
- Tìm kỳ vọng, phương sai, độ lệch tiêu chuẩn, trung vị

Bài tập thực hành 2

- Tạo ra n=1000 số ngẫu nhiên tuân theo luật phân phối Student với bậc tự do là 15
- Tìm kỳ vọng, phương sai, độ lệch tiêu chuẩn
- Tìm giá trị phân vị mức 0.9; 0.95; 0.99
- Vẽ đồ thị của hàm mật độ
- Vẽ đồ thị hàm phân phối