Common Probability Distributions

August 18, 2023

[ ]:
```
from IPython.display import Image
```

1 Discrete and Continuous Random Variables

1.1 Discrete Random Variable

Biến ngẫu nhiên được gọi là *rời rạc* nếu tập giá trị của nó là một tập hữu hạn hoặc vô hạn đếm được các phần tử.

Miến giá trị của một biến ngẫu nhiên rời rạc là một dãy số $x_1, x_2, x_3, ..., x_n, ...$ có thể có hữu hạn hoặc vô hạn phần tử

1.2 Continuous Random Variable

Biến ngẫu nhiên được gọi là *liên tục* nếu tập giá trị của nó lấp kín một khoảng trên trục số (số phần tử của tập giá trị là vô hạn không đếm được theo lý thuyết số).

Miền giá trị của biến ngẫu nhiên liên tục là một đoạn $[a, b] \subset \mathbb{R}$ hoặc chính là $\mathbb{R}$

2 Probability functions

2.1 Cummulative Distribution Function

Hàm phân phối xác suất của biến ngẫu nhiên X, ký hiệu là $F(x)$ được định nghĩa như sau:

$$F(x) = P(X < x), x \in \mathbb{R}$$

Các tính chất của $F(x)$ được trình bày dưới đây:

1. $0 \leq F(x) \leq 1$
2. $F(x)$ là hàm không giảm, có nghĩa là nếu $x_1 < x_2$ thì $F(x_1) \leq F(x_2)$
3. $P(a \leq X \leq B) = F(b) - F(a)$. Tính chất này có một hệ quả là $P(X = a) = 0$
4. $F(-\infty) = 0$, $F(+\infty) = 1$

Lưu ý:

- Đối với một biến ngẫu nhiên rời rạc, đồ thị của hàm phân phối xác suất có dạng *bậc thang*

- Việc xác định hàm phân phối xác suất đối với biến ngẫu nhiên liên tục tương đương với việc xác định toàn bộ biến ngẫu nhiên liên tục đó. Tuy nhiên, trong thực tế, gần như chúng ta không thể thực hiện được công việc này

2.2 Probability density function

Hàm mật độ xác suất của biến ngẫu nhiên X có hàm phân phối $F(x)$ khả vi (trừ hữu hạn điểm bị chặn), ký hiệu là $f(x)$, được xác định bằng đạo hàm của hàm phân phối:

$$f(x) = F'(x)$$

Với khái niệm đạo hàm và tích phân, chúng ta suy ra công thức hàm phân phối xác suất:

$$F(x) = \int_{-\infty}^{+\infty} f(x)\, dx$$

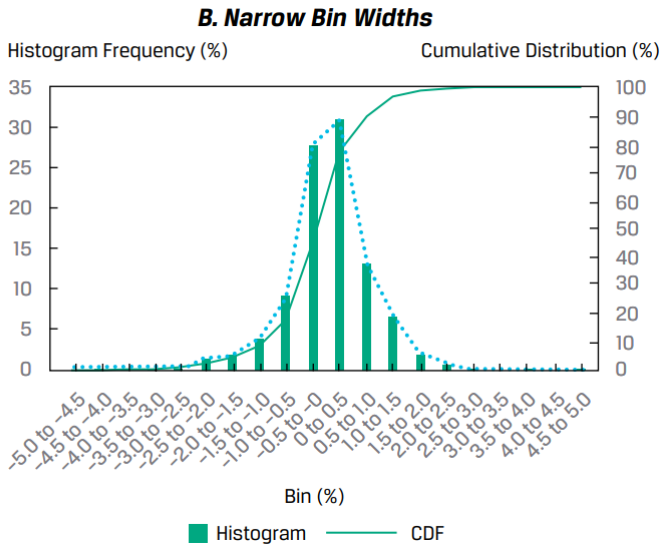
Các tính chất của hàm này có thể được kể đến như sau:

1. $f(x) \geq 0 \ \forall x$
2. $\int_{-\infty}^{+\infty} f(x)\, dx = 1$

[ ]:
```
# Hàm mật độ xác suất (PDF) và hàm phân phối xác suất (CDF)

Image(filename = "Pictures/01.png")
```

[ ]:



3 Discrete and Continuous Uniform Distribution

3.1 Discrete Uniform Distribution

Biến ngẫu nhiên X được gọi là tuân thủ theo luật phân phối đều rời rạc, nếu X có bảng phân phối xác suất như sau:

[ ]:
```
Image(filename = "Pictures/02.png")
```

[ ]:

x	1	2	...	n
$p(x)$	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

Có nghĩa là, hàm xác suất có dạng $p(i) = \frac{1}{n},\ i = \overline{1,n}$

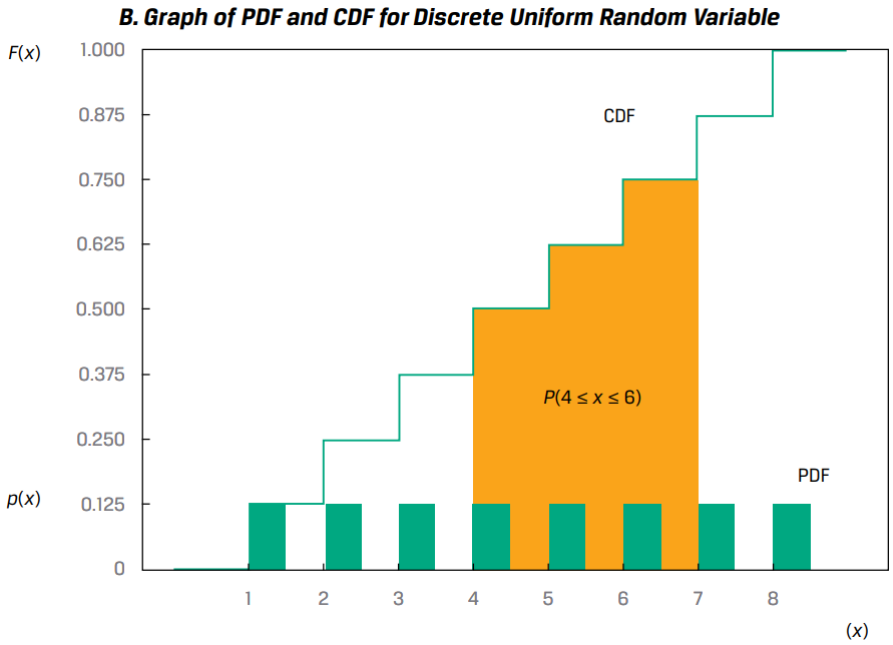
Mở rộng khái niệm phân phối đều cho biến X nhận giá trị trên một tập hữu hạn bất kỳ có n phần tử $\{x_1, x_2, x_3, ...x_n\}$, khi đó:

$$p(x_i) = \frac{1}{n},\ i = \overline{1,n}$$

Xét về đồ thị của phân phối đều rời rạc, PDF sẽ là các cột có chiều cao bằng chau, trong khi CDF có dạng bậc thang với độ cao của các bậc bằng nhau

[ ]:
```
Image(filename = "Pictures/03.png")
```

[ ]:



3.2 Continuous Uniform Variable

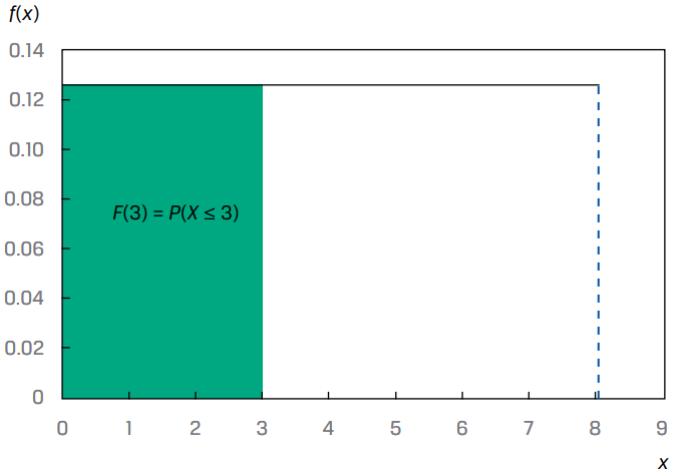
Biến ngẫu nhiên X được gọi là tuân thủ theo luật phân phối đều liên tục trên $[a, b]$ nếu X có hàm mật độ:

$$f(x) = \frac{1}{b-a},\ x \in [a, b]\ and\ f(x) = 0,\ x \notin [a, b]$$

PDF của phân phối đều liên tục trên đoạn $[a, b]$ là một hình chữ nhật được giới hạn bởi trục hoành và ba đường thẳng $y = \frac{1}{b-a}, x = a, x = b$, trong khi CDF bao gồm 3 nhánh

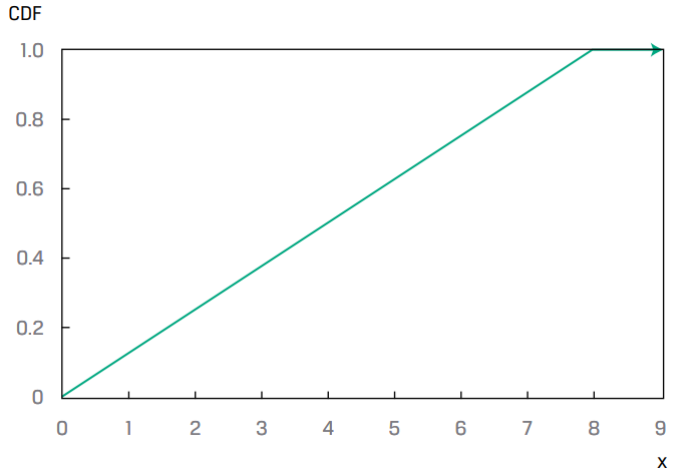
[ ]:
```
Image(filename = "Pictures/04.png")
```

[ ]:



[ ]:
```
Image(filename = "Pictures/05.png")
```

[ ]:



4 Binomal Distribution

4.1 Bernoulli Distribution

Biến ngẫu nhiên X được gọi là tuân theo luật phân phối Bernoulli, ký hiệu là $X \sim B(1, p)$ nếu hàm xác suất của nó có dạng

$$p(x) = p^x(1-p)^{1-x},\ x = 0, 1$$

Chúng ta thấy mọi phép thử chỉ có hai kết cục đều có thể mô hình hóa bằng phân phối Bernoulli. Từ hàm xác suất, chúng ta dễ dàng suy ra các tính chất của phân phối:

1. $E(X) = p$
2. $Var(X) = p(1-p)$

4.2 Binomial Distribution

Biến ngẫu nhiên X được gọi là tuân theo luật phân phối nhị thức, ký hiện là $X \sim B(n, p)$ nếu hàm xác suất của nó có dạng

$$p(x) = C_n^x p^x (1-p)^{n-x}$$

Phân phối Bernoulli ở trên rõ ràng là một trường hợp riêng của phân phối nhị thức (cụ thể là trường hợp $n = 1$.

Các điều kiện của phân phối nhị thức là:

1. Dãy các phép thử giống nhau và độc lập
2. Mỗi phép thử chỉ có hai kết cục
3. Hai tham số hằng xác định: số các phép thử n và xác suất xuất hiện một trong hai kết cục p

Các tính chất liên quan đến phân phối nhị thức có thể được nhắc đến như sau:

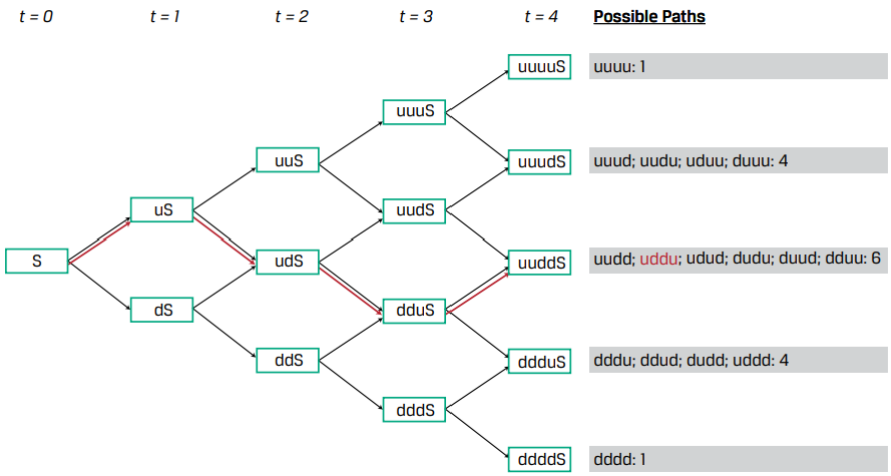
1. $E(X) = np$
2. $Var(X) = np(1-p)$
3. Đồ thị hàm mật độ đối xứng khi $p = 0.5$, lệch trái nếu $p > 0.5$ và lệch phải nếu $p < 0.5$

Lưu ý: Phân phối chuẩn có thể là một xấp xỉ tốt cho phân phối nhị thức khi p không quá gần 0 hoặc 1 đồng thời n tương đối lớn. $B(n, p)$ rất gần với $N(np, np(1-p))$ và việc sử dụng phân phối xấp xỉ sẽ rất tốt nếu $np \geq 5$ khi $p \leq 0.5$ hoặc $n(1-p) \geq 5$ khi $p \geq 0.5$

Chúng ta có thể biểu diễn phân phối nhị thức thông qua cây nhị thức, biểu đồ mật độ hoặc biểu đồ phân phối xác suất

[ ]:
```
# Cây nhị thức
Image(filename = "Pictures/06.png")
```

[ ]:



[ ]:
```
# PDF và CDF
Image(filename = "Pictures/07.png")
```

[ ]:

A. Binomial Probabilities, $n = 4$ and $p = 0.50$

Col. 1 Number of Up Moves, x	Col. 2 Implied Number of Down Moves, $n - x$	Col. 3[A] Number of Possible Ways to Reach x Up Moves	Col. 4[B] Probability for Each Way, $p(x)$	Col. 5[C] Probability for x $p(x)$	Col. 6 $F(x) = P(X \le x)$
0	4	1	0.0625	0.0625	0.0625
1	3	4	0.0625	0.2500	0.3125
2	2	6	0.0625	0.3750	0.6875
3	1	4	0.0625	0.2500	0.9375
4	0	1	0.0625	0.0625	1.0000
				1.0000	

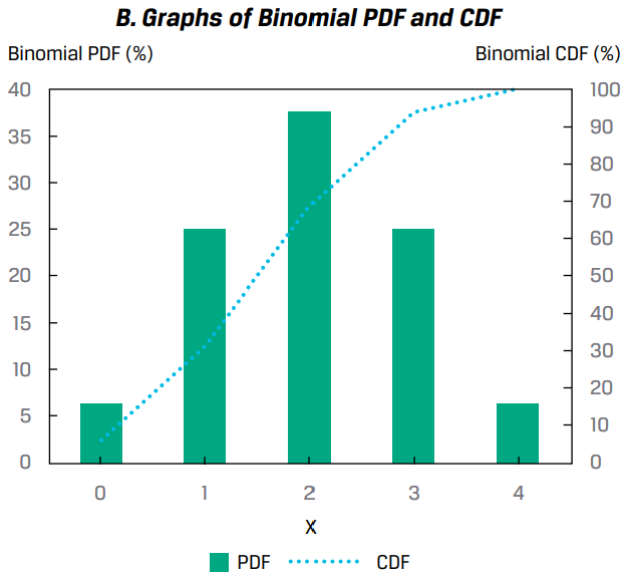
A: Column 3 = $n!$ / $[(n - x)!\, x!]$

B: Column 4 = $p^x(1 - p)^{n-x}$

C: Column 5 = Column 3 × Column 4

[ ]:
```
Image(filename = "Pictures/08.png")
```

[ ]:



5 Normal Distribution

5.1 Concepts

Biến ngẫu nhiên X được gọi là tuân theo luật phân phối chuẩn, ký hiệu là $X \sim N(\mu, \sigma^2)$ nếu hàm mật độ của nó có dạng:

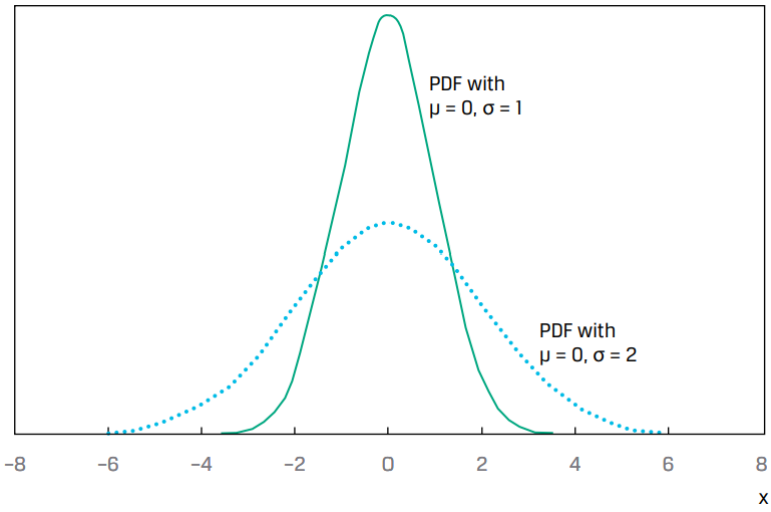
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Một phân phối chuẩn có trung bình $\mu = 0$ và phương sai $\sigma^2 = 1$ được gọi là một phân phối chuẩn chuẩn hóa (`standard normal distribution`), hoặc là phân phối chuẩn đơn vị (`unit normal distribution`). Hàm mật độ xác suất của phân phối chuẩn chuẩn hóa là:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

[ ]:
```
#  Đồ thị hàm mật độ của phân phối chuẩn
Image(filename = "Pictures/09.png")
```

[ ]:

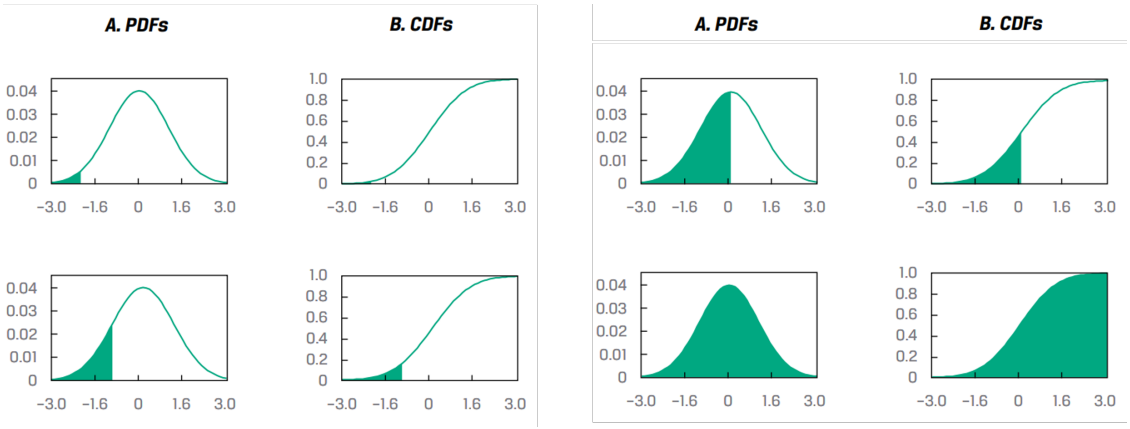


Từ hàm mật độ, chúng ta dễ thấy được các đặc điểm xác định của một phân phối chuẩn:

1. Phân phối chuẩn được xác định hoàn toàn bởi hai tham số là trung bình μ và phương sai σ^2
2. Phân phối chuẩn có dạng hình chuông, đối xứng (skewness bằng 0) và độ lồi tiêu chuẩn (kurtosis bằng 3)
3. Bất kỳ một sự kết hợp tuyến tính nào giữa các biến ngẫu nhiên cũng tuân theo phân phối chuẩn

PDF và CDF của phân phối chuẩn được thể hiện trong hình dưới đây

[ ]:
```
# PDF và CDF của phân phối chuẩn
Image(filename = "Pictures/10.png")
```

[ ]:


Các tính chất nêu trên đây chỉ xét đến một biến đơn. Phân phối một biến (`univariate distribution`) chỉ mô tả một biến đơn, trong khi phân phối đa biến (`multivariate`

`distribution`) mô tả xác suất của một nhóm biến ngẫu nhiên. Một phân phối với n biến hoàn toàn được xác định bởi ba thông số:

- n giá trị trung bình
- n giá trị phương sai
- $n(n-1)/2$ giá trị hệ số tương quan

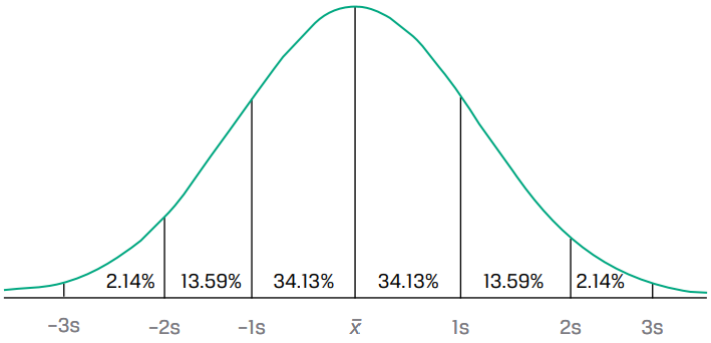
5.2 Probabilities Using Normal Distribution

Với một biến ngẫu nhiên tuân theo phân phối chuẩn $N(\mu, \sigma^2)$, dễ thấy trung bình, trung vị và mode bằng nhau và cùng bằng μ. Nếu ta xét các vùng lân cận μ, chúng ta có thể đưa ra các tuyên bố xác suất như sau:

- $P(|X - \mu| < 2/3\sigma) = 50\%$
- $P(|X - \mu| < \sigma) = 68.26\%$
- $P(|X - \mu| < 2\sigma) = 95.44\%$
- $P(|X - \mu| < 3\sigma) = 99.74\%$

[ ]:
```
# Xác suất vùng lân cận giá trị trung bình
Image(filename = "Pictures/11.png")
```

[ ]:



5.3 Standardizing a Random Variable

Quy trình chuẩn hóa một biến ngẫu nhiên X tuân theo luật phân phối chuẩn bao gồm hai bước: **(1)** trừ X cho giá trị trung bình của X và **(2)** chia kết quả vừa thu được cho độ lệch chuẩn của X

$$Z = \frac{X - \mu}{\sigma}$$

Với cách làm này, chúng ta đã chuẩn hóa $X \sim N(\mu, \sigma)$ về một phân phối chuẩn chuẩn hóa $Z \sim N(0, 1)$

5.4 Central Limit Theorem

Nội dung định lý: giả sử $\{X_n\}$ là dãy các biến ngẫu nhiên độc lập có cùng phân phối với $E(X_n) = m$ và $V(X_n) = \sigma^2\ \forall n$, khi đó:

$$\frac{\overline{X_n} - m}{\sigma}\sqrt{n} \xrightarrow[n\to\infty]{L} N(0,1)$$

với $X_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Định lý giới hạn trung tâm cho rằng khi có nhiều nhân tố ngẫu nhiên tác động (sao cho không có nhân tố nào lấn át vượt trội các nhân tố khác), thì kết quả của chúng có dạng phân phối tiệm cận chuẩn.

6 Student's t, χ^2 and F Distributions

6.1 χ^2 Distribution

Phân phối χ^2 với n bậc tự do, ký hiệu là $\chi^2(n)$, có thể được định nghĩa bằng hàm mật độ

$$f(x) = \frac{x^{\frac{n}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)},\ x > 0,\ n > 0$$

trong đó hàm gamma đã được xét trong giải tích

$$\Gamma(x) = \int_0^x t^{x-1}e^{-t}\,dt$$

$\forall i$ nguyên, hàm gamma có tính chất:

- $\Gamma(i+1) = i!\ (i \geq 0)$
- $\Gamma\left(\frac{i}{2}\right) = \left(\frac{i}{2}-1\right)\left(\frac{i}{2}-2\right)...\frac{3}{2}\frac{1}{2}\sqrt{\pi}$, với i là số lẻ lớn hơn 2
- $\Gamma(x) = (x-1)\Gamma(x-1),\ x \in \mathbb{R}$

Tuy nhiên cách định nghĩa này khá phức tạp. Chúng ta có một cách định nghĩa khác xác định rõ ràng phân phối χ^2 có xuất thân từ phân phối chuẩn.

Xét n biến ngẫu nhiên độc lập $X_i \sim N(0,1),\ i = \overline{1,n}$. Khi đó:

$$U_n = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

Các tính chất của phân phối χ^2 có thể kể đến là:

1. Đồ thị hàm mật độ nằm hoàn toàn trong góc phần tư thứ nhất của hệ trục tọa độ Descartes

2. Đồ thị hàm mật độ phi đối xứng, nhưng sẽ gần dạng chuông đối xứng hơn nếu bậc tự do tăng lên

6.2 Student's t Distribution

Với hai biến ngẫu nhiên độc lập X và Y tuân theo luật $N(0,1)$ và $\chi^2(n)$ tương ứng, khi đó:

$$T_n = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n)$$

Phân phối Student's t có một số tính chất tiêu biểu:

1. Mỗi bậc tự do khác nhau xác định một phân phối t khác nhau trong họ phân phối t (phân phối này có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1)
2. Đồ thị hàm mật độ có dạng chuông đối xứng tương tự với phân phối chuẩn chuẩn hóa
3. Độ lồi của hàm mật độ thấp hơn so với phân phối chuẩn (platykurtic), tuy nhiên sẽ gần hơn với đồ thị phân phối chuẩn chuẩn hóa khi bậc tự do tăng lên

6.3 F Distribution

Với hai biến ngẫu nhiên độc lập X và Y tuân theo luật $\chi^2(n)$ và $\chi^2(m)$ tương ứng, khi đó:

$$U = \frac{X/n}{Y/m} \sim F(n,m)$$

Phân phối F cũng có một số tính chất tương tự với phân phối χ^2. Cụ thể:

1. 1. Đồ thị hàm mật độ nằm hoàn toàn trong góc phần tư thứ nhất của hệ trục tọa độ Descartes
2. Đồ thị hàm mật độ phi đối xứng, nhưng sẽ gần với dạng chuông đối xứng hơn nếu hai bậc bậc tự do tăng lên