Xác Suất

Bởi:

Phạm Trí Cao

Xác suất biến ngẫu nhiên nhận được một giá trị cụ thể

Chúng ta thường quan tâm đến xác suất biến ngẫu nhiên nhận được một giá trị xác định. Ví dụ khi ta sắp tung một súc sắc và ta muốn biết xác suất xuất hiện Xi = 4 là bao nhiêu.

Do con súc sắc có 6 mặt và nếu không có gian lận thì khả năng xuất hiện của mỗi mặt đều như nhau nên chúng ta có thể suy ra ngay xác suất để X= 4 là: P(X=4) = 1/6.

Nguyên tắc lý do không đầy đủ(the principle of insufficient reason): Nếu có K kết quả có khả năng xảy ra như nhau thì xác suất xảy ra một kết quả là 1/K.

Không gian mẫu: Một không gian mẫu là một tập hợp tất cả các khả năng xảy ra của một phép thử, ký hiệu cho không gian mẫu là S. Mỗi khả năng xảy ra là một điểm mẫu.

Biến cố : Biến cố là một tập con của không gian mẫu.

Ví dụ 2.3. Gọi Z là tổng số điểm phép thử tung hai con súc sắc.

Không gian mẫu là S = {2;3;4;5;6;7;8;9;10;11;12}

A = {7;11}Tổng số điểm là 7 hoặc 11

B = {2;3;12}Tổng số điểm là 2 hoặc 3 hoặc 12

C = {4;5;6;8;9;10}

D = {4;5;6;7}

Là các biến cố.

Hợp của các biến cố

E = A hoặc B = $A \cup B$ = {2;3;7;11;12}

Giao của các biến cố:

F = C và D = $C \cap D$ = {4;5;6}

Các tính chất của xác suất

P(S) =1

$$0 \le P(A) \le 1$$
$$P(E) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Tần suất

Khảo sát biến X là số điểm khi tung súc sắc. Giả sử chúng ta tung n lần thì số lần xuất hiện giá trị xi là ni. Tần suất xuất hiện kết quả xi là

$$f_i = \frac{n_i}{n}$$

Nếu số phép thử đủ lớn thì tần suất xuất hiện xi tiến đến xác suất xuất hiện xi.

Định nghĩa xác suất

Xác suất biến X nhận giá trị xi là

$$P(X = xi) = \lim_{n \to \infty} \frac{n_i}{n}$$

Hàm mật độ xác suất (phân phối xác suất)

Hàm mật độ xác suất-Biến ngẫu nhiên rời rạc

X nhận các giá trị xi riêng rẽ $x_1, x_2, \ldots, x_n$. Hàm số

f(x) = P(X=xi) , với i = 1;2;..;n

= 0 , với x xi

được gọi là hàm mật độ xác suất rời rạc của X. P(X=xi) là xác suất biến X nhận giá trị xi.

Xét biến ngẫu nhiên X là số điểm của phép thử tung một con súc sắc. Hàm mật độ xác suất được biểu diễn dạng bảng như sau.

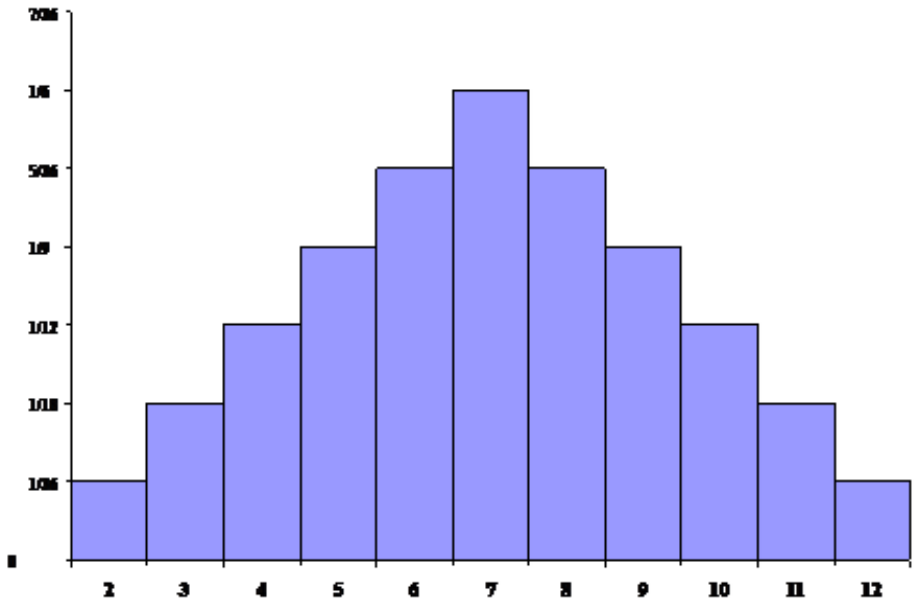
X	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6

Bảng 2.1. Mật độ xác suất của biến ngẫu nhiên rời rạc X

Xét biến Z là tổng số điểm của phép thử tung 2 con súc sắc. Hàm mật độ xác suất được biểu diễn dưới dạng bảng như sau.

z	2	3	4	5	6	7	8	9	10	11	12
P(Z=z)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Bảng 2.2. Mật độ xác suất của biến ngẫu nhiên rời rạc Z



Hình 2.1. Biểu đồ tần suất của biến ngẫu nhiên Z.

Hàm mật độ xác suất(pdf)-Biến ngẫu nhiên liên tục.

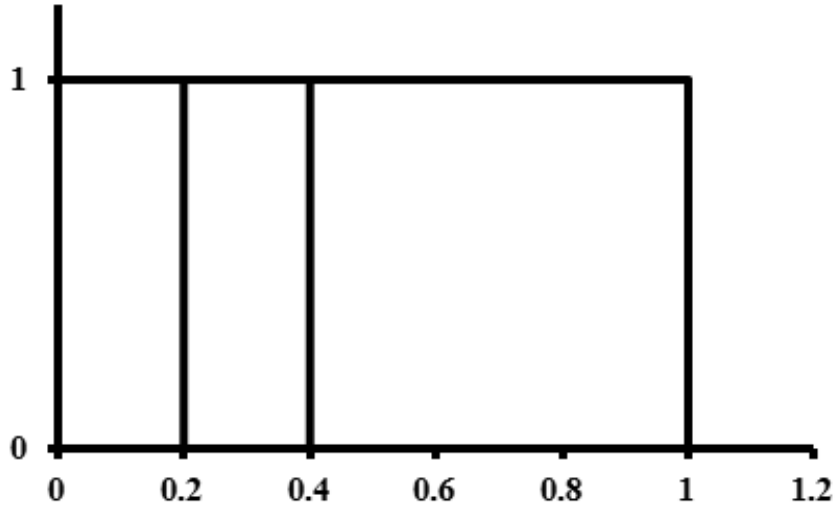
Ví dụ 2.4. Chúng ta xét biến R là con số xuất hiện khi bấm nút Rand trên máy tính cầm tay dạng tiêu biểu như Casio fx-500. R là một biến ngẫu nhiên liên tục nhận giá trị bất kỳ từ 0 đến 1. Các nhà sản xuất máy tính cam kết rằng khả năng xảy ra một giá trị cụ thể là như nhau. Chúng ta có một dạng phân phối xác suất có mật độ xác suất đều.

Hàm mật độ xác suất đều được định nghĩa như sau:f(r) =

$$\frac{1}{U-L}$$

Với L : Giá trị thấp nhất của phân phối

U: Giá trị cao nhất của phân phối



Hình 2.2. Hàm mật độ xác suất đều R.

Xác suất để R rơi vào khoảng (a; b) là P(a <r<b) =

$$\frac{b-a}{U-L}$$

.

Cụ thể xác suất để R nhận giá trị trong khoảng (0,2; 0,4) là:

P(0,2 < r < 0,4) =

$$\frac{0,4-0,2}{1-0} = 20\%$$

, đây chính là diện tích được gạch chéo trên hình 2.1.

Tổng quát, hàm mật độ xác suất của một biến ngẫu nhiên liên tục có tính chất như sau:

f(x) ≥ 0

P(a<X<b) = Diện tích nằm dưới đường pdf

P(a<X<b) =

$$\int_a^b f(x)dx$$

$$\int_S f(x)dx = 1$$

Hàm đồng mật độ xác suất -Biến ngẫu nhiên rời rạc

Ví dụ 2.5. Xét hai biến ngẫu nhiên rời rạc X và Y có xác suất đồng xảy ra X = xi và Y = yi như sau.

Y \ X	2	3	P(Y)
1	0,2	0,4	0,6
2	0,3	0,1	0,4
P(X)	0,5	0,5	1,0

Bảng 2.3. Phân phối đồng mật độ xác xuất của X và Y.

Định nghĩa :Gọi X và Y là hai biến ngẫu nhiên rời rạc. Hàm số

f(x,y) = P(X=x và Y=y)

= 0 khi X x và Y y

được gọi là hàm đồng mật độ xác suất, nó cho ta xác xuất đồng thời xảy ra X=x và Y=y.

Hàm mật độ xác suất biên

f(x) = $\sum_y f(x,y)$ hàm mật độ xác suất biên của X

f(y) = $\sum_x f(x,y)$ hàm mật độ xác suất biên của Y

Ví dụ 2.6. Ta tính hàm mật độ xác suất biên đối với số liệu cho ở ví dụ 2.5.

f(x=2) = $\sum_y f(x = 2,y)$=0,3 + 0,3 = 0,5

f(x=3) = $\sum_y f(x = 3,y)$=0,1 + 0,4 = 0,5

f(y=1) = $\sum_x f(x,y = 1)$=0,2 + 0,4 = 0,6

f(y=2) = $\sum_x f(x,y = 2)$=0,3 +0,1 = 0,4

Xác suất có điều kiện

Hàm số

f(x | y) = P(X=x | Y=y) , xác suất X nhận giá trị x với điều kiện Y nhận giá trị y,

được gọi là xác suất có điều kiện của X.

Hàm số

f(y | x) = P(Y=y | X=x) , xác suất Y nhận giá trị y với điều kiện X nhận giá trị x,

được gọi là xác suất có điều kiện của Y.

Xác suất có điều kiện được tính như sau

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

, hàm mật độ xác suất có điều kiện của X

$$f(y|x) = \frac{f(x,y)}{f(x)}$$

, hàm mật độ xác suất có điều kiện của Y

Như vậy hàm mật độ xác suất có điều kiện của một biến có thể tính được từ hàm đồng mật độ xác suất và hàm mật độ xác suất biên của biến kia.

Ví dụ 2.7. Tiếp tục ví dụ 2.5 và ví dụ 2.6.

$$f(X=2|Y=1) = \frac{f(X=2,Y=1)}{f(Y=1)} = \frac{0,2}{0,6} = \frac{1}{3}$$

$$f(Y=2|X=3) = \frac{f(X=3,Y=2)}{f(X=3)} = \frac{0,1}{0,5} = \frac{1}{5}$$

Độc lập về thống kê

Hai biến ngẫu nhiên X và Y độc lập về thống kê khi và chỉ khi

f(x,y)=f(x)f(y)

tức là hàm đồng mật độ xác suất bằng tích của các hàm mật độ xác suất biên.

Hàm đồng mật độ xác suất cho biến ngẫu nhiên liên tục

Hàm đồng mật độ xác suất của biến ngẫu nhiên liên tục X và Y là f(x,y) thỏa mãn

$f(x,y) \geq 0$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

$$\int_{a}^{b}\int_{c}^{d} f(x,y)dxdy = P(a \leq x \leq b, c \leq y \leq d)$$

Hàm mật độ xác suất biên được tính như sau

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

, hàm mật độ xác suất biên của X

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

, hàm mật độ xác suất biên của Y

Một số đặc trưng của phân phối xác suất

Giá trị kỳ vọng hay giá trị trung bình

Giá trị kỳ vọng của một biến ngẫu nhiên rời rạc

$$E(X) = \sum_{x} xf(x)$$

Giá trị kỳ vọng của một biến ngẫu nhiên liên tục

$$E(X) = \int_x xf(x)dx$$

Ví dụ 2.8. Tính giá trị kỳ vọng biến X là số điểm của phép thử tung 1 con súc sắc

$$E(X) = 1*\frac{1}{6} + 2*\frac{1}{6} + 3*\frac{1}{6} + 4*\frac{1}{6} + 5*\frac{1}{6} + 6*\frac{1}{6} = 3,5$$

Một số tính chất của giá trị kỳ vọng

E(a) = avới a là hằng số

E(a+bX) = a + bE(X)với a và b là hằng số

Nếu X và Y là độc lập thống kê thì E(XY) = E(X)E(Y)

Nếu X là một biến ngẫu nhiên có hàm mật độ xác suất f(x) thì

$$E[g(X)] = \sum_x g(X)f(x)$$

, nếu X rời rạc

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

, nếu X liên tục

Người ta thường ký hiệu kỳ vọng là μ : μ = E(X)

Phương sai

X là một biến ngẫu nhiên và μ = E(X). Độ phân tán của dữ liệu xung quanh giá trị trung bình được thể hiện bằng phương sai theo định nghĩa như sau:

$\text{var}(X) = \sigma_X^2 = E(X - \mu)^2$

Độ lệch chuẩn của X là căn bậc hai dương của σ_X^2, ký hiệu là σ_X.

Ta có thể tính phương sai theo định nghĩa như sau

$$\text{var}(X) = \sum_x (X - \mu)^2 f(x)$$

, nếu X là biến ngẫu nhiên rời rạc

$$= \int_{-\infty}^{+\infty} (X-\mu)^2 f(x)dx$$

, nếu X là biến ngẫu nhiên liên tục

Trong tính toán chúng ta sử dụng công thức sau

$var(X)=E(X^2)-[E(X)]^2$

Ví dụ 2.9. Tiếp tục ví dụ 2.8. Tính var(X)

Ta đã có E(X) = 3,5

Tính $E(X^2)$ bằng cách áp dụng tính chất (4).

$E(X^2)$ =

$$1^2 * \frac{1}{6} + 2^2 * \frac{1}{6} + 3^2 * \frac{1}{6} + 4^2 * \frac{1}{6} + 5^2 * \frac{1}{6} + 6^2 * \frac{1}{6} =$$

15,17

$var(X)=E(X^2)-[E(X)]^2 = 15,17 – 3,5^2 = 2,92$

Các tính chất của phương sai

$E(X- \mu)^2 =E(X^2)^2 - \mu^2$

var(a) = 0 với a là hằng số

$var(a+bX) = b^2 var(X)$với a và b là hằng số

Nếu X và Y là các biến ngẫu nhiên độc lập thì

var(X+Y) = var(X) + var(Y)

var(X-Y) = var(X) + var(Y)

Nếu X và Y là các biến độc lập, a và b là hằng số thì

$var(aX+bY) = a^2 var(X) + b^2 var(Y)$

Hiệp phương sai

X và Y là hai biến ngẫu nhiên với kỳ vọng tương ứng là μ_X và μ_Y. Hiệp phương sai của hai biến là

$\text{cov}(X,Y) = E[(X- \mu_X)(Y- \mu_Y)] = E(XY) - \mu_X\mu_Y$

Chúng ta có thể tính toán trực tiếp hiệp phương sai như sau

Đối với biến ngẫu nhiên rời rạc

$\text{cov}(X,Y)$

$$= \sum_y \sum_x (X - \mu_x)(Y - \mu_y) f(x,y)$$

$$= \sum_y \sum_x XYf(x,y) - \mu_x\mu_y$$

Đối với biến ngẫu nhiên liên tục

$$\text{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(X-\mu_x)(Y-\mu_y)f(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}XYf(x,y)dxdy - \mu_x\mu_y$$

Tính chất của hiệp phương sai

Nếu X và Y độc lập thống kê thì hiệp phương sai của chúng bằng 0.

$\text{cov}(X,Y) = E(XY) - \mu_X\mu_Y$

$= \mu_X\mu_Y - \mu_X\mu_Y$

$=0$

$\text{cov}(a+bX,c+dY)=bd\text{cov}(X,Y)$ với a,b,c,d là các hằng số

Nhược điểm của hiệp phương sai là nó phụ thuộc đơn vị đo lường.

Hệ số tương quan

Để khắc phục nhược điểm của hiệp phương sai là phụ thuộc vào đơn vị đo lường, người ta sử dụng hệ số tương quan được định nghĩa như sau:

$$\rho_{xy} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} = \frac{\operatorname{cov}(X,Y)}{\sigma_x \sigma_y}$$

Hệ số tương quan đo lường mối quan hệ tuyến tính giữa hai biến. ρ sẽ nhận giá trị nằm giữa -1 và 1. Nếu ρ=-1 thì mối quan hệ là nghịch biến hoàn hảo, nếu ρ=1 thì mối quan hệ là đồng biến hoàn hảo.

Từ định nghĩa ta có

$\operatorname{cov}(X,Y) = \rho\sigma_x\sigma_y$

Tính chất của biến tương quan

Gọi X và Y là hai biến có tương quan

$$\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X,Y)$$
$$= \operatorname{var}(X) + \operatorname{var}(Y) + 2\rho\ \sigma_x \sigma_y$$
$$\operatorname{var}(X-Y) = \operatorname{var}(X) + \operatorname{var}(Y) - 2\operatorname{cov}(X,Y)$$
$$= \operatorname{var}(X) + \operatorname{var}(Y) - 2\rho\ \sigma_x \sigma_y$$

Mô men của phân phối xác suất

Phương sai của biến ngẫu nhiên X là mô men bậc 2 của phân phối xác suất của X.

Tổng quát mô men bậc k của phân phối xác suất của X là

$E(X-\mu)^k$

Mô men bậc 3 và bậc 4 của phân phối được sử dụng trong hai số đo hình dạng của phân phối xác suất là skewness(độ bất cân xứng) và kurtosis(độ nhọn) mà chúng ta sẽ xem xét ở phần sau.

Một số phân phối xác suất quan trọng

Phân phối chuẩn

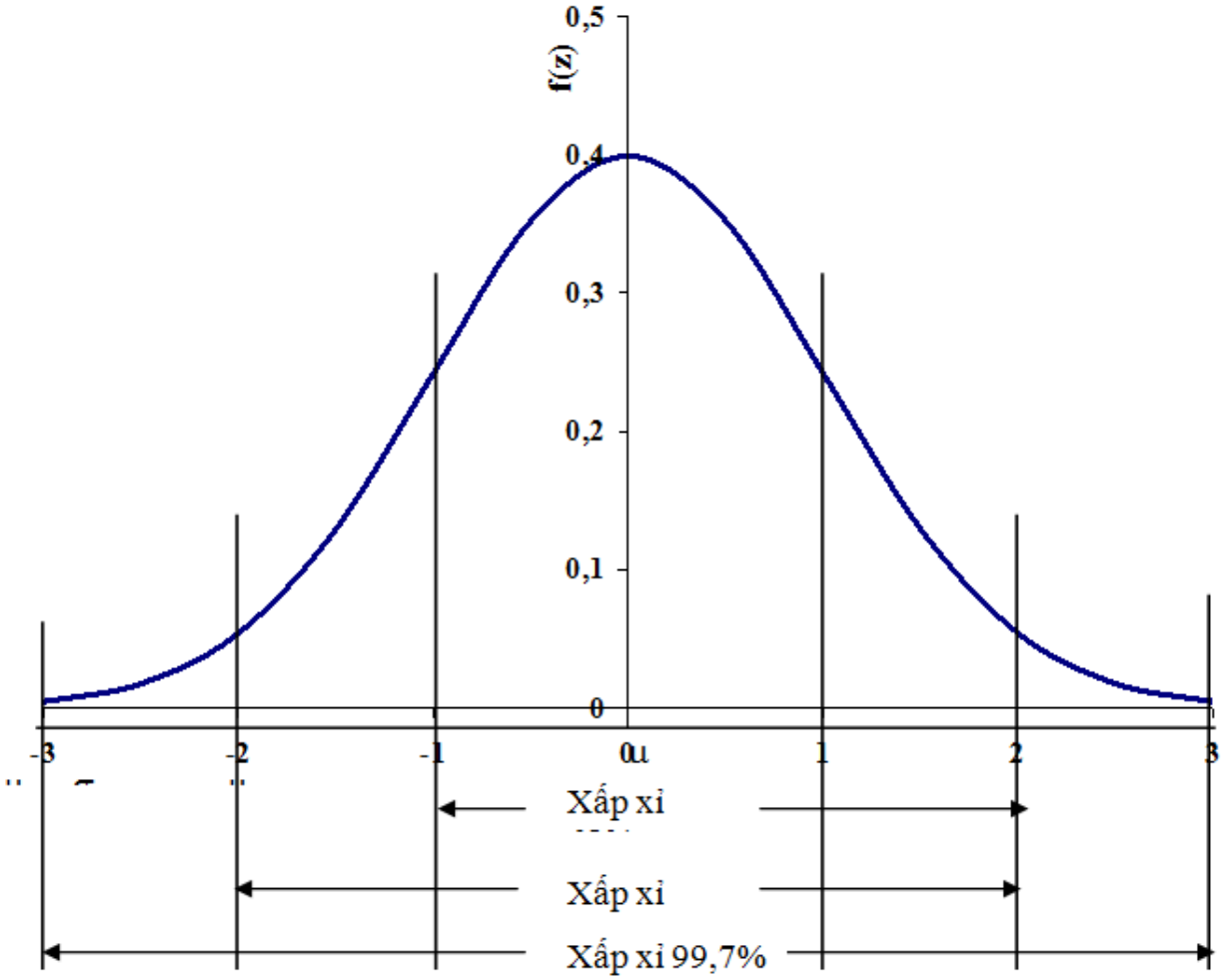
Biến ngẫu nhiên X có kỳ vọng là μ, phương sai là σ^2. Nếu X có phân phối chuẩn thì nó được ký hiệu như sau

$X \sim N(\mu,\sigma^2)$

Dạng hàm mật độ xác xuất của phân phối chuẩn như sau

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

μ- σ μ σ μ σ μ σ μ σ μ σμXấp xỉ 68%Xấp xỉ 95%



Hình 2.3. Hàm mật độ xác suất phân phối chuẩn

Tính chất của phân phối chuẩn

Hàm mật độ xác suất của đối xứng quanh giá trị trung bình.

Xấp xỉ 68% diện tích dưới đường pdf nằm trong khoảng $\mu \pm \sigma$ xấp xỉ 95% diện tích nằm dưới đường pdf nằm trong khoảng $\mu \pm 2\sigma$và xấp xỉ 99,7% diện tích nằm dưới đường pdf nằm trong khoảng $\mu \pm 3\sigma$

Nếu đặt Z =

$$(X-\mu)/\sigma$$

thì ta có Z~N(0,1). Z gọi là biến chuẩn hoá và N(0,1) được gọi là phân phối chuẩn hoá.

Định lý giới hạn trung tâm 1: Một kết hợp tuyến tính các biến có phân phối chuẩn,, trong một số điều kiện xác định cũng là một phân phối chuẩn. Ví dụ $X_1 \sim N(\mu_1,\sigma_1^2)$ và $X_2 \sim N(\mu_2,\sigma_2^2)$ thì $Y = aX_1 + bX_2$ với a và b là hằng số có phân phối $Y \sim N[(a\mu_1 + b\mu_2),(a^2\sigma_1^2 + b^2\sigma_2^2)]$.

Định lý giới hạn trung tâm 2: Dưới một số điều kiện xác định, giá trị trung bình mẫu của các một biến ngẫu nhiên sẽ gần như tuân theo phân phối chuẩn.

Mô men của phân phối chuẩn

Mô men bậc ba: $E[(X-\mu)^3]=0$

Mô men bậc bốn : $E[(X-\mu)^4]=3\sigma^4$

Đối với một phân phối chuẩn

Độ trôi (skewness):

$$S = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = 0$$

Độ nhọn(kurtosis):

$$K = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = 3$$

Dựa vào kết quả ở mục (6), người có thể kiểm định xem một biến ngẫu nhiên có tuân theo phân phối chuẩn hay không bằng cách kiểm định xem S có gần 0 và K có gần 3 hay không. Đây là nguyên tắc xây dựng kiểm định quy luật chuẩn Jarque-Bera.

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right]$$

JB tuân theo phân phối x^2với hai bậc tự do(df =2).

Phân phối x^2

Định lý : Nếu $X_1, X_2,\ldots, X_k$ là các biến ngẫu nhiên độc lập có phân phối chuẩn hoá thì $\chi_k^2 = \sum_{i=1}^{k} X_i^2$ tuân theo phân phối Chi-bình phương với k bậc tự do.

Tính chất của x^2

Phân phối x^2 là phân phối lệch về bên trái, khi bậc tự do tăng dần thì phân phối x^2 tiến gần đến phân phối chuẩn.

$\mu = k$ và $\sigma^2 = 2k$

$\chi_{k1}^2 + \chi_{k2}^2 = \chi_{k1+k2}^2$, hay tổng của hai biến có phân phối x^2cũng có phân phối x^2 với số bậc tự do bằng tổng các bậc tự do.

Phân phối Student t

Định lý: Nếu $Z \sim N(0,1)$ và χ_k^2 là độc lập thống kê thì

$$t_{(k)} = \frac{Z}{\sqrt{\chi_k^2 / k}}$$

tuân theo phân phối Student hay nói gọn là phân phối t với k bậc tự do.

Tính chất của phân phối t

Phân phối t cũng đối xứng quanh 0 như phân phối chuẩn hoá nhưng thấp hơn. Khi bậc tự do càng lớn thì phân phối t tiệm cận đến phân phối chuẩn hoá. Trong thực hành. Khi bậc tự do lớn hơn 30 người ta thay phân phối t bằng phân phối chuẩn hoá.

$\mu = 0$ và $\sigma = k/(k-2)$

Phân phối F

Định lý : Nếu χ_{k1}^2 và χ_{k2}^2 là độc lập thống kê thì

$$F_{(k1,k2)} = \frac{\chi_{k1}^2 / k_1}{\chi_{k2}^2 / k_2}$$

tuân theo phân phối F với (k_1,k_2) bậc tự do.

Tính chất của phân phối F

Phân phối F lệch về bên trái, khi bậc tự do k_1 và k_2 đủ lớn, phân phối F tiến đến phân phối chuẩn.

$\mu = k_2/(k_2-2)$ với điều kiện $k_2>2$ và

$$\sigma^2 = \frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}$$

với điều kiện $k_2>4$.

Bình phương của một phân phối t với k bậc tự do là một phân phối F với 1 và k bậc tự do $t_k^2 = F_{(1,k)}$

Nếu bậc tự do mẫu k_2 khá lớn thì

$$k_1 F_{(k_1,k_2)} = \chi_{k_1}^2$$

.

Lưu ý : Khi bậc tự do đủ lớn thì các phân phối x^2, phân phối t và phân phối F tiến đến phân phối chuẩn. Các phân phối này được gọi là phân phối có liên quan đến phân phối chuẩn