Chương 1: CƠ BẢN VỀ THỐNG KÊ BAYES

1.1. Xác xuất có điều kiện và công thức Bayes

a. Định nghĩa xác suất

Xác suất là một hàm số $f: \Omega \to [0,1]$.

Khi đó, xác suất của một biến cố A được xác định là $P(A) \geq 0$.

Một số tính chất xác suất

i. Xác suất của không gian mẫu $P(\Omega) = 1$

ii. Nếu A và B là hai biến cố xung khắc với nhau. Khi đó

$$P(A \cup B) = P(A) + P(B)$$

iii. Xác suất của biến cố rỗng $P(\emptyset) = 0$

iv. Xác suất phần bù $P(A^c) = 1 - P(A)$

v. Công thức cộng xác suất $P(A + B) = P(A) + P(B) - P(AB)$

Code R

```
mu=0;
sig=1;
n=100;
data <- rnorm(n, mu, sig);
data;
summary(data)
```

Các đặc trưng thống kê của dữ liệu

```
summary(data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.1438 -0.6153 -0.1221  0.0198  0.7730  2.3592
```

Xác suất của các giá trị $x \geq 0$

b. Xác suất có điều kiện

Nếu A là một biến cố bất kỳ.

Khi đó, có thể biểu diễn A qua các công thức sau:

i. $A = (A \cap B) \cup (A \cap B^c)$

ii. $P(A) = P(A \cap B) + P(A \cap B^c)$

Công thức xác suất có điều kiện

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Công thức xác suất có điều kiện cho các biến cố độc lập

$$P(B|A) = P(B),$$

Nếu A và B là hai biến cố độc lập.

c. Định lý Bayes

Xuất phát từ công thức xác suất có điều kiện

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)}$$

Thay các biến cố giao bằng tích các biến cố theo công thức xác suất có điều kiện, ta có định lý Bayes cho biến cố rời rạc

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A|B) \times P(B) + P(A|B^c) \times P(B^c)}$$

Ta nhận thấy, định lý Bayes trong trường hợp hai biến cố B và B^c là hai biến cố liên hợp với nhau, hay nói cách khác là hai biến cố xung khắc. Mở rộng định lý Bayes cho trường hợp tập hợp các biến cố đôi một xung khắc với nhau trong định lý tổng quát sau:

Giả sử n biến cố $B_1, \cdots, B_n$ là các biến cố:

Ví dụ 1. Hợp của các biến cố chính là không gian mẫu

$$B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$$

Ví dụ 2. Các biến cố đôi một xung khắc với nhau $B_i \cap B_j = \emptyset, i = \overline{1,n}, j = \overline{1,n}, i \neq j$.

Khi đó, xác suất có điều kiện được xác định theo công thức

$$P(B_i|A) = \frac{P(A|B_i) \times P(B_i)}{\sum_{j=1}^{n} P(A|B_j) \times P(B_j)}.$$

1.2. Thống kê Bayes

Định lý Bayes được diễn đạt theo ngôn ngữ thống kê Bayes, trong đó xác suất tiên nghiệm (the prior probability) $P(B_i)$, xác suất hợp lý (the likelihood probability) $P(A|B_i)$, xác suất hậu nghiệm (the posterior probability) $P(B_i|A)$.

Định lý Bayes được viết lại dưới dạng tỷ lệ

$$posterior \propto prior \times likelihood.$$

1.3. Suy diễn thống kê Bayes cho biến ngẫu nhiên rời rạc

Bảng phân phối xác suất đồng thời của hàm hợp lý cho biến ngẫu nhiên rời rạc:

y_j / x_i	y_1	y_2	$\cdots$	y_n
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	$\cdots$	$f(x_1, y_n)$
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	$\cdots$	$f(x_2, y_n)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_m	$f(x_m, y_1)$	$f(x_m, y_2)$	$\cdots$	$f(x_m, y_n)$

Giả sử cần tính phân phối hậu nghiệm

$$f(x_i|y_j) = \frac{g(x_i) \times f(y_j|x_i)}{\sum_{i=1}^{m} g(x_i) \times f(y_j|x_i)},$$

Trong đó $g(x_i), i = 1, 2, \cdots, m$ là hàm phân phối tiên nghiệm.

Ví dụ 3. Một hộp chứa 5 viên bi gồm hai màu đỏ và xanh, hoàn toàn không biết số bi đỏ có trong hộp.

Gọi biến ngẫu nhiên X là số bi đỏ có trong hộp. Các giá trị có thể có của X là $x_i = i$ với $i = \{0, 1, 2, 3, 4, 5\}$.

Giả sử xác suất về số bi đỏ trong tổng số bi là bằng nhau, tức là $g(0) = g(1) = \ldots = g(5) = \frac{1}{6}$

Lấy ngẫu nhiên 1 viên bi từ hộp.

Gọi biến ngẫu nhiên Y với $Y = 1$ nếu bi lấy ra là bi đỏ; $Y = 0$ nếu bi được lấy ra là bi xanh. Khi đó, ta có $P(Y = 1 | X = x_i) = \frac{i}{5}$ và $P(Y = 0 | X = x_i) = \frac{5-i}{5}$.

Bảng phân phối xác suất đồng thời của hai biến ngẫu nhiên rời rạc X và Y được cho như sau:

x_i	Xác suất tiên nghiệm $g(x_i)$	$y_1 = 0$	$y_2 = 1$
0	1/6	$\frac{1}{6} \times \frac{5-0}{5} = \frac{5}{30}$	$\frac{1}{6} \times \frac{0}{5} = 0$
1	1/6	$\frac{1}{6} \times \frac{5-1}{5} = \frac{4}{30}$	$\frac{1}{6} \times \frac{1}{5} = \frac{1}{30}$

2	1/6	$\frac{1}{6}\times\frac{5-2}{5}=\frac{3}{30}$	$\frac{1}{6}\times\frac{2}{5}=\frac{2}{30}$
3	1/6	$\frac{1}{6}\times\frac{5-3}{5}=\frac{2}{30}$	$\frac{1}{6}\times\frac{3}{5}=\frac{3}{30}$
4	1/6	$\frac{1}{6}\times\frac{5-4}{5}=\frac{1}{30}$	$\frac{1}{6}\times\frac{4}{5}=\frac{4}{30}$
5	1/6	$\frac{1}{6}\times\frac{5-5}{5}=\frac{0}{30}$	$\frac{1}{6}\times\frac{5}{5}=\frac{5}{30}$
$f(y_j)$		$f(y_1=0)=\frac{15}{30}=\frac{1}{2}$	$f(y_1=1)=\frac{15}{30}=\frac{1}{2}$

Khi đó, xác suất hậu nghiệm của X|Y = 1 là:

x_i	Xác suất tiên nghiệm $g(x_i)$	Hàm hợp lý $f(Y=1\|X=x_i)$	$g(x_i)\times f(Y=1\|X=x_i)$	Xác suất hậu nghiệm
0	$\frac{1}{6}$	$\frac{0}{5}$	$\frac{1}{6}\times\frac{0}{5}=0$	0
1	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{6}\times\frac{1}{5}=\frac{1}{30}$	$\frac{1}{30}/\frac{1}{2}=\frac{1}{15}$
2	$\frac{1}{6}$	$\frac{2}{5}$	$\frac{1}{6}\times\frac{2}{5}=\frac{2}{30}$	$\frac{2}{30}/\frac{1}{2}=\frac{2}{15}$
3	$\frac{1}{6}$	$\frac{3}{5}$	$\frac{1}{6}\times\frac{3}{5}=\frac{3}{30}$	$\frac{3}{30}/\frac{1}{2}=\frac{3}{15}$
4	$\frac{1}{6}$	$\frac{4}{5}$	$\frac{1}{6}\times\frac{4}{5}=\frac{4}{30}$	$\frac{4}{30}/\frac{1}{2}=\frac{4}{15}$
5	$\frac{1}{6}$	$\frac{5}{5}$	$\frac{1}{6}\times\frac{5}{5}=\frac{5}{30}$	$\frac{5}{30}/\frac{1}{2}=\frac{5}{15}$
$f(y_j)$			$\frac{15}{30}=\frac{1}{2}$	

Thông thường, khi sử dụng định lý Bayes, ta có thể tiến hành theo hai cách sau đây:

Cách 1. ***Phân tích mỗi lần một quan sát theo một trình tự nhất định***.

Giả sử, sau khi lấy (không hoàn lại) một viên bi ở lần thứ nhất thì được bi đỏ như ở ví dụ trên, chúng ta tiếp tục lấy ngẫu nhiên một viên bi thứ hai ra từ hộp. Giả sử bi được lấy ra ở lần thứ hai là bi xanh, nghĩa là $Y=0$. Để tính xác suất hậu nghiệm của X từ kết quả của hai lần quan sát, lần đầu là màu đỏ, lần thứ hai là màu xanh, chúng ta sẽ phân tích các quan sát theo trình tự bằng cách sử dụng định lý Bayes cho mỗi lần. Chúng ta sẽ sử dụng xác suất tiên nghiệm cho lần thử đầu tiên như trong ví dụ trước. Tuy nhiên, sau đó chúng ta sẽ sử dụng xác suất hậu nghiệm của lần thử đầu tiên làm xác suất hậu nghiệm cho lần thử thứ hai. Các kết quả được thể hiện trong Bảng sau:

x_i	Xác suất tiên nghiệm	Hàm hợp lý $f(Y=1\|X=x_i)$	$g(x_i) \times f(Y=1\|X=x_i)$	Xác suất hậu nghiệm
0	0		0	0
1	$\frac{1}{15}$	$\frac{4}{4}$	$\frac{1}{15}$	$\frac{1}{15}/\frac{1}{3}=\frac{1}{5}$
2	$\frac{2}{15}$	$\frac{3}{4}$	$\frac{1}{10}$	$\frac{1}{10}/\frac{1}{3}=\frac{3}{10}$
3	$\frac{3}{15}$	$\frac{2}{4}$	$\frac{1}{10}$	$\frac{1}{10}/\frac{1}{3}=\frac{3}{10}$
4	$\frac{4}{15}$	$\frac{1}{4}$	$\frac{1}{15}$	$\frac{1}{15}/\frac{1}{3}=\frac{1}{5}$
5	$\frac{5}{15}$	0	0	$0/\frac{1}{3}=0$
$f(y_j)$			$\frac{1}{3}$	1

Cách 2. Phân tích cùng lúc tất cả các quan sát.

Xét lại ví dụ 1, giả sử lấy ngẫu nhiên hai viên bi, mỗi lần một bi, không hoàn lại từ hộp. Để tính xác suất hậu nghiệm bằng định lý Bayes, hoàn toàn tương tự như trong phân tích như ở ví dụ 1, ta vẫn xem như xác suất tiên nghiệm về số bi đỏ có trong hộp là $g(0)=g(1)=\ldots=g(5)=\frac{1}{6}$.

Đặt Y_1, Y_2 lần lượt là kết quả lấy bi từ hộp ở lần thứ nhất và lần thứ hai. Áp dụng công thức nhân xác suất, xác suất của các quan sát với điều kiện X được cho bởi công thức:

$$f(y_1, y_2 \mid X) = f(y_1 \mid x).f(y_2 \mid y_1, x)$$

Khi đó, xác suất đồng thời của X, Y_1 và Y_2 được cho trong bảng sau:

x_i	Xác suất tiên nghiệm $g(x_i)$	$(y_1,y_2)=(0,0)$	$(y_1,y_2)=(0,1)$	$(y_1,y_2)=(1,0)$	$(y_1,y_2)=(1,1)$

0	1/6	$\frac{1}{6}\times\frac{5}{5}\times\frac{4}{4}=\frac{1}{6}$	$\frac{1}{6}\times\frac{5}{5}\times\frac{0}{4}=0$	$\frac{1}{6}\times\frac{0}{5}\times\frac{4}{4}=0$	$\frac{1}{6}\times\frac{0}{5}\times\frac{0}{4}=0$
1	1/6	$\frac{1}{6}\times\frac{4}{5}\times\frac{3}{4}=\frac{1}{10}$	$\frac{1}{6}\times\frac{4}{5}\times\frac{1}{4}=\frac{1}{30}$	$\frac{1}{6}\times\frac{1}{5}\times\frac{4}{4}=\frac{1}{30}$	$\frac{1}{6}\times\frac{1}{5}\times\frac{0}{4}=0$
2	1/6	$\frac{1}{6}\times\frac{3}{5}\times\frac{2}{4}=\frac{1}{20}$	$\frac{1}{6}\times\frac{3}{5}\times\frac{2}{4}=\frac{1}{20}$	$\frac{1}{6}\times\frac{2}{5}\times\frac{3}{4}=\frac{1}{20}$	$\frac{1}{6}\times\frac{2}{5}\times\frac{1}{4}=\frac{1}{60}$
3	1/6	$\frac{1}{6}\times\frac{2}{5}\times\frac{1}{4}=\frac{1}{60}$	$\frac{1}{6}\times\frac{2}{5}\times\frac{3}{4}=\frac{1}{20}$	$\frac{1}{6}\times\frac{3}{5}\times\frac{2}{4}=\frac{1}{20}$	$\frac{1}{6}\times\frac{3}{5}\times\frac{2}{4}=\frac{1}{20}$
4	1/6	$\frac{1}{6}\times\frac{1}{5}\times\frac{0}{4}=0$	$\frac{1}{6}\times\frac{1}{5}\times\frac{4}{4}=\frac{1}{30}$	$\frac{1}{6}\times\frac{4}{5}\times\frac{1}{4}=\frac{1}{30}$	$\frac{1}{6}\times\frac{4}{5}\times\frac{3}{4}=\frac{1}{10}$
5	1/6	$\frac{1}{6}\times\frac{0}{5}\times 0=0$	$\frac{1}{6}\times\frac{0}{5}\times\frac{4}{4}=0$	$\frac{1}{6}\times\frac{5}{5}\times\frac{0}{4}=0$	$\frac{1}{6}\times\frac{5}{5}\times\frac{4}{4}=\frac{1}{6}$
$f(y_j)$		$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$

Và xác suất hậu nghiệm của X khi $Y_1 = 1, Y_2 = 0$ là:

x_i	Xác suất tiên nghiệm $g(x_i)$	$(y_1, y_2) = (1,0)$	Xác suất hậu nghiệm
0	1/6	$\frac{1}{6}\times\frac{0}{5}\times\frac{4}{4}=0$	$0/\frac{1}{6}=0$
1	1/6	$\frac{1}{6}\times\frac{1}{5}\times\frac{4}{4}=\frac{1}{30}$	$\frac{1}{30}/\frac{1}{6}=\frac{1}{5}$
2	1/6	$\frac{1}{6}\times\frac{2}{5}\times\frac{3}{4}=\frac{1}{20}$	$\frac{1}{20}/\frac{1}{6}=\frac{3}{10}$
3	1/6	$\frac{1}{6}\times\frac{3}{5}\times\frac{2}{4}=\frac{1}{20}$	$\frac{1}{20}/\frac{1}{6}=\frac{3}{10}$
4	1/6	$\frac{1}{6}\times\frac{4}{5}\times\frac{1}{4}=\frac{1}{30}$	$\frac{1}{30}/\frac{1}{6}=\frac{1}{5}$
5	1/6	$\frac{1}{6}\times\frac{5}{5}\times\frac{0}{4}=0$	$0/\frac{1}{6}=0$
$f(y_j)$		$\frac{1}{6}$	1

1.3.1 Suy diễn Bayes cho phân phối nhị thức với tiên nghiệm rời rạc

Giả sử có n phép thử độc lập, mỗi phép thử chỉ có hai khả năng là "thành công" hoặc "thất bại". Gọi p là xác suất "thành công" và xác suất này không đổi ở mỗi lần thử. Gọi Y là tổng số "thành công" trong n lần thử. Khi đó $Y \sim B(n,p)$. Giả sử có tập hợp rời rạc I các giá trị có thể có $p_1, p_2, \dots, p_I$.

Lập bảng phân phối cho các quan sát. Trong đó, hàng thứ I tương ứng với phân phối xác suất nhị thức (n, p_i). Cột thứ j tương ứng với các giá trị $Y = j, j = \overline{0, n}$.

Trước tiên, ta cần xác định phối xác suất tiên nghiệm của tham số. Phân phối tiên nghiệm này phụ thuộc vào niềm tin trước đó của chúng ta về từng giá trị có thể có của tham số p. Nếu chúng tôi không có ý tưởng nào trước đó về tham số, chúng ta có thể chọn phân phối tiên nghiệm là tất cả các giá trị có đồng khả năng.

Phân phối xác suất đồng thời của tham số p và biến ngẫu nhiên Y được tính bằng cách nhân xác suất có điều kiện là $P(Y|p)$ với xác suất tiên nghiệm p.

Ví dụ 4. Xét biến ngẫu nhiên Y có tập các giá trị có thể có là {0,1,2,3,4} và các xác suất p lần lượt là 0.4; 0.5 và 0.6. Khi đó, bảng phân phối xác suất đồng thời được cho như sau:

p	Tiên nghiệm	0	1	2	3	4
0.4	$\frac{1}{3}$	$\frac{1}{3} \times 0.1296$	$\frac{1}{3} \times 0.3456$	$\frac{1}{3} \times 0.3456$	$\frac{1}{3} \times 0.1536$	$\frac{1}{3} \times 0.0256$
0.5	$\frac{1}{3}$	$\frac{1}{3} \times 0.0625$	$\frac{1}{3} \times 0.2500$	$\frac{1}{3} \times 0.3750$	$\frac{1}{3} \times 0.2500$	$\frac{1}{3} \times 0.0625$
0.6	$\frac{1}{3}$	$\frac{1}{3} \times 0.0256$	$\frac{1}{3} \times 0.1536$	$\frac{1}{3} \times 0.3456$	$\frac{1}{3} \times 0.3456$	$\frac{1}{3} \times 0.1296$
Tổng		0.0725	0.2497	0.3554	0.2497	0.0725

Khi đó phân phối hậu nghiệm của Y = 3 được cho trong bảng sau:

p	Tiên nghiệm	Hàm hợp lý	Tiên nghiệm×Hàm hợp lý	Hậu nghiệm
0.4	$\frac{1}{3}$	0.1536	0.0512	$\frac{0.0512}{0.2497} = 0.205$
0.5	$\frac{1}{3}$	0.2500	0.0833	$\frac{0.0833}{0.2497} = 0.334$
0.6	$\frac{1}{3}$	0.3456	0.1152	$\frac{0.1153}{0.2497} = 0.461$
P(Y = 3)			0.2497	1.000

Suy diễn Bayes cho phân phối nhị thức với tiên nghiệm rời rạc bằng phần mềm R
install.packages("LearnBayes") *library(LearnBayes)* *p = c(0.4, 0.5,0.6)* *prior = c(1, 1, 1)* *prior = prior/sum(prior)* *data = c(3, 1)* *post = pdisc(p, prior, data)* *round(cbind(p, prior, post),3)*

Kết quả thu được như sau:

```
      p  prior post
[1,] 0.4 0.333 0.205
[2,] 0.5 0.333 0.334
[3,] 0.6 0.333 0.461
```

1.3.2 Một số kết quả quan trọng của suy diễn Bayes

i. Nhân tất cả các xác suất tiên nghiệm với một hằng số không làm thay đổi kết quả của suy diễn Bayes.

ii. Nhân tất cả các hàm hợp lý với một hằng số cũng không làm thay đổi kết quả của suy diễn Bayes.

1.3.3 Suy diễn Bayes cho phân phối Poission với tiên nghiệm rời rạc

Khi quan sát có từ phân phối Poission P(μ) và cho trước một phân phối tiên nghiệm rời rạc là một vài giá trị có thể có của μ. Khi đó tương tự phần 1.3.1 chúng ta có thể lập bảng suy diễn Bayes cho phân phối Poission với tiên nghiệm rời rạc.

Ví dụ 3. Giả sử Y là một biến ngẫu nhiên theo phân phối Poission với tham số μ lần lượt là 1; 1.5; 2 và 2.5. Khi đó suy diễn Bayes cho Y = 2 với các xác suất tiên nghiệm lần lượt là $\frac{1}{6}$; $\frac{1}{3}$; $\frac{1}{3}$ và $\frac{1}{6}$ được cho trong bảng sau:

μ	Tiên nghiệm	Hàm hợp lý	Tiên nghiệm× Hàm hợp lý	Hậu nghiệm
1	$\frac{1}{6}$	$\frac{1^2 \times e^{-1}}{2!} = 0.1839$	0.0307	$\frac{0.0307}{0.2473}$
1.5	$\frac{1}{3}$	$\frac{1.5^2 \times e^{-1.5}}{2!} = 0.2510$	0.0837	$\frac{0.0837}{0.2473}$

2	$\frac{1}{3}$	$\frac{2^2 \times e^{-2}}{2!} = 0.2707$	0.0902	$\frac{0.0902}{0.2473}$
2.5	$\frac{1}{6}$	$\frac{2.5^2 \times e^{-2.5}}{2!} = 0.2565$	0.0428	$\frac{0.0428}{0.2473}$
P(Y=2)			0.2473	1.00

Suy diễn Bayes cho phân phối Poission với tiên nghiệm rời rạc bằng phần mềm R

```
mu = seq(1, 2.5, by = 0.5)
prior = c(1, 2, 2, 1)
prior = prior/sum(prior)
names(prior) = seq(1, 2.5, by = 0.5)
y=2
post = discrete.bayes(dpois, prior, y)
print(post)
```

Kết quả thu được như sau:

```
        1       1.5         2       2.5
0.1239620 0.3383404 0.3648246 0.1728729
```

Chú ý:

i. Bảng phân phối đồng thời trong suy diễn Bayes có hai chiều. Chiều dọc là các giá trị có thể có của tham số (không quan sát được). Chiều ngang là không gian mẫu và có thể quan sát giá trị nào xảy ra.

ii. Đối với biến ngẫu nhiên rời rạc với tiên nghiệm rời rạc, xác suất hậu nghiệm được tính bằng cách nhân xác suất tiên nghiệm với Hàm hợp lý và sau đó chia cho tổng của chúng.

iii. Khi dữ liệu có được theo một trình tự nào đó, chúng ta có thể sử dụng xác suất hậu nghiệm từ lần thứ nhất làm xác suất tiên nghiệm cho cho lần thứ hai.

1.4. Suy diễn thống kê Bayes cho biến ngẫu nhiên liên tục

1.4.1. Một vài phân phối liên tục

1.4.1.1. Phân phối đều

Một biến ngẫu nhiên X có phân phối đều (0,1) nếu hàm mật độ xác suất của nó có dạng:

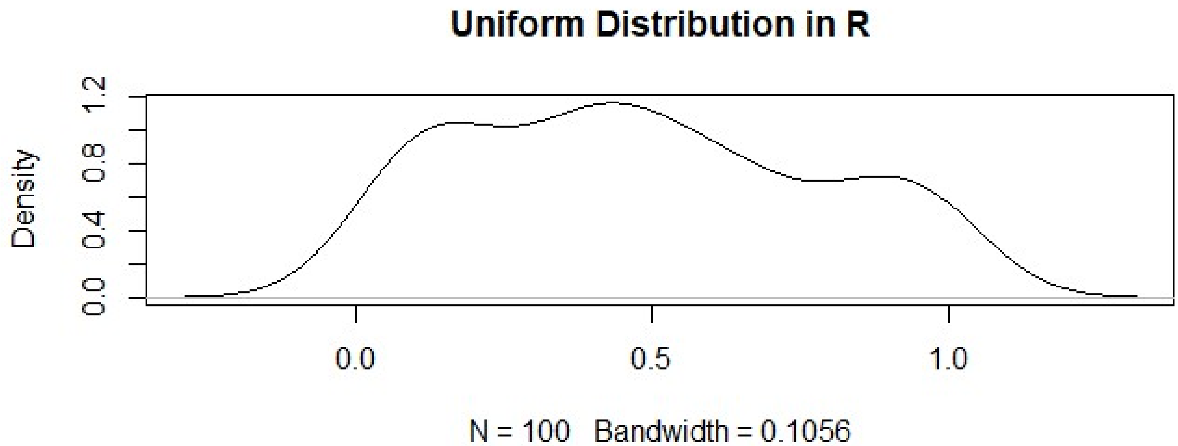
$$f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & x \notin [0,1] \end{cases}$$

Khi đó: $E(X) = \frac{1}{2}$ và $Var(X) = \frac{1}{12}$.

Mô phỏng phân phối đều

Code in R

```
data <- runif(100, min = 0, max = 1)
plot(density(data),main = "Uniform Distribution in R")
data
```




```
 [1] 0.10541237 0.75494944 0.27005219 0.96949668 0.76954071 0.07845489 0.29795688 0.15021908
 [9] 0.11175915 0.79947200 0.85913239 0.17448560 0.49956455 0.97020312 0.03200649 0.62996126
[17] 0.64864420 0.98527075 0.43669407 0.34039646 0.97781185 0.13661745 0.89893975 0.61197432
[25] 0.15080690 0.96182035 0.89696830 0.46110391 0.29561423 0.32223562 0.98850185 0.46075008
[33] 0.98379274 0.38797069 0.17362501 0.95930106 0.62236100 0.44563926 0.59721810 0.08054957
[41] 0.65923970 0.44837868 0.64820376 0.82030659 0.38940539 0.55158350 0.08597048 0.11858547
[49] 0.39886320 0.28077895 0.80926276 0.70079955 0.03672248 0.08339242 0.51995049 0.25953397
[57] 0.05996729 0.89469606 0.34869324 0.41553655 0.40445127 0.83609473 0.52393114 0.93899815
[65] 0.38518475 0.06608752 0.15131397 0.23462763 0.02846413 0.28337280 0.63550938 0.44512682
[73] 0.38352594 0.48599879 0.37150057 0.69539804 0.15674709 0.47470471 0.16723122 0.99993421
[81] 0.42649039 0.56373003 0.69239555 0.17791697 0.87095440 0.09436971 0.17366739 0.43697644
[89] 0.97239576 0.23656809 0.56035430 0.08714742 0.44388816 0.37384434 0.59523926 0.56333253
[97] 0.13400741 0.52140541 0.66856761 0.86668796
```

1.4.1.2. Phân phối của họ Bêta

Một biến ngẫu nhiên X có phân phối bêta (a,b) nếu hàm mật độ xác suất của nó có dạng:

$$f(x;a,b) = \begin{cases} k.x^{a-1}.(1-x)^{b-1}, & 0 \le x \le 1 \\ 0, & x \notin [0,1] \end{cases}$$

với $k = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ là hằng số sao cho $f(x,a,b)$ là hàm mật độ, $\Gamma(c)$ là một hàm Gamma.

Khi đó: $E(X) = \frac{a}{a+b}$; $Var(X) = \dfrac{ab}{(a+b)^2(a+b+1)}$ và $P(X \le x_0) = \int_0^{x_0} f(x;a,b)dx$.

Mô phỏng phân phối beta(1,5) trong R

Code R

```
y_rbeta <- rbeta(100, shape1 = 1, shape2 = 5)       # Draw N beta distributed values
y_rbeta                                  # Print values to RStudio console
plot(density(y_rbeta),                   # Plot of randomly drawn beta density
   main = "beta Distribution in R")
```

Các giá trị được mô phỏng từ phân phối beta

```
y_rbeta                                                    # Print values to RStudio console
 [1] 0.055243013 0.040470605 0.119498277 0.058878526 0.533222000 0.059594478 0.140433487
 [8] 0.062757443 0.007262550 0.115262498 0.052490015 0.066503661 0.019556051 0.062670384
[15] 0.015917083 0.070106359 0.106814302 0.031879448 0.123875379 0.143893192 0.124320322
[22] 0.006353812 0.152671970 0.160250878 0.265879302 0.214298219 0.079808722 0.049720739
[29] 0.009112857 0.038897511 0.063668297 0.168245259 0.151852671 0.119086455 0.216608264
[36] 0.395799947 0.208687229 0.177435192 0.265112935 0.707088495 0.195199927 0.151663940
[43] 0.167441641 0.502383368 0.312398823 0.020658080 0.204054179 0.101525373 0.038282880
[50] 0.268380355 0.095380989 0.195761497 0.060374540 0.053916948 0.129367101 0.391453264
[57] 0.056216306 0.245856477 0.218679072 0.124350791 0.256956299 0.858701795 0.107751502
[64] 0.281679963 0.270690940 0.049644980 0.063844979 0.132744882 0.024601948 0.029611728
[71] 0.083123016 0.193506544 0.132329741 0.403656053 0.114828335 0.449273994 0.126518159
[78] 0.375915347 0.041397810 0.628954089 0.126085320 0.374462042 0.047885842 0.009278074
[85] 0.120011945 0.021640172 0.026652730 0.036386348 0.258647592 0.110679605 0.063365007
[92] 0.330936869 0.012189474 0.432637508 0.022127216 0.134817482 0.399778588 0.219674490
[99] 0.073050728 0.020108191
```

1.4.1.3. Phân phối của họ Gamma

Một biến ngẫu nhiên X có phân phối Gamma (r,v) nếu hàm mật độ xác suất của nó có dạng:

$$f(x;r,v) = k.x^{r-1}.e^{-v.x}\ ;\ 0 \le x < \infty$$

với $k = \frac{v^r}{\Gamma(r)}$ là hằng số sao cho $f(x;r,v)$ là một hàm mật độ.

Khi đó: $E(X) = \frac{r}{v}$; $Var(X) = \frac{r}{v^2}$ và $P(X \le x_0) = \int_0^{x_0} f(x;r,v)\,dx$.

1.4.1.4. Phân phối chuẩn

Một biến ngẫu nhiên X có phân phối chuẩn (μ,σ^2) nếu hàm mật độ xác suất của nó có dạng:

$$f(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}.e^{\frac{-(x-\mu)^2}{2\sigma^2}}\ ; -\infty < x < \infty$$

Khi đó: $E(X) = \mu$ và $Var(X) = \sigma^2$.

Mô phỏng 1000 quan sát từ phân phối chuẩn tắc

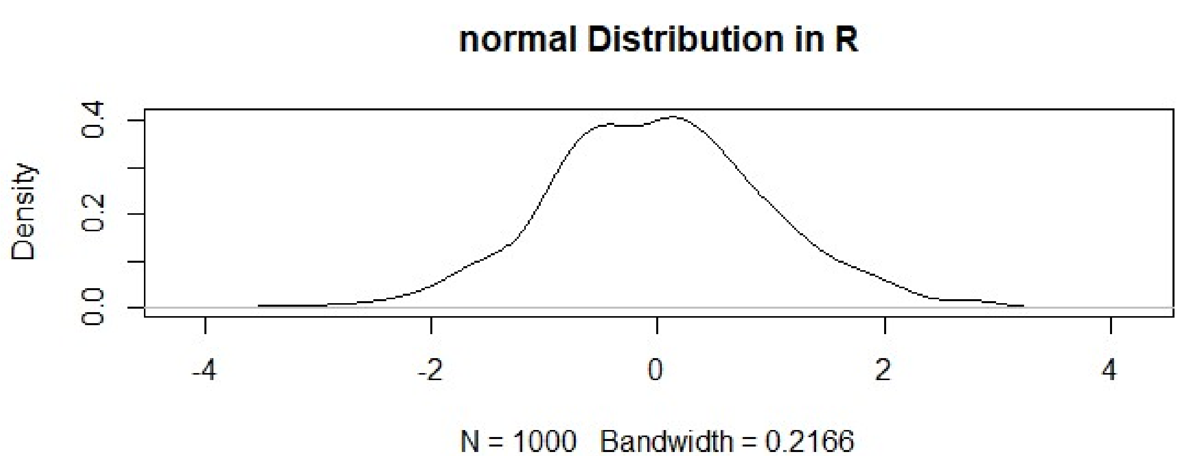
Code in R

```
y_rnorm <- rnorm(1000, 0, 1)        # Draw N normal distributed values
y_rnorm                              # Print values to RStudio console
plot(density(y_rnorm),               # Plot of randomly drawn beta density
   main = "normal Distribution in R")
```

Một phần các giá trị được mô phỏng từ phân phối chuẩn tắc

```
[865] -0.336115569  0.173692855 -0.854500307  2.774822489 -0.623461042 -0.311625689
[871] -3.284167450 -0.168185733 -0.712821511  0.490692188  0.264539994  1.381012739
[877] -0.517212075 -1.526888102 -1.134499041 -0.838459755 -0.461189138  0.703245900
[883] -0.440838291  0.763702623 -0.685491837 -0.143087626 -0.581191299  0.681654995
[889]  0.380357876  0.728690617 -0.329390452 -2.202826127  1.188779124 -0.807332175
[895]  1.097691382  0.323965574 -0.693934157  0.260466532 -2.355786979  0.326785539
[901] -0.863800005  0.802835042 -1.031857954  0.279553316  0.342820246 -0.906647102
[907]  1.178076553 -0.811015366  0.852109301 -0.648530563  0.449524283 -0.751178577
[913] -0.358830941 -1.226686423 -0.481360568  1.442829883 -0.937465754 -0.636266744
[919]  0.558239755  0.744963544  0.482366417 -0.586237175 -0.880050402 -0.246723970
[925]  0.554913366 -1.532870281  1.702848679 -0.391315367  0.834057879  0.831437609
[931]  0.078682500 -0.408910868 -0.502115162 -0.908072223 -0.130176336  0.387783395
[937] -1.346114973  1.925777644 -0.919115781 -0.474737911  0.040858268  0.492408664
[943] -0.071772917  0.125503158  0.241899061  0.291248346 -0.035107856  0.027613113
[949]  0.197517178  0.779739463 -0.661922349  0.544155642  0.860208689 -1.053173104
[955] -0.811474722 -1.535251198 -0.264752916 -0.943689258 -0.668069881 -1.757250365
[961] -0.775005369 -0.319802006  1.130957258 -0.440249408  2.083365698  0.368079000
[967] -0.159814459 -0.729061976  0.597235386 -1.990872018  0.332228090  2.847261536
[973]  0.316211281  0.029067420  1.857988835  1.169574576  0.263918005 -0.025992505
[979] -0.254060573 -0.410153108 -0.935457470 -0.111303779 -0.771345039  0.392484992
[985]  0.333867888 -0.513417493  1.154269646  0.196597281  0.323834283 -0.583583816
[991]  0.454404573  0.131551355 -0.195734482 -0.501606769 -1.073327278 -0.658242961
[997] -0.135778897 -0.064917413 -0.551986214  0.301933000
```

Đồ thị phân phối chuẩn tắc



1.4.2. Phân phối đồng thời của hai biến ngẫu nhiên liên tục

Xét hai biến ngẫu nhiên liên tục X và Y, gọi $f(x, y)$ là hàm mật độ đồng thời của chúng. Khi đó:

- Hàm mật độ biên của X: $f(x)=\int\limits_{-\infty}^{\infty} f(x,y)\,dy$
- Hàm mật độ biên của Y: $f(y)=\int\limits_{-\infty}^{\infty} f(x,y)\,dx$
- Hàm mật độ xác suất của X với điều kiện Y = y: $f(x\mid y)=\dfrac{f(x,y)}{f(y)}$
- Hàm mật độ xác suất của Y với điều kiện X = x: $f(y\mid x)=\dfrac{f(x,y)}{f(x)}$

1.4.3. Phân phối đồng thời của một biến ngẫu nhiên liên tục và một biến ngẫu nhiên rời rạc

Xét hai biến ngẫu nhiên gồm X là biến ngẫu nhiên liên tục và Y là biến ngẫu nhiên rời rạc, gọi $f(x,y_j)$ là hàm mật độ xác suất đồng thời của chúng. Khi đó:

- Hàm mật độ biên của biến ngẫu nhiên liên tục X: $f(x)=\sum\limits_{j} f(x,y_j)$,
- Hàm xác suất biên của biến ngẫu nhiên rời rạc Y: $f(y_j)=\int f(x,y_j)\,dx$
- Hàm mật độ của X với điều kiện Y = y_j: $f(x\mid y_j)=\dfrac{f(x,y_j)}{f(y_j)}=\dfrac{f(x,y_j)}{\int f(x,y_j)\,dx}$
- Hàm mật độ của Y với điều kiện X = x: $f(y_j\mid x)=\dfrac{f(x,y_j)}{f(x)}=\dfrac{f(x,y_j)}{\sum_j f(x,y_j)}$

1.4.4. Suy diễn Bayes cho tỉ lệ nhị thức

Gọi p là tỉ lệ tổng thể, Y là tổng số lần thành công trong n lần thử với xác suất thành công ở mỗi lần thử là bằng nhau và bằng p. Khi đó $Y \sim B(n,p)$ và hàm xác suất của Y với điều kiện p được cho như sau:

$$f(y\mid p)=C_n^y p^y (1-p)^{n-y}\;;\; y=0,1,2,\ldots,n$$

Trong công thức trên, chúng ta giữ nguyên p và quan sát phân phối xác suất của y qua tất cả các giá trị có thể có của nó.

Nếu chúng ta đổi vai trò của p và y, nghĩa là, chúng ta giữ nguyên y và cho p thay đổi theo tất cả các giá trị có thể có của nó, chúng ta có hàm hợp lý được cho như sau:

$$f(y\mid p)=C_n^y p^y (1-p)^{n-y}\;;\; 0\le p\le 1$$

Để sử dụng định lý Bayes, chúng ta cần một phân phối tiên nghiệm $g(p)$, là cho niềm tin của chúng ta về các giá trị có thể có của tham số p trước khi lấy dữ liệu. Như vậy hàm tiên nghiệm không được xây dựng từ dữ liệu. Khi đó, theo định lý Bayes, phân phối hậu nghiệm được cho như sau:

$$g(p \mid y) \propto g(p) \times f(y \mid p)$$

Để có phân phối hậu nghiệm thực tế, ta cần chia phân phối hậu nghiệm trên cho một hằng số k để nó trở thành một hàm mật độ xác suất. Người ta chứng minh được khi đó $k = \int_0^1 g(p) \times f(y \mid p)\,dp$ và hàm hàm phân phối hậu nghiệm được cho như sau:

$$g(p \mid y) = \frac{g(p) \times f(y \mid p)}{\int_0^1 g(p) \times f(y \mid p)\,dp}$$

1.4.4.1. Sử dụng tiên nghiệm là phân phối đều

Khi không có thông tin gì về phân phối tiên nghiệm của p, chúng ta có thể chọn tiên nghiệm là phân phối đều, nghĩa là tất cả các giá trị của p là đồng khả năng.

Khi đó, với phân phối tiên nghiệm $g(p) = 1;\ 0 \le p \le 1$, ta có phân phối hậu nghiệm được cho như sau:

$$g(p \mid y) = C_n^y p^y (1-p)^{n-y}\ ;\ 0 \le p \le 1$$

1.4.4.2. Sử dụng tiên nghiệm là phân phối Bêta

Giả sử mật độ tiên nghiệm được sử dụng cho p như sau:

$$g(p;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1};\quad 0 \le p \le 1$$

Phân phối hậu nghiệm bằng tiên nghiệm nhân với hàm hợp lý. Như đã nhận xét trong phần 1.3.2 , chúng ta có thể bỏ qua hằng số không phụ thuộc vào tham số trong tiên nghiệm và hàm hợp lý. Khi đó, ta có phân phối hậu nghiệm là một hàm của p, có dạng:

$$g(p \mid y) \propto p^{a+y-1}(1-p)^{b+n-y-1};\quad 0 \le p \le 1$$

Đây chính là phân phối beta với các tham số a’ = a + y và b’ = b + n - y. Nghĩa là, số lần thành công được thêm vào a và số lần thất bại được thêm vào b, như sau:

$$g(p \mid y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)} p^{y+a-1}(1-p)^{n-y+b-1};\ 0 \le p \le 1$$