

## Các phân phối xác suất đa biến

### Phân phối giao xác suất của hai biến ngẫu nhiên rời rạc

Cho  $X$  và  $Y$  là hai biến ngẫu nhiên rời rạc. Trong bài này, chúng ta xét các biến nói lên các tính chất khác nhau của một đơn vị quan sát cho trước. Như vậy, mỗi kết quả sẽ là một cặp  $X, Y$  tương ứng. Bảng liệt kê tất cả các kết hợp có thể có được của  $X = x_i$  và  $Y = y_j$  và các xác suất của chúng sẽ tạo nên phân phối giao xác suất (joint probability) của  $X$  và  $Y$ . Giao xác suất thường được viết dưới dạng:

$$P(X = x_i \cap Y = y_j) = P(X, Y)$$

Điều này được giải thích như sau: Nếu ta lấy một quan sát ngẫu nhiên từ tổng thể được mô tả bởi phân phối giao xác suất này, thì xác suất cho quan sát đó là:

$$X = x_i \text{ và } Y = y_j \text{ là } P(X = x_i \cap Y = y_j)$$

Hàm phân phối giao xác suất (joint pdf) là một quan hệ hàm số có dạng:

$$P(X = x_i \cap Y = y_j) = f(x_i, y_j)$$

nhưng để tiện cho việc giải thích, chúng ta sẽ dùng dạng bảng.

Hãy tưởng tượng là ta có hai biến ngẫu nhiên  $Q$  và  $H$  với các không gian mẫu như sau:

$Q_1$	=	5031,41
$Q_2$	=	8585,35
$Q_3$	=	11670,96
$Q_4$	=	16563,92
$Q_5$	=	34509,14

$H_1$	=	1,70
$H_2$	=	3,00
$H_3$	=	4,00
$H_4$	=	5,00
$H_5$	=	6,00
$H_6$	=	7,95

		<b>Q</b>					
		5031,41	8585,35	11670,35	16563,92	34509,14	P(H)
<b>H</b>	1,70	0,08	0,02	0,01	0,01	0,01	0,12
	3,00	0,03	0,03	0,02	0,02	0,02	0,12
	4,00	0,04	0,06	0,05	0,04	0,04	0,23
	5,00	0,03	0,05	0,05	0,05	0,05	0,22
	6,00	0,01	0,03	0,04	0,04	0,03	0,14
	7,95	0,01	0,02	0,03	0,05	0,05	0,16
	P(Q)	0,20	0,20	0,20	0,20	0,20	1,00

Các con số trong bảng trên cho ta xác suất của tổng thể liên quan đến các hộ gia đình có  $Q = q_i$  và  $H = h_j$  ứng với các giá trị cụ thể của  $Q$  và  $H$ .

Như vậy, xác suất của việc chọn ngẫu nhiên một hộ gia đình có  $Q = 5031,41$  và  $H = 1,70$  là 0,08. Nói cách khác là có 8% số hộ gia đình có  $Q = 5031,41$  và  $H = 1,70$ .

### Các phân phối xác suất biên

Bây giờ, giả sử là chúng ta chỉ quan tâm đến  $P(Q = q_i)$  mà không tính đến kết quả của  $H$ . Xác suất này được gọi là xác suất biên  $Q = q_i$  ... chúng ta có được nó từ phân phối giao xác suất như sau:

$$P(Q = q_1) = P(Q = q_1 \cap H = h_1) + P(Q = q_1 \cap H = h_2) + \dots + P(Q = q_1 \cap H = h_6)$$

ở dạng tổng quát:

$$P(Q = q_i) = \sum_H P(Q = q_i \cap H = h_j)$$

và

$$P(H = h_j) = \sum_Q P(Q = q_i \cap H = h_j)$$

Các phân phối xác suất biên được viết ở các lề bên ngoài của bảng phân phối giao xác suất (như ở bảng trên).

### Các phân phối xác suất có điều kiện

Có lúc chúng ta chỉ muốn biết phân phối của một biến khi có thông tin về kết quả của biến khác (biến điều kiện); ngoài ra, chúng ta có thể muốn biết phân phối của một biến sẽ thay đổi ra sao khi kết quả của biến khác thay đổi.

Cách giải quyết về vấn đề nêu trên thực chất là việc chúng ta biến không gian mẫu thành một kết quả đơn nhất từ toàn bộ dãy giá trị của biến điều kiện.

Công thức làm điều này là:

$$P(Q = q_i | H = h_1) = \frac{P(Q = q_i \cap H = h_1)}{P(H = h_1)}$$

Xét các phân phối có điều kiện sau đây được tính từ phân phối giao xác suất đã cho ở trên:

$q_i$	$P(Q = Q_i   H = h_1)$	$P(Q = q_i   H = h_6)$
$q_1$	0.64	0.08
$q_2$	0.16	0.15
$q_3$	0.08	0.28
$q_4$	0.06	0.40
$q_5$	0.06	0.45

Rõ ràng chúng ta có thể thấy rằng thông tin về kết quả của H cho chúng ta biết thông tin về phân phối của Q.

### Tính chất độc lập về thống kê

Điều gì sẽ xảy ra nếu kết quả H không mang đến thông tin về phân phối Q? Khi đó ta có:

$$P(Q = q_i | H = h_j) = P(Q = q_i) = \frac{P(Q = q_i \cap H = h_j)}{P(H = h_j)}$$

$$\Rightarrow P(Q = q_i) \times P(H = h_j) = P(Q = q_i \cap H = h_j) \text{ cho mọi } i \text{ và } j.$$

Đây chính là định nghĩa của chúng ta về tính độc lập thống kê. Nó giúp ta dễ tính giao xác suất về các kết quả của các mẫu ngẫu nhiên.

### Kỳ vọng toán học

Nếu ta có một phân phối giao xác suất, và chúng muốn tính kỳ vọng của một trong các biến số đó, thì chúng ta có thể làm điều đó từ phân phối biên:

$$E[Q] = \sum_Q q_i P(Q = q_i) = \sum_Q \sum_H q_i P(Q = q_i \cap H = h_j) = 15272,16$$

Tương tự như vậy:

$$E[H] = \sum_H h_j P(H = h_j) = \sum_H \sum_Q h_j P(Q = q_i \cap H = h_j) = 4,70$$

Chúng ta cũng có thể tính được các kỳ vọng có điều kiện:

$$E[Q | H = h_j] = \sum_Q q_i P(Q = q_i | H = h_j)$$

Từ phân phối giao xác suất trong ví dụ, ta có:

$$E[Q|H=1.70] = 8585.39$$

$$E[Q|H=3.00] = 13205.43$$

$$E[Q|H=4.00] = 14998.15$$

$$E[Q|H=5.00] = 15944.93$$

$$E[Q|H=6.00] = 17033.75$$

$$E[Q|H=7.95] = 19931.36$$

Chúng ta hãy vẽ đồ thị của các trị trung bình có điều kiện này theo các giá trị của biến điều kiện:



Lưu ý rằng theo cách này, chúng ta cho rằng hai biến khác nhau về cơ bản. Chúng ta xem biến điều kiện là cho trước (thay vì ngẫu nhiên), và biến Q là biến ngẫu nhiên.

### Bàn thêm về Kỳ vọng (các anh chị kỳ vọng điều gì?!)

Khi là nhà kinh tế lượng, chúng ta sẽ rất thường quan tâm đến tổng của các biến ngẫu nhiên, cũng như các trị trung bình và phương sai của các tổng này. Ta hãy bắt đầu tìm giá trị trung bình và phương sai của tổng hai biến ngẫu nhiên.

Để xem điều này được thực hiện như thế nào, chúng ta hãy bắt đầu bằng bảng giao xác suất và viết hai con số vào mỗi ô: giá trị của  $(x_i + y_j)$  và xác suất  $P(X = x_i \cap Y = y_j)$ .

Để tính kỳ vọng của tổng, ta phải cộng tất cả các phần tử

$$(x_i + y_j) \times P(X = x_i \cap Y = y_j)$$

Kết quả ta được:

$$E[X + Y] = \sum_X \sum_Y (x_i + y_j) \times P(X = x_i \cap Y = y_j)$$

Ở Bài tập thứ 2, các anh chị sẽ phải chứng minh rằng:

$$E[X + Y] = E[X] + E[Y]$$

Bây giờ chúng ta hãy xét phương sai của tổng:

$$\begin{aligned} \text{VAR}[X + Y] &= E[(X + Y - E[(X + Y)])^2] \\ &= E[(X - E[X] + Y - E[Y])^2] \\ &= E[(X - E[X])^2] + E[(Y - E[Y])^2] + 2E[(X - E[X])(Y - E[Y])] \end{aligned}$$

Số hạng cuối cùng có một cái tên đặc biệt: đó là hiệp phương sai của X và Y. Vì thế:

$$\text{VAR}[X + Y] = \text{VAR}[X] + \text{VAR}[Y] + 2*\text{COV}[X, Y]$$

Hiệp phương sai được tính bằng công thức:

$$\text{COV}[X, Y] = \sum_X \sum_Y (x_i - \mu_X)(y_j - \mu_Y) \times P(X = x_i \cap Y = y_j)$$

Trong trường hợp đặc biệt khi X và Y là độc lập thống kê, thì hiệp phương sai của chúng bằng 0. Bài tập 2 sẽ cho các anh chị cơ hội để chứng minh điều này.

Hiệp phương sai có một cách giải thích khá hấp dẫn bằng hình học và cách giải thích này được chúng ta trình bày trong lớp.

## Hệ số tương quan

Hiệp phương sai có nhược điểm là khó giải thích vì nó phụ thuộc vào đơn vị đo của các biến X và Y.

Hệ số tương quan là một con số không có đơn vị nên tránh được nhược điểm đã nêu trên:

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2} \sqrt{\sigma_Y^2}}$$

Người ta có thể chứng minh được:

$$-1 \leq \rho \leq 1$$

## Các biến ngẫu nhiên liên tục

Các phân phối xác suất hai biến của các biến ngẫu nhiên liên tục không thể biểu diễn được ở dạng bảng đơn giản. Chúng ta cần trình bày chúng dưới các dạng hàm phân phối xác suất phù hợp. Khi cho trước hàm phân phối giao xác suất của một cặp biến ngẫu nhiên, được ký hiệu là  $f(x,y)$ , mọi định nghĩa nêu trên đều có thể được áp dụng cho hàm này, với các toán tử tổng được thay thế bằng các tích phân.

Ngày mai, chúng ta sẽ nghiên cứu một số hàm phân phối xác suất cho một số biến ngẫu nhiên liên tục.