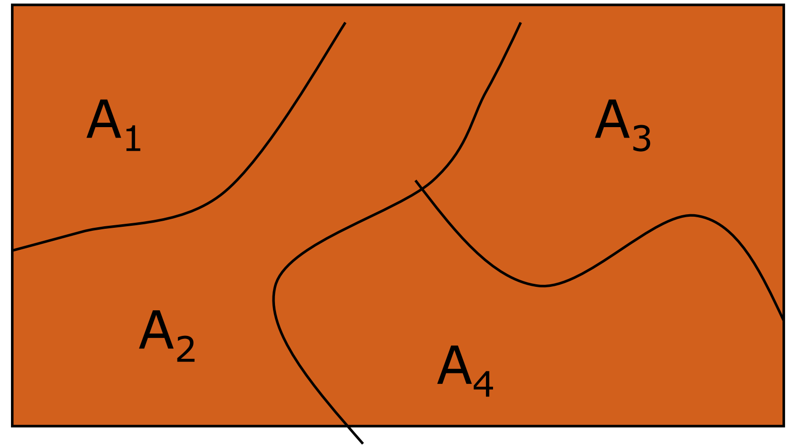
TIÊN ĐỀ CỦA XÁC SUẤT



1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$ (S là không gian mẫu)
3. Nếu $A_1, A_2, \ldots, A_n$ là các sự kiện loại trừ lẫn nhau $\left(P\left(A_i \cap A_j\right) = 0\right)$, ta có:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

Lưu ý: có thể viết: $P\left(A_i \cap A_j\right)$ dưới dạng $P\left(A_i, A_j\right)$

XÁC SUẤT TIÊN NGHIỆM

- Xác suất tiên nghiệm là xác suất của một sự kiện không có rằng buộc nào trước đó.
- Ví dụ:

  P(thi đỗ)=0.1 có nghĩa: trong trường hợp không có thêm thông tin nào khác thì chỉ có 10% là thi đỗ.

XÁC SUẤT CÓ ĐIỀU KIỆN

- Xác suất có điều kiện là xác suất của một sự kiện nào đó khi có thêm thông tin rằng buộc.
- Ví dụ:

P(thi đỗ | học sinh giỏi) = 0.8 có nghĩa: xác suất để học sinh thi đỗ khi biết đó là học sinh giỏi là 80%.

XÁC SUẤT CÓ ĐIỀU KIỆN (CONT)

- Xác suất có điều kiện có thể được định nghĩa qua xác suất không điều kiện:

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

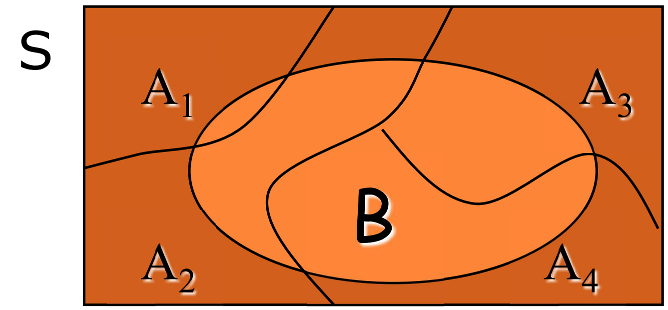
- Hay ta có: $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$

LUẬT TỔNG XÁC SUẤT

- Nếu $A_1, A_2, \ldots, A_n$ là các phần ứng với các sự kiện loại trừ lẫn nhau và B là một sự kiện nào đó, ta có:

$$P(B) = P(B/A_1)P(A_1) + P(B/A_2)P(A_2) + \ldots + P(B/A_n)P(A_n)$$

$$= \sum_{j=1}^{n} P(B/A_j)P(A_j)$$



- Trong trường hợp đặc biệt:

$$P(A) = P(A,B) + P(A,\bar{B})$$

- Sử dụng quy tắc biến đổi, ta có:

$$P(A) = P(A,B) + P(A,\bar{B}) = P(A/B)P(B) + P(A/\bar{B})P(\bar{B})$$

VÍ DỤ VỀ LUẬT TỔNG XÁC SUẤT

- My mood can take one of two values: Happy, Sad
- The weather can take one of three values: Rainy, Sunny, Cloudy.
- We can compute P(Happy) and P(Sad) as follows:

$$P(Happy) = P(Happy|Rainy) + P(Happy|Sunny) + P(Happy|Cloudy)$$

$$P(Sad) = P(Sad|Rainy) + P(Sad|Sunny) + P(Sad|Cloudy)$$

ĐỊNH LÝ BAYES

- Theo luật Bayes, ta có:

$$\boxed{P(A/B) = \frac{P(B/A)P(A)}{P(B)}}$$

trong đó,

$$P(B) = P(B,A) + P(B,\bar{A}) = P(B/A)P(A) + P(B/\bar{A})P(\bar{A})$$

VÍ DỤ VỀ ĐỊNH LÝ BAYES

- Bệnh M là nguyên nhân dẫn đến 50% có bệnh S.
- Một bệnh nhân có bệnh S, hỏi xác suất có bệnh M là bao nhiêu?
- Ta biết rằng:
  - Xác suất có bệnh M là 1/50,000.
  - Xác suất có bệnh S là 1/20.

$$P(M/S) = \frac{P(S/M)P(M)}{P(S)}$$

$$P(M|S)=0.0002$$

DẠNG TỔNG QUÁT CỦA LUẬT BAYES

- Nếu $A_1, A_2, \dots, A_n$ là các phần ứng với các sự kiện loại trừ lẫn nhau và B là một sự kiện nào đó, ta có:

$$P(A_i / B) = \frac{P(B / A_i)P(A_i)}{P(B)}$$

trong đó:

$$P(B) = \sum_{j=1}^{n} P(B / A_j)P(A_j)$$

SỰ KIỆN ĐỘC LẬP

- Hai sự kiện A và B là độc lập nếu và chỉ nếu:

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

- Từ công thức trên, chúng ta có thể thấy:

$$P(A \mid B) = P(A) \text{ và } P(B \mid A) = P(B)$$

- A và B là điều kiện độc lập theo C nếu và chỉ nếu:

$$P(A \mid B, C) = P(A \mid C)$$

- Ví dụ,

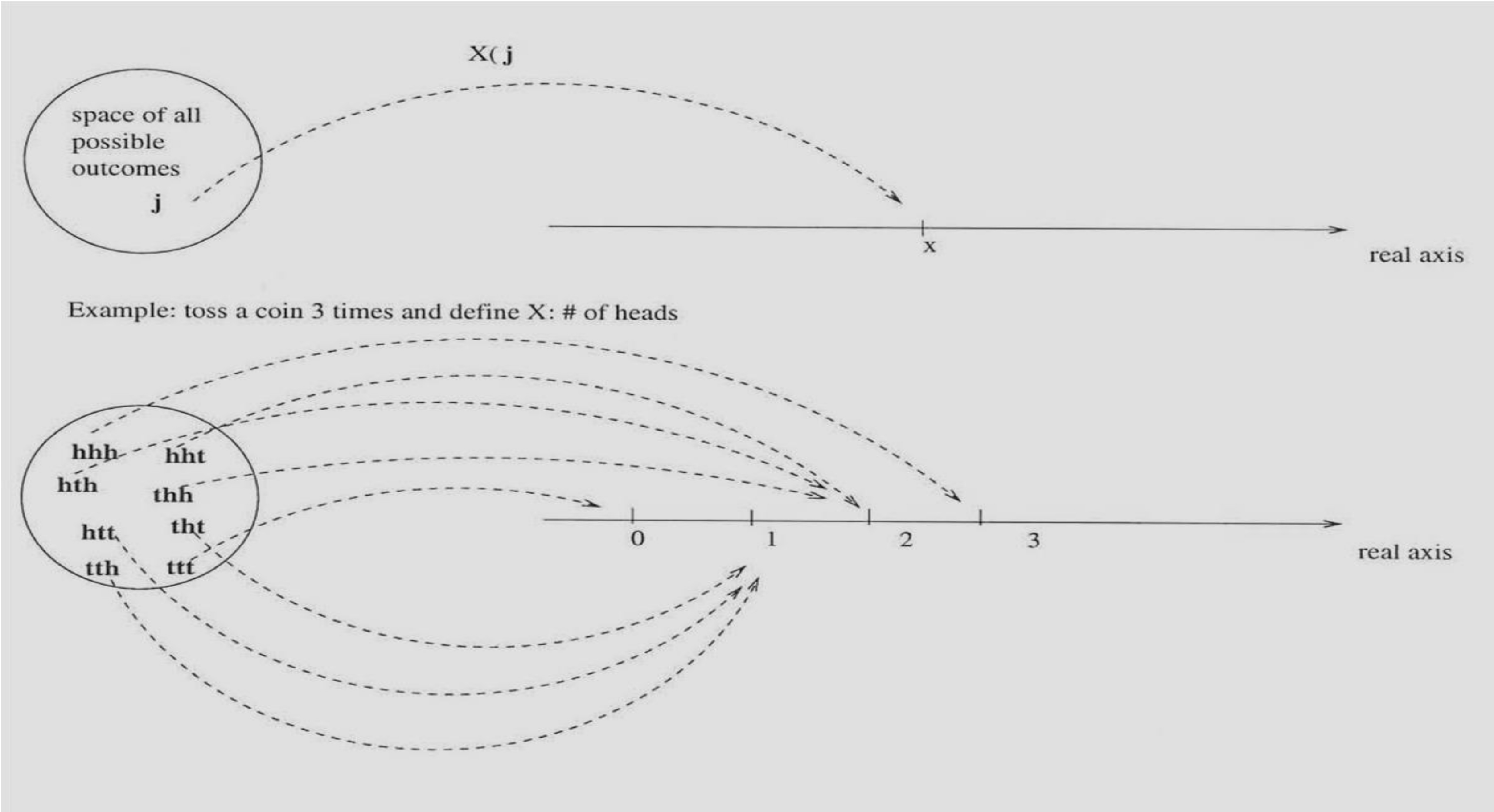
P(Wet Grass | Season, Rain)=P(Wet Grass | Rain)

Biến ngẫu nhiên

- Trong nhiều thử nghiệm, đôi khi quan tâm tới biến tổng hơn là dạng xác suất ban đầu.
- *Ví dụ*: trong một lần thăm dò dư luận, chúng ta tiến hành hỏi 50 người đồng ý hay không về một dự luật nào đó.
  - Ký hiệu "1" ứng với đồng ý, "0" ứng với không đồng ý.
  - Như vậy, không gian mẫu có 2^{50} phần tử.
  - Giả sử, ta chỉ quan tâm tới số người đồng ý.
  - Như vậy, có thể định nghĩa biến X = số số "1", có giá trị từ 0 đến 50.
  - Điều này có nghĩa, không gian mẫu nhỏ hơn, có 51 phần tử.

BIẾN NGẪU NHIÊN (CONT)

- Biến ngẫu nhiên là giá trị ta gán cho kết quả của một thử nghiệm ngẫu nhiên (hàm cho phép gán một số thực ứng với mỗi sự kiện).



Biến ngẫu nhiên (cont)

- Như vậy, làm thế nào để có hàm xác suất theo biến ngẫu nhiên từ hàm xác suất trên không gian mẫu ban đầu?
  - Giả sử ta có không gian mẫu là $S = \langle s_1, s_2, \dots, s_n \rangle$.
  - Giả sử phạm vi của biến ngẫu nhiên X nằm trong $\langle x_1, x_2, \dots, x_m \rangle$.
  - Ta quan sát thấy $X = x_j$ khi và chỉ khi kết quả của thử nghiệm ngẫu nhiên là $s_j \in S$, hay $X(s_j) = x_j$

$$\mathbf{P}(\mathbf{X} = \boldsymbol{x}_{\mathbf{j}}) = \mathbf{P}(\mathbf{s}_{\mathbf{j}} \in \mathbf{S} : \mathbf{X}(\boldsymbol{s}_{\boldsymbol{j}}) = \boldsymbol{x}_{\boldsymbol{j}})$$

  - Ví dụ: trong ví dụ trên thì P(X=2)=?

BIẾN NGẪU NHIÊN LIÊN TỤC / RỜI RẠC

- Biến ngẫu nhiên rời rạc là biến mà giá trị của nó là đếm được.
- Ví dụ: quan sát việc tung 2 con xúc xắc.
  - Gọi X là tổng các mặt của 2 con xúc xắc.
  - X=5 tương ứng với không gian có thể $A_5 = \{(1,4),(4,1),(2,3),(3,2)\}$.
  - Vậy ta có:

$$P(X = x) = P(A_x) = \sum_{s:X(s)=x} P(s)$$

Hay:

$$P(X = 5) = P\big((1,4)\big) + P\big((4,1)\big) + P\big((2,3)\big) + P\big((3,2)\big) = \frac{4}{36} = \frac{1}{9}$$

- Biến ngẫu nhiên liên tục là biến mà giá trị của nó thuộc nhóm không đếm được.

HÀM TỔNG XÁC SUẤT – HÀM MẬT ĐỘ XÁC SUẤT

- Hàm tổng xác suất - Probability mass function: là hàm cho biết xác suất của một biến ngẫu nhiên rời rạc X nào đó với giá trị x_i trong miền giá trị. Ký hiệu **pmf**.

- Hàm mật độ xác suất - Probability density function: là hàm một hàm bất kỳ *f*(*x*) mô tả mật độ xác suất theo biến đầu vào *x*. Ký hiệu **pdf**.

HÀM KHỐI XÁC SUẤT – HÀM MẬT ĐỘ XÁC SUẤT (CONT)

- Ví dụ về pmf và pdf:

$$\sum_{x} p(x) = 1; hàm\ pmf$$

$$P(a < X < b) = \sum_{k=a}^{b} p(k)\,; hàm\ pmf$$

$$\int_{-\infty}^{\infty} p(x)dx = 1; hàm\ pdf$$

$$P(a < X < b) = \int_{a}^{b} p(t)dt\,; hàm\ pdf$$

HÀM PHÂN BỐ XÁC SUẤT - PDF

- Hàm phân bố xác suất - Probability Distribution Function – ký hiệu **PDF**: là hàm được định nghĩa:

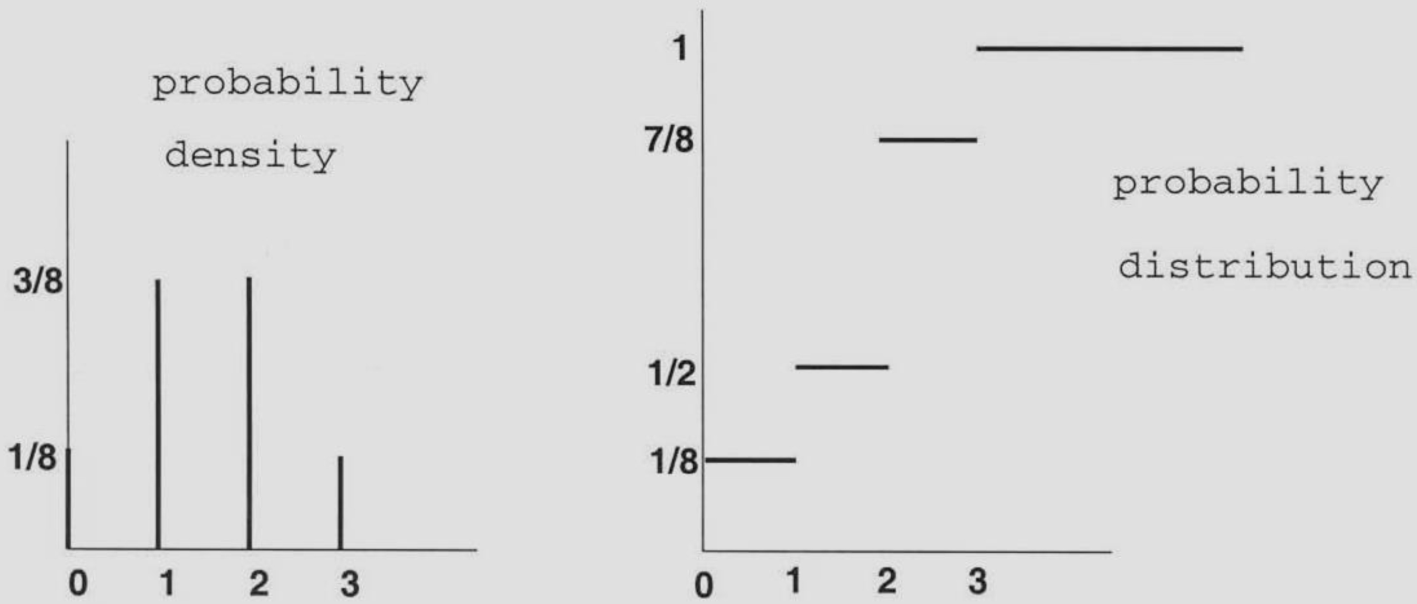
$$F(x) = P(X \leq x)$$

- Một số tính chất của hàm phân bố xác suất:
  - (1) $0 \leq F(x) \leq 1$
  - (2) *F(x)* là hàm không giảm theo biến x.
- Nếu X rời rạc, hàm phân bố xác suất được tính:

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} P(X = k) = \sum_{k=0}^{x} p(k)$$

HÀM PHÂN BỐ XÁC SUẤT – PDF (CONT)

- Ví dụ minh họa:



$$F(0) = P(X \le 0) = P(X = 0) = 1/8$$

$$F(1) = P(X \le 1) = P(X = 0) + P(X = 1) = 1/2$$

$$F(2) = P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2) = 7/8$$

$$F(3) = P(X \le 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1$$

HÀM PHÂN BỐ XÁC SUẤT – PDF (CONT)

- Nếu X là biến liên tục, PDF có thể tính:

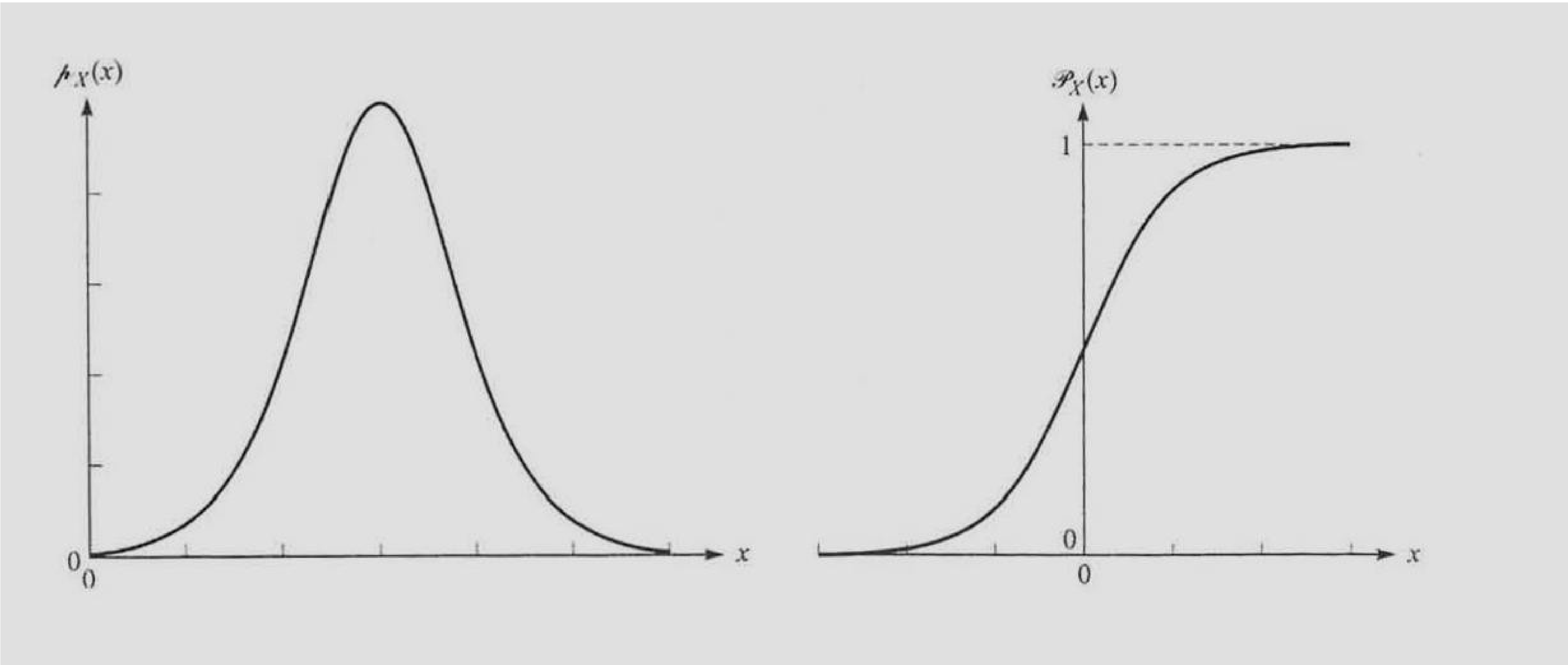
$$F(x) = \int_{-\infty}^{x} p(t)dt \ \text{với mọi}\ x$$

- Sử dụng công thức trên, ta có:

$$p(x) = \frac{dF}{dx}(x)$$

HÀM PHÂN BỐ XÁC SUẤT – PDF (CONT)

- Ví dụ về pdf và PDF của Gaussian



HÀM PMF NHIỀU BIẾN (BIẾN RỜI RẠC)

- Với hàm n biến ngẫu nhiên, khi đó ta có pmf nhiều biến được viết:

$$p(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

- Ví dụ: hàm pmf ứng với 2 biến:

P(Cavity, Toothache)

	Toothache	Not Toothache
Cavity	0.04	0.06
Not Cavity	0.01	0.89

Tổng xác suất = 1.00

HÀM PDF NHIỀU BIẾN (BIẾN LIÊN TỤC)

- Với n biến ngẫu nhiên liên tục, hàm pdf nhiều biến được tính:

$$p(x_1, x_2, \dots, x_n) \geq 0$$

$$\int_{x_1} \int_{x_2} \dots \int_{x_n} p(x_1, x_2, \dots, x_n) dx_1 \dots dx_n = 1$$

MỘT SỐ TÍNH CHẤT

- Hàm pdf có điều kiện có thể được tính từ hàm pdf nhiều biến:

$$p(y|x) = \frac{p(x,y)}{p(x)} \; hay \; p(x,y) = p(y|x)p(x)$$

- Với trường hợp nhiều biến (n biến), ta có dạng tổng quát:

$$p(x_1, x_2, \dots, x_n)$$
$$= p(x_1|x_2, \dots, x_n).p(x_2|x_3, \dots, x_n) \dots p(x_{n-1}|x_n).p(x_n)$$

MỘT SỐ TÍNH CHẤT (CONT)

- Nếu các biến là độc lập, khi đó ta có (ví dụ với 2 biến X và Y):

$$p(x,y) = p(x).p(y)$$

- Quy tắc tổng xác suất:

$$p(y) = \sum_{x} p(y|x)p(x)$$

HÀM PHÂN BỐ CHUẨN (GAUSSIAN)

- Hàm phân bố chuẩn Gaussian được định nghĩa:

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

trong đó: μ: giá trị kỳ vọng; σ: độ lệch chuẩn

- Với x là một véc tơ, ta có:

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-(x-\mu)^t\Sigma^{-1}(x-\mu)]$$

trong đó: d: số chiều; μ: kỳ vọng; Σ: ma trận hiệp phương sai.

HÀM PHÂN BỐ CHUẨN (GAUSSIAN) - CONT

- Ví dụ về phân bố chuẩn có 2 biến:

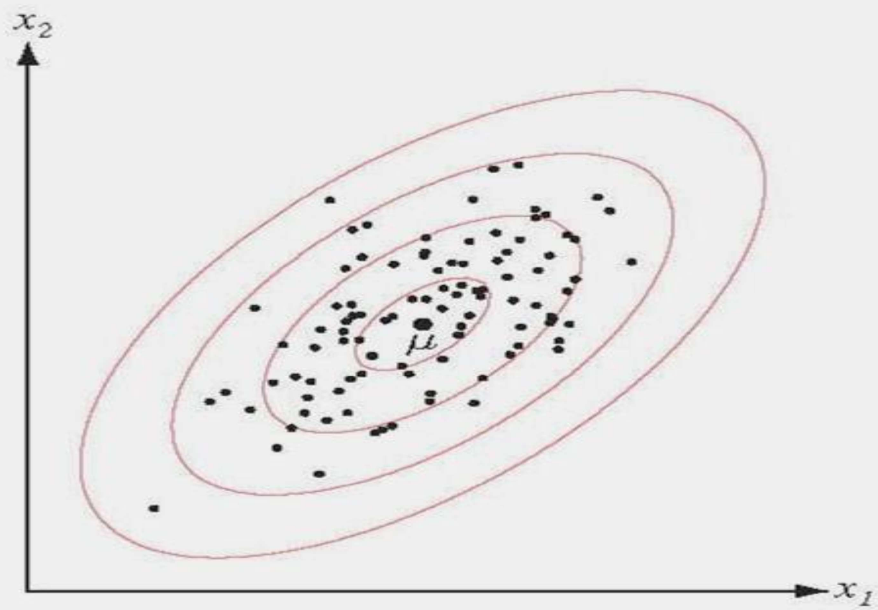


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

HÀM PHÂN BỐ CHUẨN (GAUSSIAN) - CONT

- Khoảng cách Mahalanobis:

$$r^2 = (\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu)$$

- Nếu các biến là độc lập, khi đó phân bố chuẩn có dạng:

$$p(x) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp[\frac{(x_i - \mu_i)^2}{2\sigma_i^2}]$$

GIÁ TRỊ KỲ VỌNG

- Giá trị kỳ vọng của một biến ngẫu nhiên rời rạc X được cho bởi:

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

- Ví dụ: gọi X là các mặt có thể ra của một con xúc xắc, ta có:

$$E(X) = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = 3.5.$$

- Trong trường hợp đơn giản, ta có:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

GIÁ TRỊ KỲ VỌNG (CONT)

- Mối quan hệ giữa giá trị trung bình và giá trị kỳ vọng:

$$E(X) = \lim_{n \to \infty} \bar{x}$$

- Giá trị kỳ vọng cho biến ngẫu nhiên liên tục:

$$E(X) = \int_{-\infty}^{+\infty} xp(x)dx$$

- Ví dụ, giá trị kỳ vọng của Gaussian:

$$E(X) = \mu$$

MỘT SỐ TÍNH CHẤT CỦA GIÁ TRỊ KỲ VỌNG

- Giá trị kỳ vọng của hàm $\mathrm{g(X)}$ được xác định:

$$\mathrm{E}\big(\mathrm{g(X)}\big) = \sum_{\mathrm{x}} \mathrm{g(x)p(x)} \ \text{ với T. H rời rạc}$$

$$\mathrm{E}\big(\mathrm{g(X)}\big) = \int_{-\infty}^{+\infty} \mathrm{g(x)p(x)dx} \ \text{ với T. H liên tục}$$

- Tính chất tuyến tính:

$$E\big(af(X) + bg(Y)\big) = aE\big(f(X)\big) + bE\big(g(Y)\big)$$

PHƯƠNG SAI VÀ ĐỘ LỆCH CHUẨN

- Phương sai Var(X) của biến ngẫu nhiên X:

$$\mathrm{Var(X)} = \mathrm{E}\left((\mathrm{X}-\mu)^2\right), \text{với } \mu = \mathrm{E(X)}$$

- Phương sai theo mẫu dữ liệu:

$$\overline{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Độ lệch chuẩn σ của biến ngẫu nhiên X:

$$\sigma = \sqrt{Var(X)}$$

- Ví dụ: phương sai Gaussian là σ^2

HIỆP PHƯƠNG SAI

- Hiệp phương sai của 2 biến X và Y:

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

$$\text{với } \mu_X = E(X), \mu_Y = E(Y)$$

- Hệ số tương quan ρ_{XY} giữa X và Y:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Ma trận hiệp phương sai của mẫu dữ liệu:

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i-\bar{x})(y_i-\bar{y})$$

MA TRẬN HIỆP PHƯƠNG SAI

- Với 2 biến X, Y, ma trận hiệp phương sai:

$$C_{XY} = \begin{bmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X & \mathrm{Cov}(Y,Y) \end{bmatrix}$$

$$\text{với } \mathrm{Cov}(X,X) = \mathrm{Var}(X); \mathrm{Cov}(Y,Y) = \mathrm{Var}(Y)$$

- Với trường hợp nhiều biến:

$$C_X = \begin{bmatrix} Cov(X_1,X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Cov(X_2,X_2) & \dots & Cov(X_2,X_n) \\ \dots & \dots & \dots & \dots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \dots & Cov(X_n,X_n) \end{bmatrix}$$

$$\text{where } Cov(X_i,X_j) = Cov(X_j,X_i) \text{ and } Cov(X_i,X_i) \geq 0$$