THỐNG KÊ MÁY TÍNH & ỨNG DỤNG
Bài 3
BIẾN NGẪU NHIÊN VÀ PHÂN PHỐI

Vũ Quốc Hoàng

(vqhoang@fit.hcmus.edu.vn)

FIT-HCMUS, 2018

Nội dung

- Biến ngẫu nhiên
- Phân phối của biến ngẫu nhiên
- Biến ngẫu nhiên rời rạc và hàm xác suất
- Biến ngẫu nhiên liên tục và hàm mật độ xác suất
- Hàm phân phối tích lũy
- Hàm phân vị

Biến ngẫu nhiên

- Nếu giá trị của một đại lượng/tính chất X được xác định hoàn toàn khi biết kết quả ω của một thí nghiệm T thì X được gọi là một đại lượng/biến ngẫu nhiên (liên quan đến T)
  - Trước khi biết kết quả, ta chỉ biết X có thể nhận một giá trị nào đó trong tập giá trị A
  - Sau khi biết kết quả ω, ta biết X nhận một giá trị cụ thể $x \in A$, ta kí hiệu $X(w) = x$
- Biến ngẫu nhiên (random variable) là hàm trên không gian mẫu Ω
  - $X: \Omega \rightarrow A$, gắn mỗi kết quả $\omega \in \Omega$ một giá trị $X(\omega) \in A$
  - A được gọi là tập/miền giá trị của X
  - Nếu A là tập con của tập số thực $\mathbb{R}$, ta nói X là biến số hay biến định lượng
  - Nếu A hữu hạn và không là tập con của $\mathbb{R}$, ta nói X là biến định tính

Biến ngẫu nhiên
Ví dụ

- Xét thí nghiệm: chọn ngẫu nhiên một sinh viên trong lớp
  - $\Omega = \{\text{An}, \text{Bình}, \text{Chương}, \dots\}$
  - Đo chiều cao H của sinh viên được chọn:
    - H là biến định lượng với tập giá trị là $\mathbb{R}$ (hoặc $[1.0, 2.0]$ mét)
    - $H(\text{An}) = 1.5$ mét, $H(\text{Bình}) = 1.7$ mét, …
  - Xác định giới tính G của sinh viên được chọn:
    - G là biến định tính với tập giá trị là $\{\text{Nam}, \text{Nữ}\}$ (hoặc $\{0, 1\}$)
    - $G(\text{An}) = \text{Nữ}$, $G(\text{Bình}) = \text{Nam}$, …
  - Xét điểm S của sinh viên được chọn: S là biến định lượng với tập giá trị là $\{0, 0.5, 1, 1.5, \dots, 9.5, 10\}$ (hoặc $\mathbb{R}$)
  - Xét học lực L của sinh viên được chọn: L là biến định tính với tập giá trị là $\{\text{Yếu}, \text{Kém}, \text{Trung bình}, \text{Khá}, \text{Giỏi}, \text{Xuất sắc}\}$

Biến ngẫu nhiên

- B.n.n (biến ngẫu nhiên) là phương tiện hay dùng để mô tả các biến cố
- Xét biến (số) ngẫu nhiên X liên quan đến thí nghiệm T có không gian mẫu là Ω
  - Cho $C \subset \mathbb{R}$, ta kí hiệu biến cố "X nhận giá trị trong C" là:
$$(X \in C) = \{\omega \in \Omega: X(\omega) \in C\}$$
  - Chẳng hạn, cho $x \in \mathbb{R}$ ta kí hiệu:
$$(X = x) = \{\omega \in \Omega: X(\omega) = x\}$$
$$(X \leq x) = \{\omega \in \Omega: X(\omega) \leq x\}$$
$$(X > x) = \{\omega \in \Omega: X(\omega) > x\}$$
  - Hay với hai biến X, Y ta kí hiệu:
$$(X = Y) = \{\omega \in \Omega: X(\omega) = Y(x)\}$$
$$(X \leq Y) = \{\omega \in \Omega: X(\omega) \leq Y(\omega)\}$$
  - Các biến cố này còn được gọi là biến cố liên quan đến b.n.n X, Y

Biến ngẫu nhiên
Ví dụ

- Xét thí nghiệm: gieo một xúc xắc (đồng chất) 2 lần, $\Omega = \left\{(i,j)\colon\ i,j \in \{1,2,3,4,5,6\}\right\}$, mô hình xác suất đơn giản
  - Gọi X, Y là các b.n.n "số chấm ở lần 1", "số chấm ở lần 2"
  $$X\big(\omega = (i,j)\big) = i \ \text{và}\ Y\big((i,j)\big) = j$$
  - Biến cố được "số chấm ở lần 1 là 6" là:
  $$(X = 6) = \left\{(6,j)\colon j \in \{1,2,3,4,5,6\}\right\} = \{(6,1),(6,2),\dots,(6,6)\}$$
  - Biến cố được "số chấm ở hai lần như nhau" là:
  $$(X = Y) = \{(i,j)\colon i = j\} = \{(1,1),(2,2),\dots,(6,6)\}$$
  - Xác suất để được "số chấm ở hai lần như nhau" là:
  $$P(X = Y) = |(X = Y)|/|\Omega| = 6/36 = 1/6$$
  - Xác suất để được "số chấm ở lần 1 lớn hơn số chấm ở lần 2" khi biết "số chấm ở lần 2 lớn hơn 4" là:
  $$P(X > Y \mid Y > 4) = |(X > Y > 4)|/|(Y > 4)| = 1/12$$

Phân phối của b.n.n

- Xét b.n.n X liên quan đến thí nghiệm T có không gian mẫu là Ω
  - Cho $C \subset \mathbb{R}$, ta có $P(X \in C)$ là xác suất để "X nhận giá trị trong C"
  - Tập các xác suất $\{P(X \in C): C \subset \mathbb{R}\}$ xác định một độ đo xác suất trên (không gian mẫu mới) $\mathbb{R}$ và được gọi là phân phối (distribution) của X
  - Phân phối của X cho thấy khả năng X nhận các giá trị khác nhau
  - Với phân phối của X, ta khảo sát X mà không cần để ý đến T hay Ω nữa
  - Nói chung, tập $\{P(X \in C): C \subset \mathbb{R}\}$ là "rất khó tính toán". Ta cần cách nào đó giúp xác định phân phối của X để "dễ tính toán hơn":
    - Hàm xác suất (cho b.n.n rời rạc)
    - Hàm mật độ xác suất (cho b.n.n liên tục)
    - Hàm phân phối tích lũy (chung cho các b.n.n)

Phân phối của b.n.n
Ví dụ

- B.n.n X có tập giá trị là $\{x_0\}$
  - $T(\omega) = x_0, \forall\omega \in \Omega$
  - X chỉ có 2 biến cố liên quan là $(X \neq x_0) = \emptyset$ và $(X = x_0) = \Omega$
  - Không nên gọi X là b.n.n vì ta biết giá trị của X chắc chắn là x_0 ngay cả trước khi tiến hành thí nghiệm
  - Phân phối của X rất đơn giản:

$$P(X \in C) = \begin{cases} 1 & \text{nếu } C \text{ chứa } x_0 \\ 0 & \text{nếu } C \text{ không chứa } x_0 \end{cases}$$

- Ví dụ: xét b.n.n X là "điểm tổng kết" trong thí nghiệm "bỏ thi môn TKMT&UD", X chỉ có một giá trị là 0 (điểm)

Phân phối của b.n.n
Ví dụ

- Cho biến cố A liên quan đến thí nghiệm T có không gian mẫu là Ω, ta gọi hàm đặc trưng (characteristic function) của A là hàm $I_A: \Omega \to \mathbb{R}$ được xác định bởi:

$$I_A(\omega) = \begin{cases} 1 & \text{nếu } \omega \in A \\ 0 & \text{nếu } \omega \notin A \end{cases}$$

  - I_A là b.n.n chỉ có 4 biến cố liên quan là $\emptyset, (I_A = 1) = A, (I_A = 0) = A^c$ và Ω
  - Phân phối của I_A khá đơn giản:

$$P(X \in C) = \begin{cases} 0 & \text{nếu } C \text{ không chứa cả 0 lẫn 1} \\ P(A) & \text{nếu } C \text{ chứa 1 nhưng không chứa 0} \\ 1 - P(A) & \text{nếu } C \text{ chứa 0 nhưng không chứa 1} \\ 1 & \text{nếu } C \text{ chứa cả 0 lẫn 1} \end{cases}$$

- Ví dụ: xét b.n.n X là "số lần được mặt chẵn" trong thí nghiệm gieo xúc xắc, X là hàm đặc trưng của biến cố "được mặt chẵn"
- Hàm đặc trưng giúp khảo sát biến cố như là một b.n.n

B.n.n rời rạc và hàm xác suất

- X được gọi là b.n.n rời rạc (discrete random variable) nếu tập giá trị của nó là rời rạc (hữu hạn hay vô hạn đếm được)
- Với X là b.n.n rời rạc, hàm xác suất (probability function) của X là hàm $f: \mathbb{R} \to \mathbb{R}$, được xác định bởi:

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

  - Hàm xác suất f cho biết khả năng X nhận một giá trị cụ thể
  - Tập số thực $\{x \in \mathbb{R}: f(x) > 0\}$ được gọi là tập hỗ trợ của X, kí hiệu $\mathrm{Sup}(X)$
  - Để chỉ rõ hàm xác suất của X, ta còn kí hiệu f là f_X
  - Hàm xác suất có tính chất: $f_X(x) \geq 0, \forall x \in \mathbb{R}$ và $\sum_{x \in \mathrm{Sup}(X)} f_X(x) = 1$
- Hàm xác suất xác định phân phối của b.n.n rời rạc:

$$P(X \in C) = \sum_{x \in C} f_X(x), C \subset \mathbb{R}$$

B.n.n rời rạc và hàm xác suất
Ví dụ

- Xét thí nghiệm tung một đồng xu (đồng chất) 2 lần, đặt X là số lần được mặt ngửa:
  - Tập giá trị của X là $\{0, 1, 2\}$
  - X là b.n.n rời rạc
  - Hàm xác suất của X được cho bởi:

$$f_X(x) = P(X = x) = \begin{cases} 1/4 & \text{nếu } x = 0 \\ 2/4 & \text{nếu } x = 1 \\ 1/4 & \text{nếu } x = 2 \\ 0 & \text{nếu } x \notin \{0, 1, 2\} \end{cases}$$

Hàm f_X còn được cho bởi bảng sau (gọi là bảng phân phối xác suất của X):

x	0	1	2
$P(X = x)$	1/4	1/2	1/4

B.n.n rời rạc và hàm xác suất
Phân phối rời rạc đều

- B.n.n rời rạc X được gọi là có phân phối đều (uniform distribution) trên tập n giá trị $\{x_1, x_2, \dots, x_n\}$ nếu X có hàm xác suất:

$$f_X(x) = P(X = x) = \frac{1}{n}, x \in \{x_1, x_2, \dots, x_n\}$$

  - X là kết quả của thí nghiệm "chọn ngẫu nhiên một điểm trong tập n giá trị"
- Ví dụ: xét thí nghiệm gieo một xúc xắc (đồng chất) 2 lần, gọi X, Y là các b.n.n "số chấm ở lần 1" và "số chấm ở lần 2"
  - Ta có X, Y đều là các b.n.n rời rạc có phân phối đều trên tập $\{1, 2, \dots, 6\}$
  - Tuy nhiên, "tổng số chấm ở hai lần", $Z = X + Y$, là b.n.n rời rạc với tập giá trị $\{2, 3, \dots, 11, 12\}$ có phân phối không đều

B.n.n rời rạc và hàm xác suất
Phân phối Bernoulli

- B.n.n rời rạc X được gọi là có phân phối Bernoulli (Bernoulli distribution) với tham số p nếu X có tập giá trị là $\{0, 1\}$ và:

$$f_X(x) = P(X = x) = \begin{cases} p & \text{nếu } x = 1 \\ 1 - p & \text{nếu } x = 0 \end{cases}$$

  Kí hiệu $X \sim \text{Bernoulli}(p)$

- Ví dụ:
  - Xét thí nghiệm tung một đồng xu, gọi X là b.n.n "số lần được ngửa":
    - Nếu đồng xu đồng chất: $X \sim \text{Bernoulli}(0.5)$
    - Nếu đồng xu không đồng chất với xác suất ra ngửa là 0.7: $X \sim \text{Bernoulli}(0.7)$
  - Xét thí nghiệm T với biến cố A có $P(A) = p$, khi đó $I_A \sim \text{Bernoulli}(p)$

B.n.n rời rạc và hàm xác suất
Phân phối nhị thức

- B.n.n rời rạc X được gọi là có phân phối nhị thức (binomial distribution) với tham số n, p nếu X có tập giá trị là $\{0, 1, \dots, n\}$ và:
$$f_X(x) = P(X = x) = C_n^x p^x (1-p)^{n-x}, x \in \{0, 1, \dots, n\}$$
Kí hiệu $X \sim \mathrm{Binomial}(n, p)$
- Ví dụ:
  - Xét thí nghiệm tung một đồng xu đồng chất 5 lần, gọi X là b.n.n "số lần được ngửa" thì $X \sim \mathrm{Binomial}(5, 0.5)$. Khi đó, xác suất để được không quá 1 lần ngửa là:
  $$P(X \le 1) = f_X(0) + f_X(1) = C_5^0 0.5^0 0.5^5 + C_5^1 0.5^1 0.5^4 = 0.1875$$
  - Xét thí nghiệm T với biến cố A có $P(A) = p$. Xét thí nghiệm R "thực hiện T lặp lại n lần độc lập", gọi X là b.n.n "số lần A xảy ra" thì $X \sim \mathrm{Binomial}(n, p)$

B.n.n liên tục và hàm mật độ xác suất

- X được gọi là b.n.n liên tục (continuous random variable) nếu có hàm số không âm $f: \mathbb{R} \to \mathbb{R}$ sao cho với mọi khoảng $[a, b]$ trong $\mathbb{R}$ ta có:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

  - f được gọi là hàm mật độ xác suất (probability denstity function) của X vì nó cho biết khả năng X nhận giá trị trong các khoảng rất nhỏ của trục số thực $\mathbb{R}$

$$P(a - \varepsilon \leq X \leq a + \varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\,dx \approx 2\varepsilon f(a) \text{ khi } \varepsilon \text{ rất nhỏ}$$

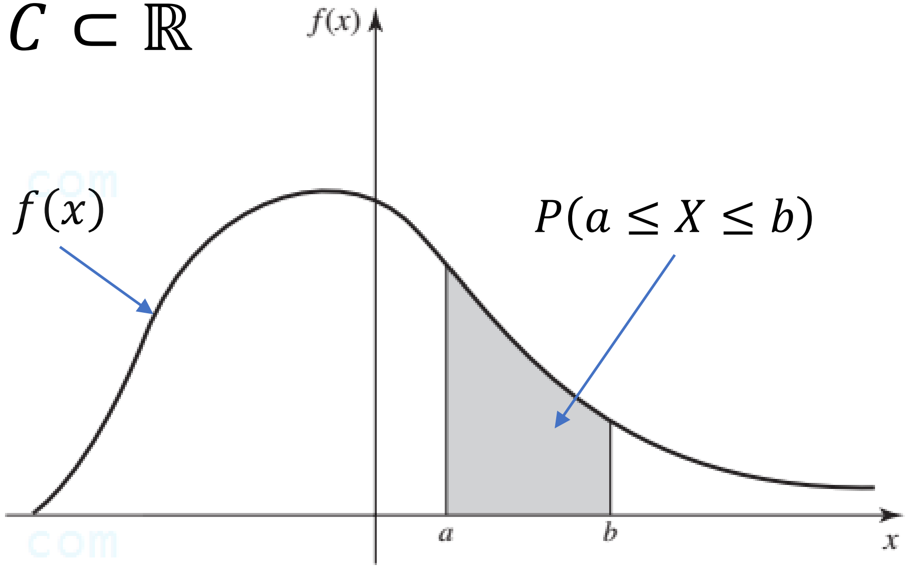
  - Tập số thực $\{x \in \mathbb{R}: f(x) > 0\}$ được gọi là tập hỗ trợ của X, kí hiệu $\mathrm{Sup}(X)$
  - Để chỉ rõ hàm mật độ xác suất của X, ta còn kí hiệu f là f_X
  - Hàm mật độ xác suất có tính chất: $f_X(x) \geq 0, \forall x \in \mathbb{R}$ và $\int_{-\infty}^{\infty} f(x)\,dx = 1$

B.n.n liên tục và hàm mật độ xác suất

- Hàm mật độ xác suất xác định phân phối của b.n.n liên tục:

$$P(X \in C) = \int_C f_X(x)\,dx\,, C \subset \mathbb{R}$$

  - $P(X = a) = \int_a^a f(x)\,dx = 0$
  - $P(X < a) = P(X \le a) = \int_{-\infty}^a f(x)\,dx$
  - $P(X > a) = P(X \ge a) = \int_a^\infty f(x)\,dx$
  - $P(a < X < b) = P(a \le X \le b) = \int_a^b f(x)\,dx$



- Lưu ý:
  - Xác suất để một b.n.n liên tục X nhận một giá trị cụ thể là 0: $P(X = a) = 0$
  - Như vậy có thể có biến cố có xác suất 0 nhưng vẫn có khả năng xảy ra (có A với $P(A) = 0$ nhưng $A \neq \emptyset$)

B.n.n liên tục và hàm mật độ xác suất
Ví dụ

- Cho X là b.n.n liên tục với hàm mật độ xác suất có dạng:

$$f_X(x) = \begin{cases} cx & \text{với } 0 < x < 4 \\ 0 & \text{khác} \end{cases}$$

- Để f_X là hàm mật độ xác suất hợp lệ, ta có điều kiện cho hệ số c là:

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1 \Longrightarrow \int_0^4 cx\,dx = 1 \Longrightarrow c\,\frac{x^2}{2}\Bigg|_{x=0}^{x=4} = 8c = 1 \Longrightarrow c = \frac{1}{8}$$

- Khi đó ta có xác suất:
  - để X nhận giá trị từ 1 đến 2 là: $P(1 \le X \le 2) = \int_1^2 f_X(x)\,dx = \int_1^2 \frac{1}{8} x\,dx = \frac{3}{16}$
  - để X nhận giá lớn hơn 2 là: $P(X > 2) = \int_2^{\infty} f_X(x)\,dx = \int_2^4 \frac{1}{8} x\,dx = \frac{3}{4}$

B.n.n liên tục và hàm mật độ xác suất Phân phối liên tục đều

- B.n.n liên tục X được gọi là có phân phối đều (uniform distribution) trên khoảng $[a, b]$ nếu X có hàm mật độ xác suất là:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{với } a \leq x \leq b \\ 0 & \text{khác} \end{cases}$$

  - X là kết quả của thí nghiệm "chọn ngẫu nhiên một điểm trong khoảng $[a, b]$"
- Ví dụ: một môn học dài 2 giờ, giáo viên điểm danh ngẫu nhiên trong thời gian học, bạn đi trễ t phút. Tính xác suất bạn được điểm danh?
  - Gọi X là thời điểm giáo viên điểm danh thì X là b.n.n liên tục có phân phối đều trên khoảng $[0, 2]$ (giờ). Xác suất bạn được điểm danh là:

$$P\left(X \geq \frac{t}{60}\right) = \int_{t/60}^{\infty} f_X(x)\,dx = \int_{t/60}^{2} \frac{1}{2}dx = \frac{1}{2}\left(2 - \frac{t}{60}\right) = 1 - \frac{t}{120}, \text{với } 0 \leq t \leq 120$$

Hàm phân phối tích lũy

- Hàm phân phối tích lũy (cumulative distribution function) của một b.n.n X là hàm số $F_X: \mathbb{R} \to \mathbb{R}$ được xác định bởi:
$$F_X(x) = P(X \leq x) = P(X \in (-\infty, x])$$
  - F_X xác định phân phối của X
  - Tính chất:
    - Tăng: nếu $x_1 \leq x_2$ thì $F(x_1) \leq F(x_2)$
    - Chuẩn hóa: $\lim_{x \to -\infty} F(x) = 0$ và $\lim_{x \to \infty} F(x) = 1$
    - Liên tục phải: $F(x) = F(x^+) = \lim_{t \to x,\, t > x} F(t)$
  - Dùng F_X để tính các xác suất:
    - $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$
    - $P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$
    - $P(X < x) = F(x^-) = \lim_{t \to x,\, t < x} F(t)$
    - $P(X = x) = P(X \leq x) - P(X < x) = F(x) - F(x^-)$

Hàm phân phối tích lũy

- X là b.n.n rời rạc:

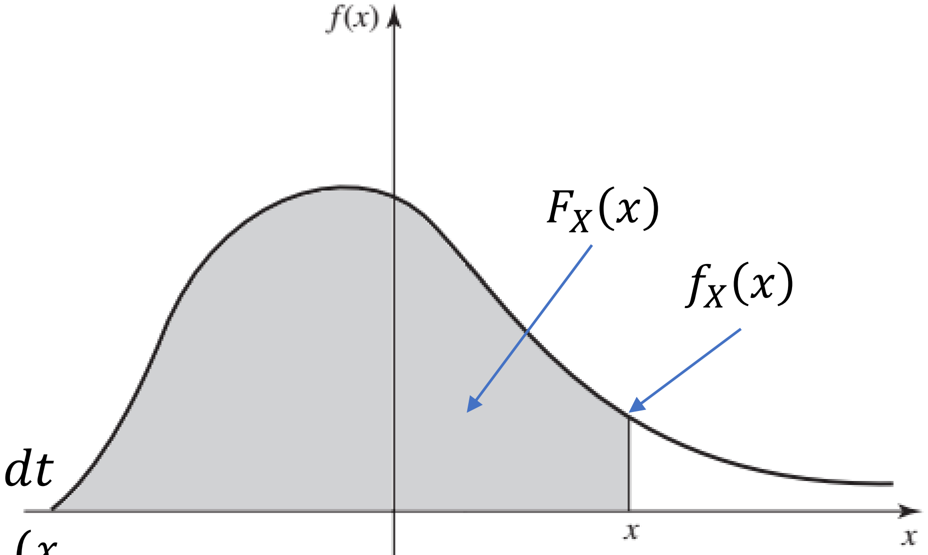
$$F_X(x) = P(X \leq x) = \sum_{t \in Sup(X),\, t \leq x} f_X(t)$$

- Ví dụ: xét thí nghiệm tung một đồng xu (đồng chất) 2 lần, đặt X là số lần được mặt ngửa. Hàm xác suất và hàm phân phối tích lũy của X là:

$$f_X(x) = \begin{cases} 1/4 & \text{nếu } x = 0 \\ 2/4 & \text{nếu } x = 1 \\ 1/4 & \text{nếu } x = 2 \\ 0 & \text{nếu } x \notin \{0, 1, 2\} \end{cases} \quad \text{và} \quad F_X(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ 1/4 & \text{nếu } 0 \leq x < 1 \\ 3/4 & \text{nếu } 1 \leq x < 2 \\ 1 & \text{nếu } 2 \leq x \end{cases}$$

x	0	1	2
$P(X = x)$	1/4	1/2	1/4
$P(X \leq x)$	1/4	3/4	1

Hàm phân phối tích lũy



- X là b.n.n liên tục:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

- Ví dụ: cho X là b.n.n liên tục với hàm mật độ xác suất: $f_X(x) = \begin{cases} \frac{x}{8} & \text{với } 0 < x < 4 \\ 0 & \text{khác} \end{cases}$

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} \int_{-\infty}^{0} 0dt = 0 & \text{nếu } x < 0 \\ \int_{0}^{x} \frac{t}{8}dt = \frac{x^2}{16} & \text{nếu } 0 \leq x < 4 \\ \int_{0}^{4} \frac{t}{8}dt = 1 & \text{nếu } 4 \leq x \end{cases}$$

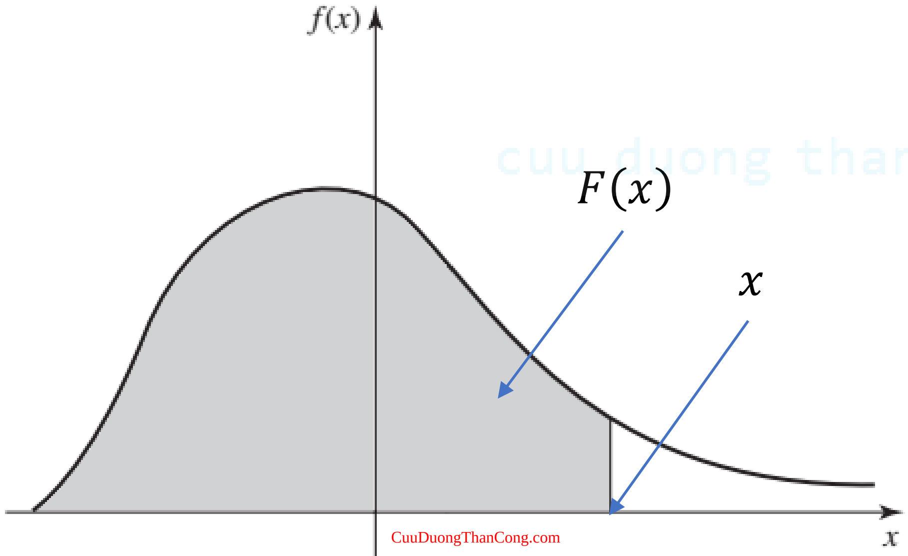
$$P(1 \leq X \leq 2) = F_X(2) - F_X(1^-) = \frac{2^2}{16} - \frac{1^2}{16} = \frac{3}{16} \text{ và } P(X > 2) = 1 - F_X(2) = 1 - \frac{2^2}{16} = \frac{3}{4}$$

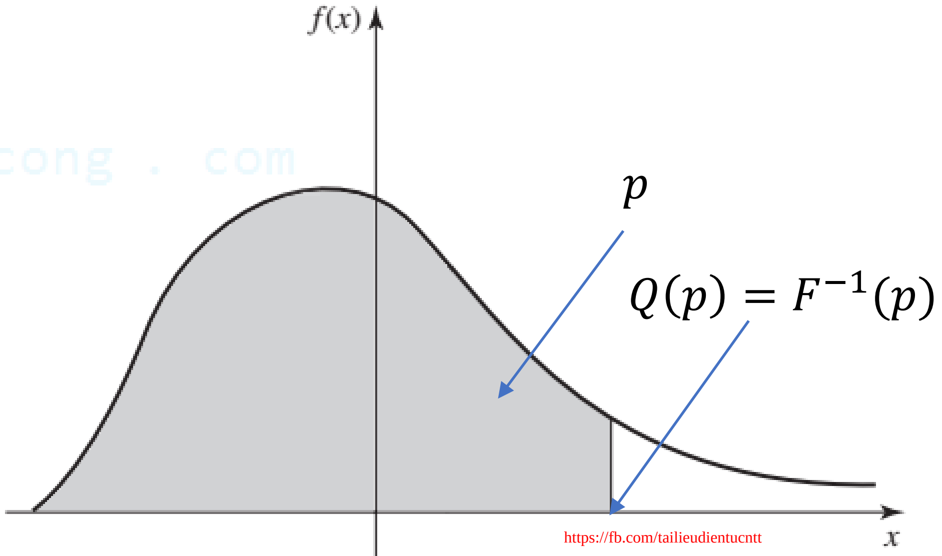
Hàm phân vị

- Cho X là b.n.n với hàm phân phối tích lũy F, hàm phân vị (quantile function) của X là hàm $Q: (0,1) \to \mathbb{R}$, được xác định bởi:

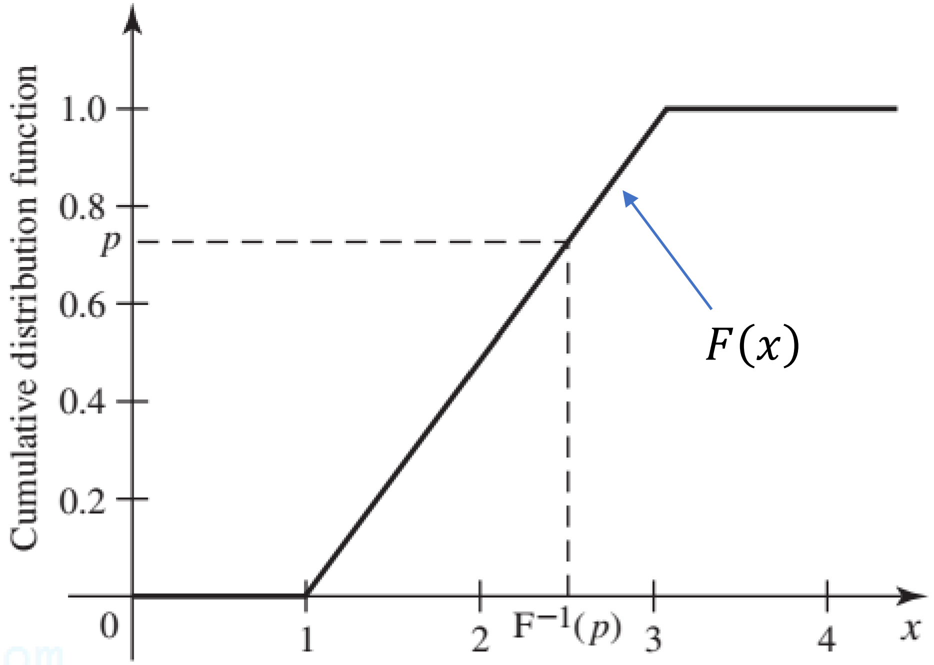
$$Q(p) = \text{"giá trị thực } x \text{ nhỏ nhất sao cho } F(x) \geq p\text{"}$$

  - $Q(p)$ được gọi là phân vị mức p của phân phối của X và thường được kí hiệu là $F^{-1}(p)$
  - Hàm phân vị Q cho biết điểm chia phân phối của X





Hàm phân vị
Ví dụ



- Xét b.n.n liên tục $X \sim \text{Uniform}(1,3)$:
  - X có hàm mật độ xác suất:

$$f(x) = \begin{cases} 1/2 & \text{với } 1 \leq x \leq 3 \\ 0 & \text{khác} \end{cases}$$

  - X có hàm phân phối tích lũy là:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} \int_{-\infty}^{1} 0dt = 0 & \text{nếu } x < 1 \\ \int_{1}^{x} \frac{1}{2}dt = \frac{x-1}{2} & \text{nếu } 1 \leq x < 3 \\ \int_{1}^{3} \frac{1}{2}dt = 1 & \text{nếu } 3 \leq x \end{cases}$$

  - X có hàm phân vị là:

$$Q(p) = x \Leftrightarrow \frac{x-1}{2} = p \Leftrightarrow x = 2p + 1$$
$$\Longrightarrow F^{-1}(p) = Q(p) = 2p + 1, 0 < p < 1$$

Hàm phân vị

- Các phân vị hay dùng:
  - Phân vị phần tư dưới (lower quartile): $Q(25\%) = Q(1/4) = Q(0.25)$
  - Phân vị giữa (median): $Q(50\%) = Q\left(\frac{1}{2}\right) = Q(0.5)$
    - Còn gọi là trung vị: là điểm chia đôi phân phối
  - Phân vị phần tư trên (upper quartile): $Q(75\%) = Q(3/4) = Q(0.75)$
- Ví dụ: $X \sim \text{Uniform}(1, 3)$ có hàm phân vị $Q(p) = 2p + 1, 0 < p < 1$
  - Phân vị phần tư dưới: $Q(25\%) = 2 \times 0.25 + 1 = 1.5$
  - Trung vị: $Q(50\%) = 2 \times 0.5 + 1 = 2$
  - Phân vị phần tư trên: $Q(75\%) = 2 \times 0.75 + 1 = 2.5$