Bài 2
Ôn tập biến ngẫu nhiên

Thống kê máy tính và ứng dụng
(Computational Statistics and Applications)

Vũ Quốc Hoàng (vqhoang@fit.hcmus.edu.vn)

FIT - HCMUS

2025

Nội dung

1. Bài toán thu thập phiếu thưởng
2. Ôn tập biến ngẫu nhiên rời rạc
3. Phương pháp xấp xỉ kỳ vọng bằng mô phỏng
4. Ôn tập biến ngẫu nhiên liên tục
5. Các định lý giới hạn
6. Phương pháp xấp xỉ phân phối bằng mô phỏng

Nội dung

1. Bài toán thu thập phiếu thưởng
2. Ôn tập biến ngẫu nhiên rời rạc
3. Phương pháp xấp xỉ kỳ vọng bằng mô phỏng
4. Ôn tập biến ngẫu nhiên liên tục
5. Các định lý giới hạn
6. Phương pháp xấp xỉ phân phối bằng mô phỏng

Bài toán thu thập phiếu thưởng

Bài toán thu thập phiếu thưởng (coupon collector's problem). Một cửa hàng phát hành n loại phiếu thưởng khác nhau. Mỗi lần khách mua hàng sẽ được cửa hàng phát một tờ phiếu thưởng ngẫu nhiên trong n loại. Khi khách thu thập được đủ n loại phiếu thưởng thì được cửa hàng tặng quà.

Hỏi: khách cần mua hàng bao nhiêu lần từ cửa hàng để được nhận quà?

Nội dung

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)

Biến ngẫu nhiên (random variable) là hàm trên không gian mẫu Ω

- $X : \Omega \to A$, gắn mỗi kết quả $\omega \in \Omega$ một giá trị $X(\omega) \in A$,
- A được gọi là **tập/miền giá trị** của X và thường là tập con của $\mathbb{R}$ (hoặc $\mathbb{R}^d$).

Biến ngẫu nhiên là phương tiện hay được 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

$$(X \in C) = \{\omega \in \Omega : X(\omega) \in C\}.$$

Phân phối của biến ngẫu nhiên

Xét biến ngẫu nhiên X liên quan đến thí nghiệm T có không gian mẫu là Ω. 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".

Biến ngẫu nhiên rời rạc và hàm xác suất

- X được gọi là **biến ngẫu nhiên rời rạc** (discrete random variable) nếu tập giá trị của nó là **rời rạc** (**hữu hạn** (finite) hoặc **vô hạn đếm được** (countably infinite)).
- Với X là biến ngẫu nhiên rời rạc, **hàm xác suất** (probability function, probability mass function) của X là hàm $f : \mathbb{R} \to \mathbb{R}$, được xác định bởi

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

  - Hàm xác suất f cho biết khả năng X nhận một giá trị cụ thể.
  - Hàm xác suất có tính chất: $f(x) \geq 0, \forall x \in \mathbb{R}$ và $\sum_x f(x) = 1$.
- Hàm xác suất xác định phân phối của biến ngẫu nhiên rời rạc

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

Các biến ngẫu nhiên độc lập

Hai biến ngẫu nhiên X, Y được gọi là **độc lập** (independent) nếu với mọi $A, B \subset \mathbb{R}$ ta có

$$P((X \in A) \cap (Y \in B)) = P(X \in A)P(Y \in B).$$

Nghĩa là việc X nhận giá trị nào cũng không ảnh hưởng đến khả năng nhận giá trị nào đó của Y (và ngược lại).

Mệnh đề. Hai biến ngẫu nhiên rời rạc X, Y độc lập khi và chỉ khi

$$P((X = x) \cap (Y = y)) = P(X = x)P(Y = y)$$

với mọi $x, y \in \mathbb{R}$.

Kì vọng của biến ngẫu nhiên

Cho biến ngẫu nhiên rời rạc X với hàm xác suất f, **kì vọng** (mean) của X, kí hiệu $E(X)$, là số thực được tính bởi ("nếu tính được")

$$\mu = E(X) = \sum_x xP(X = x) = \sum_x xf(x).$$

Kì vọng của X là giá trị trung bình của các giá trị mà X có thể nhận với trọng số là xác suất để X nhận các giá trị tương ứng đó.

Cho biến ngẫu nhiên $X : \Omega \to \mathbb{R}$ và hàm số $r : \mathbb{R} \to \mathbb{R}$, ta nói $Y : \Omega \to \mathbb{R}$ là biến ngẫu nhiên **biến đổi** (transformation) từ X qua hàm số r, kí hiệu $Y = r(X)$, nếu Y được xác định bởi

$$Y(\omega) = r(X(\omega)), \omega \in \Omega.$$

Khi đó ta có

$$E(Y) = E(r(X)) = \sum_x r(x)f(x).$$

Phương sai và độ lệch chuẩn của biến ngẫu nhiên

Cho biến ngẫu nhiên rời rạc X với hàm xác suất f và kì vọng $\mu = E(X)$, **phương sai** (variance) của X, kí hiệu $Var(X)$, là số thực được tính bởi ("nếu tính được")

$$\sigma^2 = Var(X) = E\left((X-\mu)^2\right) = \sum_x (x-\mu)^2 f(x).$$

Ta gọi $\sigma = \sqrt{Var(X)}$ là **độ lệch chuẩn** (standard deviation) của X.

Phương sai (và độ lệch chuẩn) phản ánh sự **phân tán** của phân phối của biến ngẫu nhiên.

Mệnh đề. Cho X là biến ngẫu nhiên (có phương sai), ta có

$$Var(X) = E(X^2) - (E(X))^2.$$

Các tính chất quan trọng của kì vọng và phương sai

Cho $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên (có kỳ vọng), ta có

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) \quad \textbf{(linearity of expectation)}$$

Cho X là biến ngẫu nhiên và a, b là các hằng số thực, ta có

1. $E(aX + b) = aE(X) + b$,
2. $Var(aX + b) = a^2 Var(X)$.

Cho X, Y là hai biến ngẫu nhiên độc lập, ta có

1. $E(XY) = E(X)E(Y)$,
2. $Var(X + Y) = Var(X) + Var(Y)$.

Hàm đặc trưng của biến cố

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

$$\mathbb{I}_A(\omega) = \begin{cases} 1 & \text{nếu } \omega \in A, \\ 0 & \text{nếu } \omega \notin A. \end{cases}$$

Hàm đặc trưng giúp khảo sát biến cố như là một biến ngẫu nhiên.

Mệnh đề. Với mọi biến cố A ta có

$$E(\mathbb{I}_A) = P(A).$$

Phân phối Bernoulli

Biến ngẫu nhiên rời rạc X được gọi là có **phân phối Bernoulli** (Bernoulli distribution) với tham số p $(0 \leq p \leq 1)$, kí hiệu $X \sim \text{Bernoulli}(p)$, nếu X có tập giá trị là $\{0, 1\}$ và

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

Khi đó, X có kì vọng $E(X) = p$ và phương sai $Var(X) = p(1 - p)$.

Xét thí nghiệm tung một đồng xu với xác suất ra ngửa p, gọi X là "số lần được ngửa" thì $X \sim \text{Bernoulli}(p)$. Trường hợp đồng xu đồng chất thì $X \sim \text{Bernoulli}(0.5)$.

Xét thí nghiệm T với biến cố A có $P(A) = p$, khi đó $\mathbb{I}_A \sim \text{Bernoulli}(p)$.

Phân phối nhị thức

Biến ngẫu nhiên rời rạc X được gọi là có **phân phối nhị thức** (binomial distribution) với tham số n $(n \in \mathbb{N})$, p $(0 \leq p \leq 1)$, kí hiệu $X \sim \mathcal{B}(n, p)$, nếu X có tập giá trị là $\{0, 1, ..., n\}$ và

$$f(x) = P(X = x) = C_n^x p^x (1 - p)^{n-x}, x \in \{0, 1, ..., n\}.$$

Khi đó, X có kì vọng $E(X) = np$ và phương sai $Var(X) = np(1 - p)$.

Cho 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 "số lần A xảy ra" thì $X \sim \mathcal{B}(n, p)$.

Mệnh đề. Nếu $X_1, X_2, ..., X_n$ là các biến ngẫu nhiên **độc lập và cùng phân phối** (independent and identically distributed - iid) Bernoulli với tham số p và $X = \sum_{i=1}^{n} X_i$ thì $X \sim \mathcal{B}(n, p)$.

Phân phối hình học

Biến ngẫu nhiên rời rạc X được gọi là có **phân phối hình học** (geometric distribution) với tham số p $(0 < p \leq 1)$, kí hiệu $X \sim \text{Geometric}(p)$, nếu X có tập giá trị là $\{1, 2, ...\}$ và

$$f(x) = P(X = x) = (1-p)^{x-1}p, x \in \{1, 2, ...\}.$$

Khi đó, X có kì vọng $E(X) = \frac{1}{p}$ và phương sai $Var(X) = \frac{1-p}{p^2}$.

Cho thí nghiệm T với biến cố A có $P(A) = p$. Xét thí nghiệm R "thực hiện T lặp lại nhiều lần độc lập cho đến khi A xảy ra thì dừng", gọi X là "số lần thực hiện" thì $X \sim \text{Geometric}(p)$.

Mệnh đề (tính không nhớ - memoryless). Cho $X \sim \text{Geometric}(p)$, với mọi $n = 1, 2, ...$ và mọi $k = 0, 1, ...$ ta có

$$P(X = k + n | X > k) = P(X = n).$$

Phân phối Poisson

Biến ngẫu nhiên rời rạc X được gọi là có **phân phối Poisson** (Poisson distribution) với tham số $\lambda > 0$, kí hiệu $X \sim \text{Poisson}(\lambda)$, nếu X có tập giá trị là $\{0, 1, 2, ...\}$ và

$$f(x) = P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, x \in \{0, 1, 2, ...\}.$$

Khi đó, X có kì vọng $E(X) = \lambda$ và phương sai $Var(X) = \lambda$.

Mệnh đề. Cho $X_1 \sim \text{Poisson}(\lambda_1), X_2 \sim \text{Poisson}(\lambda_2)$ và X_1, X_2 độc lập, ta có $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Mệnh đề. Cho $X \sim \mathcal{B}(n, p = \frac{\lambda}{n})$ với hằng số $\lambda > 0$, ta có

$$\lim_{n\to\infty} f_X(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \text{ với mọi } x \in \{0, 1, 2, ...\}.$$

Khi n lớn và p nhỏ, ta có thể xấp xỉ phân phối $\mathcal{B}(n, p)$ bằng phân phối $\text{Poisson}(\lambda)$.

Nội dung

Phương pháp xấp xỉ kỳ vọng bằng mô phỏng

Để xấp xỉ kì vọng $E(X)$ của một biến ngẫu nhiên X liên quan đến thí nghiệm T, ta có thể dùng phương pháp thống kê như sau

- Thực hiện lặp lại N lần (độc lập) thí nghiệm T, ghi nhận các giá trị mà X nhận $x_1, x_2, ..., x_N$ (còn gọi là **mẫu dữ liệu** - sample), và tính **trung bình mẫu**

$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N}.$$

- Khi N đủ lớn, ta có $\bar{x} \approx E(X)$.
- Việc thực hiện lặp lại nhiều lần thí nghiệm T có thể được **mô phỏng** (simulation) trên máy tính.

Bài toán thu thập phiếu thưởng - Tính toán chính xác

Gọi X là số lần khách cần mua hàng để được nhận quà, tức là số lần cần mua hàng từ đầu cho đến khi thu thập vừa đủ n loại phiếu thưởng. Gọi X_i là số lần cần mua hàng từ lúc vừa đã có $i-1$ loại phiếu thưởng cho đến khi thu thập thêm được một loại mới để có vừa đúng i loại phiếu thưởng ($i = 1, 2, ..., n$).
Ta có $X = \sum_{i=1}^{n} X_i$ và X_i có phân phối hình học với tham số

$$p_i = \frac{n-(i-1)}{n} = \frac{n-i+1}{n} \; (i = 1, 2, ..., n).$$

Từ đó ta có

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \frac{n}{n-i+1} = n\sum_{i=1}^{n} \frac{1}{i} = nH_n,$$

với $H_n = \sum_{i=1}^{n} \frac{1}{i}$ được gọi là **số điều hòa** (harmonic number) thứ n.

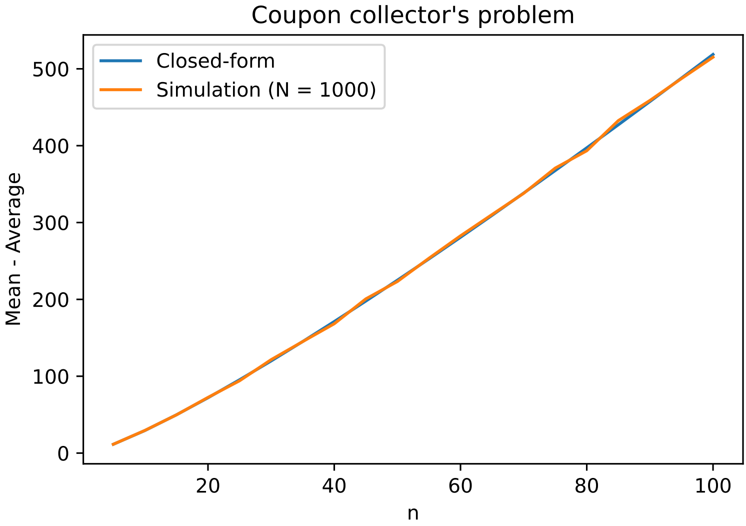
Bài toán thu thập phiếu thưởng - Mô phỏng

```python
def num_buy_to_win(n):
    coupons = []
    while len(set(coupons)) < n:
        coupons.append(random.randint(1, n))
    return len(coupons)

def average(n, N, X):
    m = sum(X(n) for _ in range (N))
    return m/N

average(10, 1000, num_buy_to_win)
#29.175
```

Bài toán thu thập phiếu thưởng - Kết quả



Nội dung

1. Bài toán thu thập phiếu thưởng
2. Ôn tập biến ngẫu nhiên rời rạc
3. Phương pháp xấp xỉ kỳ vọng bằng mô phỏng
4. **Ôn tập biến ngẫu nhiên liên tục**
5. Các định lý giới hạn
6. Phương pháp xấp xỉ phân phối bằng mô phỏng

Biến ngẫu nhiên liên tục và hàm mật độ xác suất

X được gọi là **biến ngẫu liên tục** (continuous random variable) nếu có hàm số không âm $f : \mathbb{R} \to \mathbb{R}$ sao cho với mọi $[a, b] \subset \mathbb{R}$,

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

- 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 $\mathbb{R}$

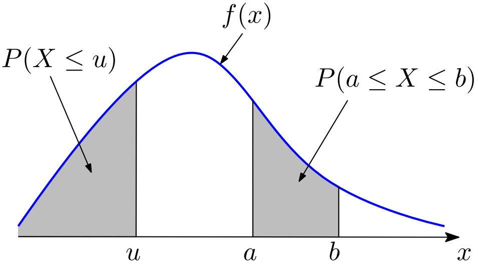
$$P(a \leq X \leq a + \epsilon) = \int_a^{a+\epsilon} f(x)dx \approx \epsilon f(a) \text{ khi } \epsilon \text{ rất nhỏ.}$$

- Hàm mật độ xác suất có tính chất: $f(x) \geq 0, \forall x \in \mathbb{R}$ và $\int_{-\infty}^{\infty} f(x)dx = 1$.

Hàm mật độ xác suất (tt)

Hàm mật độ xác suất xác định phân phối của biến ngẫu nhiên liên tục

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



Từ nhận xét $P(X = u) = 0$, ta thấy tồn tại các biến cố dù có xác suất là 0 nhưng vẫn có thể xảy ra (có E với $P(E) = 0$ nhưng $E \neq \emptyset$).

Hàm phân phối

Hàm phân phối (tích lũy) (distribution function, cumulative distribution function) của một biến ngẫu nhiên X là hàm số $F:\mathbb{R}\to\mathbb{R}$ được xác định bởi

$$F(x) = P(X \leq x) = \begin{cases} \sum_{t\leq x} f(t) & \text{nếu } X \text{ rời rạc,} \\ \int_{-\infty}^{x} f(t)dt & \text{nếu } X \text{ liên tục.} \end{cases}$$

F xác định phân phối của X.

Hàm phân phối F có các tính chất

1. Tăng: nếu $x_1 \leq x_2$ thì $F(x_1) \leq F(x_2)$,
2. Chuẩn hóa: $lim_{x\to-\infty}F(x) = 0$ và $lim_{x\to\infty}F(x) = 1$,
3. Liên tục phải: $F(x) = F(x^+) = lim_{t\to x, t>x}F(t)$.
4. Nếu X là biến ngẫu nhiên liên tục thì F là hàm liên tục và nếu F có đạo hàm tại x thì $F'(x) = f(x)$.

Hàm phân phối đồng thời

Hàm phân phối đồng thời (joint distribution function) của hai biến ngẫu nhiên X, Y là hàm số $F_{XY} : \mathbb{R}^2 \to \mathbb{R}$ được xác định bởi

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) \ (x, y \in \mathbb{R}).$$

Mệnh đề. Hai biến ngẫu nhiên X, Y độc lập khi và chỉ khi $F_{XY}(x, y) = F_X(x)F_Y(y)$ với mọi $x, y \in \mathbb{R}$.

Hai biến ngẫu nhiên X, Y được gọi là **liên tục đồng thời** (jointly continuous) nếu có hàm số không âm $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ sao cho với mọi $C \in \mathbb{R}^2$ ta có

$$P((X, Y) \in C) = \iint_C f_{XY}(x, y)dxdy.$$

Mệnh đề. Hai biến ngẫu nhiên liên tục đồng thời X, Y độc lập khi và chỉ khi $f_{XY}(x, y) = f_X(x)f_Y(y)$ với mọi $x, y \in \mathbb{R}$.

Kì vọng và phương sai

Cho biến ngẫu nhiên liên tục X với hàm mật độ xác suất f

- **Kì vọng** (mean) của X được tính bởi

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

- **Phương sai** (variance) của X được tính bởi

$$\sigma^2 = Var(X) = E\left((X-\mu)^2\right) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx,$$

- Với hàm số $r : \mathbb{R} \to \mathbb{R}$ và $Y = r(X)$

$$E(Y) = E(r(X)) = \int_{-\infty}^{\infty} r(x) f(x) dx.$$

Phân phối đều

Biến ngẫu nhiên liên tục X được gọi là có **phân phối đều** (uniform distribution) trên $[a, b]$ với $a < b$, kí hiệu $X \sim \mathcal{U}(a, b)$, nếu X có tập giá trị là $[a, b]$ và

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{nếu } a \leq x \leq b, \\ 0 & \text{khác.} \end{cases}$$

Khi đó, X có kì vọng $E(X) = \frac{a+b}{2}$ và phương sai $Var(X) = \frac{(b-a)^2}{12}$.

Gọi X là kết quả của thí nghiệm "chọn ngẫu nhiên một điểm trong khoảng $[a, b]$" thì $X \sim \mathcal{U}(a, b)$.

Mệnh đề. Cho $X \sim \mathcal{U}(a, b)$ và $d \in (a, b)$, phân phối của X khi biết $X \leq d$ là phân phối đều trên $[a, d]$, thường kí hiệu $(X|X \leq d) \sim \mathcal{U}(a, d)$.

Phân phối mũ

Biến ngẫu nhiên liên tục X được gọi là có **phân phối mũ** (exponential distribution) với tham số $\lambda > 0$, kí hiệu $X \sim \text{Exp}(\lambda)$, nếu X có tập giá trị là $[0, \infty)$ và

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{nếu } x \geq 0, \\ 0 & \text{khác.} \end{cases}$$

Khi đó, X có kì vọng $E(X) = \frac{1}{\lambda}$, phương sai $Var(X) = \frac{1}{\lambda^2}$ và hàm phân phối

$$F(x) = 1 - e^{-\lambda x}, x \geq 0.$$

Phân phối mũ có thể được xem như là "phiên bản liên tục" của phân phối hình học.

Mệnh đề (tính không nhớ - memoryless). Cho $X \sim \text{Exp}(\lambda)$, với mọi $t, s \geq 0$ ta có

$$P(X > t + s | X > s) = P(X > t).$$

Phân phối chuẩn

Biến ngẫu nhiên liên tục X được gọi là có **phân phối chuẩn** (normal distribution) với trung bình μ và phương sai σ^2 ($\sigma > 0$), kí hiệu $X \sim \mathcal{N}(\mu, \sigma^2)$, nếu X có hàm mật độ xác suất

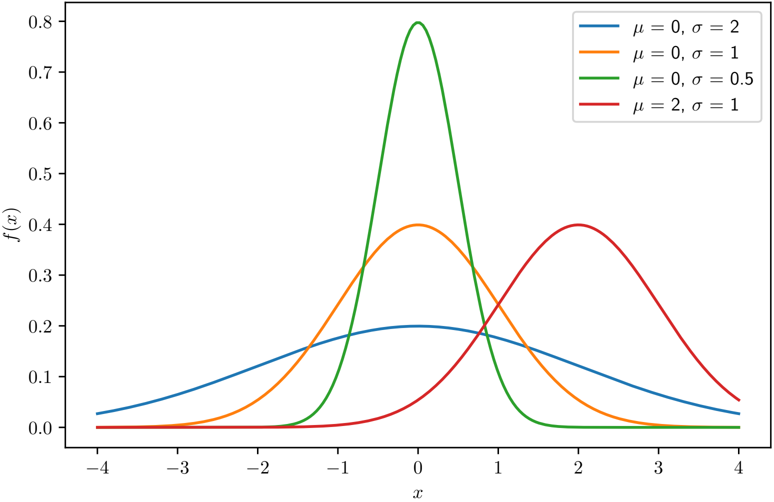
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}.$$

Khi đó, X có kì vọng $E(X) = \mu$ và phương sai $Var(X) = \sigma^2$.

Trường hợp $Z \sim \mathcal{N}(0, 1)$ thì Z được gọi là có **phân phối chuẩn tắc** (standard normal distribution). Hàm mật độ xác suất và hàm xác suất của Z thường được kí hiệu lần lượt là ϕ, Φ, tức là

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt.$$

Phân phối chuẩn (tt)



Phân phối chuẩn (tt)

Các tính chất quan trọng của phân phối chuẩn

1. Nếu $X \sim \mathcal{N}(\mu, \sigma^2)$ và $Y = aX + b$ $(a \neq 0)$ thì $X \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$,
2. Nếu $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ và X_1, X_2 độc lập thì $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$,
3. Nếu $Z \sim \mathcal{N}(0, 1)$ và $X = \sigma Z + \mu$ thì $X \sim \mathcal{N}(\mu, \sigma^2)$,
4. Nếu $X \sim \mathcal{N}(\mu, \sigma^2)$ và $Z = \frac{X-\mu}{\sigma}$ thì $Z \sim \mathcal{N}(0, 1)$, từ đó

$$F_X(x) = P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = F_Z\left(\frac{x-\mu}{\sigma}\right)$$
$$= \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Nội dung

Bất đẳng thức Markov và Chebyshev

Bất đẳng thức Markov (Markov's inequality). Nếu X là biến ngẫu nhiên không âm thì với mọi $a > 0$,

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

Bất đẳng thức Chebyshev (Chebyshev's inequality). Nếu X là biến ngẫu nhiên có kì vọng μ và phương sai σ^2 thì với mọi $k > 0$,

$$P\left(\frac{|X - \mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}.$$

Luật số lớn

Dạng mạnh của luật số lớn (strong law of large numbers). Cho $X_1, X_2, \ldots$ độc lập và cùng phân phối với kỳ vọng μ, ta có

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu \text{ (với xác suất 1).}$$

Từ đó, với n "đủ lớn", ta có $\mu \approx \frac{1}{n} \sum_{i=1}^{n} X_i$. Tổng quát hơn, cho f là hàm giá trị thực và $X_1, X_2, \ldots$ độc lập và cùng phân phối như X, ta có

$$E(f(X)) \approx \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Đặc biệt, cho biến cố A, ta có

$$P(A) = E(\mathbb{I}_A) \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_A(X_i).$$

Định lý giới hạn trung tâm

Định lý giới hạn trung tâm (central limit theorem). Cho $X_1, X_2, \ldots$ độc lập với cùng kỳ vọng μ và phương sai σ^2 hữu hạn, ta có

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1),$$

trong đó, $\xrightarrow{d}$ là kí hiệu **hội tụ theo phân phối** (convergence in distribution), nghĩa là, với mọi $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \leq x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Từ đó, với n "đủ lớn", $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$ "xấp xỉ" phân phối chuẩn tắc.

Nội dung

Phương pháp xấp xỉ phân phối bằng mô phỏng - Biến ngẫu nhiên rời rạc

Để xấp xỉ hàm xác suất f_X của một biến ngẫu nhiên rời rạc X liên quan đến thí nghiệm T, ta có thể dùng phương pháp thống kê như sau

- Thực hiện lặp lại N lần (độc lập) thí nghiệm T và tính các tần suất p_x của biến cố "X nhận giá trị x".
- Khi N đủ lớn, ta có $p_x \approx P(X = x) = f_X(x)$.
- Việc thực hiện lặp lại nhiều lần thí nghiệm T có thể được mô phỏng trên máy tính.

Phương pháp xấp xỉ phân phối bằng mô phỏng - Biến ngẫu nhiên liên tục

Để xấp xỉ hàm mật độ xác suất f_X của một biến ngẫu nhiên liên tục X liên quan đến thí nghiệm T, ta có thể dùng phương pháp thống kê như sau

- Thực hiện lặp lại N lần (độc lập) thí nghiệm T, ghi nhận các giá trị mà X nhận $x_1, x_2, ..., x_N$ (còn gọi là **mẫu dữ liệu** - sample).
- Khi N đủ lớn, ta có thể dùng **histogram** hoặc **ước lượng mật độ nhân** (kernel density estimation, KDE) trên mẫu để xấp xỉ f_X.
- Việc thực hiện lặp lại nhiều lần thí nghiệm T có thể được mô phỏng trên máy tính.

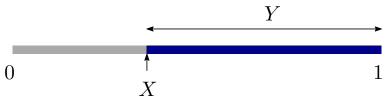
Phương pháp xấp xỉ phân phối bằng mô phỏng - Ví dụ 1

Bài toán. Chọn ngẫu nhiên một điểm trên một thanh có chiều dài 1 đơn vị, cắt tại điểm đó thành hai đoạn và giữ lại đoạn dài hơn. Tính kì vọng và tìm phân phối của chiều dài đoạn giữ lại.

Giải. Gọi X là vị trí ngẫu nhiên chọn trên thanh thì $X \sim \mathcal{U}(0,1)$. Do đó X là biến ngẫu nhiên liên tục với hàm mật độ xác suất

$$f_X(x) = \begin{cases} 1 & \text{nếu } 0 \le x \le 1, \\ 0 & \text{khác.} \end{cases}$$

Gọi Y là chiều dài của đoạn được giữ lại (tức là đoạn dài hơn) thì $Y = \max\{X, 1 - X\}$.



Phương pháp xấp xỉ phân phối bằng mô phỏng - Ví dụ 1 (tt)

Ta có kì vọng của chiều dài đoạn giữ lại

$$\begin{aligned}
E(Y) &= E\left(\max\{X, 1-X\}\right) \\
&= \int_{-\infty}^{\infty} \max\{x, 1-x\} f_X(x)dx = \int_0^1 \max\{x, 1-x\}dx \\
&= \int_0^{1/2} \max\{x, 1-x\}dx + \int_{1/2}^1 \max\{x, 1-x\}dx \\
&= \int_0^{1/2} (1-x)dx + \int_{1/2}^1 xdx \\
&= \frac{3}{4}.
\end{aligned}$$

Phương pháp xấp xỉ phân phối bằng mô phỏng - Ví dụ 1 (tt)

Hàm phân phối của $Y = \max\{X, 1-X\}$

$$F_Y(y) = P(Y \leq y) = P(\max\{X, 1-X\} \leq y), y \in \mathbb{R}.$$

Xét các trường hợp của y

1. $y < 1/2$: $(\max\{X, 1-X\} \leq y) = \varnothing$ vì $0 \leq X \leq 1$ nên $1/2 \leq \max\{X, 1-X\}$,
$$P\left(\max\{X, 1-X\} \leq y\right) = P(\varnothing) = 0.$$
2. $1/2 \leq y \leq 1$: $(\max\{X, 1-X\} \leq y) = (1-y \leq X \leq y)$,
$$P\left(\max\{X, 1-X\} \leq y\right) = P\left(1-y \leq X \leq y\right) = \int_{1-y}^{y} dy = 2y.$$
3. $y > 1$: $(\max\{X, 1-X\} \leq y) = \Omega$ vì $0 \leq X \leq 1$ nên $\max\{X, 1-X\} \leq 1$,
$$P\left(\max\{X, 1-X\} \leq y\right) = P(\Omega) = 1.$$

Phương pháp xấp xỉ phân phối bằng mô phỏng - Ví dụ 1 (tt)

Từ đó ta có

$$F_Y(y) = \begin{cases} 0 & \text{nếu } y < 1/2, \\ 2y & \text{nếu } 1/2 \leq y \leq 1, \\ 1 & \text{nếu } 1 < y. \end{cases}$$

Lấy đạo hàm của hàm phân phối, ta có hàm mật độ xác suất của Y là

$$f_Y(y) = F'_Y(y) = \begin{cases} 2 & \text{nếu } 1/2 \leq x \leq 1, \\ 0 & \text{khác.} \end{cases}$$

Như vậy Y có phân phối đều trên đoạn $[1/2, 1]$, tức là $Y \sim \mathcal{U}(1/2, 1)$. Lưu ý, vì $Y \sim \mathcal{U}(1/2, 1)$ nên $E(Y) = \frac{1/2+1}{2} = \frac{3}{4}$.

Phương pháp xấp xỉ phân phối bằng mô phỏng - Ví dụ 1 (tt)

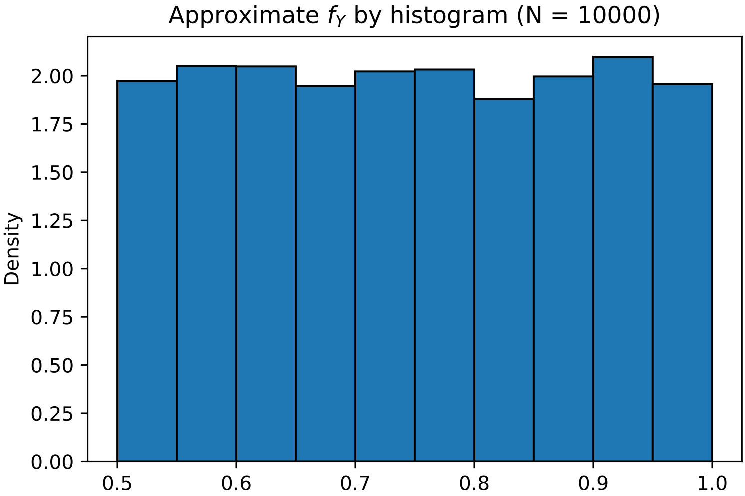
```
def greater_len(N):
    X = np.random.uniform(size=N)
    Y = np.maximum(X, 1 - X)
    return Y

N = 10000

np.mean(greater_len(N))
#0.7499721269808018

plt.hist(greater_len(N), density=True,
   edgecolor="black")
```

Phương pháp xấp xỉ phân phối bằng mô phỏng - Ví dụ 1 (tt)



Phương pháp xấp xỉ phân phối bằng mô phỏng - Ví dụ 2

Bài toán. Cho $X_1, X_2, ..., X_n$ là n biến ngẫu nhiên độc lập và cùng phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$. Đặt

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{và} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

($X_1, ..., X_n$ thường được gọi là một mẫu ngẫu nhiên cỡ n, $\bar{X}$ là trung bình mẫu và S^2 là phương sai mẫu.)
Tìm phân phối của các biến ngẫu nhiên $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ và $\frac{\bar{X}-\mu}{S/\sqrt{n}}$.

Trả lời: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ có phân phối chuẩn tắc $\mathcal{N}(0,1)$ và $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ có **phân phối Student** (Student's t-distribution) với $n-1$ bậc tự do. (`https://en.wikipedia.org/wiki/Student%27s_t-distribution`.)

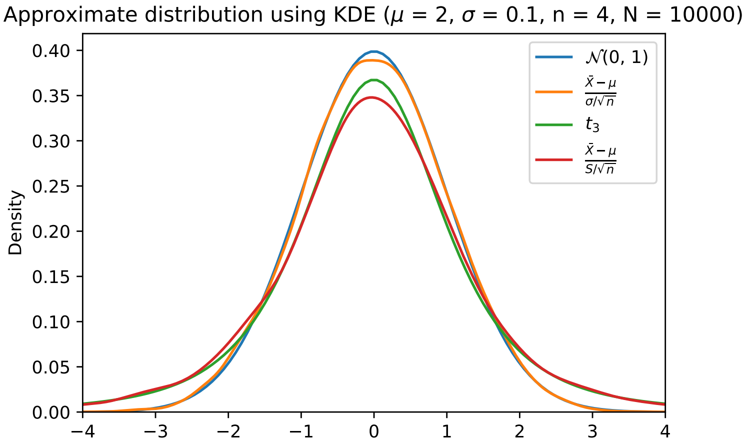
Phương pháp xấp xỉ phân phối bằng mô phỏng - Ví dụ 2 (tt)

```
def sample(mu, sigma, n, N):
    X = np.random.normal(mu, sigma, size=(N, n
        ))
    X_bar = np.mean(X, axis=1)
    S2 = np.var(X, axis=1, ddof=1)
    return X_bar, S2

X_bar, S2 = sample(mu, sigma, n, N)

plt.plot(x, scipy.stats.norm.pdf(x))
sns.kdeplot((X_bar - mu)/(sigma/np.sqrt(n)))
plt.plot(x, scipy.stats.t.pdf(x, n - 1))
sns.kdeplot((X_bar - mu)/(np.sqrt(S2)/np.sqrt(
   n)))
```

Phương pháp xấp xỉ phân phối bằng mô phỏng - Ví dụ 2 (tt)



Tài liệu tham khảo

Jupyter Notebook đi kèm.

Chapter 2. Sheldon M. Ross. *Simulation*. Elsevier, 2023.

Chapter 3-5. H. Pishro-Nik. *"Introduction to probability, statistics, and random processes"*, available at `https://www.probabilitycourse.com`. Kappa Research LLC, 2014.