

Mục lục

1	Biến cố ngẫu nhiên và xác suất	1
1.1	Khái niệm	1
1.2	Mô hình xác suất cổ điển	3
1.3	Mô hình xác suất hình học	7
1.4	Công thức cộng và nhân xác suất	9
1.5	Công thức xác suất đầy đủ và công thức Bayes	15
1.6	Dãy thử Bernoulli	17
2	Đại lượng ngẫu nhiên	22
2.1	Khái niệm	23
2.2	Hàm phân bố xác suất	26
2.3	Hàm phụ thuộc đại lượng ngẫu nhiên	28
2.4	Các đặc trưng số của đại lượng ngẫu nhiên	31
2.5	Các phân bố xác suất thường gặp	37
3	Véc tơ ngẫu nhiên	45
3.1	Khái niệm	46
3.2	Hàm phân bố xác suất đồng thời	50
3.3	Xác định luật phân bố thành phần	52
3.4	Các đại lượng ngẫu nhiên độc lập	53
3.5	Phân bố có điều kiện	55
3.6	Tổng các đại lượng ngẫu nhiên	59
3.7	Momen tương quan và Hệ số tương quan	63
4	Các định lý giới hạn	70
5	Mẫu và phân bố mẫu	74
5.1	Mẫu ngẫu nhiên đơn giản	74
5.2	Các đặc trưng mẫu	75
5.3	Các phân bố thường gặp trong thống kê	80

5.4 Phân bố mẫu	82
6 Ước lượng tham số	84
7 Kiểm định giả thuyết thống kê	86
7.1 Khái niệm	86
7.2 Kiểm định giả thuyết về giá trị trung bình và xác suất	87
7.3 Tiêu chuẩn phù hợp χ^2	94
8 Tương quan và hồi quy	107
8.1 Hồi quy	107
8.2 Hồi quy tuyến tính	108
8.3 Dữ liệu lớn và học máy	112
1 Biến cố ngẫu nhiên và xác suất	122
2 Đại lượng ngẫu nhiên	125
3 Vectơ ngẫu nhiên	129
4 Các định lý giới hạn	136
5 Mẫu và phân bố mẫu	138
7 Kiểm định giả thuyết thống kê	140
8 Tương quan và hồi quy	144
Phụ lục	145
A Python	146
A.1 Thư viện, môđun, phương thức	146

Chương 4

Các định lý giới hạn

Bất đẳng thức Chebyshev: Cho X là đại lượng ngẫu nhiên có kỳ vọng và phương sai hữu hạn. Khi đó $\forall \varepsilon > 0$

$$P(|X - EX| \geq \varepsilon) \leq \frac{DX}{\varepsilon^2}. \quad (4.1)$$

Định lý 4.1 (Luật số lớn Chebyshev). Cho X_1, X_2, \dots là dãy đại lượng ngẫu nhiên độc lập, cùng phân bố, có kỳ vọng và phương sai hữu hạn: $EX_n = a \quad \forall n$. Khi đó $\forall \varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - a\right| < \varepsilon\right) = 1.$$

Định nghĩa 4.1. Dãy đại lượng ngẫu nhiên X_1, X_2, \dots gọi là hội tụ theo xác suất tới đại lượng ngẫu nhiên X , ký hiệu $X_n \xrightarrow{p} X$ nếu:

$$\lim_{n \rightarrow \infty} P(|X_n - X| < \varepsilon) = 1 \quad \forall \varepsilon > 0.$$

Ký hiệu $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$, ta có

$$(4.1) \Leftrightarrow \bar{X} \xrightarrow{p} a. \quad (4.2)$$

Áp dụng: Đặt m = số lần A xảy ra trong n lần thử. Ký hiệu

$$X_i = \begin{cases} 1, & \text{nếu } A \text{ xảy ra ở lần thử thứ } i \\ 0, & \text{ngược lại} \end{cases}$$

$$\text{thì } m = X_1 + X_2 + \dots + X_n \Rightarrow \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{m}{n}.$$

Mặt khác, X_1, X_2, \dots là dãy đại lượng ngẫu nhiên độc lập, cùng bảng phân bố

$$\begin{array}{c|cc} X_n & 0 & 1 \\ \hline P & 1-p & p \end{array} \Rightarrow EX_n = p, \quad DX_n = p(1-p).$$

Do đó

$$(4.2) \Leftrightarrow \frac{m}{n} \xrightarrow{p} p. \quad (4.3)$$

Ví dụ 4.1. Xét $I = \int_0^1 \varphi(x) dx$. Đặt $f(x) = \begin{cases} 1, & x \in [0, 1] \\ 0, & x \notin [0, 1] \end{cases}$, là hàm mật độ của $X \sim U[0, 1]$.

Khi đó $I = \int_{-\infty}^{\infty} \varphi(x)f(x)dx = E[\varphi(X)]$.

Hãy tính gần đúng $\int_0^1 x^2 dx = \frac{1}{3}$ bởi $\overline{\varphi(X)} = \frac{\varphi(X_1) + \varphi(X_2) + \dots + \varphi(X_n)}{n}$ (thực nghiệm 10 lần với số lần thử $n = 1\,000, 10\,000, 100\,000, 1\,000\,000$).

Giải.

		Số lần thử			
		1000	10 000	100 000	1 000 000
Lần thực nghiệm	1	0.337149	0.334444	0.333024	0.332932
	2	0.337935	0.33325	0.332922	0.333245
	3	0.342481	0.333367	0.33257	0.333229
	4	0.339492	0.334291	0.333543	0.333357
	5	0.347686	0.336503	0.332915	0.333079
	6	0.323709	0.333164	0.332536	0.33376
	7	0.328597	0.335589	0.332178	0.333416
	8	0.339541	0.32789	0.333202	0.33374
	9	0.317256	0.332914	0.334396	0.332986
	10	0.327741	0.331344	0.335878	0.333006

□

Ví dụ 4.2. Chọn ngẫu nhiên một điểm trong hình vuông cho trước.

- Tìm xác suất để chọn được điểm ở miền trong hình tròn nội tiếp hình vuông.
- Xấp xỉ π thông qua tần suất $\frac{m}{n}$ của xác suất trên (thực nghiệm 10 lần với mỗi số lần thử $n = 1\,000, 10\,000, 100\,000, 1\,000\,000$).

Giải. a) Đặt a là độ dài cạnh hình vuông. Khi đó bán kính hình tròn nội tiếp hình vuông là $r = \frac{a}{2}$. Xác suất để chọn được điểm ở miền trong hình tròn nội tiếp hình vuông

$$p = \frac{S(\text{hình tròn})}{S(\text{hình vuông})} = \frac{\pi r^2}{a^2} = \frac{\pi}{4}.$$

$$b) \ p = \frac{\pi}{4} \approx \frac{m}{n} \Rightarrow \pi \approx 4 \frac{m}{n}.$$

Trong mặt phẳng Oxy , xét hình vuông $\Omega = \{(x, y) \mid -1 < x, y < 1\}$. Hình tròn nội tiếp hình vuông $A = \{(x, y) \mid x^2 + y^2 < 1\}$. Một số giá trị gần đúng của π theo số lần thử n :

Lần thực nghiệm \ Số lần thử	1000	10 000	100 000	1 000 000
1	3.164	3.1008	3.1388	3.14194
2	3.216	3.1616	3.14	3.1421
3	3.188	3.136	3.14496	3.13982
4	3.164	3.1248	3.13648	3.14131
5	3.132	3.1356	3.14908	3.14457
6	3.132	3.1404	3.1348	3.14414
7	3.136	3.1268	3.14708	3.14431
8	3.108	3.1036	3.14056	3.14123
9	3.168	3.1436	3.1424	3.1404
10	3.2	3.1636	3.13768	3.13891

□

Định lý 4.2 (Định lý giới hạn trung tâm). Cho X_1, X_2, \dots là dãy đại lượng ngẫu nhiên độc lập, cùng phân bố, có kỳ vọng và phương sai hữu hạn: $EX_n = a, DX_n = \sigma^2 \forall n$.

Khi đó $\frac{\bar{X} - a}{\sigma} \sqrt{n}$ có phân bố tiệm cận chuẩn $N(0, 1)$, tức là

$$\lim_{n \rightarrow \infty} P\left(\frac{\bar{X} - a}{\sigma} \sqrt{n} < x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt, \quad \forall x. \quad (4.4)$$

Áp dụng: $m =$ số lần A xảy ra trong n lần thử. $\frac{\bar{X} - a}{\sigma} \sqrt{n} = \frac{\frac{m}{n} - p}{\sqrt{p(1-p)}} \sqrt{n} = \frac{m - np}{\sqrt{np(1-p)}}$.

Ta có

$$\begin{aligned}
 P(k_1 \leq m \leq k_2) &= \sum_{k=k_1}^{k_2} C_n^k p^k (1-p)^{n-k} \\
 &= P\left[\frac{k_1 - np}{\sqrt{np(1-p)}} \leq \frac{m - np}{\sqrt{np(1-p)}} \leq \frac{k_2 - np}{\sqrt{np(1-p)}}\right] \\
 &\approx P\left[\frac{k_1 - np}{\sqrt{np(1-p)}} \leq N(0, 1) \leq \frac{k_2 - np}{\sqrt{np(1-p)}}\right] \\
 &= \Phi\left[\frac{k_2 - np}{\sqrt{np(1-p)}}\right] - \Phi\left[\frac{k_1 - np}{\sqrt{np(1-p)}}\right].
 \end{aligned} \quad (4.5)$$

Ví dụ 4.3. Khi giải Bài tập 1.29, ta cần tính xác suất có biểu thức $\sum_{k=51}^{400} C_{400}^k 0.1^k 0.9^{400-k} = 0.04364$. Tính gần đúng tổng trên.

Giải. $n = 400, p = 0.1$. Ta có

$$\begin{aligned} \sum_{k=51}^{400} C_{400}^k 0.1^k 0.9^{400-k} &\approx P \left[\frac{51 - np}{\sqrt{np(1-p)}} \leq N(0, 1) \leq \frac{400 - np}{\sqrt{np(1-p)}} \right] \\ &= P[1.8333 \leq N(0, 1) \leq 60] = 0.0334. \end{aligned}$$

□

Bài tập Chương 4

4.1. Gieo một xúc sắc cân đối và đồng chất 12 000 lần. Tìm xác suất để số lần xuất hiện mặt một chấm từ 1 900 tới 2 150.

4.2. Một cuộc thi trắc nghiệm gồm 100 câu hỏi. Mỗi câu có 5 lựa chọn để trả lời trong đó chỉ có 1 lựa chọn đúng. Tìm xác suất để một thí sinh hoàn toàn không biết gì về môn học đó trả lời đúng từ 20 câu đến 30 câu.

4.3. Biết rằng khối lượng của một sản phẩm lấy ngẫu nhiên là đại lượng ngẫu nhiên có phân bố chuẩn với kỳ vọng bằng 10g, độ lệch chuẩn 1g. Tìm xác suất để trong 200 sản phẩm lấy ngẫu nhiên có:

- a) Ít nhất một sản phẩm có khối lượng lớn hơn 11g.
- b) từ 60 đến 100 sản phẩm có khối lượng lớn hơn 11g.

4.4. Đại lượng ngẫu nhiên X có phân bố mũ với kỳ vọng bằng 2. Quan sát X 1000 lần. Tìm xác suất để có từ 210 đến 240 lần X nhận giá trị nhỏ hơn $\frac{1}{2}$.

4.5. Một phân xưởng may có 185 máy hoạt động. Giả sử xác suất để trong ca mỗi máy bị hỏng là 0.05.

- a) Tính xác suất trong ca có 6 máy bị hỏng, số máy bị hỏng có khả năng nhất và xác suất trong trường hợp đó.
- b) Tính xác suất để có từ 5 đến 20 máy hỏng trong ca.

Chương 4

Các định lý giới hạn

4.1 $n = 12\,000$, $p = \frac{1}{6}$, m = số lần xuất hiện mặt một chấm.

$$\begin{aligned}P(1900 \leq m \leq 2150) &= P\left[-2.4495 \leq \frac{m - np}{\sqrt{np(1-p)}} \leq 3.6742\right] \\&\approx P[-2.4495 \leq N(0, 1) \leq 3.6742] = 0.9927.\end{aligned}$$

4.2 $n = 100$, $p = \frac{1}{5}$, m = số câu trả lời đúng.

$$P(20 \leq m \leq 30) = P\left[0 \leq \frac{m - np}{\sqrt{np(1-p)}} \leq 2.5\right] \approx P[0 \leq N(0, 1) \leq 2.5] = 0.4938.$$

4.3 X = khối lượng một sản phẩm. $X \in N(\mu, \sigma^2)$: $\mu = 10$, $\sigma = 1$.

$p = P(X > 11) = 0.1587$, $n = 200$, m = số sản phẩm có khối lượng lớn hơn 11g.

a)

$$\begin{aligned}P(m \geq 1) &= P\left[-5.9477 \leq \frac{m - np}{\sqrt{np(1-p)}} \leq 32.5668\right] \\&\approx P[-5.9477 \leq N(0, 1) \leq 32.5668] = 1.\end{aligned}$$

b)

$$\begin{aligned}P(60 \leq m \leq 100) &= P\left[5.4712 \leq \frac{m - np}{\sqrt{np(1-p)}} \leq 13.2128\right] \\&\approx P[5.4712 \leq N(0, 1) \leq 13.2128] = 0.\end{aligned}$$

4.4 $p = P\left(X < \frac{1}{2}\right) = 0.2212$, $n = 1000$, m = số lần thấy $X > \frac{1}{2}$.

$$P(210 \leq m \leq 240) = P\left[-0.8533 \leq \frac{m - np}{\sqrt{np(1-p)}} \leq 1.4324\right]$$

$$\approx P[-0.8533 \leq N(0, 1) \leq 1.4324] = 0.7272.$$

4.5 a) $n = 185, p = 0.05$.

$$* P(m = 6) = C_n^6 p^6 (1 - p)^{n-6} = 0.08250.$$

$$* (n + 1)p = 9.3 \notin \mathbb{Z} \Rightarrow \text{số máy bị hỏng có khả năng nhất } k_0 = \lfloor 9.3 \rfloor = 9.$$

$$p_0 = C_n^9 p^9 (1 - p)^{n-9} = 0.1346.$$

b)

$$\begin{aligned} P(5 \leq m \leq 20) &= P\left[-1.4337 \leq \frac{m - np}{\sqrt{np(1 - p)}} \leq 3.6264\right] \\ &\approx P[-1.4337 \leq N(0, 1) \leq 3.6264] = 0.9240. \end{aligned}$$

