Let your light so shine before men, that they may see your good works, and glorify your Father which is in heaven (Matthew 5:16).

Mục lục

Thông tin về học phần

Tên học phần:	Xác suất thống kê và toán kinh tế (Probability, Statistics and Mathematical Economics)
Mã số học phần:	
Thuộc khối kiến thức/kỹ năng:	Kiến thức cơ bản
Số tín chỉ:	04
Số tiết lý thuyết:	48
Số tiết thảo luận/bài tập:	12
Số tiết thực hành:	0
Số tiết hoạt động nhóm:	0
Số tiết tự học:	120
Môn học tiên quyết:	Toán cao cấp cho các nhà kinh tế
Môn học song hành:	

1

Xác suất của biến cố và phân phối xác suất

> *"Better than a thousand days of diligent study is one day with a great teacher."*
>
> – Japanese Proverb, *http://quotationsbook.com*

Giới thiệu

Cho đến nửa đầu thế kỷ hai mươi, lý thuyết xác suất mới trở nên một khoa học được xây dựng chặt chẽ bằng hệ tiên đề Kolmogorov. Cho đến nay, lý thuyết xác suất được ứng dụng vào rất nhiều ngành khoa học khác nhau, trong đó có các ví dụ điển hình như sau:

- Lý thuyết xác suất được sử dụng trong di truyền học như là một mô hình cho sự đột biến và đóng một vai trò quan trọng trong ngành tin-sinh.
- Nhiều ngành lý thuyết phát triển cao xem các loại nhiễu trong các thiết bị điện tử và hệ thống giao tiếp như là các quá trình ngẫu nhiên.
- Rất nhiều các mô hình trong nghiên cứu biến đổi khí quyển sử dụng các khái niệm của lý thuyết xác suất.

- Xác suất đóng vai trò nền tảng của ngành lý thuyết tài chính.

- Xác suất được sử dụng để nghiên cứu các hệ thống phức hợp và cải tiến độ tin cậy của các hệ thống đó, ví dụ như các hệ thống trong thương mại hiện đại hoặc không quân.

Trong chương này chúng ta sẽ nghiên cứu các khái niệm cơ bản của lý thuyết xác suất như không gian mẫu, biến cố ngẫu nhiên, độ đo xác suất, xác suất có điều kiện, và tính độc lập. Qua đó, chúng ta sẽ từng bước xây dựng nên mô hình toán học cho các hiện tượng ngẫu nhiên và tính xác suất của các biến cố, xác suất có điều kiện thông qua các công thức tính xác suất như công thức cộng xác suất, công thức nhân xác suất, công thức xác suất toàn phần và công thức Bayes. Bên cạnh đó, chương này cũng giới thiệu những kiến thức cơ bản về biến ngẫu nhiên, vector ngẫu nhiên, hàm phân phối, các phân phối xác suất quan trọng và các đặc trưng của biến ngẫu nhiên như kỳ vọng, phương sai và độ lệch tiêu chuẩn.

Chuẩn đầu ra của chương

Mục tiêu	Mô tả CĐR	Mức độ giảng dạy
G.1.1	Trình bày được khái niệm không gian mẫu, biến cố, mối quan hệ giữa các biến cố, xác suất của các biến cố, xác suất có điều kiện, tính độc lập của các biến cố.	T,U
G.1.2	Trình bày được khái niệm biến ngẫu nhiên, biến ngẫu nhiên độc lập, bảng phân phối, hàm mật độ, hàm phân phối, vector ngẫu nhiên, các phân phối xác suất thông dụng.	T,U
G.2.1	Tính toán được xác suất của biến cố, xác suất có điều kiện thông qua việc vận dụng các công thức tính xác suất như công thức cộng xác suất, công thức nhân xác suất, công thức xác suất toàn phần, công thức Bayes.	T,U
G.2.2	Tìm được hàm phân phối, bảng phân phối, hàm mật độ, các số đặc trưng của biến ngẫu nhiên (kỳ vọng, phương sai và độ lệch tiêu chuẩn) của các biến ngẫu nhiên.	T,U
G.3.1	Có thái độ tích cực hợp tác với giáo viên và các sinh viên khác trong quá trình học và làm bài tập.	T
G.3.2	Có kế hoạch tự học, làm các bài tập về nhà, câu hỏi thảo luận một cách hiệu quả.	U
G.3.3	Có khả năng thuyết trình các vấn đề tự học ở nhà và báo cáo kết quả làm việc của bản thân.	I,T

Nội dung của chương

1.1 Biến cố và xác suất của biến cố

1.1.1 Không gian mẫu và biến cố

Phép thử ngẫu nhiên là những thí nghiệm, quan sát, hành động, ... mà kết quả xảy ra một cách ngẫu nhiên, không biết trước được. Chẳng hạn, tung một đồng xu thì ta không biết trước được mặt sấp hay mặt ngửa xuất hiện, khoảng thời gian giữa hai email mà ta nhận được ta cũng không biết trước được là bao nhiêu phút. Những thí nghiệm/quan sát này là các ví dụ về phép thử ngẫu nhiên.

Tập hợp tất cả các kết quả có thể có của phép thử ngẫu nhiên được gọi là không gian mẫu, ký hiệu là Ω. Các phần tử của Ω được gọi là các biến cố sơ cấp, thường được ký hiệu là ω. Không gian mẫu còn gọi là không gian các biến cố sơ cấp. Chúng ta xem xét các ví dụ sau đây.

Ví dụ 1.1.1. Một người tung một con xúc xắc sáu mặt. Khi đó, không gian mẫu là

$$\Omega = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}.$$

Ví dụ 1.1.2. Giả sử ta quan sát một người lạ và dự đoán ngày sinh của người đó. Khi đó, không gian mẫu sẽ là

$$\Omega = \{1/1,\ 2/1, \ldots,\ 31/12\}.$$

Ví dụ 1.1.3. Một người đưa thư đi qua một đoạn đường có ba cột đèn giao thông. Tại mỗi cột đèn người đó sẽ tiếp tục đi (c) nếu gặp đèn xanh và dừng lại (s) nếu gặp đèn đỏ. Khi đó, không gian mẫu là

$$\Omega = \{ccc,\ ccs,\ csc,\ scc,\ css,\ scs,\ ssc,\ sss\}.$$

Ví dụ 1.1.4. Số email mà một người nhận được mỗi ngày cũng là một số ngẫu nhiên. Không gian mẫu là

$$\Omega = \{0, 1,\ 2,\ 3, \ldots\}.$$

Ví dụ 1.1.5. Tuổi thọ của một loài nào đó cũng không định trước được, và nhiều khi cũng được mô hình hóa như là một số ngẫu nhiên. Trong ví dụ này, không gian mẫu là tập hợp tất cả các số thực không âm

$$\Omega = \{t | t \geq 0\}.$$

Trong lý thuyết xác suất, ta thường quan tâm đến một tập con nào đó của không gian mẫu, tức là một tập hợp các kết quả có thể nào đó của phép thử ngẫu nhiên. Một tập con của không gian mẫu được gọi là một *biến cố*, hay *sự kiện*. Chẳng hạn, trong Ví dụ 1.1.1, biến cố "xuất hiện mặt có số chấm là một số chẵn lớn hơn 3" là

$$A = \{4,\ 6\}.$$

Một biến cố sơ cấp thuộc biến cố A được gọi là kết quả (biến cố sơ cấp) thuận lợi cho biến cố A. Trong trường hợp xét ở trên, 4 và 6 là các kết quả thuận lợi cho A. Tập rỗng $\emptyset$ là biến cố không chứa kết quả nào của phép thử ngẫu nhiên, và được gọi là biến cố không thể. Trong khi đó, tập Ω là biến cố chứa tất cả các kết quả có thể của phép thử ngẫu nhiên, và được gọi là biến cố chắc chắn.

Vì các biến cố là các tập hợp nên các khái niệm và phép toán đại số về tập hợp như khái niệm tập con, phần bù, giao của hai tập hợp, hợp của hai tập hợp,... được vận dụng trực tiếp trong lý thuyết xác suất. Nếu $A \subset B$ thì sự xảy ra của biến cố A sẽ kéo theo biến cố B. Khi đó ta nói A *kéo theo* B. Trong Ví dụ 1.1.1, biến cố "xuất hiện có số chấm là một số chẵn" là

$$B = \{2,\ 4,\ 6\}.$$

Với hai biến cố A và B xét ở trên, ta thấy A kéo theo B.

Hợp của hai biến cố, $C = A \cup B$, là biến cố xảy ra khi và chỉ khi biến cố A hoặc biến cố B xảy ra. Nói cách khác, biến cố $C = A \cup B$ là tập hợp gồm các kết quả thuộc A hoặc thuộc B.

Giao của hai biến cố, $C = A \cap B$ (hoặc $C = AB$), là biến cố xảy ra khi và chỉ khi cả hai biến cố A và B đồng thời xảy ra. Nói cách khác, biến cố $C = A \cap B$ là tập hợp gồm các kết quả thuộc A và thuộc B.

Hiệu của hai biến cố, $C = A \backslash B$, là biến cố xảy ra khi và chỉ khi A xảy ra và B không xảy ra. Nói cách khác, biến cố $C = A \backslash B$ là tập hợp gồm các kết quả thuộc A nhưng không thuộc B.

Phần bù của một biến cố A, ký hiệu là $\overline{A}$, là biến cố xảy ra khi và chỉ khi biến cố A không xảy ra. Nói cách khác, $\overline{A} = \Omega \backslash A$. $\overline{A}$ được gọi là biến cố đối kháng của biến cố A. Như vậy, $\overline{A}$ là biến cố chứa tất cả những phần tử của không gian mẫu mà không thuộc A.

Nếu $A \cap B = \emptyset$, thì ta nói A và B *xung khắc* hay *rời nhau*. Rõ ràng là A và $\overline{A}$ là hai biến cố xung khắc.

Trong Ví dụ 1.1.3, nếu A là biến cố người đưa thư dừng ở đèn giao thông thứ nhất và B là biến cố người đưa thư dừng ở cột đèn thứ hai, thì

$$A = \{sss,\ ssc,\ scs,\ scc\},$$

$$B = \{sss,\ ssc,\ css,\ csc\}.$$

$A \cup$B là biến cố người đưa thư dừng ở đèn giao thông thứ nhất hoặc đèn giao thông thứ hai:

$$A \cup B = \{sss,\ ssc,\ scs,\ scc,\ css,\ csc\}.$$

$A \cap$B là biến cố người đưa thư dừng ở cả hai đèn giao thông thứ nhất và đèn giao thông thứ hai:

$$A \cup B = \{sss,\ ssc\}.$$

$\overline{A}$ là biến cố người đưa thư không dừng ở đèn giao thông thứ nhất:

$$\overline{A} = \{ccc,\ ccs,\ csc,\ css\}.$$

Nếu C là biến cố người đưa thư không dừng ở đèn giao thông nào, thì $C = \{ccc\}$. Trong trường hợp này A và C rời nhau, B và C rời nhau: $A \cap C = \emptyset$, $B \cap C = \emptyset$.

Sau đây ta liệt kê một số tính chất của lý thuyết tập hợp.

Tính giao hoán:

$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A.$$

Tính kết hợp:

$$(A \cup B) \cup C = A \cup (B \cup C),$$
$$(A \cap B) \cap C = A \cap (B \cap C).$$

Tính phân phối:

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C),$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C).$$

Công thức De Morgan:

$$\overline{A \cup B} = \overline{A} \cap \overline{B},$$
$$\overline{A \cap B} = \overline{A} \cup \overline{B}.$$

1.1.2 Định nghĩa xác suất theo quan điểm cổ điển

Trước hết ta giới thiệu định nghĩa xác suất theo quan điểm cổ điển. Giả sử một phép thử ngẫu nhiên có hữu hạn các kết quả có cùng khả năng xảy ra (gọi là các kết quả đồng khả năng). Khi đó, xác suất (probability) của biến cố A là một số đo khả năng biến cố A xuất hiện, ký hiệu là $P(A)$, được định nghĩa là

$$P(A) = \frac{\text{số các kết quả thuận lợi cho biến cố } A}{\text{số các kết quả có thể của phép thử ngẫu nhiên}}.$$

Ví dụ 1.1.6. Ví dụ xét phép thử ngẫu nhiên tung một con xúc xắc cân đối, đồng chất (để đảm bảo cho các mặt xuất hiện với khả năng như nhau). Khi đó không gian mẫu là

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Nếu gọi

A là biến cố "xuất hiện mặt có số chấm là số chẵn lớn hơn 3",

B là biến cố "xuất hiện mặt có số chấm là số chẵn",

C là biến cố "xuất hiện mặt có số chấm là số lẻ không vượt quá 2", thì

$$A = \{4, 6\},\ B = \{2, 4, 6\},\ C = \{1\}.$$

Do đó, ta có

$$P(A) = \frac{2}{6},\ P(B) = \frac{3}{6}, P(C) = \frac{1}{6}.$$

1.1.3 Định nghĩa xác suất theo hệ tiên đề

Định nghĩa xác suất theo quan điểm cổ điển yêu cầu các giả thiết là phép thử ngẫu nhiên chỉ có hữu hạn kết quả (không áp dụng được cho các Ví dụ 1.1.4 và Ví dụ 1.1.5) và các kết quả có cùng khả năng (không áp dụng được cho Ví dụ 1.1.2 vì ngày 29/2 sẽ có khả năng xuất hiện ít hơn). Ngoài ra, trong Ví dụ 1.1.6, ta đã giả thiết là mỗi mặt của con xúc xắc đều có khả năng xuất hiện là 1/6. Điều này thường được kiểm tra bằng thực nghiệm. Chẳng hạn, nhà toán học Pháp là Buffon (1707-1788) tung một đồng xu 4040 lần và quan sát thấy mặt ngửa xuất hiện 2048. Nhà thống kê Pearson (1857-1936) tung một đồng xu 12000 lần và quan sát thấy mặt ngửa xuất hiện 6019 lần. Trong một thí nghiệm khác, Pearson (1857-1936) tung một đồng xu 24000 lần và quan sát thấy mặt ngửa xuất hiện 12012 lần. Trong thời gian bị người Đức giam giữ trong chiến tranh thế giới lần thứ II, nhà toán học người Nam Phi là John Kerrich (1903-1985) đã tung một đồng xu 10000 lần và quan sát thấy mặt ngửa xuất hiện 5067 lần. Trong các thí nghiệm trên, tỉ lệ xuất hiện mặt ngửa lần lượt là 0.5049, 0.5016,0.5005, và

0.5067 lần. Từ đó, ta định nghĩa xác suất để xuất hiện mặt ngửa khi tung đồng xu là 0.5. Cách định nghĩa xác suất này được gọi là xác suất theo quan điểm tần suất: Nếu ta thực hiện một phép thử ngẫu nhiên n lần trong điều kiện giống nhau. Ký hiệu số lần biến cố A xuất hiện trong n phép thử đó là n_A. Khi đó, nếu tỉ số n_A/n tiến về một giá trị nào đó thì ta định nghĩa giá trị đó là xác suất của biến cố A. Tuy nhiên, định nghĩa này không áp dụng được đối với những phép thử không thể lặp lại nhiều lần. Định nghĩa tổng quát sau đây khắc phục được các hạn chế trên.

Một độ đo xác suất trên không gian mẫu Ω là một hàm P từ tập hợp các tập con của Ω vào tập hợp số thực (nghĩa là gán cho mỗi biến cố một số thực nào đó), thỏa mãn ba tiên đề sau đây.

- $P(\Omega) = 1$.
- $P(A) \geq 0$ với mọi biến cố A.
- Nếu A_1 và A_2 là hai biến cố rời nhau, thì

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

  Tổng quát hơn, nếu $A_1, A_2, \ldots, A_n, \ldots$ là các biến cố đôi một rời nhau thì

$$P\Big(\bigcup_{j=1}^{\infty} A_j\Big) = \sum_{j=1}^{\infty} P(A_j).$$

Tiên đề thứ nhất và thứ hai hoàn toàn dễ hiểu. Vì Ω là tập hợp tất cả các kết quả có thể của phép thử ngẫu nhiên nên $P(\Omega) = 1$. Tiên đề thứ hai phát biểu rằng xác suất của mọi biến cố đều không âm. Tiên đề thứ ba phát biểu rằng nếu hai biến cố A và B rời nhau, tức là chúng không chứa kết quả chung nào, thì $P(A \cup B) = P(A) + P(B)$. Trong ví dụ dự đoán ngày sinh của một người không quen, xác suất của biến cố "ngày sinh của người đó rơi vào ba tháng đầu tiên của năm" sẽ bằng xác suất của biến cố "ngày sinh của người đó rơi vào tháng giêng" cộng với xác suất của biến cố "ngày sinh của người đó rơi vào tháng hai" cộng với xác suất của biến cố "ngày sinh của người đó rơi vào tháng ba".

1.1.4 Tính chất của xác suất

- **Tính chất 1.** $P(\overline{A}) = 1 - P(A)$. Tính chất này được suy ra tiên đề thứ nhất và tiên đề thứ ba. Vì A và $\overline{A}$ rời nhau, và $A \cup \overline{A} = \Omega$. Do đó, $P(A) + P(\overline{A}) = P(A \cup \overline{A}) = P(\Omega) = 1$.

- **Tính chất 2.** $P(\emptyset) = 0$. Tính chất này được suy từ Tính chất 1. Vì $\emptyset = \overline{\Omega}$, nên $P(\emptyset) = P(\overline{\Omega}) = 1 - P(\Omega) = 0$.

- **Tính chất 3.** Nếu $A \subset B$, thì $P(A) \leq P(B)$. Tính chất này phát biểu rằng nếu sự xảy ra của biến cố A kéo theo sự xảy ra của biến cố B thì xác suất của biến cố B không bé hơn xác suất của biến cố A. Ví dụ, sự xảy ra của biến cố "trời mưa" kéo theo sự xảy ra của biến cố "nhiều mây". Do đó, xác suất của biến cố "trời nhiều mây" luôn lớn hơn hoặc bằng xác suất của biến cố "trời mưa". Tính chất này được chứng minh chặt chẽ như sau. Trước tiên, ta biểu diễn biến cố B bởi hợp của hai biến cố rời nhau:

  $$B = A \cup (B \backslash A) \text{ (do } A \subset B).$$

  Do đó, từ tiên đề thứ hai và thứ ba, ta có

  $$P(B) = P(A) + P(B \backslash A) \geq P(A).$$

- **Tính chất 4.** $P(A) \leq 1$ với mọi biến cố A. Tính chất này được giải thích đơn giản như sau. Vì mọi biến cố A đều là tập con của không gian mẫu Ω và $P(\Omega) = 1$ nên từ Tính chất 3 ta có $P(A) \leq 1$.

- **Tính chất 5.** Nếu $A \subset B$, thì $P(B \backslash A) = P(B) - P(A)$. Điều này được suy ra từ chứng minh của Tính chất 3. Từ cách chứng minh Tính chất 3 ta còn thấy nếu $A \subset B$, thì

  $$P(B) = P(A) + P(B \backslash A).$$

  Do các số trong đẳng thức trên là hữu hạn nên ta có $P(B \backslash A) = P(B) - P(A)$. Khi $B = \Omega$ thì điều này chính là Tính chất 1.

- **Tính chất 6.** Với A, B là hai biến cố bất kỳ, ta có

$$P(A \cup B) = P(A) + P(B) - P(AB).$$

Tính chất này được chứng minh như sau. Đầu tiên, ta biểu diễn $A \cup B$ dưới dạng hợp của hai biến cố rời nhau:

$$A \cup B = A \cup (B \backslash AB).$$

Theo tiên đề thứ ba, ta có

$$P(A \cup B) = P(A \cup (B \backslash AB)).$$

Mặt khác, do $AB \subset B$ nên

$$P(B \backslash AB) = P(B) - P(AB).$$

Từ đó

$$P(A \cup B) = P(A) + P(B \backslash AB) = P(A) + P(B) - P(AB).$$

Hoàn toàn tương tự,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC),$$

và

$$P\Big(\bigcup_{j=1}^{n} A_j\Big) = \sum_{j=1}^{n} P(A_j) - \sum_{i<j} P(A_i A_j) + \sum_{i<j<k} P(A_i A_j A_k) - \cdots + (-1)^{n+1} P(A_1 A_2 \ldots A_n).$$

Công thức tổng quát này gọi là công thức inclusion-exclusion.

Ví dụ 1.1.7. Tung một đồng xu cân đối hai lần. Khi đó không gian mẫu là

$$\Omega = \{NN, NS, SN, SS\}.$$

Gọi A là biến cố "lần thứ nhất xuất hiện mặt ngửa", B là biến cố "lần thứ hai xuất hiện mặt ngửa". Khi đó

$$A = \{NN, NS\},\ B = \{NN, SN\},\ A \cup B = \{NN, NS, SN\},\ AB = \{NN\}.$$

Ta có

$$P(A \cup B) = \frac{3}{4},\ P(A) = P(B) = \frac{2}{4},\ P(AB) = \frac{1}{4}.$$

Bạn đọc dễ kiểm tra được $P(A \cup B) = P(A) + P(B) - P(AB)$.

Ví dụ 1.1.8 (Newton-Pepys problem)**.** Năm 1693, khi nghiên cứu một trò chơi cá độ, Samuel Pepys viết cho Isaac Newton một bức thư dài hỏi về một bài toán như sau. Trong các biến cố sau đây, biến cố nào có xác suất lớn nhất.

A: Có ít nhất một mặt sáu chấm xuất hiện khi tung 6 con xúc xắc.

B: Có ít nhất hai mặt sáu chấm xuất hiện khi tung 12 con xúc xắc.

C: Có ít nhất ba mặt sáu chấm xuất hiện khi tung 18 con xúc xắc.

Pepys tin rằng C là biến cố có xác suất lớn nhất. Nhưng câu trả lời là biến cố A. Để tính xác suất của biến cố này, ta sử dụng biến cố đối kháng. Ta nhận thấy $\overline{A}$ là biến cố "không có mặt sáu chấm xuất hiện khi tung 6 con xúc xắc". Do đó

$$P(A) = 1 - P(\overline{A}) = 1 - \Big(\frac{5}{6}\Big)^6 \approx 0.6651.$$

Tương tự

$$P(B) = 1 - \Big(\frac{5}{6}\Big)^{12} - C_{12}^1\Big(\frac{5}{6}\Big)^{11}\frac{1}{6} \approx 0.6187,$$

và

$$P(C) = 1 - \Big(\frac{5}{6}\Big)^{18} - C_{12}^1\Big(\frac{5}{6}\Big)^{17}\frac{1}{6} - C_{12}^2\Big(\frac{5}{6}\Big)^{16}\Big(\frac{1}{6}\Big)^2 \approx 0.5973.$$

Có ít nhất hai chuyện thú vị của bài toán này. Trước hết, vào những thời kỳ đầu tiên của lý thuyết xác suất, Samuel Pepys, một người rất thông minh và là thư ký của Bộ trưởng hải quân Anh dưới thời các vua King Charles II và King James II đã phải viết thư hỏi Newton,

trong khi đó ngày nay ta giải được trong vòng ít phút. Thứ hai, mặc dù Newton đưa ra đáp số đúng nhưng cách lý luận của Newton lại không chính xác. Để đọc thêm về toán này, cũng như "lời giải" của Newton, ta có thể tìm hiểu tại `http://en.wikipedia.org/wiki/Newton-Pepys_problem`

Ví dụ 1.1.9 (Matching problem)**.** Đây là một bài toán rất nổi tiếng của lý thuyết xác suất, được nghiên cứu đầu tiên vào năm 1713 bởi nhà toán học người Pháp, Pierre-Remond Montmort (1678-1719), do đó còn được gọi là bài toán Montmort. Bài toán có thể phát biểu như sau. Có n quý ông uống quá say trong bữa tiệc nên khi ra về mỗi người lấy ngẫu nhiên một chiếc mũ. Tìm xác suất để ít nhất có một người lấy đúng mũ của mình. Để giải bài toán này ta sử dụng công thức inclusion-exclusion. Gọi A_k là biến cố người thứ k lấy đúng mũ của mình. Khi đó, ta cần tìm

$$P\Big(\bigcup_{k=1}^{n} A_k\Big).$$

Với mỗi $1 \leq k \leq n$,

$$P\Big(A_1 A_2 \dots A_k\Big) = \frac{(n-k)!}{n!}.$$

Do tính chất bình đẳng, ở vế phải của công thức inclusion-exclusion có C_n^k số hạng như vậy. Do đó

$$P\Big(\bigcup_{k=1}^{n} A_k\Big) = \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n+1}\frac{1}{n!} \approx 1 - \frac{1}{e} \approx 0.632.$$

1.2 Xác suất có điều kiện

Trong cuộc sống, một câu hỏi thường gặp và rất tự nhiên đó là làm thế nào để chúng ta cập nhật niềm tin/nhận định của chúng ta về một sự kiện ngẫu nhiên. Ví dụ, trước khi chưa có bất kỳ thông tin nào, niềm tin của chúng ta về sự kiện ngày mai có sương mù được biểu diễn

là một số $P(H)$. Sau đó, nếu ta xem bản tin dự báo thời tiết, thì số $P(H)$ ban đầu có thể thay đổi. Khái niệm xác suất có điều kiện giúp chúng ta giải quyết những vấn đề này.

1.2.1 Định nghĩa xác suất có điều kiện và công thức nhân xác suất

Chúng ta bắt đầu giới thiệu khái niệm xác suất có điều kiện bởi một ví dụ. Trong y học, cây địa hoàng (digitalis) được dùng như một loài thuốc để chữa bệnh suy tim. Tuy nhiên, khi sử dụng phương pháp điều trị này thì có nguy cơ nhiễm độc khá cao và nghiêm trọng hơn đó là rất khó để phát hiện cơ thể bị nhiễm độc. Để phát hiện cơ thể bị nhiễm độc, người ta đo sự tập trung của chất địa hoàng trong máu. Năm 1971, Beller và các cộng sự thực hiện một nghiên cứu về mối quan hệ giữa độ tập trung của chất địa hoàng trong máu và sự nhiễm độc địa hoàng trên 135 bệnh nhân. Ta dùng các ký hiệu sau đây.

$T^+ =$ độ tập trung của chất địa hoàng trong máu cao (xét nghiệm dương tính),

$T^- =$ độ tập trung của chất địa hoàng trong máu thấp (xét nghiệm âm tính),

$$D^+ = \text{ bị nhiễm độc (có bệnh)},$$

$$D^- = \text{ không bị nhiễm độc (không có bệnh)}.$$

Kết quả nghiên cứu cho bởi bảng sau đây.

	D^+	D^-	Tổng
T^+	25	14	39
T^-	18	78	96
Tổng	43	92	135

Xét các tỉ lệ từng khả năng trên tổng số 135 bệnh nhân, ta có bảng sau đây.

	D^+	D^-	Tổng
T^+	0.185	0.104	0.289
T^-	0.133	0.578	0.711
Tổng	0.318	0.682	1.0

Từ bảng trên, ta thấy $P(T^+) = 0.289$ và $P(D^+) = 0.318$. Bây giờ, ta giả sử rằng bác sĩ biết một xét nghiệm là dương tính, khi đó xác suất bị nhiễm độc sẽ là bao nhiêu? Lúc này chúng ta chỉ quan tâm đến dòng thứ nhất. Ta thấy có 39 bệnh nhân có xét nghiệm dương tính và có 25 người bị nhiễm độc. Chúng ta ký hiệu xác suất để một bệnh nhân bị nhiễm độc với điều kiện xét nghiệm dương tính là $P(D^+|T^+)$, và gọi là xác suất của D^+ với điều kiện T^+,

$$P(D^+|T^+) = \frac{25}{39} = 0.64.$$

Hoặc chúng ta có thể tính

$$P(D^+|T^+) = \frac{P(D^+T^+)}{P(T^+)} = \frac{0.185}{0.289} = 0.640.$$

Như vậy, ta thấy xác suất không điều kiện $P(D^+) = 0.318$, trong khi đó xác suất có điều kiện $P(D^+|T^+) = 0.64$. Do đó, biết được xét nghiệm dương tính, thì xác suất bị nhiễm độc cao hơn gấp hai lần xác suất bị nhiễm độc khi chưa có thông tin gì. Tương tự, ta cũng có thể tính $P(D^-) = 0.682$, trong khi đó $P(D^-|T^-) = 0.848$.

Trong trường hợp tổng quát, ta có định nghĩa sau đây.

Định nghĩa 1.2.1. *Cho A và B là hai biến cố với $P(B) > 0$. Khi đó xác suất có điều kiện của biến cố A cho bởi biến cố B, ký hiệu là $P(A|B)$, được định nghĩa*

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

Nội dung đằng sau định nghĩa này đó là nếu ta giả sử biến cố B đã xảy ra thì không gian mẫu là B, thay vì Ω. Trong ví dụ về nhiễm độc

digitalis ở trên, để tìm $P(D^+|T^+)$, chúng ta hạn chế không gian mẫu về 39 bệnh nhân có xét nghiệm dương tính, thay vì 135 bệnh nhân ban đầu. Vì vậy, xác suất có điều kiện có đầy đủ tính chất của một độ đo xác suất, thể hiện qua định lý sau đây. Sau này nếu không chú ý gì thêm, khi viết $P(A|B)$, ta luôn hiểu có giả thiết $P(B) > 0$.

Định lý 1.2.2. *Các phát biểu sau đây là đúng.*

- *Với mọi biến cố A, B, ta có $P(\overline{A}|B) = 1 - P(A|B)$.*
- *Với mọi biến cố B, $P(\emptyset|B) = 0$.*
- *Nếu $A_1 \subset A_2$, thì $P(A_1|B) \leq P(A_2|B)$.*
- *$P(A|B) \leq 1$ với mọi biến cố A, B.*
- *Nếu $A_1 \subset A_2$, thì $P(A_2 \backslash A_1|B) = P(A_2|B) - P(A_1|B)$.*
- *Với $A_1, A_2, \ldots, A_n, B$ là các biến cố bất kỳ, ta có*

$$P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1A_2|B).$$

Tổng quát,

$$P\Big(\bigcup_{j=1}^{n} A_j|B\Big) = \sum_{j=1}^{n} P(A_j|B) - \sum_{i<j} P(A_iA_j|B) + \sum_{i<j<k} P(A_iA_jA_k|B) - \cdots + (-1)^{n+1}P(A_1A_2\ldots A_n|B).$$

Chứng minh. Việc chứng minh các tính chất này hoàn toàn tương tự như chứng minh các tính chất của độ đo xác suất. □

Trong một số bài toán, $P(A|B)$ và $P(B)$ được tính tương đối dễ dàng. Và khi đó, ta có thể tính được $P(AB)$. Đó là nội dung của **công thức nhân xác suất**, được trình bày trong định lý sau đây.

Định lý 1.2.3 (Công thức nhân xác suất)**.** *Cho A và B là hai biến cố, với $P(B) > 0$. Khi đó,*

$$P(AB) = P(B)P(A|B).$$

Vì vai trò của A và B như nhau nên nếu $P(A) > 0$, thì ta có $P(AB) = P(A)P(B|A)$. Định lý 1.2.3 được tổng quát như sau.

Định lý 1.2.4 (Công thức nhân xác suất tổng quát)**.** *Cho $A_1, A_2, \ldots, A_n$ là các biến cố. Khi đó,*

$$P(A_1A_2 \ldots A_n) = P(A_1)P(A_2|A_1) \ldots P(A_n|A_1A_2 \ldots A_{n-1}).$$

Ví dụ 1.2.5. Một cái bình đựng 3 quả cầu đỏ và 2 quả cầu xanh. Lấy lần lượt không hoàn lại 2 quả cầu ra khỏi bình. Tìm xác suất để hai quả đều màu đỏ.

Ta thấy rằng xác suất bài toán yêu cầu ta tìm xác suất không có điều kiện. Tuy nhiên, xác suất để quả cầu lấy lần thứ hai có màu đỏ phụ thuộc vào kết quả của việc lấy quả cầu lần thứ nhất. Gợi ý này giúp ta nghĩ đến xác suất có điều kiện. Lời giải được trình bày như sau.

Ký hiệu A_1 là biến cố quả cầu lấy ra lần thứ nhất có màu đỏ, và A_2 là biến cố quả cầu lấy ra lần thứ hai có màu đỏ. Khi đó, theo công thức nhân xác suất, ta có

$$P(A_1A_2) = P(A_1)P(A_2|A_1).$$

Rõ ràng ta thấy $P(A_1) = 3/5$. Và nếu quả cầu thứ nhất lấy ra có màu đỏ, thì trong bình chỉ còn 2 quả cầu đỏ và 2 quả cầu xanh. Do đó, $P(A_2|A_1) = 2/4$. Vì vậy

$$P(A_1A_2) = \frac{3}{5} \times \frac{2}{4} = \frac{3}{10}.$$

1.2.2 Công thức xác suất toàn phần và công thức Bayes

Mục này trình bày hai công thức quan trọng xuất phát từ xác suất có điều kiện, đó là công thức xác suất toàn phần và công thức Bayes. Trong việc giải quyết một vấn đề nào đó, có một chiến lược là ta chia vấn đề ra nhiều phần nhỏ và giải quyết trên từng phần nhỏ này. Chiến lược này xuất hiện trong công thức xác suất toàn phần và công thức Bayes, thông qua qua hệ đầy đủ các biến cố.

Định nghĩa 1.2.6. *Giả sử $A_1, A_2, \ldots, A_n$ là các biến cố thỏa mãn $A_iA_j = \emptyset$ với $i \neq j$ và $\cup_{j=1}^{n} A_j = \Omega$. Khi đó ta nói $\{A_1, A_2, \ldots, A_n\}$ là một hệ đầy đủ các biến cố, hoặc là một phân hoạch của Ω.*

Trong Ví dụ 1.2.5, xác suất có điều kiện ($P(A_2|A_1)$) được sử dụng để tính xác suất không điều kiện ($P(A_1A_2)$). Công thức xác suất toàn phần (một số giáo trình còn gọi là công thức xác suất đầy đủ) cũng có ý nghĩa như vậy.

Định lý 1.2.7 (Công thức xác suất toàn phần)**.** *Giả sử $\{A_1, A_2, \ldots, A_n\}$ là hệ đầy đủ các biến cố. Khi đó, với mọi biến cố A ta có*

$$P(A) = \sum_{j=1}^{n} P(A_j)P(A|A_j).$$

Chứng minh. Trước hết, chúng ta nhận xét rằng

$$\begin{aligned} P(A) &= P(A\Omega) \\ &= P\Big(A\Big(\bigcup_{j=1}^{n} A_j\Big)\Big) \\ &= P\Big(\bigcup_{j=1}^{n} AA_j\Big). \end{aligned}$$

Vì các biến cố AA_j, $1 \leq j \leq n$, là rời nhau nên

$$\begin{aligned} P\Big(\bigcup_{j=1}^{n}(AA_j)\Big) &= \sum_{j=1}^{n} P(AA_j) \\ &= \sum_{j=1}^{n} P(A_j)P(A|A_j). \end{aligned}$$

□

Ví dụ 1.2.8. Xét Ví dụ 1.2.5, tìm xác suất để quả cầu thứ hai là màu đỏ. Ví dụ này được giải quyết bằng cách áp dụng công thức xác suất toàn phần như sau. Ký hiệu A_1 là biến cố quả cầu lấy ra lần thứ nhất có màu đỏ, và A_2 là biến cố quả cầu lấy ra lần thứ hai có màu đỏ.

Trước hết, ta nhận xét rằng $\{A_1, \overline{A_1}\}$ là một hệ đầy đủ các biến cố. Do đó, theo công thức xác suất toàn phần, ta có

$$P(A_2) = P(A_1)P(A_2|A_1) + P(\overline{A_1})P(A_2|\overline{A_1}) = \frac{3}{5} \times \frac{2}{4} + \frac{2}{5} \times \frac{3}{4} = \frac{3}{5}.$$

Công thức xác suất toàn phần có rất nhiều ứng dụng quan trọng. Ví dụ sau đây là một bài toán rất nổi tiếng trong lý thuyết trò chơi, có tên là bài toán người chơi bị phá sản.

Ví dụ 1.2.9. Có hai người tham gia một trò chơi. Người thứ nhất có xác suất thắng ở mỗi ván là p và xác suất thua ở mỗi ván là $q = 1 - p$. Nếu người nào thắng sẽ nhận được 1\$; nếu thua sẽ mất 1\$. Giả sử người thứ nhất và người thứ hai có vốn xuất phát tương ứng là i\$ và j\$. Trò chơi chỉ dừng lại khi một trong hai người hết tiền (bị phá sản). Ta hãy tìm xác suất để người chơi thứ nhất thắng toàn bộ cuộc chơi.

Giả sử tổng số tiền của hai người là $N = (i + j)$\$. Gọi F_i là biến cố người thứ nhất thắng toàn bộ cuộc chơi khi có vốn xuất phát là i\$. Khi đó, hiển nhiên là $P(F_0) = 0$ và $P(F_N) = 1$. Gọi A là biến cố người thứ nhất thắng trong ván chơi đầu tiên. Theo công thức xác suất toàn phần, ta có

$$P(F_i) = P(A)P(F_i|A) + P(\overline{A})P(F_i|\overline{A}).$$

Sau ván chơi đầu tiên, số tiền của người thứ nhất là $(i + 1)$\$ với xác suất p (A xảy ra) và là $(i - 1)$\$ với xác suất q ($\overline{A}$ xảy ra). Do đó

$$P(F_i|A) = P(F_{i+1}) \text{ và } P(F_i|\overline{A}) = P(F_{i-1}).$$

Đặt $p_i = P(F_i)$, ta có phương trình

$$p_i = pp_{i+1} + qp_{i-1}.$$

Giải phương trình sai phân này (bằng cách xét phương trình đặc trưng $px^2 - x + q = 0$) ta được

$$p_i = \begin{cases} \dfrac{1 - (q/p)^i}{1 - (q/p)^N} & \text{nếu } p \neq q, \\ \dfrac{i}{i + j} & \text{nếu } p = q. \end{cases}$$

Từ kết quả này ta thấy nếu trò chơi là công bằng ($p = q$) thì xác suất thắng càng tăng nếu số vốn xuất phát càng nhiều. Ta hãy xét trường hợp trò chơi không công bằng với $q = 0.493$ và $p = 0.507$ (xác suất có ít nhất hai người cùng ngày sinh với số người trong phòng là 23), và giả sử số tiền ban đầu của hai người chơi bằng nhau. Sau đây là một số trường hợp cụ thể: với $i = j = 10\$$, xác suất người thứ nhất thắng toàn bộ cuộc chơi là $p_i = 57\%$, với $i = j = 100\$$, $p_i = 95\%$, và với $i = j = 300\$$, $p_i = 99.998\%$.

Kết quả này cho thấy dù xác suất để người thứ nhất thắng ở mỗi ván chỉ cao hơn của người thứ hai chưa tới 1.5% thì nếu số vốn lớn (300\$), gần như là người thứ hai sẽ bị phá sản (với xác suất 99.998%).

Ví dụ 1.2.10. Giả sử rằng tính cơ động nghề nghiệp được chia thành ba cấp độ là cao (U), trung bình (M) và thấp (L). Ký hiệu U_1 là biến cố người bố có nghề nghiệp ở nhóm cấp độ cao, U_2 là biến cố người con có nghề nghiệp ở cấp độ cao, M_1 là biến cố người bố có nghề nghiệp ở nhóm cấp độ trung bình, M_2 là biến cố người con có nghề nghiệp ở cấp độ trung bình, ... Năm 1954, Glass và Hall [5] thống kê tính cơ động nghề nghiệp ở Anh và xứ Wales thể hiện bởi bảng sau:

	U_2	M_2	L_2
U_1	0.45	0.48	0.07
M_1	0.05	0.70	0.25
L_1	0.01	0.50	0.49

Bảng này được gọi là ma trận xác suất chuyển, và có nghĩa như sau: Nếu người bố có nghề nghiệp ở nhóm U, thì xác suất để con của anh ấy có nghề nghiệp ở nhóm U là 0.45, ở nhóm M là 0.48, và tương tự cho những nhóm khác. Nghĩa là, bảng trên cung cấp cho chúng ta xác suất có điều kiện: $P(U_2|U_1) = 0.45,\ P(M_2|U_1) = 0.48., \ldots$ Giả sử rằng ở thế hệ người bố, có 10% ở nhóm U, 40% ở nhóm M và 50% ở nhóm L. Khi đó xác suất để một người ở thế hệ tiếp theo có nghề nghiệp ở nhóm U là bao nhiêu? Đây là một ví dụ cho sự áp dụng công thức xác suất toàn phần. Thật vậy, do $\{U_1, M_1, L_1\}$ là hệ đầy đủ các

biến cố nên

$$\begin{aligned} P(U_2) &= P(U_1)P(U_2|U_1) + P(M_1)P(U_2|M_1) + P(L_1)P(U_2|L_1) \\ &= 0.10 \times 0.45 + 0.40 \times 0.48 + 0.50 \times 0.01 = 0.07. \end{aligned}$$

$P(M_2)$ và $P(L_2)$ cũng được tính tương tự.

Bây giờ ta đặt câu hỏi ngược lại. Nếu một người con có nghề nghiệp ở nhóm U, hỏi xác suất để người bố có nghề nghiệp ở nhóm U là bao nhiêu? Chúng ta được cho phần "kết quả" và được hỏi xác suất của phần " nguyên nhân". Trong những bài toán này, công thức Bayes mà sắp tới chúng ta sẽ trình bày rất có hiệu quả. Bây giờ, chúng ta sẽ trở lại câu hỏi. Rõ ràng, ta cần tìm $P(U_1|U_2)$. Theo định nghĩa xác suất có điều kiện,

$$\begin{aligned} P(U_1|U_2) &= \frac{P(U_1U_2)}{P(U_2)} \\ &= \frac{P(U_2|U_1)P(U_1)}{P(U_1)P(U_2|U_1) + P(M_1)P(U_2|M_1) + P(L_1)P(U_2|L_1)} \\ &\qquad \text{(áp dụng công thức nhân xác suất đối với tử số} \\ &\qquad\quad \text{và công thức xác suất toàn phần đối với mẫu số)} \\ &= \frac{0.10 \times 0.45}{0.07} = 0.64. \end{aligned}$$

Nói cách khác, nếu người con có nghề nghiệp ở nhóm cao, thì 64% khả năng là bố anh ấy có nghề nghiệp ở nhóm cao.

Bây giờ chúng ta sẽ trình bày công thức Bayes.

Định lý 1.2.11 (Công thức Bayes)**.** *Giả sử $\{A_1, A_2, \ldots, A_n\}$ là hệ đầy đủ các biến cố. Khi đó, với mọi biến cố A và với mọi $1 \leq k \leq n$, ta có*

$$P(A_k|A) = \frac{P(A_k)P(A|A_k)}{\sum_{j=1}^{n} P(A_j)P(A|A_j)}.$$

Chứng minh. Công thức Bayes được chứng minh giống như phần đã trình bày ở cuối Ví dụ 1.2.10. □

Sau đây ta trình bày một bài toán rất nổi tiếng, có tên là bài toán ba người tù. Đây là một phiên bản của bài toán Monty Hall trong lý thuyết trò chơi. Về bài toán Monty Hall, độc giả quan tâm có thể xem tại `http://en.wikipedia.org/wiki/Monty_Hall_problem`.

Ví dụ 1.2.12 (Bài toán ba người tù)**.** Có 3 người tù là A, B và C bị xử tội chết và bị nhốt trong ba phòng giam riêng biệt. Có thông tin là thống đốc bang sẽ chọn ngẫu nhiên một trong ba người để ân xá. Người cai ngục biết rõ người được ân xá là ai nhưng không được phép nói ra. Người tù A liền hỏi cai ngục tên một trong hai người tù sẽ bị xử chết: "Nếu B được ân xá, hãy nói cho tôi người bị tử hình là C, nếu C được ân xá, hãy nói cho tôi người bị tử hình là B, còn nếu tôi được ân xá, hãy tung một đồng xu rồi nói cho tôi biết là B hoặc C".

Người cai ngục nói với A rằng B sẽ bị xử chết. A rất hài lòng vì nghĩ rằng xác suất sống sót của anh ấy đã tăng từ $1/3$ lên $1/2$. Sau đó, A bí mật nói cho C biết tin tức này. Sau khi nghe tin từ A, C cũng đã rất hài lòng vì anh ấy lý luận rằng A vẫn chỉ có $1/3$ xác suất được ân xá, nhưng cơ hội cho anh ấy đã lên $2/3$. Lập luận của ai là đúng?

Đáp số là cách lập luận của C. Rất nhiều người, trong đó có nhà toán học và cũng là nhà xác suất nổi tiếng Paul Erdõs (1913-1996), đều không nghĩ rằng lời giải trên là chính xác. Tính khó thuyết phục của lời giải đối với rất nhiều người là một trong những lý do làm cho bài toán trở nên nổi tiếng.

Bây giờ ta trình bày lời giải. Gọi L_a, L_b, L_c lần lượt là các biến cố A, B, C sẽ được ân xá. Ta có

$$P(L_a) = P(L_b) = P(L_c) = \frac{1}{3}.$$

Gọi D là biến cố người cai ngục nói B sẽ bị tử hình. Ta sẽ tìm $P(L_a|D)$. Theo công thức Bayes, ta có

$$\begin{aligned} P(L_a|D) &= \frac{P(L_a)P(D|L_a)}{P(L_a)P(D|L_a) + P(L_b)P(D|L_b) + P(L_c)P(D|L_c)} \\ &= \frac{1/3 \times 1/2}{1/3 \times 1/2 + 1/3 \times 0 + 1/3 \times 1} = \frac{1}{3}. \end{aligned}$$

Như vậy, sau khi có thêm thông tin, xác suất được ân xá của A vẫn giữ nguyên 1/3.

Công thức Bayes là một thành phần toán học cơ bản của lý thuyết nhận thức chủ quan. Nhìn nhận trên đặc điểm này, nhận định hay niềm tin của một cá nhân nào đó về thế giới có thể được mã hóa bằng xác suất. Ví dụ, sự tin tưởng của một người nào đó về việc ngày mai sẽ có mưa đá có thể biểu diễn bằng xác suất $P(H)$. Xác suất này sẽ được điều chỉnh khi chúng ta có thêm thông tin. Chẳng hạn, sau khi có thông tin E (xem bản tin dự báo thời tiết), xác suất đó sẽ trở thành $P(H|E)$. Thông thường, $P(E|H)$ dễ tính toán hơn $P(H|E)$. Trong trường hợp này, áp dụng công thức Bayes, ta có

$$P(H|E) = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(\overline{H})P(E|\overline{H})}.$$

1.2.3 Tính độc lập của các biến cố

Trong mục này, chúng ta sẽ trình bày khái niệm độc lập của các biến cố, một trong những khái niệm quan trọng nhất của lý thuyết xác suất. Một cách trực giác, ta có thể hiểu rằng hai biến cố A và B được gọi là độc lập nếu sự xảy ra của biến cố A không cho chúng ta thông tin gì về việc biến cố B có xảy ra hay không, nghĩa là

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

Ta viết lại điều này dưới dạng

$$P(AB) = P(A)P(B).$$

Từ đó, ta đưa ra định nghĩa sau đây.

Định nghĩa 1.2.13. *Hai biến cố A và B được gọi là độc lập nếu*

$$P(AB) = P(A)P(B).$$

Ví dụ 1.2.14. Tung hai con xúc xắc cân đối. Gọi A là biến cố con thứ nhất xuất hiện mặt sáu chấm và B là biến cố con thứ hai xuất hiện mặt sáu chấm. Khi đó

$$P(A) = \frac{1}{6},\ P(B) = \frac{1}{6} \text{ và } P(AB) = \frac{1}{36} = P(A)P(B).$$

Như vậy, A và B là hai biến cố độc lập.

Ví dụ 1.2.15. Rút ngẫu nhiên một con bài từ một bộ bài Tây 52 con. Gọi A là biến cố rút được con At và B là biến cố rút được con Rô. Khi đó

$$P(A) = \frac{4}{52},\ P(B) = \frac{13}{52} \text{ và } P(AB) = \frac{1}{52} = P(A)P(B).$$

Như vậy, A và B là hai biến cố độc lập.

Ví dụ 1.2.16. Một hệ thống được thiết kế sao cho nó bị hỏng khi và chỉ khi bộ phận chính và bộ phận dự phòng của nó cùng bị hỏng. Giả sử hai bộ phận này độc lập với nhau và xác suất hỏng của mỗi bộ phận là p. Khi đó, xác suất hỏng của cả hệ thống là p^2. Ví dụ, với $p = 10\%$ thì xác suất hỏng của cả hệ thống chỉ là 1%. Tính chất này được ứng dụng trong lý thuyết độ tin cậy để đảm bảo cho một hệ thống không bị trục trặc.

Định lý sau đây nêu lên một đặc trưng của tính độc lập của hai biến cố.

Định lý 1.2.17. *Cho A, B là các biến cố. Khi đó, các khẳng định sau là tương đương.*

i. A độc lập với B;

ii. $P(A\overline{B}) = P(A)P(\overline{B})$;

iii. $P(\overline{A}B) = P(\overline{A})P(B)$;

iv. $P(\overline{A}\ \overline{B}) = P(\overline{A})P(\overline{B})$.

Chứng minh. Phép chứng minh được suy từ định nghĩa tính độc lập của hai biến cố và tính chất $P(\overline{A}) = 1 - P(A)$. □

Bây giờ ta quan tâm đến trường hợp có nhiều hơn hai biến cố. Chẳng hạn ta xét sự độc lập của ba biến cố A, B, C. Ngoài sự độc lập của từng cặp biến cố với nhau, tức là ngoài ba đẳng thức

$$P(AB) = P(A)P(B),\ P(BC) = P(B)P(C),\ P(AC) = P(A)P(C),$$

ta còn cần thêm đẳng thức

$$P(ABC) = P(A)P(B)P(C)$$

để đảm bảo cho tính chất sự xảy ra của hai biến cố không ảnh hưởng đến khả năng xảy ra của biến cố còn lại, chẳng hạn, $P(A|BC) = P(A)$.

Tổng quát, ta có định nghĩa sau đây.

Định nghĩa 1.2.18. *Các biến cố $\{A_1, A_2, \dots A_n\}$ được gọi là độc lập nếu với mọi họ con $\{A_{i_1}, A_{i_2}, \dots, A_{i_k}\}$ của họ $\{A_1, A_2, \dots A_n\}$, ta có*

$$P(A_{i_1}A_{i_2}\dots A_{i_k}) = P(A_{i_1})P(A_{i_2})\dots P(A_{i_k}).$$

Các biến cố $\{A_1, A_2, \dots A_n\}$ được gọi là **độc lập đôi một** *nếu*

$$P(A_iA_j) = P(A_i)P(A_j) \text{ với mọi } 1 \le i, j \le n, i \ne j.$$

Như vậy, nếu một họ các biến cố độc lập thì sẽ độc lập đôi một. Ví dụ sau đây chỉ ra sự tồn tại của họ các biến cố độc lập đôi một nhưng không độc lập.

Ví dụ 1.2.19. Tung hai đồng xu cân đối. Gọi Gọi A là biến cố đồng xu thứ nhất xuất hiện mặt ngửa, B là biến cố đồng xu thứ hai xuất hiện mặt ngửa và C là biến cố chỉ có đúng một đồng xu xuất hiện mặt ngửa. Khi đó rõ ràng A và B là độc lập, và

$$P(A) = P(B) = P(C) = \frac{1}{2}.$$

Mặt khác, ta cũng có $P(C|A) = P(\overline{B}) = 0.5 = P(C)$ và $P(C|B) = P(\overline{A}) = 0.5 = P(C)$. Như vậy, A độc lập với C và B độc lập với C. Tuy nhiên,

$$\frac{1}{8} = P(A)P(B)P(C) \neq P(ABC) = 0.$$

Như vậy, ba biến cố $\{A, B, C\}$ không độc lập.

Ví dụ 1.2.20 (Nghịch lý ngày sinh)**.** Trong một phòng có n người. Giả sử ngày sinh của những người trong nhóm là độc lập, một năm gồm 365 ngày (không tính ngày $29-2$) và 365 ngày này bình đẳng nhau (nghĩa là ta giả sử ngày sinh có phân phối đều dù cho trên thực tế, khoảng thời gian 9 tháng sau mùa nghỉ, số trẻ em sinh ra nhiều hơn). Hãy tìm n nhỏ nhất để xác suất có ít nhất hai người trùng sinh nhật vượt quá 50%.

Một năm có 365 ngày nên có thể lần đầu tiên gặp câu hỏi này, trực giác của chúng ta có thể sẽ mách bảo rằng $n = 150$ hoặc $n = 180$ hay thậm chí n lớn hơn thế nữa. Nhưng câu trả lời đúng là chỉ cần $n = 23$, xác suất để có ít nhất hai người trùng ngày sinh sẽ lớn hơn 50%; còn nếu $n = 57$, thì xác suất đó lên tới 99%! Kết quả này làm cho hầu hết mọi người rất ngạc nhiên trong lần đầu biết đến. Trong lý thuyết xác suất, bài toán này gọi là bài toán ngày sinh (birthday problem) hay còn gọi là nghịch lý ngày sinh.

Ta xét bài toán tìm xác suất để trong một nhóm n người, có ít nhất có hai người có cùng ngày sinh.

Theo nguyên lý chuồng bồ câu, nếu $n > 365$ thì xác suất cần tìm bằng 1. Như vậy, vấn đề còn lại là xét bài toán khi $n \leq 365$.

Gọi A là biến cố có ít nhất hai người trùng sinh nhật. Khi đó xác suất cần tìm là

$$P(n) = P(A) = 1 - P(\bar{A}),$$

với $\bar{A}$ là biến cố không có cặp nào trùng ngày sinh. Giả sử ta có 365 ô. Số cách để "gieo" n người vào 365 ô này là 365^n, số cách để "gieo" n người vào các ô riêng biệt là $365 \times 364 \times \cdots \times (365 - n + 1)$. Theo định nghĩa xác suất, ta thấy

$$P(\bar{A}) = \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}.$$

Do đó

$$P(A) = 1 - 1 \times (1 - \frac{1}{365}) \times (1 - \frac{2}{365}) \times (1 - \frac{n-1}{365}).$$

Các thừa số $1 - k/365$ rất gần 1 (với k bé), nhưng tích của chúng giảm rất nhanh theo tốc độ mũ. Ta liệt kê một số kết quả tương ứng với những giá trị cụ thể của n như sau:

$$n = 10,\ P(A) \approx 11.694\%;$$
$$n = 23,\ P(A) \approx 50.729\%;$$
$$n = 50,\ P(A) \approx 97\%;$$
$$n = 57,\ P(A) \approx 99\%;$$
$$n = 100,\ P(A) \approx 99.99997\%.$$

Công thức tính chính xác $P(A)$ ở trên tương đối cồng kềnh. Do đó, trong tính toán, ta có thể dùng xấp xỉ

$$1 + x \approx e^x \text{ với } x \text{ đủ gần } 0.$$

Khi đó

$$P(n) = P(A) \approx 1 - \exp\{-(1+2+\cdots+(n-1))/365\} = 1 - \exp\{-n(n-1)/730\}.$$

Từ đó, ta có, chẳng hạn

$$P(23) \approx 1 - \exp\{-253/365\} \approx 0.5;\ \ P(57) \approx 1 - \exp\{-1596/365\} \approx 0.99.$$

1.3 Biến ngẫu nhiên

1.3.1 Giới thiệu về biến ngẫu nhiên

Khi một phép thử ngẫu nhiên được thực hiện, nói chung ta thường quan tâm đến một hàm số nào đó của kết quả xảy ra. Ví dụ, khi ta tung hai con xúc xắc, ta có thể chỉ quan tâm đến tổng số chấm trên hai mặt của con xúc xắc. Nếu biến cố $(3, 2)$ hoặc $(2, 3)$ xảy ra, đều cho

ta kết quả tổng hai chấm là 5, nếu biến cố $(5,6)$ hoặc $(5,6)$ xảy ra, đều cho ta kết quả tổng hai chấm là 11. Những hàm cho tương ứng một kết quả của phép thử ngẫu nhiên với một số thực gọi là biến ngẫu nhiên.

Định nghĩa 1.3.1. *Biến ngẫu nhiên X là một hàm số xác định trên không gian mẫu Ω và nhận giá trị trong $\mathbb{R}$:*

$$X : \Omega \longrightarrow \mathbb{R}.$$

Đối với ví dụ tung hai con xúc xắc ở trên, gọi X là tổng số chấm xuất hiện. Khi đó $X((1,1)) = 2,\ X((3,2)) = 5,\ X((5,6)) = 11, \ldots$

Vì giá trị của biến ngẫu nhiên hoàn toàn phụ thuộc vào kết quả của phép thử ngẫu nhiên nên ta thường quan tâm đến các xác suất

$$P(X = x) := P(\{\omega \in \Omega : X(\omega) = x\}),$$

hoặc

$$P(a < X < b) := P(\{\omega \in \Omega : a < X(\omega) < b\}),$$

trong đó x, a, b là các số thực.

Ví dụ 1.3.2. Tung một đồng xu cân đối ba lần. Gọi X là số mặt ngửa xuất hiện. Khi đó không gian mẫu có 8 phần tử.

$$\Omega = \{(sss), (ssn), (sns), (nss), (snn), (nsn), (nns), (nnn)\}.$$

Ta có

$$P(X = 0) = P(\text{mặt ngửa xuất hiện 0 lần}) = P\Big((sss)\Big) = \frac{1}{8}.$$

Tương tự

$$P(X = 1) = \frac{3}{8},\ P(X = 2) = \frac{3}{8},\ P(X = 3) = \frac{1}{8}.$$

Khái niệm độc lập cũng được phát biểu đối với họ các biến ngẫu nhiên. Một cách trực giác, ta hiểu hai biến ngẫu nhiên X và Y được gọi là độc lập khi mọi biến cố liên quan đến biến ngẫu nhiên X độc lập với mọi biến cố liên quan đến biến ngẫu nhiên Y. Một cách chính xác, ta có định nghĩa sau đây.

Định nghĩa 1.3.3. *Giả sử I là một tập chỉ số nào đó. Các biến ngẫu nhiên $\{X_i, i \in I\}$ được gọi là* **độc lập** *nếu với mọi họ con $\{X_{k_1}, X_{k_2}, \ldots, X_{k_n}\} \subset \{X_i, i \in I\}$ đều thỏa mãn*

$$P(X_{k_1} \leq x_1, \ldots, X_{k_n} \leq x_n) = P(X_{k_1} \leq x_1) \ldots P(X_{k_n} \leq x_n), \quad \text{với mọi } x_1, \ldots, x_n \in \mathbb{R}.$$

Các biến ngẫu nhiên $\{X_i, i \in I\}$ được gọi là **độc lập đôi một** *nếu với mọi $i \neq j$, $\{X_i, X_j\}$ là hai biến ngẫu nhiên độc lập.*

1.3.2 Hàm phân phối

Khái niệm hàm phân phối được trình bày sau đây sẽ giúp ta xác định các xác suất dạng $P(X = x)$ hoặc $P(a < X < b)$. Tổng quát hơn, một biến ngẫu nhiên hoàn toàn được xác định nếu ta biết được hàm phân phối của nó.

Định nghĩa 1.3.4. *Hàm phân phối của biến ngẫu nhiên X, ký hiệu là F_X, là hàm số xác định bởi $\mathbb{R}$, cho bởi*

$$F_X(x) = P(X \leq x) := P(\{\omega \in \Omega : X(\omega) \leq x\}), \ x \in \mathbb{R}.$$

Sau này, nếu không sợ nhầm lẫn, ta ký hiệu F thay cho F_X. Trong một số bài toán thực tế, khi X là biến ngẫu nhiên chỉ thời gian sống của một thiết bị hay của một hệ thống nào đó, thì xác suất $P(X > x) = 1 - P(X \leq x)$ là xác suất để thời gian sống của hệ thống lớn hơn x.

Định nghĩa 1.3.5. *Cho X là biến ngẫu nhiên. Hàm số $\overline{F}(x) = P(X > x)$ được gọi là hàm sống sót (survival function) của X.*

Chú ý 1.3.6. Trong một số cuốn sách, Hàm phân phối được định nghĩa là $F_X(x) = P(X < x)$. Cách định nghĩa đưa ra ở trên đây được dùng trong rất nhiều cuốn được xuất bản gần đây như [3, 6, 7, 8, 9].

Ví dụ 1.3.7. Giả sử F là hàm phân phối của biến ngẫu nhiên X trong Ví dụ 1.3.2, ta có các trường hợp sau đây.

- Nếu $x < 0$, thì
$$F(x) = P(\emptyset) = 0.$$
- Nếu $0 \leq x < 1$, thì
$$F(x) = P(X = 0) = 1/8.$$
- Nếu $1 \leq x < 2$, thì
$$F(x) = P(X = 0) + P(X = 1) = 1/2.$$
- Nếu $2 \leq x < 3$, thì
$$F(x) = P(X = 0) + P(X = 1) + P(X = 2) = 7/8.$$
- Nếu $x \geq 3$, thì
$$F(x) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1.$$

Như vậy,
$$F(x) = \begin{cases} 0 & \text{nếu } x < 0; \\ 1/8 & \text{nếu } 0 \leq x < 1; \\ 1/2 & \text{nếu } 1 \leq x < 2; \\ 7/8 & \text{nếu } 2 \leq x < 3; \\ 1 & \text{nếu } x \geq 3. \end{cases}$$

Định lý sau đây trình bày một số tính chất của hàm phân phối. Trong phạm vi cuốn sách này, ta bỏ qua phép chứng minh. Bạn đọc quan tâm có thể xem chứng minh chi tiết tại [6, tr. 30].

Định lý 1.3.8. *Giả sử F là hàm phân phối của biến ngẫu nhiên X. Khi đó*

- *F là hàm không giảm, liên tục phải và có giới hạn trái;*
- $\lim_{x\to-\infty} F(x) = 0,\ \lim_{x\to+\infty} F(x) = 1;$
- *Tập các điểm gián đoạn của F là hữu hạn hoặc vô hạn đếm được. F gián đoạn tại điểm a khi và chỉ khi $P(X = a) > 0$.*

1.3.3 Biến ngẫu nhiên rời rạc và bảng phân phối xác suất

Định nghĩa 1.3.9. *Nếu tập giá trị của biến ngẫu nhiên X là một tập hữu hạn hoặc vô hạn đếm được $\{x_1, x_2, \ldots, x_n, \ldots\}$ thì ta gọi X là biến ngẫu nhiên rời rạc.*

Giả sử biến ngẫu nhiên rời rạc X nhận các giá trị $x_1, x_2, \ldots, x_n, \ldots$. Khi đó, hàm $P(X = x_1) = p_1,\ P(X = x_2) = p_2, \ldots, P(X = x_n) = p_n, \ldots$ được gọi là hàm khối lượng xác suất (probability mass function) của X. Rõ ràng $p_1 + \cdots + p_n + \cdots = 1$. Đây cũng là đặc trưng của hàm khối lượng xác suất. Nói cách khác, các số dương $p_1, p_2, \ldots, p_n, \ldots$ là hàm khối lượng của biến ngẫu nhiên nào đó khi và chỉ khi tổng của chúng bằng 1.

Biến ngẫu nhiên rời rạc X được mô tả bằng bảng sau đây được gọi là bảng phân phối.

X	x_1	x_2	$\ldots$	x_n	$\ldots$
P	p_1	p_2	$\ldots$	p_n	$\ldots$

1.3.4 Biến ngẫu nhiên liên tục và hàm mật độ xác suất

Trong thực tế, ngoài biến ngẫu nhiên rời rạc, ta còn gặp những biến ngẫu nhiên có tập giá trị là một tập phủ kín một khoảng trên trục số (gọi là tập hợp có số lượng phần tử continuum). Chẳng hạn, xét X là tuổi thọ của một thiết bị điện tử. Khi đó, tập giá trị của X là $[0, +\infty)$. Đối với những biến ngẫu nhiên dạng này, các biến cố $(X = a)$ luôn có xác suất bằng 0 với a là một điểm cụ thể nào đó trên trục số. Thay vào đó, chúng ta quan tâm đến những biến cố dạng $(a \leq X \leq b)$. Chẳng hạn, ta tính xác suất để cho một loại máy tính có tuổi thọ từ 5 năm đến 6 năm, xác suất để một người ở một vùng nào đó có chiều cao từ $1,5m$ đến $1,6m$,... Hàm số đặc trưng cho những xác suất như vậy gọi là hàm mật độ và biến ngẫu nhiên có hàm mật độ gọi là biến ngẫu nhiên liên tục.

Định nghĩa 1.3.10. *Biến ngẫu nhiên X được gọi là liên tục tuyệt đối (sau đây ta sẽ gọi là liên tục) nếu tồn tại một hàm không âm $p(x)$, khả tích trên $\mathbb{R}$, và thỏa mãn*

$$P(a \leq X \leq b) = \int_a^b p(x)dx \quad \textit{với mọi } -\infty \leq a \leq b \leq +\infty.$$

Hàm $p(x)$ như vậy gọi là hàm mật độ của biến ngẫu nhiên X.

Như vậy, nếu biến ngẫu nhiên liên tục X có hàm mật độ là p và hàm phân phối là F, thì

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(u)du.$$

Do đó, hàm F là hàm liên tục tuyệt đối trên $\mathbb{R}$. Điều này giải thích tên gọi biến ngẫu nhiên liên tục tuyệt đối. Vì F liên tục tuyệt đối trên $\mathbb{R}$ nên nó có đạo hàm hầu hết tại mọi điểm tên $\mathbb{R}$, trừ ra nhiều nhất là đếm được điểm. Đẳng thức trên cũng kéo theo

$$p(x) = F^{'}(x) \text{ nếu hàm } F \text{ khả vi tại điểm } x.$$

Chú ý rằng

$$\int_{-\infty}^{+\infty} p(x)dx = P(-\infty \leq X \leq +\infty) = 1.$$

Tương tự hàm khối lượng xác suất, đây cũng chính là đặc trưng của hàm mật độ, thể hiện qua định lý sau đây.

Định lý 1.3.11. *Giả sử $p(x)$ là một hàm số không âm, khả tích trên $\mathbb{R}$. Khi đó $p(x)$ là hàm mật độ của biến ngẫu nhiên nào đó khi và chỉ khi*

$$\int_{-\infty}^{+\infty} p(x)dx = 1.$$

Chứng minh. Nếu $p(x)$ là hàm mật độ của biến ngẫu nhiên X ta đã chỉ ra đẳng thức

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

được thỏa mãn. Bây giờ ta sẽ chứng minh điều ngược lại. Xét không gian mẫu là $\Omega = \mathbb{R}$. Bỏ qua tính chặt chẽ của lý thuyết độ đo, ta có

$$P((a,b)) = P([a,b]) = \int_a^b p(x)dx \text{ với mọi } -\infty \leq a \leq b \leq \infty$$

là một độ đo xác suất trên các tập con của không gian mẫu Ω. Khi đó hàm đồng nhất từ $\mathbb{R}$ lên $\mathbb{R}$ chính là biến ngẫu nhiên nhận $p(x)$ làm hàm mật độ. □

Ngoài ra, nếu biến ngẫu nhiên liên tục X có hàm mật độ là $p(x)$, thì với mọi số thực a ta có

$$P(X = a) = \int_a^a p(x)dx = 0.$$

Do đó, với mọi điểm $a < b$,

$$P(a \leq X \leq b) = P(a < X < b) = F(b) - F(a).$$

Ta chú ý rằng điều này không đúng đối với biến ngẫu nhiên rời rạc. Nếu hàm mật độ p liên tục tại x thì đối với $\delta > 0$ rất nhỏ, ta có

$$P\Big(x - \frac{\delta}{2} \leq X \leq x + \frac{\delta}{2}\Big) = \int_{x-\delta/2}^{x+\delta/2} p(u)du \approx \delta p(x).$$

Định nghĩa 1.3.12. *Với $0 < p < 1$, phân vị cấp p của một biến ngẫu nhiên X là số x_p thỏa mãn $P(X < x_p) \leq p$ và $P(X \leq x_p) \geq p$. Phân vị cấp $1/2$ được gọi là median hay trung vị, phân vị cấp $1/4$ và $3/4$ được gọi tương ứng là tứ phân vị dưới và tứ phân vị trên của X.*

Giả sử F là hàm phân phối của biến ngẫu nhiên liên tục X, F tăng ngặt trên một khoảng I (I có thể là khoảng vô hạn), F bằng 0 bên trái I và bằng 1 bên phải I. Khi đó hàm ngược F^{-1} hoàn toàn được xác định. Mặt khác, theo định nghĩa phân vị cấp p, trong trường hợp này ta suy ra x_p là giá trị duy nhất thỏa mãn $F(x_p) = P(X \leq x_p) = p$. Do đó, $x_p = F^{-1}(p)$. Sau đây, ta sẽ xét một ví dụ về giá trị rủi ro trong tài chính.

Ví dụ 1.3.13. Tất cả các công ti tài chính đều cần phải xác định và giám sát các nguy cơ trong đầu tư của họ. Giá trị rủi ro, ký hiệu là VaR (Value at Risk), được dùng để đo các rủi ro tiềm tàng. VaR liên quan đến hai tham số là độ tin cậy và vùng thời gian (time horizon). Ví dụ, nếu VaR của một danh mục đầu tư cổ phiếu là 1 triệu USD với độ tin cậy 95% và vùng thời gian 1 ngày, thì có nghĩa là với xác suất 5%, danh mục đầu tư đó sẽ giảm giá hơn 1 triệu USD trong một ngày nếu không có sự giao dịch. Nói một cách không chính thức, trong một giai đoạn là 20 ngày, sẽ có một ngày mà nguy cơ giảm 1 triệu USD của hạng mục đầu tư này được dự báo xảy ra (do xác suất là 5%). Bạn đọc quan tâm có thể tìm hiểu thêm về VaR tại `http://en.wikipedia.org/wiki/Value_at_risk`.

Giả sử giá trị hiện tại của danh mục đầu tư là V_0 và giá trị trong tương lai là V_1. Số $R = (V_1 - V_0)/V_0$ được xem là một biến ngẫu nhiên liên tục với hàm phân phối F_R. Giả sử độ tin cậy là $1 - \alpha$ và v^* là giá trị rủi ro VaR. Khi đó

$$\begin{aligned}\alpha &= P(V_0 - V_1 \geq v^*)\\ &= P\Big(\frac{V_1 - V_0}{V_0} \leq -\frac{v^*}{V_0}\Big)\\ &= F_R\Big(-\frac{v^*}{V_0}\Big)\end{aligned}$$

Theo nhận xét ở trên, ta có $-v^*/V_0$ là phân vị cấp α. Do đó, nếu ký hiệu r_α là phân vị cấp α của biến ngẫu nhiên R, thì giá trị rủi ro VaR sẽ là

$$v^* = -V_0 r_\alpha.$$

1.4 Một số phân phối xác suất quan trọng

Như đã trình bày ở Mục 1.3.2, một biến ngẫu nhiên hoàn toàn được xác định nếu ta biết được hàm phân phối của nó. Do đó, nhiều khi thay vì nói biến ngẫu nhiên, ta chỉ nói phân phối xác suất (hàm phân phối) của biến ngẫu nhiên đó.

1.4.1 Phân phối Bernoulli và phân phối nhị thức

Giả sử ta đang quan tâm đến một biến cố A có $P(A) = p$. Hàm chỉ tiêu của A, ký hiệu là I_A hoặc $I(A)$, là hàm số xác định bởi

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

Hàm chỉ tiêu I_A được gọi là có phân phối Bernoulli với tham số p, ký hiệu là $Bern(p)$. Như vậy, phân phối Bernoulli là biến ngẫu nhiên nhận hai giá trị là 0 (khi A không xảy ra) và 1 (khi A xảy ra) với xác suất tương ứng là $1-p$ và p. Khi A xảy ra ta nói phép thử thành công, khi A không xảy ra ta nói phép thử thất bại.

Biến ngẫu nhiên X có phân phối $Bern(p)$ thức có bảng phân phối sau đây.

X	0	1
P	$1-p$	p

Hàm phân phối của biến ngẫu nhiên $Bern(p)$ là

$$F(x) = \begin{cases} 0 & \text{nếu } x < 0; \\ 1-p & \text{nếu } 0 \le x < 1; \\ 0 & \text{nếu } x \ge 1. \end{cases}$$

Giả sử có n phép thử độc lập, với n là một số nguyên dương cho trước. Mỗi phép thử đều thành công với xác suất p và thất bại với xác suất là $1-p$. Ký hiệu X là tổng số lần thành công. Khi đó, X được gọi là có phân phối nhị thức với các tham số n và p, ký hiệu là $X \sim Bin(n,p)$. Nếu ký hiệu $\{X_1, X_2, \ldots, X_n\}$ là các biến ngẫu nhiên độc lập, cùng phân phối $Bern(p)$, thì phân phối nhị thức $Bin(n,p)$ chính là tổng của các biến ngẫu nhiên X_i:

$$X = \sum_{i=1}^{n} X_i.$$

Biến ngẫu nhiên có phân phối $B(n,p)$ có tập giá trị là $\{0,1,2,\ldots,n\}$. Ta thường ký hiệu $p(k) = P(X=k)$ với $0 \leq k \leq n$. Định lý sau đây cho ta biết hàm khối lượng xác suất của phân phối nhị thức.

Định lý 1.4.1. *Giả sử X là biến ngẫu nhiên có phân phối nhị thức với các tham số n và p. Ký hiệu $p(k) = P(X=k)$, $0 \leq k \leq n$. Khi đó*

$$p(k) = C_n^k p^k (1-p)^{n-k}.$$

Chứng minh. Một dãy cụ thể nào đó với k lần thành công xảy ra với xác suất $p^k(1-p)^{n-k}$. Số các dãy như thế là C_n^k vì có C_n^k cách để có k lần thành công trong n lần phép thử. Do đó

$$P(X=k) = C_n^k p^k (1-p)^{n-k}.$$

□

Ví dụ 1.4.2. Bệnh Tay-Sachs là một loại bệnh di truyền hiếm gặp nhưng nền y học hiện đại vẫn chưa tìm được cách chữa trị. Đây là một loại bệnh do có đồng hợp tử HEXA lặn (rr) ở cặp nhiễm sắc thể số 15. Giả sử người bố và người mẹ cùng mang dị hợp tử Rr ở cặp nhiễm sắc thể này, thì xác suất để người con mắc bệnh Tay-Sachs là 25% (mang gen lặn rr).

Giả sử có một cặp vợ chồng như vậy và họ có bốn đứa con, các đứa con được sinh ra độc lập với nhau. Gọi X là biến ngẫu nhiên chỉ số đứa con mắc bệnh Tay-Sachs. Khi đó X có phân phối xác suất $Bin(4, 0.25)$:

$$p(k) = P(X=k) = C_4^k (0.25)^k (0.75)^{4-k},\ 0 \leq k \leq 4.$$

Bảng phân phối xác suất của X như sau:

X	0	1	2	3	4
P	0.316	0.422	0.211	0.047	0.004

Trong ví dụ trên, số con mắc bệnh là 1 xảy ra với xác suất cao nhất. Tổng quát, số $0 \leq k_0 \leq n$ thỏa mãn

$$p(k_0) = \max_{0 \leq k \leq n} p(k)$$

được gọi là số có khả năng nhất. Cách tìm số có khả năng nhất hoàn toàn sơ cấp, được trình bày trong định lý sau đây.

Định lý 1.4.3. *Giả sử X là biến ngẫu nhiên có phân phối nhị thức với các tham số n và p. Khi đó*

- *Nếu $np+p \in \mathbb{Z}$, thì $p(k)$ đạt cực đại tại $k = np+p$ và $np+p-1$.*
- *Nếu $np+p \notin \mathbb{Z}$, thì $p(k)$ đạt cực đại tại $k = [np+p]$.*

Chứng minh. Kết luận của định lý được suy từ việc giải các bất phương trình ẩn k:

$$p(k) \geq p(k+1) \text{ và } p(k) \geq p(k-1).$$

□

Định lý sau đây trình bày về tổng của hai biến ngẫu nhiên có cùng phân phối nhị thức.

Định lý 1.4.4. *Giả sử X và Y là hai biến ngẫu nhiên độc lập. Nếu X có phân phối $Bin(m,p)$, Y có phân phối $Bin(n,p)$, thì $X+Y$ có phân phối $Bin(m+n,p)$.*

Chứng minh. Kết luận của định lý được suy ra từ định nghĩa phân phối nhị thức. Thật vậy, trước hết ta nhận xét rằng X là số lần thành công trong m phép thử độc lập với xác suất thành công là p, Y là số lần thành công trong n phép thử độc lập với xác suất thành công là p. Ngoài ra, do X độc lập với Y nên $m+n$ phép thử ở trên là độc lập. Từ đó, $X+Y$ chính là số lần thành công trong $m+n$ phép thử độc lập với xác suất thành công là p. Do đó, $X+Y$ có phân phối $Bin(m+n,p)$.

Chúng ta có thể đưa ra cách chứng minh khác, bằng cách nhận xét rằng

$$X = \sum_{i=1}^{m} X_i,\ Y = \sum_{j=1}^{n} Y_j,$$

trong đó $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ là $m+n$ biến ngẫu nhiên độc lập, có phân phối Bernoulli $Bern(p)$. Do đó, $X+Y$ chính là tổng của $m+n$ biến ngẫu nhiên độc lập, có cùng phân phối $Bern(p)$. Nói cách khác, $X+Y$ có phân phối $Bin(m+n,p)$. □

1.4.2 Phân phối Poisson

Trong mục nhỏ này chúng ta sẽ trình bày một loại phân phối rời rạc quan trọng của lý thuyết xác suất, đó là phân phối Poisson. Phân phối Poisson được giới thiệu lần đầu tiên bởi Siméon Denis Poisson (1781-1840) năm 1837 trong cuốn "Research on the Probability of Judgments in Criminal and Civil Matters". Năm 1898, Ladislaus Bortkiewicz (1868-1931) đã tìm ra một áp dụng của phân phối này khi ông ấy được giao một nhiệm vụ điều tra số kỵ binh trong quân đội Phổ bị chết vì ngựa đá hằng năm. Ví dụ này được Bortkiewicz công bố trong cuốn "Law of small numbers" và đã làm cho phân phối Poisson được biết đến rộng rãi. Do đó, đã có nhiều ý kiến cho rằng nên đặt tên cho loại phân phối này là phân phối Poisson-Bort kiewicz.

Giả sử rằng chúng ta có n phép thử độc lập với n rất lớn và xác suất thành công của mỗi phép thử là p rất bé sao cho $np = \lambda$ là một số cố định. Gọi X là số phép thử thành công. Khi đó, theo quy luật phân phối nhị thức, ta có

$$p(k) = P(X = k) = C_n^k p^k (1-p)^{n-k},\ 0 \le k \le n.$$

Vì $np = \lambda$ nên biểu thức này trở thành

$$\begin{aligned} p(k) &= \frac{n!}{k!(n-k)!}\Big(\frac{\lambda}{n}\Big)^k\Big(1-\frac{\lambda}{n}\Big)^{n-k} \\ &= \frac{\lambda^k}{k!}\frac{n!}{(n-k)!}\frac{1}{n^k}\Big(1-\frac{\lambda}{n}\Big)^n\Big(1-\frac{\lambda}{n}\Big)^{-k} \\ &\to \frac{e^{-\lambda}\lambda^k}{k!}. \end{aligned}$$

Như vậy, thay vì tính biểu thức cồng kềnh của $p(k)$ trong phân phối nhị thức khi n lớn, ta có thể xấp xỉ nó bằng $e^{-\lambda}\lambda^k/k!$. Ta còn nhận xét rằng

$$\sum_{k=0}^{\infty}\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = 1.$$

Định nghĩa 1.4.5. *Biến ngẫu nhiên X nhận các giá trị*

$$\{0, 1, 2, \dots, n, \dots\}$$

với hàm khối lượng xác suất

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!},\ k \geq 0$$

được gọi là có phân phối Poisson với tham số λ*, ký hiệu là* $X \sim Pois(\lambda)$.

Ví dụ sau đây cho ta thấy phân phối Poisson xấp xỉ rất tốt phân phối nhị thức khi số lượng phép thử lớn và xác suất thành công nhỏ.

Ví dụ 1.4.6. Hai con xúc xắc được tung 100 lần. Gọi X là số lần có đồng thời hai con cùng xuất hiện mặt sáu chấm. Khi đó, X có phân phối $Bin(n, p)$ với $n = 100$ và $p = 1/36 \approx 0.0278$. Vì n lớn và p bé nên ta có thể xấp xỉ phân phối nhị thức này với phân phối Poisson có tham số

$$\lambda = np = 2.78.$$

Bảng sau đây cho thấy sự xấp xỉ này là tương đối tốt.

k	Phân phối nhị thức	Phân phối Poisson
0	0.0596	0.0620
1	0.1705	0.1725
2	0.2414	0.2397
3	0.2255	0.2221
4	0.1564	0.1544
5	0.0858	0.0858
6	0.0389	0.0398
7	0.0149	0.0158
8	0.0050	0.0055
9	0.0015	0.0017
10	0.0004	0.0005
11	0.0001	0.0001

Phân phối Poisson đóng mộng vai trò quan trọng trong lý thuyết và ứng dụng, được sử dụng trong rất nhiều các lĩnh vực khác nhau, trong đó có các lĩnh vực sau đây.

- Trong lĩnh vực truyền thông, số cuộc gọi đến một tổng đài trong một đơn vị thời gian có thể được mô hình như một phân phối Poisson nếu tổng đài có một số lượng lớn khách gọi đến và các cuộc gọi là tương đối độc lập.

- Một trong những ứng dụng sớm nhất là dùng phân phối Poisson để mô hình số các hạt alpha được phóng từ nguồn phóng xạ trong một đơn vị thời gian.

- Trong các công ty bảo hiểm, phân phối Poisson được sử dụng như là một mô hình cho số những tai nạn hiếm gặp.

Ví dụ 1.4.7. Bây giờ ta xét ví dụ rất thú vị của Bortkiewicz có tên là Prussian Horse-Kick (tạm dịch là quân Phổ bị ngựa đá). Số liệu sau đây được thống kê số kỵ binh bị chết do ngựa đá ở 10 sư đoàn trong quân đội Phổ trong suốt 20 năm (từ năm 1875 đến năm 1894). Do mỗi sư đoàn được quan sát trong 20 năm nên ta xem tổng số năm quan sát là 200. Số liệu này và phân phối Poisson với tham số $\lambda = 0.61$ được cho bởi bảng dưới đây. Cột thứ nhất là số lượng kỵ binh chết do ngựa đá mỗi năm, biến thiên từ 0 đến 4. Cột thứ hai là số năm quan sát tương ứng. Cột thứ ba là tỉ lệ của từng số năm tương ứng này trên tổng số 200 năm, đó cũng chính là tần số của số kỵ binh chết mỗi năm. Cột cuối cùng là phân phối xác suất Poisson với tham số $\lambda = 0.61$. Chúng ta sẽ biết cách chọn số λ phù hợp trong các chương sau. Trong ví dụ này, ta hiểu số 0.61 chính là trung bình của số kỵ binh chết do ngựa đá mỗi năm.

Số lượng lính chết mỗi năm	Số năm quan sát	Tỉ lệ của số năm	Phân phối Poisson
0	109	0.545	0.543
1	65	0.325	0.331
2	22	0.110	0.101
3	3	0.015	0.021
4	1	0.005	0.003

Định lý sau đây phát biểu về phân phối của tổng hai biến ngẫu nhiên độc lập có phân phối Poisson.

Định lý 1.4.8. *Nếu X và Y là hai biến ngẫu nhiên độc lập, $X \sim Pois(\lambda)$, $Y \sim Pois(\mu)$, thì $X + Y \sim Pois(\lambda + \mu)$.*

Phân phối Poisson thường xuất hiện trong một mô hình có tên là quá trình Poisson. Giả sử A là một biến cố nào đó mà ta quan tâm. Quá trình Poisson với tham số λ là họ các biến ngẫu nhiên $\{N_t, t \geq 0\}$ (gọi là quá trình ngẫu nhiên có thời gian liên tục) chỉ số biến cố A xuất hiện tính đến thời điểm t trên nửa trục số, thỏa mãn các điều kiện sau:

- $N_0 \equiv 0$.
- Giả sử $S_1, S_2, \ldots, S_n$ là các khoảng con bị chặn, rời nhau của $[0, \infty)$. Khi đó số các biến cố xảy ra trong các khoảng này là các biến ngẫu nhiên độc lập, lần lượt có phân phối Poisson với các tham số $\lambda|S_1|$, $\lambda|S_2|$, $\ldots$, $\lambda|S_n|$.
- Không có hai biến cố xảy ra tại cùng một thời điểm.

Bạn đọc có thể tìm hiểu quá trình Poisson nói riêng và quá trình ngẫu nhiên nói chung tại [3].

1.4.3 Phân phối đều

Nếu chúng ta nói "chọn ngẫu nhiên một số thực trong đoạn $[0, 1]$", thì mọi số thực thuộc đoạn $[0, 1]$ đều có khả năng được chọn như nhau, và không gian mẫu trong phép thử ngẫu nhiên này là

$$\Omega = [0, 1].$$

Gọi X là số được chọn thì X được gọi là biến ngẫu nhiên có phân phối đều trên đoạn $[0, 1]$. Tập giá trị của biến ngẫu nhiên X chính là đoạn

$[0,1]$ và xác suất để X rơi vào đoạn con $[x_1, x_2]$ nào đó của đoạn $[0,1]$ chính bằng $x_2 - x_1$. Hàm số

$$p(x) = \begin{cases} 1 & \text{nếu } 0 \le x \le 1, \\ 0 & \text{nếu } x < 0 \text{ hoặc } x > 1 \end{cases}$$

thỏa mãn điều kiện

$$\int_{x_1}^{x_2} p(x)dx = x_2 - x_1 = P(x_1 \le X \le x_2) \text{ với mọi } 0 \le x_1 \le x_2 \le 1.$$

Do đó, $p(x)$ là hàm mật độ của biến ngẫu nhiên X.

Tổng quát, cho $a < b$ là các số thực. Biến ngẫu nhiên X được gọi là phân phối đều trên đoạn $[a,b]$, ký hiệu là $X \sim U[a,b]$, nếu hàm mật độ $p(x)$ của nó là hằng số trên $[a,b]$ và triệt tiêu ngoài đoạn $[a,b]$:

$$p(x) = \begin{cases} \dfrac{1}{b-a} & \text{nếu } a \le x \le b, \\ 0 & \text{nếu } x < a \text{ hoặc } x > b. \end{cases}$$

Chú ý rằng $p(x) = 1/(b-a)$ trên đoạn $[a,b]$ để đảm bảo điều kiện

$$\int_{-\infty}^{+\infty} p(x)dx = \int_a^b p(x)dx = 1.$$

Hàm phân phối của phân phối $U[a,b]$ có dạng:

$$F(x) = \begin{cases} 0 & \text{nếu } x < a, \\ \dfrac{x-a}{b-a} & \text{nếu } a \le x \le b, \\ 1 & \text{nếu } x > b. \end{cases}$$

Ví dụ 1.4.9. Giả sử $X \sim U[2,12]$. Hãy tính các xác suất $P(X < 3)$, $P(X > 6)$ và $P(3 \le X < 8)$.

Đối với ví dụ này, các xác suất cần tìm được suy ra ngay từ định nghĩa biến ngẫu nhiên liên tục và hàm mật độ. Vì $X \sim U[2,12]$ nên hàm mật độ của X là

$$p(x) = \begin{cases} \dfrac{1}{10} & \text{nếu } 2 \le x \le 12, \\ 0 & \text{nếu } x < 2 \text{ hoặc } x > 12. \end{cases}$$

Do đó

$$P(X < 3) = \int_{-\infty}^{3} p(x)dx = \int_{2}^{3} \frac{dx}{10} = \frac{1}{10}.$$

Tương tự

$$P(X > 6) = \int_{6}^{12} \frac{dx}{10} = \frac{6}{10},$$

$$P(3 \leq X < 8) = \int_{3}^{8} \frac{dx}{10} = \frac{5}{10} = \frac{1}{2}.$$

Ví dụ 1.4.10. Cứ mỗi 15 phút, xe bus đến một điểm dừng, bắt đầu từ 7 : 00. Điều này nghĩa là xe bus sẽ đến lúc, 7 : 00, 7 : 15, 7 : 30, ... Giả sử một người đến điểm chờ xe bus một cách ngẫu nhiên trong khoảng từ 7 : 00 đến 7 : 30. Tính thời gian người đó phải chờ:

(a) ít hơn 5 phút,

(b) nhiều hơn 10 phút.

Trong ví dụ này, ta quy định đơn vị tính thời gian là phút. Ký hiệu X là thời điểm người ấy đến điểm dừng xe bus. Vì X có phân phối đều trên đoạn $[0, 30]$ nên người ấy phải chờ ít hơn 5 phút khi và chỉ khi người ấy đến trong khoảng thời gian từ 7 : 10 đến 7 : 15 hoặc 7 : 25 đến 7 : 30. Do đó

$$\begin{aligned} &P(\text{người ấy phải chờ ít hơn 5 phút}) \\ &= P(10 < X < 15) + P(25 < X < 30) \\ &= 2P(10 < X < 15) = 2\int_{10}^{15} \frac{dx}{30} = \frac{1}{3}. \end{aligned}$$

Tương tự, người ấy phải chờ nhiều hơn 10 phút khi và chỉ khi người ấy đến trong khoảng từ 7 : 00 đến 7 : 05 hoặc từ 7 : 15 đến 7 : 20. Do đó

$$\begin{aligned} &P(\text{người ấy phải chờ nhiều hơn hơn 5 phút}) \\ &= P(0 < X < 5) + P(15 < X < 20) \\ &= 2P(0 < X < 5) = 2\int_{0}^{5} \frac{dx}{30} = \frac{1}{3}. \end{aligned}$$

Có thể nói biến ngẫu nhiên có phân phối đều trên đoạn $[a, b]$ là biến ngẫu nhiên liên tục đơn giản nhất. Mặc dù vậy, phân phối đều có một ứng dụng là nó có thể mô phỏng bất kỳ một loại phân phối nào khác. Điều này được thể hiện qua định lý sau. Tính chất này còn được gọi là tính phổ dụng (universality) của phân phối đều.

Định lý 1.4.11. *Giả sử U là biến ngẫu nhiên có phân phối đều trên đoạn $[0, 1]$ và F là một hàm phân phối bất kỳ. Đặt*

$$F^{-1}(u) = \inf\{x : F(x) \geq u, 0 \leq u \leq 1\}.$$

Khi đó, $F^{-1}(U)$ là biến ngẫu nhiên có hàm phân phối là F. Ngược lại, nếu X là biến ngẫu nhiên có hàm phân phối là F thì $F(X)$ là biến ngẫu nhiên có phân phối đều trên đoạn $[0, 1]$.

1.4.4 Phân phối mũ

Trước khi trình bày phân phối mũ, chúng ta nhắc lại hai loại phân phối rời rạc mà ta đã trình bày ở Mục 1.3.3, đó là phân phối nhị thức $Bin(n, p)$ và phân phối hình học $Geo(p)$. Chúng ta có một dãy các phép thử độc lập với xác suất thành công ở mỗi phép thử là một số không đổi p. Khi đó, số phép thử thành công có phân phối nhị thức $Bin(n, p)$, còn số phép thử thất bại cho đến khi có một phép thử thành công là phân phối hình học $Geo(p)$. Như vậy, phân phối hình học $Geo(p)$ là số lần phép thử thất bại giữa hai lần thành công liên tiếp cộng với 1. Khi số lượng phép thử n rất lớn và xác suất thành công rất nhỏ, phân phối nhị thức $B(n, p)$ được xấp xỉ bởi phân phối Poisson $Pois(\lambda)$ với $\lambda = np$. Chúng ta hãy xét khoảng thời gian giữa hai biến cố xảy ra kế tiếp trong quá trình Poisson. Giả sử ta có quá trình Poisson với tham số λ và tại thời điểm t_0 có một biến cố xảy ra. Ký hiệu T là độ dài khoảng thời gian từ t_0 cho tới khi một biến cố khác xảy ra. Khi đó phân phối F_T của T có thể tìm từ công thức

$$\begin{aligned} 1 - F_T(t) &= P(T > t) \\ &= P\Big(\text{không có biến cố nào xảy ra trong khoảng } (t_0, t_0 + t)\Big), \ t > 0. \end{aligned}$$

Theo định nghĩa quá trình Poisson, số các biến cố xảy ra trong khoảng $(t_0, t_0 + t)$ có phân phối Poisson với tham số λt nên

$$P\Big(\text{không có biến cố nào xảy ra trong khoảng } (t_0, t_0+t)\Big) = e^{-\lambda t},\ t > 0.$$

Từ đó,

$$F_T(t) = 1 - e^{-\lambda t},\ t > 0.$$

Vì vậy, hàm phân phối của biến ngẫu nhiên T có dạng hàm số mũ và T được gọi là có phân phối mũ. Tổng quát, ta có định nghĩa sau đây.

Định nghĩa 1.4.12. *Cho $\lambda > 0$. Biến ngẫu nhiên X được gọi là có phân phối mũ với tham số λ, ký hiệu là $X \sim Exp(\lambda)$, nếu hàm mật độ của nó có dạng:*

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{nếu } x > 0, \\ 0 & \text{nếu } x \leq 0. \end{cases}$$

Từ định nghĩa trên ta dễ dàng tìm được hàm phân phối của X là

$$F(x) = \int_{-\infty}^{x} p(u)du = \begin{cases} 1 - e^{-\lambda x} & \text{nếu } x > 0, \\ 0 & \text{nếu } x \leq 0. \end{cases}$$

Như phân tích ở phía dưới định nghĩa Định nghĩa 1.3.12, để tìm phân vị cấp p của $X \sim Exp(\lambda)$ ta giải phương trình $F(x_p) = p$. Đặc biệt, median của $X \sim Exp(\lambda)$ được tìm từ phương trình ẩn ν:

$$F(\nu) = 1 - e^{-\lambda x} = \frac{1}{2}.$$

Phương trình này cho ta

$$\nu = \frac{\log 2}{\lambda}.$$

Phân phối mũ thường được sử dụng trong các mô hình thời gian chờ tại một điểm phục vụ, hoặc thời gian sống (tuổi thọ) của những thiết bị, cấu trúc không có trí nhớ. Do đó, biến x thường được ký hiệu bởi biến t. Giả sử chúng ta xem tuổi thọ của một loại thiết bị điện

tử là một biến ngẫu nhiên có phân phối mũ. Khi đó, giả sử thiết bị này đã có tuổi thọ là s và ta cần tìm xác suất để thiết bị đó có thể sống thêm ít nhất t đơn vị thời gian nữa, tức là chúng ta cần tìm $P(T > t + s|T > s)$. Điều này được tính toán như sau.

$$\begin{aligned} P(T > t + s|T > s) &= \frac{P(T > t + s, T > s)}{P(T > s)} \\ &= \frac{P(T > t + s)}{P(T > s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} = P(X > t). \end{aligned}$$

Như vậy, ta thấy rằng thời gian để thiết bị này sống thêm ít nhất t đơn vị thời gian nữa không phụ thuộc vào s. Vì tính chất này nên phân phối mũ được gọi là phân phối không có trí nhớ (memoryless). Tính chất không có trí nhớ trên đây là đặc trưng của phân phối mũ, thể hiện qua định lý sau đây.

Định lý 1.4.13. *Nếu T là biến ngẫu nhiên liên tục, không âm, và không có trí nhớ, theo nghĩa*

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

Khi đó, T có phân phối mũ với tham số $\lambda > 0$ nào đó.

Proteins và nhiều phân tử hữu cơ khác được điều hòa theo nhiều cách khác nhau. Một số trải qua quá trình lão hóa và do đó xác suất để các phân tử này phân hủy khi chúng già sẽ cao hơn xác suất để chúng phân hủy khi còn sớm. Tuy nhiên, nếu những loại không trải qua quá trình lão hóa và xác suất để chúng phân hủy là như nhau tại mọi thời điểm thì tuổi thọ của những loại phân tử này tuân theo phân phối mũ.

1.4.5 Phân phối chuẩn

Định nghĩa 1.4.14. *Cho với μ là số thực bất kỳ và $\sigma > 0$. Biến ngẫu nhiên X được gọi là có phân phối chuẩn với các tham số μ và σ^2, ký*

hiệu là $X \sim N(\mu, \sigma^2)$*, nếu hàm mật độ của nó có dạng:*

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Phân phối chuẩn đóng một vai trò trung tâm trong lý thuyết xác suất và thống kê. Trong chương sau chúng ta sẽ nghiên cứu định lý giới hạn trung tâm, phát biểu rằng tổng của các biến ngẫu nhiên độc lập được xấp xỉ bởi phân phối chuẩn. Năm 1809, Carl Friedrich Gauss (1777-1855) sử dụng phân phối chuẩn như là một mô hình cho các sai số trong đo đạc, đồng thời Gauss sử dụng phân phối chuẩn trong một phương pháp xấp xỉ nổi tiếng của mình, có tên là phương pháp bình phương tối thiểu. Do đó, phân phối chuẩn còn được gọi là phân phối Gauss.

Hàm phân phối của phân phối $N(\mu, \sigma^2)$ có dạng:

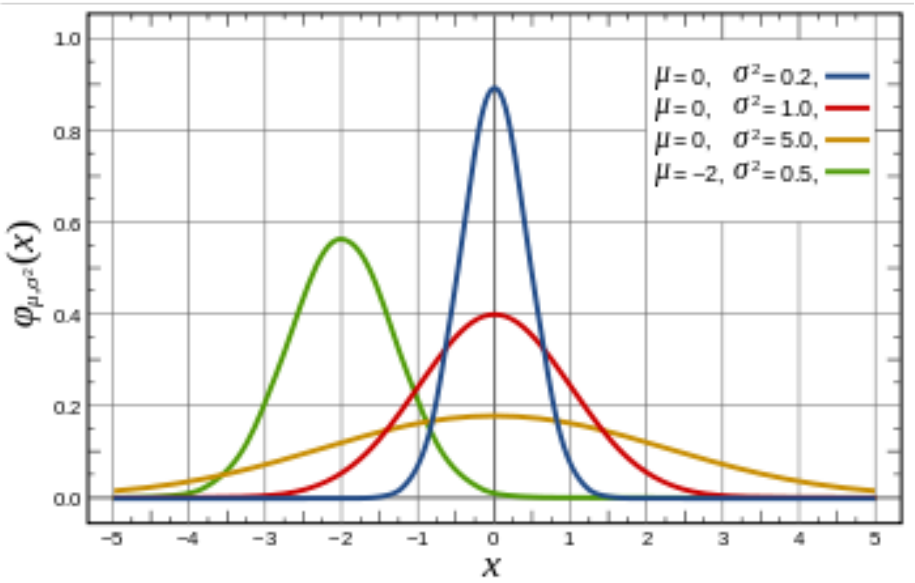
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

Khi $\mu = 0$ và $\sigma = 1$, phân phối $N(0, 1)$ được gọi là phân phối chuẩn tắc. Ta ký hiệu hàm mật độ và hàm phân phối của phân phối chuẩn tắc bởi $\varphi(x)$ và $\Phi(x)$, tương ứng.

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

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

Đồ thị hàm mật độ của phân phối chuẩn $N(\mu, \sigma^2)$ với một số giá trị cụ thể của μ và σ^2 được cho bởi hình vẽ sau đây.



Ta có thể dễ dàng chứng minh định lý sau đây.

Định lý 1.4.15. *Giả sử* $X \sim N(\mu, \sigma^2)$. *Khi đó* $aX + b \sim N(a\mu + b, a^2\sigma^2)$ *với mọi* $a, b \in \mathbb{R}, a \neq 0$.

Đặc biệt, nếu $X \sim N(\mu, \sigma^2)$ *thì*

$$\frac{X - \mu}{\sigma} \sim N(0, 1).$$

Phân phối chuẩn xuất hiện rất nhiều trong các bài toán thực tế như sai số đo đạc, chỉ số IQ, chiều cao của con người,...

Trong các bài toán thực tế, các bài toán có sự xuất hiện của phân phối chuẩn có thể được chia làm ba lớp như sau.

- Phân phối chuẩn chính xác
- Xấp xỉ phân phối chuẩn
- Có thể giả sử là phân phối chuẩn

1.5 Kỳ vọng, phương sai và độ lệch tiêu chuẩn của biến ngẫu nhiên

1.5.1 Kỳ vọng của biến ngẫu nhiên

Giả sử ở một trò chơi, ta nhận được phần thưởng là 2$ với xác suất 3/4, và 100$ với xác suất 1/4. Nếu ta chơi rất nhiều lần, dĩ nhiên ta không kỳ vọng mỗi lần đều thu về 100$, ta cũng không kỳ vọng mỗi lần thu về chỉ có 2$, mà số tiền ta kỳ vọng thu về ở mỗi lần chơi là

$$2 \times \frac{3}{4} + 100 \times \frac{1}{4} = 26.5\$.$$

Chú ý rằng nếu gọi X là số tiền ta thu về ở mỗi lần chơi thì X là biến ngẫu nhiên rời rạc có hàm khối lượng xác suất là

$$P(X = 2) = \frac{3}{4},\ P(X = 100) = \frac{1}{4}.$$

Giá trị 26.5 được gọi là kỳ vọng (hay trung bình) của số tiền ta thu về ở mỗi lần chơi. Tổng quát, ta có định nghĩa sau đây.

Định nghĩa 1.5.1. *Kỳ vọng của biến ngẫu nhiên X, ký hiệu là EX, được định nghĩa như sau:*

- *Nếu X là biến ngẫu nhiên rời rạc có hàm khối lượng xác suất $P(X = x_i) = p_i,\ i \geq 1$, thì*

$$EX := \sum_{i \geq 1} x_i p_i,$$

  với điều kiện $\sum_{i \geq 1} |x_i| p_i < \infty$.

- *Nếu X là biến ngẫu nhiên liên tục có hàm mật độ xác suất $p(x)$, thì*

$$EX := \int_{-\infty}^{\infty} x p(x) dx,$$

  với điều kiện $\int_{-\infty}^{\infty} |x| p(x) dx < \infty$.

Trong giáo trình này, khi không nói gì thêm, khi gặp biểu thức chứa kỳ vọng EX thì ta ngầm hiểu là các điều kiện trong định nghĩa đã được thỏa mãn để EX tồn tại hữu hạn.

Hai biến ngẫu nhiên được gọi là *cùng phân phối* nếu chúng có cùng hàm khối lượng xác suất (trường hợp biến ngẫu nhiên rời rạc) hoặc cùng hàm mật độ xác suất (trường hợp biến ngẫu nhiên liên tục). Từ định nghĩa Định nghĩa 1.5.1 ta dễ thấy nếu X và Y cùng phân phối thì $EX = EY$.

Sau đây, ta sẽ tính kỳ vọng của một số biến ngẫu nhiên có phân phối quen thuộc.

Ví dụ 1.5.2. Giả sử $X \sim Bern(p)$. Khi đó,

$$EX = 0P(X = 0) + 1P(X = 1) = 0 \times (1 - p) + 1 \times p = p.$$

Ví dụ 1.5.3. Giả sử $X \sim Geo(p)$. Đặt $q = 1 - p$. Khi đó,

$$\begin{aligned} EX &= \sum_{k=1}^{\infty} kP(X = k) \\ &= \sum_{k=1}^{\infty} k(1-p)^{k-1}p \\ &= p\sum_{k=1}^{\infty} kq^{k-1} \\ &= p\sum_{k=1}^{\infty} \frac{d}{dq}q^k \\ &= p\frac{d}{dq}\Big(\sum_{k=1}^{\infty} q^k\Big) \\ &= p\frac{d}{dq}\Big(\frac{q}{1-q}\Big) \\ &= \frac{p}{(1-q)^2} = \frac{1}{p}. \end{aligned}$$

Ví dụ 1.5.4. Giả sử $X \sim Pois(\lambda)$. Khi đó,

$$\begin{aligned} EX &= \sum_{k=0}^{\infty} kP(X=k) \\ &= \sum_{k=0}^{\infty} k \times e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda. \end{aligned}$$

Ví dụ 1.5.5. Giả sử $X \sim U[a,b]$. Khi đó,

$$EX = \int_a^b \frac{xdx}{a-b} = \frac{a+b}{2}.$$

Ví dụ 1.5.6. Giả sử $X \sim Exp(\lambda)$. Khi đó,

$$EX = \int_0^{\infty} \lambda x e^{-\lambda x}dx = \int_0^{\infty} xd(-e^{-\lambda x}) = -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}dx = \frac{1}{\lambda}.$$

Ví dụ 1.5.7. Giả sử $X \sim N(0,1)$. Khi đó,

$$EX = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-x^2/2}dx = 0.$$

Trong rất nhiều trường hợp, ta phải tính $Eg(X)$, trong đó g là một hàm số. Định lý sau đây nêu cách tính $Eg(X)$ khi biết hàm trọng số xác suất hoặc mật độ xác suất của X.

Định lý 1.5.8. *Giả sử $g : \mathbb{R} \to \mathbb{R}$ là một hàm số và X là biến ngẫu nhiên. Khi đó*

- *Nếu X là biến ngẫu nhiên rời rạc có hàm khối lượng xác suất $P(X = x_i) = p_i,\ i \geq 1$, thì*

$$Eg(X) = \sum_{i\geq 1} g(x_i)p_i,$$

với điều kiện $\sum_{i\geq 1} |g(x_i)|p_i < \infty$.

- *Nếu X là biến ngẫu nhiên liên tục có hàm mật độ xác suất $p(x)$, thì*
$$Eg(X) = \int_{-\infty}^{\infty} g(x)p(x)dx,$$
với điều kiện $\int_{-\infty}^{\infty} |g(x)|p(x)dx < \infty$.

Kết quả sau đây là một hệ quả trực tiếp của định lý trên.

Hệ quả 1.5.9. *Nếu X và Y là các biến ngẫu nhiên cùng phân phối thì $Eg(X) = Eg(Y)$ với mọi hàm $g : \mathbb{R} \to \mathbb{R}$.*

Ví dụ 1.5.10. Giả sử $X \sim Bern(p)$. Khi đó,

$$EX = 0^2 P(X = 0) + 1^2 P(X = 1) = 0^2 \times (1 - p) + 1^2 \times p = p.$$

Ví dụ 1.5.11. Giả sử $X \sim N(0, 1)$. Khi đó,

$$\begin{aligned} EX^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x d(-e^{-x^2/2}) \\ &= \frac{1}{\sqrt{2\pi}} \Big[- xe^{-x^2/2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2} dx \Big] \\ &= \frac{1}{\sqrt{2\pi}} \Big(0 + \sqrt{2\pi} \Big) = 1. \end{aligned}$$

Tính chất của kỳ vọng thể hiện qua định lý sau đây.

Định lý 1.5.12. *Cho X, Y là các biến ngẫu nhiên và a, b là các số thực. Khi đó*

- *(i) $E(a) = a$.*
- *(ii) Nếu $X \geq 0$ thì $E(X) \geq 0$.*
- *(iii) $E(aX + bY) = aE(X) + bE(Y)$.*
- *(iv) Nếu X và Y là các biến ngẫu nhiên độc lập, thì $E(XY) = EXEY$.*

Ví dụ 1.5.13. Giả sử $X \sim Bin(n,p)$. Khi đó,

$$X = X_1 + X_2 + \cdots + X_n,$$

trong đó $X_k, 1 \leq k \leq n$ là các biến ngẫu nhiên độc lập, có cùng phân phối $Bern(p)$. Từ đó ta có

$$\begin{aligned} EX &= E(X_1 + \ldots EX_n) \\ &= EX_1 + \cdots + EX_n \\ &= nEX_1 = np, \end{aligned}$$

và

$$\begin{aligned} EX^2 &= E(X_1 + \ldots EX_n)^2 \\ &= \sum_{k=1}^{n} EX_k^2 + 2 \sum_{1 \leq i < j \leq n} E(X_i X_j) \\ &= nEX_1^2 + n(n-1)E(X_1X_2) \\ &= np + n(n-1)EX_1EX_2 \\ &= np + n(n-1)p^2. \end{aligned}$$

1.5.2 Phương sai và độ lêch tiêu chuẩn

Nếu như kỳ vọng là một tham số biểu thị trung bình của biến ngẫu nhiên thì độ lệch tiêu chuẩn lại là tham số biểu thị cho sự phân tán của giá trị của biến ngẫu nhiên quanh giá trị trung bình. Trước hết, ta sẽ giới thiệu khái niệm phương sai, sau đó ta sẽ định nghĩa độ lệch tiêu chuẩn thông qua phương sai.

Định nghĩa 1.5.14. *Phương sai của biến ngẫu nhiên X là một số, ký hiệu bởi $D(X)$, được định nghĩa*

$$D(X) = E(X - EX)^2,$$

với điều kiện là kỳ vọng của biến ngẫu nhiên ở vế phải trong biểu thức trên tồn tại. Độ lệch tiêu chuẩn của X là $\sqrt{D(X)}$.

Từ Định nghĩa 1.5.14 ta thấy rằng $D(X)$ chính là trung bình của bình phương độ lệch của X ra khỏi giá trị trung bình $E(X)$. Nếu X có đơn vị là m, thì $D(X)$ sẽ có đơn vị là m^2 và do đó độ lệch tiêu chuẩn $\sqrt{D(X)}$ sẽ có đơn vị là m, cùng thứ nguyên với X. Ta thường ký hiệu DX là σ^2 và $\sqrt{D(X)}$ là σ.

Định lý sau đây là một hệ quả trực tiếp của Định lý 1.5.8.

Định lý 1.5.15. *Cho X là biến ngẫu nhiên với $E(X) = \mu$ và $D(X)$ tồn tại. Khi đó*

- *Nếu X là biến ngẫu nhiên rời rạc có hàm khối lượng xác suất $P(X = x_i) = p_i,\ i \geq 1$, thì*

$$D(X) = \sum_{i \geq 1} (x_i - \mu)^2 p_i.$$

- *Nếu X là biến ngẫu nhiên liên tục có hàm mật độ xác suất $p(x)$, thì*

$$D(X) = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx.$$

Trong thực tế, ta thường quan tâm đến độ lệch chuẩn nhiều hơn. Tuy nhiên, cách tính độ lệch tiêu chuẩn thông qua phương sai là một cách thuận lợi nhất. Ta sẽ xét hai ví dụ đơn giản về tính phương sai bằng cách sử dụng Định lý 1.5.15

Ví dụ 1.5.16. Giả sử $X \sim Bern(p)$. Khi đó, theo Ví dụ 1.5.2, ta đã có $EX = p$. Từ đó

$$D(X) = (0 - p)^2 \times (1 - p) + (1 - p^2 \times p = p(1 - p).$$

Ví dụ 1.5.17. Giả sử $X \sim U[a, b]$. Khi đó, theo Ví dụ 1.5.5, ta đã có

$$EX = \int_a^b \frac{x dx}{a - b} = \frac{a + b}{2}.$$

Từ đó

$$D(X) = \int_a^b \Big(x - \frac{a + b}{2}\Big) \frac{dx}{b - a} = \frac{(b - a)^2}{12}.$$

Định lý sau đây là một kết quả đơn giản, nhưng rất hay sử dụng khi tính phương sai.

Định lý 1.5.18. *Giả sử X là biến ngẫu nhiên và $D(X)$ tồn tại hữu hạn. Khi đó*

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

Ngoài ra, phương sai còn có một số tính chất sau đây.

Định lý 1.5.19. *Cho X, Y là các biến ngẫu nhiên và a là số thực. Khi đó*

- *(i) $D(X) \geq 0,\ D(a) = 0.$*
- *(i) $D(aX) = a^2 D(X).$*
- *(iii) Nếu X và Y là các biến ngẫu nhiên độc lập, thì $D(X \pm Y) = D(X) + D(Y).$*

Sau đây, ta sẽ xét một số ví dụ về tính phương sai của một số biến ngẫu nhiên có phân phối quen thuộc.

Ví dụ 1.5.20. Giả sử $X \sim Geo(p)$. Đặt $q = 1 - p$. Khi đó, theo Ví dụ 1.5.3, ta đã có

$$E(X) = \frac{1}{p}.$$

Mặt khác,

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\infty} k^2 P(X = k) \\ &= \sum_{k=1}^{\infty} k^2 (1-p)^{k-1} p \\ &= p \sum_{k=1}^{\infty} k^2 q^{k-1} \end{aligned}$$

Ta thấy

$$\sum_{k=1}^{\infty} q^k = \frac{q}{1-q}.$$

Lấy đạo hàm hai vế theo biến q, ta được

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}.$$

Điều này kéo theo

$$\sum_{k=1}^{\infty} kq^{k} = \frac{q}{(1-q)^2}.$$

Tiếp tục lấy đạo hàm hai vế của đẳng thức trên, ta có

$$\sum_{k=1}^{\infty} k^2q^{k-1} = \frac{1+q}{(1-q)^3}.$$

Do đó

$$E(X)^2 = \frac{(1+q)p}{(1-q)^3} = \frac{(1+q)}{p^2}.$$

Từ đó

$$D(X) = E(X^2) - (E(X))^2 = \frac{(1+q)}{p^2} - \frac{1}{p^2} = \frac{q}{p^2} = \frac{1-p}{p^2}.$$

Ví dụ 1.5.21. Giả sử $X \sim Pois(\lambda)$. Khi đó, theo Ví dụ 1.5.4, ta đã có

$$E(X) = \lambda.$$

Mặt khác,

$$\begin{aligned}
EX^2 &= \sum_{k=0}^{\infty} k^2 \times e^{-\lambda}\frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{k\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{d}{d\lambda}\frac{\lambda^k}{(k-1)!} \\
&= \lambda e^{-\lambda} \frac{d}{d\lambda}\Big(\sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}\Big) \\
&= \lambda e^{-\lambda} \frac{d}{d\lambda}\Big(\lambda e^{\lambda}\Big) \\
&= \lambda e^{-\lambda}(e^{\lambda} + \lambda e^{\lambda}) = \lambda + \lambda^2.
\end{aligned}$$

Từ đó

$$D(X) = E(X^2) - (E(X))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda.$$

Ví dụ 1.5.22. Giả sử $X \sim Exp(\lambda)$. Khi đó, theo Ví dụ 1.5.6, ta đã có

$$E(X) = \frac{1}{\lambda}.$$

Mặt khác

$$\begin{aligned} E(X^2) &= \int_0^\infty \lambda x^2 e^{-\lambda x} dx \\ &= \int_0^\infty x^2 d(-e^{-\lambda x}) \\ &= -x^2 e^{-\lambda x}\Big|_0^\infty + 2\int_0^\infty x e^{-\lambda x} dx \\ &= 0 + \frac{2}{\lambda}\Big[\int_0^\infty x d(-e^{-\lambda x})\Big] \\ &= \frac{2}{\lambda}\Big[-xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx\Big] \\ &= \frac{2}{\lambda}\Big[0 + \frac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty\Big] = \frac{2}{\lambda^2}. \end{aligned}$$

Từ đó

$$D(X) = E(X^2) - (E(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Ví dụ 1.5.23. Giả sử $X \sim N(\mu, \sigma^2)$. Khi đó,

$$Y = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Vì $E(Y) = 0, E(Y^2) = 1$, nên $D(Y) = 1$. Từ đó

$$E(X) = E(\sigma Y + \mu) = \sigma E(Y) + \mu = \mu,$$

$$D(X) = D(\sigma Y + \mu) = D(\sigma Y) + 0 = \sigma^2 D(Y) = \sigma^2.$$

1.6 Vector ngẫu nhiên

1.6.1 Giới thiệu

Cho số nguyên $d \geq 2$ và $\{X_k, 1 \leq k \leq d\}$ là d biến ngẫu nhiên xác định trên không gian mẫu Ω. Khi đó bộ $X = (X_1, X_2, \ldots, X_d)$ được gọi là vector ngẫu nhiên d chiều. Như vậy, vector ngẫu nhiên d chiều là một ánh xạ xác định trên không gian mẫu Ω và nhận giá trị trong $\mathbb{R}^d$.

Vector ngẫu nhiên xuất hiện trong nhiều lĩnh vực khác nhau, ta có thể nêu một số ví dụ sau.

- Trong nghiên cứu sinh thái, người ta thường phải một số loài nào đó. Số phần tử (số con) của các loài này là các biến ngẫu nhiên. Trong thực tế, loài này có thể là mồi cho loài kia. Rõ ràng, số phần tử (số con) của loài động vật ăn thịt phụ thuộc vào số thành viên con mồi.

- Khi nghiên cứu tính nhiễu của khí quyển, tọa độ vận tốc gió được mô hình bởi bộ các biến ngẫu nhiên (X, Y, Z).

- Trong nghiên cứu y học, các yếu tố sinh lý của bệnh nhân cũng được mô hình bởi các biến ngẫu nhiên và ta cần nghiên cứu bộ các biến ngẫu nhiên này.

Định nghĩa 1.6.1. *Hàm phân phối liên kết của vector ngẫu nhiên $(X_1, \ldots, X_d)$, ký hiệu là F, là một hàm số xác định trên $\mathbb{R}^d$, định nghĩa bởi:*

$$F(x_1, x_2, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d), \ (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

Như vậy, giả sử F là hàm phân phối liên kết của vector ngẫu nhiên (X, Y). Khi đó

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1).$$

Hàm phân phối liên kết của vector ngẫu nhiên có các tính chất tương tự như hàm phân phối của biến ngẫu nhiên, thể hiện qua định lý sau.

Định lý 1.6.2. *Giả sử F là hàm phân phối của vector ngẫu nhiên (X, Y), và F_X, F_Y lần lượt làm hàm phân phối của các biến ngẫu nhiên X, Y. Khi đó, ta có các khẳng định sau đây.*

- *Nếu cố định một biến, thì F là hàm không giảm, liên tục trái và có giới hạn phải theo biến còn lại.*
- $$\lim_{x\to\infty} F(x,y) = F_Y(y), \text{ với mọi } y \in \mathbb{R}; \quad \lim_{y\to\infty} F(x,y) = F_X(x), \text{ với mọi } x \in \mathbb{R}.$$
- $$\lim_{x\to-\infty} F(x,y) = 0, \text{ với mọi } y \in \mathbb{R}; \quad \lim_{y\to-\infty} F(x,y) = 0, \text{ với mọi } x \in \mathbb{R}.$$
- $$\lim_{x\to-\infty,y\to-\infty} F(x,y) = 0, \quad \lim_{x\to\infty,y\to\infty} F(x,y) = 1.$$
- *Tập các điểm gián đoạn của F là hữu hạn hoặc vô hạn đếm được.*

1.6.2 Vector ngẫu nhiên rời rạc

Nếu X, Y là các biến ngẫu nhiên rời rạc thì vector (X, Y) được gọi là vector ngẫu nhiên rời rạc. Hàm khối lượng xác suất liên kết, ký hiệu là $p(x, y)$ được định nghĩa bởi

$$p(x_i, y_j) = P(X = x_i, Y = y_j).$$

Sau đây ta sẽ xét một số ví dụ.

Ví dụ 1.6.3. Tung một đồng xu cân đối 3 lần. Gọi X là số mặt ngửa xuất hiện ở lần tung thứ nhất, và Y là tổng số mặt ngửa xuất hiện ở cả ba lần tung. Khi đó, hàm khối lượng xác suất của vector ngẫu nhiên (X, Y) thể hiện qua bảng sau đây, gọi là bảng phân phối. Trong cuốn sách này, ta quy ước cột đầu tiên của bảng là các giá trị của X, dòng đầu tiên của bảng là các giá trị của Y.

	0	1	2	3
0	1/8	2/8	1/8	0
1	0	1/8	2/8	1/8

Ví dụ 1.6.4 (Phân phối đa thức)**.** Giả sử ta có dãy gồm n phép thử độc lập. Mỗi phép thử đều có r kết quả với xác suất tương ứng là $p_1, p_2, \ldots, p_r$. Ký hiệu N_i là số biến cố dạng i xuất hiện trong n phép thử. Mỗi dãy n phép thử cụ thể cho ra kết quả $N_1 = n_1, \ldots, N_r = n_r$ với xác suất

$$p_1^{n_1} \ldots p_r^{n_r}.$$

Dễ chứng minh được có

$$\frac{n!}{n_1! \ldots n_r!}$$

dãy như trên. Do đó, hàm khối lượng xác suất liên kết của vector ngẫu nhiên $(N_1, N_2, \ldots, N_r)$ có dạng

$$P(N_1 = n_1, \ldots, N_r = n_r) = \frac{n!}{n_1! \ldots n_r!} p_1^{n_1} \ldots p_r^{n_r}.$$

Vì N_i là số biến cố dạng i xuất hiện trong n phép thử nên N_i có phân phối nhị thức $Bin(n, p_i)$. Như vậy, với $1 \leq i \leq r$, hàm khối lượng xác suất của N_i là

$$P(N_i = k) = C_n^k p_i^k (1 - p_i)^{n-k}, \ 0 \leq k \leq n.$$

Tóm tắt

Chương này giới thiệu các kiến thức cơ bản của lý thuyết xác suất như không gian mẫu, biến cố ngẫu nhiên, độ đo xác suất, xác suất có điều kiện, và tính độc lập. Qua đó, chương này giúp sinh viên tính xác suất của các biến cố, xác suất có điều kiện, lập bảng phân phối, hàm phân phối, tính kỳ vọng, phương sai, độ lệch tiêu chuẩn của biến ngẫu nhiên.

Bài tập

1. Chứng minh các đẳng thức sau đây bằng cách phân tích ý nghĩa tổ hợp của chúng.

a. $C_n^k = C_n^{n-k}$.

b. $C_n^k = C_{n-1}^{k-1} + C_{n-1}^k$.

c. $C_m^n = \sum_{k=0}^n C_n^k C_{m-n}^{n-k}$.

2. Trong một trò chơi bài, có ba người chơi và mỗi người được chia 3 con bài từ bộ bài Tây 52 con. Hỏi có bao nhiêu cách chia như vậy?

3. Trong trò chơi bài tiến lên, có bốn người chơi và mỗi người được chia 13 con bài từ bộ bài Tây 52 con. Hỏi có bao nhiêu cách chia như vậy?

4. Có 4 loại thịt, 6 loại rau và 3 loại lương thực. Hỏi có bao nhiêu cách chia để nấu một bữa ăn nếu bữa ăn gồm có một loại từ ba nhóm trên.

5. Chứng minh bất đẳng thức Bonferroni sau đây:

$$P(AB) \geq P(A) + P(B) - 1.$$

6. Chứng minh bất đẳng thức:

$$P\Big(\bigcup_{j=1}^{n} A_j\Big) \leq \sum_{j=1}^{n} P(A_j).$$

7. Bản tin dự báo thời tiết nói rằng xác suất có tuyết rơi vào tháng mười một là 50%, và xác suất có tuyết rơi vào tháng mười hai là 50%. Hỏi khẳng định có rét đậm vào ba tháng cuối năm là 100% đúng không? Tại sao?

8. Giải đặc biệt của xổ số Miền Bắc là một dãy số gồm 5 chữ số. Hai số đầu tiên của giải đặc biệt ngày hôm qua là 38. Hỏi xác suất để giải đặc biệt ngày hôm qua gồm 5 chữ số khác nhau là bao nhiêu?

9. Một người rút ngẫu nhiên 5 con bài từ bộ bài Tây 52 con. Tính xác suất của các biến cố sau đây.

a. Năm con có giá trị liên tiếp nhau.

b. Có một bộ tứ quý.

c. Có 3 con K và 2 con J.

10. Tung một đồng xu cân đối ba lần. Tính xác suất của các biến cố sau đây.

a. Có đúng hai mặt ngửa xuất hiện.

b. Có ít nhất một mặt ngửa xuất hiện.

11. Một ủy ban gồm có 5 người Châu Âu, 2 người Châu A, 3 người Châu Phi, và 6 người châu Mỹ. Người ta thành lập một ban điều hành gồm 4 người được chọn ngẫu nhiên từ ủy ban. Tính xác suất để ban điều hành có đủ đại diện cho bốn châu lục nói trên?

12. Có 12 khách đứng ở sân ga để lên 3 toa tàu một cách ngẫu nhiên. Biết rằng mỗi toa có 12 chỗ trống. Tính xác suất để

a. Cả 12 khách cùng lên cùng một toa.

b. Số khách ở mỗi toa bằng nhau.

c. Tất cả các toa đều có khách lên tàu.

13. Có n ($n \geq 4$) cặp vợ chồng đến dự tiệc và sau đó mỗi quý ông mời ngẫu nhiên một người phụ nữ để nhảy. Tính xác suất của các biến cố sau đây.

a. Có đúng 2 cặp mà bạn nhảy khác với người phối ngẫu của mình.

b. Có đúng 3 cặp mà bạn nhảy khác với người phối ngẫu của mình.

c. Có đúng 4 cặp mà bạn nhảy khác với người phối ngẫu của mình.

14. Một bộ bài Tây gồm 52 con được xếp thành một hàng. Tính xác suất để bốn con K ở cạnh nhau.

15. Tung một đồng xu cân đối 5 lần. Tính xác suất của biến cố có chứa 3 mặt ngửa và 2 mặt sấp.

16. Một bộ bài Tây gồm 52 con được xếp sấp xuống thành một hàng, sau đó người ta lật ngửa lên n con. Tính xác suất để ít nhất một con mặt người (J, Q, hoặc K) được lật lên. Tìm n nhỏ nhất để xác suất này vượt quá 0.5.

17. Có n người lên n toa tàu còn trống một cách ngẫu nhiên. Giả sử rằng mỗi toa còn chứa được n người. Tính xác suất để có đúng một toa vắng khách.

18. Có n người lên n toa tàu còn trống một cách ngẫu nhiên. Giả sử rằng mỗi toa còn chứa được n người. Tính xác suất để toa cuối cùng có j người.

19. Một người đến Casino đánh bạc với số vốn ban đầu là 1000\$, xác suất thắng ở mỗi ván là p và xác suất thua ở mỗi ván là $q = 1 - p$. Nếu thắng, anh ấy được 10\$ và nếu thua thì mất 10\$. Nếu hết tiền anh ấy phải ra về. Tính xác suất của các biến cố sau đây.

a. Tính xác suất để sau 100 ván, anh ấy giữ nguyên số vốn ban đầu.

b. Tính xác suất để sau 120 ván, anh ấy giữ nguyên số vốn ban đầu.

2

Thống kê và các kết luận thống kê

2.1 Mở đầu

Như chúng ta biết Lý thuyết xác suất được sinh ra từ việc nghiên cứu các quy luật ngâu nhiên ẩn sau các bài toán thực tế. Thông qua việc nghiên cứu các bài toán thực tế chúng ta tìm ra các quy luật ngẫu nhiên. Những quy luật đó chúng ta đã được học trong chương 1.

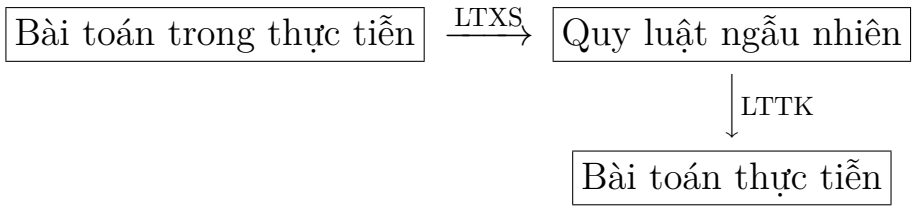
Một câu hỏi tự nhiên được đặt ra là: *Những quy luật ngẫu nhiên đó được nghiên cứu ra để làm gì? Có làm tăng sự hiểu biết của chúng ta về các hiện tượng tự nhiên và xác hội hay không?*

Câu trả lời là: Các quy luật ngẫu nhiên của lý thuyết xác suất sau khi được nghiên cứu ra thì nó được sử dụng để nghiên cứu các bài toán trong thực tế và người ta thường gọi bài toán thống kê. Vậy *Thống kê là gì?* Cho đến nay chúng ta có rất nhiều thuật ngữ thống kê khác nhau trong thực tế chẳng hạn như:

$$\begin{cases} \text{- Thống kê kinh tế} \\ \text{- Thống kê sinh học} \\ \text{- Vật lý thống kê} \\ \ldots\ldots \end{cases} \Leftarrow \text{Dùng } \textbf{thống kê toán học} \text{ làm công cụ.}$$

- Vậy *thống kê toán* học là gì? Nó có nhiệm vụ gì? Gồm những nội dung nào?

Thống kê toán học là phần ứng dụng của lý thuyết xác suất



Nhiệm vụ của thống kê Toán học

Thống kê toán học nghiên cứu các phương pháp thu thập, phân tích, xữ lý các số liệu thống kê để đưa ra các quyết định có cơ sở khoa học phục vụ cho việc quản lý xã hội.

Nội dung của thống kê toán học

Lý thuyết thống kê toán học có các nội dung cơ bản sau đây:

- Bài toán về lý thuyết chọn mẫu
- Bài toán ước lượng tham số
- Bài toán kiểm định giả thiết
- Bài toán phân tích tương quan và hồi quy

2.2 Thống kê mô tả

2.2.1 Tổng thể và mẫu ngẫu nhiên

Trong thực tế, nhiều khi ta cần quan tâm đến một số đặc điểm (định tính hoặc định lượng) của các phần tử thuộc về một tập hợp nào đó, chẳng hạn tuổi thọ của một loại sản phẩm nào đó, thu nhập trung bình củ người dân ở một quốc gia, tỉ lệ sản phẩm đạt tiêu chuẩn, tỉ lệ người dân bỏ phiếu cho một ứng cử viên nào đó, tỉ lệ cá thể nhiễm bệnh trong quần thể, ... Tập hợp các phần tử cần nghiên cứu này được gọi là *đám đông* hay *tổng thể*, ký hiệu là $\mathcal{C}$.

Việc tiến hành thu thập thông tin trên các phần tử của đám đông được gọi là *quan sát*.

Thuộc tính của một đối tượng mà chúng ta quan tâm thường là chúng ta chưa biết hoặc biết chưa đầy đủ và chúng ta coi đó là được coi như một đại lượng ngẫu nhiên, ký hiệu là X và được gọi là *đại lượng ngẫu nhiên gốc đám đông* $\mathcal{C}$. Quá trình đi nghiên cứu đám đông của $\mathcal{C}$ thực chất là quá trình đi tìm quy luật phân phối của đại lượng ngẫu nhiên X, nhiều khi đó là quá trình đi tìm các số đặc trưng của X.

Đặc điểm của đám đông (tổng thể) thường được nghiên cứu dưới hai phương diện:

◇ Phương diện định lượng: Khi ta cần quan tâm đến các giá trị về lượng của đại lượng ngẫu nhiên X như: trọng lượng, năng suất, tuổi thọ, ... và ta thường quan tâm đến hai đặc trưng:

- Kỳ vọng $\mathbb{E}X = \mu$: đặc trưng giá trị trung bình của đặc điểm định lượng cần quan tâm trên đám đông $\mathcal{C}$.

- Phương sai $\mathbb{D}X = \sigma^2$: đặc trưng cho mức độ biến động giá trị của đặc điểm định lượng cần quan tâm trên đám đông $\mathcal{C}$.

◇ Phương diện định tính: Khi ta cần quan tâm đến một tính chất A nào đó trên đám đông, các phần tử của đám đông hoặc có tính chất A hoặc không có tính chất A như: chất lượng sản phẩm, sự nẩy mầm của một giống lúa, chất độc hại trong nguồn nước, ... Giá trị mà đại lượng ngẫu nhiên X có thể nhận được

$$X = \begin{cases} 1 & \text{khi phần tử đó có tính chất } A\,; \\ 0 & \text{khi phần tử đó không có tính chất } A\,, \end{cases}$$

và ta thường quan tâm đến xác suất $\mathbb{E}X = p$.

Chúng ta khó có thể quan sát hết tất cả các phần tử của đám đông vì những lý do như thời gian, chi phí tốn kém, ... Chính vì vậy, người ta chỉ lấy ra một số phần tử đại diện cho đám đông và nghiên cứu trên tập phần tử này, tập hợp các phần tử đại diện cho đám đông đó được gọi là *mẫu*. Phương pháp nghiên cứu trên mẫu đại diện cho đám đông được gọi là *phương pháp mẫu* và cách thức thực hiện quá trình lấy mẫu được gọi là *phương pháp lấy mẫu*.

Khi cần quan tâm đến đặc điểm là đại lượng ngẫu nhiên X của đám đông $\mathcal{C}$, ta chọn ra mẫu có n phần tử, trong đó việc chọn phần tử thứ i là quá trình thực hiện một phép thử rút ngẫu nhiên một phần tử của đám đông $\mathcal{C}$, giá trị ngẫu nhiên này được gán cho đại lượng ngẫu nhiên X_i. Với cách chọn này, các đại lượng ngẫu nhiên X_i độc lập với nhau và có cùng luật phân phối với đại lượng ngẫu nhiên X. Mẫu này được gọi là *mẫu ngẫu nhiên* có kích thước n của đám đông $\mathcal{C}$. Vậy *mẫu ngẫu nhiên là gì*?.

Định nghĩa 2.2.1. *Cho $X_1, \cdots X_n$ là dãy các biến ngẫu nhiên độc lập cùng phân phối với biến ngẫu nhiên X. Khi đó véc tơ ngẫu nhiên $(X_1, \cdots X_n)$ được gọi là mẫu ngẫu nhiên cỡ n lấy từ X. Một bộ giá trị $(x_1, ..., x_n)$ của véc tơ ngẫu nhiên $(X_1, \cdots X_n)$ được gọi là một thể hiện của mẫu ngẫu nhiên hay thường gọi là* một mẫu cụ thể.

Ví dụ 2.2.2. *Thống kê về số chấm của một con xúc xắc khi gieo 5 lần. Mẫu ngẫu nhiên:* $(X_1, X_2, ..., X_5)$ *; mẫu cụ thể:* $(2, 3, 1, 6, 2)$.

Các phương pháp lấy mẫu

Việc lấy mẫu được coi là tốt nếu như thông tin thu được từ mẫu phản ánh càng gần với đặc điểm của đám đông (tính chất đại diện cao). Chính vì vậy, trong thống kê việc lấy mẫu là một công việc hết sức quan trọng. Người ta thường sử dụng một số phương pháp lấy mẫu như sau:

Lấy mẫu ngẫu nhiên đơn giản

Là phương pháp lấy mẫu thỏa mãn các điều kiện: mỗi lần chỉ được chọn một phần tử từ đám đông, khả năng được chọn của tất cả các phần tử trong đám đông đều như nhau. Có hai cách thức tiến hành chọn, đó là chọn hoàn lại và chọn không hoàn lại, tuy nhiên khi kích thước của đám đông lớn hơn nhiều so với kích thước mẫu thì có thể coi hai phương pháp chọn này là giống nhau.

Phương pháp lấy mẫu ngẫu nhiên đơn giản ở trên có tính chất đại diện cho đám đông cao, tuy nhiên nó khó thực hiện và cần nhiều thời

gian cũng như kinh phí. Ta có thể xem phương pháp lấy mẫu này là hoàn toàn ngẫu nhiên hay ngẫu nhiên không có định hướng.

Lấy mẫu ngẫu nhiên có định hướng

◇ Lấy mẫu theo nhóm: là phương pháp chia đám đông thành các nhóm thuần nhất, từ mỗi nhóm này ta lấy ra một mẫu ngẫu nhiên đơn giản với một kích thước tương ứng. Tập hợp tất cả các phần tử thu được từ các mẫu ngẫu nhiên đơn giản đó lập nên mẫu ngẫu nhiên theo nhóm.

◇ Lấy mẫu theo chùm: là phương pháp chia đám đông thành nhiều chùm (đám đông con) sao cho giữa các chùm có sự đồng đều về quy mô, từ các chùm đó ta lấy một mẫu ngẫu nhiên đơn giản. Tập hợp tất cả phần tử thu được từ các mẫu ngẫu nhiên đơn giản của các chùm lập nên mẫu ngẫu nhiên theo chùm.

Phương pháp này dễ quy hoạch, có thể tiết kiệm được thời gian và kinh phí nhưng sai số chọn mẫu cao hơn các phương pháp nói trên.

Ví dụ 2.2.3. *Chúng ta muốn đi tìm hiểu về tổng thu nhập trong một năm của toàn bộ cán bộ công chức của một tỉnh.*

- Chia đám đông này thành các nhóm theo từng cơ cấu ngành nghề: quốc phòng, an ninh, giáo dục, y tế, kinh doanh, ... Trong mỗi cơ cấu ngành nghề có sự thuần nhất về mức lương (nếu có sự sai khác về thu nhận chủ yếu là do thâm niên và chức vụ công tác). Như vậy, phương pháp lấy mẫu bằng việc gom lại các mẫu ngẫu nhiên đơn giản của từng nhóm ngành nghề chính là phương pháp lấy mẫu theo nhóm.

- Chia đám đông này theo các huyện trong tỉnh A. Giữa các huyện, có sự đồng đều về quy mô (đầy đủ các thành phần) và phương pháp lấy mẫu bằng việc gom lại các mẫu ngẫu nhiên đơn giản của từng huyện chính là phương pháp lấy mẫu theo chùm.

2.2.2 Cách biểu diễn mẫu

Bảng tần số và bảng tần suất

Ta thực hiện n lần quan sát trên đám đông $\mathcal{C}$, khi đó ta sẽ thu được mẫu cụ thể gồm k giá trị khác nhau $(x_1, x_2, ..., x_k),\ k \leqslant n$. Giá trị x_i

có n_i lần xuất hiện, n_i được gọi là *tần số xuất hiện* của x_i và tỉ số $\dfrac{n_i}{n}$ được gọi là *tần suất xuất hiện* của x_i, ký hiệu là f_i. Ta có biểu diễn kết quả của mẫu bằng bảng tần số và tần suất như sau

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

x_i	x_1	x_2	...	x_k
f_i	f_1	f_2	...	f_k

trong đó

$$n = \sum_{i=1}^{k} n_i; \quad \sum_{i=1}^{k} f_i = 1.$$

Ví dụ 2.2.4. *Thống kê điểm số kết thúc học phần của một lớp gồm 40 sinh viên, ta có*

x_i	4	5	6	7	8
n_i	5	10	12	8	5

x_i	4	5	6	7	8
f_i	5/40	10/40	12/40	8/40	5/40

Trong trường hợp mẫu cụ thể $(x_1, x_2, ..., x_n)$ có nhiều giá trị khác nhau, khi đó ta thực hiện việc ghép lớp. Nguyên tắc ghép lớp được tiến hành như sau

- Số lớp chia k được xác định trên cơ sở $k = \min\{l : 2^l > n\}$.
- Độ dài mỗi lớp: $l = \dfrac{\text{giá trị lớn nhất - giá trị nhỏ nhất}}{k}$.
- Trong 2 lớp liền nhau $x_{i-1} \to x_i$, $x_i \to x_{i+1}$ thì x_i thuộc lớp $x_{i-1} \to x_i$.

Ngoài phương pháp ghép lớp đã trình bày ở trên, còn có một số phương pháp ghép lớp khác, với những mẫu cụ thể rời rạc người ta có thể chia thành các lớp có độ dài khác nhau, các lớp được chia rời nhau. Trong phạm vi giáo trình này, chúng ta không đề cập cụ thể các kiểu ghép lớp này.

Ví dụ 2.2.5. *Thống kê về chiều cao của 30 sinh viên với chiều cao nằm trong khoảng từ 1m50 đến 1m75.*

Nhận thấy $2^5 > 30$ *và* $2^4 < 30$ *nên ta chọn* $k = 5$. *Bảng tần số, tần suất như sau:*

Lớp	*Giá trị*	*Tần số*	*Tần suất*
150-155	*152,5*	*4*	*4/30*
155-160	*157,5*	*7*	*7/30*
160-165	*162,5*	*6*	*6/30*
165-170	*167,5*	*10*	*10/30*
170-175	*172,5*	*3*	*3/30*

2.2.3 Đa giác tần số và tổ chức đồ

Đối với số liệu chưa ghép lớp

- Chấm trên mặt phẳng các điểm $(x_i, n_i), i = 1, 2, \ldots, n$.

- Nối các điểm $(x_i, 0)$ với các điểm (x_i, n_i), ta được *biểu đồ tần số hình gậy.*

- Nối liên tiếp điểm (x_i, n_i) với các điểm (x_{i+1}, n_{i+1}) ta được *biểu đồ đa giác tần số.*

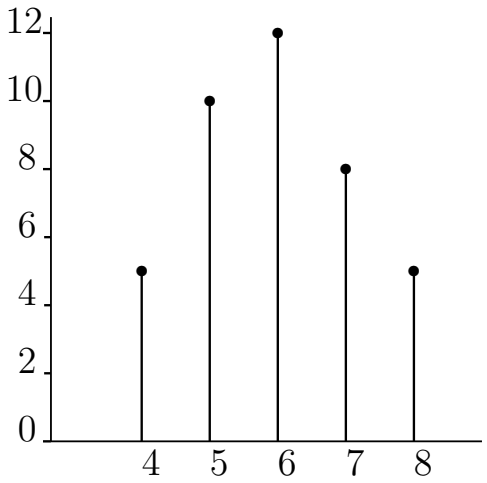
Hoàn toàn tương tự đối với tần suất:

- Chấm trên mặt phẳng các điểm $(x_i, f_i), i = 1, 2, \ldots, n$.

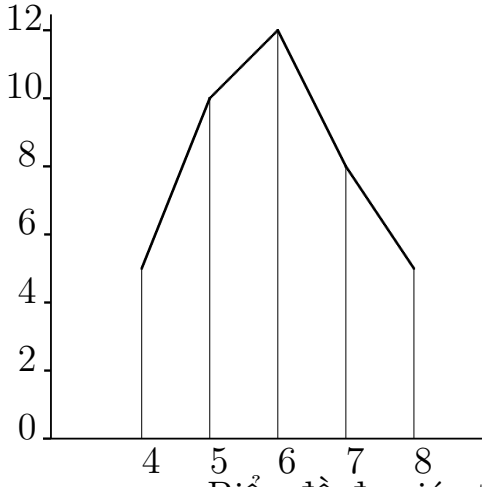
- Nối các điểm $(x_i, 0)$ với các điểm (x_i, f_i), ta được *biểu đồ tần suất hình gậy.*

- Nối liên tiếp điểm (x_i, f_i) với các điểm (x_{i+1}, f_{i+1}) ta được *biểu đồ đa giác tần suất.*

Ví dụ 2.2.6. *Minh họa số liệu của ví dụ thống kê điểm*



Biểu đồ tần số hình gậy



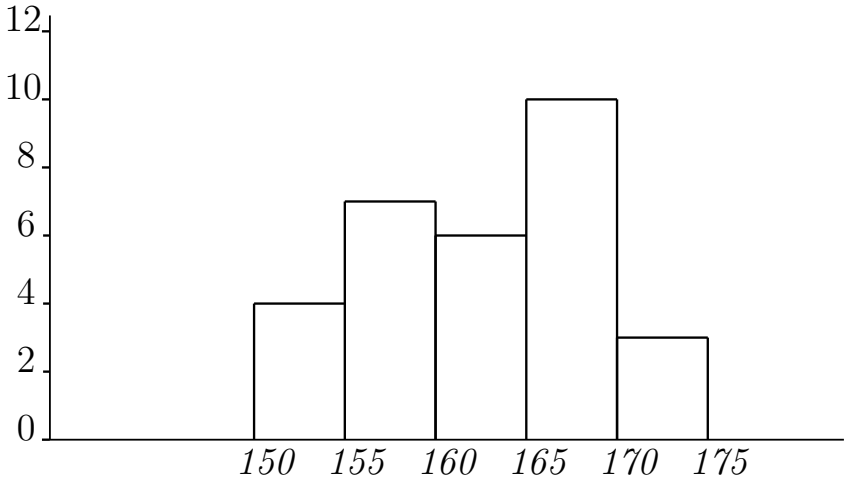
Biểu đồ đa giác tần số

Đối với số liệu đã ghép lớp.

- Trên mỗi lớp ta dựng hình chữ nhật có chiều cao bằng tần số (hay tần suất) tương ứng với lớp đó.

- Tô đậm hoặc kẻ chéo bằng các đường song song các hình chữ nhật này ta thu được *tổ chức đồ tần số* (hay *tổ chức đồ tần suất*).

Ví dụ 2.2.7. *Minh họa số liệu của Ví dụ 2.2.5.*



Biểu đồ đa giác tần số

2.2.4 Phân phối mẫu và các đặc trưng của mẫu

Trong nội dung Chương 1 chúng ta đã được làm quen với việc tính các đặc trưng của đại lượng ngẫu nhiên thông qua phân phối xác suất đã biết trước.

Tuy nhiên, trong thực tế thật khó khăn để xác định được tường minh phân phối xác suất của một đại lượng ngẫu nhiên gốc đám đông. Chính vì vậy, trên cơ sở của các thông tin thu thập được từ các mẫu, người ta đem ra một số công thức giúp chúng ta tính được các đặc trưng của mẫu.

Các giá trị này rất quan trọng và có sự tương ứng với những số đặc trưng của đại lượng ngẫu nhiên đã trình bày ở phần trước.

Hàm phân phối mẫu

X là đại lượng ngẫu nhiên gốc đám đông có hàm phân phối xác suất $F(x)$ chưa biết. Khi ta thực hiện n quan sát, gọi hàm $F_n(x) = \dfrac{m_x}{n}$ với

m_x: là số quan sát có giá trị x_i bé hơn x $(i = \overline{1,n})$ là *hàm phân phối mẫu.*

Tính chất của hàm phân phối mẫu $F_n(x)$:

$+ \ 0 \leqslant F_n(x) \leqslant 1,$

$+ \ F_n(x)$ là hàm đơn điệu tăng,

$+ \ F_n(x)$ là hàm liên tục bên trái.

Khi kích thước mẫu lớn thì phân phối mẫu $F_n(x)$ càng gần với phân phối xác suất của đại lượng ngẫu nhiên X. Khi n đủ lớn, ta có thể dùng $F_n(x)$ thay thế cho $F(x)$ chưa biết hoặc dựa vào $F_n(x)$ ta có thể sơ lược về dáng điệu của $F(x)$ và đưa ra những dự đoán về dạng của $F(x)$ cũng như tính toán các số đặc trưng có liên quan.

Ví dụ 2.2.8. *Bảng tần số từ ví dụ thống kê điểm*

x_i	*4*	*5*	*6*	*7*	*8*
n_i	*5*	*10*	*12*	*8*	*5*

Hàm phân phối mẫu

$$F_n(x) = \begin{cases} 0 & \text{với } x \leqslant 4, \\ \dfrac{5}{40} & \text{với } 4 < x \leqslant 5, \\ \dfrac{15}{40} & \text{với } 5 < x \leqslant 6, \\ \dfrac{27}{40} & \text{với } 6 < x \leqslant 7, \\ \dfrac{35}{40} & \text{với } 7 < x \leqslant 8, \\ 1 & \text{với } x > 8. \end{cases}$$

Trung bình mẫu

Định nghĩa 2.2.9. *Giả sử $n(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên có kích thước n của đám đông X, khi đó $\dfrac{1}{n}\sum_{i=1}^{n} X_i$ được gọi là trung bình mẫu và ký hiệu là $\overline{X}$.*

Trong thực hành tính toán

Đối với một mẫu cụ thể $(x_1, x_2, ..., x_n)$ thì trung bình mẫu thực nghiệm được xác định như sau $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Trường hợp mẫu cụ thể đã được ghép bộ có bảng tần số

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

thì trung bình mẫu thực nghiệm là $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\, x_i$.

Ví dụ 2.2.10. *Bảng tần số từ ví dụ thống kê điểm*

x_i	*4*	*5*	*6*	*7*	*8*
n_i	*5*	*10*	*12*	*8*	*5*

Khi đó $\overline{x} = \frac{1}{40}\sum_{i=1}^{5} n_i\, x_i = \frac{238}{40} = 5{,}95.$

Chú ý 2.2.11. *Công thức tính trung bình mẫu ở trên là dạng tổng quát, tuy nhiên do đặc trưng số nên ta thường dùng khi nghiên cứu về một đặc điểm định lượng nào đó của đám đông. Đối với đặc điểm định tính A ta có khái niệm tỉ lệ mẫu*

$$F = \frac{1}{n}\sum_{i=1}^{n} X_i$$

trong đó X_i *chỉ nhận 2 giá trị là 0 và 1 (bằng 1 nếu quan sát đó có tính chất A, bằng 0 nếu quan sát đó không có tính chất A). Với* $m = \sum_{i=1}^{n} X_i$ *chính là số quan sát có tính chất A, công thức tính tỉ lệ mẫu là* $F = \frac{m}{n}$.

Phương sai mẫu và phương sai hiệu chỉnh mẫu

Định nghĩa 2.2.12. *Giả sử* $(X_1, X_2, ..., X_n)$ *là mẫu ngẫu nhiên có kích thước n của đám đông X, khi đó* $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ *được gọi là* phương sai mẫu *và ký hiệu là* $\hat{S}^2$.

Ngoài ra, chúng ta thường dùng một đặc trưng mẫu khá quan trọng là phương sai hiệu chỉnh mẫu, *ký hiệu là* S^2, *được xác định* $S^2 = \frac{n}{n-1}\hat{S}^2$.

Mệnh đề 2.2.13. *Giả sử* $(X_1, X_2, ..., X_n)$ *là mẫu ngẫu nhiên có kích thước* n *của đám đông* X. *Ta có*

$$\hat{S}^2 = \overline{X^2} - (\overline{X})^2 \quad \textit{trong đó} \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Chứng minh.

$$\begin{aligned}\hat{S}^2 =& \frac{1}{n}\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + (\overline{X})^2\\ =& \overline{X^2} - \frac{2}{n}\overline{X}\sum_{i=1}^{n} X_i + (\overline{X})^2 = \overline{X^2} - (\overline{X})^2.\end{aligned}$$

□

Trong thực hành tính toán

Đối với một mẫu cụ thể đã được ghép bộ có bảng tần số

x_i	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k

thì *phương sai mẫu thực nghiệm* và *phương sai hiệu chỉnh mẫu thực nghiệm* được xác định như sau

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{k} n_i\big(x_i - \overline{x}\big)^2 = \overline{x^2} - \big(\overline{x}\big)^2;$$

$$s^2 = \frac{n}{n-1}\,\hat{s}^2 = \frac{n}{n-1}\big(\overline{x^2} - \big(\overline{x}\big)^2\big).$$

s được gọi là *độ lệch chuẩn mẫu.*

Việc đưa ra các khái niệm trung bình mẫu thực nghiệm (phương sai mẫu thực nghiệm, phương sai hiệu chỉnh mẫu thực nghiệm) chỉ nhằm nhấn mạnh đó là giá trị bằng số cụ thể, được xác định từ thực nghiệm.

Ví dụ 2.2.14. *Bảng tần số từ ví dụ thống kê điểm*

x_i	4	5	6	7	8
n_i	5	10	12	8	5

x_i	n_i	$n_i\, x_i$	$n_i\, x_i^2$
4	5	20	80
5	10	50	250
6	12	72	432
7	8	56	392
8	5	40	320
Tổng	***40***	***238***	***1474***

Ta có $\overline{x} = \dfrac{238}{40} = 5{,}95$; $\overline{x}^2 = \dfrac{1474}{40} = 36{,}85$.
$\hat{s}^2 = 36{,}85 - 5{,}95^2 = 1{,}4475;\ s^2 \approx 1{,}485.$

Chú ý 2.2.15. *Đối với mẫu được ghép lớp, việc tính các số đặc trưng của mẫu cũng theo trình tự tiến hành như trên, trong mỗi lớp ta sử dụng giá trị trung điểm* $x_i' = \dfrac{x_i + x_{i+1}}{2}$ *của lớp.*

2.3 Ước lượng tham số

2.3.1 Mở đầu

Giả sử đại lượng ngẫu nhiên X có luật phân phối phụ thuộc vào một tham số hoặc một véctơ tham số θ chưa biết chẳng $X \sim P(\lambda),\ N(\mu, \sigma)$; $B(n,p), \mathcal{E}(\lambda), ...$ nhưng chưa biết tham số λ, μ, σ, p .v.v. Khi đó để xác định hoàn toàn phân phối xác suất của X ta phải xác định được giá trị tham số. Để có được điều đó người ta phải quan sát X xay dựng mẫu ngâu nhiên $(X_1, X_2, ...X_n)$ từ đó tìm cách ước lượng tham số. Đây chính là *bài toán ước lượng tham số.*

Trong thực tế người ta xét 2 loại ước lượng tham số cơ bản đó là: *ước lượng điêm* và *ước lượng khoảng.*

Ước lượng điểm: Xuất phát tư mẫu ngẫu nhiên $(X_1, X_2, ...X_n)$ người ta xây dựng thống kê $\hat{\theta}(X_1, X_2, ...X_n)$ dùng để ước lượng tham số θ theo các nghĩa khác nhau như *Ước lượng không chêch (không có sai số hệ thống), ước lượng vững , ước lương hiệu quả, ước lượng hợp lý cực đai*

Ước lượng khoảng: Xuất phát tư mẫu ngẫu nhiên $(X_1, X_2, ...X_n)$ người ta xây dựng thống kê $\theta_1 := \hat{\theta}_1(X_1, X_2, ...X_n)$ và $\theta_2 := \hat{\theta}_2(X_1, X_2, ...X_n)$ sao cho

$$\mathbb{P}\{\hat{\theta}_1(X_1, X_2, ...X_n) \leq \theta \leq \hat{\theta}_2(X_1, X_2, ...X_n)\} \geq 1 - \alpha$$

trong đó α được gọi là *mức ý nghĩa* và $\beta = 1 - \alpha$ được gọi là *độ tin cậy*. Khi đó, người ta nói rằng với độ tin cậy α hay mức ý nghĩa α khoảng tin cậy đối với θ là $(\theta_1; \theta_2)$.

2.3.2 Ước lượng điểm

$\diamond$ Ước lượng $\hat{\theta}(X_1, X_2, ..., X_n)$ được gọi là *ước lượng không chệch* của θ, nếu thỏa mãn $\mathbb{E}\hat{\theta} = \theta$.

$\diamond$ Ước lượng $\hat{\theta}(X_1, X_2, ..., X_n)$ được gọi là *ước lượng vững* của θ, nếu với n lớn vô hạn thì $\hat{\theta}$ hội tụ theo xác suất về θ, nghĩa là với mọi $\varepsilon > 0$ tùy ý thì

$$\lim_{n\to\infty} \mathbb{P}[|\hat{\theta} - \theta| < \varepsilon] = 1.$$

$\diamond$ Ước lượng $\hat{\theta}(X_1, X_2, ..., X_n)$ được gọi là *ước lượng hợp lý tối đa* của θ, nếu

$$L(x, \theta) = \prod_{i=1}^{n} p(X_i, \theta)$$

đạt cực đại tại $\hat{\theta}$. $L(x, \theta)$ được gọi là *hàm hợp lý* của X, trong đó $p(x, \theta)$ là hàm mật độ xác suất hoặc là hàm tính xác suất của đại lượng ngẫu nhiên X.

$\diamond$ Ước lượng $\hat{\theta}(X_1, X_2, ..., X_n)$ được gọi là *ước lượng hiệu quả* của θ, nếu như nó là ước lượng không chệch và có phương sai bé nhất trong tất cả các ước lượng không chệch của θ.

Nếu hàm mật độ xác suất của đại lượng ngẫu nhiên X thỏa mãn thêm một số điều kiện nhất định thì ta có bất đẳng thức Cramer-Rao

$$D(\theta^*) \geqslant \frac{1}{n\mathbb{E}\Big(\dfrac{\partial \ln p(X,\,\theta)}{\partial\theta}\Big)^2}\;;\;\; \forall\theta^* \;:\;\; \mathbb{E}(\theta^*) = \theta.$$

do đó, ước lượng không chệch $\hat{\theta}$ là ước lượng hiệu quả của θ khi

$$V(\hat{\theta}) = \frac{1}{n\mathbb{E}\Big(\dfrac{\partial \ln p(X,\,\theta)}{\partial\theta}\Big)^2}.$$

Từ bất đẳng thức Cramer-Rao, ta thấy một điều lý thú đó là: đã là ước lượng thì phải chấp nhận sai số, bất đẳng thức cho ta cận dưới của sai số.

Ví dụ 2.3.1. *Giả sử X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, có kỳ vọng μ và phương sai hữu hạn, khi đó trung bình mẫu $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ chính là ước lượng không chệch, ước lượng vững, ước lượng hiệu quả, ước lượng hợp lý cực đại của μ.*

Ví dụ 2.3.2. *Giả sử X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, có phương sai $\mathbb{D}X = \sigma^2$ cần ước lượng, khi đó phương sai hiệu chỉnh mẫu $S^2 = \frac{1}{n}\sum(X_i - \bar{X})$ chính là ước lượng không chệch của σ^2. Như vậy S^2 là ước lượng không chệch của σ^2. Mặt khác $\hat{S}^2 = \dfrac{n-1}{n}S^2$ nên $\hat{S}^2$ không phải là ước lượng không chệch của σ^2. Tuy nhiên người ta chứng minh được rằng cả S^2 và $\hat{S}^2$ đều là ước lượng vững của σ^2.*

Ví dụ 2.3.3. *Giả sử X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, ta cần quan tâm đến một tính chất A có xác suất $p = \mathbb{P}(A) = \mathbb{E}X$ cần ước lượng, khi đó tỉ lệ mẫu F chính là ước lượng không chệch của xác suất p.*

Khẳng định trên là hiển nhiên vì thực chất tỉ lệ mẫu cũng là trung bình mẫu khi đặc điểm định tính được số hóa dưới dạng

$$X_i = \begin{cases} 1 & \text{khi phần tử đó có tính chất } A\,; \\ 0 & \text{khi phần tử đó không có tính chất } A\,, \end{cases}$$

và $\mathbb{E}F = \mathbb{E}\overline{X} = \mathbb{E}X = p$.

Ngoài ra người ta còn chứng minh được F *cũng chính là ước lượng vững của xác suất* p.

2.3.3 Ước lượng khoảng

Trong nội dung của phần trước, chúng ta đã đề cập đến ước lượng điểm của tham số. Do θ là tham số chưa biết nên ước lượng điểm chỉ cho ta một cách nhìn hết sức tương đối và có phần chưa thỏa đáng. Sau đây chúng ta sẽ suy nghĩ đến một cách tiếp cận khác để tìm ra miền giá trị của θ.

2.3.4 Khái niệm về khoảng tin cậy

Cho X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, có tham số θ cần ước lượng. Căn cứ vào mẫu ngẫu nhiên từ n quan sát độc lập $(X_1,\, X_2,\, ...,\, X_n)$, ta cần đưa ra khoảng $(\theta_1,\, \theta_2)$ chứa được hầu hết các giá trị θ với xác suất lớn, nghĩa là

$$\mathbb{P}(\theta_1 < \theta < \theta_2) = 1 - \alpha\,.$$

Một số khái niệm

$\diamond$ $(\theta_1,\, \theta_2)$: được gọi là *khoảng tin cậy* của ước lượng.

$\diamond$ $\theta_1 - \theta_2 = 2\varepsilon$: được gọi là *độ dài khoảng tin cậy* của ước lượng.

$\diamond$ ε: được gọi là *độ chính xác* của ước lượng.

$\diamond$ $1 - \alpha$: được gọi là *độ tin cậy* của của ước lượng.

$\diamond$ Bài toán đi tìm khoảng tin cậy cho tham số θ với độ tin cậy $1 - \alpha$ được gọi là *bài toán ước lượng khoảng tin cậy.*

2.3.5 Khoảng tin cậy cho giá trị trung bình

Cho X là đại lượng ngẫu nhiên gốc đám đông $\mathcal{C}$, có trung bình $\mathbb{E}X = \mu$ cần ước lượng và phương sai $\mathbb{D}X = \sigma^2$ (đã biết trước hoặc chưa biết), từ mẫu ngẫu nhiên $(X_1,\, X_2,\, ...,\, X_n)$ ta xác định được $\overline{X}$.

a. Ước lượng hai phía

Vấn đề đặt ra ở đây là với độ tin cậy $1-\alpha$ cho trước, tìm khoảng ước lượng $(\overline{X}-\varepsilon,\, \overline{X}+\varepsilon)$ của μ để

$$\mathbb{P}[\overline{X}-\varepsilon<\mu<\overline{X}+\varepsilon]\geq 1-\alpha.$$

Ta chia bài toán thành 3 trường hợp để giải quyết:

Trường hợp 1. Phương sai σ^2 đã biết.

Khi đó $\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}\simeq\mathcal{N}(0,1)$, đặt $t_{\alpha/2}=\varphi^{-1}(1-\frac{\alpha}{2})$, trong đó φ là hàm phân phối chuẩn $\mathcal{N}(0,1)$ và $t_{\alpha/2}$ là mức phân vị $\alpha/2$ cho phân phối chuẩn. Ta có

$$\begin{aligned}\mathbb{P}\big[-t_{\alpha/2}<\frac{\overline{X}-\mu}{\sigma}\sqrt{n}<t_{\alpha/2}\big]&=\varphi(t_{\alpha/2})-\varphi(-t_{\alpha/2})\\&=\varphi(t_{\alpha/2})-(1-\varphi(t_{\alpha/2}))=1-\alpha,\end{aligned}$$

hay

$$\mathbb{P}\Big[\overline{X}-t_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big]=1-\alpha.$$

Quy tắc thực hành

$\diamond$ Xác định mức phân vị $t_{\alpha/2}$

Tính giá trị $1-\frac{\alpha}{2}$, tra bảng hàm phân phối $\mathcal{N}(0,1)$ (xem bảng 4 phần phụ lục), tra từ giữa ra hai biên.

$\diamond$ Xác định khoảng ước lượng $(\overline{x}-\varepsilon,\overline{x}+\varepsilon)$ với độ chính xác của ước lượng

$$\varepsilon=t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Chú ý 2.3.4. *Nếu như kích thước mẫu $n<30$ cần bổ sung thêm điều kiện X tuân theo luật phân phối chuẩn, khi đó* $\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}\sim\mathcal{N}(0,1)$.

Ví dụ 2.3.5. *Tìm khoảng ước lượng cho giá trị trung bình với độ tin cậy 95% từ mẫu của một đám đông tuân theo luật phân phối chuẩn, $\sigma^2=16$. Biết mẫu đó có kích thước 16 và trung bình mẫu là 15.*

Giải. $\sigma^2 = 16$, $n = 15$; $\overline{x} = 15$; α *=0,05 tra bảng hàm phân phối chuẩn ứng với* $1 - \alpha/2$ *=0,975 được* $t_{\alpha/2}$ *=1,96. Độ chính xác của ước lượng*

$$\varepsilon = t_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1{,}96\frac{4}{\sqrt{16}} = 1{,}96.$$

Khoảng ước lượng cho giá trị trung bình:

$$(15 - 1{,}96 < \mu < 15 + 1{,}96) \quad \textit{hay} \quad (13{,}04 < \mu < 16{,}96).$$

Trường hợp 2. Phương sai σ^2 chưa biết và $n \geqslant 30$.

Khi đó $\dfrac{\overline{X} - \mu}{S}\sqrt{n} \simeq \mathcal{N}(0,1)$, việc thiết lập tương tự như ở trường hợp 1, ta được

$$\mathbb{P}\Big[\overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\Big] = 1 - \alpha\,.$$

Như vậy, với một mẫu cụ thể, ta sẽ xác định được độ chính xác của ước lượng $\varepsilon = t_{\alpha/2}\dfrac{s}{\sqrt{n}}$ và khoảng ước lượng

$$(\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}).$$

Ví dụ 2.3.6. *Để ước lượng khối lượng trung bình mỗi bao xi măng của nhà máy, người ta kiểm tra ngẫu nhiên 49 bao thu được khối lượng trung bình là 49,7kg và độ lệch chuẩn mẫu 0,5kg. Với độ tin cậy là* 94%*, hãy ước lượng khoảng khối lượng trung bình của một bao xi măng.*

Giải. α *=0,06,* $t_{\alpha/2}$ *=1,88. Độ chính xác của ước lượng*

$$\varepsilon = t_{\alpha/2}\frac{s}{\sqrt{n}} = 1{,}88\frac{0{,}5}{\sqrt{49}} = 0{,}13.$$

Khoảng ước lượng cho giá trị trung bình: $(49{,}57 < \mu < 49{,}83)$.

Trường hợp 3. Phương sai σ^2 chưa biết và $n < 30$.

Nếu $X \sim \mathcal{N}(\mu, \sigma^2)$ thì $\dfrac{\overline{X} - \mu}{S}\sqrt{n} \sim t(n-1)$. Mức phân vị $\alpha/2$ cho phân phối Student với $n-1$ bậc tự do ký hiệu là $t_{(n-1,\alpha/2)}$ là giá trị thỏa mãn $\mathbb{P}(\dfrac{\overline{X} - \mu}{S}\sqrt{n} > t_{(n-1,\alpha/2)}) = \alpha/2$. Khi đó

$$\begin{aligned}
&\mathbb{P}\big[- t_{(n-1,\alpha/2)} < \frac{\overline{X} - \mu}{S}\sqrt{n} < t_{(n-1,\alpha/2)}\big] \\
&= \mathbb{P}\big[t_{(n-1,1-\alpha/2)} < \frac{\overline{X} - \mu}{S}\sqrt{n} < t_{(n-1,\alpha/2)}\big] \\
&= 1 - \alpha/2 - \alpha/2 = 1 - \alpha.
\end{aligned}$$

Quy tắc thực hành

◇ Xác định mức phân vị $t_{(n-1,\alpha/2)}$.

Tra bảng phân phối Student (xem bảng 5 phần phụ lục), $t_{(n-1,\alpha/2)}$ là giá trị trong bảng ứng với giá trị hàng là $n-1$ và cột là $\alpha/2$.

◇ Xác định khoảng ước lượng $(\overline{x} - \varepsilon, \overline{x} + \varepsilon)$ với độ chính xác của ước lượng

$$\varepsilon = t_{(n-1,\alpha/2)}\frac{s}{\sqrt{n}}.$$

Ví dụ 2.3.7. *Độ chịu lực của mỗi tấm bê tông tuân theo luật phân phối chuẩn. Đo độ chịu lực của 20 tấm bê tông cùng loại người ta thu được trung bình mẫu độ chịu lực* $220kg/cm^2$ *và độ lệch chuẩn mẫu* $32{,}4kg/cm^2$. *Với độ tin cậy* 90%, *tìm khoảng ước lượng trung bình độ chịu lực của mỗi tấm bê tông.*

Giải. *Tra bảng hàm phân phối Student ta được* $t_{(19;0,05)}$ *=1,729. Độ chính xác của ước lượng*

$$\varepsilon = t_{(n-1,\alpha/2)}\frac{s}{\sqrt{n}} \approx 12{,}5.$$

Khoảng ước lượng cho giá trị trung bình: $(187{,}5 < \mu < 212{,}5)$.

Các dạng toán phát sinh

Xuất phát từ các công thức tương ứng với từng trường hợp

$$\varepsilon = t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,; \quad \varepsilon = t_{\alpha/2}\frac{s}{\sqrt{n}}\,; \quad \varepsilon = t_{(n-1,\,\alpha/2)}\frac{s}{\sqrt{n}}\,.$$

◇ Cho $1-\alpha$ và n. Tìm độ chính xác của ước lượng ε.

◇ Cho $1-\alpha$ và ε. Tìm kích thước mẫu n.

◇ Cho ε và n. Tìm độ tin cậy của ước lượng $1-\alpha$.

Một số trong số các vấn đề này sẽ được đề cập ở phần sau.

b. Ước lượng một phía

Vấn đề đặt ra ở đây là với độ tin cậy $1-\alpha$ cho trước, tìm khoảng ước lượng một phía:

◇ Khoảng ước lượng bên trái $(-\infty,\, \overline{X}+\varepsilon)$: $\mathbb{P}[-\infty < \mu < \overline{X}+\varepsilon] = 1-\alpha$.

◇ Khoảng ước lượng bên phải $(\overline{X}-\varepsilon,\, +\infty)$: $\mathbb{P}[\overline{X}-\varepsilon < \mu < +\infty] = 1-\alpha$.

Chú ý 2.3.8. *Khoảng tin cậy bên trái cho ta biết giá trị tối đa, khoảng tin cậy bên phải cho ta biết giá trị tối thiểu của μ với độ tin cậy $1-\alpha$.*

Ta cũng chia thành 3 trường hợp, điểm khác biệt là thay thế $\alpha/2$ bởi α.

Trường hợp 1. *Phương sai σ^2 đã biết.*

Đặt $t_\alpha = \varphi^{-1}(1-\alpha)$, ta có

$$\mathbb{P}\Big[-t_\alpha < \frac{\overline{X}-\mu}{\sigma}\sqrt{n} < +\infty\Big] = 1-\varphi(-t_\alpha) = 1-\alpha,$$
$$\mathbb{P}\Big[-\infty < \frac{\overline{X}-\mu}{\sigma}\sqrt{n} < t_\alpha\Big] = \varphi(t_\alpha) = 1-\alpha,$$

hay $\mathbb{P}\Big[-\infty < \mu < \overline{X}+t_\alpha\frac{\sigma}{\sqrt{n}}\Big] = \mathbb{P}\big[\overline{X}-t_\alpha\frac{\sigma}{\sqrt{n}} < \mu < +\infty\big] = 1-\alpha.$

Như vậy, với một mẫu cụ thể, khoảng ước lượng bên trái và bên phải lần lượt là $(-\infty, \overline{x}+\varepsilon)$, $(\overline{x}-\varepsilon, +\infty)$ trong đó $\varepsilon = t_\alpha\frac{\sigma}{\sqrt{n}}$.

Trường hợp 2. Phương sai σ^2 chưa biết và $n \geqslant 30$.

Lý luận hoàn toàn tương tự, khoảng ước lượng bên trái và bên phải lần lượt là $(-\infty, \overline{x} + \varepsilon)$, $(\overline{x} - \varepsilon, +\infty)$ trong đó $\varepsilon = t_\alpha \dfrac{s}{\sqrt{n}}$.

Trường hợp 3. Phương sai σ^2 chưa biết và $n < 30$.

Khoảng ước lượng bên trái và bên phải lần lượt là $(-\infty, \overline{x} + \varepsilon)$, $(\overline{x} - \varepsilon, +\infty)$ trong đó $\varepsilon = t_{(n-1,\alpha)} \dfrac{s}{\sqrt{n}}$.

Ước lượng khoảng cho giá trị trung bình ứng với 3 trường hợp trên được mô tả qua bảng tổng hợp sau

Loại ước lượng	ε	Độ chính xác của ước lượng: ε		
		TH1	TH2	TH3
Hai phía	$(\overline{x} - \varepsilon, \overline{x} + \varepsilon)$	$t_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$	$t_{\alpha/2} \dfrac{s}{\sqrt{n}}$	$t_{(n-1,\alpha/2)} \dfrac{s}{\sqrt{n}}$
Bên trái	$(-\infty, \overline{x} + \varepsilon)$	$t_\alpha \dfrac{\sigma}{\sqrt{n}}$	$t_\alpha \dfrac{s}{\sqrt{n}}$	$t_{(n-1,\alpha)} \dfrac{s}{\sqrt{n}}$
Bên phải	$(\overline{x} - \varepsilon, +\infty)$			

Ví dụ 2.3.9. *Để đánh giá về mức doanh thu hàng tháng tại các đại lý nhỏ trên một địa bàn, người ta lấy mẫu gồm 36 đại lý. Kết quả thu được như sau: doanh thu trung bình là 155,3 triệu đồng và độ lệch chuẩn mẫu là 16 triệu đồng. Với độ tin cậy* 99%*, ước lượng doanh thu trung bình tối đa và tối thiểu của mỗi đại lý.*

Giải. $1 - \alpha = 0{,}99$; t_α *=2,33. Độ chính xác của ước lượng*

$$\varepsilon = t_\alpha \frac{s}{\sqrt{n}} = 2{,}33 \frac{16}{\sqrt{36}} \approx 6{,}21.$$

Doanh thu tối thiểu: $\overline{x} - \varepsilon$ *=149,09;*
Doanh thu tối đa: $\overline{x} + \varepsilon$ *=161,51.*

2.3.6 Khoảng tin cậy cho tỉ lệ

a. Ước lượng hai phía

Đám đông X có tỉ lệ p cần ước lượng, từ mẫu ngẫu nhiên chúng ta xác định được tỉ lệ F, vấn đề đặt ra ở đây là với độ tin cậy $1-\alpha$ cho trước, tìm khoảng ước lượng $(F-\varepsilon,\, F+\varepsilon)$ của p để

$$\mathbb{P}[F-\varepsilon < p < F+\varepsilon] = 1-\alpha.$$

Khi n đủ lớn $\dfrac{F-p}{\sqrt{F(1-F)}}\sqrt{n} \simeq \mathcal{N}(0,\,1)$, đặt $t_{\alpha/2} = \varphi^{-1}(1-\frac{\alpha}{2})$, ta có

$$\mathbb{P}\big[-t_{\alpha/2} < \frac{F-p}{\sqrt{F(1-F)}}\sqrt{n} < t_{\alpha/2}\big] = \varphi(t_{\alpha/2}) - \varphi(-t_{\alpha/2})$$
$$= \varphi(t_{\alpha/2}) - (1-\varphi(t_{\alpha/2})) = 1-\alpha,$$

hay $\mathbb{P}\Big[F - t_{\alpha/2}\sqrt{\frac{F(1-F)}{n}} < p < F + t_{\alpha/2}\sqrt{\frac{F(1-F)}{n}}\,\Big] = 1-\alpha.$

Quy tắc thực hành: khi $nf \geqslant 10$ và $n(1-f) \geqslant 10$.

$\diamond$ Xác định mức phân vị $t_{\alpha/2}$.

$\diamond$ Xác định khoảng ước lượng $(f-\varepsilon, f+\varepsilon)$ với độ chính xác của ước lượng

$$\varepsilon = t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}.$$

Ví dụ 2.3.10. *Để ước lượng tỉ lệ phế phẩm của một kho hàng. Người ta kiểm tra 100 sản phẩm, phát hiện có 20 sản phẩm là phế phẩm. Với độ tin cậy 95%, hãy ước lượng khoảng tỉ lệ phế phẩm của kho hàng.*

Giải. *$t_{\alpha/2}$ =1,96; f =0,2; n =100. Độ chính xác của ước lượng*

$$\varepsilon = t_{\alpha/2}\sqrt{\frac{f(1-f)}{n}} = 0{,}0784.$$

Khoảng ước lượng cho tỉ lệ phế phẩm: $(0{,}1216 < p < 0{,}2784)$.

b. Ước lượng một phía

Với các bước thiết lập tương tự ta thu được khoảng ước lượng của p bên trái là $p < f+\varepsilon$ và bên phải là $p > f-\varepsilon$, trong đó $\varepsilon = t_{\alpha}\sqrt{\frac{f(1-f)}{n}}$.

Ví dụ 2.3.11. *Cho giả thiết như Ví dụ 5. Hãy ước lượng tỉ lệ phế phẩm tối đa và tối thiểu.*

Giải. t_α =1,64; f =0,2; n =100. *Độ chính xác của ước lượng*

$$\varepsilon = t_\alpha \sqrt{\frac{f(1-f)}{n}} = 0{,}0656.$$

Tỉ lệ sản phẩm tối thiểu: $f - \varepsilon$ =0,1344;
Tỉ lệ sản phẩm tối đa: $f + \varepsilon$ =0,2656.

Ví dụ 2.3.12. *Một lô hàng nhập cảng gồm 5.000 thiết bị điện tử đã qua sử dụng. Cơ quan quản lý kiểm tra ngẫu nhiên 100 thiết bị từ lô hàng thì có 82 thiết bị có thể tiếp tục sử dụng được. Với độ tin cậy* 90%, *lô hàng có tối thiểu bao nhiêu thiết bị có thể tiếp tục sử dụng được?*

Giải. $t_\alpha = 1,28$; $f = 0,82$; $n = 100$; $N = 5.000$. *Độ chính xác xủa ước lượng*

$$\varepsilon = t_\alpha \sqrt{\frac{f(1-f)}{n}} = 0{,}0492.$$

Tỉ lệ sản phẩm tối thiểu: $f - \varepsilon$ =0,7708.
Vậy, số thiết bị tối thiểu có thể tiếp tục sử dụng được: $N(f - \varepsilon) =$ 4864.

Các dạng toán phát sinh
Xuất phát từ công thức

$$\varepsilon = t_{\alpha/2} \sqrt{\frac{f(1-f)}{n}}\,.$$

◇ Cho $1 - \alpha$ và n. Tìm độ chính xác của ước lượng ε.
◇ Cho $1 - \alpha$ và ε. Tìm kích thước mẫu n.
◇ Cho ε và n. Tìm độ tin cậy của ước lượng $1 - \alpha$.

2.3.7 Độ chính xác của ước lượng

Trong các nội dung trước chúng ta đã giải quyết bài toán xây dựng ước lượng khoảng cho trung bình và ước lượng khoảng cho tỉ lệ, nghĩa

là từ mẫu cụ thể, độ tin cậy $1-\alpha$ ta sẽ xác định được khoảng ước lượng cho tham số θ là (θ_1, θ_2) trong đó độ chính xác của ước lượng $\varepsilon = \frac{\theta_2-\theta_1}{2}$.

Trong các trường hợp đã trình bày thì ε phụ thuộc vào kích thước mẫu n. Bây giờ ta đặt ra bài toán ngược: với độ tin cậy $1-\alpha$ đã biết, cho độ chính xác của ước lượng ε, tìm kích thước mẫu n cần thiết để nhận được ước lượng với độ chính xác đã cho. Chúng ta sẽ giải quyết bài toán này đối với trường hợp 1 của bài toán ước lượng khoảng trung bình. Các trường hợp còn lại là hoàn toàn tương tự (giành cho bạn đọc).

Trong trường hợp này, khoảng ước lượng là $(\overline{x}-\varepsilon, \overline{x}+\varepsilon)$ và công thức xác định độ chính xác của ước lượng $\varepsilon = t_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$. Kích thước mẫu điều tra cần thiết nếu độ chính xác của ước lượng ε_0 là

$$n = \Big[\frac{t^2_{\alpha/2}\,\sigma^2}{\varepsilon_0^2}\Big] + 1,$$

trong đó ký hiệu $[x]$ là phần nguyên của x, chẳng hạn $[20, 36] = 20$.

Ví dụ 2.3.13. *Với giả thiết như ở Ví dụ 1:* $\sigma^2 = 16$*;* $1-\alpha$ *=0,95. Muốn có ước lượng có độ chính xác là 1 thì phải điều tra mẫu có kích thước bao nhiêu?*

Giải. *Như vậy* $\varepsilon_0 = 1$*, khi đó*

$$n = \Big[\frac{t^2_{\alpha/2}\,\sigma^2}{\varepsilon_0^2}\Big] + 1 = 62.$$

Ngoài ra, chúng ta còn giải quyết được bài toán ngược dạng tìm độ tin cậy của ước lượng khi biết độ chính xác của ước lượng và kích thước mẫu. Vấn đề này được đề cập trong ví dụ sau đây:

Ví dụ 2.3.14. *Một mẫu thống kê có kích thước* $n = 36$*, có trung bình mẫu là 100 và độ lệch chuẩn mẫu là 5. Tìm độ tin cậy của ước lượng nếu khoảng ước lượng là* (99; 101)*.*

Giải. *Tính mức phân vị:* $t_{\frac{\alpha}{2}} = \dfrac{\varepsilon\sqrt{n}}{s} = 2$. *Độ tin cậy của ước lượng*

$$1 - \alpha = 2\varphi(t_{\alpha/2}) = 0{,}955.$$

2.4 Kiểm định giả thiết

2.4.1 Đặt vấn đề

Trong thực tế cuộc sống chúng ta thường gặp 2 quan điểm trái ngược nhau về một vấn đề nào đó. Chẳng hạn, các nhà sản xuất cho rằng có 95% sản phẩm của công ty đảm bảo tiêu chuẩn, trong khi đó các nhà quản lý thị trường lại cho rằng không phải như vậy thực tế thấp hơn nhiều; trước cuộc bầu cử tổng thống đảng phái A cho rằng có 65% cử tri ủng hộ UWCV của đảng phái họ, trong khi đó đảng đối lập lại cho rằng thực tế thấp hơn nhiều.

Vấn đề đặt ra là, thông qua số liệu thống kê hãy chỉ ra chấp nhận ý kiến nào trong 2 ý kiến đó với một mức ý nghĩa α cho trước.

Tổng quát: Chúng ta thường có bài toán

$$\begin{cases} H: & \text{(Giả thiết)} \quad \text{có tính chất A} \\ K: & \text{(Giả thiết)} \quad \text{không có tính chất A} \end{cases}$$

Từ số liệu thống kê hãy đưa ra kết luận cho bài toán trên.

Trong kiểm định giả thiết thường gặp 2 loại sai lầm:

- Sai lầm loại 1: Bác bỏ H trong khi H đúng
- Sai lầm loại 2: Chấp nhận H trong khi H sai.

Mục đích của các nhà thống kê là làm giảm cả 2 loại sai lầm này. Tuy vậy điều này không thể vì giảm sai lầm này thì khả năng mắc sai lầm loại kia tăng lên.

Trong thực tế thống kê người ta thấy mỗi loại sai lầm sẽ gây ra một tác hại khác nhau. Tuy vậy người ta thấy cần phải giảm sai lầm loại 1 với một xác suất xãy ra bé. Chẳng hạn như trong xã hội hiện

đại người ta cho rằng "*Kết án người vô tội nguy hiểm hơn rất nhiều so với việc tha bổng một người có tội*". .. Do đó, Neyman- Pearson đã cho rằng chúng ta chỉ xét những bài toán thống kê với

$$\mathbb{P}(\text{Sai lầm loại 1}) = \mathbb{P}(\text{Bác bỏ H}|\ \text{H đúng}) \leq \alpha$$

trong đó α là một số bé và gọi là mức ý nghĩa. Thông thường $\alpha \leq 10\%$.

2.5 Kiểm định giả thiết về giá trị trung bình và về tỉ lệ

2.5.1 Kiểm định giả thiết về giá trị trung bình

Đây là một dạng bài toán kiểm định số đặc trưng $\mathbb{E}X = \mu$ của biến ngẫu nhiên gốc đám đông X (so sánh giá trị kỳ vọng của đại lượng ngẫu nhiên X với giá trị μ_0 cho trước). Có 2 dạng bài toán kiểm định giả thiết về giá trị trung bình.

a. Kiểm định hai phía

Vấn đề đặt ra ở đây là với mức ý nghĩa α và một giá trị μ_0 cho trước, đánh giá về cặp giả thiết thống kê

$$H:\ \mu = \mu_0\ ; \quad K:\ \mu \neq \mu_0.$$

Trường hợp 1. Phương sai σ^2 đã biết.

Khoảng ước lượng của μ với độ tin cậy $1-\alpha$

$$(\overline{x} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

Chấp nhận giả thiết H khi $\mu_0 \in [\overline{x} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ;\ \overline{x} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$ hay

$$\overline{x} - t_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu_0 \leqslant \overline{x} + t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

tương đương với $\left|\dfrac{\overline{x}-\mu_0}{\sigma}\sqrt{n}\right| \leqslant t_{\alpha/2}$.

Quy tắc thực hành

◇ Từ mẫu cụ thể xác định giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}.$$

◇ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Chú ý

- Nếu như kích thước mẫu $n < 30$ thì ta cần bổ sung thêm điều kiện X tuân theo luật phân phối chuẩn.

- Như vậy miền bác bỏ $W_\alpha = (-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, +\infty)$, điều này là hợp lý. Giả sử $H:\ \mu = \mu_0$ đúng thì $T = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \simeq \mathcal{N}(0,1)$, khi đó

$$\begin{aligned}\mathbb{P}[T \in W_\alpha | H \text{ đúng}] &= \mathbb{P}\Big[\Big|\frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}\Big| > t_{\alpha/2}\Big] \\ &= 1 - \big(\varphi(t_{\alpha/2}) - \varphi(-t_{\alpha/2})\big) = \alpha.\end{aligned}$$

Nghĩa là xác suất phạm sai lầm loại 1 được ấn định bởi một giá trị tương đối nhỏ α nào đó, việc chứng minh xác suất phạm sai lầm loại 2 cực tiểu bạn đọc tham khảo tài liệu [**??**].

Ví dụ 2.5.1. *Một máy tiện tự động cho ra những trục máy có đường kính là 120mm và độ lệch chuẩn cho phép là 3mm. Kiểm tra ngẫu nhiên 50 trục máy, kết quả thu được đường kính trung bình là 119,2mm. Với mức ý nghĩa là* 10%*, máy tiện trên có hoạt động bình thường không?*

Giải. *Máy tiện được gọi là hoạt động bình thường khi nó sản xuất những trục máy với sai số không vượt quá mức cho phép. Cặp giả thiết thống kê*

$$H:\ \mu = \mu_0 = 120\ ; \quad K:\ \mu \neq \mu_0.$$

$\mu_0 = 120;\ \sigma = 3;\ \alpha = 0{,}1;\ t_{\alpha/2} = 1{,}64;\ n = 50;\ \overline{x} = 119{,}2$. *Giá trị kiểm định thực nghiệm*

$$t_{tn} = \frac{119{,}2 - 120}{3}\sqrt{50} \approx -1{,}89,$$

Vì $|t_{tn}| > t_{\alpha/2}$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận khẳng định cho rằng máy tiện trên hoạt động không bình thường.

Trường hợp 2. Phương sai σ^2 chưa biết và $n \geqslant 30$.

Tương tự như ở trường hợp 1, đặt $t_{tn} = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$, khi đó

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Trường hợp 3. Phương sai σ^2 chưa biết và $n < 30$.

Giả sử X tuân theo luật phân phối chuẩn $\mathcal{N}(0,1)$. Đặt $t_{tn} = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$, khi đó

$|t_{tn}| \leqslant t_{(n-1,\alpha/2)}$: chấp nhận H.

$|t_{tn}| > t_{(n-1,\alpha/2)}$: bác bỏ H, chấp nhận K.

Ví dụ 2.5.2. *Thể tích sơn chứa trong mỗi thùng sơn nước nhãn hiệu A là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn với trung bình 18 lít. Kiểm tra ngẫu nhiên 25 thùng thu được kết quả: thể tích trung bình là 17,92 lít và độ lệch chuẩn mẫu là 0,24 lít. Với mức ý nghĩa 5%, thể tích sơn trong các thùng sơn có đúng tiêu chuẩn không?*

Giải. *Cặp giả thiết thống kê*

$$H:\ \mu = \mu_0 = 18\ ; \quad K:\ \mu \neq \mu_0.$$

$\mu_0 = 18$; $s = 0{,}24$; $\alpha = 0,05$; $t_{(n-1,\alpha/2)} = 2{,}11$; $n = 25$; $\overline{x} = 17{,}92$.

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{17{,}92 - 18}{0{,}24}\sqrt{25} \approx -1{,}67,$$

Vì $|t_{tn}| \leqslant t_{\alpha/2}$ nên ta chấp nhận H. Nghĩa là chấp nhận khẳng định cho rằng thể tích sơn trong các thùng sơn đúng tiêu chuẩn.

b. Kiểm định một phía

Trong thực tế xuất hiện một số dạng toán về kiểm định như:

- Sau chiến dịch quảng cáo, doanh số bán ra một loại hàng có tăng lên hay không? (kiểm định lớn hơn)

- Kiểm tra xem khối lượng đóng gói các bao gạo của một kho có nhỏ hơn giá trị in trên bao bì hay không? (kiểm định nhỏ hơn)

Các dạng bài toán này được gọi là *bài toán kiểm định một phía.*

◇ Kiểm định lớn hơn: $H:\ \mu = \mu_0\ ;\quad K:\ \mu > \mu_0$.

◇ Kiểm định nhỏ hơn: $H:\ \mu = \mu_0\ ;\quad K:\ \mu < \mu_0$.

Giải quyết bài toán kiểm định một phía được phân chia các trường hợp giống như trong bài toán kiểm định hai phía. Tiêu chuẩn kiểm định ứng với 3 trường hợp của bài toán kiểm định giá trị trung bình được mô tả qua bảng tổng hợp sau đây

Trường hợp	t_{tn}	Điều kiện chấp nhận $H:\ \mu = \mu_0$		
		$K:\ \mu = \mu_0$	$K:\ \mu > \mu_0$	$K:\ \mu < \mu_0$
TH1	$\dfrac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$	$\lvert t_{tn}\rvert \leqslant t_{\alpha/2}$	$t_{tn} \leqslant t_{\alpha}$	$t_{tn} \geqslant -t_{\alpha}$
TH2	$\dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$	$\lvert t_{tn}\rvert \leqslant t_{\alpha/2}$	$t_{tn} \leqslant t_{\alpha}$	$t_{tn} \geqslant -t_{\alpha}$
TH3	$\dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$	$\lvert t_{tn}\rvert \leqslant t_{(n-1,\alpha/2)}$	$t_{tn} \leqslant t_{(n-1,\alpha)}$	$t_{tn} \geqslant -t_{(n-1,\alpha)}$

Ví dụ 2.5.3. *Một nhà máy cung cấp nước sạch cho rằng khối lượng trung bình của một loại chất độc hại trong một lít nước của nhà máy là* $14mg$. *Người ta nghi ngờ số liệu này thấp hơn thực tế. Kiểm tra ngẫu nhiên với 64 mẫu nước thu được kết quả:* $\bar{x} = 14{,}2$ *và* $s = 0{,}24$. *Hãy cho kết luận về nghi ngờ nói trên với mức ý nghĩa 8%.*

Giải. *Cặp giả thiết thống kê:* $H:\ \mu = \mu_0 = 120\ ;\quad K:\ \mu > \mu_0$.

$\mu_0 = 14;\ s = 0{,}24;\ \alpha = 0{,}08;\ t_\alpha = 1{,}4;\ n = 64;\ \overline{x} = 14{,}2$.

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{14{,}2 - 14}{0{,}24}\sqrt{64} \approx 6{,}67,$$

Vì $t_{tn} > t_\alpha$ *nên ta bác bỏ* H, *chấp nhận* K. *Nghĩa là chấp nhận nghi ngờ trên.*

2.5.2 Kiểm định giả thiết về tỉ lệ

Đây là dạng bài so sánh giá trị tỉ lệ p của đám đông X với giá trị p_0 cho trước. Có hai dạng bài toán kiểm định giả thiết về tỉ lệ.

a. Kiểm định hai phía

Vấn đề đặt ra ở đây là với mức ý nghĩa α và một giá trị p_0 cho trước, đánh giá về cặp giả thiết thống kê

$$H:\ p = p_0\ ; \quad K:\ p \neq p_0.$$

Với n đủ lớn và $H:\ p = p_0$ đúng thì $T = \dfrac{F - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \simeq \mathcal{N}(0,1)$, khi đó

$$\mathbb{P}[T \in W_\alpha | H \text{ đúng}] = \mathbb{P}\Big[\Big|\frac{F - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}\Big| > t_{\alpha/2}\Big] = \alpha,$$

trong đó miền bác bỏ $W_\alpha = (-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, +\infty)$.

Quy tắc thực hành: Khi $np_0 \geqslant 5;\ n(1-p_0) \geqslant 5$.

$\diamond$ Xác định giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}.$$

$\diamond$ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Ví dụ 2.5.4. *Một hãng sản xuất đĩa cứng công bố rằng: có 10% đĩa cứng của hãng phải bảo hành trong thời gian 2 năm đầu sử dụng. Người ta điều tra ngẫu nhiên 200 khách hàng đã sử dụng đĩa cứng của hãng thì có 29 đĩa cứng phải bảo hành trong thời gian 2 năm đầu sử dụng. Với mức ý nghĩa 5%, tỉ lệ trong công bố trên có đúng với thực tế không?*

Giải. *Cặp giả thiết thống kê:* $H:\ p = p_0 = 0{,}1\ ; \quad K:\ p \neq p_0$.

$n = 200;\ f = 0{,}145;\ p_0 = 0{,}1;\ \alpha = 0,05;\ t_{\alpha/2} = 1{,}96.$

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{0,145 - 0,1}{\sqrt{0,1 \times 0,9}}\sqrt{200} \approx 2{,}12.$$

Vì $|t_{tn}| > t_{\alpha/2}$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận khẳng định cho rằng tỉ lệ trong công bố trên không đúng với thực tế.

b. Kiểm định một phía

Tương tự như bài toán kiểm định về giá trị trung bình, bài toán kiểm định tỉ lệ cũng có hai dạng kiểm định một phía như sau:

◇ Kiểm định lớn hơn: $H:\ p = p_0\ ;\quad K:\ p > p_0$.

◇ Kiểm định nhỏ hơn: $H:\ p = p_0\ ;\quad K:\ p < p_0$.

Bảng dưới đây sẽ trình bày các tiêu chuẩn kiểm định của bài toán kiểm định tỉ lệ

t_{tn}	Điều kiện chấp nhận $H:\ p = p_0$		
	$K:\ p = p_0$	$K:\ p > p_0$	$K:\ p < p_0$
$\dfrac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$\lvert t_{tn}\rvert \leqslant t_{\alpha/2}$	$t_{tn} \leqslant t_\alpha$	$t_{tn} \geqslant -t_\alpha$

Ví dụ 2.5.5. *Một trung tâm đào tạo nghề báo cáo rằng tỷ lệ người học tại trung tâm kiếm được việc làm ngay sau khi tốt nghiệp là 70%. Một mẫu ngẫu nhiên gồm 200 người đã tốt nghiệp ở trung tâm cho thấy có 130 người kiếm được việc làm ngay sau khi tốt nghiệp. Với mức ý nghĩa 5%, kiểm định xem phải chăng tỉ lệ trong báo cáo của trung tâm là cao hơn thực tế.*

Giải. *Cặp giả thiết thống kê:* $H:\ p = p_0 = 0{,}7\ ;\quad K:\ p < p_0$.

$n = 200;\ f = 0{,}65;\ p_0 = 0{,}7;\ \alpha = 0,05;\ t_\alpha = 1{,}64$.

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{0,65 - 0,7}{\sqrt{0,7 \times 0,3}}\sqrt{200} \approx -1{,}54.$$

Vì $t_{tn} > -t_\alpha$ nên ta chấp nhận H. Nghĩa là chấp nhận ý kiến cho rằng tỉ lệ trong báo cáo của trung tâm là đúng thực tế.

2.5.3 Boài toán so sánh

Giả sử chúng ta có hai đám đông $\mathcal{C}_1$ và $\mathcal{C}_2$ có chung một đặc điểm cần nghiên cứu nào đó; hai đại lượng ngẫu nhiên gốc đám đông tương ứng lần lượt là X_1 và X_2. Trong mục này chúng ta đề cập đến dạng bài toán so sánh hai giá trị đặc trưng của hai đại lượng ngẫu nhiên này.

So sánh hai giá trị trung bình

Hai đám đông $\mathcal{C}_1$ và $\mathcal{C}_2$ có hai giá trị trung bình là $\mathbb{E}X_1 = \mu_1$ và $\mathbb{E}X_2 = \mu_2$ cần so sánh. Vấn đề đặt ra ở đây là với mức ý nghĩa α cho trước, đánh giá về cặp giả thiết thống kê

$$H:\ \mu_1 = \mu_2\ ; \quad K:\ \mu_1 \neq \mu_2.$$

Giả sử $\mathbb{D}X_1 = \sigma_1^2,\ \mathbb{D}X_2 = \sigma_2^2$. Từ hai mẫu cụ thể $(x_1, x_2, ..., x_{n_1})$ của đám đông $\mathcal{C}_1$ và $(y_1, y_2, ..., y_{n_2})$ của đám đông $\mathcal{C}_2$ chúng ta xác định được trung bình mẫu và phương sai hiệu chỉnh mẫu lần lượt là $\overline{x}_1, \overline{x}_2, s_1^2, s_2^2$.

Quy tắc thực hành

Trường hợp 1. σ_1^2, σ_2^2 đã biết.

$\diamond$ Xác định giá trị kiểm định từ thực nghiệm

$$t_{tn} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}.$$

$\diamond$ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Chú ý. Nếu như kích thước mẫu $n_1 < 30$ hoặc $n_2 < 30$ thì ta cần bổ sung thêm điều kiện X_1, X_2 tuân theo luật phân phối chuẩn.

Ví dụ 2.5.6. *Người ta muốn so sánh tuổi thọ của hai loại thiết bị điện tử (trong điều kiện hoạt động liên tục) được sản xuất bởi hai công nghệ khác nhau. Biết rằng độ lệch chuẩn tuổi thọ của thiết bị được sản xuất*

từ công nghệ thứ nhất và công nghệ thứ hai tương ứng là 120 giờ và 125 giờ. Thử nghiệm 50 thiết bị cho mỗi công nghệ trên thu được tuổi thọ trung bình của chúng tương ứng là 264 giờ và 245 giờ. Với mức ý nghĩa 5%, tuổi thọ của hai loại thiết bị điện tử được sản xuất từ hai công nghệ trên có khác nhau không?

Giải. *Cặp giả thiết thống kê:* $H:\ \mu_1 = \mu_2\ ;\quad K:\ \mu_1 \neq \mu_2$.

$\sigma_1 = 120;\ \sigma_2 = 125;\ n_1 = n_2 = 50;\ \overline{x}_1 = 264;\ \overline{x}_2 = 245$*;*

$\alpha = 0{,}05;\ t_{\alpha/2} =$*1,96.*

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{264 - 245}{\sqrt{\dfrac{120^2}{50} + \dfrac{125^2}{50}}} \approx 0{,}78.$$

Vì $|t_{tn}| \leqslant t_{\alpha/2}$ *nên ta chấp nhận* H*. Nghĩa là chấp nhận khẳng định rằng tuổi thọ của hai loại thiết bị điện tử được sản xuất từ hai công nghệ trên là giống nhau.*

Trường hợp 2. σ_1^2, σ_2^2 chưa biết và $n_1 \geqslant 30, n_2 \geqslant 30$.

$\diamond$ Xác định giá trị kiểm định từ thực nghiệm

$$t_{tn} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}.$$

$\diamond$ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Trường hợp 3. X_1, X_2 có phân phối chuẩn, $\sigma_1^2 = \sigma_2^2$ chưa biết và $n_1 < 30, n_2 < 30$.

$\diamond$ Xác định giá trị kiểm định từ thực nghiệm

$$t_{tn} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s^2\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}},\quad \text{trong đó}\quad s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}\,.$$

◇ So sánh giá trị của t_{tn} với mức phân vị $t_{(n_1+n_2-2,\alpha/2)}$, nếu

$|t_{tn}| \leqslant t_{(n_1+n_2-2,\alpha/2)}$: chấp nhận H.

$|t_{tn}| > t_{(n_1+n_2-2,\alpha/2)}$: bác bỏ H, chấp nhận K.

Ví dụ 2.5.7. *Hai máy tự động dùng cắt những thanh kim loại với cùng một yêu cầu. Từ máy thứ nhất lấy ra 12 sản phẩm thu được chiều dài trung bình là 55cm và độ lệch chuẩn mẫu là 0,3cm, từ máy thứ 2 lấy ra 18 sản phẩm có các kết quả tương ứng là : 55,2cm và 0,2cm. Với mức ý nghĩa là 0,1, đánh giá về nhận định: hai máy đó sản xuất ra các thiết bị cùng kích cỡ. Giả sử rằng kích cỡ sản phẩm từ 2 máy có phân phối chuẩn và có cùng phương sai.*

Giải. *Cặp giả thiết thống kê:* $H: \mu_1 = \mu_2$; $K: \mu_1 \neq \mu_2$.

$s_1 = 0,3cm$; $s_2 = 0,2$; $n_1 = 12$; $n_2 = 18$; $\overline{x_1} = 55cm$; $\overline{x_2} = 55,2$;

$\alpha = 0,1$; $t_{(28;0,05)} = 1,701$.

Giá trị kiểm định thực nghiệm

$$s^2 = \frac{11 \times 0,3^2 + 17 \times 0,2^2}{28} \approx 0,06;$$

$$t_{tn} = \frac{55 - 55,2}{\sqrt{0,06\left(\frac{1}{12} + \frac{1}{18}\right)}} \approx -2,2.$$

Vì $|t_{tn}| > t_{(28;0,05)}$ *nên ta bác bỏ* H*, chấp nhận* K*. Nghĩa là chấp nhận nhận định cho rằng hai máy đó sản xuất ra các thiết bị không cùng kích cỡ.*

Đối với bài toán so sánh 2 giá trị trung bình, có hai dạng bài toán kiểm định một phía như sau:

◇ Kiểm định lớn hơn: $H: \mu_1 = \mu_2$; $K: \mu_1 > \mu_2$.

◇ Kiểm định nhỏ hơn: $H: \mu_1 = \mu_2$; $K: \mu_1 < \mu_2$.

Giải quyết bài toán kiểm định một phía được phân chia các trường hợp giống như trong bài toán kiểm định hai phía. Tiêu chuẩn kiểm định ứng với 3 trường hợp được mô tả qua bảng tổng hợp sau:

Trường hợp	t_{tn}	Điều kiện chấp nhận $H:\ \mu_1 = \mu_2$	
		$K:\ \mu_1 > \mu_2$	$K:\ \mu_1 < \mu_2$
TH1	$\dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$t_{tn} \leqslant t_\alpha$	$t_{tn} \geqslant -t_\alpha$
TH2	$\dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	$t_{tn} \leqslant t_\alpha$	$t_{tn} \geqslant -t_\alpha$
TH3	$\dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{s^2\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}}$	$t_{tn} \leqslant t_{(n_1+n_2-2,\alpha)}$	$t_{tn} \geqslant -t_{(n_1+n_2-2,\alpha)}$

Ví dụ 2.5.8. *Với giả thiết như ở Ví dụ 2: $s_1 = 0{,}3cm$; $s_2 = 0{,}2$; $n_1 = 12$; $n_2 = 18$; $\overline{x_1} = 55cm$; $\overline{x_2} = 55{,}2$. Đánh giá nhận định: máy thứ hai sản xuất ra thiết bị có kích cỡ lớn hơn máy thứ nhất.*

Giải. *Cặp giả thiết thống kê: $H:\ \mu_1 = \mu_2$; $K:\ \mu_1 < \mu_2$.*

$\alpha = 0{,}1$; $t_{(28;0,1)} = 1{,}313$. Giá trị kiểm định thực nghiệm

$$s^2 \approx 0,06;\ t_{tn} \approx -2,2.$$

Vì $t_{tn} < -t_{(28;0,1)}$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận nhận định cho rằng máy thứ hai sản xuất ra thiết bị có kích cỡ lớn hơn máy thứ nhất.

So sánh hai tỉ lệ

Hai đám đông $\mathcal{C}_1$ và $\mathcal{C}_2$ có hai tỉ lệ p_1 và p_2 cần so sánh. Vấn đề đặt ra ở đây là với mức ý nghĩa α cho trước, đánh giá về cặp giả thiết thống kê

$$H:\ p_1 = p_2\ ; \quad K:\ p_1 \neq p_2.$$

Từ mẫu cụ thể kích thước n_1 của đám đông $\mathcal{C}_1$ ta xác định được k_1 phần tử có đặc điểm cần nghiên cứu, do đó tỉ lệ mẫu là $f_1 = k_1/n_1$; tương tự đối với mẫu kích thước n_2 của đám đông $\mathcal{C}_2$ ta xác định được k_2 và $f_2 = k_2/n_2$.

Quy tắc thực hành: Khi n_1, n_2 đủ lớn.

◇ Xác định giá trị kiểm định từ thực nghiệm

$$t_{tn} = \frac{f_1 - f_2}{\sqrt{f(1-f)\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}}, \quad \text{trong đó} \quad f = \frac{k_1 + k_2}{n_1 + n_2}.$$

◇ So sánh giá trị của t_{tn} với mức phân vị $t_{\alpha/2}$, nếu

$|t_{tn}| \leqslant t_{\alpha/2}$: chấp nhận H.

$|t_{tn}| > t_{\alpha/2}$: bác bỏ H, chấp nhận K.

Chú ý 2.5.9. *Khi kích thước mẫu điều tra càng lớn thì kết quả kiểm định càng chính xác, ở mức độ tương đối khái niệm n_1, n_2 đủ lớn ở đây được hiểu là thỏa mãn hai điều kiện:* $(n_1+n_2)f \geqslant 10$, $(n_1+n_2)(1-f) \geqslant 10$.

Ví dụ 2.5.10. *Người ta kiểm tra ngẫu nhiên 400 sản phẩm từ dây chuyền thứ nhất thì có 24 phế phẩm, kiểm tra 800 sản phẩm từ dây chuyền thứ hai thấy có 42 phế phẩm. Với mức ý nghĩa $\alpha = 0{,}05$, tỉ lệ phế phẩm của 2 dây chuyền trên có như nhau hay không ?*

Giải. *Cặp giả thiết thống kê:* $H: p_1 = p_2$; $K: p_1 \neq p_2$.

$t_{\alpha/2} = 1{,}96$; $f_1 = 0{,}06$; $f_2 = 0{,}0525$; $f = 0{,}055$.

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{0{,}06 - 0{,}0525}{\sqrt{0{,}055 \times 0{,}945\big(1/400 + 1/800\big)}} \approx 0{,}537.$$

Vì $|t_{tn}| < t_{\alpha/2}$ nên ta chấp nhận H. Nghĩa là chấp nhận khẳng định cho rằng tỉ lệ phế phẩm của 2 dây chuyền trên là như nhau.

Tương tự như bài toán kiểm định về giá trị trung bình, bài toán kiểm định tỉ lệ cũng có hai dạng kiểm định một phía như sau:

◇ Kiểm định lớn hơn: $H: p_1 = p_2$; $K: p_1 > p_2$.

◇ Kiểm định nhỏ hơn: $H: p_1 = p_2$; $K: p_1 < p_2$.

Bảng dưới đây sẽ trình bày các tiêu chuẩn kiểm định của bài toán kiểm định tỉ lệ:

t_{tn}	Điều kiện chấp nhận $H:\ p_1 = p_2$		
	$K:\ p_1 \neq p_2$	$K:\ p_1 > p_2$	$K:\ p_1 < p_2$
$\dfrac{f_1 - f_2}{\sqrt{f(1-f)\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}}$	$\lvert t_{tn}\rvert \leqslant t_{\alpha/2}$	$t_{tn} \leqslant t_{\alpha}$	$t_{tn} \geqslant -t_{\alpha}$

Ví dụ 2.5.11. *Dùng thuốc A cho 200 bệnh nhân thì có 160 người khỏi bệnh. Dùng thuốc B cho 300 bệnh nhân thì có 210 người khỏi bệnh. Với mức ý nghĩa $\alpha = 0{,}04$, hiệu quả của thuốc A có cao hơn thuốc B hay không?*

Giải. *Cặp giả thiết thống kê:* $H:\ p_1 = p_2;\ \ K:\ p_1 > p_2$.

$t_{\alpha} = 1{,}75;\ f_1 = 0{,}8;\ f_2 = 0{,}7;\ f = 0{,}74.$

Giá trị kiểm định thực nghiệm

$$t_{tn} = \frac{(0,8 - 0,7)}{\sqrt{0,74 \times 0,26\big(1/200 + 1/300\big)}} \approx 2,497.$$

Vì $t_{tn} > t_{\alpha}$ nên ta bác bỏ H, chấp nhận K. Nghĩa là chấp nhận khẳng định cho rằng hiệu quả của thuốc A cao hơn thuốc B.

2.6 Hồi quy và tương quan

2.6.1 Mở đầu

Trên cùng một đám đông $\mathcal{C}$ có hai đặc điểm định lượng cần nghiên cứu, hai đại lượng ngẫu nhiên gốc đám đông tương ứng lần lượt là X và Y. Bài toán đặt ra ở đây là tìm hiểu mức độ phụ thuộc giữa hai đại lượng ngẫu nhiên và tìm biểu thức biểu diễn sự liên hệ giữa chúng.

Đây là một vấn đề hoàn toàn thực tế, sự phụ thuộc của hai đại lượng ngẫu nhiên X và Y có thể phân thành ba loại:

◇ Sự phụ thuộc hàm số: tồn tại hàm φ để $Y = \varphi(X)$.

◇ Sự phụ thuộc thống kê: khi X thay đổi thì phân phối xác suất của Y cũng thay đổi.

$\diamond$ Sự phụ thuộc tương quan: X thay đổi thì kỳ vọng có điều kiện $\mathbb{E}(Y|X)$ cũng thay đổi, nghĩa là $\mathbb{E}(Y|X) = \varphi(X) \neq$ hằng số.

Nếu $\varphi(X) = AX + B$ thì ta nói X và Y có *tương quan tuyến tính*, trong trường hợp ngược lại thì ta nói X và Y có *tương quan phi tuyến.*

Phụ thuộc tương quan là trường hợp riêng của phụ thuộc thống kê, nghĩa là nếu phụ thuộc tương quan thì có sự phụ thuộc về phân phối xác suất. Khi phân tích độ phụ thuộc tương quan giữa hai đại lượng ngẫu nhiên X và Y thì ta không cần xét đến trường hợp nó độc lập với nhau.

2.6.2 Hệ số tương quan mẫu

Chúng ta đã được làm quen với khái niệm hệ số tương quan giữa hai đại lượng ngẫu nhiên X và Y

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\mathbb{D}X\,\mathbb{D}Y}} = \frac{\mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y}{\sqrt{\mathbb{D}X\,\mathbb{D}Y}}\,.$$

Đó là số đo mức độ phụ thuộc tuyến tính giữa hai đại lượng ngẫu nhiên X và Y, nhưng nếu chưa biết được phân phối xác suất thì hệ số tương quan lý thuyết $\rho(X, Y)$ chưa xác định được. Do đó ta tìm cách ước lượng $\rho(X, Y)$ bởi một giá trị thu được từ mẫu quan sát, giá trị đó được gọi là *hệ số tương quan mẫu.*

Giả sử ta có n cặp quan sát (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) của (X, Y), *hệ số tương quan mẫu* được tính theo công thức

$$r = \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2 \sum\limits_{i=1}^{n}(y_i - \overline{y})^2}}\,.$$

Do vậy

$$r = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{\sqrt{n\sum\limits_{i=1}^{n} x_i^2 - \Big(\sum\limits_{i=1}^{n} x_i\Big)^2}\ \sqrt{n\sum\limits_{i=1}^{n} y_i^2 - \Big(\sum\limits_{i=1}^{n} y_i\Big)^2}} = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\hat{s}_X\,\hat{s}_Y}\,,$$

trong đó $\overline{xy} = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i$.

Tương tự như hệ số tương quan, hệ số tương quan mẫu cũng có tính chất $|r| \leqslant 1$. Biểu diễn các cặp (x_i, y_i) của mẫu lên một mặt phẳng tọa độ tạo thành đám mây điểm. Hình ảnh của đám mây điểm thể hiện mối quan hệ giữa X và Y. Nếu đám mây điểm có xu hướng tập trung quanh một đường thẳng nào đó (có hệ số góc khác 0) thì $|r|$ càng gần 1 và ta có thể kết luận X, Y có quan hệ gần với quan hệ tuyến tính (tương quan tuyến tính), còn nếu nó phân tán thành hình tròn hay hình vuông thì $|r|$ gần bằng 0.

Ví dụ 2.6.1. *Bảng số liệu sau đây là kết quả thu thập từ một công ty về doanh thu (X) và số tiền dành cho quảng cáo (Y) của một số tháng như sau:*

X (tỉ đồng)	*5*	*7*	*8*	*11*	*9*
Y (triệu đồng)	*45*	*60*	*75*	*90*	*80*

Hãy xác định hệ số tương quan mẫu.

Giải. *Bảng tính*

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
5	*45*	*225*	*25*	*2025*
7	*60*	*420*	*49*	*3600*
8	*75*	*600*	*64*	*5625*
11	*90*	*990*	*121*	*8100*
9	*80*	*720*	*81*	*6400*
40	***350***	***2955***	***340***	***25750***

Hệ số tương quan mẫu

$$r = \frac{5 \cdot 2955 - 40 \cdot 350}{\sqrt{5 \cdot 340 - 40^2}\ \sqrt{5 \cdot 25750 - 350^2}} \approx 0{,}98.$$

Chú ý 2.6.2. *Trường hợp số liệu thu thập có kích thước lớn, dạng bảng có tần số chúng ta cũng lập bảng tính trung gian như trên sau đó sử dụng công thức:* $r = \dfrac{\overline{xy} - \overline{x}\ \overline{y}}{\hat{s}_X\, \hat{s}_Y}$

2.6.3 Phương trình hồi quy thực nghiệm

Phương trình hồi quy

Mệnh đề 2.6.3. *Trong tất cả các hàm $h(X)$ dùng để ước lượng Y thì $\varphi(X) = \mathbb{E}(Y|X)$ là hàm có sai số bình phương trung bình nhỏ nhất. Nghĩa là*

$$\mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)^2 \leqslant \mathbb{E}\big(Y - h(X)\big)^2.$$

Chứng minh.

$$\begin{aligned}\mathbb{E}\big(Y - h(X)\big) &= \mathbb{E}\big(Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) - h(X)\big)^2 \\ &= \mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)^2 + \mathbb{E}\big(\mathbb{E}(Y|X) - h(X)\big)^2 + \\ &\quad 2\mathbb{E}\Big[\big(Y - \mathbb{E}(Y|X)\big)\big(\mathbb{E}(Y|X) - h(X)\big)\Big].\end{aligned}$$

Với mọi hàm $k(X)$ ta luôn có

$$\begin{aligned}\mathbb{E}\big(k(X)\,E(Y|X)\big) &= \int \Big[k(x)\int y\,p(y|x)\,dy\Big]p_X(x)\,dx \\ &= \int\int k(x)\,y\,p(y|x)p_X(x)\,dx\,dy \\ &= \int\int k(x)\,y\,p(x,\,y)\,dx\,dy = \mathbb{E}\big(k(X)\,Y\big).\end{aligned}$$

Đặt $k(X) = \mathbb{E}(Y|X) - h(X)$, suy ra

$$\begin{aligned}&\mathbb{E}\Big[\big(Y - \mathbb{E}(Y|X)\big)\big(\mathbb{E}(Y|X) - h(X)\big)\Big] = \mathbb{E}\Big[\big(Y - \mathbb{E}(Y|X)\big)\,k(X)\Big] \\ &= \mathbb{E}\big[k(X)\,Y\big] - \mathbb{E}\big[k(X)\,\mathbb{E}(Y|X)\big] = 0.\end{aligned}$$

Do đó

$$\begin{aligned}\mathbb{E}\big(Y - h(X)\big) &= \mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)^2 + \mathbb{E}\big(\mathbb{E}(Y|X) - h(X)\big)^2 \\ &\geqslant \mathbb{E}\big(Y - \mathbb{E}(Y|X)\big)^2.\end{aligned}$$

□

Như vậy $\mathbb{E}(Y|X)$ là hàm ước lượng Y có sai số bình phương trung bình nhỏ nhất. Phương trình $\varphi(X) = \mathbb{E}(Y|X)$ được gọi là *phương trình hồi quy* của Y theo X.

2.6.4 Hệ số hồi quy tuyến tính thực nghiệm

Giả sử X là đại lượng ngẫu nhiên độc lập còn Y là đại lượng ngẫu nhiên phụ thuộc và giữa chúng có tương quan tuyến tính

$$\mathbb{E}(Y|X) = AX + B, \;\; A \neq 0,$$

trong đó A, B chưa biết và được gọi là *hệ số hồi quy lý thuyết.*

Bài toán. Căn cứ vào n cặp quan sát (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) của (X, Y), ta cần đi tìm một phương trình $y = ax + b$ ước lượng cho phương trình hồi quy tuyến tính lý thuyết $\mathbb{E}(Y|X) = AX + B$.

Phương trình $y = ax + b$ được gọi là *phương trình hồi quy tuyến tính thực nghiệm*; a và b được gọi là *hệ số hồi quy tuyến tính thực nghiệm* của Y theo X. Chúng ta sử dụng phương pháp bình phương bé nhất để xác định giá trị của a và b.

Như vậy, giữa giá trị thực nghiệm và giá trị xác định từ phương trình hồi quy tuyến tính thực nghiệm tại x_i có sai số $|y_i - (ax_i + b)|$. Tiêu chuẩn để xác định phương trình hồi quy tuyến tính thực nghiệm $y = ax + b$ là đảm bảo được yêu cầu

$$F(a, b) = \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2 \Rightarrow \min .$$

Tìm cực tiểu của $F(a, b)$ dẫn đến hệ phương trình

$$\begin{cases} \dfrac{\partial F(a,\, b)}{\partial a} = -2 \displaystyle\sum_{i=1}^{n} (y_i - ax_i - b)\, x_i = 0; \\ \dfrac{\partial F(a,\, b)}{\partial b} = -2 \displaystyle\sum_{i=1}^{n} (y_i - ax_i - b) = 0, \end{cases}$$

tương đương với hệ

$$\begin{cases} \Big(\displaystyle\sum_{i=1}^{n} x_i^2 \Big) a + \Big(\displaystyle\sum_{i=1}^{n} x_i \Big) b = \displaystyle\sum_{i=1}^{n} x_i\, y_i; \\ \Big(\displaystyle\sum_{i=1}^{n} x_i \Big) a + nb = \displaystyle\sum_{i=1}^{n} y_i. \end{cases}$$

Giải hệ phương trình bậc nhất đối với a và b, ta được

$$\begin{cases} a = \dfrac{n\sum\limits_{i=1}^{n} x_i\, y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \Big(\sum\limits_{i=1}^{n} x_i\Big)^2}; \\ b = \dfrac{\sum\limits_{i=1}^{n} y_i - a\sum\limits_{i=1}^{n} x_i}{n}. \end{cases}$$

Ngoài ra, hệ số hồi quy tuyến tính thực nghiệm còn có thể xác định nhờ công thức tương đương

$$\begin{cases} a = \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{\hat{s}_X^2}; \\ b = \overline{y} - a\,\overline{x}\,. \end{cases}$$

Ví dụ 2.6.4. *Với giả thiết như ở Ví dụ 2.6.1:*

$$n = 5;\ \sum x_i = 40;\ \sum y_i = 350;\ \sum x_i^2 = 340;\ \sum x_i y_i = 2955.$$

a. Tìm phương trình hồi quy tuyến tính thực nghiệm của y theo x.

b. Nếu doanh thu của một tháng nào đó là 10 tỉ đồng, hãy dự đoán chi phí quảng cáo của công ty tháng đó là bao nhiêu.

Giải. *a. Hệ số hồi quy tuyến tính thực nghiệm*

$$a = \frac{5\times 2955 - 40\times 350}{5\times 340 - (40)^2} = 7,75; \quad b = \frac{350 - 7,75\times 40}{5} = 8.$$

Phương trình hồi quy tuyến tính thực nghiệm: $y = 7,75\,x + 8$.

b. $x = 10$ suy ra $y = 85{,}5$. Vậy chi phí quảng cáo của tháng đó khoảng 85,5 triệu đồng.

BÀI TẬP

1. Cho ví dụ về đám đông, một số đặc điểm có thể nghiên cứu và các phương pháp thực hiện việc lấy mẫu trên đám đông đó.

2. Phân biệt sự khác nhau giữa mẫu ngẫu nhiên và mẫu cụ thể, cho ví dụ minh họa.

3. Phân biệt sự khác nhau giữa đặc điểm định lượng và đặc điểm định tính. Cho ví dụ về hai đặc điểm cùng nghiên cứu trên một đám đông.

4. Khi đo độ dài của 36 chi tiết được lấy ngẫu nhiên từ một loại sản phẩm, người ta thu được bảng số liệu sau đây:

15 14 16 14 15 12 13 16 13 12 15 13 16 13 15

13 16 13 16 13 15 12 15 15 14 14 15 15 16 15

a. Lập bảng tần số và bảng tần suất.

b. Vẽ biểu đồ đa giác tần số và tần suất.

c. Tìm hàm phân phối mẫu.

5. Dưới đây là số liệu được lấy ngẫu nhiên về thời gian đợi của các khách hành (tính bằng giây) tại quầy thanh toán tiền ở một siêu thị đối với 48 khách hàng:

3	24	34	5	14	22	3	19	13	32	19	4	24	30	48	24
14	16	3	4	5	14	19	41	43	16	48	4	58	13	10	60
12	14	14	22	3	16	14	4	34	32	4	19	12	24	13	26

a. Lập bảng tần số ghép lớp và bảng tần suất ghép lớp.

b. Vẽ bảng tổ chức đồ tần số và tần suất.

c. Tính trung bình mẫu, phương sai mẫu và phương sai hiệu chỉnh mẫu.

6. Mẫu điều tra có kích thước 35 đối với hai đặc điểm X và Y của một loại sản phẩm được kết quả cho bởi bảng số liệu dưới đây:

$X\backslash Y$	64	65	66
6-10	3	8	3
10-14	0	5	2
14-16	6	1	0
16-20	0	3	4

a. Lập bảng tần số, tần suất của Y.

b. Những sản phẩm được gọi là đạt chất lượng nếu $X \leqslant 16$ và $Y > 64$. Tính tỉ lệ sản phẩm đạt chất lượng.

c. Lập bảng tần số và tính trung bình mẫu của chỉ tiêu Y đối với các sản phẩm có $X > 10$.

7. Cơ quan quản lý thị trường lấy số liệu về giá thành bán lẻ của một loại sản phẩm tại 40 đại lý (đơn vị: ngàn), người ta thu được bảng tần số như sau:

x_i	19	20	21	22
n_i	8	16	6	10

a. Tìm hàm phân phối mẫu.

b. Tính trung bình mẫu và độ lệch chuẩn mẫu.

8. Tìm hàm phân phối mẫu, trung bình mẫu, phương sai hiệu chỉnh mẫu đối với hai mẫu cụ thể sau:

a.

x_i	19,2	19,8	20,1	20,3	20,7
n_i	6	2	4	2	6

b.

x_i	460	480	490	505
n_i	5	6	10	4

9. Điều tra ngẫu nhiên ý kiến của 2500 số khách hàng thường xuyên đi xe taxi về chất lượng phục vụ của 3 hãng taxi thu được kết quả sau đây:

Chất lượng phục vụ	Hãng taxi		
	A	B	C
Rất tốt	140	110	205
Khá	230	150	350
Bình thường	350	225	520
Kém	80	15	125

Hãy tính đặc trưng mẫu cho từng hãng taxi và nêu đánh giá sơ bộ từ số liệu điều tra trên.

10. Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên kích thước n của đám đông X có $\mathbb{E}X = \mu$. Chứng minh rằng

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

đều là các ước lượng không chệch của phương sai $\mathbb{D}X$.

11. Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên kích thước n từ phân phối với hàm mật độ là:

$$p(x, \theta) = \begin{cases} \dfrac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{với } x > 0,\ \theta > 0, \\ 0 & \text{với } x \leqslant 0. \end{cases}$$

Tìm ước lượng hiệu quả của θ.

12. Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên kích thước n từ phân phối Poisson với tham số $\mathbb{E}X = \mathbb{D}X = \lambda > 0$. Tìm ước lượng hợp lý tối đa của λ.

13. Để xác định độ chính xác của một chiếc cân, người ta tiến hành cân một quả tạ. Kết quả thu được sau 7 lần cân như sau:

159,8 159,7 160,2 159,6 160,4 159,5 160,6 (kg)

(a) Tìm ước lượng không chệch của khối lượng quả cân.

(b) Tìm ước lượng không chệch của phương sai số đo trong hai trường hợp:
- Biết khối lượng quả cân là 160 kg.
- Chưa biết khối lượng của quả cân.

14. Cơ quan quản lý thị trường lấy số liệu về giá thành bán lẻ của một loại sản phẩm tại 40 đại lý, người ta thu được bảng tần số như sau: (đơn vị: ngàn đồng)

x_i	39	40	41	42
n_i	8	16	4	12

(a) Tính trung bình mẫu $\overline{x}$ và phương sai mẫu hiệu chỉnh s^2.

(b) Với độ tin cậy 95%, ước lượng khoảng giá thành bán lẻ trung bình mỗi sản phẩm.

15. Một dây chuyền sản xuất những thanh kim loại có chiều dài tuân theo luật phân phối chuẩn. Người ta chọn ngẫu nhiên ra một số thanh và đo chiều dài (đơn vị: cm) của chúng, thu được dãy số liệu sau:

149; 151; 148; 152; 151; 152; 149; 148; 149; 151; 152; 149; 151; 149; 152.

(a) Tính trung bình mẫu $\overline{x}$ và phương sai mẫu hiệu chỉnh s^2.

(b) Với độ tin cậy 90%, ước lượng khoảng độ dài trung bình của mỗi thanh kim loại.

16. Một dây chuyền tự động đóng gói một loại bao gạo có khối lượng tuân theo luật phân phối chuẩn với độ lệch chuẩn là 0,5. Người ta cân kiểm tra 20 bao gạo, thu được bảng tần số như sau: (đơn vị: kg)

x_i	49,3	49,5	49,9	50,2
n_i	6	2	4	8

(a) Tính trung bình mẫu $\overline{x}$ và phương sai mẫu hiệu chỉnh s^2.

(b) Với độ tin cậy 98%, ước lượng khoảng khối lượng trung bình của mỗi bao gạo.

17. Nhà sản xuất muốn ước lượng khối lượng sắt trong mỗi cuộn được sản xuất từ một dây chuyền công nghệ quốc gia. Theo tiêu chuẩn của công nghệ, độ lệch chuẩn là 8 kg. Điều tra một mẫu 50 cuộn được khối lượng sắt trung bình là 97kg.

(a) Với độ tin cậy là 99%, ước lượng khối lượng sắt trung bình của một cuộn.

(b) Với độ tin cậy là 99%, ước lượng khối lượng sắt trung bình tối thiểu của một cuộn.

(c) Nếu nhà sản xuất muốn ước lượng khối lượng sắt trung bình của mỗi cuộn đảm bảo độ chính xác là 2 kg thì cần điều tra thêm bao nhiêu cuộn nữa.

18. Một công ty có 500 đại lý, để đánh giá về mức doanh thu, người ta lấy mẫu gồm 36 đại lý. Kết quả thu được như sau: doanh thu trung bình là 84,5 triệu đồng và độ lệch chuẩn mẫu là 3 triệu đồng. Với độ tin cậy 99%, hãy ước lượng doanh thu tối thiểu và tối đa của công ty.

19. Người ta đo chiều sâu của biển bằng một loại thiết bị điện tử, kết quả đo tuân theo luật phân phối chuẩn có phương sai $400m^2$. Với độ tin cậy là 95%, cần phải đo ít nhất bao nhiêu lần để kết quả có sai số không vượt quá $15m$.

20. Một mẫu thống kê có kích thước $n = 64$, tuân theo luật phân phối chuẩn với trung bình mẫu là 200, độ lệch chuẩn mẫu là 3. Tìm độ tin cậy của ước lượng nếu khoảng ước lượng là $(199, 201)$.

21. Để đánh giá hiệu quả của một loại thuốc, người ta đem sử dụng cho 1000 bệnh nhân thì có 820 người khỏi bệnh. Với độ tin cậy 96 %,

(a) Hãy ước lượng khoảng cho tỉ lệ chữa khỏi bệnh của loại thuốc trên.

(b) Hãy ước lượng tỉ lệ chữa bệnh tối đa và tối thiểu của loại thuốc trên.

22. Tỉ lệ chính phẩm của một nhà máy là 90%. Với độ tin cậy 95%, muốn ước lượng tỉ lệ chính phẩm của nhà máy với độ dài khoảng tin cậy không quá 0,02 thì phải kiểm tra ít nhất bao nhiêu sản phẩm?

23. Một kho hàng tồn gồm 10.000 chiếc bút bi. Lấy mẫu gồm 100 chiếc bút từ kho hàng ra kiểm tra thì có 75 chiếc đạt chất lượng. Với độ tin cậy 95%, hãy ước lượng khoảng tỉ lệ số bút không đạt chất lượng và suy ra khoảng tin cậy số bút không đạt chất lượng của kho.

24. Tại một bang có 4 triệu người tham gia bầu cử, người ta phỏng vấn ngẫu nhiên 1000 cử tri thì có 720 cử tri ủng hộ một ứng cử viên A. Với độ tin cậy là 95%, có ít nhất bao nhiêu cử tri của bang đó đã ủng hộ ứng cử viên A?

25. Để đánh giá trữ lượng cá trong một hồ nuôi, người ta bắt 1000 con cá và đánh dấu chúng, sau đó thả lại hồ. Lần thứ hai người ta bắt 200 con thì thấy có 30 con được đánh dấu. Với độ tin cậy là 95%,

(a) Hãy ước lượng trữ lượng cá trong hồ.

(b) Nếu muốn sai số của ước lượng giảm đi một nửa thì cần phải bắt bao nhiêu con cá.

26. Quy định của một thiết bị phải có chiều dài là 300cm và độ lệch chuẩn là 3cm. Từ một lô hàng người ta lấy ra 40 chiếc, kết quả thu được độ dài trung bình là 301,2cm. Với mức ý nghĩa 5% , lô hàng trên có đạt tiêu chuẩn hay không?

27. Trong điều kiện chăn nuôi bình thường, lượng sữa thu được trung bình hàng ngày của một loại giống bò sữa là 19,4 (đơn vị: kg/ngày).

Lấy mẫu 49 con bò sữa ở một trang trại thu được lượng sữa trung bình của một con trong một ngày là 18,9 và độ lệch chuẩn mẫu là 3,24. Với mức ý nghĩa $\alpha = 0{,}08$, lượng sữa thu được hàng ngày từ bò sữa của trang trại có đúng chuẩn không?

28. Khối lượng chuẩn của một bao gạo được đóng gói bằng dây chuyền tự động là đại lượng ngẫu nhiên có phân phối chuẩn với khối lượng mỗi bao là 50 kg. Sau một thời gian hoạt động người ta nghi ngờ khối lượng đó có xu hướng giảm sút. Cân 28 bao gạo thu được khối lượng trung bình mỗi bao là 49,8 kg và độ lệch chuẩn mẫu là 0,6 kg. Với mức ý nghĩa 1%, hãy kết luận về nghi ngờ nói trên.

29. Thời gian trước đây, số tiền gửi tiết kiệm trung bình của mỗi khách hàng vào ngân hàng A là 1000 USD. Sau đợt tăng lãi suất tiết kiệm, kiểm tra ngẫu nhiên 36 khách hàng thu được kết quả: số tiền gửi trung bình là 1060 USD và độ lệch chuẩn mẫu là 100 USD. Với mức ý nghĩa 4%, việc tăng lãi suất có làm tăng lượng tiền gửi tiết kiệm của mỗi khách hàng không?

30. Một kênh truyền thông tuyên bố rằng 30% khán giả truyền hình yêu thích các chương trình phát sóng của họ. Thăm dò ý kiến ngẫu nhiên qua mạng đối với 800 người xem truyền hình thì có 192 người yêu thích các chương trình của kênh truyền thông đó. Với mức ý nghĩa 0,08, tỉ lệ trong tuyên bố trên có đúng với thực tế không?

31. Tỉ lệ phế phẩm của một nhà máy trước đây là 10%. Sau khi cải tiến kỹ thuật, kiểm tra 400 sản phẩm thì thấy có 38 phế phẩm. Với mức ý nghĩa là 1%, kiểm tra xem việc cải tiến kỹ thuật có mang lại hiệu quả không?

32. Tỉ lệ người chữa khỏi một loại bệnh bằng loại thuốc cũ là 80%. Người ta thay thế bằng loại thuốc mới để chữa bệnh cho 1000 người thì có 820 người khỏi bệnh. Với mức ý nghĩa 1%, có thể kết luận thuốc mới tốt hơn thuốc cũ không?

33. Hai giống vịt được nuôi sau 4 tháng. Lấy mẫu $n_1 = 50$ ở giống vịt thứ nhất, được $\overline{x}_1 = 1.9kg$ và $s_1^2 = 1$. Lấy mẫu $n_2 = 80$ ở giống vịt

thứ hai, được $\overline{x}_2 = 2kg$ và $s_2^2 = 0.8$. Với mức ý nghĩa $\alpha = 10\%$, hai giống vịt này có trọng lượng trung bình bằng nhau không?

34. Chọn ngẫu nhiên 20 đại lý có áp dụng khuyến mãi thu được số lượng bán trung bình mỗi ngày là 140 sản phẩm và độ lệch chuẩn mẫu là 12; còn tại 20 đại lý không có khuyến mãi được 2 số liệu tương ứng là 120 và 10. Giả sử lượng hàng bán được có phân phối chuẩn, có cùng phương sai. Với mức ý nghĩa 5%, hình thức khuyến mãi có làm tăng số lượng hàng bán không?

35. Một công ty bán hàng muốn kiểm tra hiệu quả từ việc thay đổi kiểu đóng gói. Chọn 2 mẫu: mẫu 1 là 35 đại lý bán hàng theo loại gói cũ và mẫu 2 là 35 đại lý bán hàng theo loại gói mới để thống kê về số gói hàng bán ra sau một tháng, thu được 2 giá trị đặc trưng cho 2 mẫu tương ứng như sau: loại gói cũ: $\overline{x_1} = 560$ gói, với $s_1 = 20$; loại gói mới: $\overline{x_2} = 580$ gói, với $s_2 = 30$. Với mức ý nghĩa 1%, hãy đánh giá việc thay đổi kiểu đóng gói có đem lại hiệu quả hay không?

36. Để so sánh tỉ lệ nảy mầm của hai giống cây trong điều kiện độ ẩm thấp. Người ta đem gieo 200 hạt giống loại I thì có 150 hạt nảy mầm, gieo 300 hạt giống loại II thấy có 210 hạt nảy mầm. Với mức ý nghĩa $\alpha = 0{,}05$, tỉ lệ nảy mầm trong điều kiện độ ẩm thấp của 2 giống cây trên có như nhau không ?

37. Lấy số liệu thực tế từ các hộ gia đình vay vốn của ngân hàng nông nghiệp đối với 2 huyện. Huyện A: có 2000 hộ vay thì có 1692 hộ sử dụng tiền vay có hiệu quả; huyện B: có 1000 hộ vay thì có 810 hộ sử dụng tiền vay có hiệu quả. Với mức ý nghĩa 5%, tỉ lệ hộ sử dụng tiền vay có hiệu quả của huyện A có cao hơn ở huyện B không?

38. Để đánh giá về chất lượng sản phẩm của nhà máy do 2 dây chuyền sản xuất. Người ta kiểm tra ngẫu nhiên 200 sản phẩm từ dây chuyền thứ nhất thì có 15 phế phẩm, kiểm tra 300 sản phẩm từ dây chuyền thứ hai thấy có 21 phế phẩm. Từ số liệu thu được có thể đánh giá sơ bộ dây chuyền nào làm việc tốt hơn. Với mức ý nghĩa $\alpha = 0{,}08$, kiểm định đánh giá sơ bộ đó.

39. Có hai phương pháp gieo một loại hạt giống: theo phương pháp A, gieo 125 hạt thấy có 90 hạt nảy mầm; theo phương pháp B, gieo 100 hạt thấy có 85 hạt nảy mầm. Từ số liệu thu được có thể đánh giá sơ bộ phương pháp gieo nào tốt hơn. Với mức ý nghĩa $\alpha = 0{,}05$, kiểm định đánh giá sơ bộ đó.

40. Tại một nhà máy làm việc theo chế độ 3 ca: buổi sáng, buổi chiều và buổi tối, chọn ngẫu nhiên một số sản phẩm để kiểm tra chất lượng, thu được bảng số liệu sau

Chất lượng	Ca		
	Sáng	Chiều	Tối
Chính phẩm	84	64	70
Phế phẩm	2	8	2

Với mức ý nghĩa $\alpha = 0{,}05$, có thể kết luận chất lượng sản phẩm phụ thuộc vào ca làm việc không?

41. Tại một nhà máy có 4 phân xưởng: I, II, III, IV; cùng sản xuất ra một loại sản phẩm với 3 tiêu chí đánh giá chất lượng: Loại A (tốt), loại B (đạt), loại C (chưa đạt). Kiểm tra 1000 sản phẩm khi nhập tổng kho, thu được bảng số liệu sau

Chất lượng / Xưởng	Loại A	Loại B	Loại C
I	105	90	25
II	135	102	13
III	124	100	6
IV	146	138	16

Với mức ý nghĩa $\alpha = 0{,}01$, có thể kết luận chất lượng sản phẩm phụ thuộc vào phân xưởng sản xuất không?

42. Bảng số liệu sau đây là kết quả thống kê về tổng giá trị hàng nông sản (X) và tổng đầu tư xây dựng đường giao thông (Y) của một huyện trong 6 năm như sau: (đơn vị: tỉ đồng)

X	60	45	75	90	80	70
Y	7	5	8	11	9	10

(a) Hãy xác định hệ số tương quan mẫu.

(b) Tìm phương trình hồi quy tuyến tính thực nghiệm của y theo x.

(c) Nếu tiền đầu tư xây dựng đường giao thông của một năm nào đó là 8,6 tỉ đồng, hãy dự đoán tổng giá trị hàng nông sản năm đó là bao nhiêu ?

43. Bảng số liệu sau đây là kết quả thu được của một công ty về số tiền dành cho các hoạt động chăm sóc khách hàng (X) và doanh thu (Y) trong 6 tháng như sau:

X	8	9	7	10	9	11
Y	600	700	500	900	800	1100

(đơn vị: triệu đồng).

(a) Hãy xác định hệ số tương quan mẫu.

(b) Nếu chi phí dành cho các hoạt động chăm sóc khách hàng của một tháng nào đó là 10,5 triệu đồng, hãy dự đoán doanh thu của công ty tháng đó là bao nhiêu ?

44. Thống kê ghi lại dân số của một tỉnh qua 8 năm từ năm 1985 đến 1992 được bảng số sau

Năm	1985	1986	1987	1988	1989	1990	1991	1992
Dân số (10000)	50	51	51	53	54	56	59	60

Để thuận tiện trong tính toán ta đặt $x =$ "năm" $- 1985$ và $y =$ "dân số" $- 50$ (đơn vị 10000 người). Hãy tìm phương trình hồi quy tuyến tính thực nghiệm của y theo x.

45. Tính hệ số tương quan mẫu và phương trình hồi quy tuyến tính thực nghiệm của y theo x dựa vào bảng tần số sau:

x_i	17	14	12	15	12	20
y_i	31	33	25	29	27	40
n_i	2	4	10	3	5	6

46. Bảng số liệu sau đây chỉ ra sự phụ thuộc của năng suất thu hoạch Y theo lượng phân bón X của một loại hoa màu trên 100 thửa ruộng.

Y	X			
	20	25	30	35
400	12	5	1	1
420	6	18	3	2
450	2		10	9
490		1	10	20

Tính hệ số tương quan mẫu và phương trình hồi quy tuyến tính thực nghiệm của năng suất thu hoạch theo lượng phân bón.

3

Một số mô hình toán kinh tế

3.1 Một số mô hình toán kinh tế điển hình

3.1.1 Mô hình lập kế hoạch sản xuất

3.1.2 Mô hình bài toán vận tải

3.1.3 Mô hình bài toán khẩu phần thức ăn

3.1.4 Một số mô hình khác

3.2 Bài toán quy hoạch tuyến tính

3.2.1 Một số định nghĩa và tính chất

3.2.2 Cặp bài toán đối ngẫu và ứng dụng

3.3 Phương pháp đơn hình giải bài toán quy hoạch tuyến tính

3.3.1 Cơ sở lý luận của phương pháp đơn hình

3.3.2 Thuật toán đơn hình cho bài toán quy hoạch có cơ sở đơn vị

3.3.3 Thuật toán đơn hình cho bài toán qui hoạch

Tài liệu tham khảo

[1] Beller, G., Smith, T., Abelmann, W., and Hood, W. Digitalis intoxication: A prospective clinical study with serum level correlations. *N. Eng. J. Med.*, 284 (1971), 989-997.

[2] Diamond, G., and Forrester. J. Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. *N. Eng. J. Med.*, 300 (1979), 1350-1358.

[3] Durrett, Rick. *Probability: Theory and Examples*. Fourth edition. Cambridge University Press, 2010.

[4] Gastwirth, J. The statistical precision of medical screening procedures. *Statistical Science*. 3 (1987), 213-222.

[5] Glass, D., and Hall. J. A study of intergeneration changes in status. In *Social Mobility in Britain*, D. Glass (ed.). Glencoe, III (1954). Free Press.

[6] Gut, Allan. *Probability: a graduate course*. Second edition. Springer Texts in Statistics. Springer, New York, 2013. xxvi+600 pp.

[7] Rice, John. *Mathematical Statistics and Data Analysis*. Third edition. Duxbury, 2007.

[8] Ross, Sheldon. *A first course in probability*. Eighth edition. Pearson, 2010.

[9] Wackerly, Dennis, Mendenhall III, William, and Scheaffer, Richard. *Mathematical Statistics with Applications*. Seventh edition. Thomson, 2008.