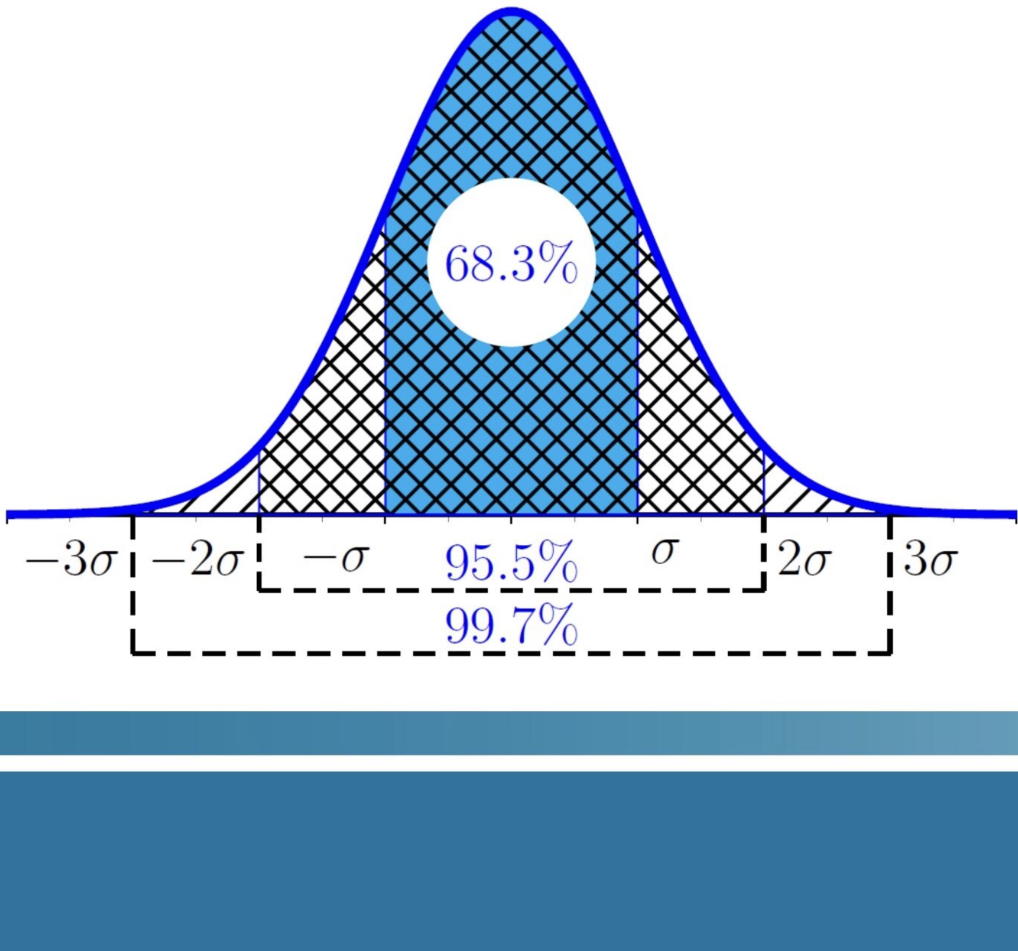
Biên soạn: TS. LÊ ANH XUÂN (Chủ biên)
ThS. DƯƠNG HÙNG MẠNH

GIÁO TRÌNH
XÁC SUẤT THỐNG KÊ

(Dùng cho chuyên ngành kỹ thuật)



NHÀ XUẤT BẢN ĐẠI HỌC CẦN THƠ
2021

TS. LÊ ANH XUÂN (Chủ biên)
GIÁO TRÌNH XÁC XUẤT THỐNG KÊ (Dùng cho chuyên ngành kỹ thuật)
2021

Lời nói đầu

Giáo sư Vũ Hà Văn (Giám đốc Khoa học của Viện Nghiên cứu Dữ liệu lớn VinBDI, Hội viên danh dự của Hiệp hội Toán thống kê thế giới năm 2020) đã từng nói rằng *"Xác suất thống kê hấp dẫn tôi không chỉ bởi vẻ đẹp toán học mà vì ý nghĩa thực sự của nó trong cuộc sống. Xác suất thống kê là nền tảng của khoa học dữ liệu và có lẽ sẽ là một trong những môn học quan trọng nhất trong tương lai"*.

Thống kê là một trong những ngành khoa học có ứng dụng nhiều nhất hiện nay với vai trò to lớn trong tất cả các nghiên cứu định lượng. Các ứng dụng, nhất là trong máy học, đem lại nhiều phát minh mới như các sản phẩm về trí tuệ nhân tạo hay AI (Artificial Intelligence). Thống kê trong nghiên cứu về gen giúp phát hiện những kiến thức mới về di truyền. Thống kê trong y học giúp phát hiện những liên quan bất ngờ giữa bệnh và thuốc. Hay thống kê trong kinh tế phát hiện những quy luật mới về tiêu dùng. Bởi chính vì lý do đó mà Xác suất thống kê là học phần quan trọng và được giảng dạy cho tất cả các ngành trong trường đại học khối kinh tế, kỹ thuật và Trường Đại học Kỹ thuật - Công nghệ Cần Thơ cũng không nằm ngoài xu hướng này.

Giáo trình Xác suất thống kê (Dùng cho chuyên ngành kỹ thuật) được biên soạn trên cơ sở Bài giảng Xác suất thống kê của Bộ môn Toán, có điều chỉnh ở phần nội dung và bổ sung thêm bài tập cho phù hợp với các đối tượng giảng dạy cũng như yêu cầu của học phần. Giáo trình bao gồm **8 chương** được khái quát dưới đây:

Chương 1: Bổ túc về giải tích tổ hợp

Chương này nhắc lại cho sinh viên những kiến thức cơ bản về giải tích tổ hợp đã được học ở trung học phổ thông nhưng được trình bày chi tiết và chứng minh chặt chẽ hơn đồng thời bổ túc thêm một số khái niệm mới như chỉnh hợp lặp và hoán vị lặp.

Chương 2: Các khái niệm cơ bản của xác suất

Nội dung chính của chương là các vấn đề cơ bản của xác suất. Chẳng hạn, khái niệm về biến cố, các loại biến cố, phép toán và tính chất; định nghĩa xác suất theo các hướng tiếp cận khác nhau, tính chất và ý nghĩa của xác suất; các công thức tính xác suất thường dùng.

Chương 3: Đại lượng ngẫu nhiên

Trong chương này sinh viên sẽ được nghiên cứu hai vấn đề là: Đại lượng ngẫu nhiên một chiều và đại lượng ngẫu nhiên hai chiều. Với đại lượng ngẫu nhiên một chiều, những kiến thức về hàm phân phối, bảng phân phối xác suất và phân phối xác suất cho hàm cùng các tham số đặc trưng sẽ được giới thiệu. Cũng với thứ tự trình bày như vậy cho đại lượng ngẫu nhiên hai chiều, tuy nhiên ở đây sinh viên còn được tìm hiểu thêm về hiệp phương sai, hệ số tương quan.

Chương 4: Một số phân phối xác suất thông dụng

Đây là một trong những chương quan trọng nhất của giáo trình. Qua chương này sinh viên sẽ được trang bị kiến thức về các luật phân phối xác suất thông dụng như phân phối nhị thức, phân phối siêu bội, phân phối Poisson,... của đại lượng ngẫu nhiên rời rạc; phân phối Chuẩn, phân phối Chi bình phương, phân phối Student,... của đại lượng ngẫu nhiên liên tục.

Chương 5: Lý thuyết mẫu

Nội dung của chương là giới thiệu về phương pháp mẫu, các đặc trưng mẫu ngẫu nhiên, mẫu cụ thể và phương pháp tính các đặc trưng mẫu cụ thể này.

Chương 6: Ước lượng tham số

Là nội dung trọng tâm của phần thống kê, gồm có ước lượng điểm và ước lượng khoảng. Trong đó, đặc biệt quan tâm đến các bài toán về ước lượng khoảng cho trung bình, tỉ lệ, phương sai và các bài toán liên quan.

Chương 7: Kiểm định giả thuyết

Nội dung của chương bao gồm phần kiến thức quan trọng của thống kê như kiểm định giả thuyết một mẫu, hai mẫu và kiểm định phi tham số.

Chương 8: Tương quan - Hồi quy

Nội dung chính của chương là phân tích tương quan tuyến tính và phân tích hồi quy.

Ở mỗi chương, đan xen giữa khái niệm, tính chất và định lí là các ví dụ được chọn lọc kỹ càng, trình bày chi tiết và dễ hiểu giúp cho sinh viên có thể tự học để nắm kiến thức vững chắc hơn nữa; trước khi kết thúc đều có phần bài tập chương cho sinh viên nghiên cứu tìm lời giải. Giáo trình được xây dựng với mục tiêu lấy người học làm trung tâm nên các nội dung đều được trình bày rõ ràng, mạch lạc, có ví dụ minh họa đầy đủ; những tính chất và định lí quan trọng đều được chứng minh để sinh viên có thể tự nghiên cứu.

Về hình thức, giáo trình được biên soạn bằng phần mềm LaTeX, các hình minh họa được vẽ bởi phần mềm GeoGebra Classic 6. Đây đều là những phần mềm mã nguồn mở được dùng phổ biến, hỗ trợ rất tốt cho công tác biên tập, xuất bản và mang tính chuyên nghiệp cao.

Cuối cùng xin gửi lời cảm ơn chân thành đến các giảng viên của Bộ môn Toán và quý đồng nghiệp cùng bạn đọc đã có những ý kiến đóng góp quý báu giúp cho giáo trình được hoàn thiện. Tuy nhiên trong quá trình biên soạn chắc chắn sẽ không tránh khỏi thiếu sót, các tác giả mong muốn tiếp tục nhận được sự đóng góp của đồng nghiệp và bạn đọc để giáo trình ngày càng hoàn hiện hơn nữa.

Cần Thơ, ngày 28 tháng 6 năm 2021
Các tác giả

Mục lục

Danh sách hình vẽ

Danh sách ký hiệu

Ký hiệu	**Ý nghĩa**
$\emptyset$	*Tập rỗng / Biến cố không thể*
Ω	*Biến cố chắc chắn / Không gian biến cố sơ cấp*
$\mathbf{N}$, $\mathbf{Z}$, $\mathbf{R}$	*Tập số tự nhiên, tập số nguyên, tập số thực*
$A \cup B$ hoặc $A + B$	*Hợp của hai tập A và B / Tổng của hai biến cố A và B*
$A \cap B$ hoặc $A \cdot B$	*Giao của hai tập A và B / Tích của hai biến cố A và B*
$A \setminus B$ hoặc $A - B$	*Hiệu của hai tập / biến cố A và B*
$A \Rightarrow B$ hoặc $A \subset B$	*A là biến cố kéo theo của B / A là tập con của B*
$A \Leftrightarrow B$ hoặc $A = B$	*A và B là hai biến cố tương đương*
$\overline{A}$	*Phần bù của tập A / Biến cố đối lập của biến cố A*
S_A	*Diện tích của miền A*
$P(A)$	*Xác suất của biến cố A*
$P(A\|B)$	*Xác suất của biến cố A xét trong điều kiện của biến cố B*
$P(a < X < b)$	*Xác suất để X nhận giá trị trong* (a, b)
$\mathrm{Mod}(X)$	*Giá trị tin chắc của đại lượng ngẫu nhiên X*
$\mathrm{E}(X)$	*Kỳ vọng của X*
$\mathrm{Var}(X)$ hoặc $\mathrm{D}(X)$	*Phương sai của X*
$\sigma(X)$	*Độ lệch chuẩn của X*
$\mathrm{Med}(X)$	*Trung vị của X*
$F(x)$ hoặc $F_X(x)$	*Hàm phân phối xác suất của X*
$f(x)$	*Hàm mật độ xác suất của X*
$F(X, Y)$	*Hàm phân phối xác suất của* (X, Y)
$f(X, Y)$	*Hàm mật độ của* (X, Y)
$F(x\|y_0)$	*Hàm phân phối xác suất của X với điều kiện* $Y = y_0$
$B(n; p)$	*Phân phối nhị thức*
$H(N; M; n)$	*Phân phối siêu bội*
$P(\lambda)$	*Phân phối Poisson*
$N(\mu; \sigma^2)$ và $N(0; 1)$	*Phân phối chuẩn và phân phối chuẩn tắc*
$\varphi(x)$ và $\Phi(x)$	*Hàm mật độ và phân phối của* $N(0; 1)$
$\chi^2(n)$	*Phân phối Chi bình phương*
$T(n)$	*Phân phối Student*
$F(m, n)$	*Phân phối Fisher*
$\overline{X}$ và S	*Trung bình mẫu và độ lệch điều chỉnh mẫu ngẫu nhiên*
$\overline{x}$ và s	*Trung bình mẫu và độ lệch điều chỉnh mẫu cụ thể*
H_0 và H_1, H_2, H_3	*Giả thuyết và các đối thuyết*
r	*Hệ số tương quan mẫu*
$\hat{y}_X$	*Hàm hồi quy của Y đối với X*
$\square$	*Kết thúc một chứng minh*

Chương 1

BỔ TÚC VỀ GIẢI TÍCH TỔ HỢP

1.1. TẬP HỢP

1.1.1. Khái niệm về tập hợp

Tập hợp là một khái niệm cơ bản của Toán học không được định nghĩa mà chỉ được mô tả qua các ví dụ, tương tự như khái niệm điểm, đường thẳng, mặt phẳng trong hình học. Ta có các ví dụ về tập hợp như: Tập hợp sinh viên của một trường đại học, tập hợp các nghiệm của phương trình $x^2 - 5x + 6 = 0$, tập hợp $\mathbf{Z}$ các số nguyên,...

Một thành phần tạo nên tập hợp gọi là một phần tử của tập hợp. Để chỉ phần tử a thuộc tập hợp A ta viết $a \in A$, ngược lại ta viết $a \notin A$ hoặc $a \overline{\in} A$ và đọc là phần tử a không thuộc tập hợp A.

Tập hợp không chứa bất kỳ phần tử nào được gọi là tập rỗng, ký hiệu $\emptyset$. Để xác định một tập hợp ta có thể dùng một trong hai cách:

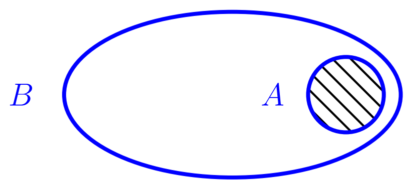
$\ast$ *Cách 1:* Xác định tập hợp bằng cách liệt kê các phần tử của nó như $A = \{0, 1, 2, ..., 9\}$, $B = \{a, b, c\}$,...

$\ast$ *Cách 2:* Xác định tập hợp bằng cách chỉ ra tính chất đặc trưng của phần tử thuộc nó. Chẳng hạn, $S = \{x \in \mathbf{R} : x^3 - 2x + 1 = 0\}$, $T = \{n \in \mathbf{N} : n$ là ước của $246\}$,...

1.1.2. Các phép toán trong tập hợp

Cho A và B là hai tập hợp tùy ý khi đó ta có các phép toán sau:

a. Tập con: Tập hợp A gọi là con của B, nếu mọi phần tử của tập A đều là phần tử của tập B, ký hiệu $A \subset B$. Ta có: $A \subset B \Leftrightarrow \{\forall x \in A \Rightarrow x \in B\}$.

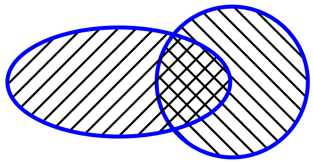


Hình 1.1: Tập hợp A là con tập hợp B

Khi đó ta còn nói A nằm trong B hoặc B chứa A.

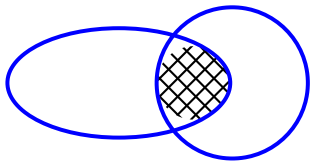
Nếu $A \subset B$ và $B \subset A$, nghĩa là mọi phần tử của A cũng là của B và ngược lại, thì ta nói A, B là hai tập bằng nhau và viết $A = B$. Ta có: $A = B \Leftrightarrow \{x \in A \Leftrightarrow x \in B\}$.

b. Phép hợp: Hợp (hay tổng) của A và B là tập gồm các phần tử hoặc thuộc A hoặc thuộc B, ký hiệu là $A \cup B$ (hay $A + B$). Ta có: $A \cup B = \{x : x \in A \vee x \in B\}$.



Hình 1.2: Hợp của hai tập A và B

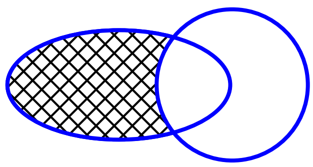
c. Phép giao: Giao của A và B là tập gồm các phần tử vừa thuộc A vừa thuộc B, ký hiệu $A \cap B$ (hay $A \cdot B$). Ta có: $A \cap B = \{x : x \in A \wedge x \in B\}$.



Hình 1.3: Giao của hai tập A và B

Nếu $A \cap B = \emptyset$ thì ta nói A và B rời nhau.

d. Phép hiệu: Hiệu của tập B đối với tập A là tập hợp gồm các phần tử thuộc B nhưng không thuộc A và ký hiệu là $B \setminus A$. Ta có: $B \setminus A = \{x : x \in B \wedge x \notin A\}$.



Hình 1.4: Hiệu của tập B đối với tập A

Đặc biệt, nếu $A \subset B$ thì $B \setminus A$ gọi là phần bù của A trong B và ký hiệu $\mathrm{C}_B A$ hoặc $\overline{A}$.

1.2. QUY TẮC ĐẾM

1.2.1. Quy tắc cộng

Giả sử một công việc nào đó có thể được thực hiện theo phương án A hoặc phương án B. Trong đó:

◇ Phương án A có n cách thực hiện,

◇ Phương án B có m cách thực hiện.

Khi đó, công việc có thể được thực hiện bởi $n + m$ cách.

Ví dụ 1.1. Có 9 quả cam và 5 quả quýt, có bao nhiêu cách chọn một trong các quả ấy?

Giải: Vì chỉ chọn một trong số những quả ấy nên ta có thể thực hiện công việc bằng 2 phương án:

+ Phương án 1: Chọn cam có 9 cách chọn,

+ Phương án 2: Chọn quýt có 5 cách chọn.

Vậy có tất cả là $9 + 5 = 14$ cách chọn.

Tổng quát: Giả sử một công việc có thể được thực hiện theo 1 trong k phương án khác nhau $A_1; A_2; \ldots; A_k$. Trong đó:

⋄ Phương án A_1 có n_1 cách thực hiện,

⋄ Phương án A_2 có n_2 cách thực hiện,

....

⋄ Phương án A_k có n_k cách thực hiện.

Khi đó, công việc được thực hiện bởi $n_1 + n_2 + \cdots + n_k = \sum_{i=1}^{k} n_i$ cách.

Ví dụ 1.2. Từ các chữ số 1, 2, 3 có thể lập được bao nhiêu số có các chữ số khác nhau?

Giải: Từ các chữ số 1, 2, 3 có thể lập được:

+ Ba số khác nhau có 1 chữ số là: 1, 2, 3.

+ Sáu số có 2 chữ số khác nhau là: 12, 21, 13, 31, 23, 32.

+ Sáu số có 3 chữ số khác nhau là: 123, 132, 213, 231, 312, 321.

Theo quy tắc cộng, có tất cả 15 số có các chữ số khác nhau lập từ các chữ số 1, 2, 3.

1.2.2. Quy tắc nhân

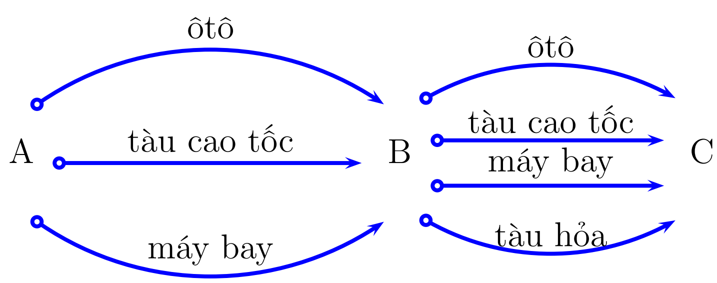
Giả sử một công việc nào đó có thể được thực hiện thông qua 2 giai đoạn A và B. Trong đó:

⋄ Giai đoạn A có n cách thực hiện,

⋄ Giai đoạn B có m cách thực hiện.

Khi đó, để hoàn thành công việc có $n \cdot m$ cách thực hiện.

Ví dụ 1.3. Để đi từ tỉnh A đến tỉnh B có thể đi bằng: Ôtô, tàu cao tốc hoặc máy bay. Đi từ tỉnh B đến tỉnh C có thể đi bằng: Ôtô, tàu cao tốc, tàu hỏa hoặc máy bay. Hỏi có bao nhiêu cách đi từ tỉnh A đến tỉnh C, bắt buộc phải ghé qua tỉnh B?



Giải: Để đi từ tỉnh A đến tỉnh C phải ghé qua tỉnh B, ta thực hiện cả 2 giai đoạn:

+ Đi từ tỉnh A đến tỉnh B có 3 cách,

+ Đi từ tỉnh B đến tỉnh C có 4 cách.

Vậy theo quy tắc nhân có tất cả là $3 \times 4 = 12$ cách đi.

Tổng quát: Giả sử một công việc được thực hiện thông qua k giai đoạn $A_1, A_2, \ldots, A_k$. Trong đó:

⋄ Giai đoạn A_1 có n_1 cách thực hiện,

⋄ Giai đoạn A_2 có n_2 cách thực hiện,

....

⋄ Giai đoạn A_k có n_k cách thực hiện.

Khi đó, để hoàn thành công việc có $n_1 \cdot n_2 \cdots n_k = \prod_{i=1}^{k} n_i$ cách thực hiện.

Ví dụ 1.4. Từ tập $\{0, 1, 2, 3, 4, 5\}$ có thể lập được bao nhiêu số tự nhiên có 4 chữ số:

a) Gồm các chữ số khác nhau?

b) Gồm các chữ số có thể giống nhau?

c) Gồm các chữ số có thể giống nhau và lớn hơn 3003?

d) Gồm các chữ số khác nhau và số đó chia hết cho 10?

Giải: Gọi số tự nhiên cần tìm có dạng $X = \overline{abcd}$.

a) Việc tìm X gồm các chữ số khác nhau phải thực hiện các giai đoạn:

+ Chọn chữ số a $(a \neq 0)$ có 5 cách chọn,

+ Chọn chữ số b $(b \neq a)$ có 5 cách chọn,

+ Chọn chữ số c $(c \neq a, b)$ có 4 cách chọn,

+ Chọn chữ số d $(d \neq a, b, c)$ có 3 cách chọn,

Theo quy tắc nhân có $5 \cdot 5 \cdot 4 \cdot 3 = 300$ số.

b) Việc tìm X gồm các chữ số có thể giống nhau phải thực hiện các giai đoạn:

+ Chọn chữ số a $(a \neq 0)$ có 5 cách chọn,

+ Chọn chữ số b có 6 cách chọn,

+ Chọn chữ số c có 6 cách chọn,

+ Chọn chữ số d có 6 cách chọn,

Theo quy tắc nhân có $5 \cdot 6 \cdot 6 \cdot 6 = 1080$ số.

c) Đầu tiên ta tìm số các số $X \geq 3000$. Khi đó để tìm X phải thực hiện các giai đoạn:

+ Chọn chữ số a $(a \geq 3)$ có 3 cách chọn,

+ Chọn chữ số b có 6 cách chọn,

+ Chọn chữ số c có 6 cách chọn,

+ Chọn chữ số d có 6 cách chọn,

Theo quy tắc nhân có $3 \cdot 6 \cdot 6 \cdot 6 = 648$ số lớn hơn hoặc bằng 3000.

Trong các số đó có 4 số nhỏ hơn hoặc bằng 3003 là 3000, 3001, 3002 và 3003.

Vậy số các số X lớn hơn 3003 là $648 - 4 = 644$ số.

d) Để X chia hết cho 10 thì $d = 0$. Khi đó, việc tìm X phải thực hiện các giai đoạn:

+ Chọn chữ số a $(a \neq 0)$ có 5 cách chọn,

+ Chọn chữ số b $(b \neq a, 0)$ có 4 cách chọn,

+ Chọn chữ số c $(c \neq a, b, 0)$ có 3 cách chọn,

Theo quy tắc nhân có $5 \cdot 4 \cdot 3 = 60$ số chia hết cho 10.

1.3. CÁC CÔNG THỨC

1.3.1. Chỉnh hợp

Định nghĩa 1.1. Giả sử có một tập hợp gồm n phần tử, mỗi cách lấy ra từ tập đó một tập con k $(0 < k \leq n)$ phần tử *có thứ tự* gọi là một chỉnh hợp chập k của n phần tử.

Định lí 1.1. *Số các chỉnh hợp chập k của một tập hợp có n phần tử, ký hiệu là A_n^k và được xác định bởi công thức sau:*

$$A_n^k = n(n-1)(n-2)\cdots\big(n-(k-1)\big).$$

Ta có thể viết gọn lại là:

$$A_n^k = \frac{n!}{(n-k)!} \quad (\textit{Quy ước } 0! = 1).$$

Chứng minh. Ta nhận thấy việc lấy một chỉnh hợp chập k của n phần tử có thể chia làm k giai đoạn như sau:

⋄ Giai đoạn thứ nhất, chọn phần tử thứ nhất có n cách chọn,

⋄ Giai đoạn thứ hai, chọn phần tử thứ hai có $n - 1$ cách chọn,

⋄ Giai đoạn thứ ba, chọn phần tử thứ ba có $n - 2$ cách chọn,

....

⋄ Giai đoạn cuối cùng, chọn phần tử thứ k có $n - (k - 1)$ cách chọn.

Vậy nên theo quy tắc nhân có $n(n-1)(n-2)\cdots\big(n-(k-1)\big)$ cách khác nhau để lập chỉnh hợp chập k của n phần tử. Tức là:

$$A_n^k = n(n-1)(n-2)\cdots\big(n-(k-1)\big).$$

□

Ví dụ 1.5. Giải vô địch bóng đá của một quốc gia có 32 đội bóng tham dự. Biết rằng, các đội sẽ thi đấu vòng tròn 2 lượt: Lượt đi trên sân khách và lượt về ở sân nhà. Hỏi để chọn được đội vô địch thì Ban điều hành giải phải tổ chức tất cả bao nhiêu trận đấu?

Giải: Do có quy định lượt đi và lượt về (phân biệt thứ tự) nên số trận đấu phải tổ chức là $A_{32}^2 = 992$ trận.

1.3.2. Hoán vị

Định nghĩa 1.2. Một hoán vị của n ($n \geq 1$) phần tử phân biệt là một cách *sắp thứ tự* của n phần tử đó.

Định lí 1.2. *Số các hoán vị của n phần tử, ký hiệu là P_n và được xác định bởi công thức:*

$$P_n = n!$$

Chứng minh. Từ định nghĩa ta thấy, hoán vị là một trường hợp đặc biệt của chỉnh hợp khi $k = n$. □

Ví dụ 1.6. Một đoàn khách du lịch đến tham quan 5 địa điểm M, N, P, Q, K ở thành phố Cần Thơ. Họ có thể tham quan theo một thứ tự nào đó, chẳng hạn M → N → P → Q → K. Như vậy, mỗi cách chọn thứ tự các địa điểm tham quan là một hoán vị của tập $\{M, N, P, Q, K\}$. Do đó, đoàn khách có tất cả là $5! = 120$ cách chọn.

Ví dụ 1.7. Trong một phòng học có 2 bàn dài, mỗi bàn có 5 ghế. Người ta muốn xếp chỗ ngồi cho 10 bạn học sinh gồm 5 nam và 5 nữ. Hỏi có bao nhiêu cách xếp nếu:

a) Các học sinh ngồi tùy ý?

b) Các học sinh nam ngồi một bàn, các học sinh nữ ngồi một bàn?

Giải:

a) Hai bàn dài, mỗi bàn có 5 ghế, như vậy có tổng cộng 10 chỗ ngồi. Mỗi cách xếp ngẫu nhiên 10 bạn vào 10 chỗ ngồi chính là một hoán vị của 10 phần tử. Vậy nên số cách xếp sẽ là $10! = 3628800$ cách.

b) Ta chia công việc xếp 10 bạn vào 2 bàn dài sao cho các học sinh nam ngồi một bàn, các học sinh nữ ngồi một bàn qua 3 giai đoạn:

+ Giai đoạn 1: Chọn bàn cho học sinh nam và bàn cho học sinh nữ có $2! = 2$ cách.

+ Giai đoạn 2: Xếp 5 bạn nam vào bàn dành cho nam có $5! = 120$ cách.

+ Giai đoạn 3: Xếp 5 bạn nữ vào bàn dành cho nữ có $5! = 120$ cách.

Vậy theo quy tắc nhân, số cách xếp sẽ là: $2 \cdot 120 \cdot 120 = 28800$ cách.

1.3.3. Tổ hợp

Định nghĩa 1.3. Giả sử có một tập hợp gồm n phần tử, mỗi cách lấy ra từ tập đó một tập con k ($0 < k \leq n$) phần tử *không kể thứ tự* gọi là một tổ hợp chập k của n phần tử.

Định lí 1.3. *Số các tổ hợp chập k của n phần tử, ký hiệu là C_n^k và được xác định bởi công thức:*

$$C_n^k = \frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!}.$$

Chứng minh. Vì không kể thứ tự nên hai tổ hợp khác nhau khi có ít nhất một phần tử khác nhau ($\{a, b\} \neq \{b, c\}$). Mỗi cách lấy không kể thứ tự tương ứng với $k!$ cách lấy có thứ tự. Như vậy số các chỉnh hợp là tích của số các tổ hợp và số các hoán vị trong mỗi tổ hợp đó. □

Tính chất 1.1. *Cho n và k là số tự nhiên, ta có:*

i) $C_n^k = C_n^{n-k}$ *với* $0 \leq k \leq n$.

ii) $C_n^k + C_n^{k-1} = C_{n+1}^k$ *với* $1 \leq k \leq n$.

Ví dụ 1.8. Cũng giả thiết như ở *Ví dụ 1.5* nhưng chỉ thi đấu vòng tròn một lượt (không phân biệt thứ tự). Để chọn ra đội vô địch thì sẽ chỉ phải tổ chức $C_{32}^2 = 496$ trận.

Ví dụ 1.9. Một người cài đặt mật khẩu cho máy tính của mình gồm 7 chữ cái. Người đó chỉ nhớ các chữ cái đó là A, B, B, C, C, C, D nhưng không nhớ thứ tự của chúng. Hỏi có thể có bao nhiêu mật khẩu được tạo thành?

Giải: Để chọn một mật khẩu người đó phải thực hiện công việc qua 4 giai đoạn:

+ Giai đoạn 1: Chọn vị trí cho chữ A, có 7 cách chọn.

+ Giai đoạn 2: Chọn vị trí cho 2 chữ B, mỗi cách chọn là chọn 2 vị trí bất kỳ (không kể thứ tự) trong 6 vị trí còn lại, có $C_6^2 = 15$ cách chọn.

+ Giai đoạn 3: Chọn vị trí cho 3 chữ C, mỗi cách chọn là chọn 3 vị trí bất kỳ (không kể thứ tự) trong 4 vị trí còn lại, có $C_4^3 = 4$ cách chọn.

+ Giai đoạn 4: Chọn vị trí cho chữ D, có 1 cách chọn.

Theo quy tắc nhân, số mật khẩu có thể được tạo thành là $7 \cdot 15 \cdot 4 \cdot 1 = 420$.

1.3.4. Nhị thức Newton

Trong chương trình toán phổ thông, ta đã biết đến các hằng đẳng thức:

$$(a+b)^2 = a^2 + 2ab + b^2 = C_2^0 a^2 b^0 + C_2^1 a^1 b^1 + C_2^2 a^0 b^2,$$
$$(a+b)^3 = a^3 + 3a^2 b + 3ab^2 + b^3 = C_3^0 a^3 b^0 + C_3^1 a^2 b^1 + C_3^2 a^1 b^2 + C_3^3 a^0 b^3.$$

Vấn đề đặt ra là: *"Với mọi số tự nhiên n thì đẳng thức $(a+b)^n$ được khai triển như thế nào?"*

Ngay từ thời trung học, Newton đã chứng minh được công thức tổng quát cho khai triển $(a+b)^n$ như sau:

$$(a+b)^n = C_n^0 a^n b^0 + C_n^1 a^{n-1} b^1 + C_n^2 a^{n-2} b^2 + \cdots + C_n^k a^{n-k} b^k + \cdots + C_n^n a^0 b^n = \sum_{k=0}^{n} C_n^k a^{n-k} b^k,$$

gọi là nhị thức Newton.

1.3.5. Chỉnh hợp lặp

Định nghĩa 1.4. Giả sử có một tập hợp gồm n phần tử, lấy *có hoàn lại* k phần tử (k có thể lớn hơn n) *có thứ tự* từ tập đó được gọi là chỉnh hợp lặp chập k của n phần tử.

Định lí 1.4. *Số các chỉnh hợp lặp chập k của n phần tử (k có thể lớn hơn n) ký hiệu là $\widetilde{A}_n^k$ và được tính theo công thức:*

$$\widetilde{A}_n^k = n^k.$$

Ví dụ 1.10. Có 7 bình gốm được xếp tùy ý vào 3 ngăn trên kệ, hỏi rằng có tất cả bao nhiêu cách sắp xếp?

Giải: Do mỗi lần xếp 1 bình gốm lên trên kệ ta xem như lấy 1 ngăn trong 3 ngăn và quá trình này sẽ lặp lại 7 lần (vì có 7 bình cần xếp). Nên số cách sắp xếp 7 bình gốm vào 3 ngăn là một chỉnh hợp lặp chập 7 của 3 phần tử. Do đó, có $\widetilde{A}_3^7 = 3^7$ cách sắp xếp bình gốm lên kệ.

Ta cũng có thể giải bằng cách sử dụng quy tắc nhân như sau:

Xếp bình thứ nhất lên kệ 3 ngăn có 3 cách, xếp bình thứ hai lên kệ 3 ngăn cũng có 3 cách,..., tương tự xếp đến bình thứ bảy lên kệ 3 ngăn cũng có 3 cách (do một ngăn có thể chứa nhiều hơn 2 bình). Theo quy tắc nhân có 3^7 cách sắp xếp.

1.3.6. Hoán vị lặp

Định nghĩa 1.5. Mỗi cách *sắp thứ tự* của n phần tử mà trong đó có k ($k \neq n$) phần tử *giống nhau* (không phân biệt) được gọi là một hoán vị lặp.

Định lí 1.5. *Số các hoán vị lặp của n phần tử, trong đó có k phần tử giống nhau, được ký hiệu là $P_n(k)$ và được xác định bởi công thức:*

$$P_n(k) = \frac{n!}{k!}.$$

Ví dụ 1.11. Từ các chữ số 1, 2, 3, 4 có thể lập được bao nhiêu số có 5 chữ số, biết rằng trong số đó chữ số 3 xuất hiện 2 lần và các chữ số còn lại đều khác nhau.

Giải: Do chữ số 3 xuất hiện 2 lần nên ta xem như lập số có 5 chữ số từ các số: 1, 2, 3, 3, 4. Tuy nhiên, trong số đó có 2 chữ số 3 vậy nên có tất cả là $\dfrac{5!}{2!} = 60$ số cần tìm.

Tổng quát: Giả sử có một tập gồm n phần tử, trong đó có r nhóm:

⋄ Nhóm 1 có k_1 phần tử giống nhau.

⋄ Nhóm 2 có k_2 phần tử giống nhau.

....

⋄ Nhóm r có k_r phần tử giống nhau ($k_1 + k_2 + \cdots + k_r \leq n$).

Thì số cách sắp có thứ tự n phần tử đó sẽ là: $\dfrac{n!}{k_1! \cdot k_2! \cdots k_r!}$.

Ví dụ 1.12. Giả thiết như ở *Ví dụ 1.9*, nhưng ta sẽ sử dụng công thức hoán vị lặp để tính số cách chọn mật khẩu. Có tất cả 7 kí tự, trong đó 2 chữ B và 3 chữ C nên số cách chọn sẽ là: $\frac{7!}{2!3!} = 420$ cách.

1.4. BÀI TẬP CHƯƠNG 1

Bài 1: Từ các chữ số 0, 1, 2, 3, 4, 5 và 6 có thể lập thành bao nhiêu số có 5 chữ số:

a) Gồm các chữ số khác nhau?

b) Gồm các chữ số khác nhau và chia hết cho 2?

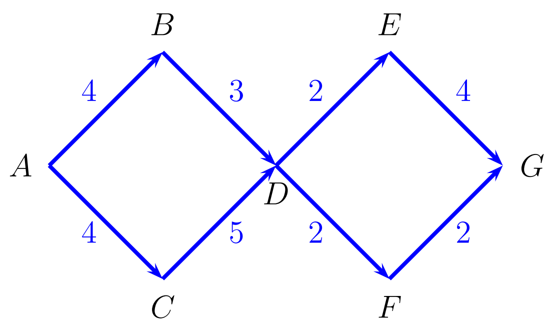
c) Gồm các chữ số khác nhau và chia hết cho 5?

d) Gồm các chữ số khác nhau và chia hết cho 9?

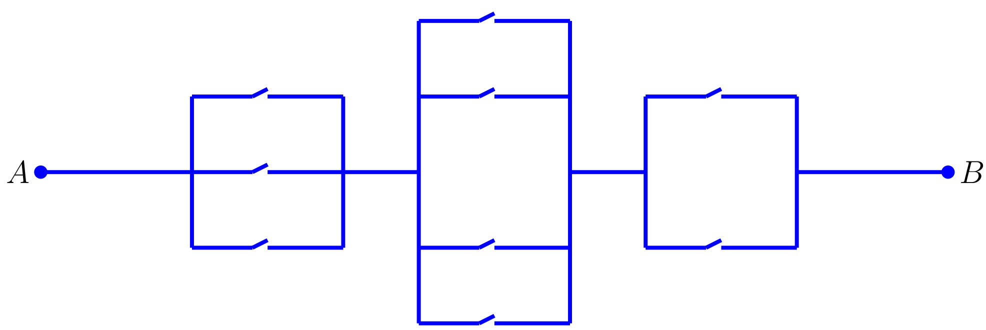
e) Gồm các chữ số có thể giống nhau?

f) Gồm các chữ số có thể giống nhau lớn hơn 43000?

Bài 2: Xét mạng đường nối các tỉnh A, B, C, D, E, F và G như hình vẽ. Trong đó, số viết trên mỗi cạnh cho biết số con đường nối hai tỉnh nằm ở hai đầu mút của cạnh. Hỏi rằng có bao nhiêu cách đi từ tỉnh A đến tỉnh G?



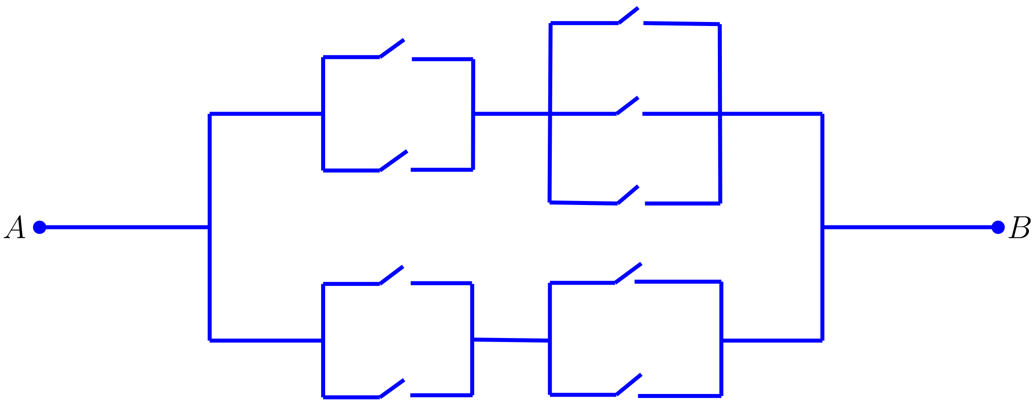
Bài 3: Có sơ đồ mạch điện như hình vẽ dưới đây với 9 công tắc khác nhau. Mỗi công tắc có hai trạng thái đóng-mở.



a) Hỏi có bao nhiêu cách đóng-mở 9 công tắc trên?

b) Hỏi có bao nhiêu cách đóng-mở 9 công tắc trên để dòng điện đi từ A đến được B?

Bài 4: Có sơ đồ mạch điện như hình vẽ dưới đây. Mỗi công tắc có hai trạng thái đóng-mở. Hỏi có bao nhiêu cách đóng-mở 9 công tắc để có dòng điện đi từ A đến B?



Bài 5: Một lớp học có 100 sinh viên. Hỏi rằng khi đó

a) Có bao nhiêu cách chọn 5 bạn vào Ban cán sự của lớp?

b) Có bao nhiêu cách chọn lớp trưởng, lớp phó học tập, lớp phó đời sống, thủ quỹ và bí thư Chi đoàn?

Bài 6: Một lô hàng có 12 sản phẩm trong số đó có 3 sản phẩm kém chất lượng. Lấy ra 2 sản phẩm theo 3 cách:

Cách 1: Lấy đồng thời ngẫu nhiên;

Cách 2: Lấy lần lượt không hoàn lại;

Cách 3: Lấy có hoàn lại.

Hỏi có bao nhiêu cách lấy:

a) 2 sản phẩm? b) 2 sản phẩm tốt? c) 1 sản phẩm tốt, 1 sản phẩm kém chất lượng?

Bài 7: Xếp chỗ ngồi ngẫu nhiên cho 7 sinh viên (3 nam, 4 nữ) vào một bàn dài 7 chỗ. Hỏi khi đó có bao nhiêu cách:

a) Xếp tùy ý?

b) Xếp sao cho ngồi 2 đầu bàn là 2 sinh viên nữ?

c) Xếp sao cho ngồi 2 đầu bàn là 1 nam và 1 nữ?

d) Xếp sao cho nam và nữ ngồi xen kẻ nhau?

Bài 8: Cần xếp 8 bình gốm và 8 lọ thủy tinh vào 4 thùng khác nhau, mỗi thùng chỉ chứa được 4 bình. Hỏi có bao nhiêu cách:

a) Xếp tùy ý?

b) Xếp sao cho thùng nào cũng chứa 2 bình gốm và 2 lọ thủy tinh?

Bài 9: Trong một buổi dạ vũ có 22 nam và 18 nữ. Hỏi có bao nhiêu cách chọn:

a) Hai người ra khiêu vũ?

b) Một nam và một nữ ra khiêu vũ?

c) Cùng lúc ba đôi nam nữ ra khiêu vũ?

Bài 10: Ở buổi liên hoan tổng kết chiến dịch Mùa hè xanh. Trong một nhóm bạn, mỗi người tham dự đều chuẩn bị những phần quà để tặng tất cả các bạn còn lại làm kỷ niệm. Người ta đếm được có tất cả 56 phần quà. Hỏi nhóm bạn đó có bao nhiêu người?

Bài 11: Người ta dùng 5 cột cờ được dựng thẳng hàng để báo hiệu trên biển. Biết rằng có tất cả 7 màu cờ khác nhau. Hỏi có bao nhiêu tín hiệu khác nhau nếu:

a) Dùng màu tùy ý?

b) Dùng 5 màu khác nhau?

c) Hai cột kế tiếp không được dùng cùng màu?

Bài 12: Giả sử 5 người cùng lên một đoàn tàu hỏa có 8 toa. Có bao nhiêu cách để:

a) Lên tùy ý?
b) Lên cùng một toa?
c) Lên 4 toa đầu?
d) Lên 5 toa khác nhau?
e) A, B lên cùng toa đầu?
f) A, B lên cùng một toa?
g) A, B lên cùng toa và không có ai khác trên toa này?

Bài 13: Có 30 câu hỏi khác nhau bao gồm 5 câu hỏi khó, 10 câu hỏi trung bình và 15 câu hỏi dễ. Từ 30 câu hỏi đó, có thể lập được bao nhiêu đề kiểm tra chứa 5 câu hỏi khác nhau? Biết rằng trong mỗi đề nhất thiết phải có đủ 3 loại câu hỏi (khó, trung bình, dễ) và số câu hỏi dễ không ít hơn 2.

Bài 14: Một công ty vận tải có 3 xe loại I (mỗi xe có thể chở được 3 kiện hàng) và 2 xe loại II (mỗi xe có thể chở được 2 kiện hàng). Công ty cùng lúc cần vận chuyển 13 kiện hàng khác nhau đến cùng một nơi. Hỏi có bao nhiêu cách sắp xếp xe-hàng tất cả?

Chương 2

CÁC KHÁI NIỆM CƠ BẢN CỦA XÁC SUẤT

2.1. BIẾN CỐ, QUAN HỆ GIỮA CÁC BIẾN CỐ

2.1.1. Thí nghiệm ngẫu nhiên

Định nghĩa xác suất được xây dựng trên cơ sở thí nghiệm (phép thử) ngẫu nhiên. Thí nghiệm ngẫu nhiên là thí nghiệm diễn ra hoàn toàn tự nhiên, không chịu sự tác động chủ quan nào từ bên ngoài và ta cũng không thể biết trước kết quả của nó. Qua đó giúp ta nghiên cứu khách quan một đối tượng hay một hiện tượng nào đó. Thí nghiệm ngẫu nhiên thường do một nhóm các điều kiện nào đó xác định, nhóm điều kiện này cần phải *rõ ràng*, *ổn định* trong suốt quá trình nghiên cứu và lặp lại bao nhiêu lần cũng được. Kết quả của thí nghiệm ngẫu nhiên được gọi là biến cố.

Ví dụ 2.1. Ta xét một vài ví dụ sau:

⋄ Gieo một con xúc xắc rồi kiểm tra số chấm xuất hiện là một thí nghiệm ngẫu nhiên. Kết quả số chấm xuất hiện là lẻ hay chẵn là biến cố.

⋄ Rút ngẫu nhiên một lá bài trong bộ bài là một thí nghiệm ngẫu nhiên. Kết quả lấy được quân át là một biến cố.

⋄ Ngắm bắn một viên đạn vào tấm bia là một thí nghiệm ngẫu nhiên, biến cố có thể xảy ra là bia bị trúng đạn hoặc bia không bị trúng đạn.

⋄ Tung một đồng xu là một thí nghiệm ngẫu nhiên, việc xuất hiện mặt sấp hay mặt ngửa là các biến cố.

2.1.2. Một số loại biến cố

a. Biến cố chắc chắn (Ω)***:*** Là biến cố luôn luôn xảy ra khi thực hiện phép thử. Chẳng hạn, khi gieo một con xúc xắc thì biến cố xuất hiện số chấm nhỏ hơn 7 là chắc chắn.

b. Biến cố không thể $(\emptyset)$***:*** Là biến cố không bao giờ xảy ra khi thực hiện phép thử. Chẳng hạn, khi gieo một con xúc xắc thì biến cố xuất hiện số chấm lớn hơn hoặc bằng 7 là biến cố không thể.

c. Biến cố ngẫu nhiên $(A, B, C, \ldots)$***:*** Là biến cố có thể xảy ra hoặc có thể không xảy ra khi thực hiện phép thử. Nó không là biến cố chắc chắn cũng không là biến cố không thể.
Chẳng hạn, khi tung một đồng xu thì biến cố xuất hiện mặt sấp là ngẫu nhiên vì nó có

thể xảy ra mà cũng có thể không xảy ra.

d. Biến cố kéo theo: Biến cố A gọi là kéo theo biến cố B nếu A xảy ra thì B cũng xảy ra. Ký hiệu: $A \Rightarrow B$ ($A \subset B$).
Chẳng hạn, gọi A là biến cố con xúc xắc xuất hiện mặt 3 chấm, B là biến cố con xúc xắc xuất hiện chấm lẻ thì $A \Rightarrow B$. Lưu ý một biến cố có thể có nhiều biến cố kéo theo nó.

e. Biến cố tương đương: Biến cố A gọi là tương đương với biến cố B nếu A xảy ra thì B cũng xảy ra và ngược lại. Có nghĩa A, B là hai biến cố kéo theo của nhau. Ký hiệu: $A \Leftrightarrow B$ ($A = B$).
Chẳng hạn, khi gieo một con xúc xắc gọi A là biến cố số chấm xuất hiện lớn hơn 5 và B là biến cố số chấm xuất hiện là 6 thì $A = B$.

f. Biến cố sơ cấp: Biến cố A gọi là biến cố sơ cấp nếu không có biến cố nào kéo theo nó ngoài chính biến cố đó. Ta có thể hình dung biến cố sơ cấp là những biến cố "nhỏ nhất", không có biến cố nào chứa trong nó nữa.
Chẳng hạn, khi gieo một con xúc xắc thì biến cố xuất hiện mặt 1 chấm, 2 chấm,... là các biến cố sơ cấp nhưng biến cố xuất hiện mặt chẵn, mặt lẻ lại không là biến cố sơ cấp.

g. Biến cố đồng khả năng: Là những biến cố có khả năng xuất hiện như nhau.
Chẳng hạn, khi gieo một đồng xu cân đối, đồng chất thì biến cố xuất hiện mặt sấp và biến cố xuất hiện mặt ngửa là hai biến cố đồng khả năng.

2.1.3. Phép toán giữa các biến cố

a. Biến cố tổng: Biến cố C được gọi là tổng của hai biến cố A và B nếu C xảy ra khi và chỉ khi A hoặc B xảy ra. Ký hiệu là $C = A + B$ hoặc $C = A \cup B$.
Chẳng hạn, hai xạ thủ cùng bắn vào một bia. Gọi A là biến cố người thứ nhất bắn trúng bia và B là biến cố người thứ hai bắn trúng bia. Nếu gọi C là biến cố bia bị trúng đạn thì khi đó $C = A + B$.

Chú ý: i) Định nghĩa tổng có thể mở rộng cho nhiều biến cố.

ii) Mọi biến cố ngẫu nhiên đều là tổng của một số biến cố sơ cấp nào đó. Đặc biệt, biến cố chắc chắn (Ω) là tổng của tất cả mọi biến cố sơ cấp có thể có nên còn gọi là không gian các biến cố sơ cấp hay không gian mẫu; biến cố không thể ($\emptyset$) không chứa bất kỳ một biến cố sơ cấp nào.

iii) Nếu $C = A + B$ thì $A \Rightarrow C$ và $B \Rightarrow C$.

b. Biến cố tích: Biến cố C gọi là tích của hai biến cố A và B nếu C xảy ra khi và chỉ khi cả A và B cùng xảy ra. Ký hiệu là $C = A \cdot B$ hoặc $C = A \cap B$.
Chẳng hạn, khi gieo một con xúc xắc. Gọi A là biến cố xuất hiện số chấm lẻ, B là biến cố xuất hiện số chấm lớn hơn 3. Nếu gọi C là biến cố xuất hiện mặt 5 chấm thì $C = A \cdot B$.

Chú ý: i) Định nghĩa tích có thể mở rộng cho nhiều biến cố.
ii) Nếu $C = A \cdot B$ thì $C \Rightarrow A$ và $C \Rightarrow B$.

c. Biến cố hiệu: Biến cố C gọi là hiệu của biến cố A với B nếu C xảy ra khi và chỉ khi A xảy ra nhưng B không xảy ra. Ký hiệu là $C = A - B$ hoặc $C = A \setminus B$.
Chẳng hạn, khi gieo một con xúc xắc. Gọi A là biến cố xuất hiện số chấm lẻ, B là biến cố xuất hiện mặt có số chấm nhỏ hơn 3 thì $A \setminus B = \{3, 5\}$.

Chú ý: Phép toán cộng và nhân có tính chất giao hoán nhưng hiệu thì không. Nghĩa là: $A \setminus B \neq B \setminus A$.

Tính chất 2.1. *Qua các khái niệm nêu ở trên ta thấy có sự tương đồng giữa phép toán biến cố với phép toán tập hợp. Cho A, B, C là các biến cố tùy ý, ta có các tính chất sau:*

- ⋄ *Giao hoán:* $A + B = B + A$, $A \cdot B = B \cdot A$.
- ⋄ *Kết hợp:* $A + (B + C) = (A + B) + C$, $(A \cdot B) \cdot C = A \cdot (B \cdot C)$.
- ⋄ *Phân phối:* $A \cdot (B + C) = AB + AC$, $A + B \cdot C = (A + B) \cdot (A + C)$.

2.1.4. Quan hệ giữa các biến cố

a. Biến cố độc lập: Hai biến cố A và B gọi là *độc lập* với nhau khi và chỉ khi biến cố này xảy ra hay không xảy ra không phụ thuộc vào biến cố kia.
Chẳng hạn, khi tung 2 đồng xu. Nếu gọi A là biến cố xuất hiện mặt sấp ở đồng xu thứ nhất, B là biến cố xuất hiện mặt sấp ở đồng xu thứ hai thì đây là hai biến cố độc lập với nhau.

b. Hai biến cố xung khắc: Hai biến cố A và B gọi là *xung khắc* với nhau nếu chúng không thể cùng xảy ra trong một phép thử. Nghĩa là: $A \cdot B = \emptyset$.
Chẳng hạn trong một lô hàng có 3 loại sản phẩm: Tốt, trung bình và xấu. Rút ngẫu nhiên một sản phẩm. Gọi A là biến cố lấy được sản phẩm tốt, B là biến cố lấy được sản phẩm trung bình. Khi đó A và B là 2 biến cố xung khắc.

c. Biến cố đối lập: Biến cố *không A* được gọi là *đối lập* của biến cố A và ký hiệu là $\overline{A}$. Vậy nên A và $\overline{A}$ là hai biến cố đối lập của nhau khi và chỉ khi A xảy ra thì $\overline{A}$ không xảy ra và ngược lại. Ta có: $A + \overline{A} = \Omega$, $A \cdot \overline{A} = \emptyset$.
Chẳng hạn khi tung một đồng xu, gọi A là biến cố xuất hiện mặt ngửa và B là biến cố xuất hiện mặt sấp. Khi đó, A và B là hai biến cố đối lập.

Nhận xét 2.1. Hai biến cố đối lập thì xung khắc nhưng ngược lại thì chưa chắc.

Tính chất 2.2. *i) Nếu $A \subset B$ thì $\overline{B} \subset \overline{A}$.* *ii)* $\overline{A + B} = \overline{A} \cdot \overline{B}$, $\overline{A \cdot B} = \overline{A} + \overline{B}$.

Ví dụ 2.2. Bắn vào bia 5 phát. Gọi A_i là biến cố bắn trúng bia ít nhất i phát, B_i là biến cố bắn trúng bia đúng i phát.

a) Diễn tả các biến cố: $\overline{A}_1, \overline{A}_2, \overline{B}_1, \overline{B}_2$.

b) Hai biến cố $\overline{A}_1$ và $\overline{B}_1$ có xung khắc với nhau không?

c) Diễn tả các biến cố: $A_1 + \overline{B}_2, B_1 + \overline{A}_2, \overline{A}_1 B_2, \overline{A}_2 B_1$.

Giải: Gọi Ω là không gian mẫu thì $\Omega = \{0, 1, 2, 3, 4, 5\}$.

a) Ta có: $A_1 = \{1, 2, 3, 4, 5\}$, $A_2 = \{2, 3, 4, 5\}$ nên $\overline{A}_1 = \{0\}$, $\overline{A}_2 = \{0, 1\}$.
Và $B_1 = \{1\}$, $B_2 = \{2\}$ nên $\overline{B}_1 = \{0, 2, 3, 4, 5\}$, $\overline{B}_2 = \{0, 1, 3, 4, 5\}$.

b) Ta thấy $\overline{A}_1 \overline{B}_1 = \{0\} \neq \emptyset$ nên $\overline{A}_1$ và $\overline{B}_1$ không xung khắc với nhau.

c) $A_1 + \overline{B}_2 = \{0, 1, 2, 3, 4, 5\} = \Omega, B_1 + \overline{A}_2 = \{0, 1\} = \overline{A}_2, \overline{A}_1 B_2 = \emptyset, \overline{A}_2 B_1 = B_1$.

Ví dụ 2.3. Kiểm tra 3 sản phẩm. Gọi A_k là biến cố sản phẩm thứ k tốt. Hãy trình bày các cách biểu diễn qua A_k các biến cố sau:

a) A: tất cả các sản phẩm đều không tốt.

b) B: có ít nhất một sản phẩm không tốt.

c) C: có ít nhất một sản phẩm tốt.

d) D: không phải tất cả sản phẩm đều tốt.

e) E: có đúng một sản phẩm không tốt.

f) F: có ít nhất 2 sản phẩm tốt.

Giải:

a) $A = \overline{A}_1 \cdot \overline{A}_2 \cdot \overline{A}_3 = \overline{A_1 + A_2 + A_3}$.

b) Ta có:

$$\begin{aligned} B &= \overline{A}_1 + \overline{A}_2 + \overline{A}_3 \\ &= \overline{A}_1 A_2 A_3 + A_1 \overline{A}_2 A_3 + A_1 A_2 \overline{A}_3 + \overline{A}_1 \overline{A}_2 A_3 + A_1 \overline{A}_2 \overline{A}_3 + \overline{A}_1 A_2 \overline{A}_3 + \overline{A}_1 \overline{A}_2 \overline{A}_3. \end{aligned}$$

c) $C = A_1 + A_2 + A_3 = A_1\overline{A}_2\overline{A}_3 + \overline{A}_1 A_2 \overline{A}_3 + \overline{A}_1\overline{A}_2 A_3 + A_1 A_2 \overline{A}_3 + A_1 \overline{A}_2 A_3 + \overline{A}_1 A_2 A_3 + A_1 A_2 A_3$.

d) $D = B$.

e) $E = A_1 A_2 \overline{A}_3 + A_1 \overline{A}_2 A_3 + \overline{A}_1 A_2 A_3$.

f) $F = A_1 A_2 \overline{A}_3 + A_1 \overline{A}_2 A_3 + \overline{A}_1 A_2 A_3 + A_1 A_2 A_3$.

Ví dụ 2.4. Quan sát 4 sinh viên làm bài thi. Ký hiệu A_j $(j = 1, 2, 3, 4)$ là biến cố sinh viên thứ j làm bài thi đạt yêu cầu. Hãy viết các biến cố sau đây:
a) Có đúng 1 sinh viên đạt yêu cầu. b) Có đúng 3 sinh viên đạt yêu cầu.
c) Có ít nhất 1 sinh viên đạt yêu cầu. d) Không có sinh viên nào đạt yêu cầu.

Giải:

a) Gọi A là biến cố có đúng 1 sinh viên đạt yêu cầu.
Ta có: $A = A_1\overline{A}_2\overline{A}_3\overline{A}_4 + \overline{A}_1 A_2 \overline{A}_3 \overline{A}_4 + \overline{A}_1 \overline{A}_2 A_3 \overline{A}_4 + \overline{A}_1 \overline{A}_2 \overline{A}_3 A_4$.

b) Gọi B là biến cố có đúng 3 sinh viên đạt yêu cầu.
Ta có: $B = \overline{A}_1 A_2 A_3 A_4 + A_1 \overline{A}_2 A_3 A_4 + A_1 A_2 \overline{A}_3 A_4 + A_1 A_2 A_3 \overline{A}_4$.

c) Gọi C là biến cố có ít nhất 1 sinh viên đạt yêu cầu. Ta có: $C = A_1 + A_2 + A_3 + A_4$.

d) Gọi D là biến cố không có sinh viên nào đạt yêu cầu. Ta có: $D = \overline{A}_1\overline{A}_2\overline{A}_3\overline{A}_4$.

2.2. XÁC SUẤT

2.2.1. Định nghĩa xác suất theo lối cổ điển

Định nghĩa 2.1. Nếu một phép thử có không gian các biến cố sơ cấp gồm n biến cố đồng khả năng. Trong đó có m biến cố sơ cấp đồng khả năng thuận lợi cho biến cố A. Khi đó, xác suất của biến cố A, ký hiệu $P(A)$ được cho bởi công thức:

$$P(A) = \frac{m}{n}.$$

Ví dụ 2.5. Phép thử là gieo một con xúc xắc. Tính xác suất xuất hiện số chấm lẻ?

Giải: Không gian các biến cố sơ cấp Ω gồm 6 trường hợp đồng khả năng. Đó là, có thể xuất hiện các mặt: 1 chấm, 2 chấm,..., 6 chấm. Nếu gọi A_i là biến cố xuất hiện mặt i chấm $(i = 1, \ldots, 6)$ thì $\Omega = \{A_1, A_2, A_3, A_4, A_5, A_6\}$.

Gọi A là biến cố xuất hiện mặt có số chấm lẻ thì có 3 trường hợp thuận lợi cho A là A_1, A_3, A_5 nên $\Omega_A = \{A_1, A_3, A_5\}$. Vậy nên $P(A) = \dfrac{|\Omega_A|}{|\Omega|} = \dfrac{3}{6} = \dfrac{1}{2}$.

Ví dụ 2.6. Gieo đồng thời 2 con xúc xắc. Tính xác suất để tổng số chấm xuất hiện của 2 con không quá 3.

Giải: Không gian các biến cố sơ cấp gồm 36 trường hợp đồng khả năng:

$$\begin{aligned}\Omega = \ &\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}.\end{aligned}$$

Gọi A là biến cố tổng số chấm xuất hiện không quá 3. Khi đó $\Omega_A = \{(1,1),(1,2),(2,1)\}$ gồm 3 biến cố thuận lợi cho A nên $P(A) = \dfrac{|\Omega_A|}{|\Omega|} = \dfrac{3}{36} = \dfrac{1}{12}$.

Ví dụ 2.7. Trong bình có 12 viên bi trong đó có 4 bi màu đỏ còn lại là bi màu trắng. Lấy ngẫu nhiên 5 bi. Tính xác suất để trong các bi lấy ra có 3 bi màu đỏ.

Giải: Lấy 5 trong 12 bi không kể thứ tự (chỉ quan tâm tổng số bi màu đỏ chứ không quan tâm bi thứ mấy màu đỏ). Nên số trường hợp đồng khả năng có thể xảy ra là: C_{12}^5.

Gọi A là biến cố trong 5 bi lấy ra có 3 bi đỏ, việc lấy bi chia thành 2 giai đoạn:

+ Lấy 3 bi đỏ trong 4 bi đỏ có C_4^3 cách.

+ Lấy 2 bi trắng trong 8 bi trắng có C_8^2 cách.

Vậy số kết quả thuận lợi cho A sẽ là: $C_4^3 \cdot C_8^2$ cách. Do đó: $P(A) = \dfrac{C_4^3 \cdot C_8^2}{C_{12}^5}$.

Nhận xét 2.2. Định nghĩa xác suất theo lối cổ điển gặp các hạn chế sau:

i) Chỉ xét được cho hệ có hữu hạn các biến cố sơ cấp.

ii) Việc xác định "đồng khả năng" không phải bao giờ cũng thực hiện được.

2.2.2. Định nghĩa xác suất theo thống kê

Định nghĩa 2.2. Giả sử phép thử được thực hiện n lần độc lập (nghĩa là kết quả của phép thử sau không phụ thuộc vào các phép thử trước đó) trong đó biến cố A xuất hiện m_A lần. Khi đó, m_A được gọi là tần số (tần số tuyệt đối) xuất hiện của biến cố A và tỉ số $\dfrac{m_A}{n} = f_n(A)$ được gọi là tần suất (tần số tương đối) của biến cố A.

Khi số lần thử n tăng lên mà tần suất luôn dao động quanh một số không đổi p và ngày càng gần với nó thì p được gọi là xác suất của biến cố A. Hay xác suất là giới hạn của tần suất khi số lần thử tăng mãi ra vô tận: $P(A) = \lim_{n\to\infty} f_n(A) = p$.

Trong thực tế khi số lần thử đủ lớn, ta lấy xác suất xấp xỉ bằng tần suất.

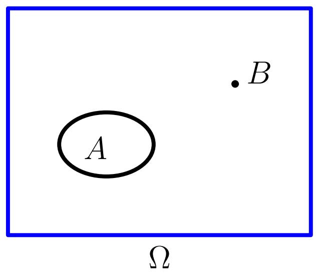
Phương pháp định nghĩa xác suất bằng thống kê được sử dụng trong thực tế khi liên quan đến số lượng lớn như: Tỉ lệ phế phẩm do máy sản suất; tỉ lệ bắn trúng bia của xạ thủ; tỉ lệ người mang nhóm máu A trong khu vực dân cư lớn;...

Ví dụ 2.8. Qua thống kê dân số, người ta tổng kết được xác suất để một em bé ra đời là trai hay gái xấp xỉ $0,5$.

2.2.3. Định nghĩa xác suất theo hình học

Định nghĩa 2.3. Giả sử có một miền Ω (là tờ giấy chẳng hạn). Ta định nghĩa:

- ⋄ Cả miền Ω xem như biến cố chắc chắn.
- ⋄ Một điểm B của miền xem như một biến cố sơ cấp.
- ⋄ Mỗi miền con A nào đó xem như biến cố ngẫu nhiên.
- ⋄ Tập rỗng (không chứa điểm nào) là biến cố không thể.



Hình 2.1: Định nghĩa xác suất theo hình học

Bây giờ ta ném vào miền Ω một chất điểm (hạt cát rất nhỏ) sao cho khả năng rơi vào bất kỳ chỗ nào trên Ω đều là như nhau. Rơi trúng biến cố nào xem như biến cố đó xảy ra. Khi đó, xác suất của một biến cố được định nghĩa như sau: $P(A) = \dfrac{S_A}{S_\Omega}$.

Nhận xét 2.3. *i)* Với B là biến cố sơ cấp (1 điểm) thì $P(B) = \dfrac{S_B}{S_\Omega} = \dfrac{0}{S_\Omega} = 0$.

Tuy nhiên khi ném chất điểm vẫn có thể trúng B, nghĩa là B vẫn có thể xảy ra. Nên biến cố có xác suất bằng 0 vẫn có thể xảy ra, gọi là biến cố "*hầu không thể*".

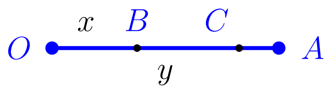
ii) Xét $\overline{B}$ là biến cố đối lập của B (miền Ω trừ điểm B), ta thấy:

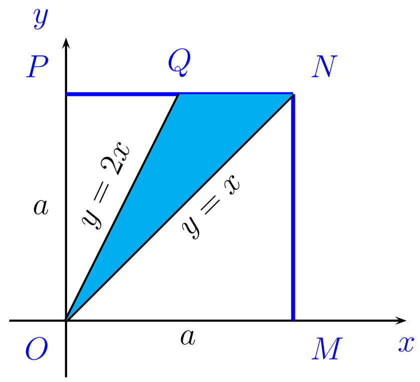
$$P(\overline{B}) = \frac{S_{\overline{B}}}{S_\Omega} = \frac{S_\Omega - S_B}{S_\Omega} = \frac{S_\Omega - 0}{S_\Omega} = 1.$$

Thế nhưng khi ném chất điểm có thể không rơi vào $\overline{B}$ mà vào B, nghĩa là biến cố có xác suất bằng 1 vẫn có thể không xảy ra, gọi là biến cố "*hầu chắc chắn*".

Ví dụ 2.9. Trên đoạn thẳng OA lấy ngẫu nhiên hai điểm B và C có tọa độ tương ứng $OB = x, OC = y$ $(x \leq y)$. Tìm xác suất để $BC < OB$.

Giải: Gọi X là biến cố "$BC < OB$".





Xét hệ trục tọa độ Oxy như hình vẽ, trong đó $OA = OM = OP = a$.

Khi đó: $0 \le x \le a, 0 \le y \le a$ và $x \le y$ là miền giới hạn bởi ΔONP nên không gian biến cố sơ cấp là $S_{\Delta ONP}$.

Do $BC < OB \Leftrightarrow OC - OB < OB \Leftrightarrow y - x < x \Leftrightarrow y < 2x$ nên không gian các biến cố thuận lợi cho X chính là $S_{\Delta OQN}$.

Suy ra:

$$P(X) = \frac{S_{\Delta OQN}}{S_{\Delta ONP}} = \frac{a^2}{4} : \frac{a^2}{2} = \frac{1}{2}.$$

2.2.4. Tính chất của xác suất

$i)$ Cho A là biến cố tùy ý, ta luôn có: $0 \le P(A) \le 1$.

$ii)$ Nếu $A \Rightarrow B$ thì $P(A) \le P(B)$. Do đó, nếu $A \Leftrightarrow B$ $(A = B)$ thì $P(A) = P(B)$.

$iii)$ Nếu $A = \Omega$ thì $P(A) = 1$ và nếu $A = \emptyset$ thì $P(A) = 0$.

2.2.5. Ý nghĩa của xác suất

Qua các định nghĩa ta thấy xác suất của một biến cố chính là một con số đo khả năng để biến cố đó xảy ra. Biến cố có xác suất càng lớn (càng gần 1) thì càng dễ xảy ra và ngược lại. Giả sử có hai cầu thủ cùng ném bóng vào rổ. Nếu xác suất ném vào rổ của cầu thủ A lớn hơn cầu thủ B thì cầu thủ A sẽ thường xuyên ném vào rổ hơn so với cầu thủ B.

Ta cũng chú ý rằng, xác suất của một biến cố còn có tính chất "trung bình". Nghĩa là, nếu cầu thủ A có xác suất ném bóng vào rổ là 70% thì không nhất thiết là cứ ném 10 quả bóng sẽ có 7 quả vào rổ mà chỉ có tính chất trung bình (trong nhiều loạt ném 10 quả thì trung bình mỗi loạt có 7 quả vào rổ).

Ngoài ra, xác suất cũng phụ thuộc vào nhóm điều kiện xác định phép thử. Khi nhóm điều kiện này thay đổi thì xác suất sẽ thay đổi theo.

2.3. CÁC CÔNG THỨC XÁC SUẤT

2.3.1. Công thức xác suất có điều kiện

Bài toán: Một nhà kho chứa n thùng hàng. Trong đó, có m thùng có dấu niêm phong hình vuông, l thùng có dấu niêm phong hình tròn, k thùng có dấu niêm phong hình vuông lẫn hình tròn. Xuất kho ngẫu nhiên 1 thùng, tính xác suất để thùng hàng xuất ra có dấu niêm phong:

a) Hình vuông. Hình tròn. Hình vuông lẫn hình tròn.

b) Hình vuông với điều kiện nó phải có hình tròn.

Giải: Gọi A là biến cố thùng hàng xuất kho có dấu niêm phong hình vuông và B là biến cố thùng hàng xuất kho có dấu niêm phong hình tròn. Khi đó, AB sẽ là biến cố thùng hàng xuất kho có dấu niêm phong hình vuông lẫn hình tròn.

a) Ta có $P(A) = \dfrac{m}{n}$, $P(B) = \dfrac{l}{n}$ và $P(AB) = \dfrac{k}{n}$.

b) Do bài toán yêu cầu tính xác suất để thùng hàng xuất ra có dấu niêm phong hình vuông với điều kiện là nó đã có dấu niêm phong hình tròn nên các kết quả thuận lợi cho biến cố này là số l thùng có dấu niêm phong hình tròn.

Trong số l thùng đó ta chọn ra những thùng có cả dấu niêm phong hình vuông, đó chính là số k thùng có dấu niêm phong hình vuông lẫn tròn.

Vậy nên xác suất cần tìm là: $\dfrac{k}{l} = \dfrac{k/n}{l/n} = \dfrac{P(AB)}{P(B)}$.

Định nghĩa 2.4. Xác suất của biến cố A được tính khi trong nhóm điều kiện xác định phép thử có bổ sung thêm một điều kiện đột xuất mới là biến cố B đã xảy ra được gọi là xác suất có điều kiện và ký hiệu là $P(A|B)$. Do đó qua bài toán trên, ta có:

$$P(A|B) = \frac{P(AB)}{P(B)} \qquad \big(P(B) \neq 0\big).$$

Ví dụ 2.10. Để làm luận văn tốt nghiệp, một lớp được chia thành 3 nhóm đi thực tập tại các công ty, xí nghiệp trong thành phố. Nhóm I có 30 sinh viên trong đó có 10 nữ, nhóm II có 25 sinh viên trong đó có 8 nữ và nhóm III có 25 sinh viên trong đó có 12 nữ. Chọn ngẫu nhiên trong lớp 1 sinh viên, tính xác suất để:

a) Sinh viên chọn được là nữ. Sinh viên chọn được thuộc nhóm II. Sinh viên chọn được là nữ trong nhóm II.

b) Biết sinh viên chọn được là nữ, tính xác suất để sinh viên đó thuộc nhóm II.

Giải:

a) Gọi A là biến cố sinh viên chọn được là nữ, B là biến cố sinh viên chọn được thuộc nhóm II. Khi đó AB sẽ là biến cố sinh viên chọn được là nữ trong nhóm II. Ta có:

$$P(A) = \frac{30}{80} = \frac{3}{8},\ P(B) = \frac{25}{80},\ P(AB) = \frac{8}{80} = \frac{1}{10}.$$

b) Theo trên có $B|A$ là biến cố sinh viên chọn được thuộc nhóm II với điều kiện là nữ, nên:

$$P(B|A) = \frac{P(BA)}{P(A)} = \frac{1}{10} : \frac{3}{8} = \frac{8}{30}.$$

2.3.2. Công thức nhân

a. Tính độc lập giữa các biến cố: *i*) Nếu hai biến cố A và B *độc lập* với nhau thì $P(A|B) = P(A), P(B|A) = P(B)$.

ii) Hệ các biến cố $\{A_1, A_2, A_3, \ldots, A_n\}$, $n > 2$ gọi là *độc lập từng đôi* nếu A_i và A_j độc lập với nhau ($\forall i \neq j$).

iii) Hệ các biến cố $\{A_1, A_2, A_3, \ldots, A_n\}$, $n > 2$ được gọi là *độc lập toàn phần* (hoàn toàn) nếu mỗi biến cố độc lập với tích của k biến cố còn lại $(1 \leq k \leq n-1)$.

b. Công thức nhân tổng quát: Cho $\{A_1, A_2, A_3, \ldots, A_n\}$ là hệ các biến cố tùy ý. Khi đó ta có công thức sau

$$P(A_1 \cdot A_2 \cdot A_3 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \cdots P(A_n|A_1A_2 \cdots A_{n-1}),$$

gọi là công thức nhân xác suất.

Chứng minh. Ta chứng minh công thức bằng phương pháp qui nạp.

▷ Với $n = 1$ ta có $P(A_1) = P(A_1)$ luôn đúng.

▷ Với $n = 2$ thì $P(A_1A_2) = P(A_1)P(A_2|A_1)$ theo công thức xác suất có điều kiện.

▷ Giả sử công thức đúng với $n = k$. Tức là:

$$P(A_1 \cdot A_2 \cdot A_3 \cdots A_k) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \cdots P(A_k|A_1A_2 \cdots A_{k-1}).$$

Ta cần chứng minh công thức đúng với $n = k + 1$. Thật vậy:

$$\begin{aligned} P(A_1 \cdot A_2 \cdots A_{k+1}) &= P\big((A_1 \cdot A_2 \cdots A_k)A_{k+1}\big) \\ &= P(A_1A_2 \cdots A_k)P(A_{k+1}|A_1A_2 \cdots A_k) \\ &= P(A_1)P(A_2|A_1) \cdots P(A_k|A_1A_2 \cdots A_{k-1})P(A_{k+1}|A_1A_2 \cdots A_k). \end{aligned}$$

Nên công thức được chứng minh. □

∗ Đặc biệt, nếu $\{A_1, A_2, A_3, \ldots, A_n\}$ *độc lập toàn phần* thì

$$P(A_1 \cdot A_2 \cdot A_3 \cdots A_n) = P(A_1)P(A_2)P(A_3) \cdots P(A_n).$$

c. Công thức nhân cho 2 hoặc 3 biến cố: Sau đây ta có công thức nhân phát biểu cho 2 và 3 biến cố.

⊛ ***Với 2 biến cố:*** Cho A và B là 2 biến cố tùy ý, ta có: $P(AB) = P(A)P(B|A)$.

∗ Đặc biệt, nếu A và B *độc lập* thì: $P(AB) = P(A)P(B)$.

Ví dụ 2.11. Trong một đợt rút thăm trúng thưởng dành cho khách hàng thân thiết của siêu thị X. Với 100 phiếu trong hộp, có 10 phiếu trúng thưởng. Tính xác suất để người thứ nhất và người thứ hai đều rút được phiếu trúng thưởng.

Giải: Gọi T_i là biến cố người thứ i rút được phiếu trúng thưởng ($i = 1, 2$). Thì T_1T_2 là biến cố cả 2 người đều rút được phiếu trúng thưởng.

Khi đó: $P(T_1T_2) = P(T_1)P(T_2|T_1) = \dfrac{10}{100} \cdot \dfrac{9}{99} = 0,009.$

Ví dụ 2.12. Tung 2 đồng xu, tính xác suất để chúng đều xuất hiện mặt ngửa.

Giải: Gọi A_i là biến cố đồng xu thứ i xuất hiện mặt ngửa ($i = 1, 2$). Khi đó A_1A_2 là biến cố để cả 2 đồng xu đều xuất hiện mặt ngửa.

Ta có: $P(A_1A_2) = P(A_1)P(A_2) = \dfrac{1}{2} \cdot \dfrac{1}{2} = 0,25.$
(Do việc xuất hiện mặt sấp hay ngửa ở đồng xu thứ 1, không ảnh hưởng đến kết quả xuất hiện mặt ngửa ở đồng xu thứ 2)

⊛ ***Với 3 biến cố:*** Cho A, B, C là 3 biến cố tùy ý: $P(ABC) = P(A)P(B|A)P(C|AB)$.

∗ Đặc biệt, nếu A, B và C *độc lập toàn phần* thì: $P(ABC) = P(A)P(B)P(C)$.

Ví dụ 2.13. Cũng từ *Ví dụ 2.11* nêu trên, hãy tính xác suất để 3 người đầu tiên đều không rút được phiếu trúng thưởng.

Giải: Với cách gọi như ở *Ví dụ 2.11* thì $\overline{T}_1\overline{T}_2\overline{T}_3$ là biến cố 3 người đầu tiên đều không rút được phiếu trúng thưởng.

Ta có: $P(\overline{T}_1\overline{T}_2\overline{T}_3) = P(\overline{T}_1)P(\overline{T}_2|\overline{T}_1)P(\overline{T}_3|\overline{T}_1\overline{T}_2) = \dfrac{90}{100} \cdot \dfrac{89}{99} \cdot \dfrac{88}{98} = 0,726.$

2.3.3. Công thức cộng

a. Quan hệ giữa các biến cố xung khắc: i) Hai biến cố A và B được gọi là *xung khắc* với nhau nếu $A \cdot B = \emptyset$.

ii) Hệ các biến cố $\{A_1, A_2, A_3, \ldots\}$ gọi là *xung khắc từng đôi* nếu $A_iA_j = \emptyset, \forall i \neq j$.

b. Công thức cộng tổng quát: Cho $\{A_1, A_2, A_3, \ldots, A_n\}$ là hệ các biến cố tùy ý. Khi đó ta có công thức sau:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1\le i<j\le n} P(A_iA_j) + \sum_{1\le i<j<k\le n} P(A_iA_jA_k) - \cdots + (-1)^{n-1}P(A_1\cdots A_n),$$

gọi là công thức cộng xác suất.

Chứng minh. Ta chứng minh công thức bằng phương pháp qui nạp.

▷ Với $n = 1$ thì $P(A_1) = P(A_1)$ hiển nhiên đúng.

▷ Với $n = 2$, ta có $P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1A_2)$. Thật vậy:

Giả sử số trường hợp có thể của phép thử là n, số trường hợp thuận lợi cho biến cố A_1, A_2 và A_1A_2 lần lượt là n_{A_1}, n_{A_2} và $n_{A_1A_2}$. Suy ra số trường hợp thuận lợi cho biến cố

$A_1 + A_2$ sẽ là $n_{A_1} + n_{A_2} - n_{A_1A_2}$. Vậy nên:

$$P(A_1 + A_2) = \frac{n_{A_1} + n_{A_2} - n_{A_1A_2}}{n} = \frac{n_{A_1}}{n} + \frac{n_{A_2}}{n} - \frac{n_{A_1A_2}}{n} = P(A_1) + P(A_2) - P(A_1A_2).$$

▷ Giả sử công thức đúng với $n = l$. Nghĩa là:

$$P\left(\sum_{i=1}^{l} A_i\right) = \sum_{i=1}^{l} P(A_i) - \sum_{1\le i<j\le l} P(A_iA_j) + \sum_{1\le i<j<k\le l} P(A_iA_jA_k) - \cdots + (-1)^{l-1}P(A_1\cdots A_l).$$

Ta chứng minh công thức đúng với $n = l + 1$. Thật vậy:

$$\begin{aligned}
&P\left(\sum_{i=1}^{l+1} A_i\right) = P\left(\sum_{i=1}^{l} A_i + A_{l+1}\right)\\
&= P\left(\sum_{i=1}^{l} A_i\right) + P(A_{l+1}) - P\left(\left(\sum_{i=1}^{l} A_i\right) A_{l+1}\right)\\
&= \left(\sum_{i=1}^{l} P(A_i) - \sum_{1\le i<j\le l} P(A_iA_j) + \cdots + (-1)^{l-1}P(A_1\cdots A_l)\right) + P(A_{l+1})\\
&\quad -P\left(\sum_{i=1}^{l} A_iA_{l+1}\right)\\
&= \sum_{i=1}^{l+1} P(A_i) - \sum_{1\le i<j\le l} P(A_iA_j) + \cdots + (-1)^{l-1}P(A_1\cdots A_l)\\
&\quad - \left(\sum_{i=1}^{l} P(A_iA_{l+1}) - \sum_{1\le i<j\le l} P(A_iA_jA_{l+1}) + \cdots + (-1)^{l-1}P(A_1\cdots A_lA_{l+1})\right)\\
&= \sum_{i=1}^{l+1} P(A_i) - \sum_{1\le i<j\le l+1} P(A_iA_j) + \cdots + (-1)^{l}P(A_1\cdots A_{l+1}).
\end{aligned}$$

Nên công thức được chứng minh. □

c. ***Công thức cộng cho 2 hoặc 3 biến cố:*** Ta có công thức cho 2 và 3 biến cố

⊛ ***Với 2 biến cố:*** Cho A và B là hai biến cố tùy ý: $P(A+B) = P(A)+P(B)-P(AB)$.

∗ Đặc biệt, nếu A và B *xung khắc* thì $P(A + B) = P(A) + P(B)$.

⊛ ***Với 3 biến cố:*** Cho A, B và C là 3 biến cố tùy ý, có:

$$P(A + B + C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(CA) + P(ABC).$$

∗ Đặc biệt, nếu A, B và C *xung khắc từng đôi* thì: $P(A+B+C) = P(A)+P(B)+P(C)$.

Ví dụ 2.14. Gieo đồng thời 2 con xúc xắc, hãy tính xác suất để:

a) Có ít nhất 1 con xuất hiện mặt 3 chấm.

b) Có đúng 1 con xuất hiện mặt 3 chấm.

<u>Giải:</u> Gọi A_i là biến cố con xúc xắc thứ i xuất hiện mặt 3 chấm.

a) Khi đó $A_1 + A_2$ là biến cố ít nhất 1 con xuất hiện mặt 3 chấm. Ta có:

$$P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1A_2) = \frac{1}{6} + \frac{1}{6} - \frac{1}{6}\cdot\frac{1}{6} = \frac{11}{36}.$$

b) Biến cố có 1 con xuất hiện mặt 3 chấm là $\overline{A}_1A_2 + A_1\overline{A}_2$. Do đó:

$$P(\overline{A}_1A_2 + A_1\overline{A}_2) = P(\overline{A}_1A_2) + P(A_1\overline{A}_2) = P(\overline{A}_1)P(A_2) + P(A_1)P(\overline{A}_2) = \frac{5}{6}\cdot\frac{1}{6} + \frac{1}{6}\cdot\frac{5}{6} = \frac{5}{18}.$$

Ví dụ 2.15. Sau kỳ thi tuyển sinh đại học tại một trường THPT có 200 em dự thi. Trong đó có 50 em thi đậu khối A, 40 em thi đậu khối B và 20 em thi đậu cả hai khối. Chọn ngẫu nhiên 1 học sinh trong số đó, tính xác suất học sinh vừa chọn được khen thưởng. Biết rằng để được thưởng học sinh đó phải thi đậu.

Giải: Gọi A là biến cố em được chọn thi đậu khối A, B là biến cố em được chọn thi đậu khối B và C là biến cố em đã chọn được khen thưởng. Ta có:

$$P(C) = P(A+B) = P(A) + P(B) - P(AB) = \frac{50}{200} + \frac{40}{200} - \frac{20}{200} = 0,35.$$

Ví dụ 2.16. Hai người cùng bắn vào mục tiêu một cách độc lập với nhau. Xác suất trúng đích của chiến sĩ A là $0,9$ còn của chiến sĩ B là $0,8$. Tìm khả năng xảy ra các tình huống sau:

a) Hai người cùng bắn trúng đích khi mỗi người bắn 1 phát.

b) Có một người bắn trúng đích khi mỗi người bắn 1 phát.

c) Chiến sĩ A bắn trúng đích ngay trong 3 phát đầu.

Giải: Gọi A_i và B_i lần lượt là biến cố chiến sĩ A và B bắn trúng đích ở phát thứ i.

a) Khi đó A_1B_1 là biến cố cả hai chiến sĩ cùng bắn trúng đích trong 1 phát.

$$P(A_1B_1) \overset{\text{ĐL}}{=} P(A_1)P(B_1) = 0,9 \cdot 0,8 = 0,72.$$

b) Nếu gọi C là biến cố có một người bắn trúng đích thì $C = \overline{A}_1B_1 + A_1\overline{B}_1$.

$$P(C) = P(\overline{A}_1B_1 + A_1\overline{B}_1) = P(\overline{A}_1B_1) + P(A_1\overline{B}_1) = 0,1 \cdot 0,8 + 0,9 \cdot 0,2 = 0,26.$$

c) Gọi D là biến cố chiến sĩ A bắn trúng đích ngay trong 3 phát đầu thì $D = A_1 + A_2 + A_3$.

$$\begin{aligned} P(D) &= P(A_1 + A_2 + A_3) \\ &= P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_2A_3) - P(A_3A_1) + P(A_1A_2A_3) \\ &= 3 \cdot 0,9 - 3 \cdot 0,9^2 + 0,9^3 = 0,999. \end{aligned}$$

Nhận xét 2.4. Giả sử $\overline{A}$ là biến cố đối lập của A. Khi đó, ta có công thức tính xác suất của biến cố đối lập: $P(\overline{A}) = 1 - P(A)$.

Ví dụ 2.17. Với câu c) ở *Ví dụ 2.16*, khi dùng công thức xác suất của biến cố đối lập, ta có thể tính ngắn gọn hơn như sau:

Vì $\overline{D} = \overline{A}_1 \cdot \overline{A}_2 \cdot \overline{A}_3$ nên $P(D) = 1 - P(\overline{D}) = 1 - 0,1 \cdot 0,1 \cdot 0,1 = 1 - 0,001 = 0,999$.

Ví dụ 2.18. Lô hàng có 100 sản phẩm, trong đó có 10 sản phẩm khuyết tật. Lấy ra đồng thời 5 sản phẩm, tính xác suất để trong những sản phẩm lấy ra có ít nhất 1 sản phẩm khuyết tật.

Giải: Gọi A_i là biến cố lấy ra được i sản phẩm khuyết tật $(i = 0, 1, \ldots, 5)$.

$*$ Cách 1: Khi đó $A_1 + A_2 + A_3 + A_4 + A_5$ là biến cố lấy ra được 1; 2; 3; 4 hoặc 5 sản phẩm khuyết tật. Vậy nên xác suất để lấy ra được ít nhất 1 sản phẩm khuyết tật sẽ là:

$$\begin{aligned} P(A_1 + A_2 + A_3 + A_4 + A_5) &= P(A_1) + P(A_2) + P(A_3) + P(A_4) + P(A_5) \\ &= \frac{C_{10}^1 \cdot C_{90}^4}{C_{100}^5} + \frac{C_{10}^2 \cdot C_{90}^3}{C_{100}^5} + \frac{C_{10}^3 \cdot C_{90}^2}{C_{100}^5} + \frac{C_{10}^4 \cdot C_{90}^1}{C_{100}^5} + \frac{C_{10}^5 \cdot C_{90}^0}{C_{100}^5} \\ &= 0,416. \end{aligned}$$

$*$ *Cách 2:* Gọi A là biến cố trong 5 sản phẩm lấy ra có ít nhất 1 sản phẩm khuyết tật thì $\overline{A} = A_0$ là biến cố trong 5 sản phẩm lấy ra không có sản phẩm khuyết tật nào.

Ta có: $P(\overline{A}) = \dfrac{C_{90}^5}{C_{100}^5} = 0,584$. Do đó: $P(A) = 1 - P(\overline{A}) = 1 - 0,584 = 0,416$.

2.3.4. Công thức xác suất đầy đủ

a. Hệ đầy đủ:

Định nghĩa 2.5. Hệ các biến cố $\{A_1, A_2, A_3, \ldots\}$ được gọi là *đầy đủ* và *xung khắc từng đôi* nếu sau khi thực hiện phép thử bắt buộc có một và chỉ một biến cố được xảy ra. Trong đó:

$\diamond$ Hệ gọi là *đầy đủ* nếu $A_1 + A_2 + A_3 + \cdots = \Omega$.

$\diamond$ Hệ gọi là *xung khắc từng đôi (đôi một)* nếu $A_i \cdot A_j = \emptyset$ với $\forall i \neq j$.

Ví dụ 2.19. Trong một bình có 8 viên bi trong đó có 3 bi đỏ, lấy ngẫu nhiên 2 bi. Nếu ta gọi A_i $(i = 0, 1, 2)$ là biến cố lấy ra được i bi đỏ thì $\{A_0, A_1, A_2\}$ là hệ đầy đủ và xung khắc từng đôi. Thật vậy: $A_0 + A_1 + A_2 = \Omega$ và $A_0A_1 = \emptyset, A_1A_2 = \emptyset, A_2A_0 = \emptyset$.

Nghĩa là, lấy 2 bi thì nhất định phải có hoặc 0, hoặc 1, hoặc 2 bi đỏ và đã xảy ra trường hợp này thì không thể xảy ra trường hợp kia.

b. Công thức: Giả sử, hệ các biến cố $\{A_1, A_2, A_3, \ldots, A_n\}$ là *đầy đủ* và *xung khắc từng đôi*. B là một biến cố bất kỳ có thể xảy ra trong phép thử. Khi đó, ta có:

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) + \cdots + P(A_n)P(B|A_n) \\ &= \textstyle\sum_{i=1}^n P(A_i)P(B|A_i), \end{aligned}$$

gọi là công thức xác suất đầy đủ.

Chứng minh. Vì $\{A_i\}_{i=1,\ldots,n}$ là hệ đầy đủ nên:

$$B = A_1B + A_2B + A_3B + \cdots + A_nB.$$

Mặt khác, do $\{A_i\}_{i=1,2,\ldots,n}$ xung khắc từng đôi nên $\{A_iB\}_{i=1,\ldots,n}$ cũng xung khác từng đôi. Áp dụng công thức cộng xác suất, ta được:

$$\begin{aligned} P(B) &= P(A_1B + A_2B + A_3B + \cdots + A_nB) \\ &= P(A_1B) + P(A_2B) + P(A_3B) + \cdots + P(A_nB) \\ &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) + \cdots + P(A_n)P(B|A_n) \\ &= \textstyle\sum_{i=1}^n P(A_i)P(B|A_i). \end{aligned}$$

□

Ví dụ 2.20. Giả sử mặt hàng mũ bảo hiểm ở một thành phố do 3 hãng H_1, H_2 và H_3 sản xuất. Trong đó tỉ lệ phần trăm sản phẩm của mỗi hãng trên thị trường lần lượt là $60\%, 10\%$ và 30% với tỉ lệ sản phẩm bị khuyết tật tương ứng là $1\%, 5\%$ và 3%. Khi mua ngẫu nhiên một chiếc mũ bảo hiểm trong thành phố, hãy tính xác suất mua phải mũ bị khuyết tật.

Giải: Gọi A_i $(i = 1, 2, 3)$ theo thứ tự là biến cố mua được mũ của hãng H_1, H_2 và H_3; B là biến cố mua phải mũ bị khuyết tật. Do các $\{A_i\}_{i=1,2,3}$ là hệ đầy đủ nên theo công thức xác suất đầy đủ, ta có:

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) \\ &= 0,6 \cdot 0,01 + 0,1 \cdot 0,05 + 0,3 \cdot 0,03 = 0,02 = 2\%. \end{aligned}$$

2.3.5. Công thức Bayes

Cũng với giả thiết như ở trên, ta có công thức sau gọi là công thức Bayes:

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

Chứng minh. Từ công thức xác suất có điều kiện và công thức xác suất đầy đủ, ta có:

$$P(A_k|B) = \frac{P(A_kB)}{P(B)} = \frac{P(A_k)P(B|A_k)}{P(B)}.$$

□

Ví dụ 2.21. Từ *Ví dụ 2.20*, giả sử đã mua phải mũ bảo hiểm bị khuyết tật. Hãy tính xác suất để mũ đó do hãng H_3 sản xuất.

Giải: Theo công thức xác suất Bayes thì xác suất để mua phải mũ do hãng H_3 sản xuất khi biết mũ đó bị khuyết tật là:

$$P(A_3|B) = \frac{P(A_3B)}{P(B)} = \frac{P(A_3)P(B|A_3)}{P(B)} = \frac{0,3 \cdot 0,03}{0,02} = 45\%.$$

2.3.6. Công thức Bernoulli

a. Dãy phép thử Bernoulli:

Định nghĩa 2.6. Dãy phép thử Bernoulli là dãy n phép thử thỏa mãn 2 điều kiện sau:

i) Các phép thử trong dãy độc lập với nhau. Nghĩa là, kết quả của phép thử sau không phụ thuộc vào các phép thử trước đó.

ii) Trong mỗi phép thử chỉ có hai biến cố A hoặc $\overline{A}$ xảy ra; Xác suất để biến cố A xảy ra trong mỗi phép thử của dãy là như nhau và $P(A) = p$ với $0 < p < 1$.

Ví dụ 2.22. Một hộp có 12 viên bi, trong đó có 4 bi màu xanh còn lại là bi màu trắng. Lần lượt rút có hoàn lại 7 bi. Gọi A là biến cố lấy được bi xanh trong mỗi lần rút.

Lúc đó ta có dãy thử Bernoulli với $n = 7, p = P(A) = \frac{4}{12} = \frac{1}{3}$.

b. Công thức: Xác suất để trong n phép thử, biến cố A xảy ra k lần với xác suất mỗi lần A xảy ra là p, được ký hiệu là $B(k, n, p)$ và cho bởi công thức sau đây:

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

gọi là công thức Bernoulli.

Ví dụ 2.23. Với giả thiết như ở *Ví dụ 2.22.* Ta tính được xác suất để trong 7 lần rút bi có 3 bi xanh là:

$$B(3, 7, 1/3) = C_7^3 \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^4 = 0,256.$$

2.4. BÀI TẬP CHƯƠNG 2

Bài 1: Xếp ngẫu nhiên thành hàng ngang 7 chữ cái từ các chữ S, X, X, U, U, Á, Ấ, C, C, C, X, T, T, S. Tìm xác suất của biến cố xếp được chữ XÁCSUẤT.

Bài 2: Bộ bài tiến lên có 52 lá, trong đó có 12 lá tây (J, Q, K). Lấy ngẫu nhiên 7 lá. Tính xác suất để trong đó có:

a) 2 hoặc 3 lá tây. b) Ít nhất một lá tây. c) Nhiều nhất 5 lá tây.

Bài 3: Phòng Kế toán của một cơ quan có 13 người, trong đó có 7 nữ. Trong số 10 người có nhà ở gần cơ quan thì có 4 nam. Theo quy định của cơ quan thì người nào là nam hoặc có nhà ở gần cơ quan thì phải trực Tết. Chọn ngẫu nhiên một người trong phòng:

a) Tính xác suất người đó phải trực Tết.

b) Biết người đó phải trực Tết. Tính xác suất để đó là nữ.

Bài 4: Một hộp bóng đèn có 15 bóng, trong đó có 3 bóng hỏng. Tính xác suất một người mua hộp bóng đèn nếu lấy đồng thời 3 bóng để kiểm tra mà:

a) Có bóng hỏng thì không mua hộp bóng đèn.

b) Số bóng hỏng nhiều hơn thì không mua hộp bóng đèn.

Bài 5: Có 4 xạ thủ cùng bắn độc lập vào một mục tiêu. Xác suất để xạ thủ thứ nhất bắn trúng là $0,6$; xạ thủ thứ hai bắn trúng $0,8$; xạ thủ thứ ba bắn là $0,9$ và xạ thủ thứ tư bắn trúng là $0,7$. Biết rằng mỗi xạ thủ chỉ bắn một viên, hãy tính xác suất để:

a) Cả 4 xạ thủ đều bắn trúng mục tiêu. b) Có đúng 1 xạ thủ bắn trượt mục tiêu.

c) Có đúng 2 xạ thủ bắn trượt mục tiêu. d) Ít nhất một xạ thủ bắn trúng mục tiêu.

Bài 6: Một hộp chứa 10 phiếu trong đó có 1 phiếu trúng thưởng. Có 10 người lần lượt rút ngẫu nhiên mỗi người 1 phiếu, hỏi người rút ở lần thứ mấy có khả năng trúng thưởng cao nhất?

Bài 7: Một hộp chứa 10 phiếu trong đó có 2 phiếu trúng thưởng. Có 10 người lần lượt lấy ngẫu nhiên mỗi người 1 phiếu.

a) Tính xác suất để người thứ 3 lấy được phiếu trúng thưởng.

b) Biết người thứ 3 lấy được phiếu trúng thưởng. Tính xác suất để trong hai người đầu có một người lấy được phiếu trúng thưởng.

Bài 8: Một hộp có 10 bi trong đó có 2 bi đỏ. Lấy ngẫu nhiên lần lượt không hoàn lại từng bi cho đến khi được 2 bi đỏ thì dừng. Tính xác suất việc lấy bi dừng ở lần thứ 3.

Bài 9: Phải gieo ít nhất bao nhiêu con xúc xắc để xác suất có ít nhất một con xuất hiện mặt 1 chấm lớn hơn hay bằng $0,99$.

Bài 10: Bỏ ngẫu nhiên 5 lá thư vào 5 phong bì đã ghi sẵn địa chỉ. Tính xác suất để có ít nhất 1 lá thư đến đúng người nhận.

Bài 11: Một sinh viên phải thi học phần Xác suất thống kê (XSTK) trước rồi đến học phần Giải tích (GT). Biết rằng xác suất thi đậu XSTK là $0,8$ và nếu trước đó đã thi đậu XSTK thì xác suất thi đậu GT sẽ là $0,9$; Còn nếu trước đó mà rớt XSTK thì xác suất thi đậu GT sẽ là $0,6$. Hãy tính xác suất để sinh viên này:

a) Thi đậu cả hai học phần.

b) Thi đậu ít nhất một học phần.

c) Thi đậu đúng một học phần.

Bài 12: Một hộp ban đầu chứa một quả cầu trắng và một quả cầu đen. Người ta rút liên tiếp các quả cầu theo nguyên tắc sau: Nếu rút được quả cầu trắng thì người ta lại trả nó cùng với hai quả cầu trắng khác. Nếu rút được quả cầu đen, người ta lại trả nó cùng với một quả cầu đen khác. Gọi A_i là biến cố quả cầu thứ i được rút là quả cầu trắng.

a) Tính $P(A_1)$, $P(A_2)$ và $P(A_3)$.

b) Biết lần thứ 2 rút được quả cầu trắng, tính xác suất để lần thứ 3 cũng rút được quả cầu trắng.

Bài 13: Bóng đèn bán ở thị trường do 3 công ty sản xuất. Tổng số bóng đèn của công ty I chiếm 30%, công ty II chiếm 50%, công ty III chiếm 20% với tỉ lệ hỏng tương ứng là 3%, 2% và 5%. Mua ngẫu nhiên một bóng đèn.

a) Tính xác suất để bóng đèn đó tốt.

b) Giả sử mua được bóng tốt, tính xác suất để bóng đèn đó do công ty III sản xuất.

Bài 14: Một thiết bị gồm 3 loại linh kiện: Loại thứ I chiếm 24% tổng số các linh kiện, loại thứ II chiếm 47% và loại thứ III chiếm 29%. Cho biết khả năng hỏng (tại thời điểm đang xét) của các linh kiện loại thứ I là 22%, loại thứ II là 23% và loại thứ III là 12%.

a) Tính khả năng thiết bị sẽ bị hỏng.

b) Nếu bị hỏng thì loại linh kiện nào có khả năng hỏng lớn nhất?

Bài 15: Nhà kho chứa 30 lô hàng, trong đó có 18 lô loại một (chứa 70% sản phẩm loại A) còn lại là lô loại hai (chứa 40% sản phẩm loại A). Chọn ngẫu nhiên 1 lô hàng và từ lô đó lấy ra 1 sản phẩm.

a) Hãy tính xác suất để sản phẩm đó là loại A.

b) Giả sử sản phẩm đó loại A. Theo bạn khả năng sản phẩm đó thuộc lô loại nào cao hơn?

Bài 16: Có hai thùng mật ong, mỗi thùng có 10 chai. Thùng thứ nhất có 4 chai loại một, thùng thứ hai có 3 chai loại một. Chọn ngẫu nhiên một thùng:

a) Từ thùng đó lấy ra 3 chai. Tính xác suất để trong 3 chai lấy ra có 1 chai loại một.

b) Từ thùng đó lấy ra 1 chai thì được chai loại một. Tính xác suất để lấy thêm 1 chai nữa trong thùng đó cũng là chai loại một.

Bài 17: Với giả thiết như ở *Bài tập 15*. Giả sử từ thùng thứ nhất lấy ra một chai bỏ sang thùng thứ hai, rồi từ thùng thứ hai lấy ra một chai.

a) Hãy tính xác suất để chai lấy ra ở thùng thứ hai là loại một.

b) Nếu chai lấy ra ở thùng hai không là loại một thì có kết luận gì về phẩm chất của chai đã bỏ vào thùng hai?

Bài 18: Có 2 lô chi tiết: Lô I gồm 12 chiếc và lô II gồm 10 chiếc, mỗi lô đều có 3 chiếc bị lỗi.

a) Chọn ngẫu nhiên một lô, rồi lấy ra 6 chi tiết. Tính xác suất để có 2 chi tiết bị lỗi.

b) Rút ngẫu nhiên 1 chi tiết từ lô I trộn vào lô II. Từ lô II rút ngẫu nhiên 1 chi tiết.

+ Tính xác suất để chi tiết đó bị lỗi.

+ Giả sử đó là chi tiết bị lỗi theo bạn khả năng chi tiết bỏ từ lô I sang lô II không bị lỗi là bao nhiêu phần trăm?

Bài 19: Một mạch điện gồm hai bộ phận mắc nối tiếp, với khả năng làm việc tốt trong một khoảng thời gian nào đó của mỗi bộ phận là $0,94$ và $0,97$. Ở một thời điểm trong khoảng thời gian đề cập ở trên người ta nhận thấy mạch điện ngưng làm việc (do bộ phận nào đó bị hỏng). Tính khả năng bộ phận thứ hai hỏng.

Bài 20: Một bài thi trắc nghiệm khách quan có 50 câu hỏi (đạt 10 điểm nếu làm đúng tất cả). Mỗi câu hỏi có 4 phương án trả lời, trong đó chỉ có 1 phương án đúng và các phương án có khả năng đúng là như nhau. Chọn ngẫu nhiên phương án trả lời cho tất cả các câu hỏi. Tính xác suất để bài làm đó được 5 điểm.

Bài 21: Một máy sản xuất lần lượt từng sản phẩm. Xác suất máy sản xuất ra phế phẩm là $0,08$. Tính xác suất để:

a) Trong 10 sản phẩm do máy sản xuất ra có 3 phế phẩm.

b) Trong 10 sản phẩm do máy sản xuất ra có phế phẩm.

c) Cần kiểm tra tối thiểu bao nhiêu sản phẩm do máy sản xuất ra để xác suất có phế phẩm lớn hơn hoặc bằng 90%?

Bài 22: Có 20 hộp sản phẩm. Mỗi hộp có 8 sản phẩm loại A và 2 sản phẩm loại B. Lấy ngẫu nhiên mỗi hộp 1 sản phẩm để kiểm tra. Tính xác suất trong số sản phẩm lấy ra có 3 sản phẩm loại B.

Chương 3

ĐẠI LƯỢNG NGẪU NHIÊN

3.1. ĐẠI LƯỢNG NGẪU NHIÊN MỘT CHIỀU

Ở Chương 2 chúng ta đã gặp khái niệm về biến cố ngẫu nhiên, chúng có thể được biểu hiện về mặt định tính: Bóng đèn sáng hay không sáng; viên bi màu xanh hay màu đỏ; đồng xu sấp hay ngửa;... và cũng có thể được biểu hiện về mặt định lượng như: Số chấm xuất hiện khi gieo con xúc xắc; số sản phẩm tốt có trong các sản phẩm lấy ra;...

Trong chương này ta chỉ xét những *"biến cố"* biểu hiện thành con số, tức là về mặt số lượng như là: Số khuyết tật của một sản phẩm vừa được đúc ra; số người được sinh ra trong một năm ở một địa phương nào đó;... những phép gán như vậy gọi là đại lượng ngẫu nhiên hay biến ngẫu nhiên.

Vậy đại lượng ngẫu nhiên là những đại lượng nhận giá trị này hay giá trị khác ở những lần thử khác nhau mà ta không khẳng định trước được khi thực hiện phép thử.

Đại lượng ngẫu nhiên thường được ký hiệu bằng những chữ in hoa như $X, Y, Z, ...$ và dùng chữ in thường như $x_1, x_2, x_3, ...$ để ký hiệu cho những giá trị cụ thể của nó.

Qua đó ta thấy, một đại lượng ngẫu nhiên được xác định nếu biết tập các giá trị của nó và các xác suất mà nó nhận giá trị thuộc tập đó.

Đại lượng ngẫu nhiên được chia thành hai loại:

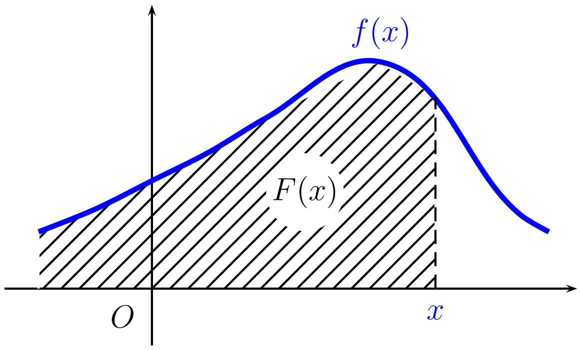
⋄ *Đại lượng ngẫu nhiên rời rạc:* Là đại lượng chỉ nhận hữu hạn hoặc vô hạn đếm được các giá trị "cách quãng" nhau. Chẳng hạn, số chấm xuất hiện khi tung một con xúc xắc; số sinh viên vắng mặt trong mỗi buổi học, số sản phẩm tốt hoặc xấu trong một lô hàng,...

⋄ *Đại lượng ngẫu nhiên liên tục:* Là đại lượng mà giá trị của nó lấp kín cả một đoạn; một khoảng hay toàn bộ trục số. Chẳng hạn, trọng lượng của một sản phẩm khi xuất xưởng; nhiệt độ ở mỗi thời điểm nào đó trong một ngày; tuổi thọ của một thiết bị đang hoạt động; sai số khi đo lường một đại lượng vật lý,...

3.1.1. Hàm phân phối xác suất

Định nghĩa 3.1. Hàm phân phối xác suất của đại lượng ngẫu nhiên X, ký hiệu $F(x)$ hoặc $F_X(x)$, là hàm số thực được xác định như sau: $F(x) = P(X < x),\ \forall x \in \mathbf{R}$.

Hàm phân phối xác suất cho ta biết xác suất để giá trị của X nằm về bên trái của số x. Khi X liên tục thì giá trị của $F(x)$ là phần diện tích ở hình vẽ dưới đây:



Hình 3.1: Giá trị của hàm phân phối

Tính chất 3.1. *Hàm phân phối xác suất của đại lượng ngẫu nhiên X có các tính chất:*

i) $0 \leq F(x) \leq 1, \ \forall x \in \mathbf{R}$.

ii) *$F(x)$ không giảm. Nghĩa là, nếu $x_1 \leq x_2$ thì $F(x_1) \leq F(x_2)$.*

iii) $\lim_{x\to-\infty} F(x) = 0, \lim_{x\to+\infty} F(x) = 1$.

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

3.1.2. Bảng phân phối xác suất

Cho đại lượng ngẫu nhiên rời rạc X có các giá trị là $x_1, x_2, \ldots, x_n$ với xác suất tương ứng là $P(X = x_1) = p_1, P(X = x_2) = p_2, \ldots, P(X = x_n) = p_n$. Khi đó, luật phân phối xác suất của X gọi là bảng phân phối xác suất và được cho như sau:

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

với $p_i \geq 0$ và $\sum_{i=1}^{n} p_i = 1$. Ở đây n có thể hữu hạn hoặc vô hạn đếm được.

Trong bảng trên chỉ cho ta xác suất của từng điểm. Muốn tính xác suất để X nhận giá trị trong một đoạn ta cộng xác suất của từng điểm trong đoạn đó lại với nhau

$$P(a \leq X \leq b) = \sum_{a \leq x_i \leq b} p_i.$$

Khi n hữu hạn thì hàm phân phối $F(x)$ có dạng:

$$F(x) = \begin{cases} 0 & (x \leq x_1), \\ p_1 & (x_1 < x \leq x_2), \\ p_1 + p_2 & (x_2 < x \leq x_3,) \\ \ldots\ldots\ldots\ldots & \\ p_1 + p_2 + \cdots + p_{k-1} & (x_{k-1} < x \leq x_k), \\ \ldots\ldots\ldots\ldots & \\ 1 & (x_n < x). \end{cases}$$

Do đó, đồ thị của $F(x)$ là hàm bậc thang có bước nhảy tại các giá trị x_i.

Ví dụ 3.1. Một hộp có 12 sản phẩm, trong đó có 3 phế phẩm. Lấy ngẫu nhiên 2 sản phẩm từ hộp để kiểm tra. Gọi X là số phế phẩm lấy được. Lập bảng phân phối xác suất của X và tính $P(-1 \leq X \leq 1)$.

Giải: Đại lượng ngẫu nhiên X có các giá trị $0, 1, 2$ nên:

$$P(X=0) = \frac{C_3^0 \cdot C_9^2}{C_{12}^2} = \frac{6}{11},\ P(X=1) = \frac{C_3^1 \cdot C_9^1}{C_{12}^2} = \frac{9}{22},\ P(X=2) = \frac{C_3^2 \cdot C_9^0}{C_{12}^2} = \frac{1}{22}.$$

Bảng phân phối xác suất của X:

X	0	1	2
P	6/11	9/22	1/22

Ta thấy:

$$\sum_{i=0}^{2} p_i = \frac{6}{11} + \frac{9}{22} + \frac{1}{22} = 1,$$

$$P(-1 \leq X \leq 1) = P(X=0) + P(X=1) = \frac{6}{11} + \frac{9}{22} = \frac{21}{22}.$$

Ví dụ 3.2. Giả thiết như ở *Ví dụ 3.1* nhưng ta lần lượt lấy có hoàn lại 2 sản phẩm. Gọi Y là số phế phẩm lấy được. Lập bảng phân phối xác suất của Y.

Giải: Do lấy lần lượt có hoàn lại các sản phẩm nên ta có dãy phép thử Bernoulli với $n = 2$ và $p = 1/4$. Khi đó:
$P(Y=0) = C_2^0 \cdot \left(\frac{1}{4}\right)^0 \cdot \left(\frac{3}{4}\right)^2 = \frac{9}{16}$, $P(Y=1) = C_2^1 \cdot \left(\frac{1}{4}\right)^1 \cdot \left(\frac{3}{4}\right)^1 = \frac{3}{8}$,
$P(Y=2) = C_2^2 \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right)^0 = \frac{1}{16}$.

Do đó, ta có bảng phân phối xác suất như sau:

Y	0	1	2
P	9/16	3/8	1/16

Ví dụ 3.3. Một phân xưởng có 2 máy hoạt động độc lập. Xác suất trong một ngày làm việc các máy đó hỏng tương ứng là $0,2$ và $0,3$. Gọi X là số máy hỏng trong một ngày làm việc. Lập hàm phân phối xác suất của X và vẽ đồ thị của nó.

Giải:

◦ Lập hàm phân phối xác suất của X.

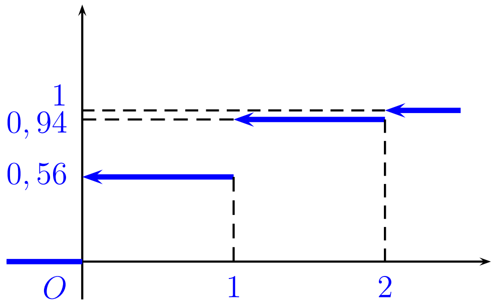
Gọi A_i là biến cố máy thứ i ($i = 1, 2$) bị hỏng. Đại lượng ngẫu nhiên X nhận các giá trị là $0; 1; 2$. Do đó:
$P(X=0) = P(\overline{A}_1\overline{A}_2) = P(\overline{A}_1)P(\overline{A}_2) = 0,8 \cdot 0,7 = 0,56$.
$P(X=1) = P(A_1\overline{A}_2 + \overline{A}_1A_2) = P(A_1)P(\overline{A}_2) + P(\overline{A}_1)P(A_2) = 0,2 \cdot 0,7 + 0,8 \cdot 0,3 = 0,38$.
$P(X=2) = P(A_1A_2) = P(A_1)P(A_2) = 0,2 \cdot 0,3 = 0,06$.

Bảng phân phối xác suất của X:

X	0	1	2
P	0,56	0,38	0,06

Từ đó ta có hàm phân phối xác suất của X: $F(x) = \begin{cases} 0 & (x \leq 0), \\ 0,56 & (0 < x \leq 1), \\ 0,94 & (1 < x \leq 2), \\ 1 & (2 < x). \end{cases}$

∘ Đồ thị:



3.1.3. Hàm mật độ xác suất

Định nghĩa 3.2. Cho đại lượng ngẫu nhiên liên tục X. Hàm số không âm $f(x)$, $x \in \mathbf{R}$ được gọi là hàm mật độ xác suất của X nếu nó thỏa điều kiện sau:

$$P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x.$$

Nhận xét 3.1. Từ định nghĩa nêu trên ta có các nhận xét sau đây:

i) $\displaystyle\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1.$

ii) $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \displaystyle\int_a^b f(x)\mathrm{d}x.$

iii) $P(X = a) = 0$ (Khi X là đại lượng ngẫu nhiên liên tục thì ta chỉ quan tâm tới xác suất để X nằm trong một khoảng nào đó chứ không quan tâm tới xác suất để X nhận một giá trị cụ thể).

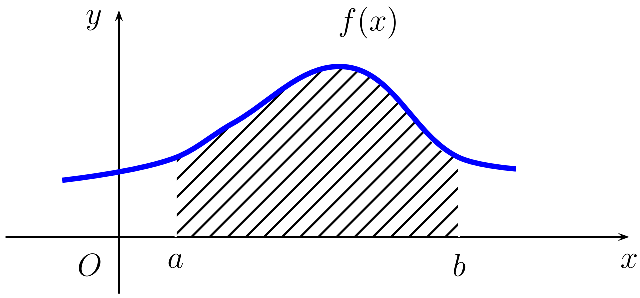
iv) Nếu X có hàm mật độ xác suất là $f(x)$ thì hàm phân phối xác suất của nó được xác định bởi:

$$F(x) = \int_{-\infty}^{x} f(x)\mathrm{d}x.$$

Nếu hàm phân phối $F(x)$ có đạo hàm liên tục trên $\mathbf{R}$ thì hàm mật độ xác suất của X được tính như sau:

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x}.$$

Ý nghĩa: Xác suất để đại lượng ngẫu nhiên liên tục X nhận giá trị trong (a, b) bằng diện tích hình thang cong giới hạn bởi $x = a, x = b, y = f(x)$ và trục Ox.



Hình 3.2: Ý nghĩa của hàm mật độ

Ví dụ 3.4. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ xác suất:

$$f(x) = \begin{cases} 0 & \text{nếu } x < 0 \text{ hoặc } x > \dfrac{\pi}{2}, \\ a\sin 2x & \text{nếu } 0 \le x \le \dfrac{\pi}{2}. \end{cases}$$

Xác định a và tính $P\left(-\frac{\pi}{4} \le X \le \frac{\pi}{4}\right)$.

Giải: Ta có:

$$1 = \int_{-\infty}^{+\infty} f(x)\mathrm{d}x$$

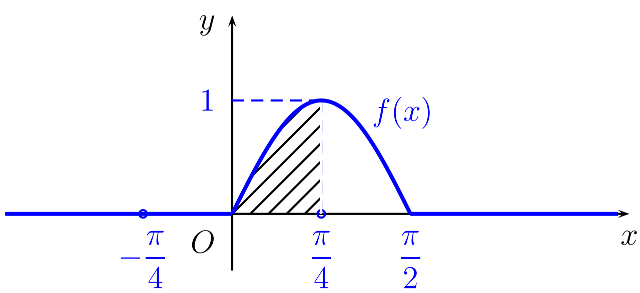
$$\Leftrightarrow 1 = \int_{-\infty}^{0} f(x)\mathrm{d}x + \int_{0}^{\frac{\pi}{2}} f(x)\mathrm{d}x + \int_{\frac{\pi}{2}}^{+\infty} f(x)\mathrm{d}x = 0 + \int_{0}^{\frac{\pi}{2}} a\sin 2x\mathrm{d}x + 0 = -a\frac{\cos 2x}{2}\Big|_0^{\frac{\pi}{2}} = a.$$

Ta nhận giá trị $a = 1$ vì khi đó $f(x) \ge 0, \ \forall x$.

Và:

$$P\left(-\frac{\pi}{4} \le X \le \frac{\pi}{4}\right) = \int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} f(x)\mathrm{d}x = \int_{-\frac{\pi}{4}}^{0} f(x)\mathrm{d}x + \int_{0}^{\frac{\pi}{4}} f(x)\mathrm{d}x$$

$$= 0 + \int_{0}^{\frac{\pi}{4}} \sin 2x\mathrm{d}x = -\frac{\cos 2x}{2}\Big|_0^{\frac{\pi}{4}} = \frac{1}{2}.$$



3.1.4. Phân phối xác suất cho hàm

Giả sử hàm $Y = \varphi(X)$, với X là đại lượng ngẫu nhiên, có phân phối xác suất nào đó. Để giải bài toán tìm phân phối xác suất cho Y, ta xét 2 trường hợp:

a. Khi X là đại lượng ngẫu nhiên rời rạc: Từ tập giá trị $\Omega(X)$ của X ta tìm tập giá trị $\Omega(Y)$ của Y. Khi đó $\forall y_i \in \Omega(Y)$, ta có:

$$P(Y = y_j) = \sum_{\varphi(x_i)=y_j} P(X = x_i).$$

Ví dụ 3.5. Cho đại lượng ngẫu nhiên X có bảng phân phối xác suất:

X	-2	-1	0	1
P	$0,2$	$0,3$	$0,15$	$0,35$

Lập bảng phân phối xác suất của $Y = 1 - X^2$.

Giải: Bảng các giá trị của Y tương ứng với các giá trị của X:

X	-2	-1	0	1
Y	-3	0	1	0

Do đó: $\Omega(Y) = \{-3, 0, 1\}$. Ta có:
$P(Y = 0) = P(X = -1) + P(X = 1) = 0,3 + 0,35 = 0,65.$
$P(Y = 1) = P(X = 0) = 0,15$ và $P(Y = -3) = P(X = -2) = 0,2.$

Bảng phân phối xác suất của Y:

Y	-3	0	1
P	$0,2$	$0,65$	$0,15$

b. Khi X là đại lượng ngẫu nhiên liên tục:

Định lí 3.1. *Nếu X là đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất $f(x)$ và $\varphi(x)$ là một hàm đơn điệu ngặt (đơn điệu tăng hoặc đơn điệu giảm), khả vi thì đại lượng ngẫu nhiên $Y = \varphi(X)$ có hàm mật độ xác suất là:*

$$g(y) = \begin{cases} f\left[\varphi^{-1}(y)\right] \cdot \left|\dfrac{\mathrm{d}\varphi^{-1}}{\mathrm{d}y}\right| & \text{nếu } \exists x : \varphi(x) = y, \\ 0 & \text{nếu } \nexists x : \varphi(x) = y. \end{cases}$$

Ví dụ 3.6. Cho đại lượng ngẫu nhiên X có hàm mật độ xác suất là $f(x)$. Tìm hàm mật độ xác suất của đại lượng ngẫu nhiên $Y = X^3$.

Giải: Ta có $y = x^3 \Rightarrow x = \sqrt[3]{y} = \varphi^{-1}(y)$. Do đó:

$$g(y) = f\left(\sqrt[3]{y}\right) \cdot \left|(\sqrt[3]{y})'\right| = \frac{1}{3\sqrt[3]{y^2}} \cdot f(\sqrt[3]{y}).$$

3.2. CÁC THAM SỐ ĐẶC TRƯNG

Ngoài luật phân phối xác suất còn có những tham số cho ta biết thông tin nhất định về đại lượng ngẫu nhiên. Chúng thường được chia thành 2 loại:

∗ Loại đặc trưng cho xu hướng trung tâm như: Mốt (Mode), kỳ vọng,...

∗ Loại đặc trưng cho độ phân tán như: Phương sai, độ lệch chuẩn,...

Ngoài ra đại lượng ngẫu nhiên còn có các tham số đặc trưng khác như: Mômen, hệ số bất đối xứng, hệ số nhọn,...

3.2.1. Giá trị tin chắc của X (Mod(X) – Mode)

Định nghĩa 3.3. Nếu X là đại lượng ngẫu nhiên rời rạc thì $\text{Mod}(X)$ là giá trị của X ứng với xác suất lớn nhất. Vậy $\text{Mod}(X)$ là giá trị của X có nhiều khả năng xảy ra nhất, nghĩa là:

$$\text{Mod}(X) = x_i \Leftrightarrow p_i = \max\{p_1, p_2, p_3, \ldots\}.$$

Nếu X là đại lượng ngẫu nhiên liên tục thì $\text{Mod}(X)$ là giá trị mà tại đó hàm mật độ xác suất đạt giá trị lớn nhất. Như vậy, $\text{Mod}(X)$ là giá trị mà X có nhiều khả năng xuất hiện trong một khoảng chứa nó nhất, tức là:

$$\text{Mod}(X) = c \Leftrightarrow f(c) = \max\{f(x), x \in \mathbf{R}\}.$$

Ví dụ 3.7. Tìm giá trị tin chắc của đại lượng ngẫu nhiên X trong *Ví dụ 3.3.*

<u>Giải:</u> Theo *Ví dụ 3.3*, ta có bảng phân phối xác suất của X như sau:

X	0	1	2
P	$0,56$	$0,38$	$0,06$

Ta thấy $P(X = 0) = 0,56$ là lớn nhất trong các mức xác suất nên $\text{Mod}(X) = 0$.

Ví dụ 3.8. Tìm giá trị tin chắc của đại lượng ngẫu nhiên X trong *Ví dụ 3.4.*

<u>Giải:</u> Theo *Ví dụ 3.4*, ta thấy hàm mật độ xác suất đạt giá trị lớn nhất tại $x = \pi/4$. Cho nên $\text{Mod}(X) = \pi/4$.

Chú ý: Đại lượng ngẫu nhiên X (rời rạc hay liên tục) có thể có nhiều hơn một giá trị tin chắc.

Ví dụ 3.9. Tìm $\text{Mod}(X)$ của đại lượng ngẫu nhiên X có bảng phân phối xác suất như sau:

X	-2	1	4	5
P	$0,1$	$0,4$	$0,1$	$0,4$

<u>Giải:</u> Do $P(X = 1) = P(X = 5) = 0,4$ là lớn nhất trong các mức xác suất nên $\text{Mod}(X) = 1$ và 5.

3.2.2. Kỳ vọng của X (E(X) – Expectation)

Định nghĩa 3.4. Nếu X đại lượng ngẫu nhiên rời rạc thì

$$\text{E}(X) = x_1p_1 + x_2p_2 + x_3p_3 + \cdots + x_np_n = \sum_{i=1}^{n} x_ip_i \quad (n \text{ có thể vô hạn}).$$

Nếu X đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất là $f(x)$ thì

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x f(x) \mathrm{d}x.$$

Nhận xét 3.2. Giả sử thực hiện n lần thí nghiệm độc lập của đại lượng ngẫu nhiên rời X, biến cố $X = x_i \quad (i = 1, 2, \ldots, k)$ xảy ra n_i lần $\left(\sum_{i=1}^{k} n_i = n\right)$. Gọi $\overline{X}$ là giá trị trung bình nhận được, ta có:

$$\begin{aligned}\overline{X} &= \frac{n_1x_1 + n_2x_2 + \cdots + n_kx_k}{n} = x_1\frac{n_1}{n} + x_2\frac{n_2}{n} + \cdots + x_k\frac{n_k}{n} \\ &= x_1f_1 + x_2f_2 + \cdots + x_kf_k.\end{aligned}$$

Vì $\lim_{n\to+\infty} f_i = p_i \quad (i = 1, 2, \ldots, k)$ nên

$$\lim_{n\to+\infty} \overline{X} = x_1p_1 + x_2p_2 + \cdots + x_kp_k = \mathrm{E}(X).$$

Ý nghĩa: Từ định nghĩa và qua nhận xét nêu trên ta rút ra ý nghĩa của kỳ vọng như sau: "Kỳ vọng là giá trị trung bình nhận được khi ta thực hiện vô hạn các phép thử (độc lập) của đại lượng ngẫu nhiên mà ta đang quan sát".

Tính chất 3.2. *Kỳ vọng có các tính chất sau:*

i) $\mathrm{E}(C) = C$ *(C là hằng số).*

ii) $\mathrm{E}(C \cdot X) = C \cdot \mathrm{E}(X)$.

iii) $\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y)$.

iv) Nếu X và Y là hai đại lượng ngẫu nhiên độc lập (hai biến cố $X < x$ và $Y < y$ độc lập $\forall x, y$) thì $\mathrm{E}(X \cdot Y) = \mathrm{E}(X) \cdot \mathrm{E}(Y)$.

v) Nếu $Y = \varphi(X)$ thì $\mathrm{E}(Y) = \begin{cases} \sum_{i=1}^{n} \varphi(x_i)p_i & \text{nếu } X \text{ rời rạc,} \\ \int_{-\infty}^{+\infty} \varphi(x)f(x)\mathrm{d}x & \text{nếu } X \text{ liên tục.} \end{cases}$

Ví dụ 3.10. Với giả thiết như ở *Ví dụ 3.3.*

a) Tìm số máy hỏng trung bình trong một ngày làm việc.

b) Giả sử mỗi lần hỏng máy phải thuê thợ sửa chữa mất 500 ngàn đồng. Tính số tiền sửa chữa máy trung bình trong một ngày làm việc.

Giải: Theo *Ví dụ 3.3*, ta có bảng phân phối xác suất của X:

X	0	1	2
P	0,56	0,38	0,06

a) $\mathrm{E}(X) = 0 \cdot 0,56 + 1 \cdot 0,38 + 2 \cdot 0,06 = 0,5$.

b) Gọi Y là số tiền phải thuê sửa máy hỏng trong một ngày làm việc. Suy ra $Y = 500X$ nên $\mathrm{E}(Y) = \mathrm{E}(500X) = 500\mathrm{E}(X) = 500 \cdot 0,5 = 250$ (ngàn đồng).

Ví dụ 3.11. Cho đại lượng ngẫu nhiên X có hàm mật độ xác suất:

$$f(x) = \begin{cases} \dfrac{2}{x^2} & \text{với } 1 \le x \le 2, \\ 0 & \text{với } x \text{ còn lại.} \end{cases}$$

a) Tính $\mathrm{E}(X)$. b) Tính $\mathrm{E}(Y)$ với $Y = X^3 - \dfrac{1}{X}$.

<u>Giải:</u> Dễ thấy:

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \int_1^2 \frac{2}{x^2}\mathrm{d}x = -\frac{2}{x}\Big|_1^2 = 1,$$

nên $f(x)$ là hàm mật độ xác suất.

a) Ta có:

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x = \int_1^2 x\frac{1}{x^2}\mathrm{d}x = \int_1^2 \frac{1}{x}\mathrm{d}x = \ln|x|\Big|_1^2 = \ln 2.$$

b) Áp dụng tính chất v), ta có:

$$\mathrm{E}(Y) = \int_{-\infty}^{+\infty} \varphi(x)f(x)\mathrm{d}x = \int_1^2 \left(x^3 - \frac{1}{x}\right)\frac{1}{x^2}\mathrm{d}x = \left[\frac{x^2}{2} + \frac{1}{2x^2}\right]_1^2 = \frac{9}{8}.$$

3.2.3. Phương sai của X (Var(X) – Variance)

Định nghĩa 3.5. Phương sai của đại lượng ngẫu nhiên X, ký hiệu là $\mathrm{D}(X)$ hoặc $\mathrm{Var}(X)$ và được xác định như sau:

$$\mathrm{Var}(X) = \mathrm{E}\Big(\big(X - \mathrm{E}(X)\big)^2\Big).$$

Định lí 3.2. *Nếu X là đại lượng ngẫu nhiên rời rạc thì*

$$\mathrm{Var}(X) = x_1^2p_1 + x_2^2p_2 + \cdots + x_n^2p_n - \big(\mathrm{E}(X)\big)^2.$$

Nếu X là đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất $f(x)$ thì

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} x^2f(x)\mathrm{d}x - \big(\mathrm{E}(X)\big)^2.$$

Chứng minh. Ta có:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\Big(\big(X - \mathrm{E}(X)\big)^2\Big) = \mathrm{E}\Big(X^2 - 2X\cdot\mathrm{E}(X) + \big(\mathrm{E}(X)\big)^2\Big)\\ &= \mathrm{E}(X^2) - 2\mathrm{E}\big(X\cdot\mathrm{E}(X)\big) + \big(\mathrm{E}(X)\big)^2\\ &= \mathrm{E}(X^2) - 2\mathrm{E}(X)\cdot\mathrm{E}(X) + \big(\mathrm{E}(X)\big)^2 = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2,\end{aligned}$$

với

$$\mathrm{E}(X^2) = \begin{cases} \sum_{i=1}^n x_i^2p_i & \text{nếu X rời rạc,} \\ \int_{-\infty}^{+\infty} x^2f(x)\mathrm{d}x & \text{nếu X liên tục.} \end{cases}$$

□

Ý nghĩa: Ta thấy, $X - \mathrm{E}(X)$ là sai số của X so với trung bình nó. Do đó, phương sai $\mathrm{E}\Big(\big(X - \mathrm{E}(X)\big)^2\Big)$ chính là trung bình của bình phương sai số đó và được gọi tắt là phương sai. Nó đo mức độ phân tán của đại lượng ngẫu nhiên quanh giá trị trung bình. Nghĩa là, phương sai nhỏ thì độ phân tán nhỏ nên độ tập trung lớn và ngược lại.

Trong kỹ thuật phương sai đặc trưng cho độ sai số của thiết bị, trong kinh doanh nó đặc trưng cho độ rủi ro của các quyết định, còn trong trồng trọt nó biểu thị cho mức độ ổn định của năng suất,...

Tính chất 3.3. *Phương sai có những tính chất sau:*

i) $\mathrm{Var}(X) \geq 0$, $\mathrm{Var}(X) = 0 \Leftrightarrow X = C$ *(C là hằng số).*

ii) $\mathrm{Var}(C \cdot X) = C^2 \cdot \mathrm{Var}(X)$.

iii) Nếu X và Y là hai đại lượng ngẫu nhiên độc lập với nhau thì

$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Ví dụ 3.12. Năng suất của hai máy tương ứng là các đại lượng ngẫu nhiên X, Y (sản phẩm/phút) và có bảng phân phối xác suất:

X	1	2	3	4
P	0,3	0,1	0,5	0,1

Y	2	3	4	5
P	0,1	0,4	0,4	0,1

Nếu cần mua một trong hai máy trên thì ta nên chọn mua máy nào?

Giải: Ta có:
$\mathrm{E}(X) = 1 \cdot 0,3 + 2 \cdot 0,1 + 3 \cdot 0,5 + 4 \cdot 0,1 = 2,4.$
$\mathrm{Var}(X) = 1^2 \cdot 0,3 + 2^2 \cdot 0,1 + 3^2 \cdot 0,5 + 4^2 \cdot 0,1 - 2,4^2 = 1,04.$
$\mathrm{E}(Y) = 2 \cdot 0,1 + 3 \cdot 0,4 + 4 \cdot 0,4 + 5 \cdot 0,1 = 3,5.$
$\mathrm{Var}(Y) = 2^2 \cdot 0,1 + 3^2 \cdot 0,4 + 4^2 \cdot 0,4 + 5^2 \cdot 0,1 - 3,5^2 = 0,65.$

Ta thấy $\mathrm{E}(Y) > \mathrm{E}(X)$: Năng suất trung bình của Y cao hơn của X và $\mathrm{Var}(Y) < \mathrm{Var}(X)$: Năng suất của Y ổn định hơn của X. Do đó, ta nên chọn mua máy Y.

3.2.4. Độ lệch chuẩn ($\sigma(X)$)

Định nghĩa 3.6. Độ lệch chuẩn của đại lượng ngẫu nhiên X, ký hiệu là $\sigma(X)$ và được xác định như sau: $\sigma(X) = \sqrt{\mathrm{Var}(X)}$.

Ví dụ 3.13. Trọng lượng của một loại sản phẩm là X (kg), có hàm mật độ xác suất như sau:

$$f(x) = \begin{cases} \dfrac{3(x^2-1)}{16} & \text{với } 2 \leq x \leq 3, \\ 0 & \text{với } x \text{ còn lại.} \end{cases}$$

Tính trọng lượng trung bình và độ lệch chuẩn của X.

Giải: Ta thấy $f(x)$ là hàm mật độ xác suất, vì:

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \int_2^3 \frac{3(x^2-1)}{16}\mathrm{d}x = 1.$$

Do đó:

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x = \int_2^3 \frac{3x(x^2-1)}{16}\mathrm{d}x = 2,578\ kg.$$

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x - \big(\mathrm{E}(X)\big)^2$$

$$= \int_2^3 \frac{3x^2(x^2-1)}{16}\mathrm{d}x - 2,578^2 = 0,0783\ kg^2.$$

Nên: $\sigma(X) = \sqrt{\mathrm{Var}(X)} = 0,2798\ kg.$

3.2.5. Trung vị (Med(X) – Median)

Định nghĩa 3.7. Trung vị là điểm chia đôi phân phối xác suất của đại lượng ngẫu nhiên X, ký hiệu $\mathrm{Med}(X)$ và được xác định như sau:

- Nếu X rời rạc thì $\mathrm{Med}(X) = m$ thì $\begin{cases} P(X > m) \le 0,5 \\ P(X < m) \le 0,5 \end{cases}$
- Nếu X liên tục thì $\mathrm{Med}(X) = m$ với $\int_{-\infty}^{m} f(x)\mathrm{d}x = 0,5.$

Ví dụ 3.14. Cho đại lượng ngẫu nhiên X có luật phân phối xác suất như bảng sau:

X	-1	0	1	2
P	$0,15$	$0,2$	$0,3$	$0,35$

Tính $\mathrm{Med}(X)$ và $\mathrm{E}(|X - \mathrm{Med}(X)|)$.

<u>Giải:</u>

Do

$$\begin{cases} P(X > 1) = P(X = 2) = 0,33 \le 0,5 \\ P(X < 1) = P(X = 0) + P(X = -1) = 0,2 + 0,15 = 0,45 \le 0,5 \end{cases}$$

nên $\mathrm{Med}(X) = 1$.

Vì $|X - \mathrm{Med}(X)| = |X - 1|$ nên các giá trị của $|X - \mathrm{Med}(X)|$ được cho như sau:

X	-1	0	1	2
$\lvert X - \mathrm{Med}(X)\rvert$	2	1	0	1

Do đó:
$P(|X - \mathrm{Med}(X)| = 0) = P(X = 1) = 0,2.$
$P(|X - \mathrm{Med}(X)| = 1) = P(X = 0) + P(X = 2) = 0,65.$
$P(|X - \mathrm{Med}(X)| = 2) = P(X = -1) = 0,15.$

Nên: $\mathrm{E}(|X - \mathrm{Med}(X)|) = 0 \cdot 0,2 + 1 \cdot 0,65 + 2 \cdot 0,15 = 0,95.$

3.2.6. Một vài tham số đặc trưng khác

a. Mômen:

Định nghĩa 3.8. Mômen cấp k đối với a của X, ký hiệu là $\mu_k(a)$ và được xác định bởi:

$$\mu_k(a) = \mathrm{E}\big((X-a)^k\big).$$

Khi $a = 0$ thì $\mu_k(0) = \mathrm{E}(X^k)$ gọi là mômen gốc cấp k.

Khi $a = \mathrm{E}(X)$ thì $\mu_k\big(\mathrm{E}(X)\big) = \mathrm{E}\Big(\big(X - \mathrm{E}(X)\big)^k\Big)$ gọi là mômem trung tâm cấp k khi đó ta ký hiệu đơn giản là μ_k. Dễ thấy: $\mu_1 = 0, \mu_2 = \mathrm{Var(X)}$.

b. Hệ số biến thiên:

Định nghĩa 3.9. Hệ số biến thiên của đại lượng ngẫu nhiên X ký hiệu là $\mathrm{CV}(X)$ và được xác định như sau:

$$\mathrm{CV}(X) = \left|\frac{\sigma(X)}{\mathrm{E}(X)}\right|.$$

Hệ số biến thiên được dùng để so sánh mức độ phân tán của các đại lượng ngẫu nhiên có kỳ vọng và phương sai khác nhau.

c. Hệ số bất đối xứng:

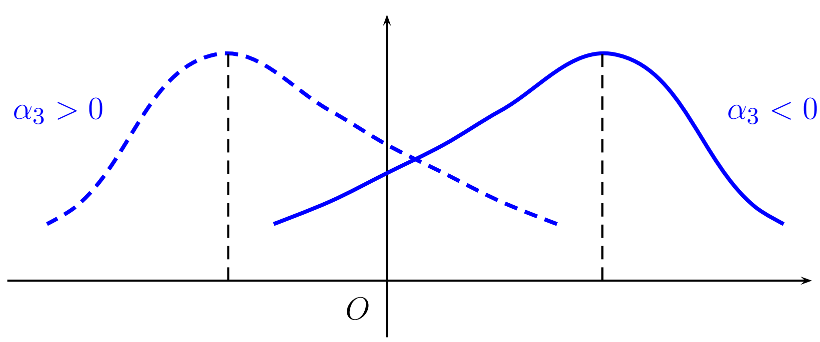
Định nghĩa 3.10. Hệ số bất đối xứng của đại lượng ngẫu nhiên X ký hiệu là α_3 và được xác định bởi biểu thức: $\alpha_3 = \dfrac{\mu_3}{\sigma^3}$.

Hệ số bất đối xứng dùng để nhận dạng đồ thị hàm mật độ xác suất của đại lượng ngẫu nhiên. Nếu

$\alpha_3 = 0:$ Đồ thị đối xứng.

$\alpha_3 > 0:$ Đồ thị lệch về bên trái và xuôi về bên phải.

$\alpha_3 < 0:$ Đồ thị lệch về bên phải và xuôi về bên trái.



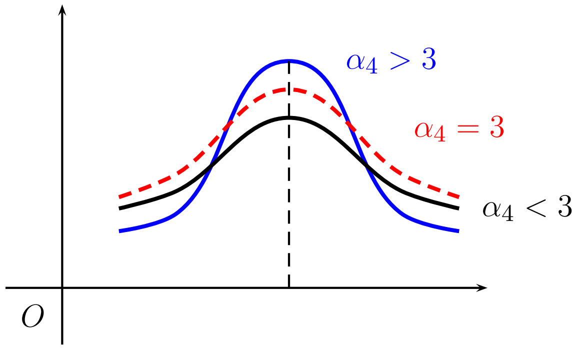
Hình 3.3: Hệ số bất đối xứng

d. Hệ số nhọn:

Định nghĩa 3.11. Hệ số nhọn của đại lượng ngẫu nhiên X ký hiệu là α_4 và được xác định bởi biểu thức: $\alpha_4 = \dfrac{\mu_4}{\sigma^4}$.

Hệ số nhọn dùng để xét độ tập trung của phân phối đại lượng ngẫu nhiên. Nếu

$\alpha_4 < 3$: Ít tập trung.
$\alpha_4 = 3$: Tập trung ở mức độ thường.
$\alpha_4 > 3$: Rất tập trung.



Hình 3.4: Hệ số nhọn

3.3. ĐẠI LƯỢNG NGẪU NHIÊN HAI CHIỀU

Chúng ta đã nghiên cứu các vấn đề cơ bản của đại lượng ngẫu nhiên một chiều: Hàm phân phối xác suất, bảng phân phối xác suất và các tham số đặc trưng,.... Tuy nhiên trong thực tế có rất nhiều bài toán để giải được ta phải xét đồng thời nhiều đại lượng ngẫu nhiên có liên quan với nhau gọi là đại lượng ngẫu nhiên nhiều chiều hay vectơ ngẫu nhiên.

Do yêu cầu của học phần không đi sâu vào lý thuyết chuyên ngành, nên trong nội dung bài giảng ta chỉ xét đại lượng ngẫu nhiên hai chiều, các kết quả của nó có thể mở rộng cho đại lượng ngẫu nhiên nhiều chiều.

Đại lượng ngẫu nhiên hai chiều là cặp (X, Y) với X và Y là các đại lượng ngẫu nhiên. Nếu X và Y đều rời rạc (liên tục) thì ta gọi đại lượng ngẫu nhiên hai chiều (X, Y) là rời rạc (liên tục). Ở đây ta chỉ xét trường hợp cả X và Y đều rời rạc hoặc đều liên tục.

3.3.1. Hàm phân phối xác suất

Định nghĩa 3.12. Hàm phân phối xác suất của đại lượng ngẫu nhiên hai chiều (X, Y) hay hàm phân phối đồng thời của X và Y là hàm hai biến được xác định như sau:

$$F(x, y) = P(X < x, Y < y), \ \forall x, y \in \mathbf{R}.$$

Tính chất 3.4. *Cũng như trường hợp một chiều, hàm phân phối xác suất $F(x, y)$ có các tính chất sau:*

i) $0 \leq F(x, y) \leq 1, \ \forall x, y \in \mathbf{R}$.

ii) *$F(x, y)$ không giảm theo từng biến số (nghĩa là nếu biến này là hằng số thì hàm không giảm đối với biến kia và ngược lại).*

iii) $\lim_{\substack{x\to-\infty \\ y\to-\infty}} F(x,y) = 0, \ \lim_{\substack{x\to+\infty \\ y\to+\infty}} F(x,y) = 1.$

iv) Từ hàm phân phối xác suất $F(x,y)$ của đại lượng ngẫu nhiên hai chiều (X,Y) ta có thể suy ra hàm phân phối xác suất của từng đại lượng ngẫu nhiên X và Y như sau:

$$\lim_{y\to+\infty} F(x,y) = F_X(x), \ \lim_{x\to+\infty} F(x,y) = F_Y(y).$$

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

Định nghĩa 3.13. Hai đại lượng ngẫu nhiên X, Y gọi là độc lập nếu như

$$F(x,y) = F_X(x)\cdot F_Y(y), \quad \forall x,y \in \mathbf{R}.$$

Nhận xét 3.3. Nếu X, Y độc lập thì hàm phân phối đồng thời của X và Y được xác định qua các hàm phân phối của X và của Y.

3.3.2. Bảng phân phối xác suất

Bảng phân phối xác suất của đại lượng ngẫu nhiên hai chiều rời rạc (X,Y) còn gọi là bảng phân phối đồng thời của X và Y được cho như sau:

$Y \backslash X$	x_1	x_2	...	x_i	...	x_n	$P(Y=y_j)$
y_1	p_{11}	p_{21}	...	p_{i1}	...	p_{1m}	q_1
y_2	p_{12}	p_{22}	...	p_{i2}	...	p_{2m}	q_2
...	...	...	...	...	...	...	...
y_j	p_{1j}	p_{2j}	...	p_{ij}	...	p_{nj}	q_j
...	...	...	...	...	...	...	...
y_m	p_{1n}	p_{2n}	...	p_{im}	...	p_{nm}	q_m
$P(X=x_i)$	p_1	p_2	...	p_i	...	p_n	1

Trong đó:

$$p_{ij} = P(X=x_i, Y=y_j) \text{ và } \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} = 1.$$

Chú ý:

1) Qua bảng phân phối xác suất của (X,Y), ta có thể suy ra các phân phối biên của X và của Y:

▷ Phân phối xác suất của X:

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

với $p_i = \sum_{j=1}^{m} p_{ij}$ (Cộng theo cột $i, i = 1,\ldots,n$).

▷ Phân phối xác suất của Y:

Y	y_1	y_2	...	y_m
P	q_1	q_2	...	q_m

với $q_j = \sum_{i=1}^{n} p_{ij}$ (Cộng theo dòng $j, j = 1,\ldots,m$).

Nhận xét 3.4. Hai đại lượng ngẫu nhiên X và Y độc lập với nhau khi và chỉ khi $p_{ij} = p_i \cdot q_j$ $(\forall i = 1, \ldots, n;\ \forall j = 1, \ldots, m)$.

2) Từ bảng phân phối xác suất của (X, Y), ta có thể suy ra phân phối có điều kiện:
▹ Của X với điều kiện $Y = y_j$:

$X \vert (Y = y_j)$	x_1	x_2	$\ldots$	x_i	$\ldots$	x_n
P	$p_{1\vert Y=y_j}$	$p_{2\vert Y=y_j}$	$\ldots$	$p_{i\vert Y=y_j}$	$\ldots$	$p_{n\vert Y=y_j}$

Trong đó: $p_{i|Y=y_j} = P\big((X = x_i)|(Y = y_j)\big) = \dfrac{p_{ij}}{q_j}$ $(\forall i = 1, \ldots, n)$.

▹ Của Y với điều kiện $X = x_i$:

$Y \vert (X = x_i)$	y_1	y_2	$\ldots$	y_j	$\ldots$	y_m
P	$q_{1\vert X=x_i}$	$q_{2\vert X=x_i}$	$\ldots$	$q_{j\vert X=x_i}$	$\ldots$	$q_{m\vert X=x_i}$

Trong đó: $q_{j|X=x_i} = P\big((Y = y_j)|(X = x_i)\big) = \dfrac{p_{ij}}{p_i}$ $(\forall j = 1, \ldots, m)$.

Ví dụ 3.15. Người ta thống kê dân số của một vùng theo hai chỉ tiêu: Giới tính (X) và học vấn (Y) được kết quả như sau:

X \ Y	Thất học: 0	Phổ thông: 1	Đại học: 2
Nam: 0	$0,12$	$0,24$	$0,15$
Nữ: 1	$0,13$	$0,22$	$0,14$

a) Lập bảng phân phối xác suất của học vấn; của giới tính.

b) Học vấn có độc lập với giới tính hay không?

c) Chọn ngẫu nhiên một người, tính xác suất để người đó không thất học.

d) Lập bảng phân phối xác suất học vấn của nữ; tính trung bình học vấn của nữ.

Giải:

a) Bảng phân phối xác suất của X và của Y.

X	0	1
P	$0,51$	$0,49$

Y	0	1	2
P	$0,25$	$0,46$	$0,29$

b) Ta có:
$p_{11} = P(X = 0, Y = 0) = 0,12$ và $P(X = 0) \cdot P(Y = 0) = 0,51 \cdot 0,25 = 0,1275$.
Nên $p_{11} \neq P(X = 0) \cdot P(Y = 0)$. Do đó, học vấn không độc lập với giới tính.

c) Xác suất để người đó không thất học là:

$$\begin{aligned} P(X \geq 0, Y > 0) &= \textstyle\sum_{x_i \geq 0} \sum_{y_j > 0} p_{ij} = 0,24 + 0,22 + 0,15 + 0,14 = 0,75 \\ &= 1 - P(X \geq 0, Y = 0). \end{aligned}$$

d) Lập bảng phân phối của Y với điều kiện $X = 1$.

$$P\big((Y = 0)|(X = 1)\big) = \frac{P(X = 1, Y = 0)}{P(X = 1)} = \frac{0,13}{0,49} = 0,2653.$$

$P\big((Y=1)|(X=1)\big) = \dfrac{P(X=1,Y=1)}{P(X=1)} = \dfrac{0,22}{0,49} = 0,449.$
$P\big((Y=2)|(X=1)\big) = \dfrac{P(X=1,Y=2)}{P(X=1)} = \dfrac{0,14}{0,49} = 0,2857.$

Bảng phân phối điều kiện:

$Y\|(X=1)$	0	1	2
P	0,2653	0,449	0,2857

3.3.3. Hàm mật độ xác suất

Định nghĩa 3.14. Cho đại lượng ngẫu nhiên hai chiều liên tục (X,Y). Hàm không âm $f(x,y)$ với $x,y\in\mathbf{R}$ được gọi là hàm mật độ xác suất đồng thời của (X,Y) nếu nó thỏa điều kiện:

$$P\big((X,Y)\in\mathbf{G}\big) = \iint_{\mathbf{G}} f(x,y)\mathrm{d}x\mathrm{d}y \quad (\mathbf{G}\subset\mathbf{R}^2).$$

Tính chất 3.5. *Hàm mật độ $f(x,y)$ có các tính chất sau:*

i) $\iint_{\mathbf{R}^2} f(x,y)\mathrm{d}x\mathrm{d}y = 1.$

ii) Nếu hàm phân phối xác suất $F(x,y)$ có các đạo hàm riêng liên tục trên $\mathbf{R}^2$ thì ta suy ra được hàm mật độ xác suất như sau:

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x\partial y}.$$

iii) Nếu có hàm mật độ xác suất $f(x,y)$ thì ta suy ra được hàm phân phối xác suất như sau:

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\mathrm{d}u\mathrm{d}v.$$

Từ hàm mật độ xác suất đồng thời $f(x,y)$ của đại lượng ngẫu nhiên hai chiều liên tục (X,Y), ta suy ra được các hàm sau đây:

⋆ ***Hàm mật độ xác suất biên:***
▷ Hàm mật độ xác suất biên của X: $f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y$.
▷ Hàm mật độ xác suất biên của Y: $f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x$.

Nhận xét 3.5. X và Y độc lập với nhau khi và chỉ khi $f(x,y) = f_X(x)\cdot f_Y(y)$.

⋆ ***Hàm phân phối xác suất có điều kiện:***
▷ Hàm phân phối xác suất của X với điều kiện $Y=y_0$, ký hiệu $F(x|y_0)$ và được định nghĩa như sau: $F(x|y_0) = \lim_{\Delta y\to 0^+} P(X<x|y_0<Y<y_0+\Delta y)$.

Ta có:

$$F(x|y_0) = \int_{-\infty}^{x} \frac{f(u,y_0)}{f_Y(y_0)}\mathrm{d}u.$$

Thật vậy:

$$\begin{aligned}\lim_{\Delta y\to 0^+} P(X<x|y_0<Y<y_0+\Delta y) &= \lim_{\Delta y\to 0^+}\frac{P(X<x;y_0<Y<y_0+\Delta y)}{P(y_0<Y<y_0+\Delta y)}\\ &= \lim_{\Delta y\to 0^+}\frac{\int_{-\infty}^{x}\int_{y_0}^{y_0+\Delta y} f(t,y)\mathrm{d}y\mathrm{d}t}{\int_{y_0}^{y_0+\Delta y} f_Y(y)\mathrm{d}y}\\ &= \lim_{\Delta y\to 0^+}\frac{\int_{-\infty}^{x}\int_{y_0}^{y_0+\Delta y} \frac{f(t,y)}{\Delta y}\mathrm{d}y\mathrm{d}t}{\int_{y_0}^{y_0+\Delta y} \frac{f_Y(y)}{\Delta y}\mathrm{d}y} = \frac{\int_{-\infty}^{x} f(t,y_0)\mathrm{d}t}{f_Y(y_0)}.\end{aligned}$$

▹ Hàm phân phối xác suất của Y với điều kiện $X=x_0$, ký hiệu $F(y|x_0)$ và được định nghĩa như sau: $F(y|x_0)=\lim_{\Delta x\to 0^+} P(Y<y|x_0<X<x_0+\Delta x)$.

Ta có:

$$F(y|x_0)=\int_{-\infty}^{y}\frac{f(x_0,v)}{f_X(x_0)}\mathrm{d}v.$$

⋆ ***Hàm mật độ xác suất có điều kiện:***

▹ Hàm mật độ xác suất của X với điều kiện $Y=y_0$:

$$f(x|y_0)=\frac{\mathrm{d}F(x|y_0)}{\mathrm{d}x}=\frac{f(x,y_0)}{f_Y(y_0)}.$$

▹ Hàm mật độ xác suất của Y với điều kiện $X=x_0$:

$$f(y|x_0)=\frac{\mathrm{d}F(y|x_0)}{\mathrm{d}y}=\frac{f(x_0,y)}{f_X(x_0)}.$$

Ví dụ 3.16. Cho hàm mật độ của đại lượng ngẫu nhiên hai chiều (X,Y):

$$f(x,y)=\begin{cases} xe^{-x(1+y)} & \text{với } x>0 \text{ và } y>0,\\ 0 & \text{các trường hợp còn lại.}\end{cases}$$

a) Tìm hàm mật độ xác suất $f_X(x)$ của X.

b) Tìm hàm mật độ xác suất $f(y|x_0)$ của Y với điều kiện $X=x_0$.

<u>Giải:</u>

a) Nếu $x>0$ thì

$$\begin{aligned}f_X(x)&=\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y=\int_{0}^{+\infty} x\cdot e^{-x(1+y)}\mathrm{d}y\\ &=-\int_{0}^{+\infty} e^{-x(1+y)}\mathrm{d}\big(-x(1+y)\big)=-e^{-x(1+y)}\Big|_0^{+\infty}=e^{-x}.\end{aligned}$$

Nếu $x\le 0$ thì

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y=\int_{-\infty}^{+\infty} 0\ \mathrm{d}y=0.$$

Nên:

$$f_X(x)=\begin{cases} e^{-x} & \text{với } x>0,\\ 0 & \text{với } x\le 0.\end{cases}$$

b) Ta chỉ xét trường hợp $x_0 > 0$:
Nếu $y > 0$ thì

$$f(y|x_0) = \frac{f(x_0, y)}{f_X(x_0)} = \frac{x_0 \cdot e^{-x_0(1+y)}}{e^{-x_0}} = x_0 \cdot e^{-x_0 y}.$$

Nếu $y \leq 0$ thì

$$f(y|x_0) = \frac{f(x_0, y)}{f_X(x_0)} = \frac{0}{e^{-x_0}} = 0.$$

Do đó:

$$f(y|x_0) = \begin{cases} x_0 e^{-x_0 y} & \text{với } y > 0, \\ 0 & \text{với } y \leq 0. \end{cases}$$

3.3.4. Phân phối xác suất cho hàm

a. Trường hợp rời rạc: Cho đại lượng ngẫu nhiên hai chiều rời rạc (X, Y) và hàm $\varphi(x, y)$. Để lập bảng phân phối xác suất của $Z = \varphi(X, Y)$ ta làm như sau:

- ◦ Tìm tập các giá trị của Z tương ứng với các giá trị có thể nhận của X, Y.
- ◦ Tìm các xác suất $P(Z = z_k) = \sum_{\varphi(x_i, y_j) = z_k} p_{ij}$.

Ví dụ 3.17. Cho bảng phân phối đồng thời của X, Y như sau:

X \ Y	0	1	2
1	0,1	0,2	0,05
2	0,15	0,25	0,25

Hãy lập bảng phân phối xác suất của $Z = 3X - 2Y + 1$.

<u>Giải:</u> Ta có:
$P(Z = 4) = P(X = 1, Y = 0) = 0,1$; $P(Z = 2) = P(X = 1, Y = 1) = 0,2$.
$P(Z = 0) = P(X = 1, Y = 2) = 0,05$; $P(Z = 7) = P(X = 2, Y = 0) = 0,15$.
$P(Z = 5) = P(X = 2, Y = 1) = 0,25$; $P(Z = 3) = P(X = 2, Y = 2) = 0,25$.

Nên bảng giá trị của Z với xác suất tương ứng là:

<table>
<tr><td colspan="2">Z</td><td colspan="3">Y</td></tr>
<tr><td colspan="2">P</td><td>0</td><td>1</td><td>2</td></tr>
<tr><td rowspan="2">X</td><td>1</td><td>4
0,1</td><td>2
0,2</td><td>0
0,05</td></tr>
<tr><td>2</td><td>7
0,15</td><td>5
0,25</td><td>3
0,25</td></tr>
</table>

Do đó bảng phân phối xác suất của Z là:

Z	0	2	3	4	5	7
P	0,05	0,2	0,25	0,1	0,25	0,15

b. Trường hợp liên tục: Giả sử (X, Y) là đại lượng ngẫu nhiên hai chiều liên tục, có hàm mật độ xác suất là $f(x, y)$; các hàm $u = u(x, y), v = v(x, y)$ có các đạo hàm riêng liên tục và định thức Jacobi:

$$J = \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} \neq 0 \quad (\forall x, y \in \mathbf{R}).$$

Khi đó hàm mật độ xác suất của đại lượng ngẫu nhiên hai chiều liên tục (U, V) (với $U = u(X, Y)$, $V = v(X, Y)$) được xác định bởi:

$$g(u, v) = f\big(x(u, v), y(u, v)\big) \cdot |J|.$$

Ví dụ 3.18. Cho đại lượng ngẫu nhiên hai chiều (X, Y) có hàm mật độ xác suất $f(x, y)$. Tìm hàm mật độ của đại lượng ngẫu nhiên $U = X + Y$.

Giải: Đặt $\begin{cases} u = x + y \\ v = y \end{cases} \Rightarrow \begin{cases} x = u - v \\ y = v \end{cases}$. Ta có: $J = \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 0 \end{vmatrix} = 1 \neq 0.$

Nên hàm mật độ xác suất của đại lượng ngẫu nhiên hai chiều liên tục (U, V) với $U = X + Y,\ V = Y$ được xác định bởi:

$$g(u, v) = f(u - v, v).$$

Khi đó hàm mật độ xác suất của U là:

$$g_U(u) = \int_{-\infty}^{+\infty} g(u, v)\mathrm{d}v = \int_{-\infty}^{+\infty} f(u - v, v)\mathrm{d}v.$$

3.3.5. Hiệp phương sai và hệ số tương quan

a. Hiệp phương sai: Khi xét một đại lượng ngẫu nhiên nhiều chiều cần có một hệ số để đo mức độ phụ thuộc tuyến tính giữa các đại lượng ngẫu nhiên với nhau, như giữa chiều cao và cân nặng, giữa số tiền đầu tư và lợi nhuận thu được,.... Hệ số đó chính là hiệp phương sai.

Định nghĩa 3.15. Hiệp phương sai của hai đại lượng ngẫu nhiên X và Y, ký hiệu $\mathrm{Cov}(X, Y)$ và được xác định như sau:

$$\mathrm{Cov}(X, Y) = \mathrm{E}\Big(\big(X - \mathrm{E}(X)\big)\big(Y - \mathrm{E}(Y)\big)\Big).$$

Nhận xét 3.6. Từ định nghĩa, ta có: $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X) \cdot \mathrm{E}(Y)$.

Thật vậy:

$$\begin{aligned} \mathrm{E}\Big(\big(X - \mathrm{E}(X)\big)\big(Y - \mathrm{E}(Y)\big)\Big) &= \mathrm{E}\Big(XY - X\mathrm{E}(Y) - Y\mathrm{E}(X) + \mathrm{E}(X)\mathrm{E}(Y)\Big) \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) - \mathrm{E}(Y)\mathrm{E}(X) + \mathrm{E}(X)\mathrm{E}(Y) \\ &= \mathrm{E}(XY) - \mathrm{E}(Y)\mathrm{E}(X) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y). \end{aligned}$$

Công thức tính $\mathrm{Cov}(X, Y)$:

▷ Nếu (X, Y) là đại lượng ngẫu nhiên hai chiều rời rạc thì

$$\mathrm{Cov}(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p_{ij} - \mathrm{E}(X)\mathrm{E}(Y).$$

▷ Nếu (X, Y) là đại lượng ngẫu nhiên hai chiều liên tục thì

$$\mathrm{Cov}(X,Y) = \iint_{\mathbf{R}^2} xyf(x,y)\mathrm{d}x\mathrm{d}y - \mathrm{E}(X)\mathrm{E}(Y).$$

Tính chất 3.6. *Hiệp phương sai có các tính chất sau:*

i) $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(\mathrm{Y}) + 2\mathrm{Cov}(X,Y)$.

ii) $|\mathrm{Cov}(X,Y)| \le \sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}$.

iii) Nếu X và Y độc lập với nhau thì $\mathrm{Cov}(X,Y) = 0$.

Các hiệp phương sai của đại lượng ngẫu nhiên nhiều chiều thường được đưa vào một ma trận gọi là *ma trận hiệp phương sai*. Đối với đại lượng ngẫu nhiên hai chiều (X, Y), ma trận hiệp phương sai được xác định bởi:

$$\mathrm{C}(X,Y) = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{bmatrix}.$$

b. Hệ số tương quan:

Định nghĩa 3.16. Hệ số tương quan của hai đại lượng ngẫu nhiên X và Y, ký hiệu $\rho(X, Y)$ và được xác định bởi:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}.$$

Nhận xét 3.7. Hệ số tương quan cũng là giá trị đo mức độ phụ thuộc tuyến tính giữa các đại lượng ngẫu nhiên. Nhưng rõ ràng hệ số tương quan tốt hơn vì nó không phụ thuộc vào đơn vị đo của X, Y như là hiệp phương sai.

Tính chất 3.7. *Hệ số tương quan có những tính chất sau:*

i) $\rho(X,Y) = 1$.

ii) $|\rho(X,Y)| \le 1$. *Và* $|\rho(X,Y)| = 1$ *khi và chỉ khi* X, Y *phụ thuộc tuyến tính.*

iii) Nếu X và Y độc lập với nhau thì $|\rho(X,Y)| = 0$ nhưng điều ngược lại không đúng. Khi $|\rho(X,Y)| = 0$, ta nói X và Y không tương quan nhau.

Giống như hiệp phương sai ta cũng có *ma trận hệ số tương quan*. Cho đại lượng ngẫu nhiên hai chiều (X, Y), ma trận hệ số tương quan được xác định như sau:

$$\mathrm{R}(X,Y) = \begin{bmatrix} 1 & \rho(X,Y) \\ \rho(X,Y) & 1 \end{bmatrix}.$$

Ví dụ 3.19. Cho đại lượng ngẫu nhiên hai chiều rời rạc (X, Y) có bảng phân phối xác suất như sau:

X \ Y	0	10	20
10	0,1	0,2	0,05
20	0,15	0,25	0,25

a) Tính hiệp phương sai và ma trận hiệp phương sai của (X, Y).

b) Tính hệ số tương quan và ma trận hệ số tương quan của (X, Y).

Giải: Ta có bảng giá trị của đại lượng XY và xác suất tương ứng:

<table>
<tr><td colspan="2" rowspan="2">XY
P</td><td colspan="3">Y P</td></tr>
<tr><td>0
0,25</td><td>10
0,45</td><td>20
0,3</td></tr>
<tr><td rowspan="2">X P</td><td>10
0,35</td><td>0
0,1</td><td>100
0,2</td><td>200
0,05</td></tr>
<tr><td>20
0,65</td><td>0
0,15</td><td>200
0,25</td><td>400
0,25</td></tr>
</table>

Từ đó:
$\mathrm{E}(XY) = 0 \cdot (0,1 + 0,15) + 100 \cdot 0,2 + 200 \cdot (0,05 + 0,25) + 400 \cdot 0,25 = 180.$
$\mathrm{E}(X) = 10 \cdot 0,35 + 20 \cdot 0,65 = 16,5.$
$\mathrm{Var}(X) = 10^2 \cdot 0,35 + 20^2 \cdot 0,65 - 16,5^2 = 22,75.$
$\mathrm{E}(Y) = 0 \cdot 0,25 + 10 \cdot 0,45 + 20 \cdot 0,3 = 10,5.$
$\mathrm{Var}(Y) = 0^2 \cdot 0,25 + 10^2 \cdot 0,45 + 20^2 \cdot 0,3 - 10,5^2 = 54,75.$

Nên: $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = 180 - 16,5 \cdot 10,5 = 6,75$
và $\rho(X, Y) = \dfrac{6,75}{\sqrt{22.75 \cdot 54,75}} = 0,1913.$

a) Ma trận hiệp phương sai: $\mathrm{C}(X, Y) = \begin{bmatrix} 22,75 & 6,75 \\ 6,75 & 22,75 \end{bmatrix}.$

b) Ma trận tương quan: $\mathrm{R}(X, Y) = \begin{bmatrix} 1 & 0,1913 \\ 0,1913 & 1 \end{bmatrix}.$

3.4. BÀI TẬP CHƯƠNG 3

Bài 1: Tung một đồng xu 2 lần độc lập. Gọi X là đại lượng ngẫu nhiên chỉ số lần xuất hiện mặt sấp.

a) Lập bảng phân phối xác suất cho X.

b) Tìm hàm phân phối xác suất của X và vẽ đồ thị của nó.

c) Tính xác suất có ít nhất 1 lần xuất hiện mặt sấp.

d) Tính kỳ vọng, phương sai, giá trị tin chắc và trung vị của X.

Bài 2: Gieo cùng lúc 2 con xúc xắc. Gọi X là đại lượng ngẫu nhiên chỉ tổng số chấm xuất hiện.

a) Lập bảng phân phối xác suất của X.

b) Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X), \mathrm{Med}(X)$.

Bài 3: Có 2 hộp đựng bi: Hộp I có 3 bi xanh và 7 bi đỏ, hộp II có 4 bi xanh và 6 bi đỏ. Một người lấy ngẫu nhiên 1 bi từ hộp I bỏ vào hộp II, sau đó lấy ra ngẫu nhiên đồng thời 3 bi từ hộp II. Gọi X là đại lượng ngẫu nhiên chỉ số bi xanh có trong 3 bi được lấy ra từ hộp II.

a) Lập bảng phân phối xác suất của X.

b) Tìm hàm phân phối xác suất của X và vẽ đồ thị của nó.

c) Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X), \mathrm{Med}(X)$.

Bài 4: Ba xạ thủ cùng bắn độc lập vào một mục tiêu. Xác suất bắn trúng tương ứng là $0,6$; $0,7$ và $0,8$. Biết rằng, mỗi xạ thủ chỉ bắn một viên.

a) Lập bảng phân phối xác suất của số viên trúng.

b) Tính xác suất có ít nhất 2 viên trúng.

c) Tìm số viên trúng mục tiêu tin chắc nhất, số viên trúng mục tiêu trung bình và phương sai của số viên trúng.

Bài 5: Có 2 lô hàng, mỗi lô chứa 10 sản phẩm. Lô A có 6 sản phẩm loại I và 4 sản phẩm loại II, lô B có 3 sản phẩm loại I và 7 sản phẩm loại II. Lấy ngẫu nhiên đồng thời 2 sản phẩm từ lô A và 3 sản phẩm từ lô B.

a) Lập bảng phân phối xác suất của đại lượng ngẫu nhiên X chỉ số sản phẩm loại I có trong 5 sản phẩm được lấy ra.

b) Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X), \mathrm{Med}(X)$.

Bài 6: Cho đại lượng ngẫu nhiên X có hàm phân phối xác suất như sau:

$$F(x) = \begin{cases} 0 & \text{nếu } x < 3, \\ (x-3)^2 & \text{nếu } 3 \le x \le 4, \\ 1 & \text{nếu } 4 < x. \end{cases}$$

a) Tìm hàm mật độ xác suất $f(x)$ và vẽ đồ thị của nó.

b) Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X), \mathrm{Med}(X)$.

Bài 7: Cho hàm số:

$$f(x) = \begin{cases} a(x-3)(5-x) & \text{nếu } x \in [3;5], \\ 0 & \text{nếu } x \notin [3;5]. \end{cases}$$

a) Tìm a để $f(x)$ là hàm mật độ xác suất và tìm $F(x)$.

b) Tính $P(X > 4)$.

c) Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X), \mathrm{Med}(X)$.

Bài 8: Cho đại lượng ngẫu nhiên X có hàm mật độ xác suất:

$$f(x) = \begin{cases} 0 & \text{nếu } x \notin [0;2], \\ kx^2 & \text{nếu } x \in (0;1), \\ k(2-x)^2 & \text{nếu } x \in [1;2). \end{cases}$$

a) Tìm k và $F(x)$.

b) Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X), \mathrm{Med}(X)$.

Bài 9: Cho hàm số:

$$f(x) = \begin{cases} a\cos x & \text{nếu } x \in (-\pi/2; \pi/2), \\ 0 & \text{nếu } x \notin (-\pi/2; \pi/2). \end{cases}$$

a) Tìm a để $f(x)$ là hàm mật độ và tìm $F(x)$.

b) Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X), \mathrm{Med}(X)$.

Bài 10: Cho đại lượng ngẫu nhiên X có bảng phân phối xác suất như sau:

X	-1	0	1	2
P	$0,1$	$0,25$	$0,3$	$0,35$

a) Lập bảng phân phối xác suất cho $|X|$. Tính $\mathrm{E}(|X|), \mathrm{Var}(|X|)$.

b) Lập bảng phân phối xác suất cho X^2. Tính $\mathrm{E}(X^2), \mathrm{Var}(X^2)$.

c) Lập bảng phân phối xác suất cho $Y = -2X^2 + 3X + 1$. Tính $\mathrm{E}(Y), \mathrm{Var}(Y)$.

Bài 11: Gọi X, Y là lợi nhuận thu được khi đầu tư 10 triệu đồng cho từng dự án và có bảng phân phối xác suất lần lượt như sau:

X (triệu đồng)	-10	0	10	20
P	$0,2$	$0,3$	$0,3$	$0,2$

Y (triệu đồng)	-5	0	5	10	15
P	$0,1$	$0,2$	$0,25$	$0,25$	$0,2$

a) Nếu phải chọn một trong hai dự án để đầu tư thì bạn chọn dự án nào? Tại sao?

b) Lập bảng phân phối xác suất của $Z = 2X+3Y+1$. Tính $\mathrm{E}(Z), \mathrm{Var}(Z), \mathrm{Mod}(Z), \mathrm{Med}(Z)$.

Bài 12: Một hộp chứa 10 viên bi. Trong đó có 4 bi mang số 1, 3 bi mang số 2 và 3 bi mang số 3. Lấy ra ngẫu nhiên đồng thời 3 viên bi. Gọi X là số bi mang số 1, Y là số bi mang số 2 có trong 3 bi được lấy ra.

a) Lập bảng phân phối xác suất đồng thời của X và Y. Kiểm tra xem X và Y có độc lập với nhau hay không?

b) Lập bảng phân phối xác suất biên của X, của Y. Tính kỳ vọng và phương sai của X và Y.

c) Lập bảng phân phối xác suất của X với điều kiện $Y = 1$.

d) Tính hiệp phương sai, hệ số tương quan của (X, Y).

Bài 13: Cho X và Y là các đại lượng ngẫu nhiên độc lập có các bảng phân phối xác suất như sau:

X	-3	0	3	6
P	$0,1$	$0,2$	$0,4$	$0,3$

Y	-1	0	1	2
P	$0,2$	$0,2$	$0,3$	$0,3$

a) Lập bảng phân phối xác suất của $Z = X \cdot Y$ và tính $\mathrm{E}(Z), \mathrm{Var}(Z)$.

b) Lập bảng phân phối xác suất của $T = X + Y$ và tính $\mathrm{E}(T), \mathrm{Var}(T)$.

c) Lập bảng phân phối xác suất của $V = 2X + Y - 1$ và tính $\mathrm{E}(V), \mathrm{Var}(V)$.

Bài 14: Cho bảng phân phối xác suất đồng thời của (X, Y) như sau:

$X \backslash Y$	-1	0	1	2
1	$0,1$		$0,1$	
0		$0,2$	$0,3$	$0,1$
3	$0,08$	$0,02$		$0,1$

a) Tìm các phân phối biên của X và của Y. Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X), \mathrm{Med}(X)$ và $\mathrm{E}(Y)$, $\mathrm{Var}(Y)$, $\mathrm{Mod}(Y)$, $\mathrm{Med}(Y)$.

b) Lập bảng phân phối xác suất của Y với điều kiện $X = 0$.

c) Lập bảng phân phối xác suất của $Z = 3X - 4Y + 6$.

d) Tính hiệp phương sai và hệ số tương quan của (X, Y).

Chương 4

MỘT SỐ PHÂN PHỐI THÔNG DỤNG

4.1. PHÂN PHỐI RỜI RẠC THÔNG DỤNG

4.1.1. Phân phối nhị thức

Bài toán: Gọi A_k là biến cố có k lần thành công khi thực hiện dãy phép thử Bernoulli, là dãy n phép thử độc lập có hai biến cố sơ cấp là S_thành công và F_thất bại với xác suất tương ứng là p, q trong đó $p + q = 1$. Tính xác suất của A_k.

Giải: Theo công thức Bernoulli: $P(A_k) = B(k, n, p) = C_n^k p^k q^{n-k} \quad (k = 0, 1, \ldots, n)$.

Định nghĩa 4.1. Đại lượng ngẫu nhiên rời rạc X được gọi là có phân phối nhị thức nếu xác suất X nhận giá trị bằng k được xác định bởi:

$$P(X = k) = B(k, n, p) = C_n^k p^k q^{n-k} \quad (k = 0, 1, \ldots, n).$$

Ký hiệu: $X \sim B(n; p)$.

Ta thấy định nghĩa trên thỏa mãn điều kiện của một phân phối xác suất vì:

$$\sum_{k=0}^{n} P(X = k) = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p + q)^n = 1.$$

Chú ý: *i*) Điều kiện quan trọng để có phân phối nhị thức là các phép thử phải độc lập với nhau và xác suất của biến cố S trong mỗi lần thử là không đổi. Nếu phép thử là lấy phần tử từ một tập hợp thì lấy có hoàn lại sẽ đảm bảo tính độc lập hoặc lấy ít từ một tập có số phần tử rất lớn thì cũng xem như độc lập.

ii) Khi $n = 1$, X được gọi là phân phối *không - một* và có bảng phân phối xác suất như sau:

X	0	1
P	q	p

Ký hiệu $X \sim A(p)$. Khi đó: $\mathrm{E}(X) = p, \mathrm{Var}(X) = pq$.

Số đặc trưng: Cho $X \sim B(n; p)$, ta có:

i) $\mathrm{E}(X) = np$.

ii) $\mathrm{Var}(X) = npq$.

iii) $np - q \leq \mathrm{Mod}(X) \leq np + p$.

Chứng minh. Gọi X_i là số lần xuất hiện S trong phép thử Bernoulli thứ i thì các X_i độc lập với nhau và $X_i \sim A(p)$ $(i = 1, \dots, n)$, $\mathrm{E}(X_i) = p, \mathrm{Var}(X_i) = pq$.

Ta có: $X = X_1 + X_2 + \cdots + X_n$. Nên:
$\mathrm{E}(X) = \mathrm{E}(X_1 + X_2 + \cdots + X_n) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n) = np$,
$\mathrm{Var}(X) = \mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) = npq$.

Giả sử $\mathrm{Mod}(X) = k$, ta có:

$$\frac{P(X=k)}{P(X=k-1)} = \frac{C_n^k p^k q^{n-k}}{C_n^{k-1} p^{k-1} q^{n-k+1}} = \frac{p(n-k+1)}{qk} \geq 1 \Leftrightarrow k \leq np + p,$$

$$\frac{P(X=k)}{P(X=k+1)} = \frac{C_n^k p^k q^{n-k}}{C_n^{k+1} p^{k+1} q^{n-k-1}} = \frac{q(k+1)}{p(n-k)} \geq 1 \Leftrightarrow k \geq np - q.$$

Từ đó suy ra iii). □

Ví dụ 4.1. Một máy sản xuất ra một loại sản phẩm với tỉ lệ phế phẩm là $0,1$. Cho máy sản xuất ra 3 sản phẩm. Gọi X là số phế phẩm có trong 3 sản phẩm đó.

a) Lập bảng phân phối xác suất của X.

b) Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X), \mathrm{Med}(X)$.

Giải:

a) Ta thấy: $X \sim B(3; 0,1)$. Do đó:
$P(X=0) = C_3^0(0,1)^0(0,9)^3 = 0,729; P(X=1) = C_3^1(0,1)^1(0,9)^2 = 0,243.$
$P(X=2) = C_3^2(0,1)^2(0,9)^1 = 0,027; P(X=3) = C_3^3(0,1)^3(0,9)^0 = 0,001.$

Nên bảng phân phối xác suất của X là:

X	0	1	2	3
P	$0,729$	$0,243$	$0,027$	$0,001$

b) Ta có: $\mathrm{E}(X) = np = 3 \cdot 0,1 = 0,3$; $\mathrm{Var}(X) = npq = 3 \cdot 0,1 \cdot 0,9 = 0,27$; $\mathrm{Med}(X) = 0$. Do $np - q = -0,6$ và $np + p = 0,4$ nên $\mathrm{Mod}(X) = 0$.

Ví dụ 4.2. Một công ty có 10 cửa hàng đặt ở các địa điểm khác nhau. Xác suất bán được hàng trong ngày ở mỗi nơi là như nhau và bằng $0,3$.

a) Tìm xác suất công ty đó bán được hàng trong 1 ngày.

b) Mỗi năm công ty đó bán hàng 300 ngày, tìm số ngày bán được hàng nhiều khả năng nhất trong 1 năm.

Giải:

a) Gọi X là số nơi bán được hàng trong 1 ngày, $X \sim B(10; 0,3)$. Xác suất công ty đó bán được hàng trong 1 ngày là:

$$P(X \geq 1) = 1 - P(X=0) = 1 - C_{10}^0(0,3)^0(0,7)^{10} = 0,972.$$

b) Gọi Y là số ngày bán được hàng trong 1 năm, $Y \sim B(300; p)$ với $p = 0,972$. Số ngày bán được hàng nhiều khả năng nhất là $\mathrm{Mod}(Y)$ với

$$\begin{aligned} 300p - q \leq \mathrm{Mod}(Y) \leq 300p + p \quad &\Leftrightarrow 300 \cdot 0,972 - 0,028 \leq \mathrm{Mod}(Y) \leq 300 \cdot 0,972 + 0,972 \\ &\Leftrightarrow 291,57 \leq \mathrm{Mod}(Y) \leq 292,57. \end{aligned}$$

Suy ra: $\mathrm{Mod}(Y) = 292$.

4.1.2. Phân phối hình học

Bài toán: Thực hiện dãy phép thử Bernoulli cho đến khi nào thành công thì dừng. Gọi A_k là biến cố phép thử dừng ở lần thứ k. Khi đó, xác suất của A_k được cho bởi công thức: $P(A_k) = pq^{k-1} \quad (k = 1, 2, \ldots)$.

Định nghĩa 4.2. Đại lượng ngẫu nhiên rời rạc X được gọi là có phân phối hình học (hay phân phối Pascal) nếu xác suất X nhận giá trị bằng k được xác định bởi:

$$P(X = k) = pq^{k-1} \quad (k = 1, 2, \ldots).$$

Ký hiệu: $X \sim Pa(p)$.

Ta thấy định nghĩa trên là một phân phối xác suất vì:

$$\sum_{k=1}^{+\infty} P(X = k) = \sum_{k=1}^{+\infty} pq^{k-1} = \lim_{n\to+\infty} p\left(\sum_{k=1}^{n} q^{k-1}\right) = \lim_{n\to+\infty} p\left(\frac{1-q^{n+1}}{1-q}\right) = 1.$$

Số đặc trưng: Cho $X \sim Pa(p)$, khi đó:

i) $\mathrm{E}(X) = \dfrac{1}{p}$.

ii) $\mathrm{Var}(X) = \dfrac{q}{p^2}$.

iii) $\mathrm{Mod}(X) = 1$.

Chứng minh. Ta có:

$$\mathrm{E}(X) = \sum_{k=1}^{+\infty} kq^{k-1}p = \sum_{k=1}^{+\infty} q^{k-1}p + \sum_{k=1}^{+\infty}(k-1)q^{k-1}p = 1 + q\mathrm{E}(X) \Rightarrow p\mathrm{E}(X) = 1 \Leftrightarrow \mathrm{E}(X) = \frac{1}{p}.$$

$$\mathrm{E}(X^2) = \sum_{k=1}^{+\infty} k^2q^{k-1}p = \sum_{k=1}^{+\infty}(k-1)^2q^{k-1}p + \sum_{k=1}^{+\infty} 2kq^{k-1}p - \sum_{k=1}^{+\infty} q^{k-1}p = q\mathrm{E}(X^2) + \frac{2}{p} - 1.$$

$$\Rightarrow p\mathrm{E}(X^2) = \frac{1+q}{p} \Leftrightarrow \mathrm{E}(X^2) = \frac{1+q}{p^2}.$$

Do đó: $\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \dfrac{1+q}{p^2} - \dfrac{1}{p^2} = \dfrac{q}{p^2}$. □

Ví dụ 4.3. Một người mỗi ngày mua 1 tờ vé số cho đến khi nào trúng độc đắc thì dừng.

a) Tính xác suất để trong cuộc đời người này được trúng độc đắc, giả sử người này sống được 30000 ngày.

b) Tính số ngày trung bình người này phải mua để được trúng độc đắc.

Giải:

a) Gọi X là đại lượng ngẫu nhiên chỉ số ngày người này mua vé số cho đến khi trúng độc đắc. Ta thấy: $X \sim Pa(10^{-6})$.

Biến cố người này trúng độc đắc trong đời chính là $X \leq 30000$. Ta có:

$$P(X \leq 30000) = \sum_{k=1}^{30000} p \cdot q^{1-k} = p \cdot \frac{1 - q^{30000}}{1 - q} = 1 - (1 - 10^{-6})^{30000} = 0,0296.$$

b) Số ngày trung bình người này phải mua để được trúng độc đắc chính là kỳ vọng E(X):

$$\mathrm{E}(X) = \frac{1}{10^{-6}} = 1000000 \text{ (ngày)} \approx 2740 \text{ (năm)}.$$

4.1.3. Phân phối siêu bội

Bài toán: Giả sử một hộp chứa N viên bi, trong đó có M viên bi màu trắng còn lại là bi màu đen. Từ hộp đó lấy ra ngẫu nhiên đồng thời n $(n \leq M)$ viên bi. Khi đó, xác suất của biến cố A_k_có k bi màu trắng trong n lấy ra là: $P(A_k) = \dfrac{C_M^k \cdot C_{N-M}^{n-k}}{C_N^n} \quad (k = 0, 1, \ldots, n)$.

Định nghĩa 4.3. Đại lượng ngẫu nhiên rời rạc X được gọi là có phân phối siêu bội nếu xác suất X nhận giá trị bằng k được xác định bởi:

$$P(X = k) = \frac{C_M^k \cdot C_{N-M}^{n-k}}{C_N^n} \quad (k = 0, 1, \ldots, n).$$

Ký hiệu: $X \sim H(N; M; n)$.

Vì $(1 + x)^N = (1 + x)^M (1 + x)^{N-M}$ nên khi khai triển nhị thức Newton thì số hạng chứa x_n ở hai vế là như nhau:

$$C_N^n = \sum_{k=0}^{n} C_M^k C_{N-M}^{n-k} \Leftrightarrow \sum_{k=0}^{n} \frac{C_M^k C_{N-M}^{n-k}}{C_N^n} \Leftrightarrow \sum_{k=0}^{n} P(X = k) = 1.$$

Vậy nên định nghĩa trên hoàn toàn thỏa điều kiện của một phân phối xác suất.

Chú ý: Phân phối siêu bội được áp dụng khi ta lấy ngẫu nhiên đồng thời n phần tử từ một tập hợp. Tuy nhiên, nếu ta thay đổi cách lấy là lần lượt có hoàn lại thì luật phân phối lúc này là phân phối nhị thức. Nhưng nếu tập N có số phần tử lớn hơn rất nhiều so với n thì xác suất của hai luật phân phối này gần như bằng nhau.

Định lí 4.1 (Xấp xỉ phân phối siêu bội bằng phân phối nhị thức)**.** *Nếu $X \sim H(N; M; n)$, n cố định, $N \longrightarrow +\infty$ và $\dfrac{M}{N} \longrightarrow p$ thì $X \longrightarrow B(n; p)$. Nghĩa là:*

$$\lim_{N \to +\infty} \frac{C_M^k \cdot C_{N-M}^{n-k}}{C_N^n} = C_n^k p^k q^{n-k}.$$

Chứng minh. Ta có:

$$\begin{aligned}
\lim_{N\to+\infty} \frac{C_M^k \cdot C_{N-M}^{n-k}}{C_N^n} &= \lim_{N\to+\infty} \frac{M!}{k!(M-k)!} \cdot \frac{(N-M)!}{(n-k)!(N-M-n-k)!} \cdot \frac{n!(N-n)!}{N!} \\
&= \lim_{N\to+\infty} \frac{n!}{k!(n-k)!} \cdot \frac{M!(N-n)!}{(M-k)!N!} \cdot \frac{(N-M)!}{(N-M-n-k)!} \\
&= C_n^k \lim_{N\to+\infty} \frac{(M-k+1)\cdots M}{(N-n+1)\cdots(N-n+k)} \cdot \lim_{N\to+\infty} \frac{(N-M-n-k+1)\cdots(N-M)}{(N-n+k+1)\cdots N} \\
&= C_n^k \left(\frac{M}{N}\right)^k \left(\frac{N-M}{N}\right)^{n-k} = C_n^k p^k q^{n-k}.
\end{aligned}$$

□

Trong thực hành khi $\dfrac{N}{n} > 20$ có thể *xấp xỉ* phân phối siêu bội bằng phân phối nhị thức.

Số đặc trưng: Cho $X \sim H(N; M; n)$, khi đó:

i) $\mathrm{E}(X) = np \quad \left(\text{với } p = \dfrac{M}{N}\right)$.

ii) $\mathrm{Var}(X) = npq\dfrac{N-n}{N-1} \quad (\text{với } q = 1-p)$.

iii) $\dfrac{Mn+M+n+1}{N+2} - 1 \leq \mathrm{Mod}(X) \leq \dfrac{Mn+M+n+1}{N+2}$.

Chứng minh. Ta có:

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{k=0}^{n} \frac{kC_M^k C_{N-M}^{n-k}}{C_N^n} = \sum_{k=1}^{n} \frac{MC_{M-1}^{k-1} C_{N-M}^{n-k}}{C_N^n} = \sum_{k=0}^{n-1} \frac{MC_{M-1}^{k} C_{N-M}^{(n-1)-k}}{C_N^n} \\
&= \frac{Mn}{N}\left(\sum_{k=0}^{n-1} \frac{C_{M-1}^{k} C_{(N-1)-(M-1)}^{(n-1)-k}}{C_(N-1)^(n-1)}\right) = \frac{Mn}{N}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(X^2) &= \sum_{k=0}^{n} \frac{k^2 C_M^k C_{N-M}^{n-k}}{C_N^n} = \sum_{k=1}^{n} \frac{k(k-1)C_M^k C_{N-M}^{n-k}}{C_N^n} = \sum_{k=0}^{n} \frac{kC_M^k C_{N-M}^{n-k}}{C_N^n} \\
&= \frac{M(M-1)n(n-1)}{N(N-1)}\left(\sum_{k=2}^{n} \frac{C_{M-2}^{k-2} C_{(N-2)-(M-2)}^{(n-2)-(k-2)}}{C_(N-2)^(n-2)}\right) + \frac{Mn}{N}.
\end{aligned}$$

$$\Rightarrow \mathrm{Var}(X) = \frac{M(M-1)n(n-1)}{N(N-1)} + \frac{Mn}{N} - \left(\frac{Mn}{N}\right)^2 = n\frac{M(N-M)(N-n)}{N^2(N-1)}.$$

Giả sử $\mathrm{Mod}(X) = k$, ta có:

$$\frac{P(X=k)}{P(X=k-1)} = \frac{(n-k+1)(M-k+1)}{k(N-M-n+k)} \geq 1 \Rightarrow k \leq \frac{Mn+M+n+1}{N+2}.$$

$$\frac{P(X=k)}{P(X=k+1)} = \frac{(k+1)(N-M-n+k+1)}{(M-k)(n-k)} \geq 1 \Rightarrow k \geq \frac{Mn+M+n+1}{N+2} - 1.$$

Từ đó suy ra iii). □

Ví dụ 4.4. Một hộp chứa 10 bóng đèn, trong đó có 4 bóng bị hỏng. Lấy ngẫu nhiên từ trong hộp ra 3 bóng. Gọi X là đại lượng ngẫu nhiên chỉ số bóng đèn hỏng có trong 3 bóng lấy ra.

a) Lập bảng phân phối xác suất của X.

b) Tính $\mathrm{E}(X), \mathrm{Var}(X), \mathrm{Mod}(X)$ và $\mathrm{Med}(X)$.

<u>Giải:</u>

a) Dễ thấy: $X \sim H(N=10; M=4; n=3)$. Nên:
$P(X=0) = \dfrac{C_4^0 C_6^3}{C_{10}^3} = \dfrac{1}{6}, P(X=1) = \dfrac{C_4^1 C_6^2}{C_{10}^3} = \dfrac{1}{2},$
$P(X=2) = \dfrac{C_4^2 C_6^1}{C_{10}^3} = \dfrac{3}{10}, P(X=3) = \dfrac{C_4^3 C_6^0}{C_{10}^3} = \dfrac{1}{30}.$

Bảng phân phối xác suất của X:

X	0	1	2	3
P	1/6	1/2	3/10	1/30

b) Ta có: $\mathrm{E}(X) = np = 3 \cdot \dfrac{4}{10} = 1,2$; $\mathrm{Var}(X) = npq\dfrac{N-n}{N-1} = 3 \cdot 0,4 \cdot 0,6 \cdot \dfrac{10-3}{10-1} = 0,56$; $\mathrm{Mod}(X) = 1$ và $\mathrm{Med}(X) = 1$.

4.1.4. Phân phối Poisson

Bài toán: Quan sát hiện tượng sợi tơ bị đứt khi quay sợi, ta thấy sợi đứt tại những thời điểm ngẫu nhiên. Xét thí nghiệm ngẫu nhiên: "Số lần sợi bị đứt diễn ra từ lúc a đến lúc b". Biết trung bình có λ lần sợi đứt trong thời đoạn (a, b), tính xác suất của biến cố A_k_có k lần sợi đứt trong thời đoạn (a, b)?

<u>Giải:</u> Để giải bài toán ta phân hoạch (a, b) thành n đoạn nhỏ bằng nhau: $I_1, I_2, \ldots, I_n$. Ta thấy rằng trong một thời đoạn vô cùng nhỏ, khả năng xuất hiện nhiều hơn một điểm đứt là không đáng kể so với khả năng xuất hiện đúng một điểm đứt.

Xem "Số điểm đứt xuất hiện trong thời đoạn I_j" là kết quả của phép thử thứ j. Gọi p_n là xác suất có điểm đứt trên đoạn I_j (không phụ thuộc vào j). Khi đó, p_n cũng có thể xem là xác suất có đúng một điểm đứt xuất hiện trên I_j.

Ta có n phép thử độc lập của cùng một thí nghiệm ngẫu nhiên với xác suất xuất hiện điểm đứt $p_n = \dfrac{\lambda}{n}$ là như nhau.

Do đó: $P(A_k) \approx B(k, n, p_n)$.

Giá trị trên càng chính xác khi $n \to +\infty$, tức là:

$$P(A_k) = \lim_{n\to+\infty} C_n^k p_n^k q_n^{n-k} = \lim_{n\to+\infty} \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \lim_{n\to+\infty} \frac{\lambda^k}{k!}\frac{(n-k+1)\cdots n}{n!}\cdot\left[\left(1-\frac{\lambda}{n}\right)^{\frac{n}{\lambda}}\right]^{\frac{(n-k)\lambda}{n}} = \frac{\lambda^k}{k!}e^{-\lambda} \quad (k=0,1,\ldots).$$

Định nghĩa 4.4. Đại lượng ngẫu nhiên rời rạc X được gọi là có luật phân phối Poisson nếu xác suất X nhận giá trị bằng k được xác định bởi:

$$P(X=k) = \frac{e^{-\lambda}\cdot\lambda^k}{k!} \quad (\lambda>0, k=0,1,\ldots).$$

Ký hiệu: $X \sim P(\lambda)$.

Số đặc trưng: Cho $X \sim P(\lambda)$, khi đó:

i) $\mathrm{E}(X) = \lambda$.

ii) $\mathrm{Var}(X) = \lambda$.

iii) $\lambda - 1 \leq \mathrm{Mod}(X) \leq \lambda$.

Chứng minh. Ta có:

$$\mathrm{E}(X) = \sum_{k=0}^{+\infty} e^{-\lambda}\frac{k\lambda^k}{k!} = \lambda e^{-\lambda}\left(\sum_{k=1}^{+\infty}\frac{\lambda^{k-1}}{(k-1)!}\right) = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

$$\mathrm{E}(X^2) = \sum_{k=0}^{+\infty} e^{-\lambda}\frac{k^2\lambda^k}{k!} = \sum_{k=1}^{+\infty} e^{-\lambda}\frac{k(k-1)\lambda^k}{k!} + \sum_{k=1}^{+\infty} e^{-\lambda}\frac{k\lambda^k}{k!} = \lambda^2 e^{-\lambda}\left(\sum_{k=2}^{+\infty}\frac{\lambda^{k-2}}{(k-2)!}\right) + \lambda = \lambda^2 + \lambda.$$

Nên: $\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Giả sử $\mathrm{Mod}(X) = k$, ta có:

$$\frac{P(X=k)}{P(X=k-1)} = \frac{\lambda}{k} \geq 1 \Leftrightarrow k \leq \lambda.$$

$$\frac{P(X=k)}{P(X=k+1)} = \frac{k+1}{\lambda} \geq 1 \Leftrightarrow k \geq \lambda - 1.$$

Từ đó suy ra iii). □

Chú ý: Phân phối Poisson được áp dụng khi:

i) Số lần xuất hiện biến cố A trong khoảng thời gian (a,b) không ảnh hưởng tới xác suất xuất hiện A trong khoảng thời gian kế tiếp.

ii) Số lần xuất hiện biến cố A trong khoảng thời gian (a,b) tỉ lệ với độ dài của khoảng thời gian đó.

Ngoài ra, ta có thể lấy gần đúng xác suất của phân phối nhị thức bằng phân phối Poisson nhờ định lí sau:

Định lí 4.2 (Xấp xỉ phân phối nhị thức bằng phân phối Poisson). *Cho $X_n \sim B(n;p)$ nếu $n \longrightarrow +\infty$ có $p \longrightarrow 0$ nhưng $np \longrightarrow \lambda$ thì $X \longrightarrow P(\lambda)$. Nghĩa là:*

$$\lim_{n\to+\infty} C_n^k p^k q^{n-k} = \frac{e^{-\lambda}\lambda!}{k!}.$$

Chứng minh. Xem *Bài toán* đã nêu ở trên. □

Ví dụ 4.5. Số khách đến mua hàng tại một quầy hàng là ngẫu nhiên, độc lập; trung bình cứ 3 phút có 1 người.

a) Tính xác suất quầy hàng có 2 khách trong 30 giây.

b) Tính số khách hàng nhiều khả năng nhất trong 30 giây.

Giải:

a) Gọi X là số khách hàng đến quầy trong 30 giây, $X \sim P(\lambda)$ với $\lambda = \dfrac{1}{6}$. Ta có:

$$P(X=2) = \frac{e^{-\frac{1}{6}}}{2!}\left(\frac{1}{6}\right)^2 = 0,0118.$$

b) Do $\dfrac{1}{6} - 1 \le \mathrm{Mod}(X) \le \dfrac{1}{6}$ nên $\mathrm{Mod}(X) = 0$.

4.2. PHÂN PHỐI LIÊN TỤC THÔNG DỤNG

4.2.1. Phân phối mũ

Bài toán: Với giả thiết như ở bài toán dẫn đến phân phối Poisson. Nhưng ta tính xác suất của biến cố: "Khoảng cách giữa hai lần sợi bị đứt kế nhau nằm trong khoảng (a,b)".

Giải: Gọi X là đại lượng ngẫu nhiên chỉ khoảng cách giữa hai lần sợi bị đứt liền nhau. Dễ thấy X là đại lượng ngẫu nhiên liên tục, ta sẽ tìm hàm mật độ xác suất $f(x)$ sao cho:

$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x.$$

Ta chia đoạn $(0,x)$ ra thành n đoạn con bằng nhau $I_1, I_2, \ldots, I_n$. Suy ra, độ dài mỗi đoạn là $\Delta x = \dfrac{x}{n}$. Số điểm đứt trung bình trên mỗi đoạn con là $\lambda_n = \dfrac{\lambda x}{n(b-a)}$.

Xem biến cố "Xuất hiện điểm đứt trong thời đoạn I_j" là kết quả của phép thử thứ j. Trong một thời đoạn vô cùng nhỏ thì khả năng xuất hiện nhiều hơn một điểm đứt là không đáng kể so với khả năng xuất hiện đúng một điểm đứt.

Ta có n phép thử độc lập với xác suất để phép thử thành công là như nhau tuân theo luật phân phối Poisson. Do đó, xác suất để mỗi đoạn con bị đứt là $\begin{cases} p_n = \lambda_n e^{-\lambda_n}, \\ q_n = e^{-\lambda_n}. \end{cases}$

Biến cố sợi bị đứt tại thời đoạn $(x-\Delta x, x)$ tuân theo luật phân phối hình học $Pa(p_n)$ nên $P(x-\Delta x < X < x) \approx q_n^{n-1}p_n$.

Mà:

$$q_n^{n-1}p_n = \lambda_n e^{-n\lambda_n} = \frac{\lambda x}{n(b-a)}e^{-\frac{\lambda x}{b-a}}.$$

Do đó:

$$P(0 < X < x) \approx \sum_{j=1}^{n} \frac{\lambda}{b-a}e^{-\frac{\lambda x_j}{b-a}} \cdot \Delta x = \sum_{j=0}^{n} \lambda' e^{\lambda' x_j} \cdot \Delta x \quad (\text{với } \lambda' = \frac{\lambda}{b-a}, x_j \in I_j).$$

Giá trị trên càng chính xác khi $n \to +\infty$, tức là:

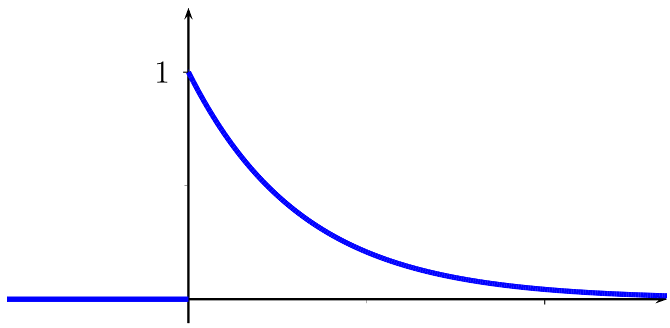
$$P(0 < X < x) = \int_0^x \lambda' e^{-\lambda' t}\mathrm{d}t.$$

Suy ra: $f(x) = \begin{cases} \lambda' e^{-\lambda' x} & \text{nếu } x \geq 0, \\ 0 & \text{nếu } x < 0. \end{cases}$

Định nghĩa 4.5. Đại lượng ngẫu nhiên liên tục X được gọi là có phân phối mũ nếu hàm mật độ xác suất của nó có dạng:

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

Ký hiệu: $X \sim E(\lambda)$.



Hình 4.1: Đồ thị hàm $f(x)$ của phân phối mũ với $\lambda = 1$

Ta thấy định nghĩa trên thỏa mãn điều kiện của một phân phối xác suất vì:

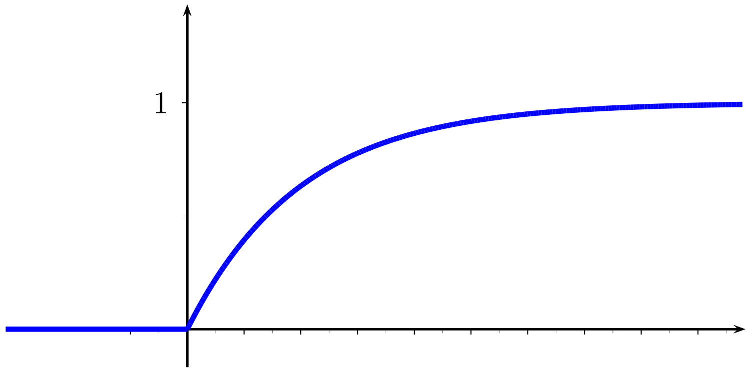
$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \int_0^{+\infty} \lambda e^{-\lambda x}\mathrm{d}x = \lim_{a\to+\infty} \left[-e^{-\lambda x}\right]_0^a = 1.$$

Ngoài ra:

$$\int_0^x \lambda e^{-\lambda t}\mathrm{d}t = -e^{-\lambda t}\Big|_0^a = 1 - e^{-\lambda x}.$$

Nên X có hàm phân phối xác suất như sau:

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



Hình 4.2: Đồ thị hàm $F(x)$ của phân phối mũ với $\lambda = 1$

Số đặc trưng: Cho $X \sim E(X)$, ta có:

i) $\mathrm{E}(X) = \dfrac{1}{\lambda}$.

ii) $\mathrm{Var}(X) = \dfrac{1}{\lambda^2}$.

iii) $\mathrm{Mod}(X) = 0$.

Chứng minh. Ta có:

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x = \int_0^{+\infty} \lambda x e^{-\lambda x}\mathrm{d}x = \lim_{a\to+\infty}\left[-xe^{-\lambda x} - \frac{e^{-\lambda x}}{\lambda}\right]_0^a = \frac{1}{\lambda}.$$

$$\mathrm{E}(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)\mathrm{d}x = \int_0^{+\infty} \lambda x^2 e^{-\lambda x}\mathrm{d}x = \lim_{a\to+\infty}\left[-x^2e^{-\lambda x}\right]_0^a + \frac{2}{\lambda}\int_0^{+\infty}\lambda x e^{-\lambda x}\mathrm{d}x = \frac{2}{\lambda^2}.$$

Do đó: $\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \dfrac{2}{\lambda^2} - \dfrac{1}{\lambda^2} = \dfrac{1}{\lambda^2}$. □

Nếu X là tuổi thọ của một thiết bị thì tính chất quên (memoryless) nói rằng: Tuổi thọ tiếp tục của một thiết bị t ($\forall t > 0$) tuổi có phân phối xác suất như tuổi thọ của một thiết bị mới. Do đó tuổi thọ của một thiết bị có phân phối mũ và chỉ có phân phối mũ mới có tính chất này. Bởi vì: $e^{-\lambda(s+t)} = e^{-\lambda s} + e^{-\lambda t}$.

Nếu số lần xuất hiện của một biến cố trong một khoảng thời gian có luật phân phối Poisson thì thời gian giữa hai lần xuất hiện biến cố đó sẽ có phân phối mũ. Chẳng hạn, thời gian làm việc liên tục của một thiết bị giữa hai lần sửa chữa.

Ví dụ 4.6. Tuổi thọ X (năm) của một mạch điện tử trong máy tính là biến ngẫu nhiên có phân phối mũ với trung bình $6,25$. Thời gian bảo hành mạch điện tử là 5 năm. Tính tỉ lệ mạch điện tử bán ra phải thay thế.

Giải: Ta có: $\lambda = \dfrac{1}{\mathrm{E}(X)} = \dfrac{1}{6,25} = 0,16$. Vậy nên:

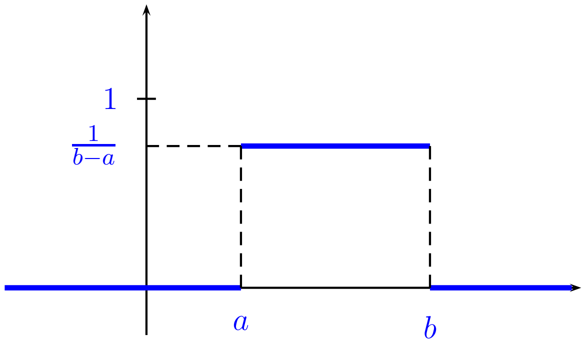
$$P(X \le 5) = \int_0^5 0,16\cdot e^{-0,16x}\mathrm{d}x = -e^{-0,16x}\Big|_0^5 = 0,551.$$

4.2.2. Phân phối đều

Định nghĩa 4.6. Đại lượng ngẫu nhiên liên tục X được gọi là có phân phối đều trong khoảng (a, b) nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{với } x \in [a, b], \\ 0 & \text{với } x \notin [a, b]. \end{cases}$$

Ký hiệu: $X \sim U(a, b)$.

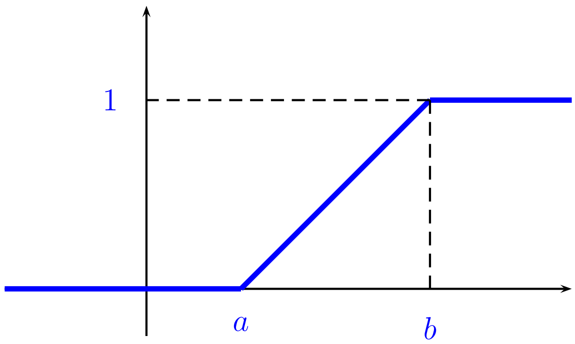


Hình 4.3: Đồ thị hàm $f(x)$ của phân phối đều

Ta thấy định nghĩa trên thỏa mãn điều kiện của một phân phối xác suất vì:

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \int_a^b \frac{1}{b-a}\mathrm{d}x = 1.$$

Hàm phân phối của X có dạng: $F(x) = \begin{cases} 0 & \text{nếu } x \leq a, \\ \dfrac{x-a}{b-a} & \text{nếu } x \in (a, b), \\ 1 & \text{nếu } x \geq b. \end{cases}$



Hình 4.4: Đồ thị hàm $F(x)$ của phân phối đều

Số đặc trưng: Cho $X \sim U(a, b)$, ta có:

i) $\mathrm{E}(X) = \dfrac{a+b}{2}$.

ii) $\mathrm{Var}(X) = \dfrac{(b-a)^2}{12}$.

iii) $\mathrm{Mod}(X) = x_0 \quad (\forall x_0 \in [a, b])$.

Chứng minh. Ta có:

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x f(x)\mathrm{d}x = \int_a^b \frac{x}{b-a}\mathrm{d}x = \frac{a+b}{2},$$

$$\mathrm{Var}(X) = \int_a^b \frac{x^2}{b-a}\mathrm{d}x - \left(\frac{a+b}{2}\right)^2 = \frac{b^3-a^3}{3(b-a)} - \left(\frac{a+b}{2}\right)^2 = \frac{(a-b)^2}{12}.$$

□

Ví dụ 4.7. Nếu bạn đến trạm xe buýt lúc 10 giờ và biết rằng thời gian xe buýt xuất hiện tại trạm từ 10 giờ đến 10 giờ 30 phút có phân phối đều thì xác suất bạn phải chờ xe hơn 10 phút là bao nhiêu?

Giải: Gọi X là số phút tính từ 10 giờ đến 10 giờ 30 xe buýt sẽ đến trạm thì $X \sim U(0, 30)$. Ta cần tính:

$$P(10 < X < 30) = \int_{10}^{30} \frac{1}{30}\mathrm{d}x = \frac{1}{30}(30 - 10) = \frac{2}{3}.$$

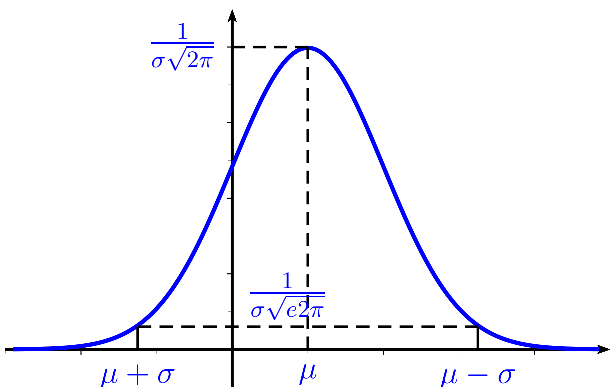
4.2.3. Phân phối chuẩn

a. Định nghĩa phân phối chuẩn:

Định nghĩa 4.7. Đại lượng ngẫu nhiên liên tục X được gọi là có luật phân phối chuẩn (phân phối Gauss) nếu hàm mật độ xác suất của nó có dạng:

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

Ký hiệu: $X \sim N(\mu; \sigma^2)$.



Hình 4.5: Đồ thị hàm $f(x)$ của phân phối chuẩn

Đồ thị của $f(x)$ có dạng hình chuông, đỉnh tại điểm $\left(\mu; \frac{1}{\sigma\sqrt{2\pi}}\right)$, nhận $x = \mu$ làm trục đối xứng, có các điểm uốn là $\left(\mu \pm \sigma; \frac{1}{\sigma\sqrt{e2\pi}}\right)$ và nhận trục hoành làm tiệm cận ngang.

Ta thấy định nghĩa trên thỏa mãn điều kiện của một phân phối xác suất vì:

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x = 1.$$

Thật vậy, đặt $t = \dfrac{x-\mu}{\sigma\sqrt{2}}$ có $\mathrm{d}t = \dfrac{1}{\sigma\sqrt{2}}\mathrm{d}x$. Suy ra:

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

Ta tính: $I = \displaystyle\iint_{\mathbf{R}^2} e^{-x^2-y^2} \mathrm{d}x\mathrm{d}y$. Gọi $S(O, R)$ là hình tròn tâm O, bán kính R. Ta có:

$$I = \lim_{R\to+\infty} \iint_{S(O,R)} e^{-x^2-y^2} \mathrm{d}x\mathrm{d}y = \lim_{R\to+\infty} \int_0^{2\pi} \mathrm{d}\varphi \int_0^R e^{-r^2} r\mathrm{d}r \lim_{R\to+\infty} \pi(1 - e^{-R^2}) = \pi.$$

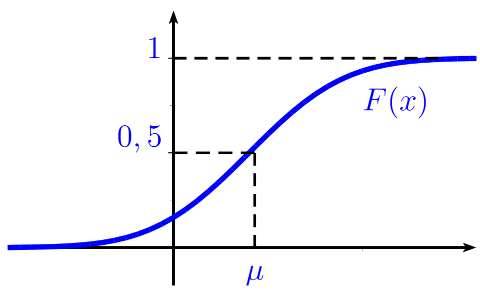
Mặt khác:

$$\iint_{\mathbf{R}^2} e^{-x^2-y^2} \mathrm{d}x\mathrm{d}y = \int_{-\infty}^{+\infty} e^{-x^2} \mathrm{d}x \cdot \int_{-\infty}^{+\infty} e^{-y^2} \mathrm{d}y = \left(\int_{-\infty}^{+\infty} e^{-x^2} \mathrm{d}x\right)^2.$$

Suy ra:

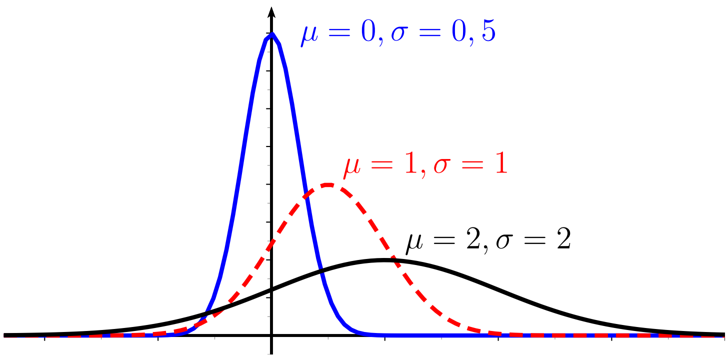
$$\left(\int_{-\infty}^{+\infty} e^{-x^2} \mathrm{d}x\right)^2 = \pi \Leftrightarrow \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}} e^{-x^2} \mathrm{d}x = 1.$$

Hàm phân phối xác suất của X có dạng: $F(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \mathrm{d}t$ nên không thể biểu diễn được thành hàm sơ cấp. Đồ thị của $F(x)$ nhận điểm $(\mu; 0,5)$ làm tâm đối xứng.



Hình 4.6: Đồ thị hàm $F(x)$ của phân phối chuẩn

Nhận xét 4.1. Khi σ càng lớn thì hàm phân phối chuẩn càng phân tán, khi đó đồ thị của hàm mật độ sẽ có đáy rộng hơn nhưng chiều cao sẽ thấp xuống.



Hình 4.7: Đồ thị của hàm mật độ với các giá trị μ và σ khác nhau

Số đặc trưng: Cho $X \sim N(\mu; \sigma^2)$, ta có:

i) $\mathrm{E}(X) = \mu$.

ii) $\mathrm{Var}(X) = \sigma^2$.

iii) $\mathrm{Mod}(X) = \mathrm{Med}(X) = \mu$.

Chứng minh.

$$\begin{aligned}\mathrm{E}(X) &= \int_{-\infty}^{+\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x = \int_{-\infty}^{+\infty} \frac{x-\mu}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x + \int_{-\infty}^{+\infty} \frac{\mu}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \\ &= \frac{2\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} te^{-t^2} \mathrm{d}t + \mu = \frac{2\sigma}{\sqrt{2\pi}} \left(\int_0^{+\infty} te^{-t^2} \mathrm{d}t - \int_0^{+\infty} te^{-t^2} \mathrm{d}t \right) + \mu = \mu.\end{aligned}$$

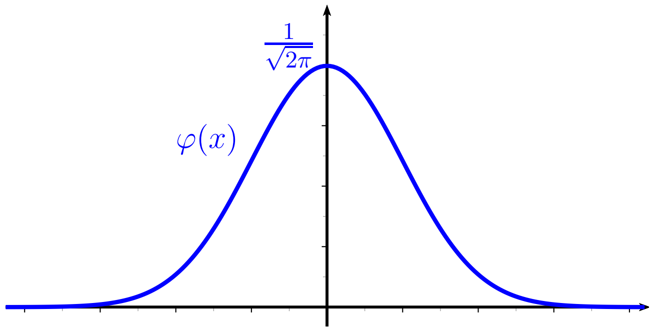
$$\begin{aligned}\mathrm{E}(X^2) &= \int_{-\infty}^{+\infty} \frac{x^2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \\ &= \int_{-\infty}^{+\infty} \frac{(x-\mu)^2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x + \int_{-\infty}^{+\infty} \frac{2\mu(x-\mu)}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x + \int_{-\infty}^{+\infty} \frac{\mu^2}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \\ &= \sigma^2 + \mu^2.\end{aligned}$$

Do đó: $\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2 = \sigma^2$. □

b. Tính xác suất cho phân phối chuẩn:

Định nghĩa 4.8. Phân phối chuẩn có kỳ vọng bằng 0 và phương sai bằng 1 được gọi là phân phối chuẩn tắc, ký hiệu $X \sim N(0; 1)$. Khi đó, hàm mật độ xác suất của X gọi là hàm Laplace có dạng:

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



Hình 4.8: Đồ thị hàm $\varphi(x)$ của phân phối chuẩn tắc

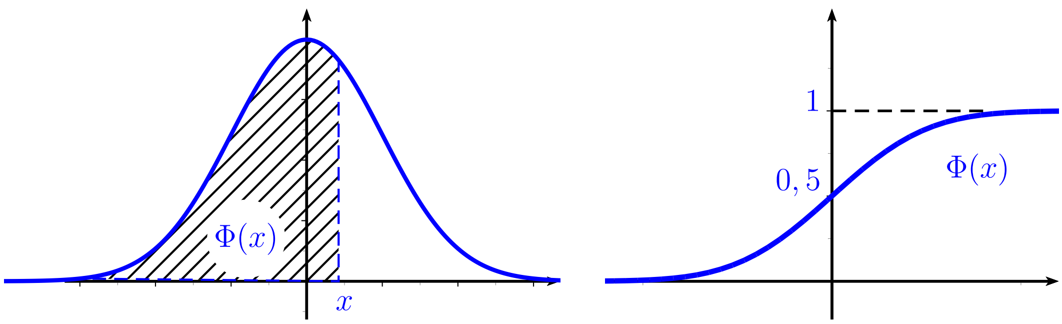
Hàm phân phối xác suất của X gọi là hàm Gauss, ký hiệu $\Phi(x)$ được xác định bởi:

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

Hàm phân phối $\Phi(x)$ có tính chất:

$$\Phi(x) + \Phi(-x) = 1 \Leftrightarrow \Phi(-x) = 1 - \Phi(x).$$

Giá trị của $\varphi(x)$ và $\Phi(x)$ được cho sẵn ở bảng Phụ lục 1 và Phụ lục 2.



Hình 4.9: Đồ thị hàm $\Phi(x)$ của phân phối chuẩn tắc

Định lí 4.3. *Nếu* $X \sim N(\mu;\sigma^2)$ *thì* $aX + b \sim N\big(a\mu + b; (|a|\sigma)^2\big)$ *(với* $a \neq 0$*).*

Chứng minh. Vì $y = ax + b$ là hàm đơn điệu nghiêm ngặt, khả vi và $x = \dfrac{y-b}{a}$ nên giả sử $f(y)$ là hàm mật độ của $Y = aX + b$, theo *Định lí 3.1* ta có:

$$f(y) = f\left(\frac{y-b}{a}\right)\left|\left(\frac{y-b}{a}\right)'\right| = \frac{1}{|a|}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-(a\mu+b))^2}{2a^2\sigma^2}}.$$

Do đó: $Y = aX + b \sim N(a\mu + b; (|a|\sigma)^2)$. □

Hệ quả 4.1. *Nếu* $X \sim N(\mu;\sigma^2)$ *thì* $\dfrac{X-\mu}{\sigma} \sim N(0;1)$.

Chứng minh. Áp dụng định lí với $a = \dfrac{1}{\sigma}, b = -\dfrac{\mu}{\sigma}$. □

Ta có công thức tính xác suất cho $X \sim N(\mu;\sigma^2)$ dựa trên phân phối chuẩn tắc:

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Từ đó suy ra các công thức sau:

i) $P(X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right)$.

ii) $P(X > a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$.

iii) $P(|X-\mu| < \epsilon) = 2\Phi\left(\frac{\epsilon}{\sigma}\right) - 1$.

Ví dụ 4.8. Thời gian X (phút) của một khách hàng chờ để được phục vụ tại một quầy hàng là đại lượng ngẫu nhiên có luật phân phối $N(4,5; 1,21)$.

a) Tính tỉ lệ khách hàng phải chờ để phục vụ từ $3,5$ phút đến 6 phút.

b) Thời gian phải chờ tối thiểu là bao nhiêu, nếu không để quá 5% khách hàng phải chờ phục vụ vượt quá thời gian đó.

Giải: Ta có: $\mu = 4,5; \sigma = \sqrt{\sigma^2} = \sqrt{1,21} = 1,1$.

a) Nên:

$$P(3,5 < X < 6) = \Phi\left(\frac{6-4,5}{1,1}\right) - \Phi\left(\frac{3,5-4,5}{1,1}\right) = \Phi(1,36) - \Phi(-0,91) = 73,17\%.$$

b) Gọi x_0 là lượng thời gian cần tìm, ta có:

$$P(X > x_0) \le 0,05 \Leftrightarrow 1 - \Phi\left(\frac{x_0-4,5}{1,1}\right) \le 0,05 \Leftrightarrow \Phi\left(\frac{x_0-4,5}{1,1}\right) \ge 0,95 = \Phi(1,65).$$

Do $\Phi(x)$ là hàm đồng biến nên $\frac{x_0-4,5}{1,1} \ge 1,65 \Leftrightarrow x_0 \ge 6,315$. Vậy nên thời gian tối thiểu phải chờ là $6,315$ phút.

Ví dụ 4.9. Một doanh nghiệp cần mua một loại trục máy có đường kính từ $1,18$ *cm* đến $1,22$ *cm*. Có hai nhà máy sản xuất loại trục này và đường kính của trục máy được sản xuất ra là đại lượng ngẫu nhiên có phân phối chuẩn với các thông số cho ở bảng:

	Đường kính trung bình	Độ lệnh chuẩn	Giá bán
Nhà máy 1	$1,2$ *cm*	$0,01$ *cm*	3 triệu đồng/1 hộp/100 chiếc
Nhà máy 2	$1,2$ *cm*	$0,015$ *cm*	$2,7$ triệu đồng/1 hộp/100 chiếc

Hỏi doanh nghiệp nên mua trục máy của nhà máy nào?

Giải: Gọi X_i $(i = 1,2)$ là đường kính trục máy do nhà máy i sản xuất. Tỉ lệ trục máy của các nhà máy sản xuất ra thỏa mãn yêu cầu của doanh nghiệp là:

$P(1,18 \leq X_1 \leq 1,22) = \Phi(|X_1 - 1,2| < 0,02) = 2\Phi\left(\frac{0,02}{0,01}\right) - 1 = 2 \cdot 0,9772 - 1 = 0,9544.$
$P(1,18 \leq X_2 \leq 1,22) = \Phi(|X_2 - 1,2| < 0,02) = 2\Phi\left(\frac{0,02}{0,015}\right) - 1 = 2 \cdot 0,9082 - 1 = 0,8164$

Do đó, số trục máy trung bình sử dụng được khi mua 1 hộp của các nhà máy và số tiền chi cho 1 trục máy sử dụng được là:

+ Nhà máy 1 khoảng $95,44$ chiếc với giá $\dfrac{3000000}{95,44} = 31433$ đồng.

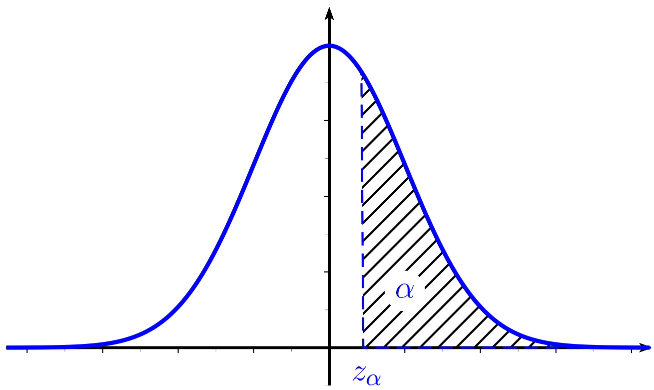
+ Nhà máy 2 khoảng $81,64$ chiếc với giá $\dfrac{2700000}{81,64} = 33072$ đồng.

Vậy doanh nghiệp nên mua sản phẩm của nhà máy 1.

c. Phân vị chuẩn: Giá trị z_α được gọi là phân vị mức α của phân phối chuẩn tắc X nếu:

$$P\left(X > z_\alpha\right) = \alpha.$$

Như vậy z_α là giá trị sao cho $\Phi(z_\alpha) = 1 - \alpha$. Để tìm z_α ta tra bảng Phụ lục 3. Ngoài ra, ta còn có công thức sau: $z_\alpha = -z_{(1-\alpha)}$.



Hình 4.10: Phân vị mức α của phân phối chuẩn tắc

Ví dụ 4.10. Tra bảng Phụ lục 3, ta có: $z_{0,025} = 1,96$; $z_{0,25} = 0,67$; $z_{0,5} = 0$; $z_{0,67} = -0,44$.

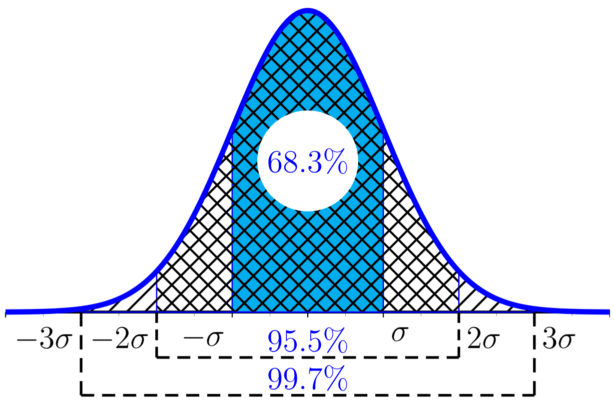
Quy tắc 2σ và 3σ: Nếu $X \sim N(\mu;\sigma^2)$ thì:

$P(|X - \mu| < \sigma) = 2 \cdot \Phi(1) - 1 = 68,27\%.$

$P(|X - \mu| < 2\sigma) = 2 \cdot \Phi(2) - 1 = 95,45\%.$

$P(|X - \mu| < 3\sigma) = 2 \cdot \Phi(3) - 1 = 99,73\%.$

Ý nghĩa của quy tắc: Trong thực hành dù chưa biết luật phân phối xác suất của đại lượng ngẫu nhiên X nhưng nếu thỏa mãn quy tắc 2σ hoặc 3σ thì ta coi như X có phân phối chuẩn. Mặt khác, quy tắc này cũng rất hay được dùng trong thống kê phần ước lượng và kiểm định giả thuyết.



Hình 4.11: Các mức xác suất trong quy tắc 2σ và 3σ

d. Các định lí:

Định lí 4.4 (Định lí giới hạn địa phương Moivre - Laplace)**.** *Cho dãy phép thử Bernoulli $B(n;p)$, khi $n \to +\infty, npq \nrightarrow 0$ ta có công thức xấp xỉ:*

$$\lim_{n\to+\infty} \frac{B(k,n,p)\sqrt{npq}}{\varphi(x_k)} = 1 \qquad \left(\text{với } \varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \text{ và } x_k = \frac{k-np}{\sqrt{npq}}\right).$$

Chứng minh. Do $x_k = \dfrac{k-np}{\sqrt{npq}}$ nên $k = np + x_k\sqrt{npq} \to +\infty$.
Đặt $j = n-k = nq - x_k\sqrt{npq} \to +\infty$. Theo công thức Stirling $m! = \sqrt{2\pi m}\cdot m^m\cdot e^{-m}\cdot e^{\theta_m}$ $\left(0 < \theta_m < \frac{1}{12m}\right)$, ta có:

$$B(k,n,p) = \frac{n!}{k!j!}p^kq^j = \frac{\sqrt{2\pi n}\cdot n!e^{-n}}{\sqrt{2\pi k}\cdot k!e^{-k}\sqrt{2\pi j}\cdot j!e^{-j}}p^kq^je^{\theta_n-\theta_k-\theta_j} = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{kj}}\left(\frac{np}{k}\right)^k\left(\frac{nq}{j}\right)^j e^{\theta},$$

với $|\theta| < \frac{1}{12}\left(\frac{1}{n}+\frac{1}{k}+\frac{1}{j}\right)$.

Mặt khác:

$$\frac{kj}{n} = n\left(p + x_k\sqrt{\frac{pq}{n}}\right)\left(q - x_k\sqrt{\frac{pq}{n}}\right) \to npq,$$

và:

$$\begin{aligned}
\ln\left[\left(\frac{np}{k}\right)^k\left(\frac{nq}{j}\right)^j\right] &= -(np + x_k\sqrt{npq})\left[x_k\sqrt{\frac{q}{np}} - \frac{1}{2}\frac{qx_k^2}{np} + O\left(\frac{1}{n^{2/3}}\right)\right]\\
&\quad - (nq - x_k\sqrt{npq})\left[-x_k\sqrt{\frac{p}{np}} - \frac{1}{2}\frac{px_k^2}{nq} - O\left(\frac{1}{n^{2/3}}\right)\right]\\
&= -\frac{x_k^2}{2} + O\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Do đó:

$$B(k,n,p)\sqrt{npq} \approx \frac{1}{2\pi}\left(\frac{np}{k}\right)^k\left(\frac{nq}{j}\right)^j \to \frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}}.$$

Ta thấy, giá trị xấp xỉ càng tốt khi: i) npq lớn; ii) x_k nhỏ hơn $(npq)^{1/6}$. $\square$

Định lí 4.5 (Định lí giới hạn Moivre - Laplace)**.** *Giả sử* $X \sim B(n;p)$ *và* $S_n = \dfrac{X - np}{\sqrt{npq}}$, *khi* $n \to +\infty$ *thì* $S_n \to N(0;1)$.

Chứng minh. Ta cần chứng minh $P(a \le S_n \le b) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} \mathrm{d}x \quad (\forall [a,b] \subset \mathbf{R})$.

Giả sử $x_k = \dfrac{k - np}{\sqrt{npq}}$. Theo *Định lí 4.4*, ta có:

$$P(S_n = x_k) \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{x_k^2}{2}}.$$

Suy ra:

$$P(a \le S_n \le b) = \sum_k P(S_n = x_k) \approx \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} \sum_k e^{-\frac{x_k^2}{2}}.$$

Biểu thức cuối cùng là tổng Riemann của tích phân $\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_a^b e^{-\frac{x^2}{2}} \mathrm{d}x$. □

Chú ý: Công thức trên xấp xỉ tốt khi $n \ge 30, np \ge 5$ và $nq \ge 5$. Khi đó $S_n \simeq N(0;1)$ nên $X \simeq N(np;npq)$. Do đó, ta có công thức hiệu chỉnh tính liên tục:

$$P(a \le X < b) \approx \Phi\left(\frac{b - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - np}{\sqrt{npq}}\right).$$

Ví dụ 4.11. Một lô bóng đèn điện tử gồm 10000 bóng. Xác suất để mỗi bóng đèn bị hỏng là $0,005$. Tính xác suất để trong lô có từ 40 đến 60 bóng hỏng.

Giải: Gọi X là số bóng hỏng trong lô hàng. Dễ thấy $X \sim B(10000; 0,005)$ với $np = 50$, $npq = 49,75$. Theo công thức tính gần đúng ở trên, ta có:

$$P(40 \le X \le 60) = P(40 \le X < 61) \approx \Phi\left(\frac{61 - 50}{\sqrt{49,75}}\right) - \Phi\left(\frac{40 - 50}{\sqrt{49,75}}\right) = 86,25\%.$$

Định lí giới hạn Moivre - Laplace là trường hợp riêng của định lí sau đây:

Định lí 4.6 (Định lí giới hạn trung tâm)**.** *Nếu dãy* $X_1, X_2, X_3, \ldots, X_n, \ldots$ *các đại lượng ngẫu nhiên độc lập và có cùng luật phân phối xác suất, với* $\mathrm{E}(X_n) = \mu, \mathrm{Var}(X_n) = \sigma^2 \ (\forall n)$ *thì:*

$$S_n = \frac{\sum_{k=1}^n X_k - n\mu}{\sigma\sqrt{n}} \overset{n \to +\infty}{\longrightarrow} N(0;1).$$

Định lí giới hạn trung tâm là một định lí nổi tiếng và có vai trò quan trọng trong lý thuyết xác suất thống kê. Định lí này được nhà toán học Laplace chứng minh vào năm 1810, trong phạm vi của giáo trình ta sẽ không chứng minh.

Ý nghĩa của định lí là, nếu $X_1, X_2, X_3, \ldots, X_n$ độc lập có cùng luật phân phối xác suất với n đủ lớn ($n \ge 30$) thì $S_n \simeq N(0;1)$. Do đó: $X_1 + X_2 + X_3 + \cdots + X_n \simeq N(n\mu; n\sigma^2)$.

Qua đó ta thấy: Tổng của một số rất lớn các đại lượng ngẫu nhiên độc lập có cùng luật phân phối mà sự đóng góp của mỗi thành phần trong tổng đó rất nhỏ thì tổng sẽ có luật phân phối xác suất xấp xỉ phân phối chuẩn.

Ví dụ 4.12. Trọng lượng của một loại sản phẩm là đại lượng ngẫu nhiên có trung bình 50 g, độ lệch chuẩn 10 g. Các sản phẩm được đóng thành hộp, mỗi hộp 100 sản phẩm. Hộp có trọng lượng trên 4,58 kg là đạt chuẩn. Tính tỉ lệ hộp đạt chuẩn.

Giải: Gọi X_j $(j = 1, \ldots, 100)$ là đại lượng ngẫu nhiên chỉ trọng lượng của sản phẩm thứ j. Ta thấy, các X_j độc lập và có cùng phân phối với $\mathrm{E}(X_j) = 50\ g, \mathrm{Var}(X_j) = 100\ g^2$.

Trọng lượng mỗi hộp: $X = X_1 + X_2 + \cdots + X_{100}$, $X \simeq N(100 \cdot 50; 100 \cdot 100) = N(5\ kg; 0,1\ kg^2)$. Do đó, tỉ lệ sản phẩm đạt tiêu chuẩn là:

$$P(X > 4,85) = 1 - \Phi\left(\frac{4,85 - 5}{0,1}\right) = 1 - \big(1 - \Phi(1,5)\big) = 93,32\%.$$

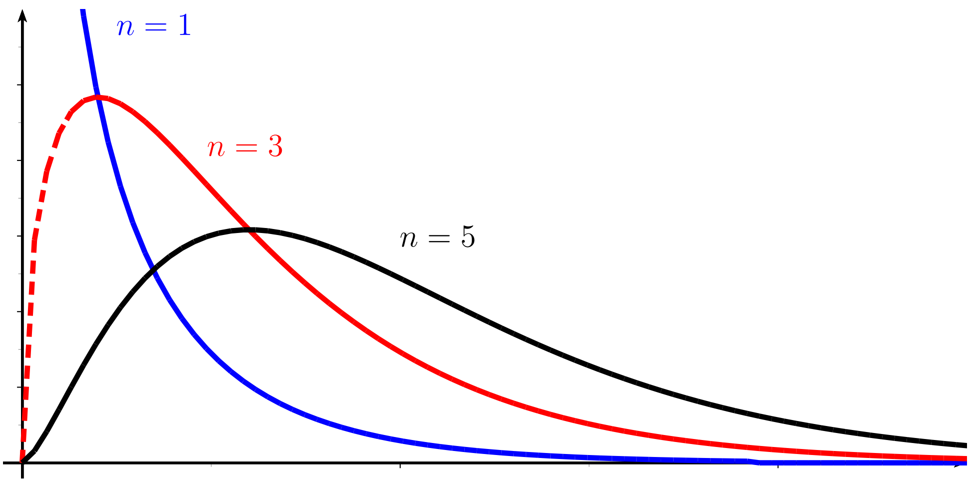
4.2.4. Phân phối Chi bình phương $\chi^2(n)$

Định nghĩa 4.9. Cho X_i $(i = 1, 2, \ldots, n)$ là các đại lượng ngẫu nhiên ngẫu nhiên độc lập và $X_i \sim N(0; 1)$. Khi đó đại lượng ngẫu nhiên $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ được gọi là có phân phối Chi bình phương, bậc tự do n. Ký hiệu: $Y \sim \chi^2(n)$.

Nếu $Y \sim \chi^2(n)$ thì hàm mật độ xác suất của nó là:

$$f(x) = \begin{cases} \dfrac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \cdot e^{-\frac{x}{2}} & \text{với } x \geq 0, \\ 0 & \text{với } x < 0. \end{cases}$$

Trong đó: $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t$ là hàm Gamma.



Hình 4.12: Đồ thị của hàm mật độ $f(x)$ với các bậc tự do khác nhau

Chú ý: Hàm Gamma có tính chất $\Gamma(a+1) = a\Gamma(a)$ vì:

$$\Gamma(a+1) = \int_0^{+\infty} x^a e^{-x}\mathrm{d}x = -x^2e^{-x}\Big|_0^{+\infty} + a\int_0^{+\infty} x^{a-1}e^{-x}\mathrm{d}x = 0 + a\Gamma(a).$$

Dễ thấy:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},\ \Gamma(1) = 1.$$

Nên ta có công thức truy hồi:

$$\Gamma(n+1) = n! \text{ và } \Gamma\left(n+\frac{1}{2}\right) = \frac{(2n-1)!!}{2^n}\sqrt{\pi} \quad (n \in \mathbf{N}).$$

Số đặc trưng: Cho $X \sim \chi^2(n)$, khi đó:

i) $\mathrm{E}(X) = n$.

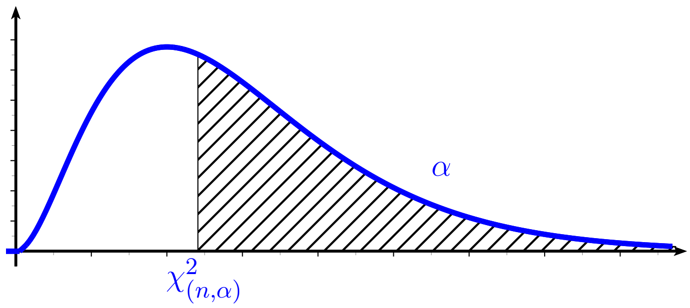
ii) $\mathrm{Var}(X) = 2n$.

iii) $\mathrm{Mod}(X) = \max\{0, n-2\}$.

Phân vị Chi bình phương: Giá trị $\chi^2_{(n,\alpha)}$ được gọi là phân vị mức xác suất α, bậc tự do n của phân phối Chi bình phương nếu:

$$P\left(X > \chi^2_{(n,\alpha)}\right) = \alpha.$$

Phân vị của phân phối Chi bình phương được cho trong bản Phụ lục 4.



Hình 4.13: Phân vị mức xác suất α, bậc tự do n của phân phối Chi bình phương

4.2.5. Phân phối Student $T(n)$

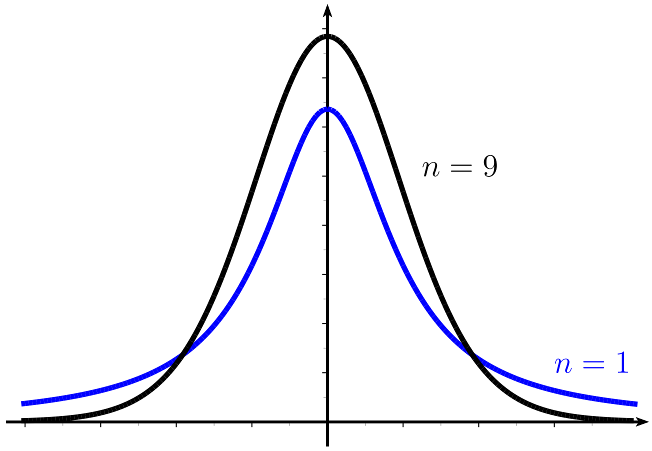
Định nghĩa 4.10. Cho X, Y là các biến ngẫu nhiên độc lập và $X \sim N(0;1)$, $Y \sim \chi^2(n)$. Khi đó đại lượng ngẫu nhiên $Z = \dfrac{X}{\sqrt{Y/n}}$ được gọi là có phân phối Student, bậc tự do n. Ký hiệu: $Z \sim T(n)$.

Nếu $Z \sim T(n)$ thì hàm mật độ xác suất của nó là:

$$f(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad (\forall x \in \mathbf{R}).$$

Do hàm mật độ là hàm chẵn nên đồ thị đối xứng qua trục tung. Khi bậc tự do tăng thì phân phối Student hội tụ rất nhanh về phân phối chuẩn tắc $N(0;1)$. Nên khi n đủ lớn ($n \geq 30$) ta có thể xấp xỉ phân phối Student bằng phân phối chuẩn tắc. Tuy nhiên khi n nhỏ ($n < 30$) việc xấp xỉ như vậy sẽ gặp sai số lớn.

Đồ thị của hàm mật độ có dạng hình chuông tương tự như đồ thị của phân phối chuẩn nhưng đỉnh thấp hơn và hai phần đuôi cao hơn.



Hình 4.14: Đồ thị của hàm mật độ $f(x)$ với các bậc tự do khác nhau

Số đặc trưng: Cho $X \sim T(n)$, khi đó:

i) $\mathrm{E}(X) = 0$ ($n > 1$, với $n = 1$ không xác định).

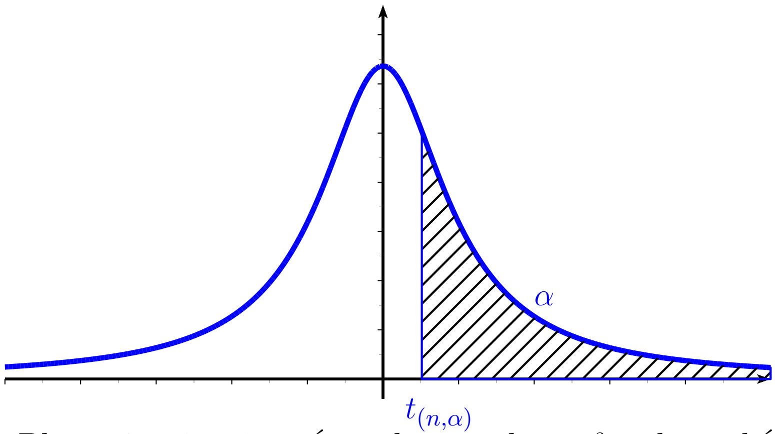
ii) $\mathrm{Var}(X) = \dfrac{n}{n-2}$ ($n > 2$, với $n = 1$ và $n = 2$ không xác định).

iii) $\mathrm{Mod}(X) = 0$.

Phân vị Student: Giá trị $t_{(n,\alpha)}$ được gọi là phân vị mức xác suất α, bậc tự do n của phân phối Student nếu:

$$P\left(X > t_{(n,\alpha)}\right) = \alpha.$$

Phân vị của phân phối Student được cho trong bảng Phụ lục 5. Tương tự như phân phối chuẩn tắc ta cũng có tính chất: $t_{(n,1-\alpha)} = -t_{(n,\alpha)}$.



Hình 4.15: Phân vị mức xác suất α, bậc tự do n của phân phối Student

4.2.6. Phân phối Fisher - Snedecor $F(m, n)$

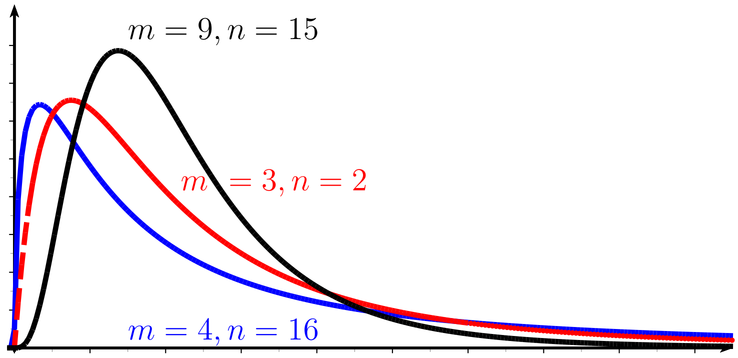
Định nghĩa 4.11. Cho X_1, X_2 là các biến ngẫu nhiên độc lập và $X_1 \sim \chi^2(m)$, $X_2 \sim \chi^2(n)$. Đại lượng ngẫu nhiên $X = \dfrac{X_1/m}{X_2/n}$ được gọi là có phân phối Fisher - Snedecor với (m, n)

bậc tự do. Ký hiệu: $X \sim F(m,n)$.

Nếu $X \sim F(m,n)$ thì hàm mật độ xác suất của nó là:

$$f(x) = \begin{cases} \dfrac{\left(\frac{m}{n}\right)^{\frac{m}{2}} \Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \cdot x^{\frac{m}{2}-1} \cdot \left(\frac{m}{n}x+1\right)^{-\frac{m+n}{2}} & \text{với } x > 0, \\ 0 & \text{với } x \le 0. \end{cases}$$

Đồ thị của $f(x)$ có dạng gần giống với đồ thị hàm mật độ của phân phối Chi bình phương.



Hình 4.16: Đồ thị hàm $f(x)$ của phân phối Fisher

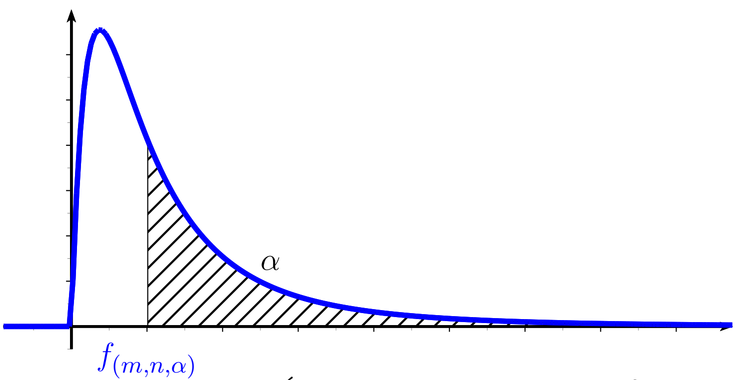
Số đặc trưng: Cho $X \sim F(m,n)$, khi đó:

i) $\mathrm{E}(X) = \dfrac{n}{n-2}$ (với $n > 2$).

ii) $\mathrm{Var}(X) = \dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ (với $n > 4$).

iii) $\mathrm{Mod}(X) = \dfrac{m-2}{2} \cdot \dfrac{n}{n+2}$.

Phân vị Fisher: Giá trị $f_{(m,n,\alpha)}$ được gọi là phân vị mức xác suất α, bậc tự do (m,n) của phân phối Fisher nếu: $P\left(X > f_{(m,n,\alpha)}\right) = \alpha$.



Hình 4.17: Phân vị mức xác suất α, bậc tự do (m,n) của phân phối Fisher

Phân vị của phân phối Fisher được cho ở bảng Phụ lục 6 và có tính chất:

$$f_{(m,n,\alpha)} = \frac{1}{f_{(m,n,1-\alpha)}}.$$

4.3. BÀI TẬP CHƯƠNG 4

Bài 1: Một kiện hàng có 5 sản phẩm tốt và 3 sản phẩm xấu. Lấy ngẫu nhiên từ kiện hàng ra 3 sản phẩm.

a) Lập bảng phân phối xác suất của số sản phẩm xấu lấy được.

b) Tính kỳ vọng, phương sai, giá trị tin chắc của số sản phẩm xấu.

Bài 2: Trong một rổ chứa 20 quả bóng bàn có 5 quả bị móp. Lấy ra ngẫu nhiên 3 quả.

a) Lập bảng phân phối xác suất số quả móp lấy được.

b) Xác định số quả móp tin chắc nhất.

c) Tính xác suất lấy được ít nhất 1 quả móp, nhiều nhất 2 quả móp.

Bài 3: Một người nuôi 20 con gà mái. Biết xác suất để một con gà đẻ trứng trong ngày là $0,7$.

a) Tính xác suất trong một ngày không có con nào đẻ, cả 20 con đều đẻ.

b) Tính xác suất trong một ngày có ít nhất 1 con đẻ, có nhiều nhất 2 con đẻ.

c) Để mỗi ngày có trung bình 100 trứng thì người đó phải nuôi bao nhiêu con gà?

Bài 4: Một người nuôi 180 con gà mái. Biết xác suất để một con đẻ trứng trong ngày là $0,8$.

a) Tính xác suất để người nuôi có được ít nhất 150 trứng trong ngày.

b) Nếu mỗi quả trứng bán được 2000 đồng và tiền mua thức ăn cho mỗi con gà trong ngày là 1200 đồng. Tính số tiền lời trung bình người nuôi thu được trong ngày?

Bài 5: Ở một tổng đài điện thoại, các cuộc điện thoại gọi đến xuất hiện ngẫu nhiên, độc lập với nhau và trung bình có 2 cuộc gọi trong 1 phút. Tính xác suất để:

a) Có đúng 5 cuộc điện thoại trong 2 phút.

b) Không có cuộc điện thoại nào trong khoảng thời gian 30 giây.

c) Có ít nhất 1 cuộc điện thoại trong khoảng thời gian 10 giây.

Bài 6: Trung bình tại bến cảng có 5 tàu cập bến trong một ngày bất kỳ. Tính xác suất để trong một ngày mà ta xét có:

a) Không tàu nào cập bến. b) Đúng 5 tàu cập bến.
c) Từ 3 đến 7 tàu cập bến. d) Có ít nhất 2 tàu cập bến.

Bài 7: Cho $X \sim N(3;4)$. Tính:

a) $P(1 \leq X < 6)$
b) $P(X > 4)$
c) $P(X^2 < 4)$
d) $P(|X-3| < 4)$
e) $P(|X-2| \geq 1)$
f) $P(X^2 - 3X < -2)$

Bài 8: Đo trọng lượng của một loại trái cây, người ta thấy trong 3000 quả chọn ra có 500 quả nặng hơn 1800 g và 2000 quả nặng dưới 1500 g. Hãy tính trọng lượng trung bình và độ lệch tiêu chuẩn trọng lượng của loại trái cây trên. Biết trọng lượng trái cây có phân phối chuẩn.

Bài 9: Trọng lượng của một gói đường (đóng gói bằng máy) có phân phối chuẩn. Trong 1000 gói được kiểm tra có 70 gói trọng lượng lớn hơn 1015 g. Theo bạn trong đó có bao nhiêu gói trọng lượng nhỏ hơn 1008 g? Biết trọng lượng trung bình của gói đường là 1012 g.

Bài 10: Cho X có phân phối chuẩn, $\mathrm{Var}(X) = 25$. Tính $\mathrm{E}(X)$ nếu biết $P(X \geq 20) = 0,38$.

Bài 11: Lãi suất đầu tư vào một dự án được coi như một đại lượng ngẫu nhiên có phân phối chuẩn. Theo đánh giá của Ủy ban đầu tư thì lãi suất lớn hơn 20% có xác suất là $0,1587$ và lãi suất cao hơn 25% có xác suất là $0,0228$. Vậy khả năng đầu tư không bị lỗ là bao nhiêu phần trăm?

Bài 12: Lượng điện sử dụng hàng tháng của các hộ gia đình ở thành phố A là đại lượng ngẫu nhiên có phân phối chuẩn với trung bình là 200 kWh, độ lệch chuẩn là 50 kWh. Tính xác suất để chọn ngẫu nhiên một hộ gia đình trong thành phố thì hộ đó:

a) Có mức sử dụng điện hàng tháng trên 275 kWh.

b) Có mức sử dụng điện hàng tháng dưới 180 kWh.

Bài 13: Khi nghiên cứu về chiều cao của nam giới trưởng thành, người ta nhận thấy chiều cao đó tuân theo luật phân phối chuẩn với trung bình là 168 cm và độ lệch chuẩn 10 cm. Hãy xác định:

a) Tỉ lệ người trưởng thành có chiều cao lớn hơn 173 cm.

b) Tỉ lệ người trưởng thành có chiều cao từ 158 cm đến 175 cm.

c) Giá trị h_0, biết 40,13% người trưởng thành có chiều cao dưới mức h_0.

Bài 14: Tuổi thọ của một loại chíp máy vi tính là $X \sim N(3,5; 1,44)$ (năm). Tính xác suất trong 100 chip loại này có ít nhất 20 chip mà tuổi thọ của nó nhỏ hơn 2 năm.

Bài 15: Tuổi thọ của một loại bóng đèn là $X \sim N(4,2; 2,25)$ (năm). Khi bán một bóng đèn lời được 100000 đồng nhưng nếu bóng đèn phải bảo hành thì lỗ 200000 đồng. Nếu muốn tiền lãi trung bình khi bán mỗi bóng đèn là 30000 đồng thì thời gian bảo hành là bao lâu?

Chương 5

LÝ THUYẾT MẪU

5.1. PHƯƠNG PHÁP MẪU

5.1.1. Giới thiệu về thống kê toán học

Thống kê toán học là một phương pháp khoa học phân tích và xử lý dữ liệu có được nhờ các thí nghiệm, các cuộc điều tra nghiên cứu những hiện tượng tự nhiên, các vấn đề kỹ thuật cũng như các vấn đề xã hội. Từ đó đưa ra các kết luận, dự báo những xu hướng trong tương lai của hiện tượng đang quan sát.

Mỗi dữ liệu có hệ thống các kết quả của dãy thí nghiệm ngẫu nhiên lập thành tập hợp các số liệu thống kê đối với hiện tượng đang được nghiên cứu. Những dữ liệu này có thể mang tính chất định lượng hoặc định tính. Từ những số liệu thống kê, dựa vào các quy luật xác suất người ta nghiên cứu khả năng thu được những kết luận tin cậy và xây dựng phương pháp để rút ra kết luận ấy.

Đặc điểm khác của lý thuyết thống kê là tính ổn định thống kê. Thật vậy, mặc dầu ta không thể đoán trước được các kết quả có thể xảy ra trong các thí nghiệm ngẫu nhiên riêng lẻ vì có những ràng buộc lẫn nhau mà ta không thể tính hết được. Tuy nhiên, nếu không xét một vài thí nghiệm mà xét trong một số lớn thí nghiệm thì có một hiện tượng cực kỳ quan trọng, đó là: Những kết quả tính theo trung bình của nhiều thí nghiệm lại có tính ổn định.

Việc nghiên cứu một hiện tượng nào đó bằng phương pháp thống kê toán học được tiến hành theo các giai đoạn sau:

+ Điều tra và thu thập số liệu có tính hệ thống, tìm số đặc trưng của dãy số liệu.

+ Những đặc trưng từ các số liệu thu thập được chưa phản ánh đầy đủ tất cả các đặc trưng đáng lẽ phải có nên chưa hẳn chúng đã phản ánh chính xác, đầy đủ những đặc điểm của hiện tượng. Vậy nên, chúng ta cần phải phân tích, so sánh sao cho từ những đặc điểm được phát hiện qua các số liệu đã có, mà rút ra tính quy luật khách quan của hiện tượng.

Bài toán phân tích thống kê có thể bao gồm các bước sau:

+ Thu thập thông tin từ số lượng đã quan sát.

+ Thiết lập mô hình quan sát, tức là chọn qui luật phân phối.

+ Ước lượng các tham số của mô hình.

+ Đánh giá sự phù hợp giữa mô hình và quan sát thực nghiệm.

+ Giải bài toán thực tế bằng cách ước lượng các tham số với các tiêu chuẩn phù hợp.

5.1.2. Khái niệm tổng thể và mẫu

Tổng thể được hiểu là tập hợp N có số phần tử rất lớn, bao gồm tất cả những phần tử mang những thông tin liên quan đến một vấn đề cần khảo sát nào đó. Thí dụ, khi khảo sát tỷ lệ sinh viên khá giỏi của một trường đại học thì tổng thể nghiên cứu là tập hợp tất cả sinh viên đang theo học ở trường đó.

Trong thực tế, khó có thể (và thường là không thể) khảo sát tất cả các phần tử của tổng thể. Từ tổng thể người ta kiểm tra, quan sát trên một nhóm n phần tử đại diện gọi là *mẫu*. Với các thông tin thu được từ mẫu, bằng các phương pháp khoa học người ta rút ra kết luận chung cho cả tổng thể. Phương pháp làm việc như vậy được gọi là phương pháp mẫu.

Để đưa ra kết luận chính xác, việc chọn mẫu phải được tiến hành một cách khách quan. Có hai phương thức cơ bản thường được dùng để lấy một mẫu đại diện từ tổng thể là:

◦ *Lấy mẫu có hoàn lại:* Chọn ngẫu nhiên từ tổng thể ra một phần tử để kiểm tra, ghi lại các thông tin đặc trưng cần thiết của phần tử đó, rồi trả nó trở lại tổng thể trước khi chọn tiếp ngẫu nhiên lần sau. Khi đó kết quả ở các lần chọn độc lập với nhau.

◦ *Lấy mẫu không hoàn lại:* Tương tự như trên, chỉ khác ở chỗ là các phần tử được chọn ra sẽ không được trả lại tập ban đầu. Khi đó kết quả ở các lần chọn phụ thuộc nhau nhưng nếu lấy ít từ tập có số phần tử rất lớn thì cũng có thể xem như độc lập.

Ngoài hai phương pháp lấy mẫu ở trên, đối với những lĩnh vực riêng biệt như nông nghiệp, y học, xã hội học,... lại có những phương pháp lấy mẫu mang tính đặc thù riêng của nó.

5.1.3. Mẫu ngẫu nhiên, mẫu cụ thể

Định nghĩa 5.1. Tiến hành n quan sát độc lập về đại lượng ngẫu nhiên X (xác định cho dấu hiệu nào đó trên các phần tử của tổng thể), gọi X_i là đại lượng ngẫu nhiên chỉ kết quả quan sát ở lần thứ i của X. Khi đó $W = (X_1, X_2, \ldots, X_n)$ được gọi là *mẫu ngẫu nhiên*, n gọi là *cỡ mẫu* (kích thước mẫu).

Như vậy mẫu ngẫu nhiên kích thước n thực chất là n đại lượng ngẫu nhiên độc lập $X_1, X_2, \ldots, X_n$ được lập ra từ biến ngẫu nhiên X và có cùng luật phân phối xác suất với X. Khi đó X được gọi là đại lượng ngẫu nhiên gốc ứng với tổng thể nghiên cứu. Kỳ vọng $\mathrm{E}(X) = \mu$ và phương sai $\mathrm{Var}(X) = \sigma^2$ của đại lượng ngẫu nhiên gốc còn gọi là trung bình và phương sai của tổng thể.

Ta gọi x_i là giá trị thu được trong lần quan sát thứ i của X_i. Khi đó $w = (x_1, x_2, \ldots, x_n)$ gọi là một *mẫu cụ thể* mà mẫu ngẫu nhiên W nhận được.

5.1.4. Biểu đồ, tổ chức đồ

Giả sử X là đại lượng ngẫu nhiên rời rạc có phân phối xác suất $P(X = x_i) = p_i$. Tập hợp các điểm $(x_i; p_i)$ trên mặt phẳng tọa độ gọi là biểu đồ xác suất của X. Cho mẫu cỡ n, gọi n_i là tần số của giá trị x_i $(i = 1, \dots, k)$ với $n_1 + n_2 + \cdots + n_k = n$. Khi đó, tập hợp các điểm $(x_i; f_i = n_i/n)$ trong mặt phẳng tọa độ gọi là *biểu đồ tần suất*, đường gấp khúc nối các điểm đó gọi là *đa giác tần suất*.

Ví dụ 5.1. Khi tiến hành đếm số lô hàng còn lại trong 5 kho của một nhà máy ta có bảng:

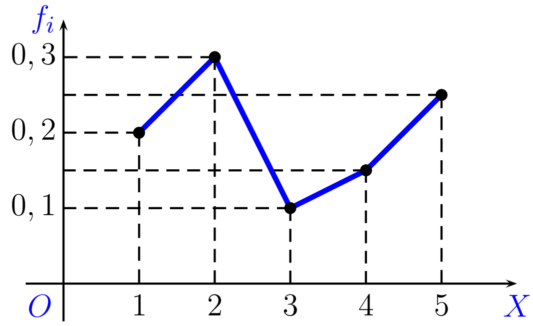
Kho số	1	2	3	4	5
Số lô hàng	20	30	10	15	25

Hãy xây dựng biểu đồ tần suất và đa giác tần suất.

Giải: Ta có bảng phân phối tần suất:

X	1	2	3	4	5
f_i	0,2	0,3	0,1	0,15	0,25

Biểu đồ tần suất và đa giác tần suất:

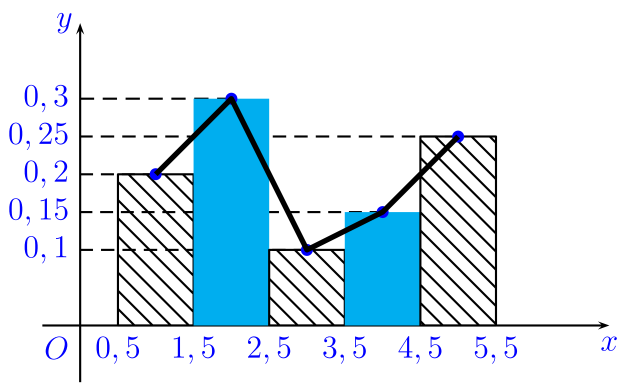


Giả sử X là đại lượng ngẫu nhiên liên tục với hàm mật độ xác suất $f(x)$ chưa biết. Ta chia khoảng biến thiên của mẫu thành các khoảng con bằng nhau có độ dài $2h$ bởi các điểm chia $x_i - h; x_i + h; \dots$ $(i = 1, 2, \dots, k)$. Trong mặt phẳng tọa độ, trên mỗi khoảng con ta vẽ hình chữ nhật có cạnh đáy là $(x_i - h; x_i + h)$ và cạnh còn lại có độ dài là $\dfrac{f_i}{2h}$. Khi đó biểu đồ được tạo thành bởi các hình chữ nhật gọi là *tổ chức đồ tần suất* và đa giác tần suất là đường gấp khúc đi qua trung điểm của các đáy trên.

Ta thấy, các hình chữ nhật này có diện tích là f_i và tổng diện tích của chúng bằng 1.

Khi xấp xỉ đa giác tần suất bởi một đường cong nào đó thì đường cong này cho ta hình ảnh hình học ban đầu về đường cong của hàm mật độ.

Ví dụ 5.2. Ta có tổ chức đồ tần suất và đa giác tần suất ở *Ví dụ 5.1* như hình vẽ sau:



Chú ý: Từ mẫu cụ thể $(x_1, x_2, \dots, x_n)$ của mẫu ngẫu nhiên $(X_1, X_2, \dots, X_n)$ lập từ X, ta đặt:

$$F_n(x) = \frac{\sum_{x_i<x} n_i}{n} = \sum_{x_i<x} f_i,$$

khi đó $F_n \overset{n\to+\infty}{\longrightarrow} F(x) = P(X < x)$.

Định lí 5.1 (Định lí Glivenko). *Giả sử $F(x)$ là hàm phân phối của biến ngẫu nhiên X mà ta đang tìm, $F_n(x)$ là hàm phân phối nhận được từ mẫu ngẫu nhiên cỡ n. Khi đó:*

$$\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| \longrightarrow 0 \textit{ khi } n \longrightarrow +\infty.$$

Như vậy, với n đủ lớn thì $F_n(x) \approx F(x)$.

5.2. CÁC ĐẶC TRƯNG MẪU

5.2.1. Các đặc trưng mẫu ngẫu nhiên

Hàm của mẫu ngẫu nhiên được gọi là *thống kê*, ký hiệu là: $\varphi(X_1, X_2, \ldots, X_n)$.

Giả sử tổng thể X có kỳ vọng $\mathrm{E}(X) = \mu$, phương sai $\mathrm{Var}(X) = \sigma^2$ và tỉ lệ p. Cho mẫu ngẫu nhiên $W = (X_1, X_2, \ldots, X_n)$, ta có các thống kê sau đây gọi là các đặc trưng mẫu ngẫu nhiên:

Trung bình mẫu ngẫu nhiên: $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Khi đó:
$\mathrm{E}(\overline{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}(X_i) = \mathrm{E}(X) = \mu$.
$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{n\mathrm{Var}(X)}{n^2} = \frac{\sigma^2}{n}$.

Phương sai mẫu ngẫu nhiên: $S_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2$. Khi đó:

$$\begin{aligned}\mathrm{E}(S_n^2) &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}(X_i^2) - \mathrm{E}(\overline{X}^2) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}(X_i^2 - \mu^2) - \mathrm{E}(\overline{X}^2 - \mu^2)\\ &= \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(X_i) - \mathrm{Var}(\overline{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.\end{aligned}$$

Độ lệch chuẩn mẫu ngẫu nhiên: $S_n = \sqrt{S_n^2}$.

Phương sai mẫu điều chỉnh mẫu ngẫu nhiên: $S^2 = \frac{n}{n-1}S_n^2$.
Khi đó: $\mathrm{E}(S^2) = \frac{n}{n-1}\mathrm{E}(S_n^2) = \sigma^2$.

Độ lệch điều chỉnh mẫu ngẫu nhiên: $S = \sqrt{S^2}$.

Tỉ lệ mẫu ngẫu nhiên: $F = \frac{1}{n}\sum_{i=1}^n X_i$, $X_i \sim A(p)$. Khi đó: $\mathrm{E}(F) = p, \mathrm{Var}(F) = \frac{pq}{n}$.

5.2.2. Các đặc trưng mẫu cụ thể

Cho mẫu cụ thể $w = (x_1, \ldots, x_n)$ là một giá trị của mẫu ngẫu nhiên $W = (X_1, X_2, \ldots, X_n)$. Khi đó, có các đặc trưng (tham số) mẫu cụ thể sau:

Trung bình mẫu cụ thể: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Phương sai mẫu cụ thể: $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$.

Độ lệch chuẩn mẫu cụ thể: $s_n = \sqrt{s_n^2}$.

Phương sai điều chỉnh mẫu cụ thể: $s^2 = \frac{n}{n-1}s_n^2$.

Độ lệch điều chỉnh mẫu cụ thể: $s = \sqrt{s^2}$.

Chú ý: Nếu mẫu cụ thể được cho dưới dạng tần số thì ta có công thức tính như sau:

Trung bình mẫu cụ thể: $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$.

Phương sai mẫu cụ thể: $s_n^2 = \frac{1}{n}\sum_{i=1}^{k}(n_i x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$.

5.2.3. Cách tính các đặc trưng mẫu cụ thể

Trường hợp $w = (x_1, x_2, \ldots, x_n)$ có kích thước n tương đối lớn thì việc xác định các giá trị $\overline{x}, s_n, s$ theo công thức ở trên sẽ mất nhiều công sức. Khi đó, nếu không cần độ chính xác cao, ta sắp xếp lại các số liệu ban đầu của mẫu, lập nên một mẫu các điểm đại diện, rồi xác định các đặc trưng mẫu dựa vào các điểm đại diện đó. Cách làm cụ thể như sau:

Chia các giá trị trong w thành các nửa khoảng giá trị $(x'_i; x'_{i+1}]$ (thường chia thành các nửa khoảng có độ dài như nhau) và xác định tần số n_i tương ứng ở mỗi nửa khoảng (số giá trị trong w thuộc $(x'_i; x'_{i+1}]$). Khi đó, ta được một mẫu dưới dạng các nửa khoảng:

Các khoảng	$(x'_1; x'_2]$	$(x'_2; x'_3]$	$\ldots$	$(x'_k; x'_{k+1}]$
Tần số	n_1	n_2	$\ldots$	n_k

Trong đó: $\sum_{i=1}^{k} n_i = n$.

Từ các nửa khoảng $(x'_i; x'_{i+1}]$ ta chọn các điểm đại diện x_i $\left(\text{thường lấy } x_i = \frac{x'_i + x'_{i+1}}{2}\right)$. Sau đó ta lập bảng tính như sau:

Lớp	x_i	n_i	$n_i x_i$	$n_i x_i^2$
$(x'_1; x'_2]$	x_1	n_1	$n_1 x_1$	$n_1 x_1^2$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$(x'_k; x'_{k+1}]$	x_k	n_k	$n_k x_k$	$n_k x_k^2$
	Σ	n	$\Sigma n_i x_i$	$\Sigma n_i x_i^2$

Khi đó, ta được: $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$ và $s_n^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2$.

Ví dụ 5.3. Lượng xăng hao phí (X) của một xe tải đi từ thành phố A đến thành phố B sau 50 lần chạy, cho kết quả trong bảng sau:

$X\ (l)$	$(10,4; 10,6]$	$(10,6; 10,8]$	$(10,8; 11]$	$(11; 11,2]$	$(11,2; 11,4]$	$(11,4; 11,6]$
Số lần	10	5	8	9	7	11

Hãy tính $\overline{x}, s_n^2, s_n, s^2$ và s.

Giải: Ta lập bảng tính như sau:

Lớp	x_i	n_i	n_ix_i	$n_ix_i^2$
(10,4;10,6]	10,5	10	105	1102,5
(10,6;10,8]	10,7	5	53,5	572,45
(10,8;11]	10,9	8	87,2	950,48
(11;11,2]	11,1	9	99,9	1108,89
(11,2;11,4]	11,3	7	79,1	893,83
(11,4;11,6]	11,5	11	126,5	1454,75
		$n = 50$	$\Sigma n_ix_i = 551,2$	$\Sigma n_ix_i^2 = 6082,9$

Do đó: $\overline{x} = \dfrac{1}{n}\sum n_ix_i = \dfrac{551,2}{50} = 11,024\ l.$

$s_n^2 = \dfrac{1}{n}\sum n_ix_i^2 - \overline{x}^2 = \dfrac{6082,9}{50} - 11,024^2 = 0,1294 \Rightarrow s_n = \sqrt{0,1294} = 0,3598\ l.$

$s^2 = \dfrac{n}{n-1}s_n^2 = \dfrac{50}{49}\cdot 0,1294 = 0,1321 \Rightarrow s = \sqrt{0,1321} = 0,3634\ l.$

5.3. BÀI TẬP CHƯƠNG 5

Bài 1: Kiểm tra ngẫu nhiên chiều cao của 50 học sinh nam ở một trường trung học phổ thông ta được bảng số liệu sau (được đo chính xác đến cm):

157	164	170	153	170	169	178	153	171	164
173	165	159	177	157	162	165	162	170	170
158	163	164	176	156	160	170	166	167	179
172	156	158	168	165	160	171	160	156	180
159	163	172	173	165	156	175	170	166	159

Hãy tính trung bình, phương sai, phương sai điều chỉnh, độ lệch chuẩn, độ lệch điều chỉnh của mẫu trên.

Bài 2: Thống kê tỉ giá của một mẫu gồm 50 chứng khoán trên thị trường trong một ngày được cho như sau:

7	9	8	6	12	6	9	15	9	16
8	5	14	8	7	6	10	8	11	4
10	6	16	5	10	12	7	10	15	7
10	8	8	10	18	8	10	11	7	10
7	8	8	23	13	9	8	9	9	13

Hãy tính trung bình, phương sai, phương sai điều chỉnh, độ lệch chuẩn, độ lệch điều chỉnh của mẫu trên.

Bài 3: Điều tra thu nhập hàng tháng (X) của 100 người ta có bảng số liệu sau:

X (triệu đồng)	≤ 1	$(1;2]$	$(2;3]$	$(3;4]$	$(4;5]$	$(5;6]$
Số người	9	18	24	34	12	3

Hãy tính trung bình, phương sai, phương sai điều chỉnh, độ lệch chuẩn, độ lệch điều chỉnh của mẫu trên.

Bài 4: Thống kê số sản phẩm (SP) bán được trong một năm (365 ngày) của một cửa hàng, ta có bảng số liệu sau đây:

Số SP	(0;5]	(5;10]	(10;15]	(15;20]	(25;25]	(25;30]	(30;35]	(35;40]	(40;45]
Số ngày	15	24	37	50	59	57	52	47	24

Hãy tính trung bình, phương sai, phương sai điều chỉnh, độ lệch chuẩn, độ lệch điều chỉnh của mẫu trên.

Chương 6

ƯỚC LƯỢNG THAM SỐ

6.1. ƯỚC LƯỢNG ĐIỂM

6.1.1. Bài toán ước lượng điểm

Nghiên cứu đại lượng ngẫu nhiên X, giả sử ta biết được luật phân phối xác suất của X phụ thuộc vào một hay vài tham số θ nào đó: $X = F(x, \theta)$. Khi đó để xác định hoàn toàn phân phối của X ta phải xác định được tham số θ mà phân phối đó nhận.

Ngay trong trường hợp chưa biết rõ về phân phối của X thì việc biết được các tham số đặc trưng này cũng rất có ý nghĩa. Do đó bài toán đi tìm ước lượng cho các tham ẩn của phân phối hoặc ước lượng các số đặc trưng của biến ngẫu nhiên là bài toán cần thiết.

Bài toán: Giả sử đại lượng ngẫu nhiên X có tham số θ (trung bình, tỷ lệ, phương sai,...) chưa biết. Dựa vào mẫu ngẫu nhiên $W = (X_1, X_2, \ldots, X_n)$ ta xây dựng thống kê $\varphi = \varphi(W) = \varphi(X_1, X_2, \ldots, X_n)$. Từ đó đưa ra giá trị $\theta_0 = \varphi(w) = \varphi(x_1, x_2, \ldots, x_n)$ để ước lượng (dự đoán) cho θ chưa biết "chấp nhận được" gọi là bài toán ước lượng điểm của tham số θ.

Do θ chưa biết nên không thể so sánh θ_0 với θ để đánh giá chất lượng của thống kê $\varphi = \varphi(W)$ cho nên ta dựa vào những tiêu chuẩn khác nhau để đưa ra các dạng ước lượng khác nhau cho tham số θ.

6.1.2. Ước lượng không chệch

Định nghĩa 6.1. Giá trị $\theta_0 = \varphi(w)$ được gọi là *ước lượng không chệch* của tham số θ nếu: $\mathrm{E}\big(\varphi(W)\big) = \theta$.

Chú ý: Từ định nghĩa nêu trên, ta có:

i) f là ước lượng không chệch cho p.

ii) $\overline{x}$ là ước lượng không chệch cho μ.

iii) s^2 là ước lượng không chệch cho σ^2.

Ví dụ 6.1. Thống kê 100 trái cây của nông trường ta thu được bảng số liệu sau đây:

Trọng lượng (g)	$[45; 55]$	$[55; 65)$	$[65; 75)$	$[75; 85)$	$[85; 95)$	$[95; 105)$	$[105; 115)$
Số trái	3	11	19	34	21	9	3

a) Tìm ước lượng không chệch cho trọng lượng trung bình của trái cây ở nông trường.

b) Tìm ước lượng không chệch cho độ lệch tiêu chuẩn của trái cây ở nông trường.

c) Xem trái cây có trọng lượng không quá 75 g là trái loại II. Tìm ước lượng không chệch của tỷ lệ trái loại II trong nông trường.

Giải:

a) Ước lượng không chệch cho trọng lượng trung bình là: $\overline{x} = 79,98\ g$.

b) Ước lượng không chệch cho độ lệch tiêu chuẩn là: $s = 13,256\ g$.

c) Ước lượng không chệch cho tỷ lệ trái cây loại II là: $f = \dfrac{33}{100} = 0,33$.

6.1.3. Ước lượng hiệu quả

Định nghĩa 6.2. Giá trị $\theta_0 = \varphi(w)$ được gọi là *ước lượng hiệu quả* của θ nếu

i) θ_0 là ước lượng không chệch của tham số θ,

ii) $\theta' = \psi(w)$ là một ước lượng không chệch khác của tham số θ thì

$$\mathrm{E}\big(\varphi(W) - \theta\big)^2 \leq \mathrm{E}\big(\psi(W) - \theta\big)^2.$$

Chú ý: Bởi vì $\mathrm{E}\big(\varphi(W)\big) = \theta$ và $\mathrm{E}\big(\psi(W)\big) = \theta$ cho nên $\mathrm{E}\big(\varphi(W) - \theta\big)^2 = \mathrm{Var}\big(\varphi(W)\big)$ và $\mathrm{E}\big(\psi(W) - \theta\big)^2 = \mathrm{Var}\big(\psi(W)\big)$. Do đó, ước lượng hiệu quả còn được gọi là ước lượng không chệch có phương sai bé nhất.

Khi hàm mật độ xác suất $f(x,\theta)$ (với θ là tham số) của X thỏa mãn một số điều kiện nhất định thì ta có bất đẳng thức Cramer - Rao:

$$\mathrm{Var}\big(\varphi(W)\big) \geq n\mathrm{E}\left(\frac{\partial \ln f(x,\theta)}{\partial \theta}\right)^{-2}.$$

Ta thấy, dấu đẳng thức xảy ra khi $\varphi(w)$ là ước lượng hiệu quả của tham số θ.

Ví dụ 6.2. Nếu đại lượng ngẫu nhiên $X \sim N(\mu;\sigma^2)$ thì trung bình mẫu $\overline{x}$ là ước lượng hiệu quả của μ.

Thật vậy, bởi vì $\overline{X} \sim N\left(\mu; \dfrac{\sigma^2}{n}\right)$ nên $\mathrm{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$. Hàm mật độ của X là:

$$f(x,\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow \frac{\partial \ln f(x,\mu)}{\partial \mu} = \frac{x-\mu}{\sigma^2}.$$

Do đó:

$$n\mathrm{E}\left(\frac{\partial \ln f(x,\mu)}{\partial \mu}\right)^2 = n\mathrm{E}\left(\frac{x-\mu}{\sigma^2}\right)^2 = \frac{n}{\sigma^2}\mathrm{E}\left(\frac{x-\mu}{\sigma}\right)^2 = \frac{n}{\sigma^2} = \frac{1}{\mathrm{Var}(\overline{X})}.$$

Suy ra $\overline{x}$ là ước lượng hiệu quả của μ.

6.1.4. Ước lượng vững

Định nghĩa 6.3. Giá trị $\theta_0 = \varphi(w)$ được gọi là *ước lượng vững* của θ nếu $\forall\epsilon > 0$ bé tùy ý có: $\lim_{n\to+\infty} P(|\varphi(W) - \theta| < \epsilon) = 1$.

Ý nghĩa: Nếu thống kê $\varphi(W)$ là ước lượng vững của tham số θ thì khi n đủ lớn sự sai khác giữa $\varphi(W)$ và θ là không đáng kể. Nghĩa là: $\varphi \overset{n\to+\infty}{\longrightarrow} \theta$.

Ví dụ 6.3. Cho đại lượng ngẫu nhiên X có $\mathrm{E}(X) = \mu, \mathrm{Var}(X) = \sigma^2$. Khi đó $\overline{x}$ là ước lượng vững của μ.

Thật vậy, do $\mathrm{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$ nên theo bất đẳng thức Chebyshev ta có:

$$P\left(|\overline{X} - \mu| < \epsilon\right) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

Mà $\lim_{n\to+\infty}\left(1 - \dfrac{\sigma^2}{n\epsilon^2}\right) = 1$ nên $\lim_{n\to+\infty} P(|\overline{X} - \mu| < \epsilon) = 1$.

6.1.5. Ước lượng hợp lí tối đa

Giả sử $W = (X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên được lấy từ X, có mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$ và thống kê $\varphi = \varphi(X_1, X_2, \ldots, X_n)$.

Hàm hợp lí $L(x_1, x_2, \ldots, x_n, \theta)$ của đối số θ được xác định như sau:

▷ Nếu X rời rạc thì

$$L(x_1, \ldots, x_n, \theta) = P(X_1 = x_1|\theta, \ldots, X_n = x_n|\theta) = \prod_{i=1}^{n} P(X_i = x_i, \theta),$$

với $L(x_1, x_2, \ldots, x_n, \theta)$ là xác suất để nhận được mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$.

▷ Nếu X liên tục có hàm mật độ xác suất là $f(x, \theta)$ thì

$$L(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta),$$

với $L(x_1, x_2, \ldots, x_n, \theta)$ là giá trị của hàm mật độ xác suất tại điểm $w = (x_1, x_2, \ldots, x_n)$.

Giá trị $\theta_0 = \varphi(x_1, x_2, \ldots, x_n)$ được gọi là *ước lượng hợp lí tối đa* nếu ứng với giá trị này của θ hàm hợp lí đạt cực đại.

Vì hàm L và $\ln L$ đạt cực đại tại cùng một giá trị θ nên khi tìm ước lượng hợp lí tối đa ta xét $\ln L$ thay vì xét L. Ta thực hiện qua các bước sau:

- Tìm $\dfrac{\partial \ln L}{\partial \theta}$.
- Giải phương trình hợp lí $\dfrac{\partial \ln L}{\partial \theta} = 0$. Giả sử $\theta_0 = \varphi(x_1, x_2, \ldots, x_n)$ là nghiệm.

∘ Tìm đạo hàm cấp hai $\dfrac{\partial^2 \ln L}{\partial^2 \theta}$.

Nếu tại θ_0 mà $\dfrac{\partial^2 \ln L}{\partial^2 \theta} < 0$ thì $\ln L$ đạt cực đại. Khi đó $\theta_0 = \varphi(x_1, x_2, \ldots, x_n)$ là ước lượng điểm hợp lí tối đa của θ.

Ví dụ 6.4. Cho $X \sim A(p)$. Tìm ước lượng hợp lí tối đa cho p.

Giải: Do $X \sim A(p)$ nên $P(X_i = x_i, p) = p^{x_i}(1-p)^{1-x_i}$ với $x_i \in \{0, 1\}$. Ta có:

$$L(x_1, x_2, \ldots, x_n, p) = \prod_{i=1}^{n} P(X_i = x_i, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$\Leftrightarrow \ln L(x_1, x_2, \ldots, x_n, p) = \sum_{i=1}^{n} \big(x_i \ln p + (1 - x_i) \ln (1-p)\big)$$

$$\Leftrightarrow \frac{\partial \ln L(x_1, x_2, \ldots, x_n, p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\sum_{i=1}^{n}(1 - x_i).$$

Giải phương trình $\dfrac{\partial \ln L}{\partial p} = 0$, ta được:

$$\sum_{i=1}^{n} x_i \left(\frac{1}{p} + \frac{1}{1-p}\right) - \sum_{i=1}^{n} \left(\frac{1}{1-p}\right) = 0$$

$$\Leftrightarrow \frac{1}{p(1-p)} \sum_{i=1}^{n} x_i = \frac{1}{1-p} \cdot n \Leftrightarrow \frac{1}{p}\sum_{i=1}^{n} x_i = n$$

$$\Leftrightarrow p = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \quad (= f).$$

Vậy nên trung bình mẫu $\overline{x}$ (và tần số mẫu f) là ước lượng hợp lí tối đa cho p.

Ví dụ 6.5. Cho $X \sim N(\mu, \sigma^2)$. Tìm ước lượng hợp lí tối đa cho μ, σ^2.

Giải: Hàm hợp lí có dạng:

$$L(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2}$$

$$\Leftrightarrow \ln L(x_1, x_2, \ldots, x_n, \theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\ln(2\pi\sigma^2).$$

Suy ra:

$$\begin{cases} \dfrac{\partial \ln L}{\partial \mu} = \dfrac{n\overline{x} - n\mu}{\sigma^2} = 0, \\ \dfrac{\partial \ln L}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^4} = 0. \end{cases} \Leftrightarrow \begin{cases} \mu = \overline{x}, \\ \sigma^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = s_n^2. \end{cases}$$

Vậy $\overline{x}$ là ước lượng hợp lý tối đa cho μ và s_n^2 là ước lượng hợp lý tối đa cho σ^2.

6.2. ƯỚC LƯỢNG KHOẢNG

6.2.1. Bài toán ước lượng khoảng

Ước lượng điểm sử dụng một giá trị đơn lẻ để đưa ra về kết luận tham số của tổng thể nên thường gặp sai lệch, đặc biệt là khi mẫu có kích thước nhỏ. Để khắc phục nhược điểm này người ta xây dựng ước lượng khoảng.

Giả sử tổng thể nghiên cứu có tham số θ nào đó chưa biết cần phải ước lượng. Mục đích của phương pháp khoảng tin cậy là chỉ ra cách xác định khoảng giá trị (θ_1, θ_2) để ước lượng cho tham số θ. Khoảng giá trị (θ_1, θ_2) được gọi là *khoảng tin cậy* của ước lượng.

Điều kiện đặt ra đối với khoảng tin cậy (θ_1, θ_2) là phải chứa tham số θ với xác suất khá cao. Trong phương pháp này, người ta thường ấn định một số α $(0 < \alpha < 1)$ rồi buộc xác suất để tham số θ rơi vào khoảng tin cậy (θ_1, θ_2) phải đạt là $1 - \alpha$, tức là $P(\theta_1 < \theta < \theta_2) = 1 - \alpha$. Khi đó θ_1, θ_2 lần lượt gọi là giới hạn tin cậy dưới và giới hạn tin cậy trên của ước lượng.

Giá trị xác suất cho trước α gọi là *mức ý nghĩa* của ước lượng và $1 - \alpha$ gọi là *độ tin cậy* của ước lượng. Rõ ràng khi α càng bé thì $P(\theta_1 < \theta < \theta_2)$ càng lớn, nghĩa là độ tin cậy của ước lượng càng cao.

Các giá trị θ_1 và θ_2 tạo nên khoảng tin cậy được xác định bằng phương pháp mẫu. Từ tổng thể lập mẫu ngẫu nhiên $W = (X_1, X_2, \ldots, X_n)$, các thống kê $\varphi_1(X_1, X_2, \ldots, X_n)$ và $\varphi_2(X_1, X_2, \ldots, X_n)$. Thay mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$ vào các thống kê ta tính được $\theta_1 = \varphi_1(x_1, x_2, \ldots, x_n)$ và $\theta_2 = \varphi_2(x_1, x_2, \ldots, x_n)$.

Bằng cách thiết lập các thống kê $\varphi_1(X_1, X_2, \ldots, X_n)$, $\varphi_2(X_1, X_2, \ldots, X_n)$ phù hợp thì các giá trị θ_1, θ_2 tính được từ mẫu cụ thể w sẽ thỏa mãn điều kiện $P(\theta_1 < \theta < \theta_2) = 1-\alpha$.

Để ước lượng khoảng ta thực hiện các bước sau đây:

Bước 1: Xây dựng thống kê $G = G(X_1, X_2, \ldots, X_n, \theta)$ sao cho có thể xấp xỉ luật phân phối của G về một trong những luật phân phối thường gặp như phân phối Chuẩn tắc, phân phối Student hay phân phối Chi bình phương,...

Bước 2: Với mức ý nghĩa α cho trước, tìm cặp giá trị g_1, g_2 thỏa: $P(g_1 < G < g_2) = 1-\alpha$. Biến đổi về dạng $P(\theta_1 < \theta < \theta_2) = 1 - \alpha$. Khi đó (θ_1, θ_2) chính là khoảng tin cậy cần tìm.

Tùy theo yêu cầu của bài toán mà ta có các dạng ước lượng khác nhau:

▷ *Ước lượng đối xứng:* Là khoảng ước lượng có dạng $(\theta_0 - \varepsilon; \theta_0 + \varepsilon)$, thỏa điều kiện: $P(\theta_0 - \varepsilon < \theta < \theta_0 + \varepsilon) = 1 - \alpha$. Trong đó, θ_0 là ước lượng điểm của tham số lý thuyết θ.

Như vậy nếu ε tăng thì độ tin cậy của ước lượng cũng tăng và ngược lại nếu ε giảm thì độ tin cậy của ước lượng cũng giảm; ε được gọi là *độ chính xác* của ước lượng.

▷ *Ước lượng trái (hay ước lượng tối đa):* Là khoảng ước lượng có dạng $(-\infty; \theta_0 + \varepsilon)$, thỏa mãn điều kiện: $P(\theta < \theta_0 + \varepsilon) = 1 - \alpha$.

▷ *Ước lượng phải (hay ước lượng tối thiểu):* Là khoảng ước lượng có dạng $(\theta_0 - \varepsilon; +\infty)$, thỏa mãn điều kiện: $P(\theta_0 - \varepsilon < \theta) = 1 - \alpha$.

6.2.2. Ước lượng trung bình

Bài toán: Giả sử tổng thể $X \sim N(\mu;\sigma^2)$, $\mathrm{E}(X)=\mu$ chưa biết. Từ mẫu cụ thể $w=(x_1,x_2,\ldots,x_n)$ và với độ tin cậy $1-\alpha$ cho trước, ta xây dựng khoảng tin cậy $(\mu_1;\mu_2)$ để ước lượng trung bình cho μ sao cho $P(\mu_1<\mu<\mu_2)=1-\alpha$.

Tùy thuộc vào đặc điểm của đại lượng ngẫu nhiên X đã biết phương sai hay chưa và kích thước mẫu n có lớn hay không mà ta có công thức xác lập nên khoảng tin cậy $(\mu_1;\mu_2)$ khác nhau. Xét các trường hợp sau:

Trường hợp 1: Phương sai (σ^2) đã biết.

Chọn thống kê $U=\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}\sim N(0,1)$ để ước lượng trung bình. Trong đó $\mathrm{E}(X)=\mu$ chưa biết, $\mathrm{D}(X)=\sigma^2$ đã biết.

◦ *Khoảng ước lượng đối xứng:*

$$1-\alpha=P(|U|<z_{\alpha/2})=P(-z_{\alpha/2}<U<z_{\alpha/2})=P\left(-z_{\alpha/2}<\frac{\overline{X}-\mu}{\sigma}\sqrt{n}<z_{\alpha/2}\right)$$

$$\Leftrightarrow 1-\alpha=P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\overline{X}-\mu<z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)=P\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Vậy khoảng ước lượng cần tìm là $(\overline{x}-\varepsilon;\overline{x}+\varepsilon)$ với $\varepsilon=z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.

◦ *Khoảng ước lượng trái:*

$$1-\alpha=P(-z_\alpha<U)=P\left(-z_\alpha<\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\right)$$

$$\Leftrightarrow 1-\alpha=P\left(-z_\alpha\frac{\sigma}{\sqrt{n}}<\overline{X}-\mu\right)=P\left(\mu<\overline{X}+z_\alpha\frac{\sigma}{\sqrt{n}}\right).$$

Vậy khoảng ước lượng cần tìm là $(-\infty;\overline{x}+\varepsilon)$ với $\varepsilon=z_\alpha\dfrac{\sigma}{\sqrt{n}}$.

◦ *Khoảng ước lượng phải:*

$$1-\alpha=P(U<z_\alpha)=P\left(\frac{\overline{X}-\mu}{\sigma}\sqrt{n}<z_\alpha\right)$$

$$\Leftrightarrow 1-\alpha=P\left(\overline{X}-\mu<z_\alpha\frac{\sigma}{\sqrt{n}}\right)=P\left(\overline{X}-z_\alpha\frac{\sigma}{\sqrt{n}}<\mu\right).$$

Vậy khoảng ước lượng cần tìm là $(\overline{x}-\varepsilon;+\infty)$ với $\varepsilon=z_\alpha\dfrac{\sigma}{\sqrt{n}}$.

Trường hợp 2: Phương sai σ^2 chưa biết, cỡ mẫu $n\geq 30$.

Chọn thống kê $U=\dfrac{\overline{X}-\mu}{S}\sqrt{n}\sim T(n-1)$ để ước lượng trung bình. Trong đó μ và σ^2 chưa biết. Vì $n\geq 30$ nên ta xấp xỉ $T(n-1)$ với $N(0,1)$. Hoàn toàn tương tự như trên có:

- *Khoảng ước lượng đối xứng:* $(\overline{x}-\varepsilon;\overline{x}+\varepsilon)$ với $\varepsilon = z_{\alpha/2}\dfrac{s}{\sqrt{n}}$.
- *Khoảng ước lượng trái:* $(-\infty;\overline{x}+\varepsilon)$ với $\varepsilon = z_{\alpha}\dfrac{s}{\sqrt{n}}$.
- *Khoảng ước lượng phải:* $(\overline{x}-\varepsilon;+\infty)$ với $\varepsilon = z_{\alpha}\dfrac{s}{\sqrt{n}}$.

Trường hợp 3: Phương sai σ^2 chưa biết, cỡ mẫu $n < 30$.

Chọn thống kê $T = \dfrac{\overline{X}-\mu}{S}\sqrt{n} \sim T(n-1)$ để ước lượng trung bình. Trong đó μ và σ^2 chưa biết. Do hàm mật độ của phân phối Student cũng là hàm chẵn nên tương tự như trên có:

- *Khoảng ước lượng đối xứng:* $(\overline{x}-\varepsilon;\overline{x}+\varepsilon)$ với $\varepsilon = t_{(n-1,\alpha/2)}\dfrac{s}{\sqrt{n}}$.
- *Khoảng ước lượng trái:* $(-\infty;\overline{x}+\varepsilon)$ với $\varepsilon = t_{(n-1,\alpha)}\dfrac{s}{\sqrt{n}}$.
- *Khoảng ước lượng phải:* $(\overline{x}-\varepsilon;+\infty)$ với $\varepsilon = t_{(n-1,\alpha)}\dfrac{s}{\sqrt{n}}$.

Các dạng toán liên quan: Có 2 dạng.

Dạng 1: Nếu biết ε, n $(n \geq 30)$ và s hay σ của ước lượng đối xứng thì có thể xác định độ tin cậy

$$1-\alpha = 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)-1 \text{ hay } 1-\alpha = 2\Phi\left(\frac{\varepsilon\sqrt{n}}{s}\right)-1.$$

Dạng 2: Nếu biết $\varepsilon, 1-\alpha$ và s hay σ của ước lượng đối xứng thì có thể xác định cỡ mẫu

$$n = \left[\left(z_{\alpha/2}\frac{\sigma}{\varepsilon}\right)^2\right]+1 \text{ hay } n = \left[\left(z_{\alpha/2}\frac{s}{\varepsilon}\right)^2\right]+1.$$

Trong đó, $[a]$ là phần nguyên của số a.

Ví dụ 6.6. Tiến hành thống kê giá của một nguyên liệu A trong khoảng thời gian 50 ngày, ta có bảng số liệu sau đây:

Giá (nghìn đồng/kg)	$(29;31]$	$(31;33]$	$(33;35]$	$(35;37]$	$(37;39]$	$(39;41]$
Số ngày	3	9	14	15	7	2

Với độ tin cậy 95% có thể nói giá trung bình của nguyên liệu A nằm trong khoảng nào?

a) Biết $\sigma^2 = 5,0625$.

b) Chưa biết σ^2.

c) Giả thiết như câu a), để có độ chính xác là $0,5$ thì độ tin cậy lúc đó là bao nhiêu?

d) Giả thiết như câu a), để độ tin cậy là 99% và độ chính xác là $0,6$ thì cần lấy mẫu thêm bao nhiêu ngày nữa?

Giải:

a) Chọn thống kê $U = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$ để ước lượng trung bình.

Với mẫu đã cho có: $\overline{x} = 34,8; \sigma = \sqrt{5,0625} = 2,25; n = 50$. Theo giả thiết: $1-\alpha = 0,95$ nên $\alpha/2 = 0,025$ tra Phụ lục 3, ta được: $z_{\alpha/2} = z_{0,025} = 1,96$.

Độ chính xác: $\varepsilon = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} = 1,96\dfrac{2,25}{\sqrt{50}} = 0,624$.

Vậy với độ tin cậy 95% ta có thể nói giá của nguyên liệu A trong khoảng $(\overline{x}-\varepsilon; \overline{x}+\varepsilon) = (34,8-0,624; 34,8-0,624) = (34,176; 35,424)$ (nghìn đồng/kg).

b) Chọn thống kê $U = \dfrac{\overline{X} - \mu}{S}\sqrt{n} \sim N(0,1)$ để ước lượng trung bình.

Với mẫu đã cho có: $\overline{x} = 34,8; s = 2,458; n = 50$. Theo giả thiết: $1-\alpha = 0,95$ nên $\alpha/2 = 0,05$ tra Phụ lục 3 có $z_{\alpha/2} = z_{0,025} = 1,96$.

Độ chính xác: $\varepsilon = z_{\alpha/2}\dfrac{s}{\sqrt{n}} = 1,96\dfrac{2,458}{\sqrt{50}} = 0,681$.

Vậy với độ tin cậy 95% ta có thể nói giá của nguyên liệu A nằm trong khoảng $(\overline{x}-\varepsilon; \overline{x}+\varepsilon) = (34,8-0,681; 34,8+0,681) = (34,119; 35,481)$ (nghìn đồng/kg).

c) Theo giả thiết có $\varepsilon = 0,5; n = 50; \sigma = 2,25$ ta cần tìm độ tin cậy:

$$1-\alpha = 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1 = 2\Phi\left(\frac{0,5\sqrt{50}}{2,25}\right) - 1 = 2\Phi(1,571) - 1 = 88,38\%.$$

d) Theo giả thiết có $\varepsilon = 0,6; \sigma = 2,25; 1-\alpha = 0,99$ ta tìm n:

$$n = \left[\left(z_{\alpha/2}\frac{\sigma}{\varepsilon}\right)^2\right] + 1 = \left[\left(z_{0,005}\frac{2,25}{0,6}\right)^2\right] + 1 = \left[\left(2,57\frac{2,25}{0,6}\right)^2\right] + 1 = [93,606] + 1 = 94.$$

Do đó, số ngày cần kiểm tra thêm là $94 - 50 = 44$ ngày.

Ví dụ 6.7. Đo đường kính 23 trục máy do một máy tiện sản xuất tự động, được kết quả (đơn vị tính: mm).

250	249	251	253	248	250	252	257
245	248	247	249	250	280	250	247
253	256	252	254	254	251	253	

Giả sử đường kính của trục máy là đại lượng ngẫu nhiên X có phân phối chuẩn.

a) Hãy ước lượng đường kính trung bình các trục máy với độ tin cậy 98%.

b) Hãy ước lượng trung bình tối thiểu đường kính các trục máy với độ tin cậy 90%.

Giải: Đây là bài toán ước lượng trung bình, phương sai chưa biết, $n = 23 < 30$.

Chọn thống kê $T = \dfrac{\overline{X} - \mu}{S}\sqrt{n} \sim T(n-1)$ để ước lượng trung bình.

a) Với mẫu đã cho, ta có: $\overline{x} = 252,13; s = 6,75; n = 23$. Độ tin cậy $1-\alpha = 0,98$ nên $\alpha/2 = 0,01$ tra Phụ lục 5 được $t_{(n-1;\alpha/2)} = t_{(22;0,01)} = 2,508$.

Độ chính xác: $\varepsilon = t_{(22;0,01)}\dfrac{s}{\sqrt{n}} = 2,508\dfrac{6,75}{\sqrt{23}} = 3,53$.

Với độ tin cậy 98% thì đường kính trung bình của trục máy nằm trong khoảng $(\overline{x}-\varepsilon;\overline{x}+\varepsilon) = (252,13-3,53;252,13+3,53) = (248,6;255,66)$ (mm).

b) Với mẫu đã cho, ta có: $\overline{x} = 252,13; s = 6,75; n = 23$. Độ tin cậy $1-\alpha = 0,9$ nên $\alpha = 0,1$ tra Phụ lục 5 có $t_{(n-1;\alpha)} = t_{(22;0,1)} = 1,321$.

Độ chính xác: $\varepsilon = t_{(22;0,1)}\dfrac{s}{\sqrt{n}} = 1,321\dfrac{6,75}{\sqrt{23}} = 1,859$.

Vậy nên khoảng ước lượng cho trung bình tối thiểu đường kính của trục máy là $(\overline{x}-\varepsilon;+\infty) = (252,13-1,321;+\infty) = (250,271;+\infty)$ (mm).

6.2.3. Ước lượng tỉ lệ

Bài toán: Giả sử tổng thể X có tỉ lệ p các phần tử mang dấu hiệu A nào đó chưa biết. Từ mẫu cụ thể $w = (x_1, x_2, ..., x_n)$ và độ tin cậy $1-\alpha$ cho trước, ta xây dựng khoảng tin cậy $(p_1;p_2)$ để ước lượng tỉ lệ cho p sao cho $P(p_1 < p < p_2) = 1-\alpha$.

Giả sử mẫu n quan sát có m phần tử mang dấu hiệu A. Khi đó tần suất mẫu $f = \dfrac{m}{n}$.

Chọn thống kê $U = \dfrac{F-p}{\sqrt{F(1-F)}}\sqrt{n} \simeq N(0,1)$ để ước lượng tỉ lệ. Trong đó p là tỉ lệ chưa biết, n là cỡ mẫu và F là thống kê nhận giá trị bằng tần suất mẫu. Tương tự ước lượng trung bình, ta có:

- *Khoảng ước lượng đối xứng:*

$$1-\alpha = P(|U| < z_{\alpha/2}) = P\left(F - z_{\alpha/2}\sqrt{\frac{F(1-F)}{n}} < p < F + z_{\alpha/2}\sqrt{\frac{F(1-F)}{n}}\right).$$

Vậy khoảng tin cậy cần tìm là $(f-\varepsilon;f+\varepsilon)$ với $\varepsilon = z_{\alpha/2}\sqrt{\dfrac{f(1-f)}{n}}$. Để ước lượng được tốt thì f không quá gần 0 hoặc 1 hay thỏa mãn $nf \geq 10$ và $n(1-f) \geq 10$.

- *Khoảng ước lượng trái:* $[0;f+\varepsilon)$ với $\varepsilon = z_\alpha\sqrt{\dfrac{f(1-f)}{n}}$.

- *Khoảng ước lượng phải:* $(f-\varepsilon;1]$ với $\varepsilon = z_\alpha\sqrt{\dfrac{f(1-f)}{n}}$.

Các dạng toán liên quan: Có 2 dạng.

Dạng 1: Nếu biết ε, n, f của ước lượng đối xứng thì có thể xác định độ tin cậy

$$1-\alpha = 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{f(1-f)}}\right) - 1.$$

Dạng 2: Nếu biết $\varepsilon, 1-\alpha, f$ của ước lượng đối xứng thì có thể xác định cỡ mẫu

$$n = \left[\left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 f(1-f)\right] + 1.$$

Chú ý: Giả sử N là kích thước của tổng thể.

i) Với mẫu không hoàn lại có cỡ mẫu $n > \dfrac{N}{10}$.

+ Để tính độ chính xác ta cần nhân thêm hệ số hiệu chỉnh:

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

+ Kích thước mẫu cần điều tra được tính theo công thức:

$$n = \left[\frac{n_1 N}{n_1 + (N-1)}\right] + 1 \text{ với } n_1 = \left[\left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 f(1-f)\right] + 1.$$

ii) Khoảng ước lượng cho số phần tử mang dấu hiệu A là:

$$\Big([Nf - N\varepsilon]; [Nf + N\varepsilon] + 1\Big).$$

Ví dụ 6.8. Giám đốc một ngân hàng muốn xác định số khách hàng vay tiền chậm trả lãi suất tại ngân hàng. Lấy mẫu ngẫu nhiên 120 khách thì có 40 người chưa trả lãi theo đúng kỳ hạn quy định. Biết ngân hàng đang có 2000 khách vay tiền.

a) Với độ tin cậy 90%, hãy ước lượng số khách hàng chậm trả lãi suất.

b) Nếu muốn ước lượng tỉ lệ khách hàng chậm trả lãi suất với độ tin cậy như câu a) và độ chính xác $\varepsilon = 0,05$ thì cần kích thước mẫu là bao nhiêu?

c) Giải câu b), trong trường hợp ngân hàng chỉ có 1000 khách đang vay tiền.

d) Với mẫu đã cho, nếu muốn ước lượng tỉ lệ khách hàng chậm trả lãi suất có độ chính xác $\varepsilon = 0,05$ thì độ tin cậy lúc đó là bao nhiêu?

Giải:

a) Chọn thống kê $U = \dfrac{F-p}{\sqrt{F(1-F)}}\sqrt{n} \simeq N(0,1)$ để ước lượng tỉ lệ.

Với mẫu đã cho, có: $f = \dfrac{40}{120} = 0,333$. Độ tin cậy $1-\alpha = 0,9$ nên $\alpha/2 = 0,05$ tra Phụ lục 3 được $z_{\alpha/2} = z_{0,05} = 1,64$.

Do ngân hàng có 2000 khách vay tiền: $\dfrac{N}{10} = 200 > 120 = n$. Cho nên độ chính xác: $\varepsilon = z_{\alpha/2}\sqrt{\dfrac{f(1-f)}{n}} = 1,64\sqrt{\dfrac{0,333\cdot 0,667}{120}} = 0,0706.$

Khoảng ước lượng tỉ lệ khách hàng chậm trả lãi suất là $(f-\varepsilon; f+\varepsilon) = (0,2624; 0,4036)$.

Ta có: $[N(f-\varepsilon)] = [2000\cdot 0,2624] = 524$ và $[N(f+\varepsilon)] + 1 = [2000\cdot 0,4036] + 1 = 808$.

Vậy nên có từ 524 đến 808 khách hàng chậm trả lãi suất.

b) Với độ chính xác $\varepsilon = 0,05$ và độ tin cậy $1 - \alpha = 0,9$. Ta có:

$$n = \left[\left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 f(1-f) \right] + 1 = \left[\left(\frac{1,64}{0,05} \right)^2 0,333 \cdot 0,667 \right] + 1 = 239.$$

Nên kích thước mẫu cần điều tra là 239 người.

c) Nếu công ty chỉ có 1000 khách vay: $\dfrac{N}{10} = 100 < 120$ thì cỡ mẫu cần điều tra là:

$$n = \left[\frac{239 \cdot 1000}{239 + (1000-1)} \right] + 1 = 194.$$

d) Ta có: $\varepsilon = 0,05; f = 0,333$ và $n = 120$ nên

$$1 - \alpha = 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{f(1-f)}} \right) - 1 = 2\Phi\left(\frac{0,05\sqrt{120}}{\sqrt{0,333 \cdot 0,667}} \right) - 1 = 2 \cdot \Phi(1,162) - 1 = 0,7548.$$

Vậy độ tin cậy lúc này là $75,48\%$.

6.2.4. Ước lượng phương sai

Bài toán: Giả sử tổng thể $X \sim N(\mu, \sigma^2)$, $\text{Var}(X) = \sigma^2$ chưa biết. Từ mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$ và độ tin cậy $1 - \alpha$ cho trước, hãy tìm khoảng tin cậy $(\sigma_1^2; \sigma_2^2)$ để ước lượng cho σ^2 sao cho $P(\sigma_1^2 < \sigma^2 < \sigma_2^2) = 1 - \alpha$. Để ước lượng ta xét 2 trường hợp:

Trường hợp 1: Trung bình μ chưa biết.

Chọn thống kê $K = \dfrac{n-1}{\sigma^2} S^2 \sim \chi^2_{(n-1)}$ để ước lượng phương sai σ^2.

∘ *Khoảng tin cậy đối xứng:*

$$\begin{aligned} 1 - \alpha &= P\left(\chi^2_{(n-1,1-\alpha/2)} < \frac{n-1}{\sigma^2} S^2 < \chi^2_{(n-1,1-\alpha/2)} \right) \\ &= P\left(\frac{\chi^2_{(n-1,1-\alpha/2)}}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{(n-1,1-\alpha/2)}}{(n-1)S^2} \right) = P\left(\frac{(n-1)S^2}{\chi^2_{(n-1,1-\alpha/2)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1,1-\alpha/2)}} \right). \end{aligned}$$

Vậy khoảng ước lượng cần tìm là $(\sigma_1^2; \sigma_2^2)$ với $\sigma_1^2 = \dfrac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha/2)}}$ và $\sigma_2^2 = \dfrac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha/2)}}$.

∘ *Khoảng ước lượng trái:*

$$1 - \alpha = P\left(\chi^2_{(n-1,1-\alpha)} < \frac{n-1}{\sigma^2} S^2 \right) = P\left(\sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1,1-\alpha)}} \right).$$

Vậy khoảng ước lượng cần tìm là $\left(0; \dfrac{(n-1)s^2}{\chi^2_{(n-1,1-\alpha)}} \right)$.

∘ *Khoảng ước lượng phải:*

$$1-\alpha = P\left(\frac{n-1}{\sigma^2}S^2 < \chi^2_{(n-1,\alpha)}\right) = P\left(\sigma^2 > \frac{(n-1)S^2}{\chi^2_{(n-1,\alpha)}}\right).$$

Vậy khoảng tin cậy cần tìm là $\left(\dfrac{(n-1)s^2}{\chi^2_{(n-1,\alpha)}}; +\infty\right)$.

Trường hợp 2: Trung bình μ đã biết.

Chọn thống kê $K = \dfrac{nS^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2_{(n)}$ để ước lượng phương sai σ^2.

∘ *Khoảng ước lượng đối xứng:* $(\sigma_1^2; \sigma_2^2)$ với $\sigma_1^2 = \dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{(n,\alpha/2)}}$ và $\sigma_2^2 = \dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{(n,1-\alpha/2)}}$.

∘ *Khoảng ước lượng trái:* $\left(0; \dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{(n,1-\alpha)}}\right)$.

∘ *Khoảng ước lượng phải:* $\left(\dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{(n,\alpha)}}; +\infty\right)$.

Ví dụ 6.9. Mức hao phí nguyên liệu cho một sản phẩm X có luật phân phối chuẩn. Thống kê nguyên liệu hao phí của một mẫu gồm 30 sản phẩm X, được kết quả cho trong bảng sau:

Nguyên liệu hao phí (g)	19,5	20,0	20,5
Số sản phẩm	7	18	5

Với độ tin cậy 90% hãy tìm khoảng tin cậy cho độ lệch chuẩn của mức hao phí nguyên liệu trên trong 2 trường hợp:

a) E(X) chưa biết.

b) E(X) $= 20$ g.

Giải: Với mẫu đã cho, ta có:

$$\overline{x} = \frac{1}{30}\sum_{i=1}^{3} n_i x_i = \frac{1}{30}(7.19,5 + 18.20 + 5.20,5) = 19,97.$$

$$\sum_{i=1}^{3} n_i(x_i-\mu)^2 = 7(19,5-20)^2 + 18(20-20)^2 + 5(20,5-20)^2 = 3.$$

$$(n-1)s^2 = \sum_{i=1}^{3} n_i(x_i-\overline{x})^2 = 7(19,5-19,97)^2 + 18(20-19,97)^2 + 5(20,5-19,97)^2 = 2,967.$$

a) Có $1-\alpha = 0,9 \Rightarrow \alpha/2 = 0,05$ tra bảng Phụ lục 4 được $\chi_{(n-1;\alpha/2)} = \chi^2_{(29;0,05)} = 42,557$ và $\chi_{(n-1;1-\alpha/2)} = \chi^2_{(29;0,95)} = 17,708$.

Khoảng ước lượng cho σ^2: $\left(\dfrac{2,967}{42,557}; \dfrac{2,967}{17,708}\right) = (0,06972; 0,1676)\ (g^2)$.

Khoảng ước lượng cho độ lệch chuẩn σ: $(0,2633; 0,4093)\ (g)$.

b) Ta có: $\chi_{(n;\alpha/2)} = \chi^2_{(30;0,05)} = 43,773$ và $\chi_{(n;1-\alpha/2)} = \chi^2_{(30;0,95)} = 18,439$.

Khoảng ước lượng cho σ^2: $\left(\dfrac{3}{43,773}; \dfrac{3}{18,439}\right) = (0,0685; 0,1627)\ (g^2)$.

Khoảng ước lượng cho độ lệch chuẩn σ: $(0,2618; 0,4033)\ (g)$.

6.3. BÀI TẬP CHƯƠNG 6

Bài 1: Kiểm tra ngẫu nhiên 66 sản phẩm của một lô hàng, ta thu được kết quả sau:

Trọng lượng (g)	$[3500; 3520]$	$(3520; 3540]$	$(3540; 3560]$	$(3560; 3580]$	$(3580; 3600]$
Số sản phẩm	9	14	16	20	7

Với độ tin cậy 90% hãy ước lượng đối xứng trung bình khối lượng sản phẩm của lô hàng đó trong các trường hợp sau:

a) Đã biết $\sigma = 23\ g$.

b) Chưa biết σ.

Bài 2: Khảo sát hàm lượng kẽm có trong nước ở một số con sông, thu được số liệu:

Hàm lượng (mg/l)	$(0; 1]$	$(1; 2]$	$(2; 3]$	$(3; 4]$
Số nơi khảo sát	10	17	15	5

Với độ tin cậy 97%, hãy cho biết hàm lượng kẽm trung bình tối đa và tối thiểu có trong nước sông.

Bài 3: Thống kê số lượng tiền cho vay trong tháng ở ngân hàng, được bảng số liệu:

Tiền (triệu đồng)	$(5; 15]$	$(15; 25]$	$(25; 35]$	$(35; 45]$
Số lượt cho vay	8	10	7	3

Biết đại lượng ngẫu nhiên chỉ số tiền cho vay có luật phân phối chuẩn.

a) Với độ tin cậy 90%, hãy cho biết số tiền trung bình mỗi lượt vay ở ngân hàng đó.

b) Biết trung bình một lượt cho vay của ngân hàng đó là 20 triệu đồng, hãy ước lượng phương sai của số tiền cho vay với độ tin cậy 95%.

Bài 4: Đo độ pH trong đất nông nghiệp ở một địa phương, được kết quả cho ở bảng:

Độ pH	4,5	4,6	4,7	4,8	4,9	5,0	5,1	5,2
Số mẫu thử	3	7	10	9	7	5	6	2

a) Hãy đánh giá độ pH trung bình trong đất nông nghiệp ở địa phương đó với độ tin cậy 98%.

b) Nếu muốn độ tin cậy 99%, sai số không quá 0,05 thì cần lấy thêm bao nhiêu mẫu thử nữa?

Bài 5: Lấy mẫu cá giống được nuôi trong hồ, ta có bảng số liệu như sau:

Trọng lượng (g)	Số cá
$(600; 620]$	5
$(620; 640]$	42
$(640; 660]$	18
$(660; 680]$	27
$(680; 700]$	8

a) Tìm ước lượng trung bình tối đa và trung bình tối thiểu của những con cá giống trong hồ với độ tin cậy 95%.

b) Cá giống loại A là cá giống có trọng lượng lớn hơn 660 g. Hãy ước lượng tỷ lệ cá loại A trong hồ; ước lượng tối đa và tối thiểu của tỷ lệ đó với độ tin cậy 99%.

Bài 6: Chọn ngẫu nhiên 36 công nhân của một xí nghiệp thì thấy lương trung bình là 5 triệu 230 ngàn đồng/tháng. Giả sử lương công nhân có luật phân phối chuẩn với $\sigma = 120$ ngàn đồng. Với độ tin cậy 95%, hãy ước lượng mức lương trung bình của công nhân trong toàn xí nghiệp.

Bài 7: Điểm trung bình học phần Xác suất thống kê của 80 sinh viên trong lớp là 5 với độ lệch điều chỉnh là 2.

a) Ước lượng trung bình điểm học phần Xác suất thống kê của toàn thể sinh viên trong trường với độ tin cậy là 95%.

b) Để sai số là 0,25 điểm thì độ tin cậy là bao nhiêu.

Bài 8: Tuổi thọ trung bình của một loại bóng đèn có độ lệch chuẩn 100 giờ.

a) Chọn ngẫu nhiên 100 bóng để thử nghiệm, thấy tuổi thọ trung bình là 1000 giờ. Hãy ước lượng tuổi thọ trung bình của bóng đèn với độ tin cậy 95%.

b) Với độ chính xác là 15 giờ. Hãy xác định độ tin cậy.

c) Để độ chính xác là 25 giờ và độ tin cậy là 95% thì cần thử nghiệm bao nhiêu bóng?

Bài 9: Sau 5 năm trồng mới rừng tràm, kiểm tra 20 cây ở lâm trường người ta thấy chiều cao trung bình là 120 cm và độ lệch điều chỉnh mẫu là 20 cm. Với độ tin cậy 95%, hãy ước lượng chiều cao trung bình tối đa và tối thiểu của cây tràm ở lâm trường.

Bài 10: Để ước lượng tỷ lệ đồ hộp quá hạn dùng trong một kho chứa, người ta kiểm tra ngẫu nhiên 100 hộp thấy có 11 hộp quá hạn dùng.

a) Ước lượng tỷ lệ đồ hộp quá hạn dùng của kho với độ tin cậy 94%.

b) Với sai số cho phép là 3%, hãy xác định độ tin cậy.

Bài 11: Một lô hàng có 5000 sản phẩm. Chọn ngẫu nhiên 400 sản phẩm từ lô hàng để kiểm tra thì thấy có 360 sản phẩm loại A.

a) Hãy ước lượng số sản phẩm loại A có trong lô hàng với độ tin cậy 96%.

b) Nếu muốn ước lượng số sản phẩm loại A của lô hàng đạt được độ chính xác 150 sản phẩm và độ tin cậy 99% thì phải kiểm tra bao nhiêu sản phẩm?

Bài 12: Lô trái cây của một chủ hàng được đóng thành sọt, mỗi sọt có 100 trái. Kiểm tra 50 sọt thấy có 450 trái không đạt tiêu chuẩn.

a) Ước lượng tỷ lệ trái cây không đạt tiêu chuẩn của lô hàng với độ tin cậy 95%.

b) Muốn ước lượng tỷ lệ trái cây không đạt tiêu chuẩn với độ chính xác 1% thì độ tin cậy đạt được là bao nhiêu?

c) Muốn ước lượng tỷ lệ trái cây không đạt tiêu chuẩn với độ tin cậy 99% và độ chính xác 1% thì cần kiểm tra bao nhiêu sọt?

d) Muốn ước lượng tỷ lệ trái cây không đạt tiêu chuẩn với độ tin cậy $99,7\%$ thì độ chính xác đạt được là bao nhiêu?

Bài 13: Điều tra năng suất lúa trên diện tích 100 hecta trồng lúa của một vùng, ta thu được bảng số liệu như sau:

Năng suất (tấn/hecta)	6	$6,5$	7	$7,5$	8	$8,5$	9
Số hecta	10	20	30	15	10	10	5

a) Hãy ước lượng năng suất lúa trung bình của vùng đó với độ tin cậy 95%.

b) Những thửa ruộng có năng suất từ 8 tấn/hecta trở lên là những thửa có năng suất cao. Hãy ước lượng tỷ lệ diện tích có năng suất cao trong vùng với độ tin cậy 97%.

Bài 14: Đo đường kính của 100 chi tiết do một máy sản xuất, được kết quả như sau:

Đường kính (mm)	Số chi tiết
$(19,80; 19,85]$	3
$(19,85; 19,90]$	5
$(19,90; 19,95]$	16
$(19,95; 20,00]$	28
$(20,00; 20,05]$	23
$(20,05; 20,10]$	14
$(20,10; 20,15]$	7
$(20,15; 20,20]$	4

Quy định những chi tiết có đường kính lớn hơn $19,90$ mm đến nhỏ hơn hoặc bằng $20,10$ mm là những chi tiết đạt tiêu chuẩn.

a) Ước lượng đường kính trung bình của chi tiết đạt tiêu chuẩn với độ tin cậy 98%.

b) Ước lượng tỷ lệ chi tiết đạt tiêu chuẩn với độ tin cậy 95%.

Bài 15: Khảo sát trọng lượng trẻ sơ sinh ở một địa phương, được bảng số liệu sau:

Trọng lượng (kg)	$(2; 3]$	$(3; 4]$	$(4; 5]$
Số trẻ	12	24	4

a) Tìm khoảng tin cậy 95% của trung bình trọng lượng trẻ sơ sinh ở địa phương đó.

b) Để có độ chính xác $0,5$ kg, độ tin cậy 99%, thì có cần lấy thêm mẫu hay không?

Bài 16: Kiểm tra trọng lượng 16 cái bánh sau khi làm ra, ta thu được kết quả sau:

Trọng lượng bánh (g)	$(90;95]$	$(95;100]$	$(100;105]$	$(105;110]$
Số bánh	3	4	5	4

Với độ tin cậy 90% hãy ước lượng phương sai của trọng lượng bánh trong 2 trường hợp sau:

a) Biết $\mu = 95\ g$.

b) Chưa biết μ.

Bài 17: Trong một mẫu gồm có 30 chi tiết máy được kiểm tra, ta tính được $s = 0,32\ cm$. Tìm khoảng ước lượng cho phương sai và độ lệch tiêu chuẩn của kích thước của toàn bộ các chi tiết máy với độ tin cậy 95%.

Bài 18: Kiểm tra về sức chịu lực R (kg/cm^2) của một sản phẩm mới. Người ta lấy một mẫu gồm có 28 sản phẩm để kiểm tra, được kết quả như sau:

10,0	13,0	13,7	11,5	11,0	13,5	12,2
13,0	10,0	11,0	13,5	11,5	13,0	12,2
13,5	10,0	10,0	11,5	13,0	13,7	14,0
13,0	13,7	13,0	11,5	10,0	11,0	13,0

Với độ tin cậy 95% hãy ước lượng:

a) Sức chịu lực trung bình của sản phẩm đó.

b) Phương sai của sức chịu lực.

Chương 7

KIỂM ĐỊNH GIẢ THUYẾT

7.1. KIỂM ĐỊNH GIẢ THUYẾT

Kiểm định giả thuyết là một vấn đề quan trọng của thống kê. Nội dung của công việc kiểm định là căn cứ các số liệu thu được để đưa ra kết luận về một giả thuyết thống kê nào đó.

Để giải quyết bài toán trên, thông tin duy nhất mà chúng ta có là một mẫu ngẫu nhiên. Trong quá trình kiểm định, người kiểm định cần liên hệ với thực tế để có thể đưa ra các kết luận sao cho ngày càng chính xác hơn.

7.1.1. Giả thuyết thống kê

Giả thuyết thống kê là giả thuyết nói về dạng phân phối xác suất; về giá trị của các tham số hoặc là về tính độc lập giữa các đại lượng ngẫu nhiên.

Giả thuyết không (Null hypothesis), ký hiệu H_0, là mệnh đề xác định giá trị cụ thể cho tham số θ chưa biết. Mệnh đề đối lập với giả thuyết H_0 được gọi là *đối thuyết* (alternative hypothesis), có ba dạng đối thuyết là $H_1 : \theta \neq \theta_0$, $H_2 : \theta > \theta_0$ và $H_3 : \theta < \theta_0$ với θ_0 là giá trị lý thuyết của các tham số.

Ví dụ 7.1. Một nhà sản xuất tuyên bố rằng bóng đèn của hãng được thiết kế có tuổi thọ trung bình là 12000 giờ. Một mẫu gồm n bóng đèn loại này được chọn để kiểm tra tuổi thọ. Hãy phát biểu giả thuyết không và đối thuyết.

Giải: Gọi μ là tuổi thọ trung bình của bóng đèn.

Giả thuyết không $H_0 : \mu = 12000$ giờ; đối thuyết $H_3 : \mu < 12000$ giờ.

Ví dụ 7.2. Chủ một nhà hàng kinh doanh thức ăn nhanh tuyên bố rằng thời gian chờ phục vụ của khách hàng là không quá 60 giây. Một mẫu đo thời gian chờ phục vụ của n khách khi đến nhà hàng đó. Hãy phát biểu giả thuyết không và đối thuyết.

Giải: Gọi μ là trung bình thời gian chờ phục vụ của khách hàng.

Giả thuyết không $H_0 : \mu = 60$ giây; đối thuyết $H_2 : \mu > 60$ giây.

Ví dụ 7.3. Chủ một trang trại chăn nuôi cho rằng trọng lượng trung bình của lợn chín tháng tuổi ở trang trại là 100 kg. Một mẫu gồm n con lợn cùng độ tuổi nêu trên ở trang trại được cân thử. Hãy phát biểu giả thuyết không và đối thuyết.

Giải: Gọi μ là trọng lượng trung bình của lợn chín tháng tuổi.

Giả thuyết không $H_0 : \mu = 100\ kg$; đối thuyết $H_1 : \mu \neq 100\ kg$.

7.1.2. Kiểm định giả thuyết thống kê

Kiểm định giả thuyết thống kê là việc thừa nhận hay không thừa nhận (bác bỏ) một giả thuyết nào đó dựa vào những kết quả thống kê có được. Trong kiểm định, nếu H_0 được chấp nhận thì đối thuyết của nó bị bác bỏ và ngược lại nếu H_0 bị bác bỏ thì đối thuyết của nó được chấp nhận.

Khi kiểm định giả thuyết thống kê có thể mắc phải một trong hai loại sai lầm sau:

∘ *Sai lầm loại 1:* Khi giả thuyết H_0 là đúng mà ta bác bỏ. Xác suất xảy ra sai lầm loại một, ký hiệu là α.

∘ *Sai lầm loại 2:* Khi giả thuyết H_0 là sai mà ta chấp nhận. Xác suất xảy ra sai lầm loại hai, ký hiệu là β.

Giá trị α còn được gọi là *mức ý nghĩa* của kiểm định và nó được ấn định trước khi tiến hành kiểm định. Trong thực tế người ta thường chọn giá trị α rất nhỏ: $0,01; 0,05; 0,1; \ldots$

7.1.3. Các bước kiểm định giả thuyết

Để kiểm định giả thuyết thống kê người ta thường dựa vào một tiêu chuẩn kiểm định giả thuyết là thống kê $G = G(X_1, X_2, \ldots, X_n, \theta_0)$ nào đó được lập từ mẫu ngẫu nhiên kích thước n thỏa điều kiện: *"Khi H_0 là đúng thì luật phân phối xác suất của G hoàn toàn xác định"*.

Khi có tiêu chuẩn kiểm định G với mức ý nghĩa α, ta thiết lập *miền bác bỏ* giả thuyết W_α thỏa mãn điều kiện: $P\big(G \in W_\alpha | H_0 \text{ đúng}\big) = \alpha$.

Vì α rất nhỏ nên có thể coi như biến cố $(G \in W_\alpha)$ không xảy ra. Nên với mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$ nếu $(G(x_1, x_2, \ldots, x_n) \in W_\alpha)$ xảy ra thì H_0 là sai: Ta bác bỏ H_0, chấp nhận đối thuyết. Nếu $(G(x_1, x_2, \ldots, x_n) \in W_\alpha)$ không xảy ra: Chưa có cơ sở bác bỏ H_0.

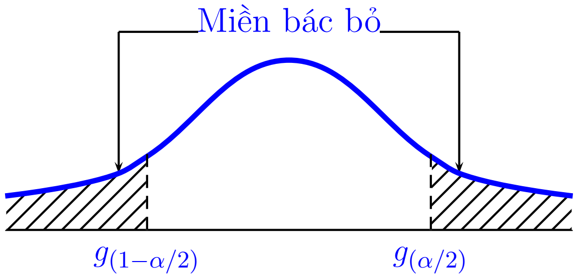
Đối với bài toán kiểm định tham số, giả thuyết $H_0 : \theta = \theta_0$ có ba dạng miền bác bỏ ứng với các đối thuyết được trình bày như sau:

▷ Trường hợp đối thuyết hai phía H_1: $\theta \neq \theta_0$.

Khi H_0 đúng, ta chọn $g_{(1-\alpha/2)}, g_{(\alpha/2)}$ là các phân vị mức $1 - \alpha/2$ và $\alpha/2$ của G sao cho

$$P\left(G > g_{(1-\alpha/2)} | H_0\right) = 1 - \frac{\alpha}{2} \Leftrightarrow P\left(G < g_{(1-\alpha/2)} | H_0\right) = \frac{\alpha}{2} \text{ và } P\left(G > g_{(\alpha/2)} | H_0\right) = \frac{\alpha}{2}$$

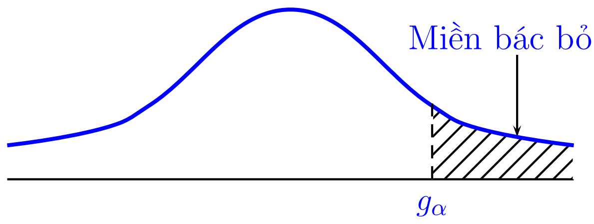
nên miền bác bỏ là $W_\alpha = \big(-\infty; g_{(1-\alpha/2)}\big) \cup \big(g_{(\alpha/2)}; +\infty\big)$.



Hình 7.1: Miền bác bỏ ứng với đối thuyết H_1

▹ Trường hợp đối thuyết lệch phải $H_2 : \theta > \theta_0$.

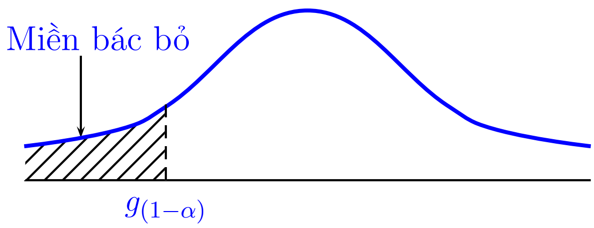
Khi H_0 đúng, ta chọn g_α sao cho $P\left(G > g_\alpha | H_0\right) = \alpha$ nên miền bác bỏ $W_\alpha = (g_\alpha; +\infty)$.



Hình 7.2: Miền bác bỏ ứng với đối thuyết H_2

▹ Trường hợp đối thuyết lệch trái $H_3 : \theta < \theta_0$.

Khi H_0 đúng, chọn $g_{(1-\alpha)}$ sao cho $P\left(G > g_{(1-\alpha)}|H_0\right) = 1 - \alpha \Leftrightarrow P\left(G < g_{(1-\alpha)}|H_0\right) = \alpha$ nên miền bác bỏ $W_\alpha = \left(-\infty; g_{(1-\alpha)}\right)$.



Hình 7.3: Miền bác bỏ ứng với đối thuyết H_3

7.2. KIỂM ĐỊNH GIẢ THUYẾT MỘT MẪU

7.2.1. Kiểm định giả thuyết về trung bình

Bài toán: Giả sử đại lượng ngẫu nhiên X có trung bình $\mathrm{E}(X) = \mu$ chưa biết. Với mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$ có kích thức n và mức ý nghĩa α, hãy kiểm định giả thuyết $H_0 : \mu = \mu_0$ với một trong các đối thuyết $H_1 : \mu \neq \mu_0; H_2 : \mu > \mu_0; H_3 : \mu < \mu_0$.

Để xác định chọn phân phối chuẩn (Z-test) hay phân phối Student (T-test) khi kiểm định chúng ta cần lần lượt trả lời các câu hỏi sau:

1) Tổng thể có luật phân phối chuẩn hay không?

2) Phương sai của tổng thể đã biết hay chưa?

3) Kích thước mẫu có lớn không ($n \geq 30$)?

Nếu phân phối của tổng thể là phân phối chuẩn và phương sai σ^2 của tổng thể đã biết thì ta sử dụng phân phối chuẩn để kiểm định. Chú ý rằng, nếu thỏa mãn cả hai điều kiện này ta không cần sử dụng Định lí giới hạn trung tâm.

Nếu phân phối của tổng thể không phải là phân phối chuẩn thì phải đảm bảo cỡ mẫu phải lớn để sử dụng Định lí giới hạn trung tâm đưa về xấp xỉ phân phối chuẩn.

Trường hợp 1: X có phân phối chuẩn, phương sai (σ^2) đã biết. Hoặc X không có phân phối chuẩn, phương sai (σ^2) đã biết và $n \geq 30$.

Chọn thống kê $Z = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \sim N(0,1)$ làm tiêu chuẩn kiểm định.

Với mẫu cụ thể ta có giá trị kiểm định: $z_0 = \dfrac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$.

Khi đó tùy theo đối thuyết mà ta xem xét bác bỏ giả thuyết H_0 như sau:

∘ Nếu đối thuyết là $H_1 : \mu \neq \mu_0$ thì ta bác bỏ H_0 khi $|z_0| > z_{\alpha/2}$.

∘ Nếu đối thuyết là $H_2 : \mu > \mu_0$ thì ta bác bỏ H_0 khi $z_0 > z_\alpha$.

∘ Nếu đối thuyết là $H_3 : \mu < \mu_0$ thì ta bác bỏ H_0 khi $z_0 < -z_\alpha$.

Trường hợp 2: Phương sai (σ^2) chưa biết và $n \geq 30$.

Chọn thống kê $Z = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n} \simeq N(0,1)$ làm tiêu chuẩn kiểm định.

Với mẫu cụ thể ta có giá trị kiểm định: $z_0 = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$.

Việc xem xét bác bỏ giả thuyết H_0 hoàn toàn tương tự *Trường hợp 1.*

Trường hợp 3: X có phân phối chuẩn, phương sai (σ^2) chưa biết và $n < 30$.

Chọn thống kê $T = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n} \sim T(n-1)$ làm tiêu chuẩn kiểm định.

Với mẫu cụ thể ta có giá trị kiểm định: $t_0 = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$.

Sau đó tùy theo đối thuyết mà ta xem xét bác bỏ giả thuyết H_0 như sau:

∘ Nếu đối thuyết là $H_1 : \mu \neq \mu_0$ thì ta bác bỏ H_0 khi $|t_0| > t_{(n-1;\alpha/2)}$.

∘ Nếu đối thuyết là $H_2 : \mu > \mu_0$ thì ta bác bỏ H_0 khi $t_0 > t_{(n-1;\alpha)}$.

∘ Nếu đối thuyết là $H_3 : \mu < \mu_0$ thì ta bác bỏ H_0 khi $t_0 < -t_{(n-1;\alpha)}$.

Ví dụ 7.4. Trọng lượng của một hộp sản phẩm do dây chuyền tự động đóng gói là 100 g, độ lệch chuẩn $\sigma = 0,8$ g. Sau một thời gian sản xuất, người ta nghi ngờ khối lượng sản phẩm có xu hướng tăng lên. Kiểm tra 60 sản phẩm tính được trung bình mẫu $\overline{x} = 100,2$ g.

a) Với độ tin cậy 99%, hãy kết luận về nghi ngờ ở trên.

b) Độ tin cậy lớn nhất có thể được là bao nhiêu để kết luận điều nghi ngờ nói trên là đúng?

Giải:

a) Xét giả thuyết $H_0 : \mu = 100$ g và đối thuyết $H_2 : \mu > 100$ g.

Chọn thống kê $Z = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n} \sim N(0,1)$ làm tiêu chuẩn kiểm định.

Với mẫu đã cho có $\mu_0 = 100; \sigma = 0,8; n = 60; \overline{x} = 100,2$ nên giá trị kiểm định:

$$z_0 = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = \frac{100,2 - 100}{0,8}\sqrt{60} = 1,936.$$

Độ tin cậy $1 - \alpha = 0,99$ nên $\alpha = 0,01$ tra Phụ lục 3, ta được: $z_\alpha = z_{0,01} = 2,327$.

Do $z_\alpha > z_0$ nên chấp nhận H_0, bác bỏ H_2. Vì thế chưa đủ cơ sở cho rằng khối lượng sản phẩm tăng lên.

b) Giả sử với độ tin cậy $1 - \alpha$ điều nghi ngờ ở trên là đúng. Nghĩa là, $z_0 > z_\alpha$ suy ra $\Phi(z_0) > \Phi(z_\alpha) = 1 - \alpha$ hay $0,9736 > 1 - \alpha$.

Vậy độ tin cậy lớn nhất để chấp nhận ý kiến nói trên là $97,36\%$.

Ví dụ 7.5. Trọng lượng của một bịch bột giặt do máy tự động đóng gói theo quy định là 6 kg. Sau một thời gian sản xuất, người ta tiến hành kiểm tra n bịch được trọng lượng trung bình $\overline{x} = 5,975$ kg và $s^2 = 5,7596$ kg^2. Biết trọng lượng của bịch bột giặt có phân phối chuẩn, với mức ý nghĩa là 5%, hãy cho kết luận về sự hoạt động của máy trong hai trường hợp:

a) Cỡ mẫu $n = 121$.　　　　b) Cỡ mẫu $n = 25$.

Giải:

a) Xét giả thuyết $H_0 : \mu = 6$ và đối thuyết $H_1 : \mu \neq 6$.

Chọn thống kê $Z = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n} \simeq N(0,1)$ làm tiêu chuẩn kiểm định.

Với mẫu đã cho có $\mu_0 = 6; n = 121; \overline{x} = 5,975; s = 2,3999$ nên giá trị kiểm định là:

$$z_0 = \frac{\overline{x} - \mu_0}{s}\sqrt{n} = \frac{5,975 - 6}{2,3999}\sqrt{121} = -0,1146.$$

Mức ý nghĩa $\alpha = 0,05$ nên $\alpha/2 = 0,025$ tra Phụ lục 3 được $z_{\alpha/2} = z_{0,025} = 1,96$.

Do $z_{\alpha/2} > |z_0|$ nên chấp nhận H_0, bác bỏ H_1. Vì vậy chưa đủ cơ sở cho rằng máy hoạt động không bình thường.

b) Xét giả thiết $H_0 : \mu = 6$ và đối thiết $H_1 : \mu \neq 6$.

Chọn thống kê $T = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n} \sim T(n-1)$ làm tiêu chuẩn kiểm định.

Với mẫu đã cho có $\mu_0 = 6; n = 25; \overline{x} = 5,975; s = 2,3999$ nên giá trị kiểm định:

$$t_0 = \frac{\overline{x} - \mu_0}{s}\sqrt{n} = \frac{5,975 - 6}{\sqrt{5,7596}}\sqrt{25} = -0,052.$$

Mức ý nghĩa $\alpha = 0,05$ nên $\dfrac{\alpha}{2} = 0,025$ tra Phụ lục 5 có $t_{(n-1,\alpha/2)} = t_{(24;0,025)} = 2,064$.

Do đó $t_{(n-1,\alpha/2)} > |t_0|$ nên chấp nhận H_0, bác bỏ H_1. Vì vậy chưa đủ cơ sở cho rằng máy hoạt động không bình thường.

Phương pháp p-value: Bên cạnh đó, người ta còn sử dụng phương pháp p-value được trình bày dưới đây khi kiểm định giả thuyết về tham số.

▷ *Bài toán kiểm định giả thuyết $H_0 : \mu = \mu_0$, đối thuyết $H_1 : \mu \neq \mu_0$ với mức ý nghĩa α cho trước.*

Như đã biết giả thuyết H_0 sẽ bị bác bỏ nếu giá trị kiểm định z_0 thỏa điều kiện: $|z_0| > z_{\alpha/2}$. Và p-value của mẫu là giá trị được xác định như sau:

$$p\text{-value} = P(|Z| > |z_0|) = P(Z > |z_0|) + P(Z < -|z_0|) = 2P(Z > |z_0|).$$

▷ *Bài toán kiểm định giả thuyết $H_0 : \mu = \mu_0$, đối thuyết $H_2 : \mu > \mu_0$ với mức ý nghĩa α cho trước.*

Ta có p-value của mẫu là giá trị được xác định như sau:

$$p\text{-value} = P(Z > z_0).$$

▷ *Bài toán kiểm định giả thuyết $H_0 : \mu = \mu_0$, đối thuyết $H_3 : \mu < \mu_0$ với mức ý nghĩa α cho trước.*

Ta có p-value của mẫu là giá trị được xác định như sau:

$$p\text{-value} = P(Z < z_0).$$

Ở cả ba bài toán kiểm định nêu trên, giả thuyết H_0 sẽ bị bác bỏ nếu p-value $< \alpha$.

Chú ý: Nếu mức ý nghĩa α không cho trước thì ta có thể xét các mức cụ thể sau đây:

◦ p-value $> 0,05$: Không đủ cơ sở bác bỏ H_0.

◦ $0,01 <$ p-value $< 0,05$: Có đủ cơ sở bác bỏ H_0.

◦ p-value $< 0,01$: Có cơ sở rất mạnh để bác bỏ H_0.

Ví dụ 7.6. Giải *Ví dụ 7.4 a)* bằng phương pháp p-value ta làm như sau:

Xét giả thuyết $H_0 : \mu = 100$ và đối thuyết $H_2 : \mu > 100$.

Ta có: $z_0 = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n} = \dfrac{100,2 - 100}{0,8}\sqrt{60} = 1,936.$

Khi đó: p-value $= P(Z > z_0) = P(Z > 1,936) = 1 - \Phi(1,936) = 0,02643 > 0,01 = \alpha$.

Nên ta chấp nhận H_0, bác bỏ H_2.

Ví dụ 7.7. Giải *Ví dụ 7.5 a)* bằng phương pháp p-value ta làm như sau:

Xét giả thuyết $H_0 : \mu = 6$ và đối thuyết $H_1 : \mu \neq 6$.

Ta có: $z_0 = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n} = \dfrac{5,975 - 6}{2,3999}\sqrt{121} = -0,1146.$

Khi đó: p-value $= 2P(Z > |z_0|) = 2P(Z > 0,1146) = 2\big(1 - \Phi(0,1146)\big) = 0,9088 > 0,05 = \alpha$.

Vậy nên ta chấp nhận H_0, bác bỏ H_1.

7.2.2. Kiểm định giả thuyết về tỉ lệ

Bài toán: Giả sử tổng thể X có tỉ lệ các phần tử mang dấu hiệu A là p chưa biết. Với mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$ có kích thước n và mức ý nghĩa α cho trước, hãy kiểm định giả thuyết $H_0 : p = p_0$ với một trong các đối thuyết $H_1 : p \neq p_0; H_2 : p > p_0; H_3 : p < p_0$.

Chọn thống kê $Z = \dfrac{F - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \simeq N(0,1)$ để kiểm định tỉ lệ p khi $np_0 \geq 5$ và $n(1-p_0) \geq 5$.

Với mẫu cụ thể ta được giá trị kiểm định: $z_0 = \dfrac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$ (f là tần suất mẫu).

Tùy theo đối thuyết mà ta xem xét bác bỏ giả thuyết H_0 như sau:

- Nếu đối thuyết là $H_1 : p \neq p_0$ thì ta bác bỏ H_0 khi $|z_0| > z_{\alpha/2}$.
- Nếu đối thuyết là $H_2 : p > p_0$ thì ta bác bỏ H_0 khi $z_0 > z_\alpha$.
- Nếu đối thuyết là $H_3 : p < p_0$ thì ta bác bỏ H_0 khi $z_0 < -z_\alpha$.

Ví dụ 7.8. Một hãng sản xuất tuyên bố rằng 40% dân chúng ưa thích các sản phẩm do hãng sản xuất. Khi tiến hành điều tra 500 người thì có 150 người ưa thích sản phẩm do hãng sản xuất. Với mức ý nghĩa 5%, hãy kiểm định xem tuyên bố trên có cao hơn so với thực tế không?

Giải: Đặt giả thuyết $H_0 : p = 0,4$ và đối thuyết $H_3 : p < 0,4$.

Chọn thống kê $Z = \dfrac{F - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \simeq N(0,1)$ để kiểm định tỉ lệ p.

Với mẫu đã cho có $n = 500; m = 150$ nên $f = \dfrac{150}{500} = 0,3$ và giá trị kiểm định:

$$z_0 = \frac{0,3 - 0,4}{\sqrt{0,4.0,6}}\sqrt{500} = -4,564.$$

Mức ý nghĩa $\alpha = 0,05$ tra Phụ lục 3 được $z_\alpha = z_{0,05} = 1,645$.

Do $z_0 < -z_\alpha$ nên bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_3. Vì vậy có thể cho rằng tỉ lệ trong lời tuyên bố ở trên là cao hơn so với thực tế.

Ví dụ 7.9. Trước đợt bầu cử Tổng thống của một quốc gia, một Đảng chính trị tuyên bố rằng 51% cử tri sẽ bỏ phiếu cho ứng cử viên của họ. Chọn ngẫu nhiên 250 người hỏi ý kiến thì có 90 người sẽ bầu cho ứng cử viên đó. Với mức ý nghĩa 5%, hãy nhận định xem tuyên bố trên có đúng hay không?

Giải: Đặt giả thuyết $H_0 : p = 0,51$ và đối thuyết $H_1 : p \neq 0,51$.

Chọn thống kê $Z = \dfrac{F - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \simeq N(0,1)$ để kiểm định tỉ lệ p.

Với mẫu đã cho có $n = 250; m = 90$ nên $f = \dfrac{90}{250} = 0,36$ và giá trị kiểm định:

$$z_0 = \frac{0,36 - 0,51}{\sqrt{0,51.0,49}}\sqrt{250} = -4,744.$$

Mức ý nghĩa $\alpha = 0,05$ nên $\alpha/2 = 0,025$ tra Phụ lục 3 được $z_{\alpha/2} = 1,96$.

Do $|z_0| = 4,744 > z_{\alpha/2}$ nên bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1. Vậy với mức ý nghĩa 5% có thể cho rằng lời tuyên bố ở trên là không đúng.

7.2.3. Kiểm định giả thuyết về phương sai

Bài toán: Giả sử tổng thể $X \sim N(\mu, \sigma^2)$, phương sai $\mathrm{Var}(X) = \sigma^2$ chưa biết. Với mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$ có kích thước n và mức ý nghĩa α cho trước, hãy kiểm định giả thuyết $H_0 : \sigma^2 = \sigma_0^2$ với một trong các đối thuyết $H_1 : \sigma \neq \sigma_0; H_2 : \sigma > \sigma_0; H_3 : \sigma < \sigma_0$.

Trường hợp 1: Trung bình $\mathrm{E}(X) = \mu$ chưa biết.

Chọn thống kê $K = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$ làm tiêu chuẩn kiểm định.

Từ mẫu cụ thể, ta tính được giá trị kiểm định: $k_0 = \dfrac{(n-1)s^2}{\sigma_0^2}$.

Tùy theo đối thuyết mà ta xem xét bác bỏ giả thuyết H_0 như sau:

∘ Nếu đối thuyết là $H_1 : \sigma^2 \neq \sigma_0^2$ thì ta bác bỏ H_0 khi $\left[\begin{array}{l} k_0 > \chi^2_{(n-1,\alpha/2)}. \\ k_0 < \chi^2_{(n-1,1-\alpha/2)}. \end{array}\right.$

∘ Nếu đối thuyết là $H_2 : \sigma^2 > \sigma_0^2$ thì ta bác bỏ H_0 khi $k_0 > \chi^2_{(n-1,\alpha)}$.

∘ Nếu đối thuyết là $H_3 : \sigma^2 < \sigma_0^2$ thì ta bác bỏ H_0 khi $k_0 < \chi^2_{(n-1,1-\alpha)}$.

Trường hợp 2: Trung bình $\mathrm{E}(X) = \mu$ đã biết.

Chọn thống kê $K = \dfrac{\sum(X_i - \mu)^2}{\sigma_0^2} \sim \chi^2(n)$ làm tiêu chuẩn kiểm định.

Từ mẫu cụ thể, ta tính được giá trị kiểm định: $k_0 = \dfrac{\sum(x_i - \mu)^2}{\sigma_0^2}$.

Tùy theo đối thuyết mà ta xem xét bác bỏ giả thuyết H_0 như sau:

∘ Nếu đối thuyết là $H_1 : \sigma^2 \neq \sigma_0^2$ thì ta bác bỏ H_0 khi $\left[\begin{array}{l} k_0 > \chi^2_{(n,\alpha/2)}. \\ k_0 < \chi^2_{(n,1-\alpha/2)}. \end{array}\right.$

∘ Nếu đối thuyết là $H_2 : \sigma^2 > \sigma_0^2$ thì ta bác bỏ H_0 khi $k_0 > \chi^2_{(n,\alpha)}$.

∘ Nếu đối thuyết là $H_3 : \sigma^2 < \sigma_0^2$ thì ta bác bỏ H_0 khi $k_0 < \chi^2_{(n,1-\alpha)}$.

Ví dụ 7.10. Nếu một máy hoạt động bình thường thì kích thước của một loại sản phẩm là đại lượng ngẫu nhiên có phân phối chuẩn với phương sai $\sigma^2 = 25\ cm^2$. Nghi ngờ máy hoạt động không bình thường, người ta đo thử 20 sản phẩm và tính được $s^2 = 27,5\ cm^2$. Với mức ý nghĩa 2% hãy kết luận về điều nghi ngờ trên.

Giải: Xét giả thuyết $H_0 : \sigma^2 = 25$ và đối thuyết $H_1 : \sigma^2 \neq 25$.

Chọn thống kê $K = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$ làm tiêu chuẩn kiểm định.

Với mẫu đã cho có $n = 20; s^2 = 27,5$ nên giá trị kiểm định:

$$k_0 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(20-1)\cdot 27,5}{25} = 20,9.$$

Mức ý nghĩa $\alpha = 0,02$ nên tra bảng Phụ lục 4, ta được $\chi^2_{(n-1,1-\alpha/2)} = \chi^2_{(19;0,99)} = 7,633$ và $\chi^2_{(n-1,\alpha/2)} = \chi^2_{(19;0,01)} = 36,191$.

Do $\chi^2_{(n-1,1-\alpha/2)} < k_0 < \chi^2_{(n-1,\alpha/2)}$ nên ta chấp nhận giả thuyết H_0, bác bỏ đối thuyết H_1. Vậy với mức ý nghĩa 2%, nghi ngờ máy hoạt động không bình thường là không có cơ sở.

7.3. KIỂM ĐỊNH GIẢ THUYẾT HAI MẪU

7.3.1. So sánh hai giá trị trung bình

Bài toán: Giả sử tổng thể X có $\mathrm{E}(X) = \mu_1$ và tổng thể Y có $\mathrm{E}(Y) = \mu_2$ chưa biết. Từ hai mẫu cụ thể độc lập $w_X = (x_1, x_2, \ldots, x_m)$ kích thước m của X và $w_Y = (y_1, y_2, \ldots, y_n)$ kích thước n của Y, với mức ý nghĩa α cho trước, hãy kiểm định giả thuyết $H_0 : \mu_1 = \mu_2$ với một trong ba đối thuyết $H_1 : \mu_1 \neq \mu_2$, $H_2 : \mu_1 > \mu_2$ và $H_3 : \mu_1 < \mu_2$.

Giả sử $\sigma_1^2 = \mathrm{Var}(X), \sigma_2^2 = \mathrm{Var}(Y)$ khi đó ta xét các trường hợp sau:

Trường hợp 1: Đã biết phương sai (σ_1^2 và σ_2^2) của hai tổng thể, hai mẫu độc lập, X và Y có phân phối chuẩn.
Hoặc đã biết phương sai của hai tổng thể, hai mẫu là độc lập, X và Y không có phân phối chuẩn, cỡ mẫu lớn.

Chọn thống kê $Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \simeq N(0,1)$ làm tiêu chuẩn kiểm định. Trong đó, $\overline{X}$ và $\overline{Y}$ lần lượt là các thống kê nhận giá trị bằng trung bình mẫu cụ thể $\overline{x}$ và $\overline{y}$.

Từ mẫu cụ thể w_X và w_Y ta tính được giá trị kiểm định: $z_0 = \dfrac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$.

Tùy theo đối thuyết mà ta xem xét bác bỏ giả thuyết H_0 như sau:

- Nếu đối thuyết là $H_1 : \mu \neq \mu_0$ thì ta bác bỏ H_0 khi $|z_0| > z_{\alpha/2}$.
- Nếu đối thuyết là $H_2 : \mu > \mu_0$ thì ta bác bỏ H_0 khi $z_0 > z_\alpha$.
- Nếu đối thuyết là $H_3 : \mu < \mu_0$ thì ta bác bỏ H_0 khi $z_0 < -z_\alpha$.

Trường hợp 2: Chưa biết phương sai (σ_1^2 và σ_2^2) của hai tổng thể, hai mẫu độc lập, X và Y có phân phối chuẩn, cỡ mẫu lớn.
Hoặc chưa biết phương sai của hai tổng thể, hai mẫu là độc lập, X và Y không có phân phối chuẩn, cỡ mẫu lớn.

Chọn thống kê $Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \simeq N(0,1)$ làm tiêu chuẩn kiểm định. Trong đó, S_1^2 và S_2^2 lần lượt là các thống kê nhận giá trị bằng phương sai điều chỉnh mẫu cụ thể s_1^2 và s_2^2.

Từ mẫu cụ thể w_X và w_Y ta tính được giá trị kiểm định: $z_0 = \dfrac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$.

Việc xem xét bác bỏ giả thuyết H_0 hoàn toàn tương tự *Trường hợp 1*.

Trường hợp 3: Chưa biết phương sai của hai tổng thể nhưng biết $\sigma_1^2 = \sigma_2^2$, hai mẫu độc lập, X và Y có phân phối chuẩn, cỡ mẫu nhỏ.

Chọn thống kê $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{S^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim T(m+n-2)$ làm tiêu chuẩn kiểm định. Trong đó $S^2 = \dfrac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$ với S_1^2, S_2^2 lần lượt là các thống kê nhận giá trị bằng phương sai điều chỉnh mẫu cụ thể s_1^2, s_2^2.

Từ mẫu cụ thể w_X, w_Y ta tính được $s^2 = \dfrac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$ nên giá trị kiểm định:

$$t_0 = \frac{\overline{x} - \overline{y}}{\sqrt{s^2\left(\frac{1}{m} + \frac{1}{n}\right)}}.$$

Tùy theo đối thuyết mà ta xem xét bác bỏ giả thuyết H_0 như sau:

- Nếu đối thuyết là $H_1 : \mu \neq \mu_0$ thì ta bác bỏ H_0 khi $|t_0| > t_{(m+n-2;\alpha/2)}$.
- Nếu đối thuyết là $H_2 : \mu > \mu_0$ thì ta bác bỏ H_0 khi $t_0 > t_{(m+n-2;\alpha)}$.
- Nếu đối thuyết là $H_3 : \mu < \mu_0$ thì ta bác bỏ H_0 khi $t_0 < -t_{(m+n-2;\alpha)}$.

Ví dụ 7.11. Giám đốc một xí nghiệp muốn xác định xem có sự khác nhau về năng suất giữa ca ngày và ca tối không. Một mẫu gồm 100 công nhân ca ngày sản xuất được $\overline{x} = 74,3$ (sản phẩm) với độ lệch điều chỉnh $s_1 = 16$ (sản phẩm); một mẫu khác gồm 80 công nhân ca tối sản xuất được $\overline{y} = 69,7$ (sản phẩm) với $s_2 = 18$ (sản phẩm). Với mức ý nghĩa 0,1 hãy xem có sự khác nhau về năng xuất giữa hai ca không?

Giải: Gọi μ_1, μ_2 là trung bình số sản phẩm của mỗi công nhân làm được trong ca sáng, ca tối. Xét giả thuyết $H_0 : \mu_1 = \mu_2$, đối thuyết $H_1 : \mu_1 \neq \mu_2$.

Chọn thống kê $Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \simeq N(0,1)$ làm tiêu chuẩn kiểm định.

Với hai mẫu đã cho, ta tính được giá trị kiểm định:

$$z_0 = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{74,39 - 69,7}{\sqrt{\frac{16^2}{100} + \frac{18^2}{80}}} = 1,789.$$

Mức ý nghĩa $\alpha = 0,1$ nên $\alpha/2 = 0,05$ tra Phụ lục 3 có $z_{\alpha/2} = z_{0,05} = 1,645$.

Do $|z_0| > z_{\alpha/2}$ nên bác bỏ H_0, chấp nhận H_1. Nghĩa là với mức ý nghĩa $0,1$ chưa có bằng chứng về sự khác nhau giữa năng suất của hai ca.

Ví dụ 7.12. Giám đốc của một công ty thực phẩm muốn xác định liệu loại bao bì mới có làm tăng sản lượng bán hàng hay không. Chọn ngẫu nhiên 10 quầy bán hàng bao bì mới và 8 quầy bán hàng bao bì cũ. Trung bình mỗi quầy bán hàng bao bì mới bán được $\overline{x} = 130$ (sản phẩm) với $s_1 = 10$ (sản phẩm); mỗi quầy bán hàng bao bì cũ bán được $\overline{y} = 117$ (sản phẩm) với $s_2 = 12$ (sản phẩm). Với mức ý nghĩa $\alpha = 5\%$, hãy xem kiểu bao bì mới có làm tăng sản lượng bán hàng không. Giả sử lượng hàng bán được có phân phối chuẩn, cùng phương sai.

Giải: Gọi lượng hàng trung bình bán được tại các quầy tương ứng với loại bao bì mới và cũ là μ_1 và μ_2. Đặt giả thuyết $H_0 : \mu_1 = \mu_2$ đối thuyết $H_2 : \mu_1 > \mu_2$.

Chọn thống kê $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{S^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim T(m+n-2)$ làm tiêu chuẩn kiểm định.

Với hai mẫu đã cho có:

$$s^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} = \frac{9 \cdot 10^2 + 7 \cdot 12^2}{10+8-2} = 119,25$$

nên giá trị kiểm định:

$$t_0 = \frac{\overline{x} - \overline{y}}{\sqrt{s^2\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{130 - 117}{\sqrt{119,25\left(\frac{1}{10} + \frac{1}{8}\right)}} = 2,5097.$$

Mức ý nghĩa $\alpha = 0,05$ tra Phụ lục 5 có $t_{(m+n-2,\alpha)} = t_{(16;0,05)} = 1,745$.

Do $t_0 > t_{(m+n-2,\alpha)}$ nên bác bỏ H_0, chấp nhận H_2. Vậy với mức ý nghĩa 5% ta có thể cho rằng loại bao bì mới làm tăng sản lượng bán hàng.

7.3.2. So sánh hai tỉ lệ

Bài toán: Xét hai tổng thể X và Y có p_1, p_2 lần lượt là tỉ lệ các phần tử mang dấu hiệu A trong X và Y. Từ hai mẫu cụ thể độc lập $w_X = (x_1, x_2, \ldots, x_m)$ kích thước m của X và $w_Y = (y_1, y_2, \ldots, y_n)$ kích thước n của Y. Với mức ý nghĩa α, hãy kiểm định giả thuyết $H_0 : p_1 = p_2$ với một trong ba đối thuyết $H_1 : p_1 \neq p_2$, $H_2 : p_1 > p_2$ và $H_3 : p_1 < p_2$.

Chọn thống kê $Z = \dfrac{F_1 - F_2}{\sqrt{F(1-F)\left(\frac{1}{m} + \frac{1}{n}\right)}} \simeq N(0,1)$ làm tiêu chuẩn kiểm định. Trong đó, F_1 và F_2 là các thống kê nhận giá trị bằng tỉ lệ mẫu cụ thể f_1 của X và f_2 của Y, $F = \dfrac{mF_1 + nF_2}{m+n}$ là thống kê tỉ lệ chung của hai mẫu.

Từ các mẫu cụ thể w_X của X và w_Y của Y thỏa mãn điều kiện $mf_1 + nf_2 \geq 10$ và $m(1-f_1) + n(1-f_2) \geq 10$, ta có giá trị kiểm định:

$$z_0 = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{m} + \frac{1}{n}\right)}} \text{ với } f = \frac{mf_1 + nf_2}{m+n}.$$

Tùy theo dạng của đối thuyết mà ta có quy tắc bác bỏ giả thuyết H_0 như sau:

∘ Nếu đối thuyết là $H_1 : p_1 \neq p_2$ thì ta bác bỏ H_0 khi $|z_0| > z_{\alpha/2}$.

∘ Nếu đối thuyết là $H_2 : p_1 > p_2$ thì ta bác bỏ H_0 khi $z_0 > z_\alpha$.

∘ Nếu đối thuyết là $H_3 : p_1 < p_2$ thì ta bác bỏ H_0 khi $z_0 < -z_\alpha$.

Ví dụ 7.13. Khi kiểm tra 300 người ở thành phố A thì phát hiện 60 người bị bệnh đường hô hấp. Kiểm tra 200 người ở thành phố B thì phát hiện 45 người bị bệnh đường hô hấp. Với mức nghĩa 0,05 hỏi có thể khẳng định rằng có sự khác nhau về tỉ lệ người bị bệnh đường hô hấp ở hai thành phố hay không?

Giải: Gọi p_1, p_2 lần lượt là tỉ lệ người bị bệnh đường hô hấp ở hai thành phố. Xét giả thuyết $H_0 : p_1 = p_2$ và đối thuyết $H_1 : p_1 \neq p_2$.

Chọn thống kê $Z = \dfrac{F_1 - F_2}{\sqrt{F(1-F)\left(\frac{1}{m} + \frac{1}{n}\right)}} \simeq N(0,1)$ làm tiêu chuẩn kiểm định.

Với hai mẫu đã cho có $f_1 = \dfrac{60}{300} = 0,2; f_2 = \dfrac{45}{200} = 0,225; f = \dfrac{60+45}{500} = 0,21$ thỏa điều kiện $mf_1 + nf_2 = 105 \geq 10$, $m(1-f_1) + n(1-f_2) = 395 \geq 10$ nên giá trị kiểm định:

$$z_0 = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{0,2 - 0,225}{\sqrt{0,21(1-0,21)\left(\frac{1}{300} + \frac{1}{200}\right)}} = -0,672.$$

Mức ý nghĩa $\alpha = 0,05$ nên $\alpha/2 = 0,025$ tra Phụ lục 3 có $z_{\alpha/2} = z_{0,025} = 1,96$.

Do $|z_0| < z_{\alpha/2}$ nên chấp nhận giả thuyết H_0, bác bỏ đối thuyết H_1. Nghĩa là, chưa đủ cơ sở cho rằng tỉ lệ người bị bệnh về đường hô hấp ở hai thành phố là khác nhau.

7.3.3. So sánh hai phương sai

Bài toán: Giả sử $X \sim N(\mu_1, \sigma_1^2)$ và $Y \sim N(\mu_2, \sigma_2^2)$. Từ hai mẫu cụ thể độc lập $w_X = (x_1, x_2, \ldots, x_m)$ kích thước m của X và $w_Y = (y_1, y_2, \ldots, y_n)$ kích thước n của Y. Với mức ý nghĩa α, hãy kiểm định giả thuyết $H_0 : \sigma_1^2 \neq \sigma_2^2$ với một trong các đối thuyết $H_1 : \sigma_1^2 = \sigma_2^2$, $H_2 : \sigma_1^2 > \sigma_2^2$ và $H_3 : \sigma_1^2 < \sigma_2^2$.

Chọn thống kê $F = \dfrac{S_1^2}{S_2^2} \sim F(m-1, n-1)$ làm tiêu chuẩn kiểm định. Trong đó, S_1^2 và S_2^2 là các thống kê nhận giá trị bằng phương sai điều chỉnh mẫu cụ thể s_1^2 và s_2^2.

Từ hai mẫu cụ thể, ta có giá trị kiểm định: $f_0 = \dfrac{s_1^2}{s_2^2}$.

Sau đó tùy theo dạng của đối thuyết mà ta có quy tắc bác bỏ giả thuyết H_0 như sau:

∘ Nếu đối thuyết là $H_1 : \sigma_1^2 \neq \sigma_2^2$ thì ta bác bỏ H_0 khi $\left[\begin{array}{l} f_0 > f_{(m-1,n-1,\alpha/2)}. \\ f_0 < f_{(m-1,n-1,1-\alpha/2)}. \end{array}\right.$

∘ Nếu đối thuyết là $H_2 : \sigma^2 > \sigma_0^2$ thì ta bác bỏ H_0 khi $f_0 > f_{(m-1,n-1,\alpha)}$.

∘ Nếu đối thuyết là $H_3 : \sigma^2 < \sigma_0^2$ thì ta bác bỏ H_0 khi $f_0 < f_{(m-1,n-1,1-\alpha)}$.

Ví dụ 7.14. Người ta dùng phương sai làm độ đo đánh giá sự rủi ro của cổ phiếu. Điều tra ngẫu nhiên giá cổ phiếu của công ty A trong 25 ngày tính được $s_1^2 = 6,52$; của công ty B trong 22 ngày tính được $s_2^2 = 3,47$. Với mức ý nghĩa 5% có thể cho rằng độ rủi ro cổ phiếu của công ty A cao hơn công ty B hay không?

Giải: Gọi σ_1^2, σ_2^2 lần lượt là phương sai giá cổ phiếu của A, B. Xét giả thuyết $H_0 : \sigma_1^2 = \sigma_2^2$, đối thuyết $H_2 : \sigma_1^2 > \sigma_2^2$.

Chọn thống kê $F = \dfrac{S_1^2}{S_2^2} \sim F(m-1, n-1)$ làm tiêu chuẩn kiểm định.

Với các mẫu đã cho có $s_1^2 = 6,52; s_2^2 = 3,47$ nên giá trị kiểm định:

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{6,52}{3,47} = 1,879.$$

Mức ý nghĩa $\alpha = 0,05$ tra Phụ lục 6 có $f_{(m-1,n-1,\alpha)} = f_{(24;21;0,05)} = 2,054$.

Do $f_0 < f_{(m-1,n-1,\alpha)}$ nên chấp nhận H_0, bác bỏ H_2. Vậy với mức ý nghĩa 5% chưa đủ cơ sở cho rằng cổ phiếu của công ty A nhiều rủi ro hơn công ty B.

7.4. KIỂM ĐỊNH PHI THAM SỐ

7.4.1. Kiểm định giả thuyết về luật phân phối

Bài toán: Giả sử tổng thể X có luật phân phối xác suất $F_X(x)$ chưa biết. Với mức ý nghĩa α, từ mẫu cụ thể $w_X = (x_1, x_2, \ldots, x_n)$ hãy kiểm định giả thuyết $H_0 : F_X(x) = F^*(x)$ với đối thuyết $H_1 : F_X(x) \neq F^*(x)$ ($F^*(x)$ đã biết luật phân phối).

Cũng như kiểm định tham số, ta phải dựa vào một tiêu chuẩn để kiểm định, tiêu chuẩn này gọi là tiêu chuẩn phù hợp. Ở đây ta xét tiêu chuẩn khá thông dụng do Pearson đề xuất, nó được xây dựng dựa trên cơ sở so sánh tần số mẫu cụ thể và tần số lý thuyết của phân phối xác suất cần kiểm định. Tiêu chuẩn Pearson như sau:

Chọn thống kê $K = \sum_{i=1}^{k} \dfrac{(n_i - np_i)^2}{np_i} = n \sum_{i=1}^{k} \dfrac{\left(\frac{n_i}{n} - p_i\right)^2}{p_i}$, khi $n \to +\infty$ thì K xấp xỉ phân phối Chi bình phương $\chi^2_{(k-r-1)}$, với r là số tham số của luật phân phối cần kiểm định (Chẳng hạn, phân phối chuẩn $N(\mu, \sigma^2)$ thì $r = 2$, phân phối Poisson $P(\lambda)$ thì $r = 1$ và các tham số này được ước lượng theo phương pháp hợp lí tối đa).

Với mẫu cụ thể $w = (x_1, x_2, \ldots, x_n)$ có kích thước $n \geq 50$, ta lập bảng phân bố tần số:

x_i	x_1	x_2	$\ldots$	x_k
n_i	n_1	n_2	$\ldots$	n_k

Trong đó, $n_1 + n_2 + \cdots + n_k = n$ thỏa mãn điều kiện $n_i \geq 5$. Nếu tồn tại $n_i < 5$ thì phải ghép các giá trị hay các khoảng giá trị của mẫu lại với nhau để tăng n_i lên.

Khi H_0 đúng, ta tính $p_i = \begin{cases} P(X = x_i) & \text{nếu } X \text{ rời rạc,} \\ P(x_i < X < x_{i+1}) & \text{nếu } X \text{ liên tục.} \end{cases}$

Giá trị kiểm định: $k_0 = \sum_{i=1}^{k} \dfrac{(n_i - np_i)^2}{np_i}$.

Nếu $k_0 > \chi^2_{(k-r-1,\alpha)}$ thì bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1.

Ví dụ 7.15. Để tìm hiểu số thiết bị hỏng trong một tháng của một hệ thống máy, người ta theo dõi 60 tháng liên tục và được số thiết bị hỏng cho ở bảng sau:

x_i	0	1	2	3	4	6	8	9
n_i	7	8	12	10	14	9	3	7

Với mức ý nghĩa 5% có thể cho rằng số thiết bị hỏng X tuân theo luật phân phối Poisson $P(\lambda)$ hay không?

Giải: Bằng phương pháp hợp lí tối đa ta ước lượng $\lambda = \overline{x} = 3,9$ (Xem *Ví dụ 6.4*).

Xét giả thuyết $H_0 : X \sim P(3,9)$, đối thuyết $H_1 : X \nsim P(3,9)$.

Chọn thống kê $K = \sum_{i=1}^{k} \dfrac{(n_i - np_i)^2}{np_i} \simeq \chi^2_{(k-r-1)}$, với $r = 1$, làm tiêu chuẩn kiểm định.

Do trong bảng phân bố tần số tồn tại $n_i < 5$ nên ta sắp xếp lại như sau:

x_i	0	1	2	3	4 và 5	6	≥ 7
n_i	7	8	12	10	14	9	10

Khi giả thuyết H_0 là đúng thì: $P(X = k) = \dfrac{e^{-3,9} \cdot 3,9^k}{k!}$. Ta có:

$p_1 = P(X = 0) = 0,0202$; $p_2 = P(X = 1) = 0,0789$;
$p_3 = P(X = 2) = 0,1539$; $p_4 = P(X = 3) = 0,2001$;
$p_5 = P(4 \leq X \leq 5) = P(X = 4) + P(X = 5) = 0,1951 + 0,1522 = 0,3473$;
$p_6 = P(X = 6) = 0,0989$; $p_7 = P(7 \leq X) = 1 - P(X < 7) = 0,1006$.

Ta có bảng:

x_i	n_i	p_i	np_i	$\dfrac{(n_i - np_i)^2}{np_i}$
0	7	$0,0202$	$1,212$	$27,641$
1	8	$0,0789$	$4,734$	$2,2532$
2	12	$0,1539$	$9,234$	$0,8285$
3	10	$0,2001$	$12,006$	$0,3352$
4 và 5	14	$0,3473$	$20,838$	$2,2439$
6	9	$0,0989$	$5,934$	$1,5842$
≥ 7	10	$0,1005$	$6,03$	$2,6137$
Σ	$n = 60$			$37,0497$

Giá trị kiểm định: $k_0 = \sum_{i=1}^{7} \dfrac{(n_i - np_i)^2}{np_i} = 37,0497$.

Mức ý nghĩa $\alpha = 0,05$ tra Phụ lục 4 có $\chi^2_{(k-r-1,\alpha)} = \chi^2_{(6-1-1;0,05)} = \chi^2_{(4;0,05)} = 9,488$.

Do $k_0 > \chi^2_{(k-r-1,\alpha)}$ nên bác bỏ H_0, chấp nhận H_1. Nghĩa là, chưa có cơ sở cho rằng X có phân phối Poisson với $\lambda = 3,9$.

Ví dụ 7.16. Khi kiểm tra năng suất lúa trồng thử nghiệm trên 100 hecta cho kết quả:

Năng suất (tấn/hecta)	8 – 9	9 – 10	10 – 11	11 – 12	12 – 13	13 – 14	14 – 15
Số hecta	8	15	21	23	16	9	8

Với mức ý nghĩa 1%, xét xem X có phân phối chuẩn $N(\mu;\sigma^2)$ hay không?

Giải: Bằng phương pháp hợp lí tối đa ta ước lượng $\mu = \overline{x} = 11,33$ và $\sigma^2 = s^2 = 2,7688$ (Xem *Ví dụ 6.5*).

Xét giả thuyết $H_0 : X \sim N(11,33; 2,7688)$, đối thuyết $H_1 : X \nsim N(11,33; 2,7688)$.

Chọn thống kê $K = \sum_{i=1}^{k} \dfrac{(n_i - np_i)^2}{np_i} \sim \chi^2_{(k-r-1)}$, với $r = 2$, làm tiêu chuẩn kiểm định.

Khi giả thuyết H_0 là đúng thì

$$P(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Do đó:

$$p_1 = P(X < 9) = \Phi\left(\frac{9-11,33}{1,664}\right) - \Phi(-\infty) = \Phi(-1,4) = 1 - \Phi(1,4) = 0,0807.$$

$$p_2 = P(9 < X < 10) = \Phi\left(\frac{10-11,33}{1,664}\right) - \Phi\left(\frac{9-11,33}{1,664}\right) = \Phi(1,4) - \Phi(0,8) = 0,1313.$$

$$p_3 = P(10 < X < 11) = \Phi\left(\frac{11-11,33}{1,664}\right) - \Phi\left(\frac{10-11,33}{1,664}\right) = \Phi(0,8) - \Phi(0,2) = 0,2093.$$

$$p_4 = P(11 < X < 12) = \Phi\left(\frac{12-11,33}{1,664}\right) - \Phi\left(\frac{11-11,33}{1,664}\right) = \Phi(0,4) - \Phi(-0,2) = 0,235.$$

$$p_5 = P(12 < X < 13) = \Phi\left(\frac{13-11,33}{1,664}\right) - \Phi\left(\frac{12-11,33}{1,664}\right) = \Phi(1) - \Phi(0,4) = 0,1858.$$

$$p_6 = P(13 < X < 1) = \Phi\left(\frac{14-11,33}{1,664}\right) - \Phi\left(\frac{13-11,33}{1,664}\right) = \Phi(1,62) - \Phi(1) = 0,1034.$$

$$p_7 = P(X > 14) = 1 - \Phi\left(\frac{14-11,33}{1,664}\right) = 1 - \Phi(1,62) = 0,0543.$$

Ta có bảng:

x_i	n_i	p_i	np_i	$\dfrac{(n_i - np_i)^2}{np_i}$
< 9	8	$0,0807$	$8,07$	$0,0000061$
9-10	15	$0,1313$	$13,13$	$0,2663329$
10-11	21	$0,2093$	$20,93$	$0,0002341$
11-12	23	$0,235$	$23,5$	$0,0106383$
12-13	16	$0,1858$	$18,58$	$0,3582562$
13-14	9	$0,1034$	$10,34$	$0,1736557$
> 14	8	$0,0543$	$5,43$	$1,216372$
Σ	$n = 100$			$2,0254953$

Giá trị kiểm định: $k_0 = \sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} = 2,0255$.

Mức ý nghĩa $\alpha = 0,01$ tra Phụ lục 4 có $\chi^2_{(k-r-1,\alpha)} = \chi^2_{(7-2-1;0,01)} = \chi^2_{(4;0,01)} = 13,277$.

Do $k_0 < \chi^2_{(k-r-1,\alpha)}$ nên chấp nhận giả thuyết H_0. Nghĩa là, có thể xem như X có luật phân phối chuẩn $N(11,33;2,7688)$.

7.4.2. Kiểm định sự độc lập

Bài toán: Giả sử cần nghiên cứu đồng thời hai tổng thể X và Y. Trong đó X có k dấu hiệu thành phần $x_1,\ldots,x_k$; Y có m dấu hiệu thành phần $y_1,\ldots,y_m$. Với mức ý nghĩa α, hãy kiểm định giả thuyết $H_0: X$ và Y độc lập với đối thuyết $H_1: X$ và Y phụ thuộc.

Từ mẫu kích thước n quan sát đồng thời hai dấu hiệu (X,Y) ta có bảng:

$X \backslash Y$	y_1	y_2	$\ldots$	y_j	$\ldots$	y_r	Σ
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\ldots$	n_{1r}	n_1
x_2	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2r}	n_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\ldots$	n_{ir}	n_i
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_k	n_{k1}	n_{k2}	$\ldots$	n_{kj}	$\ldots$	n_{kr}	n_k
Σ	m_1	m_2	$\ldots$	m_j	$\ldots$	m_r	n

Trong đó:

$n_i = \sum_{j=1}^{r} n_{ij}$: Tổng các tần số ứng với dấu hiệu thành phần x_i.

$m_j = \sum_{i=1}^{k} n_{ij}$: Tổng các tần số ứng với dấu hiệu thành phần y_j.

n_{ij}: Tần số ứng với các phần tử mang đồng thời dấu hiệu (x_i,y_j).
Với n đủ lớn, ta có:

$P(X = x_i, Y = y_j) \approx \frac{n_{ij}}{n}$: Xác suất xuất hiện đồng thời (x_i,y_j).

$P(X = x_i) \approx \frac{n_i}{n}$: Xác suất xuất hiện x_i.

$P(Y = y_j) \approx \frac{m_j}{n}$: Xác suất xuất hiện y_j.
Khi giả thuyết H_0 là đúng thì X,Y độc lập nên $P(X = x_i, Y = y_j) = P(X = x_i)\cdot P(Y = y_j)$.

Chọn thống kê $K = n \sum_{\substack{i=1,\ldots,k\\ j=1,\ldots,r}} \frac{\left(\frac{n_{ij}}{n} - \frac{n_i m_j}{n^2}\right)^2}{\frac{n_i m_j}{n^2}} = n\left(\sum_{\substack{i=1,\ldots,k\\ j=1,\ldots,r}} \frac{n_{ij}^2}{n_i m_j} - 1\right)$ làm tiêu chuẩn kiểm định, khi $n \to +\infty$ thì K xấp xỉ phân phối $\chi^2_{\left((k-1)(m-1)\right)}$.

Giá trị kiểm định: $k_0 = n\left(\sum_{\substack{i=1,\ldots,k\\ j=1,\ldots,r}} \frac{n_{ij}^2}{n_i m_j} - 1\right)$.

Nếu $k_0 > \chi^2_{\left((k-1)(m-1),\alpha\right)}$ thì bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1.

Ví dụ 7.17. Điều tra ngẫu nhiên thu nhập (triệu đồng/năm) của 400 công nhân ở một công ty, người ta thu được bảng kết quả như sau:

Giới tính \ Thu nhập	< 5	$[5;10)$	$[10;15]$	Σ
Nữ	36	50	38	124
Nam	69	105	102	276
Σ	105	155	140	400

Với mức ý nghĩa 5%, hãy kiểm định xem thu nhập của công nhân có phụ thuộc vào giới tính hay không?

Giải: Giả thuyết $H_0 : X$ và Y độc lập, đối thuyết $H_1 : X$ và Y phụ thuộc.

Chọn thống kê $K = n\sum_{\substack{i=1,\dots,k\\ j=1,\dots,r}} \dfrac{\left(\frac{n_{ij}}{n} - \frac{n_i m_j}{n^2}\right)^2}{\frac{n_i m_j}{n^2}} \simeq \chi^2_{\left((k-1)(m-1)\right)}$ làm tiêu chuẩn kiểm định. Giá trị kiểm định:

$$\begin{aligned}
k_0 &= n\left(\sum_{\substack{i=1,2\\ j=1,2,3}} \frac{n_{ij}^2}{n_i m_j} - 1\right)\\
&= 400\left(\frac{36^2}{105\cdot 124} + \frac{50^2}{155\cdot 124} + \frac{38^2}{140\cdot 124} + \frac{69^2}{105\cdot 267} + \frac{105^2}{155\cdot 267} + \frac{102^2}{140\cdot 267} - 1\right)\\
&= 10,9387.
\end{aligned}$$

Mức ý nghĩa $\alpha = 0,05$ tra Phụ lục 4 có $\chi^2_{\left((2-1)(3-1);0,05\right)} = \chi^2_{(2;0,05)} = 5,991.$

Vì thế $k_0 > \chi^2_{\left((2-1)(3-1);0,05\right)}$ nên bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1. Nghĩa là thu nhập có phụ thuộc vào giới tính.

7.5. BÀI TẬP CHƯƠNG 7

Bài 1: Những thống kê trong năm ngoái cho thấy một người Mỹ du lịch ở Châu Âu trong vòng 3 tuần sẽ chi hết trung bình 1010 USD cho việc mua sắm. Một cuộc nghiên cứu dự định tiến hành cuối năm nay để xác định xem có sự thay đổi gì trong việc tiêu tiền mua sắm khi du lịch của người Mỹ hay không. Giả sử thống kê trên 50 khách du lịch cho thấy số tiền trung bình họ tiêu là 1090 USD, với độ lệch tiêu chuẩn là 300 USD. Với mức ý nghĩa $\alpha = 0,01$ hãy nhận định xem có sự thay đổi về tiêu tiền khi du lịch ở Châu Âu của người Mỹ hay không.

Bài 2: Trong điều kiện chăn nuôi bình thường thì một con bò trung bình cung cấp 3 lít sữa mỗi ngày. Người chủ thay đổi thức ăn cho bò, sau đó tiến hành kiểm tra sản lượng sữa của 20 con bò thì phát hiện trung bình mỗi con cho 3,2 lít. Có ý kiến cho rằng loại thức ăn đó làm tăng sản lượng sữa bò. Với mức ý nghĩa $\alpha = 0,05$ hãy nhận định xem ý kiến ở trên đúng hay không, biết sản lượng sữa của bò là một đại lượng ngẫu nhiên có phân phối chuẩn với độ lệch chuẩn $\sigma = 0,3$ lít?

Bài 3: Kiểm tra ngẫu nhiên 81 sản phẩm thấy khối lượng trung bình là 210 g với độ lệch điều chỉnh là 36 g. Với mức ý nghĩa 1% nếu nói rằng trọng lượng thật sự của các sản phẩm là thấp hơn mức ghi ngoài bao bì 225 g thì đúng hay sai?

Bài 4: Một nhà sản xuất săm lốp ôtô tuyên bố rằng tuổi thọ trung bình của một chiếc lốp ôtô của nhà máy là 30000 dặm. Cơ quan giám định nghi ngờ tuyên bố này nên đã kiểm tra ngẫu nhiên 100 chiếc lốp và tính được trung bình $\overline{x} = 29000$ dặm, độ lệch điều chỉnh $s = 5000$ dặm.

a) Với mức ý nghĩa 5% cơ quan giám định chất lượng cần rút ra kết luận gì?

b) Cũng với câu hỏi trên nhưng với mức ý nghĩa 2%.

Bài 5: Một công ty phần mềm ở thành phố X cho biết lương kỹ sư trung bình của công ty là 4800 USD/năm. Một kỹ sư dự định xin vào làm ở công ty này đã thăm dò 12 kỹ sư thì thấy lương trung bình của họ là 4585 USD/năm, độ lệch điều chỉnh là 630 USD. Với mức ý nghĩa 5% người này nên rút ra kết luận gì?

Bài 6: Một công ty xe buýt công bố rằng, không quá 5 phút sẽ có một chuyến xe dừng lại trạm chờ. Chọn ngẫu nhiên 8 thời điểm và ghi lại thời gian giữa hai chuyến xe buýt ta thu được số liệu sau đây:

5,3 5,3 4,8 5,1 4,3 4,8 4,9 4,7

Với mức ý nghĩa 5%, nhận định xem công bố của công ty xe buýt có đúng hay không?

Bài 7: Một phân xưởng sản xuất có tỉ lệ phế phẩm là 5%. Sau khi lắp đặt một thiết bị mới, kiểm tra ngẫu nhiên 200 sản phẩm thì phát hiện có 8 phế phẩm. Với mức ý nghĩa $\alpha = 0,05$ hãy cho biết tỉ lệ phế phẩm do phân xưởng sản xuất đã giảm hay chưa?

Bài 8: Biết rằng 18% số gia đình ở thành phố A có máy vi tính. Trong số 80 gia đình có trẻ em đang đi học, thì có 22 gia đình có máy vi tính. Với mức ý nghĩa 2% hãy nhận định xem có phải trong số các gia đình có trẻ em đang đi học, tỉ lệ gia đình có máy vi tính cao hơn tỉ lệ chung hay không?

Bài 9: Một công ty sản xuất bánh kẹo quảng cáo rằng ít nhất 2/3 số trẻ em thích ăn bánh kẹo của công ty. Trong một cuộc thăm dò ý kiến của 100 em, có 55 em nói rằng thích ăn bánh kẹo của công ty này. Sử dụng mức ý nghĩa 5%, có nhận xét gì về lời quảng cáo của công ty?

Bài 10: Một máy sản xuất tự động, lúc đầu tỉ lệ sản phẩm loại A là 20%. Sau khi cải tiến phương pháp sản xuất, người ta lấy 40 mẫu, mỗi mẫu 10 sản phẩm để kiểm tra. Được kết quả cho ở bảng sau:

Sản phẩm loại A	1	2	3	4	5	6	7	8	9	10
Số mẫu	2	0	4	6	8	10	4	5	1	0

Với mức ý nghĩa 5% hãy kết luận về phương pháp sản xuất mới này.

Bài 11: Một báo cáo nói rằng 90% số người điều khiển ô tô tham gia giao thông là có bằng lái theo quy định. Cảnh sát giao thông kiểm tra ngẫu nhiên 150 người thì phát hiện có 30 người không có bằng lái.

a) Phát biểu giả thuyết H_0, đối thuyết H_1 và kiểm định báo cáo trên với $\alpha = 2\%$.

b) Giả thuyết H_0 bị bác bỏ ở mức ý nghĩa thấp nhất là bao nhiêu?

Bài 12: Đo đường kính 12 sản phẩm do một dây chuyền sản xuất, người kỹ sư kiểm tra chất lượng tính được $s = 4\ mm$. Biết rằng nếu độ biến động của các sản phẩm lớn hơn 3 mm thì dây chuyền sản xuất phải dừng hoạt động để điều chỉnh. Với mức ý nghĩa 5%, người kỹ sư nên kết luận gì?

Bài 13: Trọng lượng của một loại sản phẩm do máy sản xuất ra là đại lượng ngẫu nhiên có phân phối chuẩn. Nghi ngờ độ đồng đều về trọng lượng sản phẩm có xu hướng giảm sút người ta cân thử 12 sản phẩm được $s^2 = 11,41\ g^2$. Với mức ý nghĩa 5% hãy kết luận về nghi ngờ trên, biết rằng bình thường độ phân tán của trọng lượng sản phẩm là $\sigma^2 = 10\ g^2$.

Bài 14: Trọng lượng sản phẩm do hai nhà máy sản xuất là các đại lượng ngẫu nhiên có luật phân phối chuẩn và có cùng độ lệch tiêu chuẩn là $\sigma = 1\ kg$. Với mức ý nghĩa $\alpha = 0,05$ có thể xem trọng lượng trung bình của các sản phẩm do hai nhà máy sản xuất ra là như nhau được không? Nếu cân thử 25 sản phẩm của nhà máy A ta được trung bình là 50 kg; cân thử 20 sản phẩm của nhà máy B ta được trung bình là $50,6\ kg$.

Bài 15: Kiểm tra những sản phẩm được chọn ngẫu nhiên ở hai nhà máy cùng sản xuất một loại sản phẩm. Ta có các số liệu sau:

Nhà máy	Số sản phẩm	Số phế phẩm
X	1000	20
Y	900	30

Có thể cho rằng tỷ lệ phế phẩm của hai nhà máy là như nhau hay không? ($\alpha = 4\%$)

Bài 16: Một doanh nghiệp kinh doanh trò chơi điện tử cần kết nối internet, khu vực đó có hai nhà mạng viễn thông A và B cùng cung cấp gói dịch vụ có tốc độ cam kết là 10 Mbs. Để đánh giá tốc độ đường truyền, hai mẫu được đo theo tốc độ đường truyền (Mbs) của nhà mạng A và B tại các thời điểm và vị trí như nhau ta được bảng số liệu như sau:

Tốc độ	< 5	$[5;6)$	$[6;7)$	$[7;8)$	$[8;9)$	≥ 9
Tần số nhà mạng A	2	4	6	15	17	16
Tần số nhà mạng B	4	3	7	12	15	9

Có thể cho rằng tốc độ đường truyền của hai nhà mạng như nhau hay không? ($\alpha = 0,05$)

Bài 17: Giám đốc một doanh nghiệp muốn so sánh chi phí tiếp khách giữa hai bộ phận bán hàng và sản xuất. Một mẫu điều tra 15 hóa đơn chi phí mỗi nơi kết quả cho như sau:

Bán hàng	Sản xuất
$\overline{x}_1 = 7358750$ đồng	$\overline{x}_2 = 5438125$ đồng
$s_1 = 1583750$ đồng	$s_2 = 1403500$ đồng

a) Với mức ý nghĩa 1%, có sự khác nhau về chi phí tiếp khách trung bình giữa hai bộ phận trên hay không? Giả sử các mức chi phí có cùng độ lệch chuẩn.

b) Với mức ý nghĩa 10% và số liệu như trên hãy cho kết luận về phương sai của chi phí tiếp khách ở hai bộ phận. Giả sử chi phí của hai bộ phận có cùng phân phối chuẩn.

Bài 18: Theo dõi số lượng mũ bảo hiểm bán được trong tháng ở một cửa hàng, ta có kết quả:

Số mũ bán được	1	3	4	5	6	8
Số ngày	2	5	8	10	2	3

Số lượng mũ bảo hiểm bán được trong tháng có phân phối Poisson hay không? ($\alpha = 5\%$)

Bài 19: Để tìm hiểu lượng mủ cao su mỗi cây cho trong một ngày ở năm đầu khai thác, ghi nhận ở 100 cây ta có số liệu:

Lượng mủ (g)	200-210	210-220	220-230	230-240	240-250	250-260	260-270
Số cây	2	8	14	30	25	12	9

Số lượng mủ cao su một cây cho trong ngày có luật phân phối chuẩn hay không? ($\alpha = 10\%$)

Bài 20: Trong 78 người dùng cà phê có 30 người bị mất ngủ. Trong 90 người không dùng cà phê có 15 người bị mất ngủ. Với mức ý nghĩa 5% xét xem cà phê có gây mất ngủ hay không?

Bài 21: Kiểm tra chất lượng một loại sản phẩm do 3 nhà máy sản xuất hàng xuất khẩu (XK) ra ta ghi được bảng số liệu:

Nhà máy \ Chất lượng	Đạt yêu cầu XK	Tốt	Cần sửa thêm	Thứ phẩm
Nhà máy A	125	40	18	17
Nhà máy B	91	29	14	16
Nhà máy C	84	31	18	17

Hãy xét xem chất lượng sản phẩm của 3 nhà máy có như nhau không? ($\alpha = 5\%$)

★ BÀI TẬP TỔNG HỢP

Bài 1: Tại một nông trường, để điều tra trọng lượng của một loại trái cây, người ta cân thử một số trái và được kết quả cho trong bảng sau:

Trọng lượng (g)	Số trái
$(45; 50]$	2
$(50; 55]$	11
$(55; 60]$	25
$(60; 65]$	74
$(65; 70]$	187
$(70; 75]$	43
$(75; 80]$	16
$(80; 85]$	2
$(85; 90]$	1

a) Ước lượng trọng lượng trung bình của loại trái cây đó ở nông trường với độ tin cậy 99%.

b) Muốn dùng trung bình của mẫu này để ước lượng trung bình của tổng thể với độ tin cậy 99% và độ chính xác $0,7$ thì cần cân thêm ít nhất bao nhiêu trái nữa?

c) Người ta quy ước những trái có trọng lượng nhỏ hơn hoặc bằng 60 g là loại II. Hãy ước lượng tỉ lệ trái loại II với độ tin cậy 95%.

d) Biết rằng trước đợt kiểm tra trọng lượng trung bình của một trái là 70 g. Sau đó người ta bón thêm một loại phân hóa học mới thì có số liệu như trên. Hãy cho kết luận về tác dụng của loại phân này ($\alpha = 1\%$).

Bài 2: Đo đường kính của 1000 chi tiết máy, người ta được kết quả cho ở bảng sau:

Đường kính X (mm)	98	$98,5$	99	$99,5$	100	$100,5$	101	$101,5$	102	$102,5$
Số chi tiết	21	47	87	158	181	201	142	97	40	26

a) Tìm khoảng ước lượng cho đường kính trung bình của chi tiết máy với $\alpha = 98\%$.

b) Muốn dùng trung bình đường kính trong mẫu để ước lượng trung bình đường kính tổng thể thì độ tin cậy là bao nhiêu để có độ chính xác là $0,05$ mm.

c) Như câu b), nhưng muốn độ tin cậy 68% thì độ chính xác là bao nhiêu?

d) Gọi các chi tiết máy có đường kính dưới 100 mm là loại nhỏ. Hãy ước lượng tỉ lệ chi tiết máy loại nhỏ với độ tin cậy 97%.

e) Muốn ước lượng tỉ lệ chi tiết máy loại nhỏ có độ chính xác 5% thì cần đo thêm bao nhiêu chi tiết nữa?

Bài 3: Hàm lượng dầu trung bình trong một loại trái cây lúc đầu là 5 g. Người ta chăm bón bằng loại phân K và sau một thời gian, kiểm tra hàm lượng dầu ở một số trái thu được kết quả:

Hàm lượng dầu (g)	Số trái
$(1; 5]$	51
$(5; 9]$	47
$(9; 13]$	39
$(13; 17]$	36
$(17; 21]$	32
$(21; 25]$	8
$(25; 29]$	7
$(29; 33]$	3
$(33; 37]$	2

a) Hãy cho kết luận về tác dụng của loại phân K nói trên ($\alpha = 1\%$).

b) Tìm ước lượng cho hàm lượng dầu trung bình của loại trái cây đó sau khi chăm bón với độ tin cậy 97%.

c) Giả sử với số liệu điều tra ở trên, muốn ước lượng hàm lượng dầu trung bình với độ chính xác $0,8$ g thì độ tin cậy đạt được là bao nhiêu.

d) Những trái có hàm lượng dầu lớn hơn 21 g gọi là loại A. Hãy ước lượng tỉ lệ những trái loại A với độ tin cậy 95%.

Chương 8

TƯƠNG QUAN - HỒI QUY

8.1. MỞ ĐẦU

Giả sử X và Y là hai đại lượng ngẫu nhiên. Khi đó có hay không quan hệ phụ thuộc giữa X và Y. Nếu chúng độc lập thì ta xét riêng từng đại lượng nhưng nếu chúng phụ thuộc thì sự phụ thuộc và mức độ phụ thuộc ra sao. Giữa X và Y có các loại phụ thuộc sau:

◦ *Sự phụ thuộc hàm số:* Nghĩa là tồn tại hàm f sao cho $Y = f(X)$;

◦ *Sự phụ thuộc thống kê:* Khi X thay đổi thì phân phối xác suất của Y cũng thay đổi;

◦ *Sự phụ thuộc tương quan:* Khi X thay đổi thì trung bình có điều kiện $\mathrm{E}(Y|X)$ cũng thay đổi, nghĩa là $\mathrm{E}(Y|X) = \varphi(X) \neq C$_hằng số. Do đó, sự phụ thuộc tương quan cũng là sự phụ thuộc thống kê. Ở đây ta chỉ xét sự phụ thuộc tương quan mà thôi.

Phương trình $\mathrm{E}(Y|X) = \varphi(X)$ còn gọi là phương trình hồi quy của Y theo X.

+ Nếu $\varphi(X) = AX + B$ thì giữa Y và X có tương quan tuyến tính.

+ Nếu $\varphi(X)$ không tuyến tính thì giữa Y và X có tương quan phi tuyến tính.

Trong chương này ta sẽ nghiên cứu hai bài toán sau đây:

◦ *Phân tích tương quan tuyến tính:* Xem xét và đánh giá mức độ chặt chẽ của sự phụ thuộc tương quan giữa Y và X.

◦ *Phân tích hồi quy:* Tìm dạng $\varphi(X)$ và với mẫu cụ thể tìm $\varphi(X)$ để dự báo giá trị Y.

8.2. PHÂN TÍCH TƯƠNG QUAN TUYẾN TÍNH

Ở Chương 3, ta đã biết hệ số tương quan $\rho(X, Y)$ được dùng để đo sự phụ thuộc tương quan tuyến tính giữa hai biến ngẫu nhiên X và Y. Tuy nhiên, để tính được $\rho(X, Y)$ ta cần phải biết luật phân phối của $\rho(X, Y)$ mà điều này trong thực tế là rất khó khăn.

Vì thế ta sẽ dựa vào các mẫu quan sát $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ của (X, Y) để ước lượng hay kiểm định giả thuyết về $\rho(X, Y)$.

8.2.1. Hệ số tương quan mẫu

Khi tính toán ta xấp xỉ $\rho(X,Y)$ bằng hệ số tương quan mẫu r được xác định như sau:

$$r = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} = \frac{\overline{xy}-\overline{x}\cdot\overline{y}}{s_n(X)\cdot s_n(Y)}.$$

Trong đó: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ và $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$; $s_n(X)$ và $s_n(Y)$ lần lượt là độ lệch chuẩn mẫu cụ thể của X và Y.

Nhận xét 8.1. Ta thấy rằng $|r| \leq 1$.

Ví dụ 8.1. Một công ty tiến hành phân tích hiệu quả quảng cáo và thu được số liệu trong hai quý như sau:

X (triệu đồng)	5	8	10	15	22
Y (chục triệu đồng)	6	15	20	30	39

Với X là chi phí cho quảng cáo, Y là tổng doanh thu. Hãy xác định hệ số tương quan mẫu.

<u>Giải:</u>

Từ mẫu đã cho, ta có: $\overline{x} = 12$; $\overline{y} = 22$; $\overline{xy} = 331,6$ và $s_n(X) = 5,966$; $s_n(Y) = 11,5065$. Cho nên hệ số tương quan mẫu là: $r = \frac{331,6 - 12\cdot 22}{5,966\cdot 11,5065} = 0,9847$.

Chú ý: Đối với mẫu cho ở dạng bảng số liệu hai chiều (Xem mục *7.4.2*). Ta có:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i,\ \overline{y} = \frac{1}{n}\sum_{j=1}^{r} m_j y_j,\ \overline{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij} x_i y_j,$$

$$s_n^2(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i - \overline{x}^2,\ s_n^2(Y) = \frac{1}{n}\sum_{j=1}^{r} m_j y_j - \overline{y}^2.$$

Ví dụ 8.2. Tiến hành 100 quan sát độc lập để theo dõi sự phụ thuộc giữa lượng phân bón X và năng suất Y của một giống lúa ta được kết quả cho ở bảng sau:

X \ Y	10	12	14	16	18	20	Σ
10	9	4	1				14
30	1	10	9	3			23
50		2	6	14	6		28
70			1	10	18	6	35
Σ	10	16	17	27	24	6	100

Hãy tính hệ số tương quan mẫu.

Giải: Ta có:

$$\overline{x} = \frac{1}{100}\sum_{i=1}^{4} n_i x_i = \frac{1}{100}(10 \cdot 14 + 30 \cdot 23 + 50 \cdot 28 + 70 \cdot 35) = 46,8.$$

$$\overline{y} = \frac{1}{100}\sum_{j=1}^{6} m_j y_j = \frac{1}{100}(10 \cdot 10 + 12 \cdot 16 + 14 \cdot 17 + 16 \cdot 27 + 18 \cdot 24 + 20 \cdot 6) = 15,14.$$

$$\overline{xy} = \frac{1}{100}\sum_{i=1}^{4}\sum_{j=1}^{6} n_{ij} x_i y_j = 759.$$

$$s_n^2(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i - (\overline{x})^2 = 445,76 \Rightarrow s_n(X) = \sqrt{445,76} = 21,113.$$

$$s_n^2(Y) = \frac{1}{n}\sum_{j=1}^{r} m_j y_j - (\overline{y})^2 = 8,02 \Rightarrow s_n(Y) = \sqrt{8,02} = 2,832.$$

Nên hệ số tương quan mẫu: $r = \dfrac{\overline{xy} - \overline{x} \cdot \overline{y}}{s_n(X) \cdot s_n(Y)} = \dfrac{759 - 46,8 \cdot 15,14}{21,113 \cdot 2,832} = 0,844.$

8.2.2. Kiểm định giả thuyết về hệ số tương quan mẫu

Bài toán 1 (*Xét có sự tương quan giữa* X *và* Y *hay không*): Với mức ý nghĩa α đã cho, hãy kiểm định giả thuyết $H_0 : \rho = 0$ với đối thuyết $H_1 : \rho \neq 0$.

Khi (X, Y) có phân phối chuẩn hai chiều và giả thuyết H_0 là đúng.

Chọn thống kê $T = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim T(n-2)$ làm tiêu chuẩn kiểm định.

Với mẫu cụ thể có giá trị kiểm định: $t_0 = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

Nếu $|t_0| > t_{(n-2,\alpha/2)}$ thì ta bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1.

Nhận xét 8.2. Giá trị $\rho = 0$ cho ta biết *không* có mối liên hệ tuyến tính của X và Y chứ chưa có nghĩa là giữa X và Y không có mối liên hệ, vì có thể chúng có mối liên hệ phi tuyến.

Ví dụ 8.3. Một nhà xã hội học tuyên bố rằng kết quả học tập X của sinh viên (đo bằng điểm thi tốt nghiệp) không có liên hệ tuyết tính tới thu nhập Y của gia đình họ. Lấy ngẫu nhiên 20 sinh viên tính được $r = 0,4$. Với mức ý nghĩa 1%, hãy cho kết luận về tuyên bố trên.

Giải: Xét giả thuyết $H_0 : \rho = 0$ và đối thuyết $H_1 : \rho \neq 0$.

Chọn thống kê $T = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim T(n-2)$ làm tiêu chuẩn kiểm định.

Với mẫu cụ thể có giá trị kiểm định: $t_0 = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{0,4\sqrt{20-2}}{\sqrt{1-0,4^2}} = 1,852.$

Mức ý nghĩa $\alpha = 0,01$ nên $\alpha/2 = 0,005$ tra Phụ lục 5 có $t_{(n-2;\alpha/2)} = t_{(18;0,005)} = 2,878.$

Do $|t_0| < t_{(n-2;\alpha/2)}$ nên chấp nhận giả thuyết H_0. Vậy với mức ý nghĩa 1%, có thể cho rằng tuyên bố ở trên là đúng.

Bài toán 2 (*Xét mức độ tương quan tuyến tính của* X *và* Y): Với mức ý nghĩa α cho trước, hãy kiểm định giả thuyết $H_0 : \rho = \rho_0$ với đối thuyết $H_1 : \rho \neq \rho_0$.

Ta xây dựng thống kê $Z = \dfrac{1}{2}\ln\dfrac{1+r}{1-r}$, Fisher đã chứng minh được khi $n \to +\infty$ thì $Z \simeq N\left(\dfrac{1}{2}\ln\dfrac{1+\rho}{1-\rho}; \dfrac{1}{n-3}\right)$. Nên khi giả thuyết H_0 là đúng, ta chọn thống kê $Z = \dfrac{1}{2}\left(\ln\dfrac{1+r}{1-r} - \ln\dfrac{1+\rho}{1-\rho}\right)\cdot\sqrt{n-3} \simeq N(0,1)$ làm tiêu chuẩn kiểm định.

Với mẫu cụ thể ta có giá trị kiểm định: $z_0 = \dfrac{1}{2}\left(\ln\dfrac{1+r}{1-r} - \ln\dfrac{1+\rho_0}{1-\rho_0}\right)\cdot\sqrt{n-3}$.

Nếu $|z_0| > z_{\alpha/2}$ thì ta bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1.

Ví dụ 8.4. Cho mẫu kích thước $n = 35$ rút ra từ (X, Y) và tính được $r = 0,8$. Với mức ý nghĩa $\alpha = 5\%$, hãy kiểm định giả thuyết $\rho = 0,9$.

Giải: Xét giả thuyết $H_0 : \rho = 0,9$ với đối thuyết $H_1 : \rho \neq 0,9$.

Chọn thống kê $Z = \dfrac{1}{2}\left(\ln\dfrac{1+r}{1-r} - \ln\dfrac{1+\rho}{1-\rho}\right)\cdot\sqrt{n-3} \simeq N(0,1)$ làm tiêu chuẩn kiểm định. Với mẫu cụ thể, ta có giá trị kiểm định:

$$z_0 = \frac{1}{2}\left(\ln\frac{1+r}{1-r} - \ln\frac{1+\rho_0}{1-\rho_0}\right)\cdot\sqrt{n-3} = \frac{1}{2}\left(\ln\frac{1+0,8}{1-0,8} - \ln\frac{1+0,9}{1-0,9}\right)\cdot\sqrt{35-3} = -2,113.$$

Mức ý nghĩa $\alpha = 0,05$ nên $\alpha/2 = 0,025$ tra Phụ lục 3 có $z_{\alpha/2} = z_{0,025} = 1,96$.

Do đó $|z_0| > z_{\alpha/2}$ nên bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1. Vậy với mức ý nghĩa $\alpha = 5\%$, chưa đủ cơ sở cho rằng $\rho = 0,9$.

8.3. PHÂN TÍCH HỒI QUY

8.3.1. Hàm hồi quy

Giả sử có hai đại lượng ngẫu nhiên X và Y. Cần tìm một hàm φ nào đó của X sao cho nó xấp xỉ Y một cách tốt nhất theo nghĩa tổng bình phương sai số bé nhất.

Giả sử $m(X) = \mathrm{E}(Y|X = x)$. Người ta chứng minh được rằng

$$\mathrm{E}\Big(\big(Y - m(X)\big)^2\Big) = \min_{\varphi}\mathrm{E}\Big(\big(Y - \varphi(X)\big)^2\Big).$$

Nên hàm φ cần tìm là $m(X)$, đó cũng chính là kỳ vọng có điều kiện của Y với điều kiện X đã cho: $\mathrm{E}(Y|X) = m(X)$.

$\mathrm{E}(Y|X)$ được gọi là *hàm hồi quy* hay chính xác hơn là hàm hồi quy đối với giá trị trung bình của Y và ta gọi tắt là hàm hồi quy kỳ vọng. Nó xấp xỉ tốt nhất của Y theo X. Nói chung hai đường hồi quy $\mathrm{E}(Y|X)$ và $\mathrm{E}(X|Y)$ không trùng nhau.

Định lí 8.1. *Nếu phân phối đồng thời của (X,Y) là chuẩn thì các hàm hồi quy $\mathrm{E}(Y|X)$ và $\mathrm{E}(X|Y)$ là tuyến tính, cụ thể:*

$$\mathrm{E}(Y|X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X),\ \mathrm{E}(X|Y) = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(Y-\mu_Y).$$

Trong đó, ρ là hệ số tương quan giữa X và Y.

8.3.2. Hàm hồi quy tuyến tính mẫu

Sau đây ta sẽ tìm hàm hồi quy mẫu trong trường hợp đơn giản nhất, hàm hồi quy tuyến tính: $\varphi(x) = Ax + B$.

Khi đó hàm hồi quy của Y đối với X có dạng: $\widehat{y}_X = \overline{y} + r\dfrac{s_Y}{s_X}(x-\overline{x})$.

Tương tự, hàm hồi quy của X đối với Y có dạng: $\widehat{x}_Y = \overline{x} + r\dfrac{s_X}{s_Y}(y-\overline{y})$.

Sai số bình phương trung bình mẫu là: $s^2_{Y|X} = s^2_Y(1-r^2)$.

Ta thấy rằng $s_{Y|X} = 0$, tức là mọi giá trị quan sát nằm trên đường hồi quy, khi và chỉ khi $r^2 = 1$.

Ví dụ 8.5. Tìm phương trình hồi quy tuyến tính mẫu của Y theo X với số liệu cho ở *Ví dụ 8.2.*

<u>Giải:</u> Ta có: $\overline{x} = 46,8; \overline{y} = 15,14; s_X = 21,113; s_Y = 2,832$ và $r = 0,844$.

Do đó phương trình hồi quy tuyến tính mẫu của Y theo X là:

$$\widehat{y}_X = \overline{y} + r\frac{s_Y}{s_X}(x-\overline{x}) = 15,14 + 0,844\cdot\frac{2,832}{21,113}(x-46,8) = 0,113x + 9,842.$$

8.4. BÀI TẬP CHƯƠNG 8

<u>Bài 1:</u> Một mẫu gồm 7 sinh viên được chọn để nghiên cứu mối quan hệ giữa trung bình điểm thi đại học (X) và trung bình điểm tích lũy ở năm thứ nhất đại học (Y) (thang điểm 10).

X	4	2	6	8	8	8	8
Y	5	4	4	6	7	8	8

a) Tính hệ số tương quan mẫu.

b) Tìm phương trình hồi quy tuyến tính mẫu của Y theo X.

c) Nếu trung bình điểm thi vào đại học của một sinh viên là $6,5$ thì dự đoán trung bình điểm tích lũy ở năm thứ nhất đại học của sinh viên này là bao nhiêu?

Bài 2: Theo dõi mật độ X (số cây/100 m^2) và sản lượng Y (kg/cây) của một loại cây, ta được kết quả:

X	5,8	6,8	7,7	6,5	6,6	5,6	5,7	6,6	5,7	9,4	7,3	7,5	8,4
Y	44	56	62	68	66	52	34	68	65	90	69	90	62

a) Giữa X, Y có mối liên hệ tương quan tuyến tính hay không với mức ý nghĩa 5%.

b) Tìm phương trình hồi quy của Y theo X.

Bài 3: Thống kê nhiệt độ (^{0}C) và độ ẩm (%) các tháng trong một năm ở Hà Nội, người ta có bảng số liệu như sau:

Tháng	1	2	3	4	5	6	7	8	9	10	11	12
Nhiệt độ	17,7	17,2	19,9	25,3	29,3	30,1	29,5	28,9	29,2	27	22,9	17,6
Độ ẩm	71	79	87	88	77	80	81	82	78	73	79	67

a) Tính hệ số tương quan mẫu và sai số bình phương trung bình mẫu.

b) Lập phương trình hồi quy tuyến tính của Y theo X.

c) Với mức ý nghĩa 5%, hãy xét xem có sự phụ thuộc tuyến tính giữa nhiệt độ và độ ẩm hay không?

Bài 4: Điều tra thu nhập một mẫu gồm 10 người trong một công ty, X (năm) là thời gian công tác, Y (triệu đồng) là thu nhập hàng tháng ta thu được số liệu sau:

X	3	3	4	5	5	6	7	7	8	9
Y	3	4	4	6	7	6	7	8	8	8

Tìm hệ số tương quan mẫu và lập phương trình hồi quy tuyến tính của Y theo X.

Bài 5: Bảng dưới đây chỉ ra các kết quả thu hoạch Y (kg) theo lượng phân bón X (kg) của một loại ngũ cốc trồng trên 100 thửa ruộng có cùng diện tích.

X \ Y	1	2	3	4	5	Σ
14	10	8				18
15		12	7			19
16			28	6		34
17				8	9	17
18					12	12
Σ	10	20	35	14	21	100

a) Tìm hệ số tương quan r.

b) Tìm hàm hồi quy tuyến tính mẫu của Y theo X.

Bài 6: Để nghiên cứu sự phát triển của một loại cây trồng một năm tuổi, người ta tiến hành đo chiều cao Y (cm) và đường kính thân X (cm) của một số cây. Kết quả như sau:

X \ Y	30	40	50	60	70	80
$3,5$	2	5				
4		3	11			
$4,5$			8	15	10	
5			4	17	6	
$5,5$					7	12

a) Tìm hệ số tương quan r.

b) Lập phương trình hồi quy tuyến tính mẫu của X đối với Y.

PHỤ LỤC

Phụ lục 1: Giá trị hàm Gauss $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3986	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1319	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0388	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0031	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0006	0006	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003

3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001
x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

Phụ lục 2: Giá trị hàm Laplace $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0,1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0,2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0,3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0,4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0,5	6915	6950	6985	7019	7019	7088	7123	7157	7190	7224
0,6	7257	7291	7324	7357	7389	7421	7454	7486	7517	7549
0,7	7580	7611	7642	7673	7703	7734	7764	7793	7823	7852
0,8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8132
0,9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1,0	8413	8437	8461	8485	8508	8531	8554	8577	8599	8621
1,1	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1,2	8849	8869	8888	8906	8925	8943	8962	8980	8997	9015
1,3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
1,4	9192	9207	9222	9236	9251	9265	9279	9292	9306	9319
1,5	9332	9345	9357	9370	9382	9394	9406	9418	9429	9441
1,6	9452	9463	9474	9483	9495	9505	9515	9525	9535	9545
1,7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1,8	9641	9648	9656	9664	9671	9678	9686	9693	9699	9706
1,9	9713	9719	9726	9732	9738	9744	9750	9756	9761	9767
2,0	9773	9778	9783	9788	9793	9798	9803	9808	9812	9817
2,1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2,2	9861	9864	9868	9871	9875	9878	9881	9884	9887	9890
2,3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2,4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2,5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2,6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964
2,7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2,8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2,9	9981	9982	9983	9983	9984	9984	9985	9985	9986	9986
3,0	9987		**3,1**	9990	**3,2**	9993	**3,3**	9995	**3,4**	9996
3,5	9997		**3,6**	9998	**3,7**	9999	**3,8**	9999	**3,9**	9999
4,0	999968									
4,5	999997									
5,0	99999997									

Phụ lục 3: Phân vị Chuẩn tắc $P(X > z_\alpha) = \alpha$

α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$
0.001	3.090	0.041	1.739	0.081	1.398	0.121	1.170	0.161	0.990	0.201	0.838
0.002	2.878	0.042	1.728	0.082	1.392	0.122	1.165	0.162	0.986	0.202	0.834
0.003	2.748	0.043	1.717	0.083	1.385	0.123	1.160	0.163	0.982	0.203	0.831
0.004	2.652	0.044	1.706	0.084	1.379	0.124	1.155	0.164	0.978	0.204	0.827
0.005	2.576	0.045	1.695	0.085	1.372	0.125	1.150	0.165	0.974	0.205	0.824
0.006	2.512	0.046	1.685	0.086	1.366	0.126	1.146	0.166	0.970	0.206	0.820
0.007	2.457	0.047	1.675	0.087	1.359	0.127	1.141	0.167	0.966	0.207	0.817
0.008	2.409	0.048	1.665	0.088	1.353	0.128	1.136	0.168	0.962	0.208	0.813
0.009	2.366	0.049	1.655	0.089	1.347	0.129	1.131	0.169	0.958	0.209	0.810
0.010	2.326	0.050	1.645	0.090	1.341	0.130	1.126	0.170	0.954	0.210	0.806
0.011	2.290	0.051	1.635	0.091	1.335	0.131	1.122	0.171	0.950	0.211	0.803
0.012	2.257	0.052	1.626	0.092	1.329	0.132	1.117	0.172	0.946	0.212	0.800
0.013	2.226	0.053	1.616	0.093	1.323	0.133	1.112	0.173	0.942	0.213	0.796
0.014	2.197	0.054	1.607	0.094	1.317	0.134	1.108	0.174	0.938	0.214	0.793
0.015	2.170	0.055	1.598	0.095	1.311	0.135	1.103	0.175	0.935	0.215	0.789
0.016	2.144	0.056	1.589	0.096	1.305	0.136	1.098	0.176	0.931	0.216	0.786
0.017	2.120	0.057	1.580	0.097	1.299	0.137	1.094	0.177	0.927	0.217	0.782
0.018	2.097	0.058	1.572	0.098	1.293	0.138	1.089	0.178	0.923	0.218	0.779
0.019	2.075	0.059	1.563	0.099	1.287	0.139	1.085	0.179	0.919	0.219	0.776
0.020	2.054	0.060	1.555	0.100	1.282	0.140	1.080	0.180	0.915	0.220	0.772
0.021	2.034	0.061	1.546	0.101	1.276	0.141	1.076	0.181	0.912	0.221	0.769
0.022	2.014	0.062	1.538	0.102	1.270	0.142	1.071	0.182	0.908	0.222	0.765
0.023	1.995	0.063	1.530	0.103	1.265	0.143	1.067	0.183	0.904	0.223	0.762
0.024	1.977	0.064	1.522	0.104	1.259	0.144	1.063	0.184	0.900	0.224	0.759
0.025	1.960	0.065	1.514	0.105	1.254	0.145	1.058	0.185	0.896	0.225	0.755
0.026	1.943	0.066	1.506	0.106	1.248	0.146	1.054	0.186	0.893	0.226	0.752
0.027	1.927	0.067	1.499	0.107	1.243	0.147	1.049	0.187	0.889	0.227	0.749
0.028	1.911	0.068	1.491	0.108	1.237	0.148	1.045	0.188	0.885	0.228	0.745
0.029	1.896	0.069	1.483	0.109	1.232	0.149	1.041	0.189	0.882	0.229	0.742
0.030	1.881	0.070	1.476	0.110	1.227	0.150	1.036	0.190	0.878	0.230	0.739
0.031	1.866	0.071	1.468	0.111	1.221	0.151	1.032	0.191	0.874	0.231	0.736
0.032	1.852	0.072	1.461	0.112	1.216	0.152	1.028	0.192	0.871	0.232	0.732
0.033	1.838	0.073	1.454	0.113	1.211	0.153	1.024	0.193	0.867	0.233	0.729
0.034	1.825	0.074	1.447	0.114	1.206	0.154	1.019	0.194	0.863	0.234	0.726
0.035	1.812	0.075	1.440	0.115	1.200	0.155	1.015	0.195	0.860	0.235	0.722
0.036	1.799	0.076	1.433	0.116	1.195	0.156	1.011	0.196	0.856	0.236	0.719
0.037	1.787	0.077	1.426	0.117	1.190	0.157	1.007	0.197	0.852	0.237	0.716
0.038	1.774	0.078	1.419	0.118	1.185	0.158	1.003	0.198	0.849	0.238	0.713
0.039	1.762	0.079	1.412	0.119	1.180	0.159	0.999	0.199	0.845	0.239	0.710
0.040	1.751	0.080	1.405	0.120	1.175	0.160	0.994	0.200	0.842	0.240	0.706

α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$	α	$\mathbf{z}_\alpha$
0.241	0.703	0.285	0.568	0.329	0.443	0.372	0.327	0.415	0.215	0.458	0.105
0.242	0.700	0.286	0.565	0.330	0.440	0.373	0.324	0.416	0.212	0.459	0.103
0.243	0.697	0.287	0.562	0.331	0.437	0.374	0.321	0.417	0.210	0.460	0.100
0.244	0.693	0.288	0.559	0.332	0.434	0.375	0.319	0.418	0.207	0.461	0.098
0.245	0.690	0.289	0.556	0.333	0.432	0.376	0.316	0.419	0.204	0.462	0.095
0.246	0.687	0.290	0.553	0.334	0.429	0.377	0.313	0.420	0.202	0.463	0.093
0.247	0.684	0.291	0.550	0.335	0.426	0.378	0.311	0.421	0.199	0.464	0.090
0.248	0.681	0.292	0.548	0.336	0.423	0.379	0.308	0.422	0.197	0.465	0.088
0.249	0.678	0.293	0.545	0.337	0.421	0.380	0.305	0.423	0.194	0.466	0.085
0.250	0.674	0.294	0.542	0.338	0.418	0.381	0.303	0.424	0.192	0.467	0.083
0.251	0.671	0.295	0.539	0.339	0.415	0.382	0.300	0.425	0.189	0.468	0.080
0.252	0.668	0.296	0.536	0.340	0.412	0.383	0.298	0.426	0.187	0.469	0.078
0.253	0.665	0.297	0.533	0.341	0.410	0.384	0.295	0.427	0.184	0.470	0.075
0.254	0.662	0.298	0.530	0.342	0.407	0.385	0.292	0.428	0.181	0.471	0.073
0.255	0.659	0.299	0.527	0.343	0.404	0.386	0.290	0.429	0.179	0.472	0.070
0.256	0.656	0.300	0.524	0.344	0.402	0.387	0.287	0.430	0.176	0.473	0.068
0.257	0.653	0.301	0.522	0.345	0.399	0.388	0.285	0.431	0.174	0.474	0.065
0.258	0.650	0.302	0.519	0.346	0.396	0.389	0.282	0.432	0.171	0.475	0.063
0.259	0.646	0.303	0.516	0.347	0.393	0.390	0.279	0.433	0.169	0.476	0.060
0.260	0.643	0.304	0.513	0.348	0.391	0.391	0.277	0.434	0.166	0.477	0.058
0.261	0.640	0.305	0.510	0.349	0.388	0.392	0.274	0.435	0.164	0.478	0.055
0.262	0.637	0.306	0.507	0.350	0.385	0.393	0.272	0.436	0.161	0.479	0.053
0.263	0.634	0.307	0.504	0.351	0.383	0.394	0.269	0.437	0.159	0.480	0.050
0.264	0.631	0.308	0.502	0.352	0.380	0.395	0.266	0.438	0.156	0.481	0.048
0.265	0.628	0.309	0.499	0.353	0.377	0.396	0.264	0.439	0.154	0.482	0.045
0.266	0.625	0.310	0.496	0.354	0.375	0.397	0.261	0.440	0.151	0.483	0.043
0.267	0.622	0.311	0.493	0.355	0.372	0.398	0.259	0.441	0.148	0.484	0.040
0.268	0.619	0.312	0.490	0.356	0.369	0.399	0.256	0.442	0.146	0.485	0.038
0.269	0.616	0.313	0.487	0.357	0.366	0.400	0.253	0.443	0.143	0.486	0.035
0.270	0.613	0.314	0.485	0.358	0.364	0.401	0.251	0.444	0.141	0.487	0.033
0.271	0.610	0.315	0.482	0.359	0.361	0.402	0.248	0.445	0.138	0.488	0.030
0.272	0.607	0.316	0.479	0.360	0.358	0.403	0.246	0.446	0.136	0.489	0.028
0.273	0.604	0.317	0.476	0.361	0.356	0.404	0.243	0.447	0.133	0.490	0.025
0.274	0.601	0.318	0.473	0.362	0.353	0.405	0.240	0.448	0.131	0.491	0.023
0.275	0.598	0.319	0.470	0.363	0.350	0.406	0.238	0.449	0.128	0.492	0.020
0.276	0.595	0.320	0.468	0.364	0.348	0.407	0.235	0.450	0.126	0.493	0.018
0.277	0.592	0.321	0.465	0.365	0.345	0.408	0.233	0.451	0.123	0.494	0.015
0.278	0.589	0.322	0.462	0.366	0.342	0.409	0.230	0.452	0.121	0.495	0.013
0.279	0.586	0.323	0.459	0.367	0.340	0.410	0.228	0.453	0.118	0.496	0.010
0.280	0.583	0.324	0.457	0.368	0.337	0.411	0.225	0.454	0.116	0.497	0.008
0.281	0.580	0.325	0.454	0.369	0.335	0.412	0.222	0.455	0.113	0.498	0.005
0.282	0.577	0.326	0.451	0.370	0.332	0.413	0.220	0.456	0.111	0.499	0.003
0.283	0.574	0.327	0.448	0.371	0.329	0.414	0.217	0.457	0.108	0.500	0.000
0.284	0.571	0.328	0.445								

Phụ lục 4: Phân vị Chi bình phương $P(X > \chi^2_{(n,\alpha)}) = \alpha$

n \ α	0,995	0,99	0,975	0,95	0,05	0,025	0,01	0,005
1	0,000039	0,000157	0,00098	0,00393	3,841	5,024	6,635	7,879
2	0,0100	0,0201	0,0506	0,103	5,991	7,378	9,210	10,597
3	0,0717	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	11,071	12,833	15,086	16,749
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,590
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,758
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,299
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,820
14	4,075	4,660	5,629	6,571	23,685	26,119	29,142	31,320
15	4,601	5,229	6,262	7,261	24,996	27,489	30,578	32,802
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,268
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,717
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,853	36,191	38,581
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,400
22	8,643	9,542	10,982	12,338	33,926	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,075	41,638	44,184
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,930
26	11,160	12,198	13,844	15,379	38,885	41,924	45,643	48,290
27	11,808	12,878	14,573	16,151	40,113	43,195	46,963	49,647
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,338
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,673

Phụ lục 5: Phân vị Student $P(X > t_{(n,\alpha)}) = \alpha$

n \ α	0,1	0,05	0,025	0,01	0,005	0,0005
1	3,078	6,314	12,706	31,821	63,657	363,6
2	1,886	2,920	4,303	6,965	9,925	31,600
3	1,638	2,353	3,182	4,541	5,841	12,922
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,869
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,408
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,769	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,080	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,767
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,310	1,697	2,042	2,457	2,750	3,646
∞	1,282	1,645	1,960	2,326	2,576	3,291

Phụ lục 6: Phân vị Fisher $P(X > f_{(m,n,\alpha)}) = \alpha = 0,05$

n \ m	1	2	3	4	5	6	7	8	9
1	161,41	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96
$+\infty$	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88

m / n	10	12	15	20	24	30	40	60	120	$+\infty$
1	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,65	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
9	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,578	2,54
10	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
11	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
12	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
13	2,60	2,53	2,46	2,39	2,35	2,31	2,34	2,30	2,25	2,21
14	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
15	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,10	2,06	2,02
16	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
17	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
18	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
19	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
20	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
21	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
22	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
23	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
24	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
25	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
30	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
40	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
60	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
120	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

TÀI LIỆU THAM KHẢO

[1] **Lê Sĩ Đồng**, *Xác suất thống kê và ứng dụng*, NXB Giáo dục, 2011.

[2] **Đinh Văn Gắng**, *Lí thuyết Xác suất và thống kê*, NXB Giáo dục, 2012.

[3] **Đinh Văn Gắng**, *Bài tập Xác suất và thống kê*, NXB Giáo dục, 2014.

[4] **Đào Hữu Hồ**, *Xác suất thống kê*, NXB ĐHQG Hà Nội, 2010.

[5] **Đặng Hấn**, *Xác suất thống kê*, NXB Thống kê, 1996.

[6] **Đặng Hấn**, *Bài tập Xác suất thống kê*, NXB Thống kê, 1996.

[7] **Phạm Văn Kiều**, *Xác suất thống kê*, NXB ĐH Sư phạm, 2000.

[8] **N. Balakrishnan, Markos V. Koutras, Konstadinos G. Politis**, *Introduction to Probability: Models and Applications*, John Wiley & Sons, Inc., 2020.

[9] **Géza Schay**, *Introduction to Probability with Statistical Applications*, Birkhäuser Basel, 2016.