

TỔNG ĐÌNH QUỲ

GIÁO TRÌNH
XÁC SUẤT
THỐNG KÊ

(Tái bản lần thứ năm)

NHÀ XUẤT BẢN BÁCH KHOA – HÀ NỘI

LỜI NÓI ĐẦU

Lý thuyết xác suất và thống kê toán học là một ngành khoa học đang giữ vị trí quan trọng trong các lĩnh vực ứng dụng rộng rãi và phong phú của đời sống con người. Cùng với sự phát triển mạnh mẽ của khoa học và công nghệ, nhu cầu hiểu biết và sử dụng các công cụ ngẫu nhiên trong phân tích và xử lý thông tin ngày càng trở nên đặc biệt cần thiết. Các kiến thức và phương pháp của xác suất và thống kê đã hỗ trợ hữu hiệu các nhà nghiên cứu trong nhiều lĩnh vực khoa học khác nhau như vật lý, hóa học, sinh y học, nông học, kinh tế học, xã hội học, ngôn ngữ học...

Trong một chục năm gần đây, giáo trình xác suất thống kê đã trở thành cơ sở của nhiều ngành học trong các trường đại học và cao đẳng, từ đó xuất hiện nhu cầu học tập và nghiên cứu ứng dụng rất lớn, nhất là đối với sinh viên các ngành khoa học không chuyên về toán. Để thỏa mãn yêu cầu đó, giáo trình này cố gắng đáp ứng đòi hỏi của đông đảo sinh viên nhằm hiểu biết sâu sắc hơn các khái niệm và phương pháp tính xác suất và thống kê để học tập đạt hiệu quả cao hơn cũng như ứng dụng môn học vào ngành học và môn học khác.

Giáo trình xác suất thống kê được viết cho thời gian giảng dạy là 60 tiết học. Do đối tượng sinh viên rất đa dạng với trình độ toán cơ bản khác nhau, chúng tôi đã cố gắng tìm những cách tiếp cận đơn giản và hợp lý, và như vậy đã buộc phải bớt đi phần nào sự chặt chẽ hình thức (vốn rất đặc trưng cho toán học) để giúp bạn đọc tiếp cận dễ dàng hơn bản chất xác suất của các vấn đề đặt ra và tăng cường kỹ năng phân tích, xử lý các tình huống, từ đó dần dần hình thành một hệ thống khái niệm khá đầy đủ để đi sâu giải quyết các bài toán ngày càng phức tạp hơn.

Giáo trình được chia thành 6 chương gồm 3 chương dành cho phần xác suất và 3 chương cho phần phân tích thống kê. Những khái niệm và công thức cơ bản được trình bày tương đối đơn giản, dễ hiểu và được

minh hoạ bằng nhiều thí dụ áp dụng. Các chứng minh khó được lược bớt có chọn lọc để giáo trình không quá cồng kềnh, mặc dù vậy các công thức và vấn đề liên quan đều được nhắc đến đầy đủ để tiện không chỉ cho học tập sâu hơn, mà còn có ích cho những bạn đọc muốn tra cứu, tìm tòi phục vụ cho ứng dụng và tính toán thống kê. Cuối mỗi chương có một loạt bài tập dành để bạn đọc tự giải nhằm hiểu biết sâu sắc hơn lý thuyết và rèn luyện kỹ năng thực hành.

Hy vọng rằng giáo trình có ích cho bạn đọc xa gần, các sinh viên, cán bộ giảng dạy ở các trường đại học và cao đẳng, các cán bộ khoa học và kinh tế muốn tự học và tự nghiên cứu xác suất thống kê – môn học thường được coi là khó tiếp thu. Tác giả cũng cảm ơn mọi ý kiến góp ý để quyển sách sẽ ngày càng được hoàn thiện hơn để góp phần nâng cao chất lượng dạy và học môn học này.

Trong lần tái bản này tại Nhà xuất bản Bách Khoa – Hà Nội, một số lỗi chế bản đã được sửa chữa. Tác giả một lần nữa tỏ lời cảm ơn đến những ý kiến góp ý của đông đảo bạn đọc để cải tiến giáo trình trong lần tái bản tiếp theo.

TÁC GIẢ

Chương I

SỰ KIẾN NGẪU NHIÊN VÀ PHÉP TÍNH XÁC SUẤT

§1. KHÁI NIỆM MỞ ĐẦU

1.1. Sự kiện ngẫu nhiên

Khái niệm thường gặp trong lý thuyết xác suất là *sự kiện* (mà không thể định nghĩa chặt chẽ). Sự kiện được hiểu như là một sự việc, một hiện tượng nào đó của cuộc sống tự nhiên và xã hội.

Khi thực hiện một tập hợp điều kiện xác định, nói tắt là bộ điều kiện, gọi là một *phép thử*, có thể có nhiều kết cục khác nhau.

Thí dụ 1.1. Gieo một con xúc sắc đồng chất trên một mặt phẳng (phép thử). Phép thử này có 6 kết cục là: xuất hiện mặt 1, mặt 2, ..., mặt 6 chấm. Mỗi kết cục này cùng với các kết quả phức tạp hơn như: xuất hiện mặt có số chấm chẵn, mặt có số chấm bội 3, đều có thể coi là các sự kiện.

Như vậy kết cục của một phép thử là một trường hợp riêng của sự kiện. Để cho tiện lợi sau này, ta ký hiệu sự kiện bằng các chữ cái in hoa $A, B, C, \dots$. Sự kiện được gọi là *tất yếu*, nếu nó chắc chắn xảy ra, và được gọi là *bất khả*, nếu nó không thể xảy ra khi thực hiện phép thử. Còn nếu sự kiện có thể xảy ra hoặc không sẽ được gọi là *sự kiện ngẫu nhiên*. Từ đó, theo một nghĩa nào đó, có thể coi các sự kiện tất yếu, ký hiệu là U , và bất khả, ký hiệu là V , như các trường hợp riêng của sự kiện ngẫu nhiên. Thí dụ, dưới những điều kiện xác định, nước đóng băng ở 0°C là sự kiện tất yếu; khi gieo một con xúc sắc, việc xuất hiện mặt bảy chấm là sự kiện bất khả...

Để mô tả một phép thử người ta xác định tập hợp các kết cục có thể có. Tập hợp tất cả các kết cục của một phép thử (được gọi là các *sự kiện sơ cấp*, ký hiệu là ω_i) tạo thành không gian các sự kiện sơ cấp, ký hiệu là $\Omega = \{\omega_i, i \in I\}$, I là tập chỉ số, có thể vô hạn (đếm được hoặc không đếm được). Dễ thấy trong thí dụ 1.1, nếu ký hiệu A_i – sự kiện xuất hiện mặt i chấm ($i = \overline{1,6}$) thì $\Omega = \{A_1, A_2, A_3, A_4, A_5, A_6\} = \{A_i, i = \overline{1,6}\}$.

Trong nhiều hiện tượng hàng loạt khi thực hiện nhiều lần cùng một phép thử, ta thấy tần suất xuất hiện một sự kiện A nào đó chênh lệch không nhiều so với một số đặc trưng cho khả năng xuất hiện A . Số đó được gọi là *xác suất* xuất hiện A và được ký hiệu là $P(A)$. Như vậy nếu viết $P(A) = p$ có nghĩa là xác suất xảy ra sự kiện A là bằng p .

Một câu hỏi tự nhiên là. Do đâu có sự kiện ngẫu nhiên? Và chúng ta có thể nhận biết được chúng không? Thực ra mỗi sự kiện đều xảy ra theo quy luật nào đó; song do điều kiện thiếu tri thức, thông tin và phương tiện cần thiết (cả về kinh phí, thiết bị lẫn thời gian) nên ta không có khả năng nhận thức đầy đủ về sự kiện đó. Vấn đề càng trở nên khó khăn hơn khi chỉ cần có một sự thay đổi bất ngờ dù rất nhỏ của bộ điều kiện đã làm thay đổi kết cục của phép thử. Cho nên bài toán xác định bản chất xác suất của một sự kiện bất kỳ trong một phép thử tùy ý là không thể giải được.

1.2. Phép toán và quan hệ của các sự kiện

Về mặt toán học, việc nghiên cứu quan hệ và phép toán trên tập các sự kiện cho phép ta xác định chúng thực chất hơn.

(i) *Tổng* của A và B , ký hiệu là $A + B$, chỉ sự kiện khi có xuất hiện ít nhất một trong hai sự kiện trên.

(ii) *Tích* của A và B , ký hiệu là AB , chỉ sự kiện khi có xuất hiện đồng thời cả hai sự kiện trên.

(iii) *Đôi lập* của A , ký hiệu là $\bar{A}$, chỉ sự kiện không xuất hiện A . Rõ ràng đôi lập có tính tương hỗ $\bar{\bar{A}} = A$ và $A + \bar{A} = U$, $A\bar{A} = V$, $\bar{U} = V$.

(iv) *Xung khắc*: hai sự kiện A và B được gọi là xung khắc nếu chúng không thể đồng thời xảy ra, tức là $AB = V$.

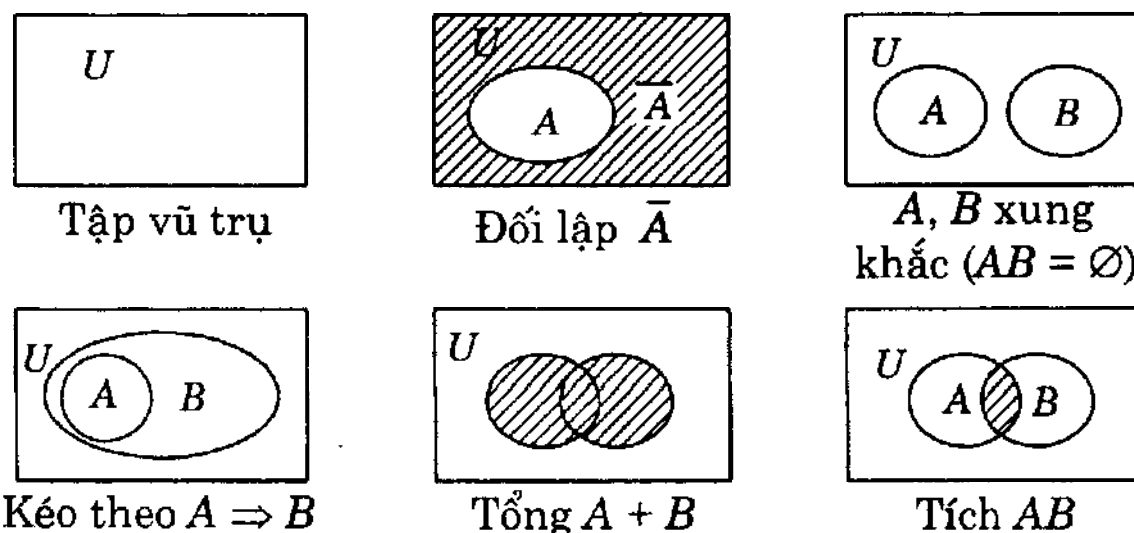
(v) *Kéo theo*, ký hiệu $A \Rightarrow B$, chỉ nếu xuất hiện A thì xuất hiện B .

(vi) *Tương đương*, ký hiệu $A = B$, chỉ việc nếu xuất hiện A thì xuất hiện B và ngược lại.

(vii) *Hiệu* của A và B , ký hiệu $A - B$ (hoặc $A \setminus B$), chỉ sự kiện xuất hiện A nhưng không xuất hiện B , tức là $A - B = A\bar{B}$.

Các khái niệm cho thấy tính đôi lập, tổng, tích và hiệu của hai kiện tương ứng với bù, hợp, giao và hiệu của hai tập hợp. Như vậy có thể sử dụng các tính chất của các phép toán trên tập hợp cho các phép toán trên sự kiện, chẳng hạn dùng sơ đồ Ven trong thí dụ sau đây.

Thí dụ 1.2. Ký hiệu U là tập vũ trụ, V là tập $\emptyset$ (rỗng). Khi đó A và B sẽ là các tập con của U và các phép toán trên A và B có thể minh họa bằng sơ đồ Ven (xem hình 1.1).



Hình 1.1

Từ đó, dễ dàng chỉ ra các công thức sau:

$$A + B = B + A, AB = BA \text{ (giao hoán);}$$

$$A + (B + C) = (A + B) + C, A(BC) = (AB)C \text{ (kết hợp);}$$

$$A(B + C) = AB + AC \text{ (phân phối);}$$

$$A + U = U, A + V = A, A + A = A;$$

$$AU = A, AV = V, AA = A.$$

Thí dụ 1.3. Chọn từ một lô hàng ra 5 sản phẩm và ta quan tâm đến số phế phẩm trong 5 sản phẩm đó (phép thử).

a) Xác định các sự kiện sơ cấp.

b) Biểu diễn các sự kiện sau theo các sự kiện sơ cấp: có nhiều nhất 1 phế phẩm; có không quá 4 phế phẩm, có ít nhất 1 phế phẩm.

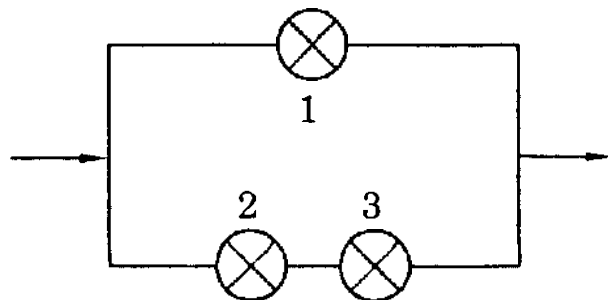
Giải. a) Ký hiệu A_i – trong 5 sản phẩm có i phế phẩm. Rõ ràng $i = 0, 5$ và $\Omega = \{A_0, A_1, A_2, A_3, A_4, A_5\}$.

b) Gọi A , B và C là các sự kiện tương ứng. Dễ dàng biểu diễn $A = A_0 + A_1$, $B = A_0 + A_1 + A_2 + A_3 + A_4 = \overline{A_5}$, $C = A_1 + A_2 + A_3 + A_4 + A_5 = \overline{A_0}$.

Thí dụ 1.4. Cho sơ đồ mạng điện trên hình 1.2 gồm 3 bóng đèn. Việc mạng mất điện (sự kiện A) chỉ có thể xảy ra do cháy các bóng đèn (ký hiệu là A_1, A_2, A_3). Hãy biểu diễn A theo các $A_i, i = 1, 2, 3$.

Giải. A xuất hiện khi xảy ra một trong 3 trường hợp:

- (i) cả ba bóng cháy,
- (ii) cháy hai bóng 1 và 2,
- (iii) cháy hai bóng 1 và 3.



Hình 1.2

Từ đó ta có $A = A_1 A_2 A_3 + A_1 A_2 \overline{A_3} + A_1 \overline{A_2} A_3$.

Có thể dùng tính chất của mạng song song và nối tiếp để có một biểu diễn khác gọn hơn:

$$A = A_1(A_2 + A_3).$$

Trong nhiều bài tập, việc xác định số lượng các sự kiện sơ cấp đưa đến sử dụng các kết quả của lý thuyết tổ hợp.

1.3. Giải tích kết hợp

Việc đếm số các kết cục của một phép thử dựa vào mô hình: chọn hú họa ra k phần tử từ n phần tử cho trước. Nếu phân biệt thứ tự các phần tử chọn ra, ta có khái niệm chỉnh hợp; nếu thứ tự không phân biệt, ta có tổ hợp.

(i) *Chỉnh hợp*: chỉnh hợp chập k từ n là một nhóm có thứ tự gồm k phần tử lấy từ n đã cho. Đó chính là một nhóm gồm k phần tử khác nhau được xếp theo thứ tự nhất định. Số các chỉnh hợp như vậy, ký hiệu là A_n^k .

$$A_n^k = n(n-1)\dots(n-k+1) = \frac{n!}{(n-k)!}. \quad (1.1)$$

(ii) *Chỉnh hợp lặp*: chỉnh hợp lặp chập k từ n là một nhóm có thứ tự gồm k phần tử có thể giống nhau lấy từ n đã cho. Đó chính là một nhóm gồm k phần tử có thể lặp lại và được xếp theo thứ tự nhất định. Số các chỉnh hợp lặp như vậy, ký hiệu là $\tilde{A}_n^k$.

$$\tilde{A}_n^k = n^k. \quad (1.2)$$

(iii) *Hoán vị*: hoán vị của n là một nhóm gồm n phần tử được sắp xếp theo một thứ tự nào đó. Rõ ràng số các hoán vị như vậy, ký hiệu là P_n , chính là số các chỉnh hợp A_n^n và

$$P_n = n! \quad (1.3)$$

(iv) *Tổ hợp*: tổ hợp chập k từ n là một nhóm (không phân biệt thứ tự) gồm k phần tử khác nhau lấy từ n đã cho. Số các tổ hợp như vậy, ký hiệu là C_n^k .

$$C_n^k = \frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!}. \quad (1.4)$$

Thí dụ 1.5. Cho một tập hợp gồm 3 phần tử $\{a, b, c\}$. Có thể tạo ra bao nhiêu nhóm gồm 2 phần tử chọn từ tập trên?

Giải:

(i) Nếu ta để ý đến thứ tự các phần tử và mỗi phần tử chỉ được chọn một lần, số nhóm thu được sẽ là $A_3^2 = 3.2 = 6$; đó là $\{a, b\}; \{b, a\}; \{a, c\}; \{c, a\}; \{b, c\}, \{c, b\}$.

(ii) Nếu vẫn để ý đến thứ tự, nhưng mỗi phần tử được chọn nhiều lần, số nhóm thu được trở thành $\tilde{A}_3^2 = 3^2 = 9$; đó là:

$\{a, b\}; \{b, a\}; \{a, c\}; \{c, a\}; \{b, c\}, \{c, b\}; \{a, a\}; \{b, b\}; \{c, c\}$.

(iii) Nếu không để ý đến thứ tự các phần tử và chúng chỉ được chọn một lần, số nhóm thu được trở thành $C_3^2 = 3$; đó là

$\{a, b\}; \{a, c\}; \{b, c\}$.

Thí dụ 1.6. Một lớp phải học 6 môn trong học kỳ, mỗi ngày học 3 môn. Hỏi có bao nhiêu cách xếp thời khóa biểu trong 1 ngày?

Giải. Số cách xếp cần tìm chính là số cách ghép 3 môn từ 6 môn, trong đó các cách ghép sẽ khác nhau nếu có ít nhất một môn khác nhau hoặc thứ tự môn khác nhau. Từ đó theo (1.1) ta có số cách cần tìm là $A_6^3 = 6.5.4 = 120$.

Thí dụ 1.7. Có thể đánh số được bao nhiêu xe nếu chỉ dùng 3 con số từ 1 đến 5?

Giải. Mỗi số thứ tự của một xe dễ thấy là chỉnh hợp lặp chập 3 từ 5. Từ đó theo (1.2) ta có số lượng xe được đánh số sẽ là

$$\tilde{A}_5^3 = 5^3 = 125.$$

Thí dụ 1.8. Có bao nhiêu cách lập một hội đồng gồm 3 người chọn trong số 8 người?

Giải. Hội đồng là một nhóm 3 người lấy từ 8 người, do đó theo (1.4) sẽ có $C_8^3 = 8!/(3!5!) = 56$ cách lập.

Cuối cùng, để ý là ta đã rất quen thuộc với khái niệm tổ hợp được dùng trong công thức nhị thức Niu-tơn

$$(x + a)^n = C_n^0 x^n + C_n^1 x^{n-1} a + \dots + C_n^k x^{n-k} a^k + \dots + C_n^n a^n.$$

Từ đó có thể dễ dàng chứng minh (để ý $C_n^0 = C_n^n = 1$)

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

§2. CÁC ĐỊNH NGHĨA CỦA XÁC SUẤT

2.1. Định nghĩa cổ điển

Trong mục này ta làm việc với các phép thử có kết cục *đồng khả năng*. Khái niệm đồng khả năng đóng vai trò chủ đạo và khó có thể định nghĩa một cách hình thức. Xét thí dụ đơn giản sau đây:

Thí dụ 2.1. Trong một hộp có n viên bi giống nhau về kích cỡ và chỉ khác nhau về màu sắc, trong đó có m bi trắng và $n - m$ bi đỏ. Rút hù họa ra một viên bi (phép thử). Do số viên bi là n nên tổng số các kết cục khác nhau sẽ là n , và vì tính giống nhau của chúng nên mỗi viên bi có cùng khả năng được rút. Bây giờ nếu gọi A là sự kiện rút được bi trắng thì trong số n kết cục đồng khả năng có m kết cục thuận lợi cho A . Vì vậy trực giác cho thấy nên chọn tỷ số m/n làm xác suất của việc xuất hiện A .

Định nghĩa. Cho một phép thử với n kết cục đồng khả năng, trong đó có m kết cục thuận lợi cho A , khi đó

$$P(A) = \frac{m}{n} = \frac{\text{số kết cục thuận lợi cho } A}{\text{tổng số kết cục có thể}}. \quad (2.1)$$

Định nghĩa trên được gọi là *định nghĩa cổ điển* của xác suất. Cách tính xác suất theo (2.1) có ưu điểm là tương đối đơn giản và trực quan, tuy nhiên phạm vi áp dụng rất hạn chế chỉ cho các loại phép thử gồm hữu hạn kết cục đồng khả năng. Trong tính toán thường sử dụng các kết quả (1.1) – (1.4).

Thí dụ 2.2. Gieo đồng thời 2 con xúc sắc giống nhau. Tính xác suất để tổng số chấm thu được bằng 6.

Giải. Phép thử có $6.6 = 36$ kết cục (sự kiện sơ cấp) khác nhau đồng khả năng. Gọi A là sự kiện “tổng số chấm bằng 6”, thì có tất cả 5 kết cục thuận lợi cho A là $\{1,5\}$, $\{2,4\}$, $\{3,3\}$, $\{4,2\}$ và $\{5,1\}$ (số thứ nhất chỉ số chấm của con xúc sắc 1, số thứ 2 – số chấm của con xúc sắc 2). Vậy $P(A) = 5/36$.

Thí dụ 2.3. Trong hộp có 4 viên bi trắng và 6 viên bi đỏ cùng kích cỡ. Rút hú họa ra 2 bi, tính các xác suất để trong đó có:

- a) hai viên trắng;
- b) ít nhất 1 viên đỏ;
- c) viên thứ hai đỏ.

Giải. Ta dùng định nghĩa cổ điển ở trên.

a) Tổng số cách để rút ra 2 bi có quan tâm đến thứ tự là $A_{10}^2 = 10.9 = 90$, trong đó số cách thuận lợi cho A – rút được 2 bi trắng – là $A_4^2 = 4.3 = 12$; vậy xác suất cần tìm $P(A) = 12/90 = 2/15$. Có thể sử dụng khái niệm tổ hợp để tính xác suất: tổng số cách lấy ra 2 bi từ 10 viên bi là C_{10}^2 (không quan tâm đến thứ tự), trong đó để rút ra 2 bi trắng có C_4^2 cách. Từ đó ta có cùng kết quả như trên.

b) Có thể tính trực tiếp xác suất của B – sự kiện rút được ít nhất 1 bi đỏ (tức là hoặc được 1 hoặc cả 2 bi đỏ). Dễ thấy sự kiện đối lập $\bar{B}$ – cả 2 bi đều trắng – đã có xác suất hiện bằng $2/15$. Từ đó $P(B) = 1 - P(\bar{B}) = 13/15$ (xem tính chất của xác suất ngay dưới đây).

c) Gọi C là sự kiện viên bi thứ hai màu đỏ. Số cách thuận lợi cho C bao gồm (có quan tâm đến thứ tự): $6 \cdot 5 = 30$ cách đối với trường hợp viên bi đầu màu đỏ và $4 \cdot 6 = 24$ cách đối với trường hợp viên bi đầu màu trắng. Từ đó $P(C) = (30 + 24)/90 = 3/5$. Có thể lý luận đơn giản hơn như sau: do viên bi đầu không biết màu sắc nên thông tin về tỷ lệ màu không thay đổi với viên bi thứ hai. Vậy sự kiện C sẽ có cùng xác suất với việc rút hủ họa ra 1 bi đỏ từ hộp 10 viên ban đầu và xác suất của sự kiện đó rất dễ tính là $6/10 = 3/5$.

Dùng công thức (2.1) dễ dàng chứng minh các tính chất sau đây của xác suất (đúng cho cả các trường hợp định nghĩa khác):

- (i) $1 \geq P(A) \geq 0$;
- (ii) $P(U) = 1$; $P(V) = 0$;
- (iii) Nếu A, B xung khắc thì $P(A + B) = P(A) + P(B)$;
- (iv) $P(\bar{A}) = 1 - P(A)$;
- (v) Nếu $A \Rightarrow B$ thì $P(A) \leq P(B)$.

Để khắc phục hạn chế của (2.1) chỉ áp dụng cho các phép thử có hữu hạn kết cục, người ta đưa ra *định nghĩa hình học* của xác suất. Gải sử tập hợp (vô hạn) các kết cục đồng khả năng của một phép thử có thể biểu thị bởi một miền hình học G (chẳng hạn đoạn thẳng, một miền mặt cong hoặc khối không gian...), còn tập các kết cục thuận lợi cho A bởi một miền con nào đó $S \subseteq G$. Sẽ rất hợp lý nếu ta định nghĩa xác suất bằng tỷ số độ đo của S với G (phụ thuộc vào S và G mà độ đo có thể là độ dài, diện tích hoặc thể tích...). Như vậy ta có $P(A)$ bằng xác suất để điểm gieo rơi vào S , với giả thiết nó có thể rơi đồng khả năng vào các điểm của G và

$$P(A) = \frac{\text{độ đo } S}{\text{độ đo } G}. \quad (2.2)$$

Khái niệm “rơi đồng khả năng vào G ” có nghĩa là điểm gieo có thể rơi vào bất kỳ điểm nào của G và xác suất để nó rơi vào một miền con nào đó của G tỷ lệ với độ đo của miền ấy, mà không phụ thuộc vào vị trí và hình dạng của miền.

Thí dụ 2.4. Đường dây điện thoại ngầm nối một tổng đài với một trạm dài 1km. Tính xác suất để dây đứt tại nơi cách tổng đài không quá 100m.

Giải. Rõ ràng nếu dây điện thoại đồng chất, khả năng nó bị đứt tại một điểm bất kỳ là như nhau, nên tập hợp các kết cục đồng khả năng có thể biểu thị bằng đoạn thẳng nối tổng đài với trạm. Các kết cục thuận lợi cho A – sự kiện chỗ đứt cách tổng đài không quá 100m – được biểu thị bằng đoạn thẳng có độ dài 100m. Từ đó theo (2.2) $P(A) = 100/1000 = 0,1$.

Một số bài toán thực tế khác có thể đưa về mô hình dạng trên. Chú ý rằng theo cách định nghĩa này thì sự kiện có xác suất bằng 0 vẫn có thể xảy ra (chẳng hạn mũi tên bắn trúng một điểm cho trước...). Tính chất này rất đặc trưng cho các biến ngẫu nhiên liên tục sẽ nghiên cứu ở chương II.

2.2. Định nghĩa thống kê

Điều kiện đồng khả năng của các kết cục một phép thử không phải lúc nào cũng được bảo đảm. Có nhiều hiện tượng xảy ra không theo các yêu cầu của định nghĩa cổ điển, chẳng hạn khi tính xác suất một đứa trẻ sắp sinh là con trai, ngày mai trời mưa vào lúc chính Ngọ, v.v...

Có một cách khác để xác định xác suất của một sự kiện. Giả sử tiến hành một loạt n_1 phép thử cùng loại, nếu sự kiện A nào đó xuất hiện trong m_1 phép thử thì ta gọi m_1/n_1 là tần suất xuất hiện A trong loạt phép thử đã cho. Tương tự với loại phép thử thứ hai, thứ ba... ta có các tần suất tương ứng $m_2/n_2, m_3/n_3, \dots$

Trên cơ sở quan sát lâu dài các thí nghiệm khác nhau người ta nhận thấy tần suất xuất hiện một sự kiện có tính ổn định, thay đổi rất ít trong các loạt phép thử khác nhau và dao động xung quanh một hằng số xác định. Sự khác biệt đó càng ít khi số phép thử tăng nhiều lên. Hơn nữa đối với các phép thử xét ở mục 2.1 hằng số xác định đó trùng với xác suất theo định nghĩa cổ điển. Đặc tính ổn định của tần suất khi số phép thử tăng lên khá lớn cho phép ta định nghĩa xác suất của sự kiện là trị số ổn định đó của tần suất xuất hiện sự kiện. Nhưng do hằng số đó chưa biết, nên người ta lấy ngay tần suất khi số phép thử đủ lớn làm xác suất của sự kiện. Cách hiểu như vậy được gọi là *định nghĩa thống kê* của xác suất.

Như vậy xác suất ở đây là một giá trị gần đúng và nhiều người cho rằng đó không phải là một định nghĩa thật sự. Tuy nhiên, trong nhiều ngành khoa học thực nghiệm xác suất được xác định theo cách này đạt độ chính xác khá lớn và rất phù hợp với thực tế cũng như với tính toán lý thuyết, nhiều khi sai số phạm phải bé hơn nhiều so với sai số đo của thí nghiệm. Vì thế định nghĩa thống kê vẫn được thừa nhận rộng rãi và rất có ý nghĩa. Ta có thể định nghĩa chặt chẽ hơn về mặt toán học như sau: xác suất của sự kiện là giới hạn của tần suất xuất hiện sự kiện đó khi số phép thử tăng vô hạn. Sự hợp lý của định nghĩa được minh chứng không chỉ bằng thực nghiệm mà cả bằng lý thuyết (sau này ta sẽ thấy rõ trong luật số lớn Béc-nu-li).

Có nhiều thí dụ minh họa tính ổn định của tần suất khi số phép thử khá lớn. Ta có thể tham khảo dưới đây các tần suất xuất hiện mặt sấp khi gieo một đồng tiền nhiều lần:

<i>Người thí nghiệm</i>	<i>Số lần gieo</i>	<i>Số lần sấp</i>	<i>Tần suất</i>
Buýt-phông	4040	2048	0,5080
Piéc-xơn	12000	6019	0,5016
Piéc-xơn	24000	12012	0,5005

Một thí dụ khác: có thể cho rằng xác suất phân rã của một nguyên tử Ra^{226} sau 100 năm là 0,04184 (với độ chính xác tới 5 chữ số sau dấu phẩy); ở đây số lượng nguyên tử tham gia thí nghiệm rất lớn (cỡ $10^{23} - 10^{24}$).

Có thể kiểm tra được rằng xác suất định nghĩa theo thống kê thỏa mãn các tính chất trình bày ở mục trước. Chú ý là trong định nghĩa phải có điều kiện các phép thử lặp lại như nhau, điều này trên thực tế không dễ bảo đảm nên tần suất có thể phụ thuộc vào thời gian. Mặc dù vậy phương pháp xác định xác suất theo tần suất có phạm vi ứng dụng rất lớn trong nhiều ngành khoa học và kỹ thuật. Mặt khác, điểm xuất phát để xây dựng lý thuyết xác suất như là một khoa học cũng chính là việc quan sát tính ổn định thống kê của các tần suất của vô vàn các hiện tượng thực tế. Từ đó dễ hiểu vì sao có thể định nghĩa lý thuyết xác suất như là một khoa học nghiên cứu các mô hình toán học của các hiện tượng ngẫu nhiên có tần suất ổn định.

2.3. Định nghĩa tiên đề

Các định nghĩa cổ điển và thống kê của xác suất có nhiều hạn chế để xây dựng một lý thuyết tổng quát. Khái niệm cổ điển không dùng được trong trường hợp không thể xây dựng một hệ thống đầy đủ các sự kiện đồng khả năng. Trong khi đó, tần suất chỉ là một giá trị xấp xỉ để đánh giá xác suất, chưa kể đòi hỏi là số quan sát phải rất lớn và giá trị tần suất tìm được phải lớn hơn nhiều sai số đo và cả sai số tính toán.

Chúng ta bắt đầu từ hệ thống các tiên đề dưới dạng do Kôn-mô-gô-rốp phát biểu. Các tiên đề đó (giống như các tiên đề toán học khác) được thừa nhận là đúng đắn, tất nhiên căn cứ vào kinh nghiệm cuộc sống và hoạt động thực tiễn. Cách tiếp cận này liên hệ chặt chẽ lý thuyết xác suất với lý thuyết hàm số và tập hợp. Cách xác định xác suất theo tiên đề sẽ chứa

trong nó các định nghĩa cổ điển và thống kê của xác suất như là các trường hợp riêng.

Ta quay trở lại không gian các sự kiện sơ cấp Ω (xem §1), còn bản thân các phần tử là gì không quan trọng. Tiếp theo xác định hệ thống $\mathcal{A}$ các tập hợp con của Ω , các phần tử của $\mathcal{A}$ được gọi là các sự kiện ngẫu nhiên. Ta đặt cho $\mathcal{A}$ các yêu cầu hợp lý sau:

(i) $\mathcal{A}$ chứa Ω .

(ii) Nếu A và $B \in \mathcal{A}$ thì $\bar{A}, \bar{B}, A + B, AB \in \mathcal{A}$.

Hệ thống $\mathcal{A}$ thỏa mãn các điều kiện trên được gọi là *đại số Bun*. Nếu ta yêu cầu thêm

(iii) Nếu $A_1, A_2, \dots, A_n, \dots$ là các phần tử của $\mathcal{A}$, thì tổng và tích vô hạn $A_1 + A_2 + \dots + A_n + \dots, A_1 A_2 \dots A_n \dots$ cũng thuộc $\mathcal{A}$. Nếu $\mathcal{A}$ thỏa mãn thêm điều kiện (iii) ta có một *trường Bô-ren*, hay σ -*đại số*.

Bây giờ ta đã có thể định nghĩa xác suất:

Định nghĩa. Ta gọi xác suất trên $(\Omega, \mathcal{A})$ là một hàm số xác định trên $\mathcal{A}$ có giá trị trong $[0; 1]$ và thỏa mãn 3 tiên đề

$$(T_1) P(\Omega) = 1;$$

$$(T_2) P(A + B) = P(A) + P(B) \text{ (} A, B \text{ xung khắc);}$$

$$(T_3) \text{ Nếu dãy } \{A_n\} \text{ có tính chất } A_j \supseteq A_i, \forall i \leq j \text{ và}$$

$$A_1 A_2 \dots A_n \dots = V, \text{ thì } P(A_n) \xrightarrow{n \rightarrow \infty} 0.$$

Xuất phát từ hệ tiên đề trên có thể chứng minh được các tính chất của xác suất đã trình bày ở §1, hoặc chính chúng đã là các tính chất đó (tiên đề 1 và 2). Chú ý rằng hệ tiên đề này chưa đầy đủ: ứng với một tập Ω có thể chọn xác suất theo nhiều cách khác nhau. Người ta có thể thay tiên đề 2 và 3 bằng một tiên đề có tên là *tiên đề cộng mở rộng*:

(T₄) Nếu dãy $\{A_n\}$ có tính chất xung khắc từng đôi và $A = \sum_{n=1}^{\infty} A_n \in \mathcal{A}$ thì

$$P(A) = P(A_1) + P(A_2) + \dots P(A_n) + \dots = \sum_{n=1}^{\infty} P(A_n).$$

Để kết luận, có thể nói rằng cách định nghĩa xác suất ở đây nhìn từ quan điểm của lý thuyết tập hợp chính là sự đưa vào cùng với Ω một độ đo không âm, trực chuẩn, cộng tính, xác định cho mọi phần tử của tập $\mathcal{A}$. Như vậy khi định nghĩa xác suất chúng ta phải có không chỉ tập Ω các sự kiện sơ cấp ban đầu, mà còn phải có tập các sự kiện ngẫu nhiên $\mathcal{A}$ và hàm số P xác định trên đó. Tổ hợp $\{\Omega, \mathcal{A}, P\}$ sau này thường được gọi là *không gian xác suất*.

§3. XÁC SUẤT CÓ ĐIỀU KIỆN

3.1. Khái niệm

Thực ra mọi xác suất $P(A)$ đều là có điều kiện, vì sự kiện A xảy ra khi thực hiện một bộ điều kiện xác định. Tuy nhiên, nếu ngoài bộ điều kiện đó ra còn có thêm điều kiện khác thể hiện bằng việc xuất hiện B nào đó, thì người ta đưa ra một khái niệm mới: *xác suất có điều kiện của A biết rằng đã xảy ra B* , ký hiệu là $P(A|B)$. Bằng trực giác ta cũng thấy rằng khi có B với $P(B) > 0$ thì nói chung “khả năng” xuất hiện A cũng thay đổi; đặc biệt nếu $AB = V$ khả năng đó triệt tiêu, còn nếu $B \Rightarrow A$ thì khả năng trở thành tất yếu. Vậy là, với điều kiện đã có B , người ta xác định một cách tự nhiên khả năng xuất hiện A nào đó bằng một số tỷ lệ với $P(AB)$, tức là số có dạng $kP(AB)$, $k > 0$. Để xác định hằng số k đó, do $P(A|B) = kP(AB)$ là một xác suất và ta chọn $A = B$, $P(B|B) = 1$, nên $kP(B) = 1$. Từ đó

$$k = \frac{1}{P(B)}.$$

Định nghĩa 1. Giả sử trong một phép thử ta có $P(B) > 0$. Khi đó xác suất có điều kiện của sự kiện A nào đó, biết rằng đã có B , sẽ là một số không âm, ký hiệu là:

$$P(A|B) = \frac{P(AB)}{P(B)}. \quad (3.1)$$

Để ý rằng nói chung $P(A) \neq P(A|B)$. Ngoài ra xác suất có điều kiện có mọi tính chất của một xác suất bình thường.

Thí dụ 3.1. Gieo 2 con xúc sắc giống nhau. Tính xác suất để ta có tổng số chấm thu được bằng 6, biết rằng tổng đó là một số chẵn.

Giải. Ta đã biết $P(A) = 5/36$ (xem thí dụ 2.2, A là sự kiện xuất hiện tổng chấm bằng 6). Nếu ký hiệu B là sự kiện xuất hiện tổng chấm chẵn, thì điều kiện để tính $P(A|B)$ đã thay đổi, tổng số chẵn chỉ tương ứng với 18 kết cục của phép thử gieo 2 con xúc sắc. Từ đó $P(A|B) = 5/18$.

Thí dụ 3.2. Rút từ bộ bài tú lơ khơ 52 con lần lượt ra 2 con bài. Tìm xác suất để con thứ hai là át, biết rằng con thứ nhất đã là át.

Giải. Dễ thấy nếu ký hiệu A_i là sự kiện con thứ i là át ($i = 1, 2$), thì $P(A_2|A_1) = \frac{3}{51} = \frac{1}{17}$, tương đương với việc do đã có A_1 , việc tính xác suất sự kiện A_2 đưa về tính trong trường hợp chỉ còn 51 con bài với 3 con át trong đó.

Định nghĩa 2. Ta nói rằng A và B *độc lập* (thống kê), nếu

$$P(A|B) = P(A) \text{ hoặc } P(B|A) = P(B). \quad (3.2)$$

Như vậy nếu A, B độc lập việc xuất hiện sự kiện này không làm thay đổi xác suất của sự kiện kia. Tuy nhiên việc kiểm tra tính chất (3.2) trong thực tiễn rất khó khăn và trong nhiều

trường hợp là không thể. Vì vậy dựa vào thực tế và trực giác mà ta thừa nhận các sự kiện độc lập trong các bài tập sau này. Công thức tương đương của (3.2), có để ý đến (3.1) là:

$$P(AB) = P(A)P(B). \quad (3.3)$$

Định nghĩa 3. Ta nói bộ sự kiện $A_1, A_2, \dots, A_n$ *độc lập* (hay *độc lập trong tổng thể*) nếu

$$P(A_{i_1} A_{i_2} \dots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_k}) \quad (3.4)$$

với mọi dãy $(i_1, \dots, i_k)$ gồm các số nguyên khác nhau lấy từ $\{1, 2, \dots, n\}$.

Thí dụ 3.3. Gieo hai lần một đồng tiền và ta có 4 kết cục đồng khả năng (S – ký hiệu mặt sấp, N – mặt ngửa)

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

Rõ ràng các sự kiện $A = SS + SN$, $B = SS + NS$, $C = SS + NN$ là độc lập từng đôi do $P(A) = P(B) = P(C) = \frac{1}{2}$; còn $P(AB) =$

$P(AC) = P(BC) = \frac{1}{4}$ thỏa mãn (3.3). Tuy nhiên chúng không độc lập trong tổng thể do

$$P(ABC) = \frac{1}{4} \neq P(A)P(B)P(C) = \frac{1}{8}.$$

Như vậy không nên nhầm lẫn hai khái niệm độc lập trong các định nghĩa 2 và 3. Khái niệm độc lập trong tổng thể kéo theo độc lập từng đôi (do (3.3) là trường hợp riêng của (3.4) khi $k = 2$), nhưng ngược lại nói chung không đúng.

3.2. Công thức cộng và nhân xác suất

1. Công thức nhân xác suất

$$P(AB) = P(A)P(B|A) = P(B)P(A|B). \quad (3.5)$$

Đó là hệ quả trực tiếp suy ra từ (3.1). Từ (3.5) có thể dẫn ra các kết quả quan trọng:

(i) Nếu A, B độc lập thì $P(AB) = P(A)P(B)$ (xem 3.3)).

(ii) Mở rộng cho tích n sự kiện

$$\begin{aligned} P(A_1 A_2 \dots A_n) &= \\ &= P(A_1)P(A_2 | A_1)P(A_3 | A_1 A_2) \dots P(A_n | A_1 A_2 \dots A_{n-1}). \end{aligned} \quad (3.6)$$

(iii) Nếu $A_1, A_2, \dots, A_n$ độc lập trong tổng thể, thì:

$$P\left(\prod_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i).$$

2. Công thức cộng xác suất

$$P(A + B) = P(A) + P(B) - P(AB). \quad (3.7)$$

Việc chứng minh công thức trên không có gì quá phức tạp (nhất là từ các tiên đề của mục 2.3). Từ (3.7) có thể dẫn ra các kết quả sau:

(i) Nếu A, B xung khắc, thì $P(A + B) = P(A) + P(B)$.

(ii) Mở rộng cho tổng n sự kiện

$$\begin{aligned} P\left(\sum_{i=1}^n A_i\right) &= \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i A_j) + \sum_{i < j < k} P(A_i A_j A_k) - \dots \\ &\quad + (-1)^{n-1} P(A_1 A_2 \dots A_n). \end{aligned} \quad (3.8)$$

(iii) Nếu $A_1, A_2, \dots, A_n$ xung khắc từng đôi

$$P\left(\sum_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i).$$

Các công thức (3.5) – (3.8) cho ta các công cụ hiệu quả để tính xác suất các sự kiện phức tạp qua xác suất các sự kiện đơn giản hơn.

Thí dụ 3.4. Hai cọc bài được lấy từ một bộ bài tứ lơ khơ, cọc thứ nhất gồm 4 con át, cọc thứ hai gồm 4 con ka. Rút ngẫu nhiên từ mỗi cọc bài ra một con bài, tính các xác suất để

- a) cả 2 con là con cơ,
- b) có ít nhất 1 con cơ.

Cũng câu hỏi như vậy nhưng thay điều kiện đầu bài: trộn cộc bài và rút hú họa từ đó ra 2 con bài.

Giải. Gọi A – con bài thứ nhất là cơ, B – con bài thứ hai là cơ. Để ý rằng thuật ngữ “thứ nhất”... chỉ để phân biệt hai con bài chứ không để chỉ thứ tự nào cả. Trong trường hợp hai cộc bài riêng rẽ, dễ thấy A và B độc lập. Từ đó

- a) Xác suất cần tìm là $P(AB)$, để ý đến (3.3) ta có:

$$P(AB) = P(A)P(B) = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}.$$

- b) Sự kiện ta quan tâm là $A + B$, theo (3.7):

$$P(A + B) = P(A) + P(B) - P(AB) = \frac{1}{4} + \frac{1}{4} - \frac{1}{16} = \frac{7}{16}.$$

Trường hợp trộn lẫn hai cộc bài thành một thì A, B không còn độc lập nữa. Tuy nhiên các xác suất $P(A)$ và $P(B)$ đều bằng $2/8 = 1/4$ do vai trò hai quân bài như nhau. Từ đó:

- a) Dùng công thức (3.5):

$$P(AB) = P(A)P(B|A) = \frac{1}{4} \cdot \frac{1}{7} = \frac{1}{28}.$$

- b) Một lần nữa theo (3.7):

$$P(A + B) = P(A) + P(B) - P(AB) = \frac{1}{4} + \frac{1}{4} - \frac{1}{28} = \frac{13}{28}.$$

Thí dụ 3.5. Ba xạ thủ mỗi người bắn một viên đạn với xác suất bắn trúng của từng người tương ứng là 0,7; 0,8 và 0,9. Tính các xác suất:

- a) có hai người bắn trúng,
- b) có ít nhất một người bắn trượt.

Giải. Gọi A_i là sự kiện xạ thủ thứ i bắn trúng ($i = 1, 2, 3$) và $P(A_1) = 0,7; P(A_2) = 0,8; P(A_3) = 0,9$.

a) Nếu gọi A là sự kiện có đúng 2 người bắn trúng thì:

$$A = A_1 A_2 \overline{A_3} + A_1 \overline{A_2} A_3 + \overline{A_1} A_2 A_3.$$

Dùng tính xung khắc của các số hạng và tính độc lập của các A_i và $\overline{A_j}$ ($j \neq i$), ta có:

$$\begin{aligned} P(A) &= P(A_1 A_2 \overline{A_3}) + P(A_1 \overline{A_2} A_3) + P(\overline{A_1} A_2 A_3) \\ &= P(A_1)P(A_2)P(\overline{A_3}) + P(A_1)P(\overline{A_2})P(A_3) + P(\overline{A_1})P(A_2)P(A_3) \\ &= 0,7.0,8.(1 - 0,9) + 0,7.(1 - 0,8).0,9 + (1 - 0,7).0,8.0,9 \\ &= 0,398. \end{aligned}$$

b) Nếu gọi B là sự kiện có ít nhất một người bắn trượt, thì $\overline{B}$ là sự kiện không có ai bắn trượt hay cả ba đều bắn trúng. Rõ ràng việc tính $P(\overline{B})$ dễ dàng hơn nhiều so với tính $P(B)$ theo cách trực tiếp, từ đó

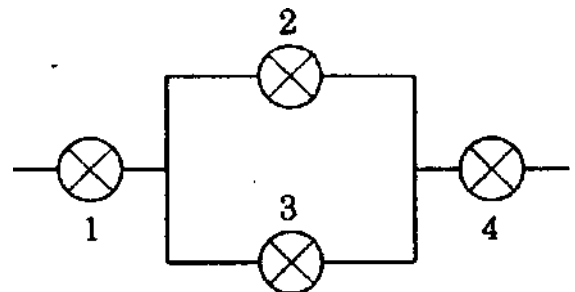
$$\begin{aligned} P(B) &= 1 - P(\overline{B}) = 1 - P(A_1 A_2 A_3) \\ &= 1 - 0,7.0,8.0,9 = 0,496. \end{aligned}$$

Thí dụ 3.6. Cho một mạch điện gồm 4 linh kiện như hình 1.3, trong đó xác suất hỏng của từng linh kiện trong một khoảng thời gian nào đó tương ứng là 0,2; 0,1; 0,05 và 0,02. Tìm xác suất để mạng hoạt động tốt trong khoảng thời gian đó, với giả thiết là các linh kiện làm việc độc lập với nhau và các dây luôn tốt.

Giải. Gọi A_i là sự kiện linh kiện thứ i làm việc tốt ($i = 1, 4$). Sử dụng các tính chất của mạng song song và nối tiếp, gọi A là sự kiện mạng hoạt động tốt, khi đó $A = A_1(A_2 + A_3)A_4$.

Để ý rằng từ giả thiết đầu bài ta luôn có A_1 , A_4 và $A_2 + A_3$ độc lập, nên:

$$P(A) = P(A_1)P(A_2 + A_3)P(A_4). \quad (3.9)$$



Hình 1.3

Ta cần tính $P(A_2 + A_3)$, và do A_2, A_3 không xung khắc, nên

$$P(A_2 + A_3) = P(A_2) + P(A_3) - P(A_2 A_3).$$

Thay vào (3.9), để ý rằng $P(A_2 A_3) = P(A_2)P(A_3)$ và giả thiết của đầu bài

$$\begin{aligned} P(A) &= P(A_1)[P(A_2) + P(A_3) - P(A_2)P(A_3)]P(A_4) \\ &= 0,8.(0,9 + 0,95 - 0,9.0,95).0,98 \\ &= 0,78008. \end{aligned}$$

Chú ý rằng nếu ta khai triển $A = A_1 A_2 A_4 + A_1 A_3 A_4$ sau đó dùng các công thức (3.6) – (3.7) để tính $P(A)$ thì sẽ phức tạp hơn một chút, bạn đọc hãy tự giải theo cách này.

Thí dụ 3.7. Một gia đình có 6 con. Tìm xác suất để gia đình đó có số con trai nhiều hơn số con gái.

Giải. Ta chấp nhận xác suất sinh con trai bằng xác suất sinh con gái và bằng 0,5, ngoài ra kết quả mỗi lần sinh được coi là độc lập với nhau. Gọi A là sự kiện số con trai nhiều hơn con gái, khi đó việc tính trực tiếp $P(A)$ đưa về xác định các trường hợp: hoặc 6 trai, hoặc 5 trai 1 gái, hoặc 4 trai 2 gái. Tuy nhiên có thể dùng cách khác. Gọi B là sự kiện số gái nhiều hơn trai, còn C là sự kiện số trai và số gái như nhau. Dễ thấy

$$A + B + C = U \text{ và } P(A) + P(B) + P(C) = 1.$$

Do tính đối xứng của việc sinh con trai và con gái, nên $P(A) = P(B)$, từ đó:

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

và ta cần phải tính $P(C)$ – xác suất để trong gia đình có 3 con trai, 3 con gái. Một trường hợp như vậy có xác suất $\frac{1}{2^6}$ và có tất

cả $C_6^3 = 20$ khả năng khác nhau, từ đó $P(C) = 20/64 = \frac{5}{16}$ và

$$P(A) = \frac{1 - \frac{5}{16}}{2} = \frac{11}{32}.$$

Thí dụ 3.8. Một người viết n lá thư cho n người khác nhau, bỏ ngẫu nhiên vào n phong bì đã có sẵn địa chỉ. Tìm xác suất để có ít nhất một lá thư bỏ vào đúng phong bì.

Giải. Gọi A_i là sự kiện là thư thứ i bỏ đúng phong bì ($i = \overline{1, n}$), A – là sự kiện cần tìm xác suất, ta có $A = A_1 + A_2 + \dots + A_n$. Do các A_i không xung khắc, nên ta dùng công thức (3.8). Dễ thấy

$$P(A_i) = \frac{1}{n} = \frac{(n-1)!}{n!};$$

$$P(A_i A_j) = P(A_i)P(A_j|A_i) = \frac{1}{n} \cdot \frac{1}{n-1} = \frac{(n-2)!}{n!};$$

$$P(A_i A_j A_k) = P(A_i)P(A_j|A_i)P(A_k|A_i A_j) = \frac{(n-3)!}{n!};$$

.....

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2|A_1) \dots P(A_n|A_1 A_2 \dots A_{n-1}) = \frac{1}{n!}.$$

Từ đó theo (3.8)

$$\begin{aligned} P(A) &= \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i A_j) + \sum_{i < j < k} P(A_i A_j A_k) - \dots \\ &\quad + (-1)^{n-1} P(A_1 A_2 \dots A_n) \\ &= C_n^1 \frac{(n-1)!}{n!} - C_n^2 \frac{(n-2)!}{n!} + C_n^3 \frac{(n-3)!}{n!} - \dots + (-1)^{n-1} \frac{1}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n-1} \frac{1}{n!}. \end{aligned}$$

Khi n khá lớn xác suất cần tìm $\approx 1 - \frac{1}{e}$.

Thí dụ 3.9. Tìm xác suất để xuất hiện ít nhất 1 lần 2 mặt chấm khi gieo n lần 2 con xúc sắc.

Giải. Xác suất để trong 1 lần gieo 2 con xúc sắc ta có hai mặt 6 chấm sẽ là $\frac{1}{36}$, và không có hai mặt 6 chấm sẽ là $1 - \frac{1}{36}$. Nếu đặt A là sự kiện cần tìm, rõ ràng $\bar{A}$ là sự kiện gieo n lần 2 con xúc sắc mà không lần nào có 2 mặt 6 chấm. Từ đó

$$P(\bar{A}) = \left(1 - \frac{1}{36}\right)^n \text{ và } P(A) = 1 - \left(\frac{35}{36}\right)^n.$$

3.3. Công thức Béc-nu-li

Xét một dãy n phép thử độc lập giống nhau, trong mỗi phép thử chỉ có hai kết cục hoặc xảy ra A hoặc không và $P(A) = p$, $P(\bar{A}) = 1 - p = q$ không phụ thuộc vào số thứ tự của phép thử. Những bài toán thỏa mãn các yêu cầu trên được gọi là tuân theo *lược đồ Béc-nu-li* và hay gặp trong nhiều lĩnh vực ứng dụng.

Ta quan tâm đến xác suất để trong dãy n phép thử độc lập nói trên sự kiện A xuất hiện đúng k lần, ký hiệu là $P_n(k)$. Gọi B là sự kiện “trong dãy n phép thử Béc-nu-li sự kiện A xuất hiện đúng k lần”, ta thấy B có thể xảy ra theo nhiều phương án khác nhau, miễn sao trong dãy các kết cục của n phép thử sự kiện A có mặt đúng k lần. Rõ ràng B sẽ là tổng của C_n^k các phương án như vậy. Còn xác suất để xảy ra một phương án, do trong dãy n phép thử độc lập sự kiện A xuất hiện đúng k lần, $\bar{A}$ xuất hiện $n - k$ lần, nên sẽ bằng $p^k q^{n-k}$. Từ đó ta có *công thức Béc-nu-li*

$$P(B) = P_n(k) = C_n^k p^k q^{n-k}, \quad k = 0, 1, \dots, n. \quad (3.10)$$

Việc sử dụng công thức (3.10) sẽ đơn giản hơn nhiều việc dùng các công thức (3.5) – (3.8) và vì vậy nó có ý nghĩa thực tiễn rất lớn.

Thí dụ 3.10. Một thiết bị có 10 chi tiết đối với độ tin cậy (xác suất làm việc tốt trong một khoảng thời gian nào đó) của

mỗi chi tiết là 0,9. Tìm xác suất để trong khoảng thời gian ấy có đúng 2 chi tiết làm việc tốt.

Giải. Rõ ràng ta có lược đồ Béc-nu-li, với $n = 10$, $p = 0,9$ và $k = 2$, áp dụng (3.10) ta có xác suất cần tìm là:

$$P_{10}(2) = C_{10}^2 \cdot (0,9)^2 \cdot (0,1)^8 = 3645 \cdot 10^{-10}.$$

Thí dụ 3.11. Một bác sỹ có xác suất chữa khỏi bệnh là 0,8. Có người nói rằng cứ 10 người đến chữa thì có chắc chắn 8 người khỏi bệnh; điều đó có đúng không?

Giải. Câu khẳng định là sai. Ở đây có thể coi việc chữa bệnh cho 10 người là dãy 10 phép thử, trong đó A là sự kiện được chữa khỏi bệnh có $P(A) = 0,8$. Từ đó xác suất để trong 10 bệnh nhân đến chữa có 8 người khỏi là:

$$P_{10}(8) = C_{10}^8 \cdot 0,8^8 \cdot 0,2^2 \approx 0,3108.$$

Thí dụ 3.12. Tỷ lệ phế phẩm của một lô hàng là 1%. Hỏi cỡ mẫu cần chọn ra là bao nhiêu (có hoàn lại) sao cho trong mẫu có ít nhất 1 phế phẩm với xác suất lớn hơn 0,95?

Giải. Giả sử mẫu chọn ra có kích cỡ là n và việc chọn ra một sản phẩm có hoàn lại là một phép thử Béc-nu-li với $p = 0,01$. Rõ ràng xác suất để trong mẫu có ít nhất 1 phế phẩm sẽ là:

$$1 - (1 - p)^n = 1 - 0,99^n.$$

Theo yêu cầu của đầu bài

$$1 - 0,99^n > 0,95 \Leftrightarrow 0,05 > 0,99^n$$

$$\Rightarrow n > \frac{\log 0,05}{\log 0,99} \approx 296.$$

Nhiều khi ta muốn tìm xác suất để trong dãy n phép thử Béc-nu-li sự kiện A xuất hiện với số lần từ k_1 đến k_2 ; để thấy xác suất cần tìm, ký hiệu là $P_n(k_1, k_2)$, sẽ là:

$$P_n(k_1; k_2) = \sum_{k=k_1}^{k_2} P_n(k) = \sum_{k=k_1}^{k_2} C_n^k p^k q^{n-k}. \quad (3.11)$$

Ta có nhận xét rằng khi n và k khá lớn, việc tính toán xác suất theo (3.10) và (3.11) rất cồng kềnh và khó khăn; vì vậy người ta tìm cách tính gần đúng các xác suất đó. Có thể sử dụng các cách xấp xỉ sau đây:

(i) Nếu n rất lớn, trong khi p rất nhỏ, xác suất theo công thức (3.10) có thể xấp xỉ bằng (xấp xỉ Poa-xông)

$$P_n(k) \approx \frac{(np)^k}{k!} e^{-np}. \quad (3.12)$$

(ii) Nếu n lớn, nhưng p không quá bé và quá lớn, ta có xấp xỉ chuẩn (định lý giới hạn địa phương Moa-vơ – Láp-la-xơ)

$$P_n(k) \approx \frac{\varphi(x_k)}{\sqrt{npq}}, \quad x_k = \frac{k - np}{\sqrt{npq}}, \quad (3.13)$$

trong đó $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ là hàm Gao-xơ (xem bảng 1).

(iii) Nếu n lớn, nhưng p không quá bé hoặc quá lớn thì xác suất trong (3.11) có thể xấp xỉ bằng (định lý giới hạn tích phân Moa-vơ – Láp-la-xơ)

$$P_n(k_1; k_2) \approx \phi(x_2) - \phi(x_1), \quad x_j = \frac{k_j - np}{\sqrt{npq}}, \quad j = 1, 2, \quad (3.14)$$

và trong đó $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ là hàm Láp-la-xơ (xem bảng 2).

Thí dụ 3.13. Xác suất sản xuất ra phế phẩm của một máy là 0,005. Tìm xác suất để trong 800 sản phẩm của máy đó có đúng 3 phế phẩm.

Giải. Rõ ràng có thể dùng xấp xỉ Poa-xông theo (3.12), với $np = 4$

$$P_{800}(3) \approx \frac{e^{-4} \cdot 4^3}{3!} = 0,1954.$$

Thí dụ 3.14. Xác suất ném trúng rổ của một cầu thủ là 0,8. Tìm xác suất để trong 100 lần cầu thủ đó:

- a) ném trúng 75 lần;
- b) ném trúng không ít hơn 75 lần.

Giải. Việc tính theo công thức (3.10) hoặc (3.11) của lược đồ Béc-nu-li sẽ khá phức tạp. Ta sẽ tính xấp xỉ theo (3.13) và (3.14):

$$a) P_{100}(75) \approx \frac{\varphi\left(\frac{75 - 0,8 \cdot 100}{\sqrt{100 \cdot 0,8 \cdot 0,2}}\right)}{\sqrt{100 \cdot 0,8 \cdot 0,2}} = \frac{\varphi(-1,25)}{4} = 0,04565.$$

$$b) P_{100}(75; 100) \approx \phi(5) + \phi(1,25) = 0,8943.$$

§4. CÔNG THỨC BAY-ÉT

4.1. Khái niệm nhóm đầy đủ

Định nghĩa. Nhóm các sự kiện $A_1, A_2, \dots, A_n$ ($n \geq 2$) của một phép thử được gọi là (hay tạo thành) một *nhóm đầy đủ*, nếu

- (i) $A_i A_j = V, \forall i \neq j$ (xung khắc từng đôi),
- (ii) $A_1 + A_2 + \dots + A_n = U$.

Theo định nghĩa này ở phép thử đang xét chỉ có thể xuất hiện một sự kiện trong số n sự kiện $A_1, \dots, A_n$ (và phải có một sự kiện). Nhóm $A_1, \dots, A_n$ có các tính chất trên còn được gọi là một hệ thống đầy đủ.

Thí dụ 4.1. Xét phép thử gieo một con xúc sắc. Nếu ký hiệu A_i là sự kiện xuất hiện mặt i chấm ($i = \overline{1,6}$), ta có một nhóm đầy đủ $\{A_i, i = \overline{1,6}\}$. Có thể tạo thành nhiều nhóm đầy đủ khác cho phép thử này, chẳng hạn đặt $A = A_6$, từ đó $\overline{A} = A_1 + A_2 + \dots + A_5 = \overline{A_6}$ và nhóm $\{A, \overline{A}\}$ chính là một nhóm đầy đủ.

Như vậy dễ thấy tập hợp tất cả các sự kiện sơ cấp tạo nên một nhóm đầy đủ. Tổng quát hơn tập các sự kiện tạo nên một phân hoạch của không gian Ω các sự kiện sơ cấp cũng là một nhóm đầy đủ. Tập $\{A, \bar{A}\}$, với A là sự kiện tùy ý là nhóm đầy đủ bé nhất (chỉ có 2 phần tử). Để ý $\{U, V\}$ cũng tạo nên một nhóm đầy đủ và được gọi là nhóm đầy đủ *tầm thường*.

4.2. Công thức xác suất đầy đủ

Giả sử ta có một nhóm đầy đủ các sự kiện $A_1, A_2, \dots, A_n$ và đồng thời xét một sự kiện H nào đó. Nếu đã biết các $P(A_i)$ và $P(H|A_i)$, ta có thể tính được $P(H)$. Rõ ràng từ giả thiết về nhóm đầy đủ:

$$H = A_1H + A_2H + \dots + A_nH = \sum_{i=1}^n A_iH.$$

Từ đó $P(H) = P\left(\sum_{i=1}^n A_iH\right) = \sum_{i=1}^n P(A_iH)$ (do xung khắc),

và áp dụng công thức nhân (3.5):

$$P(H) = \sum_{i=1}^n P(A_i)P(H|A_i). \quad (4.1)$$

Công thức (4.1) có tên gọi là *công thức xác suất đầy đủ* (hay *xác suất toàn phần*).

Thí dụ 4.2. Một phân xưởng có 3 máy sản xuất cùng loại sản phẩm với tỷ lệ phế phẩm tương ứng 1%; 0,5% và 0,2%. Biết rằng máy I sản xuất ra 35%, máy II – 45% và máy III – 20% sản phẩm. Chọn hủ họa ra một sản phẩm, tìm xác suất đó là phế phẩm.

Giải. Đặt M_1, M_2 và M_3 tương ứng là sự kiện sản phẩm chọn ra do máy I, II và III sản xuất. Dễ thấy $\{M_i, i = \overline{1,3}\}$ tạo nên một nhóm đầy đủ và $P(M_1) = 0,35$; $P(M_2) = 0,45$; $P(M_3) = 0,20$. Gọi H sự kiện rút được phế phẩm, áp dụng (4.1) để ý rằng $P(H|M_1) = 1\%$; $P(H|M_2) = 0,5\%$; $P(H|M_3) = 0,2\%$, ta có

$$\begin{aligned}
 P(H) &= \sum_{i=1}^3 P(M_i)P(H|M_i) = \\
 &= 0,35.1\% + 0,45.0,5\% + 0,20.0,2\% = 0,615\%.
 \end{aligned}$$

Ý nghĩa của xác suất này là tỷ lệ phế phẩm của phân xưởng.

Thí dụ 4.3. Có hai hộp áo, hộp I có 10 áo trong đó có 1 phế phẩm, hộp II có 8 áo trong đó có 2 phế phẩm. Lấy hù họa 1 áo từ hộp I bỏ sang hộp II, sau đó từ hộp này chọn hù họa ra 2 áo. Tìm xác suất để cả 2 áo đó đều là phế phẩm.

Giải. Ta lập nhóm đầy đủ để làm rõ thông tin về chất lượng chiếc áo mang từ hộp I sang; gọi A – áo đó là phế phẩm, $\bar{A}$ – áo tốt. Đặt H là sự kiện 2 áo cuối chọn ra đều là phế phẩm. Rõ ràng $P(A) = \frac{1}{10}$; $P(\bar{A}) = \frac{9}{10}$; ta còn cần tính $P(H|A)$

và $P(H|\bar{A})$. Dùng định nghĩa xác suất:

$$P(H|A) = \frac{C_3^2}{C_9^2} = \frac{3.2}{9.8} = \frac{1}{12};$$

$$P(H|\bar{A}) = \frac{1}{C_9^2} = \frac{2}{9.8} = \frac{1}{36}.$$

Từ đó dùng (4.1)

$$P(H) = P(A)P(H|A) + P(\bar{A})P(H|\bar{A}) = \frac{1}{10} \cdot \frac{1}{12} + \frac{9}{10} \cdot \frac{1}{36} = \frac{1}{30}.$$

4.3. Công thức Bay-ét

Giả sử ta có một nhóm đầy đủ $A_1, A_2, \dots, A_n$, sau đó có thêm sự kiện H nào đó. Đôi khi ta muốn xác định xác suất $P(A_i|H)$, i là một số nào đó trong $\{1, 2, \dots, n\}$. Theo công thức nhân (3.5) ta có

$$P(A_i H) = P(A_i)P(H|A_i) = P(H)P(A_i|H).$$

$$\text{Từ đó} \quad P(A_i|H) = \frac{P(A_i)P(H|A_i)}{P(H)} \quad (4.2)$$

và thay (4.1) vào (4.2)

$$P(A_i|H) = \frac{P(A_i)P(H|A_i)}{\sum_{i=1}^n P(A_i)P(H|A_i)}. \quad (4.3)$$

Công thức (4.3) có tên gọi là *công thức Bay-ét*. Các xác suất $P(A_i)$, $i = \overline{1, n}$, đã được xác định từ trước, thường được gọi là xác suất *tiên nghiệm*; còn các xác suất $P(A_i|H)$, $i = \overline{1, n}$, được xác định sau khi đã có kết quả thí nghiệm nào đó thể hiện qua sự xuất hiện của H , thường được gọi là xác suất *hậu nghiệm*. Như vậy công thức Bay-ét cho phép đánh giá lại xác suất xảy ra các A_i sau khi đã có thêm thông tin về H . Cần phải nhấn mạnh rằng nếu muốn dùng các công thức (4.1) hoặc (4.3), nhất thiết phải có nhóm đầy đủ. Ngoài ra nếu (4.1) cho ta xác suất không có điều kiện, thì (4.3) cho phép tính xác suất có điều kiện, trong đó sự kiện A_i cần tính xác suất phải là một thành viên của nhóm đầy đủ đang xét. Từ đó thấy rằng việc dùng công thức Bay-ét để tính xác suất có điều kiện đã gợi ý cho ta cách chọn nhóm đầy đủ sao cho sự kiện quan tâm phải là thành viên. Trong trường hợp không có (hoặc rất khó xác định) nhóm đầy đủ, nên dùng công thức (4.2), trong trường hợp này việc tính $P(H)$ sẽ khó hơn là dùng công thức (4.1).

Thí dụ 4.4. Một mạch điện gồm 2 bộ phận mắc nối tiếp, với xác suất làm việc tốt trong một khoảng thời gian nào đó của mỗi bộ phận là 0,95 và 0,98. Ở một thời điểm trong khoảng thời gian trên người ta thấy mạch điện ngừng làm việc (do bộ phận nào đó hỏng); tìm xác suất để chỉ bộ phận thứ hai hỏng.

Giải. Do hai bộ phận mắc nối tiếp nên chỉ cần một bộ phận hỏng là mạch ngừng làm việc. Gọi A_i ($i = 1, 2$) là sự kiện bộ phận thứ i tốt; khi đó có thể xảy ra 4 khả năng khác nhau:

B_0 – cả hai bộ phận đều tốt; B_1 – bộ phận I tốt, II hỏng; B_2 – bộ phận II tốt, I hỏng; B_3 – cả hai bộ phận đều hỏng. Dễ thấy các $B_i, i = \overline{0,3}$, tạo nên một nhóm đầy đủ và do tính độc lập

$$P(B_0) = P(A_1 A_2) = 0,95.0,98 = 0,931;$$

$$P(B_1) = P(A_1 \overline{A_2}) = 0,95.0,02 = 0,019;$$

$$P(B_2) = P(\overline{A_1} A_2) = 0,05.0,98 = 0,049;$$

$$P(B_3) = P(\overline{A_1} \overline{A_2}) = 0,05.0,02 = 0,001.$$

Gọi H – sự kiện mạch không làm việc, ta có:

$$P(H|B_0) = 0; P(H|B_1) = P(H|B_2) = P(H|B_3) = 1.$$

Từ đó theo công thức Bay-ét (4.3):

$$\begin{aligned} P(B_1|H) &= \frac{P(B_1)P(H|B_1)}{\sum_{i=0}^3 P(B_i)P(H|B_i)} = \frac{0,019}{0,019 + 0,049 + 0,001} \\ &= \frac{19}{69}. \end{aligned}$$

Để ý rằng ta có thể dùng (4.2) để tính $P(B_1|H)$. Để làm điều đó ta viết:

$$H = A_1 \overline{A_2} + \overline{A_1} A_2 + \overline{A_1} \overline{A_2}.$$

Do tính xung khắc và độc lập của các sự kiện tương ứng ta có $P(H) = P(A_1)P(\overline{A_2}) + P(\overline{A_1})P(A_2) + P(\overline{A_1})P(\overline{A_2}) = 0,069$. Mặt khác $B_1 H = A_1 \overline{A_2}$ (nhân B_1 vào công thức của H và để ý $A_1 \overline{A_1} = V$), nên tử số của (4.2) sẽ là 0,019; từ đó ta có lại kết quả cần tìm mà không cần đến nhóm đầy đủ. Tuy nhiên mọi khó khăn rơi vào việc tính trực tiếp $P(H)$.

Thí dụ 4.5. Tại một phòng khám chuyên khoa tỷ lệ người đến khám có bệnh là 83%. Theo thống kê biết rằng nếu chẩn đoán có bệnh thì đúng tới 90%, còn nếu chẩn đoán không bệnh thì chỉ đúng 80%.

a) Tính xác suất chẩn đoán đúng.

b) Biết có một trường hợp chẩn đoán đúng; tìm xác suất người được chẩn đoán đúng có bệnh.

Giải. Gọi H sự kiện chẩn đoán đúng, vậy $\bar{H}$ – chẩn đoán sai; A – người có bệnh, $\bar{A}$ – người không có bệnh; B – chẩn đoán bệnh, $\bar{B}$ – chẩn đoán không bệnh.

a) Để tính $P(H)$, ta thử dùng công thức (do $A, \bar{A}$ – nhóm đầy đủ):

$$P(H) = P(A)P(H|A) + P(\bar{A})P(H|\bar{A}),$$

tuy nhiên $P(H|A)$ – xác suất để khi chẩn đoán người có bệnh thì đúng – chưa biết (chú ý phân biệt với xác suất chẩn đoán có bệnh thì đúng là $P(H|B)$). Vì vậy ta tìm cách dùng công thức thứ hai (do B và $\bar{B}$ tạo ra nhóm đầy đủ).

$$P(H) = P(B)P(H|B) + P(\bar{B})P(H|\bar{B}). \quad (4.4)$$

Nhưng $P(B)$ (và $P(\bar{B})$ nữa) lại chưa biết, tuy nhiên ta có thể khai thác công thức:

$$P(A) = P(B)P(A|B) + P(\bar{B})P(A|\bar{B}). \quad (4.5)$$

Theo giả thiết đầu bài $P(A) = 0,83$; ngoài ra dễ thấy:

$$P(A|B) = P(H|B) = 0,9;$$

$$P(A|\bar{B}) = P(\bar{H}|\bar{B}) = 1 - P(H|\bar{B}) = 1 - 0,8 = 0,2.$$

Từ đó nếu đặt $P(B) = x$, $P(\bar{B}) = 1 - P(B) = 1 - x$ và thay tất cả vào (4.5).

$$0,83 = 0,9x + 0,2(1 - x) \Rightarrow x = P(B) = 0,9.$$

Từ đó thay các kết quả trên vào (4.4)

$$P(H) = 0,9.0,9 + 0,1.0,8 = 0,89.$$

b) Xác suất cần tìm là $P(A|H)$. Áp dụng công thức (4.2):

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

Mặt khác dựa vào ý nghĩa các sự kiện và lại dùng tiếp (4.2)

$$P(H|A) = P(B|A) = \frac{P(B)P(A|B)}{P(A)},$$

từ đó thay vào công thức trên:

$$P(A|H) = \frac{P(B)P(A|B)}{P(H)} = \frac{0,9 \cdot 0,9}{0,89} \approx 0,91.$$

BÀI TẬP

1. Cho 4 sản phẩm và gọi A là sự kiện có ít nhất một phế phẩm, B – cả 4 đều tốt. Cho biết ý nghĩa của các sự kiện sau: $\bar{A}$, $\bar{B}$, $A + B$, AB , $A\bar{B}$, $\bar{A}B$, $\bar{A} + B$, $A + \bar{B}$, $\bar{A} + \bar{B}$, $\bar{A}\bar{B}$.
2. Chứng minh công thức Đơ Moóc-găng:
$$\overline{A + B} = \bar{A}\bar{B}, \quad \overline{AB} = \bar{A} + \bar{B}.$$
3. Có bao nhiêu số tự nhiên mà mỗi số có 4 chữ số?
4. Tìm sự kiện X từ đẳng thức $\overline{X + A} + \overline{X + \bar{A}} = B$.
5. Một giải bóng đá gồm 16 đội. Hỏi phải tổ chức bao nhiêu trận đấu, biết rằng mỗi đội gặp nhau 2 lần?
6. Có bao nhiêu cách xếp 10 quả bóng vào 2 hộp?
7. Có bao nhiêu số điện thoại có các chữ số khác nhau ở một tổng đài nội bộ với các số chỉ có 4 chữ số? Có bao nhiêu số điện thoại có đúng 1 cặp số trùng?
8. Có bao nhiêu cách xếp 5 người ngồi quanh một bàn tròn sao cho hai người định trước ngồi cạnh nhau? Cũng câu hỏi như vậy nhưng thay bàn tròn bằng bàn dài.
9. Một lô hàng có N sản phẩm trong đó có M phế phẩm. Có bao nhiêu cách chọn ra n sản phẩm để trong đó có m phế phẩm?
10. Có bao nhiêu cách để 8 người lên tầng của một tòa nhà có 4 tầng lầu?

11. Xếp ngẫu nhiên một bộ sách gồm 6 tập lên giá sách, tìm xác suất để bộ sách được xếp đúng thứ tự.
12. Một cậu bé có 10 bi, trong đó có 6 đỏ và 4 xanh. Một hôm cậu thấy mất một viên bi, tìm xác suất để nếu rút hù họa ra 1 bi trong số còn lại thì đó là bi đỏ.
13. Tìm xác suất để khi rút hù họa ra n con bài từ cỗ bài tú lơ khơ 52 con thì chúng có giá trị khác nhau (không để ý đến chất).
14. Một lớp học sinh có 30 sinh viên trong đó có 4 giỏi, 8 khá và 10 trung bình. Chọn hù họa ra 3 người, tính các xác suất:
 - a) cả ba đều là học sinh yếu;
 - b) có ít nhất một học sinh giỏi;
 - c) có đúng một học sinh giỏi.
15. Gieo đồng thời 4 đồng tiền cân đối đồng chất, tìm các xác suất:
 - a) cả 4 mặt giống nhau xuất hiện;
 - b) có đúng 2 mặt sấp.
16. Tìm xác suất khi chia đôi một bộ tam cúc thì mỗi phần có đúng một nửa là quân đỏ.
17. Bẻ ngẫu nhiên một thanh gỗ có độ dài l thành 3 đoạn. Tìm xác suất để ba đoạn đó tạo được một tam giác.
18. Tìm xác suất để khi lấy hù họa ra một số có hai chữ số thì nó là bội số của 2 và 3.
19. Bài toán Buýt-phông. Trên mặt phẳng đã kẻ sẵn các đường song song cách đều nhau một khoảng có độ dài $2a$ gieo ngẫu nhiên một kim dài $2l$ ($l < a$). Tính xác suất để chiếc kim cắt một đường thẳng nào đó.
20. Bài toán Ba-nắc. Một người có trong túi 2 bao diêm, mỗi bao có n que. Mỗi khi cần diêm anh ta rút hù họa ra một bao. Tìm xác suất sao cho người đó lần đầu rút phải bao rỗng thì trong bao kia còn đúng k que ($k = 1, 2, \dots, n$).

21. Xác suất trúng đích của một lần bắn là 0,4. Cần phải bắn bao nhiêu phát để xác suất có ít nhất một viên trúng sẽ lớn hơn 0,95?
22. Một xí nghiệp có 3 xe tải với xác suất hỏng trong ngày của mỗi xe tương ứng là 0,01; 0,005 và 0,002. Tìm xác suất để trong ngày:
- a) có 2 xe bị hỏng;
 - b) có ít nhất một xe hỏng.
23. Xếp ngẫu nhiên 10 quyển sách vào 2 ngăn kéo. Tính các xác suất:
- a) ngăn kéo nào cũng có sách;
 - b) ngăn kéo thứ nhất có 2 quyển sách và ngăn thứ hai có 6 quyển sách.
24. Chứng minh rằng nếu A và B độc lập thì các cặp sự kiện sau cũng độc lập: A và $\bar{B}$, $\bar{A}$ và B , $\bar{A}$ và $\bar{B}$.
25. Một gia đình có 6 con. Giả sử xác suất sinh con trai là 0,5, tính các xác suất để trong 6 con có:
- a) đúng 3 con trai;
 - b) có không quá 3 con trai;
 - c) có nhiều nhất 4 con trai.
26. Một xạ thủ phải bắn cho đến khi nào trúng thì thôi. Tìm xác suất để anh ta phải bắn không quá 4 lần, biết rằng xác suất trúng của mỗi lần bắn là 0,6.
27. Trong thời gian có dịch ở 1 vùng dân cư cứ 100 người bị dịch thì có 10 người phải đi cấp cứu. Xác suất gặp một người phải cấp cứu vì mắc bệnh dịch ở vùng đó là 0,06. Tìm tỉ lệ mắc bệnh dịch của vùng dân cư.
28. Một công nhân đứng máy 1000 ống sợi. Xác suất mỗi ống bị đứt trong vòng một giờ là 0,005. Tính xác suất để trong vòng 1 giờ có: a) 40 ống sợi bị đứt; b) không quá 40 ống sợi bị đứt.

29. Tỷ lệ hút thuốc ở một vùng là 35%. Theo thống kê biết rằng tỷ lệ viêm họng trong số người hút thuốc là 60%, còn trong số không hút là 30%. Khám ngẫu nhiên một người thì thấy anh ta bị viêm họng; tìm xác suất đó là người hút thuốc. Nếu anh ta không bị viêm họng thì xác suất đó bằng bao nhiêu?
30. Một xạ thủ bắn 4 phát đạn với xác suất bắn trúng của mỗi viên đạn là 0,7. Biết rằng có hai viên trúng, tìm xác suất để viên thứ nhất đã trúng đích.
31. Một phân xưởng có 3 máy với xác suất trục trặc trong ngày của từng máy là 0,1; 0,05 và 0,2. Cuối ngày thấy có 2 máy trục trặc, tính xác suất đó là máy thứ hai và ba.
32. Một người có 3 chỗ ưa thích như nhau để câu cá. Xác suất để câu được cá mỗi lần thả câu ở từng nơi tương ứng là 0,2; 0,3 và 0,4. Biết rằng ở một chỗ anh ta thả câu 3 lần và chỉ câu được 1 con cá, tìm xác suất để đó là chỗ thứ nhất.
33. Ở một bệnh viện tỷ lệ mắc bệnh A là 15%. Để chẩn đoán xác định người ta phải làm phản ứng miễn dịch, nếu không bị bệnh thì phản ứng dương tính chỉ có 10%. Mặt khác biết rằng khi phản ứng là dương tính thì xác suất bị bệnh là 60%.
- a) Tính xác suất phản ứng dương tính của nhóm có bệnh.
- b) Tính xác suất chẩn đoán đúng.

Chương II

BIẾN NGẪU NHIÊN VÀ LUẬT PHÂN PHỐI XÁC SUẤT

§1. KHÁI NIỆM BIẾN NGẪU NHIÊN

1.1. Khái niệm

Tính toán bằng số vốn đã quen thuộc và dễ sử dụng trong ứng dụng, nhất là có dùng tới máy tính. Khi nghiên cứu các sự kiện ngẫu nhiên, rất bất tiện khi mô tả và làm tính với các sự kiện.

Khái niệm biến số (đại lượng biến thiên) đã rất thông dụng trong giải tích toán. Chính vì thế ta tìm cách đưa vào khái niệm *biến số ngẫu nhiên* như là một đại lượng phụ thuộc vào kết cục của một phép thử ngẫu nhiên nào đó.

Thí dụ 1.1. Gieo một con xúc sắc. Nếu ta gọi biến ngẫu nhiên là “số chấm xuất hiện”, rõ ràng nó phụ thuộc vào kết cục của phép thử và nhận các giá trị nguyên từ 1 đến 6.

Thí dụ 1.2. Nghiên cứu biến ngẫu nhiên “nhiệt độ” của một phản ứng hóa học trong một khoảng thời gian nào đó. Rõ ràng nhiệt độ đó nhận giá trị trong một khoảng $[t; T]$, trong đó t và T là các nhiệt độ thấp nhất và cao nhất của phản ứng trong khoảng thời gian trên.

Về mặt hình thức, có thể định nghĩa biến ngẫu nhiên như là một hàm số có giá trị thực xác định trên không gian các sự kiện sơ cấp (sao cho nghịch ảnh của một khoảng số là một sự kiện). Để phân biệt sau này ta kí hiệu $X, Y, \dots$ là các biến ngẫu nhiên, còn $x, y, \dots$ là giá trị của các biến ngẫu nhiên đó. Như

vậy, X mang tính ngẫu nhiên, còn x là giá trị cụ thể quan sát được khi phép thử đã tiến hành (trong thống kê được gọi là *thể hiện* của X).

Việc xác định một biến ngẫu nhiên bằng tập các giá trị của nó rõ ràng là chưa đủ. Bước tiếp theo là phải xác định xác suất của từng giá trị hoặc từng tập các giá trị. Vì thế ở tiết sau ta sẽ phải dùng tới khái niệm về phân phối xác suất của biến ngẫu nhiên X .

1.2. Phân loại

Biến ngẫu nhiên được gọi là *rời rạc*, nếu tập giá trị của nó là một tập hữu hạn hoặc vô hạn đếm được các phần tử. Thí dụ: số điểm thi của một học sinh, số cuộc gọi điện thoại của một tổng đài trong một đơn vị thời gian, số tai nạn giao thông, ...

Biến ngẫu nhiên được gọi là *liên tục*, nếu tập giá trị của nó lấp kín một khoảng trên trục số (số phần tử của tập giá trị là vô hạn không đếm được theo lý thuyết số). Thí dụ: huyết áp của một bệnh nhân, độ dài của chi tiết máy, tuổi thọ của một loại bóng đèn điện tử, ...

Như vậy miền giá trị của một biến rời rạc sẽ là một dãy số $x_1, x_2, \dots, x_n, \dots$ có thể hữu hạn hoặc vô hạn. Miền giá trị của một biến liên tục sẽ là một đoạn $[a; b] \subset \mathbf{R}$ hoặc là chính $\mathbf{R} = (-\infty, +\infty)$.

§2. LUẬT PHÂN PHỐI XÁC SUẤT

2.1. Bảng phân phối xác suất và hàm xác suất

Đối với biến ngẫu nhiên rời rạc, mỗi giá trị của nó được gán với một xác suất đặc trưng cho khả năng biến ngẫu nhiên nhận giá trị đó $p_i = P(X = x_i)$. Như vậy ta đã xác định:

Định nghĩa 1. *Bảng phân phối xác suất* của biến ngẫu nhiên X là

$X = x$	x_1	x_2		x_n	
$p(x)$	p_1	p_2		P_n	

trong đó $\{x_1, x_2, \dots, x_n, \dots\}$ là tập các giá trị của X (đã sắp xếp theo thứ tự tăng); còn $p_n = p(x_n) = P(X = x_n)$.

Thí dụ 2.1. Bảng phân phối xác suất của thí dụ 1.1 §1 sẽ là:

x	1	2	3	4	5	6
$p(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Để thấy các $p(x)$, $x = \overline{1,6}$, đều bằng nhau; hay X có phân phối đều trên tập số $\{1, 2, \dots, 6\}$. Chú ý rằng $p(x) = 0$ với mọi x không nằm trong tập giá trị trên của X , chẳng hạn $p(8) = 0$.

Thí dụ 2.2. Một xạ thủ chỉ có 3 viên đạn. Anh ta được yêu cầu bắn từng phát cho đến khi trúng mục tiêu thì dừng bắn, biết rằng xác suất trúng của mỗi lần bắn là 0,6. Hãy lập bảng phân phối xác suất của số đạn cần bắn.

Giải. Rõ ràng số đạn cần bắn, ký hiệu là X , là một biến ngẫu nhiên rời rạc và từ yêu cầu của bài toán sẽ có 3 giá trị là 1, 2 và 3. $X = 1$ là sự kiện phát thứ nhất trúng và $p_1 = P(X = 1) = 0,6$; $X = 2$ là sự kiện phát thứ nhất trượt, còn phát thứ hai trúng và do độc lập nên $p_2 = P(X = 2) = 0,4 \cdot 0,6 = 0,24$; cuối cùng nếu viên thứ hai vẫn trượt, thì dù viên thứ ba kết quả thế nào, p_3 vẫn bằng $P(X = 3) = 0,4^2 = 0,16$. Từ đó bảng phân phối cần tìm:

X	1	2	3
$p(x)$	0,6	0,24	0,16

Thí dụ 2.3. Một xạ thủ bắn 3 phát, xác suất bắn trúng mục tiêu của mỗi phát là 0,6. Hãy lập bảng phân phối xác suất của số đạn trúng mục tiêu.

Giải. Nếu gọi X là số đạn bắn trúng, ta có tập giá trị là $\{0, 1, 2, 3\}$. Ta tính xác suất $P(X = k) = p(k)$ bằng công thức Béc-nu-li $p(k) = C_n^k p^k q^{n-k}$; $n = 3, p = 0,6$; từ đó bảng phân phối cần tìm:

X	0	1	2	3
$p(x)$	0,064	0,288	0,432	0,216

Hàm số $p(x) = P(X = x)$, $x \in$ tập giá trị của X , thường được gọi là *hàm xác suất* của X ; nó có hai tính chất cơ bản:

- (i) $p(x) \geq 0 \forall x$;
- (ii) $\sum_{\text{mọi } x} p(x) = 1$.

Bạn đọc có thể kiểm tra dễ dàng các tính chất này trong 3 thí dụ trên. Ngoài ra có thể thấy rằng hàm của một hoặc nhiều biến ngẫu nhiên vẫn tiếp tục là một biến ngẫu nhiên. Trong trường hợp biến rời rạc việc tìm luật phân phối của một biến hàm như vậy thường dễ hơn so với biến liên tục.

Thí dụ 2.4. Cho hai biến X và Y có bảng phân phối tương ứng:

x	-1	0	1
$p(x)$	0,3	0,4	0,3

và

y	1	2
$p(y)$	0,3	0,7

Hãy lập bảng phân phối xác suất của: a) X^2 ; b) $X + Y$.

Giải. a) Biến $Z = X^2$ rõ ràng chỉ có hai giá trị 0 và 1, từ đó bảng phân phối xác suất của nó:

z	0	1
$p(z)$	0,4	0,6

b) Biến $Z = X + Y$ có các giá trị sau: 0, 1, 2 và 3. Để ý rằng

$$P(Z = z_k) = P(X + Y = z_k) = \sum_{x_i + y_j = z_k} P(X = x_i; Y = y_j).$$

trong đó tổng hiểu theo nghĩa lấy theo mọi giá trị x_i và y_j của X và Y sao cho $x_i + y_j = z_k$; còn $P(X = x_i; Y = y_j)$ là xác suất

để đồng thời $X = x_i$ và $Y = y_j$. Nếu X và Y không có quan hệ gì (tức độc lập, sẽ nói đến ở chương III) thì rõ ràng xác suất $P(X = x_i; Y = y_j) = P(X = x_i)P(Y = y_j)$. Từ đó bảng phân phối của Z :

z	0	1	2	3
$p(z)$	0,09	0,33	0,37	0,21

(chẳng hạn $P(Z = 2) = P(X = 0; Y = 2) + P(X = 1; Y = 1) = 0,4.0,7 + 0,3.0,3 = 0,37$).

2.2. Hàm phân phối xác suất

Bảng phân phối xác suất có một hạn chế cơ bản là chưa đủ tổng quát để đặc trưng cho một biến ngẫu nhiên tùy ý, nhất là trường hợp biến liên tục. Vì vậy người ta đưa ra khái niệm sau:

Định nghĩa 2. *Hàm phân phối xác suất* của biến ngẫu nhiên X , ký hiệu là $F(x)$, được xác định như sau:

$$F(x) = P(X < x), \quad x \in \mathbf{R} \quad (2.1)$$

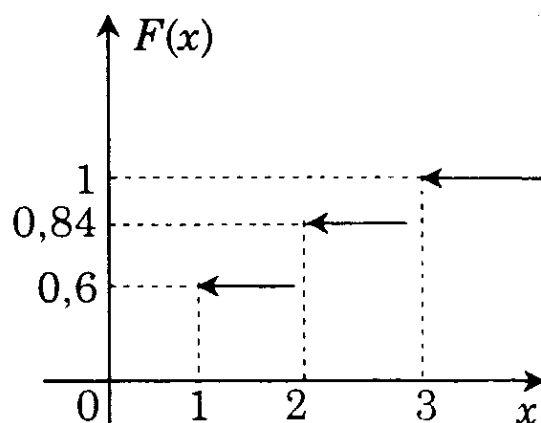
Từ định nghĩa trên, $F(x)$ phản ánh độ tập trung xác suất ở bên phải của số thực x . Trong trường hợp biến ngẫu nhiên rời rạc, (2.1) cho ta một hàm còn được gọi là hàm phân phối *tích lũy* (hay xác suất tích lũy).

Thí dụ 2.5. Từ bảng phân phối của thí dụ 2.2 và dùng (2.1) ta sẽ có $F(x) = \sum_{x_i < x} p(x_i)$ và

$$F(x) = \begin{cases} 0, & x \leq 1, \\ 0,6, & 1 < x \leq 2, \\ 0,84, & 2 < x \leq 3, \\ 1, & x > 3. \end{cases}$$

Đồ thị của hàm phân phối xác suất này là hàm bậc thang. Để ý là X có bao nhiêu giá trị thì $F(x)$ có bằng ấy điểm gián đoạn loại 1 (xem hình 2.1).

Hàm phân phối xác suất có vai trò quan trọng khi nghiên cứu các biến ngẫu nhiên liên tục. Nếu ta biết được hàm phân phối xác suất có nghĩa là xác định hoàn toàn biến ngẫu nhiên. Tuy nhiên trong thực tế cũng phải thấy rằng việc tìm được $F(x)$ là rất khó, nếu không nói là hầu như không thể làm được.



Hình 2.1

Có thể nêu ra một vài tính chất của hàm $F(x)$:

(i) $1 \geq F(x) \geq 0$.

(ii) $F(x)$ là một hàm không giảm, tức là nếu $x_2 > x_1$ thì $F(x_2) \geq F(x_1)$.

(iii) $P(\alpha \leq X \leq \beta) = F(\beta) - F(\alpha)$. (2.2)

Hệ quả hiển nhiên: nếu X liên tục và $F(x)$ liên tục tại α thì $P(X = \alpha) = 0$.

(iv) $F(+\infty) = 1$; $F(-\infty) = 0$.

Việc chứng minh các tính chất trên có thể dựa vào định nghĩa (2.1). Cũng từ định nghĩa ấy ta thấy $F(x)$ ít nhất phải là hàm liên tục trái (xem ví dụ 2.5 ở trên), còn trong trường hợp X liên tục thì $F(x)$ nói chung là một hàm liên tục. Trong tính chất (iv) $F(+\infty)$ ký hiệu $\lim_{x \rightarrow +\infty} F(x)$, tương tự đối với $F(-\infty)$. Cuối

cùng để ý là (2.2) luôn đúng với mọi biến X liên tục hay rời rạc, trong trường hợp $F(x)$ liên tục có hệ quả hiển nhiên:

$$P(\alpha \leq X \leq \beta) = P(\alpha < X < \beta) = P(\alpha < X \leq \beta) = P(\alpha \leq X < \beta).$$

Thí dụ 2.6. Cho hàm phân phối xác suất của một biến ngẫu nhiên liên tục X có dạng:

$$F(x) = \begin{cases} 0, & x \leq 2, \\ a(x-2)^2, & 2 < x \leq 4, \\ 1, & x > 4. \end{cases}$$

Xác định hằng số a và tính $P(2 \leq X < 3)$.

Giải. Do $F(x)$ liên tục, nên tại $x = 4$ ta phải có $a(4-2)^2 = 1$, từ đó $a = \frac{1}{4}$. Dùng (2.2) ta có:

$$P(2 \leq X < 3) = F(3) - F(2) = \frac{1}{4}(3-2)^2 - 0 = \frac{1}{4}.$$

2.3. Hàm mật độ xác suất

Hàm phân phối $F(x)$ còn một hạn chế (mà bảng phân phối không có) là không cho biết rõ phân phối xác suất ở lân cận một điểm nào đó trên trục số. Vì vậy đối với các biến ngẫu nhiên liên tục, có $F(x)$ khả vi, người ta đưa ra khái niệm sau đây.

Định nghĩa 3. *Hàm mật độ xác suất* của biến ngẫu nhiên X , ký hiệu là $f(x)$, có hàm phân phối $F(x)$ khả vi (trừ ở một số hữu hạn điểm gián đoạn bị chặn), được xác định bằng

$$f(x) = F'(x). \quad (2.3a)$$

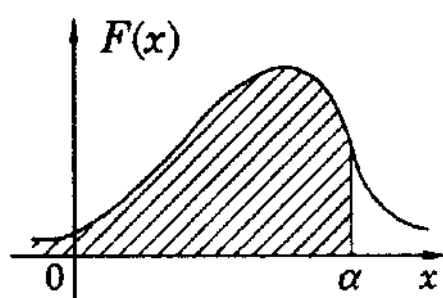
Từ công thức định nghĩa (2.3a) và các khái niệm đạo hàm và tích phân, ta có ngay do tích phân là phép toán ngược của đạo hàm

$$F(x) = \int_{-\infty}^x f(t)dt. \quad (2.3b)$$

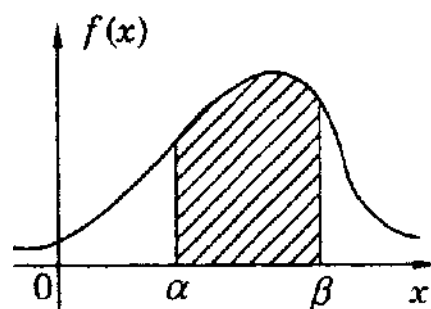
Từ đó công thức (2.2) sẽ tương đương với:

$$P(\alpha \leq X < \beta) = \int_{\alpha}^{\beta} f(x)dx. \quad (2.4)$$

Về mặt hình học (2.3b) và (2.4) cho ta diện tích phần mặt phẳng chắn bởi đường cong $y = f(x)$, trục Ox và các đường thẳng tương ứng (xem hình 2.2 và 2.3).



Hình 2.2. Giá trị của $F(\alpha)$



Hình 2.3. Xác suất $P(\alpha \leq X < \beta)$

Hàm mật độ xác suất của một biến liên tục có hai tính chất cơ bản giống như hàm xác suất ở mục 2.1 là

(i) $f(x) \geq 0 \quad \forall x$;

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

Từ định nghĩa (2.1) và khái niệm đạo hàm, ta có thể thấy ở nơi nào giá trị của $f(x)$ lớn thì tại lân cận điểm đó có độ tập trung xác suất cao, điều đó giải thích tên gọi mật độ xác suất.

Thí dụ 2.7. Cho hàm mật độ của biến ngẫu nhiên X có dạng :

$$f(x) = \begin{cases} a \cos x, & x \in \left[-\frac{\pi}{2}; \frac{\pi}{2}\right], \\ 0, & x \notin \left[-\frac{\pi}{2}; \frac{\pi}{2}\right]. \end{cases}$$

a) Tìm a và xác định hàm phân phối xác suất $F(x)$ của X .

b) Tính xác suất để X nhận giá trị trong khoảng $\left(\frac{\pi}{4}, \pi\right)$.

Giải. a) Dùng tính chất (ii) của hàm mật độ:

$$\int_{-\infty}^{+\infty} f(x) dx = a \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos x dx = 2a = 1,$$

từ đó rút ra $a = \frac{1}{2}$. Việc tìm $F(x)$ dựa vào công thức (2.3b). Ta có:

Với $x \leq -\frac{\pi}{2}$ thì $\int_{-\infty}^x f(x)dx = 0;$

Với $-\frac{\pi}{2} < x \leq \frac{\pi}{2}$ thì $\int_{-\infty}^x f(x)dx =$

$$= \int_{-\infty}^{-\frac{\pi}{2}} 0dx + \int_{-\frac{\pi}{2}}^x \frac{1}{2} \cos x dx = \frac{1}{2}(\sin x + 1);$$

Với $x > \frac{\pi}{2}$ thì $\int_{-\infty}^x f(x)dx = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{2} \cos x dx = 1.$

$$\text{Từ đó } F(x) = \begin{cases} 0, & x \leq -\frac{\pi}{2}, \\ \frac{1}{2}(\sin x + 1), & -\frac{\pi}{2} < x \leq \frac{\pi}{2}, \\ 1, & x > \frac{\pi}{2}. \end{cases}$$

b) Theo (2.2):

$$\begin{aligned} P\left(\frac{\pi}{4} < X < \pi\right) &= P\left(\frac{\pi}{4} \leq X < \pi\right) = F(\pi) - F\left(\frac{\pi}{4}\right) \\ &= 1 - \frac{1}{2}\left(\sin \frac{\pi}{4} + 1\right) = \frac{1}{2} - \frac{\sqrt{2}}{4}. \end{aligned}$$

Thí dụ 2.8. Cho xác suất phân rã của một nguyên tử chất phóng xạ trong khoảng thời gian dt khá bé là λdt (giả sử sự phân rã đó không phụ thuộc vào quá khứ). Hãy xác định:

a) Xác suất để nguyên tử đó phân rã trong khoảng thời gian t ;

b) Hàm mật độ xác suất của thời điểm phân rã của nguyên tử.

Giải.

a) Dễ thấy xác suất không phân rã của nguyên tử trong khoảng thời gian dt là $1 - \lambda dt$. Chia khoảng thời gian t thành t/dt các khoảng con có độ dài dt ; từ đó xác suất để nguyên tử không phân rã trong khoảng thời gian đó xấp xỉ là (do có giả thuyết độc lập) $(1 - \lambda dt)^{t/dt}$. Lấy giới hạn khi $dt \rightarrow 0$, ta có xác suất cần tìm là $1 - e^{-\lambda t}$ (bằng $1 -$ xác suất nguyên tử không phân rã trong khoảng thời gian t).

b) Gọi T là thời điểm phân rã của nguyên tử và $f(t)$ là hàm mật độ của T . Rõ ràng xác suất để nguyên tử phân rã ở thời điểm trong khoảng thời gian từ t đến $t + dt$ sẽ bằng xác suất không phân rã trong khoảng thời gian t trước đó nhân với xác suất phân rã trong khoảng thời gian dt , từ đó:

$$P(t \leq T < t + dt) = f(t)dt = e^{-\lambda t} \lambda dt.$$

$$\text{Vậy ta có } f(t) = \begin{cases} 0, & t \leq 0, \\ \lambda e^{-\lambda t}, & t > 0. \end{cases}$$

Đây chính là hàm mật độ của biến ngẫu nhiên tuân theo *luật phân phối mũ*, ký hiệu ở đây $T \sim \mathcal{E}(\lambda)$.

§3. CÁC SỐ ĐẶC TRƯNG CỦA BIẾN NGẪU NHIÊN

Dẫu biết rằng hàm phân phối xác suất cho ta thông tin đầy đủ nhất về biến ngẫu nhiên, nhưng trong thực tế ta không thể xác định được nó; từ đó dẫn đến việc tìm một vài đặc trưng quan trọng, thông thường là đặc trưng về vị trí và về độ phân tán. Trong 3 số đặc trưng về vị trí, đầu tiên ta xét về kỳ vọng, hai số khác là mô men và trung vị sẽ xét ở mục 3.3.

3.1. Kỳ vọng

Định nghĩa 1. *Kỳ vọng* của biến ngẫu nhiên X , ký hiệu là EX , được xác định như sau:

– nếu X là biến rời rạc có hàm xác suất $p(x_i) = p_i, i = 1, 2, \dots$ thì:

$$EX = \sum_{\forall i} x_i p_i; \quad (3.1a)$$

– nếu X là biến liên tục có hàm mật độ $f(x), x \in \mathbf{R}$, thì:

$$EX = \int_{-\infty}^{\infty} xf(x)dx. \quad (3.1b)$$

Từ (3.1) ta thấy kỳ vọng chính là tổng có trọng số của tất cả các giá trị của X , hay còn là trị trung bình của biến ngẫu nhiên (phân biệt với trung bình cộng của các giá trị). Trong thực tế, nếu quan sát các giá trị của X nhiều lần và lấy trung bình cộng, thì khi số quan sát càng lớn số trung bình đó càng gần tới kỳ vọng EX , vì vậy kỳ vọng còn được gọi là *trị trung bình* của biến X mà không sợ nhầm lẫn.

Thí dụ 3.1. Xét lại thí dụ 2.1 với X là số chấm xuất hiện khi gieo một con xúc sắc. Theo (3.1a)

$$EX = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3,5.$$

Như vậy trong trường hợp xác suất được phân phối đều trên tập giá trị, kỳ vọng chính là trung bình cộng của các giá trị ấy. $EX = 3,5$ còn có nghĩa là nếu gieo nhiều lần số chấm trung bình thu được sẽ là 3,5.

Thí dụ 3.2. Tìm kỳ vọng của biến X trong thí dụ 2.3.

Giải. Theo (3.1a) ta có:

$$EX = 0.0,064 + 1.0,288 + 2.0,432 + 3.0,216 = 1,8.$$

Thí dụ 3.3. Tìm kỳ vọng của biến X trong thí dụ 2.6.

Giải. Trước hết ta phải tìm hàm mật độ của X , theo (2.3a)

$$f(x) = \begin{cases} \frac{1}{2}(x - 2), & x \in [2; 4], \\ 0, & x \notin [2; 4]. \end{cases}$$

Từ đó theo (3.1b):

$$EX = \int_{-\infty}^{+\infty} xf(x)dx = \int_2^4 \frac{x}{2}(x-2)dx = \frac{10}{3}.$$

Thí dụ 3.4. Một người mua 10000 đồng xổ số lô tô 2 số với luật chơi như sau: anh ta sẽ thắng 700000 đồng (gấp 70 lần tiền mua) nếu số mua trùng với 2 số cuối của giải độc đắc gần nhất sắp tới (và không được đồng nào nếu không trùng). Hãy tìm số tiền thắng trung bình của một lần chơi như vậy.

Giải. Gọi X là số tiền thắng của một lần chơi, rõ ràng X nhận các giá trị 0^d và 700000^d với các tần suất (và coi luôn là xác suất cũng không sợ làm mất tổng quát) tương ứng là 99% và 1%. Từ đó số tiền thắng trung bình chính là:

$$EX = 0^d.99\% + 700000^d.1\% = 7000 \text{ đồng}.$$

Mặc dù $EX > 0$, nhưng chớ quên rằng anh ta đã bỏ ra 10000 đồng để mua xổ số. Như vậy trong thực tế mỗi lần chơi anh ta mất trung bình 3000 đồng.

Ta phát biểu một số *tính chất* của kỳ vọng:

(i) $Ec = c$ (c – hằng số);

(ii) $E(cX) = cEX$;

(iii) $E(X + Y) = EX + EY$;

(iv) Nếu X, Y độc lập thì $E(XY) = EX.EY$

(để ý rằng khái niệm độc lập sẽ được làm rõ hơn ở chương III);

(v) Nếu $Y = \varphi(X)$, thì phụ thuộc vào X rời rạc hay liên tục

ta có: $EY = \sum_i \varphi(x_i)p_i$ hoặc $EY = \int_{-\infty}^{+\infty} \varphi(x)f(x)dx$, trong đó các

$p(x)$ và $f(x)$ là các hàm xác suất hoặc mật độ tương ứng.

Thí dụ 3.5. Gieo đồng thời 2 con xúc sắc. Tìm tổng số chấm trung bình.

Giải. Gọi X_i là số chấm xuất hiện của con xúc sắc thứ i ($i = 1, 2$), dễ thấy từ thí dụ 3.1 $EX_1 = EX_2 = 3,5$. Mặt khác tổng số chấm của 2 con xúc sắc sẽ là $X_1 + X_2$, từ đó dùng tính chất (iii) của kỳ vọng, ta có $E(X_1 + X_2) = 3,5 + 3,5 = 7$.

3.2. Phương sai

1. Dùng phép lấy kỳ vọng ở mục trước, ta có thể định nghĩa khái niệm phương sai.

Định nghĩa 2. *Phương sai* của biến ngẫu nhiên X , ký hiệu là VX , được định nghĩa như sau:

$$VX = E[(X - EX)^2]. \quad (3.2)$$

Trong (3.2) ta thấy $X - EX$ chính là độ lệch của biến X so với trung bình của nó, từ đó phương sai chính là trung bình của bình phương độ lệch đó. Vậy phương sai đặc trưng cho độ phân tán của biến ngẫu nhiên quanh trị trung bình của biến đó. Cũng theo ý nghĩa đó phương sai càng lớn thì độ bất định của biến tương ứng càng lớn.

Trong tính toán, phụ thuộc vào X là rời rạc (với hàm xác suất $p(x)$) hay liên tục (với hàm mật độ $f(x)$), ta có hai công thức tính phương sai:

$$VX = \sum_{\forall i} (x_i - EX)^2 p_i. \quad (3.3a)$$

hoặc:
$$VX = \int_{-\infty}^{\infty} (x - EX)^2 f(x) dx. \quad (3.3b)$$

Tuy nhiên việc tính theo (3.3) khá phức tạp, vì vậy, dùng các tính chất của kỳ vọng, ta có thể biến đổi (3.2) về dạng tương đương, khá dễ dàng để tính toán

$$VX = E(X^2) - (EX)^2, \quad (3.4)$$

với các phương án tính ứng với X rời rạc hay liên tục như trong (3.3):

$$VX = \sum_{\forall i} x_i^2 p_i - \left(\sum_{\forall i} x_i p_i \right)^2 \quad (3.4a)$$

$$VX = \int_{-\infty}^{+\infty} x^2 f(x) dx - \left(\int_{-\infty}^{+\infty} x f(x) dx \right)^2 \quad (3.4b)$$

Thí dụ 3.6. Bảng phân phối của biến ngẫu nhiên X trong thí dụ 2.3 có dạng:

x	0	1	2	3
$p(x)$	0,064	0,288	0,432	0,216

Hãy tính VX .

Giải. Ta đã tính $EX = 1,8$ trong thí dụ 3.2. Rõ ràng việc tính theo (3.3a) khá phức tạp. Ta sẽ dùng công thức (3.4a)

$$\begin{aligned} VX &= 0^2 \cdot 0,064 + 1^2 \cdot 0,288 + 2^2 \cdot 0,432 + 3^2 \cdot 0,216 - 1,8^2 \\ &= 3,96 - 3,24 = 0,72. \end{aligned}$$

Thí dụ 3.7. Cho hàm mật độ của biến ngẫu nhiên X tuân theo phân phối mũ (xem thí dụ 2.8, để ý $\lambda > 0$)

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

Hãy tính phương sai của X .

Giải. Đầu tiên ta tính kỳ vọng theo (3.1)

$$EX = \int_{-\infty}^{+\infty} x f(x) dx = \lambda \int_0^{+\infty} x e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Từ đó dùng (3.4b) ta có:

$$VX = \lambda \int_0^{+\infty} x^2 e^{-\lambda x} dx - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

2. Để ý rằng phương sai VX luôn là một số không âm. Từ định nghĩa ta cũng thấy rằng về mặt vật lý VX không cùng thứ nguyên (cùng đơn vị đo) đối với X , vì vậy ta đưa vào khái niệm sau đây:

Định nghĩa 3. *Độ lệch chuẩn* của biến ngẫu nhiên X , ký hiệu là $\sigma(X)$, được định nghĩa như sau:

$$\sigma(X) = \sqrt{VX}. \quad (3.5)$$

Từ định nghĩa (3.5), nhiều khi người ta ký hiệu phương sai là $\sigma^2(X)$ hoặc σ^2 (nếu đã biết rõ là phương sai của biến nào). Độ lệch chuẩn được dùng thường xuyên hơn phương sai do có cùng đơn vị đo với chính biến X .

3. Cuối cùng ta phát biểu một số *tính chất* của phương sai và độ lệch chuẩn:

(i) $Vc = 0$ (c – hằng số);

(ii) $V(cX) = c^2 VX$; $\sigma(cX) = |c| \sigma(X)$;

(iii) Nếu X, Y độc lập thì $V(X+Y) = VX + VY$;

$$\sigma(X+Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)}.$$

Chú ý điều kiện độc lập là khá chặt, sau này ở chương III ta thấy có thể giảm nhẹ. Từ ba tính chất trên, ta có thể dẫn ra 2 hệ quả quan trọng:

$$- V(X + c) = VX.$$

– Phương sai của trung bình cộng n biến ngẫu nhiên độc lập cùng ph. n phối sẽ bé hơn n lần phương sai của các biến thành phần, tức là nếu $VX_i = \sigma^2 \forall i = \overline{1, n}$, thì:

$$V \bar{X} = V \left(\frac{X_1 + X_2 + \dots + X_n}{n} \right) = \frac{\sigma^2}{n}.$$

Đây chính là lý do khi đo đạc các đại lượng vật lý người ta thường đo nhiều lần rồi lấy trung bình cộng các kết quả.

3.3. Một số đặc số khác

1. *Mốt* là giá trị của biến ngẫu nhiên X có khả năng xuất hiện lớn nhất trong một lần cận nào đó của nó. Như vậy đối với biến rời rạc *mốt* là giá trị của X ứng với xác suất lớn nhất, còn đối với biến liên tục *mốt* là giá trị làm hàm mật độ đạt *max*. Như vậy *mốt* có thể chỉ là cực đại địa phương và một biến ngẫu nhiên có thể có một *mốt* hoặc nhiều *mốt*.

Thí dụ 3.8. Cho hàm mật độ của biến ngẫu nhiên X tuân theo phân phối Vây-bun

$$f(x) = \begin{cases} 0, & x \leq 0, \\ \frac{x}{2} e^{-\frac{x^2}{4}}, & x > 0. \end{cases}$$

Hãy xác định *mốt* của X .

Giải. *Mốt* của X sẽ là nghiệm của phương trình:

$$f'(x) = \frac{1}{2} e^{-\frac{x^2}{4}} - \frac{x^2}{4} e^{-\frac{x^2}{4}} = 0.$$

Từ đó *mốt* sẽ là nghiệm của $1 - \frac{x^2}{2} = 0$. Nhưng do $x > 0$,

suy ra *mốt* $= \sqrt{2} \approx 1,414$.

2. *Trung vị* là giá trị của biến ngẫu nhiên X chia phân phối thành hai phần có xác suất giống nhau, tức là nếu ký hiệu *trung vị* là $medX$ thì:

$$P(X < medX) = P(X \geq medX) = \frac{1}{2}.$$

Từ định nghĩa hàm phân phối, rõ ràng để tìm *trung vị* ta chỉ cần giải $F(x) = \frac{1}{2}$. Trong nhiều trường hợp ứng dụng, *trung vị* là đặc trưng vị trí rất tốt, nhiều khi tốt hơn cả kỳ vọng, nhất là khi trong số liệu có những sai sót thái quá.

Trung vị còn có tên gọi là phân vị 50% của phân phối. *Phân vị* là một điểm (giá trị của X) sao cho xác suất để biến ngẫu nhiên nhận giá trị bé hơn nó sẽ bằng số phần trăm cho trước của tổng xác suất phân phối, chẳng hạn ta nói rằng 2 là phân vị 72% của X nếu $F(2) = 0,72$. Thông thường người ta hay xét các phân vị 25%, 50% (trung vị), 75%, 95%,...

Thí dụ 3.9. Tìm trung vị của biến X trong thí dụ 3.8.

Giải. Rõ ràng trung vị là nghiệm của phương trình:

$$\int_0^{medX} f(x)dx = 0,5 \text{ hay } 1 - e^{-\frac{(medX)^2}{4}} = 0,5;$$

từ đó suy ra $medX = 1,665$.

Nói chung, ba số đặc trưng kỳ vọng, một và trung vị không trùng nhau, chẳng hạn từ thí dụ 3.8 và 3.9 và tính thêm trung bình, ta có $EX = 1,772$; một = 1,414 và $med X = 1,665$. Tuy nhiên trong trường hợp phân phối đối xứng và chỉ có một thì cả ba đặc trưng đó trùng nhau.

3. Mômen là khái niệm tổng quát hơn so với kỳ vọng và phương sai.

Định nghĩa 4. *Mômen cấp k đối với a của biến ngẫu nhiên X là một số xác định như sau:*

$$\nu_k(a) = E[(X - a)^k]. \quad (3.6)$$

Nếu $a = 0$, ta ký hiệu $\nu_k = \nu_k(0) = E(X^k)$ và gọi nó là *mômen gốc cấp k* . Rõ ràng kỳ vọng chính là mômen gốc cấp 1 $EX = \nu_1$. Nếu $a = EX$, ta ký hiệu $\mu_k = \nu_k(EX) = E[(X - EX)^k]$ và gọi nó là *mômen trung tâm cấp k* ; cũng rõ ràng phương sai là mômen trung tâm cấp 2 $VX = \mu_2$.

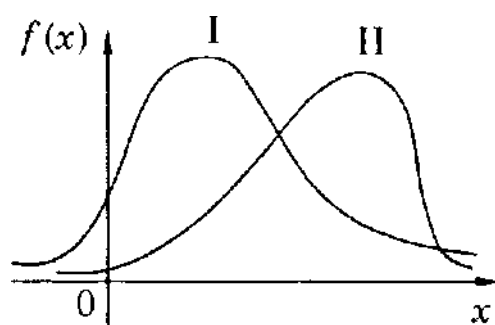
Mômen có vai trò quan trọng trong thống kê và ứng dụng xác suất. Giữa chúng (mômen gốc và mômen trung tâm) có các liên hệ sau:

$$\mu_2 = v_2 - v_1^2 = \sigma^2,$$

$$\mu_3 = v_3 - 3v_2v_1 + 2v_1^3,$$

$$\mu_4 = v_4 - 4v_3v_1 + 6v_2v_1^2 - 3v_1^4, \dots$$

Người ta còn dùng các mômen để đặc trưng cho hình dạng của hàm mật độ phân phối:



Hình 3.1

– Hệ số bất đối xứng là tỷ số $\beta_1 = \frac{\mu_3}{\sigma_3}$; nếu $\beta_1 = 0$ đường cong mật độ đối xứng, nếu nó âm hay dương đường cong đó sẽ bất đối xứng tương ứng với các đường I và II trên hình 3.1.

– Hệ số nhọn là tỷ số $\beta_2 = \frac{\mu_4}{\sigma_4}$. Nếu tỷ số này càng lớn đường cong có đỉnh càng nhọn hơn. Đường cong mật độ của phân phối chuẩn (xét ở mục sau) có $\beta_2 = 3$.

§4. MỘT SỐ PHÂN PHỐI THÔNG DỤNG

4.1. Phân phối đều

1. Phân phối đều rời rạc

Định nghĩa 1. Biến ngẫu nhiên X được gọi là tuân theo luật đều rời rạc với tham số n , ký hiệu là $X \sim \mathcal{U}(n)$, nếu X có bảng phân phối xác suất

x	1	2	...	n
$p(x)$	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

(4.1)

Như vậy hàm xác suất sẽ có dạng $p(i) = \frac{1}{n}$, $i = \overline{1, n}$. Người ta còn mở rộng khái niệm phân phối đều cho biến X nhận giá trị trên một tập hữu hạn bất kỳ có n phần tử $\{x_1, x_2, \dots, x_n\}$; khi đó:

$$p(x_i) = \frac{1}{n}, i = \overline{1, n}.$$

Dễ dàng, nếu $X \sim \mathcal{U}(n)$ và từ (4.1), ta có ngay:

$$EX = \frac{n+1}{2}; VX = \frac{n^2-1}{12}.$$

2. Phân phối đều liên tục

Định nghĩa 2. Biến ngẫu nhiên X được gọi là tuân theo luật phân phối đều liên tục trên $[a; b]$ ký hiệu là $X \sim \mathcal{U}([a; b])$, nếu X có hàm mật độ ($a < b$):

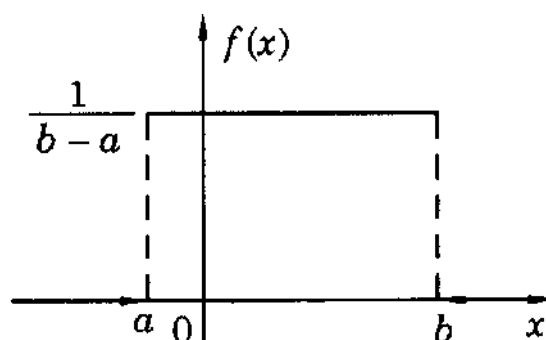
$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a; b], \\ 0, & x \notin [a; b]. \end{cases} \quad (4.2)$$

Đồ thị của hàm $f(x)$ cho trên hình 4.1. Bằng tính toán đơn giản có thể tìm được: nếu $X \sim \mathcal{U}([a; b])$ thì:

$$EX = \frac{a+b}{2};$$

$$VX = \frac{(b-a)^2}{12}.$$

Phân phối đều $\mathcal{U}([0; 1])$ có vai trò rất quan trọng trong mô phỏng các số ngẫu nhiên.



Hình 4.1

4.2. Phân phối nhị thức

1. Phân phối Béc-nu-li

Định nghĩa 3. Biến ngẫu nhiên X được gọi là tuân theo luật phân phối Béc-nu-li, ký hiệu là $X \sim \mathcal{B}(1, p)$, nếu hàm xác suất của nó có dạng:

$$p(x) = p^x(1-p)^{1-x}, x = 0 \text{ và } 1. \quad (4.3)$$

Ta thấy mọi phép thử chỉ có hai kết cục đều có thể mô hình hóa bằng phân phối này. Chẳng hạn một phép thử chỉ có kết cục A với xác suất p và $\bar{A}$ với xác suất $q = 1 - p$. Xây dựng biến ngẫu nhiên X sao cho $P(X = 1) = P(A) = p$ và $P(X = 0) = P(\bar{A}) = q$, ta có $X \sim \mathcal{B}(1, p)$.

Dễ dàng từ (4.3), bảng phân phối xác suất của X :

x	0	1
$p(x)$	q	p

$$EX = 0 \cdot q + 1 \cdot p = p,$$

$$VX = 0^2 \cdot q + 1^2 \cdot p - p^2 = p(1 - p) = pq.$$

Trong thực tế phân phối Béc-nu-li ít được sử dụng (có thể do nó quá đơn giản), tuy nhiên nó được dùng làm cơ sở để tìm luật phân phối của các biến ngẫu nhiên khác.

2. Phân phối nhị thức

Đây là một trong các phân phối rất hay dùng trong thống kê hiện đại. Ở chương I ta đã làm quen với lược đồ Béc-nu-li khi xét dãy n phép thử độc lập, giống nhau, trong mỗi phép thử sự kiện A xuất hiện với xác suất p . Nếu gọi X là số lần xuất hiện A trong dãy n phép thử đó, ta đã biết X có các giá trị từ 0 đến n với các xác suất tương ứng ($q = 1 - p$):

$$p(x) = P_n(x) = C_n^x p^x q^{n-x}, \quad x = \overline{0, n}. \quad (4.4)$$

Định nghĩa 4. Biến ngẫu nhiên X được gọi là tuân theo *luật phân phối nhị thức*, ký hiệu $X \sim \mathcal{B}(n, p)$ nếu hàm xác suất của nó có dạng (4.4).

Bạn đọc hãy tự xây dựng bảng phân phối xác suất của $X \sim \mathcal{B}(n, p)$. Rõ ràng phân phối Béc-nu-li ở trên là một trường hợp riêng của phân phối nhị thức khi $n = 1$. Cần nhắc lại các điều kiện để có phân phối nhị thức:

- dãy các phép thử giống nhau, độc lập;
- trong mỗi phép thử chỉ có 2 kết cục (có và không);
- hai tham số hằng xác định: số các phép thử n và xác suất xuất hiện 1 trong 2 kết cục trên là p .

Thí dụ 4.1. Cho $X \sim \mathcal{B}(5; 0,25)$. Hãy xây dựng bảng phân phối xác suất của X , sau đó tính các xác suất:

- a) $X > 3$; b) $X \geq 1$; c) $X \leq 4$.

Giải. Về mặt ý nghĩa X là số lần xuất hiện sự kiện A nào đó trong dãy 5 phép thử độc lập, biết rằng trong mỗi phép thử sự kiện A có xác suất $P(A) = 0,25$. Dùng công thức (4.4) với $n = 5$; $p = 0,25$, ta sẽ có:

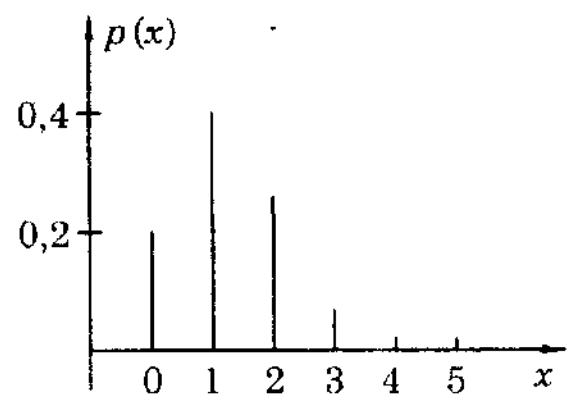
x	0	1	2	3	4	5
$p(x)$	0,2373	0,3955	0,2637	0,0879	0,0146	0,0010

Quan sát bảng số này ta thấy $X = 1$ là giá trị có xác suất lớn nhất, vậy 1 là một của X , trong ứng dụng người ta gọi là *số lần xuất hiện chắc chắn nhất*. Việc tìm các xác suất tương ứng dựa vào bảng số trên:

- a) $P(X > 3) = p(4) + p(5) = 0,0156$;
b) $P(X \geq 1) = 1 - P(X < 1) = 1 - p(0) = 0,7627$;
c) $P(X \leq 4) = 1 - P(X > 4) = 1 - p(5) = 0,9990$.

Có thể dựng biểu đồ (đồ thị của hàm rời rạc) của $p(x)$ như hình 4.2.

Bây giờ ta tính kỳ vọng và phương sai của phân phối nhị thức. Rõ ràng nếu $X \sim \mathcal{B}(n, p)$ thì X là số lần xuất hiện sự kiện A nào đó trong dãy n phép thử Béc-nu-li. Gọi X_i là số lần xuất hiện sự kiện A đó trong phép thử



Hình 4.2

thứ i , $i = \overline{1, n}$. Ta thấy X_i chỉ có hai giá trị 0 và 1 và $P(A) = p = P(X_i = 1)$, $X = X_1 + X_2 + \dots + X_n$. Do các X_i độc lập, mặt khác $EX_i = p$, $VX_i = pq$, nên ta có:

$$EX = \sum_{i=1}^n EX_i = np;$$

$$VX = \sum_{i=1}^n VX_i = npq.$$

Chú ý rằng khi n khá lớn mức độ đối xứng (đối với kỳ vọng) của hàm xác suất càng rõ rệt. Nói chung việc tính xác suất theo công thức (4.4) khá phức tạp, tuy nhiên bằng các chương trình máy tính thì không có vấn đề gì lớn. Việc tính xấp xỉ các xác suất đó đã xét ở §3 chương I và ở các mục dưới đây. Ngoài ra dễ dàng chứng minh hai kết quả sau:

(i) Nếu $X \sim \mathcal{B}(n, p)$ thì $Y = n - X \sim \mathcal{B}(n, 1-p)$.

(ii) Nếu $X_1 \sim \mathcal{B}(n_1, p)$, $X_2 \sim \mathcal{B}(n_2, p)$, thì $X_1 + X_2 \sim \mathcal{B}(n_1 + n_2, p)$.

4.3. Phân phối Poa-xông

Định nghĩa 5. Biến ngẫu nhiên X được gọi là tuân theo *luật phân phối Poa-xông*, ký hiệu là $X \sim \mathcal{P}(\lambda)$, nếu hàm xác suất của nó có dạng:

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots \quad (4.5)$$

Phân phối Poa-xông có nhiều ứng dụng trong lý thuyết phục vụ đám đông, kiểm tra chất lượng sản phẩm... Chẳng hạn số cuộc gọi điện thoại của một tổng đài trong 1 ngày, số lượng khách hàng của một nhà băng trong 1 giờ... đều là các biến ngẫu nhiên có phân phối Poa-xông.

Có thể chứng minh rằng $C_n^x p^x (1-p)^{n-x}$, khi $n \rightarrow +\infty$, $p \rightarrow 0$ sao cho $np \rightarrow \lambda = \text{hằng số}$, có giới hạn $\frac{\lambda^x e^{-\lambda}}{x!}$. Trong thực hành, nếu n khá lớn và p khá bé, thì ($\lambda = np$):

$$P_n(x) = C_n^x p^x (1-p)^{n-x} \approx \frac{\lambda^x e^{-\lambda}}{x!}.$$

Thí dụ 4.2. Người ta vận chuyển 5000 chai rượu vào kho với xác suất vỡ của mỗi chai là 0,0004. Tính xác suất để khi vận chuyển có không quá 1 chai bị vỡ.

Giải. Có thể dùng lược đồ Béc-nu-li (phân phối nhị thức), nhưng $n = 5000$ rất lớn, còn $p = 0,0004$ quá bé. Nếu gọi X là số chai bị vỡ khi vận chuyển, có thể coi phân phối của X xấp xỉ với phân phối Poa-xông với $\lambda = np = 2$. Từ đó theo (4.5):

$$P(0 \leq X \leq 1) = e^{-2} \frac{2^0}{0!} + e^{-2} \frac{2^1}{1!} = \frac{3}{e^2} \approx 0,406.$$

Ta đi tính các số đặc trưng của $X \sim \mathcal{P}(\lambda)$:

$$\begin{aligned} EX &= \sum_{x=0}^{+\infty} xp(x) = \sum_{x=0}^{\infty} xe^{-\lambda} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda; \end{aligned}$$

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

Đôi khi người ta yêu cầu tính *mốt* của X . Người ta đã chứng minh $\lambda - 1 \leq \text{mốt } X \leq \lambda$. Nếu λ nguyên, ta có 2 *mốt* là λ và $\lambda - 1$; còn nếu λ không nguyên, *mốt* sẽ là giá trị nguyên nằm giữa $\lambda - 1$ và λ . Trong thí dụ 4.2 *mốt* của X là 1 và 2, đó là số chai có khả năng vỡ nhiều nhất khi vận chuyển (xác suất vỡ bằng 0,2707 cho mỗi trường hợp $X = 1$ hoặc $X = 2$).

4.4. Các phân phối rời rạc khác

1. Một trong các giả thiết của phân phối nhị thức là sự độc lập của các phép thử thành viên trong dãy. Trong nhiều bài toán thực tế giả thiết đó không được thỏa mãn. Một trường hợp cổ điển là việc chọn mẫu không hoàn lại, trong đó xác suất

không còn là hằng số nữa. Ví dụ ta có N sản phẩm trong đó có m phế phẩm; nếu ta chọn không hoàn lại ra n sản phẩm và gọi X là số phế phẩm trong đó thì $P(X = x)$ sẽ không còn được tính theo (4.4) được nữa (để ý muốn tính theo (4.4) ta phải chọn có hoàn lại). Theo định nghĩa cổ điển, xác suất để trong n sản phẩm có đúng x phế phẩm chính là:

$$P_n(x) = \frac{C_m^x C_{N-m}^{n-x}}{C_N^n}, \quad x = 0, 1, \dots, n.$$

Định nghĩa 6. Biến ngẫu nhiên X được gọi là tuân theo *luật phân phối siêu hình học*, ký hiệu là $X \sim \mathcal{H}(N, n, p)$, nếu hàm xác suất được xác định theo công thức:

$$p(x) = \frac{C_{Np}^x C_{N-Np}^{n-x}}{C_N^n}, \quad x = 0, 1, \dots, n. \quad (4.6)$$

Để ý rằng trong công thức (4.6), nếu lưu ý đến thí dụ bên trên định nghĩa, ta có $p = \frac{m}{N}$ là tỷ lệ phế phẩm lúc ban đầu, và nếu đặt $q = 1 - p$ thì (4.6) sẽ trở thành:

$$p(x) = \frac{C_{Np}^x C_{Nq}^{n-x}}{C_N^n}, \quad x = 0, 1, \dots, n.$$

Khi N rất lớn, xác suất p sẽ ít thay đổi và khi đó ta có thể dùng lại (4.4) để xấp xỉ cho (4.6) và giả thiết p là hằng xác định không bị thay đổi đáng kể.

Thí dụ 4.3. Trong một hộp đèn 15 bóng có 5 bóng kém chất lượng. Chọn ngẫu nhiên ra 10 bóng (tất nhiên không hoàn lại), hãy lập bảng phân phối xác suất của số bóng kém chất lượng trong mẫu chọn ra.

Giải. Rõ ràng X tuân theo phân phối siêu hình học với $N = 15$, $n = 10$ và $p = \frac{1}{3}$. Dùng phần mềm để tính theo (4.6), ta có bảng phân phối như sau:

x	0	1	2	3	4	5
$p(x)$	0,00033	0,01665	0,14985	0,39960	0,34965	0,08392

x	6	7	8	9	10
$p(x)$	0	0	0	0	0

Trong trường hợp này ta không thể xấp xỉ các xác suất bằng phân phối nhị thức được, chẳng hạn tính theo (4.4) với $n = 10$, $p = \frac{1}{3}$, ta có $P_{10}(3) = 0,26012$; $P_{10}(7) = 0,01626$, ... Trong thực hành khi $N > 10n$ người ta mới chấp nhận xấp xỉ bằng phân phối nhị thức.

Có thể chứng minh được rằng nếu $X \sim \mathcal{N}(N, n, p)$:

$$EX = np; VX = npq \frac{N-n}{N-1}.$$

Ngoài ra khi $N \rightarrow \infty$ sao cho $\frac{n}{N} \rightarrow 0$ ta có:

$$\lim_{\frac{n}{N} \rightarrow 0} \frac{C_{Np}^x C_{Nq}^{n-x}}{C_N^n} = C_n^x p^x q^{n-x},$$

với $p = \frac{m}{N}$. Điều đó giải thích cho việc xấp xỉ phân phối siêu hình bằng phân phối nhị thức khi N khá lớn.

Trong mục này ta định nghĩa thêm hai phân phối rời rạc lấy cơ sở của phép thử Béc-nu-li.

2. Định nghĩa 7. Biến ngẫu nhiên X được gọi là tuân theo luật phân phối hình học, ký hiệu là $X \sim \mathcal{G}(p)$, nếu hàm xác suất của nó có dạng:

$$p(x) = p(1-p)^x, \quad x = 0, 1, 2, \dots \quad (4.7)$$

Từ đó ta thấy X chính là số lần không xuất hiện trước lần xuất hiện đầu tiên của một sự kiện A nào đó (trong dãy Béc-nu-li với $P(A) = p$). Dễ dàng chứng minh khi $X \sim \mathcal{G}(p)$.

$$EX = \frac{1-p}{p} = \frac{q}{p}; \quad VX = \frac{1-p}{p^2} = \frac{q}{p^2}.$$

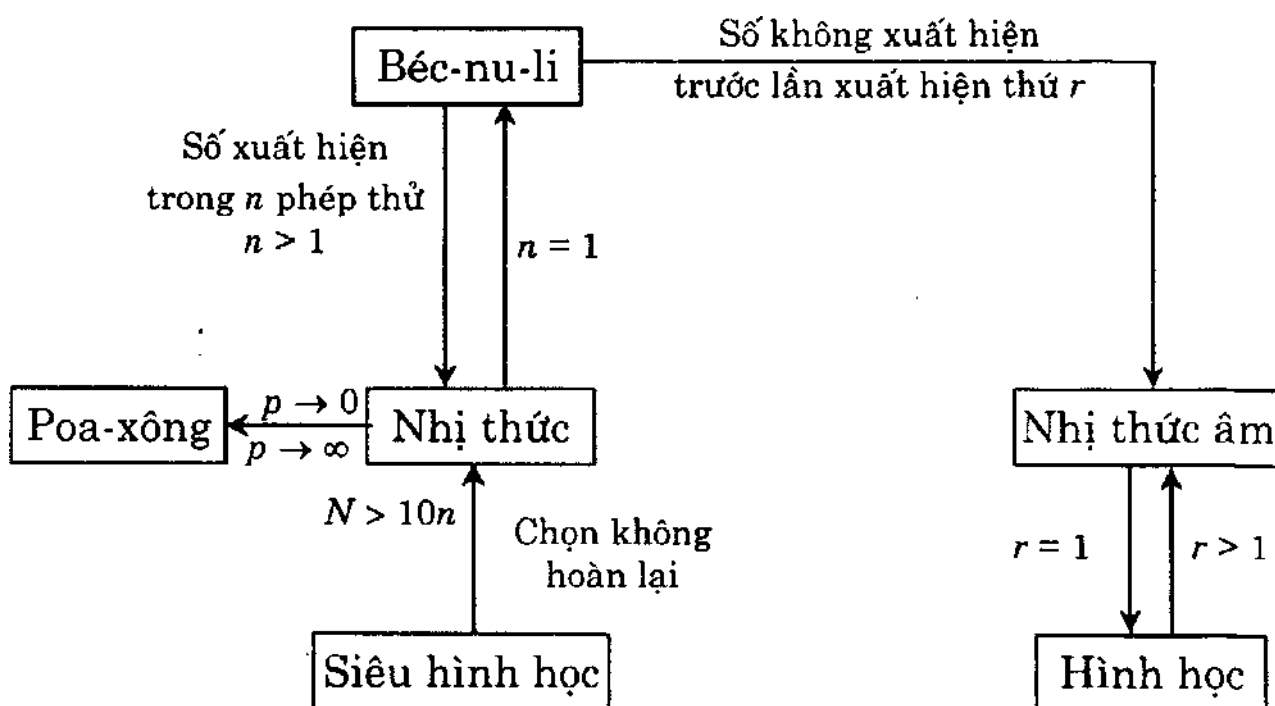
3. Định nghĩa 8. Biến ngẫu nhiên X được gọi là tuân theo *luật phân phối nhị thức âm*, ký hiệu là $X \sim \mathcal{NB}(r, p)$, nếu hàm xác suất của nó có dạng:

$$p(x) = C_{r+x-1}^x p^r (1-p)^x, \quad x = 0, 1, 2, \dots \quad (4.8)$$

Ý nghĩa của X chính là số lần không xuất hiện trước lần xuất hiện thứ r ($r > 0$) của một sự kiện A nào đó (trong dãy Béc-nu-li với $P(A) = p$). So sánh (4.7) và (4.8) ta thấy rằng phân phối hình học là trường hợp riêng của phân phối nhị thức âm khi $r = 1$. Cũng có thể chứng tỏ khi $X \sim \mathcal{NB}(r, p)$.

$$EX = \frac{rq}{p}; \quad VX = \frac{rq}{p^2} \quad (q = 1 - p).$$

Có thể tóm tắt các quan hệ của các phân phối rời rạc ở trên bằng sơ đồ sau:



Sơ đồ quan hệ giữa các phân phối rời rạc

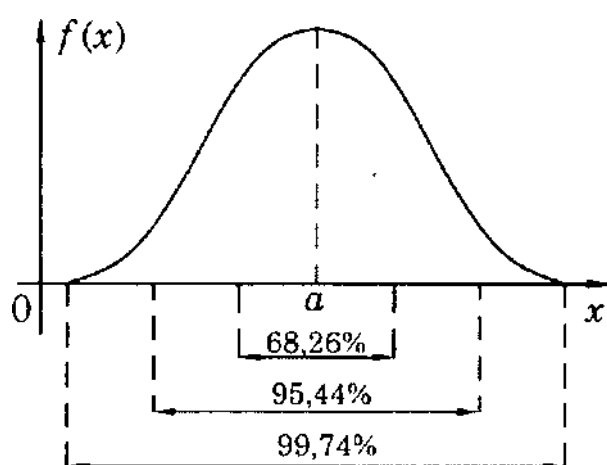
4.5. Phân phối chuẩn

Đây là phân phối liên tục quan trọng và có ứng dụng rộng rãi nhất, còn có tên gọi là *phân phối Gao-xơ*.

Định nghĩa 9. Biến ngẫu nhiên X được gọi là tuân theo *luật phân phối chuẩn*, ký hiệu là $X \sim \mathcal{N}(a, \sigma^2)$, nếu hàm mật độ của nó có dạng

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad x \in \mathbf{R}. \quad (4.9)$$

Dễ thấy hai tham số trong (4.9) là a và σ^2 cũng chính là hai số đặc trưng quan trọng EX và VX , còn σ là độ lệch chuẩn của X . Về mặt đồ thị, đường cong (4.9) có dạng hình chuông (xem hình 4.3). Từ hình vẽ 4.3 ta thấy hàm $f(x)$ trong (4.9) đối xứng qua $EX = a$, từ đó $medX = a$, đồng thời $modX = a$



Hình 4.3

do hàm đạt max tại $x = a$. Nếu ta lấy lân cận σ của a , thì phần diện tích chắn bởi $f(x)$, trục hoành và các đường $x = a \pm \sigma$ sẽ có diện tích bằng 68,26% (đơn vị diện tích). Đó cũng chính là $P(|X - a| < \sigma) = 68,26\%$. Tương tự ta cũng có

$$P(|X - a| < 2\sigma) = 95,44\%;$$

$$P(|X - a| < 3\sigma) = 99,74\%. \quad (4.10)$$

Công thức (4.10) cho ta thấy hầu chắc chắn biến ngẫu nhiên $X \sim \mathcal{N}(a, \sigma^2)$ sẽ nhận giá trị trong lân cận 3σ của kỳ vọng, sự kiện đó mang tên gọi *quy tắc 3 σ* rất quen thuộc trong các tính toán kỹ thuật. Qua hình 4.3 ta cũng thấy rõ nếu EX là đặc trưng định vị của phân phối, thì VX là đặc trưng độ tán xạ.

Nếu σ^2 càng lớn $f(x)$ phân tán nhiều hơn, đỉnh đồ thị càng thấp và tù hơn, đường cong tiệm cận tới trục hoành chậm hơn (chú ý là tổng diện tích chắn bởi $f(x)$ và trục Ox luôn bằng 1).

Thí dụ 4.4. Độ dài một chi tiết máy giả sử tuân theo luật phân phối chuẩn với trị trung bình 20cm và độ lệch chuẩn là 0,5. Hãy tính xác suất khi chọn ngẫu nhiên ra một chi tiết thì độ dài của nó:

- a) lớn hơn 20cm;
- b) bé hơn 19,5cm;
- c) lớn hơn 21,5cm.

Giải. Gọi X là độ dài của chi tiết máy chọn ra, rõ ràng $X \sim \mathcal{N}(20; 0,5^2)$.

a) Do phân phối đối xứng qua kỳ vọng nên $P(X > 20) = 0,5$.

b) Do $P(19,5 \leq X \leq 20,5) = 68,26\%$ (quy tắc 1σ) nên xác suất để X nằm ngoài khoảng đó là 31,74%. Do tính đối xứng $P(X < 19,5) = 15,87\%$ (và cũng bằng $P(X > 20,5)$).

c) Do cùng lý do như trên và dùng quy tắc 3σ ta có $P(X > 21,5) = (1 - 99,74\%)/2 = 0,13\% = 0,0013$ (xác suất không đáng kể).

Tuy nhiên trong thí dụ trên khó tìm được xác suất để độ dài X nằm trong một khoảng tùy ý. Có hai cách giải quyết hoặc dùng máy tính với các phần mềm tương ứng, hoặc sử dụng các bảng số có sẵn. Ở chương I §3 ta đã đưa vào khái niệm hàm Láp-la-xơ

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt. \quad (4.11)$$

Ta sẽ tìm cách biến đổi (4.9) và hàm phân phối tương ứng của $X \sim \mathcal{N}(a; \sigma^2)$ để có thể dùng được bảng số hàm trên. Để thấy từ (4.9), hàm phân phối của X có dạng:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-a)^2}{2\sigma^2}} dt. \quad (4.12)$$

Dùng phép biến đổi $z = \frac{t-a}{\sigma}$ ta có thể đưa (4.12) về dạng

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-a}{\sigma}} e^{-\frac{z^2}{2}} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{z^2}{2}} dz + \frac{1}{\sqrt{2\pi}} \int_0^{\frac{x-a}{\sigma}} e^{-\frac{z^2}{2}} dz. \end{aligned} \quad (4.13)$$

Mặt khác phép biến đổi biến trên sẽ ứng với phép biến đổi:

$$Z = \frac{X-a}{\sigma}; \quad (4.14)$$

từ đó do $X \sim \mathcal{N}(a; \sigma^2)$, nên $Z \sim \mathcal{N}(0; 1)$ với hàm mật độ:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \text{ (hàm Gao-xơ).}$$

Phân phối $\mathcal{N}(0; 1)$ sẽ có tên gọi là phân phối *chuẩn rút gọn*, hay phân phối *chuẩn chuẩn tắc* và động tác biến đổi (4.14) được gọi là *quy chuẩn*. Do kỳ vọng của Z bằng 0 nên

$$P(Z < 0) = \int_{-\infty}^0 f(z) dz = 0,5$$

và từ (4.13) ta có:

$$F(x) = \frac{1}{2} + \phi\left(\frac{x-a}{\sigma}\right). \quad (4.15)$$

Việc biết được hàm phân phối trong (4.15) cho phép chúng ta tính được mọi xác suất của X thông qua hàm Láp-la-xơ trong (4.11) được xác định trong bảng số 2. Chú ý rằng Z chỉ có phân

phôi chuẩn khi biến X tương ứng tuân theo luật chuẩn. tuy nhiên Z luôn có kỳ vọng 0 và phương sai 1.

Bây giờ giả sử ta muốn tính $P(\alpha \leq X < \beta)$, biết rằng $X \sim \mathcal{N}(\alpha; \sigma^2)$. Dùng tính chất của hàm phân phối của X ta có ngay, có tính đến (4.15):

$$P(\alpha \leq X < \beta) = F(\beta) - F(\alpha) = \Phi\left(\frac{\beta - \alpha}{\sigma}\right) - \Phi\left(\frac{\alpha - \alpha}{\sigma}\right). \quad (4.16)$$

Trong trường hợp đặc biệt nếu ta muốn tính $P(|X - \alpha| < \varepsilon)$ tùy ý, viết lại $|X - \alpha| < \varepsilon$ thành $\alpha - \varepsilon < X < \alpha + \varepsilon$ và từ đó:

$$P(|X - \alpha| < \varepsilon) = \Phi\left(\frac{\varepsilon}{\sigma}\right) - \Phi\left(-\frac{\varepsilon}{\sigma}\right) = 2\Phi\left(\frac{\varepsilon}{\sigma}\right).$$

Với sự kiện này, các kết quả được tính trong thí dụ 4.4 là các trường hợp riêng ứng với $\varepsilon = \sigma$ và 3σ . Nếu ta chọn ε tùy ý, chẳng hạn $\varepsilon = 1,25$ (dung sai của máy) và muốn tính xác suất $P(|X - 20| < 1,25)$ với $\alpha = 20$ chính là độ dài quy định, khi đó theo công thức trên:

$$\begin{aligned} P(|X - 20| < 1,25) &= 2\Phi\left(\frac{1,25}{0,5}\right) = 2\Phi(2,5) \\ &= 2 \cdot 0,4938 = 0,9876. \end{aligned}$$

Ở đây xác suất này có ý nghĩa là tỷ lệ chính phẩm của chiếc máy đã cho bằng 98,76%.

Tổng của n biến ngẫu nhiên độc lập cùng có phân phối chuẩn vẫn là một biến ngẫu nhiên chuẩn (mà ta có thể chứng minh bằng các kết quả ở chương III). Từ đó nếu $X_i \sim \mathcal{N}(\alpha; \sigma^2)$ $\forall i = \overline{1, n}$, và độc lập, theo các tính chất của kỳ vọng và phương sai:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \sim \mathcal{N}\left(\alpha, \frac{\sigma^2}{n}\right); \quad (4.17a)$$

tương đương với điều đó nếu đặt:

$$Z = \frac{\bar{X} - a}{\sigma} \sqrt{n}, \text{ thì } Z \sim \mathcal{N}(0; 1). \quad (4.17b)$$

Phân phối chuẩn có thể dùng xấp xỉ khá tốt cho một số phân phối rời rạc. Đối với phân phối nhị thức, khi tham số p không quá gần 0 hoặc 1 và n khá lớn, $\mathcal{B}(n, p)$ sẽ rất gần với $\mathcal{N}(np; npq)$; việc xấp xỉ sẽ rất tốt nếu $np \geq 5$ khi $p \leq 0,5$ hoặc $n(1 - p) \geq 5$ khi $p \geq 0,5$. Từ đó nếu $X \sim \mathcal{B}(n, p)$ và có các điều kiện ở trên thì (xem §3 chương I):

$$P(\alpha \leq X \leq \beta) \approx \Phi\left(\frac{\beta - np}{\sqrt{npq}}\right) - \Phi\left(\frac{\alpha - np}{\sqrt{npq}}\right). \quad (4.18)$$

$$P(X = \alpha) \approx \frac{\varphi\left(\frac{\alpha - np}{\sqrt{npq}}\right)}{\sqrt{npq}}, \quad \varphi - \text{hàm Gao-xơ}.$$

Trong trường hợp như vậy người ta nói rằng luật nhị thức *hội tụ theo luật* đến luật chuẩn chuẩn tắc và viết:

$$\frac{X - np}{\sqrt{npq}} \xrightarrow[n \rightarrow \infty]{L} \mathcal{N}(0; 1).$$

Thí dụ 4.5. Xét $X \sim \mathcal{B}(20; 0,4)$, tính $P(4 \leq X \leq 13)$.

Giải. Theo (4.18):

$$\begin{aligned} P(4 \leq X \leq 13) &\approx \Phi\left(\frac{13 - 8}{\sqrt{4,8}}\right) - \Phi\left(\frac{4 - 8}{\sqrt{4,8}}\right) = \Phi(2,28) + \Phi(1,83) \\ &= 0,4884 + 0,4664 = 0,9548. \end{aligned}$$

Nhưng do $n = 20$ vẫn chưa thật lớn, trong thực hành người ta hiệu chỉnh (4.18) như sau:

$$P(\alpha \leq X < \beta) = \Phi\left(\frac{\beta + 0,5 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{\alpha - 0,5 - np}{\sqrt{npq}}\right).$$

Việc cộng thêm vào +0,5 và -0,5 chính là yếu tố hiệu chỉnh khi xấp xỉ một biến rời rạc bằng biến liên tục. Từ đó

$$P(4 \leq X \leq 13) = \phi(2,51) + \phi(1,60) = 0,9743.$$

Để ý là kết quả thật của xác suất này là 0,978.

Người ta cùng chứng minh được rằng, nếu $X \sim \mathcal{P}(\lambda)$ thì:

$$\frac{X - \lambda}{\sqrt{\lambda}} \xrightarrow[\lambda \rightarrow \infty]{L} \mathcal{N}(0; 1).$$

4.6. Các phân phối liên tục khác

Người ta đã thấy rằng nhiều phân phối liên tục được cảm sinh trực tiếp bởi phân phối chuẩn (kể cả chuẩn). Trong mục này ta sẽ xét một số phân phối quan trọng hay dùng trong thống kê. Các phân phối khác có thể tham khảo trong bảng thống kê ở cuối tiết này.

1. *Phân phối χ^2* với n bậc tự do, ký hiệu là $\chi^2(n)$, có thể được định nghĩa bằng việc xác định hàm mật độ:

$$f(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, \quad x > 0, n > 0, \quad (4.19)$$

trong đó hàm gam-ma đã được xét trong giải tích

$$\Gamma(x) = \int_0^x t^{x-1} e^{-t} dt, \quad x > 0$$

có các tính chất $\forall i$ nguyên

$$(i) \quad \Gamma(i+1) = i! \quad (i \geq 0);$$

$$(ii) \quad \Gamma\left(\frac{i}{2}\right) = \left(\frac{i}{2} - 1\right) \left(\frac{i}{2} - 2\right) \dots \frac{3}{2} \cdot \frac{1}{2} \sqrt{\pi} \quad (i > 2, \text{ lẻ});$$

$$(iii) \quad \Gamma(x) = (x-1)\Gamma(x-1), \quad x \in \mathbf{R}.$$

Tuy nhiên cách định nghĩa này khá phức tạp và không cho ta cách xác định rõ ràng phân phối χ^2 xuất phát từ phân phối chuẩn.

Định nghĩa 10. Xét n biến ngẫu nhiên độc lập $X_i \sim \mathcal{N}(0; 1)$, $i = \overline{1, n}$. Khi đó biến ngẫu nhiên:

$$U_n = \sum_{i=1}^n X_i^2 \sim \chi^2(n). \quad (4.20)$$

Rõ ràng (4.20) cho ta cách nhận biết đơn giản một biến có phân phối khi bình phương xuất phát từ n biến độc lập cùng phân phối chuẩn chuẩn tắc. Dạng đồ thị của hàm mật độ (4.19) cho ở hình 4.4. Các số đặc trưng quan trọng là

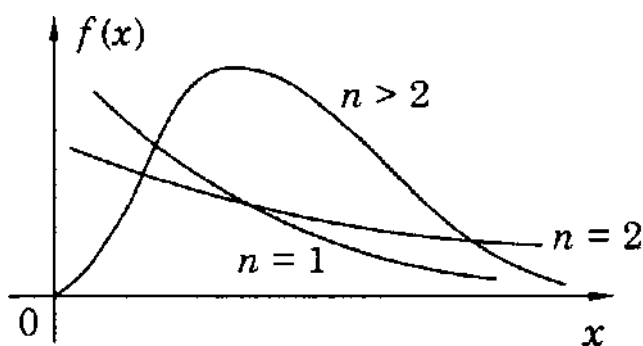
$$EU_n = n,$$

$$VU_n = 2n.$$

Phân phối χ^2 có một vài tính chất quan trọng:

a) Nếu $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$, và độc lập $\Rightarrow X + Y \sim \chi^2(n + m)$.

b) $\frac{U_n - n}{\sqrt{2n}} \xrightarrow[n \rightarrow \infty]{L} \mathcal{N}(0; 1).$



Hình 4.4

Ngoài ra có một hệ quả quan trọng sẽ được dùng nhiều trong thống kê: Nếu ta có n biến độc lập $X_i \sim \mathcal{N}(a; \sigma^2)$; $i = \overline{1, n}$,

và $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ thì $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1)$. (4.21)

Trong (4.21) do ta thay thế a bằng $\bar{X}$, vì vậy bậc tự do của phân phối χ^2 đã bớt đi 1. Việc tính toán với phân phối $\chi^2(n)$ đưa về sử dụng bảng 4 trong phụ lục hoặc dùng máy tính.

2. Ta sẽ dùng cách định nghĩa ở trên để xác định luật phân phối Stiu-đơn với n bậc tự do, ký hiệu là $t(n)$.

Định nghĩa 11. Cho X và Y là hai biến ngẫu nhiên độc lập tuân theo luật $\mathcal{N}(0; 1)$ và $\chi^2(n)$ tương ứng. Khi đó biến

$$T_n = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n). \quad (4.22)$$

Hàm mật độ của phân phối $t(n)$ cho ở bảng cuối tiết, đồ thị của nó có dạng rất giống với đường cong chuẩn. Các số đặc trưng của T_n là (chú ý hàm mật độ đối xứng):

$$ET_n = 0 \quad (n > 1);$$

$$VT_n = \frac{n}{n-2} \quad (n > 2).$$

Phân phối Stiu-đơn có tính chất quan trọng:

$$T_n \xrightarrow[n \rightarrow \infty]{L} \mathcal{N}(0; 1).$$

Trong thực hành, khi $n \geq 30$, đồ thị của đường cong mật độ phân phối $t(n)$ đã rất gần với $\mathcal{N}(0; 1)$. Chú ý khi $n = 1$, ta có phân phối Cô-si, đó là phân phối không có mô men nào. Bảng phân vị $t(n)$ cho ở phần phụ lục (bảng 3).

3. Tỷ số của hai biến ngẫu nhiên độc lập có phân phối χ^2 cho ta một phân phối mới (ký hiệu là $\mathcal{F}(n, m)$ – phân phối Phi-sơ – Sne-đơ-co với n và m bậc tự do).

Định nghĩa 12. Cho X và Y là hai biến ngẫu nhiên độc lập tuân theo luật $\chi^2(n)$ và $\chi^2(m)$ tương ứng. Khi đó biến

$$U = \frac{X/n}{Y/m} \sim \mathcal{F}(n, m). \quad (4.23)$$

Hàm mật độ của phân phối $\mathcal{F}(n, m)$ cho ở bảng cuối tiết. Đồ thị của hàm đó có dạng gần giống với đường cong mật độ χ^2 . Biến có các đặc trưng:

$$EU = \frac{m}{m-2} \quad (m > 2),$$

$$VU = \frac{2m^2(n+m-2)}{n(m-4)(m-2)^2} (m > 4).$$

Để ý từ (4.23), do vai trò của X và Y có thể đổi cho nhau nên nếu $U \sim \mathcal{F}(n, m)$, $V \sim \mathcal{F}(m, n)$ thì U và $\frac{1}{V}$ có cùng phân phối. Ngoài ra nếu $n = 1$, từ (4.22) thấy ngay rằng T_m tuân theo luật $\mathcal{F}(1, m)$.

4. Định nghĩa 13. X tuân theo luật *phân phối Gam-ma*, ký hiệu là $X \sim \gamma(r, \lambda)$, nếu hàm mật độ có dạng:

$$f(x) = \frac{\lambda}{\Gamma(x)} e^{-\lambda x} x^{r-1}, \quad r > 0, \lambda > 0, x > 0. \quad (4.24)$$

(hàm $\Gamma(x)$ đã xác định ở trên).

Các số đặc trưng của $X \sim \gamma(r, \lambda)$:

$$EX = \frac{r}{\lambda}; \quad VX = \frac{r}{\lambda^2}.$$

Ta đề ý một số tính chất quan trọng của phân phối Gam-ma

- a) Nếu $X \sim \gamma(p, \lambda)$, $Y \sim \gamma(q, \lambda)$ và độc lập $\Rightarrow X + Y \sim \gamma(q + p, \lambda)$.
- b) Nếu $X \sim \gamma(r, 1)$ thì

$$\frac{X - r}{\sqrt{r}} \xrightarrow[r \rightarrow \infty]{L} \mathcal{N}(0; 1).$$

Để ý nếu $r = 1$, ta có phân phối mũ $\mathcal{E}(\lambda)$ (xem thí dụ 2.8 chương này) có nhiều ứng dụng trong lý thuyết độ tin cậy.

5. Bảng tổng kết các phân phối liên tục.

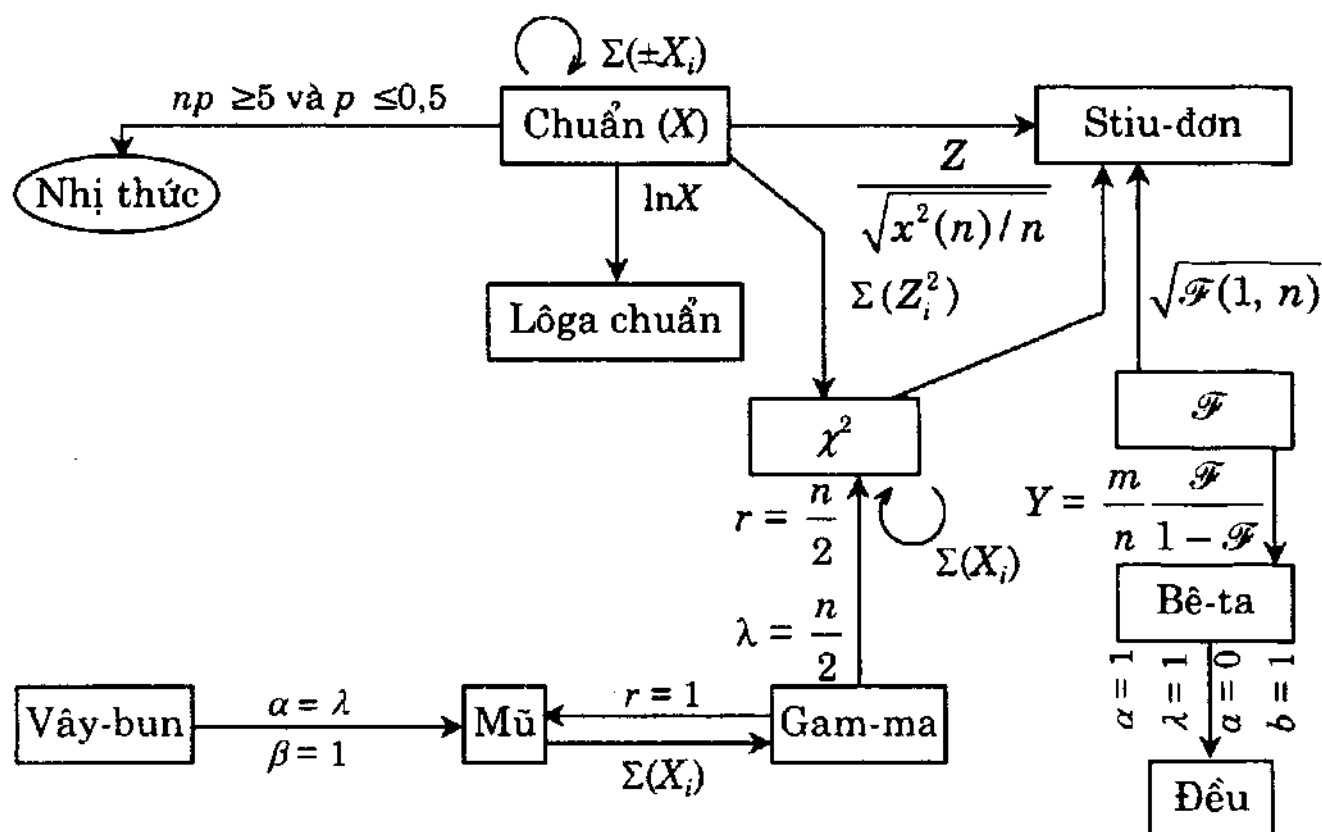
Bảng tổng kết phân phối liên tục

	Hàm mật độ $f(x)$	EX	VX	
n	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2\right], \sigma > 0$	a	σ^2	σ
	$\frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, x > 0, n > 0$	n	$2n$	
m	$\frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\left[\frac{n+1}{2}\right]}, n > 0$	$0 (n > 1)$	$\frac{n}{n-2} (n > 2)$	
σ	$\frac{n^{\frac{n}{2}} m^{\frac{m}{2}} \Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{m}{2}\right)} x^{\frac{n-2}{2}} (m+nx)^{-\frac{n+m}{2}}$ $x > 0$ $m, n > 0$	$\frac{m}{m-2} (m > 2)$	$\frac{2m^2 (n+m-2)}{n(m-2)^2 (m-4)}$ $(m > 4)$	

Bảng tổng kết phối liên tục (tiếp)

na	$\frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x}, \lambda, r > 0$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
	$\lambda e^{-\lambda x}, \lambda, x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
	$\frac{1}{b-a} \quad a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	$\alpha \lambda x^{\lambda-1} \exp(-\alpha x^\lambda) \quad \alpha, \lambda > 0$	$\alpha^{-\frac{1}{\lambda}} \Gamma\left(1 + \frac{1}{\lambda}\right)$	$\alpha^{-\frac{2}{\lambda}} \left[\Gamma\left(1 + \frac{2}{\lambda}\right) - \Gamma^2\left(1 + \frac{1}{\lambda}\right) \right]$
	$\frac{1}{\sigma\sqrt{2\pi}} x^{-1} \exp\left[-\frac{(\ln x - \alpha)^2}{2\sigma^2}\right] \quad \begin{matrix} x > 0 \\ \sigma > 0 \end{matrix}$	$\exp\left(\alpha + \frac{\sigma^2}{2}\right)$	$\exp(2\alpha + \sigma^2) [\exp(\sigma^2) - 1]$
	$\frac{\Gamma(\alpha + \lambda)}{\Gamma(\alpha)\Gamma(\lambda)} x^{\alpha-1} (1-x)^{\lambda-1}, \quad 0 < x < 1, \alpha, \lambda > 0$	$\frac{\alpha}{\alpha + \lambda}$	$\frac{\alpha\lambda}{(\alpha + \lambda)^2 (\alpha + \lambda + 1)}$

Từ bảng tổng kết trên, có thể xây dựng sơ đồ quan hệ sau:



Sơ đồ quan hệ giữa các phân phối liên tục

BÀI TẬP

- Một xí nghiệp có 3 xe ô tô với các xác suất làm việc tốt trong ngày là 0,99; 0,995 và 0,999. Tìm bảng phân phối xác suất của số xe hỏng trong ngày.
- Hai cầu thủ thay nhau ném bóng vào rổ cho đến khi nào trúng rổ thì dừng ném, biết rằng xác suất ném trúng của mỗi người tương ứng là 0,6 và 0,7 (trong mỗi lần ném). Tìm luật phân phối xác suất của:
 - số lần ném của cầu thủ thứ nhất;
 - số lần ném của cả hai cầu thủ.
- Một tổ có 6 nam và 4 nữ. Chọn ngẫu nhiên ra 3 người. Tìm luật phân phối của số nữ trong nhóm được chọn.

4. Xác suất chữa khỏi bệnh A của một bác sĩ là 0,8. Tìm luật phân phối của số được chữa khỏi bệnh trong một nhóm bệnh nhân gồm 5 người do bác sĩ đó điều trị.
5. Cho bảng phân phối xác suất của một biến X nào đó có dạng:

x	1	2	3	4	5
$p(x)$	a	$2a$	a	$3a$	$2a$

- (a là tham số). Hãy xác định: a) tham số a ; b) giá trị k nhỏ nhất sao cho $P(X \leq k) > \frac{1}{2}$.
6. Một vùng dân cư có tỷ lệ sốt rét là 5%. Cần chọn ra ít nhất bao nhiêu người để với xác suất 95% trong số đó có ít nhất 1 người mắc bệnh sốt rét?
7. Xác suất bắn trúng đích của một khẩu súng là p . Tiến hành bắn liên tiếp trong điều kiện như nhau đến khi trúng thì dừng bắn. Tìm số đạn trung bình phải bắn.
8. Cho hàm mật độ của biến ngẫu nhiên X có dạng:

$$f(x) = \frac{e^{-x}}{(1 + e^{-x}) \ln 2}, \quad x > 0.$$

Hãy tính EX và VX .

9. Cho biến X có hàm phân phối có dạng:

$$F(x) = \begin{cases} 0, & x \leq -2, \\ a + b \arcsin \frac{x}{2}, & -2 < x \leq 2, \\ 1, & x > 2. \end{cases}$$

- a) Xác định a và b ; b) Tìm hàm mật độ $f(x)$; tìm các số đặc trưng EX , VX , $mod X$, $med X$.
10. Năng suất lúa ở một địa phương là biến ngẫu nhiên có phân phối chuẩn với kỳ vọng 42 tạ/ha và $\sigma = 3$ tạ/ha. Tìm xác suất

để khi gặt ngẫu nhiên 3 thửa ruộng thì có 2 thửa có năng suất sai lệch so với trung bình không quá 1 tạ/ha.

11. Kiểm tra chất lượng 100 sản phẩm với tỷ lệ chính phẩm 0,95. Tìm xác suất để số sản phẩm đạt tiêu chuẩn nằm trong khoảng từ 900 đến 980.
12. Ở một thửa ruộng trung bình trong một giờ tìm được 60 con sâu. Tìm xác suất trong vòng 1 phút không tìm thấy con sâu nào.
13. Tìm mật của biến X tuân theo luật nhị thức.
14. Một viên đạn có tầm xa trung bình là 300 m. Giả sử tầm xa đó là một biến ngẫu nhiên tuân theo luật chuẩn với $\sigma = 10$. Hãy tìm tỉ lệ đạn bay quá tầm xa trung bình từ 15 đến 30 m.
15. Biên độ dao động của thành tàu thủy là biến ngẫu nhiên X tuân theo luật phân phối Rê-le

$$f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} \quad (x > 0).$$

Tìm xác suất để biên độ dao động lớn hơn trung bình của nó.

16. Từ kết quả 2 lần thí nghiệm ta có 2 đại lượng ngẫu nhiên độc lập cùng phân phối χ^2 với bậc tự do tương ứng là 4 và 6. Tìm xác suất để đại lượng thứ nhất bé hơn 3 lần đại lượng thứ hai.
17. Cho các $X_i \sim \mathcal{N}\left(0, \frac{1}{3}\right)$, $i = \overline{1,5}$; $Y_i \sim \mathcal{N}\left(0, \frac{1}{4}\right)$, $i = \overline{1,11}$; và giả sử chúng độc lập. Tính $P\left(3\sum_{i=1}^5 X_i^2 > 2\sum_{i=1}^{11} Y_i^2\right)$.
18. Cho $X \sim \mathcal{N}(3, 1)$, $Y \sim \mathcal{N}(4, 2)$ độc lập. Tìm các xác suất
 - a) $X > Y$;
 - b) $X > 2Y$.

Chương III

BIẾN NGẪU NHIÊN NHIỀU CHIỀU

§1. LUẬT PHÂN PHỐI CỦA BIẾN NGẪU NHIÊN NHIỀU CHIỀU

1.1. Các khái niệm cơ sở

1. Ở hai chương trước ta đã nghiên cứu bản chất xác suất của một biến ngẫu nhiên riêng rẽ. Nhưng trong thực tế nhiều khi phải xét đồng thời nhiều biến khác nhau có quan hệ tương hỗ và dẫn tới khái niệm *véc tơ ngẫu nhiên* hay *biến ngẫu nhiên nhiều chiều*. Những thí dụ về các biến nhiều chiều rất phổ biến, chẳng hạn khi nghiên cứu một chỉ tiết máy, ta quan tâm đồng thời đến nhiều khía cạnh khác nhau như trọng lượng, kích thước (riêng nó đã là nhiều chiều), chất lượng, chất liệu... Việc nghiên cứu riêng rẽ từng khía cạnh có thể cho ta các thông tin không đầy đủ.

Để cho đơn giản, ta nghiên cứu biến ngẫu nhiên 2 chiều (X, Y) , trong đó X và Y là các biến một chiều. Hầu hết các kết quả có thể mở rộng khá dễ dàng cho biến n chiều. Nếu X và Y là rời rạc, ta có biến ngẫu nhiên hai chiều rời rạc; nếu chúng liên tục, ta có biến hai chiều liên tục. Sẽ phức tạp hơn một chút là một biến rời rạc và một biến liên tục mà ta không xét ở đây.

2. Ta phát triển khái niệm hàm phân phối xác suất cho biến ngẫu nhiên hai chiều. Xét hai sự kiện $A = \{X < x\}$ và $B = \{Y < y\}$.

Định nghĩa 1. *Hàm phân phối xác suất của biến hai chiều (X, Y) được xác định như sau:*

$$F(x, y) = P(AB) = P(X < x; Y < y), \quad x, y \in \mathbb{R}. \quad (1.1)$$

Trong nhiều tài liệu hàm $F(x, y)$ trong (1.1) được gọi là *hàm phân phối đồng thời* của hai biến X và Y . Đây là một hàm thực hai biến và về mặt hình học ta có thể biểu diễn tập xác định của $F(x, y)$ bằng các điểm trên mặt phẳng tọa độ Đề-các.

Tương tự như trường hợp một chiều, ta có thể dẫn ra một số tính chất của hàm phân phối hai chiều

(i) $1 \geq F(x, y) \geq 0$;

(ii) $F(x, y)$ không giảm theo từng đôi số;

(iii) $F(-\infty, y) = F(x, -\infty) = 0$; $F(+\infty, +\infty) = 1$ (giá trị $\pm\infty$ hiểu theo nghĩa lấy giới hạn);

(iv) Với $x_1 < x_2$; $y_1 \leq y_2$ ta luôn có

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

Đó chính là xác suất để điểm ngẫu nhiên (X, Y) rơi vào miền chữ nhật ABCD (xem hình 1.1).

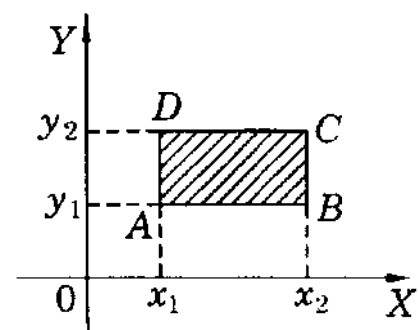
Để ý rằng

$$F(x; +\infty) = P(X < x; Y < +\infty) = P(X < x) = F_1(x);$$

$$F(+\infty; y) = P(X < +\infty; Y < y) = P(Y < y) = F_2(y)$$

là các phân phối của riêng từng thành phần X và Y tương ứng; chúng được gọi là các *phân phối biên* của biến hai chiều (X, Y) . Đó cũng chính là các phân phối (một chiều) thông thường của X và Y .

3. Ở chương I ta đã làm quen với khái niệm độc lập của hai sự kiện A và B : chúng được gọi là độc lập nếu $P(AB) = P(A)P(B)$. Áp dụng khái niệm này vào (1.1) ta có



Hình 1.1

Định nghĩa 2. Hai biến ngẫu nhiên X và Y được gọi là *độc lập* nếu

$$F(x, y) = F_1(x) F_2(y). \quad (1.2)$$

Tất nhiên nếu X và Y độc lập, ta có thể nghiên cứu riêng rẽ từng biến theo các phương pháp đã có và từ các phân phối biên của X và Y có thể xác định được phân phối của (X, Y) theo (1.2). Tuy nhiên chúng không đủ để xác định phân phối đồng thời nếu X và Y không độc lập.

1.2. Phân phối xác suất của biến ngẫu nhiên hai chiều rời rạc

1. Giống như trường hợp một chiều ta tìm cách xác định biến hai chiều rời rạc qua bảng phân phối xác suất.

Định nghĩa 3. *Bảng phân phối xác suất* của biến (X, Y) rời rạc là

$x \backslash y$	y_1	y_2	...	y_j	...	y_m	$\sum_j$
x_1	p_{11}	p_{12}	...	p_{1j}	...	p_{1m}	$p_1(x_1)$
x_2	p_{21}	p_{22}		p_{2j}	...	p_{2m}	$p_2(x_2)$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
x_i	p_{i1}	p_{i2}	...	p_{ij}		p_{im}	$p_i(x_i)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_n	p_{n1}	p_{n2}	...	p_{nj}	...	p_{nm}	$p_n(x_n)$
$\sum_i$	$p_2(y_1)$	$p_2(y_2)$	...	$p_2(y_j)$	...	$p_2(y_m)$	1

trong đó $p_{ij} = P(X = x_i; Y = y_j)$ là xác suất đồng thời để X lấy giá trị x_i , $i = \overline{1, n}$, và Y lấy giá y_j , $j = \overline{1, m}$. Bảng này có thể trở thành vô hạn khi n, m nhận giá trị ∞ .

Giống như trong trường hợp một chiều, ta xác định hàm xác suất $p(x, y)$ sao cho $p(x_i, y_j) = p_{ij}$, $i = \overline{1, n}$, $j = \overline{1, m}$. Hàm này có tính chất:

$$(i) p_{ij} \geq 0 \forall i, j;$$

$$(ii) \sum_i \sum_j p_{ij} = 1 \text{ (tổng hiệu theo nghĩa lấy theo } \forall i, j).$$

Từ định nghĩa 3, ta có thể tìm được hàm phân phối xác suất được đưa vào bằng (1.1):

$$F(x, y) = \sum_{x_i < x} \sum_{y_j < y} p_{ij}. \quad (1.3)$$

Các phân phối biên của biến hai chiều đang xét được xác định từ:

$$P(X = x_i) = p_1(x_i) = \sum_j p_{ij}, i = \overline{1, n}; \quad (1.4a)$$

$$P(Y = y_j) = p_2(y_j) = \sum_i p_{ij}, j = \overline{1, m} \quad (1.4b)$$

Thí dụ 1.1. Cho bảng phân phối đồng thời của X và Y :

$x \backslash y$		1	2	3
1		0,10	0,25	0,10
2		0,15	0,05	0,35

Tìm luật phân phối xác suất của các biến X và Y , sau đó tính $F(2, 3)$.

Giải. Lấy tổng hàng và tổng cột tương ứng của bảng số, ta có các phân phối biên cần tìm (xem (1.4)):

x	1	2	y	1	2	3
$p_1(x)$	0,45	0,55	$p_2(y)$	0,25	0,30	0,45

Việc tính $F(2, 3)$ dựa vào (1.3):

$$F(2, 3) = \sum_{x_i < 2} \sum_{y_j < 3} p_{ij} = p_{11} + p_{12} = 0,35.$$

Từ định nghĩa 2, hai biến rời rạc X, Y được gọi là độc lập nếu với mọi cặp giá trị x_i, y_j , ta luôn có

$$p_{ij} = p_1(x_i)p_2(y_j), i = \overline{1, n}, j = \overline{1, m} \quad (1.5)$$

Rõ ràng trong thí dụ 1.1 ta thấy $p_{11} = 0,10 \neq p_1(1)p_2(1) = 0,1125$; vậy hai biến X và Y ở đây không độc lập do (1.5) bị phá khi $i = j = 1$. Có thể chứng tỏ (1.2) và (1.5) là tương đương.

2. Bây giờ giả sử Y lấy một giá trị cố định nào đó và ta muốn quan tâm đến luật phân phối xác suất của X có bị ảnh hưởng không. Theo công thức xác suất có điều kiện ở chương I

$$P(X = x_i | Y = y_k) = \frac{P(X = x_i; Y = y_k)}{P(Y = y_k)}, i = \overline{1, n}. \quad (1.6)$$

Như vậy (1.6) cho phép ta định nghĩa luật phân phối có điều kiện của X biết Y nhận giá trị y_k cụ thể. Tương tự có thể xác định luật phân phối có điều kiện của Y biết X nhận một giá trị cụ thể nào đó.

Thí dụ 1.2. Tìm phân phối có điều kiện của X biết rằng $Y = 1$ trong bài toán ở thí dụ 1.1.

Giải. Theo (1.6)

$$P(X = 1 | Y = 1) = \frac{P(X = 1, Y = 1)}{P(Y = 1)} = \frac{p_{11}}{p_2(1)} = \frac{0,10}{0,25} = 0,4;$$

$$P(X = 2 | Y = 1) = \frac{p_{21}}{p_2(1)} = \frac{0,15}{0,25} = 0,6.$$

Bảng phân phối xác suất có điều kiện của X biết $Y = 1$ là:

x	1	2
$p(x Y = 1)$	0,40	0,60

Tổng quát, nếu ta biết một điều kiện C_Y nào đó của Y , thì phân phối có điều kiện của X biết C_Y sẽ là:

$$P(X = x | C_Y) = \frac{P(X = x; C_Y)}{P(C_Y)}.$$

Chẳng hạn nếu ta biết $y_1 \leq Y \leq y_2$ với y_1 và y_2 nào đó, thì:

$$P(X = x | y_1 \leq Y \leq y_2) = \frac{P(X = x; y_1 \leq Y \leq y_2)}{P(y_1 \leq Y \leq y_2)}. \quad (1.7)$$

Để ý rằng trong (1.7) biến ngẫu nhiên Y có thể rời rạc hoặc liên tục.

1.3. Phân phối xác suất của biến ngẫu nhiên hai chiều liên tục

1. Khái niệm hàm phân phối xác suất của biến hai chiều (X, Y) liên tục đã được xét ở định nghĩa 1 (công thức (1.1)). Ta sẽ đưa ra khái niệm hàm mật độ của (X, Y) như sau

Định nghĩa 4. Nếu hàm phân phối $F(x, y)$ của biến hai chiều (X, Y) có dạng:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv, \quad (1.8a)$$

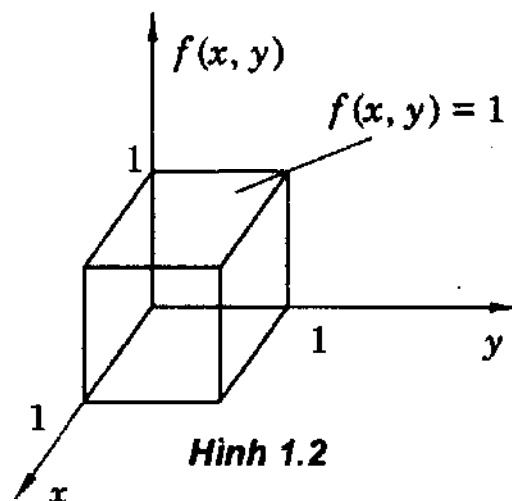
trong đó $f(x, y) > 0$, thì hàm $f(x, y)$ được gọi là *hàm mật độ* của biến (X, Y) (hay *hàm mật độ đồng thời* của X và Y).

Về mặt hình học, hàm $f(x, y)$ có thể xem như là một mặt cong trong $\mathbb{R}^3$ và được gọi là mặt phân phối xác suất. Nếu $f(x, y)$ liên tục theo cả hai biến thì:

$$f(x, y) = \frac{\partial F(x, y)}{\partial x \partial y}. \quad (1.8b)$$

Thí dụ 1.3. Cho hàm mật độ đồng thời của X và Y là $f(x, y) = 1$, với $0 \leq x, y \leq 1$. Vẽ hàm $f(x, y)$ và tính hàm phân phối đồng thời $F(x, y)$.

Giải. Mặt cong phân phối cho trên hình 1.2. Để ý là $f(x, y) \neq 0$ chỉ với các (x, y) thuộc khoảng vuông $[0; 1] \times [0; 1]$. Hàm phân phối $F(x, y)$ được tính theo (1.8a):



$$F(x, y) = \begin{cases} 0, & \text{nếu } x \leq 0 \text{ hoặc } y \leq 0; \\ xy, & \text{nếu } 0 \leq x \leq 1 \text{ và } 0 \leq y \leq 1; \\ x, & \text{nếu } 0 \leq x \leq 1 \text{ và } y > 1; \\ y, & \text{nếu } x > 1 \text{ và } 0 \leq y \leq 1; \\ 1, & \text{nếu } x > 1 \text{ và } y > 1. \end{cases}$$

Dạng hàm phân phối thường khá phức tạp, nên người ta hay dùng hàm mật độ. Đây là thí dụ về *phân phối đều hai chiều*, tổng quát hóa phân phối đều liên tục $\mathcal{U}([0; 1])$ đã xét ở chương II.

Hàm mật độ của biến hai chiều (X, Y) có các tính chất quan trọng sau:

(i) $f(x, y) \geq 0$;

(ii) $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$;

(iii) $P[(X, Y) \in \mathcal{D}] = \iint_{\mathcal{D}} f(x, y) dx dy$.

Chẳng hạn, trong thí dụ 1.3, ta muốn tính $P(0,2 \leq X < 0,7; 0,25 \leq Y < 0,45)$, đó chính là tích phân kép của $f(x, y)$

$$\int_{0,2}^{0,7} \int_{0,25}^{0,45} dx dy = (0,7 - 0,2)(0,45 - 0,25) = 0,1.$$

Về mặt hình học, đó là thể tích một hộp chữ nhật có đáy trên nằm trong mặt phân phối $f(x, y) = 1$. Trong trường hợp tổng quát, $\mathcal{D}$ sẽ là một miền nào đó thuộc mặt xOy và $P[(X, Y) \in \mathcal{D}]$ bằng thể tích của hộp chữ nhật cong giới hạn bởi phần mặt xác suất $f(x, y)$ và có đáy là hình chiếu của mặt đó trên mặt xOy (chính là miền $\mathcal{D}$).

Tương tự như ở mục 1.1, ta xác định các *hàm mật độ biên* của biến (X, Y) :

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{+\infty} f(x, y) dy; \\ f_2(y) &= \int_{-\infty}^{+\infty} f(x, y) dx; \end{aligned} \quad (1.9)$$

Để ý $f_1(x)$ cũng chính bằng $\frac{\partial F_1}{\partial x}$ và là mật độ của biến thành phần X , tương tự đối với $f_2(y)$.

Thí dụ 1.4. Tìm các hàm mật độ biên của biến (X, Y) có hàm mật độ hai chiều $f(x, y) = \frac{1}{\pi^2 (1+x^2)(1+y^2)}$, $x, y \in \mathbf{R}$.

Giải. Dễ thấy theo (1.9)

$$f_1(x) = \frac{1}{\pi^2} \int_{-\infty}^{+\infty} \frac{1}{(1+x^2)(1+y^2)} dy = \frac{1}{\pi(1+x^2)}.$$

Do tính đối xứng, ta có ngay $f_2(y) = \frac{1}{\pi(1+y^2)}$.

2. Tương tự như (1.2), hai biến ngẫu nhiên được gọi là *độc lập*, nếu

$$f(x, y) = f_1(x)f_2(y). \quad (1.10)$$

Nếu mật độ đồng thời của X và Y không bằng tích các mật độ biên f_1 và f_2 , ta nói X và Y không độc lập. Trong trường hợp đó có thể đưa vào khái niệm *hàm mật độ có điều kiện* của thành phần X biết $Y = y$, ký hiệu là

$$\varphi(x | y) = \frac{f(x, y)}{f_2(y)} = \frac{f(x, y)}{\int_{-\infty}^{+\infty} f(x, y) dx}; \quad (1.11)$$

tương tự như $\psi(y|x)$ là hàm mật độ có điều kiện của Y biết $X = x$ cụ thể nào đó, nó sẽ bằng $f(x, y)/f_1(x)$. Chú ý rằng các mật độ có điều kiện cũng thoả mãn các tính chất của hàm mật độ bình thường.

Thí dụ 1.5. Cho hàm mật độ đồng thời $f(x) = x + y$, $0 \leq x, y \leq 1$. Xác định các hàm mật độ có điều kiện.

Giải. Để có thể dùng được (1.11), trước hết ta phải tính các $f_1(x)$ và $f_2(y)$ (là các mật độ biên, xem (1.9)):

$$f_1(x) = \int_0^1 (x + y) dy = x + \frac{1}{2}, \quad 0 \leq x \leq 1;$$

tương tự $f_2(y) = y + \frac{1}{2}$, $0 \leq y \leq 1$. Từ đó theo (1.11), với $0 \leq y \leq 1$

$$\varphi(x | y) = \begin{cases} \frac{x + y}{y + 0,5}, & 0 \leq x \leq 1, \\ 0, & x \notin [0; 1], \end{cases}$$

và với $0 \leq x \leq 1$:

$$\psi(y | x) = \begin{cases} \frac{x + y}{x + 0,5}, & 0 \leq y \leq 1, \\ 0, & y \notin [0; 1]. \end{cases}$$

Để ý là hàm mật độ có điều kiện $\varphi(x | y)$ là hàm của x , đồng thời nếu coi y là tham số thì nó cũng là hàm của y . Từ (1.11) ta có $f(x, y) = f_2(y)\varphi(x | y) = f_1(x)\psi(y | x)$ và rõ ràng nếu:

$$\varphi(x|y) = f_1(x) \text{ (hoặc } \psi(y|x) = f_2(y))$$

ta có lại điều kiện độc lập (1.10).

Cuối cùng có thể dẫn ra các công thức tổng quát sau đây (xem (1.7) và (1.8)):

$$\varphi(x|y) = \frac{\int_{-\infty}^x f(u, y) du}{f_2(y)} \quad (\text{phân phối có điều kiện});$$

$$\varphi(x|y_1 \leq Y \leq y_2) = P(X < x | y_1 \leq Y \leq y_2) = \frac{\int_{-\infty}^x du \int_{y_1}^{y_2} f(u, v) dv}{\int_{-\infty}^{\infty} du \int_{y_1}^{y_2} f(u, v) dv};$$

$$\varphi(x|y_1 \leq Y \leq y_2) = \frac{\int_{y_1}^{y_2} f(x, v) dv}{\int_{-\infty}^{+\infty} du \int_{y_1}^{y_2} f(u, v) dv}. \quad (1.12)$$

Để ý trong các công thức trên cần bảo đảm để mẫu số khác không.

Thí dụ 1.6. Lấy hàm mật độ của thí dụ 1.5, hãy tính các hàm mật độ có điều kiện của X biết $Y \in [0,5; 0,75]$; biết $Y = 0,5$.

Giải. Theo công thức (1.12) ta có:

$$\varphi(x|Y \in [0,5; 0,75]) = \frac{\int_{0,5}^{0,75} (x+y) dy}{\int_0^{0,5} \int_{0,5}^{0,75} (x+y) dx dy} = \frac{8}{9}x + \frac{5}{9},$$

để ý là $0 \leq x \leq 1$; nếu $x \notin [0; 1]$ thì $\varphi(x \mid 0,5 \leq y \leq 0,75) = 0$. Trong trường hợp biết $Y = 0,5$;

$$\varphi(x \mid Y = 0,5) = \frac{x + 0,5}{0,5 + 0,5} = x + 0,5, 0 \leq x \leq 1.$$

§2. CÁC SỐ ĐẶC TRƯNG CỦA BIẾN NGẪU NHIÊN HAI CHIỀU

2.1. Các số đặc trưng của các biến thành phần

Các biến X và Y đã có các số đặc trưng quan trọng là kỳ vọng và phương sai. Ở đây ta nhắc lại kết quả đã biết có để ý đến các khái niệm mới ở chương này, các công thức chỉ viết cho biến X , đối với Y hoàn toàn tương tự.

Nếu X là biến rời rạc:

$$EX = \sum_i x_i p_1(x_i) = \sum_i \sum_j x_i p(x_i, y_j);$$

$$VX = \sum_i (x_i - EX)^2 p_1(x_i) = \sum_i \sum_j x_i^2 p(x_i, y_j) - (EX)^2.$$

Còn nếu X là biến liên tục

$$EX = \int_{-\infty}^{+\infty} x f_1(x) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy;$$

$$VX = \int_{-\infty}^{+\infty} (x - EX)^2 f_1(x) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 f(x, y) dx dy - (EX)^2.$$

Mở rộng phép lấy kỳ vọng, ta có thể dẫn ra các công thức tổng quát hơn. Chẳng hạn nếu (X, Y) có phân phối đã biết và ta xác định biến mới $Z = g(X, Y)$ (g là hàm đo được), khi đó:

$$E \{ g (X, Y) \} = \sum_i \sum_j g (x_i, y_j) p (x_i, y_j) \text{ (biến rời rạc);}$$

$$= \iint_{\forall x, y} g (x, y) f (x, y) dx dy \text{ (biến liên tục).} \quad (2.1)$$

Khi đó, để tính EX ta chỉ cần đặt $g = X$ và thay vào công thức (2.1); để tính VX đặt $g = (X - EX)^2$.

2.2. Hiệp phương sai và hệ số tương quan

Trong (2.1), nếu thay $g(X, Y) = (X - EX)(Y - EY)$, ta có định nghĩa hiệp phương sai của hai biến X và Y , ký hiệu là μ_{XY}

$$\mu_{XY} = E[(X - EX)(Y - EY)] = E(XY) - EX.EY. \quad (2.2)$$

Chú ý là phép toán lấy E ở bên ngoài dấu ngoặc móc hiểu theo nghĩa (2.1) và không giống như trường hợp biến một chiều. Phụ thuộc vào (X, Y) là rời rạc hay liên tục, ta có

$$\mu_{XY} = \sum_i \sum_j x_i x_j p(x_i, x_j) - EX.EY, \quad (2.2a)$$

$$\mu_{XY} = \int_{-\infty}^x \int_{-\infty}^x xy f(x, y) dx dy - EX.EY. \quad (2.2b)$$

Dễ thấy phương sai là trường hợp riêng của hiệp phương sai khi $X = Y$ và $VX = \mu_{XX}$. Hiệp phương sai được dùng làm độ đo quan hệ giữa hai biến X và Y ; nếu chúng đồng biến cùng nhau thì hiệp phương sai dương, nếu chúng nghịch biến ta có hiệp phương sai âm.

Ta biết nếu X và Y độc lập thì $E(XY) = EX.EY$, điều đó kéo theo hiệp phương sai của hai biến độc lập bằng 0. Nhưng điều ngược lại không chắc đúng. Vì vậy ta đưa vào khái niệm mới:

Định nghĩa 1. Nếu $\mu_{XY} = 0$, ta nói rằng X và Y *không tương quan*.

Rõ ràng khái niệm độc lập là mạnh hơn không tương quan. Nhiều khi để đơn giản ký hiệu, người ta tập hợp các hiệp

phương sai của một véctơ ngẫu nhiên vào một ma trận gọi là *ma trận hiệp phương sai*; trong trường hợp biến 2 chiều (X, Y) đó là:

$$\Gamma = \begin{bmatrix} VX & \mu_{XY} \\ \mu_{YX} & VY \end{bmatrix}.$$

Trên đường chéo chính là các phương sai, và do $\mu_{XY} = \mu_{YX}$ nên ma trận này đối xứng.

Hiệp phương sai có hạn chế cơ bản là khó xác định được miền biến thiên, nó thay đổi từ cặp biến này sang cặp biến khác. Chưa kể về mặt vật lý nó có đơn vị đo bằng bình phương đơn vị đo của X và Y (nếu chúng cùng đơn vị đo). Vì thế người ta đưa ra một số đặc trưng khác gọi là *hệ số tương quan*, ký hiệu là ρ_{XY} , được xác định như sau:

$$\rho_{XY} = \frac{\mu_{XY}}{\sigma_X \sigma_Y}. \quad (2.3)$$

Có thể chứng minh rằng $|\rho_{XY}| \leq 1$. Nếu $\rho_{XY} = \pm 1$, ta có hai biến X và Y tương quan dạng tuyến tính (tức là tồn tại a và b sao cho $Y = aX + b$); còn nếu $\rho_{XY} = 0$ thì X và Y không tương quan. Nói chung $0 < |\rho_{XY}| < 1$, trong trường hợp này ta nói rằng hai biến X và Y tương quan với nhau. Hai biến tương quan thì phụ thuộc (không độc lập), nhưng không tương quan thì chưa chắc độc lập.

Thí dụ 2.1. Tính hiệp phương sai và hệ số tương quan của X và Y trong thí dụ 1.1.

Giải. Ta phải tính VX , VY và $E(XY) - EX.EY$. Hai bảng phân phối biên đã tìm được trong thí dụ 1.1.

x	1	2	y	1	2	3
$p_1(x)$	0,45	0,55	$p_2(y)$	0,25	0,30	0,45

Ta có ngay $EX = 1,55$; $EY = 2,20$; $VX = 0,2475$; $VY = 0,66$. Tính

$$\begin{aligned} E(XY) &= \sum_i \sum_j x_i x_j p(x_i, y_j) = \\ &= 1.1.0,10 + 1.2.0,25 + 1.3.0,10 + 2.1.0,15 + \\ &\quad + 2.2.0,05 + 2.3.0,35 = 3,50. \end{aligned}$$

Từ đó hiệp phương sai

$$\mu_{XY} = E(XY) - EX \cdot EY = 3,50 - 1,55 \cdot 2,20 = 0,09.$$

Hệ số tương quan được tính theo (2.3)

$$\rho_{XY} = \frac{\mu_{XY}}{\sqrt{VX \cdot VY}} = \frac{0,09}{\sqrt{0,2475 \cdot 0,66}} \approx 0,22.$$

Thí dụ 2.2. Biến ngẫu nhiên hai chiều (X, Y) có hàm mật độ:

$$f(x, y) = \begin{cases} \frac{1}{2\pi}, & 4x^2 + y^2 \leq 4, \\ 0, & 4x^2 + y^2 > 4. \end{cases}$$

Chứng tỏ X và Y phụ thuộc và tính hiệp phương sai μ_{XY} .

Giải. Các hàm mật độ biên được tính theo (1.9):

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x, y) dy = \begin{cases} \frac{2}{\pi} \sqrt{1-x^2}, & |x| \leq 1, \\ 0, & |x| > 1, \end{cases} \\ f_2(y) &= \int_{-\infty}^{\infty} f(x, y) dx = \begin{cases} \frac{1}{2\pi} \sqrt{4-y^2}, & |y| \leq 2, \\ 0, & |y| > 2. \end{cases} \end{aligned}$$

Do $f(x, y) \neq f_1(x)f_2(y)$ nên X và Y phụ thuộc. Để tính μ_{XY} ta dùng công thức (2.2). Vì $f_1(x)$ và $f_2(x)$ là các hàm chẵn nên đồ thị đối xứng qua các trục tương ứng và $EX = EY = 0$, từ đó:

$$\begin{aligned} \mu_{XY} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dx dy \\ &= \frac{1}{2\pi} \int_{-1}^1 x dx \int_{-2\sqrt{1-x^2}}^{2\sqrt{1-x^2}} y dy = 0. \end{aligned}$$

(tích phân trong lấy theo hàm lẻ có cận đối xứng). Rõ ràng X và Y không tương quan, nhưng vẫn phụ thuộc nhau.

2.3. Các số đặc trưng có điều kiện

Dùng các khái niệm xác suất có điều kiện (xem (1.6)) và hàm mật độ có điều kiện (xem (1.11)), ta có thể định nghĩa kỳ vọng có điều kiện của biến ngẫu nhiên X với $Y = y$ là một giá trị xác định như sau:

$$E(X|y_k) = \sum_i x_i P(X = x_i | Y = y_k) \quad (X \text{ rời rạc}),$$

$$E(X|y) = \int_{-\infty}^{\infty} x \varphi(x|y) dx \quad (X \text{ liên tục}).$$

Tương tự có thể định nghĩa $E(Y|x)$ và các phương sai tương ứng.

Kỳ vọng có điều kiện $E(Y|x)$ là một hàm phụ thuộc x , và trong thống kê người ta gọi là *hàm hồi quy* của Y đối với X . Đồ thị của hàm đó trên mặt phẳng tọa độ Đề-các có tên gọi là *đường hồi quy*. Sau này ta sẽ dùng hồi quy để biểu diễn sự phụ thuộc tương quan giữa các biến ngẫu nhiên (xem chương VI).

Để ý là các kỳ vọng có điều kiện $E(X|Y)$, $E(Y|X)$ (cũng như các số đặc trưng có điều kiện khác) là các biến ngẫu nhiên nên đến lượt mình nó lại có thể có những đặc số tương ứng.

Thí dụ 2.3. Cho bảng phân phối của biến (X, Y) :

$\begin{smallmatrix} y \\ \backslash x \end{smallmatrix}$	1	2	3
2	0,15	0,08	0,27
4	0,10	0,20	0,20

Tính các kỳ vọng có điều kiện $E(X|y_1)$; $E(Y|x_2)$.

Giải. Dùng (1.3) và (1.6) ta có

$$P(X = 2 | Y = 1) = \frac{p_{11}}{p_2(1)} = \frac{0,15}{0,25} = 0,6;$$

$$P(X = 4 | Y = 1) = \frac{p_{21}}{p_2(1)} = \frac{0,10}{0,25} = 0,4.$$

Từ đó $E(X | Y = 1) = 2.0,6 + 4.0,4 = 2,8$. Tương tự

$$P(Y = 1 | X = 4) = \frac{p_{21}}{p_1(4)} = \frac{0,10}{0,50} = 0,2;$$

$$P(Y = 2 | X = 4) = \frac{p_{22}}{p_1(4)} = \frac{0,20}{0,50} = 0,4;$$

$$P(Y = 3 | X = 4) = \frac{p_{23}}{p_1(4)} = \frac{0,20}{0,50} = 0,4.$$

và từ đó suy ra $E(Y | X = 4) = 1. 0,2 + 2.0,4 + 3. 0,4 = 2,2$.

Cuối cùng, lưu ý đến một số tính chất của kỳ vọng có điều kiện $E(Y | X)$:

- (i) với mọi g liên tục $E[g(X)Y | X] = g(X)E(Y | X)$;
- (ii) $E(X_1 + X_2 | X) = E(X_1 | X) + E(X_2 | X)$;
- (iii) Nếu X, Y độc lập $E(Y | X) = E(Y)$;
- (iv) $E[E(Y | X)] = EY$.

2.4. Phân phối chuẩn hai chiều

Để cho gọn, ta dùng các ký hiệu sau:

$$\alpha_X = EX; \alpha_Y = EY; \sigma_X^2 = VX; \sigma_Y^2 = VY; \rho = \rho_{XY} \text{ và } \mu = \mu_{XY}.$$

Định nghĩa 2. Biến ngẫu nhiên hai chiều (X, Y) được gọi là tuân theo *luật phân phối chuẩn*, ký hiệu là $\mathcal{N}(\alpha_X, \alpha_Y, \sigma_X^2, \sigma_Y^2, \rho)$, nếu hàm mật độ đồng thời của X và Y có dạng

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times$$

$$\times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-a_X}{\sigma_X} \right)^2 + \left(\frac{y-a_Y}{\sigma_Y} \right)^2 - 2\rho \frac{(x-a_X)(y-a_Y)}{\sigma_X \sigma_Y} \right] \right\}. \quad (2.4)$$

Có thể chỉ ra dễ dàng nếu X, Y không tương quan ($\rho = 0$) thì giả thiết chuẩn cho phép kết luận chúng là độc lập. Bạn đọc có thể chứng minh trong trường hợp này $f(x, y) = f_1(x)f_2(y)$.

Dùng ma trận hiệp phương sai Γ và vectơ $\mathbf{x}$ xác định như sau:

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}; \Gamma = \begin{pmatrix} \sigma_X^2 & \mu \\ \mu & \sigma_Y^2 \end{pmatrix},$$

ta có thể biểu diễn hàm mật độ chuẩn (2.4) dưới dạng gọn hơn

$$f(\mathbf{x}) = f(x, y) = \frac{1}{2\pi\sqrt{\det \Gamma}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - E\mathbf{x})^t \Gamma^{-1} (\mathbf{x} - E\mathbf{x}) \right\},$$

trong đó $\det$ là ký hiệu định thức, t – phép chuyển vị, còn $E\mathbf{x}$ hiểu theo nghĩa là vectơ có các thành phần EX và EY (hay a_X và a_Y).

Thí dụ 2.4. Cho $(X, Y) \sim \mathcal{N}(a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Hãy tính các kỳ vọng có điều kiện và phương sai có điều kiện.

Giải. Bạn đọc có thể tính được dễ dàng từ (2.4)

$$f_2(y) = \frac{1}{\sigma_Y \sqrt{2\pi}} e^{-\frac{1}{2\sigma_Y^2}(y-a_Y)^2} \quad \text{hay } Y \sim \mathcal{N}(a_Y, \sigma_Y^2),$$

từ đó

$$\begin{aligned} \varphi(x|y) &= \frac{f(x, y)}{f_2(y)} = \\ &= \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2\sigma_X^2(1-\rho^2)} \left[x - a_X - \rho \frac{\sigma_X}{\sigma_Y} (Y - a_Y) \right]^2 \right\}. \end{aligned}$$

Biểu thức trên chính là hàm mật độ của phân phối chuẩn

$$\mathcal{N}\left(a_X + \rho \frac{\sigma_X}{\sigma_Y}(y - a_Y); \sigma_X^2(1 - \rho^2)\right),$$

từ đó $E(X|Y=y) = a_X + \rho \frac{\sigma_X}{\sigma_Y}(y - a_Y);$

$$V(X|Y=y) = \sigma_X^2(1 - \rho^2).$$

Hoàn toàn tương tự đối biến Y (do tính đối xứng của hai biến):

$$E(Y|X=x) = a_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - a_X);$$

$$V(Y|X=x) = \sigma_Y^2(1 - \rho^2).$$

§3. HÀM CỦA CÁC BIẾN NGẪU NHIÊN

3.1. Hàm của một biến ngẫu nhiên

Nếu ta xác định $Z = g(X)$ là một hàm của biến ngẫu nhiên X thì Z trở thành một biến ngẫu nhiên mới. Vấn đề đặt ra là tìm cách xác định luật phân phối của Z qua luật phân phối đã biết của X . Ở đây ta chỉ xét các trường hợp đơn giản khi hàm g không quá phức tạp. Xét trường hợp rời rạc.

Thí dụ 3.1. Cho biến ngẫu nhiên X có luật phân phối

x	-2	-1	0	1	2
$p(x)$	0,1	0,2	0,3	0,2	0,2

Xác định luật phân phối của $Z = X^2$ và tìm kỳ vọng của Z .

Giải. Dễ thấy $P(Z=0) = P(X=0) = 0,3;$

$Z=1 \Leftrightarrow X=-1$ hoặc $X=1$, suy ra $P(Z=1) = 0,2 + 0,2 = 0,4;$

$Z=4 \Leftrightarrow X=-2$ hoặc $X=2$, suy ra $P(Z=4) = 0,1 + 0,2 = 0,3.$

Từ đó bảng phân phối của $Z = X^2$ là:

z	0	1	4
$p(z)$	0,3	0,4	0,3

Xuất phát từ luật phân phối trên:

$$EZ = \sum_i z_i p(z_i) = 0.0,3 + 1.0,4 + 4.0,3 = 1,6.$$

Trong trường hợp $Z = g(X)$ tổng quát, ta có thể tính trực tiếp kỳ vọng của Z không cần qua luật phân phối (xem tính chất kỳ vọng ở chương II):

$$EZ = \sum_i g(x_i) P(X = x_i) = \sum_i g(x_i) p_i.$$

Trong thí dụ 3.1, dễ thấy $EZ = (-2)^2.0,1 + (-1)^2.0,2 + 0^2.0,3 + 1^2.0,2 + 2^2.0,2 = 1,6$.

Khi X là biến ngẫu nhiên liên tục, vấn đề sẽ phức tạp hơn. Giả sử X có hàm mật độ $f(x)$ đã biết và $Z = g(X)$, trong đó g là hàm đơn điệu sao cho tồn tại hàm ngược duy nhất $X = \psi(Z) = g^{-1}(Z)$. Khi đó hàm mật độ của $\varphi(z)$ của biến $Z = g(X)$ sẽ là:

$$\varphi(z) = f[\psi(z)] |\psi'(z)|. \quad (3.1)$$

Thí dụ 3.2. Biến ngẫu nhiên X tuân theo luật phân phối mũ với tham số $\lambda = \frac{1}{2}$. Tìm luật phân phối xác suất của biến $Z = X^3$.

Giải. Vì hàm $y = x^3$ là đơn điệu tăng và khả vi, do đó có thể áp dụng công thức (3.1). Dễ thấy $\psi(y) = y^{\frac{1}{3}} = x$, mặt khác do hàm mật độ của X có dạng:

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0,$$

nên nếu đặt $\varphi(z)$ là hàm mật độ của $Z = X^3$, ta có

$$\varphi(z) = f[\psi(z)] |\psi'(z)| = \lambda e^{-\lambda \sqrt[3]{z}} \cdot \frac{1}{3} z^{-\frac{2}{3}} = \frac{\lambda}{3\sqrt[3]{z^2}} e^{-\lambda \sqrt[3]{z}}, \quad z > 0.$$

Thí dụ 3.3. Cho biến ngẫu nhiên $X \sim \mathcal{N}(m, \sigma^2)$. Tìm luật phân phối xác suất của biến $Y = aX + b$ trong đó $a, b \in \mathbf{R}$.

Giải. Hàm số $y = ax + b$ khả vi đơn điệu, có hàm ngược là $x = \psi(y) = \frac{y-b}{a} (a \neq 0)$. Từ đó:

$$|\psi'(y)| = \frac{1}{|a|},$$

$$\begin{aligned} \text{còn } f[\psi(y)] &= \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{\left(\frac{y-b}{a} - m\right)^2}{2\sigma^2} \right] \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(y-am-b)^2}{2(a\sigma)^2} \right]. \end{aligned}$$

Từ đó dùng công thức (3.1) ta có hàm mật độ của $Y = aX + b$ là:

$$\varphi(y) = f[\psi(y)] |\psi'(y)| = \frac{1}{|a|\sigma\sqrt{2\pi}} \exp \left\{ -\frac{[y-(am+b)]^2}{2(a\sigma)^2} \right\}.$$

Đó chính là hàm mật độ của luật phân phối chuẩn với hai tham số $EY = am + b$ và $VY = a^2\sigma^2$. Như vậy một hàm tuyến tính của biến ngẫu nhiên chuẩn vẫn bảo toàn tính phân phối chuẩn.

3.2. Hàm của hai biến ngẫu nhiên

Xét biến ngẫu nhiên $Z = g(X, Y)$, trong đó (X, Y) là biến ngẫu nhiên 2 chiều đã biết luật phân phối xác suất. Nếu g là

một hàm tùy ý thì bài toán xác định luật phân phối của Z qua luật phân phối của (X, Y) sẽ rất phức tạp. Ta sẽ xét một trường hợp đơn giản khi $g(X, Y) = X + Y$.

1. Trường hợp các biến X, Y rời rạc

Theo công thức xác suất đầy đủ (tổng lấy theo i sao cho $x_i + y_i = z_k$)

$$\begin{aligned} P(Z = z_k) &= \sum_i P(X = x_i; Y = z_k - x_i) \\ &= \sum_i P(X = x_i) P(Y = z_k - x_i | X = x_i). \end{aligned}$$

Nếu X và Y độc lập:

$$P(Z = z_k) = \sum_i P(X = x_i) P(Y = z_k - x_i). \quad (3.2)$$

Thí dụ 3.3. Cho luật phân phối của (X, Y) có dạng.

$x \backslash y$	2	3	4
1	0	0,15	0,05
2	0,20	0,10	0
3	0,25	0,05	0,20

Xác định luật phân phối xác suất của $X + Y$.

Giải. Tập giá trị của $Z = X + Y$ là $\{3, 4, 5, 6, 7\}$, từ đó

$$P(Z = 3) = P(X = 1, Y = 2) = 0;$$

$$P(Z = 4) = P(X = 1, Y = 3) + P(X = 2, Y = 2) = 0,15 + 0,20 = 0,35;$$

$$\begin{aligned} P(Z = 5) &= P(X = 1, Y = 4) + P(X = 2, Y = 3) + P(X = 3, Y = 2) \\ &= 0,05 + 0,10 + 0,25 = 0,40; \end{aligned}$$

$$P(Z = 6) = P(X = 2, Y = 4) + P(X = 3, Y = 3) = 0 + 0,05 = 0,05;$$

$$P(Z = 7) = P(X = 3, Y = 4) = 0,20.$$

Từ bảng phân phối xác suất của Z là:

z	4	5	6	7
$p(z)$	0,35	0,40	0,05	0,20

Thí dụ 3.4. Cho X và Y là hai biến ngẫu nhiên độc lập tuân theo luật Poa-xông với các tham số tương ứng λ và μ . Tìm luật phân phối của $Z = X + Y$.

Giải. Theo công thức (3.2)

$$P(Z=z) = \sum_{x=0}^z e^{-\lambda} \frac{\lambda^x}{x!} e^{-\mu} \frac{\mu^{z-x}}{(z-x)!}; x, y, z \in \mathbb{N}.$$

Nhân và chia vế phải với $z!$

$$P(Z=z) = \frac{e^{-(\lambda+\mu)}}{z!} \sum_{x=0}^z C_z^x \lambda^x \mu^{z-x} = e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^z}{z!}.$$

Hệ thức cuối cho thấy $Z = X + Y$ cũng tuân theo luật Poa-xông với tham số $\lambda + \mu$.

2. Trường hợp các biến X, Y liên tục

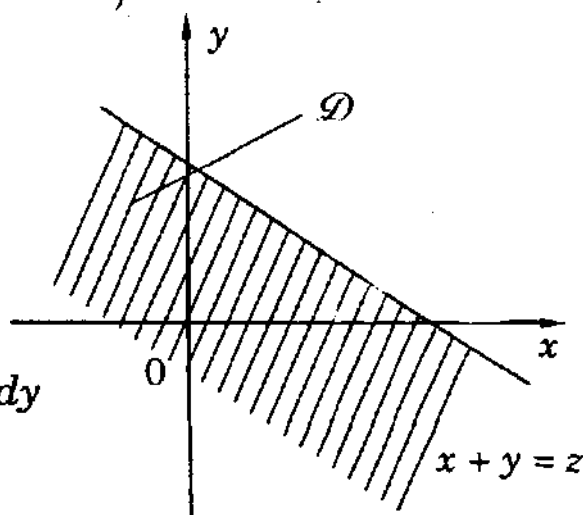
Gọi $\phi(z)$ là hàm phân phối của biến $Z = X + Y$, ta có:

$$\phi(z) = P(Z < z) = P(X + Y < z).$$

Biểu diễn tập giá trị của (X, Y) sao cho $X + Y < z$ là miền $\mathcal{D}$ (miền gạch chéo trên hình 3.1) nên theo tính chất (iii) của hàm mật độ 2 chiều:

$$\phi(z) = P((X, Y) \in \mathcal{D}) = \iint_{\mathcal{D}} f(x, y) dx dy$$

$$= \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{z-x} f(x, y) dy \right\} dx.$$



Hình 3.1

Lấy đạo hàm hai vế theo z và gọi $\varphi(z) = \phi'(z)$ là mật độ của Z :

$$\varphi(z) = \int_{-x}^{+\infty} f(x, z-x) dx. \quad (3.3a)$$

Tương tự nếu ta thay đổi trình tự lấy tích phân:

$$\varphi(z) = \int_{-y}^{+\infty} f(z-y, y) dy. \quad (3.3b)$$

Trong trường hợp riêng, khi X và Y độc lập, từ (3.3) ta có:

$$\varphi(z) = \int_{-x}^{+\infty} f_1(x) f_2(z-x) dx = \int_{-y}^{+\infty} f_1(z-y) f_2(y) dy, \quad (3.4)$$

trong đó f_1 và f_2 là các hàm mật độ biên của X và Y tương ứng. Biểu thức (3.4) mô tả một phép toán liên hệ hai hàm f_1 và f_2 ; nó được gọi là *tích chập*, ký hiệu là $f_1 * f_2$.

Từ dụ 3.5. Cho hai biến X, Y độc lập cùng có phân phối đều trên đoạn $[0; 1]$ (tức hàm mật độ $f_1(x) = f_2(x) = 1$ khi $x \in [0; 1]$). Tìm hàm mật độ và hàm phân phối của $Z = X + Y$.

Giải. Để ý rằng cả hai hàm mật độ của X và Y đều bằng 0 khi đối số nằm ngoài $[0; 1]$. Gọi $\varphi(z)$ là hàm mật độ của $Z = X + Y$, ta có theo (3.4):

$$\varphi(z) = \int_0^1 f_1(x) f_2(z-x) dx. \quad (3.5)$$

Do $f_1(x)$ và $f_2(x)$ có giá trị khác không trên $[0; 1]$, nên $\varphi(z)$ chỉ có thể có giá trị khác không trên $[0; 2]$. Ta tính lần lượt:

– nếu $z \leq 0$, $\varphi(z) = 0$;

– nếu $0 < z \leq 1$, (3.5) được viết lại:

$$\varphi(z) = \int_0^z f_1(x) f_2(z-x) dx + \int_z^1 f_1(x) f_2(z-x) dx;$$

do tích phân thứ hai bằng 0 bởi $f_2 \equiv 0$, suy ra $\varphi(z) = \int_0^z dx = z$;

– nếu $1 < z \leq 2$, hay $0 < z - 1 \leq 1$, (3.5) trở thành

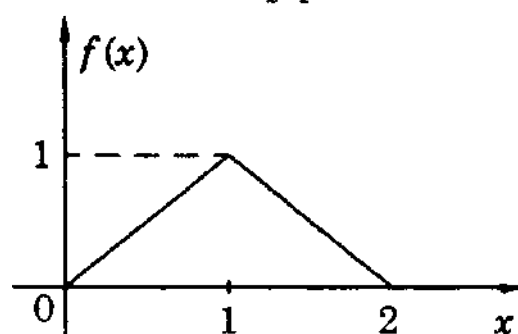
$$\varphi(z) = \int_0^{z-1} f_1(x)f_2(z-x)dx + \int_{z-1}^1 f_1(x)f_2(z-x)dx;$$

do tích phân thứ nhất bằng 0 bởi $f_2 \equiv 0$, suy ra $\varphi(z) = \int_{z-1}^1 dx = 2 - z$;

– nếu $z > 2$, $\varphi(z) = 0$.

Từ đó mật độ của $Z = X + Y$ sẽ là

$$\varphi(z) = \begin{cases} 0, & z \leq 0, \\ z, & 0 < z \leq 1, \\ 2-z, & 1 < z \leq 2, \\ 0, & z > 2. \end{cases}$$



Hình 3.2

Biến ngẫu nhiên Z có hàm mật độ như trên được gọi là tuân theo *luật phân phối tam giác*, hay *phân phối Xim-xon* (xem hình 3.2). Việc tìm hàm phân phối của Z không gì phức tạp:

$$\phi(z) = \begin{cases} 0, & z \leq 0, \\ \frac{z^2}{2}, & 0 < z \leq 1, \\ 1 - \frac{(2-z)^2}{2}, & 1 < z \leq 2, \\ 1, & z > 2. \end{cases}$$

3.3. Các số đặc trưng của hàm của các biến ngẫu nhiên

Khi muốn tính các số đặc trưng (kỳ vọng, phương sai, ...) của biến ngẫu nhiên $X = g(X, Y)$, đã biết luật phân phối xác suất của Z , ta không gặp trở ngại gì lớn. Tuy nhiên việc xác định luật phân phối của Z khá phức tạp. Trong thực tế nhiều khi ta chỉ cần quan tâm đến các số đặc trưng của Z là đủ.

Chẳng hạn trong trường hợp khi X và Y là các biến rời rạc và đã biết phân phối đồng thời $p(x_i; y_j)$

$$EZ = E[g(X, Y)] = \sum_i \sum_j g(x_i; y_j) p(x_i; y_j). \quad (3.6a)$$

Nếu (X, Y) liên tục có hàm mật độ $f(x, y)$, thì

$$EZ = \int_{\mathbb{R}^2} g(x, y) f(x, y) dx dy. \quad (3.6b)$$

Tất nhiên (3.6) chỉ có giá trị khi tổng và tích phân tồn tại.

Đề ý là (3.6) cho phép ta chứng minh chặt chẽ nhiều tính chất của kỳ vọng như:

$$+ E(X + Y) = EX + EY;$$

$$+ \text{nếu } X \text{ và } Y \text{ độc lập, } E(XY) = EX \cdot EY.$$

Bạn đọc thử thiết lập công thức sau đây cho phương sai:

$$+ V(X + Y) = VX + VY - \mu_{XY}.$$

§4. CÁC ĐỊNH LÝ GIỚI HẠN VÀ LUẬT SỐ LỚN

Các định lý giới hạn và luật số lớn rất có ý nghĩa trong thực tiễn. Nó tạo ra cơ sở cho các ứng dụng của thống kê toán học sau này.

4.1. Sự hội tụ của dãy biến ngẫu nhiên

1. Hội tụ hầu chắc chắn

Ta nói rằng dãy biến ngẫu nhiên $\{X_n\}$ *hội tụ hầu chắc chắn* (hay *hội tụ mạnh*) đến biến X , ký hiệu là $X_n \xrightarrow[n \rightarrow \infty]{hcc} X$, nếu

$$P\left(\lim_{n \rightarrow \infty} X_n \neq X\right) = 0.$$

Có thể dùng một tiêu chuẩn khác để xác định hội tụ hầu chắc chắn: Điều kiện cần và đủ để $X_n \xrightarrow[n \rightarrow \infty]{hcc} X$, là $\forall \varepsilon > 0$:

$$P\left(\left\{\sup_{m \geq n} |X_m - X| > \varepsilon\right\}\right) \xrightarrow{n \rightarrow \infty} 0. \quad (4.1)$$

Như vậy hội tụ hầu chắc chắn trùng với hội tụ thường đối với sự kiện có xác suất 1. Chú ý là có thể thay (4.1) bằng các điều kiện tương đương

$$P\left(\sum_{m \geq n} \{|X_m - X| > \varepsilon\}\right) \xrightarrow{n \rightarrow \infty} 0.$$

$$P\left(\prod_{m \geq n} \{|X_m - X| \leq \varepsilon\}\right) \xrightarrow{n \rightarrow \infty} 1.$$

2. Hội tụ theo xác suất

Ta nói rằng dãy $\{X_n\}$ *hội tụ theo xác suất* đến X , ký hiệu $X_n \xrightarrow[n \rightarrow \infty]{XS} X$, nếu

$$\forall \varepsilon > 0: P(\{|X_n - X| > \varepsilon\}) \xrightarrow{n \rightarrow \infty} 0. \quad (4.2)$$

Rõ ràng hội tụ chắc chắn (xem (4.1)) kéo theo hội tụ theo xác suất (4.2), điều ngược lại nói chung không đúng.

3. Hội tụ theo luật

Ta nói rằng dãy $\{X_n\}$ *hội tụ theo luật* đến X , ký hiệu là $X_n \xrightarrow[n \rightarrow \infty]{L} X$, nếu dãy hàm phân phối $F_n(x)$ của X_n hội tụ đến hàm phân phối $F(x)$ của biến X tại mọi điểm liên tục của hàm $F(x)$.

Để ý đối với biến ngẫu nhiên rời rạc sự hội tụ theo luật được diễn đạt bởi hệ thức

$$p_n(x) = P(X_n = x) \xrightarrow{n \rightarrow \infty} P(X = x) = p(x).$$

Người ta chứng minh được rằng hội tụ theo xác suất kéo theo hội tụ theo luật. Đây là kiểu hội tụ yếu nhất, tuy nhiên lại hay dùng nhất. Ở chương II ta đã sử dụng kiểu hội tụ này, thí dụ trong các công thức xấp xỉ chuẩn. Chẳng hạn nếu $Y_n \sim \mathcal{B}(n, p)$, với p không quá gần 0 hoặc 1, ta đã có

$$X_n = \frac{Y_n - np}{\sqrt{np(1-p)}} \xrightarrow[n \rightarrow \infty]{L} \mathcal{N}(0, 1) \quad (4.3)$$

$\mathcal{N}(0, 1)$ hiểu theo nghĩa là biến ngẫu nhiên có phân phối chuẩn chuẩn tắc).

4. Hội tụ trung bình cấp k

Ta nói rằng dãy $\{X_n\}$ hội tụ trung bình cấp k đến X , ký hiệu là $X_n \xrightarrow[n \rightarrow \infty]{tbk} X$, nếu $E[|X_n - X|^k] \xrightarrow[n \rightarrow \infty]{} 0$, (với điều kiện kỳ vọng đó tồn tại). Hội tụ trung bình cấp k (thường hay dùng với $k = 2$ – hội tụ trung bình bình phương) kéo theo hội tụ theo xác suất.

4.2. Các định lý giới hạn

1. Các định lý giới hạn Moa-vơ – Láp-la-xơ

Sử dụng kết quả (4.3) dễ dàng suy ra định lý giới hạn địa phương Moa-vơ – Láp-la-xơ (xem (3.13) chương I)

$$P_n(k) \approx \frac{\varphi(x_k)}{\sqrt{npq}}, \quad x_k = \frac{k - np}{\sqrt{npq}}; \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}; \quad (4.4)$$

và định lý giới hạn tích phân (xem (3.14) chương I, (4.18) chương II)

$$P_n(k_1, k_2) \approx \Phi(x_2) - \Phi(x_1), \quad x_i = \frac{k_i - np}{\sqrt{npq}}, \quad i = 1, 2, \quad (4.5)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt = \int_0^x \varphi(t) dt.$$

Công thức xấp xỉ (4.5) sẽ khá tốt khi $np \geq 5$ hoặc $nq \geq 5$. Nếu p càng gần 0,5 đồ thị của phân phối nhị thức càng rất gần chuẩn.

2. Định lý giới hạn trung tâm

Lin-đơ-bóc – Lê-vi đã mở rộng định lý giới hạn Moa-vơ – Láp-la-xơ từ năm 1922 và kết quả đó mang tên *định lý giới*

hạn trung tâm: giả sử $\{X_n\}$ là dãy các biến ngẫu nhiên độc lập có cùng phân phối với $EX_n = m$ và $VX_n = \sigma^2 \forall n$, khi đó

$$\frac{\bar{X}_n - m}{\sigma} \sqrt{n} \xrightarrow[n \rightarrow \infty]{L} \mathcal{N}(0; 1), \text{ với } \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i. \quad (4.6)$$

Ý nghĩa của định lý giới hạn trung tâm là khi có nhiều nhân tố ngẫu nhiên tác động (sao cho không có nhân tố nào vượt trội lấn át các nhân tố khác) thì kết quả của chúng có dạng phân phối tiệm cận chuẩn.

Thí dụ 4.1. Một quả đậu có trọng lượng trung bình là 15 gam với độ lệch chuẩn là 3 gam. Một túi gồm 100 quả đậu cùng loại được gọi là đạt loại A nếu trọng lượng ít nhất phải đạt 1,5 kg.

a) Lấy ra ngẫu nhiên một túi, tìm xác suất để túi đó đạt loại A.

b) Chọn ngẫu nhiên ra 40 túi đậu, tìm xác suất để số túi loại A không vượt quá 15.

Giải. Gọi X_i – trọng lượng quả đậu thứ i trong túi ($i = \overline{1, 100}$), rõ ràng trọng lượng của túi là $S_{100} = X_1 + X_2 + \dots + X_{100}$ và theo định lý giới hạn trung tâm (4.6) $S_n \xrightarrow[\text{xấp xỉ}]{L} \mathcal{N}(ES_n, \sigma_{s_n}^2)$. Dễ thấy ở đây $ES_{100} = 100.15 \text{ gam} = 1,5 \text{ kg}$, $VS_{100} = \sigma_{S_{100}}^2 = 100.3^2 = 900 \text{ gam}^2$. Từ đó có thể coi S_{100} có phân phối xấp xỉ chuẩn.

$$S_{100} \sim \mathcal{N}(1,5; 0,9).$$

a) Rõ ràng do S_{100} có phân phối chuẩn nên $P(S_{100} \geq ES_{100} = 1,5 \text{ kg}) = 0,5$.

b) Chọn hú họa ra 40 túi và gọi $p = 0,5$ là xác suất để một túi đạt loại A, suy ra số túi loại A trong loạt túi 40 túi đó, ký hiệu là X , tuân theo luật nhị thức $\mathcal{B}(40; 0,5)$. Từ đó ta cần phải tính $P(X \leq 15) = P_{40}(0; 15)$. Ở đây $np = 40.0,5 = 20 > 5$, ta áp dụng công thức (4.5)

$$\begin{aligned}
P_{50}(0; 15) &\approx \phi\left(\frac{15 - np}{\sqrt{npq}}\right) - \phi\left(\frac{0 - np}{\sqrt{npq}}\right) \\
&= \phi\left(\frac{-5}{\sqrt{0}}\right) - \phi\left(\frac{-20}{\sqrt{10}}\right) = \phi(2\sqrt{10}) - \phi\left(\frac{\sqrt{10}}{2}\right). \\
&\approx \phi(6,32) - \phi(1,58) = 0,5 - 0,443 = 0,017.
\end{aligned}$$

4.3. Luật số lớn

Một lớp các định lý giới hạn đặc biệt có tên gọi là *luật số lớn*. Để ý là trong các kết quả sau này ta sử dụng khái niệm hội tụ theo xác suất (mạnh hơn (4.4) – (4.6) dùng hội tụ theo luật phân phối).

1. Bất đẳng thức Trê-bư-sép

Định lý 1. Nếu biến ngẫu nhiên X có kỳ vọng $EX = a$ và phương sai $VX = \sigma^2$ hữu hạn thì:

$$P(|X - a| \geq \varepsilon) < \frac{\sigma^2}{\varepsilon^2}, \quad \forall \varepsilon > 0. \quad (4.7)$$

Chứng minh. Ta sẽ chứng minh cho trường hợp biến X liên tục. Việc chứng minh cho trường hợp X rời rạc dành cho bạn đọc. Đặt $f(x)$ là hàm mật độ của X , theo tính chất của hàm mật độ

$$P(|X - a| \geq \varepsilon) = \int_{|x-a| \geq \varepsilon} f(x) dx.$$

Trong miền lấy tích phân dễ thấy $(x - a)^2 \geq \varepsilon^2$, nên:

$$\begin{aligned}
\int_{|x-a| \geq \varepsilon} f(x) dx &\leq \frac{1}{\varepsilon^2} \int_{|x-a| \geq \varepsilon} (x - a)^2 f(x) dx \\
&\leq \frac{1}{\varepsilon^2} \int_{\mathbf{R}} (x - a)^2 f(x) dx = \frac{\sigma^2}{\varepsilon^2} \quad (\text{đpcm}).
\end{aligned}$$

Bất đẳng thức (4.7) có thể chuyển về dạng tương đương

$$P(|X - a| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}, \quad \forall \varepsilon > 0. \quad (4.8)$$

Mặc dù (4.7) – (4.8) được chứng minh khá đơn giản, song chúng có ý nghĩa rất to lớn để dùng làm cơ sở cho các ứng dụng của thống kê. Để ý nếu chọn ε khá bé, chẳng hạn $\varepsilon < \sigma$, bất đẳng thức Trê-bư-sép trở nên tầm thường; nếu chọn $\varepsilon = 3\sigma$ ta có $P(|X - a| < 3\varepsilon) \geq 1 - \frac{1}{9} \approx 0,9$ (ít nhất bằng 0,9, ở chương II ta đã biết nếu $X \sim \mathcal{N}(\alpha; \sigma^2)$ thì xác suất trên $\approx 0,9973$).

2. Luật số lớn Trê-bư-sép

Định lý 2. Nếu dãy các biến ngẫu nhiên $X_1, X_2, \dots, X_n, \dots$ độc lập có kỳ vọng hữu hạn và phương sai bị chặn đều (tức là $VX_i \leq C \forall i$), khi đó với mọi $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n EX_i\right| < \varepsilon\right) = 1. \quad (4.9)$$

Chứng minh. Đặt $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, từ đó

$$E\bar{X} = \frac{1}{n} \sum_{i=1}^n EX_i \text{ hữu hạn;}$$

$$V\bar{X} = \frac{1}{n^2} \sum_{i=1}^n VX_i \leq \frac{C}{n}.$$

Từ đó áp dụng (4.8) cho biến ngẫu nhiên $\bar{X}$

$$P(|\bar{X} - E\bar{X}| < \varepsilon) \geq 1 - \frac{V\bar{X}}{\varepsilon^2} \geq 1 - \frac{C}{n\varepsilon^2}. \quad (4.10)$$

Do xác suất không vượt quá 1, nên khi chuyển qua giới hạn $n \rightarrow \infty$ ta có kết quả cần chứng minh (4.9).

Để ý đến (4.2) và (4.9), rõ ràng

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{XS} \frac{1}{n} \sum_{i=1}^n EX_i.$$

Điều đó có nghĩa là khi n đủ lớn thì trung bình cộng của các biến ngẫu nhiên sẽ có giá trị lệch rất ít so với trung bình cộng của các kỳ vọng. Một hệ quả quan trọng của định lý 2 là nếu đưa thêm giả thiết là các $X_i, i = 1, 2, \dots$ có cùng vọng số (tức là $EX_i = a, i = 1, 2, \dots$) thì (4.9) sẽ trở thành

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - a\right| < \varepsilon\right) \xrightarrow{n \rightarrow \infty} 1, \forall \varepsilon > 0.$$

Sự kiện này cho phép ta ước lượng kỳ vọng bằng trung bình cộng các kết quả đo đặc độc lập của biến ngẫu nhiên có kỳ vọng đó. Ngoài ra công thức (4.10) cung cấp một đánh giá khá tốt xác suất $P(|\bar{X} - E\bar{X}| < \varepsilon)$, nhất là khi n đủ lớn.

3. Luật số lớn Béc-nu-li

Định lý 3. Nếu ta có dãy n phép thử độc lập Béc-nu-li, với $p = P(A)$ và m là số lần xuất hiện A trong dãy phép thử đó, thì $\forall \varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1. \quad (4.11)$$

Việc chứng minh (4.11) không quá phức tạp vì nó là trường hợp riêng của (4.9), nếu ta ký hiệu X_i là số lần xuất hiện A trong phép thử thứ $i, i = \overline{1, n}$. Rõ ràng X_i tuân theo luật Béc-nu-li và

$$EX_i = p, VX_i = p(1-p) < 1, m = \sum_{i=1}^n X_i;$$

các điều kiện của định lý 2 đã được thỏa mãn và ta suy ra ngay (4.11). Kết quả này cho ta:

$$\frac{m}{n} \xrightarrow[n \rightarrow \infty]{XS} p = P(A);$$

đó chính là cơ sở cho định nghĩa thống kê của xác suất đã đưa ra ở chương I.

Như vậy tổng của một số khá lớn các biến ngẫu nhiên tương đối tùy ý lại trở nên tuân theo một số quy luật xác định. Điều này cho phép chúng ta ứng dụng rộng rãi các kết quả của xác suất và thống kê vào nhiều lĩnh vực khác nhau của khoa học và đời sống.

BÀI TẬP

1. Cho biến ngẫu nhiên hai chiều (X, Y) có bảng phân phối như sau

$x \backslash y$	y_1	y_2
x_1	0,18	0,08
x_2	0,22	0,16
x_3	0,16	0,20

Xác định luật phân phối biên của từng biến X, Y .

2. Người ta tiến hành 3 thí nghiệm với xác suất thành công của mỗi lần là 0,7. Tìm luật phân phối đồng thời của cặp biến X, Y với X là số thí nghiệm thành công, còn Y là số thất bại.
3. Luật phân phối của biến (X, Y) cho bởi bảng

$x \backslash y$	20	40	60
10	λ	λ	0
20	2λ	λ	λ
30	3λ	λ	λ

Xác định λ và các phân phối biên của X và của Y .

4. Luật phân phối đồng thời của số lỗi vẽ màu X và số lỗi đục Y của một loại sản phẩm nhựa ở một công ty cho bởi

$x \backslash y$	0	1	2
0	0,58	0,10	0,06
1	0,06	0,05	0,05
2	0,02	0,04	0,01
3	0,02	0,01	0,00

Hai biến X và Y độc lập không? Tính xác suất để tổng số các lỗi vẽ màu và lỗi đúc lớn hơn 4. Nếu ta biết trên sản phẩm có 2 lỗi vẽ màu thì xác suất để không có lỗi đúc bằng bao nhiêu?

5. Cho luật phân phối của biến hai chiều (X, Y) như sau:

$x \backslash y$	2	3	5
1	0,1	0	0,1
4	0,2	0,5	0,1

Tìm luật phân phối xác suất của hàm $X + Y$ và XY sau đó tính các kỳ vọng và phương sai.

6. Biến ngẫu nhiên (X, Y) có hàm mật độ đồng thời

$$f(x, y) = a(x^2 + y^2), x^2 + y^2 \leq 4.$$

Xác định hệ số a , các kỳ vọng thành phần và hiệp phương sai.

7. Cho biến hai chiều (X, Y) có phân phối đều trong mặt tròn tâm ở gốc tọa độ và bán kính r . Hãy xác định hàm phân phối biên của X và Y , sau đó tìm hàm mật độ có điều kiện $\varphi(x | y)$.

8. Cho hàm mật độ đồng thời của X và Y

$$f(x, y) = cxy, 0 \leq x \leq 4; 0 \leq y \leq 5.$$

Xác định hằng số c , sau đó tìm các hàm mật độ biên và hàm mật độ có điều kiện của Y biết $0,5 \leq X \leq 2$.

9. Cho hàm mật độ đồng thời của X và Y

$$f(x, y) = ae^{-\frac{1}{2}(x^2 + 2xy + 5y^2)}.$$

Xác định hằng số a , sau đó tìm các hàm mật độ có điều kiện.

10. Hai máy tự động làm việc độc lập, xác suất để từng máy sản xuất ra sản phẩm tốt tương ứng là p_1 và p_2 . Giả sử mỗi máy làm được 2 sản phẩm và gọi X và Y tương ứng là số sản phẩm tốt của từng máy. Hãy tìm bảng phân phối xác suất của biến hai chiều (X, Y) .

11. Tính hiệp phương sai và hệ số tương quan của X và Y cho trong bài tập 4.
12. Các tọa độ (X, Y) của một điểm ngẫu nhiên trên mặt phẳng tuân theo luật phân phối có hàm mật độ

$$f(x, y) = \frac{1}{2\pi ab} e^{-\frac{1}{2}\left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right)}, a, b \in \mathbf{R}^+.$$

Tìm xác suất để điểm đó nằm trong một elíp có các bán trục bằng ka và kb nằm trên các trục tọa độ Ox và Oy .

13. Tính hệ số tương quan của X và Y có hàm mật độ đồng thời

$$f(x, y) = \frac{2}{\pi(x^2 + y^2 + 1)^3}.$$

14. Cho hai biến ngẫu nhiên X và Y độc lập, có cùng phân phối chuẩn $\mathcal{N}(0, \sigma^2)$. Tính các xác suất của các sự kiện sau: $X < Y$; $|X| > Y$; đồng thời $X < 1$ và $Y < 1$.
15. Cho hai biến ngẫu nhiên X và Y độc lập, có cùng phân phối đều trên $[a; b]$. Xác định hàm phân phối của $Z = X + Y$; sau đó tính kỳ vọng và phương sai của Z^2 .
16. Xác suất để có lỗ hổng trong một vật đúc là 0,2. Tìm xác suất để trong 1000 vật đúc độ lệch của số vật đúc tốt (không có lỗ hổng) so với 800 không vượt quá 5%.
17. Cho độ lệch chuẩn của mỗi biến trong số 2500 biến ngẫu nhiên độc lập không quá 3 (đơn vị). Tìm xác suất để độ lệch tuyệt đối của trung bình cộng các biến đó so với trung bình cộng các kỳ vọng của chúng không vượt quá 0,3.
18. Gieo 1000 lần một đồng tiền cân đối đồng chất. Hãy đánh giá xác suất để tần suất xuất hiện mặt sấp lệch khỏi 0,5 sẽ không vượt quá 0,1. Tìm khoảng dao động của số lần xuất hiện mặt sấp tương ứng.

Chương IV

MẪU THỐNG KÊ VÀ ƯỚC LƯỢNG THAM SỐ

Từ chương này ta bắt đầu nghiên cứu thống kê, một lĩnh vực rộng tới mức khó có thể đưa ra một định nghĩa chung. Mặc dù vậy cũng có thể tóm tắt thống kê như là một khoa học về *phân tích dữ liệu* (bao gồm cả thu nhập và xử lý) nhằm thu nhận thông tin chân thực về đối tượng nghiên cứu với một độ tin cậy nhất định và rút ra những kết luận hợp lý. Những quyết định thống kê có ứng dụng to lớn như: dự báo, chẩn đoán, điều khiển ngẫu nhiên, kiểm tra chất lượng sản phẩm, thăm dò dư luận...

Cũng cần lưu ý rằng các vấn đề thống kê xuất hiện nếu có hai điều kiện: (i) có nhiều tình huống cần phải lựa chọn (chọn một hoặc một số); (ii) có các thông tin về các tình huống thông qua các dữ liệu thống kê. Trong giáo trình này chúng ta chủ yếu nghiên cứu việc xử lý dữ liệu số mà ta hay gọi là *xử lý số liệu*.

§1. MẪU VÀ THỐNG KÊ MÔ TẢ

1.1. Mẫu và tập đám đông

Trong công việc hàng ngày ta phải làm việc với các dãy số liệu. Chúng có thể là kết quả của việc đếm khi quan sát, của đo đạc nhờ các thiết bị đo, của tính toán trước đó... và cần được thu thập, lưu trữ và phân tích. Để làm được điều đó ta cần sắp xếp lại các số, tổng hợp và xử lý bước đầu nhằm tìm kiếm các thông tin quan trọng của tập số liệu. Phần công việc này và vấn đề thu thập các số liệu được mang tên gọi là *thống kê mô tả*.

Dãy số liệu thống kê thường được gọi là *mẫu*. Nó có nguồn gốc từ một tập lớn hơn mà ta sẽ gọi là *tập đám đông* hay *tập nền*. Chính vì thế mẫu sẽ mang thông tin nào đó về tập nền, mặc dù các thông tin đó có thể khác nhau ở những mẫu khác nhau. Sau này để cho xác định, ta giả sử rằng cả tập nền lẫn mẫu đều là tập các số cùng bản chất, đặc trưng cho một số khía cạnh nào đó của các đối tượng quan tâm. Các số đó chính là các giá trị khác nhau của một biến số. Nếu tập giá trị có thể có của biến số có số lượng hữu hạn, ta có biến rời rạc. Đối với các biến liên tục, số lượng giá trị là vô hạn không đếm được và tập số liệu chỉ phản ánh tập nền với một độ chính xác nhất định.

Muốn có đầy đủ thông tin về đối tượng nào đó, ta phải làm việc với tập nền. Tuy nhiên việc nghiên cứu tập nền sẽ vô cùng khó khăn vì:

- do nó quá lớn dẫn đến đòi hỏi quá nhiều chi phí vật chất và thời gian;

- do trình độ tổ chức và nghiên cứu hạn chế của đội ngũ khi làm việc với quy mô lớn, không nắm bắt và kiểm soát được quá trình nghiên cứu;

- do nhiều khi không thể làm được nếu tập nền biến động nhanh, các phần tử thay đổi thường xuyên, v.v...

Như vậy việc nghiên cứu trên tập nền, trừ các tập đủ bé, thường không thể thực hiện được. Từ đó đặt ra vấn đề chọn mẫu và nghiên cứu trên tập mẫu. Nếu mẫu được chọn ngẫu nhiên và với số lượng đủ, chúng ta hy vọng rằng việc xử lý chúng sẽ cho ta kết quả vừa nhanh vừa đỡ tốn kém mà vẫn đạt được độ chính xác và tin cậy cần thiết.

1.2. Vấn đề chọn mẫu

Ta mong muốn mẫu có tính đại diện tốt cho tập nền bởi vì việc nghiên cứu với mẫu như vậy cho ta độ tin cậy cao. Hiện nay

có nhiều phương pháp khác nhau để chọn mẫu, nhưng khó có thể nói rằng phương pháp nào là tốt nhất. Việc chọn phương pháp lấy mẫu phù hợp phụ thuộc vào chính tập đối tượng cụ thể và vào thói quen sở trường của nhà nghiên cứu.

1. Chọn mẫu ngẫu nhiên

Trong phương pháp chọn mẫu ngẫu nhiên, mỗi phần tử của tập nền đã có xác suất chọn xác định từ trước cả khi chọn mẫu. Mẫu ngẫu nhiên cho phép đánh giá khách quan hơn các đặc trưng của tập nền. Có 3 cách chọn như sau:

a) *Chọn mẫu ngẫu nhiên đơn giản* là phương pháp chọn mẫu có tính chất: mọi mẫu có cùng kích cỡ (cùng số phần tử) có cùng xác suất được chọn và mọi phần tử của tập nền có đồng khả năng lọt vào mẫu. Để việc chọn hoàn toàn ngẫu nhiên, ta có thể tiến hành theo kiểu bốc thăm hoặc dùng bảng số ngẫu nhiên. Ở đây để ý có hai phương thức chọn là *không hoàn lại* (mỗi phần tử chỉ được chọn một lần) và *có hoàn lại*. Nếu số lượng phần tử của mẫu khá bé so với tập nền thì kết quả lấy mẫu theo hai phương thức sai lệch không đáng kể. Do tính ngẫu nhiên nên mẫu có tính đại diện cao và tin cậy. Tuy nhiên phương pháp đòi hỏi phải biết toàn bộ tập nền và vì thế chi phí chọn mẫu khá lớn.

b) *Chọn mẫu phân nhóm*: Đầu tiên ta chia tập nền thành các nhóm tương đối thuần nhất, sau đó từ mỗi nhóm trích ra một mẫu ngẫu nhiên; tập hợp tất cả các mẫu đó cho ta một mẫu (ngẫu nhiên) phân nhóm. Người ta dùng phương pháp này khi trong nội bộ tập nền có những sai khác lớn. Nhà nghiên cứu phải có hiểu biết nhất định về cấu trúc tập nền để phân chia nhóm hợp lý. Sau này mỗi nhóm sẽ có vai trò khác nhau phụ thuộc vào độ quan trọng của chúng trong tập nền. Hạn chế của phương pháp là tính chủ quan khi phân chia nhóm. Nhưng nó vẫn hay được dùng do cách thức đơn giản khi làm việc với các nhóm đã khá bé và thuần nhất.

c) *Chọn mẫu chùm* chính là chọn một mẫu ngẫu nhiên của các tập con của tập nền, được gọi là các chùm. Ta cũng giả sử rằng các phần tử của mỗi chùm mang tính đại diện cho tập nền. Ngoài ra ta cố gắng sao cho mỗi chùm vẫn có độ phân tán cao như tập nền và đồng đều nhau về quy mô. Chẳng hạn ta muốn nghiên cứu nhu cầu tiêu thụ một mặt hàng nào đó bằng phương pháp chọn mẫu chùm: đầu tiên ta chia thành phố thành các khu dân cư, sau đó chọn ra một số khu làm phần tử của mẫu, cuối cùng ta nghiên cứu tất cả các gia đình sống trong các khu dân được chọn. Phương pháp này cho ta tiết kiệm kinh phí và thời gian (vì không phải di chuyển trên toàn thành phố), nhưng sai số có thể lớn hơn hai phương pháp trên.

2. *Chọn mẫu có suy luận*

Phương pháp chọn mẫu này dựa trên ý kiến các chuyên gia về đối tượng nghiên cứu. Như vậy việc chọn mẫu dựa trên hiểu biết và kinh nghiệm của một vài nhà chuyên môn. Tuy nhiên phương pháp này cũng có hạn chế cơ bản: Khi không có sự tham gia của các công cụ thống kê vào việc chọn mẫu tính khách quan rất khó được bảo đảm, từ đó kéo theo các kết luận mang nặng tính chủ quan. Tất nhiên điều đó không có nghĩa là không nên dùng các phương pháp chuyên gia. Rất rõ ràng chất lượng mẫu phụ thuộc nhiều vào trình độ của nhà nghiên cứu và kinh nghiệm của họ hy vọng trở thành một công cụ hữu hiệu.

1.3. *Phân loại và mô tả số liệu mẫu*

1. *Phân loại.* Giả sử từ một tập nền có N phần tử, ta chọn ra một mẫu có kích thước n , các phần tử của mẫu được ký hiệu là x_i , $i = \overline{1, n}$. Tập n giá trị $x_1, x_2, \dots, x_n$ tạo ra một *mẫu đơn*. Nhiều khi trong mẫu có nhiều giá trị giống nhau: chẳng hạn giá trị x_1 xuất hiện n_1 lần, x_2 xuất hiện n_2 lần, ..., x_k xuất hiện n_k lần; khi đó $n_1 + n_2 + \dots + n_k = n$. Trong thực hành có nhiều số liệu cho dưới dạng khoảng:

Thí dụ 1.1. Chiều cao của 300 học sinh 12 tuổi cho bởi bảng số liệu:

Ta để ý là trong bộ số liệu đó các khoảng có độ dài đều nhau (tuy nhiên nói chung độ dài đó có thể không đều). Trong trường hợp này ta có *mẫu lớp* (mẫu cho dưới dạng nhiều lớp là các khoảng không cắt nhau).

Chiều cao (cm)	Số lượng
117,5 – 122,5	9
122,5 – 127,5	33
127,5 – 132,5	74
132,5 – 137,5	93
137,5 – 142,5	64
142,5 – 147,5	21
147,5 – 152,5	6

2. Tần số và bảng tần số

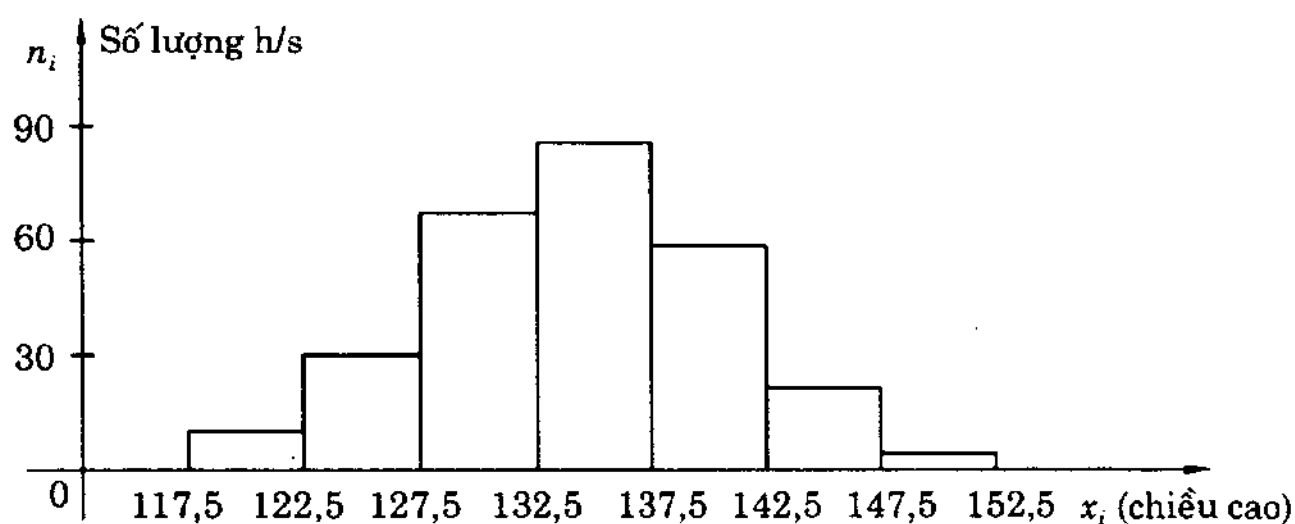
Số lần xuất hiện x_i hoặc một lớp thứ i nào đó, ký hiệu là n_i , được gọi là *tần số*. Sau khi sắp xếp số liệu theo thứ tự tăng của giá trị mẫu, ta có thể xây dựng *bảng tần số*. Bảng số trong thí dụ 1.1 chính là một bảng tần số (hay còn gọi là *phân phối tần số*). Bảng này bao gồm 7 lớp, mỗi lớp có độ dài 5 cm và toàn bộ có 300 số liệu đo chia thành các tần số của các lớp. Thông thường người ta hay chia các số liệu vào từ 5 đến 15 lớp phụ thuộc vào nhiều yếu tố như số lượng số liệu, mục đích xử lý... Để ý là nếu số lớp nhiều hơn, có thể làm tốt hơn các phân tích, nhưng việc cải thiện đó không nhiều, ngược lại nếu số lớp ít quá, có khả năng sẽ bị mất mát nhiều thông tin. Mỗi số liệu chỉ có mặt trong một lớp, độ dài mỗi lớp chính là hiệu của các giá trị lớn nhất và bé nhất.

Thông thường người ta hay biểu diễn phân phối tần số bằng đồ thị để quan sát và nghiên cứu trực giác hơn. Có hai dạng biểu diễn đồ thị hay dùng là biểu đồ và đa giác tần số.

a) Biểu đồ

Biểu đồ bao gồm các hình chữ nhật cạnh nhau có đáy bằng độ dài và chiều cao bằng số quan sát của lớp số liệu tương ứng.

Trên hình 1.1 cho ta biểu đồ ứng với bảng tần suất trong thí dụ 1.1.



Hình 1.1. Biểu đồ tần số

Rõ ràng diện tích các hình chữ nhật tỷ lệ với tần số của các lớp tương ứng.

b) Đa giác tần số

Đa giác tần số là đường gấp khúc nối các điểm có hoành độ x_i và tung độ n_i (hoặc các điểm có hoành độ ở giữa lớp số liệu thứ i và tung độ n_i). Đa giác tần số của thí dụ 1.1 vẽ trên hình 1.2.



Hình 1.2. Đa giác tần số

Ta thấy đa giác tần số dễ xây dựng hơn và dễ dùng hơn biểu đồ. Ngoài ra khi hiệu giữa hai hoành độ liên tiếp khá bé, đường gấp khúc sẽ càng ngày càng trơn và dần tiến tới dạng hàm mật độ xác suất.

3. Tần suất và phân phối thực nghiệm

Từ bảng tần số

x_i	x_1	x_2		x_i		x_k
n_i	n_1	n_2		n_i		n_k

nếu ta đặt $f_i = \frac{n_i}{n}$, $i = \overline{1, k}$, là *tần suất* xuất hiện giá trị x_i ở trong mẫu thì ta có thể mô tả bằng tần suất tương ứng. Rõ ràng từ định nghĩa f_i ta có $f_1 + f_2 + \dots + f_k = 1$ và bảng tần suất đó là

x_i	x_1	x_2		x_i		x_k
f_i	f_1	f_2		f_i		f_k

rất giống với bảng phân phối xác suất của một biến ngẫu nhiên rời rạc.

Nếu đặt w_i , $i = \overline{1, k}$, là tần số tích lũy của x_i và $F_n(x_i)$ là tần suất tích lũy của x_i , ta sẽ có

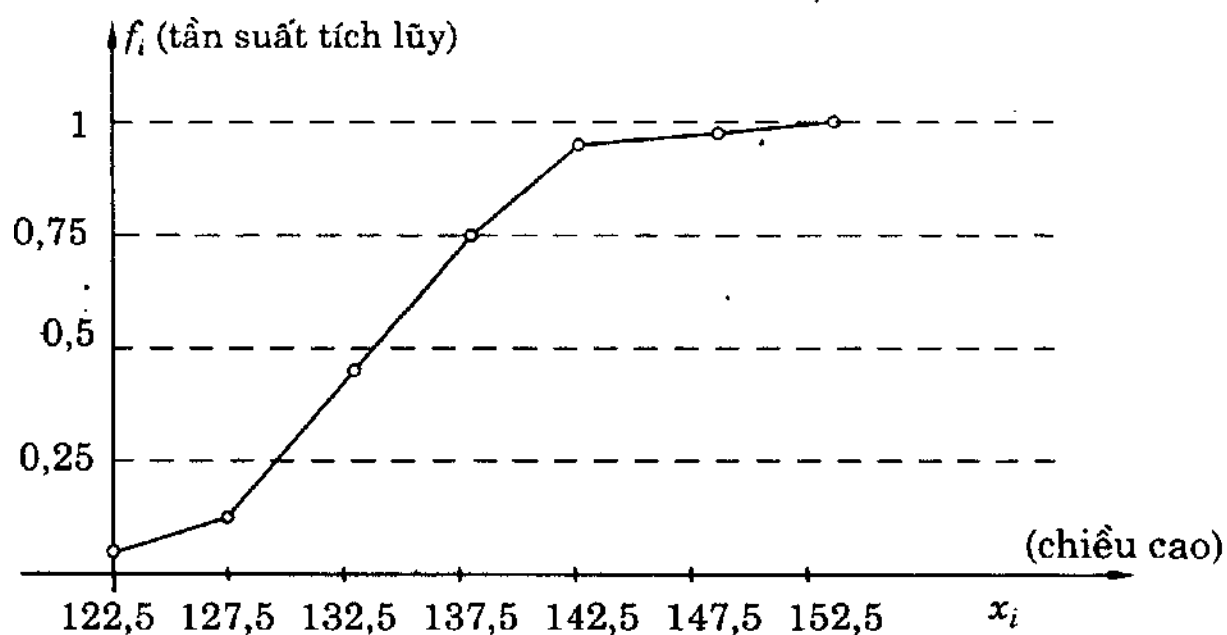
$$w_i = \sum_{x_j < x_i} n_j; F_n(x_i) = \frac{w_i}{n} = \sum_{x_j < x_i} f_j,$$

thì $F_n(x_i)$ là một hàm của x_i và được gọi là *hàm phân phối thực nghiệm* của mẫu hay là *hàm phân phối mẫu*. Chú ý rằng theo luật số lớn (định lý Béc-nu-li) $F_n(x) \xrightarrow[n \rightarrow \infty]{XS} F(x) = P(X < x)$, trong đó X là biến ngẫu nhiên gốc cảm sinh ra tập đám đông (và cả tập mẫu). Như vậy hàm phân phối mẫu có thể dùng để xấp xỉ luật phân phối của tập nền.

Thí dụ 1.2. Ta xây dựng bảng tần suất và tần suất tích lũy ứng với bộ số liệu của thí dụ 1.1.

Lớp	Tần số	Tần số tích lũy	Tần suất	Tần suất tích lũy
117,5 – 122,5	9	9	0,030	0,030
122,5 – 127,5	33	42	0,110	0,140
127,5 – 132,5	74	116	0,247	0,387
132,5 – 137,5	93	209	0,310	0,697
137,5 – 142,5	64	273	0,213	0,910
142,5 – 147,5	21	294	0,070	0,980
147,5 – 152,5	6	300	0,020	1,000

Tương tự như trên ta có thể xây dựng biểu đồ tần suất và đa giác tần suất tương ứng. Ngoài ra có thể vẽ được đồ thị của đa giác tần suất tích lũy hoặc tần số tích lũy (xem hình 1.3).



Hình 1.3. Đa giác tần suất tích lũy

§2. MẪU NGẪU NHIÊN VÀ CÁC ĐẶC TRƯNG MẪU

2.1. Mẫu ngẫu nhiên từ một tập nền

1. Mẫu ngẫu nhiên

Trong phân tích thống kê cổ điển người ta chấp nhận giả thiết rằng các phần tử của một tập đám đông nào đó đều được cảm sinh bởi một biến ngẫu nhiên gốc. Trong thực hành biến ngẫu nhiên gốc thường tuân theo luật phân phối chuẩn $\mathcal{N}(a, \sigma^2)$, hoặc chưa biết rõ dạng, hoặc chưa biết các tham số. Việc phân tích để xác định phân phối của tập nền sẽ dựa trên các số liệu mẫu.

Giả sử bây giờ ta tiến hành n phép thử độc lập để xác định các giá trị mẫu (biến ngẫu nhiên gốc của tập nền sẽ ký hiệu là X). Gọi X_i là biến ngẫu nhiên chỉ giá trị sẽ thu được ở phép thử thứ i , $i = \overline{1, n}$; rõ ràng các X_i sẽ tạo nên tập các biến ngẫu nhiên độc lập có cùng phân phối với X . Sau khi thử nghiệm, mỗi X_i sẽ có một giá trị xác định x_i , được gọi là các giá trị quan sát hay *thể hiện* của mẫu. Để đảm bảo tính đại diện của tập $\{x_1, \dots, x_n\}$ cho tập nền, ta cần dựa trên khái niệm mẫu ngẫu nhiên.

Định nghĩa 1. Ta gọi *mẫu ngẫu nhiên* kích thước n từ tập nền có biến ngẫu nhiên gốc X là một tập các biến $X_1, X_2, \dots, X_n$ thỏa mãn điều kiện:

- (i) độc lập thống kê,
- (ii) có cùng phân phối xác suất với biến X .

Các X_i thỏa mãn hai tính chất trên sẽ được gọi là các biến ngẫu nhiên độc lập và đồng phân phối. Như vậy khái niệm mẫu mà ta đưa vào tiết trước có thể hiểu như là một thể hiện của một mẫu ngẫu nhiên.

Để ý rằng giả thiết độc lập cho phép làm đơn giản rất nhiều các tính toán sau này. Chẳng hạn nếu biến gốc X rời rạc, có hàm xác suất $p(x)$, thì hàm xác suất đồng thời của $(X_1, X_2, \dots, X_n)$ sẽ là

$$p_n(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n p(x_i). \quad (2.1a)$$

Tương tự nếu biến X liên tục có mật độ $f(x)$ thì

$$f_n(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i). \quad (2.1b)$$

Như vậy từ (2.1) các phân phối đồng thời đã được biểu diễn đơn giản qua các phân phối biến thành phần. Ngoài ra luật phân phối đồng thời còn có thể phụ thuộc vào các tham số chưa biết.

2. Thống kê

Định nghĩa 2. Một hàm nào đó $Y = g(X_1, X_2, \dots, X_n)$ phụ thuộc vào tập giá trị của mẫu ngẫu nhiên được gọi là *một thống kê*.

Chú ý thống kê là một hàm đo được (khái niệm của lý thuyết hàm) và không phụ thuộc vào các tham số chưa biết. Do X_i nhận các giá trị tương ứng x_i , nên hàm $g(x_1, \dots, x_n)$ cũng được gọi là thống kê.

Thí dụ 2.1. Xét tập hợp giá trị mẫu $(x_1, x_2, \dots, x_n)$, các hàm sau đây sẽ được gọi là các thống kê:

$$a) \ g(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{X};$$

$$b) \ g(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2;$$

$$c) \ g(x_1, \dots, x_n) = (x_{(1)}, \dots, x_{(n)}), \text{ trong đó } x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Thống kê trong c) có tên gọi là *thống kê hạng* (trong đó $\{x_{(i)}\}$ là dãy các x_i đã được sắp thứ tự), và giá trị của một thống kê trong định nghĩa 2 có thể là một véc tơ (g là véc tơ hàm).

2.2. Các đặc trưng mẫu

Từ nay về sau, trong các công thức liên quan đến đặc trưng mẫu, thay vì X_i ta hay dùng x_i do nhiều lý do. Thứ nhất, đó là các công thức tính toán làm việc trực tiếp với các quan sát mẫu cụ thể. Thứ hai, nếu dùng quá nhiều ký hiệu khác nhau sẽ gây nhầm lẫn (hơn nữa về mặt biện chứng trong ngẫu nhiên có tất định và ngược lại). Thứ ba, các công thức chủ yếu dùng để tính toán, còn trong các trường hợp chứng minh các tính chất lý thuyết, ta dễ dàng (và nên cũng để tránh nhầm) thay trở lại các giá trị mẫu x_i bằng X_i .

Một mẫu, như ta đã biết ở tiết 1, có thể mô tả bằng bảng phân phối tần số hoặc bằng chính dãy số liệu

$$a) x_1, x_2, \dots, x_n \quad (2.2)$$

$$b) \begin{array}{c|cccc} x_i & x_1 & \dots & x_k \\ \hline n_i & n_1 & \dots & n_k \end{array} \quad (n_1 + n_2 + \dots + n_k = n) \quad (2.3)$$

Trong trường hợp mẫu lớp, nhiều khi thay khoảng giá trị bằng giá trị trung bình của khoảng; khi đó ta đưa về mẫu đơn dạng (2.3).

1. Trung bình mẫu (hay kỳ vọng mẫu)

Nếu mẫu cho dưới dạng (2.2) thì trung bình mẫu ký hiệu là $\bar{X}$, được xác định như sau:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (2.4)$$

Để ý là khi chứng minh lý thuyết, ta sẽ thay các x_i bằng X_i là biến ngẫu nhiên cảm sinh ra quan sát x_i , có cùng phân phối với X gốc. Nếu số liệu cho dưới dạng (2.3), ta có

$$\bar{X} = \frac{1}{n} \sum_{i=1}^k x_i n_i. \quad (2.5)$$

Về mặt bản chất (2.4) và (2.5) là một (nếu $k = n$, thì $n_i = 1 \forall i$), mặc dù vậy trên hình thức ta vẫn để riêng dưới dạng hai công thức khác nhau.

Rõ ràng $\bar{X}$ theo cách hiểu lý thuyết sẽ là một biến ngẫu nhiên (do các X_i là biến ngẫu nhiên), nên có thể tìm các số đặc trưng của $\bar{X}$. Giả sử biến ngẫu nhiên gốc X có $EX = a$ và $VX = \sigma^2$; khi đó

$$E\bar{X} = a, V\bar{X} = \frac{\sigma^2}{n}. \quad (2.6)$$

Ta chứng minh công thức bên trái: do $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ nên dùng tính chất của kỳ vọng $E\bar{X} = \frac{1}{n}(EX_1 + \dots + EX_n)$; từ định nghĩa mẫu ngẫu nhiên các X_i có cùng phân phối với X nên $EX_i = EX = a$, suy ra $E\bar{X} = \frac{1}{n}(na) = a$. Công thức bên phải của (2.6) đã được chứng minh ở phần tính chất của phương sai. Từ (2.6), do phương sai $V\bar{X}$ bé hơn n lần VX , nên các giá trị có thể có của $\bar{X}$ sẽ ổn định quanh kỳ vọng hơn các giá trị của X .

Chú ý rằng nếu tập nền có kích thước bé (N bé) và ta chọn mẫu không hoàn lại, công thức $V\bar{X}$ trong (2.6) phải nhân thêm với thừa số hiệu chỉnh $(N - n)/(N - 1)$:

$$V\bar{X} = \frac{\sigma^2}{n} \cdot \frac{N - n}{N - 1}. \quad (2.7)$$

Ta xét ý nghĩa của (2.7) trong các trường hợp đặc biệt. Nếu chọn mẫu có $n = N$, tức là lấy toàn bộ các phần tử của tập nền, khi đó mọi thông tin của tập nền đã biết và rõ ràng $V\bar{X} = 0$. Trong mọi trường hợp ta chọn ra mẫu rất bé so với tập nền (chẳng hạn N vô hạn hoặc chọn mẫu có hoàn lại với trường hợp N lớn và hữu hạn), rõ ràng $V\bar{X}$ trở thành như trong (2.6) vì

$$\lim_{N \rightarrow \infty} \frac{N - n}{N - 1} = 1.$$

Thí dụ 2.2. Ta có năm mảnh bìa được đánh số từ 1 đến 5. Nếu gọi X số thu được khi rút hủ họa ra một mảnh bìa thì rõ ràng phân phối của X là

x	1	2	3	4	5
$p(x)$	0,2	0,2	0,2	0,2	0,2

Giả sử bây giờ ta lấy ra một mẫu 2 mảnh bìa không hoàn lại và thu được số X_1 và X_2 . Hãy tìm phân phối của $\bar{X}$ và các số đặc trưng của nó.

Giải. Rõ ràng $X \sim \mathcal{U}(5)$ với $EX = 3$ và $VX = 2$ (xem §4 chương II). Mặt khác đặt $\bar{X} = (X_1 + X_2)/2$, có thể tính được luật phân phối của $\bar{X}$

$\bar{x}$	1,5	2	2,5	3	3,5	4	4,5
$p(x)$	0,1	0,1	0,2	0,2	0,2	0,1	0,1

Dễ dàng tính được

$$E\bar{X} = 1,5 \cdot 0,1 + 2 \cdot 0,1 + 2,5 \cdot 0,2 + 3 \cdot 0,2 + 3,5 \cdot 0,2 + 4 \cdot 0,1 + 4,5 \cdot 0,1 = 3;$$

$$V\bar{X} = (1,5 - 3)^2 \cdot 0,1 + (2 - 3)^2 \cdot 0,1 + (2,5 - 3)^2 \cdot 0,2 + (3 - 3)^2 \cdot 0,2 + (3,5 - 3)^2 \cdot 0,2 + (4 - 3)^2 \cdot 0,1 + (4,5 - 3)^2 \cdot 0,1 = 0,75.$$

Ta thấy $E\bar{X} = EX$; $V\bar{X} = 0,75 < VX$. Để ý nếu chọn có hoàn lại, ta có phương sai được tính theo (2.6) và bằng 1. Từ đó theo (2.7) $V\bar{X} = 1 \cdot \frac{N-n}{N-1} = 1 \cdot \frac{5-2}{5-1} = 0,75$ như ở trên. Cũng lưu ý rằng

khi chọn mẫu không hoàn lại, X_2 đã không cùng phân phối như X nữa nên việc áp dụng (2.6) là không được phép.

2. Phương sai mẫu

Nếu mẫu cho dưới dạng (2.2), phương sai mẫu, ký hiệu là S^2 , được xác định như sau:

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2; \quad (2.8)$$

với $\bar{X}$ xác định theo (2.4). Nếu mẫu cho dưới dạng (2.3), ta có

$$S^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{X})^2 n_i; \quad (2.9)$$

với $\bar{X}$ xác định theo (2.5). Do S^2 là biến ngẫu nhiên, ta tìm số đặc trưng ES^2 :

$$\begin{aligned} \text{Ta viết } S^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 \\ &= \frac{n-1}{n^2} \sum_{i=1}^n X_i^2 - \frac{1}{n^2} \sum_{i \neq j} X_i X_j; \end{aligned}$$

do $X_i, i = \overline{1, n}$, độc lập đồng phân phối với X , nên $E(X_i X_j) = EX_i \cdot EX_j = (EX)^2$ và $E(X_i^2) = E(X^2)$ nên

$$ES^2 = \frac{n-1}{n^2} \cdot nE(X^2) - \frac{n(n-1)}{n^2} (EX)^2 = \frac{n-1}{n} VX = \frac{n-1}{n} \sigma^2. \quad (2.10)$$

Chính vì $ES^2 \neq \sigma^2$, nên người ta đưa vào đặc trưng mẫu thứ hai của phương sai với tên gọi là *phương sai mẫu hiệu chỉnh*, ký hiệu là s^2 , như sau (so sánh với (2.8) và (2.9)):

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2; \quad (2.11)$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i - \bar{X})^2 n_i; \quad (2.12)$$

Rõ ràng $s^2 = \frac{1}{n-1} S^2$ và $Es^2 = \frac{n}{n-1} ES^2 = \sigma^2$ (xem (2.10)).

Ngoài trung bình mẫu và các phương sai mẫu, ta còn có thể xác định các đặc trưng mẫu khác:

– mô men mẫu cấp k $M_k = \frac{1}{n} \sum_{i=1}^n x_i^k$;

– mô men trung tâm mẫu cấp k $S_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^k$;

– trung vị mẫu, một mẫu...

3. Về luật phân phối của các đặc trưng mẫu

Nếu biến ngẫu nhiên gốc tuân theo luật phân phối chuẩn $X \sim \mathcal{N}(a, \sigma^2)$, khi đó $\bar{X}$ và S^2 độc lập với nhau và

$$\text{a) } \bar{X} \sim \mathcal{N}\left(a, \frac{\sigma^2}{n}\right) \text{ hay } \frac{\bar{X} - a}{\sigma} \sqrt{n} \sim \mathcal{N}(0; 1); \quad (2.13)$$

$$\text{b) } \frac{nS^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1); \quad (2.14)$$

$$\text{c) } \frac{\bar{X} - a}{s} \sqrt{n} = \frac{\bar{X} - a}{S} \sqrt{n-1} \sim t(n-1). \quad (2.15)$$

Các kết quả này đã được nhắc tới trong các công thức (4.17), (4.21)... của chương II. Ngoài ra, nếu ta xét hai mẫu khác nhau cảm sinh bởi hai biến ngẫu nhiên chuẩn và gọi s_1^2 và s_2^2 là các phương sai mẫu hiệu chỉnh của các mẫu tương ứng (với kích thước n_1 và n_2) thì với giả thiết $\sigma_1^2 = \sigma_2^2$

$$\text{d) } \frac{s_1^2}{s_2^2} \sim \mathcal{F}(n_1 - 1; n_2 - 1) \quad (2.16)$$

$\left(\text{nếu } \sigma_1^2 \neq \sigma_2^2, \frac{s_2^2}{\sigma_1^2} \bigg/ \frac{s_2^2}{\sigma_2^2} \sim \mathcal{F}(n_1 - 1; n_2 - 1) \right)$. Cuối cùng để ý

trong (2.14) nếu thay $\bar{X}$ bằng a ta sẽ có $\sum_{i=1}^n (x_i - a)^2 / \sigma^2 \sim \chi^2(n)$.

Với các giả thiết tồn tại các giới hạn hoặc mômen tương ứng và dùng các kết quả của luật số lớn hoặc định lý giới hạn trung tâm, khi $n \rightarrow \infty$ ta sẽ có

$$a) \bar{X} \xrightarrow{hcc} a; S^2 \xrightarrow{hcc} \sigma^2, s^2 \xrightarrow{hcc} \sigma^2, \text{ từ đó suy ra}$$

$$\bar{X} \xrightarrow{XS} a; S^2 \xrightarrow{XS} \sigma^2, s^2 \xrightarrow{XS} \sigma^2;$$

$$b) \frac{\bar{X} - a}{\sigma} \sqrt{n} \xrightarrow{L} \mathcal{N}(0, 1),$$

$$\frac{\bar{X} - a}{s} \sqrt{n} \xrightarrow{L} \mathcal{N}(0, 1);$$

$$c) \frac{S^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \sqrt{n} \xrightarrow{L} \mathcal{N}(0, 1),$$

$$\frac{s^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \sqrt{n} \xrightarrow{L} \mathcal{N}(0, 1),$$

$$d) (s - \sigma) \sqrt{n} \xrightarrow{L} \mathcal{N}\left(0, \frac{1}{2\sigma} \sqrt{\mu_4 - \sigma^4}\right).$$

Các kết quả trên sẽ rất có ích trong thực hành vì không cần đến giả thiết chuẩn của biến ngẫu nhiên gốc và trong nhiều trường hợp ta đã có thể chấp nhận kết quả với n không quá lớn. Chẳng hạn với $n > 30$, kết quả (b) đã có thể chấp nhận được. Ngoài ra theo định lý Gli-ven-cô – Can-te-li, khi n đủ lớn hàm phân phối thực nghiệm đã khá gần với hàm lý thuyết.

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n \rightarrow \infty]{hcc} 0.$$

2.3. Vấn đề tính toán các đặc trưng mẫu

1. Mẫu đơn

Nếu mẫu cho dưới dạng (2.2) $x_1, x_2, \dots, x_n$, ta tính $\bar{X}$ và S^2 , s^2 theo các công thức (2.4), (2.8) và (2.11). Trong nhiều trường hợp, người ta thay tổng trong (2.8) và (2.11) bằng

$$\sum_{i=1}^n x_i^2 - n(\bar{X})^2$$

để tính hơn. Trung vị thực nghiệm chính là giá trị thứ $\frac{n+1}{2}$ của tập mẫu đã sắp xếp (nếu n lẻ thì đó là giá trị chính giữa dãy số liệu, nếu n chẵn ta lấy trung bình cộng của hai giá trị chính giữa).

Nếu mẫu cho dưới dạng (2.3), tức là các giá trị mẫu có tần số xuất hiện khác 1, khi đó ta dùng các công thức (2.5), (2.9) và (2.12). Trong cách tính trực tiếp, giống như ở trên ta thay tổng trong (2.9) và (2.12) bằng

$$\sum_{i=1}^k x_i^2 n_i - n(\bar{X})^2.$$

Tuy nhiên có thể dùng một cách tính rút gọn hơn theo các bước sau:

B1. Chọn một giá trị trung bình tùy ý x_0 .

B2. Gọi h là khoảng cách đều giữa hai số liệu liên tiếp và tính $d_i = \frac{x_i - x_0}{h}$.

B3. Tính các tổng $\sum_{i=1}^k d_i n_i$ và $\sum_{i=1}^k d_i^2 n_i$.

B4. Tính $\bar{X} = x_0 + \frac{h}{n} \sum d_i n_i$; (2.17)

$$s^2 = \frac{h^2}{n-1} \left[\sum d_i^2 n_i - \frac{(\sum d_i n_i)^2}{n} \right] \quad (2.18a)$$

hoặc
$$S^2 = \frac{h^2}{n} \left[\sum d_i^2 n_i - \frac{(\sum d_i n_i)^2}{n} \right] \quad (2.18b)$$

Hạn chế của cách tính này là thường đòi hỏi số liệu cách đều (nhưng trong nhiều bài toán thực tế lại chấp nhận được). Các kết quả trung gian được đưa vào một bảng tính nên việc kiểm tra lại kết quả khá dễ dàng và tiện lợi.

Thí dụ 2.3. Người ta cân 150 con vịt của một giống mới, kết quả như sau

Cân nặng	1,25	1,50	1,75	2,00	2,25	2,50	2,75	3,00
Số con	2	6	24	35	39	24	14	6

Hãy tính các đặc trưng mẫu của trọng lượng vịt.

Giải. Ta chọn $x_0 = 2,25$, $h = 0,25$ và

x_i	n_i	d_i	$n_i d_i$	$d_i^2 n_i$
1,25	2	-4	-8	32
1,50	6	-3	-18	54
1,75	24	-2	-48	96
2,00	35	-1	-35	35
2,25	39	0	0	0
2,50	24	1	24	24
2,75	14	2	28	56
3,00	6	3	18	54
Σ	150		-39	351

$$\text{Từ đó } \bar{X} = x_0 + \frac{h}{n} \sum d_i n_i = 2,25 - \frac{0,25}{150} \cdot 39 = 2,185;$$

$$\begin{aligned} S^2 &= \frac{h^2}{n} \left[\sum d_i^2 n_i - \frac{1}{n} \left(\sum d_i n_i \right)^2 \right] \\ &= \frac{0,25^2}{150} \left(351 - \frac{39^2}{150} \right) = 0,142025. \end{aligned}$$

Dễ dàng thấy trung vị mẫu là 2,25, đồng thời đó cũng là giá trị thực nghiệm của một.

2. Mẫu lớp

Mẫu lớp được cho dưới dạng

$$\frac{[x_0, x_1] \quad [x_1, x_2] \quad \dots \quad [x_{k-1}, x_k]}{n_1 \quad n_2 \quad \dots \quad n_k} \quad (2.19)$$

trong đó giá trị mẫu là một khoảng số từ x_{i-1} đến x_i . Trong trường hợp này các đặc trưng $\bar{X}$ và S^2, s^2 chỉ có thể được tính gần đúng. Ta sẽ chuyển mẫu từ dạng (2.19) về dạng (2.3) bằng cách thay các khoảng số bằng giá trị trung bình của khoảng. Như vậy, việc tính $\bar{X}$ và S^2, s^2 đưa về trường hợp mẫu đơn.

Trong thực hành, đối với một và trung vị mẫu, người ta sử dụng các công thức sau đây (ký hiệu một và trung vị là *Mod* và *Med*)

$$Mod = x_{mo} + \frac{d_t}{d_t + d_s} h, \quad (2.20)$$

trong đó x_{mo} – điểm đầu của khoảng một,

d_t – hiệu tần số của khoảng một và khoảng trước,

d_s – hiệu tần số của khoảng một và khoảng sau,

h – độ dài khoảng;

$$Med = x_{me} + \frac{n/2 - n_{tl}}{n_{me}} h, \quad (2.21)$$

trong đó x_{me} – điểm đầu của trung vị,

n_{tl} – tần số tích lũy trước khoảng trung vị,

n_{me} – tần số khoảng trung vị,

h – độ dài khoảng;

n – tổng tần số hay kích thước mẫu.

Kết quả tính toán được minh họa trong thí dụ sau:

Thí dụ 2.4. Tính các đặc trưng nẫu của thí dụ 1.1.

Giải. Ta lập bảng tính (chọn $x_0 = 135$ trong công thức (2.17), còn $h = 5$)

Khoảng	TB	n_i	d_i	$d_i n_i$	$d_i^2 n_i$	n_{tl}
117,5 – 122,5	120	9	–3	–27	81	9
122,5 – 127,5	125	33	–2	–66	132	42
127,5 – 132,5	130	74	–1	–74	74	116
132,5 – 137,5	135	93	0	0	0	209
137,5 – 142,5	140	64	1	64	64	273
142,5 – 147,5	145	21	2	42	84	294
147,5 – 152,5	150	6	3	18	54	300
Σ		150		–43	489	

Theo các công thức (2.17), (2.18)

$$\bar{X} = 135 + \frac{5}{300}(-43) = 134,2823;$$

$$S^2 = \frac{5^2}{300} \left(489 - \frac{43^2}{300} \right) \approx 40,2364.$$

Để tính mốt và trung vị mẫu theo (2.20) và (2.21) ta thấy

$$x_{mo} = 132,5; d_l = 93 - 74 = 9; d_s = 93 - 64 = 19;$$

$$x_{me} = 132,5; n_{tl} = 116; n_{me} = 93;$$

từ đó

$$Mod = 132,5 + \frac{9}{9+19} \cdot 5 \approx 134,1072;$$

$$Med = 132,5 + \frac{150 - 116}{93} \cdot 5 \approx 134,3279.$$

§3. ƯỚC LƯỢNG ĐIỂM

3.1. Ước lượng tham số

Khái niệm ước lượng thường được dùng trong thực tế, chẳng hạn để đánh giá trình độ học sinh ta tính điểm trung bình. Đó là một ước lượng của điểm số học sinh ấy, nó dựa trên thông tin quá khứ là các điểm mà học sinh đã nhận được trong học kỳ.

Bài toán *ước lượng tham số* có thể phát biểu tổng quát như sau: Cho biến ngẫu nhiên gốc X có luật phân phối xác suất đã biết nhưng chưa biết tham số θ nào đó; ta phải xác định giá trị của θ dựa trên các thông tin thu được từ một mẫu quan sát $x_1, x_2, \dots, x_n$ của X . Quá trình đi xác định một tham số θ chưa biết được gọi là quá trình ước lượng tham số. Giá trị tìm được trong quá trình ấy, ký hiệu là $\hat{\theta}$, sẽ được gọi là *ước lượng* của θ , ở đây do $\hat{\theta}$ là một giá trị số nên nó được gọi là *ước lượng điểm*, sau này ta còn có ước lượng khoảng hay khoảng tin cậy. Chú ý là θ sau này có thể nhiều chiều và $\hat{\theta}$ sẽ là một điểm trong không gian nhiều chiều tương ứng.

Rõ ràng $\hat{\theta} = g(x_1, x_2, \dots, x_n)$ là một hàm của các giá trị mẫu, hay là một thống kê. Để đánh giá một ước lượng là tốt hay không, ta phải so sánh nó với giá trị θ thật, nhưng θ chưa biết. Vì vậy sau này phải đưa ra các tiêu chuẩn để đánh giá chất lượng của thống kê $\hat{\theta}$ như là một xấp xỉ tốt nhất của θ . Những tiêu chuẩn như vậy cho ta các *nguyên lý thống kê* khác nhau.

Nói chung, do nhiều lý do, ta không thể xác định được θ chính xác. Việc chọn một ước lượng $\hat{\theta}$ nào đó khó có thể gọi là tối ưu, bao giờ ta cũng phải chịu một *tổn thất*. Trong thống kê, người ta thường lấy hàm tổn thất dạng bình phương $L(g, \theta) = (g - \theta)^2$. Trong nhiều bài toán thực tế việc chọn hàm tổn thất như trên bảo đảm được yêu cầu cần thiết. Nếu hàm tổn thất L

có dạng khác, ta hoàn toàn có thể xấp xỉ nó bằng dạng bình phương như trên với những giả thiết về tính lồi trong một lân cận nào đó của θ (cùng với giả thiết về liên tục và khả vi hai lần); khi đó ta có thể khai triển L tại lân cận đó của θ .

$$L(g, \theta) = L(g_0, \theta) + \frac{\partial L(g_0, \theta)}{\partial g}(g - \theta) + \frac{1}{2} \frac{\partial^2 L(g_1, \theta)}{\partial g^2}(g - \theta)^2, \quad (3.1)$$

trong đó g_1 nằm giữa g và θ . Rõ ràng

+ $L(g_0, \theta) = 0$ (tổng thất cực tiểu, nếu $g_0 = \theta$);

+ $\frac{\partial L(g_0, \theta)}{\partial g} = 0$ tại $g_0 = \theta$ do muốn tổng thất cực tiểu;

+ với $g_1 \in$ lân cận θ , đạo hàm cấp 2 ngặt dương ở g_1 (do giả thiết lồi của L), từ đó $L(g, \theta)$ hoàn toàn có thể xấp xỉ bằng $(g - \theta)^2$, ít ra ở lân cận của θ . Chú ý là $g = g(X_1, X_2, \dots, X_n)$ nên sau này người ta thường làm cực tiểu *hàm rủi ro*

$$R(g, \theta) = E[L(g, \theta)].$$

3.2. Các tính chất của ước lượng điểm

Ở đây ta quan tâm đến ước lượng điểm của θ , ký hiệu là $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$ hay về mặt lý thuyết $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$.

1. Ước lượng không chệch

Định nghĩa 1. Thống kê $\hat{\theta}$ được gọi là *ước lượng không chệch* của θ nếu $E\hat{\theta} = \theta$.

Từ định nghĩa trên ta thấy $E(\hat{\theta} - \theta) = 0$, điều đó có nghĩa là trung bình độ lệch của ước lượng so với giá trị thật bằng 0. Nếu độ lệch có trung bình khác 0, ta có ước lượng chệch. Một sai số nào đó có trung bình khác không sẽ được gọi là *sai số hệ thống*; ngược lại sẽ là *sai số ngẫu nhiên*. Như vậy một ước lượng sẽ được gọi là không chệch khi độ lệch so với giá trị thật (sai số ước lượng) là sai số ngẫu nhiên.

Dựa vào các kết quả của mục 2.2 rõ ràng ta có

- trung bình mẫu là ước lượng không chệch của kỳ vọng,
- phương sai mẫu hiệu chỉnh là ước lượng không chệch của phương sai,

- tần suất mẫu là ước lượng không chệch của xác suất xuất hiện sự kiện A nào đó (nếu X có phân phối Béc-nu-li và việc lấy mẫu có hoàn lại),

- phương sai tính theo công thức $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ là ước lượng

không chệch của phương sai, trong khi đó S^2 là ước lượng chệch.

2. Ước lượng vững

Định nghĩa 2. Thống kê $\hat{\theta}$ được gọi là *ước lượng vững* của θ , nếu $\hat{\theta} (x_1, \dots, x_n) \xrightarrow[n \rightarrow \infty]{XS} \theta$.

Sử dụng khái niệm hội tụ theo xác suất ở chương III, ta có thể thấy rằng nếu $\hat{\theta}$ là ước lượng tiệm cận không chệch của θ (tức là $\lim_{n \rightarrow \infty} E\hat{\theta} = \theta$) và $\lim_{n \rightarrow \infty} V\hat{\theta} = 0$ thì $\hat{\theta}$ sẽ là ước lượng vững.

Rõ ràng $\bar{X}$ và S^2 (hoặc s^2) là các ước lượng vững tương ứng của EX và VX , tần suất mẫu là ước lượng vững của xác suất tương ứng.

3. Ước lượng hiệu quả

Trong lớp các ước lượng không chệch của θ việc so sánh hai ước lượng theo nghĩa tổn thất đưa về so sánh hai phương sai.

Định nghĩa 3. Thống kê $\hat{\theta}$ được gọi là *ước lượng hiệu quả* của θ , nếu nó là ước lượng không chệch có phương sai bé nhất.

Người ta đã chứng minh được rằng nếu $\hat{\theta}$ là ước lượng hiệu quả của θ thì phương sai của nó là

$$V\hat{\theta} = \frac{1}{nE \left[\frac{\partial \ln f(x, \theta)}{\partial \theta} \right]^2}, \quad (3.2)$$

trong đó $f(x, \theta)$ là hàm mật độ của biến ngẫu nhiên gốc cảm sinh ra tập mẫu đang xét. Như vậy với mọi ước lượng không chệch bất kỳ của θ ta luôn có phương sai lớn hơn $V\hat{\theta}$ trong (3.2), sau này (3.2) được gọi là *giới hạn Cra-me – Rao*.

Thí dụ 3.1. Nếu biến ngẫu nhiên gốc $X \sim \mathcal{N}(a, \sigma^2)$ thì trung bình mẫu $\bar{X}$ là ước lượng hiệu quả của kỳ vọng $EX = a$.

Giải. Ta đã biết $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(a, \frac{\sigma^2}{n}\right)$.

Mặt khác X có phân phối chuẩn, nên nếu $f(x, a)$ là hàm mật độ của X ,

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

$$\frac{\partial}{\partial a} \ln f(x, a) = \frac{x-a}{\sigma^2}.$$

$$\text{Vậy } nE\left[\frac{\partial \ln f(x, a)}{\partial a}\right]^2 = nE\left(\frac{x-a}{\sigma^2}\right)^2 = \frac{n}{\sigma^2}.$$

và $V\bar{X}$ chính bằng nghịch đảo σ^2/n . Vậy $\bar{X}$ là ước lượng hiệu quả của a .

Bạn đọc hãy chứng minh tần suất mẫu f là ước lượng hiệu quả của xác suất biến ngẫu nhiên gốc X tuân theo luật Bé-nu-li. Để ý là nghịch đảo của (3.2) thường mang tên gọi là *lượng thông tin Phi-sơ* của mẫu tương ứng hay dùng trong lý thuyết thông tin.

3.3. Các phương pháp ước lượng

1. Sử dụng các đặc trưng mẫu

Cách ước lượng này đưa về sử dụng các đặc trưng đi nhắc tới ở §2 là trung bình mẫu, phương sai mẫu hiệu chỉnh, tộ lệch chuẩn mẫu hiệu chỉnh...

2. Phương pháp mômen

Đây là phương pháp thực nghiệm dựa trên sự kiện lý thuyết là các mô men mẫu của một tập mẫu ngẫu nhiên có biến gốc X hội tụ hầu chắc chắn về các mô men lý thuyết tương ứng của X . Như vậy nếu ký hiệu θ bây giờ là véc tơ k chiều $\theta = (\theta_1, \dots, \theta_k)$, $m_j(\theta)$ là mô men lý thuyết cấp j , $m_j(e, n)$ – mômen thực nghiệm cấp j , ước lượng theo phương pháp mô men của véc tơ tham số θ được tìm bằng cách giải hệ phương trình

$$\begin{cases} m_1(\theta) = m_1(e, n), \\ \vdots \\ m_k(\theta) = m_k(e, n). \end{cases}$$

Thí dụ 3.2. Cho biến gốc X tuân theo luật gam-ma $X \sim \gamma(r, \lambda)$. Dùng phương pháp mô men tìm ước lượng của r và λ .

Giải. Từ kết quả chương II, ta đưa về giải hệ

$$\begin{cases} EX = \frac{r}{\lambda} = \bar{X}; \\ VX = \frac{r}{\lambda^2} = S^2. \end{cases}$$

Suy ra các ước lượng cần tìm là

$$\hat{\lambda} = \frac{\bar{X}}{S^2}; \quad \hat{r} = \frac{(\bar{X})^2}{S^2}.$$

3. Phương pháp hợp lý nhất

Nguyên lý hợp lý nhất là tìm giá trị của θ – hàm của quan sát $(x_1, \dots, x_n)$ sao cho bảo đảm xác suất thu được các quan sát đó lớn nhất. Giả sử biến gốc X có phân phối (hàm mật độ) là $f(x, \theta)$; khi đó hàm hợp lý, ký hiệu là $L(x, \theta)$, x ở đây là véc tơ $(x_1, x_2, \dots, x_n)$, θ cũng có thể là véc tơ,

$$L(x, \theta) = \prod_{i=1}^n f(x_i, \theta). \quad (3.3)$$

Để ý là hàm hợp lý $L(x, \theta)$ có thể không khả vi đối với θ . Ta gọi $\hat{\theta}$ là *ước lượng hợp lý nhất* của θ nếu $\forall \theta$ (thuộc tập tham số nào đó)

$$L(x, \hat{\theta}) \geq L(x, \theta). \quad (3.4)$$

Việc tìm $\hat{\theta}$ thỏa mãn (3.4) rất khó khăn do hàm hợp lý (3.3) không là hàm lồi và tất nhiên thường phi tuyến. Không có lý do nào để đảm bảo cho $\hat{\theta}$ thỏa mãn (3.4) là duy nhất, hoặc là không chệch (và vì thế không thể hiệu quả).

Nếu đảm bảo các giả thiết về khả vi hai lần của hàm hợp lý, ta có tìm hiểu điều kiện cần để có cực trị:

$$\frac{\partial L(x, \theta)}{\partial \theta} = 0$$

hoặc tương đương với nó

$$\frac{\partial \ln L(x, \theta)}{\partial \theta} = 0. \quad (3.5)$$

(3.5) có tên gọi là *phương trình hợp lý nhất*, nhưng nghiệm của nó không duy nhất và vì vậy chưa chắc đã là nghiệm cần tìm. Vì vậy ta cần kiểm tra điều kiện đủ

$$\left. \frac{\partial^2 \ln L(x, \theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} < 0. \quad (3.5)$$

Để ý trong trường hợp rời rạc, $f(x_i, \theta)$ trong (3.3) phải được thay bằng hàm xác suất $p(x_i, \theta)$. Nếu θ là một véc tơ tham số, các đạo hàm trong (3.5) phải hiểu đạo hàm theo véc tơ. May sao người ta đã chứng minh được rằng nếu phương trình (3.5) có nghiệm duy nhất thì khi đó không cần kiểm tra điều kiện đủ (3.6).

Thí dụ 3.3. Tìm ước lượng hợp lý nhất của tham số λ trong phân phối Poa-xông $\mathcal{P}(\lambda)$.

$$\text{Giải. } L(x_1, \dots, x_n, \lambda) = e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} \frac{1}{\prod_{i=1}^n x_i!} \quad (\text{theo 3.3});$$

$$\frac{\partial \ln L(x_1, \dots, x_n, \lambda)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^n x_i = 0$$

có nghiệm duy nhất $\Rightarrow \hat{\lambda} = \bar{X}$.

Cũng dễ dàng kiểm tra (để ý các x_i đều dương):

$$\left. \frac{\partial^2 \ln L}{\partial \lambda^2} \right|_{\lambda=\bar{X}} = -n \frac{\bar{X}}{\lambda^2} \Big|_{\lambda=\bar{X}} = -\frac{n}{\bar{X}} < 0.$$

Thí dụ 3.4. Tìm ước lượng hợp lý nhất của các tham số a và σ^2 của phân phối chuẩn $\mathcal{N}(a, \sigma^2)$.

$$\text{Giải. } L(x_1, \dots, x_n, a, \sigma^2) = (2\sigma^2\pi)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^n (x_i - a)^2}{2\sigma^2}}$$

$$\text{Từ đó } \begin{cases} \frac{\partial \ln L}{\partial a} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - a) = 0, \\ \frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - a)^2 = 0. \end{cases}$$

Giải hệ phương trình trên và do tính duy nhất nghiệm, ta có các ước lượng hợp lý nhất

$$\hat{a} = \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$\hat{\sigma}^2 = S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2.$$

Ngoài ra ta có các phương pháp ước lượng khác không xét ở đây như *Bay-ét, độ lệch bé nhất...* Cuối cùng nếu hàm L trong (3.3) phức tạp, việc tìm $\hat{\theta}$ theo (3.5)–(3.6) rất khó khăn; khi đó ta phải dùng các thuật toán phi tuyến xấp xỉ dạng lặp hoặc Niu-ơn – Ráp-xơn cải biên.

§4. KHOẢNG TIN CÂY

4.1. Ước lượng khoảng

Ước lượng điểm có một nhược điểm cơ bản là không thể biết được độ chính xác cũng như xác suất để ước lượng đó chính xác. Nhất là khi kích thước mẫu nhỏ sự sai lệch của ước lượng so với giá trị thật khá lớn và chỉ với một số khó đánh giá được khả năng mắc sai lầm khi ước lượng là bao nhiêu. Để khắc phục các hạn chế đó, người ta dựa vào khái niệm ước lượng bằng một khoảng giá trị. Rõ ràng ước lượng khoảng có độ tin cậy cao hơn nhiều và cho phép xác định khách quan sai số ước lượng. Tất nhiên một khoảng ước lượng vẫn có thể sai, giống như mọi ước lượng khác, nhưng khác với ước lượng điểm, xác suất sai lầm có thể biết và trong chừng mực nào đó có thể hy vọng kiểm soát được. Nói như vậy không có nghĩa là không nên dùng ước lượng điểm nữa. Nó vẫn cho ta một thông tin quan trọng và ước lượng khoảng sẽ được xây dựng xung quanh ước lượng điểm.

Từ đó, để ước lượng một tham gia số θ , phương pháp này chủ trương xây dựng một thống kê nào đó có luật phân phối xác định không phụ thuộc θ (nhưng thống kê lại phụ thuộc). Nếu dựa vào thống kê đó ta tìm được khoảng giá trị (θ_1, θ_2) trong đó θ_1 và θ_2 phụ thuộc vào thống kê trên, sao cho với một xác suất cho trước tham số θ rơi vào khoảng đó, thì khoảng (θ_1, θ_2) sẽ được gọi là *khoảng tin cậy* với *độ tin cậy* đã cho. Như vậy nếu đặt $1 - \alpha = \gamma$ là độ tin cậy cho trước, ta cần xác định θ_1 và θ_2 sao cho

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha. \quad (4.1)$$

Độ dài $\theta_2 - \theta_1$ sẽ được gọi là *độ dài của khoảng tin cậy*.

Để làm được điều đó quy tắc chung như sau: Đầu tiên tìm một thống kê $G = G(x_1, \dots, x_n, \theta)$ sao cho phân phối của G xác định hoàn toàn (không chứa tham số θ nữa). Khi đó với độ tin cậy $1 - \alpha = \gamma$ cho trước, ta tìm cặp giá trị α_1 và α_2 sao cho

$\alpha_1 + \alpha_2 = \alpha$ (để ý tất cả chúng đều dương) và tương ứng với chúng là các phân vị g_{α_1} và $g_{1-\alpha_2}$ thoả mãn điều kiện

$$P(G < g_{\alpha_1}) = \alpha_1 \text{ và } P(G > g_{1-\alpha_2}) = \alpha_2. \quad (4.2)$$

Rõ ràng

$$P(g_{\alpha_1} < G(x_1, \dots, x_n, \theta) < g_{1-\alpha_2}) = 1 - \alpha_1 - \alpha_2 = 1 - \alpha. \quad (4.3)$$

Bằng các phép biến đổi tương đương ta đưa bất đẳng thức trong (4.3) về dạng $\theta_1 < \theta < \theta_2$ và

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha;$$

đó chính là khoảng tin cậy cần tìm. Trong thực tế người ta thường chọn độ tin cậy khá lớn $1 - \alpha = 0,95, 0,99$ hoặc $0,999$. Khả năng mắc sai lầm khi dùng các ước lượng khoảng ở đây bằng α .

4.2. Khoảng tin cậy cho kỳ vọng

Đầu tiên ta giả sử biến gốc $X \sim \mathcal{N}(a, \sigma^2)$ và tham số a chưa biết, ngoài ra ta biết được mẫu quan sát được cảm sinh từ X là $x_1, x_2, \dots, x_n$. Bài toán đặt ra là tìm khoảng tin cậy cho $EX = a$ với độ tin cậy $1 - \alpha$ cho trước.

1. Bài toán 1 (phương sai $\sigma^2 = \sigma_0^2$ đã biết)

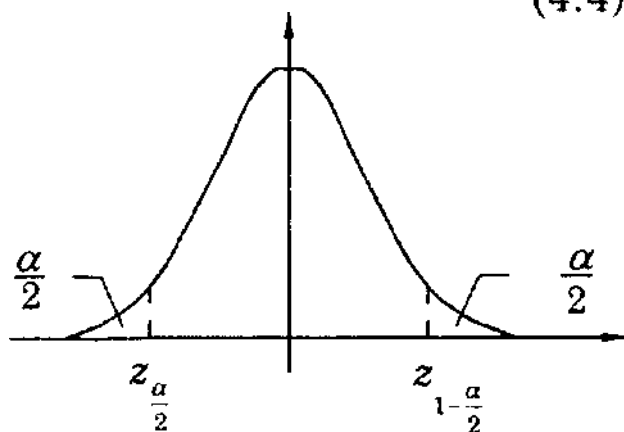
Ta chọn thống kê

$$G = Z = \frac{\bar{X} - a}{\sigma_0} \sqrt{n}. \quad (4.4)$$

Từ giả thiết chuẩn của X ta thấy $Z \sim \mathcal{N}(0, 1)$. Chọn cặp α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tìm các phân vị (xem (4.2))

$$P(Z < z_{\alpha_1}) = \alpha_1,$$

$$P(Z < z_{1-\alpha_2}) = 1 - \alpha_2.$$



Hình 4.1. Phân phối của Z

Do phân vị chuẩn có tính chất $z_{\alpha_1} = -z_{1-\alpha_1}$ nên

$$P(-z_{1-\alpha_1} < Z < z_{1-\alpha_2}) = 1 - \alpha. \quad (4.5)$$

Để ý đến (4.4) và giải hệ bất phương trình trong (4.5) đối với α , ta thu được khoảng tin cậy cần tìm

$$\bar{X} - \frac{\sigma_0}{\sqrt{n}} z_{1-\alpha_2} < a < \bar{X} + \frac{\sigma_0}{\sqrt{n}} z_{1-\alpha_1}. \quad (4.6)$$

Như vậy đối với độ tin cậy $1 - \alpha$ cho trước, ta sẽ có vô số cặp α_1, α_2 thỏa mãn $\alpha_1 + \alpha_2 = \alpha$ và tương ứng có vô số khoảng tin cậy. Ta xét một số trường hợp đặc biệt:

a) *Khoảng tin cậy đối xứng*: Nếu ta chọn $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ (xem hình 4.1); từ (4.6) nếu đặt $z_b = z_{1-\frac{\alpha}{2}}$ (tra từ bảng số 2) ta có khoảng

$$\bar{X} - \frac{\sigma_0}{\sqrt{n}} z_b < a < \bar{X} + \frac{\sigma_0}{\sqrt{n}} z_b. \quad (4.7)$$

Đại lượng $\varepsilon = \frac{\sigma_0}{\sqrt{n}} z_b$ được gọi là *độ chính xác* của ước lượng, nó phản ánh độ lệch của trung bình mẫu so với kỳ vọng lý thuyết với độ tin cậy $1 - \alpha$.

b) *Khoảng tin cậy phải*: Nếu chọn $\alpha_1 = 0, \alpha_2 = \alpha$ thì $z_{1-\alpha_1} = +\infty$ đặt $z_b = z_{1-\alpha}$ ta có khoảng cần tìm

$$\left(\bar{X} - \frac{\sigma_0}{\sqrt{n}} z_b, +\infty \right). \quad (4.8)$$

c) *Khoảng tin cậy trái*: Nếu chọn $\alpha_1 = \alpha, \alpha_2 = 0$ thì $z_{1-\alpha_2} = +\infty$ đặt $z_b = z_{1-\alpha}$ ta có khoảng cần

$$\left(-\infty, \bar{X} + \frac{\sigma_0}{\sqrt{n}} z_b \right). \quad (4.9)$$

Để ý z_b trong (4.8) và (4.9) đều là phân vị chuẩn $1 - \alpha$. Khi tra bảng hàm Láp-la-xơ lưu ý $\phi(z_b) = \phi(z_{1-\alpha}) = \frac{1}{2} - \alpha$. Trong khi đó

$$z_b \text{ của (4.7) thỏa mãn } \phi(z_b) = \phi\left(z_{1-\frac{\alpha}{2}}\right) = \frac{1-\alpha}{2}.$$

Với cùng độ tin cậy $1 - \alpha$, rõ ràng khoảng tin cậy càng ngắn càng tốt. Theo nghĩa đó khoảng (4.7) là tốt nhất, chưa kể đến sự đối xứng của nó đối với kỳ vọng mẫu. Để ý trong trường hợp này độ dài khoảng tin cậy sẽ là

$$2\varepsilon = \frac{2\sigma_0}{\sqrt{n}} z_{1-\frac{\alpha}{2}}. \quad (4.10)$$

Công thức (4.10) cho ta thấy quan hệ giữa độ tin cậy $1 - \alpha$, dung lượng mẫu n và độ chính xác ε (hay độ dài khoảng tin cậy 2ε). Nếu biết 2 trong số 3 tham số, ta hoàn toàn xác định được biến thứ ba.

Thí dụ 4.1. Một phân xưởng muốn ước lượng thời gian trung bình để sản xuất 1 ram giấy. Giả sử lượng thời gian đó tuân theo luật chuẩn với $\sigma = 0,3$ phút. Trên một tập mẫu gồm 36 ram thời gian trung bình tính được là 1,2 phút/ram. Tính khoảng tin cậy 95% cho thời gian sản xuất trung bình trên.

Giải. Thông tin đầu vào $\bar{X} = 1,2$; $\sigma_0 = 0,3$; $n = 36$ và $\alpha = 1 - 95\% = 5\%$. Ta chọn khoảng tin cậy đối xứng (4.7), trước tiên tra bảng $\phi(z_b) = \frac{1-\alpha}{2} = 0,475$ để có $z_b = 1,96$; từ đó

$$\begin{aligned} \bar{X} - \frac{\sigma_0}{\sqrt{n}} z_b < EX < \bar{X} + \frac{\sigma_0}{\sqrt{n}} z_b \\ \Rightarrow \left(1,2 - \frac{0,3}{\sqrt{36}} 1,96; 1,2 + \frac{0,3}{\sqrt{36}} 1,96 \right) &\Leftrightarrow (1,102; 1,298). \end{aligned}$$

Thí dụ 4.2. Trong thí dụ 4.1 nếu ta muốn độ chính xác của ước lượng tăng gấp đôi nhưng độ tin cậy không đổi $= 0,95$ thì cần nghiên cứu mẫu có kích thước bằng bao nhiêu?

Giải. Do ở thí dụ 4.1, độ chính xác của ước lượng bằng 0,098; nên để nó tăng gấp đôi ta cần có $\varepsilon = 0,049$. Theo (4.10) ta cần mẫu có dung lượng

$$n \geq \frac{\sigma_0^2}{\varepsilon^2} z_{0,475}^2 = \frac{(0,3)^2}{(0,049)^2} (1,96)^2 \approx 142.$$

Cuối cùng từ (4.10) ta có hai nhận xét:

– Khi kích thước mẫu tăng và độ tin cậy giữ nguyên thì ε giảm hay độ chính xác của ước lượng tăng.

– Ngược lại nếu tăng độ tin cậy và giữ nguyên kích thước mẫu, do giá trị phân vị chuẩn tăng nên ε tăng làm cho độ chính xác của ước lượng giảm đi.

2. Bài toán 2 (phương sai σ^2 chưa biết)

Trong trường hợp này đầu tiên ta phải ước lượng σ^2 bằng phương sai mẫu hiệu chỉnh, sau đó chọn thống kê

$$G = T = \frac{\bar{X} - a}{s} \sqrt{n}. \quad (4.11)$$

Theo (2.15) ta biết thống kê T tuân theo luật Stiu-đơn với $n - 1$ bậc tự do, mặt khác hình dạng của mật độ phân phối này rất gần với chuẩn, nên cách ước lượng rất giống với bài toán 1. Ta tìm phân vị

$$P(T < t_{n-1, \alpha_1}) = \alpha_1 ; P(T < t_{n-1, 1-\alpha_2}) = 1 - \alpha_2$$

và do $\alpha_1 + \alpha_2 = \alpha$ nên

$$P(T < t_{n-1, \alpha_1}) = \alpha_1 ; P(T < t_{n-1, 1-\alpha_2}) = 1 - \alpha_2. \quad (4.12)$$

So sánh (4.13) dưới đây với (4.6) ta thấy chỉ khác nhau ở hai chỗ: thay σ_0 bằng s và thay giá trị bảng Láp-la-xơ bằng bảng Stiu-đơn. Từ đó giống như trong (4.7) – (4.9):

a) *Khoảng tin cậy đối xứng*: tra bảng tính $t_b = t_{n-1, 1-\frac{\alpha}{2}}$ và

ta có

$$\left(\bar{X} - \frac{s}{\sqrt{n}} t_b; \bar{X} + \frac{s}{\sqrt{n}} t_b \right). \quad (4.13)$$

b) *Khoảng tin cậy phải*: tra bảng tìm $t_b = t_{n-1, 1-\alpha}$ và

$$\left(\bar{X} - \frac{s}{\sqrt{n}} t_b; +\infty \right).$$

c) *Khoảng tin cậy trái*: với cùng giá trị bảng ở phần b)

$$\left(-\infty, \bar{X} + \frac{s}{\sqrt{n}} t_b \right).$$

Thí dụ 4.3. Một lò bánh muốn ước lượng trọng lượng trung bình của số bột dùng hàng ngày (giả sử lượng bột tuân theo luật chuẩn). Với kết quả thống kê của 14 ngày ta có ước lượng điểm của α là 17,3kg với $s = 4,5$ kg. Xây dựng khoảng tin cậy 99% cho trọng lượng trung bình α .

Giải. Số liệu đầu vào $\bar{X} = 17,3$; $s = 4,5$; $n = 14$ và $1 - \alpha = 99\%$. Ta tra bảng Stiu-đơn $t_b = t_{13; 0,995} = 3,012$. Từ đó khoảng tin cậy 99% sẽ là

$$\left(\bar{X} - \frac{4,5}{14} \cdot 3,012, \bar{X} + \frac{4,5}{14} \cdot 3,012 \right) = (136,77; 209,23).$$

Thí dụ 4.4. Ta muốn đánh giá nhiệt độ lớn nhất trung bình ở tỉnh Lâm Đồng vào ngày 25 tháng 9 (giả sử nhiệt độ đó tuân theo luật chuẩn). Nhiệt độ cao nhất ở 5 vùng của tỉnh đo được trong ngày hôm đó là 25, 27, 29, 32 và 33°C. Hãy xác định khoảng tin cậy 95% cho nhiệt độ cao nhất trung bình trong ngày đang xét.

Giải: Gọi X là nhiệt độ cao nhất ở Lâm Đồng vào ngày 25/9, ta đã có $X \sim \mathcal{N}(\alpha, \sigma^2)$. Do chưa có các đặc trưng mẫu nên ta cần tính

$$\bar{X} = \frac{146}{5} = 29,2; s = \sqrt{\frac{44,8}{4}} = 3,35$$

x	$x - \bar{X}$	$(x - \bar{X})^2$
25	-4,2	17,64
27	-2,2	8,84
29	-0,2	0,04
32	2,8	7,84
33	3,8	14,44
146		44,8

Tra bảng Stiu-đơn $t_b = t_{4,0,975} = 2,776$, ta có

$$\left(\bar{X} - \frac{s}{\sqrt{n}} t_b, \bar{X} + \frac{s}{\sqrt{n}} t_b \right) = \left(29,2 - \frac{33,5}{\sqrt{5}} \cdot 2,776; 29,2 + \frac{33,5}{\sqrt{5}} \cdot 2,776 \right) \\ = (25,04; 33,36).$$

Để ý đây là khoảng tin cậy 95% tính trên bộ số liệu cụ thể của thí dụ, nó hoàn toàn không có nghĩa là xác suất để trung bình thật rơi vào khoảng tin cậy trên là 0,95. Bởi vậy không nên quên rằng độ tin cậy 95% của một khoảng nào đó được hiểu theo nghĩa thống kê (tức là nếu cứ làm thí nghiệm 100 lần với các khoảng tin cậy 95% thì có khoảng 95 lần giá trị trung bình thật nằm trong khoảng đó).

Nếu dung lượng mẫu $n > 30$, thống kê T trong (4.11) sẽ có phân phối tiệm cận chuẩn $\mathcal{N}(0, 1)$, và việc tìm khoảng ước lượng với độ tin cậy $1 - \alpha$ được làm giống như bài toán 1, với σ_0 được thay bằng độ lệch chuẩn mẫu hiệu chỉnh s . Lưu ý là trong các bài toán và thí dụ ở đây, ta luôn luôn có giả thuyết chuẩn của phân phối gốc.

4.3. Khoảng tin cậy cho tỷ lệ

Nếu biến ngẫu nhiên gốc không tuân theo luật phân phối chuẩn, việc xác định khoảng tin cậy cho EX sẽ rất phức tạp và đòi hỏi các kỹ thuật hiện đại hơn. Tuy nhiên trong trường hợp n đủ lớn, cả hai thống kê Z trong (4.4) và T trong (4.11) đều có phân phối xấp xỉ chuẩn $\mathcal{N}(0, 1)$. Do đó các thủ tục ước lượng khoảng làm giống như bài toán 1 đã nói đến ở mục trên.

Ta xét một trường hợp cụ thể khi dấu hiệu $X \sim \mathcal{B}(1, p)$ (phân phối Béc-nu-li). Khi đó nếu ta chọn ra phần tử từ tập nền (theo dạng mẫu ngẫu nhiên) thì số lần xuất hiện dấu hiệu quan tâm X_i cùng phân phối với X . Như vậy $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ chính là tần suất ước lượng điểm của xác suất hay tỉ lệ $p = EX$. Mặt khác từ kết quả chương II, $n\bar{X}$ sẽ có phân phối nhị thức

$\mathcal{B}(n, p)$, từ đó $E\bar{X} = p$ và $V\bar{X} = \frac{p(1-p)}{n}$. Nếu ta chọn thống kê

(với $f = \frac{m}{n}$ là tần suất mẫu xuất hiện dấu hiệu quan tâm)

$$Z = \frac{f - p}{\sqrt{p(1-p)}} \sqrt{n} \quad (4.14)$$

thì khi n khá lớn $Z \xrightarrow{L} \mathcal{N}(0, 1)$.

Bài toán 3 (tìm thấy khoảng tin cậy $1 - \alpha$ cho tỷ lệ (xác suất))

Dựa vào (4.14) ta có hai cách đi tìm khoảng tin cậy khi n đủ lớn.

1) Theo cách làm ở trên chọn $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ (xem bài toán 1)

và từ (4.5) – (4.6) ta có

$$-z_b \leq \frac{f - p}{\sqrt{p(1-p)}} \sqrt{n} < z_b$$

với $z_b = z_{1-\frac{\alpha}{2}}$ (khoảng tin cậy đối xứng). Giải hệ bất phương

trình trên đối với p

$$\begin{aligned} n(f - p)^2 &< p(1 - p)z_b^2 \\ \Leftrightarrow (n + z_b^2)p^2 - (2nf + z_b^2)p + nf^2 &< 0. \end{aligned}$$

(Giải và tìm nghiệm phương trình bậc 2 ở vế trái, ta có 2 nghiệm

$$p_1, p_2 = \frac{nf + \frac{1}{2}z_b^2 \pm z_b \sqrt{nf(1-f) + \frac{1}{4}z_b^2}}{n + z_b^2} \quad (4.15)$$

và khoảng tin cậy cần tìm sẽ là (p_1, p_2) , với $p_1 < p_2$. Tuy nhiên việc tính toán theo (4.15) sẽ khá khó khăn.

2. Ta tìm ước lượng khoảng gần đúng theo cách khác. Để ý nếu n khá lớn, thống kê

$$Z = \frac{\hat{f} - p}{\sqrt{\hat{f}(1 - \hat{f})}} \sqrt{n} \xrightarrow{L} \mathcal{N}(0, 1)$$

với $VX = p(1 - p)$ được thay bằng ước lượng điểm $\hat{f}(1 - \hat{f})$. Bây giờ quy trình giải bài toán 1 đã có thể được áp dụng ($\bar{X}$ thay bằng $\hat{f}$, σ_0^2 thay bằng $\hat{f}(1 - \hat{f})$, ...)

$$\hat{f} - \sqrt{\frac{\hat{f}(1 - \hat{f})}{n}} z_{1-\alpha_2} < p < \hat{f} + \sqrt{\frac{\hat{f}(1 - \hat{f})}{n}} z_{1-\alpha_1}.$$

Từ đó (xem (4.7) – (4.9))

a) *Khoảng tin cậy đối xứng*: $z_b = z_{1-\frac{\alpha}{2}}$

$$\left(\hat{f} - z_b \sqrt{\frac{\hat{f}(1 - \hat{f})}{n}}; \hat{f} + z_b \sqrt{\frac{\hat{f}(1 - \hat{f})}{n}} \right). \quad (4.16)$$

b) *Khoảng tin cậy phải*: $z_b = z_{1-\alpha}$ và

$$\left(\hat{f} - z_b \sqrt{\frac{\hat{f}(1 - \hat{f})}{n}}; +\infty \right). \quad (4.17a)$$

c) *Khoảng tin cậy trái*: với z_b như trên

$$\left(-\infty; \hat{f} + z_b \sqrt{\frac{\hat{f}(1 - \hat{f})}{n}} \right). \quad (4.17b)$$

Cuối cùng, nếu ký hiệu ε là độ chính xác của ước lượng khoảng đối xứng, ta có quan hệ (xem (4.10)):

$$\varepsilon^2 = \frac{\hat{f}(1 - \hat{f})}{n} z_{1-\frac{\alpha}{2}}^2.$$

Thí dụ 4.5. Kiểm tra ngẫu nhiên 600 sản phẩm của một máy dập thấy có 24 phế phẩm. Với độ tin cậy $1 - \alpha = 95\%$ hãy ước lượng tỷ lệ phế phẩm tối đa của máy đó.

Giải. Gọi p là xác suất ra phế phẩm của máy trên hay p là xác suất xuất hiện dấu hiệu phế phẩm của sản phẩm nào đó và ta có thể dùng quy trình bài toán 3. Ở đây $n = 600$ (khá lớn), tỷ lệ phế phẩm mẫu $f = 24/600 = 0,04$. Ta sẽ dùng khoảng tin cậy trái (xem 4.17b); trước tiên tìm phân vị chuẩn $z_b = z_{1-\alpha}$ (nếu tra bảng Láp-la-xơ $\phi(z_b) = \frac{1}{2} - \alpha = 0,45$) ta tìm được $z_b = 1,64$; từ đó

$$\left(-\infty; f + \sqrt{\frac{f(1-f)}{n}} z_b \right) = \left(-\infty; 0,04 + \sqrt{\frac{0,04 \cdot 0,96}{600}} \cdot 1,64 \right)$$

hay tỷ lệ phế phẩm tối đa là $0,05312 = 5,312\%$.

Thí dụ 4.6. Phỏng vấn 400 người ở một khu vực 300000 người thấy có 240 người ủng hộ dự luật A. Với độ tin cậy 0,95 hãy ước lượng số người ủng hộ dự luật A trong khu vực bằng khoảng tin cậy đối xứng.

Giải. Gọi p là tỷ lệ người ủng hộ dự luật A và ta sẽ sử dụng kết quả (4.16). Theo đầu bài $f = 240/400 = 0,6$; $\alpha = 0,05$; phân vị chuẩn $z_{1-\frac{\alpha}{2}} = z_{0,975} = 1,96$ ($\phi(1,96) = 0,475$), vậy khoảng tin cậy cần tìm của p là

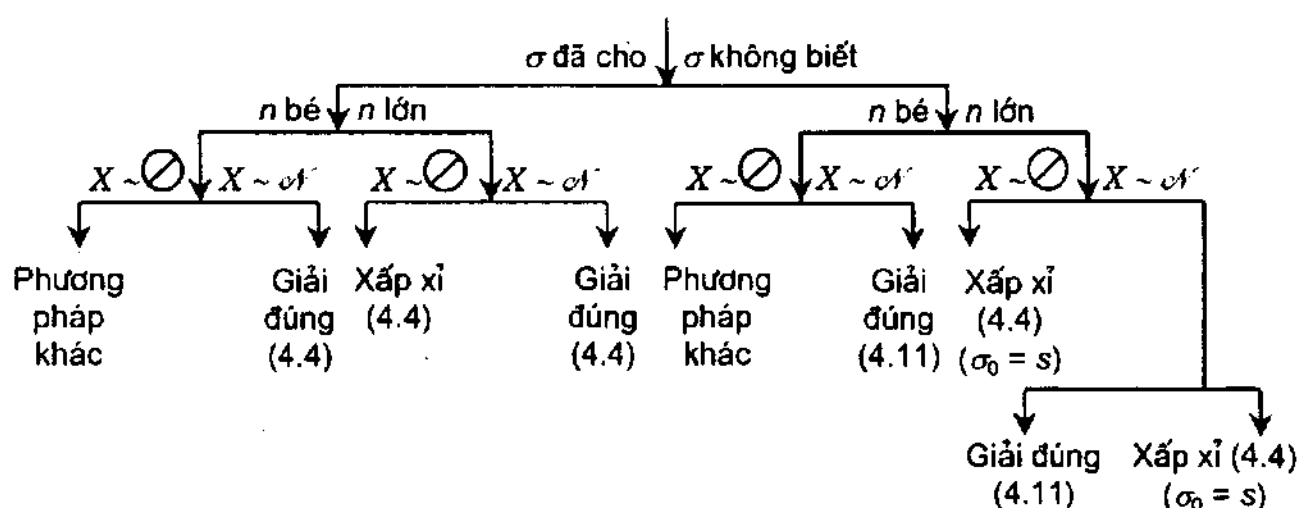
$$\begin{aligned} & \left(0,6 - \sqrt{\frac{0,6 \cdot 0,4}{400}} \cdot 1,96; 0,6 + \sqrt{\frac{0,6 \cdot 0,4}{400}} \cdot 1,96 \right) \\ & \Rightarrow 0,5522 < p < 0,6478. \end{aligned}$$

Do đó khoảng tin cậy của số người ủng hộ dự luật A ở vùng đó là

$$(300000 \cdot 0,5522; 300000 \cdot 0,6478) = (165660; 194340).$$

Nếu ta sử dụng công thức (4.15) khoảng tin cậy sẽ là $(0,5513; 0,6468)$ đối với p và $(165390; 194040)$ đối với số người ủng hộ dự luật A.

Ta có thể tóm tắt các kết quả của mục 4.2 và 4.3 như sau:



4.4. Khoảng tin cậy cho phương sai

Bài toán 4. Giả sử $X \sim \mathcal{N}(a, \sigma^2)$ và độ tin cậy $1 - \alpha$ đã cho. Ta cần xác định khoảng tin cậy tương ứng cho $VX = \sigma^2$ dựa trên mẫu $x_1, x_2, \dots, x_n$ được cảm sinh bởi biến gốc X .

Quy trình xây dựng khoảng tin cậy dựa trên sự kiện

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1); \quad (4.18)$$

$$\chi^2 = \frac{\sum_{i=1}^n (x_i - a)^2}{\sigma^2} \sim \chi^2(n). \quad (4.19)$$

Từ đó ta làm giống như trong mục 4.2, chia bài toán thành hai trường hợp:

1) Nếu $a = a_0$ đã biết, sử dụng thống kê (4.19) và chọn α_1, α_2 , sao cho $\alpha_1 + \alpha_2 = \alpha$, sau đó tìm các phân vị

$$P(\chi^2 < \chi_{n, \alpha_1}^2) = \alpha_1; \quad P(\chi^2 < \chi_{n, 1-\alpha_2}^2) = 1 - \alpha_2.$$

Từ đó suy ra $P(\chi_{n, \alpha_1}^2 < \chi^2 < \chi_{n, 1-\alpha_2}^2) = 1 - \alpha$

hay khoảng tin cậy $1 - \alpha$ của σ^2 là

$$\left(\frac{\sum_{i=1}^n (x_i - a_0)^2}{\chi_{n, 1-\alpha_2}^2}; \frac{\sum_{i=1}^n (x_i - a_0)^2}{\chi_{n, \alpha_1}^2} \right). \quad (4.20)$$

Ta xét một số trường hợp cụ thể của (4.20):

a) Nếu $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, khoảng (4.20) trở thành

$$\left(\frac{\sum_{i=1}^n (x_i - a_0)^2}{\chi_{n, 1-\frac{\alpha}{2}}^2}; \frac{\sum_{i=1}^n (x_i - a_0)^2}{\chi_{n, \frac{\alpha}{2}}^2} \right) \quad (4.21)$$

b) Nếu $\alpha_1 = 0$, $\alpha_2 = \alpha$, ta có khoảng tin cậy phải

$$\left(\frac{\sum_{i=1}^n (x_i - a_0)^2}{\chi_{n, 1-\alpha}^2}; +\infty \right). \quad (4.21a)$$

c) Nếu $\alpha_1 = \alpha$, $\alpha_2 = 0$, ta có khoảng tin cậy trái

$$\left(-\infty; \frac{\sum_{i=1}^n (x_i - a_0)^2}{\chi_{n, \alpha}^2} \right). \quad (4.21b)$$

2) Nếu a chưa biết, ta thay nó bằng ước lượng $\bar{X}$ và sử dụng thống kê (4.18). Cách làm giống như ở trên và bạn đọc tự tìm ra kết quả. Chẳng hạn trường hợp chọn $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ thì (4.21) trở thành

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1, 1-\frac{\alpha}{2}}}; \frac{(n-1)s^2}{\chi^2_{n-1, \frac{\alpha}{2}}} \right). \quad (4.22)$$

Thí dụ 4.7. Cho khối lượng một loại sản phẩm tuân theo luật phân phối chuẩn. Cân thử từng sản phẩm của một mẫu ngẫu nhiên gồm 25 đơn vị, ta có kết quả

Khối lượng	29,3	29,7	30	30,5	30,7
Số sản phẩm	4	5	8	5	3

Với độ tin cậy 95% hãy tìm khoảng tin cậy cho phương sai của khối lượng sản phẩm trong 2 trường hợp: a) biết kỳ vọng $\alpha = 30$; b) không biết kỳ vọng.

Giải. Đầu tiên xác định các đặc trưng mẫu

$$\begin{aligned} \bar{X} &= \frac{1}{25} \sum_{i=1}^5 x_i n_i = \frac{1}{25} (29,3 \cdot 4 + 29,7 \cdot 5 + 30 \cdot 8 + 30,5 \cdot 5 + 30,7 \cdot 3) \\ &= 30,012; \end{aligned}$$

$$\sum_{i=1}^5 (x_i - 30)^2 n_i = 5,13;$$

$$24s^2 = \sum_{i=1}^5 (x_i - 30,012)^2 n_i = 5,1264.$$

a) Ta dùng (4.21), với $\alpha_0 = 30$, và khoảng tin cậy cần tìm là

$$\left(\frac{5,13}{\chi^2_{25; 0,975}}; \frac{5,13}{\chi^2_{25; 0,025}} \right) = \left(\frac{5,13}{40,65}; \frac{5,13}{13,12} \right) = (0,1262; 0,3910).$$

b) Dùng (4.22), khoảng tin cậy cần tìm

$$\left(\frac{5,1264}{\chi^2_{24; 0,975}}; \frac{5,1264}{\chi^2_{24; 0,025}} \right) = \left(\frac{5,1264}{39,36}; \frac{5,1254}{12,40} \right) = (0,1302; 0,4134).$$

BÀI TẬP

1. Điểm thi tiếng Anh của một lớp học sinh như sau:

68	52	49	69	56	41
74	79	59	81	57	42
88	87	60	65	55	47
65	68	50	78	61	90
85	66	65	72	63	95

- a) Xác định bảng phân phối thực nghiệm và đa giác tần suất tương ứng.
- b) Tính các số đặc trưng mẫu: trung bình, mốt, trung vị, phương sai.
2. Thống kê số km đã chạy của 100 xe tải của một hãng trong năm 1997:

Số km	Số xe tải
10000 – 14000	5
14000 – 18000	10
18000 – 22000	11
22000 – 26000	20
26000 – 30000	25
30000 – 34000	13
34000 – 38000	12
38000 – 42000	4

- a) Dựng biểu đồ và đa giác tần suất.
- b) Tính các số đặc trưng mẫu: trung bình, mốt và trung vị. Cho biết ý nghĩa của chúng.
- c) Phân phối mẫu không đối xứng, hãy giải thích tại sao?

3. Thu nhập gia đình năm của hai nhóm dân ở hai làng một huyện nhỏ là:

Thu nhập năm (đồng)	Số gia đình	
	nhóm A	nhóm B
1250000 – 1300000	4	2
1300000 – 1350000	40	20
1350000 – 1400000	73	32
1400000 – 1450000	52	58
1450000 – 1500000	23	34
1500000 – 1550000	8	31
1550000 – 1600000	–	15
1600000 trở lên	–	8

- a) Tính thu nhập trung bình năm của hai nhóm gia đình trên.
b) Tính một của thu nhập của hai nhóm gia đình.
c) So sánh và phân tích tình trạng thu nhập của nhóm A và B.
4. Số km phải đi của 6 học sinh một lớp buổi tối như sau:

Học sinh	1	2	3	4	5	6
Số km	1	4	9	8	6	5

- a) Tính khoảng cách trung bình.
b) Tính các số đặc trưng mẫu của khoảng cách: một và trung vị.
5. Giá của một loại bất động sản ở một vùng dân cư là

180	165	151	148	145	121	110	110	105	100
100	100	100	98	95	95	90	90	90	85
84	83	82	80	80	75	72	72	68	65
61	61	60	60	60	58	57	56	55	55
54	54	53	52	51	50	50	50	50	50
50	49	46	45	45	41	41	40	40	38
38	36	35	35						

a) Xác định bảng phân phối thực nghiệm (lấy độ dài khoảng bằng 30 và bắt đầu từ giá trị 30).

b) Xây dựng biểu đồ và đa giác tần suất.

c) Tính các số đặc trưng mẫu theo bảng phân phối ở mục a, sau đó so sánh với kết quả tính trực tiếp (trung bình, phương sai, môđ và trung vị).

6. Số lượng khách hàng đến mua ở một cửa hàng trong vòng 1 ngày được thống kê như sau:

Số khách	Số ngày
95 – 105	4
105 – 115	14
115 – 125	18
125 – 135	15
135 – 145	10
145 – 155	7
155 – 165	2

Hãy xác định các số đặc trưng mẫu (trung bình, phương sai, độ lệch chuẩn, môđ, trung vị, mômen cấp 1 và 2) của số khách.

7. Điều tra 1600 gia đình có 4 con ta có kết quả

Số con trai	0	1	2	3	4
Số gia đình	111	367	576	428	118

Xác định kỳ vọng mẫu và phương sai mẫu hiệu chỉnh.

8. Cho 8 số liệu đo cùng một đại lượng thiết bị đo không có sai số hệ thống

369 378 365 420 385 401 372 383

Hãy tính ước lượng không chệch của phương sai sai số đo trong hai trường hợp: a) biết số đo thật bằng 375; b) không biết số đo thật.

9. Theo dõi thời gian hoàn thành một sản phẩm của hai nhóm công nhân

– Nhóm 1	Thời gian	43	44	50	55	60	65
	Số người	2	5	15	20	5	3
– Nhóm 2	Thời gian	45	49	53	60		
	Số người	2	41	5	1		

Tính trung bình mẫu và phương sai mẫu hiệu chỉnh của thời gian hoàn thành sản phẩm của từng nhóm và bình luận kết quả.

10. Một lô hàng gồm n linh kiện (n rất lớn). Người ta chọn ngẫu nhiên ra m linh kiện, đánh dấu chúng rồi trả lại lô hàng. Sau khi trộn kỹ, chọn ngẫu nhiên ra k linh kiện thì thấy có l chiếc bị đánh dấu (k khá bé so với n). Hãy xác định ước lượng hợp lý nhất cho số lượng n .
11. Ở một công trường xây dựng lớn lương trung bình của một công nhân là 600000 đồng với độ lệch chuẩn là 50000 đồng. Tính xác suất để lương trung bình của một nhóm 50 công nhân chọn ngẫu nhiên nằm trong khoảng từ 610000 đến 650000 đồng.
12. Biết tỷ lệ phế phẩm của một lô hàng là 5%. Tìm xác suất để khi chọn ra 400 sản phẩm từ lô hàng trên (với số lượng rất lớn) thì có trên 9% phế phẩm.
13. Trong 3500 sinh viên năm thứ nhất của trường đại học Bách khoa có 28% muốn học ngành điện tử – viễn thông. Chọn ngẫu nhiên ra một nhóm sinh viên 350 người (của năm thứ nhất đó). Tính tỷ lệ trung bình của số sinh viên muốn học ngành điện tử – viễn thông trong nhóm sinh viên trên.
14. Nhiệt độ của 24 thành phố Việt Nam ở cùng một giờ và một ngày trong tháng 7 như sau:

36	30	31	32	31	40	37	29
41	37	35	34	34	35	32	33
35	33	33	31	34	34	35	32

Xây dựng khoảng tin cậy 99% cho nhiệt độ trung bình trên.

15. Người ta muốn ước lượng số lần gọi trung bình của một tổng đài điện thoại trong vòng 1 ngày. Thống kê trong vòng 50 ngày cho số lần gọi trung bình là 525 với $s = 52$. Hãy xác định khoảng tin cậy 90% cho số lần gọi trung bình đó.
16. Chiều cao trung bình của một nhóm học sinh gồm 20 em là 1,65 m với độ lệch chuẩn mẫu là 0,2 m. Xây dựng khoảng tin cậy 95% cho chiều cao trung bình của toàn bộ học sinh.
17. Chọn ngẫu nhiên 50 sinh viên ở một trường đại học thì thấy có 21 nữ. Hãy ước lượng tỷ lệ nữ ở trường đại học đó với độ tin cậy 90%.
18. Một thiết bị đo có hai dung sai là 0,2 cm. Thống kê 25 lần đo các chi tiết cùng loại ta có độ dài trung bình là 15,2 cm. Hãy ước lượng độ dài trung bình của loại chi tiết trên với độ tin cậy 99% (giả sử sai số đo không có tính hệ thống).
19. Kiểm tra ngẫu nhiên 500 sản phẩm của một nhà máy thì thấy có 240 sản phẩm loại A. Hãy ước lượng tỷ lệ sản phẩm loại A tối thiểu của nhà máy với độ tin cậy 95%.
20. Theo dõi 100 sinh viên để xác định số giờ tự học (X), kết quả như sau: $\bar{X} = 4,01$ với $s = 3,51$. Hãy tìm khoảng tin cậy 95% cho số giờ tự học trung bình của sinh viên. Thử ước lượng tỷ lệ sinh viên không tự học.
21. Trên một mẫu gồm 26 số liệu người ta tính được độ dài trung bình $\bar{X} = 30,2$, với $s^2 = 6,25$. Tìm khoảng tin cậy 95% cho phương sai.
22. Để ước lượng xác suất mắc bệnh A với độ tin cậy 95% và sai số không vượt quá 2% thì cần khám bao nhiêu người, biết rằng tỷ lệ mắc bệnh A thực nghiệm đã cho bằng 0,8.

Chương V

KIỂM ĐỊNH GIẢ THUYẾT

§1. GIẢ THUYẾT THỐNG KÊ VÀ QUY TẮC KIỂM ĐỊNH

1.1. Giả thuyết thống kê

Trong nhiều lĩnh vực đời sống kinh tế – xã hội chúng ta hay nêu ra các nhận xét khác nhau về các đối tượng quan tâm. Những nhận xét như vậy thường được coi là các *giả thuyết*, chúng có thể đúng và cũng có thể sai. Vấn đề xác định đúng sai của một giả thuyết sẽ được gọi là *kiểm định*.

Trong thống kê chúng ta xuất phát từ một mẫu $x_1, x_2, \dots, x_n$ chọn từ một tập nền chưa biết phân phối hoặc có phân phối $F(x, \theta)$ nhưng chưa biết tham số θ . Ta có thể phát biểu nhiều nhận xét khác nhau về các yếu tố chưa biết – đó là các giả thuyết thống kê (thí dụ phân phối tập nền có dạng chuẩn, tham số kỳ vọng bằng một số cho trước...). Nếu tham số θ chưa biết và giả thuyết θ bằng giá trị cụ thể θ_0 được đưa ra, ta nói rằng có một *giả thuyết đơn*; nếu khác đi, ta có *giả thuyết phức*. Việc kiểm định một giả thuyết đơn thường dễ dàng hơn.

Giả thuyết được đưa ra kiểm định được gọi là *giả thuyết gốc*, ký hiệu là H_0 ; nó thường là giả thuyết đơn trong các bài toán kiểm định tham số. Các giả thuyết khác với gốc được gọi là *giả thuyết đối* hay *đối thuyết* (có thể đơn hoặc phức), ký hiệu là H_1 . Ta thừa nhận khi đã chọn cặp H_1, H_0 thì việc chấp nhận H_0 sẽ chính là bác bỏ H_1 và ngược lại. Việc kiểm định một giả thuyết là đúng hay sai dựa trên thông tin mẫu sẽ được gọi là *kiểm định thống kê*.

Chẳng hạn khi nghiên cứu thu nhập của cư dân một thành phố nào đó, ta có thể đưa ra nhiều giả thuyết khác nhau:

– Thu nhập của cư dân tuân theo luật phân phối chuẩn (H_0) hoặc không tuân theo luật đó (H_1).

– Thu nhập trung bình năm là 50 triệu đồng (H_0) với nhiều dạng đối thuyết khác nhau: thí dụ $\neq 50$ triệu, > 50 triệu hoặc < 50 triệu đồng...

1.2. Quy tắc kiểm định giả thuyết

Nguyên tắc chung của kiểm định giả thuyết thống kê là dựa trên *nguyên lý xác suất nhỏ*: một sự kiện có xác suất xuất hiện khá bé thì có thể coi rằng nó không xảy ra khi thực hiện một phép thử có liên quan đến sự kiện đó. Tuy nhiên trong thực tế, vấn đề phức tạp và tế nhị hơn nhiều.

1. Tiêu chuẩn kiểm định

Tiêu chuẩn được xây dựng rõ ràng phải đơn giản và dựa trên các thông tin mẫu $x_1, x_2, \dots, x_n$. Thông thường người ta chọn một thống kê

$$K = K(x_1, x_2, \dots, x_n) \quad (1.1)$$

có thể phụ thuộc vào tham số đã biết trong giả thuyết H_0 . Nếu giả thuyết H_0 đúng thì luật phân phối của K phải hoàn toàn xác định. Một thống kê như vậy được gọi là *tiêu chuẩn kiểm định*.

2. Quy tắc kiểm định

Nếu ta thành công trong việc chia miền xác định của tiêu chuẩn (1.1) thành hai phần B_α và $\overline{B_\alpha}$ trong đó B_α là *miền bác bỏ* H_0 , còn $\overline{B_\alpha}$ là *miền chấp nhận* H_0 , thì quy tắc kiểm định khá đơn giản: Nếu K tính trên mẫu có giá trị thuộc miền B_α ta bác bỏ H_0 ; nếu ngược lại ta chấp nhận H_0 . Miền bác bỏ H_0 được gọi là *miền tới hạn* của tiêu chuẩn K .

Như vậy, nếu ta dùng quy tắc như trên, có thể mắc hai loại sai lầm sau đây:

- Sai lầm loại 1: bác bỏ một giả thuyết đúng;
- Sai lầm loại 2: chấp nhận một giả thuyết sai.

Do giả thiết K có phân phối xác định khi H_0 đúng và nếu gọi α là xác suất để xảy ra sai lầm loại 1 thì

$$\alpha = P(K_{tn} \in B_\alpha | H_0 \text{ đúng}), \quad (1.2)$$

trong đó K_{tn} chính là giá trị của K trên mẫu cụ thể đang xét. Tương tự nếu gọi β là xác suất phạm sai lầm loại 2, thì

$$\beta = P(K_{tn} \in \bar{B}_\alpha | H_0 \text{ sai}). \quad (1.3)$$

Người ta hay gọi xác suất bác bỏ giả thuyết sai $1 - \beta$ là *lực lượng* của tiêu chuẩn K .

Tất nhiên chúng ta mong muốn cả hai xác suất (1.2) và (1.3) càng bé càng tốt. Trong thực tế ta không thể đồng thời làm giảm cả hai xác suất đó, bởi vì cứ α giảm thì β tăng và ngược lại. Thông thường do sai lầm loại 1 dễ kiểm soát và (1.2) dễ tính hơn nên người ta hay lựa chọn trước α như là một ngưỡng để xác suất phạm sai lầm loại 1 luôn nhỏ hơn α đủ bé đó. Các giá trị của α có thể là 0,1; 0,05; 0,01; 0,001... phụ thuộc vào yêu cầu của thực tế và nhà nghiên cứu; giá trị α được gọi là *mức ý nghĩa* của quy tắc kiểm định (hay của tiêu chuẩn kiểm định tương ứng). Quy tắc với mức ý nghĩa α được gọi là *mạnh nhất* nếu nó có lực lượng lớn nhất.

Thí dụ 1.1. Theo Nây-man – Piéc-xơn: “Nếu $\mathcal{X} = (x_1, \dots, x_n)$ có phân phối $f_0(.)$ dưới giả thuyết $H_0: \theta = \theta_0$ và phân phối $f_1(.)$ dưới đối thuyết $H_1: \theta = \theta_1 \neq \theta_0$, ngoài ra cho $0 < \alpha < 1$ sao cho tồn tại một số k_α để $P(f_1(.) > k_\alpha f_0(.) | H_0) = \alpha$; thì $B_\alpha = \{\mathcal{X} : f_0(\mathcal{X}) > k_\alpha f_1(\mathcal{X})\}$ sẽ là miền bác bỏ H_0 của tiêu chuẩn mạnh nhất cho bài toán kiểm định giả thuyết đơn H_0 đối với thuyết đơn H_1 ”. Áp dụng vào bài toán kiểm định $H_0: \theta = 3$, đối thuyết $H_1: \theta = 4$; có $n = 9$, $\bar{X} = 2,35$ và phân phối nền $\mathcal{N}(0, 25)$ ($\alpha = 0,05$).

Giải. Do tập nên có phân phối chuẩn $\mathcal{N}(\theta, \sigma^2)$ nên

$$f_j(\mathcal{X}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta_j)^2}; j = 0; 1.$$

Từ biểu thức xác định miền tới hạn B_α trong bổ đề và công thức trên suy ra

$$(\theta_1 - \theta_0) \sum_{i=1}^n x_i - \frac{n}{2}(\theta_1^2 - \theta_0^2) \geq \sigma^2 \ln k_\alpha$$

và

$$B_\alpha = \left\{ \mathcal{X} = (x_1, \dots, x_n) : \bar{X} > \frac{2\sigma^2 \ln k_\alpha + n(\theta_1^2 - \theta_0^2)}{2n(\theta_1 - \theta_0)} = A_0 \right\}, \quad (1.4)$$

trong đó A_0 xác định từ $P(\bar{X} > A_0 | H_0) = \alpha$, $\bar{X} = \frac{1}{n}(x_1 + \dots + x_n)$.

Trong bài toán áp dụng $\bar{X} = 2,35$ còn A_0 được xác định dựa vào giả thuyết chuẩn

$$P(\bar{X} > A_0 | H_0) = \left(\frac{\bar{X} - 3}{5} \sqrt{9} > \frac{A_0 - 3}{5} \sqrt{9} \right) = 0,05.$$

Từ bảng Láp-la-xơ ta có $\frac{A_0 - 3}{5} \cdot 3 = 1,645$, từ đó $A_0 = 5,74$. Rõ ràng $\bar{X} = 2,35 < A_0$ và $\mathcal{X} \in \bar{B}_\alpha$ nên không có cơ sở để bác bỏ giả thuyết $H_0: \theta = 3$.

Chú ý là từ thí dụ trên ta thấy khi xác định được tiêu chuẩn kiểm định, về mặt nguyên tắc có thể tính được các xác suất (1.2) và (1.3) nếu biết được ngưỡng phân chia miền bác bỏ H_0 với miền chấp nhận nó.

Thí dụ 1.2. Tìm các xác suất phạm sai lầm loại 1 và loại 2 trong thí dụ 1.1, nếu chọn ngưỡng $A_0 = 5,5$.

Giải. Từ công thức (1.4), do $\bar{X} \sim \mathcal{N}(3; 25)$ khi H_0 đúng

$$\alpha = P(\bar{X} > 5,5 | H_0) = P\left(\frac{\bar{X} - 3}{5} \cdot 3 > \frac{5,5 - 3}{5} \cdot 3\right)$$

và do $Z = \frac{\bar{X} - 3}{5} \cdot 3 \sim \mathcal{N}(0, 1)$ nên dễ thấy $\alpha = P(Z > 1,5) = 0,5 - \Phi(1,5) = 0,5 - 0,4332 = 0,0668$. Còn xác suất phạm sai lầm loại hai

$$\beta = P(\bar{X} \leq 5,5 | H_1),$$

do khi H_1 đúng $Z = \frac{\bar{X} - 4}{5} \cdot 3 \sim \mathcal{N}(0, 1)$ nên $\beta = P(Z \leq 0,9) = 0,5 + \Phi(0,9) = 0,5 + 0,3159 = 0,8159$ khá lớn.

1.3. Các dạng miền tới hạn

Trong thực tế, người ta chọn miền tới hạn của tiêu chuẩn K phụ thuộc vào cặp giả thuyết $H_0; H_1$ như sau:

1) Nếu H_1 là đối lập ($H_0 = \theta = \theta_0$ với $H_1: \theta \neq \theta_0$), ta chọn các phân vị $K_{\alpha/2}$ và $K_{1-\alpha/2}$ (xem hình 5.1) sao cho (dựa vào phân phối của K khi H_0 đúng)

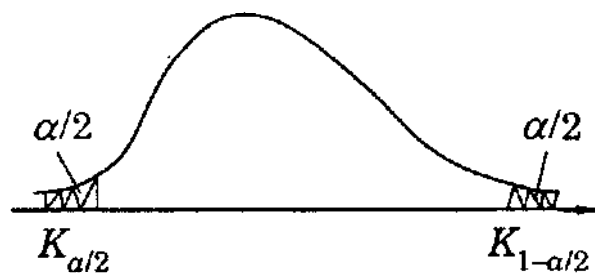
$$P(K > K_{1-\alpha/2} | H_0) = \frac{\alpha}{2};$$

$$P(K < K_{\alpha/2} | H_0) = \frac{\alpha}{2}$$

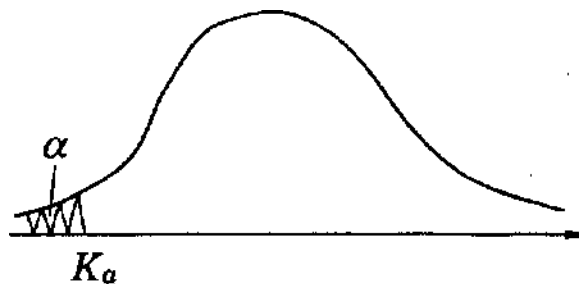
Khi đó miền tới hạn

$$B_\alpha = (-\infty, K_{\alpha/2}) \cup (K_{1-\alpha/2}, +\infty).$$

2) Nếu H_1 bất đối xứng lệch về trái (thí dụ $H_1: \theta < \theta_0$), ta chọn miền tới hạn lệch về bên trái (hình 5.2). Dựa vào phân phối của K khi H_0 đúng, ta xác định phân vị K_α sao cho



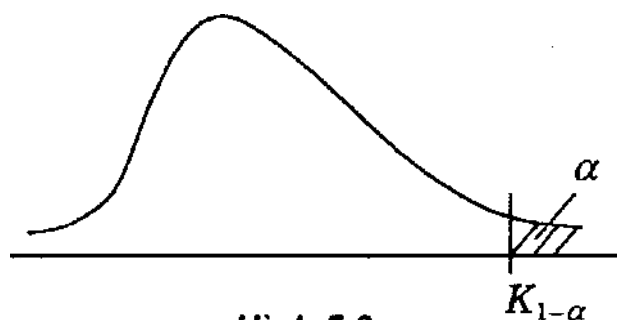
Hình 5.1



Hình 5.2

$P(K < K_\alpha | H_0 = \alpha)$ và miền tới hạn B_α là $(-\infty, K_\alpha)$.

3) Tương tự nếu H_1 bất đối xứng lệch về phải (thí dụ $H_1: \theta > \theta_0$), ta chọn B_α lệch về bên phải (hình 5.3). Dựa vào phân phối của K khi H_0 đúng, ta xác định $K_{1-\alpha}$ sao cho $P(K > K_{1-\alpha} | H_0) = \alpha$ và miền tới hạn B_α là $(K_{1-\alpha}, +\infty)$.



Hình 5.3

§2. CÁC KIỂM ĐỊNH DÙNG MỘT MẪU

2.1. Kiểm định về kỳ vọng

Giả sử mẫu $x_1, x_2, \dots, x_n$ được chọn từ biến gốc $X \sim (\mu, \sigma^2)$.

Bài toán đặt ra là với mức ý nghĩa α cho trước hãy kiểm định giả thuyết $H_0: \mu = \mu_0$ (μ_0 đã cho).

1. Bài toán 1 (phương sai $\sigma^2 = \sigma_0^2$ đã biết)

Ta chọn tiêu chuẩn

$$K = \frac{\bar{X} - \mu_0}{\sigma_0} \sqrt{n}. \quad (2.1)$$

Rõ ràng nếu H_0 đúng, $K \sim \mathcal{N}(0, 1)$ xác định hoàn toàn. Phụ thuộc vào đối thuyết H_1 , ta có miền tới hạn khác nhau.

a) *Kiểm định hai phía* sẽ được dùng, nếu $H_1: \mu \neq \mu_0$. Khi đó do tính đối xứng của phân phối chuẩn $\mathcal{N}(0, 1)$, hai phân vị sẽ đối xứng qua gốc tọa độ và nếu đặt $z_b = K_{1-\alpha/2}$ (hay $\phi(z_b) = 0,5 - \alpha/2$) ta có miền tới hạn cần tìm

$$B_\alpha = (-\infty; -z_b) \text{ hoặc } (z_b; +\infty). \quad (2.2)$$

Tức là nếu $|K_{tn}| \leq z_b$ ta chấp nhận H_0 (không có cơ sở để bác bỏ H_0); $|K_{tn}| > z_b$ ta bác bỏ H_0 .

b) *Kiểm định một phía*

+ Nếu $H_1: \theta < \theta_0$, rõ ràng nên chọn miền tới hạn lệch trái (xem hình 5.2). Từ đó nếu đặt $z_b = K_\alpha$ (hay $\phi(z_b) = \alpha - 0,5$) ta có miền bác bỏ H_0

$$B_\alpha = (-\infty; z_b) \quad (2.3)$$

tức là nếu $K_{tn} < z_b$ (chú ý là z_b là số âm) ta bác bỏ H_0 , ngược lại ta chấp nhận nó.

+ Nếu $H_1: \theta > \theta_0$, miền tới hạn sẽ lệch phải (xem hình 5.3):

$$B_\alpha = (z_b, +\infty) \quad (2.4)$$

với $z_b = K_{1-\alpha}$ (hay $\phi(z_b) = 0,5 - \alpha$).

Để ý rằng miền chấp nhận H_0 (tính đối với thống kê K) chính là khoảng tin cậy với độ tin cậy $1 - \alpha$ cho kỳ vọng a_0 (khi biết phương sai $\sigma^2 = \sigma_0^2$). Người đọc có thể dễ dàng kiểm tra được khi so sánh (2.1) và (2.2) – (2.4) với các công thức tương ứng ở chương IV, §4 (các công thức (4.4), (4.7) – (4.9)).

Thí dụ 2.1. Một hãng bảo hiểm thông báo rằng số tiền trung bình hãng chi trả cho khách hàng bị tai nạn ô tô là 8500 đô la. Để kiểm tra lại, người ta kiểm tra ngẫu nhiên hồ sơ chi trả của 25 trường hợp thì thấy trung bình mẫu là 8900 đô la. Giả sử rằng số tiền chi trả tuân theo luật chuẩn với $\sigma = 2600$; hãy kiểm định lại thông báo của hãng bảo hiểm trên ($\alpha = 0,05$).

Giải. Ta chọn $H_0: a = 8500$, với $H_1: a \neq 8500$; ở đây mẫu được cảm sinh bởi $X \sim \mathcal{N}(a, \sigma^2) = \mathcal{N}(a, 2600^2)$. Miền bác bỏ giả thuyết H_0 là (xem (2.2)) $(-\infty, -z_b)$ hoặc $(z_b, +\infty)$ với z_b tìm từ bản phân phối: $\phi(z_b) - 0,5 - 0,025 = 0,475$, suy ra $z_b = 1,96$. Tính thống kê thực nghiệm

$$K_{tn} = \frac{\bar{X} - a_0}{\sigma_0} \sqrt{n} = \frac{8900 - 8500}{2600} \sqrt{25} \approx 0,77.$$

Rõ ràng $|0,77| = 0,77 < 1,96$; ta không có cơ sở để bác bỏ thông báo của hãng bảo hiểm.

Thí dụ 2.2. Một ông chủ cửa hàng thùng cho rằng dung tích trung bình của thùng là 55 lít ($\sigma = 6$ lít). Do kích thước tôn mua về đã cố định nên không có khả năng đóng được thùng có dung tích lớn hơn nữa. Hãy kiểm định lại ý kiến của ông chủ trên, biết rằng khi kiểm tra 36 thùng ta thấy dung tích trung bình chỉ có 49 lít ($\alpha = 0,001$).

Giải. Độ điều kiện của đầu bài, sẽ hợp lý nếu ta chọn $H_0: a = 55$ với $H_1: a < 55$ (α ký hiệu là dung tích trung bình lý thuyết). Để ý đến kết quả của chương IV, thống kê (2.1) sẽ có phân phối xấp xỉ chuẩn, ngay cả trong trường hợp chưa biết phân phối của biến gốc X . Từ đó miền tới hạn của cặp H_0, H_1 trên sẽ là $B_a = (-\infty, z_b)$ với z_b tìm từ $\Phi(z_b) = \alpha - 0,5 = -0,499$; suy ra $z_b = -3,09$. Mặt khác thống kê thực nghiệm

$$K_{tn} = \frac{\bar{X} - a_0}{\sigma_0} \sqrt{n} = \frac{49 - 55}{6} \sqrt{36} = -6.$$

Ta thấy $-6 < -3,09$, không có cơ sở để chấp nhận ý kiến của ông chủ cửa hàng thùng.

2. Bài toán 2 (phương sai σ^2 chưa biết)

Ta thay thống kê (2.1) bằng

$$K = \frac{\bar{X} - a_0}{s} \sqrt{n}, \quad (2.5)$$

trong đó s^2 là phương sai mẫu hiệu chỉnh. Khi H_0 đúng, ta đã biết K sẽ có phân phối Stiu-đơn $t(n-1)$. Mặt khác do tính đối xứng của phân phối này qua gốc tọa độ, nên cách làm rất giống với bài toán 1 ở trên.

a) Kiểm định hai phía với $H_0: a = a_0$ và $H_1: a \neq a_0$. Đặt $z_b = K_{1-\alpha/2} = t_{n-1, 1-\frac{\alpha}{2}}$ ta có miền tới hạn tương ứng

$$B_a = \{K_{tn} : |K_{tn}| > z_b\}. \quad (2.6)$$

b) Kiểm định một phía.

+ Nếu $H_1: \theta < \theta_0$; tìm $z_b = K_\alpha = t_{n-1, \alpha}$ và miền tới hạn sẽ là

$$B_\alpha = \{K_{tn} : K_{tn} < z_b\}. \quad (2.7)$$

+ Nếu $H_1: \theta > \theta_0$; tìm $z_b = K_{1-\alpha} = t_{n-1, 1-\alpha}$ và miền tới hạn sẽ là

$$B_\alpha = \{K_{tn} : K_{tn} > z_b\}. \quad (2.8)$$

Bạn đọc hãy so sánh (2.5) – (2.8) với (2.1) – (2.4) để thấy rõ sự giống và khác nhau giữa hai bài toán 1 và 2.

Thí dụ 2.3. Một nhà nhân chủng học cho rằng chiều cao trung bình của một bộ tộc người thiểu số là 160 cm. Người ta chọn ngẫu nhiên ra 16 người lớn của bộ tộc người đó thì thấy chiều cao trung bình là 164,25 cm với độ lệch chuẩn mẫu hiệu chỉnh là 6,25 cm. Có thể cho rằng bộ tộc người đó có chiều cao trung bình lớn hơn 160 cm hay không (giả sử chiều cao tuân theo luật phân phối chuẩn và α chọn bằng 0,05)?

Giải. Ở đây ta chọn $H_0: a = 160$ với $H_1: a > 160$ cùng với giả thiết chiều cao X tuân theo luật chuẩn. Với $\alpha = 0,05$, $n = 16$ ta có $z_b = t_{15; 0,95} = 1,753$. Mặt khác

$$K_{tn} = \frac{\bar{X} - a_0}{s} \sqrt{n} = \frac{164,25 - 160}{6,25} \sqrt{16} \approx 1,36.$$

Do $1,36 < 1,753$, ta không có cơ sở để bác bỏ H_0 , có nghĩa là ý kiến của nhà nhân chủng học là có thể tin được.

Chú ý rằng khi $n > 30$ việc tìm z_b trong các công thức (2.6) – (2.8) sẽ đưa về tra bảng Láp-la-xơ do tính xấp xỉ chuẩn của phân phối Stiu-đơn. Thậm chí người ta có thể bỏ qua cả giả thiết chuẩn của biến gốc X . Tuy nhiên các kết quả trong cả hai trường hợp đều chỉ là gần đúng (nhưng đòi hỏi mẫu lớn).

2.2. Kiểm định về tỷ lệ

Giống như bài toán 3 ở phần khoảng tin cậy, ta đi giải quyết bài toán kiểm định về tỷ lệ sau:

Bài toán 3. Với mức ý nghĩa α , hãy kiểm định giả thuyết $H_0: p = p_0$, biết rằng p là tham số phân phối $\mathcal{B}(1, p)$.

Ở chương IV ta đã biết nếu dung lượng mẫu n lớn và p không quá gần 0 hoặc 1 (tức là $np \geq 5$ hoặc $n(1 - p) \geq 5$) thì phân phối chuẩn có thể được dùng xấp xỉ phân phối nhị thức $\mathcal{B}(n, p)$. Nếu gọi $\hat{p} = m/n = f$ là tần suất mẫu – ước lượng của xác suất p , thì $\hat{p}$ sẽ có phân phối xấp xỉ chuẩn với kỳ vọng bằng p và phương sai $p(1 - p)/n$. Từ đó bài toán kiểm định về tỷ lệ không có khác biệt căn bản so với kiểm định về kỳ vọng.

Bạn đọc hãy tự tìm lấy các quy tắc kiểm định tương ứng (để ý đến mục 4.3 của chương IV). Chẳng hạn nếu chọn đối thuyết $H_1: p \neq p_0$ thì tiêu chuẩn kiểm định sẽ là

$$K = \frac{f - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} \quad (2.9)$$

và ta sẽ bác bỏ H_0 nếu $|K_{tn}| > z_b$ với $\phi(z_b) = 0,5 - \frac{\alpha}{2}$.

Thí dụ 2.4. Một toà báo thanh niên thông báo có 25% học sinh phổ thông trung học là độc giả thường xuyên. Một mẫu ngẫu nhiên gồm 200 học sinh được chọn cho thấy có 45 em đọc báo đó thường xuyên. Kiểm định tính chính xác của thông báo trên với mức ý nghĩa 0,05.

Giải. Rõ ràng nên chọn $H_0: p = 0,25$ với $H_1: p \neq 0,25$. Với $\alpha = 0,05$ giá trị tra bảng $z_b = 1,96$. Mặt khác theo (2.9)

$$K_{tn} = \frac{f - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} = \frac{(45/200) - 0,25}{\sqrt{0,25 \cdot (1 - 0,25)}} \sqrt{200} \approx -0,806$$

Từ đó do $|-0,806| < 1,96$ ta không có cơ sở để bác bỏ thông báo của tòa báo đó.

Thí dụ 2.5. Một hiệu làm đầu cho rằng 90% khách hàng của họ hài lòng với chất lượng phục vụ. Nghi ngờ chủ hiệu nói quá lên, một nhà điều tra xã hội học phỏng vấn 150 khách hàng của hiệu làm đầu thì thấy 132 người nói hài lòng. Với mức $\alpha = 0,05$; có thể trả lời thế nào cho nghi ngờ trên?

Giải. Ở đây ta nên chọn $H_0: p = 0,9$ với $H_1: p < 0,9$. Với $\alpha = 0,05$, giá trị z_b tìm được bằng các tra bảng $\phi(z_b) = \alpha - 0,5 = -0,45$, suy ra $z_b = -1,645$. Mặt khác

$$K_{tn} = \frac{f - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} = \frac{(132/150) - 0,9}{\sqrt{0,9 \cdot (1 - 0,9)}} \cdot \sqrt{150} \approx -0,833.$$

Từ đó do $-0,833 > -1,645$; ta không có cơ sở bác bỏ ý kiến của hiệu làm đầu.

2.3. Kiểm định về phương sai

Với giả thuyết chuẩn của biến gốc X và xuất phát từ một mẫu $x_1, x_2, \dots, x_n$, ta phải kiểm định giả thuyết sau:

Bài toán 4. Kiểm định $H_0: VX = \sigma_0^2$ (σ_0 đã biết) với mức ý nghĩa α . Để kiểm định giả thuyết trên ta dùng thống kê

$$K = \frac{(n-1)s^2}{\sigma_0^2}. \quad (2.10)$$

Nếu giả thuyết H_0 đúng thì từ chương IV ta biết $K \sim \chi^2(n-1)$

(chú ý rằng nếu thay $(n-1)s^2 = \sum_{i=1}^n (x_i - a_0)^2$, với $a_0 = EX$ đã

cho, thì thống kê trong (2.10) sẽ tuân theo luật $\chi^2(n)$ và cách làm sẽ giống như trường hợp trên với $n-1$ được thay bằng n). Từ đó phụ thuộc vào H_1 ta có các miền tới hạn khác nhau:

a) $H_1: \sigma^2 \neq \sigma_0^2$;

$$B_\alpha = \{K_{tn} : K_{tn} < \chi_{n-1, \alpha/2}^2 \text{ hoặc } K_{tn} > \chi_{n-1, 1-\alpha/2}^2\} \quad (2.11)$$

b) $H_1: \sigma^2 < \sigma_0^2$;

$$B_\alpha = \{K_{tn} : K_{tn} < \chi_{n-1, \alpha}^2\} \quad (2.12a)$$

c) $H_1: \sigma^2 > \sigma_0^2$

$$B_\alpha = \{K_{tn} : K_{tn} > \chi_{n-1, 1-\alpha}^2\} \quad (2.12a)$$

Bạn đọc hãy so sánh (2.11) – (2.12) với các công thức tương ứng của mục 4.4, chương IV.

Thí dụ 2.6. Chủ hãng sản xuất một loại thiết bị đo cho biết độ lệch chuẩn của sai số đo (giả sử nó tuân theo luật chuẩn) là 5 mm. Kiểm tra mẫu gồm 19 thiết bị đo thì thấy $s^2 = 33 \text{ mm}^2$. Với $\alpha = 0,05$ có thể kết luận gì về ý kiến của chủ hãng?

Giải. Ta chọn $H_0: \sigma^2 = 25$, còn đối thuyết hoặc $H_1: \sigma^2 \neq 25$, hoặc $H_1: \sigma^2 > 25$. Trong cả hai trường hợp

$$K_{tn} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{18,33}{25} \approx 23,76.$$

Mặt khác nếu cho $H_1: \sigma^2 \neq 25$ ta phải tra hai lần bảng

$$\chi_{18; 0,025}^2 = 8,2; \chi_{18; 0,975}^2 = 31,5;$$

Còn nếu $H_1: \sigma^2 > 25$ thì $\chi_{18; 0,95}^2 = 28,9$. Như vậy trong cả hai trường hợp ý kiến của chủ hãng đều có thể chấp nhận được (do $8,2 < 23,76 < 31,5$ hoặc $23,76 < 28,9$).

Thí dụ 2.7. Thử độ chịu lực của 35 chốt khoá thì thấy độ lệch chuẩn mẫu hiệu chỉnh là 3,5 pao (1 pao cỡ 450g). Có thể cho rằng bảo đảm của người sản xuất là độ lệch chuẩn thật bằng 3 pao được không?

Giải. Ta chọn $H_0: \sigma = 3$ với đối thuyết $H_1: \sigma > 3$. Tất nhiên có thể đưa về bài toán 4, nhưng ở đây dung lượng mẫu $n = 35$ nên ta sử dụng sự kiện (xem §2, chương IV) s có phân phối xấp xỉ chuẩn $\mathcal{N}\left(\sigma_0; \frac{\sigma_0^2}{2n}\right)$ nếu H_0 đúng. Từ đó ta đưa về dùng (2.4).

Với $\alpha = 0,05$, giá trị $z_b = 1,645$. Mặt khác

$$K_{tn} = \frac{s - \sigma_0}{\sigma_0} \sqrt{2n} = \frac{3,5 - 3}{3} \sqrt{70} \approx 1,39.$$

Do $1,39 < 1,645$ nên không có cơ sở để bác bỏ bảo đảm của nhà sản xuất.

§3. CÁC KIỂM ĐỊNH DÙNG NHIỀU MẪU

3.1. So sánh hai kỳ vọng

Giả sử ta có hai tập nền với hai biến gốc tương ứng $X \sim \mathcal{N}(a_1, \sigma_1^2)$ và $Y \sim \mathcal{N}(a_2, \sigma_2^2)$. Nếu muốn so sánh a_1 và a_2 người ta đưa ra giả thuyết $H_0: a_1 = a_2$. Thông tin mẫu gồm hai tập mẫu tương ứng $x_1, \dots, x_{n_1}$ và $y_1, \dots, y_{n_2}$.

Bài toán 1. Với mức α hãy kiểm định $H_0: a_1 = a_2$.

Ý tưởng lý thuyết là đưa về kiểm định $a_1 - a_2 = 0 = E(X - Y)$.

1) Nếu biết σ_1 và σ_2 ta sử dụng tiêu chuẩn

$$K = \frac{(\bar{X} - \bar{Y}) - (a_1 - a_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \quad (3.1)$$

Để kiểm tra từ chương IV, ta thấy thống kê (3.1) có phân phối $\mathcal{N}(0, 1)$ và nếu H_0 đúng thì $a_1 - a_2 = 0$. Từ đó giống như bài toán 1 của §2 ta có (trong đó z_α là phân vị chuẩn α).

a) Nếu $H_1: a_1 \neq a_2$, miền tới hạn một phía là (xem (2.2))

$$B_\alpha = \{K_{tn}: |K_{tn}| > z_{1-\frac{\alpha}{2}}\}. \quad (3.2)$$

b) Nếu $H_1: a_1 < a_2$, miền tới hạn một phía là (xem (2.3))

$$B_\alpha = \{K_{tn}: K_{tn} < z_\alpha\} \quad (3.3)$$

c) Nếu $H_1: a_1 > a_2$, miền tới hạn hai phía là (xem (2.4))

$$B_\alpha = \{K_{tn}: K_{tn} > z_{1-\alpha}\} \quad (3.4)$$

Để ý trong (3.3) và (3.4) $z_\alpha = -z_{1-\alpha}$.

2) Nếu σ_1^2 và σ_2^2 chưa biết ta lưu ý hai trường hợp:

+ Nếu n_1 và n_2 đủ lớn (>30) ta có thể tính toán xấp xỉ bằng cách dùng thống kê (3.1), nhưng các σ_1^2 và σ_2^2 thay bằng các ước lượng không chệch tương ứng của chúng là s_1^2 và s_2^2 . Bạn đọc tự viết các miền tới hạn tương ứng (xem (3.2). (3.4)).

+ Nếu n_1 và n_2 khá bé, vấn đề sẽ phức tạp hơn một chút. Ta sẽ sử dụng tiêu chuẩn sau:

$$K = \frac{(\bar{X} - \bar{Y}) - (a_1 - a_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (3.5)$$

Nếu thêm giả thiết hai biến gốc có phương sai giống nhau thì nếu H_0 đúng, thống kê $K \sim t(n_1 + n_2 - 2)$. Từ đó cách làm sẽ giống như bài toán 2 của §2.

$$a) H_1: a_1 \neq a_2 \text{ thì } B_\alpha = \{K_{tn} : |K_{tn}| > t_{n_1+n_2-2; 1-\frac{\alpha}{2}}\}; \quad (3.6)$$

$$b) H_1: a_1 < a_2 \text{ thì } B_\alpha = \{K_{tn} : K_{tn} < t_{n_1+n_2-2;\alpha}\}; \quad (3.7)$$

$$c) H_1: a_1 > a_2 \text{ thì } B_\alpha = \{K_{tn} : K_{tn} > t_{n_1+n_2-2; 1-\alpha}\}; \quad (3.8)$$

Bạn đọc thử so sánh (3.6) – (3.8) với (2.6) – (2.8) và sau đó với (3.2) – (3.4).

Thí dụ 3.1. Nghiên cứu trọng lượng sơ sinh của hai nhóm trẻ có mẹ không hút thuốc và hút thuốc trên 2 mẫu tương ứng, ta có

$$n_1 = 15; \bar{X}_1 = 3,5933; s_1 = 0,3707;$$

$$n_2 = 14; \bar{X}_2 = 3,2029; s_2 = 0,4927;$$

Giả sử trọng lượng trẻ ở các nhóm có phân phối chuẩn cùng phương sai. Với mức $\alpha = 0,05$ có thể cho rằng trẻ sơ sinh ở nhóm mẹ hút thuốc nhẹ cân hơn của nhóm mẹ không hút thuốc không?

Giải: Ta chọn $H_0: a_1 = a_2$ với đối thuyết $H_1: a_1 > a_2$. Theo (3.8) giá trị bảng $z_b = t_{n_1 + n_2 - 2; 1-\alpha} = t_{27; 0,95} = 1,703$. Mặt khác

$$K_{tn} = \frac{3,5933 - 3,2029}{\sqrt{\frac{14 \cdot 0,3707^2 + 13 \cdot 0,4927^2}{15 + 14 - 2} \left(\frac{1}{15} + \frac{1}{14} \right)}} \approx 2,42.$$

Theo (3.8) do $2,42 > 1,703$, có cơ sở để cho rằng trẻ ở nhóm mẹ không hút thuốc nặng hơn.

Thí dụ 3.2. Người ta nghiên cứu năng suất lúa mì ở hai vùng khác nhau. Ở vùng thứ nhất có 9 thửa ruộng được chọn với năng suất bình quân $\bar{X}_1 = 24,6$ tạ/ha và $s_1^2 = 0,24$; còn ở vùng thứ hai có 16 thửa ruộng với năng suất bình quân $\bar{X}_2 = 25,8$ tạ/ha và $s_2^2 = 0,16$. Với $\alpha = 0,05$ hỏi có sự sai khác đáng kể giữa các năng suất trung bình ở hai vùng không (giả sử năng suất là các biến ngẫu nhiên tuân theo luật chuẩn cùng phương sai)?

Giải. Để so sánh năng suất trung bình, ta chọn $H_0: a_1 = a_2$ với $H_1: a_1 \neq a_2$. Tra bảng Stiu-đơn cho ta $t_{23; 0,975} = 2,069$. Mặt khác

$$K_{in} = \frac{24,6 - 25,8}{\sqrt{\frac{8 \cdot 0,24 + 15 \cdot 0,16}{9 + 16 - 2 \left(\frac{1}{9} + \frac{1}{16} \right)}}} \approx -6,67.$$

Do $|-6,67| > 2,069$ không có cơ sở chấp nhận H_0 , hay năng suất trung bình có thể coi là khác nhau.

Chú ý:

– Nếu n_1 và n_2 khá lớn, ta có thể bỏ giả thiết chuẩn của đầu bài.

– Để ý là hai đối thuyết $a_1 > a_2$ và $a_1 < a_2$ dễ dàng chuyển đổi cho nhau bằng cách thay đổi thứ tự của hai mẫu.

– Trường hợp mẫu cặp (x_i, y_i) , $i = \overline{1, n}$, nên thiết lập hiệu $z_i = x_i - y_i$ và đưa về kiểm định giả thuyết $H_0: EZ = 0$ (mà trong §2 đã xét).

3.2. So sánh hai tỷ lệ

Cho hai tập nền có các biến gốc $X \sim \mathcal{B}(1, p_1)$ và $Y \sim \mathcal{B}(1, p_2)$. Ta có thể so sánh các xác suất p_1 và p_2 bằng kiểm định giả thuyết.

Bài toán 2. Với mức ý nghĩa α , hãy kiểm định $H_0: p_1 = p_2$.

Tiêu chuẩn kiểm định sẽ là

$$K = \frac{(f_1 - f_2) - (p_1 - p_2)}{\sqrt{\bar{f}(1 - \bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

trong đó f_i là tần suất mẫu tương ứng với hai mẫu $x_1, x_2, \dots, x_{n_1}$, và $y_1, y_2, \dots, y_{n_2}$; $\bar{f} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$. Nếu H_0 đúng và n_1, n_2

khá lớn thì

$$K = \frac{f_1 - f_2}{\sqrt{\bar{f}(1 - \bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (3.9)$$

có phân phối $\mathcal{N}(0, 1)$. Từ đó cách làm giống như bài toán 3 §2.

a) Nếu $H_1: p_1 \neq p_2$, tra bảng $\phi(z_b) = 0,5 - \frac{\alpha}{2}$ để tìm z_b và

$$B_\alpha = \{K_{tn} : |K_{tn}| > z_b\}. \quad (3.10)$$

b) Nếu $H_1: p_1 > p_2$, tra bảng $\phi(z_b) = 0,5 - \alpha$ để tìm z_b và

$$B_\alpha = \{K_{tn} : K_{tn} > z_b\}. \quad (3.11)$$

Nếu $H_1: p_1 < p_2$, ta có thể đổi số thứ tự hai mẫu để đưa về miền tới hạn (3.11).

Thí dụ 3.3. Kiểm tra chất lượng hai lô sản phẩm, người ta thấy trong lô thứ nhất gồm 500 sản phẩm có 50 phế phẩm, còn trong lô thứ hai gồm 400 sản phẩm thì có 60 phế phẩm. Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận gì so sánh chất lượng hai lô sản phẩm?

Giải. Gọi p_1, p_2 là các xác suất gặp phế phẩm của các lô hàng tương ứng. Ta cần kiểm định giả thuyết $H_0: p_1 = p_2$ với $H_1: p_1 \neq p_2$ ($\alpha = 0,05$). Ta sử dụng tiêu chuẩn (3.9):

$$f_1 = \frac{50}{500} = 0,1; f_2 = \frac{60}{400} = 0,15; \bar{f} = \frac{50 + 60}{500 + 400} \approx 0,12,$$

khi đó

$$K_{in} = \frac{f_1 - f_2}{\sqrt{\bar{f}(1-\bar{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0,1 - 0,15}{\sqrt{0,12 \cdot 0,88 \cdot \left(\frac{1}{500} + \frac{1}{400}\right)}} \approx -2,276.$$

Tra bảng tìm $z_b = 1,96$ và do $|-2,276| > 1,96$ ta không có cơ sở để chấp nhận H_0 . Chú ý rằng do mẫu lớn nên để ý đến phần 2 của bài toán 1 ta có thể tính xấp xỉ

$$K_{tn} = \frac{f_1 - f_2}{\sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}} = \frac{0,1 - 0,15}{\sqrt{\frac{0,1 \cdot 0,9}{500} + \frac{0,15 \cdot 0,85}{400}}} \approx -2,56$$

và H_0 càng bị bác bỏ. Nhưng để kết luận lô thứ nhất có chất lượng tốt hơn thì chưa đủ. Bây giờ ta chọn $H_1: p_1 < p_2$ và tự nhận xét theo (3.11), ta tìm z_b từ $\phi(z_b) = 0,5 - \alpha$, suy ra $z_b = 1,654$. Do $2,276 > 1,645$ nên không có cơ sở để chấp nhận H_0 ; ta chấp nhận H_1 , tức là tỷ lệ phế phẩm của lô hàng thứ nhất bé hơn đáng kể so với lô hàng thứ hai.

3.3. So sánh hai phương sai

Cho hai tập nền với hai biến gốc $X \sim \mathcal{N}(a_1, \sigma_1^2)$ và $Y \sim \mathcal{N}(a_2, \sigma_2^2)$. Từ hai mẫu tương ứng $x_1, x_2, \dots, x_{n_1}$ và $y_1, y_2, \dots, y_{n_2}$ ta muốn so sánh hai phương sai lý thuyết σ_1^2 và σ_2^2 .

Bài toán 3. Với mức ý nghĩa α hãy kiểm định $H_0: \sigma_1^2 = \sigma_2^2$.

Ta chọn thống kê

$$K = \frac{s_1^2 \cdot \sigma_2^2}{s_2^2 \cdot \sigma_1^2}$$

nếu $s_1^2 > s_2^2$. Phân phối của K đã xét ở chương IV. Nếu H_0 đúng tiêu chuẩn kiểm định trở thành

$$K = \frac{s_1^2}{s_2^2} \tag{3.12}$$

và $K \sim \mathcal{F}(n_1 - 1, n_2 - 1)$. Từ đó phụ thuộc vào đối thuyết H_1 ta có:

a) Nếu $H_1: \sigma_1^2 \neq \sigma_2^2$, ta có miền tới hạn

$$B_\alpha = \{K_{tn} : K_{tn} < F_{n_1-1, n_2-1, \frac{\alpha}{2}} \text{ hoặc } K_{tn} > F_{n_1-1, n_2-1, 1-\frac{\alpha}{2}}\}. \quad (3.13)$$

b) Nếu $H_1: \sigma_1^2 > \sigma_2^2$, ta có tương ứng

$$B_\alpha = \{K_{tn} : K_{tn} > F_{n_1-1, n_2-1, 1-\alpha}\}. \quad (3.14)$$

Thí dụ 3.4. Người ta đo tốc độ xuất phát của đạn khi súng phát hỏa khi thử nghiệm hai mẫu đạn của hai công ty khác nhau. Số liệu thử nghiệm của mẫu thứ nhất là $n_1 = 10$, $\bar{X}_1 = 1210$ và $s_1^2 = 2500$, còn của mẫu thứ hai $n_2 = 10$, $\bar{X}_2 = 1175$ và $s_2^2 = 3600$. Với mức $\alpha = 0,05$ có thể kết luận gì về chất lượng giống nhau của hai mẫu đạn (giả sử các biến X_1 và X_2 tuân theo luật chuẩn)?

Giải. Muốn đưa về mô hình so sánh kỳ vọng, ta phải có giả thiết là X_1 và X_2 cùng phương sai. Giả thiết đó có thể được thừa nhận dựa vào bài toán 3: kiểm định $H_0: \sigma_1^2 = \sigma_2^2$ với $H_1: \sigma_1^2 > \sigma_2^2$, xem (3.12) – (3.14). Tra bảng Phi-sơ ta có $F_{9;9;0,95} = 3,18$ và

$$K_{tn} = \frac{s_2^2}{s_1^2} = \frac{3600}{2500} \approx 1,41.$$

Từ đó do $1,41 < 3,18$ nên giả thuyết về sự bằng nhau của hai phương sai chấp nhận được.

Bây giờ ta kiểm định $H_0: \mu_1 = \mu_2$ với $H_1: \mu_1 \neq \mu_2$. Ta sẽ tính theo (3.5)

$$K_{tn} = \frac{1210 - 1175}{\sqrt{\frac{9 \cdot 2500 + 9 \cdot 3600}{10 + 10 - 2} \left(\frac{1}{10} + \frac{1}{10} \right)}} \approx 1,42.$$

Trong khi đó $t_{18; 0,975} = 2,101$ và do $|1,42| < 2,101$ giả thuyết $H_0: \mu_1 = \mu_2$ được chấp nhận. Chú ý trong thực hành khi n_1 và $n_2 > 30$ người ta còn xấp xỉ

$$s^2 \sim \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n}\right),$$

từ đó $s_1^2 - s_2^2 \sim \mathcal{N}\left(\sigma_1^2 - \sigma_2^2, \frac{2s_1^4}{n_1} + \frac{2s_2^4}{n_2}\right)$ và miền tới hạn hai phía của quy tắc kiểm định H_0 sẽ là (xem (3.2))

$$B_\alpha = \left\{ K_{tn} = \frac{s_1^2 - s_2^2}{\sqrt{\frac{2s_1^4}{n_1} + \frac{2s_2^4}{n_2}}} : |K_{tn}| > z_{1-\frac{\alpha}{2}} \right\}.$$

3.4. So sánh nhiều trung bình (phân tích phương sai)

Ta xét một trường hợp đơn giản là bài toán phân tích phương sai một nhân tố. Giả sử ta có k biến ngẫu nhiên gốc (ứng với k tập nền) $X_j \sim \mathcal{N}(a_j, \sigma^2)$, $j = \overline{1, k}$, với các tham số chưa biết. Để có thể so sánh các trung bình dựa trên k bộ số liệu mẫu x_{ij} , $i = \overline{1, n_j}$, $j = \overline{1, k}$ ta cần giải bài toán sau:

Bài toán 4. Với mức ý nghĩa α hãy kiểm định $H_0: a_1 = a_2 = \dots = a_k$ với đối thuyết H_1 : “tồn tại j_1 và j_2 sao cho $a_{j_1} \neq a_{j_2}$ ”.

Lưu ý là việc tách bài toán 4 thành nhiều bài toán 1 cho sai số rất lớn và khối lượng tính toán rất đồ sộ khi k lớn. Vì vậy ta sẽ dùng một kỹ thuật mới là *phân tích phương sai*, về mặt lý thuyết có hơi phức tạp, nhưng về mặt thực hành khá đơn giản. Để ý là các mẫu theo giả thiết đều có phân phối chuẩn cùng phương sai, và do nhiều mẫu nên ta có nhiều cách ước lượng phương sai đó.

Trước hết ta tính các đặc trưng mẫu trên cơ sở các số liệu x_{ij} , chỉ số i là thứ tự quan sát trong nội bộ mẫu của nhóm thứ j (gồm n_j số liệu), chỉ số j là số thứ tự nhóm (gồm k nhóm). Gọi n là tổng số các quan sát, từ đó

$$n = n_1 + n_2 + \dots + n_k = \sum_{j=1}^k n_j.$$

$$\overline{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}; \quad \bar{X} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij},$$

trong đó $\overline{X}_j$ là trung bình mẫu của nhóm j , còn $\bar{X}$ là trung bình chung.

Từ mỗi nhóm ta có thể xác định phương sai mẫu hiệu chỉnh của nhóm

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ij} - \overline{X}_j)^2;$$

và tính tổng bình phương các độ lệch riêng của các nhóm so với $\bar{X}$

$$S_r^2 = \sum_{j=1}^k (\overline{X}_j - \bar{X})^2 n_j. \quad (3.15)$$

Tổng bình phương các độ lệch được tính theo công thức

$$S_c^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{X})^2.$$

Bạn đọc có thể chứng minh được (để ý (3.15))

$$S_c^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{X})^2 = \sum_{j=1}^k (\overline{X}_j - \bar{X})^2 n_j + \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \overline{X}_j)^2; \quad (3.16)$$

trong đó tổng thứ nhất bên phải đặc trưng cho sự khác nhau giữa các nhóm, còn tổng thứ hai – giữa các số liệu trong nội bộ các nhóm. Bậc tự do của S_c^2 là $n - 1$, của S_r^2 là $k - 1$, dẫn đến

bậc tự do của $\sum_i \sum_j (x_{ij} - \overline{X}_j)^2$ sẽ là $n - k$ (bậc tự do của phân phối χ^2). Từ đó ta có hai ước lượng phương sai σ^2

$$S_1^2 = \frac{1}{k-1} S_r^2 = \frac{1}{k-1} \sum_{j=1}^k (\bar{X}_j - \bar{X})^2 n_j; \quad (3.17a)$$

$$S_2^2 = \frac{1}{n-k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_j)^2. \quad (3.17b)$$

Người ta chứng minh được rằng nếu H_0 đúng thì tỷ số S_1^2/S_2^2 tuân theo luật Phi-sơ – Sne-đơ-co với các bậc tự do $k-1$ và $n-k$. Từ đó miền tới hạn của quy tắc kiểm định sẽ là

$$B_\alpha = \left\{ K_{tn} = \frac{S_1^2}{S_2^2}; K_{tn} > F_{k-1, n-k, 1-\alpha} \right\}, \quad (3.18)$$

trong đó S_1^2/S_2^2 được xác định từ (3.17), $F_{k-1, n-k, 1-\alpha}$ là phân vị $1-\alpha$ của phân phối $\mathcal{F}(n-1, n-k)$.

Thí dụ 3.5. Người ta đo nồng độ haemoglobin ở 3 nhóm bệnh nhân mắc 3 dạng bệnh khác nhau A, B, C, kết quả đo bởi bảng:

Nhóm	n_j	x_{ij}									
A	16	7,2	7,7	8,0	8,1	8,3	8,4	8,4	8,5	8,6	
		8,7	9,1	9,1	9,1	9,8	10,1	10,3			
B	10	8,1	9,2	10,0	10,4	10,6	10,9	11,1	11,9	12,0	
		12,1									
C	15	10,7	11,3	11,5	11,6	11,7	11,8	12,0	12,1	12,3	
		12,6	12,6	13,3	13,3	13,8	13,9				

Hãy so sánh các nồng độ trung bình của các nhóm ($\alpha = 0,05$).

Giải. $n = 16 + 10 + 15 = 41$; $k = 3$; $\bar{X}_1 = 8,7425$; $\bar{X}_2 = 10,6300$; $\bar{X}_3 = 12,3000$; $s_1 = 0,8445$, $s_2 = 1,2841$, $s_3 = 0,9419$.

$$\sum_i \sum_j x_{ij} = 7,2 + 7,7 + \dots + 13,9 = 430,2;$$

$$\sum_i \sum_j x_{ij}^2 = 7,2^2 + 7,7^2 + \dots + 13,9^2 = 4651,80;$$

từ đó tổng bình phương

$$\begin{aligned}\sum_{i,j} (x_{ij} - \bar{X})^2 &= \sum_{ij} x_{ij}^2 - \left(\sum_{i,j} x_{ij} \right)^2 / n \\ &= 4651,80 - 430,2^2 / 41 = 137,85.\end{aligned}$$

Ta tính S_1^2 và S_2^2 :

$$\begin{aligned}S_1^2 &= \frac{1}{k-1} \sum_j n_j (\bar{X}_j - \bar{X})^2 = \frac{1}{k-1} \left[\sum_j n_j \bar{X}_j^2 - \left(\sum_{i,j} x_{ij} \right)^2 / n \right] \\ &= \frac{1}{2} (16.8,7125^2 + 10.10,6300^2 + 15.12,3000^2 - 430,2^2 / 41) \\ &= \frac{99,89}{2} = 49,94; \\ S_2^2 &= \frac{1}{n-k} \sum_{ij} (x_{ij} - \bar{X}_j)^2 = \frac{1}{n-k} \sum_j (n_j - 1) s_j^2 \\ &= \frac{1}{38} (15.0,8445^2 + 9.1,2841^2 + 14.0,9419^2) = \frac{37,96}{38} = 0,99.\end{aligned}$$

Cuối cùng từ (3.18) $K_{in} = \frac{49,94}{0,99} \approx 50,5$;

mặt khác nếu chọn $\alpha = 0,05$ thì $F_{2; 38; 0,95} = 3,24 < 50,5$. Như vậy không có cơ sở để chấp nhận H_0 , hay nồng độ haemoglobin của các nhóm bệnh khác nhau đáng kể.

§4. KIỂM ĐỊNH PHI THAM SỐ

4.1. Kiểm định giả thiết về luật phân phối

Trong nhiều bài toán thống kê, ta hay có giả thiết biến gốc X có phân phối chuẩn, phân phối Béc-nu-li... Trong thực tế nói chung không thể biết được X có phân phối nào từ đó dẫn đến bài

toán kiểm định tính đúng đắn của những giả thiết về phân phối. Cách giải quyết các bài toán dạng này làm giống như kiểm định tham số. Đầu tiên ta xác định giả thuyết, thí dụ như X tuân theo luật chuẩn, luật đều, luật Poa-xông..., và đối thuyết là X không có phân phối tương ứng đó. Sau đó dựa vào một tiêu chuẩn kiểm định và tính nó trên tập mẫu đã có để quyết định. Loại tiêu chuẩn ở đây được gọi là *tiêu chuẩn phù hợp*.

Có nhiều loại tiêu chuẩn phù hợp khác nhau. Trong mục này ta chỉ xét một tiêu chuẩn khá thông dụng mang tên Piéc-xơn và dùng tới phân phối χ^2 . Nó được xây dựng dựa trên sự so sánh tần suất thực nghiệm và xác suất lý thuyết của phân phối xác suất giả định.

Giả sử ta có một tập mẫu đã được phân lớp

x_1	$x_0 - x_1$	$x_1 - x_2$	$\dots$	$x_{i-1} - x_i$	$\dots$	$x_{k-1} - x_k$
n_1	n_1	n_2	$\dots$	n_i	$\dots$	n_k

với kích thước mẫu $n = n_1 + n_2 + \dots + n_k$ và $x_1 < x_2 < \dots < x_k$. Thông thường độ dài các khoảng chia bằng nhau (có thể khác nhau) và giá trị n_i không quá bé (≥ 5 , có thể chấp nhận ngoại lệ cho khoảng đầu và cuối). Giả thuyết đưa ra kiểm định có dạng H_0 : " X – có hàm phân phối xác suất $F(x)$ " với đối thuyết H_1 đối lập với H_0 . Lưu ý rằng nếu $F(x)$ phụ thuộc vào các tham số chưa biết, ta phải thay thế chúng bằng các ước lượng hợp lý nhất.

Nếu H_0 đúng, tỷ số n_i/n sẽ gần với xác suất p_i để biến X nhận giá trị trong khoảng thứ i (chú ý p_i hoàn toàn tính được dựa vào $F(x)$ đã biết). Từ đó Piéc-xơn đưa ra tiêu chuẩn

$$K = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}. \quad (4.1)$$

Rõ ràng K càng bé thì phân phối xác suất của X càng gần $F(x)$. Người ta đã chứng minh được rằng $n \rightarrow \infty$ thì phân phối xác suất của K không phụ thuộc vào phân phối của biến gốc X sẽ

xấp xỉ tới phân phối $\chi^2(k-r-1)$, với r là tham số chưa biết cần phải ước lượng.

Từ đó với mức ý nghĩa α , ta có thể xác định miền tới hạn cho tiêu chuẩn K trong (4.1)

$$B_\alpha = \{K_{tn} : K_{tn} > \chi_{k-r-1; 1-\alpha}^2\} \quad (4.2)$$

Chú ý khi tính $p_i = P(x_{i-1} < X < x_i) = F(x_i) - F(x_{i-1})$ cho các biến X liên tục thì x_0 chọn bằng $-\infty$ và $x_k = +\infty$. Nếu biến X là rời rạc việc tính p_i dựa vào hàm xác suất tương ứng với H_0 .

Thí dụ 4.1. Quan sát một thiết bị có 10 trạng thái tất cả 75 lần ta thu được kết quả

Trạng thái	1	2	3	4	5	6	7	8	9	10
Số lần n_i	5	8	3	11	4	5	4	14	13	8

Với $\alpha = 0,05$ có thể cho rằng vai trò các trạng thái như nhau hay không?

Giải. Nếu vai trò các trạng thái là như nhau thì số lần xuất hiện của chúng phải bằng nhau. Từ đó nếu gọi X là biến ngẫu nhiên chỉ số thứ tự của trạng thái thì X phải tuân theo luật phân phối đều rời rạc với $p_i = 0,1$, $i = \overline{1,10}$. Bài toán đưa về kiểm định H_0 : “ X có phân phối đều”, với $\alpha = 0,05$. Tính với $k = 10$, $n = 75$,

$$K_{tn} = \sum_{i=1}^{10} \frac{(n_i - np_i)^2}{np_i} = 19,0.$$

Tra bảng χ^2 tìm $\chi_{9;0,95}^2 = 16,92 < 19,0$, từ đó $K_{tn} \in B_{0,05}$ và giả thuyết H_0 không có cơ sở được chấp nhận.

Thí dụ 4.2. Quan sát số lượng ký sinh trùng trong hồng cầu của bệnh nhân mắc một loại bệnh về máu, ta có kết quả

Số lượng ký sinh trùng	0	1	2	3	4	≥ 5
Số người bệnh	40000	8621	1259	99	21	0

4.2. Kiểm định giả thuyết độc lập

Tiêu chuẩn χ^2 còn có thể dùng để kiểm định tính độc lập của hai đặc tính nào đó của các đối tượng ta quan tâm. Để kiểm định giả thuyết trên ta lập bảng như sau: Giả sử X và Y là hai đặc tính đang được quan tâm, lần lượt chúng có r và s thuộc tính, từ một mẫu dung lượng n ta có

$x_i \backslash y_j$	y_1	y_2	$\dots$	y_j	$\dots$	y_s	Σ
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1s}	l_1
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2s}	l_2
$\vdots$							
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{is}	l_i
$\vdots$							
x_r	n_{r1}	n_{r2}	$\dots$	n_{rj}	$\dots$	n_{rs}	l_r
Σ_i	m_1	m_2	$\dots$	m_j	$\dots$	m_s	n

n_{ij} là số lần quan sát đối tượng cùng có thuộc tính i của đặc tính X và thuộc tính j của đặc tính Y . Nếu ký hiệu $p(i, j)$, $p_x(i)$, $p_y(j)$ là các xác suất có đồng thời các thuộc tính i và j , có thuộc tính i , có thuộc tính j (của các đặc tính tương ứng), thì tính độc lập tương đương với (xem chương III)

$$p(i, j) = p_x(i)p_y(j).$$

Đặt $n_i = \sum_{j=1}^s n_{ij}$; $m_j = \sum_{i=1}^r n_{ij}$; $n = \sum_{i=1}^r \sum_{j=1}^s n_{ij}$, ta có thể ước lượng các xác suất trên bằng

$$\hat{p}_x(i) = \frac{n_i}{n}; \hat{p}_y(j) = \frac{m_j}{n}.$$

Rõ ràng nếu X và Y độc lập thì $\frac{n_{ij}}{n} \approx \frac{n_i \cdot m_j}{n^2}$, $i = \overline{1, r}$; $j = \overline{1, s}$. Từ đó nếu H_0 : “ X và Y độc lập” đúng, tiêu chuẩn

$$K = \sum_{i=1}^r \sum_{j=1}^s \frac{[n_{ij} - np(i, j)]^2}{np(i, j)} \quad (4.3)$$

sẽ là độ đo sự gần nhau giữa xác suất “lý thuyết” và thực nghiệm. Để ý K tuân theo luật χ^2 với số bậc tự do $(r-1)(s-1)$ khi n khá lớn. Vì vậy với α cho trước miền tới hạn của tiêu chuẩn K sẽ là

$$B_\alpha = \left\{ K_{tn} = n \left(\sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{n_i m_j} - 1 \right) : K_{tn} > \chi_{(r-1)(s-1); 1-\alpha}^2 \right\} \quad (4.4)$$

Thí dụ 4.4. Khảo sát màu mắt và tóc của 6800 người Pháp cho ta kết quả

Mắt \ Tóc					Σ
	Vàng	Nâu	Đen	Hung	
Xanh	1768	807	189	47	2811
Ghi	946	1387	746	53	3132
Nâu	115	438	288	16	857
Σ	2829	2632	1223	116	6800

Hỏi màu mắt và màu tóc có độc lập với nhau hay không ($\alpha = 0,05$)?

Giải: Theo (4.4) miền bác bỏ H_0 : “Màu mắt và màu tóc độc lập” là

$$\begin{aligned} B_\alpha &= \left\{ K_{tn} : K_{tn} > \chi_{(3-1)(4-1); 0,95}^2 \right\} \\ &= \left\{ K_{tn} : K_{tn} > 12,59 \right\}. \end{aligned}$$

Có thể dùng (4.3) hoặc (4.4) để tính tiêu chuẩn $K_{tn} = 1075$. Rõ ràng hai đặc tính trên không thể độc lập với nhau.

Thí dụ 4.5. Người ta tiến hành thăm dò về 3 ứng cử viên vào chức thị trưởng (là các ông A_1, A_2, A_3) ở ba quận (quận B_1, B_2, B_3). Kết quả thăm dò như sau (trên tổng số 310 người):

4.2. Kiểm định giả thuyết độc lập

Tiêu chuẩn χ^2 còn có thể dùng để kiểm định tính độc lập của hai đặc tính nào đó của các đối tượng ta quan tâm. Để kiểm định giả thuyết trên ta lập bảng như sau: Giả sử X và Y là hai đặc tính đang được quan tâm, lần lượt chúng có r và s thuộc tính, từ một mẫu dung lượng n ta có

$x_i \backslash y_j$	y_1	y_2	$\dots$	y_j	$\dots$	y_s	$\sum_j$
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1s}	n_1
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2s}	n_2
$\vdots$							$\vdots$
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{is}	n_i
$\vdots$							$\vdots$
x_r	n_{r1}	n_{r2}	$\dots$	n_{rj}	$\dots$	n_{rs}	n_r
$\sum_i$	m_1	m_2	$\dots$	m_j	$\dots$	m_s	n

n_{ij} là số lần quan sát đối tượng cùng có thuộc tính i của đặc tính X và thuộc tính j của đặc tính Y . Nếu ký hiệu $p(i, j)$, $p_x(i)$, $p_y(j)$ là các xác suất có đồng thời các thuộc tính i và j , có thuộc tính i , có thuộc tính j (của các đặc tính tương ứng), thì tính độc lập tương đương với (xem chương III)

$$p(i, j) = p_x(i)p_y(j).$$

Đặt $n_i = \sum_{j=1}^s n_{ij}$; $m_j = \sum_{i=1}^r n_{ij}$; $n = \sum_{i=1}^r \sum_{j=1}^s n_{ij}$, ta có thể ước lượng các xác suất trên bằng

$$\hat{p}_x(i) = \frac{n_i}{n}; \hat{p}_y(j) = \frac{m_j}{n}.$$

Rõ ràng nếu X và Y độc lập thì $\frac{n_{ij}}{n} \approx \frac{n_i \cdot m_j}{n^2}$, $i = \overline{1, r}$; $j = \overline{1, s}$. Từ đó nếu H_0 : “ X và Y độc lập” đúng, tiêu chuẩn

$$K = \sum_{i=1}^r \sum_{j=1}^s \frac{[n_{ij} - np(i, j)]^2}{np(i, j)} \quad (4.3)$$

sẽ là độ đo sự gần nhau giữa xác suất “lý thuyết” và thực nghiệm. Để ý K tuân theo luật χ^2 với số bậc tự do $(r-1)(s-1)$ khi n khá lớn. Vì vậy với α cho trước miền tới hạn của tiêu chuẩn K sẽ là

$$B_\alpha = \left\{ K_{tn} = n \left(\sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{n_i m_j} - 1 \right) : K_{tn} > \chi_{(r-1)(s-1); 1-\alpha}^2 \right\} \quad (4.4)$$

Thí dụ 4.4. Khảo sát màu mắt và tóc của 6800 người Pháp cho ta kết quả

Mắt \ Tóc					Σ
	Vàng	Nâu	Đen	Hung	
Xanh	1768	807	189	47	2811
Ghi	946	1387	746	53	3132
Nâu	115	438	288	16	857
Σ	2829	2632	1223	116	6800

Hỏi màu mắt và màu tóc có độc lập với nhau hay không ($\alpha = 0,05$)?

Giải: Theo (4.4) miền bác bỏ H_0 : “Màu mắt và màu tóc độc lập” là

$$\begin{aligned} B_\alpha &= \left\{ K_{tn} : K_{tn} > \chi_{(3-1)(4-1); 0,95}^2 \right\} \\ &= \left\{ K_{tn} : K_{tn} > 12,59 \right\}. \end{aligned}$$

Có thể dùng (4.3) hoặc (4.4) để tính tiêu chuẩn $K_{tn} = 1075$. Rõ ràng hai đặc tính trên không thể độc lập với nhau.

Thí dụ 4.5. Người ta tiến hành thăm dò về 3 ứng cử viên vào chức thị trưởng (là các ông A_1, A_2, A_3) ở ba quận (quận B_1, B_2, B_3). Kết quả thăm dò như sau (trên tổng số 310 người):

Ứng cử viên \ Quận	Quận			Σ
	B_1	B_2	B_3	
A_1	50	40	35	125
A_2	30	45	25	100
A_3	20	45	20	85
Σ	100	130	80	310

Hỏi có sự khác biệt đáng kể giữa các quận về sự tín nhiệm của cử tri hay không ($\alpha = 0,05$)?

Giải. Rõ ràng nên chọn giả thuyết gốc là tỷ lệ tín nhiệm của các cử tri đối với các ứng cử viên là như nhau ở cả 3 quận. Như vậy đối thuyết sẽ là tỷ lệ đó không giống nhau. Tuy nhiên bài toán so sánh 3 tỷ lệ không đơn giản, vì vậy ta chọn giả thuyết gốc H_0 tương đương như sau: “Mức tín nhiệm của cử tri độc lập với việc họ ở quận nào” và ta đưa về bài toán đã xét ở mục này (đối thuyết trong trường hợp này là có tồn tại quan hệ giữa mức tín nhiệm của cử tri với nơi ở của họ).

Trong công thức (4.3), n_{ij} chính là tần số người được thăm dò ở quận B_j bầu cho ứng cử viên A_i ; đây là tần số thực nghiệm. Còn $np(i, j)$ được hiểu là tần số “lý thuyết” tính trên tổng số người được thăm dò nếu giả thuyết H_0 đúng; nó sẽ bằng $n_i m_j / n$. Từ đó ta có bảng tần số mong muốn để H_0 đúng

Ứng cử viên \ Quận	Quận		
	B_1	B_2	B_3
A_1	40,32	52,42	32,26
A_2	32,26	41,94	25,81
A_3	27,42	35,65	21,94

Chú ý là các tổng hàng và cột của bảng số mới vẫn giống như bảng số cũ. Bây giờ ta đã có thể tính tiêu chuẩn

$$K_{tn} = \sum_{i=1}^3 \sum_{j=1}^3 \frac{\left(n_{ij} - \frac{n_i m_j}{n} \right)^2}{\frac{n_i m_j}{n}} \approx 10,539.$$

Bạn đọc có thể thiết lập bảng tính tương ứng, để ý là $\frac{n_i m_j}{n}$ đã được tính trong bảng vừa lập ở trên. Bây giờ ta chỉ còn việc tra bảng χ^2 để tìm ngưỡng của miền tới hạn $\chi^2_{(3-1)(3-1); 0,95} = 9,488$. Do $K_{tn} = 10,539 > 9,488$ ta không có cơ sở để chấp nhận giả thuyết H_0 và có thể kết luận rằng mức độ tín nhiệm của cử tri ở 3 quận là khác nhau.

BÀI TẬP

1. Một loại bóng đèn có tuổi thọ tuân theo luật phân phối chuẩn $\mathcal{N}(a, \sigma^2)$ với $\sigma = 150$. Cho giả thuyết gốc $H_0: a = 3600$ với đối thuyết $H_1: a = 3500$ và $\alpha = 0,01$. Nếu muốn xác suất phạm sai lầm loại hai bằng 0,05 thì cần đòi hỏi kích thước mẫu bằng bao nhiêu?
2. Một thầy giáo nghĩ rằng chỉ có 33% học sinh có làm bài tập ở nhà. Nhưng một cậu học sinh lại cho rằng thầy giáo có phần bi quan. Cậu chọn một nhóm ngẫu nhiên gồm 49 học sinh và thấy có 17 làm bài tập ở nhà. Với mức $\alpha = 0,05$ bạn xác định xem thầy giáo hay học sinh có lý hơn.
3. Một lô gà được thông báo là có trọng lượng trung bình là 1,6 kg. Nghi ngờ trọng lượng trung bình không đạt mức ấy, một người lấy ngẫu nhiên ra 24 con gà thì thấy giá trị trung bình mẫu là 1,5 kg với $s = 0,1$ kg. Với $\alpha = 0,01$, hãy kiểm định lại nghi ngờ trên.
4. Thông thường một máy đóng gói được coi là đạt yêu cầu nếu 90% sản phẩm đạt một trọng lượng quy định nào đó. Chọn hú họa ra 100 sản phẩm thì thấy có 87 đạt trọng

lượng quy định. Hãy xác định xem, với $\alpha = 0,05$, máy hoạt động đạt yêu cầu hay không?

5. Một hãng điều tra dư luận cho biết có 68% cử tri sẽ bỏ phiếu cho ứng cử viên A. Chọn ngẫu nhiên ra 36 cử tri thì thấy có 26 người bỏ phiếu cho ứng cử viên A. Với $\alpha = 0,05$ bạn có kết luận gì về kết quả điều tra của hãng trên.
6. Một dây chuyền sản xuất bóng đèn có tuổi thọ 750 giờ. Nghi ngờ do dây chuyền hoạt động đã lâu nên sản xuất kém chất lượng, người ta chọn ngẫu nhiên ra 10 bóng thì thấy tuổi thọ trung bình đạt 740 giờ với $s = 40$ giờ. Với mức $\alpha = 0,1$ có thể kết luận rằng chất lượng của dây chuyền trên có kém hơn hay không?
7. Một hãng truyền hình cho biết 70% khán giả xem chương trình phim truyện của hãng vào tối thứ bảy hàng tuần. Một hãng khác nghi ngờ tính chân thực của tuyên bố ấy đã làm một cuộc điều tra trên mẫu gồm 200 khán giả thì chỉ có 130 người nói có xem chương trình phim truyện trên. Với ngưỡng $\alpha = 0,05$ có thể cho rằng tuyên bố của hãng truyền hình đầu là nói hơi quá lên không?
8. Một máy tiện sản xuất ra một loại chi tiết có đường kính trung bình là 1,5 cm (giả sử đường kính đó tuân theo luật chuẩn), biết rằng độ lệch chuẩn của toàn bộ số chi tiết sản xuất ra là 0,01 cm. Người ta chọn ngẫu nhiên ra 25 chi tiết thì thấy đường kính trung bình là 1,501cm. Với ngưỡng $\alpha = 0,05$ có thể cho rằng máy tiện trên vẫn đạt yêu cầu hay không?
9. Để xác định độ béo của một loại pho mát, người ta chọn ngẫu nhiên ra 10 miếng, cắt đôi và hai nửa được gửi cho hai phòng thí nghiệm A và B. Kết quả xét nghiệm như sau :

Thứ tự miếng	1	2	3	4	5	6	7	8	9	10
A	40	39	40,2	38,2	39,7	37,7	41,4	36,5	40,7	38,9
B	41,9	39	40,7	39,3	39,2	38,2	41,3	38,5	39,8	38,7

Giả sử các số đo tuân theo luật chuẩn. Với $\alpha = 0,01$ có thể cho rằng các kết quả xét nghiệm của hai phòng thí nghiệm khác nhau cơ bản hay không?

10. Theo phương pháp nuôi thứ nhất có 12 con gà con bị chết trong số 200 con. Nuôi đối chứng 100 con theo cách nuôi thứ hai thì có 5 con bị chết. Với $\alpha = 0,05$ có thể kết luận phương pháp nuôi thứ hai tốt hơn không?
11. Chọn ngẫu nhiên 47 vòng bi cùng loại thì thấy độ lệch chuẩn trung bình của đường kính $s = 0,003$. Theo quảng cáo thì độ lệch chuẩn thật không vượt quá 0,0025. Vậy ta có thể kết luận gì ($\alpha = 0,05$)?
12. Nhà sản xuất đình tán cho biết đường kính định của ông ta có độ lệch chuẩn 0,01 cm. Chọn một mẫu ngẫu nhiên gồm 10 đình tán thì thấy $s = 0,018$ cm. Bạn sẽ nói gì về ý kiến của nhà sản xuất?
13. Một hãng sản xuất cho rằng chi phí trung bình cho một chuyến công tác đến nước A của nhân viên hãng đó là 1700\$. Nghiên cứu ngẫu nhiên chi phí của 10 lần công tác như vậy cho kết quả (\$)

1750 1693 1710 1730 1650
1720 1688 1703 1680 1760

Với $\alpha = 0,05$, kiểm định xem chi phí trung bình của một lần công tác có quá cao hay không?

14. Khảo sát hai siêu thị ở thành phố X, người ta thấy độ lệch chuẩn của số tiền mua hàng ở siêu thị tương ứng là 30000 đồng và 20000 đồng. Nghiên cứu 2 mẫu khách hàng ở hai siêu thị trên với $n_1 = 44$ và $n_2 = 15$ ta thấy số tiền trung bình chi để mua hàng là 150000 và 135000 đồng. Với $\alpha = 0,05$ hỏi có sự khác biệt cơ bản về chi phí mua hàng trung bình của khách hàng hai siêu thị trên hay không?

15. Hai máy cắt dây thép có các độ lệch chuẩn tương ứng $\sigma_1 = 0,26$ cm và $\sigma_2 = 0,31$ cm. Để kiểm tra xem hai máy có cắt dây cùng độ dài hay không, người ta chọn ngẫu nhiên 50 dây thép do mỗi máy cắt ra thì thấy có các độ dài trung bình mẫu tương ứng $\overline{X}_1 = 142,6$ cm, $\overline{X}_2 = 142,30$ cm. Hãy kiểm định với $\alpha = 0,05$.
16. Người ta chia các vận động viên thành hai nhóm: nhóm thứ nhất gồm 130 người được uống vitamin X, nhóm thứ hai gồm 128 người được uống thuốc giả (placebo). Sau một mùa thi đấu số người nhiễm cúm ở mỗi nhóm tương ứng là 30 và 39. Với $\alpha = 0,05$ ta có thể cho rằng vitamin X làm tăng đáng kể khả năng chống cúm của các vận động viên không? Có thể cùng kết luận như trên với $\alpha = 0,01$ không? Giải thích tại sao.
17. Để xác định xem người vùng núi cao có tuổi thọ trung bình cao hơn người ở vùng biển hay không, người ta chọn ngẫu nhiên ra hai mẫu. Ở mẫu thứ nhất gồm người vùng núi khi xét 50 giấy khai tử thấy tuổi thọ trung bình là 70 năm với độ lệch chuẩn $s_1 = 11,2$ năm; còn ở mẫu thứ hai (người vùng biển) 100 giấy khai tử cho thấy tuổi thọ trung bình là 65 năm với độ lệch chuẩn $s_2 = 12$ năm. Với mức $\alpha = 0,05$ có thể cho rằng người vùng núi thọ hơn người vùng biển không?
19. Người ta khảo sát 15 sinh viên để nghiên cứu hiệu quả của việc giảng dạy theo phương pháp mới. Trước khi học, sinh viên sẽ làm một bài kiểm tra (điểm tối đa bằng 100), sau khi học sẽ làm bài kiểm tra thứ hai. Kết quả điểm của từng học sinh như sau:

Số thứ tự	1	2	3	4	5	6	7	8
Trước học	54	79	91	75	68	43	33	85
Sau học	66	85	83	88	93	40	58	91

Số thứ tự	9	10	11	12	13	14	15
Trước học	22	56	73	63	29	75	87
Sau học	34	62	59	80	54	83	81

Với mức $\alpha = 0,05$ bạn có nhận xét gì về sự khác nhau giữa hai dãy điểm trên? Có thể coi rằng việc học theo phương pháp mới có hiệu quả hơn hay không?

20. Có hai máy tự động sản xuất cùng một loại sản phẩm. Từ các lô sản phẩm của mỗi máy ta chọn ra 10 sản phẩm và kết quả đo độ dài của các mẫu đó như sau:

Mẫu 1:	39,37	49,88	49,91	49,33	49,77
	49,81	50,01	50,14	49,75	50,15
Mẫu 2:	49,68	49,75	50,12	48,99	49,67
	49,99	50,20	50,11	50,02	49,72

Hỏi có sự khác nhau đáng kể giữa các độ dài trung bình của hai máy trên hay không (chọn $\alpha = 0,05$)?

21. Dùng một dụng cụ đo 200 lần ta tính được phương sai mẫu hiệu chỉnh $s_1^2 = 3,54$. Đo 16 lần bằng dụng cụ đo thứ hai cho ta $s_2^2 = 2,04$. Có thể cho rằng dụng cụ đo thứ hai chính xác hơn không ($\alpha = 0,05$)?
22. Khảo sát sở thích về 6 mác cà phê của 300 khách hàng thì thấy

Mác	A	B	C	D	E	F
Khách hàng	41	56	60	60	59	46

Với mức $\alpha = 0,01$, có thể cho rằng không có sự sai khác đáng kể về sở thích các loại cà phê khác nhau hay không?

23. Một công ty lắp máy mua linh kiện của hai nhà máy. Mỗi linh kiện có thể có 4 loại lỗi khác nhau. Sau khi thử nghiệm một số linh kiện người ta thu được kết quả (hai mẫu cùng kích thước)

Nhà máy \ Lỗi loại	1	2	3	4
A	60	50	40	30
B	30	32	25	25

Với mức $\alpha = 0,01$ có thể chấp nhận giả thuyết về sự giống nhau giữa các tỷ lệ lỗi của các linh kiện của hai nhà máy?

24. Một hãng bảo hiểm nghiên cứu tần số tai nạn tại gia trong các gia đình có từ hai con trở lên. Một mẫu gồm 200 gia đình được chọn và kết quả thống kê cho thấy:

Số tai nạn	0	1	2	3	4	5	7
Số gia đình	25	54	59	36	18	6	2

Theo bạn số tai nạn trên phù hợp với phân phối xác suất nào? Kiểm định điều nhận xét của bạn với mức $\alpha = 0,05$.

25. Người ta tiến hành bắn thử 100 loạt vào bia, mỗi loạt 10 viên vào 1 bia. Bảng số liệu quan sát như sau:

Số đạn trúng	0	1	2	3	4	5	6	7	8	9	10
Số bia	0	1	3	5	20	22	25	16	6	2	0

Hỏi số đạn bắn trúng 1 bia có tuân theo luật nhị thức không ($\alpha = 0,05$; tham số p được ước lượng bằng tần suất)?

26. Nghiên cứu 1000 đối tượng ta có bộ số liệu

x_i	5-15	15-25	25-35	35-45	45-55	55-65	65-75
n_i	45	197	308	212	198	22	18

Có thể cho rằng bộ số liệu được cảm sinh bởi một biến ngẫu nhiên chuẩn hay không ($\alpha = 0,05$)?

27. Cho bộ số liệu về chiều cao của một giống cây 2 tuần tuổi

Độ cao	5	7	9	11	13	15	17	19	21
Số cây	10	26	27	33	25	22	24	20	13

Với $\alpha = 0,01$ hãy kiểm định giả thuyết về phân phối chuẩn của độ cao trên.

28. Giám đốc thương mại của một hãng đồ chơi muốn nghiên cứu ý kiến khác hàng về một loại đồ chơi mới ở 4 vùng. Kết quả điều tra như sau:

Vùng	Không biết gì về đồ chơi	Giá đồ chơi vừa phải	Giá cao	Tổng
1	64	28	106	198
2	84	42	76	202
3	56	14	130	200
4	60	20	120	200
Tổng	264	104	432	800

Với mức $\alpha = 0,01$ có thể cho rằng các câu trả lời là giống nhau giữa 4 vùng trên?

Chương VI

PHÂN TÍCH HỒI QUY

§1. PHÂN TÍCH TƯƠNG QUAN

1.1. Hiệp phương sai và hệ số tương quan

Trong nhiều bài toán thực tế người ta quan tâm đến quan hệ của hai (hoặc nhiều) biến ngẫu nhiên. Tuy nhiên do thiếu thông tin, ta không thể nghiên cứu đầy đủ mọi đặc trưng của mối quan hệ đó. Thông thường ta chỉ có một bộ số liệu cặp $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ được xem như là cặp quan sát của hai biến ngẫu nhiên X, Y .

Nội dung chính của chương này là xác định sự phụ thuộc giữa các biến ngẫu nhiên. Rõ ràng nếu hai biến ngẫu nhiên độc lập, ta có thể nghiên cứu chúng riêng rẽ. Trong trường hợp chúng không độc lập, cần xác định mức độ phụ thuộc và quan hệ hàm giữa các biến.

Ta đã biết ở chương III một số đặc trưng quan trọng của cặp biến (X, Y) là hiệp phương sai $\mu_{XY} = cov(X, Y)$. Nếu $\mu_{XY} \neq 0$ ta có thể cho rằng hai biến X và Y có mối quan hệ nào đó, hay là chúng phụ thuộc ngẫu nhiên (còn gọi là tương quan). Do những hạn chế của khái niệm hiệp phương sai, ta đã đưa vào định nghĩa hệ số tương quan, ký hiệu là ρ_{XY} hay ρ nếu không sợ nhầm lẫn,

$$\rho_{XY} = \frac{\mu_{XY}}{\sigma_X \sigma_Y}, \quad (1.1)$$

trong đó ρ_X, ρ_Y là các độ lệch chuẩn tương ứng của X và Y . Số đặc trưng ρ_{XY} có các tính chất (xem chương III):

$$(i) |\rho_{XY}| \leq 1;$$

(ii) $|\rho_{XY}| = 1$ khi và chỉ khi $Y = aX + b$, trong đó a và b là các hằng số bất định;

(iii) Nếu X, Y độc lập thì $\rho_{XY} = 0$ (ngược lại nói chung không đúng).

Như vậy ta có thể dùng ρ để đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên. Khi $|\rho| = 1$ chúng có *quan hệ tuyến tính*; nếu $\rho = 0$ hai biến đó *không tương quan*; khi ρ khá gần 0 ta nói rằng chúng *tương quan yếu*, còn nếu $|\rho|$ khá gần 1 chúng *tương quan chặt* (hiểu theo nghĩa gần với tuyến tính). Nếu có thêm giả thiết chuẩn của X và Y thì $\rho_{XY} = 0$ tương đương với khẳng định X và Y độc lập.

1.2. Hệ số tương quan mẫu

Trong thực hành, ta không thể tính ρ chính xác được. Người ta thường xấp xỉ ρ bằng *hệ số tương quan mẫu*, ký hiệu là r , như sau:

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2 \sum_{i=1}^n (y_i - \bar{Y})^2}} \\ &= \frac{\sum_{i=1}^n x_i y_i - n \bar{X} \bar{Y}}{\left(\sum_{i=1}^n x_i^2 - n \bar{X}^2 \right) \left(\sum_{i=1}^n y_i^2 - n \bar{Y}^2 \right)}. \end{aligned} \quad (1.2)$$

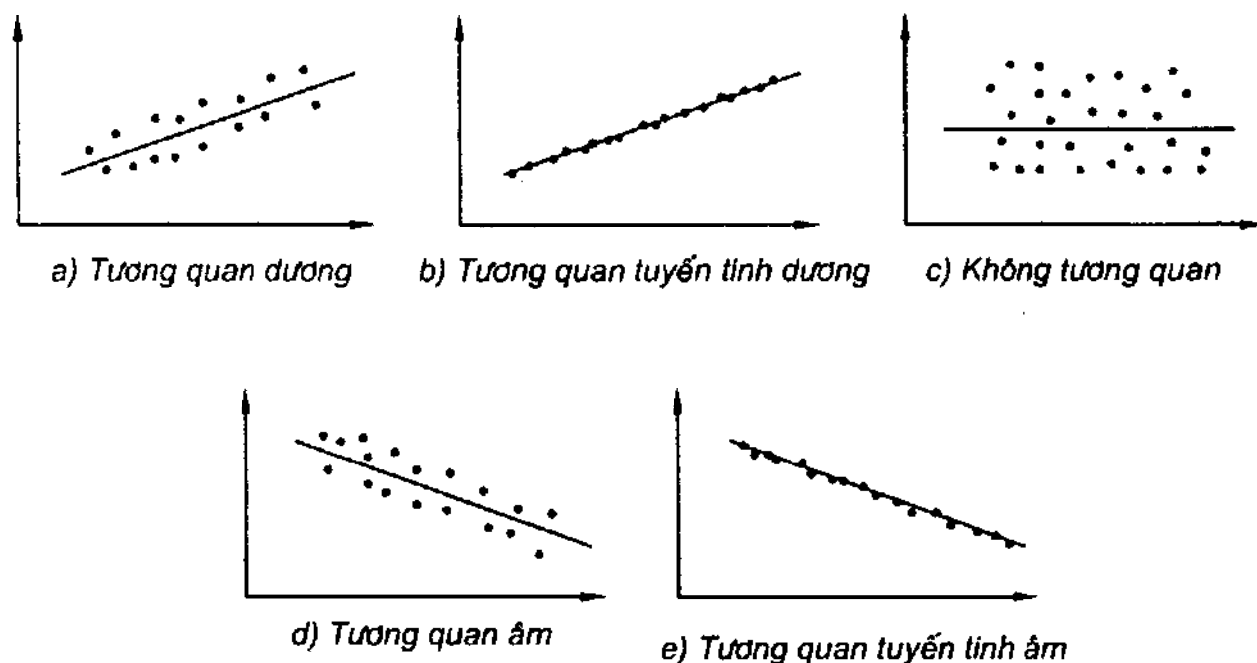
Tất nhiên ta có thể tính hiệp phương sai mẫu hiệu chỉnh:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}) = \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - n \bar{X} \bar{Y} \right).$$

và (1.2) sẽ trở thành gọn hơn (s_x^2 và s_y^2 là phương sai mẫu hiệu chỉnh tương ứng của X và Y)

$$r = \frac{s_{xy}}{s_x s_y}.$$

Giống như ρ , hệ số mẫu r là độ đo mức quan hệ cũng như chiều hướng của quan hệ giữa các giá trị x_i và y_i . Nếu ta biểu diễn các cặp số (x_i, y_i) là các điểm trên mặt phẳng tọa độ. Để các, ta sẽ có một đám mây điểm. Khi đó nếu $|r| = 1$ đám mây điểm sẽ tập trung trên một đường sẽ tập trung trên một đường thẳng. Nếu $r > 0$, đa số các giá trị lớn của y_i ứng với các giá trị lớn của x_i và ngược lại; ta nói tới *tương quan dương* hay tỷ lệ thuận. *Tương quan âm* sẽ có khi $r < 0$ và ta có quan hệ với khuynh hướng tỷ lệ nghịch (xem hình 1.1). Trong thực tế, do sai số quan sát, đo đạc hoặc tính toán mà



Hình 1.1. Các dạng tương quan

r rất khó bằng ± 1 (hoặc 0). Vì vậy nếu trong thực hành nếu $|r| > 0,8$, ta đã có thể coi là có mối quan hệ dạng tuyến tính (xấp xỉ tuyến tính) giữa hai biến đang xét.

Trong một số tài liệu, người ta còn xét *hệ số xác định mẫu* β_{xy} (viết tắt là β) được định nghĩa là $\beta = r^2$, với r đã xét trong (1.2). Rõ ràng β có miền xác định $0 \leq \beta \leq 1$ và là trường hợp riêng của khái niệm tương ứng dùng cho các hệ động học, nhiều chiều và phi tuyến.

Thí dụ 1.1. Tính các đặc trưng mẫu của bộ số liệu (x_i, y_i) , $i = \overline{1, 15}$, thể hiện ở cột 1 và cột 2 của bảng số dưới đây.

x_i	y_i	y'_i	x_i^2	$y_i'^2$	$x_i y'_i$
7,9	70,3	28,1	62,41	789,61	221,99
0,9	85,0	42,8	0,81	1831,84	38,52
3,7	100,0	57,8	13,69	3340,84	213,86
8,1	78,1	35,9	65,61	1288,81	290,79
6,9	77,9	35,7	47,61	1274,49	246,33
0,8	98,4	56,2	0,64	3158,44	44,96
6,0	59,2	17,0	36,00	289,00	102,00
7,2	86,8	44,6	51,84	1989,16	321,12
8,8	70,1	27,9	77,44	778,41	245,52
10,2	42,2	0,0	104,04	0,00	0,00
11,2	81,9	39,7	125,44	1576,09	444,64
0,5	97,1	54,9	0,25	3014,01	27,45
4,6	68,2	26,0	21,16	676,00	119,60
9,7	92,1	49,9	94,09	2490,01	484,03
1,0	91,2	49,0	1,00	2401,00	49,00
87,5	1198,5	565,5	702,03	24897,71	2489,81

Giải. Để tính các đặc trưng mẫu của bộ số liệu cặp (x_i, y_i) , ta phải tính các tổng sau:

$$\sum_i x_i, \sum_i y_i, \sum_i x_i^2, \sum_i y_i^2, \sum_i x_i y_i.$$

Để tránh phải tính toán với các số quá lớn, các giá trị của y được trừ đi 42,2 và ta thu được y' . Để ý là:

$$y' = y_i - 42,2; \bar{Y}' = \bar{Y} - 42,2; s_{y'}^2 = s_y^2; r_{xy'} = r_{xy}; \beta_{xy'} = \beta_{xy}.$$

Các kết quả tính trung gian được đưa ra trong bảng số (cột 3 – 6). Từ đó:

$$\bar{X} = \frac{87,5}{15} = 5,83; \bar{Y}' = \frac{565,5}{15} = 37,7;$$

$$\bar{Y} = 37,7 + 42,2 = 79,9;$$

$$s_x^2 = \frac{1}{14} (702,03 - 15 \cdot 5,83^2) = 13,7; s_x = 3,7;$$

$$s_y^2 = s_{y'}^2 = \frac{1}{14} (24897,71 - 15 \cdot 37,7^2) = 255,6; s_y = 16,0;$$

$$r_{xy} = r_{xy'} = \frac{2849,81 - 15 \cdot 5,83 \cdot 37,7}{14 \cdot 3,7 \cdot 16,0} = -0,5417;$$

$$\beta_{xy} = 0,2934.$$

Về mặt tính toán, khi x_i và y_i lớn và mẫu có kích thước lớn (n lớn), ta sẽ gặp khá nhiều khó khăn. Để đơn giản hơn, người ta đầu tiên sắp xếp số liệu dưới dạng bảng hai chiều. Giả sử trong bảng đó cặp giá trị (x_i, y_j) xuất hiện n_{ij} lần ($\sum \sum n_{ij} = n$).

Khi đó ta đổi biến số giống như đã làm ở §2, chương IV:

B1. Chọn x_0, y_0 với h_1, h_2 tương ứng;

B2. Tính các $u_i = \frac{x_i - x_0}{h_1}; v_j = \frac{y_j - y_0}{h_2};$

B3. Tính $\sum_i u_i n_{ui}; \sum_i u_i^2 n_{ui}; \sum_j v_j n_{vj}; \sum_j v_j^2 n_{vj}; \sum_i u_i v_i n_{uvi};$

B4. Tính $\bar{X} = \bar{u}h_1 + x_0; \bar{Y} = \bar{v}h_2 + y_0;$

$$S_x^2 = h_1^2 S_u^2; S_y^2 = h_2^2 S_v^2;$$

$$r = \frac{\sum_i u_i v_i n_{uvi} - n \bar{u} \bar{v}}{n \sqrt{[u^2 - (\bar{u})^2][v^2 - (\bar{v})^2]}}.$$

Ta xét quy trình tính toán trên một thí dụ cụ thể sau đây:

Thí dụ 1.2. Xác định các đặc trưng mẫu từ bảng số liệu sau:

$\begin{matrix} y_i \\ x_i \end{matrix}$	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	n_x
0 – 0,2	4	–	–	–	–	–	4
0,2 – 0,4	2	2	–	–	–	–	4
0,4 – 0,6	–	–	2	–	–	–	2
0,6 – 0,8	–	6	–	4	4	–	14
0,8 – 1,0	–	–	–	–	6	6	12
1,0 – 1,2	–	–	–	–	–	4	4
n_y	6	8	2	4	10	40	40

Giải. Đối với biến X ta chọn $x_0 = 0,7$; $h_1 = 0,2$ và biến mới $u_i = \frac{x_i - 0,7}{0,2}$, còn đối với Y chọn $y_0 = 27,5$; $h_2 = 5$ và biến mới $v_i = \frac{y_i - 27,5}{5}$. Bảng số trên trở thành

$\begin{matrix} v_i \\ u_i \end{matrix}$	-3	-2	-1	0	1	1	n_u
-3	4	–	–	–	–	–	4
-2	2	2	–	–	–	–	4
-1	–	–	2	–	–	–	2
0	–	6	–	4	4	–	14
1	–	–	–	–	6	6	12
2	–	–	–	–	–	4	4
n_v	6	8	2	4	10	10	40

Như vậy thay vì làm việc với các x_i không nguyên và y_i khá lớn, ta tính toán với các u_i và v_i đều nguyên và khá bé. Theo các công thức ở trên:

$$\bar{u} = \frac{1}{40} \sum_i u_i n_{ui} = \frac{1}{40} (-3.4 - 2.4 - 1.2 + 0.14 + 1.12 + 2.4) = -0,05;$$

$$\bar{v} = \frac{1}{40} \sum_i v_i n_{vi} = \frac{1}{40} (-3.6 - 2.8 - 1.2 + 0,4 + 1.10 + 2.10) = -0,15;$$

$$\overline{u^2} = \frac{1}{40} \sum_i u_i^2 n_{ui} = \frac{1}{40} (9.4 + 4.4 + 1.2 + 0.14 + 1.12 + 4.4) = 2,05;$$

$$\overline{v^2} = \frac{1}{40} \sum_i v_i^2 n_{vi} = \frac{1}{40} (9.6 + 4.8 + 1.2 + 0.4 + 1.10 + 4.10) = 3,45;$$

$$\sum_i u_i v_i n_{uvi} = (9.4 + 6.2 + 4.2 + 1.2 + 1.6 + 2.6 + 4.4) = 92.$$

$$\text{Từ đó } r = \frac{92 - 40.0,15.0,05}{40\sqrt{(2,05 - 0,0025)(3,45 - 0,0225)}} = \frac{91,7}{40.1,85.1,43} = 0,86;$$

$$\bar{X} = -0,05.0,2 + 0,7 = 0,69; \quad \bar{Y} = -0,15.5 + 27,5 = 26,75;$$

$$S_x = 1,43.0,2 = 0,286; \quad S_y = 1,85.5 = 9,25.$$

1.3. Tiêu chuẩn độc lập của hai biến ngẫu nhiên

Trong thực hành, như đã nói ở trên, r cũng hiếm khi bằng 0 và kết luận ngay cả X, Y không tương quan cũng đã gặp khó khăn. Nhưng nếu có thêm giả thiết chuẩn của phân phối (điều mà trong nhiều bài toán thực tế có thể chấp nhận được) thì điều kiện $\rho_{XY} = 0$ sẽ tương đương với sự kiện X và Y độc lập. Như vậy nếu ta thiết lập giả thuyết gốc H_0 : “ X và Y độc lập” (với đối thuyết H_1 đối lập), thì nếu có giả thiết X, Y tuân theo luật chuẩn, giả thuyết trên tương đương với H_0 : “ $\rho_{XY} = 0$ ”.

Tiêu chuẩn để kiểm định $H_0: \rho_{XY} = 0$ ($H_1: \rho_{XY} \neq 0$) là:

$$K = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad (1.3)$$

với r là hệ số tương quan tính trên tập mẫu ngẫu nhiên (x_i, y_j) , $i = \overline{1, n}$. Nếu giả thuyết H_0 đúng người ta đã chứng minh rằng $K \sim t(n-2)$; từ đó miền tới hạn của tiêu chuẩn k sẽ là:

$$B_\alpha = \left\{ K_{tn} : |K_{tn}| > t_{n-2; 1-\frac{\alpha}{2}} \right\}. \quad (1.4)$$

Nếu giả thuyết về tính độc lập của X và Y chấp nhận được, ít có lý do để xem xét đồng thời hai biến đó. Trong trường hợp ngược lại, ta sẽ quan tâm đến quan hệ của chúng.

Thí dụ 1.3. Cho cặp biến (X, Y) tuân theo luật chuẩn và bộ số liệu quan sát như sau:

x_i	12,0	16,5	15,2	11,7	18,3	10,9	14,4	16,0
y_i	2,75	3,37	2,86	2,62	2,76	3,49	3,12	3,05

Với $\alpha = 0,05$ hãy kiểm định tính độc lập của hai biến X và Y đó.

Giải. Trước hết ta tính hệ số tương quan mẫu theo (1.2)

$$n = 8; \sum x_i = 12,0 + 16,5 + \dots + 16,0 = 115;$$

$$\sum y_i = 2,75 + 3,37 + \dots + 3,05 = 24,02;$$

$$\sum x_i^2 = 1701,25; \sum y_i^2 = 72,798; \sum x_i y_i = 345,008;$$

$$\sum (x_i - \bar{X})^2 = \sum x_i^2 - \frac{1}{n} (\sum x_i)^2 = 1701,25 - \frac{115^2}{8} = 48,125;$$

$$\sum (y_i - \bar{Y})^2 = \sum y_i^2 - \frac{1}{n} (\sum y_i)^2 = 72,798 - \frac{24,02^2}{8} = 0,6780;$$

$$\begin{aligned} \sum (x_i - \bar{X})(y_i - \bar{Y}) &= \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i = 345,008 - \frac{115 \cdot 24,02}{8} \\ &\approx -0,2975; \end{aligned}$$

$$r = \frac{-0,2975}{\sqrt{48,125 \cdot 0,678}} \approx -0,0489.$$

Bài toán kiểm định tính độc lập của X và Y đưa về kiểm định giả thuyết

$$H_0: \rho_{XY} = 0; H_1: \rho_{XY} \neq 0 \ (\alpha = 0,05).$$

Theo (1.3), ta tính

$$K_{tn} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0,0489 \cdot \sqrt{6}}{\sqrt{1-0,0489^2}} \approx -0,1199.$$

Do từ bảng Stiu-đơn $t_{6;0,975} = 2,447 > |K_{tn}| = 0,1199$ nên theo (1.4) giả thuyết H_0 chấp nhận được.

Thí dụ 1.4. Kiểm định tính độc lập của hai biến X và Y cảm sinh ra bộ số liệu mẫu trong thí dụ 1.1 ($\alpha = 0,01$; X và Y tuân theo luật chuẩn).

Giải. Từ kết quả thực nghiệm $r = -0,5417$, ta có tiêu chuẩn (1.3) tính trên mẫu là:

$$K_{tn} = \frac{-0,5417\sqrt{3}}{\sqrt{1-0,5417^2}} = -2,32.$$

Mặt khác, việc tra bảng cho ta $t_{13;0,995} = 3,01$. Do $|K_{tn}| = 2,32 < 3,01$ nên ta không có cơ sở để bác bỏ H_0 . Ở đây ta thấy $r = -0,5417$ khác khá xa 0 mà ta vẫn chưa thể khẳng định là giữa X và Y có quan hệ nào đó. Nguyên nhân cũng có thể là kích thước mẫu quá bé chẳng? Lưu ý là nếu chọn $\alpha = 0,05$, ta có $t_{13;0,975} = 2,16$ và kết luận của kiểm định lại là bác bỏ H_0 . Như vậy có thể thấy rằng ước lượng hệ số tương quan phụ thuộc tới mức độ nào vào kích thước mẫu và những kết luận không dựa trên những tiêu chuẩn thống kê chính xác và hợp lý sẽ dẫn tới những sai lầm nguy hiểm.

1.4. Kiểm định giả thuyết về hệ số tương quan

1. Kiểm định $H_0: \rho = \rho_0$ cho trước

Theo một nghĩa nào đó, giả thuyết H_0 ở đây là trường hợp tổng quát hóa kết quả của mục trên. Người ta đưa ra thống kê

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}. \quad (1.5)$$

Theo Phi-sơ, nếu H_0 đúng thống kê Z sẽ tiệm cận tới phân phối chuẩn (khi $n \rightarrow \infty$) với các số đặc trưng xấp xỉ

$$EZ = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2(n-1)};$$

$$VZ = \frac{1}{n-3}.$$

Trong thực hành với $n > 50$ đã có thể chấp nhận kết quả trên. Từ đó nếu chọn tiêu chuẩn của quy tắc kiểm định là:

$$K = \frac{Z - EZ}{\sqrt{VZ}} = (Z - EZ) \sqrt{n-3} \sim \mathcal{N}(0, 1),$$

thì miền tới hạn của quy tắc sẽ là (miền đối xứng)

$$B_\alpha = \left\{ K_{in} : |K_{in}| > z_b \text{ với } \phi(z_b) = \frac{1-\alpha}{2} \right\}. \quad (1.6)$$

Thí dụ 1.5. Từ bộ số liệu thủy văn gồm 150 cặp, người ta tính được hệ số tương quan mẫu $r = 0,5273$. Với $\alpha = 0,05$; có thể cho rằng hệ số tương quan thật là 0,5 được không?

Giải. Theo (1.5) ta xác định giá trị thực nghiệm của Z

$$Z_{in} = \frac{1}{2} \ln \frac{1+0,5273}{1-0,5273} = 0,5862,$$

và các số đặc trưng tương ứng

$$EZ = \frac{1}{2} \ln \frac{1+0,5}{1-0,5} + \frac{0,5}{2 \cdot 149} = 0,5510;$$

$$\sqrt{VZ} = \frac{1}{\sqrt{147}} = 0,082,$$

từ đó:

$$K_{tn} = \frac{0,5862 - 0,5510}{0,082} = 0,43.$$

Với $\alpha = 0,05$, ta có $\phi(1,96) = \frac{1 - 0,05}{2} = 0,475$. Do $0,43 < 1,96$,

không có cơ sở để bác bỏ H_0 và chấp nhận rằng hệ số tương quan lý thuyết của tập nền là 0,5.

2. So sánh hai hệ số tương quan

Bài toán đưa về kiểm định $H_0: \rho_1 = \rho_2$ dựa trên hai bộ số liệu mẫu cặp (có kích thước tương ứng là n_1 và n_2) của hai cặp biến (X_1, Y_1) và (X_2, Y_2) . Bằng cách xác định hai thống kê Z_1 và Z_2 như trong (1.5), người ta đưa ra tiêu chuẩn kiểm định

$$K = \frac{Z_1 - Z_2}{\sqrt{VZ_1 + VZ_2}}.$$

Hàm số này có phân phối tiệm cận chuẩn $\mathcal{N}(0, 1)$ và ta có thể dùng lại quy tắc như trong (1.6) cho miền tới hạn đối xứng.

§2. HỒI QUY

2.1. Mô hình tuyến tính

Khi hai biến X và Y phụ thuộc, ta quan tâm đầu tiên đến quan hệ hàm $Y = f(x)$. Nếu hàm f tùy ý, đây là quan hệ rất phức tạp. Trong phần này ta giới hạn vào trường hợp f có dạng tuyến tính

$$Y = aX + b, \tag{2.1}$$

trong đó a, b là các hằng số thực cần xác định. Tuy nhiên do X và Y đều là các biến ngẫu nhiên, quan hệ (2.1) không giống

như quan hệ hàm theo nghĩa thông thường của giải tích. Về mặt lý thuyết người ta đưa vào khái niệm hồi quy tuyến tính (xem chương III) thông qua kỳ vọng có điều kiện

$$E(Y/X = x) = ax + b. \quad (2.1a)$$

Về mặt thực hành, để cho đơn giản, ta sẽ tất định hóa biến X , và sau này sẽ chuyển cách viết thành x , và gọi nó là *biến độc lập* (x là tất định theo nghĩa ta kiểm soát nó hoàn toàn). Y vẫn là *biến phụ thuộc* và là biến ngẫu nhiên, thể hiện của nó y_i là đáp ứng đối với giá trị x_i . Ta vẫn có bộ mẫu cặp kích thước n là (x_i, y_i) , $i = \overline{1, n}$. Với những đơn giản đó, công thức (2.1a) sẽ trở thành:

$$EY = ax + b.$$

Trong công thức (2.1b) chưa xuất hiện các yếu tố ngẫu nhiên gây ra tính bất định của biến Y . Vì thế để cho chặt chẽ và đầy đủ, người ta đưa vào khái niệm nhiễu, ký hiệu là ε , và thiết lập *mô hình tuyến tính*

$$Y_i = ax_i + b + \varepsilon_i, i = \overline{1, n}, \quad (2.2)$$

với ε_i là các biến ngẫu nhiên liên quan trực tiếp và gây ra sự bất định của Y_i . Ta sẽ yêu cầu ε_i thỏa mãn 2 điều kiện

$$(i) \mathcal{H}_1: E\varepsilon_i = 0 \quad \forall i = \overline{1, n}; \quad (2.3a)$$

$$(ii) \mathcal{H}_2: E(\varepsilon_i \varepsilon_j) = \sigma^2 \delta_{ij}, i, j = \overline{1, n} \quad (2.3b)$$

và sẽ gọi là nhiễu trắng (ký hiệu $\delta_{ij} = 0$ nếu $i \neq j$ và $= 1$ nếu $i = j$). Giả thiết $\mathcal{H}_1$ cho thấy ε_i có dạng sai số ngẫu nhiên, còn $\mathcal{H}_2$ yêu cầu chúng tạo ra dãy không tương quan. Như vậy trong mô hình (2.2), a và b là hai hệ số hồi quy chưa biết và sau này phải ước lượng, x_i là các hằng số đã biết, còn y_i là thể hiện của biến ngẫu nhiên phụ thuộc vào x_i . Ngoài ra tham số σ^2 đóng vai trò phương sai hằng của các nhiễu trắng ε_i và nó cũng chưa biết.

Trong thực tế, có thể việc giả sử là các x_i được xác định chính xác là không thật hợp lý. Tuy nhiên có thể yêu cầu tính

bất định của biến X là không đáng kể so với Y (mà thực nghiệm có thể chấp nhận được). Hơn nữa, vế phải của (2.2) đã có yếu tố nhiễu ngẫu nhiên, những khía cạnh ngẫu nhiên của X có thể trong một chừng mực nào đó chuyển sang cho nhiễu.

Tóm lại, bài toán đặt ra là trên cơ sở bộ số liệu quan sát $(x_i, y_i), i = \overline{1, n}$, hãy:

a) Ước lượng các hệ số hồi quy, tức là tìm $\hat{a}, \hat{b}$ và sau đó cả $\hat{\sigma}^2$;

b) Kiểm định tính phù hợp của mô hình (2.2) đối với bộ số liệu đã cho.

Khi $\hat{a}, \hat{b}$ đã xác định, ta có *đường hồi quy tuyến tính mẫu*

$$y = \hat{a}x + \hat{b}.$$

2.2. Ước lượng hệ số hồi quy

1. Phương pháp bình phương cực tiểu

Trong thực hành tập điểm số liệu nằm xung quanh dọc theo đường hồi quy (lý thuyết hoặc mẫu). Nếu trong (2.2) thay các biến ngẫu nhiên Y_i bằng các quan sát, ta có:

$$y_i = ax_i + b + \varepsilon_i, i = \overline{1, n}. \quad (2.4)$$

Trong (2.4) biến ε_i còn có thể hiểu như là *sai số* khi ta dùng mô hình tuyến tính để xấp xỉ quan hệ giữa 2 biến đang xét. Rõ ràng nếu phương sai sai số σ^2 càng nhỏ thì mô hình (2.4) càng phù hợp để mô tả quan hệ đó và tập số liệu đã cho.

Để ước lượng các hệ số a và b ta dùng định nghĩa sau:

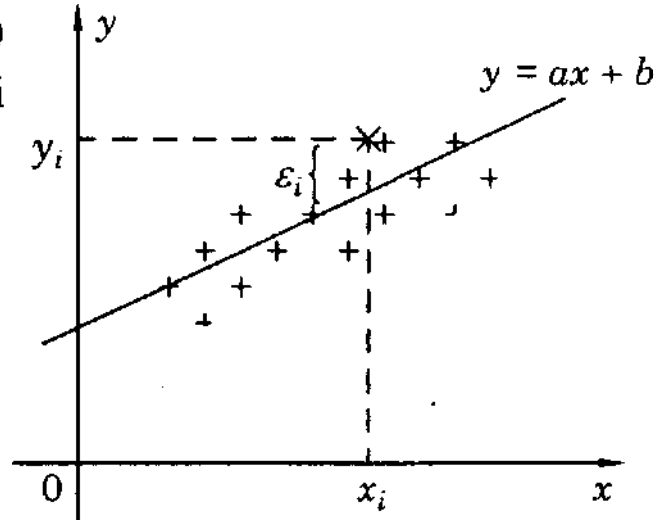
Định nghĩa. Các $\hat{a}$ và $\hat{b}$ được gọi là *ước lượng bình phương cực tiểu* của a và b , nếu:

$$\sum_{i=1}^n (y_i - \hat{a}x_i - \hat{b})^2 = \min Q(a, b) = \min \sum_{i=1}^n (y_i - ax_i - b)^2. \quad (2.5)$$

Phương pháp này có tên gọi là bình phương cực tiểu, còn hàm $Q(a, b) = \sum_i (y_i - ax_i - b)^2 = \sum_i \varepsilon_i^2$ chính là tổng bình phương sai số mô hình (2.4). Về mặt hình học đó chính là tổng bình phương khoảng cách (theo phương song song với trục tung) từ các điểm (x_i, y_i) đến đường thẳng hồi quy $y = ax + b$ (xem hình 2.1).

Việc cực tiểu hàm $Q(a, b)$ trong (2.5) đưa về giải hệ hai phương trình $\frac{\partial Q}{\partial a} = 0$ và $\frac{\partial Q}{\partial b} = 0$, suy ra:

$$\begin{cases} \hat{a}\bar{X} + \hat{b} = \bar{Y} \\ \hat{a} \sum_i x_i^2 + \hat{b} \sum_i x_i = \sum_i x_i y_i \end{cases}$$



Hình 2.1

Ta sẽ có:

$$\hat{a} = \frac{\sum_i (x_i - \bar{X})(y_i - \bar{Y})}{\sum_i (x_i - \bar{X})^2} = \frac{\sum_i x_i y_i - n\bar{X}\bar{Y}}{\sum_i x_i^2 - n\bar{X}^2}; \quad (2.6a)$$

$$\hat{b} = \bar{Y} - \hat{a}\bar{X}. \quad (2.6b)$$

Có nhận xét rằng

$$\sum_i (x_i - \bar{X})(y_i - \bar{Y}) = \sum_i (x_i - \bar{X})y_i = \sum_i (y_i - \bar{Y})x_i.$$

Việc kiểm tra điều kiện đủ không cần thiết do Q là hàm dạng bình phương. Đường hồi quy mẫu $y = \hat{a}x + \hat{b}$ sẽ đi qua trọng tâm của tập điểm, tức là điểm $(\bar{X}, \bar{Y})$ sẽ nằm trên đường thẳng đó.

Để ý là trong (2.2) Y_i là biến ngẫu nhiên có:

$$EY_i = ax_i + b; \quad (2.7a)$$

$$VY_i = V\varepsilon_i = \sigma^2; \quad (2.7b)$$

và σ^2 có thể dễ ước lượng bằng phương sai mẫu của Y_i

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{Y})^2. \quad (2.8)$$

Sau khi đã có $\hat{a}$ và $\hat{b}$, ta đặt $\hat{y}_i = \hat{a}x_i + \hat{b}$ và hiệu $\hat{\varepsilon}_i = y_i - \hat{y}_i$ sẽ được gọi là *phần dư* (sai số thực nghiệm) của mô hình để ý:

$$\hat{\varepsilon}_i = y_i - \hat{a}x_i - \hat{b} = (a - \hat{a})x_i + b - \hat{b} + \varepsilon_i.$$

Người ta cũng chứng minh được rằng ước lượng không chệch của σ^2 theo phương pháp bình phương cực tiểu không phải là (2.8) mà là:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{a}x_i - \hat{b})^2. \quad (2.9)$$

2. Phương pháp hợp lý nhất

Nếu ta thêm vào giả thiết chuẩn của ε_i

$$(iii) \mathcal{H}_3 : \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i = \overline{1, n}, \quad (2.3c)$$

thì dễ dàng chứng tỏ $Y_i \sim \mathcal{N}(ax_i + b, \sigma^2)$. Khi đó *ước lượng hợp lý nhất* của a và b sẽ hoàn toàn giống như (2.6) (bạn đọc có thể tự chứng minh, sử dụng hàm hợp lý xây dựng trên các quan sát y_i của biến ngẫu nhiên chuẩn Y_i (xem chương IV). Lưu ý ước lượng hợp lý nhất của phương sai σ^2 sẽ là:

$$S^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{a}x_i - \hat{b})^2. \quad (2.10)$$

Cũng chú ý là khi dùng (2.9) hoặc (2.10), tổng các bình phương ước lượng sai số sẽ được tính như sau:

$$\sum \hat{\varepsilon}_i^2 = \sum (y_i - \bar{Y})^2 - \hat{a} \sum (x_i - \bar{X})^2 = n(S_y^2 - \hat{a}S_x^2). \quad (2.11)$$

Thí dụ 2.1. Kết quả nghiên cứu thực nghiệm trên 8 người đàn ông như sau:

Trọng lượng (kg)	58,0	70,0	74,0	63,5	62,0	70,5	71,0	66,0
Huyết tương (l)	2,75	2,86	3,37	2,76	2,62	3,49	3,05	3,12

Hãy xây dựng đường hồi quy tuyến tính mẫu của huyết tương với trọng lượng.

Giải. Gọi X là trọng lượng cơ thể, còn Y là lượng huyết tương. Ở đây $n = 8$ và các tổng lấy theo $i = \overline{1, 8}$:

$$\sum x_i = 535; \sum y_i = 24,02; \bar{X} = 66,875; \bar{Y} = 3,0025;$$

$$\sum x_i^2 = 35983,5; \sum y_i^2 = 72,7980; \sum x_i y_i = 1615,295.$$

Từ đó

$$\sum (x_i - \bar{X})(y_i - \bar{Y}) = 1615,295 - 535 \cdot 24,02/8 = 8,96;$$

$$\sum (x_i - \bar{X})^2 = 35983,5 - 535^2/8 = 205,38;$$

$$\sum (y_i - \bar{Y})^2 = 72,7980 - 20,02^2/8 = 0,6780.$$

và

$$r = \frac{8,96}{\sqrt{205,38 \cdot 0,6780}} = 0,76;$$

$$\hat{a} = \frac{8,96}{205,38} = 0,043615;$$

$$\hat{b} = 3,0025 - 0,043615 \cdot 66,875 = 0,0857.$$

Như vậy sự phụ thuộc của lượng huyết tương vào trọng lượng cơ thể có thể được mô tả bằng (hồi quy mẫu):

$$y = 0,0436x + 0,0857.$$

Ta có thể tích ước lượng không chệch của phương sai sai số mô hình trên theo (2.9) có để ý đến (2.11)

$$s^2 = \frac{1}{6} (0,6780 - 0,0436^2 \cdot 205,38) = 0,047929.$$

Thí dụ 2.2. Xây dựng đường hồi quy tuyến tính mẫu theo thí dụ 1.2.

Giải. Áp dụng công thức (2.6) và kết quả tính toán của thí dụ 1.2

$$\hat{a} = r \frac{S_y}{S_x} = 0,86 \cdot \frac{9,25}{0,286} = 27,8147;$$

$$\hat{b} = \bar{Y} - \hat{a}\bar{X} = 26,75 - 27,8147 \cdot 0,69 = -7,5578;$$

và đường hồi quy cần tìm là:

$$y = 27,81x - 7,56.$$

3. Các tính chất của ước lượng bình phương cực tiểu

a) Các tính chất thống kê

(i) $\hat{a}$ và $\hat{b}$ là ước lượng không chệch của a và b .

Ta chứng minh cho $\hat{a}$, việc chứng minh đối với $\hat{b}$ rất đơn giản. Để ý đến (2.4)

$$y_i = ax_i + b + \varepsilon_i, i = \overline{1, n},$$

ta có:

$$\bar{Y} = a\bar{X} + b + \bar{\varepsilon};$$

($\bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i$). Từ đó $y_i - \bar{Y} = a(x_i - \bar{X}) + \varepsilon_i - \bar{\varepsilon}$ và thay vào công thức (2.6a):

$$\hat{a} = a + \frac{\sum_{i=1}^n (x_i - \bar{X}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{X})^2},$$

từ đó $E(\hat{a}) = a$. Nếu ký hiệu $\hat{y}_i = \hat{a}x_i + \hat{b}$, dễ thấy:

$$E(\hat{y}_i - y_i) = 0 \quad \forall i = \overline{1, n}.$$

$$\text{cov}(\hat{a}, \hat{b}) = \mu_{ab} = -\frac{\sigma^2 \bar{X}}{\sum_{i=1}^n (x_i - \bar{X})^2}.$$

b) Ý nghĩa hình học

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}; x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \text{ và } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$
$$y = ax + be + \varepsilon.$$


Việc tìm ước lượng bình phương cực tiểu dẫn đến xấp xỉ tốt nhất của y trong mặt phẳng P sinh bởi hai vectơ e và x (xem hình 2.2). Để ý $\hat{y}$ là hình chiếu vuông góc của y trên P , ngọn của $\hat{y}$ là điểm của mặt phẳng gần với (ngọn của) y nhất. Hình chiếu vuông góc của y lên e sẽ là $\bar{Y}e$. Đồng thời dựng từ gốc vectơ $x - \bar{X}e$, ta có định lý Ta-lét:

$$\hat{y} - \bar{Y}e = \hat{a}(x - \bar{X}e).$$

Vectơ vuông góc với mặt phẳng chính là vectơ phần dư $\hat{\varepsilon} = y - \hat{y}$. Trong tam giác vuông $(y, \hat{y}, \bar{y})$, theo định lý Pi-ta-go

$$\sum_i (y_i - \bar{Y})^2 = \sum_i \hat{\varepsilon}_i^2 + \sum_i (\hat{y}_i - \bar{\hat{y}})^2, \quad (2.13)$$

(để ý $\bar{\hat{y}} = \bar{Y}$). Do $\sum \hat{\varepsilon}_i = 0$ nên $\bar{\hat{\varepsilon}} = 0$ và $\sum \hat{\varepsilon}_i^2 = \sum (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2$.

Chia hai vế của (2.13) cho n , ta thu được:

$$\frac{1}{n} \sum_i (y_i - \bar{Y})^2 = \frac{1}{n} \sum_i (\hat{y}_i - \bar{\hat{y}})^2 + \frac{1}{n} \sum_i (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2, \quad (2.14)$$

trong đó:

$\frac{1}{n} \sum_i (y_i - \bar{Y})^2 = S_y^2$ là tổng phương sai của tập mẫu $(y_1, \dots, y_n)$,

$\frac{1}{n} \sum_i (\hat{y}_i - \bar{\hat{y}})^2$ là phương sai của tập điều chỉnh $(\hat{y}_1, \dots, \hat{y}_n)$,

$\frac{1}{n} \sum_i (\varepsilon_i - \bar{\varepsilon})^2$ là phương sai dư.

Như vậy (2.14) chính là phương trình phân tích phương sai (xem (3.16) bài toán 4, §3 chương V). Cũng theo Ta-lét:

$$\sum_i (y_i - \bar{Y})^2 = \sum_i \hat{\varepsilon}_i^2 + \hat{a}^2 \sum_i (x_i - \bar{X})^2,$$

hay $S_y^2 = S_{\hat{\varepsilon}}^2 + \hat{a}^2 S_x^2$.

2.3. Trường hợp có giả thiết chuẩn

1. Phân phối của ước lượng

Mô hình (2.4) với các giả thiết $\mathcal{H}_1 - \mathcal{H}_3$ có thể tóm tắt lại (xem (2.3a – c))

$$\begin{aligned}y_i &= ax_i + b + \varepsilon_i, i = \overline{1, n}, \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma^2); E(\varepsilon_i \varepsilon_j) = 0, i \neq j.\end{aligned}$$

Khi đó ta có thể xác định luật phân phối xác suất của các ước lượng $\hat{a}$, $\hat{b}$ và s^2 (trong (2.9)). Trước hết:

$$\hat{a} \sim \mathcal{N}(0, \sigma_a^2); \hat{b} \sim \mathcal{N}(b, \sigma_b^2), \quad (2.15)$$

với σ_a^2 và σ_b^2 đã xác định trong (2.12). Tương tự:

$$(n-2) \frac{s^2}{\sigma^2} \sim \chi^2(n-2), \quad (2.16)$$

với $v(s^2) = \frac{2\sigma^4}{n-2}$. Để ý là $\hat{a}$ và $\hat{b}$ về mặt lý thuyết là các biến ngẫu nhiên độc lập với $\varepsilon_i, i = \overline{1, n}$. Ngoài ra lưu ý đến công thức của $cov(\hat{a}, \hat{b})$ trong (2.12), ta có phân phối đồng thời của $\hat{a}$ và $\hat{b}$ là:

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} a \\ b \end{pmatrix}; \sigma_a^2 \begin{pmatrix} 1 & -\bar{X} \\ -\bar{X} & \frac{\sum_i x_i^2}{n} \end{pmatrix} \right). \quad (2.17)$$

Các phân phối (2.15) – (2.17) đặt cơ sở cho các bài toán kiểm định giả thuyết hoặc tìm khoảng tin cậy cho các ước lượng hoặc cho các dự báo dùng hồi quy trong thực hành.

Người ta cũng chứng minh được rằng $\hat{a}$ và $\hat{b}$ cũng là ước lượng hiệu quả của a và b .

2. Khoảng tin cậy

Đầu tiên ta tìm khoảng tin cậy $1 - \alpha$ cho ước lượng $\hat{a}$. Theo (2.15) và (2.16):

$$\frac{\hat{a} - a}{\sigma / \sqrt{\sum_i (x_i - \bar{X})^2}} \sim \mathcal{N}(0, 1);$$

$$(n-2) \frac{s^2}{\sigma^2} \sim \chi^2(n-2),$$

nên thống kê:

$$\frac{\hat{a} - a}{s / \sqrt{\sum_i (x_i - \bar{X})^2}} \sim t(n-2).$$

Kết quả §4, chương V, cho ta khoảng tin cậy $1 - \alpha$ (xét khoảng đối xứng, các khoảng dạng khác bạn đọc dễ dàng tự tìm được):

$$\hat{a} - \frac{st_{n-2, 1-\alpha/2}}{\sqrt{\sum_i (x_i - \bar{X})^2}} < a < \hat{a} + \frac{st_{n-2, 1-\alpha/2}}{\sqrt{\sum_i (x_i - \bar{X})^2}}. \quad (2.18a)$$

Tương tự:

$$\hat{b} - st_{n-2, 1-\alpha/2} \sqrt{\frac{\sum_j x_j^2}{n \sum_i (x_i - \bar{X})^2}} < b < \hat{b} + st_{n-2, 1-\alpha/2} \sqrt{\frac{\sum_j x_j^2}{n \sum_i (x_i - \bar{X})^2}}. \quad (2.18b)$$

$$(n-2) \frac{s^2}{\chi_{n-2, 1-\alpha/2}^2} < \sigma^2 < (n-2) \frac{s^2}{\chi_{n-2, \alpha/2}^2}. \quad (2.18c)$$

Thí dụ 2.3. Tìm các khoảng tin cậy 95% cho các tham số trong thí dụ 2.1.

Giải. Trước hết lưu ý rằng phải có giả thiết chuẩn thì các khoảng (2.18a-c) mới dùng được. Bây giờ ta tra các bảng phân vị $t(6)$ và $\chi^2(6)$

$$t_{6, 0,975} = 2,447; \chi^2_{6,0,975} = 14,449; \chi^2_{6, 0,025} = 1,237.$$

Từ đó ta có các khoảng tin cậy 95% tương ứng cho a , b và σ^2 :

$$\left(0,0436 - \frac{0,2189.2,447}{\sqrt{205,38}}; 0,0436 + \frac{0,2189.2,447}{\sqrt{205,38}} \right) \\ = (0,0062; 0,0810)$$

$$\left(0,0857 - 0,2189.2,447\sqrt{\frac{35893,5}{8.205,38}}; \right. \\ \left. 0,0857 + 0,2189.2,447\sqrt{\frac{35893,5}{8.205,38}} \right) \\ = (-2,8049; 2,9763);$$

$$\left(\frac{0,2867}{14,449}; \frac{0,2876}{1,237} \right) = (0,0199; 0,2325).$$

Ta có thể lưu ý rằng:

– Phân phối đồng thời (2.17) cho phép xây dựng miền tin cậy cho vectơ tham số (miền tin cậy đồng thời của a và b , đó là một hình e-líp);

– Mô hình (2.4) có thể dùng để dự báo giá trị y nếu biết x tương ứng và ta có thể tìm khoảng tin cậy đối với giá trị đó.

3. Kiểm định giả thuyết

Bằng các lý luận giống như ở trên và để ý đến chương V ta có thể xét các kiểm định giả thuyết về tham số. Chẳng hạn xét bài toán kiểm định:

$$H_0: a = a_0 \text{ với } H_1: a \neq a_0.$$

Thông thường σ^2 chưa biết và dùng tiêu chuẩn:

$$K = \frac{\hat{a} - a_0}{s} \sqrt{\sum_i (x_i - \bar{X})^2},$$

ta có miền tới hạn với mức α (miền đối xứng):

$$B_\alpha = \{K_{tn}: |K_{tn}| > t_{n-2; 1-\alpha/2}\}. \quad (2.19)$$

Bạn đọc có thể tìm các kết quả khác cho trường hợp kiểm định một phía; tương tự cho kiểm định về tham số b và σ^2 .

Thí dụ 2.4. Với $\alpha = 0,05$ hãy kiểm định giả thuyết $H_0: a = 0$ với đối thuyết $H_1: a \neq 0$ (số liệu của thí dụ 2.1).

Giải. Như trong thí dụ 2.3 ở đây ta giả sử có giả thiết chuẩn. Khi đó giá trị ngưỡng của bài toán $t_{6; 0,975} = 2,447$. Tính thống kê thực nghiệm (2.19), dựa vào tiêu chuẩn tương ứng:

$$K_{tn} = \frac{0,0436 - 0}{0,2189} \sqrt{205,38} \approx 2,8542.$$

Do $2,447 < 2,8542$ giả thuyết H_0 bị bác bỏ.

Để ý rằng giả thuyết $H_0: a = 0$ có ý nghĩa rất quan trọng vì nó cho phép chấp nhận hay bác bỏ sự có mặt của biến X trong mô hình đang xét. Ngoài ra ở đây:

$$K = \frac{\hat{a}}{s / \sqrt{\sum_i (x_i - \bar{X})^2}}, \text{ suy ra } K^2 = \frac{\hat{a}^2}{s^2 / \sum_{i=1}^n (x_i - \bar{X})^2}.$$

và do $\frac{\hat{a}^2}{s^2} \sum (x_i - \bar{X})^2 = \frac{(n-2)r^2}{1-r^2} \sim \mathcal{F}(1; n-2)$ ta có thể dùng tiêu chuẩn này và phân phối $\mathcal{F}$ để kiểm định $H_0: a = 0$.

2.4. Hệ số xác định

Để đánh giá sự phù hợp của mô hình tuyến tính người ta sử dụng nhiều cách khác nhau, chẳng hạn dùng phương sai sai số mô hình, khoảng tin cậy của các hệ số dùng các kiểm định tương ứng, hệ số tương quan mẫu gần ± 1 ... Khái niệm hệ số xác định cũng rất có ích để đánh giá chất lượng của mô hình tuyến tính.

Từ công thức (2.13) ta đã thấy:

$$\sum (y_i - \bar{Y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{Y})^2.$$

với vế trái là tổng bình phương độ lệch của y khỏi $\bar{Y}$ (độ lệch tiên nhiệm); hai số hạng vế phải lần lượt là tổng bình phương

độ lệch của y so với hồi quy (hay tổng phương sai dư, tổng bình phương sai số hồi quy) và tổng bình phương sai số cảm sinh bởi hồi quy. Nếu ta đem tổng thứ ba chia cho vế trái thì:

$$r^2 = \frac{\sum (\hat{y}_i - \bar{Y})^2}{\sum (y_i - \bar{Y})^2} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{Y})^2} \quad (2.20)$$

sẽ được gọi *hệ số xác định mẫu* của mô hình hồi quy tuyến tính (2.4), giá trị mà ta đã biết ở §1 như là bình phương của hệ số tương quan. Để ý là nếu $r^2 = 1$, (2.20) sẽ cho ta $\sum (y_i - \hat{y}_i)^2 = 0$ hay trong mọi trường hợp $y_i = \hat{y}_i$ (mô hình chính xác). Nói chung, r^2 cho thấy tỷ lệ tổng bình phương sai số tiên nghiệm được giải thích bởi mô hình tuyến tính (bởi biến X). Để ý rằng từ đó $\sum (y_i - \hat{y})^2$ cho ta phần của tổng tiên nghiệm không được giải thích bởi mô hình tuyến tính.

Cuối cùng ước lượng (không có điều kiện hay tiên nghiệm) của phương sai của Y , như ta biết chính là:

$$S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{Y})^2.$$

Còn
$$\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = S_{y/x}^2$$

là ước lượng có điều kiện của phương sai EY biết giá trị tương ứng $X = x$. Đây cũng là ước lượng tốt nhất của $VY = \sigma^2$ hiểu theo nghĩa không chệch mà ta đã xét trong (2.9) và ký hiệu là s^2 .

2.5. Hồi quy phi tuyến

Nếu ta mô hình hóa quan hệ giữa hai biến X và Y bằng hàm f tùy ý (chẳng hạn $Y = f(X) = P_n(X)$ đa thức cấp $n > 1$), thì việc xác định f sẽ cực kỳ phức tạp (bài toán ước lượng hàm).

Để cho đơn giản, thông thường ta giả sử đã biết dạng hàm, khi đó bài toán đưa về ước lượng các tham số của một hàm đã biết. Thí dụ cho hàm f có dạng đa thức bậc hai:

$$f(x) = a_0 + a_1x + a_2x^2.$$

Việc xác định đường hồi quy phi tuyến mẫu lại dựa vào phương pháp bình phương cực tiểu đã xét ở trên. Ở đây ta đi tìm các ước lượng $\hat{a}_0$, $\hat{a}_1$ và $\hat{a}_2$ làm cực tiểu hàm mục tiêu:

$$Q(a_0, a_1, a_2) = \sum_{i=1}^n (y_i - a_0 - a_1x_i - a_2x_i^2)^2, \quad (2.21)$$

với (x_i, y_i) là các cặp số liệu $(i = \overline{1, n})$. Để tìm các ước lượng trên, ta phải lấy các đạo hàm riêng của $Q(a_0, a_1, a_2)$ trong (2.21) và cho chúng = 0. Khi đó vấn đề đưa về giải một hệ phương trình tuyến tính:

$$\begin{cases} \hat{a}_0 \sum x_i^2 + \hat{a}_1 \sum x_i^3 + \hat{a}_2 \sum x_i^4 = \sum x_i^2 y_i \\ \hat{a}_0 \sum x_i + \hat{a}_1 \sum x_i^2 + \hat{a}_2 \sum x_i^3 = \sum x_i y_i \\ \hat{a}_0 n + \hat{a}_1 \sum x_i + \hat{a}_2 \sum x_i^2 = \sum y_i \end{cases} \quad (2.22)$$

Việc giải hệ (2.22) với 3 phương trình 3 ẩn cũng không quá phức tạp. Tuy nhiên các tính chất thống kê đẹp đẽ của các ước lượng bình phương cực tiểu ở các mục trên sẽ không còn đúng nữa, dù ta có thể đưa vào giả thiết chuẩn của nhiễu ε_i trong mô hình phi tuyến tương tự với dạng (2.4) là $y_i = f(x_i) + \varepsilon_i, i = \overline{1, n}$.

Hoàn toàn tương tự ta có thể ước lượng các tham số cho các dạng hàm phi tuyến khác (chẳng hạn hàm hy-péc-bôn $y = b + a/x$, dạng lũy thừa, dạng lô-ga-rít, dạng mũ...). Phương sai mẫu của sai số mô hình phi tuyến thường được tính theo công thức:

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2,$$

trong đó k là tham số chưa biết của f mà ta cần ước lượng, còn $\hat{f}$ có dạng giống f nhưng các tham số được thay bằng các ước lượng của chúng.

Trong một số trường hợp, ta có thể sử dụng thuật toán hồi quy tuyến tính cho phi tuyến, nhưng tất nhiên trước đó phải làm các biến đổi sơ bộ tương ứng:

Hàm f	Phương trình	Biến đổi sơ bộ	Phương trình sau biến đổi
hy-péc-bôn	$y = \frac{x}{a + bx}$	$z = \frac{1}{y}; t = \frac{1}{x}$	$z = at + b$
mũ	$y = be^{ax}$	$z = \ln y;$	$z = ax + \ln b$
lũy thừa	$y = bx^a$	$z = \ln y; t = \ln x$	$z = at + \ln b$
mũ ngược	$y = be^{a/x}$	$z = \ln y; t = \frac{1}{x}$	$z = at + \ln b$
mũ giả	$y = \frac{1}{b + ae^{-x}}$	$z = \frac{1}{y}; t = e^{-x}$	$z = at + b$

Thí dụ 2.5. Tìm đường hồi quy dạng đa thức bậc hai của y đối với x dựa trên bộ số liệu sau đây:

$y_i \backslash x_i$	1	2	3	4	5	6	n_x
1	2	1	—	—	—	—	3
2	1	2	3	1	—	—	7
3	—	—	1	3	2	—	6
4	—	—	—	1	2	1	4
5	—	—	—	—	2	1	3
6	—	—	—	—	—	1	2
n_y	3	3	4	5	6	3	25

Giải. Ở đây $n = 25$, nhưng để ý có nhiều cặp số liệu xuất hiện nhiều hơn 1 lần. Đường hồi quy mẫu cần tìm có dạng :

$$y = \hat{a}x^2 + \hat{b}x + \hat{c},$$

trong đó $\hat{a}$, $\hat{b}$, $\hat{c}$ thỏa mãn hệ (2.22), nhưng tổng không nên lấy chạy từ 1 đến 25, mà chỉ đến 6, với các điều chỉnh tương ứng:

$$\hat{a} \sum n_{x_i} x_i^4 + \hat{b} \sum n_{x_i} x_i^3 + \hat{c} \sum n_{x_i} x_i^2 = \sum n_{x_i} x_i^2 \bar{y}_{x_i};$$

$$\hat{a} \sum n_{x_i} x_i^3 + \hat{b} \sum n_{x_i} x_i^2 + \hat{c} \sum n_{x_i} x_i = \sum n_{x_i} x_i \bar{y}_{x_i};$$

$$\hat{a} \sum n_{x_i} x_i^2 + \hat{b} \sum n_{x_i} x_i + \hat{c} n = \sum n_{x_i} \bar{y}_{x_i};$$

trong đó $\bar{y}_{x_i}$ là trung bình cộng của các giá trị y_j ứng với x_i cụ thể, chẳng hạn ở đây $\bar{y}_5 = \frac{1}{3}(2.5 + 1.6) = 16/3$. Bảng tính được thiết lập như sau:

x_i	n_{x_i}	$n_{x_i} x_i$	$n_{x_i} x_i^2$	$n_{x_i} x_i^3$	$n_{x_i} x_i^4$	$\bar{y}_{x_i}$	$n_{x_i} \bar{y}_{x_i}$	$n_{x_i} x_i \bar{y}_{x_i}$	$n_{x_i} x_i^2 \bar{y}_{x_i}$
1	3	3	3	3	3	1,33	3,99	3,99	3,99
2	7	14	28	56	112	2,57	17,99	35,98	71,96
3	6	18	54	162	486	4,17	25,02	75,06	225,18
4	4	16	64	256	1024	5,00	20,00	80,00	320,00
5	3	15	75	375	1875	5,33	15,99	79,95	339,75
6	2	12	72	432	2592	5,50	11,00	66,00	396,00
Σ	25	78	296	1284	6092		93,99	340,98	1416,88

Từ đó hệ phương trình trở thành:

$$\begin{cases} 296\hat{a} + 78\hat{b} + 25\hat{c} = 93,99 \\ 1284\hat{a} + 296\hat{b} + 78\hat{c} = 340,98 \\ 6092\hat{a} + 1284\hat{b} + 296\hat{c} = 1416,88. \end{cases}$$

Giải hệ này ta thu được $\hat{a} \approx -0,19$; $\hat{b} \approx 2,21$ và $\hat{c} \approx -0,98$ và đường hồi quy phi tuyến mẫu sẽ là:

$$y = -0,19x^2 + 2,21x - 0,89.$$

§3. HỒI QUY BỘI

3.1. Mô hình hồi quy bội tuyến tính

1. Mô hình

Khi xét đồng thời biến phụ thuộc Y với nhiều biến độc lập $X_1, \dots, X_k$, ta có thể mở rộng mô hình tuyến tính (2.1). Giả sử ta có bộ số liệu có kích thước n ($y_i, x_{i1}, \dots, x_{ik}$), $i = \overline{1, n}$. Ký hiệu X là ma trận số liệu của các biến $X_1, \dots, X_k$.

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2j} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nj} & \cdots & x_{nk} \end{pmatrix},$$

còn y , ε và a và vectơ tương ứng với các biến y_i , ε_i , a_j ($i = \overline{1, n}$; $j = \overline{0, k}$); khi đó mô hình hồi quy bội tuyến tính biểu diễn theo các quan sát sẽ là:

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \dots + a_k x_{ik} + \varepsilon_i, i = \overline{1, n}; \quad (3.1)$$

hay viết gọn dưới dạng vectơ ma trận:

$$y = Xa + \varepsilon.$$

Để ý trong mô hình (3.1) các ε_i là nhiễu trắng thỏa mãn các giả thiết đã xét ở §2; các hệ số $a_0, a_1, \dots, a_k$ là các tham số hồi quy cần ước lượng cùng với phương sai của sai số mô hình σ^2 .

Để cho đơn giản, ta xét trường hợp $k = 2$. Khi đó mặt hồi quy bội sẽ có dạng:

$$y = a_0 + a_1 x_1 + a_2 x_2,$$

là phương trình mô tả một mặt phẳng trong không gian 3 chiều. Mô hình hồi quy tuyến tính bội 2 sẽ có dạng

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + \varepsilon_i, i = \overline{1, n}. \quad (3.2)$$

Nếu dùng các ký hiệu:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

thì mô hình có thể viết gọn lại:

$$y = Xa + \varepsilon.$$

với $E\varepsilon = 0$ (véc tơ không); $V\varepsilon = \sigma^2 I_n$ (I_n ma trận đơn vị cấp n).

2. Ước lượng tham số hồi quy

Ta lại dùng phương pháp bình phương cực tiểu:

$$\min_{a_0, a_1, \dots, a_k} \sum_{i=1}^n (y_i - a_0 - a_1 x_{i1} - \dots - a_k x_{ik})^2. \quad (3.4a)$$

Dưới dạng ma trận, ta có thể viết hàm mục tiêu:

$$\begin{aligned} Q(a) &= (y - Xa)'(y - Xa) = \\ &= y'y - a'X'y - yXa + a'X'Xa. \end{aligned} \quad (3.4b)$$

(dấu t chỉ phép chuyển vị). Như vậy, nếu ký hiệu $\hat{a}$ là véc tơ các ước lượng của $a_0, a_1, \dots, a_k$, ta có ngay (lấy đạo hàm (3.4b) theo véc tơ a và cho bằng 0):

$$-2X'y + 2X'Xa = 0, \quad (3.5)$$

từ đó

$$\hat{a} = (X'X)^{-1}X'y. \quad (3.6)$$

Để ý trong (3.6) giả sử ma trận $X'X$ không suy biến. Trong thực hành khi $k = 2$ việc tính các ước lượng $\hat{a}_0, \hat{a}_1$ và $\hat{a}_2$ đưa về giải hệ phương trình đại số tuyến tính (3.5) gồm 3 phương trình 3 ẩn số khá đơn giản. Bằng các phương pháp số việc giải các hệ như vậy không đặt ra nhiều khó khăn lớn. Ở đây (3.5) sẽ có dạng cụ thể:

$$\begin{cases} a_0 \sum x_{i2} + a_1 \sum x_{i1}x_{i2} + a_2 \sum x_{i2}^2 = \sum x_{i2}y_i, \\ a_0 \sum x_{i1} + a_1 \sum x_{i1}^2 + a_2 \sum x_{i1}x_{i2} = \sum x_{i1}y_i, \\ a_0 n + a_1 \sum x_{i1} + a_2 \sum x_{i2} = \sum y_i. \end{cases}$$

với các tổng lấy theo i từ 1 đến n .

3. Các tính chất của ước lượng bình phương cực tiểu

Bạn đọc có thể tự chứng minh các tính chất sau:

- (i) $\hat{a}$ là ước lượng không chệch của véc tơ tham số a :
- (ii) $V\hat{a} = \sigma^2(X'X)^{-1}$.
- (iii) $\hat{a}$ và $\hat{\varepsilon} = y - \hat{y} = y - X\hat{a}$ không tương quan.
- (iv) Ước lượng không chệch của σ^2 sẽ là:

$$s^2 = \frac{1}{n - k - 1} \sum_{i=1}^n \hat{\varepsilon}_i^2. \quad (3.7)$$

4. Trường hợp có giả thiết chuẩn

Mô hình (3.1) với giả thiết $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ có nhiều tính chất thống kê khá tốt (xem tiết trước). Chẳng hạn ước lượng hợp lý nhất của a sẽ trùng với $\hat{a}$ được xác định trong (3.6); còn ước lượng hợp lý nhất của σ^2 sẽ có dạng (so sánh (3.7) ở trên):

$$\frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2.$$

Cũng do có giả thiết chuẩn của ε nên:

$$\begin{aligned} Y &\sim \mathcal{N}(Xa, \sigma^2 I_n) \text{ (} Y \text{ là véc tơ cột gồm các } y_i \text{);} \\ \hat{a} &\sim \mathcal{N}(a, \sigma^2(X'X)^{-1}); \end{aligned} \quad (3.8)$$

$$(n - k - 1) \frac{s^2}{\sigma^2} \sim \chi^2(n - k - 1). \quad (3.9)$$

Ngoài ra, $\hat{a}$ còn là ước lượng hiệu quả của a . Các kết quả trên cho phép ta xác định các khoảng tin cậy hoặc làm các kiểm định giả thuyết tương ứng.

Chú ý, do (3.8) nên:

$$\frac{(\hat{a} - a)' X' X (\hat{a} - a)}{\sigma^2} \sim \chi^2(k + 1),$$

và do tính chất (iii) ở trên, có dễ ý đến (3.9):

$$\frac{(\hat{a} - a)' X' X (\hat{a} - a)}{(k + 1)s^2} \sim \mathcal{F}(k + 1, n - k - 1).$$

Từ đó ta có thể kiểm định đồng thời nhiều giả thuyết đơn dạng

$$H_0: a = a^{(0)} \quad (H_1: a \neq a^{(0)}).$$

Khi đó miền tới hạn của quy tắc kiểm định sẽ là:

$$B_a = \left\{ K_{\alpha} : K_{\alpha} \frac{(\hat{a} - a)^{(0)'} X' X (\hat{a} - a)^{(0)}}{(k + 1)s^2} > F_{k+1, n-k-1, 1-\alpha} \right\}.$$

3.2. Tương quan bội và tương quan riêng

1. Tương quan riêng

Khi xét đồng thời 3 biến X_1 , X_2 , và Y ta có thể sử dụng các (hệ số) tương quan mẫu:

$$r_{x_1 x_2}; r_{x_1 y}; r_{x_2 y}. \quad (3.10)$$

Tuy nhiên, $r_{x_1 y}$ chẳng hạn khi mô tả quan hệ giữa hai biến X_1 và Y , rõ ràng quan hệ đó không chỉ phụ thuộc vào bản thân X_1 và Y , mà còn bị ảnh hưởng bởi tác động của biến số thứ ba là X_2 . Vì vậy để loại trừ ảnh hưởng đó, người ta đưa ra khái niệm *hệ số tương quan riêng mẫu*, ký hiệu là:

$$r_{x_1 x_2 \cdot y}; r_{x_1 y \cdot x_2}; r_{x_2 y \cdot x_1}. \quad (3.11)$$

Khái niệm này dễ dàng mở rộng cho trường hợp có nhiều hơn 3 biến. Công thức tính hệ số tương quan riêng (3.11) theo các hệ số tương quan đơn (3.10) có dạng:

$$r_{x_1 x_2 \cdot y} = \frac{r_{x_1 x_2} - r_{x_1 y} \cdot r_{x_2 y}}{\sqrt{(1 - r_{x_1 y}^2)(1 - r_{x_2 y}^2)}}.$$

Do tính đối xứng của ba biến nên bằng cách thay đổi vị trí của chúng, bạn đọc dễ dàng tìm được hai công thức còn lại. Để ý hệ

số tương quan riêng cũng có tính chất chỉ nhận giá trị từ -1 đến $+1$. Ta hoàn toàn có thể định nghĩa hệ số xác định riêng giống như ở §1.

2. Tương quan bội

Khái niệm *hệ số tương quan bội* được đưa vào để đo mối phụ thuộc giữa một biến nào đó với tập các biến khác. Ở đây ta có thể xác định hệ số tương quan bội mẫu:

$$r_{y.x_1x_2}; r_{x_1.yx_2}; r_{x_2.yx_1}, \quad (3.12a)$$

và các hệ số xác định bội mẫu tương ứng:

$$\beta_{y.x_1x_2}; \beta_{x_1.yx_2}; \beta_{x_2.yx_1}. \quad (3.12b)$$

Rõ ràng ta luôn có $0 \leq \beta_{y.x_1x_2} \leq 1$ và $-1 \leq r_{y.x_1x_2} \leq 1$. Khi $|r_{y.x_1x_2}|$ càng gần 1, biến Y càng có tương quan chặt (gần với tuyến tính bội) với cặp biến X_1 và X_2 . Có thể chứng minh:

$$\begin{aligned} \beta_{y.x_1x_2} &= \\ &= \frac{1}{\sum_{i=1}^n (y_i - \bar{Y})^2} \left[\hat{a}_1 \sum_{i=1}^n (x_{i1} - \bar{X}_1)(y_i - \bar{Y}) + \hat{a}_2 \sum_{i=1}^n (x_{i2} - \bar{X}_2)(y_i - \bar{Y}) \right]. \end{aligned}$$

Trong thực hành người ta hay dùng công thức sau:

$$\beta_{y.x_1x_2} = \frac{r_{x_1y}^2 + r_{x_2y}^2 - 2r_{x_1x_2}r_{x_1y}r_{x_2y}}{1 - r_{x_1x_2}^2}.$$

Như ở §2 $\beta_{y.x_1x_2}$ cho ta tỷ lệ của tổng bình phương sai số được giải thích bởi mô hình hồi quy bội đã chọn. Khái niệm hệ số tương quan bội là tổng quát hóa của tương quan đơn đã xét từ trước đến nay.

BÀI TẬP

1. Khảo sát chi phí sản xuất (X) và sản lượng (Y) của 10 công ty cùng loại ta có bộ số liệu:

STT công ty	Chi phí (triệu đồng)	Sản lượng (nghìn tấn)
1	150	40
2	140	38
3	160	48
4	170	56
5	150	62
6	162	75
7	180	70
8	190	110
9	165	90
10	185	120

- a) Xây dựng đường hồi quy tuyến tính mẫu.
 b) Đánh giá sự phù hợp của mô hình tuyến tính đối với bộ số liệu.
 c) Xác định hệ số xác định mẫu và cho biết ý nghĩa.
2. Khảo sát hai biến ngẫu nhiên, ta thu được kết quả:

$y_i \backslash x_i$	1	3	5	7	9
10	4	—	—	—	—
15	7	10	—	—	—
20	—	15	26	10	2
25	—	—	35	8	5
30	—	—	3	18	6
35	—	—	—	6	1

- a) Đánh giá mức độ phụ thuộc của hai biến trên.
- b) Xây dựng đường hồi quy tuyến tính mẫu (của y theo x).
3. Một hãng quảng cáo nhận thấy có mối liên hệ giữa ngân sách quảng cáo (Y) và doanh số của các công ty (X). Điều tra 8 công ty người ta thu được:

STT công ty	Chi phí (tỷ đồng)	Ngân sách quảng cáo (triệu đồng)
1	6	45
2	7	80
3	9	70
4	9	85
5	7	60
6	8	55
7	6	75
8	12	90

- a) Xác định đường hồi quy tuyến tính mẫu.
- b) Đánh giá sự phù hợp của mô hình tuyến tính đã chọn.
- c) Tìm khoảng tin cậy 95% cho hệ số góc của đường hồi quy. Có thể cho rằng hệ số đó khác không đáng kể không?
4. Nghiên cứu về lượng prô-tê-in chứa trong hạt lúa mỳ và năng suất lúa trên 10 thửa ruộng cho ta kết quả:

Năng suất (x_i)	9,9	10,2	11,0	11,6	11,8	12,5	12,8	13,5	14,3	14,4
Tỷ lệ prôtêin (y_i)	10,7	10,8	12,1	12,5	12,2	12,8	12,4	11,8	11,8	12,6

- a) Xác định đường hồi quy tuyến tính của y theo x ; sau đó của x theo y . Bạn có nhận xét gì về hai đường hồi quy đó.
- b) Có nên dùng mô hình phi tuyến không? Tại sao?

5. Trong một nghiên cứu về tai nạn giao thông, người ta đã thống kê giá trị thiệt hại (y_i) và tốc độ va chạm của phương tiện (đã quy chuẩn, kí hiệu x_i):

x_i	1	6	11	16	2	7	12	17	3	8
y_i	41	61	89	129	44	66	94	134	48	70

x_i	13	18	4	9	14	19	5	10	15
y	96	142	50	75	106	147	58	81	118

Xác định đường hồi quy tuyến tính mẫu. Bạn có nhận xét gì về mô hình đó và có ý kiến gì để cải tiến mô hình?

6. Nghiên cứu độ bền x_i của các dây kim loại có đường kính y_i , người ta thu được các số liệu:

x_i	0,6	2	2,2	2,45	2,6
y_i	500	560	690	760	900

Giả sử giữa y và x có liên hệ dạng đa thức bậc hai, hãy xây dựng đường hồi quy thực nghiệm.

7. Số liệu điều tra về tỷ lệ cơ giới hóa (x_i) và giá trị một đơn vị sản phẩm (y_i) như sau:

$x_i \backslash y_i$	1,5 – 2,1	2,1 – 2,7	2,7 – 3,3	3,3 – 3,9	3,9 – 4,5
50 – 60	–	–	1	1	1
60 – 70	1	4	1	–	–
70 – 80	3	6	1	–	–
80 – 90	6	3	–	–	–
90 – 100	10	3	3	–	–

Tìm đường hồi quy phi tuyến thực nghiệm dạng $y = \frac{a}{x} + b$

và đánh giá sai số mô hình.

8. Khảo sát nhiệt độ một phản ứng hóa học (y) cùng với nồng độ của bốn hóa chất khác nhau ($x_i, i = \overline{1,4}$) ta có số liệu:

y	x	x_2	x_3	x_4
78,5	7	26	6	60
74,5	1	29	15	52
104,3	11	56	8	20
87,6	11	31	8	47
95,9	7	52	6	33
109,2	11	55	9	22
102,7	3	71	17	6
72,5	1	31	22	44
93,1	2	54	18	22
115,9	21	47	4	26
83,8	1	40	23	34
113,3	11	66	9	12
109,4	10	68	8	12

- Tính ma trận tương quan mẫu và bình luận kết quả.
- Xây dựng các mô hình hồi quy bội 2 (tuyến tính) thực nghiệm và so sánh kết quả trên phương diện sai số mô hình và hệ số tương quan bội và riêng.
- Mô hình hồi quy bội 3 có tốt hơn mô hình bội 2 hay không? Tại sao?
- Xây dựng mô hình hồi quy mẫu bội 4 và đánh giá tính phù hợp của mô hình. Theo bạn mô hình bội cấp cao có tốt hơn không? Tại sao?

Phụ lục

CÁC BẢNG SỐ

1. Bảng hàm Gao-xơ $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

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	9653	3637	3621	3605	3589	3572	3555	3538
0.5	3521	3503	3485	3467	3448	3929	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	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	1219	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

1. Bảng hàm Gao-xơ $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ (tiếp theo)

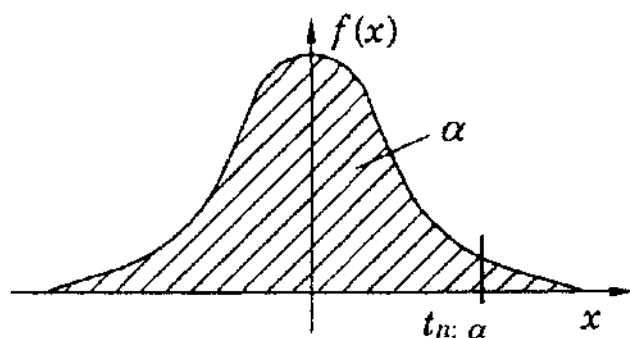
2.0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2.1	0440	0431	0422	0413	0404	0396	0388	0379	0371	0363
2.2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2.3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2.4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2.5	0175	0171	0167	0163	0158	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	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3.1	0033	0032	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

2 Bảng hàm Láp-la-xơ $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$

x	0	1	2	3	4	5	6	7	8	9
0.0	0,0000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0.1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0.2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0.3	11791	12172	12556	12930	13307	13683	14058	14431	14803	15173
0.4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0.5	19146	19497	19847	20194	20194	20884	21226	21566	21904	22240
0.6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0.7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0.8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0.9	31594	31859	32121	32881	32639	32894	33147	33398	33646	33891
1.0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214
1.1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1.2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1.3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774
1.4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1.5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1.6	44520	44630	44738	44815	44950	45053	45154	45254	45352	45449
1.7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1.8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1.9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2.0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2.1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2.2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2.3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2.4	49180	49202	49224	49245	49266	49285	49305	49324	49343	49361
2.5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2.6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2.7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2.8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2.9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3.0	0,49865		3,1	49903	3,2	49931	3,3	49952	3,4	49966
3.5	49977		3,6	49984	3,7	49989	3,8	49993	3,9	49995
4.0	499968									
4.5	499997									
5.0	49999997									

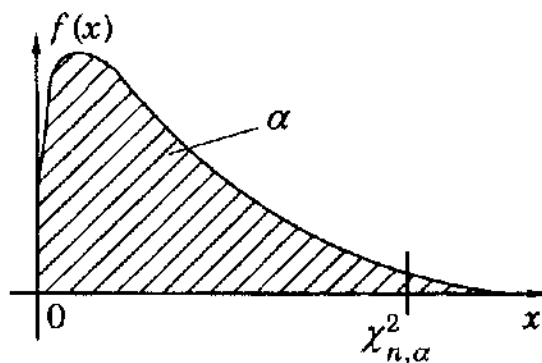
3. Bảng phân vị Stiu-đơn $P(X < t_{n, \alpha}) = \alpha$ với $X \sim t(n)$

n	α					
	0,90	0,95	0,975	0,99	0,995	0,9995
1	3.078	6.314	12.706	31.820	63.526	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	1.796	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	2.074	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
∞	1.282	1.645	1.960	2.326	2.576	3.291



4. Bảng phân vị χ^2 $P(X < \chi^2_{n,\alpha}) = \alpha$ với $X \sim \chi^2(n)$

n	α							
	0.005	0.001	0.025	0.05	0.95	0.975	0.99	0.995
1	0.0000393	0.000157	0.000982	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	32.801
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



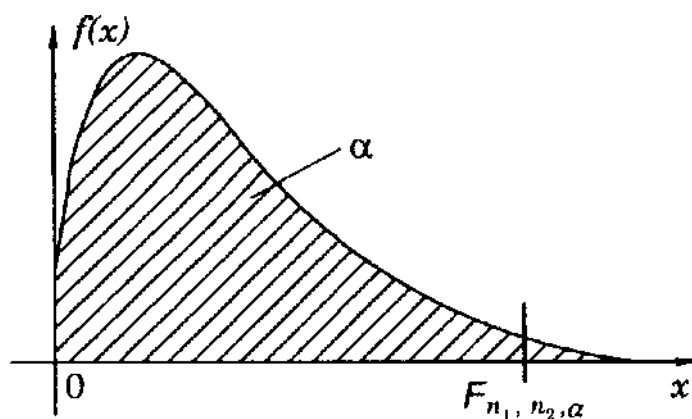
5. Bảng phân vị Phi-sơ

$$P(X < F_{n_1, n_2, \alpha}) = \alpha = 0,95 \text{ với } X \sim \mathcal{F}(n_1, n_2)$$

n_2	n_1								
	1	2	3	4	5	6	7	8	9
1	161.14	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
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

5. Bảng phân vị Phi-sơ (tiếp theo)

n_1									
10	12	15	20	24	30	40	60	120	∞
241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	9.49	19.50
8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
3.65	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.578	2.54
2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
2.60	2.53	2.46	2.39	2.35	2.31	2.34	2.30	2.25	2.21
2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.10	2.06	2.02
2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00



6. Hướng dẫn sử dụng các bảng số

1. Nếu muốn tính $\phi(1,25)$, dóng hàng "1,2" và cột "5" ta thấy 1826, suy ra $\phi(1,25) = 0,1826$. Với các giá trị không có trong bảng ($x > 3,9$) coi $\phi(x) = 0$. Chú ý $\phi(\cdot)$ là hàm chẵn $\phi(-x) = \phi(x)$.

2. Việc tra bảng tính $\phi(x)$ làm giống như trên. Chẳng hạn nếu muốn tính $\phi(1,25)$, dóng hàng "1,2" và cột "5" ta gặp 39435 nên $\phi(1,25) = 0,39425$. Với $x > 5$ coi $\phi(x) = 0,5$. Chú ý rằng $\phi(x)$ là hàm lẻ $\phi(-x) = -\phi(x)$.

Ý nghĩa xác suất của $\phi(x)$ là rõ ràng: nếu $X \sim \mathcal{N}(0, 1)$ thì $\phi(x_0) = P(0 < X < x_0)$. Ngoài ra hàm $F(x) = \phi(x) + 0,5$ sẽ làm hàm phân phối xác suất của X nói trên, tức là: ($x_0 > 0$)

$$P(|X| < x_0) = P(-x_0 < X < x_0) = 2\phi(x_0),$$

$$P(X < x_0) = \phi(x_0) + 0,5 = F(x_0),$$

$$P(X < -x_0) = 0,5 - \phi(x_0),$$

$$P(X > x_0) = 0,5 - \phi(x_0) \dots$$

Nếu $X \sim \mathcal{N}(a, \sigma^2)$, nên làm phép biến đổi $Y = \frac{X - a}{\sigma}$ và việc tra bảng đối với biến X chuyển thành đối với biến $Y \sim \mathcal{N}(0, 1)$.

Nếu đã biết giá trị $\phi(x_0)$, muốn tìm lại x_0 , quá trình tra bảng ngược lại với bên trên.

3. Để tìm giá trị $t_{n,\alpha}$ sao cho $P(X < t_{n,\alpha}) = \alpha$ biết rằng $X \sim t(n)$ việc tra bảng cũng đơn giản: dóng hàng " n " và cột " α " tương ứng (chẳng hạn $t(8; 0,95) = 1,860$). Việc tìm $\theta_b > 0$ sao cho $P(X > \theta_b) = \alpha$ với $X \sim t(n)$ tương đương với việc tra bảng tìm $\theta_b = t_{n, 1-\alpha}$. Do tính đối xứng, nếu muốn tìm $\theta_b < 0$ sao cho $P(X > \theta_b) = \alpha$, ta tra bảng tìm $t_{n, 1-\alpha}$, sau đó $\theta_b = -t_{n, 1-\alpha}$.

Trong ba tham số n , α và $t_{n,\alpha}$ nếu biết hai ta có thể tìm được tham số thứ 3.

Cuối cùng, nếu $n > 30$, thay vì tìm $\theta_b = t_{n, 1-\alpha}$ ta sẽ tìm θ_b sao cho $\phi(\theta_b) = \frac{1-\alpha}{2}$ từ bảng Láp-la-xơ.

4. Để tìm giá trị $\chi^2_{n,\alpha}$ sao cho $P(X < \chi^2_{n,\alpha}) = \alpha$ biết rằng $X \sim \chi^2(n)$ ta làm giống phần 3: dóng hàng “ n ” và cột “ α ”.

5. Việc tra bảng tìm giá trị $F(n_1, n_2; 0,95)$ cũng đơn giản: dóng cột “ n_1 ” và hàng “ n_2 ”.

TÀI LIỆU THAM KHẢO

1. **Barnes J.W.** *Statistical analysis for engineers and scientists*. McGraw – Hill, 1994.
2. **Cramer H.** *Mathematical methods of statistics*. Princeton Univ. Press, Princeton, NJ, 1946.
3. **Feller W.** *An introduction to probability theory and its applications*. John Wiley & Sons, NY, vol. 1, 1950; vol. 2. 1966.
4. **Gnedenko B.V.** *Giáo trình lý thuyết xác suất*. “Khoa học”, Moskva, 1965 (tiếng Nga).
5. **Hald A.** *Statistical theory with engineering applications*. John Wiley & Sons, NY, 1966.
6. **Kirkwood B.R.** *Essentials of medical statistics*. Blackwell Scient. Publ., 1988.
7. **Monfort A.** *Cours de probabilités*. Economica, Paris, 1980.
8. **Monfort A.** *Cours de statistique mathématique*. Economica, Paris, 1982.
9. **Sanders D.H. and F. Allard.** *Statistics: A fresh approach*. McGraw – Hill, 1990.
10. **Tassi F.** *Méthodes statistiques*. Economica, Paris, 1989.
11. **Trần Tuấn Diệu, Lý Hoàng Tú.** *Giáo trình lý thuyết xác suất và thống kê toán học*, NXB Đại học và THCN, Hà Nội, 1977.

MỤC LỤC

Lời nói đầu	3
CHƯƠNG I. SỰ NGẪU NHIÊN VÀ PHÉP TÍNH XÁC SUẤT	5
§1. Khái niệm mở đầu	5
1.1. Sự kiện ngẫu nhiên	5
1.2. Phép toán và quan hệ của các sự kiện	6
1.3. Giải tích kết hợp	9
§2. Các định nghĩa của xác suất.....	11
2.1. Định nghĩa cổ điển	11
2.2. Định nghĩa thống kê	14
2.3. Định nghĩa tiên đề	16
§3. Xác suất có điều kiện	18
3.1. Khái niệm	18
3.2. Công thức cộng và nhân xác suất.....	20
3.3. Công thức Béc-nu-li.....	26
§4. Công thức Bay-ét	29
4.1. Khái niệm nhóm đầy đủ	29
4.2. Công thức xác suất đầy đủ.....	30
4.3. Công thức Bay-ét.....	31
Bài tập	35
CHƯƠNG II. BIẾN NGẪU NHIÊN VÀ LUẬT PHÂN PHỐI XÁC SUẤT	39
§1. Khái niệm biến ngẫu nhiên	39
1.1. Khái niệm	39
1.2. Phân loại	40
§2. Luật phân phối xác suất.....	40
2.1. Bảng phân phối xác suất và hàm xác suất	40
2.2. Hàm phân phối xác suất.....	43
2.3. Hàm mật độ xác suất	45

§3. Các số đặc trưng của biến ngẫu nhiên	48
3.1. Kỳ vọng	48
3.2. Phương sai	51
3.3. Một số đặc số khác.....	54
§4. Một số phân phối thông dụng	56
4.1. Phân phối đều.....	56
4.2. Phân phối nhị thức.....	57
4.3. Phân phối Poa-xông	60
4.4. Các phân phối rời rạc khác	61
4.5. Phân phối chuẩn.....	65
4.6. Các phân phối liên tục khác	70
Bài tập	76
CHƯƠNG III. BIẾN NGẪU NHIÊN NHIỀU CHIỀU	79
§1. Luật phân phối của biến ngẫu nhiên nhiều chiều.....	79
1.1. Các khái niệm cơ sở.....	79
1.2. Phân phối xác suất của biến ngẫu nhiên hai chiều rời rạc	81
1.3. Phân phối xác suất của biến ngẫu nhiên hai chiều liên tục.....	84
§2. Các số đặc trưng của biến ngẫu nhiên hai chiều.....	89
2.1. Các số đặc trưng của các biến thành phần	89
2.2. Hiệp phương sai và hệ số tương quan	90
2.3. Các số đặc trưng có điều kiện	93
2.4. Phân phối chuẩn hai chiều	94
§3. Hàm của các biến ngẫu nhiên	96
3.1. Hàm của một biến ngẫu nhiên	96
3.2. Hàm của hai biến ngẫu nhiên	98
3.3. Các số đặc trưng của hàm của các biến ngẫu nhiên..	102
§4. Các định lý giới hạn và luật số lớn	103
4.1. Sự hội tụ của dãy biến ngẫu nhiên.....	103
4.2. Các định lý giới hạn	105
4.3. Luật số lớn	107
Bài tập	110

CHƯƠNG IV. MẪU THỐNG KÊ VÀ ƯỚC LƯỢNG THAM SỐ.....	113
§1. Mẫu và thống kê mô tả.....	113
1.1. Mẫu và tập đám đông	113
1.2. Vấn đề chọn mẫu.....	114
1.3. Phân loại và mô tả số liệu mẫu	116
§2. Mẫu ngẫu nhiên và các đặc trưng mẫu.....	121
2.1. Mẫu ngẫu nhiên từ một tập nền	121
2.2. Các đặc trưng mẫu	123
2.3. Vấn đề tính toán các dạng đặc trưng mẫu.....	128
§3. Ước lượng điểm	133
3.1. Ước lượng tham số.....	133
3.2. Các tính chất của ước lượng điểm	134
3.3. Các phương pháp ước lượng.....	136
§4. Khoảng tin cậy.....	140
4.1. Ước lượng khoảng.....	140
4.2. Khoảng tin cậy cho kỳ vọng.....	141
4.3. Khoảng tin cậy cho tỷ lệ.....	146
4.4. Khoảng tin cậy cho phương sai.....	150
Bài tập	153
CHƯƠNG V. KIỂM ĐỊNH GIẢ THUYẾT	158
§1. Giả thuyết thống kê và quy tắc kiểm định	158
1.1. Giả thuyết thống kê	158
1.2. Quy tắc kiểm định giả thuyết.....	159
1.3. Các dạng miền tới hạn	162
§2. Các kiểm định dùng một mẫu	163
2.1. Kiểm định về kỳ vọng	163
2.2. Kiểm định về tỷ lệ	166
2.3. Kiểm định về phương sai	168
§3. Các kiểm định dùng nhiều mẫu	170
3.1. So sánh hai kỳ vọng	170
3.2. So sánh hai tỷ lệ.....	172
3.3. So sánh hai phương sai	174
3.4. So sánh nhiều trung bình (phân tích phương sai)	176

§4. Kiểm định phi tham số	179
4.1. Kiểm định giả thiết về luật phân phối.....	179
4.2. Kiểm định giả thuyết độc lập	184
Bài tập	188
CHƯƠNG VI. PHÂN TÍCH HỒI QUY	194
§1. Phân tích tương quan	194
1.1. Hiệp phương sai và hệ số tương quan.....	194
1.2. Hệ số tương quan mẫu	195
1.3. Tiêu chuẩn độc lập của hai biến ngẫu nhiên.....	200
1.4. Kiểm định giả thuyết về hệ số tương quan.....	203
§2. Hồi quy	204
2.1. Mô hình tuyến tính	204
2.2. Ước lượng hệ số hồi quy	206
2.3. Trường hợp có giả thiết chuẩn.....	213
2.4. Hệ số xác định	216
2.5. Hồi quy phi tuyến.....	217
§3. Hồi quy bội	221
3.1. Mô hình hồi quy bội tuyến tính.....	221
3.2. Tương quan bội và tương quan riêng	224
Bài tập	227
PHỤ LỤC. CÁC BẢNG SỐ	230
1. Bảng hàm Gao-xơ $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$	230
2. Bảng hàm Láp-la-xơ $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$	232
3. Bảng phân vị Stiu-đơn $P(X < t_{n,\alpha}) = \alpha$ với $X \sim t(n)$	233
4. Bảng phân vị χ^2 $P(X < \chi^2_{n,\alpha}) = \alpha$ với $X \sim \chi^2(n)$	234
5. Bảng phân vị Phi-sơ	235
6. Hướng dẫn sử dụng các bảng số	237
Tài liệu tham khảo	239