MỤC LỤC

LÍ THUYẾT XÁC SUẤT

VÀ

THỐNG KÊ TOÁN

NGUYỄN NGỌC CỪ

Sách dùng cho sinh viên trường Đại học xây dựng
và sinh viên các trường Đại học kĩ thuật

Chương 1

Biến cố ngẫu nhiên và xác suất

1.1 Biến cố ngẫu nhiên và các phép toán

1.1.1 Phép thử ngẫu nhiên

Mọi vấn đề về lí thuyết xác suất đều liên quan tới một phép thử ngẫu nhiên nào đó. *Phép thử ngẫu nhiên* có thể hiểu như là một thí nghiệm hoặc sự quan sát một hiện tượng tự nhiên nào đó mà kết quả của thí nghiệm hoặc quan sát đó không phụ thuộc vào ý muốn chủ quan của chúng ta. Nói một cách khác phép thử ngẫu nhiên là một thí nghiệm hoặc sự quan sát một hiện tượng tự nhiên dưới những điều kiện nhất định mà kết quả của nó không được xác định duy nhất.

Các ví dụ về phép thử ngẫu nhiên như: gieo đồng xu để quan sát mặt sấp hay mặt ngửa xuất hiện, gieo xúc xắc, quan sát số lượng xe cộ qua lại ở một nút giao thông nào đó trong một khoảng thời gian xác định, số lượng phế phẩm khi chọn một lượng nhất định các sản phẩm từ một lô hàng nào đó để kiểm tra chất lượng sản phẩm, quan sát mực nước một con sông tại một địa điểm vào một thời điểm xác định...

Mỗi phép thử ngẫu nhiên có thể có hữu hạn hoặc vô hạn các kết quả có thể xảy ra. Ví dụ khi gieo xúc xắc, 6 khả năng có thể xảy ra: mặt một chấm, mặt 2 chấm,... mặt 6 chấm xuất hiện. Khi phải chọn 3 người trong một trong một tổ gồm 10 người để làm một công việc nào đó (ta tiến hành cách chọn như sau: làm 10 phiếu đánh số từ 1 đến 10 tương ứng với tên của

10 người và chọn ngẫu nhiên ra 3 phiếu). Có thể coi đây là phép thử ngẫu nhiên với

$$C_{10}^{3} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$$

khả năng có thể xảy ra.

Nếu phép thử ngẫu nhiên là đo mực nước sông ta có thể coi có bao nhiêu khả năng khác nhau về mực nước sông là bấy nhiêu kết quả có thể xảy ra của phép thử ngẫu nhiên, hay có thể nói kết quả là bất cứ số dương nào từ giá trị m đến giới hạn trên M. (Trong thực tế người ta đo mức nước sông chỉ cần chính xác đến milimét, tuy nhiên về mặt lí thuyết có thể coi mọi số thực từ m đến M là kết quả có thể xảy ra của phép thử ngẫu nhiên).

Mỗi kết quả có thể xảy ra ở một phép thử ngẫu nhiên nào đó được gọi là biến cố ngẫu nhiên cơ bản. Ví dụ khi gieo xúc xắc, ta có 6 biến cố ngẫu nhiên cơ bản: mặt 1 chấm, mặt 2 chấm, ... hoặc mặt 6 chấm xuất hiện. Trong mục sau ta sẽ đưa ra khái niệm về biến cố ngẫu nhiên. Ví dụ khi gieo xúc xắc, mặt chẵn xuất hiện là một biến cố ngẫu nhiên, tuy nhiên nó không phải là biến cố ngẫu nhiên cơ bản. Biến cố ngẫu nhiên cơ bản phải là kết quả nhỏ nhất, không thể phân chia được có thể xảy ra ở một phép thử ngẫu nhiên nào đó. Một ví dụ khác đã nói ở trên: chọn ngẫu nhiên 3 người trong một tổ gồm 10 người để làm một công việc nào đó, ta có 120 biến cố ngẫu nhiên cơ bản và ở phép thử ngẫu nhiên đo mực nước sông ta có vô hạn biến cố ngẫu nhiên cơ bản. Các biến cố ngẫu nhiên cơ bản thường được kí hiệu là ω và tập hợp toàn bộ các biến cố ngẫu nhiên cơ bản được gọi là *không gian các biến cố ngẫu nhiên cơ bản*, kí hiệu là Ω.

1.1.2 Biến cố ngẫu nhiên

Khi tiến hành một phép thử ngẫu nhiên, ngoài các biến cố ngẫu nhiên cơ bản, chúng ta còn đưa vào các khái niệm biến cố ngẫu nhiên khác. Ví dụ khi gieo xúc xắc, mặt chẵn xuất hiện , khi đo mực nước sông Hồng kết quả lớn hơn 11 mét hoặc tỉ lệ phế phẩm khi chọn các sản phẩm từ một lô hàng nào đó để kiểm tra chất lượng sản phẩm nằm trong khoảng từ 2 đến 3%. Các biến cố như vậy có thể coi như một tập hợp con nào đó gồm các biến cố ngẫu nhiên cơ bản của không gian các biến cố Ω. Ví dụ khi gieo xúc

xắc nếu ta kí hiệu $\omega_1, \omega_2, ..., \omega_6$ là các biến cố: mặt 1 chấm, mặt 2 chấm, ..., mặt 6 chấm xuất hiện. Khi đó biến cố mặt chẵn xuất hiện có thể đặt tương ứng với tập con $\{\omega_2, \omega_4, \omega_6\}$ của không gian các biến cố ngẫu nhiên cơ bản $\Omega = \{\omega_1, \omega_2, ..., \omega_6\}$.

Ta nói một biến cố ngẫu nhiên xảy ra (hay xuất hiện) nếu kết quả của phép thử ngẫu nhiên - biến cố ngẫu nhiên cơ bản - là phần tử của tập hợp tương ứng với biến cố ngẫu nhiên đó. Ví dụ khi gieo xúc xắc nếu kết quả là mặt 2 chấm, khi đó ta nói mặt chẵn xuất hiện. Tất nhiên nếu kết quả là mặt 4 chấm (hoặc mặt 6 chấm) ta cũng nói mặt chẵn xuất hiện.

Như vậy về mặt toán học biến cố ngẫu nhiên thực chất là một tập con của không gian các biến cố ngẫu nhiên Ω. Người ta thường kí hiệu các biến cố ngẫu nhiên (từ nay ta sẽ nói gọn là biến cố) là các chữ in hoa $A, B, C...$ Các biến cố ngẫu nhiên cơ bản do vậy cũng được coi là các biến cố, chúng là các tập con gồm một phần tử của không gian các biến cố Ω. Trong số các tập con của Ω có hai tập hợp đặc biệt, đó chính là Ω và tập trống $\emptyset$. Tập Ω được gọi là *biến cố chắc chắn xảy ra* vì trong mọi trường hợp kết quả của phép thử ngẫu nhiên - biến cố ngẫu nhiên cơ bản - luôn luôn là phần tử của Ω, còn tập trống $\emptyset$ được gọi là *biến cố không thể xảy ra.*

1.1.3 Quan hệ và các phép toán giữa các biến cố ngẫu nhiên

Do biến cố ngẫu nhiên có thể đồng nhất với một tập hợp, nên ta có thể đưa vào các phép toán cũng như mối quan hệ giữa các biến cố như các phép toán của các tập hợp.

1. Cho hai biến cố A và B. Ta nói biến cố A kéo theo biến cố B nếu A xuất hiện kéo theo B cũng xuất hiện. Điều này có nghĩa là mọi phần tử ω của tập A - biến cố ngẫu nhiên cơ bản thuộc A - cũng là phần tử của B, nói cách khác A là tập con của B. Chính vì vậy ta kí hiệu $A \Rightarrow$ B (hoặc $A \subset B$) để diễn tả mối quan hệ biến cố A kéo theo biến cố B.

   Ví dụ khi gieo xúc xắc biến cố mặt 6 chấm xuất hiện kéo theo biến cố mặt chẵn xuất hiện.

2. Tổng của hai biến cố A và B là biến cố (kí hiệu $A + B$) xuất hiện khi

và chỉ khi ít nhất một trong hai biến cố A hoặc B xuất hiện. Về quan điểm tập hợp $A+B$ chính là hợp của hai tập hợp A và B, vì một phần tử bất kì - biến cố ngẫu nhiên cơ bản - của $A+B$ phải là một phần tử của A hoặc của B. Do vậy ta cũng kí hiệu tổng của hai biến cố A và B là $A \cup B$.

Ví dụ một xạ thủ dự kiểm tra sát hạch bằng việc bắn hai phát đạn vào bia, xạ thủ đó bắn không đạt yêu cầu nếu ít nhất một lần bắn trượt. Gọi A là biến cố lần bắn đầu xạ thủ đó bắn trượt, B là biến cố lần bắn thứ hai xạ thủ đó bắn trượt. Khi đó $A+B$ là biến cố xạ thủ đó bắn không đạt yêu cầu.

3. Tích của hai biến cố A và B là biến cố (kí hiệu AB) xuất hiện khi và chỉ khi cả hai biến cố A, B đồng thời xuất hiện. Về quan điểm tập hợp AB chính là giao của hai tập hợp A và B, vì một phần tử bất kì - biến cố ngẫu nhiên cơ bản - của AB phải là phần tử của cả A và B. Do vậy ta cũng kí hiệu tích của hai biến cố A và B là $A \cap B$.

   Ví dụ khi gieo xúc xắc, gọi A là biến cố mặt chẵn xuất hiện và B là biến cố mặt xuất hiện có số chấm chia hết cho 3. Khi đó biến cố tích AB là biến cố mặt 6 chấm xuất hiện.

   Tương tự như trong lí thuyết tập hợp ta có thể mở rộng định nghĩa của tổng và tích nhiều biến cố: nếu $A_1, A_2, A_3, \ldots$ là một họ hữu hạn hoặc vô hạn các biến cố ngẫu nhiên, khi đó tổng

$$\sum_i A_i \quad (\textit{hoặc kí hiệu } \bigcup_i A_i)$$

   là biến cố ít nhất một trong các biến cố $A_1, A_2, A_3, \ldots$ xuất hiện và tích $\prod_i A_i$ (hoặc kí hiệu $\bigcap_i A_i$) là biến cố: các biến cố $A_1, A_2, A_3, \ldots$ đồng thời cùng xuất hiện.

4. Hoàn toàn như trong lí thuyết tập hợp ta có thể định nghĩa các phép toán cũng như các mối quan hệ khác giữa các biến cố:

   - Hai biến cố A và B được gọi là xung khắc nhau nếu $AB = \emptyset$. Nói cách khác hai biến cố A và B được gọi là xung khắc nhau nếu xảy ra biến cố này thì không xảy ra biến cố kia và ngược lại.

- Hiệu của hai biến cố A và B, kí hiệu $A\backslash B$, là biến cố xảy ra khi và chỉ khi biến cố A xảy ra và biến cố B không xảy ra.

- Biến cố đối lập với biến cố A, kí hiệu $\overline{A}$, là biến cố xảy ra khi và chỉ khi A không xảy ra.

Dễ dàng nhận thấy $\overline{A} = \Omega\backslash A$ và A $\backslash B = A\overline{B}$.

5. Tương ứng với biến cố đối lập trong lí thuyết tập hợp là khái niệm phần bù. Biến cố $\overline{A}$ là phần bù của A trong Ω. Do vậy luật De Morgan cũng đúng với các biến cố ngẫu nhiên

$$\overline{\sum_i A_i} = \prod_i \overline{A_i}$$

$$\overline{\prod_i A_i} = \sum_i \overline{A_i}$$

6. Trong lí thuyết xác suất, ta thường gặp khái niệm các biến cố

$$\{A_1, A_2, A_3, \cdots, A_n\}$$

lập thành một hệ đầy đủ các biến cố, nếu tổng của chúng là biến cố chắc chắn xảy ra và đôi một xung khắc nhau:

$$\sum_{i=1}^{n} A_i = \Omega \quad \text{và} \quad A_iA_j = \emptyset \quad \text{với mọi} \quad i \neq j$$

Ví dụ khi gieo xúc xắc các biến cố $\{\omega_1, \omega_2, ..., \omega_6\}$ lập thành một hệ đầy đủ.

Tổng quát hơn ta có định nghĩa sau

Định nghĩa 1.1.1 *Một hệ gồm vô hạn các biến cố ngẫu nhiên*

$$A_1, A_2, ..., A_n, ...$$

được gọi là một hệ đầy đủ nếu

$$\sum_{i=1}^{\infty} A_i = \Omega \qquad \textit{và}$$

$$A_iA_j = \emptyset \quad \textit{với mọi} \quad i \geq 1, j \geq 1 \quad \textit{và} \quad i \neq j.$$

Ví dụ 1.1.1

Khi gieo liên tiếp một xúc xắc, gọi A_k là biến cố lần đầu tiên mặt 6 chấm xuất hiện ở lần gieo thứ k. Nói cách khác $k-1$ lần gieo đầu không xuất hiện mặt 6 chấm và ở lần gieo thứ k biến cố "mặt 6 chấm" mới xuất hiện. Như vậy với $k = 1, 2, ...$ các biến cố A_k đôi một xung khắc nhau

$$A_i A_k = \emptyset \quad \text{với mọi} \quad i \neq k$$

Ta gọi B là biến cố "mặt 6 chấm" xuất hiện sau một số lần gieo nào đó. Nói cách khác mặt 6 chấm có thể xảy ra ở lần gieo thứ nhất hoặc thứ hai, hoặc thứ ba... Điều đó tương đương với việc một biến cố nào đó trong dãy các biến cố $A_k, \quad k = 1, 2, 3, ...$ xảy ra. Như vậy biến cố B bằng tổng của các biến cố A_k

$$B = \sum_{k=1}^{\infty} A_k$$

Sau này trong nhận xét 3 mục 1.3 ta sẽ chỉ ra biến cố B là biến cố chắc chắn xảy ra $B = \Omega$. Do vậy hệ vô hạn các biến cố ngẫu nhiên

$$A_1, A_2, ..., A_n, ...$$

là một hệ đầy đủ theo nghĩa tổng quát nói trên.

Ví dụ 1.1.2

Cho A và B là hai biến cố ngẫu nhiên. Kí hiệu C_n là biến cố

$$C_n = \begin{cases} A & \text{nếu } n \text{ lẻ} \\ B & \text{nếu } n \text{ chẵn} \end{cases}$$

Hãy xác định các biến cố

$$\sum_{m=1}^{\infty} \prod_{i=m}^{\infty} C_i$$

$$\prod_{m=1}^{\infty} \sum_{i=m}^{\infty} C_i$$

Từ định nghĩa của biến cố C_n, dễ dàng nhận thấy

$$\prod_{i=m}^{\infty} C_i = AB \quad \text{cũng như} \quad \sum_{i=m}^{\infty} C_i = A + B \quad \text{với mọi } m,$$

suy ra

$$\sum_{m=1}^{\infty} \prod_{i=m}^{\infty} C_i = AB$$

$$\prod_{m=1}^{\infty} \sum_{i=m}^{\infty} C_i = A + B$$

Nhận xét rằng nếu $A_1, A_2, ..., A_n, ...$ là một dãy vô hạn các biến cố giảm dần $A_1 \supset A_2 \supset A_3 \supset ...$ Khi đó

$$\prod_{i=1}^{\infty} A_i = \prod_{i=2}^{\infty} A_i = \prod_{i=3}^{\infty} A_i = ...$$

Suy ra

$$\sum_{m=1}^{\infty} \prod_{i=m}^{\infty} A_i = \prod_{i=1}^{\infty} A_i$$

Tương tự nếu $A_1, A_2, ..., A_n, ...$ là một dãy vô hạn các biến cố tăng dần $A_1 \subset A_2 \subset A_3 \subset ...$ Khi đó

$$\prod_{m=1}^{\infty} \sum_{i=m}^{\infty} A_i = \sum_{i=1}^{\infty} A_i$$

Một cách tổng quát, ta có định nghĩa sau

Định nghĩa 1.1.2 *Với một dãy vô hạn các biến cố ngẫu nhiên bất kì*

$$A_1, A_2, ..., A_n, ...$$

Người ta gọi $\sum_{m=1}^{\infty} \prod_{i=m}^{\infty} A_i$ *là* $lim\, inf$ *và tương tự* $\prod_{m=1}^{\infty} \sum_{i=m}^{\infty} A_i$ *là* $lim\, sup$ *của dãy các biến cố ngẫu nhiên* A_n :

$$\sum_{m=1}^{\infty} \prod_{i=m}^{\infty} A_i = lim\, inf A_i.$$

$$\prod_{m=1}^{\infty}\sum_{i=m}^{\infty} A_i = lim\, sup A_i.$$

Ta có thể thấy ý nghĩa của các biến cố $lim\, inf A_i$ và $lim\, sup A_i$. Chẳng hạn $lim\, sup A_i$ là biến cố xảy ra nếu đồng thời xảy ra *vô hạn các biến cố* trong dãy các biến cố ngẫu nhiên ban đầu $A_1, A_2, ...$ còn $lim\, inf A_i$ là biến cố xảy ra khi *xuất hiện mọi biến cố trừ hữu hạn biến cố* trong dãy các biến cố $A_1, A_2, ...$ Hiển nhiên

$$lim\, inf A_i \subset lim\, sup A_i$$

Đặc biệt nếu $A_1, A_2, ..., A_n, ...$ là một dãy vô hạn các biến cố tăng dần, khi đó

$$lim\, inf A_i = lim\, sup A_i = \sum_{i=1}^{\infty} A_i.$$

Tương tự nếu $A_1, A_2, ..., A_n, ...$ là một dãy vô hạn các biến cố giảm dần, khi đó

$$lim\, inf A_i = lim\, sup A_i = \prod_{i=1}^{\infty} A_i.$$

Một cách tổng quát, nếu dãy vô hạn các biến cố thoả mãn

$$lim\, inf A_i = lim\, sup A_i$$

khi đó ta nói dãy $A_1, A_2, ...$ hội tụ và kí hiệu

$$lim A_i = lim\, inf A_i = lim\, sup A_i.$$

Như vậy đối với dãy các biến cố tăng dần (hoặc giảm dần)

$$lim A_i = \sum_{i=1}^{\infty} A_i \quad (\text{hoặc} \quad lim A_i = \prod_{i=1}^{\infty} A_i).$$

Ví dụ 1.1.3

Giả sử Ω gồm n biến cố ngẫu nhiên cơ bản. Chứng tỏ rằng số các biến cố ngẫu nhiên của nó bằng 2^n.

Thật vậy, gọi $\mathfrak{A}$ là tập hợp gồm tất cả các biến cố ngẫu nhiên (các tập con của Ω) suy ra $\mathfrak{A}$ là tập hữu hạn phần tử và số các phần tử của $\mathfrak{A}$ dễ dàng tính được như sau:

Số các tập con gồm đúng k phần tử của Ω bằng $C_n^k (1 \leq k \leq n)$. Khi đó tổng $\sum_{k=1}^{n} C_n^k$ là số các tập con khác trống của Ω. Vậy toàn bộ số các tập con (kể cả tập trống) bằng

$$1 + \sum_{k=1}^{n} C_n^k = \sum_{k=0}^{n} C_n^k = (1+1)^n = 2^n$$

(Theo khai triển nhị thức Newton). Vậy số các phần tử của $\mathfrak{A}$, hay số các biến cố ngẫu nhiên cần tìm bằng 2^n.

Chẳng hạn khi gieo xúc xắc, Ω gồm 6 biến cố ngẫu nhiên cơ bản, suy ra số các biến cố ngẫu nhiên là $2^6 = 64$.

1.2 Tần suất và xác suất

1.2.1 Khái niệm về tần suất và xác suất của các biến cố ngẫu nhiên

Ta tiến hành phép thử ngẫu nhiên T để quan sát biến cố ngẫu nhiên A nào đó. Nói chung chúng ta không có cơ sở nào để dự đoán kết quả của phép thử T cũng như biến cố A có xảy ra hay không. Nếu phép thử T được tiến hành nhiều lần và các lần tiến hành phép thử độc lập với nhau, các kết quả của chúng dường như có vẻ hỗn độn không theo một quy luật nào cả. Thực ra dãy các kết quả của các phép thử ngẫu nhiên đó chứa các quy luật mà ta sẽ bàn đến trong mục này cũng như trong lí thuyết xác suất nói chung.

Giả sử trong số n lần tiến hành phép thử T, có đúng k lần xuất hiện biến cố A. Khi đó tỉ số $\frac{k}{n}$ được gọi là *tần suất* xuất hiện của biến cố A. Tần suất xuất hiện của biến cố A nói chung không phải là hằng số trong dãy các phép thử ngẫu nhiên. Tuy nhiên độ dao động của nó càng nhỏ khi tăng số lần tiến hành phép thử. Nói cách khác, tần suất ổn định và dao động xung

quanh một số xác định nào đó hay tỉ số $\frac{k}{n}$ sẽ tiến dần tới một giá trị xác định khi n dần tới vô hạn. Quy luật đó trong lí thuyết xác suất được gọi là luật số lớn, được phát hiện và nghiên cứu từ thế kỉ XVII. Nó đặt nền móng cho việc phát triển và ứng dụng môn lí thuyết xác suất. Chúng ta nhắc đến ở đây các thí nghiệm của Buffon và Pearson để nghiên cứu quy luật sự ổn định của tần suất từ thế kỉ XVIII.

Buffon đã 4040 lần gieo đồng xu và ghi lại 2048 lần mặt sấp xuất hiện. Như vậy tần suất xuất hiện mặt sấp của đồng xu là 0,5069.

Pearson đã tiến hành gieo đồng xu nhiều lần hơn, ông gieo 24000 lần và nhận được kết quả tần suất xuất hiện mặt sấp của đồng xu là 0,5005. (Có thể coi đồng xu là đồng chất, đối xứng và các lần gieo độc lập với nhau, do vậy vai trò của hai biến cố mặt sấp xuất hiện và mặt ngửa xuất hiện là như nhau, suy ra tần suất xuất hiện mặt sấp của đồng xu dao động xung quanh số $\frac{1}{2}$).

Số thực mà tần suất xuất hiện của biến cố A dao động quanh nó được gọi là xác suất của biến cố A. Người ta thường kí hiệu xác suất của biến cố A là $P(A)$. Ví dụ khi gieo đồng xu xác suất của biến cố mặt sấp xuất hiện là $\frac{1}{2}$. Khái niệm về *xác suất* của biến cố ngẫu nhiên A như vậy chưa là một định nghĩa chính xác về mặt toán học. Nó chỉ đưa ra một sự thực mà việc xây dựng các khái niệm của lí thuyết xác suất cần đạt được. Cũng như mọi lĩnh vực khác của toán học, lí thuyết xác suất được xây dựng trên một hệ tiên đề mà các khái niệm về tần suất, xác suất cũng như quy luật về sự ổn định của tần suất kể trên được bảo đảm.

Vì vậy trước khi phát biểu một cách chính xác hơn khái niệm toán học về xác suất, chúng ta cần nói đến một số tính chất cơ bản của tần suất:

1. Nếu tiến hành một phép thử ngẫu nhiên n lần để quan sát biến cố A nào đó. Giả sử trong dãy phép thử nói trên có đúng k lần suất hiện biến cố A, khi đó tần suất xuất hiện của biến cố A bông $\frac{k}{n}$. Do $0 \leq k \leq n$, suy ra $0 \leq \frac{k}{n} \leq 1$ với mọi n. Như vậy tần suất xuất hiện của biến cố A bất kì là số thực nôm giữa 0 và 1.

2. Biến cố chắc chắn xảy ra Ω luôn luôn có tần suất bằng $\frac{n}{n} = 1$, còn biến cố không thể xảy ra $\emptyset$ có tần suất bằng 0.

3. Khi tiến hành một phép thử ngẫu nhiên n lần, biến cố A, biến cố B là hai biến cố xung khắc nhau ($AB = \emptyset$). Gọi k_A, k_B, k_{A+B} là số lần xuất hiện biến cố A, biến cố B, biến cố $A + B$ tương ứng trong dãy phép

thử ngẫu nhiên nói trên. Khi đó hiển nhiên $k_{A+B} = k_A + k_B$. Chia cả hai vế cho n, ta được

$$\frac{k_{A+B}}{n} = \frac{k_A}{n} + \frac{k_B}{n}$$

Suy ra tần suất của tổng hai biến cố xung khắc nhau $A + B$ bông tổng các tần suất của hai biến cố đó. kết quả cũng tương tự như vậy đối với hữu hạn hoặc vô hạn đếm được các biến cố đôi một xung khắc nhau. Nếu chúng ta coi khái niệm xác suất (một cách thô thiển) như là giới hạn của tần suất, khi đó mọi hệ tiên đề về xác suất phải kế thừa được các tính chất trên về tần suất. Bây giờ chúng ta có thể phát biểu hệ tiên đề về xác suất, cơ sở toán học chặt chẽ của nó được nhà toán học Nga A. N. Kolmogorov đưa ra năm 1933.

1.2.2 Các tiên đề về xác suất của biến cố ngẫu nhiên

Trước hết chúng ta đưa vào khái niệm một họ $\mathfrak{A}$ các tập con nào đó của không gian các biến cố ngẫu nhiên cơ bản Ω được gọi là σ-đại số nếu:

1. $\Omega \in \mathfrak{A}$

2. $A \in \mathfrak{A}$ suy ra $\Omega \backslash A \in \mathfrak{A}$

3. Nếu $A_1, A_2, ...$ là dãy các tập hợp thuộc $\mathfrak{A}$, khi đó hợp của chúng $\bigcup_i A_i$ cũng thuộc $\mathfrak{A}$.

Trong lí thuyết xác xuất, tập các biến cố ngẫu nhiên là một σ-đại số $\mathfrak{A}$. Một ánh xạ P từ $\mathfrak{A}$ vào tập các số thực $\mathbb{R}$

$$P : \mathfrak{A} \to R$$

thực chất là một phép gán mỗi biến cố ngẫu nhiên A cho một số thực $P(A)$, số đó được gọi là *xác suất* của biến cố A nếu thoả mãn các *tiên đề* sau:

1. Với mọi $A \in \mathfrak{A} \quad 0 \leq P(A) \leq 1$

2. $P(\Omega) = 1$

3. Nếu $A_1, A_2, ..., A_i, ...$ là các biến cố ngẫu nhiên đôi một xung khắc nhau thuộc $\mathfrak{A}$, khi đó

$$P\left(\sum_i A_i\right) = \sum_i P(A_i)$$

Ánh xạ P được gọi là *phân bố xác suất* còn $P(A)$ được gọi là *xác suất* của biến cố ngẫu nhiên A.

Dựa vào hệ tiên đề này, người ta đã chứng minh được nhiều kết quả có tầm quan trọng rất lớn, trong đó có luật số lớn. Một trong các kết quả của luật số lớn là tần suất xuất hiện của biến cố ngẫu nhiên tiến dần tới xác suất của biến cố ngẫu nhiên đó khi n (số lần tiến hành phép thử ngẫu nhiên) tăng dần ra vô hạn.

Chú ý rằng theo hệ tiên đề trên, trên không gian các biến cố ngẫu nhiên cơ bản Ω, ta có thể cho nhiều phân bố xác suất, tuy nhiên việc xác định phân bố xác suất nào phù hợp với thực tế là nhiệm vụ của lí thuyết thống kê sau này.

Chẳng hạn khi gieo một đồng xu , tập $\mathfrak{A}$ các biến cố ngẫu nhiên gồm 4 phần tử $\mathfrak{A}$ ={S: mặt sấp, N: mặt ngửa, $\emptyset$ và Ω}. Các phân bố xác suất có thể có trên $\mathfrak{A}$ là:

$$P(S) = p, P(N) = 1 - p, P(\emptyset) = 0 \text{ và} P(\Omega) = 1$$

trong đó p là số thực bất kì thuộc khoảng $[0, 1]$. Dễ dàng kiểm tra ánh xạ

$$P : \mathfrak{A} \to \mathbb{R}$$

nói trên thoả mãn hệ tiên đề Kolmogorov. Như vậy tương ứng với mỗi số thực $p \in [0, 1]$ ta có một phân bố xác suất trên $\mathfrak{A}$. Trong thực tế, nếu đồng xu là đối xứng, đồng chất ta thường gán cho các biến cố S (mặt sấp), N (mặt ngửa) các xác suất $P(S) = \frac{1}{2}, P(N) = \frac{1}{2}$, còn nếu đồng xu không đối xứng hoặc không đồng chất, phân bố xác suất thực trên $\mathfrak{A}$ phải là các số $p \in [0, 1]$ nào đó (có thể $p \neq \frac{1}{2}$) cần xác định cho phù hợp với thực tế.

Một ví dụ khác, khi gieo xúc xắc gọi $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6$ là các biến cố ngẫu nhiên cơ bản tương ứng với các mặt 1 chấm, 2 chấm, 3 chấm, 4 chấm, 5 chấm, 6 chấm xuất hiện. Gán cho mỗi biến cố ω_i xác suất $P(\omega_i) = p_i \in [0, 1]$ với $\forall i = 1, 2, ..., 6$ sao cho

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1.$$

$\mathfrak{A}$ là tập hợp tất cả các tập con của $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$. Dễ dàng nhận thấy $\mathfrak{A}$ là σ-đại số và phép gán (ánh xạ)

$$P(A) = \sum_{i:\omega_i \in A} p_i$$

thoả mãn hệ tiên đề Kolmogorov $\Rightarrow P$ là phân bố xác suất trên $\mathfrak{A}$. Như vậy có thể có nhiều cách gán là phân bố xác suất, tuy nhiên thực tiễn sẽ quyết định cách gán nào phù hợp, đúng đắn. Chẳng hạn khi xúc xắc là đồng chất, đối xứng người ta thường coi khả năng xuất hiện của các biến cố $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6$ là như nhau, vì vậy hiển nhiên các xác suất

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$$

là phù hợp hơn cả.

1.3 Các định lí cơ bản và phương pháp tính xác suất các biến cố ngẫu nhiên

1.3.1 Các tính chất cơ bản của xác suất

Từ hệ tiên đề trên ta có thể chứng minh một số kết quả cơ bản sau

Định lí 1.3.1 *Nếu $A \Rightarrow B$ (hay $A \subset B$) khi đó $P(A) \leq P(B)$.*

Chứng minh. Từ $A \subset B$, suy ra $B = A + C$, trong đó $C = B \backslash A$ và do vậy $AC = \emptyset$. Theo tiên đề 3 trong hệ tiên đề Kolmogorov $P(B) = P(A) + P(C)$. Mặt khác theo tiên đề 1

$$P(C) \geq 0 \quad \text{suy ra} \quad P(A) \leq P(B).$$

Do A và $\overline{A}$ lập thành một hệ đầy đủ: $A + \overline{A} = \Omega$ và $A\overline{A} = \emptyset$. Từ tiên đề 2. của hệ tiên đề, suy ra

Định lí 1.3.2 *Với mọi biến cố ngẫu nhiên A*

$$P(A) + P(\overline{A}) = 1.$$

Nhận xét rằng, tổng quát hơn định lí 1.3.2, nếu $A_1, A_2, ..., A_n$ là hệ đầy đủ các biến cố ngẫu nhiên, khi đó:

$$P(A_1) + P(A_2) + ... + P(A_n) = 1.$$

Định lí 1.3.3 (Định lí cộng xác suất) *Gọi A và B là hai biến cố bất kì. Khi đó*

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

Chứng minh. Từ hệ thức $A + B = A + (B\backslash AB)$ và $A \cap (B\backslash AB) = \emptyset$ $\Rightarrow$ $P(A+B) = P(A) + P(B\backslash AB)$. Mặt khác $AB \subset B$, vì vậy theo Định lí 1

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

Nhận xét 1.3.1 *Sử dụng định lí 1.3.3, chúng ta có thể mở rộng cho tổng của 3 hoặc nhiều hơn các biến cố. Chẳng hạn, xét 3 biến cố A, B, C tùy ý, khi đó:*

$$\begin{aligned}
&P(A + B + C) = P\left(A + (B + C)\right) = \\
=&P(A) + P(B + C) - P(A(B + C)) = \\
=&P(A) + P(B + C) - P(AB + AC) = \\
=&P(A) + P(B) + P(C) - P(BC) - \big(P(AB) + P(AC) - P(ABAC)\big) \\
=&P(A) + P(B) + P(C) - P(BC) - P(AB) - P(AC) + P(ABC).
\end{aligned}$$

Bằng quy nạp ta dễ dàng chứng minh định lí sau:

Định lí 1.3.4 *Với các biến cố $A_1, A_2, ..., A_n$ tùy ý ta luôn có bất đẳng thức*

$$P(A_1 + A_2 + ... + A_n) \leq P(A_1) + P(A_2) + ... + P(A_n).$$

Định lí 1.3.5 *Cho một dãy vô hạn các biến cố $A_1, A_2, ..., A_n, ...$ thoả mãn điều kiện giảm dần $A_1 \supset A_2 \supset A_3 \supset ....$ Khi đó*

$$\lim_{n\to\infty} P(A_n) = P(\prod_{i=1}^{\infty} A_i).$$

Chứng minh. Giả sử $A = \prod_{i=1}^{\infty} A_i$. Đặt $B_k = A_k \backslash A_{k+1} \quad k = 1, 2, 3...$ Khi đó dãy các biến cố B_k đôi một xung khắc nhau và $A_1 = A + \sum_{i=1}^{\infty} B_i$. Theo tiên đề 3.

$$P(A_1) = P(A) + \sum_{i=1}^{\infty} P(B_i) = P(A) + \lim_{n\to\infty} (P(A_1) - P(A_n))$$

$$\Rightarrow \quad \lim_{n\to\infty} P(A_n) = P(A) = P(\prod_{i=1}^{\infty} A_i).$$

Nhận xét 1.3.2 *Nếu $A_1, A_2, ..., A_n, ...$ là một dãy các biến cố thoả mãn điều kiện $A_1 \subset A_2 \subset A_3 \subset ...$ Áp dụng định lí 1.3.5, với dãy các biến cố giảm $\overline{A_i}, \quad i = 1, 2, 3...$ ta được kết quả tương tự*

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

Hệ quả 1.3.1 *Với một dãy vô hạn các biến cố ngẫu nhiên bất kì*

$$A_1, A_2, ..., A_n, ...$$

ta luôn có:

$$P\left(\sum_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$$

Thật vậy theo định lí 1.3.4

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

Mặt khác $B_n = \sum_{i=1}^{n} A_i$ là dãy các biến cố tăng

$$B_1 \subset B_2 \subset B_3, ... \qquad B = \sum_{i=1}^{\infty} B_i = \sum_{i=1}^{\infty} A_i$$

Theo nhận xét trên

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} P\left(\sum_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i) \qquad \text{(đ.p.c.m.)}$$

1.3.2 Các phương pháp để tính xác suất

Phương pháp cổ điển

Phương pháp cổ điển để giải các bài toán trong đó phép thử ngẫu nhiên chỉ gồm hữu hạn các khả năng có thể xảy ra và khả năng xảy ra của chúng như nhau. Nói cách khác không gian các biến cố cơ bản Ω chỉ gồm hữu hạn các biến cố $\{\omega_1, \omega_2, ..., \omega_n\}$. Chúng ta giả thiết rằng **các biến cố ngẫu nhiên cơ bản** $\omega_1, \omega_2, ..., \omega_n$ **đồng khả năng**, hay xác suất để xảy ra các biến cố $\omega_1, \omega_2, ..., \omega_n$ là như nhau:

$$P(\omega_1) = P(\omega_2) = ... = P(\omega_n)$$

Từ hệ tiên đề xác suất, $P(\Omega) = 1$, ta suy ra

$$P(\omega_1) = P(\omega_2) = ... = P(\omega_n) = \frac{1}{n}$$

Vì vậy nếu biến cố ngẫu nhiên A (tập con của Ω) gồm m biến cố ngẫu nhiên cơ bản đồng khả năng

$$A = \{\omega_{n_1}, \omega_{n_2}, ..., \omega_{n_m}\}$$

khi đó cũng theo tiên đề về xác suất

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

m còn được gọi là **số trường hợp thuận lợi** cho biến cố A, n là **số trường hợp đồng khả năng**. Theo cách viết truyền thống của phương pháp cổ điển, xác suất của biến cố ngẫu nhiên A được tính theo công thức:

$$\boxed{P(A) = \frac{\text{số trường hợp thuận lợi}}{\text{số trường hợp đồng khả năng}}.}$$

Ví dụ 1.3.1

Hãy tìm xác suất để khi chọn ngẫu nhiên các chữ cái A, G, H, H, I, O, P, N và lần lượt xếp chúng lại cạnh nhau theo thứ tự đã chọn, ta được chữ $HAIPHONG$.

Các biến cố ngẫu nhiên cơ bản trong ví dụ này là các hoán vị của 8 chữ cái trên. Do vậy số trường hợp đồng khả năng chính là số hoán vị của 8 phần tử: 8!=40320. Chữ $HAIPHONG$ không thay đổi nếu ta hoán vị hai chữ cái H cho nhau. Vậy số trường hợp thuận lợi bằng 2. Suy ra xác suất cần tìm là $\frac{2}{40320} = \frac{1}{20160}$.

Ví dụ 1.3.2

Một nhóm 9 người ngồi trong một cửa hàng giải khát. Họ gọi tất cả 3 chai bia, 4 hộp cola và 2 cà phê (mỗi người chỉ uống một thứ). Người phục vụ không nhớ ai đặt gì , do vậy đưa ngẫu nhiên cho mỗi người khách một thứ đồ uống. Tính xác suất để mọi người đều nhận đúng thứ mình yêu cầu.

Các biến cố ngẫu nhiên cơ bản trong ví dụ này cũng là các hoán vị của 9 loại đồ uống. Vậy số trường hợp đồng khả năng bằng 9!. Số trường hợp thuận lợi, để mọi người đều nhận đúng thứ đồ uống mình yêu cầu, hiển nhiên là số hoán vị lặp 3!4!2!. Vậy xác suất cần tìm bằng

$$\frac{3!4!2!}{9!}$$

Ví dụ 1.3.3

Gieo đồng thời 3 xúc xắc. Tìm xác suất để cả ba xúc xắc xuất hiện ba số đôi một khác nhau.

Các biến cố ngẫu nhiên cơ bản khi gieo đồng thời 3 xúc xắc là các chỉnh hợp lặp chập 3 của 6 phần tử. Do vậy số trường hợp đồng khả năng là 6^3 và số trường hợp thuận lợi cho biến cố cả ba xúc xắc xuất hiện ba số đôi một khác nhau bằng số chỉnh hợp chập 3 của 6 phần tử: $A_6^3 = 6 \cdot 5 \cdot 4$. Vậy xác suất cần tìm bằng: $\frac{6 \cdot 5 \cdot 4}{6^3}$.

Ví dụ 1.3.4

Gieo một xúc xắc n lần. Tìm xác suất để không xảy ra biến cố: có hai lần gieo liên tiếp nào đó xuất hiện cùng một số.

Tương tự như trong ví dụ trên số các trường hợp đồng khả năng (số các biến cố ngẫu nhiên cơ bản) là 6^n và số trường hợp thuận lợi bằng $6 \cdot 5^n$ (vì ở lần gieo đầu có thể xảy ra 6 khả năng tùy ý, ở các lần gieo sau chỉ có 5 khả năng thuận lợi cho biến cố không có hai lần gieo liên tiếp nào xuất hiện cùng một số). Vậy xác suất cần tìm bằng: $\frac{6 \cdot 5^{n-1}}{6^n} = \left(\frac{5}{6}\right)^{n-1}$.

Ví dụ 1.3.5

Gọi A_k là biến cố khi gieo liên tiếp k lần một xúc xắc, lần đầu tiên mặt 6 chấm xuất hiện ở lần gieo thứ k.

(a) Tính $P(A_k)$

(b) Hãy tính xác suất để sau một số lần gieo ít hơn 5, xuất hiện mặt 6 chấm.

(c) Hãy tính xác suất để khi gieo xúc xắc liên tiếp, sau một số lần gieo nào đó mặt 6 chấm xuất hiện.

(d) Hãy tính xác suất để sau một số chẵn lần gieo xúc xắc, mặt 6 chấm xuất hiện.

(a) Các biến cố ngẫu nhiên cơ bản là các chỉnh hợp lặp chập k của 6 phần tử. Vậy không gian các biến cố ngẫu nhiên cơ bản gồm 6^k biến cố (do chúng ta luôn quy ước các biến cố ngẫu nhiên cơ bản đồng khả năng xuất hiện nên 6^k cũng là số trường hợp đồng khả năng trong công thức tính xác suất bằng phương pháp cổ điển). Số trường hợp thuận lợi cho biến cố A_k được xác định như sau: k-1 lần gieo đầu không xuất hiện mặt 6 chấm và lần gieo thứ k mặt 6 chấm xuất hiện. Như vậy ở k-1 lần gieo đầu có 5^{k-1} khả năng khác nhau $\Rightarrow$ Số trường hợp thuận lợi cho biến cố A_k là $5^{k-1} \cdot 1 = 5^{k-1}$. Vì vậy

$$P(A_k) = \frac{5^{k-1}}{6^k} = \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{k-1}$$

(b) Gọi B_4 là biến cố sau một số lần gieo ít hơn 5, xuất hiện mặt 6 chấm. Như vậy biến cố B_4 xuất hiện khi và chỉ khi mặt 6 chấm xuất hiện ở lần gieo đầu (biến cố A_1) hoặc lần gieo thứ nhất mặt 6 chấm không xuất hiện và lần gieo thứ hai mặt 6 chấm xuất hiện(biến cố A_2) hoặc hai lần gieo đầu mặt 6 chấm không xuất hiện và lần gieo thứ ba mặt 6 chấm xuất hiện (biến cố A_3) hoặc biến cố A_4 xảy ra. Nói cách khác

$$B_4 = A_1 + A_2 + A_3 + A_4.$$

Mặt khác các biến cố A_1, A_2, A_3, A_4 đôi một xung khắc nhau

$$\Rightarrow P(B_4) = P(A_1) + P(A_2) + P(A_3) + P(A_4)$$
$$= \frac{1}{6}\left(1 + \frac{5}{6} + \frac{5^2}{6^2} + \frac{5^3}{6^3}\right) = \frac{671}{6^4} = 0,517747$$

(c) Gọi B là biến cố khi gieo liên tiếp một xúc xắc, sau một số lần gieo nào đó mặt 6 chấm xuất hiện. Lập luận như trên $B = \sum_{i=1}^{\infty} A_i$
Mặt khác các biến cố $A_1, A_2, A_3, ...$ đôi một xung khắc nhau

$$\Rightarrow P(B) = \sum_{i=1}^{\infty} P(A_i) = \frac{1}{6}\sum_{i=1}^{\infty} \frac{5^{i-1}}{6^{i-1}} = \frac{1}{6}\frac{1}{1-\frac{5}{6}} = 1$$

($\sum_{i=1}^{\infty} \frac{5^{i-1}}{6^{i-1}}$ là tổng của chuỗi cấp số nhân)

Nhận xét 1.3.3 *Trong ví dụ này biến cố B, khi gieo liên tiếp một xúc xắc, sau một số lần gieo nào đó mặt 6 chấm xuất hiện, có xác suất bằng 1: $P(B) = 1$, điều đó có nghĩa là biến cố B là biến cố chắc chắn xảy ra. Nói cách khác khi gieo liên tiếp một xúc xắc, sớm hay muộn chắc chắn mặt 6 chấm sẽ xuất hiện một lúc nào đó.*

(d) Nếu kí hiệu C là biến cố sau một số chẵn lần gieo xúc xắc, mặt 6 chấm xuất hiện. Khi đó C có thể được biểu diễn như sau:

$$C = \sum_{i=1}^{\infty} A_{2i} = A_2 + A_4 + A_6 + ... + ...$$

$$\Rightarrow \quad P(C) = \sum_{i=1}^{\infty} P(A_{2i}) = \frac{1}{6}\sum_{i=1}^{\infty} \frac{5^{2i-1}}{6^{2i-1}} = \frac{1}{5} \cdot \frac{25}{11} = \frac{5}{11}$$

Ví dụ 1.3.6

Chọn ngẫu nhiên từ cỗ bài tú lơ khơ 52 quân ra 5 quân bài. Tìm xác suất để trong đó có đúng 2 quân K và 1 quân bài át. Các biến cố ngẫu nhiên cơ bản là các tổ hợp chập 5 của 52 phần tử. Vậy số trường hợp đồng khả năng là $C_{52}^5 = 2598960$. Gọi A là biến cố trong 5 quân bài được chọn ra có 2 quân bài K và 1 quân bài át. Số trường hợp thuận lợi cho biến cố A là $C_4^2 C_4^1 C_{44}^2 = 22704$

$$\Rightarrow \quad P(A) = \frac{C_4^2 C_4^1 C_{44}^2}{C_{52}^5} = \frac{22704}{2598960} = \frac{473}{54145} = 0,00874$$

Ví dụ 1.3.7

Một nhóm gồm $2N$ nam và $2N$ nữ bị tách ngẫu nhiên thành 2 nhóm nhỏ với số người bằng nhau. Hãy tính xác suất để số nam và số nữ trong mỗi nhóm nhỏ đều bằng nhau.

Tổng số người trong cả nhóm là $4N$, như vậy việc tách thành 2 nhóm cũng có nghĩa là việc chọn ngẫu nhiên $2N$ người từ cả nhóm $4N$ người. Vậy số trường hợp đồng khả năng là C_{4N}^{2N}.

Để số nam và số nữ trong mỗi nhóm bằng nhau khi chọn $2N$ người từ cả nhóm, ta phải lấy N nam (trong tổng số $2N$ nam) và N nữ (trong tổng số $2N$ nữ). Suy ra số trường hợp thuận lợi bằng $\left(C_{2N}^{N}\right)^2$. Vậy xác suất cần tìm là

$$\frac{\left(C_{2N}^{N}\right)^2}{C_{4N}^{2N}}$$

Ví dụ 1.3.8

Hãy tính xác suất để sinh nhật của 12 người được chọn bất kì rơi vào các tháng khác nhau. (Giả sử xác suất để mỗi người sinh vào các tháng là như nhau).

Số các trường hợp đồng khả năng là số chỉnh hợp lặp chập 12 của 12. Số trường hợp thuận lợi chính là số hoán vị các tháng sinh khác nhau của 12 người. Vậy xác suất để sinh nhật của 12 người vào các tháng khác nhau bằng

$$\frac{12!}{12^{12}} = \frac{11!}{12^{11}}.$$

Phương pháp hình học

Khi phép thử ngẫu nhiên gồm vô hạn các khả năng có thể xảy ra (không gian các biến cố ngẫu nhiên cơ bản Ω gồm vô hạn các phần tử), hiển nhiên ta không thể áp dụng phương pháp cổ điển nói trên để tính xác suất các biến cố ngẫu nhiên. Khi đó phương pháp hình học để tính xác suất của các biến cố ngẫu nhiên có thể áp dụng trong trường hợp sau đây: phép thử ngẫu nhiên T có thể coi như một thí nghiệm chọn ngẫu nhiên một điểm trong một miền hình học Ω nào đó

(miền hình học được hiểu là một đoạn thẳng, một cung đường cong, một miền phẳng hoặc một miền trong không gian). Giả sử miền hình học Ω là một miền có độ đo μ hữu hạn (độ đo của miền hình học được hiểu theo nghĩa rộng: là độ dài đoạn thẳng, độ dài cung đường cong, diện tích hoặc thể tích tương ứng với các miền hình học được nhắc đến ở trên). Phương pháp hình học để tính xác suất của các biến cố dựa trên một giả thiết cơ bản là: xác suất để điểm được chọn ngẫu nhiên rơi vào miền A (A là tập con của $\Omega : A \subset \Omega$) tỉ lệ với độ đo $\mu(A)$ của A. Như vậy các biến cố ngẫu nhiên cơ bản (điểm chọn ngẫu nhiên trong Ω) là các điểm thuộc Ω và họ các biến cố ngẫu nhiên (σ-đại số $\mathfrak{A}$) gồm các tập đo được trong Ω (khái niệm đo được ở đây là khái niệm độ dài, diện tích hoặc thể tích như đã nhắc đến ở trên). $P(A)$ tỉ lệ với độ đo của A, nghĩa là tồn tại hông số c sao cho $P(A) = c\mu(A)\ \forall A \in \mathfrak{A}$. Do $P(\Omega) = 1$, suy ra $c = \frac{1}{\mu(\Omega)}$. Vì vậy:

$$P(A) = \frac{\mu(A)}{\mu(\Omega)}.$$

Công thức trên là công thức để tính xác suất biến cố A bằng *phương pháp hình học*. Một cách tổng quát, người ta còn viết công thức trên dưới dạng

$$\boxed{P(A) = \frac{\text{độ đo của } A}{\text{độ đo của } \Omega}.}$$

Chú ý Xác suất của biến cố: *điểm được chọn ngẫu nhiên rơi vào miền A (để cho tiện ta cũng kí hiệu biến cố đó là biến cố A) không phụ thuộc vào hình dáng và vị trí của miền A mà chỉ phụ thuộc vào độ đo của A. Phân bố xác suất như vậy được gọi là* phân bố đều trên Ω. *Chúng ta sẽ nghiên cứu kĩ hơn phân bố này ở chương sau.*

Ví dụ 1.3.9

Một người quên không lên dây cót đồng hồ, do vậy sau một thời gian nhất định đồng hồ bị ''chết''. Tìm xác suất để kim giờ sẽ dừng giữa số 2 và số 4.

Chúng ta hoàn toàn có thể quan niệm sau khi chạy hết dây cót, các kim đồng hồ dừng lại một cách ngẫu nhiên. (Trong thực tế

người ta đã sử dụng nguyên lí đó để tạo các số ngẫu nhiên như sử dụng vành xe đạp để ''quay'' xổ số). Kim giờ có thể dừng lại chỉ bất cứ vị trí nào trên đường tròn mặt số của đồng hồ. Gọi A là biến cố kim giờ sẽ dừng giữa số 2 và số 4 khi đó xác suất của A bằng tỉ số giữa độ dài cung tròn nối các số 2 và 4 với chu vi đường tròn, hay $P(A) = \frac{1}{6}$.

Chú ý *Nhận xét rằng nếu kim đồng hồ chỉ biết dừng ở các vị trí phút: phút thứ nhất, phút thứ hai, . . . , phút thứ 60 khi đó có thể giải bài toán trên theo phương pháp cổ điển: không gian các biến cố ngẫu nhiên cơ bản (gồm 60 phần tử có xác suất như nhau và đều bằng $\frac{1}{60}$. Biến cố A kim giờ dừng giữa số 2 và số 4 là tập hợp gồm 10 biến cố ngẫu nhiên cơ bản* $\Rightarrow$ *$P(A) = \frac{10}{60} = \frac{1}{6}$. Nói chung các bài toán giải được bằng phương pháp hình học đều có thể chuyển về phương pháp cổ điển bằng cách chia Ω thành n biến cố ngẫu nhiên cơ bản đồng khả năng, sau đó cho n tăng ra ∞. Đó cũng chính là mối liên quan giữa phương pháp cổ điển và phương pháp hình học trong việc giải các bài toán xác suất.*

Ví dụ 1.3.10 (Bài toán gặp gỡ)

Hai người bạn hẹn gặp nhau tại một địa điểm X trong khong thời gian nào đó (ví dụ: từ 19 giờ đến 20 giờ). Họ quy ước người đến trước chờ người sau không quá 20 phút. Hãy tìm xác suất để hai người gặp được nhau.

Các khả năng có thể xảy ra ở đây là các cặp (t_1, t_2), trong đó t_1 chỉ thời điểm mà người thứ nhất tới điểm hẹn và t_2 là thời điểm mà người thứ hai tới điểm hẹn (ví dụ cặp (19 giờ 15', 19 giờ 40') chỉ rõ thời điểm mà người thứ nhất tới điểm hẹn là 19 giờ 15' và thời điểm mà người thứ hai tới điểm hẹn là 19 giờ 40'). Họ gặp được nhau nếu $|t_1 - t_2| \leq 20$ phút $= \frac{1}{3}$ giờ. Để thuận tiện hơn về mặt kí hiệu ta lập một tưng ứng một-một giữa các thời điểm trong khoảng từ 19 giờ đến 20 giờ với các số thực trong khoảng [0,1] (chẳng hạn chọn ánh xạ t $\rightarrow$ t-19, t là thời điểm từ 19 giờ đến 20 giờ). Khi đó cặp (19 giờ 15', 19 giờ 40') nói trên tương ứng với cặp $(\frac{1}{4}, \frac{2}{3})$ gồm hai số trong khoảng [0, 1]. Vậy các thời điểm mà 2 người tới điểm hẹn bây giờ có thể được biểu diễn như là các cặp (t_1, t_2), trong đó $t_i \in [0, 1]$ và cặp (t_1, t_2) được coi là một điểm được chọn ngẫu nhiên trong hình vuông $[0, 1] \times [0, 1]$. Gọi A là biến cố hai người bạn gặp được nhau, A xảy ra khi và

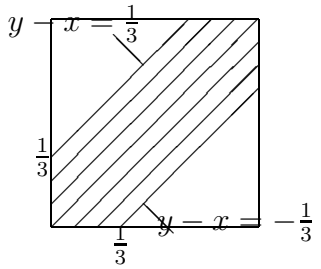
chỉ khi

$$|t_1 - t_2| \leq \frac{1}{3}.$$

Để thuận tiện cho việc minh hoạ bằng hình vẽ, thay vì sử dụng t_1 và t_2 bây giờ ta kí hiệu x là thời điểm người thứ nhất tới điểm hẹn và y là thời điểm người thứ hai tới điểm hẹn. Tập hợp những điểm (x, y) của hình vuông $[0, 1] \times [0, 1]$ thoả mãn bất đẳng thức

$$|x - y| \leq \frac{1}{3}$$

được mô tả trong hình vẽ dưới đây (phần gạch chéo) - cũng được kí hiệu là A - ứng với biến cố ngẫu nhiên hai người bạn gặp được nhau. Miền phẳng A có diện tích bằng $\frac{5}{9}$, hình vuông $[0, 1] \times [0, 1]$ có diện tích là 1. Vậy $P(A) = \frac{5}{9}$.



Ví dụ 1.3.11

Chọn ngẫu nhiên 3 điểm trên đoạn [0,1]. Tìm xác suất để các khoảng cách từ các điểm đã chọn tới 0 là độ dài của 3 cạnh một tam giác nào đó.

Lần lượt gọi các khoảng cách từ các điểm đã chọn tới 0 là x, y, z. Ta coi bộ 3 số (x, y, z) như một điểm được chọn ngẫu nhiên trong hình hộp

$$\{(x, y, z) : \quad 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1\}$$

Các khoảng cách x, y, z là độ dài của 3 cạnh một tam giác nào đó, nếu chúng thỏa mãn các bất đẳng thức

$$x + y > z, y + z > x, z + x > y$$

Dễ dàng nhận thấy trong hình hộp các điểm thỏa mãn các bất đẳng thức trên là phần không gian $OABCD$ trong đó các đỉnh O, A, B, C, D lần lượt có tọa độ $O(0,0,0); A(1,0,1); B(0,1,1); C(1,1,0); D(1,1,1)$. Xác suất cần tìm bằng tỉ số thể tích giữa khối đa diện $OABCD$ và hình hộp. Tỉ số đó bằng $\frac{1}{2}$.

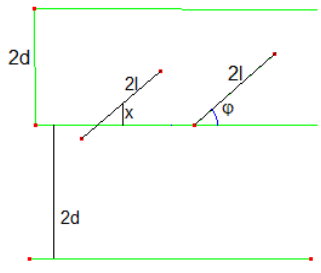
Ví dụ 1.3.12 (Bài toán gieo cái kim của Buffon)

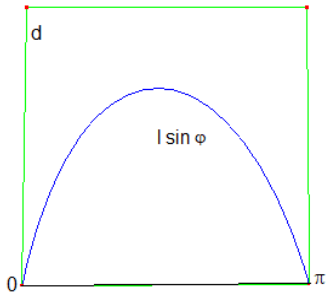
Người ta gieo một cái kim có độ dài $2l$ xuống mặt phẳng sàn nhà. Mặt phẳng sàn nhà chứa các đường thẳng song song cách đều nhau một khoảng bằng $2d$ (giả thiết $l < d$). Hãy tính xác suất để kim cắt một đường thẳng nào đó.

Chúng ta sẽ giải bài toán như sau: để xác định kim cắt hay không cắt một đường thẳng nào đó trong họ các đường thẳng song song, trước hết ta xác định một hướng cho họ các đường thẳng song song đó. Vị trí của kim sẽ được xác định bởi 2 yếu tố : thứ nhất là góc φ giữa kim và chiều dương của họ các đường thẳng song song ($\varphi \in [0, \pi]$) và thứ hai là khoảng cách x từ trung điểm của kim tới đường thẳng gần nhất ($x \in [0, d]$). Vậy khi gieo kim xuống mặt phẳng sàn nhà, các cặp số (φ, x) có thể được biểu diễn như là một điểm được chọn ngẫu nhiên trong hình chữ nhật $[0, \pi] \times [0, d]$. Gọi A là biến cố kim cắt một đường thẳng nào đó, biến cố A xảy ra khi và chỉ khi

$$0 \leq x \leq l \sin \varphi$$

Tập hợp những điểm (φ, x) của hình chữ nhật $[0, \pi] \times [0, d]$ thoả mãn bất đẳng thức trên - để thuận tiện - ta cũng kí hiệu là A.





Phương pháp hình học dựa trên giả thiết: xác suất của A bằng tỉ số giữa diện tích miền phẳng A với diện tích hình chữ nhật $[0, \pi] \times [0, d]$. Điều đó về trực giác có thể được diễn đạt là mọi hướng của kim có khả năng xảy ra như nhau và mọi vị trí của tâm kim cũng vậy. Diện tích của miền A bằng

$$\int_0^\pi l \sin\varphi \, d\varphi = 2l \qquad \Rightarrow \qquad P(A) = \frac{2l}{\pi d}.$$

Từ ví dụ 1.3.12 ta có nhận xét sau:

Nhận xét 1.3.4 *Nếu chúng ta tiến hành phép thử ngẫu nhiên gieo kim n lần và kí hiệu số lần xảy ra biến cố A (số lần kim cắt một đường thẳng nào đó) là k. Khi đó tỉ số $\frac{k}{n}$ là tần suất xuất hiện của biến cố A và theo luật số lớn*

$$\frac{k}{n} \cong P(A) = \frac{2l}{\pi d} \qquad \textit{khi } n \textit{ đủ lớn}$$

Từ đây suy ra có thể tính gần đúng số $\pi \cong \frac{2nl}{kd}$.

Chú ý rằng để tăng độ chính xác của số π trong cách tính này, ta phi tăng số lần gieo kim lên hàng nghìn lần và do vậy trong thực tế người ta sử dụng những phưng pháp khác đơn giản hơn rất nhiều để tính số π. Tuy nhiên ngày nay với sự giúp đỡ của máy tính, công việc gieo kim hàng nghìn lần tương đương với việc tạo hàng nghìn số ngẫu nhiên, có thể được hoàn thành trong vài giây và việc giải các bài toán phức tạp được tính toán gần đúng bằng các phương pháp ngẫu nhiên tương tự (chẳng hạn phương pháp Monte-Carlo có thể giải các bài toán có độ phức tạp cao).

1.4 Xác suất có điều kiện

Trước khi dẫn vào khái niệm quan trọng xác suất có điều kiện, chúng ta hãy nghiên cứu tần suất có điều kiện. Giả sử A và B là hai biến cố ngẫu nhiên, trong đó biến cố B có xác suất lớn hơn 0: $P(B) > 0$. Tiến hành n lần phép

thử ngẫu nhiên để quan sát tần suất xuất hiện của các biến cố đó. Giả sử k_A, k_B kí hiệu số lần xuất hiện của biến cố A và biến cố B tương ứng. Bây giờ ta hạn chế chỉ xét *trong dãy kết quả k_B lần xuất hiện biến cố B, gọi k là số lần xuất hiện biến cố A.* Hiển nhiên k cũng là số lần xuất hiện biến cố tích AB trong dãy kết quả n phép thử ngẫu nhiên ban đầu: $k_{AB} = k$. Tỉ số

$$\frac{k_{AB}}{k_B}$$

được gọi là *tần suất xuất hiện của biến cố A với điều kiện biến cố B xảy ra.* Tần suất có điều kiện đó còn có thể biểu diễn dưới dạng:

$$\frac{k_{AB}}{k_B} = \frac{k_{AB} : n}{k_B : n}$$

ở mục trước, khi dẫn vào khái niệm xác suất của biến cố ngẫu nhiên, chúng ta biết rằng các tần suất $\frac{k_{AB}}{n}$, $\frac{k_B}{n}$ tiến dần đến các xác suất $P(AB)$ và $P(B)$ tương ứng khi $n \to \infty$. Do vậy tần suất có điều kiện $\frac{k_{AB}}{k_B} \to \frac{P(AB)}{P(B)}$ và người ta sử dụng biểu thức $\frac{P(AB)}{P(B)}$ như là định nghĩa của xác suất có điều kiện.

Định nghĩa 1.4.1 *Xác suất của biến cố A với điều kiện biến cố B xảy ra được kí hiệu $P(A/B)$ và*

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

Ví dụ 1.4.1

Trong tổng số 1000 học sinh gồm 600 nam và 400 nữ của một trường học có 100 học sinh giỏi là nam và 100 học sinh giỏi là nữ. Chọn ngẫu nhiên một học sinh của trường. Tính xác suất để học sinh được chọn ra là nữ, với điều kiện học sinh đó là học sinh giỏi.

Gọi A là biến cố học sinh được chọn ra là nữ và B là biến cố học sinh được chọn ra là học sinh giỏi. Khi đó AB là biến cố học sinh được chọn ra là nữ học sinh giỏi. Theo định nghĩa xác suất có điều kiện

$$P(A/B) = \frac{P(AB)}{P(B)} = \frac{100 : 1000}{200 : 1000} = \frac{1}{2}$$

Ví dụ 1.4.2

Từ một hộp đựng 4 bi đỏ và 6 bi trắng, lần lượt rút ngẫu nhiên ra hai viên bi. Tìm xác suất để viên bi được rút ra lần thứ hai là viên bi trắng, với điều kiện lần đầu rút được bi đỏ .

Gọi A là biến cố viên bi được rút ra lần thứ hai là viên bi trắng và B là biến cố viên bi được rút ra lần đầu là bi đỏ . Theo yêu cầu đầu bài, ta phải tính xác suất có điều kiện $P(A/B)$. Dễ dàng thấy $P(B) = \frac{4}{10}$. Theo định nghĩa xác suất có điều kiện, bây giờ ta cần tính $P(AB)$. ở bài toán này, phép thử ngẫu nhiên là lần lượt chọn ra 2 viên bi từ một hộp đựng 10 viên bi, do vậy số các biến cố ngẫu nhiên cơ bản đồng khả năng là số các chỉnh hợp chập 2 của 10: $A_{10}^2 = 10 \cdot 9$. Số trường hợp thuận lợi cho biến cố AB là $4 \cdot 6 = 24$. Vậy $P(AB) = \frac{24}{90}$. Suy ra xác suất cần tìm

$$P(A/B) = \frac{\frac{24}{90}}{\frac{4}{10}} = \frac{2}{3}.$$

Nhận xét 1.4.1 *Dễ dàng nhận thấy xác suất có điều kiện thoả mãn các tính chất sau:*

Đặt $P^*(A) = P(A/B)$ với A là biến cố ngẫu nhiên bất kì.

1. *Với mọi biến cố ngẫu nhiên A* $\quad 0 \leq P^*(A) \leq 1$

2. $P^*(\Omega) = 1$

3. *Nếu $A_1, A_2, ..., A_i, ...$ là các biến cố ngẫu nhiên đôi một xung khắc nhau, khi đó*

$$P^*\left(\sum_i A_i\right) = \sum_i P^*(A_i)$$

Như vậy xác suất có điều kiện thoả mãn các tiên đề về xác suất, suy ra nó cũng lập nên một phân bố xác suất khác trên không gian các biến cố ngẫu nhiên $\mathfrak{A}$. Do vậy trong nhiều bài toán, để tính xác suất có điều kiện của các biến cố ngẫu nhiên một cách nhanh, gọn hơn thay vì sử dụng định nghĩa $P(A/B) = \frac{P(AB)}{P(B)}$, người ta xét bài toán trong một không gian xác suất khác - không gian xác suất có điều kiện - mà phân bố xác suất trên nó thoả mãn các tính chất nêu trên.

Ví dụ khi tính xác suất biến cố ngẫu nhiên bằng phương pháp cổ điển, Giả sử Ω gồm n biến cố ngẫu nhiên cơ bản đồng khả năng. Biến cố B là tập con của Ω gồm n_1 phần tử cũng đồng khả năng. Biến cố AB là tập con của A gồm m phần tử (ta cũng có thể nói m là số phần tử thuộc B của A). Khi đó

$$P(AB) = \frac{m}{n} \quad \text{và} \quad P(B) = \frac{n_1}{n} \qquad \Rightarrow$$
$$P(A/B) = (P^*(A)) = \frac{\frac{m}{n}}{\frac{n_1}{n}} = \frac{m}{n_1}$$

Tỉ số $\frac{m}{n_1}$ cũng có thể được coi là tỉ số giữa số phần tử của A (nằm trong B) với số các phần tử của B:

$$P(A/B) = \frac{\text{số phần tử của } AB}{\text{số phần tử của } B}$$

Chẳng hạn ở ví dụ 1.4.1, xác suất để chọn ngẫu nhiên được học sinh nữ, biết rằng đó là học sinh giỏi là:

$$P(A/B) = \frac{\text{số nữ học sinh giỏi}}{\text{số học sinh giỏi}} = \frac{100}{200} = \frac{1}{2}.$$

Ở ví dụ 1.4.2, biến cố AB: ''lần chọn thứ nhất được bi đỏ và lần chọn thứ hai được bi trắng'' gồm $4 \cdot 6$ biến cố ngẫu nhiên cơ bản, biến cố B: ''lần chọn thứ nhất được bi đỏ'' ($\Rightarrow$ lần chọn thứ hai chọn tùy ý 1 trong 9 viên bi còn lại) gồm $4 \cdot 9$ biến cố ngẫu nhiên cơ bản.Vì vậy:

$$P(A/B) = \frac{4 \cdot 6}{4 \cdot 9} = \frac{6}{9} = \frac{2}{3}.$$

Nhận xét 1.4.2 Ở *ví dụ thứ hai, cách giải trên tương đương với việc xét trên không gian xác suất mà biến cố B đã xảy ra, do vậy trước lần rút thứ hai trong hộp còn 9 viên bi, trong đó có 6 bi trắng và 3 bi đỏ. Ta suy ra*

$$P(A/B) = \frac{6}{9} = \frac{2}{3}.$$

Từ định nghĩa xác suất có điều kiện, hiển nhiên ta có *quy tắc nhân xác suất* $P(AB) = P(A/B)P(B)$. Quy tắc nhân xác suất có thể tổng quát lên tích n biến cố ngẫu nhiên:

Định lí 1.4.1 (Quy tắc nhân xác suất) *Nếu dãy gồm n các biến cố ngẫu nhiên $A_1, A_2, ..., A_n$ thoả mãn điều kiện $P(A_1A_2\cdots A_{n-1}) > 0$. Khi đó:*

$$P(A_1A_2\cdots A_n) =$$

$$= P(A_n/A_1A_2\cdots A_{n-1})P(A_{n-1}/A_1A_2\cdots A_{n-2})\cdots P(A_2/A_1)P(A_1)$$

Chứng minh. Áp dụng quy tắc nhân xác suất nêu trên $n-1$ lần ta được điều phải chứng minh.

$$P(A_1A_2\cdots A_n) = P(A_n/A_1A_2\cdots A_{n-1})P(A_1A_2\cdots A_{n-1}) =$$
$$=P(A_n/A_1A_2\cdots A_{n-1})P(A_{n-1}/A_1A_2\cdots A_{n-2})P(A_1A_2\cdots A_{n-2})$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$=P(A_n/A_1A_2\cdots A_{n-1})P(A_{n-1}/A_1A_2\cdots A_{n-2})\cdots P(A_2/A_1)P(A_1)$$

Ví dụ 1.4.3

Chọn ngẫu nhiên từ cỗ bài tú lơ khơ 52 quân lần lượt 4 quân bài. Tìm xác suất để 2 quân bài chọn ra đầu tiên là 2 quân đỏ (''cơ'' hoặc ''dô''), 2 quân bài chọn ra sau cùng là 2 quân ''nhép''.

Gọi A_1, A_2, A_3, A_4 lần lượt là các biến cố: quân bài chọn ra lần thứ nhất là quân đỏ, quân bài chọn ra lần thứ hai là quân đỏ, quân bài chọn ra lần thứ ba là quân ''nhép'' và quân bài chọn ra lần cuối cùng là quân ''nhép''. Ta có

$$P(A_1) = \frac{26}{52}$$
$$P(A_2/A_1) = \frac{25}{51}$$
$$P(A_3/A_1A_2) = \frac{13}{50}$$
$$P(A_4/A_1A_2A_3) = \frac{12}{49}$$

Áp dụng quy tắc nhân xác suất

$$P(A_1A_2A_3A_4) = P(A_4/A_1A_2A_3)\cdot P(A_3/A_1A_2)\cdot P(A_2/A_1)\cdot P(A_1)$$
$$= \frac{12}{49}\cdot\frac{13}{50}\cdot\frac{25}{51}\cdot\frac{26}{52} = \frac{13}{833} = 0,0156.$$

Ví dụ 1.4.4

Một cỗ bài tú lơ khơ gồm 3 quân ''nhép'' và 5 quân ''dô". Hai người chơi một trò chơi như sau: mỗi người lần lượt chọn ngẫu nhiên từ cỗ bài đó 1 quân (quân bài đã rút ra không đặt trở lại cỗ bài). Họ quy ước ai rút được quân ''dô" trước thì thắng cuộc. Hãy tìm xác suất người rút trước thắng cuộc (Giả sử người thứ nhất là người rút trước).

Gọi A là biến cố người rút trước thắng cuộc. Khi đó biến cố đối lập $\overline{A}$ là biến cố người rút sau thắng cuộc. Do cỗ bài chỉ có 3 quân ''nhép'' nên trò chơi sẽ kết thúc sau tối đa 3 lần rút ngẫu nhiên.

Gọi A_1 là biến cố người thứ nhất rút lần đầu được ngay quân ''dô"

A_2 là biến cố người thứ nhất rút lần thứ hai được quân ''dô"

B_1 là biến cố người thứ hai rút lần đầu được quân ''dô".

Biến cố $\overline{A_1} \cdot \overline{B_1} \cdot A_2$ là biến cố: người thứ nhất rút lần thứ nhất được quân ''nhép'', người thứ hai rút lần thứ nhất được quân ''nhép'' và sau đó người thứ nhất rút lần thứ hai được quân ''dô''. Nói cách khác $\overline{A_1} \cdot \overline{B_1} \cdot A_2$ là biến cố ''người thứ nhất thắng cuộc ở lần rút thứ hai của người đó''.

Do tổng số quân ''nhép'' trong cỗ bài bằng 3, suy ra

$$A = A_1 + \overline{A_1} \cdot \overline{B_1} \cdot A_2$$

là biến cố người thứ nhất (người rút trước) thắng cuộc. Đây cũng là tổng của 2 biến cố xung khắc nhau

$$\Rightarrow P(A) = P(A_1) + P(\overline{A_1} \cdot \overline{B_1} \cdot A_2)$$

Hiển nhiên $P(A_1) = \frac{5}{8}$, áp dụng quy tắc nhân xác suất

$$P(\overline{A_1} \cdot \overline{B_1} \cdot A_2) = P(A_2/\overline{A_1} \cdot \overline{B_1})P(\overline{B_1}/\overline{A_1})P(\overline{A_1}) = \frac{5}{6} \cdot \frac{2}{7} \cdot \frac{3}{8}$$

$$\Rightarrow \ P(A) = \frac{5}{8} + \frac{5}{6} \cdot \frac{2}{7} \cdot \frac{3}{8} = \frac{5}{7}.$$

Từ khái niệm xác suất có điều kiện, ta có định lí quan trọng sau đây

Định lí 1.4.2 (Định lí xác suất đầy đủ) *Nếu $A_1, A_2, ..., A_n$ là hệ đầy đủ các biến cố ngẫu nhiên, trong đó $P(A_i) > 0, \quad i = 1, 2, ..., n$. A là biến cố ngẫu nhiên bất kì, khi đó*

$$P(A) = \sum_{i=1}^{n} P(A/A_i)P(A_i)$$

Chứng minh. Do $A_1, A_2, ..., A_n$ là hệ đầy đủ các biến cố ngẫu nhiên

$$\Omega = \sum_{i=1}^{n} A_i \quad \Rightarrow \quad A\Omega = \sum_{i=1}^{n} AA_i,$$

trong đó các biến cố $AA_1, AA_2, ..., AA_n$ đôi một xung khắc nhau. Suy ra

$$P(A) = \sum_{i=1}^{n} P(AA_i)$$

Áp dụng quy tắc nhân xác suất để tính $P(AA_i)$ ta được

$$P(A) = \sum_{i=1}^{n} P(A/A_i)P(A_i), \quad \text{đ.p.c.m.}$$

Nhận xét rằng định lí có thể mở rộng cho hệ đầy đủ vô hạn các biến cố ngẫu nhiên $A_1, A_2, ...$

$$P(A) = \sum_{i=1}^{\infty} P(A/A_i)P(A_i).$$

Ví dụ 1.4.5

Một nhà máy gồm 3 phân xưởng cùng sản xuất bóng đèn. Phân xưởng I sản xuất 50%, phân xưởng II sản xuất 30% và phân xưởng III sản xuất 20% tổng số bóng đèn của toàn nhà máy. Tỉ lệ phế phẩm của các phân xưởng tương ứng là 2%, 3%, 4%. Hãy tính tỉ lệ phế phẩm chung của toàn nhà máy.

Trước hết chúng ta cần làm rõ khái niệm tỉ lệ phế phẩm theo ngôn ngữ xác suất. Chẳng hạn, ta nói tỉ lệ phế phẩm của phân xưởng I là 2% chẳng hạn, điều đó có nghĩa là nếu chọn ngẫu nhiên một sản phẩm từ lô

hàng do phân xưởng I sản xuất, xác suất để sản phẩm được chọn ra là phế phẩm bằng 0,02 hoặc tương tự nếu chọn ngẫu nhiên một sản phẩm từ lô hàng chung của toàn nhà máy (tức là các sản phẩm của các phân xưởng đã được trộn lẫn và dán mác sản xuất của nhà máy - không phân biệt sản phẩm đó là do phân xưởng nào sản xuất) khi đó xác suất để sản phẩm được chọn ra là phế phẩm chính là tỉ lệ phế phẩm chung của toàn nhà máy.

Quay lại ví dụ của chúng ta, gọi A là biến cố chọn ngẫu nhiên một sản phẩm từ lô hàng chung của toàn nhà máy, ta được phế phẩm. Theo đề bài, ta phải tính $P(A)$.

Giả sử A_1 là biến cố chọn ngẫu nhiên một sản phẩm từ lô hàng chung của toàn nhà máy, ta được sản phẩm chọn ra là sản phẩm do phân xưởng I sản xuất. A_2, A_3 là các biến cố khi chọn ngẫu nhiên một sản phẩm từ lô hàng chung của toàn nhà máy, ta được sản phẩm chọn ra là sản phẩm của phân xưởng II, III tương ứng. Khi đó hiển nhiên các biến cố A_1, A_2, A_3 lập thành một hệ đầy đủ và theo bài ra

$$P(A_1) = 0,5 \quad P(A_2) = 0,3 \quad P(A_3) = 0,2$$

$$P(A/A_1) = 0,02 \quad P(A/A_2) = 0,03 \quad P(A/A_3) = 0,04$$

Theo định lí xác suất đầy đủ

$$P(A) = P(A/A_1)P(A_1) + P(A/A_2)P(A_2) + P(A/A_3)P(A_3) = \\ = 0,5 \cdot 0,02 + 0,3 \cdot 0,03 + 0,2 \cdot 0,04 = 0,027$$

Vậy tỉ lệ phế phẩm chung của toàn nhà máy là 2,7%.

Ví dụ 1.4.6

Có hai hộp: Hộp I đựng 7 bi đỏ và 3 bi trắng. Hộp II đựng 6 bi đỏ, 3 bi trắng. Rút ngẫu nhiên từ hộp I ra một viên bi và bỏ sang hộp II, sau đó lấy ngẫu nhiên từ hộp II ra một viên bi và bỏ sang hộp I. Cuối cùng rút ngẫu nhiên từ hộp I ra một viên bi.

a) Tìm xác suất để viên bi được rút ra lần thứ hai (từ hộp II) là viên bi trắng.

b) Tìm xác suất để viên bi được rút ra lần thứ nhất (từ hộp I) là viên bi đỏ, với điều kiện viên bi được rút ra lần thứ hai (từ hộp II) là viên bi trắng.

c) Tìm xác suất để viên bi được rút ra lần cuối cùng (từ hộp I) là viên bi đỏ.

Gọi A_i là các biến cố ở lần rút thứ $i,\ (i = 1, 2, 3)$ ta rút được bi đỏ và B_i là các biến cố ở lần rút thứ $i,\ (i = 1, 2, 3)$ ta rút được bi trắng.

a) Ta phải tính xác suất của biến cố B_2. Các biến cố A_1 (rút lần thứ nhất được bi đỏ), B_1 (rút lần thứ nhất được bi trắng) lập thành một hệ đầy đủ.

$$P(A_1) = \frac{7}{10}, P(B_1) = \frac{3}{10}$$

Dễ dàng tính các xác suất điều kiện

$$P(B_2/A_1) = \frac{3}{10} \quad P(B_2/B_1) = \frac{4}{10}.$$

Áp dụng định lí xác suất đầy đủ

$$P(B_2) = P(B_2/A_1)P(A_1) + P(B_2/B_1)P(B_1) =$$

$$= \frac{3}{10} \cdot \frac{7}{10} + \frac{4}{10} \cdot \frac{3}{10} = \frac{33}{100}$$

b) Ta phải tính xác suất có điều kiện $P(A_1/B_2)$. Xác suất có điều kiện $P(A_1/B_2)$ không dễ dàng tính trực tiếp như xác suất $P(B_2/A_1)$ ở phần a). Tuy nhiên ta có thể tính nó thông qua $P(B_2/A_1)$ bằng cách áp dụng quy tắc nhân xác suất

$$P(A_1/B_2) = \frac{P(B_2/A_1)P(A_1)}{P(B_2)} = \left(\frac{3}{10} \cdot \frac{7}{10}\right) : \frac{33}{100} = \frac{7}{11}.$$

c) Hiển nhiên các biến cố $A_1A_2, B_1B_2, A_1B_2, B_1A_2$ cũng lập thành một hệ đầy đủ.

$$P(A_1A_2) = P(A_2/A_1)P(A_1) = \frac{7}{10} \cdot \frac{7}{10} = \frac{49}{100}$$

$$P(B_1B_2) = P(B_2/B_1)P(B_1) = \frac{4}{10} \cdot \frac{3}{10} = \frac{12}{100}$$

$$P(A_1B_2) = P(B_2/A_1)P(A_1) = \frac{3}{10} \cdot \frac{7}{10} = \frac{21}{100}$$

$$P(B_1A_2) = P(A_2/B_1)P(B_1) = \frac{6}{10} \cdot \frac{3}{10} = \frac{18}{100}$$

Theo công thức xác suất đầy đủ:

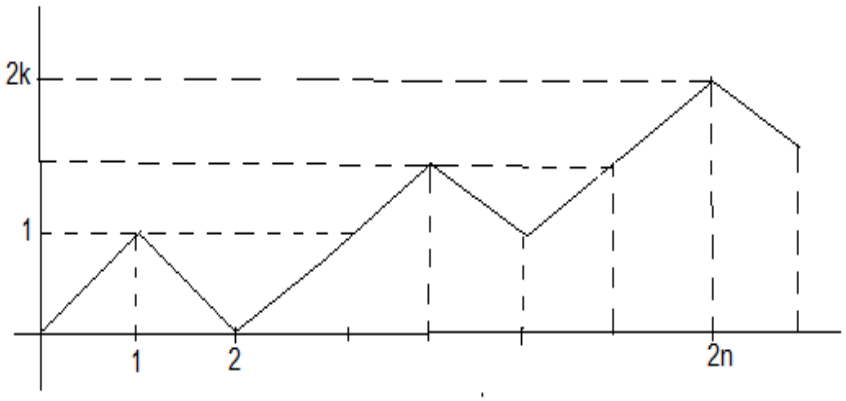
$$\begin{aligned} P(A_3) &= P(A_3/A_1A_2)P(A_1A_2) + P(A_3/B_1B_2)P(B_1B_2) + \\ &+ P(A_3/A_1B_2)P(A_1B_2) + P(A_3/B_1A_2)P(B_1A_2) = \\ &= \frac{7}{10} \cdot \frac{49}{100} + \frac{7}{10} \cdot \frac{12}{100} + \frac{6}{10} \cdot \frac{21}{100} + \frac{8}{10} \cdot \frac{18}{100} = 0,697. \end{aligned}$$

Ví dụ 1.4.7

Một điểm M nằm ở vị trí có hoành độ là các số nguyên trên trục số (Giả sử ban đầu M nằm ở vị trí $x = 0$ trên trục số). M dịch chuyển sang trái hoặc sang phải một đơn vị với xác suất như nhau. (Nói cách khác nó bước sang trái với xác suất $\frac{1}{2}$ và bước sang phải cũng với xác suất bằng $\frac{1}{2}$). Tìm xác suất để sau n bước M sẽ tới vị trí $x = k$.

Chú ý rằng n và k phải cùng chẵn hoặc cùng lẻ. Để thuận tiện về mặt kí hiệu, ta chỉ giải bài toán tìm xác suất để sau $2n$ bước M sẽ tới vị trí $x = 2k$.

Ta sẽ minh hoạ vị trí của điểm M lên mặt phẳng toạ độ: M có toạ độ (x, y) nếu sau x bước M dịch chuyển từ điểm gốc 0 tới điểm y trên trục số. Ban đầu M nằm tại gốc toạ độ $O(0, 0)$ và gọi A là biến cố sau $2n$ bước M dịch chuyển tới điểm $(2n, 2k)$.



Số trường hợp đồng khả năng của biến cố A chính là số đường đi có thể có từ O tới điểm $(2n, 2k)$: số đường đi đó bằng C_{2n}^{n-k} ($n + k$ bước

sang phải và $n-k$ bước sang trái). Tổng số đường đi lang thang có thể là 2^{2n}. Vậy tỉ số của chúng là xác suất cần tìm: $P(A) = \frac{C_{2n}^{n-k}}{2^{2n}}$.

Bài toán trên thuộc lớp các bài toán "lang thang ngẫu nhiên". Bây giờ ta xét một bài toán lang thang ngẫu nhiên khác.

Ví dụ 1.4.8

Giả sử ban đầu điểm M nằm ở vị trí $x = k, 0 < k < n$, k và n là các số tự nhiên. M dịch chuyển sang trái hoặc sang phải một đơn vị với xác suất như nhau và bằng $\frac{1}{2}$. Nếu sau một số bước nào đó điểm M đạt tới vị trí $x = 0$ hoặc $x = n$ trên trục số, nó sẽ dừng tại đó và không tiếp tục lang thang nữa (người ta gọi hai vị trí 0 và n là các *''điểm hút''*). Hãy tính xác suất để điểm M dịch chuyển tới *''điểm hút''* 0.

Gọi A là biến cố sau một số bước nào đó M từ vị trí k ban đầu dịch chuyển tới ''điểm hút'' 0 trên trục số, B là biến cố M bước sang trái. Kí hiệu $p_k = P(A)$ là xác suất cần tìm. Theo định lí xác suất đầy đủ:

$$p_k = P(A) = P(A/B)P(B) + P(A/\overline{B})P(\overline{B})$$

Chú ý rằng do B là biến cố M bước sang trái, nên xác suất có điều kiện $P(A/B)$ chính là xác suất để M dịch chuyển từ vị trí $k-1$ tới "điểm hút" 0, hay $p_{k-1} = P(A/B)$. Tương tự $p_{k+1} = P(A/\overline{B})$. Đẳng thức trên có thể viết dưới dạng sau

$$p_k = \frac{1}{2}p_{k-1} + \frac{1}{2}p_{k+1} \quad \text{hay} \quad p_{k-1} - p_k = p_k - p_{k+1}$$

Như vậy với mọi $k = 1, 2, ..., n-1$:

$$p_0 - p_1 = p_1 - p_2 = \cdots = p_{k-1} - p_k = \cdots = p_{n-1} - p_n$$

Lưu ý rằng các vị trí 0 và n là các "điểm hút" nên

$$p_0 = 1, p_n = 0 \quad \Rightarrow$$

$$p_{k-1} - p_k = \frac{p_0 - p_n}{n} \quad \text{hay} \quad p_k = 1 - \frac{k}{n}.$$

Nhận xét 1.4.3 *Bài toán lang thang ngẫu nhiên trên còn có tên là bài toán người đánh bạc phá sản:*

Giả sử một người có số tiền ban đầu là k đồng. Người đó đánh bạc với hy vọng sẽ kiếm đủ n đồng thì dừng. Thắng mỗi ván bài người đó được 1 đồng và nếu thua người đó mất 1 đồng. (Người chơi bạc bị phá sản nếu sau một số ván chơi người đó hết sạch tiền. Như vậy 0 và n là hai điểm hút như đã định nghĩa trong ví dụ 1.4.8). Theo kết quả trên, xác suất để người đó phá sản (về điểm hút 0) bằng

$$p_k = 1 - \frac{k}{n}.$$

Hiển nhiên nếu số tiền ban đầu (k đồng) càng ít so với số tiền cần kiếm (n đồng) thì khả năng phá sản của người đó càng lớn.

Một định lí có nhiều ứng dụng trong các lĩnh vực khác nhau và gắn liền với định lí xác suất đầy đủ:

Định lí 1.4.3 (Định lí Bayes) *Nếu các biến cố $A_1, A_2, ..., A_n$ lập thành một hệ đầy đủ, A là biến cố bất kì. Khi đó với mọi $k = 1, .., n$:*

$$P(A_k/A) = \frac{P(A/A_k)P(A_k)}{\sum_{i=1}^{n} P(A/A_i)P(A_i)}$$

Chứng minh. Từ định nghĩa xác suất có điều kiện suy ra

$$P(A_k/A) = \frac{P(A/A_k)P(A_k)}{P(A)}$$

Mặt khác theo định lí xác suất đầy đủ

$$P(A) = \sum_{i=1}^{n} P(A/A_i)P(A_i) \quad \Rightarrow$$

$$P(A_k/A) = \frac{P(A/A_k)P(A_k)}{\sum_{i=1}^{n} P(A/A_i)P(A_i)} \qquad \text{đ.p.c.m.}$$

Ví dụ 1.4.9

Quay trở lại ví dụ 1.4.5, một nhà máy gồm 3 phân xưởng cùng sản xuất bóng đèn. Phân xưởng I sản xuất 50%, phân xưởng II sản xuất 30% và phân xưởng III sản xuất 20% tổng số bóng đèn của toàn nhà máy. Tỉ lệ phế phẩm của các phân xưởng tương ứng là 2%, 3%, 4%.

Giả sử các bóng đèn của các phân xưởng được trộn đều và nhập vào kho chung của toàn nhà máy. Lấy ngẫu nhiên một sản phẩm từ kho chung của nhà máy để kiểm tra, người ta chọn phải bóng đèn phế phẩm. Hãy tính xác suất để bóng đèn phế phẩm đó là do phân xưởng I sản xuất.

Giữ nguyên các kí hiệu như trong ví dụ 1.4.5, gọi A là biến cố bóng được chọn ra là phế phẩm. A_1, A_2, A_3 là biến cố sản phẩm được chọn ra là sản phẩm do phân xưởng I, II, III tương ứng sản xuất. Yêu cầu của bài toán phải tính xác suất có điều kiện $P(A_1/A)$. Theo định lí Bayes

$$P(A_1/A) = \frac{P(A/A_1)P(A_1)}{\sum_{i=1}^{3} P(A/A_i)P(A_i)} = \\ = \frac{P(A/A_1)P(A_1)}{P(A)} = \frac{0,5 \cdot 0,02}{0,027} = 0,37037...$$

Hoàn toàn tương tự, ta có thể tính các xác suất

$$P(A_2/A) = 0,333..., P(A_3/A) = 0,296...$$

Định lí Bayes cho ta phương pháp tính xác suất có điều kiện $P(A_i/A)$, sau khi có thêm thông tin biến cố A xảy ra. Định lí còn mang tên ''*định lí về các nguyên nhân*''. Xuất xứ tên của định lí từ các ứng dụng của định lí đó. Người ta thường áp dụng định lí Bayes khi muốn tính xác suất của các giả thiết (hypothesis) A_i, sau khi biết thông tin: biến cố A đã xảy ra. Nói cách khác người ta muốn khảo sát việc xảy ra của biến cố A đã tác động như thế nào đến các giả thiết A_i. Các xác suất $P(A_i)$ được gọi là các xác suất *tiên nghiệm (priori)* và các xác suất $P(A_i/A)$ được gọi là các xác suất *hậu nghiệm (posteri)*. Trong nhiều ứng dụng thực tế các xác suất $P(A_i)$ nói chung chưa biết, do vậy người ta thường gán cho nó các giá trị nào đó để áp dụng định lí Bayes.

1.5 Sự độc lập của các biến cố ngẫu nhiên

Cho A và B là hai biến cố ngẫu nhiên, giả thiết rằng $P(A) > 0$ và $P(B) > 0$. Trong mục trước ta đã định nghĩa xác suất có điều kiện $P(A/B)$, nói chung xác suất đó khác $P(A)$. Trường hợp khi

$$P(A/B) = P(A),$$

ta nói biến cố A *độc lập* với biến cố B. Từ định nghĩa xác suất có điều kiện

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

ta suy ra nếu

$$P(A/B) = P(A) \text{ thì } P(B/A) = P(B)$$

Nói một cách đơn giản hai biến cố ngẫu nhiên A và B *độc lập với nhau* khi và chỉ khi $P(A/B) = P(A)$ hoặc $P(AB) = P(A)P(B)$.

Nhận xét rằng nếu hai biến cố A và B độc lập với nhau, khi đó A và $\overline{B}$ (tương tự $\overline{A}$ và B, hoặc $\overline{A}$ và $\overline{B}$) cũng độc lập với nhau. Thật vậy

$$\begin{aligned} P(A\overline{B}) &= P(A \backslash AB) = P(A) - P(AB) = \\ &= P(A) - P(A)P(B) = P(A)P(\overline{B}) \end{aligned}$$

Chúng ta sẽ mở rộng khái niệm độc lập cho nhiều biến cố ngẫu nhiên. Tuy nhiên ta có nhận xét rằng nếu A, B, C là các biến cố ngẫu nhiên độc lập từng đôi một:

$$P(AB) = P(A)P(B); P(AC) = P(A)P(C); P(BC) = P(B)P(C)$$

khi đó ta chưa đủ cơ sở để suy ra biến cố AB và biến cố C độc lập với nhau. Ví dụ sau minh họa cho nhận xét đó.

Ví dụ 1.5.1

Ví dụ xét bài toán gieo đồng thời hai xúc xắc, gọi A là biến cố xúc xắc thứ nhất xuất hiện mặt chẵn, B là biến cố xúc xắc thứ hai xuất hiện mặt lẻ và C là biến cố cả hai xúc xắc cùng xuất hiện mặt chẵn hoặc mặt lẻ. Dễ dàng tính được các xác suất sau

$$P(A) = P(B) = P(C) = \frac{1}{2}; P(AB) = P(BC) = P(AC) = \frac{1}{4}$$

Theo định nghĩa sự độc lập của hai biến cố ngẫu nhiên, A, B, C độc lập từng đôi một. Mặt khác, biến cố tích $ABC = \emptyset$ nên

$$P(ABC) = 0 \quad \Rightarrow \quad P((AB)C) \neq P(AB)P(C)$$

Từ đó cũng suy ra

$$P(ABC) \neq P(A)P(B)P(C).$$

Ta dẫn vào định nghĩa sau:

Định nghĩa 1.5.1 *A, B, C được gọi là các biến cố ngẫu nhiên độc lập (hoặc hoàn toàn độc lập) nếu*

$$\begin{cases} P(AB) = P(A)P(B) \\ P(BC) = P(B)P(C) \\ P(AC) = P(A)P(C) \\ P(ABC) = P(A)P(B)P(C). \end{cases}$$

Ba đẳng thức đầu khẳng định các biến cố A, B, C đôi một độc lập và đẳng thức cuối khẳng định một biến cố bất kì độc lập với tích của hai biến cố kia. Một cách tổng quát ta có

Định nghĩa 1.5.2 *Các biến cố ngẫu nhiên $A_1, A_2, ..., A_n$ được gọi là độc lập (hoặc hoàn toàn độc lập) nếu với bất kì k biến cố đôi một khác nhau $A_{i_1}, A_{i_2}, ..., A_{i_k} \quad k = 2, 3, ..n$ trong số n biến cố đã cho, ta luôn có*

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

Định lí 1.5.1 *giả sử các biến cố ngẫu nhiên $A_1, A_2, ..., A_n$ hoàn toàn độc lập. Nếu thay một biến cố bất kì trong số đó bằng biến cố đối lập, khi đó ta vẫn được một hệ các biến cố hoàn toàn độc lập. Nói cách khác với mọi $k = 1, 2, ...n$*

$$A_1, ..., A_{k-1}, \overline{A_k}, A_{k+1}, ..., A_n$$

là hệ các biến cố độc lập.

(*Chú ý rằng từ định lí trên suy ra nếu thay một số hữu hạn các biến cố bất kì bằng các biến cố đối lập với chúng trong hệ các biến cố độc lập, ta vẫn được một hệ độc lập*).

Chứng minh. Dựa vào định nghĩa hệ các biến cố độc lập nói trên, định lí được suy ra từ nhận xét sau:

giả sử $B_1, B_2, ..., B_r, B$ độc lập:

$$P(B_1B_2...B_rB) = P(B_1)P(B_2)...P(B_r)P(B)$$

Do

$$B_1B_2...B_rB + B_1B_2...B_r\overline{B} = B_1B_2...B_r(B + \overline{B}) = B_1B_2...B_r$$

suy ra

$$\begin{aligned} &P(B_1B_2...B_r\overline{B}) = P(B_1B_2...B_r) - P(B_1B_2...B_rB) = \\ &= P(B_1)P(B_2)...P(B_r) - P(B_1)P(B_2)...P(B_r)P(B) = \\ &= P(B_1)P(B_2)...P(B_r)(1 - P(B)) = P(B_1)P(B_2)...P(B_r)P(\overline{B}) \end{aligned}$$

Chú ý rằng chúng ta thường sử dụng tính độc lập định nghĩa trên đây để xác định xác suất của biến cố tích các biến cố ngẫu nhiên độc lập.

Chẳng hạn khi gieo đồng thời 2 xúc xắc, gọi A là biến cố mặt chẵn xuất hiện ở xúc xắc thứ nhất và B là biến cố số chấm lớn hơn 4 xuất hiện ở xúc xắc thứ hai. Hiển nhiên biến cố A có xác suất $P(A) = \frac{1}{2}$ và biến cố B có xác suất$P(B) = \frac{2}{6} = \frac{1}{3}$. Nếu chúng ta giả thiết rằng kết quả xuất hiện ở hai xúc xắc độc lập nhau, khi đó biến cố tích AB có xác suất

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

Điều này dựa vào một thực tế thường xuyên được sử dụng để tính xác suất sau đây:

Nếu ta tiến hành hai phép thử ngẫu nhiên độc lập với nhau, A là biến cố liên quan tới phép thử thứ nhất và B là biến cố liên quan tới phép thử thứ hai, khi đó chúng ta sẽ coi hai biến cố A và B độc lập với nhau. Suy ra

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

Điều đó hoàn toàn phù hợp với thực tế và dễ dàng kiểm tra bằng việc so sánh các tần suất xuất hiện của chúng. Khi chúng ta nói các phép thử ngẫu nhiên được tiến hành ''độc lập'' với nhau, ta hiểu khái niệm ''độc lập'' theo nghĩa thông thường: chúng không có mối quan hệ, không liên quan gì với nhau. Trong ứng dụng thực tế hệ các biến cố mà mỗi biến cố liên quan tới một phép thử ngẫu nhiên trong dãy các phép thử ngẫu nhiên được tiến hành độc lập tạo thành hệ các biến cố hoàn toàn độc lập. Đây chính là cầu nối giữa lí thuyết và ứng dụng thực tế.

Ví dụ 1.5.2

Một xạ thủ bắn súng bắn hai phát đạn vào bia. giả sử xác suất bắn trúng bia trong mỗi lần bắn của xạ thủ đó là $0,8$. Hãy tìm xác suất để trong hai lần bắn đó có ít nhất một viên đạn trúng bia.

Gọi A là biến cố xạ thủ bắn trúng bia ở lần bắn thứ nhất và B là biến cố xạ thủ bắn trúng bia ở lần bắn thứ hai, khi đó $A+B$ là biến cố có ít nhất một viên đạn trúng bia.

$$P(A+B) = P(A) + P(B) - P(AB)$$

Chúng ta có thể giả thiết hai lần bắn của xạ thủ đó độc lập nhau, suy ra biến cố A và B là hai biến cố độc lập. Do đó

$$P(AB) = P(A)P(B) = 0,8 \cdot 0,8 = 0,64 \quad \text{vậy}$$

$$P(A+B) = 0,8 + 0,8 - 0,64 = 0,96.$$

Ví dụ 1.5.3

Một xạ thủ tập bắn bia, bắn liên tục vào bia cho đến khi trúng bia thì dừng. Giả sử xác suất bắn trúng bia của xạ thủ đó bằng p. Hãy tìm xác suất để xạ thủ đã sử dụng n viên đạn trong lần bắn tập đó.

Biến cố xạ thủ đã sử dụng n viên đạn được hiểu là $n-1$ lần bắn đầu, xạ thủ đó bắn trượt và ở lần bắn thứ n xạ thủ đó mới bắn trúng bia. Gọi $A_k, k = 1, 2, ..., n$ là biến cố lần bắn thứ k xạ thủ bắn trúng bia. Khi đó biến cố xạ thủ sử dụng n viên đạn chính là biến cố tích $\overline{A_1} \cdot \overline{A_2} \cdots \overline{A_{n-1}} \cdot A_n$, trong đó $\overline{A_i}$ là biến cố đối lập với biến cố A_i: lần bắn thứ i xạ thủ đó bắn trượt. Theo giả thiết $P(A_i) = p \Rightarrow P(\overline{A_i}) = q = 1-p$. Do các lần bắn độc lập nhau (kết quả của lần bắn này

không ảnh hưởng đến kết quả của lần bắn sau), ta coi các biến cố $\overline{A_1}, \cdots, \overline{A_{n-1}}, A_n$ độc lập hoàn toàn với nhau, suy ra xác suất cần tìm bằng

$$P(\overline{A_1}) \cdot P(\overline{A_2}) \cdots P(\overline{A_{n-1}}) \cdot P(A_n) = q^{n-1}p$$

trong đó $p + q = 1$.

Ví dụ 1.5.4

Hai người chơi một trò chơi như sau: gieo đồng thời 2 xúc xắc, nếu tổng bằng 7 hoặc 11, người thứ nhất thắng cuộc, nếu tổng bằng 2,3 hoặc 12, người thứ hai thắng cuộc. Các trường hợp còn lại, lặp lại trò chơi cho đến khi có người thắng người thua. Tìm xác suất để người thứ nhất thắng.

Bài toán có nhiều cách giải. Ta xét 2 cách giải sau đây:

Cách 1. Gọi A_1 là biến cố tổng số chấm xuất hiện khi gieo hai xúc xắc bằng 7 hoặc 11 (người thứ nhất thắng), A_2 là biến cố tổng số chấm xuất hiện khi gieo hai xúc xắc bằng $2, 3$ hoặc 12 (người thứ hai thắng) và A_3 là biến cố đối lập với tổng 2 biến cố trên (tổng số chấm xuất hiện bằng $4, 5, 6, 8, 9$ hoặc 10). Dễ dàng nhận thấy các biến cố A_1, A_2, A_3 lập thành hệ đầy đủ và

$$P(A_1) = \frac{8}{36}, P(A_2) = \frac{4}{36}, P(A_3) = \frac{24}{36}$$

Gọi A là biến cố người thứ nhất thắng cuộc. Theo định lí xác suất đầy đủ

$$P(A) = P(A/A_1)P(A_1) + P(A/A_2)P(A_2) + P(A/A_3)P(A_3)$$

Hiển nhiên $P(A/A_1) = 1, P(A/A_2) = 0$ trong khi $P(A/A_3) = P(A)$ vì nếu biến cố A_3 xảy ra, trò chơi bắt đầu lại từ đầu bằng việc gieo hai xúc xắc. Vậy

$$P(A) = P(A_1) + P(A)P(A_3) = \frac{8}{36} + P(A)\frac{24}{36}.$$

Suy ra xác suất cần tính (người thứ nhất thắng cuộc) bằng

$$P(A) = \frac{8}{36} : \left(1 - \frac{24}{36}\right) = \frac{8}{12} = \frac{2}{3}.$$

Cách 2. Duy trì các kí hiệu như trên và để thuận tiện ta đặt

$$P(A_1) = p_1, P(A_2) = p_2, P(A_3) = \alpha \quad (p_1 + p_2 + \alpha = 1)$$

trong đó $p_1 = \frac{8}{36}, p_2 = \frac{4}{36}$. Ta nhận thấy người thứ nhất có thể thắng ở các lần gieo thứ nhất, thứ hai, ... hoặc lần gieo xúc xắc thứ n, $n = 1, 2, ...$ Chính xác hơn người thứ nhất thắng cuộc ở lần gieo thứ n nếu biến cố A_3 xảy ra ở các lần gieo thứ nhất, thứ hai..., thứ $n-1$ và ở lần gieo thứ n biến cố A_1 xuất hiện. Xác suất của biến cố đó, lập luận như ví dụ 1.5.3, bằng

$$P(A_3)^{n-1}P(A_1) = \alpha^{n-1}p_1$$

Xác suất để người thứ nhất thắng cuộc bằng tổng các xác suất nói trên

$$P(A) = \sum_{n=1}^{\infty} \alpha^{n-1}p_1 = \frac{p_1}{1-\alpha} = \frac{p_1}{p_1+p_2} = \frac{2}{3}.$$

1.6 Công thức Bernoulli

Công thức Bernoulli có xuất xứ từ bài toán cơ bản sau đây:

Giả sử A là biến cố ngẫu nhiên có xác suất bằng p. Tiến hành n phép thử ngẫu nhiên độc lập nhau để quan sát biến cố A. Tìm xác suất để có đúng k lần xảy ra biến cố A trong số n lần tiến hành phép thử ngẫu nhiên kể trên.

Để giải bài toán này, tươngtự như ví dụ 1.5.3, ta đưa vào các kí hiệu sau: A_k là biến cố ngẫu nhiên ở lần thử thứ k biến cố A xuất hiện $(k = 1, 2, ..., n)$. Xác suất để lần thử thứ $i_1, i_2, ..., i_k$ biến cố A xảy ra và $n-k$ lần còn lại (lần thử thứ $j_1, j_2, ..., j_{n-k}$) biến cố A không xảy ra, bằng xác suất của biến cố tích

$$A_{i_1}A_{i_2}...A_{i_k}\overline{A}_{j_1}\overline{A}_{j_2}...\overline{A}_{j_{n-k}}$$

Do các lần tiến hành phép thử ngẫu nhiên độc lập nhau nên các biến cố $A_{i_1}, A_{i_2}, ..., A_{i_k}, \overline{A}_{j_1}, \overline{A}_{j_2}, ..., \overline{A}_{j_{n-k}}$ là hệ các biến cố hoàn toàn độc lập. Mặt khác $P(A_i) = p, P(\overline{A_i}) = q \quad (p+q=1)$, vì vậy

$$P(A_{i_1}A_{i_2}...A_{i_k}\overline{A}_{j_1}\overline{A}_{j_2}...\overline{A}_{j_{n-k}}) = p^k q^{n-k}$$

Lưu ý rằng chúng ta không quan tâm tới thứ tự của các lần xảy ra biến cố A trong dãy các phép thử ngẫu nhiên nói trên mà chỉ quan tâm tới số lần xảy ra biến cố A. Để có đúng k lần xảy ra biến cố A, số các khả năng có thể có, chính là số dãy các biến cố

$$A_{i_1}, A_{i_2}, ..., A_{i_k}, \overline{A}_{j_1}, \overline{A}_{j_2}, ..., \overline{A}_{j_{n-k}}$$

hay cũng có thể nói là số cách chọn k chỉ số $\{i_1, i_2, ..., i_k\}$ trong tập các chỉ số $\{1, 2, ..., n\}$. Số cách chọn đó bằng C_n^k (số tổ hợp chập k của n). Vậy xác suất để có đúng k lần xảy ra biến cố A, kí hiệu $P_{k;n}$ bằng

$$P_{k;n} = C_n^k p^k q^{n-k}.$$

Ví dụ 1.6.1

Một lô hàng có tỉ lệ phế phẩm là p. Chọn ngẫu nhiên một sản phẩm để kiểm tra, rồi đặt lại sản phẩm đó vào lô hàng. Ta tiến hành công việc kiểm tra đó n lần (trong thống kê, người ta gọi đó là phép chọn mẫu có hoàn lại). Hãy tìm xác suất để trong số n sản phẩm được chọn có hoàn lại như trên, có đúng k phế phẩm.

Trong ví dụ này, việc chọn có hoàn lại n lần cũng đồng nghĩa với việc tiến hành n lần phép thử ngẫu nhiên độc lập nhau. Từ giả thiết tỉ lệ phế phẩm bằng p, tức là xác suất để mỗi lần chọn được phế phẩm bằng p, áp dụng Công thức Bernulli suy ra xác suất cần tìm: có đúng k phế phẩm trong số n sản phẩm được chọn ra là: $P_{k;n} = C_n^k p^k q^{n-k}$

Ví dụ 1.6.2

Hai đấu thủ bóng bàn vào chung kết tranh giải vô địch. Họ phải đấu với nhau tối đa là 5 ván (người nào thắng trước đối phương 3 ván sẽ là người đoạt chức vô địch). Giả sử xác suất để đấu thủ thứ nhất thắng đấu thủ thứ hai ở mỗi ván là p. Tính xác suất để đấu thủ thứ nhất đoạt chức vô địch (giả thiết rằng các ván họ đấu với nhau là độc lập).

Gọi A là biến cố đấu thủ thứ nhất đoạt chức vô địch. Ba khả năng có thể xảy ra:

a) Hai đấu thủ chỉ đấu với nhau 3 ván và đấu thủ thứ nhất thắng cả 3 ván đó. Gọi A_1 là biến cố như vậy.

b) Hai đấu thủ đấu với nhau 4 ván và đấu thủ thứ nhất thắng 3 trong số 4 ván họ đấu với nhau.

c) Hai đấu thủ phải đấu với nhau 5 ván và đấu thủ thứ nhất thắng 3 trong số 5 ván đó.

Hiển nhiên $P(A_1) = p^3$. Gọi A_2, A_3 là hai biến cố tương ứng với trường hợp b) và c) xảy ra. Chú ý rằng ở cả 3 trường hợp trên đấu thủ thứ nhất luôn luôn thắng đấu thủ thứ hai ở ván cuối cùng. Do vậy để tính xác suất của A_2 ta cần viết A_2 dưới dạng tích của 2 biến cố: $A_2 = BC$ trong đó B là biến cố đấu thủ thứ nhất thắng 2 trong số 3 ván đầu họ đấu với nhau, còn C là biến cố đấu thủ thứ nhất thắng đấu thủ thứ hai ở ván thứ 4. Hiển nhiên B và C là 2 biến cố độc lập, $P(C) = p$. Áp dụng Công thức Bernoulli cho biến cố B:

$$P(B) = C_3^2 p^2 q \quad \Rightarrow \quad P(A_2) = P(B)P(C) = C_3^2 p^2 q.p = C_3^2 p^3 q$$

Tương tự $P(A_3) = C_4^2 p^3 q^2$

Mặt khác A_1, A_2, A_3 là ba biến cố đôi một xung khắc nhau

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

Vậy xác suất để đấu thủ thứ nhất đoạt chức vô địch

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

$$= p^3 + C_3^2 p^3 q + C_4^2 p^3 q^2 = p^3(1 + 3q + 6q^2).$$

Bảng sau cho một số kết quả tính toán chi tiết với p là xác suất để đấu thủ thứ nhất thắng đấu thủ thứ hai ở mỗi ván và $P(A)$ là xác suất để đấu thủ thứ nhất đoạt chức vô địch.

p	$P(A) = p^3(1 + 3q + 6q^2)$
0,3	0,16308
0,4	0,31744
0,5	0,5
0,6	0,68256
0,7	0,83692

BÀI TẬP CHƯƠNG I

1. Một hộp đựng 3 bi đỏ, 3 bi trắng và 3 bi xanh. Chọn ngẫu nhiên ra 6 viên bi. Tính xác suất để có đủ 3 màu trong số 6 viên bi chọn ra.

2. Chứng minh rằng $C_N^n = \sum_{k=1}^{N-n+1} C_{N-k}^{n-1} \quad (n \leq N)$.

3. Gieo đồng thời 3 xúc xắc. Tìm xác suất để tổng bình phương của các số xuất hiện chia hết cho 3.

4. Gieo đồng thời n xúc xắc. Tìm xác suất để tổng các số xuất hiện chia hết cho 6.

5. Gieo đồng thời 10 xúc xắc. Tìm xác suất để tổng các số xuất hiện lớn hơn hoặc bằng 58.

6. Hãy tính xác suất để sinh nhật của k người ($k < 12$) vào các tháng khác nhau. (Xác suất sinh vào các tháng như nhau).

7. Hãy tính xác suất để sinh nhật của 2 người vào cùng một tháng. (Coi tháng Hai 28 ngày và xác suất sinh vào các ngày trong năm bằng nhau và bằng $\frac{1}{365}$).

8. Một người viết thư cho 5 người quen. Sau khi ghi tên và địa chỉ lên các phong bì, do đãng trí người đó bỏ lẫn lộn các thư vào các phong bì. Tính xác suất để

   (a) Có ít nhất 3 người nhận được đúng thư gửi cho mình.

   (b) Có ít nhất 2 người nhận được đúng thư gửi cho mình.

9. A và B chơi một trò chơi như sau: A gieo xúc xắc, kết quả giả sử mặt k chấm xuất hiện. A gieo tiếp đồng thời 2 đồng xu k lần. Nếu ít nhất có một lần xảy ra biến cố cả hai đồng xu cùng xuất hiện mặt ngửa, khi đó A thắng cuộc, ngược lại A bị thua. Hỏi trò chơi đó có lợi cho A hay cho B?

10. Cho n hộp, mỗi hộp chứa đúng a bi trắng và b bi đỏ. Lấy ngẫu nhiên 1 viên bi từ hộp thứ nhất và bỏ sang hộp thứ hai, sau đó lấy tiếp 1

viên bi từ hộp thứ hai và bỏ sang hộp thứ ba,... Cuối cùng lấy 1 viên bi từ hộp thứ n. Gọi A là biến cố viên bi lấy từ hộp thứ nhất bỏ sang hộp thứ hai là viên bi trắng, B là biến cố viên bi lấy từ hộp thứ n là viên bi trắng. Kí hiệu $p_n = P(B/A)$. Chứng minh rằng

$$p_n = \frac{a}{a+b} + \frac{b}{a+b}(a+b+1)^{1-n}.$$

11. Cho phương trình $x^2 + ax + b = 0$, trong đó a, b là các điểm ngẫu nhiên lấy trên khoảng $(-1, 1)$ theo phân bố đều. Tìm xác suất để phương trình có các nghiệm thực.

12. Chọn ngẫu nhiên N điểm trong hình cầu bán kính R. Tìm xác suất để khoảng cách của điểm gần tâm nhất tới tâm hình cầu lớn hơn r. Tìm giới hạn của xác suất đó khi $N \to \infty, R \to \infty$, với giả thiết

$$\lim_{N\to\infty, R\to\infty} \frac{Nr^3}{R^3} = c.$$

13. Trên đường tròn chọn n điểm, chúng tạo thành một đa giác (lồi). Tìm xác suất để tâm hình tròn nằm trong đa giác đó.

14. Rải ngẫu nhiên N viên bi vào n hộp (sao cho n^N khả năng có xác suất như nhau). Với điều kiện một hộp xác định từ trước nào đó (ví dụ hộp thứ nhất) không rỗng, tìm xác suất để hộp đó có đúng K viên bi (K không nhỏ hơn 1).

15. Một bia hình tròn được chia thành $2n$ hình quạt bằng nhau. Một xạ thủ bắn 2 phát đạn trúng bia. Tìm xác suất để 2 viên đạn trúng vào 2 hình quạt đối xứng nhau.

16. Bốn người bạn tiến hành bốc thăm để chọn một người đi nghỉ mát. Họ làm 4 phiếu, trong đó có 1 phiếu ghi được đi nghỉ mát, 3 phiếu còn lại ghi KHÔNG. Từng người lần lượt chọn ngẫu nhiên 1 phiếu, hỏi rằng để bốc được phiếu đi nghỉ mát, thứ tự bốc thăm có quan trọng không?

17. Các hộp được đánh số $0, 1, 2, ..., N$ và hộp mang số k chứa k bi đỏ, $N - k$ bi trắng $(k = 0, 1, 2, ..., N)$. Chọn ngẫu nhiên một hộp và từ hộp này chọn lần lượt có hoàn lại từng viên bi. Gọi A_n là biến cố lần chọn thứ n lấy được viên bi đỏ.

    (a) Tính $P(A_3/A_1A_2)$

    (b) Giả sử từ hộp đã chọn ngẫu nhiên chọn lần lượt hai viên bi không hoàn lại. Tìm xác suất để cả hai bi đã chọn là bi đỏ.

18. Sửa lại bài tập số 16. như sau: hộp mang số k chứa k bi đỏ và N bi trắng (chứa $k + N$ viên bi). Tìm xác suất để cả hai bi đã chọn từ hộp chọn ngẫu nhiên là bi đỏ trong các trường hợp

    (a) Phép chọn có hoàn lại.

    (b) Phép chọn không hoàn lại.

    Tìm giới hạn của các xác suất đó khi $N \to \infty$.

19. Tại một quầy bán vé có 10 người xếp hàng, giá vé vào cổng viên 500 đồng. Giả sử 4 người trong số họ chỉ có tiền loại 1000 đồng và 6 người chỉ có loại tiền 500 đồng. Hãy tìm xác suất để việc bán vé không phải dừng lại giữa chừng để đổi tiền (hay nói cách khác không xảy ra trường hợp, quầy bán vé chỉ có loại tiền 1 nghìn đồng, trong khi đến lượt người mua vé cũng chỉ có loại tiền đó). Giả thiết ban đầu quầy bán vé không có tiền.

20. Một hộp chứa n bi đ ỏ và m bi trắng. Lần lượt chọn ngẫu nhiên các viên bi từ hộp đó (bi đã chọn không hoàn lại). Tìm xác suất để sau một số lần rút nào đó, số bi đỏ và bi trắng còn lại trong hộp bằng nhau. (Gi thiết ban đầu $n > m$.)

ĐÁP SỐ VÀ HƯỚNG DẪN

1. $1 - \frac{3}{C_9^6} = \frac{27}{28}$.

2. $C_N^n = \sum_{k=1}^{N-n+1} C_{N-k}^{n-1} \quad (n \leq N)$.

   (Hướng dẫn: Gọi p_k là xác suất để số nhỏ nhất trong n số chọn ra (chọn ngẫu nhiên) từ tập {1,2,...N} bằng k, $p_k = \frac{C_{N-k}^{n-1}}{C_N^n}$).

3. $\frac{1}{3}$.

   (Hướng dẫn: hoặc 3 số chọn từ tập $\{1, 2, 4, 5\}$ hoặc 3 số chọn từ $\{3, 6\}$.

4. $\frac{1}{6}$.

   (Hướng dẫn: $n - 1$ xúc xắc đầu tùy ý, xúc xắc thứ n xảy ra duy nhất 1 trong 6 khả năng).

5. $\frac{1}{6^{10}} + \frac{10}{6^{10}} + \frac{10}{6^{10}} + \frac{C_{10}^2}{6^{10}} = \frac{11}{6^9}$.

6. $\frac{C_{12}^k k!}{12^k}$.

7. $7\left(\frac{31}{365}\right)^2 + 4\left(\frac{30}{365}\right)^2 + \left(\frac{28}{365}\right)^2 \approx 0,09$.

8. (a) $\frac{C_5^3+1}{5!} = \frac{11}{120}$.

   (b) $\frac{C_5^3+1+2C_5^2}{5!} = \frac{31}{120}$.

9. Sử dụng công thức xác suất đầy đủ, xác suất để A thắng:

$$\frac{1}{2}\left(1 + \frac{3^6}{4^6}\right) > \frac{1}{2}.$$

10. Chứng minh bằng quy nạp. Dùng công thức xác suất đầy đủ:

$$p_{n+1} = \frac{a+1}{a+b+1} p_n + \frac{a}{a+b+1}(1 - p_n).$$

11. $\dfrac{13}{24}$.

12. $\lim\limits_{N\to\infty, R\to\infty}\left(1-\dfrac{r^3}{R^3}\right)^N = e^{-c}$.

13. $1-\dfrac{n}{2^{n-1}}$.

14. $\dfrac{C_K^N \frac{(n-1)^{N-K}}{n^N}}{1-\frac{(n-1)^N}{n^N}}$.

15. $\dfrac{1}{2n}$.

16. Thứ tự bốc thăm KHÔNG quan trọng. Sử dụng định lí nhân xác suất ta sẽ dễ dàng chỉ ra xác suất để người bốc đầu tiên cũng như người bốc thứ hai trúng phiếu đi nghỉ mát bằng $\frac{1}{4}$. Tương tự người bốc thứ ba, rồi thứ tư trúng phiếu đi nghỉ mát đều bằng $\frac{1}{4}$.

17. (a) $P(A_1A_2A_3) = \sum_{k=1}^{N}\left(\frac{k}{N}\right)^3\frac{1}{N+1} = \frac{N+1}{4N}$.
$P(A_1A_2) = \dfrac{2N+1}{6N} \to P(A_3/A_1A_2) = \dfrac{3(N+1)}{2(2N+1)}$

(b) $P(A_1A_2 = \sum_{k=2}^{N}\frac{k(k-1)}{N(N-1)}\cdot\frac{1}{N+1} = \frac{1}{3}$.

18. Giới hạn trong cả hai trường hợp đều bằng
$$\frac{3}{2} - 2\log 2.$$

19. Có thể coi 10 người này xếp một cách ngẫu nhiên vào hàng trước quầy bán vé. Đến lượt một người có loại tiền 500 đồng có thể coi như điểm dịch chuyển ngẫu nhiên trong bài toán lang thang ngẫu nhiên bước sang phải và ngược lại nếu đến lượt một người có loại tiền 1000 đồng có thể coi như điểm ngẫu nhiên bước sang trái. Bài toán dẫn đến hãy tìm xác suất để sau 10 bước điểm ngẫu nhiên dịch chuyển từ O (ban đầu quầy bán vé không có tiền) tới điểm (10, 6-4) (là điểm (10, 2)) và không chạm đường thẳng $f(x) = -1$. Tổng số đường đi có thể có từ $O(0,0)$ tới điểm (10, 2) bằng C_{10}^4 và số trường hợp thuận lợi là số đường đi không cắt đường thẳng $f(x) = -1$ bằng $C_{10}^4 - C_{10}^3$ (số đường đi có thể có từ $(0,-2)$ tới $(10,2)$).
Vậy xác suất cần tìm là $\frac{C_{10}^4 - C_{10}^3}{C_{10}^4} = \frac{3}{7}$.

20. Gọi A là biến cố cần tìm xác suất, khi đó biến cố đối lập $\overline{A}$ là biến cố ở bất kì thời điểm nào trong quá trình rút bi (trừ lần cuối cùng thứ $n+m$), số bi đỏ và bi trắng còn lại trong hộp không bằng nhau (điều đó cũng có nghĩa là số bi đỏ luôn luôn nhiều hơn số bi trắng ở các lần rút thứ i mà $i < n+m$). Để tính xác suất của $\overline{A}$, trước hết ta tính số trường hợp đồng khả năng của nó. Tổng số đường đi có thể có từ $O(0,0)$ tới điểm $(n+m, n-m)$ bằng C^n_{n+m} và số trường hợp thuận lợi là số đường đi từ $O(0,0)$ tới điểm $(n+m-1, n-m-1)$ mà đường đi không cắt đường thẳng $f(x) = n-m$. Số đường đi từ $O(0,0)$ tới điểm $(n+m-1, n-m-1)$ mà cắt đường thẳng $f(x) = n-m$ bằng số đường đi từ $(0, 2n-2m)$ tới điểm $(n+m-1, n-m-1)$

$$C^{n-1}_{n+m-1} = C^m_{n+m-1}$$

Số trường hợp thuận lợi của $\overline{A}$ bằng

$$C^n_{n+m} - C^{n-1}_{n+m-1} = C^n_{n+m-1}$$

Xác suất cần tìm là

$$P(A) = 1 - \frac{C^n_{n+m-1}}{C^n_{n+m}}.$$

Chương 2

Đại lượng ngẫu nhiên và phân bố xác suất

2.1 Khái niệm về đại lượng ngẫu nhiên

Cho đến nay chúng ta mới chỉ nghiên cứu các biến cố liên quan tới phép thử ngẫu nhiên có xảy ra hay không và xác suất của chúng. Phần lớn các phép thử ngẫu nhiên, cùng với các kết quả của nó (các biến cố ngẫu nhiên cơ bản) là các giá trị số đi kèm. Chẳng hạn khi gieo xúc xắc, ta nói đến số chấm trên mặt xuất hiện của xúc xắc hoặc số cuộc gọi tới tổng đài điện thoại trong một khoảng thời gian T nào đó. Nói cách khác ta cần các con số để mô tả, diễn tả các biến cố ngẫu nhiên. Nó vừa gọn, đảm bảo chính xác và thuận tiện cho việc tính toán sau này. Các giá trị số đó nói chung phụ thuộc vào kết quả của phép thử ngẫu nhiên, nói cách khác chúng phụ thuộc vào các biến cố ngẫu nhiên cơ bản. Do vậy nếu ứng với mỗi biến cố ngẫu nhiên cơ bản, ta gán cho một số thực nào đó, khi đó hầu hết các biến cố ngẫu nhiên có thể biểu diễn bằng một cách nào đó thông qua phép gán (ánh xạ) này.

Một ánh xạ từ tập các biến cố ngẫu nhiên cơ bản vào tập các số thực được gọi là đại lượng ngẫu nhiên.

$$X : \Omega \to \mathbb{R}$$

Người ta thường kí hiệu đại lượng ngẫu nhiên bằng các chữ in hoa: $X, Y, Z, ...$ hoặc các các chữ cái Hylạp: $\xi, \eta ...$

Nếu miền giá trị của đại lượng ngẫu nhiên (ánh xạ $X : \Omega \to \mathbb{R}$) là tập hữu hạn hoặc vô hạn đếm dược, ta gọi X là *đại lượng ngẫu nhiên rời rạc*.

Trên không gian các biến cố ngẫu nhiên cơ bản có thể xác định rất nhiều các đại lượng ngẫu nhiên. Tuy nhiên sẽ chỉ có một số nhất định có ý nghĩa thực tiễn.

Ví dụ 2.1.1

Gieo đồng thời hai xúc xắc, không gian các biến cố ngẫu nhiên cơ bản Ω gồm các cặp số $(i, j) \quad i, j = 1, 2, ..., 6$ (i là số chấm xuất hiện ở xúc xắc thứ nhất và j là số chấm xuất hiện ở xúc xắc thứ hai). Gọi X là tổng số chấm xuất hiện ở hai xúc xắc. Khi đó X thực chất là ánh xạ

$$X : \Omega \to \mathbb{R}$$

$$X(i, j) = i + j$$

Do vậy X là một đại lượng ngẫu nhiên (rời rạc). Chúng ta có thể mô-tả nhiều biến cố ngẫu nhiên thông qua đại lượng ngẫu nhiên X, chẳng hạn $\{X > 10\}$ kí hiệu biến cố tổng số chấm xuất hiện ở hai xúc xắc lớn hơn 10. Ta cũng có thể biểu diễn biến cố đó thành tổng của hai biến cố $\{X > 10\} = \{X = 11\} + \{X = 12\}$. Tương tự như thế biến cố $\{X \leq 3\}$ kí hiệu biến cố tổng số chấm xuất hiện ở hai xúc xắc không lớn hơn 3, có thể biểu diễn thành tổng của 3 biến cố

$$\{X \leq 3\} = \{X = 1\} + \{X = 2\} + \{X = 3\}$$

Ngoài ra chúng ta còn có thể dẫn vào các đại lượng ngẫu nhiên khác như

$$Y(i, j) = 2i - 3j \quad \text{hoặc} \quad Z(i, j) = i^2 + j^2$$

Ví dụ 2.1.2

Một xạ thủ bắn súng vào bia. Giả sử bia là một hình tròn có bán kính bằng 1 và xạ thủ chắc chắn bắn trúng bia. Mỗi điểm thuộc bia (điểm viên đạn chạm bia) được coi là một biến cố ngẫu nhiên cơ bản. Không gian các biến cố ngẫu nhiên cơ bản Ω gồm tất cả các điểm thuộc hình tròn đó. Đưa vào hệ trục toạ độ Đề các uOv sao cho tâm bia trùng với gốc toạ độ, ta có

$$\Omega = \{(u, v)/u^2 + v^2 \leq 1\}$$

Bây giờ chúng ta có thể xác định các đại lượng ngẫu nhiên trên Ω, chẳng hạn ξ là khoảng cách từ điểm viên đạn chạm bia tới tâm bia

$$\xi = \sqrt{u^2 + v^2}$$

Nếu chia bia thành 10 phần bởi các vòng tròn đồng tâm (tâm O) bán kính bằng $\frac{1}{10}, \frac{2}{10}, ..., \frac{9}{10}, 1$, khi đó biến cố $\{\xi < \frac{1}{10}\}$ xảy ra đồng nghĩa với việc xạ thủ đó bắn giỏi (vòng 10), hoặc biến cố $\{\xi > \frac{1}{2}\}$ xảy ra tức là xạ thủ đó bắn không đạt yêu cầu (điểm dưới 5).

2.2 Phân bố xác suất của đại lượng ngẫu nhiên

Trước hết chúng ta dẫn vào khái niệm *bảng phân bố xác suất* của đại lượng ngẫu nhiên rời rạc.

Giả sử X là đại lượng ngẫu nhiên rời rạc. Như đã trình bày ở trên, X thực chất là một ánh xạ

$$X : \Omega \to \mathbb{R}$$

miền giá trị của X là tập hợp hữu hạn hoặc vô hạn đếm được gồm các số $\{x_i, i \in I\}$. Gọi A_i là biến cố $\{X = x_i\}$ và $P(A_i) = p_i$. Hiển nhiên $A_m A_n = \emptyset$ với $m \neq n \quad (m, n \in I)$. Mặt khác $\{x_i, i \in I\}$ là tất cả các giá trị có thể có của X nên hệ các biến cố ngẫu nhiên $\{A_i, i \in I\}$ lập thành một hệ đầy đủ. Nói cách khác $\sum_{i \in I} P(A_i) = \sum_{i \in I} p_i = 1$.

Khi tiến hành phép thử ngẫu nhiên, trong nhiều ứng dụng chúng ta không quan tâm tới biến cố A_n có xảy ra hay không xảy ra (hoặc nói cách khác đại lượng ngẫu nhiên X nhận giá trị x_n hay không) mà chúng ta muốn biết xác suất xảy ra các biến cố đó. Các xác suất $p_i, i \in I$ được gọi là *phân bố xác suất* của đại lượng ngẫu nhiên X. Phân bố xác suất của đại lượng ngẫu nhiên rời rạc thường được cho dưới dạng bảng như sau

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

$\big(n \in I\big)$

trong đó x_n là các giá trị có thể có của X và

$$p_n = P(X = x_n), \ \big(n \in I\big), \quad \sum_{n \in I} p_n = 1.$$

Trong thực tế nếu biết phân bố xác suất của X, ta có thể tính được xác suất của các biến cố như $\{a < X < b\}$ hay tổng quát hơn $\{X \in E\}$, trong đó E là tập các số thực bất kì nào đó.

Ví dụ nếu kí hiệu X là số chấm khi gieo một xúc xắc đối xứng, đồng chất. $A_n = \{X = n\}$ là biến cố mặt n chấm ($n = 1, 2, ..., 6$) xuất hiện. Bảng phân bố của X thường được cho dưới dạng sau:

X	1	2	3	4	5	6
P	p_1	p_2	p_3	p_4	p_5	p_6

trong đó $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$.

Gọi A là biến cố số chấm xuất hiện nhỏ hơn 5 và lớn hơn hoặc bằng 2

$$A = \{2 \leq X < 5\} = \{X = 2\} + \{X = 3\} + \{X = 4\} = A_2 + A_3 + A_4$$

khi đó dựa vào bảng phân bố trên, xác suất của A bằng

$$P(A) = P(A_2) + P(A_3) + P(A_4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Ví dụ 2.2.1

Quay lại ví dụ 1.5.3 chương I, một xạ thủ bắn bia, xác suất bắn trúng bia của xạ thủ đó bằng p. Xạ thủ bắn liên tục vào bia cho đến khi trúng bia thì dừng. Gọi X là số đạn chi phí (số đạn xạ thủ đã sử dụng). Rõ ràng X là đại lượng ngẫu nhiên có thể nhận mọi giá trị là các số tự nhiên $1, 2, ...$ Biến cố $\{X = n\}$ là biến cố "xạ thủ bắn trượt ở tất cả $n - 1$ lần bắn đầu và ở lần bắn thứ n xạ thủ đó mới bắn trúng bia". Ở chương I chúng ta đã tính xác suất của biến cố đó

$$P(X = n) = q^{n-1}p$$

Vậy bảng phân bố của đại lượng ngẫu nhiên X là

X	1	2	...	n	...
P	p	qp	...	$q^{n-1}p$	...

trong đó $p + q = 1$. Ta cũng nhận xét rằng

$$\sum_{n=1}^{\infty} P(X = n) = \sum_{n=1}^{\infty} q^{n-1}p = 1.$$

Lưu ý rằng miền giá trị của đại lượng ngẫu nhiên X là tập vô hạn không đếm được, khi đó ta không thể liệt kê thành bảng phân bố như trên.

Chẳng hạn xét một đại lượng ngẫu nhiên ξ bất kì (có thể là đại lượng ngẫu nhiên rời rạc hoặc không). ξ luôn tạo ra một phân bố xác suất trên tập các số thực theo nghĩa sau đây. Gọi E là tập các số thực nào đó và kí hiệu A là biến cố "đại lượng ngẫu nhiên ξ nhận các giá trị trong E".

$$A = \{\xi \in E\}$$

Khi E biến thiên, biến cố A cũng thay đổi theo. Xác suất $P(A)$ còn được kí hiệu là

$$P_\xi(E)$$

và được gọi là *phân bố xác suất của* ξ. Đặc biệt khi tập E có dạng

$$(-\infty, x) = \{u \in \mathbb{R}/u < x\} \subset \mathbb{R}$$

trong đó x là số thực bất kì. Hàm

$$F(x) = P\big(\xi \in (-\infty, x)\big) = P(\xi < x) = P_\xi(-\infty, x)$$

được gọi là *hàm phân bố xác suất của đại lượng ngẫu nhiên* ξ.

Về mặt lí thuyết khi đã biết hàm phân bố xác suất của ξ, người ta có thể tính được các xác suất $P(\xi \in E)$ với $E \subset \mathbb{R}$ là một tập con xác định nào đó.

Nếu ξ là đại lượng ngẫu nhiên rời rạc và biết bảng phân bố của ξ như sau:

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

Khi đó hàm $F(x) = P(\xi < x)$ (xác suất của biến cố "ξ nhận các giá trị nhỏ hơn x") được tính bằng

$$F(x) = \sum_{n:x_n<x} p_n$$

Ngược lại hiển nhiên từ hàm phân bố $F(x)$ của đại lượng ngẫu nhiên rời rạc, ta có thể lập bảng phân bố xác suất của đại lượng ngẫu nhiên đó.

Ví dụ 2.2.2

Từ một hộp đựng 3 bi đỏ và 5 bi trắng, chọn ngẫu nhiên ra 3 viên bi. Gọi X là số bi trắng chọn được.

i) Hãy lập bảng phân bố xác suất của X.

ii) Tính xác suất để có ít nhất một bi đỏ trong số 3 viên bi chọn ra.

iii) Tìm hàm phân bố xác suất của X.

Giải: i) Đại lượng ngẫu nhiên X có thể nhận các giá trị $0, 1, 2, 3$.

$$P(X=0) = \frac{C_3^3}{C_8^3} = \frac{1}{56}$$
$$P(X=1) = \frac{C_3^2 \cdot C_5^1}{C_8^3} = \frac{15}{56}$$
$$P(X=2) = \frac{C_3^1 \cdot C_5^2}{C_8^3} = \frac{30}{56}$$
$$P(X=3) = \frac{C_5^3}{C_8^3} = \frac{10}{56}$$

Vậy bảng phân bố xác suất của X là:

X	0	1	2	3
P	$\frac{1}{56}$	$\frac{15}{56}$	$\frac{30}{56}$	$\frac{10}{56}$

ii) Gọi A là biến cố "có ít nhất một bi đỏ trong số 3 viên bi chọn ra'". Như vậy A cũng chính là biến cố "có nhiều nhất 2 bi trắng trong số 3 viên bi chọn ra".

$$A = \{X=0\} + \{X=1\} + \{X=2\}. \quad \text{Suy ra}$$

$$P(A) = \frac{1}{56} + \frac{15}{56} + \frac{30}{56} = \frac{23}{28}.$$

Nhận xét rằng ta có thể tính $P(A)$ thông qua biến cố đối lập $\overline{A}$: cả 3 viên bi chọn ra đều là bi trắng. Vì vậy

$$P(\overline{A}) = P(X=3) = \frac{10}{56} \quad \Rightarrow \quad P(A) = 1 - P(\overline{A}) = \frac{23}{28}.$$

iii) Dựa vào bảng phân bố xác suất của X, suy ra hàm phân bố $F(x)$ bằng

$$F(x) = \begin{cases} P(X<0)=0 & \text{nếu } x \le 0 \\ P(X=0)=\frac{1}{56} & \text{nếu } 0 < x \le 1 \\ P(X=0)+P(X=1)=\frac{16}{56} & \text{nếu } 1 < x \le 2 \\ P(X\le 2)=\frac{46}{56} & \text{nếu } 2 < x \le 3 \\ P(X>3)=1 & \text{nếu } x > 3 \end{cases}$$

2.3 Tính chất của hàm phân bố

Ta nhắc lại định nghĩa về hàm phân bố

Định nghĩa 2.3.1 *Giả sử ξ là đại lượng ngẫu nhiên bất kì, khi đó hàm*

$$F(x) = P(\xi < x) \quad \textit{với mọi} \quad x \in \mathbb{R}$$

được gọi là hàm phân bố của đại lượng ngẫu nhiên ξ.

Hàm phân bố $F(x)$ có các tính chất sau:

a) $F(x)$ là hàm đơn điệu tăng

$$F(x_1) \le F(x_2), \quad \text{nếu} \quad x_1 < x_2$$

vì $\{\xi < x_1\} \subset \{\xi < x_2\} \Rightarrow F(x_1) = P(\xi < x_1) \le P(\xi < x_2) = F(x_2)$

b) $\lim_{x\to-\infty} F(x) = 0, \quad \lim_{x\to\infty} F(x) = 1$

Thật vậy, gọi $x_1 < x_2 < ... < x_n < ...$ là dãy các số thực tăng và $\lim_{n\to\infty} x_n = \infty$. Kí hiệu $A_n = \{\xi < x_n\}$, khi đó $A_1 \subset A_2 \subset ... \subset A_n \subset ...$ là dãy các biến cố tăng và $\sum_{n=1}^{\infty} A_n = \Omega$. Theo nhận xét sau định lí 1.3.5

$$F(x_n) = P(\xi < x_n) = P(A_n) \to P(\Omega) = 1, \text{khi} \quad n \to \infty$$

Tương tự xét dãy các số thực giảm bất kì $\{x_n\}, \lim_{n\to\infty} x_n = -\infty$, khi đó dãy các biến cố $A_n = \{\xi < x_n\}$ đơn điệu giảm và

$$\prod_{n=1}^{\infty} A_n = \emptyset.$$

Suy ra

$$\lim_{x\to-\infty} F(x) = 0$$

c) $F(x)$ liên tục trái tại $x = a$ với mọi $a \in \mathbb{R}$

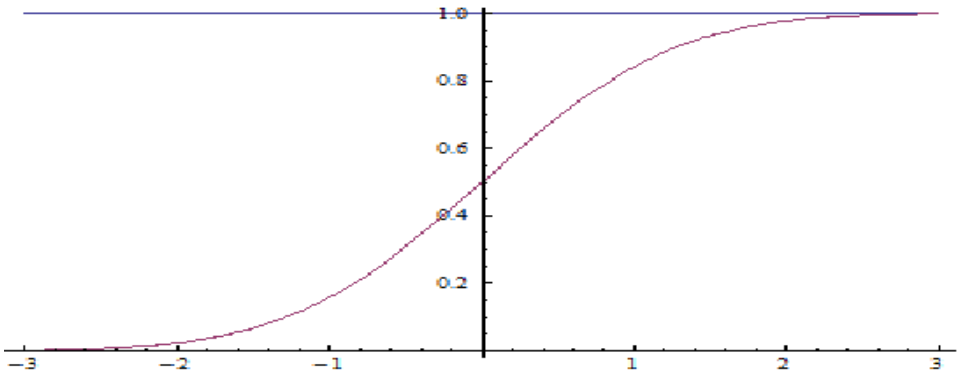
$$\lim_{x\to a-0} F(x) = F(a)$$

Chứng minh tương tự như trên, gọi $x_1 < x_2 < ... < x_n < ...$ là dãy các số thực tăng và $\lim_{n\to\infty} x_n = a$. Kí hiệu $A = \{\xi < a\}$ và $A_n = \{\xi < x_n\}$, khi đó $A_1 \subset A_2 \subset ... \subset A_n \subset ...$ là dãy các biến cố tăng, $\sum_{n=1}^{\infty} A_n = A$, suy ra

$$F(x_n) = P(A_n) \to P(A) = P(\xi < a) = F(a), \quad \text{khi} \quad n \to \infty$$

Ngược lại, người ta cũng chứng minh được rằng nếu một hàm $F(x)$ thoả mãn các tính chất a), b), c) nêu trên, khi đó tồn tại đại lượng ngẫu nhiên ξ sao cho $F(x)$ là hàm phân bố của ξ.

Đồ thị của hàm phân bố có dạng



Hệ quả 2.3.1 *Từ định nghĩa và tính chất của hàm phân bố $F(x)$, suy ra với mọi $a, b \in \mathbb{R}$*

i) $P(a \le \xi < b) = F(b) - F(a)$

ii) $P(\xi = a) = F(a+0) - F(a)$ *iii)* $P(a < \xi < b) = F(b) - F(a+0)$

iv) $P(a \le \xi \le b) = F(b+0) - F(a)$

v) $P(a < \xi \le b) = F(b+0) - F(a+0)$

(Chú ý rằng $F(x)$ là hàm đơn điệu tăng nên với mọi $a, b \in \mathbb{R}$ luôn tồn tại các giới hạn phải $F(a+0), F(b+0)$)

Thật vậy hệ quả i) được chứng minh như sau. Do
$\{a \leq \xi < b)\} = \{\xi < b\} \backslash \{\xi < a\}$ suy ra

$$\begin{aligned} P(a \leq \xi < b) &= P(\{\xi < b\} \backslash \{\xi < a\}) \\ &= P(\{\xi < b\} - P(\{\xi < a\}) \\ &= F(b) - F(a) \end{aligned}$$

Hệ quả ii) $P(\xi = a) = F(a + 0) - F(a)$ được suy ra từ

$$\{\xi = a\} = \bigcap_{n=1}^{\infty} \left(\{a \leq \xi < a + \frac{1}{n}\} \right)$$

tích (giao) của dãy các biến cố giảm dần.

Các phần còn lại của hệ quả được chứng minh tương tự.

Nhận xét 2.3.1 *Trong các ứng dụng của lí thuyết xác suất ta thường phải tính xác suất $P(\xi \in E)$, trong đó E là hợp của các khoảng mở (a, b) hoặc nửa đóng, nửa mở $(a, b]$. Nếu chúng đôi một không có điểm chung khi đó nhờ hệ quả trên ta có thể xác định được $P(E)$.*

Chú ý rằng nếu hàm phân bố $F(x)$ liên tục tại $x = a$, khi đó theo ii) $P(\xi = a) = F(a + 0) - F(a) = 0$.

2.4 Đại lượng ngẫu nhiên liên tục

Định nghĩa 2.4.1 *X được gọi là đại lượng ngẫu nhiên liên tục nếu tồn tại một hàm không âm*

$$f : \mathbb{R} \to [0, \infty]$$

sao cho hàm phân bố $F(x)$ của đại lượng ngẫu nhiên X thoả mãn

$$F(b) - F(a) = P(a \leq X < b) = \int_a^b f(x)\, dx$$

với mọi $a < b \in \mathbb{R}$. Khi đó hàm f được gọi là mật độ xác suất (hay nói tắt là hàm mật độ) của đại lượng ngẫu nhiên X.

Chú ý rằng các cận tích phân a, b có thể bằng $-\infty, +\infty$. Cụ thể nếu $a = -\infty$, khi đó

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)\,dt,$$

suy ra (do tính chất của tích phân) hàm phân bố $F(x)$ liên tục trên $\mathbb{R}$. Như vậy theo hệ quả 1, các biến cố $\{a \le X < b\}, \{a < X < b\}, \{a \le X \le b\}$ có xác suất bằng nhau và bằng $\int_a^b f(x)\,dx = F(b) - F(a)$

Đặc biệt khi hàm mật độ $f(x)$ liên tục trên $\mathbb{R}$, từ định lí đạo hàm theo cận trên suy ra hàm phân bố $F(x)$ khả vi và

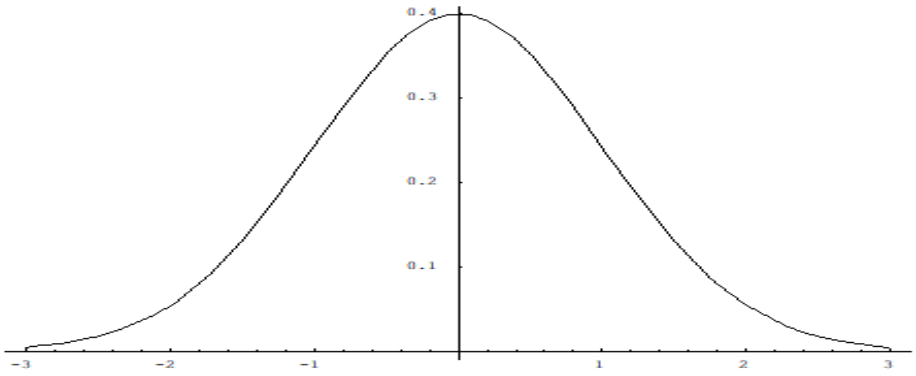
$$F'(x) = f(x)$$

Chọn $a = -\infty, b = +\infty$ trong định nghĩa trên, ta thấy hàm mật độ $f(x)$ có tính chất

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

Ngược lại, người ta cũng chứng minh được rằng nếu $f : \mathbb{R} \to [0, \infty]$ là hàm không âm và $\int_{-\infty}^{+\infty} f(x)\,dx = 1$, khi đó $f(x)$ là hàm mật độ của đại lượng ngẫu nhiên nào đó.

Đồ thị của hàm mật độ có dạng



Ví dụ 2.4.1

Trong hầu hết các chương trình của máy vi tính, lệnh "RANDOM" tạo ra một số thực ngẫu nhiên trong khoảng [0,1]. Kí hiệu X là số thực

ngẫu nhiên tạo bởi lệnh "RANDOM". Tìm hàm phân bố, hàm mật độ của X và tính xác suất để $\{\frac{1}{2} \leq X \leq \frac{2}{3}\}$

Giải: Ở ví dụ này, xác suất của các biến cố được tính bằng phương pháp hình học. Cụ thể hơn, xác suất để $\{X \in [\frac{1}{2}, \frac{2}{3}]\}$ bằng tỉ số giữa độ dài khoảng $[\frac{1}{2}, \frac{2}{3}]$ và độ dài đoạn $[0, 1]$

$$P(X \in [\frac{1}{2}, \frac{2}{3}]) = \frac{1}{6}$$

Tổng quát hơn, hàm phân bố của X được xác định như sau:

$$F(x) = \begin{cases} P(X < x) = 0 & \text{nếu } x \leq 0 \\ P(X < x) = P(0 \leq X < x) = \frac{x}{1} = x & \text{nếu } 0 < x \leq 1 \\ P(X < x) = P(0 \leq X < 1) = \frac{1}{1} = 1 & \text{nếu } x > 1 \end{cases}$$

Dễ dàng nhận thấy

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

trong đó

$$f(x) = \begin{cases} 1 & \text{nếu } 0 < x \leq 1 \\ 0 & \text{nếu } x \leq 0 \text{ hoặc } x > 1 \end{cases} \qquad \forall x \in \mathbb{R}.$$

Vậy $f(x)$ là hàm mật độ của đại lượng ngẫu nhiên X.

Chú ý rằng theo định nghiã 2.4.1, xác suất của biến cố $\{X \in [\frac{1}{2}, \frac{2}{3}]\}$ có thể tính thông qua hàm mật độ $f(t)$

$$P(\frac{1}{2} \leq X \leq \frac{2}{3}) = \int_{\frac{1}{2}}^{\frac{2}{3}} f(t)\, dt = \int_{\frac{1}{2}}^{\frac{2}{3}} dt = \frac{1}{6}.$$

Định nghĩa 2.4.2 *Người ta gọi ξ là đại lượng ngẫu nhiên có phân bố đều trên đoạn $[a, b]$ nếu hàm mật độ của ξ*

$$f(x) = \begin{cases} \frac{1}{b-a} & \textit{nếu } a < x \leq b \\ 0 & \textit{nếu } x \leq a \textit{ hoặc } x > b \end{cases}$$

Suy ra X trong ví dụ 2.4.1, phân bố đều trên đoạn $[0,1]$. Theo định nghĩa 2.4.1, xác suất để một đại lượng ngẫu nhiên bất kì nhận các giá trị thuộc khoảng (c,d) bằng

$$P(c \le X < d) = \int_c^d f(x)\,dx.$$

Vậy nếu ξ là đại lượng ngẫu nhiên có phân bố đều trên đoạn $[a,b]$, khi đó với $(c,d) \subset (a,b)$

$$P(c < X < d) = \int_c^d f(x)\,dx = \int_c^d \frac{1}{b-a}\,dx = \frac{d-c}{b-a}.$$

Từ nhận xét này suy ra phương pháp hình học để tính xác suất (đã trình bày trong chương I) chỉ áp dụng cho các biến cố tạo thành từ các đại lượng ngẫu nhiên có phân bố đều.

Ví dụ 2.4.2

Trở lại ví dụ 2.1.2, một xạ thủ bắn bia. Bia là một hình tròn có bán kính bằng 1 và giả thiết xác suất để viên đạn rơi vào một miền A bất kì bằng tỉ lệ giữa diện tích miền A với diện tích hình tròn bia. Không gian các biến cố ngẫu nhiên cơ bản

$$\Omega = \{(u,v)/u^2 + v^2 \le 1\}$$

Gọi ξ là khoảng cách từ điểm viên đạn chạm bia tới tâm bia

$$\xi = \sqrt{u^2 + v^2}$$

Tìm hàm phân bố, hàm mật độ của ξ

Giải: Biến cố $\{\xi < x\}$ là biến cố viên đạn rơi vào hình tròn tâm trùng với tâm của bia và bán kính bằng x. Suy ra hàm phân bố

$$F(x) = P(\xi < x) = \frac{\pi \cdot x^2}{\pi \cdot 1^2} = x^2 \quad \text{nếu } 0 \le x \le 1$$

(Hiển nhiên $F(x) = 0$ nếu $x < 0$ và $F(x) = 1$ nếu $x > 1$). Mặt khác $F'(x) = 2x$ nếu $0 \le x \le 1$ hay

$$F(x) = \int_{-\infty}^x f(t)\,dt$$

trong đó

$$f(x) = \begin{cases} 2x & \text{nếu } 0 < x \leq 1 \\ 0 & \text{nếu } x \leq 0 \text{ hoặc } x > 1 \end{cases}$$

Suy ra $f(x)$ là hàm mật độ của ξ.

Trong nhiều bài toán, biết hàm mật độ của đại lượng ngẫu nhiên ξ, chúng ta cần tìm hàm mật độ (hàm phân bố) của một đại lượng ngẫu nhiên khác $\eta = \varphi(\xi)$. Định lí sau cho ta khả năng như vậy

Định lí 2.4.1 *Giả sử $f(x)$ là hàm mật độ của ξ, I là miền giá trị của đại lượng ngẫu nhiên ξ và hàm $\varphi : I \to \mathbb{R}$ khả vi, đơn điệu thực sự (tăng hoặc giảm) trên I. Khi đó hàm mật độ của $\eta = \varphi(\xi)$ bằng*

$$g(y) = f\left(\varphi^{-1}(y)\right) \left| \frac{d\varphi^{-1}(y)}{dy} \right|$$

Chứng minh. Theo giả thiết hàm $\varphi : I \to \mathbb{R}$ khả vi và đơn điệu thực sự (trước tiên ta xét trường hợp φ là hàm tăng) suy ra hàm phân bố của η là

$$P(\eta < y) = P\left(\varphi(\xi) < y\right) = P\left(\xi < \varphi^{-1}(y)\right) = F\left(\varphi^{-1}(y)\right).$$

trong đó F là hàm phân bố của ξ. Từ đây đạo hàm theo y ta được điều phải chứng minh.

Chứng minh tương tự trong trường hợp φ là hàm giảm

$$P(\eta < y) = P\left(\varphi(\xi) < y\right) = P\left(\xi > \varphi^{-1}(y)\right) = 1 - F\left(\varphi^{-1}(y)\right).$$

Trong trường hợp này do φ là hàm giảm nên đạo hàm $\frac{d\varphi^{-1}(y)}{dy} < 0$. Vậy đạo hàm hàm phân phối $G(y) = P(\eta < y)$ theo y ta được kết quả trên.

Ví dụ 2.4.3

Nếu $y = \varphi(x) = ax + b (a \neq 0), \eta = \varphi(\xi)$. Khi đó $\varphi^{-1}(y) = \frac{y-b}{a}$, suy ra hàm mật độ của η

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right)$$

Ví dụ 2.4.4

Quay lại ví dụ 2.4.2, một xạ thủ bắn bia. Không gian biến cố ngẫu nhiên cơ bản là hình tròn $\Omega = \{(u,v)/u^2+v^2 \leq 1\}$. Gọi ξ là khoảng cách từ điểm viên đạn chạm bia tới tâm bia

$$\xi = \sqrt{u^2+v^2}$$

Tìm hàm mật độ của $\eta = \xi^2$

Giải: Trong ví dụ 2.4.2, ta đã tính hàm mật độ của ξ

$$f(x) = \begin{cases} 2x & \text{nếu } 0 < x \leq 1 \\ 0 & \text{nếu } x \leq 0 \text{ hoặc } x > 1 \end{cases}$$

Áp dụng định lí 2.4.1, với $\varphi^{-1}(y) = \sqrt{y} \quad (\varphi(\xi) = \xi^2)$ suy ra hàm mật độ của $\eta = \xi^2$ là

$$g(y) = \begin{cases} 2\varphi^{-1}(y)\frac{d\varphi^{-1}(y)}{dy} = 1 & \text{nếu } 0 < y \leq 1 \\ 0 & \text{nếu } y \leq 0 \text{ hoặc } y > 1 \end{cases}$$

Vậy η là đại lượng ngẫu nhiên có phân bố đều trên đoạn $[0,1]$.

Chú ý rằng công thức trong định lí 2.4.1, chỉ được sử dụng trong trường hợp φ là hàm đơn điệu thực sự trên miền giá trị của đại lượng ngẫu nhiên ξ.

Trường hợp φ không đơn điệu trên miền giá trị I của ξ, trước hết ta phải tính phân bố xác suất của η

$$G(y) = P(\eta < y) = P(\varphi(\xi) < y),$$

sau đó đạo hàm hàm $G(y)$ để tìm hàm mật độ $g(y) = G'(y)$ của η.

Ví dụ nếu ξ là đại lượng ngẫu nhiên phân bố đều trên đoạn $[-1,1]$. Xét đại lượng ngẫu nhiên $\varphi(\xi) = \xi^2$. Rõ ràng φ không là hàm đơn điệu trên đoạn $[-1,1]$. Để tìm hàm mật độ của $\eta = \xi^2$, ta có với $\forall y \in (0;1)$ hàm phân bố của η là:

$$G(y) = P(\eta < y) = P(\xi^2 < y) = P(-\sqrt{y} < \xi < \sqrt{y}) = \sqrt{y}.$$

Suy ra hàm mật độ $g(y)$ của η bằng

$$g(y) = \begin{cases} \frac{1}{2} \cdot \frac{1}{\sqrt{y}} & \text{nếu } 0 < y \leq 1 \\ 0 & \text{nếu } y \leq 0 \text{ hoặc } y > 1. \end{cases}$$

2.5 Các đặc trưng cơ bản: Kì vọng và phương sai của đại lượng ngẫu nhiên

2.5.1 Kì vọng

Một trong các đặc trưng quan trọng của đại lượng ngẫu nhiên là kì vọng. Để giúp ta hiểu khái niệm kì vọng của đại lượng ngẫu nhiên, ta xét ví dụ sau

Một người chơi một trò chơi may rủi. Giả sử trò chơi đó có thể xảy ra r khả năng và xác suất xảy ra các khả năng (biến cố) đó lần lượt là

$$p_1, p_2, ..., p_r.$$

Nếu khả năng (biến cố ngẫu nhiên) thứ i xảy ra, người đó thu được số tiền $x_i, i = 1, 2, ..., r$. (Chú ý rằng x_i có thể là số âm, điều đó đồng nghĩa với việc người chơi bị thua).

Giả sử người đó chơi n ván, gọi k_i là số lần xảy ra biến cố thứ i. (k_i là số ván mà mỗi ván thu được số tiền x_i). Như vậy tổng số tiền thu được sau n ván là

$$k_1x_1 + k_2x_2 + ... + k_rx_r,$$

hay trung bình một ván người đó được

$$\frac{k_1x_1 + k_2x_2 + ... + k_rx_r}{n}.$$

Do $\frac{k_i}{n}$ là tần suất của biến cố thứ i, suy ra $\frac{k_i}{n} \approx p_i$. Vậy trung bình mỗi ván người đó thu được

$$\frac{k_1}{n}x_1 + \frac{k_2}{n}x_2 + ... + \frac{k_r}{n}x_r \approx p_1x_1 + p_2x_2 + ... + p_rx_r.$$

Biểu thức $p_1x_1 + p_2x_2 + ... + p_rx_r$ (số tiền thu được trung bình mỗi ván khi n đủ lớn) được gọi là *kì vọng*. Hiển nhiên nếu kì vọng là số dương trò chơi có lợi cho người chơi và nếu kì vọng âm trò chơi sẽ bất lợi cho người đó. Ta dẫn vào định nghĩa sau

Định nghĩa 2.5.1 *Gọi X là đại lượng ngẫu nhiên rời rạc, bảng phân bố của X được cho như sau*

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

Khi đó $\sum_{i=1}^{\infty} x_i p_i$ *được gọi là kì vọng (hay còn gọi là giá trị trung bình) của* X*, nếu chuỗi hội tụ tuyệt đối. Người ta thường kí hiệu* $E(X)$ *là kì vọng của đại lượng ngẫu nhiên* X*.*

Nhận xét rằng nếu miền giá trị của đại lượng ngẫu nhiên X là tập hữu hạn khi đó luôn luôn tồn tại kì vọng $E(X)$. Từ định nghĩa trên ta thấy $E(X)$ chỉ phụ thuộc vào phân bố của X, do vậy người ta còn gọi $E(X)$ là *kì vọng của hàm phân bố* của X.

Ví dụ 2.5.1

Số trẻ em mới sinh ở một bệnh viện trong 1 ngày là đại lượng ngẫu nhiên X có phân bố xác suất

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

khi đó kì vọng của X (hay trung bình số trẻ em mới sinh trong 1 ngày) bằng

$$E(X) = 0 \cdot 0,3 + 1 \cdot 0,4 + 2 \cdot 0,2 + 3 \cdot 0,1 = 1,1$$

Định nghĩa 2.5.2 *Nếu* X *là đại lượng ngẫu nhiên liên tục,* $f(x)$ *là hàm mật độ. Khi đó kì vọng (giá trị trung bình) của* X *được xác định*

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

nếu tích phân hội tụ tuyệt đối.

Chú ý rằng nếu ta chia $\mathbb{R}$ thành các khoảng nhỏ bởi các điểm chia

$$-\infty < ... < m_{-2} < m_{-1} < m_0 < m_{-2} < m_{-2} < ... < \infty$$

Tích phân trong định nghĩa 2.5.2

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx = \sum_{i=-\infty}^{+\infty} \int_{i=m_i}^{m_{i+1}} x f(x)\, dx$$

Kí hiệu

$$p_i = \int_{i=m_i}^{m_{i+1}} f(x)\,dx = P\left(X \in (m_i, m_{i+1})\right)$$

khi đó tồn tại $x_i \in (m_i, m_{i+1})$ trong khoảng cận lấy tích phân sao cho

$$\int_{i=m_i}^{m_{i+1}} xf(x)\,dx = x_i pi$$

suy ra

$$E(X) = \sum_{i=-\infty}^{+\infty} x_i pi$$

Do vậy có thể coi định nghĩa 2.5.2 là mở rộng của định nghĩa 2.5.3 trong trường hợp X là đại lượng ngẫu nhiên liên tục.

Ví dụ 2.5.2

Trở lại ví dụ 2.4.2 một xạ thủ bắn bia. Bia là một hình tròn có bán kính bằng 1. Không gian các biến cố ngẫu nhiên cơ bản

$$\Omega = \{(u, v)/u^2 + v^2 \leq 1\}$$

Gọi ξ là khoảng cách từ điểm viên đạn chạm bia tới tâm bia

$$\xi = \sqrt{u^2 + v^2}$$

Hàm mật độ của ξ đã được xác định trong ví dụ 2.4.2

$$f(x) = \begin{cases} 2x & \text{nếu } 0 < x \leq 1 \\ 0 & \text{nếu } x \leq 0 \text{ hoặc } x > 1 \end{cases}$$

Hãy tìm kì vọng của ξ.

Giải: ξ là đại lượng ngẫu nhiên liên tục, theo định nghĩa 2.5.2

$$E(\xi) = \int_{-\infty}^{+\infty} xf(x)\,dx = \int_0^1 x \cdot 2x\,dx = \frac{2}{3}.$$

Kì vọng của đại lượng ngẫu nhiên có một vài tính chất cơ bản sau đây

i) Nếu X bị chặn, $|X| \leq K$, khi đó tồn tại kì vọng và

$$|E(X)| \leq K$$

ii) Nếu đại lượng ngẫu nhiên X là hằng số ($X \equiv C$), khi đó

$$E(X) = C$$

iii) Giả sử đại lượng ngẫu nhiên X có kì vọng, c là số thực bất kì, khi đó tồn tại $E(cX)$ và

$$E(cX) = cE(X)$$

Thật vậy i) và ii) là hiển nhiên, iii) được suy ra từ

$$E(cX) = \sum_{i=1}^{\infty} cx_i p_i = c\sum_{i=1}^{\infty} x_i p_i = cE(X)$$

nếu X là đại lượng ngẫu nhiên rời rạc có phân bố

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

Trường hợp X liên tục, với $f(x)$ là hàm mật độ, khi đó (theo ví dụ 2.4.3) hàm mật độ của cX bằng

$$g(y) = \frac{1}{|c|} f\left(\frac{y}{c}\right)$$

suy ra

$$E(cX) = \int_{-\infty}^{\infty} yg(y)\,dy = \int_{-\infty}^{\infty} \frac{y}{|c|} f(\frac{y}{c})\,dy = \int_{-\infty}^{\infty} f(u)du = cE(X)$$

Tương tự ta có

iv) Với số thực a bất kì

$$E(X + a) = E(X) + a$$

Định lí 2.5.1 *Nếu X và Y là hai đại lượng ngẫu nhiên có kì vọng, khi đó tồn tại $E(X + Y)$ và*

$$E(X + Y) = E(X) + E(Y).$$

Chứng minh Trường hợp X và Y là các đại lượng ngẫu nhiên liên tục, việc chứng minh định lí cần các kiến thức sâu hơn về giải tích, vì vậy ta hạn chế chỉ chứng minh trong trường hợp X và Y là hai đại lượng ngẫu nhiên rời rạc.

Kí hiệu $x_i, y_k, i, k = 1, 2, ...$ là các giá trị có thể có của X và Y tương ứng, A_{ik} là biến cố tích $\{X = x_i\} \cdot \{Y = y_k\}$. Hiển nhiên

$$\sum_i P(A_{ik}) = P(Y = y_k)$$

$$\sum_k P(A_{ik}) = P(X = x_i)$$

Mặt khác nếu z là giá trị nào đó của $X + Y$, khi đó z có thể biểu diễn dưới dạng $z = x_i + y_k$. Với mỗi z như vậy

$$zP(X + Y = z) = z \sum_{i,k: x_i + y_k = z} P(A_{ik}) = \sum_{x_i + y_k = z} (x_i + y_k) P(A_{ik}).$$

Theo định nghĩa kì vọng $E(X + Y) = \sum zP(X + Y = z)$. Suy ra

$$E(X + Y) = \sum_i \sum_k (x_i + y_k) P(A_{ik}) = E(X) + E(Y),$$

đ.p.c.m.

Từ định lí 2.5.1, bằng quy nạp ta có

Định lí 2.5.2 *Nếu tồn tại $E(X_i), i = 1, 2, ..., n$, khi đó $E(X_1 + X_2 + ... + X_n)$ cũng tồn tại và*

$$E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n).$$

Giả sử X là đại lượng ngẫu nhiên và φ là một hàm cho trước sao cho $Y = \varphi(X)$ là đại lượng ngẫu nhiên được hoàn toàn xác định. Nếu X là đại lượng ngẫu nhiên rời rạc có phân bố

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

khi đó $Y = \varphi(X)$ cũng là đại lượng ngẫu nhiên rời rạc có phân bố sau

Y	$\varphi(x_1)$	$\varphi(x_2)$	...	$\varphi(x_n)$	...
P	p_1	p_2	...	p_n	...

suy ra

$$E(Y) = E\left(\varphi(X)\right) = \sum_i \varphi(x_i)p_i.$$

Nếu X là đại lượng ngẫu nhiên liên tục với $f(x)$ là hàm mật độ. Ta có định lí sau

Định lí 2.5.3

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

Ta công nhận không chứng minh định lí này. Nhận xét rằng, để tính kì vọng của hàm một đại lượng ngẫu nhiên $Y = \varphi(X)$, ta có thể tính theo hai cách:

i) Áp dụng định nghĩa kì vọng $E(Y) = E(\varphi(X)) = \int_{-\infty}^{\infty} yg(y)dy$ trong đó $g(y)$ là hàm mật độ của Y.

ii) Áp dụng định lí 2.5.3 ở trên $E(Y) = \int_{-\infty}^{\infty} \varphi(x)f(x)dx$ trong đó $f(x)$ là hàm mật độ của X.

Ví dụ 2.5.3

Quay lại ví dụ 2.4.4, một xạ thủ bắn bia. Không gian các biến cố ngẫu nhiên cơ bản là hình tròn $\Omega = \{(u, v)/u^2 + v^2 \leq 1\}$. Gọi ξ là khoảng cách từ điểm viên đạn chạm bia tới tâm bia

$$\xi = \sqrt{u^2 + v^2}$$

Hãy tìm kì vọng của đại lượng ngẫu nhiên $\eta = \xi^2$

Giải: Trong ví dụ 2.4.4, chúng ta đã khẳng định một kết qủa: $\eta = \xi^2$ có phân bố đều trên đoạn [0,1]. Cụ thể hàm mật độ của $\eta = \xi^2$

$$g(y) = \begin{cases} 1 & \text{nếu } 0 < y \leq 1 \\ 0 & \text{nếu } y \leq 0 \text{ hoặc } y > 1 \end{cases}$$

Theo cách thứ nhất i) nêu trên

$$E(\xi^2) = \int_{-\infty}^{\infty} yg(y)dy = \int_0^1 ydy = \frac{1}{2}.$$

Áp dụng cách thứ hai ii)

$$E(\xi^2) = \int_{-\infty}^{\infty} x^2 f(x)dx,$$

trong đó hàm mật độ của ξ đã được tính trong ví dụ 2.4.2,

$$f(x) = \begin{cases} 2x & \text{nếu } 0 < x \leq 1 \\ 0 & \text{nếu } x \leq 0 \text{ hoặc } x > 1 \end{cases}$$

ta cũng được kết quả

$$E(\xi^2) = \int_0^1 x^2 \cdot 2x\, dx = \int_0^1 2x^3 dx = \frac{1}{2}.$$

Ví dụ 2.5.4

Chọn ngẫu nhiên một điểm trên đoạn thẳng có độ dài đơn vị (bằng 1). Điểm đó chia đoạn thẳng đã cho thành hai đoạn nhỏ. Hãy tìm độ dài trung bình của đoạn bé hơn trong số 2 đoạn thẳng tạo thành.

Giải: Trước hết chúng ta tìm hàm phân bố xác suất của độ dài đoạn bé. Gọi X là độ dài đoạn bé hơn trong số 2 đoạn thẳng tạo thành. Dễ dàng tính được hàm phân bố, hàm mật độ của X (bạn đọc tự chứng minh, chẳng hạn bằng phương pháp hình học):

$$F(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ 2x & \text{nếu } 0 < x \leq \frac{1}{2} \\ 1 & \text{nếu } x \geq \frac{1}{2} \end{cases} \qquad f(x) = \begin{cases} 2 & \text{nếu } 0 < x \leq \frac{1}{2} \\ 0 & \text{nếu } x \notin (0, \frac{1}{2}] \end{cases}$$

Từ định nghĩa 2.5.2 về kì vọng, suy ra độ dài trung bình của đoạn bé

$$E(X) = \int_0^{\frac{1}{2}} 2x\, dx = \frac{1}{4}.$$

Nhận xét rằng từ hàm mật độ $f(x)$ ở trên, X trong ví dụ này phân bố đều trên đoạn $[0, \frac{1}{2}]$.

2.5.2 Phương sai

Một đặc trưng khác hết sức quan trọng của đại lượng ngẫu nhiên đó là *phương sai*. Người ta muốn đo độ lệch của đại lượng ngẫu nhiên X so với kì vọng (giá trị trung bình) $E(X)$ cuả đại lượng ngẫu nhiên đó.

Tuy nhiên độ lệch giữa X và $E(X)$ có kì vọng

$$E(X - E(X)) = E(X) - E(X) = 0$$

(phần lớn hơn và nhỏ hơn của X so với giá trị trung bình $E(X)$ cân bằng nhau). Vì vậy hiển nhiên cần đưa vào giá trị trung bình

$$E|X - E(X)|$$

để nghiên cứu độ lệch $|X - E(X)|$. Tuy nhiên biểu thức đó về mặt toán học xử lí không đơn giản. Do vậy thay cho việc tính

$$E(|X - E(X)|),$$

người ta đưa vào khái niệm sau

Định nghĩa 2.5.3 *Giả sử $E(X)$ là kì vọng của X. Nếu tồn tại kì vọng của đại lượng ngẫu nhiên $(X - E(X))^2$, khi đó*

$$E\left(X - E(X)\right)^2$$

được gọi là phương sai của đại lượng ngẫu nhiên X và kí hiệu

$$D(X) = E\left(X - E(X)\right)^2$$

Căn bậc hai của phương sai, kí hiệu là

$$\sigma_X = \sqrt{D(X)}$$

được gọi là độ lệch tiêu chuẩn của đại lượng ngẫu nhiên X (hoặc gọi tắt là độ lệch chuẩn của đại lượng ngẫu nhiên X).

Áp dụng định lí 2.5.2 và định lí 2.5.3 suy ra công thức tính phương sai trong các trường hợp

i) Nếu X là đại lượng ngẫu nhiên rời rạc có phân bố

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

khi đó

$$D(X) = \sum_n p_n(x_n - E(X))^2$$

ii) Nếu X là đại lượng ngẫu nhiên liên tục với $f(x)$ là hàm mật độ. Ta có

$$D(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)dx$$

Nhận xét 2.5.1 *Do tính tuyến tính của kì vọng $E(.)$, suy ra:*

$$D(X) = E(X - E(X))^2 = E(X^2 - 2XE(X) + E(X^2)) =$$

$$= E(X^2) - 2[E(X)]^2 + E(X^2)) = E(X^2) - [E(X)]^2.$$

Sử dụng nhận xét này ta cũng có các công thức sau để tính phương sai

i) Trường hợp rời rạc

$$D(X) = \sum_n p_n x_n^2 - (\sum_n p_n x_n)^2$$

ii) Trường hợp liên tục

$$D(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - (E(X))^2$$

Phương sai của đại lượng ngẫu nhiên có các tính chất cơ bản sau đây

i) Nếu đại lượng ngẫu nhiên X là hằng số ($X \equiv C$), khi đó

$$D(X) = 0$$

ii) Giả sử đại lượng ngẫu nhiên X có phương sai, c là số thực bất kì, khi đó tồn tại $D(cX)$ và

$$D(cX) = c^2 D(X)$$

Thật vậy i) là hiển nhiên, ii) được suy ra từ định nghĩa

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

2.5.3 Kì vọng, phương sai của một vài phân bố thường gặp và các ví dụ

1. *Phân bố theo luật 0-1*

X là đại lượng ngẫu nhiên có phân bố theo luật 0-1 nếu bảng phân bố của X có dạng

X	1	0
P	p	q

trong đó $p+q=1,\ p>0, q>0$. Hiển nhiên

$$E(X)=p \quad D(X)=pq$$

2. *Phân bố nhị thức*

Gọi A là biến cố ngẫu nhiên có xác suất bằng p. Tiến hành n lần phép thử ngẫu nhiên độc lập nhau để quan sát biến cố A. Kí hiệu X là số lần xảy ra biến cố A trong số n lần tiến hành phép thử ngẫu nhiên kể trên. X là đại lượng ngẫu nhiên rời rạc, theo công thức Bernoulli (chương I.) phân bố của X có dạng

X	0	1	...	k	...	n
P	p_0	p_1	...	p_k	...	p_n

trong đó $p_k=C_n^k p^k q^{n-k}, \quad p+q=1,\ p>0, q>0, \quad k=0,1,...,n$

Định nghĩa 2.5.4 *Đại lượng ngẫu nhiên có bảng phân bố như trên được gọi là đại lượng ngẫu nhiên có phân bố nhị thức.*

i) Kì vọng của phân bố nhị thức

$$E(X)=\sum_{k=0}^{n} kC_n^k p^k q^{n-k}=$$

$$\sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k q^{n-k}=\sum_{k=1}^{n} n\frac{(n-1)!}{(k-1)!(n-k)!}p^k q^{n-k}=$$

$$np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}q^{n-k}=np\sum_{i=0}^{n-1}C_{n-1}^i p^i q^{n-1-i}=$$

$$np(p+q)^{n-1}=np$$

Nhận xét 2.5.2 *Ta có thể tính $E(X)$ gọn hơn như sau:*

Nếu kí hiệu X_i là đại lượng ngẫu nhiên nhận các giá trị 1 hoặc 0 tùy theo ở lần thử thứ i biến cố A xảy ra hay không

$$P(X_i = 1) = P(A) = p, P(X_i = 0) = P(\overline{A}) = q \quad i = 1, 2, ..., n$$

Hiển nhiên X_i phân bố theo luật 0-1 và $X_1 + X_2 + ... + X_n = k$ khi và chỉ khi có đúng k số 1 trong số n số hạng

$$X_1, X_2, ..., X_n$$

Nói cách khác biến cố $X_1 + X_2 + ... + X_n = k$ là biến cố có đúng k lần xảy ra A. Suy ra

$$X = X_1 + X_2 + ... + X_n$$

là đại lượng ngẫu nhiên có phân bố nhị thức. Do X_i phân bố theo luật 0-1, $E(X_i) = p$. Theo định lí 2.5.2

$$E(X) = E(X_1) + E(X_2) + ... + E(X_n) = np.$$

ii) Phương sai của phân bố nhị thức

Trước hết ta tính

$$\begin{aligned}
E(X^2) &= \sum_{k=0}^{n} k^2 C_n^k p^k q^{n-k} = \\
&= \sum_{k=0}^{n} (k + k(k-1)) \frac{n!}{k!(n-k)!} p^k q^{n-k} = \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k} + \\
&= \sum_{k=0}^{n} k(k-1) \frac{n!}{k!(n-k)!} p^k q^{n-k} = \\
&= E(X) + n(n-1) \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} p^k q^{n-2-(k-2)} = \\
&= E(X) + n(n-1)p^2 \sum_{i=0}^{n-2} C_{n-2}^i p^i q^{n-2-i} = \\
&= E(X) + n(n-1)p^2 (p+q)^{n-2} = np + n(n-1)p^2 = npq + n^2 p^2
\end{aligned}$$

Vậy

$$D(X) = E(X^2) - (EX)^2 = npq + n^2p^2 - (np)^2 = npq$$

Trở lại ví dụ 1.6.2 chương I.

Ví dụ 2.5.5

Hai đấu thủ bóng bàn vào chung kết tranh giải vô địch. Gọi p là xác suất để đấu thủ thứ nhất thắng đấu thủ thứ hai ở mỗi ván. Họ phải đấu với nhau tối đa 5 ván (người nào thắng trước 3 ván, là người đoạt chức vô địch). Tính số ván trung bình hai đấu thủ thi đấu với nhau (giả thiết rằng các ván họ đấu với nhau là độc lập).

Giả sử X là số ván hai đấu thủ thi đấu với nhau. Duy trì kí hiệu như trong ví dụ 30. chương I, gọi A_1 là biến cố "hai đấu thủ chỉ đấu với nhau 3 ván và đấu thủ thứ nhất thắng cả 3 ván đó"

$$P(A_1) = p^3$$

hoàn toàn tương tự, gọi B_1 là biến cố "hai đấu thủ chỉ đấu với nhau 3 ván và đấu thủ thứ hai thắng 3 ván đó"

$$P(B_1) = q^3 \qquad (q = 1 - p)$$

Biến cố $\{X = 3\}$ là biến cố "hai đấu thủ đấu với nhau 3 ván", hiển nhiên

$$P(X = 3) = P(A_1 + B_1) = P(A_1) + P(B_1) = p^3 + q^3 = 1 - 3pq$$

Biến cố $\{X = 4\}$ là biến cố "hai đấu thủ đấu với nhau 4 ván", cũng lập luận tương tự như trên, dựa theo kết quả ở ví dụ 1.6.2

$$P(X = 4) = C_3^2 p^3 q + C_3^2 q^3 p = 3pq(p^2 + q^2) = 3pq(1 - 2pq)$$

và xác suất của biến cố $\{X = 5\}$

$$P(X = 5) = C_4^2 p^3 q^2 + C_4^2 q^3 p^2 = 6p^2 q^2 (p + q) = 6p^2 q^2.$$

Vậy số ván trung bình hai đấu thủ thi đấu với nhau

$$\begin{aligned} E(X) &= 3P(X=3) + 4P(X=4) + 5P(X=5) \\ &= 3(1-3pq) + 12pq(1-2pq) + 30p^2q^2 \\ &= 3(1+pq+2p^2q^2). \end{aligned}$$

Nhận xét rằng $E(X)$ là một hàm của pq và do $p+q=1$ nên $E(X)$ đạt giá trị lớn nhất khi và chỉ khi $p=q=\frac{1}{2}$. Trường hợp $p=q=\frac{1}{2}$ ta nói hai đấu thủ ngang sức và số ván trung bình họ đấu với nhau

$$E(X) = 3(1+pq+2p^2q^2) = \frac{33}{8} = 4,125.$$

Như vậy nếu hai đấu thủ chênh lệch về trình độ, số ván trung bình cần đấu sẽ ít hơn.

Để đo độ lệch giữa số ván họ đấu với nhau và kì vọng $E(X)$, ta cần tính cả phương sai $D(X)$. Ở ví dụ này, dễ dàng tính được phương sai

$$D(X) = 3^2P(X=3) + 4^2P(X=4) + 5^2P(X=5) - (EX)^2 \leq \frac{39}{64}$$

Phương sai sẽ đạt giá trị lớn nhất $D(X) = \frac{39}{64} \approx 0,61$ khi và chỉ khi $p=q=\frac{1}{2}$

Ví dụ 2.5.6

Một bài thi trắc nghiệm khách quan gồm 20 câu hỏi. Mỗi câu hỏi có 4 phương án chọn lựa, trong đó có duy nhất một phương án đúng mà thí sinh phải chỉ ra khi làm bài. Nếu một thí sinh không học bài, chọn ngẫu nhiên một trong 4 phương án ở tất cả các câu. Hãy tính trung bình thí sinh làm được bao nhiêu câu đúng? Tính xác suất để thí sinh làm đúng không quá 8 câu.

Gọi X là số câu thí sinh làm đúng. X có phân bố nhị thức với $n=20$ và $p=\frac{1}{4}$. Vậy trung bình số câu đúng thí sinh làm được là

$$E(X) = np = 20\frac{1}{4} = 5 \quad \text{câu.}$$

Xác suất để thí sinh làm đúng không quá 8 câu bằng

$$P(X \leq 8) = \sum_{i=0}^{8} C_{20}^i \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{20-i} \approx 0,959075.$$

Tra bảng phân bố nhị thức hoặc sử dụng các phần mềm tính toán ta tính một vài giá trị hàm phân bố của X

Số câu đúng	≤ 5	≤ 6	≤ 7	≤ 8
Xác suất	0,617173	0,785782	0,898188	0,959075

Như vậy có thể nói hầu như chắc chắn (với xác suất 0,9590) nếu làm bài theo kiểu chọn ngẫu nhiên một phương án, thí sinh làm được không quá 8 câu đúng (trên tổng số 20 câu toàn bài thi). Căn cứ vào kết quả trên ta có thể định ra thang điểm để đánh giá đúng năng lực của học sinh khi cho thi dưới hình thức trắc nghiệm khách quan.

Ví dụ 2.5.7 (Bài toán bao diêm của Banach)

Giáo sư Banach nghiện thuốc lá và rất hay đãng trí. Ông để hai bao diêm, mỗi bao trong một túi áo và mỗi lần hút thuốc ông chọn ngẫu nhiên một bao (bên túi áo trái hoặc túi áo phải). Cho tới một lần khi lấy diêm để hút thuốc ông chọn phải bao đã hết diêm. Hỏi khi đó bao diêm ở túi áo bên kia còn lại trung bình bao nhiêu que diêm? (Biết rằng ban đầu mỗi bao gồm n que.)

Gọi X là số que diêm còn lại ở túi áo bên kia. Xét hai trường hợp, thứ nhất khi giáo sư chọn phải bao đã hết diêm trong túi áo bên trái. Gọi A là biến cố giáo sư chọn phải bao đã hết diêm trong túi áo bên trái và bao diêm trong túi áo phải còn đúng k que. Như vậy ông đã $n-k$ lần lấy diêm từ túi áo phải và $n+1$ lần lấy diêm từ túi áo trái (lần cuối cùng ông phát hiện ra bao diêm túi bên trái đã hết).

Có thể coi vấn đề cũng giống như bài toán lang thang ngẫu nhiên: số trường hợp đồng khả năng $= 2^{2n-k+1}$ và số trường hợp thuận lợi cho biến cố A bằng C^n_{2n-k}. Vậy xác suất của A

$$P(A) = \frac{C^n_{2n-k}}{2^{2n-k+1}}$$

Trường hợp thứ hai xác suất cũng đúng bằng như vậy nếu giáo sư chọn phải bao đã hết diêm trong túi áo bên phải và bao diêm trong túi áo trái còn lại đúng k que. Suy ra

$$P(X=k) = 2P(A) = \frac{C^n_{2n-k}}{2^{2n-k}} \quad \text{với} \quad k = 0, 1, 2, ..., n$$

Hiển nhiên

$$\sum_{k=0}^{n} P(X=k) = \sum_{k=0}^{n} \frac{C_{2n-k}^{n}}{2^{2n-k}} = 1$$

Theo định nghĩa kì vọng, trung bình số que diêm còn lại trong túi áo bên kia là

$$E(X) = \sum_{k=0}^{n} kP(X=k) = \sum_{k=0}^{n} k\frac{C_{2n-k}^{n}}{2^{2n-k}}$$

Để tính tổng này ta biến đổi như sau

$$n - E(X) = \sum_{k=0}^{n-1}(n-k)\frac{C_{2n-k}^{n}}{2^{2n-k}} = \sum_{k=0}^{n-1}(n-k)\frac{C_{2n-k}^{n-k}}{2^{2n-k}} =$$

$$= \sum_{k=0}^{n-1}(2n-k)\frac{C_{2n-k-1}^{n-k-1}}{2^{2n-k}} = \sum_{k=0}^{n-1}((2n+1)-(k+1))\frac{C_{2n-k-1}^{n-k-1}}{2^{2n-k}} =$$

$$= \frac{2n+1}{2}\sum_{k=0}^{n-1}\frac{C_{2n-k-1}^{n-k-1}}{2^{2n-k-1}} - \frac{1}{2}\sum_{k=0}^{n-1}(k+1)\frac{C_{2n-k-1}^{n-k-1}}{2^{2n-k-1}} =$$

$$= \frac{2n+1}{2}\sum_{k=1}^{n}\frac{C_{2n-k}^{n-k}}{2^{2n-k}} - \frac{1}{2}\sum_{k=1}^{n}k\frac{C_{2n-k}^{n-k}}{2^{2n-k}}$$

Sử dụng $\sum_{k=0}^{n} P(X=k) = \sum_{k=0}^{n}\frac{C_{2n-k}^{n-k}}{2^{2n-k}} = 1$, suy ra

$$n - E(X) = \frac{2n+1}{2}(1 - \frac{C_{2n}^{n}}{2^{2n}}) - \frac{E(X)}{2}$$

hay

$$E(X) = 2n - (2n+1)\left(1 - \frac{C_{2n}^{n}}{2^{2n}}\right) = \frac{2n+1}{2^{2n}}C_{2n}^{n} - 1.$$

Nhận xét rằng nếu sử dụng công thức *Stirling*

$$n! \sim n^n e^{-n}\sqrt{2\pi n}$$

ta có thể chứng minh

$$E(X) \approx \frac{2}{\sqrt{\pi}}\sqrt{n} \quad \text{khi} \quad n \to \infty \qquad \left(\frac{2}{\sqrt{\pi}} = 1,128..\right)$$

Đặc biệt khi mỗi bao gồm $n = 50$ que diêm, khi đó $E(X) \approx \frac{2}{\sqrt{\pi}}\sqrt{50} = 7,978...$ Nói cách khác nếu một bao rỗng thì bao diêm kia trung bình còn lại khoảng 8 que diêm.

3. *Phân bố Poisson*

Ta dẫn vào định nghĩa sau

Định nghĩa 2.5.5 *Đại lượng ngẫu nhiên X có phân bố*

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \textit{với mọi } k = 0, 1, 2, ... \textit{và } \lambda > 0$$

được gọi là đại lượng ngẫu nhiên có phân bố Poisson.

Ta xét vấn đề sau:

Một biến cố A xảy ra ở những thời điểm ngẫu nhiên (chẳng hạn như cuộc gọi tới một tổng đài điện thoại nào đó, sự phân rã của nguyên tử phóng xạ,...) thoả mãn các yêu cầu:

i) Số lần xuất hiện biến cố A trong khoảng thời gian (t_1, t_2) độc lập với số lần xuất hiện biến cố A trong khoảng thời gian kế tiếp (t_2, t_3).

ii) Xác suất để trong khoảng thời gian (t_1, t_2) có đúng k lần xảy ra biến cố A chỉ phụ thuộc vào $t_2 - t_1$

iii) Khi $t = t_2 - t_1$ đủ nhỏ ($t \to 0$) xác suất để có ít nhất một lần xảy ra biến cố A và xác suất để có đúng một lần xảy ra biến cố A trong khoảng thời gian (t_1, t_2) là hai vô cùng bé tương đương.

Khi đó người ta chứng minh được rằng số lần xảy ra biến cố A trong một khoảng thời gian nào đó là đại lượng ngẫu nhiên có phân bố Poisson.

Vì vậy trong thực tế ứng dụng người ta coi số các cuộc gọi tới một tổng đài điện thoại trong một khoảng thời gian nào đó, hay số người tới một cửa hàng bách hoá mua hàng trong một ngày,... là các đại lượng ngẫu nhiên có phân bố Poisson.

i) Kì vọng của phân bố Poisson

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

ii) Phương sai của phân bố Poisson

Trước hết ta tính

$$E(X^2) = \sum_{k=0}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=0}^{\infty}(k + k(k-1))e^{-\lambda}\frac{\lambda^k}{(k-1)!}$$
$$= \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} + \sum_{k=0}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} =$$
$$= E(X) + e^{-\lambda}\lambda^2 \sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \lambda + \lambda^2$$

Do vậy

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

Nhận xét rằng do $E(X) = D(X) = \lambda$ nên tham số λ trong phân bố Poisson vừa là giá trị trung bình, vừa là phương sai của đại lượng ngẫu nhiên X.

Ví dụ 2.5.8

Trung bình số các cuộc gọi tới một tổng đài điện thoại A nào đó trong khoảng thời gian 1 phút bằng 3. Hãy tìm xác suất để

i) Trong khoảng thời gian 2 phút có đúng 5 cuộc gọi.

ii) Trong khoảng thời gian 30 giây có nhiều nhất 3 cuộc gọi.

Giải:

i) Do số các cuộc gọi tới tổng đài trong khoảng thời gian 1 phút là đại lượng ngẫu nhiên có phân bố Poisson. Kì vọng (giá trị trung bình) của đại lượng ngẫu nhiên, theo giả thiết bằng 3. Suy ra số các cuộc gọi tới tổng đài trong khoảng thời gian 2 phút cũng là đại lượng ngẫu nhiên (kí hiệu X) có phân bố Poisson với tham số $\lambda = 3 \cdot 2 = 6$. Vậy

$$P(X = 5) = e^{-6}\frac{6^5}{5!} = 0,160623$$

ii) Số cuộc gọi tới tổng đài trong khoảng thời gian 30 giây (0,5 phút) cũng là đại lượng ngẫu nhiên (kí hiệu Y) có phân bố Poisson với tham số $\lambda = 3 \cdot 0,5 = 1,5$. Vậy xác suất để có nhiều nhất 3 cuộc gọi bằng

$$P(Y \le 3) = e^{-1,5} + e^{-1,5}\frac{1,5}{1!} + e^{-1,5}\frac{1,5^2}{2!} + e^{-1,5}\frac{1,5^3}{3!} \approx 0,934358$$

4. *Phân bố hình học*

Trước hết ta xét ví dụ sau: Xác suất bắn trúng bia của một xạ thủ là p. Giả sử kết quả các lần bắn độc lập nhau. Để bắn trúng bia, hỏi trung bình xạ thủ đó cần bắn bao nhiêu viên đạn?

Gọi X là số đạn xạ thủ đó đã sử dụng. Biến cố $\{X = n\}$ là biến cố "$n-1$ lần bắn đầu, xạ thủ đó bắn trượt và ở lần bắn thứ n xạ thủ đó mới bắn trúng bia". Ở ví dụ 2.2.1 chương II chúng ta đã tính xác suất cuả biến cố đó

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

Một cách tổng quát ta có định nghĩa sau:

Định nghĩa 2.5.6 *Đại lượng ngẫu nhiên X có phân bố*

X	1	2	...	n	...
P	p	qp	...	$q^{n-1}p$	...

trong đó $p + q = 1$ được gọi là đại lượng ngẫu nhiên có phân bố hình học.

i) Kì vọng của phân bố hình học

Trong lí thuyết về chuỗi hàm ta đã biết với $\forall x \in (-1, 1)$

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

Đạo hàm hai vế theo x, ta được

$$\sum_{n=1}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$$

Áp dụng để tính kì vọng

$$E(X) = \sum_{n=1}^{\infty} nq^{n-1}p = p\frac{1}{(1-q)^2} = p\frac{1}{p^2} = \frac{1}{p}$$

Vậy trong ví dụ trên, để bắn trúng bia, số đạn trung bình xạ thủ đó cần bắn (số đạn chi phí) là $\frac{1}{p}$ viên.

ii) Phương sai của phân bố hình học

Đạo hàm theo x chuỗi hàm $\sum_{n=1}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$ một lần nữa

$$\sum_{n=2}^{\infty} n(n-1)x^{n-2} = \frac{2}{(1-x)^3}$$

Áp dụng để tính kì vọng

$$\begin{aligned} E(X^2) &= \sum_{n=1}^{\infty} n^2 q^{n-1} p = \sum_{n=1}^{\infty} (n + n(n-1)) q^{n-1} p \\ &= E(X) + \sum_{n=2}^{\infty} n(n-1) q^{n-1} p = \frac{1}{p} + \frac{2pq}{(1-q)^3} = \frac{1+q}{p^2} \end{aligned}$$

Suy ra phương sai của X

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

Ví dụ 2.5.9

Một xạ thủ tập bắn bia, bắn liên tục vào bia cho đến khi có đủ 3 viên đạn trúng bia thì dừng. Giả sử xác suất bắn trúng bia (mỗi lần bắn) của xạ thủ đó bằng p và các lần bắn độc lập nhau. Hãy tính số đạn trung bình mà xạ thủ đã sử dụng.

Gọi số đạn mà xạ thủ đó đã sử dụng là đại lượng ngẫu nhiên X. Biến cố $X = n$ xảy ra đồng nghĩa với tích 2 biến cố: "trong $n-1$ lần bắn đầu có đúng 2 viên đạn trúng bia" và "lần bắn thứ n xạ thủ bắn trúng bia". Do các lần bắn độc lập nhau xác suất của tích 2 biến cố trên bằng tích các xác suất. Mặt khác xác suất của biến cố "có đúng 2 viên đạn trúng bia trong $n-1$ lần bắn đầu" được tính bằng công thức Bernoulli

$$C_{n-1}^2 p^2 q^{n-1-2} \qquad n = 3, 4, ...$$

Vì vậy

$$P(X = n) = C_{n-1}^2 p^2 q^{n-1-2} \cdot p = C_{n-1}^2 p^3 q^{n-3}.$$

(Nhận xét rằng đại lượng ngẫu nhiên với phân bố

$$p_n = P(X = n) = C^2_{n-1}p^3q^{n-3}$$

được gọi là đại lượng ngẫu nhiên có *phân bố hình học cấp 3* sẽ được đề cập tới trong nhận xét 3, chương III).

Do đó số đạn trung bình hay kì vọng của X bằng

$$E(X) = \sum_{n=3}^{\infty} nC^2_{n-1}p^3q^{n-3} = \frac{3}{p}\sum_{n=3}^{\infty} C^3_n p^4 q^{n-3}$$

Sử dụng chuỗi nhị thức đã biết trong giải tích (hoặc bằng con đường tính xác suất tương tự như trên)

$$\sum_{n=3}^{\infty} C^3_n p^4 q^{n-3} = 1$$

suy ra

$$E(X) = \frac{3}{p}.$$

Điều này cũng phù hợp với thực tế nếu xác suất bắn trúng bia của xạ thủ bằng p, khi đó để 3 lần trúng bia, trung bình xạ thủ đó phải bắn $\frac{3}{p}$ lần.

Ví dụ 2.5.10

Một người đánh bạc đặt tiền vào một cửa (ví dụ người đó đặt tiền vào mặt 6 chấm khi gieo xúc xắc), nếu mặt 6 chấm không xuất hiện người đó nâng tiền đặt lên gấp đôi ở lần gieo sau. Giả sử ban đầu số tiền đặt của người đó là 1 đồng, nếu mặt 6 chấm không xuất hiện người đó nâng tiền đặt lên 2 đồng ở lần thứ hai, rồi lên 4 đồng ở lần thứ ba,... Sau $n-1$ lần chơi không xuất hiện mặt 6 chấm, khi đó ở lần chơi thứ n người đó tăng số tiền đặt lên thành 2^{n-1} đồng. Người đánh bạc chơi "chung thân" theo kiểu "gấp thếp" như vậy cho đến khi thắng thì dừng. Hãy tính trung bình số tiền lãi của người đó.

Gọi X là số lần gieo xúc xắc cho tới khi mặt 6 chấm xuất hiện. Hiển nhiên X là đại lượng ngẫu nhiên có phân bố hình học.

Gọi Y là số tiền lãi của người đánh bạc. Nếu biến cố $X = n$ xảy ra, số tiền người đánh bạc thua ở $n-1$ lần chơi đầu là

$$1 + 2^1 + 2^2 + ... + 2^{n-2} = 2^{n-1} - 1$$

và lần thứ n được bạc (2^{n-1} đồng). Do vậy số tiền lãi bằng

$$2^{n-1} - (2^{n-1} - 1) = 1.$$

Như vậy đại lượng ngẫu nhiên Y là hằng số (luôn luôn bằng 1), suy ra $E(Y) = 1$.

Bài toán người đánh bạc nói trên không chỉ đúng với gieo xúc xắc mà với cả các trò chơi quay số ở các khu vui chơi giải trí, thậm chí cả chơi đề. Với chiến lược chơi "gấp thếp" như trên, số tiền lãi trung bình của người chơi $E(Y)$ là số dương, cách chơi dường như có vẻ đảm bảo phần thắng cho người đánh bạc. Thực ra điều khẳng định đó chỉ đúng nếu người đánh bạc có số tiền vô hạn. Bây giờ ta sẽ giải bài toán trên trong trường hợp người đánh bạc chỉ có một lượng tiền hữu hạn.

Giả sử ban đầu người đánh bạc có số tiền T_0, $\quad 2^{n-1} - 1 \leq T_0 < 2^n - 1$. Nếu biến cố $A = \{X < n\}$ xảy ra số tiền lãi của người đó như đã tính ở trên $Y = 1$, nếu biến cố A không xảy ra, tức là ở lần gieo thứ $n-1$ biến cố mặt 6 chấm vẫn chưa xuất hiện, người đó thua

$$1 + 2^1 + 2^2 + ... + 2^{n-2} = 2^{n-1} - 1 \quad \text{đồng}$$

hay $Y = -(2^{n-1} - 1)$. Do số tiền của người đánh bạc không còn đủ để đặt chơi tiếp ($2^{n-1} - 1 \leq T_0 < 2^n - 1$), người đó phải dừng chơi suy ra số tiền lãi (đại lượng ngẫu nhiên Y) được xác định như sau

$$Y = \begin{cases} 1 & \text{nếu } A \text{ xảy ra} \\ -(2^{n-1} - 1) & \text{nếu } A \text{ không xảy ra} \end{cases}$$

Do X có phân bố hình học với tham số $p = \frac{1}{6}$ nên

$$P(A) = P(X < n) = \sum_{k=0}^{n-1} q^k p = 1 - \frac{5}{6^{n-1}} \Rightarrow P(\overline{A}) = \frac{5}{6^{n-1}}$$

Vậy số tiền lãi trung bình

$$E(Y) = 1 \cdot P(A) - (2^{n-1} - 1) \cdot P(\overline{A}) = 1 - \left(\frac{10}{6}\right)^{n-1} < 0.$$

Như vậy nếu số lượng tiền T_0 là hữu hạn (T_0 có thể rất lớn) kì vọng luôn luôn âm.

Trường hợp người đánh bạc áp dụng chiến lược trên cho trò chơi "số đề", dễ dàng thấy kì vọng $E(Y)$ là số âm có trị tuyệt đối rất lớn.

5. *Phân bố siêu bội*

Ta xét bài toán dẫn đến phân bố siêu bội sau đây.

Một lô hàng có N sản phẩm, trong đó có M phế phẩm ($M < N$). Lần lượt chọn ngẫu nhiên ra n sản phẩm, các sản phẩm đã chọn ra không đặt trở lại. Nói cách khác đây là bài toán chọn ngẫu nhiên không hoàn lại. Gọi X là số phế phẩm trong số n sản phẩm chọn ra. (Giả thiết $n < M, n < N - M$)

Khi đó xác suất để có đúng k phế phẩm trong số n sản phẩm chọn ra bằng

$$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$$

(*Nhấn mạnh lại một lần nữa đây là bài toán chọn ngẫu nhiên không hoàn lại. Nếu bài toán gắn với giả thiết chọn mẫu có hoàn lại, khi đó phân bố của X là phân bố nhị thức.*)

Hiển nhiên các biến cố $P(X = k), \quad k = 0, 1, .., n$ lập thành một hệ đầy đủ, suy ra

$$\sum_{k=0}^{n} \frac{C_M^k C_{N-M}^{n-k}}{C_N^n} = 1.$$

Ta dẫn vào định nghĩa sau

Định nghĩa 2.5.7 *Đại lượng ngẫu nhiên X có phân bố*

$$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n} \quad (k = 0, 1, .., n)$$

được gọi là đại lượng ngẫu nhiên có phân bố siêu bội.

i) Kì vọng của phân bố siêu bội

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} k\frac{C_M^k C_{N-M}^{n-k}}{C_N^n} = \sum_{k=1}^{n} \frac{(kC_M^k)C_{N-M}^{n-k}}{C_N^n} \\ &= \sum_{k=1}^{n} \frac{MC_{M-1}^{k-1}C_{N-M}^{(n-1)-(k-1)}}{\frac{N}{n}C_{N-1}^{n-1}} \\ &= n\frac{M}{N}\sum_{k=0}^{n-1} \frac{C_{M-1}^k C_{(N-1)-(M-1)}^{(n-1)-k}}{C_{N-1}^{n-1}} \end{aligned}$$

Do $\frac{C_{M-1}^k C_{(N-1)-(M-1)}^{(n-1)-k}}{C_{N-1}^{n-1}}$ $(k = 0, 1, .., n-1)$ cũng là các số hạng của phân bố siêu bội nên

$$\sum_{k=0}^{n-1} \frac{C_{M-1}^k C_{(N-1)-(M-1)}^{(n-1)-k}}{C_{N-1}^{n-1}} = 1$$

Vậy kì vọng của phân bố siêu bội:

$$E(X) = n\frac{M}{N}.$$

ii) Phương sai của phân bố siêu bội

Phương sai của phân bố siêu bội cũng được tính tương tự như tính phương sai của các phân bố nhị thức, phân bố Poisson:

$$D(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\left(1 - \frac{n-1}{N-1}\right).$$

Đặt $p = \frac{M}{N}, q = 1 - p$, khi đó

$$D(X) = npq\left(1 - \frac{n-1}{N-1}\right).$$

Nhớ lại rằng phương sai của phân bố nhị thức bằng npq, như vậy phương sai của phân bố siêu bội nhỏ hơn phương sai của phân bố nhị thức. Nếu N tăng ra vô hạn trong khi n tăng chậm hơn, khi đó sự chênh lệch giữa

phương sai của phân bố siêu bội và phương sai của phân bố nhị thức trở thành không đáng kể. (Trong trường hợp đó người ta không phân biệt giữa chọn mẫu có hoàn lại và chọn mẫu không hoàn lại).

Ví dụ 2.5.11

Để ước lượng số cá trong hồ, người ta sử dụng phương pháp sau: bắt M con cá trong hồ và bằng một phương pháp nào đó để đánh dấu M con cá đó (đeo vòng vào đuôi cá chẳng hạn), sau đó thả trả lại hồ. Sau một thời gian để các con cá đã đánh dấu hoàn toàn trộn đều với các con khác, ta bắt một đợt mới n con cá từ hồ. Gọi X là số cá đã đánh dấu trong số n con mới bắt lên. Khi đó số cá trong hồ được ước lượng vào khoảng $M\frac{n}{X}$ con. Tuy nhiên X có thể nhận giá trị 0, vì vậy thay cho $E\left(M\frac{n}{X}\right)$ người ta ước lượng số cá trong hồ bằng $E\left(M\frac{n}{X+1}\right)$. Hãy tìm kì vọng đó.

Giả sử số cá trong hồ ban đầu gồm N con. Khi đó X là đại lượng ngẫu nhiên có phân bố siêu bội

$$P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, \quad (k=0,1,..,n)$$

Vậy

$$E\left(M\frac{n}{X+1}\right)=M\sum_{k=0}^{n}\frac{n}{k+1}\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}=$$

$$=M\sum_{k=0}^{n}\frac{n}{M+1}\frac{C_{M+1}^{k+1} C_{N-M}^{n-k}}{C_N^n}=$$

$$=M\frac{n}{M+1}\sum_{k=0}^{n}\frac{N+1}{n+1}\frac{C_{M+1}^{k+1} C_{(N+1)-(M+1)}^{(n+1)-(k+1)}}{C_{N+1}^{n+1}}=$$

$$=(N+1)\frac{M}{M+1}\frac{n}{n+1}\sum_{i=1}^{n+1}\frac{C_{M+1}^{i} C_{(N+1)-(M+1)}^{(n+1)-i}}{C_{N+1}^{n+1}}$$

Do $\frac{C_{M+1}^{i} C_{(N+1)-(M+1)}^{(n+1)-i}}{C_{N+1}^{n+1}} \quad (i=0,1,..,n+1)$ cũng là các số hạng của phân bố siêu bội nên

$$\sum_{i=1}^{n+1}\frac{C_{M+1}^{i} C_{(N+1)-(M+1)}^{(n+1)-i}}{C_{N+1}^{n+1}}=1-\frac{C_{N-M}^{n+1}}{C_{N+1}^{n+1}}$$

Vậy kì vọng của $M\frac{n}{X+1}$:

$$E\left(\frac{Mn}{X+1}\right) = (N+1)\frac{n}{n+1}\frac{M}{M+1}\left(1 - \frac{C_{N-M}^{n+1}}{C_{N+1}^{n+1}}\right)$$

xấp xỉ với số cá trong hồ (N con).

6. *Phân bố đều*

Xét đại lượng ngẫu nhiên ξ có phân bố đều trên đoạn [a,b]. Kji đó ξ là đại lượng ngẫu nhiên liên tục với hàm mật độ

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

Kì vọng của ξ

$$E(\xi) = \int_{-\infty}^{+\infty} xf(x)\,dx = \int_a^b \frac{x}{b-a}\,dx = \frac{a+b}{2}.$$

Để tính phương sai của ξ, trước hết ta tính

$$E(\xi^2) = \int_{-\infty}^{+\infty} x^2 f(x)\,dx = \int_a^b \frac{x^2}{b-a}\,dx = \frac{a^2+b^2+ab}{3}.$$

Vậy

$$D(\xi) = E(\xi^2) - (E\xi)^2 = \frac{a^2+b^2+ab}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(a-b)^2}{12}.$$

7. *Phân bố mũ*

Trước hết chúng ta nghiên cứu bài toán về khoảng thời gian giữa hai cuộc gọi điện thoại liên tiếp. Đó là đại lượng ngẫu nhiên và gọi X là đại lượng ngẫu nhiên đó. Kí hiệu $A_s = \{X > s\}$ là biến cố trong khoảng thời gian $(0, s)$ không có cuộc gọi điện thoại nào. Khi đó người ta nghiệm thấy rằng xác suất để không có cuộc gọi điện thoại nào trong khoảng thời gian kế tiếp $(s, s+t)$, với điều kiện A_s xảy ra, chỉ phụ thuộc vào t (độ dài của khoảng thời gian $(s, s+t)$) và không phụ thuộc vào thời điểm s trước đó.

Nói cách khác $P(A_{s+t}/A_s) = P(A_t)$. Gọi $F(t) = P(X < t)$ là hàm phân bố của X, kí hiệu $G(t) = 1 - F(t)$, khi đó $G(t) = P(A_t)$ và $G(0) = 1$. Theo định nghĩa xác suất có điều kiện

$$P(A_{s+t}) = P(A_{s+t}/A_s)P(A_s)$$

hay

$$G(s+t) = G(s)G(t) \quad s > 0, t > 0.$$

suy ra

$$\frac{G(t+\Delta t) - G(t)}{\Delta t} = G(t)\frac{G(\Delta t) - 1}{\Delta t}$$

Nếu ta giả thiết rằng hàm $G(t)$ khả vi tại $t = 0$, cho $\Delta t \to 0$, khi đó

$$G'(t) = G'(0)G(t).$$

Do hàm $G(t)$ đơn điệu giảm, nên $G'(0) < 0$. Đặt $G'(0) = -\lambda$ ($\lambda > 0$) và sử dụng điều kiện đầu $G(0) = 1$, phương trình trên có nghiệm

$$G(t) = e^{-\lambda t}$$

Vậy hàm phân bố của X bằng

$$F(t) = 1 - G(t) = 1 - e^{-\lambda t} \quad \text{với} \quad t > 0$$

và hàm mật độ $f(x) = F'(x) = \lambda e^{-\lambda x}$ với $x > 0$ vì

$$F(t) = \int_0^t \lambda e^{-\lambda x}\, dx \quad \text{với} \quad t > 0$$

Nhận xét rằng ta có thể chứng minh kết quả trên không sử dụng tính khả vi tại $t = 0$ của $G(t)$, thay vào đó chỉ cần giả thiết $G(t)$ là hàm đơn điệu.

Thật vậy dễ dàng suy ra từ giả thiết

$$G(s+t) = G(s)G(t) \quad s > 0, t > 0$$

rằng với mọi số tự nhiên m, n

$$G\left(\frac{m}{n}t\right) = (G(t))^{\frac{m}{n}}$$

hay

$$G(r) = (G(1))^r$$

với bất kì số hữu tỉ $r \geq 0$. Do $G(1) < 1$ đặt $G(1) = e^{-\lambda}$ và sử dụng tính đơn điệu của $G(t)$, suy ra điều phải chứng minh

$$G(t) = e^{-\lambda t} \quad \text{với mọi số thực} \quad t \geq 0.$$

Định nghĩa 2.5.8 *Đại lượng ngẫu nhiên X có hàm mật độ*

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

trong đó λ là tham số dương, được gọi là đại lượng ngẫu nhiên có phân bố mũ.

Qua ví dụ trên, những vấn đề tương tự như thời gian hoạt động liên tục của một hệ thống giữa hai lần hỏng hóc, tuổi thọ của các linh kiện máy, khoảng thời gian để một nguyên tử radium tự phân hủy... là các đại lượng ngẫu nhiên có phân bố mũ.

i) Kì vọng của phân bố mũ

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx = \int_0^{+\infty} x\lambda e^{-\lambda x}\, dx$$

Sử dụng tích phân từng phần

$$E(X) = \left[-xe^{-\lambda x}\right]_0^{\infty} + \int_0^{+\infty} e^{-\lambda x}\, dx = \frac{1}{\lambda}.$$

ii) Phương sai của phân bố mũ

Áp dụng công thức $D(X) = E(X^2) - (EX)^2$ để tìm phương sai, trước hết tích phân từng phần

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{+\infty} x^2 f(x)\, dx = \int_0^{+\infty} x^2 \lambda e^{-\lambda x}\, dx \\ &= \left[-x^2 e^{-\lambda x}\right]_0^{\infty} + 2\int_0^{+\infty} xe^{-\lambda x}\, dx = \frac{2}{\lambda^2}. \end{aligned}$$

Suy ra phương sai của phân bố mũ

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

Ví dụ 2.5.12

Ta gọi X là tuổi thọ của một nguyên tử radium. X là đại lượng ngẫu nhiên có phân bố mũ. Xác suất để một nguyên tử nào đó cho đến thời điểm t chưa tan rã (nói cách khác tuổi thọ của nguyên tử đó X lớn hơn t) bằng

$$P(X > t) = 1 - F(t) = e^{-\lambda t}$$

Biết rằng chu kì bán rã của radium bằng $T = 1580$ năm. Diễn đạt theo ngôn ngữ xác suất điều đó có nghĩa là trong khoảng thời gian $T = 1580$ năm, xác suất để một nguyên tử nào đó tan rã bằng $\frac{1}{2}$. Hay nói cách khác

$$F(T) = \frac{1}{2} \quad \Rightarrow \quad e^{-\lambda T} = \frac{1}{2}$$

suy ra

$$\lambda = \frac{\ln 2}{T} = \frac{0,69315}{1580} \quad \text{hay} \quad \frac{1}{\lambda} = 2279.$$

Như trên chúng ta đã tính kì vọng của phân bố mũ bằng $E(X) = \frac{1}{\lambda}$. Vậy tuổi thọ trung bình của một nguyên tử radium bằng 2279 năm.

Ví dụ 2.5.13

Tuổi thọ trung bình của một loại thiết bị là 5 năm. Biết tuổi thọ của thiết bị đó là đại lượng ngẫu nhiên có phân bố mũ. Nhà máy sản xuất thiết bị bảo hành 1 năm cho sản phẩm của mình. Hỏi trong thời gian bảo hành bao nhiêu phần trăm sản phẩm bán ra phải thay thế?

Gọi X là tuổi thọ của thiết bị đó. Tham số $\lambda = \frac{1}{E(X)} = \frac{1}{5} = 0,2$.

Xác suất để thiết bị hỏng phải thay thế trong khoảng thời gian bảo hành 1 năm là

$$P(X \leq 1) = F(1) = 1 - e^{-\lambda} = 1 - e^{-0,2} \approx 0,1813$$

Vậy số sản phẩm phải thay thế trong thời gian bảo hành chiếm 18%.

Ví dụ 2.5.14

Một ví dụ khác, giả sử tuổi thọ của một loại bóng đèn thắp sáng là đại lượng ngẫu nhiên có phân bố mũ. Biết tuổi thọ trung bình của bóng đèn là 8 tháng. Người ta thắp sáng một khu vực với 6 bóng đèn loại đó và chỉ thay thế bóng hỏng khi tất cả 6 bóng đều bị cháy. Hãy tìm xác suất để trong vòng 1 năm không phải thay bóng đèn, giả thiết rằng tuổi thọ của các bóng đèn là các đại lượng ngẫu nhiên độc lập.

Gọi X là tuổi thọ của bóng đèn. Theo giả thiết tham số λ của phân bố mũ

$$\lambda = \frac{1}{E(X)} = \frac{1}{8} = 0,125$$

Xác suất để một bóng đèn bị cháy trong khoảng thời gian 1 năm (12 tháng) là

$$P(X < 12) = 1 - e^{-12 \cdot 0,125} \approx 0,77687$$

Vậy xác suất để cả 6 bóng đèn bị cháy trong vòng 1 năm bằng

$$(P(X < 12))^6 = 0,77687^6 = 0,219831$$

Biến cố trong vòng 1 năm không phải thay bóng đèn chính là biến cố đối lập với biến cố nói trên. Do vậy xác suất để trong vòng 1 năm không phải thay bóng đèn bằng

$$1 - (P(X < 12))^6 = 0,780169.$$

Ví dụ 2.5.15

Xét ví dụ về hiện tượng đứt sợi trong công nghiệp dệt. Sợi dệt có thể bị đứt vào một thời điểm bất kì, và người ta nhận xét rằng biến cố đó không phụ thuộc vào thời gian máy dệt làm việc liên tục không bị gián đoạn trước đó. Nói cách khác nếu kí hiệu X là khoảng thời gian từ lúc bắt đầu tới thời điểm đứt sợi lần đầu. Khi đó X là đại lượng ngẫu nhiên có phân bố mũ với tham số λ nào đó.

Giả sử máy làm việc liên tục trong thời gian dài, sợi bị đứt lập tức được nối lại và để đơn giản ta coi thời gian nối sợi là đủ nhỏ có thể bỏ qua. Gọi ν_T là số lần sợi bị đứt trong khoảng thời gian từ lúc bắt đầu tới thời điểm T. Người ta chứng minh được rằng ν_T là đại lượng ngẫu nhiên có phân bố Poisson:

$$P(\nu_T = n) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}.$$

Kết quả đó được ứng dụng trong khá nhiều bài toán thực tế. Chẳng hạn khoảng thời gian (đơn vị thời gian tính bằng tháng) để một xe máy phải vá săm là đại lượng ngẫu nhiên có phân bố mũ với tham số $\lambda = 0,2$. Suy ra thời gian trung bình để một xe máy phải một lần vá săm là

$$E(X) = \frac{1}{\lambda} = 5 \quad \text{tháng.}$$

Gọi ν là số lần vá săm xe máy trong khoảng thời gian 1 năm. Do ν có phân bố Poisson với tham số $\lambda T = 0,2 \cdot 12 = 2,4$, vì vậy xác suất để trong một năm phải thay hoặc vá săm ít nhất 2 lần bằng

$$P(\nu \geq 2) = 1 - \left(\frac{e^{-\lambda T}}{0!} + e^{-\lambda T} \frac{\lambda T}{1!} \right) = 0,69156.$$

Tương tự xác suất để trong một năm phải thay hoặc vá săm ít nhất 3 lần bằng

$$P(\nu \geq 3) = 1 - \left(\frac{e^{-\lambda T}}{0!} + e^{-\lambda T} \frac{\lambda T}{1!} + e^{-\lambda T} \frac{(\lambda T)^2}{2!} \right) = 0,43029.$$

8. *Phân bố chuẩn*

Trong lí thuyết xác suất cũng như thống kê, hàm phân bố thường gặp nhất và cũng có ý nghĩa lớn nhất là phân bố chuẩn. Nó xuất hiện ở rất nhiều lĩnh vực từ nghiên cứu đến ứng dụng thực tế: chẳng hạn các sai số trong đo đạc, các chỉ tiêu trong kinh tế, các số liệu về y tế, xã hội, giáo dục,... Định lí giới hạn trung tâm ta sẽ nghiên cứu ở chương IV giải thích lí do sự xuất hiện trong rất nhiều lĩnh vực của phân bố này.

Định nghĩa 2.5.9 *Đại lượng ngẫu nhiên được gọi là đại lượng ngẫu nhiên có phân bố chuẩn nếu hàm mật độ của nó có dạng*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

trong đó m là số thực và σ là hằng số dương. Người ta kí hiệu $N(m, \sigma^2)$ là lớp các phân bố chuẩn với hai tham số m và σ^2.

Nhận xét 2.5.3 *Để chứng minh hàm $f(x)$ định nghĩa trên thực sự là hàm mật độ, ta cần phải chứng minh tích phân của nó trên toàn bộ trục số bằng 1*

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = 1.$$

Lưu ý rằng mặc dù phân bố chuẩn là phân bố thường gặp nhất trong các lĩnh vực ứng dụng của lí thuyết xác suất hay thống kê toán sau này, hàm mật độ của phân bố chuẩn nêu trong định nghĩa trên là hàm mà nguyên hàm không có biểu diễn dưới dạng sơ cấp. Chính vì lí do đó, các sách về ứng dụng của lí thuyết xác suất hay thống kê đều phải in kèm một số trang phụ lục liệt kê một số giá trị của các hàm phân bố, hàm mật độ của phân bố chuẩn với các tham số đặc biệt nào đó ($m = 0, \sigma = 1$ chẳng hạn). Đặt $I = \int_0^{\infty} e^{-x^2}\, dx$. Trước hết ta chứng minh

$$I = \int_0^{\infty} e^{-x^2}\, dx = \frac{\sqrt{\pi}}{2}.$$

Thật vậy, xét tích phân bội sau

$$\int_0^{\infty}\int_0^{\infty} e^{-(x^2+y^2)}\, dxdy = \int_0^{\infty} e^{-x^2}\, dx \int_0^{\infty} e^{-y^2}\, dy = I^2$$

Chuyển sang hệ toạ độ cực bằng phép đổi biến $x = r\cos\varphi, y = r\sin\varphi$, ta được

$$I^2 = \int_0^{\frac{\pi}{2}}\int_0^{\infty} re^{-r^2}\, drd\varphi = \frac{\pi}{2}\int_0^{\infty} re^{-r^2}\, dr = \frac{\pi}{4}$$

Suy ra

$$I = \int_0^{\infty} e^{-x^2}\, dx = \frac{\sqrt{\pi}}{2}, \quad \text{đ.p.c.m.}$$

Trở lại tích phân của hàm $f(x)$, sử dụng phép đổi biến $t = \frac{x-m}{\sqrt{2}\sigma}$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-t^2}\, dt = \frac{2}{\sqrt{\pi}}\int_0^{\infty} e^{-t^2}\, dt = 1.$$

i) Kì vọng của phân bố chuẩn

• Giả sử $X \in N(0,1)$ là đại lượng ngẫu nhiên tương ứng với hai tham số $m = 0$ và $\sigma = 1$. Do hàm mật độ của phân bố chuẩn thuộc lớp $N(0,1)$

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

là hàm chẵn, suy ra trong biểu thức tính kì vọng hàm dưới dấu tích phân là hàm lẻ. Vậy

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 0$$

• Trường hợp $Z \in N(m, \sigma^2)$, sử dụng tính chất của hàm mật độ

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{(x-m)^2}{2\sigma^2}}\, dx = 1$$

và phép đổi biến $t = \frac{x-m}{\sigma}$

$$\begin{aligned} E(Z) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - m + m) e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-m) e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} m e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{\frac{-t^2}{2}}\, dt + m = m. \end{aligned}$$

Vậy tham số m cũng đồng thời là kì vọng của phân bố chuẩn.

ii) Phương sai của phân bố chuẩn

• Trước hết ta tính phương sai của đại lượng ngẫu nhiên có phân bố chuẩn thuộc lớp $N(0,1)$, sử dụng tích phân từng phần ta có

$$\begin{aligned} D(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\, dx \\ &= -\frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} |_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\, dx = 1. \end{aligned}$$

• Trường hợp tổng quát $Z \in N(m, \sigma^2)$, sử dụng phép đổi biến $t = \frac{x-m}{\sigma}$

$$D(Z) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}}\, dt = \sigma^2.$$

Vậy tham số σ^2 cũng đồng thời là phương sai của phân bố chuẩn. Tham số σ là độ lệch tiêu chuẩn của nó.

Áp dụng ví dụ 2.4.3 chương này ta có nhận xét quan trọng sau

Nếu X là đại lượng ngẫu nhiên có phân bố chuẩn với kì vọng bằng m và phương sai bằng σ^2, khi đó

$$Z = \frac{X-m}{\sigma} \in N(0,1)$$

Tổng quát hơn nếu $X \in N(m, \sigma)$, khi đó đại lượng ngẫu nhiên $Y = aX + b$ cũng có phân bố chuẩn

$$Y = aX + b \in N(am + b, a^2\sigma^2)$$

với kì vọng $E(Y) = am + b$ và phương sai $D(Y) = a^2\sigma^2$

Người ta thường kí hiệu hàm mật độ, hàm phân bố của $X \in N(0,1)$ là $\varphi(.)$ và $\Phi(.)$ tương ứng

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du$$

Do nguyên hàm của hàm mật độ phân bố chuẩn không biểu diễn được dưới dạng sơ cấp nên việc tính toán gần đúng các giá trị của hàm $\Phi(.)$ thông qua các bảng ở cuối sách hoặc được tính sẵn trong nhiều phần mềm tính toán (Phụ lục 1). Một số tài liệu không ghi ra các giá trị hàm phân bố $\Phi(.)$ mà chỉ đưa vào bảng các giá trị tích phân $\frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{u^2}{2}}\, du,\ \forall x > 0.$

Trong thực hành người ta hay sử dụng các công thức sau

Quy tắc $2\sigma, 3\sigma, 4\sigma$

Giả sử $X \in N(m, \sigma)$, khi đó

$$P(m - \lambda\sigma \leq X \leq m + \lambda\sigma) = P(|X - m| \leq \lambda\sigma) =$$

$$= P(|\frac{X - m}{\sigma}| \leq \lambda) = \frac{1}{\sqrt{2\pi}} \int_{-\lambda}^{\lambda} e^{-\frac{x^2}{2}}\, dx = 2\Phi(\lambda) - 1$$

Đặc biệt khi $\lambda = 2, \lambda = 3$ thậm chí $\lambda = 4$, tra bảng phân bố chuẩn

$$2\Phi(\lambda) - 1 = \begin{cases} 0,95450 & \text{nếu } \lambda = 2 \\ 0,99730 & \text{nếu } \lambda = 3 \\ 0,99994 & \text{nếu } \lambda = 4 \end{cases}$$

Như vậy tùy theo yêu cầu về độ tin cậy của vấn đề, ta có thể coi các biến cố

$$\{m - \lambda\sigma \leq X \leq m + \lambda\sigma\} \quad \text{với} \quad \lambda = 2, \lambda = 3, \lambda = 4$$

hầu như chắc chắn xảy ra. Các công thức

$$P(m - \lambda\sigma \leq X \leq m + \lambda\sigma) = 2\Phi(\lambda) - 1$$

với $\lambda = 2, \lambda = 3, \lambda = 4$ được gọi là các quy tắc $2\sigma, 3\sigma, 4\sigma$ tương ứng.

Ví dụ 2.5.16

Trong một trường đại học gồm 2000 học sinh có 50 học sinh cao trên 1 mét 75. Giả thiết chiều cao của thanh niên là đại lượng ngẫu nhiên có phân bố chuẩn với chiều cao trung bình là lm63. Hãy tính số người có chiều cao trên 1 mét 70 trong trường đó.

Giải:

Gọi X là chiều cao của thanh niên. X là đại lượng ngẫu nhiên có phân bố chuẩn với kì vọng bằng $m = 1,63$ và phương sai σ^2 chưa biết cần ước lượng.

Theo giả thiết có 50 học sinh cao trên 1 mét 75. Vậy xác suất biến cố

$$P(X > 1,75) \approx \frac{50}{2000} = 0,025$$

Mặt khác do $\frac{X-m}{\sigma}$ là đại lượng ngẫu nhiên có phân bố chuẩn thuộc lớp $N(0,1)$, nói cách khác nó có hàm phân bố là $\Phi(.)$

$$P(X > 1,75) = 1 - \Phi(\frac{1,75-m}{\sigma}) = 1 - \Phi(\frac{0,12}{\sigma})$$

Suy ra

$$\Phi(\frac{0,12}{\sigma}) = 1 - P(X > 1,75) \approx 0,975 = \Phi(1,95996)$$

hay $\sigma = \frac{0,12}{1,95996} = 0,0612257$. Vậy

$$\begin{aligned} P(X > 1,70) &= 1 - P(X < 1,70) \approx \Phi(\frac{1,70-1,63}{0,0612257}) \\ &= 1 - \Phi(1,14331) = 1 - 0,874 = 0,126. \end{aligned}$$

Do đó số người có chiều cao trên 1 mét 70 trong trường xấp xỉ bằng

$$2000 \times 0,126 = 252 \qquad \text{(người)}.$$

2.6 Các đặc trưng khác của đại lượng ngẫu nhiên

i) **Mode** Mode của đại lượng ngẫu nhiên X, kí hiệu là $mod(X)$ được xác định như sau:

* Nếu X là đại lượng ngẫu nhiên rời rạc khi đó $mod(X)$ là trị x_{i_0} của đại lượng ngẫu nhiên sao cho xác suất tương ứng $p_{i_0} = P(X = x_{i_0})$ đạt giá trị lớn nhất.

$$p_{i_0} = P(X = x_{i_0}) = \max\{P(X = x_i)\}$$

* Nếu X là đại lượng ngẫu nhiên liên tục với hàm mật độ $f(x)$ khi đó $mod(X)$ là giá trị x_0 mà tại đó hàm $f(x)$ đạt cực đại.

Ví dụ 2.6.1

X là đại lượng ngẫu nhiên có phân bố nhị thức:

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

Ta nhận xét rằng

$$P(X=k) > P(X=k-1) \quad \text{khi và chỉ khi} \quad (n+1)p > k$$

hay

$$P(X=0) < P(X=1) < ... < P(X=[(n+1)p]) =$$
$$= P(X=k_0) \geq P(X=k_0+1) > ... > P(X=n)$$

suy ra $mod(X) = [(n+1)p]$ (phần nguyên của số thực $(n+1)p$).

Ví dụ 2.6.2

Nếu X là đại lượng ngẫu nhiên có phân bố chuẩn với hàm mật độ

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Hiển nhiên $f(x)$ đạt cực đại tại $x = m$, suy ra $mod(X) = m$.

ii) **Trung vị** *Trung vị* (hay còn gọi là Median) của đại lượng ngẫu nhiên X, là giá trị m sao cho

$$P(X < m) = P(X > m)$$

hoặc tương đương với nó

$$F(m) \leq \frac{1}{2}, F(m+0) \geq \frac{1}{2}$$

trong đó $F(x)$ là hàm phân bố của X.

Nếu X là đại lượng ngẫu nhiên liên tục có hàm phân bố tăng nghiêm ngặt, khi đó trung vị của X là giá trị m sao cho

$$F(m) = \frac{1}{2}$$

Về mặt hình học nếu m là trung vị của X khi đó đường thẳng $x = m$ chia diện tích hình thang cong giới hạn bởi hàm mật độ và trục hoành thành hai phần diện tích bằng nhau (và $= \frac{1}{2}$).

Định nghĩa 2.6.1 *Người ta nói X là đại lượng ngẫu nhiên có phân bố đối xứng qua giá trị 0 nếu X và $-X$ có phân bố trùng nhau.*

Phân bố của đại lượng ngẫu nhiên Y được gọi là phân bố đối xứng qua giá trị m nếu $Y - m$ có phân bố đối xứng qua 0.

Giả sử $F(x)$ là hàm phân bố của X. Dễ dàng suy ra điều kiện cần và đủ để X có phân bố đối xứng qua giá trị m là

$$F(m - x) = 1 - F(m + x + 0) \quad \text{với mọi } x$$

Nhận xét rằng nếu X là đại lượng ngẫu nhiên liên tục có phân bố đối xứng qua giá trị 0, $f(x)$ là hàm mật độ, khi đó $f(x)$ là hàm chẵn. Ví dụ đại lượng ngẫu nhiên có phân bố chuẩn thuộc lớp $N(0, 1)$ có phân bố đối xứng qua 0.

Trường hợp X có phân bố đối xứng qua giá trị m, khi đó m cũng là median (trung vị) và kì vọng $E(X)$ của X (nếu chúng tồn tại) bằng m.

iii) **Phân vị** Tổng quát hơn khái niệm trung vị là phân vị

Định nghĩa 2.6.2 *Phân vị mức p của đại lượng ngẫu nhiên X là số x_p sao cho*

$$P(X < x_p) \leq p, P(X \leq x_p) \geq p$$

Nếu X là đại lượng ngẫu nhiên liên tục có hàm phân bố tăng nghiêm ngặt, khi đó tồn tại duy nhất một giá trị x_p sao cho

$$P(X < x_p) = p \quad \text{hay} \quad F(x_p) = p$$

Giá trị x_p đó là phân vị mức p của đại lượng ngẫu nhiên X.

Nhận xét rằng phân vị mức $p = \frac{1}{2}$ chính là *median* của X.

iv) **Mô men** Giả sử X là đại lượng ngẫu nhiên tồn tại $E(X)$. Khi đó

$$m_k = E(X^k)$$

được gọi là *mô men cấp k* và

$$\alpha_k = E((X - E(X))^k)$$

được gọi là *mô men quy tâm cấp k* của đại lượng ngẫu nhiên X.

Ví dụ 2.6.3

Giả sử X là đại lượng ngẫu nhiên có phân bố chuẩn $X \in N(m, \sigma^2)$. Hãy tính các mô men quy tâm cấp 2, cấp 3, cấp 4 của X.

Mô men quy tâm cấp 2 của X

$$\alpha_2 = E((X - E(X))^2) = E((X - m)^2) = D(X)$$

chính là phương sai của đại lượng ngẫu nhiên X.

Mô men quy tâm cấp 3 của X:

Hiển nhiên đại lượng ngẫu nhiên

$$Y = \frac{X - m}{\sigma}$$

có phân bố chuẩn thuộc lớp $N(0, 1)$ và

$$\alpha_3 = E((X - m)^3) = \sigma^3 E((\frac{X - m}{\sigma})^3) = \sigma^3 E(Y^3)$$

Mô men quy tâm cấp 3 của Y bằng:

$$E(Y^3) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^3 e^{-\frac{t^2}{2}}\, dt = 0$$

do hàm dưới dấu tích phân là hàm lẻ. Vậy $\alpha_3 = 0$.

Tương tự *mô men quy tâm cấp 4* của X

$$\alpha_4 = E((X - m)^4) = \sigma^4 E(Y^4).$$

Bây giờ ta tính mô men quy tâm cấp 4 của Y, bằng phương pháp tích phân từng phần:

$$\begin{aligned} E(Y^4) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^4 e^{-\frac{t^2}{2}}\, dt = \frac{-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^3 de^{-\frac{t^2}{2}} \\ &= 3\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}}\, dt = 3E(Y^2) = 3. \end{aligned}$$

Vậy mô men quy tâm cấp 4 của X, $\alpha_4 = 3\sigma^4$.

BÀI TẬP CHƯƠNG II

1. Hãy xác định trong số các hàm liệt kê dưới đây, hàm nào là hàm phân bố xác suất

   (a) $F(x) = \frac{3}{4} + \frac{1}{2\pi}\arctan x$

   (b) $F(x) = \begin{cases} \dfrac{x}{1+x} & \text{nếu } x > 0 \\ 0 & \text{nếu } x \leq 0 \end{cases}$

   (c) $F(x) = \begin{cases} 1 - \dfrac{1-e^{-x}}{x} & \text{nếu } x \geq 0 \\ 0 & \text{nếu } x < 0 \end{cases}$

2. Chứng minh rằng nếu $F(x)$ là hàm phân bố xác suất và $\alpha > 0$, khi đó $[F(x)]^\alpha$ cũng là hàm phân bố.

3. Hãy xác định trong số các hàm liệt kê dưới đây, hàm nào là hàm mật độ xác suất

   (a) $f(x) = \begin{cases} \frac{1}{3} & \text{nếu } 0 < x < 1 \\ 0 & \text{nếu } x \notin (0;1) \end{cases}$

   (b) $f(x) = \begin{cases} \frac{\sin x}{2} & \text{nếu } 0 < x < 1 \\ 0 & \text{nếu } x \notin (0;1) \end{cases}$

   (c) $f(x) = \begin{cases} \frac{1}{x^2} & \text{nếu } x > 1 \\ 0 & \text{nếu } x \leq 1 \end{cases}$

   (d) $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ với mọi $-\infty < x < \infty$

   (e) $f(x) = \frac{1}{2}e^{-|x|}$ với mọi $-\infty < x < \infty$

   (f) $f(x) = \begin{cases} \frac{x}{x+1} & \text{nếu } x > 0 \\ 0 & \text{nếu } x \leq 0 \end{cases}$

   (g) $f(x) = \begin{cases} 4x^3e^{-x^4} & \text{nếu } x > 0 \\ 0 & \text{nếu } x \leq 0 \end{cases}$

   (h) $f(x) = \begin{cases} 2e^{-x}(1-e^{-x}) & \text{nếu } x > 0 \\ 0 & \text{nếu } x \leq 0 \end{cases}$

4. Với các giá trị nào của a, b, c hàm sau là hàm mật độ

$$f(x) = \frac{a}{1 + b(x - c)^2} \quad \text{với mọi} \quad -\infty < x < \infty$$

5. Chọn ngẫu nhiên một điểm trên đoạn thẳng có độ dài đơn vị (bằng 1). Điểm đó chia đoạn thẳng đã cho thành hai đoạn nhỏ. Hãy tìm hàm phân bố độ dài của đoạn bé hơn trong số 2 đoạn thẳng tạo thành.

6. Chọn ngẫu nhiên một điểm trên đoạn (0, 1) của trục hoành Ox trong hệ trục toạ độ đề các xOy. Gọi ξ là khoảng cách từ điểm chọn ngẫu nhiên tới điểm (0;1) trên mặt phẳng toạ độ. Tìm hàm mật độ của ξ.

7. Giả sử ξ là đại lượng ngẫu nhiên tồn tại kì vọng và nhận các giá trị nguyên, không âm. Chứng minh rằng

$$E(\xi) = \sum_{i=1}^{\infty} P(\xi \geq i).$$

8. Gieo đồng xu liên tục cho đến khi có 2 lần liên tiếp cùng xuất hiện mặt sấp hoặc mặt ngửa. Hãy tính trung bình số lần gieo.

9. Một điểm M lang thang trên trục số. Nó bước sang trái hoặc sang phải một đơn vị với xác suất bằng nhau và bằng $\frac{1}{2}$. Giả sử ban đầu M ở gốc toạ độ. Hãy tính phương sai của M sau 4 bước.

10. Chọn ngẫu nhiên ba điểm trên đoạn (0, 1) của trục số. Hãy tìm hàm phân bố của điểm nằm giữa.

11. Một người hút thuốc lá, mỗi ngày hút 10 hoặc 11 điếu thuốc với xác suất p hoặc $q, (q = 1 - p)$ tương ứng. Hãy tính trung bình người đó hút bao nhiêu điếu thuốc trong một tháng (30 ngày). Tính phương sai của số điếu thuốc hút trong một tháng của người đó.

12. Chọn ngẫu nhiên một điểm trong hình vuông đơn vị. Gọi X là khoảng cách từ đó tới cạnh gần nhất. Hãy tìm hàm phân bố của X.

13. Chọn ngẫu nhiên một điểm trong hình vuông đơn vị. Gọi ξ là khoảng cách từ đó tới đỉnh hình vuông gần nhất. Hãy tìm hàm mật độ của ξ.

14. Hãy tìm số hạng lớn nhất của phân bố nhị thức.

15. Giả sử chúng ta đang chờ trước trạm điện thoại công cộng. Bên trong trạm một người đang nói chuyện. Giả thiết rằng thời gian nói chuyện của người đó là đại lượng ngẫu nhiên có hàm mật độ (tính theo phút)

$$\frac{1}{3}e^{-\frac{x}{3}}, x > 0$$

    (a) Hãy tìm xác suất để người đó gọi nhiều hơn 3 phút.

    (b) Tìm xác suất để người đó nói chuyện tiếp hơn 3 phút nữa, với điều kiện thời gian nói chuyện của người đó đã hơn 3 phút rồi.

16. Hãy tính $E(X^2 - 1)$, trong đó X là đại lượng ngẫu nhiên có phân bố hình học.

17. Hãy tính $E\left(\frac{1}{1+X}\right)$, trong đó X là đại lượng ngẫu nhiên có phân bố nhị thức.

18. Một người chơi một trò chơi như sau: gieo đồng thời 2 xúc xắc, nếu tổng bằng 7 hoặc 11, người đó thắng, nếu tổng bằng $2, 3$ hoặc 12, người đó thua cuộc. Các trường hợp còn lại, người chơi lặp lại trò chơi cho đến khi:

    (a) Tổng số chấm xuất hiện ở hai xúc xắc bằng 7. Trong trường hợp này người đó thắng.

    (b) Tổng số chấm xuất hiện ở hai xúc xắc bằng tổng số chấm ở lần gieo đầu tiên. Trong trường hợp này người đó thua.

    Tìm xác suất để người chơi thắng cuộc và hãy tính trung bình người đó gieo xúc xắc bao nhiêu lần trong trò chơi đó.

19. Trong mặt phẳng xOy, một đường thẳng d đi qua $A(-3, 3)$ quay tròn xung quanh A và dừng lại một cách ngẫu nhiên (khi đó ta coi góc φ của d phân bố đều trong khoảng $[0, \pi]$). Kí hiệu đại lượng ngẫu nhiên ξ là khoảng cách từ gốc toạ độ $O(0, 0)$ tới đường thẳng d. Tìm hàm mật độ của ξ.

ĐÁP SỐ VÀ HƯỚNG DẪN

1.

(a) Không là hàm phân bố.

(b) Là hàm phân bố.

(c) Là hàm phân bố.

2. Sử dụng tính chất hàm phân bố.

3.

(a) Không là hàm mật độ.

(b) Không là hàm mật độ.

(c) Là hàm mật độ.

(d) Là hàm mật độ.

(e) Là hàm mật độ.

(f) Không là hàm mật độ.

(g) Là hàm mật độ.

(h) Là hàm mật độ.

4. c tùy ý, $b > 0$ và $a = \dfrac{\sqrt{b}}{\pi}$.

5. $F(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ 2x & \text{nếu } 0 < x \leq \frac{1}{2} \\ 1 & \text{nếu } x \geq \frac{1}{2} \end{cases}$

6. $f(x) = \begin{cases} \frac{x}{\sqrt{x^2-1}} & \text{nếu } 1 < x < \sqrt{2} \\ 0 & \text{nếu } x \leq 1 \text{ hoặc } x \geq \sqrt{2} \end{cases}$

7. $E(\xi) = \sum_{i=1}^{\infty} iP(\xi = i) = \sum_{i=1}^{\infty} P(\xi \geq i)$.

8. Số lần gieo trung bình bằng 3.

9. Phương sai của M bằng 2.

10. $F(x)$ là xác suất của biến cố "có đúng hai điểm nhỏ hơn x và 1 điểm lớn hơn x hoặc cả 3 điểm nhỏ hơn x":

$$F(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ 3x^2 - 2x^3 & \text{nếu } 0 < x \leq 1 \\ 1 & \text{nếu } x > 1 \end{cases}$$

11. Trong một tháng trung bình người đó hút

$$330 - 30p \quad \text{điếu thuốc với phương sai} \quad 30pq.$$

12. $$F(x) = \begin{cases} 0 & \text{nếu } x \leq 0 \\ 1 - (1-2x)^2 & \text{nếu } 0 < x \leq \frac{1}{2} \\ 1 & \text{nếu } x \geq \frac{1}{2} \end{cases}$$

13. $$f(x) = \begin{cases} 2\pi x & \text{nếu } 0 < x \leq \frac{1}{2} \\ 2\pi x - 8x \arccos \frac{1}{2x} & \text{nếu } \frac{1}{2} < x \leq \frac{\sqrt{2}}{2} \\ 0 & x \leq 0 \text{ hoặc } x \geq \frac{\sqrt{2}}{2} \end{cases}$$

14. Số hạng thứ k_0 trong đó $k_0 = [(n+1)p]$. Nếu $(n+1)p$ là số nguyên, khi đó

$$C_n^{k_0} p^{k_0} q^{n-k_0} = C_n^{k_0-1} p^{k_0-1} q^{n-k_0+1}.$$

15. Gọi X là thời gian nói chuyện của người đó. Theo giả thiết X là đại lượng ngẫu nhiên có phân bố mũ

(a) $P(X > 3) = e^{-1}$.

(b) $P(X > 6/X > 3) = \frac{e^{-6/3}}{e^{-3/3}} = e^{-1}$.

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

17. $E\left(\frac{1}{1+X}\right) = \frac{1}{(n+1)p}(1 - q^{n+1})$.

18. Xác suất để người chơi thắng cuộc $\frac{34}{55}$

Số lần gieo trung bình $\frac{557}{165} \approx 3,375$.

19. Phương trình đường thẳng đi qua $A(-3,3)$:

$$y = (x+3)\tan\varphi + 3$$

Khoảng cách từ gốc toạ độ $O(0,0)$ tới đường thẳng:

$$\xi = 3|\sin\varphi + \cos\varphi|.$$

Vậy hàm phân bố của ξ:

$$F(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ \frac{2}{\pi}\cdot \arcsin\frac{x}{3\sqrt{2}} & \text{nếu } 0 < x < 3\sqrt{2} \\ 1 & \text{nếu } x > 3\sqrt{2}. \end{cases}$$

Hàm mật độ của ξ là

$$f(x) = \begin{cases} \frac{2}{\pi\sqrt{18-x^2}} & \text{nếu } 0 < x < 3\sqrt{2} \\ 0 & \text{nếu } x \notin (0, 3\sqrt{2}) \end{cases}$$

Ta có thể giải cách khác: Nếu ta gọi α là góc giữa đường thẳng d đi qua $A(-3,3)$ và đường thẳng OA, khi đó α phân bố đều trên $(0, \frac{\pi}{2})$ và khoảng cách từ gốc toạ độ $O(0,0)$ tới đường thẳng d:

$$\xi = 3\sqrt{2}\sin\alpha.$$

Từ đây suy ra hàm phân bố, mật độ của ξ.

Chương 3

Đại lượng ngẫu nhiên hai chiều

3.1 Phân bố của hai đại lượng ngẫu nhiên

Khi khảo sát nhiều đại lượng ngẫu nhiên, ngoài các hàm phân bố của từng đại lượng ngẫu nhiên chúng ta cũng cần biết thêm các thông tin về mối quan hệ giữa các đại lượng ngẫu nhiên đó.

Ví dụ ξ và η là 2 đại lượng ngẫu nhiên với các bảng phân bố của chúng như sau:

ξ	1	2
P	$\frac{1}{2}$	$\frac{1}{2}$

η	0	1	2
P	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Các bảng phân bố này chưa cho chúng ta thông tin gì về các xác suất $P(\xi = i, \eta = j)$. Chúng ta kí hiệu $p_{ij} = P(\xi = i, \eta = j)$ là xác suất của biến cố tích $\{\xi = i\} \cdot \{\eta = j\}$:

$$P(\xi = 1, \eta = 0) = p_{10}, P(\xi = 1, \eta = 1) = p_{11}, P(\xi = 1, \eta = 2) = p_{12}$$

$$P(\xi = 2, \eta = 0) = p_{20}, P(\xi = 2, \eta = 1) = p_{21}, P(\xi = 2, \eta = 2) = p_{22}.$$

Như vậy để có đầy đủ thông tin hơn, chúng ta cần biểu diễn các xác suất $p_{ij} = P(\xi = i, \eta = j),\ \forall i, j$ dưới dạng bảng phân bố xác suất như sau

η \ ξ	1	2
0	p_{10}	p_{20}
1	p_{11}	p_{21}
2	p_{12}	p_{22}

Chẳng hạn giả sử các xác suất p_{ij} được cho trong bảng sau:

η \ ξ	1	2
0	$\frac{1}{6}$	$\frac{1}{6}$
1	$\frac{1}{12}$	$\frac{1}{4}$
2	$\frac{1}{4}$	$\frac{1}{12}$

Từ bảng phân bố này chúng ta có thể lập được các bảng phân bố xác suất của ξ và η. Ví dụ bảng phân bố của ξ được tính như sau:

$$\begin{aligned} P(\xi = 1) &= P(\xi = 1, \eta = 0) + P(\xi = 1, \eta = 1) + P(\xi = 1, \eta = 2) \\ &= \frac{1}{6} + \frac{1}{12} + \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

$$\begin{aligned} P(\xi = 2) &= P(\xi = 2, \eta = 0) + P(\xi = 2, \eta = 1) + P(\xi = 2, \eta = 2) \\ &= \frac{1}{6} + \frac{1}{4} + \frac{1}{12} = \frac{1}{2}. \end{aligned}$$

Ta gọi một cặp hai đại lượng ngẫu nhiên (ξ, η) là một véc tơ ngẫu nhiên hai chiều. Nếu chúng ta đồng thời khảo sát hai đại lượng ngẫu nhiên ξ và η, chúng ta sẽ coi chúng như các toạ độ của một véc tơ ngẫu nhiên (hay một điểm ngẫu nhiên) (ξ, η). Các giá trị có thể có của nó là các điểm (x, y) trong mặt phẳng toạ độ xOy. Gọi tập E là một miền phẳng bất kì $E \subset \mathbb{R}^2$ và $P_{\xi,\eta}(E) = P((\xi, \eta) \in E)$ là xác suất để điểm ngẫu nhiên (ξ, η) rơi vào tập E. *Người ta gọi $P_{\xi,\eta}(E)$, với mọi $E \subset \mathbb{R}^2$ là độ đo xác suất của các tập hợp trên mặt phẳng sinh bởi véc tơ ngẫu nhiên (ξ, η).*

Giả sử E_1 là tích Đề các của hai tập hợp $(-\infty, x)$ và $(-\infty, y)$:

$$E_1 = (-\infty, x) \times (-\infty, y) \subset \mathbb{R}^2,$$

khi đó biến cố $\{(\xi,\eta)\in E_1\}$ là biến cố tích

$$\{(\xi,\eta)\in E_1\}=\{\xi\in(-\infty,x)\}\cdot\{\eta\in(-\infty,y)\}.$$

Người ta thường kí hiệu xác suất $P((\xi,\eta)\in E_1)$ dưới dạng

$$P((\xi,\eta)\in E_1)=P(\xi<x,\eta<y).$$

Ta dẫn vào định nghĩa sau

Định nghĩa 3.1.1 *Hàm*

$$H(x,y)=P(\xi<x,\eta<y)=P(\{\xi\in(-\infty,x)\}\cdot\{\eta\in(-\infty,y)\})$$

với mọi $x,y\in\mathbb{R}$ là hàm phân bố chung của hai đại lượng ngẫu nhiên ξ và η (hay còn gọi là hàm phân bố đồng thời của véc tơ ngẫu nhiên (ξ,η)).

Ta có nhận xét sau:

Nhận xét 3.1.1 *i) Nếu tập E_2 là hình chữ nhật*

$$E_2=[a,b)\times[c,d)=\{(x,y)/a\le x<b,c\le y<d\}$$

khi đó xác suất $P((\xi,\eta)\in E_2)$ có thể tính theo hàm phân bố $H(x,y)$:

$$\begin{aligned}P((\xi,\eta)\in E_2)&=P((\xi,\eta)\in[a,b)\times[c,d))\\&=H(b,d)-H(a,d)-H(b,c)+H(a,c).\end{aligned}$$

ii) Nếu ξ và η là 2 đại lượng ngẫu nhiên rời rạc nhận các giá trị x_i,y_j tương ứng. Đặt $p_{ij}=P(\xi=x_i,\eta=y_j)\quad\forall i,j$ khi đó các xác suất p_{ij} xác định phân bố của (ξ,η) trên mặt phẳng và

$$P((\xi,\eta)\in E)=\sum_{i,j:(x_i,y_j)\in E}p_{ij}.$$

Định nghĩa 3.1.2 *Nếu tồn tại một hàm không âm $h(x,y)\ge 0,\ \forall x,y\in\mathbb{R}$ sao cho*

$$P((\xi,\eta)\in E)=\iint_E h(x,y)\,dxdy$$

với mọi miền E của mặt phẳng. Khi đó ta nói $h(x,y)$ là hàm mật độ của véc tơ ngẫu nhiên (ξ,η). (Hay còn gọi $h(x,y)$ là hàm mật độ chung của ξ và η).

Định nghĩa 3.1.3 *Trường hợp đặc biệt, nếu hàm mật độ của véc tơ ngẫu nhiên (ξ, η) có dạng:*

$$h(x,y) = \begin{cases} \frac{1}{S(D)} & \text{nếu } (x,y) \in D \\ 0 & \text{nếu } (x,y) \notin D \end{cases}$$

trong đó $S(D)$ là diện tích của miền D. Khi đó ta nói véc tơ ngẫu nhiên (ξ, η) phân bố đều trên miền D và hiển nhiên

$$P((\xi,\eta) \in E) = \iint_E h(x,y)\,dxdy = \frac{S(E \cap D)}{S(D)}.$$

Như vậy nếu E là tập con của D và điểm ngẫu nhiên (ξ, η) có phân bố đều trên miền D, khi đó xác suất để điểm ngẫu nhiên thuộc E tỉ lệ với diện tích miền E:

$$P((\xi,\eta) \in E) = \frac{S(E)}{S(D)}.$$

Đây cũng chính là phương pháp hình học để tính xác suất của các biến cố ngẫu nhiên đã giới thiệu ở chương I. Phương pháp hình học thực chất là phương pháp để tính xác suất của các biến cố sinh bởi phân bố đều.

Nhận xét 3.1.2 *i) Hàm phân bố của véc tơ ngẫu nhiên (ξ, η) có thể tính theo hàm mật độ*

$$H(x,y) = P(\xi < x, \eta < y) = \int_{-\infty}^{x}\int_{-\infty}^{y} h(u,v)\,dudv$$

suy ra

$$h(x,y) = \frac{\partial^2 H(x,y)}{\partial x \partial y}.$$

ii) Từ nhận xét trên suy ra

$$P((\xi,\eta) \in [a,b) \times [c,d)) = H(b,d) - H(a,d) - H(b,c) + H(a,c) =$$

$$= \int_a^b \int_c^d h(u,v)\,dudv.$$

Thật vậy

$$\int_a^b \int_c^d h(x,y)\,dxdy = \int_a^b \int_c^d \frac{\partial^2 H(x,y)}{\partial x \partial y} dxdy =$$

$$= \int_a^b \frac{\partial}{\partial x}(H(x,d) - H(x,c))dx = H(b,d) - H(a,d) - H(b,c) + H(a,c).$$

3.2 Tính chất Hàm Phân bố và Hàm Mật độ của đại lượng ngẫu nhiên hai chiều

Từ định nghĩa hàm phân bố

$$H(x,y) = P(\{\xi \in (-\infty, x)\} \cdot \{\eta \in (-\infty, y)\}) = P(\xi < x, \eta < y)$$

Suy ra $H(x,+\infty) = P(\xi < x)$ là hàm phân bố $F(x)$ của ξ và tương tự $H(+\infty, y) = P(\eta < y)$ là hàm phân bố $G(y)$ của η.

Chứng minh tương tự như phân bố của đại lượng ngẫu nhiên một chiều, hàm phân bố $H(x,y)$ và hàm mật độ $h(x,y)$ có các tính chất sau:

a) $H(x,y)$ là hàm đơn điệu tăng theo từng biến x và biến y.

b) $H(x,-\infty) = H(-\infty, y) = 0$.

c) $H(+\infty,+\infty) = 1$.

d) Hàm mật độ $h(x,y)$ không âm $h(x,y) \geq 0 \quad \forall x, y$.

Ngoài ra từ các nhận xét ở mục 1. và từ định nghĩa hàm mật độ

e) $H(x+\Delta x, y+\Delta y) - H(x+\Delta x, y) - H(x, y+\Delta y) + H(x,y) \geq 0$

với mọi $\Delta x \geq 0, \Delta y \geq 0$.

f)

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y)\,dxdy = 1.$$

Chú ý rằng nếu các hàm $H(x,y)$ (hoặc $h(x,y)$) bất kì thoả mãn các tính chất nêu trên tương ứng với chúng, khi đó chúng là các hàm phân bố (hoặc hàm mật độ) của một véc tơ ngẫu nhiên hai chiều nào đó.

Do mối quan hệ giữa hàm phân bố và hàm mật độ đồng thời

$$H(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} h(u,v)\,dudv$$

suy ra

$$F(x) = H(x, \infty) = \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} h(u, v)\, dv \right) du.$$

Tương tự

$$G(y) = H(\infty, y) = \int_{-\infty}^{y} \left(\int_{-\infty}^{\infty} h(u, v)\, du \right) dv.$$

Vậy

$$F'(x) = f(x) = \int_{-\infty}^{\infty} h(x, v)\, dv$$

là hàm mật độ của ξ và hàm mật độ của η bằng

$$G'(y) = g(y) = \int_{-\infty}^{\infty} h(u, y)\, du.$$

Để thuận tiện về mặt kí hiệu, từ đây về sau ta thường biểu diễn các hàm mật độ $f(x)$ và $g(y)$ dưới dạng:

$$f(x) = \int_{-\infty}^{\infty} h(x, y)\, dy$$

và

$$g(y) = \int_{-\infty}^{\infty} h(x, y)\, dx.$$

Ví dụ 3.2.1

Giả sử hàm mật độ chung của X và Y là

$$h(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & \text{nếu } 0 < x < 1, 0 < y < 1 \\ 0 & \text{trong trường hợp ngược lại.} \end{cases}$$

khi đó áp dụng công thức $f(x) = \int_{-\infty}^{\infty} h(x, y)\, dy$

$$f(x) = \begin{cases} \frac{6}{5} \int_0^1 (x + y^2)\, dy = \frac{6}{5}(x + \frac{1}{3}) & \text{nếu } 0 < x < 1 \\ 0 & \text{nếu } x \notin (0, 1) \end{cases}$$

là hàm mật độ của X. Tương tự

$$g(y) = \begin{cases} \frac{6}{5} \int_0^1 (x + y^2)\, dx = \frac{6}{5}(\frac{1}{2} + y^2) & \text{nếu } 0 < y < 1 \\ 0 & \text{nếu } y \notin (0, 1) \end{cases}$$

là hàm mật độ của Y.

Ví dụ 3.2.2

Giả sử véc tơ ngẫu nhiên (X, Y) có phân bố đều trên miền hình chữ nhật $[a, b] \times [c, d]$. Hãy tìm hàm mật độ, kì vọng và phương sai của X.

Giải: Hàm mật độ chung của X và Y là

$$h(x,y) = \begin{cases} \frac{1}{(b-a)(d-c)} & \text{nếu } a < x < b, c < y < d \\ 0 & \text{nếu } x \notin (a,b) \text{ hoặc } y \notin (c,d). \end{cases}$$

Vậy hàm mật độ của X bằng

$$f(x) = \int_c^d \frac{1}{(b-a)(d-c)}\, dy = \begin{cases} \frac{1}{b-a} & \text{nếu } a < x < b \\ 0 & \text{nếu } x \notin (a,b) \end{cases}$$

hay X phân bố đều trên (a, b). Kì vọng và phương sai của phân bố đều đã được tính ở chương II:

$$E(X) = \frac{a+b}{2},\ D(X) = \frac{(a-b)^2}{12}.$$

Tương tự đại lượng ngẫu nhiên Y có phân bố đều trên (c, d).

Ví dụ 3.2.3

Giả sử véc tơ ngẫu nhiên (X, Y) có phân bố đều trên hình tròn tâm $I(0, R)$ bán kính R. Hãy tìm hàm mật độ, kì vọng và phương sai của đại lượng ngẫu nhiên Y.

Giải: Hàm mật độ chung của X và Y là

$$h(x,y) = \begin{cases} \frac{1}{\pi R^2} & \text{nếu } x^2 + (y-R)^2 < R^2 \\ 0 & \text{trong trường hợp ngược lại.} \end{cases}$$

Vậy hàm mật độ của Y bằng

$$g(y) = \int_{-\infty}^{\infty} h(x,y)\, dx =$$

$$= \begin{cases} \int_{-\sqrt{R^2-(y-R)^2}}^{+\sqrt{R^2-(y-R)^2}} \frac{1}{\pi R^2}\, dx = \frac{2\sqrt{R^2-(y-R)^2}}{\pi R^2} & \text{nếu} \quad 0 < y < 2R \\ 0 & \text{nếu} \quad y \notin (0, 2R). \end{cases}$$

Khác với ví dụ trước, nếu (X, Y) phân bố đều trên miền hình chữ nhật, khi đó Y có phân bố đều. Ở ví dụ này (X, Y) phân bố đều trên hình tròn và phân bố của Y như đã tính trên không là phân bố đều. Kì vọng của Y

$$E(Y) = \int_{-\infty}^{\infty} yg(y)\,dy = \int_{0}^{2R} \frac{2y\sqrt{R^2-(y-R)^2}}{\pi R^2}\,dy = R.$$

Phương sai của Y

$$\begin{aligned} D(Y) = E(Y^2) - (EY)^2 &= \int_{0}^{2R} \frac{2y^2\sqrt{R^2-(y-R)^2}}{\pi R^2}\,dy - R^2 \\ &= \frac{5R^2}{4} - R^2 = \frac{R^2}{4}. \end{aligned}$$

Tương tự như định lí 2.5.3 chương II, để tính kì vọng của hàm của hai đại lượng ngẫu nhiên X và Y, người ta sử dụng định lí sau

Định lí 3.2.1 *Giả sử $\varphi(x, y)$ là một hàm hai biến. Gọi $h(x, y)$ làm mật độ chung của X và Y. Khi đó*

$$E\left(\varphi(X, Y)\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x, y)h(x, y)\,dxdy.$$

Ta công nhận không chứng minh định lí này. Nhận xét rằng, nhiều kết quả quan trọng được chứng minh nhờ định lí trên.

Xét định lí 2.5.1 chương II trong trường hợp X, Y là các đại lượng ngẫu nhiên liên tục có $f(x), g(y)$ là các hàm mật độ tương ứng. Gọi $h(x, y)$ là hàm mật độ chung của chúng. Ta sẽ chứng minh

$$E(X + Y) = EX + EY.$$

Thật vậy, áp dụng định lí trên cho tổng của hai đại lượng ngẫu nhiên X và Y, hàm $\varphi(.,.)$ có dạng

$$\varphi(x, y) = x + y.$$

$$\begin{aligned}E(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)h(x,y)\,dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xh(x,y)\,dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yh(x,y)\,dxdy \\ &= \int_{-\infty}^{\infty}x\int_{-\infty}^{\infty}h(x,y)\,dydx + \int_{-\infty}^{\infty}y\int_{-\infty}^{\infty}h(x,y)\,dxdy \\ &= \int_{-\infty}^{\infty}xf(x)\,dx + \int_{-\infty}^{\infty}yg(y)\,dy = EX + EY, \text{ đ.p.c.m.}\end{aligned}$$

3.3 Phân bố có điều kiện

Giả sử A là biến cố có xác suất $P(A) > 0$ và X là đại lượng ngẫu nhiên tùy ý. Ta có định nghĩa sau

Định nghĩa 3.3.1 *Người ta gọi hàm*

$$F(x/A) = P(X < x/A) \quad \textit{với} \quad \forall x \in \mathbb{R}$$

là hàm phân bố có điều kiện của X với điều kiện biến cố A xảy ra. Nếu $F(x/A)$ khả vi, kí hiệu $f(x/A) = F'(x/A)$ và

$$F(x/A) = P(X < x/A) = \int_{-\infty}^{x} f(t/A)\,dt \quad \textit{với} \quad \forall x$$

khi đó $f(x/A)$ được gọi là hàm mật độ có điều kiện của X với điều kiện biến cố A xảy ra (hoặc nói tắt là hàm mật độ của X với điều kiện A).

Ta có nhận xét rằng nếu $A_i, i = 1, 2, ...$ là một hệ đầy đủ các biến cố. Khi đó theo công thức xác suất đầy đủ, hàm phân bố của X có thể biểu diễn theo các hàm phân bố có điều kiện:

$$F(x) = P(X < x) = \sum_i P(X < x/A_i)P(A_i) = \sum_i F(x/A_i)P(A_i)$$

Đạo hàm cả hai vế theo x, ta cũng có kết quả tương tự cho hàm mật độ có điều kiện

$$f(x) = \sum_i f(x/A_i)P(A_i).$$

Chú ý: Khi X là đại lượng ngẫu nhiên rời rạc, thay vì xét các hàm phân bố có điều kiện $F(x/A)$, trong các ví dụ minh hoạ người ta thường xét xác suất có điều kiện của các biến cố $\{X = x_i\}$ với điều kiện A:

$$p_i = P(X = x_i/A) = \frac{P(A \cdot \{X = x_i\})}{P(A)} \quad \text{với } \forall i.$$

Ví dụ 3.3.1

Quay lại ví dụ 2.2.1 chương II, một xạ thủ bắn bia, xác suất bắn trúng bia của xạ thủ đó bằng p. Xạ thủ bắn liên tục vào bia cho đến khi lần đầu tiên bắn trúng bia. Gọi X là số đạn chi phí (số đạn xạ thủ đã bắn) và A là biến cố "trong 10 lần bắn đầu tiên, có đúng 1 lần bắn trúng bia". Tìm phân bố của X với điều kiện A xảy ra.

Giải: Hiển nhiên xác suất có điều kiện

$$P(X = n/A) = 0 \quad \text{nếu} \quad n > 10.$$

Xét trường hợp $n \leq 10$

$$P(X = n/A) = \frac{P(\{X = n\} \cdot A)}{P(A)}$$

Theo công thức Bernoulli $P(A) = C_{10}^1 pq^9$. Biến cố $A \cdot \{X = n\}$ là biến cố trong 10 lần bắn, duy nhất lần bắn thứ n xạ thủ đó bắn trúng bia. Xác suất

$$P(A \cdot \{X = n\}) = q^{n-1}pq^{10-n} = pq^9.$$

Vậy

$$P(X = n/A) = \frac{P(A \cdot \{X = n\})}{P(A)} = \frac{pq^9}{C_{10}^1 pq^9} = \frac{1}{10}, \forall n = 1, ..., 10$$

Do đó ta có thể nói phân bố có điều kiện của X với điều kiện A xảy ra, là *phân bố đều.*

$$P(X = 1/A) = P(X = 2/A) = ... = P(X = 10/A) = \frac{1}{10}.$$

Ví dụ 3.3.2

Số hạt giống hỏng trong 1 bao hạt giống là đại lượng ngẫu nhiên có phân bố Poisson với tham số λ. Trước khi gieo trồng người ta kiểm tra lại để loại các hạt giống bị hỏng ra khỏi bao hạt giống đó. Giả sử xác suất để người kiểm tra phát hiện có đúng một hạt giống hỏng là p. Gọi X là số hạt giống hỏng còn lại trong bao sau khi đã được kiểm tra. Tìm phân bố của X và tính trung bình số hạt giống hỏng còn lại sau khi đã được kiểm tra trước khi gieo trồng.

Giải: Gọi Y là số hạt giống hỏng có trong bao trước khi kiểm tra. Theo giả thiết Y có phân bố Poisson với tham số λ.

$$P(Y=i) = e^{-\lambda}\frac{\lambda^i}{i!}$$

Hiển nhiên $\{Y=i\}, \quad i=0,1,2,...$ là một hệ đầy đủ các biến cố. Theo công thức xác suất đầy đủ

$$P(X=n) = \sum_{i=0}^{\infty} P(X=n/Y=i)P(Y=i) = \\ = \sum_{i=n}^{\infty} P(X=n/Y=i)P(Y=i).$$

Xác suất $P(X=n/Y=i)$ là xác suất để người kiểm tra chỉ phát hiện đúng $i-n$ hạt giống hỏng với điều kiện trong bao có i hạt giống hỏng. Do xác suất để phát hiện có đúng một hạt giống hỏng là p, theo công thức Bernoulli

$$P(X=n/Y=i) = C_i^n p^{i-n} q^n \quad (q=1-p).$$

Suy ra

$$P(X=n) = \sum_{i=n}^{\infty} C_i^n p^{i-n} q^n e^{-\lambda}\frac{\lambda^i}{i!} \\ = e^{-\lambda} q^n \sum_{i=n}^{\infty} \frac{i! p^{i-n}}{n!(i-n)!}\frac{\lambda^i}{i!} = e^{-\lambda}\frac{q^n\lambda^n}{n!}\sum_{i=n}^{\infty}\frac{(p\lambda)^{i-n}}{(i-n)!} \\ = e^{-\lambda}\frac{q^n\lambda^n}{n!}\sum_{k=0}^{\infty}\frac{(p\lambda)^k}{k!} = e^{-\lambda}\frac{(q\lambda)^n}{n!}e^{p\lambda} = e^{-q\lambda}\frac{(q\lambda)^n}{n!}.$$

Vậy X là đại lượng ngẫu nhiên có phân bố Poisson với tham số $q\lambda$. Trong chương II, chúng ta đã chứng minh kì vọng của phân bố Poisson với tham số $q\lambda$ bằng tham số đó

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

Vậy số trung bình số hạt giống hỏng còn lại sau khi đã được kiểm tra trước khi gieo trồng là $E(X) = q\lambda$.

Cho đến lúc này chúng ta chỉ định nghĩa xác suất có điều kiện $P(A/B)$ nếu $P(B) > 0$. Tuy nhiên bây giờ ta sẽ mở rộng khái niệm xác suất có điều kiện cho cả trường hợp $P(B) = 0$.

Giả thiết (X, Y) là véc tơ ngẫu nhiên có $h(x, y)$ là hàm mật độ chung. Khi đó Y là đại lượng ngẫu nhiên liên tục, hàm mật độ của Y là

$$g(y) = \int_{-\infty}^{\infty} h(x, y)\, dx.$$

Theo nhận xét **??** chương II, các biến cố ngẫu nhiên $B = \{Y = y\}$ có xác suất bằng 0 vì Y là đại lượng ngẫu nhiên liên tục có hàm mật độ:

$$P(B) = P(Y = y) = 0$$

Ta định nghĩa xác suất có điều kiện của biến cố $\{X < x\}$ với điều kiện $Y = y$ như là giới hạn của $P(X < x/y \le Y < y + \Delta y)$ khi Δy dần tới 0.

Hàm

$$F(x/y) = \lim_{\Delta y \to 0} P(X < x/y \le Y < y + \Delta y)$$

được gọi là *hàm phân bố có điều kiện của X với điều kiện $Y = y$*, tất nhiên với giả thiết tồn tại giới hạn trên.

Do định nghĩa xác suất có điều kiện và tính chất của hàm phân bố chung

$$\begin{aligned} P(X < x/y \le Y < y + \Delta y) &= \frac{P(X < x, y \le Y < y + \Delta y)}{P(y \le Y < y + \Delta y)} \\ &= \frac{H(x, y + \Delta y) - H(x, y)}{G(y + \Delta y) - G(y)}. \end{aligned}$$

($H(x,y)$ là hàm phân bố chung của X và Y, $G(y)$ là hàm phân bố của Y). Chia cả tử và mẫu cho Δy, chuyển qua giới hạn khi $\Delta y \to 0$ ta được

$$F(x/y) = \frac{\frac{\partial}{\partial y}H(x,y)}{g(y)}.$$

Do $h(x,y)$ là hàm mật độ chung của véc tơ ngẫu nhiên (X,Y), khi đó tồn tại đạo hàm riêng hàm số ở vế phải của đẳng thức trên, kí hiệu

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

$f(x/y)$ được gọi là *hàm mật độ có điều kiện của X với điều kiện $Y = y$*. Ta có

$$f(x/y) = \frac{\frac{\partial^2}{\partial x \partial y}H(x,y)}{g(y)} = \frac{h(x,y)}{g(y)}.$$

Chú ý rằng các hàm mật độ có điều kiện cũng như phân bố có điều kiện ở đây chỉ được xác định tại y mà $g(y) > 0$. Tại những điểm mà $g(y) = 0$, hàm mật độ $f(x/y)$ được xác định tùy ý (để đơn giản, tại đó người ta thường gán cho $f(x/y)$ giá trị 0). Viết chính xác hơn, mật độ có điều kiện

$$f(x/y) = \begin{cases} \frac{h(x,y)}{g(y)} & \text{nếu} \quad g(y) > 0 \\ 0 & \text{nếu} \quad g(y) = 0. \end{cases}$$

Tương tự hàm mật độ có điều kiện của Y với điều kiện $X = x$

$$g(y/x) = \begin{cases} \frac{h(x,y)}{f(x)} & \text{nếu} \quad f(x) > 0 \\ 0 & \text{nếu} \quad f(x) = 0. \end{cases}$$

Suy ra $h(x,y) = f(x/y)g(y) = g(y/x)f(x)$. Kết hợp với các đẳng thức ở mục 2 chương này:

$$f(x) = \int_{-\infty}^{\infty} h(x,y)\,dy$$

$$g(y) = \int_{-\infty}^{\infty} h(x,y)\,dx$$

ta nhận được các công thức tương tự như công thức xác suất đầy đủ

$$f(x) = \int_{-\infty}^{\infty} f(x/y)g(y)\,dy$$

$$g(y) = \int_{-\infty}^{\infty} g(y/x)f(x)\,dx.$$

Tương tự như công thức Bayes trong trường hợp liên tục là các công thức sau

$$f(x/y) = \frac{g(y/x)f(x)}{\int_{-\infty}^{\infty} g(y/x)f(x)\,dx}$$

$$g(y/x) = \frac{f(x/y)g(y)}{\int_{-\infty}^{\infty} f(x/y)g(y)\,dy}.$$

Trở lại ví dụ 3.2.3

Ví dụ 3.3.3

Giả sử véc tơ ngẫu nhiên (X, Y) có phân bố đều trên hình tròn tâm $I(0, R)$ bán kính R. Hãy tìm hàm mật độ có điều kiện của X với điều kiện $Y = y$.

Giải: Hàm mật độ chung của X và Y là

$$h(x,y) = \begin{cases} \frac{1}{\pi R^2} & \text{nếu } x^2 + (y-R)^2 < R^2 \\ 0 & \text{trong trường hợp ngược lại} \end{cases}$$

Trong ví dụ 3.2.3 ta đã biết hàm mật độ của Y bằng

$$g(y) = \begin{cases} \frac{2\sqrt{R^2-(y-R)^2}}{\pi R^2} & \text{nếu} \quad 0 < y < 2R \\ 0 & \text{nếu} \quad y \notin (0, 2R). \end{cases}$$

Theo công thức tính hàm mật độ có điều kiện $f(x/y) = \frac{h(x,y)}{g(y)}$, hàm mật độ $f(x/y)$ của X với điều kiện $Y = y$ là

$$f(x/y) = \begin{cases} \frac{1}{2\sqrt{R^2-(y-R)^2}} & \text{nếu} \quad |x| < \sqrt{R^2-(y-R)^2} \\ 0 & \text{nếu} \quad |x| \geq \sqrt{R^2-(y-R)^2}. \end{cases}$$

Do $f(x/y)$ là hằng số (không phụ thuộc vào x) trong khoảng

$$\left(-\sqrt{R^2-(y-R)^2}, \sqrt{R^2-(y-R)^2}\right)$$

suy ra phân bố của X với điều kiện $Y=y$ là phân bố đều trên khoảng $\left(-\sqrt{R^2-(y-R)^2}, \sqrt{R^2-(y-R)^2}\right)$.

Nhận xét rằng nếu tồn tại hàm mật độ đồng thời $h(x,y)$ của X và Y khi đó tồn tại hàm mật độ có điều kiện $f(x/y)$ của X với điều kiện $Y=y$ và

$$P(X<x/Y=y)=F(x/y)=\int_{-\infty}^{x} f(t/y)\,dt.$$

Mặt khác một tính chất đặc trưng của hàm mật độ là

$$P(X\in A)=\int_A f(x)\,dx \quad \text{với mọi tập} \quad A\subset\mathbb{R}.$$

Do vậy ta có thể mở rộng khái niệm hàm phân bố có điều kiện như sau

$$P(X\in A/Y=y)=\int_A f(x/y)\,dx$$

được gọi là xác suất của biến cố $\{X\in A\}$ *với điều kiện* $Y=y$ *và kí hiệu*

$$P_X(A/y)=\int_A f(x/y)\,dx.$$

Nhân cả hai vế với hàm mật độ $g(y)$ của Y sau đó tích phân theo biến y ta được

$$\int_{-\infty}^{\infty} P_X(A/y)dy=\int_{-\infty}^{\infty}\int_A f(x/y)g(y)\,dxdy=$$

$$=\int_A\int_{-\infty}^{\infty} f(x/y)g(y)\,dydx=\int_A f(x)dx=P(X\in A).$$

Như vậy khái niệm xác suất có điều kiện $P(A/B)$ định nghĩa trong chương I đã dược mở rộng cho cả trường hợp biến cố B là biến cố $\{Y=y\}$ có xác suất bằng 0 (do Y là đại lượng ngẫu nhiên liên tục nên P(Y=y)=0). Đặc

trưng cơ bản của khái niệm xác suất có điều kiện mở rộng $P_X(A/y)$ là hệ thức

$$P(X \in A) = \int_{-\infty}^{\infty} P_X(A/y)dy.$$

Ví dụ 3.3.4

Lần lượt chọn ngẫu nhiên hai điểm: một điểm C trên biên của một hình vuông và một điểm B bên trong hình vuông đó. Các đoạn thẳng nối điểm C (chọn trên biên) với các đỉnh hình vuông chia hình vuông thành ba tam giác. Hãy tìm xác suất để điểm B thuộc tam giác bé nhất (giả thiết rằng hai lần chọn độc lập với nhau).

Giải: Không mất tính tổng quát ta giả thiết rằng cạnh của hình vuông bằng 1. Cố định một đỉnh bất kì của hình vuông (đỉnh D) và quy ước một hướng dọc theo biên hình vuông, gọi Y là độ dài dọc theo hướng đó từ đỉnh D tới điểm chọn ngẫu nhiên trên biên, A là biến cố "điểm B thuộc tam giác bé nhất". Do Y phân bố đều trên biên nên hàm mật độ của Y:

$$g(y) = \begin{cases} \frac{1}{4} & \text{nếu} \quad 0 < y < 4 \\ 0 & \text{nếu} \quad y \notin (0,4). \end{cases}$$

Theo công thức trên

$$P(A) = \int_0^4 \frac{1}{4} P(A/y)dy.$$

Do tính đối xứng của hình vuông, dễ dàng tính được xác suất có điều kiện

$$P(A/y) = \frac{y}{2} \quad \text{nếu} \quad 0 < y < \frac{1}{2}.$$

Suy ra

$$P(A) = \int_0^{\frac{1}{2}} \frac{8}{4} P(A/y)dy = \int_0^{\frac{1}{2}} ydy = \frac{1}{8}.$$

3.4 Sự độc lập của các đại lượng ngẫu nhiên

Định nghĩa 3.4.1 *Người ta gọi X và Y là hai đại lượng ngẫu nhiên độc lập nhau, nếu với các số thực bất kì $a < b, c < d$, ta luôn có*

$$P(a \leq X < b, c \leq Y < d) = P(a \leq X < b)P(c \leq Y < d)$$

Dựa vào định lí 1.3.5 ở chương I và các tính chất cơ bản của xác suất ta dễ dàng chứng minh rằng định nghĩa trên tương đương với

$$H(x, y) = F(x)G(y) \quad \text{với mọi} \quad x, y \in \mathbb{R}$$

Trong đó $F(x), G(y), H(x, y)$ là các hàm phân bố của X, Y cũng như phân bố chung của (X, Y) tương ứng.

Đặc biệt khi ξ là đại lượng ngẫu nhiên rời rạc có thể nhận các giá trị $x_i, i = 1, 2, ...$ và η cũng là đại lượng ngẫu nhiên rời rạc nhận các giá trị $y_j, j = 1, 2, ...$ Khi đó sự độc lập của ξ và η tương đương với các đẳng thức sau

$$P(\xi = x_i, \eta = y_j) = P(\xi = x_i)P(\eta = y_j) \quad \text{với mọi} \quad i, j.$$

Trường hợp đại lượng ngẫu nhiên X, Y liên tục, $f(x), g(y), h(x, y)$ là các hàm mật độ của X, Y cũng như mật độ chung của (X, Y). Lần lượt đạo hàm hai vế

$$H(x, y) = F(x)G(y)$$

theo x và theo y, ta được

$$h(x, y) = f(x)g(y) \quad \text{với mọi} \quad x, y \in \mathbb{R}.$$

Ngược lại từ đẳng thức $h(x, y) = f(x)g(y),\ \forall x, y \in \mathbb{R}$ suy ra

$$H(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} h(u, v)\, dudv = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u)g(v)\, dudv$$

$$= \int_{-\infty}^{x} f(u)du \cdot \int_{-\infty}^{y} g(v)\, dv = F(x) \cdot G(y).$$

Vậy định nghĩa trên có thể phát biểu dưới dạng tương đương:

X và Y là hai đại lượng ngẫu nhiên độc lập, nếu hàm mật độ chung của chúng thoả mãn

$$h(x,y) = f(x)g(y) \quad \text{với mọi} \quad x, y \in \mathbb{R}.$$

Liên quan tới sự độc lập của các đại lượng ngẫu nhiên, ta có thể nói đúng như đã nói về sự độc lập của các biến cố ngẫu nhiên. Trong nhiều bài toán ứng dụng chúng ta cần khẳng định xem các đại lượng ngẫu nhiên ξ và η có độc lập với nhau không. Ví dụ thu nhập của các gia đình và số người trong gia đình có là các đại lượng ngẫu nhiên độc lập với nhau không, hay cũng hỏi như vậy với chiều cao và trọng lượng của một người... Những vấn đề này thống kê xác suất sẽ cho câu trả lời phụ thuộc vào các số liệu thống kê thu thập được. Ở những trường hợp khác chúng ta sử dụng giả thiết hai đại lượng ngẫu nhiên ξ và η độc lập với nhau và dựa vào đó để suy ra các xác suất của các biến cố liên quan. Chẳng hạn nếu mỗi đại lượng ngẫu nhiên gắn với một phép thử khác nhau, mà các phép thử ngẫu nhiên đó không có mối liên quan, không ảnh hưởng hoặc phụ thuộc một chút nào vào nhau. Như đã nói đến trong chương I về các biến cố độc lập, khi đó ta xem chúng là các đại lượng ngẫu nhiên độc lập.

Định lí 3.4.1 *Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập, tồn tại kì vọng EX, EY. Khi đó đại lượng ngẫu nhiên tích XY cũng tồn tại kì vọng và*

$$E(XY) = E(X)E(Y).$$

Chứng minh Nếu X, Y là các đại lượng ngẫu nhiên rời rạc. Kí hiệu

$$p_i = P(X = x_i), q_k = P(Y = y_k), r_{ik} = P(X = x_i, Y = y_k)$$

Do X và Y độc lập $r_{ik} = p_i q_k$, theo định nghĩa kì vọng của đại lượng ngẫu nhiên

$$E(XY) = \sum_i \sum_k x_i y_k r_{ik} = \sum_i \sum_k x_i y_k p_i q_k =$$

$$= \left(\sum_i x_i p_i\right)\left(\sum_k y_k q_k\right) = E(X)E(Y).$$

Trường hợp X, Y là các đại lượng ngẫu nhiên liên tục có $f(x), g(y)$ là các hàm mật độ tương ứng. Khi đó $h(x,y) = f(x)g(y)$ là hàm mật độ chung của chúng. Áp dụng định lí 3.2.1 đối với hàm $\varphi(x,y) = xy$ của hai đại lượng ngẫu nhiên X và Y

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(xy)h(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x)g(y)dxdy =$$

$$= \int_{-\infty}^{\infty}xf(x)\,dx\int_{-\infty}^{\infty}yg(y)\,dy = E(X)E(Y).$$

Bằng quy nạp ta suy ra nếu $X_1, X_2, ..., X_n$ là các đại lượng ngẫu nhiên độc lập, tồn tại kì vọng $E(X_i), \quad i = 1, 2, ..., n$. Khi đó

$$E(X_1X_2...X_n) = E(X_1)E(X_2)...E(X_n).$$

Chúng ta công nhận không chứng minh định lí sau

Định lí 3.4.2 *Giả sử ξ và η là hai đại lượng ngẫu nhiên độc lập, φ và ψ là hai hàm thực tùy ý. Khi đó $\varphi(\xi)$ và $\psi(\eta)$ cũng độc lập với nhau.*

Nhận xét rằng khi đó theo định lí 3.4.1 nếu tồn tại kì vọng $E(\varphi(\xi))$ cũng như $E(\psi(\eta))$ suy ra

$$E(\varphi(\xi)\psi(\eta)) = E(\varphi(\xi))E(\psi(\eta)).$$

Theo một nghĩa nhất định nào đó, hầu hết các tập hợp số thực đều được "sinh ra" bởi các tập hợp $[a,b)$, trong đó $a, b \in \mathbb{R}$ (chính xác hơn hầu hết các tập hợp số thực đều được biểu diễn bởi các phép toán tập hợp như hợp, giao hoặc lấy phần bù,... các tập có dạng $[a_i, b_i)$). Do vậy định nghĩa 3.4.1 cũng tương đương với khẳng định sau:

X và Y độc lập, nếu với bất kì hai tập E, F trong $\mathbb{R}$, ta luôn có

$$P(X \in E, Y \in F) = P(X \in E) \cdot P(Y \in F).$$

Nói cách khác $\{X \in E\}$ và $\{Y \in F\}$ là hai biến cố độc lập.

Vì vậy trong định lí trên do giả thiết ξ và η độc lập nên

$$\{a \leq \varphi(\xi) < b\} = \{\xi \in \varphi^{-1}([a,b)\} = \{\xi \in E\}$$

và

$$\{c \leq \psi(\eta) < d\} = \{\eta \in \psi^{-1}([c,d)\} = \{\eta \in F\}$$

cũng là hai biến cố độc lập. Suy ra $\varphi(\xi)$ và $\psi(\eta)$ là hai đại lượng ngẫu nhiên độc lập.

Hệ quả 3.4.1 *Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập, tồn tại phương sai $D(X), D(Y)$. Khi đó tổng của hai đại lượng ngẫu nhiên $X+Y$ cũng tồn tại phương sai và*

$$D(X+Y) = D(X) + D(Y).$$

Chứng minh. Theo định nghĩa

$$\begin{aligned} D(X+Y) &= E\left((X+Y-(EX+EY))^2\right) \\ &= E\left((X-EX)^2 + (Y-EY)^2 + 2(X-EX)(Y-EY)\right) \\ &= D(X) + D(Y) + 2E((X-EX)(Y-EY)) \\ &= D(X) + D(Y) + 2E(X-EX)E(Y-EY). \end{aligned}$$

$E((X-EX)(Y-EY)) = E(X-EX)E(Y-EY)$ do áp dụng các định lí 3.4.1 và 3.4.2 đối với hai đại lượng ngẫu nhiên độc lập X và Y, mặt khác $E(X-EX) = E(Y-EY) = 0$, suy ra

$$D(X+Y) = D(X) + D(Y), \quad \text{đ.p.c.m.}$$

Bằng quy nạp ta suy ra nếu $X_1, X_2, ..., X_n$ là các đại lượng ngẫu nhiên độc lập, tồn tại phương sai $D(X_i), \quad i = 1, 2, ..., n$, khi đó

$$D(X_1 + X_2 + ... + X_n) = D(X_1) + D(X_2) + ... + D(X_n).$$

Trong mục 2.5.3 chương II khi xét phân bố nhị thức, chúng ta đã tính phương sai của đại lượng ngẫu nhiên có phân bố nhị thức. Giả sử X có

phân bố nhị thức với 2 tham số p và n, khi đó phương sai $D(X) = npq$. Ta nhớ lại rằng X có thể biểu diễn như tổng của các các đại lượng ngẫu nhiên X_i, trong đó X_i là đại lượng ngẫu nhiên nhận các giá trị 1 hoặc 0 tùy theo ở lần thử thứ i biến cố A xảy ra hay không

$$P(X_i = 1) = P(A) = p, P(X_i = 0) = P(\overline{A}) = q \quad i = 1, 2, ..., n$$

$$X = X_1 + X_2 + ... + X_n$$

Dễ dàng tính được $D(X_i) = pq$. Mặt khác do $X_i \quad i = 1, 2, ..., n$ chỉ phụ thuộc vào lần thử thứ i nên chúng là các đại lượng ngẫu nhiên độc lập. Sử dụng hệ quả nêu trên, bây giờ ta có thể tính phương sai $D(X)$ của phân bố nhị thức một cách gọn hơn

$$D(X) = D(X_1) + D(X_2) + ... + D(X_n) = npq.$$

Ví dụ 3.4.1

Giả sử véc tơ ngẫu nhiên (X, Y) có phân bố đều trên miền hình chữ nhật $[a, b] \times [c, d]$. Ta có thể khẳng định X và Y độc lập với nhau.

Thật vậy như đã xét trong ví dụ 3.2.3 hàm mật độ chung của (X, Y) và các hàm mật độ thành phần của X, Y tương ứng là

$$h(x, y) = \begin{cases} \frac{1}{(b-a)(d-c)} & \text{nếu } a < x < b, c < y < d \\ 0 & \text{nếu } x \notin (a, b) \text{ hoặc } y \notin (c, d) \end{cases}$$

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

$$g(y) = \begin{cases} \frac{1}{d-c} & \text{nếu } c < y < d \\ 0 & \text{nếu } y \notin (c, d) \end{cases}$$

Hiển nhiên $h(x, y) = f(x)g(y)$ $\forall x, y \in \mathbb{R}$, suy ra X và Y độc lập.

Ví dụ 3.4.2

Giả sử véc tơ ngẫu nhiên (X, Y) có phân bố đều trên hình tròn tâm $I(0, 0)$ bán kính R. Dễ dàng chứng minh được hai đại lượng ngẫu

nhiên X và Y không độc lập. Tuy nhiên nếu kí hiệu (φ, r) là toạ độ cực của điểm ngẫu nhiên (X, Y)

$$X = r\cos\varphi, Y = r\sin\varphi \quad 0 \leq \varphi \leq 2\pi, 0 \leq r \leq R$$

Khi đó φ, r là hai đại lượng ngẫu nhiên độc lập.

Giải: Trước hết ta tính xác suất biến cố tích

$$\{\varphi_1 \leq \varphi < \varphi_2\} \cdot \{r_1 \leq r < r_2\}$$

với $0 < \varphi_1 < \varphi_2 < 2\pi, 0 < r_1 < r_2 < R$.

Xét trong hệ toạ độ Đề các, đây là một phần của hình quạt tròn giới hạn bởi 2 hình tròn $r = r_1, r = r_2$ và 2 tia $\varphi = \varphi_1, \varphi = \varphi_2$. Do X, Y có phân bố đều, xác suất để X, Y thuộc vào miền đó bằng tỉ số giữa diện tích của miền phẳng đó và diện tích hình tròn bán kính $2R$

$$P(\varphi_1 \leq \varphi < \varphi_2, r_1 \leq r < r_2) = \frac{r_2^2(\varphi_2 - \varphi_1) - r_1^2(\varphi_2 - \varphi_1)}{2\pi R^2} =$$

$$= \frac{r_2^2 - r_1^2}{R^2} \cdot \frac{\varphi_2 - \varphi_1}{2\pi}$$

Đặc biệt khi $r_1 = 0, r_2 = R$

$$P(\varphi_1 \leq \varphi < \varphi_2) = P(\varphi_1 \leq \varphi < \varphi_2, 0 \leq r < R) = \frac{\varphi_2 - \varphi_1}{2\pi}.$$

Tương tự

$$P(r_1 \leq r < r_2) = P(0 \leq \varphi < 2\pi, r_1 \leq r < r_2) = \frac{r_2^2 - r_1^2}{R^2}.$$

Vậy định nghĩa 3.4.1 được thoả mãn

$$P(\varphi_1 \leq \varphi < \varphi_2, r_1 \leq r < r_2) = P(\varphi_1 \leq \varphi < \varphi_2)P(r_1 \leq r < r_2).$$

Với các trường hợp còn lại $\varphi_1, \varphi_2 \notin (0, 2\pi)$ hoặc $r_1, r_2 \notin (0, R)$, hiển nhiên đẳng thức vẫn đúng. Suy ra φ và r độc lập.

Nhận xét rằng do đẳng thức $P(\varphi_1 \leq \varphi < \varphi_2) = \frac{\varphi_2 - \varphi_1}{2\pi}$, suy ra φ có phân bố đều trên khoảng $(0, 2\pi)$.

Phân bố của r được suy ra từ đẳng thức

$$P(r_1 \leq r < r_2) = \frac{r_2^2 - r_1^2}{R^2},$$

nếu thay vào đẳng thức đó $r_2 = u, r_1 = 0$, với $0 < u < R$:

$$P(r < u) = \frac{u^2}{R^2}.$$

Suy ra hàm mật độ của r bằng $\frac{2u}{R^2}$. Vậy hàm mật độ chung của (r, φ)

$$g(u, v) = \frac{1}{2\pi}\frac{2u}{R^2} = \frac{u}{\pi R^2}$$

với $0 \leq u \leq R, 0 \leq v \leq 2\pi$ và $g(u, v) = 0$ trong trường hợp ngược lại.

Định lí 2.4.1 của chương II đã cho ta công thức tính hàm mật độ của một đại lượng ngẫu nhiên là hàm của đại lượng ngẫu nhiên khác. Vấn đề cũng được đặt ra tương tự đối với véc tơ ngẫu nhiên hai chiều.

Giả sử φ là một song ánh

$$\varphi : D \to T \quad D \subset R^2, T \subset R^2$$

khả vi tại mọi điểm thuộc D. (X, Y) là véc tơ ngẫu nhiên nhận các giá trị trong D và $h(x, y)$ là hàm mật độ đồng thời của véc tơ ngẫu nhiên đó. Định lí sau sẽ chỉ ra cách tính hàm mật độ chung của hàm $\varphi(X, Y)$ của hai đại lượng ngẫu nhiên X, Y.

Định lí 3.4.3 *Với các giả thiết như đã nêu trên, khi đó hàm mật độ của* $(U, V) = \varphi(X, Y)$ *bằng*

$$g(u, v) = h\left(\varphi^{-1}(u, v)\right) |J(u, v)|$$

trong đó $J(u, v)$ *là Jacobien của* φ^{-1}. *(Ta nhắc lại rằng trong lí thuyết hàm nhiều biến, nếu kí hiệu* $(x, y) = \varphi^{-1}(u, v)$ *là ánh xạ ngược của* φ, *khi đó Jacobien của* φ^{-1} *bằng*

$$J(u, v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u})$$

Chứng minh. Xét $E \subset D$ là tập con bất kì của D. Hàm phân bố của $(U,V) = \varphi(X,Y)$ là

$$P((U,V) \in E) = P((X,Y) \in \varphi^{-1}(E)) = \iint_{\varphi^{-1}(E)} h(x,y)\,dxdy$$

Sử dụng phép đổi biến $(x,y) = \varphi^{-1}(u,v)$, trong đó

$$\varphi^{-1} : T \to D$$

là ánh xạ ngược của φ. Như đã biết trong phép đổi biến của tích phân bội

$$\iint_{\varphi^{-1}(E)} h(x,y)\,dxdy = \iint_E h(\varphi^{-1}(u,v))|J(u,v)|\,dudv$$

trong đó $J(u,v)$ là Jacobien của φ^{-1}. Kết hợp với đẳng thức trên

$$P((U,V) \in E) = \iint_E h(\varphi^{-1}(u,v))|J(u,v)|\,dudv$$

trong đó E là tập con bất kì của D. Suy ra $h\left(\varphi^{-1}(u,v)\right)|J(u,v)|$ là hàm mật độ của $(U,V) = \varphi(X,Y)$, đ.p.c.m.

Ví dụ 3.4.3

Áp dụng vào ví dụ 3.4.2 vừa xét, giả sử η là song ánh từ hình tròn tâm $I(0,0)$ lên hình chữ nhật $[0,R] \times [0,2\pi]$, sao cho ánh xạ ngược $(x,y) = \eta^{-1}(u,v)$ là phép chuyển sang toạ độ cực.

$$x = u\cos v, y = u\sin v \quad \Rightarrow \quad |J(u,v)| = u$$

Mặt khác h là hàm mật độ của phân bố đều trên hình tròn bán kính R nên

$$h(x,y) = \frac{1}{\pi R^2} \quad \text{nếu} \quad x^2 + y^2 \leq R^2.$$

Suy ra

$$h\left(\eta^{-1}(u,v)\right) = \frac{1}{\pi R^2} \quad \text{với} \quad (u,v) \in [0,R] \times [0,2\pi]$$

Theo định lí 4.4.2 hàm mật độ chung của (r,φ) bằng

$$g(u,v) = h\left(\eta^{-1}(u,v)\right)|J(u,v)| = \frac{1}{\pi R^2}u$$

với $(u, v) \in [0, R] \times [0, 2\pi]$, hay

$$g(u,v) = \begin{cases} \frac{u}{\pi R^2} & \text{nếu} \quad (u,v) \in [0,R] \times [0,2\pi] \\ 0 & \text{nếu} \quad (u,v) \notin [0,R] \times [0,2\pi] \end{cases}$$

phù hợp với nhận xét sau ví dụ 3.4.2 về hàm mật độ chung của φ và r, tức là của u, v - bởi cách kí hiệu ở đây.

Ví dụ 3.4.4

Cho X và Y là hai đại lượng ngẫu nhiên độc lập, X có phân bố mũ với tham số $\lambda_1 = 1$, Y cũng là đại lượng ngẫu nhiên có phân bố mũ với tham số $\lambda_2 = 2$. Gọi $U = X + 2Y, V = \frac{X}{X+2Y}$. Hãy tìm hàm mật độ chung của U và V. Từ đó suy ra U và V độc lập và hãy tìm các hàm mật độ của U, của V.

Hàm mật độ của X

$$f_X = \begin{cases} e^{-x} & \text{nếu} \quad x > 0 \\ 0 & \text{nếu} \quad x \leq 0 \end{cases}$$

và hàm mật độ của Y

$$f_Y = \begin{cases} 2e^{-2y} & \text{nếu} \quad y > 0 \\ 0 & \text{nếu} \quad y \leq 0 \end{cases}$$

Theo giả thiết X và Y độc lập, suy ra hàm mật độ đồng thời của véc tơ ngẫu nhiên (X, Y) là

$$h(x,y) = \begin{cases} 2e^{-x-2y} & \text{nếu} \quad x > 0, y > 0 \\ 0 & \text{nếu trái lại.} \end{cases}$$

Dễ dàng chứng minh được ánh xạ

$$\varphi(x,y) = (u,v) = \left(x + 2y, \frac{x}{x+2y} \right)$$

là một song ánh từ $D = \{(x,y)/ \quad x > 0, y > 0\} \subset \mathbb{R}^2$ lên tập hợp $T = \{(u,v)/ \quad u > 0, 0 < v < 1\} \subset \mathbb{R}^2$. Ánh xạ ngược

$$\varphi^{-1}(u,v) = \left(uv, \frac{u - uv}{2} \right)$$

có Jacobien bằng

$$J(u,v) = \begin{vmatrix} v & u \\ \frac{1-v}{2} & \frac{-u}{2} \end{vmatrix} = -\frac{u}{2}$$

Do $h\left(\varphi^{-1}(u,v)\right)|J(u,v)| = 2e^{-u} \cdot \frac{u}{2} = ue^{-u}$, theo định lí 4. hàm mật độ chung $t(u,v)$ của (U,V) bằng

$$t(u,v) = \begin{cases} ue^{-u} & \text{nếu} \quad u > 0, 0 < v < 1 \\ 0 & \text{nếu trái lại.} \end{cases}$$

Dễ dàng tính được hàm mật độ của U bằng

$$f(u) = \int_{-\infty}^{\infty} t(u,v)dv = \int_0^1 t(u,v)dv = \begin{cases} ue^{-u} & \text{nếu} \quad u > 0 \\ 0 & \text{nếu} \quad u \leq 0 \end{cases}$$

và hàm mật độ của V bằng

$$g(v) = \int_{-\infty}^{\infty} t(u,v)du = \int_0^{\infty} t(u,v)du = \begin{cases} 1 & \text{nếu} \quad 0 < v < 1 \\ 0 & \text{nếu} \quad v \notin (0,1). \end{cases}$$

Vậy $t(u,v) = f(u)g(v)$ suy ra U và V độc lập đồng thời từ hàm mật độ của V ta thấy V là đại lượng ngẫu nhiên có *phân bố đều trên đoạn* $[0,1]$.

Ví dụ 3.4.5

Ta xét một ứng dụng của định lí 4.4.2 trong trường hợp ánh xạ φ là phép biến đổi tuyến tính không suy biến, $(x,y) = \varphi^{-1}(u,v)$ có dạng

$$x = a_{11}u + a_{12}v$$

$$y = a_{21}u + a_{22}v$$

Gọi $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ là ma trận của ánh xạ tuyến tính đó.

Khi đó Jacobien của φ^{-1} bằng $det(A)$, do vậy hàm mật độ chung của U và V, với $(U,V) = \varphi(X,Y)$ là

$$g(u,v) = h\left(a_{11}u + a_{12}v, a_{21}u + a_{22}v\right)|det(A)|$$

Ta viết lại kết quả trên dưới dạng hệ quả sau:

Hệ quả 3.4.2 *$h(x,y)$ là hàm mật độ đồng thời của véc tơ ngẫu nhiên (X,Y). A là ma trận vuông cấp hai không suy biến. Véc tơ ngẫu nhiên (U,V) được xác định bởi hệ thức*

$$X = a_{11}U + a_{12}V$$
$$Y = a_{21}U + a_{22}V$$

Khi đó hàm mật độ đồng thời của véc tơ ngẫu nhiên (U,V) bằng

$$g(u,v) = h\left(a_{11}u + a_{12}v, a_{21}u + a_{22}v\right)|det(A)|$$

3.5 Tổng của hai đại lượng ngẫu nhiên độc lập

Trong thực hành người ta rất hay gặp bài toán tìm hàm phân bố (hoặc hàm mật độ) của tổng hai đại lượng ngẫu nhiên độc lập X và Y

$$V = X + Y.$$

Chẳng hạn khi gieo đồng thời hai xúc xắc, X và Y là số chấm xuất hiện ở mỗi xúc xắc, V là tổng số chấm xuất hiện ở hai xúc xắc. Người ta muốn biết phân bố xác suất của V.

Chúng ta chỉ giới hạn bài toán trong trường hợp cả hai đại lượng ngẫu nhiên độc lập X và Y cùng rời rạc hoặc cùng liên tục. Trước hết chúng ta giả thiết X và Y là hai đại lượng ngẫu nhiên độc lập, chỉ nhận các giá trị nguyên. Kí hiệu

$$p_i = P(X = i),$$
$$q_k = P(Y = k),$$
$$r_n = P(V = X + Y = n)$$

Chúng ta sẽ tìm công thức để tính xác suất $r_n = P(V = X + Y = n)$ theo các xác suất p_i, q_k của X và Y. Rõ ràng biến cố $\{V = n\}$ là tổng của các biến cố đôi một xung khắc nhau được liệt kê sau đây

$$\begin{gathered}
\ldots\ldots\ldots\ldots\ldots\ldots \\
\{X = n+2\} \cdot \{Y = -2\} \\
\{X = n+1\} \cdot \{Y = -1\} \\
\{X = n\} \cdot \{Y = 0\} \\
\{X = n-1\} \cdot \{Y = 1\} \\
\ldots\ldots\ldots\ldots\ldots\ldots
\end{gathered} \tag{3.1}$$

Suy ra

$$r_n = P(V = n) = \sum_{k=-\infty}^{\infty} P(X = n-k, Y = k)$$

Do X và Y độc lập, nên

$$P(X = n-k, Y = k) = P(X = n-k)P(Y = k) = p_{n-k}q_k$$

Vậy

$$r_n = P(V = n) = \sum_{k=-\infty}^{\infty} p_{n-k}q_k$$

Đặc biệt khi hai đại lượng ngẫu nhiên X và Y chỉ nhận các giá trị nguyên không âm. Khi đó $p_k = q_k = 0$ với $k = -1, -2, \ldots$.
Suy ra xác suất của biến cố $\{X + Y = n\}$ bằng

$$r_n = P(X + Y = n) = \sum_{k=0}^{n} p_{n-k}q_k.$$

Áp dụng cho tổng của hai đại lượng ngẫu nhiên độc lập, có phân bố Poisson X và Y với tham số λ_1 và λ_2 tương ứng

$$r_n = P(X + Y = n) = \sum_{k=0}^{n} p_{n-k}q_k =$$

$$= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^{n-k}}{(n-k)!} e^{-\lambda_2} \frac{\lambda_2^k}{k!} =$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!}\sum_{k=0}^{n} C_n^k \lambda_2^k \lambda_1^{n-k} = \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}.$$

Đây cũng là một số hạng của phân bố Poisson. Vậy ta có hệ quả sau

Hệ quả 3.5.1 *Gọi X và Y là hai đại lượng ngẫu nhiên độc lập, có phân bố Poisson với tham số λ_1 và λ_2 tương ứng. Khi đó tổng $X+Y$ cũng là đại lượng ngẫu nhiên có phân bố Poisson với tham số $\lambda_1+\lambda_2$.*

Bây giờ ta xét trường hợp cả hai đại lượng ngẫu nhiên X và Y liên tục với $f(x)$ và $g(y)$ là các hàm mật độ tương ứng. Giả thiết tiếp rằng X và Y độc lập với nhau. Để tính hàm mật độ của $V = X+Y$ ta dẫn vào phép biến đổi tuyến tính sau

$$X = U$$
$$Y = -U + V$$

Jacobien của phép biến đổi tuyến tính A

$$J(u,v) = det(A) = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1$$

Hàm mật độ chung của (X,Y) bằng $h(x,y) = f(x)g(y)$. Theo hệ quả 3.4.2 mục trước, mật độ chung của (U,V) là

$$g_1(u,v) = h(u,-u+v)|det(A)| = f(u)g(v-u)$$

Áp dụng công thức tích phân theo từng biến của hàm mật độ chung $g_1(u,v)$ của (U,V), ta sẽ nhận được các hàm mật độ thành phần (đã trình bày trong mục 2. chương này). Suy ra hàm mật độ của $V = X+Y$ là

$$r(v) = \int_{-\infty}^{\infty} g_1(u,v)\,du = \int_{-\infty}^{\infty} f(u)g(v-u)\,du$$

Công thức $\int_{-\infty}^{\infty} f(u)g(v-u)\,du$ còn được gọi là *tích chập* của hai hàm f và g. Vai trò của X và Y trong biểu thức $V = X+Y$ như nhau (dẫn đến vai

trò của f và g cũng vậy). Lập luận tương tự, hàm mật độ của $V = X + Y$ còn có thể viết dưới dạng

$$r(v) = \int_{-\infty}^{\infty} f(v-u)g(u)\,du$$

Ta tóm tắt kết quả quan trọng vừa thu được trong hệ quả sau

Hệ quả 3.5.2 *Gọi $f(x)$ và $g(y)$ là các hàm mật độ của hai đại lượng ngẫu nhiên độc lập X và Y. Khi đó hàm mật độ của tổng $X + Y$ bằng*

$$r(v) = \int_{-\infty}^{\infty} f(u)g(v-u)\,du = \int_{-\infty}^{\infty} f(v-u)g(u)\,du.$$

Đặc biệt khi hai đại lượng ngẫu nhiên X và Y chỉ nhận các giá trị không âm $X \geq 0, Y \geq 0$, khi đó

$$f(x) = g(x) = 0 \quad \text{nếu} \quad x < 0.$$

Vì vậy công thức trong hệ quả trên được rút gọn lại

$$r(v) = \begin{cases} \displaystyle\int_0^v f(v-u)g(u)\,du = \int_0^v g(v-u)f(u)\,du & \text{nếu} \quad v > 0 \\ 0 & \text{nếu} \quad v \leq 0. \end{cases}$$

Nhận xét rằng ta có thể tính hàm mật độ của tổng hai đại lượng ngẫu nhiên độc lập $X + Y$ bằng phương pháp sau đây.

Hàm phân bố của tổng $X + Y$ bằng

$$R(z) = P(X + Y < z) = \iint_{(x,y):x+y<z} f(x)g(y)\,dxdy$$

hay

$$R(z) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} f(x)g(y)\,dy.$$

Đổi biến $y = v - x$, ta được

$$R(z) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z} f(x)g(v-x)\,dv = \int_{-\infty}^{z} \left(\int_{-\infty}^{\infty} f(x)g(v-x)\,dx \right) dv.$$

Đạo hàm hàm phân bố $R(z)$, ta được điều phải chứng minh

$$r(z) = R'(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,dx.$$

Ví dụ 3.5.1

Một thiết bị làm việc hàng ngày liên tục trong một khoảng thời gian T nào đó (ví dụ $T = 2$ giờ). Ngày nào thiết bị cũng làm việc trong khoảng thời gian đó. Gọi u_1 là xác suất để thiết bị hoạt động bình thường trong ngày và $v_1 = 1 - u_1$ là xác suất để thiết bị bị hỏng, X là số ngày làm việc liên tục của thiết bị cho đến khi bị hỏng.

Nếu thiết bị bị hỏng, nó được thay bằng một thiết bị khác tương tự. Giả sử xác suất để thiết bị thay thế hoạt động bình thường trong ngày là u_2 và xác suất để thiết bị bị hỏng $v_2 = 1 - u_2$ (u_1 và u_2 có thể bằng nhau hoặc khác nhau). Gọi Y là số ngày làm việc liên tục của thiết bị thay thế (cho đến khi nó cũng bị hỏng). Giả thiết rằng trong các ngày chúng làm việc, hoạt động của chúng độc lập nhau. Hãy tìm phân bố xác suất của tổng số ngày cả hai thiết bị đó làm việc $X + Y$.

Giải: Xác suất để số ngày làm việc liên tục của thiết bị thứ nhất bằng k là

$$p_k = P(X = k) = u_1^k v_1, \quad k = 0, 1, 2, ...$$

Tương tự xác suất để số ngày làm việc liên tục của thiết bị thứ hai bằng k là

$$p_k = P(Y = k) = u_2^k v_2, \quad k = 0, 1, 2, ...$$

Áp dụng công thức xác suất của tổng hai đại lượng ngẫu nhiên rời rạc, độc lập trình bày trên

$$r_n = P(X + Y = n) = \sum_{k=0}^{n} u_1^{n-k} v_1 u_2^k v_2 =$$

$$= u_1^n v_1 v_2 \sum_{k=0}^{n} \left(\frac{u_2}{u_1}\right)^k = u_1^n v_1 v_2 \frac{\left(\frac{u_2}{u_1}\right)^{n+1} - 1}{\frac{u_2}{u_1} - 1} =$$

$$= v_1 v_2 \frac{u_2^{n+1} - u_1^{n+1}}{u_2} u_1, \quad n = 0, 1, 2, ...$$

với $u_1 \neq u_2$.

Trường hợp $u_1 = u_2 = p$, dễ dàng tính được

$$r_n = (n+1)q^2p^n, \quad n = 0, 1, 2, ...$$

trong đó $q = 1 - p$.

Nhận xét 3.5.1 *Nhận xét rằng X và Y trong ví dụ 3.5.1 là các đại lượng ngẫu nhiên có phân bố hình học với các tham số u_1 và u_2 tương ứng (chính xác hơn $X + 1$ và $Y + 1$ là hai đại lượng ngẫu nhiên có phân bố hình học). Như đã chứng minh trong ví dụ đó nếu $u_1 = u_2 = q$, khi đó tổng của 2 đại lượng ngẫu nhiên độc lập X và Y có phân bố*

$$r_n = P(X + Y = n) = (n+1)p^2q^n, \quad n = 0, 1, 2, ...$$

Người ta gọi phân bố đó là phân bố hình học cấp 2. Bằng quy nạp ta có thể định nghĩa phân bố hình học cấp r là phân bố của tổng r đại lượng ngẫu nhiên độc lập có cùng phân bố hình học. Một cách tổng quát ta gọi ξ là đại lượng ngẫu nhiên có phân bố hình học cấp r, nếu

$$P(\xi = r + n) = C_{n+r-1}^{r-1} p^r q^n \qquad (n = 0, 1, 2, ...)$$

Ví dụ 3.5.2

Quay lại ví dụ 2.5.9 chương II, ta có cách tính sau đây đơn giản hơn cách tính kì vọng nêu trong ví dụ đó:

Một xạ thủ tập bắn bia, bắn liên tục vào bia cho đến khi có đủ 3 viên đạn trúng bia thì dừng. Giả sử xác suất bắn trúng bia của xạ thủ đó bằng p và các lần bắn độc lập nhau. Hãy tính số đạn trung bình mà xạ thủ đã sử dụng.

Áp dụng nhận xét trên ta có thể coi phân bố của số đạn xạ thủ-đã sử dụng cũng là phân bố của tổng 3 đại lượng ngẫu nhiên độc lập có cùng phân bố hình học với tham số p. Mặt khác ở chương II mục 2.5.4 ta đã chỉ ra kì vọng, phương sai của phân bố hình học bằng $\frac{1}{p}$ và $\frac{q}{p^2}$ tương ứng. Vậy giá trị trung bình của ξ trong ví dụ này (số đạn trung bình mà xạ thủ đã bắn) bằng

$$E(\xi) = \frac{3}{p}.$$

Ngoài ra do phương sai của tổng các đại lượng ngẫu nhiên độc lập bằng tổng các phương sai. Suy ra

$$D(\xi) = \frac{3q}{p^2}.$$

Chú ý rằng việc tính trực tiếp phương sai trong ví dụ 2.5.9 chương II phức tạp hơn rất nhiều so với phương pháp nêu trên, do vậy ta không đề cập đến trong chương trước.

Ví dụ 3.5.3

Số khách vào mua hàng trong một ngày tại một cửa hàng là đại lượng ngẫu nhiên có phân bố Poisson. Một công ty có hai cửa hàng đặt tại hai địa điểm A và B khác nhau trong thành phố. Trung bình một ngày cửa hàng A có 4 khách đến đặt mua hàng và cửa hàng B có 5 khách đến đặt mua hàng. Giả sử số khách đến các cửa hàng của công ty để mua hàng là các đại lượng ngẫu nhiên độc lập với nhau. Hãy tính xác suất để số khách đến đặt mua hàng ở cả hai cửa hàng của công ty trong một ngày lớn hơn 10.

Giải: Gọi X và Y là số khách đến đặt mua hàng ở cửa hàng A và B của công ty trong ngày. Theo giả thiết X và Y là các đại lượng ngẫu nhiên độc lập, có phân bố Poisson với tham số $\lambda_A = 4$ và $\lambda_B = 5$ tương ứng. Do hệ quả 3, $X + Y$ cũng có phân bố Poisson với tham số

$$E(X + Y) = \lambda = \lambda_A + \lambda_B = 4 + 5 = 9.$$

Điều đó cũng có nghiã rằng trung bình trong ngày cả hai cửa hàng có 9 khách. Xác suất để số khách đến đặt mua hàng ở cả hai cửa hàng lớn hơn 10 bằng

$$P(X + Y > 10) = 1 - P(X + Y \leq 10) = 1 - 0,706 = 0,294.$$

(Xác suất $P(X + Y \leq 10)$ có thể tính bằng cách tra bảng phân bố Poisson với tham số $\lambda = 9$ hoặc tính trực tiếp $\sum_{k=0}^{10} e^{-9}\frac{9^k}{k!}$)

Ví dụ 3.5.4

Tuổi thọ của một linh kiện điện tử là đại lượng ngẫu nhiên có phân bố mũ với tham số λ. Người ta sử dụng một thiết bị chứa linh kiện đó,

nếu linh kiện trong thiết bị bị hỏng, người ta thay linh kiện đó bằng một linh kiện khác cùng loại. Gọi X là tuổi thọ của linh kiện thứ nhất và Y là tuổi thọ của linh kiện thay thế (giả thiết rằng X và Y độc lập với nhau). Hãy tìm hàm mật độ, kì vọng (giá trị trung bình) của tổng $X+Y$.

Giải: X và Y là các đại lượng ngẫu nhiên độc lập cùng phân bố, hàm mật độ của chúng là

$$f(u) = \lambda e^{-\lambda u}, \quad u > 0$$

Do X và Y là các đại lượng ngẫu nhiên có phân bố mũ, chúng chỉ nhận các giá trị dương, áp dụng công thức trên để tính hàm mật độ $r(z)$ của tổng $X+Y$

$$r(z) = \int_0^z f(z-u)f(u)\,du = \int_0^z \lambda e^{-\lambda(z-u)}\lambda e^{-\lambda u}\,du = \lambda^2 z e^{-\lambda z}$$

với $z > 0$ và $r(z) = 0$ nếu $z \leq 0$.

Trong chương II khi xét phân bố mũ, ta đã chỉ ra phân bố mũ với tham số λ có kì vọng bằng $\frac{1}{\lambda}$. Vậy giá trị trung bình của tổng $X+Y$ là

$$E(X+Y) = \frac{1}{\lambda} + \frac{1}{\lambda} = \frac{2}{\lambda}$$

Nhận xét sau đây rất có ích cho việc tìm các hàm mật độ của đại lượng ngẫu nhiên mà đại lượng ngẫu nhiên đó là hàm của hai (hoặc nhiều hơn) các đại lượng ngẫu nhiên độc lập.

Nhận xét 3.5.2 *1. Nếu X và Y là hai đại lượng ngẫu nhiên độc lập, khi đó hàm phân bố có điều kiện của X với điều kiện $Y = y$ trùng với hàm phân bố của X, (không phụ thuộc vào điều kiện $Y = y$)*

$$F(x/Y = y) = P(X < x/Y = y) = P(X < x) = F(x).$$

2. Tổng quát hơn, giả sử $\varphi(x, y)$ là một hàm hai biến bất kì, X và Y là hai đại lượng ngẫu nhiên độc lập. Khi đó hàm phân bố có điều kiện của $\varphi(X, Y)$ với điều kiện $Y = y$ trùng với hàm phân bố của $\varphi(X, y)$

$$P(\varphi(X,Y) < x/Y = y) = P(\varphi(X,y) < x).$$

Chẳng hạn hệ quả 3.5.2, để tính hàm mật độ của tổng hai đại lượng ngẫu nhiên độc lập X và Y, có thể suy ra từ nhận xét trên như sau:

Xét $Z = \varphi(X, Y) = X + Y$, hàm phân bố có điều kiện của Z với điều kiện $Y = y$ (kí hiệu $H(z/y)$), theo nhận xét trên bằng hàm phân bố của $\varphi(X, y) = X + y$:

$$H(z/y) = P(X + y < z) = F(z - y)$$

Đạo hàm hai vế theo z để xác định hàm mật độ, ta được mật độ có điều kiện của Z với điều kiện $Y = y$ (kí hiệu $h(z/y)$)

$$h(z/y) = f(z - y)$$

Áp dụng công thức "xác suất đầy đủ mở rộng" trong mục 3.3 để tính hàm mật độ của Z (kí hiệu $r(z)$), ta được

$$r(z) = \int_{-\infty}^{\infty} h(z/y)g(y)\,dy = \int_{-\infty}^{\infty} f(z - y)g(y)\,dy.$$

Đây chính là công thức xác định hàm mật độ của tổng hai đại lượng ngẫu nhiên độc lập nêu trong hệ quả 3.5.2.

Hoàn toàn tương tự ta có thể thiết lập được các hàm mật độ của XY và X/Y, nếu X, Y độc lập nhau. Bạn đọc tự chứng minh các kết quả sau:

a. Hàm mật độ của XY bằng

$$s(z) = \int_{-\infty}^{\infty} \frac{1}{|y|} f(\frac{z}{y})g(y)\,dy.$$

b. Hàm mật độ của $\frac{X}{Y}$ bằng

$$t(z) = \int_{-\infty}^{\infty} |y| f(zy))g(y)\,dy.$$

Chẳng hạn ta phác qua cách dẫn dắt đến kết quả a. để tính hàm mật độ của XY. Trước tiên ta tìm hàm phân bố có điều kiện của XY với điều kiện $Y = y$ (xét hai trường hợp $y > 0$ và $y < 0$). Theo nhận xét 3.5.1 hàm

phân bố có điều kiện đó bằng hàm phân bố của $y \cdot X$ (không điều kiện). Áp dụng ví dụ 2.4.3. chương II, hàm mật độ của $y \cdot X$ bằng $\frac{1}{|y|} f(\frac{z}{y})$ suy ra hàm mật độ của XY

$$s(z) = \int_{-\infty}^{\infty} \frac{1}{|y|} f(\frac{z}{y}) g(y)\, dy.$$

Ví dụ 3.5.5

Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập có cùng phân bố đều trên $(-1, 1)$. Hãy tính hàm mật độ của $X + Y$.

Giải: Hàm mật độ của X và Y

$$f(x) = \begin{cases} \frac{1}{2} & \text{nếu } -1 < x < 1 \\ 0 & \text{nếu } x \notin (-1, 1). \end{cases}$$

Áp dụng công thức tính hàm mật độ của tổng hai đại lượng ngẫu nhiên độc lập, ta có mật độ của $X + Y$ bằng

$$h(x) = \begin{cases} \frac{x+2}{4} & \text{nếu } -2 < x \leq 0 \\ \frac{2-x}{4} & \text{nếu } 0 < x \leq 2 \\ 0 & \text{nếu } x \notin (-2, 2). \end{cases}$$

Hàm phân bố có mật độ $h(x)$ được gọi là *phân bố Simpson*.

Nếu X, Y, Z là ba đại lượng ngẫu nhiên độc lập có cùng phân bố đều trên $(-1, 1)$. Khi đó hoàn toàn tương tự, hàm mật độ của $X + Y + Z$ bằng

$$r(x) = \begin{cases} \frac{(3-|x|)^2}{16} & \text{nếu } 1 \leq |x| < 3 \\ \frac{3-x^2}{8} & \text{nếu } |x| < 1 \\ 0 & \text{nếu } |x| \geq 3. \end{cases}$$

Ví dụ 3.5.6

Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập có cùng phân bố đều trên (a, b) (để đơn giản ta giả thiết a, b là các số dương $0 < a < b$). Hãy tính hàm mật độ của đại lượng ngẫu nhiên tích XY.

Giải: Do mật độ của phân bố đều bằng 0 tại các điểm không thuộc (a,b), áp dụng công thức trên, hàm mật độ của XY bằng

$$s(z) = \frac{1}{b-a}\int_a^b \frac{1}{y} f\left(\frac{z}{y}\right)\, dy$$

nếu $a^2 < z < b^2$ và bằng 0 trong trường hợp ngược lại.

Để tính tích phân trên ta chú ý rằng khi biến tích phân y biến thiên từ a đến b, $\frac{z}{y}$ biến thiên từ $\frac{z}{b}$ đến $\frac{z}{a}$ do vậy hàm mật độ $f\left(\frac{z}{y}\right) = \frac{1}{b-a}$ khi $y \in \left(\frac{z}{b}, \frac{z}{a}\right) \cap (a,b)$. Mặt khác giao của hai khoảng này phụ thuộc vào giá trị của z, cụ thể

* Nếu $a^2 < z < ab$ khi đó $\left(\frac{z}{b}, \frac{z}{a}\right) \cap (a,b) = \left(a, \frac{z}{a}\right)$

* Nếu $ab < z < b^2$ khi đó $\left(\frac{z}{b}, \frac{z}{a}\right) \cap (a,b) = \left(\frac{z}{b}, b\right)$

Suy ra

$$s(z) = \begin{cases} \frac{1}{(b-a)^2}\displaystyle\int_a^{\frac{z}{a}} \frac{1}{y}dy = \frac{1}{(b-a)^2}\ln\frac{z}{a^2} & \text{nếu} \quad a^2 < z < ab \\ \frac{1}{(b-a)^2}\displaystyle\int_{\frac{z}{b}}^b \frac{1}{y}dy = \frac{1}{(b-a)^2}\ln\frac{b^2}{z} & \text{nếu} \quad ab < z < b^2 \\ 0 & \text{nếu} \quad z \notin (a^2, b^2). \end{cases}$$

Cuối cùng ta xét đến một ứng dụng khác của hệ quả 3.5.2 vào lớp các hàm phân bố chuẩn được phát biểu trong định lí sau

Định lí 3.5.1 *Nếu X và Y là hai đại lượng ngẫu nhiên độc lập có phân bố chuẩn với cùng phương sai*

$$X \in N(m_1, \sigma^2), \quad Y \in N(m_2, \sigma^2).$$

Khi đó tổng $X+Y$ cũng là đại lượng ngẫu nhiên có phân bố chuẩn

$$X + Y \in N(m_1 + m_2, 2\sigma^2).$$

Chứng minh: Áp dụng công thức tính hàm mật độ của tổng hai đại lượng ngẫu nhiên độc lập

$$r(z) = \int_{-\infty}^{\infty} f(u)g(z-u)\, du,$$

trong đó

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m_1)^2}{2\sigma^2}}$$

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m_2)^2}{2\sigma^2}}.$$

Ta có mật độ của $Z = X + Y$ bằng

$$\begin{aligned} r(z) &= \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(u-m_1)^2}{2\sigma^2} - \frac{(z-u-m_2)^2}{2\sigma^2}} du \\ &= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{u-m_1}{\sigma} + \frac{z-u-m_2}{\sigma}\right)^2 + \frac{(u-m_1)(z-u-m_2)}{\sigma^2}} du \\ &= \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(z-(m_1+m_2))^2} \int_{-\infty}^{\infty} e^{\frac{(u-m_1)(z-u-m_2)}{\sigma^2}} du. \end{aligned}$$

Biến đổi

$$\frac{(u-m_1)(z-u-m_2)}{\sigma^2} = -\left(\frac{u}{\sigma} - \frac{z+m_1-m_2}{2\sigma}\right)^2 + \left(\frac{z-m_1-m_2}{2\sigma}\right)^2$$

và sử dụng phép biến đổi quen thuộc

$$\int_{-\infty}^{\infty} e^{\frac{(u-m_1)(z-u-m_2)}{\sigma^2}} du = \sqrt{\pi}\sigma e^{\frac{1}{4\sigma^2}(z-(m_1+m_2))^2}.$$

Vậy mật độ của $Z = X + Y$ bằng

$$\begin{aligned} r(z) &= \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(z-(m_1+m_2))^2} \sqrt{\pi}\sigma e^{\frac{1}{4\sigma^2}(z-(m_1+m_2))^2} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{2}\sigma} e^{-\frac{(z-(m_1+m_2))^2}{2(\sqrt{2}\sigma)^2}}. \end{aligned}$$

Đây là hàm mật độ của phân bố chuẩn với kì vọng

$$E(X+Y) = m_1 + m_2 \quad \text{và phương sai} \quad D(X+Y) = 2\sigma^2, \text{ đ.p.c.m.}$$

Tổng quát hơn, người ta cũng chứng minh được kết quả sau

Nếu X và Y là hai đại lượng ngẫu nhiên độc lập có phân bố chuẩn (không nhất thiết phương sai cuả chúng phải bằng nhau)

$$X \in N(m_1, \sigma_1^2), \quad Y \in N(m_2, \sigma_2^2).$$

Khi đó tổng $X + Y$ cũng là đại lượng ngẫu nhiên có phân bố chuẩn

$$X + Y \in N(m_1 + m_2, \sigma_1^2 + \sigma_2^2).$$

Nhận xét 3.5.3 *Sử dụng công cụ sâu hơn của lí thuyết xác suất (hàm đặc trưng), người ta chứng minh được điều ngược lại cũng đúng (xem tài liệu tham khảo 1. hoặc 5.)*

Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập và tổng $X+Y$ có phân bố chuẩn, khi đó cả X, cả Y đều là các đại lượng ngẫu nhiên phân bố chuẩn.

3.6 Kì vọng có điều kiện

Giả sử A là biến cố có xác suất $P(A) > 0$ và X là đại lượng ngẫu nhiên tùy ý. Tương tự như định nghĩa kì vọng của đại lượng ngẫu nhiên trong chương II, ta có định nghĩa sau

Định nghĩa 3.6.1 *Nếu X là đại lượng ngẫu nhiên rời rạc chỉ nhận các giá trị $x_i, i = 1, 2, ...,$ khi đó*

$$E(X/A) = \sum_i x_i P(X = x_i/A)$$

được gọi là kì vọng có điều kiện của X với điều kiện biến cố A xảy ra.

Trường hợp X là đại lượng ngẫu nhiên liên tục với $f(x/A)$ là hàm mật độ có điều kiện, khi đó

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

được gọi là kì vọng có điều kiện của X với điều kiện biến cố A xảy ra.

Ta có định lí sau

Định lí 3.6.1 *Giả sử $A_i, i = 1, 2, ...$ là một hệ đầy đủ các biến cố. Khi đó*

$$E(X) = \sum_i E(X/A_i)P(A_i)$$

Chứng minh. Định lí được chứng minh dựa trên công thức xác suất đầy đủ. Do cách chứng minh hoàn toàn tương tự, ta chỉ hạn chế trong trường hợp X là đại lượng ngẫu nhiên liên tục với $f(x), f(x/A_i)$ là các hàm mật độ cũng như hàm mật độ có điều kiện.

Theo công thức xác suất đầy đủ

$$F(x) = P(X < x) = \sum_i P(X < x/A_i)P(A_i) = \sum_i F(x/A_i)P(A_i)$$

Đạo hàm cả hai vế theo x

$$f(x) = \sum_i f(x/A_i)P(A_i).$$

Vậy kì vọng của X là

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{-\infty}^{\infty} \sum_i xf(x/A_i)P(A_i)dx =$$

$$= \sum_i \left(\int_{-\infty}^{\infty} xf(x/A_i)dx \right) P(A_i) = \sum_i E(X/A_i)P(A_i).$$

Ta trở lại ví dụ 3.3.2 trong chương này

Ví dụ 3.6.1

Số hạt giống hỏng trong 1 bao hạt giống là đại lượng ngẫu nhiên có phân bố Poisson với tham số λ. Trước khi gieo trồng người ta kiểm tra lại để loại ra các hạt giống bị hỏng ra khỏi bao hạt giống đó. Giả sử xác suất để người kiểm tra phát hiện đúng một hạt giống hỏng là p. Gọi X là số hạt giống hỏng còn lại trong bao sau khi đã được kiểm tra. Tính trung bình số hạt giống hỏng còn lại sau khi đã được kiểm tra trước khi gieo trồng.

Giải: Duy trì kí hiệu như trong ví dụ 3.3.2, gọi Y là số hạt giống hỏng có trong bao trước khi kiểm tra. Theo giả thiết Y có phân bố Poisson, suy ra $A_i = \{Y = i\}, \quad i = 0, 1, 2, ...$ là một hệ đầy đủ các biến cố. Hàm phân bố của X với điều kiện biến cố $\{Y = i\}$ xảy ra, như đã chứng minh trong ví dụ 3.3.2, là phân bố nhị thức với tham số $q = 1-p$ (xác suất $P(X = n/Y = i)$ là xác suất để người kiểm tra chỉ phát hiện đúng $i-n$ hạt giống hỏng với điều kiện trong bao có i hạt giống hỏng, theo công thức Bernoulli $P(X = n/Y = i) = C_i^n p^{i-n} q^n$. Suy ra

$$E(X/Y = i) = iq.$$

Áp dụng định lí 3.6.1 số trung bình các hạt giống hỏng còn lại sau kiểm tra

$$\begin{aligned} E(X) &= \sum_{i=0}^{\infty} E(X/Y = i)P(Y = i) \\ &= \sum_{i=0}^{\infty} iqe^{-\lambda}\frac{\lambda^i}{i!} = q\sum_{i=0}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!}. \end{aligned}$$

Mặt khác $\sum_{i=0}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!}$ là kì vọng của Y nên

$$\sum_{i=0}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!} = E(Y) = \lambda.$$

Vậy số trung bình các hạt giống hỏng còn lại sau kiểm tra là

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

Nhận xét: Kì vọng có điều kiện $E(X/A_i)$ được coi như một đại lượng ngẫu nhiên mà nó nhận giá trị $E(X/A_i)$ khi và chỉ khi A_i xảy ra. Với quan niệm này, $\sum_i E(X/A_i)P(A_i)$ trong định lí 3.6.1 là kì vọng của $E(X/A_i)$. Mặt khác gọi Y là đại lượng ngẫu nhiên mà giá trị của Y phụ thuộc vào biến cố A_i: $Y = i$ khi và chỉ khi A_i xảy ra. Khi đó thay cho $E(X/A_i)$, ta có thể viết $E(X/Y)$. Suy ra định lí 3.6.1 có thể diễn đạt dưới dạng rất hay được sử dụng trong lí thuyết xác suất:

$$E(X) = E\left(E(X/Y)\right).$$

Bây giờ ta xét X và Y là hai đại lượng ngẫu nhiên liên tục với $f(x)$ và $g(x)$ là các hàm mật độ của chúng. Gọi $f(x/y)$ là hàm mật độ có điều kiện như đã định nghĩa trong mục 2. Cũng như khái niệm kì vọng ta có định nghĩa sau

Định nghĩa 3.6.2 *Kì vọng của X với điều kiện $Y = y$, (đại lượng ngẫu nhiên Y nhận giá trị y) được kí hiệu $E(X/Y = y)$ là tích phân*

$$E(X/Y = y) = \int_{-\infty}^{\infty} xf(x/y)\,dx,$$

nếu tích phân tồn tại và hội tụ tuyệt đối.

Nhận xét rằng $E(X/Y = y)$ là một số thực và khi Y nhận các giá trị y khác nhau, kì vọng có điều kiện $E(X/Y = y)$ là hàm của biến y. Hàm đó được gọi là hàm *hồi quy* của X với điều kiện $Y = y$.

Khi tiến hành phép thử ngẫu nhiên, biến cố ngẫu nhiên cơ bản nào đó xảy ra. Tương ứng với biến cố ngẫu nhiên cơ bản đó là giá trị y của đại lượng ngẫu nhiên Y và ứng với giá trị y đó là kì vọng có điều kiện $E(X/Y = y)$. Như vậy có thể coi kì vọng có điều kiện như một đại lượng ngẫu nhiên và thay cho điều kiện $Y = y$ ta kí hiệu $E(X/Y)$. Điều này chẳng khác gì việc coi kì vọng có điều kiện như một hàm $h(Y)$ của đại lượng ngẫu nhiên Y, trong đó

$$h(y) = E(X/Y = y).$$

Đại lượng ngẫu nhiên $E(X/Y)$ cũng được gọi là hàm *hồi quy* của X đối với Y. Định lí sau cho ta mối quan hệ giữa kì vọng của X và kì vọng của đại lượng ngẫu nhiên $h(Y) = E(X/Y)$.

Định lí 3.6.2 *Giả thiết rằng X và Y là hai đại lượng ngẫu nhiên liên tục, tồn tại kì vọng có điều kiện của X đối với Y, khi đó*

$$E(X) = E(E(X/Y)).$$

Chứng minh. Áp dụng định lí 2.5.3 chương II để tính kì vọng đó với hàm $h(y) = E(X/Y = y)$ của đại lượng ngẫu nhiên Y

$$\begin{aligned} E(h(Y)) &= \int_{-\infty}^{\infty} h(y)g(y)\,dy = \int_{-\infty}^{\infty} E(X/Y = y)g(y)\,dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} xf(x/y)\,dx \right) g(y)\,dy = \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f(x/y)g(y)dy \right) dx. \end{aligned}$$

Mặt khác do $f(x) = \int_{-\infty}^{\infty} f(x/y)g(y)\,dy$ nên

$$E(h(Y)) = E(E(X/Y)) = \int_{-\infty}^{\infty} xf(x)\,dx = E(X), \quad \text{đ.p.c.m.}$$

3.7 Tương quan của hai đại lượng ngẫu nhiên

Khi khảo sát hai đại lượng ngẫu nhiên, người ta muốn biết chúng độc lập với nhau hay không. Hơn nữa chúng ta cũng muốn biết về mối quan hệ giữa các đại lượng ngẫu nhiên đó. Để đo mức độ phụ thuộc của chúng, ta dẫn vào khái niệm sau

Định nghĩa 3.7.1 *Nếu X và Y là hai đại lượng ngẫu nhiên tồn tại kì vọng $E(X)$ và $E(Y)$, khi đó*

$$cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

được gọi là covarian (hay còn gọi là mô men tương quan) của X và Y.

Hiển nhiên nếu X và Y độc lập, khi đó

$$\begin{aligned} cov(X,Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(X - E(X)) \cdot E(Y - E(Y)) = 0. \end{aligned}$$

Trường hợp $X = Y$, khi đó covarian $cov(X,X) = D(X)$.

Mô men tương quan cuả hai đại lượng ngẫu nhiên có các tính chất sau

i) $cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$

ii) $cov(\alpha X, Y) = cov(X, \alpha Y) = \alpha \cdot cov(X,Y)$

iii) Nếu $h(x,y)$ là hàm mật độ chung của (X,Y) khi đó

$$cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - E(X))(y - E(Y)h(x,y)dxdy$$

và cũng bằng

$$cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyh(x,y)dxdy - E(X)E(Y).$$

iv) Trường hợp X là đại lượng ngẫu nhiên rời rạc

$$cov(X,Y) = \sum_i \sum_k ((x_i - E(X))((y_k - E(Y))r_{ik} =$$

$$= \sum_i \sum_k x_i y_k r_{ik} - E(X)E(Y)$$

trong đó $r_{ik} = P(X = x_i, Y = y_k)$.

v) Kí hiệu $\sigma_x = \sqrt{D(X)}$ và $\sigma_y = \sqrt{D(Y)}$ là các độ lệch tiêu chuẩn của X và Y. Khi đó

$$|cov(X,Y)| \leq \sigma_x \sigma_y.$$

Thật vậy xét

$$E[(Y - tX)^2] = E(Y^2 - 2tXY + t^2Y^2) =$$

$$= E(Y^2) - 2E(XY)t + E(Y^2)t^2 \geq 0$$

với mọi t. Đây là tam thức bậc hai không âm với mọi t, suy ra

$$[E(XY)]^2 \leq E(X^2)E(Y^2) \text{hay } |E(XY)| \leq \sqrt{E(X^2)}\sqrt{E(Y^2)}.$$

Áp dụng bất đẳng thức trên với $X - E(X)$ và $Y - E(Y)$ thay cho X và Y

$$|cov(X,Y)| = |E[(X - E(X))(Y - E(Y))]| \leq$$

$$\leq \sqrt{D(X)}\sqrt{D(X)} = \sigma_x \sigma_y.$$

Nhận xét rằng từ chứng minh trên suy ra

$$|cov(X,Y)| = \sigma_x \sigma_y$$

khi và chỉ khi tồn tại một số thực t_0 sao cho

$$Y - E(Y) = t_0(X - E(X)$$

hay nói cách khác Y là một hàm bậc nhất của X:

$$Y = aX + b.$$

Định nghĩa 3.7.2

$$\varrho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{D(X)}\sqrt{D(X)}}$$

được gọi là hệ số tương quan của X và Y.

Hiển nhiên hệ số tương quan có các tính chất

i) $-1 \leq \varrho(X,Y) \leq 1$. Dấu bằng xảy ra khi và chỉ khi $Y = aX + b$ (hoặc $X = aY + b$)

ii) Nếu X và Y độc lập, khi đó hệ số tương quan $\varrho(X,Y) = 0$

Hệ số tương quan đo mức độ phụ thuộc tuyến tính giữa Y và X. Nếu $|\varrho(X,Y)|$ xấp xỉ 1 khi đó các điểm ngẫu nhiên (X,Y) gần như tạo thành một đường thẳng trên mặt phẳng toạ độ. Khi $\varrho(X,Y) = 0$ ta nói X và Y *không tương quan*. Chú ý rằng nếu X và Y độc lập khi đó chúng không tương quan, ngược lại từ sự không tương quan của X và Y không suy ra chúng độc lập với nhau.

Ví dụ 3.7.1

Gieo 2 lần một xúc xắc đồng chất, đối xứng. Hãy tìm hệ số tương quan giữa kết quả lần gieo đầu và tổng số chấm xuất hiện ở 2 lần gieo.

Gọi X là số chấm xuất hiện ở lần gieo đầu và Y là số chấm xuất hiện ở lần gieo thứ hai. Hệ số tương quan của X và $X+Y$, theo định nghĩa bằng

$$\begin{aligned}\varrho(X, X+Y) &= \frac{E[(X - E(X))(X + Y - E(X+Y))]}{\sqrt{D(X)}\sqrt{D(X+Y)}} \\ &= \frac{E(X^2 + XY) - E(X)E(X+Y)}{\sqrt{D(X)}\sqrt{2D(X)}}.\end{aligned}$$

Do X và Y độc lập, cùng phân bố, suy ra

$$E(X^2 + XY) - E(X)E(X+Y) = E(X^2) - (EX)^2 = D(X)$$

Mặt khác $\sqrt{D(X)}\sqrt{2D(X)} = D(X)\sqrt{2}$. Vậy hệ số tương quan

$$\varrho(X, X+Y) = \frac{1}{\sqrt{2}}.$$

Ví dụ 3.7.2

φ là đại lượng ngẫu nhiên phân bố đều trên $(0, 2\pi)$. Chứng tỏ rằng

$$X = \cos\varphi, Y = \sin\varphi$$

không tương quan và chúng cũng không độc lập.

Giải: X và Y là hai đại lượng ngẫu nhiên không độc lập. Thật vậy

$$P(X < -\frac{\sqrt{2}}{2}, Y < -\frac{\sqrt{2}}{2}) = P(\cos\varphi < -\frac{\sqrt{2}}{2}, \sin\varphi < -\frac{\sqrt{2}}{2}) = 0$$

trong khi tích các xác suất

$$P(X < -\frac{\sqrt{2}}{2})P(Y < -\frac{\sqrt{2}}{2}) =$$

$$= P(\cos\varphi < -\frac{\sqrt{2}}{2})P(\sin\varphi < -\frac{\sqrt{2}}{2}) = \frac{\pi^2}{4^2}.$$

Chúng không thoả mãn định nghĩa về sự độc lập của hai đại lượng ngẫu nhiên

$$P(X < -\frac{\sqrt{2}}{2}, Y < -\frac{\sqrt{2}}{2}) = P(X < -\frac{\sqrt{2}}{2})P(Y < -\frac{\sqrt{2}}{2}).$$

Hiển nhiên

$$E(X) = \int_0^{2\pi} \frac{1}{2\pi} \sin x \, dx = 0$$

và tương tự $E(Y) = 0$. Do vậy theo tính chất ii) dễ dàng tính được mô men tương quan cuả X và Y

$$cov(X, Y) = E(XY) - E(X)E(Y) = \int_0^{2\pi} \cos\varphi \sin\varphi d\varphi = 0.$$

Định nghĩa 3.7.3 *Kí hiệu $c = cov(X, Y)$ là mô men tương quan của X và Y. Khi đó ma trận*

$$C = \begin{pmatrix} D(X) & c \\ c & D(Y) \end{pmatrix}$$

được gọi là ma trận covarian (ma trận tương quan) của X và Y.

Duy trì các kí hiệu σ_x, σ_y là các độ lệch tiêu chuẩn của X và Y, ϱ là hệ số tương quan của X và Y. Từ định nghĩa hệ số tương quan suy ra $c = \varrho\sigma_x\sigma_y$. Khi đó ma trận covarian có thể viết dưới dạng

$$C = \begin{pmatrix} \sigma_x^2 & \varrho\sigma_x\sigma_y \\ \varrho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

Do $|\varrho| \leq 1$ nên

$$det(C) = \begin{vmatrix} \sigma_x^2 & \varrho\sigma_x\sigma_y \\ \varrho\sigma_x\sigma_y & \sigma_y^2 \end{vmatrix} = (1 - \varrho^2)\sigma_x^2\sigma_y^2 \geq 0$$

Phương sai của đại lượng ngẫu nhiên cho ta biết độ lệch giữa đại lượng ngẫu nhiên và giá trị trung bình của đại lượng ngẫu nhiên đó. Ma trận các hệ số tương quan cũng đóng vai trò tương tự như phương sai khi xét độ dao động của véc tơ ngẫu nhiên.

Giả sử d là đường thẳng đi qua (EX, EY) (giá trị trung bình của véc tơ ngẫu nhiên (X, Y)) và $\overrightarrow{n}(\alpha, \beta)$ là véc tơ đơn vị chỉ phương của d. Gọi

$$Z = \alpha(X - EX) + \beta(Y - EY)$$

là hình chiếu vuông góc của $(X - EX, Y - EY)$ lên đường thẳng d. Phương sai của Z sẽ được tính thông qua ma trận covarian C như sau

$$D(Z) = \alpha^2 E(X - EX)^2 + \beta^2(Y - EY)^2 + 2\alpha\beta E(X - EX)E(Y - EY) =$$

$$= \alpha^2\sigma_x^2 + \beta^2\sigma_y^2 + 2\alpha\beta\varrho\sigma_x\sigma_y.$$

Nhận xét rằng phương sai của Z là dạng toàn phương với ma trận covarian C là ma trận của dạng toàn phương đó. Do $det(C) \geq 0$, nói chung C là ma trận bán xác định dương. Nếu X và Y độc lập tuyến tính ($|\varrho| < 1$), khi đó C là ma trận xác định dương thực sự.

Nhận xét 3.7.1 *Sử dụng các phép toán đối với ma trận, ta có thể mở rộng khái niệm ma trận covarian cho nhiều đại lượng ngẫu nhiên*

$$X_i, E(X_i) = m_i,\ cov(X_i, X_j) = \sigma_{ij}, \quad i, j = 1, 2, ..., n$$

Khi đó ma trận covarian của $(X_1, X_2, ..., X_n)$ *là*

$$C(X) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ & \cdots & \cdots & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

Giả sử $a_i, i = 1, 2, ...n$ *là các số thực bất kì. Khi đó*

$$D(\sum_{i=1}^{n} a_i X_i) = E\left(\sum_{i=1}^{n} a_i (X_i - m_i)\right)^2 =$$

$$= \sum_i \sum_j a_i a_j \sigma_{ij}$$

Tương tự

$$cov(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{n} b_i X_i) = \sum_i \sum_j a_i b_j \sigma_{ij}.$$

Kí hiệu ***A,B,X,M*** *là các véc tơ cột với các thành phần* a_i, b_i, X_i, m_i *tương ứng.* ***C(X)*** *là ma trận covarian của* ***X****. Từ các đẳng thức trên suy ra*

$$E(A^T X) = A^T E(X) = A^T M$$

$$D(A^T X) = A^T C(X) A$$
$$cov(A^T X, B^T X) = A^T C(X) B = B^T C(X) A.$$

Định nghĩa 3.7.4 *Cặp hai đại lượng ngẫu nhiên* (X, Y) *được gọi là có phân bố chuẩn hai chiều nếu hàm mật độ chung* $h(x, y)$ *có dạng*

$$h(x,y) = \frac{\sqrt{|A|}}{2\pi} e^{-\frac{1}{2}(a_{11}(x-m_1)^2 + 2a_{12}(x-m_1)(y-m_2) + a_{22}(y-m_2)^2)}$$

trong đó m_1, m_2 *là hai số thực tùy ý,* A *là ma trận đối xứng xác định dương.*

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$$

Trước hết ta chứng minh hàm $h(x,y)$ trong định nghĩa trên là hàm mật độ. Thật vậy gọi P là ma trận trực giao ($det(P)=1$)

$$P=\begin{pmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{pmatrix}$$

sao cho $P^{-1}AP$ có dạng chéo. Do A là ma trận đối xứng xác định dương, ma trận $P^{-1}AP$ có dạng

$$P^{-1}AP=\begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}$$

Véc tơ ngẫu nhiên (U,V) được xác định bởi hệ thức

$$X-m_1=p_{11}U+p_{12}V$$
$$Y-m_2=p_{21}U+p_{22}V$$

Khi đó hàm mật độ đồng thời của véc tơ ngẫu nhiên (U,V) theo hệ quả 3.4.2 bằng

$$g(u,v)=h\left(a_{11}u+a_{12}v,a_{21}u+a_{22}v\right)|det(P)|=$$

$$=\frac{1}{2\pi\sigma_1\sigma_2}e^{-\frac{1}{2}(\frac{1}{\sigma_1^2}u^2+\frac{1}{\sigma_2^2}v^2)}$$

Vì sau khi đổi biến, dạng toàn phương trong biểu thức của hàm $h(x,y)$ có dạng chính tắc

$$a_{11}(x-m_1)^2+2a_{12}(x-m_1)(y-m_2)+a_{22}(y-m_2)^2=\frac{1}{\sigma_1^2}u^2+\frac{1}{\sigma_2^2}v^2$$

còn $|A|$ bằng

$$det(A)=det(P^{-1}AP)=\frac{1}{\sigma_1^2\sigma_2^2}$$

Vậy $g(u,v)$ là tích của hai hàm mật độ có phân bố chuẩn

$$g(u,v)=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{u^2}{2\sigma_1^2}}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{v^2}{2\sigma_2^2}}$$

Suy ra $h(x,y)$ là hàm mật độ, hơn nữa U và V là hai đại lượng ngẫu nhiên độc lập, có phân bố chuẩn ($U \in N(0,\sigma_1^2), V \in N(0,\sigma_2^2)$).

(Lưu ý rằng do $g(u,v)$ là hàm mật độ nên tích phân của $g(u,v)$ trên $\mathbb{R}^2$ bằng 1 và do vậy tích phân của $h(x,y)$ trên $\mathbb{R}^2$ cũng bằng 1. Vậy $h(x,y)$ là hàm mật độ. Mặt khác từ định lí 3.5.1 chương này suy ra đại lượng ngẫu nhiên $X = p_{11}U + p_{12}V + m_1$ cũng có phân bố chuẩn).

Hiển nhiên ma trận covarian của (U,V) là

$$C(U,V) = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = P^T A^{-1} P$$

Suy ra ma trận covarian của (X,Y), theo nhận xét trên bằng

$$C(X,Y) = P \cdot C(U,V) \cdot P^T = PP^T A^{-1} PP^T = A^{-1}.$$

trong đó, ta nhớ lại rằng X, Y được cho dưới dạng như sau

$$X = p_{11}U + p_{12}V + m_1$$
$$Y = p_{21}U + p_{22}V + m_2$$

Chú ý rằng cách trình bày trên đây có thể mở rộng cho khái niệm phân bố chuẩn nhiều chiều. Trường hợp phân bố chuẩn hai chiều, hàm mật độ thường được cho dưới dạng như sau

$$h(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} e^{-\frac{1}{2(1-\varrho^2)}\left(\frac{(x-m_1)^2}{\sigma_1^2} - 2\varrho\frac{x-m_1}{\sigma_1}\frac{y-m_2}{\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right)}$$

Áp dụng các kết quả trên (hoặc tính trực tiếp) ta được

$$D(X) = \sigma_1^2, D(Y) = \sigma_2^2, \varrho(X,Y) = \varrho$$

trong đó ϱ là hệ số tương quan của X và Y. Nhận xét rằng nếu (X,Y) có phân bố chuẩn hai chiều, khi đó điều kiện cần và đủ để X và Y độc lập là $\varrho(X,Y) = 0$

Cuối cùng ta tính hàm mật độ của X với điều kiện $Y = y$

$$f(x/y) = \frac{h(x,y)}{g(y)}$$

Dễ dàng tính được

$$g(y) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}}$$

Vậy

$$f(x/y) = \frac{1}{\sigma_1\sqrt{2\pi(1-\varrho^2)}} e^{-\frac{1}{2\sigma_1^2(1-\varrho^2)}(x-m_1-\frac{\sigma_2}{\sigma_1}\varrho(y-m_2))^2}$$

Nói cách khác hàm mật độ của X với điều kiện $Y = y$ cũng là mật độ của phân bố chuẩn với kì vọng bằng

$$m_1 + \frac{\sigma_2}{\sigma_1}\varrho(y - m_2)$$

và phương sai $\sigma_1^2(1 - \varrho^2)$.

Suy ra hồi quy của X đối với Y

$$E(X/Y = y) = m_1 + \frac{\sigma_2}{\sigma_1}\varrho(y - m_2)$$

là hàm bậc nhất của y. Đây là tính chất đặc biệt của phân bố chuẩn hai chiều.

BÀI TẬP CHƯƠNG III

1. Giả sử hàm mật độ chung của 2 đại lượng ngẫu nhiên X và Y là

$$h(x,y) = \begin{cases} \frac{4}{5}(x+xy+y) & \text{nếu } 0<x<1, 0<y<1 \\ 0 & \text{với các trường hợp khác.} \end{cases}$$

   Hãy xác định hàm mật độ của X, của Y.

2. Giả sử hàm mật độ chung của 2 đại lượng ngẫu nhiên X và Y là

$$h(x,y) = \begin{cases} x+y & \text{nếu } 0<x<1, 0<y<1 \\ 0 & x \notin (0,1) \text{ hoặc } y \notin (0,1) \end{cases}$$

   Hãy xác định hệ số tương quan của X và Y.

3. Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập, cùng phân bố đều trên đoạn $[a,b]$. Gọi $\xi = max(X,Y)$ và $\eta = min(X,Y)$. Hãy xác định các hàm mật độ của ξ và η, hàm mật độ chung, và tính kì vọng, phương sai của $\frac{\xi+\eta}{2}$.

4. Hãy tìm các hàm mật độ của X và của Y, biết rằng phân bố đồng thời của chúng là phân bố chuẩn 2 chiều.

5. Chọn ngẫu nhiên 2 điểm M và N trên đoạn [0,1], 2 điểm M, N đó chia đoạn [0,1] thành 3 phần, gọi các độ dài của 3 đoạn thẳng đó tương ứng là các đại lượng ngẫu nhiên X_1, X_2 và X_3.

   (a) Hãy tìm các hàm mật độ của X_1, X_2 và X_3.

   (b) Hãy tính các kì vọng $E(X_1), E(X_2)$ và $E(X_3)$.

6. Gọi X là đại lượng ngẫu nhiên có $F(x)$ là hàm phân bố. Giả sử-hàm phân bố $F(x)$ đơn điệu tăng (nghiêm ngặt) và liên tục trên $\mathbb{R}$. Chứng minh rằng đại lượng ngẫu nhiên $Y = F(X)$ phân bố đều trên đoạn $[0,1]$.

7. Giả sử X và Y là 2 đại lượng ngẫu độc lập, có cùng phân bố mũ (với cùng tham số). Hãy tính xác suất

$$P(X - \frac{1}{X} < Y - \frac{1}{Y}).$$

8. Chứng minh rằng nếu X và Y là 2 đại lượng ngẫu nhiên liên tục, độc lập và có cùng phân bố, khi đó

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

9. Giả sử X là đại lượng ngẫu nhiên liên tục tồn tại kì vọng và gọi $F(x)$ là hàm phân bố của X. Chứng minh rằng

   (a) $E(X) = \displaystyle\int_0^\infty (1 - F(y))\, dy - \int_{-\infty}^0 F(y)\, dy.$

   (b) Nếu giả thiết tiếp X tồn tại phương sai. Chứng minh rằng

$$E(X^2) = 2\int_0^\infty y(1 - F(y) + F(-y))\, dy.$$

10. Gọi $X_1, X_2, X_3, \cdots, X_n$ là các đại lượng ngẫu nhiên độc lập cùng phân bố. Chứng tỏ rằng

$$E\left(\frac{X_1 + X_2 + \cdots + X_k}{X_1 + X_2 + \cdots + X_n}\right) = \frac{k}{n} \quad (1 \le k \le n)$$

11. Gọi X và Y là hai đại lượng ngẫu nhiên độc lập, có phân bố Poisson với các tham số λ và μ tương ứng. Tìm hệ số tương quan của X và $X + Y$.

12. Gọi X và Y là hai đại lượng ngẫu nhiên độc lập, có cùng phân bố mũ với tham số λ. Kí hiệu $g(y/x)$ là hàm mật độ của $X + Y$ với điều kiện $X = x$ và $f(x/y)$ là hàm mật độ của X với điều kiện $X + Y = y$. Hãy xác định các hàm mật độ có điều kiện $g(y/x)$ và $f(x/y)$.

13. Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập có cùng phân bố mũ với tham số λ. Hãy tìm hàm mật độ của $|X - Y|$.

14. Giả sử $X = (X_1, X_2)$ và $Y = (Y_1, Y_2)$ là hai điểm chọn ngẫu nhiên (theo phân bố đều) độc lập nhau trên đường tròn đơn vị:

$$x^2 + y^2 = 1.$$

Hãy tìm hàm mật độ của đại lượng ngẫu nhiên

$$Z = \begin{vmatrix} X_1 & X_2 \\ Y_1 & Y_2 \end{vmatrix}$$

15. Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập có cùng phân bố chuẩn $N(0,1)$.

(a) Hãy tìm hàm mật độ của $Z = |X| \cdot signY$.

(b) Chứng minh rằng $X + Y$ và $X - Y$ cũng độc lập.

16. Một cửa hàng bảo dưỡng xe máy có thể nhận bảo dưỡng đồng thời 9 xe máy. Giả thiết rằng bảo dưỡng cho mỗi xe máy kéo dài trong 20 phút. Một người khách vào cửa hàng và thấy cả 9 xe máy đang được bảo dưỡng, ngoài ra còn 3 khách đang ngồi chờ. Hỏi người đó còn phải chờ bao lâu để được phục vụ. (Giả thiết rằng thời điểm kết thúc việc bảo dưỡng cho mỗi xe độc lập và phân bố đều trong khoảng (0, 20)).

17. Giả sử X và Y là các đại lượng ngẫu nhiên có phân bố chuẩn thuộc lớp $N(0,1)$, phân bố đồng thời của chúng cũng là phân bố chuẩn với r là hệ số tương quan. Hãy xác định xác suất

$$P(X \geq 0, Y \geq 0).$$

18. Giả sử X là đại lượng ngẫu nhiên có phân bố chuẩn $X \in N(m, \sigma^2)$. Hãy tính các mô men cấp 1, cấp 2, cấp 3, cấp 4 của X.

19. Cho véc tơ ngẫu nhiên (X, Y) phân bố đều trong tam giác với các đỉnh $O(0,0), A(0,4), B(4,0)$.

(a) Hãy xác định hàm mật độ chung của véc tơ ngẫu nhiên (X, Y).

(b) Xác định kì vọng và phương sai của X

(c) Tìm hàm mật độ của đại lượng ngẫu nhiên $X + Y$

(d) Xác định hàm mật độ của các đại lượng ngẫu nhiên

$$max(X, Y) \quad \text{và} \quad min(X, Y).$$

20. Cho véc tơ ngẫu nhiên (X, Y), biết hàm mật độ chung của (X, Y) có dạng: $h(x, y) = \begin{cases} a(x - xy + y) & \text{nếu } 0 < x < 1, 0 < y < 1 \\ 0 & \text{nếu trái lại.} \end{cases}$

(a) Hãy xác định a để $h(x, y)$ là hàm mật độ chung.

(b) Xác định hàm mật độ của đại lượng ngẫu nhiên X.

(c) Xác định kì vọng và phương sai của X.

(d) Tính xác suất $P(X < 3Y)$.

21. Một xạ thủ bắn súng mang theo người 5 viên đạn, xác suất bắn trúng đích của xạ thủ đó là $p = 0,6$. Xạ thủ bắn liên tục vào đích cho đến khi có 3 viên trúng đích thì dừng (chú ý rằng nếu bắn hết cả 5 viên mà số lần trúng đích nhỏ hơn 3 thì cũng phải dừng vì hết đạn). Lập bảng phân bố số đạn chi phí (đã bắn) và tính số đạn trung bình mà xạ thủ đãsử dụng.

22. Giả sử X và Y là 2 đại lượng ngẫu nhiên độc lập, phân bố đều trong khoảng $[-\frac{1}{2}, \frac{1}{2}]$.

(a) Hãy tính $E(X^2 + Y^2)$ và $D(X^2 + Y^2)$.

(b) Hãy tìm hàm mật độ của $X + Y$.

23. Cho biết chiều cao của thanh niên là đại lượng ngẫu nhiên tuân theo quy luật chuẩn với chiều cao trung bình (kì vọng) $m = 1,6$ (mét) và độ lệch tiêu chuẩn $\sigma = 0,07$. Tìm xác suất để khi chọn ngẫu nhiên một thanh niên:

(a) Người đó cao trên 1,7 mét.

(b) Người đó cao từ 1,6 đến 1,7 mét.

24. Cho véc tơ ngẫu nhiên (X, Y) có hàm mật độ chung :

$$h(x,y) = \frac{1}{4\pi} e^{-\frac{x^2+y^2-4x+4}{4}}.$$

(a) Hãy xác định kì vọng và phương sai của X.

(b) Hãy tìm hàm mật độ của $X + Y$.

ĐÁP SỐ VÀ HƯỚNG DẪN

1. $f(x) = g(x) = \begin{cases} \frac{2}{5}(3x+1) & \text{nếu } 0 < x < 1 \\ 0 & \text{nếu } x \notin (0,1). \end{cases}$

2. $-\frac{1}{11}$.

3. $\frac{2}{b-a}\left(\frac{x-a}{b-a}\right), \quad \frac{2}{b-a}\left(\frac{b-x}{b-a}\right).$
$E\left(\frac{\xi+\eta}{2}\right) = \frac{a+b}{2}, \quad D\left(\frac{\xi+\eta}{2}\right) = \frac{(b-a)^2}{24}.$

4. Giả sử hàm mật độ đồng thời của X và Y có dạng

$$h(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} e^{-\frac{1}{2(1-\varrho^2)}\left(\frac{(x-m_1)^2}{\sigma_1^2} - 2\varrho\frac{x-m_1}{\sigma_1}\frac{y-m_2}{\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right)}$$

Khi đó

$$X \in N(m_1, \sigma_1^2), Y \in N(m_2, \sigma_2^2), \varrho(X,Y) = \varrho.$$

5. (a) Các hàm mật độ của X_1, X_2 và X_3 bằng nhau và bằng $2(1-x)$ nếu $0 < x < 1$.

(b) $E(X_1) = E(X_2) = E(X_3) = \frac{1}{3}$.

6. Hãy tính xác suất

$$P(Y < y) = P(X < F^{-1}(y)) = F(F^{-1}(y)) = y$$

với mọi $0 < y < 1$.

7. $\frac{1}{2}$.

8. $X - Y$ là đại lượng ngẫu nhiên có phân bố đối xứng qua $x = 0$. Suy ra
$$P(X < Y) = \frac{1}{2}.$$

9. Sử dụng tích phân từng phần và chứng minh
$$\lim_{x \to +\infty} x(1 - F(x)) = 0.$$

10. Chứng minh các đại lượng ngẫu nhiên
$$\frac{X_k}{X_1 + X_2 + \cdots + X_n} \quad \text{với mọi} \quad k = 1, 2, ..., n$$
cùng phân bố và sử dụng
$$\sum_{k=1}^{n} \frac{X_k}{X_1 + X_2 + \cdots + X_n} = \frac{X_1 + X_2 + \cdots + X_n}{X_1 + X_2 + \cdots + X_n} = 1.$$

11. $\sqrt{\frac{\lambda}{\lambda+\mu}}$.

12. Hàm mật độ có điều kiện $g(y/x) = \lambda e^{-\lambda(y-x)}$ với $0 < x < y$ và
$$f(x/y) = \frac{1}{y} \qquad \text{nếu} \qquad 0 < x < y.$$

13. Hàm mật độ của $|X - Y|$ cũng là mật độ của phân bố mũ

14.
$$Z = \begin{vmatrix} X_1 & X_2 \\ Y_1 & Y_2 \end{vmatrix}$$
$|Z|$ là diện tích hình bình hành căng bởi 2 véc tơ OM và ON hay $Z = \sin\alpha$. Chú ý rằng Z có phân bố đối xứng qua 0. Vậy hàm mật độ của đại lượng ngẫu nhiên Z bằng
$$f(z) = \begin{cases} \frac{1}{\pi\sqrt{1-z^2}} & \text{nếu } -1 < z < 1 \\ 0 & \text{nếu } z \notin (-1, 1). \end{cases}$$

15. (a) Hàm mật độ của $Z = |X| \cdot signY$ thuộc lớp $N(0,1)$.

    (b) Để chứng minh $X+Y$ và $X-Y$ độc lập, hãy tính hàm mật độ đồng thời của $X+Y$ và $X-Y$.

16. Với giả thiết thời điểm kết thúc việc bảo dưỡng cho mỗi xe máy độc lập và phân bố đều, gọi thời gian chờ đợi của người thứ nhất, người thứ hai, người thứ ba là X_1^*, X_2^*, X_3^* tương ứng. Khi đó thời gian chờ đợi của người khách mới đến là X_4^*.

17. $P(X \geq 0, Y \geq 0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin r$.

18. *Mô men cấp 1* của X

$$m_1 = E(X) = m$$

Mô men cấp 2 của X

$$m_2 = E(X^2) = \sigma^2 + m^2$$

Mô men cấp 3 của X

$$m_3 = E(X^3) = m^3 + 3m\sigma^2.$$

Mô men cấp 4 của X

$$m_4 = E(X^4) = m^4 + 6m^2\sigma^2 + 3\sigma^4.$$

19. (a) Hàm mật độ chung của véc tơ ngẫu nhiên (X,Y):

$$h(x,y) = \begin{cases} \frac{1}{8} & \text{nếu } (x,y) \in D \\ 0 & \text{nếu } (x,y) \notin D. \end{cases}$$

Trong đó miền D là tam giác OAB và diện tích của miền D bằng 8.

(b) Do hàm mật độ của X

$$f(x) = \begin{cases} \frac{4-x}{8} & \text{nếu } 0 < x < 4 \\ 0 & \text{nếu } x \notin (0,4) \end{cases}$$

suy ra $E(X) = \frac{4}{3}, \quad D(X) = \frac{8}{9}$.

(c) Ta tính mật độ của $X + Y$

$$u(t) = \begin{cases} \frac{t}{8} & \text{nếu } 0 < t < 4 \\ 0 & \text{nếu } t \notin (0,4) \end{cases}$$

(d) Hàm mật độ của $max(X, Y)$:

$$M(t) = \begin{cases} \frac{t}{4} & \text{nếu } 0 < t < 2 \\ \frac{4-t}{4} & \text{nếu } 2 < t < 4 \\ 0 & \text{nếu } t \notin (0,4). \end{cases}$$

Hàm mật độ của $min(X, Y)$:

$$m(t) = \begin{cases} \frac{2-t}{2} & \text{nếu } 0 < x < 2 \\ 0 & \text{nếu } t \notin (0,2). \end{cases}$$

20. (a) $a = \frac{4}{3}$.

(b) Hàm mật độ của X

$$f(x) = \begin{cases} \frac{2(x+1)}{3} & \text{nếu } 0 < x < 1 \\ 0 & \text{nếu } x \notin (0,1). \end{cases}$$

(c) Kì vọng $E(X) = \frac{5}{9}$ và phương sai $D(X) = \frac{13}{162}$.

(d) $P(X < 3Y) = \frac{137}{162}$.

21. (a) Bảng phân bố củaX là:

X	3	4	5
$P(X = k)$	0,216	0,2592	0,5248

(b) $E(X) = 4,3088$.

22. (a) $E(X^2 + y^2) = \frac{1}{6} \quad D(X^2 + y^2) = \frac{1}{90}$

(b) Hàm mật độ của $X + Y$:

$$r(z) = \begin{cases} z + 1 & \text{nếu } -1 < z < 0 \\ 1 - z & \text{nếu } 0 < z < 1. \end{cases}$$

23. $P(X > 1,7) = 0,076564$ và $P(1,6 < X < 1,7) = 0,423436$.

24. (a) Kì vọng $EX = 2$, phương $DX = 2$.

    (b) Hàm mật độ của $X + Y$:

$$r(z) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(z-2)^2}{8}}.$$

Chương 4

Luật số lớn và định lí giới hạn trung tâm

4.1 Khái niệm về luật số lớn

Trong chương I chúng ta đã nói về sự ổn định của tần suất của biến cố ngẫu nhiên A. Cụ thể nếu ta tiến hành n lần phép thử ngẫu nhiên để quan sát biến cố A và gọi $\frac{k}{n}$ là tần suất xuất hiện của biến cố A. Khi đó với n tăng ra vô cùng, tần suất $\frac{k}{n}$ sẽ tiến dần tới xác suất của biến cố A. Sự ổn định của tần suất xung quanh xác suất của biến cố ngẫu nhiên như vậy được gọi là *luật số lớn*.

Chúng ta sẽ minh hoạ chi tiết hơn sự ổn định nêu trên của tần suất. Gọi X_i là đại lượng ngẫu nhiên chỉ nhận các giá trị 1 hoặc 0 tùy theo biến cố A xảy ra hoặc không xảy ra ở phép thử thứ i, $i = 1, 2, ..., n$. Hiển nhiên tần suất của biến cố A có thể biểu diễn theo các đại lượng ngẫu nhiên X_i:

$$\frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{k}{n}$$

Vậy luật số lớn khẳng định

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to P(A) \quad \text{khi} \quad n \to \infty$$

Một cách tổng quát, cho một dãy các đại lượng ngẫu nhiên $Y_i, \quad i = 1, 2, ...$ Người ta gọi các định lí khẳng định sự hội tụ tới hằng số C (theo một nghĩa

nào đó) trung bình cộng của n đại lượng ngẫu nhiên đầu tiên $Y_i, i = 1, 2, ..., n$ trong dãy đó:

$$\frac{Y_1 + Y_2 + \cdots + Y_n}{n} \to C \quad \text{khi} \quad n \to \infty$$

là *luật số lớn*. Như vậy sự ổn định của tần suất xung quanh xác suất nêu trên cũng là một trong số các định lí về luật số lớn.

4.2 Bất đẳng thức Máckốp và bất đẳng thức Trêbusép

Định lí 4.2.1 (Máckốp) *Cho X là đại lượng ngẫu nhiên không âm. Khi đó với mọi $\epsilon > 0$ ta có:*

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

Chứng minh

Ta chỉ chứng minh định lí trong hai trường hợp X rời rạc hoặc X liên tục.

Trường hợp X là đại lượng ngẫu nhiên rời rạc nhận các giá trị x_i với xác suất p_i

$$E(X) = \sum_i x_i p_i \geq \sum_{x_i \geq \epsilon} x_i p_i \geq \epsilon \sum_{x_i \geq \epsilon} p_i = \epsilon P(X \geq \epsilon).$$

Suy ra

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

Trường hợp X là đại lượng ngẫu nhiên có $f(x)$ là hàm mật độ. Ta có

$$E(X) = \int_0^\infty x f(x)dx \geq \int_\epsilon^\infty x f(x)dx \geq \epsilon \int_\epsilon^\infty f(x)dx = \epsilon P(X \geq \epsilon).$$

Vậy trong cả hai trường hợp

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

Từ định lí Máckốp, ta có

Định lí 4.2.2 (Trêbusép) *Giả sử X là đại lượng ngẫu nhiên tồn tại kì vọng $m = E(X)$ và phương sai $\sigma^2 = D(X)$. Khi đó với mọi $\epsilon > 0$ ta có:*

$$P(|X - m| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Chứng minh

Áp dụng định lí Máckốp cho đại lượng ngẫu nhiên $Y = (X - m)^2$ và ϵ^2 thay cho ϵ, ta được

$$P(|X - m| \geq \epsilon) = P(|X - m|^2 \geq \epsilon^2) \leq$$

$$\leq \frac{E(|X - m|^2)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}.$$

Mặc dù định lí Trêbusép (hay ta còn gọi là bất đẳng thức Trêbusép) được chứng minh khá đơn giản, song nó có ý nghĩa lớn về mặt ứng dụng. Trước hết phải nói đến ứng dụng của bất đẳng thức Trêbusép vào việc chứng minh luật yếu số lớn sẽ đề cập tới trong mục 4.4. Mặt khác nó cho ta một vài ước lượng xác suất có nhiều ý nghĩa trong thực tế.

Ví dụ 4.2.1

Trong các tính toán gần đúng nếu mọi số được vê tròn đến 2 chữ số sau dấu phảy (chẳng hạn 1,236645 được làm tròn thành 1,24), khi đó ta coi sai số X_i của việc làm tròn số thứ i là đại lượng ngẫu nhiên có phân bố đều trên đoạn [-0,005; 0,005]. Gọi X là sai số khi tính trung bình cộng n số đã được làm tròn số. Hiển nhiên

$$E(X_i) = 0, D(X_i) = \frac{1}{12 \cdot 10^2}$$

Ta có thể giả thiết $X_i, i = 1, 2, ..., n$ độc lập, suy ra

$$E(X) = E(\frac{X_1 + X_2 + \cdots + X_n}{n}) = 0, D(X) = \frac{1}{12 \cdot 10^2 n}$$

Theo định lí Trêbusép, xác suất để sai số của phép tính trung bình cộng vượt quá ϵ được ước lượng

$$P(|X| \geq \epsilon) \leq \frac{1}{12 \cdot 10^2 n\epsilon^2}$$

Chẳng hạn với $n > 8, \epsilon = 0,1$, sai số của phép tính trung bình cộng vượt quá ϵ có xác suất nhỏ hơn 0,01:

$$P(|X| \geq \epsilon) \leq \frac{1}{12 \cdot 10^2 n\epsilon^2} < \frac{1}{12n} < \frac{1}{100}.$$

4.3 Hội tụ theo xác suất và hội tụ hầu chắc chắn

Định nghĩa 4.3.1 *Cho một dãy các đại lượng ngẫu nhiên* $Y_n, \quad n = 1, 2, ...$ *Ta nói* Y_n *hội tụ theo xác suất tới đại lượng ngẫu nhiên* Y*, nếu với bất kì* $\epsilon > 0$ *nhỏ tùy ý*

$$\lim_{n\to\infty} P(|Y_n - Y| > \epsilon) = 0.$$

Chúng ta nhắc lại rằng các đại lượng ngẫu nhiên Y_n là các ánh xạ xác định trên không gian các biến cố ngẫu nhiên cơ bản Ω

$$Y_n : \Omega \to \mathbb{R}$$

Do vậy biến cố $|Y_n - Y| > \epsilon$ trong định nghĩa trên là biến cố gồm các biến cố ngẫu nhiên cơ bản ω thoả mãn

$$\omega : |Y_n(\omega) - Y(\omega)| > \epsilon.$$

Định nghĩa 1 đưa ra khái niệm hội tụ theo xác suất nếu biến cố ngẫu nhiên $\{\omega : |Y_n(\omega) - Y(\omega)| > \epsilon\}$ có xác suất dần tới 0 khi n tiến ra vô cùng.

Bây giờ ta nói đến khái niệm hội tụ hầu chắc chắn. Kí hiệu $\lim\limits_{n\to\infty} Y_n = Y$ là biến cố gồm các biến cố ngẫu nhiên cơ bản ω thoả mãn

$$\omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega),$$

ta có định nghĩa sau

Định nghĩa 4.3.2 *Dãy các đại lượng ngẫu nhiên* $Y_n, \quad n = 1, 2, ...$ *hội tụ hầu chắc chắn tới đại lượng ngẫu nhiên* Y*, nếu*

$$P(\lim_{n\to\infty} Y_n = Y) = 1.$$

Chú ý rằng trong cả hai định nghĩa trên, đại lượng ngẫu nhiên Y có thể là hằng số. Trong hai khái niệm hội tụ nêu trên hội tụ hầu chắc chắn mạnh hơn hội tụ theo xác suất. Nói một cách chính xác, người ta chứng minh được rằng nếu dãy Y_n hội tụ hầu chắc chắn tới Y, khi đó Y_n cũng hội tụ theo xác suất tới Y. Điều ngược lại nói chung không đúng. Người ta còn gọi luật số lớn liên quan tới sự hội tụ theo xác suất là *luật yếu số lớn* và luật số lớn liên quan tới sự hội tụ hầu chắc chắn là *luật mạnh số lớn*. Chúng ta sẽ phát biểu chi tiết các luật số lớn này trong mục 4.4 dưới đây.

4.4 Luật số lớn

Tiến hành phép thử ngẫu nhiên vô cùng lần độc lập với nhau. Kí hiệu X_i là đại lượng ngẫu nhiên nhận các giá trị 1 hoặc 0 tùy theo ở lần thử thứ i biến cố A xảy ra hay không

$$P(X_i = 1) = P(A) = p, P(X_i = 0) = P(\overline{A}) = q,\ i = 1, 2, ..., n$$

Hiển nhiên X_i phân bố theo luật 0-1 và $X_1 + X_2 + ... + X_n = k$ khi và chỉ khi có đúng k số 1 trong số n số hạng

$$X_1, X_2, ..., X_n$$

Nói cách khác biến cố $X_1 + X_2 + ... + X_n = k$ là biến cố có đúng k lần xảy ra A. Suy ra

$$X = \frac{X_1 + X_2 + ... + X_n}{n}$$

là tần suất xuất hiện của biến cố A trong n phép thử đầu tiên. Trước hết ta phát biểu và chứng minh luật yếu số lớn dạng Bernoulli

Định lí 4.4.1 (Bernoulli) *Giả sử biến cố ngẫu nhiên A có xác suất $P(A) = p$. Khi đó tần suất của biến cố A hội tụ theo xác suất tới xác suất p của biến cố đó. Nói cách khác*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + ... + X_n}{n} - p\right| > \epsilon\right) = 0$$

Chứng minh

Do X_i phân bố theo luật 0-1, nên $E(X_i) = p, D(X_i) = pq$, suy ra

$$E\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = p, D\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{pq}{n}$$

Áp dụng bất đẳng thức Trêbưsép cho $\frac{X_1+X_2+...+X_n}{n}$, ta có

$$P\left(\left|\frac{X_1 + X_2 + ... + X_n}{n} - p\right| > \epsilon\right) \leq \frac{pq}{n\epsilon^2},$$

từ đây suy ra điều phải chứng minh.

Định lí Bernoulli là trường hợp đặc biệt của định lí sau

Định lí 4.4.2 *Giả sử $X_1, X_2, ...$ là một dãy các đại lượng ngẫu nhiên độc lập có cùng kì vọng và phương sai*

$$E(X_i) = m, D(X_i) = \sigma^2, \quad i = 1, 2, ...,$$

Khi đó $\frac{X_1+X_2+...+X_n}{n}$ hội tụ theo xác suất tới kì vọng của các đại lượng ngẫu nhiên đó. Nói cách khác

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + ... + X_n}{n} - m\right| > \epsilon\right) = 0$$

Chứng minh

Hoàn toàn tương tự như chứng minh định lí Bernoulli, với $\epsilon > 0$ tùy ý

$$P\left(\left|\frac{X_1 + X_2 + ... + X_n}{n} - m\right| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Khi $n \to \infty$, hiển nhiên $n\epsilon^2 \to \infty$, suy ra điều phải chứng minh.

Nhận xét rằng trong chứng minh trên bất đẳng thức Trêbưsép không chỉ chứng minh sự hội tụ theo xác suất của $\frac{X_1+X_2+...+X_n}{n}$ tới m mà còn cho ta ước lượng xác suất về độ lệch của nó với kì vọng m nữa.

Ngoài ra Bernstein còn chứng minh được rằng nếu các đại lượng ngẫu nhiên X_i không độc lập và thay vào đó hệ số tương quan giữa X_i và X_j đủ nhỏ khi i hoặc j dần ra vô cùng, phương sai của X_i bị chặn, khi đó dãy các

đại lượng ngẫu nhiên X_i tuân theo luật yếu số lớn (*xem tài liệu tham khảo 5.*)

Như trên đã trình bày, khái niệm hội tụ hầu chắc chắn mạnh hơn, nó kéo theo hội tụ theo xác suất. Các định lí nghiên cứu sự hội tụ hầu chắc chắn của trung bình cộng của n đại lượng ngẫu nhiên đầu tiên của một dãy các đại lượng ngẫu nhiên nào đó được gọi là *luật mạnh số lớn*. Việc chứng minh các luật mạnh số lớn đòi hỏi các kiến thức sâu hơn về lí thuyết độ đo, vì vậy ta chỉ phát biểu, không chứng minh.

Định lí 4.4.3 *Giả sử $X_1, X_2, ...$ là một dãy các đại lượng ngẫu nhiên độc lập có cùng hàm phân bố, tồn tại kì vọng*

$$E(X_i) = m, \quad i = 1, 2, ...$$

Khi đó $\frac{X_1+X_2+...+X_n}{n}$ hội tụ hầu chắc chắn tới kì vọng của các đại lượng ngẫu nhiên đó. Nói cách khác

$$P\left(\lim_{n\to\infty} \frac{X_1 + X_2 + ... + X_n}{n} = m\right) = 1$$

Thậm chí người ta còn chứng minh được rằng nếu X_i độc lập, có cùng phân bố, điều kiện cần và đủ để các đại lượng ngẫu nhiên

$$\frac{X_1 + X_2 + ... + X_n}{n}$$

hội tụ hầu chắc chắn tới hằng số C nào đó là tồn tại kì vọng $E(X_i)$ và suy ra $E(X_i) = C$.

Ví dụ 4.4.1 (Phương pháp Monte-Carlo)

Cho ánh xạ $g(x)$ liên tục, xác định trên đoạn [0, 1]. Ta tính tích phân

$$\int_0^1 g(x)\,dx$$

bằng phương pháp như sau. Giả sử $X_1, X_2, ...$ là một dãy các đại lượng ngẫu nhiên độc lập có cùng *phân bố đều* trên đoạn [0, 1]. Khi đó dãy các đại lượng ngẫu nhiên

$$g(X_1), g(X_2), ..., g(X_n), ...$$

cũng độc lập, cùng hàm phân bố, có kì vọng (theo định lí 4 chương II) bằng

$$E(g(X_i)) = \int_0^1 g(x)\,dx, \quad i = 1, 2, ...$$

Áp dụng định lí 5 luật số lớn nói trên

$$\frac{g(X_1) + g(X_2) + \cdots + g(X_n)}{n} \quad \rightarrow \quad \int_0^1 g(x)\,dx$$

hội tụ hầu chắc chắn. Trong thực hành, để tính gần đúng giá trị tích phân, ta sử dụng

$$\frac{g(X_1) + g(X_2) + \cdots + g(X_n)}{n} \approx \int_0^1 g(x)\,dx.$$

Dãy các đại lượng ngẫu nhiên độc lập, phân bố đều trên đoạn [0, 1]

$$X_1, X_2, ...$$

có thể tạo được chẳng hạn bằng lệnh "RAND" trong chương trình EXCEL thông dụng.

Xét ví dụ, tính tích phân sau

$$\int_0^1 e^{x^2}\,dx$$

Tạo 35 số ngẫu nhiên bằng lệnh "RAND", ta được các số X_i sau:

{0,876766; 0,0242577; 0,135319; 0,270109; 0,916834; 0,331621; 0,634461; 0,202422; 0,487615; 0,777944; 0,741303; 0,417968; 0,544738; 0,995114; 0,2672570; 0,514089; 0,912009; 0,1319020; 0,659020; 0,719162; 0,1500810; 0,715338; 0,915038; 0,592031; 0,2733150; 0,69108; 0,7797190; 0,321922; 0,356482; 0,35946; 0,145258; 0,1195; 0,868867; 0,581515; 0,403955}

Các số $g(X_i) = e^{X_i^2}$ tương ứng là:

{2,157; 1,00059; 1,01848; 1,07569; 2,31772; 1,11625; 1,49562; 1,04183; 1,26842; 1,83161; 1,73244; 1,19089; 1,34546; 2,69191; 1,07404; 1,3025; 2,29736; 1,01755; 1,54389; 1,67732; 1,02278; 1,66814; 2,31011; 1,41978; 1,07756; 1,61219; 1,83668; 1,10919; 1,13551; 1,13793; 1,02132; 1,01438; 2,12746; 1,40236; 1,17725}

Trung bình cộng của các số $g(X_i)$ bằng

$$\frac{\sum_{i=1}^{35} g(X_i)}{35} = 1,46483.$$

Vậy

$$\int_0^1 e^{x^2}\,dx \approx 1,46483.$$

Phương pháp nêu trên để tính tích phân được gọi là *phương pháp Monte-Carlo*, nó đặc biệt hiệu quả để tính tích phân bội cũng như nhiều bài toán phức tạp khác.

4.5 Định lí giới hạn trung tâm

Nếu luật số lớn ở mục trước khẳng định sự ổn định của tần suất xung quanh xác suất của biến cố ngẫu nhiên thì ở mục này các định lí giới hạn cũng như định lí giới hạn trung tâm giải thích sự xuất hiện thường xuyên của phân bố chuẩn hoặc xấp xỉ chuẩn trong nhiều lĩnh vực ứng dụng. Trước hết ta phát biểu định lí sau:

Định lí 4.5.1 *Giả sử X là đại lượng ngẫu nhiên có phân bố nhị thức*

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

Kí hiệu

$$x = \frac{k - np}{\sqrt{npq}} \quad \textit{và giả sử} \quad |x| \le A$$

Khi đó

$$P(X = k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{npq}} (1 + O(\frac{1}{\sqrt{n}}))$$

trong đó $|O(\frac{1}{\sqrt{n}})| \le \frac{C}{\sqrt{n}}$, hằng số C phụ thuộc chỉ vào A.

Như vậy xác suất $P(X = k) = C_n^k p^k q^{n-k}$ xấp xỉ với biểu thức

$$P(X = k) = C_n^k p^k q^{n-k} \approx \frac{1}{\sqrt{npq}} f(x)$$

khi n đủ lớn, trong đú

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

Nhận xét rằng $f(x)$ chính là hàm mật độ của phân bố chuẩn thuộc lớp $N(0,1)$, định lí trên được gọi là *định lí giới hạn địa phương.*

Ví dụ 4.5.1

Hãy tính xác suất để biến cố A xảy ra đúng 45 lần khi tiến hành phép thử ngẫu nhiên 200 lần độc lập nhau để quan sát biến cố A. Giả thiết rằng xác suất xảy ra biến cố A trong một phép thử ngẫu nhiên là $p = P(A) = 0,2$. Theo công thức Bernoulli, xác suất đó bằng

$$C_{200}^{45}(0,2)^{45}(0,8)^{155}$$

Việc tính trực tiếp xác suất theo công thức trên là không thể, vì số các phép tính quá nhiều. Áp dụng định lí trên ta có

$$C_{200}^{45}(0,2)^{45}(0,8)^{155} \approx \frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

trong đó

$$np = 200 \cdot 0,2 = 40, \sqrt{npq} = 5,65685$$

$$x = \frac{45 - np}{\sqrt{npq}} = \frac{5}{5,65685} = 0,883883$$

$$f(x) = f(0,883883) = 0,269938$$

Vậy xác suất cần tìm xấp xỉ bằng

$$C_{200}^{45} p^{45} q^{155} \approx \frac{1}{\sqrt{npq}} f(x) = \frac{0,269938}{5,65685} = 0,048.$$

Định lí 4.5.2 (Moivre-Laplace) *Giả sử X là đại lượng ngẫu nhiên có phân bố nhị thức*

$$P(X = k) = C_n^k p^k q^{n-k}, 0 < p < 1, q = 1 - p, 0 \leq k \leq n$$

Khi đó

$$\lim_{n\to\infty} P\left(\frac{X - np}{\sqrt{npq}} < x\right) = \Phi(x)$$

trong đó

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

là hàm phân bố của phân bố chuẩn thuộc lớp $N(0,1)$.

Định lí Moivre-Laplace được chứng minh khá phức tạp, ta phát biểu một dạng khác của định lí có nhiều ý nghĩa trong thực hành:

$$P(A \le X \le B) = \Phi(b) - \Phi(a) + R$$

trong đó A, B là hai số nguyên $0 \le A < B \le n$, a và b được xác định qua A, B:

$$a = \frac{A - np - \frac{1}{2}}{\sqrt{npq}}, b = \frac{B - np + \frac{1}{2}}{\sqrt{npq}}$$

phần dư R được ước lượng như sau:

$$R = \frac{q-p}{6\sqrt{npq}} \left((1-b^2)e^{-\frac{b^2}{2}} - (1-a^2)e^{-\frac{a^2}{2}} \right) + O(\frac{1}{n})$$

như vậy với n đủ lớn ($n \to \infty$), ta có thể bỏ số hạng dư R để tính gần đúng xác suất $P(A \le X \le B)$ thông qua hàm phân bố $\Phi(x)$ của phân bố chuẩn:

$$P(A \le X \le B) \approx \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

Chi tiết hơn

$$\boxed{P(A \le X \le B) \approx \Phi\left(\frac{B - np + \frac{1}{2}}{\sqrt{npq}}\right) - \Phi\left(\frac{A - np - \frac{1}{2}}{\sqrt{npq}}\right).}$$

Ví dụ 4.5.2

Qua số liệu thống kê nhiều năm người ta biết rằng khi bắt đầu chuyển sang mùa hè, tỉ lệ trẻ em dưới 5 tuổi bị viêm phế quản là 25%.

(a) Tìm xác suất để số trẻ em bị viêm phế quản trong một nhà trẻ gồm 100 cháu dao động từ 10 đến 20.

(b) Hãy tìm xác suất để một trường mẫu giáo gồm 500 em có không quá 120 em bị viêm phế quản.

Giải:

(a) Gọi X là số lượng trẻ em bị viêm phế quản trong nhà trẻ. (Nhà trẻ gồm 100 cháu). Xác suất để 1 cháu mắc bệnh viêm phế quản theo giả thiết là 25%=0,25. Mặt khác bệnh viêm phế quản ở các cháu độc lập nhau. Vì vậy đại lượng ngẫu nhiên X có phân bố nhị thức với tham số p =0,25 $n = 100$.

$$P(X = k) = C_n^k p^k q^{n-k} = C_{100}^k (0,25)^k (0,75)^{100-k}$$

trong đó $0 \leq k \leq 100$. Suy ra xác suất cần tìm bằng

$$P(10 \leq X \leq 20) = \sum_{k=10}^{20} C_{100}^k (0,25)^k (0,75)^{100-k}.$$

Cơ sở của việc tính gần đúng xác suất trên là định lí Moivre-Laplace. Theo đó đại lượng ngẫu nhiên X có phân bố xấp xỉ chuẩn và

$$P(10 \leq X \leq 20) \approx \Phi(b) - \Phi(a).$$

trong đó

$$a = \frac{10 - 100 \cdot 0,25 - \frac{1}{2}}{\sqrt{100 \cdot 0,25 \cdot 0,75}} = -3,58$$

$$\Phi(a) = \Phi(-3,58) = 0,000172$$

và

$$b = \frac{20 - 100 \cdot 0,25 + \frac{1}{2}}{\sqrt{100 \cdot 0,25 \cdot 0,75}} = -1,04.$$

$$\Phi(b) = \Phi(-1,04) = 0,14917$$

Vậy

$$P(10 \leq X \leq 20) \approx 0,14917 - 0,000172 = 0,148998.$$

(b) Tương tự như phần (a) xác suất cần tìm xấp xỉ

$$\Phi(\frac{120 - 500 \cdot 0,25 + \frac{1}{2}}{\sqrt{500 \cdot 0,25 \cdot 0,75}}) = \Phi(-0,465) = 0,32.$$

Bây giờ chúng ta phát biểu định lí *giới hạn trung tâm*, mở rộng của định lí Moivre-Laplace và là một trong các kết quả quan trọng của lí thuyết xác suất.

Định lí 4.5.3 (Định lí giới hạn trung tâm) *Giả sử $X_1, X_2, ...$ là một dãy các đại lượng ngẫu nhiên độc lập có cùng phân bố với kì vọng và phương sai*

$$E(X_i) = m, D(X_i) = \sigma^2, \quad i = 1, 2, ...,$$

Khi đó

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + ... + X_n - nm}{\sigma\sqrt{n}} < x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt.$$

Nhận xét rằng nm và $\sigma\sqrt{n}$ trong định lí chính là kì vọng và độ lệch chuẩn của đại lượng ngẫu nhiên

$$X_1 + X_2 + ... + X_n.$$

Trường hợp đặc biệt khi X_i độc lập có cùng phân bố theo luật 0-1, đại lượng ngẫu nhiên

$$X_1 + X_2 + ... + X_n$$

là đại lượng ngẫu nhiên có phân bố nhị thức với kì vọng bằng np, phương sai bằng npq. Do vậy định lí giới hạn trung tâm trong trường hợp đặc biệt này trở thành định lí Moivre-Laplace:

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + ... + X_n - np}{\sqrt{npq}} < x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt.$$

Định lí trên vẫn đúng dưới các điều kiện tổng quát hơn nhiều. Ta phát biểu không chứng minh định lí sau (mang tên định lí giới hạn trung tâm của Liapunov)

Định lí 4.5.4 (Định lí Liapunov) *Giả sử $X_1, X_2, ..., X_n, ...$ là một dãy các đại lượng ngẫu nhiên độc lập (không giả thiết chúng cùng phân bố) có kì vọng, phương sai và mô men quy tâm cấp 3*

$$E(X_i) = m_i, D(X_i) = \sigma_i^2, E(|X_i - m_i|^3) = h_i^3 \quad i = 1, 2, ...,$$

Kí hiệu

$$S_n = X_1 + X_2 + \cdots + X_n$$

$$s_n = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}$$

$$k_n = \sqrt{h_1^3 + h_2^3 + \cdots + h_n^3}.$$

Giả thiết rằng

$$\lim_{n\to\infty} \frac{k_n}{s_n} = 0,$$

khi đó

$$\lim_{n\to\infty} P\left(\frac{S_n - E(S_n)}{\sqrt{E(S_n)}} < x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt.$$

Bài tập 2. chương này cho ta một ví dụ về dãy các đại lượng ngẫu nhiên độc lập không tuân theo luật số lớn, song sử dụng định lí Liapunov, ta có thể chứng minh rằng trung bình cộng của n số hạng đầu của dãy có phân bố xấp xỉ chuẩn khi n đủ lớn.

4.6 Xấp xỉ phân bố nhị thức với ph.bố Poisson

Trong mục trước ta đã giới thiệu vai trò quan trọng của phân bố chuẩn, đặc biệt có thể xấp xỉ phân bố nhị thức với phân bố chuẩn. Bây giờ chúng ta sẽ chứng minh một định lí giới hạn khác liên quan tới phân bố Poisson, trong đó cũng chỉ rõ với các điều kiện nào phân bố nhị thức có thể xấp xỉ được với phân bố Poisson.

Xét các số hạng sau của phân bố nhị thức

$$P_n^{(k)} = C_n^k p^k (1-p)^{n-k}, \quad k = 0, 1, ..., n.$$

Chúng ta giả thiết n trong phân bố nhị thức nêu trên khá lớn và xác suất p rất nhỏ. Trước mắt để thuận tiện về mặt kí hiệu ta giả thiết rằng n tiến dần ra vô cùng, trong khi p tiến dần tới 0 sao cho tích np là hằng số:

$$np = \lambda > 0, \quad \text{hay} \quad p = \frac{\lambda}{n}.$$

Xác suất $P_n^{(k)}$ được biến đổi như sau:

$$P_n^{(k)} = C_n^k p^k (1-p)^{n-k} = C_n^k (\frac{\lambda}{n})^k (1-\frac{\lambda}{n})^{n-k} =$$

$$= \frac{n(n-1)\cdots(n-k+1)}{k!} \frac{\lambda^k}{n^k} (1-\frac{\lambda}{n})^n (1-\frac{\lambda}{n})^{-k} =$$

$$= \frac{\lambda^k}{k!} (1-\frac{\lambda}{n})^n \frac{n(n-1)\cdots(n-k+1)}{n^k} (1-\frac{\lambda}{n})^{-k}.$$

Với k cố định và n tiến dần ra vô cùng, khi đú

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} = (1-\frac{1}{n})(1-\frac{2}{n})\cdots(1-\frac{k-1}{n}) \to 1$$

$$(1-\frac{\lambda}{n})^{-k} \to 1, \quad (1-\frac{\lambda}{n})^n \to e^{-\lambda}.$$

Vậy giới hạn của phân bố nhị thức bằng

$$\lim_{n\to\infty} P_n^{(k)} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, ...$$

Như đã định nghĩa trong chương II, $e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, ...$ là các số hạng của phân bố Poisson. Ngoài ra ta có nhận xét rằng chứng minh trên vẫn đúng nếu thay cho giả thiết tích $np = \lambda > 0$ là hằng số, ta giả thiết rằng xác suất $p = p(n)$ (p là hàm số phụ thuộc n) tiến dần tới 0 sao cho

$$\lim_{n\to\infty} np = \lambda > 0.$$

Tóm tắt lại kết quả thu được , ta có định lí giới hạn sau

Định lí 4.6.1 *Khi n khá lớn và xác suất p đủ nhỏ, cùng với giả thiết* $\lim_{n\to\infty} np = \lambda > 0$*, các số hạng của phân bố nhị thức tiến dần tới các số hạng của phân bố Poisson, khi* $n \to \infty$

$$\lim_{n\to\infty} P_n^{(k)} = \lim_{n\to\infty} C_n^k p^k (1-p)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Nói cách khác khi n lớn và p nhỏ, ta xấp xỉ các số hạng của phân bố nhị thức

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

với xác suất

$$e^{-np}\frac{(np)^k}{k!}.$$

Câu hỏi được đặt ra là sai số của xấp xỉ đó khi n, k, p cho trước được đánh giá ra sao? Người ta chứng minh được kết quả sau. Tỉ số

$$C_n^k p^k (1-p)^{n-k} / e^{-np}\frac{(np)^k}{k!} = e^A$$

trong đó

$$A = \frac{kp^2}{2} - \frac{(k-np)^2}{2n} - \frac{k^3}{2n^2} + R,$$

$$|R| < \frac{k^2(3k+4)}{12(n-k)^2} + \frac{n-k}{3}\left(\frac{p}{1-p}\right)^3.$$

Như vậy xấp xỉ các số hạng của phân bố nhị thức với phân bố Poisson khá tốt khi n lớn, k và p đủ nhỏ. Để minh hoạ ta lập bảng so sánh 10 số hạng đầu của phân bố nhị thức với các số hạng tương ứng của phân bố Poisson trong trường hợp $p = \frac{1}{32}$ và $n = 64$.

k	Phân bố nhị thức	Phân bố Poisson
0	0,131	0,135
1	0,271	0,271
2	0,275	0,271
3	0,183	0,180
4	0,090	0,090
5	0,035	0,036
6	0,011	0,012
7	0,003	0,003
8	0,001	0,001
9	0,000	0,000

(*Các kết quả trên được tính bằng Mathematica 4.0*)

Ví dụ 4.6.1

Biết rằng tại một công ty may mặc xác suất để một áo sơ mi bị lỗi là $p = 0,012$. (Nói cách khác tỉ lệ phế phẩm bằng $1,2\%$). Hãy tính xác suất để

(a) Trong một lô hàng gồm 500 áo không có chiếc áo nào bị lỗi.

(b) Trong một lô hàng gồm 500 áo, số áo lỗi không vượt quá 11.

Giải:

(a) Gọi X là số áo sơ mi bị lỗi trong lô hàng gồm 500 chiếc. X là đại lượng ngẫu nhiên có phân bố nhị thức. Xác suất cần tìm bằng

$$P(X = 0) = C_{500}^{0}(1 - 0,012)^{500} = (0,988)^{500}.$$

Để tính xác suất $P(X = 0)$, do $n = 500$ tương đối lớn và $p = 0,012$ khá bé, ta xấp xỉ X với với đại lượng ngẫu nhiên có phân bố Poisson với tham số $\lambda = 500 \cdot 0,012 = 6$.

$$P(X = 0) \approx e^{-np}\frac{(np)^k}{k!} = e^{-6} = 0,002479.$$

(b) Ta xấp xỉ

$$P(X \leq 11) \approx \sum_{k=0}^{11} e^{-np}\frac{(np)^k}{k!} = 0,98.$$

Vậy hầu như chắc chắn (với xác suất 0,98) số áo lỗi trong lô hàng gồm 500 chiếc không vượt quá 11.

BÀI TẬP CHƯƠNG IV

1. Giả sử $X_1, X_2, ..., X_6$ là các đại lượng ngẫu nhiên độc lập có cùng kì vọng và phương sai

$$E(X_i) = 15, D(X_i) = 1, \quad i = 1, 2, ..., 6.$$

Sử dụng bất đẳng thức Trêbusép, hãy ước lượng xác suất

$$P\left(80 < X_1 + X_2 + ... + X_6 < 100\right).$$

2. Giả sử $X_1, X_2, ..., X_n, ...$ là một dãy các đại lượng ngẫu nhiên độc lập, phân bố của chúng được cho như sau

$$P(X_n = \sqrt{n}) = P(X_n = -\sqrt{n}) = \frac{1}{2} \quad \text{với mọi} \quad n = 1, 2, ...$$

Đặt $S_n = \frac{X_1 + X_2 + ... + X_n}{n}$.

(a) Hãy tính kì vọng và phương sai của S_n.

(b) Chứng tỏ rằng

$$S_n = \frac{X_1 + X_2 + ... + X_n}{n}$$

không tuân theo luật số lớn.

(c) Chứng minh rằng phân bố của S_n tiến dần tới phân bố chuẩn

$$\lim_{n\to\infty} P(S_n < x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt.$$

3. Giả sử $X_1, X_2, ...$ là một dãy các đại lượng ngẫu nhiên đôi một độc lập, tồn tại kì vọng và phương sai

$$E(X_i) = m_i, D(X_i) = \sigma_i^2, \quad i = 1, 2, ...,$$

Ngoài ra ta giả thiết tiếp rằng

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m_i = m \quad \text{hữu hạn}$$

$$\lim_{n\to\infty}\frac{\sqrt{\sum_{i=1}^{n}\sigma_i^2}}{n} = 0.$$

Khi đó $\frac{X_1+X_2+...+X_n}{n}$ hội tụ theo xác suất tới m. Nói cách khác

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+...+X_n}{n} - m\right| > \epsilon\right) = 0.$$

4. Gọi k_n là tần suất xuất hiện mặt ngửa khi gieo đồng xu đối xứng, đồng chất n lần. Hãy xác định số lần gieo n sao cho

$$\left|k_n - \frac{1}{2}\right| < 0,01$$

với xác suất 0,95.

5. Hãy tính tích phân sau

$$\int_0^{\frac{\pi}{2}} \sin(x^2)\,dx.$$

ĐÁP SỐ VÀ HƯỚNG DẪN

1. $P\left(80 < X_1 + X_2 + ... + X_6 < 100\right) \le 1 - \frac{6}{100} = 0,94.$

2. Kì vọng $E(S_n) = 0$ và phương sai $D(S_n) = n.$

3. Chứng minh như chứng minh định lí 4.4.2

4. $n \geq 9604$.

Hướng dẫn: Sử dụng định lí giới hạn trung tâm, chứng minh rằng nếu

$$n \geq u_\alpha^2 \frac{\sigma^2}{\epsilon^2}.$$

Khi đó

$$P\left(\left|\frac{X_1 + X_2 + ... + X_n}{n} - m\right| < \epsilon\right) \geq 1 - \alpha.$$

Trong đó u_α được xác định từ hệ thức

$$\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt = 1 - \alpha.$$

5. *Hướng dẫn*: Sử dụng phương pháp Monte-Carlo như trong ví dụ 4.4.1.

Chương 5

Thống kê toán

5.1 Cơ sở của thống kê toán

5.1.1 Khái niệm về mẫu ngẫu nhiên

Tiến hành phép thử ngẫu nhiên T nào đó để quan sát biến cố ngẫu nhiên A hoặc đại lượng ngẫu nhiên X. Kết quả thu được có thể coi như một thể hiện của đại lượng ngẫu nhiên. Để có các kết luận về biến cố A hoặc đại lượng ngẫu nhiên X, người ta phải tiến hành không chỉ một lần phép thử ngẫu nhiên T đó.

Giả sử phép thử ngẫu nhiên T được tiến hành n lần độc lập với nhau. Kí hiệu X_i là kết quả của lần thử thứ i về đại lượng ngẫu nhiên X. Khi đó

$$(X_1, X_2, ..., X_n)$$

được gọi là *mẫu ngẫu nhiên* và n là *kích thước mẫu (hay số phần tử mẫu)* của mẫu ngẫu nhiên đó.

Do các phép thử ngẫu nhiên được tiến hành độc lập với nhau nên

$$X_1, X_2, ..., X_n$$

là các đại lượng ngẫu nhiên độc lập và có cùng phân bố như phân bố của X.

Các thể hiện cụ thể của mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

là các bộ n số $(x_1, x_2, ..., x_n)$. Nếu người khác hoặc lúc khác tiến hành n lần độc lập nhau phép thử ngẫu nhiên T, ta có thể nhận được các bộ số khác nhau.

Nhiệm vụ của thống kê toán là đưa ra các kết luận liên quan tới phân bố của X dựa trên mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

(hay nói chính xác hơn là dựa trên các thể hiện cụ thể mẫu ngẫu nhiên $(x_1, x_2, ..., x_n)$, và đó cũng là các thông tin duy nhất mà ta dựa vào để rút ra các kết luận cần thiết).

Chúng ta nhắc lại rằng từ nay về sau nếu nói

$$(X_1, X_2, ..., X_n)$$

là mẫu ngẫu nhiên, khi đó $X_1, X_2, ..., X_n$ là các đại lượng ngẫu nhiên độc lập và có cùng phân bố.

Nhận xét rằng nếu chúng ta cần có các kết luận về hai đại lượng ngẫu nhiên X và Y (chẳng hạn hệ số tương quan giữa chúng), khi đó mẫu ngẫu nhiên tương ứng với X và Y là

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$$

trong đó $(X_i, Y_i) \quad i = 1, 2, ..., n$ là các véc tơ ngẫu nhiên độc lập và có cùng phân bố như phân bố của (X, Y).

Như trên đã nói mẫu ngẫu nhiên là thông tin duy nhất mà ta dựa vào để rút ra các kết luận thống kê do vậy mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

phải phản ánh, đại diện một cách trung thực cho đại lượng ngẫu nhiên X. Việc tổ chức lấy mẫu, phương pháp chọn mẫu là những vấn đề nằm ngoài khuôn khổ cuốn sách này.

Mục tiêu cuối cùng của thống kê toán là đưa ra các kết luận thống kê. Bản chất của các kết luận thống kê cũng giống như các kết luận lôgic khác trong toán, tuy nhiên có sự khác nhau: các kết luận thống kê chỉ khẳng định một vấn đề nào đó *đúng với xác suất gần* 1 *hoặc xấp xỉ với* 1. Chẳng hạn khi

ta đưa ra một kết luận K với xác suất 95%, khi đó kết luận ấy có thể sai. Tuy nhiên trung bình trong 100 trường hợp có 95 trường hợp kết luận K là đúng. Với các hiện tượng ngẫu nhiên chúng ta phải chấp nhận và hài lòng với các kết luận kiểu như vậy.

5.1.2 Phân bố mẫu (hoặc hàm phân bố thực nghiệm)

Xét một mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

tương ứng với đại lượng ngẫu nhiên X. Gọi $F(x)$ là hàm phân bố của đại lượng ngẫu nhiên X. Kí hiệu $(x_1, x_2, ..., x_n)$ là thể hiện cụ thể của mẫu đó. Giả sử ξ là đại lượng ngẫu nhiên nhận các giá trị $x_i, \quad i = 1, 2, ..., n$ với xác suất $\frac{1}{n}$:

$$P(\xi = x_i) = \frac{1}{n} \quad \text{với mọi} \quad i = 1, 2, ..., n.$$

Khi đó hàm phân bố của ξ được gọi là *hàm phân bố mẫu (hay hàm phân bố thực nghiệm)*. Nếu kí hiệu ξ_x là số các giá trị x_i trong mẫu

$$(x_1, x_2, ..., x_n)$$

mà $x_i < x$, khi đó hàm phân bố của ξ (phân bố thực nghiệm) kí hiệu $F_n(x) = P(\xi < x)$ bằng

$$F_n(x) = \frac{\xi_x}{n}.$$

Hàm phân bố mẫu phụ thuộc vào các giá trị cụ thể của mẫu ngẫu nhiên nên bản thân phân bố mẫu $F_n(x)$ cũng là đại lượng nhẫu nhiên. Đại lượng nhẫu nhiên $F_n(x)$ nhận các giá trị $\frac{k}{n}$ với xác suất bằng xác suất của biến cố $\{\xi_x = k\}$ (đồng thời cũng là biến cố có đúng k giá trị x_i trong mẫu $(x_1, x_2, ..., x_n)$ nhỏ hơn x). Do $P(X < x) = F(x)$ nên ξ_x là đại lượng ngẫu nhiên có phân bố nhị thức

$$P(F_n(x) = \frac{k}{n}) = P(\xi_x = k) = C_n^k (F(x))^k (1 - F(x))^{n-k}.$$

Suy ra

$$E(F_n(x)) = F(x).$$

Vậy theo luật số lớn hàm phân bố mẫu $F_n(x)$ hội tụ theo xác suất tới hàm phân bố lí thuyết $F(x)$.

Tổng quát hơn ta phát biểu định lí sau, còn được gọi là định lí cơ bản của thống kê:

Định lí 5.1.1 (Định lí Glivenko) *Giả sử $F(x)$ là hàm phân bố của đại lượng ngẫu nhiên X, $F_n(x)$ là hàm phân bố mẫu nhận được từ mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ tương ứng với X. Khi đó*

$$P(\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0) = 1.$$

Như vậy hàm phân bố mẫu $F_n(x)$ xấp xỉ với hàm phân bố lí thuyết $F(x)$ mà ta cần đưa ra các kết luận thống kê về nó.

5.1.3 Các đặc trưng mẫu

Xét một mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

tương ứng với đại lượng ngẫu nhiên X. Giả thiết rằng đại lượng ngẫu nhiên X tồn tại kì vọng và phương sai

$$E(X) = m, \quad D(X) = \sigma^2.$$

Ở mục trên ta đã dẫn vào đại lượng ngẫu nhiên ξ:

$$P(\xi = x_i) = \frac{1}{n} \quad \text{với mọi} \quad i = 1, 2, ..., n.$$

Các đặc trưng của đại lượng ngẫu nhiên ξ (kì vọng $E(\xi)$, phương sai $D(\xi)$, mô men) được gọi là các *đặc trưng mẫu*. Người ta thường kí hiệu $\overline{X} = E(\xi)$ là *kì vọng mẫu* và $S^2 = D(\xi)$ là *phương sai mẫu*. Hiển nhiên

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

và

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

Do $(X_1, X_2, ..., X_n)$ là các đại lượng ngẫu nhiên độc lập và có cùng phân bố với X nên

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = m$$

$$D(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n}.$$

Để tính kì vọng của phương sai mẫu, ta sử dụng

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

Suy ra

$$E(S^2) = \frac{1}{n}E\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(m^2 + \sigma^2) - \left(m^2 + \frac{\sigma^2}{n}\right) = \frac{n-1}{n}\sigma^2.$$

Ta dẫn vào đại lượng ngẫu nhiên

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Khi đó

$$E(S^{*2}) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2.$$

S^{*2} được gọi là *phương sai mẫu điều chỉnh*. Phương sai mẫu điều chỉnh sai lệch rất ít so với phương sai mẫu, khi kích thước mẫu n lớn. Do mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

tương ứng với đại lượng ngẫu nhiên X có kì vọng bằng m, phương sai bằng σ^2 và

$$E(\overline{X}) = m = E(X), E(S^{*2}) = \sigma^2 = D(X),$$

người ta thường dùng $\overline{X}$ và S^{*2} làm các ước lượng cho m, σ^2.

Ta có nhận xét rằng, theo luật số lớn các đặc trưng mẫu:

Nhận xét 5.1.1

1. $\overline{X}$ không những hội tụ theo xác suất mà hội tụ hầu chắc chắn tới $m = E(X)$.

2. S^2, S^{*2} hội tụ hầu chắc chắn (suy ra cũng hội tụ theo xác suất) tới σ^2 khi n tăng ra vô cùng.

5.2 Các phân bố thường gặp trong thống kê

5.2.1 Hàm Gamma, hàm Beta

Các hàm mật độ của các phân bố xác suất thường gặp trong thống kê nói chung là các hàm phức tạp, nó được biểu diễn thông qua các hàm đặc biệt: Hàm Gamma, hàm Beta. Do vậy trước tiên chúng ta cần làm quen với các hàm Gamma, Beta và một vài tính chất của chúng.

Bổ đề 5.2.1 *Tích phân suy rộng $\int_0^{+\infty} e^{-t}t^{x-1}dt$ hội tụ với mọi số thực $x > 0$.*

Chứng minh.

• Với $0 < x < 1$ xét hai tích phân $I_1 = \int_0^1 e^{-t}t^{x-1}dt$ và $I_2 = \int_1^{+\infty} e^{-t}t^{x-1}dt$.

Tích phân I_1 hội tụ vì với $0 < x < 1, 0 < t \leq 1$, ta có $e^{-t}t^{x-1} < \dfrac{1}{t^{1-x}}$ và trong Giải tích 1 ta đã biết tích phân $\int_0^1 \dfrac{1}{t^{1-x}}dt$ hội tụ.

Tích phân I_2 hội tụ vì $\lim\limits_{t\to+\infty} e^{-t}t^{x+1} = 0$, suy ra với t đủ lớn $e^{-t}t^{x-1} < \frac{1}{t^2}$ và tích phân $\int_1^{+\infty} \dfrac{1}{t^2}\,dt$ hội tụ. Vậy $I_1 = \int_0^1 e^{-t}t^{x-1}dt = I_1 + I_2$ hội tụ.

• Với $x \geq 1$ tương tự như trên, với t đủ lớn $e^t > t^{x+1} \Rightarrow e^{-t}t^{x-1} < \dfrac{1}{t^2}$. Vậy tích phân đã cho hội tụ với mọi $x > 0$.

Từ bổ đề trên, người ta định nghĩa

Định nghĩa 5.2.1 *Hàm Gamma $\Gamma(x) = \int_0^{+\infty} e^{-t}t^{x-1}dt$ được xác định với mọi số thực $x > 0$.*

Hàm Gamma có các tính chất quan trọng dưới đây. Chúng ta dựa vào các tính chất này để tính giá trị của chúng khi cần.

1. $\Gamma(1) = \int_0^{+\infty} e^{-t}dt = 1.$

2. $\Gamma(x+1) = x\Gamma(x)$ với mọi $x > 0$.

   Thật vậy, bằng cách tính tích phân từng phần

$$\begin{aligned}\Gamma(x+1) &= \int_0^{+\infty} e^{-t}t^x dt = -\int_0^{+\infty} t^x de^{-t} \\ &= -t^x e^{-t}\Big|_0^{+\infty} + \int_0^{+\infty} xt^{x-1}e^{-t}dt = 0 + x\Gamma(x).\end{aligned}$$

   Từ hai tính chất trên, bằng quy nạp ta có

3. Với $x - k > 0$, k là số tự nhiên bất kì

$$\Gamma(x) = (x-1)(x-2)\cdots(x-k)\Gamma(x-k).$$

   Suy ra

$$\Gamma(n) = (n-1)! \text{ với mọi số tự nhiên } \ n = 1, 2...$$

4. Chú ý rằng $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Thật vậy đổi biến $u = \sqrt{t}$

$$\Gamma(\frac{1}{2}) = \int_0^{+\infty} \frac{e^{-t}}{\sqrt{t}}dt = 2\int_0^{+\infty} e^{-u^2}du = \sqrt{\pi}.$$

   Suy ra với mọi số tự nhiên $n \in \mathbb{N}^*$

$$\Gamma(n + \frac{1}{2}) = \frac{1\cdot 3\cdots(2n-1)}{2^n}\sqrt{\pi} = \frac{(2n-1)!!}{2^n}\sqrt{\pi}$$

5. $\lim\limits_{x\to 0+} \Gamma(x) = \lim\limits_{x\to 0+} \dfrac{\Gamma(x+1)}{x} = +\infty.$

Bổ đề 5.2.2 *Tích phân suy rộng* $\int_0^1 t^{x-1}(1-t)^{y-1}dt$ *hội tụ với mọi số thực* $x > 0, y > 0$.

Chứng minh.

• Trường hợp $0 < x, y < 1$, xét hai tích phân sau, với $\forall c \in (0,1)$

$I_1 = \int_0^c t^{x-1}(1-t)^{y-1}dt$ và $I_2 = \int_c^1 t^{x-1}(1-t)^{y-1}dt$.

Tích phân I_1 hội tụ vì $t^{x-1}(1-t)^{y-1} < \dfrac{K}{t^{1-x}}$ với số K thích hợp nào đó và tích phân $\int_0^c \dfrac{K}{t^{1-x}}\,dt$ hội tụ do $x > 0$.

Tích phân I_2 hội tụ được chứng minh tương tự. Suy ra tích phân đã cho $\int_0^1 t^{x-1}(1-t)^{y-1}dt = I_1 + I_2$ hội tụ.

• Trường hợp ngược lại hoặc $x \geq 1$ hoặc $y \geq 1$ bổ đề trở thành hiển nhiên.

Từ bổ đề trên, người ta định nghĩa

Định nghĩa 5.2.2 *Hàm Beta* $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$ *được xác định với mọi số thực* $x > 0, y > 0$.

Hàm Beta có các tính chất sau

1. $B(x,y) > 0$ với mọi số thực $x > 0, y > 0$.
2. $B(x,y) = B(y,x)$ vì các hàm $t^{x-1}(1-t)^{y-1}$ và $t^{y-1}(1-t)^{x-1}$ có đồ thị đối xứng nhau qua đường thẳng $t = \frac{1}{2}$.
3. $B(x,1) = \int_0^1 t^{x-1}dt = \dfrac{1}{x}$ với mọi số thực $x > 0$.
4. $B(\dfrac{1}{2}, \dfrac{n}{2}) = 2\int_0^{\pi/2} \cos^{n-1} u\,du = \int_{-\pi/2}^{\pi/2} \cos^{n-1} u\,du$ với $\forall n \in \mathbb{N}^*$.

   Thật vậy với phép đổi biến $t = \sin^2 u$

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \int_0^{\pi/2} \sin^{2x-2} u \cos^{2y-2} u \cdot 2\sin u \cos u\,du$$
$$= 2\int_0^{\pi/2} \sin^{2x-1} u \cos^{2y-1} u\,du.$$

   Thay $x = \frac{1}{2}, y = \frac{n}{2}$ ta được $B(\dfrac{1}{2}, \dfrac{n}{2}) = 2\int_0^{\pi/2} \cos^{n-1} u\,du$.

5. Với $x > 0, y > 1$

$$B(x,y) = \frac{y-1}{x+y-1} B(x,y-1).$$

Thật vậy sử dụng tích phân từng phần

$$\begin{aligned} B(x,y) &= \frac{y-1}{x} \int_0^1 t^x (1-t)^{y-2} dt \\ &= \frac{y-1}{x} \int_0^1 t^{x-1} (1-t)^{y-2} dt - \frac{y-1}{x} \int_0^1 t^{x-1} (1-t)^{y-1} dt \\ &= \frac{y-1}{x} B(x,y-1) - \frac{y-1}{x} B(x,y). \end{aligned}$$

Suy ra $\frac{x+y-1}{x} B(x,y) = \frac{y-1}{x} B(x,y-1)$. Từ đây suy ra đ.p.c.m.

6. Với số thực $x > 0$ và y là số tự nhiên bất kì, sử dụng tính chất trên liên tiếp và tính chất 3, ta được

$$B(x,y) = \frac{y-1}{x+y-1} B(x,y-1) = \frac{y-1}{x+y-1} \cdot \frac{y-2}{x+y-2} B(x,y-2) = ..$$

$$= \frac{y-1}{x+y-1} \cdot \frac{y-2}{x+y-2} .. \frac{2}{x+2} \cdot \frac{1}{x+1} B(x,1) = \frac{(y-1)!}{x(x+1)...(x+y-1)}$$

7. Với m, n là hai số tự nhiên bất kì, theo tính chất trên và tính chất hàm Gamma

$$B(m,n) = \frac{(m-1)!(n-1)!}{(m+n-1)!} = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}.$$

8. Ta thừa nhận kết quả sau với mọi số thực $x > 0, y > 0$

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Tính chất này là sự mở rộng tính chất 7 ở trên cho hai số thực dương bất kì.

5.2.2 Hàm phân bố Gamma

Trong mục này chúng ta sẽ cần đến 2 định lí được suy ra từ định lí 3.4.3 trong chương III, phần lí thuyết xác suất nói về hàm mật độ và mật độ chung của các đại lượng ngẫu nhiên. Ta viết lại 2 định lí này trong phần này để người đọc tiện so sánh, theo dõi.

Định lí 5.2.1 *Giả sử $f(x)$ là hàm mật độ của ξ. Khi đó hàm mật độ của $\eta = \varphi(\xi)$ bằng*

$$g(y) = f\left(\varphi^{-1}(y)\right) \cdot \left|\left(\varphi^{-1}(y)\right)'\right|.$$

Giả thiết rằng φ là một song ánh và khả vi trên miền giá trị của đại lượng ngẫu nhiên ξ.

Chẳng hạn nếu $y = \varphi(x) = ax + b$ $(a \neq 0)$ là hàm bậc nhất và ξ là đại lượng ngẫu nhiên với $f(x)$ là hàm mật độ của ξ. Khi đó $\varphi^{-1}(y) = \dfrac{y-b}{a}$ và hàm mật độ $g(y)$ của $\eta = \varphi(\xi)$ theo định lí trên

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right).$$

Tương tự, ta có kết quả sau cho đại lượng ngẫu nhiên 2 chiều

Định lí 5.2.2 *Giả sử φ là một song ánh*

$$\varphi : D \to T \quad D \subset R^2, T \subset R^2$$

khả vi tại mọi điểm thuộc miền D. (X, Y) là véc tơ ngẫu nhiên nhận các giá trị trong D và $h(x, y)$ là hàm mật độ của véc tơ ngẫu nhiên đó. Khi đó hàm mật độ của $(U, V) = \varphi(X, Y)$ bằng

$$g(u, v) = h\left(\varphi^{-1}(u, v)\right) \cdot |J(u, v)|$$

trong đó $J(u, v)$ là Jacobien của φ^{-1}.

Chú ý rằng hàm mật độ của véc tơ ngẫu nhiên (X, Y) còn được gọi là mật độ *đồng thời* hoặc *mật độ chung* của hai đại lượng ngẫu nhiên X và Y.

Jacobien của φ^{-1} được xác định như sau:
Kí hiệu $(x, y) = \varphi^{-1}(u, v)$, Jacobien của (x, y) theo (u, v)

$$J(u,v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Chứng minh. Định lí được chứng minh dựa trên định lí đổi biến trong tích phân kép. Xét $E \subset D$ là tập con bất kì của D. Sử dụng phép đổi biến $(x, y) = \varphi^{-1}(u, v)$ ta có xác suất để điểm ngẫu nhiên (U, V) thuộc tập E bằng

$$\begin{aligned} P((U,V) \in E) = P((X,Y) \in \varphi^{-1}(E)) &= \iint_{\varphi^{-1}(E)} h(x,y)\,dxdy \\ &= \iint_E h(\varphi^{-1}(u,v)) \cdot |J(u,v)|\,dudv. \end{aligned}$$

Do $E \subset D$ là tập con bất kì của D suy ra $h\left(\varphi^{-1}(u, v)\right) \cdot |J(u, v)|$ là hàm *mật độ chung* của U và V. ■

Nhận xét rằng sử dụng định lí 5.2.1, ta có thể dễ dàng tìm được hàm mật độ của $Y = X^2$ với X là đại lượng ngẫu nhiên có phân bố chuẩn $X \in N(0, 1)$.

Thật vậy, hàm mật độ của X : $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, suy ra mật độ của

$$\xi = |X| : \ f(x) = \begin{cases} \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} & \text{nếu } x > 0 \\ 0 & \text{nếu } x \leq 0 \end{cases}.$$

Áp dụng định lí 1.2.1, ta được hàm mật độ của $Y = X^2 = \xi^2$

$$g(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} y^{-\frac{1}{2}} \text{ với } \ y > 0.$$

Hàm mật độ $g(y)$ ở trên là trường hợp đặc biệt của phân bố Gamma được định nghĩa dưới đây:

Định nghĩa 5.2.3 *Đại lượng ngẫu nhiên X được gọi là đại lượng ngẫu nhiên có phân bố Gamma nếu X có hàm mật độ*

$$G(x,\alpha,p)=\frac{\alpha^p}{\Gamma(p)}\cdot e^{-\alpha x}x^{p-1},\quad \alpha>0,p>0,x>0$$

trong đó $\alpha>0,p>0$ là 2 tham số dương, $x>0$ là biến của hàm mật độ $G(x,\alpha,p)$.

Hàm mật độ của phân bố Gamma có thể viết dưới dạng khác đầy đủ hơn

$$G(x,\alpha,p)=\begin{cases}\frac{\alpha^p}{\Gamma(p)}\cdot e^{-\alpha x}x^{p-1} & \text{nếu } x>0\\ \\ 0 & \text{nếu } x\le 0.\end{cases}$$

Nhận xét rằng tích phân $\displaystyle\int_0^{+\infty}e^{-\alpha x}x^{p-1}dx=\frac{\Gamma(p)}{\alpha^p}$, suy ra hàm mật độ luôn không âm và có tích phân trên $\mathbb{R}$ bằng 1. Do tính chất này của hàm mật độ, từ nay về sau ta viết hàm mật độ của phân bố Gamma dưới dạng gọn hơn

$$G(x,\alpha,p)=c\cdot e^{-\alpha x}x^{p-1},\ \text{trong đó } c \text{ là hằng số thích hợp.}$$

Hằng số c trong công thức trên bằng giá trị của tích phân $\displaystyle\int_0^{+\infty}e^{-\alpha x}x^{p-1}dx$ và để thuận tiện từ nay về sau ta kí hiệu $X\in G(\alpha,p)$ để nói X là đại lượng ngẫu nhiên có phân bố Gamma với 2 tham số α và p.

Mô men cấp k của phân bố Gamma

$$m_k=\int_0^{+\infty}x^k\frac{\alpha^p}{\Gamma(p)}\cdot e^{-\alpha x}x^{p-1}dx=\int_0^{+\infty}\frac{\alpha^p}{\Gamma(p)}\cdot e^{-\alpha x}x^{k+p-1}dx=\frac{\Gamma(p+k)}{\alpha^k\Gamma(p)}.$$

Vì vậy kì vọng và phương sai của phân bố Gamma lần lượt bằng

$$m=\frac{p}{\alpha},\quad \sigma^2=m_2-m_1^2=\frac{\Gamma(p+2)}{\alpha^2\Gamma(p)}-\frac{p^2}{\alpha^2}=\frac{p}{\alpha^2}. \tag{5.1}$$

Định lí ngay sau đây sẽ là cơ sở để ta trình bày tiếp các phân bố thường gặp hơn (phân bố χ^2, phân bố F, phân bố t) trong thống kê.

Định lí 5.2.3 *Nếu $X \in G(\alpha, p_1), Y \in G(\alpha, p_2)$ là 2 đại lượng ngẫu nhiên độc lập có cùng tham số α, khi đó $r = X + Y$ và $f = \frac{X}{Y}$ cũng độc lập. Ngoài ra $r \in G(\alpha, p_1 + p_2)$ và hàm mật độ của f bằng*

$$\frac{\Gamma(p_1 + p_2)}{\Gamma(p_1)\Gamma(p_2)} \cdot \frac{f^{p_1 - 1}}{(1 + f)^{p_1 + p_2}}.$$

Chứng minh. Hàm mật độ của (X, Y) bằng

$$c \cdot e^{-\alpha x - \alpha y} x^{p_1 - 1} y^{p_2 - 1}.$$

Đổi biến $x = r \sin^2 \varphi, y = r \cos^2 \varphi, \quad 0 < r < +\infty, 0 < \varphi < \frac{\pi}{2}$, khi đó Jacobien của (x, y) bằng $J(r, \varphi) = r \sin 2\varphi$. Theo định lí 5.2.2, mật độ của (r, φ) bằng

$$c' \cdot e^{-\alpha r} r^{p_1 + p_2 - 1} (\sin \varphi)^{2p_1 - 1} (\cos \varphi)^{2p_2 - 1}, \tag{5.2}$$

điều đó chứng tỏ r và φ độc lập. Suy ra $r = X + Y$ và $f = \frac{X}{Y} = tg^2 \varphi$ cũng độc lập. Từ biểu thức (5.2) hiển nhiên $r \in G(\alpha, p_1 + p_2)$.

Cũng từ hàm mật độ chung trong biểu thức (5.2), hàm mật độ của φ có dạng

$$c \cdot (\sin \varphi)^{2p_1 - 1} (\cos \varphi)^{2p_2 - 1}. \tag{5.3}$$

Để xác định hàm mật độ của f, ta sử dụng định lí 5.2.1 và đổi biến $\varphi = \operatorname{arctg} \sqrt{f}$ (hay $f = \operatorname{tg}^2 \varphi$

$$\cos \varphi = \sqrt{\frac{1}{1 + tg^2 \varphi}} = \sqrt{\frac{1}{1 + f}}, \qquad \sin \varphi = tg\varphi \sqrt{\frac{1}{1 + tg^2 \varphi}} = \sqrt{\frac{f}{1 + f}}$$

thay vào (5.3), ta thu được hàm mật độ của f bằng

$$c \cdot \frac{f^{p_1 - 1}}{(1 + f)^{p_1 + p_2}}$$

Để tính chính xác hệ số c trong biểu thức hàm mật độ của f, ta sử dụng phép biến đổi $u = \dfrac{1}{1 + f}$, khi đó

$$c = \int_0^\infty \frac{f^{p_1 - 1}}{(1 + f)^{p_1 + p_2}} df = \int_0^1 u^{p_2 - 1} (1 - u)^{p_1 - 1} du = B(p_1, p_2) = \frac{\Gamma(p_1)\Gamma(p_2)}{\Gamma(p_1 + p_2)}.$$

Vậy hàm mật độ của $f = \frac{X}{Y}$ (thương của 2 đại lượng ngẫu nhiên độc lập có phân bố Gamma và cùng chung tham số α) bằng

$$\frac{\Gamma(p_1+p_2)}{\Gamma(p_1)\Gamma(p_2)} \cdot \frac{f^{p_1-1}}{(1+f)^{p_1+p_2}}. \text{ đ.p.c.m.} \qquad \blacksquare$$

Trong mục này chúng ta sẽ lần lượt dẫn ra biểu thức hàm mật độ của các phân bố thường hay gặp nhất trong thống kê. Đó là phân bố χ^2, phân bố F và phân bố Student.

5.2.3 Phân bố χ^2

Nếu $X_i \in N(0,1),\ i = 1, 2, ..., n$ là n đại lượng ngẫu nhiên độc lập có cùng phân bố chuẩn, khi đó phân bố của $X_1^2 + X_2^2 + \cdots + X_n^2$ được gọi là *phân bố χ^2 với n bậc tự do*. Người ta thường kí hiệu $\chi^2(n)$ (hoặc χ_n^2) là lớp các đại lượng ngẫu nhiên có phân bố χ^2 với n bậc tự do.

Trong mục trước chúng ta đã chỉ ra nếu $X_i \in N(0,1)$ thì X_i^2 là đại lượng ngẫu nhiên có phân bố Gamma với các tham số $\alpha = \frac{1}{2}$ và $p = \frac{1}{2}$. Theo phần đầu định lí 5.2.3, $X_1^2 + X_2^2 + \cdots + X_n^2$ cũng có phân bố Gamma với cùng tham số $\alpha = \frac{1}{2}$, tham số còn lại $p = \underbrace{\frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2}}_{n \text{ số hạng}} = \frac{n}{2}$. Từ đây suy ra hàm mật độ của phân bố χ^2 với n bậc tự do chính là phân bố Gamma với các tham số $\alpha = \frac{1}{2}$ và $p = \frac{n}{2}$

$$G(x, \frac{1}{2}, \frac{n}{2}) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \cdot e^{-\frac{x}{2}} x^{\frac{n}{2}-1}, \quad x > 0.$$

Do đẳng thức (5.1), kì vọng và phương sai của phân bố $\chi^2(n)$ lần lượt bằng

$$E(\chi_n^2) = n, \quad D(\chi_n^2) = 2n.$$

Nhận xét sau khá quan trọng trong thực hành. Theo định lí giới hạn trung tâm, hàm phân bố của đại lượng ngẫu nhiên dưới đây, với $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$,

$$\frac{\chi^2 - n}{\sqrt{2n}}$$

tiến dần tới hàm phân bố chuẩn $\Phi(x)$*, nói cách khác* χ^2 *xấp xỉ phân bố chuẩn* $N(n, (\sqrt{2n})^2)$*. Mặt khác*

$$P\left(\sqrt{2\chi^2} - \sqrt{2n} < x\right) = P\left(\frac{\chi^2 - n}{\sqrt{2n}} < x + \frac{x^2}{2\sqrt{2n}}\right) \to \Phi(x) \quad \text{khi} \quad n \to \infty$$

Vậy $\sqrt{2\chi^2}$ *xấp xỉ phân bố chuẩn* $N(\sqrt{2n}, 1)$*. Tuy nhiên người ta chứng minh được xấp xỉ sau còn tốt hơn, do vậy nó thường được sử dụng hơn trong thực hành*

$$\sqrt{2\chi^2} \approx N(\sqrt{2n-1}, 1).$$

5.2.4 Phân bố F

Nếu $X_1 \in \chi^2(m), X_2 \in \chi^2(n)$ là hai đại lượng ngẫu nhiên độc lập, có phân bố χ^2 với m và n bậc tự do tương ứng, khi đó phân bố của

$$F = \frac{\frac{1}{m}X_1}{\frac{1}{n}X_2}$$

được gọi là *phân bố* F *với* (m, n) *bậc tự do.*

Theo phần thứ hai của định lí 5.2.3, mật độ của $\frac{X_1}{X_2}$ bằng

$$\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{f^{\frac{m}{2}-1}}{(1+f)^{\frac{m+n}{2}}}.$$

Như vậy $\displaystyle\int_0^\infty \frac{f^{\frac{m}{2}-1}}{(1+f)^{\frac{m+n}{2}}}df = \frac{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}{\Gamma(\frac{m+n}{2})}$, suy ra kì vọng của $\frac{X_1}{X_2}$ bằng

$$\begin{aligned} E\left(\frac{X_1}{X_2}\right) &= \int_0^\infty \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{f^{\frac{m}{2}}}{(1+f)^{\frac{m+n}{2}}}df \\ &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty \frac{f^{\frac{m+2}{2}-1}}{(1+f)^{\frac{(m+2)+(n-2)}{2}}}df \\ &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{\Gamma(\frac{m+2}{2})\Gamma(\frac{n-2}{2})}{\Gamma(\frac{m+n}{2})} = \frac{m}{n-2} \end{aligned}$$

Suy ra kì vọng của phân bố F với (m,n) bậc tự do

$$E(F) = \left(\frac{n}{m} \cdot \frac{X_1}{X_2}\right) = \frac{n}{m} \cdot \frac{m}{n-2} = \frac{n}{n-2}.$$

Để tính phương sai của phân bố F với (m,n) bậc tự do, ta có

$$\begin{aligned} E\left(\frac{X_1^2}{X_2^2}\right) &= \int_0^\infty \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{f^2 \cdot f^{\frac{m}{2}-1}}{(1+f)^{\frac{m+n}{2}}}df \\ &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty \frac{f^{\frac{m+4}{2}-1}}{(1+f)^{\frac{(m+4)+(n-4)}{2}}}df \\ &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{\Gamma(\frac{m+4}{2})\Gamma(\frac{n-4}{2})}{\Gamma(\frac{m+n}{2})} = \frac{m(m+2)}{(n-2)(n-4)}. \end{aligned}$$

Vì vậy

$$D\left(\frac{X_1}{X_2}\right) = \frac{m(m+2)}{(n-2)(n-4)} - \frac{m^2}{(n-2)^2} = \frac{m}{n-2} \cdot \frac{2(m+n-2)}{(n-4)(n-2)}.$$

Phương sai của phân bố F với (m,n) bậc tự do, theo đó bằng

$$D(F) = \frac{n^2}{m^2} \cdot \frac{m}{n-2} \cdot \frac{2(m+n-2)}{(n-4)(n-2)} = \frac{2n^2(m+n-2)}{m(n-4)(n-2)^2}.$$

Phân bố F được coi là thương của hai đại lượng ngẫu nhiên độc lập có phân bố χ^2. Áp dụng định lí 5.2.1, mật độ của phân bố F với (m,n) bậc tự do bằng

$$\left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{x^{\frac{m}{2}-1}}{(1+\frac{mx}{n})^{\frac{m+n}{2}}}.$$

5.2.5 Phân bố Student (hay còn gọi là phân bố T hoặc t)

Nếu $X \in \chi^2(n)$ và $Y \in N(0,1)$ là hai đại lượng ngẫu nhiên độc lập, khi đó phân bố của

$$T = \frac{Y}{\sqrt{X}}\sqrt{n}$$

được gọi là *phân bố* T (*hay phân bố Student*) với n bậc tự do. Phân bố đồng thời của (Y, X) bằng

$$c \cdot e^{-\frac{y^2}{2}} e^{-\frac{x}{2}} x^{\frac{n}{2}-1}.$$

Đổi biến $y = r\sin\varphi, x = r^2\cos^2\varphi, \quad 0 < r < +\infty, -\frac{\pi}{2} < \varphi < \frac{\pi}{2}$, khi đó Jacobien của (x, y) bằng $J(r, \varphi) = 2r^2\cos\varphi$. Theo định lí 5.2.2, mật độ của (r, φ) bằng

$$c' \cdot e^{-\frac{r^2}{2}} r^n (\cos\varphi)^{n-1},$$

điều đó chứng tỏ r và φ độc lập. *Chú ý rằng theo tính chất 4 của hàm Beta, hàm mật độ của* φ *bằng*

$$\frac{1}{B(\frac{1}{2}, \frac{n}{2})} \cdot (\cos\varphi)^{n-1}.$$

Để xác định hàm mật độ của T, ta sử dụng định lí 5.2.1 và đổi biến

$$t = \frac{\sqrt{n} \cdot y}{\sqrt{x}} = \sqrt{n}\,\mathrm{tg}\,\varphi \quad \text{hay} \quad \varphi = \mathrm{arctg}\,\frac{t}{\sqrt{n}},$$

ta được hàm mật độ của phân bố T với n bậc tự do

$$S(t, n) = \left[\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)\right]^{-1} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

Sử dụng công thức Sterling ta có thể chứng minh được hàm mật độ $S(t, n)$ *tiến tới mật độ của phân bố chuẩn thuộc lớp* $N(0, 1)$ *khi* $n \to \infty$.

Nếu $\frac{X}{\sigma^2} \in \chi^2(n)$ và $Y \in N(m, \sigma^2)$ độc lập, khi đó

$$T = \frac{Y - m}{\sqrt{X}}\sqrt{n}$$

có phân bố Student với n bậc tự do.

Kí hiệu $S(n)$ là lớp các đại lượng ngẫu nhiên có phân bố Student với n bậc tự do. Phân bố Student là phân bố đối xứng nên nó luôn có *kì vọng bằng 0* (với $n \geq 2$).

Chú ý rằng sử dụng phép biến đổi $u = \frac{t^2}{n}$ và hàm Beta ta cũng có thể tính chính xác hệ số của hàm mật độ của phân bố T với n bậc tự do. Thật vậy xét

$$I = \int_0^\infty \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} dt = \int_0^\infty (1+u)^{-\frac{n+1}{2}} \frac{\sqrt{n}}{2} \cdot \frac{1}{\sqrt{u}} du.$$

Đổi biến tiếp $x = \frac{1}{1+u}$, khi đó

$$I = \frac{\sqrt{n}}{2} \int_0^1 x^{\frac{n}{2}-1}(1-x)^{-\frac{1}{2}} dx = \frac{\sqrt{n}}{2} B(\frac{1}{2}, \frac{n}{2})$$

Vậy hàm mật độ của phân bố T với n bậc tự do có hệ số $\left[\sqrt{n}B(\frac{1}{2}, \frac{n}{2})\right]^{-1}$

$$S(t,n) = \left[\sqrt{n}B\left(\frac{1}{2}, \frac{n}{2}\right)\right]^{-1} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

Để tính phương sai của T, xét tích phân (vẫn sử dụng phép biến đổi $u = \frac{t^2}{n}$)

$$\begin{aligned} I_1 = \int_0^\infty t^2 \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} dt &= \int_0^\infty nu(1+u)^{-\frac{n+1}{2}} \frac{\sqrt{n}}{2} \cdot \frac{1}{\sqrt{u}} du \\ &= \frac{n\sqrt{n}}{2} \int_0^\infty u^{\frac{1}{2}}(1+u)^{-\frac{n+1}{2}} du. \end{aligned}$$

Đổi biến tiếp $x = \frac{1}{1+u}$, khi đó với $n > 2$

$$I_1 = \frac{n\sqrt{n}}{2} \int_0^1 x^{\frac{n}{2}-2}(1-x)^{\frac{1}{2}} dx = \frac{n\sqrt{n}}{2} B(\frac{3}{2}, \frac{n}{2} - 1) = \frac{n\sqrt{\pi n}}{4} \cdot \frac{\Gamma(\frac{n}{2}-1)}{\Gamma(\frac{n+1}{2})}.$$

Phương sai của phân bố T (với $n > 2$)

$$\begin{aligned} D(T) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})} \int_{-\infty}^\infty t^2 \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} dt &= 2 \cdot \frac{n \cdot \Gamma(\frac{n}{2}-1)}{4 \cdot \Gamma(\frac{n}{2})} = \frac{n}{n-2} \\ &= \frac{n\sqrt{\pi n}}{2} \cdot \frac{\Gamma(\frac{n}{2}-1)}{\Gamma(\frac{n+1}{2})} \end{aligned}$$

Nhận xét rằng T^2 có phân bố F với $(1,n)$ bậc tự do, từ kì vọng của F ta cũng suy ra $D(T) = \frac{n}{n-2}$.

5.2.6 Phân bố của trung bình mẫu và phương sai mẫu

Giả thiết $X_i \in N(m, \sigma^2),\ i = 1, 2, ..., n$ là các đại lượng ngẫu nhiên độc lập có cùng phân bố chuẩn (với kì vọng là m và phương sai bằng σ^2). Nói cách khác $(X_1, X_2, ..., X_n)$ là một mẫu ngẫu nhiên đơn giản có phân bố chuẩn. Khi đó người ta đã chỉ ra trong lí thuyết xác suất, *kì vọng mẫu* (hoặc còn gọi là *trung bình mẫu*)

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

cũng có phân bố chuẩn với kì vọng bằng m và phương sai bằng $\frac{\sigma^2}{n}$. Kí hiệu $\overline{X} \in N(m, N\left(m, \frac{\sigma^2}{n}\right)$. Người ta sử dụng kì vọng mẫu làm ước lượng hiệu quả cho giá trị trung bình, tham số m.

Một đặc trưng mẫu thứ hai rất quan trọng trong thống kê, nó được sử dụng làm ước lượng cho tham số σ^2. Đó là *phương sai mẫu*

$$S^2 = \frac{\sum\limits_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

và *phương sai mẫu điều chỉnh*

$$S^{*2} = \frac{\sum\limits_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

Hiển nhiên $\frac{n}{\sigma^2}S^2 = \frac{n-1}{\sigma^2}S^{*2}$. Ta dễ dàng chỉ ra phương sai mẫu điều chỉnh là một ước lượng *không chệch* cho tham số σ^2.

$$E(S^{*2}) = \frac{1}{n-1}\sum_{i=1}^{n} E(X_i^2) - \frac{n}{n-1}E(\overline{X}^2) = \frac{n\sigma^2 - \sigma^2}{n-1} = \sigma^2.$$

Định lí sau sẽ chỉ ra phân bố của đặc trưng mẫu thứ hai, phương sai mẫu (cũng như phương sai mẫu điều chỉnh) vừa nói ở trên.

Định lí 5.2.4 *Giả thiết $X_i \in N(m, \sigma^2)$, $i = 1, ..., n$ là các đại lượng ngẫu nhiên độc lập có cùng phân bố chuẩn với kì vọng là m và phương sai bằng σ^2. Khi đó*

$$\frac{n}{\sigma^2}S^2 = \frac{n-1}{\sigma^2}S^{*2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \in \chi^2(n-1).$$

Nói cách khác đại lượng ngẫu nhiên $\frac{n}{\sigma^2}S^2$ có phân bố χ^2 với $n-1$ bậc tự do.

Chứng minh. Kí hiệu $\mathbf{X} = (X_1, ..., X_n)^T$ là véc tơ cột n thành phần và xét phép biến đổi trực giao với $\mathbf{A}$ là ma trận trực giao bất kì sao cho $(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}})$ là hàng thứ nhất của $\mathbf{A}$. Kí hiệu $\mathbf{Y} = (Y_1, ..., Y_n)^T = \mathbf{AX}$ là ảnh của $\mathbf{X}$ qua phép biến đổi trực giao đó ($\mathbf{Y}$ là tích của ma trận trực giao $\mathbf{A}$ và ma trận cột $\mathbf{X}$.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \\ Y_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \\ X_n \end{pmatrix}$$

Từ tính chất của phép biến đổi trực giao ta suy ra các kết quả sau

1. $Y_1 = \dfrac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} = \overline{X}\sqrt{n}.$

2. Do phép biến đổi trực giao bảo toàn độ dài véc tơ nên

$$Y_1^2 + \cdots + Y_n^2 = X_1^2 + \cdots + X_n^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n\overline{X}^2$$

hay

$$Y_1^2 + \cdots + Y_n^2 = X_1^2 + \cdots + X_n^2 = nS^2 + Y_1^2.$$

Suy ra $Y_2^2 + \cdots + Y_n^2 = nS^2 = (n-1)S^{*2}$

3. Kí hiệu $\mathbf{m} = (m, m, ..., m)^T$ là véc tơ cột có các thành phần là kì vọng m. Khi đó $\mathbf{m}$ trực giao với các hàng thứ hai, thứ 3,.., thứ n của ma trận $\mathbf{A}$, ta có

$$\mathbf{A}(\mathbf{X} - \mathbf{m}) = \mathbf{Y} - (m\sqrt{n}, 0, ..., 0)^T = (Y_1 - m\sqrt{n}, Y_2, ..., Y_n)^T.$$

Suy ra

$$(Y_1 - m\sqrt{n})^2 + Y_2 + \cdots + Y_n^2 = (X_1 - m)^2 + (X_2 - m)^2 + \cdots + (X_n - m)^2.$$

Biết hàm mật độ của $\mathbf{X}$ bằng $c \cdot e^{-\frac{\sum (x_i - m)^2}{2\sigma^2}}$. Vậy mật độ của $\mathbf{Y}$ bằng

$$c \cdot e^{-\frac{(y_1 - m\sqrt{n})^2 + y_2^2 + \cdots + y_n^2}{2\sigma^2}}.$$

Điều đó chứng tỏ $Y_1 = \overline{X}\sqrt{n} \in N(m\sqrt{n}, \sigma^2), Y_i \in N(0, \sigma^2), i = 2, ..., n$ độc lập và do đó

$$\frac{Y_2^2 + \cdots + Y_n^2}{\sigma^2} = \frac{nS^2}{\sigma^2} = \frac{(n-1)S^{*2}}{\sigma^2} \in \chi^2(n-1)$$

là đại lượng ngẫu nhiên có phân bố χ^2 với $n - 1$ bậc tự do. ■

Theo chứng minh trên $Y_1, Y_2, ..., Y_n$ là các đại lượng ngẫu nhiên độc lập có phân bố chuẩn. Ta suy ra hệ quả quan trọng sau

Hệ quả 5.2.1 *Đại lượng ngẫu nhiên*

$$\frac{\overline{X} - m}{S^*}\sqrt{n} = \frac{\overline{X} - m}{S}\sqrt{n-1}$$

có phân bố Student với $n - 1$ bậc tự do.

Thật vậy $\dfrac{\overline{X} - m}{S}\sqrt{n-1} = \sqrt{n-1}\dfrac{\overline{X} - m}{\sigma}\sqrt{n} : \dfrac{S\sqrt{n}}{\sigma}$ là thương của 2 đại lượng ngẫu nhiên độc lập $\frac{\overline{X} - m}{\sigma}\sqrt{n} \in N(0,1)$ và $\frac{nS^2}{\sigma^2} \in \chi^2(n-1)$. Do vậy đại lượng ngẫu nhiên trong hệ quả, kí hiệu $T = \dfrac{\overline{X} - m}{S}\sqrt{n-1}$ có phân bố Student với $n - 1$ bậc tự do.

Tổng quát hơn nếu giả thiết các đại lượng ngẫu nhiên

$$X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n$$

độc lập nhau, $(X_1, X_2, ..., X_m)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $X \in N(m_1, \sigma^2)$ và $(Y_1, Y_2, ..., Y_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $Y \in N(m_1, \sigma^2)$. Khi đó

$$t = \sqrt{\frac{(m+n-2)mn}{m+n}} \cdot \frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sqrt{nS_X^2 + mS_Y^2}}$$

có phân bố Student với $m + n - 2$ bậc tự do.

5.2.7 Phụ lục

Ta nghiên cứu thêm một phân bố khác, nó đặt cơ sở cho việc xử lí dữ liệu (các phần tử mẫu) trong các bài toán thống kê. Giả thiết $u_1, u_2, ..., u_n \in N(0, \sigma^2)$ độc lập. Khi đó đại lượng ngẫu nhiên $\tau = \dfrac{u_1}{\sqrt{\frac{1}{n}\sum_1^n u_i^2}}$ $(|\tau| \leq \sqrt{n})$ không là phân bố Student (tử số và mẫu số không độc lập). Tuy nhiên ta có thể xác định được hàm mật độ của τ. Đại lượng ngẫu nhiên

$$v = \sqrt{\frac{n-1}{n}} \cdot \frac{\tau}{\sqrt{1 - \frac{\tau^2}{n}}} = \frac{u_1}{\sqrt{\frac{1}{n-1}\sum_{i=2}^n u_i^2}}$$

có phân bố Student với $n-1$ bậc tự do. Khi τ tăng từ $-\sqrt{n}$ đến $\sqrt{n}$, biến v cũng tăng từ $-\infty$ tới ∞, do đó ta có thể xác định được hàm mật độ của τ (thông qua hàm phân bố của v). Hàm phân bố của τ bằng

$$P(\tau < x) = P\left(v < \sqrt{\frac{n-1}{n}} \cdot \frac{x}{\sqrt{1 - \frac{x^2}{n}}}\right)$$

Đạo hàm theo x ta được hàm mật độ của τ

$$\frac{1}{\sqrt{n\pi}} \cdot \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \left(1 - \frac{x^2}{n}\right)^{\frac{n-3}{2}}.$$

Định lí 5.2.5 *Với kí hiệu $\tau = \frac{X_1 - \overline{X}}{S}$, khi đó*

$$\frac{\tau\sqrt{n-2}}{\sqrt{n-1-\tau^2}}$$

có phân bố Student với $n-2$ bậc tự do.

Chứng minh. Để chứng minh định lí, xét phép biến đổi trực giao $\mathbf{Y} = \mathbf{AX}$ với

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \sqrt{\frac{n-1}{n}} & \frac{-1}{\sqrt{n(n-1)}} & \cdots & \frac{-1}{\sqrt{n(n-1)}} \\ * & * & * & * \\ * & * & * & * \end{pmatrix}$$

Như đã biết ở trên $Y_1 = \overline{X}\sqrt{n}$, $Y_2^2 + \cdots + Y_n^2 = nS^2$ và $Y_i \in N(0, \sigma^2), i = 2, ..., n$ độc lập. Đồng thời

$$Y_2 = \sqrt{\frac{n-1}{n}}X_1 - \frac{1}{\sqrt{n(n-1)}}X_2 - \cdots - \frac{1}{\sqrt{n(n-1)}}X_n = \sqrt{\frac{n}{n-1}}(X_1 - \overline{X})$$

Hiển nhiên

$$\frac{Y_2}{\sqrt{\frac{1}{n-1}\sum_2^n Y_i^2}} = \frac{X_1 - \overline{X}}{S} = \tau \quad (|\tau| \leq \sqrt{n-1}).$$

Mặc dù tử số và mẫu số của τ không độc lập nhau, xét thống kê

$$v = \sqrt{\frac{n-2}{n-1}} \cdot \frac{\tau}{\sqrt{1 - \frac{\tau^2}{n-1}}} = \frac{\tau\sqrt{n-2}}{\sqrt{n-1-\tau^2}} = \frac{Y_2}{\sqrt{\frac{1}{n-2}\sum_{i=3}^n Y_i^2}} \in S(n-2)$$

Hiển nhiên các đại lượng ngẫu nhiên

$$\frac{X_i - \overline{X}}{S}, \quad i = 1, 2, ..., n$$

có cùng phân bố như phân bố của τ. Người ta sử dụng kết quả trên để bỏ đi những phần tử mẫu nằm quá xa giá trị trung bình mẫu.

Tổng quát hơn, kí hiệu $\overline{X}_k = \frac{X_1 + X_2 + \cdots + X_k}{k}$ và $\tau_k = \frac{\overline{X}_k - \overline{X}}{S}$, khi đó

$$\tau_k\sqrt{\frac{k(n-1)}{n-k}} \qquad \text{có cùng phân bố như của} \quad \tau = \frac{X_1 - \overline{X}}{S}.$$

Suy ra

$$\frac{\tau_k\sqrt{k(n-2)}}{\sqrt{n-k-k\tau_k^2}}$$

có phân bố Student với $n-2$ bậc tự do.

5.3 Ước lượng thống kê

5.3.1 Khái niệm về ước lượng

Giả sử chúng ta đã biết dạng phân bố của đại lượng ngẫu nhiên X (chẳng hạn X có phân bố mũ hay phân bố nhị thức,...). Gọi $F(x, \Theta)$ (Θ là tham số) là hàm phân bố của X. Nói chung ta chưa biết giá trị thực của tham số Θ. Chúng ta cần xác định tham số đó dựa trên mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ của X. Nói cách khác ta sử dụng một hàm

$$\widehat{\Theta} = \widehat{\Theta}(X_1, X_2, ..., X_n)$$

của mẫu ngẫu nhiên làm ước lượng cho tham số "thực" của Θ.

Định nghĩa 5.3.1

$\widehat{\Theta}$ được gọi là ước lượng không chệch của Θ, nếu kì vọng của $\widehat{\Theta}$ bằng Θ với bất kì giá trị nào của Θ:

$$E(\widehat{\Theta}) = E(\widehat{\Theta}(X_1, X_2, ..., X_n)) = \Theta.$$

Nếu $\widehat{\Theta_1}$ và $\widehat{\Theta_2}$ là hai ước lượng không chệch của Θ và

$$D(\widehat{\Theta_1}) < D(\widehat{\Theta_2}),$$

khi đó ta nói ước lượng $\widehat{\Theta_1}$ hiệu quả hơn ước lượng $\widehat{\Theta_2}$

Hiển nhiên trong hai ước lượng không chệch $\widehat{\Theta_1}$ và $\widehat{\Theta_2}$, ta coi ước lượng nào có phương sai bé hơn (hiệu quả hơn) là tốt hơn.

Định nghĩa 5.3.2

Trường hợp tồn tại một ước lượng không chệch $\widehat{\Theta_0}$ của Θ mà phương sai của nó nhỏ nhất trong số tất cả các ước lượng không chệch:

$$D(\widehat{\Theta_0}) \leq D(\widehat{\Theta}) \quad \text{với mọi ước lượng không chệch} \quad \widehat{\Theta}.$$

Khi đó ta gọi ước lượng $\widehat{\Theta_0}$ là ước lượng hiệu quả.

Ví dụ 5.3.1

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên X. Giả thiết rằng tồn tại kì vọng và phương sai

$$E(X) = m, \quad D(X) = \sigma^2.$$

(a) Do

$$E(\overline{X}) = m, E(S^{*2}) = \sigma^2$$

suy ra kì vọng mẫu và phương sai mẫu điều chỉnh là các ước lượng không chệch của m và σ^2

(b) Đặc biệt khi giả thiết $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên có phân bố chuẩn, khi đó người ta chứng minh được rằng kì vọng mẫu $\overline{X}$ là ước lượng hiệu quả của m.

5.3.2 Ước lượng tham số bằng phương pháp hợp lí cực đại

Phương pháp hợp lí cực đại là phương pháp xây dựng các ước lượng thống kê mà các ước lượng đó hội tụ được nhiều tính chất đã nêu trong các định nghĩa 4, định nghĩa 5. Thực chất của phương pháp hợp lí cực đại được mô tả như sau:

Giả sử X là đại lượng ngẫu nhiên rời rạc, kí hiệu

$$P(X = x) = p(x, \Theta),$$

trong đó Θ là tham số cần ước lượng. Gọi $(x_1, x_2, \cdots, x_n)$ là một thể hiện của một mẫu ngẫu nhiên của X. Xét hàm

$$L(\Theta) = p(x_1, \Theta)p(x_2, \Theta)\cdots p(x_n, \Theta)$$

tương ứng với mẫu ngẫu nhiên đó. Khi đó $L(\Theta)$ là xác suất để ta nhận được $(x_1, x_2, \cdots, x_n)$ trong quá trình chọn mẫu. Giả sử hàm $L(\Theta)$ đạt giá trị lớn nhất tại $\Theta = \widehat{\Theta} = \widehat{\Theta}(x_1, x_2, \cdots, x_n)$. Khi đó ta coi thống kê

$$\widehat{\Theta} = \widehat{\Theta}(x_1, x_2, \cdots, x_n)$$

là ước lượng cho giá trị thực của tham số Θ và ước lượng nhận được bằng cách đó được gọi là ước lượng theo *phương pháp hợp lí cực đại*.

Như vậy thực chất của phương pháp hợp lí cực đại là xây dựng các ước lượng dựa trên mẫu ngẫu nhiên $(X_1, X_2, \cdots, X_n)$ sao cho với xác suất lớn nhất có thể mẫu ngẫu nhiên đó nhận giá trị

$$(x_1, x_2, \cdots, x_n)$$

trong quá trình chọn mẫu.

Trường hợp X là đại lượng ngẫu nhiên liên tục với $f(x, \Theta)$ là hàm mật độ, thay cho $p(x_1, \Theta)p(x_2, \Theta)\cdots p(x_n, \Theta)$ ta xét hàm

$$L(\Theta) = f(x_1, \Theta)f(x_2, \Theta)\cdots f(x_n, \Theta)$$

và tương tự như trên, ước lượng

$$\widehat{\Theta} = \widehat{\Theta}(x_1, x_2, \cdots, x_n)$$

được xác định sao cho $L(\Theta)$ đạt giá trị lớn nhất.

Giả sử hàm $L(\Theta)$ khả vi và đạt giá trị lớn nhất. Để tìm giá trị $\Theta = \widehat{\Theta}$ mà tại đó hàm đạt giá trị lớn nhất, do tính đơn điệu của hàm $\log$, thay cho $L(\Theta)$ ta xét hàm $\log L(\Theta)$. Bằng phương pháp xét cực trị hàm số đã quen thuộc trong giải tích, $\Theta = \widehat{\Theta}$ là nghiệm của phương trình

$$\frac{\partial \log L(\Theta)}{\partial \Theta} = \sum_{i=1}^{n} \frac{\partial \log f(x_i, \Theta)}{\partial \Theta} = 0.$$

Phương trình trên được gọi là *phương trình hợp lí*, trong đó Θ là ẩn cần xác định. Trường hợp X là đại lượng ngẫu nhiên rời rạc, phương trình hợp lí có dạng

$$\sum_{i=1}^{n} \frac{\partial \log p(x_i, \Theta)}{\partial \Theta} = 0.$$

Để thấy rõ ý nghĩa của ước lượng được tìm bằng phương pháp hợp lí cực đại, ta pháp biểu, không chứng minh định lí sau

Định lí 5.3.1 *Nếu tồn tại ước lượng hiệu quả $\widehat{\Theta}$ của tham số Θ, khi đó phương trình hợp lí có duy nhất một nghiệm và nghiệm đó trùng với ước lượng hiệu quả $\widehat{\Theta}$.*

Ta có nhận xét rằng trường hợp tham số Θ là véc tơ, ví dụ $\Theta = (\Theta_1, \Theta_2, \cdots, \Theta_k)$ (hay ta còn nói hàm phân bố của X phụ thuộc k tham số $(\Theta_1, \Theta_2, \cdots, \Theta_k)$), phương pháp ước lượng hợp lí cực đại về cơ bản không thay đổi. Khi đó thay cho việc giải phương trình hợp lí $\frac{\partial \log L(\Theta)}{\partial \Theta} = 0$ ta phải giải hệ phương trình hợp lí sau

$$\frac{\partial \log L(\Theta)}{\partial \Theta_1} = 0; \cdots; \frac{\partial \log L(\Theta)}{\partial \Theta_k} = 0.$$

Để minh hoạ phương pháp tìm ước lượng hợp lí cực đại, ta xét các ví dụ sau

Ví dụ 5.3.2

Hãy tìm ước lượng hợp lí cực đại cho tham số Θ $(0 < \Theta < 1)$ của phân bố nhị thức

$$P(x, \Theta) = C_x^N \Theta^x (1 - \Theta)^{N-x}, x = 0, 1, ..., N.$$

Hàm hợp lí

$$L(\Theta) = \prod_{i=1}^{n} C_{x_i}^N \Theta^{x_i} (1 - \Theta)^{N - x_i},$$

suy ra phương trình hợp lí

$$\begin{aligned}\frac{\partial \log L(\Theta)}{\partial \Theta} &= \frac{\partial}{\partial \Theta} \sum_{i=1}^{n} \left(\log C_{x_i}^N + x_i \log \Theta + (N - x_i) \log(1 - \Theta) \right) = \\ &= \left(\frac{x_i}{\Theta} - \frac{N - x_i}{1 - \Theta} \right) = \sum_{i=1}^{n} \frac{(1 - \Theta)x_i - \Theta(N - x_i)}{\Theta(1 - \Theta)} = \\ &= \frac{1}{\Theta(1 - \Theta)} \sum_{i=1}^{n} (x_i - \Theta N) = 0.\end{aligned}$$

Vậy nghiệm của phương trình hợp lí

$$\widehat{\Theta} = \frac{x_1 + x_2 + \cdots + x_n}{nN}$$

chính là ước lượng hợp lí cực đại cho tham số Θ của phân bố nhị thức.

Chú ý rằng theo ví dụ 1, $\frac{x_1+x_2+\cdots+x_n}{n}$ là ước lượng không chệch của kì vọng của phân bố nhị thức. Mặt khác phân bố nhị thức có kì vọng bằng $N\Theta$, suy ra ước lượng không chệch của Θ bằng

$$\widehat{\Theta} = \frac{x_1 + x_2 + \cdots + x_n}{nN},$$

hoàn toàn phù hợp với ước lượng của Θ tìm bằng phương pháp hợp lí cực đại trong ví dụ này.

Ví dụ 5.3.3

Hãy tìm ước lượng hợp lí cực đại cho tham số m và σ^2 của phân bố chuẩn. Kí hiệu $\Theta = \sigma^2$, hàm mật độ của phân bố chuẩn

$$f(x, m, \Theta) = \frac{1}{\sqrt{2\pi\Theta}} e^{-\frac{(x-m)^2}{2\Theta}}.$$

Suy ra lôgarit của hàm hợp lí bằng

$$\log L(m, \Theta) = \log \prod_{i=1}^{n} f(x_i, m, \Theta) =$$

$$-\frac{n}{2}(\log 2\pi + \log \Theta) - \sum_{i=1}^{n} \frac{(x_i - m)^2}{2\Theta}.$$

Đạo hàm riêng theo m và Θ ta được hệ các phương trình hợp lí sau

$$\begin{cases} \frac{\partial \log L(m,\Theta)}{\partial m} & = \sum_{i=1}^{n} \frac{x_i - m}{\Theta} = 0 \\ \frac{\partial \log L(m,\Theta)}{\partial \Theta} & = -\frac{n}{2\Theta} + \sum_{i=1}^{n} \frac{(x_i-m)^2}{2\Theta^2} = 0. \end{cases}$$

Dễ dàng tính được nghiệm của hệ

$$\widehat{m} = \sum_{i=1}^{n} \frac{x_i}{n} = \overline{X}$$

và

$$\widehat{\Theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{m})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = S^2.$$

Như vậy các ước lượng hợp lí cực đại cho m và σ^2 cũng chính là các kì vọng mẫu và phương sai mẫu mà ta đã xét trong mục 5.1

5.3.3 Ước lượng khoảng tin cậy cho kì vọng (giá trị trung bình) của phân bố chuẩn

Trường hợp σ đã biết

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $X \in N(m, \sigma^2)$, trong đó hai tham số m chưa biết cần ước lượng và σ^2 đã biết. Chúng ta cần tìm khoảng tin cậy cho tham số m (giá trị trung bình của X).

Ta đã biết

$$u = \frac{\overline{X} - m}{\sigma}\sqrt{n}$$

là đại lượng ngẫu nhiên có phân bố chuẩn N(0,1). Tra bảng phân vị của phân bố chuẩn $u \in N(0,1)$, ta tìm được $u_\alpha > 0$ sao cho:

$$P(|u| > u_\alpha) = \alpha.$$

Điều này tương đương với

$$P(u < u_\alpha) = 1 - \frac{\alpha}{2} \quad \text{hoặc} \quad P(u < -u_\alpha) = \frac{\alpha}{2}.$$

hay

$$\Phi(u_\alpha) = 1 - \frac{\alpha}{2} \quad \text{trong đó} \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

Khi đó

$$P\left(\left|\frac{\overline{X} - m}{\sigma}\sqrt{n}\right| < u_\alpha\right) = 1 - \alpha$$

hay

$$P\left(\overline{X} - u_\alpha\frac{\sigma}{\sqrt{n}} < m < \overline{X} + u_\alpha\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Người ta gọi $\left(\overline{X} - u_\alpha\frac{\sigma}{\sqrt{n}}\,;\, \overline{X} + u_\alpha\frac{\sigma}{\sqrt{n}}\right)$ là *khoảng tin cậy* của tham số m và xác suất $1 - \alpha$ là *độ tin cậy* của khoảng đó. Người ta còn viết khoảng tin cậy của m với độ tin cậy $1 - \alpha$ dưới dạng

$$\boxed{\overline{X} - u_\alpha\frac{\sigma}{\sqrt{n}} < m < \overline{X} + u_\alpha\frac{\sigma}{\sqrt{n}}.}$$

Trong thực hành ta thường tìm khoảng tin cậy với độ tin cậy

$$1-\alpha = 0,95; 0,96; 0,97; 0,98; 0,99.$$

Ví dụ 5.3.4

Xét mẫu ngẫu nhiên sau của phân bố chuẩn, giả sử rằng do một lí do nào đó ta biết độ lệch chuẩn $\sigma = 4,09$:

{6,9 ; 10,2 ; 5,3 ; 6,6 ; 18,4 ; 2,0 ; 5,9 ; 10,1 ; 9,6 ; 8,7 ; 6,5 ; 11,1 ; 8,9 ; 9,9 ; 1,6 ; 4,6 ; 2,9 , 13,6 ; 9,9.}

Tính kì vọng mẫu $\overline{X} = 8,03684$. Mẫu trên có kích thước $n = 19$. Để tìm khoảng tin cậy cho giá trị trung bình với độ tin cậy $1-\alpha = 95\%$ (hay $\alpha = 0,05$), tra bảng phân vị của phân bố chuẩn để xác định u_α từ hệ thức

$$\Phi(u_\alpha) = 1 - \frac{\alpha}{2} = 0,975.$$

Phân vị đó bằng

$$u_{0,05} = 1,959961.$$

Áp dụng công thức

$$\overline{X} - u_\alpha \frac{\sigma}{\sqrt{n}} < m < \overline{X} + t_\alpha \frac{\sigma}{\sqrt{n}},$$

trong đó $\sigma = 0,05$ ta có khoảng tin cậy của kì vọng với độ tin cậy 95% bằng

$$(6,197789 \ ; \ 9,875892).$$

Trường hợp σ chưa biết

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $X \in N(m, \sigma^2)$, trong đó hai tham số m và σ^2 đều chưa biết. Chúng ta cần tìm khoảng tin cậy cho tham số m (giá trị trung bình của X) với độ tin cậy $1-\alpha$ cho trước.

Ta đã biết

$$t = \frac{\overline{X} - m}{S^*}\sqrt{n}$$

có phân bố Student với $n-1$ bậc tự do. Tra bảng phân vị của phân bố Student với $n-1$ bậc tự do, ta tìm được $t_\alpha > 0$ sao cho:

$$P(|t| > t_\alpha) = \alpha.$$

Điều này tương đương với

$$P(t < t_\alpha) = 1 - \frac{\alpha}{2} \quad \text{hoặc} \quad P(t < -t_\alpha) = \frac{\alpha}{2}.$$

Khi đó

$$P\left(\left|\frac{\overline{X} - m}{S^*}\sqrt{n}\right| < t_\alpha\right) = 1 - \alpha$$

hay

$$P\left(\overline{X} - t_\alpha\frac{S^*}{\sqrt{n}} < m < \overline{X} + t_\alpha\frac{S^*}{\sqrt{n}}\right) = 1 - \alpha.$$

Người ta gọi $\left(\overline{X} - t_\alpha\frac{S^*}{\sqrt{n}}\,;\, \overline{X} + t_\alpha\frac{S^*}{\sqrt{n}}\right)$ là *khoảng tin cậy* của tham số m và xác suất $1-\alpha$ là *độ tin cậy* của khoảng đó. Người ta còn viết khoảng tin cậy của m với độ tin cậy $1-\alpha$ dưới dạng

$$\boxed{\overline{X} - t_\alpha\frac{S^*}{\sqrt{n}} < m < \overline{X} + t_\alpha\frac{S^*}{\sqrt{n}}}.$$

Các ví dụ sau cho các mẫu ngẫu nhiên tương ứng với phân bố chuẩn, và tìm khoảng tin cậy cho giá trị trung bình của phân bố đó

Ví dụ 5.3.5

Cho mẫu ngẫu nhiên sau của đại lượng ngẫu nhiên có phân bố chuẩn với kích thước mẫu $n = 22$:

{10,0 ; 8,5 ; 6,4 ; 12,8 ; 7,1 ; 3,6 ; 8,1 ; 3,9 ; 13, ; 7,8 ; 10,7 ; 10,6 ; 12,7 ; 11,1 ; 6,1 ; 10,4 ; 8,0 ; 6,9 ; 5,9 ; 10,6 ; 4,9 ; 8,2}

Ta tính các đặc trưng:

Kì vọng mẫu

$$\overline{X} = \frac{1}{22}\sum_{i=1}^{22} X_i = 8,51364$$

và phương sai mẫu điều chỉnh

$$S^{*2} = \frac{1}{21}\sum_{i=1}^{22}(X_i - \overline{X})^2 = (2,78795)^2.$$

vậy độ lệch chuẩn $S^* = 2,78795$.

Để tìm khoảng tin cậy cho giá trị trung bình với độ tin cậy $1-\alpha = 95\%$ (hay $\alpha = 0,05$), tra bảng phân vị của phân bố Student 21 bậc tự do ứng với $\alpha = 0,05$. Phân vị đó bằng

$$t_\alpha = 2,07961.$$

Áp dụng công thức

$$\overline{X} - t_\alpha\frac{S^*}{\sqrt{n}} < m < \overline{X} + t_\alpha\frac{S^*}{\sqrt{n}},$$

ta có khoảng tin cậy của kì vọng với độ tin cậy 95% bằng

$$(7,27753\ ;\ 9,74974).$$

Ví dụ 5.3.6

Cho mẫu ngẫu nhiên sau của đại lượng ngẫu nhiên có phân bố chuẩn với kích thước mẫu $n = 17$:

{5,2 ; 4,2 ; 3,1 ; 5,9 ; 3,3 ; 5,1 ; 10,2 ; 3,8 ; 0,1 ; 2,5 ; -4,6 ; 3,4 ; 0 ; 6,7 ; 7,9 ; 4,4 ; 5,1}

Hãy tìm khoảng tin cậy cho giá trị trung bình của mẫu với độ tin cậy 95% .

Ta tính các đặc trưng:

Kì vọng mẫu

$$\overline{X} = \frac{1}{17}\sum_{i=1}^{17} X_i = 3,9$$

và phương sai mẫu điều chỉnh

$$S^{*2} = \frac{1}{16}\sum_{i=1}^{17}(X_i - \overline{X})^2 = (3,33129)^2.$$

hay độ lệch chuẩn $S^* = 3,33129$.

Để tìm khoảng tin cậy cho giá trị trung bình với độ tin cậy 95% (hay $\alpha = 0,05$) tra bảng phân vị của phân bố Student 16 bậc tự do ứng với $\alpha = 0,05$. Phân vị đó bằng

$$t_\alpha = 2,11991.$$

Áp dụng công thức

$$\overline{X} - t_\alpha \frac{S^*}{\sqrt{n}} < m < \overline{X} + t_\alpha \frac{S^*}{\sqrt{n}}$$

ta có khoảng tin cậy của kì vọng với độ tin cậy95% bằng

$$(2,18721\ ; 5,61279).$$

Ví dụ 5.3.7

Cho mẫu ngẫu nhiên sau của đại lượng ngẫu nhiên có phân bố chuẩn với kích thước mẫu $n = 23$:

{7,0 ; 5,5 ; 5,3 ; 6,6 ; 10,5 ; 5,0 ; 3,3 ; 6,1 ; 4,3 ; 7,6 ; 4,8 ; 5,2 ; 8,0 ; 5,3 ; 5,9 ; 5,7 ; 5,8 ; 8, ; 9,2 ; 4,8 ; 3,8 ; 4,3 ; 5,0.}

Hãy tìm khoảng tin cậy cho giá trị trung bình của mẫu với độ tin cậy 98% .

Ta tính các đặc trưng:

Kì vọng mẫu

$$\overline{X} = \frac{1}{23} \sum_{i=1}^{23} X_i = 5,95652$$

và phương sai mẫu điều chỉnh

$$S^{*2} = \frac{1}{22} \sum_{i=1}^{23} (X_i - \overline{X})^2 = (1,74742)^2.$$

Vậy độ lệch chuẩn $S^* = 1,74742$.

Để tìm khoảng tin cậy cho giá trị trung bình với độ tin cậy 98% ($\alpha = 0,02$) tra bảng phân vị của phân bố Student 22 bậc tự do ứng với $\alpha = 0,02$. Phân vị đó bằng

$$t_\alpha = 2,508323.$$

Áp dụng công thức

$$\overline{X} - t_\alpha \frac{S^*}{\sqrt{n}} < m < \overline{X} + t_\alpha \frac{S^*}{\sqrt{n}}$$

ta có khoảng tin cậy của kì vọng với độ tin cậy 98% bằng

$$(5,02204\ ; 6,89100).$$

Chú ý rằng khoảng tin cậy ứng với độ tin cậy 95% là (5,20088 ; 6,71216). Như vậy khi tăng độ tin cậy khoảng tin cậy cũng lớn theo, vì vậy trong thực hành người ta không chọn độ tin cậy quá gần 1.

Ví dụ 5.3.8

Một ví dụ khác, xét mẫu sau của phân bố chuẩn:

{6,9 ; 10,2 ; 5,3 ; 6,6 ; 18,4 ; 2,0 ; 5,9 ; 10,1 ; 9,6 ; 8,7 ; 6,5 ; 11,1 ; 8,9 ; 9,9 ; 1,6 ; 4,6 ; 2,9 , 13,6 ; 9,9.}

Ta có

$$\overline{X} = 8,03684\ ; S^* = 4,08985.$$

Mẫu trên có kích thước $n = 19$. Phân vị Student 18 bậc tự do với độ tin cậy 95% bằng :

$$t_\alpha = 2,10092.$$

Vậy khoảng tin cậy của kì vọng với độ tin cậy 95% là

$$(6,0656\ ; 10,0081).$$

Ví dụ 5.3.9

Đo chiều cao của 100 thanh niên ở lứa tuổi trưởng thành, kết quả được cho trong bảng dưới đây. Biết chiều cao của thanh niên là đại lượng ngẫu nhiên có phân bố chuẩn. Hãy tìm khoảng tin cậy cho chiều cao trung bình của thanh niên với độ tin cậy 95%

Chiều cao (mét)	số người (n_i)
1,55 -1,57	5
1,58 -1,60	12
1,61 -1,63	26
1,64 -1,66	25
1,67 -1,69	20
1,70 -1,72	7
1,73 -1,75	4
1,76	1
Tổng cộng	100

Trong bảng trên số thanh niên có chiều cao từ 1,55 đến 1, 57 mét bằng 5, số thanh niên có chiều cao từ 1,58 đến 1, 60 mét bằng 12,... Trong thực hành ta coi mẫu ngẫu nhiên trên gồm 100 phần tử mẫu, trong đó:

5 phần tử mẫu bằng nhau và bằng trung điểm của khoảng thứ nhất: $\frac{1,55+1,57}{2} = 1,56$ (mét).

12 phần tử mẫu bằng nhau và bằng trung điểm của khoảng thứ hai: $\frac{1,58+1,60}{2} = 1,59$ (mét)...

Ta có bảng đưới đây để mô tả mẫu ngẫu nhiên nói trên và dựa vào đó để tính các đặc trưng mẫu:

Khoảng	Chiều cao x_i	n_i
1,55 -1,57	1,56	5
1,58 -1,60	1,59	12
1,61 -1,63	1,62	26
1,64 -1,66	1,65	25
1,67 -1,69	1,68	20
1,70 -1,72	1,71	7
1,73 -1,75	1,74	4
1,76	1,76	1
Tổng cộng		100

Ta tính các đặc trưng:

Kì vọng mẫu

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i n_i = \frac{1}{100}(1,56 \cdot 5 + 1,59 \cdot 12 + 1,62 \cdot 26 +$$

$$+1,65 \cdot 25 + 1,68 \cdot 20 + 1,71 \cdot 7 + 1,74 \cdot 4 + 1,76 \cdot 1) = 1,6454$$

và tương tự phương sai mẫu điều chỉnh

$$S^{*2} = \frac{1}{99}\sum_{i=1}^{100}(X_i - \overline{X})^2 = (0,044116925)^2.$$

Vậy độ lệch chuẩn $S^* = 0,044116925$.

Để tìm khoảng tin cậy cho giá trị trung bình với độ tin cậy 95% ($\alpha = 0,05$) ta tính phân vị của phân bố Student 99 bậc tự do ứng với $\alpha = 0,05$ (Chẳng hạn phân vị đó được tính bằng cách tra hàm ngược của hàm phân bố Student trong phần mềm $Excel$ với lệnh $= TINV(0.05, 99)$). Ta được

$$t_\alpha = 1,984217.$$

Áp dụng công thức

$$\overline{X} - t_\alpha \frac{S^*}{\sqrt{n}} < m < \overline{X} + t_\alpha \frac{S^*}{\sqrt{n}}$$

ta có khoảng tin cậy cho chiều cao trung bình của thanh niên với độ tin cậy 95% bằng

$$(1,636646\ ; 1,654154).$$

5.3.4 Khoảng tin cậy cho phương sai của đại lượng ngẫu nhiên có phân bố chuẩn

Gọi

$$(X_1, X_2, ..., X_n)$$

là một mẫu ngẫu nhiên của phân bố chuẩn. Ta đã biết $\frac{nS^2}{\sigma^2}$ có phân bố χ^2 với $n-1$ bậc tự do. Tra bảng phân vị của phân bố χ^2 với $n-1$ bậc tự do

$$P\left(\frac{nS^2}{\sigma^2} > \chi^2_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

và

$$P\left(\frac{nS^2}{\sigma^2} > \chi^2_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$

Khi đó

$$P\left(\chi^2_{1-\frac{\alpha}{2}} < \frac{nS^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Từ đây suy ra khoảng tin cậy của phương sai σ^2 với độ tin cậy $1 - \alpha$ là

$$\boxed{\frac{nS^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}}.}$$

5.3.5 Khoảng tin cậy cho tham số p của đại lượng ngẫu nhiên có phân bố nhị thức (khoảng tin cậy cho xác suất)

Giả sử A là biến cố ngẫu nhiên có xác suất $P(A) = p$ chưa biết. Ta sử dụng ước lượng

$$\widehat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

trong đó X_i bằng 1 hoặc 0 tuỳ theo biến cố A xảy ra hoặc không xảy ra ở phép thử ngẫu nhiên thứ i. Khi đó $n\widehat{p}$ có phân bố nhị thức với

$$E(n\widehat{p}) = np$$

$$D(n\widehat{p}) = npq, \quad q = 1 - p$$

Theo định lí giới hạn trung tâm

$$\frac{n\widehat{p} - np}{\sqrt{npq}} = \sqrt{n}\frac{\widehat{p} - p}{\sqrt{pq}}$$

có phân bố xấp xỉ chuẩn $\approx N(0, 1)$ khi n đủ lớn

$$P\left(\sqrt{n}\frac{\widehat{p} - p}{\sqrt{pq}} \leq \lambda\right) \approx 2\Phi(\lambda) - 1$$

Bất đẳng thức $\sqrt{n}\frac{\widehat{p}-p}{\sqrt{pq}} \leq \lambda$ tương đương với bất phương trình bậc hai

$$(\widehat{p}-p)^2 \leq \lambda^2 \frac{pq}{n}$$

Vậy giải bất phương trình đó ta được nghiệm

$$\widehat{p_1} \leq p \leq \widehat{p_2}$$

trong đó

$$\widehat{p_1} = \frac{\widehat{p} + \frac{\lambda^2}{2n} - \frac{\lambda}{\sqrt{n}}\sqrt{\widehat{p}(1-\widehat{p}) + \frac{\lambda^2}{4n}}}{1+\frac{\lambda^2}{n}} \tag{5.4}$$

$$\widehat{p_2} = \frac{\widehat{p} + \frac{\lambda^2}{2n} + \frac{\lambda}{\sqrt{n}}\sqrt{\widehat{p}(1-\widehat{p}) + \frac{\lambda^2}{4n}}}{1+\frac{\lambda^2}{n}} \tag{5.5}$$

Nhận xét rằng khi n đủ lớn $\frac{1}{n}$ là vô cùng bé cấp cao hơn so với $\frac{1}{\sqrt{n}}$, do vậy nếu bỏ đi các vô cùng bé cấp cao hơn $\frac{1}{\sqrt{n}}$, ta được công thức xấp xỉ sau:

$$\begin{aligned} \widehat{p_1} &\approx \widehat{p} - \frac{\lambda}{\sqrt{n}}\sqrt{\widehat{p}(1-\widehat{p})} \\ \widehat{p_2} &\approx \widehat{p} + \frac{\lambda}{\sqrt{n}}\sqrt{\widehat{p}(1-\widehat{p})} \end{aligned} \tag{5.6}$$

Vậy $(\widehat{p_1};\widehat{p_2})$ là khoảng tin cậy cho tham số p với độ tin cậy $2\Phi(\lambda)-1$:

$$\boxed{P\left(\widehat{p_1} \leq p \leq \widehat{p_2}\right) = 2\Phi(\lambda) - 1.}$$

Ví dụ 5.3.10

Qua điều tra 1500 trẻ em dưới 5 tuổi khi bắt đầu chuyển sang mùa hè, số trẻ em bị viêm phế quản là 620 em. Với độ tin cậy 95% hãy tính khoảng tin cậy cho tỉ lệ bệnh viêm phế quản ở trẻ em.

$$\widehat{p} = \frac{620}{1500} = 0,41333$$

Sử dụng công thức (5.4) với $\lambda = 1,96$ ta được

$$\widehat{p_1} = 0,388665; \quad \widehat{p_2} = 0,438444$$

Nếu sử dụng công thức rút gọn (5.6), ta có kết quả

$$\widehat{p_1} = 0,388413; \quad \widehat{p_2} = 0,438254$$

Ví dụ 5.3.11

Trong đợt vận động bầu cử, điều tra 1600 người có 960 người ủng hộ ứng cử viên A. Tìm khoảng tin cậy với độ tin cậy 95% tỉ lệ số người bỏ phiếu cho A.

Sử dụng công thức (5.4) với $\lambda = 1,96$ ta được khoảng tin cậy với độ tin cậy 95% tỉ lệ số người bỏ phiếu cho A là

$$(0,575783\ ; 0,623738)$$

Chú ý rằng sử dụng công thức rút gọn (5.6) ta được kết quả xấp xỉ

$$(0,576\ ; 0,624).$$

5.3.6 Khoảng tin cậy cho hiệu các giá trị trung bình của phân bố chuẩn

Tương tự như các mục trên, dựa vào nhận xét ở cuối mục 5.2.6, ta có thể đưa ra khoảng tin cậy cho hiệu các giá trị trung bình của phân bố chuẩn.

Giả thiết mẫu $\{X_i\}_{i=1}^m \in N(m_1, \sigma^2), \quad \{Y_i\}_{i=1}^n \in N(m_2, \sigma^2)$, có phân bố chuẩn với phương sai σ^2 chưa biết. Giả thiết các phần tử mẫu đó độc lập nhau. Với độ tin cậy $1 - \alpha$, khoảng tin cậy cho $m_1 - m_2$

$$\boxed{(\overline{X} - \overline{Y}) - S \cdot t_\alpha \sqrt{\frac{m+n}{m \cdot n}} < m_1 - m_2 < (\overline{X} - \overline{Y}) + S \cdot t_\alpha \sqrt{\frac{m+n}{m \cdot n}}}$$

trong đó kí hiệu

$$S^2 = \frac{m \cdot S_X^2 + n \cdot S_Y^2}{m+n-2}$$
$$S_X^2 = \frac{\sum\limits_{i=1}^{m}(X_i - \overline{X})^2}{m}$$
$$S_Y^2 = \frac{\sum\limits_{i=1}^{n}(Y_i - \overline{Y})^2}{n}$$

là các phương sai mẫu tương ứng và t_α là phân vị mức α được xác định từ hệ thức $P(|t| \geq t_\alpha) = \alpha$, t có phân bố Student với $m+n-2$ bậc tự do.

5.4 Kiểm định giả thiết thống kê

Trong mục này ta giả thiết đại lượng ngẫu nhiên X có hàm phân bố $F(x, \Theta)$, trong đó tham số Θ, chưa biết. Người ta đưa ra các giả thiết nào đó về tham số Θ. Chẳng hạn ta xét hai giả thiết có thể về tham số Θ như sau:

(H): $\Theta = \Theta_0$ và

(K): $\Theta \neq \Theta_0$

trong đó Θ_0 là một giá trị xác định nào đó. Giả thiết (H) thường được gọi là *giả thiết không* và giả thiết (K) được gọi là *đối thiết* với giả thiết (H).

Xét một mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ tương ứng với đại lượng ngẫu nhiên X. Nhiệm vụ của kiểm định là dựa vào mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$, xác định một miền $W \subset R^n$ sao cho khi

$(X_1, X_2, ..., X_n) \in W$ ta bác bỏ giả thiết (H)

còn ngược lại nếu

$(X_1, X_2, ..., X_n) \notin W$ ta chưa có cơ sở để bác bỏ giả thiết (H).

Nói cách khác ta tạm thời chấp nhận (H) cho đến khi có thêm các thông tin mới. Miền W như vậy được gọi là *miền bác bỏ* hay còn gọi là *miền tiêu chuẩn*.

Khi bác bỏ hoặc chấp nhận giả thiết (H) ta có thể mắc các sai lầm:

Sai lầm loại I: bác bỏ (H) (hay $(X_1, X_2, ..., X_n) \in W$) trong khi thực ra giả thiết (H) là đúng.

Sai lầm loại II: chấp nhận (H) (hay $(X_1, X_2, ..., X_n) \notin W$) trong khi (H) sai.

Xác suất xảy ra sai lầm loại I chính là xác suất của biến cố: mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ thuộc miền bác bỏ W với điều kiện giả thiết (H) đúng, ta kí hiệu xác suất này

$$P(W/H) = P((X_1, X_2, ..., X_n) \in W/H).$$

Tương tự xác suất xảy ra sai lầm loại II là xác suất của biến cố: mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ không thuộc miền bác bỏ W với điều kiện đối thiết (K) đúng, ta kí hiệu xác suất này

$$P(\overline{W}/K) = P((X_1, X_2, ..., X_n) \notin W/K).$$

Về nguyên tắc chúng ta muốn xây dựng kiểm định (xác định miền bác bỏ W) sao cho cả hai sai lầm loại I và loại II là nhỏ nhất có thể, song khi kích thước mẫu cố định, điều mong muốn đó không thể thực hiện được. Vì vậy người ta thường giới hạn xác suất để sai lầm loại I xảy ra nhỏ hơn hoặc bằng một số α cho trước ($\alpha > 0$ nhỏ tùy ý) và xác định miền bác bỏ W sao cho sai lầm loại II là nhỏ nhất có thể:

$P(\overline{W}/K)$ đạt giá trị nhỏ nhất trong số tất cả các kiểm định mà $P(\overline{W}/K) \leq \alpha$ cho trước.

Số $\alpha > 0$ cho trước nói trên được gọi là *mức ý nghĩa của kiểm định.*

Nhằm mục đích để độc giả tiện theo dõi các lập luận về bài toán kiểm định giả thiết thống kê, chúng ta nghiên cứu chi tiết bài toán kiểm định giả thiết về giá trị trung bình dưới đây. Các bài toán còn lại cũng dựa trên các cơ sở lí luận tương tự.

5.4.1 Kiểm định giả thiết về giá trị trung bình (trường hợp phương sai σ^2 đã biết)

Gỉa sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $X \in N(m, \sigma^2)$, trong đó tham số m chưa biết cần phải kiểm định và σ^2 là tham số đã biết. Lần lượt xét các bài toán sau

Bài toán 1

Chúng ta cần kiểm định:

(H): $m = m_0$ với đối thiết (K): $m \neq m_0$ theo mức ý nghĩa α cho trước (m_0 là giá trị cho trước nào đó). Ta đã biết rằng thống kê

$$u = \frac{\overline{X} - m_0}{\sigma}\sqrt{n}$$

có phân bố chuẩn N(0,1) nếu giả thiết (H): $m = m_0$ đúng. Tra bảng phân vị của phân bố chuẩn N(0,1) để xác định u_α:

$$P(|u| \leq u_\alpha) = 1 - \alpha$$

trong đó $u \in N(0,1)$.

* Nếu $|u| > u_\alpha$ ta bác bỏ giả thiết (H), điều đó dựa trên cơ sở: xác suất để xảy ra biến cố $|u| > u_\alpha$ tương đối nhỏ nếu giả thiết (H): $m = m_0$ đúng:

$$P(|u| > u_\alpha / H) = \alpha$$

** Ngược lại nếu $|u| < u_\alpha$ ta chưa có đủ cơ sở bác bỏ giả thiết (H), nói cách khác mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ không mâu thuẫn gì với giả thiết (H).

Như vậy ở **Bài toán 1 miền bác bỏ** là

$$\boxed{W = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / \left|\frac{\overline{X} - m_0}{\sigma}\sqrt{n}\right| > u_\alpha\}.}$$

và

$$P(|u| > u_\alpha / H) = P(W/H) = \alpha$$

là mức ý nghĩa của kiểm định.

Khi $|u| \leq u_\alpha$, ta chấp nhận (H), điều này tương đương với biến cố

$$m_0 - u_\alpha \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq m_0 - u_\alpha \frac{\sigma}{\sqrt{n}}$$

xảy ra, hay

$$\frac{m_0 - m}{\sigma}\sqrt{n} - u_\alpha \leq \frac{\overline{X} - m}{\sigma}\sqrt{n} \leq \frac{m_0 - m}{\sigma}\sqrt{n} + u_\alpha.$$

Do $\frac{\overline{X}-m}{\sigma}\sqrt{n} \in N(0,1)$, vậy

$$\Phi\left(\frac{m_0-m}{\sigma}\sqrt{n}+u_\alpha\right)-\Phi\left(\frac{m_0-m}{\sigma}\sqrt{n}-u_\alpha\right)$$

là xác suất để xảy ra sai lầm loại II (khi đối thiết (K) đúng). Với kích thước mẫu n tăng dần ra vô hạn

$$\Phi\left(\frac{m_0-m}{\sigma}\sqrt{n}+u_\alpha\right)-\Phi\left(\frac{m_0-m}{\sigma}\sqrt{n}-u_\alpha\right)\to 0.$$

Nói cách khác xác suất để xảy ra sai lầm loại hai có thể làm bé tùy ý khi tăng kích thước mẫu. Một quy tắc kiểm định như vậy còn được gọi là *kiểm định vững*. Chú ý rằng các bài toán kiểm định trình bày trong chương này đều là các bài toán kiểm định vững.

Bài toán 2

Chúng ta kiểm định

(H): $m = m_0$ với đối thiết (K): $m > m_0$, với mức ý nghĩa α cho trước.

Bài toán 3

Kiểm định

(H): $m = m_0$ với đối thiết (K): $m < m_0$, mức ý nghĩa α cho trước.

Cả hai bài toán này còn được gọi là các bài toán kiểm định một phía. Các miền bác bỏ được xây dựng gần giống như miền bác bỏ ở bài toán 1.

Miền bác bỏ ở bài toán 2

$$\boxed{W_2=\{(X_1,X_2,...,X_n)\in\mathbb{R}^n/\frac{\overline{X}-m_0}{\sigma}\sqrt{n}>u_\alpha\}.}$$

trong đó u_α được xác định từ bảng phân vị của phân bố chuẩn N(0,1):

$$P(u>u_\alpha/H)=P(W/H)=\alpha,$$

α là mức ý nghĩa của kiểm định, u là đại lượng ngẫu nhiên có phân bố chuẩn thuộc lớp $u \in N(0,1)$.

* Nếu $u > u_\alpha$ ta bác bỏ giả thiết (H), điều đó dựa trên cơ sở: xác suất để xảy ra biến cố $u > u_\alpha$ tương đối nhỏ nếu giả thiết (H): $m = m_0$ đúng:

$$P(u>u_\alpha/H)=\alpha$$

** Ngược lại nếu $u < u_\alpha$ ta ta chấp nhận giả thiết (H).

(Lưu ý rằng $u = \frac{\overline{X}-m_0}{\sigma}\sqrt{n}$ vẫn là thống kê xác định trong bài toán 1.)

Tương tự **Miền bác bỏ ở bài toán 3**

$$\boxed{W_3 = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / \frac{\overline{X} - m_0}{\sigma}\sqrt{n} < -u_\alpha\}.}$$

trong đó u_α được xác định từ bảng phân vị của phân bố chuẩn N(0,1):

$$P(u > u_\alpha / H) = \alpha, \quad u \in N(0,1)$$

* Nếu $u < -u_\alpha$ ta bác bỏ giả thiết (H), điều đó dựa trên cơ sở: xác suất để xảy ra biến cố $u < -u_\alpha$ tương đối nhỏ nếu giả thiết (H): $m = m_0$ đúng:

$$P(u < -u_\alpha / H) = P(u > u_\alpha / H) = \alpha$$

** Ngược lại nếu $u \geq -u_\alpha$ ta ta chấp nhận giả thiết (H).

Neyman-Pearson đã chứng minh rằng đối với các bài toán *kiểm định một phía* (bài toán 2, bài toán 3), quy tắc nêu trên là *quy tắc mạnh đều nhất*: xác suất xảy ra sai lầm loại II đạt giá trị nhỏ nhất khi mức ý nghĩa của kiểm định là α (tức là xác suất xảy ra sai lầm loại I nhỏ hơn hoặc bằng α).

Chú ý rằng ở cả ba bài toán trên, với cùng một giá trị α, các giá trị tương ứng u_α trong các bài toán 1, bài toán 2 là khác nhau:

Giả sử u là đại lượng ngẫu nhiên có phân bố chuẩn $u \in N(0,1)$

Trong bài toán 1, u_α được xác định từ hệ thức

$$P(|u| > u_\alpha) = \alpha \quad \text{hay} \quad P(|u| \leq u_\alpha) = 1 - \alpha.$$

Trong bài toán 2, bài toán 3, u_α được xác định từ hệ thức

$$P(u > u_\alpha) = \alpha.$$

Chú ý rằng giá trị u được tính sau khi thay kì vọng mẫu $\overline{X}$ vào thống kê

$$u = \frac{\overline{X} - m_0}{\sigma}\sqrt{n},$$

còn được gọi là *u-quan sát* và kí hiệu u_{qs}:

$$u_{qs} = \frac{\overline{X} - m_0}{\sigma}\sqrt{n}.$$

Người ta thường tóm tắt các kết luận của kiểm định dưới dạng:

Miền bác bỏ ở bài toán 1: $|u_{qs}| > u_\alpha$.

Miền bác bỏ ở bài toán 2: $u_{qs} > u_\alpha$.

Miền bác bỏ ở bài toán 3: $u_{qs} < -u_\alpha$.

Lưu ý rằng như đã nhắc đến ở trên, với cùng một giá trị α, u_α trong các bài toán 2 và bài toán 3 là như nhau, trong khi u_α trong bài toán 1 được xác định khác.

5.4.2 Kiểm định giả thiết về giá trị trung bình (trường hợp phương sai σ^2 chưa biết)

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên có phân bố chuẩn $X \in N(m, \sigma^2)$. Giả thiết rằng cả 2 tham số m và σ^2 đều chưa biết. Ta xét bài toán sau

Bài toán 1

Kiểm định giả thiết:

(H): $m = m_0$ với đối thiết (K): $m \neq m_0$, mức ý nghĩa $\alpha > 0$ cho trước.

Ta biết rằng thống kê

$$t = \frac{\overline{X} - m_0}{S^*}\sqrt{n}$$

có phân bố Student với $n - 1$ bậc tự do, nếu giả thiết (H) đúng. Tra bảng phân vị của phân bố Student để xác định t_α:

$$P(|t| \leq t_\alpha) = 1 - \alpha \quad \text{hay} \quad P(|t| \geq t_\alpha) = \alpha.$$

* Nếu $|t| > t_\alpha$ ta bác bỏ giả thiết (H). Như đã trình bày trong mục trước, trong trường hợp này ta có thể mắc sai lầm (loại I), tuy nhiên xác suất mắc sai lầm đó tương đối nhỏ, bằng α:

$$P(|t| > t_\alpha / H) = \alpha$$

** Ngược lại nếu $|t| \leq t_\alpha$ ta chấp nhận giả thiết (H), nói cách khác mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ không mâu thuẫn gì với giả thiết (H).

Như vậy ở bài toán 1 miền bác bỏ là

$$\boxed{W_1 = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / \left| \frac{\overline{X} - m_0}{S^*} \sqrt{n} \right| > t_\alpha\}.}$$

Hoàn toàn tương tự như trường hợp σ đã biết, xét 2 bài toán kiểm định một phía với mức ý nghĩa α cho trước:

Bài toán 2

(H): $m = m_0$ với đối thiết (K): $m > m_0$.

Bài toán 3

(H): $m = m_0$ với đối thiết (K): $m < m_0$.

Miền bác bỏ của bài toán 2: $t_{qs} > t_\alpha$ hay:

$$\boxed{W_2 = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / \frac{\overline{X} - m_0}{S^*} \sqrt{n} > t_\alpha\}.}$$

Miền bác bỏ của bài toán 3: $t_{qs} < -t_\alpha$ hay:

$$\boxed{W_3 = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / \frac{\overline{X} - m_0}{S^*} \sqrt{n} < -t_\alpha\}.}$$

Ở 2 bài toán này t_α được xác định từ hệ thức

$$P(t > t_\alpha) = \alpha$$

trong đó t có phân bố Student với $n - 1$ bậc tự do.

Ví dụ 5.4.1

Cho mẫu ngẫu nhiên sau của đại lượng ngẫu nhiên có phân bố chuẩn với kích thước mẫu $n = 20$:

{4,7 ; 3,01 ; 4,65 ; 3,52 ; 3 ; 3,26 ; 3,1 ; 3,41 ; 5,64 ; 2,97 ; 4,89 ; 4,55 ; 4,89 ; 2,72 ; 4,57 ; 3,94 ; 3,61 ; 6,22 ; 3,02 ; 2,47}

Hãy kiểm định giả thiết (H) giá trị trung bình của đại lượng ngẫu nhiên đó bằng 4,4 với đối thiết (K): $m \neq 4,4$. Cho mức ý nghĩa của kiểm định $\alpha = 0,04$.

Ta tính các đặc trưng:

Kì vọng mẫu

$$\overline{X} = \frac{1}{20}\sum_{i=1}^{20} X_i = 3,907$$

và phương sai mẫu điều chỉnh

$$S^{*2} = \frac{1}{19}\sum_{i=1}^{20}(X_i - \overline{X})^2 = 1,086401 = (1,042306)^2.$$

Vậy độ lệch chuẩn $S^* = 1,042306$.

Tra bảng phân vị của phân bố Student 19 bậc tự do ứng với $\alpha = 0,04$, phân vị đó bằng

$$t_\alpha = 2,20470 \quad (P(|t| > 2,20470) = 0,04).$$

Áp dụng công thức

$$t = t_{qs} = \frac{\overline{X} - 4,4}{S^*}\sqrt{20} = -2,115275.$$

Do $|t_{qs}| = 2,115275 < t_\alpha = 2,20470$, mẫu không thuộc miền bác bỏ, vậy ta chấp nhận giả thiết giá trị trung bình của đại lượng ngẫu nhiên bằng 4,4.

Với mẫu trên nếu xét bài toán kiểm định một phĩa (bài toán 3), kiểm định:

(H): Giá trị trung bình bằng 4,4 với đối thiết (K): $m < 4,4$. Mức ý nghĩa của kiểm định $\alpha = 0,04$.

Trước hết ta tra t_α từ hệ thức $P(t > t_\alpha) = 0,04$, ta được $t_{0,04} = 1,84953$.

Do $t_{qs} = -2,115275 < -t_{0,04} = -1,84953$, mẫu thuộc miền bác bỏ của bài toán 3, vậy ta bác bỏ giả thiết (H), chấp nhận đối thiết $m < 4,4$.

Ví dụ 5.4.2

Quay lại ví dụ 9 chương này, đo chiều cao của 100 thanh niên ở lứa tuổi trưởng thành, kết quả được cho trong bảng dưới đây. Biết chiều cao của thanh niên là đại lượng ngẫu nhiên có phân bố chuẩn.

Khoảng	Chiều cao x_i	n_i
1,55 -1,57	1,56	5
1,58 -1,60	1,59	12
1,61 -1,63	1,62	26
1,64 -1,66	1,65	25
1,67 -1,69	1,68	20
1,70 -1,72	1,71	7
1,73 -1,75	1,74	4
1,76	1,76	1
Tổng cộng		100

Hãy kiểm định giả thiết (H) chiều cao trung bình của thanh niên bằng 1,63 với đối thiết (K): $m > 1,63$. Cho mức ý nghĩa của kiểm định $\alpha = 0,05$.

Đây là bài toán kiểm định một phĩa (bài toán 2), tra bảng phân vị của phân bố Student 99 bậc tự do từ hệ thức

$P(t > t_\alpha) = 0,05$, ta được $t_{0,05} = 1,660392$.

Như đã tính trong ví dụ 9, kì vọng mẫu $\overline{X} = 1,6454$ và độ lệch chuẩn $S^* = 0,044116925$. Vậy giá trị quan sát

$$t_{qs} = \frac{\overline{X} - 1,63}{S^*}\sqrt{100} = 3,4907.$$

Do $t_{qs} = 3,4907 > t_{0,05} = 1,660392$, mẫu thuộc miền bác bỏ của bài toán 2, vậy ta bác bỏ giả thiết (H), chấp nhận đối thiết chiều cao trung bình lớn hơn 1,63.

5.4.3 Kiểm định giả thiết về sự bằng nhau của các giá trị trung bình

Gọi $(X_1, X_2, ..., X_m)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $X \in N(m_1, \sigma_2^2)$, $(Y_1, Y_2, ..., Y_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $Y \in N(m_2, \sigma_2^2)$. Các tham số m_1, m_2 và σ_1^2, σ_2^2 là các tham số chưa biết, tuy nhiên ta giả thiết rằng $\sigma_1^2 = \sigma_2^2$. Giả thiết tiếp các đại lượng ngẫu nhiên

$$X_1, X_2, ..., X_m, \quad Y_1, Y_2, ..., Y_n$$

độc lập nhau.

Xét bài toán kiểm định sau

(H): $m_1 = m_2$ với đối thiết (K): $m_1 \neq m_2$ với mức ý nghĩa α cho trước.

Ta biết rằng với các giả thiết trên

$$\boxed{t = \sqrt{\frac{mn(m+n-2)}{m+n}} \cdot \frac{\overline{X} - \overline{Y}}{\sqrt{nS_X^2 + mS_Y^2}}}$$

có phân bố Student với $m + n - 2$ bậc tự do, nếu giả thiết (H): $m_1 = m_2$ đúng. Tra bảng phân vị của phân bố Student với $m + n - 2$ bậc tự do để xác định t_α:

$$P(|t| \leq t_\alpha) = 1 - \alpha \quad \text{hay} \quad P(|t| > t_\alpha) = \alpha.$$

Giá trị t được tính trong thống kê trên cũng được gọi là *t-quan sát* và kí hiệu t_{qs}

* Nếu $|t_{qs}| > t_\alpha$ ta bác bỏ giả thiết (H), điều đó dựa trên cơ sở: xác suất để xảy ra biến cố $|t_{qs}| > t_\alpha$ tương đối nhỏ nếu giả thiết (H): $m_1 = m_2$ đúng:

$$P(|t| > t_\alpha / H) = \alpha.$$

** Ngược lại nếu $|t| \leq t_\alpha$ ta chấp nhận giả thiết (H).

Chú ý rằng tương tự như các bài toán nêu trên, vấn đề này cũng có các bài toán kiểm định một phía.

Ví dụ 5.4.3

Người ta muốn kiểm định xem một phương pháp mới phải chăng có cải thiện chất lượng bê tông tốt hơn so với phương pháp trước đây (tăng cường độ chịu nén của bê tông). Để nhằm mục đích đó, người ta chọn 6 mẫu bê tông theo phương pháp sản xuất cũ và chọn 6 mẫu bê tông theo phương pháp sản xuất mới. Kí hiệu X_i là cường độ chịu nén của bê tông sản xuất theo phương pháp cũ và Y_i là cường độ bê tông sản xuất theo phương pháp mới.

$X_i(kg/cm^2)$	$Y_i(kg/cm^2)$
300	305
301	317
303	308
288	300
294	314
296	316

Gọi $m_1 = EX$ là cường độ chịu nén trung bình của bê tông theo phương pháp sản xuất cũ và $m_2 = EY$ là cường độ chịu nén trung bình của bê tông theo phương pháp sản xuất mới. Hãy kiểm định

(H): $m_1 = m_2$ với đối thiết (K): $m_1 < m_2$ với mức ý nghĩa $\alpha = 0,01$.

Đây là bài toán kiểm định một phía (tương ứng với bài toán 3), miền bác bỏ có dạng

$$t_{qs} = \sqrt{\frac{mn(m+n-2)}{m+n}} \cdot \frac{\overline{X} - \overline{Y}}{\sqrt{nS_X^2 + mS_Y^2}} < -t_\alpha$$

trong đó t_α là phân vị của phân bố Student với 6+6-2= 10 bậc tự do

$$P(t > t_\alpha) = \alpha = 0,01 \quad \text{(phân vị một phía).}$$

Tra bảng phân vị một phía phân bố Student 10 bậc tự do ta được

$$t_{0,01} = 2,76377.$$

Trung bình mẫu $\overline{X} = 297, \overline{Y} = 310$.

Phương sai mẫu $S_X^2 = 25,333, S_Y^2 = 38,333$

Giá trị quan sát $t_{qs} = -3,64311$

Do $t_{qs} = -3,64311 < -t_{0,01} = -2,76377$, ta bác bỏ(H) và chấp nhận đối thiết (K): $m_1 < m_2$ ở mức ý nghĩa $\alpha = 0,01$.

5.4.4 Kiểm định giả thiết về sự bằng nhau của các phương sai

Gọi $(X_1, X_2, ..., X_m)$ là mẫu ngẫu nhiên tương ứng với đại lượng ngẫu nhiên $X \in N(m_1, \sigma_2^2)$, $(Y_1, Y_2, ..., Y_n)$ là mẫu ngẫu nhiên tương ứng với đại lượng

ngẫu nhiên $Y \in N(m_2, \sigma_2^2)$. Các tham số m_1, m_2, σ_1^2 σ_2^2 là các tham số chưa biết. Tuy nhiên ta giả thiết rằng các đại lượng ngẫu nhiên

$$X_1, X_2, ..., X_m, \quad Y_1, Y_2, ..., Y_n$$

độc lập nhau.

Xét bài toán kiểm định với mức ý nghĩa α cho trước.

(H): $\sigma_1 = \sigma_2$ với đối thiết (K): $\sigma_1 \neq \sigma_2$. Ta biết rằng theo các kết quả đã chứng minh trong mục 5.1.5 chương này

$$\frac{1}{\sigma_1^2}\sum_{i=1}^{m} X_i^2 = \frac{m}{\sigma_1^2}S_X^2 = \frac{m-1}{\sigma_1^2}{S_X^*}^2$$

có phân bố χ^2 với $m-1$ bậc tự do.

$$\frac{1}{\sigma_2^2}\sum_{i=1}^{n} Y_i^2 = \frac{n}{\sigma_2^2}S_Y^2 = \frac{n-1}{\sigma_2^2}{S_Y^*}^2$$

có phân bố χ^2 với $n-1$ bậc tự do. Mặt khác do

$$X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n$$

độc lập, suy ra nếu giả thiết (H): $\sigma_1 = \sigma_2$ đúng

$$F = \frac{n-1}{m-1} \cdot \frac{\frac{m-1}{\sigma_1^2}{S_X^*}^2}{\frac{n-1}{\sigma_2^2}{S_Y^*}^2} = \frac{{S_X^*}^2}{{S_Y^*}^2}$$

là đại lượng ngẫu nhiên có phân bố-F với $m-1, n-1$ bậc tự do.

Việc xây dựng miền bác bỏ (tương tự như các trường hợp đã giới thiệu trong mục trước), là chỉ ra một khoảng sao cho giá trị quan sát F_{qs} nằm ngoài khoảng đó với xác suất α khá bé cho trước. Ta kí hiệu $\frac{1}{F_1}$ là cận dưới và F_2 là cận trên của khoảng đó. F_1 và F_2 được xác định từ hệ thức

$$P(F > F_2) = \frac{\alpha}{2}. \quad \text{Tương tự}$$

$$P(F < \frac{1}{F_1}) = \frac{\alpha}{2} \quad \text{hay} \quad P(\frac{1}{F} > F_1) = \frac{\alpha}{2}$$

trong đó F là đại lượng ngẫu nhiên phân bố-F với $m-1, n-1$ bậc tự do. Khi đó

$$P\left(F \notin \left(\frac{1}{F_1}, F_2\right)\right) = \alpha.$$

Vậy miền bác bỏ của bài toán là

$$\boxed{F_{qs} = \frac{{S_X^*}^2}{{S_Y^*}^2} \notin \left(\frac{1}{F_1}, F_2\right).}$$

Chú ý rằng khi α tương đối nhỏ, cả F_1 và F_2 đều lớn hơn 1. Vì vậy trong thực hành ta thường sắp xếp các mẫu X_i, Y_j sao cho $\frac{{S_X^*}^2}{{S_Y^*}^2} > 1$, khi đó miền bác bỏ thực chất là

$$\boxed{F_{qs} = \frac{{S_X^*}^2}{{S_Y^*}^2} > F_2.}$$

Ví dụ 5.4.4

Quay lại ví dụ 14, về các phương pháp sản xuất bê tông. Chọn 6 mẫu bê tông theo phương pháp sản xuất cũ và chọn 6 mẫu bê tông theo phương pháp sản xuất mới. Kí hiệu X_i là cường độ chịu nén của bê tông sản xuất theo phương pháp cũ và Y_i là cường độ bê tông sản xuất theo phương pháp mới.

$X_i(kg/cm^2)$	$Y_i(kg/cm^2)$
300	305
301	317
303	308
288	300
294	314
296	316

Hãy kiểm định giả thiết về sự bằng nhau của các phương sai với mức ý nghĩa $\alpha = 0,05$. Các phương sai mẫu điều chỉnh:

$${S_X^*}^2 = 30,4 \quad {S_Y^*}^2 = 46$$

Giá trị quan sát $F_{qs} = \frac{{S_Y^*}^2}{{S_X^*}^2} = \frac{46}{30,4} = 1,51316.$

Tra bảng phân vị phân bố F với $m-1=5, n-1=5$ bậc tự do, F_2 xác định từ hệ thức

$$P(F > F_2) = \frac{\alpha}{2} = 0,025$$

ta được $F_2 = 7,14636$. Giá trị quan sát

$$F_{qs} = 1,51316 < F_2 = 7,14636,$$

suy ra ta không có cơ sở bác bỏ (H).

5.4.5 Kiểm định giả thiết về xác suất của biến cố ngẫu nhiên

Giả sử A là biến cố ngẫu nhiên có xác suất $P(A) = p$ chưa biết. Ta sử dụng ước lượng

$$\widehat{p} = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

trong đó X_i bằng 1 hoặc 0 tùy theo biến cố A xảy ra hoặc không xảy ra ở phép thử ngẫu nhiên thứ $i, i = 1, 2, ..., n$. ($\widehat{p}$ thực chất là tần suất xuất hiện của biến cố A). Khi đó $n\widehat{p}$ có phân bố nhị thức với

$$E(n\widehat{p}) = np$$

$$D(n\widehat{p}) = npq, q = 1 - p$$

Xét bài toán kiểm định sau

Bài toán 1

(H): $p = p_0$ với đối thiết (K): $p \neq p_0$ với mức ý nghĩa α cho trước ($0 < p_0 < 1$ là giá trị cho trước nào đó).

Ta đã biết, theo định lí giới hạn trung tâm

$$\frac{n\widehat{p} - np}{\sqrt{npq}} = \sqrt{n}\frac{\widehat{p} - p}{\sqrt{pq}}$$

có phân bố xấp xỉ chuẩn ($\approx N(0,1)$) khi n đủ lớn. Vì vậy sử dụng thống kê

$$u = u_{qs} = \sqrt{n}\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}},$$

u có phân bố xấp xỉ chuẩn N(0,1), khi giả thiết (H): $p = p_0$ đúng. Tra bảng phân vị của phân bố chuẩn N(0,1) để xác định u_α:

$$P(|u| \leq u_\alpha) = 1 - \alpha$$

trong đó $u \in N(0,1)$

* Nếu $|u_{qs}| > u_\alpha$ ta bác bỏ giả thiết (H), chấp nhận đối thiết (K). Điều đó dựa trên cơ sở: xác suất để xảy ra biến cố$|u_{qs}| > u_\alpha$ tương đối nhỏ nếu giả thiết (H): $p = p_0$ đúng:

$$P(|u| > u_\alpha / H) = \alpha$$

** Ngược lại nếu $|u_{qs}| \leq u_\alpha$ ta chấp nhận giả thiết (H).

Như vậy ở bài toán này miền bác bỏ

$$\boxed{W_1 = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / \left| \sqrt{n} \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}} \right| > u_\alpha\}.}$$

và

$$P(|u| > u_\alpha / H) = P(W/H) = \alpha$$

là mức ý nghĩa của kiểm định.

Xét các bài toán kiểm định một phía. Các miền bác bỏ của chúng được xây dựng gần giống như miền bác bỏ ở bài toán 1.

Bài toán 2

Chúng ta kiểm định: (H): $p = p_0$ với đối thiết (K): $p > p_0$ với mức ý nghĩa α cho trước.

Miền bác bỏ có dạng

$$\boxed{W_2 = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / \sqrt{n} \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}} > u_\alpha\}.}$$

Bài toán 3

Kiểm định:

(H): $p = p_0$ với đối thiết (K): $p < p_0$ với mức ý nghĩa α cho trước.

Miền bác bỏ

$$\boxed{W_3 = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / = \sqrt{n}\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}} < -u_\alpha\}.}$$

Trong bài toán 2, bài toán 3, u_α được xác định từ hệ thức

$$P(u > u_\alpha) = \alpha$$

trong đó u là đại lượng ngẫu nhiên có phân bố chuẩn: $u \in N(0,1)$.

5.4.6 Kiểm định giả thiết về tính phù hợp của hàm phân bố

Người ta sớm sử dụng đại lượng ngẫu nhiên có phân bố χ^2 để kiểm định các bài toán về tính phù hợp của hàm phân bố. Xét bài toán kiểm định giả thiết:

(H): Một đại lượng ngẫu nhiên X nào đó có phân bố dạng $F(x, \Theta)$. (Phân bố $F(x, \Theta)$ chẳng hạn như dạng phân bố chuẩn, phân bố Poisson, phân bố mũ hoặc đều, . . .) với đối thiết ngược lại.

Để giải bài toán đó, người ta chọn một mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

tương ứng với đại lượng ngẫu nhiên X và chia các phần tử mẫu vào r nhóm: mỗi nhóm chứa n_i phần tử mẫu, mỗi phần tử mẫu chỉ thuộc một nhóm duy nhất

$$n = n_1 + n_2 + ... + n_r = \sum_{i=1}^{r} n_i.$$

Giả sử p_i là xác suất để đại lượng ngẫu nhiên X nhận các giá trị thuộc nhóm thứ $i, i = 1, 2, ..., r$ với điều kiện giả thiết (H) đúng. Khi đó

$$1 = p_1 + p_2 + ... + p_r$$

Hiển nhiên n_i là đại lượng ngẫu nhiên có phân bố nhị thức với kì vọng $E(n_i) = np_i$. Xét thống kê

$$Q^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}.$$

a) Trường hợp phân bố $F(x, \Theta) = F(x)$ không phụ thuộc tham số Θ, nói cách khác các xác suất $p_i, i = 1, 2, ..., r$ được xác định cụ thể. Khi đó người ta đã chứng minh được rằng với n đủ lớn và giả thiết (H) là đúng, Q^2 sẽ có phân bố xấp xỉ phân bố χ^2 với $r - 1$ bậc tự do.

b) Trường hợp phân bố dạng $F(x, \Theta)$ phụ thuộc tham số Θ.

Giả sử tham số $\Theta = (\Theta_1, \Theta_2, ..., \Theta_k)$ là véc tơ, gồm k tham số tạo thành (chẳng hạn như dạng phân bố chuẩn $F(x, \Theta) = F(x, m, \sigma^2) \in N(m, \sigma^2)$ gồm 2 tham số thành phần). Khi đó thay vì

$$Q^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

ta xét thống kê

$$Q^2 = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p_i})^2}{n\widehat{p_i}}$$

trong đó $\widehat{p_i}, i = 1, 2, ..., r$ là xác suất để X nhận các giá trị thuộc nhóm thứ i, xác suất đó được tính thông qua hàm phân bố $F(x, \widehat{\Theta})$ mà $\widehat{\Theta} = (\widehat{\Theta_1}, \widehat{\Theta_2}, ..., \widehat{\Theta_k})$ là các ước lượng hợp lí cực đại của các tham số $\Theta_1, \Theta_2, ..., \Theta_k$.

Người ta đã chứng minh được rằng với n đủ lớn và giả thiết (H) là đúng khi đó Q^2 sẽ có phân bố xấp xỉ phân bố χ^2 với $r - k - 1$ bậc tự do, k là số tham số của phân bố $F(x, \Theta)$ trong giả thiết (H).

(Giả sử phân bố $F(x, \Theta)$ là phân bố chuẩn $N(m, \sigma^2)$, Θ được coi như véc tơ (m, σ^2) và số tham số của phân bố bằng $k = 2$, trường hợp $F(x, \lambda)$ là phân bố mũ chẳng hạn số tham số của phân bố là $k = 1$,...)

Miền bác bỏ của kiểm định do vậy là

$$\boxed{W = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / \sum_{i=1}^{r} \frac{(n_i - n\widehat{p_i})^2}{n\widehat{p_i}} > \chi^2_\alpha\}.}$$

trong đó χ^2_α được xác định từ hệ thức $P(\chi^2 > \chi^2_\alpha) = \alpha$, ($\chi^2$ là đại lượng ngẫu nhiên có phân bố χ^2 với $r - k - 1$ bậc tự do).

Ví dụ 5.4.5

Hãy kiểm định tính đối xứng của đồng xu thông qua việc gieo đồng xu 24000 lần, số lần xuất hiện mặt sấp là 12139 và số lần xuất hiện mặt ngửa 11861 lần.

Nếu gọi p là xác suất mặt sấp xuất hiện khi gieo đồng xu. Bài toán dẫn đến kiểm định giả thiết sau:

(H): $p = 0,5$ với đối thiết (K) $p \neq 0,5$ sử dụng quy tắc kiểm định χ^2 nêu trên với mức ý nghĩa $\alpha = 0,05$.

Tính giá trị quan sát

$$Q^2 = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p_i})^2}{n\widehat{p_i}} = \frac{(n_1 - 0,5n)^2}{0,5n} + \frac{(n_2 - 0,5n)^2}{0,5n} = 3,22017$$

trong đó

$$n = 24000, n_1 = 12139, n_2 = 11861, p_1 = p_2 = 0,5.$$

Tra bảng phân vị phân bố χ^2 với 1 bậc tự do: $\chi^2_{0,05} = 3,84146$

$$Q^2 = 3,22017 < \chi^2_{0,05} = 3,84146.$$

Ta chấp nhận giả thiết (H) : đồng xu đối xứng theo nghĩa xác suất mặt sấp xuất hiện bằng xác suất mặt ngửa xuất hiện và bằng $0,5$.

Chú ý rằng với mức ý nghĩa $\alpha = 0,01$, phân vị bằng $\chi^2_{0,01} = 6,6369$ càng phù hợp.

Ví dụ 5.4.6

Một nguyên tố chất phóng xạ bắn ra các hạt α vào một vùng không gian K nào đó trong khoảng thời gian t. Người ta sử dụng máy đo và thực hiện $n = 800$ lần đo, mỗi lần kéo dài trong thời gian 7 giây. Bảng sau cho ta số lần (n_k) xảy ra biến cố: có đúng k hạt bắn vào vùng không gian đó trong khoảng thời gian t.

k	n_k	$\widehat{p_k}$	$n\widehat{p_k}$
0	18	0,021094	16,8755
1	65	0,081398	65,1183
2	121	0,157047	125,638
3	160	0,202002	161,601
4	162	0,194868	155,895
5	118	0,15039	120,312
6	82	0,096719	77,3755
7	45	0,053317	42,6532
8	16	0,025717	20,5735
9	8	0,011026	8,8209
≥ 10	5	0,006422	5,1371
Tổng	800	1,0	800,000

Hãy kiểm định giả thiết sau với mức ý nghĩa 0,05:

(H): Số hạt bắn vào vùng K trong khoảng thời gian t có phân bố Poisson.

Ước lượng hợp lí cực đại của tham số λ:

$$\overline{X} = \frac{\sum_{k=0}^{9} kn_k + 11n_{10}}{800} = \frac{3087}{800} = 3,85875$$

hay

$$\widehat{\lambda} = 3,85875.$$

Cột 3 của bảng trên là các xác suất $\widehat{p_k}$ của phân bố Poisson sau khi thay $\widehat{\lambda} = \overline{X} = 3,85875$ vào biểu thức $e^{-\lambda}\frac{(\lambda)^k}{k!}$. Suy ra

$$Q^2 = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p_i})^2}{n\widehat{p_i}} = 2,04801.$$

Số nhóm trong ví dụ này là 11, phân bố Poisson có 1 tham số. Tra bảng phân vị phân bố χ^2 với 11-1-1=9 bậc tự do:

$$\chi^2_{0,05} = 16,91896.$$

Ta có

$$Q^2 = 2,04801 < \chi^2_{0,05} = 16,91896.$$

Vậy ta chấp nhận giả thiết số hạt bắn vào vùng không gian đó trong khoảng thời gian t có phân bố Poisson.

5.4.7 Kiểm định về tính độc lập

Người ta có thể kiểm định về tính độc lập của các biến cố ngẫu nhiên, các đại lượng ngẫu nhiên. Chúng ta trình bày vấn đề dưới dạng sau đây:

Cho hai hệ đầy đủ các biến cố $A_1, A_2, ..., A_r$; $B_1, B_2, ..., B_s$. Hãy kiểm định giả thiết hai hệ đó độc lập:

(H): $P(A_iB_j) = P(A_i)P(B_j)$ với mọi $i = 1, 2, ..., r; j = 1, 2, ..., s$.

Xét một mẫu ngẫu nhiên cỡ n (mẫu gồm n phần tử mẫu). Ta đưa vào các kí hiệu sau:

n_{ij} là số lần xảy ra biến cố tích A_iB_j trong tập hợp các phần tử mẫu.

$n_{i.} = \sum_{j=1}^{s} n_{ij}$ là số lần xảy ra biến cố A_i.

$n_{.j} = \sum_{i=1}^{r} n_{ij}$ là số lần xảy ra biến cố B_j.

Hiển nhiên $\sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = n$ và $\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = n$.

Các số n_{ij} được xếp vào bảng sau đây:

j / i	1	2	. . .	s	Tổng
1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2.}$
.			$\cdots$	.	
.			$\cdots$	.	
.			$\cdots$	.	
r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r.}$
Tổng	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.s}$	n

Nếu giả thiết (H) đúng, khi đó $P(A_iB_j) = P(A_i)P(B_j)$ với mọi i, j và

$$P(A_iB_j) \approx \frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n^2}.$$

Người ta chứng minh được rằng khi n tăng ra vô cùng, thống kê

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

có phân bố xấp xỉ phân bố χ^2 với $(r-1)(s-1)$ bậc tự do. Miền bác bỏ của

kiểm định do vậy là

$$W = \{(X_1, X_2, ..., X_n) \in \mathbb{R}^n / \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} > \chi^2_\alpha\}$$

trong đó χ^2_α được xác định từ hệ thức $P(\chi^2 > \chi^2_\alpha) = \alpha$, χ^2 là đại lượng ngẫu nhiên có phân bố χ^2 với $(r-1)(s-1)$ bậc tự do.

Ví dụ 5.4.7

Người ta kiểm tra đường kính vòng ngoài và đường kính vòng trong của các ổ bi, họ phân các ổ bi thành ba loại: loại I là loại tốt, loại II là loại cần sửa chút ít và loại III là loại phế phẩm. Chọn ngẫu nhiên ra 200 ổ bi để kiểm tra, kết quả ghi nhận trong bảng sau

ĐK ngoài / ĐK trong	loại I	loại II	loại III	Tổng
loại I	169	8	1	178
loại II	9	4	1	14
loại III	1	3	4	8
Tổng	179	15	6	200

Hãy kiểm định giả thiết các đặc điểm loại I, loại II và loại III thuộc về đường kính vòng ngoài và đường kính vòng trong của các ổ bi độc lập với nhau. Cho trước mức ý nghĩa của kiểm định là $\alpha = 0,05$.

Trong ví dụ này $r = s = 3$, suy ra phân vị $\chi^2_{0,05}$ được xác định từ hệ thức

$$P(\chi^2 > \chi^2_{0,05}) = 0,05,$$

trong đó χ^2 là đại lượng ngẫu nhiên có phân bố χ^2 với $(r-1)(s-1) = 4$ bậc tự do. Phân vị đó bằng

$$\chi^2_{0,05} = 9,488.$$

Mặt khác

$$Q^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = 90,15$$

lớn hơn rất nhiều so với phân vị $\chi^2_{0,05} = 9,488$ ở trên. Vậy ta bác bỏ giả thiết về tính độc lập của chúng.

5.5 Tương quan và hồi quy

5.5.1 Hồi quy và hồi quy bình phương trung bình tuyến tính

Trong thực tế ta thường gặp các cặp hai đại lượng ngẫu nhiên chẳng hạn X và Y chẳng những không độc lập nhau mà còn có quan hệ phụ thuộc ngẫu nhiên vào nhau. Ví dụ X là chiều cao và Y là trọng lượng của một người nào đó, (hoặc X là lượng phân bón và Y là năng suất của một giống lúa tương úng...) khi đó nhằm mục đích dự báo đại lượng ngẫu nhiên Y trên cơ sở đã biết về X, người ta tìm cách xấp xỉ Y bởi một hàm $f(X)$ của đại lượng ngẫu nhiên X.

Trong chương III ta đã định nghĩa hàm hồi quy của Y đối với X:

$$h(X) = E(Y/X)$$

Bây giờ ta sẽ phát biểu và chứng minh định lí sau:

Định lí 5.5.1 *Cho X và Y là hai đại lượng ngẫu nhiên. Giả sử tồn tại phương sai $D(Y)$. Khi đó*

$$E(Y - f(X))^2$$

đạt giá trị nhỏ nhất khi $f(X) = E(Y/X)$ là hàm hồi quy của Y đối với X.

Chứng minh

Đặt $M = E(Y/X)$. Xét biểu thức

$$E(Y - f(X))^2 = E(Y - M + M - f(X))^2 =$$

$$= E(Y - M)^2 + E(M - f(X))^2 + 2E[(Y - M)(M - f(X))].$$

Theo định lí 3.6.2 chương III

$$E[(Y - M)(M - f(X))] = E[(Y - M)(M - f(X))/X] =$$

$$= (M - f(X))E(Y - M/X) = (M - f(X))[E(Y/X) - M] = 0$$

vì $(M - f(X))$ là hàm của X nên có thể đưa ra ngoài dấu kì vọng có điều kiện. Suy ra

$$E(Y - f(X))^2 = E(Y - M)^2 + E(M - f(X))^2 \geq E(Y - E(Y/X))^2.$$

Dấu bằng xảy ra khi và chỉ khi số hạng

$$E(M - f(X))^2 = 0,$$

hay $f(X) = E(Y/X)$ đ.p.c.m.

Chú ý rằng cũng chứng minh như trên

$$cov(f(X), Y) = E[(f - Ef)M] = cov(f(X), M).$$

Đặc biệt khi $f(X) = E(Y/X) = M$, mô men tương quan $cov(M, Y)$ bằng phương sai của M: $cov(M, Y) = \sigma_M^2$, suy ra hệ số tương quan giữa Y và hồi quy M đạt giá trị lớn nhất

$$\varrho^2(f(X), Y) = \varrho^2(f(X), M)\varrho^2(Y, M) \leq \varrho^2(Y, M).$$

Từ định lí trên ta suy ra rằng nếu ta xấp xỉ Y với hàm hồi quy $f(X) = E(Y/X)$ khi đó bình phương của sai số đạt giá trị bé nhất. Chính xác hơn bình phương của sai số có kì vọng $E(Y - f(X))^2$ bé nhất.

Tuy nhiên trong thực tế việc tìm hàm hồi quy lí thuyết $f(X) = E(Y/X)$ không đơn giản. Vì vậy trong thực hành nếu sự phụ thuộc tuyến tính của X và Y tương đối chặt, khi đó ta xấp xỉ Y với một hàm bậc nhất $\alpha X + \beta$ của đại lượng ngẫu nhiên X theo nghĩa làm nhỏ nhất sai số bình phương trung bình

$$E(Y - \alpha X - \beta)^2.$$

Xét

$$E(Y - \alpha X - \beta)^2 = E[(Y - EY) - \alpha(X - EX) + EY - \alpha EX - \beta]^2 =$$

$$= D(Y) + \alpha^2 D(X) - 2\alpha cov(X, Y) + (EY - \alpha EX - \beta)^2 + 0 + 0 \geq$$

$$\geq D(Y) + \alpha^2 D(X) - 2\alpha cov(X, Y).$$

Tam thức bậc hai $\alpha^2 D(X) - 2\alpha cov(X, Y) + D(Y)$ đạt giá trị bé nhất khi

$$\alpha = cov(X, Y)/D(X) = \varrho \frac{\sqrt{DY}}{\sqrt{DX}}.$$

Điều kiện đó cùng với $\beta = EY - \alpha EX$ làm nhỏ nhất sai số bình phương trung bình $E(Y - \alpha X - \beta)^2$. Vậy hàm bậc nhất cần tìm

$$\alpha X + b = \varrho\frac{\sqrt{DY}}{\sqrt{DX}}X + EY - \varrho\frac{\sqrt{DY}}{\sqrt{X}}EX =$$

$$= \varrho\frac{\sqrt{DY}}{\sqrt{DX}}(X - EX) + EY.$$

Hàm đó được gọi là *hàm hồi quy bình phương trung bình tuyến tính* của Y đối với X hoặc nói tắt là *hàm hồi quy tuyến tính lí thuyết.*

Để dễ nhớ người ta còn viết hàm hồi quy bình phương trung bình tuyến tính dưới dạng

$$y = \alpha x + \beta = \varrho\frac{\sqrt{DY}}{\sqrt{DX}}(x - EX) + EY$$

hay

$$\boxed{y - EY = \varrho\frac{\sqrt{DY}}{\sqrt{DX}}(x - EX).}$$

Sai số bình phương trung bình

$$E(Y - \alpha X - \beta)^2 = D(Y) + \alpha^2 D(X) - 2\alpha cov(X, Y) = DY(1 - \varrho^2).$$

Ta thường kí hiệu sai số đó là $\sigma^2_{Y/X}$:

$$\sigma^2_{Y/X} = DY(1 - \varrho^2).$$

Tương tự hàm hồi quy bình phương trung bình tuyến tính của X đối với Y

$$X - EX = \varrho\frac{\sqrt{DX}}{\sqrt{DY}}(Y - EY).$$

Sai số bình phương trung bình

$$\sigma^2_{X/Y} = DX(1 - \varrho^2).$$

Chú ý rằng hồi quy bình phương trung bình tuyến tính chỉ nên dùng khi $|\varrho|$ đủ lớn ($|\varrho| > 0,7$).

5.5.2 Hồi quy bình phương trung bình tuyến tính thực nghiệm

Để xấp xỉ Y với một hàm bậc nhất $aX + b$ của đại lượng ngẫu nhiên X, theo mục trước ta dùng công thức

$$\boxed{Y - EY = \varrho\frac{\sqrt{DY}}{\sqrt{DX}}(X - EX)}$$

hay

$$y = EY + \varrho\frac{\sqrt{DY}}{\sqrt{DX}}(x - EX).$$

Song các giá trị lí thuyết kì vọng, phương sai EX, DX hoặc hệ sốtương quan ϱ ta chưa biết. Do vậy ta chỉ có thể xác định hồi quy thông qua các đặc trưng mẫu.

Giả sử

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$$

là mẫu ngẫu nhiên tương ứng với X và Y. Ta hạn chế chỉ xét trường hợp $E(Y/X)$ là hàm tuyến tính đối với X

$$E(Y/X) = \alpha X + \beta.$$

Khi đó hàm hồi quy bình phương trung bình tuyến tính có dạng

$$\boxed{y = \overline{Y} + r\frac{S_Y}{S_X}(x - \overline{X}).}$$

Trong đó $\overline{X}, \overline{Y}$ là kì vọng mẫu của X, Y

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

S_X^2, S_Y^2 là phương sai mẫu của X, Y

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2,$$

$$S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2.$$

Hệ số tương quan mẫu r bằng

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{S_X S_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\cdot\overline{Y}}{S_X S_Y}.$$

Sai số bình phương trung bình

$$\sigma_{Y/X}^2 = S_Y^2(1 - r^2).$$

Tương tự hàm hồi quy bình phương trung bình tuyến tính của X đối với Y

$$\boxed{x = \overline{X} + r\frac{S_X}{S_Y}(y - \overline{Y}).}$$

Ví dụ 5.5.1

Bảng sau cho mẫu ngẫu nhiên của (X, Y) với kích thước mẫu $n = 7$. Hãy tìm đường thẳng hồi quy của Y đối với X.

X	1	2	3	4	5	6	7
Y	0,203	0,815	6,05	10,3	11,6	15,6	17

Trước hết ta tính kì vọng mẫu của X và Y

$$\overline{X} = 4 \quad \overline{Y} = 8,795429$$

Phương sai mẫu của X được tính theo công thức

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = 4,$$

hay $S_X = 2$. Tương tự $S_Y = 6,196785$. Suy ra hệ số tương quan mẫu

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\cdot\overline{Y}}{S_X S_Y} = 0,985661$$

rất gần với 1. Áp dụng công thức

$$y = \overline{Y} + r\frac{S_Y}{S_X}(x - \overline{X}),$$

ta tính được đường thẳng hồi quy của Y đối với X

$$y = 3,053964x - 3,42043$$

với sai số bình phương trung bình

$$\sigma^2_{Y/X} = S^2_Y(1 - r^2) = 1,09335.$$

Ví dụ 5.5.2

Bảng sau cho mẫu ngẫu nhiên của (X, Y) với kích thước mẫu $n = 15$. Hãy tìm đường thẳng hồi quy của Y đối với X.

X	Y
4,94	15,2
5,17	16,5
3,79	11
3,74	11,7
3,56	10,4
6,75	22,9
5,03	13,7
5,76	18,1
3,82	12,7
6,91	23,2
4,49	14,3
2,58	11
4,67	17,9
4,34	14,4
7	20,8

Kì vọng mẫu, Phương sai mẫu của X và Y

$$\overline{X} = 4,83708 \quad S_X = 1,26838$$

$$\overline{Y} = 15.5808 \quad S_Y = 4.08811$$

Hệ số tương quan mẫu $r = 0,938428$.

Vậy đường thẳng hồi quy của Y đối với X

$$y = 3,025x + 0,95.$$

Sai số bình phương trung bình $\sigma^2_{Y/X} = 1,99469$.

5.5.3 Kiểm định và ước lượng hệ số hồi quy

Bài toán tìm một hàm của đại lượng ngẫu nhiên X xấp xỉ một cách tốt nhất cho đại lượng ngẫu nhiên Y (theo nghĩa làm cực tiểu bình phương sai số) là một bài toán lớn trong thống kê. Trên đây ta đã chỉ ra phương pháp tìm đường hồi quy trung bình tuyến tính thực nghiệm $y = ax + b$.

Về mặt lí thuyết trong mục 5.4.1 chúng ta đã chỉ ra đường hồi quy trung bình tuyến tính, gọi phương trình của đường hồi quy lí thuyết đó là

$$y = \alpha x + \beta = \varrho\frac{\sqrt{DY}}{\sqrt{DX}}x + EY - \varrho\frac{\sqrt{DY}}{\sqrt{X}}EX$$

Thực ra nếu tính tuyến tính giữa Y và X quá yếu thì xấp xỉ bằng hồi quy trung bình tuyến tính thực nghiệm sẽ kém và do vậy trong thực tế ta phải hủy bỏ không thể chấp nhận xấp xỉ đó.

Ở phần cuối mục 3.7 chương III, chúng ta đã chứng minh rằng nếu (X, Y) có phân bố chuẩn đồng thời, khi đó hồi quy lí thuyết $E(Y/x)$ là hàm tuyến tính. Trong trường hợp này phương pháp ước lượng nêu trên khá tốt cho công tác dự báo.

Tuy nhiên trong thực tế, dựa vào giả thiết (X, Y) có phân bố chuẩn đồng thời, người ta cần giải quyết một số vấn đề liên quan tới bài toán hồi quy tuyến tính nhằm tăng độ tin cậy của dự báo bằng hồi quy.

1. **Kiểm định quan hệ tuyến tính của hàm hồi quy**

   Kiểm định giả thuyết

   (H): $\alpha = 0$ với đối thiết (K): $\alpha \neq 0$

   Nếu ta không có cơ sở bác bỏ giả thiết (H), ta coi hai đại lượng ngẫu nhiên X và Y không có mối quan hệ tuyến tính (khi đó chúng là các đại lượng ngẫu nhiên không tương quan hoặc coi chúng độc lập nhau). Trường hợp đối thiết (K) đúng ta có thể kiểm tra tiếp giả thiết $\alpha = \alpha_0$ nào đó.

   Dựa vào phép phân tích phương sai ta giải bài toán kiểm định nêu trên như sau:

Với kí hiệu $R_0^2 = ns_{Y.X}^2 = nS_Y^2(1-r^2)$,

$$\frac{R_0^2}{n-2} = \frac{nS_Y^2(1-r^2)}{n-2}$$

là ước lượng không chệch của đại lượng ngẫu nhiên. Kí hiệu $\sigma_a^2 = \frac{R_0^2}{n-2}$, ta có

$$E(\sigma_a^2) = \sigma^2.$$

($\frac{R_0^2}{\sigma^2(n-2)}$ có phân bố χ^2 với $n-2$ bậc tự do. σ_a được gọi là sai số chuẩn (standard error) của giá trị dự báo y với mỗi giá trị x.)

Vậy nếu giả thiết (H): $\alpha = 0$ đúng, thống kê

$$F = \frac{S_Y^2 - R_0^2}{\frac{R_0^2}{n-2}} = n(n-2)\frac{S_Y^2 r^2}{R_0^2} = \frac{(n-2)r^2}{1-r^2}$$

có phân bố-F với 1 và $n-2$ bậc tự do. Tương tự như kiểm định giả thiết về sự bằng nhau của các phương sai trong mục 5.3.4, tùy theo mức ý nghĩa tra bảng phân vị phân bố F với 1 và $n-2$ bậc tự do để xác định F_2. Nếu giá trị quan sát F kể trên nhỏ hơn F_2, ta không có cơ sở bác bỏ giả thiết (H), tức là không có mối quan hệ tuyến tính giữa Y và X.

Nhận xét rằng tương đương với việc dùng thống kê F, người ta dẫn vào thống kê

$$t = \frac{r\frac{S_Y}{S_X}}{\frac{\sigma_a}{\sqrt{n}S_X}} = \frac{r\frac{S_Y}{S_X}}{\frac{R_0}{\sqrt{n(n-2)}S_X}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Nếu giả thiết (H): $\alpha = 0$ đúng, thống kê t có phân bố Student với $n-2$ bậc tự do. Phương pháp kiểm định giả thiết sử dụng thống kê t đã được trình bày trong mục 5.3.2. Nếu giá trị quan sát t_{qs} có trị tuyệt đối lớn hơn phân vị t_ϵ, ta bác bỏ giả thiết (H): $\alpha = 0$ và ngược lại.

2. Khoảng tin cậy cho hệ số góc α của đường thẳng hồi quy

Ta công nhận các kết quả sau. Thống kê

$$t = \frac{(a-\alpha)S_X\sqrt{n-2}}{S_Y\sqrt{1-r^2}}$$

cũng có phân bố Student với $n-2$ bậc tự do (xem C.R.RAO) Do vậy áp dụng phương pháp ước lượng khoảng tin cậy cho giá trị trung bình đã được trình bày trong mục 5.3.3, ta nhận được khoảng tin cậy của α (hệ số góc của đường thẳng hồi quy lí thuyết) với độ tin cậy ϵ là

$$a - t_\epsilon \frac{S_Y\sqrt{1-r^2}}{S_X\sqrt{n-2}} < \alpha < a + t_\epsilon \frac{S_Y\sqrt{1-r^2}}{S_X\sqrt{n-2}}.$$

Phương sai của α và β

$$D(\alpha) = \frac{S_Y^2(1-r^2)}{(n-2)S_X^2} = \frac{\sigma_a^2}{nS_X^2}$$

$$D(\beta) = \frac{(1-r^2)S_Y^2(S_X^2+\overline{X}^2)}{(n-2)S_X^2} = \frac{\sigma_a^2(\sum_{i=1}^n X_i^2)}{n^2S_X^2}.$$

Trong thực hành ta coi $\sqrt{D(\alpha)}$ và $\sqrt{D(\beta)}$ là các sai số trung bình khi ước lượng các hệ số α và β của đường thẳng hồi quy lí thuyết $y = \alpha x + \beta$ bằng các hệ số a và b. Ta nhắc lại công thức tính các hệ số a và b trong đường thẳng hồi quy thực nghiệm $y = ax + b$.

$$a = r\frac{S_Y}{S_X} = S_Y\frac{\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})(Y_i-\overline{Y})}{S_X^2}$$

$$b = \overline{Y} - r\overline{X}\frac{S_Y}{S_X} = \overline{Y} - \overline{X}\frac{\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})(Y_i-\overline{Y})}{S_X^2}.$$

Chú ý rằng thống kê t trên đây cũng bằng

$$t = \frac{a-\alpha}{\sqrt{D(\alpha)}}.$$

Do vậy khoảng tin cậy của α còn có thể viết dưới dạng

$$a - t_\epsilon\sqrt{D(\alpha)} < \alpha < a + t_\epsilon\sqrt{D(\alpha)}.$$

3. Khoảng tin cậy cho đường thẳng hồi quy

Chúng ta kí hiệu Y là đại lượng ngẫu nhiên cần dự báo và $y(x)$ là giá trị hàm hồi quy tại x, trong đó hồi quy bình phương trung bình tuyến tính có dạng

$$y = \overline{Y} + a(x - \overline{X}).$$

Khi đó phương sai của hiệu $Y - y(x)$ bằng

$$D(Y - y) = \sigma_a^2 + D(y).$$

Mặt khác

$$D(y) = D(\overline{Y}) + (x - \overline{X})^2 \frac{\sigma_a^2}{S_X^2} = \sigma_a^2 \left(\frac{(x - \overline{X})^2}{S_X^2} + \frac{1}{n} \right)$$

Suy ra đại lượng ngẫu nhiên

$$t = \frac{Y - y}{\sigma_a \sqrt{1 + \frac{(x-\overline{X})^2}{S_X^2} + \frac{1}{n}}}$$

có phân bố Student với $n - 2$ bậc tự do.

Do vậy khoảng tin cậy cho Y tại x như đã được trình bày trong mục 5.2.3 bằng

$$y - t_\epsilon \sigma_a \sqrt{1 + \frac{(x - \overline{X})^2}{S_X^2} + \frac{1}{n}} < Y < y + t_\epsilon \sigma_a \sqrt{1 + \frac{(x - \overline{X})^2}{S_X^2} + \frac{1}{n}}.$$

Ta có nhận xét rằng nếu bài toán đặt ra dưới dạng tìm khoảng tin cậy cho giá trị trung bình với điều kiện $E(Y/X = x)$ tại $x = x^*$. Khi đó

$$t = \frac{E(Y/x^*) - y^*}{\sigma_a \sqrt{\frac{(x^*-\overline{X})^2}{S_X^2} + \frac{1}{n}}}$$

có phân bố Student với $n - 2$ bậc tự do ($y^* = \overline{Y} + a(x^* - \overline{X})$ là giá trị hàm hồi quy tại x^*). Vậy khoảng tin cậy cho kì vọng có điều kiện $E(Y/x^*)$ bằng

$$\left(y^* - t_\epsilon \sigma_a \sqrt{\frac{(x^* - \overline{X})^2}{S_X^2} + \frac{1}{n}} \quad ; \quad y^* + t_\epsilon \sigma_a \sqrt{\frac{(x^* - \overline{X})^2}{S_X^2} + \frac{1}{n}} \right).$$

VÍ DỤ ÁP DỤNG VÀO BÀI TOÁN DỰ BÁO LŨ

trong phần cuối của chương, chúng tôi dành cho việc trình bày một ví dụ sử dụng hồi quy vào bài toán dự báo. Giả sử chúng ta có các số liệu thống kê về tình hình lũ trên sông Hồng tại Hà nội từ năm 1969 đến 1992.

STT	Năm	Lượng mưa (Xmm)	Đỉnh lũ(Ycm)
1	1969	720	1405
2	1970	720	1405
3	1971	730	1439
4	1972	590	1133
5	1973	660	1272
6	1974	780	1519
7	1975	770	1524
8	1976	710	1364
9	1977	640	1253
10	1978	670	1324
11	1979	520	1002
12	1980	660	1303
13	1981	690	1337
14	1982	500	960
15	1983	460	879
16	1984	610	1176
17	1985	710	1382
18	1986	620	1178
19	1987	660	1271
20	1988	620	1194
21	1989	590	1161
22	1990	740	1449
23	1991	640	1225
24	1992	805	1377

Ta xét bài toán dự báo đỉnh lũ hàng năm trên sông Hồng tại Hà nội. Người ta coi đỉnh lũ tại Hà nội phụ thuộc tuyến tính vào lượng mưa trong tháng Sáu trên thượng nguồn sông Hồng (nằm cả trên đất Trung quốc và nước ta). Vấn đề dự báo lũ là công việc lớn đòi hỏi công việc nghiên cứu cơ bản và toàn diện. Ở nhiều nước trên thế giới người ta cùng hợp tác đưa ra các số liệu chính xác và sử dụng nhiều công cụ khác nhau để dự báo, trong

đó công cụ chính là hồi quy nhiều chiều. Trong ví dụ này chúng tôi chỉ đưa ra một ví dụ giả định nhằm giúp độc giả nghiên cứu cách sử dụng hồi quy đơn giản trong công việc dự báo. Việc nghiên cứu hồi quy nhiều chiều sẽ được trình bày riêng trong chương sau, chương cuối cùng. Ta tạm thời hài lòng với ví dụ giả định trên.

Kì vọng mẫu, phương sai mẫu của X và Y được tính trong bảng sau

$\overline{X}$	$\overline{Y}$	S_X^2	S_Y^2
$\frac{1}{n}\sum_{i=1}^n X_i$	$\frac{1}{n}\sum_{i=1}^n Y_i$	$\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2$	S_Y^2
658,95833	1272,16667	$85,02425^2$	$163,5071^2$

Hệ số tương quan mẫu

$$r = \frac{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})(Y_i-\overline{Y})}{S_X S_Y} = 0,97045.$$

Áp dụng công thức để tính các hệ số a và b của đường thẳng hồi quy $y = ax+b$

$$a = r\frac{S_Y}{S_X} = 1,86623$$

$$b = \overline{Y} - r\overline{X}\frac{S_Y}{S_X} = 42,39808.$$

Vậy đường thẳng hồi quy của Y đối với X

$$y = 1,86623x + 42,39808.$$

Sai số trung bình

$$\sigma_{Y/X} = S_Y\sqrt{1-r^2} = 39,45667087.$$

1. *Kiểm định quan hệ tuyến tính của hàm hồi quy*

Như đã trình bày ở trên, kiểm định về mối liên quan tuyến tính tương đương với kiểm định giả thuyết

(H): $\alpha = 0$ với đối thiết (K): $\alpha \neq 0$

Khi giả thiết (H): $\alpha = 0$ đúng, giá trị quan sát của thống kê

$$F_{qs} = \frac{(24-2)r^2}{1-r^2} = 355,7938$$

Với mức ý nghĩa $\epsilon = 0,05$ tra bảng phân vị phân bố F với 1 và $n-2 = 22$ bậc tự do, ta xác định

$$F_2 = 5,78632$$

Giá trị quan sát $F_{qs} = 355,7938$ lớn hơn rất nhiều so với $F_2 = 5,78632$, ta bác bỏ giả thiết (H): $\alpha = 0$, tức là mối quan hệ tuyến tính giữa Y và X khá chặt.

2. *Khoảng tin cậy cho hệ số góc α của đường hồi quy*

Thống kê

$$t = \frac{(a-\alpha)S_X\sqrt{n-2}}{S_Y\sqrt{1-r^2}}$$

có phân bố Student với 22 bậc tự do. Áp dụng công thức tìm khoảng tin cậy cho hệ số góc α

$$a - t_\epsilon \frac{S_Y\sqrt{1-r^2}}{S_X\sqrt{n-2}} < \alpha < a + t_\epsilon \frac{S_Y\sqrt{1-r^2}}{S_X\sqrt{n-2}}$$

Phân vị $t_{0,05} = 2,405468$, suy ra khoảng tin cậy cho hệ số góc α là

$$(1,628237 \quad ; \quad 2,104225)$$

3. *Sai số khi ước lượng các hệ số a và b của đường hồi quy*

Ta biết rằng

$$D(\alpha) = \frac{S_Y^2(1-r^2)}{(n-2)S_X^2}$$

$$D(\beta) = \frac{(1-r^2)S_Y^2(S_X^2 + \overline{X}^2)}{(n-2)S_X^2}.$$

Thay vào tính ta sẽ được các sai số khi ước lượng a và b. Sai số trung bình của a

$$\sqrt{D(\alpha)} = 0,098939$$

Sai số của b

$$\sqrt{D(\beta)} = 65,73696$$

4. *Khoảng tin cậy cho đường thẳng hồi quy*

Khoảng tin cậy cho giá trị trung bình với điều kiện $E(Y/X = x)$ tại $x = x^*$ của đường thẳng hồi quy, với độ tin cậy 95%

$$\left(y^* - t_{0,05}\sigma_a\sqrt{\frac{(x^* - \overline{X})^2}{S_X^2} + \frac{1}{n}};\quad y^* + t_{0,05}\sigma_a\sqrt{\frac{(x^* - \overline{X})^2}{S_X^2} + \frac{1}{n}} \right)$$

Nếu chọn $x^* = 800$, thay vào công thức trên, khoảng tin cậy cần tìm bằng

$$(1106,48 \quad ; \quad 1437,85)$$

Tuy nhiên khi x^* khá gần với giá trị trung bình $\overline{X} = 658,95833$, chẳng hạn $x^* = 600$ khoảng tin cậy cần nhận được sẽ thu hẹp hơn

$$(1200,51 \quad ; \quad 1343,82).$$

BÀI TẬP CHƯƠNG V

1. Cho mẫu ngẫu nhiên

$$(X_1, X_2, ..., X_n)$$

tương ứng với đại lượng ngẫu nhiên X. Chứng minh rằng phương sai mẫu

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

2. Cho mẫu ngẫu nhiên 2 chiều

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$$

của véc tơ ngẫu nhiên (X, Y). Chứng minh rằng

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{X}\cdot\overline{Y}.$$

3. Cho mẫu ngẫu nhiên sau của đại lượng ngẫu nhiên có phân bố chuẩn với kích thước mẫu $n = 30$:

   {4,18; 4,28; 4,86; 5,79; 6,55; 4,3; 4,51; 4,74; 4,1; 4,44; 5,54; 5,66; 4,18; 4,48; 5,01; 3,96; 4,85; 5,33; 4,95; 4,67; 6,18; 5,05; 5,04; 4,97; 4,7; 5,59; 4,06; 4,16; 5,41; 4,57}

   Hãy tìm khoảng tin cậy cho giá trị trung bình của phân bố đó với độ tin cậy $1 - \alpha = 95\%$.

4. Cho mẫu ngẫu nhiên sau của đại lượng ngẫu nhiên có phân bố chuẩn với kích thước mẫu $n = 23$:

   {6,34 ; 6,55 ; 6,84 ; 6,11 ; 7,34 ; 7,59 ; 6,77 ; 7,08 ; 7,45 ; 7,18 ; 6,94 ; 6,89 ; 6,86 ; 7,24 ; 7,73 ; 6,43 ; 7,53 ; 6,7 ; 6,85 ; 7,48 ; 7,09 ; 6,68 ; 7,46}

   Hãy tìm khoảng tin cậy cho giá trị trung bình của phân bố đó với độ tin cậy $1 - \alpha = 96\%$.

5. Cũng với bài trên, hãy kiểm định giả thiết (H): giá trị trung bình của đại lượng ngẫu nhiên đó bằng 6,8 (đối thiết (K): $m \neq 6,8$) với mức ý nghĩa của kiểm định $\alpha = 0,04$.

6. Giả sử đại lượng ngẫu nhiên X nhận các giá trị $1, 2, 3, ..., 12$ với các xác suất tương ứng là $p_1, p_2, p_3, ..., p_{12}$. Hãy kiểm định với mức ý nghĩa 5% tính phân bố đều của X trên miền giá trị đó của X.

   (Kiểm định giả thiết $P(X = i) = \frac{1}{12}$ với mọi $i = 1, 2, ..., 12$.)

   Biết các số liệu thống kê sau về X:

Giá trị của X	1	2	3	4	5	6
Số lần n_i	17	20	16	21	19	21

Giá trị của X	7	8	9	10	11	12
Số lần n_i	19	14	18	15	18	19

7. Gieo một xúc xắc 400 lần, ta thấy

Số chấm xuất hiện	1	2	3	4	5	6
Số lần n_i	70	72	60	65	66	67

   Hãy kiểm định với mức ý nghĩa 5%, xúc xắc đó đồng chất, đối xứng.

8. Đo chiều cao của 100 cây bạch đàn, kết quả được cho trong bảng dưới đây. (Chọn các giá trị mẫu là các trung điểm của các khoảng tương ứng). Hãy kiểm định với mức ý nghĩa 5% chiều cao của bạch đàn là đại lượng ngẫu nhiên có phân bố chuẩn.

Khoảng	Chiều cao x_i	n_i
<8,425	8,4	7
8,425 -8,475	8,45	5
8,475 -8,525	8,50	8
8,525 -8,575	8,55	10
8,575 -8,625	8,60	18
8,625 -8,675	8,65	17
8,675 -8,725	8,70	12
8,725 -8,775	8,75	9
8,775 -8,825	8,80	7
>8,825	8,85	7
Tổng cộng		100

9. Bảng sau cho mẫu ngẫu nhiên của (X, Y) với kích thước mẫu $n = 10$. Hãy tìm đường thẳng hồi quy của Y đối với X.

X	1	2	3	4	5
Y	1,9167	2,0201	6,91657	0,807614	-7,70526

X	6	7	8	9	10
Y	-0,590358	10,9637	-15,1728	-10,4491	-11,6077

10. Cho X là đại lượng ngẫu nhiên phân bố đều trên khoảng (a, b). Ta biết rằng $EX = \frac{a+b}{2}$. Bài toán đặt ra là dựa trên một mẫu cỡ n của X, hãy ước lượng $\frac{a+b}{2}$, trong đó a và b giả sử chưa biết. Gọi

$$X_1, X_2, ..., X_n$$

là mẫu ngẫu nhiên đó. Dễ dàng thấy kì vọng mẫu $\overline{X}$ là ước lượng không chệch của $\frac{a+b}{2}$. Chứng minh rằng

$$\frac{maxX_i + minX_i}{2}$$

cũng là ước lượng không chệch của $\frac{a+b}{2}$, đồng thời ước lượng đó hiệu quả hơn (có phương sai bé hơn) kì vọng mẫu $\overline{X}$.

ĐÁP SỐ VÀ HƯỚNG DẪN

1. Bạn đọc tự chứng minh.

2. Bạn đọc tự chứng minh.

3. Kì vọng mẫu $\overline{X} = 4,87033$

   Độ lệch chuẩn $S^* = 0,654588$

   Phân vị phân bố Student 29 bậc tự do ứng với $\alpha = 0,05$:

   $$t_\alpha = 2,04523.$$

   Khoảng tin cậy bằng

   $$(4,62591; 5,11476).$$

4. Kì vọng mẫu $\overline{X} = 7,00565$

   Độ lệch chuẩn $S^* = 0,432202$

   Phân vị phân bố Student 22 bậc tự do ứng với $\alpha = 0,04$:

   $$t_\alpha = 2,07387.$$

   Khoảng tin cậy bằng

   $$(6,80893; 7,20238).$$

5. Với giả thiết (H) đúng, tính giá trị quan sát:

   $$t = t_{qs} = 2,28197.$$

   Do $|t_{qs}| = 2,28197 > t_\alpha = 2,07387$, bác bỏ giả thiết giá trị trung bình của đại lượng ngẫu nhiên bằng 6,8.

6. Sử dụng tiêu chuẩn phù hợp χ^2:

$$q_{qs} = Q^2 = 3,03687.$$

Kích thước mẫu $n = 217$, Phân vị Phân bố χ^2 với 11 bậc tự do $\chi^2_{0,05} = 19,6751$. Do $q_{qs} = Q^2 = 3,03687 < \chi^2_{0,05} = 19,6751$, Chấp nhận giả thiết về tính phân bố đều của X

$$P(X = i) = \frac{1}{12} \quad \text{với mọi} \quad i = 1, 2, ..., 12.$$

7. Áp dụng tiêu chuẩn phù hợp χ^2:

$$q_{qs} = Q^2 = 1,31.$$

Phân vị Phân bố χ^2 với 5 bậc tự do $\chi^2_{0,05} = 11,07048$. Vậy ta chấp nhận giả thiết về tính phân bố đều của xúc xắc.

8. Kì vọng mẫu $\overline{X} = 8,631$ và độ lệch chuẩn $S = 0,121815$ là các ước lượng hợp lí cực đại của các tham số m và σ. Dựa vào đó, ta tính được giá trị quan sát

$$q_{qs} = Q^2 = 3,464337.$$

Phân vị Phân bố χ^2 với 10-2-1=7 bậc tự do $\chi^2_{0,05} = 14,06713$.

Do $q_{qs} = Q^2 = 3,464337 < \chi^2_{0,05} = 14,06713$, ta kết luận về tính phân bố chuẩn của chiều cao bạch đàn.

9. Kì vọng mẫu của X và Y

$$\overline{X} = 5,5 \quad \overline{Y} = -4,48279$$

Phương sai mẫu của X được tính theo công thức

$$S_X = 2,872281; \quad S_Y = 7,136875.$$

Suy ra hệ số tương quan mẫu

$$r = -0,84796.$$

Áp dụng công thức

$$y = \overline{Y} + r\frac{S_Y}{S_X}(x - \overline{X}),$$

ta tính được đường thẳng hồi quy của Y đối với X

$$y = -2,10697x + 7,105536.$$

với sai số bình phương trung bình

$$\sigma^2_{Y/X} = S^2_Y(1 - r^2) = 14,66116.$$

10. Gọi

$$\widehat{\Theta} = \frac{maxX_i + minX_i}{2}.$$

Khi đó tính

$$D(\widehat{\Theta}) = \frac{6(a-b)^2}{12(n+1)(n+2)}$$

trong khi $D(\overline{X}) = \frac{(a-b)^2}{12}$.

Chương 6

Tương quan bội và hồi quy bội

6.1 Hệ số tương quan

Trước tiên ta ôn lại một vài khái niệm quan trọng sẽ được sử dụng nhiều trong chương này. Trong lí thuyết xác suất, chúng ta biết rằng để đo mối quan hệ giữa hai hoặc nhiều đại lượng ngẫu nhiên, người ta thường tính các hệ số tương quan giữa chúng.

$$\varrho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{D(X)}\sqrt{D(X)}}.$$

Người ta chứng minh hệ số tương quan $\varrho(X,Y)$ là một số thực thuộc đoạn $[-1,1]$. Nếu X và Y là hai đại lượng ngẫu nhiên độc lập khi đó hệ số tương quan $\varrho(X,Y) = 0$. Khi $|\varrho(X,Y)|$ càng gần với 1 thì sự phụ thuộc tuyến tính giữa X và Y càng mạnh. Trường hợp $|\varrho(X,Y)| = 1$, giữa X và Y có mối quan hệ phụ thuộc tuyến tính $Y = aX + b$.

Trong thống kê, thay vì hai đại lượng ngẫu nhiên X, Y ta xét mẫu ngẫu nhiên

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n).$$

Có thể coi chúng như các điểm ngẫu nhiên trên mặt phẳng toạ độ. Chúng được phân bố theo phân bố của cặp 2 đại lượng ngẫu nhiên (X,Y). *Hệ số tương quan mẫu* được định nghĩa

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{S_X S_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{x} \cdot \overline{Y}}{S_X S_Y}$$

S_X^2, S_Y^2 là phương sai mẫu của X, Y tương ứng

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2, S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2.$$

Dễ dàng chứng minh được

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{S_X^* S_Y^*} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\cdot\overline{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2\right)}}.$$

Chú ý rằng $C(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}$ *được gọi là covarian mẫu.* Theo đó hệ số tương quan mẫu

$$r = \frac{C(X,Y)}{S_X \cdot S_Y}.$$

Phương sai mẫu của đại lượng ngẫu nhiên X cũng là covarian của X với chính nó: $S_X^2 = C(X,X)$. Lệnh $COVAR$ trong EXCEL được sử dụng để tính covarian mẫu (và do vậy cả phương sai mẫu của đại lượng ngẫu nhiên).

Quay trở lại với ví dụ về bài toán dự báo lũ sông Hồng tại Hà nội ở cuối chương trước, trang 265. Các số liệu giả định đã được cho trong bảng sau

STT	Năm	Lượng mưa (X)	Đỉnh lũ (Y)
1	1969	720	1405
2	1970	720	1405
3	1971	730	1439
4	1972	590	1133
5	1973	660	1272
6	1974	780	1519
7	1975	770	1524
8	1976	710	1364
9	1977	640	1253
10	1978	670	1324
11	1979	520	1002
12	1980	660	1303

STT	Năm	Lượng mưa (X)	Đỉnh lũ (Y)
13	1981	690	1337
14	1982	500	960
15	1983	460	879
16	1984	610	1176
17	1985	710	1382
18	1986	620	1178
19	1987	660	1271
20	1988	620	1194
21	1989	590	1161
22	1990	740	1449
23	1991	640	1225
24	1992	805	1377

Nếu ta minh hoạ các cặp số liệu $(x_i, y_i), i = 1, 2, ..., 24$ trong bảng trên bằng các điểm trên mặt phẳng, chúng ta cảm nhận thấy một mối liên hệ giữa lượng mưa (X) hàng năm và đỉnh lũ tại Hà nội (Y), lượng mưa càng lớn thì lũ do mưa gây nên càng cao. Hệ số tương quan mẫu sẽ giải thích mối quan hệ giữa hai đại lượng: lượng mưa hàng năm và đỉnh lũ tại Hà nội. Để tính hệ số tương quan mẫu giữa chúng, ta tính các đặc trưng mẫu của hai đại lượng ngẫu nhiên X và Y

$\overline{x}$	$\overline{y}$	S_x^2	S_y^2	$C(X,Y)$
$\frac{1}{n}\sum_{i=1}^n x_i$	$\frac{1}{n}\sum_{i=1}^n y_i$	$\frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2$	$\frac{1}{n}\sum_{i=1}^n (y_i-\overline{y})^2$	
658,95833	1272,16667	$85,02425^2$	$163,5071^2$	$13491,215$

Hệ số tương quan mẫu do vậy bằng

$$r = \frac{\frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})}{S_x S_y} = \frac{13491,215}{85,02425 \times 163,5071} = 0,97045.$$

(Trong EXCEL lệnh $CORREL(X_i, Y_i)$ cho kết quả là hệ số tương quan mẫu của X và Y). Dựa vào hệ số tương quan mẫu, sau này người ta giải thích được mức độ liên hệ giữa hai đại lượng ngẫu nhiên X và Y khi biểu diễn chúng thông qua mối quan hệ tuyến tính.

Lưu ý rằng để xây dựng quy tắc kiểm định về việc 2 đại lượng ngẫu nhiên có tương quan hay không, người ta đã chứng minh nếu (X_i, Y_i) có phân bố chuẩn 2 chiều, khi đó với giả thiết $\varrho(X,Y) = 0$

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \textit{có phân bố Student với} \quad n-2 \quad \textit{bậc tự do.}$$

6.2 Tương quan bội và hồi quy tuyến tính

6.2.1 Phương trình mặt phẳng hồi quy

Giả sử ta có $k+1$ đại lượng ngẫu nhiên $\eta, \xi_1, \xi_2, ..., \xi_k$ mô tả $k+1$ yếu tố ngẫu nhiên của một hiện tượng nào đó. Chúng ta sẽ dự đoán chẳng hạn η theo các đại lượng ngẫu nhiên còn lại $\xi_1, \xi_2, \xi_3, ..., \xi_k$. Như đã biết dự báo tốt nhất là hàm hồi quy và trong mục này ta chỉ dự đoán η bằng hàm tuyến tính của các đại lượng ngẫu nhiên còn lại. (Nếu $(\eta, \xi_1, \xi_2, ..., \xi_k)$ có phân bố chuẩn khi đó hàm hồi quy là hàm tuyến tính). Chúng ta cũng giả thiết $m = E(\eta) = 0, m_i = E(\xi_i) = 0$ với mọi $i = 1, 2, ...k$. (Trường hợp ngược lại ta sẽ tịnh tiến hệ trục toạ độ tới điểm $(m, m_1, m_2, ..., m_k)$ trong $\mathbb{R}^{k+1}$). Bài toán dự báo thực chất là tìm các hệ số b_i sao cho

$$E(\eta - b_1\xi_1 - b_2\xi_2 - ... - b_k\xi_k)^2 \to \min.$$

(Đây chính là *phương pháp bình phương bé nhất* để xác định các hệ số b_i), $y = b_1x_1 + b_{12}x_2 + ... + b_{1k}x_k$ được gọi là *mặt phẳng hồi quy tuyến tính*, các hệ số b_i được gọi là *các hệ số hồi quy*.

Giả sử rằng các đại lượng ngẫu nhiên $\eta, \xi_1, \xi_2, ..., \xi_k$ tồn tại phương sai (nói cách khác chúng thuộc không gian L_2 với tích vô hướng $\langle \xi, \eta \rangle = E(\xi\eta) = cov(\xi, \eta)$). Khi đó hình chiếu vuông góc của η lên không gian con sinh bởi $\xi_1, \xi_2, ..., \xi_k$ làm cho biểu thức $E(\eta - b_1\xi_1 - b_2\xi_2 - ... - b_k\xi_k)^2$ đạt giá trị bé nhất.

Gọi $\hat{\eta} = b_1\xi_1 + \cdots + b_k\xi_k$ là hình chiếu vuông góc của η lên không gian con sinh bởi $\xi_1, \xi_2, ..., \xi_k$, ta có:

$$\langle \xi_i, \eta - \hat{\eta} \rangle = \langle \xi_i, \eta - b_1\xi_1 - b_2\xi_2 - ... - b_k\xi_k \rangle = 0, \quad \forall\, i = 1, ..., k. \qquad (6.1)$$

Kí hiệu $\mathbf{c} = (c_{ij}) = (\sigma_i \sigma_j \varrho_{ij})$ là ma trận covarian (cấp $k+1$) của $\eta, \xi_1, \xi_2, ..., \xi_k$ và $\mathbf{A}$ là ma trận covarian (cấp k) của $\xi_1, \xi_2, ..., \xi_k$

$$\mathbf{c} = \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{10} & c_{11} & \cdots & c_{1k} \\ \cdots & \cdots & \cdots & \cdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{pmatrix} \quad A = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ c_{k1} & c_{k2} & \cdots & c_{kk} \end{pmatrix}$$

Gọi C_{ij} là phần phụ đại số tương ứng với c_{ij} của ma trận $\mathbf{c}$, R_{ij} là phần phụ đại số tương ứng với ϱ_{ij} của ma trận các hệ số tương quan $\mathbf{r} = (\varrho_{ij})$. Tất nhiên ta có thể giả thiết tiếp $C_{11} = \det A \neq 0$, điều đó luôn đúng nếu $\xi_1, \xi_2, ..., \xi_k$ độc lập tuyến tính. Khi đó hệ phương trình (6.1) có thể viết

$$\begin{cases} c_{11}b_1 + c_{12}b_2 + \cdots + c_{1k}b_k = c_{01} \\ c_{21}b_1 + c_{22}b_2 + \cdots + c_{2k}b_k = c_{01} \\ \quad \cdots \quad \cdots \quad \cdots \\ c_{k1}b_1 + c_{k2}b_2 + \cdots + c_{kk}b_k = c_{0k} \end{cases}$$

hoặc dưới dạng ma trận

$$\mathbf{Ab} = \mathbf{c_1}, \tag{6.2}$$

$\mathbf{b} = (b_1, \cdots, b_k)$ là véc tơ ẩn số, $\mathbf{c_1} = (c_{01}, \cdots, c_{0k})$ là covarian của η với các đại lượng ngẫu nhiên $\xi_1, \xi_2, ..., \xi_k$. Phương trình (6.2) có nghiệm duy nhất

$$\boxed{\mathbf{b} = \mathbf{A^{-1}c_1} \quad \text{hay} \quad b_i = -\frac{C_{0i}}{C_{00}} \quad i = 1, ..., k.} \tag{6.3}$$

Thật vậy, nhận xét rằng do $\det(\mathbf{c})\mathbf{c}^{-1} = (C_{ij})^T = (C_{ij})$ *hay* $(C_{ij})\mathbf{c} = \det(\mathbf{c})E$, *hàng thứ nhất của* (C_{ij}) : $(C_{00}, \mathbf{h}) = (C_{00}, C_{01}, ..., C_{0k})$ *vuông với cột thứ* $i, i \geq 1$ *của* $\mathbf{c}$, *suy ra* $\mathbf{Ah} = -C_{00}\mathbf{c_1}$ *hay*

$$\mathbf{b} = \mathbf{A^{-1}c_1} = -\frac{1}{C_{00}}\mathbf{h} \Leftrightarrow b_i = -\frac{C_{0i}}{C_{00}}, \quad i = 1, ..., k. \tag{6.4}$$

Các hệ số hồi quy $b_1, b_2, ..., b_k$ được tính thông qua ma trận covarian A nhờ công thức (6.3). Vậy phương trình của mặt phẳng hồi quy tuyến tính

$$y = \sum_{i=1}^{k} b_i x_i = -\sum_{i=1}^{k} \frac{C_{0i}}{C_{00}} x_i$$

Trường hợp tổng quát (m_i có thể khác 0)

$$\boxed{y = m + \sum_{i=1}^{k} b_i(x_i - m_i) = m - \sum_{i=1}^{k} \frac{C_{0i}}{C_{00}}(x_i - m_i).} \tag{6.5}$$

Như vậy phương trình mặt phẳng hồi quy có dạng $y = a + \sum_{i=1}^{k} b_i x_i$, trong đó hệ số tự do của mặt phẳng hồi quy theo công thức này $a = m - \sum_{i=1}^{k} b_i m_i$.

Người ta sử dụng mặt phẳng hồi quy để dự báo đại lượng ngẫu nhiên η theo các đại lượng ngẫu nhiên còn lại $\xi_1, \xi_2, \xi_3, ..., \xi_k$ bằng cách thay các giá trị của $\xi_1, \xi_2, \xi_3, ..., \xi_k$ vào mặt phẳng hồi quy

$$\eta \approx m + \sum_{i=1}^{k} b_i(\xi_i - m_i), \quad \text{trong đó} \quad m = E(\eta), m_i = E(\xi_i).$$

Các sai số của dự báo cũng như sai số của các hệ số hồi quy sẽ được trình bày trong mục sau.

Nhận xét rằng ta cũng có thể tính các hệ số hồi quy $b_1, b_2, ..., b_k$ thông qua các hệ số tương quan $\varrho_{ij} = \varrho(\xi_i, \xi_j)$ của ma trận $\mathbf{r} = (\varrho_{ij})$. Do $c_{ij} = \sigma_i \sigma_j \varrho_{ij}$, trong đó $\sigma_i^2 = D(\xi_i)$ là phương sai của ξ_i, suy ra

$$C_{0i} = (-1)^i \frac{\sigma_0^2 \sigma_1^2 \cdots \sigma_k^2}{\sigma_0 \sigma_i} \begin{vmatrix} \varrho_{10} & \varrho_{11} & \cdots & \varrho_{1i-1} & \varrho_{1i+1} & \cdots & \varrho_{1k} \\ \varrho_{20} & \varrho_{21} & \cdots & \varrho_{2i-1} & \varrho_{2i+1} & \cdots & \varrho_{2k} \\ \cdots & \cdots & \cdots & \cdots & & & \\ \varrho_{k0} & \varrho_{k1} & \cdots & \varrho_{ki-1} & \varrho_{ki+1} & \cdots & \varrho_{kk} \end{vmatrix}$$

$$= \frac{\sigma_0^2 \sigma_1^2 \cdots \sigma_k^2}{\sigma_0 \sigma_i} R_{0i}, \quad R_{ij} \text{ là phần phụ đại số ứng với } \varrho_{ij}.$$

Vậy

$$\boxed{b_i = -\frac{C_{0i}}{C_{00}} = -\frac{\sigma_0}{\sigma_i} \cdot \frac{R_{0i}}{R_{00}} \quad i = 1, ..., k.}$$

6.2.2 Cách tính mặt phẳng hồi quy

Trong thống kê, thay cho các giá trị chưa biết của các đại lượng ngẫu nhiên $\eta, \xi_1, \xi_2, ..., \xi_k$, người ta xét một mẫu ngẫu nhiên kích thước n

$$(y_i, x_{1i}, x_{2i}, ..., x_{ki}), \quad i = 1, 2, ..., n.$$

Phương trình mặt phẳng hồi quy sẽ được tính dựa trên các phần tử mẫu. Ma trận covarian A trong công thức (6.3) là ma trận các covarian mẫu và các kì vọng m, m_i trong công thức (6.5) là các kì vọng mẫu của ξ, ξ_i tương ứng.

Xét một ví dụ sau về mối quan hệ giữa sản lượng của một loại cây trồng (y) với chi phí đầu tư ban đầu (x_2) và lượng mưa trong cả đợt gieo trồng đó (x_1). Để tìm hồi quy tuyến tính của y theo x_1 và x_2, người ta dựa vào bảng số liệu quan sát về sản lượng của giống cây đó tại nhiều địa phương có thổ nhưỡng, khí hậu khác nhau.

STT	Y	x_1	x_2	STT	Y	x_1	x_2
1	590	58	405	14	710	62	560
2	660	52	450	15	620	54	420
3	780	133	350	16	660	48	620
4	770	179	285	17	620	86	390
5	710	98	330	18	590	74	350
6	640	72	400	19	740	95	570
7	670	72	550	20	730	44	710
8	520	43	480	21	720	53	700
9	660	62	450	22	720	77	580
10	690	67	610	23	640	46	700
11	500	64	380	24	805	123	560
12	460	33	460	25	510	26	370
13	610	57	425	26	673	62	430

Cột SST chỉ 26 địa phương khác nhau trồng giống cây đó.

Sử dụng lệnh $COVAR(Y, X)$ trong EXCEL để lập ma trận covarian

$$\mathbf{c} = \begin{pmatrix} 7507.100592 & 1852.139053 & 2870.872781 \\ 1852.139053 & 1060.408284 & -1448.16568 \\ 2870.872781 & -1448.16568 & 14221.48669 \end{pmatrix}$$

$$A = \begin{pmatrix} 1060.408284 & -1448.16568 \\ -1448.16568 & 14221.48669 \end{pmatrix}$$

Theo (6.3) các hệ số b_1, b_2 của mặt phẳng hồi quy được tính thông qua ma trận nghịch đảo A^{-1}. Sử dụng lệnh $MINVERSE$ để tính ma trận nghich đảo ta được

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \mathbf{A^{-1}c_1} = \begin{pmatrix} 0.0010954 & 0.0001115 \\ 0.0001115 & 0.00008167 \end{pmatrix} \begin{pmatrix} 1852.139053 \\ 2870.872781 \end{pmatrix} = \begin{pmatrix} 2.348974 \\ 0.441063 \end{pmatrix}$$

Như vậy các hệ số hồi quy

$$b_1 = 2.348974,\ \ b_2 = 0.441063.$$

Để tính hệ số tự do trong phương trình hồi quy $y = a + b_1x_1 + b_2x_2$, trong công thức (6.5), thay cho m, m_1, m_2 là các kì vọng mẫu $\overline{y} = AVERAGE(y_1, ..., y_n)$, và $\overline{x_1},\ \overline{x_2}$

$$\overline{y} = 653.7692, \overline{x_1} = 70.7692, \overline{x_2} = 482.1154 \ \Rightarrow\ a = \overline{y} - \sum_{i=1}^{2} b_i\overline{x_i} = 274.8907$$

Vậy phương trình mặt phẳng hồi quy $y = 2.348974x_1 + 0.441063x_2 + 274.8907$.

6.2.3 Hệ số tương quan bội và tương quan riêng

Như đã trình bày trong mục đầu, người ta sử dụng phương trình mặt phẳng hồi quy để dự báo η khi biết các giá trị $\xi_i,\ i = 1, 2, ..., k$.

Bài toán dự báo dựa trên giả thiết $\eta = \alpha + \beta_1\xi_1 + \beta_2\xi_2 + \cdots + \beta_k\xi_k + \varepsilon$ trong đó $E(\varepsilon) = 0$ và $D(\varepsilon) = \sigma^2$. Các hệ số hồi quy được tính toán (như ở ví dụ trên) là các ước lượng cho các tham số thực $\alpha, \beta_1, \beta_2, ..., \beta_k$

$$a = \hat{\alpha}, b_1 = \hat{\beta}_1, b_2 = \hat{\beta}_2, ..., b_k = \hat{\beta}_k$$

của hàm hồi quy $y = \alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k$. Đại lượng ngẫu nhiên $\hat{\eta} = a + b_1\xi_1 + \cdots + b_k\xi_k$ đưa ra giá trị dự báo, ta cũng gọi nó là giá trị hàm hồi quy. Khi đó *phần dư* $\eta - \hat{\eta}$ là sai số của dự báo. Hệ số tương quan giữa

η và $\hat{\eta}$ được gọi là *hệ số tương quan bội*, nó đo mức độ tác dụng tuyến tính của $\xi_1, \cdots, \xi_k$ lên η.

Nhận xét rằng khi tính các hệ số hồi quy, trong phần đầu của mục này ta đã coi các đại lượng ngẫu nhiên $\eta, \xi_1, \xi_2, ..., \xi_k$ như các phần tử thuộc không gian L_2 với tích vô hướng $\langle \xi, \eta \rangle = E(\xi\eta) = cov(\xi, \eta)$). Khi đó hệ số tương quan bội (tương quan giữa η và $\hat{\eta}$), kí hiệu $R = \varrho(\eta, \hat{\eta})$ bằng

$$R = \frac{cov(\eta, \hat{\eta})}{||\eta|| \cdot ||\hat{\eta}||} = \frac{\langle \hat{\eta}, \eta \rangle}{||\hat{\eta}|| \cdot ||\eta||} = \frac{\langle \hat{\eta}, \hat{\eta} + \eta - \hat{\eta} \rangle}{||\hat{\eta}|| \cdot ||\eta||} = \frac{\langle \hat{\eta}, \hat{\eta} \rangle}{||\hat{\eta}|| \cdot ||\eta||} = \frac{||\hat{\eta}||}{||\eta||}.$$

(*Trong không gian L_2 người ta thường kí hiệu $||\eta|| = \sqrt{\langle \eta, \eta \rangle}$ và gọi đó là chuẩn của η. Hiển nhiên $||\eta||^2$ chính là phương sai của η.* Nói cách khác

$$R^2 = \frac{D(\hat{\eta})}{D(\eta)}. \tag{6.6}$$

Trong một số tài liệu thống kê người ta gọi R^2 là *hệ số xác định* của hồi quy. Nó đo tỉ lệ phụ thuộc tuyến tính của η lên các biến ngẫu nhiên phụ thuộc $\xi_1, \cdots, \xi_k$. Hệ số xác định R^2 được tính thông qua ma trận covarian $\mathbf{c}$ và các phần phụ đại số tương ứng của $\mathbf{c}$. Thật vậy phương sai của phần dư $E(\eta - \hat{\eta})^2 = ||\eta - \hat{\eta}||^2 = \langle \eta - \hat{\eta}, \eta - \hat{\eta} \rangle = \langle \eta - \hat{\eta}, \eta \rangle$, áp dụng (6.4) bằng

$$\langle \eta + \frac{C_{01}\xi_1}{C_{00}} + \cdots + \frac{C_{0k}\xi_k}{C_{00}}, \eta \rangle = \frac{C_{00}c_{00}}{C_{00}} + \frac{C_{01}c_{01}}{C_{00}} + \cdots + \frac{C_{0k}c_{0k}}{C_{00}} = \frac{\det \mathbf{c}}{C_{00}} = \frac{\det \mathbf{c}}{\det A}$$

Suy ra

$$\boxed{R^2 = \frac{||\hat{\eta}||^2}{||\eta||^2} = 1 - \frac{||\eta - \hat{\eta}||^2}{||\eta||^2} = 1 - \frac{\det \mathbf{c}}{c_{00}C_{00}}.} \tag{6.7}$$

Lưu ý rằng người ta đã chứng minh phương sai của phần dư $E(\eta - \hat{\eta})^2 = \frac{n-k-1}{n}\sigma^2$. Do vậy

$$\boxed{s_e^2 = \frac{n}{n-k-1}||\eta - \hat{\eta}||^2 = \frac{n \det \mathbf{c}}{(n-k-1)\det A}} \tag{6.8}$$

là ước lượng không chệch của σ^2 và ta gọi $s_e = \sqrt{s_e^2}$ là *sai số tiêu chuẩn* của hồi quy. (*Ta cũng kí hiệu $\hat{\sigma}^2 = s_e^2$*).

Người ta cũng đã chứng minh các hệ số hồi quy $b_1, ..., b_k$ là ước lượng không chệch của $\beta_1, ..., \beta_k$ và do đó sai số của các hệ số hồi quy được suy ra từ ma trận covarian của $\mathbf{b}$

$$\boxed{cov(\mathbf{b}) = \mathbf{A^{-1}} cov(\mathbf{c_1}) \mathbf{A^{-1}} = \frac{\sigma^2}{n} \mathbf{A^{-1}}.} \tag{6.9}$$

Khi khảo sát mối tương quan ta tính hệ số tương quan giữa các đại lượng ngẫu nhiên, chẳng hạn $\varrho_{ij} = \varrho_{ij}(\xi_i, \xi_j)$. Đó là độ đo toàn phần mối tương quan giữa chúng (có kể đến mối quan hệ thông qua các biến ngẫu nhiên khác: $\xi_1, ..., \xi_k$). Như trên ta biết rằng có thể phân tích một đại lượng ngẫu nhiên thành tổng của hai đại lượng ngẫu nhiên không tương quan, chẳng hạn

$$\eta = \hat{\eta} + (\eta - \hat{\eta}) = \hat{\eta} + \eta_{0.23...k}, \quad \xi_1 = \hat{\xi_1} + (\xi_1 - \hat{\xi_1}) = \hat{\xi_1} + \eta_{1.23...k}$$

$\big(\hat{\eta}$ và $\hat{\xi_1}$ là hình chiếu vuông góc của η và ξ_1 xuống$L_2(\xi_2, ..., \xi_k)\big)$

Ta coi $\eta_{0.23...k} = \eta - \hat{\eta}$ là phần còn lại của η sau khi đã loại đi các tác động tuyến tính của $\xi_2, ..., \xi_k$ vào η. Tương tự $\eta_{1.23...k} = \xi_1 - \hat{\xi_1}$ là phần còn lại của ξ_1 sau khi đã loại đi các tác động tuyến tính của $\xi_2, ..., \xi_k$ vào ξ_1. Khi đó hệ số tương quan giữa hai phần dư $\xi_1 - \hat{\xi_1}$ và $\eta - \hat{\eta}$ được gọi là *hệ số tương quan riêng* (mối quan hệ nội tại, không phụ thuộc vào các đại lượng ngẫu nhiên khác: $\xi_2, ..., \xi_k$) giữa ξ_1 và η. Kí hiệu

$$\varrho_{01.(23...k)} = \varrho(\xi_1 - \hat{\xi_1}, \eta - \hat{\eta}).$$

Ta có thể chứng minh (như đã tính hệ số tương quan bội), hệ số tương quan riêng giữa η và ξ_1

$$\boxed{\varrho_{01.(23...k)} = \varrho(\xi_1 - \hat{\xi_1}, \eta - \hat{\eta}) = \frac{-C_{10}}{\sqrt{C_{00} C_{11}}}} \tag{6.10}$$

Một cách tổng quát hệ số tương quan riêng giữa ξ_i và ξ_j bằng

$$\boxed{\text{HS tương quan riêng:} \quad \varrho_{ij.(...)} = \frac{-C_{ij}}{\sqrt{C_{ii} C_{jj}}}} \tag{6.11}$$

(*Trong đó C_{ij} là phần phụ đại số tương ứng với c_{ij} của ma trận covarian* $\mathbf{c}$).

Trở lại với ví dụ về sản lượng giống cây trồng trong phần đầu mục 6.2.2, áp dụng các công thức từ (6.7) đến (6.11) trong mục này, ta có thể tính hệ số tương quan bội và các hệ số tương quan riêng giữa sản lượng giống cây và các nhân tố khác như lượng mưa, chi phí đầu tư ban đầu.

Ma trận các phần phụ đại số của ma trận covarian **c** (tính bằng EXCEL, $\det \mathbf{c} = 24541694726$)

$$(C_{ij}) = \det \mathbf{c} \cdot \mathbf{c}^{-1} = \begin{pmatrix} 12983398.46 & -30497670.32 & -5726501.492 \\ -30497670.32 & 98520220.59 & 16188781.03 \\ -5726501.492 & 16188781.03 & 4530172.584 \end{pmatrix}$$

Hệ số xác định $R^2 = 1 - \dfrac{\det \mathbf{c}}{c_{00}C_{00}} = 1 - \dfrac{24541694726}{7507.100592 * 12983398.46} = 0.748$ và do đó hệ số tương quan bội $R = \sqrt{0.748} = 0.865$.

Sai số tiêu chuẩn của hồi quy, theo (6.8)

$$s_e = \sqrt{\frac{n \det \mathbf{c}}{(n-k-1)\det A}} = \sqrt{\frac{26 \det \mathbf{c}}{23 \det A}} = 46.2254$$

Để tính sai số của các ước lượng hệ số hồi quy, ta sử dụng công thức (6.9), thay σ^2 bằng ước lượng s_e^2

$$cov(\mathbf{b}) = \frac{s_e^2}{n}\mathbf{A^{-1}} = \frac{s_e^2}{26}\begin{pmatrix} 1060.408284 & -1448.16568 \\ -1448.16568 & 14221.48669 \end{pmatrix}^{-1} =$$

$$= \frac{s_e^2}{26}\begin{pmatrix} 0.001095359 & 0.00011154 \\ 0.00011154 & 8.16742E-05 \end{pmatrix}$$

Thay $s_e = 46.2254$, suy ra sai số của các hệ số b_1, b_2

$$\sqrt{D(b_1)} = \frac{46.2254}{\sqrt{26}}\sqrt{0.001095359} = 0.300035$$

$$\sqrt{D(b_2)} = \frac{46.2254}{\sqrt{26}}\sqrt{8.16742E-05} = 0.08193$$

Hệ số tương quan riêng giữa Y và X_1, sử dụng công thức (6.10)

$$\varrho(Y, X_1) = \varrho_{01.(2)} = \frac{-C_{10}}{\sqrt{C_{00}C_{11}}} = 0.8527.$$

Chú ý rằng ta cũng có thể tính hệ số tương quan riêng bằng định nghĩa $\varrho_{01.(2)} = \varrho(Y - \hat{Y}, X_1 - \hat{X}_1)$.

Một cách khác để tính các hệ số hồi quy, hệ số tương quan bội cũng như các sai số khác là sử dụng lệnh $\{= LINEST(Y, X, 1, 1)\}$ trong EXCEL (nhấn đồng thời các phím CTRL+SHIFT+ENTER).

Kết quả của lệnh trên là bảng số

0.441063	2.348974	274.89068
0.08193	0.300035	52.1415458
0.7482	46.2254	
34.1724	23	
146038.4642	49146.151	

Hàng thứ nhất là các hệ số hồi quy $a = 274.89068, b_1 = 2.348974$ và $b_2 = 0.441063$

$$y = 274.89068 + 2.348974x_1 + 0.441063x_2$$

Sai số trung bình của các hệ số hồi quy a và b trong hàng thứ hai.

$$\sqrt{D(b_1)} = 0.300035 \quad \sqrt{D(b_2)} = 10.08193, \sqrt{D(a)} = 52.1415458$$

Hàng thứ ba là hệ số xác định $R^2 = 0.7482$, do vậy hệ số tương quan bội $R = 0.86499$ và sai số chuẩn (standard error) $s_e = 46.2254$.

Hàng thứ tư cho giá trị quan sát $F_{qs} = 34.1724$ của phân bố F với $(k, 23)$ bậc tự do. (Trong ví dụ này $k = 2$).

Hàng thứ năm là các tổng bình phương hồi quy theo Y, thường được kí hiệu là $SSR = 146038.4642$ và phần dư $SSE = 49146.151$.

6.2.4 Khoảng tin cậy và kiểm định giả thiết cho các tham số của hồi quy

Sử dụng các lệnh Excel, bạn đọc hãy tự giải

Bài tập Bảng sau cho ta số liệu quan sát được về kết quả học tập của học sinh. Giả thiết mô hình hồi quy giữa chúng

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

trong đó

Y là điểm trung bình chung của học sinh cuối năm thứ nhất.

x_1 là điểm thi tốt nghiệp phổ thông trung học của học sinh.

x_2 là điểm thi tuyển vào đại học của học sinh.

x_3 là điểm thi môn toán kì I của học sinh.

STT	x_1	x_2	x_3	Y
1	45	25	6	5.88
2	43	24.5	7	6.63
3	50	26	7	7.57
4	46	22	8	7.79
5	46	21	5	5.5
6	51	26	8	8.39
7	48	27	9	8.44
8	43	25	8	7.75
9	52	23	6	6.48
10	50	23.5	8	7.81
11	48	25	7	7.12
12	51	22.5	9	8.87
13	55	24	6	6.9

1. Viết phương trình mặt phẳng hồi quy Y theo x_1, x_2, x_3 và dự báo điểm trung bình chung cuối năm thứ nhất cho một học sinh nếu điểm thi tốt nghiệp phổ thông trung học $x_1 = 53$, điểm thi tuyển vào đại học $x_2 = 28$, và điểm thi môn toán kì I của học sinh đó $x_3 = 8$.

2. Hãy tính hệ số tương quan bội và hệ số tương quan riêng giữa điểm trung bình chung cuối năm thứ nhất và điểm thi tuyển vào đại học.

3. Hãy tính khoảng tin cậy cho β_1 với độ tin cậy 96%. Kiểm định giả thiết $\beta_2 = 0$ với mức ý nghĩa 5%.

Các vấn đề về khoảng tin cậy và kiểm định giả thiết cho các tham số của hồi quy dựa vào các kết quả sau

Với các giả thiết thêm rằng các đại lượng ngẫu nhiên có phân bố chuẩn. Kí hiệu $s_{b_k}, s_{b_{k-1}}, ..., s_{b_2}, s_{b_1} s_a$ *là các sai số chuẩn của các hệ số hồi quy* $b_k, b_{k-1}, ..., b_2, b_1, a$, *khi đó*

$$t_a = \frac{a - \alpha}{s_a}, \quad t_{b_i} = \frac{b_i - \beta_i}{s_{b_i}}, \quad i = 1, 2, ..., k$$

là các đại lượng ngẫu nhiên có phân bố Student với $n - k - 1$ *bậc tự do.*

Chẳng hạn trong bài tập trên, hệ số hồi quy của x_1 (điểm thi tốt nghiệp phổ thông) được ước lượng bằng $b_1 = 0.770966$ với độ lệch tiêu chuẩn $s_{b_1} = 0.054249$. Đại lượng ngẫu nhiên tương ứng $t_{b_i} = \frac{b_i - \beta_i}{s_{b_i}}$ có phân bố Student với $n - k - 1 = 9$ bậc tự do. Vậy khoảng tin cậy cho β_1 với độ tin cậy $1 - \alpha$ cho trước được tính theo công thức

$$b_1 - s_{b_1} t_\alpha \leqslant \beta_1 \leqslant b_1 + s_{b_1} t_\alpha$$

Trong bài tập trên, khoảng tin cậy cho β_1 với độ tin cậy 96% bằng $(0.64, 0.901)$

Kiểm định giả thiết cho mỗi tham số của hồi quy

Cũng dựa trên cơ sở t_{b_i} có phân bố Student với $n - k - 1$ bậc tự do, ta có thể kiểm định các giả thiết

$$H_0: \quad \beta_i = \beta_{i,0} \quad \text{hoặc} \qquad H_0: \quad \beta_i \leqslant \beta_{i,0} \quad \text{hoặc} \qquad H_0: \quad \beta_i \geqslant \beta_{i,0}$$

với đối thiết tương ứng.

Bài toán (1): Kiểm định giả thiết $H_0: \quad \beta_i = \beta_{i,0}$ *với đối thiết*

$$H_1: \qquad \beta_i \neq \beta_{i,0},$$

$$\text{theo quy tắc bác bỏ} \quad H_0 \quad \text{nếu} \qquad |t_{qs}| = \left| \frac{b_i - \beta_{i,0}}{s_{b_i}} \right| > t_\alpha.$$

Bài toán (2): Kiểm định giả thiết

$$H_0: \quad \beta_i = \beta_{i,0} \qquad \textit{hoặc} \qquad H_0: \quad \beta_i \leqslant \beta_{i,0}$$

với đối thiết

$$H_1: \qquad \beta_i > \beta_{i,0},$$

$$\text{theo quy tắc bác bỏ} \quad H_0 \quad \text{nếu} \qquad t_{qs} = \frac{b_i - \beta_{i,0}}{s_{b_i}} > t_\alpha.$$

Bài toán (3): Kiểm định giả thiết

$$H_0: \quad \beta_i = \beta_{i,0} \qquad \textit{hoặc} \qquad H_0: \quad \beta_i \geq \beta_{i,0}$$

với đối thiết

$$H_1: \qquad \beta_i < \beta_{i,0},$$

$$\text{theo quy tắc bác bỏ} \quad H_0 \quad \text{nếu} \qquad t_{qs} = \frac{b_i - \beta_{i,0}}{s_{b_i}} < -t_\alpha.$$

Đặc biệt nếu giá trị thực của $\beta_1 = 0$, $Y_i = \alpha + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$ không bị ảnh hưởng bởi biến độc lập X_1 khi các biến $X_2, ..., X_k$ nhận các giá trị cố định cho trước. Nói cách khác X_1 không góp phần vào giải thích mối quan hệ tuyến tính giữa biến phụ thuộc với các biến độc lập.

Trong bài tập về các loại điểm thi của học sinh ở đầu mục này, xét bài toán kiểm định $H_0: \quad \beta_2 = 0$ với đối thiết $H_1: \quad \beta_2 \neq 0$

$$t_{qs} = \frac{b_2 - 0}{s_{b_2}} = \frac{0.011599}{0.038539} = 0.30098 < t_{0.05} = 2.262,$$

ta chưa có cơ sở bác bỏ $H_0: \ \beta_2 = 0$ ở mức 0.5%.

Tính hệ số tương quan riêng giữa điểm trung bình chung cuối năm thứ nhất (Y) và điểm thi tuyển vào đại học (x_2). Ta được hệ số tương quan riêng đó khá bé $r = 0.0998$. Trong trường hợp này ta chấp nhận giả thiết $H_0: \ \beta_2 = 0$, và tìm hồi quy Y theo 2 biến còn lại: điểm thi tốt nghiệp phổ thông trung học và điểm thi môn toán kì I của học sinh.

Kiểm định giả thiết đồng thời cho các tham số của hồi quy

$$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

với đối thiết

$$H_1: \qquad \text{Tồn tại ít nhất một } i: \beta_i \neq 0.$$

Nếu giả thiết H_0 đúng, $Y_i = \alpha + \varepsilon_i$, nên $E(Y_i/X) = \alpha$ là hằng số. Các biến độc lập X_i không có ảnh hưởng (tuyến tính) tới Y. Kiểm định giả thiết H_0 thực chất nhằm bác bỏ tính phụ thuộc tuyến tính giữa các biến. Ta biết rằng $SST = SSR + SSE$, trong đó SSR nhằm giải thích sự biến động của hồi quy (sự phụ thuộc tuyến tính của biến phụ thuộc vào các biến độc lập), còn SSE là phần biến động ngoài hồi quy. Do vậy nếu giữa các biến ngẫu nhiên không tồn tại quan hệ tuyến tính khi đó SSR tương đối nhỏ so với SSE, nói cách khác tỉ số giữa SSR và SSE càng lớn, khả năng bác bỏ giả thiết không (quan hệ tuyến tính) càng cao. Vì thế để tạo ra một thống kê như vậy người ta sử dụng kết quả sau:

Nếu giả thiết $H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0$ đúng và ε_i có phân bố chuẩn, khi đó

$$F = \frac{SSR/k}{SSE/(n-k-1)}$$

có phân bố F với $(k, n-k-1)$ bậc tự do. Vậy ta có quy tắc ở mức α

$$\text{Bác bỏ} \quad H_0 \quad \text{nếu} \qquad F_{qs} = \frac{SSR/k}{SSE/(n-k-1)} > F_{k,n-k-1,\alpha},$$

trong đó

$$P(F_{k,n-k-1} > F_{k,n-k-1,\alpha}) = \alpha.$$

Nhận xét rằng do $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$, *suy ra*

$$F = \frac{SSR/k}{SSE/(n-k-1)} = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2}.$$

Kiểm định giả thiết đồng thời cho một tập con các tham số của hồi quy

Giả thiết rằng ta cần kiểm định k_1 tham số đầu tiên của hồi quy bằng 0.

$$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_{k_1} = 0$$

(Với đối thiết H_1: Tồn tại ít nhất một $i, 1 \leqslant i \leqslant k_1 : \beta_i \neq 0$.)

Nếu giả thiết H_0 đúng, các biến $X_1, X_2, ..., X_{k_1}$ không có ảnh hưởng gì tới Y, do vậy ta tiến hành ước lượng hồi quy của Y chỉ thông qua các biến $X_{k_1+1}, X_{k_1+2}, ..., X_k$

$$Y_i = \alpha^* + \beta^*_{k_1+1} x_{k_1+1,i} + \cdots + \beta^*_k x_{ki} + \varepsilon^*_i$$

Khi đó ta hy vọng SSE của mẫu hồi quy cũ khác nhiều so với SSE^* của mẫu hồi quy mới.

Thống kê

$$F = \frac{(SSR^* - SSE)/k_1}{SSE/(n-k-1)}$$

có phân bố F với $(k_1, n-k-1)$ bậc tự do. Vậy ta có quy tắc ở mức α

$$\text{Bác bỏ} \quad H_0 \quad \text{nếu} \qquad F_{qs} = \frac{(SSE^* - SSE)/k_1}{SSE/(n-k-1)} > F_{k_1,n-k-1,\alpha}.$$

Dự báo

Với mẫu hồi quy như đã nói ở trên, kí hiệu $a, b_1, b_2, ..., b_k$ là các ước lượng theo phương pháp bình phương bé nhất các hệ số hồi quy, khi đó với mẫu thứ $n+1$ của các biến độc lập:

$$(x_{1,n+1}, x_{2,n+1}, ..., x_{k,n+1})$$

dự báo của biến phụ thuộc ($Y_{n+1} = \alpha + \beta_1 x_{1,n+1} + \cdots + \beta_k x_{k,n+1} + \varepsilon_{n+1}$)

$$\hat{Y}_{n+1} = a + b_1 x_{1,n+1} + b_2 x_{2,n+1} + \cdots + b_k x_{k,n+1}$$

là ước lượng tuyến tính không chệch tốt nhất của Y_{n+1}.

Trở lại với bài tập về điểm trung bình chung cuối năm thứ nhất của học sinh, nếu điểm thi tốt nghiệp phổ thông trung học $x_1 = 53$, điểm thi tuyển

vào đại học $x_2 = 28$, và điểm thi môn toán kì I của học sinh đó $x_3 = 8$, khi đó điểm trung bình chung cuối năm thứ nhất của học sinh được dự báo là

$$\hat{Y}_{n+1} = a + b_1 x_{1,n+1} + b_2 x_{2,n+1} + b_3 x_{3,n+1} = 8.32$$

Ngoài ra nếu giả thiết ε_i có phân bố chuẩn khi đó chúng ta có thể tính các khoảng tin cậy cho các dự báo $\hat{Y}_{n+1}$.

6.3 Một vài lệnh EXCEL sử dụng trong các bài toán thống kê

Dưới đây chúng ta liệt kê một vài câu lệnh EXCEL để tính hàm phân bố, hàm mật độ của các đại lượng ngẫu nhiên quen thuộc cũng như tính các đặc trưng mẫu tương ứng với các đại lượng ngẫu nhiên đó.

1. $AVERAGE(x_1, x_2, , , , , x_n)$ cho giá trị trung bình mẫu $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

2. $DEVSQ(x_1, x_2, , , , , x_n)$ cho giá trị bằng tổng bình phương độ lệch $ns^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$. Vậy phương sai mẫu s^2 được tính bằng

$$s^2 = \frac{1}{n} DEVSQ(x_1, x_2, , , , , x_n).$$

3. $COVAR(\{x_1, x_2, ..., x_n\}, \{x_1, x_2, ..., x_n\})$ cho giá trị phương sai mẫu $s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

4. $VAR(x_1, x_2, , , , , x_n)$ cho giá trị phương sai mẫu điều chỉnh $s^{*2} = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

5. $STDEV(x_1, x_2, ..., x_n)$ cho giá trị độ lệch mẫu điều chỉnh

$$s^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

6. $CONFIDENCE(\alpha, \sigma, n)$ cho giá trị $u_\alpha \frac{\sigma}{\sqrt{n}}$ trong công thức tính khoảng tin cậy của m, trường hợp σ đã biết

$$\left(\overline{X} - u_\alpha \frac{\sigma}{\sqrt{n}}\,,\, \overline{X} + u_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

Nói cách khác lệnh $CONFIDENCE(\alpha, \sigma, n)$ để tính *bán kính khoảng tin cậy*. Như vậy khoảng

$$\left(\overline{X} - CONFIDENCE(\alpha, \sigma, n)\,,\, \overline{X} + CONFIDENCE(\alpha, \sigma, n)\right)$$

là khoảng tin cậy cho giá trị trung bình m với độ tin cậy $1 - \alpha$.

Các lệnh EXCEL dưới đây để tính giá trị các hàm phân bố và các phân vị của chúng

1. $NORMSDIST(x)$ cho giá trị hàm phân bố chuẩn tại x của đại lượng ngẫu nhiên có phân bố chuẩn $u \in N(0, 1)$.

$$NORMSDIST(x) = P(u < x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} dx$$

2. $NORMSINV(p)$ cho giá trị u_p sao cho $\Phi(u_p) = p$. Nói cách khác lệnh $NORMSINV(p)$ là hàm ngược của hàm phân bố chuẩn $\Phi(x)$.

Người ta sử dụng lệnh này để tìm phân vị u_α trong công thức lập khoảng tin cậy cho giá trị trung bình m với độ tin cậy $1 - \alpha$ trường hợp phương sai σ^2 đã biết, mục 5.3.3. Từ tính chất hàm phân bố chuẩn $\Phi(x)$, ta có

$$P(|u| < u_\alpha) = 1 - \alpha = \frac{1}{\sqrt{2\pi}} \int_{-u_\alpha}^{u_\alpha} e^{-\frac{x^2}{2}} dx \Leftrightarrow u_\alpha = NORMSINV(1 - \frac{\alpha}{2}).$$

3. $NORMDIST(x, m, \sigma, 1)$ cho giá trị hàm phân bố chuẩn tại x của đại lượng ngẫu nhiên phân bố chuẩn $X \in N(m, \sigma^2)$ có kì vọng bằng m và phương sai σ^2.

$$NORMDIST(x) = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

Lưu ý rằng lệnh $NORMDIST(x, m, \sigma, 0)$ cho giá trị hàm mật độ phân bố chuẩn tại x.

4. $GAMMADIST(x, p, \frac{1}{\alpha}, 0)$ cho giá trị hàm mật độ của phân bố Gamma với hai tham số dương α và p.

$$GAMMADIST(x, p, \frac{1}{\alpha}, 0) = G(x, \alpha, p) = \frac{\alpha^p}{\Gamma(p)} \cdot e^{-\alpha x} x^{p-1}.$$

5. $GAMMADIST(x, p, \frac{1}{\alpha}, 1)$ cho giá trị hàm phân bố Gamma

$$GAMMADIST(x, p, \frac{1}{\alpha}, 1) = P(X < x) = \frac{\alpha^p}{\Gamma(p)} \int_{0}^{x} e^{-\alpha x} x^{p-1} dx.$$

6. $CHIDIST(x, n)$ cho giá trị phía đuôi hàm phân bố χ^2 với n bậc tự do

$$CHIDIST(x, n) = P(X > x) = \frac{\alpha^p}{\Gamma(p)} \int_{x}^{+\infty} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} dx.$$

Lưu ý rằng trong mục 1.3 ta đã biết phân bố χ^2 với n bậc tự do là trường hợp đặc biệt của phân bố Gamma với $\alpha = \frac{1}{2}, p = \frac{n}{2}$.

7. $CHIINV(\alpha, n)$ cho giá trị là phân vị χ^2_α mức α xác định từ hệ thức

$$P(\chi^2 > \chi^2_\alpha) = \alpha$$

đã được sử dụng trong 5.4.6 và 5.4.7, mục kiểm định tính phù hợp hay kiểm định tính độc lập. Lệnh này thực chất chính là hàm ngược của hàm lệnh $CHIDIST(x, n)$.

8. $FDIST(x, m, n)$ cho giá trị phía đuôi hàm phân bố F với m, n bậc tự do

$$FDIST(x, m, n) = P(X > x) = \int_{x}^{+\infty} \left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{x^{\frac{m}{2}-1}}{(1 + \frac{mx}{n})^{\frac{m+n}{2}}}$$

9. $FINV(\alpha, m, n)$ cho giá trị là phân vị F_α mức α xác định từ hệ thức $P(X > F_\alpha) = \alpha$. Lệnh này thực chất chính là hàm ngược của hàm lệnh $FDIST(x, m, n)$.

10. $TDIST(x, n)$ cho giá trị phía đuôi "kép" của hàm phân bố Student với n bậc tự do

$$TDIST(x, n) = P(|X| > x) = 1 - 2\int_0^x \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n}\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} dt$$

11. $TINV(\alpha, n)$ cho giá trị là phân vị t_α mức α xác định từ hệ thức $P(|X| > t_\alpha) = \alpha$.
Lưu ý rằng trong mục 1.3 ta đã nhắc tới mật độ của phân bố Student tiến dần đến mật độ của phân bố chuẩn $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ *khi* n *dần ra vô cùng. Suy ra với* n *đủ lớn*

$$TINV(\alpha, n) \approx NORMSINV(1 - \frac{\alpha}{2}).$$

12. $LINEST(Y, X, 1, 1)$ sử dụng trong hồi quy bội, cho kết quả là một bảng số mà ý nghĩa của chúng được chỉ ra trong mục 6.2, trang 286.

Hàm phân bố chuẩn lớp $N(0,1)$.

Bảng sau cho giá trị hàm phân bố chuẩn $N(0,1)$

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

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,500	0,35	0,637	0,70	0,758	1,05	0,853	1,40	0,919
0,01	0,504	0,36	0,641	0,71	0,761	1,06	0,855	1,41	0,921
0,02	0,508	0,37	0,644	0,72	0,764	1,07	0,858	1,42	0,922
0,03	0,512	0,38	0,648	0,73	0,767	1,08	0,860	1,43	0,924
0,04	0,516	0,39	0,652	0,74	0,770	1,09	0,862	1,44	0,925
0,05	0,520	0,40	0,655	0,75	0,773	1,10	0,864	1,45	0,926
0,06	0,524	0,41	0,659	0,76	0,776	1,11	0,867	1,46	0,928
0,07	0,528	0,42	0,663	0,77	0,779	1,12	0,869	1,47	0,929
0,08	0,532	0,43	0,666	0,78	0,782	1,13	0,871	1,48	0,931
0,09	0,536	0,44	0,670	0,79	0,785	1,14	0,873	1,49	0,932
0,10	0,540	0,45	0,674	0,80	0,788	1,15	0,875	1,50	0,933
0,11	0,544	0,46	0,677	0,81	0,791	1,16	0,877	1,51	0,934
0,12	0,548	0,47	0,681	0,82	0,794	1,17	0,879	1,52	0,936
0,13	0,552	0,48	0,684	0,83	0,797	1,18	0,881	1,53	0,937
0,14	0,556	0,49	0,688	0,84	0,800	1,19	0,883	1,54	0,938
0,15	0,560	0,50	0,691	0,85	0,802	1,20	0,885	1,55	0,939
0,16	0,564	0,51	0,695	0,86	0,805	1,21	0,887	1,56	0,941
0,17	0,567	0,52	0,698	0,87	0,808	1,22	0,889	1,57	0,942
0,18	0,571	0,53	0,702	0,88	0,811	1,23	0,891	1,58	0,943
0,19	0,575	0,54	0,705	0,89	0,813	1,24	0,893	1,59	0,944
0,20	0,579	0,55	0,709	0,90	0,816	1,25	0,894	1,60	0,945
0,21	0,583	0,56	0,712	0,91	0,819	1,26	0,896	1,61	0,946
0,22	0,587	0,57	0,716	0,92	0,821	1,27	0,898	1,62	0,947
0,23	0,591	0,58	0,719	0,93	0,824	1,28	0,900	1,63	0,948
0,24	0,595	0,59	0,722	0,94	0,826	1,29	0,901	1,64	0,949
0,25	0,599	0,60	0,726	0,95	0,829	1,30	0,903	1,65	0,951
0,26	0,603	0,61	0,729	0,96	0,831	1,31	0,905	1,66	0,952
0,27	0,606	0,62	0,732	0,97	0,834	1,32	0,907	1,67	0,953
0,28	0,610	0,63	0,736	0,98	0,836	1,33	0,908	1,68	0,954
0,29	0,614	0,64	0,739	0,99	0,839	1,34	0,910	1,69	0,954
0,30	0,618	0,65	0,742	1,00	0,841	1,35	0,911	1,70	0,955
0,31	0,622	0,66	0,745	1,01	0,844	1,36	0,913	1,71	0,956
0,32	0,626	0,67	0,749	1,02	0,846	1,37	0,915	1,72	0,957
0,33	0,629	0,68	0,752	1,03	0,848	1,38	0,916	1,73	0,958
0,34	0,633	0,69	0,755	1,04	0,851	1,39	0,918	1,74	0,959

Hàm phân bố chuẩn lớp $N(0,1)$

Bảng sau cho giá trị hàm phân bố chuẩn $N(0,1)$

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

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,75	0,960	2,10	0,982	2,45	0,993	2,80	0,997	3,15	0,999
1,76	0,961	2,11	0,983	2,46	0,993	2,81	0,998	3,16	0,999
1,77	0,962	2,12	0,983	2,47	0,993	2,82	0,998	3,17	0,999
1,78	0,962	2,13	0,983	2,48	0,993	2,83	0,998	3,18	0,999
1,79	0,963	2,14	0,984	2,49	0,994	2,84	0,998	3,19	0,999
1,80	0,964	2,15	0,984	2,50	0,994	2,85	0,998	3,20	0,999
1,81	0,965	2,16	0,985	2,51	0,994	2,86	0,998	3,21	0,999
1,82	0,966	2,17	0,985	2,52	0,994	2,87	0,998	3,22	0,999
1,83	0,966	2,18	0,985	2,53	0,994	2,88	0,998	3,23	0,999
1,84	0,967	2,19	0,986	2,54	0,994	2,89	0,998	3,24	0,999
1,85	0,968	2,20	0,986	2,55	0,995	2,90	0,998	3,25	0,999
1,86	0,969	2,21	0,986	2,56	0,995	2,91	0,998	3,26	0,999
1,87	0,969	2,22	0,987	2,57	0,995	2,92	0,998	3,27	0,999
1,88	0,970	2,23	0,987	2,58	0,995	2,93	0,998	3,28	0,999
1,89	0,971	2,24	0,987	2,59	0,995	2,94	0,998	3,29	0,999
1,90	0,971	2,25	0,988	2,60	0,995	2,95	0,998	3,30	1,000
1,91	0,972	2,26	0,988	2,61	0,995	2,96	0,998	3,31	1,000
1,92	0,973	2,27	0,988	2,62	0,996	2,97	0,999	3,32	1,000
1,93	0,973	2,28	0,989	2,63	0,996	2,98	0,999	3,33	1,000
1,94	0,974	2,29	0,989	2,64	0,996	2,99	0,999	3,34	1,000
1,95	0,974	2,30	0,989	2,65	0,996	3,00	0,999	3,35	1,000
1,96	0,975	2,31	0,990	2,66	0,996	3,01	0,999	3,36	1,000
1,97	0,976	2,32	0,990	2,67	0,996	3,02	0,999	3,37	1,000
1,98	0,976	2,33	0,990	2,68	0,996	3,03	0,999	3,38	1,000
1,99	0,977	2,34	0,990	2,69	0,996	3,04	0,999	3,39	1,000
2,00	0,977	2,35	0,991	2,70	0,997	3,05	0,999	3,40	1,000
2,01	0,978	2,36	0,991	2,71	0,997	3,06	0,999	3,41	1,000
2,02	0,978	2,37	0,991	2,72	0,997	3,07	0,999	3,42	1,000
2,03	0,979	2,38	0,991	2,73	0,997	3,08	0,999	3,43	1,000
2,04	0,979	2,39	0,992	2,74	0,997	3,09	0,999	3,44	1,000
2,05	0,980	2,40	0,992	2,75	0,997	3,10	0,999	3,45	1,000
2,06	0,980	2,41	0,992	2,76	0,997	3,11	0,999	3,46	1,000
2,07	0,981	2,42	0,992	2,77	0,997	3,12	0,999	3,47	1,000
2,08	0,981	2,43	0,992	2,78	0,997	3,13	0,999	3,48	1,000
2,09	0,982	2,44	0,993	2,79	0,997	3,14	0,999	3,49	1,000

Bảng phân vị phân bố Student với k bậc tự do.

Với k và p cho trước, t_p được xác định từ hệ thức sau

$$P(|t| > t_p) = p.$$

p / k	0,11	0,10	0,09	0,08	0,07	0,06	0,05
1	5,730	6,314	7,026	7,916	9,058	10,579	12,706
2	2,760	2,920	3,104	3,320	3,578	3,896	4,303
3	2,249	2,353	2,471	2,605	2,763	2,951	3,182
4	2,048	2,132	2,226	2,333	2,456	2,601	2,776
5	1,941	2,015	2,098	2,191	2,297	2,422	2,571
6	1,874	1,943	2,019	2,104	2,201	2,313	2,447
7	1,830	1,895	1,966	2,046	2,136	2,241	2,365
8	1,797	1,860	1,928	2,004	2,090	2,189	2,306
9	1,773	1,833	1,899	1,973	2,055	2,150	2,262
10	1,754	1,812	1,877	1,948	2,028	2,120	2,228
11	1,738	1,796	1,859	1,928	2,007	2,096	2,201
12	1,726	1,782	1,844	1,912	1,989	2,076	2,179
13	1,715	1,771	1,832	1,899	1,974	2,060	2,160
14	1,706	1,761	1,821	1,887	1,962	2,046	2,145
15	1,699	1,753	1,812	1,878	1,951	2,034	2,131
16	1,692	1,746	1,805	1,869	1,942	2,024	2,120
17	1,686	1,740	1,798	1,862	1,934	2,015	2,110
18	1,681	1,734	1,792	1,855	1,926	2,007	2,101
19	1,677	1,729	1,786	1,850	1,920	2,000	2,093
20	1,672	1,725	1,782	1,844	1,914	1,994	2,086
21	1,669	1,721	1,777	1,840	1,909	1,988	2,080
22	1,665	1,717	1,773	1,835	1,905	1,983	2,074
23	1,662	1,714	1,770	1,832	1,900	1,978	2,069
24	1,660	1,711	1,767	1,828	1,896	1,974	2,064
25	1,657	1,708	1,764	1,825	1,893	1,970	2,060
26	1,655	1,706	1,761	1,822	1,890	1,967	2,056
27	1,653	1,703	1,758	1,819	1,887	1,963	2,052
28	1,651	1,701	1,756	1,817	1,884	1,960	2,048
29	1,649	1,699	1,754	1,814	1,881	1,957	2,045
30	1,647	1,697	1,752	1,812	1,879	1,955	2,042
31	1,645	1,696	1,750	1,810	1,877	1,952	2,040
32	1,644	1,694	1,748	1,808	1,875	1,950	2,037
33	1,642	1,692	1,747	1,806	1,873	1,948	2,035
34	1,641	1,691	1,745	1,805	1,871	1,946	2,032

Bảng phân vị phân bố Student với k bậc tự do.

Với k và p cho trước, t_p được xác định từ hệ thức sau

$$P(|t| > t_p) = p.$$

p / k	0,040	0,035	0,030	0,025	0,020	0,015	0,010
1	15,894	18,171	21,205	25,452	31,821	42,433	63,656
2	4,849	5,204	5,643	6,205	6,965	8,073	9,925
3	3,482	3,670	3,896	4,177	4,541	5,047	5,841
4	2,999	3,135	3,298	3,495	3,747	4,088	4,604
5	2,757	2,870	3,003	3,163	3,365	3,634	4,032
6	2,612	2,712	2,829	2,969	3,143	3,372	3,707
7	2,517	2,608	2,715	2,841	2,998	3,203	3,499
8	2,449	2,535	2,634	2,752	2,896	3,085	3,355
9	2,398	2,480	2,574	2,685	2,821	2,998	3,250
10	2,359	2,437	2,527	2,634	2,764	2,932	3,169
11	2,328	2,404	2,491	2,593	2,718	2,879	3,106
12	2,303	2,376	2,461	2,560	2,681	2,836	3,055
13	2,282	2,353	2,436	2,533	2,650	2,801	3,012
14	2,264	2,334	2,415	2,510	2,624	2,771	2,977
15	2,249	2,318	2,397	2,490	2,602	2,746	2,947
16	2,235	2,304	2,382	2,473	2,583	2,724	2,921
17	2,224	2,291	2,368	2,458	2,567	2,706	2,898
18	2,214	2,280	2,356	2,445	2,552	2,689	2,878
19	2,205	2,271	2,346	2,433	2,539	2,674	2,861
20	2,197	2,262	2,336	2,423	2,528	2,661	2,845
21	2,189	2,254	2,328	2,414	2,518	2,649	2,831
22	2,183	2,247	2,320	2,405	2,508	2,639	2,819
23	2,177	2,241	2,313	2,398	2,500	2,629	2,807
24	2,172	2,235	2,307	2,391	2,492	2,620	2,797
25	2,167	2,229	2,301	2,385	2,485	2,612	2,787
26	2,162	2,225	2,296	2,379	2,479	2,605	2,779
27	2,158	2,220	2,291	2,373	2,473	2,598	2,771
28	2,154	2,216	2,286	2,368	2,467	2,592	2,763
29	2,150	2,212	2,282	2,364	2,462	2,586	2,756
30	2,147	2,208	2,278	2,360	2,457	2,581	2,750
31	2,144	2,205	2,275	2,356	2,453	2,576	2,744
32	2,141	2,202	2,271	2,352	2,449	2,571	2,738
33	2,138	2,199	2,268	2,348	2,445	2,566	2,733
34	2,136	2,196	2,265	2,345	2,441	2,562	2,728

Bảng phân vị phân bố χ^2 với k bậc tự do.

Với k và p cho trước, χ^2_p được xác định từ hệ thức sau

$$P(\chi^2 > \chi^2_p) = p.$$

k \ p	0,030	0,025	0,020	0,015	0,010	0,005
1	4,709	5,024	5,412	5,916	6,635	7,879
2	7,013	7,378	7,824	8,399	9,210	10,597
3	8,947	9,348	9,837	10,465	11,345	12,838
4	10,712	11,143	11,668	12,339	13,277	14,860
5	12,375	12,832	13,388	14,098	15,086	16,750
6	13,968	14,449	15,033	15,777	16,812	18,548
7	15,509	16,013	16,622	17,398	18,475	20,278
8	17,011	17,535	18,168	18,974	20,090	21,955
9	18,480	19,023	19,679	20,512	21,666	23,589
10	19,922	20,483	21,161	22,021	23,209	25,188
11	21,342	21,920	22,618	23,503	24,725	26,757
12	22,742	23,337	24,054	24,963	26,217	28,300
13	24,125	24,736	25,471	26,403	27,688	29,819
14	25,493	26,119	26,873	27,827	29,141	31,319
15	26,848	27,488	28,259	29,235	30,578	32,801
16	28,191	28,845	29,633	30,629	32,000	34,267
17	29,523	30,191	30,995	32,011	33,409	35,718
18	30,845	31,526	32,346	33,382	34,805	37,156
19	32,158	32,852	33,687	34,742	36,191	38,582
20	33,462	34,170	35,020	36,093	37,566	39,997
21	34,759	35,479	36,343	37,434	38,932	41,401
22	36,049	36,781	37,659	38,768	40,289	42,796
23	37,332	38,076	38,968	40,094	41,638	44,181
24	38,609	39,364	40,270	41,413	42,980	45,558
25	39,880	40,646	41,566	42,725	44,314	46,928
26	41,146	41,923	42,856	44,031	45,642	48,290
27	42,407	43,195	44,140	45,331	46,963	49,645
28	43,662	44,461	45,419	46,626	48,278	50,994
29	44,913	45,722	46,693	47,915	49,588	52,335
30	46,160	46,979	47,962	49,199	50,892	53,672
31	47,402	48,232	49,226	50,478	52,191	55,002
32	48,641	49,480	50,487	51,753	53,486	56,328
33	49,876	50,725	51,743	53,024	54,775	57,648
34	51,107	51,966	52,995	54,290	56,061	58,964

Bảng phân vị phân bố χ^2 với k bậc tự do.

Với k và p cho trước, χ^2_p được xác định từ hệ thức sau

$$P(\chi^2 > \chi^2_p) = p.$$

k \ p	0,990	0,985	0,980	0,975	0,970	0,965
1	0,000	0,000	0,001	0,001	0,001	0,002
2	0,020	0,030	0,040	0,051	0,061	0,071
3	0,115	0,152	0,185	0,216	0,245	0,273
4	0,297	0,368	0,429	0,484	0,535	0,582
5	0,554	0,662	0,752	0,831	0,903	0,969
6	0,872	1,016	1,134	1,237	1,330	1,414
7	1,239	1,418	1,564	1,690	1,802	1,903
8	1,647	1,860	2,032	2,180	2,310	2,428
9	2,088	2,335	2,532	2,700	2,848	2,982
10	2,558	2,837	3,059	3,247	3,412	3,561
11	3,053	3,363	3,609	3,816	3,997	4,160
12	3,571	3,910	4,178	4,404	4,601	4,778
13	4,107	4,476	4,765	5,009	5,221	5,411
14	4,660	5,057	5,368	5,629	5,856	6,058
15	5,229	5,653	5,985	6,262	6,503	6,718
16	5,812	6,263	6,614	6,908	7,163	7,390
17	6,408	6,884	7,255	7,564	7,832	8,071
18	7,015	7,516	7,906	8,231	8,512	8,762
19	7,633	8,159	8,567	8,907	9,200	9,462
20	8,260	8,811	9,237	9,591	9,897	10,169
21	8,897	9,471	9,915	10,283	10,601	10,884
22	9,542	10,139	10,600	10,982	11,313	11,605
23	10,196	10,815	11,293	11,689	12,030	12,333
24	10,856	11,497	11,992	12,401	12,754	13,067
25	11,524	12,187	12,697	13,120	13,484	13,807
26	12,198	12,882	13,409	13,844	14,219	14,551
27	12,878	13,583	14,125	14,573	14,959	15,301
28	13,565	14,290	14,847	15,308	15,704	16,055
29	14,256	15,002	15,574	16,047	16,454	16,813
30	14,953	15,719	16,306	16,791	17,208	17,576
31	15,655	16,440	17,042	17,539	17,966	18,343
32	16,362	17,166	17,783	18,291	18,727	19,113
33	17,073	17,897	18,527	19,047	19,493	19,887
34	17,789	18,631	19,275	19,806	20,262	20,665

Bảng phân vị phân bố F với m, n bậc tự do.

Với m, n và p cho trước, F_p được xác định từ hệ thức sau

$$P(F > F_p) = p.$$

m	n	0,035	0,030	0,025	0,020	0,015	0,010
2	2	27,571	32,333	39,000	49,000	65,666	99,000
2	3	12,519	14,036	16,044	18,858	23,162	30,816
2	4	8,690	9,547	10,649	12,142	14,330	18,000
2	5	7,057	7,665	8,434	9,454	10,912	13,274
2	6	6,171	6,655	7,260	8,052	9,164	10,925
2	7	5,621	6,032	6,542	7,203	8,119	9,547
3	2	27,736	32,499	39,166	49,165	65,833	99,164
3	3	12,093	13,534	15,439	18,110	22,194	29,457
3	4	8,189	8,972	9,979	11,343	13,342	16,694
3	5	6,540	7,080	7,764	8,670	9,964	12,060
3	6	5,652	6,073	6,599	7,287	8,253	9,780
3	7	5,103	5,454	5,890	6,454	7,236	8,451
4	2	27,819	32,582	39,248	49,249	65,917	99,251
4	3	11,852	13,251	15,101	17,694	21,659	28,710
4	4	7,905	8,648	9,604	10,899	12,796	15,977
4	5	6,247	6,751	7,388	8,233	9,439	11,392
4	6	5,357	5,744	6,227	6,859	7,746	9,148
4	7	4,808	5,127	5,523	6,035	6,744	7,847
5	2	27,869	32,631	39,298	49,298	65,966	99,302
5	3	11,697	13,069	14,885	17,429	21,319	28,237
5	4	7,722	8,440	9,364	10,616	12,448	15,522
5	5	6,057	6,538	7,146	7,953	9,104	10,967
5	6	5,166	5,531	5,988	6,585	7,422	8,746
5	7	4,616	4,915	5,285	5,765	6,429	7,460
6	2	27,902	32,665	39,331	49,332	66,000	99,331
6	3	11,589	12,943	14,735	17,245	21,084	27,911
6	4	7,594	8,295	9,197	10,419	12,207	15,207
6	5	5,924	6,390	6,978	7,758	8,871	10,672
6	6	5,031	5,382	5,820	6,393	7,196	8,466
6	7	4,481	4,766	5,119	5,576	6,208	7,191
7	2	27,926	32,688	39,356	49,356	66,022	99,357
7	3	11,509	12,850	14,624	17,110	20,912	27,671
7	4	7,500	8,188	9,074	10,274	12,030	14,976
7	5	5,826	6,280	6,853	7,614	8,699	10,456
7	6	4,931	5,271	5,695	6,251	7,029	8,260

Hàm phân bố Poisson với λ là tham số.

Bảng sau cho giá trị hàm phân bố

$$\sum_{i=0}^{k} P(X=i) = \sum_{i=0}^{k} e^{-\lambda}\frac{\lambda^i}{i!}.$$

k / λ	0	1	2	3	4	5	6
0,2	0,819	0,982	0,999	1,000	1,000	1,000	1,000
0,3	0,741	0,963	0,996	1,000	1,000	1,000	1,000
0,4	0,670	0,938	0,992	0,999	1,000	1,000	1,000
0,5	0,607	0,910	0,986	0,998	1,000	1,000	1,000
0,6	0,549	0,878	0,977	0,997	1,000	1,000	1,000
0,7	0,497	0,844	0,966	0,994	0,999	1,000	1,000
0,8	0,449	0,809	0,953	0,991	0,999	1,000	1,000
0,9	0,407	0,772	0,937	0,987	0,998	1,000	1,000
1	0,368	0,736	0,920	0,981	0,996	0,999	1,000
1,1	0,333	0,699	0,900	0,974	0,995	0,999	1,000
1,2	0,301	0,663	0,879	0,966	0,992	0,998	1,000
1,3	0,273	0,627	0,857	0,957	0,989	0,998	1,000
1,4	0,247	0,592	0,833	0,946	0,986	0,997	0,999
1,5	0,223	0,558	0,809	0,934	0,981	0,996	0,999
1,6	0,202	0,525	0,783	0,921	0,976	0,994	0,999
1,7	0,183	0,493	0,757	0,907	0,970	0,992	0,998
1,8	0,165	0,463	0,731	0,891	0,964	0,990	0,997
1,9	0,150	0,434	0,704	0,875	0,956	0,987	0,997
2	0,135	0,406	0,677	0,857	0,947	0,983	0,995
2,1	0,122	0,380	0,650	0,839	0,938	0,980	0,994
2,2	0,111	0,355	0,623	0,819	0,928	0,975	0,993
2,3	0,100	0,331	0,596	0,799	0,916	0,970	0,991
2,4	0,091	0,308	0,570	0,779	0,904	0,964	0,988
2,5	0,082	0,287	0,544	0,758	0,891	0,958	0,986
2,6	0,074	0,267	0,518	0,736	0,877	0,951	0,983
2,7	0,067	0,249	0,494	0,714	0,863	0,943	0,979
2,8	0,061	0,231	0,469	0,692	0,848	0,935	0,976
2,9	0,055	0,215	0,446	0,670	0,832	0,926	0,971
3	0,050	0,199	0,423	0,647	0,815	0,916	0,966
3,1	0,045	0,185	0,401	0,625	0,798	0,906	0,961
3,2	0,041	0,171	0,380	0,603	0,781	0,895	0,955
3,3	0,037	0,159	0,359	0,580	0,763	0,883	0,949
3,4	0,033	0,147	0,340	0,558	0,744	0,871	0,942
3,5	0,030	0,136	0,321	0,537	0,725	0,858	0,935

Hàm phân bố Poisson với λ là tham số.

Bảng sau cho giá trị hàm phân bố

$$\sum_{i=0}^{k} P(X=i) = \sum_{i=0}^{k} e^{-\lambda}\frac{\lambda^i}{i!}.$$

k / λ	6	7	8	9	10	11	12
2,6	0,983	0,995	0,999	1,000	1,000	1,000	1,000
2,7	0,979	0,993	0,998	0,999	1,000	1,000	1,000
2,8	0,976	0,992	0,998	0,999	1,000	1,000	1,000
2,9	0,971	0,990	0,997	0,999	1,000	1,000	1,000
3	0,966	0,988	0,996	0,999	1,000	1,000	1,000
3,1	0,961	0,986	0,995	0,999	1,000	1,000	1,000
3,2	0,955	0,983	0,994	0,998	1,000	1,000	1,000
3,3	0,949	0,980	0,993	0,998	0,999	1,000	1,000
3,4	0,942	0,977	0,992	0,997	0,999	1,000	1,000
3,5	0,935	0,973	0,990	0,997	0,999	1,000	1,000
3,6	0,927	0,969	0,988	0,996	0,999	1,000	1,000
3,7	0,918	0,965	0,986	0,995	0,998	1,000	1,000
3,8	0,909	0,960	0,984	0,994	0,998	0,999	1,000
3,9	0,899	0,955	0,981	0,993	0,998	0,999	1,000
4	0,889	0,949	0,979	0,992	0,997	0,999	1,000
4,1	0,879	0,943	0,976	0,990	0,997	0,999	1,000
4,2	0,867	0,936	0,972	0,989	0,996	0,999	1,000
4,3	0,856	0,929	0,968	0,987	0,995	0,998	0,999
4,4	0,844	0,921	0,964	0,985	0,994	0,998	0,999
4,5	0,831	0,913	0,960	0,983	0,993	0,998	0,999
4,6	0,818	0,905	0,955	0,980	0,992	0,997	0,999
4,7	0,805	0,896	0,950	0,978	0,991	0,997	0,999
4,8	0,791	0,887	0,944	0,975	0,990	0,996	0,999
4,9	0,777	0,877	0,938	0,972	0,988	0,995	0,998
5	0,762	0,867	0,932	0,968	0,986	0,995	0,998
5,1	0,747	0,856	0,925	0,964	0,984	0,994	0,998
5,2	0,732	0,845	0,918	0,960	0,982	0,993	0,997
5,3	0,717	0,833	0,911	0,956	0,980	0,992	0,997
5,4	0,702	0,822	0,903	0,951	0,977	0,990	0,996
5,5	0,686	0,809	0,894	0,946	0,975	0,989	0,996
5,6	0,670	0,797	0,886	0,941	0,972	0,988	0,995
5,7	0,654	0,784	0,877	0,935	0,969	0,986	0,994
5,8	0,638	0,771	0,867	0,929	0,965	0,984	0,993
5,9	0,622	0,758	0,857	0,923	0,961	0,982	0,992

CHỈ DẪN

TÀI LIỆU THAM KHẢO

1. CRAMER, H.: Mathematical methods of statistics. Princeton, 1951.

2. FELLER, W.: An introduction to probability theory and its applications. New York, Wiley 1950.

3. GNEDENKO, B.: The Theory of Probability. Moscow, Mir 1973.

4. ĐẶNG HÙNG THẮNG.: Mở đầu về lý thuyết xác suất và các ứng dụng. Nhà xuất bản Giáo dục, 1997.

5. TRẦN TUẤN ĐIỆP - LÝ HOÀNG TÚ: Lý thuyết xác suất và Thống kê toán học. Nhà xuất bản Giáo dục, 1999.

6. C.RADHAKRISHNA RAO: Linear statistical inference and its applications. Moscow, Nayka 1968.