Chủởng 1

NHỮNG KHÁI NIỆM CƠ BẢN VỀ XÁC SUẤT

1. BỒ TÚC VỀ GIẢI TÍCH TỒ HỢP

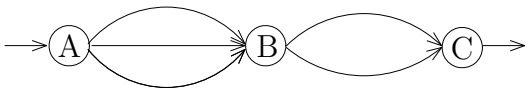
1.1 Qui tắc nhân

Giả sủ một công việc nào đó đuợc chia thành k giai đoạn. Có n_1 cách thục hiện giai đoạn thuí nhất, n_2 cách thục hiện giai đoạn thuí hai,...,n_k cách thục hiện giai đoạn thuí k. Khi đó ta có

$$n = n_1.n_2 \dots n_k$$

cách thục hiện công việc.

• **Ví dụ 1** *Giả sử để đi từ A đến C ta bắt buộc phải đi qua điểm B. Có 3 đường khác nhau để đi từ A đến B và có 2 đường khác nhau để đi từ B đến C. Vậy có* $n = 3.2$ *cách khác nhau để đi từ A đến C.*



1.2 Chỉnh hợp

□ **Định nghĩa 1** *Chỉnh hợp chập k của n phần tử* $(k \leq n)$ *là một nhóm (bộ) có thứ tự gồm k phần tử khác nhau chọn từ n phần tử đã cho.*

Số chỉnh hợp chập k của n phần tử kí hiệu là A_n^k.

⊙ **Công thức tính:** $\boxed{A_n^k = \dfrac{n!}{(n-k)!} = n(n-1)\dots(n-k+1)}$

• **Ví dụ 2** *Một buổi họp gồm 12 người tham dự. Hỏi có mấy cách chọn một chủ tọa và một thư ký?*

Giải

Mỗi cách chọn một chủ tọa và một thuu ký tù 12 người tham dụ buổi họp là một chỉnh hợp chập k của 12 phần tủ.

Do đó số cách chọn là $A_{12}^2 = 12.11 = 132$.

• **Ví dụ 3** *Với các chữ số 0,1,2,3,4,5 có thể lập được bao nhiêu số khác nhau gồm 4 chữ số.*

Giải

Các số bắt đầu bằng chữ số 0 (0123, 0234,...) không phải là số gồm 4 chữ số.

Chữ số đầu tiên phải chọn trong các chữ số 1,2,3,4,5. Do đó có 5 cách chọn chữ số đầu tiên.

Ba chữ số kế tiếp có thể chọn tùy ý trong 5 chữ số còn lại. Có A_5^3 cách chọn.

Vậy số cách chọn là $5.A_5^3 = 5.(5.4.3) = 300$

1.3 Chỉnh hợp lặp

□ **Định nghĩa 2** *Chỉnh hợp lặp chập k của n phần tử là một nhóm có thứ tự gồm k phần tử chọn từ n phần tử đã cho, trong đó mỗi phần tử có thể có mặt 1,2,...,k lần trong nhóm.*

Số chỉnh hợp lặp chặp k của n phần tử được kí hiệu B_n^k.

⊙ **Công thức tính**

$$\boxed{B_n^k = n^k}$$

• **Ví dụ 4** *Xếp 5 cuốn sách vào 3 ngăn. Hỏi có bao nhiêu cách xếp ?*

Giải

Mỗi cách xếp 5 cuốn sách vào 3 ngăn là một chỉnh hợp lặp chập 5 của 3 (Mỗi lần xếp 1 cuốn sách vào 1 ngăn xem như chọn 1 ngăn trong 3 ngăn. Do có 5 cuốn sách nên việc chọn ngăn được tiến hành 5 lần).

Vậy số cách xếp là $B_3^5 = 3^5 = 243$.

1.4 Hoán vị

□ **Định nghĩa 3** *Hoán vị của m phần tử là một nhóm có thứ tự gồm đủ mặt m phần tử đã cho.*

Số hoán vị của m phần tử được kí hiệu là P_m.

⊙ **Công thức tính**

$$\boxed{P_m = m!}$$

• **Ví dụ 5** *Một bàn có 4 học sinh. Hỏi có mấy cách xếp chỗ ngồi ?*

Giải

Mỗi cách xếp chỗ của 4 học sinh ở một bàn là một hoán vị của 4 phần tử. Do đó số cách xếp là $P_4 = 4! = 24$.

1.5 Tổ hợp

□ **Định nghĩa 4** *Tổ hợp chập k của n phần tử ($k \leq n$) là một nhóm không phân biệt thứ tự, gồm k phần tử khác nhau chọn từ n phần tử đã cho.*

Số tổ hợp chập k của n phần tử kí hiệu là C_n^k.

⊙ **Công thức tính**

$$\boxed{C_n^k = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\dots(n-k+1)}{k!}}$$

⊙ **Chú ý**

i) Qui ước $0! = 1$.
ii) $C_n^k = C_n^{n-k}$.
iii) $C_n^k = C_{n-1}^{k-1} + C_{n-1}^k$.

• **Ví dụ 6** *Mỗi đề thi gồm 3 câu hỏi lấy trong 25 câu hỏi cho trước. Hỏi có thể lập nên bao nhiêu đề thi khác nhau ?*

Giải

Số đề thi có thể lập nên là $C_{25}^3 = \dfrac{25!}{3!.(22)!} = \dfrac{25.24.23}{1.2.3} = 2.300.$

• **Ví dụ 7** *Một máy tính có 16 cổng. Giả sử tại mỗi thời điểm bất kỳ mỗi cổng hoặc trong sử dụng hoặc không trong sử dụng nhưng có thể hoạt động hoặc không thể hoạt động. Hỏi có bao nhiêu cấu hình (cách chọn) trong đó 10 cổng trong sử dụng, 4 không trong sử dụng nhưng có thể hoạt động và 2 không hoạt động?*

Giải

Để xác định số cách chọn ta qua 3 bước:

Bước 1: Chọn 10 cổng sử dụng: có $C_{16}^{10} = 8008$ cách.

Bước 2: Chọn 4 cổng không trong sử dụng nhưng có thể hoạt động trong 6 cổng còn lại: có $C_6^4 = 15$ cách.

Bước 3: Chọn 2 cổng không thể hoạt động: có $C_2^2 = 1$ cách.

Theo qui tắc nhân, ta có $C_{16}^{10}.C_6^4.C_2^2 = (8008).(15).(1) = 120.120$ cách.

1.6 Nhị thức Newton

Ở phổ thông ta đã biết các hằng đẳng thức đáng nhớ

$$\begin{aligned} a+b &= a^1 + b^1 \\ (a+b)^2 &= a^2 + 2a^1b^1 + b^2 \\ (a+b)^3 &= a^3 + 3a^2b^1 + 3a^1b^2 + b^3 \end{aligned}$$

Các hệ số trong các hằng đẳng thức trên có thể xác định từ tam giác Pascal

$$\begin{array}{ccccccccc} 1 & & 1 & & & & & & \\ 1 & & 2 & & 1 & & & & \\ 1 & & 3 & & 3 & & 1 & & \\ 1 & & 4 & & 6 & & 4 & & 1 \\ \\ C_n^0 & & C_n^1 & & C_n^2 & & C_n^3 & \ldots & C_n^{n-1} & & C_n^n \end{array}$$

Newton đã chứng minh được công thức tổng quát sau (*Nhị thức Newton*):

$$\begin{aligned} (a+b)^n &= \sum_{k=o}^{n} C_n^k a^{n-k} b^k \\ &= C_n^0 a^n + C_n^1 a^{n-1} b + C_n^2 a^{n-2} b^2 + \ldots + C_n^k a^{n-k} b^k + \ldots + C_n^{n-1} a b^{n-1} + C_n^n b^n \end{aligned}$$

(a,b là các số thực; n là số tự nhiên)

2. BIẾN CỐ VÀ QUAN HỆ GIỮA CÁC BIẾN CỐ

2.1 Phép thử và biến cố

Việc thực hiện một nhóm các điều kiện cơ bản để quan sát một hiện tượng nào đó được gọi một phép thử. Các kết quả có thể xảy ra của phép thử được gọi là biến cố (sự kiện).

• Ví dụ 8

i) Tung đồng tiền lên là một phép thử. Đồng tiền lật mặt nào đó (xấp, ngửa) là một biến cố.

ii) Bắn một phát súng vào một cái bia là một phép thử. Việc viên đạn trúng (trật) bia là một biến cố.

2.2 Các biến cố và quan hệ giữa các biến cố

i) Quan hệ kéo theo

Biến cố A được gọi là kéo theo biến cố B, kí hiệu $A \subset B$, nếu A xảy ra thì B xảy ra.

ii) Quan hệ tương đương

Hai biến cố A và B được gọi là tương đương với nhau nếu $A \subset B$ và $B \subset A$, kí hiệu $A = B$.

iii) Biến cố sơ cấp

Biến cố sơ cấp là biến cố không thể phân tích được nữa được nua.

iv) Biến cố chắc chắn

Là biến cố nhất định sẽ xảy ra khi thực hiện phép thử. Kí hiệu Ω.

• **Ví dụ 9** *Tung một con xúc xắc. Biến cố mặt con xúc xắc có số chấm bé hơn 7 là biến cố chắc chắn.*

v) Biến cố không thể

Là biến cố nhất định không xảy ra khi thực hiện phép thử. Kí hiệu $\emptyset$.

$\oplus$ **Nhận xét** Biến cố không thể $\emptyset$ không bao hàm một biến cố sơ cấp nào, nghĩa là không có biến cố sơ cấp nào thuận lợi cho biên cố không thể.

vi) Biến cố ngẫu nhiên

Là biến cố có thể xảy ra hoặc không xảy ra khi thực hiện phép thử. Phép thử mà các kết quả của nó là các biến cố ngẫu nhiên được gọi là phép thử ngẫu nhiên.

vii) Biến cố tổng

Biến cố C được gọi là tổng của hai biến cố A và B, kí hiệu $C = A + B$, nếu C xảy ra khi và chỉ khi ít nhất một trong hai biến cố A và B xảy ra.

• **Ví dụ 10** *Hai người thợ săn cùng bắn vào một con thú. Nếu gọi A là biến cố người thứ nhất bắn trúng con thú và B là biến cố người thứ hai bắn trúng con thú thì $C = A+B$ là biến cố con thú bị bắn trúng.*

$\odot$ **Chú ý**

i) Mọi biến cố ngẫu nhiên A đều biểu diễn được dưới dạng tổng của một số biến cố sơ cấp nào đó. Các biến cố sơ cấp trong tổng này được gọi là *các biến cố thuận lợi* cho biến cố A.

ii) Biến cố chắc chắn Ω là tổng của mọi biến cố sơ cấp có thể, nghĩa là mọi biến cố sơ cấp đều thuận lợi cho Ω. Do đó Ω còn được gọi là *không gian các biến cố sơ cấp.*

• **Ví dụ 11** *Tung một con xúc xắc. Ta có 6 biến cố sơ cấp $A_1, A_2, A_3, A_4, A_5, A_6$, trong đó A_j là biến cố xuất hiện mặt j chấm $j = 1, 2, \ldots, 6$.*

Gọi A là biến cố xuất hiện mặt với số chấm chẵn thì A có 3 biến cố thuận lợi là A_2, A_4, A_6.

Ta có $A = A_2 + A_4 + A_6$

Gọi B là biến cố xuất hiện mặt với số chấm chia hết cho 3 thì B có 2 biến cố thuận lợi là A_3, A_6.

Ta có $B = A_3 + A_6$

viii) Biến cố tích

Biến cố C được gọi là tích của hai biến cố A và B, kí hiệu AB, nếu C xảy ra khi và chỉ khi cả A lẫn B cùng xảy ra.

• **Ví dụ 12** *Hai người cùng bắn vào một con thú.*

Gọi A là biến cố người thứ nhất bắn trượt, B là biến cố người thứ hai bắn trượt thì $C = AB$ là biến cố con thú không bị bắn trúng.

ix) Biến cố hiệu

Hiệu của biến cố A và biến cố B, kí hiệu A \ B là biến cố xảy ra khi và chỉ khi A xảy ra nhưng B không xảy ra.

x) Biến cố xung khắc

Hai biến cố A và B được gọi là hai biến cố xung khắc nếu chúng không đồng thời xảy ra trong một phép thử.

• **Ví dụ 13** *Tung một đồng tiền.*

Gọi A là biến cố xuất hiện mặt xấp, B là biến cố xuất hiện mặt ngửa thì $AB = \emptyset$.

xi) Biến cố đối lập

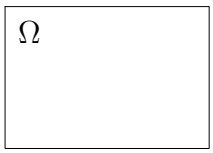
Biến cố *không xảy ra biến cố A* được gọi là biến cố đối lập với biến cố A. Kí hiệu $\overline{A}$. Ta có

$$A + \overline{A} = \Omega, \qquad A\overline{A} = \emptyset$$

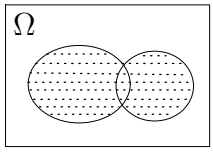
⊕ **Nhận xét**

Qua các khái niệm trên ta thấy các biến cố tổng, tích, hiệu, đối lập tương ứng với tập hợp, giao, hiệu, phần bù của lý thuyết tập hợp. Do đó ta có thể sử dụng các phép toán trên các tập hợp cho các phép toán trên các biến cố.

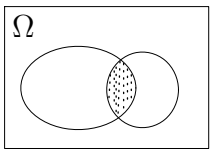
Ta có thể dùng biểu đồ Venn để miêu tả các biến cố.



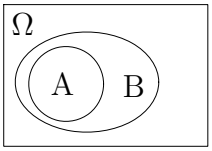
Bc chắc chắn



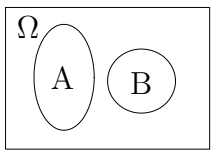
A+B



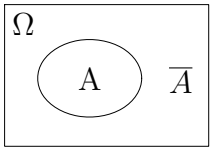
AB



A⟹B



A,B xung khắc



Đối lập $\overline{A}$

3. XÁC SUẤT

3.1 Định nghĩa xác suất theo lối cổ điển

□ **Định nghĩa 5** *Giả sử phép thử có n biến cố đồng khả năng có thể xảy ra, trong đó có m biến cố đồng khả năng thuận lợi cho biến cố A (A là tổng của m biến cố sơ cấp này). Khi đó xác suất của biến cố A, kí hiệu P(A) được định nghĩa bằng công thức sau:*

$$\boxed{P(A) = \frac{m}{n} = \frac{\textit{Số trường hợp thuận lợi cho } A}{\textit{Số trường hợp có thể xảy ra}}}$$

• **Ví dụ 14** *Gieo một con xúc xắc cân đối, đồng chất. Tính xác suất xuất hiện mặt chẵn.*

Giải

Gọi A_i là biến cố xuất hiện mặt i chấm và A là biến cố xuất hiện mặt chẵn thì

$$A = A_2 + A_4 + A_6$$

Ta thấy phép thử có 6 biến cố sơ cấp đồng khả năng có thể xảy ra trong đó có 3 biến cố thuận lợi cho A.

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

• **Ví dụ 15** *Một người gọi điện thoại nhưng lại quên 2 số cuối của số điện thoại cần gọi mà chỉ nhớ là 2 số đó khác nhau. Tìm xác suất để người đó quay ngẫu nhiên một lần trúng số cần gọi.*

Giải

Gọi A là biến cố người đó quay ngẫu nhiên một lần trúng số cần gọi.

Số biến cố sơ cấp đồng khả năng có thể xảy ra (số cách gọi 2 số cuối) là $n = A_{10}^2 = 90$.

Số biến cố thuận lợi cho A là $m = 1$.

Vậy $P(A) = \frac{1}{90}$.

• **Ví dụ 16** *Trong hộp có 6 bi trắng, 4 bi đen. Tìm xác suất để lấy từ hộp ra được*
i) 1 viên bi đen.
ii) 2 viên bi trắng.

Giải

Gọi A là biến cố lấy tủ hộp ra được 1 viên bi đen và B là biến cố lấy tủ hộp ra 2 viên bi trắng.

Ta có

i) $P(A) = \dfrac{C_4^1}{C_{10}^1} = \dfrac{2}{5}$

ii) $P(B) = \dfrac{C_6^2}{C_{10}^2} = \dfrac{1}{3}$

• **Ví dụ 17** *Rút ngẫu nhiên từ một cỗ bài tú lơ khơ 52 lá ra 5 lá. Tìm xác suất sao cho trong 5 lá rút ra có*

a) 3 lá đỏ và 2 lá đen.
b) 2 con cơ, 1 con rô, 2 con chuồn.

Giải

Gọi A là biến cố rút ra được 3 lá đỏ và 2 lá đen.
B là biến cố rút ra được 2 con cơ, 1 con rô, 2 con chuồn.

Số biến cố có thể xảy ra khi rút 5 lá bài là C_{52}^5.

a) Số biến cố thuận lợi cho A là $C_{26}^3.C_{26}^2$.

$$P(A) = \frac{C_{26}^3.C_{26}^2}{C_{52}^5} = \frac{845000}{2598960} = 0,3251$$

b) Số biến cố thuận lợi cho B là $C_{13}^2.C_{13}^1.C_{13}^2$

$$P(B) = \frac{C_{13}^2.C_{13}^1.C_{13}^2}{C_{52}^5} = \frac{79092}{2598960} = 0,30432$$

• **Ví dụ 18** (`Bài toán ngày sinh`) *Một nhóm gồn n người. Tìm xác suất để có ít nhất hai người có cùng ngày sinh (cùng ngày và cùng tháng).*

Giải

Gọi S là tập hợp các danh sách ngày sinh có thể của n người và E là biến cố có ít nhất hai người trong nhóm có cùng ngày sinh trong năm.

Ta có $\overline{E}$ là biến cố không có hai người bất kỳ trong nhóm có cùng ngày sinh.

Số các trường hợp của S là

$$n(S) = \underbrace{365.365\ldots365}_{n} = 365^n$$

Số trường hợp thuận lợi cho $\overline{E}$ là

$$\begin{aligned} n(\overline{E}) &= 365.364.363.\ldots.[365-(n-1)] \\ &= \frac{[365.364.363.\ldots.(366-n)](365-n)!}{(365-n)!} \\ &= \tfrac{365!}{(365-n)!} \end{aligned}$$

Vì các biên cố đồng khả năng nên

$$P(\overline{E}) = \frac{n(\overline{E})}{n(S)} = \frac{\frac{365!}{(365-n)!}}{365^n} = \frac{365!}{365^n.(365-n)!}$$

Do đó xác suất để ít nhất có hai người có cùng ngày sinh là

$$P(E) = 1 - P(\overline{E}) = 1 - \frac{\frac{365!}{(365-n)!}}{365^n} = \frac{365!}{365^n.(365-n)!}$$

Số người trong nhóm n	*Xác suất có ít nhất 2 người có cùng ngày sinh* $P(E)$
5	0,027
10	0,117
15	0,253
20	0,411
23	0,507
30	0,706
40	0,891
50	0,970
60	0,994
70	0,999

Bảng bài toán ngày sinh

⊙ **Chú ý** Định nghĩa xác suất theo lối cổ điển có một số hạn chế:

i) Nó chỉ xét cho hệ hữu hạn các biến cố sơ cấp.

ii) Không phải lúc nào việc "đồng khả năng" cũng xảy ra.

3.2 Định nghĩa xác suất theo lối thống kê

□ **Định nghĩa 6** *Thực hiện phép thử n lần. Giả sử biến cố A xuất hiện m lần. Khi đó m được gọi là tần số của biến cố A và tỷ số* $\frac{m}{n}$ *được gọi là tần suất xuất hiện biến cố A trong loạt phép thử.*

Cho số phép thử tăng lên vô hạn, tần suất xuất hiện biến cố A dần về một số xác định gọi là xác suất của biến cố A.

$$\boxed{P(A) = \lim_{n\to\infty} \frac{m}{n}}$$

• **Ví dụ 19** *Một xạ thủ bắn 1000 viên đạn vào bia. Có xấp xỉ 50 viên trúng bia. Khi đó xác suất để xạ thủ bắn trúng bia là* $\frac{50}{1000} = 5\%$.

• **Ví dụ 20** *Để nghiên cứu khả năng xuất hiện mặt sấp khi tung một đồng tiền, người ta tiến hành tung đồng tiền nhiều lần và thu được kết quả cho ở bảng dưới đây:*

Người làm thí nghiệm	Số lần tung	Số lần được mặt sấp	Tần suất $f(A)$
Buyffon	4040	2.048	0,5069
Pearson	12.000	6.019	0,5016
Pearson	24.000	12.012	0,5005

3.3 Định nghĩa xác suất theo quan điểm hình học

□ **Định nghĩa 7** *Xét một phép thử có không gian các biến cố sơ cấp Ω được biểu diễn bởi miền hình học Ω có độ đo (độ dài, diện tích, thể tích) hữu hạn khác 0, biến cố A được biểu diễn bởi miền hình học A. Khi đó xác suất của biến cố A được xác định bởi:*

$$\boxed{P(A) = \frac{\text{Độ đo của miền } A}{\text{Độ đo của miền } \Omega}}$$

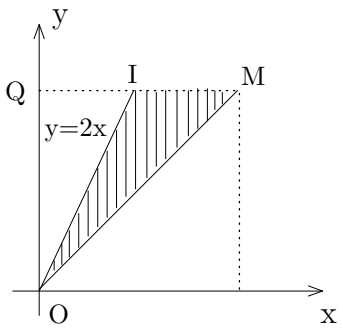
• **Ví dụ 21** *Trên đoạn thẳng OA ta gieo ngẫu nhiên hai điểm B và C có tọa độ tương ứng $OB = x$, $OC = y$ $(y \geq x)$. Tìm xác suất sao cho độ dài của đoạn BC bé hơn độ dài của đoạn OB.*

Giải

Giả sử $OA = l$. Các tọa độ x và y phải thỏa mãn các điều kiện:

$$0 \leq x \leq l, \quad 0 \leq y \leq l, \quad y \geq x \qquad (*)$$

Biểu diễn x và y lên hệ trục tọa độ vuông góc. Các điểm có tọa độ thỏa mãn (*) thuộc tam giác OMQ (có thể xem như biến cố chắc chắn).



Mặt khác, theo yêu cầu bài toán ta phải có $y - x < x$ hay $y < 2x$ (**). Những điểm có tọa độ thỏa mãn (*) và (**) thuộc miền có gạch. Miền thuận lợi cho biến cố cần tìm là tam giác OMI. Vậy xác suất cần tính

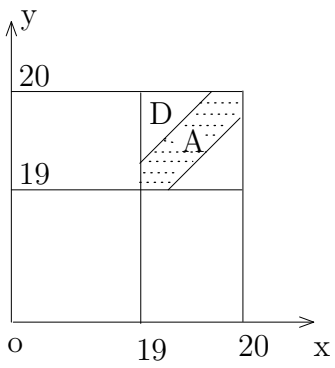
$$p = \frac{\text{diện tích } OMI}{\text{diện tích } OMQ} = \frac{1}{2}$$

• **Ví dụ 22** `(Bài toán hai người gặp nhau)`

Hai người hẹn gặp nhau ở một địa điểm xác định vào khoảng từ 19 giờ đến 20 giờ. Mỗi người đến (chắc chắn sẽ đến) điểm hẹn trong khoảng thời gian trên một cách độc lập với nhau, chờ trong 20 phút, nếu không thấy người kia đến sẽ bỏ đi. Tìm xác suất để hai người gặp nhau.

Giải

Gọi x, y là thời gian đến điểm hẹn của mỗi người và A là biến cố hai người gặp nhau. Rõ ràng x, y là một điểm ngẫu nhiên trong khoảng $[19, 20]$, ta có $19 \le x \le 20$;
$19 \le y \le 20$.



Để hai người gặp nhau thì

$|x - y| \le 20$ phút $= \frac{1}{3}$ giờ.

Do đó

$$\Omega = \{(x, y) : \ 19 \le x20, \ 19 \le y \le 20\}$$

$$A = \{(x, y) : \ |x - y| \le \frac{1}{3}\}$$

Diện tích của miền Ω bằng 1.

Diện tích của miền A bằng $1 - 2.\frac{1}{2}.\frac{2}{3}.\frac{2}{3} = \frac{5}{9}$

Vậy $P(A) = \dfrac{\text{diện tích A}}{\text{diện tích } \Omega} = \dfrac{5/9}{1} = 0,555.$

3.4 Định nghĩa xác suất theo tiên đề

Giả sử Ω là biến cố chắc chắn. Gọi $\mathcal{A}$ là họ các tập con của Ω thỏa các điều kiện sau:

i) $\mathcal{A}$ chứa Ω.

ii) Nếu $A, B \in \mathcal{A}$ thì $\overline{A}, A + B, AB$ thuộc $\mathcal{A}$.

Họ $\mathcal{A}$ thỏa các tiên đề i) và ii) thì $\mathcal{A}$ được gọi là đại số.

iii) Nếu $A_1, A_2, \ldots, A_n, \ldots$ là các phần tử của $\mathcal{A}$ thì tổng và tích vô hạn $A_1 + A_2 + \ldots + A_n$ và $A_1 A_2 \ldots A_n \ldots$ cũng thuộc $\mathcal{A}$.

Nếu $\mathcal{A}$ thỏa các điều kiện i), ii), iii) thì $\mathcal{A}$ được gọi là σ đại số.

□ **Định nghĩa 8** *Ta gọi xác suất trên $(\Omega, \mathcal{A})$ là một hàm P số xác định trên $\mathcal{A}$ có giá trị trong [0,1] và thỏa mãn 3 tiên đề sau:*

i) $P(\Omega) = 1$.

ii) $P(A + B) = P(A) + P(B)$ (với A, B xung khắc).

iii) Nếu dãy $\{A_n\}$ có tính chất $A_1 \supset A_2 \supset \ldots \supset A_n \supset \ldots$ và $A_1 A_2 \ldots A_n \ldots = \emptyset$ thì $\lim\limits_{n \to \infty} P(A_n) = 0$.

3.5 Các tính chất của xác suất

i) $0 \leq P(A) \leq 1$ với mọi biến cố A
ii) $P(\Omega) = 1$
iii) $P(\emptyset) = 0$
iv) Nếu $A \subset B$ thì $P(A) \leq P(B)$.
v) $P(A) + P(\overline{A}) = 1$.
vi) $P(A) = P(AB) + P(A\overline{B})$.

4. MỘT SỐ CÔNG THỨC TÍNH XÁC SUẤT

4.1 Công thức cộng xác suất

⊙ Công thức 1

Giả sử A và B là hai biến cố xung khắc ($AB = \emptyset$). Ta có

$$\boxed{P(A+B) = P(A) + P(B)}$$

Chứng minh

Giả sử phép thử có n biến cố đồng khả năng có thể xảy ra, trong đó có m_A biến cố thuận lợi cho biến cố A và m_B biến cố thuận lợi cho biến cố B. Khi đó số biến cố thuận lợi cho biến cố $A + B$ là $m = m_A + m_B$.

Do đó

$$P(A+B) = \frac{m_A + m_B}{n} = \frac{m_A}{n} + \frac{m_B}{n} = P(A) + P(B)$$

□ Định nghĩa 9

i) Các biến cố $A_1, A_2, \ldots, A_n$ được gọi là nhóm các biến cố đầy đủ xung khắc từng đôi nếu chúng xung khắc từng đôi và tổng của chúng là biến cố chắc chắn. Ta có

$$A_1 + A_2 + \ldots + A_n = \Omega, \quad A_i A_j = \emptyset$$

ii) Hai biến cố A và B được gọi là hai biến cố độc lập nếu sự tồn tại hay không tồn tại của biến cố này không ảnh hưởng đến sự tồn tại hay không tồn tại của biến cố kia.

iii) Các biến cố $A_1, A_2, \ldots, A_n$ được gọi độc lập toàn phần nếu mỗi biến cố độc lập với tích của một tổ hợp bất kỳ trong các biến cố còn lại.

△ Hệ quả 1

i) Nếu $A_1, A_2, \ldots, A_n$ là biến cố xung khắc từng đôi thì

$$P(A_1 + A_2 + \ldots + A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

ii) Nếu $A_1, A_2, \ldots, A_n$ là nhóm các biến cố đầy đủ xung khắc từng đôi thì

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

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

⊙ **Công thức 2**

$$\boxed{P(A+B) = P(A) + P(B) - P(AB)}$$

Chứng minh

Giả sử phép thử có n biến cố đồng khả năng có thể xảy ra, trong đó có m_A biến cố thuận lợi cho biến cố A, m_B biến cố thuận lợi cho biến cố B và k biến cố thuận lợi cho biến cố AB. Khi đó số biến cố thuận lợi cho biến cố $A+B$ là $m_A + m_B - k$.

Do đó

$$P(A+B) = \frac{m_A + m_B - k}{n} = \frac{m_A}{n} + \frac{m_B}{n} - \frac{k}{n} = P(A) + P(B) - P(AB).$$

△ **Hệ quả 2**

i) $P(A_1 + A_2 + \ldots, +A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_iA_j) + \sum_{i<j<k} P(A_iA_jA_k) + \ldots + (-1)^{n-1} P(A_1A_2 \ldots A_n)$.

ii) Nếu $A_1, A_2, \ldots, A_n$ là các biến cố độc lập toàn phần thì

$$P(A_1 + A_2 + \ldots + A_n) = 1 - P(\overline{A_1}).P(\overline{A_2}) \ldots P(\overline{A_n}).$$

• **Ví dụ 23** *Một lô hàng gồm 10 sản phẩm, trong đó có 2 phế phẩm. Lấy ngẫu nhiên không hoàn lại từ lô hàng ra 6 sản phẩm. Tìm xác suất để có không quá 1 phế phẩm trong 6 sản phẩm được lấy ra.*

Giải

Gọi

A là biến cố không có phế phẩm trong 6 sản phẩm lấy ra.
B là biến cố có đúng 1 phế phẩm.
C là biến cố có không quá một phế phẩm

thì A và B là hai biến cố xung khắc và $C = A + B$.

Ta có

$$P(A) = \frac{C_8^6}{C_{10}^6} = \frac{28}{210} = \frac{2}{15}$$

$$P(B) = \frac{C_2^1.C_8^5}{C_{10}^6} = \frac{112}{210} = \frac{8}{15}$$

Do đó

$$P(C) = P(A) + P(B) = \frac{2}{15} + \frac{8}{15} = \frac{2}{3}$$

• **Ví dụ 24** *Một lớp có 100 sinh viên, trong đó có 40 sinh viên giỏi ngoại ngữ, 30 sinh viên giỏi tin học, 20 sinh viên giỏi cả ngoại ngữ lẫn tin học. Sinh viên nào giỏi ít nhất một trong hai môn sẽ được thêm điểm trong kết quả học tập của học kỳ. Chọn ngẫu nhiên một sinh viên trong lớp. Tìm xác suất để sinh viên đó được tăng điểm.*

Giải

Gọi

A là biến cố gọi được sinh viên được tăng điểm.
N là biến cố gọi được sinh viên giỏi ngoại ngữ.
T là biến cố gọi được sinh viên giỏi tin học

thì $A = T + N$.

Ta có

$$P(A) = P(T) + P(N) - P(TN) = \frac{30}{100} + \frac{40}{100} - \frac{20}{100} = \frac{50}{100} = 0,5$$

4.2 Xác suất có điều kiện và công thức nhân xác suất

a) Xác suất có điều kiện

□ **Định nghĩa 10** *Xác suất của biến cố A với điều kiện biến cố B xảy ra được gọi là xác có điều kiện của biến cố A. Kí hiệu $P(A/B)$.*

• **Ví dụ 25** *Trong hộp có 5 viên bi trắng, 3 viên bi đen. Lấy lần lượt ra 2 viên bi (không hoàn lại). Tìm xác suất để lần thứ hai lấy được viên bi trắng biết lần thứ nhất đã lấy được viên bi trắng.*

Giải

Gọi A là biến cố lần thứ hai lấy được viên bi trắng
B là biến cố lần thứ nhất lấy được viên bi trắng.

Ta tìm $P(A/B)$.

Ta thấy lần thứ nhất lấy được viên bi trắng (B đã xảy ra) nên trong hợp còn 7 viên bi trong đ ó có 4 viên bi trắng. Do đó

$$P(A/B) = \frac{C_4^1}{C_7^1} = \frac{4}{7}$$

⊙ Công thức

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

Chứng minh

Giả sử phép thử có n biến cố đồng khả năng có thể xảy ra trong đó có m_A biến có thuận lợi cho biến cố A, m_B biến cố thuận lợi cho biến cố B và k biến cố thuận lợi cho biến cố AB.

Theo định nghĩa xác suất theo lối cổ điển ta có

$$P(AB) = \frac{k}{n}, \quad P(B) = \frac{m_B}{n}$$

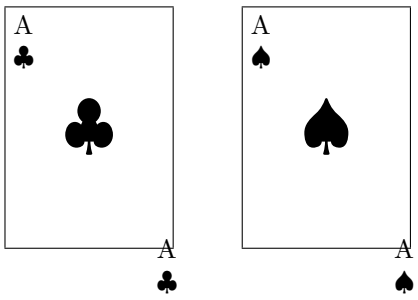
Ta tìm $P(A/B)$. Vì biến cố B đã xảy ra nên biến cố đồng khả năng của A là m_B, biến cố thuận lợi cho A là k. Do đó

$$P(A/B) = \frac{k}{m_B} = \frac{\frac{k}{n}}{\frac{m_B}{n}} = \frac{P(AB)}{P(B)}.$$

• **Ví dụ 26** *Một bộ bài có 52 lá. Rút ngẫu nhiên 1 lá bài. Tìm xác suất để rút được con "át" biết rằng lá bài rút ra là lá bài màu đen.*

Giải

Gọi A là biến cố rút được con "át"
B là biến cố rút được lá bài màu đen.



Ta thấy trong bộ bài có

26 lá bài đen nên $P(B) = \frac{26}{52}$

2 con "át" đen nên $P(AB) = \frac{2}{52}$.

Do đó $P(A/B) = \dfrac{P(AB)}{P(B)} = \dfrac{2/52}{26/52} = \dfrac{1}{13}$

b) Công thức nhân xác suất

Từ công thức xác suất có điều kiện ta có

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

ii) Nếu A, B là hai biến cố độc lập thì $P(AB) = P(A).P(B)$.

iii) $P(ABC) = P(A).P(B/A).P(C/AB)$
$P(A_1A_2 \dots A_n) = P(A_1)P(A_2/A_1) \dots P(A_n/A_1A_2 \dots A_{n-1})$.

• **Ví dụ 27** *Hộp thứ nhất có 2 bi trắng và 10 bi đen. Hộp thứ hai có 8 bi trắng và 4 bi đen. Từ mỗi hộp lấy ra 1 viên bi. Tìm xác suất để*

a) Cả 2 viên bi đều trắng,
b) 1 bi trắng, 1 bi đen.

Giải

Gọi T là biến cố lấy ra được cả 2 bi trắng
T_1 là biến cố lấy được bi trắng tù hộp thứ nhất
T_2 là biến cố lấy được bi trắng tù hộp thứ hai

thì T_1, T_2 là 2 biến cố độc lập và $T = T_1T_2$. Ta có

$$P(T_1) = \frac{1}{6}, \quad P(T_2) = \frac{2}{3}$$

Do đó $P(T) = P(T_1T_2) = P(T_1).P(T_2) = \frac{1}{6}.\frac{2}{3} = \frac{1}{9}$.

b) Gọi T_1, T_2 là biến cố lấy được bi trắng ở hộp thứ nhất, thứ hai
D_1, D_2 là biến cố lấy được bi đen ở hộp thứ nhất, thứ hai
T_1D_2 là biến cố lấy được bi trắng ở hộp thứ nhất và bi đen ở hộp thứ hai
T_2D_1 là biến cố lấy được bi trắng ở hộp thứ hai và bi đe n ở hộp thứ nhất

thì $A = T_1D_2 + T_2D_1$.

Ta có

$$P(T_1) = \frac{1}{6}, \quad P(T_2) = \frac{2}{3}$$

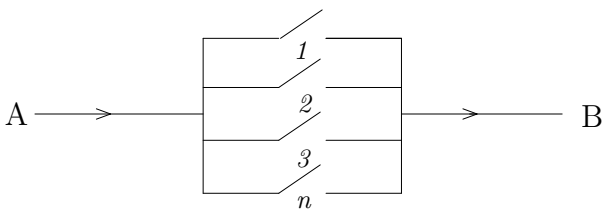
$$P(D_1) = 1 - P(T_1) = \frac{5}{6} \quad P(D_2) = 1 - P(T_2) = \frac{1}{3}$$

Suy ra

$$P(A) = P(T_1D_2) + P(T_2D_1) = P(T_1).P(D_2) + P(T_2).P(T_1)$$

$$= \frac{1}{6}.\frac{1}{3} + \frac{2}{3}.\frac{5}{6} = \frac{11}{8}$$

• **Ví dụ 28** *Một hệ thống được cấu thành bởi n thành phần riêng lẻ được gọi là một hệ thống song song nếu nó hoạt động khi ít nhất một thành phần hoạt động. Thành phần thứ i (độc lập với các thành phần khác) hoạt động với xác suất p_i. Tìm xác suất để hệ thống song song hoạt động.*



Giải

Gọi

A là biến cố hệ thống hoạt động.

A_i là biến cố thành phần thứ i hoạt động.

Ta có

$$\begin{aligned} \mathrm{P(A)} &= 1 - P(\overline{A}) \\ &= 1 - P(\overline{A_1}.\overline{A_2}\ldots\overline{A_n}) \\ &= 1 - \prod_{i=1}^{n} \overline{P(A_i)} \\ &= 1 - \prod_{i=1}^{n} (1 - p_i) \end{aligned}$$

• **Ví dụ 29** (`Hệ xích`) *Xét một hệ thống gồm hai thành phần. Hệ thống hoạt động khi và chỉ khi cả hai thành phần hoạt động (các thành phần được nối theo xích).*

Độ tin cậy R(t) của một thành phần của hệ thống là xác suất mà thành phần có thể hoạt động ít nhất khoảng thời gian t.

Nếu kí hiệu biến cố "thành phần hoạt động ít nhất t đơn vị thời gian" bởi $T > t$ thì

$$R(t) = P(T > t)$$

Gọi P_A và P_B là độ tin cậy của thành phần A và B, nghĩa là

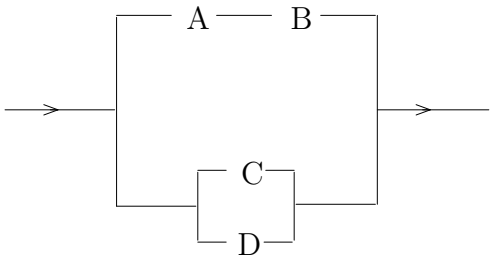
$P_A = P(A$ hoạt động ít nhất t đơn vị thời gian),

$P_B = P(B$ hoạt động ít nhất t đơn vị thời gian).

Nếu các thành phần hoạt động độc lập thì độ tin cậy của hệ thống là $R = p_A.p_B$.

• **Ví dụ 30**

Xét độ tin cậy của hệ thống cho bởi hình bên. Thành phần nối A và B trên đỉnh có thể thay bởi thành phần đơn với độ tin cậy $p_A.p_B$. Thành phần song song của ngắt C và D có thể thay bởi ngắt đơn với độ tin cậy $1-(1-p_C).(1-p_D)$.



Độ tin cậy của hệ thống song song này là

$$1 - (1 - p_A.p_B)[1 - (1 - (1 - p_C).(1 - p_D))]$$

4.3 Công thức xác suất đầy đủ và công thức Bayes

a) Công thức xác suất đầy đủ

⊙ **Công thức**

Giả sử $A_1, A_2, \ldots, A_n$ là nhóm các biến cố đầy đủ xung khắc từng đôi và B là biến cố bất kỳ có thể xảy ra trong phép thử. Khi đó ta có

$$\boxed{P(B) = \sum_{i=1}^{n} P(A_i).P(B/A_i)}$$

Chứng minh

Vì $A_1 + A_2 + \ldots + A_n = \Omega$ nên

$$B = B(A_1 + A_2 + \ldots + A_n) = BA_1 + B_2 + \ldots + BA_n$$

Do các biến cố $A_1, A_2, \ldots, A_n$ xung khắc từng đôi nên các biến cố tích $BA_1, BA_2, \ldots, BA_n$ cũng xung khắc từng đôi.

Theo định lý cộng xác suất ta có $P(B) = \sum_{i=1}^{n} P(BA_i)$.

Mặt khác theo công thức nhân xác suất thì $P(BA_i) = P(A_i).P(B/A_i)$.

Do đó $P(B) = \sum_{i=1}^{n} P(A_i).P(B/A_i)$.

⊙ **Chú ý** Công thức trên còn đúng nếu ta thay điều kiện $A_1 + A_2 + \ldots + A_n = \Omega$ bởi $B \subset A_1 + A_2 + \ldots + A_n$.

• **Ví dụ 31** *Xét một lô sản phẩm trong đó số sản phẩm do nhà máy I sản xuất chiếm 20%, nhà máy II sản xuất chiếm 30%, nhà máy III sản xuất chiếm 50%. Xác suất phế phẩm của nhà máy I là 0,001; nhà máy II là 0,005; nhà máy III là 0,006. Tìm xác suất để lấy ngẫu nhiên được đúng 1 phế phẩm.*

Giải

Gọi B là biến cố sản phẩm lấy ra là phế phẩm

A_1, A_2, A_3 là biến cố lấy được sản phẩm của nhà máy I, II, III

thì A_1, A_2, A_3 là nhóm các biến cố xung khắc từng đôi. Ta có

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

$$P(B/A_1) = 0,001; \quad P(B/A_2) = 0,005; \quad P(B/A_3) = 0,006$$

Do đó

$$\begin{aligned} P(B) &= P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + P(A_3).P(B/A_3) \\ &= 0,2.0,001 + 0,3.0,005 + 0,5.0,006 \\ &= 0,0065 \end{aligned}$$

• **Ví dụ 32** *Một hộp chứa 4 bi trắng, 3 bi vàng và 1 bi xanh. Lấy lần lượt (không hoàn lại) từ hộp ra 2 bi. Tìm xác suất để lấy được 1 bi trắng và 1 bi vàng.*

Giải

Gọi T là biến cố lấy được bi trắng, V là biến cố lấy được bi vàng.

Ta có

$$P(T) = \frac{4}{8} = \frac{1}{2}; \qquad P(V) = \frac{3}{8};$$
$$P(V/T) = \frac{3}{7}; \qquad P(T/V) = \frac{4}{7}$$

Xác xuất để lấy được 1 bi trắng và 1 bi vàng là

$$P(TV) = P(T).P(V/T) + P(V).P(T/V) = \frac{1}{2}.\frac{3}{7} + \frac{3}{8}.\frac{4}{7} = \frac{3}{7}.$$

□ **Cây xác suất**

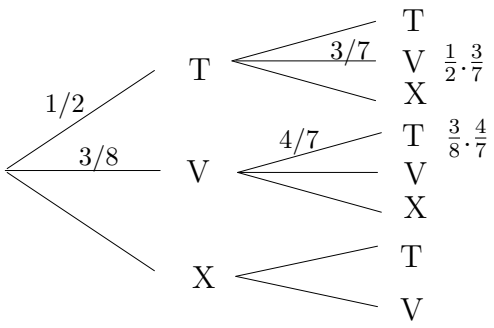
Trong thực tế có nhiều phép thử chứa một dãy nhiều biến cố. *Cây xác suất* cung cấp cho ta một công cụ thuận lợi cho việc xác định cấu trúc các quan hệ bên trong các phép thử khi tính xác suất.

Cấu trúc của cây xác suất được xác định như sau:

i) Vẽ biểu đồ cây xác suất tương ứng với các kết quả của dãy phép thử.

ii) Gán mỗi xác suất với mỗi nhánh.

Cây xác suất sau minh họa cho ví dụ 32.



b) Công thức Bayes

⊙ **Công thức**

Giả sử $A_1, A_2, \ldots, A_n$ là nhóm các biến cố đầy đủ xung khắc từng đôi và B là biến cố bất kỳ có thể xảy ra trong phép thử. Khi đó ta có

$$\boxed{P(A_i/B) = \frac{P(A_i).P(B/A_i)}{\sum_{i=1}^{n} P(A_i).P(B/A_i)}} \quad i = 1, 2, \ldots, n$$

Chứng minh

Theo công thức xác suất có điều kiện ta có

$$P(A_i/B) = \frac{(A_iB)}{P(B)} = \frac{P(A_i).P(B/A_i)}{P(B)}$$

Mặt khác theo công thức xác suất đầy đủ thì $P(B) = \sum_{i=1}^{n} P(A_i).P(B/A_i)$.

Do đó $P(A_i/B) = \frac{P(A_i).P(B/A_i)}{\sum_{i=1}^{n} P(A_i).P(B/A_i)}$.

• **Ví dụ 33** *Giả sử có 4 hộp như nhau đựng cùng một chi tiết máy, trong đó có một hộp 3 chi tiết xấu, 5 chi tiết tốt do máy I sản suất; còn ba hộp còn lại mỗi hộp đựng 4 chi tiết xấu, 6 chi tiết tốt do máy II sản suất. Lấy ngẫu nhiên một hộp rồi từ hộp đó lấy ra một chi tiết máy.*

a) Tìm xác suất để chi tiết máy lấy ra là tốt.

b) Với chi tiết tốt ở câu a, tìm xác suất để nó được lấy ra từ hộp của máy I.

Giải

Gọi B là biến cố lấy được chi tiết tốt

A_1, A_2 là biến cố lấy được hộp đựng chi tiết máy của máy I, II

thì A_1, A_2 là nhóm các biến cố xung khắc từng đôi.

a)

$$P(B) = P(A_1).P(B/A_1) + P(A_2).P(B/A_2)$$

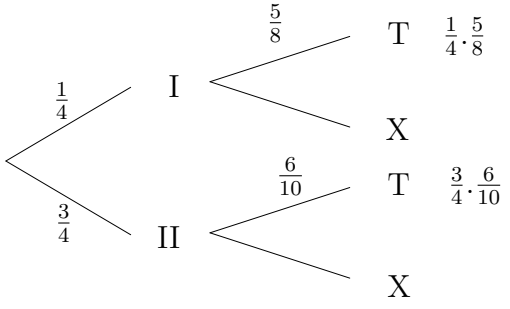
$$P(A_1) = \frac{1}{4}; \quad P(B/A_1) = \frac{5}{8}; \quad P(A_2) = \frac{3}{4}; \quad P(B/A_2) = \frac{6}{10}$$

Do đó

$$P(B) = \frac{1}{4}.\frac{5}{8} + \frac{3}{4}.\frac{6}{10} = \frac{97}{160}$$

b) $P(A_1/B) = \frac{P(A_1).P(B/A_1)}{P(B)} = \frac{\frac{1}{4}.\frac{5}{8}}{\frac{97}{160}} = \frac{26}{97}$

* Cây xác suất của câu a) cho bởi



• **Ví dụ 34** *Một hộp có 4 sản phẩm tốt được trộn lẫn với 2 sản phẩm xấu. Lấy ngẫu nhiên lần lượt từ hộp ra 2 sản phẩm. Biết sản phẩm lấy ra ở lần hai là sản phẩm tốt. Tìm xác suất để sản phẩm lấy ra ở lần thứ nhất cũng là sản phẩm tốt.*

Giải

Gọi A là biến cố sản phẩm lấy ra lần thứ nhất là sản phẩm tốt.

B là biến cố sản phẩm lấy ra lần thứ hai là sản phẩm tốt.

Ta có

$$P(A) = \frac{4}{6}, \quad P(B|A) = \frac{3}{5}, \quad P(\overline{A}) = \frac{2}{6}, \quad P(B|\overline{A}) = \frac{4}{5}$$

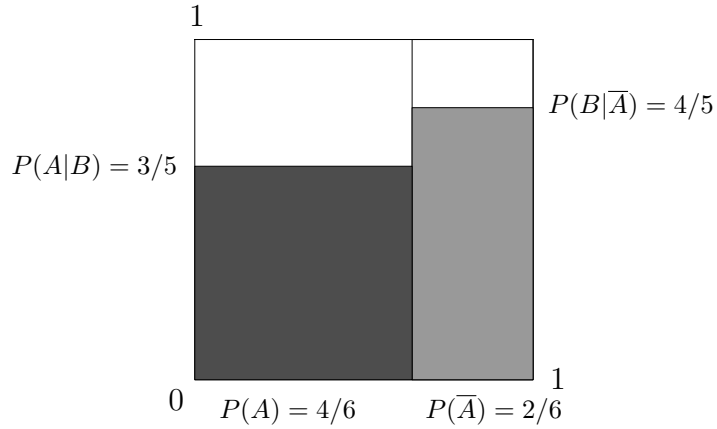
Theo định lý Bayes thì xác suất cần tìm là

$$P(A|B) = \frac{P(A).P(B|A)}{P(A).P(B|A) + P(\overline{A}).P(B|\overline{A})} = \frac{\frac{4}{6}.\frac{3}{5}}{\frac{4}{6}.\frac{3}{5} + \frac{2}{6}.\frac{4}{5}} = \frac{3}{5}.$$

⊙ **Chú ý** Ta có thể nhìn định lý Bayes theo cách hình học thông qua việc việc minh họa ví dụ trên như sau:

Vẽ một hình vuông cạnh 1. Chia trục hoành theo các tỉ số
$P(A) = \frac{4}{6}$, $P(\overline{A}) = \frac{2}{6}$.
Trục tung chỉ các xác suất có điều kiện
$P(B|A) = \frac{3}{5}$, $P(B|\overline{A}) = \frac{4}{5}$.

Vùng sậm nhiều trên $P(A)$ chỉ $P(A).P(B|A)$.
Vùng sậm toàn bộ chỉ
$P(B) = \frac{4}{6}.\frac{3}{5} + \frac{2}{6}.\frac{4}{5} = \frac{2}{3}$.



Xác suất $P(A|B) = \frac{\frac{4}{6}.\frac{3}{5}}{\frac{4}{6}.\frac{3}{5}+\frac{2}{6}.\frac{4}{5}} = \frac{3}{5}$ là tỉ số giữa vùng sậm nhiều và vùng sậm toàn bộ.

• **Ví dụ 35** (Theo thời báo New York ngày 5/9/1987)

Một "test" kiểm tra sự hiện diện của virus HIV (human immunodeficiency virus) cho kết quả dương tính nếu bệnh nhân thực sự nhiễm virus. Tuy nhiên, test này cũng có sai sót. Đôi khi cho kết quả dương tính đối với người không bị nhiễm virus, tỷ lệ sai sót là 1/20000. Giả sử kiểm tra ngẫu nhiên 10.000 người thì có 1 người nhiễm virus. Tìm tỷ lệ người có kết quả dương tính thực sự nhiễm virus.

Giải

Gọi A là biến có người bệnh bị nhiễm virus và

T^+ là biến có test cho kết quả dương tính

thì $P(A) = 0,0001;\quad P(T^+/A) = 1;\quad P(T^+/\overline{A}) = \frac{1}{20000}$

Theo định lý Bayes ta có

$$\begin{aligned} P(A/T^+) &= \frac{P(A).P(T^+/A)}{P(A).P(T^+/A) + P(A).P(T^+/\overline{A})} \\ &= \frac{(0,0001).1}{(0,0001).1 + (0,9999).\frac{1}{20000}} \\ &= \frac{20000}{29999} \end{aligned}$$

5. DÃY PHÉP THỬ BERNOULLI

□ **Định nghĩa 11** *Tiến hành n phép thử độc lập. Giả sử trong mỗi phép thử chỉ có thể xảy ra một trong hai trường hợp: hoặc biến cố A xảy ra hoặc biến cố A không xảy ra. Xác suất để A xảy ra trong mỗi phép thử đều bằng p. Dãy phép thử thỏa mãn các điều kiện trên được gọi là dãy phép thử Bernoulli.*

⊙ **Công thức Bernoulli**

Xác suất để biến cố A xuất hiện k lần trong n phép thử của dãy phép thử Bernoulli cho bởi

$$\boxed{P_n(k) = C_n^k p^k q^{n-k}} \qquad (q = 1 - p;\ k = 0, 1, 2, \ldots, n)$$

Chứng minh. Xác suất của một dãy n phép thử độc lập bất kỳ trong đó biến cố A xảy ra k lần (biến cố A không xảy ra $n - k$ lần) bằng $p^k q^{n-k}$. Vì có C_n^k dãy như vậy nên xác suất để biến cố A xảy ra k lần trong n phép thử là $P_n(k) = C_n^k p^k q^{n-k}$ $(q = 1 - p;\ k = 0, 1, 2, \ldots, n)$ □

• **Ví dụ 36** *Một bác sĩ có xác suất chữa khỏi bệnh là 0,8. Có người nói rằng cứ 10 người đến chữa thì chắc chắn có 8 người khỏi bệnh. Điều khẳng định đó có đúng không?*

Giải

Điều khẳng định trên là sai. Ta có xem việc chữa bệnh cho 10 người là một dãy của 10 phép thử độc lập. Gọi A là biến cố chữa khỏi bệnh cho một người thì $P(A) = 0,8$.

Do đó xác suất để trong 10 người đến chữa có 8 người khỏi bệnh là

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

• **Ví dụ 37** *Bắn 5 viên đạn độc lập với nhau vào cùng một bia, xác suất trúng đích các lần bắn như nhau và bằng 0,2. Muốn bắn hỏng bia phải có ít nhất 3 viên đạn bắn trúng đích. Tìm xác suất để bia bị hỏng.*

Giải

Gọi k là số đạn bắn trúng bia thì xác suất để bia bị hỏng là

$$\begin{aligned} P(k \geq 3) &= P_5(3) + P_5(4) + P_5(5) \\ &= C_5^3 p^3 q^2 + C_5^4 p^4 q + C_5^5 p^5 \\ &= 0{,}0512{+}0{,}0064{+}0{,}0003 \\ &= 0{,}0579 \end{aligned}$$

6. BÀI TẬP

1. Gieo đồng thời hai con xúc sắc. Tìm xác suất để:
   (a) Tổng số nốt xuất hiện trên hai con là 7.
   (b) Tổng số nốt xuất hiện trên hai con là 8.
   (c) Số nốt xuất hiện hai con hơn kém nhau 2.

2. Có 12 hành khách lên một tàu điện có 4 toa một cách ngẫu nhiên. Tìm xác suất để:
   (a) Mỗi toa có 3 hành khách;
   (b) Một toa có 6 hành khách, một toa có 4 hành khách, hai toa còn lại mỗi toa có 1 hành khách.

3. Có 10 tấm thẻ được đánh số từ 0 đến 9. Lấy ngẫu nhiên hai tấm thẻ xếp thành một số gồm 2 chữ số. Tìm xác suất để số đó chia hết cho 18.

4. Trong hộp có 6 bi đen và 4 bi trắng. Rút ngẫu nhiên từ hộp ra 2 bi. Tìm xác suất để được:
   (a) 2 bi đen,
   (b) ít nhất 1 bi đen,
   (c) bi thứ hai màu đen.

5. Cho ba biến cố A, B, C có các xác suất

   $P(A) = 0,525, \quad P(B) = 0,302, \quad P(C) = 0,480,$

   $P(AB) = 0,052, \quad P(BC) = 0,076, \quad P(CA) = 0,147, \quad P(ABC) = 0,030.$

   Chứng minh rằng các số liệu đã cho không chính xác.

6. Trong tủ có 8 đôi giày. Lấy ngẫu nhiên ra 4 chiếc giày. Tìm xác suất sao cho trong các chiếc giày lấy ra
   (a) không lập thành một đôi nào cả.
   (b) có đúng 1 đôi giày.

7. Một người bỏ ngẫu nhiên 3 lá thư vào 3 chiếc phong bì đã ghi địa chỉ. Tính xác suất để ít nhất có một lá thư bỏ đúng phong bì của nó.

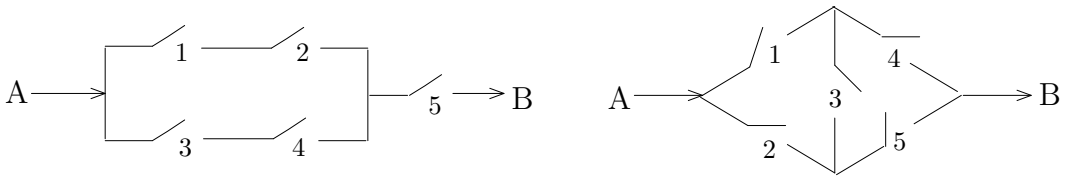
8. Một phòng điều trị có 3 bệnh nhân với xác suất cần cấp cứu trong một ca trực là 0,7; 0,8 và 0,9. Tìm xác suất sao cho trong một ca trực:

   (a) Có 2 bệnh nhân cần cấp cứu.

   (b) Có ít nhất 1 bệnh không cần cấp cứu.

9. Biết xác suất để một học sinh đạt yêu cầu ở lần thi thứ i là p_i $(i = 1, 2)$. Tìm xác suất để học sinh đó đạt yêu cầu trong kỳ thi biết rằng mỗi học sinh được phép thi tối đa 2 lần.

10. Cho 2 mạch điện như hình vẽ



(a) (b)

Giả sử xác suất để dòng điện qua ngắt i là p_i. Tìm xác suất có dòng điện đi từ A đến B.

11. Gieo đồng thời hai con xúc xắc cân đối đồng chất 20 lần liên tiếp. Tìm xác suất để xuất hiện ít nhất một lần 2 mặt trên cùng có 6 nốt.

12. Một sọt cam rất lớn được phân loại theo cách sau. Chọn ngẫu nhiên 20 quả cam làm mẫu đại diện. Nếu mẫu không có quả cam hỏng nào thì sọt cam được xếp loại 1. Nếu mẫu có một hoặc hai quả hỏng thì sọt cam được ees p loại 2. Trong trường hợp còn lại (có từ 3 quả hỏng trở lên) thì sọt cam được xếp loại 3.

    Giả sử tỉ lệ cam hỏng của sọt cam là 3%. Hãy tính xác suất để:

    (a) Sọt cam được xếp loại 1.

    (b) Sọt cam được xếp loại 2.

    (c) Sọt cam được xếp loại 3.

13. Một nhà máy sản xuất tivi có90% sản phẩm đạt tiêu chuẩn kỹ thuật. Trong quá trình kiểm nghiệm, xác suất để chấp nhận một sản phẩm đạt tiêu chuẩn kỹ thuật là 0,95 và xác suất để chấp nhận một sản phẩm không đạt kỹ thuật là 0,08. Tìm xác suất để một sản phẩm đạt tiêu chuẩn kỹ thuật qua kiểm nghiệm được chấp nhận.

14. Một công ty lớn A hợp đồng sản xuất bo mạch, 40% đối với công ty B và 60 % đối với công ty C. Công ty B lại hợp đồng 70% bo mạch nó nhận được từ công ty A với công ty D và 30% đối với công ty E. Khi bo mạch được hoàn thành từ các công ty C, D và E, chúng được đưa đến công ty A để gắn vào các model khác

nhau của máy tính. Người ta nhận thấy 1,5%, 1% và 5% tương ứng của các bo mạch của công ty D, C và E hư trong vòng 90 ngày bảo hành sau khi bán. Tìm xác suất bo mạch của máy tính bị hư trong khoảng thời gian 90 ngày đ ược bảo hành.

15. Biết rằng một người có nhóm máu AB có thể nhận máu của bất kỳ nhóm máu nào. Nếu người đó có nhóm máu còn lại (A, B hoặc O) thì chỉ có thể nhận máu của người có cùng nhóm máu với mình hoặc nhóm máu O.

    Cho biết tỷ lệ người có nhóm máu O, A, B và AB tương ứng là 33,7%; 37,5%; 20,9% và 7,9%.

    (a) Chọn ngẫu nhiên một người cần tiếp máu và một người cho máu. Tính xác suất để sự truyền máu được thực hiện.

    (b) Chọn ngẫu nhiên một người cần tiếp máu và hai người cho máu. Tính xác suất để sự truyền máu được thực hiện.

16. Lô hàng thứ I có 5 chính phẩm và 3 phế phẩm. Lô hàng thứ II có 3 chính phẩm và 2 phế phẩm.

    (a) Lấy ngẫu nhiên từ mỗi lô hàng ra 1 sản phẩm.

        i) Tìm xác suất để lấy được 2 chính phẩm.

        ii) Tìm xác suất để lấy được 1 chính phẩm và 1 phế phẩm.

        iii) Giả sử lấy được 1 chính phẩm và 1 phế phẩm. Tìm xác suất để phế phẩm là của lô hàng thứ I.

    (b) Chọn ngẫu nhiên một lô hàng rồi từ đó lấy ra 2 sản phẩm. Tìm xác suất để lấy được 2 chính phẩm.

◙ TRẢ LỜI BÀI TẬP

1. (a) $\frac{1}{6}$, (b) $\frac{5}{36}$, (c) $\frac{2}{9}$.
2. (a) $\frac{12!}{(3!)^4.4^{12}}$, (b) $\frac{12!}{6!4!4^{12}}$
3. $\frac{1}{8}$.
4. (a) $\frac{1}{3}$, (b) $\frac{3}{5}$, (c) $\frac{3}{5}$.
6. (a) 0,6154 ; (b) 0,3692.
7. $\frac{2}{3}$.
8. (a) 0,398; (b) 0,496.
9. $p_1 + (1 - p_1)p_2$.
10. $1 - (\frac{35}{36})^{20}$.
12. (a) $p = (0,97)^{20} = 0,5438$,

    (b) $p = 20(0,03)(0,97)^{19} + 190(0,03)^2.(0,97)^{18} = 0,4352$,

    (c) $1 - 0,54338 - 0,4352 = 0,021$
13. 0,99
14. $p = 0,4.0,7.0,015 + 0,4.0,3.0,01 + 0,6.0,005 = 0,0084$.

15. (a) 0,5737; (b) 0,7777.

16. (a) i) $\frac{3}{8}$, ii) $\frac{19}{40}$, iii) $\frac{9}{19}$, (b) $\frac{23}{70}$.

Chủởng 2

ĐẠI LƯỢNG NGẪU NHIÊN VÀ PHÂN PHỐI XÁC SUẤT

1. ĐẠI LƯỢNG NGẪU NHIÊN

1.1 Khái niệm đại lượng ngẫu nhiên

□ **Định nghĩa 1** *Đại lượng ngẫu nhiên là đại lượng biến đổi biểu thị giá trị kết qủa của một phép thử ngẫu nhiên.*

Ta dùng các chữ cái hoa như X, Y, Z, ... để kí hiệu đại lượng ngẫu nhiên.

• **Ví dụ 1** *Tung một con xúc xắc. Gọi X là số chấm xuất hiện trên mặt con xúc xắc thì X là một đại lượng ngẫu nhiên nhận các giá trị có thể là 1, 2, 3, 4, 5, 6.*

1.2 Đại lượng ngẫu nhiên rời rạc

a) Đại lượng ngẫu nhiên rời rạc

□ **Định nghĩa 2** *Đại lượng ngẫu nhiên được gọi là rời rạc nếu nó chỉ nhận một số hữu hạn hoặc một số vô hạn đếm được các giá trị.*

Ta có thể liệt kê các giá trị của đại lượng ngẫu nhiên rời rạc $x_1, x_2, \ldots, x_n$.

Ta kí hiệu đại lượng ngẫu nhiên X nhận giá trị x_n là $X = x_n$ và xác suất để X nhận giá trị x_n là $P(X = x_n)$.

• **Ví dụ 2** *Số chấm xuất hiện trên mặt con xúc xắc, số học sinh vắng mặt trong một buổi học...là các đại lượng ngẫu nhiên rời rạc.*

b) Bảng phân phối xác suất

Bảng phân phối xác suất dùng để thiết lập luật phân phối xác suất của đại lượng ngẫu nhiên rời rạc, nó gồm 2 hàng: hàng thứ nhất liệt kê các giá trị có thể $x_1, x_2, \ldots, x_n$ của đại lượng ngẫu nhiên X và hàng thứ hai liệt kê các xác suất tương ứng $p_1, p_2, \ldots, p_n$ của các giá trị có thể đó.

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

Nếu các giá trị có thể của đại lượng ngẫu nhiên X gồm hữu hạn số $x_1, x_2, \ldots, x_n$ thì các biến cố $X = x_1, X = x_2, \ldots, X = x_n$ lập thành một nhóm các biến cố đầy đủ xung khắc từng đôi.

Do đó $\sum_{i=1}^{n} p_i = 1$.

• **Ví dụ 3** *Tung một con xúc xắc đồng chất. Gọi X là số chấm xuất hiện trên mặt con xúc xắc thì X là đại lượng ngẫu nhiên rời rạc có phân phối xác suất cho bởi:*

X	*1*	*2*	*3*	*4*	*5*	*6*
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

1.3 Đại lượng ngẫu nhiên liên tục và hàm mật độ xác suất

a) Đại lượng ngẫu nhiên liên tục

□ **Định nghĩa 3** *Đại lượng ngẫu nhiên được gọi là liên tục nếu các giá trị có thể của nó lấp đầy một khoảng trên trục số.*

• **Ví dụ 4**

- *Nhiệt độ không khí ở mỗi thời điểm nào đó.*
- *Sai số khi khi đo lường một đại lượng vật lý.*
- *Khoảng thời gian giữa hai ca cấp cứu của một bệnh viện.*

b) Hàm mật độ xác suất

□ **Định nghĩa 4** *Hàm mật độ xác suất của đại lượng ngẫu nhiên liên tục X là hàm không âm f(x), xác định với mọi $x \in (-\infty, +\infty)$ thỏa mãn*

$$P(X \in B) = \int_B f(x)dx$$

với mọi tập số thực B.

◇ **Tính chất** Hàm mật độ xác suất có các tính chất sau

i) $f(x) \geq 0, \ \forall x \in (-\infty, +\infty)$

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

⊙ **Ý nghĩa của hàm mật độ**

Tù định nghĩa của hàm mật độ ta có $P(x \leq X \leq x + \triangle x) \sim f(x).\triangle x$

Do đó ta thấy xác suất để X nhận giá trị thuộc lân cận khá bé $(x, x + \triangle x)$ gần như tỉ lệ với f(x).

1.4 Hàm phân phối xác suất

□ **Định nghĩa 5** *Hàm phân phối xác suất của đại lượng ngẫu nhiên X, kí hiệu F(x), là hàm được xác định như sau*

$$F(x) = P(X < x)$$

** Nếu X là đại lượng ngẫu nhiên rời rạc nhận các giá trị có thể $x_1, x_2, \ldots, x_n$ thì*

$$F(x) = \sum_{x_i<x} P(X = x_i) = \sum_{x_i<x} p_i \qquad (\text{với } p_i = P(X = x_i))$$

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

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

◇ **Tính chất** Ta có thể chứng minh được các công thức sau

i) $0 \le F(x) \le 1; \quad \forall x.$

ii) F(x) là hàm không giảm ($x_1 \le x_2 \Longrightarrow F(x_1) \le F(x_2)$).

iii) $\lim\limits_{x\to-\infty} F(x) = 0; \quad \lim\limits_{x\to+\infty} F(x) = 1.$

iv) $F'(x) = f(x), \ \forall x.$

⊙ **Ý nghĩa của hàm phân phối xác suất**

Hàm phân phối xác suất F(x) phản ánh mức độ tập trung xác suất về bên trái của điểm x.

• **Ví dụ 5** *Cho đại lượng ngẫu nhiên rời rạc X có bảng phân phối xác suất*

X	*1*	*3*	*6*
P	*0,3*	*0,1*	*0,6*

Tìm hàm phân phối xác suất của X và vẽ đồ thị của hàm này.

Giải

Nếu $x \le 1$ thì $F(x) = 0.$

Nếu $1 < x \le 3$ thì $F(x) = 0,3.$

Nếu $3 < x \le 6$ thì $F(x) = 0,3 + 0,1 = 0,4.$

Nếu $x > 6$ thì $F(x) = 0,3 + 0,1 + 0,6 = 1.$

$$F(x) = \begin{cases} 0 & ; \quad x \le 1 \\ 0,3 & ; \quad 1 < x \le 3 \\ 0,4 & ; \quad 3 < x \le 6 \\ 1 & ; \quad x > 6 \end{cases}$$

• **Ví dụ 6** *Cho X là đại lượng ngẫu nhiên liên tục có hàm mật độ*

$$f(x) = \begin{cases} 0 & \text{nếu} \quad x < 0 \\ \frac{6}{5}x & \text{nếu} \quad 0 \le x \le 1 \\ \frac{6}{5x^4} & \text{nếu} \quad x > 1 \end{cases}$$

Tìm hàm phân phối xác suất F(x).

Giải

Khi $x < 0$ thì $F(x) = \int\limits_{-\infty}^{x} f(t)dt = 0$

Khi $0 \le x \le 1$ thì $F(x) = \int\limits_{-\infty}^{x} f(t)dt = \int\limits_{0}^{x} \frac{6}{5}tdt = \frac{3}{5}x^2.$

Khi $x > 1$ thì

$$F(x) = \int\limits_{-\infty}^{x} f(t)dt = \int\limits_{0}^{1} \frac{6}{5}tdt + \int\limits_{1}^{x} \frac{6}{5t^4}dt = \frac{3}{5} + \left[-\frac{2}{5t^3}\right]_1^x = 1 - \frac{2}{5x^3}$$

Vậy $F(x) = \begin{cases} 0 & ; \quad x < 0 \\ \frac{3}{5}x^2 & ; \quad 0 \le x \le 1 \\ 1 - \frac{2}{5x^3} & ; \quad x > 1 \end{cases}$

2. CÁC THAM SỐ ĐẶC TRƯNG CỦA ĐẠI LƯỢNG NGẪU NHIÊN

2.1 Kỳ vọng (Expectation)

□ Định nghĩa 6

** Giả sử X là đại lượng ngẫu nhiên rời rạc có thể nhận các giá trị $x_1, x_2, \ldots, x_n$ với các xác suất tương ứng $p_1, p_2, \ldots, p_n$. Kỳ vọng của đại lượng ngẫu nhiên X, kí hiệu E(X) (hay M(X)), là số được xác định bởi*

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

** Giả sử X là đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất $f(x)$. Kỳ vọng của đại lượng ngẫu nhiên X được xác định bởi*

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

• **Ví dụ 7** *Tìm kỳ vọng của đại lượng ngẫu nhiên có bảng phân phối xác suất sau*

X	5	6	7	8	9	10	11
P	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Ta có

$E(X) = 5.\frac{1}{12} + 6.\frac{2}{12} + 7.\frac{3}{12} + 8.\frac{2}{12} + 9.\frac{2}{12} + 10.\frac{1}{12} + 11.\frac{1}{12} = \frac{93}{12} = \frac{31}{4} = 7,75.$

• **Ví dụ 8** *Cho X là đại lượng ngẫu nhiên liên tục có hàm mật độ*

$$f(x) = \begin{cases} 2.e^{-2x} & \text{nếu } 0 < x < 2 \\ 0 & \text{nếu } x \notin (0,2) \end{cases}$$

Tìm E(X).

Giải

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{2} x.(\frac{1}{2}x) dx = \frac{x^3}{6}\Big|_0^2 = \frac{4}{3}$$

◇ **Tính chất**

i) $E(C) = C$, C là hằng.

ii) $E(cX) = c.E(X)$.

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

iv) Nếu X và Y là hai đại lượng ngẫu nhiên độc lập thì $E(XY) = E(X).E(Y)$.

⊙ **Ý nghĩa của kỳ vọng**

Tiến hành n phép thử. Giả sử X là đại lượng ngẫu nhiên nhận các giá trị có thể $x_1, x_2, \ldots, x_n$ với số lần nhận $k_1, k_2, \ldots, k_n$.

Giá trị trung bình của đại lượng ngẫu nhiên X trong n phép thử là

$$\overline{x} = \frac{k_1 x_1 + k_2 x_2 + \ldots + k_n x_n}{n} = \frac{k_1}{x} x_1 + \frac{k_2}{n} x_2 + \ldots + \frac{k_n}{n} x_n = f_1 x_1 + f_2 x_2 + \ldots + f_n k_n$$

với $f_i = \frac{k_i}{n}$ là tần suất để X nhận giá trị x_i.

Theo định nghĩa xác suất theo lối thống kê ta có $\lim_{n\to\infty} f_i = p_i$. Vì vậy với n đủ lớn ta có

$$\overline{x} \approx p_1x_1 + p_2x_2 + \ldots + p_nx_n = E(X)$$

Ta thấy kỳ vọng của đại lượng ngẫu nhiên xấp xỉ với trung bình số học các giá trị quan sát của đại lượng ngẫu nhiên.

Do đó có thể nói *kỳ vọng của đại lượng ngẫu nhiên chính là giá trị trung bình (theo xác suất) của đại lượng ngẫu nhiên. Nó phản ánh giá trị trung tâm của phân phối xác suất*

2.2 Phương sai (Variance)

□ **Định nghĩa 7** *Phương sai (độ lệch bình phương trung bình) của đại lượng ngẫu nhiên X, kí hiệu Var(X) hay D(X), được định nghĩa bằng công thức*

$$Var(X) = E\{[X - E(X)]^2\}$$

** Nếu X là đại lượng ngẫu nhiên rời rạc nhận các giá trị có thể $x_1, x_2, \ldots, x_n$ với các xác suất tương ứng $p_1, p_2, \ldots, p_n$ thì*

$$Var(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 p_i$$

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

$$Var(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x)dx$$

⊙ **Chú ý** Trong thực tế ta thường tính phương sai bằng công thức

$$\boxed{Var(X) = E(X^2) - [E(X)]^2}$$

Thật vậy, ta có

$$\begin{aligned} Var(X) &= E\{X - E(X)]^2\} \\ &= E\{X^2 - 2X.E(X) + [E(X)]^2\} \\ &= E(X^2) - 2E(X).E(X) + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

• **Ví dụ 9** *Cho đại lượng ngẫu nhiên rời rạc X có bảng phân phối xác suất sau*

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

Tìm phương sai của X.

Giải

E(X)=1.0,1+3.0,4+5.0,5=3,8

$E(X^2) = 1^2.0,1 + 3^2.0,4 + 5^2.0,5 = 16,2$

Do đó $Var(X) = E(X^2) - [E(X)]^2 = 16,2 - 14,44 = 1,76.$

• **Ví dụ 10** *Cho đại lượng ngẫunhiên X có hàm mật độ*

$$f(x) = \begin{cases} cx^3 & \text{với } 0 \le x \le 3 \\ 0 & \text{với } x \notin [0,3] \end{cases}$$

Hãy tìm

i) Hằng số c.

ii) Kỳ vọng.

iii) Phương sai

Giải

i) Ta có $1 = \int\limits_0^3 cx^3 dx = c\left[\frac{x^4}{4}\right]_0^3 = \frac{81}{4}c.$

Suy ra $c = \frac{4}{81}$.

ii) $E(X) = \int\limits_0^3 x\frac{4}{81}x^3 dx = \frac{4}{81}\left[\frac{x^5}{5}\right]_0^3 = 2,4.$

iii) Ta có

$$E(X^2) = \int\limits_{-\infty}^{\infty} x^2 f(x)dx = \int\limits_0^3 x^2\frac{4}{81}x^3 dx = \frac{4}{81}\left[\frac{x^6}{6}\right]_0^3 = 6$$

Vậy $Var(X) = E(X^2) - [E(X)]^2 = 6 - (2,4)^2 = 0,24.$

◇ **Tính chất**

i) Var(C)=0; (C không đổi).

ii) $Var(cX) = c^2.Var(X)$.

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

* $Var(X+Y) = Var(X) + Var(Y)$;

* Var(X-Y)=Var(X)+Var(Y);

* Var(C+X)=Var(X).

⊙ **Ý nghĩa của phương sai**

Ta thấy $X - E(X)$ là độ lệch khỏi giá trị trung bình nên $Var(X) = E\{[X - E(X)]^2\}$ là độ lệch bình phương trung bình. Do đó phương sai phản ánh mức độ phân tán các giá trị của đại lượng ngẫu nhiên chung quanh giá trị trung bình.

2.3 Độ lệch tiêu chuẩn

Đơn vị đo của phương sai bằng bình phương đơn vị đo của đại lượng ngẫu nhiên. Khi cần đánh giá mức độ phân tán các giá trị của đại lượng ngẫu nhiên theo đơn vị của nó, người ta dùng một đặc trưng mới đó là độ lệch tiêu chuẩn.

□ **Định nghĩa 8** *Độ lệch tiêu chuẩn của đại lượng ngẫu nhiên X, kí hiệu là $\sigma(X)$, được định nghĩa như sau:*

$$\sigma(X) = \sqrt{Var(X)}$$

2.4 Mode

□ **Định nghĩa 9** *Mod(X) là giá trị của đại lượng ngẫu nhiên X có khả năng xuất hiện lớn nhất trong một lân cận nào đó của nó.*

Đối với đại lượng ngẫu nhiên rời rạc mod(X) là giá trị của X ứng với xác suất lớn nhất, còn đối với đại lượng ngẫu nhiên liên tục thì mod(X) là giá trị của X tại đó hàm mật độ đạt giá trị cực đại.

⊙ **Chú ý** Một đại lượng ngẫu nhiên có thể có một mode hoặc nhiều mode.

• **Ví dụ 11** *Giả sử X là điểm trung bình của sinh viên trong trường thì mod(X) là điểm mà nhiều sinh viên đạt được nhất.*

• **Ví dụ 12** *Cho đại lượng ngẫu nhiên liên tục có phân phối Vây−bun với hàm mật độ*

$$f(x) = \begin{cases} 0 & \text{nếu} \quad x \le 0 \\ \frac{x}{2}e^{-\frac{x^2}{4}} & \text{nếu} \quad x > 0 \end{cases}$$

Hãy xác định mod(X).

Giải

mod(X) là nghiệm của phương trình

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

Suy ra mod(X) là nghiệm của phương trình $1 - \dfrac{x^2}{2} = 0$. Do $mod(X) > 0$ nên $mod(X) = \sqrt{2} = 1,414$.

2.5 Trung vị

□ **Định nghĩa 10** *Trung vị của đại lượng ngẫu nhiên X là giá trị của X chia phân phối xác suất thành hai phần có xác suất giống nhau. Kí hiệu med(X).*

Ta có $P(X < med(X)) = P(X \ge med(X)) = \frac{1}{2}$

⊕ **Nhận xét** Từ định nghĩa ta thấy để tìm trung vị chỉ cần giải phương trình $F(x) = \frac{1}{2}$. Trong ứng dụng, trung vị là đặc trưng vị trí tốt nhất, nhiều khi tốt hơn cả kỳ vọng, nhất là khi trong số liệu có nhiều sai sót. Trung vị còn được gọi là *phân vị 50% của phân phối.*

• **Ví dụ 13** *Tìm med(X) trong ví dụ (12).*

Giải

med(X) là nghiệm của phương trình

$$\int_0^{med(X)} f(x)dx = 0,5 \quad \text{hay} \quad 1 - e^{-\frac{[med(X)]^2}{4}} = 0,5$$

Suy ra $med(X) = 1,665$.

⊙ **Chú ý** Nói chung, ba số đặc trưng kỳ vọng, mode và trung vị không trùng nhau. Chẳng hạn, từ các ví dụ (12), (13) và tính thêm kỳ vọng ta có $E(X) = 1,772$; $mod(X) = 1,414$ và $med(X) = 1,665$. Tuy nhiên nếu phân phối đối xứng và chỉ có một mode thì cả ba đặc trưng đó trùng nhau.

2.6 Moment

□ **Định nghĩa 11**

** Moment cấp k của đại lượng ngẫu nhiên X là số $m_k = E(X^k)$.*
** Moment qui tâm cấp k của đại lượng ngẫu nhiên X là số $\alpha_k = E\{[X - E(X)]^k\}$.*

⊕ **Nhận xét**

i) Moment cấp 1 của X là kỳ vọng của X ($m_1 = E(X)$).

ii) Moment qui tâm cấp hai của X là phương sai của X ($\alpha_2 = m_2 - m_1^2 = Var(X)$).

iii) $\alpha_3 = m_3 - 3m_2m_1 + 2m_1^3$.

2.7 Hàm moment sinh

□ **Định nghĩa 12** *Hàm moment sinh của đại lượng ngẫu nhiên X là hàm xác định trong $(-\infty, +\infty)$ cho bởi*

$$\phi(t) = E(e^{tX}) = \begin{cases} \sum\limits_x e^{tx}p(x) & \text{nếu } X \text{ rời rạc} \\ \int\limits_{-\infty}^{+\infty} e^{tx}p(x)dx & \text{nếu } X \text{ liên tục} \end{cases}$$

◇ **Tính chất**

i) $\phi'(0) = E(X)$.

ii) $\phi''(0) = E(X^2)$.

iii) Tổng quát: $\phi^{(n)}(0) = E(X^n)$, $\forall n \geq 1$.

Chứng minh.

i) $\phi'(t) = \dfrac{d}{dt}E(e^{tX}) = E\left(\dfrac{d}{dt}(e^{tX})\right) = E(Xe^{tX})$.

Suy ra $\phi'(0) = E(X)$.

ii) $\phi''(t) = \dfrac{d}{dt}\phi'(t) = \dfrac{d}{dt}E(Xe^{tX}) = E\left(\dfrac{d}{dt}(Xe^{tX})\right) = E(X^2e^{tX})$.

Suy ra $\phi''(0) = E(X^2)$. □

⊙ **Chú ý**

i) Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập có hàm moment sinh tương ứng là $\phi_X(t)$ và $\phi_Y(t)$. Khi đó hàm moment sinh của $X+Y$ cho bởi

$$\phi_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY}) = \phi_X(t)\phi_Y(t)$$

(đẳng thức gần cuối có được do e^{tX} và e^{tY} độc lập)

ii) Có tương ứng 1−1 giữa hàm moment sinh và hàm phân phối xác suất của đại lượng ngẫu nhiên X.

3. MỘT SỐ QUI LUẬT PHÂN PHỐI XÁC SUẤT

3.1 Phân phối nhị thức (Binomial Distribution)

□ **Định nghĩa 13** *Đại lượng ngẫu nhiên rời rạc X nhận môt trong các giá trị 0,1,2,...,n với các xác suất tương ứng được tính theo công thức Bernoulli*

$$P_x = P(X = x) = C_n^x p^x q^{n-x} \tag{2.1}$$

gọi là có phân phối nhị thức với tham số n và p. Kí hiệu $X \in B(n,p)$ (hay $X \sim B(n,p)$).

⊙ **Công thức**

Với h nguyên dương và $h \le n - x$, ta có

$$\boxed{P(x \le X \le x+h) = P_x + P_{x+1} + \ldots + P_{x+h}} \tag{2.2}$$

• **Ví dụ 14** *Tỷ lệ phế phẩm trong lô sản phẩm là 3%. Lấy ngẫu nhiên 100 sản phẩm để kiểm tra. Tìm xác suất để trong đó*

i) Có 3 phế phẩm.

ii) Có không quá 3 phế phẩm.

Giải

Ta thấy mỗi lần kiểm tra một sản phẩm là thực hiện một phép thử. Do đó ta có n=100 phép thử.

Gọi A là biến cố sản phẩm lấy ra là phế phẩm thì trong mỗi phép thủ. Ta có $p = p(A) = 0,03$.

Đặt X là tổng số phế phẩm trong 100 sản phẩm thì $X \in B(100; 0,03)$.

i) $P(X = 3) = C^3_{100}(0,03)^3.(0,97)^{97} = 0,2274.$

ii) $$\begin{aligned} P(0 \le X \le 3) &= P_0 + P_1 + P_2 + P_3 \\ &= C^0_{100}(0,03)^0(0,97)^{100} + C^1_{100}(0,03)^1(0,97)^{99} \\ &\quad + C^2_{100}(0,03)^2(0,97)^{98} + C^3_{100}(0,03)^3(0,97)^{97} \\ &= 0,647. \end{aligned}$$

⊙ **Chú ý** Khi n khá lớn thì xác suất p không quá gần 0 và 1. Khi đó ta có thể áp dụng công thức xấp xỉ sau

i)

$$\boxed{P_x = C^x_n p^x q^{n-x} \approx \frac{1}{\sqrt{npq}} f(u)} \tag{2.3}$$

trong đó

$$u = \frac{x - np}{\sqrt{npq}}; \quad f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}};$$

(2.3) được gọi công thức địa phương Laplace.

ii)

$$\boxed{P(x \le X \le x + h) \approx \varphi(u_2) - \varphi(u_1)} \tag{2.4}$$

trong đó

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{t^2}{2}} dt \quad \text{(Hàm Laplace);}$$

$$u_1 = \frac{x - np}{\sqrt{npq}}; \quad u_2 = \frac{x + h - np}{\sqrt{npq}}$$

(2.4) được gọi là công thức tích phân Laplace.

⊙ **Các tham số đặc trưng**

Nếu $X \in B(n, p)$ thì ta có

i) $E(X) = np.$

ii) $Var(X) = npq.$

iii) $np - q \le mod(X) \le np + p.$

Chứng minh. Xét đại lượng ngẫu nhiên X có phân phối nhị thức với các tham số n và p biểu diễn phép thử biến cố A xảy ra, mỗi phép thử có cùng xác suất xảy ra biến cố A là p.

Ta có thể biểu diễn X như sau:

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

trong đó $X_i = \begin{cases} 1 & \text{nếu ở phép thử thứ } i \text{ biến cố } A \text{ xảy ra} \\ 0 & \text{nếu ngược lại} \end{cases}$

Vì $X_i,\ i = 1, 2, \ldots, n$ là các đại lượng ngẫu nhiên độc lập có phân phối nhị thức nên

$$E(X_i) = P(X_i = 1) = p$$

$$Var(X_i) = E(X_i^2) - p^2 = p(1-p) = pq \qquad (X_i^2 = X_i)$$

Do đó

$$E(X) = \sum_{i=1}^{n} E(X_i) = np$$

$$Var(X) = \sum_{i=1}^{n} Var(X_i) = npq$$

□

• **Ví dụ 15** *Một máy sản xuất được 200 sản phẩm trong một ngày. Xác suất để máy sản xuất ra phế phẩm là* $0,05$. *Tìm số phế phẩm trung bình và số phế phẩm có khả năng tin chác của máy đó trong một ngày.*

Giải

Gọi X là số phế phẩm của máy trong một ngày thì $X \in B(200; 0,05)$.

Số phế phẩm trung bình của máy trong một ngày là

$$E(X) = np = 200 \times 0,05 = 10$$

Số phế phẩm tin chắc trong ngày là mod(X). Ta có
$np - q = 200 \times 0,05 - 0,95 = 9,05$
$np + p = 200 \times 0,05 + 0,05 = 10,05$

$\Longrightarrow 9,05 \leq mod(X) \leq 10,05$

Vì $X \in B(200; 0,05)$ nên $mod(X) \in Z$. Do đó $mod(X) = 10$.

3.2 Phân phối Poisson

⊙ Công thức Poisson

Giả sử X là đại lượng ngẫu nhiên có phân phối nhị thức với tham số (n, p) và $a = np$ trong đó n khá lớn và p khá bé.

Ta có

$$\begin{aligned} P(X = k) &= \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \\ &= \frac{n!}{(n-k)!k!} . (\frac{a}{n})^k . (1 - \frac{a}{n})^{n-k} \\ &= \frac{n(n-1)\ldots(n-k+1)}{n^k} . \frac{a^k}{k!} . \frac{(1-\frac{a}{n})^n}{(1-\frac{a}{n})^k} \end{aligned}$$

Do n khá lớn và p khá bé nên

$$(1-\frac{a}{n})^n \approx e^{-a}, \qquad \frac{n(n-1)\dots(n-k+1)}{n^k} \approx 1, \qquad (1-\frac{a}{n})^k \approx 1$$

Do đó $P(X=k) \approx e^{-a}\frac{a^k}{k!}$

Vậy từ công thức Bernoulli ta có công thức xấp xỉ

$$P_k = P(X=k) = C_n^k p^k q^{n-k} \approx \frac{a^k}{k!}e^{-a}$$

Khi đó ta có thể thay công thức Bernoulli bởi công thức Poisson

$$\boxed{P_k = P(X=k) = \frac{a^k}{k!}e^{-a}} \tag{2.5}$$

□ **Định nghĩa 14** *Đại lượng ngẫu nhiên rời rạc X nhận một trong các giá trị 0,1,...,n với các xác suất tương ứng được tính theo công thức (2.5) được gọi là có phân phối Poisson với tham số a. Kí hiệu $X \in \mathcal{P}(a)$ (hay $X \sim \mathcal{P}(a)$).*

⊙ **Chú ý**

$$P(k \le X \le k+h) = P_k + P_{k+1} + \dots + P_{k+h} \quad \text{với} \quad P_k = \frac{a^k}{k!}e^{-a}.$$

• **Ví dụ 16** *Một máy dệt có 1000 ống sợi, Xác suất để một giờ máy hoạt động có 1 ống sợi bị đứt là 0,002. Tìm xác suất để trong một giờ máy hoạt động có không quá 2 ống sợi bị đứt.*

Giải

Việc quan sát một ống sợi có bị đứt hay không trong một giờ máy hoạt động là một phép thử. Máy dệt có 1000 ống sợi nên ta có n = 1000 phép thử độc lập.

Gọi A là biến cố ống sợi bị đứt và X là số ống sợi bị đứt trong một giờ máy hoạt động thì $p = P(A) = 0,002$ và $X \in B(1000; 0,002)$.

Vì $n = 1000$ khá lớn và $np = 2$ không đổi nên ta có thể xem $X \in \mathcal{P}(a)$.

Do đó xác suất để có không quá 2 ống sợi bị đứt trong một giờ là

$P(0 \le X \le 2) = P_0 + P_1 + P_2$

$P_0 = P(X=0) = \frac{2^0}{0!}e^{-2}$

$P_1 = P(X=1) = \frac{2^1}{1!}e^{-2}$

$P_2 = P(X=2) = \frac{2^2}{2!}e^{-2}$

Do đó $P(0 \le X \le 2) = (1+2+2)e^{-2} = 5(2,71)^{-2} = 0,6808.$

⊙ **Các tham số đặc trưng**

Nếu $X \in \mathcal{P}(a)$ thì $E(X) = Var(X) = a$ và $a - 1 \leq modX \leq a$.

Chứng minh. Để nhận được kỳ vọng và phương sai của đại lượng ngẫu nhiên có phân phối Poisson ta xác định hàm moment sinh

$$\psi(t) = E(e^{tX})$$

Ta có

$$\psi(t) = \sum_{k=0}^{\infty} e^{tk} e^{-a} \frac{a^k}{k!} = e^{-a} \sum_{k=0}^{\infty} \frac{(ae^t)^k}{k!} = e^{-a} e^{ae^t} = e^{a(e^t-1)}$$

$\psi'(t) = ae^t e^{a(e^t-1)}$

$\psi''(t) = (ae^t)^2 e^{a(e^t-1)} + ae^t e^{a(e^t-1)}$

Do đó

$$E(X) = \psi'(0) = a$$

$$Var(X) = \psi''(0) - [E(X)]^2 = a^2 + a - a^2 = a$$

□

⊙ **Ứng dụng**

Một vài đại lượng ngẫu nhiên có phân phối Poisson:

i) Số lỗi in sai trong một trang (hoặc một số trang) của một cuốn sách.

ii) Số người trong một cộng đồng sống cho tới 100 tuổi.

iii) Số cuộc điện thoại gọi sai trong một ngày.

iv) Số transitor hư trong ngày đầu tiên sử dụng.

v) Số khách hàng vào bưu điện trong một ngày.

vi) Số hạt α phát ra từ cát hạt phóng xạ trong một chu kỳ.

3.3 Phân phối siêu bội

a) Công thức siêu bội

Xét một tập hợp gồm N phần tử, trong đó có M phần tử có tính chất A nào đó. Lấy ngẫu nhiên (không hoàn lại) từ tập hợp ra n phần tử. Gọi X là số phần tử có tính chất A có trong n phần tử lấy ra. Ta có

$$\boxed{P_x = P(X = x) = \frac{C_M^x C_{N-M}^{n-x}}{C_N^n}} \quad (x = 0, 1, \ldots, n) \tag{2.6}$$

b) Phân phối siêu bội

□ **Định nghĩa 15** *Đại lượng ngẫu nhiên rời rạc X nhận một trong các giá trị 0,1,...,n với các xác suất tương ứng được tính theo công thức (2.6) được gọi là có phân phối siêu bội với tham số N, M, n. Kí hiệu* $X \in H(N,M,n)$ *(hay* $X \sim H(N,M,n)$*).*

• **Ví dụ 17** *Một lô hàng có 10 sản phẩm, trong đó có 6 sản phẩm tốt. Lấy ngẫu nhiên (không hoàn lại) từ lô hàng ra 4 sản phẩm. Tìm xác suất để có 3 sản phẩm tốt trong 4 sản phẩm được lấy ra.*

Giải

Gọi X là số sản phẩm tốt có trong 4 sản phẩm lấy ra thì X là đại lượng ngẫu nhiên có phân phối siêu bội với tham số $N = 10, M = 6, n = 4$.

Xác suất để có 3 sản phẩm tốt trong 4 sản phẩm lấy ra là

$$P(X=3) = \frac{C_6^3.C_4^1}{C_{10}^4} = \frac{8}{21} = 0,3809$$

⊙ **Chú ý**

Khi n khá bé so với N thì $\dfrac{C_M^x C_{N-M}^{n-x}}{C_N^n} \approx C_n^x p^x q^{n-x} \quad (p = \dfrac{M}{N},\ q = 1-p)$

Gọi X là số phần tử có tính chất A nào đó trong n phần tử lấy ra thì ta có thể xem $X \in B(n,p)$ với p là tỉ lệ phần tử có tính chất A của tập hợp.

c) Các tham số đặc trưng

Nếu $X \in H(N,M,n)$ thì ta có

$$E(X) = np \quad (\text{với } p = \frac{M}{N})$$

$$Var(X) = npq\frac{N-n}{N-1} \quad (\text{với } q = 1-p).$$

Bảng tổng kết các phân phối rời rạc

Phân phối	Kí hiệu	Xác suất $P(X=k)$	$E(X)$	$Var(X)$
Nhị thức	$B(n,p)$	$C_n^k p^k (1-p)^{n-k}$	np	npq
Poisson	$\mathcal{P}(a)$	$\dfrac{a^k}{k!}e^{-a}$	a	a
Siêu bội	$H(N,M,n)$	$\dfrac{C_M^k.C_{N-M}^{n-k}}{C_N^n}$	$np\ (p=\frac{M}{N})$	$npq\dfrac{N-n}{N-1}$

3.4 Phân phối mũ

□ **Định nghĩa 16** *Đại lượng ngẫu nhiên X được gọi là có phân phối mũ với tham số $\lambda > 0$ nếu nó có hàm mật độ xác suất*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{nếu } x > 0 \\ 0 & \text{nếu } x \le 0 \end{cases}$$

⊕ **Nhận xét** Nếu X có phân phối mũ với tham số λ thì hàm phân phối xác suất của X là

$F(x) = \int_0^x \lambda e^{-\lambda x} dt = 1 - e^{-\lambda x}$ với $x > 0$

và

$F(x) = 0$ với $x \le 0$.

⊙ **Các tham số đặc trưng**

Nếu X là đại lượng ngẫu nhiên có phân phối mũ với tham số $\lambda > 0$ thì

i) Kỳ vọng của X là

$$E(X) = \lambda \int_0^{+\infty} x e^{-\lambda x} dx = \left[-x e^{-\lambda x}\right]_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x} dx = \frac{1}{\lambda}$$

ii) Phương sai của X là

$$Var(X) = \int_0^{+\infty} x^2 \lambda e^{-\lambda x} dx - \frac{1}{\lambda^2}$$

Tích phân từng phần ta được $\int_0^{+\infty} x^2 \lambda e^{-\lambda x} dx = \left[-x^2 e^{-\lambda x}\right]_0^{+\infty} + 2 \int_0^{+\infty} \lambda x e^{-\lambda x} dx = \frac{2}{\lambda^2}$.

Do đó $Var(X) = \frac{1}{\lambda^2}$.

• **Ví dụ 18** *Giả sử tuổi thọ (tính bằng năm) của một mạch điện tử trong máy tính là một đại lượng ngẫu nhiên có phân phối mũ với kỳ vọng là 6,25. Thời gian bảo hành của mạch điện tử này là 5 năm.*

Hỏi có bao nhiêu phần trăm mạch điện tử bán ra phải thay thế trong thời gian bảo hành?

Giải

Gọi X là tuổi thọ của mạch. Thì X có phân phối mũ

Ta có $\lambda = \frac{1}{E(X)} = \frac{1}{6,25}$

$P(X \le 5) = F(5) = 1 - e^{-\lambda.5} = 1 - e^{-\frac{5}{6,25}} = 1 - e^{-0,8} = 1 - 0,449 = 0,5506$

Vậy có khoảng 55% số mạch điện tử bán ra phải thay thế trong thời gian bảo hành.

⊙ **Ứng dụng trong thực tế**

Khoảng thời gian giữa hai lần xuất hiện của một biến có phân phối mũ. Chẳng hạn khoảng thời gian giữa hai ca cấp cứu ở một bệnh viện, giữa hai lần hỏng hóc của một cái máy, giữa hai trận lụt hay động đất là nhữnng đại lượng ngẫu nhiên có phân phối mũ.

3.5 Phân phối đều

□ **Định nghĩa 17** *Đại lượng ngẫu nhiên liên tục X được gọi là có phân phối đều trên đoạn [a,b] nếu hàm mật độ xác suất có dạng*

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{nếu } x \in [a,b] \\ 0 & \text{nếu } x \notin [a,b] \end{cases}$$

⊕ **Nhận xét** Nếu X có phân phối đều trên [a,b] thì hàm phân phối của X cho bởi

$F(x) = 0$ nếu $x < a$

$F(x) = \int\limits_{-\infty}^{x} f(x)dx = \int\limits_{a}^{x} \dfrac{dx}{b-a} = \dfrac{x-a}{b-a}$ nếu $a \le x \le b$

$F(x) = 1$ nếu $x > b$.

⊙ **Chú ý** Giả sử $(\alpha, \beta) \subset [a,b]$. Xác suất để X rơi vào (α, β) là

$$P(\alpha < X < \beta) = \int\limits_{\alpha}^{\beta} f(x)dx = \frac{\beta - \alpha}{b-a}$$

⊙ **Các tham số đăc trưng**

i) $E(X) = \int\limits_{a}^{b} \dfrac{xdx}{b-a} = \dfrac{1}{b-a}\dfrac{b^2-a^2}{2} = \dfrac{a+b}{2}$ (kỳ vọng là trung điểm của [a,b]).

ii) $Var(X) = \int\limits_{a}^{b} \dfrac{x^2dx}{b-a} - [E(X)]^2 = \dfrac{1}{b-a}\left[\dfrac{x^3}{3}\right]_a^b - \left(\dfrac{a+b}{2}\right)$

$= \dfrac{b^2+ab+a^2}{3} - \dfrac{(a+b)^2}{4} = \dfrac{(b-a)^2}{12}$

iii) modX là bất cứ điểm nào trên [a,b].

• **Ví dụ 19** *Lịch chạy của xe buýt tại một trạm xe buýt như sau: chiếc xe buýt đầu tiên trong ngày sẽ khởi hành từ trạm này vào lúc 7 giờ, cứ sau mỗi 15 phút sẽ có một xe khác đến trạm. Giả sử một hành khách đến trạm trong khoảng thời gian từ 7 giờ đến 7 giờ 30. Tìm xác suất để hành khách này chờ*

a) Ít hơn 5 phút.

b) Ít nhất 12 phút.

Giải

Gọi X là số phút sau 7 giờ mà hành khách đến trạm thì X là đại lượng ngẫu nhiên có phân phối đều trong khoảng $(0, 30)$.

a) Hành khách sẽ chờ ít hơn 5 phút nếu đến trạm giữa 7 giờ 10 và 7 giờ 15 hoặc giữa 7 giờ 25 và 7 giờ 30. Do đó xác suất cần tìm là

$$P(10 < X < 15) + P(25 < X < 30) = \frac{5}{30} + \frac{5}{30} = \frac{1}{3}$$

b) Hành khách chờ ít nhất 12 phút nếu đến trạm giữa 7giờ và 7 giờ 3 phút hoặc giữa 7 giờ 15 phút và 7 giờ 18 phút. Xác suất cần tìm là

$$P(0 < X < 3) + P(15 < X < 18) = \frac{3}{30} + \frac{3}{30} = \frac{1}{5}$$

3.6 Phân phối chuẩn (Karl Gauss)

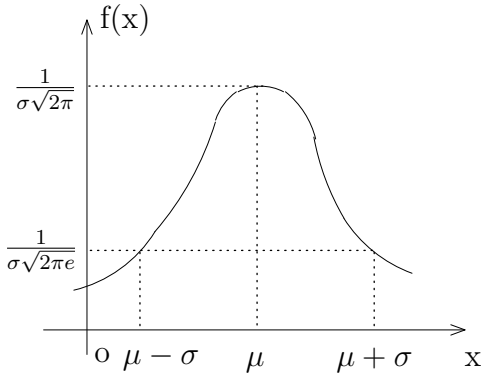
a) Phân phối chuẩn

□ **Định nghĩa 18**

Đại lượng ngẫu nhiên liên tục X nhận giá trị trong khoảng $(-\infty, +\infty)$ được gọi là có phân phối chuẩn nếu hàm mật độ xác suất có dạng

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

trong đó μ, σ là hằng số, $\sigma > 0$, $-\infty < x < \infty$.



Kí hiệu $X \in N(\mu, \sigma^2)$ hay $(X \sim N(\mu, \sigma^2))$.

b) Các tham số đặc trưng

Nếu $X \in N(\mu, \sigma^2)$ thì $E(X) = \mu$ và $Var(X) = \sigma^2$.

Chứng minh. Xét hàm moment sinh

$$\phi(t) = E(e^{tX}) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tx}.e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Đặt $y = \frac{x-\mu}{\sigma}$ thì

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{\mu t} \int_{-\infty}^{+\infty} e^{tx} e^{-\frac{y^2}{2}} dy = \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{y^2 - 2t\sigma y}{2}} dy$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(y-t\sigma)^2}{2} + \frac{t^2\sigma^2}{2}} dy = e^{\mu t + \frac{\sigma^2 t^2}{2}} \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(y-t\sigma)^2}{2}} dy$$

Vì $f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-t\sigma)^2}{2}}$ là hàm mật độ của phân phối chuẩn với tham số $t\sigma$ và 1 nên $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(y-t\sigma)^2}{2}} dy = 1$.

Do đó $\phi(t) = e^{\mu t + \frac{\sigma^2 + t^2}{2}}$.

Lấy các đạo hàm ta được

$$\phi'(t) = (\mu + t\sigma^2) e^{\mu t + \sigma^2 \frac{t^2}{2}}, \quad \phi''(t) = \sigma^2 e^{\mu t + \sigma^2 \frac{t^2}{2}} . (\mu + t\sigma^2)$$

Khi đó

$E(X) = \phi'(0) = \mu$

$E(X^2) = \phi''(0) = \sigma^2 + \mu^2 \quad \Longrightarrow \quad Var(X) = E(X^2) - [E(X)]^2 = \sigma^2$ □

c) Phân phối chuẩn hóa

□ **Định nghĩa 19** *Đại lượng ngẫu nhiên X được gọi là có phân phối chuẩn hóa nếu nó có phân phối chuẩn với $\mu = 0$ và $\sigma^2 = 1$. Kí hiệu $X \in N(0,1)$ hay $X \sim N(0,1)$.*

⊕ **Nhận xét** Nếu $X \in N(\mu, \sigma^2)$ thì $U = \frac{X - \mu}{\sigma} \in N(0,1)$.

d) Phân vị chuẩn

Phân vị chuẩn mức α, kí hiệu u_α, là giá trị của đại lượng ngẫu nhiên U có phân phối chuẩn hóa thỏa mãn điều kiện

$$P(U < u_\alpha) = \alpha.$$

Với α cho trước có thể tính được các giá trị của u_α. Các giá trị của u_α được tính sẵn thành bảng.

e) Công thức

Nếu $X \in N(\mu, \sigma^2)$ thì ta có

i) $P(x_1 \le X \le x_2) = \varphi(\frac{x_2-\mu}{\sigma}) - \varphi(\frac{x_1-\mu}{\sigma})$

ii) $P(|X-\mu| < \varepsilon) = 2\varphi(\frac{\varepsilon}{\sigma})$

trong đó $\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ (hàm Laplace).

• **Ví dụ 20** *Trọng lượng của một loại sản phẩm là đại lượng ngẫu nhiên có phân phối chuẩn với trọng lượng trung bình $\mu = 5kg$ và độ lệch tiêu chuẩn $\sigma = 0,1$. Tính tỉ lệ những sản phẩm có trọng lượng từ 4,9 kg đến 5,2 kg.*

Giải

Gọi X là trọng lượng của sản phẩm thì $X \in N(5; 0,1)$.

Tỉ lệ sản phẩm có trọng lượng tù 4,9 kg đến 5,2 kg là

$$\begin{aligned} P(4,9 \le X \le 5,2) &= \varphi(\tfrac{5,2-5}{0,1}) - \varphi(\tfrac{4,9-5}{0,1}) \\ &= \varphi(2) - \varphi(-1) \\ &= 0,4772 - (-0,3413) \\ &= 0,8185 \end{aligned}$$

f) Qui tăc "k−σ"

Trong công thức $P(|X-\mu| < \varepsilon) = 2\varphi(\frac{\varepsilon}{\sigma})$ nếu lấy $\varepsilon = k\sigma$ thì $P(|X-\mu| < \varepsilon) = 2\varphi(k)$.

Trong thực tế ta thường dùng qui tắc $1,96\sigma$, $2,58\sigma$ và 3σ với nội dung là:

"Nếu $X \in N(\mu, \sigma^2)$ thì xác suất để X nhận giá trị sai lệch so với kỳ vọng không quá $1,96\sigma; 2,58\sigma$ và 3σ là 95 %, 99% và 99% ".

g) Ứng dụng

Các đại lượng ngẫu nhiên sau có phân phối chuẩn:

- Kích thước chi tiết máy do máy sản suất ra.
- Trọng lượng của nhều sản phẩm cùng loại.
- Năng suất của một loại cây trồng trên những thửa ruộng khác nhau.

3.7 Phân phối χ^2

□ **Định nghĩa 20** *Giả sử X_i (i=1,2,...,n) là các đại lượng ngẫu nhiên độc lập cùng có phân phối chuẩn hóa.*

Đại lượng ngẫu nhiên $\chi^2 = \sum_{i=1}^{n} X_i^2$ *được gọi là có phân phối* χ^2 *(khi−bình phương) với* n *bậc tự do. Kí hiệu* $\chi^2 \in \chi^2(n)$ *(hay* $\chi^2 \sim \chi^2(n)$*).*

⊕ **Nhận xét**

Hàm mật độ xác suất của χ^2 có dạng

$$f_n(x) = \begin{cases} \dfrac{e^{-\frac{x}{2}}.x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})} & \text{với} \quad x > 0 \\ 0 & \text{với} \quad x \leq 0 \end{cases}$$

trong đó $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$ (Hàm Gamma)

Hàm mật độ xác suất của χ^2 với n bậc tự do

⊙ **Các tham số đặc trưng**

Nếu $\chi^2 \in \chi^2(n)$ thì $E(\chi^2) = n$ và $Var(\chi^2) = 2n$.

⊙ **Phân vị** χ^2

Phân vị χ^2 mức α, kí hiệu χ^2_α, là giá trị của đại lượng χ^2_α có phân phối "khi−bình phương" với n bậc tụ do thỏa mãn

$$P(\chi^2 < \chi^2_\alpha) = \alpha$$

Các giá trị của χ^2_α được tính sẵn thành bảng.

⊙ **Chú ý** Khi bậc n tăng lên thì phân phối χ^2 xấp xỉ với phân phối chuẩn.

3.8 Phân phối Student (G.S Gosset)

□ **Định nghĩa 21** *Giả sử* U *là đại lượng ngẫu nhiên có phân phối chuẩn hóa và* V *là đại lượng ngẫu nhiên độc lập với* U *có phân phối* χ^2 *với* n *bậc tự do. Khi đó đại lượng ngẫu nhiên*

$$T = \frac{U\sqrt{n}}{\sqrt{V}}$$

được gọi là có phân phối Student với n *bậc tự do. Kí hiệu* $T \in T(n)$ *(hay* $T \sim T(n)$*).*

⊕ **Nhận xét** Hàm mật độ của đại lượng ngẫu nhiên có phân phối Student với n bậc tụ do có dạng

$$f_n(t) = \frac{\Gamma(\frac{n+1}{2})(1+\frac{t^2}{n})^{-\frac{n+1}{2}}}{\Gamma(\frac{n}{2})\sqrt{n\pi}}; \quad (-\infty < t < +\infty)$$

trong đó $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$ (Hàm Gamma)

⊙ **Các tham số đặc trưng**

Nếu $T \in T(n)$ thì $E(T) = 0$ và $Var(T) = \dfrac{n}{n-2}$.

• **Phân vị Student**

Phân vị Student mức α, kí hiệu t_α là giá trị của đại lượng ngẫu nhiên $T \in T(n)$ thỏa mãn $P(T < t_\alpha) = \alpha$.

Ta có $t_\alpha = -t_{1-\alpha}$.

⊙ **Chú ý**

Phân phối Student có cùng dạng và tính đối xứng như phân phối chuẩn nhưng nó phản ánh tính biến đổi của phân phối sâu sắc hơn. Các biến có về giá và thời gian thường giới hạn một cách nghiêm ngặt kích thước của mẫu. Chính vì thế phân phối chuẩn không thể dùng để xấp xỉ phân phối khi mẫu có kích thước nhỏ. Trong trường hợp này ta dùng phân phối Student.

Khi bậc tự do n tăng lên ($n > 30$) thì phân phối Student tiến nhanh về phân phối chuẩn. Do đó khi $n > 30$ ta có thể dùng phân phối chuẩn thay cho phân phối Student.

3.9 Phân phối F (Fisher−Snedecor)

□ **Định nghĩa 22** *Nếu χ^2_n và χ^2_m là hai đại lượng ngẫu nhiên có phân phối "khi bình phương" với n và m bậc tự do thì đại lượng ngẫu nhiên $F_{n,m}$ xác định bởi*

$$F_{n,m} = \frac{\chi^2_n/n}{\chi^2_m/m}$$

được gọi là có phân phối F với n và m bậc tự do.

⊕ **Nhận xét** Hàm mật độ của phân phối F có dạng

$$p(x) = \begin{cases} 0 & ; \quad x \le 0 \\ \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2}).\Gamma(\frac{m}{2})}(\frac{n}{m})^{\frac{n}{2}} \frac{x^{\frac{n}{2}-1}}{(1+\frac{n}{m}x)^{\frac{n+m}{2}}} & ; \quad x > 0 \end{cases}$$

• **Các tham số đặc trưng**

$E(F_{n,m}) = \dfrac{m}{m-2}$ với $m > 2$

$Var(F_{n,m}) = \dfrac{m^2(2m+2n-4)}{n(m-2)^2(m-4)}$ với $m > 4$

3.10 Phân phối Gamma

□ **Định nghĩa 23** *Đại lượng ngẫu nhiên X được gọi là có phân phối Gamma với các tham số (α, λ), kí hiệu $X \in \gamma(\alpha, \lambda)$, nếu hàm mật độ xác suất có dạng*

$$f(x) = \begin{cases} \dfrac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & ; \quad x \geq 0 \\ 0 & ; \quad x < 0 \end{cases}$$

trong đó $\Gamma(\alpha) = \int\limits_0^\infty \lambda e^{-\lambda x}(\lambda x)^{\alpha-1}dx = \int\limits_0^\infty e^{-y}y^{\alpha-1}dy \quad (y = \lambda x).$

⊙ **Các tham số đặc trưng**

Nếu $X \in \gamma(\alpha, \lambda)$ thì $E(X) = \dfrac{\alpha}{\lambda}$ và $Var(X) = \dfrac{\alpha}{\lambda^2}$.

◇ **Tính chất** Nếu $X \in \gamma(\alpha, \lambda)$ và $Y \in \gamma(\beta, \lambda)$ thì $X + Y \in \gamma(\alpha + \beta, \lambda)$.

Bảng tổng kết các phân phối liên tục

Phân phối	Kí hiệu	Hàm mật độ $f(x)$	$E(X)$	$Var(X)$
Mũ		$\lambda e^{-\lambda x}$ $(x > 0)$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
Đều		$\dfrac{1}{b-a}$ $(a \leq x \leq b)$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Chuẩn	$N(\sigma^2, \mu)$	$\dfrac{1}{\sigma\sqrt{2\pi}} exp\left[-\dfrac{(x-\mu)^2}{2\sigma^2}\right]$	μ	σ^2
Khi bình phương	$\chi^2(n)$	$\dfrac{e^{-\frac{x}{2}}.x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}.\Gamma(\frac{n}{2})}$ $(x > 0, n > 0$	n	2n
Student	$T(n)$	$\dfrac{\Gamma(\frac{n+1}{2})(1+\frac{x^2}{n})^{-\frac{n+1}{2}}}{\Gamma(\frac{n}{2})\sqrt{n\pi}}$ $(n > 0)$	0 $(n > 1)$	$\dfrac{n}{n-2}$
Gamma	$\gamma(\alpha, \lambda)$	$\dfrac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}$	$\dfrac{\alpha}{\lambda}$	$\dfrac{\alpha}{\lambda^2}$

4. ĐẠI LƯỢNG NGẪU NHIÊN HAI CHIỀU

4.1 Khái niệm về đại lượng ngẫu nhiên hai chiều

Đại lượng ngẫu nhiên hai chiều là đại lượng ngẫu nhiên mà các giá trị có thể của nó được xác định bằng hai số. Kí hiệu (X, Y).

(X, Y được gọi là các thành phần của đại lượng ngẫu nhiên hai chiều)

Đại lượng ngẫu nhiên hai chiều được gọi là rời rạc (liên tục) nếu các thành phần của nó là các đại lượng ngẫu nhiên rời rạc (liên tục).

4.2 Phân phối xác suất của đại lượng ngẫu nhiên hai chiều

a) Bảng phân phối xác suất

X\Y	y_1	y_2	...	y_j	...	y_m
x_1	$P(x_1,y_1)$	$P(x_2,y_2)$	...	$P(x_1,y_j)$	...	$P(x_1,y_m)$
x_2	$P(x_2,y_1)$	$P(x_2,y_2)$	...	$P(x_2,y_j)$	...	$P(x_2,y_m)$
$\vdots$	...	...	...	...	...	...
x_i	$P(x_i,y_1)$	$P(x_i,y_2)$	...	$P(x_i,y_j)$	...	$P(x_i,y_m$
$\vdots$	...	...	...	...	...	...
x_n	$P(x_n,y_1)$	$P(x_n,y_2)$	...	$P(x_n,y_j)$	...	$P(x_n,y_m)$

trong đó

$x_i\,(i=\overline{1,n})$ là các giá trị có thể của thành phần X

$y_j\,(j=\overline{1,m})$ là các giá trị có thể của thành phần Y

$P(x_i,y_j)=P(\,(X,Y)=(x_i,y_j)\,)=P(X=x_i,Y=y_j),\quad i=\overline{1,n},j=\overline{1,m}$

$\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}P(x_i,y_j)=1$

b) Hàm mật độ xác suất

□ **Định nghĩa 24** *Hàm không âm, liên tục $f(x,y)$ được gọi là hàm mật độ xác suất của đại lượng ngẫu nhiên hai chiều (X,Y) nếu nó thỏa mãn*

$$P(X\in A,Y\in B)=\int\limits_A dx\int\limits_B f(x,y)dy$$

với A, B là các tập số thực.

c) Hàm phân phối xác suất

□ **Định nghĩa 25** *Hàm phân phối xác suất của đại lượng ngẫu nhiên hai chiều (X,Y), kí hiệu $F(x,y)$, là hàm được xác định như sau*

$$F(x,y)=P(X<x,Y<y)$$

⊙ **Nhận xét**

Ta có $F(x,y)=P(X<x,Y<y)=\int\limits_{-\infty}^{x}\left(\int\limits_{-\infty}^{y}f(x,y)dy\right)dx$ nên

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

4.3 Kỳ vọng và phưởng sai của các thành phần

i) Trường hợp (X,Y) rời rạc

$$E(X) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i P(x_i, y_j); \quad E(Y) = \sum_{j=1}^{m}\sum_{i=1}^{n} y_j P(x_i, y_j)$$

$$Var(X) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i^2 P(x_i, y_j) - [E(X)]^2, \quad Var(Y) = \sum_{j=1}^{m}\sum_{i=1}^{n} y_j^2 P(x_i, y_j) - [E(Y)]^2$$

ii) Trường hợp (X, Y) liên tục

$$E(X) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x f(x,y)dxdy, \quad E(Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y f(x,y)dxdy.$$

$$Var(X) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^2 f(x,y)dxdy - [E(X)]^2, \quad Var(Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y^2 f(x,y)dxdy - [E(Y)]^2$$

5. PHÂN PHỐI XÁC SUẤT CỦA HÀM CÁC ĐẠI LƯỢNG NGẪU NHIÊN

5.1 Hàm của một đại lượng ngẫu nhiên

□ **Định nghĩa 26** *Nếu mỗi giá trị có thể của đại lượng ngẫu nhiên X tương ứng với một giá trị có thể của đại lượng ngẫu nhiên Y thì Y được gọi là hàm của đại lượng ngẫu nhiên X. Kí hiệu* $Y = \varphi(X)$.

◇ **Tính chất**

i) Nếu X là đại lượng ngẫu nhiên rời rạc và $Y = \varphi(X)$ thì ứng với các giá trị khác nhau của X ta có các giá trị khác nhau của Y và có

$$P(Y = \varphi(x_i)) = P(X = x_i)$$

ii) Giả sử X là đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất $f(x)$ và $Y = \varphi(X)$.

Nếu $y = \varphi(x)$ là hàm khả vi, đơn điệu, có hàm ngược là $x = \psi(y)$ thì hàm mật độ xác suất $g(y)$ của đại lượng ngẫu nhiên Y được xác định bởi

$$g(y) = f(\psi(y)).\psi'(y)$$

• **Ví dụ 21** *Giả sử X là đại lượng ngẫu nhiên rời rạc có bảng phân phối xác suất*

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

Tìm qui luật phân phối xác suất của $Y = X^2$.

Giải

Các giá trị Y có thể nhận là $y_1 = 1^2 = 1$; $y_2 = 3^2 = 9$; $y_3 = 4^2 = 16$. Vậy phân phối xác suất của Y có thể cho bởi

Y	1	9	16
P	0,3	0,5	0,2

• **Các tham số**

i) Nếu X là đại lượng ngẫu nhiên rời rạc nhận một trong các giá trị $x_1, x_2, \ldots, x_n$ với các xác suất tương ứng $p_1, p_2, \ldots, p_n$ thì

$$E(Y) = E[\varphi(X)] = \sum_{i=1}^{n} \varphi(x_i)p_i$$

$$Var(Y) = Var[\varphi(X)] = \sum_{i=1}^{n} \varphi^2(x_i)p_i - [E(Y)]^2$$

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

$$E(Y) = E[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi(x)f(x)dx$$

$$Var(Y) = Var[\varphi(X)] = \int_{-\infty}^{+\infty} \varphi^2(x)f(x)dx - [E(Y)]^2$$

5.2 Hàm của đại lượng ngẫu nhiên hai chiều

□ **Định nghĩa 27** *Nếu mỗi cặp giá trị có thể các đại lượng X và Y tương ứng với một giá trị có thể của Z thì Z được gọi là hàm của hai đại lượng ngẫu nhiên X, Y. Kí hiệu* $Z = \varphi(X, Y)$.

⊙ **Chú ý** Việc xác định phân phối xác suất của $Z = \varphi(X, Y)$ thường rất phức tạp. Ta xét trường hợp đơn giản $Z = X + Y$ thông qua ví dụ dưới đây.

• **Ví dụ 22** *Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập có bảng phân phối xác suất*

X	*1*	*2*
P	*0,3*	*0,7*

Y	*3*	*4*
P	*0,2*	*0,8*

Tìm phân phối xác suất của $Z = X + Y$.

Giải

Các giá trị có thể của Z là tổng của một giá trị của X và một giá trị có thể của Y.

Do đó Z nhận các giá trị có thể

$z_1 = 1 + 3 = 4$; $z_2 = 1 + 4 = 5$; $z_3 = 2 + 3 = 5$; $z_4 = 2 + 4 = 6$

Các xác suất tương ứng là

$P(Z = 4) = P(X = 1).P(Y = 3) = 0,3 \times 0,2 = 0,06$

$P(Z = 5) = P(X = 1, Y = 4) + P(X = 2, Y = 3)$

$= P(X = 1).P(Y = 4) + P(X = 2).P(Y = 3)$

$= 0,3 \times 0,8 + 0,7 \times 0,2 = 0,38$

$P(Z = 6) = P(X = 2).P(Y = 4) = 0.7 \times 0,8 = 0,56$

Vậy Z có phân phối xác suất

Z	4	5	6
P	0,006	0,38	0,56

6. LUẬT SỐ LỚN

6.1 Bất đẳng thức Markov

Δ Định lý 1 *Nếu X là đại lượng ngẫu nhiên nhận giá trị không âm thì $\forall \varepsilon > 0$ ta có*

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Chứng minh. Ta chứng minh trong trường hợp X là đại lượng ngẫu nhiên liên tục có hàm mật độ $f(x)$.

$$E(X) = \int_0^{+\infty} xf(x)dx = \int_0^{a} xf(x)dx + \int_a^{+\infty} xf(x)dx$$

$$\geq \int_a^{+\infty} xf(x)dx \geq \int_a^{+\infty} af(x)dx = a\int_a^{+\infty} = aP(X \geq a).$$

□

6.2 Bất đẳng thức Tchebyshev

Δ Định lý 2 *Nếu X là đại lượng ngẫu nhiên có kỳ vọng μ và phương sai σ^2 hữu hạn thì $\forall \varepsilon > 0$ bé tùy ý ta có*

$$P(|X - \mu| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}$$

hay

$$P(|X - \mu| < \varepsilon) > 1 - \frac{Var(X)}{\varepsilon^2}$$

Chứng minh.

Ta thấy $(X - \mu)^2$ là đại lượng ngẫu nhiên nhận giá trị không âm.

Áp dụng bất đẳng thức Tchebyshev với $a = \varepsilon^2$ ta được

$$P[(X - \mu)^2 \geq \varepsilon^2] \leq \frac{E[(X - \mu)^2]}{\varepsilon^2} = \frac{Var(X)}{\varepsilon^2}$$

Vì $(X-\mu)^2 \geq \varepsilon^2$ khi và chỉ khi $|X-\mu| \geq \varepsilon$ nên

$$P(|X-\mu| \geq \varepsilon) \geq \frac{Var(X)}{\varepsilon^2}$$

□

⊙ **Chú ý** Bất đẳng thức Markov và Tchebuchev giúp ta phương tiện thấy được giới hạn của xác suất khi kỳ vọng và phương sai của phân phối xác suất chưa biết.

• **Ví dụ 23** *Giả sử số sản phẩm được sản xuất của một nhà máy trong một tuần là một đại lượng ngẫu nhiên với kỳ vọng $\mu = 50$.*

a) Có thể nói gì về xác suất sản phẩm của tuần này vượt quá 75.

b) Nếu phương sai của sản phẩm trong tuần này là $\sigma^2 = 25$ thì có thể nói gì về xác suất sản phẩm tuần này sẽ ở giữa 40 và 60.

Giải

a) Theo bất đẳng thức Markov

$$P(X > 75) \geq \frac{E(X)}{75} = \frac{50}{75} = \frac{2}{3}$$

b) Theo bất đẳng thức Tchebyshev

$$P(|X-50| \geq 10) \leq \frac{\sigma^2}{10^2} = \frac{25}{100} = \frac{1}{4}$$

Do đó

$$P(40 < X < 60) = P(|X-50| < 10) > 1 - \frac{1}{4} = \frac{3}{4}$$

6.3 Định lý Tchebyshev

Δ Định lý 3 (`Định lý Tchebyshev`) *Nếu các đại lượng ngẫu nhiên $X_1, X_2, \ldots, X_n$ độc lập từng đôi, có kỳ vọng hữu hạn và các phương sai đều bị chặn trên bởi số C thì $\forall \varepsilon > 0$ bétùy ý ta có*

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)\right| < \varepsilon)\right) = 1$$

Đặc biệt, khi $E(X_i) = a$; $(i = \overline{1,n})$ thì $\lim_{n\to\infty}(|\frac{1}{n}\sum_{i=1}^{n} X_i - a| < \varepsilon) = 1$

Chứng minh. Ta chứng minh trong trường hợp đặc biệt $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ $(i = 1, 2 \ldots, n)$. Ta có

$$E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \mu, \quad Var(\frac{1}{n}\sum_{i=1}^{n}) = \frac{\sigma^2}{n}$$

Theo bất đẳng thức Tchebyshev

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|\right)\leq\frac{\sigma^2}{n\varepsilon^2}$$

□

• Ý nghĩa

Mặc dù từng đại lượng ngẫu nhiên độc lập có thể nhận giá trị sai khác nhiều so với kỳ vọng của chúng, nhưng trung bình số học của một số lớn đại lượng ngẫu nhiên lại nhận giá trị gần bằng trung bình số học của các kỳ vọng của chúng. Điều này cho phép ta dự đoán giá trị trung bình số học của các đại lượng ngẫu nhiên.

6.4 Định lý Bernoulli

Δ Định lý 4 (Định lý Bernoulli) *Nếu f_n là tần suất xuất hiện biến cố A trong n phép thử độc lập và p là xác suất xuất hiện biến cố A trong mỗi phép thử thì $\forall\varepsilon>0$ bé tùy ý ta có*

$$\lim_{n\to\infty}P(|f_n-p|<\varepsilon)=1$$

• Ý nghĩa

Tần suất xuất hiện biến cố trong n phép thử độc lập dần về xác suất xuất hiện biến cố trong mỗi phép thử khi số phép thử tăng lên vô hạn.

7. BÀI TẬP

1. Một nhóm có 10 người gồm 6 nam và 4 nữ. Chọn ngẫu nhiên ra 3 người. Gọi X là số nữ ở trong nhóm. Lập bảng phân phối xác suất của X và tính $E(X), Var(X), mod(X)$.

2. Gieo đồng thời hai con xúc sắc cân đối đồng chất. Gọi X là tổng số nốt xuất hiện trên hai mặt con xúc sắc. lập bảng qui luật phân phối xác suất của X. Tính $E(X)$ và $Var(X)$.

3. Trong một cái hộp có 5 bóng đèn trong đó có 2 bóng tốt và 3 bóng hỏng. Chọn ngẫu nhiên từng bóng đem thử (thử xong không trả lại) cho đến khi thu được 2 bóng tốt. Gọi X là số lần thử cần thiết. Tìm phân phối xác suất của X. Trung bình cần thử bao nhiêu lần?

4. Một đợt xổ số phát hành N vé. Trong đó có m_i vé trúng k_i đồng một vé ($i=1,2,\ldots,n$). Hỏi giá của mỗi vé số là bao nhiêu để cho trung bình của tiền thưởng cho mỗi vé bằng một nửa giá tiền của một vé?

5. Tuổi thọ của một loại côn trùng nào đó là một đại lượng ngẫu nhiên liên tục X (đơn vị là tháng) có hàm mật độ

$$f(x) = \begin{cases} kx^2(4-x) & \text{nếu } 0 \le x \le 4 \\ 0 & \text{nếu ngược lại} \end{cases}$$

a) Tìm hằng số k.
b) Tìm $mod(X)$.
c) Tính xác suất để côn trùng chết trước khi nó được 1 tháng tuổi.

6. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ

$$f(x) = \begin{cases} kx^2e^{-2x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

a) Tìm hằng số k.
b) Tìm hàm phân phối của X.
c) Tìm $mod(X)$.
d) Tìm $E(X)$ và $Var(X)$.

7. Một xí nghiệp sản xuất máy tính có xác suất làm ra phế phẩm là 0,02. Chọn ngẫu nhiên 250 máy tính để kiểm tra. Tìm xác suất để:
a) Có đúng 2 phế phẩm.
b) Có không quá 2 phế phẩm.

8. `(Bài toán Samuel−Pepys)` Pepys đã đưa ra bài toán sau cho Newton: *Biến cố nào trong các biến có sau đây có xác suất lớn nhất?*

a) Có ít nhất một lần xuất hiện mặt 6 khi tung một con xúc xắc 6 lần.
b) Có ít nhất 2 lần xuất hiện mặt 6 khi tung con xúc xắc 12 lần.
c) Có ít nhất 3 lần xuất hiện mặt 6 khi tung con xúc xắc 18 lần.

9. Xác suất một người bị phản ứng tù việc tiêm huyết thanh là 0,001. Tìm xác suất sao cho trong 2000 người có đúng 3 người, có nhiều hơn 2 người bị phản ứng.

10. Một lô hàng có 500 sản phẩm (trong đó có 400 sản phẩm loại A). Lấy ngẫu nhiên tù lô hàng đó ra 200 sản phẩm để kiểm tra. Gọi X là số sản phẩm loại A có trong 200 sản phẩm lấy ra kiểm tra. Tìm kỳ vọng và phương sai của X.

11. Một trung tâm bưu điện nhận được trung bình 300 lần gọi điện thoại trong một giờ. Tìm xác suất để trung tâm này nhận được đúng 2 lần gọi trong 1 phút.

12. Trọng lượng của một con bò là một đại lượng ngẫu nhiên có phân phối chuẩn với giá trị trung bình $250kg$ và độ lệch tiêu chuẩn là $40kg$. Tìm xác suất để một con bò chọn ngẫu nhiên có trọng lượng:

a) Nặng hơn $300kg$.
b) Nhẹ hơn $175kg$.
c) Nằm trong khoảng tù $260kg$ đến $270kg$.

13. Chiều cao của 300 sinh viên là một đại lượng ngẫu nhiên có phân phối chuẩn với trung bình $172cm$ và độ lệch tiêu chuẩn $8cm$. Có bao nhiêu sinh viên có chiều cao:

    a) lớn hơn $184cm$,
    b) nhỏ hơn hoặc bằng $160cm$,
    c) giữa $164cm$ và $180cm$,
    d) bằng $172cm$.

14. Cho hai đại lượng ngẫu nhiên độc lập X, Y có bảng phân phối xác suất như sau:

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

Y	2	4
P	0,4	0,6

    Tìm phân phối xác suất của $Z = X + Y$.

15. Cho đại lượng ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau:

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

    Tìm kỳ vọng và phương sai của đại lượng ngẫu nhiên $Y = \varphi(X) = X^2 + 1$.

16. Gieo một con xúc xắc cân đối n lần. Gọi X là số lần xuất hiện mặt lục. Chứng minh rằng

$$P(\frac{n}{6} - \sqrt{n} < X < \frac{n}{6} + \sqrt{n}) \geq \frac{31}{36}$$

◙ TRẢ LỜI BÀI TẬP

1.

X	0	1	2	3
P	$\frac{5}{30}$	$\frac{15}{30}$	$\frac{9}{30}$	$\frac{1}{30}$

$E(X) = 1,2$, $Var(X) = 0,56$, $mod(X) = 1$.

2.

X	2	3	4	5	6	7	8	9	10	11	12
P	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$E(X) = 7$, $Var(X) = 5,833$.

3. $P(X = 2) = \frac{2}{5}.\frac{1}{4} = \frac{1}{10}$.

$P(X = 3) = \frac{3}{5}.\frac{2}{4}.\frac{1}{3} + \frac{2}{5}.\frac{3}{4}.\frac{1}{3} = \frac{2}{10}$.

$P(X = 4) = \frac{3}{5}.\frac{2}{4}.\frac{2}{3}.\frac{1}{2} + \frac{3}{5}.\frac{2}{4}.\frac{2}{3}.\frac{1}{2} + \frac{2}{5}.\frac{3}{4}.\frac{2}{3}.\frac{1}{2} = \frac{3}{10}$.

$P(X = 5) = 1 - (\frac{2}{20} + \frac{4}{20} + \frac{6}{20}) = \frac{4}{10}$.

Trung bình cần $E(X) = 4$ lần thử.

4. $\frac{2}{N}\sum_{i=1}^{n} k_i m_i$.

5. a) Vì $\int_0^4 x^2(4-x)dx = \frac{64}{3}$ suy ra $k = \frac{3}{64}$, b) $mod(X) = \frac{8}{3}$,

c) $P(X<1)=\frac{3}{64}\int\limits_0^1 x^2(4-x)dx=\frac{13}{256}$.

6. a) $k=4$, b) $F(x)=\begin{cases}1-e^{-2x}(2x^2+2x+1) & \text{nếu } x>0\\ 0 & \text{nếu } x<0\end{cases}$

c) $mod(X)=1$, d) $E(X)=\frac{3}{2}$, $Var(X)=\frac{3}{4}$.

7. $X\in B(250,2\%)$ a) $P(X=2)=0,0842$, b) $P(x\le 2)=0,1247$.

8. a) $P=0,665$, b) $P=0,619$, c) $P=0,597$.

9. $P(X=x)=\frac{a^x.e^{-a}}{x!}$ với $a=np=(2000).(0,001)=2$.

$P(X=3)=0,18$, $P(X>2)=0,323$.

10. $E(X)=160$, $Var(X)=19,238$. **11.** $P=0,09$.

12. a) $P(X>300)=1-\phi(1,25)==0,1056$,

b) $P(X,175)==\phi(-1,875)=0,0303$,

c) $P(260<X<270)=\phi(0,5)-\phi(0,25)=0,0928$.

13. a) 18, b) 22, c) 213, d) 14.

14.

Z	3	4	5	6	7
P	0,08	0,12	0,32	0,18	0,3

15. $E(Y)=13,2$, $Var(Y)=79,36$.

16. X có phân phối nhị thức với $P=\frac{1}{6}$ nên $E(X)=\frac{n}{6}$. Áp dụng bất đẳng thức Tchebyshev ta được bất đẳng thức cần chứng minh.

Chủởng 3

TỔNG THỂ VÀ MẪU

1. TỔNG THỂ VÀ MẪU

1.1 Tổng thể

Khi nghiên cứu về một vấn đề người ta thuờng khảo sát trên một dấu hiệu nào đó, các dấu hiệu này thể hiện trên nhiều phần tủ. Tập hợp các phần tủ mang dấu hiệu đuợc gọi là *tổng thể* hay *đám đông* (*population*).

• **Ví dụ 1** *Nghiên cứu tập hợp gà trong một trại chăn nuôi ta quan tâm đến dấu hiệu trọng lượng. Nghiên cứu chất lượng học tập của sinh viên trong một trường đại học ta quan tâm đến dấu hiệu điểm.*

⊙ **Chú ý** Trong phần này ta sủ dụng một số khái niệm và kí hiệu sau:

1. N: số phần tủ của tổng thể, đuợc gọi là kích thuớc của tổng thể.
2. X^*: dấu hiệu mà ta khảo sát.
3. $x_i\,(i=\overline{1,k})$: giá trị của dấu hiệu X^* đo đuợc trên phần tủ của tổng thể (x_i là thông tin mà ta quan tâm, còn các phần tủ của tổng thể là vật mang thông tin).
4. $N_i\,(i=\overline{1,k})$: tần số của x_i (số phần tủ có chung giá trị x_i).
5. $p_i=\frac{N_i}{N}$: tần suất của x_i.

⊙ **Bảng cơ cấu của tổng thể**

Sụ tuởng ứng giữa các giá trị x_i và tần suất p_i đuợc biểu diễn bởi bảng cơ cấu tổng thể theo dấu hiệu X^* nhu sau:

Giá trị của X^*	x_1	x_2	...	x_k
Tần suất p_i	p_1	p_2	...	p_k

• **Các đặc trưng của tổng thể**

1. Trung bình của dấu hiệu X^* (trung bình của tổng thể) $m = \sum_{i=1}^{k} x_i p_i$.

2. Phương sai của dấu hiệu X^* (phương sai của tổng thể) $\sigma^2 = \sum_{i=1}^{k} (x_i - m)^2 p_i$.

3. Độ lệch tiêu chuẩn của dấu hiệu X^* (độ lệch tiêu chuẩn của tổng thể)

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{k} (x_i - m)^2 p_i}$$

1.2 Mẫu

- Từ tổng thể lấy ra n phần tử và đo lường dấu hiệu X^* trên chúng. Khi đó n phần tử này lập nên một mẫu (*sample*). Số phần tử của mẫu được gọi là *kích thước của mẫu*.
- Vì từ mẫu suy ra kết luận cho tổng thể nên mẫu phải đại diện cho tổng thể và phải được chọn một cách khách quan.
- Việc lấy mẫu được tiến hành theo hai phương thức: lấy mẫu có hoàn lại và lấy mẫu không hoàn lại.

2. MÔ HÌNH XÁC SUẤT CỦA TỔNG THỂ VÀ MẪU

2.1 Đại lượng ngẫu nhiên gốc và phân phối gốc

Lấy tùy ý từ tổng thể ra một phần tử. Gọi X là giá trị của X^* đo được trên phần tử lấy ra thì X là đại lượng ngẫu nhiên có phân phối xác suất

X	x_1	x_2	...	x_i	...	x_k
P	p_1	p_2	...	p_i	...	p_k

Ta thấy dấu hiệu X^* được mô hình hóa bởi đại lượng ngẫu nhiên X. Khi đó X được gọi là đại lượng ngẫu nhiên gốc và phân phối xác suất của X được gọi là phân phối gốc.

2.2 Các tham số của đại lượng ngẫu nhiên gốc

$$E(X) = \sum_{i=1}^{k} x_i p_i.$$

$$Var(X) = \sum_{i=1}^{k} [x_i - E(X)]^2 p_i$$

2.3 Mẫu ngẫu nhiên

Lấy n phần tử của tổng thể theo phương pháp hoàn lại để quan sát. Gọi X_i là giá trị của X^* đo được trên phần tử thứ i $(i = \overline{1,n})$ thì $X_1, X_2, \ldots, X_n$ là các đại lượng ngẫu nhiên độc lập có cùng phân phối như X. Khi đó bộ $(X_1, X_2, \ldots, X_n)$ được gọi là một *mẫu ngẫu nhiên* kích thước n được tạo nên từ đại lượng ngẫu nhiên gốc X. Kí hiệu $W_X = (X_1, X_2, \ldots, X_n)$.

Giả sử X_i nhận giá trị x_i $(i = \overline{1,n})$. Khi đó $(x_1, x_2, \ldots, x_n)$ là một giá trị cụ thể của mẫu ngẫu nhiên W_X, được gọi là *mẫu cụ thể*. Kí hiệu $w_x = (x_1, x_2, \ldots, x_n)$.

• **Ví dụ 2** *Kết quả điểm môn Toán của một lớp gồm 100 sinh viên cho bởi bảng sau*

Điểm	3	4	5	6	7
Số sinh viên có điểm tương ứng	25	20	40	10	5

Gọi X là điểm môn Toán của một sinh viên được chọn ngẫu nhiên trong danh sách lớp thì X là đại lượng ngẫu nhiên có phân phối

X	3	4	5	6	7
P	0,25	0,2	0,4	0,1	0,05

Chọn ngẫu nhiên 5 sinh viên trong danh sách lớp để xem điểm. Gọi X_i là điểm của sinh viên thứ i. Ta có mẫu ngẫu nhiên kích thước $n = 5$ được xây dựng từ đại lượng ngẫu nhiên X

$$W_X = (X_1, X_2, \ldots, X_n)$$

Giả sử sinh viên thứ nhất được 4 điểm, thứ hai được 3 điểm, thứ ba được 6 điểm thứ tư được 7 điểm và thứ năm được 5 điểm. Ta được mẫu cụ thể

$$w_x = (4, 3, 6, 7, 5)$$

3. THỐNG KÊ

Trong thống kê (*statistics*), việc tổng hợp mẫu $W_X = (X_1, X_2, \ldots, X_n)$ được thực hiện dưới dạng hàm $G = f(X_1, X_2, \ldots, X_n)$ của các đại lượng ngẫu nhiên $X_1, X_2, \ldots, X_n$. Khi đó G được gọi là một thống kê.

3.1 Trung bình mẫu ngẫu nhiên

□ **Định nghĩa 1** *Trung bình của mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ là một thống kê, kí hiệu $\overline{X}$, được xác định bởi*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.1}$$

⊙ **Chú ý**

i) Vì $X_1, X_2, \ldots, X_n$ là các đại lượng ngẫu nhiên nên $\overline{X}$ cũng là đại lượng ngẫu nhiên.

ii) Nếu mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ có mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$ thì $\overline{X}$ sẽ nhận giá trị $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ và $\overline{x}$ được gọi là trung bình của mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$.

◇ **Tính chất**

Nếu đại lượng ngẫu nhiên gốc X có kỳ vọng $E(X) = m$ và phương sai $Var(X) = \sigma^2$ thì $E(\overline{X}) = m$ và $Var(X) = \dfrac{\sigma^2}{n}$.

⊙ **Phân phối xác suất của $\overline{X}$**

i) Nếu $X \in B(n,p)$ thì $\overline{X} \in B(n,p)$.

ii) Nếu $X \in \mathcal{P}(a)$ thì $\overline{X} \in \mathcal{P}(a)$.

iii) Nếu $X \in N(\mu, \sigma^2)$ thì $\overline{X} \in N(\mu, \frac{\sigma^2}{n})$.

iv) Nếu $X \in \chi^2(n)$ thì $\overline{X} \in \chi^2(n)$.

3.2 Phương sai của mẫu ngẫu nhiên

□ **Định nghĩa 2** *Phương sai của mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ là một thống kê, kí hiệu S^2, được xác định bởi*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

trong đó $\overline{X}$ là trung bình của mẫu ngẫu nhiên.

⊙ **Chú ý**

i) Vì $X_1, X_2, \ldots, X_n$ là các đại lượng ngẫu nhiên nên S^2 cũng là đại lượng ngẫu nhiên.

ii) Nếu mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ có mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$ thì S^2 nhận giá trị $s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Khi đó s^2 được gọi là phương sai của mẫu cụ thể.

◇ **Tính chất** Nếu $Var(X) = \sigma^2$ thì $E(S^2) = \dfrac{n-1}{n}\sigma^2$.

⊙ **Phương sai điều chỉnh**

Đặt $S'^2 = \dfrac{n}{n-1}S^2$ thì ta có $E(S'^2) = \sigma^2$.

S'^2 được gọi là *phương sai điều chỉnh* của mẫu ngẫu nhiên W_X.

Với mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$ thì S'^2 sẽ nhận giá trị

$$s'^2 = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

s'^2 được gọi là *phương sai điều chỉnh* của mẫu cụ thể.

⊙ **Phân phối xác suất**

Giả sử $W_X = (X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên được xây dựng từ đại lượng ngẫu nhiên X có phân phối chuẩn với $E(X) = m$ và $Var(X) = \sigma^2$. Khi đó

i) $\dfrac{nS^2}{\sigma^2} = \displaystyle\sum_{i=1}^{n}\frac{(X_i - \overline{X})^2}{\sigma^2} \in \chi^2(n-1)$.

ii) $\displaystyle\sum_{i=1}^{n}\frac{(X_i - m)^2}{\sigma^2} \in \chi^2(n)$

3.3 Độ lệch tiêu chuẩn và độ lệch tiêu chuẩn điều chỉnh

i) Độ lệch tiêu chuẩn của mẫu ngẫu nhiên W_X là $S = \sqrt{S^2}$.

Độ lệch tiêu chuẩn của mẫu cụ thể w_x là $s = \sqrt{s^2}$, trong đó s là giá trị của S.

ii) Độ lệch tiêu chuẩn điều chỉnh của mẫu ngẫu nhiên W_X là $S' = \sqrt{S'^2}$.

Độ lệch tiêu chuẩn điều chỉnh của mẫu cụ thể w_x là $s' = \sqrt{s'^2}$, trong đó s' là giá trị của S'.

4. SẮP XẾP SỐ LIỆU

Quá trình nghiên cứu thống kê thường trải qua 2 khâu: thu thập các số liệu liên quan đến việc nghiên cứu và xử lý số liệu. Để việc xử lý được thuận lợi ta cần phải sắp xếp lại số liệu.

4.1 Trường hợp mẫu có kích thước nhỏ

Giả sử mẫu có kích thước n và đại lượng ngẫu nhiên gốc X nhận các giá trị có thể x_i $(i = \overline{1,k})$ với số lần lặp lại (tần số) n_i $(i = \overline{1,k})$. Ta thường lập bảng như sau:

x_i	n_i
x_i	n_1
x_2	n_2
$\ldots$	$\ldots$
x_k	n_k

Chú ý $\displaystyle\sum_{i=1}^{k} n_i = n$.

• **Ví dụ 3** *Tiến hành thu thập dữ liệu số trẻ ở lứa tuổi đến trường của 30 gia đình ở một huyện ta được kết quả cho bởi bảng*

0	3	0	0	3	0
2	2	0	1	2	1
0	0	1	2	4	0
4	2	1	0	1	0
0	2	0	1	3	2

Sắp xếp số liệu lại ta có bảng sau

Số trẻ ở lứa tuổi đến trường	n_i
0	12
1	6
2	7
3	3
4	2

4.2 Trường hợp mẫu có kích thước lớn

Ta chia mẫu thành các khoảng (lớp), trong mỗi khoảng ta chọn một giá trị đại diện. Người ta thường chia thành các khoảng đều nhau (có thể khoảng đầu hoặc cuối có độ dài khác với độ dài của các khoảng còn lại) và chọn giá trị đại diện là giá trị trung tâm của khoảng. Ta qui ước đầu mút bên phải của mỗi khoảng thuộc khoảng đó mà không thuộc khoảng tiếp theo khi tính tần số của mỗi khoảng.

• **Ví dụ 4** *Chiều cao của 400 cây sao được chia thành các khoảng được xếp trong bảng sau:*

Khoảng chiều cao	*Tần số* n_i	*Độ dài của khoảng*
5,5 – 8,5	*18*	*3*
8,5 – 12,5	*58*	*4*
12,5 – 16,5	*62*	*4*
16,5 – 20,5	*72*	*4*
20,5 – 24,5	*57*	*4*
24,5 – 28,5	*42*	*4*
28,5 – 32,5	*36*	*4*
32,5 – 36,5	*10*	*4*

5. BẢNG TÍNH $\overline{x}$, s^2

5.1 Tính trực tiếp

Ta dùng công thức

$$\boxed{\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i \\ s^2 &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - (\overline{x})^2 \end{aligned}} \tag{3.2}$$

trong đó x_i $(i = \overline{1,k})$ là các giá trị của X^*.

• **Ví dụ 5** *Số xe hơi bán được trung bình trong một tuần ở mỗi đại lý trong 45 đại lý cho bởi*

Số xe hơi được bán trong tuần / đại lý	n_i
1	15
2	12
3	9
4	5
5	3
6	1

Ta lập bảng tính như sau

x_i	n_i	$n_i x_i$	$n_i x_i^2$
1	15	15	15
2	12	24	48
3	9	27	81
4	5	20	80
5	3	15	75
6	1	6	36
$\sum$	$n = 45$	107	335

Ta có

$$\overline{x} = \frac{107}{45} = 2,38$$
$$s^2 = \frac{335}{45} - (2,38)^2 = 7,444 - 5,664 = 1,78.$$

• **Ví dụ 6** *Theo dõi 336 trường hợp tàu cập cảng, người ta thấy khoảng thời gian ngắn nhất giữa hai lần tàu vào cảng liên tiếp là 4 giờ, thời gian dài nhất là 80 giờ.*

Vì số liệu nhiều nên ta sắp xếp thành lớp có độ dài 8 và thay mỗi lớp bởi giá trị trung tâm $x_i^0 = \dfrac{x_{min} + x_{max}}{2}$.

Ta có bảng tính sau

$x_i - x_{i+1}$	x_i^0	n_i	$n_i x_i^0$	$n_i {x_i^0}^2$
4 – 12	8	143	1144	9152
12 – 20	16	75	1200	19200
20 – 28	24	53	1272	30528
28 – 36	32	27	864	27648
36 – 44	40	14	560	22400
44 – 52	48	9	432	20736
52 – 60	56	5	280	15680
60 – 68	64	4	256	16384
68 – 76	72	3	216	15552
76 – 80	78	3	234	18252
$\sum$		336	6458	195532

Ta có

$\overline{x} = \frac{6458}{336} = 19,22$

$s^2 = \frac{195532}{336} - (19,22)^2 = 212,532.$

5.2 Tính theo phương pháp đổi biến

Ta dùng phương pháp này khi x_i hoặc giá trị trung tâm x_i^0 của khoảng khá lớn.

Đặt $\boxed{u_i = \frac{x_i - x_0}{h}}$

trong đó x_i là giá trị của dấu hiệu X^*; x_0 và h là những giá trị tùy ý.

Ta thường chọn x_0 là giá trị x_i (hoặc x_i^0) ứng với tần số lớn nhất và h là độ dài của khoảng.

Khi đó

$$\boxed{\begin{array}{rcl} \overline{x} & = & x_0 + h\overline{u} \\ s^2 & = & h^2\left[\frac{1}{n}\sum_{i=1}^{k} n_i u_i^2 - (\overline{u})^2\right] \end{array}}$$

• **Ví dụ 7** *Tính $\overline{x}$ và s^2 từ số liệu cho ở bảng của ví dụ trước.*

Ta chọn

$x_0 = 8$ (ứng với tần số $n_i = 143$ lớn nhất)
$h = 8$ (độ dài của lớp)

$x_i - x_{i+1}$	x_i^0	n_i	u_i	$n_i u_i$	$n_i u_i^2$
4 – 12	8	143	0	0	0
12 – 20	16	75	1	75	75
20 – 28	24	53	2	106	212
28 – 36	32	27	3	81	243
36 – 44	40	14	4	56	224
44 – 52	48	9	5	45	225
52 – 60	56	5	6	30	180
60 – 68	64	4	7	28	196
68 – 76	72	3	8	24	192
76 – 80	78	3	8,75	26,25	229,6875
$\sum$		336		471,25	1176,6875

Áp dụng công thức ta có

$\overline{x} = 8.\frac{471,25}{336} + 8 = 19,22$

$s^2 = 8^2.[\frac{1776,6875}{336} - (\frac{471,25}{336})^2] = 212,5229$

6. BÀI TẬP

1. Chiều cao của 40 sinh viên nam ở một trường đại học cho bởi bảng dưới đây. Hãy sắp xếp các số liệu trên thành bảng bằng cách chia số liệu thành các khoảng thích hợp.

52	68	60	48	55	45	59	61
57	64	54	55	49	58	60	66
70	48	52	73	67	51	62	69
56	73	53	57	51	61	54	59
66	57	49	64	60	70	73	67

2. Theo dõi năng suất của 100 hecta lúa ở một vùng, người ta thu được kết quả cho ở bảng sau:

Năng suất *(tạ/ha)*	Diện tích *(ha)*
30 – 35	7
35 – 40	12
40 – 45	18
45 – 50	27
50 – 55	20
55 – 60	8
60 – 65	5
65 – 70	3

Tính giá trị trung bình, phương sai và phương sai điều chỉnh của mẫu cụ thể này.

3. Quan sát về thời gian cần thiết để sản xuất một chi tiết máy ta thu được các số liệu cho ở bảng sau:

Khoảng thời gian *(phút)*	Số quan sát
20 – 25	2
25 – 30	14
30 – 35	26
35 – 40	32
40 – 45	14
45 – 50	8
50 – 55	4

Tính giá trị trung bình, phương sai và phương sai điều chỉnh của mẫu.

4. Thống kê số hàng bán được trong một ngày và số ngày bán được số lượng hàng tương ứng, ta có bảng số liệu sau:

Lượng hàng bán trong 1 ngày kg	Số ngày (n_i)
$100 - 200$	5
$200 - 250$	12
$250 - 300$	56
$300 - 350$	107
$350 - 400$	75
$400 - 450$	70
$450 - 500$	35
$500 - 550$	30
$550 - 700$	10

Tính giá trị trung bình mẫu và nêu ý nghĩa của nó.

◙ TRẢ LỜI BÀI TẬP

2. $\overline{x} = 47,5$ tạ/ha, $s^2 = 68,5$, $s'^2 = 69,192$.

3. $\overline{x} = 36,6$ phút, $s^2 = 44,69$, $s'^2 = 45,14$.

4. $\overline{x} = 375,3kg$

Chương 4

ƯỚC LƯỢNG THAM SỐ CỦA ĐẠI LƯỢNG NGẪU NHIÊN

Giả sử đại lượng ngẫu nhiên X có tham số θ chưa biết. Ước lượng tham số θ là dựa vào mẫu ngẫu nhiên $W_x = (X_1, X_2, \ldots, X_n)$ ta đưa ra thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ để ước lượng (dự đoán) θ.

Có 2 phương pháp ước lượng:

i) Ước lượng điểm: chỉ ra $\theta = \theta_0$ nào đó để ước lượng θ.

ii) Ước lượng khoảng: chỉ ra một khoảng (θ_1, θ_2) chứa θ sao cho $P(\theta_1 < \theta < \theta_2) = 1 - \alpha$ cho trước ($1 - \alpha$ gọi là độ tin cậy của ước lượng).

1. CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG ĐIỂM

1.1 Phương pháp hàm ước lượng

• Mô tả phương pháp

Giả sử cần ước lượng tham số θ của đại lượng ngẫu nhiên X. Từ X ta lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$.

Chọn thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$. Ta gọi $\hat{\theta}$ là *hàm ước lượng* của X.

Thực hiện phép thử ta được mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$. Khi đó ước lượng điểm của θ là giá trị $\theta_0 = \hat{\theta}(x_1, x_2, \ldots, x_n)$.

a) Ước lượng không chệch

□ **Định nghĩa 1** *Thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ được gọi là ước lượng không chệch của tham số θ nếu $E(\hat{\theta}) = \theta$.*

⊙ **Ý nghĩa**

Giả sử $\hat{\theta}$ là ước lượng không chệch của tham số θ. Ta có

$$E(\hat{\theta} - \theta) = E(\hat{\theta}) - E(\theta) = \theta - \theta = 0$$

Vậu uốc lượng không chệch là uốc lượng có sai số trung bình bằng 0.

⊕ **Nhận xét**

i) Trung bình của mẫu ngẫu nhiên $\overline{X}$ là uốc lượng không chệch của trung bình của tổng thể $\theta = E(X) = m$ vì $E(\overline{X}) = m$.

ii) Phương sai điều chỉnh của mẫu ngẫu nhiên S'^2 là uốc lượng không chệch của phương sai của tổng thể σ^2 vì $E(S'^2) = \sigma^2$.

• **Ví dụ 1** *Chiều cao của 50 cây lim được cho bởi*

Khoảng chiều cao (mét)	*số cây lim*	x_i^0	u_i	$n_i u_i$	$n_i u_i^2$
$[6,25 - 6,75)$	1	6,5	-4	-4	16
$[6,75 - 7,25)$	2	7,0	-3	-6	18
$[7,25 - 7,75)$	5	7,5	-2	-10	20
$[7,75 - 8,25)$	11	8	-1	-11	11
$[8,25 - 8,75)$	18	8,5	0	0	0
$[8,75 - 9,25)$	9	9	1	9	9
$[9,25 - 9,75)$	3	9,5	2	6	12
$[9,75 - 10,2)$	1	10	3	3	9
$\sum$	50			-13	95

Gọi X là chiều cao của cây lim

a) Hãy chỉ ra uốc lượng điểm cho chiều cao trung bình của các cây lim.

b) Hãy chỉ ra uốc lượng điểm cho độ tản mát của các chiều cao cây lim so với chiều cao trung bình.

c) Gọi $p = P(7,75 \leq X \leq 8,75)$. Hãy chỉ ra uốc lượng điểm cho p.

Giải

Ta lập bảng tính cho $\overline{x}$ và s^2.

Thực hiện phép đổi biến $u_i = \dfrac{x_i^0 - 8,5}{0,5}$ $\qquad (x_0 = 8,5;\ h = 0,5)$

Ta có $\overline{u} = -\frac{13}{50} = -0,26$. Suy ra

$\overline{x} = 8,5 + 0,5.(-0,26) = 8,37$

$s^2 = (0,5)^2.\left[\dfrac{95}{50} - (-0,26)^2\right] = 0,4581 \sim (0,68)^2.$

a) Chiều cao trung bình được uốc lượng là 8,37 mét.

b) Độ tản mát được uốc lượng là $s = 0,68$ mét hoặc $\hat{s} = \sqrt{\frac{50}{50-1}0,4581} \sim 0,684$

c) Trong 50 quan sát đã cho có $11+18 = 29$ quan sát cho chiều cao lim thuộc khoảng $[7,5 - 8,5)$

Vậy uốc lượng điểm cho p là $p^* = \frac{29}{50} = 0,58$.

b) Ước lượng hiệu quả

⊕ **Nhận xét** Giả sử $\hat{\theta}$ là ước lượng không chệch của tham số θ. Theo bất đẳng thức Tchebychev ta có

$$P(|\hat{\theta} - E(\hat{\theta})| < \varepsilon) > 1 - \frac{Var(\hat{\theta})}{\varepsilon^2}$$

Vì $E(\hat{\theta}) = \theta$ nên $P(|\hat{\theta} - \theta| < \varepsilon) > 1 - \frac{Var(\hat{\theta})}{\varepsilon^2}$.

Ta thấy nếu $Var(\hat{\theta})$ càng nhỏ thì $P(|\hat{\theta} - \theta| < \varepsilon)$ càng gần 1. Do đó ta sẽ chọn $\hat{\theta}$ với $Var(\hat{\theta})$ nhỏ nhất.

□ **Định nghĩa 2** *Ước lượng không chệch $\hat{\theta}$ được gọi là ước lượng có hiệu quả của tham số θ nếu $Var(\hat{\theta})$ nhỏ nhất trong các ước lượng của θ.*

⊙ **Chú ý** Người ta chứng minh được rằng nếu $\hat{\theta}$ là ước lượng hiệu quả của θ thì phương sai của nó là

$$Var(\hat{\theta}) = \frac{1}{n.E(\frac{\partial lnf(x,\theta)}{\partial \theta})^2} \tag{4.1}$$

trong đó $f(x, \theta)$ là hàm mật độ xác suất của đại lượng ngẫu nhiên gốc. Mọi ước lượng không chệch θ luôn có phương sai lớn hơn $Var(\hat{\theta})$ trong (4.1). Ta gọi (4.1) là *giới hạn Crame-Rao.*

⊕ **Nhận xét** Nếu đại lượng ngẫu nhiên gốc $X \in N(\mu, \frac{\sigma^2}{n})$ thì trung bình mẫu $\overline{X}$ là ước lượng hiệu quả của kỳ vọng $E(X) = \mu$.

Thật vậy, ta biết $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \in N(\mu, \frac{\sigma^2}{n})$

Mặt khác do X có phân phối chuẩn nên nếu $f(x, \mu)$ là hàm mật độ của X_i thì

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

Ta có $\frac{\partial}{\partial \mu} lnf(x, \mu) = \frac{x - \mu}{\sigma^2}$.

Suy ra $nE\left[\frac{\partial lnf(x,\mu)}{\partial \mu}\right]^2 = nE\left(\frac{x-\mu}{\sigma^2}\right)^2 = \frac{n}{\sigma^2}$. Do đó $Var(\overline{X})$ chính bằng nghịch đảo σ^2/n.

Vậy $\overline{X}$ là ước lượng hiệu quả của μ.

c) Ước lượng vững

□ **Định nghĩa 3** *Thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ được gọi là ước lượng vững của tham số θ nếu $\forall \varepsilon > 0$ ta có*

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$

⊙ Điều kiện đủ của ước lượng vững

Nếu $\hat{\theta}$ là ước lượng không chệch của θ và $\lim\limits_{n\to\infty} Var(\hat{\theta}) = 0$ thì $\hat{\theta}$ là ước lượng vững của θ.

1.2 Phương pháp ước lượng hợp lý tối đa

Giả sử $W_X = (X_1, X_2, \ldots, X_n)$ là mẫu ngẫu nhiên được tạo nên từ đại lượng ngẫu nhiên X có mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$ và $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$.

Xét hàm hàm hợp lý $L(x_1, \ldots, x_n, \theta)$ của đối số θ xác định như sau:

- Nếu X rời rạc:

$$\begin{aligned} L(x_1, \ldots, x_n, \theta) &= P(X_1 = x_1/\theta, \ldots, X_n = x_n/\theta) && (4.2)\\ &= \prod_{i=1}^{n} P(X_i = x_i/\theta) && (4.3)\end{aligned}$$

  $L(x_1, \ldots, x_n, \theta)$ là xác suất để ta nhận được mẫu cụ thể $W_x = (x_1, \ldots, x_n)$

- Nếu X liên tục có hàm mật độ xác suất $f(x, \theta)$

$$L(x_1, \ldots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta)$$

  $L(x_1, x_2, \ldots, x_n, \theta)$ là mật độ của xác suất tại điểm $w_x(x_1, x_2, \ldots, x_n)$

  Giá trị $\theta_0 = \hat{\theta}(x_1, x_2, \ldots, x_n)$ được gọi là ước lượng hợp lý tối đa nếu ứng với giá trị này của θ hàm hợp lý đạt cực đại.

⊙ Phương pháp tìm

Vì hàm L và lnL đạt cực đại tại cùng một giá trị θ nên ta xét lnL thay vì xét L.

Bước 1: Tìm $\dfrac{\partial lnL}{\partial \theta}$

Bước 2: Giải phương trình $\dfrac{\partial lnL}{\partial \theta}$ (Phương trình hợp lý)

Giả sử phương trình có nghiệm là $\theta_0 = \hat{\theta}(x_1, x_2, \ldots, x_n)$

Bước 3: Tìm đạo hàm cấp hai $\dfrac{\partial^2 lnL}{\partial \theta}$

Nếu tại θ_0 mà $\dfrac{\partial^2 lnL}{\partial \theta} < 0$ thì lnL đạt cực đại. Khi đó $\theta_0 = \hat{\theta}(x_1, x_2, \ldots, x_n)$ là ước lượng điểm hợp lý tối đa của θ.

2. PHƯƠNG PHÁP KHOẢNG TIN CẬY

2.1 Mô tả phương pháp

Giả sử tổng thể có tham số θ chưa biết. Ta tìm khoảng (θ_1, θ_2) chứa θ sao cho $P(\theta_1 < \theta < \theta_2) = 1 - \alpha$ cho trước.

Từ đại lượng ngẫu nhiên gốc X lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$. Chọn thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ có phân phối xác suất xác định dù chưa biết θ.

Với α_1 khá bé ($\alpha_1 < \alpha$) ta tìm được phân vị θ_{α_1} của $\hat{\theta}$ (tức là $P(\hat{\theta} < \theta_{\alpha_1}) = \alpha_1$).

Với α_2 mà $\alpha_1 + \alpha_2 = \alpha$ khá bé (thường lấy $\alpha \leq 0,05$) ta tìm được phân vị $\theta_{1-\alpha_2}$ của $\hat{\theta}$ (tức là $P(\hat{\theta} < \theta_{1-\alpha_2}) = 1 - \alpha_2$).

Khi đó

$$P(\theta_{\alpha_1} \leq \hat{\theta} \leq \theta_{1-\alpha_2}) = P(\hat{\theta} < \theta_{1-\alpha_2}) - P(\hat{\theta} < \theta_{\alpha_1}) = 1 - \alpha_2 - \alpha_1 = 1 - \alpha \qquad (*)$$

Từ (*) ta giải ra được θ. Khi đó (*) được đưa về dạng $P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha$.

Vì xác suất $1 - \alpha$ gần bằng 1, nên biến cố $(\hat{\theta}_1 < \theta < \hat{\theta}_2)$ hầu như xảy ra. Thực hiện một phép thử đối với mẫu ngẫu nhiên W_X ta thu được mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$. Từ mẫu cụ thể này ta tính được giá trị $\theta_1 = \hat{\theta}_1(x_1, x_2, \ldots, x_n)$, $\theta_2 = \hat{\theta}_2(x_1, x_2, \ldots, x_n)$.

Vậy với $1 - \alpha$ cho trước, qua mẫu cụ thể w_x ta tìm được khoảng (θ_1, θ_2) chứa θ sao cho $P(\theta_1 < \theta < \theta_2) = 1 - \alpha$.

- Khoảng (θ_1, θ_2) được gọi là khoảng tin cậy.
- $1 - \alpha$ được gọi là độ tin cậy của ước lượng.
- $|\theta_2 - \theta_1|$ được gọi là độ dài khoảng tin cậy.

2.2 Ước lượng trung bình

Giả sử trung bình của tổng thể $E(X) = m$ chưa biết. Ta tìm khoảng (m_1, m_2) chứa m sao cho $P(m_1 < m < m_2) = 1 - \alpha$, với $1 - \alpha$ là độ tin cậy cho trước.

i) Trường hợp 1

$$\begin{cases} \text{Biết } Var(X) = \sigma^2 \\ n \geq 30 \quad \text{hoặc} \quad (n < 30 \text{ nhưng X có phân phối chuẩn}) \end{cases}$$

Chọn thống kê

$$U = \frac{(\overline{X} - m)\sqrt{n}}{\sigma} \qquad (4.4)$$

Ta thấy $U \in N(0, 1)$.

Chọn cặp α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tìm các phân vị

$$P(U < u_{\alpha_1}) = \alpha_1, \quad P(U < u_{\alpha_2}) = 1 - \alpha_2$$

Do phân vị chuẩn có tính chất $u_{\alpha_1} = -u_{1-\alpha_1}$ nên

$$P(-u_{1-\alpha_1} < U < u_{1-\alpha_2}) = 1 - \alpha \tag{4.5}$$

Dựa vào (4.4) và giải hệ bất phương trình trong (4.5) ta được

$$\overline{X} - \frac{\sigma}{\sqrt{n}} u_{1-\alpha_2} < m < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha_1}$$

Để được khoảng tin cậy đối xứng ta chọn $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ và đặt $\gamma = 1 - \frac{\alpha}{2}$ thì

$$\overline{X} - \frac{\sigma}{\sqrt{n}} u_{\gamma} < m < \overline{X} + \frac{\sigma}{\sqrt{n}} u_{\gamma}$$

Tóm lại, ta tìm được khoảng tin cậy $(\overline{x} - \varepsilon, \overline{x} + \varepsilon)$, trong đó

* $\overline{x}$ là trung bình của mẫu ngẫu nhiên.

* $\boxed{\varepsilon = u_\gamma \frac{\sigma}{\sqrt{n}}}$ (độ chính xác) với u_γ là phân vị chuẩn mức $\gamma = 1 - \frac{\alpha}{2}$

• **Ví dụ 2** *Khối lượng sản phẩm là đại lượng ngẫu nhiên X có phân phối chuẩn với độ lệch tiêu chuẩn $\sigma = 1$. Cân thử 25 sản phẩm ta thu được kết quả sau*

X (khối lượng)	*18*	*19*	*20*	*21*
n_i (số lượng	*3*	*5*	*15*	*2*

Hãy ước lượng trung bình khối lượng của sản phẩm với độ tin cậy 95 %.

Giải

x_i	n_i	$x_i n_i$
18	3	54
19	5	95
20	15	300
21	2	42
$\sum$	25	491

Ta có $\overline{x} = \frac{491}{25} = 19,64kg$.

Độ tin cậy $1 - \alpha = 0,95 \implies \alpha = 0,025 \implies \gamma = 1 - \frac{\alpha}{2} = 0,975$ Ta tìm được phân vị chuẩn $u_\gamma = u_{0,975} = 1,96$. Do đó

$$\varepsilon = u_{0,975} \frac{1}{\sqrt{25}} = 1,96.\frac{1}{5} = 0.39$$

$$x_1 = \overline{x} - \varepsilon = 19,6 - 0,39 = 19,25$$

$$x_2 = \overline{x} + \varepsilon = 19,6 + 0,39 = 20,03$$

Vậy khoảng tin cậy là $(19,25; 20,03)$.

ii) Trường hợp 2

$\begin{cases} \sigma^2 \text{ chưa biết} \\ n \geq 30 \end{cases}$

Trường hợp này kích thước mẫu lớn ($n \geq 30$) có thể dùng ước lượng của S'^2 thay cho σ^2 chưa biết ($E(S'^2) = \sigma^2$), ta tìm được khoảng tin cậy $(\overline{x} - \varepsilon, \overline{x} + \varepsilon)$ trong đó

* $\overline{x}$ là trung bình của mẫu cụ thể.

* $\boxed{\varepsilon = u_\gamma \frac{s'}{\sqrt{n}}}$ với u_γ là phân vị chuẩn mức $\gamma = 1 - \frac{\alpha}{2}$ và s' là độ lệch tiêu chuẩn điều chỉnh của mẫu cụ thể.

• **Ví dụ 3** *Người ta tiến hành nghiên cứu ở một trường đại học xem trong một tháng trung bình một sinh viên tiêu hết bao nhiêu tiền gọi điện thoại. Lấy một mẫu ngẫu nhiên gồm 59 sinh viên thu được kết quả sau:*

14	*18*	*22*	*30*	*36*	*28*	*42*	*79*	*36*	*52*	*15*	*47*
95	*16*	*27*	*111*	*37*	*63*	*127*	*23*	*31*	*70*	*27*	*11*
30	*147*	*72*	*37*	*25*	*7*	*33*	*29*	*35*	*41*	*48*	*15*
29	*73*	*26*	*15*	*26*	*31*	*57*	*40*	*18*	*85*	*28*	*32*
22	*36*	*60*	*41*	*35*	*26*	*20*	*58*	*33*	*23*	*35*	

Hãy ước lượng khoảng tin cậy 95% cho số tiền gọi điện thoại trung bình hàng tháng của một sinh viên.

Giải

Từ các số liệu đã cho, ta có

$$n = 59; \quad \overline{x} = 41,05; \quad s' = 27,99$$

Độ tin cậy $1 - \alpha = 0,95 \implies 1 - \frac{\alpha}{2} = 0,975$. Tra bảng phân vị chuẩn ta có $u_{0,975} = 1,96$.

Do đó $\varepsilon = 1,96.\frac{27,99}{\sqrt{59}} = 7,13$.

$$\overline{x} - 7,13 = 33,92; \quad \overline{x} + 7,13 = 48,18$$

Vậy khoảng tin cậy của ước lượng là (33,92; 48,18).

iii) Trường hợp 3

$\begin{cases} \sigma^2 \text{ chưa biết} \\ n < 30 \text{ và X có phân phối chuẩn} \end{cases}$

Chọn thống kê $T = \dfrac{(\overline{X} - m)\sqrt{n}}{S'} \quad \in \quad T(n-1)$.

Ta tìm được khoảng tin cậy $(\overline{x}-\varepsilon, \overline{x}+\varepsilon)$ trong đó $\boxed{\varepsilon = t_\gamma \frac{S'}{\sqrt{n}}}$

với t_γ là phân vị Student mức $\gamma = 1 - \frac{\alpha}{2}$ với $n-1$ bậc tự do và s' là độ lệch tiêu chuẩn điều chỉnh của mẫu cụ thể.

• **Ví dụ 4** *Dioxide Sulfur và Oxide Nitrogen là các hóa chất được khai thác từ lòng đất. Các chất này được gió mang đi rất xa, kết hợp thành acid và rơi trở lại mặt đất tạo thành mưa acid. Người ta đo độ đậm đặc của Dioxide Sulfur* $(\mu g/m^3)$ *trong khu rừng Bavarian của nước Đức. Số liệu cho bởi bảng dưới đây:*

52,7	*43,9*	*41,7*	*71,5*	*47,6*	*55,1*
62,2	*56,5*	*33,4*	*61,8*	*54,3*	*50,0*
45,3	*63,4*	*53,9*	*65,5*	*66,6*	*70,0*
52,4	*38,6*	*46,1*	*44,4*	*60,7*	*56,4*

Hãy ước lượng độ đậm đặc trung bình của Dioxide Sulsfur với độ tin cậy 95%.

Giải

Ta tính được $\overline{x} = 53,92 \mu g/m^3, \quad s' = 10,07 \mu g/m^3$.

Độ tin cậy $1-\alpha = 0,95 \implies \alpha = 0,025 \implies 1-\frac{\alpha}{2} = 0,975$. Tra bảng phân vị student mức 0,975 bậc $n-1 = 23$ ta được $t_{23;0,975} = 2,069$.

Do đó $\varepsilon = 2,069\frac{10,07}{\sqrt{24}} = 4,25$.

$$\overline{x} - \varepsilon = 53,92 - 4,25 = 49,67, \quad \overline{x} + \varepsilon = 53,92 + 4,25 = 58,17$$

Vậy khoảng tin cậy là (49,67; 58,17).

Người ta biết được nếu độ đậm đặc của Dioxide Sulfur trong một khu vực lớn hơn $20 \mu g/m^3$ thì môi trường trong khu vực bị phá hoại bởi mưa acid. Qua ví dụ này các nhà khoa học đã tìm ra được nguyên nhân rừng Bavarian bị phá hoại trầm trọng năm 1983 là do mưa acid .

⊙ **Chú ý** (Xác định kích thước mẫu)

Nếu muốn độ tin cậy $1-\alpha$ và độ chính xác ε đạt ở mức cho trước thì ta cần xác định kích thước n của mẫu.

i) *Trường hợp biết* $Var(X) = \sigma^2$:

Từ công thức $\varepsilon = u_\gamma^2 \frac{\sigma}{\sqrt{n}}$ ta suy ra

$$\boxed{n = u_\gamma^2 \frac{\sigma^2}{\varepsilon^2}}$$

ii) *Trường hợp chưa biết* σ^2:

Dựa và mẫu cụ thể đã cho (nếu chưa có mẫu thì ta có thể tiến hành lấy mẫu lần đầu với kích thước $n_1 \geq 30$) để tính s'^2. Từ đó xác định được

$$\boxed{n = u_\gamma^2 \frac{s'^2}{\varepsilon^2}}$$

Kích thước mẫu n phải là số nguyên. Nếu khi tính n theo các công thức trên được giá trị không nguyên thì ta lấy phần nguyên của nó cộng thêm với 1.

Tức là $n = \left[u_\gamma^2 \frac{\sigma^2}{\varepsilon^2}\right] + 1$ hoặc $n = \left[u_\gamma^2 \frac{s'^2}{\varepsilon^2}\right] + 1$.

2.3 Ước lượng tỷ lệ

Giả sử tổng thể được chia ra làm hai loại phần tử. Tỷ lệ phần tử có tính chất A là p chưa biết. Ước lượng tỷ lệ là chỉ ra khoảng (f_1, f_2) chứa p sao cho $P(f_1 < p < f_2) = 1-\alpha$.

Để cho việc giải bài toán được đơn giản, ta chọn mẫu với kích thước n khá lớn.

Gọi X là số phần tử có tính chất A khi lấy ngẫu nhiên một phần tử từ tổng thể thì X là đại lượng ngẫu nhiên có phân phối xác suất

X	0	1
P	1-p	p

Gọi X_i $(i = \overline{1,n})$ là số phần tử có tính chất A trong lần lấy thứ i.

Ta có $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ chính là tần suất ước lượng điểm của $p = E(X)$. Mặt khác, theo chương 2, $n\overline{X}$ có phân phối nhị thức $B(n,p)$. Từ đó $E(\overline{X}) = p$ và $Var(\overline{X}) = \frac{p(1-p)}{n}$.

Chọn thống kê $U = \frac{(f-p)\sqrt{n}}{\sqrt{p(1-p)}}$, trong đó f là tỷ lệ các phần tử của mẫu có tính chất A.

Khi n khá lớn thì $U \in N(0,1)$. Giải quyết bài toán tương tự như ở ước lượng trung bình, thay $\overline{X}$ bởi f, σ^2 bởi $f(1-f)$... ta được

$$f - u_\gamma\sqrt{\frac{f(1-f)}{n}} < p < f + u_\gamma\sqrt{\frac{f(1-f)}{n}}$$

Tóm lại, ta xác định được khoảng tin cậy $(f_1, f_2) = (f-\varepsilon, f+\varepsilon)$, trong đó

f là tỷ lệ các phần tử của mẫu có tính chất A

$$\boxed{\varepsilon = u_\gamma\sqrt{\frac{f(1-f)}{n}}} \quad \text{(độ chính xác)} \tag{4.6}$$

với u_γ là phân vị chuẩn mức $1-\frac{\alpha}{2}$.

Từ (4.6) ta có

$$\boxed{u_\gamma = \frac{\varepsilon\sqrt{n}}{\sqrt{f(1-f)}}}$$

$$\boxed{n = u^2_{1-\frac{\alpha}{2}}\frac{f(1-f)}{\varepsilon^2}}$$

⊙ **Chú ý** Ta có thể tìm khoảng tin cậy của p bằng cách khác như sau:

Từ khoảng tin cậy của p:

$$\left(f - u_\gamma\sqrt{\frac{p(1-p)}{n}} < p < f + u_\gamma\sqrt{\frac{p(1-p)}{n}}\right) \quad \text{hay} \quad \left(|f-p| < u_\gamma\sqrt{\frac{p(1-p)}{n}}\right)$$

Giải bất phương trìnhnày ta tìm được

$$p_1 = \frac{nf + 0,5u^2_\gamma - \sqrt{0,25u^2_\gamma - nf(1-f)}}{n+u^2_\gamma}, \quad p_2 = \frac{nf + 0,5u^2_\gamma + \sqrt{0,25u^2_\gamma - nf(1-f)}}{n+u^2_\gamma}$$

Khi đó (p_1, p_2) là khoảng tin cậy của p với độ tin cậy $1-\alpha$.

• **Ví dụ 5** *Kiểm tra 100 sản phẩm trong lô hàng thấy có 20 phế phẩm.*

i) Hãy ước lượng tỷ lệ phế phẩm có độ tin cậy 99 %.

ii) Nếu độ chính xác $\varepsilon = 0,04$ thì độ tin cậy của ước lượng là bao nhiêu?

iii) Nếu muốn có độ tin cậy 99% và độ chính xác 0,04 thì phải kiểm tra bao nhiêu sản phẩm?

Giải

i) $n = 100, \quad f = \frac{20}{100} = 0.2$

Xét $U = \frac{(f-p)\sqrt{100}}{\sqrt{pq}} \in N(0,1).$

Ta có

$$1-\alpha = 0,99 \implies \alpha = 0,01 \implies 1-\frac{\alpha}{2} = 1-0,005 = 0,995$$

$$\varepsilon = u_{0,995}\frac{\sqrt{0,2.0,8}}{\sqrt{100}} = 2,58.\frac{0,4}{10} = 0,1$$

$$f_1 = f - \varepsilon = 0,2 - 0,1 = 0,1$$

$$f_2 = f + \varepsilon = 0,2 + 0,1 = 0,3$$

Vậy khoảng tin cậy là $(0,1; 0,3)$.

ii) $u_{1-\frac{\alpha}{2}} = \dfrac{0,04.\sqrt{100}}{\sqrt{0,2.0,8}} = 1$

Tìm được

$$1 - \frac{\alpha}{2} = 0,84 \quad \Longrightarrow \quad 1 - \alpha = 0,68$$

Vậy độ tin cậy là 68%.

iii)$1-\alpha = 0,99 \quad \Longrightarrow \quad \alpha = 0,01 \quad \Longrightarrow \quad 1-\frac{\alpha}{2} = 0,995$. Tìm được $u_{0,995} = 2,576$.

Do đó

$$n \approx \frac{(2,576)^2.0,2.0,8}{(0,04)^2} = 6,635.100 = 663,5$$

Vậy $n = 664$

2.4 Ước lượng phương sai

Giả sử đại lượng ngẫu nhiên X có phân phối chuẩn với phương sai $Var(X) = \sigma^2$ chưa biết. Cho $0 < \alpha < 0.05$. Ước lượng phương sai $Var(X)$ là chỉ ra khoảng (σ_1^2, σ_2^2) chứa σ^2 sao cho $P(\sigma_1^2 < \sigma^2 < \sigma_2^2) = 1 - \alpha$.

Từ X lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ và xét các trường hợp

a) Biết $E(X) = \mu$.

Chọn thống kê $\chi^2 = \sum_{i=1}^{n} \dfrac{(X_i - \mu)^2}{\sigma^2}$

Ta thấy χ^2 có phân phối "khi-bình phương" với n bậc tự do.

Chọn α_1 và α_2 khá bé sao cho $\alpha_1 + \alpha_2 = \alpha$. Ta tìm được các phân vị $\chi^2_{\alpha_1}$ và $\chi^2_{1-\alpha_2}$ thỏa mãn

$$P(\chi^2_{\alpha_1} < \chi^2 < \chi^2_{1-\alpha_2}) = 1 - \alpha \tag{4.7}$$

Thay biểu thức của χ^2 vào (4.7) và giải ra ta được

$$\frac{\sum(X_i - \mu)^2}{\chi^2_{1-\alpha_2}} < \sigma^2 < \frac{\sum(X_i - \mu)^2}{\chi^2_{\alpha_1}}$$

Chọn $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ thì

$$\frac{\sum(X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{\sum(X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2}}} \tag{4.8}$$

Với mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$, tính các tổng $\sum(x_i - \mu)^2$ và dựa vào (4.8) ta tìm được khoảng tin cậy (σ_1^2, σ_2^2), trong đó

$$\boxed{\begin{aligned}\sigma_1^2 &= \frac{\sum(x_i-\mu)^2 n_i}{\chi^2_{n,1-\frac{\alpha}{2}}}\\ \sigma_2^2 &= \frac{\sum(x_i-\mu)^2 n_i}{\chi^2_{n,\frac{\alpha}{2}}}\end{aligned}}$$

với

$\chi^2_{n,1-\frac{\alpha}{2}}$ là phân vị "khi−bình phương" mức $1-\frac{\alpha}{2}$ với n bậc tự do.

$\chi^2_{n,\frac{\alpha}{2}}$ là phân vị "khi−bình phương" mức $\frac{\alpha}{2}$ với n bậc tự do.

b) Chưa biết $E(X)$.

Chọn thống kê $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2}$

Thống kê này có phân phối "khi−bình phương với $n-1$ bậc tự do. Tương tự như trên ta tìm được khoảng tin cậy (σ_1^2, σ_2^2) với

$$\boxed{\sigma_1^2 = \frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}};\quad \sigma_2^2 = \frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}}$$

• **Ví dụ 6** *Mức hao phí nhiên liệu cho một đơn vị sản phẩm là đại lượng ngẫu nhiên có phân phối chuẩn. Xét trên 25 sản phẩm ta thu được kết quả sau:*

X	*19,5*	*20*	*20,5*
n_i	*5*	*18*	*2*

Hãy ước lượng phương sai với độ tin cậy 90 % trong các trường hợp sau:

i) Biết kỳ vọng $\mu = 20g$.

ii) Chưa biết kỳ vọng.

Giải

i) Biết $\mu = 20g$.

x_i	n_i	$x_i - 20$	$(x_i-20)^2$	$(x_i-20)^2 n_i$
19,5	5	-0,5	0,25	1,25
20	18	0	0	0
20,5	2	0,5	0,25	0,5
$\sum$	n=25			1,75

Độ tin cậy $1-\alpha = 0,9 \implies \alpha = 0,1 \implies \dfrac{\alpha}{2} = 0,05 \implies 1-\dfrac{\alpha}{2} = 0.95$

Tra bảng phân vị χ^2 với $n = 25$ bậc tự do ta được

$$\chi^2_{25;0,05} = 14,6; \qquad \chi^2_{25;0,95} = 37,7$$

Do đó

$$\sigma_1^2 = \frac{\sum(x_i - 20)^2 n_i}{\chi^2_{25;0,95}} = \frac{1,75}{37,7} = 0,046$$

$$\sigma_2^2 = \frac{\sum(x_i - 20)^2 n_i}{\chi^2_{25;0,05}} = \frac{1,75}{14,6} = 0,12$$

Vậy khoảng tin cậy là $(0,046;\ 0,12)$.

ii) Khi chưa biết kỳ vọng ta tìm $s'^2 = 0,0692$.

Tra bảng phân vị khi bình phương với bậc tự do $n - 1 = 24$.

$$\chi^2_{0,05} = 13,85; \quad \chi^2_{0,95} = 36,4$$

và tính

$$\sigma_1^2 = \frac{24s'^2}{\chi^2_{0,95}} = \frac{24 \times 0,0692}{36,4} = 0,046$$

$$\sigma_2^2 = \frac{24s'^2}{\chi^2_{0,05}} = \frac{24 \times 0,0692}{13,85} = 0,12$$

Vậy khoảng tin cậy là $(0,046;\ 0,12)$.

3. BÀI TẬP

1. Một mẫu các trọng lượng tương ứng là 8,3; 10,6; 9,7; 8,8; 10,2 và 9,4 kg. Xác định ước lượng không chệch của

   a) trung bình của tổng thể,
   b) phương sai của tổng thể.

2. Một mẫu độ đo 5 đường kính của quả cầu là 6,33; 6,37; 6,36; 6,32 và 6,37cm. Xác định ước lượng không chệch của trung bình và phương sai của đường kính quả cầu.

3. Để xác định độ chính xác của một chiếc cân tạ không có sai số hệ thống, người ta tiến hành 5 lần cân độc lập (cùng một vật), kết quả như sau:

   $$94,1 \quad 94,8 \quad 96,0 \quad 95,2\ kg$$

   Xác định ước lượng không chệch của phương sai số đo trong hai trường hợp:

   a) biết khối lượng vật cân là 95kg;
   b) không biết khối lượng vật cân.

4. Đường kính của một mẫu ngẫu nhiên của 200 viên bi được sản xuất bởi một máy trong một tuần có trung bình 20,9mm và độ lệch tiêu chuẩn 1,07mm. Ước lượng trung bình đường kính của viên bi với độ tin cậy (a) 95%, (b) 99%.

5. Để khảo sát sức bền chịu lực của một loại ống công nghiệp người ta tiến hành đo 9 ống và thu được các số liệu sau

   4500 6500 5000 5200 4800 4900 5125 6200 5375

   Từ kinh nghiệm nghề nghiệp người ta biết rằng sức bền đó có phân phối chuẩn với độ lệch chuẩn $\sigma = 300$. Xác định khoảng tin cậy 95% cho sức bền trung bình của loại ống trên.

6. Tại một vùng rừng nguyên sinh, người ta đeo vòng cho 1000 con chim. Sau một thời gian, bắt lại 200 con thì thấy có 40 con có đeo vòng. Thử ước lượng số chim trong vùng rừng đó với độ tin cậy 99%.

7. Biết tỷ lệ nảy mầm của một loại hạt giống là 0,9. Với độ tin cậy 0,95, nếu ta muốn độ dài khoảng tin cậy của tỷ lệ nảy mầm không vượt quá 0,02 thì cần phải gieo bao nhiêu hạt?

8. Kết quả quan sát về hàm lượng vitamine C của một loại trái cây cho ở bảng sau:

Hàm lượng vitamine C (%)	Số trái
6 – 7	5
7 – 8	10
8 – 9	20
9 – 10	35
10 – 11	25
11 – 12	5

   a) Hãy ước lượng hàm lượng vitamine C trung bình trong một trái với độ tin cậy 95%.
   b) Qui ước những trái có hàm lượng vitamine C trên 10% là trái loại A. Ước lượng tỷ lệ trái loại A với độ tin cậy 90%.
   c) Muốn độ chính xác khi ước lượng hàm lượng vitamine C trung bình là 0,1 và độ chính xác khi ước lượng tỷ lệ trái loại A là 5% với cùng độ tin cậy 95% thì cần quan sát thêm bao nhiêu trái nữa? A

9. Đo đường kính của 100 chi tiết máy do một phân xưởng sản xuất, ta được kết quả cho ở bảng sau:

Đường kính (mm)	Số chi tiết máy
9,85	8
9,90	12
9,95	20
10,00	30
10,05	14
10,10	10
10,15	6

Theo qui định, những chi tiết có đường kính từ $9,9mm$ đến $10,1mm$ là những chi tiết đạt tiêu chuẩn kỹ thuật.

a) Ước lượng tỷ lệ và ước lượng trung bình đường kính của những chi tiết đạt tiêu chuẩn với cùng độ tin cậy 95%?
b) Để độ chính xác khi ước lượng đường kính trung bình của những chi tiết đạt tiêu chuẩn là $0,02mm$ và độ chính xác khi ước lượng tỷ lệ chi tiết đạt tiêu chuẩn là 5% với cùng độ tin cậy 99% thì cần đo thêm ít nhất bao nhiêu chi tiết nữa?

10. Độ dài của bản kim loại tuân theo luật chuẩn. Đo 10 bản kim loại đó ta thu được số liệu sau:

$$4,1 \quad 3,9 \quad 4,7 \quad 4,4 \quad 4,0 \quad 3,8 \quad 4,4 \quad 4,2 \quad 4,4 \quad 5,0$$

Hãy xác định

a) Khoảng tin cậy 90% cho độ dài trung bình trên;
b) Khoảng tin cậjy 95% cho phương sai của độ dài đó.

11. Người ta đo chiều sâu của biển, sai lệch ngẫu nhiên được giả thiết phân phối theo qui luật chuẩn với độ lệch tiêu chuẩn là $20m$. Cần đo bao nhiêu lần để xác định chiều sâu của biển với sai lệch không quá $15m$ và độ tin cậy đạt được 95%?

12. Theo dõi số hàng bán được trong một ngày ở một cửa hàng, ta được kết quả ghi ở bảng sau:

Số hàng bán được ($kg/ngày$)	Số ngày
1900 − 1950	2
1950 − 2000	10
2000 − 2050	8
2050 − 2100	5

Hãy ước lượng phương sai của lượng hàng bán được mỗi ngày với độ tin cậy 95%? (cho biết $\alpha_1 = \alpha_2$).

◙ TRẢ LỜI BÀI TẬP

1. a) $9,5kg$, b) $0,74kg^2$

2. $\overline{x} = 6,35cm$, $s^2 = 0,00055cm^2$.

3. a) Trung bình khối lượng $m = 95kg$. Ước lượng không chệch của phương sai là

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - 95)^2 = 0,41$$

b) $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\sum_{i=1}^{5} x_i = 95,5$

Ước lượng không chệch của phương sai là

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - 95,5)^2 = 0,7rff$$

4. (a) $20,9 \pm 0,148mm$, (b) $20,9 \pm 0,195mm$.

5. $(5092,89\,;\,5484,89)$.

6. $0,1271 < p < 0,2729$

Tổng số chim trong vùng rừng nằm trong khoảng $(\frac{1000}{0,2729}, \frac{1000}{0,1271})$

7. $2 \times 1,96\sqrt{\frac{0,9\times 0,1}{n}} < 0,02$. Giải bất phương trình ta có $n > 3457$.

8. a) $9,06; 9,54)$, c) 467 trái.

9. a) $(0,792 < p < 0,928)$; $(9,982 < m < 10,006)$. b) 221.

10. a) $(4,09\,;\,4,49)$, b) $(0,064\,;\,0,456)$.

11. 7 lần.

12. $(1253,8 < \sigma^2 < 3983,8)$.

Chương 5

KIỂM ĐỊNH GIẢ THIẾT THỐNG KÊ

1. CÁC KHÁI NIỆM

1.1 Giả thiết thống kê

Khi nghiên cứu về các lĩnh vực nào đó trong thực tế ta thường đưa ra các nhận xét khác nhau về các đối tượng quan tâm. Những nhận xét như vậy thường được coi là các *giả thiết*, chúng có thể đúng và cũng có thể sai. Việc sai định tính đúng sai của một giả thiết được gọi là *kiểm định*.

Giả sử cần nghiên cứu tham số θ của đại lượng ngẫu nhiên X, người ta đưa ra giả thiết cần kiểm định

$$H\ :\ \theta = \theta_0$$

Gọi $\overline{H}$ là giả thiết đối của H thì $\overline{H}\ :\ \theta \neq \theta_0$.

Từ mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ ta chọn thống kê $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ sao cho nếu H đúng thì $\hat{\theta}$ có phân phối xác suất hoàn toàn xác định và với mẫu cụ thể thì giá trị của $\hat{\theta}$ sẽ tính được. $\hat{\theta}$ được gọi là *tiêu chuẩn kiểm định giả thiết* H.

Với α bé tùy ý cho trước ($\alpha \in (0,01; 0,05)$) ta tìm được miền W_α sao cho $P(\hat{\theta} \in W_\alpha) = \alpha$.

W_α được gọi là *miền bác bỏ* , α được gọi là *mức ý nghĩa của kiểm định*.

Thực hiện phép thủ đối với mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ ta được mẫu cụ thể $w_x = (x_1, x_2, \ldots, x_n)$. Tính giá trị của $\hat{\theta}$ tại $w_x = (x_1, x_2, \ldots, x_n)$ ta được $\theta_0 = \hat{\theta}(x_1, x_2, \ldots, x_n)$ (θ_0 được gọi là *giá trị quan sát*).

- Nếu $\theta_0 \in W_\alpha$ thì bác bỏ giả thiết H và thừa nhận giả thiết đối $\overline{H}$.
- Nếu $\theta_0 \notin W_\alpha$ thì chấp nhận giả thiết H.

⊙ **Chú ý**

Có trường hợp giả thiết kiểm định và giả thiết đối được nêu cụ thể hơn. Chẳng hạn:

$$H\text{: } \theta \leq \theta_0; \qquad \overline{H}\text{: } \theta > \theta_0$$

Khi đó ta có kiểm định một phía.

1.2 Sai lầm loại 1 và loại 2

Khi kiểm định giả thiết thống kê, ta có thể mắc phải một trong hai loại sai lầm sau:

i) *Sai lầm loại 1:* là sai lầm mắc phải khi ta bác bỏ một giả thiết H trong khi H đúng.

Xác suất mắc phải sai lầm loại 1 bằng $P(\hat{\theta} \in W_\alpha) = \alpha$.

ii) *Sai lầm loại 2:* là sai lầm mắc phải khi ta thừa nhận giả thiết H trong khi H sai.

Xác suất mắc phải sai lầm loại 2 bằng $P(\hat{\theta} \notin W_\alpha)$.

⊙ **Chú ý**

Nếu ta muốn giảm xác suất sai lầm loại 1 thì sẽ làm tăng xác suất sai lầm loại 2 và ngược lại.

Đối với một tiêu chuẩn kiểm định $\hat{\theta}$ và với mức ý nghĩa α ta có thể tìm được vô số miền bác bỏ W_α. Thường người ta ấn định trước xác suất sai lầm loại 1 (tức cho trước mức ý nghĩa α) chọn miền bác bỏ W_α nào đó có xác suất sai lầm loại 2 nhỏ nhất.

2. KIỂM ĐỊNH GIẢ THIẾT VỀ TRUNG BÌNH

Đại lượng ngẫu nhiên X có trung bình $E(X) = m$ chưa biết. Người ta đưa ra giả thiết

$$H : m = m_0 \quad (\overline{H} : m \neq m_0)$$

2.1 Trường hợp 1:

$$\begin{cases} Var(X) = \sigma^2 \text{ đã biết} \\ n \geq 30 \quad \text{hoặc} \quad (n < 30 \text{ và X có phân phối chuẩn}) \end{cases}$$

Chọn thống kê $U = \dfrac{(\overline{X} - m_0)\sqrt{n}}{\sigma}$. Nếu H_0 đúng thì $U \in N(0,1)$

Với mức ý nghĩa α cho trước, xác định phân vị chuẩn $u_{1-\frac{\alpha}{2}}$. Ta tìm được miền bác bỏ

$$\boxed{W_\alpha = \{u : |u| > u_{1-\frac{\alpha}{2}}\} = (-\infty; -u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty)}$$

Vì

$$\begin{aligned} P(U \in W_\alpha) &= P(U < -u_{1-\frac{\alpha}{2}} + P(U > u_{1-\frac{\alpha}{2}}) \\ &= P(U < u_{\frac{\alpha}{2}}) + 1 - P(U > u_{1-\frac{\alpha}{2}}) \\ &= \frac{\alpha}{2} + 1 - (1 - \frac{\alpha}{2}) = \alpha \end{aligned}$$

Lấy mẫu cụ thể và tính giá trị quan sát $\boxed{u_0 = \dfrac{|\overline{x} - m_0|}{\sigma}\sqrt{n}}$.

So sánh u_0 và $u_{1-\frac{\alpha}{2}}$.

- Nếu $u_0 > u_{1-\frac{\alpha}{2}}$ $(u_0 \in W_\alpha)$ thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.
- Nếu $u_0 < u_{1-\frac{\alpha}{2}}$ $(u_0 \notin W_\alpha)$ thì chấp nhận H_0.

• **Ví dụ 1** *Một tín hiệu của giá trị m được gởi từ địa điểm A và được nhận ở địa điểm B có phân phối chuẩn với trung bình m và độ lệch tiêu chuẩn $\sigma = 2$. Tin rằng giá trị của tín hiệu $m = 8$ được gởi mỗi ngày. Người ta tiến hành kiểm tra giả thiết này bằng cách gởi 5 tín hiệu một cách độc lập trong ngày thì thấy gía trị trung bình nhận được tại địa điểm B là $\overline{X} = 9,5$. Với độ tin cậy 95%, hãy kiểm tra giả thiết $m = 8$ đúng hay không?*

Giải

Ta cần kiểm định giả thiết $H : m_0 = 8 \quad (\overline{H} : m_0 \neq 8)$

Ta có $n = 5 < 30$. Độ tin cậy $1 - \alpha = 0,95 \implies 1 - \frac{\alpha}{2} = 0,975$

Phân vị chuẩn $u_{0,975} = 1,96$.

Miền bác bỏ là $W_\alpha = (-\infty; -1,96) \cup (1,96; +\infty)$.

Giá trị quan sát $u_0 = \dfrac{|\overline{x} - m_0|}{\sigma}\sqrt{n} = \dfrac{9,5 - 8}{2}\sqrt{5} = 1,68.$

Ta thấy $m_0 \notin W_\alpha$ nên giả thiết H được chấp nhận.

2.2 Trường hợp 2:

$\begin{cases} \sigma^2 \text{ chưa biết} \\ n \geq 30 \end{cases}$

Trong trường hợp này ta vẫn chọn thống kê như trên trong đó độ lệch tiêu chuẩn σ được thay bởi độ lệch tiêu chuẩn của mẫu ngẫu nhiên S'.

$$U = \frac{(\overline{X} - m_0)}{S'}\sqrt{n}$$

Nếu H đúng thì $U \in N(0,1)$. Tương tự như trên ta có miền bác bỏ là

$$W_\alpha = \{u : |u| > u_{1-\frac{\alpha}{2}}\} = (-\infty; u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty)$$

Lấy mẫu cụ thể và ta tính giá trị quan sát $\boxed{u_0 = \dfrac{|\overline{x} - m_0|}{s'}\sqrt{n}}$.

So sánh u_0 và $u_{1-\frac{\alpha}{2}}$.

- Nếu $u_0 > u_{1-\frac{\alpha}{2}}$ $(u_0 \in W_\alpha)$ thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.
- Nếu $u_0 < u_{1-\frac{\alpha}{2}}$ $(u_0 \notin W_\alpha)$ thì chấp nhận H_0.

• **Ví dụ 2** *Một nhóm nghiên cứu tuyên bố rằng trung bình một người vào siêu thị X tiêu hết 140 ngàn đồng. Chọn một mẫu ngẫu nhiên gồm 50 người mua hàng, tính được số tiền trung bình họ tiêu là 154 ngàn đồng với độ lệch tiêu chuẩn điều chỉnh của mẫu là $S' = 62$. Với mức ý nghĩa 0,02 hãy kiểm định xem tuyên bố của nhóm nghiên cứu có đúng hay không?*

Giải

Ta cần kiểm định giả thiết $H : m = 140 \quad (\overline{H} : m \neq 140)$

Ta có $n = 50 > 30$ và $1 - \frac{\alpha}{2} = 0,99$.

Phân ví chuẩn $u_{0,99} = 2,33$.

Miền bác bỏ $W_\alpha = (-\infty; -2,33) \cup (2,33; +\infty)$

Giá trị quan sát $u_0 = \dfrac{|\overline{x} - m_0|}{S'}\sqrt{n} = \dfrac{154 - 140}{62}\sqrt{50} = 1,59.$

Ta thấy $u_0 \notin W_\alpha$ nên chưa có cơ sở để loại bỏ H. Tạm thời chấp nhận rằng báo cáo của nhóm nghiên cứu là đúng.

2.3 Trường hợp 3:

$\begin{cases} \sigma^2 \text{ chưa biết} \\ n < 30 \text{ và X có phân phối chuẩn} \end{cases}$

Chọn thống kê

$$T = \frac{(\overline{X} - m_0)}{S'}\sqrt{n}$$

Nếu H đúng thì $T \in T(n-1)$

Với mức ý nghĩa α cho trước, ta xác định phân vị Student $(n-1)$ bậc tự do mức $1 - \frac{\alpha}{2}$ là $t_{1-\frac{\alpha}{2}}$.

Khi đó miền bác bỏ là

$$W_\alpha = \{t : |t| > t_{1-\frac{\alpha}{2}}\} = (-\infty; -t_{1-\frac{\alpha}{2}}) \cup (t_{1-\frac{\alpha}{2}}; +\infty)$$

Lấy mẫu cụ thể và tính giá trị quan sát $\boxed{t_0 = \dfrac{|\overline{x} - m_0|}{s'}\sqrt{n}}$.

- Nếu $t_0 > t_{1-\frac{\alpha}{2}} \quad (t_0 \in W_\alpha)$ thì bác bỏ giả thiết H và chấp nhận H.
- Nếu $t_0 < t_{1-\frac{\alpha}{2}} \quad (t_0 \notin W_\alpha)$ thì chấp nhận H.

• **Ví dụ 3** *Trọng lượng của các bao gạo là đại lượng ngẫu nhiên có phân phối chuẩn với trọng lượng trung bình là 50kg. Sau một khoảng thời gian hoạt động người ta nghi ngờ trọng lượng các bao gạo có thay đổi. Cân 25 bao gạo thu được các kết quả sau*

X(khối lượng)	n_i*(số bao)*
$48 - 48,5$	*2*
$48,5 - 49$	*5*
$49 - 49,5$	*10*
$49,5 - 50$	*6*
$50 - 50,5$	*2*

Với độ tin cậy 99%, hãy kết luận về điều nghi ngờ nói trên.

Giải

Xét giả thiết $H : m = 50$

$$T = \frac{(\overline{X} - 50)\sqrt{25}}{S'} \in T(24)$$

$x_i - x_{i+1}$	x_i^0	n_i(số bao)	$u_i n_i$	$x_i^2 n_i$
$48 - 48,5$	48,25	2	96,5	4656,125
$48,5 - 49$	48,75	5	243,75	11882,812
$49 - 49,5$	49,25	10	492,5	24255,625
$49,5 - 50$	49,75	6	298,5	14850,375
$50 - 50,5$	50,25	2	100,5	5050,125
$\sum$		25	1231,75	60695,062

Ta có $1 - \alpha = 0,99 \quad \Longrightarrow 1 - \frac{\alpha}{2} = 0,995$

Phân vị Student mức 0,995 với 24 bậc tụ do là $t_{1-\frac{\alpha}{2}} = u_{0,995} = 2,797$

Miền bác bỏ là $W_\alpha = (-\infty; -2,797) \cup (2,797; \infty)$

$\overline{x} = \frac{1231,75}{25} = 49,27.$

$s^2 = \frac{60695,06}{25} - (49,27)^2 = 2427,8 - 2427,53 = 0,27$

$s'^2 = \frac{25}{24}0,27 = 0,2812 \quad \Longrightarrow s' = 0,53$

Giá trị quan sát $t_0 = \frac{|(49,27-50)|\sqrt{25}}{0,53} = 6,886$

Ta thấy $t_0 \in W_\alpha$, nên giả thiết bị bác bỏ. Vậy điều nghi ngờ là đúng.

3. KIỂM ĐỊNH GIẢ THIẾT VỀ TỶ LỆ

Giả sử tổng thể có hai loại phần tử có tính chất A và không có tính chất A, trong đó tỷ lệ phần tử có tính chất A là p_0 chưa biết. Ta đưa ra thiết

$$H : \quad p = p_0$$

Lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ và tính tỷ lệ f các phần tử của mẫu có tính chất A.

Với mức ý nghĩa α cho trước, xác định phân vị chuẩn $u_{1-\frac{\alpha}{2}}$. Miền bác bỏ là

$$W_\alpha = \{u : |u| > u_{1-\frac{\alpha}{2}}\} = (-\infty; u_{1-\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty)$$

Lấy mẫu cụ thể và tính giá trị quan sát $\boxed{u_0 = \dfrac{|f - p_0|\sqrt{n}}{\sqrt{p_0 q_0}}}$

- Nếu $u_0 > u_{1-\frac{\alpha}{2}} \quad (u_0 \in W_\alpha)$ thì bác bỏ H và chấp nhận $\overline{H}$.
- Nếu $u_0 < u_{1-\frac{\alpha}{2}} \quad (u_0 \notin W_\alpha)$ thì chấp nhận H.

• **Ví dụ 4** *Tỷ lệ phế phẩm ở một nhà máy cần đạt là 10%. Sau khi cải tiến, kiểm tra 400 sản phẩm thì thấy có 32 phế phẩm với độ tin cậy 99%. Hãy xét xem việc cải tiến kỹ thuật có kết quả hay không?*

Giải

Ta có $n = 400$

Gọi p là tỷ lệ phế phẩm của nhà máy .Ta kiểm định giả thiết

$$H : p = 0,1. \quad (\text{giả thiết đối } \overline{H} : p < 0,1)$$

Tỷ lệ phế phẩm trong 400 sản phẩm là $f = \frac{32}{400} = 0,08$

Độ tin cậy $1 - \alpha = 0,99 \implies 1 - \frac{\alpha}{2} = 0,995 \implies u_{0,995} = 2,576$

Miền bác bỏ là $W_\alpha = (-\infty; -2,576) \cup (2,576; +\infty)$

Giá trị quan sát $u_0 = \frac{(|0,08-0,1|)\sqrt{400}}{\sqrt{0,1.0,9}} = 1,333 \notin W_\alpha$.

Do đó chấp nhận H_0.

Vậy việc cải tiến có hiệu quả.

4. KIỂM ĐỊNH GIẢ THIẾT VỀ PHƯƠNG SAI

Giả sử X là đại lượng ngẫu nhiên có phân phối chuẩn với phương sai $Var(X)$ chưa biết. Ta đưa ra giả thiết

$$H : Var(X) = \sigma_0^2$$

Lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ và chọn thống kê

$$\chi^2 = \frac{(n-1)S'^2}{\sigma_0^2}$$

Nếu H đúng thì χ^2 có phân phối " khi−bình phương " với $n-1$ bậc tự do.

Với mức ý nghĩa α cho trước, ta xác định các phân vị "khi−bình phương" $\chi^2_{n-1,\frac{\alpha}{2}}$, $\chi^2_{n-1,1-\frac{\alpha}{2}}$ $(n-1)$ bậc tự do, mức $\frac{\alpha}{2}$, $1 - \frac{\alpha}{2}$. Khi đó miền bác bỏ là

$$W_\alpha = \{t : t < \chi^2_{n-1,\frac{\alpha}{2}} \text{ hoặc } t > \chi^2_{n-1,1-\frac{\alpha}{2}}\} = (-\infty; \chi^2_{n-1,\frac{\alpha}{2}}) \cup (\chi^2_{n-1,1-\frac{\alpha}{2}}; +\infty)$$

Lấy mẫu cụ thể và tính giá trị quan sát $\boxed{\chi^2_0 = \dfrac{(n-1)s'^2}{\sigma_0^2}}$.

- Nếu $\chi^2_0 < \chi^2_{n-1,\frac{\alpha}{2}}$ hoặc $\chi^2_0 > \chi^2_{n-1,1-\frac{\alpha}{2}}$ $(\chi^2_0 \in W_\alpha)$ thì bác bỏ H và chấp nhận $\overline{H}$.
- Nếu $\chi^2_{n-1,\frac{\alpha}{2}} < \chi^2_0 < \chi^2_{n-1,1-\frac{\alpha}{2}}$ $(\chi^2_0 \notin W_\alpha)$ thì chấp nhận H.

• **Ví dụ 5** *Nếu máy móc hoạt động bình thường thì trọng lượng của sản phẩm là đại lượng ngẫu nhiên X có phân phối chuẩn với $D(X) = 12$. Nghi ngờ máy hoạt động không bình thường người ta cân thử 13 sản phẩm và tính được $s'^2 = 14,6$. Với mức ý nghĩa $\alpha = 0,05$. Hãy kết luận điều nghi ngờ trên có đúng hay không?*

Giải

Ta kiểm định giả thiết $H : Var(X) = 12$; $\overline{H} : Var(X) \neq 12$.

Từ các số liệu của bài toán ta tìm được $\chi^2_0 = \frac{(13-1)14,6}{12} = 14,6$

Với $\alpha = 0,05$, tra bảng phân vị χ^2 với $(n-1) = 12$ bậc tự do ta được

$$\chi^2_{\frac{\alpha}{2}} = \chi^2_{0,025} = 4,4 \quad \text{và} \quad \chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0,975} = 23,3$$

Ta thấy $4,4 < 14,6 < 23,3$ nên chấp nhận giả thiết H.

Vậy điều nghi ngờ trên là không đúng. Máy vẫn hoạt động bình thường.

5. KIỂM ĐỊNH MỘT PHÍA

Trong các bài toán trên ta chỉ xét giả thiết đối có dạng $\overline{H} : \theta \neq \theta_0$. Ta cũng có thể giải bài toán kiểm định với giả thiết đối có dạng: $\overline{H} : \theta < \theta_0$ hoặc $\overline{H} : \theta > \theta_0$. Khi giải các bài toán này ta cũng áp dụng các qui tắc đã được trình bày với chú ý là:

i) Khi tính gía trị quan sát u_0 (hoặc t_0) trong các qui tắc kiểm định trên ta bỏ dấu trị tuyệt đối ở tử số và thay bằng dấu ngoặc đơn (...). Chẳng hạn $u_0 = \dfrac{(\overline{x} - \mu_0)}{\sigma}\sqrt{n}$.

ii) Nếu giả thiết đối có dạng $\overline{H} : \theta > \theta_0$ thì ta so sánh gía trị quan sát u_0 với $u_\gamma = u_{1-\alpha}$ (hoặc $t_\gamma = t_{1-\alpha}$, hoặc $\chi^2_{1-\alpha}$).

Nếu $u_0 > u_\gamma$ (hoặc $t_0 > t_\gamma, \chi^2_0 > \chi^2_{1-\alpha}$) thì bác bỏ H và thừa nhận H. Nếu ngược lại thì chấp nhận H.

iii) Nếu giả thiết đối có dạng $H : \theta < \theta_0$ thì ta so sánh u_0 với $u_\gamma = -u_{1-\alpha}$, (hoặc $t_\gamma = -t_{1-\alpha}$, hoặc χ^2_α).

Nếu $u_0 < -u_{1-\alpha}$;(hoặc $t_0 < -t_{1-\alpha}, \chi^2_0 < \chi^2_\alpha$) thì bác bỏ H.Nếu ngược lại thì chấp nhận H.

• **Ví dụ 6** *Một nhà sản xuất thuốc chống dị ứng thực phẩm tuyên bố rằng 90% người dùng thuốc thấy thuốc có tác dụng trong vòng 8 giờ. Kiểm tra 200 người bị dị ứng thực phẩm thì thấy trong vòng 8 giờ thuốc làm giảm bớt dị ứng đối với 160 người. Hãy kiểm định xem lời tuyên bố trên của nhà sản xuất có đúng hay không với mức ý nghĩa* $\alpha = 0,01$.

Giải

Ta đưa ra giả thiết $H : p_0 = 0,9 \quad (\overline{H} < 0,9)$

$\alpha = 0,01 \longrightarrow 1 - \alpha = 0,99 \Longrightarrow -u_{1-\alpha} = -2,326$

$f = \dfrac{160}{200} = 0,8$

$u_0 = \dfrac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \dfrac{0,8 - 0,9}{\sqrt{0,9 \times 0,1}}\sqrt{200} = -\dfrac{0,1}{0,3}.14,14 = -4,75$

Ta thấy $u_0 < -u_{1-\alpha}$ nên bác bỏ giả thiết H.

Vậy lời tuyên bố của nhà sản xuất là không đúng sự thật.

6. KIỂM ĐỊNH GIẢ THIẾT VỀ SỰ BẰNG NHAU GIỮA HAI TRUNG BÌNH

Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập có cùng phân phối chuẩn với $E(X)$ và $E(Y)$ chưa biết. Ta cần kiểm định giả thiết

$$H : \ E(X) = E(Y) \qquad (\overline{H} : \ E(X) \neq E(Y))$$

Lấy mẫu ngẫu nhiên kích thước n đối X và mẫu ngẫu nhiên kích thước m đối với Y và xét các trường hợp:

i) Trường hợp biết $Var(x) = \sigma_x^2, Var(y) = \sigma_y^2$

Tính giá trị quan sát $u_0 = \dfrac{|\overline{x} - \overline{y}|}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$.

ii) Trường hợp chưa biết $Var(X), Var(Y)$.

Tính giá trị quan sát $u_0 = \dfrac{|\overline{x} - \overline{y}|}{\sqrt{\frac{s_x'^2}{n} + \frac{s_y'^2}{m}}}$.

Với mức ý nghĩa α cho trước, xác định phân vị chuẩn $u_{1-\frac{\alpha}{2}}$.

Ta tìm được miền bác bỏ $W_\alpha = \{ u : |u| > u_{1-\frac{\alpha}{2}} \}$.

So sánh u_0 và $u_{1-\frac{\alpha}{2}}$

* Nếu $u_0 > u_{1-\frac{\alpha}{2}}$ thì bác bỏ giả thiết H và thừa nhận $\overline{H}$.

* Nếu $u_0 < u_{1-\frac{\alpha}{2}}$ thì thừa nhận H.

• **Ví dụ 7** *Trọng lượng sản phẩm do hai nhà máy sản xuất là các đại lượng ngẫu nhiên có phân phối chuẩn và có cùng độ lệch tiêu chuẩn là $\sigma = 1kg$. Với mức ý nghĩa $\alpha = 0,05$, có thể xem trọng lượng trung bình của sản phẩm do hai nhà máy sản xuất là như nhau hay không? Nếu cân thử 25 sản phẩm của nhà máy A ta tính được $\overline{x} = 50kg$, cân 20 sản phẩm của nhà máy B thì tính được $\overline{y} = 50,6kg$.*

Giải

Gọi trọng lượng của nhà máy A là X; trọng lượng của nhà máy B là Y thì X, Y là các đại lượng ngẫu nhiên có phân phối chuẩn với $Var(X) = Var(Y) = 1$.

Ta kiểm tra giả thiết $H: \ E(X) = E(Y); \quad (E(X) \neq E(Y))$

Với mức ý nghĩa $\alpha = 0,05$ thì $u_{1-\frac{\alpha}{2}} = 1,96$.

Tính $u_0 = \frac{|50-50,6|}{\sqrt{\frac{1}{25}+\frac{1}{20}}} = 2$.

Ta thấy $u_0 > u_{1-\frac{\alpha}{2}}$ nên bác bỏ giả thiết H, tức là trọng lượng trung bình của sản phẩm sản xuất ở hai nhà máy là khác nhau.

7. KIỂM ĐỊNH GIẢ THIẾT VỀ SỰ BẰNG NHAU CỦA HAI TỶ LỆ

Giả sử p_1, p_2 tương ứng là tỷ lệ các phần tử mang dấu hiệu nào đó của tổng thể thứnhất, tổng thể thứ hai. Ta cần kiểm định giả thiết

$$H: \ p_1 = p_2 = p_0 \qquad (H: \ p_1 \neq p_2)$$

i) Trường hợp chưa biết p_0.

Chọn thống kê $U = \dfrac{(P^* - p_1) - (p^* - p_2)}{\sqrt{p^*(1-p^*)(\frac{1}{n_1}+\frac{1}{n_2})}}$.

với $p^* = \dfrac{n_1.f_{n_1} + n_2.f_{n_2}}{n_1 + n_2}$ (ước lượng hợp lý tối đa của p_0)

trong đó

f_{n_1} là tỷ lệ phần tử có dấu hiệu của mẫu thứ nhất với kích thước n_1.

f_{n_2} là tỷ lệ phần tử có dấu hiệu của mẫu thứ hai với kích thước n_2.

Với n_1, n_2 khá lớn thì U có phân phối chuẩn hóa.

ii) Trường hợp biết p_0.

Chọn thống kê $U = \dfrac{f_{n_1} - f_{n_2}}{\sqrt{p_0(1-p_0)(\frac{1}{n_1}+\frac{1}{n_2})}}$

*** Qui tắc kiểm định**

Lấy hai mẫu ngẫu nhiên kích thước n_1, n_2 và tính

$$u_0 = \frac{|f_{n_1} - f_{n_2}|}{\sqrt{p^*(1-p^*)(\frac{1}{n_1}+\frac{1}{n_2})}} \qquad (p^* = \frac{n_1.f_{n_1} + n_2.f_{n_2}}{n_1+n_2}) \text{ nếu chưa biết } p_0$$

hoặc

$$u_0 = \frac{|f_{n_1} - f_{n_2}}{\sqrt{p_0(1-p_0)(\frac{1}{n_1}+\frac{1}{n_2})}} \text{ nếu biết } p_0.$$

Với mức ý nghĩa α cho trước, xác định phân vị chuẩn $u_{1-\frac{\alpha}{2}}$.

Ta tìm được miền bác bỏ $W_\alpha = \{\, u : \ |u|.u_{1-\frac{\alpha}{2}} \,\}$.

So sánh u_0 và $u_{1-\frac{\alpha}{2}}$

* Nếu $u_0 > u_{1-\frac{\alpha}{2}}$ thì bác bỏ giả thiết H.

* Nếu $u_0 < u_{1-\frac{\alpha}{2}}$ thì thừa nhận giả thiết H.

• **Ví dụ 8** *Kiểm tra các sản phẩm được chọn ngẫu nhiên ở hai nhà máy sản xuất ta được các số liệu sau:*

Nhà máy I	*Số sản phẩm được kiểm tra*	*Số phế phẩm*
I	$n_1 = 100$	*20*
II	$n_2 = 120$	*36*

Với mức ý nghĩa $\alpha = 0,01$; có thể coi tỷ lệ phế phẩm của hai nhà máy là như nhau không?

Giải

Gọi p_1, p_2 tương ứng là tỷ lệ phế phẩm của nhà máy I, II.

Ta kiểm tra giả thiết $H : \ p_1 = p_2 \qquad (H : \ p_1 \neq p_2)$.

Với mức ý nghĩa $\alpha = 0,01$ thì $u_{1-\frac{\alpha}{2}} = u_{0,995} = 2,58$.

Từ các số liệu đã cho ta có

$$f_{n_1} = \frac{20}{100} = 0,2; \qquad f_{n_2} = \frac{36}{120} = 0,3$$

$$p^* = \frac{100 \times 0,2 + 120 \times 0,3}{100+120} = 0,227 \quad \Longrightarrow \quad 1 - p^* = 0,773$$

Do đó $u_0 = \dfrac{|0,2-0,3|}{\sqrt{0,227 \times 0,773(\frac{1}{100}+\frac{1}{120})}} \approx 1,763.$

Ta thấy $u_0 < u_{1-\frac{\alpha}{2}}$ nên chấp nhận giả thiết H, tức là tỷ lệ phế phẩm của hai nhà máy là như nhau.

8. KIỂM ĐỊNH GIẢ THIẾT VỀ SỰ BẰNG NHAU GIỮA HAI PHƯƠNG SAI

Giả sử X, Y là hai đại lượng ngẫu nhiên độc lập có phân phối chuẩn với các tham số tương ứng σ_x^2, σ_y^2 chưa biết. Ta cần kiểm định giả thiết

$$H:\ \sigma_x^2 = \sigma_y^2 \qquad (\text{giả thiết đối } \overline{H}: \sigma_x^2 \neq \sigma_y^2)$$

Lấy mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$, $W_Y = (Y_1, Y_2, \ldots, Y_n)$ đối với X, Y.

Chọn các thống kê

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1} \qquad S_y^2 = \frac{\sum_{i=1}^m (Y_j - \overline{X})^2}{m-1}$$

Ta thấy $\dfrac{(n-1)S_x^2}{\sigma_x^2}$ và $\dfrac{(m-1)S_y^2}{\sigma_y^2}$ là các đại lượng ngẫu nhiên độc lập có phân phối χ^2 với $n-1$ và $m-1$ bậc tự do. Do đó $\dfrac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$ có phân phối F với các tham số $n-1$ và $m-1$.

Khi H đúng thì $S_x^2/S_y^2 \in F_{\alpha/2,n-1,m-1}$ và có

$$P(F_{1-\alpha/2,n-1,m-1} < S_x^2/S_y^2 < F_{\alpha/2,n-1,m-1}) = 1-\alpha$$

Ta tìm được

* Miền bác bỏ $W_\alpha = (-\infty, F_{1-\alpha/2,n-1,m-1}) \cup (F_{\alpha/2,n-1,m-1}, +\infty)$.

* Giá trị quan sát $v = \frac{S_x^2}{S_y^2}$

Do đó

- Nếu $v \in W_\alpha$ thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.
- Nếu $v \notin W_\alpha$ thì chấp nhận giả= thiết H.

⊙ **Chú ý** Kiểm định ở trên bị ảnh hưởng bởi giá trị quan sát $v = S_x^2/S_y^2$ và xác suất $P(F_{n-1,m-1} < v)$ trong đó $F_{n-1,m-1}$ là đại lượng ngẫu nhiên có phân phối F với các tham số $n-1, m-1$. Nếu xác suất nhỏ hơn $\frac{\alpha}{2}$ (xảy ra khi S_x^2 nhỏ hơn S_y^2) hoặc lớn hơn $1-\alpha/2$ (xảy ra khi S_x^2 lớn hơn S_y^2) thì giả thiết bị tù chối.

Nếu đặt

$$p - \text{giá trị} = 2\min[P(F_{n-1,m-1<v}), 1 - P(F_{n-1,m-1})]$$

thì giả thiết bị tù chối khi mức ý nghĩa α t lớn hơn $p-$giá trị.

• **Ví dụ 9** *Có hai cách chọn chất xúc tác khác nhau để kích thích một phản ứng hóa học. Để kiểm định phương sai sản sinh ra có giống nhau hay không người ta lấy mẫu gồm 10 nhóm dùng cho chất xúc tác thứ nhất và 12 nhóm dùng cho chất xúc tác thứ hai.*

Dữ liệu cho kết quả $S_1^2 = 0,14$ *và* $S_2^2 = 0,28$. *Với mức ý nghĩa 5%, hãy kiểm định giả thiết trên.*

Giải

Ta cần kiểm định giả thiết $H : \sigma_1^2 = \sigma_2^2$.

Ta có $v = \frac{S_1^2}{S_2^2} = \frac{0,14}{0,28} = 0,5$ và $P(F_{9,11<0,5}) = 0,1539$.

Do đó $p-$giá trị $= 2\min(0,1539; 0,8461) = 0,3074$.

Ta thấy $\alpha = 0,05 < p-$giá trịnên giả thiết về sự bằng nhau của hai phương sai được chấp nhận.

9. BÀI TẬP

1. Độ bền của một loại dây thép sản xuất theo công nghệ cũ là 150. Sau khi cải tiến kỹ thuật người ta lấy mẫu gồm 100 sợi dây thép để thử độ bền thì thấy độ bền trung bình là 185 và $s = 25$. Với mức ý nghĩa $\alpha = 0,05$, hỏi công nghệ mới có tốt hơn công nghệ cũ hay không?

2. Độ dày của một chi tiết mấy do một máy sản xuất là một đại lượng ngẫu nhiên phân phối theo qui luật chuẩn với độ dày trung bìng $1,25mm$. Nghi ngờ máy hoạt động không bình thường người ta kiểm tra 10 chi tiết máy thì thấy độ dài trung bình là $1,325$ với độ lệch tiêu chuẩn $0,075mm$. Với mức ý nghĩa $\alpha = 0,01$, hãy kết luận về điều nghi ngờ nói trên?

3. Trọng lượng của một loại sản phẩm do một nhà máy sản xuất là đại lượng ngẫu nhiên phân phối theo qui luật chuẩn với trọng lượng trung bình là 500 gr. Nghi ngờ trọng lượng của loại sản phẩm này có xu hướng giảm sút, người ta cân thử 25 sản phẩm và thu được kết quả cho ở bảng sau:

Trọng lượng (gr)	480	485	490	495	500	510
Số sản phẩm	2	3	8	5	3	4

Với mức ý nghĩa $\alpha = 0,05$, hãy kết luận về điều nghi ngờ nói trên?

4. Năng suất lúa trung bình trong vụ trước là 4,5 tấn/ha. Vụ lúa năm nay người ta áp dụng một biện pháp kỹ thuật mới cho toàn bộ diện tích trồng lúa ở trong vùng. Theo dõi năng suất lúa ở 100 hecta ta có bảng số liệu sau:

Năng suất (tạ/ha)	Diện tích (ha)
30 – 35	7
35 – 40	12
40 – 45	18
45 – 50	27
50 – 55	20
55 – 60	8
60 – 65	5
65 – 70	3

Hãy cho kết luận về biện pháp kỹ thuật mới này?

5. Tuổi thọ trung bình của một mẫu gồm 100 bóng đèn được sản xuất ở một nhà máy là 1570 giờ với độ lệch tiêu chuẩn 120 giờ. Gọi μ là tuổi thọ trung bình của tất cả bóng đèn nhà máy sản xuất ra. Với mức ý nghĩa $\alpha = 0,05$, hãy kiểm tra giả thiết $H_0 : \mu = 1600$ giờ với giả thiết đối $H_1 : \mu < 1600$ giờ.

6. Một hãng dược phẩm sản xuất một loại thuốc trị dị ứng thực phẩm tuyên bố rằng thuốc có tác dụng giảm dị ứng trong 8 giờ đối với 90% người dùng. Kiểm tra 200 người bị dị ứng dùng thì thấy thuốc có tác dụng đối với 160 người . Với mức ý nghĩa $\alpha = 0,01$, kiểm tra xem lời tuyên bố trên có đúng không?

7. Tỷ lệ phế phẩm của một nhà máy trước đây là 5%. Năm nay nhà máy áp dụng một biện pháp kỹ thuật mới. Để xem biện pháp kỹ thuật mới có tác dụng làm giảm tỷ lệ phế phẩm của nhà máy hay không, người ta lấy một mẫu gồm 800 sản phẩm để kiểm tra và thấy có 24 phế phẩm trong mẫu này.

   a) Với mức ý nghĩa $\alpha = 0,01$, hãy cho kết luận về biện pháp kỹ thuật mới đó?

   b) Nếu nhà máy báo cáo tỷ lệ phế phẩm sau khi áp dụng biện pháp kỹ thuật mới đã giảm xuống 2% (vơs i mức ý nghĩa $\alpha = 0,05$) thì có chấp nhận được không?

8. Giám đốc một nhà máy tuyên bố 90% máy móc của nhà máy đạt tiêu chuẩn kỹ thuật quốc tế. Người ta tiến hành kiểm tra 200 máy thì thấy có 168 máy đạt tiêu chuẩn kỹ thuật quốc tế. Với mức ý nghĩa $\alpha = 0,05$, hãy kết luận về lời tuyên bố trên?

9. Nếu máy móc làm việc bình thường thì kích thước của một loại sản phẩm là đại lượng ngẫu nhiên phân phối theo qui luật chuẩn với $Var(X) = 0,25$. Nghi ngờ máy làm việc không bình thường, người ta tiến hành đo thử 28 sản phẩm và thu được kết quả cho ở bảng sau:

Kích thước (cm)	19,0	19,5	19,8	20,4	20,6
Số sản phẩm	2	4	5	12	5

Với mức ý nghĩa $\alpha = 0,02$, hãy kết luận về điều nghi ngờ nói trên?

10. Trọng lượng của gói hàng được đóng bao bởi một máy trước đây là 1135 gram với độ lệch tiêu chuẩn là 7,1 gram. Nghi ngờ máy hoạt động không tốt, người ta tiên hành kiểm tra 20 gói hàng thì thấy độ lệch tiêu chuẩn là 9,1 gram. Với mức ý nghĩa $\alpha = 0,05$, hãy kiểm tra giả thiết ($H_0 : \sigma = 7,1$ gram) với giả thiết đối ($H_1 : \sigma > 7,1$ gram).

11. Theo dõi số tai nạn lao động của hai phân xưởng, ta có số liệu sau: phân xưởng I: 20/200 công nhân, phân xưởng II: 120/800 công nhân. Với mức ý nghĩa $\alpha = 0,005$ hỏi có sự khác nhau đáng kể về chất lượng công tác bảo hộ lao động ở hai phân xưởng trên hay không?

12. Để nghiên cứu ảnh hưởng của một loại thuốc, người ta cho 10 bệnh nhân uống thuốc. Lần khác họ cũng cho bệnh nhân uống thuốc nhưng là thuốc giả (thuốc không có tác dụng). Kết quả thí nghiệm thu được như sau:

Bệnh nhân	1	2	3	4	5	6	7	8	9	10
Số giờ ngủ có thuốc	6,1	7,0	8,2	7,6	6,5	8,4	6,9	6,7	7,4	5,8
Số giờ ngủ với thuốc giả	5,2	7,9	3,9	4,7	5,3	5,4	4,2	6,1	3,8	6,3

Giả sử số giờ ngủ của các bệnh nhân có qui luật chuẩn. Với mức ý nghĩa $\alpha = 0,05$, hãy kết luận về ảnh hưởng của loại thuốc ngủ trên?

TRẢ LỜI BÀI TẬP

1. $u_0 = 14 > 1,645$ nên việc cải tiến kỹ thuật là có hiệu quả.

2. Vì $u_0 = 3 < 3,25$ nên điều nghi ngờ trên là sai.

3. $t_0 = 3,37$. Điều nghi ngờ là đúng.

4. Biện pháp kỹ thuật mới có tác dụng làm tăng năng suất lúa trung bình của toàn vùng.

5. Vì $u_0 = -2,5 < -1,645$ nên bác bỏ H_0.

6. $u_0 = 4,73$. Lời tuyên bố không đúng.

8. Lời tuyên bố là sai.

9. Nghi ngờ sai. Máy làm việc bình thường.

10. $\chi_0^2 = 32,86 > 30,1$ nên bác bỏ H_0.

11. Do $1,82 < 1,96$ nên không có cơ sở cho rằng sự khác biệt đáng kể về chất lượng công tác bảo hộ lao động ở hai phân xưởng.

12. Loại thuốc ngủ trên có tác dụng.

Chủởng 6

LÝ THUYẾT TƯỞNG QUAN VÀ HÀM HỒI QUI

1. MỐI QUAN HỆ GIỮA HAI ĐẠI LƯỢNG NGẪU NHIÊN

Khi khảo sát hai đại lượng ngẫu nhiên X, Y ta thấy giữa chúng có thể có một số quan hệ sau:

i) X và Y độc lập với nhau, tức là việc nhận giá trị của đại lượng ngẫu nhiên này không ảnh hưởng đến việc nhận giá trị của đại lượng ngẫu nhiên kia.

ii) X và Y có mối phụ thuộc hàm số $Y = \varphi(X)$.

iii) X và Y có sự phụ thuộc tương quan và phụ thuộc không tương quan.

2. HỆ SỐ TƯỞNG QUAN

2.1 Moment tưởng quan (Covarian)

□ **Định nghĩa 1**

** Moment tưởng quan (hiệp phương sai) của hai đại lượng ngẫu nhiên X và Y, kí hiệu $cov(X,Y)$ hay μ_{XY}, là số được xác định như sau*

$$cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

** Nếu $cov(X,Y) = 0$ thì ta nói hai đại lượng ngẫu nhiên X và Y không tưởng quan.*

⊙ **Chú ý**

$$\boxed{cov(X,Y) = E(XY) - E(X).E(Y)}$$

Thật vậy, ta có

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

⊕ **Nhận xét 1**

* Nếu (X, Y) rời rạc thì

$$cov(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(x_i, y_j) - E(X)E(Y)$$

* Nếu (X, Y) liên tục thì

$$cov(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy - E(X)E(Y)$$

⊕ **Nhận xét**

i) Nếu X và Y là hai đại lượng ngẫu nhiên độc lập thì chúng không tương quan.

ii) Cov(X,X)=Var(X).

2.2 Hệ số tương quan

□ **Định nghĩa 2** *Hệ số tương quan của hai đại lượng ngẫu nhiên X và Y, kí hiệu r_{XY}, là số được xác định như sau*

$$\boxed{r_{XY} = \frac{cov(X,Y)}{S_X.S_Y}}$$

với S_x, S_Y là độ lệch tiêu chuẩn của X, Y.

• **Ý nghĩa của hệ số tương quan**

Hệ số tương quan đo mức độ phụ thuộc tuyến tính giữa X và Y. Khi $|r_{XY}|$ càng gần 1 thì mối quan hệ tuyến tính càng chặt, khi $|r_{XY}|$ càng gần 0 thì quan hệ tuyến tính càng "lỏng lẻo".

2.3 Ước lượng hệ số tương quan

Lập mẫu ngẫu nhiên $W_{XY} = [(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)]$.

Để ước lượng hệ số tương quan $r_{XY} = \dfrac{E(XY) - E(X).E(Y)}{S_X.S_Y}$ ta dùng thống kê

$$R = \frac{\overline{XY} - \overline{X}.\overline{Y}}{S_X.S_Y}$$

trong đó

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \overline{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Với mẫu cụ thể, ta tính được giá trị của R là

$$\boxed{r_{XY} = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x.s_y}}$$

trong đó

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2, \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\overline{y})^2$$

Ta có

$$\boxed{r_{XY} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}.\sqrt{n(\sum y^2) - (\sum y)^2}}}$$

2.4 Tính chất của hệ số tương quan

Hệ số tương quan $r = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{s_x.s_y}$ được dùng để đánh giá mức độ chặt chẽ của sự phụ thuộc tương quan tuyến tính giữa hai đại lượng ngẫu nhiên X và Y, nó có các tính chất sau đây:

i) $|r| \leq 1$.

ii) Nếu $|r| = 1$ thì X và Y có quan hệ tuyến tính.

iii) Nếu $|r|$ càng lớn thì sự phụ thuộc tương quan tuyến tính giữa X và Y càng chặt chẽ.

iv) Nếu $|r| = 0$ thì giữa X và Y không có phụ thuộc tuyến tính tương quan.

v) Nếu $r > 0$ thì X và Y có tương quan thuận (X tăng thì Y tăng). Nếu $r < 0$ thì X và Y có tương quan nghịch (X giảm thì Y giảm).

• **Ví dụ 1** *Từ số liệu được cho bởi bảng sau, hãy xác định hệ số tương quan của Y và X*

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Giải

Ta lập bảng sau

x_i	y_i	x_i^2	x_iy_i	y_i^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\sum x = 56$	$\sum y = 40$	$\sum x^2 = 524$	$\sum xy = 364$	$\sum y^2 = 256$

Hệ số tương quan của X và Y là

$$r_{XY} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}.\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$= \frac{8.364 - (56).(40)}{\sqrt{8.524 - (56)^2}.\sqrt{8.256 - (40)^2}} = \frac{672}{687,81} = 0,977$$

2.5 Tỷ số tương quan

Để đánh giá mức độ chặt chẻ của sự phụ thuộc tương quan phi tuyến, người ta dùng *tỷ số tương quan*:

$$\boxed{\eta_{Y/X} = \frac{s_{\overline{y}}}{s_y}}$$

trong đó

$$s_{\overline{y}} = \sqrt{\frac{1}{n}\sum n_i.(\overline{y_{x_i}} - \overline{y})^2}; \qquad s_y = \sqrt{\frac{1}{n}\sum m_j.(y_j - \overline{y})^2}$$

Tỷ số tương quan có các tính chất sau:

i) $0 \le \eta_{Y/X} \le 1$.

ii) $\eta_{Y/X} = 0$ khi và chỉ khi Y và X không có phụ thuộc tương quan.

iii) $\eta_{Y/X} = 1$ khi và chỉ khi Y và X phụ thuộc hàm số.

iv) $\eta_{Y/X} \ge |r|$.

Nếu $\eta_{Y/X} = |r|$ thì sự phụ thuộc tương quan của Y và X có dạng tuyến tính.

2.6 Hệ số xác định mẫu

Trong thống kê, để đánh giá chất lượng của mô hình tuyến tính người ta còn xét *hệ số xác định mẫu* $\beta = r^2$ với r là hệ số tương quan. Ta có $0 \le \beta \le 1$.

3. HỒI QUI

3.1 Kỳ vọng có điều kiện

i) Đại lượng ngẫu nhiên rời rạc

* Kỳ vọng có điều kiện của đại lượng ngẫu nhiên rời rạc Y với điều kiện $X = x$ là

$$E(Y/x) = \sum_{j=1}^{m} y_j P(X = x, Y = y_j)$$

* Tương tự, kỳ vọng có điều kiện của đại lượng ngẫu nhiên rời rạc X với điều kiện $Y = y$ là

$$E(X/y) = \sum_{i=1}^{n} x_i P(X = x_i, Y = y)$$

ii) Đại lượng ngẫu nhiên liên tục

$$E(Y/x) = \int_{-\infty}^{+\infty} y f(y/x) dy$$

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

trong đó

$f(y/x) = f(x, y)$ với x không đổi

$f(x/y) = f(x, y)$ với y không đổi

3.2 Hàm hồi qui

* Hàm hồi qui của Y đối với X là $f(x) = E(Y/x)$.

* Hàm hồi qui của X đối với Y là $f(y) = E(X/y)$.

Trong thực tế ta thường gặp hai đại lượng ngẫu nhiên X, Y có mối liên hệ với nhau, trong đó việc khảo sát X thì dễ còn khảo sát Y thì khó hơn thậm chí không thể khảo sát được. Người ta muốn tìm mối liên hệ $\varphi(X)$ nào đó giữa X và Y để biết X ta có thể dự đoán được Y.

Giả sử biết X, nếu dự đoán Y bằng $\varphi(X)$ thì sai số phạm phải là $E[Y - \varphi(X)]^2$. Vấn đề được đặt ra là tìm $\varphi(X)$ như thế nào để $E[Y - \varphi(X)]^2$ là nhỏ nhất.

Ta sẽ chứng minh khi chọn $\varphi(X) = E(Y/X)$ (với $\varphi(x) = E(Y/x)$) thì $E[Y - \varphi(X)]^2$ sẽ nhỏ nhất.

Thật vậy, ta có

$$\begin{aligned} E[Y - \varphi(X)]^2 &= E\{([Y - E(Y/X)] + [E(Y/X) - \varphi(X)])^2\} \\ &= E\{[Y - E(Y/X)]^2\} + E\{[E(Y/X) - \varphi(X)]^2\} \\ &\quad +2E\{[Y - E(Y/X)][E(Y/X) - \varphi(X)]\} \end{aligned}$$

Ta thấy $E(Y/X)$ chỉ phụ thuộc vào X nên có thể đặt $T(X) = E(Y/X) - \varphi(X)$.

Vì $E[E(Y/X)T(X)] = E[YT(X)]$ nên

$$\begin{aligned} 2E[Y - E(Y/X)][E(Y/X) - \varphi(X)] &= 2E\{[Y - E(Y/X)]T(X)\} \\ &= 2E[YT(X)] - 2E[E(Y/X)T(X)] = 0 \end{aligned}$$

Do đó

$$E\{[Y - \varphi(X)]^2\} = E\{[Y - E(Y/X)]^2\} + E\{E(Y/X) - \varphi(X)]^2$$

nhỏ nhất khi

$$E\{[(Y/X) - \varphi(X)]^2 = 0$$

Ta chỉ cần chọn

$$\varphi(X) = E(Y/X) \tag{6.1}$$

Phương trình (6.1) được gọi là *phương trình tương quan* hay *phương trình hồi qui*.

3.3 Xác định hàm hồi qui

a) Trường hợp ít số liệu (tương quan cặp)

Giả sử giữa hai đại lượng ngẫu nhiên X và Y có tương quan tuyến tính, tức là $E(Y/X) = AX + B$.

Dựa vào n cặp giá trị $(x_1, x_2), (x_2, y_2), \ldots, (x_n, y_n)$ của (X, Y) ta tìm hàm

$$\overline{y_x} = y = ax + b \quad (*)$$

để ước lượng hàm $Y = AX + B$.

(*) được gọi là *hồi qui tuyến tính mẫu*.

Vì các cặp giá trị trên là trị xấp xỉ của x và y nên thỏa (*) một cách xấp xỉ.

Do đó $y_i = ax_i + b + \varepsilon_i$ hay $\varepsilon_i = y_i - ax_i - b$.

Ta tìm a, b sao cho các sai số ε_i $(i = \overline{1, n})$ có trị tuyệt đối nhỏ nhất hay hàm

$$S(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

đạt cực tiểu. Phương pháp tìm này được gọi là *phương pháp bình phương bé nhất*.

Ta thấy S sẽ đạt giá trị nhỏ nhất tại điểm dừng thỏa mãn

$$0 = \frac{\partial S}{\partial a} = -2 \sum_{i=1}^{n} x_i (y_i - ax_i - b)$$

$$0 = \frac{\partial S}{\partial b} = -2 \sum_{i=1}^{n} (y_i - ax_i - b)$$

hay

$$\boxed{\begin{array}{rcl}\left(\sum_{i=1}^{n} x_i^2\right).a + \left(\sum_{i=1}^{n} x_i\right).b &=& \sum_{i=1}^{n} x_i y_i \\ \left(\sum_{i=1}^{n} x_i\right).a + nb &=& \sum_{i=1}^{n} y_i\end{array}} \tag{6.2}$$

Hệ trên có định thức

$$D = \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix} = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2$$

Vì các x_i khác nhau nên theo bất đẳng thức Bunhiakovsky ta có $\left(\sum_{i=1}^{n} x_i\right)^2 < n\sum_{i=1}^{n} x_i^2$. Do đó $D > 0$. Suy ra hệ trên có nghiệm duy nhất

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Nếu đặt

$$\overline{x} = \frac{1}{n}.\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}.\sum_{i=1}^{n} y_i, \quad \overline{xy} = \frac{1}{n}.\sum_{i=1}^{n} x_i y_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

thì nghiệm của hệ có thể viết lại dưới dạng

$$a = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2}; \qquad b = \frac{\overline{x^2}.\overline{y} - \overline{x}.\overline{xy}}{\overline{x^2} - (\overline{x})^2} = \frac{\overline{x^2}.\overline{y} - \overline{x}.\overline{xy}}{s_x^2}$$

Tóm lại, ta có thể tìm hàm $\overline{y_x} = ax + b$ từ các công thức

$$\boxed{\begin{array}{c} a = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2} = \dfrac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \\ b = \overline{y} - a.\overline{x}\end{array}}$$

⊙ **Chú ý**

-bb-error =

Đường gấp khúc nối các điểm (x_1, y_1), (x_2, y_2) , ... , (x_n, y_n) được gọi là *đường hồi qui thực nghiệm.*

Đường thẳng $y = ax + b$ nhận được bởi công thức bình phương bé nhất không đi qua được tất cả các điểm nhưng là đường thẳng "gần" các điểm đó nhất được gọi là *đường thẳng hồi qui* và thủ tục làm thích hợp đường thẳng thông qua các điểm dữ liệu cho trước được gọi là *hồi qui tuyến tính.*

Theo trên ta có $b = \overline{y} - a.\overline{x}$, do đó điểm $(\overline{x}, \overline{y})$ luôn nằm trên đường thẳng hồi qui.

• **Ví dụ 2** *Ước lượng hàm hồi qui tuyến tính mẫu xủa Y theo X trên cơ sở bảng tương quan cặp sau*

X	15	38	23	16	16	13	20	24
Y	145	228	150	130	160	114	142	265

Giải

Ta lập bảng sau

x_i	y_i	x_i^2	x_iy_i
15	145	225	3175
38	228	1444	8664
23	150	529	3450
16	130	256	2080
16	160	256	2560
13	114	169	1482
20	142	400	2840
24	265	576	6360
$\sum x = 165$	$\sum y = 1334$	$\sum x^2 = 3855$	$\sum xy = 29611$

Ta có

$$a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$= \frac{8(19611) - (165)(1334)}{8(3855)(165)^2} = \frac{16778}{3615} = 4,64$$

$$b = \overline{y} - a\overline{x} = \frac{1334}{8} - \left(\frac{16778}{3615}\right)\left(\frac{165}{8}\right) = 71$$

Vậy hàm hồi qui tuyến tính mẫu là $\overline{y_x} = 4,64x + 71$.

• **Ví dụ 3** *Độ ẩm của không khí ảnh hưởng đến sự bay hơi của nước trong sơn khi phun ra. Người ta tiến hành nghiên cứu mối liên hệ giữa độ ẩm của không khí X và độ bay hơi Y. Sự hiểu biết về mối quan hệ này sẽ giúp ta tiết kiệm được lượng sơn bằng cách chỉnh súng phun sơn một cách thích hợp. Tiến hành 25 quan sát ta được các số liệu sau:*

Quan sát	Độ ẩm (%)	Độ bay hơi (%)	Quan sát	Độ ẩm (%)	Độ bay hơi (%)
1	35,3	11,0	14	39,1	9,6
2	29,7	11,1	15	46,8	10,9
3	30,8	12,5	16	48,5	9,6
4	58,8	8,4	17	59,3	10,1
5	61,4	9,3	18	70,0	8,1
6	71,3	8,7	19	70,0	6,8
7	74,4	6,4	20	74,4	8,9
8	76,7	8,5	21	72,1	7,7
9	70,7	7,8	22	58,1	8,5
10	57,5	9,1	23	44,6	8,9
11	46,4	8,2	24	33,4	10,4
12	28,9	12,2	25	28,6	11,1
13	28,1	11,9			

Hãy tìm hàm hồi qui tuyến tính mẫu $\overline{y_x} = ax + b$.

Giải

Ta có

$$n = 25 \qquad \sum x = 1314,9 \qquad \sum y = 235,7$$

$$\sum x^2 = 76308,53 \qquad \sum y^2 = 2286,07$$

$$\sum xy = 11824,44$$

Do đó

$$a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{25 \times 11824,44 - (1314,9 \times 235,7)}{25 \times 76308,53 - (1314,9)^2} = -0,08$$

$$b = \overline{y} - a\overline{x} = 9,43 - (-0,08) \times 52,6 = 13,64$$

Vậy hàm hồi qui tuyến tính mẫu là $\overline{y_x} = -0,08x + 13,64$

b) Trường hợp nhiều số liệu (tương quan bảng)

Giả sử

X nhận các giá trị x_i với tần suất n_i $i = \overline{1,k}$,

Y nhận các giá trị y_j với tần suất m_j $j = \overline{1,h}$,

XY nhận các giá trị x_iy_j với tần suất n_{ij} $i = \overline{1,k}$, $j = \overline{1,h}$,

Ta tìm hồi qui tuyến tính mẫu $\overline{y_x} = ax + b$ trong trường hợp có nhiều số liệu. Theo (6.2) ta có

$$\begin{aligned}\left(\sum_{i=1}^{k} n_i x_i^2\right).a + \left(\sum_{i=1}^{k} n_i x_i\right).b &= \sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} x_i y_j \\ \left(\sum_{i=1}^{k} n_i x_i\right).a + nb &= \sum_{j=1}^{h} m_j y_j\end{aligned} \tag{6.3}$$

Thay $\sum_{i=1}^{k} n_i x_i = n\overline{x}$, $\sum_{j=1}^{h} m_j y_j = n\overline{y}$, $\sum_{i=1}^{k} n_i x_i^2 = n\overline{x^2}$, $\sum_{j=1}^{h} m_j y_j^2 = n\overline{y^2}$,

$\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} x_i y_j = n\overline{xy}$ vào (6.3) ta được

$$\begin{aligned}\overline{x^2}.a + \overline{x}.b &= \overline{xy} \quad (i) \\ \overline{x}.a + nb &= \overline{y} \quad (ii)\end{aligned}$$

Từ (ii) ta có $b = \overline{y} - a.\overline{x}$

Thay b vào $\overline{y_x} = ax + b$ ta suy ra

$$\overline{y_x} - \overline{y} = a(x - \overline{x}) \tag{6.4}$$

Ta tìm a bởi

$$a = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} x_i y_j - (\sum_{i=1}^{k} n_i x_i)(\sum_{j=1}^{h} m_j y_j)}{n\sum_{i=1}^{k} n_i x_i^2 - (\sum_{i=1}^{k} n_i x_i)^2} = \frac{n^2\overline{xy} - n\overline{x}.n\overline{y}}{n.n\overline{x^2} - (n\overline{x})^2}$$

$$= \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2}$$

Tóm lại, ta tìm hồi qui tuyến tính mẫu $\overline{y_x} = ax + b$ với $\boxed{a = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2},\ \ b = \overline{y} - a\overline{x}}$.

⊙ **Chú ý**

i) Ta biết hệ số tương quan $r_{XY} = \dfrac{\overline{xy} - \overline{xy}}{s_x.s_y}$ nên $a = r_{XY}\dfrac{s_y}{s_x}$

Thay a vào (6.4) ta có

$$\overline{y_x} - \overline{y} = r_{XY}\frac{s_y}{s_x}(x - \overline{x})$$

hay

$$\frac{\overline{y_x} - \overline{y}}{s_y} = r_{XY}\frac{(x - \overline{x})}{s_x}$$

Từ phương trình này ta có thể suy ra phương trình hồi qui tuyến tính mẫu $\overline{y_x} = ax + b$ một cách thuận lợi hơn vì thông qua việc tìm r_{XY} ta đã tính s_x, s_y.

ii) Khi các giá trị của X, Y khá lớn, ta có thể dùng phép đổi biến

$$u_i = \frac{x_i - x_0}{h_x} \quad (\forall i = \overline{1,k}); \qquad v_j = \frac{y_j - y_0}{h_y} \quad (\forall j = \overline{1,h})$$

trong đó

* x_0, y_0 là những giá trị tùy ý (thường chọn x_0, y_0 là giá trị của X, Y ứng với tần số n_{ij} lớn nhất trong bảng tương quan thực nghiệm),

* h_x, h_y là các giá trị tùy ý (thường chọn h_x, h_y là khoảng cách các giá trị kế tiếp nhau của X, Y).

Lập bảng tương quan đối với các biến mới U, V và tính toán các giá trị cần thiết ta tìm được hàm hồi qui tuyến tính mẫu

$$\overline{v_u} = a_0.u + b_0$$

trong đó

$$a_0 = \frac{\overline{uv} - \overline{u}.\overline{v}}{s_u^2}, \quad b_0 = \overline{v} - a_0.\overline{u}$$

Khi đó ta suy ra hàm $\overline{y_x} = ax + b$ với a, b được tìm bởi công thức

$$a = a_0\frac{h_y}{h_x}, \qquad b = y_0 + b_0.h_y - a_0.\frac{h_y}{h_x}.x_0$$

• **Ví dụ 4** *Xác định hệ số tương quan và hàm hồi qui tuyến tính mẫu* $\overline{y_x} = ax + b$ *của các đại lương ngẫu nhiên X và Y cho bởi bảng tương quan thực nghiệm sau:*

X / *Y*	*1*	*2*	*3*
10	*20*		
20		*30*	*1*
30		*1*	*48*

Giải

Ta lập bảng sau

X / Y	1	2	3	m_j	$m_j y_j$	$m_j y_j^2$
10	200 $\lvert\overline{20}$			20	200	2000
20		1200 $\lvert\overline{30}$	60 $\lvert\overline{1}$	31	620	12400
30		60 $\lvert\overline{1}$	4320 $\lvert\overline{48}$	49	1470	44100
n_i	20	31	49	n=100	$\sum y = 2290$	$\sum y^2 = 58500$
$n_i x_i$	20	62	147	$\sum x = 229$		
$n_i x_i^2$	20	124	441	$\sum x^2 = 585$		$\sum xy = 5840$

$$\sum xy = 200 + 1200 + 60 + 60 + 4320 = 5840$$

Phần trên góc trái của ô ghi các tích $n_{ij}x_iy_j$. Ta có

$$\overline{x} = \frac{229}{100} = 2,29; \qquad \overline{y} = \frac{2290}{100} = 22,9;$$

$$\overline{x^2} = \frac{585}{100} = 5,58; \qquad \overline{y^2} = \frac{58500}{100} = 585 \qquad \overline{xy} = \frac{5840}{100} = 58,4;$$

$$s_x^2 = \overline{x^2} - (\overline{x})^2 = 5,85 - (2,29)^2 \approx 0,6059 \quad \Longrightarrow \quad s_x \approx 0,78$$

$$s_y = \sqrt{\overline{y^2} - (\overline{y})^2} = \sqrt{585 - (22,9)^2} \approx 7,78$$

Do đó

$$a = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x^2} = \frac{58,4 - 2,29 \times 22,9}{0,6059} = 9,835$$

$$b = \overline{y} - a.\overline{x} = 22,9 - 9,835 \times 2,29 = 0,378$$

Hàm hồi qui tuyến tính mẫu là $\overline{y_x} = 9,835x + 0,378$

Hệ số tương quan là

$$r_{xy} = \frac{\overline{xy} - \overline{x}.\overline{y}}{s_x.s_y} = \frac{58,4 - 2,29 \times 22,9}{0,78 \times 7,78} \approx 0,982$$

4. BÀI TẬP

1. Cho các giá trị quan sát của hai đại lượng ngẫu nhiên X và Y ở bảng sau:

X	5	10	10	10	15	15	15	20	20	20
Y	20	20	30	30	30	40	50	50	60	60

   Giả sử X và Y có sự phụ thuộc tương quan tuyến tính. Tìm hàm hồi qui tuyến tính mẫu: $\overline{y}_x = ax + b$.

2. Người ta đo chiều dài vật đúc và khuôn thì thấy chúng lệch khỏi qui định nhưsau:

X	0.90	1,22	1,32	0,77	1,30	1,20	1,32	0,95	0,45	1,30	1,20
Y	-0,30	0,10	0,70	-0,28	0,25	0,02	0,37	-0,70	0,55	0,35	0,32

   Trong đó X, Y là các độ lệch.

   Xác định hệ số tương quan.

3. Số liệu thống kê nhằm nghiên cứu quan hệ giữa tổng sản phẩm nông nghiệp Y với tổng giá trị tài sản cố định X của 10 nông trại (tính trên 100 ha) như sau:

X	11,3	12,9	13,6	16,8	18,8	20,0	22,2	23,7	26,6	27,5
Y	13,2	15,6	17,2	18,8	20,2	23,9	22,4	23,0	24,4	24,6

Xác định đường hồi qui tuyến tính mẫu $\overline{y}_x = ax + b$. Sau đó tìm phương sai sai số thực nghiệm và khoảng tin cậy 95% cho hệ số góc của đường hồi qui trên.

4. Đo chiều cao X (cm) và trọng lượng Y (kg) của 100 học sinh, ta được kết quả sau:

X / Y	145 − 150	150 − 155	155 − 160	160 − 165	165 − 170
35 − 40	3				
40 − 45	5	10			
45 − 50		14	20	6	
50 − 55			15	12	5
55 − 60				6	4

Giả thuyết X và Y có mố phụ thuộc tương quan tuyến tính. Tìm các hàm hồi qui

a) $\overline{y}_x = ax + b$;

b) $\overline{x}_y = cy + d$

5. Theo dõi lượng phân bón và năng suất lúa của 100 hecta lúa ở một vùng, ta thu được bảng số liệu sau:

X / Y	120	140	160	180	200
2,2	2				
2,6	5	3			
3,0		11	8	4	
3,4			15	17	
3,8			10	6	7
4,2					12

Trong đó X là phân bón (kg/ha) và Y là năng suất lúa (tấn/ha).

a) Hãy ước lượng hệ số tương quan tuyến tính r.
b) Tìm phương trình tương quan tuyến tính: $\overline{y}_x = ax + b$.

6. Đo chiều cao và đường kính của một loại cây, ta được kết quả cho bở bảng sau:

X / Y	6	8	10	12	14
30	2	17	9	3	
35		10	17	9	
40		3	24	16	13
45			6	24	12
50			2	11	22

Trong đó X là đường kính (cm) và Y là chiều cao (m).

a) Xác định hệ số tương quan tuyến tính mẫu r.
b) Tìm các phương trình hồi qui tuyến tính mẫu.
c) Các phương trình trên sẽ thay đổi như thế nào nếu X được tính theo đơn vị là mét (m)?

TRẢ LỜI BÀI TẬP

1. $\overline{x} = 14$, $\overline{y} = 39$, $\overline{y}_x = \frac{8}{3}x + \frac{5}{3}$.

2. $r = -0,3096$.

3. $\overline{y}_x = 0,67x + 7,18$, $\sigma^2 = 1,126$, $(0,6280\,;\,0,7176)$.

4. a) $\overline{y}_x = 0,7018x - 61,5537$, b) $\overline{x}_y = 0,91y + 112,96$.

5. $r = 0,8165$; $\overline{y}_x = 0,017x + 0,5622$.

6. a) $r = 0,69$, b) $\overline{y}_x = 0,218x + 2,434$, $\overline{x}_y = 2,18y + 15,87$.

c) $\overline{y}_{x'} = 21,8x' + 2,434$, $\overline{x}_y = 0,0218y' + 0,1587$.

Chủởng 7

KIỂM TRA CHẤT LƯỢNG SẢN PHẨM

Trong mỗi quá trình sản xuất thuờng có sụ thay đổi giữa các sản phẩm gây ra tác động xấu lên chất lượng của sản phẩm. Sụ thay đổi này có thể đuợc gây nên bởi sụ sui hư hỏng của máy móc, chất lượng xấu của nguyên liệu thô cung cấp cho sản xuất, phần mềm quản lý không chính xác hoặc do sai lầm của con ngườ̀i khi điều khiển quá trình.

Việc nhận biết khi nào thì quá trình đi ra ngoài sụ kiểm soát đuợc xác định bởi *biểu đồ kiểm soát*. Biểu đồ này đuợc xác định bởi hai giá trị: giới hạn kiểm soát duới LCL (*lower control limit*) và giới hạn kiểm soát trên UCL (*upper control limit*). Dữ liệu sản xuất đuợc chia thành nhữ̃ng nhóm con và thống kê của nhóm con, nhu trung bình nhóm con và độ lệch tiêu chuẩn nhóm con. Khi thống kê nhóm con không rơi vào giữa giới hạn kiểm soát duới và giới hạn kiểm soát trên thì ta kết luận *quá trình đi ra ngoài kiểm soát*.

1. BIỂU ĐỒ KIỂM SOÁT CHO GIÁ TRỊ TRUNG BÌNH

1.1 Trùờng hợp biết μ và σ

Giả sủ khi quá trình trong sụ kiểm soát các sản phẩm liên tiếp đuợc sản xuất ra có các đặc trung số đo đuợc là đại lượng ngẫu nhiên chuẩn, độc lập với trung bình μ và phuơng sai σ^2. Tuy nhiên, vì một tình huống đặc biệt nào đó quá trình đi ra ngoài sụ kiểm soát và bắt đầu sản xuất ra sản phẩm có phân phối khác. Ta cần nhận biết khi nào thì điều này xảy ra để ngừng quá trình, tìm ra sụ cố và khắc phục nó.

Giả sủ $X_1, X_2, \ldots$ là các đặc trung đo đuợc của các sản phẩm liên tiếp. Ta chia dữ liệu ra thành các nhóm con có kích thuớc n xác định. Giá trị n đuợc chọn sao cho trong mỗi nhóm con sản phẩm cótính chất nhu nhau. Chẳng hạn, n có thể đuợc chọn sao cho tất cả sản phẩm bên trong một nhóm con đuợc sản xuất trong cùng một ngày, hoặc cùng một ca, hoặc cùng một cách sắp đặt,...Các giá trị tiêu biểu của n là $4, 5$ hoặc 6.

Gọi $\overline{X}_i$, $i = 1, 2, \ldots$ là giá trị trung bình của nhóm thứ i. Tức là

$$\overline{X}_1 = \frac{X_1 + \ldots + X_n}{n}$$

$$\overline{X}_2 = \frac{X_{n+1} + \ldots + X_{2n}}{n}$$

$$\overline{X}_3 = \frac{X_{2n+1} + \ldots + X_{3n}}{n}$$

Vì khi trong sụ kiểm soát, mỗi X_i có trung bình μ và phương sai σ^2 nên

$$E(\overline{X}_i) = \mu, \qquad Var(\overline{X}_i) = \frac{\sigma^2}{n}$$

Do đó $\dfrac{\overline{X}_i - \mu}{\sqrt{\frac{\sigma^2}{n}}}$ có phân phối chuẩn hóa.

Ta biết một đại lượng ngẫu nhiên Z có phân phối chuẩn hóa hầu như nhận giá trị giữa -3 và 3 (vì $P(-3 < Z < 3) = 0,9973$).

Do đó

$$-3 < \sqrt{n}\frac{\overline{X}_i - \mu}{\sigma} < 3$$

hay

$$\mu - \frac{3\sigma}{\sqrt{n}} < \overline{X}_i < \mu + \frac{3\sigma}{\sqrt{n}}$$

Giá trị

$$LCL \equiv \mu - \frac{3\sigma}{\sqrt{n}} \qquad \text{và} \qquad UCL \equiv \mu + \frac{3\sigma}{\sqrt{n}}$$

được gọi là *giới hạn kiểm soát dưới và giới hạn kiểm soát trên.*

Biểu đồ *kiểm soát–*$\overline{X}$ được tạo nên để nhận biết sụ thay đổi của hàng hóa được sản xuất, và nhận được bằng cách đưa vào các trung bình nhóm con liên tiếp $\overline{X}_i$. Biểu đồ cho biết quá trình đi ra ngoài sụ kiểm soát ở lần đầu tiên X_i không rơi vào giữa LCL và UCL.

• **Ví dụ 1** *Một nhà máy sản xuất một chi tiết máy bằng thép có đường kính là đại lượng ngẫu nhiên có phân phối chuẩn với trung bình* $3mm$ *và độ lệch tiêu chuẩn* $0,1mm$*. Các mẫu liên tiếp của 4 chi tiết có trung bình mẫu tính bằng milimet như sau:*

Mẫu	$\overline{X}$	*Mẫu*	$\overline{X}$
1	3,01	6	3,02
2	2,97	7	3,10
3	3,12	8	3,14
4	2,99	9	3,09
5	3,03	10	3,20

Hãy kết luận về sự kiểm soát của quá trình.

Giải

Khi trong sụ kiểm soát các đường kính của các chi tiết liên tiếp có trung bình $\mu = 3$ và độ lệch tiêu chuẩn $\sigma = 0,1$. Với $n = 4$ thì các giới hạn kiểm soát là

$$LCL = 3 - \frac{3.1}{4} = 2,85, \qquad UCL = 3 + \frac{3.1}{4} = 3,15$$

Tù mẫu số 6 đến mẫu số 10 cho thấy đường kính của chi tiết máy có xu hướng tăng và ở mẫu số 10 thì đường kính ở phía trên giới hạn kiểm soát trên. Điều này cho ta nhận thấy bắt đầu tù mẫu số 10 quá trình ra ngoài sụ kiểm soát và đường kính trung bình của chi tiết máy bắt đầu khác $3mm$.

⊙ **Chú ý** Giả sử quá trình vùa ra ngoài sụ kiểm soát bởi sụ thay đổi giá trị trung bình của sản phẩm tù μ tới $\mu + a$ với $a > 0$. Phải mất bao lâu tới khi biểu đồ nhận thấy quá trình đi ra ngoài kiểm soát?

Ta thấy trung bình của nhóm con ở trong giới hạn kiểm soát nếu

$$-3 < \sqrt{n}\frac{\overline{X} - \mu}{\sigma} < 3$$

$$\iff \qquad -3 - \frac{a\sqrt{n}}{\sigma} < \sqrt{n}\frac{\overline{X} - \mu}{\sigma} - \frac{a\sqrt{n}}{\sigma} < 3 - \frac{a\sqrt{n}}{\sigma}$$

hay

$$-3 - \frac{a\sqrt{n}}{\sigma} < \sqrt{n}\frac{\overline{X} - \mu - a}{\sigma} < 3 - \frac{a\sqrt{n}}{\sigma}$$

Vì $\overline{X}$có phân phối chuẩn với trung bình $\mu + a$ và phưởng sai $\frac{\sigma^2}{n}$ nên $\sqrt{n}\frac{\overline{X} - \mu - a}{\sigma}$ có phân phối chuẩn hóa. Xác suất để nó rơi và giới hạn kiểm soát là

$$P\left(-3 - \frac{a\sqrt{n}}{\sigma} < Z < 3 - \frac{a\sqrt{n}}{\sigma}\right) = \phi\left(3 - \frac{a\sqrt{n}}{\sigma}\right) - \phi\left(-3 - \frac{a\sqrt{n}}{\sigma}\right) \approx \phi\left(3 - \frac{a\sqrt{n}}{\sigma}\right)$$

Do đó xác suất để nó rơi ra ngoài xấp xỉ $1 - \phi(3 - \frac{a\sqrt{n}}{\sigma})$.

1.2 Trường hợp chủa biết μ và σ

Ta sẽ ước lượng μ và σ bằng cách chọn k nhóm con với $k \geq 20$ và $nk \geq 100$.

Nếu $\overline{X}_i,\ i = 1, 2, \ldots, k$ là trung bình của nhóm con thứ i thì ta ước lượng μ bởi

$$\overline{\overline{X}} = \frac{\overline{X}_1 + \ldots + \overline{X}_k}{k}$$

Để ước lượng σ ta gọi S_i là độ lệch tiêu chuẩn mẫu của nhóm thứ i $(i = 1, 2, \ldots, k)$, tức là

$$S_1 = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \overline{X}_1)^2}{n-1}}$$

$$S_2 = \sqrt{\sum_{i=1}^{n} \frac{(X_{n+i} - \overline{X}_2)^2}{n-1}}$$

$$\vdots$$

$$S_k = \sqrt{\sum_{i=1}^{n} \frac{(X_{(k-1)n+i} - \overline{X}_k)^2}{n-1}}$$

Đặt

$$\overline{S} = \frac{S_1 + \ldots + S_k}{k}$$

Thống kê $\overline{S}$ không là ước lượng không chệch của σ vì $E(\overline{S}) \neq \sigma$. Để chuyển nó thành ước lượng không chệch cần phải tính $E(\overline{S})$. Ta có

$$E(\overline{S}) = \frac{E(S_1) + \ldots + E(S_k)}{k} = E(S_1) \tag{7.1}$$

(do $S_1, \ldots, S_k$ độc lập và có phân phối đồng nhất nên có cùng giá trị trung bình).

Để tính $E(S_1)$ ta dùng các kết quả sau:

* *Kết quả 1:*

$$\frac{(n-1)S_1^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \in \chi^2_{n-1} \tag{7.2}$$

* *Kết quả 2:* Với $Y \in \chi^2_{n-1}$ thì

$$E(Y) = \sqrt{2}\frac{\Gamma(\frac{n}{2})}{\Gamma\frac{n-1}{2}} \tag{7.3}$$

Ta có

$$E(Y) = \int_0^{+\infty} \sqrt{y} f_{\chi^2_{n-1}}(y)dy = \int_0^{+\infty} \frac{e^{-\frac{y}{2}}.y^{\frac{n-1}{2}-1}}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}dy = \int_0^{+\infty} \frac{e^{-\frac{y}{2}}.y^{\frac{n}{2}-1}}{2^{\frac{n-1}{2}}.\Gamma(\frac{n-1}{2})}dy$$

Đặt $x = \dfrac{y}{2}$ thì $E(Y) = \sqrt{2}\dfrac{\Gamma(\frac{n}{2})}{\Gamma\frac{n-1}{2}}$.

Vì $\left(\sqrt{\dfrac{(n-1)S_1^2}{\sigma^2}}\right) = \sqrt{n-1}\dfrac{E(S_1)}{\sigma}$ nên từ (7.2) và (7.3) ta có

$$E(S_1) = \frac{\sqrt{2}\Gamma(\frac{n}{2})\sigma}{\sqrt{n-1}\Gamma(\frac{n-1}{2})}$$

Đặt

$$c(n) = \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\sqrt{n-1}\Gamma(\frac{n-1}{2})}$$

Bảng giá trị của $c(n)$
c(2)=0,7978849
c(3)=0,8862266
c(4)=0,9213181
c(5)=0,9399851
c(6)=0,9515332
c(7)=0,9593684
c(8)=0,9650309
c(9)=0,9693103
c(10)=0,9726596

thì theo (7.1) ta thấy $\dfrac{\overline{S}}{c(n)}$ là ước lượng không chệch của σ.

Ước lượng cho μ và σ ở trên chỉ hợp lý nếu quá trình trong sự kiểm soát.

Các giới hạn kiểm soát trong trường hợp này là

$$LCL = \overline{\overline{X}} - \frac{3\overline{S}}{\sqrt{n}c(n)} \qquad UCL = \overline{\overline{X}} + \frac{3\overline{S}}{\sqrt{n}c(n)}$$

Ta sẽ thực hiện việc kiểm tra trung bình của các nhóm con. Nếu nhóm con nào mà giá trị trung bình không rơi vào giữa các giới hạn kiểm soát thì ta loại ra và thực hiện ước lượng lại. Tiếp tục kiểm tra lần nữa sao cho giá trị trung bình của các nhóm con rơi vào giữa các giới hạn kiểm soát. Nếu có quá nhiều giá trị trung bình của các nhóm con rơi ra ngoài các giới hạn kiểm soát thì rõ ràng sự kiểm soát không được thiết lập.

• **Ví dụ 2** *Xét lại ví dụ (1) dưới giả thiết mới rằng quá trình mới bắt đầu với μ và σ chưa biết. Giả sử độ lệch tiêu chuẩn được cho:*

	$\overline{X}$	S		$\overline{X}$	S
1	3,01	0,12	6	3,02	0,08
2	2,97	0,14	7	3,10	0,15
3	3,12	0,08	8	3,14	0,16
4	2,99	0,11	9	3,09	0,13
5	3,03	0,09	10	3,20	0,16

Vì $\overline{\overline{X}} = 3,067,\ S = 0,122,\ c(4) = 0,9213$ nên các giới hạn kiểm soát là

$$LCL = 3,067 - \frac{3 \times 0,122}{2 \times 0,9213} = 2,868$$

$$UCL = 3,067 + \frac{3 \times 0,122}{2 \times 0,9213} = 3,266$$

Ta thấy tất cả $\overline{X}_i$ đều rơi vào giữa các giới hạn kiểm soát nên có thể xem quá trình trong sự kiểm soát với $\mu = 3,067$ và $\sigma = \frac{\overline{S}}{c(4)} = 0,1324$.

Bây giờ giả sử quá trình vẫn duy trì trong sự kiểm soát và các ước lượng của μ và σ là đúng. Vấn đề đặt ra là xác định tỷ lệ sản phẩm rơi vào $3 \pm 0,1$.

Khi $\mu = 3,067$ và $\sigma = 0,1324$ ta có

$$\begin{aligned} P(2,9 \le X \le 3,1) &= P(\frac{2,9-3,067}{0,1324} \le \frac{X-3,067}{0,1324} \le \frac{3,1-3,067}{0,1324}) \\ &= \Phi(0,2492) - \Phi(-1,2613) \\ &= 0{,}5984 \text{ -(1-0,8964)} \\ &= 0{,}4948 \end{aligned}$$

Vậy 49% các sản phẩm rơi vào $3 \pm 0,1$.

2. BIỂU ĐỒ KIỂM SOÁT S

Trong phần này ta xây dựng biểu đồ kiểm soát sự thay đổi phương sai của tổng thể.

Giả sử khi trong sự kiểm soát, các sản phẩm được tạo ra có đặc trưng đo được là đại lượng ngẫu nhiên có phân phối chuẩn với trung bình μ và phương sai σ^2. Nếu S_i là độ lệch tiêu chuẩn mẫu của nhóm con thứ i thì

$$S_i = \sqrt{\sum_{j=1}^{n} \frac{(X_{(i-1)n+j} - \overline{X}_i)^2}{n-1}}$$

thì theo mục **1.** ta có

$$E(S_i) = c(n)\sigma \tag{7.4}$$

và

$$\begin{aligned} Var(S_i) &= E(S_i^2) - [E(S_i)]^2 && (7.5) \\ &= \sigma^2 - c^2(n)\sigma^2 && (7.6) \\ &= \sigma^2[1 - c^2(n)] && (7.7) \end{aligned}$$

(7.7) có tỉ lệ (7.2) và dựa vào tính chất kỳ vọng của đại lượng ngẫu nhiên có phân phối "khi−bình phương" thì bằng với bậc tự do của nó.

Khi trong sự điều khiển S_i có phân phối của một hằng (bằng $\frac{\sigma}{\sqrt{n-1}}$) nhân với căn bậc hai của đại lượng ngẫu nhiên có phân phối "khi−bình phương" với $n-1$ bậc tự do. Có thể thấy S_i ở trong độ lệch tiêu chuẩn 3 của kỳ vọng của nó với xác suất gần bằng 1.

$$P\left(E(S_i) - 3\sqrt{Var(S_i)} < S_i < E(S_i) + 3\sqrt{Var(S_i)}\right) \approx 0,99$$

Dùng công thức (7.4) và (7.5) cho $E(S_i)$ và $Var(S_i)$ thì ta có giới hạn kiểm soát dưới và giới hạn kiểm soát trên của biểu đồ S là

$$LCL = \sigma[c(n) - 3\sqrt{1 - c^2(n)}]$$

$$UCL = \sigma[c(n) + 3\sqrt{1 - c^2(n)}]$$

Các gía trị liên tiếp của S_i được đưa vào đảm bảo chúng rơi vào giữa giới hạn kiểm soát dưới và giới hạn kiểm soát trên. Khi một giá trị rơi ra ngoài, quá trình phải dừng và được khai báo ra ngoài sự kiểm soát.

⊙ **Chú ý** Khi σ chưa biết, ta có thể ước lượng σ từ $\frac{\overline{S}}{c(n)}$. Tương tự như trên, ta có thể ước lượng các giới các giới hạn kiểm soát

$$LCL = \overline{S}\left[1 - 3\sqrt{\frac{1}{c^2(n)} - 1}\right]$$

$$UCL = \overline{S}\left[1 + 3\sqrt{\frac{1}{c^2(n)} - 1}\right]$$

Khi lập biểu đồ kiểm soát $\overline{X}$, phải kiểm tra rằng k độ lệch tiêu chuẩn $S_1, S_2, \ldots, S_k$ của các nhóm con phải rơi vào trong các giới hạn kiểm soát. Nếu giá trị nào trong chúng rơi ra ngoài thì loại bỏ nhóm con đó và tính lại $\overline{S}$.

• **Ví dụ 3** *Các giá trị của $\overline{X}$ và S của 20 nhóm con kích thước 5 của quá trình mới bắt đầu cho bởi*

Nhóm con	$\overline{X}$	S	Nhóm con	$\overline{X}$	S	Nhóm con	$\overline{X}$	S
1	35,1	4,2	8	38,4	5,1	15	43,2	3,5
2	33,2	4,4	9	35,7	3,8	16	41,3	8,2
3	31,7	2,5	10	27,2	6,2	17	35,7	8,1
4	35,4	3,2	11	38,1	4,2	18	36,3	4,2
5	34,5	2,6	12	37,6	3,9	19	35,4	4,1
6	36,4	4,5	13	38,8	3,2	20	34,6	3,7
7	35,9	3,4	14	34,3	4,0			

Vì $\overline{\overline{X}} = 35,94$, $\overline{S} = 4,35$, $c(5) = 0,9400$ nên giới hạn kiểm soát dưới và giới hạn kiểm soát trên của $\overline{X}$ và S là

$$LCL(\overline{X}) = 29,731; \qquad UCL(\overline{X}) = 42,149$$

$$LCL(S) = -0,386; \qquad UCL(S) = 9,087$$

Biểu đồ S

Biểu đồ $\overline{X}$

Ta thấy $\overline{X}_{10}$ và $\overline{X}_{15}$ rơi ra ngoài giới hạn kiểm soát của $\overline{X}$ nên các nhóm con này

phải được loại ra và các giới hạn kiểm soát phải được tính lại. Việc tính lại xem như bài tập, các bạn tự giải.

3. BIỂU ĐỒ KIỂM SOÁT CHO TỶ LỆ KHIẾM KHUYẾT

Biểu đồ kiểm soát $\overline{X}$ và S được dùng khi dữ liệu là các đại lượng đo được. Có trường hợp sản phẩm được sản xuất có đặc trưng về chất (tính chất nào đó) được phân loại không xảy ra (*ta gọi là khuyết*) hoặc xảy ra. Biểu đồ kiểm soát cũng được dùng cho trường hợp này.

Giả sử khi quá trình trong trong sự kiểm soát mỗi sản phẩm được tạo ra *khuyết* một cách độc lập với xác suất p.

Nếu gọi X là số sản phẩm *khuyết* trong một nhóm con kích thước n thì X là đại lượng ngẫu nhiên có phân phối nhị thức với tham số n và p.

Nếu $F = \dfrac{X}{n}$ là tỷ số của nhóm con bị *khuyết* thì trung bình và độ lệch tiêu chuẩn của nó được cho bởi

$$E(F) = \frac{E(X)}{n} = \frac{np}{n} = p$$

$$\sqrt{Var(F)} = \sqrt{\frac{Var(X)}{n^2}} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}$$

Do đó khi quá trình trong sự kiểm soát tỷ lệ *khuyết* trong một nhóm con của n sản phẩm có xác suất nằm giữa các giới hạn

$$LCL = p - 3\sqrt{\frac{p(1-p)}{n}}; \qquad UCL = p + \sqrt{\frac{p(1-p)}{n}}$$

⊙ **Chú ý** Kích thước n của nhóm nhóm con thường lớn hơn nhiều so với các giá trị tiêu biểu từ 4 đến 10 được dùng trong biểu đồ kiểm soát $\overline{X}$ và S. Lý do chính của điều này là nếu p nhỏ và n là kích thước không hợp lý thì hầu hết các nhóm con sẽ có *khuyết* zero thậm chí khi quá trình ra ngoài sự kiểm soát. Vì vậy n phải được chọn lớn hơn sao cho np không gần 0 để có thể nhận ra sự thay đổi chất lượng của sản phẩm.

Để bắt đầu biểu đồ kiểm soát như vật trước hết phải ước lượng p. Ta chọn k nhóm con với $k \geq 20$ và gọi F_i là tỷ số của nhóm thứ i bị *khuyết*. Ước lượng của p cho bởi

$$\overline{F} = \frac{F_1 + \ldots + F_k}{k}$$

Vì nF_i bằng số của các *khuyết* trong nhóm i nên có thể xem

$$\overline{F} = \frac{nF_1 + \ldots + nF_k}{k} = \frac{\textit{tổng số các khuyết trong tất cả các nhóm con}}{\textit{số sản phẩm trong các nhóm con}}$$

Giới hạn kiểm soát dưới và giới hạn kiểm soát trên cho bởi

$$LCL = \overline{F} - 3\sqrt{\frac{\overline{F}(1-\overline{F})}{n}}; \qquad UCL = \overline{F} + 3\sqrt{\frac{\overline{F}(1-\overline{F})}{n}}$$

Bây giờ ta kiểm tra xem tỷ số nhóm con $F_1, F_2, \ldots, F_k$ có rơi vào giữa các giới hạn kiểm soát không? Nếu giá trị nào rơi ra ngoài thì nhóm con tương ứng với nó sẽ bị loại bỏ và $\overline{F}$ được tính lại.

• **Ví dụ 4** *Các mẫu liên tiếp của 50 đinh ốc được lấy ra từ một máy sản xuất đinh ốc tự động. Mỗi đinh ốc có tính chất nào đómà ta quan tâm nó xảy ra hoặc không xảy ra khuyết. Quan sát tính chất trên 20 sản phẩm ta có kết quả sau:*

Nhóm con	Khuyết	F	Nhóm con	Khuyết	F
1	6	0.12	11	1	0.02
2	5	0.10	12	3	0.06
3	3	0.06	13	2	0.04
4	0	0.00	14	0	0.00
5	1	0.02	15	1	0.02
6	2	0.04	16	1	0.02
7	1	0.02	17	0	0.00
8	0	0.00	18	2	0.04
9	2	0.04	19	1	0.02
10	1	0.02	20	2	0.04

Ta có

$$\overline{F} = \frac{\textit{Tổng các khuyết}}{\textit{Tổng các sản phẩm}} = \frac{34}{1000} = 0,034$$

Do đó

$$LCL = 0,034 - 3\sqrt{\frac{0,034.0.966}{50}} = -0,0429$$

$$UCL = 0,034 + 3\sqrt{\frac{0,034.0,966}{50}} = 0,1109$$

Vì tỷ số các *khuyết* trong nhóm đầu tiên rơi ra ngoài giới hạn trên nên ta loại nhóm con này ra và tính lại $\overline{F}$ nhưsau:

$$\overline{F} = \frac{34-6}{950} = 0,0295$$

Các giới hạn kiểm soát mới là

$$LCL = 0,0295 - \sqrt{\frac{0,0295(1-0,0295)}{50}} = -0,0423$$

$$UCL = 0,0295 + 3\sqrt{\frac{0,0295(1-0,0295)}{50}} = 0,1013$$

Ta thấy các nhóm con còn lại có tỷ số các khuyết rơi vào trong các giới hạn kiểm soát. Ta thừa nhận rằng khi trong sự kiểm soát tỷ số các sản phẩm bị khuyết trong một nhóm con phải dưới 0,1013.

4. BIỂU ĐỒ SỐ CÁC KHUYẾT

Trong phần này ta xét trường hợp dữ liệu bao gồm số các khuyết trong một đơn vị chứa một sản phẩm hoặc một nhóm các sản phẩm. Ví dụ số các đinh ốc bị khuyết trong một cánh máy bay hoặc số các *chip* máy tính bị khuyết được sản xuất của một nhà máy. Trường hợp thông thường có một số lớn các sản phẩm bị khuyết, trong đó mỗi sản phẩm bị khuyết với xác suất nhỏ. Do dó ta có thể xem khi quá trình trong sự kiểm soát thì số các khuyết có phân phối Poisson với trung bình λ.

Gọi X_i là số các khuyết trong đơn vị thứ i. Vì phương sai của đại lượng ngẫu nhiên có phân phối Poisson bằng với trung bình của nó nên

$$E(X_i) = \lambda, \qquad Var(X_i) = \lambda$$

Do đó khi trong sự kiểm soát mỗi X_i có xác suất cao trong $\lambda \pm 3\sqrt{\lambda}$. Vì vậy giới hạn kiểm soát dưới và giới hạn kiểm soát trên cho bởi

$$LCL = \lambda - 3\sqrt{\lambda}; \qquad UCL = \lambda + 3\sqrt{\lambda}$$

⊙ **Chú ý** Giống như phần trước, khi biểu đồ kiểm soát bắt đầu mà λ chưa biết ta chọn một mẫu của k đơn vị và ước lượng λ bởi

$$\overline{X} = \frac{X_1 + \ldots + X_k}{k}$$

Ta được các giới hạn kiểm soát dưới và trên

$$\overline{X} - 3\sqrt{\overline{X}}; \qquad \overline{X} + 3\sqrt{\overline{X}}$$

Nếu tất cả X_i, $i = 1, \ldots, k$ rơi vào phía trong các giới hạn này ta giả thiết quá trình trong sự kiểm soát với $\lambda = \overline{X}$. Nếu một vài giá trị rơi ra ngoài thì các giá trị này bị loại bỏ và ta tính lại $\overline{X}$.

Trong trường hợp số các khuyết trung bình trên sản phẩm nhỏ, ta kết hợp các sản phẩm lại và dùng như dữ liệu số các khuyết n đã cho. Vì tổng của các đại lượng ngẫu nhiên có phân phối Poisson cũng là đại lượng ngẫu nhiên có phân phối Poisson với cùng trung bình λ Sự kết hợp các sản phẩm như vậy có tác dụng khi số các khuyết trên sản phẩm ít hơn 25.

• **Ví dụ 5** *Số các khuyết sau được phát hiện tại một nhà máy trên các đơn vị của mỗi 10 ô tô:*

Ô tô	Các khuyết	Ô tô	Các khuyết	Ôtô	Các khuyết
1	141	8	95	15	94
2	162	9	76	16	68
3	150	10	68	17	95
4	111	11	63	18	81
5	92	12	74	19	102
6	74	13	103	20	73
7	85	14	81		

Xét xem quá trình sản xuất có trong sự kiểm soát không?

Giải

Ta có $\overline{X} = 94,4$ nên các giới hạn kiểm soát thủ là

$$LCL = 94,4 - 3\sqrt{94,4} = 65,25$$

$$UCL = 94,4 + 3\sqrt{94,4} = 123,55$$

Vì ba giá trị dữ liệu đầu tiên lớn hơn UCL nên chúng bị loại đi và trung bình mẫu được tính lại. Ta được

$$\overline{X} = \frac{94,4.20 - (141 + 162 + 150)}{17} = 84,41$$

và các giới hạn kiểm soát thủ mới là

$$LCL = 84,41 - 3\sqrt{84,41} = 56,85$$

$$UCL = 84,41 + 3\sqrt{84,41} = 111,97$$

Ta thấy 17 giá trị dữ liệu còn lại rơi vào trong các giới hạn kiểm soát. Do đó có thể nói rằng bây giờ quá trình trong sự kiểm soát với giá trị trung bình $84,41$. Tuy nhiên dường như các giá trị trung bình của các khuyết cao tù ban đầu trước khi ổn định để đi vào sự kiểm soát, dường như có vẻ tin tưởng rằng giá trị dữ liệu X_4 cũng cao trước khi đi vào sự kiểm soát. Trong trường hợp này,đẻ thận trọng ta loại bỏ X_4 và tính lại. Dựa vào việc tính lại 16 dữ liệu này ta nhận được

$$\overline{X} = 82,56$$

$$LCL = 82,56 - 3\sqrt{82,56} = 55,30$$

$$UCL = 82,56 + 3\sqrt{82,56} = 109,82$$

và quá trình trong sự kiểm soát với giá trị trung bình 82,56.

5. BÀI TẬP

1. Xét các dữ liệu về giá của 10 mẫu cho dưới đây

Mẫu	Giá			
1	10,6	10,1	11,3	9,1
2	10,2	11,6	10,5	10,5
3	10,1	9,8	8,8	9,3
4	10,1	9,5	10,3	10,6
5	8,7	11,6	9,7	9,3
6	10,1	9,8	10,8	8,9
7	11,2	11,5	10,9	11,6
8	10,6	9,6	10,3	9,9
9	9,8	7,7	9,4	9,9
10	10,0	8,4	10,6	8,8

Hãy tìm giới hạn kiểm soát trên và giới hạn kiểm soát dưới cho $\overline{X}$.

2. Giả sử các sản phẩm được sản xuất có phân phối chuẩn với trung bình 35 và độ lệch tiêu chuẩn 3. Để giám sát quá trình ta chọn mẫu các nhóm con kích thước 5. Trung bình của 20 nhóm con đầu tiên cho bởi bảng sau:

Số nhóm con	$\overline{X}$	Số nhóm con	$\overline{X}$
1	34,0	11	35,8
2	31,6	12	35,8
3	30,8	13	34,0
4	33,0	14	35,0
5	35,0	15	33,8
6	32,2	16	31,6
7	33,0	17	33,0
8	32,6	18	33,2
9	33,8	19	31,8
10	35,8	20	35,6

Hỏi quá trình có trong sự kiểm soát hay không?

3. Các giá trị của $\overline{X}$ và S đối với 20 nhóm con kích thước 5 cho bởi bảng sau

Nhóm con	$\overline{X}$	S	Nhóm con	$\overline{X}$	S
1	33,8	5,1	11	29,7	5,1
2	37,2	5,4	12	31,6	5,3
3	40,4	6,1	13	38,4	5,8
4	39,3	5,5	14	40,2	6,4
5	41,1	5,2	15	35,6	4,8
6	40,4	4,8	16	36,4	4,6
7	35,0	5,0	17	37,2	6,1
8	36,1	4,1	18	31,3	5,7
9	38,2	7,3	19	33,6	5,5
10	32,4	6,6	20	36,7	4,2

a) Xác định các giới hạn kiểm soát cho $\overline{X}$.

b) Xác định các giới hạn kiểm soát cho S.

4. Dữ liệu sau giới thiệu số khiếm khuyết của "con chip" điện tử được sản xuất trong 15 ngày gần đây: 121, 133, 98, 85, 101, 78, 66, 82, 90, 78, 85, 81, 100, 75, 89. Hãy kết luận xem quá trình có trong sự kiểm soát hay không? Hãy chỉ ra các giới hạn kiểm soát cho các sản phẩm trong tương lai?

◙ TRẢ LỜI BÀI TẬP

1. 8,8292 ; 11,2458.

2. Không.

4. LCL = 57,5 ; UCL= 112,9.