TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BỘ MÔN KHOA HỌC TỰ NHIÊN, KHOA KHOA HỌC CƠ BẢN

XÁC SUẤT THỐNG KÊ

(TÀI LIỆU LƯU HÀNH NỘI BỘ)

THÁI NGUYÊN- 2022

Mục lục

Lời nói đầu

Lý thuyết xác suất và thống kê toán là một ngành khoa học ngày càng giữ vị trí quan trọng trong các lĩnh vực khoa học cũng như trong cuộc sống con người. Lý thuyết xác suất nghiên cứu những qui luật ngẫu nhiên của các hiện tượng số lớn, việc nắm bắt các qui luật này sẽ cho phép dự báo các hiện tượng ngẫu nhiên đó sẽ xảy ra như thế nào. Dựa vào các thành tựu của lý thuyết xác suất, thống kê toán xây dựng các phương pháp thu thập và xử lý các số liệu thống kê nhằm rút ra các kết luận khoa học và thực tiễn.

Với mục đích cung cấp tài liệu học tập cũng như tài liệu tham khảo phù hợp cho sinh viên ngành kỹ thuật mà cụ thể là sinh viên trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên, bài giảng được biên soạn theo quan điểm ứng dụng dựa trên chương trình chi tiết của học phần Xác suất thống kê. Mỗi khái niệm, vấn đề hay phương pháp đều được minh họa bằng nhiều ví dụ tương ứng. Lượng kiến thức được trình bày gọn nhẹ không quá phức tạp, với các bài tập đa dạng giúp sinh viên luyện tập và củng cố lý thuyết.

Bài giảng chia thành 5 chương:

Chương 1: Các khái niệm cơ bản về xác suất.

Chương 2: Đại lượng ngẫu nhiên.

Chương 3: Lý thuyết mẫu và ước lượng.

Chương 4: Kiểm định giả thiết thống kê.

Chương 5: Phân tích tương quan và hồi quy.

Trong quá trình biên soạn chúng tôi đã nhận được nhiều ý kiến đóng góp quí báu của các đồng nghiệp thuộc Bộ môn Khoa học tự nhiên trường ĐH CNTT và TT, chúng tôi xin chân thành cảm ơn. Tuy vậy chắc chắn cuốn bài giảng này không tránh khỏi những hạn chế và thiếu sót. Chúng tôi rất mong nhận được các ý kiến góp ý để tiếp tục hoàn thiện nội dung cuốn bài giảng cho phù hợp với các ngành đào tạo trong trường ĐH CNTT và TT.

Chúng tôi xin chân thành cảm ơn!

Thái Nguyên, ngày 02 tháng 03 năm 2022

Bộ môn Khoa Học Tự Nhiên

Bảng kí hiệu

Trong bài giảng này ta dùng những kí hiệu với các ý nghĩa xác định trong bảng dưới đây

Ω	Không gian mẫu
C_n^k	Số tổ hợp chập k của n phần tử
A_n^k	Số chỉnh hợp chập k của n phần tử
$P\left(A\right)$	Xác suất của biến cố A
$E\left(X\right)$	Kỳ vọng của đại lượng ngẫu nhiên X
$D\left(X\right)$	Phương sai của đại lượng ngẫu nhiên X
F_X	Hàm phân phối của đại lượng ngẫu nhiên X
σ_X	Độ lệch tiêu chuẩn của đại lượng ngẫu nhiên X
$Cov(X, Y)$	Hiệp phương sai của đại lượng ngẫu nhiên hai chiều
$\rho(X, Y)$	Hệ số thương quan của đại lượng ngẫu nhiên hai chiều
μ	Trung bình tổng thể
σ	Phương sai của tổng thể
$\overline{X}$	Trung bình mẫu
S^2	Phương sai mẫu
S'^2	Phương sai mẫu hiệu chỉnh
S	Độ lệch mẫu
S'	Độ lệch mẫu hiệu chỉnh

Chương 1
NHỮNG KHÁI NIỆM CƠ BẢN VỀ XÁC SUẤT

Chương này cung cấp những kiến thức cơ bản nhất về đại số tổ hợp và lý thuyết xác suất. Giúp sinh viên hiểu rõ hơn về xác suất theo các quan điểm cổ điển, quan điểm thống kê và quan điểm hình học. Ngoài ra, chương này cũng xây dựng các công thức tính xác suất giúp sinh viên có thể giải quyết được các bài toán lý thuyết, các bài toán thực tế và cung cấp các kiến thức để giải quyết các vấn đề trong các chương tiếp theo.

1.1. Giải tích tổ hợp

1.1.1. Quy tắc cộng và quy tắc nhân

Quy tắc cộng: Để hoàn thành một việc, ta có thể thực hiện bởi một trong nhiều phương án. Phương án thứ nhất có m_1 cách thực hiện; phương án thứ hai có m_2 cách thực hiên,..., phương án thứ k có m_k cách thực hiện. Khi đó số cách hoàn thành công việc đó là

$$m_1 + m_2 + ... + m_k.$$

Ví dụ 1.1. Đội tuyển văn nghệ của trường đại học X có 5 sinh viên nam và 8 sinh viên nữ. Hỏi có bao nhiêu cách chọn một em trong đội để làm đội trưởng, (biết rằng em nào cũng có thể làm đội trưởng).

Giải. Để chọn một em làm đội trưởng ta có thể chọn một em nam hoặc một em nữ. Áp quy tắc cộng ta có $5 + 8 = 13$ (cách).

Quy tắc nhân: Giả sử một công việc nào đó sẽ được hoàn thành nếu ta thực hiện liên tiếp k công đoạn. Nếu công đoạn thứ nhất có m_1 cách thực hiện, công đoạn thứ 2 có m_2 cách thực hiện, ..., công đoạn thứ k có m_k cách thực hiện. Khi đó, số cách hoàn thành công việc là

$$m_1.m_2...m_k.$$

Ví dụ 1.2. Đội tuyển văn nghệ của trường đại học X có 5 sinh viên nam và 8 sinh viên nữ. Hỏi có bao nhiêu cách chọn một cặp nam nữ trong đội để hát song ca ?

Giải. Để chọn một cặp nam nữ trong đội để hát song ca thì ta phải thực hiện liên tiếp hai công đoạn là chọn một sinh viên nam và chọn một sinh viên nữ. Áp dụng quy tắc nhân ta có $5.8 = 40$ (cách).

1.1.2. Hoán vị, chỉnh hợp, tổ hợp

Định nghĩa 1.3. ***(Hoán vị)*** *Cho tập A có n phần tử, mỗi cách xếp n phần tử của tập A vào n vị trí khác nhau được gọi là một hoán vị của n phần tử.*

Vậy hai hoán vị là khác nhau nếu thứ tự sắp xếp các phần tử của chúng là khác nhau.
Số hoán vị của n phần tử là $p_n = n! = n(n-1).(n-2)...2.1.$

Ví dụ 1.4. Một đội tuyển học sinh giỏi có 6 học sinh trong đó có hai bạn A và B.
a. Hỏi có bao nhiêu cách xếp 6 học sinh trong đội vào một hàng ghế gồm 6 chiếc xếp theo hàng ngang sao cho mỗi ghế có đúng một học sinh ngồi.
b. Hỏi có bao nhiêu cách xếp 6 học sinh trong đội vào một hàng ghế gồm 6 chiếc xếp theo hàng ngang sao cho mỗi ghế có đúng một học sinh ngồi và hai bạn A, B ngồi cạnh nhau.

Giải. a. Mỗi cách xếp 6 học sinh của đội tuyển vào hàng ghế trên là một hoán vị của 6 phần tử. Vậy số cách xếp là $6! = 720$ (cách).

b.

• Ghép hai bạn A, B với nhau có 2 cách.

• Xếp 2 bạn A, B đã ghép và 4 bạn còn lại vào hàng ghế có $5! = 120$ cách.

Áp dụng quy tắc nhân ta có $2.120 = 240$ (cách).

Định nghĩa 1.5. ***(Chỉnh hợp)*** *Cho tập A có n phần tử. Mỗi một cách chọn k phần tử từ n phần tử của A (với $0 < k \leq n$) và xếp vào k vị trí khác nhau được gọi là một chỉnh hợp chập k của n phần tử của A.*

Vậy hai chỉnh hợp chập k của n phần tử là khác nhau nếu các phần tử của chúng khác nhau hoặc thứ tự sắp xếp của các phần tử là khác nhau. Số chỉnh hợp chập k của n phần tử của tập A là

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

Định nghĩa 1.6. ***(Tổ hợp)*** *Cho tập A có n phần tử, mỗi một tập con có k phần tử của tập A (với $0 \leq k \leq n$) được gọi là một tổ hợp chập k của n phần tử.*

Mỗi một cách chọn k phần tử từ n phần tử của tập A cho ta một tập con có k phần tử của tập A. Do đó một tổ hợp chập k của n phần tử chính là một cách chọn k phần tử từ n phần tử của tập A. Số tổ hợp chập k của n phần tử là $C_n^k = \dfrac{n!}{k!(n-k)!}$.

Tính chất của C_n^k

• $C_n^k = C_n^{n-k},\ \forall 0 \leq k \leq n;\ \ k,\ n \in \mathbb{Z}.$

• $C_n^k = C_{n-1}^{k-1} + C_{n-1}^k, \forall 1 \leq k \leq n-1;\ \ k,\ n \in \mathbb{Z}.$

Ví dụ 1.7. Một lô hàng gồm 8 sản phẩm. Từ lô hàng đó lấy ra cùng một lúc 3 sản phẩm. Hỏi có bao nhiêu cách lấy ?

Giải. Mỗi một cách lấy đồng thời 3 sản phẩm từ 8 sản phẩm là một tổ hợp chập 3 của 8 phần tử. Vậy số cách lấy là

$$C_8^3 = \frac{8!}{3!(8-3)!} = 56.$$

Ví dụ 1.8. Một lô linh kiện điện tử có 100 linh kiện tốt và 10 linh kiện bị hỏng. Người ta lấy đồng thời 3 linh kiện trong lô hàng để kiểm tra.
a. Hỏi có bao nhiêu cách lấy sao cho không có linh kiện nào bị hỏng ?
b. Hỏi có bao nhiêu cách lấy sao cho có cả linh kiện bị hỏng và linh kiện tốt ?
c. Hỏi có bao nhiêu cách lấy sao cho có ít nhất một linh kiện tốt ?

Giải. a. Số cách lấy 3 linh kiện trong lô hàng sao cho không có linh kiện nào bị hỏng là $C_{100}^3 = 161700$ cách.
b. Số cách lấy 3 linh kiện trong lô hàng sao cho có cả linh kiện bị hỏng và linh kiện tốt là $C_{110}^3 - C_{100}^3 - C_{10}^3 = 54000$ cách.
c. Số cách lấy 3 linh kiện trong lô hàng sao cho có ít nhất một linh kiện tốt là $C_{110}^3 - C_{10}^3 = 215700$ cách.

1.2. Biến cố và mối quan hệ của biến cố

1.2.1. Phép thử và biến cố

Định nghĩa 1.9. *Việc thực hiện một nhóm các điều kiện cơ bản để quan sát một hiện tượng nào đó có xảy ra hay không được gọi là thực hiện một phép thử. Các kết quả có thể xảy ra của phép thử được gọi là biến cố (sự kiện).*
• *Phép thử ngẫu nhiên là phép thử mà kết qủa của nó không thể dự báo trước.*
• *Tập hợp tất cả các kết quả có thể có của phép thử được gọi là không gian mẫu của phép thử và thường kí hiệu bởi Ω.*

Ví dụ 1.10. 1. Sản xuất 1 sản phẩm là thực hiện 1 phép thử. Kết quả 1 sản phẩm sản xuất ra đạt tiêu chuẩn hay không đạt tiêu chuẩn là các biến cố.

2. Tung 1 đồng xu là thực hiện 1 phép thử. Hiện tượng đồng xu xuất hiện mặt sấp hay xuất hiện mặt ngửa là các biến cố. Trong phép thử này ta có $\Omega = \{S, N\}$.
3. Gieo 1 con xúc sắc là thực hiện một phép thử. Hiện tượng xúc sắc xuất hiện mặt 3 chấm; xúc sắc xuất hiện mặt 6 chấm; xúc sắc xuất hiện mặt có số chấm lớn hơn 6; xúc sắc xuất hiện mặt có số chấm nhỏ hơn 4 là các biến cố. Trong phép thử này ta có $\Omega = \{1, 2, 3, 4, 5, 6\}$.

1.2.2. Các loại biến cố

Một biến cố chỉ có thể xảy ra khi một phép thử gắn liền với nó được thực hiện. Trong thực tế có thể xảy ra các loại biến cố sau đây:

• ***Biến cố ngẫu nhiên:*** Là biến cố có thể xảy ra hoặc không xảy ra khi phép thử được thực hiện. Các biến cố ngẫu nhiên thường ký hiệu $A, B, C, A_1, A_2, B_1, B_2$...

• ***Biến cố chắc chắn:*** Là biến cố luôn xảy ra khi phép thử được thực hiện, ký hiệu là Ω hoặc U.

• ***Biến cố không thể:*** Là biến cố không bao giờ xảy ra khi phép thử được thực hiện, ký hiệu là $\emptyset$ hoặc V.

• ***Biến cố sơ cấp:*** Là biến cố không thể phân tích được nữa.

Ví dụ 1.11. Xét phép thử: Gieo một con xúc sắc một lần. Khi đó

Biến cố A:" Con xúc sắc xuất hiện mặt có số chấm là chẵn" là biến cố ngẫu nhiên. Các kết quả của phép thử là $\{2, 4, 6\}$ làm biến cố A xảy ra được gọi là các kết quả thận lợi của biến cố A.

Biến cố B: "Con xúc sắc xuất hiện mặt có số chấm là số tự nhiên nhỏ hơn 7" là biến cố chắc chắn.

Biến cố C "Con xúc sắc xuất hiện mặt có số chấm lớn hơn 7" là biến cố không thể.

Nếu gọi A_i là biến cố: "con xúc xắc xuất hiện mặt i chấm $(i = 1, .., 6)$ " thì

$A_1, A_2, A_3, A_4, A_5, A_6$ là các biến cố sơ cấp.

1.2.3. Mối quan hệ giữa các biến cố

a. ***Biến cố giao (Biến cố tích)***

Định nghĩa 1.12. *Biến cố chỉ xảy ra khi tất cả các biến cố* $A_1, A_2, ..., A_n$ *cùng xảy ra được gọi là biến cố giao (hay biến cố tích) của các biến cố* $A_1, A_2, ..., A_n$ *và được ký hiệu là* $A_1.A_2...A_n$ *hoặc* $A_1 \cap A_2 \cap ... \cap A_n$.

Ví dụ 1.13. Hai xạ thủ mỗi người bắn một viên đạn vào bia.
Gọi A_1:"Xạ thủ thứ nhất bắn trúng bia"; A_2:"Xạ thủ thứ hai bắn trúng bia" và A: " Cả hai xạ thủ cùng bắn trúng bia". Biến cố A chỉ xảy ra khi cả hai biến cố A_1 và A_2 cùng xảy ra. Do đó $A = A_1.A_2$.

b. ***Biến cố tổng (biến cố hợp)***

Định nghĩa 1.14. *Biến cố chỉ xảy ra khi có ít nhất một trong các biến cố* A_1 *hoặc* A_2 *... hoặc* A_n *xảy ra được gọi là biến cố tổng của các biến cố* $A_1, A_2, ..., A_n$ *và được ký hiệu* $A_1 + A_2 + ... + A_n$ *hoặc* $A_1 \cup A_2 \cup ... \cup A_n$.

Ví dụ 1.15. Hai người thợ săn cùng bắn vào một con thú. Gọi A là biến cố "người thứ nhất bắn trúng", B là biến cố "người thứ hai bắn trúng" và C là biến cố "con thú bị bắn trúng". Khi đó C chỉ xảy ra khi có ít nhất một trong hai biến cố A hoặc B xảy ra. Do đó $C = A + B$.

Ví dụ 1.16. Xét phép thử: Sản xuất 3 sản phẩm.
Gọi A_i là biến cố sản phẩm thứ i sản xuất ra đạt tiêu chuẩn (i =1,2,3). Gọi A là biến cố cả 3 sản phẩm sản xuất ra đều đạt tiêu chuẩn, B là biến cố có ít nhất 1 sản phẩm sản xuất ra đạt tiêu chuẩn, Theo định nghĩa của tổng và tích các biến cố, ta có thể biểu diễn các biến cố A và B theo các biến cố A_1, A_2, A_3 như sau: $A = A_1.A_2.A_3$ và $B = A_1 + A_2 + A_3$.

c. ***Biến cố xung khắc***

Định nghĩa 1.17. *Hai biến cố A và B gọi là xung khắc nếu chúng không thể đồng thời xảy ra trong một phép thử. Nói cách khác, nếu biến cố A xảy ra thì biến cố B không xảy ra và ngược lại. Như vậy nếu A và B là hai biến cố xung khắc thì $A.B = \emptyset$.*

• *Nhóm n biến cố $A_1, A_2, ..., A_n$ gọi là xung khắc từng đôi nếu bất kỳ hai biến cố nào trong nhóm này cũng xung khắc với nhau, nghĩa là $A_i.A_j = \emptyset$ với $i \neq j$.*

• *Các biến cố $A_1, A_2, ..., A_n$ được gọi là nhóm biến cố đầy đủ nếu trong kết quả của phép thử sẽ xảy ra một và chỉ một trong các biến cố đó. Nói cách khác, các biến cố nói trên sẽ tạo nên một nhóm đầy đủ các biến cố nếu chúng xung khắc từng đôi với nhau và tổng của chúng là một biến cố chắc chắn. Như vậy, các biến cố $A_1, A_2, ..., A_n$ là nhóm biến cố đầy đủ nếu*

$$\begin{cases} A_1 + A_2 + ... + A_n = \Omega, \\ A_1.A_2....A_n = \phi. \end{cases}$$

Ví dụ 1.18. Một đội tuyển văn nghệ của trường THPT A có 5 em khối 10, 7 em khố 11 và 10 em khối 12. Chọn ngẫu nhiên 1 em trong đội.
Gọi A là biến cố: "Chọn được em khối 10"; B là biến cố: "Chọn được em khối 11"; C là biến cố: "Chọn được em khối 12".

• Ta có A và B là 2 biến cố xung khắc.

• Các biến cố A, B, C là 3 biến cố đôi một xung khắc.

• Các biến cố A, B, C lập thành một nhóm biến cố đầy đủ.

d. ***Biến cố đối***

Định nghĩa 1.19. *Hai biến cố A và $\bar{A}$ gọi là đối lập với nhau nếu chúng tạo nên một nhóm đầy đủ các biến cố, tức là*

$$\begin{cases} A + \overline{A} = \Omega, \\ A.\overline{A} = \phi. \end{cases}$$

Ví dụ 1.20. 1. Sản xuất 1 sản phẩm. Gọi A là biến cố “sản phẩm sản xuất ra đạt tiêu chuẩn” thì $\bar{A}$ là biến cố “sản phẩm sản xuất ra không đạt tiêu chuẩn”.

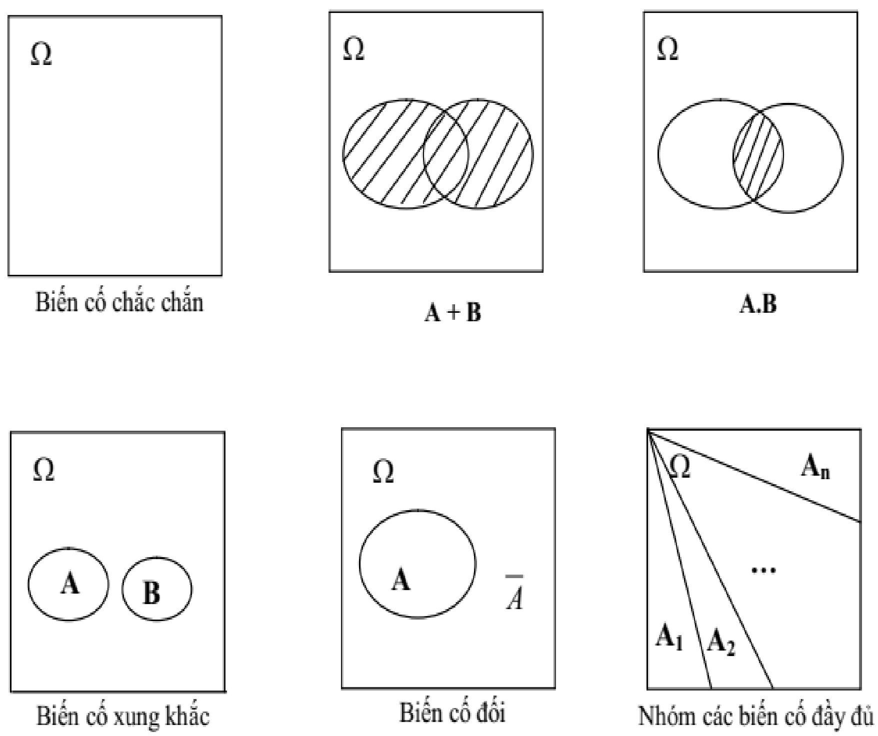
2. Một xạ thủ bắn liên tiếp 3 viên đạn vào bia. Gọi A là biến cố :“Có ít nhất một viên đạn trúng bia”, khi đó $\bar{A}$ là biến cố :“Không có viên đạn nào trúng bia ”.

Chú ý

• Từ định nghĩa biến cố đối lập và biến cố xung khắc ta suy ra rằng nếu hai biến cố đối lập thì chúng xung khắc. Điều ngược lại chưa chắc đúng.

• Với hai biến cố A và B ta có $\overline{A+B} = \overline{A}.\overline{B}$ và $\overline{A.B} = \overline{A} + \overline{B}$.

Ta có thể sử dụng biểu đồ Ven để biểu diễn mối quan hệ của các biến cố như sau:



Hình 1.1 Mối quan hệ giữa các biến cố

1.3. Các định nghĩa về xác suất

Như trên đã thấy, việc biến cố ngẫu nhiên xảy ra hay không xảy ra trong kết quả của phép thử là điều không thể đoán trước được, tuy nhiên, bằng trực quan ta có thể nhận thấy các biến cố ngẫu nhiên khác nhau có những khả năng xảy ra khác nhau. Hơn nữa, khi lặp đi lặp lại nhiều lần cùng một phép thử trong những điều kiện như nhau, người ta thấy tính chất ngẫu nhiên của biến cố mất dần đi và khả năng xảy ra của biến cố sẽ được thể hiện theo những quy luật nhất định. Từ đó ta thấy có khả năng định lượng (đo lường) khả năng khách quan xuất hiện một biến cố nào đó.

Xác suất của một biến cố là một con số đặc trưng khả năng khách quan xuất hiện biến cố đó khi thực hiện phép thử.

Như vậy bản chất xác suất của một biến cố là một con số xác định. Để tính xác suất của một biến cố, người ta xây dựng các định nghĩa và định lý sau đây.

1.3.1. Định nghĩa xác suất theo quan điểm cổ điển

Định nghĩa 1.21. *Xét một phép thử có hữu hạn kết quả và đồng khả năng xảy ra. Xác suất xuất hiện biến cố A là một số được ký hiệu là $P(A)$ và được xác định bởi công thức*

$$P(A) = \frac{n(A)}{n(\Omega)},$$

trong đó $n(\Omega)$ là số phần tử của không gian mẫu và $n(A)$ là số trờng hợp thuận lợi của biến cố A.

Ví dụ 1.22. Tung 1 con xúc sắc cân đối và đồng chất. Tính xác suất để con xúc xắc xuất hiện mặt chẵn chấm.

Giải. Ta có $\Omega = \{1; 2; 3; 4; 5; 6\} \Rightarrow n(\Omega) = 6.$

Gọi A là biến cố: “con xúc xắc xuất hiện mặt chẵn chấm”.

Các kết quả thuận lợi của A là $\{2; 4; 6\}$, suy ra $(A) = 3$.
Do đó $P(A) = \dfrac{n(A)}{n(\Omega)} = \dfrac{3}{6} = \dfrac{1}{2}$.

Ví dụ 1.23. Một người đặt mật khẩu nhưng lại quên 2 số cuối của mật khẩu, chỉ nhớ đó là 2 chữ số khác nhau. Tìm xác xuất để người đó gõ ngẫu nhiên một lần đúng mật khẩu.

Giải. Số kết quả có thể xảy ra của phép thử là: $n(\Omega) = A_{10}^2 = 90$.
Gọi A là biến cố: "người đó gõ ngẫu nhiên một lần đúng mật khẩu". Ta có $n(A) = 1$, suy ra $P(A) = \dfrac{1}{90}$.

Ví dụ 1.24. Một đội tuyển học sinh giỏi có 5 em khối 12, 4 em khối 11 và 3 em khối 10. Chọn ngẫu nhiên 4 em trong đội tuyển để tham dự kỳ thi học sinh giỏi quốc gia. Tính xác suất sao cho :

a. Bốn em được chọn có đúng 2 em khối 12 ?

b. Bốn em được chọn có đủ cả 3 khối ?

Giải. Chọn đồng thời 4 em trong đội tuyển có $C_{12}^4 = 495$ cách, suy ra $n(\Omega) = 495$.

a. Gọi A là biến cố : " Bốn em được chọn có đúng hai em khối 12".
Ta có $n(A) = C_5^2.C_7^2 = 210$, suy ra $P(A) = \dfrac{210}{495} = \dfrac{4}{33}$.

b. Gọi B là biến cố : " Bốn em được chọn có đủ cả 3 khối ".
Ta có $n(B) = C_5^2.C_4^1.C_3^1 + C_5^1.C_4^2.C_3^1 + C_5^1.C_4^1.C_3^2 = 270$, suy ra
$P(B) = \dfrac{270}{495} = \dfrac{6}{11}$.

Ví dụ 1.25. Trong hộp có 10 sản phẩm trong đó có 6 chính phẩm và phế phẩm. Lấy ngẫu nhiên 3 sản phẩm cùng một lúc. Tính xác suất để:

a. Ba sản phẩm lấy ra đều là chính phẩm

b. Trong 3 sản phẩm lấy ra có 2 chính phẩm.

Giải. Ta có $n(\Omega) = C_{10}^3 = 120$.

a. Gọi A là biến cố "cả 3 sản phẩm lấy ra đều là chính phẩm".

Số trường hợp thuận lợi cho việc xuất hiện biến cố A là $n(A) = C_6^3 = 20$, suy ra $P(A) = \frac{20}{120} = \frac{1}{6}$.
b. Gọi B biến cố "trong 3 sản phẩm lấy ra có 2 chính phẩm".
Số trường hợp thuận lợi cho việc xuất hiện biến cố B là $n(B) = C_6^2.C_4^1 = 60$, suy ra $P(B) = \frac{60}{120} = \frac{1}{2}$.

Ưu điểm và hạn chế của định nghĩa xác suất theo lối cổ điển

• Để tìm xác suất của biến cố không cần phải tiến hành phép thử (phép thử chỉ tiến hành một cách giả định) mà ta vẫn có thể tìm được một cách chính xác giá trị của xác suất.

• Tuy nhiên, định nghĩa cổ điển về xác suất đòi hỏi số kết quả của phép thử phải là hữu hạn và đồng khả năng xảy ra. Trong thực tế nhiều phép thử không có số kết quả là hữu hạn và cũng không đồng khả năng xảy ra.

Vì những lý do trên mà ngoài định nghĩa cổ điển về xác suất, trong thực tế người ta còn sử dụng định nghĩa xác suất theo quan điểm thống kê sau đây.

1.3.2. Định nghĩa xác suất theo quan điểm thống kê

Định nghĩa 1.26. *Thực hiện n lần phép thử T. Giả sử biến cố A xuất hiện m lần, khi đó m được gọi là tần số của biến cố A. Tỷ số $\frac{m}{n}$ gọi là tần suất xuất hiện biến cố A trong loạt phép thử và được ký hiệu $f(A)$.*

Khi 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.
Vậy $P(A) = \lim\limits_{n\to\infty} \frac{m}{n}$.

Ví dụ 1.27. 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 $\frac{50}{1000} = 5\%$.

Người ta nhận thấy nếu tiến hành các thí nghiệm trong những điều kiện như nhau và số phép thử khá lớn thì tần suất thể hiện tính ổn định của nó khá rõ ràng. Tính chất này thể hiện ở chỗ là trong những loạt thí nghiệm khác nhau nhưng *số phép thử khá lớn thì tần suất dao động rất ít xung quanh một giá trị nào đó.*

Ví dụ 1.28. 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 thử nghiệm	Số lần tung (n)	Số lần xuất hiện mặt sấp	Tần suất xuất hiện mặt sấp
Buyffon	4040	2.048	0,5069
Pearson	12000	6019	0,5016
Pearson	24000	12012	0,5016

Qua thí dụ trên ta thấy khi số phép thử tăng lên thì tần suất xuất hiện mặt sấp sẽ dao động ngày càng ít hơn xung quanh giá trị không đổi là $0,5$. Điều đó cho phép hy vọng rằng khi số phép thử tăng lên vô hạn, tần suất sẽ hội tụ về giá trị $0,5$. Như vậy về mặt thực tế, với số phép thử n đủ lớn ta có thể lấy $P(A) \approx f(A)$.

Dễ dàng nhận thấy rằng định nghĩa xác suất theo quan điểm thống kê, xác suất cũng có đầy đủ những tính chất như trong định nghĩa cổ điển.

Định nghĩa thống kê của xác suất đã khắc phục được một nhược điểm của định nghĩa cổ điển (định nghĩa này không dùng đến khái niệm đồng khả năng), vì vậy định nghĩa này được sử dụng nhiều trong thực tế. Bên cạnh ưu điểm trên, ta nhận thấy định nghiã này còn có những hạn chế, nó không giúp ta tìm được giá trị chính xác của xác suất, mà chỉ tìm được giá trị gần đúng.

Tuy nhiên bằng định nghĩa này, người ta đã tìm được xác suất để sinh con trai trong mỗi lần sinh là $0,518$, số con này hầu như không thay đổi theo thời gian, địa phương và chủng tộc. Nhà toán học Laplatxơ (Laplace) trong

10 năm liền theo dõi ở các thành phố Pêtecbua, Luân đôn và Beclin thấy tỷ số đó là $\frac{22}{43}$. Ông cũng đã theo dõi 40 năm liền ở Pari thấy tỷ số là $\frac{25}{49}$. Nhà toán học Crame theo dõi ở Thụy điển trong năm 1935 cũng thấy tỷ số đó là 0,518.

Các định nghĩa trên đây của xác suất đã giúp ta một cách tích cực trong việc tính xác suất, nhưng mỗi định nghĩa đều có nhược điểm của nó. Để khắc phục những nhược điểm đó, năm 1933 nhà toán học Xô viết Canmơgơrôp đã đưa ra xác suất theo phương pháp tiên đề. Trong phạm vi của giáo trình, ta không đề cập đến định nghĩa đó.

Nguyên lý xác suất lớn và xác suất nhỏ

• Nếu một biến cố có xác suất rất nhỏ thì thực tế có thể cho rằng trong một phép thử biến cố đó sẽ không xảy ra.

• Nếu biến cố ngẫu nhiên có xác suất gần bằng 1 thì thực tế có thể cho rằng biến cố đó sẽ xảy ra trong một phép thử.

Tính chất của xác suất

• $P(\emptyset) = 0$ và $P(\Omega) = 1$.

• $0 \leq P(A) \leq 1$.

1.4. Các công thức tính xác suất

1.4.1. Xác suất có điều kiện, công thức nhân xác suất

a. Xác suất có điều kiện

Định nghĩa 1.29. *Trong thực tế có những trường hợp ta phải tính xác suất của biến cố A khi biến cố B đã xảy ra, xác suất đó ký hiệu là $P(A/B)$ và được gọi là xác suất có điều kiện của biến cố A với điều kiện biến cố B đã xảy ra.*

Ví dụ 1.30. 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. Vậy xác suất cần tính là $P(A/B)$. Ta thấy lần thứ nhất lấy được viên bi trắng (B đã xảy ra) nên trong hộp còn 7 viên bi trong đó có 4 viên bi trắng. Do đó $P(A/B) = \dfrac{C_4^1}{C_7^1} = \dfrac{4}{7}$.

Ví dụ 1.31. Có hai hộp linh kiện, hộp 1 có 10 linh kiện tốt và 4 linh kiện hỏng; hộp 2 có 8 linh kiện tốt và 4 linh kiện hỏng. Lấy ngẫu nhiên một linh kiện từ hộp 1 bỏ vào hộp 2 sau đó lấy một linh kiện từ hộp 2 ra ngoài. Tính xác suất để linh kiện lấy ra từ hộp 2 là linh kiện tốt biết rằng linh kiện lấy từ hộp 1 bỏ vào hộp 2 là linh kiện hỏng.

Giải. Gọi B là biến cố: " Linh kiện lấy từ hộp 1 bỏ vào hộp 2 là linh kiện hỏng " và A là biến cố: " Linh kiện lấy từ hộp 2 ra ngoài là linh kiện tốt". Khi đó xác suất cần tính là $P(A/B)$.
Khi B xảy ra thì hộp 2 có 8 linh kiện tốt và 5 linh kiện hỏng. Do đó xác suất để lấy được 1 linh kiện tốt từ hộp 2 là $\dfrac{8}{13}$. Vậy, $P(A/B) = \dfrac{8}{13}$. Chúng ta có thể chứng minh được công thức xác suất có điều kiện được xác định như sau:

$$P(A/B) = \frac{P(AB)}{P(B)}, (P(B) \neq 0)$$

và

$$P(B/A) = \frac{P(AB)}{P(A)}, (P(A) \neq 0).$$

b. Biến cố độc lập

Định nghĩa 1.32. *Hai biến cố A và B được gọi là độc lập với nhau, nếu biến cố này xuất hiện hay không, cũng không ảnh hưởng đến xác suất xuất hiện của biến cố kia.*

Nhận xét:

• Hai biến cố A, B độc lập nếu $P(A/B) = P(A/\bar{B}) = P(A)$ và $P(B/A) = P(B/\bar{A}) = P(B)$.

• Nếu A và B là hai biến cố độc lập với nhau thì các cặp biến cố sau cũng độc lập với nhau: A và $\bar{B}$; $\bar{A}$ và B; $\bar{A}$ và $\bar{B}$.

• Để xác định tính độc lập của các biến cố, trong thực tế ít khi người ta dùng cách kiểm nghiệm xem những đẳng thức trên có được thực hiện hay không, mà thông thường người ta căn cứ vào kinh nghiệm, vào trực giác. Chẳng hạn, khi tung hai đồng xu phân biệt, rõ ràng đồng xu này có xuất hiện mặt sấp hay không cũng không ảnh hưởng tới xác suất để đồng xu kia xuất hiện mặt sấp (hay ngửa); Việc bà mẹ này sinh con trai (hay gái) cũng không ảnh hưởng tới xác suất sinh con trai (gái) của bà mẹ khác...Vậy các biến cố độc lập thường xuất hiện trên các chủ thể riêng biệt nhau.

• Các biến cố $A_1, A_2, ..., A_n$ gọi là độc lập từng đôi với nhau nếu mỗi cặp hai trong n biến cố độc lập với nhau.

• Các biến cố $A_1, A_2, ..., A_n$ gọi là độc lập toàn phần nếu sự xảy ra hay không xảy ra của biến cố này là không ảnh hưởng đến xác suất xảy ra của c biến cố c lại.

c. Công thức nhân xác suất

Định lí 1.33. *Với hai biến cố A và B ta có*

$$P\left(A.B\right) = P\left(B\right).P\left(A/B\right) = P\left(A\right).P\left(B/A\right).$$

Tổng quát, với n biến cố $A_1, A_2, ..., A_n$ ta có

$$P(A_1.A_2...A_n) = P(A_1.P(A_2/A_1)...P(A_n/A_1.A_2...A_{n-1}).$$

Hệ quả 1.34. *1. Nếu A và B là hai biến cố độc lập với nhau thì ta có*

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

2. Nếu $A_1, A_2, ..., A_n$ là các biến cố độc lập toàn phần thì ta có

$$P\left(A_1.A_2...A_n\right) = P\left(A_1\right).P\left(A_2\right)...P\left(A_n\right).$$

Ví dụ 1.35. Có hai máy hoạt động độc lập nhau. Xác suất để máy thứ nhất, máy thứ hai không bị hỏng trong 1 ca làm việc lần lượt là 0,9 và 0,8.

a. Tính xác suất để cả hai máy đều không bị hỏng trong 1 ca làm việc ?

b. Tính xác suất để có ít nhất một máy không bị hỏng trong 1 ca làm việc ?

Giải. a. Gọi A là biến cố "cả hai máy đều không bị hỏng trong 1 ca làm việc", A_i là biến cố "máy thứ i không hỏng trong 1 ca làm việc" (i = 1,2).
Khi đó ta có $A = A_1.A_2$. Theo giả thiết A_1 và A_2 là hai biến cố độc lập nhau nên

$$P(A) = P(A_1.A_2) = P(A_1).P(A_2) = 0,9.0,8 = 0,72.$$

b. Gọi B là biến cố: "Có ít nhất một máy không bị hỏng trong 1 ca làm việc", khi đó $\bar{B}$ là biến cố:"Cả hai máy đều bị hỏng trong 1 ca làm việc". Ta có $\bar{B} = \bar{A}_1.\bar{A}_2$. Áp dụng quy tắc nhân xác suất với hai biến cố $\bar{A}_1$ và $\bar{A}_2$ độc lập ta có: $P(\bar{B}) = P\left(\bar{A}_1\right).P\left(\bar{A}_2\right) = 0,1.0,2 = 0,02.$
Vậy $P(B) = 1 - 0,02 = 0,98.$

Ví dụ 1.36. Một tập gồm 10 chứng từ trong đó có 2 chứng từ không hợp lệ. Một cán bộ kế toán rút ngẫu nhiên lần lượt không hoàn lại 2 chứng từ để kiểm tra.
a. Tính xác suất để cả hai chứng từ rút ra đều hợp lệ ?
b. Người đó rút tiếp chứng từ thứ ba. Tính xác suất để trong 3 chứng từ rút ra chỉ có chứng từ thứ 3 không hợp lệ ?

Giải. a. Gọi A là biến cố "cả 2 chứng từ rút ra đều hợp lệ"
Gọi A_i là biến cố "chứng từ rút ra lần i là hợp lệ" (i =1,2,3) thì $\bar{A}_i$ là biến cố "Chứng từ rút ra lần i là không hợp lệ". Khi đó $A = A_1.A_2$, áp dụng công thức nhân xác suất ta được

$$P(A) = P(A_1.A_2) = P(A_1)P(A_2/A_1) = \frac{8}{10}.\frac{7}{9} = \frac{56}{90} = \frac{28}{45}.$$

b. Gọi B là biến cố "trong 3 chứng từ rút ra, chỉ có chứng từ thứ 3 không hợp lệ". Khi đó $B = A_1A_2\bar{A}_3$, áp dụng công thức nhân xác suất ta được

$$P(B) = P(A_1A_2\bar{A}_3) = P(A_1).P(A_2/A_1).P(\bar{A}_3/A_1.A_2) = \frac{8}{10}.\frac{7}{9}.\frac{2}{8} = \frac{7}{45}.$$

1.4.2. Công thức cộng xác suất

Định lí 1.37. *Với hai biến cố A và B ta có*

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

Tổng quát, với n biến cố $A_1, A_2, ..., A_n$ ta có:

$$P(A_1+A_2+...+A_n)=\sum_{1\le i_1\le n}P(A_{i_1})-\sum_{1\le i_1<i_2\le n}P(A_{i_1}.A_{i_2})+...$$
$$+(-1)^{k+1}\sum_{1\le i_1<i_2<...<i_k\le n}P(A_{i_1}.A_{i_{2...A_k}})+...+(-1)^{n+1}P(A_1.A_2...A_n).$$

Hệ quả 1.38. *1. Nếu A và B là hai biến cố xung khắc thì*

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

2. Nếu $A_1, A_2, ..., A_n$ là một hệ các biến cố xung khắc từng đôi một thì

$$P(A1+A2+...+An)=P(A1)+P(A2)+...+P(An).$$

3. Nếu $A_1, A_2, ..., A_n$ là một nhóm các biến cố đầy đủ thì

$$P(A1)+P(A2)+...+P(An)=1.$$

4. $P\left(\bar{A}\right)=1-P(A).$

Ví dụ 1.39. Ba vận động viên bóng rổ, mỗi người ném 1 quả vào rổ. Giả sử xác suất ném trúng rổ của vận đội viên thứ 1, 2, 3 lần lượt là 0,7; 0,8; 0,9. Tính xác suất để”:

a. Có đúng 1 vận động viên ném trúng rổ ?

b. Có đúng 2 vận động viên ném trúng rổ ?

c. Có ít nhất 1 vận động viên ném trúng rổ ?

d. Có vận động viên thứ 1 ném trúng rổ biết rằng có 2 người ném trúng rổ ?

Giải. a. Gọi A_i là biến cố "Vận động viên thứ i ném trúng rổ" (i=1,2,3), và A là biến cố "Có đúng 1 vận động viên ném trúng rổ".

Khi đó, $A = A_1\bar{A}_2\bar{A}_3 + \bar{A}_1A_2\bar{A}_3 + \bar{A}_1\bar{A}_2A_3$, áp dụng công thức cộng xác suất ta có

$$\begin{aligned}
P(A) &= P(A_1\bar{A}_2\bar{A}_3 + \bar{A}_1A_2\bar{A}_3 + \bar{A}_1\bar{A}_2A_3)\\
&= P(A_1\bar{A}_2\bar{A}_3) + P(\bar{A}_1A_2\bar{A}_3) + P(\bar{A}_1\bar{A}_2A_3)\\
&\overset{dl}{=} P(A_1)P(\bar{A}_2)P(\bar{A}_3) + P(\bar{A}_1)P(A_2)P(\bar{A}_3) + P(\bar{A}_1)P(\bar{A}_2)P(A_3)\\
&= 0,092.
\end{aligned}$$

b. Gọi B là biến cố "Có đúng 2 vận động viên ném trúng rổ".

Khi đó $B = A_1A_2\bar{A}_3 + A_1\bar{A}_2A_3 + \bar{A}_1A_2A_3$ làm tương tự như trên ta được $P(B) = 0,398$.

c. Gọi C là biến cố "Có ít nhất 1 vận động viên ném trúng rổ". Khi đó $\overline{C}$ là biến cố: "Không có vận động viên nào ném trúng rổ". Ta có $\bar{C} = \bar{A}_1\bar{A}_2\bar{A}_3$. Áp dụng quy tắc nhân xác suất ta được:

$$P(\bar{C}) = P(\bar{A}_1\bar{A}_2\bar{A}_3) = P(\bar{A}_1)P(\bar{A}_2)P(\bar{A}_3) = 0,3.0,2.0,1 = 0,006.$$

Vậy $P(C) = 1 - P(\bar{C}) = 1 - 0,006 = 0,994$. d. Xác suất cần tính là $P(A_1/B) = \dfrac{P(A_1B)}{P(B)}$, trong đó $P(B) = 0,398$.

Ta có $A_1B = A_1A_2\bar{A}_3 + A_1\bar{A}_2A_3$, do đó

$$\begin{aligned}
P(A_1B) &= P(A_1A_2\bar{A}_3 + A_1\bar{A}_2A_3)\\
&\overset{dl}{=} P(A_1)P(A_2)P(\bar{A}_3) + P(A_1)P(\bar{A}_2)P(A_3)\\
&= 0,7.0,8.(1-0,9) + 0,7(1-0,8).0,9\\
&= 0,182.
\end{aligned}$$

Ví dụ 1.40. Một người bắn 1 viên đạn vào một bia đã được chia làm 3 vòng. Giả sử xác suất để người đó bắn trúng vòng 10, vòng 9, vòng 8 của bia lần lượt là 0,2; 0,3 và 0,4. Tính xác suất để người đó:

a. Bắn trúng vòng 10 hoặc vòng 9 của bia ?

b. Bắn trúng bia ?

c. Bắn không trúng bia ?

Giải. a. Gọi A_i là biến cố “bắn trúng vòng i của bia” (i =10, 9, 8), A là biến cố “bắn trúng vòng 10 hoặc vòng 9 của bia”. Khi đó $A = A_1 + A_2$, áp dụng công thức cộng xác suất với hai biến cố A_1 và A_2 xung khắc ta được:

$$P(A) = P(A_1 + A_2) = P(A_1) + P(A_2) = 0,2 + 0,3 = 0,5.$$

b.Gọi B là biến cố "Người đó bắn trúng bia", khi đó $B = A_1 + A_2 + A_3$. Áp dụng công thức tính xác suất với các biến cố A_1, A_2, A_3 đôi một xung khắc ta có: $P(B) = P(A1+A2+A3) = P(A1)+P(A2)+P(A3) = 0,2+0,3+0,4 = 0,9$.
c. Gọi C là biến cố "Người đó bắn không trúng bia", khi đó $C = \bar{B}$, do đó $P(C) = 1 - P(B) = 0,1$.

1.4.3. Công thức xác suất toàn phần và công thức Bayes

a. Công thức xác suất toàn phần

Định lí 1.41. *Giả sử $A_1, A_2, ..., A_n$ là nhóm các biến cố đầy đủ và A là một biến cố bất kỳ nào đó có thể xảy ra đồng thời với một và chỉ một trong các biến cố trên. Khi đó xác suất của biến cố A được tính bởi công thức:*

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

Chứng minh: Vì $A_1, A_2, ..., A_n$ lập thành hệ đầy đủ nên ta có $A_1 + A_2 + \ldots + A_n = \Omega$ và A có thể xảy ra đồng thời với một trong các biến cố đó nên $A = A.\Omega = A.\sum_{i=1}^{n} A_i = \sum_{i=1}^{n} A_iA$.

Các biến cố $A_i, (i = 1, ..., n)$ xung khắc từng đôi nên các biến cố $A_i.A, (i = 1, ..., n)$ cũng xung khắc từng đôi. Áp dụng công thức cộng và nhân xác suất ta được:

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

Ví dụ 1.42. Có hai lô sản phẩm: Lô 1 có 2 sản phẩm tốt và 5 sản phẩm xấu. Lô 2 có 3 sản phẩm tốt và 6 sản phẩm xấu. Lấy ngẫu nhiên một sản phẩm từ lô 1 bỏ vào lô 2 rồi lấy ngẫu nhiên 1 sản phẩm từ lô 2 ra ngoài. Tính xác suất để sản phẩm lấy ra từ lô 2 là sản phẩm tốt?

Giải. Gọi A là biến cố "sản phẩm lấy ra từ lô 2 là tốt"; A_1 là biến cố "sản phẩm lấy từ lô1 bỏ vào lô 2 là tốt"; A_2 là biến cố "sản phẩm lấy từ lô 1 bỏ vào lô 2 là xấu".

Khi đó A_1, A_2 lập thành nhóm biến cố đầy đủ và A chỉ xảy ra đồng thời với 1 trong 2 biến cố đó. Áp dụng công thức xác suất toàn phần ta có $P(A) = P(A1).P(A/A1) + P(A2).P(A/A2)$. Trong đó, ta tính được $P\left(A_1\right) = \dfrac{2}{7};\ P\left(A_2\right) = \dfrac{5}{7}$ và $P\left(A/A_1\right) = \dfrac{4}{10} = \dfrac{2}{5}; P\left(A/A_2\right) = \dfrac{3}{10}$.

Vậy $P\left(A\right) = \dfrac{2}{7}.\dfrac{2}{5} + \dfrac{5}{7}.\dfrac{3}{10} = \dfrac{23}{70}$.

Ví dụ 1.43. Một nhà máy có 3 phân xưởng sản xuất ra cùng một loại sản phẩm. Xác suất để phân xưởng 1, phân xưởng 2, và phân xưởng 3 sản xuất được sản phẩm loại I lần lượt là 0,7; 0,8; 0,6. Từ 1 lô hàng gồm 20% sản phẩm của phân xưởng 1; 50% sản phẩm của phân xưởng 2; 30% sản phẩm của phân xưởng 3 người ta lấy ra 1 sản phẩm để kiểm tra. Tính xác suất để sản phẩm được kiểm tra là loại 1

Giải. Gọi A là biến cố "sản phẩm được kiểm tra là loại 1", A_i là biến cố "sản phẩm được kiểm tra do phân xưởng i sản xuất" $(i = 1, 2, 3)$.

Ta có A_1, A_2, A_3 lập thành một nhóm đầy đủ các biến cố và A chỉ xảy ra đồng thời với 1 trong 3 biến cố đó. Áp dụng công thức xác suất toàn phần ta có

$$P(A) = P(A1)P(A/A1) + P(A2)p(A/A2) + P(A3)P(A/A3).$$

Trong đó, ta tính được $P(A_1) = 20\% = 0,2; P(A/A_1) = 0,7; P(A_2) = 50\% = 0,5; P(A/A_2) = 0,8;\ P(A_3) = 30\% = 0,3; P(A/A_3) = 0,6.$

Vậy $P(A) = 0,2.0,7 + 0,5.0,8 + 0,3.0,6 = 0,72.$

b. Công thức Bayes

Định lí 1.44. *Giả sử $A_1, A_2, ..., A_n$ là nhóm các biến cố đầy đủ và A là một biến cố bất kỳ nào đó có thể xảy ra đồng thời với một trong các biến cố trên. Khi đó:*

$$P\left(A_i/A\right) = \frac{P\left(A_i\right).P\left(A/A_i\right)}{P\left(A\right)},\ i = 1, 2, ..., n,$$

trong đó $P(A) = \sum\limits_{i=1}^{n} P(A_i)P(A/A_i)$.

Chứng minh: Áp dụng công thức xác suất có điều kiện ta có :

$$P(A_k/A) = \frac{P(A_kA)}{P(A)} = \frac{P(A_k)P(A/A_k)}{P(A)} = \frac{P(A_k)P(A/A_k)}{\sum\limits_{i=1}^{n} P(A_i)P(A/A_i)}.$$

Vậy ta có điều phải chứng minh.

Ví dụ 1.45. Một nhà máy có 3 phân xưởng sản xuất ra cùng một loại sản phẩm. Xác suất để phân xưởng 1, phân xưởng 2, và phân xưởng 3 sản xuất được sản phẩm loại I lần lượt là 0,7; 0,8; 0,6. Từ 1 lô hàng gồm 20% sản phẩm của phân xưởng 1; 50% sản phẩm của phân xưởng 2; 30% sản phẩm của phân xưởng 3 người ta lấy ra 1 sản phẩm để kiểm tra. Biết sản phẩm được kiểm tra là loại 1.

a. Tính xác suất để sản phẩm đó thuộc phân xưởng 2?

b. Hỏi sản phẩm đó có khả năng thuộc phân xưởng nào nhất?

Giải. Gọi A là biến cố "sản phẩm được kiểm tra là loại 1", A_i là biến cố "sản phẩm được kiểm tra do phân xưởng i sản xuất" $(i = 1, 2, 3)$.

Ta có A_1, A_2, A_3 lập thành một nhóm các biến cố đầy đủ và A chỉ xảy ra đồng thời với 1 trong 3 biến cố đó.

Theo Ví dụ 1.43 ta có $P(A_1) = 20\% = 0,2; P(A/A_1) = 0,7;\ P(A_2) = 50\% =$

$0,5; P(A/A_2) = 0,8; P(A_3) = 30\% = 0,3; P(A/A_3) = 0,6.$ và $P(A) = 0,72.$

a. Áp dụng công thức Bayes ta có

$$P(A_2/A) = \frac{P(A_2)P(A/A_2)}{P(A)} = \frac{0,5.0,8}{0,72} = \frac{20}{36}.$$

b. Áp dụng công thức Bayes ta có

$$P(A_1/A) = \frac{P(A_1)P(A/A_1)}{P(A)} = \frac{0,2.0,7}{0,72} = \frac{7}{36},$$

$$P(A_2/A) = \frac{P(A_2)P(A/A_2)}{P(A)} = \frac{0,5.0,8}{0,72} = \frac{20}{36},$$

$$P(A_3/A) = \frac{P(A_3)P(A/A_3)}{P(A)} = \frac{0,3.0,6}{0,72} = \frac{9}{36}.$$

Ta thấy $P(A_2/A) > P(A_3/A) > P(A_1/A)$ do đó sản phẩm loại 1 do có khả năng thuộc phân xưởng 2 nhiều hơn.

1.4.4. Công thức Bernoulli

a. Dãy phép thử độc lập

Định nghĩa 1.46. *Một dãy các phép thử được gọi là độc lập với nhau nếu xác suất để xảy ra một biến cố nào đó trong từng phép thử sẽ không phụ thuộc vào việc biến cố đó có xáy ra ở các phép thử khác hay không.*

Ví dụ 1.47. 1. Tung nhiều lần một đồng xu sẽ tạo nên các phép thử độc lập.
2. Lấy nhiều lần sản phẩm từ một lô sản phẩm theo phương thức có hoàn lại cũng tạo nên các phép thử độc lập.

b. Dãy phép thử Bernoulli

Định nghĩa 1.48. *Tiến hành một dãy n phép thử độc lập. 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, xác suất không xảy ra của biến cố A trong mỗi phép thử đều bằng q =1-p. Dãy phép thử thỏa mãn cả 3 điều giả thiết trên được gọi là dãy phép thử Bernoulli hay lược đồ Bernoulli .*

Ví dụ 1.49. 1. Xét một dãy phép thử độc lập: gieo một đồng xu 10 lần. Trong mỗi phép thử, xác suất xảy ra biến cố A : "Đồng xu xuất hiện mặt sấp" đều như nhau và bằng $0,5$, Xác suất để biến cố A không xảy ra, tức là đồng xu xuất hiện mặt ngửa trong mỗi pheps thử cũng đều bằng $0,5$. Do đó dãy 10 phép thử trên là một dãy phép thử Bernoulli.
2. Một đề thi trắc nghiệm môn Toán có 50 câu, mỗi câu có 4 phương án trả lời trong đó có đúng một phương án trả lời đúng. Một học sinh X làm hết bài thi bằng cách chọn đáp án ngẫu nhiên. Khi đó học sinh X đã thực hiện 50 phép thử độc lập. Trong mỗi phép thử này, xác suất trả lời đúng của mỗi câu là như nhau và bằng $\frac{1}{4}$, còn xác xuất trả lời sai đều bằng $\frac{3}{4}$. Do đó dãy phép thử này là dãy phép thử Bernoulli.

c. Công thức Bernoulli

Định lí 1.50. *Cho một dãy n phép thử Bernoulli với xác suất xuất hiện biến cố A trong mỗi phép thử là p. Khi đó:*
1. Xác suất để biến cố A xuất hiện đúng k lần trong dãy n phép thử Bernoulli là $P_n(k) = C_n^k.p^k q^{n-k}$, với $k = 0, 1, ..., n$.
2. Xác suất để biến cố A xuất hiện từ k_1 đến k_2 lần trong dãy n phép thử Bernoulli là $P_n(k_1, k_2) = \sum_{k=k_1}^{k_2} C_n^k p^k q^{n-k}$.

Ví dụ 1.51. Trong phân xưởng có 5 máy hoạt động độc lập. Xác suất để trong mỗi ca mỗi máy bị hỏng đều bằng 0,1. Tìm xác suất để trong ca đó có đúng 2 máy hỏng.

Giải. Coi sự hoạt động của mỗi máy là một phép thử, ta có 5 phép thử độc lập. Trong mỗi phép thử chỉ có hai trường hợp hoặc máy hỏng, hoặc máy chạy tốt. Xác suất hỏng của mỗi máy đều bằng 0,1. Như vậy bài toán thoả mãn lược đồ Bernoulli. Do đó xác suất để trong ca có đúng 2 máy hỏng được tính bằng công thức Bernoulli như sau:

$$P_5(2) = C_5^2.p^2q^3 = C_5^2(0,1)^2.(0,9)^3 = 10.(0,1)^2(0,9)^3 = 0,0729.$$

Ví dụ 1.52. Một bác sĩ có xác suất chữa khỏi bệnh là 0,8. Có người nói rằng cứ 10 người đến chữa bệnh thì chắc chắn có 8 người khỏi bệnh. Điều khẳng định đó có đúng không?

Giải. Ta xem việc chữa bệnh cho 10 người là một dãy 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:

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

Vậy, điều khẳng định đó là sai.

Ví dụ 1.53. Một xạ thủ bắn lần lượt 6 viên đạn vào bia một cách ộc lập với xác suất trúng đích của mỗi viên là 0,7. Tìm xác suất để

a. Có 3 viên trúng bia.

b. Có ít nhất 3 viên trúng bia.

Giải. Xạ thủ bắn 6 viên đạn vào bia chính là thực hiện một dãy 6 phép thử Bernoulli.

a. Áp dụng công thức Bernoulli: $P_6(3) = C_6^3(0,7)^3(0,3)^3 = 0,18522.$

b. Áp dụng công thức Bernoulli:

$$\begin{aligned}
P_6(3;6) &= P_6(3) + P_6(4) + P_6(5) + P_6(6)\\
&= C_6^3(0,7)^3(0,3)^3 + C_6^4(0,7)^4(0,3)^2\\
&= 0,83229.
\end{aligned}$$

d. Số có khả năng nhất (trong lược đồ Bernoulli)

Định nghĩa 1.54. *Số x_0 mà tại đó xác suất đạt giá trị lớn nhất gọi là số có khả năng (hay số lần xuất hiện chắc chắn nhất).*

Định lí 1.55. *Cho một dãy n phép thử Bernoulli với xác suất xuất hiện biến cố A trong mỗi phép thử là p và xác suất không xuất hiện biến cố A trong mỗi phép thử là $q = 1 - p$. Khi đó:*

• *Nếu $np - q \in \mathbb{Z}$ thì số có khả năng nhất được xác định là $x_0 = np - q$ hoặc $x_0 = np - q + 1$.*

• *Nếu $np - q \notin \mathbb{Z}$ thì số có khả năng nhất được xác định là $x_0 = [np - q] + 1$, trong đó $[np - q]$ là phần nguyên của $np - q$.*

Ví dụ 1.56. Xác suất để mỗi con lợn khi tiêm phòng bằng một loại vacxin được miễn dịch là 0,9. Có 50 con lợn được tiêm phòng. Tìm số lợn được miễn dịch có khả năng nhiều nhất?

Giải. Tiêm phòng cho 50 con lơn chính là thực hiện một dãy 50 phép thử Bernoulli với $p = 0,9$ và $q = 0,1$. Ta có $np - q = 50.0,9 - 0,1 = 44,9 \notin \mathbb{Z}$ nên $x_0 = [np - q] + 1 = [44,9] + 1 = 45$.

Vậy số lợn có khả năng miễn dịch nhất là 45 con.

1.5. Bài tập chương 1

Bài 1.1. Một nhóm học sinh gồm 12 em trong đó có 5 em lớp 10A, 4 học sinh lớp 10B và 3 học sinh lớp 10C. Chọn ngẫu nhiên đồng thời 4 em trong nhóm.
a. Tính xác suất sao cho 4 em được chọn không thuộc cùng một lớp.
b. Tính xác suất sao cho 4 em được chọn có số học sinh lớp 10A bằng số học sinh lớp 10C .

Bài 1.2. Một hộp đựng 5 quả cầu xanh, 4 quả cầu đỏ, 3 quả cầu vàng. Lấy ngẫu nhiên đồng thời 4 quả trong hộp.
a. Tính xác suất để 4 quả lấy ra có 2 quả đỏ, 1 quả xanh và 1 quả vàng ?
b. Tính xác suất để 4 quả lấy được thuộc cả 3 màu nói trên ?

Bài 1.3. Có 2 hộp phân biệt đựng các quả cầu. Hộp I đựng 4 quả cầu xanh, 3 quả cầu đỏ. Hộp II đựng 3 quả cầu xanh, 5 quả cầu đỏ. Lấy ngẫu nhiên mỗi hộp một quả cầu.

a. Tính xác suất để hai quả lấy được đều là màu xanh?

b. Tính xác suất để hai quả lấy được có 1 quả màu xanh và 1 quả màu đỏ?

Bài 1.4. Một hộp kín đựng 16 tấm thẻ được đánh số từ 1 đến 16. Lấy ngẫu nhiên đồng thời 4 tấm thẻ trong hộp.

a. Tính xác suất sao cho 4 tấm thẻ lấy ra đều được đánh số lẻ?

b. Tính xác suất sao cho 4 tấm thẻ lấy ra có đúng 2 tấm được đánh số chẵn?

Bài 1.5. Một chiếc kim của bánh xe trong trò chơi " Chiếc nón kỳ diệu" có thể dừng lại ở một trong 7 vị trí với khả năng như nhau.

a.Tính xác suất để trong 3 lần quay liên tiếp chiếc kim của bánh xe đó lần lượt dừng lại ở 3 vị trí khác nhau ?

b.Tính xác suất để trong 3 lần quay liên tiếp chiếc kim của bánh xe đó dừng lại ở đúng một vị trí.

Bài 1.6. Ba vận động viên ném bóng rổ, mỗi người ném một quả với xác suất ném trúng rổ của từng người là 0.7, 0.9, 0.8 tương ứng. Tính xác suất để:

a. Có đúng 2 người ném trúng rổ ?

b. Có ít nhất một nười ném trúng rổ ?

c. Người thứ 2 ném trúng rổ biết rằng có 2 người ném trúng rổ ?

Bài 1.7. Trong một phân xưởng có 3 máy (I, II, III) hoạt động độc lập nhau với xác suất hỏng trong một ca làm việc của các máy tương ứng là 0,01; 0.1; 0,5. Tính xác suất để trong một ca làm việc có:

a. Ít nhất một máy bị hỏng ?

b. Có đúng một máy không bị hỏng ?

c. Giả sử trong ca làm việc đó có đúng 1 máy không bị hỏng, tính xác suất để đó là máy I.

Bài 1.8. Ba xạ thủ, mỗi người bắn một viên đạn vào một mục tiêu một cách độc lập. Xác suất bắn trúng đích của mỗi người tương ứng là 0,8; 0,85 và

0,9.

a. Tính xác suất để có hai viên đạn trúng đích ?

b. Giả sử mục tiêu bị tiêu diệt với xác suất 0,7 nếu có 1 viên đạn trúng đích; 0,9 nếu có 2 viên đạn trúng đích và chắc chắn bị tiêu diệt nếu cả ba viên trúng đích. Tính xác suất để mục tiêu bị tiêu diệt ?

Bài 1.9. Một thành phố có 45% dân số là nam và 55% dân số là nữ. Biết rằng nam giới trong thành phố có 20% tốt nghiệp đại học còn nữ giới là 15%. Vào thành phố bạn gặp 1 người.

a. Tính xác suất để người đó tốt nghiệp đại học ?

b. Tính xác suất để người đó là nam biết rằng người đó đã tốt nghiệp đại học ?

Bài 1.10. Có 2 lô sản phẩm cùng loại, cùng kích thước. Lô I có 12 chính phẩm và 3 phế phẩm, lô II có 15 chính phẩm và 3 phế phẩm. Lấy ngẫu nhiên 1 sản phẩm từ lô I bỏ sang lô II, sau đó từ lô II lấy ra 2 sản phẩm. Tính xác suất sao cho 2 sản phẩm lấy ra từ lô II đều là chính phẩm ?

Bài 1.11. Tỷ lệ hút thuốc lá ở một vùng là 35%. Theo thống kê cho biết, tỷ lệ viêm họng trong số người hút thuốc là 65%, còn trong số không hút thuốc là 30%. Khám ngẫu nhiên một người thì thấy anh ta bị viêm họng, tính xác suất để anh ta là người hút thuốc lá ?

Bài 1.12. Một thành phố có 45% dân số là nam và 55% dân số là nữ. Biết rằng nam giới trong thành phố có 20% tốt nghiệp đại học còn nữ giới là 15%. Vào thành phố bạn gặp 1 người. Tính xác suất để người đó tốt nghiệp đại học ?

Bài 1.13. Một máy phát đi 5 tín hiệu với xác suất phát thành công một tín hiệu là 0.99. Tại một máy thu, xác suất nhận đúng tín hiệu phát là 0.95. Tính xác suất để trong 5 tín hiệu đã phát có 4 tín hiệu phát thành công và máy thu nhận được 3 tín hiệu chính xác ?

Bài 1.14. Có hai thư mục. Thư mục thứ nhất có 3 file .doc và 3 file .xls. Thư mục thứ hai có 6 file .doc và 4 file .xls. Di chuyển ngẫu nhiên 4file từ thư mục thứ nhất sang thư mục thứ hai rồi chọn ngẫu nhiên một 1 file ở thư mục thứ hai. Tính xác suất để chọn được file .xls ở thư mục thứ hai ?

Bài 1.15. Một trạm chỉ phát hai loại tín hiệu A và B với xác suất tương ứng 0,8 và 0,2. Do có nhiễu trên đường truyền nên 1/5 tín hiệu A bị méo và được thu như là tín hiệu B, còn 1/8 tín hiệu B bị méo thành tín hiệu A.
a. Tìm xác suất thu được tín hiệu A ? b. Giả sử thu được tín hiệu A, tìm xác suất để thu được đúng tín hiệu lúc phát ?

Bài 1.16. Tỷ lệ hút thuốc lá ở một vùng là 35%. Theo thống kê cho biết, tỷ lệ viêm họng trong số người hút thuốc là 65%, còn trong số không hút thuốc là 30%. Khám ngẫu nhiên một người thì thấy anh ta bị viêm họng, tính xác suất để anh ta là người hút thuốc. Nếu anh ta không bị viêm họng thì xác suất đó bằng bao nhiêu ?

Bài 1.17. Một công ty bảo hiểm chia đối tượng bảo hiểm làm 3 loại: ít rủi ro (chiếm 20%), rủi ro trung bình (chiếm 50%), rủi ro cao (chiếm 30%). Biết tỷ lệ khách hàng gặp rủi ro trong một năm tương ứng với các đối tượng trên là: 0,05; 0,15 và 0,3.
a. Tính tỷ lệ khách hàng gặp rủi ro trong 1 năm ?
b. Gặp một khách hàng bị rủi ro, tính xác suất để người đó thuộc loại ít rủi ro ?

Bài 1.18. Một xí nghiệp may có hai phân xưởng với tỷ lệ phế phẩm tương ứng là: $1\%; 2\%$,
- Phân xưởng 1: sản xuất 40% sản phẩm
- Phân xưởng 2: sản xuất 60% sản phẩm

a. Tìm xác suất để từ 1 lô sản phẩm của xí nghiệp chọn ngẫu nhiên được 1 phế phẩm ?

b. Giả sử lấy được 1 phế phẩm. Tìm xác suất để phế phẩm này do phân xưởng 1 sản xuất; phân xưởng 2 sản xuất ?

Bài 1.19. Một thành phố có 45% dân số là nam và 55% dân số là nữ. Biết rằng nam giới trong thành phố có 20% tốt nghiệp đại học còn nữ giới là 15%. Vào thành phố bạn gặp 1 người. Vào thành phố gặp được người tốt nghiệp đại học. Tính xác suất để người đó là nữ giới ?

Bài 1.20. Có 2 lô hàng, lô I có 30 chính phẩm và 10 phế phẩm. Lô II có 25 chính phẩm và 15 phế phẩm. Từ lô I lấy ngẫu nhiên ra 1 sản phẩm bỏ sang lô II, sau đó từ lô II lấy ngẫu nhiên ra 2 sản phẩm.

a. Tính xác suất để được cả 2 chính phẩm ?

b. Tính xác suất để lấy được 1 phế phẩm và 1 chính phẩm ?

Bài 1.21. Một xưởng có hai máy cùng sản xuất một loại sản phẩm. Xác suất để sản xuất ra phế phẩm của mỗi máy tương ứng là 0,1 và 0,2. Đồng thời công suất của máy 2 gấp đôi công suất máy 1.

a. Tính tỷ lệ phế phẩm của xưởng đó ?

b. Lấy ngẫu nhiên 2 sản phẩm của xưởng thì thấy chỉ có một sản phẩm tốt. Tính xác suất để sản phẩm tốt đó là của máy 1 ?

Bài 1.22. Một nhà máy có 3 phân xưởng 1, 2, 3. Biết phân xưởng 1 chiếm 40% sản phẩm của nhà máy, phân xưởng 2 chiếm 30% sản phẩm của nhà máy và phân xưởng 3 chiếm 30% sản phẩm của nhà máy. Biết tỉ lệ phế phẩm của phân xưởng 1, 2, 3 lần lượt là $10\%, 20\%$ và 15%. Lấy ngẫu nhiên một sản phẩm của nhà máy và được một phế phẩm, tính xác suất sao cho sản phẩm lấy được là thuộc phân xưởng 1 ?

Bài 1.23. Một nhà máy có 3 phân xưởng cùng sản xuất ra một loại sản phẩm. Sản phẩm của phân xưởng I chiếm 40% , phân xưởng II chiếm 30% sản lượng của nhà máy. Tỷ lệ chính phẩm của từng phân xưởng tương ứng là $94\%, 96\%, 95\%$. Lấy ngẫu nhiên 1 sản phẩm của nhà máy, tính xác suất để sản phẩm đó là chính phẩm ?

Bài 1.24. Một đề thi trắc nghiệm gồm 10 câu hỏi, mỗi câu có 4 phương án trả lời, trong đó chỉ có một phương án trả lời đúng. Mỗi câu trả lời đúng được 5 điểm, mỗi câu trả lời sai bị trừ 2 điểm. Một thí sinh chọn cách trả lời một cách hoàn toàn hú họa.

a. Tính xác suất để thí sinh được 15 điểm ?

b. Tính xác suất để thí sinh đó đỗ, biết rằng để đỗ phải được ít nhất 28 điểm?

Bài 1.25. Một người say rượu bước 8 bước. Mỗi bước anh ta tiến lên phía trước 1 mét hoặc lùi lại phía sau 1 mét với xác suất như nhau. Tính xác suất để sau 8 bước:

a. Anh ta trở lại điểm xuất phát ?

b. Anh ta cách điểm xuất phát hơn 4m ?

Chương 2
ĐẠI LƯỢNG NGẪU NHIÊN

Trong chương này ta khảo sát các biến cố của đại lượng nhận các giá trị nào đó, khi các giá trị này thay đổi ta được các biến ngẫu nhiên. Khái niệm biến ngẫu nhiên và các đặc trưng của chúng là những khái niệm quan trọng của lý thuyết xác suất.

2.1. Đại lượng ngẫu nhiên và bảng phân phối xác suất

2.1.1. Đại lượng ngẫu nhiên và phân loại

Định nghĩa 2.1. *Đại lượng ngẫu nhiên (ĐLNN) là đại lượng biến đổi biểu thị giá trị kết quả của một phép thử ngẫu nhiên.*

Ký hiệu: $X, Y, Z...$ hoặc $X_1, ..., X_n, ..Y_1, Y_2, ..., Y_n, ...$
Các giá trị có thể có của chúng được ký hiệu $x_1, x_2, ..., x_n, y_1, y_2, ..., y_n$.

Chú ý: X gọi là đại lượng ngẫu nhiên vì trước khi tiến hành phép thử ta chưa có thể nói một cách chắc chắn nó sẽ nhận giá trị bằng bao nhiêu, mà chỉ dự đoán điều đó với một xác suất nhất định. Nói cách khác, việc X nhận một giá trị nào đó $(X = x_1); (X = x_2); ...; (X = x_n)$ về thực chất là các biến cố ngẫu nhiên. Hơn nữa, trong kết quả của phép thử đại lượng ngẫu nhiên X nhất định sẽ nhận một và chỉ một trong các giá trị có thể có của nó, do đó các biến cố $(X = x_1); (X = x_2); ...; (X = x_n)$ tạo nên nhóm đầy đủ các biến cố.

Ví dụ 2.2. Tung một con xúc xắc. Gọi X là "số chấm xuất hiện" thì X là một đại lượng ngẫu nhiên vì trong kết quả của phép thử nó sẽ nhận 1 trong 6 giá trị có thể có là 1,2,3,4,5,6.

Ví dụ 2.3. Trong hộp có 4 bi trắng, 2 bi đen. Lấy ngẫu nhiên 2 bi, gọi X là số bi trắng thì Xlà đại lượng ngẫu nhiên vì X có thể là 0, 1, 2.

Đại lượng ngẫu nhiên được chia ra làm hai loại: Đại lượng ngẫu nhiên rời rạc và đại lượng ngẫu nhiên liên tục.

★ Đại lượng ngẫu nhiên rời rạc

Định nghĩa 2.4. *Đại lượng ngẫu nhiên gọi là rời rạc nếu nó chỉ nhận một số hữu hạn hoặc một số vô hạn đếm được các giá trị.*

Nói cách khác đại lượng ngẫu nhiên sẽ là rời rạc nếu ta có thể liệt kê được tất cả các giá trị có thể có của nó.

Ta ký 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 2.5. 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.

Ví dụ 2.6. Gọi Y là "số người vào mua hàng tại một cửa hàng trong một ngày". Y là đại lượng ngẫu nhiên rời rạc với các giá trị có thể có của nó lập nên một tập đếm được $Y = 0, 1, 2, ...$

Ví dụ 2.7. Một phân xưởng có 5 máy hoạt động. Gọi X là "số máy hỏng trong một ca". X là đại lượng ngẫu nhiên rời rạc với các giá trị có thể có của nó là $X = 0, 1, 2, 3, 4, 5.$

★ Đại lượng ngẫu nhiên liên tục

Định nghĩa 2.8. *Đại lượng ngẫu nhiên gọi là liên tục nếu các giá trị có thể có của nó lấp đầy một khoảng trên trục số.*

Như vậy, đối với đại lượng ngẫu nhiên liên tục ta không thể liệt kê được tất cả các giá trị có thể có của nó.

Ví dụ 2.9. Gọi X là thời gian đến điểm hẹn của A từ 7 đến 8 giờ thì X là đại lượng ngẫu nhiên liên tục vì ta không thể kể ra được tất cả các giá trị có thể có của nó. Ta có thể nói rằng các giá trị có thể có của Xnằm trong một khoảng $(7;8)$.

Ví dụ 2.10. Gọi Y là "năng suất lúa vụ mùa một tỉnh", Y là đại lượng ngẫu nhiên liên tục.

2.1.2. Bảng phân phối xác suất

Bảng phân phối xác suất chỉ dùng để thiết lập quy luật phân phối xác suất của đại lượng ngẫu nhiên rời rạc.

Định nghĩa 2.11. *Giả sử đại lượng ngẫu nhiên rời rạc X có thể nhận một trong các giá trị có thể có là $x_1, x_2, ..., x_n$ với các xác suất tương ứng là $p_1, p_2, ..., p_n$. Khi đó bảng phân phối các xác suất của đại lượng ngẫu nhiên rời rạc X có dạng:*

X	x_1	x_2	x_3	...	x_n
$P(X = x_i)$	p_1	p_2	p_3	...	p_n

Nếu các giá trị có thể có 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. Do đó: $0 \leq p_i \leq 1$ với mọi $i = 1, 2, ..., n$ và $\sum\limits_{i=1}^{n} p_i = 1$.

Ví dụ 2.12. Tung một con xúc sắc. Gọi X là "số chấm xuất hiện". Lập bảng phân phối xác suất của X.

Giải. Vì X là đại lượng ngẫu nhiên rời rạc với các giá trị có thể có 1,2,3,4,5,6 với các xác suất tương ứng đều bằng 1/6, do đó bảng phân phối xác suất của

X có dạng

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

Kiểm tra: $\sum_{i=1}^{6} p_i = 1$.

Ví dụ 2.13. Trong hộp có 10 sản phẩm trong đó có 6 chính phẩm. Lấy ngẫu nhiên 2 sản phẩm. Lập bảng phân phối xác suất của số chính phẩm được lấy ra.

Giải. Gọi Y là "số chính phẩm được lấy ra trong 2 sản phẩm, Y là đại lượng ngẫu nhiên rời rạc với các giá trị có thể có là 0,1,2 . Ta tính xác suất tương ứng:

$$P(Y=0) = \frac{C_4^2}{C_{10}^2} = \frac{2}{15}; P(Y=1) = \frac{C_6^1.C_4^1}{C_{10}^2} = \frac{8}{15}; P(Y=2) = \frac{C_6^2}{C_{10}^2} = \frac{5}{15}.$$

Vậy quy luật phân phối xác suất của Y là

Y	0	1	2
P(Y)	$\frac{2}{15}$	$\frac{8}{15}$	$\frac{5}{15}$

Kiểm tra: $2/15 + 8/15 + 5/15 = 1$.

Ví dụ 2.14. Xác suất để xạ thủ bắn trúng bia là 0,8. Xạ thủ được phát từng viên đạn để bắn cho đến khi trúng bia. Lập bảng phân phối xác suất của số viên đạn được phát.

Giải. Gọi X là "số viên đạn mà xạ thủ được phát". X là đại lượng ngẫu nhiên rời rạc với các giá trị có thể có là $1, 2, 3, ..., n, ..$ Tính xác suất tương ứng $p(X=1) = 0,8; p(X=2) = 0,2.0,8; ...; p(X=n) = (0,2)^{n-1}.0,8...$ Vậy bảng phân phối xác suất của X là

X	1	2	...	n	...
P(X)	0,8	0,2.0,8	...	$(0,2)^{n-1}.0,8$	...

Kiểm tra: $0,8 + 0,2.0,8 + ... + (0,2)^{n-1}.0,8...$

Đây là tổng của một cấp số nhân lùi vô hạn với công bội $q = 0,2$, do đó $\sum_{i=1}^{\infty}(0,2)^{n-1}.0,8 = \frac{0,8}{1-0,2} = 1.$

Hạn chế của bảng phân phối xác suất là nó chỉ thiết lập được quy luật phân phối xác suất của đại lượng ngẫu nhiên rời rạc mà thôi.

2.2. Hàm phân phối xác suất

2.2.1. Định nghĩa hàm phân phối xác suất

Khái niệm hàm phân phối xác suất áp dụng được đối với cả đại lượng ngẫu nhiên rời rạc và liên tục. Giả sử X là đại lượng ngẫu nhiên bất kỳ, x là một số thực nào đó. Xét biến cố "đại lượng ngẫu nhiên X nhận giá trị nhỏ hơn x", ký hiệu $(X < x)$. Hiển nhiên khi x thay đổi thì xác suất P$(X < x)$ cũng thay đổi. Như vậy xác suất này là một hàm số của x.

Định nghĩa 2.15. *Hàm phân phối xác suất của đại lượng ngẫu nhiên X, ký hiệu là $F_X(x)$ là hàm được xác định như sau:* $F_X(x) = P(X < x)$.

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

$$F(x) = \begin{cases} 0 & khi\ x \le x_1, \\ p_1 + p_2 + ... + p_{i-1} & khi\, x_{i-1} < x \le x_i, \\ 1 & khi\ x > x_i. \end{cases}$$

- Nếu X là đại lượng ngẫu nhiên liên tục thì hàm phân phối xác suất được xác định bởi công thức:

$$F_X(x) = \int_{-\infty}^{x} f(u)du,$$

trong đó $f(u)$ là hàm mật độ xác suất của đại lượng ngẫu nhiên liên tục X.

Ví dụ 2.16. Đại lượng ngẫu nhiên rời rạc X có bảng phân phối xác suất

X	*1*	*3*	*4*
P(X)	*0,1*	*0,5*	*0,4*

Hãy xây dựng hàm phân phối xác suất và vẽ đồ thị.

Giải. - Nếu $x \le 1$ thì biến cố $(X < x)$ là biến cố không thể có do đó:

$$F_X(x) = P(X < x) = P(\phi) = 0$$

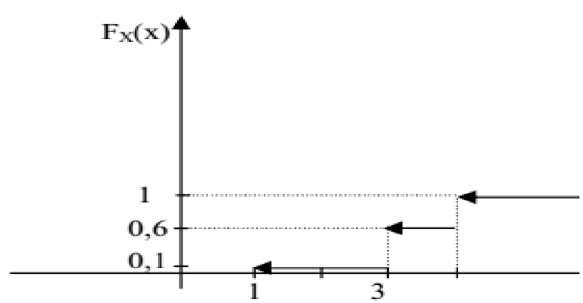
- Nếu $1 < x \le 3$ thì $F_X(x) = P(X < x) = P(x = 1) = 0,1.$

- Nếu $3 < x \le 4$ thì $F_X(x) = P(X < x) = P(x = 1) + P(x = 3) = 0,1 + 0,5.$

- Nếu $x > 4$ thì $F_X(x) = P(X < x) = P(x = 1) + P(x = 3) + P(x = 4) = 0,1 + 0,5 + 0,4 = 1.$

Vậy hàm phân phối xác suất của X là

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



Hình 2.1 Đồ thị của F(x)

Như vậy đồ thị của hàm phân phối xác suất của đại lượng ngẫu nhiên rời rạc có dạng bậc thang với số điểm gián đoạn chính bằng số giá trị có thể có của X. Nếu X là đại lượng ngẫu nhiên liên tục thì hàm phân phối xác suất của nó liên tục và khả vi tại mọi điểm của X, do đó đồ thị của nó là một đường cong liên tục.

2.2.2. Các tính chất của hàm phân phối xác suất

Tính chất 1: $0 \leq F_X(x) \leq 1$ với mọi x.

Thật vậy: Vì $F_X(x) = P(X < x)$, suy ra điều phải chứng minh.

Tính chất 2: Hàm phân phối xác suất là hàm không giảm. Tức là nếu $x_1 < x_2$ thì $F_X(x_1) \leq F_X(x_2)$.

Thật vậy: Giả sử $x_1 < x_2$. Xét biến cố

$$(X < x_2) = (X < x_1) + (x_1 \leq X < x_2).$$

Khi đó

$$P(X < x_2) = P[(X < x_1) + (x_1 \leq X < x_2)] = P(X < x_1) + P(x_1 \leq X < x_2).$$

Hay $P(X < x_2) - P(X < x_1) = P(x_1 \leq X < x_2) \geq 0$,

hay $F_X(x_2) - F_X(x_1) \geq 0$ hay $F_X(x_2) \geq F_X(x_1)$ (đpcm).

Hệ quả 1: $P(a \leq X < b) = F_X(b) - F_X(a)$.

Hệ quả này suy ra trực tiếp từ quá trình chứng minh tính chất 2.

Hệ quả 2: Nếu X là đại lượng ngẫu nhiên liên tục thì $P(X = x) = 0$.

Thật vậy, nếu đặt $a = x, b = x + \Delta x$ thì

$$P(a \leq x < b) = P(x \leq X < x + \Delta x) = F_X(x + \Delta x) - F_X(x).$$

Do đó

$$\lim_{\Delta x \to 0} P(x \leq X < x + \Delta x) = \lim_{\Delta x \to 0} F_X(x + \Delta x) - F_X(x) = F(x) - F(x) = 0.$$

(Vì X liên tục nên $F_X(x)$ liên tục $\lim\limits_{x \to 0} F(x + \Delta x) = F(x)$).

Hệ quả 3: Nếu X là đại lượng ngẫu nhiên liên tục thì

$$P(a \leq x < b) = P(a \leq x \leq b) = P(a < x < b) = P(a < x \leq b).$$

Hệ quả này được suy ra từ hệ quả 2

Tính chất 3: $\lim\limits_{x\to+\infty} F(x) = 1;\ \lim\limits_{x\to+-\infty} F(x) = 0.$

Thật vậy:

$$F_X(-\infty) = P(X < -\infty) = P(\phi) = 0, F_X(+\infty) = P(X < +\infty) = P(\Omega) = 1.$$

2.2.3. Ý nghĩa của hàm phân phối

Hàm phân phối xác suất phản ánh mức độ tập trung xác suất về phía bên trái một số thực x nào đó.

Ví dụ 2.17. Đại lượng ngẫu nhiên X có hàm phân phối xác suất

$$F_X(x) = \begin{cases} 0 & khi\ x \le -1, \\ \dfrac{3}{4}x + \dfrac{3}{4} & khi\ \ -1 < x \le \dfrac{1}{3}, \\ 1 & khi\ x > \dfrac{1}{3}. \end{cases}$$

Tìm xác suất để trong kết quả của phép thử X nhận giá trị trong khoảng $[0, 1/3)$.

Giải. Theo tính chất 2 ta có

$$P(0 \le x < 1/3) = F_X(1/3) - F_X(0) = [\frac{3}{4}.\frac{1}{3} + \frac{3}{4}] - [\frac{3}{4}.0 + \frac{3}{4}] = \frac{1}{4}.$$

Ví dụ 2.18. Cho đại lượng ngẫu nhiên liên tục X có hàm phân phối

$$F_X(x) = \begin{cases} 0 & khi\ x \le -a, \\ A + B.\arcsin\dfrac{x}{a} & khi\ \ -a < x \le a, \\ 1 & khi\ x > a, \end{cases}$$

trong đó $a > 0$. Hãy tìm A, B?

Giải. Theo tính chất của hàm phân phối xác suất $0 \le F_X(x) \le 1$.

$$0 \le A + B\arcsin\frac{x}{a} \le 1$$

Mặt khác vì X liên tục nên $F_X(x)$ liên tục. Do đó

$$\lim_{x\to -a+0} F_X(x) = \lim_{x\to -a-0} F_X(x) = F_X(-a) \Leftrightarrow \lim_{x\to -a+0} (A + B\arcsin\frac{x}{a}) = 0$$

$$\Leftrightarrow A + B(-\frac{\pi}{2}) = 0. \quad \textbf{(2.1)}$$

$$\lim_{x\to a+0} F_X(x) = \lim_{x\to a-0} F_X(x) = F_X(a) \Leftrightarrow \lim_{x\to a-0} (A + B\arcsin\frac{x}{a}) = 1$$

$$\Leftrightarrow A + B(\frac{\pi}{2}) = 1. \quad \textbf{(2.2)}$$

Từ (2.1) và (2.2) suy ra $A = \frac{1}{2}; B = \frac{1}{\pi}$.

2.3. Hàm mật độ xác suất

2.3.1. Định nghĩa hàm mật độ xác suất và ví dụ

Hàm mật độ xác suất đặc trưng cho quy luật phân phối của đại lượng ngẫu nhiên liên tục X.

Định nghĩa 2.19. *Hàm mật độ xác suất của đại lượng ngẫu nhiên liên tục X (ký hiệu là $f(x)$) là đạo hàm bậc nhất của hàm phân phối xác suất của đại lượng ngẫu nhiên đó.*

Vậy: $f(x) = F_X^{'}(x)$.

2.3.2. Các tính chất của hàm mật độ xác suất

Tính chất 1: $f(x) \geq 0$ với mọi x.

Thật vậy, vì $F(x)$ là hàm không giảm, do đó $F'(x) = f(x)$ là hàm không âm. Về mặt hình học, đồ thị của hàm $f(x)$ không nằm thấp hơn trục Ox.

Tính chất 2: $P(a < X < b) = \int\limits_a^b f(x)dx$.

Thật vậy, ta có

$$P(a < X < b) \overset{Xlt}{=} P(a \leq X < b) = F_X(b) - F_X(a)$$
$$\overset{dn}{=} \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx.$$

Tính chất 3: $F(x) = \int\limits_{-\infty}^{x} f(x)dx.$

Thật vậy: Theo định nghĩa hàm phân phối xác suất ta có:

$$F(x) = P(X < x) = P(-\infty < X < x).$$

Theo tính chất 2, ta đặt $a = -\infty$ và $b = x$, ta có:

$$P(-\infty < X < x) = \int_{-\infty}^{x} f(x)dx.$$

Công thức trên cho phép tìm hàm phân phối xác suất của đại lượng ngẫu nhiên liên tục khi đã biết hàm mật độ xác suất của nó.

Tính chất 4: $\int\limits_{-\infty}^{+\infty} f(x)dx = 1.$

Thật vậy: $\int\limits_{-\infty}^{+\infty} f(x)dx = P(-\infty < X < +\infty) = P(\Omega) - P(\phi) = 1.$

Chú ý: Để $f(x)$ là hàm mật độ xác suất của đại lượng ngẫu nhiên liên tục nào đó thì nó phải thoả mãn hai điều kiện:

$$\begin{cases} f(x) & \geq 0, \\ \int\limits_{-\infty}^{+\infty} f(x)dx & = 1. \end{cases}$$

Ví dụ 2.20. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ xác suất là:

$$f(x) = \begin{cases} k(3x^2 + 2) & \textbf{\textit{khi}} \ \ x \in [0, 1], \\ 0 & \textbf{\textit{khi}} \ x \notin [0, 1]. \end{cases}$$

a. Tìm hằng số k.

b. Xây dựng hàm phân phối xác suất của X.

Giải. a. Vì $f(x)$ là hàm mật độ xác suất của đại lượng ngẫu nhiên liên tục X nên nó phải thoả mãn:

$$\begin{cases} f(x) & \geq 0, \\ \int\limits_{-\infty}^{+\infty} f(x)dx & = 1. \end{cases}$$

- Điều kiện thứ nhất $\Leftrightarrow k \geq 0$.

- Điều kiện thứ hai $\Leftrightarrow \int\limits_{-\infty}^{0} f(x)\,dx + \int\limits_{0}^{1} f(x)\,dx + \int\limits_{0}^{+\infty} f(x)\,dx = 1$

$\Leftrightarrow \int\limits_{0}^{1} k\left(3x^2+2\right)dx = 1 \Leftrightarrow k = \dfrac{1}{\int_0^1 (3x^2+2)\,dx} = \dfrac{1}{3}$ (thỏa mãn).

b. - Nếu $x \leq 0 \Rightarrow F_X(x) = \int\limits_{-\infty}^{x} f(x)\,dx = 0$.

- Nếu $0 < x \leq 1 \Rightarrow F_X(x) = \int\limits_{-\infty}^{x} f(x)\,dx = \int\limits_{-\infty}^{0} f(x)\,dx + \int\limits_{0}^{x} f(x)\,dx$

$$= \int\limits_{0}^{x} \frac{1}{3}\left(3x^2+2\right)dx = \frac{1}{3}\left(x^3+2x\right).$$

- Nếu $x > 1 \Rightarrow F_X(x) = \int\limits_{-\infty}^{x} f(x)\,dx = \int\limits_{-\infty}^{0} f(x)\,dx + \int\limits_{0}^{1} f(x)\,dx + \int\limits_{1}^{x} f(x)\,dx$

$$= \int\limits_{0}^{1} \frac{1}{3}\left(3x^2+2\right)dx = 1.$$

Vậy

$$F(x) = \begin{cases} 0 & \textbf{khi } x \leq 0, \\ \dfrac{1}{3}(x^3+2x) & \textbf{khi } 0 < x \leq 1, \\ 1 & \textbf{khi } x > 1. \end{cases}$$

Ví dụ 2.21. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ xác suất là:

$$f(x) = \begin{cases} C cosx & \textbf{\textit{khi}} \ \ x \in [\frac{-\pi}{2}, \frac{\pi}{2}], \\ 0 & \textbf{\textit{khi}} \ x \notin [\frac{-\pi}{2}, \frac{\pi}{2}]. \end{cases}$$

a. Tìm hằng số C.

b. Xây dựng hàm phân phối xác suất của X.

c. Tìm xác suất để đại lượng ngẫu nhiên X nhận giá trị trong $(0; \pi/4)$.

Giải. a. Vì $f(x)$ là hàm mật độ xác suất của đại lượng ngẫu nhiên liên tục X nên nó phải thoả mãn:

$$\begin{cases} f(x) & \geq 0, \\ \int\limits_{-\infty}^{+\infty} f(x)dx & = 1. \end{cases}$$

- Điều kiện thứ nhất $\Leftrightarrow C \geq 0$.

Giải điều kiện thứ hai ta có :

$$1 = \int\limits_{-\infty}^{+\infty} f(x)dx = \int\limits_{-\infty}^{-\frac{\pi}{2}} 0dx + \int\limits_{-\frac{\pi}{2}}^{\frac{\pi}{2}} C\cos x dx + \int\limits_{\frac{\pi}{2}}^{+\infty} 0dx = C \sin x \Big|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = 2C \Leftrightarrow C = \frac{1}{2}.$$

Kết hợp hai điều kiện, vậy $C = \dfrac{1}{2}$.

b. $F_X(x) = \int\limits_{-\infty}^{x} f(u)du$, trong đó,

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

Nếu $x \leq -\dfrac{\pi}{2}$ thì $F_X(x) = \int\limits_{-\infty}^{x} 0dx = 0$.

Nếu $-\dfrac{\pi}{2} < x \leq \dfrac{\pi}{2}$ thì $F_X(x) = \int\limits_{-\infty}^{x} f(u)du = \int\limits_{-\infty}^{-\frac{\pi}{2}} 0du + \int\limits_{-\frac{\pi}{2}}^{x} \frac{1}{2}\cos u du$

$= \frac{1}{2}\sin u\Big|_{-\frac{\pi}{2}}^{x} = \frac{1}{2}(\sin x + 1).$

Nếu $x > \frac{\pi}{2}$ **thì** $F_X(x) = \int\limits_{-\infty}^{x} f(u)du = \int\limits_{-\infty}^{-\frac{\pi}{2}} 0du + \int\limits_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{2}\cos u du + \int\limits_{\frac{\pi}{2}}^{x} 0du$

$$= \frac{1}{2}\sin u\Big|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = \frac{1}{2}(1+1) = 1.$$

Vậy hàm phân phối xác suất của X là

$$F(x) = \begin{cases} 0 & \textbf{khi } x \le \frac{-\pi}{2}, \\ \frac{1}{2}(sinx + 1) & \textbf{khi } \frac{-\pi}{2} < x \le \frac{\pi}{2}, \\ 1 & \textbf{khi } x > \frac{\pi}{2}. \end{cases}$$

c. Tính $P(0 < X < \frac{\pi}{4}) = F(\frac{\pi}{4}) - F(0) = \frac{1}{2}(\sin\frac{\pi}{4} + 1) - \frac{1}{2}(\sin 0 + 1) = \frac{\sqrt{2}}{4}.$

Cách khác: $P(0 < X < \frac{\pi}{4}) = \int\limits_{0}^{\frac{\pi}{4}} f(x)dx = \int\limits_{0}^{\frac{\pi}{4}} \frac{1}{2}\cos x dx = \frac{1}{2}\sin x\Big|_{0}^{\frac{\pi}{4}} = \frac{\sqrt{2}}{4}.$

Ví dụ 2.22. Hàm phân phối xác suất của đại lượng ngẫu nhiên liên tục X có dạng $F_X(x) = a + b.arctgx$ ***với*** $-\infty < x < +\infty$.

a. Tìm hệ số a, b.

b. Tìm hàm mật độ xác suất $f(x)$.

c. Tìm xác suất để khi tiến hành 3 phép thử độc lập có 2 lần X nhận giá trị trong khoảng $(-1, 1)$.

Giải. a. Theo tính chất của hàm phân phối xác suất ta có

$$\begin{cases} F_X(-\infty) = 0, \\ F_X(+\infty) = 0 \end{cases} \Leftrightarrow \begin{cases} a + b.\arctan(-\infty) = 0, \\ a + b.\arctan(+\infty) = 1 \end{cases}$$

$$\Leftrightarrow \begin{cases} a + b.\left(-\frac{\pi}{2}\right) = 0, \\ a + b.\left(\frac{\pi}{2}\right) \quad = 1 \end{cases} \Leftrightarrow \begin{cases} a = \frac{1}{2}, \\ b = \frac{1}{\pi}. \end{cases}$$

b. Ta có

$$f(x) = F'_X(x) = \left(\frac{1}{2} + \frac{1}{\pi}arctgx\right)' = \frac{1}{\pi}.\frac{1}{(1+x^2)} = \frac{1}{\pi(1+x^2)}.$$

c. Bài toán thoả mãn lược đồ Bernoulli, ta phải tính $P_3(2) = C_3^2 p^2 q^1$, trong đó

$$p = P(-1 < X < 1) \overset{tc}{=} \int_{-1}^{1} \frac{1}{\pi}\frac{dx}{1+x^2} = \frac{1}{\pi}arctgx\,\Big|_{-1}^{1} = \frac{1}{2}.$$

Vậy $P_3(2) = C_3^2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 = \frac{3}{8} = 0,375.$

Ý nghĩa của hàm mật độ xác suất

Hàm mật độ xác suất của đại lượng ngẫu nhiên liên tục X tại mỗi điểm x cho biết mức độ tập trung xác suất tại điểm đó.

2.4. Các tham số đặc trưng

2.4.1. Kỳ vọng

Định nghĩa 2.23. *Nếu X là đại lượng ngẫu nhiên rời rạc có bảng phân phối xác suất là*

X	x_1	x_2	x_3	...	x_n
$P(X = x_i)$	p_1	p_2	p_3	...	p_n

thì kỳ vọng toán của X được ký hiệu và xác định bởi công thức $E(X) = \sum_{i=1}^{n} x_i p_i$.

- Trường hợp X rời rạc có vô hạn đếm được phần tử thì $E(X) = \sum_{i=1}^{+\infty} x_i p_i$ (với điều kiện chuỗi này hội tụ tuyệt đối)

- Nếu X là đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất là $f(x)$ thì kỳ vọng toán của X là $E(X) = \int_{-\infty}^{+\infty} x.f(x)dx$.

Chú ý: Nếu $f(x)$ chỉ dương trong khoảng (a, b) thì

$$E(X) = \int_a^b x.f(x)dx.$$

Ví dụ 2.24. Tìm kỳ vọng toán của đại lượng ngẫu nhiên rời rạc X có bảng phân phối xác suất như sau

X	*1*	*3*	*4*
P(X)	*0,1*	*0,5*	*0,4*

Giải. Theo định nghĩa kỳ vọng toán của đại lượng ngẫu nhiên rời rạc ta có $E(X) = \sum x_i p_i = 1.0,1 + 3.0,5 + 4.0,4 = 3,2.$

Ví dụ 2.25. Tìm kỳ vọng toán của đại lượng nngẫu nhiên liên tục X có hàm mật độ xác suất:

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

Giải. Theo định nghĩa ta có

$$E(X) \overset{lt}{=} \int_a^b x.\frac{1}{b-a}dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{a+b}{2}$$

Ta thấy, $E(X)$ trùng với điểm giữa của $[a, b]$.

Chú ý: Kỳ vọng toán của đại lượng ngẫu nhiên là số xác định.

★ **Các tính chất của kỳ vọng toán**

Tính chất 1: $E(C) = C$(C là hằng số).

Thật vậy, ta coi C như một đại lượng ngẫu nhiên rời rạc, đặc biệt, với một giá trị có thể có là C và xác suất tương ứng bằng 1. Theo định nghĩa của kỳ vọng toán ta có $E(C) = C$.

Tính chất 2: $E(C.X) = C.E(X)$ (C - hằng số).

Thật vậy, giả sử X là đại lượng ngẫu nhiên rời rạc có bảng phân phối xác suất là

X	x_1	x_2	x_3	...	x_n
P	p_1	p_2	p_3	...	p_n

Do đó $C.X$ sẽ là một đại lượng ngẫu nhiên rời rạc mà các giá trị có thể có của nó là $Cx_1, ..., Cx_n$. Với xác suất tương ứng là:

$$p(cX = cx_1) = p(X = x_1) = p_1, ..., p(cX = cx_n) = p(X = x_n) = p_n.$$

Vậy bảng phân phối xác suất của đại lượng ngẫu nhiên cX có dạng

cX	cx_1	cx_2	cx_3	...	cx_n
P	p_1	p_2	p_3	...	p_n

Khi đó theo định nghĩa kỳ vọng ta có:

$$E(cX) = cx_1p_1 + cx_2p_2 + ... + cx_np_n = c(x_1p_1 + ...x_np_n) = c.E(X).$$

Tính chất 3: $E(X + Y) = E(X) + E(Y)$.

Thật vậy, giả sử các đại lượng ngẫu nhiên X, Y rời rạc có bảng phân phối xác suất như sau:

X	x_1	x_2	x_3	...	x_n
P	p_1	p_2	p_3	...	p_n

Y	y_1	y_2	y_3	...	y_m
q	q_1	q_2	q_3	...	q_m

lúc đó bảng phân phối xác suất của X +Y là

$X + Y$	$x_1 + y_1$	$x_2 + y_2$	$x_3 + y_3$	...	$x_n + y_m$
P	p_{11}	p_{12}	p_{13}	...	p_{nm}

Trong đó p_{ij} là xác suất để $X + Y$ nhận giá trị bằng

$$x_i + y_j; (i = 1, ..., n; j = 1, ..., m)$$

Theo định nghĩa kỳ vọng ta có

$$E(X+Y) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i+y_j)p_{ij}$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{m}x_ip_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m}y_jp_{ij}$$
$$= \sum_{i=1}^{n}x_i\sum_{j=1}^{m}p_{ij} + \sum_{j=1}^{m}y_j\sum_{i=1}^{n}p_{ij} = \sum_{i=1}^{n}x_ip_i + \sum_{j=1}^{m}y_jq_j = E(X)+E(Y)$$

(Trong đó: $\sum_{i=1}^{n}p_{ij} = q_j; \sum_{j=1}^{m}p_{ij} = p_i$.)

Hệ quả: $E(\sum_{i=1}^{n}x_i) = \sum_{i=1}^{n}E(x_i)$.

Tính chất 4: $E(X.Y) = E(X).E(Y)$, với X, Y là hai đại lượng ngẫu nhiên độc lập.

Thật vậy, giả sử X, Y rời rạc, độc lập có bảng phân phối xác suất sau

X	x_1	x_2	...	x_n	Y	y_1	y_2	...	y_m
p	p_1	p_2	...	p_n	q	q_1	q_2	...	q_m

lúc đó X, Y có quy luật phân phối như sau (vì X, Y độc lập)

XY	x_1y_1	x_2y_2	...	x_ny_m
P	p_1q_1	p_2q_2	...	p_nq_m

Ta có $E(X.Y) = \sum_{i=1}^{n}\sum_{j=1}^{m}x_iy_jp_iq_j = \sum_{i=1}^{n}x_ip_i.\sum_{j=1}^{m}y_jq_j = E(X).E(Y)$.

Hệ quả:

$E(\prod_{i=1}^{n}X_i) = \prod_{i=1}^{n}E(X_i)$, với $X_1, X_2, .., X_n$ độc lập với nhau.

★ Bản chất và ý nghĩa của kỳ vọng toán

Giả sử đối với đại lượng ngẫu nhiên X, tiến hành n phép thử, trong đó n_1 lần X nhận giá trị x_1, n_2 lần X nhận giá trị $x_2, ..., n_k$ lần X nhận giá trị $x_k(\sum_{i=1}^{k}n_i = n)$.

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

$$\bar{X} = \frac{n_1x_1 + n_2x_2 + ... + n_kx_k}{n} = x_1\frac{n_1}{n} + x_2\frac{n_2}{n} + ... + x_k\frac{n_k}{n}.$$

Trong đó $\frac{n_i}{n}$(i =1,..,k) chính là tần suất xuất hiện các giá trị xi trong n phép thử trên, do đó

$$\bar{X} = x_1 f_1 + x_2 f_2 + ... + x_k f_k.$$

Theo định nghĩa thống kê về xác suất khi $n \to \infty$, các tần suất sẽ hội tụ theo xác suất về các xác suất tương ứng, do đó với n đủ lớn ta có thể viết

$$\bar{X} \approx x_1 p_1 + x_2 p_2 + ... + x_k p_k = E(X).$$

Vậy kỳ vọng toán của đại lượng ngẫu nhiên gần bằng trung bình số học của các giá trị quan sát của đại lượng ngẫu nhiên. Nó phản ánh giá trị trung tâm của phân phối xác suất của đại lượng ngẫu nhiên.

Ví dụ 2.26. Tung con xúc sắc n lần. Tìm kỳ vọng toán của tổng số chấm thu được.

Giải. Gọi $X_i (i = 1, 2, ..., n)$ là số chấm thu được ở lần tung thứ i và gọi X là tổng số chấm thu được trong n lần tung. Vậy $X = \sum_{i=1}^{n} X_i$. Theo tính chất của kỳ vọng $E(X) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$. Mỗi đại lượng X_i đều có bảng phân phối xác suất

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

Do đó $E(X_i) = 1/6(1 + 2 + 3 + 4 + 5 + 6) = 7/2$.

$$E(X) = \sum_{i=1}^{n} E(X_i) = \frac{7}{2}.n.$$

2.4.2. Phương sai và độ lệch tiêu chuẩn

Trong thực tế nhiều khi chỉ xác định kỳ vọng toán của đại lượng ngẫu nhiên thì chưa đủ để xác định đại lượng ngẫu nhiên đó. Ta còn phải xác định mức độ phân tán của các giá trị của đại lượng ngẫu nhiên xung quanh các giá trị trung bình của nó nữa.

Định nghĩa 2.27. *Phương sai của đại lượng ngẫu nhiên X, ký hiệu là D(X) là kỳ vọng toán của bình phương độ lệnh của đại lượng ngẫu nhiên so với kỳ vọng của nó.*

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

Nếu X là đại lượng ngẫu nhiên rời rạc có bảng phân phối xác suất là

X	x_1	x_2	x_3	...	x_n
P	p_1	p_2	p_3	...	p_n

thì phương sai được tính bởi công thức $D(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 p_i$.
Nếu X là đại lượng ngẫu nhiên liên tục có hàm mật độ xác suất f(x) thì phương sai được thì phương sai được tính bởi công thức.

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

Trong thực tế việc tính phương sai bằng công thức định nghĩa trên có thể gặp khó khăn. Người ta thường tính phương sai bằng công thức sau

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

Thật vậy, theo định nghĩa:

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

Nếu X rời rạc thì $D(X) = \sum_{i=1}^{n} x_i^2 p_i - (\sum_{i=1}^{n} x_i p_i)^2$.
Nếu X liên tục thì $D(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - (\int_{-\infty}^{+\infty} x f(x)dx)^2$.

Ví dụ 2.28. Tìm phương sai của ĐLNN rời rạc X có bảng phân phối xác suất như sau

X	*1*	*3*	*4*
P(X)	*0,1*	*0,5*	*0,4*

Giải. Áp dụng công thức: $D(X) = E(X^2) - [E(X)]^2$.

Tính $E(X) = 1.0,1 + 3.0,5 + 4.0,4 = 3,2$,

$E(X^2) = 1^2.0,1 + 3^2.0,5 + 4^2.0,4 = 11$.

Vậy $D(X) = 11 - (3,2)^2 = 0,76$.

Ví dụ 2.29. Đại lượng ngẫu nhiên liên tục X có hàm mật độ xác suất như sau

$$f(x) = \begin{cases} 3x^2 & \textbf{\textit{khi}} \ \ x \in (0,1), \\ 0 & \textbf{\textit{khi}} \ x \notin (0,1). \end{cases}$$

Tìm phương sai của đại lượng ngẫu nhiên X?

Giải. Áp dụng công thức $D(X) = E(X^2) - [E(X)]^2$

Trong đó

$$E(X) \overset{lt}{=} \int\limits_{-\infty}^{+\infty} x f(x) dx = \int\limits_{0}^{1} x.3x^2 dx = \frac{3x^4}{4} \Big|_0^1 = \frac{3}{4},$$

$$E(X^2) \overset{lt}{=} \int\limits_{-\infty}^{+\infty} x^2 f(x) dx = \int\limits_{0}^{1} x^2.3x^2 dx = \frac{3x^5}{5} \Big|_0^1 = \frac{3}{5}.$$

Vậy $D(X) = \dfrac{3}{5} - \left(\dfrac{3}{4}\right)^2 = 0.0375$.

Chú ý: Phương sai của đại lượng ngẫu nhiên là một giá trị xác định không âm.

★ Các tính chất của phương sai

Tính chất 1: $D(C) = 0$ (C = const).

Thật vậy, theo định nghĩa của phương sai

$$D(C) = E[C - E(C)]^2 = E(C - C) = 0.$$

Tính chất 2: $D(C.X) = C2D(X)$ (C = const).

Thật vậy, $D(C.X) = E[CX - E(CX)]^2 = E[CX - CE(X)]^2$

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

Tính chất 3: $D(X+Y) = D(X) + D(Y)$ với X, Y là hai đại lượng ngẫu nhiên độc lập.

Thật vậy :

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

Hệ quả 1: $D(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} D(X_i)$, với $X_1, X_2, ..., X_n$ là các đại lượng ngẫu nhiên độc lập.

Hệ quả 2: $D(C+X) = D(X)$ (C= const).

Hệ quả 3: $D(X-Y) = D(X) + D(Y)$.

★ Ý nghĩa của phương sai

Xuất phát từ định nghĩa của phương sai, ta thấy phương sai chính là trung bình số học của bình phương các độ lệnh giữa các giá trị có thể có của đại lượng ngẫu nhiên so với giá trị trung bình của của các giá trị đó. Do đó nó phản ánh mức độ phân tán của các giá trị của đại lượng ngẫu nhiên xung quanh giá trị trung tâm của nó là kỳ vọng toán.

Ví dụ 2.30. Tung một con xúc sắc n lần. Tìm phương sai của tổng số chấm thu được

Giải. Gọi $X_i (i = 1, 2, ..., n)$ là “số chấm thu được ở con xúc sắc thứ i” Gọi X là “tổng số chấm thu được” ở cả n con xúc xắc. Khi đó

$$X = \sum_{i=1}^{n} X_i \Rightarrow D(X) \overset{dl}{=} \sum_{i=1}^{n} D(X_i)$$

với $D(X_i) = E(X_i^2) - (E(X_i))^2$. Mỗi đại lượng ngẫu nhiên X_i có bảng phân phối xác suất như sau

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

Do đó $E(X_i) = 1/6(1+2+3+4+5+6) = 7/2.$
$E(X_i^2) = \dfrac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) = \dfrac{91}{6}.$
Suy ra $D(X_i) = \dfrac{91}{6} - \left(\dfrac{7}{2}\right)^2 = \dfrac{35}{12}$, (i =1,...,n).
Vậy $D(X) = \sum\limits_{i=1}^{n} D(X_i) = \dfrac{35}{12}.n.$

★ **Độ lệch tiêu chuẩn** $\sigma(x)$

Ngoài phương sai, người ta còn sử dụng một vài tham số để đặc trưng cho mức độ phân tán của đại lượng ngẫu nhiên, trong đó có độ lệch tiêu chuẩn.
Độ lệch tiêu chuẩn của đại lượng ngẫu nhiên X: $\sigma_{(X)} = \sqrt{D(X)}.$
Ta thấy đơn vị đo của phương sai bằng bình phương đơn vị đo của đại lượng ngẫu nhiên. Vì vậy khi phải đánh giá mức độ phân tán của đại lượng ngẫu nhiên theo đơn vị đo của nó người ta thường tính độ lệch tiêu chuẩn, vì nó có cùng đơn vị đo với đại lượng ngẫu nhiên cần nghiên cứu.
Độ lệch tiêu chuẩn đặc trưng cho mức độ phân tán trung bình của các giá trị của đại lượng ngẫu nhiên xung quanh kỳ vọng của nó.

Ví dụ 2.31. Tiến hành n phép thử độc lập, xác suất xuất hiện biến cố A trong mỗi phép thử đều bằng p và $P\left(\overline{A}\right) = 1 - p = q$. Gọi X là số lần xuất hiện biến cố A trong n phép thử trên. Hãy tính $E(X); D(X); \sigma(x).$

Giải. Gọi X_i là số lần xuất hiện A trong phép thử thứ $i(i = 1, ..., n)$. Ta có:
$E\left(X_i\right) \ = \ 0.q \ + \ 1.p \ = \ p,$
$D\left(X_i\right) \ = \ \left(0^2.q \ + 1^2.p\right) \ \ p^2 = \ p\left(1-p\right) \ = \ p.q.$
Suy ra $E\left(X\right) \ = \ E\left(X_1 + ... + X_n\right) \ = \ E\left(X_1\right) \ + ... + E\left(X_n\right) \ = \ np,$
$D\left(X\right) \ = \ D\left(X_1 + ... + X_n\right) \ = \ D\left(X_1\right) \ + ... + D\left(X_n\right) \ = npq,$
$\sigma_{(X)} = \sqrt{D(X)} = \sqrt{npq}.$

2.4.3. Mode

Định nghĩa 2.32. *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 2.33. Giả sử X là điểm trung bình của sinh viên trong trường thì Mod(X) là điểm mà nhiều sinh viên đạt được nhất.

2.5. Một số quy luật phân phối xác suất thường gặp

2.5.1. Phân phối không – một

Định nghĩa 2.34. *Đại lượng ngẫu nhiên rời rạc X phân phối theo quy luật không – một với tham số p nếu nó có bảng phân phối xác suất (với p + q = 1).*

X,	*0*	*1*
P	*p*	*q*

★ Các tham số đặc trưng của quy luật không – một

Theo bảng phân phối xác suất của X, dễ dàng tính được

$E(X) = p, D(X) = pq,$

Suy ra $\sigma_{(X)} = \sqrt{pq}$.

Ví dụ 2.35. Một máy tiến hành sản xuất thử 1 sản phẩm. Xác suất để sản phẩm đó đạt tiêu chuẩn là 0,7. Gọi X là số sản phẩm không đạt tiêu chuẩn. X có phân phối gì? Trung bình có mấy sản phẩm không đạt tiêu chuẩn. Tính phương sai của X.

Giải. X nhận giá trị 0, 1. Với xác suất tương ứng

$P(X = 0) = 1 - 0,7 = 0,3, P(X = 1) = 0,7.$

Vậy X có quy luật không - một với tham số $p = 0,3$.

Khi đó $E(X) = p = 0,3,$ **(E(X) là trung bình)**
$D(X) = p.q = 0,3.0,7 = 0,21.$

2.5.2. Phân phối nhị thức

Định nghĩa 2.36. *Đại lượng ngẫu nhiên rời rạc X được gọi là có phân phối nhị thức với tham số* n, p *và kí hiệu* $X \sim B(n,p)$*, ở đó* $n \in N$ *và* $0 < p < 1$*, nếu* $X(\Omega) = 0, 1, ..., n$ *và* $p_k = P(X = k) = C_n^k p^k q^{n-k}$ *với* $q = 1 - p.$

Như vậy, bảng phân phối xác suất của đại lượng ngẫu nhiên phân phối theo quy luật nhị thức có dạng:

X	0	1	2	...	k	...	n
P	$C_n^0 p^0 q^n$	$C_n^1 p^1 q^{n-1}$	$C_n^2 p^2 q^{n-2}$	...	$C_n^k p^k q^{n-k}$	...	$C_n^n p^n q^0$

★ **Các tham số đặc trưng của quy luật nhị thức**
Nếu $X \sim B(n,p)$ thì ta có:

$$E(X) = np, D(X) = npq,$$

$$\sigma_{(X)} = \sqrt{npq}$$

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

Ví dụ 2.37. Xác suất để một người bắn trúng bia là 0,8. Gọi X là số viên đạn trúng bia khi người đó bắn 6 phát đạn. Hỏi X có phân phối gì? Trung bình có mấy viên đạn trúng bia? Số viên đạn có khả năng trúng bia nhiều nhất là bao nhiêu?

Giải. Bài toán thoả mãn lược đồ Bernoulli, do đó X có phân phối nhị thức với tham số n = 6, p = 0,8.
Số viên đạn trúng bia trung bình chính là kỳ vọng của X.
Suy ra $E(X) = np = 6.0,8 = 4,8.$ Số viên đạn có khả năng trúng bia nhiều nhất chính là giá trị mode. Ta có

$np - q \leq mod(X) \leq np + p, n = 6, p = 0.8, q = 1 - 0.8 = 0.2 \Rightarrow 6.0,8 - 0,2 \leq mod(X) \leq 6.0,8 + 0,8.$

Do đó $4,6 \leq mod(X) \leq 5,6.$

Vậy số viên đạn có khả năng trúng bia nhiều nhất là 5 viên.

Ví dụ 2.38. 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:

- Có 3 phế phẩm
- 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 trong mỗi phép thử.

Ta có $p = p(A) = 0,03$. Gọi X là tổng số phế phẩm trong 100 sản phẩm thì $X \sim B(100; 0,03)$.

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

$P(0 \leq X \leq 3) = P_0 + P_1 + P_2 + P_3 = \sum\limits_{i=0}^{3} C_{100}^{i}(0,03)^{i}(0,97)^{100-i} = 0,647.$

Chú ý: Khi n khá lớn việc sử dụng các công thức trên gặp nhiều khó khăn, xác suất p không quá gần 0 và 1. Khi đó ta có thể áp dụng công thức xấp xỉ sau:

i) $P_x = C_n^x p^x q^{n-x} \approx \dfrac{1}{\sqrt{npq}}\varphi(u).$ (2.3)

trong đó

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

Công thức (2.3) gọi là công thức địa phương Laplace.

ii) $P(x \leq X \leq x + h) \approx \Phi(u_2) - \Phi(u_1).$

Trong đó:

$$u_1 = \frac{x - np}{\sqrt{npq}} \;;\; u_2 = \frac{x + h - np}{\sqrt{npq}} ;\; \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_0^u e^{-\frac{t^2}{2}} dt$$

(Hàm Laplace)

2.5.3. Phân phối Poisson - $P(\lambda)$

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

$$\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}; \quad \frac{n\,(n-1)\,...\,(n-k+1)}{n^k} \approx 1; \quad \left(1 - \frac{\lambda}{n}\right)^k \approx 1.$$

Do đó : $P\,(X = k) \approx e^{-\lambda}\dfrac{\lambda^k}{k!}$.
Vậy từ công thức Bernoulli ta có công thức xấp xỉ

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

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

$$P_k = P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}.$$

Định nghĩa 2.39. *Đại lượng ngẫu nhiên rời rạc X có phân phối Poisson với tham số λ và kí hiệu $X \sim Poisson(\lambda)$ nếu $X\,(\Omega) = \{0\,,\ 1\,,2,...\}$ và $P(X = k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$ với λ là một số dương cho trước.*

Chú ý: Phân phối Poát – xông xuất hiện trong lược đồ Becnuli khi số phép thử rất lớn và xác suất p rất nhỏ, với $np = \lambda$ không đổi.

Ví dụ 2.40. 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 bị đứt hay không 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 $X \sim B(1000; 0,002)$.

Vì $n = 1000$ khá lớn và $np = 2$ không đổi nên ta có thể xem $X \in P(2)$ Do đó xác suất để có không quá 2 ống sợi bị đứt trong một giờ là

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

$$P_0 = P(X = 0) = e^{-2}\frac{2^0}{0!}; P_1 = P(X = 1) = e^{-2}\frac{2^1}{1!}; P_2 = P(X = 2) = e^{-2}\frac{2^2}{2!}$$

Do đó $P\left(0 \leq X \leq 2\right) = 5.e^{-2} \approx 0,6808.$

★ **Các tham số đặc trưng**

Nếu $X \in P(\lambda)$ thì $E(X) = D(X) = \lambda$ và $\lambda - 1 \leq modX \leq \lambda.$

★ **Ứng dụng** Một vài đại lượng ngẫu nhiên có phân phối Poisson:

- Số lỗi in sai trong một trang (hoặc một số trang) của một cuốn sách.
- Số người trong một cộng đồng sống cho tới 100 tuổi.
- Số cuộc điện thoại gọi sai trong một ngày.
- Số khách hàng vào bưu điện trong một ngày.

Ví dụ 2.41. Xác suất trong khi vận chuyển mỗi chai rượu bị vỡ là 0,001. Người ta tiến hành vận chuyển 2000 chai đến cửa hàng. Tìm số chai vỡ có khả năng nhiều nhất khi vận chuyển. Trung bình có mấy chai vỡ khi vận chuyển.

Giải. - Số chai vỡ có khả năng xảy ra nhiều nhất là giá trị môt x_0. Bài toán thoả mãn lược đồ Becnuli; vì n = 2000 rất lớn và $p = 0,001$ khá bé; tích $np = 2000.0,001 = 2$ không đổi, do đó nếu gọi X là số chai rượu bị vỡ khi vận chuyển thì X là đại lượng ngẫu nhiên phân phối theo quy luật Poat – xông. Vì $np = 2 \in Z$. Giá trị môt $x_0 = \lambda = 2$ và $x_0 = \lambda - 1 = 2\check{}1 = 1.$

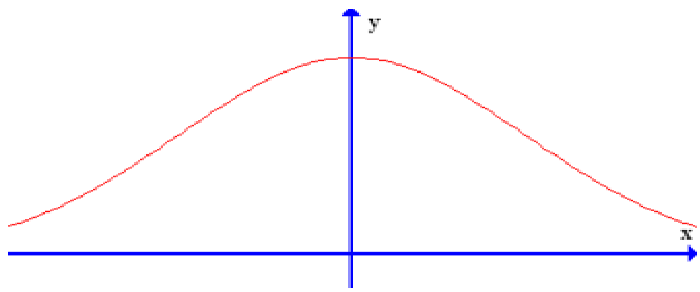
Vậy số chai vỡ có khả năng xảy ra nhiều nhất là 1 và 2.

- Số chai vỡ trung bình chính là kỳ vọng toán của X. Ta có $E(X) = \lambda = np = 2$ chai.

2.5.4. Phân phối chuẩn

Định nghĩa 2.42. ***(Phân phối chuẩn tắc)*** *ĐLNN Z được gọi là một ĐLNN* *có phân phối chuẩn tắc nếu hàm mật độ của nó có dạng*

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



Hình 2.2 Đồ thị của hàm mật độ phân phối chuẩn tắc

Đó là một đường cong đối xứng qua trục tung, có điểm cực đại tại $x = 0$. Các điểm uốn của đồ thị là $x = \pm 1$. Hàm phân phối của Z kí hiệu bởi $\Phi(x)$, là $\Phi(x) = \int\limits_{-\infty}^{x} f(t)\,dt = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2}\,dt.$
Tuy nhiên hàm $\Phi(x)$ không biểu diễn được thông qua các hàm sơ cấp đã biết. Người ta lập bảng tính sẵn các giá trị của $\Phi(x)$.
Các đặc trưng của ĐLNN Z là EZ $= 0$, DZ $= 1$. Khi đó $Z \sim N(0,1)$. Nếu $Z \sim N(0,1)$ thì

$$P(Z < a) = \Phi(a)$$

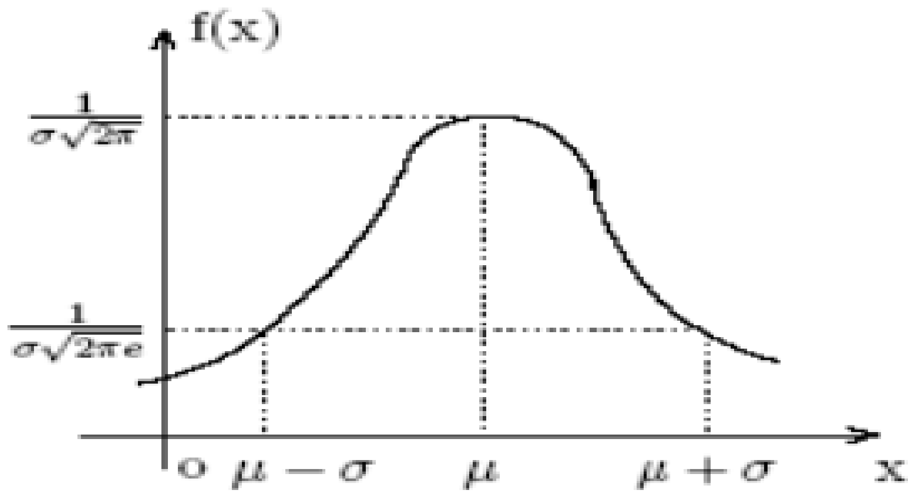
$$P(Z > a) = 1 - P(Z < a) = 1 - \Phi(a)$$

$$P(a < Z < b) = \Phi(b) - \Phi(a)$$

Định nghĩa 2.43. ***(Phân phối chuẩn)*** *ĐLNN X được gọi là một ĐLNN có phân phối chuẩn với tham số μ và σ^2 $(\sigma > 0)$ nếu ĐLNN $Z = \frac{X - \mu}{\sigma}$ phân phối chuẩn tắc. Hàm mật độ của X có dạng $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.*

Hàm phân phối của X kí hiệu bởi $\Phi_X(x)$ là

$$\Phi_X(x) = \int\limits_{-\infty}^{x} f_X(t)\,dt = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt.$$



Hình 2.3 Đồ thị của hàm mật độ phân phối chuẩn

Ta có liên hệ $\Phi_X(x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)$. Các đặc trưng của X là EX $= \mu, DX = \sigma^2$. Khi đó kí hiệu $X \sim N\left(\mu, \sigma^2\right)$. Khi đó

$$P(X < a) = P(Z < \frac{a-\mu}{\sigma}) = \Phi(\frac{a-\mu}{\sigma})$$
$$P(X > a) = 1 - P(X < a) = 1 - \Phi(\frac{a-\mu}{\sigma})$$
$$P(a < X < b) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

Ví dụ 2.44. Kích thước của các chi tiết do một máy sản xuất ra là đại lượng ngẫu nhiên phân phối chuẩn với kích thước trung bình $\mu = 5$ cm và độ lệnh tiêu chuẩn $\sigma = 0,9cm$. Tìm xác suất để lấy ngẫu nhiên 1 chi tiết có kích thước nằm trong khoảng từ 4 cm đến 7 cm.

Giải. Gọi X là kích thước chi tiết được lấy ra. Theo giả thiết X phân phối chuẩn $N(\mu, \sigma^2) = N(5; (0,9)^2)$. Do đó

$$P(4 \leq X \leq 7) = \Phi\left(\frac{7-5}{0,9}\right) - \Phi\left(\frac{4-5}{0,9}\right) = \Phi\left(2,22\right) - \Phi\left(-1,11\right).$$

Tra bảng ta có $\Phi(2,22) = 0,4868, \Phi(1,11) = 0,3665.$

Vậy $p(4 \leq X \leq 7) = 0,4868 + 0,3665 = 0,8533.$

Ví dụ 2.45. Trọng lượng của một loại sản phẩm là đại lượng ngẫu nhiên có phân phối chuẩn với trọng lượng trung bình $\mu = 5$kg và độ lệch tiêu chuẩn $\sigma = 0,1$. Tính tỷ lệ những sản phẩm có trọng lượng từ 4,9kg đến 5,2kg.

Giải. Gọi X là trọng lượng của sản phẩm thì $X \sim N(5; 0,01)$.
Tỷ lệ sản phẩm có trọng lượng từ 4,9kg đến 5,2kg là
$P(4,9 \leq X \leq 5,2) = \phi(\frac{5,2-5}{0,1}) - \phi(\frac{4,9-5}{0,1})$
$= \phi(2) - \phi(-1)$ = 0,9773 - 0,1587 = 0,8186.

2.5.5. Phân phối Student

Định nghĩa 2.46. *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^2với 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. Ký hiệu $T \in T(n)$.*

Nhận xét: Hàm mật độ của đại lượng ngẫu nhiên có phân phối Student với n bậc tự do có dạng

$$f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}},$$

trong đó $\Gamma(x) = \int\limits_0^{+\infty} t^{x-1}e^{-t}dt.$

★ Phân vị Student

Phân vị Student mức α, ký 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 cố 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 đẻ 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.

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

Hình 2.4 Bảng tông kết các hàm phân phối liên tục

2.6. Một số kiến thức về biến ngẫu nhiên nhiều chiều

2.6.1. Định nghĩa và phân loại

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. Ký 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).

Ví dụ 2.47. Một máy sản xuất một loại sản phẩm. Nếu kích thước của sản phẩm được đo bằng chiều dài X và chiều rộng Y thì ta có đại lượng ngẫu nhiên hai chiều, nếu tính thêm cả chiều cao Z nữa thì ta có đại lượng ngẫu nhiên 3 chiều.

2.6.2. Bảng phân phối xác suất

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

Trong đó:

$x_i (i = 1, 2, \ldots, n)$ là các giá trị có thể của thành phần X;

$y_j (j = 1, 2, \ldots, m)$ là các giá trị có thể của thành phần Y;

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

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

2.6.3. Hàm phân phối xác suất

Hàm phân phối xác suất của đại lượng ngẫu nhiên hai chiều (X, Y), ký hiệu $F(x, y)$ là hàm được xác định như sau:

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

Nhận xét:

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

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

2.6.4. Hàm mật độ phân phối xác suất

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 tập số thực.

2.6.5. Các đặc trưng số của đại lượng ngẫu nhiên 2 chiều

★ Kỳ vọng và phương sai của các thành phần

• Trường hợp (X, Y) rời rạc

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

$D(X) = \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} x_i^2 P(x_i, y_j) - [E(X)]^2,$

$D(Y) = \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} y_j^2 P(x_i, y_j) - [E(Y)]^2.$

• Trường hợp (X, Y) liên tục

$E(X) = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} xf(x,y)dydx,\ E(Y) = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} yf(x,y)dydx,$

$D(X) = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} x^2 f(x,y)dydx - [E(X)]^2,$

$D(Y) = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} y^2 f(x,y)dydx - [E(Y)]^2,$

★ Kỳ vọng của các đại lượng ngẫu nhiên hai chiều

- Kỳ vọng của đại lượng ngẫu nhiên hai chiều là một vectơ gồm 2 thành phần đượng xác định như sau:

$$E(X,Y) = (E(X), E(Y)),$$

trong đó $E(X), E(Y)$ là kỳ vọng của các thành phần X, Y tương ứng.

★ Hiệp phương sai (Covariance) của đại lượng ngẫu nhiên hai chiều

- Cho đại lượng ngẫu nhiên hai chiều (X, Y), hiệp phương sai của đại lượng ngẫu nhiên được ký hiệu là $Cov(X,Y)$ và được xác định bởi công thức:

$$Cov(X,Y) = E[X - E(X)][Y - E(Y)] = E(X.Y) - E(X).E(Y).$$

Trong đó:

$$E(X,Y) = \begin{cases} \sum_{i=1}^{n} x_i.y_j.P(x_i, y_j) & \text{nếu (X, Y) rời rạc,} \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x.y.f(x,y)dxdy & \text{nếu (X, Y) liên tục.} \end{cases}$$

và $E(X), E(Y)$ là kỳ vọng của các thành phần.

★ Hệ số tương quan (Corelation) của đại lượng ngẫu nhiên hai chiều

- Hệ số tương quan của đại lượng ngẫu nhiên hai chiều (X,Y) được xác định bởi:

$$\rho = \rho(X,Y) = \frac{E(X.Y) - E(X).E(Y)}{\sigma(X).\sigma(Y)} = \frac{Cov(X,Y)}{\sigma(X).\sigma(Y)}.$$

- Hệ số tương quan cho biết mối liên hệ giữa hai thành phần X và Y

Tính chất:

i, $|\rho(| \leq 1$.

ii, $\rho > 0$ ta nói X và Y có mối tương quan dương (tương quan đồng biến).

iii, $\rho < 0$ ta nói X và Y có mối tương quan âm (tương quan nghịch biến).

iv, $\rho = 0$ ta nói X và Y không tương quan với nhau.

v, Mối tương quan giữa X và Y tỷ lệ thuận với giá trị $|\rho(|$.Đặc biệt: Khi $|\rho(| = 1$, ta nói X và Y có mối tương quan tuyến tính. (Có thể biểu diễn mối quan hệ giữa X và Y bằng phương trình đường thẳng Y = AX + B hoặc X = A'Y + B').

Ví dụ 2.48. Cho đại lượng ngẫu nhiên 2 chiều rời rạc (X,Y) có bảng phân phối xác suất đồng thời:

$X \setminus Y$	*1*	*2*	*3*
1	*0,1*	*0,2*	*0,15*
2	*0,15*	*0,1*	*0,3*

a. Xác định phân phối xác suất cho từng thành phần.

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

Giải. a. - Phân phối xác suất cho thành phần X:

Các giá trị mà X nhận là 1, 2. Với các xác suất tương ứng bằng:

$P(X = 1) = 0,1 + 0,2 + 0,15 = 0,45$

$P(X = 2) = 0,15 + 0,1 + 0,3 = 0,55$

Bảng phân phối xác suất của X là:

x_i	1	2
p_i	0.45	0.55

- Phân phối xác suất cho thành phần Y:

Các giá trị mà Y nhận là $1, 2, 3$. Với các xác suất tương ứng bằng:

$P(Y = 1) = 0,1 + 0,15 = 0,25,$

$P(Y = 2) = 0,2 + 0,1 = 0,3,$

$P(Y = 3) = 0,15 + 0,3 = 0,45.$

Bảng phân phối xác suất của Y là:

y_j	1	2	3
q_j	0,25	0,3	0,45

b. Kỳ vọng:

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

Với $E(X) = 1.0,45 + 2.0,55 = 1.55;$

$E(Y) = 1.0,25 + 2.0,3 + 3.0,45 = 2,2.$

Suy ra $E(X, Y) = (1.55, 2.2).$

- Hiệp phương sai

$$Cov(X, Y) = E(X.Y) - E(X).E(Y),$$

với $E(X.Y) = 1.1.0,1 + 1.2.0,2 + 1.3.0,15 + 2.1.0,15 + 2.2.0,1 + 2.3.0,3 = 3.45.$

Vậy: $Cov(X, Y) = 3,45 - 1,55.2,2 = 0,04.$

- Hệ số tương quan:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma(X).\sigma(Y)},$$

trong đó:

$$D(X) = E(X^2) - [E(X)]^2 = 2,65 - (1,55)^2 = 0,2475 \Rightarrow \sigma(X) = 0,4975.$$

$$D(Y) = E(Y^2) - [E(Y)]^2 = 5,5 - (2,2)^2 = 0,66 \Rightarrow \sigma(Y) = 0,8124.$$

Vậy: $\rho(X, Y) = \dfrac{0,04}{0,4975.0,8124} = 0,09897.$

Thấy rằng mối tương quan giữa X và Y là rất yếu (Gần như không tương quan với nhau)

2.7. Luật số lớn

2.7.1. Bất đẳng thức Markov

Định lí 2.49. *: 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: Trong trường hợp X là đại lượng ngẫu nhiên liên tục có hàm mật độ f(x)

$E(X) = \int\limits_0^{+\infty} xf(x)dx = \int\limits_0^{a} xf(x)dx + \int\limits_a^{+\infty} xf(x)dx$

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

2.7.2. Bất đẳng thức Tchebyshev

Định lí 2.50. *Nếu X là đại lượng ngẫu nhiên có kỳ vọng μ và phương sai σ^2 hữu hạn thì $\forall \varepsilon \geq 0$ bé tùy ý ta có:* $P(|X-\mu| \geq \varepsilon) \leq \dfrac{D(X)}{\varepsilon^2}$ *hay* $P(|X-\mu| < \varepsilon) > 1 - \dfrac{D(X)}{\varepsilon^2}$.

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

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

Vì $(X-\mu)^2 \geq \varepsilon^2$ khi và chỉ khi $|X-\mu| \geq \varepsilon$ nên $P(|X-\mu| \geq \varepsilon) \geq \dfrac{D(X)}{\varepsilon^2}$.

Chú ý: Bất đẳng thức Markov và Tchebysev 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 2.51. Gỉa 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 nói gì về xác suất sản phẩm tuần này sẽ ở giữa 40 và 60.

Giải. a) Theo bất đẳng thức Markov $P(X > 75) \geq \dfrac{E(X)}{75} = \dfrac{50}{75} = \dfrac{2}{3}$.

b) Theo bất đẳng thức Tchebysev $P(|X-50| \geq 10) \leq \dfrac{\sigma^2}{10^2} = \dfrac{25}{100} = \dfrac{1}{4}$.

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

2.7.3. Định lý Tchebyshev

Định lí 2.52. *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* $\forall\varepsilon > 0$ *bé tùy ý ta có:* $\lim\limits_{n\to\infty} P(\left|\frac{1}{n}\sum\limits_{i=1}^{n} X_i - \frac{1}{n}\sum\limits_{i=1}^{n} E(X_i)\right| < \varepsilon) = 1.$
Đặc biệt, khi $E(X_i) = a; (i = 1, 2, \ldots, n)$ *thì*

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

Chứng minh:
Ta chứng minh trong trường hợp đặc biệt $E(Xi) = \mu, D(Xi) = \sigma^2$ $(i = \overline{1, n})$. Ta có $E(\frac{1}{n}\sum\limits_{i=1}^{n} X_i) = \mu$, $D(\frac{1}{n}\sum\limits_{i=1}^{n} X_i) = \frac{\sigma^2}{n}$. Theo bất đẳng thức Tchebysev $P(\left|\frac{1}{n}\sum\limits_{i=1}^{n} X_i - \mu\right|) \leq \frac{\sigma^2}{n\varepsilon^2}$.
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 lương 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.

2.7.4. Định lý Bernoulli

Định lí 2.53. *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 với mọi* $\varepsilon > 0$ *bé tùy ý ta có* $\lim\limits_{n\to\infty} P\left(|f_n - p| < \varepsilon\right) = 1.$

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.

2.8. Bài tập chương 2

Bài 2.1. Gieo đồng thời 2 con xúc xắc cân đối đồng chất. Gọi X là tổng số chấm ở mặt trên của 2 con xúc xắc.

a. Lập bảng phân phối xác suất của X.

b. Lập hàm phân phối xác suất của X

Bài 2.2. Một thiết bị gồm 3 bộ phận hoạt động độc lập với nhau, xác suất trong khoảng thời gian t các bộ phận hỏng tương ứng bằng 0,2 ; 0,3 ; 0,25. Gọi X là số bộ phận bị hỏng trong khoảng t.

a. Lập bảng phân phối xác suất của X.

b. Viết biểu thức hàm phân phối xác suất của X

c. Tính $P(0 < X \leq 4)$ bằng cách tính trực tiếp và bằng cách thông qua hàm phân phối.

Bài 2.3. Một cửa hàng có 10 bóng đèn trong đó có 8 bóng tốt và 2 bóng xấu. Nam mua ngẫu nhiên 2 bóng. Gọi X là số bóng tốt mà Nam mua được.

a. Lập bảng phân phối xác suất của X.

b. Tính kỳ vọng E(X) và phương sai D(X).

Bài 2.4. Cho ĐLNN rời rạc X có bảng phân phối xác suất

X	0	1	2	3	4	5	6	7
p	0	k	k	k	k	$2k^2$	k^2	$2k^2$

a. Xác định k và tính EX.

b. Tính xác suất $P(X \geq 5)$ và $P(X < 3)$.

Bài 2.5. Trong một hộp kín có 3 quả táo và 27 quả lê. Lấy ngẫu nhiên liên tiếp từ trong hộp ra từng quả cho đến khi lấy được quả lê thì dừng. Gọi X là số quả táo đã lấy ra. Lập bảng phân phối xác suất của X trong trường hợp:

a. Lấy không hoàn lại

b. Lấy có hoàn lại

Bài 2.6. Một hộp kín đựng 5 quả cầu xanh, 4 quả cầu đỏ, 3 quả cầu vàng (Các quả cầu đôi một khác nhau và mỗi quả chỉ mang một màu). Lấy ngẫu nhiên đồng thời 4 quả trong hộp. Gọi X là số quả cầu vàng lấy được.

a. Lập bảng phân phối xác suất của X.

b. Tính kỳ vọng E(X) và phương sai D(X).

Bài 2.7. Một nhóm học sinh gồm 12 em trong đó có 5 em lớp 10A, 4 học sinh lớp 10B và 3 học sinh lớp 10C. Chọn ngẫu nhiên đồng thời 4 em trong nhóm. Gọi X là số em lớp 10C được chọn.

a. Lập bảng phân phối xác suất của X.

b. Tính kỳ vọng E(X) và phương sai D(X).

Bài 2.8. Một lô hàng có 8 chính phẩm và 4 phế phẩm. Lấy ngẫu nhiên liên tiếp không hoàn lại 3 sản phẩm. Gọi X là số phế phẩm lấy được.

a. Lập bảng phân phối xác suất của X.

b. Tính kỳ vọng E(X) và phương sai D(X).

Bài 2.9. Một lô hàng có 6 chính phẩm và 4 phế phẩm. Lấy ngẫu nhiên liên tiếp có hoàn lại 3 sản phẩm. Gọi X là số phế phẩm lấy được.

a. Lập bảng phân phối xác suất của X.

b. Tính kỳ vọng E(X) và phương sai D(X).

Bài 2.10. Một xạ thủ đem 5 viên đạn để bắn kiểm tra trước ngày thi bắn. Xạ thủ bắn từng viên vào bia với xác suất trúng vòng 10 là 0,8. Nếu bắn được 3 viên liên tiếp trúng vòng 10 thì thôi không bắn nữa. Gọi X là số viên đạn đã bắn.

a. Lập bảng phân phối xác suất của X.

b. Lập hàm phân phối xác suất của X. Tính E(X) và D(X).

Bài 2.11. Ba vận động viên ném bóng rổ, mỗi người ném một quả với xác suất ném trúng rổ của từng người là 0.7, 0.9, 0.8 tương ứng. Gọi X là số quả ném trúng rổ.

a. Lập bảng phân phối xác suất của X.

b. Tính kỳ vọng E(X) và phương sai D(X).

Bài 2.12. . Một phân xưởng có 3 máy sản xuất. Xác suất để trong thời gian T các máy hoạt động tốt là 0.9, 0.7, 0.6. Gọi X là số máy không hoạt động tốt trong khoảng thời gian T.

a. Lập bảng phân phối xác suất của X.

b. Tính kỳ vọng E(X) và phương sai D(X).

Bài 2.13. Trong một lô hàng linh kiện máy tính cùng loại với tỷ lệ không đạt tiêu chuẩn là 10%. Lấy ngẫu nhiên có hoàn lại từ trong lô ra 5 linh kiện. Gọi X là số linh kiện đạt tiêu chuẩn được lấy ra.

a. X tuân theo luật phân phối gì ? Lập bảng phân phối xác suất cho X.

b. Tính kỳ vọng, phương sai và Mod cho X.

Bài 2.14. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

$$f(x) = \begin{cases} Kcosx & \textbf{khi } x \in [\frac{-\pi}{4}, \frac{\pi}{4}], \\ 0 & \textbf{khi } x \notin [\frac{-\pi}{4}, \frac{\pi}{4}]. \end{cases}$$

a. Xác định hệ số K .

b. Tính kỳ vọng E(X), phương sai D(X).

c. Tính xác suất $P(-\infty \leq X \leq 0)$.

Bài 2.15. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

$$f(x) = \begin{cases} Ccosx & \textbf{khi } x \in [0, \frac{\pi}{2}], \\ 0 & \textbf{khi } x \notin [0, \frac{\pi}{2}]. \end{cases}$$

a. Xác định hệ số C và xây dựng hàm phân phối F(X).

b. Tính kỳ vọng E(X), phương sai D(X).

Bài 2.16. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

$$f(x) = \begin{cases} \frac{C}{\sqrt{4-x^2}} & \textbf{khi } x \in (-2, 2), \\ 0 & \textbf{khi } x \notin (-2, 2). \end{cases}$$

a. Xác định hệ số C .

b. Tính $P\{-1 < X < 1\}$

c. Tìm hàm phân phối F(x).

Bài 2.17. Cho biến ngẫu nhiên X có hàm mật độ xác định bởi

$$f(x) = \begin{cases} k(x - x^2) & \textbf{khi } (0, 1), \\ 0 & \textbf{khi } x \notin (0, 1). \end{cases}$$

a. Tìm hằng số k và xác định hàm phân phối xác suất $F(x)$.

b. Tính $P(X > 0,5)$.

Bài 2.18. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

$$f(x) = \begin{cases} 1 + x & \textbf{khi } -1 \le x \le 0, \\ 1 - x & \textbf{khi } 0 \le x \le 1, \\ 0 & \textbf{khi } x < -1, x > 1. \end{cases}$$

a. Xây dựng hàm F(x)

b. Tính kỳ vọng, phương sai cho X

c. Tính xác suất $P(-\frac{1}{2} \le X \le \frac{1}{2})$.

Bài 2.19. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

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

a. Xác định hệ số k.

b. Tính kỳ vọng và phương sai cho X.

Bài 2.20. Cho biến ngẫu nhiên liên tục X với hàm mật độ như sau

$$f(x) = \begin{cases} kx^2(4 - x) & \textbf{khi } x \in [0, 4], \\ 0 & \textbf{khi } x \notin [0, 4]. \end{cases}$$

a. Tìm hằng số k và xác định hàm phân phối xác suất $F(x)$.

b. Tính kỳ vọng E(X) và phương sai D(X).

Bài 2.21. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

$$f(x) = \begin{cases} Asinx & \textbf{khi } x \in [0, \pi], \\ 0 & \textbf{khi } x \notin [0, \pi]. \end{cases}$$

a. Xác định hệ số A và xây dựng hàm phân phối F(X)

b. Tính kỳ vọng E(X), phương sai D(X).

Bài 2.22. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

$$f(x) = \frac{C}{1+x^2} \quad \forall x \in R$$

a. Xác định hệ số C .

b. Xây dựng hàm phân phối F(x)

c. Tính xác suất để trong 4 lần thực hiện phép thử độc lập có 2 lần X nhận giá trị trong khoảng (-1,1).

Bài 2.23. Cho đại lượng ngẫu nhiên liên tục X có hàm phân phối

$$F(x) = \begin{cases} 1 - e^{\frac{-x^2}{2\sigma}^2} & \textbf{khi } x \geq 0, \\ 0 & \textbf{khi } x < 0. \end{cases}$$

σ là hằng số

a. Tìm hàm mật độ xác suất f(x)

b. Tính $P(0 \leq X < \sigma)$.

Bài 2.24. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

$$f(x) = \begin{cases} 0 & \textbf{khi } x \leq 0, x \geq 2, \\ ax^2 & \textbf{khi } 0 < x < 1, \\ a(2-x)^2 & \textbf{khi } 1 \leq x < 2. \end{cases}$$

a. Tìm a để f(x) là hàm mật độ

b. Tìm hàm phân phối xác suất tương ứng

c. Tính các giá trị E(X) và D(X).

Bài 2.25. Cho ĐLNN liên tục X có hàm mật độ

$$f(x) = \begin{cases} 2a(x-1) & \textbf{khi } x \in [1,3], \\ 0 & \textbf{khi } x \notin [1,3]. \end{cases}$$

a. Tìm hằng số a.

b. Xây dựng hàm phân phối xác suất $F(x)$.

c. Tính xác suất $P(0 \le X < 2)$.

Bài 2.26. Cho ĐLNN liên tục X có hàm mật độ

$$f(x) = \begin{cases} K(1-x^2) & \textbf{khi } |x| < 1, \\ 0 & \textbf{khi } |\mathbf{x}| \ge 1. \end{cases}$$

a. Tìm hằng số K.

b. Tính các giá trị E(X), D(X).

c. Tìm hàm phân phối F(x) và vẽ đồ thị tương ứng.

d. Tính $P(-2 < X \le -1/2)$.

Bài 2.27. Tuổi thọ X của một loại thiết bị điện tử là 1 ĐLNN có hàm mật độ như sau:

$$f(x) = \begin{cases} Cxe^{\frac{-x}{2}} & \textbf{khi } x > 0, \\ 0 & \textbf{khi } x \le 0. \end{cases}$$

a. Tìm hằng số C.

b. Tính $P\{X \ge 4\}$ và $P\{X \ge 4/X \ge 2\}$.

c. Tìm kỳ vọng và phương sai của X.

Bài 2.28. ĐLNN X có hàm mật độ như sau:

$$f(x) = Ke^{-|x|}$$

a. Tìm hằng số K và tính $P\{X < 0\}$.

b. Tìm hàm phân phối xác suất của X.

c. Tìm kỳ vọng và phương sai của X.

Bài 2.29. Cho ĐLNN liên tục X có hàm mật độ như sau:

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

a. Tìm hằng số C.

b. Tìm kỳ vọng EX.

c. Tìm hàm phân phối xác suất F(x) của ĐLNN X.

Bài 2.30. Cho ĐLNN liên tục X có hàm mật độ như sau:

$$f(x) = \begin{cases} A\,|\mathrm{sinx}| & \text{khi } |x| < \dfrac{\pi}{2}, \\ 0 & \text{khi } |x| \geq \dfrac{\pi}{2}. \end{cases}$$

a. Tìm hằng số A.

b. Tìm hàm phân phối xác suất của X .

Bài 2.31. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

$$f(x) = \begin{cases} K(sinx + cosx) & \text{khi } x \in [\dfrac{-\pi}{4}, \dfrac{\pi}{4}], \\ 0 & \text{khi } x \notin [\dfrac{-\pi}{4}, \dfrac{\pi}{4}]. \end{cases}$$

a. Xác định hệ số K .

b. Xây dựng phân phối xác suất F(x) của X

c. Tính kỳ vọng E(X), phương sai D(X).

Bài 2.32. Cho đại lượng ngẫu nhiên liên tục X có hàm mật độ:

$$f(x) = \begin{cases} \dfrac{C}{2x^2} & \text{khi } x \in [2, 4], \\ 0 & \text{khi } x \notin [2, 4]. \end{cases}$$

a. Xác định hệ số C .

b. Tìm hàm phân phối F(x).

c. Tính kỳ vọng E(X), phương sai D(X).

Bài 2.33. Cho ĐLNN liên tục X có hàm phân phối xác suất

$$F(x) = \begin{cases} 0 & \textbf{khi } x < -2, \\ \dfrac{1}{2} + \dfrac{1}{\pi}\arcsin\dfrac{x}{2} & \textbf{khi } -2 \le x \le 2, \\ 1 & \textbf{khi } x > 2. \end{cases}$$

a. Tìm hàm mật độ.

b. Tính $P(-1 < x \le 1)$ và EX, DX

c. Tính xác suất để trong 5 lần thực hiện phép thử thì có 2 lần X nhận giá trị trong khoảng (-1,1).

Bài 2.34. Cho X là lợi nhuận doanh nghiệp trong năm của một doanh nghiệp (đơn vị: tỉ) có hàm mật độ như sau:

$$f(x) = \begin{cases} 0 & \textbf{khi } x < -1, \\ k(x+1)^2 & \textbf{khi } -1 \le x \le 3, \\ 0 & \textbf{khi } x > 3. \end{cases}$$

a. Tìm k. Tìm lợi nhuận trung bình năm của doanh nghiệp

b. Độ lệch chuẩn của lợi nhuận bao nhiêu?

c. Xác suất doanh nghiệp làm ăn có lãi bao nhiêu?

Bài 2.35. Cho ĐLNN liên tục X có hàm mật độ như sau:

$$f(x) = \begin{cases} 0 & x < 5, \\ \dfrac{K}{x^3} & \textbf{khi } x \ge 5. \end{cases}$$

a. Tìm K? Tính tuổi thọ trung bình của sản phẩm?

b. Nếu dự định tỷ lệ sản phẩm sẽ phải bảo hành là 20%, vậy phải quy định thời hạn bảo hành là bao nhiêu?

Bài 2.36. Cho ĐLNN liên tục X có hàm mật độ như sau:

$$f(x) = \begin{cases} a\,|x| & \text{khi } x \in [-2,2], \\ 0 & \text{khi } x \notin [-2,2]. \end{cases}$$

a. Tìm a để f(x) là hàm mật độ xác suất

b. Tính E(X) và D(X).

Chương 3
LÝ THUYẾT MẪU VÀ ƯỚC LƯỢNG

Thống kê toán là bộ môn toán học nghiên cứu các quy luật của hiện tượng ngẫu nhiên có tính chất số lớn trên cơ sở thu thập và xử lý số liệu thống kê các kết quả quan sát về hiện tượng ngẫu nhiên này. Nếu ta thu thập được các số liệu liên quan đến tất cả các đối tượng cần nghiên cứu thì ta có thể biết được đối tượng này. Tuy nhiên trong thực tế điều đó không thể thực hiện được vì quy mô của các đối tượng cần nghiên cứu quá lớn hoặc trong quá trình nghiên cứu đối tượng nghiên cứu bị phá hủy. Lý thuyết mẫu cung cấp phương pháp nghiên cứu tổng thể thông qua mẫu. Dựa trên mẫu cùng với lý thuyết ước lượng chúng ta có thể đưa ra những đánh giá về các tham số mà ta quan tâm trong tổng thể.

Chương này trình bày về phương pháp lấy mẫu ngẫu nhiên, xử lý mẫu số liệu và ước lượng cho các tham số.

3.1. Tổng quan về lý thuyết mẫu

3.1.1. Tổng thể và mẫu

★ Tổng thể

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

• Ký hiệu N là số phần tử của tổng thể còn gọi là kích thước của tổng thể.

• X^*: Dấu hiệu khảo sát.

• $x_i\,(i = 1, 2, ..., k)$: Giá trị của dấu hiệu X^* đo được trên phần tử của tổng thể (x_i là thông tin ta quan tâm, các phần tử của tổng thể là vật mang thông tin).

• $N_i\,(i = 1, 2, ..., k)$: Tần số của x_i (số phần tử có chung giá trị x_i).

• $p_i = \dfrac{N_i}{N}$: Tần suất của x_i.

Bảng cơ cấu của tổng thể

Sự tương ứng giữa các giá trị x_i và tần suất p_i được biểu diễn bởi bảng cơ cấu tổng thể theo dấu hiệu X^* như sau:

Giá trị của X^*	x_1	x_2	$\ldots$	x_k
Tần suất p_i	p_1	p_2	$\ldots$	p_k

Trong đó, $0 \le p_i \le 1$ với mọi p_i và $p_1 + p_2 + ... + p_k = 1$.

Về mặt hình thức, bảng cơ cấu của tổng thể tương tự như bảng phân phối xác suất của đại lương ngẫu nhiên rời rạc.

Các tham số đặc trưng của tổng thể

Các tham số đặc trưng của tổng thể biểu diễn các thông tin tổng hợp phản ánh những khía cạnh quan trọng nhất của tổng thể.

(i) Trung bình của tổng thể: $\mu = \sum\limits_{i=1}^{k} x_i p_i$. Nếu xem dấu hiệu nghiên cứu như đại lượng ngẫu nhiên X thì trung bình tổng thể chính là kỳ vọng của dấu hiệu đó, nó chính là trung bình số học của các giá trị của dấu hiệu X^* trong tổng thể.

(ii) Phương sai của tổng thể: $\sigma^2 = \sum\limits_{i=1}^{k} (x_i - \mu)^2 p_i$. Phương sai của tổng thể chính là phương sai của dấu hiệu nghiên cứu trong tổng thể đó. Nó phản ánh mức độ phân tán của các giá trị của dấu hiệu X^* xung quanh giá trị trung bình tổng thể.

Ví dụ 3.1. Một trại chăn nuôi có 1000 con gà. Người ta tiến hành khảo sát chất lượng chăn nuôi gà.

• Tổng thể: Tất cả các con gà của trại chăn nuôi.

• Kích thước của tổng thể: $N = 1000$.

• Dấu hiệu nghiên cứu X^*: Trọng lượng của gà.

• Ta có bảng theo dõi trọng lượng của gà như sau:

Trọng lượng(kg)	*1*	*2*	*3*	*4*
Số con gà (n_i))	*100*	*400*	*300*	*200*

Bảng cơ cấu của tổng thể

X^*	1	2	3	4
p_i	$\frac{1}{10}$	$\frac{2}{5}$	$\frac{3}{10}$	$\frac{1}{5}$

★ **Mẫu**

Mục đích của chương này là nghiên cứu tổng thể thông qua mẫu. Nếu nghiên cứu tổng thể có nghĩa là ta nghiên cứu một hoặc một số đặc trưng nào đó, khi đó không thể đem tất cả các phần tử trong đám đông nghiên cứu vì số lượng lớn và các phần tử bị hỏng nên vừa không đạt mục đích vừa không kinh tế. Vì vậy, ta chỉ lấy một số phần tử trong đám đông ra nghiên cứu và làm sao qua việc nghiên cứu này có thể kết luận được về một hoặc một số đặc trưng của tổng thể mà ta quan tâm lúc đầu.

Ví dụ 3.2. Một nhà máy sản xuất 10.000 hộp sữa. Để kiểm tra chất lượng những hộp sữa đó có đạt tiêu chuẩn qui định không, ta không thể kiểm tra tất cả mà chỉ lấy ngẫu nhiên 50 hộp kiểm tra và qua đó kết luận được về chất lượng của 10.000 hộp sữa đó.

Trong ví dụ này 10.000 hộp sữa được gọi là tổng thể, 50 hộp sữa lấy ra gọi là mẫu.

Định nghĩa 3.3. *Tập hợp các phần tử lấy ra từ một tổng thể được gọi là mẫu. Số lượng các phần tử của mẫu gọi là cỡ mẫu (kích thước mẫu).*

★ Mẫu ngẫu nhiên

Vì có thể mô hình hóa dấu hiệu X^* bằng một đại lượng ngẫu nhiên X với một quy luật phân phối xác suất nào đó nên việc chọn mẫu kích thước n theo nguyên tắc trên có thể xem như tiến hành n phép thử độc lập đối với X và các giá trị X_i của dấu hiệu có thể xem như các đại lượng ngẫu nhiên thu được qua việc tiến hành n phép thử độc lập đối với đại lượng ngẫu nhiên X.

Định nghĩa 3.4. *Mẫu ngẫu nhiên kích thước n là tập hợp của n đại lượng ngẫu nhiên độc lập $X_1, X_2, ..., X_n$ được thành lập từ đại lượng ngẫu nhiên X và có cùng quy luật phân phối xác suất với X.*

Kí hiệu mẫu ngẫu nhiên là $W = (X_1, X_2, ..., X_n)$.

Thực hiện một phép thử đối với mẫu ngẫu nhiên W, tức là thực hiện một phép thử đối với mỗi thành phần X_i của mẫu, giả sử X_1 nhận giá trị x_1, X_2 nhận giá trị x_2,. . . ,X_n nhận giá trị x_n tập hợp n giá trị $x_1, x_2, ...x_n$ tạo thành một giá trị của mẫu ngẫu nhiên (hay mẫu cụ thể) : $w = (x_1, x_2, ..., x_n)$.

Ví dụ 3.5. Gọi X là “số chấm xuất hiện khi gieo một con xúc xắc ”, X là đại lượng ngẫu nhiên có bảng phân phối xác suất như sau:

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

Giải. Nếu gieo con xúc sắc 3 lần. Gọi X_i là “số chấm xuất hiện trong lần gieo thứ i” (i=1,2,3) thì ta có 3 đại lượng ngẫu nhiên độc lập có cùng qui luật phân phối xác suất với X. Vậy ta có mẫu ngẫu nhiên cỡ $n = 3$: $W = (X_1, X_2, X_3)$ được xây dựng từ đại lượng ngẫu nhiên gốc X.

Thực hiện một phép thử đối với mẫu ngẫu nhiên này (tức là gieo 3 lần một con xúc xắc) giả sử lần 1 xuất hiện mặt 6; lần 2 xuất hiện mặt 3; lần 3 xuất hiện mặt 1 thì ta có 1 giá trị của mẫu ngẫu nhiên w = (6, 3, 1).

Chú ý:

- Muốn cho mẫu phản ánh tương đối chính xác tổng thể, thì mẫu phải tiêu biểu tức là mọi phần tử của tổng thể đều có thể được chọn vào mẫu với xác suất như nhau.

- Có hai cách lấy mẫu ngẫu nhiên: Lấy có hoàn lại và lấy không hoàn lại.

3.1.2. Các đặc trưng của mẫu ngẫu nhiên

★ Kỳ vọng của mẫu ngẫu nhiên (Trung bình mẫu ngẫu nhiên)

Định nghĩa 3.6. *Kỳ vọng của mẫu ngẫu nhiên* $W = (X_1, X_2, ..., X_n)$ *là một thống kê ký hiệu và xác định bởi* $\overline{X} = \dfrac{1}{n}(X_1 + X_2 + ...X_n) = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$.

Chú ý:

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

• Nếu mẫu ngẫu nhiên $W = (X_1, X_2, ...X_n)$ có mẫu cụ thể $w_X = (x_1, x_2, ...x_n)$ thì $\overline{X}$ sẽ nhận giá trị $\overline{x} = \dfrac{1}{n}\sum\limits_{i=1}^{k} n_i x_i$ (n_i là tần số xuất hiện các giá trị x_i, $i = 1, 2, \ldots, n$.)

• Nếu đại lượng ngẫu nhiên gốc X có kỳ vọng $E(X) = \mu$ và phương sai $D(X) = \sigma^2$ thì thống kê $\overline{X}$ có $E(\overline{X}) = \mu$ và $D(\overline{X}) = \dfrac{\sigma^2}{n}$ nhỏ hơn phương sai của đại lượng ngẫu nhiên gốc n lần, nghĩa là các giá trị có thể có của $\overline{X}$ ổn định quanh kỳ vọng μ hơn các giá trị có thể có của X.

★ Phương sai của mẫu ngẫu nhiên

Định nghĩa 3.7. *Phương sai của mẫu ngẫu nhiên* $W = (X_1, X_2, ..., X_n)$ *là một thống kê, ký hiệu* S^2 *và được xác định bởi* $S^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)^2$.

Chú ý:

• Vì $X_1, X_2, ..., X_n$ là các đại lượng ngẫu nhiên nên S^2 cũng là đại lượng ngẫu nhiên.

• Nếu mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$ có mẫu cụ thể $w = (x_1, x_2, ..., x_n)$ và các x_i xuất hiện với tần số n_i thì S^2 sẽ nhận giá trị

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

Nếu $E(X) = \mu$; $D(X) = \sigma^2$ thì $E(S^2) = \dfrac{n-1}{n}\sigma^2$.

Trong thực tế để tiện cho việc tính toán người ta sử dụng công thức sau:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2.$$

Trong thống kê, ngoài phương sai mẫu người ta còn thường dùng loại phương sai sau đây:

Phương sai điều chỉnh của mẫu ngẫu nhiên

Đặt $S'^2 = \dfrac{n}{n-1}S^2; S'^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ thì ta có $E(S'^2) = \sigma^2$. Khi đó S'^2 được gọi là *phương sai điều chỉnh* của mẫu ngẫu nhiên W.

Với mẫu cụ thể $w = (x_1, x_2, ..., x_n)$ và các giá trị x_i xuất hiện với tần số n_i thì S'^2 sẽ nhận giá trị

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

★ Độ lệch tiêu chuẩn và độ lệch tiêu chuẩn điều chỉnh của mẫu ngẫu nhiên

• Độ lệch tiêu chuẩn của mẫu ngẫu nhiên ký hiệu và xác định bởi

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

• Độ lệch tiêu chuẩn điều chỉnh của mẫu ngẫu nhiên ký hiệu và xác định bởi

$$S' = \sqrt{S'^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}.$$

★ Tần suất của mẫu ngẫu nhiên

Trường hợp đặc biệt, trong tổng thể ta nghiên cứu chỉ có hai đối tượng A và $\overline{A}$. Tỷ lệ đối tượng A là p chưa biết. Cho mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$ trong đó X_i là số lần xuất hiện A ở lần lấy thứ $i (i = 1, 2, ..., n)$. Nếu A xuất hiện thì X_i nhận giá trị 1. Gọi m là số lần xuất hiện A trong mẫu ngẫu nhiên, khi đó tần suất mẫu là một thống kê ký hiệu và xác định bởi $f = \dfrac{m}{n}$.

3.1.3. Thống kê và phân phối xác suất của các thống kê

Để nghiên cứu đại lượng ngẫu nhiên gốc X, nếu chỉ rút ra một mẫu ngẫu nhiên $W = (X_1, X_2, ..., X_n)$ thì rõ ràng chưa giải quyết được vấn đề gì, bởi các đại lượng ngẫu nhiên X_i có cùng qui luật phân phối xác suất với X mà ta chưa biết được hoàn toàn. Vì vậy, ta phải liên kết hay tổng hợp các đại lượng $X_1, X_2, ..., X_n$ lại sao cho đại lượng ngẫu nhiên mới thu được có những tính chất mới, có thể đáp ứng được yêu cầu giải những bài toán khác nhau về đại lượng ngẫu nhiên X.

Định nghĩa 3.8. *Việc tổng hợp mẫu $W = (X_1, X_2, ..., X_n)$ được thực hiện dưới dạng một hàm nào đó của các đại lượng ngẫu nhiên $X_1, X_2, ..., X_n$ được gọi là một thống kê, ký hiệu là G. Như vậy $G = f(X_1, X_2, ..., X_n)$. Nếu mẫu ngẫu nhiên có giá trị $w = (x_1, x_2, ..., x_n)$, thì $g = f(x_1, x_2, ..., x_n)$ là giá trị của thống kê G (hay còn gọi là g quan sát).*

Sau đây ta xét một số thống kê thông dụng nhất.

★ Quy luật phân phối xác suất của $\overline{X}$

Định lí 3.9. *Nếu $X_1, X_2, ..., X_n$ được lấy từ đại lượng ngẫu nhiên X có quy luật phân phối chuẩn $N(\mu, \sigma^2)$ thì:*

a. $\overline{X}$ có phân phối chuẩn $N(\mu, \dfrac{\sigma^2}{n})$.

b. $U = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n}$ *có phân phối chuẩn* $N(0,1)$.

Chú ý:

• Với kích thước mẫu lớn ($n \geq 30$). Giả sử tổng thể mà dấu hiệu X^* được mô hình hoá bởi đại lượng ngẫu nhiên X có $E(X) = \mu$và $D(X) = \sigma^2$, lấy một mẫu cỡ $n \geq 30$, người ta chứng minh được rằng $\overline{X}$ là đại lượng thống kê có phân phối gần đúng phân phối chuẩn với $E(\overline{X}) = \mu$, $D(\overline{X}) = \dfrac{\sigma^2}{n}$.

• Khi nghiên cứu quy luật của trung bình mẫu $\overline{X}$, nếu không biết phương sai của tổng thể và mẫu cỡ n tương đối nhỏ ($n < 30$) thì ta phải sử dụng phân phối Student.

Người ta đã chứng minh được rằng đại lượng $T = \dfrac{\overline{X} - a}{S'}\sqrt{n}$ có phân phối Student với $n - 1$ bậc tự do.

(Phân phối Student có bảng tính sẵn ở phần phụ lục).

Phân phối Student dần đến phân phối chuẩn khi $n \to \infty$. Thực tế khi $n \geq 30$ ta có thể coi T có phân phối chuẩn.

★ Quy luật phân phối của tần suất mẫu

Trong trường hợp n đủ lớn người ta xây dựng thống kê

$$U = \frac{f - p}{\sqrt{f\,(1 - f)}}\sqrt{n}.$$

Khi đó U cũng có phân phối xấp xỉ $N(0,1)$.

Trong trường hợp từ hai tổng thể rút ra hai mẫu ngẫu nhiên độc lập W_1 kích thước n_1 và W_2 có kích thước là n_2. Từ đó xác định được các tần suất mẫu tương ứng là f_1 và f_2. Xét thống kê

$U = \dfrac{f_1 - f_2}{\sqrt{\overline{f}(1 - \overline{f})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ trong đó $\overline{f} = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$.

Có thể chứng minh được rằng thống kê U sẽ có phân phối xấp xỉ $N(0,1)$ khi n_1 và n_2 khá lớn.

★ Quy luật phân phối của xác suất của S^2

Nếu mẫu ngẫu nhiên được xây dựng từ đại lượng ngẫu nhiên chuẩn với $E(X) = a$ và $D(X) = \sigma^2$ thì người ta đã chứng minh được rằng: $\chi^2 = \dfrac{nS^2}{\sigma^2} = \sum\limits_{i=1}^{n} \dfrac{\left(X_i - \overline{X}\right)^2}{\sigma^2}$ có phân phối χ (khi bình phương), $n-1$ bậc tự do χ^2_{n-1}. Còn thống kê $\chi^2 = \sum\limits_{i=1}^{n} \dfrac{(X_i - a)^2}{\sigma^2}$ có phân phối "khi bình phương" n bậc tự do χ^2_n *.(Phân phối "khi bình phương" có bảng tính sẵn ở phần phụ lục).*

3.1.4. 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.

a.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 $x_1 (i = 1, 2, ..., k)$ với số lần lặp lại (tần số) $n_i (i = 1, 2, ..., k)$. Ta thường lập bảng như sau,

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

trong đó $n_1 + n_2 + ... + n_k = n$.

b. Trường hợp mẫu có kích thước lớn

Ta chia mẫu thành các khoảng (lớp), trong mỗi khoảng ta chọn một giá trị đại diện cho khoảng đó (thường chia thành các khoảng đều nhau), sau đó chọn các giá trị đại diện cho khoảng (thường chọn giá trị đại diện là giá trị trung tâm của khoảng).

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 3.10. Phòng Công tác HSSV tiến hành điều tra số giờ chơi Game trong một tháng của 336 sinh viên được kết quả như sau:

Thời gian	Số Sv	Thời gian	Số Sv
4-12	143	44-52	9
12-20	75	52-60	5
20-28	53	60-68	4
28-36	27	68-76	3
36-44	14	76-80	3

Với các số liệu của mẫu nói trên cho phép ta có thể xác định được các thống kê đặc trưng của mẫu như trung bình mẫu (Số giờ chơi Game trung bình), phương sai mẫu.

Giải. Để tiện cho việc tính toán người ta thường lập bảng tính như sau:

Khoảng	n_i	x_i	n_ix_i	$n_ix_i^2$
4-12	143	8	1144	9152
12-20	75	16	1200	19200
20-28	53	24	1272	30528
28-36	27	32	864	27648
36-44	14	40	560	22400
44-52	9	48	432	20736
52-60	5	56	280	15680
60-68	4	64	256	16384
68-76	3	72	216	15552
76-80	3	78	234	18252
Tổng	336		6458	195532

Trong đó:

• Khoảng: Là khoảng thời gian;

• n_i là số sinh viên;

• x_i là giá trị đại diện cho khoảng.

Khi đó áp dụng các công thức: $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$ và $s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2$.

Ta có kết quả $\overline{x} = 19,92$ và $s^2 = 212,532$.

Việc tính toán các số liệu trên sẽ dễ dàng hơn nếu ta dùng phương pháp đổi biến như sau:

Phương pháp đổi biến

Đặt $u_i = \frac{x_i - x_0}{h}$ trong đó x_0 là giá trị x_i tương ứng với tần số n_i lớn nhất và h là độ dài của khoảng. Khi đó:

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{k} n_i.u_i;\ \ S_u^2 = \frac{1}{n}\sum_{i=1}^{k} n_i.u_i^2 - (\overline{u})^2.$$

Trở lại số liệu ban đầu ta có: $\overline{x} = x_0 + h.\overline{u};\ s_x^2 = h^2.s_u^2$.

Áp dụng cho ví dụ 3.10 ta có bảng sau

Khoảng	n_i	x_i	u_i	n_iu_i	$n_ix_i^2$
4-12	143	8	0	0	0
12-20	75	16	1	75	75
20-28	53	24	2	106	212
28-36	27	32	3	81	243
36-44	14	40	4	56	224
44-52	9	48	5	45	225
52-60	5	56	6	30	180
60-68	4	64	7	28	196
68-76	3	72	8	24	192
76-80	3	78	8,75	26,25	229,69
Tổng	336			471,25	1176,69

Ta cũng có các kết quả như trên.

3.2. Lý thuyết ước lượng

3.2.1. Bài toán ước lượng

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 $\theta' = \theta'(X_1, X_2, \ldots, X_n)$ để ước lượng (dự đoán) θ Có hai 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).

3.2.2. Các loại ước lượng điểm và tiêu chuẩn lựa chọn hàm ước lượng

Phương pháp ước lượng điểm chủ trương là dùng một giá trị đã biết thay thế cho tham số θ chưa biết của tổng thể. Thông thường giá trị được chọn là một thống kê G nào đó của mẫu ngẫu nhiên. Có nhiều cách chọn thống kê G khác nhau tạo nên những phương pháp ước lượng điểm khác nhau.

Giả sử cần ước lượng tham số θ của đại lượng ngẫu nhiên X, từ tổng thể lập mẫu ngẫu nhiên cỡ n: $W = (X_1, X_2, ..., X_n)$. Chọn lập thống kê $G = f(X_1, X_2, ..., X_n)$. Ta gọi G là hàm ước lượng của θ. Có vô số cách chọn hàm f, do đó có vô số thống kê G. Vì vậy, cần đưa ra tiêu chí để đánh giá chất lượng của các thống kê G, từ đó lựa chọn được thống kê "xấp xỉ một cách tốt nhất" tham số cần ước lượng.

• Các tiêu chuẩn lựa chọn hàm ước lượng

a. Ước lượng không chệch

Định nghĩa 3.11. *Thống kê G gọi là ước lượng không chệch của tham số θ*

của đại lượng ngẫu nhiên X nếu $E(G) = \theta$.
Ngược lại nếu $E(G) \neq \theta$ thì G là ước lượng chệch của θ.

Ý nghĩa Giả sử G là ước lượng không chệch của tham số θ. Ta có:

$$E(G - \theta) = E(G) - E(\theta) = \theta - \theta = 0.$$

Vậy ước lượng không chệch là ước lượng có sai số trung bình bằng 0.

Chú ý: G là ước lượng không chệch của θ không có nghĩa là mọi giá trị của G đều trùng khít với θ, mà chỉ có nghĩa trung bình các giá trị của G bằng θ . Một giá trị của G có thể lệch rất lớn so với θ.

Một số kết quả

(i) Trung bình của mẫu ngẫu nhiên $\overline{X}$ là ước lượng không chệch của trung bình của tổng thể $\theta = E(X) = \mu$ vì $E(\overline{X}) = \mu$.

(ii) Phương sai điều chỉnh của mẫu ngẫu nhiên S'^2 là ước lượng không chệch của phương sai của tổng thể σ^2 vì $E(S'^2) = \sigma^2$.

Ví dụ 3.12. Chiều cao trung bình của 50 cây lim được cho bởi:

Khoảng chiều cao (mét)	Số cây lim	x_i^0	u_i	n_iu_i	$n_ix_i^2$
[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	-1	20
[7,75;8,25)	11	8,0	-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
Tổng	50			-13	95

Gọi X là chiều cao của cây lim.

(i) Hãy chỉ ra ước lượng điểm cho chiều cao trung bình của các cây lim.

(ii) Hãy chỉ ra ước lượng điểm cho độ tản mát của các chiều cao cây lim so với chiều cao trung bình.

(iii) Gọi $p = P(7,75 \leq \overline{X} \leq 8,75)$. Hãy chỉ ra ước lượng điểm cho p.

Giải. Ta lập bảng tính cho $\overline{x}$ và s^2
Thực hiện phép đổi biến $u_i = \dfrac{x_i^0 - 8,5}{0,5}$ $(x_0 = 8,5; h = 0,5)$.
Ta có $\overline{u} = \dfrac{-13}{50} = -0,26$. Suy ra $\overline{x} = 8,5 + 0,5.(-0,26) = 8,37$

$$s^2 = (0,5)^2.[\frac{95}{50} - (-0,26)^2] = 0,4581 \approx (0,68)^2$$

$$s'^2 = \frac{50}{50-1}.s^2 \approx (0,684)^2.$$

Vậy:
• Chiều cao trung bình được ước lượng là $8,37$ mét.
• Độ tản mát được ước lượng là $s = 0,68$ mét .
• Trong 50 quan sát đã cho có $11 + 18 = 29$ quan sát cho chiều cao lim thuộc khoảng $[7,75; 8,75]$. Vậy ước lượng điểm cho p là $p^* = \dfrac{29}{50} = 0,58.$

b. Ước lượng hiệu quả

Nhận xét. Giả sử G là ước lượng không chệch của tham số θ. Theo bất đẳng thức Tchebychev ta có:

$$P\left(|G - E(G| < \varepsilon\right) > 1 - \frac{D(G)}{\varepsilon^2}.$$

Vì $E(G) = \theta$ nên

$$P(|G - \theta| < \varepsilon) > 1 - \frac{D(G)}{\varepsilon^2}.$$

Ta thấy, nếu $D(G)$ càng nhỏ thì $P(|G - \theta| < \varepsilon)$ càng gần 1. Do đó ta sẽ chọn G với $D(G)$ nhỏ nhất.

Định nghĩa 3.13. *Thống kê G của mẫu được gọi là ước lượng hiệu quả của tham số θ của đại lượng ngẫu nhiên X nếu nó có phương sai nhỏ nhất so với mọi thống kê khác được xây dựng trên cùng mẫu đó.*

Hay ước lượng không chệch G được gọi là ước lượng có hiệu quả của tham số θ nếu $D(G)$ nhỏ nhất trong các ước lượng của θ.

Như vậy để xét xem G có phải là ước lượng hiệu quả của θ hay không, vấn đề đặt ra là phải tìm được cận dưới của phương sai các hàm ước lượng.

Chú ý: Người ta chứng minh được rằng nếu G là ước lượng hiệu quả của θ thì phương sai của nó là

$$D\left(G\right)=\frac{1}{n.E\left(\frac{\partial \ln f(x,\theta)}{\partial \theta}\right)^{2}}.$$

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 $D(G)$ trong công thức trên. Ta gọi công thức trên là *giới hạn Crame-Rao*.

Kết quả: Nếu đại lượng ngẫu nhiên gốc $X \in N\left(\mu,\sigma^2\right)$ thì trung bình mẫu $\overline{X}$ là ước lượng hiệu quả của kỳ vọng $E(X)=\mu$.

c. Ước lượng vững

Khi xét những mẫu có kích thước lớn thì nảy sinh vấn đề là mẫu càng lớn thì thống kê G của mẫu phải càng gần tham số θ cần ước lượng. Từ đó ta có định nghĩa:

Định nghĩa 3.14. *Thống kê $G=f(X_1,X_2,\ldots,X_n)$ gọi là ước lượng vững của tham số θ nếu $\forall\varepsilon>0$ bé tuỳ ý cho trước ta đều có $\lim\limits_{n\to\infty}P\left(|G-\theta|<\varepsilon\right)=1$, hay nói cách khác G hội tụ theo xác suất đến θ.*

Điều kiện đủ của ước lượng vững

Nếu G là ước lượng không chệch của θ và $\lim D(G)=0 (n\to\infty)$ thì G là ước lượng vững của θ.

Một số kết quả:

• Trung bình mẫu $\overline{X}$ là ước lượng vững của tham số μ .

• Tần suất mẫu f là ước lượng vững của xác suất p của đại lượng ngẫu nhiên gốc X.

Một vài kết luận của phương pháp ước lượng điểm

Dùng những tiêu chuẩn nêu trên để đánh giá các thống kê đặc trưng mẫu khác nhau cho phép ta lựa chọn được những thống kê tốt nhất. Từ đó đưa ra một số kết luận sau:

• Vì trung bình mẫu $\overline{X}$ là ước lượng không chệch , hiệu quả và ước lượng vững của trung bình tổng thể μ, do đó *nếu chưa biết μ có thể dùng $\overline{X}$ để ước lượng*.

• Vì tần suất mẫu f là ước lượng không chệch, ước lượng hiều quả và ước lượng vững của tần suất tổng thể p do đó *nếu chưa biết p có thể dùng f để ước lượng*.

• Vì phương sai điều chỉnh mẫu S'^2 và phương sai S^2 đều là các ước lượng không chệch của phương sai tổng thể σ^2 , do đó *nếu chưa biết phương sai σ^2 có thể dùng S'^2 hoặc S^2 để ước lượng*.

• Người ta cũng chứng minh được rằng $\overline{X}$ là ước lượng vững của $E(x) = \mu$ và $S^2; S'^2$ là ước lượng vững của $D(x) = \sigma^2$.

Nhận xét. Các phương pháp ước lượng điểm có một nhược điểm cơ bản là khi kích thước mẫu nhỏ thì ước lượng điểm tìm được có thể sai lệch rất nhiều so với giá trị của tham số cần ước lượng túc là sai số của ước lượng có thể rất lớn. Mặt khác dùng các phương pháp trên không thể đánh giá được khả năng mắc sai lầm khi ước lượng bằng bao nhiêu. Do đó khi kích thước mẫu nhỏ người ta thường sử dụng phương pháp ước lượng bằng khoảng tin cậy.

3.3. Các trường hợp ước lượng khoảng

3.3.1. Mô tả phương pháp

Giả sử tổng thể có tham số θ chưa biết. Ta tìm khoảng (G_1, G_2) chứa θ sao cho $P\left(G_1 < \theta < G_2\right) = 1 - \alpha$ cho trước. Xác suất $1 - \alpha$ được gọi là độ tin cậy của ước lượng, $|G_2 - G_1|$ được gọi là độ dài của khoảng tin cậy.
Như vậy, vấn đề chủ yếu của phương pháp ước lượng bằng khoảng tin cậy là làm thế nào để xác định được khoảng tin cậy (G_1, G_2) thoả mãn

$$P(G_1 < \theta < G_2) = 1 - \alpha.$$

Để làm điều đó người ta tiến hành như sau:
Từ tổng thể lập mẫu ngẫu nhiên $W_X = (X_1, X_2, .., X_n)$. Chọn thống kê $G = f(X_1, X_2, .., X_n, \theta)$ có phân phối xác suất xác định dù chưa biết θ.
Với độ tin cậy $1 - \alpha$ cho trước có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và tương ứng với chúng tìm được cặp giá trị $g\alpha_1$ và $g\alpha_2$ (thường là các phân vị của G) thoả mãn điều kiện: $P(G < g\alpha_1) = \alpha_1$ và $P(G > g\alpha_2) = \alpha_2$
.
Từ đó

$$P\left(g\alpha_1 < G < g\alpha_2\right) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha \qquad (3.1)$$

Như vậy, với độ tin cậy $1-\alpha$ ta đã xây dựng được khoảng tin cậy $(g\alpha_1; g\alpha_2)$ cho G. Bằng các phép biến đổi tương đương ta có thể đưa (3.1) về dạng: $P(G_1 < \theta < G_2) = 1 - \alpha$.

Như đã nói trên do α khá nhỏ nên $1 - \alpha$ khá lớn, thông thường trong thực tế thường yêu cầu $1 - \alpha = \gamma \geq 0,95$ để có thể áp dụng nguyên lý xác suất lớn cho biến cố $(G_1 < \theta < G_2)$.

Gọi $I = G_2 - G_1$ là độ dài khoảng tin cậy (I có thể là hằng số, có thể là đại lượng ngẫu nhiên).

Do xác suất $1-\alpha=\gamma$ khá lớn nên biến cố $(G_1<\theta<G_2)$ hầu như chắc chắn xảy ra trong một phép thử. Thực hiện một phép thử đối với mẫu ngẫu nhiên $w_X=(X_1,X_2,..,X_n)$ thu được mẫu cụ thể $w_X=(x_1,x_2,...,x_n)$, do đó tính được các giá trị của G_1 và G_2 ứng với mẫu cụ thể này, ký hiệu là g_1,g_2. Như vậy có thể kết luận là với độ tin cậy $1-\alpha=\gamma,\theta$ nằm trong khoảng (g_1,g_2) tức là $(g_1<\theta<g_2)$.

Nhận xét. Phương pháp ước lượng bằng khoảng tin cậy khắc phục được những nhược điểm của phương pháp ước lượng điểm. Chẳng những nó làm tăng độ chính xác của ước lượng mà còn đánh giá được mức độ tin cậy của ước lượng đó. Tuy nhiên nó cũng chứa đựng khả năng mắc sai lầm bằng α.

Dưới đây sẽ áp dụng phương pháp này để ước lượng một vài tham số đặc trưng cơ bản của đại lượng ngẫu nhiên X trong tổng thể. Giả sử trung bình của tổng thể $E(X)=\mu$ chưa biết. Ta tìm khoảng (m_1,m_2) chứa μ sao cho $P(m_1<\mu<m_2)=1-\alpha$, với $1-\alpha$ là độ tin cậy cho trước.

a. Trường hợp 1: Biết $D(X)=\sigma^2$ và cỡ mẫu $n\geq 30$ hoặc ($n<30$ nhưng X có phân phối chuẩn)

Chọn lập thống kê: $U=\dfrac{(\overline{X}-\mu)}{\sigma}.\sqrt{n}$. Ta có $U\in N(0,1)$.

Chọn cặp α_1 và α_2 sao cho $\alpha_1+\alpha_2=\alpha$ và tìm các phân vị chuẩn tương ứng là u_{α_1} và $u_{1-\alpha_2}$ thỏa mãn các điều kiện $P(U<u_{\alpha_1})=\alpha_1, P(U>u_{1-\alpha_2})=\alpha_2$.

Từ đó, $P(u_{\alpha_1}<U<u_{1-\alpha_2})=1-(\alpha_1+\alpha_2)=1-\alpha$.

Do phân vị chuẩn có tính chất $u_{\alpha_1}=-u_{1-\alpha_1}$ nên

$$P\left(-u_{1-\alpha_1}<U<u_{1-\alpha_2}\right)=1-\alpha \tag{3.2}$$

Thay biểu thức của U vào (3.2) và giải ra μ, ta thu được

$$\overline{X}-\frac{\sigma}{\sqrt{n}}.u_{1-\alpha_2}<\mu<\overline{X}+\frac{\sigma}{\sqrt{n}}.u_{1-\alpha_1} \tag{3.3}$$

(3.3) mới chỉ ra một khoảng tin cậy tổng quát. Với độ tin cậy $(1-\alpha)$ cho trước có thể tìm được vô số cặp giá trị α_1 và α_2 thỏa mãn điều kiện $\alpha_1+\alpha_2=\alpha$, từ đó sẽ có vô số khoảng tin cậy tương ứng. Trong thực tế người ta thường chỉ sử dụng một số trường hợp đặc biệt sau:

Khoảng tin cậy đối xứng

Để được khoảng tin cậy đối xứng ta chọn $\alpha_1 = \alpha_2$ và đặt $\gamma = 1 - \dfrac{\alpha}{2}$ thì

$$\overline{X} - \frac{\sigma}{\sqrt{n}}.u_\gamma < \mu < \overline{X} - \frac{\sigma}{\sqrt{n}}.u_\gamma.$$

Tóm lại, ta tìm được khoảng tin cậy $(\overline{x} - \varepsilon; \overline{x} + \varepsilon)$, trong đó :

(i) $\overline{x}$ là trung bình của mẫu cụ thể.

(ii) $\varepsilon = u_\gamma \dfrac{\sigma}{\sqrt{n}}$ với $u_g amma$ là phân vị chuẩn mức $\gamma = 1 - \dfrac{\alpha}{2}$, ε được gọi là độ chính xác của ước lượng. Nó phản ánh mức độ sai lệch của trung bình mẫu so với trung bình tổng thể với xác suất $(1 - \alpha)$ cho trước.

Khoảng tin cậy bên phải *(Ước lượng giá trị tối thiểu).*

Lấy $\alpha_1 = 0, \alpha_2 = \alpha$ thì $u_{1-\alpha_1} = u_1 = +\infty$, do đó khoảng tin cậy bên phải của μ là:

$$\left(\overline{x} - \frac{\sigma}{\sqrt{n}}.u_{1-\alpha}; +\infty\right).$$

Khoảng tin cậy bên trái *(ước lượng giá trị tối đa).*

Lấy $\alpha_1 = \alpha, \alpha_2 = 0$ thì $u_{1-\alpha_2} = u_1 = +\infty$, do đó khoảng tin cậy bên trái của μ là:

$$\left(-\infty; \overline{x} + \frac{\sigma}{\sqrt{n}}.u_{1-\alpha}\right).$$

Nhận xét. Với cùng độ tin cậy $1 - \alpha$ hiển nhiên khoảng tin cậy nào ngắn hơn sẽ tốt hơn. Trong trường hợp này độ dài khoảng tin cậy $I = (\overline{x} + \varepsilon) - (\overline{x} - \varepsilon) = 2\varepsilon$ sẽ là ngắn nhất khi khoảng tin cậy là đối xứng.

Xác định cỡ mẫu

Xuất phát từ khoảng tin cậy đối xứng, dễ dàng thấy được một ước lượng tốt nếu $(1 - \alpha)$ lớn còn ε khá nhỏ. Nếu yêu cầu một chất lượng ước lượng với $(1 - \alpha)$ và ε ấn định trước, bằng cách nào có thể thoả mãn yêu cầu này? Giải pháp duy nhất là lựa chọn cỡ mẫu cho phù hợp. Trong trường hợp này ta có thể chọn $n = u_\gamma^2 \dfrac{\sigma^2}{\varepsilon^2}$.

Ví dụ 3.15. Khối lượng sản phẩm là đại lượng ngẫu nhiên 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

(i) Hãy ước lượng trung bình khối lượng của sản phẩm với độ tin cậy 95%.

(ii) Nếu yêu cầu độ chính xác là 0,1, giữ nguyên độ tin cậy 95% thì cỡ mẫu là bao nhiêu thì phù hợp ?

Giải.

x_i	n_i	$n_i x_i$
18	3	54
19	5	95
20	15	300
21	2	42
$\sum$	25	491

Ta có $\overline{x} = \dfrac{491}{25} = 19,64kg$.

Độ tin cậy $1 - \alpha = 0,95$ do đó $\alpha = 0,05$. Suy ra $\gamma = 1 - \dfrac{\alpha}{2} = 0,975$.

Ta tìm được phân vị chuẩn $u_\gamma = u_0,975 = 1,96$.

Do đó, $\varepsilon = u_0,975.\dfrac{1}{\sqrt{25}} = 1,96.\dfrac{1}{5} = 0,39$

$x_1 = \overline{x} - \varepsilon = 19,6 - 0,39 = 19,25$,

$x_2 = \overline{x} + \varepsilon = 19,6 + 0,39 = 20,03$.

Vậy khoảng tin cậy là (19,25 ; 20,03)

Kết luận: Trọng lượng trung bình của loại sản phẩm nói trên từ 19,25kg đến

20,03kg với độ tin cậy 95%.

Để độ chính xác chỉ là 0,1, giữ nguyên độ tin cậy 95% thì cỡ mẫu là

$$n = u_\gamma^2 \frac{\sigma^2}{\varepsilon^2} = 1,96^2.\frac{1}{0,01} = 385.$$

b. Trường hợp 2: Chưa biết σ^2 và cỡ mẫu $n \geq 30$

Trường hợp này kích thước mẫu lớn $(n \geq 30)$ có thể dùng ước lượng của S'^2 thay cho σ^2 chưa biết (vì $E(S'^2) = \sigma^2$).

• **Khoảng tin cậy đối xứng :** $(\overline{x} - \varepsilon; \overline{x} + \varepsilon)$

trong đó: $\varepsilon = u_\gamma \frac{s'}{\sqrt{n}}$ (độ chính xác) 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.

• **Khoảng tin cậy bên phải :**$(\overline{x} - \frac{s'}{\sqrt{n}}.u_{1-\alpha}; +\infty)$.

• **Khoảng tin cậy bên trái :** $(-\infty; \overline{x} + \frac{s'}{\sqrt{n}}.u_{1-\alpha})$.

• **Xác định cỡ mẫu :** $n = u_\gamma^2 \frac{s'^2}{\varepsilon^2}$.

Ví dụ 3.16. 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: $n = 59, \overline{x} = 41,05, s' = 27,99$.

Độ tin cậy $1 - \alpha = 0,95$ do đó $\alpha = 0,05$. Suy ra $\gamma = 1 - \frac{\alpha}{2} = 0,975$. Ta tìm được phân vị chuẩn $u_\gamma = u_{0,975} = 1,96$. Do đó $\varepsilon = 1,96.\dfrac{27,99}{\sqrt{59}} \approx 7,13$.

$\overline{x} - 7,13 = 33,92$, $\overline{x} + 7,13 = 48,18$.

Vậy khoảng tin cậy ước lượng là $(33,92; 48,18)$.

c. Trường hợp 3: Chưa biết phương sai $D(X) = \sigma^2$ và $n < 30$ và X có phân phối chuẩn

Chọn lập thống kê

$$T = \frac{(\overline{X} - \mu)}{S'}\sqrt{n} \in T(n-1)$$

.

• **Khoảng tin cậy đối xứng** $\left(\overline{x} - \dfrac{s'}{\sqrt{n}}.t_\gamma^{n-1} < \mu < \overline{x} + \dfrac{s'}{\sqrt{n}}.t_\gamma^{n-1}\right)$, trong đó $\gamma = 1 - \dfrac{\alpha}{2}$. Sai số của ước lượng $\varepsilon = t_\gamma^{n-1}\dfrac{s'}{\sqrt{n}}$.

• **Ước lượng giá trị tối thiểu** $\left(\overline{x} - t_{1-\alpha}^{n-1}\dfrac{s'}{\sqrt{n}}; +\infty\right)$.

• **Ước lượng giá trị tối đa** $\left(-\infty; \overline{x} + t_{1-\alpha}^{n-1}\dfrac{s'}{\sqrt{n}}\right)$.

Phương pháp mẫu kép: Xác định kích thước mẫu tối thiểu n sao cho với độ tin cậy $(1-\alpha)$ cho trước, độ dài khoảng tin cậy không vượt quá giá trị I_o cho trước:

(i) Trước hết điều tra một mẫu sơ bộ kích thước $m \geq 2$: $W_1 = (X_1, X_2, \ldots, X_m)$ từ đó tìm được phương sai điều chỉnh mẫu của mẫu sơ bộ đó: $S'^2 = \dfrac{1}{m-1}\sum\limits_{i=1}^{m}\left(X_i - \overline{X}\right)^2$.

(ii) Sau đó lập mẫu thứ hai kích thước $n - m : W_2 = (X_{m+1}, \ldots, X_n)$ $I_o = \dfrac{2S'}{\sqrt{n}}t^{m-1}_{1-\dfrac{\alpha}{2}}$. Từ đó kích thước mẫu cần tìm là :$n \geq \left[\dfrac{4S'^2}{I_0^2}t_{1-\frac{\alpha}{2}}^{m-1}\right]$

Ví dụ 3.17. Theo dõi mức xăng hao phí (X) cho một loại ôtô đi từ A đến B thu được bảng số liệu sau:

Mức xăng hao phí (l)	*19,0-19,5*	*19,5-20,0*	*20,0-20,5*	*20,5-21,0*
Số lần đi	*2*	*10*	*8*	*5*

Với độ tin cậy $1-\alpha = 0,95$. Hãy tính mức xăng hao phí trung bình tối thiểu khi đi từ A đến B biết X có phân phối chuẩn.

Giải. Đây là bài toán ước lượng khoảng của kỳ vọng (giá trị tối thiểu), trong phân phối chuẩn trường hợp chưa biết phương sai, mẫu cỡ $n = 25 < 30$.

Áp dụng: $\left(\overline{x} - t_{1-\alpha}^{n-1}\dfrac{s'}{\sqrt{n}}; +\infty\right)$, trong đó $t_{1-\alpha}^{n-1} = t_{24}^{0,975} = 1,711$ (tra bảng Student).

Với mẫu cụ thể trên, ta lập bảng tính toán $\overline{x}$ và s': Chọn $x_0 = 19,75; h = 0,5$.

X(lit)	x_i	n_i	u_i	n_iu_i	$n_1u_i^2$
19,0-19,5	19,25	2	-1	-2	2
19,5-20,0	19,75	10	0	0	0
20,0-20,5	20,25	8	1	8	8
20,5-21,0	20,75	5	2	10	20
$\sum$		$n = 25$		16	30

Vậy $\overline{x} = 19,75 + \dfrac{0,5}{25}16 = 20,07$,

$$S^2 = (0,5)^2\left[\frac{1}{25}30 - \left(\frac{1}{25}.16\right)^2\right] = 0,25[1,2-0,4096] = 0,1976,$$

$$S'^2 = \frac{25}{24}.0,1976 = 0,2058 \rightarrow S' = 0,45.$$

Với độ tin cậy $(1-\alpha) = 0,95$, qua mẫu này một giá trị cho mức tối thiểu của μ là $(20,07 - 1,711\dfrac{0,45}{\sqrt{n}}; +\infty)$ hay $(19,92 < \mu < +\infty)$.

Kết luận: Mức xăng hao phí trung bình tối thiểu khi đi từ A đến B là 19,92lít với độ tin cậy 0,95.

Ví dụ 3.18. Để nghiên cứu năng suất của giống lúa A tại một vùng, người ta gặt ngẫu nhiên tại 100 điểm của vùng đó thu được bảng số liệu sau:

Năng suất(tạ/ha)	40- 42	42- 44	44- 46	46- 48	48- 50	50- 52
Số điểm gặt tương ứng	7	13	25	35	15	5

Hãy ước lượng năng suất trung bình của giống lúa đó trong vùng bằng khoảng tin cậy đối xứng với độ tin cậy $(1-\alpha) = 0,95$. Biết năng suất lúa phân phối chuẩn.

Giải. Đây là bài toán ước lượng khoảng của kỳ vọng(khoảng đối xứng) trong phân phối chuẩn, trường hợp chưa biết phương sai, mẫu cỡ $n = 100 > 30$.

Áp dụng khoảng $(\overline{x} - u_{1-\frac{\alpha}{2}}.\frac{S'}{\sqrt{n}}; \overline{x} + u_{1-\frac{\alpha}{2}}.\frac{S'}{\sqrt{n}})$.

Trong đó $u_{1-\frac{\alpha}{2}} = 1,96$ (tra bảng hàm số Laplace) với mẫu cụ thể trên, ta lập bảng tính toán để tính $\overline{x}$ và S'

X(tạ/ha)	x_i	n_i	u_i	n_iu_i	$n_iu_i^2$
40-42	41	7	-3	-21	63
42-44	43	13	-2	-26	52
44-46	45	24	-1	-25	25
46-48	47	35	0	0	0
48-50	49	15	1	15	15
$\sum$		$n = 100$		-47	175

Chọn $x_0 = 47; h = 2$,

$\overline{x} = 47 + \frac{2}{100}(-47) = 46,06;$

$$s^2 = 2^2.\left[1100175 - \left(\frac{1}{100}(-0,47)^2\right)\right] = 6,1164;$$

$s'^2 = \frac{100}{99}.6,1164 = 6,17818$ suy ra $s' \approx 2,48$.

Với độ tin cậy 0,95 qua mẫu cụ thể này, khoảng tin cậy đối xứng của μ là:

$$\left(46,06 - 1,96.\frac{2,48}{\sqrt{100}}; 46,06 + 1,96.\frac{2,48}{\sqrt{100}}\right).$$

Hay $(45,57 < \mu < 46,55)$.

Kết luận: Năng suất trung bình của giống lúa A trong vùng từ $45,57 \to 46,55$ tạ/ha với độ tin cậy 0,95.

Ví dụ 3.19. Phỏng vấn 5 gia đình có con học đại học về chi phí hàng tháng cho nhu cầu sinh hoạt được các số liệu sau: 150 ngàn, 200 ngàn, 250 ngàn, 300 ngàn. Hỏi phải phỏng vấn bao nhiêu gia đình để với độ tin cậy 95% sai số của việc ước lượng chi phí trung bình hàng tháng không vượt quá 30 ngàn. Giả thiết chi phí hàng tháng là ĐLNN phân phối chuẩn.

Giải. Gọi X: Chi phí sinh hoạt hàng tháng, $X \in N(\mu, \sigma^2)$.

Vậy chi phí trung bình chính là giá trị μ . Đây là bài toán xác định kích thước mẫu tối thiểu cho việc ước lượng tham số μ của phân phối $N(\mu, \sigma^2))$ khi chưa biết phương sai σ^2 .

Theo phương pháp mẫu kép, từ mẫu sơ bộ kích thước $n = 5$ ta tìm được: $\overline{x} = 216$ ngàn, $s'^2 = 3530$ ngàn, $I_0 = 2.30$ ngàn $= 60$ ngàn,$t^4_{0,975} = 2,776$.

Do đó $n \geq \dfrac{4.3530}{60^2}(2,776)^2 = 31$ gia đình.

Như vậy phải phỏng vấn thêm 31 - 5 = 26 gia đình nữa.

3.3.2. Ước lượng khoảng cho tỷ lệ (xác suất)

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 ĐLNN có phân phối xác suất.

X	0	1
P	$1-p$	p

Gọi $X_i, (i = 1, 2, .., n)$ là số phần tử có tính chất A trong lần lấy thứ i. Ta có $\overline{X}$ chính là tần suất ước lượng điểm của $p = E(X)$. Mặt khác, theo chương 2 thì nX có phân phối nhị thức $B(n, p)$.

Từ đó $E(\overline{X}) = p$ và $D(\overline{X}) = \dfrac{p(1-p)}{n}$.

Chọn thống kê $U = \dfrac{(f-p).\sqrt{n}}{\sqrt{p(1-p)}}$, trong đóf 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:

$$f - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} < p < f - u_{1-\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}.$$

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.
$\varepsilon = u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{f(1-f)}{n}}$(sai số).

Từ công thức trên ta có: $u_{1-\frac{\alpha}{2}} = \dfrac{\varepsilon.\sqrt{n}}{\sqrt{f(1-f)}}$.

• ***Khoảng tin cậy đối xứng*** $\left(f - u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{f(1-f)}{n}} < p < f - u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{f(1-f)}{n}}\right)$.

Độ chính xác của ước lượng $\varepsilon = u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{f(1-f)}{n}}$. Độ dài khoảng tin cậy đối xứng là $I = 2\varepsilon$.

• ***Ước lượng giá trị tối thiểu***: $f - u_{1-\alpha}\sqrt{\dfrac{f(1-f)}{n}} < p < +\infty$.

• ***Ước lượng giá trị tối đa***: $-\infty < p < f + u_{1-\alpha}\sqrt{\dfrac{f(1-f)}{n}}$.

• ***Kích thước mẫu*** : Được xác định bởi: $n = u^2_{1-\frac{\alpha}{2}}\dfrac{f(1-f)}{\varepsilon^2}$.

Ví dụ 3.20. Điều tra nhu cầu tiêu dùng một loại hàng trong 100 gia đình thấy 60 gia đình có nhu cầu loại hàng hoá nói trên. Với độ tin cậy $1-\alpha = 0,95$, hãy tìm khoảng tin cậy đối xứng của tỷ lệ gia đình có nhu cầu loại hàng đó.

Giải. Đây là bài toán ước lượng khoảng (khoảng đối xứng) của xác suất, áp dụng khoảng trong đó $u_{1-\frac{\alpha}{2}} = 1,96$ (tra bảng Laplace).

Từ mẫu cụ thể trên ta có $f = \dfrac{60}{100} = 0,6, 1-f = 1-0,6 = 0,4$.

Với độ tin cậy 0,95 khoảng tin cậy đối xứng của p là
$(0,6 - 1,96\sqrt{\dfrac{0,6.0,4}{100}}; 0,6 + 1,96\sqrt{\dfrac{0,6.0,4}{100}})$ hay $(0,504 < p < 0,696)$.

Kết luận: Tỷ lệ gia đình có nhu cầu loại hàng đó từ 50,4% đến 69,6% với độ tin cậy là 0,95.

Ví dụ 3.21. 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 = 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. Độ tin cậy $1-\alpha = 99\% \to 1 - \dfrac{\alpha}{2} = 0,995; n = 100, f = \dfrac{20}{100} = 0,2.$

Xét $U = \dfrac{(f-p)\sqrt{100}}{\sqrt{pq}} \in N(0,1),$

Ta có $\varepsilon = u_{0,995}\dfrac{\sqrt{0,2.0,8}}{\sqrt{100}} = 2,58.\dfrac{0,4}{10} = 0,1;$

$f_1 = f - \varepsilon = 0,2 - 0,1 = 0,1; f_2 = f + \varepsilon = 0,2 + 0,1 = 0,3.$

Vậy khoảng tin cậy là $(0,1; 0,3)$.

$$u_{1-\frac{\alpha}{2}}\frac{0,04.\sqrt{100}}{0,2.0,8} = 1 \Rightarrow 1 - \frac{\alpha}{2} = 0,84 \Rightarrow 1-\alpha = 0,68.$$

Vậy độ tin cậy là 68%.

$$n \approx \frac{(2,576)^2.0,2.0,8}{(0,04)^2} = 6,635.100 = 663,5.$$

Vậy $n = 664$.

3.4. Bài tập chương 3

Bài 3.1. Theo dõi mức hao phí nguyên liệu cho một đơn vị sản phẩm tại nhà máy A thu được số liệu sau:

Mức hao phí nguyên liệu	19 - 20	20 - 21	21 - 22	22 - 23
Số sản phẩm	5	15	10	6

Với độ tin cậy 95%, hãy xác định mức hao phí nguyên liệu trung bình tối đa trên một đơn vị sản phẩm của nhà máy trên biết mức hao phí nguyên liệu là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn.

Bài 3.2. Năng suất của một loại ngô mới tại một vùng trồng thử nghiệm là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn. Kiểm tra năng suất tại 100 thửa ruộng thu được kết quả sau:

Năng suất (Tạ/ha)	Số thửa
30-35	7
35-40	12
40-45	18
45-50	27
50-55	20
55-60	8
60-65	5
65-70	3

a. Với độ tin cậy 95%, hãy ước lượng cho năng suất trung bình tối thiểu của loại ngô mới tại vùng đó.

b. Bằng khoảng tin cậy đối xứng hãy ước lượng cho tỷ lệ thửa ruộng có năng suất trên 50 tạ/ha với độ tin cậy 99%.

Bài 3.3. Để xác định lượng cá trong một chiếc hồ rộng người ta bắt ngẫu nhiên 4000 con, đánh dấu rồi thả xuống hồ. Sau một thời gian người ta bắt ngẫu nhiên 200 con trong hồ thấy có 75 con được đánh dấu. Với độ tin cậy 99,5%, hãy ước lượng số cá tối thiểu có trong hồ.

Bài 3.4. Một vùng có 2000 hộ gia đình. Để điều tra nhu cầu tiêu dùng loại sản phẩm A tại vùng đó, người ta phỏng vấn ngẫu nhiên 100 gia đình và thấy có 65 gia đình có nhu cầu về loại sản phẩm trên. Với độ tin cậy 95% hãy ước lượng số gia đình có nhu cầu về loại sản phẩm đó bằng khoảng tin cậy đối xứng

Bài 3.5. Theo dõi điểm thi môn Tin học của 100 học sinh của một trường nghề thu được số liệu sau:

Điểm	1	2	3	4	5	6	7	8	9	10
Số SV	2	3	4	12	18	25	20	10	5	1

Hãy xác định khoảng tin cậy đối xứng với độ tin cậy 95% cho điểm trung bình của học sinh trường đó. Biết điểm thi môn Tin học của học sinh trường đó là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn.

Bài 3.6. Kiểm tra hàm lượng vitamin C trong một loại quả thu được số liệu sau:

Hàm lượng vitamin C (%)	Số trái
6-7	5
7-8	10
8-9	20
9-10	35
10-11	25
11-12	5

Với độ tin cậy 95%, hãy ước lượng hàm lượng Vitamin C trung bình trong loại quả trên bằng khoảng tin cậy đối xứng. Biết hàm lượng vitamin C trong loại quả đó là đại lượng ngẫu nhiên có phân phối chuẩn.

Bài 3.7. Điều tra năng suất giống lúa mới tại một khu thử nghiệm người ta thu hoạch 25 ha và có kết quả như sau:

Năng suất (Tạ/ha)	18	20	21	24	27
Số ha	2	5	6	8	4

Với độ tin cậy 95% , hãy ước lượng năng suất trung bình của giống lúa đó biết rằng năng suất giống lúa đó là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn với độ lệch tiêu chuẩn là 2 tạ/ha.

Bài 3.8. Theo dõi mức xăng hao phí cho một loại ô tô trên đoạn đường từ A đến B thu được số liệu sau:

Mức xăng hao phí	19,0 - 19,5	19,5 - 20,0	20,0 - 20,5	20,5 - 21,0
Số lượng xe	2	10	8	5

Với độ tin cậy 95%, hãy xác định mức xăng hao phí trung bình tối thiểu (tối đa) của loại xe trên khi đi từ A đến B biết mức xăng hao phí là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn.

Bài 3.9. Theo dõi trọng lượng của một máy tự động, người ta cân thử 100 gói và thu được số liệu

Trọng lượng (gam)	10	11	13	15	20
Số gói	25	5	35	20	15

Hãy ước lượng cho trọng lượng trung bình của loại sản phẩm trên bằng khoảng tin cậy đối xứng với độ tin cậy 95% biết trọng lượng của sản phẩm là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn.

Bài 3.10. Trước khi quyết định đầu tư cho một dự án phát triển phần mềm của một công ty máy tính, người ta tiến hành thăm dò ý kiến của 1800 người thì thấy có 1180 người ủng hộ đầu tư cho dự án. Với độ tin cậy 99%, hỏi sẽ có tối thiểu bao nhiêu phần trăm người ủng hộ ? và tối đa là bao nhiêu?

Bài 3.11. Điều tra 200 sinh viên khoa CNTT sử dụng thời gian để chơi

Game online trong một ngày được kết quả như sau:

Thời gian (phút)	Số sinh viên
190 - 200	10
200 - 210	26
210 - 220	56
220 - 230	64
230 - 240	30
240 - 250	14

a. Hãy ước lượng thời gian trung bình sinh viên đã sử dụng vào việc chơi Game online với độ tin cậy 95%.

b. Với độ tin cậy 0,95 hãy ước lượng tỷ lệ sinh viên chơi Game online nhiều hơn 220 phút một ngày.

Bài 3.12. Cân thử 100 quả trứng ta có kết quả sau

X_i (g)	150	160	165	170	180	185
n_i	4	16	25	30	15	10

a. Tìm khoảng ước lượng cho khối lượng trung bình của trứng với độ tin cậy 99%.

b. Trứng có khối lượng nhỏ hơn 165 gam được coi là trứng loại 2. Tìm khoảng ước lượng cho tỷ lệ trứng loại 2 với độ tin cậy 95%.

Bài 3.13. Số liệu sau đây là chỉ số thông minh IQ của 28 sinh viên một trường đại học (giả sử chỉ số đó tuân theo luật phân phối chuẩn):

130	135	121	123	121	117	135
117	121	123	117	130	112	112
108	121	117	112	117	121	123
117	108	121	121	112	123	130

Xây dựng khoảng tin cậy 95% cho kỳ vọng của chỉ số trên. Có những cách nào để tăng độ chính xác của ước lượng?

Bài 3.14. Kết quả đo chiều cao của 24 trẻ em 2 tuổi cho bởi bảng sau (cm):

84,4	89,9	89,0	81,9	87,0	78,5
84,1	86,3	80,6	80,0	81,3	86,8
83,4	89,4	85,4	80,6	85,0	82,5
80,7	84,3	85,4	85,0	85,5	81,6

Với độ tin cậy 99% cho biết khoảng tin cậy của chiều cao trung bình của trẻ em 2 tuổi? Giả thiết chiều cao của trẻ em 2 tuổi có phân phối chuẩn.

Bài 3.15. Điều tra tỷ lệ sinh viên khoa CNTT sử dụng Internet, người ta phỏng vấn ngẫu nhiên 400 sinh viên thì thấy có 368 sinh viên sử dụng Internet. Hỏi với mức ý nghĩa $\alpha = 5\%$ có bao nhiêu phần trăm sinh viên khoa CNTT sử dụng Internet.

Bài 3.16. Trước cuộc bầu cử tổng thống người ta phỏng vấn ngẫu nhiên 1800 cử tri thì thấy có 1220 người ủng hộ một ứng cử viên A. Với độ tin cậy 95% hỏi ứng cử viên đó thu được bao nhiêu phần trăm số phiếu?

Bài 3.17. Để xác định chất lượng vải của một nhà máy dệt, người ta lấy ra 400mét vải của các phân xưởng thì thấy 280mét vải đạt loại 1. Xác định khoảng tin cậy khi ước lượng tỷ lệ vải đạt loại 1 của toàn nhà máy với độ tin cậy 90%.

Bài 3.18. Để xác định chiều cao trung bình của các cây bạch đàn trong một vườn ươm, người ta tiến hành đo ngẫu nhiên 35 cây. Kết quả được biểu thị bởi bảng sau:

Chiều cao(cm)	6,5	7	7,5	8	8,5	9
Số cây	2	4	10	11	5	3

Giả thiết rằng chiều cao của các cây bạch đàn có phân phối chuẩn. Hãy ước lượng chiều cao trung bình của các cây bạch đàn trong vườn trên với độ tin cậy 95%.

Bài 3.19. Để xác định lượng xăng trung bình trung bình tiêu phí (lít/100km) của một loại ôtô người ta kiểm tra 36 chiếc trên đoạn đường 100km và tính được kỳ vọng mẫu 5,38 lít với phương sai mẫu hiệu chỉnh là 0,285. Với độ tin cậy 95% hãy ước lượng mức hao phí trung bình của loại xe này.

Bài 3.20. Kết quả thi môn xác suất thống kê của lớp 30 sinh viên được cho như sau:

8	9	7	6	8	5	10	8	8	5
10	8	4	5	7	4	9	6	3	10
10	9	4	7	6	8	9	6	10	3

Với độ tin cậy 90% hãy ước lượng cho điểm trung bình môn XSTK của lớp đạt được. Biết điểm thi có phân phối chuẩn.

Bài 3.21. Điều tra doanh số hàng tháng của 100 hộ kinh doanh một ngành nào đó ta thu được số liệu sau (triệu đồng):

Doanh số X	10,1	10,2	10,4	10,5	10,7	10,8	10,9	11	11,3	11,4
Số hộ	1	3	8	13	25	20	12	10	6	2

a. Với độ tin cậy 95% thì doanh số trung bình của các hộ kinh doanh tối thiểu (tối đa) là bao nhiêu? Biết doanh số có phân phối chuẩn với phương sai $\sigma^2 = 0,01$

b. Với độ tin cậy 95% hãy ước lượng cho tỷ lệ các hộ kinh doanh có doanh số lớn hơn 10,8 triệu đồng?

Bài 3.22. Điều tra 365 điểm trồng lúa của một huyện ta thu được số liệu sau:

Năng suất X (tạ/ha)	25	30	33	34	35	36	37	39	40
Số điểm trồng	6	13	38	74	106	85	30	10	3

Với độ tin cậy 95% hãy ước lượng cho năng suất trung bình của huyện đó. Biết năng suất có phân phối chuẩn và có phương sai $\sigma^2 = 4$.

Bài 3.23. Trong cuộc điều tra Glucoza trong máu ở 100 người ta thu được kết quả như sau (mg%)

Khoảng Glucoza	Số người	Khoảng Glucoza	Số người
65-70	1	100-105	17
70-75	0	105-110	16
75-80	2	110-115	9
80-85	5	115-120	5
85-90	8	120-125	2
90-95	16	125-130	1
95-100	18		

a. Với độ tin cậy 95% hãy ước lượng cho lượng Glucoza trung bình trong máu của một người. Biết lượng Glucoza có phân phối chuẩn với $\sigma = 10$.

b. Với độ tin cậy 95%, hãy ước lượng cho tỷ lệ người có lượng Glucoza trung bình trong máu trong khoảng từ 90 đến 115?

Bài 3.24. Một tổ gồm 50 công nhân xây dựng trong một tuần thực hiện lương khoán theo sản phẩm được cho như sau (đơn vị: nghìn đồng)

Khoảng	Số người	Khoảng	Số người
475-525	2	725-775	7
525-575	2	775-825	4
575-625	2	825-875	7
625-675	4	875-925	9
675-725	5	925-975	8

Với độ tin cậy 95% hãy ước lượng cho thu nhập trung bình của công nhân trong tổ. Biết thu nhập của công nhân có phân phối chuẩn với độ lệch tiêu chuẩn $\sigma = 10$.

Bài 3.25. Điều tra doanh số hàng tháng của 100 hộ kinh doanh của một loại hàng, ta có bảng số liệu sau (triệu đồng)

Doanh số (Triệu đồng)	11,5	11,6	11,7	11,8	11,9	12,0
Số hộ tương ứng	10	15	20	30	15	10

a. Với độ tin cậy 95% hãy ước lượng cho doanh số trung bình của các hộ kinh doanh.

b. Với độ tin cậy 95% tối thiểu (tối đa) có bao nhiêu phần trăm các hộ kinh doanh có doanh số nhỏ hơn 11,8 triệu?

Bài 3.26. Để xác định giá trung bình đối với một loại hàng hóa trên thị trường, người ta điều tra ngẫu nhiên tại 100 cửa hàng thu được bảng số liệu:

Giá X (đồng)	83	85	87	89	91	93	95	97	99	101
Số cửa hàng	6	7	12	15	30	10	8	6	4	2

a. Với độ tin cậy 95% hãy ước lượng giá trung bình của loại hàng đó tại thời điểm đang xét.

b. Với độ tin cậy 95% hãy ước lượng cho tỷ lệ cửa hàng có giá lớn hơn 90 đồng.

Bài 3.27. Công ty may vừa chuẩn bị cho xuất xưởng một số lượng lớn áo jacket. Lấy ngẫu nhiên ra 36 chiếc, kiểm tra đếm số khuyết tật nhỏ trên các áo này. Kết quả ghi ở bảng sau (mỗi số của bảng là số khuyết tật trên áo tương ứng).

0	4	2	2	3	0	5	1	2	4	3	3
0	4	5	5	0	1	2	6	0	1	3	3
0	2	4	5	4	1	3	3	0	4	4	3

Với độ tin cậy 95% hãy ước lượng cho số khuyết tật trung bình trên mỗi áo.

Bài 3.28. Số liệu thống kê về doanh số (triệu đồng/ngày) bán của một siêu thị trong một số ngày như sau:

Doanh số	24	30	36	42	48	54	60	65	70
Số ngày	5	12	25	35	24	15	12	10	6

Hãy ước lượng doanh số bán trung bình trong 1 ngày của siêu thị với độ tin cậy 95%.

Bài 3.29. Cho X là năng suất lúa ở một khu vực (tạ/ha). Điều tra ở một số thửa ruộng ta có:

X	30-35	35-40	40-45	45-50	50-55
N (Số thửa ruộng)	8	18	28	40	16

Với độ tin cậy 96%, hãy ước lượng năng suất lúa trung bình của toàn vùng?

Bài 3.30. Kết quả thi môn xác suất thống kê của lớp 27 sinh viên được cho như sau:

9	7	6	8	5	10	8	8	5
8	4	5	7	4	9	6	3	10
9	4	7	6	8	9	6	10	3

Với độ tin cậy 95% hãy ước lượng cho điểm trung bình môn Xác suất thống kê. Biết điểm thi có phân phối chuẩn.

Bài 3.31. Nghiên cứu trọng lượng trẻ sơ sinh của nhóm với mẹ không hút thuốc lá, ta có kết quả (kg) như sau:

3,99	3,79	3,60	3,21	3,60	4,08	3,61
3,83	3,31	4,13	3,26	2,72	3,54	3,51

Giả sử rằng trọng lượng trẻ có phân phối chuẩn. Với độ tin cậy 95%, hãy ước lượng trọng lượng trung bình của trẻ sơ sinh ở nhóm với mẹ không hút thuốc lá.

Bài 3.32. Để kiểm tra tư duy Toán học của sinh viên Khoa CNTT, người ta tiến hành chọn ngẫu nhiên 16 sinh viên khoa CNTT làm một bài thi trắc nghiệm, được kết quả điểm như sau:

400	379	360	373	329	373	320	360
385	361	383	325	413	326	354	351

Hãy ước lượng điểm trung bình của sinh viên khoa CNTT với mức ý nghĩa 5%. Biết điểm thi có phân phối chuẩn.

Bài 3.33. Để thử nghiệm hiệu quả của một loại thuốc trừ sâu người ta áp dụng thử loại thuốc này đối với 5 thửa ruộng đang bị sâu phá hoại. Số lượng sâu bắt được sau khi dùng loại thuốc trừ sâu nói trên được cho:

Thửa ruộng	1	2	3	4	5
Số lượng sâu	107	72	82	100	92

Với độ tin cậy 95% hãy ước lượng số lượng sâu trung bình sau khi phun thuốc, biết số lượng sâu có phân phối chuẩn.

Bài 3.34. Khoa Công nghệ thông tin muốn đánh giá số giờ tự học của sinh viên trong tuần. Phòng Đào tạo chọn ngẫu nhiên 25 sinh viên và nhận được kết quả sau:

9	8	7	6	7	8	9
4	7	6	6	4	11	
5	4	3	7	8	8	
7	6	2	2	8	6	

Với độ tin cậy 95% , hãy ước lượng số giờ tự học trung bình của sinh viên trong tuần. Giả thiết số giờ tự học trong tuần tuân theo qui luật phân phối chuẩn.

Chương 4
KIỂM ĐỊNH GIẢ THIẾT THỐNG KÊ

Ở chương 3 ta đã nghiên cứu các tham số đặc trưng của tổng thể trên cơ sở thông tin của mẫu bằng phương pháp ước lượng. Chương này tiếp tục nghiên cứu dấu hiệu của tổng thể bằng một phương pháp khác là kiểm định giả thiết thống kê. Đây là phương pháp quan trọng cho phép ta giải quyết nhiều bài toán trong thực tế. Nội dung của phương pháp kiểm định giả thiết thống kê là dựa vào mẫu cụ thể và các quy tắc hay thủ tục quyết định dẫn đến bác bỏ hay chấp nhận giả thiết của tổng thể.

4.1. Bài toán kiểm định giả thiết thống kê

4.1.1. Giả thiết thống kê

Chương này ta nghiên cứu đại lượng ngẫu nhiên trong trường hợp thông tin không đầy đủ thể hiện ở nhiều mặt, cụ thể là:

- Chưa biết chính xác tham số θ hoặc qui luật phân phối xác suất của đại lượng ngẫu nhiên X, nhưng có cơ sở nào đó để nêu lên giả thiết, chẳng hạn: $\theta = \theta_0$ (θ_0 là hằng số đã biết), hay X tuân theo qui luật chuẩn.

- Khi nghiên cứu hai hay nhiều đại lượng ngẫu nhiên, một trong những vấn đề cần quan tâm nhất là: Các đại lượng này độc lập với nhau hay có sự phụ thuộc tương quan? Hơn nữa các tham số của chúng có bằng nhau hay không? Những câu hỏi này thường chưa được trả lời khẳng định mà mới chỉ nêu lên như một giả thiết.

Định nghĩa 4.1. *Bất kỳ giả thiết nào nói về tham số, dạng qui luật phân phối hoặc tính độc lập của đại lượng ngẫu nhiên, đều gọi là giả thiết thống* *kê. Những giả thiết đó có thể đúng hoặc cũng có thể sai. Việc xác định tính đúng sai của một giả thiết được gọi là kiểm định giả thiết thống kê.*

Trong khuôn khổ của giáo trình ta chỉ đề cập đến các giả thiết về tham số của đại lượng ngẫu nhiên.

Giả sử cần nghiên cứu tham số θ của đại lượng ngẫu nhiên X, người ta đưa ra giả thiết cần kiểm định: $H : \theta = \theta_0$ (H: Hypothes). Gọi $\overline{H}$ là giả thiết đối của H thì $\overline{H} : \theta \neq \theta_0$.

4.1.2. Mức ý nghĩa, miền bác bỏ

Từ mẫu ngẫu nhiên $W_X = (X_1, X_2, ...X_n)$ ta chọn thống kê $G = f(X_1, X_2, ...X_n)$ sao cho nếu H đúng thì G có phân phối xác suất hoàn toàn xác định và với mẫu cụ thể thì giá trị của G sẽ tính được. Khi đó G được gọi là *tiêu chuẩn kiểm định giả thiết* H.

Với α bé tuỳ ý cho trước (thường $\alpha \in (0,01; 0,05)$) ta tìm được miền W_α sao cho $P(G \in W_\alpha) = \alpha$. Khi đó W_α được gọi là *miền bác bỏ,* α được gọi là *mức ý nghĩa của kiểm định.*

- Thực hiện phép thử đối với mẫu ngẫu nhiên $W_X = (X_1, X_2, ...X_n)$ ta được mẫu cụ thể. Tính giá trị của G tại $w_X = (x_1, x_2, ...x_n)$ ta được $\theta_0 = f(x_1, x_2, ...x_n), \theta_0$ 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:

$$\begin{cases} H : \theta = \theta_0 \\ \overline{H} : \theta > \theta_0 \end{cases}$$

Khi đó ta có kiểm định một phía.

4.1.3. Sai lầm loại 1 và sai lầm loại 2

Khi kiểm định giả thiết thống kê, ta có thể mắc phải một trong hai loại sai lầm sau:

- Sai lầm loại 1: Là sai lầm mắc phải khi ta bác bỏ một giả thiết H trong khi H lại đúng. Xác suất mắc phải sai lầm loại 1 bằng $P(G \in W_\alpha) = \alpha$.

- 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 $P(G \notin W_\alpha)$.

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 G 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.

4.2. Các trường hợp kiểm định giả thiết thống kê

4.2.1. So sánh trung bình mẫu với trung bình lý thuyết

Giả sử ĐLNN X có kỳ vọng chưa biết. Có 3 bài toán kiểm định:

2.1.1 Trường hợp 1: $D(X) = \sigma^2$ đã biết

Giả thiết : X có phân phối chuẩn hoặc cỡ mẫu n đủ lớn ($n \geq 30$). Chọn thống kê $U = \dfrac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma}$ làm tiêu chuẩn kiểm định. Nếu giả thiết H đúng thì $U \in N(0,1)$.

a. Bài toán (1): Với mức ý nghĩa α cho trước, xác định phân vị chuẩn $u_{1-\frac{\alpha}{2}}$. Ta tìm được miền bác bỏ

$$W_\alpha = \left\{u : |u| \geq u_{1-\frac{\alpha}{2}}\right\} = \left(-\infty; -u_{1-\frac{\alpha}{2}}\right] \cup \left[u_{1-\frac{\alpha}{2}}; +\infty\right).$$

$$\begin{aligned}P(U \in W_\alpha) &= P(U < -u_{1-\frac{\alpha}{2}}) + P(U > u_{1-\frac{\alpha}{2}})\\ &= \mathrm{P(U < u_{1-\frac{\alpha}{2}})} + 1 - P(U > u_{1-\frac{\alpha}{2}})\\ &= \frac{\alpha}{2} + 1 - (1 - \frac{\alpha}{2}) = \alpha.\end{aligned}$$

Lấy mẫu cụ thể và tính giá trị quan sát $u_{qs} = \dfrac{|\overline{x} - \mu_0|}{\sigma}\sqrt{n}$. So sánh u_{qs} và $u_{1-\frac{\alpha}{2}}$:

-Nếu $|u_{qs}| \geq u_{1-\frac{\alpha}{2}}$ ($u_{qs} \in W_\alpha$) thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.

-Nếu $|u_{qs}| < u_{1-\frac{\alpha}{2}}$ ($u_{qs} \notin W_\alpha$) thì chấp nhận H.

b. Bài toán (2) : $(\overline{H} : \mu > \mu_0)$. Với mức ý nghĩa α cho trước, xác định phân vị chuẩn $u_{1-\alpha}$. Ta tìm được miền bác bỏ

$$W_\alpha = \{u : u \geq u_{1-\alpha}\} = [u_{1-\alpha}; +\infty).$$

c. Bài toán (3):$\left(\overline{H} : \mu < \mu_0\right)$. Với mức ý nghĩa α cho trước, xác định phân vị chuẩn $u_{1-\alpha}$. Ta tìm được miền bác bỏ

$$W_\alpha = \{u : u \leq -u_{1-\alpha}\} = (-\infty; -u_{1-\alpha}]\,.$$

Ví dụ 4.2. Một tín hiệu được gửi từ địa điểm A và được nhận ở địa điểm B có phân phối chuẩn với trung bình μ và độ lệch tiêu chuẩn $\sigma = 2$. Tin rằng giá trị của tín hiệu $\mu = 8$ được gửi mỗi ngày. Người ta tiến hành kiểm tra giả thiết này bằng cách gửi 5 tín hiệu một cách độc lập trong ngày thì thấy giá trị trung bình nhận được tại địa điểm B là $\overline{x} = 9,5$. Với độ tin cậy 95%, hãy kiểm tra giả thiết $\mu = 8$ đúng hay không?

Giải. Bài toán kiểm định giả thiết thống kê với cặp giả thiết:

$$\begin{cases} H : \mu = 8;\\ \overline{H} :\ \mu \neq 8.\end{cases}$$

Ta có $n = 5 < 30$, độ tin cậy $1 - \alpha = 0,95; 1 - \dfrac{\alpha}{2} = 0,975.$

Phân vị chuẩn : $u_{0,975} = 1,96.$

Miền bác bỏ là : $(-\infty; -1,96] \cup [1,96; +\infty)$.

Giá trị quan sát $u_{qs} = \dfrac{|\overline{x} - \mu_0|}{\sigma}\sqrt{n} = \dfrac{9,5-8}{2}\sqrt{5} = 1,68$.

Ta thấy $u_{qs} \notin W_\alpha$ nên giả thiết $\mu = 8$ được chấp nhận.

2.1.2 Trường hợp 2 : $D(X) = \sigma^2$ chưa biết, giả thiết X có phân phối chuẩn.

Chọn thống kê $T = \dfrac{(\overline{X} - \mu_0)}{S'}\sqrt{n}$. Nếu H đúng thì $T \in T(n-1)$.

a. Bài toán (1): $(\overline{H} : \mu \neq \mu_0)$. Với mức ý nghĩa α cho trước, ta xác định phân vị Student (n-1) bậc tự do mức $1 - \dfrac{\alpha}{2}$.

Khi đó miền bác bỏ là:

$$W_\alpha = \left\{t : |t| \geq t^{(n-1)}_{1-\frac{\alpha}{2}}\right\} = \left(-\infty; -t^{(n-1)}_{1-\frac{\alpha}{2}}\right] \cup \left[t^{(n-1)}_{1-\frac{\alpha}{2}}; +\infty\right).$$

Lấy mẫu cụ thể và tính giá trị quan sát : $t_{qs} = \dfrac{|\overline{x} - \mu_0|}{s'}\sqrt{n}$.

So sánh t_{qs} và $t^{(n-1)}_{1-\frac{\alpha}{2}}$:

- Nếu $|t_{qs}| \geq t^{(n-1)}_{1-\frac{\alpha}{2}}$ ($t_{qs} \in W_\alpha$) thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.
- Nếu $|t_{qs}| < t^{(n-1)}_{1-\frac{\alpha}{2}}$ ($t_{qs} \notin W_\alpha$) thì chấp nhận H.

b. Bài toán (2):$(\overline{H} : \mu > \mu_0)$. Với mức ý nghĩa α cho trước, ta xác định phân vị Student (n-1) bậc tự do mức $1 - \alpha$.

Khi đó miền bác bỏ là: $W_\alpha = \left\{t : t \geq t^{(n-1)}_{1-\alpha}\right\} = \left[t^{(n-1)}_{1-\alpha}; +\infty\right)$.

Lấy mẫu cụ thể và tính giá trị quan sát : $t_{qs} = \dfrac{\overline{x} - \mu_0}{s'}\sqrt{n}$.

So sánh t_{qs} và $t^{(n-1)}_{1-\alpha}$:

- Nếu $t_{qs} \geq t^{(n-1)}_{1-\alpha}$ ($t_{qs} \in W_\alpha$) thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.
- Nếu $t_{qs} < t^{(n-1)}_{1-\alpha}$ ($t_{qs} \notin W_\alpha$) thì chấp nhận H.

c. Bài toán (3) : Với mức ý nghĩa α cho trước, ta xác định phân vị Student (n-1) bậc tự do mức $1 - \alpha$.

Khi đó miền bác bỏ là: $W_\alpha = \left\{t : t \leq -t^{(n-1)}_{1-\alpha}\right\} = (-\infty; -t^{(n-1)}_{1-\alpha}]$.

Lấy mẫu cụ thể và tính giá trị quan sát: $t_{qs} = \dfrac{\overline{x} - \mu_0}{s'}\sqrt{n}$.

So sánh t_{qs} và $-t^{(n-1)}_{1-\alpha}$:

- Nếu $t_{qs} \leq -t^{(n-1)}_{1-\alpha}$ ($t_{qs} \in W_\alpha$) thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.

- Nếu $t_{qs} > -t_{1-\alpha}^{(n-1)}$ ($t_{qs} \notin W_\alpha$) thì chấp nhận H.

Ví dụ 4.3. 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 của các bao gạo có sự 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*
49,0 - 49,5	*10*
49,5 - 50,0	*6*
50,0-50,5	*2*

*V*ới độ tin cậy 99%, hãy kết luận về điều nghi ngờ nói trên.

Giải. Bài toán kiểm định giả thiết thống kê với cặp giả thiết:

$$\begin{cases} H : \mu = 50; \\ \overline{H} : \mu \neq 50. \end{cases}$$

$T = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S'} \in T(24)$.

Ta có $1 - \alpha = 0,99 \Rightarrow 1 - \dfrac{\alpha}{2} = 0,995 \Rightarrow t_{0,995}^{24} = 2,797$.

Miền bác bỏ là: $W_\alpha = (-\infty; -2,797] \cup [2,797; +\infty)$.

$$\overline{x} = \frac{1231,75}{25} = 49,27.$$
$$s^2 = \frac{60695,06}{25} - (49,27)^2 = 2427,8 - 2427,53 = 0,27.$$
$$s'^2 = \frac{25}{24}0,27 = 0,2812 \Rightarrow s' = 0,53.$$

Giá trị quan sát $t_{qs} = \dfrac{|49,27 - 50|\sqrt{25}}{0,53} = 6,886$.

Ta thấy $t_{qs} \notin W_\alpha$, nên giả thiết H bị bác bỏ, chấp nhận $\overline{H}$. Vậy điều nghi ngờ là đúng.

Ví dụ 4.4. Để xác định trọng lượng trung bình của các bao bột mì ta lấy ngẫu nhiên 17 bao đem cân được kết quả: Trọng lượng trung bình $\overline{x} = 3,98$kg, $s'^2 = 0,16$. Hãy kiểm định giả thiết: $H : \mu = 40; \overline{H} : \mu \neq 40$ với mức ý nghĩa $\alpha = 0,05$.

Giải. Đây là bài toán kiểm định giả thiết của kỳ vọng trong phân phối chuẩn, trường hợp chưa biết phương sai, mẫu cỡ $n = 17 < 30$ với cặp giả thiết

$$\begin{cases} H : \mu = 40; \\ \overline{H} : \mu \neq 40. \end{cases}$$

Lập mẫu ngẫu nhiên $W_X = (X_1, X_2, \ldots, X_n)$ cỡ $n = 17$.

Chọn thống kê $T = \dfrac{(\overline{X} - \mu_0)}{S'}\sqrt{n}$ làm tiêu chuẩn kiểm định.

Xây dựng miền bác bỏ W_α. Vì T có phân phối Student $(n-1)$ bậc tự do nên với mức ý nghĩa $\alpha = 0,05$, tra bảng Student có $t_{1-\alpha}^{(n-1)} = t_{0,95}^{16} = 2,12$.

Vậy $W_\alpha = (-\infty; -2,12] \cup [2,12; +\infty)$.

Tính $t_{qs} = \dfrac{(39,8 - 40)}{0,4}\sqrt{17} = -2,06$.

Rõ ràng $t_{qs} \notin \mathrm{W}_\alpha$ nên chấp nhận giả thiết H. Nghĩa là trọng lượng bình quân của những bao bột mì là 40kg.

2.1.3 Trường hợp 3: $D(X) = \sigma^2$ **chưa biết và** $n \geq 30$.

Trong trường hợp này ta vẫn chọn thống kê như trên trong đó độ lệch tiêu chuẩn σ được ước lượng bằng độ lệch tiêu chuẩn hiệu chỉnh của mẫu ngẫu nhiên S'. Tiêu chuẩn kiểm định $U = \dfrac{\left(\overline{X} - \mu_0\right)\sqrt{n}}{S'}$.

Nếu H đúng thì $U \in N(0,1)$. Ta có miền bác bỏ của bài toán 1 là

$$W_\alpha = \left\{u : |u| > u_{1-\frac{\alpha}{2}}\right\} = \left(-\infty; -u_{1-\frac{\alpha}{2}}\right) \cup \left(u_{1-\frac{\alpha}{2}}; +\infty\right).$$

Lấy mẫu cụ thể và ta tính giá trị quan sát $u_{qs} = \dfrac{|\overline{x} - \mu_0|}{s'}\sqrt{n}$.

So sánh t_{qs} và $u_{1-\frac{\alpha}{2}}$

- Nếu $|t_{qs}| \geq t_{1-\frac{\alpha}{2}}^{(n-1)}$ $(t_{qs} \in W_\alpha)$ thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.

- Nếu $|t_{qs}| < t^{(n-1)}_{1-\frac{\alpha}{2}}$ ($t_{qs} \notin W_\alpha$) thì chấp nhận H.

Tương tự như trên ta có miền bác bỏ của bài toán 2:

$$W_\alpha = \{u : u \geq u_{1-\alpha}\} = [u_{1-\alpha}; +\infty) .$$

Tương tự như trên ta có miền bác bỏ của bài toán 3:

$$W_\alpha = \{u : u \leq -u_{1-\alpha}\} = (-\infty; -u_{1-\alpha}] .$$

Ví dụ 4.5. Một nhóm nghiên cứu tuyên bố: Trung bình một người vào siêu thị X tiêu nhiều hơn 140 ngàn đồng. Chọn một mẫu ngẫu nhiên gồm 50 người mua hàng, tính được số tiền trung bình họ tiêu là 158 ngàn đồng với độ lệch tiêu chuẩn điều chỉnh của mẫu là $s' = 62$. Với mức ý nghĩa $0,05$ hãy kiểm định xem tuyên bố của nhóm nghiên cứu có đúng hay không?

Giải. Bài toán kiểm định giả thiết thống kê với cặp giả thiết

$$\begin{cases} H : \mu = 140; \\ \overline{H} : \mu > 140. \end{cases}$$

Ta có $n = 50 > 30$ và $1 - \alpha = 0,95$. Ta có $u_{1-\alpha} = u_{0,95} = 1,65$.

Miền bác bỏ: $W_\alpha = [1,65; +\infty)$.

Giá trị quan sát $u_{qs} = \dfrac{|\overline{x} - \mu_0|}{s'}\sqrt{n} = \dfrac{158 - 140}{62}\sqrt{50} = 1,59$.

Ta thấy $u_{qs} \in \mathrm{W}_\alpha$ nên bác bỏ H, chấp nhận $\overline{H}$. Vậy báo cáo của nhóm nghiên cứu là đúng.

4.2.2. So sánh tỷ lệ mẫu với tỷ lệ lý thuyết

Xét phép thử C và biến cố A liên quan: $P(A) = p$, trong đó p là tham số chưa biết. Thông thường đưa ra 3 bài toán kiểm định giả thiết sau:

$$(1) \begin{cases} \mathrm{H} : \mathrm{p} = \mathrm{p}_0 \\ \overline{\mathrm{H}} : \mathrm{p} \neq \mathrm{p}_0 \end{cases} ; \ (2) \begin{cases} \mathrm{H} : \mathrm{p} = \mathrm{p}_0 \\ \overline{\mathrm{H}} : \mathrm{p} > \mathrm{p}_0 \end{cases} ; \ (3) \begin{cases} \mathrm{H} : \mathrm{p} = \mathrm{p}_0 \\ \overline{\mathrm{H}} : \mathrm{p} < \mathrm{p}_0 \end{cases}$$

Thực hiện n phép thử C trong đó biến cố A xuất hiện m lần. Tính tỷ lệ $f = \dfrac{m}{n}$ là ước lượng điểm cho tham số p.

Chọn thống kê làm tiêu chuẩn kiểm định. Nếu H đúng thì $U \in N(0,1)$.

Bài toán 1:

$$\begin{cases} \mathrm{H : p = p_0;} \\ \overline{\mathrm{H}}\mathrm{: p \neq p_0.} \end{cases}$$

Với mức ý nghĩa α cho trước, có miền bác bỏ của bài toán 1 là

$$W_\alpha = \left\{u : |u| \geq u_{1-\frac{\alpha}{2}}\right\} = \left(-\infty; -u_{1-\frac{\alpha}{2}}\right] \cup \left[u_{1-\frac{\alpha}{2}}; +\infty\right).$$

Như vậy:

- Lấy mẫu cụ thể và tính giá trị quan sát $u_{qs} = \dfrac{|f - p_0|}{\sqrt{p_0(1-p_0)}}.\sqrt{n}$.

- So sánh u_{qs} và $u_{1-\frac{\alpha}{2}}$:

• Nếu $|u_{qs}| \geq u_{1-\frac{\alpha}{2}}$ ($u_{qs} \in W_\alpha$) thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.

• Nếu $|u_{qs}| < u_{1-\frac{\alpha}{2}}$ ($u_{qs} \notin W_\alpha$) thì chấp nhận H.

Bài toán 2:

$$\begin{cases} \mathrm{H : p = p_0;} \\ \overline{\mathrm{H}}\mathrm{: p > p_0.} \end{cases}$$

Với mức ý nghĩa α cho trước, có miền bác bỏ của bài toán 2 là

$$W_\alpha = \{u : u \geq u_{1-\alpha}\} = [u_{1-\alpha}; +\infty).$$

- Lấy mẫu cụ thể và tính giá trị quan sát $u_{qs} = \dfrac{(f - p_0)}{\sqrt{p_0(1-p_0)}}.\sqrt{n}$

- So sánh u_{qs} và $u_{1-\alpha}$:

• Nếu $u_{qs} \geq u_{1-\alpha}$ ($u_{qs} \in W_\alpha$) thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.

• Nếu $u_{qs} < u_{1-\alpha}$ ($u_{qs} \notin W_\alpha$) thì chấp nhận H.

Bài toán 3:

$$\begin{cases} \mathrm{H : p = p_0;} \\ \overline{\mathrm{H}}\mathrm{: p < p_0.} \end{cases}$$

Với mức ý nghĩa α cho trước, có miền bác bỏ của bài toán 3 là

$$W_\alpha = \{u : u \leq -u_{1-\alpha}\} = (-\infty; -u_{1-\alpha}].$$

Như vậy:

- Lấy mẫu cụ thể và tính giá trị quan sát $u_{qs} = \dfrac{(f - p_0)}{\sqrt{p_0(1-p_0)}}.\sqrt{n}$.
- So sánh u_{qs} và $u_{1-\alpha}$:

• Nếu $u_{qs} \le -u_{1-\alpha}$ ($u_{qs} \in W_\alpha$) thì bác bỏ giả thiết H và chấp nhận $\overline{H}$.

• Nếu $u_{qs} > -u_{1-\alpha}$ ($u_{qs} \notin W_\alpha$) thì chấp nhận H.

Ví dụ 4.6. Tỷ lệ phế phẩm ở một nhà máy là 10%. Sau khi đã cải tiến kỹ thuật, điều tra 400 sản phẩm thì thấy 32 phế phẩm. Với độ tin cậy 99% hãy xét xem việc cải tiến kỹ thuật có làm giảm tỷ lệ phế phẩm hay không?

Giải. Đây là bài toán kiểm định giả thiết về tỷ lệ (hay xác suất), với mẫu cỡ $n = 400$ khá lớn. Gọi p là tỷ lệ phế phẩm của nhà máy. Ta kiểm định giả thiết thống kê với cặp giả thiết $\begin{cases} H : p = 0,1; \\ \overline{H} : p < 0,1. \end{cases}$

Tỷ lệ phế phẩm trong 400 sản phẩm là $f = \dfrac{32}{400} = 0,08$.

Độ tin cậy $1 - \alpha = 0,99$. Do đó $1 - \dfrac{\alpha}{2} = 0,995; u_{0,995} = 2,576$.

Miền bác bỏ là $W_\alpha = (-\infty; -2,576] \cup [2,576; +\infty)$.

Giá trị quan sát $u_{qs} = \dfrac{|0,08 - 0,1|\sqrt{400}}{\sqrt{0,1.0,9}} = 1,333 \notin W_\alpha$.

Do đó chấp nhận H. Vậy việc cải tiến có hiệu quả.

Ví dụ 4.7. Một nhà sản xuất thuốc chống dị ứng thực phẩm tuyên bố rằng 90% người dùng 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 suất có đúng hay không với mức ý nghĩa $\alpha = 0,01$.

Giải. Ta đưa ra bài toán kiểm định giả thiết thống kê $\begin{cases} H : p = 0,9; \\ \overline{H} : \ p < 0,9. \end{cases}$

Vì $\alpha = 0,01$ nên $1 - \alpha = 0,99, -u_{1-\alpha} = -2,326$.

$f = 160/200 = 0,8$.

$$u_{qs} = \frac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{0,8 - 0,9}{\sqrt{0,9.0,1}}\sqrt{200} = -4,75.$$

Ta thấy $u_{qs} < -u_{1-\alpha}$ nên bác bỏ giả thiết H. Vậy lời tuyên bố của nhà sản xuất là không đúng sự thật.

4.2.3. So sánh 2 giá trị trung bình

Giả sử ta có hai mẫu ngẫu nhiên: $W_X = (X_1, X_2, \dots)$ được rút ra từ ĐLNN X và $W_Y = (Y_1, Y_2, \dots)$ được rút ra từ ĐLNN Y. Bài toán tổng quát đặt ra: Ta muốn kiểm tra xem hai mẫu trên có phải được rút ra từ một phân phối hay không?

Trong bài này ta xét bài toán đơn giản hơn là so sánh 2 giá trị trung bình EX và EY.

Kí hiệu: $EX = \mu_1$; $EY = \mu_2$; $DX = \sigma_1^2$; $DY = \sigma_2^2$. Có thể đưa ra 2 bài toán kiểm định giả thiết:

$$(1)\begin{cases} \text{H}:\mu_1 = \mu_2; \\ \overline{\text{H}}:\mu_1 \neq \mu_2. \end{cases} ;\quad (2)\begin{cases} \text{H}:\mu_1 = \mu_2; \\ \overline{\text{H}}:\mu_1 > \mu_2. \end{cases}$$

2.3.1 Trường hợp 1: Biết σ_1^2, σ_2^2

Giả sử giả thiết H đúng. Khi đó $U = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \to N(0,1)$.

Như vậy: $P\left\{|u| \geq u_{1-\frac{\alpha}{2}}\right\} = \alpha, P\left\{u \geq u_{1-\alpha}\right\} = \alpha$.

Với mức ý nghĩa α cho trước, miền bác bỏ của bài toán 1 là

$$W_\alpha = \left\{u : |u| \geq u_{1-\frac{\alpha}{2}}\right\} = \left(-\infty; -u_{1-\frac{\alpha}{2}}\right] \cup \left[u_{1-\frac{\alpha}{2}}; +\infty\right).$$

Với mức ý nghĩa α cho trước, có miền bác bỏ của bài toán 2 là

$$W_\alpha = \{u : u \geq u_{1-\alpha}\} = [u_{1-\alpha}; +\infty).$$

2.3.2 Trường hợp 2: Chưa biết σ_1^2, σ_2^2

Trong trường hợp này ta phải giả thiết $X \sim N(\mu_1, \sigma_1^2)$; $Y \sim N(\mu_2, \sigma_2^2)$

trong đó $\sigma_1^2 = \sigma_2^2$.

Khi đó các bước thực hiện như sau:

- Từ mẫu đã cho tính $\overline{X}$, $\overline{Y}$, S_X^2, S_Y^2.
- Tính $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{n_1 S_X^2 + n_2 S_X^2}{n_1 + n_2 - 2}}.\sqrt{\dfrac{n_1 + n_2}{n_1.n_2}}}$.
- Với α đã cho tra bảng phân phối Student tìm $t_{1-\frac{\alpha}{2}}^{(n_1+n_2-2)}$ hoặc $t_{1-\alpha}^{(n_1+n_2-2)}$.

Với mức ý nghĩa α cho trước, miền bác bỏ của bài toán 1 là

$$W_\alpha = \left\{t : |t| \geq t_{1-\frac{\alpha}{2}}^{n_1+n_2-2}\right\} = \left(-\infty; -t_{1-\frac{\alpha}{2}}^{n_1+n_2-2}\right] \cup \left[t_{1-\frac{\alpha}{2}}^{n_1+n_2-2}; +\infty\right).$$

Với mức ý nghĩa α cho trước, có miền bác bỏ của bài toán 2 là

$$W_\alpha = \left\{t : t \geq t_{1-\alpha}^{(n_1+n_2-2)}\right\} = \left[t_{1-\alpha}^{(n_1+n_2-2)}; +\infty\right).$$

Nhận xét: Trong trường hợp σ_1, σ_2 chưa biết, ta phải giả thiết về tính chuẩn của hai mẫu và $\sigma_1 = \sigma_2$. Song nếu n_1, n_2 đủ lớn ($n_1 \geq 30$, $n_2 \geq 30$) thì mặc dù các giả thiết trên không thỏa mãn ta có thể xấp xỉ σ_1 bởi S_1', σ_2 bởi S_2'. Do đó ta có thể áp dụng như trường hợp đã biết $U = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}}}$.

Ví dụ 4.8. Trọng lượng sản phẩm do hai máy sản xuất ra đều là các đại lượng ngẫu nhiên tuân theo quy luật phân phối chuẩn và cùng có độ lệnh tiêu chuẩn là $\sigma = 1$kg. Với mức ý nghĩa $\alpha = 0,05$ có thể xem trọng lượng trung bình của sản phẩm do 2 máy sản xuất ra là như nhau hay không? Nếu cân 25 sản phẩm của máy I thấy trọng lượng của chúng là 1250kg, còn 20 sản phẩm của máy II có trọng lượng 1012kg.

Giải. Gọi trọng lượng sản xuất ra ở máy I là X; ở máy II là Y. Theo giả thiết X, Y có phân phối chuẩn với $D(X) = D(Y) = 1$. Đây là bài toán kiểm định giả thiết về 2 kỳ vọng trong phân phối chuẩn, trường hợp đã biết phương sai.

Ta kiểm tra giả thiết $H : E(X) = E(Y)(\overline{H} : E(X) \neq E(Y))$.

Từ mẫu cụ thể để tính u_{qs} ta phải tính $\overline{X}, \overline{Y}$.

Theo bài ra: $\overline{x} = \dfrac{1250}{25} = 50\ (kg);\ \overline{y} = \dfrac{1012}{20} = 50,6\ (kg)$.

Vậy: $u_{qs} = \dfrac{|50 - 50,6|}{\sqrt{\dfrac{1}{25} + \dfrac{1}{20}}} = 2$.

Với mức ý nghĩa $\alpha = 0,05$ thì $u_{1-\frac{\alpha}{2}} = 1,96$.

Ta thấy $u_{qs} \geq u_{1-\frac{\alpha}{2}}$ nên bác bỏ giả thiết H, tức là trọng lượng trung bình của sản phẩm sản xuất ở hai máy là khác nhau.

Ví dụ 4.9. Người ta cân trẻ sơ sinh ở hai khu vực thành trị và nông thôn thu được kết quả sau:

Khu vực	*Số trẻ cân*	*Trọng lượng trung bình*	*Phương sai hiệu chỉnh*
Nông thôn	*2500*	*3,0*	*200*
Thành thị	*500*	*3,1*	*5*

Với mức ý nghĩa $\alpha = 1\%$ có coi trọng lượng trung bình của trẻ sơ sinh ở 2 khu vực như nhau không?

Giải. Gọi X là trọng lượng trẻ sơ sinh ở nông thôn, Y là trọng lượng trẻ sơ sinh ở thành thị.

Đây là bài toán kiểm định giả thiết về 2 kỳ vọng, trường hợp chưa biết phương sai, mẫu cỡ $n = 2500, m = 500$ khá lớn.

Ta kiểm tra giả thiết $H : E(X) = E(Y)(\overline{H} : E(X) \neq E(Y))$.

Với mức ý nghĩa $\alpha = 0,01$ thì $u_{1-\frac{\alpha}{2}} = u_{0,995} = 2,58$.

Vậy $W_\alpha = (-\infty; -2,58] \cup [2,58; +\infty)$.

Từ mẫu cụ thể tính $u_{qs} = \dfrac{|3,0 - 3,1|}{\sqrt{\dfrac{200}{2500} + \dfrac{5}{500}}} = \dfrac{1}{3} = 0,33$.

Vậy $u_{qs} \notin W_\alpha$, chưa có cơ sở bác bỏ H. Ta có thể coi rằng trọng lượng của trẻ sơ sinh ở thành thị và nông thôn là như nhau.

4.2.4. So sánh 2 tỷ lệ

Giả sử ta có hai mẫu ngẫu nhiên: $(X_1,\ X_2, ...,\ X_{n_1})$ và $(Y_1,\ Y_2, ...,\ Y_{n_2})$, trong đó

$$P\{X_i = 1\} = p_1\ ;\ P\{X_i = 0\} = 1 - p_1 = q_1\ ;\ m_1 = \sum_{i=1}^{n_1} X_i;$$

$$P\{Y_i = 1\} = p_2\ ;\ P\{Y_i = 0\} = 1 - p_2 = q_2\ ;\ m_2 = \sum_{i=1}^{n_2} Y_i.$$

Có thể đưa ra các bài toán kiểm định giả thiết:

$$(1)\ \begin{cases} \mathrm{H} : \mathrm{p_1{=}p_2} \\ \overline{\mathrm{H}} : \mathrm{p_1 \neq p_2} \end{cases};\ (2)\ \begin{cases} \mathrm{H} : \mathrm{p_1{=}p_2} \\ \overline{\mathrm{H}} : \mathrm{p_1{>}p_2} \end{cases}.$$

Giả sử giả thiết H đúng, khi đó $DX_i = DY_j = \sigma^2$. Như vậy $D\left(\overline{X} - \overline{Y}\right) = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$. Để ước lượng phương sai chung σ^2 từ hai mẫu đã cho ta gộp lại thành một mẫu cỡ $n_1 + n_2$. Ta nhận được ước lượng cho σ^2 :

$$\frac{m_1 + m_2}{n_1 + n_2} \times \left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right).$$

Với n_1, n_2 đủ lớn ta xấp xỉ phân phối $\dfrac{\overline{X} - \overline{Y}}{\sqrt{D(\overline{X} - \overline{Y})}}$ bởi phân phối chuẩn $N(0, 1)$.

ĐLNN $U = \dfrac{\dfrac{m_1}{n_1} - \dfrac{m_2}{n_2}}{\sqrt{\dfrac{m_1 + m_2}{n_1 + n_2} \times \left(1 - \dfrac{m_1 + m_2}{n_1 + n_2}\right)}.\sqrt{\dfrac{n_1 + n_2}{n_1.n_2}}} \to N\left(0, 1\right).$

Với mức ý nghĩa α cho trước, miền bác bỏ của bài toán 1 (đối thiết $\overline{H}$: $p_1 \neq p_2$) là

$$W_\alpha = \{u : u \geq u_{1-\alpha}\} = [u_{1-\alpha}; +\infty).$$

Với mức ý nghĩa α cho trước, miền bác bỏ của bài toán 2 (đối thiết $\overline{H}$: $p_1 > p_2$) là

$$W_\alpha = \{u : u \geq u_{1-\alpha}\} = [u_{1-\alpha}; +\infty).$$

Ví dụ 4.10. Kiểm tra những sản phẩm được chọn ngẫu nhiên ở hai nhà máy sản xuất cùng 1 loại hàng ta có số liệu

Nhà máy	*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 do hai nhà máy sản xuất ra là như nhau không?

Giải. Gọi p_1, p_2 lần lượt là tỷ lệ phế phẩm của nhà máy I, II. Đây là bài toán kiểm định giả thiết về hai tỷ lệ, với mẫu cỡ lớn.

Ta kiểm tra giả thiết $H : p_1 = p_2; \overline{H} : p_1 \neq p_2$.

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: $n_1 = 100; n_2 = 120; m_1 = 20; m_2 = 36$. Do đó

$$u_{qs} = \frac{\dfrac{m_1}{n_1} - \dfrac{m_2}{n_2}}{\sqrt{\dfrac{m_1+m_2}{n_1+n_2} \times \left(1 - \dfrac{m_1+m_2}{n_1+n_2}\right)} . \sqrt{\dfrac{n_1+n_2}{n_1.n_2}}}$$

$$= \frac{\dfrac{20}{100} - \dfrac{36}{120}}{\sqrt{\dfrac{56}{220} \times \left(1 - \frac{56}{220}\right)} . \sqrt{\dfrac{220}{100.120}}} = -1.695.$$

Ta thấy $|u_{qs}| < u_{1-\frac{\alpha}{2}}$ nên chưa có cơ sở bác bỏ H. Tức là coi tỷ lệ phế phẩm do 2 nhà máy sản xuất ra là như nhau.

4.3. Bài tập chương 4

Bài 4.1. Kết quả thi môn xác suất thống kê của lớp 30 sinh viên được cho như sau:

8	9	7	6	8	5	10	8	8	5
10	8	4	5	7	4	9	6	3	10
10	9	4	7	6	8	9	6	10	3

Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: Điểm trung bình của môn XSTK là 7,5 điểm được không? Biết điểm thi có phân phối chuẩn.

Bài 4.2. Điều tra doanh số hàng tháng của 100 hộ kinh doanh một ngành nào đó ta thu được số liệu sau:

Doanh số (triệu đồng)	10,1	10,2	10,4	10,5	10,7	10,8	10,9	11,0	11,3	11,4
Số hộ	2	3	8	13	25	20	12	10	6	1

Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: Doanh thu trung bình của các hộ kinh doanh là 10,8 triệu đồng được không? Biết doanh thu có phân phối chuẩn và có phương sai $\sigma = 0,04$.

Bài 4.3. Điều tra 365 điểm trồng lúa của một huyện ta thu được số liệu sau:

Năng suất X (tạ/ha)	25	30	33	34	35	36	37	39	40
Số điểm trồng	6	14	38	74	106	84	30	10	3

Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: Năng suất trung bình của huyện đó là 35,5 tạ/ha được không? Biết năng suất có phân phối chuẩn.

Bài 4.4. Trong cuộc điều tra Glucoza trong máu ở 100 người ta thu được kết quả như sau (mg$\%$)

Khoảng Glucoza	Số người	Khoảng Glucoza	Số người
70-75	1	100-105	18
75-80	2	105-110	16
80-85	6	110-115	8
85-90	7	115-120	5
90-95	18	120-125	2
95-100	16	125-130	1

a.Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: Lượng Glucoza trung bình trong máu là 98 được không nếu biết lượng Glucoza có phân phối chuẩn với $\sigma = 10$.

b. Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: Lượng Glucoza trung bình trong máu là 98 được không?

Bài 4.5. Một tổ gồm 50 công nhân xây dựng trong một tuần thực hiện lương khoán theo sản phẩm được cho như sau (đơn vị: nghìn đồng)

Khoảng	Số người	Khoảng	Số người
47,5-52,5	2	72,5-77,5	7
52,5-57,5	2	77,5-82,5	4
57,5-62,5	2	82,5-87,5	8
62,5-67,5	5	87,5-92,5	9
67,5-72,5	4	92,5-97,5	7

Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: thu nhập trung bình của công nhân là 79 nghìn đồng được không?

Bài 4.6. Để xác định giá trung bình đối với một loại hàng hóa trên thị trường, người ta điều tra ngẫu nhiên tại 100 cửa hàng thu được bảng số liệu:

Giá X (đồng)	83	85	87	89	91	93	95	97	99	10
Số cửa hàng	6	7	12	15	30	10	8	6	4	2

a. Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: giá trung bình của loại hàng hóa đó là 92 đồng được hay không? Biết giá tiền có phân phối chuẩn.

b. Với độ tin cậy 95% hãy kiểm định ý kiến cho rằng tỷ lệ cửa hàng có giá lớn hơn 90 nghìn đồng là 70%.

Bài 4.7. Điều tra doanh số hàng tháng của 100 hộ kinh doanh của một loại hàng, ta có bảng số liệu sau (triệu đồng)

Doanh số (triệu đồng)	Số hộ tương ứng
11,5	10
11,6	15
11,7	20
11,8	30
11,9	15
12,0	10

a. Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: Doanh số trung bình của các hộ kinh doanh là 11,8 triệu được hay không? Biết doanh số có phân phối chuẩn.

b. Với độ tin cậy 95% hãy kiểm định lại ý kiến cho rằng: doanh số trung bình của các hộ kinh doanh lớn hơn 11,8 triệu đồng?

c. Với mức ý nghĩa 5% có thể chấp nhận ý kiến cho rằng: tỷ lệ các hộ kinh doanh có doanh số thấp hơn 11,8 triệu là 60% được hay không?

Bài 4.8. Tại một trạm phát, trong điều kiện bình thường trung bình có 25 tín hiệu được phát trong 1 phút. Sau cơn bão người ta nghi ngờ rằng tốc độ tín hiệu tại trạm phát đó đã thay đổi. Kiểm tra ngẫu nhiên 30 phút thấy trung bình tốc độ là 23,5 tín hiệu/phút với độ lệch tiêu chuẩn hiệu chỉnh của mẫu là 5. Với mức ý nghĩa 5%, hãy kết luận về điều nghi ngờ nói trên, biết số tín hiệu được phát trong một phút là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn.

Bài 4.9. Một hệ thống máy tính trong một công ty có thể xử lý 1300 hoá đơn trong 1 giờ. Công ty mới thay thế một số trang thiết bị của hệ thống và khi chạy kiểm tra trong 40 giờ cho thấy số hoá đơn xử lý trung bình trong 1 giờ là 1310 với độ lệch tiêu chuẩn 215. Với mức ý nghĩa $\alpha = 2,5\%$ hãy nhận định xem sau khi thay thế, hệ thống máy tính có hoạt động tốt hơn hay không?

Bài 4.10. Tỷ lệ cửa hàng ăn không đạt tiêu chuẩn tại vùng A theo báo cáo là 5%. Một đoàn kiểm tra khả sát ngẫu nhiên 400 cửa hàng thấy có 25 cửa hàng không đạt tiêu chuẩn. Đoàn kiểm tra cho rằng tỷ lệ cửa hàng ăn không đạt tiêu chuẩn tại vùng A cao hơn trong báo cáo. Với mức ý nghĩa 2.5%, hãy kết luận về nhận định của đoàn kiểm tra?

Bài 4.11. Nhịp mạch trung bình của một người là 72 lần/phút. Kiểm tra 64 người làm việc thường xuyên với máy tính thấy nhịp mạch trung bình của họ là 74 lần/phút với độ lệch tiêu chuẩn điều chỉnh của mẫu là 9 lần/phút. Với mức ý nghĩa 5%. hãy xét xem làm việc thường xuyên với máy tính có làm tăng nhịp mạch hay không?

Bài 4.12. Định mức thời gian để sản xuất ra một loại sản phẩm là 1 giờ. Sau cải tiến công nghệ người ta sản xuất thử 150 sản phẩm thì thấy thời gian trung bình để được một sản phẩm là 50 phút với độ lệch mẫu là 4 phút. Hỏi với độ tin cậy 95% có thể kết luận rằng công nghệ mới thực sự giúp làm giảm chi phí thời gian sản xuất hay không?

Bài 4.13. Công ty may vừa chuẩn bị cho xuất xưởng một số lượng lớn sản phẩm. Lấy ngẫu nhiên ra 36 chiếc, kiểm tra đếm số khuyết tật nhỏ trên các sản phẩm này. Kết quả ghi ở bảng sau (mỗi số của bảng là số khuyết tật trên sản phẩm tương ứng).

0	4	2	2	3	0	5	1	2	4	3	3
0	4	5	5	0	1	2	6	0	1	3	3
0	2	4	5	4	1	3	3	0	4	4	3

Phòng kiểm tra chất lượng nêu giả thuyết về số khuyết tật trung bình trên mỗi sản phẩm là: Giả thuyết $H : \mu = 2,2$ với độ tin cậy 95%. Hỏi rằng giả thuyết này có chấp nhận được hay không?

Bài 4.14. Điểm thi môn Tin (thang điểm 10) của 100 thí sinh (chọn ngẫu nhiên) trong kỳ thi vừa qua được cho trong bảng sau đây:

Điểm	1	2	3	4	5	6	7	8	9	10
Tần số	2	3	5	8	13	19	24	14	10	2

Với độ tin cậy 95% hỏi có thể kết luận điểm trung bình của môn Tin trong kỳ thi vừa qua là 6,5 điểm được hay không?

Bài 4.15. Giám đốc một công ty tuyên bố 90% động cơ của công ty đạt chuẩn quốc gia. Một hãng độc lập kiểm tra 400 động cơ của công ty thì thấy có 320 động cơ đạt yêu cầu. Với mức $\alpha = 0,05$ ta có thể kết luận gì về tuyên bố trên?

Bài 4.16. Một chủ trại gà tuyên bố: Lô gà này có trọng lượng trung bình là 1,8kg. Nghi ngờ tuyên bố này khách hàng bèn cân thử 25 con thì tính được trọng lượng trung bình 1,7 kg với phương sai mẫu là 0,01. Biết trọng lượng gà tuân theo luật phân phối chuẩn. Với mức ý nghĩa điều nghi ngờ trên có đúng không?

Bài 4.17. Kết quả thi môn xác suất thống kê của lớp 26 sinh viên được cho như sau:

$$\begin{array}{ccccccccc} 8 & 9 & 7 & 6 & 8 & 5 & 10 & 8 & 8 \\ 10 & 8 & 4 & 5 & 7 & 4 & 9 & 6 & 3 \\ 10 & 9 & 4 & 7 & 6 & 8 & 9 & 6 & \end{array}$$

Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: Điểm trung bình của lớp là 7 điểm được không? Biết điểm thi có phân phối chuẩn.

Bài 4.18. Điều tra 360 điểm trồng lúa của huyện A ta thu được số liệu sau:

Năng suất(tạ/ha)	25	30	33	34	35	36	37	39	40
Số điểm trồng	6	14	38	74	100	85	30	10	3

Với độ tin cậy 95% có thể chấp nhận ý kiến cho rằng: Năng suất trung bình của huyện đó là 35,5 tạ/ha được không? Biết năng suất có phân phối chuẩn và có phương sai $\sigma^2 = 25$.

Bài 4.19. Để xác định chiều cao trung bình của các cây bạch đàn trong một vườn ươm, người ta tiến hành đo ngẫu nhiên 35 cây. Kết quả được biểu thị bởi bảng sau:

Chiều cao (cm)	6,5	7,0	7,5	8	8,5	9,0
Số cây	2	4	10	11	5	3

Giả thiết rằng chiều cao của các cây bạch đàn có phân phối chuẩn dạng $N(\mu, \sigma)$. Người ta dự đoán chiều cao trung bình của các cây bạch đàn trong vườn ươm đó là 8 cm. Hãy kiểm định dự đoán trên ở mức ý nghĩa $\alpha = 0,05$.

Bài 4.20. Đo chỉ số mỡ sữa của 130 con bò lai ta được bảng số liệu sau:

Chỉ số	3,0-3,6	3,6-4,2	4,2-4,8	4,8-5,4	5,4-6,0	6,0-6,6	6,6-7,2
Số bò	2	8	35	43	22	15	5

Biết rằng chỉ số mỡ sữa trung bình của giống bò thuần chủng là 4,95. Với mức ý nghĩa 0,05 , hãy cho kết luận về hiệu quả của việc lai giống?

Bài 4.21. Để nghiên cứu tuổi thọ của một loại bóng đèn người ta thắp thử 100 bóng và thu được bảng số liệu sau:

Tuổi thọ (h)	Số bóng
1000-1050	5
1050-1100	25
1100-1150	40
1150-1200	15
1200-1250	10
1250-1300	5

Có ý kiến cho rằng tuổi thọ trung bình của loại bóng đèn này nhỏ hơn 1200 giờ. Với mức ý nghĩa 0,05, hãy kết luận về ý kiến trên. Biết rằng tuổi thọ của bóng đèn đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn

Bài 4.22. Một công ty dự định mở một siêu thị tại khu dân cư. Để đánh giá khả năng mua hàng của dân cư trong khu vực, người ta tiến hành điều tra thu nhập của 100 hộ dân trong khu vực và có bảng số liệu sau

Thu nhập bình quân (nghìn/người/tuần))	150	200	250	300	350
Số hộ	8	15	38	22	17

Theo bộ phận tiếp thị thì siêu thị chỉ hoạt động hiệu quả tại khu vực này khi thu nhập bình quân hàng tháng của các hộ tối thiểu là vào khoảng 250 nghìn/người/tuần. Vậy theo điều tra trên, công ty có nên mở siêu thị tại khu vực này không với mức ý nghĩa 5%?

Bài 4.23. Đo chiều cao của 100 nữ sinh trường đại học Công nghệ thông tin và truyền thông ta thu được bảng sau:

Khoảng chiều cao(m)	1,52-1,56	1,56-1,6	1,6-1,64	1,64-1,68	1,68-1,72
Số nữ sinh	20	30	18	7	15

Với mức ý nghĩa 5%, có thể kết luận chiều cao trung bình của nữ sinh trường Công nghệ thông tin và truyền thông là 1,62m được hay không?

Bài 4.24. Trọng lượng một loại sản phẩm của công ty kinh doanh đồ ăn khi sản xuất được quy định là 35g, độ lệch tiêu chuẩn là 5g. Nghi ngờ việc sản xuất không đáp ứng được yêu cầu đề ra, lấy ngẫu nhiên 100 sản phẩm kiểm tra và thu được trọng lượng bình quân là 34,5g. Hãy xét xem điều nghi ngờ trên có đúng không? Mức ý nghĩa 0,05. Biết trọng của mỗi sản phẩm tuân theo quy luật phân phối chuẩn.

Bài 4.25. Số liệu thống kê về doanh số (triệu đồng/ngày) bán của một siêu thị trong một số ngày như sau:

Doanh số	24	30	36	42	48	54	60	65	70
Số ngày	5	12	25	35	24	15	12	10	6

Trước đây doanh số bán trung bình của siêu thị là 35 triệu/ ngày. Với mức ý nghĩa 5% hãy cho nhận xét về phương thức bán hàng mới.

Bài 4.26. Khoa Công nghệ thông tin muốn đánh giá số giờ tự học của sinh viên trong tuần. Phòng Đào tạo chọn ngẫu nhiên 25 sinh viên và nhận được kết quả sau:

$$\begin{array}{ccccccccc} 9 & 8 & 7 & 6 & 7 & 8 & 9 & 4 & \\ 7 & 6 & 6 & 4 & 11 & 4 & 5 & 3 & \\ 7 & 8 & 8 & 7 & 6 & 2 & 2 & 8 & 6 \end{array}$$

Theo kết quả điều tra cũ thì số giờ tự học trung bình của sinh viên trong tuần là 8. Với mức ý nghĩa 5%, hãy so sánh kết quả mới điều tra với kết quả cũ.

Bài 4.27. Kết quả đo chiều cao của 24 trẻ em 2 tuổi như sau sau (cm):

$$\begin{array}{cccccc} 84,4 & 89,9 & 89,0 & 81,9 & 87,0 & 78,5 \\ 84,1 & 86,3 & 80,6 & 80,0 & 81,3 & 86,8 \\ 83,4 & 89,4 & 85,4 & 80,6 & 85,0 & 82,5 \\ 80,7 & 84,3 & 85,4 & 85,0 & 85,5 & 81,6 \end{array}$$

Kiểm định lại ý kiến cho rằng chiều cao trung bình của trẻ 2 tuổi là 84 cm với mức ý nghĩa 5%. Giả thiết chiều cao của trẻ em 2 tuổi có phân phối chuẩn

Bài 4.28. Quan sát định mức thời gian hoàn thành của một sản phẩm (quả bóng) đạt tiêu chuẩn Quốc tế của công ty Động Lực là 14 phút. Khi sản xuất nghi ngờ định mức trên chưa sát với thực tế, theo dõi thời gian hoàn thành một quả bóng của 25 công nhân ta được

Thời gian hoàn thành 1 sản phẩm	11	13	15	17	19
Số công nhân	2	6	10	4	3

Hãy xét xem điều nghi ngờ trên có đúng không, với mức ý nghĩa 0,05. Biết thời gian hoàn thành một quả bóng được xem tuân theo quy luật phân phối chuẩn.

Bài 4.29. Số liệu sau đây là chỉ số thông minh IQ của 28 sinh viên một trường đại học (giả sử chỉ số đó tuân theo luật phân phối chuẩn):

130	135	121	123	121	117	135
117	121	123	117	130	112	112
108	121	117	112	117	121	123
117	10	121	121	112	123	130

Có ý kiến cho rằng chỉ số IQ của sinh viên trường đó cao hơn 121. Với độ tin cậy 95%, hãy kiểm định lại ý kiến trên?

Bài 4.30. Cho X là năng suất lúa ở một khu vực (tạ/ha). Điều tra ở một số thửa ruộng trong khu vực ta có:

Năng suất (X)	30-35	35-40	40-45	45-50	50-55
Số thửa ruộng	6	18	28	40	16

Những thửa ruộng đạt năng suất trên 45 tạ/ha là những thửa ruộng đạt năng suất cao. Người ta nhận định tỷ lệ những thửa ruộng đạt năng suất cao chiếm hơn 50%. Nhận định đó có đúng không, $\alpha = 5\%$,?

Bài 4.31. Cô giáo cho rằng có đến 30% học sinh không làm bài tập về nhà, học sinh lại cho rằng cô giáo quá bi quan. Người ta kiểm tra 50 em học sinh của trường đó thì thấy 15 em không làm bài tập về nhà. Với mức ý nghĩa $\alpha = 5\%$, hãy kiểm tra nhận định nào đúng?

Bài 4.32. Điều tra năng suất lúa trên diện tích 100 ha trồng lúa của một vùng, ta thu được bảng số liệu sau:

Năng suất (tạ/ha)	41	44	45	46	48	52	54
Số ha	10	20	30	15	10	10	5

Những thửa ruộng có năng suất từ 48 tạ/ha trở lên là những thửa ruộng có năng suất cao. Hãy kiểm định ý kiến cho rằng tỷ lệ diện tích có năng suất cao trong vùng chiếm hơn 20% với độ tin cậy 97%.

Bài 4.33. Giám đốc một công ty tuyên bố 85% động cơ của công ty đạt chuẩn quốc gia. Một hãng độc lập kiểm tra 200 động cơ của công ty thì thấy có 160 động cơ đạt yêu cầu. Với mức $\alpha = 5\%,$ ta có thể kết luận gì về tuyên bố trên?

Bài 4.34. Kiểm tra những sản phẩm được chọn ngẫu nhiên ở hai nhà máy sản xuất cùng một loại hàng ta có số liệu sau:

Nhà máy	Số sản phẩm kiểm tra	Số phế phẩm
I	4000	120
II	800	20

Với mức ý nghĩa $\alpha = 5\%$, có thể coi tỷ lệ phế phẩm do nhà máy I sản xuất cao hơn hay không?

Bài 4.35. Trọng lượng sản phẩm do 2 máy sản xuất là đại lượng ngẫu nhiên tuân theo luật phân phối chuẩn với cùng độ lệch tiêu chuẩn là 1.2 kg. Với mức ý nghĩa $\alpha = 5\%$, có thể nói rằng trọng lượng trung bình của sản phẩm do 2 máy sản xuất là như nhau hay không nếu biết khi cân 25 sản phẩm do máy I sản xuất thấy tổng trọng lượng là 1240, còn cân 20 sản phẩm của máy II thì tổng trọng lượng là 1010.

Bài 4.36. Theo dõi số tai nạn lao động tại 2 phân xưởng, ta có số liệu sau:
Phân xưởng I: $20/200$ công nhân bị tai nạn.
Phân xưởng II: $120/800$ công nhân bị tai nạn.
Với mức ý nghĩa $\alpha = 0,005$ có thể nói rằng phân xưởng I có chất lượng bảo hộ lao động tốt hơn hay không?

Bài 4.37. Để kiểm tra tư duy Toán học của sinh viên Khoa CNTT và sinh viên khoa KHTN, người ta tiến hành chọn ngẫu nhiên 15 sinh viên khoa CNTT và 14 sinh viên khoa KHTN làm một bài kiểm tra trắc nghiệm, được kết quả điểm như sau:
Với mức ý nghĩa $\alpha = 5\%,$ hãy kết luận xem tư duy Toán học của sinh viên

SV	1	2	3	4	5	6	7	8	9	10	11	2	13	14	15
CNTT	399	378	360	374	321	360	408	361	384	330	413	326	354	350	270
KHTN	318	284	290	328	385	352	324	276	360	375	360	362	238	234	

hai khoa có tương đương với nhau hay không? Biết điểm thi tuân theo phân phối chuẩn.

Bài 4.38. Người ta cân trẻ sơ sinh ở 2 khu vực thành thị và nông thôn, thu được kết quả sau:

Khu vực	Số trẻ cân	Trọng lượng trung bình (kg)	Phương sai hiệu chỉnh
Nông thôn	2500	3,0	0,09
Thành thị	500	3,2	0,04

Với mức ý nghĩa 1%, có thể nói trọng lượng trung binh của trẻ sơ sinh ở 2 khu vực trên là như nhau không?

Bài 4.39. Kiểm tra những sản phẩm được chọn ngẫu nhiên ở hai nhà máy sản xuất cùng một loại hàng ta có số liệu sau:

Nhà máy	Số sản phẩm kiểm tra	Số phế phẩm
A	1000	20
B	900	30

Với mức ý nghĩa $\alpha = 5\%$, có thể coi tỷ lệ phế phẩm do hai nhà máy trên sản xuất ra có như nhau hay không?

Bài 4.40. Nghiên cứu trọng lượng trẻ sơ sinh của hai nhóm với mẹ nghiện thuốc lá và không hút thuốc lá, ta có kết quả (kg) như sau:
Giả sử rằng trọng lượng trẻ của từng nhóm đều có phân phối chuẩn và có cùng phương sai. Với mức $\alpha = 5\%$ có thể coi rằng trẻ sơ sinh của nhóm mẹ nghiện thuốc nhẹ cân hơn của nhóm mẹ không nghiện thuốc không? Giả thiết trọng lượng trẻ phân phối chuẩn.

Không hút thuốc	4	3,8	3,6	3,7	3,2	3,6	4,1	3,6	3,8	3,31	4,1	3,3	3,5	3,5	2,7
Nghiện thuốc lá	3,2	2,8	2,9	3,3	3,9	3,5	3,3	2,8	3,6	3,7	3,6	3,7	2,4	2,3	

Bài 4.41. Để thử nghiệm hiệu quả của một loại thuốc trừ sâu người ta áp dụng thử loại thuốc này đối với 5 thửa ruộng đang bị sâu phá hoại. Số lượng sâu bắt được trước và sau khi dùng loại thuốc trừ sâu nói trên được cho ở bảng sau:

Thửa ruộng	1	2	3	4	5
Trước khi phun thuốc	110	70	90	115	100
Sau khi phun thuốc	100	65	80	95	90

Với độ tin cậy 95% có thể kết luận số lượng sâu sau khi phun thuốc đã giảm so với trước khi phun thuốc hay không? Giả sử số lượng sâu tại mỗi thửa ruộng đều phân phối chuẩn.

Bài 4.42. Nghiên cứu ảnh hưởng của chế độ dinh dưỡng của bà mẹ khi mang thai đối với sự phát triển của thai nhi, người ta khảo sát trọng lượng trẻ sơ sinh ở hai vùng nông thôn và thành phố, ta có kết quả (kg) như sau:

Nông thôn	2,34	2,38	3,63	3,59	3,60	3,75	2,80	3,23	3,45	2,80	2,90	3,18	
Thành phố	4,08	3,79	3,60	3,75	3,90	3,21	2,71	3,26	3,31	4,14	3,52	3,54	3,6

Với độ tin cậy là 95% có thể kết luận trẻ sơ sinh ở nông thôn nhẹ cân hơn trẻ sơ sinh ở thành phố? Giả sử trọng lượng của từng nhóm trẻ đều có phân phối chuẩn và có cùng phương sai.

Chương 5

PHÂN TÍCH TƯƠNG QUAN HỒI QUY

Trong nhiều bài toán, với một đối tượng cần nghiên cứu, ta không chỉ xét một chỉ tiêu riêng biệt mà cần kết hợp nhiều chỉ tiêu với nhau thì mới có được kết quả mong muốn. Các chỉ tiêu này có thể mang tính chất định tính hoặc định lượng. Ví dụ: Kỹ sư phần mềm muốn dự đoán được thời gian xử lý để đánh giá nhiệt động lực được chạy trên máy tính đặc biệt thì cần có thông tin về thời gian xử lý, dung lượng RAM đã dùng, tổng đầu vào, tổng đầu ra.

Như vậy, để giải quyết một vấn đề cụ thể người ta phải nghiên cứu, phân tích nhiều chỉ tiêu là các biến ngẫu nhiên, các nhân tố, dấu hiệu. Ta có thể làm rõ bản chất của hiện tượng, sự việc cần nghiên cứu để tìm ra quy luật, dự đoán được xu thế biến động của hiện tượng, sự việc đó trong tương lai bằng phương pháp phân tích mối quan hệ phụ thuộc giữa các chỉ tiêu phản ánh nội dung của hiện tượng, sự việc cần nghiên cứu (gọi là phương pháp phân tích tương quan và hồi quy). Để đơn giản ta xét mối tương quan giữa 2 chỉ tiêu của một đối tượng nghiên cứu (biến ngẫu nhiên 2 chiều).

Sự phụ thuộc giữa X và Y được chia thành 2 loại:

- Phụ thuộc hàm: Hai biến ngẫu nhiên X và Y được gọi là phụ thuộc hàm nếu tồn tại hàm f sao cho $Y = f(X)$.

- Phụ thuộc thống kê: hai biến ngẫu nhiên X và Y được gọi là phụ thuộc thống kê nếu mỗi giá trị x của biến X tương ứng với hàm phân phối (hoặc hàm mật độ) có điều kiện của biến Y: $F(y/x)$ hoặc $f(y/x)$.

5.1. Phân tích tương quan

5.1.1. Hiệp phương sai của các biến ngẫu nhiên

Định nghĩa 5.1. *Hiệp phương sai (covariance) giữa X và Y ký hiệu là $Cov(X,Y)$ và là kỳ vọng toán của tích các sai lệch của các biến ngẫu nhiên X,Y với các kỳ vọng toán của chúng. Cụ thể: $Cov(X,Y)$ được xác định theo công thức:*

$$Cov(X,Y) = E\left[(\mathrm{X}-\mathrm{E}(\mathrm{X}))\left(Y-E(Y)\right)\right] = E(X.Y) - E(X).E(Y),$$

trong đó:

$E(X), E(Y)$ là kỳ vọng của từng biến ngẫu nhiên X,Y;

$E(X.Y)$ được tính theo công thức:

-Nếu (X,Y) là biến ngẫu nhiên rời rạc, có phân phối xác suất đồng thời $P(X=x_i, Y=y_j) = pij$ thì $E(X.Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j p_{ij}$.

-Nếu (X,Y) là biến ngẫu nhiên liên tục, có hàm mật độ xác suất đồng thời $f(x,y)$ thì $E(X.Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy$.

Ví dụ 5.2. Cho số liệu

x_i \ y_j	0,5	1	1,5
0	0,1	0,3	0,2
1	0,06	0,18	0,16

Tính $Cov(X,Y)$.

Giải. Ta tính được: $E(X) = 0.0,6 + 1.0,4 = 0,4$.

$E(Y) = 0,5.0,16 + 1.0,48 + 1,5.0,36 = 1,1$.

$E(X.Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j p_{ij} = 0,45.$

Từ đó: $Cov(X,Y) = 0,45 - 0,4.1,1 = 0,01.$

5.1.2. Hệ số tương quan

Định nghĩa 5.3. *Hệ số tương quan giữa 2 biến ngẫu nhiên X,Y được ký hiệu là ρ_{xy} là tỷ số giữa hiệp phương sai và tích các độ lệch chuẩn của các biến ngẫu nhiên đó* $\rho_{xy} = \dfrac{Cov(X,Y)}{\sigma_{\mathrm{X}}.\sigma_Y}$.

Hệ số tương quan không có đơn vị đo và có các tính chất sau:

1.$\rho_{xy} = \rho_{yx}$,

2. $|\rho_{xy}| \leq 1$,

3. Nếu X và Y độc lập thì $\rho_{xy} = 0$,

4. Nếu $\rho_{xy} = \pm 1$ thì X và Y phụ thuộc hàm với nhau.

Hiệp phương sai và hệ số tương quan được dùng để đánh giá mức độ chặt chẽ của mối liên hệ phụ thuộc giữa các biến ngẫu nhiên X và Y: Hai biến ngẫu nhiên gọi là tương quan với nhau nếu hiệp phương sai (hay hệ số tương quan) khác không và chúng không tương quan nếu hiệp phương sai hay hệ số tương quan bằng không.

Chú ý: Hai biến ngẫu nhiên tương quan thì phụ thuộc nhau nhưng điều ngược lại chưa hẳn đã đúng.

Ví dụ 5.4. Với số liệu cho như ở ví dụ 5.1, xét tính độc lập và xác định mối tương quan giữa X và Y.

Giải. Ta thấy mối tương quan giữa X và Y được thể hiện bởi:

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_{\mathrm{X}}.\sigma_Y},$$

trong đó:

$\sigma_X = \sqrt{0^2.0,6 + 1^2.0,4 - (0,4)^2} = \sqrt{0,24} = 0,4899$

$\sigma_Y = \sqrt{(0,5)^2.0,16 + 1^2.0,48 + (1,5)^2.0.36 - (1,1)^2} = 0,3464.$

Thay số ta được:

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_{\mathrm{X}}.\sigma_Y} = \frac{0,01}{0,4899.0,3464} = 0,0599.$$

Do ρ_{XY} khá gần 0 nên X và Y có tương quan song mối tương quan giữa chúng không chặt chẽ (tương quan yếu).

Ví dụ 5.5. Cho biến ngẫu nhiên 2 chiều liên tục (X,Y) có hàm mật độ đồng thời:

$$f(x,y) = \begin{cases} \dfrac{1}{6\pi} \quad khi \quad \dfrac{x^2}{9} + \dfrac{y^2}{4} \leq 1 \\ \\ 0 \qquad khi \quad \dfrac{x^2}{9} + \dfrac{y^2}{4} > 1 \end{cases}.$$

Xét tính độc lập và xác định mối tương quan giữa X và Y.

Giải. Hàm mật độ xác suất của từng thành phần được xác định bằng:

$$f_1(x) = \int\limits_{-\infty}^{+\infty} f(x,y)dy = \begin{cases} \dfrac{2}{9\pi}\sqrt{9-x^2} \; khi \quad |x| \leq 3 \\ \\ 0 \qquad\qquad khi \; |x| > 3 \end{cases}$$

$$f_2(y) = \int\limits_{-\infty}^{+\infty} f(x,y)dx = \begin{cases} \dfrac{1}{2\pi}\sqrt{4-y^2} \; khi \quad |x| \leq 2 \\ \\ 0 \qquad\qquad khi \; |x| > 2 \end{cases}.$$

Dễ thấy $f(x,y) \neq f_1(x).f_2(y)$ nên X và Y không độc lập.Để xét mối tương quan ta cần xác định $Cov(X,Y)$, ta có:

$$Cov(X,Y) = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} xyf(x,y)dxdy - E(X).E(Y).$$

Do $f_1(x), f_2(y)$ là các hàm đối xứng qua các trục toạ độ nên $E(X) = 0$ và $E(Y) = 0$. Từ đó ta được:

$$Cov(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy = \frac{1}{6\pi}\int_{-\frac{2}{3}\sqrt{9-x^2}}^{\frac{2}{3}\sqrt{9-x^2}} y\left(\int_{-\frac{3}{2}\sqrt{4-x^2}}^{\frac{3}{2}\sqrt{4-x^2}} xdx\right)dy = 0.$$

Do tích phân trong ngoặc có hàm dưới dấu tích phân là hàm lẻ, cận lấy tích phân đối xứng qua gốc toạ độ. Như vậy X và Y không tương quan.

Chú ý:

- Khi $\rho_{XY} = \pm 1$ ta nói X và Y có mối tương quan tuyến tính. Khi đó, biến ngẫu nhiên X có thể biểu diễn bằng một hàm tuyến tính theo Y và ngược lại.

- Nếu $\rho_{XY} > 0$ thì ta nói X và Y có mối tương quan đồng biến (tương quan dương), ngược lại thì X và Y có mối tương quan nghịch biến (tương quan âm).

5.1.3. Hệ số tương quan mẫu

Định nghĩa 5.6. *Giả sử (X_i, Y_i), $i = 1, 2, ..., n$ là một mẫu ngẫu nhiên của biến ngẫu nhiên 2 chiều (X, Y) thì hệ số tương quan mẫu của X và Y được xác định bởi:*

$$R_{xy} = \frac{\overline{X.Y} - \overline{X}.\overline{Y}}{S_{\mathrm{X}}.S_Y},$$

trong đó:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i; \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i; \overline{\mathrm{X}.Y} = \frac{1}{n}\sum_{i=1}^{n} X_i.Y_i;$$

$$S_X = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}; S_Y = \sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}.$$

Khi tiến hành nghiên cứu, khảo sát (thu thập số liệu), ta thu được mẫu cụ thể (x_i, y_i), $i = 1, 2, ..., n$.

Khi đó ta có hệ số tương quan mẫu được xác định bởi công thức:

$$r_{xy} = \frac{\overline{x.y} - \overline{x}.\overline{y}}{s_{\mathrm{X}}.s_Y},$$

với $\overline{x.y}, \overline{x}, \overline{y}, s_X, s_Y$ là các đặc trưng của mẫu cụ thể và được tính theo các trường hợp:

a. Nếu số liệu được cho dưới dạng:

x_i	x_1	x_2	$\ldots$	x_n
y_i	y_1	y_2	$\ldots$	y_n

thì các đặc trưng của mẫu cụ thể được tính bởi công thức:

$$\overline{x.y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i; \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i;$$

$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2}; s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\overline{y})^2}.$$

Ví dụ 5.7. Từ số liệu đường kính (X) và trọng lượng vỏ khô (Y) của 10 cây quế được cho dưới đây:

Đường kính (cm)	10	12	14	16	18	19	20	22	23	24
Trọng lượng vỏ khô (kg)	10	9	15	15	17	16	15	18	20	17

Hãy xác định hệ số tương quan mẫu.

Giải. Lập bảng:

x_i	y_i	$x_i.y_i$	x_i^2	y_i^2
10	10	100	100	100
12	9	108	144	81
14	15	210	196	225
16	15	240	256	225
18	17	306	324	289
19	16	304	361	256
20	15	300	400	225
22	18	396	484	324
23	20	460	529	400
24	17	408	576	289
Tổng = 178	152	2832	3370	2414

Từ đó:

$$\overline{x.y} = 283,2; \overline{x} = 17,8; \overline{y} = 15,2;$$

$$s_X = \sqrt{337 - 17.8^2} = 4,49; s_Y = \sqrt{241.4 - 15.2^2} = 3,22.$$

Thay số ta được:

$$r_{xy} = \frac{\overline{x.y} - \overline{x}.\overline{y}}{s_{\text{X}}.s_Y} = \frac{283,2 - 17,8.15.2}{4,49.3,22} = 0,874.$$

b. Nếu số liệu được cho dưới dạng:

x_i	x_1	x_2	...	x_k
y_i	y_1	y_2	...	y_k
n_i	n_1	n_2	...	n_k

thì các đặc trưng của mẫu cụ thể được tính bởi công thức:

$$\overline{x.y} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i y_i; \overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i; \bar{y} = \frac{1}{n}\sum_{i=1}^{k} n_i y_i;$$

$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i x_i - (\overline{x})^2}; s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i y_i - (\bar{y})^2}.$$

Ví dụ 5.8. Nghiên cứu mối liên hệ giữa X là số tiền đầu tư cho việc phòng bệnh (nghìn đồng) và Y là tỷ lệ người mắc bệnh (%) ở 50 địa phương ta có số liệu:

X \ Y	2	2.5	3	3.5	4
1000				2	3
2000			3	6	2
3000		4	6	3	
4000	1	6	4	1	
5000	6	3			

Tính hệ số tương quan và nhận xét về sự phụ thuộc giữa X và Y.

Giải. Ta lập bảng

x_i	y_i	n_i	$n_i.x_i$	$n_i.y_i$	$n_i.x_i.y_i$	$n_i.x_i^2$	$n_i.y_i^2$
1000	3.5	2	2000	7	7000	2000000	24,5
1000	4	3	3000	12	12000	3000000	48
2000	3	3	6000	9	18000	12000000	27
2000	3,5	6	12000	21	42000	24000000	73,5
2000	4	2	4000	8	16000	8000000	32
3000	2,5	4	12000	10	30000	36000000	25
3000	3	6	18000	18	54000	54000000	54
3000	3,5	3	9000	10,5	31500	27000000	36,75
4000	2	1	4000	2	8000	16000000	4
4000	2,5	6	24000	15	60000	96000000	37,5

4000	3	4	16000	12	48000	64000000	36
4000	3,5	1	4000	3,5	14000	16000000	12,25
5000	2	6	30000	12	60000	150000000	24
5000	2,5	3	15000	7,5	37500	75000000	18,75
Tổng		$n = 50$	159000	147,5	438000	583000000	453,25

Từ đó tính được:

$$\overline{x.y} = 8760; \overline{x} = 3180; \overline{y} = 2,95; s_X = 1244,026; s_Y = 0,602.$$

Thay số ta được:

$$r_{xy} = \frac{\overline{x.y} - \overline{x}.\overline{y}}{s_X.s_Y} = \frac{8760 - 3180.2,95}{1244,026.0,602} = -0,83.$$

Nhận xét: X và Y có mối tương quan tuyến tính và do hệ số tương quan âm nên mối tương quan giữa X và Y là tương quan nghịch biến. Nghĩa là nếu số tiền đầu tư cho việc phòng bệnh tăng lên thì tỷ lệ mắc bệnh sẽ giảm đi và ngược lại.

c. Nếu số liệu được cho dưới dạng:

Y \ X	...	x_i	...	m_j	$m_j.y_j$	$m_j.y_j^2$	$\sum x_i y_j n_{ij}$
...							
y_j	...	n_{ij} ($n_{ij}.x_i.y_j$)					
...							
n_i				n	$\Sigma m_j.y_j$	$\Sigma m_j.y_j^2$	
$x_i.n_i$				$\Sigma n_i x_i$			
$x_i^2.n_i$				$\Sigma n_i x_i^2$			$\Sigma\Sigma=$

thì các đặc trưng của mẫu cụ thể được tính bởi công thức:

$$\bar{x} = \frac{1}{n}\sum x_i n_i; \overline{y} = \frac{1}{n}\sum y_j m_j; \overline{x.y} = \frac{1}{n}\sum\sum x_i y_j n_{ij};$$

$$s_X = \sqrt{\frac{1}{n}\sum x_i^2 n_i - (\overline{x})^2}; s_Y = \sqrt{\frac{1}{n}\sum y_j^2 m_j - (\bar{y})^2}.$$

Chú ý: Hệ số tương quan thực nghiệm có các tính chất của hệ số tương quan chỉ với chú ý rằng khi $|r_{XY}| \geq 0,7$, ta có thể coi giữa X và Y có mối tương quan tuyến tính chặt.

5.2. Phân tích hồi quy

Hồi quy được dùng để xem xét mối liên hệ tuyến tính giữa hai biến ngẫu nhiên X và Y, trong đó X được xem là biến độc lập, Y là biến phụ thuộc. Mục tiêu của phân tích hồi quy là mô hình hoá mối liên hệ (xây dựng một mô hình toán học nhằm thể hiện một cách tốt nhất mối liên hệ giữa 2 biến X và Y). Phân tích hồi quy xác định sự liện quan định lượng giữa X và Y, kết quả của phân tích hồi quy được dùng cho dự đoán, ngược lại phân tích tương quan khảo sát khuynh hướng và mức độ của sự liên quan, được dùng để đo lường tính bền vững của mối liên hệ giữa các biến đặc biệt là biến định lượng.

5.2.1. Mô hình hồi quy tuyến tính đơn giản của tổng thể

Giả sử có 2 biến ngẫu nhiên X và Y, trong đó Y được xem là biến phụ thuộc tuyến tính vào X. Với một giá trị x_i nào đó của biến X, giá trị tương ứng $\widehat{y_i}$ của biến Y được thể hiện bằng công thức:

$$\widehat{y_i} = \alpha + \beta x_i + \varepsilon_i,$$

trong đó:

α, β là các hằng số: α thể hiện giá trị ước lượng của Y khi giá trị của biến Xbằng không, β là độ dốc của đường hồi quy, thể hiện mức tăng lên của Y khi X tăng một đơn vị.

ε_i là sai số ngẫu nhiên thể hiện ảnh hưởng của các yếu tố khác (các yếu tố không được nghiên cứu) tới Y. ε_i được xem là các biến ngẫu nhiên độc lập, có phân phối chuẩn với trung bình bằng không, phương sai bằng nhau.

Một cách tổng quát mô hình hồi quy tuyến tính đơn giản của tổng thể thể hiện mối liên hệ tuyến tính giữa Y và X là $y = \alpha + \beta x + \varepsilon$.

Trước tiên ta xét mối liên hệ giữa các chỉ tiêu, nếu số liệu quan sát về các chỉ tiêu đó gợi ý xác lập một đường thẳng để mô tả mối liên hệ này thì ta xác định phương trình đường thẳng. Ngược lại, ta phải sử dụng một phương trình khác hay phương pháp phân tích khác để đưa số liệu đã có về dạng tuyến tính (tuyến tính hoá).

Trong thực tế ta không thể xác định một cách chính xác các hằng số của phương trình hồi quy tuyến tính của tổng thể mà chỉ có thể ước lượng chúng từ các giá trị quan sát.

5.2.2. Phương trình hồi quy của mẫu theo phương pháp bình phương tối thiểu

Giả sử khi quan sát các biến ngẫu nhiên X và Y ta thu được một mẫu cỡ n :

$$(x_1, y_1)\,;(x_2, y_2)\,;(x_3, y_3)\,;...;(x_n, y_n)\,.$$

Ta sẽ tìm các giá trị a, b để ước lượng cho các hằng số α, β. Do các điểm trên đồ thị quan sát không nằm trên một đường thẳng nên ta thực hiện theo cách sau: Kẻ nhiều đường thẳng xuyên qua các điểm quan sát được và ta chọn ra một đường thẳng mô tả tốt nhất xu hướng liên hệ giữa X và Y. Phương pháp dùng để xác định đường thẳng này là phương pháp bình phương tối thiểu. Phương pháp này tìm ra được một đường thẳng làm cực tiểu hoá được tổng các độ lệch bình phương giữa tung độ của các điểm quan sát được và đường thẳng. Một đường thẳng được xem là "thích hợp nhất" khi tổng phương giữa các giá trị thực tế y_i với giá trị $\widehat{y_i}$ là nhỏ nhất. Tức là: với đường thẳng $y_i = a + bx_i + e_i$ ta cần xác định các hệ số a, b thoả mãn:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min .$$

Từ đó, ta tính được:

$$b = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum\limits_{i=1}^{n} x_i^2 - n(\overline{x})^2} = \frac{\overline{xy} - \overline{x}\overline{y}}{s_X^2} = r_{XY}\frac{s_Y}{s_X};$$

$$a = \overline{y} - b\overline{x}.$$

Đường hồi quy tìm được là: $\widehat{y} = a + bx$.

Ví dụ 5.9. Khảo sát ngẫu nhiên 30 khách hàng nữ đi siêu thị ta thu được số liệu về mức thu nhập hộ trung bình/tháng (X-triệu đồng) và số món hàng mua ngoài dự định trong tháng (Y) như sau:

STT	x_i	y_i	*STT*	x_i	y_i	*STT*	x_i	y_i
1	*3*	*2*	*11*	*9*	*5*	*21*	*4,5*	*4*
2	*3,5*	*2*	*12*	*7*	*6*	*22*	*3,5*	*2*
3	*5*	*4*	*13*	*6*	*4*	*23*	*4*	*3*
4	*4*	*3*	*14*	*6*	*5*	*24*	*4,5*	*5*
5	*6,5*	*5*	*15*	*2*	*2*	*25*	*6,5*	*4*
6	*3*	*1*	*16*	*3*	*3*	*26*	*7*	*4*
7	*6*	*3*	*17*	*3*	*2*	*27*	*6*	*5*
8	*5*	*2*	*18*	*6*	*4*	*28*	*4,2*	*2*
9	*2*	*0*	*19*	*3,5*	*3*	*29*	*4*	*2*
10	*3,5*	*1*	*20*	*4*	*2*	*30*	*4,4*	*3*

a. Xác định hệ số tương quan mẫu giữa X và Y.

b. Xác định đường hồi quy tuyến tính thực nghiệm của Y đối với X (nếu có).

Giải. a. Hệ số tương quan mẫu đượng tính theo công thức:

$$r_{xy} = \frac{\overline{x.y} - \overline{x}.\overline{y}}{s_X.s_Y}.$$

Từ số liệu đã cho ta dễ dàng tính được:

$$\overline{xy} = \frac{1}{30}\sum_{i=1}^{30} x_i y_i = \frac{1}{30}.487,5 = 16,253;$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{30} x_i = \frac{1}{30}.139,6 = 4,653; \overline{y} = \frac{1}{n}\sum_{i=1}^{30} y_i = \frac{1}{30}.93 = 3,1;$$

$$s_X = \sqrt{\frac{1}{30}.728 - 4,563^2} = 1,6166; s_Y = \sqrt{\frac{1}{30}.349 - 3,1^2} = 1,4224.$$

Thay số ta được:

$$r_{xy} = \frac{\overline{x.y} - \overline{x}.\overline{y}}{s_X.s_Y} = \frac{16,253 - 4,653.3,1}{1,6166.1,4224} = 0,795.$$

Do hệ số tương quan bằng 0.795> 0.7 nên mối tương quan giữa mức thu nhập hộ trung bình/tháng và số món hàng mua ngoài dự định trong tháng là mối tương quan tuyến tính chặt.

b. Đường hồi quy tuyến tính thực nghiệm được xác định từ công thức

$$\widehat{y} = a + bx,$$

với các hệ số:

$$b = r_{XY}\frac{s_Y}{s_X} = 0,795.\frac{7,79}{8,85} = 0,6995;$$

$$a = 3,1 - 0,6995.4,653 = -0,1548.$$

Vậy đường hồi quy tuyến tính thực nghiệm cần tìm là

$$\widehat{y} = -0.1548 + 0.6995x.$$

Chú ý: Hệ số tương quan chỉ đo được mức độ liên hệ tuyến tính giữa hai biến ngẫu nhiên X và Y. Để đo mức độ phụ thuộc hàm bất kỳ giữa chúng ta phải dùng đến khái niệm tỷ số tương quan của Y theo X $\eta^2_{Y/X}$ hoặc tương tự là tỷ số tương quan của X theo Y $\eta^2_{X/Y}$.

5.3. Bài tập chương 5

Bài 5.1. Quan sát một mẫu cây công nghiệp với 2 dấu hiệu quan sát X : Đường kính (cm), Y : Chiều cao (m). Ta có số liệu quan sát cho trong bảng sau:

Y \ X	20-22	22-24	24-26	26-28	28-30
3	2				
4	5	3			
5		11	8	4	
6			15	17	
7			10	6	7
8					12

a. Tính hệ số tương quan giữa X và Y.

b. Lập phương trình tương quan tuyến tính của Y theo X.

Bài 5.2. Quan sát một mẫu thì thấy có $X(\%)$ cây có chiều cao Y(m)

Y \ X	2.5	7.5	12.5	17.5	22.5
115-125	7				
125-135	12	8			
135-145		20	10	2	
145-155		19	15	9	5
155-165			16	8	3

a. Tính hệ số tương quan giữa X và Y.

b. Lập phương trình tương quan tuyến tính của Y theo X.

Bài 5.3. Với 10 cặp quan sát về X và Y như sau:

X	80	82	79	60	65	92	90	81	70	68
Y	110	111	102	87	92	112	110	100	81	92

Trong đó X là vòng bụng và Y là chiều dài (cả hai đều đo bằng cm) do một cửa hàng may đo đã đo khách hàng để cắt quần.

a. Tìm hệ số tương quan và lập phương trình tương quan tuyến tính của Y theo X.

b. Một chiếc quần có vòng bụng 82.5cm thì chiều dài cỡ bao nhiêu?

Bài 5.4. Đo một chiều cao Y (cm) và đường kính X (cm) của một loại cây ta có bảng kết quả sau:

X	5	6	5	6	10	5	7	8	9	10
Y	28	28	24	30	60	30	32	42	43	49

a. Tìm hệ số tương quan và lập phương trình tương quan tuyến tính của Y theo X.

b. Cho $X = 5,5cm$. Dự đoán chiều cao Y.

Bài 5.5. Đo một chiều cao Y (cm) và đường kính X (cm) của một loại cây ta có bảng kết quả sau:

X	5	6	5	6	10	5	7	8	9	10
Y	28	28	24	30	60	30	32	42	43	49

a. Tìm hệ số tương quan và lập phương trình tương quan tuyến tính của X theo Y.

b. Cho $Y = 29$ cm. Dự đoán đường kính X.

Bài 5.6. X (đơn vị: kg) và Y (đơn vị: cm) là hai chỉ tiêu của một loại sản phẩm. Điều tra một mẫu ta có bảng số liệu sau:

a. Viết công thức và tính hệ số tương quan mẫu r_{XY}.

b. Tìm hàm hồi quy tuyến tính mẫu của Y theo X.

Y \ X	80-84	84-88	88-92	92-96	96-100
1	8				
3	12	9	4		
5		11	15	10	
7			7	3	3

Bài 5.7. Cho bảng quan sát về X và Y như sau:

Y	32	29	26	23	20	17	14	11
X	7	11	15	19	23	27	31	35

a. Tìm hệ số tương quan giữa X và Y.

b. Tìm đường hồi quy bình phương trung bình tuyến tính thực nghiệm của Y theo X.

Bài 5.8. Cho bảng quan sát về X và Y như sau:

X	1	3	5	7	9	11	13	15	17
Y	6	11	16	21	26	31	36	41	46

a. Tìm hệ số tương quan giữa X và Y.

b. Tìm đường hồi quy bình phương trung bình tuyến tính thực nghiệm của Y theo X.

Bài 5.9. Sản lượng khai thác than ở một công ty than được ghi lại qua 9 năm:

Năm X	1988	1989	1990	1991	1992	1993	1994	1995	1996
Sản lượng Y	60	61	64	65	66	66	69	70	72

a. Tính hệ số tương quan mẫu giữa X và Y.

b. Tìm đường hồi quy bình phương trung bình tuyến tính của Y theo X. Hãy dự đoán số than khai thác được vào năm 2012.

Bài 5.10. Khảo sát sự tương quan giữa X và Y ta có số liệu

X	-5	-3	1	6	8	12	18
Y	4	7	9	15	20	26	38

a. Tính hệ số tương quan mẫu giữa X và Y.

b.Tìm đường hồi quy bình phương trung bình tuyến tính của Y theo X.

Bài 5.11. Hai đại lượng ngẫu nhiên X và Y. có mối liên hệ cho dưới bảng sau:

a. Tính hệ số tương quan mẫu giữa X và Y.

b.Tìm đường hồi quy bình phương trung bình tuyến tính của Y theo X.

Phụ lục

Bảng 1. **GIÁ TRỊ PHÂN PHỐI POISSON P(λ)** $P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda}$

x \ λ	1	2	3	4	5	6	7	8	9	10
0	0.3679	0.1353	0.0498	0.0183	0.0067	0.0025	0.0009	0.0003	0.0001	0.0000
1	0.3679	0.2707	0.1494	0.0733	0.0337	0.0149	0.0064	0.0027	0.0011	0.0005
2	0.1839	0.2707	0.2240	0.1465	0.0842	0.0446	0.0223	0.0107	0.0050	0.0023
3	0.0613	0.1804	0.2240	0.1954	0.1404	0.0892	0.0521	0.0286	0.0150	0.0076
4	0.0153	0.0902	0.1680	0.1954	0.1755	0.1339	0.0912	0.0573	0.0337	0.0189
5	0.0031	0.0361	0.1008	0.1563	0.1755	0.1606	0.1277	0.0916	0.0607	0.0378
6	0.0005	0.0120	0.0504	0.1042	0.1462	0.1606	0.1490	0.1221	0.0911	0.0631
7	0.0001	0.0034	0.0216	0.0595	0.1044	0.1377	0.1490	0.1396	0.1171	0.0901
8	0.0000	0.0009	0.0081	0.0298	0.0653	0.1033	0.1304	0.1396	0.1318	0.1126
9		0.0002	0.0027	0.0132	0.0363	0.0688	0.1014	0.1241	0.1318	0.1251
10		0.0000	0.0008	0.0053	0.0181	0.0413	0.0710	0.0993	0.1186	0.1251
11			0.0002	0.0019	0.0082	0.0225	0.0452	0.0722	0.0970	0.1137
12			0.0001	0.0006	0.0034	0.0113	0.0263	0.0481	0.0728	0.0948
13			0.0000	0.0002	0.0013	0.0052	0.0142	0.0296	0.0504	0.0729
14				0.0001	0.0005	0.0022	0.0071	0.0169	0.0324	0.0521
15				0.0000	0.0002	0.0009	0.0033	0.0090	0.0194	0.0347
16					0.0000	0.0003	0.0014	0.0045	0.0109	0.0217
17						0.0001	0.0006	0.0021	0.0058	0.0128
18						0.0000	0.0002	0.0009	0.0029	0.0071
19							0.0001	0.0004	0.0014	0.0037
20							0.0000	0.0002	0.0006	0.0019
21								0.0001	0.0003	0.0009
22								0.0000	0.0001	0.0004
23									0.0000	0.0002
24										0.0001
25										0.0000

Bảng 2. GIÁ TRỊ HÀM MẬT ĐỘ PHÂN PHỐI CHUẨN HÓA $\varphi(u)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.3989	0.3950	0.3910	0.3872	0.3833	0.3795	0.3757	0.3720	0.3683	0.3646
0.1	0.3610	0.3574	0.3538	0.3503	0.3468	0.3434	0.3400	0.3366	0.3332	0.3299
0.2	0.3266	0.3234	0.3202	0.3170	0.3138	0.3107	0.3076	0.3045	0.3015	0.2985
0.3	0.2955	0.2926	0.2897	0.2868	0.2840	0.2811	0.2783	0.2756	0.2728	0.2701
0.4	0.2674	0.2648	0.2621	0.2595	0.2569	0.2544	0.2518	0.2493	0.2469	0.2444
0.5	0.2420	0.2396	0.2372	0.2348	0.2325	0.2302	0.2279	0.2256	0.2234	0.2211
0.6	0.2189	0.2168	0.2146	0.2125	0.2104	0.2083	0.2062	0.2041	0.2021	0.2001
0.7	0.1981	0.1961	0.1942	0.1923	0.1903	0.1884	0.1866	0.1847	0.1829	0.1811
0.8	0.1793	0.1775	0.1757	0.1740	0.1722	0.1705	0.1688	0.1671	0.1655	0.1638
0.9	0.1622	0.1606	0.1590	0.1574	0.1558	0.1543	0.1528	0.1512	0.1497	0.1482
1.0	0.1468	0.1453	0.1439	0.1424	0.1410	0.1396	0.1382	0.1368	0.1355	0.1341
1.1	0.1328	0.1315	0.1302	0.1289	0.1276	0.1263	0.1251	0.1238	0.1226	0.1214
1.2	0.1202	0.1190	0.1178	0.1166	0.1154	0.1143	0.1132	0.1120	0.1109	0.1098
1.3	0.1087	0.1076	0.1066	0.1055	0.1045	0.1034	0.1024	0.1014	0.1004	0.0994
1.4	0.0984	0.0974	0.0964	0.0955	0.0945	0.0936	0.0926	0.0917	0.0908	0.0899
1.5	0.0890	0.0881	0.0873	0.0864	0.0855	0.0847	0.0838	0.0830	0.0822	0.0814
1.6	0.0805	0.0797	0.0790	0.0782	0.0774	0.0766	0.0759	0.0751	0.0744	0.0736
1.7	0.0729	0.0722	0.0714	0.0707	0.0700	0.0693	0.0686	0.0680	0.0673	0.0666
1.8	0.0659	0.0653	0.0646	0.0640	0.0634	0.0627	0.0621	0.0615	0.0609	0.0603
1.9	0.0597	0.0591	0.0585	0.0579	0.0573	0.0568	0.0562	0.0556	0.0551	0.0545
2.0	0.0540	0.0535	0.0529	0.0524	0.0519	0.0514	0.0508	0.0503	0.0498	0.0493
2.1	0.0489	0.0484	0.0479	0.0474	0.0469	0.0465	0.0460	0.0456	0.0451	0.0446
2.2	0.0442	0.0438	0.0433	0.0429	0.0425	0.0420	0.0416	0.0412	0.0408	0.0404
2.3	0.0400	0.0396	0.0392	0.0388	0.0384	0.0380	0.0377	0.0373	0.0369	0.0366
2.4	0.0362	0.0358	0.0355	0.0351	0.0348	0.0344	0.0341	0.0337	0.0334	0.0331
2.5	0.0327	0.0324	0.0321	0.0318	0.0315	0.0312	0.0308	0.0305	0.0302	0.0299
2.6	0.0296	0.0293	0.0290	0.0288	0.0285	0.0282	0.0279	0.0276	0.0274	0.0271
2.7	0.0268	0.0265	0.0263	0.0260	0.0258	0.0255	0.0252	0.0250	0.0247	0.0245
2.8	0.0243	0.0240	0.0238	0.0235	0.0233	0.0231	0.0228	0.0226	0.0224	0.0222
2.9	0.0220	0.0217	0.0215	0.0213	0.0211	0.0209	0.0207	0.0205	0.0203	0.0201
3.0	0.0199	0.0197	0.0195	0.0193	0.0191	0.0189	0.0187	0.0185	0.0183	0.0182
3.1	0.0180	0.0178	0.0176	0.0174	0.0173	0.0171	0.0169	0.0168	0.0166	0.0164
3.2	0.0163	0.0161	0.0159	0.0158	0.0156	0.0155	0.0153	0.0152	0.0150	0.0149
3.3	0.0147	0.0146	0.0144	0.0143	0.0141	0.0140	0.0139	0.0137	0.0136	0.0134
3.4	0.0133	0.0132	0.0131	0.0129	0.0128	0.0127	0.0125	0.0124	0.0123	0.0122
3.5	0.0120	0.0119	0.0118	0.0117	0.0116	0.0115	0.0113	0.0112	0.0111	0.0110
3.6	0.0109	0.0108	0.0107	0.0106	0.0105	0.0104	0.0103	0.0102	0.0101	0.0100
3.7	0.0099	0.0098	0.0097	0.0096	0.0095	0.0094	0.0093	0.0092	0.0091	0.0090
3.8	0.0089	0.0088	0.0087	0.0087	0.0086	0.0085	0.0084	0.0083	0.0082	0.0082
3.9	0.0081	0.0080	0.0079	0.0078	0.0078	0.0077	0.0076	0.0075	0.0075	0.0074
4.0	0.0073	0.0072	0.0072	0.0071	0.0070	0.0070	0.0069	0.0068	0.0067	0.0067
4.1	0.0066	0.0065	0.0065	0.0064	0.0064	0.0063	0.0062	0.0062	0.0061	0.0060
4.2	0.0060	0.0059	0.0059	0.0058	0.0057	0.0057	0.0056	0.0056	0.0055	0.0055
4.3	0.0054	0.0054	0.0053	0.0053	0.0052	0.0051	0.0051	0.0050	0.0050	0.0049
4.4	0.0049	0.0048	0.0048	0.0048	0.0047	0.0047	0.0046	0.0046	0.0045	0.0045

Bảng 2. GIÁ TRỊ HÀM MẬT ĐỘ PHÂN PHỐI CHUẨN HÓA (tiếp)

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
4.5	0.0044	0.0044	0.0043	0.0043	0.0043	0.0042	0.0042	0.0041	0.0041	0.0041
4.6	0.0040	0.0040	0.0039	0.0039	0.0039	0.0038	0.0038	0.0037	0.0037	0.0037
4.7	0.0036	0.0036	0.0036	0.0035	0.0035	0.0035	0.0034	0.0034	0.0033	0.0033
4.8	0.0033	0.0033	0.0032	0.0032	0.0032	0.0031	0.0031	0.0031	0.0030	0.0030
4.9	0.0030	0.0029	0.0029	0.0029	0.0029	0.0028	0.0028	0.0028	0.0027	0.0027
5.0	0.0027	0.0027	0.0026	0.0026	0.0026	0.0026	0.0025	0.0025	0.0025	0.0025
5.1	0.0024	0.0024	0.0024	0.0024	0.0023	0.0023	0.0023	0.0023	0.0022	0.0022
5.2	0.0022	0.0022	0.0022	0.0021	0.0021	0.0021	0.0021	0.0021	0.0020	0.0020
5.3	0.0020	0.0020	0.0020	0.0019	0.0019	0.0019	0.0019	0.0019	0.0018	0.0018
5.4	0.0018	0.0018	0.0018	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0016
5.5	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0015	0.0015	0.0015	0.0015
5.6	0.0015	0.0015	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0014	0.0013
5.7	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0013	0.0012	0.0012	0.0012
5.8	0.0012	0.0012	0.0012	0.0012	0.0012	0.0011	0.0011	0.0011	0.0011	0.0011
5.9	0.0011	0.0011	0.0011	0.0011	0.0011	0.0010	0.0010	0.0010	0.0010	0.0010
6.0	0.0010	0.0010	0.0010	0.0010	0.0010	0.0009	0.0009	0.0009	0.0009	0.0009
6.1	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008
6.2	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
6.3	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
6.4	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
6.5	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005
6.6	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005
6.7	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004
6.8	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
6.9	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
7.0	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
7.1	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
7.2	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
7.3	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002
7.4	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
7.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
7.6	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
7.7	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
7.8	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001
7.9	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.1	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.2	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.3	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.4	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.5	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.6	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8.9	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
9.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Bảng 3. **GIÁ TRỊ HÀM** $\Phi_0(u)=\int\limits_0^u \varphi(z)dz$ $\Phi_0(-u)=-\Phi_0(u)$

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
4.0	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Bảng 4. **GIÁ TRỊ HÀM Φ(u) = P(U < u)** $\Phi(u) = \int_{-\infty}^{u} \varphi(z)dz$

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Bảng 5. **GIÁ TRỊ XÁC SUẤT $P(U > u)$; GIÁ TRỊ TỚI HẠN CHUẨN u_α**

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Tài liệu tham khảo

[1] Đậu Thế Cấp, *Xác suất Thống kê – Lý thuyết và các bài tập*, NXB Giáo dục, 2006.

[2] Đào Hữu Hồ, *Xác suất thống kê*, NXB Đại học Quốc gia Hà Nội, 2008.

[3] Tống Đình Quỳ , *Giáo trình xác suất thống kê*, NXB Bách khoa Hà Nội, 2015.

[4] Nguyễn Cao Văn, Trần Thái Ninh, *Giáo trình lý thuyết xác suất và thống kê toán* , NXB Thống kê, 2005.

[5] F.M.Dekking, *A modern introduction to Probability and Statistics*, Springer Publication, 2005.